Novel Approaches Toward Area- and Energy-Efficient Embedded Memories

THÈSE N^{O} 6074 (2014)

PRÉSENTÉE LE 7 FÉVRIER 2014

À LA FACULTÉ DES SCIENCES ET TECHNIQUES DE L'INGÉNIEUR
LABORATOIRE DE CIRCUITS POUR TÉLÉCOMMUNICATIONS
PROGRAMME DOCTORAL EN MICROSYSTÈMES ET MICROÉLECTRONIQUE

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

POUR L'OBTENTION DU GRADE DE DOCTEUR ÈS SCIENCES

PAR

Pascal Andreas MEINERZHAGEN

acceptée sur proposition du jury:

Prof. D. Atienza Alonso, président du jury
Prof. A. P. Burg, Prof. Y. Leblebici, directeurs de thèse
Prof. C. Enz, rapporteur
Prof. A. Fish, rapporteur
Prof. J. Rodrigues, rapporteur

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Suisse
2014

Acknowledgments

First of all, I am deeply grateful to my thesis advisor Prof. Dr. Andreas Burg for his continuous support, guidance through the world of digital VLSI circuits, systems, and applications, providing an excellent research environment, providing the possibility of manufacturing a large number of test chips in various technology nodes, guidance and support to build up a chip measurement lab, and his steady encouragement to publish at and attend conferences. He was always available and willing to advise and support me, even at late hours and on weekends, face to face or remotely. I would also like to thank my thesis co-advisor Prof. Dr. Yusuf Leblebici who was available for several insightful and pathbreaking discussions and generously provided his lab facilities for chip measurements. Many thanks to both of them. Many thanks go to Prof. Dr. Joachim Rodrigues from Lund University for his initiative to use standard-cell based memories in sub-V_{T} systems, for his guidance through the world and challenges of sub-V_{T} circuit and system design, for continuing pushing the optimization of sub-V_{T} memories, providing the possibility of manufacturing a large number of chips in ST65nm CMOS, and for hosting me at Lund University several times. Many special thanks go to Prof. Dr. Alexander Fish, who not only provided me with many technical advises and insights into the fields of several emerging memory technologies and advanced adaptive bulk biasing schemes for CMOS memory arrays, but also taught me various important lessons concerning the writing of journal papers and research proposals. In addition, I am grateful to Prof. Dr. Christian Enz for kindly serving as an internal expert in my PhD defense committee.

I am truly grateful to a number of individuals I met during the first two years of my PhD curriculum at the the Integrated Systems Laboratory (IIS) at ETHZ. Prof. Dr. Hubert Käslin and Dr. Norbert Felber provided useful comments and critical feedback during review meetings of various early design ideas; in addition, Hubert provided excellent EDA and PDK support through the Microelectronics Design Center (DZ), while Norbert provided valuable guidance concerning the use of test boards and an industrial digital tester, as well as funding for the tapeout of several student projects. Dr. Frank Gürkaynak and Beat Muheim from the DZ at ETHZ provided a truly excellent EDA and PDK support without which it would have hardly been possible to manufacture and measure many of the prototype chips covered in this thesis; with their tremendous patience and kindness, they even helped me not only with the standard digital design flow, but also with many special needs like separate power domains and special power pads. Frank even helped us measure chips on their equipment at ETHZ while we were already at EPFL. Dr. Jürg Treichler advised me in many questions in the domain of analog IC design and also co-advised many of my students working on the analog aspects

of digital memory circuits; beyond that, Jürg always had the right script or advise to fix any potential problem. Dr. Christoph Studer generously provided the baseline LDPC decoder design which was used as application example for the use of standard-cell based memories in this thesis. I enjoyed many interesting discussions with Dr. Luca Henzen about cryptographic systems and their memory requirements. Christoph Roth worked with me on the integration of standard-cell based memories into his LDPC decoder architecture and, even after we left ETHZ, he was still available to support us while we evaluated the idea of using refresh-free dynamic memories in the LDPC decoder. Onur Andic pioneered the work on multilevel gain-cell memories and designed a complete multilevel GC-eDRAM array, while Markus Schulz taped out the first single-bit-per-cell gain-cell array. I would also like to thank Dario Carnelli, Schekeb Fateh, and Dr. Christian Benkeser for several interesting technical discussions. I am also grateful to many other colleagues I collaborated with at ETHZ, as well as MSc and BSc students and interns I supervised at the IIS at ETHZ. Many thanks to all of them.

I am deeply grateful to many individuals at which I met at EPFL. Dr. Alain Vachoux from the Microelectronics Systems Laboratory (LSM) provided a continuous support with EDA tools and PDKs. Many thanks go to Christian Senning as a colleague and for his excellent support and help with the IT infrastructure in our Telecommunications Systems Laboratory (TCL) at EPFL. Moreover, I would like to thank Dr. Georgios Karakonstantis for his friendship and many interesting technical discussions and collaborations. Furthermore, Jeremy Constantin, Dr. Pavle Belanovic, Alexios Balatsoukas-Stimming, and Nicholas Preyss brought more life to our lab by organizing lunch groups and many social events. I am also grateful to all MSc and BSc students as well as interns which I have advised at EPFL for the countless design tasks and simulations they have carried out. In particular, Rashid Iqbal analyzed the impact of voltage scaling on gain-cell memories and even taped out a test chip, while Muhammad Umer Khalid proposed and evaluated a replica column for fast read and write access to multilevel gain-cell memories under PVT variations. Andrea Bonetti did a truly excellent job by designing a dynamic storage cell and taping out a full LDPC decoder chip containing dynamic standard-cell based memories in a short time. Ibrahim Kazi designed non-volatile flip-flop topologies and optimized them for subthreshold operation. My special thanks also go to Dr. Pierre-Emmanuel Gaillardon from the Integrated Systems Laboratory (LSI) at EPFL and to Dr. Davide Sacchetto from LSM at EPFL for interesting and truly enriching collaborations in the field of emerging memory technologies, and the integration of such emerging devices into CMOS circuits. Moreover, I am thankful to Radisav Cojbasic, Nikola Katic, Alessandro Cevrero, and Clemens Nyffeler from LSM at EPFL for many interesting technical as well as non-technical discussions and coffee breaks. I am also grateful to Dr. Ahmed Dogan, an alumni of the Embedded Systems Laboratory (ESL) at EPFL for providing interesting applications for our sub-V_{T} memories in the field of biomedical signal processing, and for the collaboration in the field of standard-cell library characterization. Many thanks to all of them.

I am very grateful to several individuals from Lund University, Sweden which I had the chance to visit several times during my PhD curriculum. I am deeply grateful to Yasser Sherazi for an amazingly efficient and productive collaboration on the comparative analysis of sub-V_{T} standard-cell based memories, and for teaching me their sub-V_{T} characterization flow.

Moreover, I am addressing many thanks to Oskar Andersson and Babak Mohammadi for various great collaborations on the design, manufacturing, and measurement of many sub-V_T memories based on custom-designed standard-cells. Many thanks to all of them.
I am also deeply grateful to several people from Ben-Gurion University and Bar-Ilan University, Israel, which I could both visit several times during the completion of my PhD studies. I am truly and deeply grateful to Adam (Adi) Teman, who I extensively collaborated with in the field of gain-cell memories during the last years of my PhD curriculum, which resulted in many great contributions to this thesis; Adi taught me many circuit techniques for use in CMOS memory arrays, how to prioritize tasks and get work done efficiently, as well as a positive and healthy attitude toward research, teaching, networking, and life. I am also grateful to Anatoli Mordakhay who devised and measured several gain-cell test circuits. I would also like to thank Robert Giterman who helped us tremendously with the measurement of various gain-cell eDRAM test chips and who continued research into low-power gain-cell memories. Many thanks to all of them.
In addition, I would like to truly thank a number of individuals from Intel Labs, Intel Corporation who, in numerous truly enriching discussions provided me with a fresh view on my PhD work from and industry perspective, and helped me to identify future industry-relevant research directions. First of all, I would like to thank my direct advisor Dr. Jaydeep Kulkarni, who dedicated an amazing amount of his time to me (up to three 1:1 meetings per week), and from whom I learned tons in the field of analog and digital IC design, particularly in the field of power management and SRAM. Second, I would like to thank my manager James (Jim) Tschanz, who, in many personal meetings, helped me understand the relevance of my research in a larger context and taught me tons on the activities and the structure of Intel. Furthermore, I am deeply grateful to Dr. Vivek De for the unique possibility to carry out a truly enriching and unforgettable internship in his Circuit Research Laboratory (CRL). In addition, I would like to thank Dr. Ulrich Bretthauer who served as my Intel mentor and showed me the Intel Braunschweig Laboratories. Moreover, I would like to thank Dr. Dinesh Somasekhar, Dr. Muhammad Khellah, Dr. Badarinath Kommandur, and Dr. Anant Deval, for various technical discussions and valuable feedback on our work. Last but not least, I thank my colleagues Dr. Amin Khajeh, Alicia Klinefelter, Rangharajan Venkatesan, and Farah Yahya for the many interesting, technical discussions and social activities.
Finally, I would like to truly and deeply thank my parents Andreas and Anita Meinerzhagen, as well as my sisters Manuela Bregy-Meinerzhagen and Sarah Meinerzhagen for their continuous support during many years; they often helped me remember what really matters in life, beyond work, and helped me gain perspectives for future professional development. I truly enjoyed playing with my godson Loan Bregy, and of course also with Jael and Enea, as a delighting relief and change from my work and everyday life. Many thanks to all of them. I am also truly and deeply grateful to my girlfriend Maricel Montezuma for her continuous support, her unconditional love, the many nice moments she gifted me, and for always being here for me when I needed her. Many thanks to her.

Lausanne, January 15, 2014 Pascal Meinerzhagen

Abstract

Embedded memories consume an increasingly dominant share of the overall area and power of very large scale integration (VLSI) systems-on-chip (SoCs) targeted toward applications ranging from microprocessors, to wireless communications, to biomedical implants. Static random-access memory (SRAM) is the predominant embedded memory technology used in most VLSI SoCs, while conventional embedded dynamic-random access memory (eDRAM) is sometimes used for higher storage density. Unfortunately, SRAM encounters several design challenges when operated at ultra-low supply voltages or if implemented in aggressively scaled complementary metal-oxide-semiconductor (CMOS) technologies, while conventional eDRAM based on the 1-transistor-1-capacitor (1T-1C) bitcell is incompatible with standard digital CMOS technologies. This thesis investigates and proposes interesting alternatives to SRAMs and eDRAMs for the implementation of embedded memories, namely standard-cell based memories (SCMs) and gain-cell based eDRAM (GC-eDRAM).

SCMs can be synthesized from commercial standard-cell libraries (SCLs) and function reliably in any VLSI system, even if operated at ultra-low voltages or when implemented in aggressively scaled CMOS nodes, where conventional 6-transistor (6T)-bitcell SRAM would fail. This thesis presents an extensive comparative analysis of possible SCM topologies based on commercial SCLs and identifies the border in storage capacity up to which SCMs are still smaller than SRAM macrocells, despite the larger storage cell (latch or flip-flop), due to less peripheral circuits. In addition, the enormous benefits of the design and integration of custom standard-cells to meet the specific needs of various VLSI SoCs with very different memory requirements are demonstrated and verified by various application examples and the manufacturing and measurement of several test chips. For example, all internal memories of a low-density parity-check (LDPC) decoder, extensively used in wireless communications, can be implemented as refresh-free, dynamic SCMs (D-SCMs) due to frequent and periodic write updates; the use of custom-designed dynamic latches instead of commercial static latches leads to dramatic area savings. Moreover, subthreshold (sub-V_T) SCMs are especially interesting for ultra-low power VLSI systems such as biomedical implants due to the lack of good sub-V_T SRAM macrocell compilers; silicon measurements show that the design of a single ultra-low leakage standard-cell and its integration into the SCM compilation flow lead to unprecedentedly low leakage power and access energy per bit. Finally, a non-volatile flip-flop topology, based on emerging ReRAM device technology, which can operate and wake-up at sub-V_T voltages is proposed for future low-power VLSI SoCs with zero standby leakage.

GC-eDRAM is an interesting alternative to both SRAM and conventional 1T-1C eDRAM, since

it combines the main advantages of both SRAM and eDRAM, while it avoids most of their drawbacks. In fact, a gain-cell, built from 2–4 MOS transistors, is smaller than any SRAM bitcell and exhibits less leakage current, while it is fully compatible with standard digital CMOS technology, and allows for non-destructive read (as opposed to 1T-1C eDRAM). Moreover, any gain-cell can simultaneously and independently be optimized for robust read and write access (as opposed to both 6T SRAM and 1T-1C eDRAM) and allows for two-port memory implementations at virtually no overhead compared to single-port implementations. The main drawback of GC-eDRAM is the degraded retention time compared to 1T-1C eDRAM and the need for periodic, power-consuming refresh cycles. In this thesis, the impact of supply voltage scaling on the behavior of 2-transistor (2T)-bitcell GC-eDRAM is analyzed in detail; counter to intuition, the retention time of GC-eDRAM can be improved by voltage scaling for given memory access statistics and a given write bit-line (WBL) control scheme, identifying near-threshold (near-V_T) GC-eDRAMs as an interesting and feasible memory type for use in low-power, medium-performance VLSI SoCs. Furthermore, two novel techniques to further improve the retention time and reduce the data retention power of near-V_T GC-eDRAM are proposed and verified by silicon measurements: 1) reverse body biasing (RBB) of the storage array for reduced subthreshold conduction of the write transistor; and 2) replica techniques for optimum refresh timing under varying environmental conditions (process-voltage-temperature) and for varying write-access disturb frequencies. Moreover, as a high-density counterpart to large 8–14 transistor (8–14T) sub-V_T SRAM bitcells, the feasibility of sub-V_T GC-eDRAM is investigated for the first time; we find that sub-V_T operation is a viable option leading to sufficiently high array availability for read and write access in a mature CMOS node, while we recommend near-V_T operation in aggressively scaled nodes due to increased parametric variations and lower achievable storage node capacitance. Finally, the feasibility of multilevel gain-cells is investigated for the first time; such multilevel GC-eDRAM is identified as convenient means to trade circuit reliability for the benefit of higher storage density in error-resilient VLSI systems (such as many wireless communications systems).

Keywords: Embedded memories, VLSI systems, SoC, ASIC, CMOS, SRAM, eDRAM, standard-cell based memory, dynamic storage cells, voltage scaling, near-threshold operation, sub-threshold operation, ultra-low power, reliability, non-volatile memory, ReRAM, OxRAM, gain-cells, gain-cell based eDRAM (GC-eDRAM), retention time improvement, refresh power reduction, body biasing, replica techniques, technology scaling, multilevel gain-cell

Zusammenfassung

Integrierte Datenspeicherbausteine verbrauchen einen stetig wachsenden Anteil des Flächenbedarfs und des gesamten Energieverbrauchs von VLSI Systemen (SoCs) welche in Mikroprozessoren, drahtlosen Kommunikationssystemen, biomedizinischen Implantaten und für viele andere Anwendungen gebraucht werden. Die meisten dieser VLSI Systemen bedienen sich der dominanten und meist genutzten SRAM Speichertechnologie, welche nur selten durch konventionelle eDRAM Technologie ersetzt wird um höhere Speicherdichten zu erreichen. Leider ist es problematisch SRAM Speichereinheiten zuverlässig mit tiefen Versorgungsspannungen zu betreiben oder in den modernsten, extrem skalierten CMOS Technologieprozessen zu implementieren. Zudem ist die konventionelle eDRAM Technologie, welche auf der 1-Transistor-1-Kondensator (1T-1C) Speicherzelle beruht, nicht gänzlich kompatibel mit normalen, digitalen CMOS Technologien. Diese Dissertation untersucht Speicherbausteine basierend auf Standardzellen (SCMs) und eDRAM basierend auf so gennanten “Gain-Cells” (GC-eDRAM) als vielversprechende Alternativen zu den konventionellen SRAM und 1T-1C eDRAM Technologien und schlägt viele konkrete Implementierungen in verschiedenen CMOS Technologien vor.

In der Tat können SCMs mit Hilfe von kommerziell zugänglichen Standardzellenbibliotheken (SCLs) einfach synthetisiert und in einem beliebigen VLSI System zuverlässig in Betrieb genommen werden, sogar bei extrem tiefen Versorgungsspannungen oder in stark skalierten CMOS Technologien wo konventionelles SRAM (basierend auf der 6T-Speicherzelle) normalerweise nicht mehr zuverlässig funktionieren würde. Diese Dissertation präsentiert eine detaillierte Studie und einen umfangreichen Vergleich von vielen möglichen SCM Topologien welche auf kommerziellen SCLs basieren; ausserdem wird genau untersucht, bis zu welcher Speicherkapazität SCMs flächenmassig noch kleiner sind als SRAM Speichereinheiten, trotz der grösseren Speicherzelle (bistabile Kippschaltung anstelle der 6T SRAM-Zelle) und dank weniger Peripherieschaltungen. Zudem wird anhand von verschiedenen Anwendungsbeispielen und durch das Ausmessen von mehreren fabrizierten Mikrochips aufgezeigt, wie die Spezialanfertigung und Integration von eigens entwickelten Standardzellen gezielt die teilweise sehr unterschiedlichen Speicherbedürfnisse verschiedener VLSI SoCs befriedigen können und eine bestimmte Kennzahl (wie etwa die Siliziumfläche oder den Energieverbrauch) massgebend verbessern können. Beispielsweise können alle internen Speicherelemente von einem LDPC Dekoder, ein Bauteil welches oft in der drahtlosen Kommunikation gebraucht wird, als dynamische SCMs (D-SCMs) implementiert werden–sogar ohne die übliche, periodische, energieverbrauchende Refresh-Operation–, dank der häufigen und periodischen Schreibzugriffen. Die so eingesetz-

ten, eigens dafür entwickelten dynamischen Speicherzellen führen zu einer signifikanten Reduktion des Flächenbedarfs im Vergleich zu den kommerziellen, statischen Speicherzellen. Ein weiteres Anwendungsbeispiel sind VLSI Systeme mit extrem geringem Energieverbrauch ("ultra-low power VLSI SoCs") wie etwa biomedizinische Implantate, wo der Einsatz von zuverlässigen sub-V_T SCMs besonders interessant und praktisch ist, da es keine guten Kompiler für sub-V_T SRAM Makrozellen gibt[1]. Messungen von eigens dafür hergestelleten Mikrochips zeigen, dass die Entwicklung und Integration von einer einzigen Standardzelle gekennzeichnet durch einen sehr tiefen Leckstrom, zu der tiefsten jemals in einer 65 nm CMOS Technologie gemessenen Leistungsaufnahme im Standby-Modus und zum tiefsten Energieverbrauch für Lese- und Schreibzugriffe führen. Schlussendlich werden in dieser Dissertation auch zum ersten Mal Flip-Flops vorgestellt welche mit sub-V_T Versorgungsspannungen auskommen und dank neuartigen ReRAM Speicherelementen ihre Daten sogar nach der Entfernung der Versorgunsspannung beibehalten, wodurch zukünftige VLSI Systeme mit bereits geringem Energieverbrauch sogar in einen Standby-Modus ganz ohne Stromverbrauch versetzt werden können.

Die zweite in dieser Dissertation untersuchte Art von Speichertechnologien, namentlich GC-eDRAM, kombiniert die meisten Vorteile von SRAM und konventioneller eDRAM Technologie, während die meisten Nachteile von diesen konventionellen Technologien vermieden werden. In der Tat besteht eine "Gain-Cell" (GC) aus 2–4 MOS Transistoren, ist damit wesentlich kleiner und hat weniger Leckströme als alle bekannten SRAM Speicherzellen, kann direkt in jeder digitalen CMOS Technologie gebaut werden (ohne zusätzliche Prozessschritte) und hat einen nicht-destruktiven Lesezugriff (im Gegensatz zu der 1T-1C eDRAM Technologie). Zudem kann jede GC gleichzeitig und unabhängig für zuverlässige Lese- und Schreibzugriffe optimiert werden (was bei 6T SRAM und 1T-1C eDRAM Speicherzellen nicht möglich ist) und erlaubt auch das Bauen von Speichermakrozellen mit einem separaten Lese- und Schreibzugang, welche nur unwesentlich grösser sind als Makrozellen mit einem einzigen Zugang. Der bedeutendste Nachteil von GC-eDRAMs ist die kurze Datenspeicherzeit verglichen mit 1T-1C eDRAM und die daraus folgenden, frequenten, periodischen, energieverbrauchenden Refresh-Operationen. Diese Dissertation präsentiert Forschungsergebnisse, welche den Einfluss einer verringerten Versorgungsspannung auf das Verhalten von GC-eDRAM, basierend auf einer 2T Speicherzelle, aufzeigen; entgegen allen Erwartungen kann die Datenspeicherzeit durch eine Verringerung der Versorgungsspannung gesteigert werden, falls kritische Schaltungsknoten, namentlich die "write bit-lines" (WBLs), dank seltenen Lesezugriffen gezielt kontrolliert werden können. Diese Analyse zeigt, dass near-V_T GC-eDRAM eine interessante und durchaus realisierbare Speichertechnologie für energie-effiziente VLSI Systeme mit mittelmässig hohem Datendurchsatz darstellen[2]. Des Weiteren werden zwei neuartige Methoden vorgeschlagen und durch Messungen von entsprechenden Mikrochips bestätigt, um die Datenspeicherzeiten von near-V_T GC-eDRAM weiter zu verlängern. Erstens wird gezeigt, dass "reverse body biasing"

[1] Der Begriff "sub-V_T" bezieht sich auf extrem tiefe Versorgungsspannungen, welche unter der Schwellenspannung (V_T) der Transistoren liegen.

[2] Der Begriff "near-V_T" bezieht sich auf tiefe Versorgungsspannungen, welche nur leicht über der Schwellenspannung (V_T) der Transistoren liegen.

(RBB) den unerwünschten Leckstrom durch den Lesezugrifftransistoren der GC reduziert. Zweitens kann der ideale Zeitpunkt für eine Refresh-Operation durch eine Replika-Technik bestimmt werden, sogar bei Prozess-, Spannungs- und Temperaturvariationen und für unterschiedlich häufige auftretende Störungen durch Lesezugriffe. Des Weiteren wird zum ersten Mal die Machbarkeit von sub-V_T GC-eDRAM untersucht, welcher mit nur 2 Transistoren pro Speicherzelle eine bedeutend höhere Speicherdichte aufweisen kann als sub-V_T SRAM Zellen, welche auf 8–14 Transistoren basieren. Unsere Analysen zeigen, dass GC-eDRAM im sub-V_T Bereich betrieben werden kann falls auf ältere CMOS Technologien zurückgegriffen wird, während die Versorgungsspannung nur bis in den near-V_T Bereich verringert werden sollte für die modernsten, stark skalierten CMOS Technologien um eine genügend hohe Verfügbarkeit für Lese- und Schreibzugriffe zu erreichen. Schlussendlich untersucht diese Dissertation zum ersten Mal die Machbarkeit von GC-eDRAMs, welche mehrere Bits pro Zelle speichern ("multilevel GC-eDRAM"). Es wird aufgezeigt dass solche multilevel GC-eDRAMs eine angebrachte Speichertechnologie darstellen um höhere Speicherdichten zu erreichen in fehlertoleranten VLSI Systemen (wie zum Beispiel drahtlose Kommunikationssysteme), welche eine kleine Anzahl von Schaltungsfehlern tolerieren können.

Schlüsselwörter: Integrierte Speicher, VLSI Systeme, SoC, ASIC, CMOS, SRAM, eDRAM, Standardzellenspeicher (SCM), dynamische Speicherzellen, Spannungsreduktion, near-threshold Operation, subthreshold Operation, ultra-tiefer Energieverbrauch, Zuverlässigkeit, nichtflüchtige Speicher, ReRAM, OxRAM, "gain-cells", GC-eDRAM, Datenspeicherzeiterhöhung, Reduktion der Datenspeicherleistungsaufnahme, "body biasing", Replikatechnik, Technologieskalierung, mehrstufige gain-cell

Résumé

Les mémoires embarquées consomment une part de plus en plus importante de la surface totale et de la consommation des systèmes sur puces (System-on-Chip SoC) VLSI (Very Large Scale Integration) au sein d'un large domaine d'applications telles que les microprocesseurs, les systèmes de communications sans fil ou encore les implants biomédicaux. La technologie mémoire prédominante dans la plupart des systèmes VLSI est la SRAM, tandis que la technologie eDRAM conventionnelle s'utilise quelquefois pour atteindre des densités de stockage plus élevées. Malheureusement, la technologie SRAM se voit confrontée à plusieurs défis en cas d'opération à des tensions d'alimentation très basse et/ou de sa réalisation dans des technologies CMOS très avancées, alors que la technologie eDRAM conventionnelle basée sur la cellule 1-transistor-1-condensateur (1T-1C) n'est pas entièrement compatible avec les technologies CMOS numériques standards. Cette thèse de doctorat analyse et propose des nouvelles technologies pour l'implémentation des mémoires embarquées, avec notamment des mémoires à cellules de standard (SCMs) et des mémoires dynamiques basées sur des cellules à gain (GC-eDRAM).

Les mémoires SCM peuvent être synthétisées à partir de bibliothèques de cellules standard (Standard Cell Libraries SCLs) commerciales et fonctionnent de manière fiable dans tous les systèmes VLSI, même à des tension d'alimentation très basses et dans les technologie CMOS les plus avancées, où les mémoires SRAM conventionnelles, s'appuyant sur la cellule à 6 transistors, cessent de fonctionner correctement. Cette thèse de doctorat présente une analyse comparative approfondie des topologies SCM basées sur des SCLs commerciales et identifie les limites en terme de capacité de stockage pour laquelle les SCMs présentent un gain en surface par rapport à un équivalent SRAM. Bien que la cellule élémentaire soit plus grande (bascule bistable au lieu de la cellule SRAM 6T), le gain en vient de la réduction des besoins en circuits périphériques. En outre, plusieurs exemples d'application ainsi que la fabrication et des mesures de plusieurs puces de prototype montrent les avantages énormes qui résultent de la conception et de l'intégration des cellules standard faites sur mesure afin de répondre aux besoins spécifiques de différentes classes de systèmes VLSI. Par exemple, toutes les mémoires internes d'un décodeur LDPC (un dispositif qui est fréquemment utilisé dans les systèmes de communication sans fil), peuvent être implémentées comme des SCMs dynamiques (D-SCMs) grâce aux accès d'écriture fréquents et périodiques. En effet, l'utilisation des cellules de stockage dynamiques faites sur mesure donne lieu à une remarquable réduction de surface en comparaison de l'utilisation des cellules de stockage statiques commerciales. De plus, les SCM travaillant sous le seuil (sub-V_T) sont particulièrement intéressantes pour les systèmes VLSI de

très faible puissance (ultra-low power VLSI systems) tels que les implants biomédicaux puisque de bons compilateurs de mémoires SRAM sub-V_T ne sont normalement pas disponibles. En effet, les mesures sur prototypes montrent que la conception et l'intégration dans la procédure de compilation SCM d'une seule cellule standard caractérisée par un courant de fuite très bas offrent une consommation au repas (stand-by) et l'énergie d'accès normalisée les plus basses jamais mesurées dans une technologie CMOS 65nm. Finalement, cette thèse de doctorat propose une nouvelle topologie de bascule bistable rémanente (non volatile), basée sur une technologie ReRAM émergente, qui peut être alimentée par une tension très faible (dans le domaine sub-V_T) pour toutes les opérations régulières (sauf écriture de la partie rémanente). Cette bascule bistable rémanente permettra des modes stand-by sans aucun courant de fuite dans les futurs systèmes VLSI.

La technologie de mémoire GC-eDRAM proposée unit les avantages principaux des technologies SRAM et eDRAM conventionnelles, tout en évitant la plupart de leurs inconvénients. En fait, une cellule à gain (Gain-Cell GC), construite avec 2–4 transistors MOS, est plus compacte et présente un courant de fuite plus faible que n'importe quelle cellule SRAM, tandis qu'elle est entièrement compatible avec les technologies CMOS numériques standards et permet des accès en lecture non destructifs (ce qui n'est pas le cas pour la technologie eDRAM conventionnelle). De plus, il est possible d'optimiser une cellule à gain indépendamment pour des accès en lecture et des accès en écriture fiables en même temps, ce qui n'est pas possible ni pour les cellules SRAM, ni pour les cellules eDRAM 1T-1C conventionnelles. Aussi, les cellules à gain permettent de facilement construire des mémoires à double ports avec un très faible surcoût en surface par rapport à des mémoires à port unique. Toutefois, le désavantage principal des mémoires GC-eDRAM est le temps de rétention de données réduit par rapport aux mémoires eDRAM conventionnelles et donc la nécessité de cycles de rafraîchissement (refresh cycles) périodiques consommant de l'énergie. Dans cette thèse de doctorat, le comportement des GC-eDRAM, basée sur une cellule de stockage à deux transistors, est analysé en détail dans des conditions d'utilisation à faible tensions d'alimentation : contrairement à toutes attentes, le temps de rétention de données peut être augmenté par une réduction de la tension d'alimentation, au cas où quelques noeuds particuliers (notamment les write bit-lines WBLs) peuvent être librement contrôlés grâce à des accès en écriture peu fréquents. Cette analyse fait émerger les GC-eDRAMs near-V_T comme un type de mémoire pertinent pour les systèmes VLSI à faible consommation et débit de données moyen. En outre, deux nouvelles techniques pour améliorer encore plus le temps de rétention et pour réduire la consommation des GC-eDRAMs near-V_T sont proposées et vérifiées par mesure de puces de prototype : 1) la technique de "reverse body biasing" (RBB) réduit le courant de fuite du transistor à accès en écriture avec succès ; et 2) l'utilisation de cellules répliques permet de trouver l'instant idéal pour les cycles de rafraîchissement même pour des circonstances environnementales (procédés de fabrications, tension, et température) fluctuantes et sous influence de différentes fréquences de perturbation par accès en écriture. Par ailleurs, cette thèse de doctorat étudie pour la première fois la possibilité d'alimenter des GC-eDRAM avec des tensions ultra basses (se trouvant dans le domaine sub-V_T), afin de proposer une alternative aux cellules SRAM sub-V_T avec 8–14 transistors, pour les densité de stockage élevées. Nous constatons que l'alimentation avec des

tensions sub-V_T est possible pour implémentation des GC-eDRAMs dans des technologies CMOS mûres et entraîne une disponibilité suffisante pour les accès à la mémoire, tandis que nous recommandons des tensions d'alimentation se trouvant dans le domaine near-V_T pour l'implémentation des GC-eDRAMs dans les technologies CMOS les plus avancées. Finalement, la faisabilité des cellules à gain à multiples niveaux (multilevel gain-cell) est évaluée pour la première fois ; ce genre de mémoire est identifié comme une solution optimale pour augmenter la densité de stockage, au prix d'une fiabilité plus basse. Cette perte de fiabilité est acceptable dans les systèmes VLSI naturellement résistants à quelques erreurs matérielles (comme, par exemple, beaucoup de systèmes de communication sans fil).

Mots-clés : Mémoires intégrées, systèmes VLSI, SoC, ASIC, CMOS, SRAM, eDRAM, mémoire à cellules standards (SCM), cellules de stockage dynamique, reduction de tension d'alimentation, opération near-threshold, opération subthreshold, consommation ultra-basse, fiabilité, mémoire rémanente (non volatile), ReRAM, OxRAM, cellule à gain ("gain-cell"), GC-eDRAM, amélioration du temps de rétention de données, réduction de la puissance de rafraîchissement, "body biasing", techniques de répliques, réduction de technologie, cellule de gain de différents niveaux

Contents

List of Figures

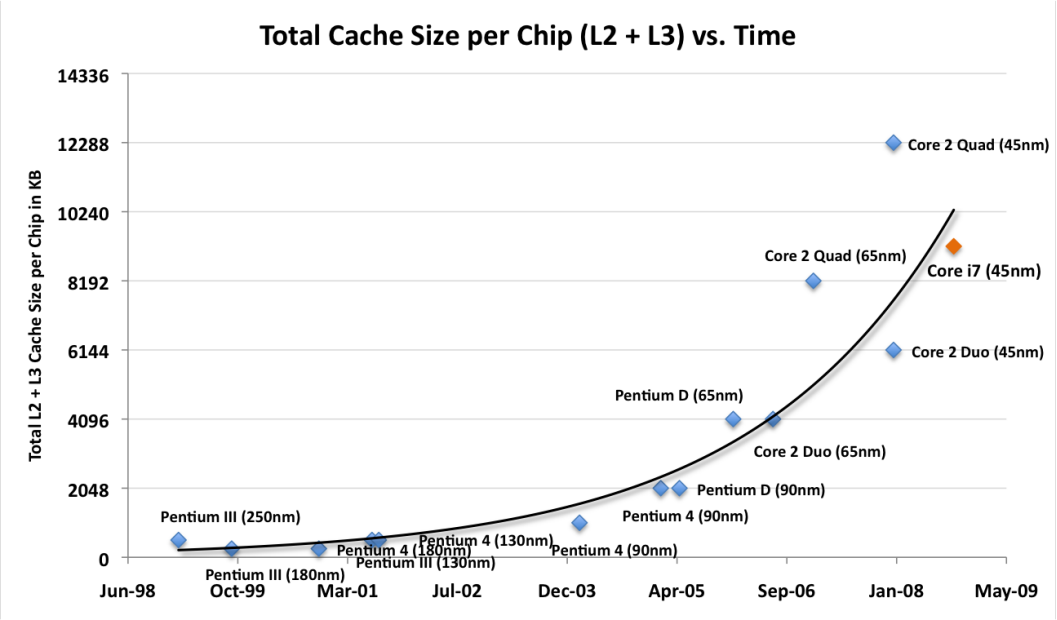
List of Tables

1 Introduction

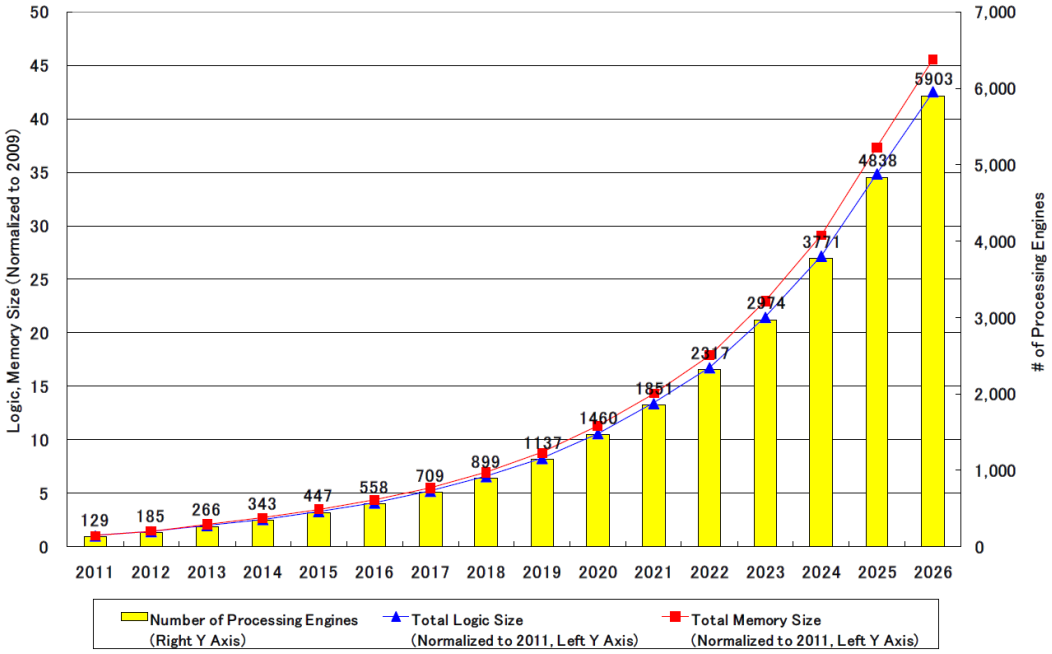
1.1 Increasing Need for Embedded Memories in VLSI SoCs

There is a steadily increasing need for *embedded memories* in very large scale integration (VLSI) system-on-chip (SoC) designs targeted toward microprocessors (for servers; personal computers; laptop computers; tablets; and smartphones); biomedical implants; wireless communications systems; and many other applications. Such embedded memories are required to temporarily store data and/or instructions. From a system level perspective, it is clearly advantageous to have always more memories embedded on-chip rather than relying on external memory chips due to a number of reasons: 1) embedded memories allow higher system-level integration densities; and 2) going off-chip through I/O pads and capacitive lines on printed circuit boards (PCBs) entails severe speed and power penalties compared to on-chip connections [8]. As shown in Fig. 1.1a, the total cache size requirement in microprocessors has increased by around 5× in a time interval as short as 4 years: back in 2005, an Intel® Pentium® D microprocessor used around 2 MB of cache memory, while the Intel® Core™ i7 released in 2009 requires almost 10 MB of cache memory [9]. In accordance with this past, quickly increasing demand for embedded memories, the International Technology Roadmap for Semiconductors (ITRS) predicted in its 2011 Edition that the total embedded memory size for general SoC applications will increase by almost 50× over the next 15 years [1], as shown in Fig. 1.1b.

Already nowadays, embedded memories consume around or even more than 50% of the total area and power budget of a VLSI SoC [1]. Fig. 1.2 illustrates this showing the layout pictures or the chip microphotographs of various VLSI systems, ranging from high-end microprocessors, to wireless communications systems, to ultra-low power (sub-V_T) microprocessors for health monitoring: the embedded memories, in form of static random-access memory (SRAM) macrocells, are visible as regular tiles. Especially in case of the sub-V_T microprocessor shown in Fig. 1.2d, the embedded memories, visible as yellow tiles, consume a dominant area share compared to the logic core which is in the center of the chip. Also the 4-stream 802.11n baseband transceiver [12], whose chip microphotograph is shown in Fig. 1.2c, contains a large



(a) Evolution of total cache size in microprocessors since 1998 [9].

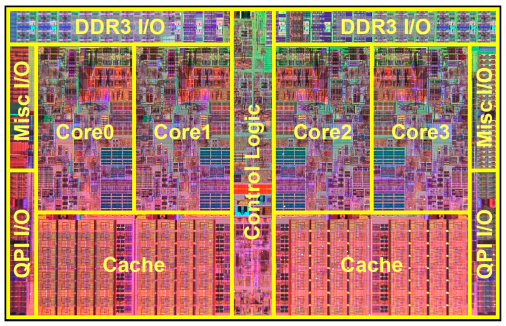


(b) Predicted evolution of total memory size in SoCs [1].

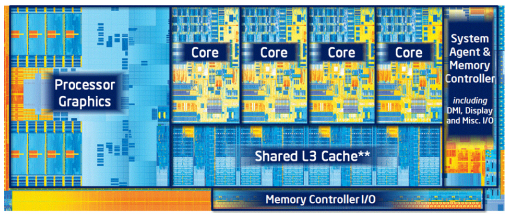
Figure 1.1: (a) Past; and (b) predicted future evolution of embedded memory size.

number of SRAM macrocells which are highlighted as dark areas.

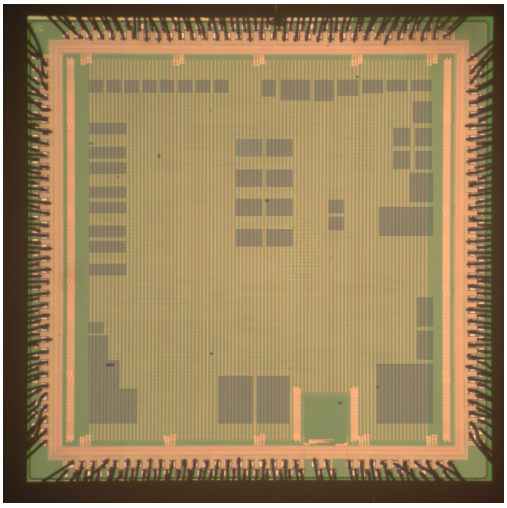
Beside the large area share, embedded memories are also responsible for a large power share of most VLSI SoCs. For example, the embedded memories of TamaRISC-CS, an ultra-low power application-specific processor for compressed sensing [13], consume 70–95% of the



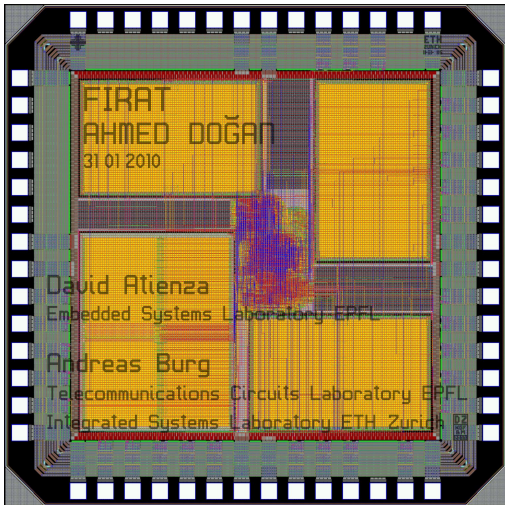
(a) Layout picture of 45 nm Intel® Core™ i7 processor (Nehalem) [10].



(b) Layout picture of 22 nm Intel® processor (a multi-CPU and GPU SoC) codenamed Ivy Bridge [11].



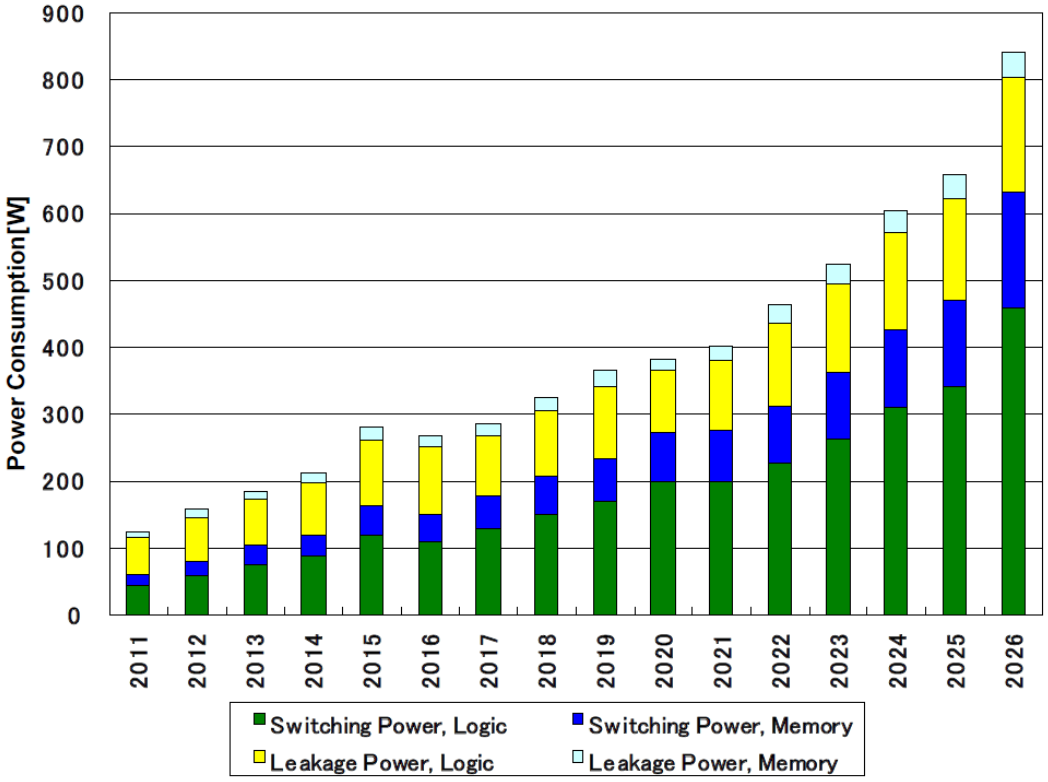
(c) Chip microphotograph of 4-Stream 802.11n baseband transceiver [12].



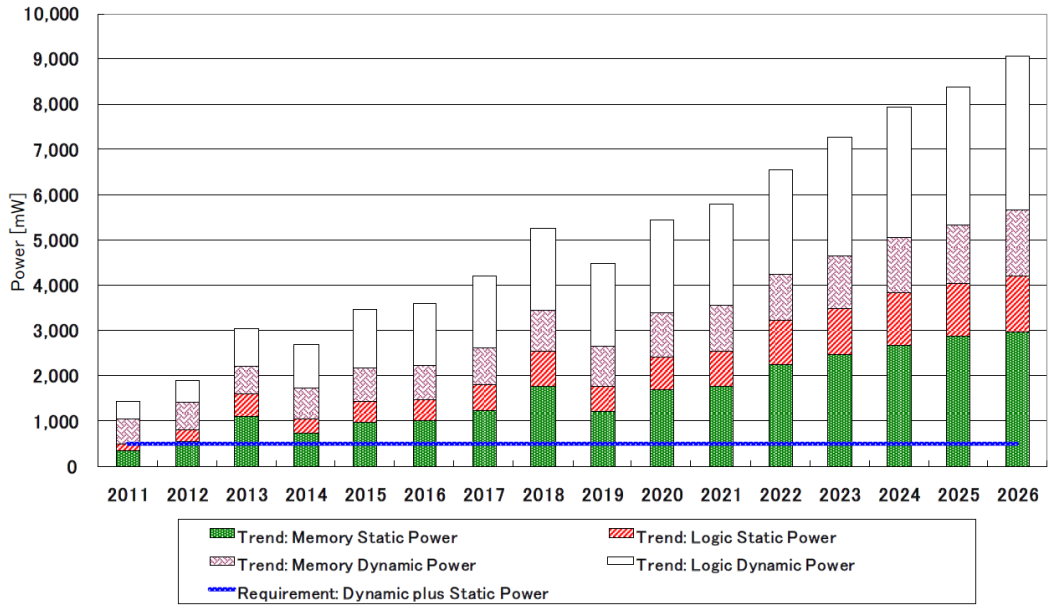
(d) Layout picture of an ultra-low power, sub-V_T microprocessor for biomedical applications.

Figure 1.2: Layout pictures and/or chip microphotographs of high-end microprocessors (a–b), a baseband transceiver (c), and a low-power processor for biomedical signals (d). All these VLSI SoCs require a significant amount of embedded memories, which are visible as regular tiles in the layout.

total power, depending on the mode of operation. As a further example, in a configurable high-throughput decoder for quasi-cyclic LDPC codes [14], the embedded memories are responsible for 68% of the total power consumption. Also, as illustrated in Fig. 1.3a, as of today, VLSI SoCs for stationary applications typically have a total power consumption of up to 100 W, corresponding to the total of dynamic and static power consumptions of logic and embedded memories [1]. As opposed to this, Fig. 1.3b shows that VLSI SoC processors for portable applications have a considerably lower total power budget of 0.5 W, a requirement established by the ITRS in 2009. Especially for portable applications, the power consumption of embedded memories is expected to further increase and consume almost 50% of the total power budget of processors in the next 15 years (see Fig. 1.3b). Reducing the power consumption of embedded memories is of utmost importance for all applications, for example



(a) Power breakdown of stationary consumer SoCs [1].



(b) Power breakdown of portable consumer SoCs [1].

Figure 1.3: Predicted power breakdowns of VLSI SoCs for (a) stationary; and (b) portable consumer electronics [1].

to ensure runtimes of several years for ultra-low power systems such as biomedical implants, to continue ensuring runtimes of one day for always more complex portable computing devices (including smartphones and tablet computers), or to reduce cooling costs for server and data center applications [15].

In addition to consuming dominant area and power shares of VLSI SoCs, embedded memories are normally the first point of failure under voltage and technology down-scaling, due to the extremely high replication count of the same basic bitcell (the 6-transistor SRAM bitcell in most cases). For example, if the supply voltage (V_{DD}) is scaled from its nominal value to the near-threshold (near-V_T) domain, the functional failure rate of embedded memories increases by 5 orders of magnitude [15]. As a consequence, under voltage and technology scaling, embedded memories typically limit the overall manufacturing yield of VLSI SoCs.

1.2 Memory Requirements of Various VLSI Systems

Conventional personal computers and servers exhibit a deep memory hierarchy, ranging from on-chip, ultra-high speed, low storage capacity register files and cache memories, to off-chip, fast, larger capacity random-access memory (RAM), to off-chip, large capacity, non-volatile data storage. Traversing this memory hierarchy, the predominant, mainstream memory technologies are: 1) distributed flip-flops and latches; 2) 6-transistor (6T)-bitcell SRAM; 3) external, conventional 1-transistor-1-capacitor (1T-1C) dynamic random-access memory (DRAM); 4) Flash memory using a floating-gate transistor as bitcell; and 5) mechanical hard disk drives, which are currently being replaced more and more with solid-state drives. Note that only the register files and cache memories are embedded within the microprocessor chip, while the remaining part of the computer memory hierarchy is off-chip. Beside personal computers, laptop computers, and servers, battery-powered mobile computing devices such as smartphones and tablet computers impose extremely challenging requirements on embedded memory solutions due to the increasing power awareness (to extend the runtime on a single battery charge) accompanied by an ever increasing demand for higher integration density and higher speed performance.

In addition to microprocessors for computers, a large number of target applications in the broad field of VLSI SoCs often have diametrically opposite requirements on embedded memories, compared to each other, as shown in Table 1.1. For example, on the one hand, embedded memories for ultra-low power (ULP) VLSI SoCs for biomedical or remote sensing applications (such as [16, 17]) require ultra-low leakage power and access energy and entail significant engineering effort to ensure high robustness, while the area and speed are only secondary concerns. Therefore, such ULP VLSI systems, including their embedded memories, are often operated at ultra-low voltages (ULV), typically residing in the subthreshold (sub-V_T) domain. On the other hand, power-aware high-performance VLSI SoCs often used in wireless communications (e.g., channel decoders) or in smartphones need high-capacity, high-density, high-speed embedded memories operated at nominal supply voltages. Rather than using robust, upsized SRAM bit-

cells, to cope with manufacturing defects (such as shorts and opens), one-time programmable address decoders, if desired in combination with spare rows or columns to maintain the storage capacity, are commonly used [18]. Moreover, to cope with soft errors (for example caused by alpha-particle impacts), redundant memory cells in conjunction with error detection and correction codes are often employed, a prominent example being the single-error-correction-double-error-detection (SECDED) code [19, 20]. Furthermore, as a new research direction, people have recently started to argue that the memory reliability can even be deliberately relaxed for VLSI systems which are inherently resilient to a small number of hardware defects. Examples of such inherently error-resilient systems include high-speed packet access (HSPA) systems [21] and wireless body sensor network (WBSN) nodes [22]. Moreover, an increasing number of VLSI systems supports dynamic voltage and frequency scaling (DVFS), in order to support different operating modes according to varying workloads, and/or reduce voltage and frequency guardbands for improved energy-efficiency and speed performance, respectively. Such systems employing DVFS ideally contain embedded memories which are fully functional over the same voltage and frequency ranges. Besides the well-known Razor [23] technique, as a further prominent example in the category of power-aware, high-performance VLSI SoCs supporting DVFS, Intel has presented an experimental, error-resilient processor (codenamed Palisades) which has built-in mechanisms to detect and correct timing errors, allowing higher performance (by means of over-clocking) or higher energy-efficiency (by means of voltage scaling) than a traditional processor with frequency and voltage guardbands, while it still computes correctly [24]. In between the two extreme categories of ultra-low power VLSI SoCs operating in the sub-V_T domain and high-performance, power-aware, potentially error-resilient VLSI SoCs operating at nominal voltage, there is a third class corresponding to low-power, medium-performance SoCs (see Table 1.1). These SoCs and their embedded memories are typically operated at near-threshold (near-V_T) supply voltages. Near-threshold computing (NTC) retains much of the energy savings of sub-V_T operation but has much more favorable performance and variability characteristics [15]. An experimental, near-threshold voltage IA-32 microprocessor is able to successfully boot Windows XP™ while being supplied from a small solar panel providing only 10−20 mW of power [25, 26]. As a further example, Diet SODA [27] is a power-efficient processor for digital cameras relying on near-threshold circuit operation.

1.3 Brief Review of the State of the Art

Broadly speaking, embedded memories can be classified into two main categories: 1) SRAM; and 2) embedded DRAM (eDRAM). SRAM uses a cross-coupled inverter pair to retain the stored data indefinitely (as long as a power supply voltage is provided). The eDRAM technology stores data in form of electric charge on a capacitor; unfortunately, the stored data is compromised due to leakage currents, which requires a periodic refresh operation.

As shown in Table 1.1, latches and flip-flops (mostly implemented as static storage cells) are commonly used as pipeline registers or in small, distributed, synthesized storage arrays within

Table 1.1: Memory requirements of different classes of VLSI SoCs, from ultra-low power to power-aware, high-performance systems.

<table>
<tr><th></th><th>Ultra-low power</th><th>Low-power, medium-performance</th><th>Power-aware, high-performance</th></tr>
<tr><td>Application fields</td><td>Biomedical implants, remote sensors</td><td>Near-threshold computing, complex sensor nodes, simple handheld devices</td><td>Wireless communications, tablet computers, smartphones</td></tr>
<tr><td>Robustness</td><td>Robust</td><td colspan="2">Potentially unreliable (Detect+Correct, or Error-Resilient)</td></tr>
<tr><td>Area priority</td><td>Secondary</td><td></td><td>High</td></tr>
<tr><td>Supply voltage V_{DD}</td><td>Subthreshold (sub-V_T), e.g., 400 mV</td><td>Slightly scaled, near-threshold (near-V_T), e.g., 600 mV</td><td>Nominal, e.g., 1 V</td></tr>
<tr><td>Power</td><td>Ultra low, fW−pW</td><td></td><td>High, mW−W</td></tr>
<tr><td>Speed</td><td>Very slow, kHz−MHz</td><td></td><td>Fast, 100MHz−GHz</td></tr>
<tr><td rowspan="2">State of the art</td><td colspan="3">Bistables (latches, flip-flops), pipeline registers</td></tr>
<tr><td>8T, 10T, ..., 14T-bitcell SRAM
No good compilers!</td><td colspan="2">6T-bitcell SRAM, compilers
1T-1C eDRAM: special technology, extra cost
Gain-cells: logic-compatible</td></tr>
<tr><td rowspan="3">Contributions</td><td colspan="3">Standard-cell based memories (SCMs)</td></tr>
<tr><td>Low-leakage latches, ReRAM-based NVFF</td><td>Commercial library</td><td>Dynamic latches</td></tr>
<tr><td colspan="3">Gain-cell based eDRAM (GC-eDRAM)</td></tr>
<tr><td></td><td>2T sub-V_T</td><td>2T near-V_T</td><td>3T multilevel</td></tr>
</table>

datapaths. Static latches and flip-flops are reliably operated at a large range of supply voltages, including sub-V_T voltages. Memory macrocells based on the conventional 6T SRAM bitcell can be used for all applications running at nominal or slightly scaled supply voltages. In fact, almost invariably, SRAM has been the mainstream solution for on-chip embedded memories for virtually all VLSI SoC target applications for the last decades [8]. This unquestioned dominance of SRAM technology for on-chip storage mostly arises from their fast write and read accesses and their robust operation (at least in mature CMOS nodes and at nominal supply voltage V_{DD}). Also, for most process nodes, SRAM memory compilers are readily available. However, the footprint of the 6T SRAM bitcell is relatively large. In order to increase the storage density, eDRAM macrocells are an interesting alternative to SRAM macrocells. We distinguish two types of eDRAM: 1) conventional, 1T-1C eDRAMs whose basic bitcell is built from a special, high-density, 3D capacitor and an access transistor; and 2) gain-cell based eDRAMs (e.g., [28]) whose basic bitcell is built from 2–4 MOS transistors [29]. Conventional 1T-

1C eDRAMs typically require special process options to build high-density stacked or trench capacitors [30] and are therefore not compatible with standard digital CMOS technologies. Such process options become only available at an extra cost. As opposed to this, gain-cell based eDRAMs are fully compatible with baseline digital CMOS technologies and can easily be integrated into any SoC at no extra cost. The main drawback of gain-cells is the small storage node capacitor (compared to the dedicated DRAM capacitors) and the low retention time. From a functional perspective, all types of dynamic memories usually require refresh cycles that are costly in terms of access bandwidth and power.

6T-bitcell SRAM fails to operate reliably at aggressively scaled supply voltages. As shown in Table 1.1, alternative SRAM bitcells consisting of 8, 10, or even more transistors are required to ensure reliable sub-V_T operation [31]. In addition to large, alternative SRAM bitcells, various low-voltage write and read assist techniques have recently been proposed. Unfortunately, good memory compilers yielding robust sub-V_T SRAM macrocells are not commercially available.

1.4 Contributions

This PhD dissertation makes many contributions to the field of embedded memories for use in a large range of VLSI SoCs. In general, most contributions aim at either improving the area-efficiency or reducing the power consumption of embedded memories by using novel CMOS-compatible memory technologies and circuit techniques. In addition, some of the proposals made in this thesis allow and/or simplify the use of highly robust memories under extreme operating conditions (such as subthreshold circuit operation). Furthermore, additional contributions are memory optimizations (e.g., refresh-free eDRAM) for VLSI systems characterized by special memory usage (e.g., frequent write updates). The various contributions of this PhD thesis are in two main research areas, namely the fields of standard-cell based memories (SCMs) and gain-cell based eDRAM (GC-eDRAM) design, as expatiated on below.

Standard-Cell Based Memories (SCMs)

While SRAM macrocells are the unquestionable mainstream solution, synthesized latch or flip-flop arrays have also been used since a long time to implement small storage arrays distributed in datapaths. Unfortunately, there are no previous studies which systematically compare all possible architectural variants of latch and flip-flop arrays. This PhD dissertation uses *standard-cell based memories (SCMs)* as an umbrella term for all types of latch and flip-flop arrays and makes the following specific contributions in the field of SCMs.

SCM Architectures for Above-V_T and Sub-V_T Applications For the first time, this thesis systematically investigates and compares all architectural variants for the write logic, read logic, and storage cell implementation of SCMs. As shown in Table 1.1, targeting a large range

of applications, the comparative analysis of SCM topologies is carried out both in the above-threshold (above-V_T) domain at nominal supply voltage and in the sub-V_T domain in order to identify the respective best-practice implementations. Initially, we consider only SCMs synthesized from commercially available standard-cell libraries (SCLs), for straightforward integration of the proposed SCM topologies into any VLSI system. In order to draw conclusions as general as possible, various different technology nodes, semiconductor fabrication lines ("fabs"), and SCL provideres are considered.

Detailed Comparison with SRAM It is intuitively clear that SCMs are smaller than SRAM macrocells for small storage capacities (due to less peripheral circuits), but become significantly larger for larger storage capacities (due to the larger bitcell). In this dissertation, we systematically investigate this area comparison between SCMs and SRAM macrocells and describe the border line below which SCMs are still more area-efficient than SRAM macrocells. This analysis is carried out for SCMs based on commercially available, robust, static latches, as well as various custom-designed, high-density, dynamic latches. A case study of a low-density parity-check (LDPC) decoder, extensively used in wireless communications, shows how SCMs promote higher data locality and lower power consumption than SRAM macrocells (at the cost of area for the considered memory sizes, in case of using static SCMs).

Customization of Standard-Cells In a further main contribution, the design of custom standard-cells and their integration into SCMs is proposed in order to address the specific requirements of given target applications. In all cases, we are able to dramatically improve a given target metric (such as leakage power or storage density) by designing only one custom standard-cell and integrating it into the SCM compilation flow.

Ultra-Low Leakage Sub-V_T SCMs In ultra-low power (ULP) systems, the leakage currents of sub-V_T memories typically dominate the total power budget. The major leakage contributors of sub-V_T SCMs are identified to be the latches and the read multiplexers; the design of a single, custom-designed standard-cell latch with a tri-state output buffer addresses all major leakage contributors at once and enables a significant leakage power reduction. Using such ultra-low leakage sub-V_T SCMs, we demonstrate the lowest ever measured leakage power and access energy per bit among all sub-V_T memories in a 65 nm CMOS node. The sub-V_T SCM compilation flow which integrates custom-designed, ultra-low leakage standard-cells is a convenient tool for many low-power SoC designers especially when considering the lack of good, commercially available sub-V_T memory compilers.

OxRAM Based Non-Volatile Flip-Flops While the proposed sub-V_T SCMs exhibit extremely low leakage power and access energy (outperforming all previous works on sub-V_T SRAMs in the same technology node), emerging non-volatile memory technologies such as resistive

memories (e.g., oxide stacks) bare the potential for zero-leakage sleep states. For the first time, we investigate how oxide memory (OxRAM) devices can be interfaced with CMOS circuits and reliably read out at sub-V_T voltages. We propose the first non-volatile flip-flop topology which can be operated at sub-V_T voltages (only writing the OxRAM device requires a nominal voltage).

Dynamic SCMs As a further, completely opposite example to demonstrate the high benefit of standard-cell customization, we propose to investigate the access statistics of all internal memories of a LDPC decoder. Since all internal memories are frequently and periodically updated, it is possible to use our proposed dynamic SCMs (D-SCMs), i.e., storage arrays synthesized from dynamic latches, even without the need for a power-hungry refresh operation. Compared to their static counterpart, the D-SCMs enable a significant reduction of the silicon area of the LDPC decoder chip. In addition, silicon measurements show a slight reduction in power consumption, enabled by the D-SCMs, and confirm our proposed circuit techniques avoiding short-circuit currents.

Gain-Cell Based eDRAM (GC-eDRAM) Design

Most previous works on gain-cell based eDRAM (GC-eDRAM) try to promote gain-cells as high-density alternative to SRAM bitcells in cache memories of microprocessors. Therefore, the focus of almost all previous works was on achieving high speed and access bandwidth (beside high storage density) while operating at nominal supply voltages. Unfortunately, there are no previous studies on the behavior of GC-eDRAM under supply voltage scaling and no previous initiatives to promote GC-eDRAM for low-voltage applications. Moreover, there are no silicon-proven GC-eDRAM implementations in nodes below 65 nm CMOS. This PhD thesis makes the following specific contributions in the field of GC-eDRAM design.

Voltage Scaling for GC-eDRAM In this PhD dissertation, we investigate for the first time systematically the impact of voltage scaling on the retention time of gain-cells. It is shown that in some cases, depending on the write access statistics and the write-bit line (WBL) control scheme, surprisingly, the retention time can be increased by means of voltage down-scaling, favoring near-threshold (near-V_T) operation for GC-eDRAMs (see Table 1.1).

GC-eDRAM Retention Time Improvement by Reverse Body Biasing With the ultimate goal of reducing the refresh power of low-voltage, near-V_T GC-eDRAM arrays, different techniques to improve the retention time are proposed. Silicon measurements show that the retention time is dramatically improved when switching from a slight forward body bias (FBB), used for fast memory access, to a reverse body bias (RBB).

Replica Technique for Refresh Power Reduction In addition, a replica technique is proposed in order to automatically determine the optimum refresh rate for varying process-voltage-technology (PVT) conditions and according to write access statistics (which impact the retention time). Silicon measurements show that the proposed replica technique successfully tracks the effective retention time of the GC-eDRAM array and leads up to a 5× retention time extension, which results in a significant refresh power reduction compared to conventional retention time guardbanding.

Sub-V_T GC-eDRAM In addition to various techniques to reduce the refresh power of near-V_T GC-eDRAMs, we demonstrated for the first time the feasibility of sub-V_T operation of GC-eDRAMs in a mature CMOS node. In fact, despite heavily degraded on-to-off current ratios in the sub-V_T domain, resulting in low access time to retention time ratios, it is possible to achieve high array availability for random write and read access.

Low-Voltage GC-eDRAM in Deeply Scaled CMOS A simulation-based study to find the minimum recommended supply voltage for sufficiently high array availability is conducted at deeply scaled CMOS nodes, as well. Unfortunately, high aggregated leakage currents from the storage node and low in-cell storage capacitance in aggressively scaled CMOS nodes limit the amount of voltage scaling. While near-V_T operation is still viable in a 40 nm node, we show that sub-V_T operation should be avoided due to prohibitively short retention times compared to the access times.

Multilevel Gain-Cells Finally, in the context of error-resilient VLSI systems (such as wireless communications systems), which are able to tolerate a small amount of hardware defects in general and memory read failures in particular, we investigate the trade-off between reliability and storage density in GC-eDRAM. More precisely, 100% correct circuit operation is traded off for the benefit of higher storage density by proposing multilevel gain-cells where up to 4 voltage levels, corresponding to 2 bits, are stored in a single cell. In order to locally generate several voltage levels for data storage, as well as reference voltage levels for sensing at a low area overhead, charge sharing among precharged and pre-discharged bitline segments is used. In order to read out the multilevel gain-cells, a successive approximation algorithm is used, comparing one data level to various reference levels. Post-layout circuit simulations indicate 2% of read failures after a 10 µs retention time for a multilevel GC-eDRAM implementation in 90 nm CMOS. Moreover, as an additional contribution, a replica bitline (BL) technique is proposed to improve both the read and write access times of a multilevel GC-eDRAM macrocell under varying PVT conditions.

1.5 Thesis Outline

The remainder of this PhD dissertation is organized as follows.

Chapter 2 is dedicated to standard-cell based memories (SCMs) operated at nominal voltage and targeted toward high-performance VLSI systems. Section 2.1 provides background information, motivates the use of SCMs, and lists all advantages as well as drawbacks of SCMs. Section 2.2 introduces and compares SCM topologies based on commercially available standard-cell libraries (SCLs), and compares the best-practice SCM implementation to SRAM macrocells, as well, before presenting a case study where SCMs are used in a low-power low-density parity-check (LDPC) decoder. Section 2.3 discusses the integration of custom-designed, dynamic latches into the SCM compilation flow for high storage density, and, in a second case study, proposes the use of such dynamic SCMs (D-SCMs), operated without refresh cycles, in a LDPC decoder which frequently updates all internal memories.

Chapter 3 is dedicated to SCMs operated at aggressively scaled supply voltages residing in the sub-V_T domain and targeted toward ultra-low power applications. Section 3.1 explains the various challenges of low-voltage SRAM operation and presents a review of the state of the art of sub-V_T SRAM design. Section 3.2 evaluates and compares all SCM topologies built from commercial SCLs for operation in the sub-V_T regime. In Section 3.3, the so identified best-practice sub-V_T SCM topology is further optimized by the integration of an ultra-low leakage, custom-designed standard-cell. Section 3.4 presents non-volatile flip-flop topologies, for use in SCMs or as state registers, based on emerging OxRAM devices, in order to enable zero-leakage standby modes in future low-power applications.

Chapter 4 is dedicated to gain-cell based eDRAM (GC-eDRAM) for both low-voltage, low-power applications and high-performance VLSI systems. Section 4.1 reviews various types of eDRAM, explains the assets and drawbacks of GC-eDRAM, and presents a detailed review of the state of the art of GC-eDRAM design. Section 4.2 analyzes the impact of voltage down-scaling on the retention time and refresh power of a 2-transistor (2T)-bitcell GC-eDRAM array. Section 4.3 proposes several techniques to enhance the retention time and reduce the refresh power of near-V_T GC-eDRAM macrocells, including reverse body biasing and replica techniques. Section 4.4 studies the feasibility of sub-V_T operation for GC-eDRAM in light of technology scaling. Finally, Section 4.5 presents the design of a multilevel GC-eDRAM implementation for high-density data storage in error-resilient VLSI systems, including replica techniques for optimally fast write and read access under PVT variations.

Conclusions are drawn in Chapter 5.

1.6 Selected Publications

This PhD dissertation is mostly based on the following journal articles and conference papers. A complete list of book chapters, journal articles, conference papers, and invention disclosures can be found in the attached curriculum vitae.

Journals

"Comparative Analysis of ReRAM-Based Non-Volatile Flip-Flop Topologies with Sub-VT Read and CMOS Voltage-Compatible Write," I. Kazi, P. Meinerzhagen, P.-E. Gaillardon, D. Sacchetto, A. Burg, and G. De Micheli, IEEE Transactions on Circuits and Systems I (T-CASI), *in preparation*

"Comparative Analysis of Energy, Area, and Failure Probability of Dynamic Latch Topologies," P. Meinerzhagen, A. Bonetti, G. Karakonstantis, and A. Burg, IEEE Transactions on Very Large Scale Integration Systems (T-VLSI), *under internal review*

"Area-Efficient Low-Density Parity Check (LDPC) Decoder with Refresh-Free eDRAMs," P. Meinerzhagen, A. Bonetti, G. Karakonstantis, C. Roth, F. Gürkaynak, and A. Burg, IEEE Transactions on Circuits and Systems II (TCAS-II), *under review*

"Replica Technique for Adaptive Refresh Timing of Gain Cell embedded DRAM," A. Teman, P. Meinerzhagen, R. Giterman, A. Fish, and A. Burg, IEEE Transactions on Circuits and Systems II (TCAS-II), *accepted*

"On the Impact of Body Biasing on the Retention Time of Gain-Cell Memories," P. Meinerzhagen, A. Teman, A. Burg, and A. Fish, Journal of Engineering (JoE), August 2013

"Exploration of Sub-VT and Near-VT 2T Gain-Cell Memories for Ultra-Low Power Applications under Technology Scaling," P. Meinerzhagen, A. Teman, R. Giterman, A. Burg, and A. Fish, Journal of Low Power Electronics and Applications (JLPEA), April 2013, **invited article**

"Benchmarking of Standard-Cell Based Memories in the Sub-VT Domain in 65-nm CMOS Technology," P. Meinerzhagen, Y. Sherazi, A. Burg, J. Rodrigues, in IEEE Journal on Emerging and Selected Topics in Circuits and Systems (JETCAS), June 2011

Conferences (All Peer-Reviewed, 4+ Pages Papers)

"Dual-VT 4kb Sub-VT Memories with <1 pW/bit Leakage in 65nm CMOS," O. Andersson, B. Mohammadi, P. Meinerzhagen, A. Burg, and J. N. Rodrigues, IEEE European Solid-State Circuits Conference (ESSCIRC), September 2013

"ReRAM-Based Non-Volatile Flip-Flop with Sub-VT Read and CMOS Voltage-Compatible Write," I. Kazi, P. Meinerzhagen, P.-E. Gaillardon, D. Sacchetto, A. Burg, and G. De Micheli, IEEE International NEWCAS Conference, June 2013

"Review and Classification of Gain Cell eDRAM Implementations," Adam Teman, P. Meinerzhagen, A. Burg, and A. Fish, IEEE Convention of Electrical and Electronics Engineering in Israel (IEEEI), November 2012, **invited paper**

"A Sub-VT 2T Gain-Cell Memory for Biomedical Applications," P. Meinerzhagen, A. Teman, A.

Mordakhay, A. Burg, and A. Fish, in Proc. IEEE Subthreshold Microelectronics Conference, October 2012

"A 500 fW/bit 14 fJ/bit-access 4kb Standard-Cell Based Sub-VT Memory in 65nm CMOS," P. Meinerzhagen, Oskar Andersson, Babak Mohammadi, Yasser Sherazi, Andreas Burg, and Joachim Neves Rodrigues, in Proc. IEEE European Solid-State Circuits Conference (ESSCIRC), September 2012

"Replica Bit-Line Technique for Embedded Multilevel Gain-Cell DRAM," U. Khalid, P. Meinerzhagen, A. Burg, in Proc. IEEE International NEWCAS Conference, June 2012

"Two-Port Low-Power Gain-Cell Storage Array: Voltage Scaling and Retention Time," R. Iqbal, P. Meinerzhagen, A. Burg, in Proc. IEEE International Symposium on Circuits and Systems (ISCAS), May 2012

"Synthesis Strategies for Sub-VT Systems," P. Meinerzhagen, O. Andersson, Y. Sherazi, A. Burg, J. Rodrigues, in Proc. IEEE European Conference on Circuit Theory and Design (ECCTD), August 2011, **invited paper**

"Design and Failure Analysis of Logic-Compatible Multilevel Gain-Cell-Based DRAM for Fault-Tolerant VLSI Systems," P. Meinerzhagen, O. Andic, J. Treichler, A. Burg, in Proc. ACM/IEEE GLSVLSI, May 2011

"A 15.8 pJ/bit/iter Quasi-Cyclic LDPC Decoder for IEEE 802.11n in 90 nm CMOS," C. Roth, P. Meinerzhagen, C. Studer, A. Burg, in Proc. IEEE A-SSCC, November 2010

"Towards generic low-power area-efficient standard cell based memory architectures," P. Meinerzhagen, C. Roth, A. Burg, in Proc. IEEE International Midwest Symposium on Circuits & Systems, August 2010, **nomination student paper contest**

2 Standard-Cell Based Memories (SCMs) for High-Performance VLSI Systems

Standard-cell based memories (SCMs) are random-access memories (RAMs) which can be synthesized from standard-cell libraries (SCLs). Topologically, SCMs are latch or flip-flop arrays with logic circuits to control random write and read access. As an alternative to SRAM macrocells, SCMs are immediately functional in any VLSI system, even if operated at ultra-low voltages or if implemented in deeply scaled CMOS nodes, and considerably simplify any digital design flow, since they can be synthesized and placed-and-routed together with logic blocks. The use of SCMs is especially interesting for applications requiring many small memory blocks distributed within datapaths. This Chapter presents, compares, and optimizes SCMs for use in high-performance VLSI SoCs. In order to cope with high speed performance requirements, the analyses presented in this Chapter are limited to circuit operation at high, typically nominal supply voltage (V_{DD}). We consider both the case of synthesizing SCMs from commercially available SCLs exclusively, for maximum portability and short design times, as well as the case of relying on dynamic, custom-designed latches for high storage density. In this Chapter, SCMs are primarily developed and studied as stand-alone entities to be used in a large variety of digital VLSI systems; however, we also consider SCMs as building blocks of wireless channel decoders throughout this Chapter in order to carefully study and understand the benefits which SCMs enable at a higher integration level, namely the digital VLSI system-on-chip level.

Digital IC designers predominantly use SRAM macrocells to implement on-chip memory functionality; in Section 2.1 we argue that in many situations, SCMs can have advantages over SRAM macrocells. In particular, for reasonably small storage capacities, SCMs might be an interesting alternative to SRAM macrocells in order to improve area- and energy-efficiency, amongst others. Section 2.1 also introduces the application example used throughout this Chapter, namely low-density parity-check (LDPC) decoders, and identifies unique advantages enabled by SCMs in such decoders.

Section 2.2 introduces and compares a large variety of SCM topologies and presents an application example in the field of wireless communications. In Section 2.2.1, various ways to implement SCMs based on commercial SCLs are presented and compared to each other for different CMOS technology nodes, semiconductor fabrication lines (“fabs”), and various

SCL library providers; in addition, the best-practice SCM implementations are compared to corresponding SRAM macrocells, aiming for finding the most adequate memory option for each application. In Section 2.2.2, the benefits and drawbacks of SCMs compared to SRAM macrocells are illustrated with the example of a low-power LDPC decoder. The LDPC decoder using SCMs was manufactured in a 90 nm CMOS node and silicon measurement results are presented, as well.

Section 2.3 discusses the advantages and drawbacks of standard-cell customization, using dynamic latches for high storage density, and presents a further application example in the field of decoders for wireless communication channels. In Section 2.3.1, various dynamic latch topologies are introduced and compared in terms silicon area, reliability, and ease of integration into a digital design flow. Section 2.3.2 analyzes the access patterns of all internal memories of an LDPC decoder and proposes the use of high-density, dynamic SCMs (D-SCMs) which can be operated without the need for a refresh operation due to frequent and periodic write updates.

Section 2.4 draws conclusions from this Chapter.

This Chapter is partially based on our previous publications [32, 33, 2].

2.1 Introduction

As opposed to microprocessors requiring large cache memories, preferentially implemented as SRAM macrocells for high storage density and compatibility with logic CMOS technologies, many applications in the broad field of integrated circuit (IC) design require many small storage arrays distributed within datapaths. A few examples of such VLSI systems include channel decoders for wireless communications (e.g., Turbo, Viterbi, or LDPC decoders), VLSI implementations of FFT algorithms, and many other digital signal processing (DSP) systems. In order to integrate a large number of small memory arrays and seamlessly merge them with logic circuits, these embedded memories can conveniently and preferably be implemented as SCMs rather than SRAM macrocells. In the following, the advantages and potential drawbacks of SCMs are discussed in detail.

Advantages and Drawbacks of SCMs

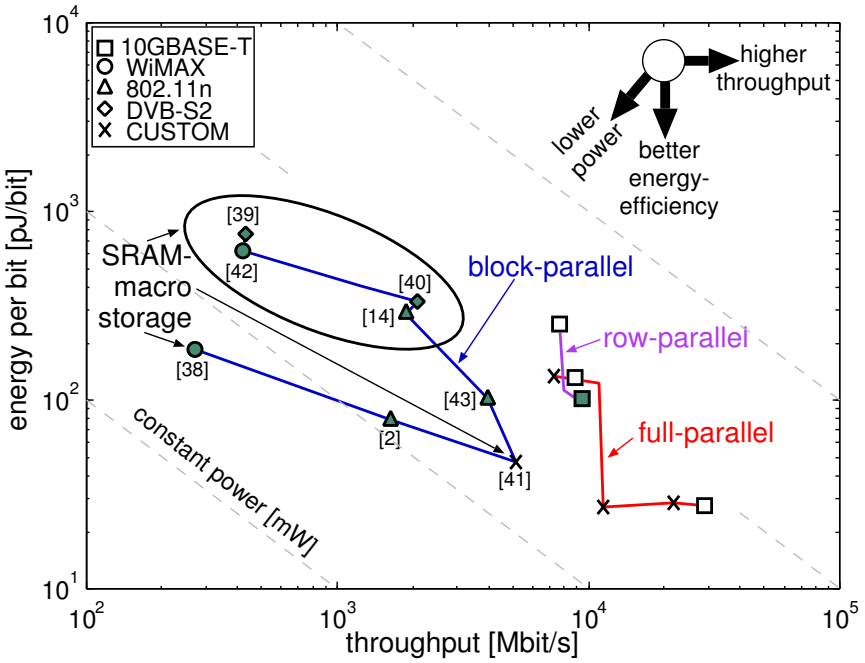
The use of SCMs described in a hardware description language (HDL), such as VHDL or Verilog, eases the portability of a design to other technologies. SRAM macrocells need to be created again for each new technology node or process design kit (PDK), using a dedicated memory compiler which might generate cells that are not fully compatible with the original design. Also, SCMs can be described in a generic way, which renders it easy to modify the *number of words* or the *number of bits per word* at design time; also, any desired numbers can be chosen, which is not the case for typical SRAM compilers. Furthermore, designs comprising SCMs can

be placed fully automatically using the standard placement tool, whereas SRAM macrocells need to be placed manually or by a specifically written script. Consequently, SCMs can be merged with logic blocks, which improves data locality and thus can reduce routing overhead. The one-bit storage cell of SCMs (i.e., a flip-flop or a latch) is clearly bigger than the one of SRAM macrocells (typically the 6-transistor SRAM bitcell). However, SRAM macrocells require more peripheral circuitry such as precharge circuitry and sense amplifiers [34] than SCMs. For SRAM macrocells with small storage capacity, the area overhead due to peripheral circuitry can be significant. Hence, SCMs can outperform SRAM macrocells in terms of silicon area for small storage capacities, but become much bigger for large storage capacities. Moreover, the use of SCMs can reduce routing, which leads to a reduction in active (switching) power consumption. Also, traditional 6T-bitcell SRAM exhibits high failure rates under voltage scaling [15] and does not work reliably in the near-threshold (near-V_T) domain [35, 36]. As opposed to this, SCMs directly support voltage scaling and can even be reliably operated in the sub-threshold (sub-V_T) regime without the need for fullcustom design, as expatiated on in Chapter 3. For these reasons, SCMs are a promising, lower-power alternative to conventional 6T-bitcell SRAM macrocells. SCMs can share the power and ground rings with the rest of the chip (i.e., with logic blocks), while SRAM macrocells typically have extra rings. For reconfigurable designs targeting low power consumption, memories are preferably organized in many small blocks which can be individually clock-gated and/or power-gated. In the context of such fine-granular memory organizations, SCMs provide more flexibility at design time, might result in smaller overall area due to the lack of separate rings and less peripheral circuitry, and are more adequate to reduce the overall power consumption. In summary, SCMs entail a minimum design effort, simplify the digital design flow, are immediately functional in any VLSI system, and operate reliably at any supply voltage. The only apparent drawback of SCMs is their large silicon area exceeding the one of SRAM macrocells for large storage capacities (unless small, dynamic latches are used, as proposed in Section 2.3).

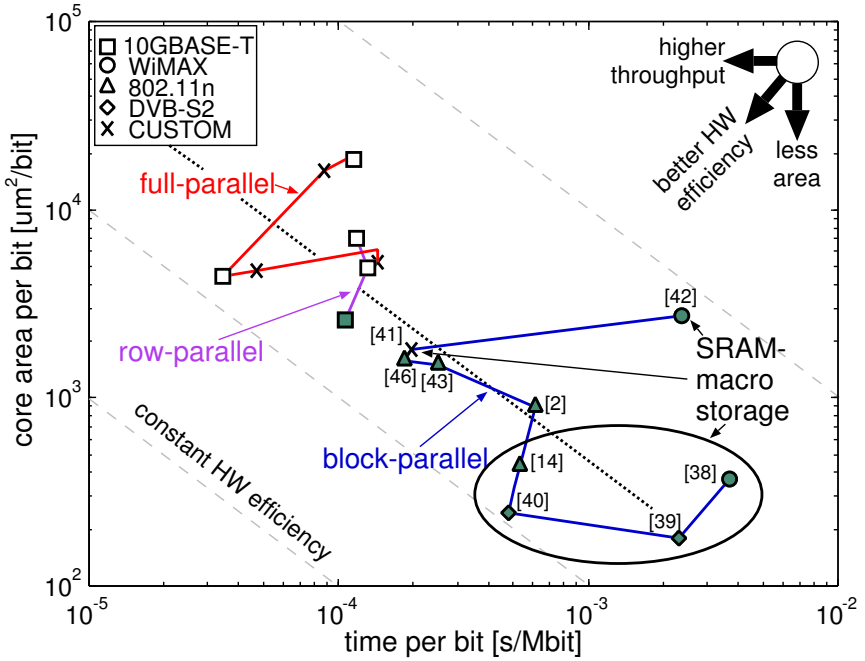
While many digital IC designers have previously used SCMs (our case study below exemplifies this), and while SCMs have several assets compared to SRAM macrocells, there are unfortunately no previous studies comparing all possible SCM topologies. In the following Section 2.2, we present a systematic comparative analysis of all possible SCM topologies. The best-practice SCM topology is then also compared with 6T-bitcell SRAM macrocells.

Application Example: LDPC Decoders

Low-density parity-check (LDPC) decoders are a good representative example of VLSI systems requiring many small, distributed storage arrays. We will therefore use such decoders as a case study throughout this Chapter to illustrate the advantages and drawbacks of SCMs. A tutorial paper [37] published in 2011, reviewing and analyzing the best LDPC decoder architectures known at that time, unveils that, in terms of embedded memories, the block-parallel LDPC decoder design community has adopted two different solutions: 1) one part of the community uses SRAM macrocells as internal memories [14, 38, 39, 40, 41, 42]; while 2) the other part



(a) Energy-efficiency vs. throughput of prior-art LDPC decoders [37].



(b) Area vs. time per bit of prior-art LDPC decoders [37].

Figure 2.1: (a) Energy-efficiency and throughput; and (b) area-efficiency and time per bit of state-of-the-art LDPC decoder implementations as of 2011. Decoder implementations based on SRAM are circled or highlighted by arrows, while all other block-parallel implementations are based on SCMs.

of the community prefers SCMs [2, 43, 44, 45, 46]. As shown in Fig. 2.1a, most block-parallel LDPC decoders using SRAM macrocells (circled or highlighted by arrows) have worse energy efficiency than the block-parallel decoders using SCMs (all remaining, not highlighted marks). The better energy efficiency of the latter decoder implementations can be attributed to two factors: 1) SCMs can be more energy-efficient than SRAM macrocells; and 2) SCMs merge better with logic blocks, result in less routing overhead (shorter wires), and lead to lower switching power. As a second observation from Fig. 2.1a, block-parallel LDPC decoders based on SCMs are generally faster and allow for higher decoding throughputs than decoders using SRAM macrocells. Unfortunately, as shown in Fig. 2.1b, the use of SCMs synthesized from commercial SCL, almost exclusively providing large, static latches and flip-flops, leads to a lower area-efficiency in LDPC decoders than the use of SRAM macrocells. In summary, if power awareness and high decoding performance are of high importance while silicon area is only a secondary concern, it is definitely beneficial to employ SCMs instead of SRAM macrocells in LDPC decoders.

2.2 SCMs Based on Commercial Standard-Cell Libraries (SCLs)

2.2.1 SCM Architectural Choices and Comparison

This Section introduces and discusses architectural choices for SCMs, before rigorously comparing all possible SCM architectures. The comparative SCM architecture analyses are carried out at three different CMOS nodes (180 nm, 130 nm, and 90 nm) and for different fabs and SCL providers in each node (resulting in a total of five different cases) to draw conclusions as generic as possible. The best-practice SCM architecture is then compared in detail with SRAM macrocells, as well. The remainder of this Chapter, as well as Chapter 3 assume memories (both SCMs and SRAM macrocells) with a separate read and write port, a word access scheme (as opposed to sub-word/byte access or bit-wise access), and a read and write latency of one, which are typical requirements for memories distributed within dedicated datapaths. As shown in Fig. 2.2a, any such SCM has the following building blocks: 1) a write logic, 2) a read logic, and 3) an array of storage cells. Different ways to implement the write and read logic are presented in the Sections “Write Logic” and “Read Logic” below, respectively, assuming flip-flops as storage cells. The use of latches instead of flip-flops as storage cells is discussed in the subsequent Section “Array of Storage Cells”.

Write Logic

Consider an array of $R \times C$ flip-flops, where R and C denote the number of rows (words) and the number of columns (bits per word), respectively. Assuming a word access scheme and a write latency of one cycle, the write logic needs to select one out of R words, according to the given write address, and update the content of the corresponding flip-flops on the next active clock edge. To accomplish this, the *write address decoder* (WAD) produces one-hot encoded

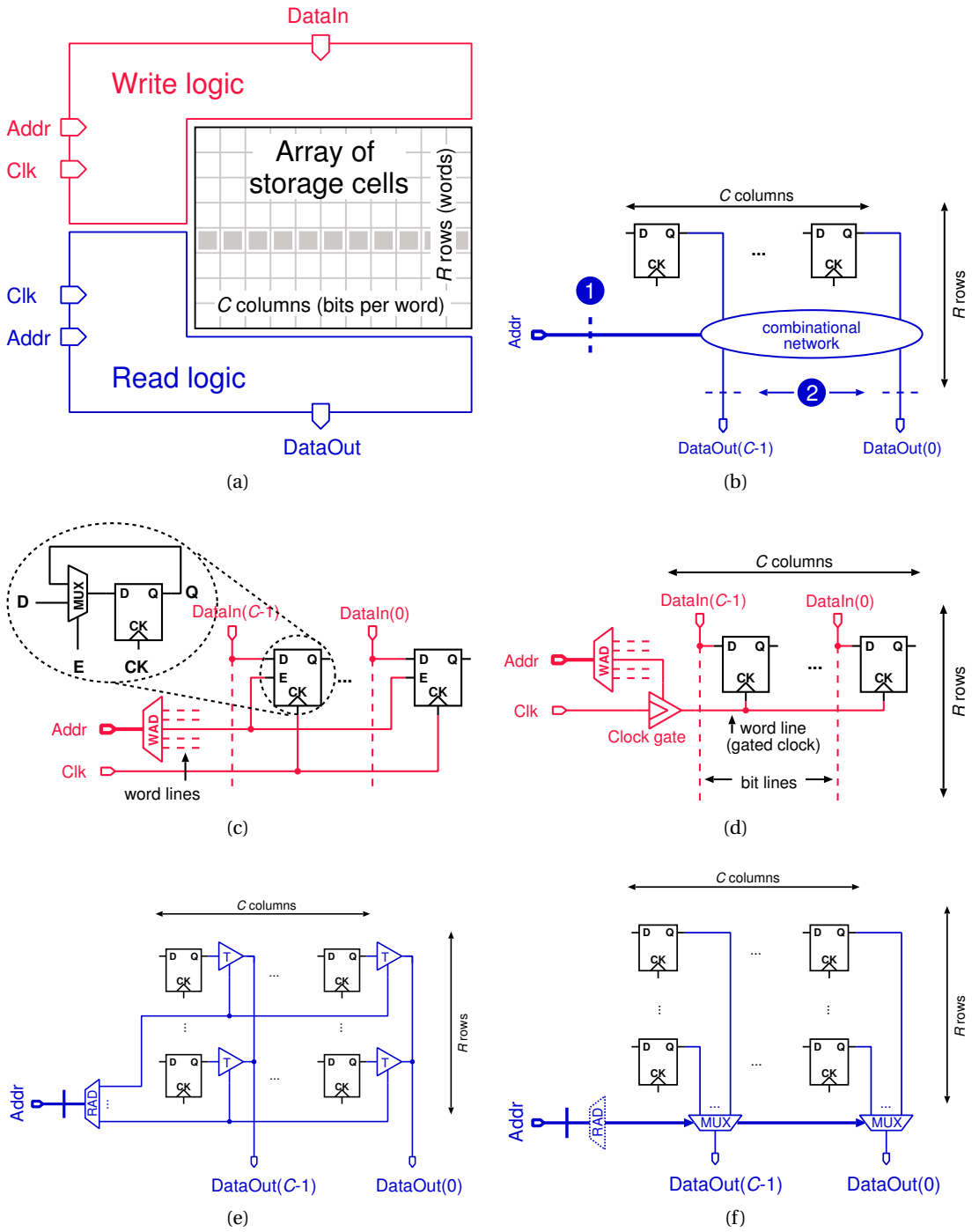


Figure 2.2: (a) Building blocks of a generic standard-cell based memory architecture. (b) Achieving typical one-cycle read latency. (c) Write logic relying on enable flip-flops, and (d) basic flip-flops in conjunction with clock-gates. (e) Read logic relying on tri-state buffers, and (f) CMOS multiplexers.

row select signals, which select one row of the flip-flop array. Next, the flip-flops in the selected row need to update their state according to the data to be written. One possibility consists in using flip-flops with an enable feature or with a corresponding logic (*FFE* architecture), as shown in Fig. 2.2c; all flip-flops in one row are enabled by the same row select signal. Another possibility consists in using basic flip-flops in conjunction with clock gates (*CG* architecture), as shown in Fig. 2.2d. In this case, a separate clock signal is generated for each row, and only the currently selected row receives a clock pulse, thereby sampling the provided data, while all other rows receive a silenced clock, thereby keeping their previous data.

Synthesis results using different CMOS technology nodes, different semiconductor fabs, and different standard-cell library providers show that the CG architecture yields smaller SCMs than the FFE architecture for $C \geq 4$ in most cases, and $C \geq 2$ in few cases. This result is almost always independent of R.

It is clear that the CG architecture consumes less power than the FFE architecture, as the latter distributes the clock signal to each storage cell, while the former silences the clock signal of all but the selected rows. Furthermore, the 2-to-1 multiplexer inside the enable flip-flop consumes additional power which can be avoided by the CG architecture.

Read Logic

As shown in Fig. 2.2b, the read logic can be purely combinational or contain sequential elements, which leads to a read latency. Assuming a word access scheme, one out of R words needs to be routed to the data output, according to the read address. The typical one-cycle latency is obtained by inserting flip-flops either at the read address input, see case (1) in Fig. 2.2b, or at the data output, see case (2) in Fig. 2.2b. The former and latter case require $\text{ceil}(\log_2(R))$ and C additional flip-flops, impose gentle and hard read address setup-time requirements, and cause considerable and negligible output delays, respectively. The task of routing one out of R words to the output is accomplished using either tri-state buffers or multiplexers.

Tri-State Buffer Based Read Logic This approach asks for a *read address decoder* (RAD) to produce one-hot encoded row select signals, and $R \cdot C$ tri-state buffers, i.e., exactly one per storage cell, as shown in Fig. 2.2e. R tri-state buffer outputs connect to one bit-line (BL), which has a large lumped capacitance if R is big. In fact, beside the gate capacitance in the fanout and interconnect parasitic capacitance, a large portion of the total BL capacitance arises from the junction capacitance of the tri-state buffers. In order to drive this large BL capacitance, tri-state buffers with high driving capability are required. If R increases, stronger buffers, exhibiting larger parasitic junction capacitance, are required, which further increases the lumped BL capacitance and thus requires even stronger buffers. Therefore, increasing the tri-state buffer's driving strength provides only a limited advantage to increase the read speed. Furthermore, it is generally difficult to buffer tri-state buses [47], which might be necessary

to maintain reasonable slew rates if these buses are routed over long distances. Also, if two or more row select signals accidentally overlap, DC paths from V_{DD} to ground can arise and short-circuit power is consumed. In summary, employing tri-state buffers is expected to result in a large overall area and a high power consumption, while it is challenging to achieve fast read operations.

Multiplexer Based Read Logic C parallel R-to-1 multiplexers are required to route an entire word to the output, as shown in Fig. 2.2f. The R-to-1 multiplexer itself can be implemented in many ways. Most multiplexer architectures, such as binary selection tree multiplexers, do not require one-hot encoded row select signals and can therefore save the RAD. However, there is an energy efficient multiplexer architecture which accepts one-hot encoded row select signals, performs a logic AND operation between each row select signal and the corresponding data bit, and finally performs a logic OR operation on all AND-gate outputs. For this particular multiplexer architecture, and assuming a proper, i.e., a non-overlapping one-hot code at the selection inputs, any glitch or activity on an unselected data input will die out after the first logic stage. As opposed to this, some glitches or activity on unselected data inputs of a binary selection tree multiplexer can propagate all the way to the input of the last stage, giving rise to unnecessary power consumption. In summary, intuitively, it is best for low power operation to use a glitch-free RAD to mask (AND operation) unselected data at the leaf-level of an OR-tree to realize the multiplexer functionality. Luckily, most logic synthesizers yield multiplexers similar to the AND-then-OR multiplexer, typically employing dedicated multiplexer cells in the back-most logic stages.

Post-Layout Simulation Results Comparing Write and Read Logic Implementations

Flip-flop based SCMs using clock gates for the write logic, and using either multiplexers or tri-state buffers for the read logic are synthesized, placed, and routed for different memory dimensions $R \times C$ (see Table 2.1) as well as for different CMOS technologies (various nodes and various fabs for the same node) and different standard cell libraries (see Table 2.2). For the voltage-change dump (VCD)-based post-layout power analyses, random data is written to random addresses, while data is read from random addresses, for 1000 cycles at a clock frequency of 100 MHz. All inputs of the SCMs can be driven by buffers of standard driving strength; highly capacitive nets such as the bit lines are buffered inside the SCMs.

The post-layout simulation results show that the multiplexer based SCMs always have smaller area and lower power consumption than the tri-state buffer based SCMs. However, the power estimation of the tri-state buffer based SCMs is rather optimistic as short-circuit power due to DC paths through tri-state buffers is not accounted for in the simulations.

Table 2.1: Flip-flop based SCM, CG write logic, 0.13 µm CMOS: area and power for multiplexer and 3-state read logic for different configurations $R \times C$.

		Area [µm^2]		Power [mW]	
R	C	MUX	3-state	MUX	3-state
16	8	6k	6k	0.8	0.8
16	128	67k	76k	5.3	6.9
32	8	10k	11k	1.0	1.3
32	128	135k	170k	8.1	14.1
64	8	20k	28k	2.4	4.2
64	128	274k	397k	19.5	38.4
128	8	39k	56k	4.5	9.1
128	128	557k	850k	38.0	93.8

Table 2.2: Flip-flop based SCM, CG write logic, $R = 16$, $C = 128$: area and power for multiplexer and 3-state read logic for different technologies and standard cell libraries.

	Area [µm^2]		Power [mW]	
Tech. & lib.	MUX	3-state	MUX	3-state
180 nm i)	132k	170k	11.0	19.8
180 nm ii)	126k	160k	12.5	17.0
130 nm i)	67k	76k	5.3	6.9
130 nm ii)	72k	83k	4.1	4.9
90 nm	36k	41k	1.9	3.5

Array of Storage Cells

Instead of flip-flops, latches can be used as storage cells. The previous discussions on the write and read logic remain valid when latches are used as storage cells. However, setup-time requirements on the write port become considerably more stringent when using latches. In fact, sticking to a single-edge-triggered one-phase clocking discipline and a duty cycle of 50%, the WAD together with the clock gates get only the first half of a clock period to generate one clock pulse and $R-1$ silenced clocks, which will make—during the second half of the clock period—the latches in one out of R rows transparent and keep the latches in all other rows non-transparent, respectively. Those latches which have received a clock pulse store the applied input data on the next active clock edge.

Furthermore, if the currently transparent latches are also selected by the output multiplexers, the SCM becomes transparent from its data input to its data output, and combinatorial loops through external logic can arise. To avoid this problem, a restriction on the choice of read and write addresses needs to be imposed. If such a restriction is not desired, latches which are

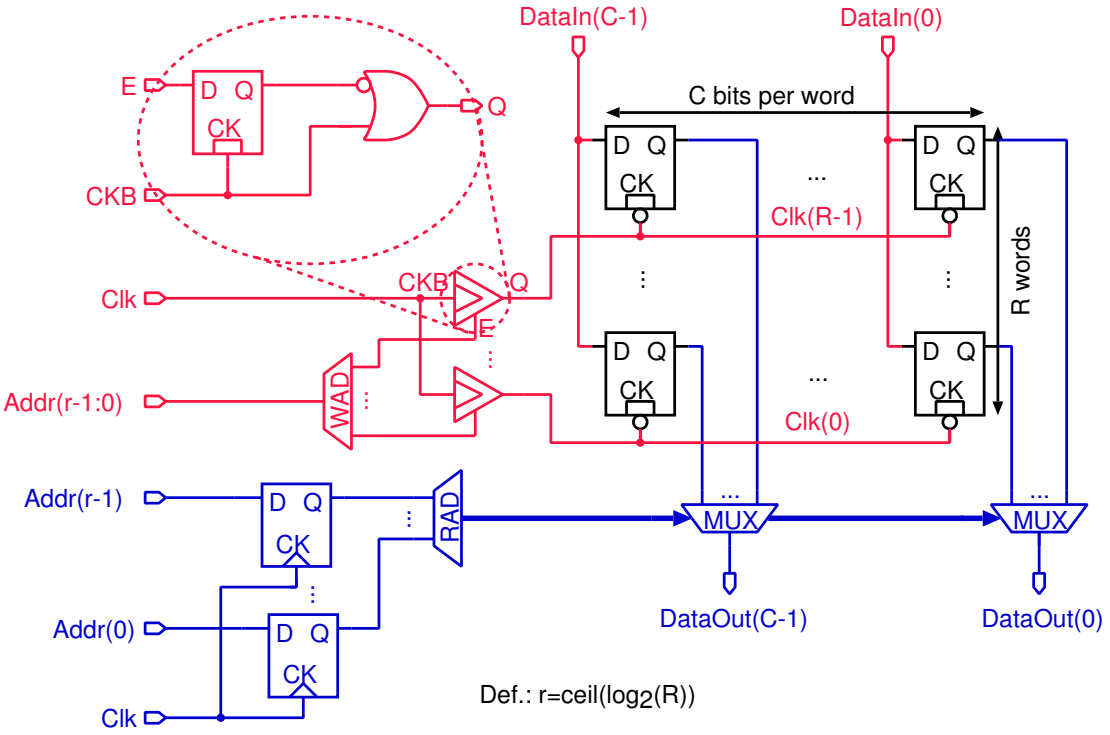


Figure 2.3: Schematic of latch based SCM with clock-gates for the write logic and multiplexers for the read logic.

non-transparent during the second half of the clock period need to be inferred at either the SCM's data input or output, or alternatively, registers need to be inserted into any path feeding the SCM's data output back to the data input.

Averaging across different CMOS technologies (nodes and fabs) and standard cell libraries, we find that the area of a basic latch with given drive strength is 77% of the area of a corresponding flip-flop. Even though the total storage cell area can be reduced by 23% on average when replacing flip-flops with latches, the total SCM area shrinks less, as write and read logic remain the same. In fact, for a 0.13 µm technology, averaging over 49 samples corresponding to $R = 2^3, 2^4, \ldots, 2^9$, $C = 2^1, 2^2, \ldots, 2^7$, latch based SCMs are only 13% smaller than flip-flop based SCMs.

In latch based SCMs, the WAD together with the clock-gates get only half a clock period to select one out of R words, while in flip-flop based SCMs, they get a full clock period. This is why flip-flop based SCMs qualify better for high-speed applications where address generation involves a long combinational path which cannot be pipelined.

Fig. 2.3 shows the schematic of the proposed standard-cell based memory, which uses latches without enable feature as storage cells, clock-gates for the write logic, and flip-flops at the read address input in conjunction with multiplexers for the read logic.

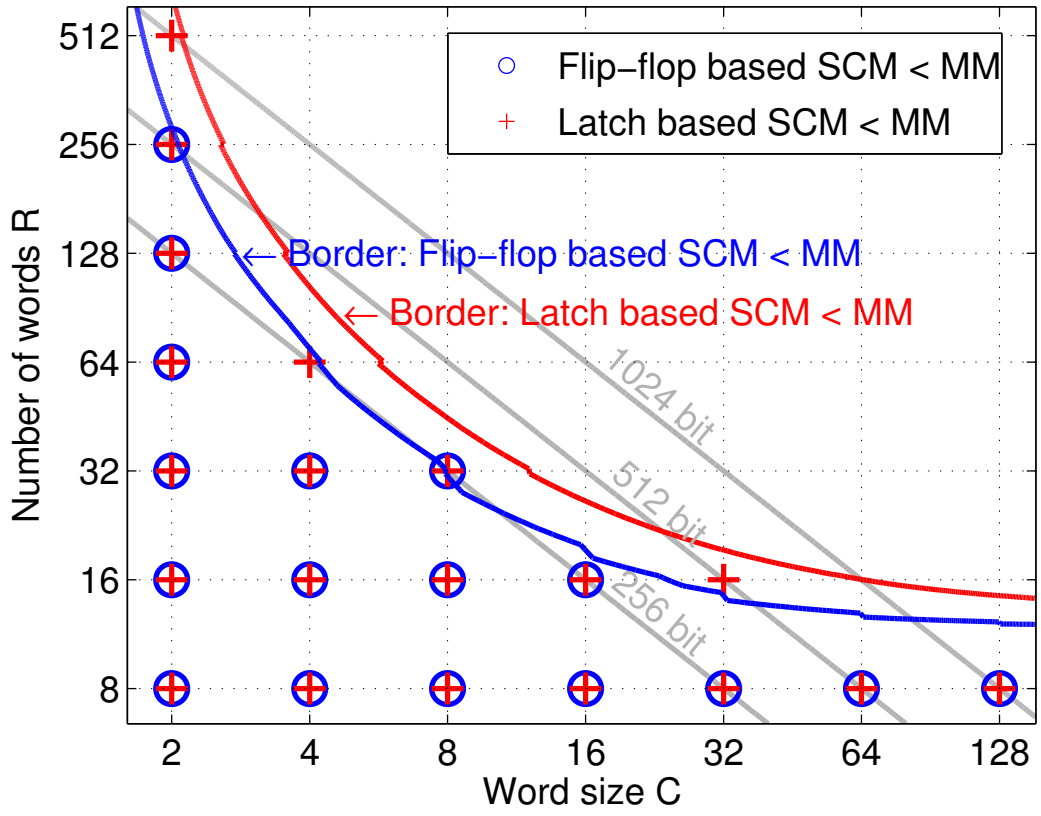


Figure 2.4: Flip-flop and latch based SCMs versus SRAM memory macros (MM): sampled data points and intersection lines of regression functions.

SCM Versus SRAM Area Comparison

For the smallest flip-flop and latch based SCM architectures, as well as for the SRAM memory macro (MM), 49 samples corresponding to $R = 2^3, 2^4, \ldots, 2^9$, $C = 2^1, 2^2, \ldots, 2^7$ have been synthesized in a 0.13 µm CMOS technology. Fig. 2.4 shows all points in the $C \times R$ plane—using a log-log scale—for which the SCMs are smaller than the corresponding SRAM macrocells. The sampled data points are interpolated in the least squares sense, and the intersection lines SCM = MM of the resulting surfaces are plotted, as well. Those intersection lines show the border up to which the SCMs are smaller than SRAM macrocells. Of course, changing from flip-flop based to latch based SCMs pushes the intersection line toward slightly bigger storage capacities $R \cdot C$. The gray lines show all memory configurations $C \times R$ with constant storage capacity $R \cdot C$. Flip-flop and latch based SCMs are smaller than SRAM macrocells for storage capacities of up to around 512 and 1024 bits, respectively, considering rather high but still very applicable C/R ratios.

Table 2.3: Area and power of SCM vs. SRAM based decoder.

	Dec. w/ SRAM	Dec. w/ SCM	SCM gain/penalty
Power [mW]	144.32	91.58	−36.54%
Area [mm^2]	1.37	2.06	+49.97%

2.2.2 Application Example: Low-Power LDPC Decoder

This Section investigates the use of the best-practice SCM in a low-power LDPC decoder. First, the impact of replacing all internal SRAM macrocells with SCMs on the silicon area and the power consumption of the LDPC decoder are evaluated for a 0.13 µm CMOS technology. Second, to demonstrate the simple design portability enabled by SCMs, the optimized low-power LDPC decoder architecture using SCMs is taped-out in a 90 nm CMOS node, and silicon measurement results are presented and compared with prior-art LDPC decoder implementations.

LDPC decoders used in modern communication systems require a considerable amount of memories, which often consume a dominant part of the total power. Furthermore, most wireless communication standards define several operating modes, which asks for a fine-granular memory organization if low power consumption is targeted. The employment of SCMs is thus a promising way for designing portable low-power LDPC decoder intellectual properties (IPs), even without the need for third-party SRAM macrocell IPs.

In the following, two versions of an IEEE 802.11n-compliant low-power LDPC decoder based on [14] are compared. The first version uses SRAM macrocells and the second one uses several instances of the previously proposed, best-practice, latch based SCM (see Fig. 2.3). Both decoders contain three separate memories, named Q-, T-, and R-memory, and some combinational blocks between them. The R-memory is divided into an $(R, C) = (88, 135)$ always-on block and two $(R, C) = (88, 135)$ blocks which can be turned off (clock-gated) separately, depending on the decoder's operating mode. Similarly, both Q- and T-memories are divided into three $(R, C) = (24, 135)$ blocks.

Considering that $R < C$ for all employed SCMs, flip-flops are inserted at the read address input rather than at the data output. Each multiplexer selection signal has a fan-out of $C = 135$, which requires buffering and causes a non-negligible delay. In fact, it turns out that the paths through the SCM output multiplexers are the most timing critical paths of the LDPC decoder design.

The two decoder versions are synthesized, placed, and routed in a 0.13 µm CMOS technology for a target clock period of 6 ns, which is required to achieve the throughput demanded by the IEEE 802.11n standard. Table 2.3 shows the core area and the VCD-based post-layout power analysis results for both decoder implementations.

The power analyses show that the SCM based decoder consumes 37% less power than the

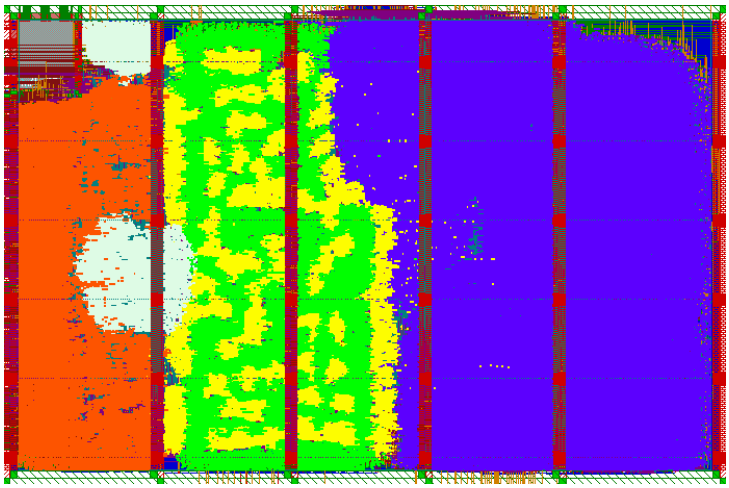


Figure 2.5: Layout of SCM based low-power LDPC decoder in 0.13 µm CMOS technology. The Q- and the R-memory are located on the left-hand and right-hand side, respectively, while the T-memory is located in the middle, merged with and surrounded by combinational logic blocks.

corresponding SRAM based decoder. The main part of the decoder's power reduction can be attributed directly to the lower power consumption of the employed SCMs as compared to the SRAM macrocells. Furthermore, power analyses and placement results show that SCMs enable a more local placement and routing, which leads to lower switching power. Fig. 2.5 for example shows that the T-memory in the SCM based decoder is completely merged into the main combinational block by the placement tool. This high data locality enables the routing tool to use shorter and lower-layer wires at these locations. Also, the fact that SCMs are not limited to a rectangular shape allows the placement tool to wrap the memories around the connected logic (see left part of Fig. 2.5), thereby minimizing wire lengths at the interfaces, which leads to a further reduction in switching power. For both decoder implementations in the considered 0.13 µm CMOS technology, the leakage power is less than 1% of the total power.

All memory sub-blocks resulting from dividing the Q-, T-, and R-memory have a capacity > 3kb, which is too high for SCMs to outperform SRAM macrocells also in terms of area (see Fig. 2.4). However, for the considered low-power LDPC decoder, an increased silicon area is acceptable for the benefit of lower power consumption.

After the comparative study of the SCM and SRAM based LDPC decoders, identifying the SCM based version as a promising, lower-power alternative to the SRAM based version, the decoder design was ported to a 90 nm CMOS node. The final design [2] is optimized for low power consumption at all design levels: 1) at the algorithmic level, an early termination mechanism avoids additional decoding iterations and power consumption if the likelihood of successfully decoding a data packet is low; 2) at the architectural level, a memory bypass mechanism avoids power-hungry memory accesses in case the data is used again shortly. Moreover, employing a sign-magnitude instead of a 2's complement number representation

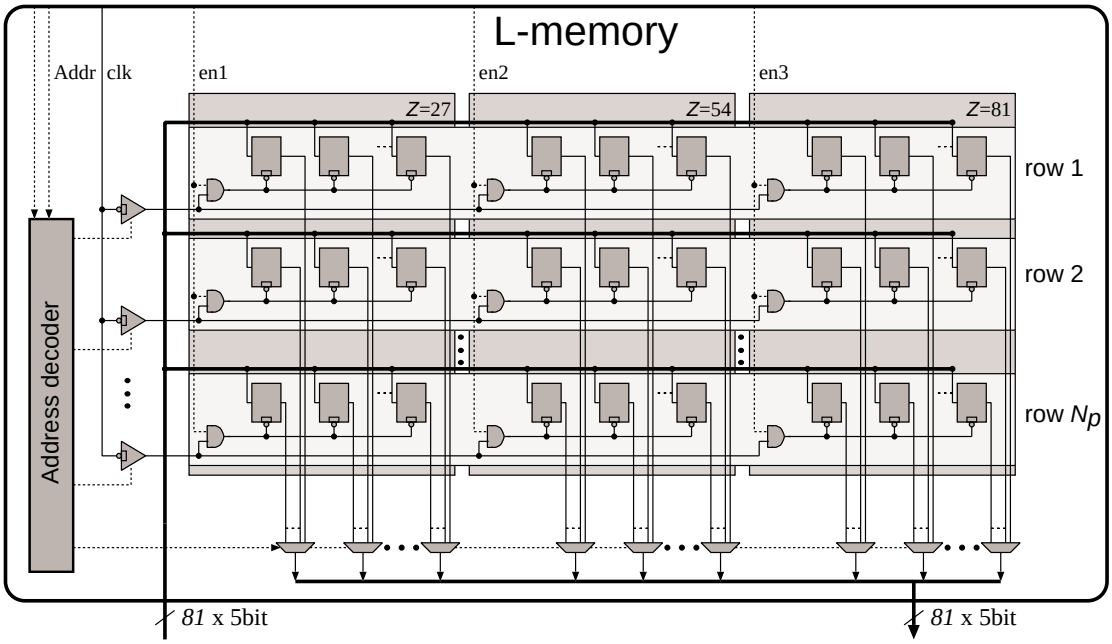


Figure 2.6: Modified SCM architecture with in-word clock-gating to support different LDPC code configurations.

reduces switching activity; and 3) at the circuit level, the previously presented, best-practice SCM implementation is further refined to enable in-word clock-gating according to the LDPC code configuration. In fact, some configurations do not require the full word length, and, as show in 2.6, unused blocks are clock-gated to avoid unnecessary switching power.

The chip microphotograph of the low-power LDPC decoder implementation in 90 nm CMOS technology is shown in Fig. 2.7. The core occupies a silicon area of 1.77 mm^2 with an active cell area of 398 kGE. Silicon measurements at a supply voltage of 1.0 V show a maximum clock frequency of 346 MHz which translates into a throughput of 680 Mbps (information bits) at 10 decoding iterations with the rate-5/6, $Z = 81$ code [48]. Further measurement results are summarized in Table 2.4, which also provides a comparison with prior-art quasi-cyclic (QC)-LDPC decoders. To account for differences in process technology, we scale the results to 90 nm and 1.0 V supply voltage. The proposed decoder exhibits a 2.4× and 1.9× better energy-efficiency than the decoders presented in [38] and [49], respectively, and our circuit is more hardware-efficient when taking technology scaling into account. Compared to the work in [14], which served as a reference design for the presented implementation, we were able to improve the energy-efficiency by a factor of 7.8 at the cost of a lower hardware efficiency. Beside algorithmic and architectural optimizations, a portion of these significant energy savings can doubtlessly be attributed to the use of SCMs instead of SRAM macrocells as in [14], while, unfortunately, the lower area efficiency also arises from the SCMs. The next Section investigates the use of custom-designed, dynamic latches in SCMs for higher area efficiency.

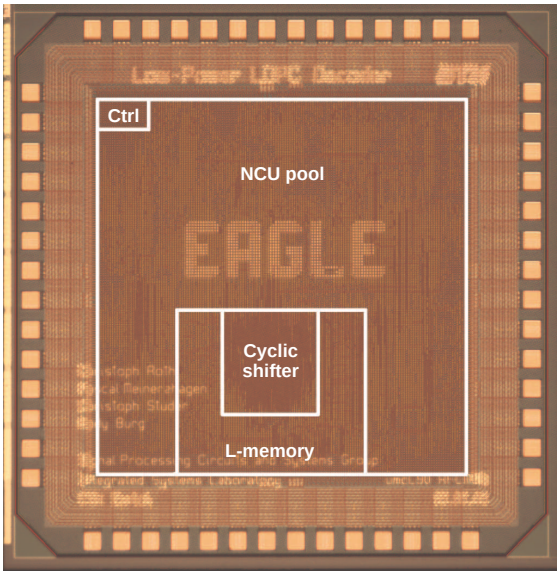


Figure 2.7: Chip microphotograph of the fabricated LDPC decoder using static SCMs.

Table 2.4: Comparison of quasi-cyclic (QC)-LDPC decoder implementations.

Publications	[38]	[43]	[49]	[14]	This work [2]
Technology [nm]	180	90	130	180	90
V_{dd} [V]	1.8	1.0	1.2	1.8	1.0
Basis of results	post-layout simulations		ASIC measurements		
$Z_{\max}$	96	96	64	81	81
Core area [mm^2]	3.39	3.5	2.46	3.39	1.77
Max. throughput^a in [Mbps]	57 (113^c)	1667	115 (166^c)	390 (780^c)	679
Hardware eff.^{a} in [µm^2/Mbps]	59.8 (7.5^c)	2.1	21.5 (7.7^c)	8.7 (1.1^c)	2.6
Energy eff.^{b} in [pJ/bit/iter]	243 (37.5^c)	34.2	63.2 (30.4^c)	800 (124^c)	15.8

^{a}at 10 iterations, $r = 5/6$.

^{b}measured at nominal supply voltage.

^{c}Technology scaling to 90 nm, $V_{dd} = 1.0$ V: $t_{pd} \sim 1/s$, $A \sim 1/s^2$, $P \sim 1/s \cdot (V'_{dd}/V_{dd})^2$.

2.3 High-Density Dynamic SCMs (D-SCMs)

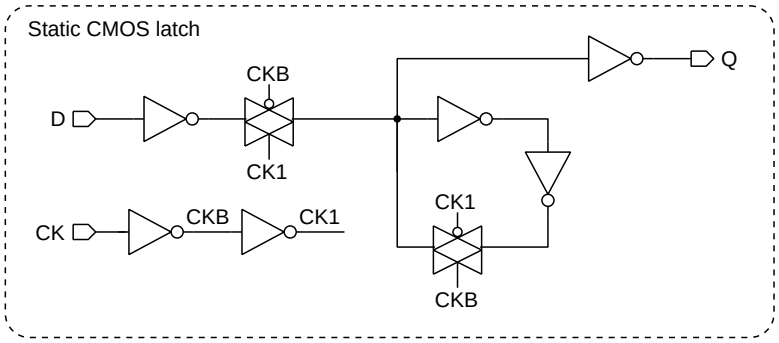
2.3.1 Integration of Custom-Designed Dynamic Latches

Thus far, even though considering various CMOS technology nodes and different fabs, the analysis of SCM topologies has been limited to the use of commercially available standard-cell

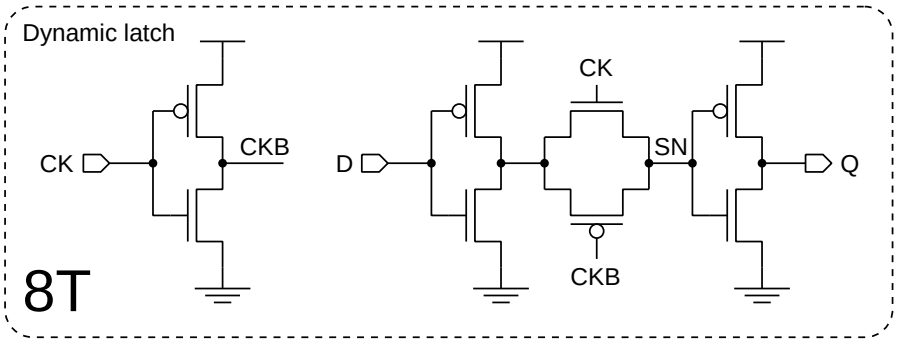
libraries (SCLs). Typically, such commercial SCLs provide only static latches and flip-flops which are optimized for high speed performance and high robustness at nominal supply voltage, serving the predominant needs and requirements in the broad field of VLSI design. In particular, the operating frequency of microprocessors and other VLSI systems is steadily increasing, and SCL providers follow this common trend by primarily focusing on flip-flops and latches with short insertion delay (i.e., the sum of setup time and clock-to-output propagation delay) to enable ever shorter clock periods; in fact, the insertion delay is the amount of time which the register takes out of the clock cycle, limiting the remaining, available propagation delay for logic circuits in a pipeline. Equally importantly, most SCL designers focus on highly robust circuit operation in a high volume manufacturing (HVM) context and under process-voltage-temperature (PVT) variations. Ensuring high circuit reliability is especially problematic for flip-flops due to their extremely high replication count in most VLSI SoCs; in fact, there are typically significantly more flip-flops in a given design than any other type of standard-cell. Due to these reasons, most library providers only offer static flip-flop and latch topologies, such as the one shown in Fig. 2.8a, while, unfortunately, they do not provide higher-density, dynamic flip-flop or latch topologies which are considered too error-prone and can retain data only for a limited time.

VLSI Systems Which Can Benefit from D-SCMs

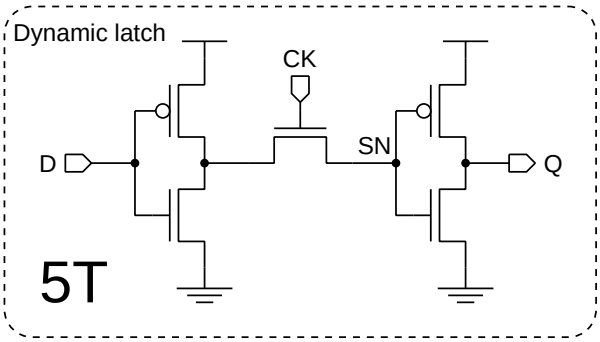
There is a large variety of VLSI systems which have two interesting properties which favor the use of dynamic latches for high area efficiency: 1) low data retention time requirements (from tens of ns to tens of µs); and 2) resilience to a given, typically small amount of hardware defects in general and memory bitcell failures in particular. VLSI implementations of wireless communications systems such as WLAN and high speed-packet access (HSPA+) systems are a typical class of applications requiring only short data retention times. As a concrete example, the previously presented LDPC decoder [2] (see Section 2.2.2) requires retention times as low as 288 ns, before new data is written to all internal memories anyway. Moreover, the LDPC decoder presented in [50] has a data retention time requirement of only 20 ns and can therefore use dynamic gain-cell based eDRAM macrocells. Such low retention time requirements and the periodic write accesses do not only allow to skip the power-hungry refresh cycles, but also allow to trade retention time of dynamic bitcells for the benefit of faster access or smaller silicon area. Beside short retention time requirements, several recent studies unveil that various VLSI systems, in the fields of multimedia [51], wireless communications [52, 21, 53], and data mining [54], just to name a few, are resilient to a small amount of hardware defects, such as broken memory cells. A general trend to such fault-tolerant VLSI systems [55, 56] is taking place mainly due to increasing process parameter variations and high defect levels in nanometric CMOS technologies. Exploiting fault tolerance is particularly intuitive and interesting for wireless communications systems, which are primarily designed to deal with channel-induced noise, but continue to work if they are build from slightly unreliable hardware. As an example, the work in [21] presents simulation results of a complete HSPA+ system with errors being injected in the hybrid automatic repeat request (HARQ) memory. It is shown that



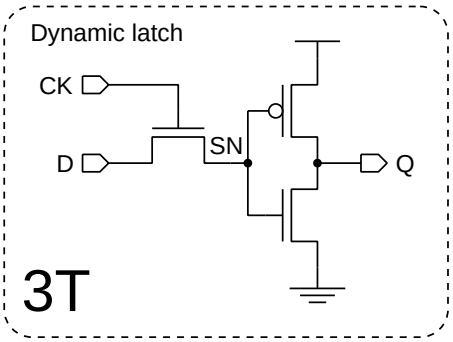
(a) Commonly used static latch.



(b) 8T dynamic latch topology.



(c) 5T dynamic latch topology.



(d) 3T dynamic latch topology.

Figure 2.8: (a) Conventional static latch topology used in most commercial SCLs. In newer SCLs for aggressively scaled CMOS nodes, it is increasingly more common to replace the inverter followed by a transmission-gate with a tri-state inverter for lower leakage; and various dynamic latch topologies, consisting of (b) 8 transistors, (c) 5 transistors, and (d) 3 transistors, respectively.

with a bitcell failure rate of 1% the system still achieves the required throughput. Moreover, if the four most significant bits (MSBs) of the log-likelihood ratios are stored in robust 8-transistor (8T) SRAM bitcells, the remaining bits can be stored in unreliable memory cells with a defect rate of up to 10% for an overall system throughput which is only slightly degraded compared to completely error-free hardware [21].

In the context of such VLSI systems which require only short data retention time and/or can tolerate a small amount of failing memory cells, we propose to use dynamic instead of static latch topologies for the benefit of higher storage density. Such custom-designed dynamic latches are characterized as standard-cells and integrated into the digital design flow, alongside with commercial SCLs. This approach leads to synthesized storage arrays consisting of dynamic latches, which we refer to as dynamic SCMs (D-SCMs) in the following. D-SCMs keep all the advantages of static SCMs compared to SRAM macrocells, as discussed earlier in Section 2.1, except for the straightforward portability among technology nodes and semiconductor fabs. In fact, the custom-designed standard-cells need to be designed again in each new target technology. However, this small, additional design effort is easily justified by the tremendous area savings which D-SCMs enable compared to the use of static SCMs, as will be seen in an application example in Section 2.3.2.

Dynamic Latch Topologies

Fig. 2.8a shows a commonly used static latch topology consisting of 16 transistors. This topology with an inverter and a transmission gate on both the data input-to-output path and the internal feedback path was traditionally (and still is) heavily used by many commercial SCL providers. In advanced, aggressively scaled CMOS nodes where leakage current becomes an increasingly dominant problem, and where leakage power becomes significant compared to switching power, it is more common to replace the inverter and the transmission gate with a single tri-state inverter for the benefit of lower leakage current at an equal transistor count (i.e., a comparable silicon area cost) and a comparable robustness.

As opposed to such static latch topologies, a variety of dynamic latch topologies are discussed next, focusing on their area cost, reliability, and ease of integration into a digital design flow. First of all, the highly robust, general-purpose, commonly used static latch topology in Fig. 2.8a is converted into the dynamic 8-transistor (8T) latch topology in Fig. 2.8b by removing the keeper part and the second clock inverter. Without the keeper part which resembles an SRAM bitcell during the non-transparent phase of the latch, data is now stored in form of charge on the parasitic storage node (SN) capacitance, which is primarily formed by gate (MOSCAP), diffusion, and interconnect parasitic capacitance. The deletion of the second clock inverter is justified by the fact that the additional capacitive load to be driven by the clock network, i.e., the capacitive input load of the clock (CK) port, is only small: one additional transistor's gate capacitance compared to two in case of the static latch. Except for its dynamic storage mechanism, the 8T dynamic latch topology is still very robust due to a number of reasons:

1. The full transmission-gate is able to transfer full high and low logic voltage levels to the SN, i.e., the levels are not deteriorated due to voltage drop across a single pass transistor. Even charge injection from the PMOS and NMOS device and clock-feedthrough occurring while the latch changes from the transparent to the non-transparent phase are well balanced and result only in a small voltage disturb on the SN.

2. There is an inverter at the data input which ensures a strong drive of the SN through the transmission-gate. Otherwise, the SN might be driven only weakly through an a-priori unknown, complex, distributed RC network external to the latch. Moreover, the input capacitance of this latch is constant, which simplifies the characterization of the cell and its integration into a standard-cell based synthesis flow.

Of course, as in case of static latches, the 8T dynamic latch shown in Fig. 2.8b can also be implemented with a tri-state inverter for lower leakage currents. Furthermore, depending on the process design kit (PDK), various threshold voltage (V_T) options, including low-V_T (LVT), standard-V_T (SVT), and high-V_T (HVT), can be explored. For example, in order to improve the data retention time and eventually the read robustness at the cost of a longer setup time, HVT transistors can be used in the transmission gate instead of SVT or LVT transistors (which are preferably used for short setup times and short insertion delays).

While all 8T latch topologies are still relatively large and have a strong SN drive, smaller yet less reliable topologies are introduced next, namely 5-transistor (5T) and 3-transistor (3T) dynamic latches. Using a single pass transistor, either an NMOS or a PMOS device, instead of the full transmission gate allows to get rid of the clock inverter, and saves also one transistor from the transmission gate, leading to the 5T topology shown in Fig. 2.8c (example of an NMOS pass transistor). However, the reduced silicon area comes at the cost of a degraded SN drive. In case of an NMOS pass transistor, it is difficult (or even impossible) to transfer a strong logic '1' level to SN in a short time, due to the threshold voltage drop across the NMOS device. Similarly, a PMOS pass transistor cannot pass a strong logic '0' level to the SN in a short time. Gate overdrive (above V_{DD}) for an NMOS pass transistor and gate underdrive (below ground) for a PMOS pass transistor would remedy the threshold voltage drop problem. However, this technique is not adopted here to allow a simple integration of the latches into synthesized SCMs with a single power supply. A potential problem of the 5T latch topologies with a single pass transistor is also the occurrence of short-circuit currents in the output buffer: with a weak '1' or a weak '0' level on the SN, the PMOS or the NMOS transistor in the output buffer is on the edge of turning on, respectively, while the complementary transistor is already turned on as well. Within-die process parameter variations can aggravate this problem; for example, a large (larger than nominal) V_T of the NMOS pass transistor in combination with a small (smaller than nominal) V_T of the PMOS device in the output buffer is likely to result in short-circuit current already early. Of course, using an LVT pass transistor and HVT transistors in the output inverter provides a comfortable margin for charge leakage from SN before the onset of short-circuit current. However, LVT transistors are typically so leaky that the minimum required retention time (several tens of ns, for systems like [50]) cannot even be achieved. Often, in most CMOS technologies, PMOS devices have a higher absolute value of V_T than NMOS devices. Therefore, it is easier to avoid short-circuit current in case of using an NMOS pass transistor, as opposed to using a PMOS pass transistor. The probability and especially the magnitude of a short-circuit current in the output inverter are reduced if it is implemented with HVT devices.

Finally, in order to further reduce the area cost, the cell-internal SN driver (or the input inverter) can be removed, which results in a 3T latch as shown in Fig. 2.8d (example of using SVT devices and an NMOS pass transistor). For the 3T latches, the driver of the SN is shared between all latches in the same column of the SCM array. If many latches connect to the same bit line (BL), it is challenging to drive the SN of a given latch through a complex RC network in a short time. Moreover, the input impedance of the 3T latch depends on the clock phase: a single diffusion capacitance for the low clock phase, and a C-R-C network for the high clock phase. This property complicates the characterization of the cell and its integration into a standard digital design flow.

D-SCM Versus SRAM Area Comparison

Fig. 2.9 shows how the integration of custom-designed, dynamic latches into SCMs affects the area comparison of SCMs with 6T-bitcell SRAM. Recall from Fig. 2.4 in Section 2.2.1 that static SCMs are smaller than corresponding SRAM macrocells for storage capacities up to around 1 kb. While the transistor count of a standard-cell does not directly represent its silicon footprint, we found by layout drawing in various technology nodes that the area of the 8T dynamic latch topology is indeed around 50% of the area of the 16T static latch topology. This reduced storage cell footprint favors the SCM versus SRAM macrocell comparison: as shown in Fig. 2.9, the 8T-bitcell D-SCM architecture is smaller than corresponding SRAM macrocells for storage capacities up to around 2 kb. The area comparison does further evolve in favor of SCMs for the 5T and 3T dynamic latches, which comes, however, at the cost of lower circuit reliability and more challenges for the integration into the digital design flow. The 3T dynamic latch can even be smaller than a 6T SRAM bitcell, promoting D-SCMs which are smaller than SRAM macrocells irrespective of the storage capacity. The following Section discusses in detail the integration of D-SCMs using a 3T dynamic latch into the previously discussed LDPC decoder and quantifies the area (and power) savings resulting from such a custom-designed standard-cell.

2.3.2 Application Example: LDPC Decoder with Refresh-Free D-SCMs

In this Section, we reconsider the same low-power LDPC decoder architecture as before in Section 2.2.2, analyze the access patterns of all internal memories, and demonstrate that the static SCMs can be substituted with D-SCMs for dramatically improved area-efficiency. In fact, all memories are frequently and periodically written with new data, which allows us to use area-efficient D-SCMs even without the need for power-hungry refresh cycles. The D-SCMs are designed to retain data just long enough to guarantee reliable circuit operation. Note that the use of refresh-free dynamic memories leads to the requirement for a minimum operating frequency. The low-power LDPC decoder architecture with refresh-free D-SCMs was implemented in a 90 nm CMOS process, and silicon measurements show full functionality and an information bit throughput of up to 600 Mbps (as required by the IEEE 802.11n standard). Silicon measurements show an improved energy metric (energy per bit per iteration), as well,

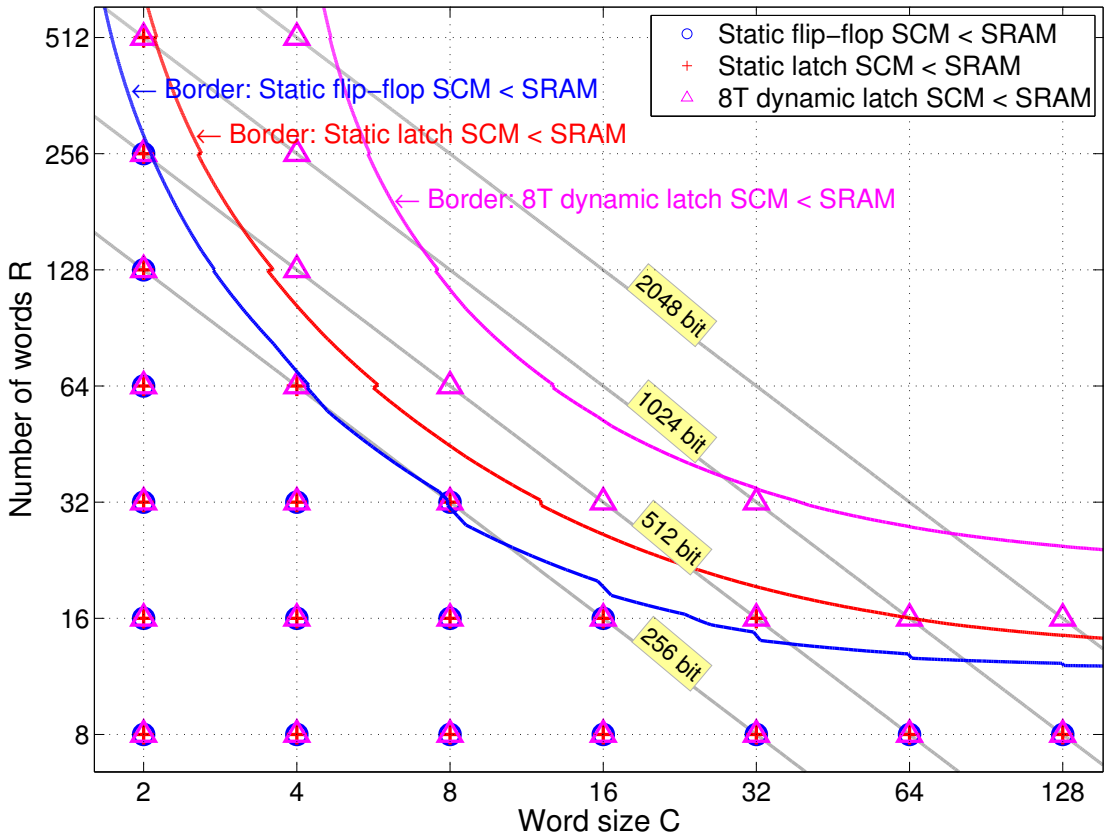


Figure 2.9: Area efficiency of 1) static flip-flop SCM (blue); 2) static latch SCM (red); and 3) 8T dynamic latch SCM (magenta) compared to 6T-bitcell SRAM macrocells. SCM implementations below the blue, red, and magenta lines are smaller than corresponding SRAM macrocells.

compared to a the previous implementation based on static SCMs.

As a reminder from Section 2.2.2, replacing conventional SRAM macrocells with SCMs was shown to entail a considerable 37% power reduction due to the ability to merge memories with logic, better data locality, less routing, and consequently lower active power consumption. However, the energy savings provided by SCMs came at the cost of an increased decoder area; in fact, the silicon area of the decoder became 50% larger compared to the case of using SRAM hardmacros. As a further alternative to SRAM macrocells, a recent work [50] proposes to use gain-cell based eDRAMs in a high-throughput LDPC decoder. These eDRAM macrocells can be operated without a refresh operation and lead to an overall better area and energy efficiency. In this Section, we propose to combine all the advantages of SCMs (see Section 2.1), especially the high data locality and the low switching power, with the high storage density of dynamic bitcells. This approach reduces the area penalty of previous LDPC decoders using static SCMs (such as [57, 2]), and can be safely adopted even without the need for explicit refresh cycles. Refresh-free operation is possible as the write and read access statistics of all internal memories of the considered LDPC decoder are known *a priori* to exhibit frequent

write updates.

In the following, the architecture of the previously presented LDPC decoder is first reviewed in more detail in order to properly understand the memory access patterns and the required retention times, before the dynamic bitcell design providing just enough data retention time is expatiated on. The Section closes by presenting silicon measurement results of the LDPC decoder using D-SCMs and by comparing it with prior-art implementations.

QC-LDPC Decoder Architecture

LDPC codes and in particular quasi-cyclic (QC)-LDPC codes [58, 59] are among the most popular and capable error-correcting codes adopted in many modern standards including DVB-S2 [60] and IEEE 802.11n [61]. The decoding of a QC-LDPC code is in general performed by iterative message passing between variable nodes which represent the code bits and check nodes which represent the parity check equations of the code-specific parity-check matrix **H**. The messages going from variable nodes to check nodes are denoted as Q-messages and the messages exchanged in the other direction as R-messages. In addition, an L-value is associated with each variable node representing the reliability information for the corresponding code bit in the form of an estimate of the a-posteriori log-likelihood ratio (LLR). In this work, we use an LDPC decoder based on the offset-min-sum (OMS) message-update rules combined with the layered decoding schedule in order to profit from a good balance between convergence speed and VLSI implementation complexity [59, 37].

Architecture Details The considered LDPC decoder architecture [2] is shown in Fig. 2.10. The decoder starts by initializing the L-memory with the initial LLRs of the code bits obtained from the baseband receiver and continues with the sequential processing of the loaded parity-check matrix. To this end, Z node computation units (NCUs) sequentially execute the OMS algorithm for each layer of **H**, where Z denotes the number of parity-check equations per layer. Each NCU follows a two-step procedure. In the first step, the MIN unit iteratively computes all Q-messages and other intermediate data of the current layer using the corresponding R-messages and L-values from the previous iteration. During this process, the cyclic shifter shifts the Z successive L-values fetched from the L-memory according to the quasi-cyclic property of **H** in order to feed all MIN units with the proper values. In the second step, the SEL unit iteratively updates the R-messages and L-values based on the old R-messages and on the buffered Q-messages and intermediate data provided by the MIN unit. This process is repeated for all layers of **H** and until a predefined number of iterations has been reached or an online stopping-criterion has been triggered. As shown in Fig. 2.10, several NCUs are grouped together with the corresponding Q-memory and R-memory sub-blocks in order to maximize data locality [2].

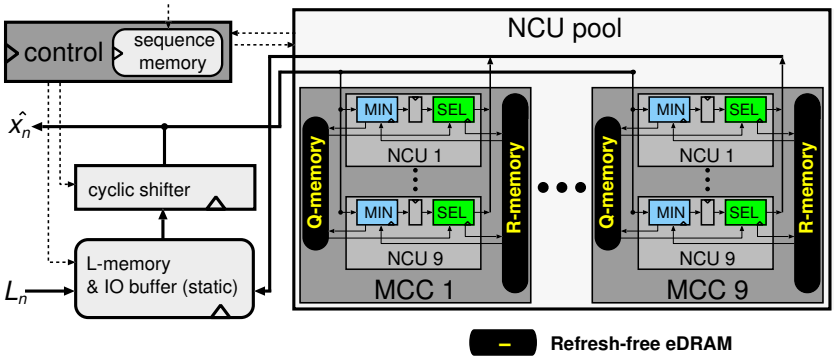


Figure 2.10: Architecture of the quasi-cyclic LDPC decoder with refresh-free dynamic memories (highlighted in yellow).

Table 2.5: Memory sizes, retention times, and update rates.

	R Memory	L Memory	Q Memory
Size [bits]	35640	9720	9720
t_{ret} [ns], clock cycles	287.8, 88	287.8, 88	78.5, 24
t_{up} [ns], clock cycles	287.8, 88	287.8, 88	287.8, 88

Memory Requirements and Characteristics Interestingly, we observe that the Q- and R-messages as well as the L-values need to be stored only for a short time, before the corresponding memories are updated again with new data, which allows us to use refresh-free dynamic storage elements. Three types of memories are required in the considered decoder architecture: the R, Q, and L memories store the R-messages, Q-messages, and L-values, respectively. The total size requirement of each memory type for the highest-rate parity-check matrix with $Z = 81$ specified for the IEEE 802.11n standard is shown in Table 2.5. In order to enable refresh-free operation, the memories are characterized according to the following definitions: 1) the *retention time* t_{ret} denotes the time interval between the first write access and the last corresponding read access to a memory block; and 2) the *update rate* t_{up} is defined as the time interval between a write access to a word and the next write access to the same word. Note that at a time t_{ret} after writing, all addresses must still read out correctly while up to a time t_{up} after writing, the data levels in the dynamic storage cell should still be strong enough to avoid short-circuit currents (unless they can be avoided by circuit techniques). Table 2.5 shows the retention time (t_{ret}) requirements as well as the effective, guaranteed update rates (t_{up}) of all memory types contained in the QC-LDPC decoder architecture. The table assumes an operating frequency of 305.8 MHz which is necessary to achieve an information bit throughput of 600 Mbps, required by the highest-rate mode of the IEEE 802.11n standard.

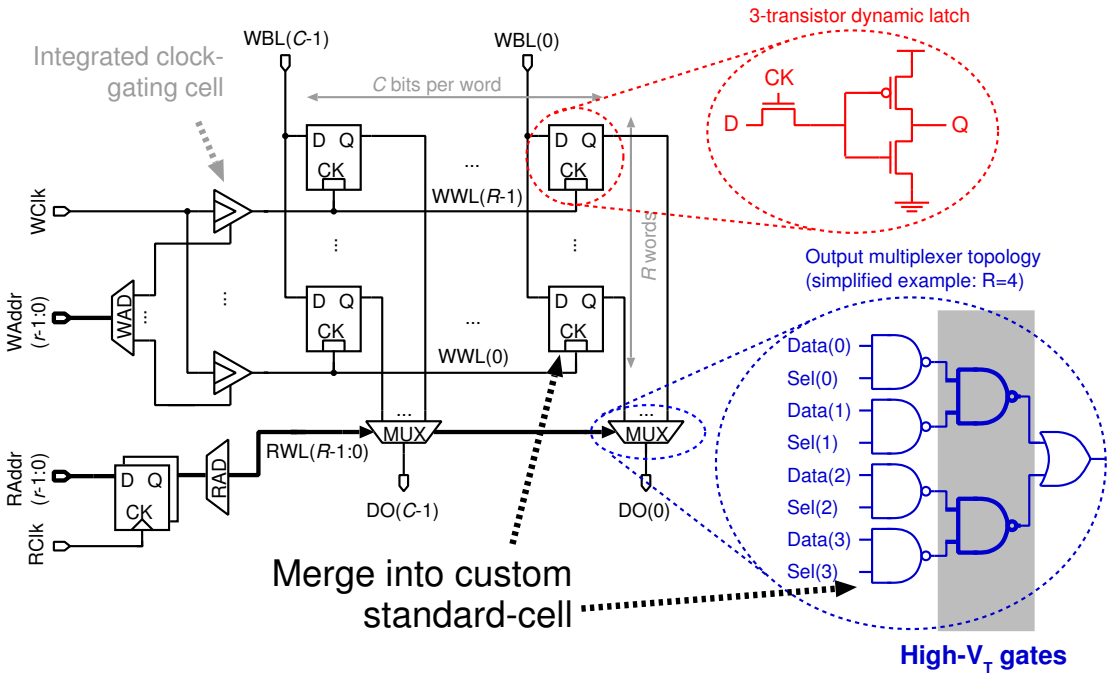


Figure 2.11: Architecture of dynamic standard-cell based memory (D-SCM).

Dynamic Standard-Cell Based Memory Design

As explained earlier in Section 2.2, an SCM architecture based on latches as basic storage cells, integrated clock-gating cells for the generation of write select pulses, and static CMOS multiplexers for the readout of the selected word is most suitable in terms of area efficiency, power consumption, and speed. This SCM architecture is drawn again in Fig. 2.11. Due to the frequent write updates, a custom-designed dynamic latch is proposed as basic storage cell rather than a commercially available static latch. In order to aggressively push for minimum area, a 3-transistor (3T) dynamic latch topology is adopted as starting point, as shown in Fig. 2.11 in the top-right corner. To further improve the area efficiency, the 3T latch is merged with the first stage of the read multiplexer, namely a NAND gate, into a single, custom-designed standard-cell. The conceptual schematic of this standard-cell is shown in Fig. 2.12(a).

As a protection mechanism against excessive leakage in case of potentially weak output levels of the dynamic storage cell, the second stage of the readout multiplexer (i.e., all logic gates directly following the basic storage cell with NAND functionality) is implemented with high threshold-voltage (high-V_T) gates, as shown in Fig. 2.11. Since only one cell in a long combinatorial path is replaced with a high-V_T cell the impact on the speed is negligible.

Bitcell Optimization to Avoid Short-Circuit Currents The initial cell shown in Fig. 2.12(a) uses a single NMOS transistor to transfer a logic level from the write bit-line (WBL) to the storage node (SN) as soon as a write operation is initiated by rising the write word-line (WWL).

While logic '0' levels are properly transferred to the SN, logic '1' levels are degraded by the threshold voltage drop across the NMOS write transistor (MW), as we do not use a WWL overdrive voltage for straightforward integration of this cell into a design with a single core supply voltage. Charge injection and clock feedthrough further deteriorate the logic '1' level during de-assertion of the WWL. These deteriorated logic '1' levels bare the risk for short-circuit currents during readout of the cell, i.e., as soon as the read word-line (RWL) is asserted and goes high. To avoid such excessive short-circuit currents which would last for an entire clock cycle, the PMOS transistor connected to SN is removed, as shown in Fig. 2.12(b). This results in a cell that operates similarly to domino logic [30, 62]: prior to a read access, the output node Q is precharged to V_{DD}, since at that time the RWL is still de-asserted and low. During the read access, the RWL is high, and the output node Q is safely discharged even with a deteriorated, weak logic '1' level on the storage node, while the node Q remains in its pre-charged state if SN holds a logic '0'. In addition to avoiding short-circuit current during read, there is no risk for short-circuit currents during non-read cycles (including potential standby times of the LDPC decoder) either. In fact, the output node Q of the domino-like dynamic bitcell is always properly charged to V_{DD} during non-read cycles, which circumvents short-circuit currents in its output stage and in subsequent logic gates. This property distinguishes the presented cell from conventional, dynamic memory and logic cells.

Increasing Read Robustness Transistor MSN in Fig. 2.12(b) suffers from the body effect: its positive source-to-body voltage V_{SB} increases its threshold voltage V_T, which aggravates the readout process of an already deteriorated logic '1'. Similarly to a common practice in gain-cell based eDRAM design [63], adding a coupling capacitor in form of a MOS capacitor (MCP) between the SN and the RWL, as shown in Fig. 2.12(c), was found to considerably improve the read '1' robustness of our bitcell, as well. The positive RWL transition during the onset of a read operation couples onto the SN and temporarily rises the SN level, thereby strengthening the logic '1' and leading to a faster read operation. Note that this MOS capacitor exhibits a channel formation during a write '1' operation, while it turns off for a write '0' operation. Therefore, the more critical '1' level preferentially receives a larger SN capacitance and SN boost during readout, whereas a logic '0' level is hardly affected by the additional MCP device (only the gate-over-diffusion overlap capacitors are added to the SN).

Simulation Results The final cell shown in Fig. 2.12(c) has been extensively simulated and verified under pessimistic assumptions prior to tape-out. By assuming that the state of WBL is always opposite to the data stored on SN, a worst-case memory access condition is created. Moreover, the simulated readout always occurs after the maximum considered retention time of 287.8 ns. Furthermore, temperatures of up to 50°C and a supply voltage of 1.0 V are considered. Under these conditions, Monte Carlo simulations accounting for local, within-die parametric variations in different global process corners indicate robust read '1' (and read '0') operations. The layout of the basic storage cell with NAND functionality is shown on the left-hand side of Fig. 2.13. Compared with a minimum-drive, minimum-size, static latch

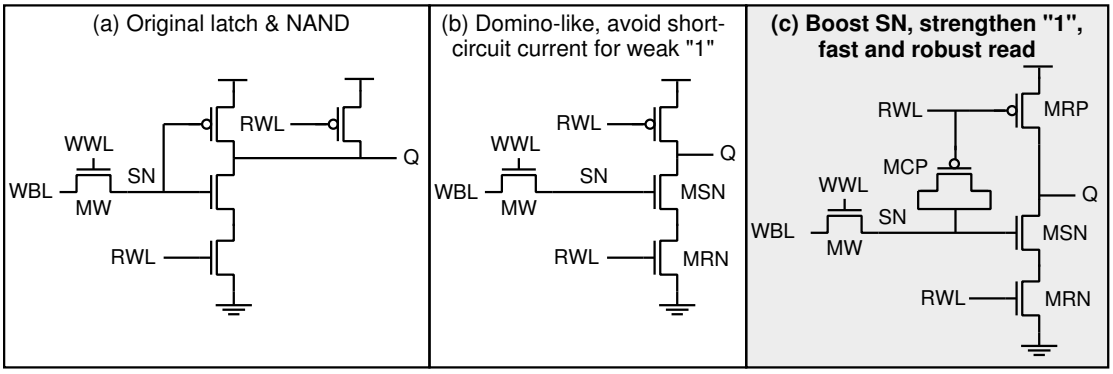


Figure 2.12: Design exploration of custom standard-cells combining dynamic latch and NAND functionality.

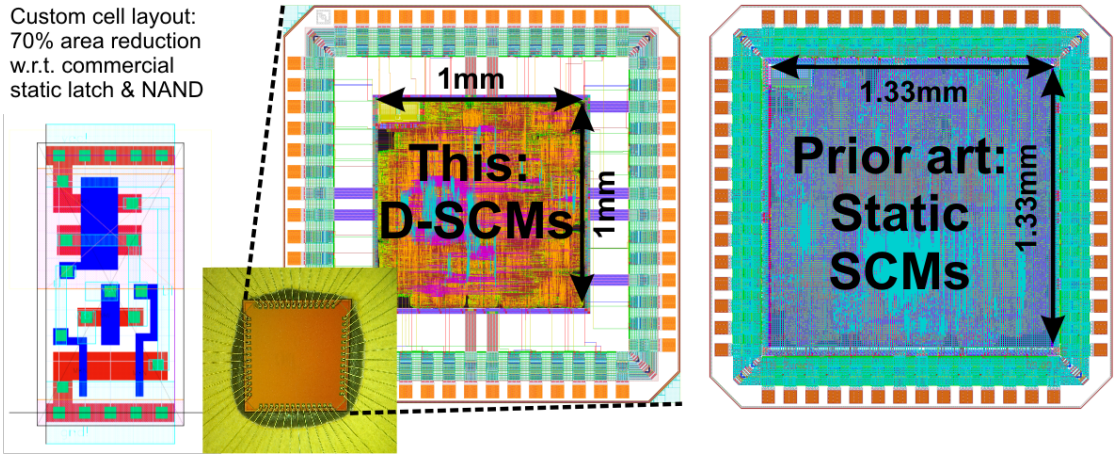


Figure 2.13: (Left) Layout of custom standard-cell; (Middle) Chip microphotograph and layout picture of the proposed LDPC decoder using D-SCMs; and (Right) Layout picture of the same LDPC decoder architecture using static SCMs.

and NAND gate from a commercial standard-cell library, the silicon area of the proposed, custom-designed, multifunctional standard-cell is reduced by 70%.

Silicon Measurement Results

The above-described QC-LDPC decoder was manufactured in a 90nm CMOS technology. A chip microphotograph and a complete layout picture of the decoder core, surrounded by a pad-frame are shown in the middle of Fig. 2.13. A total of 8 packaged dies were verified on a HP93000 digital tester; all measured dies were fully functional within the expected voltage and frequency range.

Frequency and Voltage Characterization Fig. 2.14 shows the percentage of failing chips as a function of the frequency and the supply voltage V_{DD}. As expected, there is a maximum and a

minimum operating frequency, together defining a frequency range for valid circuit operation. The maximum frequency is determined by the critical path delay and decreases with the supply voltage. There is a sharp transition from 0 to 100% failing chips, which means that die-to-die variations between the 8 measured dies (all from the same wafer) do not significantly affect the critical path delay. The need for a minimum operating frequency (below which the chips fails) arises from the dynamic memories, which are designed to retain data only for the minimum required time of 287.8 ns. We observe a rather slow transition from 0 to 100% failing chips when gradually slowing down the clock frequency for a given V_{DD}. Compared to a few critical timing paths whose delays are determined by the transistor's on-current (I_{on}) that varies only slightly from die to die, the minimum retention time of the dynamic storage cells is determined by several leakage mechanisms and is much more sensitive to parametric variations. This behavior is well aligned with previous reports on gain-cell based eDRAMs whose retention time is very sensitive even to within-die parametric variations [64]. Supply voltage scaling has two complementary effects on the retention time of the considered dynamic bitcell: 1) weakened leakage currents (e.g., the subthreshold conduction of MW decreases with V_{DS}, which in turn decreases with V_{DD}); and 2) lower noise margins (i.e., less headroom for deterioration of logic storage levels due to leakage). According to the measurements shown in Fig. 2.14, the weaker leakage currents at lower V_{DD} are the dominant effect, allowing longer retention times and lower frequencies at lower V_{DD}. The same behavior, i.e., improved retention times at scaled voltages, has also been observed in logic-compatible, gain-cell based eDRAMs [65] (see Section 4.2 for more details). For all voltages between 0.8 and 1.2 V, there is a large range of frequencies where all measured LDPC decoder chips function correctly. Within these admissible voltage and frequency ranges, the decoder supports different throughput modes, as exemplified by the markers in Fig. 2.14.

Comparison with Prior-Art Implementations The 70% area reduction of the multifunctional, dynamic standard-cell results in a considerable 44.4% reduction in the area cost of the LDPC decoder[1], compared to its previous implementation with static SCMs. In fact, the core size of the proposed decoder is only 1.00 mm^2, while it is 1.77 mm^2 with static SCMs, as shown on the right-hand side of Fig. 2.13. According to post-layout simulations, the dynamic storage cell leads to a 31.0% total power reduction in the R memory, and a 15.4% power reduction in the Q memory. The modified bitcell has a higher impact on the bigger R memory (88×45 bits) than on the small Q memory (24×45 bits) whose power consumption is more influenced by peripheral circuits. These power savings at the memory level are reflected in a simulated 11.3% power reduction at the LDPC decoder level.

The leakage current of the presented decoder architecture is dominated by the leakage current

[1] Note that the L memory is not only used during decoding, but also as an I/O memory to load data to and from the decoder chip. During this I/O operational phase, it requires much higher retention time than during decoding. For this reason, the L memory was implemented as a static memory. However, using an extra SRAM to handle I/O operations (which are not part of the decoding), the L memory could be implemented with dynamic bitcells, as well, and the decoder area would be even smaller.

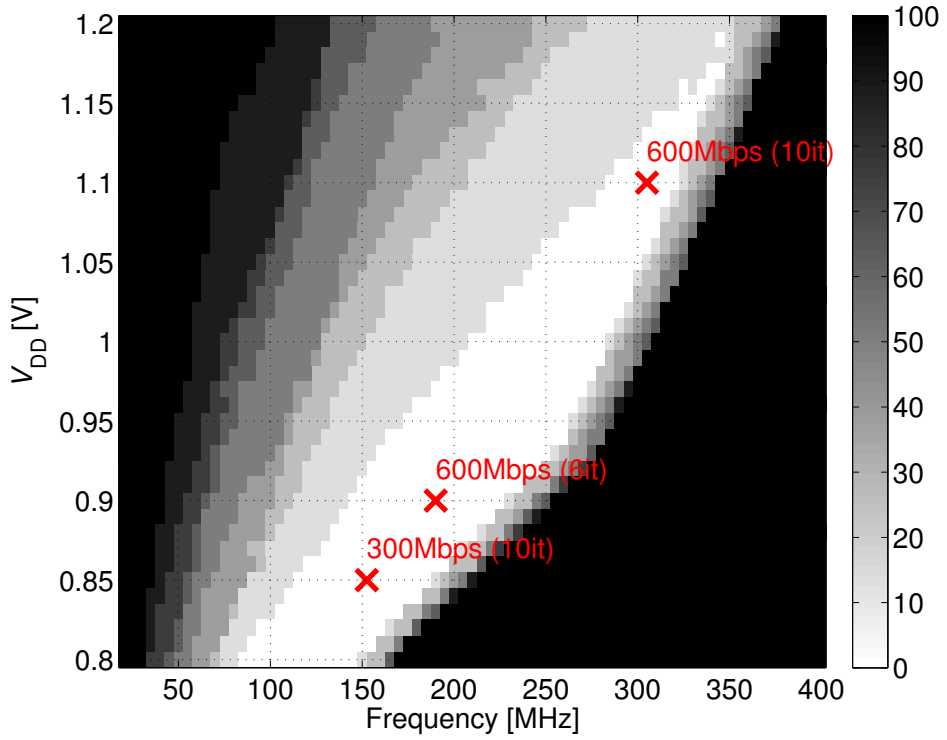


Figure 2.14: Percentage of failing chips as a function of the frequency and V_{DD}.

of the embedded memories. Replacing the static SCMs with the proposed D-SCMs (which have a built-in mechanism to avoid short-circuit currents) results in an average decoder's leakage current reduction of 55% compared to the decoder using static SCMs, as illustrated in Fig. 2.15 presenting silicon measurement results. However, as the leakage current is small compared to the switching current, and as all parts other than the basic storage cell remain unchanged, the proposed decoder implementation exhibits only a small total power reduction of 5.5% on average (among all measured dies) compared to the same decoder architecture using static SCMs. The corresponding, average decoding energy is 14.7 pJ/bit/iteration (measured at 1.0 V, 305.8 MHz, for 10 iterations, computed over the coded throughput, 600 Mbps information bit throughput, averaged over 8 dies) in case of D-SCMs and 15.5 pJ/bit/iteration in case of static SCMs. Finally, as shown in Table 2.6, the proposed LDPC decoder is compared with a selection of the best—in terms of hardware efficiency A [mm^2/Gbps] and energy efficiency E [pJ/bit/iter]—, recent, silicon-proven LDPC decoders for the IEEE 802.11n or the WiMAX standards. All metrics are scaled to the 90 nm CMOS node and reported in parenthesis, in addition to the original values. The proposed decoder compares favorably with prior art by achieving both good hardware and energy efficiency. Only one work [66] has slightly better hardware efficiency, at the cost of worse energy efficiency.

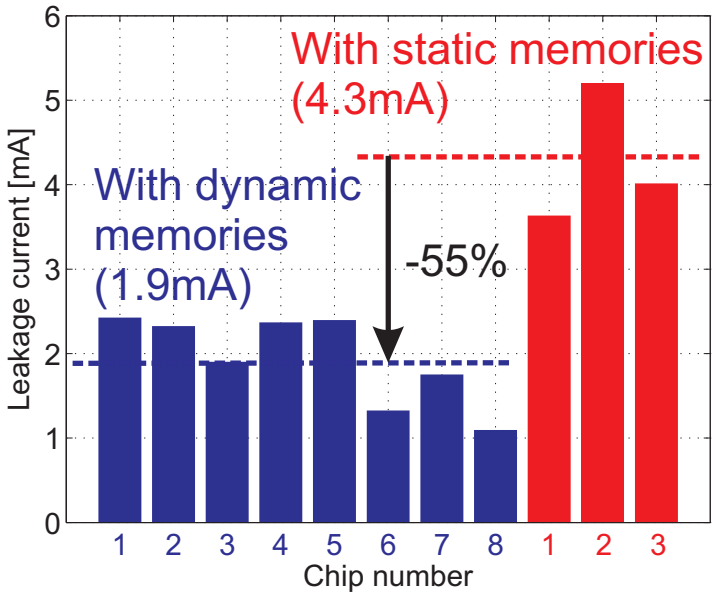


Figure 2.15: Leakage current comparison of the proposed QC-LDPC decoder implementation based on D-SCMs (8 measured dies, blue bars) with the same decoder architecture using static SCMs (3 measured dies, red bars) [2].

Table 2.6: Comparison with prior-art LDPC decoder implementations.

Publications	[2]	[41]	[66]	[67]	**This**
Technology [nm]	90	180	130	65	90
V_{DD} [V]	1.0	1.8	1.2	1.2	1.1
Core area [mm^2]	1.77	14.3	3.03	3.36	1.0
Maximum throughput T [Mpbs]	679	640 (1280)	728 (1052)	1056 (763)	600
Hardware efficiency A [mm^2/Gbps]	2.6	22.3 (2.8)	4.16 (1.4)	3.2 (2.2)	**1.7**
Energy efficiency E [pJ/bit/iter]	15.8	123 (19.0)	39 (18.8)	10.9 (15.1)	**14.7**

Scaling to 90 nm, 1.0 V: $T \sim s$, $A \sim 1/s^2$, $E \sim 1/s \cdot (1.0\,\text{V}/V_{DD})^2$

2.4 Conclusions

In this Chapter, it was shown that SCMs can bring various benefits compared to SRAM macro-cells, such as ease of portability (especially if working only with commercial standard-cell libraries), modifications at design time, ability to merge storage with logic, potentially less routing, lack of separate voltage supply rings, and more flexibility for fine-granular memory organizations. As for the write logic of SCMs, using basic flip-flops or latches as storage cells in conjunction with clock-gates leads to smaller area and lower power consumption than using flip-flops or latches with enable feature. As for the read logic, multiplexer based implementations lead to smaller area and lower power consumption than tri-state buffer based implementations. Latch based SCMs are only slightly smaller than flip-flop based SCMs.

Flip-flop based SCMs, however, are more convenient for high-speed applications than latch based SCMs.

In our first case study, a low-power LDPC decoder, which has 9 memory blocks with capacity > 3kb, becomes bigger when replacing SRAM macrocells with SCMs, but its power consumption is significantly reduced. Besides post-layout circuit simulations, this result is verified by means of silicon measurements of a LDPC decoder ASIC manufactured in a 90 nm CMOS technology. Back in 2010, the proposed LDPC decoder architecture employing SCMs achieved the best energy-efficiency across comparable designs reported in the open literature. For applications requiring memories with storage capacity < 1kb, replacing SRAM macrocells by SCMs can be profitable for both power and area.

The introduction of custom-designed, dynamic latches is an efficient way to address the area bottleneck of SCMs synthesized from commercial standard-cell libraries, especially for VLSI systems which require only short data retention times and/or which can tolerate a small amount of hardware defects (such as many wireless communications systems). A robust dynamic latch topology uses 8 transistors, while the smallest possible topology uses only 3 transistors. The large 8T dynamic latch topologies are still rather robust, while the smaller 5T and 3T topologies become more and more error-prone (in terms of retention time and read failures) and are always more difficult to integrate into a digital, standard-cell based design flow. However, the 3T dynamic latch can be smaller than a 6T SRAM bitcell, and dynamic SCMs (D-SCMs) based on a 3T storage cell can be smaller than SRAM macrocells irrespective of the storage capacity (while static SCMs are smaller than SRAM macrocells only up to around 1 kb).

In our second case study, all embedded memories of the previously presented low-power LDPC decoder are implemented using area-efficient, dynamic storage cells, operated without refresh cycles due to frequent and periodic write updates. At the decoder level, the newly proposed and seamlessly integrated dynamic, standard-cell based memories lead to a silicon area and leakage current reduction of 44.4% and 55.0%, respectively. The proposed multifunctional, dynamic storage cell avoids short-circuit currents by changing the read logic from CMOS to domino style and is optimized for robust read by inserting a coupling capacitor between the storage node (SN) and the read word-line (RWL). Beside the considerable area reduction, the total power consumption of the decoder is reduced by 5.5%. A potential drawback of the proposed decoder is the need of a minimum operating frequency, below which the refresh-free dynamic storage elements start to loose their data. However, all measured dies have a large range of safe operating frequencies compatible with various throughput modes. The manufactured and silicon-proven LDPC decoder exhibits a core area of 1.0 mm^2 in a 90 nm CMOS node, dissipates an energy of 14.7 pJ/bit/iteration, and runs at all frequencies from 85 to 345 MHz for a voltage range from 0.8 to 1.2 V.

In summary, SCMs are a straightforward approach and interesting alternative to SRAM macrocells for the implementation embedded memories, especially for small, distributed memory

blocks of several kb. SCMs work reliably in any target system, even at aggressively scaled voltages (see Chapter 3 for more details) and in the most advanced, deeply scaled, nanometric CMOS nodes. In fact, as soon as a standard-cell library for digital design is available in such a node, it is also possible to synthesized SCMs. High-density dynamic SCMs with retention times of several hundreds of ns were successfully demonstrated in a 90 nm CMOS node; such D-SCMs can still be used in more deeply scaled CMOS nodes for temporary data storage, but the retention times will be even lower due to higher leakage currents, unless adopting aggressive leakage reduction techniques (e.g., using high-V_T transistors) or using metal stacks to increase the storage node capacitance.

3 Ultra-Low-Power Standard-Cell Based Memories (SCMs)

Devices such as hearing aids, medical implants [68], and remote sensors impose severe constraints on size and energy dissipation. Supply voltage scaling is an efficient low-power technique which reduces both active energy dissipation and leakage power [69]. When applied aggressively, voltage scaling leads to sub-threshold (sub-V_T) operation [70]. In this regime, severely degraded on/off current ratios I_{on}/I_{off} and increased sensitivity to process variations are the main challenges for sub-V_T circuit design [71] in 65 nm CMOS technologies and below.

As an alternative to variation-tolerant full-custom circuit design, [72, 73, 74] promote the design of sub-V_T circuits based on conventional standard-cell libraries. In such conventional standard-cell based designs, embedded memory macros may limit the scalability of the supply voltage[1], and thus the minimum achievable energy per operation, as the noise margins gradually decrease with the supply voltage, which leads to write and read failures in the sub-V_T regime [75], or even already in the near-threshold (near-V_T) domain.

The main options for embedded memories which may be operated reliably in the sub-V_T domain are: 1) specially designed SRAM macros; and 2) standard-cell based memories (SCMs). Standard 6-transistor (6T)-bitcell SRAM designs require non-trivial modifications to function reliably in the sub-V_T regime [6, 76, 71, 77, 78, 79, 80]. However, SCMs, originally intended for above-V_T operation (see Chapter 2), and easily synthesized with standard digital design tools may directly be adopted in the sub-V_T domain, where they are still fully functional.

While Chapter 2 has investigated SCMs operated at nominal supply voltage and for use in high-performance VLSI systems such as channel decoders for wireless communications, this Chapter focuses on ultra-low-power SCMs operated at aggressively scaled voltages, typically residing in the subthreshold (sub-V_T) regime, for use in ultra-low power systems such as wireless sensor nodes or biomedical implants. Again, in a first step, for short design times and straightforward implementation in any technology node, all previously introduced SCM architectures (see Chapter 2) based exclusively on commercial standard-cell libraries (SCLs)

[1] Of course, it is also possible to operate the embedded memories at a higher voltage than the logic blocks, which, however, requires an additional power distribution network and the insertion of level shifters.

are evaluated and compared in the sub-V_T domain. Then, in a second step, the integration of custom-designed, ultra-low leakage standard-cells for even lower SCM leakage power and access energy is proposed. Finally, even though the sub-V_T compilation flow using low-leakage cells yields unprecedentedly low standby leakage power and access energy in a 65 nm node, non-volatile flip-flop (NVFF) topologies based on emerging memory device technology (oxide stacks, OxRAM, or "memristors") are investigated to enable zero standby leakage power for future ultra-low power VLSI systems; for the first time, we propose an OxRAM-based NVFF topology with sub-V_T read operation.

The remainder of this Chapter is structured as follows. Section 3.1 explains the various failure mechanisms of 6-transistor (6T)-bitcell SRAM under scaled supply voltages and reviews alternative SRAM bitcells (consisting of 8, 10, or more transistors) operating reliably at scaled voltages, as well as various low-voltage write and read assist techniques. Next, Section 3.2 proposes sub-V_T SCMs based on commercial SCLs as an affordable, straightforward, and interesting alternative to custom-designed sub-V_T SRAM macrocells; first of all, various sub-V_T design strategies applicable to any digital design and to stand-alone SCMs entities are quickly reviewed, before a detailed comparative analysis of all SCM topologies operated in the sub-V_T domain is presented. After identifying the best-practice SCM topology using commercial SCLs, further optimizations for ultra-low leakage power and access energy achieved by standard-cell customization are presented in Section 3.3; silicon measurement results from various test chips are presented in this Section, as well. Finally, Section 3.4 presents non-volatile flip-flop topologies capable of operating in the sub-V_T regime (except for the write operation) for zero standby leakage, before Section 3.5 concludes this Chapter.

This Chapter is mostly based on our previous publications [33, 81, 82, 7, 83].

3.1 Challenges and Review of Prior-Art Low-Voltage SRAM Design

As SRAM has been the mainstream solution for embedded memories for many decades, there has been a considerable amount of research on improving yield and robustness of SRAM arrays operated under scaled supply voltages (including sub-V_T voltages) or implemented in aggressively scaled CMOS nodes. Many new SRAM bitcell designs and various low-voltage write and read assist techniques have been proposed to deal with a series of problems which conventional 6-transistor (6T) SRAM suffers from: 1) write failures; 2) read failures; 3) hold failures; and 4) read-access time failures [84, 85]. All these failures are primarily caused by process parameter variations and are seriously aggravated by voltage scaling. Write failures result from the incapability of switching the SRAM cell due to an unusually strong PMOS keeper device, while read failures arise from the voltage dividing effect between the access device and the NMOS keeper device which may switch the cell while reading in the occurrence of within-die (WID) process parameter variations. Hold failures represent the inability of keeping the content of a bitcell under typically aggressively scaled supply voltages during standby modes. Read-access time failures result from the inability of reading data in a previously

define maximum access time, and are less critical than the other three failure mechanisms for ultra-low voltage (ULV) systems operating at moderate frequencies. Unfortunately, optimizing (by transistor sizing) a conventional 6T SRAM bitcell for good write-ability has a negative impact on the read-ability, and vice versa. In other words, improving the write-ability and the read-ability of a conventional 6T SRAM bitcell are conflicting requirements [84].

Several innovative SRAM bitcell topologies dealing with the above mentioned read and write failure mechanisms have been proposed in the recent years [31]. For example, the well-known 8-transistor (8T) bitcell shown in Fig. 3.1a includes a separate read buffer to avoid the voltage dividing effect [3], thereby improving read-ability. Moreover, a 9-transistor (9T) bitcell (see Fig. 3.1b) uses, in addition to the read buffer, a cell-internal supply feedback transistor to weaken the pull-up current for a more robust write operation at low voltages [4], thereby improving the write-ability. Furthermore, a 10-transistor (10T) bitcell topology (see Fig. 3.1c) contains two additional transistors (compared to the 8T SRAM bitcell with read buffer) to convert one of the cell-internal cross-coupled inverters into a tri-state inverter [5], allowing cutting the pull-up path and easily writing a logic '0' level to the cell (without contention). A more straightforward technique to improve the robustness of an SRAM bitcell consists in transistor up-sizing [6]. Among various low-voltage write and read assist techniques, Intel has presented a voltage collapse scheme to temporarily lower the bitcell supply voltage to a value below the data-retention voltage and thereby dramatically weaken the PMOS keeper during write access [86], thereby improving write-ability. In order to counteract read disturbs, a popular read assist technique consists in raising the bitcell supply voltage above the voltage levels of the bit-line (BL) and the word line (WL) [87]. The work presented in [88] uses an integrated charge pump in order to selectively boost the write word-lines (WWLs) and the read word-lines (RWLs) of and 8T-bitcell SRAM register file (RF) which allows to reduce the main supply voltage of the RF (Vmin reduction), thereby reducing the overall power consumption (despite boosting the voltage of a few critical circuit nodes).

Using the above mentioned examples and similar techniques, a large variety of full-custom SRAM macrocells reliably operating in the near-threshold (near-V_T) and even in the sub-V_T domain have been designed in the last decade [6, 76, 77, 78, 89, 90]. However, all these techniques lead to large SRAM bitcells consisting of 8-14 transistors or a considerable overhead for low-voltage read and write assist circuits [31], which further aggravates the already dominant area share of embedded memories in SoCs and often results in a standby leakage power which dominates the overall power budget of ULV/ultra-low power (ULP) systems. To remedy excessive leakage currents, [89] has proposed a 14-transistor (14T) bitcell using high-threshold voltage (high-V_T) I/O transistors, stack forcing, and channel length stretching. As an interesting architectural technique to minimize standby power without the need for DC/DC voltage converters (such as low dropout (LDO) regulators), the work in [91] proposes voltage stacking between two SRAM sub-arrays, i.e., a series instead of a parallel connection of the sub-arrays between the power and ground rails, which efficiently reduces leakage current by 88%. Moreover, the works presented in [92, 93] introduce an improved adaptive bulk biasing control (AB2C) scheme for reduction of leakage currents during standby periods of SRAM (and

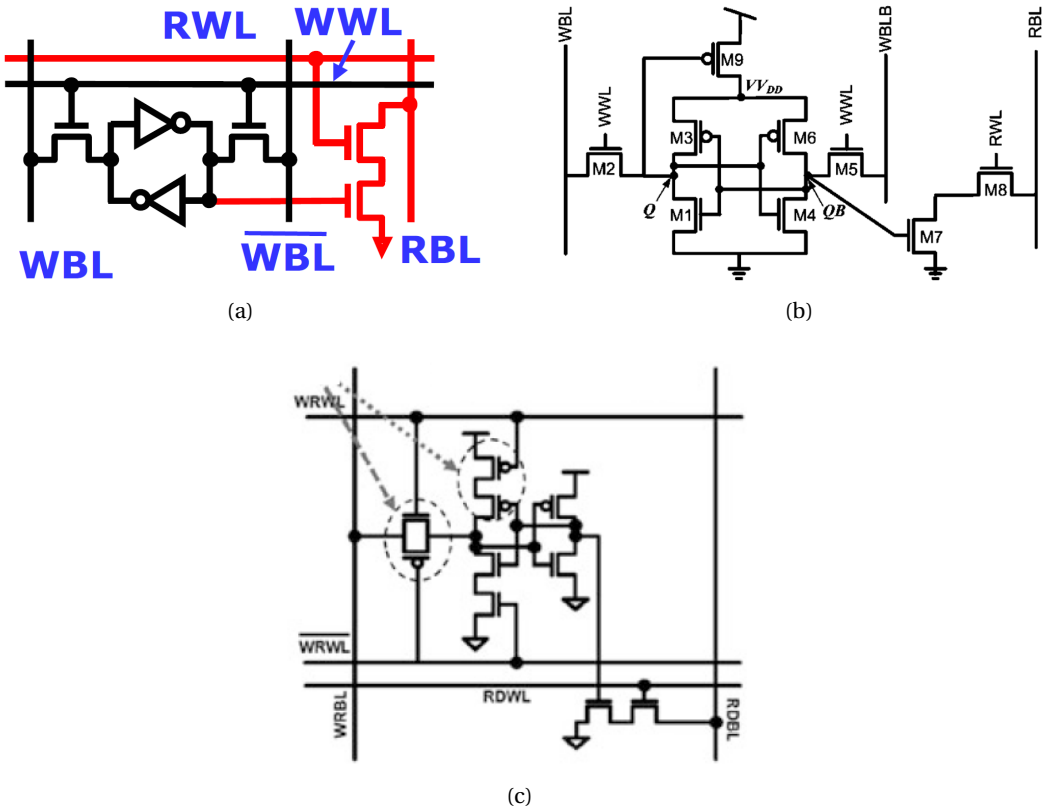


Figure 3.1: Robust low-voltage SRAM bitcells: (a) 8T [3], (b) 9T [4], and (c) 10T [5].

also of CMOS image sensor) arrays, while enabling device acceleration during active cycles. A main bottleneck inhibiting a wide acceptance of all above mentioned near-threshold and subthreshold SRAM macrocells among the digital ULV/ULP design community is the lack of good, fully automated memory compilers. To fill this gap, the following Section proposes the use of a fully automated sub-V_T SCM compilation flow. Also, ULV/ULP biomedical implants and sensor nodes typically require small memories of a few kb, a range of storage capacities where SCMs can even be more area-efficient than SRAM macrocells, while previous work on reliable subthreshold memories targets several hundreds of kb.

3.2 SCMs Based on Commercial SCLs Operated in Sub-V_T Regime

In this Section, SCMs are proposed as an alternative to full-custom sub-V_T SRAM macrocells for ULP systems requiring small memory blocks. The energy per memory access as well as the maximum achievable throughput in the sub-V_T domain of various SCM architectures are evaluated by means of a gate-level sub-V_T characterization model, building on data extracted from fully placed, routed, and back-annotated netlists. The reliable operation at the energy-minimum voltage of the various SCM architectures in a 65 nm CMOS technology considering

within-die (WID) process parameter variations is demonstrated by means of Monte Carlo circuit simulation. Finally, the energy per memory access, the achievable throughput, and the area of the best SCM architecture are compared to recent sub-V_T SRAM designs.

Section 3.2.1 presents the employed sub-V_T design and characterization flow, before Section 3.2.2 evaluates all SCM architectures for operation in the sub-V_T domain. Section 3.2.3 verifies the reliability of the best-practice sub-V_T SCM topology, while Section 3.2.4 compares it with prior-art sub-V_T SRAM macrocells.

3.2.1 Sub-V_T Design and Modeling Flow

The works in [72, 73, 74] promote the design of sub-V_T circuits based on conventional standard-cell libraries (SCLs), an approach which we follow and evaluate in this Section, as an alternative to full-custom sub-V_T circuit design. However, most commercial SCLs are designed for the above-V_T domain, meaning that a) they are mainly optimized for speed performance, as speed performance has been the main concern for above-V_T circuit design over the last few decades, and that b) physical models describing the timing and the power consumption of the standard-cells are readily available only for the nominal supply voltage. Instead of using commercial SCLs optimized for above-V_T operation, standard-cell based sub-V_T design would ideally rely on SCLs which are especially optimized for sub-V_T operation [94, 95], meaning that more emphasis is given to leakage reduction and robustness than to performance while designing the standard-cells. If the development of a dedicated sub-V_T SCL is not economic—which corresponds to the viewpoint adopted in this Section—, a commercial SCL, optimized for above-V_T operation, can still be re-characterized to at least generate the physical timing and power models valid for sub-V_T supply voltages. Beside SCLs, virtually all logic synthesis tools as well as place-and-route (P&R) tools have been developed for regular digital VLSI design in the above-V_T domain, and therefore use sophisticated timing-driven optimization algorithms, whereas they are less well suited to directly optimize a design for *minimum energy dissipation per operation* (including the evaluation at different voltages and for different switching activities to find the minimum-energy point), which is an important metric for energy-constrained systems. This Section outlines different synthesis and analysis strategies for sub-V_T system design using commercial SCLs and commercial logic synthesis as well as P&R tools. The focus is on energy-constrained sub-V_T systems, which are optimized to perform a given operation with the lowest possible energy dissipation, assuming that the system might be power-gated or turned off after task completion. For more details on the various sub-V_T synthesis strategies and a detailed case study, the reader is referred to [81]. Note that while the various sub-V_T design and analysis flows discussed hereinafter will be applied to SCMs in this Chapter, they can also be used for any other synthesizable digital design.

Synthesis and Analysis Methods

Above-V_T Synthesis with Sub-V_T Analysis Due to the predominance of SCLs and design tools developed for regular above-V_T synthesis, it might be convenient to synthesize different architectural variants of a system (SCM or any other digital design), with different constraints on timing and power, in the above-V_T domain, and subsequently analyze and compare the energy dissipation and throughput of the various resulting designs in the sub-V_T domain. To this end, two methods to analyze the sub-V_T behavior of designs which have previously been synthesized in the above-V_T domain are presented and compared next.

Analytical sub-V_T model: As shown in Fig. 3.2a, the first method starts from a regular static timing analysis (STA) and voltage-change dump (VCD)-based power analysis of a fully placed, routed, and back-annotated netlist in the above-V_T domain. An analytical model [96, 97], summarized in Appendix A, is then used to scale timing and power quantities to the sub-V_T domain. A main advantage of this analytical model is the ability to immediately find the energy minimum voltage (EMV), sometimes also referred to as minimum-energy point (MEP), i.e., the supply voltage which minimizes the energy per operation [98]. The analytical sub-V_T frequency model in [96, 97] makes the assumption that the propagation delay(s) of all standard-cells slow down at the same pace as the propagation delay of a basic inverter when the supply voltage V_{DD} is gradually scaled down. In a dedicated study [81], we verified the accuracy of this assumption by analog circuit simulation of all standard-cells used in a benchmark design [97]. The analytical model was found to slightly underestimate the critical path delay in the sub-V_T domain for the considered 65 nm CMOS SCL. A more time-consuming but more precise (in terms of timing) sub-V_T analysis method is discussed next.

Evaluation using sub-V_T characterized SCLs: The second method, shown in Fig. 3.2b, consists of characterizing the original SCL again for many different supply voltages in the sub-V_T domain (from 250 mV to 400 mV in steps of 10 mV in the current case[2]), and then repeating the STA and the VCD-based power analysis using these re-characterized SCLs. For an accurate VCD-based power analysis, the standard delay format (SDF) file generation from the RC-annotated netlist, and the VCD dump from the gate-level simulation must be repeated for each supply voltage.

Comparison of sub-V_T analysis methods: The results of the two sub-V_T analysis methods (analytical sub-V_T model and evaluation using re-characterized sub-V_T SCLs) are compared by applying them to a reference design [97] which has previously been synthesized, placed, and routed at nominal supply voltage using a 65-nm CMOS SCL. Concerning the estimation of the energy dissipation per clock cycle for operation at a constant clock frequency, both sub-V_T analysis methods coincide fairly well, as shown in Fig. 3.3. This means that the sub-V_T model [96, 97] does accurately predict the active energy and the leakage power.

[2] Since the considered low-power (LP) high threshold-voltage (HVT) NMOS and PMOS transistors in a 65 nm CMOS technology have absolute threshold-voltage values above 450 mV, the considered voltage range is clearly in the sub-V_T domain.

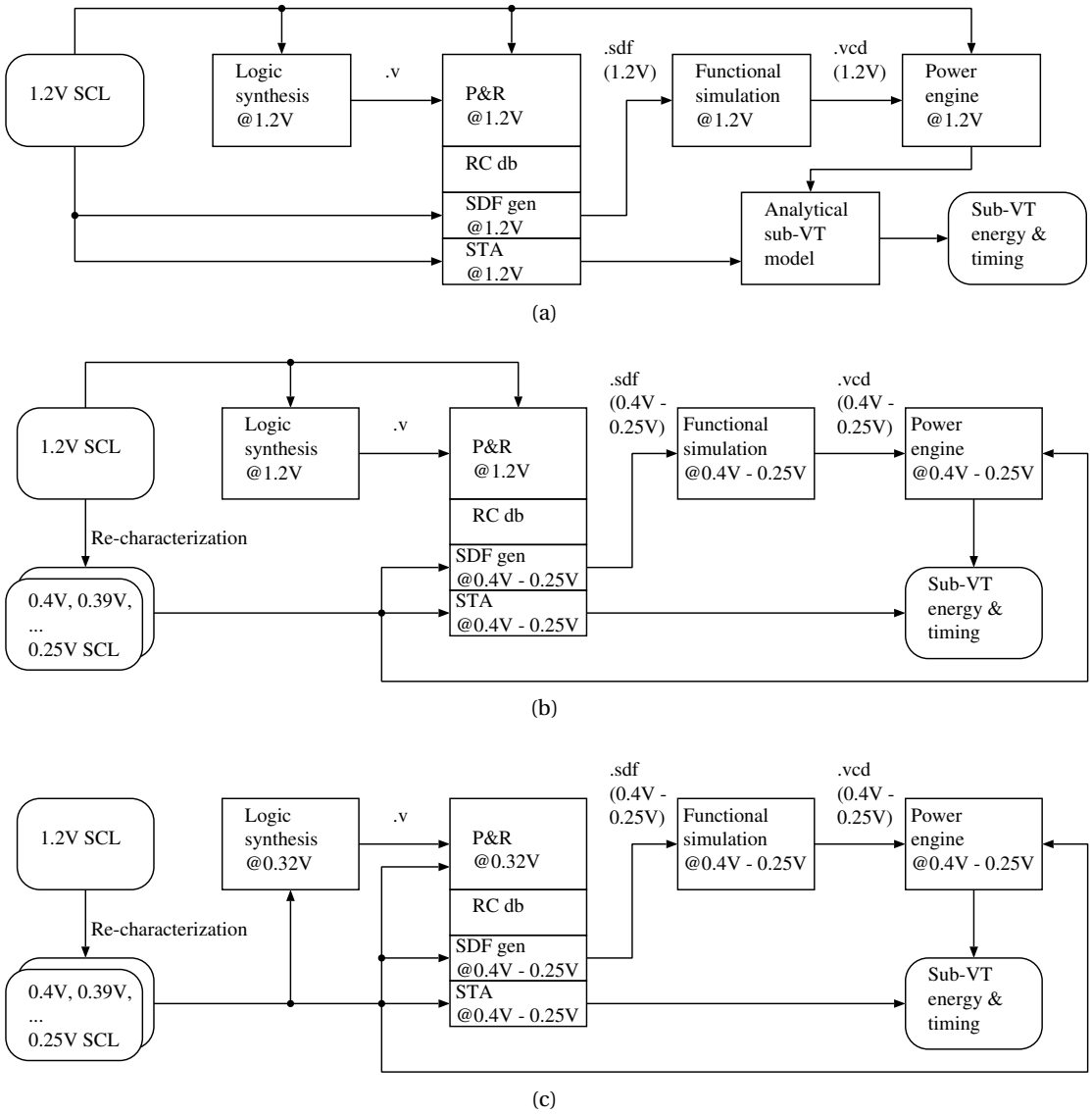


Figure 3.2: Sub-V_T design and analysis flows: (a) Above-V_T synthesis, STA, and power analysis. Analytical sub-V_T model. (b) Above-V_T synthesis. Sub-V_T STA and power analysis. (c) Sub-V_T synthesis, STA, and power analysis.

The analytical sub-V_T model is thus very convenient to quickly and reasonably precisely estimate the leakage power consumption and the active energy dissipation in the sub-V_T domain, and to quickly have a reasonable guess of EMV. For a more precise maximum frequency and EMV estimation, it is important to re-characterize the SCL and repeat the STA in the sub-V_T domain. In the remainder of this Section, for an extensive design space exploration of many SCM topologies operated in the sub-V_T domain, we use the fast (in terms of CPU time) and reasonably precise flow based on the analytical sub-V_T model, as shown in Fig. 3.2a.

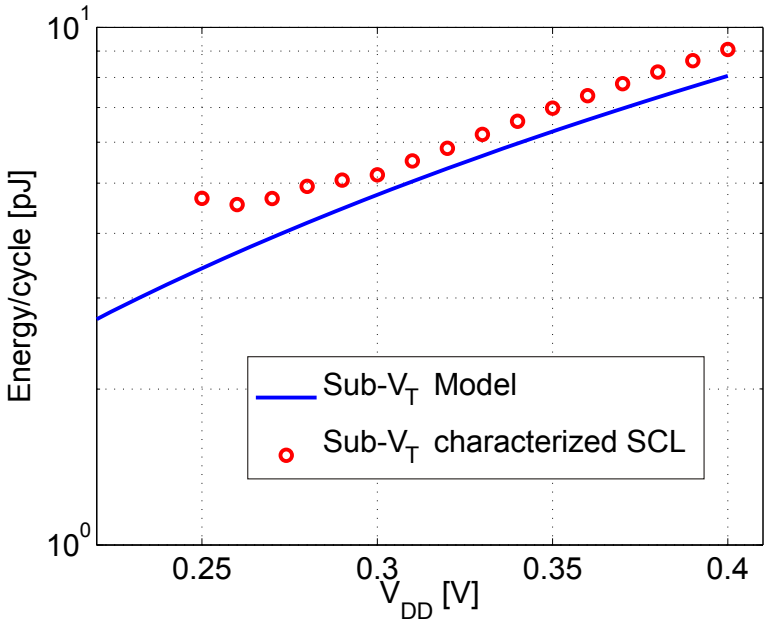


Figure 3.3: Comparison of two sub-V_T analysis methods (analytical sub-V_T model and evaluation using sub-V_T SCLs): energy dissipation for operation at a constant frequency of 1kHz.

Direct Sub-V_T Synthesis For voltage-constrained sub-V_T systems, or if the approximate EMV is already known from a previous above-V_T synthesis, it might be desirable to directly synthesize in the sub-V_T domain, which allows to specify meaningful timing constraints, and to directly obtain timing and power figures for the considered supply voltage from STA and the power engine, respectively. Fig. 3.2c shows a direct sub-V_T synthesis and analysis flow, which, in addition to the supply voltage at which the logic synthesis and P&R are performed, gives the energy dissipation and timing metrics of the resulting design for the entire sub-V_T range, allowing to find the true EMV. Since we do not know *a priori* the EMV of the various SCM architectures or of the target system, we do not perform direct sub-V_T synthesis in the following. Rather, we will perform above-V_T synthesis followed by the analytical sub-V_T modeling to see and compare the behavior of all SCM architectures in the entire sub-V_T domain.

3.2.2 Sub-V_T SCM Architecture Evaluation

We now aim at identifying the SCM architecture that performs best in the sub-V_T domain in terms of energy, but also in terms of throughput, and silicon area. To this end, the SCM architectures originally introduced in Section 2.2.1 (see Fig. 2.2) are evaluated for operation in the sub-V_T domain using the previously explained design and analysis flow shown in Fig. 3.2a. All SCMs are mapped to a 65 nm CMOS technology with low-power (LP) high threshold-voltage (HVT) transistors (V_T is above 450 mV) and the results are based on fully synthesized, placed, and routed netlists with back-annotated layout parasitics. The average switching activity μ_e is

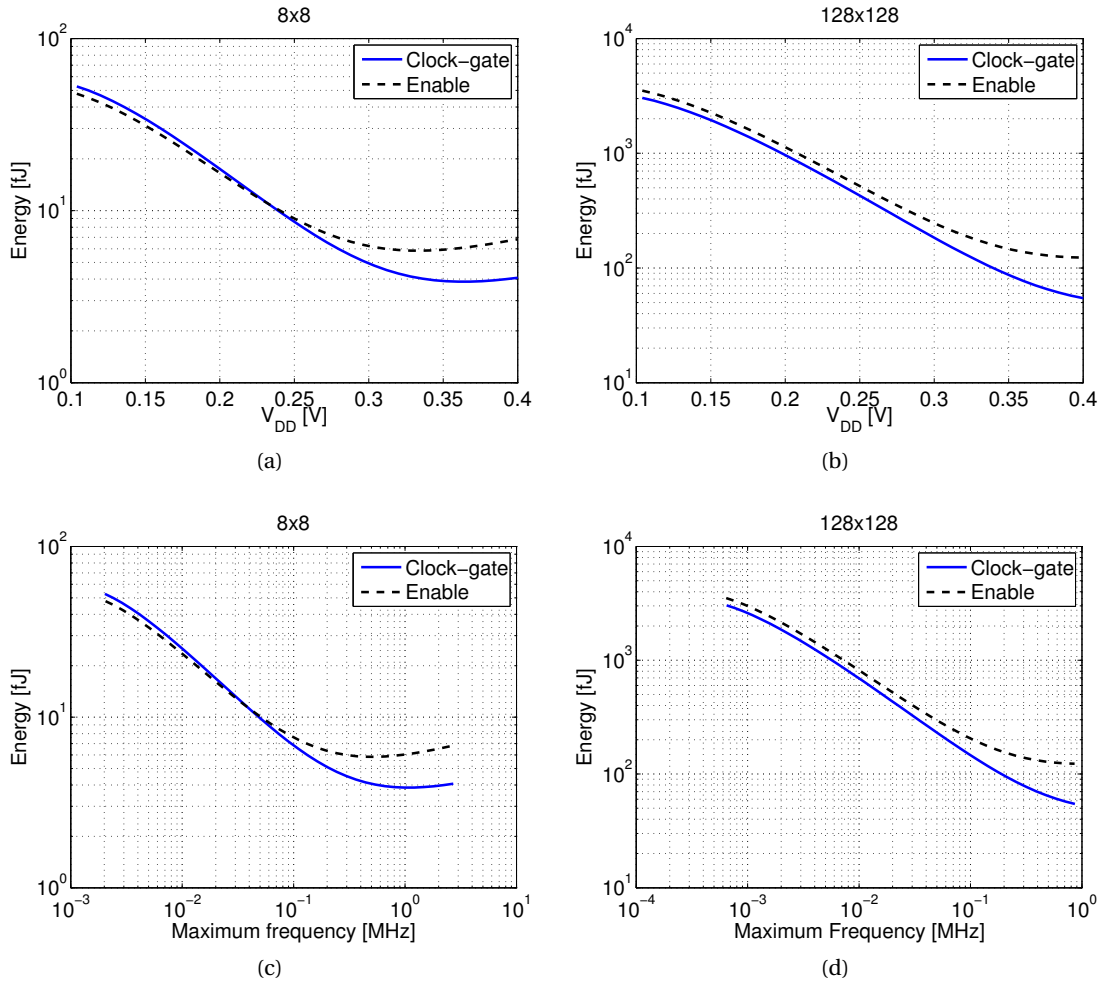


Figure 3.4: Energy versus V_{DD} for different write logic implementations, namely *enable flip-flops* and *basic flip-flops in conjunction with clock-gates*, assuming a multiplexer based read logic, for (a) $R = 8$ and $C = 8$ as well as for (b) $R = 128$ and $C = 128$. Energy versus maximum achievable frequency for the same memory architectures and sizes is shown in (c) and (d).

obtained using voltage change dumps (VCDs) for 1000 write and read cycles. All inputs of the SCMs are driven by buffers of standard driving strength, and all highly capacitive nets such as the bit lines (BLs) are buffered inside the SCMs. For the comparisons between SCMs of different sizes $R \times C$, energy figures are reported as *energy per written bit* and *energy per read bit*, commonly referred to as *energy per accessed bit*. In paragraphs "Comparison of Write Logic Implementations" and "Comparison of Read Logic Implementations" below the different implementations of the write and read ports are compared and in paragraph "Comparison of Storage Cell Implementations" flip-flop arrays are compared with latch arrays.

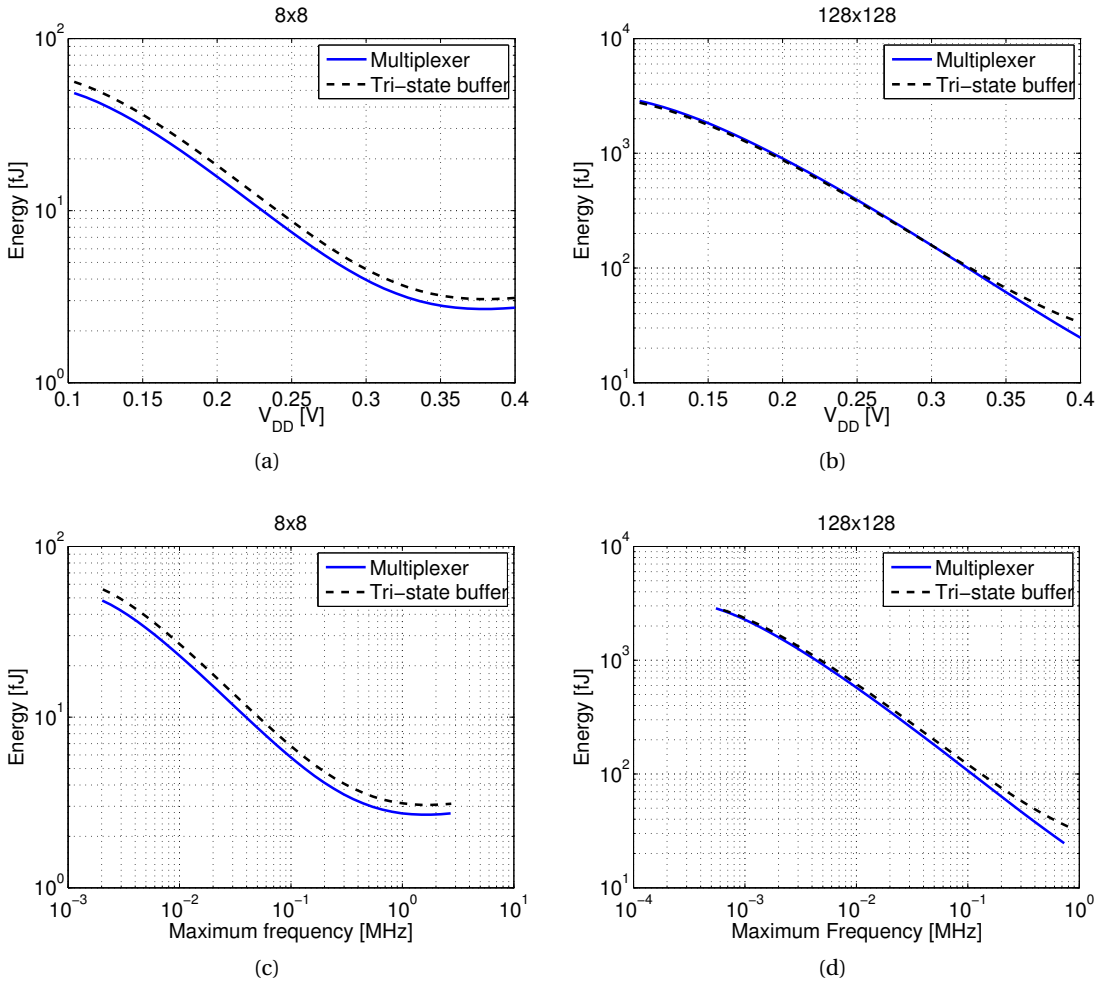


Figure 3.5: Energy versus V_{DD} for different read logic implementations, namely *tri-state buffers* and *multiplexers*, assuming a clock-gate based write logic and latches as storage cells, for (a) $R = 8$ and $C = 8$ as well as for (b) $R = 128$ and $C = 128$. Energy versus maximum achievable frequency for the same memory architectures and sizes is shown in (c) and (d).

Comparison of Write Logic Implementations

In order to compare different write logic implementations, we choose a multiplexer-based read logic and flip-flops as storage cells. We consider two memory configurations ($R = 8$, $C = 8$ and $R = 128$, $C = 128$) which are expected to have a smaller and to full-custom sub-V_T SRAM designs comparable area cost, respectively.

Fig. 3.4a and Fig. 3.4b show the energy per written bit as a function of the supply voltage V_{DD} for the small and the larger memory configuration, respectively. In both cases, the write logic relying on clock-gates in addition to basic flip-flops exhibits lower energy per written bit than the architecture that employs flip-flops with enable, for the range around the energy-minimum supply voltage (EMV). In the sub-V_T regime, there are two main reason for this

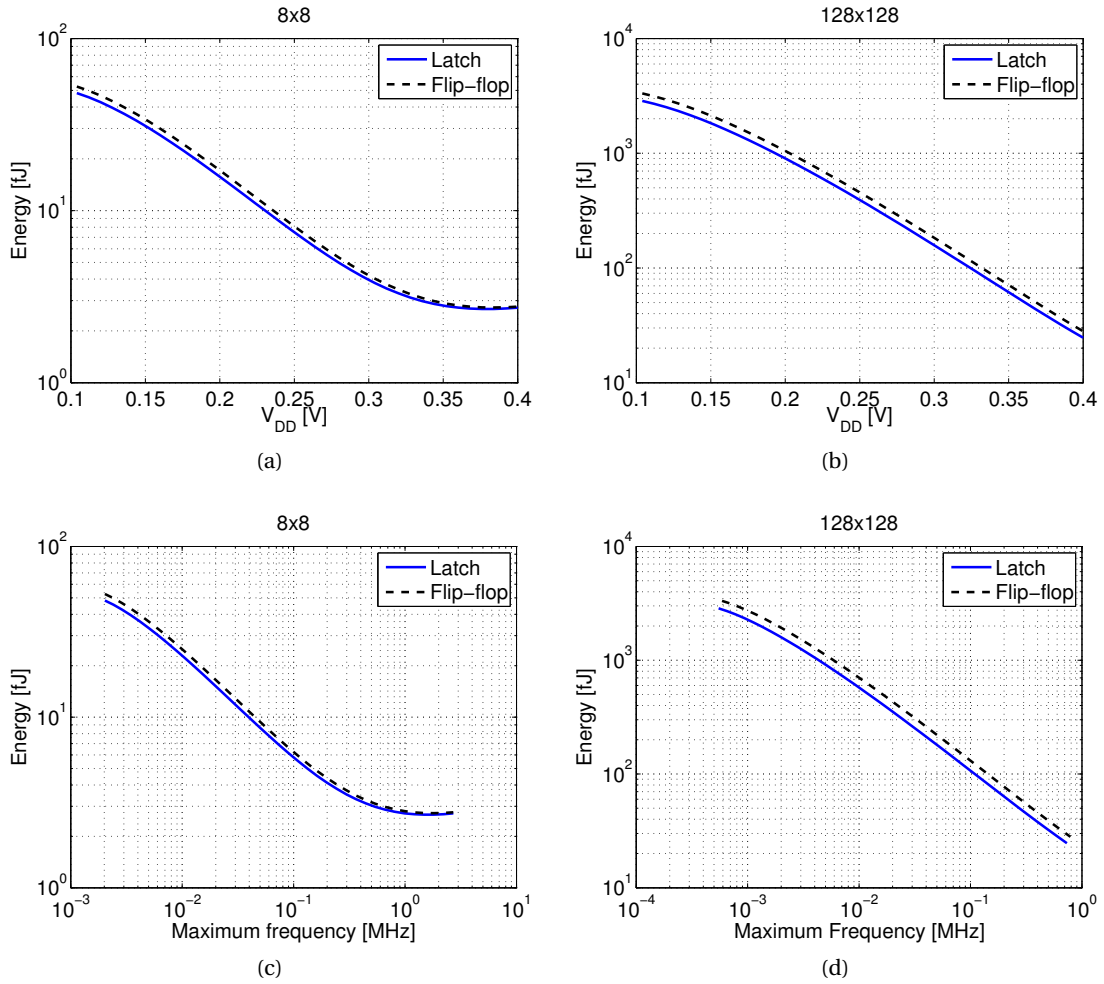


Figure 3.6: Energy versus V_{DD} for different storage cell implementations, namely *latches* and *flip-flops*, assuming a clock-gate based write logic and a multiplexer based read logic, for (a) $R = 8$ and $C = 8$ as well as for (b) $R = 128$ and $C = 128$. Energy versus maximum achievable frequency for the same memory architectures and sizes is shown in (c) and (d).

behavior: First, the architecture based on clock-gates dissipates less active energy than the architecture based on enable flip-flops, as the latter distributes the clock signal to each storage cell, while the former silences the clock signal of all, but the selected row. The second reason is more visible for the larger storage array whose energy dissipation is dominated by leakage. This leakage is larger for the case of the more complex storage cells that require additional circuitry to realize the enable for each cell in a standard-cell based implementation.

For systems that require a constrained memory bandwidth, the energy dissipation at a given frequency may also be of interest. Fig. 3.4c and Fig. 3.4d show the energy per written bit as a function of the maximum achievable operating frequency of the corresponding SCM. The frequency range on the x-axis is obtained by sweeping V_{DD} from 0.1 V to 0.4 V. It can be seen

that both architectures have the same maximum operating frequencies, as the critical path is in the read logic through the output multiplexers.

With respect to area, we remind from Section 2.2.1 that the clock-gate architecture yields smaller SCMs than the enable architecture if only $C \geq 4$. This statement is true for many different CMOS technologies and standard-cell libraries.

In summary, for sub-V_T memory implementations, the clock-gate architecture exhibits lower energy, equal throughput, and smaller area compared to the enable architecture and is therefore generally preferred.

Comparison of Read Logic Implementations

In order to compare different read logic implementations, we choose the clock-gate based write logic and a latch-based storage array for again a small and a larger SCM configuration. Fig. 3.5a and Fig. 3.5b show that the multiplexer based read logic with a read address decoder (RAD) has a small advantage over the tri-state buffer based read logic in terms of energy per read bit, at least around the energy-minimum supply voltage. Fig. 3.5c and Fig. 3.5d show that there is no significant difference between the two read logic implementations as far as the maximum achievable operating frequency is concerned. Indeed, the delay of the tri-state buffer is quite long and comparable to the delay through the entire multiplexer as all R tri-state buffers in one column are connected to the same net, which consequently has a high capacitance.

In summary, multiplexer based SCMs have a small energy and an area advantage [32] (see Section 2.2.1), compared to the tri-state buffer approach and are therefore preferred.

Comparison of Storage Cell Implementations

In order to compare different storage cell implementations, the best write and read logic implementations and again a small and a larger SCM block are considered. Fig. 3.6a and Fig. 3.6b show that latch arrays have less energy per accessed bit than flip-flop arrays, due to smaller leakage currents drained in each storage cell and due to lower active energy of the latch implementation. However, the energy savings of using latches instead of flip-flops are only small: a latch has around 2/3 the leakage of a flip-flop in the considered standard-cell library, but only around 2/3 of all cells in an SCM are storage cells, which accounts for the approximately 22 % energy reduction visible from Fig. 3.6d.

Fig. 3.6c and Fig. 3.6d show that there is no significant difference in terms of maximum frequency. In fact, the storage cells are not in the critical path, since the critical path of any SCM is through the RAD and the tri-state buffers or the multiplexers. However, flip-flops as sotrage cells allow for shorter write address setup-times than latches, as previously described in Section 2.2.1.

Table 3.1: Standard-cell area A_{SC} and area $A_{P\&R}$ of fully placed and routed latch and flip-flop arrays for different configurations $R \times C$, clock-gate based write logic, and multiplexer based read logic.

		Latch array		Flip-flop array	
R	C	A_{SC} [μm^2]	$A_{P\&R}$ [μm^2]	A_{SC} [μm^2]	$A_{P\&R}$ [μm^2]
8	8	738	984	811	1.1k
8	32	2.5k	3.3k	2.8k	3.7k
8	128	9.5k	12.7k	10.6k	14.1k
32	8	2.9k	3.8k	3.1k	4.2k
32	32	9.9k	13.2k	10.9k	14.6k
32	128	37.9k	50.6k	42.1k	56.2k
128	8	11.2k	15.0k	12.3k	16.4k
128	32	39.4k	52.5k	43.7k	58.3k
128	128	152.2k	202.9k	169.0k	225.4k

Latch arrays have only slightly smaller area than flip-flop arrays [32]. Table 3.1 shows the standard-cell area A_{SC} and the area $A_{P\&R}$ of fully placed and routed latch and flip-flop arrays for different configurations $R \times C$, the clock-gate based write logic, and the multiplexer based read logic. Notice that $A_{P\&R} = A_{SC}/0.75$, as the SCMs have been successfully placed and routed with a typical initial floorplan utilization of 75 %. An approximation of the area $A(R,C)$ for an arbitrary memory configuration $R \times C$ can be found according to

$$A(R,C) = \beta_1 + \beta_2 R + \beta_3 C + \beta_4 RC + \beta_5 \text{ceil}(\log_2(R)) + \beta_6 \text{ceil}(\log_2(C)). \quad (3.1)$$

The coefficients $\beta_1 \ldots \beta_6$ are obtained through a least squares fit to a set of reference configurations in the technology under consideration such as the ones provided in Table 3.1.

To summarize, sub-V_T latch arrays have slightly less energy per accessed bit, achieve the same frequency, and are smaller compared to sub-V_T flip-flop arrays.

Best Practice Implementation

Fig. 3.7 shows the schematic of the best sub-V_T SCM architecture. This architecture uses latches without enable feature as storage cells, clock-gates for the write logic, and multiplexers for the read logic. Note that this topology coincides with the best-practice implementation which was previously identified for above-V_T operation (see Section 2.2.1). Therefore, if working exclusively with commercially available standard-cell libraries, and avoiding standard-cell optimization for high-density (see Section 2.3) or ultra-low leakage (see subsequent Section 3.3), the SCM topology shown in Fig. 3.7 is the optimum choice irrespective of the targeted supply voltage.

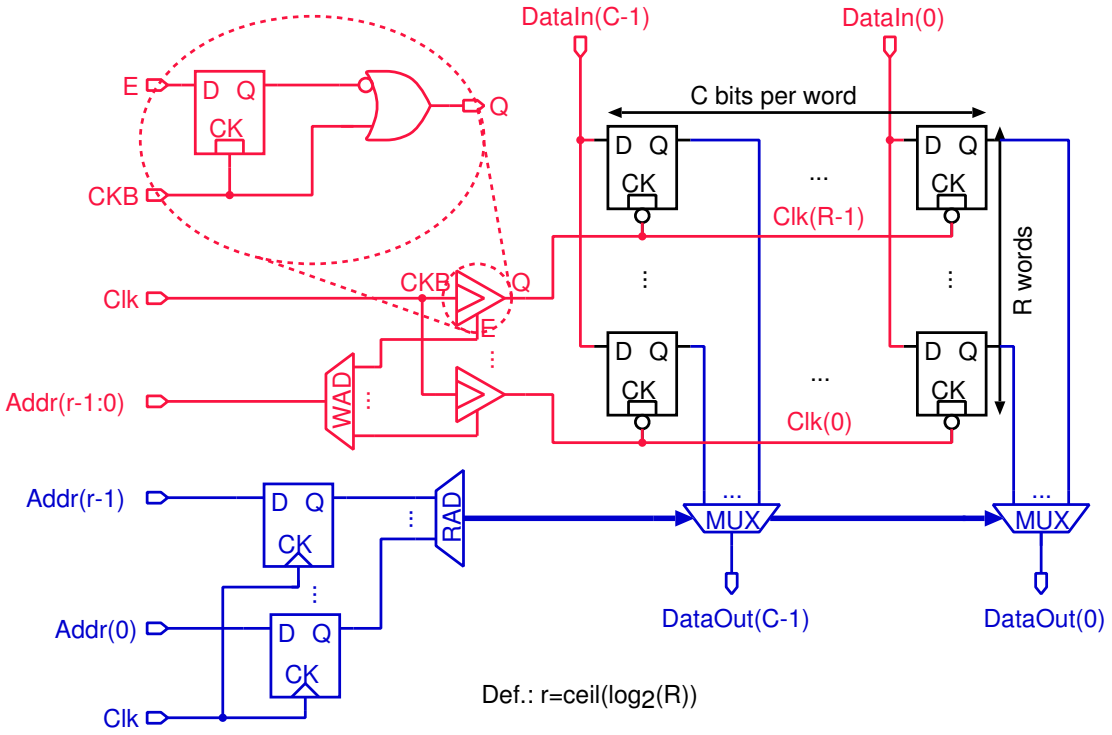


Figure 3.7: Best-practice sub-V_T SCM topology: latch based SCM with clock-gates for the write logic and multiplexers for the read logic.

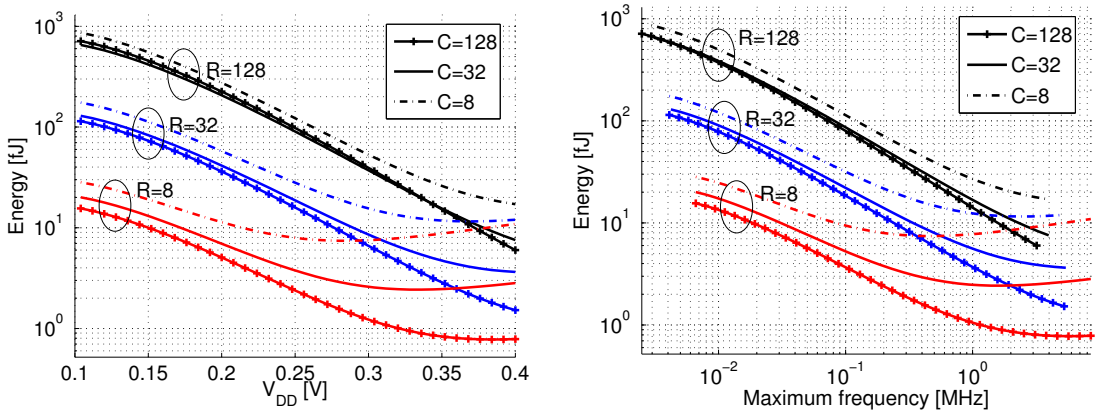


Figure 3.8: Energy versus V_{DD} (a) and energy versus frequency (b) for the *latch multiplexer clock-gate* architecture for different memory configurations.

With respect to the energy efficiency, it is clear that a significant switching activity is required to find an energy-minimum, which occurs only for the smallest memory configurations. However, for the large memory configurations, the overall switching activity is very low and the energy dissipation is clearly dominated by the integration of the leakage power over the access time,

which decreases with increasing V_{DD} if always operating at maximum speed. Consequently, the energy-minimum supply voltage within the sub-V_T domain approaches the threshold voltage V_T when increasing the memory size.

For different memory configurations with the same storage capacity ($R \cdot C$ = const.), we observe from Fig. 3.8a and Fig. 3.8b that the energy-efficiency improves for a larger number of columns C and a smaller number of rows R. The reason for this behavior is that the maximum operating frequency increases as R decreases which again reduces the contribution of the energy consumed due to leakage power in each access cycle.

3.2.3 Reliability Analysis

One of the limiting factors with respect to voltage scaling in the sub-V_T domain is the reliability of the circuit. Reliability issues arise mainly from within-die process variations and are aggravated in nanometric CMOS technologies. Consequently, ensuring robust operation in the sub-V_T regime has been one of the most important concerns in the design of full-custom sub-V_T storage arrays (refer back to Secion 3.1 for more details).

Compared to full-custom designs, SCMs are compiled from conventional combinational CMOS logic gates, such as NAND, NOR, or AOI gates, and from sequential elements, i.e., latches and/or flip-flops. The reliability issue therefore corresponds to the discussion down to which supply voltage a given standard-cell library can operate reliably. This point limits in the same way the operation of the combinational and sequential logic and of the embedded SCMs for a given process corner.

To determine the range of reliable operation of the SCMs, we distinguish between the combinational and the sequential cells in the library, used to construct the storage array. Previous work shows that when gradually scaling down the supply voltage, the sequential cells fail earlier then the combinational CMOS logic gates [73], provided that the combinational logic is built without transmission gates. Therefore, the focus is on the analysis of the sequential elements in the following.

The peripherals of SCM storage arrays, i.e., the read and write logic, are built from combinational CMOS gates and are thus less sensitive to process variation than the array of storage cells itself. Also, delay variations in SCM peripherals induced by process variation are unproblematic due to the used single-edge-triggered one-phase clocking discipline where path delays do not necessarily need to be matched. Compared to SCM peripherals, the peripherals of SRAM arrays are more sensitive to process variation: delay variations may cause the sense amplifiers to be triggered at the wrong time, and mismatch in the sense amplifiers can further compromise reliability, especially at very low supply voltages.

Sensitivity of SCMs to Variations

Reliability issues in both sequential standard-cells and in dedicated SRAM storage cells essentially arise from mismatch between carefully sized transistors due to *within-die* process variations [99]. Remember from Section 3.1 that in a conventional 6T SRAM bitcell, such mismatch manifests itself in four types of failures: a) read failures, b) write failures, c) hold failures, and d) read-access time failures. The read failures result from the direct access of the read bit line to the storage node, a situation which does not occur in a standard latch design such as the one shown in Fig. 3.9, where the output is isolated from the internal node with a separate buffer. The write failures in a 6T SRAM bitcell are caused by the inability to flip the storage nodes that suffer from an unusually strong keeper. The standard-cell latch avoids this issue by turning off the feedback path during write operation. Read-access time failures are a concern in high-speed systems, but are not problematic in ultra-low power sub-V_T systems where speed performance is only a secondary concern. The only remaining issue are hold failures which occur in the non-transparent phase of a latch during which the circuit behavior essentially resembles that of a basic 6T SRAM bitcell. Hence, a conventional standard-cell latch may be viewed as a very conservative SRAM cell design [6] where the reliability is determined by the risk of experiencing hold failures.

Hold Failure Analysis

Fig. 3.9 shows a simplified schematic of the latch which was chosen by the logic synthesizer from a commercial standard-cell library in order to minimize leakage and area of the latch arrays described in this paper. A latch needs to be able to hold data in the non-transparent phase. In this phase, INV2 and INV3 in Fig. 3.9 act as a cross-coupled inverter pair. The stability of the state of this pair is usually defined by the *static noise margin* (SNM) that is required to hold data in the presence of voltage noise on the storage nodes [100]. This SNM is extracted as the side of the largest embedded square of the butterfly curves shown in Fig. 3.10 for different supply voltages in the sub-V_T domain. For each butterfly curve, there is an SNM associated with the top-left and the bottom-right eye, referred to as *SNM high* and *SNM low*. The probability distribution functions on the right-hand side of Fig. 3.10 are always for the minimum of *SNM high* and *SNM low*. The butterfly curves and the corresponding minimum SNM distributions are obtained from 1000-point Monte Carlo circuit simulation assuming within-die process parameter variations for the typical process corner at a temperature of 25 °C. All common parameters of the BSIM4 transistor simulation models are subject to variation according to statistical distributions provided by the foundry.

The distributions in Fig. 3.10 show that the SNM values decrease with the supply voltage. As can be seen in Fig. 3.10a, there is a clear separation between the voltage transfer characteristic (VTC) of inverter INV2 and the inverse VTC of inverter INV3 corresponding to a comfortable SNM for a supply voltage of 400 mV, which also corresponds to the energy optimum supply voltage for most SCM architectures and sizes. Fig. 3.10b and Fig. 3.10c show that there is still a separation between the VTCs even at lower supply voltages, indicating that operation is still

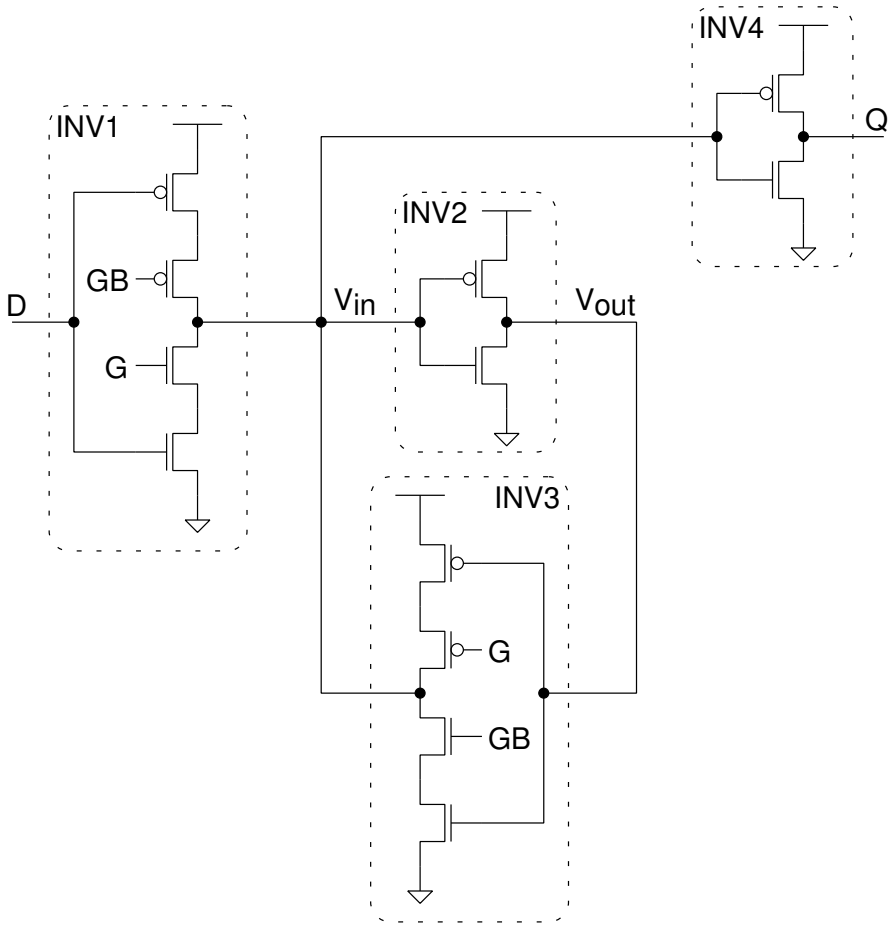
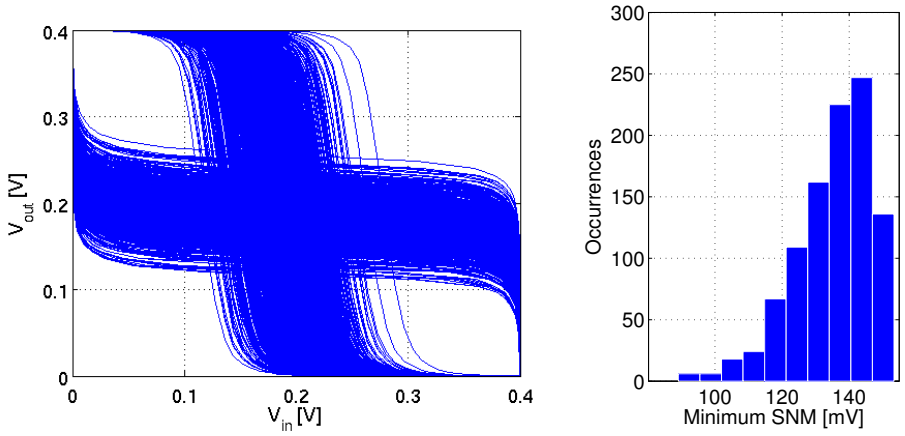


Figure 3.9: Simplified schematic of the latch used in the best sub-V_T SCM architecture.

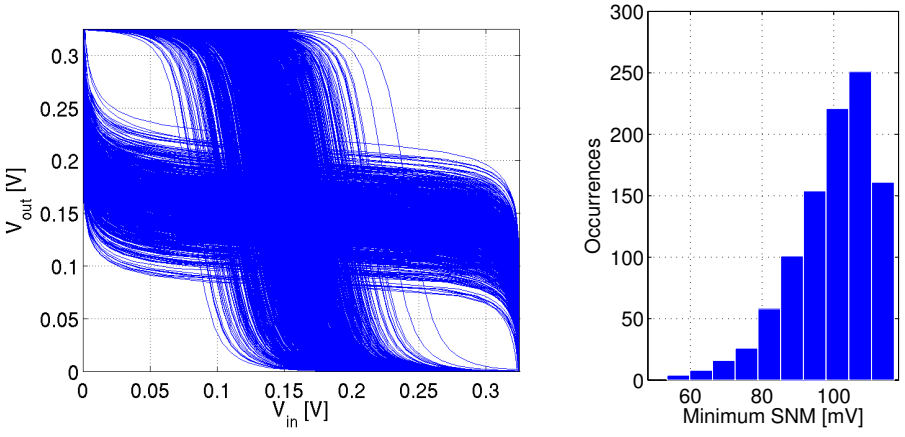
possible, but the SNMs are small and reliability clearly starts to become critical at 250 mV, limiting the range of operation.

3.2.4 Comparison with Sub-V_T SRAM Designs

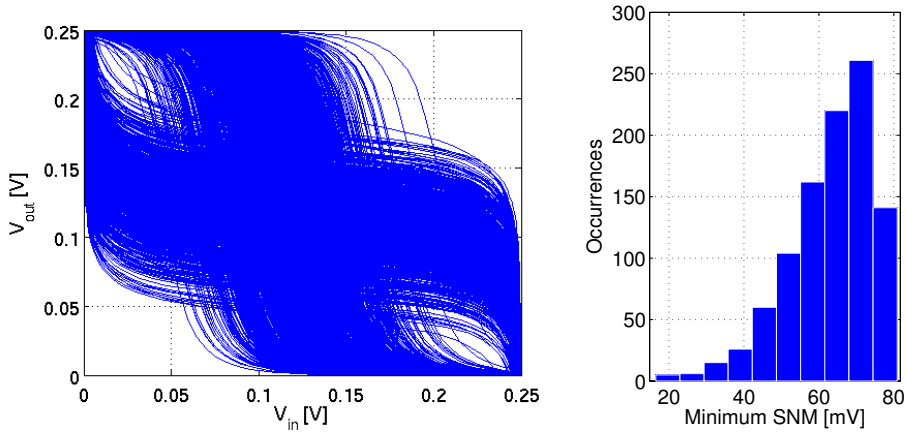
In this section, the performance and cost of sub-V_T SCMs is compared to a selection of sub-V_T SRAM designs in literature [6, 76, 77, 78, 80]. The paragraph "Overview" below gives an overview of recent sub-V_T memory implementations including this work. The paragraph "Energy and throughput" compares in detail the energy and throughput of the smallest SCM architecture with a prominent sub-V_T SRAM design, while the paragraph "Area" compares their area.



(a)



(b)



(c)

Figure 3.10: Butterfly curves (left) and distribution of minimum hold SNM (right) of the latch used in the best sub-V_T SCM architecture for (a) $V_{DD} = 400\,mV$, (b) $V_{DD} = 325\,mV$, and (c) $V_{DD} = 250\,mV$.

Table 3.2: Comparison of sub-V_T memories.

Publication	[6]	[76]	[77]	[78]	[80]	This [33]
Capacity [kb]	256	256	64	8	480	32
Tech. [nm]	65	65	65	90	130	65
Basis of results	ASIC measurements					Post-layout
V_{DDmin} [mV]	380[a]	350[c]	300	160	200	300
f_{max} [kHz]	475 (0.4 V)	25	20 (0.25 V)	200	120	1 000 (0.4 V)
Energy [fJ/bit]	65.6 (0.4 V)	884.4	86.0[d] (0.4 V)	750[e]	4.2	32.7 (0.4 V)
Area [μm^2/bit]	2.9[b]	4.0[b]	7.0[b]	19.5	12.8	12.5

[a]One redundant row and column per 32-kb block are assumed to guarantee reliable operation at this supply voltage.
[b]Area estimated from die photograph.
[c]Plus 50 mV for boosting of word line drivers.
[d]Estimation extracted from a graph.
[e]Includes the energy dissipation of the package.

Overview

Table 3.2 presents a selection of recently published sub-V_T memories. V_{DDmin} is defined as the minimum supply voltage which guarantees reliable write, hold, and read operations. Unless otherwise stated, the maximum operating frequency f_{max} is given for $V_{DD} = V_{DDmin}$. The reported energy includes both active energy for a read operation and the leakage energy of the memory array during the access time. Furthermore, the total energy value is normalized by the width of the data IO bus, thereby reporting the total energy per read bit. Unless otherwise stated, the energy is given for f_{max} at V_{DDmin}.

All sub-V_T SRAM designs [6, 76, 77] realized in a 65 nm CMOS technology have $V_{DDmin} \geq$ 300 mV. Monte Carlo simulations indicate that SCMs mapped to the same technology should operate reliably at least down to the same minimum supply voltage. Two SRAM designs [78, 80] fabricated in older technologies are less sensitive to process parameter variations and are reported to have an even lower V_{DDmin}, i.e., 160 mV and 200 mV, respectively.

At the same technology node and supply voltage V_{DD}, SCMs are faster than SRAM designs, which bares the potential to lower energy dissipation per memory access if 1) speed is traded against energy, or 2) early task completion is honored by power gating. Obviously, older technologies exhibit lower leakage currents which may lead to lower energy per memory access.

With respect to area, the use of robust latches, available from conventional standard-cell

libraries, instead of 8T or 10T SRAM cells in the same CMOS technology node, is clearly paid for by a larger area per bit for SCMs.

Energy and Throughput

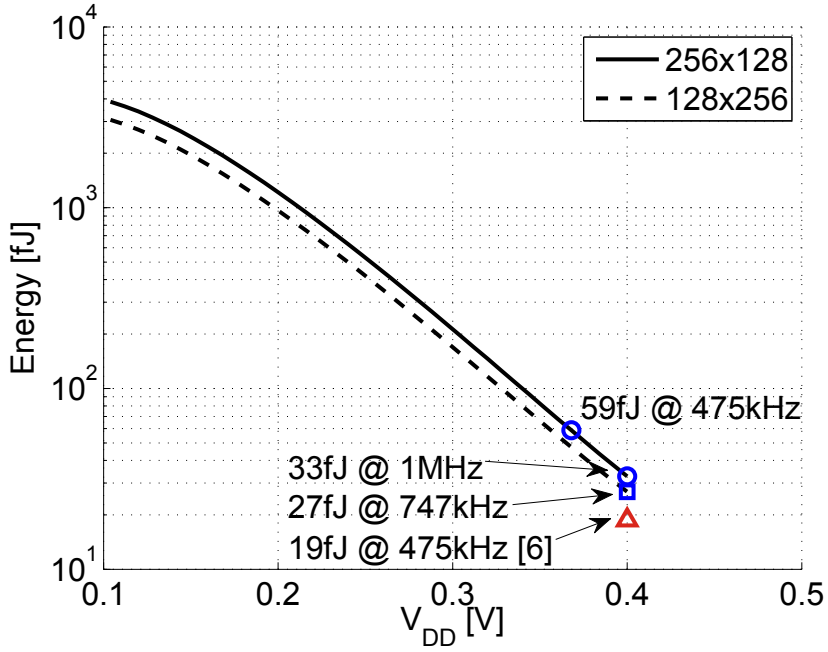
A well-cited 256-kb 10T-bitcell sub-V_T SRAM [6] in 65 nm CMOS has 8 32-kb blocks ($R = 256$, $C = 128$), which are served by a single 128-bit data IO bus. The leakage energy of this SRAM macro is divided by 8 to compare one block with the proposed 32-kb SCM block, while the active energy is taken as is, since only one block is accessed at a time. At 400 mV, the SRAM macro is reported to be operational at $f_{max} = 475\,kHz$, and a single 32-kb block dissipates 19 fJ per accessed bit, as indicated by the triangle in Fig. 3.11.

For comparison, Fig. 3.11a, and Fig. 3.11b, show the energy per accessed bit of the smallest SCM architecture as a function of V_{DD} and f_{max}, respectively. Considering an SCM block with $R = 256$ and $C = 128$, $f_{max} = 475\,kHz$ is already achieved at $V_{DD} = 370\,mV$ and the energy per accessed bit for this operating point is 59 fJ, which is more than for the full-custom SRAM macro. However, when operated at the same supply voltage ($V_{DD} = 400\,mV$), the SCM is able to operate at $f_{max} = 1\,MHz$, with an energy dissipation of 33 fJ per accessed bit, which is only 1.7× higher compared to the full-custom design. The energy savings compared to the initial operating point are achieved due to a higher possible clock frequency combined with power gating after earlier completion of a task.

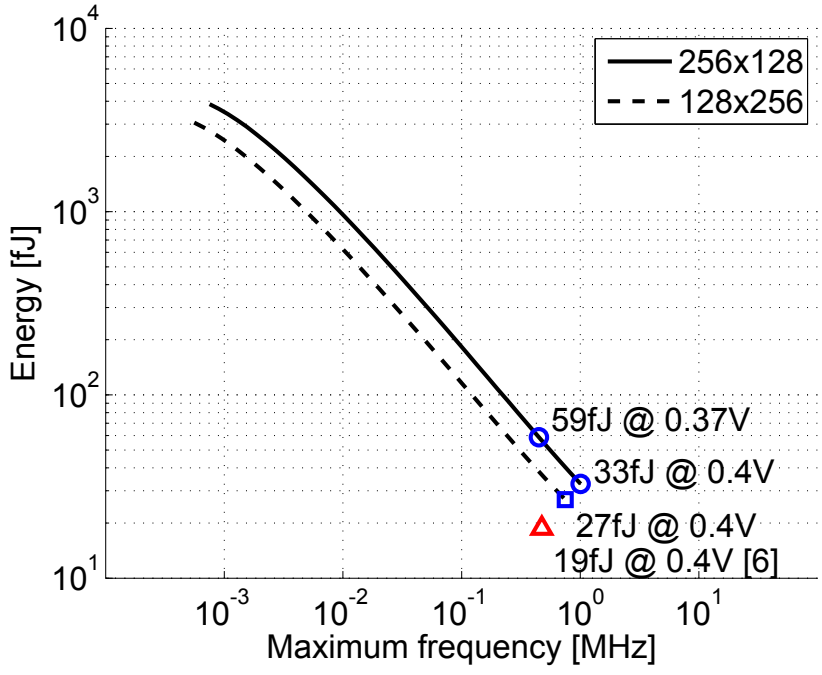
Changing the SCM configuration to $R = 128$ and $C = 256$ while keeping a constant storage capacity $R{\cdot}C$, the energy per accessed bit of the SCM is further reduced. As shown by the square marker in Fig. 3.11, this new SCM configuration is able to run at 747 kHz for $V_{DD} = 400\,mV$, and dissipates 27 fJ per read bit in this operating point, which is only 1.4× higher than for the full-custom design. This change in the SCM configuration results in lower energy and doubled memory bandwidth at the price of a higher routing congestion during system integration.

Area

The bitcell of SCMs (flip-flop or latch) is clearly larger than the SRAM bitcell. However, SRAM macrocells have an overhead to accommodate the peripheral circuitry, i.e., precharge circuitry and sense amplifiers [34]. For SRAM macrocells with small storage capacity, this area overhead may be significant. Hence, SCMs may outperform SRAM macrocells in terms of area for small storage capacities, but become bigger for large storage capacities. In Section 2.2.1, it was shown that the border up to which static above-V_T SCMs are still smaller than 6T-bitcell SRAM macrocells depends on the *number of words* and the *number of bits per word*, and may be as large as 1 kb. However, the analysis in Section 2.2.1, considered only circuit implementations for above-V_T operation, i.e., SRAM macros based on the 6T bitcell and SCMs synthesized with a given timing constraint. When considering circuit implementations specifically optimized for sub-V_T operation, SRAM macrocells become significantly larger due to the need for 8T [76],



(a)



(b)

Figure 3.11: Energy versus V_{DD} (a) and energy versus frequency (b) for the *latch multiplexer clock-gate* architecture for $R = 256$, $C = 128$ and for $R = 128$, $C = 256$. The red triangle corresponds to [6].

10T [6], or even 14T [89] bitcells and the additional assist circuits required for reliable sub-V_T operation. As opposed to this, SCMs may be synthesized with relaxed timing constraints (and still reach 1 MHz in the current study) as speed is not of major concern for typical ultra-low-power applications and may therefore have a reduced area cost compared to above-V_T implementations.

In the present case, considering a storage capacity of 32 kb, the SCM is 4.3 times larger than a corresponding SRAM block [6]. For some applications, this area increase may be acceptable for the benefit of lower energy per memory access and higher throughput.

3.3 Ultra-Low Leakage Sub-V_T SCMs

Thus far, in Section 3.2, the design and analysis of robust sub-V_T SCM topologies was limited to the use of commercial standard-cell libraries (SCLs), which, unfortunately, are typically optimized for high speed performance at nominal supply voltage in the above-V_T domain. However, for ultra-low power (ULP)/ultra-low voltage (ULV) systems which are operated in the sub-V_T domain, speed performance is only a secondary concern, while the design of the standard-cells should rather focus on low leakage currents, which, eventually, leads to low leakage power and low access energy of the SCMs. To this end, this Section identifies the major leakage contributors of the best-practice SCM found in Section 3.2.2 and shows the significant leakage current savings which can be achieved by the design and integration of custom-designed standard-cells. As opposed to previous work [101], the proposed SCM design flow does not restrict the leakage minimization to the bitcells, but extends it to the peripheral circuits by using a 3-state read logic, accepting a speed degradation for the benefit of ultra-low leakage.

More precisely, this Section presents an ultra-low-leakage 4 kb SCM manufactured in 65 nm CMOS technology. To minimize leakage power during standby, a single custom-designed standard-cell (D-latch with 3-state output buffer) addressing all major leakage contributors of SCMs is seamlessly integrated into the fully automated SCM compilation flow. Silicon measurements of the 4 kb SCM indicate a leakage power of 500 fW per stored bit (at a data-retention voltage of 220 mV) and a total energy of 14 fJ per accessed bit (at energy-minimum voltage of 500mV), corresponding to the lowest values in 65 nm CMOS reported to date, among all previous sub-V_T memory implementations.

3.3.1 Ultra-Low Leakage Standard-Cell Design

Custom Low-Leakage Latch Design

Approximately 66% of the leakage power of SCMs are consumed by the latches, whereas the read multiplexers dominate the remaining power. Hereinafter, the most dominant leakage contributors are addressed by a custom low-leakage latch design. Latch topologies using

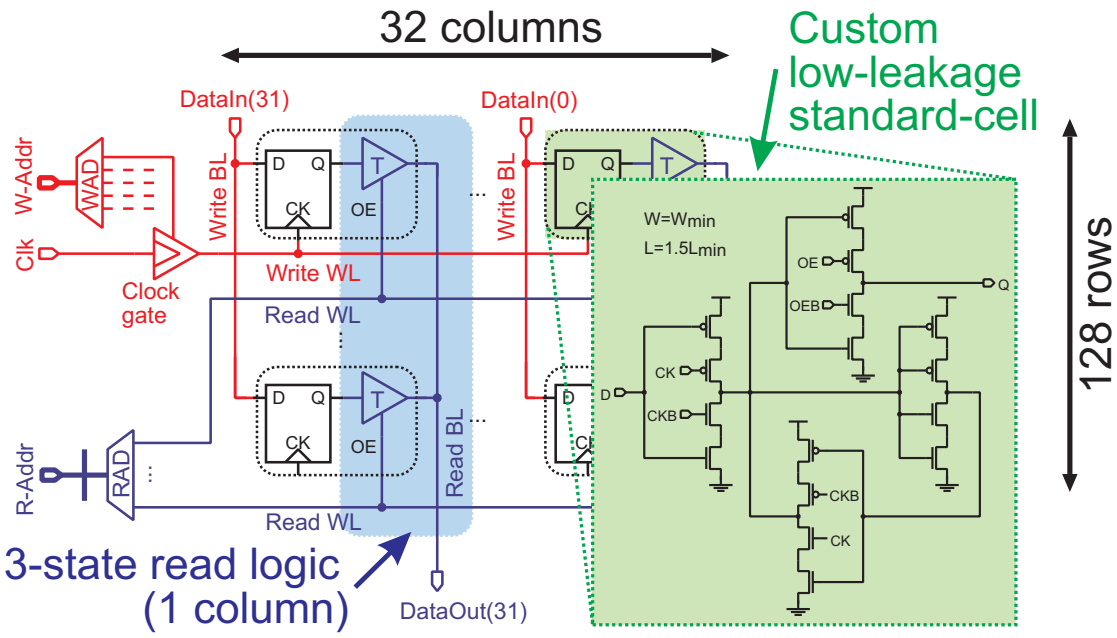


Figure 3.12: Architecture of ultra-low-leakage 4 kb standard-cell based memory (SCM): the write logic uses clock-gates, while the 3-state inverters used for the read functionality are integrated in the low-leakage latch design.

3-state buffers inherently have transistor stacks and consequently low leakage currents, while topologies using transmission-gates and static-CMOS gates suffer from higher leakage currents. The best latch topology exhibiting the lowest leakage current has 1) the lowest number of paths from V_{DD} to ground, and 2) the highest resistance on each such paths, directly leading to a topology with 3-state buffers only. Having identified the best latch topology, transistor stacking (for parts of the latch which do not yet have transistor stacks) and channel length stretching are applied to further reduce leakage currents. The stacking factor is strictly limited to 2 since higher factors suffer from diminishing returns in leakage reduction and compromise reliability for sub-V_T operation. Moreover, the point of diminishing returns of channel length stretching, where the area increases with a negligible reduction in leakage, is found to be 1.5–2× the minimum channel length. The right-hand side of Fig. 3.12 shows the transistor-level schematic of the final custom-designed standard-cell latch (with 3-state output buffer, the assets of which are discussed in the following), while the left-hand side shows the SCM architecture.

Low-Leakage 3-State Read Logic

The read multiplexers of SCMs, routing the selected word to the data output, are an integral part of the read logic and can be implemented with 3-state buffers instead of combinational CMOS gates and/or dedicated multiplexer standard-cells (see Fig. 2.2 in Section 2.2.1). Choosing a 3-state read logic and integrating a 3-state inverter into the basic storage cell

allows to address all dominant leakage contributors of SCMs by designing only one custom standard-cell. Relying on a CMOS multiplexer read logic would require a larger number of custom-designed, low-leakage standard-cells for an overall SCM leakage reduction. Moreover, the 3-state-enabled latch allows for a compact placement of the storage array on a regular grid (not easily achieved with CMOS multiplexers), reducing the SCM-internal routing and thus active energy, while it still provides the freedom to spread the SCM and merge it with logic blocks, in case the interface to the memory is more critical. As previously discussed, it is beneficial in terms of leakage current to apply transistor stacking to each branch of the latch, including the output buffer (or, more precisely, the output inverter). This already stacked output inverter of the custom-designed D latch is easily converted into a 3-state inverter, thereby addressing all major SCM leakage contributors by designing a single custom standard-cell.

The remainder of this section aims at finding the optimum transistor sizing of the 3-state drivers to simultaneously reduce overall leakage and improve speed, which is not contradictory in the sub-V_T regime, as expatiated on below. The presented 4 kb SCM consists of 128 rows and 32 columns, as shown in Fig. 3.12. Thus, 128 3-state buffers are connected to the same read bit-line (RBL). During a read operation, the 3-state buffer in the selected word has to drive the RBL against 127 unselected, yet leaking 3-state buffers. To investigate the impact of the 3-state drive strength on the RBL (dis-)charge delay, a strong and a weak driver, defined in Table 3.3, are considered. For a compact layout fitting nicely onto the standard-cell grid, and symmetric rise and fall times being only a secondary goal for the targeted low-speed ULV/ULP applications, the 3-state drivers are non-symmetric with equal NMOS and PMOS transistor sizes. As a result, RBL rise times are always longer than RBL fall times. Moreover, the reported RBL pull-up delays correspond to a worst-case data scenario where the initially discharged RBL needs to be pulled up to V_{DD} through the selected 3-state driver, while the input voltages of all unselected 3-state drivers are set to pull the RBL low (see Fig. 3.12), thereby maximizing the total leakage current working against the active current. Table 3.3 shows this worst-case, 50%-to-50% rising-RBL propagation delay of the selected 3-state driver for the typical-typical (TT) process corner at 27 °C, for both above-V_T and sub-V_T supply voltages, and for both drive strengths. The considered low-power (LP) high threshold-voltage (HVT) 65nm CMOS technology has a nominal V_{DD} and a threshold-voltage of 1.2 V and 650 mV, respectively. Thus, a V_{DD} of 400 mV is already deep in the sub-V_T domain. Simulation results indicate that the stronger 3-state driver is faster for operation at nominal V_{DD} where on-to-off current ratios (I_{on}/I_{off}) are as high as 10^7 (for both NMOS and PMOS transistors), whereas the weaker 3-state driver is faster for sub-V_T operation, due to much lower I_{on}/I_{off} ratios of around 10^4 and the resulting non-negligible impact of the leakage current of unselected 3-state drivers. Furthermore, the impact of the input voltage of unselected 3-states on the RBL delay is completely negligible in the above-V_T domain, whereas it is slightly visible in the sub-V_T domain. Of course, the weaker drivers have lower leakage currents in both the sub-V_T and the above-V_T regime, compared to the stronger drivers.

Table 3.3: Read bit-line (RBL) delay, TT corner, 27 °C.

Drive strength	**Strong**	**Weak**
W/W_{min}, L/L_{min}	15, 1	1, 2
V_{DD}	**RBL delay**	
1.2 V	1.064 ns	2.126 ns
400 mV	3.336 µs	2.688 µs

Reliability Analysis

While bitcell read failures and write failures are avoided by using a read buffer and by disabling the bitcell-internal keeper, respectively, hold failures limit V_{DD} down-scaling, as previously explained in Section 3.2.3 in more detail. To assess the minimum V_{DD} (V_{DDhold}) required for the ultra-low leakage latch to hold data, the minimum V_{DD} for which both static noise margin (SNM) values (corresponding to data '1' and '0', or, in other words, to top and bottom eye of the butterfly curve [100]) are still positive are extracted from a 1k-point Monte Carlo (MC) circuit simulation (accounting for within-die (WID) parametric variations, in the TT corner, at 27 °C). Fig. 3.13 shows the hold failure probability as a function of V_{DD}, while the inset shows the corresponding distribution of V_{DDhold}. The first hold failure occurs at 200 mV, corresponding to a worst (maximum) value of V_{DDhold} equal to 210 mV.

Due to the strong impact of parametric variations and low I_{on}/I_{off} ratios in the sub-V_T regime, the total leakage current from a large number of disabled 3-state buffers might become high enough, compared to the active drive-current of a single, weak 3-state buffer, to compromise the reliability of the 3-state read logic. This leakage current issue limits the maximum number of words per RBL for reliable operation. Nevertheless, 1k MC runs accounting for WID parametric variations in the slow-slow (SS) process corner at 27 °C indicate that for up to 128 words per RBL, a single 3-state driver successfully drives the RBL at a V_{DD} as low as 400 mV.

3.3.2 Silicon Measurements of 4 kb Sub-V_T SCM

Fig. 3.14 shows the chip microphotograph and the layout picture of the 4 kb SCM based on 3-state-enabled low-leakage latches and manufactured in 65 nm CMOS technology with LP-HVT transistors. The silicon area of the 4 kb SCM block is 315 x 165 µm^2, corresponding to 12.7 µm^2 per bit. Functionality is verified by writing and reading back checker-board and random data patterns using a scan-chain test interface. Unless stated differently, the environmental temperature is carefully controlled to 27 °C with an oven for all silicon measurements. Moreover, self-heating effects are insignificant due to the extremely low power consumption of this memory chip.

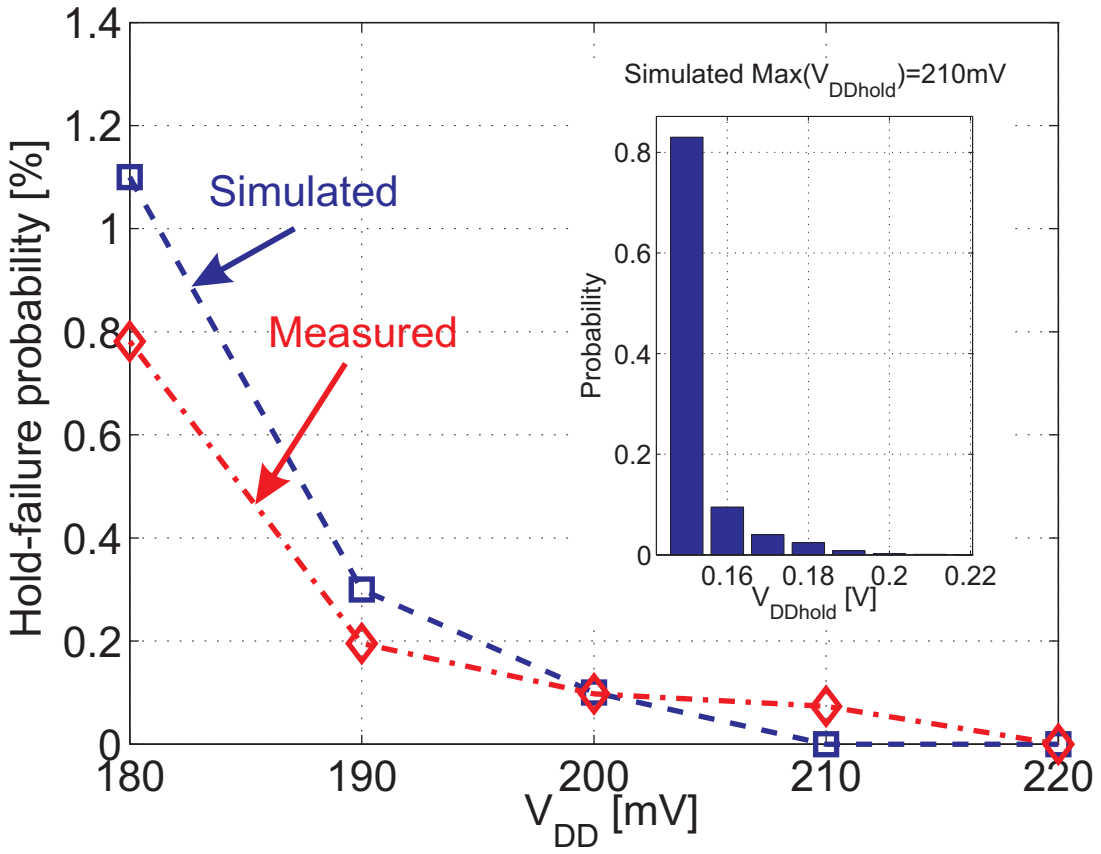


Figure 3.13: Simulated and measured hold failure probability versus V_{DD}. Inset: Simulated distribution of V_{DDhold}.

Minimum V_{DD} for Data Retention and Memory Access

The measured minimum required supply voltages to guarantee correct hold, write, and read functionality are 220, 300, and 420 mV, respectively. The measured value of V_{DDhold} (220 mV) is in good agreement with the aforementioned simulated value (210 mV), as shown in Fig. 3.13. It is apparent that the low-leakage 3-state read logic limits the minimum voltage for read/write access (V_{DDmin}). For a closer inspection of the onset of read failures, Fig. 3.15 shows error maps: a green (bright) marker indicates correct access to a bitcell, while a red (dark) marker indicates an access failure. For $V_{DD} = 380$ mV, it is apparent that failures occur column-wise, confirming that the 3-stated RBLs are the first point of failure under V_{DD} scaling. Completely error-free access is measured at $V_{DDmin} = 420$ mV. Fig. 3.16 shows the the number of inoperative columns, i.e., columns containing at least one bitcell with access failure, as a function of V_{DD}, while the inset shows the total number of bitcell read failures versus V_{DD}.

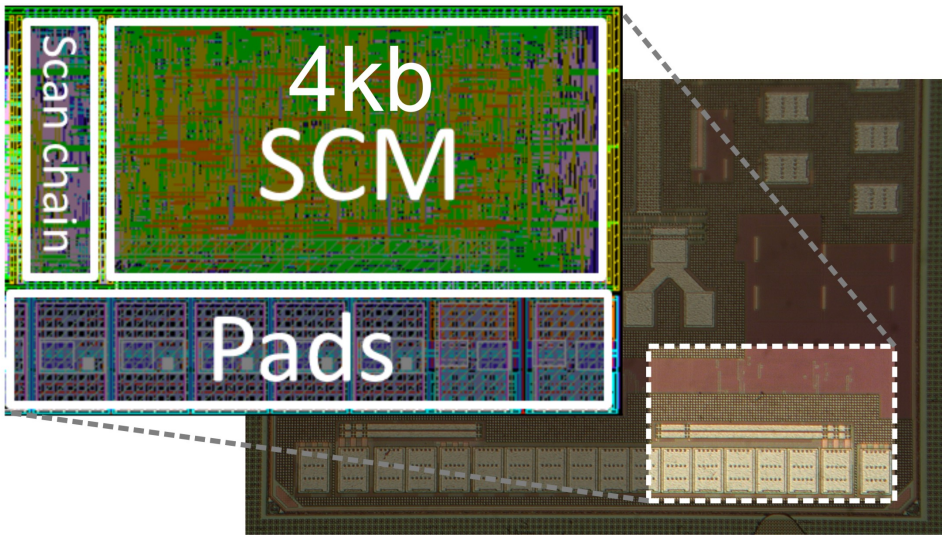


Figure 3.14: Chip microphotograph and zoomed-in layout of sub-V_T SCM test chip; the 4 kb SCM block, the test interface, and the I/O pads are highlighted.

Access Energy, Frequency, and Leakage Power

Fig. 3.17 shows the measured energy per bit-access performed at maximum speed versus V_{DD}. The measured energy-minimum voltage is located at 500 mV, while the minimum energy dissipation per bit access is 14 fJ. At 675, 500, and 420 mV (V_{DDmin}), the maximum measured operating frequencies are 1.5 MHz, 110 kHz, and 10 kHz, respectively. The 3-state read logic limits V_{DDmin} and the read-access time, but satisfies the ambition of ultra-low leakage power and access energy, while the energy-minimum voltage is still higher than V_{DDmin}. At V_{DDhold} = 220 mV, data is correctly held with a leakage power of 425-500 fW per bit (best and worst dies), as shown in Fig. 3.18.

Measurements at Human-Body Temperature

Biomedical implants encounter a typical working temperature of 37 °C. At 37 °C, the first completely error-free read access to the entire SCM array is measured at already 400 mV, as compared to 420 mV for a temperature of 27 °C. As a desirable effect of higher temperatures, the maximum operating frequency doubles when heating the chips from 27 to 37 °C (measured at V_{DD} = 420 mV). Unfortunately, the leakage power increases as well with increasing temperature, as shown in Fig. 3.18.

3.3.3 Comparison with Prior-Art Sub-V_T Memories

Compared to our previous study on SCMs considering only commercially available standard-cell libraries (presented in Section 3.2, based on simulation results), designing merely one custom standard-cell (3-state-enabled low-leakage latch) cuts the leakage power and the

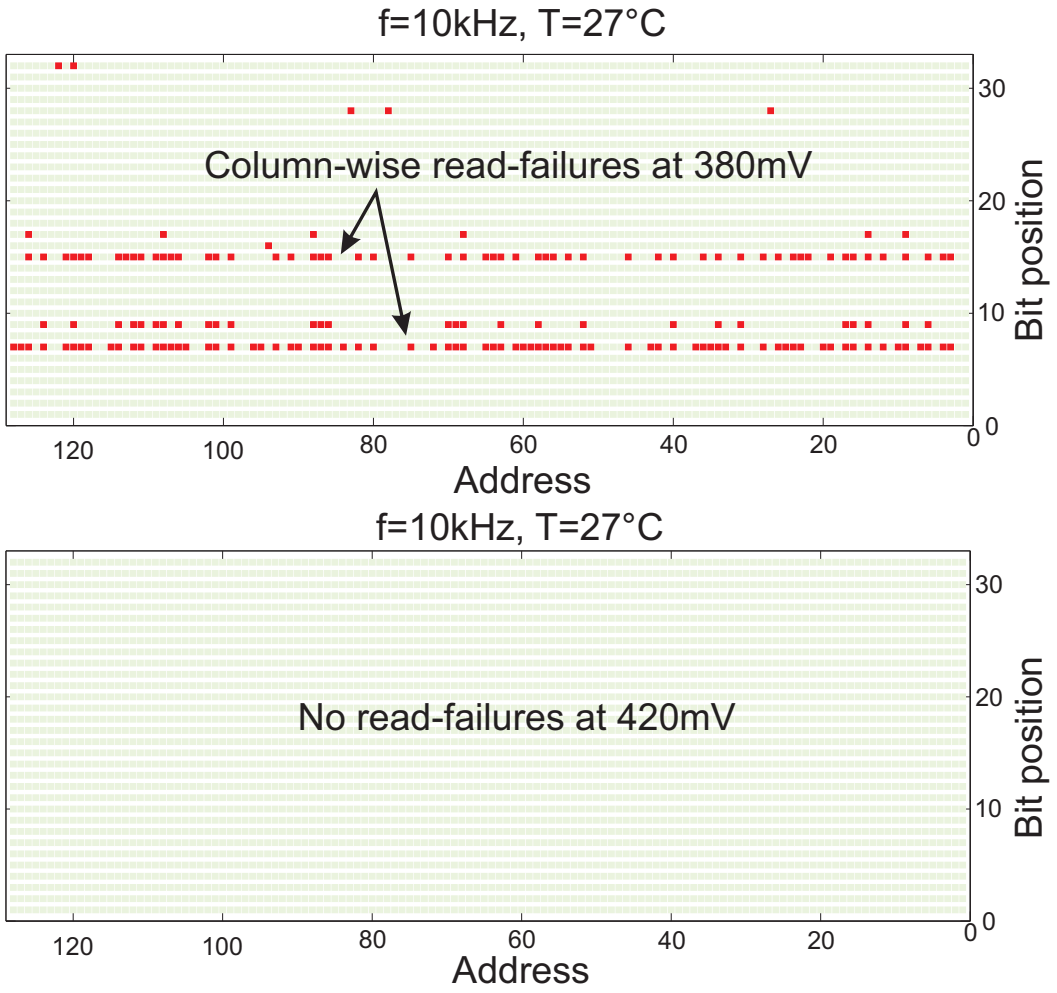


Figure 3.15: Measured error maps for V_{DD} of 380 mV (top) and 420 mV (bottom).

energy per bit-access into half while maintaining the same silicon area.

Table 3.4 shows the best (in terms of access energy and leakage power) silicon-proven sub-V_T memories in 65 nm CMOS reported until the day of writing. The energy figures ($E_{tot/bit}$) correspond to the total (active and leakage) energy per memory access performed at maximum speed, normalized to the size of the data I/O bus. Unless stated in parentheses, $E_{tot/bit}$ is given for V_{DDmin}. The power figures ($P_{leak/bit}$) correspond to the leakage power of the memory macro (including peripheral circuits) during standby, normalized to the macro's storage capacity. Unless stated in parentheses, $P_{leak/bit}$ is given for V_{DDhold}.

In [101], the standby leakage of the SRAM macro is dominated by the leakage of peripheral circuits, due to the aggressive reduction of array leakage. In our approach, not only the bitcell (latch), but also the leakage-dominant peripheral circuits (read multiplexers) are leakage-optimized, which clearly pays off compared to [101] (see Table 3.4).

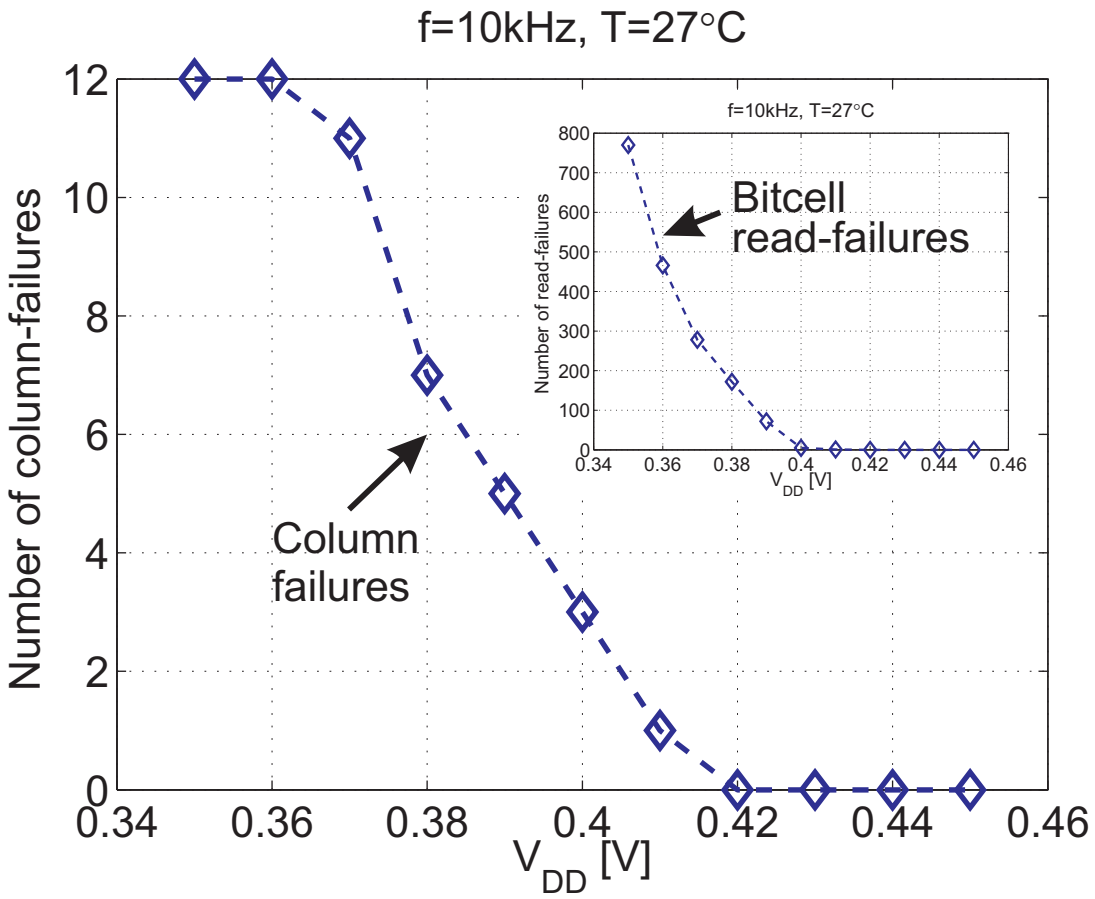


Figure 3.16: Measured number of inoperative columns versus V_{DD}. Inset: Total number of read-failures versus V_{DD}.

With a total energy dissipation of 14 fJ per accessed bit and a leakage power of 500 fW per stored bit, the presented work outperforms all previous works in 65nm CMOS nodes. The reported clock frequencies are suitable for a wide range of biomedical applications, while most previously reported sub-V_T SRAMs are overdesigned. Nevertheless, silicon measurements from a further test chip (not expatiated on in this thesis) show that the the frequency can be improved by 5× (reaching 100 kHz at 0.45V) at only a small area and leakage power overhead (600 fW/bit instead of 500 fW/bit) if the read bit-line (RBL) is segmented, limiting the number of tri-state drivers per segment to 8, and using conventional CMOS multiplexers to choose a segment [83]. Moreover, if using custom-designed low-leakage latches and only CMOS multiplexers, integrating the first stage (a NAND gate) of the multiplexer as output buffer of the storage cell (latch), the frequency is even improved to 200 kHz (at 0.45V) at the cost of higher area cost and leakage power (700 pW/bit) [83]. Finally, the silicon area of SCMs is smaller compared to sub-V_T SRAM hardmacros for storage capacities of up to several kb, due to less area for peripheral circuits (see Section 3.2.4). However, for several tens of kb, an area-increase of roughly 4× [33], stemming from the larger bitcell, is often acceptable for the

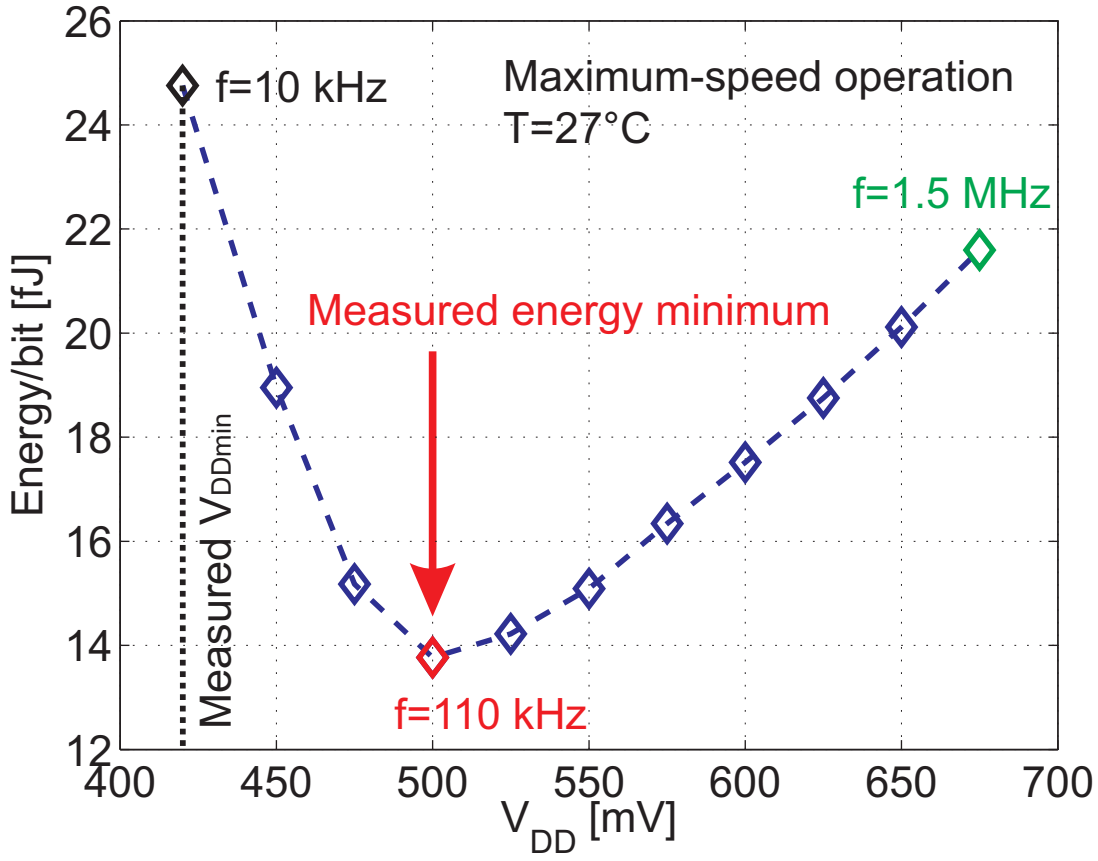


Figure 3.17: Measured energy per bit-access.

Table 3.4: Comparison with prior-art sub-V_T memories in 65 nm CMOS.

	[6]	[77]	[101]	**This work [82]**
V_{DDmin} [mV]	380	250	700	**420**
V_{DDhold} [mV]	230	250	500	**220**
$E_{tot/bit}$ [fJ/bit]	54 (0.4V)	86 (0.4V)	-	**14 (0.5V)**
$P_{leak/bit}$ [pW/bit]	7.6 (0.3V)	6.1	6.0, 1.0[a]	**0.5**

[a] Leakage-power of bitcell only

benefit of the clearly lower leakage power and access energy. While this Section presented the lowest ever measured data retention power per bit in a conventional 65 nm CMOS technology, the following Section investigates the integration of an emerging memory device (an oxide stack, or "memristor") into a non-volatile CMOS flip-flop for zero-leakage standby states in future ULP/ULV systems.

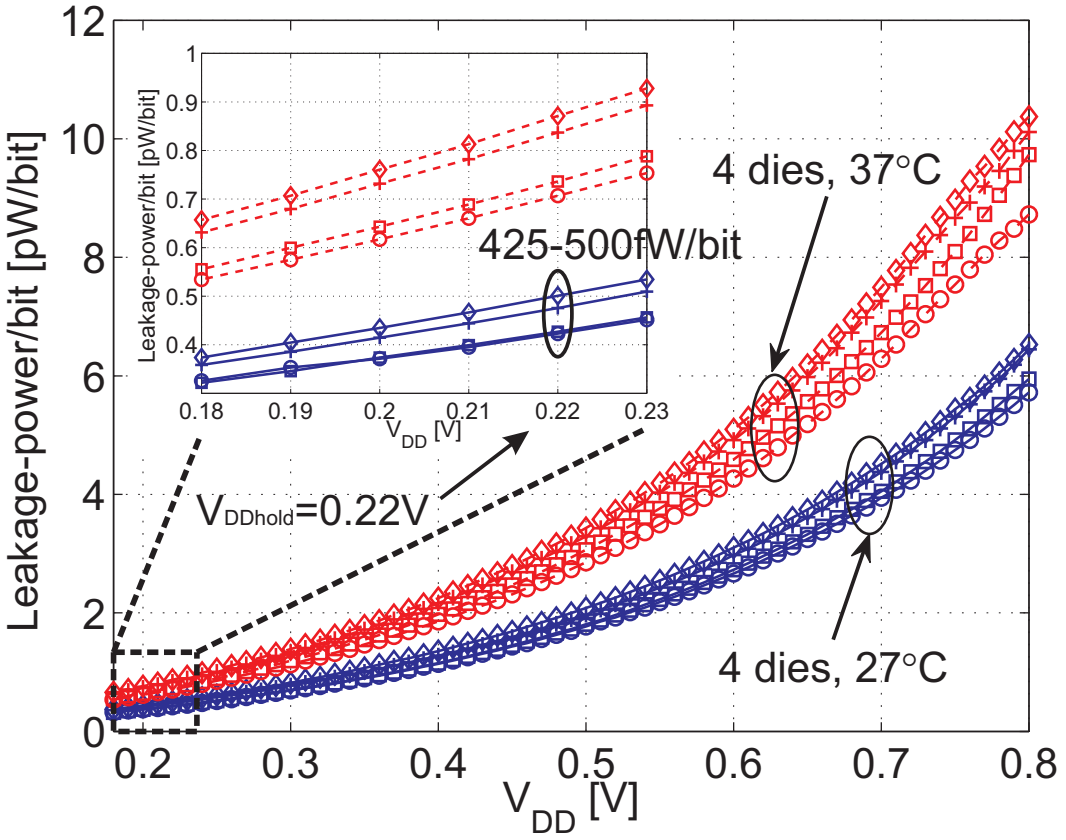


Figure 3.18: Measured leakage power per bit, including overhead of peripheral circuits, measured for 4 dies, at 27 and 37 °C. Inset: Zoom around V_{DDhold}.

3.4 ReRAM-Based Non-Volatile Flip-Flop (NVFF) Topologies

While near-threshold (near-V_{T}) and subthreshold (sub-V_{T}) circuit operation enables extremely low leakage power for on-chip storage elements, emerging device technologies allowing the integration of non-volatile memory devices on CMOS chips bear the potential of zero-leakage sleep states [102]. Such non-volatile, zero-leakage storage elements are especially important for systems characterized by only short active periods and long sleep periods requiring data and program state retention, whose total power budget is otherwise dominated by the leakage power of retentive, volatile memories in CMOS technology. Among many technological options, oxide memories (OxRAMs) [103] are a promising candidate for next generation, CMOS-compatible, non-volatile memory arrays. Compared to traditional Flash memories, OxRAMs have better scalability and faster programming time. While a lot of research effort targets OxRAM-based stand-alone memories, the hereinafter presented work focuses on the seamless integration of OxRAM devices into CMOS flip-flops for use in zero-leakage, non-volatile SCMs or in state registers.

Previous works on non-volatile flip-flops were based on the "memristor" [104, 102] from Hewlett Packard (HP), on bipolar OxRAM [105], and magnetic tunneling junction (MTJ) devices [106, 107, 108]. All these works considered circuit operation at a high supply voltage, normally corresponding to the CMOS technology's nominal voltage.

In the remainder of this Section, we design a non-volatile flip-flop exploring the benefits of OxRAM devices. We combine for the first time the advantages of sub-V_T and near-V_T circuit operation with OxRAMs, thereby enabling VLSI systems with ultra-low active energy dissipation in addition to non-volatile memory storage with zero leakage. The ULP and especially the biomedical design community often prefers to use mature technology nodes for 1) high reliability; 2) low leakage currents; and 3) low cost. Therefore, this study adopts a mature 0.18 µm CMOS process. The proposed non-volatile flip-flop, to be used within standard-cell based memories (SCMs) or yet in status and/or pipeline registers, operates reliably in the sub-V_T regime. Indeed it reliably recovers the saved data on wake-up with a sub-V_T supply voltage and a standard deviation of up to 5% of the nominal value of the ReRAM resistance. In the proposed design, write energy is ReRAM technology dependent while the read energy can be optimized at circuit level. Thanks to sub-V_T operation, the read energy has been drastically reduced down to 5.4% of the total read+write energy. Beside the main novelty of designing hybrid CMOS/OxRAM circuits for reliable operation in the sub-V_T and near-V_T regime, a number of additional factors distinguishes this work from previous works: 1) All simulations are based on real CMOS technology data (while some previous work used predictive technology models); 2) the OxRAM devices considered in this study have been fabricated, characterized, and modeled in-house by our research partners at EPFL; 3) parametric variations are considered not only for the MOS transistors, but also for the OxRAM devices; 4) energy characterization has been done for read and write operations. In the following, Section 3.4.1 reviews the manufactured ReRAM stacks that serve as the starting point for this circuit-level work, while Section 3.4.2 discusses the proposed NVFF architecture, before detailed simulation results are presented in Section 3.4.3.

3.4.1 ReRAM Manufacturing Process and Switching Characteristics

Among many ReRAM candidates, OxRAMs base their working principle on the change in resistance of an oxide layer. Different physical mechanisms can be identified in the switching of ReRAMs [103]. In the following, we will focus only on the bipolar resistive switching (BRS) [109], related to the O_2 vacancy redistribution in TiO_2 layers upon application of a voltage across the oxide. We realized memory stack prototypes of Al/TiO_2/Al from bulk-Si wafers passivated by a 100 nm thick Al_2O_3 layer. 70 nm thick bottom electrode (BE) lines were patterned by lift-off and e-beam evaporation. Then, a 50 nm thick TiO_2 layer was deposited by atomic layer deposition (ALD) at 200°C. Finally, vertical top electrode (TE) lines were defined with a second lift-off step together with contact areas used for electrical characterization. Such nodes are expected to be embedded within standard top-layer metal vias.

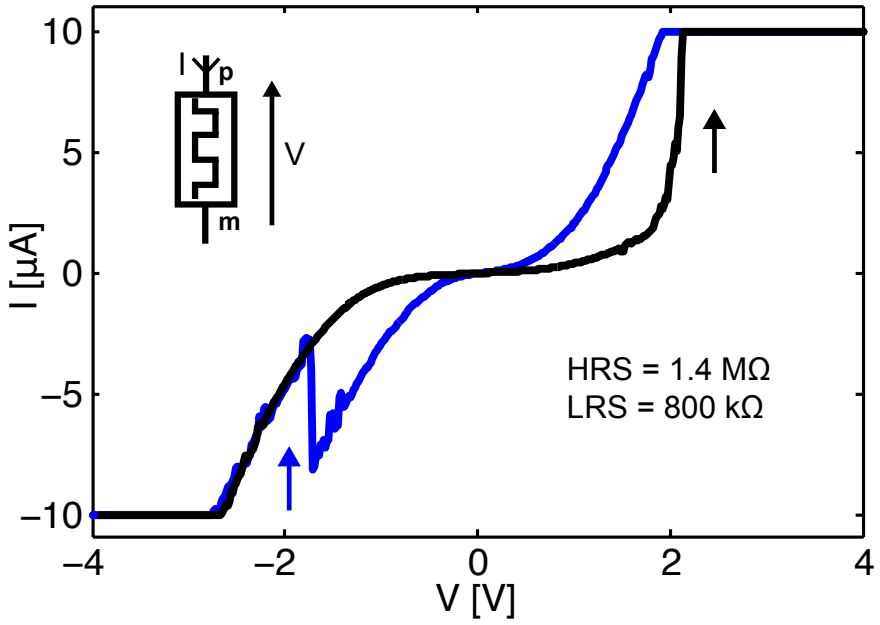


Figure 3.19: 1.5 µm^2 Al/TiO_2/Al ReRAM stack switching under 10 µA current compliance [7].

As opposed to most ReRAMs, the devices used in this work do not require a forming operation. Instead, the resistive switching functionality is obtained by cycling the memory. After 50 cycles, the resistive switching behavior stabilizes to the behavior shown in Fig. 3.19. Consistent BRS with a high resistance state (HRS) and a low resistance state (LRS) is achieved. The SET and RESET threshold voltages range from −2 V to +2 V. Moreover, the switching operation is limited by a low current compliance of 10 µA, allowing the use of small (close to minimum size) programming transistors. As opposed to this, most previously reported ReRAMs require much higher current (around 1–10 mA) to switch successfully, which needs prohibitively wide transistors to drop a sufficiently high voltage across the ReRAM.

3.4.2 Non-Volatile Flip Flop Architecture and Operation

This Section explains the design and the operating principle of the proposed ReRAM-based non-volatile flip-flop. A first design is suitable for operation at nominal and near-V_T supply voltages, while a second version is specifically optimized for robust operation in the sub-V_T domain.

Architecture

A conventional master-slave flip-flop based on tri-state inverters serves as a starting point, as shown in Fig. 3.20 in blue color. In order to add non-volatility to this basic CMOS flip-flop, two ReRAM devices are inserted in the current sink of the cross-coupled inverter pair in the slave latch [110, 111]. These ReRAM devices are used in a complementary way, i.e., one

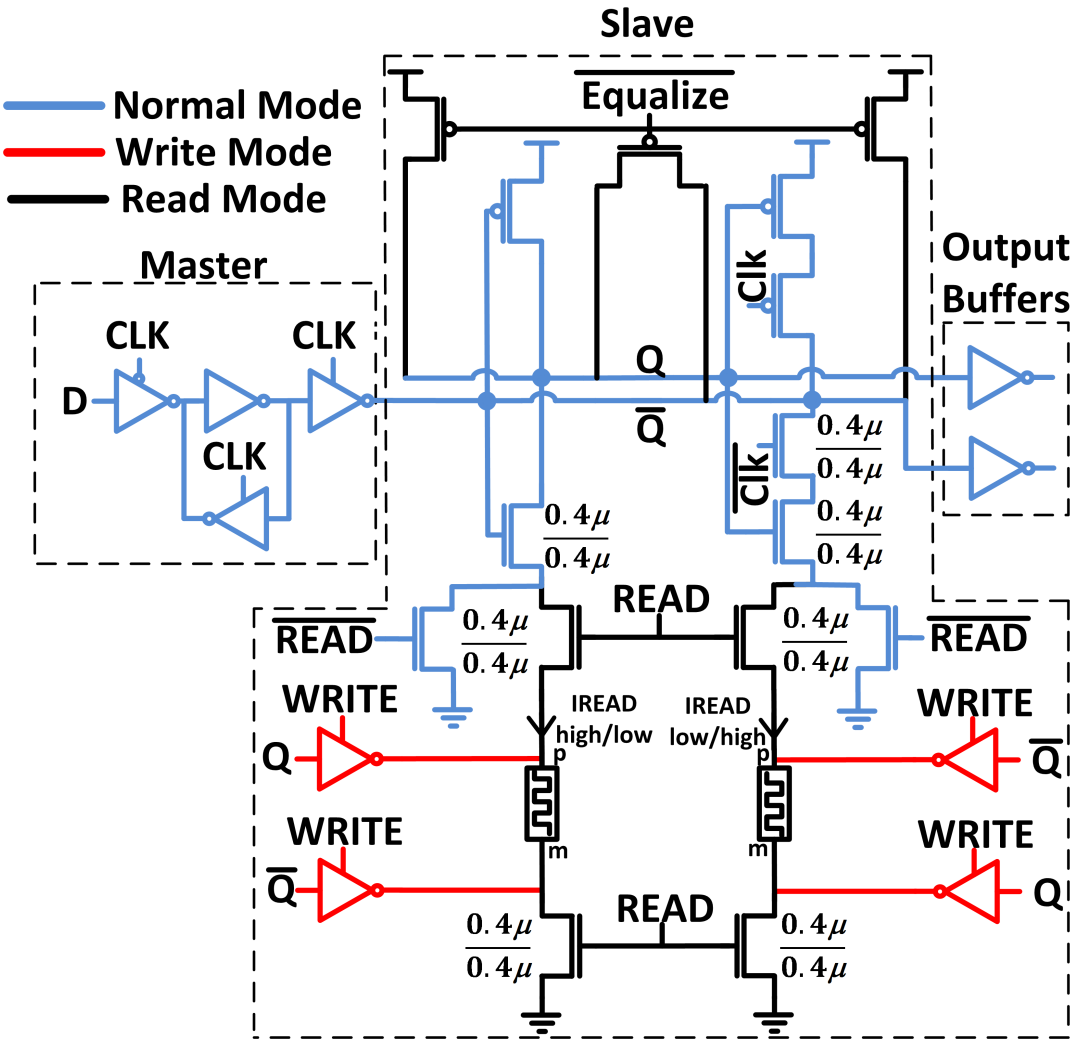


Figure 3.20: ReRAM-based non-volatile flip-flop for above-V_T operation; circuit parts are highlighted in colors according to their activation for different operating modes.

device is programmed to the HRS, while the other one is programmed to the LRS. Dedicated programming (or ReRAM write) circuits are highlighted in red color, while dedicated restore (or ReRAM read) circuits are shown in black color.

During normal operation, all ReRAM write and read circuits are disabled, and both branches of the slave latch are properly grounded through two NMOS transistors (controlled by $\overline{\text{READ}}$). Consequently, the hybrid CMOS/ReRAM non-volatile flip-flop fully relies on CMOS transistors during normal operation, which are known to exhibit high endurance. The part of the circuit containing ReRAMs is only activated during the preparation of a sleep state or during wake-up. Therefore, the ReRAMs, whose endurance is not yet comparable with the one of CMOS transistors, do not switch very frequently, which guarantees high overall system lifetime.

ReRAM Write Operation, or Flip-Flop Store

For the entire duration of ReRAM write and read, the clock needs to be silenced and kept low, as shown in Fig. 3.21, in order for the slave latch to be non-transparent and isolated from the master. During write, the ReRAMs are disconnected from the slave latch and from the read circuits, so that the voltage drop across their terminals can be set by the write drivers (highlighted in red in Fig. 3.20). The write drivers are controlled by the internal nodes Q and $\overline{\text{Q}}$. A write pulse width of 10 ns is used to program the ReRAMs. As illustrated in Fig. 3.19, a voltage of +2 V or −2 V is required for successful switching. To be able to use small programming transistors (with a non-negligible voltage drop across their channel) and limit the programming current, the write drivers are supplied with a voltage as high as 2.4 V. This voltage is only slightly above the nominal supply voltage range of the core transistors in the considered 0.18 µm CMOS technology and does neither seriously enhance the risk of oxide break-down, nor compromise junction reliability, nor considerably accelerate aging (in particular due to the infrequent and short write cycles to the ReRAM device).

Two architectural alternatives for the distribution of the 2.4 V supply may be adopted: 1) the supply voltage of all non-volatile flip-flops in the VLSI system is temporarily increased. This can safely be done without the need for level shifters, even if the rest of the system is biased in the sub-V_{T} domain, as the slave latch already holds data and the clock signal is constantly low. The energy overhead is kept small by rising only the supply of the non-volatile flip-flops; or 2) the entire VLSI system as well as the CMOS part of the flip-flop and the read circuits are constantly biased at a low supply voltage, while the CMOS write drivers are constantly supplied with 2.4 V. This alternative avoids the energy overhead associated with dynamically charging the capacitive power distribution network, but requires a level shifter in each flip-flop if the main power supply is considerably lower than 2.4 V. In this study, we adopt the first approach of dynamically rising the supply voltage during a write operation, as shown in Fig. 3.21. Once the ReRAMs are programmed, the power supply can be completely turned off, enabling a zero-leakage sleep state.

ReRAM Read Operation, or Flip-Flop Restore

During system wake-up (power-on), the slave latch would ideally be directly restored, based on the data stored in the ReRAMs, during ramp-up or connection of the power supply. However, this is impossible due to a number of reasons: 1) the clock and the READ signal are not controlled yet; 2) there might be uncontrolled, residual charges on the internal nodes Q and $\overline{\text{Q}}$; and 3) different power-gating approaches (mechanical, footer and/or header transistors, driving the supply to ground level) result in different wake-up scenarios. Therefore, the following wake-up sequence is proposed, as shown in Fig. 3.21: 1) turn on the power supply; 2) at the system level, silence the clock signal to low; 3) enable the READ and the EQUALIZE signals; and 4) upon de-assertion of EQUALIZE, the slave latch is correctly restored based on the value of the ReRAMs. Both nodes Q and $\overline{\text{Q}}$ are pre-charged and equalized using three dedicated PMOS transistors controlled by $\overline{\text{EQUALIZE}}$.

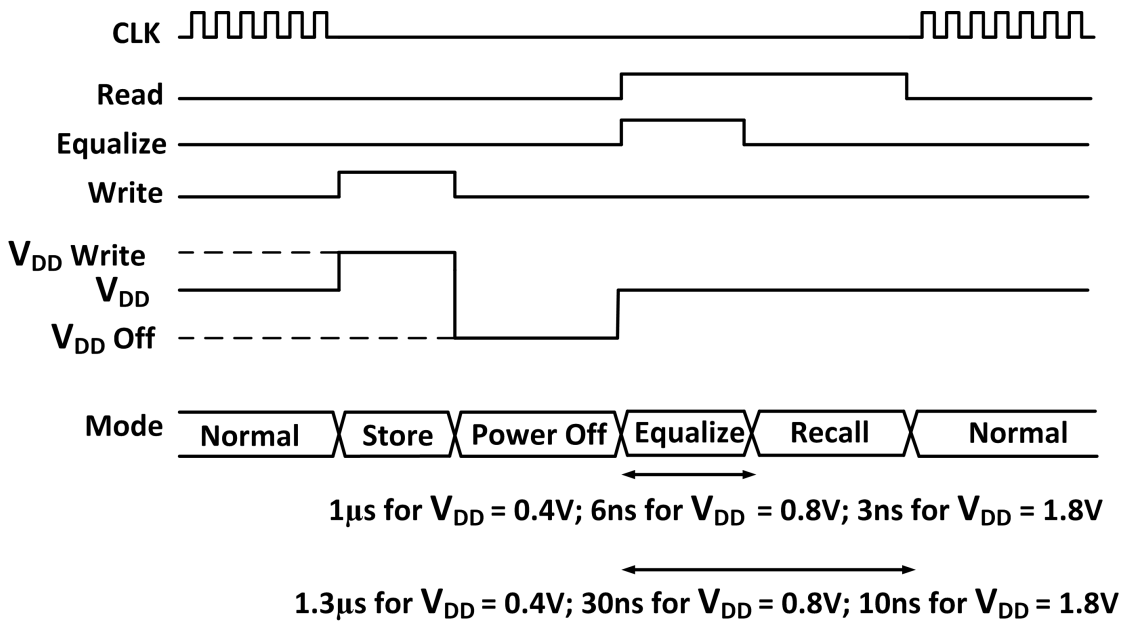


Figure 3.21: Control signals sequence for ReRAM read and write operations.

Following this pre-charge phase, the READ is asserted. At this time, the pre-charged, internal nodes Q and $\overline{\text{Q}}$ are connected to ground through the ReRAMs. The complementary resistance state of the two ReRAMs modulates the discharge currents (the branch with HRS has a lower discharge current than the branch with LRS), starting a race condition. As soon as one internal node is discharged to $V_{\text{DD}} - V_{\text{T,PMOS}}$, the PMOS transistor driven by that node turns on and starts to pull up the other internal node. This decides the race, before the feedback of the latch restores full logic levels.

Modifications for Robust Sub-V_{T} Operation

A correct read depends on the modulation of the discharge current by the complementary ReRAMs. However, referring to Fig. 3.20, the discharge current might be altered due to other reasons: 1) different pull down networks in the two branches due to the use of a simple inverter on one side and a tri-state inverter on the other side; and 2) mismatch between transistor pairs (in the inverters and in the dedicated read transistors) and ReRAMs, caused by local variations. For high supply voltages (0.8–1.8 V), the rather small ratio between the HRS and the LRS (around 2) is still high enough to overcome these alterations in discharge current. In fact, the circuit shown in Fig. 3.20 reads correctly under within-die parametric variations (for both the MOS transistors and the ReRAMs), provided that the pull-down strength of the single inverter is engineered to match the one of the tri-state inverter under nominal conditions. However, for operation in the sub-V_{T} domain (for example at 0.4 V), the following modifications are necessary to ensure correct read, as shown in Fig. 3.22: 1) the circuit needs to be fully symmetric; to this end, two always-on transistors (D_n and D_p) are inserted into

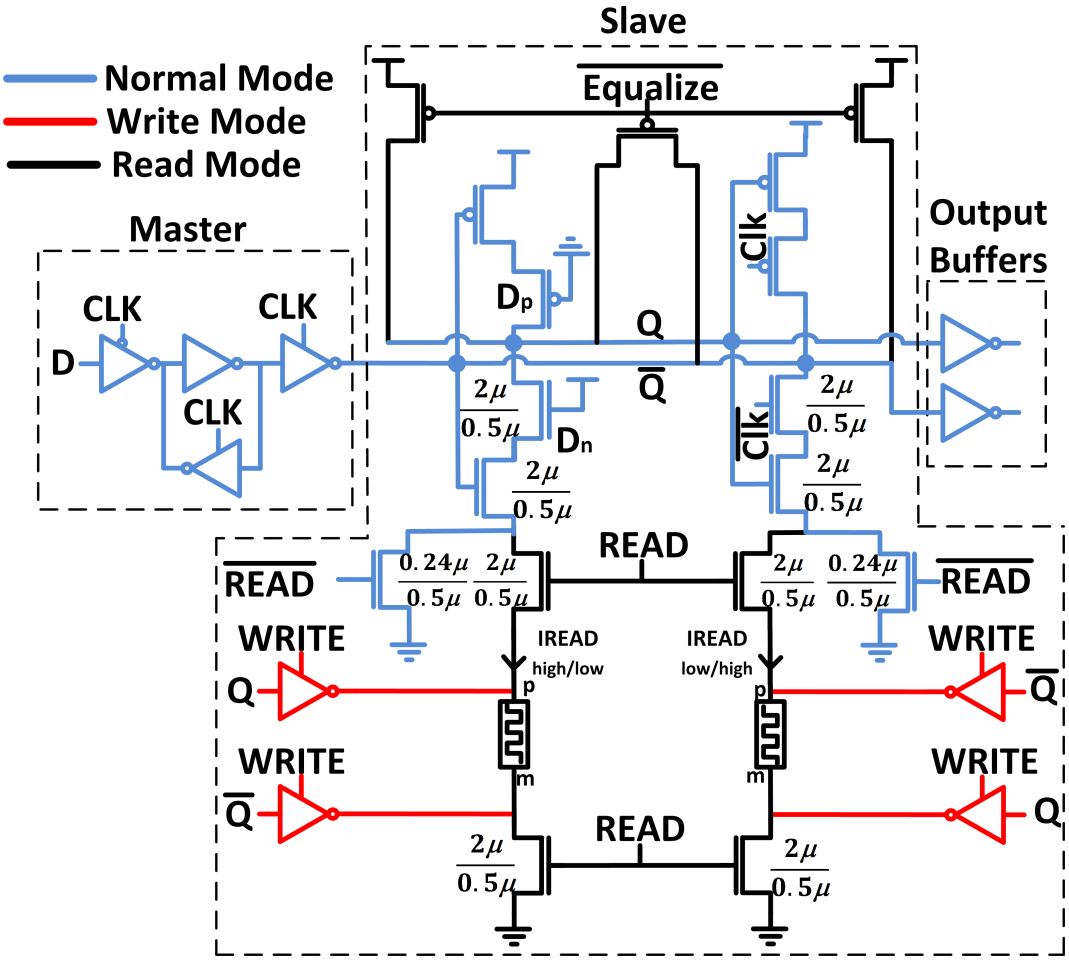


Figure 3.22: ReRAM-based non-volatile flip-flop optimized for robust sub-V_T operation; circuit parts are highlighted in colors according to their activation for different operating modes.

the simple inverter to mimic the tri-state inverter, 2) all transistor pairs are upsized for better matching.

3.4.3 Simulation Results

This Section verifies the robustness, with special emphasis on the ReRAM read operation, and characterizes the energy for both previously introduced non-volatile flip-flop architectures, optimized for above-V_T and sub-V_T operation, respectively. All simulations run by Spectre assume a typical-typical (TT) process corner at 27°C. A dynamically adjustable power supply is presumed, switching between 2.4 V for write operations, and a lower value ($V_{DD,read}$) for read as well as normal operation (flip-flop sampling operation). $V_{DD,read}$ assumes the technology's nominal value (1.8 V), a near-V_T value (0.8 V), and a sub-V_T value (0.4 V). Monte Carlo circuit simulations (1000 runs) account for local parametric variations of all MOS transistors, according to statistical distributions provided by the foundry. While sophisticated statistical models

of the ReRAMs are not available yet, we assume that the HRS and the LRS follow a Gaussian distribution. The measured, nominal values of HRS (1.4 MΩ) and LRS (800 kΩ) are taken as mean values, denoted by μ(HRS) and μ(LRS), respectively. The values 40 kΩ, 80 kΩ, and 160 kΩ corresponding to 5%, 10% and 20% of μ(LRS), respectively, are taken for the standard deviations, denoted by σ(HRS) and σ(LRS).

Sub-V_T Robustness Analysis

Among normal sampling, write, and read operations, read is the most critical one. Studies have shown that normal operation of CMOS flip-flops can be robust in the sub-V_T domain [33], while the write operation uses an elevated supply voltage. The read operation of the non-volatile flip-flop topology built for above-V_T operation (see Fig. 3.20) is simulated at 0.8 V, while the topology optimized for sub-V_T operation (see Fig. 3.22) is evaluated at both 0.8 V and 0.4 V. An appropriate metric to assess the read robustness is the initial discharge current (I_{read}) flowing through the two branches of the slave latch right after the de-assertion of the EQUALIZE signal. Fig. 3.23 shows the distributions of I_{read} for the sub-V_T-optimized topology, at 0.4 V, for different standard deviations of HRS and LRS. For a well-controlled, repeatable ReRAM process with σ(HRS) = σ(LRS) = 40 kΩ, the discharge current flowing through the branch containing the ReRAM in the HRS is clearly lower than the current flowing through the other branch (non-overlapping I_{read} distributions). This results in zero read failures out of 1k Monte Carlo runs, as shown in Fig. 3.24. For a less precisely controlled ReRAM process with higher standard deviation of the resistance (σ(HRS) = σ(LRS) = 160 kΩ), the distributions of I_{read} start to overlap, which results in a small read failure probability of around 4%. Finally, Fig. 3.24 illustrates the high effectiveness of the proposed circuit optimizations for robust sub-V_T operation: the optimized circuit, supplied with 0.4 V, exhibits a much lower read failure probability than the initial, unoptimized circuit, even if the latter is supplied with a higher voltage of 0.8 V. For a badly controlled ReRAM process, rising the supply voltage of the optimized circuit from 0.4 V to 0.8 V yields a virtually zero read failure probability, while, of course, the read failure probability remains zero for a well-controlled ReRAM process.

Energy Characterization

Fig. 3.25 shows the energy dissipation of a single read and write operation of the non-volatile sub-V_T flip-flop (see Fig. 3.22). The main power supply V_{DD} (used for read and normal operations) is swept from 1.8 V to 0.4 V. Prior to a write operation the power supply is always risen to 2.4 V. For each V_{DD}, the read operation is performed at maximum speed, with the minimum required pulse widths for EQUALIZE and READ signals, given in Fig. 3.21. Initially, voltage scaling from 1.8 V to 0.8 V considerably reduces the read energy; however, the active energy benefits of further scaling are offset by longer pulse widths at 0.4 V (in the order of µs instead of tens of ns) and the associated integration of leakage currents.

For a main V_{DD} of 1.8 V and 0.4 V, the supply needs to be risen by 0.6 V and 2 V for a write

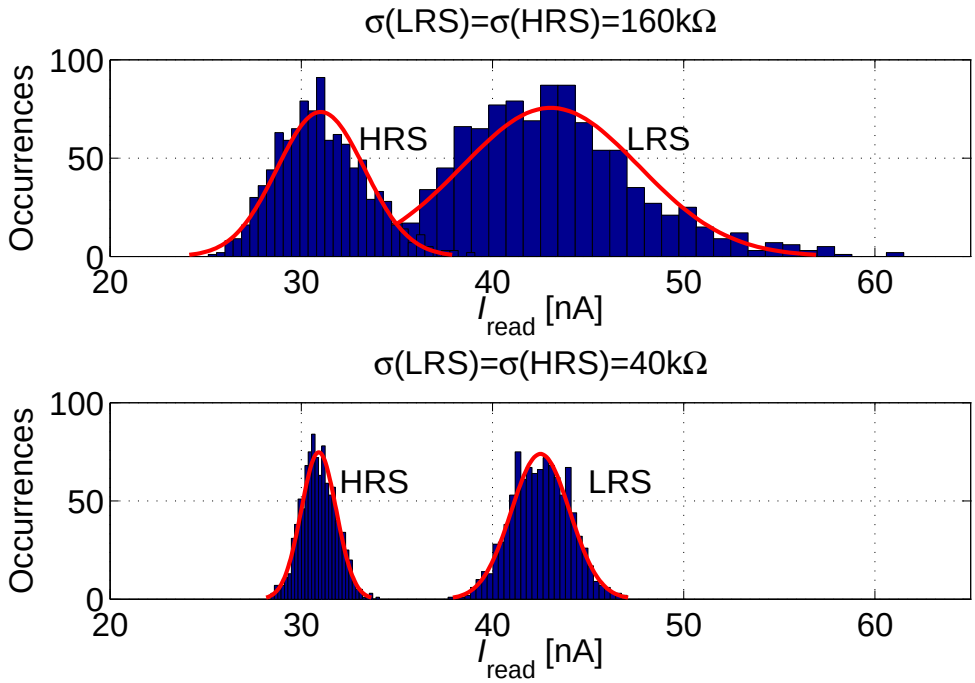


Figure 3.23: Statistical distribution of the discharge current (I_{read}) through the two branches of the slave latch of the sub-V_T-optimized non-volatile flip-flop, for 0.4 V, given for two different standard deviations of the ReRAM's resistance.

operation, respectively. As illustrated in Fig. 3.25, the lower the main V_{DD} is, the larger the transition to 2.4 V, and the larger the write energy. For comparison, Fig. 3.25 also shows the energy cost of 5 normal sampling operations at 100 MHz, 1 MHz, and 100 kHz for 1.8 V, 0.8 V, and 0.4 V, respectively. Finally, the minimum total energy for sleep preparation and wake-up, found at 0.8 V, is 735 fJ. The write energy mostly depends on the ReRAM stack, whereas the read energy depends on the circuit topology. A direct comparison with previous work is difficult due to missing energy reports and a multitude of different ReRAMs. However, the total read+write energy of the sub-V_T-optimized circuit is compared with the energy of the leakage-optimized latch from Section 3.3. This shows that the sub-V_T-optimized non-volatile flip-flop is more energy efficient for system sleep times longer than 1.47 s.

3.5 Conclusions

This Chapter has addressed the lack of good ultra-low power (ULP) sub-V_T memory compilers by utilizing a fully automated standard-cell based memory (SCM) compilation flow, especially interesting for ULP systems (such as biomedical systems) requiring only small storage capacities of several kb.

In fact, for standard-cell based ultra-low power designs which need to operate in the sub-V_T

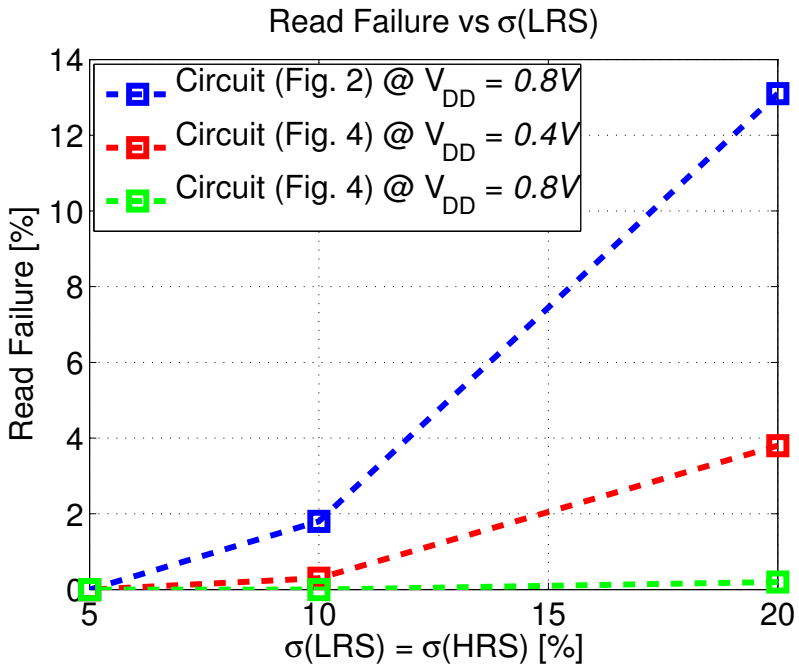


Figure 3.24: Read failure probability for a ReRAM resistance's standard deviation of 5%, 10%, and 20% of the nominal LRS value. Parametric variations of MOS transistors are also accounted for, according to statistical distributions provided by the foundry.

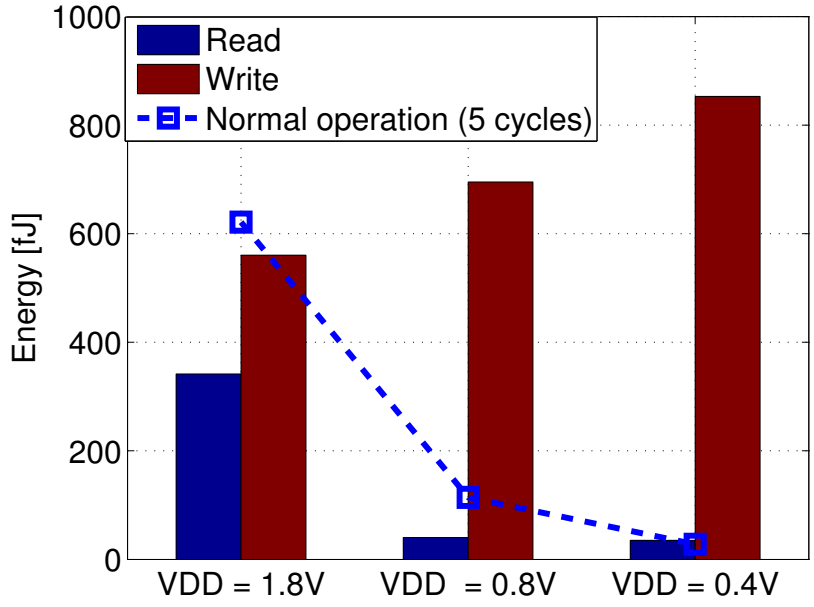


Figure 3.25: Energy for read, write and five clock cycles of normal operation of the sub-V_T-optimized non-volatile flip-flop.

regime, standard-cell based memories (SCMs) are an interesting alternative to full-custom SRAM macrocells which must be specifically optimized to guarantee reliable operation by using 8T, 10T, ..., 14T bitcells and/or low-voltage write and read assist circuits. The main advantages of SCMs exclusively synthesized from commercial standard-cell libraries (SCLs) are the reduced design effort, reliable operation for the same voltage range as the associated logic, high speed (when compared to corresponding full-custom sub-V_T SRAM macrocells), and reasonably good energy efficiency for maximum-speed operation. The drawbacks are the area penalty (for storage arrays larger than a few kb) and a loss in energy efficiency compared to full-custom designs when operating at the same clock frequency.

Energy-efficient sub-V_T SCM design is driven by the fact that most of the energy is consumed due to leakage while active energy plays only a minor role, especially for large configurations. Considering only commercial SCLs, a design based on latches using clock-gates for the write logic and glitch-free multiplexers for the read logic achieves the best energy efficiency and has the smallest silicon area. For the same maximum throughput but smaller write address setup-times, the latches may be replaced by flip-flops. If the analysis is limited to commercial SCLs for minimum design effort, the best-practice SCM implementations for above-V_T operation (previously identified in Chapter 2) and for sub-V_T operation coincide. This means that the chosen SCM topology supports dynamic voltage and frequency scaling (DVFS) while remaining the optimum topology irrespective of the supply voltage.

Unfortunately, commercial SCLs are primarily optimized for high speed at nominal voltage, but not for low leakage or high robustness at sub-V_T voltages. In order to aggressively push for ultra-low leakage power and access energy (at the cost of a speed degradation) with a small extra effort for only one custom-designed standard-cell, it is best to use a latch with tri-state inverters, transistor stacks, and channel length stretching, as well as a tri-state read logic. In fact, the design and integration into the SCM compilation flow of a single standard-cell (D-latch with 3-state output buffer) addresses all dominant SCM leakage contributors at once and cuts the leakage power of SCMs into half compared to using only commercial SCLs. Silicon measurements show that a 3-state read logic with up to 128 words per bit-line operates reliably in the sub-V_T regime down to 420 mV. Counter to intuition, weaker 3-state buffers not only reduce leakage, but also shorten the bit-line delay compared to stronger 3-state buffers as the total leakage current of all disabled 3-state buffers becomes significant compared to the active drive current in the sub-V_T regime. A 4kb SCM manufactured in 65 nm CMOS technology consumes a leakage power of 500 fW per stored bit (at data-retention voltage of 220 mV) and dissipates a total (active and leakage) energy of 14 fJ per accessed bit (at energy-minimum voltage of 500 mV), thereby outperforming all previously reported sub-V_T memories in 65 nm CMOS technology.

While the sub-V_T SCMs based on custom-designed ultra-low-leakage latches and read logic already exhibit an extremely low standby leakage power, non-volatile flip-flop (NVFF) circuits based on emerging ReRAM technology have been proposed, as well. These NVFFs leverage the use of sub-V_T operation to enable future energy-efficient VLSI systems with zero-leakage

sleep states. The considered oxide stacks switch their resistive state with a 0.18 µm CMOS-compatible voltage of 2 V and under a low current compliance of 10 µA. The write energy is mostly ReRAM technology dependent. Thanks to sub-V_T and near-V_T operation the read energy is brought down to 5.4% of the total read+write energy. The read energy improvement saturates between near-V_T and sub-V_T due to the increase in the minimum required READ pulse time. With the currently used OxRAM technology, the break-even sleep time for which the use of the sub-V_T NVFF circuit results in net energy savings compared to our retentive 500 fW leakage-power latch is 1.47 s. Monte Carlo simulations demonstrate a robust restore operation (ReRAM read operation) at 0.4 V, accounting for parametric variations in both ReRAM devices and MOS transistors. Robustness can be further increased by having a larger ratio between the high and low resistance values of the ReRAM.

Briefly, sub-V_T SCMs are very convenient to implement robust embedded memories operated at ultra-low voltages, due to the lack of good sub-V_T SRAM compilers. Beside voltage scaling, SCMs also support technology down-scaling and can easily be adopted as soon as a commercial standard-cell library becomes available. Thanks to the design of a dedicated, ultra-low leakage standard-cell, our silicon-proven sub-V_T SCMs exhibit lower leakage power and access energy per bit compared to all prior-art sub-V_T SRAMs in a 65 nm CMOS node. Our sub-V_T non-volatile flip-flop based on oxide memory (OxRAM) devices enables energy savings for relatively long sleep times in the order of seconds; this type of embedded, non-volatile storage element can be interesting for environmental monitoring, sensor networks, or periodic health monitoring systems which perform a sensor readout every hour or so. While conventional, purely CMOS based, volatile SCMs can immediately and reliably be used in every VLSI system requiring small storage arrays of several kb, the large adoption of ReRAM-based, non-volatile SCMs will become interesting in future VLSI SoCs as soon as the ReRAM manufacturing processes (e.g., for oxide stacks) become mature enough to guarantee high yield.

4 Gain-Cell Based eDRAMs (GC-eDRAMs)

While 6T-bitcell SRAM macrocells are the mainstream solution for memories embedded in VLSI SoCs, and while standard-cell based memories (SCMs) can be an interesting replacement for SRAM in many cases (as discussed in Chapter 2 and Chapter 3), embedded dynamic random-access memory (eDRAM) is a further alternative to implement embedded memories. The conventional eDRAM bitcell uses a dedicated storage capacitor to store information in form of electric charge and a MOS transistor to access the basic bitcell for read and write operations; unfortunately, such conventional 1-transistor-1-capacitor (1T-1C)-bitcell eDRAM requires special processing steps to manufacture high-density stacked or trench capacitors and is therefore not directly compatible with standard digital CMOS technologies. As opposed to conventional 1T-1C eDRAM, gain-cell (GC) based eDRAM (GC-eDRAM) is fully logic-compatible since it is built exclusively from MOS transistors, used as access transistors and MOSCAPs, and optionally by the readily available metal stack and vias for enhanced storage node capacitance. As such, GC-eDRAM is an interesting alternative to 6T-bitcell SRAM and 1T-1C eDRAM, since it combines many of the advantages of SRAM (e.g., the logic compatibility) and 1T-1C eDRAM (e.g., higher density than SRAM), while it avoids most of the drawbacks of SRAM (e.g., large bitcell) and of 1T-1C eDRAM (e.g., destructive read, write-back operation, and extra cost for special processes). The main drawback of GC-eDRAM compared to SRAM is the need for a periodic refresh operation, unless the entire memory block is frequently and periodically updated with new data.

Section 4.1 discusses in detail the advantages and potential drawbacks of GC-eDRAM compared to SRAM and 1T-1C eDRAM, and provides a detailed review of the field of GC-eDRAM design, identifying not only bitcell and peripheral circuit techniques, but also the main target applications. All following Sections present our particular gain-cell bitcell and GC-eDRAM macrocell designs and analyses, targeting a large range of applications from robust low-voltage/low-power gain-cell storage arrays with extended retention times and low refresh power for systems operated at near-V_T and even sub-V_T supply voltages, to high-density storage arrays for high-performance, potentially error-resilient VLSI systems (operated at nominal voltage). More precisely, Section 4.2 studies the impact of voltage scaling on the retention

time of GC-eDRAM and shows that, counter to intuition, voltage scaling can improve the retention time in some cases, depending on the write access statistics and the write bit-line (WBL) control scheme. With this encouraging results, Section 4.3 considers voltage scaling to the near-threshold (near-V_T) domain, and proposes several techniques to extend the retention time of near-V_T GC-eDRAM, including reverse body biasing and replica cells for optimum refresh timing, in order to reduce the data retention power. Silicon measurements of a test chip containing several GC-eDRAM arrays verify the effectiveness of the various proposed retention time extension techniques. Next, Section 4.4 goes a step further in terms of voltage scaling and, for the first time, investigates the feasibility of GC-eDRAMs operated at sub-V_T supply voltages. It is shown that sub-V_T GC-eDRAM is a viable option in mature CMOS nodes (which are especially interesting for ultra-low power systems), while high leakage currents and low in-cell storage capacitors (built from the metal stack) lead to prohibitively short retention times in deeply scaled CMOS nodes. Therefore, the supply voltage should only be scaled down to the near-V_T domain for viable operation of GC-eDRAM in sub-40 nm CMOS nodes. Finally, Section 4.5 presents the design and analysis of a multilevel gain-cell eDRAM storing several bits per basic gain-cell for high storage density at the cost of a small read failure probability which can be tolerated by some error-resilient systems. Moreover, since access times of this multilevel GC-eDRAM are rather long, replica techniques for frequency guardband reduction, eventually leading to faster access times are presented, as well. Conclusions are drawn at the end of each Section.

This Chapter is mostly based on our previous publications [29, 63, 112, 65, 113, 114, 115].

4.1 Introduction to GC-eDRAM

A gain-cell is a dynamic memory cell built exclusively from MOS transistors, either used as write and read access transistors or as MOSCAPs, and optionally from parasitic capacitors between metal lines and vias to increase the in-cell storage capacitor. Therefore, a gain-cell is fully compatible with mainstream digital CMOS technologies, and GC-eDRAM macrocells can readily be integrated with any digital system at no additional manufacturing cost for special process options. A large variety of different gain-cell topologies has been proposed in the last decade, consisting of 2−4 transistors. All of them exhibit a write access device (MW) to access the capacitive storage node (SN) and deposit charge on it. Moreover, all gain-cell topologies have an SN capacitor which consists of a dedicated MOSCAP, the junction capacitance of MW, and in some cases of sidewall and parallel-plate capacitors built above the cell footprint with the available metal lines and vias. In the smallest 2-transistor (2T) gain-cell configuration, the dedicated storage transistor (MOSCAP) is also used as read transistor (MR); the 3-transistor (3T) gain-cell configuration exhibits a more robust read operation by using a separate MR. Some 4-transistor (4T) gain-cells use an additional MOSCAP to increase the SN capacitor and to capacitively couple the read bit-line (RBL) to the SN for increased read robustness. The term "gain-cell" stems from the transconduction gain of the read transistor MR, which translates a voltage level on the SN, or, equivalently, the gate voltage of MR into an output sense current

(the drain current of MR). From a similar viewpoint, the term “gain” can also relate to the fact that a small amount of charge on the SN leads to a large charge flow on the read bit-line (RBL) during readout thanks to the use of MR [116].

4.1.1 Advantages and Drawbacks of GC-eDRAM

GC-eDRAM has several advantages compared to both SRAM and 1T-1C eDRAM. In fact, a gain-cell is significantly smaller than a 6T SRAM bitcell; typically, area savings of at least 50% can be achieved by employing gain-cells instead of SRAM bitcells. Moreover, gain-cells have much lower aggregated bitcell leakage than SRAM bitcells. This reduced bitcell leakage current can even lead to lower data retention power, i.e., leakage power and active refresh power, for GC-eDRAM compared to the static leakage power of a corresponding SRAM macrocell [117]. Compared to conventional 1T-1C eDRAM, GC-eDRAM does not require any special processing steps to build high-density trench or stacked capacitors [30], which would require 4 to 6 extra masks and would add cost to a digital CMOS process [8]. As a further advantage compared to 1T-1C eDRAM, gain-cells enable a non-destructive read operation and thereby avoid the need for a write-back (restore) operation. Furthermore, compared to both the 6T SRAM bitcell and the 1T-1C bitcell, all gain-cell topologies have a separate read and write port, which allows to build two-port GC-eDRAM macrocells at virtually no area overhead compared to single-port macrocells. Both the 6T SRAM bitcell and the 1T-1C bitcell share the same bit-line(s) (BL) and word-line (WL) for both write and read accesses; additional hardware is required in each basic storage cell to allow simultaneous write and read access to a storage array built from SRAM or conventional DRAM cells. The use of two-port GC-eDRAM macrocells is appealing to ensure high memory bandwidth compared to single-port macrocells [118]; this can be especially interesting to recover some of the speed penalty resulting from voltage scaling (for low power consumption), or simply to ensure high access bandwidth for GC-eDRAMs used as caches in high-performance microprocessors. Finally, the separate write and read ports of all gain-cell topologies allow to independently and simultaneously optimize the bitcell for good write-ability and read-ability, which is especially important for the implementation of embedded memories in aggressively scaled CMOS nodes (characterized by high parametric variations) and/or operated at low voltages (in which case parametric variations become problematic due to degraded on/off current ratios). Note that the possibility to simultaneously and independently size the transistors in a gain-cell for robust read and robust write is a unique property of gain-cells which cannot be found in the 6T SRAM bitcell or in the 1T-1C eDRAM cell. In fact, in case of SRAM bitcells, additional transistors are required to avoid write contention and to improve read-ability. These various advantages of gain-cells compared to the traditional 6T SRAM and 1T-1C bitcells motivate the analysis and optimization of GC-eDRAM for use as embedded memories in a large variety of future VLSI SoCs implemented in scaled CMOS nodes and operated a scaled voltages.

Beside this long list of advantages, the main drawback of GC-eDRAM, compared to SRAM, is the dynamic storage mechanism, which requires periodic, power-consuming refresh cycles

(unless the memory block is anyway periodically updated, such as the internal memories of the LDPC decoder presented in Section 2.3.2). Compared to the conventional 1T-1C eDRAM bitcell, the total in-cell storage capacitor of gain-cells is considerably smaller, which leads to shorter retention times and requires more frequent refresh cycles. Also, there is a large variability of per-cell retention time across a GC-eDRAM array [64, 113], and, unfortunately, the global refresh rate needs to be set according to the gain-cell with the worst retention time, unless spare rows or columns in conjunction with programmable address decoders are used [18]. Later in this Chapter, in Section 4.2 and Section 4.3, we present several techniques to improve the retention time of GC-eDRAM in order to render it even more attractive for use in future VLSI systems. Before presenting our specific GC-eDRAM designs, a detailed review of the field of GC-eDRAM is presented in the following, which also positions our own work with respect to prior-art GC-eDRAM implementations.

4.1.2 Review of GC-eDRAM Target Applications and Circuit Techniques

Categorization of GC-eDRAM Implementations

From the large number of recent publications on GC-eDRAM, it is possible to identify four main categories of target applications: 1) high-end processors requiring large embedded cache memories; 2) general system-on-chip designs; 3) low-voltage low-power systems, such as biomedical systems; and 4) fault-tolerant systems including channel decoders for wireless communications.

Gain-Cells for High-End Processors The vast majority of recent research on GC-eDRAM is dedicated to large embedded cache memories for microprocessors [119, 120, 116, 121, 28, 122, 123, 124, 125, 126, 117]. In fact, GC memories are considered to be an interesting alternative to SRAM, which has been the dominant solution for cache memories for decades. This is due to the GC-eDRAM's higher density, increased speed, and potentially lower leakage power. Besides the obvious advantage of high integration density, the main design goal for GC memories in this application category are high speed operation and high memory bandwith, especially for industrial players like IBM [121] and Intel [28, 122], and recently also for academia [126, 117]. A smaller number of research groups specify low power consumption as their primary design goal [124, 125]. A recent study shows that in fact, as mentioned before, GC memories can potentially consume less data retention power (i.e., the sum of leakage power and refresh power) than SRAM arrays (leakage power only) [117].

General Systems-on-Chip (SoCs) Several authors are not very specific about their target applications [127, 128, 129], as they only mention general SoCs. However, they follow the same trend as the aforementioned processor community by proposing GC memories as a replacement for the mainstream 6T-bitcell SRAM solution. For these SoC applications, the main drivers are the potential for higher density and lower power consumption than SRAM.

Gain-Cells for Ultra-Low Power (Biomedical) Systems While the previously described target applications require relatively high memory bandwidth, several recent GC memory publications target low-voltage low-power applications (mostly in the biomedical domain). A GC memory implemented in a mature low-leakage 180 nm CMOS process achieves low retention power through voltage scaling well below the the nominal supply voltage [130]. The positive impact of supply voltage scaling on retention time for given access statistics and a given write bit-line control scheme is demonstrated in our own work [65] and expatiated on in Section 4.2, proposing near-threshold (near-V_T) operation for long retention times and therefore low retention power. We have also proposed reverse body biasing (RBB) [113] and replica techniques in order to further enhance the retention time and reduce the power consumption of near-V_T GC-eDRAM macrocells, as will be shown in Section 4.3. Moreover, our recent studies [114, 115] show that the supply voltage of GC arrays can even be scaled down to the subthreshold (sub-V_T) domain, while still guaranteeing robust operation and high memory availability for read and write operations; more details on these studies follow in Section 4.4.

Gain-Cells for Wireless Communications Systems A small number of recently presented GC memory designs, including some of our own designs, are fundamentally different from the aforementioned works, as they are specifically built and optimized for systems which require only short retention times, and in some cases, are tolerant to a small number of hardware defects (read failures) [21]. The refresh-free GC memory used in a recently published low-density parity-check (LDPC) decoder is periodically updated with new data, and therefore requires a retention time of only 20 ns [50]. Besides safely skipping power-hungry refresh cycles and designing for low retention times, our own works in [63, 112], presented in more detail in Section 4.5, also exploit the fact that wireless communications systems and other fault-tolerant systems are inherently resilient to a small number of hardware defects. In fact, by proposing memories based on multilevel GCs, the storage density of GC memories is further increased at the price of a small number of read failures which do not significantly impede the system performance [63, 112].

Comparison of State-of-the-Art Implementations Fig. 4.1 shows the bandwidth and the technology node of state-of-the-art GC memory implementations, highlighted according to target application categories. References appearing multiple times correspond to different operating modes or operating points of the same design. The figure shows a difference of more than four orders-of-magnitude in the achieved memory bandwidth among the various implementations. GC memories designed as cache memory for processors achieve around 10 Gb/s if implemented in older technologies and over 100 Gb/s if implemented in a more advanced 65 nm CMOS node. Most memories designed for wireless communications systems or generally for SoCs still achieve bandwidths between 1 and 10 Gb/s. Only the high-density multilevel GC array has a lower bandwidth due to a slow successive approximation multilevel read operation [112]. GC memories targeted towards biomedical systems are preferably implemented in a mature, reliable 180 nm CMOS node and achieve sufficiently high bandwidths

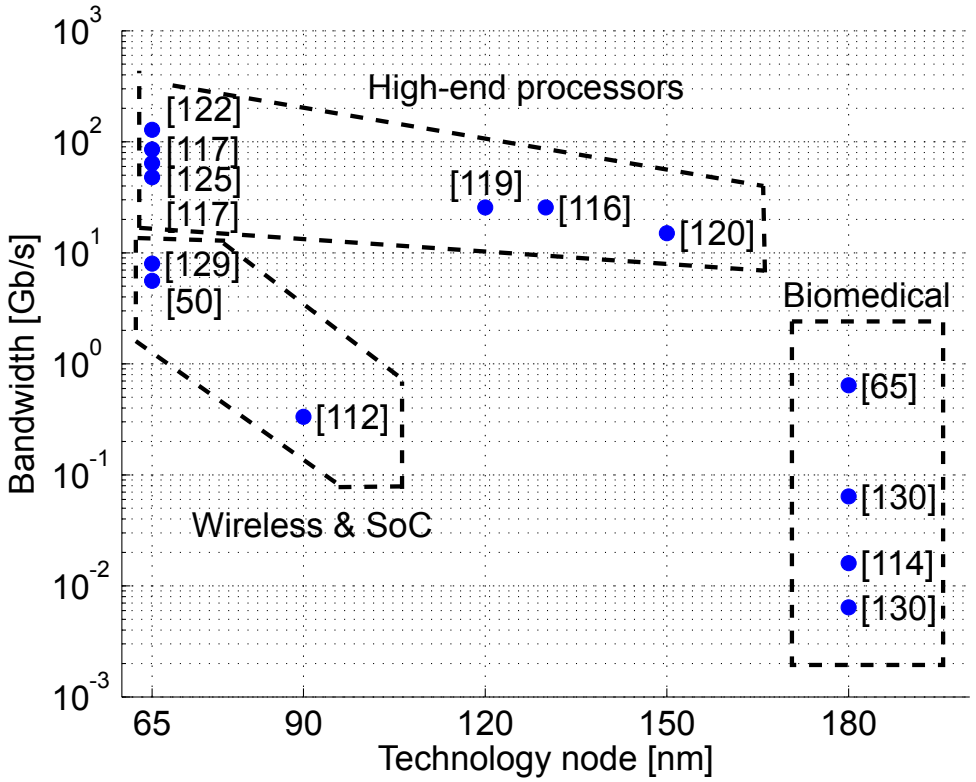


Figure 4.1: Bandwidth vs. technology node of several published GC-eDRAM implementations.

between 10 Mb/s and several 100 Mb/s at near-V_T or sub-V_T supply voltages.

Fig. 4.2 plots the retention power (i.e., the sum of refresh power and leakage power) of previously reported GC memories versus their retention time. For energy-constrained biomedical systems, long retention times of 1–10 ms are a key design goal in order to achieve low retention power between 600 fW/bit and 10 pW/bit. The memory banks of the LDPC decoder have a nominal retention time of 1.6 µs [50], which is around four orders-of-magnitude lower than that of the arrays targeted at biomedical systems. Even though the reported power consumption of 5 µW/bit corresponds to active power [50], it is fair to compare it to the retention power of other implementations, as data would anyway need to be refreshed at the same rate as new data is written. Interestingly, the power consumption per bit of this refresh-free eDRAM is almost seven orders-of-magnitude higher than the retention power per bit of the most efficient eDRAM implementation for biomedical systems. The retention time and retention power of GC memories for processors are in between the values for the wireless and biomedical application domains. Overall, of course, it is clearly visible that enhancing the retention time is an efficient way to lower the retention power.

The area cost per bit (ACPB) is defined as the silicon area of the entire memory macro (including peripheral circuits), divided by the storage capacity. As opposed to the simple bitcell size

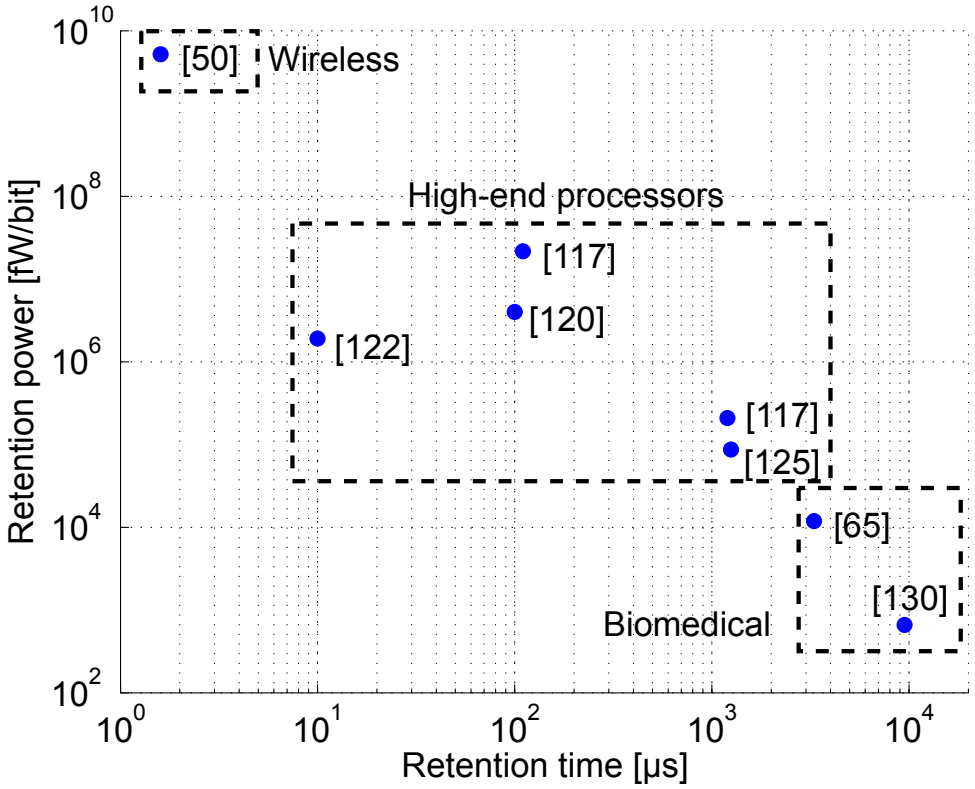


Figure 4.2: Retention power vs. retention time for several published GC-eDRAM implementations.

metric, ACPB accounts for the area overhead of peripheral circuits and is a more suitable metric to compare different memory implementations. Moreover, we define the array efficiency as the bitcell size divided by the ACPB; note that the array efficiency is a technology-independent metric. Fig. 4.3 shows the comparably higher ACPB of biomedical GC memories due to the use of a mature 180 nm CMOS node. However, despite their small storage capacity requirements, these implementations achieve a high array efficiency of over 0.5, by using small yet slow peripheral circuits [130]. On the other hand, none of the GC memories targeted toward processors, wireless communications, or SoC applications achieves an array efficiency as high as 0.5, meaning that over half of the area of those macrocells is occupied by peripheral circuits.

Circuit Techniques for Target Applications

GC-eDRAMs have been shown to be an attractive alternative to traditional SRAM arrays for large caches, wireless communication systems, and ultra-low power systems. Hereinafter, we will take a closer look at the circuits used in these GC-eDRAM implementations, and analyze the compatibility of these techniques with their target metrics.

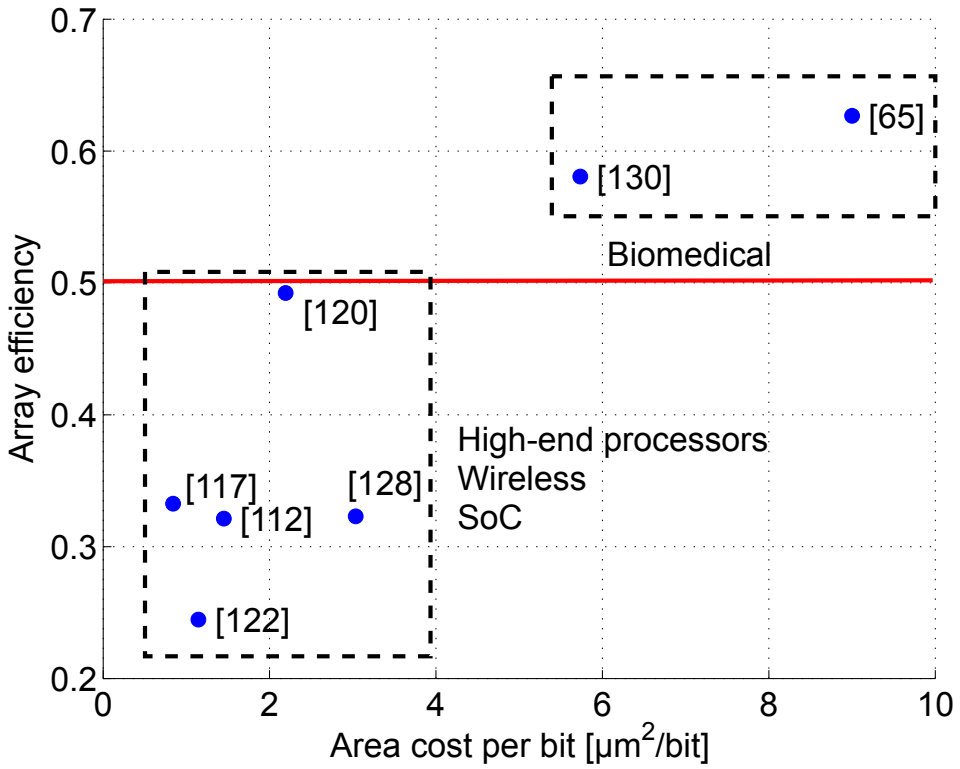
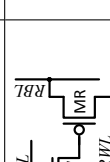
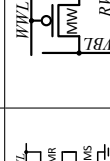
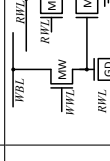
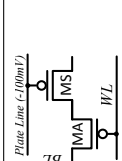
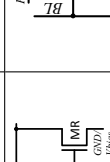
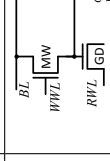
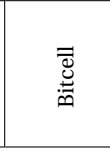
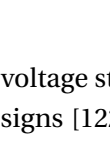
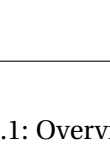


Figure 4.3: Array efficiency vs. area cost per bit (ACPB) for several published GC-eDRAM implementations.

Gain-Cell Topologies An extensive comparison between recent GC topologies is presented in Table 4.1. The common feature for all these circuits is their reduced transistor count, as compared to traditional SRAM circuits. The highest device count appears in [121], comprising three transistors and a "gated diode" (MOS transistor acting as storage device and amplifier), with all other proposals made up of three [119, 116, 129, 124, 125, 128, 63, 112, 50] or two [120, 28, 122, 117, 130, 65, 114, 115] transistors. The obvious implication of the transistor count is the bitcell size; however, the choice of the topology is application dependent, as well. The simple structure of the 2-transistor (2T) topologies usually includes a write transistor (MW) and a combined storage and read transistor (MR). MW connects the write bit-line (WBL) to the storage node (SN) when the write word-line (WWL) is asserted, and MR amplifies the stored charge signal by driving a current through the read bit-line (RBL) when the read word-line (RWL) is asserted. The 2T structure results in coupling effects between the control lines and the SN, which can affect the data integrity and degrade performance. Therefore, a third device is often added, primarily to avoid disturbing couplings from the RWL onto the SN and to reduce RBL leakage. These 3-transistor (3T) gain-cell configurations give up some of the density advantage of gain-cells for the benefit of enhanced speed performance, robustness, and/or retention time. The boosted 3T topology of [125] utilizes the coupling effect to extend the retention time by connecting MR to RWL rather than ground, thereby negating some of the

Table 4.1: Overview of gain-cell circuit techniques according to target applications.

Category	High Performance Processor Caches					
Publication	**[119, 116, 129]**	**[120]**	**[121]**	**[28, 122]**	**[124, 125]**	**[117]**
Bitcell						
Tech. Node	0.12 µm, 0.13 µm, 65 nm PTM	0.15 µm	90 nm	65 nm	65 nm	65 nm
Techniques	Gated Diode, Footer Power Gating, Foot Driver	Multi-Level Bitlines, Hybrid open bitline architecture	Gated Diode Sense Amplifier	RBL Clamping, Pipelined Architecture	Boosted 3T, PVT tracking read reference feedback, Regulated WBL	Half Swing WBL, Stepped WWL
Main Design Metric	400 MHz, 70 µs retention, 100 kb	400 MHz, 100 µs retention, 1 Mb	up to 2 GHz, 110 µs retention, 40 kb	2 GHz, 10 µs retention, 2 Mb	500 MHz, up to 1.25 ms ret., 64 kb	667 MHz, 110 µs ret., 192 kb

Category	General SoC		Wireless	Low Power Biomedical Systems		
Publication	**[128]**	**[63, 112]**	**[50]**	**[130]**	**[65]**	**[114, 115]**
Bitcell						
Tech. Node	90 nm	90 nm	65 nm	0.18 µm	0.18 µm	0.18 µm
Techniques	Forced Feedback, Write Echo Refresh	Multi Level Bitcell, PVT Replica Column	Refresh Free, Sequential Decoding	I/O Write Transistor, Low Area Sense Buffer	Low Area Sense Buffer	Hybrid Cell with I/O MW, Sense Buffer
Main Design Metric	V_{DD} =0.5 V, 180 µA ref. power, 5 MHz	2–50 µs retention, 1.45 µm^2/bit density	32× 1 kb arrays, 700 MHz, 170 ns retention	V_{DD}=0.75 V, up to 306 ms ret., 0.1–1 MHz, 662 fW/bit ret. power	V_{DD}=0.75 V, 3.3 ms retention, 11.9 pW/bit ret. power	V_{DD} =400 mV, over 40 ms ret., 500 kHz

positive SN voltage step inherent to the PMOS MW configurations. Interestingly, large cache memory designs [122, 120, 117] prefer the 2T topology at the cost of additional peripheral

hardware to retain high speed performance. An interesting choice of the 2T topology is used in [114] even though the target application is a small array for ultra-low power (biomedical) systems. In this case, the stacked readout path of the 3T topology proved to be too slow under sub-V_T biases.

Device Choices The majority of today's CMOS process technologies provide several device choices, manipulating the oxide thickness and channel implants to create several threshold voltage (V_T) and maximum voltage tolerance options. Careful choice of the appropriate device (PMOS/NMOS, standard/high/low V_T) can provide orders-of-magnitude improvement in GC performance, as apparent in Table 4.1. PMOS devices suffer from lower drive strength than their NMOS counterparts, but have substantially lower subthreshold conduction and gate leakage. For most of the common process technologies, the primary cause of storage node charge loss is subthreshold conduction through MW, and therefore the ultra-low power implementations [130, 114] employ a high-V_T or I/O PMOS to substantially extend retention time. Gate leakage is a substantial contributor in thin oxide nodes, and so the all-PMOS 2T configuration [122] balances the subthreshold conduction and the gate leakage out of and in to the storage node to improve retention time. The decoder system of [50] requires high performance with very short retention times, and therefore an all NMOS low-V_T circuit is used. Low-V_T devices are used in the readout path of several other publications [117, 128] in order to improve the read speed without increasing the static power, as there is a zero drain-to-source voltage drop across MR during write and standby cycles.

The device choices affect the capacitive couplings to and charge injection onto the SN. WWL access significantly modifies the initial level of the storage node, depending on several factors. A PMOS write transistor passes a weak '0', and an NMOS passes a weak '1'; therefore an underdrive (for PMOS MW) or boosted (for NMOS MW) access voltage of WWL is necessary to pass a full level to the storage node. However, the larger the WWL swing is, the larger the capacitively coupled voltage step on the storage node during WWL deassertion . A PMOS MW is cut-off by the rising edge of WWL, resulting in both capacitive coupling and charge injection to the storage node. Therefore, the initial '0' value will always be significantly higher than ground for a PMOS MW, and the initial '1' value will be significantly lower than V_{DD} for an NMOS device. This limits the storage node range and degrades both the readout overdrive, as well as the retention time. In a 2T gain-cell, using the same device option for MR as for MW induces an additional step in the same direction during read access, further impeding the performance. A hybrid cell, mixing NMOS and PMOS devices [114, 128, 117, 63, 112], can be used to combat these effects, at a small area overhead for two different wells within each bitcell.

Peripheral Circuit Techniques In addition to the choice of a gain-cell topology and device options, several peripheral circuit techniques have been demonstrated to further improve system performance according to the target application. One simple and efficient technique

is the employment of a sense buffer in place of a standard sense amplifier (SA) in low-power systems [130, 114, 65]. This implementation requires a larger RBL swing, trading off speed for area and PVT sensitivity. The area trade-off is apparent in Fig. 4.3 as [130] shows exceptionally high area efficiency. Several other SA configurations have been demonstrated to deal with various design challenges. Chun et al. [117] overcome the problem of small RBL voltage swing by using a current mode SA featuring a cross-coupled PMOS latch and pseudo-PMOS diode pairs. Other SA designs include p-type gated diodes [119, 116, 121], offset compensating amplifiers [120], single-ended thyristors [50], and standard latches [122]. The most complex sensing scheme is used for multilevel gain-cells in [63, 112]: to decipher the four data levels, a successive approximation sensing scheme is used.

Several publications [130, 114, 128, 65] discharge WBL during non-write operations to extend retention time that is worse for a stored '0' than a '1' with a PMOS WM. A "write echo refresh" technique was employed by Ichihashi et al. [128] to further reduce the WBL='1' disturbance. In this technique, the number of '1' write-back operations during refresh are counted and oppositely biased to combat the disturbance. The authors of [125] recognized that the steady state level of a '1' and '0' is common, so they monitor this level and use it as the WBL voltage for writing a '1'. This minimizes the '0' level disturbance without impeding the worst-case '1' level. For the system proposed in [117], WBL switching speed is the performance bottleneck, and therefore a half-swing WBL is employed, improving the write speed and reducing the write power.

An issue that is rarely discussed in 2T bitcell implementations is the voltage saturation of RBL during readout. Depending on the implementation of MR, readout is achieved by either charging (NMOS) or discharging (PMOS) RBL. However, once RBL crosses a threshold (depending on the current ratio of the selected bitcell and the number of off unselected cells), a steady state is reached. This phenomena not only limits the swing available for RBL sensing, but also causes static current dissipation that is present throughout the entire read operation. This is one of the phenomena which should be considered when choosing the appropriate V_{DD} for a low-power GC. Somasekhar et al. [122] combat the self clamping of RBL by explicitly clamping its voltage with designated devices.

Summary and Conclusions

We reviewed and compared recently proposed GC memories, categorizing them according to target applications and overviewing the characteristics that make them appropriate for these applications. A closer look into the circuit design of these arrays provided further insight into the methods used to achieve the required design metrics through the use of different bitcell topologies, device options, technology nodes, and peripheral circuit implementations. To summarize briefly, the following best-practice guidelines should be used when designing GC arrays for future applications:

- High-V_T write access transistors for long retention times and low refresh power, in

conjunction with area-efficient sense buffers for high array efficiency are most suitable to meet the storage requirements of ultra-low power (biomedical) systems.

- High-speed applications should use sensitive sense amplifiers to overcome small voltage differences, and should consider the use of low-V_T readout transistors for improved read access speed.

- Frequently updating systems can trade off high-speed access for limited retention time to achieve improved bandwidth.

4.2 GC-eDRAMs Operated at Scaled Supply Voltages

While almost all previous works on GC-eDRAM considered operation at nominal supply voltage for high speed performance and high memory bandwidth (see Section 4.1), this Section investigates the impact of voltage scaling on the retention time and power consumption of a 2-transistor (2T)-bitcell GC-eDRAM. Targeting near-threshold computing (NTC) [15] systems (see middle column of Table 1.1 in Section 1.2) which are characterized by low power consumption at still relatively high speed performance, we investigate the limit of voltage scaling for GC-eDRAM such that all operations still rely on on-currents of the inherent transistors (avoiding the use of subthreshold conduction for active operations, which is addressed later in Section 4.4). This voltage limit for the main supply which still ensures fast circuit operation is derived for the case of using an underdrive voltage for the write word-line (WWL) and for the case of using a single, main supply for the entire GC-eDRAM macrocell. Interestingly, the retention time can be increased when scaling down the supply voltage for given memory access statistics and a given write bit-line (WBL) control scheme. Moreover, for a given supply voltage, the retention time can be further increased by controlling the WBL to a voltage level between the supply rails during idle and read states (which, however, has a considerable overhead for voltage generation). These two concepts are proved by means of Spectre simulation of a GC-eDRAM macrocell implemented in 180 nm CMOS technology and operated at only 40 % of the nominal supply voltage. In order to maintain high memory bandwidth even for reduced operating frequencies at scaled voltages, we show that a 2T-bitcell GC-eDRAM macrocell can easily be implemented as a two-port memory at a negligible area overhead compared to a single-port memory implementation.

Section 3.1 has reviewed specially designed SRAM macrocells operating reliably at scaled supply voltages at the price of relatively large 8-transistor (8T) [90], 10T [6], or even 14T [89] bitcells. The entire Chapter 3 was dedicated to synthesized latch arrays and flip-flop arrays which are a more straightforward approach to reliable low-voltage storage arrays than SRAMs but have an even larger area cost for storage capacities higher than a few kb [33]. In conventional 1-transistor-1-capacitor (1T-1C) embedded DRAM (eDRAM), the offset voltage of the sense amplifier limits voltage downscaling, unless dedicated offset cancellation techniques are used [131]. Another major obstacle in low-voltage 1T-1C eDRAM is the degradation of the data retention time, which requires power-consuming refresh operations more frequently [131].

Furthermore, as expatiated on in Section 4.1, conventional 1T-1C eDRAMs require special process options to build high-density 3D capacitors, which adds cost to standard digital CMOS technologies. As a further attractive option for building embedded storage arrays operated at scaled voltages, gain-cells are smaller than any SRAM bitcell, latches, and flip-flops, while they are fully compatible with standard digital CMOS technologies. While most previous works promote GC-eDRAM as denser successor of SRAM for on-die caches in high-end processors [125, 28] (see Section 4.1), only a small number of works investigate GC-eDRAM operation at scaled voltages: 1) a dual threshold-voltage (dual-V_T) GC storage array [130] is operated at a fraction of the nominal supply voltage; the circuit increases the retention time by using a high threshold-voltage (high-V_T) write access transistor (WT); and 2) another storage macro based on a boosted 3-transistor (3T) GC [132] is operable in a supply voltage range from 1.2 down to 0.7 V and uses preferential storage node boosting at the time of reading to increase the retention time (and the read speed).

Previously reported GC-eDRAM macrocells are not clearly classified as either single-port or two-port implementations. Furthermore, while previous work on GC storage arrays targets a given supply voltage (or supply voltage range) and presents dedicated techniques to increase the retention time, the impact of supply voltage scaling on the retention time has not been systematically investigated yet. Moreover, previous publications do not clearly state the assumed write access statistics for the measurement of the retention time, while frequent write accesses may in fact significantly degrade the retention time.

Therefore, the remainder of this Section reviews why GCs are inherently suitable for two-port memory implementations with a negligible area-overhead compared to single-port implementations. The limit to supply voltage scaling in 2T-bitcell GC-eDRAM in the occurrence of process parameter variation is then discussed, avoiding relying on subthreshold conduction to achieve medium speed performance in NTC systems. Next, the impact of supply voltage downscaling on the retention time under well-defined memory access statistics is investigated, allowing for finding the optimum supply voltage for lowest power consumption and highest retention time. Finally, a simple technique to further improve the retention time at any given supply voltage is presented.

4.2.1 2T Low-Voltage GC-eDRAM Array Architecture

Two-Port Implementation

Concurrent read/write access is an effective method for achieving high memory bandwidth [118]. Two-port memories have a separate read and write port to enable such access. In conventional 1T-1C DRAM and conventional SRAM, the same word-lines (WLs) and bit-lines (BLs) are used for both the read and the write operation; enabling two-port operation is non-trivial and requires additional hardware in each cell. As opposed to this, gain-cells (GCs) are inherently well suited for two-port operation, as they already have a separate read port consisting of the read word-line (RWL) terminal and the read bit-line (RBL) terminal as well as a separate write

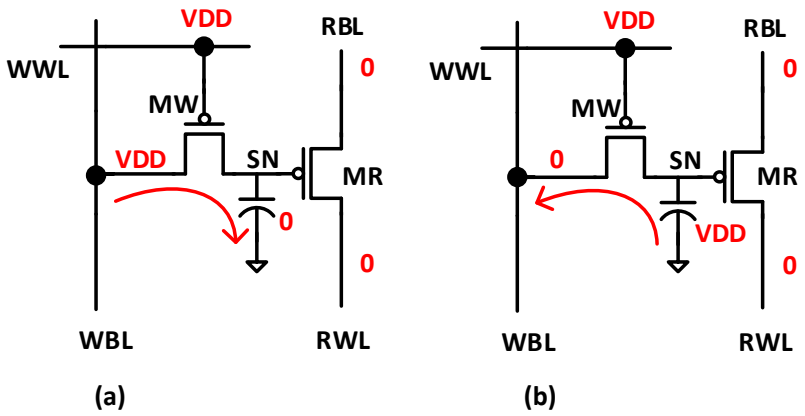


Figure 4.4: 2-PMOS gain-cell; worst write bit-line (WBL) state for retention of (a) logic '0' and (b) logic '1'.

port consisting of the write word-line (WWL) terminal and the write bit-line (WBL) terminal, as shown in Fig. 4.4. It is therefore straightforward to enable two-port operation in GC-based storage arrays and benefit from the resulting high memory bandwidth.

In the two-port memory architecture adopted in this work, there are two address decoders: one for the write address, and another one for the read address. A single-port implementation would save one address decoder, but it would require additional logic circuits—comparable in size to a single decoder—to distribute the decoded address to either the write port or the read port, while silencing the other port.

Array and Gain-Cell Implementation

Apart from the explicit two-port configuration, the memory architecture serving as a basis for the presented analyses is mostly adopted from [130]. As shown in Fig. 4.5, the storage array consists of 32 rows and 64 columns. Moreover, the conventional sense amplifiers are replaced with simple sense inverters to improve area-efficiency [130]. To allow for conclusions as general as possible, the basic 2-PMOS GC with regular threshold-voltage (regular-V_T) transistors from [28] is adopted in this work, as the high-V_T transistors used in [130] might not be available in all technologies. Notice, however, that high-V_T transistors may reduce subthreshold conduction by more than 2 orders of magnitude compared to regular-V_T transistors [130], and therefore allow for considerably longer retention times (as will be seen in Section 4.3 and Section 4.4).

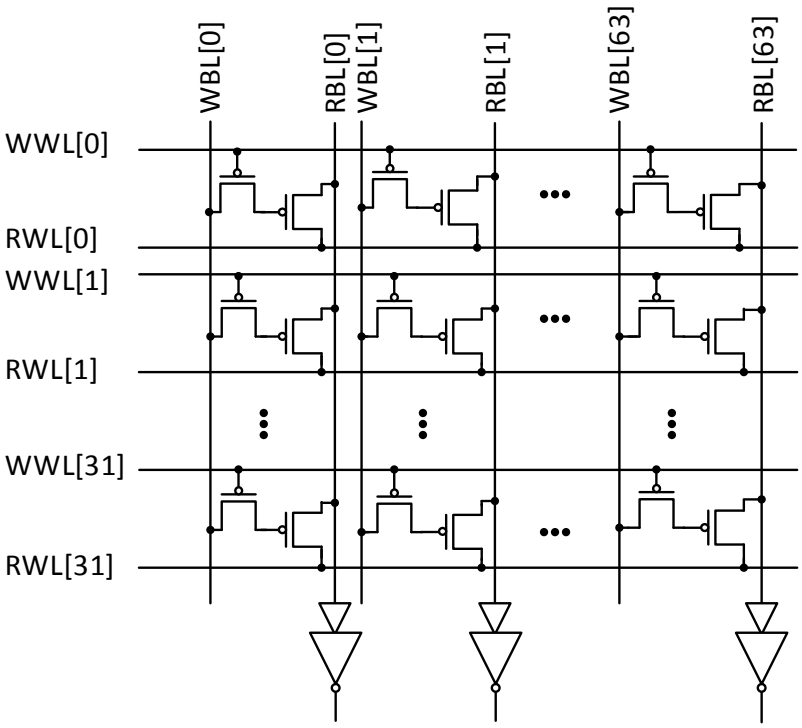


Figure 4.5: 2T-bitcell GC-eDRAM storage array with area-efficient sense inverters.

4.2.2 Operation Principle

Hold, Write, and Read Operations

In each 2-transistor (2T) gain-cell, data is stored in form of charge on the storage node (SN) capacitor, which is formed by the gate capacitance of the storage/read transistor (MR) and junction/wire parasitic capacitance. The parasitic SN capacitor is explicitly shown in Fig. 4.4. During a write operation, the write transistor (MW) of the selected GC is turned on to transfer the new data level from the WBL to the SN. To allow the transfer of a clean logic '0', an underdrive voltage of −500 mV is applied to the selected WWL. At the beginning of a read operation, all RBLs are discharged to ground. Next, the selected RWL is pulled high to V_{DD}. If a GC stores a logic '1', its MR remains off and the connected RBL remains at ground. However, if the GC stores a logic '0', the RBL starts to charge through MR. The sense inverter must switch before RBL is charged to the threshold voltage of MR (V_T^{MR}), as at this time read transistors MR in unselected cells storing logic '0' turn on, which provides a current path to ground and prevents a further voltage rise on the RBL.

Limit to Supply Voltage Scaling for Fast Access

The minimum supply voltage for reasonably fast memory access is determined by the ability of writing, holding, and reading two distinct data levels while not relying on subthreshold conduction for active circuit operation. Considering the 2-PMOS GC and avoiding any underdrive voltage, MW can easily transfer a high voltage level equal to V_{DD} to the SN. However, the lowest data level which can be transferred in a reasonable time, i.e., not relying on subthreshold conduction, is equal to the threshold-voltage of MW (V_T^{MW}). When turning off MW, charge injection and clock feedthrough rise the voltage on the SN (V_{SN}) by ΔV_{SN}, which depends on the SN capacitance, the voltage level being transferred, and many other factors. After writing a logic '0' level, $V_{SN} = V_T^{MW} + \Delta V_{SN}$. Holding a data level on the SN during a small amount of time is possible regardless of V_{DD}. To tell a logic '0' from a logic '1' at the time of reading, V_{SN} must be smaller than $V_{DD} - V_T^{MR}$ in order to still be able to turn on the RT:

$$V_T^{MW} + \Delta V_{SN} < V_{DD} - V_T^{MR} \tag{4.1}$$

Equation (4.1) is rearranged to show the lower limit for V_{DD}:

$$V_T^{MW} + V_T^{MR} + \Delta V_{SN} < V_{DD} \tag{4.2}$$

To account for process parameter variations (die-to-die and within-die variations), Equation (4.2) is rewritten as follows, where $\mu(X)$ and $\sigma(X)$ denote the mean and the standard deviation of the random variable X.

$$\left(\mu(V_T^{MW}) + N\sigma(V_T^{MW})\right) + \left(\mu(V_T^{MR}) + N\sigma(V_T^{MR})\right) + \Delta V_{SN} < V_{DD} \tag{4.3}$$

The parameter N is chosen depending on the desired yield. For small storage arrays of several kb, $N = 3$ is reasonable.

Assuming a WWL underdrive, a clean ground level can be transferred to the SN, and V_{DD} can be further reduced, with its lower limit now given by:

$$\left(\mu(V_T^{MR}) + N\sigma(V_T^{MR})\right) + \Delta V_{SN} < V_{DD} \tag{4.4}$$

It is usually beneficial in terms of energy to have a WWL underdrive, as most parts of the circuit can be operated from a lower V_{DD}, while the underdrive voltage is only applied to the write address decoder and the WWL drivers.

In the current case, using an underdrive voltage of −500 mV, and with $\mu(V_T^{MR}) = 500\,\text{mV}$, $\sigma(V_T^{MR}) = 25\,\text{mV}$, $N = 3$, $\Delta V_{SN} \approx 100\,\text{mV}$ (extracted from circuit simulations), and a small margin for uncertainty in ΔV_{SN}, the lowest V_{DD} for reliable operation and reasonable yield is 700 mV, which is only 40 % of nominal V_{DD} (1.8 V).

4.2.3 Impact of Supply Voltage Scaling on Retention Time

Low-voltage low-to-medium speed VLSI systems (such as microprocessors) are best implemented in older, low-leakage CMOS technology nodes (such as 180 nm) to minimize energy dissipation, especially if leakage-reduction techniques such as power gating switches are applied [133]. The considered GC storage array is therefore implemented in a commercial 180 nm CMOS technology. Among many leakage mechanisms, the subthreshold conduction of MW is clearly the dominant mechanism corrupting the stored data. This subthreshold conduction and consequently the data retention time strongly depend on the voltage level encountered on the WBL, denoted by V_{WBL}.

Assuming that a GC has just been written to and is now holding its data, there are two possible scenarios:

1. Further write operations are performed to GCs on the same WBL, meaning that V_{WBL} is data-dependent and cannot be controlled.

2. The memory remains in idle state (no data accesses) or only read accesses are performed. During idle and read states, V_{WBL} can be controlled to any desired voltage to minimize subthreshold conduction of MW.

Fig. 4.4 shows the *worst-case access* scenario in terms of retention time where the opposite data level is permanently written to GCs on the same WBL after writing a given data level to the first GC. The *retention mode* scenario presumes an application where a relatively small storage array (with only few GCs per WBL) is fully written in a negligibly short time, whereafter the memory is kept in idle or read states and the WBL can be controlled to either V_{DD} or ground. Very short write access times, compared to the read access time, may be achieved in two-port memories. Under the retention mode scenario, the potential of controlling the WBL to a voltage level between the supply rails will be evaluated, as well.

Worst-Case Access

Assuming the worst-case access scenario where V_{WBL} is permanently opposite to the stored data level, the retention time for a logic '0' ('1'), denoted by t_{ret0} (t_{ret1}), is defined as the time it takes for V_{SN} to rise (fall) to $V_{\mathrm{DD}} - V_{\mathrm{T}}^{\mathrm{MR}}$. At nominal V_{DD}, t_{ret1} is longer than t_{ret0}: the more the logic '1' voltage level decays, the more positive the gate-to-source voltage V_{GS} and the higher the reverse body biasing (RBB) of MW, both suppressing the subthreshold conduction harder [28].

As shown in Fig. 4.6, when V_{DD} is gradually scaled down, the storage range for a logic '0', given by $V_{\mathrm{DD}} - V_{\mathrm{T}}^{\mathrm{MR}}$ (if neglecting charge sharing and clock feedthrough for simplicity), becomes smaller, while the storage range for a logic '1', given by $V_{\mathrm{T}}^{\mathrm{MR}}$, remains unchanged. At the same

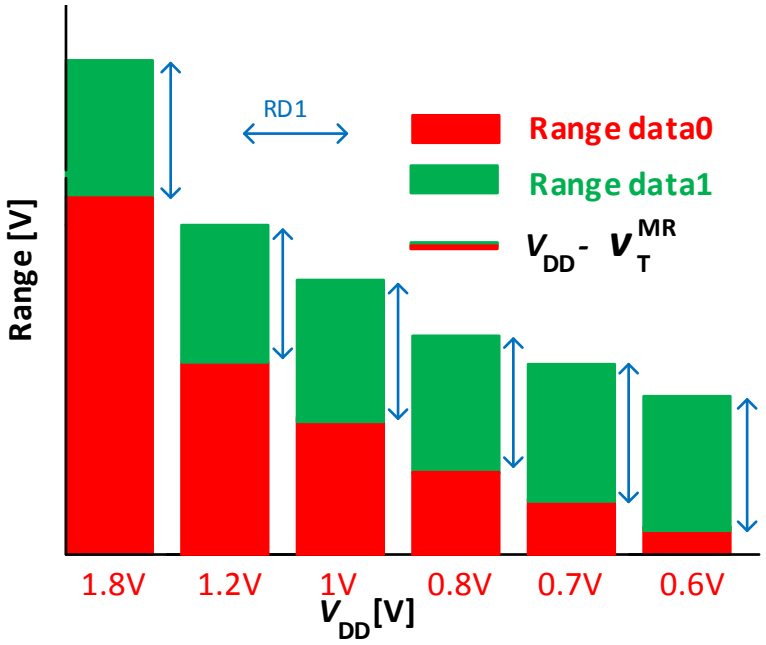


Figure 4.6: Storage ranges (voltage ranges) for data '0' and '1' versus main supply voltage V_{DD}.

time, when V_{DD} is scaled down, the subthreshold conduction of MW becomes smaller due to its exponential dependence on V_{GS} and the drain-to-source voltage V_{DS}.

As a consequence, t_{ret1} increases with decreasing V_{DD}, as shown by the Spectre simulation results in Fig. 4.7. However, Fig. 4.7 also shows that t_{ret0} decreases with decreasing V_{DD}, as the always smaller storage range has the higher impact than the decreasing strength of the subthreshold conduction.

Retention Mode

WBL Control to Ground If the access scenario is now changed, assuming only idle and read states after initially writing the entire storage array, V_{WBL} can be controlled to ground, in order to avoid the decay of a logic '0'. In this case, the data retention time of the storage array is given by t_{ret1}. When scaling V_{DD} from its nominal value of 1.8 V down to 700 mV, the data retention time increases by 4× (see Fig. 4.7). At the same time, the power consumption is considerably reduced, due to 1) lower V_{DD}, and 2) fewer required refresh cycles. Briefly, if the GC-eDRAM is kept in idle/hold or read states after an initial write access, supply voltage scaling improves both retention time and energy-efficiency.

WBL Control for Enhanced Retention Time Still presuming the retention mode scenario, but now considering that V_{WBL} can be controlled to any desired voltage level between the

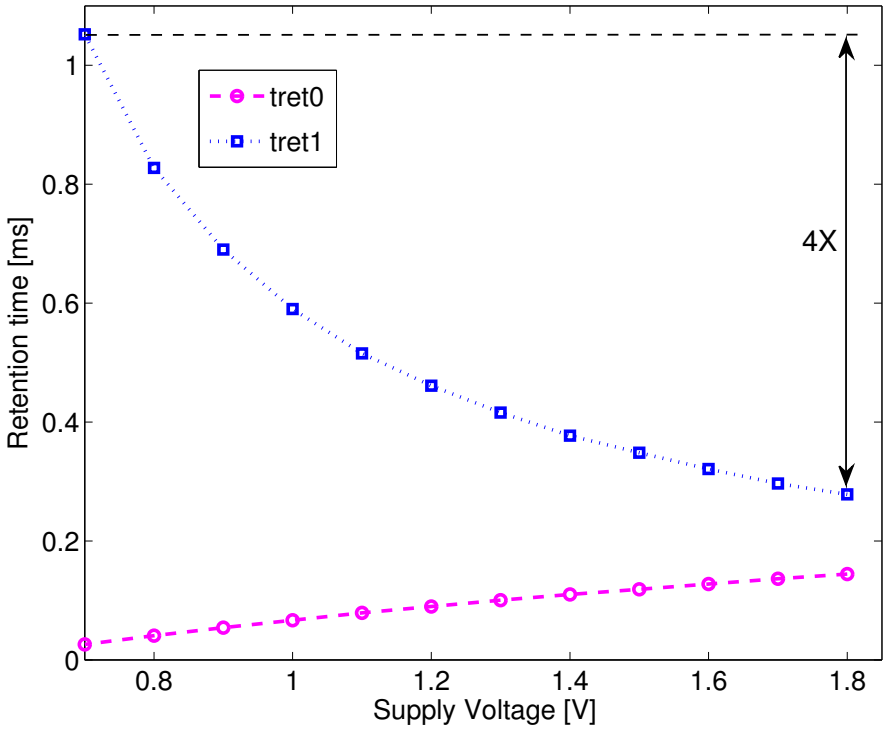


Figure 4.7: Retention time versus V_{DD} for worst-case WBL state (always opposite to stored data).

supply rails[1] to reduce subthreshold conduction, the retention time for any V_{DD} can be further increased compared to the previously mentioned WBL-discharge control.

Fig. 4.8 shows t_{ret1} and t_{ret0} as a function of V_{WBL}, for different values of V_{DD}. Clearly, t_{ret0} increases with decreasing V_{WBL} for any considered V_{DD}, due to a constant storage range and decreasing strength of the subthreshold conduction. For the same reasons, t_{ret1} increases with increasing V_{WBL}. The highest retention times are reached when V_{WBL} approaches $V_{DD} - V_T^{MR}$, and t_{ret1} (t_{ret0}) becomes infinitely long for V_{WBL} higher (lower) than $V_{DD} - V_T^{MR}$. However, the slopes in this region are very steep, so that any noise on V_{WBL} considerably degrades the retention time. At $V_{DD} = 700\,mV$, choosing $V_{WBL} = 200\,mV$, a retention time of 3.3 ms is achieved, corresponding to a 3.3× improvement compared the case where V_{WBL} is controlled to ground.

[1] Of course, controlling V_{WBL} to a voltage level between the main supply rails requires additional circuits (DC-DC voltage converters) whose use can only be justified for large GC-eDRAM storage arrays.

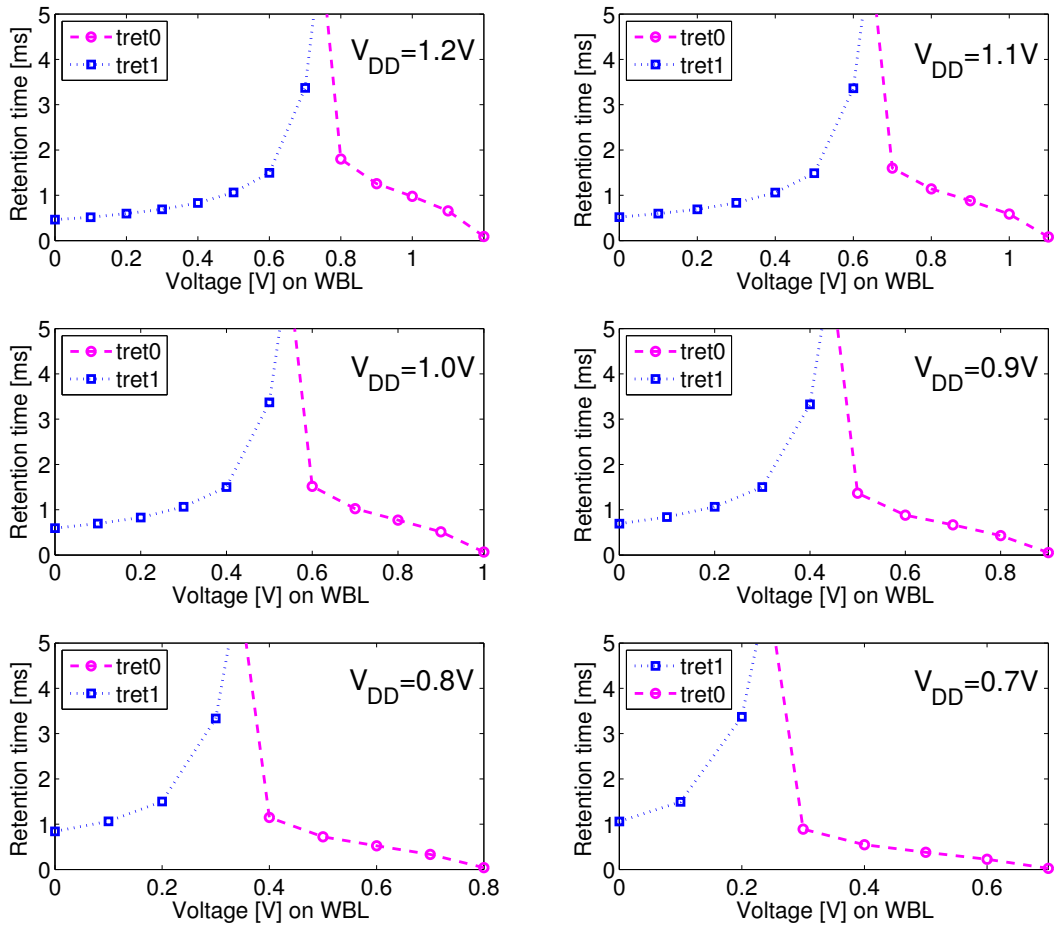


Figure 4.8: WBL control for enhanced retention time.

4.2.4 Macrocell Implementation Results

In the retention mode, an overall improvement of 13.2× in retention time and a considerable reduction in power consumption are obtained by supply voltage scaling and the controlled WBL technique. The active refresh power of the presented 2 kb macro is 10.8 pW/bit, while the leakage power is 1.1 pW/bit, amounting to a total data retention power of 11.9 pW/bit.

Table 4.2 compares this work to a selection of GC storage arrays in literature [28, 130, 124]. All retention time and retention power values are given for a temperature of 25 °C, unless otherwise stated.

For the same technology node (180 nm), Table 4.2 shows the effectiveness of a high-V_T write transistor (MW) [130] (if available and economic) to improve the retention time by around 100×. For smaller technology nodes (65 nm), [124] manages to keep a good retention time using a low-leakage process (and circuit-level techniques); however, in a native 65 nm logic process [28] (design optimized for high bandwidth), the retention time is degraded by around 100×.

Table 4.2: Comparison of low-voltage GC-eDRAM storage arrays.

Publication	[28]	[130]	[124]	This [65]
Technology node [nm]	65	180	65	180
V_{DD} [V]	1.1	0.75	0.9	0.7
Retention Time [ms]	0.01	306[a]	1.25[b]	3.3
Retention Power [pW/bit]	-	0.662	87.1 (85 °C)	11.9

[a] High-V_T transistor reduces leakage by more than 2 orders of magnitude [130]
[b] Low-leakage CMOS technology

In the presented study relying on a commercial 180 nm CMOS technology, the active refresh power is clearly dominant compared to the leakage power, meaning that any effort to increase the retention time also significantly reduces the total data retention power (see Table 4.2). Therefore, the focus of the following Section 4.3 will be on novel techniques to extend the retention time. Reference [124] reports higher refresh power in 65 nm CMOS, but also uses a slightly higher supply voltage and measures at a temperature of 85 °C.

4.2.5 Conclusions

Gain-cell storage arrays are an interesting alternative to SRAM macros in low-power/low-voltage (near-V_T) VLSI SoCs and microprocessors. Gain-cells are inherently suitable for building two-port memories (as opposed to SRAM and conventional eDRAM). 2-PMOS gain-cell storage arrays can be reliably operated at low supply voltages close to the threshold voltage if a few critical circuit nodes (namely the WWLs) receive an underdrive voltage.

The data retention time improves by 4× when scaling down the supply voltage from 1.8 to 0.7 V, provided that write access is unfrequent and short. In addition to this, another 3.3× improvement in retention time is achieved by controlling the voltage on the WBL to a value between the supply rails during idle and read states. This overall improvement in retention time of 13.2× combined with operation at less than 40 % of the nominal V_{DD} leads to a data retention power of 11.9 pW/bit. The data retention power was found to be dominated by active refresh power, while leakage power plays only a minor role. Therefore, the next Section presents several techniques to enhance the retention time of near-V_T GC-eDRAM arrays for reduced data retention power.

4.3 Near-V_T GC-eDRAM Implementations with Extended Retention Times

As explained in the previous Section, supply voltage scaling to the near-threshold domain is beneficial to improve the retention time of GC-eDRAMs, provided that write access occurs

only seldom and that the write bit-lines (WBLs) can therefore be controlled to a desired voltage level during most of the time. In this Section, two techniques to further enhance the retention time of near-V_T GC-eDRAMs are presented: 1) reverse body biasing (RBB) in order to suppress the subthreshold conduction of the write transistor MW (see Section 4.3.1 below); and 2) replica gain-cells to track the data integrity of the actual gain-cell array across process-voltage-temperature (PVT) corners and across varying write access statistics (accounting for write disturbs, see Section 4.3.3).

4.3.1 Impact of Body Biasing (BB) on the Retention Time

Reverse body biasing (RBB) is a well-known technique to suppress leakage current and is extensively and industrially used in conventional 1T-1C DRAM technology. In fact, in most DRAM chips, the p-well is biased to a negative voltage to improve the data retention time, a technique also referred to as back bias control. However, there are no previous studies on applying RBB to fully logic-compatible GC-eDRAM in order to improve its retention time and reduce its data retention power. In the following, we measure the impact of body biasing as a control factor to improve the retention time of a 2 kb GC-eDRAM macrocell, and also examine the distribution of the retention time across the entire gain-cell array. The concept is demonstrated through silicon measurements of a test chip manufactured in a logic-compatible 0.18 µm CMOS process. While there is a large retention time spread across the measured 2 kb gain-cell array, the minimum, average, and maximum retention times are all improved by up to 2 orders of magnitude when sweeping the body voltage over a range of 375 mV.

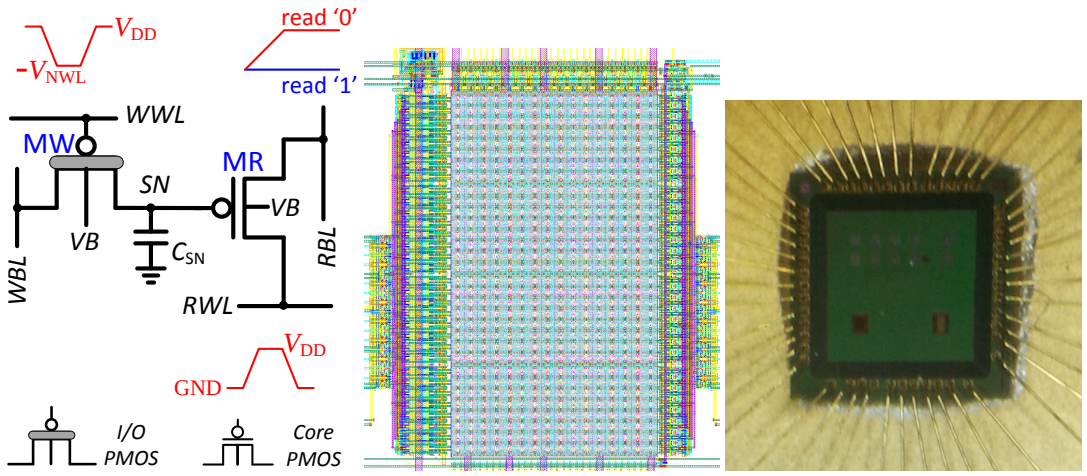
As already mentioned in Section 4.1, the main drawback of GC-eDRAMs is the need for periodic refresh cycles, which results in a considerable amount of power consumption and limits the read/write availability of the memory array. Therefore, to improve the competitiveness of gain-cell eDRAM, it is crucial to extend the data retention time. Data levels in GC-eDRAMs are stored as charge on the capacitive storage node (SN), whose equivalent capacitance is referred to as C_{SN}, and therefore data retention is limited by the time it takes for this charge to leak away. Several simple measures can be taken to extend the retention time, such as: 1) increasing C_{SN} through layout techniques (increasing the write transistor's diffusion area and the storage transistor's gate area, as well as employing the metal stack and vias readily available in digital CMOS technologies to gain additional in-cell capacitance [114, 115]); 2) minimizing the subthreshold conduction through the write access transistor (MW) by using low-leakage MOS transistors [130]; and 3) employing write bit-line (WBL) control schemes to minimize charge loss through MW (see previous Section 4.2 and [65]). An additional technique that has not yet been applied to gain-cells is threshold voltage (V_T) adjustment through body biasing. While the application of a reverse body bias (RBB) raises V_T and therefore reduces the charge loss through subthreshold conduction, this means of control can also improve the array availability by applying a forward body bias (FBB) during refresh cycles to reduce access time [134]. Our main contributions can be summarized as follows: 1) For the first time, we propose reverse body biasing as a technique to improve the retention time of GC-eDRAM and

demonstrate its high effectiveness by means of silicon measurements; and 2) moreover, the retention time penalty of forward body biasing, used for fast memory access and short refresh times, is evaluated.

Bitcell Design

Fig. 4.9a shows the schematic and the basic operation of the two-transistor (2T) all-PMOS gain-cell used in this study (a similar cell has previously been proposed in [130]). Other than the high-V_T I/O PMOS write transistor (MW) requiring a larger underdrive voltage, the cell operation is equal as for the gain-cell considered in Section 4.2 and is therefore recalled only briefly. MW is used to transfer the data driven onto the WBL to C_{SN}. MR is the read access transistor, used to read out the data level stored in the bitcell. A write access is initiated by applying an underdrive voltage ($-V_{NWL}$) to the write word-line (WWL) in order to properly transfer a logic '0' level (V_{SS}) from WBL to SN in a short time. A read access is initiated by pre-discharging the read bit-line (RBL) and subsequently raising the read word-line (RWL). If a logic '0' is stored on C_{SN}, MR will charge RBL past a detectable threshold, whereas if a logic '1' (V_{DD}) has been written to the SN, RBL will remain discharged. The basic C_{SN} is increased by building up side-wall capacitors between the SN and a constant potential (V_{DD}) atop the bitcell footprint, using all 6 available metal layers in the considered 0.18 µm CMOS process.

The dominant leakage mechanism that causes the deterioration of the stored data levels is clearly the subthreshold conduction of MW. This is especially true for mature CMOS nodes, such as the 0.18 µm process used in this study, but also holds for a deeply scaled 40 nm CMOS node [115] (as will be seen in Section 4.4 focusing on aggressive voltage and technology scaling). In order to achieve the longest possible retention time, an I/O PMOS transistor is used to implement MW, as this device features the lowest subthreshold conduction among all devices offered in the chosen 0.18 µm CMOS technology [114]. By implementing MR with a PMOS device, as well, the entire array resides in an equi-potential n-well, enabling simple control over the body voltage (V_B) of the bitcells. Reverse biasing the n-well at a voltage above V_{DD} increases the V_T of the transistors, thereby suppressing the subthreshold conduction of MW and improving the retention time. Likewise, forward biasing V_B below V_{DD} lowers the V_T of the transistors, resulting in faster read and write access times. The variable ΔV_B is used to express the amount of body biasing, according to $V_B = V_{DD} + \Delta V_B$, where a positive and a negative value of ΔV_B correspond to RBB and FBB, respectively. In this study, a biasing range of $-250\,\text{mV} < \Delta V_B < 125\,\text{mV}$ is considered, corresponding to a V_T range of $-770\,\text{mV} < V_T < -625\,\text{mV}$ for a PMOS I/O device under otherwise nominal conditions with $V_{DD} = 750\,\text{mV}$ (corresponding to a near-V_T supply voltage for core transistors).



(a) (b) (c)

Figure 4.9: (a) 2T gain-cell design and basic operation, (b) layout of 2 kb GC-eDRAM macrocell, and (c) microphotograph of test chip.

4.3.2 GC-eDRAM with BB: Silicon Measurements

Macrocell Architecture and Test Chip Design

Fig. 4.9b shows the layout of the 2 kb gain-cell macro memory in 0.18 µm CMOS technology. The core bitcell array consists of 64×32 gain-cells, all sharing the same n-well and V_B. N-well contacts are provided every 16 rows, as well as at the top and the bottom of the array. In addition to the bitcell array, the macrocell comprises the following peripheral circuits: 1) A write address pre- and post-decoder that drives V_{DD} or $-V_{NWL}$ onto WWL; a read address pre- and post-decoder that drives V_{DD} or V_{SS} onto RWL; level-shifters; WBL drivers; readout sense buffers; and timing control units.

Fig. 4.9c shows a microphotograph of the manufactured test chip. In addition to the 2 kb GC-eDRAM macrocell in the lower-left corner, the chip contains a built-in self test (BIST) unit. The main features of the BIST can be summarized as follows: 1) address sequence generation (increasing, decreasing, pseudo-random); 2) data pattern generation (checkerboard, pseudo-random, all-'1', all-'0'); 3) programmable refresh period of the memory under test (MUT); 4) pass/fail decision during readout of the MUT; 5) embedded SRAM for storing maps of MUT retention time, read failures, or write failures; and 6) support for two-port operation of the MUT. Finally, the test chip also contains scan chains for full access to the MUT with any data or address sequence pattern independent of the BIST.

Silicon Measurement Results

The packaged test chips were mounted on a test board by means of a burn-in socket and connected to a TMPC PG3A pattern generator and a Tektronix TLA6403 logic analyzer. The

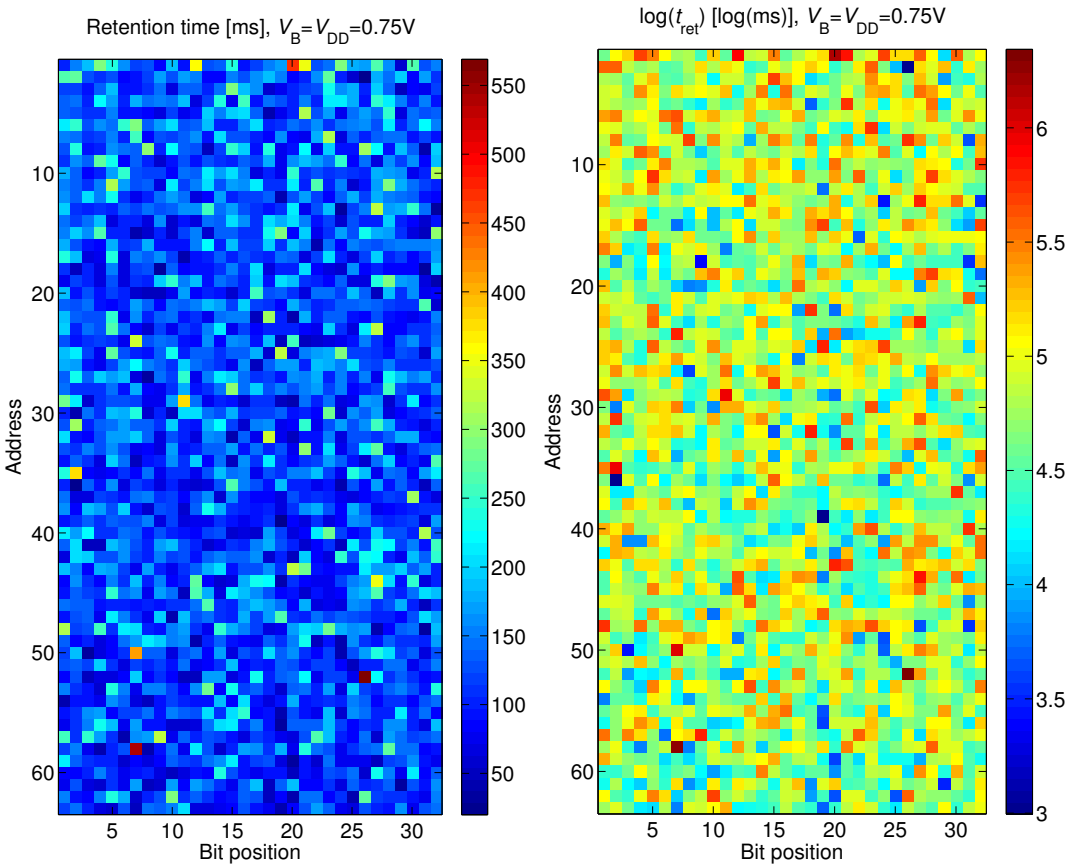
Table 4.3: Measurement setup for GC-eDRAM test chip with adaptive body bias control.

V_{DD}	750 mV
ΔV_B	−250 to 125 mV
Write access time	1 µs
Read access time	1 µs
Write-'1' disturb activity	25%
Temperature	Room temperature (uncontrolled)

main supply of the memory macrocell was set to 750 mV, and the body voltage V_B was swept from 500 to 875 mV to analyze the impact of body biasing. A separate negative voltage of −1.5 V was supplied to the macrocell for the WWL underdrive. The BIST and other digital control units were supplied with the technology's nominal voltage of 1.8 V. Both the write and read access times were set to 1 µs for robust write and read operations, even at the low V_{DD} of 750 mV. This ensured that the measured failures relate to retention time, and were not caused by incomplete writes or erroneous reads due to insufficient access time. Table 4.3 summarizes the primary specifications of the measurement setup.

Measurements indicate that the 2-PMOS gain-cell retains logic '1' levels for extensive periods (>1 s), even when the WBL is held at 0 V (which maximizes the subthreshold conduction of MW). This coincides with previous reports that logic '1' levels decay very slowly due to the increasing reverse gate overdrive and body effect of MW as the SN voltage drops [114]. Therefore, the gain-cell's retention time is almost exclusively limited by its ability to hold a logic '0' level. The decay of a cell's logic '0' level is heavily dependent on the state of the WBL. On the one hand, when WBL is low, subthreshold conduction through MW discharges the SN, reinforcing a stored logic '0' level. On the other hand, when WBL is high, a worst-case condition occurs, as leakage through MW causes accelerated decay of a stored logic '0' level. Our measurement setup assumes a 50% write duty cycle (i.e., there is a write access during 50% of the time) and that the probability of writing a '1' (which requires pulling WBL up to V_{DD}) is 50% as well. Overall, this leads to a write-'1' disturb activity factor ($\alpha_{disturb}$) of 25%.

Using the measurement setup described above, retention time was measured for the entire 2 kb array under standard biasing conditions (i.e., $V_B = V_{DD} = 750$ mV) at room temperature (temperature was not controlled). The results of this measurement are shown in Fig. 4.10a. The minimum and maximum retention times (t_{ret}) of 2048 measured gain-cells were found to be 23 and 569 ms, respectively, corresponding to a ratio of 25 between the maximum and minimum value. A recent study [64] reports an even higher ratio of over 50 between the maximum and minimum measured retention times in an 1 kb array implemented in 65 nm CMOS. In the present study, the majority of the cells exhibited retention times in the range of 20 to 200 ms (dark and light blue color), whereas a small number of cells exhibited considerably higher retention times (yellow, orange, and red colors). In order to better visualize the differences among the lower retention times (20–200 ms), Fig. 4.10a plots t_{ret} on a logarithmic scale. There is no systematic pattern, indicating that the retention time variability arises from local



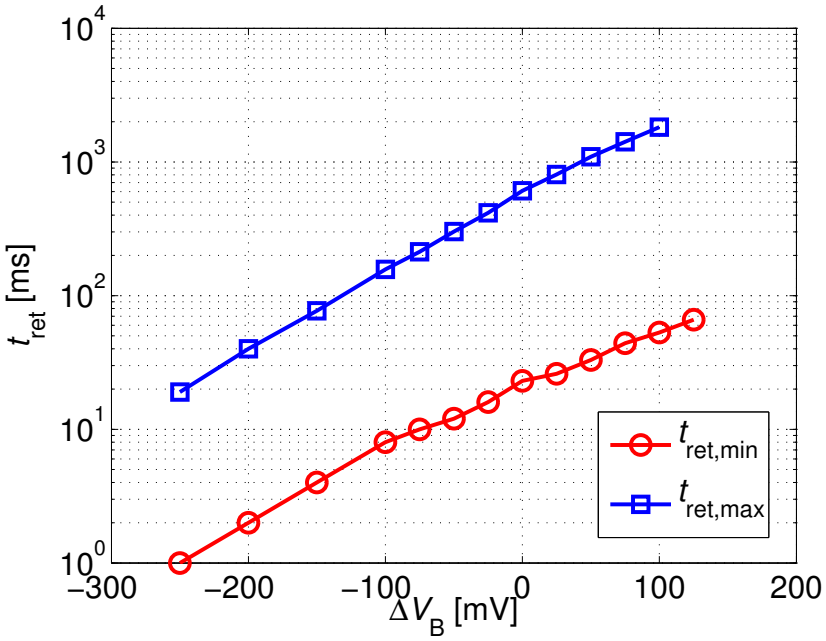
(a) (b)

Figure 4.10: (a) Retention time (t_{ret}) map of 2 kb 2T gain-cell array with standard body bias and α_{disturb}=25% at room temperature, and (b) map of log(t_{ret}).

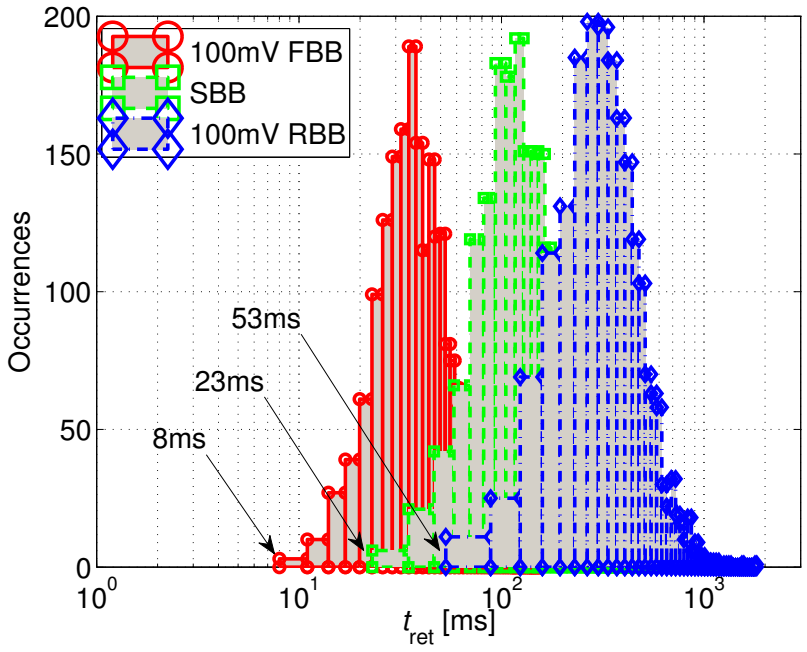
(within-die), random process parameter variations.

The impact of body biasing on the measured retention times was evaluated by sweeping V_{B} from 500 mV to 875 mV (-250 mV< ΔV_{B} <125 mV). The minimum and maximum measured retention times across the entire array are plotted in Fig. 4.11a. This figure clearly shows that the minimum and maximum retention times change by up to 2 orders of magnitude over this 375 mV V_{B} range. As expected, the best cells with the highest retention time remain at the same location under varying V_{B} (not shown in the figure).

Finally, Fig. 4.11b shows the distributions of the retention time across the 2k measured cells, for three biasing conditions: 100 mV FBB, standard body biasing (i.e., $V_{\text{B}} = V_{\text{DD}}$), and 100 mV RBB. The minimum retention time for each biasing condition is annotated, as well. The spread of retention time across the array is large; however, there is a clear improvement in the minimum, as well as in the average retention times with each 100 mV increase in the body bias, illustrating the effectiveness of the proposed technique.



(a)



(b)

Figure 4.11: V_{DD} = 750 mV with $\alpha_{disturb}$=25% at room temperature: (a) Minimum ($t_{ret,min}$) and maximum ($t_{ret,max}$) retention times across the entire 2 kb array, as a function of ΔV_B, and (b) retention time distributions of 2048 measured gain-cells for 100 mV FBB, standard body biasing (SBB), and 100 mV RBB.

Conclusions

This study showed the impact of body biasing on the retention time of an all-PMOS 2T gain-cell topology in a mature 0.18 µm CMOS technology. The measured retention time of a 2 kb GC-eDRAM macrocell is improved by 2.3× (from 23 to 53 ms) with a reverse body bias (RBB) of only 100 mV. The cell-to-cell retention time variability is high, ranging from 23 to 569 ms under standard body bias; the absence of a systematic pattern in the measured retention time maps suggests that the high variability is due to local parametric variations, which are particularly high in memory arrays due to the use of minimum-sized devices [56]. Moreover, the process parameters of I/O devices, used to achieve high retention times, may be less carefully controlled than those of core transistors. Nevertheless, RBB is an attractive technique to improve the minimum (as well as the average) retention time.

At the same time, the retention time penalty for FBB (used for fast memory access) is high, exhibiting a 2.9× reduction for 100 mV FBB. However, a possible control scheme could dynamically apply an RBB during retention periods and an FBB during refresh cycles to maximize the array availability. Overall, sweeping the body voltage over a range of 375 mV provides an interesting trade-off between access and retention time, with the retention time range spanning almost 2 orders of magnitude.

4.3.3 Replica Technique for Optimum Refresh Timing

The primary component of power consumption in GC-eDRAMs is the dynamic power consumed during periodic refresh operations. Refresh timing is traditionally set according to a worst-case evaluation of the retention time, under extreme environmental variations, namely process-voltage-temperature (PVT) variations, and worst-case access statistics, leading to frequent, power-hungry refresh cycles. In this Section, we present a replica technique for automatically tracking the retention time of a GC-eDRAM macrocell according to PVT variations and operating statistics, thereby reducing the data retention power of the array. A 2 kb array was designed and fabricated in a mature 0.18 µm CMOS process, appropriate for integration in ultra-low power applications such as biomedical sensors. Silicon measurements show efficient retention time tracking across a range of supply voltages and access statistics, reducing the refresh frequency by more than 5× compared to traditional worst-case design.

Replica Technique for Auto-Refresh Timing

Retention Time of a 2T Gain Cell In order to demonstrate the replica technique for optimum refresh timing, we consider the same all-PMOS 2T GC topology as for the adaptive body biasing study presented in Section 4.3.1. This GC topology is shown again in Fig. 4.12, which also illustrates the basic operating principle. The leakage power of this GC circuit, shown to be dominated by subthreshold conduction for implementation in submicron and even nanometric CMOS nodes [115], is extremely low, since during standby and write, the drain-to-

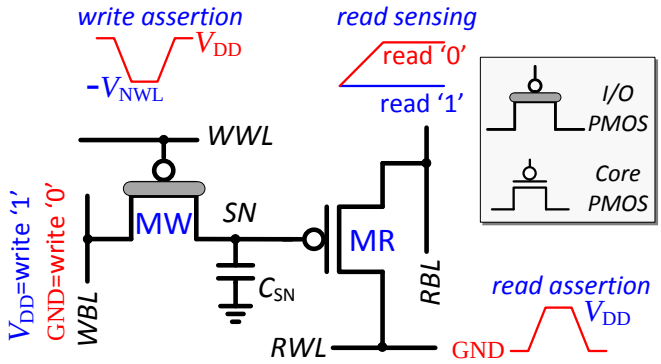


Figure 4.12: Schematic of the all-PMOS 2T gain cell with I/O write transistor (MW), including waveforms for write and read operations.

source voltage (V_{DS}) of the read transistor MR is zero, and the subthreshold leakage through the write transistor MW is limited to (dis)charging the storage node capacitor C_{SN}. The obvious issue is that any leakage to or from the storage node SN results in a degradation of the stored data level, requiring periodic refresh cycles. Therefore, the standby, or data retention power of a GC-eDRAM macrocell is given by (4.5):

$$P_{\text{retention}} = P_{\text{leakage}} + P_{\text{refresh}} = V_{DD} I_{\text{leak}} + \frac{E_{\text{refresh}}}{t_{\text{refresh}}} \tag{4.5}$$

where I_{leak} is the standby leakage current, E_{refresh} is the energy required to refresh the entire array, and t_{refresh} is the time between refresh operations. Clearly, in order to minimize the retention power, t_{refresh} must be maximized; however, in order to ensure data integrity, this parameter must be set lower than the estimated data retention time t_{ret}. Therefore, an accurate estimation of t_{ret} is required to achieve low power operation.

Various metrics have been used for simulating the retention time t_{ret} of a bitcell [130, 29, 114], but the unequivocal definition of this important parameter is the time at which the voltage written to C_{SN} degrades to the point where it results in an incorrect readout. This time is set by four primary factors: 1) the initial level stored on C_{SN} following a write; 2) the size of C_{SN}; 3) the leakage currents to and from SN; and 4) the readout mechanism. All of these factors are significantly affected by both environmental and manufacturing variations, as demonstrated by silicon measurements in [130]. This results in a large spread of the per-cell retention time across the GC-eDRAM array [64, 113] (see Fig. 4.10a in Section 4.3.2), and as with any memory array, necessitates design for the worst cell. However, in addition to the effects of PVT variations, SN leakage currents are highly sensitive to the biasing level of WBL. For a stored '1', the highest discharge leakage occurs when WBL is low, while the worst case for a stored '0' occurs when WBL is high. As shown in [130, 115, 29], the worst-case biasing for a stored '0' leads to a much lower retention time than that for a '1' in an all-PMOS 2T

cell. Consequently, in order to determine t_{refresh}, the retention time needs to be calculated assuming that WBL is constantly held high. However, this situation would only occur if a write '1' operation was executed on a given column during every clock cycle, leading to early, power-consuming refresh operations in any typical scenario.

Replica Technique Concept The design for worst-case conditions, coupled with the wide spread of t_{ret} due to PVT variations and write access disturbs, almost always results in the initiation of refresh cycles when the stored data is still at strong levels. By implementing a replica technique to track the global parametric variations, changes in the supply voltage, environmental conditions (such as the temperature), and acute operating characteristics (specifically write accesses), a significant amount of the refresh power can be saved. An additional post-silicon calibration step is implemented to adjust the tracking mechanism for each manufactured die to handle local parametric variations.

The foundation of the proposed technique relies on the superiority of the retention time for data '1' in the all-PMOS 2T gain-cell. This superiority is due to a number of factors, starting with the PMOS write transistor that easily passes a high level to SN, as opposed to a low level, which requires a WWL underdrive to completely discharge C_{SN} in a reasonable amount of time. Subsequently, both charge injection from MW and the coupling capacitance between WWL and SN drive charge onto C_{SN} during the rising edge of WWL (i.e., during the de-assertion of WWL at the culmination of a write operation), causing a slight voltage rise on SN, resulting in a degraded initial '0' state and an overcharged initial '1' state. Moreover, the decay of a '1' level due to subthreshold conduction of MW is self-limited due to the steady increase of the reverse gate overdrive and the increasing body effect of MW with progressing decay. A more detailed description of this self-limiting effect will be provided in Section 4.4.

Two primary mechanisms are incorporated to simultaneously extend the retention time of the entire array while maintaining data stability. First, during all non-write cycles, WBL is driven low, thereby enhancing the level of a stored '0' bit while minimally affecting the level of a stored '1'. Second, several replica cells are integrated within an extra column in the GC-eDRAM array and are periodically read out to analyze the state of the storage array's data retention. These replica cells are standard all-PMOS 2T bitcells, designed with slightly reduced C_{SN} (less metal stacking above the bitcells) to make them fail before the data cells, while tracking the PVT variations of the fabricated array. In addition, the replica column is designed to track the access statistics of the array, rather than assuming unlikely worst-case conditions (i.e., write operations during every clock cycle). Immediately prior to an array refresh, data '0' is written to all of the replica cells, and during read and standby cycles, the WBL of the replica column is driven low, exactly as the WBLs of all columns in the main storage array. Significant data level degradation only occurs when the WBL is high, which can only happen to a cell storing a '0' when a '1' is written to a cell on the same column. Therefore, during write cycles, the WBL of the replica column is driven high, thereby applying worst-case conditions only when they can actually occur. In this way, the retention time of the replica cells is always slightly worse

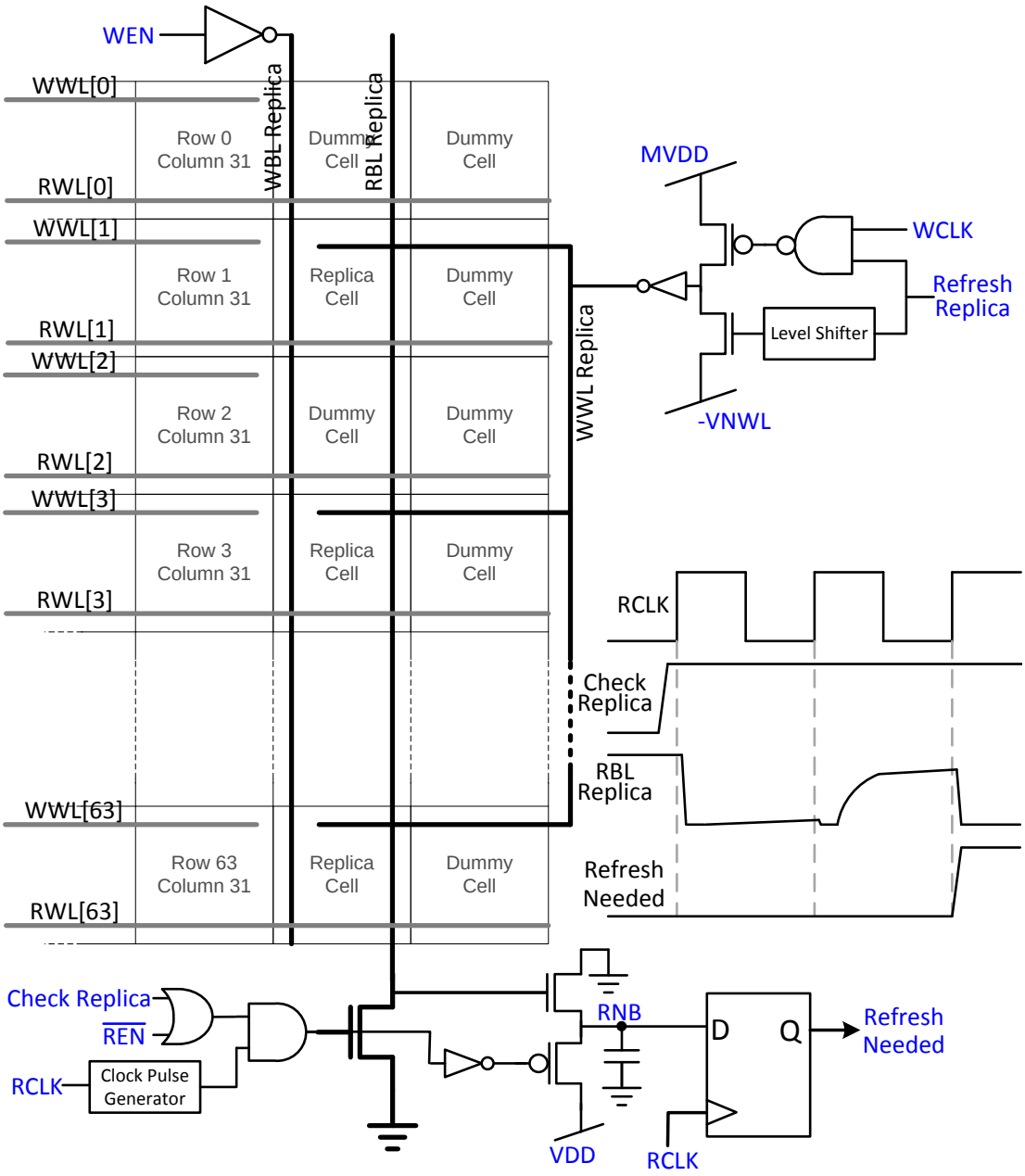


Figure 4.13: Schematic illustration of the read and write circuitry for operation and control of the proposed replica technique, including timing diagrams.

than for a data cell in a column with cells that were repeatedly written as '1' over the retention period. However, instead of assuming an extreme worst case of tying WBL to '1' (which would only occur if the array was written to during every clock cycle), this setup tracks the actual frequency of write operations. Therefore, the replica cells track the access statistics of the array (i.e., the relationship between non-write and write operations), while still ensuring that the replica cells will fail before the real data is lost.

While the mechanisms described above accurately track the global variations and access statistics of the array, local variations may result in a worst-case retention time lower than that of the worst-case replica cell. Therefore, a post-silicon calibration is used to skew the retention time of the replica cells below the measured worst-case retention time of the array. This is done by employing periodic *pseudo-write* cycles to the replica column. During these operations, the WBL of the replica column is charged, causing the replica cells to degrade at a slightly higher rate than dictated by the write statistics, thereby ensuring the initiation of an array refresh prior to a data loss in the worst cell of the array[2].

Replica Technique Integration into Gain-Cell Array The proposed replica technique was integrated into a 2 kb all-PMOS GC-eDRAM array in an 0.18 µm CMOS technology according to the schematic illustration in Fig. 4.13. A total of 32 replica cells were placed in an additional column to deal with the large distribution of local variations [64]. In order to maintain the mirrored-column symmetry of the array, a dummy column was attached to the replica column. All replica cells are written with data '0' upon the assertion of the external *RefreshReplica* signal within a single clock cycle, independent of the operation of the rest of the array. The same write mechanism as used for the data bit WWL drivers is incorporated for driving the negative write voltage to the replica cells. In order to track the write statistics of the array, the WBL of the replica column is tied to the write enable (WEN) signal. The layout of the replica cells is almost identical to the one of standard storage cells; only one layer of the metal stack is removed to reduce the C_{SN} of the replica cells and ensure a slightly lower retention time than for the regular storage cells.

Readout of the replica cells is achieved through a mechanism similar to the readout of the data cells with the addition of a designated *CheckReplica* signal. As the replica cells were designed to fail due to the deterioration of a stored '0' level, reading out a '1' from the replica column indicates the need for a refresh cycle. Therefore, the readout of such an erroneous level is propagated to the control block as the *RefreshNeeded* signal.

Testing and Characterization Procedure Testing and characterization of the replica technique was implemented with an on-chip controller, incorporating the finite-state machine (FSM) illustrated in Fig. 4.14. This controller initially writes data to the entire array, and subsequently proceeds into an *Idle* (standby) state for a configurable time period. In *Idle*, the controller initiate one of two operations. To measure tracking of write statistics, the controller initiates periodic *Disturb* cycles, during which a row of '1's (0xFFFFFFFF) is written to a pre-

[2] In an extremely unlikely case this calibration would be insufficient. This would happen if a continuous write '1' operation was applied to a column with a bitcell with worse retention time than the worst replica cell. However, this scenario would hardly ever occur in any real application. Otherwise, it is still possible to impose a write access policy to the array, which, for example, allows to write to the array only every second clock cycle. In addition, to avoid such a write access policy, it is possible to limit the WBL pulse time for the storage array, while using a pulse width equal to a full clock cycle for the replica columns.

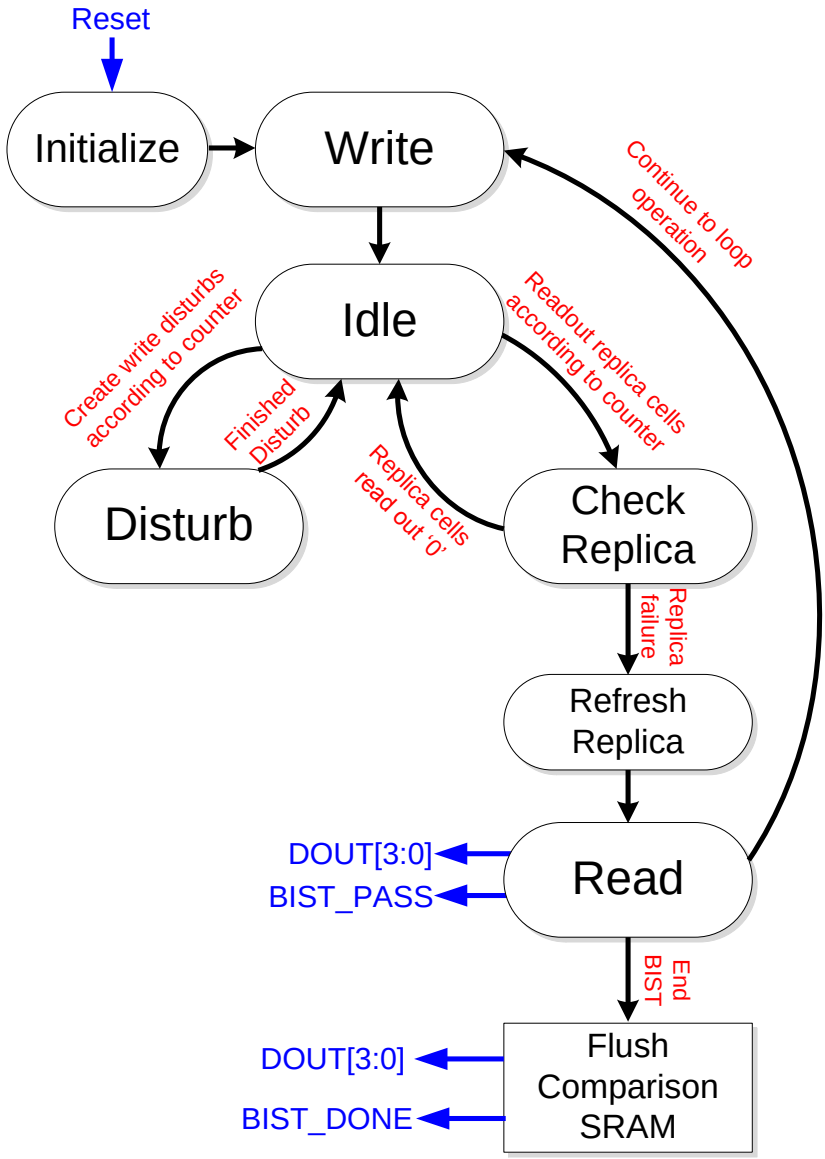


Figure 4.14: State machine of the test controller.

determined “victim" address[3]. This *Disturb* operation drives the WBL of all columns high, thereby causing deterioration of stored ‘0’ bits in the entire array. A similar mechanism is incorporated through a post-silicon calibration to further deteriorate the replica cells in order to account for local variations that may otherwise skew the retention time of the worst cell in the array below the retention time of the worst replica cell.

The second operation which can be periodically initiated from within the *Idle* state is the *CheckReplica* sequence, during which the 32 replica cells are serially read out to determine the

[3]The victim address will always store 0xFFFFFFFF, and therefore is not considered for comparison with expected responses.

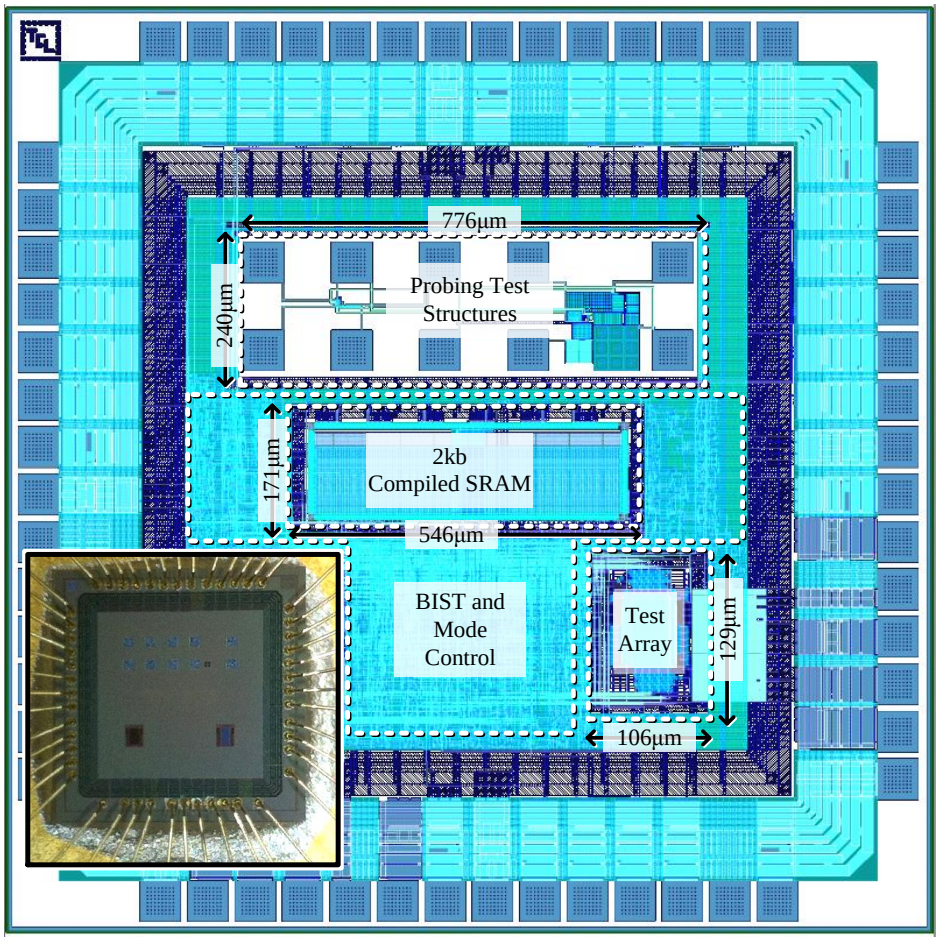


Figure 4.15: Full layout of the replica GC-eDRAM test chip with major components.

onset of a refresh operation. If the *RefreshNeeded* signal is asserted (i.e., the data in at least one of the replica cells reads out erroneously), the controller proceeds to refreshing the replica cells, before refreshing the actual storage array and looping back to the *Idle* state for another retention period.

The *Read* state (part of the array refresh sequence) of the test controller provides important measurement data for analysis. The read-out data is compared with the originally written data to ensure equality and the per-bit comparison results are stored in an on-chip 2 kb SRAM. Concurrently, the one-bit comparison result of the currently read row is driven off-chip via the *BIST_PASS* signal, and the four MSBs are propagated to the external *DOUT*[3:0] pads to enable further observation. An external interrupt signal can break the refresh loop, sending the controller into its termination state, during which the comparison data can be analyzed. In this state, the *BIST_DONE* signal is raised, and subsequently, the full, per-bit comparison data that was stored in the SRAM is flushed out to the *DOUT*[3:0] pads by means of scan chains. This control scheme enables at-speed testing of the GC-eDRAM array, including the ability to observe the functionality of the replica technique under various write disturb statistics.

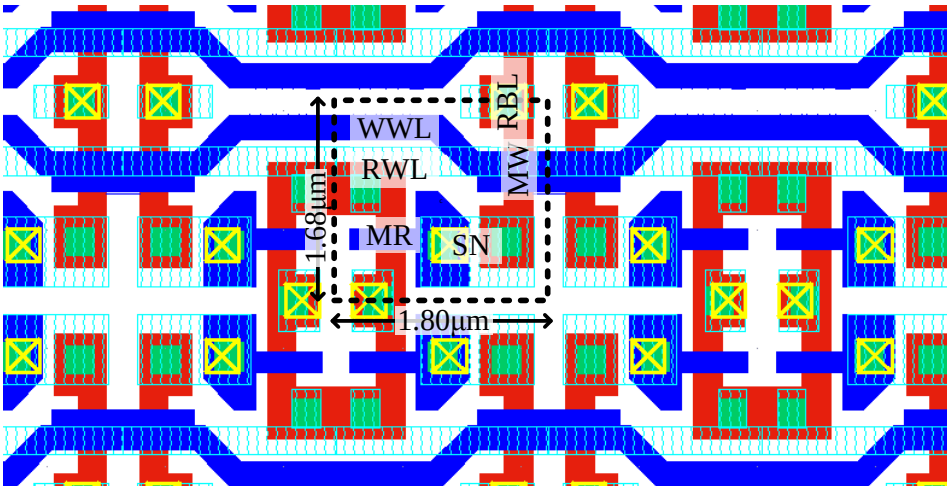


Figure 4.16: Small section of the GC-eDRAM array layout showing the dimensions of the unit cell.

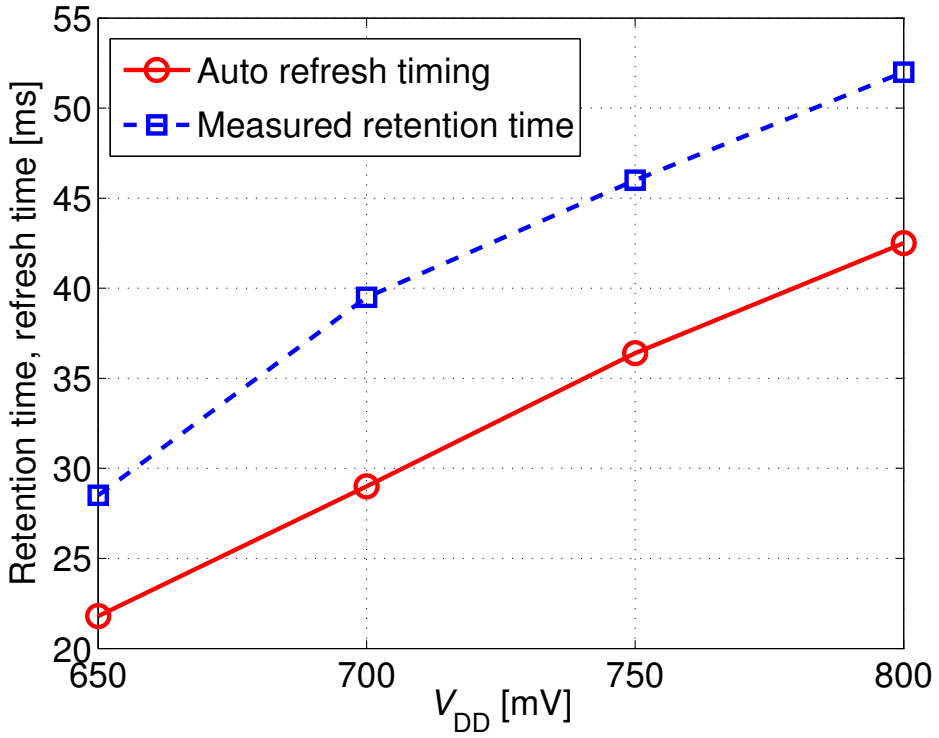


Figure 4.17: Automatic refresh timing vs. measured retention time for a range of supply voltages.

4.3.4 Replica GC-eDRAM: Silicon Measurements

A 2 kb (64×32) GC-eDRAM array with integrated replica technique was designed and fabricated in a commercial 0.18 µm CMOS technology, as part of the test chip shown in Fig. 4.15. In

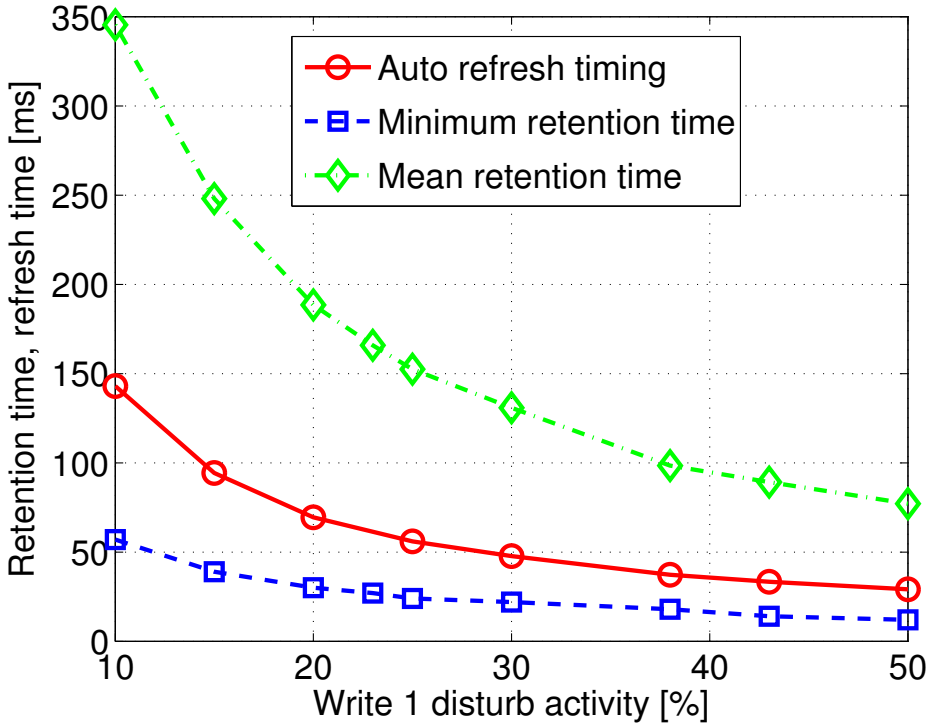


Figure 4.18: Automatic refresh timing vs. measured retention time for a varying degree of write disturbs.

addition to the array, the test chip included the on-chip test controller, the 2 kb SRAM for data comparison, and several other test components. The test chip was designed to enable three primary test modes: full, at-speed, controller testing; array operation through scan chain configuration; and external direct access to the array. A combination of these three modes was used to test the functionality of the array and produce the measurement data shown below.

The GC-eDRAM bitcell was laid out in a compact array with mirrored rows and columns, as shown in Fig. 4.16, with a unit cell size of 3.024 μm^2 (1.8 µm×1.68 µm). The array, including peripheral circuits, occupies 0.013 mm^2 (106 µm×129 µm) and is biased by a separate, low-voltage supply (MVDD) different than the supply (VDD) of the BIST and the other digital peripheral circuits of the test chip. In addition, an external negative voltage is supplied for write underdrive.

Fig. 4.17 illustrates the ability of the replica technique to automatically track the retention time of the array. The figure shows the automatically triggered refresh period for various supply voltages, as compared to the minimum retention time measured at this voltage, following a post-silicon adjustment in the write disturb frequency to account for local variations. Refresh is consistently initiated just prior to the array's minimum retention time for a range of supply voltages.

Tracking of the write statistics is shown in Fig. 4.18. This figure plots the automatic refresh

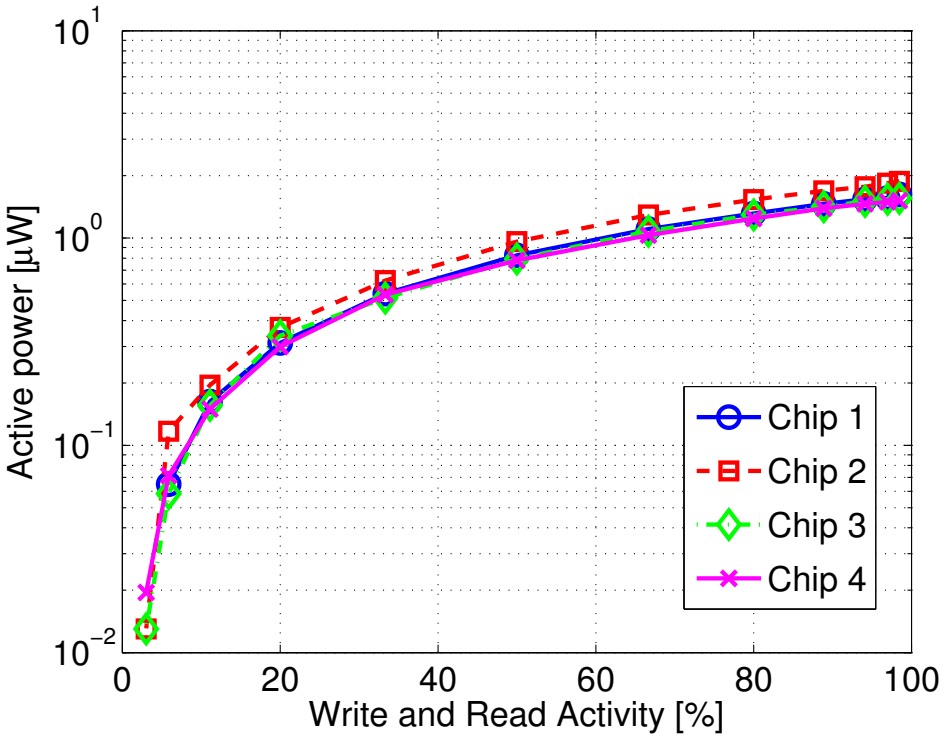


Figure 4.19: Dynamic power consumption of 2 kb GC-eDRAM array as a function of the write and read activity factor for several measured chips.

timing, as compared to the measured retention time of the array with a given frequency of write operations. The figure shows both the mean and minimum retention time of the array. For the shown die, the minimum retention time is lower than the uncalibrated automatic refresh timing; however, the write activity tracking mechanism is shown to work correctly, such that the post-silicon calibration can easily skew the refresh time to a value below the worst-case retention time. This plot emphasizes the efficiency of integrating the replica technique. Traditional worst-case design assumes 100% write activity, resulting in a refresh period of well below 10 ms, even for this typical die. Application of the replica technique adapts this period, refreshing at a more than 5× lower frequency for 10% write activity.

Fig. 4.19 shows the dynamic power consumption of the array, as a function of the write and read activity. For a retention period of 20 ms, the active refresh power of the array is 635 fW/bit, which is comparable with previous low-power GC-eDRAM implementations [130].

Conclusions

In this Section, we proposed a replica technique for tracking the PVT variations and operating statistics of a near-V_T GC-eDRAM array for efficient data retention time extension. The technique was implemented on a 2 kb all-PMOS 2T array in a commercial 0.18 µm CMOS process

along with an advanced control scheme for extensive testing and measurement. The replica technique was shown to effectively track the retention time of the array across various supply voltages and write activity frequencies, enabling as much as a 5× improvement in retention time, thereby significantly reducing the frequency of power-hungry refresh operations.

4.4 Aggressive Technology and Voltage Scaling (to Sub-V_T Domain)

This Section considers the design and operation of GC-eDRAMs at aggressively scaled supply voltages (residing in the subthreshold regime if possible) and under aggressive technology scaling (down to 40 nm CMOS nodes). In fact, ultra-low power applications often require several kb of embedded memory and are typically operated at the lowest possible operating voltage (V_{DD}) to minimize both dynamic and static power consumption. Embedded memories can easily dominate the overall silicon area of these systems, and their leakage currents often dominate the total power consumption. Gain-cell based embedded DRAM arrays provide a high-density, low-leakage alternative to SRAM for such systems; however, they are typically designed for operation at nominal or only slightly scaled supply voltages, sometimes including near-V_T voltages (see all previous Sections in this Chapter). This Section presents a gain-cell array which, for the first time, targets aggressively scaled supply voltages, down into the subthreshold (sub-V_T) domain. Minimum V_{DD} design of gain-cell arrays is evaluated in light of technology scaling, considering both a mature 0.18 µm CMOS node, as well as a scaled 40 nm node. We first analyze the trade-offs that characterize the bitcell design in both nodes, arriving at a best-practice design methodology for both mature and scaled technologies. Following this analysis, we propose full gain-cell arrays for each of the nodes, operated at a minimum V_{DD}. We find that an 0.18 µm CMOS gain-cell array can be robustly operated at a sub-V_T supply voltage of 400 mV, providing read/write availability over 99% of the time, despite refresh cycles. This is demonstrated on a 2 kb array, operated at 1 MHz, exhibiting full functionality under parametric variations. As opposed to sub-V_T operation at the mature node, we find that the scaled 40 nm node requires a near-threshold 600 mV supply to achieve at least 97% read/write availability due to higher leakage currents that limit the bitcell's retention time. Monte Carlo simulations show that a 600 mV 2 kb 40 nm gain-cell array is fully functional at frequencies higher than 50 MHz. Briefly, GC-eDRAMs implemented in mature CMOS nodes can successfully be operated at at aggressively scaled sub-V_T voltages, whereas voltage scaling for GC-eDRAM implementations in aggressively scaled CMOS nodes is best limited to the near-V_T domain.

4.4.1 Introduction

Many ultra-low power (ULP) systems, such as biomedical sensor nodes and implants, are expected to run on a single cubic-millimeter battery charge for days or even for years, and therefore are required to operate with extremely low power budgets. Aggressive supply voltage scaling, leading to near-threshold (near-V_T) or even to subthreshold (sub-V_T) circuit operation,

is widely used in this context to lower both active energy dissipation and leakage power consumption; albeit, at the price of severely degraded on/off current ratios (I_{on}/I_{off}) and increased sensitivity to process variations [71]. The majority of these biomedical systems require a considerable amount of embedded memory for data and instruction storage, often amounting to a dominant share of the overall silicon area and power. Typical storage capacity requirements range from several kb for low-complexity systems [89] to several tens of kb for more sophisticated systems [13]. Over the last decade, robust, low-leakage, low-power sub-V_T memories have been heavily researched [6, 4, 82]. In order to guarantee reliable operation in the sub-V_T domain, many new SRAM bitcells consisting of 8 [90, 77], 9 [4, 135], 10 [6], and up to 14 [89] transistors have been proposed (see Section 3.1 for more details). These bitcells utilize the additional devices to solve the predominant problems of write contention and bit-flips during read, and, in addition, some of the designs reduce leakage by using transistor stacks. All these state-of-the-art sub-V_T memories are based on static bitcells, while the advantages and drawbacks of dynamic bitcells for operation in the sub-V_T regime have not yet been studied.

Remember that conventional 1-transistor-1-capacitor (1T-1C) embedded DRAM (eDRAM) is incompatible with standard digital CMOS technologies due to the need for high-density stacked or trench capacitors. Therefore, it cannot easily be integrated into a ULP system-on-chip (SoC) at low cost. Moreover, low-voltage operation is inhibited by the offset voltage of the required sense amplifier, unless special offset cancellation techniques are used [131].

Gain-cells are a promising alternative to SRAM and to conventional 1T-1C eDRAM, as they are both smaller than any SRAM bitcell, as well as fully logic-compatible. Recall from Section 4.1 that much of the previous work on GC-eDRAMs focuses on high-speed operation, in order to use gain-cells as a dense alternative to SRAM in on-chip processor caches [125, 28], while only a few publications deal with the design of low-power near-V_T gain-cell arrays [130, 132]. The possibility of operating gain-cell arrays in the sub-V_T regime for high-density, low-leakage, and voltage-compatible data storage in ULP sub-V_T systems has not been exploited yet. One of the main objections to sub-V_T gain-cells are the degraded I_{on}/I_{off} current ratios, leading to rather short data retention times compared to the achievable data access times. However, in the following we show that these current ratios are still high enough in the sub-V_T regime to achieve short access and refresh cycles and high memory availability, at least down to 0.18 µm CMOS nodes. While gain-cells are considerably smaller than robust sub-V_T 8−14T SRAM bitcells, they also exhibit lower leakage currents, especially in mature CMOS nodes where sub-V_T conduction is the dominant leakage mechanism. Recent studies for above-V_T, high-speed caches show that gain-cell arrays can even have lower retention power (leakage power plus refresh power) than SRAM (leakage power only) [117]. However, a direct power comparison between GC-eDRAM and SRAM is difficult and not within the scope of this study; for example, an ultra-low power sub-V_T SRAM implementation [89] employs power gating of all peripheral circuits and of the read-buffer in the bitcell, while most power reports for gain-cell eDRAMs include the overhead of peripherals. Compared to SRAM, gain-cells are naturally suitable for two-port memory implementation, which provides an advantage in terms of memory bandwidth, and enables simultaneous and independent optimization of

write and read reliability. Finally, while local parametric variations directly compromise the reliability of the SRAM bitcell (write contention, and data loss during read), such parametric variations only impact the access time (and the retention time) of gain-cells, which is not a severe issue when targeting the typically low speed requirements of ULP applications, such as sub-V_T sensor nodes or biomedical implants.

To start with, in this Section, we consider sub-V_T GC-eDRAM design in a mature 0.18 µm CMOS node, which is typically used to: 1) easily fulfill the high reliability requirements of ULP systems; 2) reach the highest energy-efficiency of such ULP systems, typically requiring low frequencies and duty cycles [133]; and 3) achieve low manufacturing costs. In a second step, we investigate the feasibility of sub-V_T gain-cell eDRAMs under the aspect of technology scaling. In particular, in addition to the mature 0.18 µm CMOS node, we analyze low voltage gain-cell operation in a 40 nm CMOS technology node. We show that deep-nanoscale gain-cell arrays are still feasible, despite the reduced retention times inherent to these nodes. Due to high refresh rates, we identify that the minimum supply voltage (V_{DDmin}) that ensures an array availability of 97% is in the near-V_T domain.

Contributions The contributions of the work presented in this Section can be summarized as follows:

- We investigate the minimum achievable supply voltage for ultra-low power gain-cell operation.
- We analyze gain-cell arrays from a technology scaling perspective, examining the design trade-offs that arise due to the inherent characteristics of various technology nodes.
- For the first time, we present a fully functional gain-cell array at a deeply scaled technology node, as low as 40 nm.
- For the first time, we present a gain-cell array operated in the sub-V_T domain.

Outline Section 4.4.2 explains the best-practice 2T gain-cell design in light of technology scaling, emphasizing the optimum choices of the write access transistor, read access transistor, storage node capacitance, and word line underdrive voltage for different nodes. Sections 4.4.3 and 4.4.4 present detailed implementation results of a 2 kb gain-cell memory in a 0.18 µm and in a 40 nm CMOS node, respectively, before Section 4.4.5 summarizes the all findings.

4.4.2 Two-Transistor (2T) Sub-V_T Gain-Cell Design

Previously reported gain-cell topologies include either two or three transistors and an optional MOSCAP [29]. While the basic two-transistor (2T) bitcell has the smallest area cost, it limits the number of cells which can connect to the same read bitline (RBL) due to leakage currents from

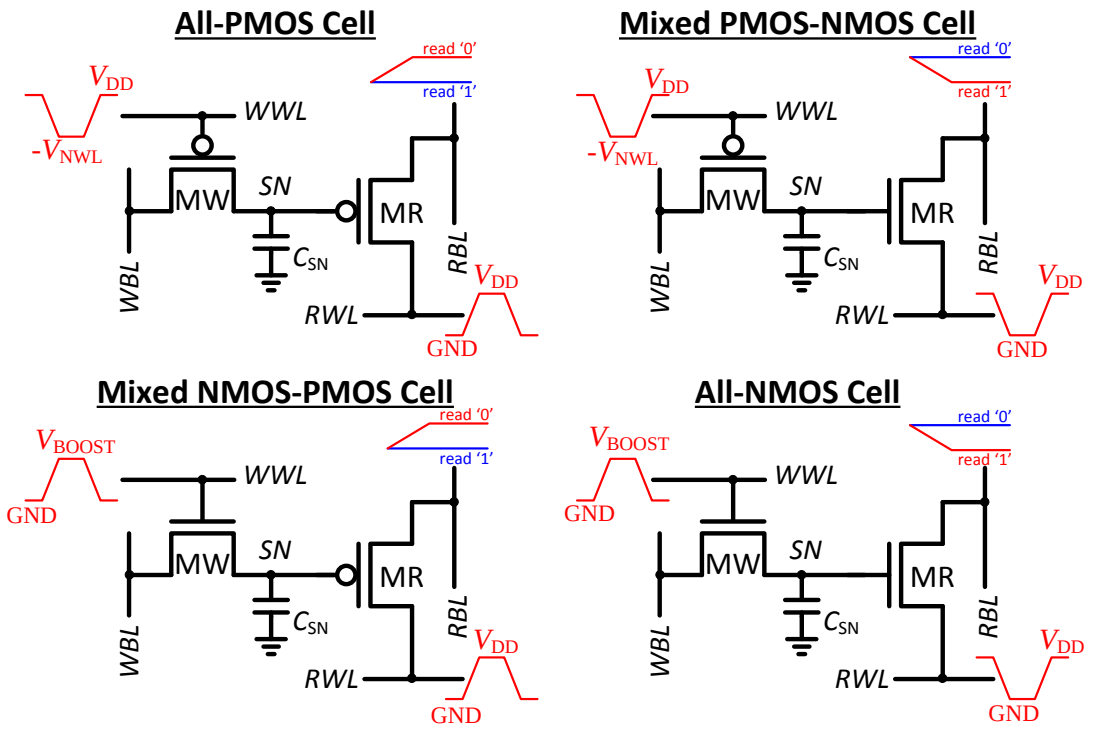


Figure 4.20: 2T gain-cell implementation options including the schematic waveforms.

unselected cells masking the sense current [63]. However, as many ULP systems require only small memory arrays with relatively few cells per RBL, in the following section, we consider the implementation of a 2T bitcell as a viable low-voltage option and propose a best-practice 2T bitcell design for the considered technology nodes (0.18 µm and 40 nm).

2T Gain-Cell Implementation Alternatives

Fig. 4.20 shows the four basic options for implementing a 2T gain-cell, allowing both the write transistor (MW) and the combined storage and read transistor (MR) to be implemented with either an NMOS or a PMOS device. These standard topologies require the following control schemes to achieve robust write and read operations. A boosted write wordline (WWL) voltage is required during write access due to V_T drop across MW; above V_{DD} for the NMOS option (V_{BOOST}) and below V_{SS} for the PMOS option (V_{NWL}). For a read operation with a PMOS MR, the parasitic RBL capacitance is pre-discharged, and the read wordline (RWL) is subsequently raised. If the selected bitcell's storage node (SN) holds a '0', MR is conducting and charges RBL past a detectable sensing threshold. If SN holds a '1', MR is cut off, such that RBL remains discharged below the sensing threshold. Using an NMOS transistor to implement MR provides the exact opposite operation, i.e., RBL is pre-charged and RWL is lowered to initiate a read.

In the considered 0.18 µm CMOS technology, both MW and MR can be implemented with

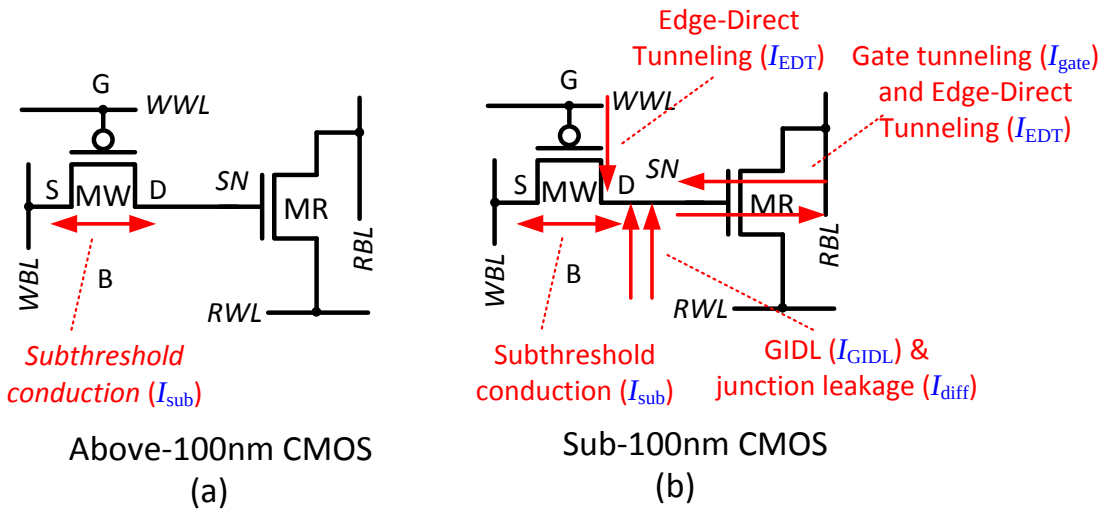


Figure 4.21: Leakage components which are considered for the choice of the best-practice write and read transistor implementations, for (a) mature CMOS nodes, and (b) scaled CMOS nodes.

either standard-V_T core or high-V_T I/O devices. In more advanced technology nodes, typically starting with the 130 nm or 90 nm node for most semiconductor foundries, several V_T options become available for core devices, most commonly low-V_T (LVT), standard-V_T (SVT), and high-V_T (HVT) devices. One of the primary considerations for gain-cell implementation is achieving high retention time, i.e., the time it takes for the level stored on SN to deteriorate through leakage currents. In mature, above-100 nm CMOS nodes, subthreshold conduction is the dominant leakage mechanism, compromising data retention in any 2T gain-cell through the channel of MW, as shown in Fig. 4.21(a). Therefore, the primary selection criterion for the device type of MW is to minimize subthreshold conduction. Note that subthreshold conduction of MW weakens both a logic '1' and a logic '0' level, whenever the write bitline (WBL) voltage is opposite to the SN voltage.

In more advanced, sub-100 nm CMOS nodes, there are other significant leakage mechanisms that can compromise data integrity[4]. Only leakage components that bring charge onto the SN or take charge away from SN need to be considered in terms of retention time, while other leakage components are merely undesirable in terms of static power consumption. Fig. 4.21(b) schematically shows the main leakage components that can compromise the stored level in sub-100 nm nodes, including reverse-biased pn-junction leakage (I_{diff}), gate-induced drain leakage (I_{GIDL}), gate tunneling leakage (I_{gate}), edge-direct tunneling current (I_{EDT}), and subthreshold conduction (I_{sub}). When employing a PMOS MW, the bulk-to-drain leakages (I_{diff} and I_{GIDL}) weaken a logic '0' and strengthen a logic '1', but have the opposite impact (strengthen a logic '0' and weaken a logic '1') when MW is implemented with an

[4] Note that in the sub-V_T region, these mechanisms are still negligible, as compared to subthreshold conduction. However, as shown in Section 4.4.2, at near-V_T supplies, some of the mechanisms still must be considered.

NMOS device. During standby, MW is always off and has no channel; therefore, forward gate tunneling (I_{gate}) from the gate into the channel region and into the two diffusion areas that would occur in a turned-on MOS device is of no concern here. Only the edge-direct tunneling current, from the diffusion connected to the SN in the absence of a strongly inverted channel, compromises data integrity. When using an NMOS MW, edge-direct tunneling discharges a logic '1', while it charges a logic '0' for a PMOS MW.

The only leakage through MR that affects the stored data level is gate tunneling. During standby, there is no channel formation in MR, no matter what the stored data level is. For example, if using an NMOS MR, both RWL and RBL are charged to V_{DD} during standby, such that even a logic '1' level results in zero gate overdrive. In this case, both diffusion areas of MR are at the same potential as the SN, eliminating tunneling currents between the diffusions and the gate ($I_{EDT} = 0$). However, tunneling might occur from the gate directly into the grounded bulk (I_{gate}), weakening a logic '1'. If the same cell stores a logic '0', tunneling between the gate and bulk is avoided ($I_{gate} = 0$), while reverse tunneling from the diffusions (I_{EDT}) into the gate can charge the logic '0' level. The exact opposite biasing conditions and corresponding tunneling mechanisms are found when implementing MR with a PMOS.

Best-Practice Write Transistor Implementation

Mature 0.18 µm CMOS Node For the ULP sub-V_T applications, long retention times that minimize the number of power-consuming refresh cycles are of much higher importance than fast write access. Therefore, low subthreshold conduction becomes the primary factor in the choice of a best practice write transistor in the 0.18 µm node. The subthreshold conduction of NMOS and PMOS, core and I/O devices offered in this process are shown in Fig. 4.22a. Clearly, the I/O PMOS device has the lowest subthreshold conduction I_{sub} ($V_{GS} = 0\,V$, $V_{DS} = -V_{DD}$) among all device options and across all standard process corners, leading to the longest retention time. At a 400 mV sub-V_T V_{DD}, the on-current I_{on} ($V_{GS} = -V_{DD}$, $V_{DS} = -V_{DD}$) of this preferred I/O PMOS device is still four orders of magnitude larger than I_{sub}, as shown in Fig. 4.22b, which results in sufficiently fast write and refresh operations compared to the achievable retention time. This holds for temperatures up to 37 °C, which is considered a maximum, worst-case temperature for ULP systems that are often targeted at biomedical applications, typically attached to the human body, and hardly suffer from self-heating due to low computational complexity. Nevertheless, for temperatures as high as 125 °C, a sufficiently high I_{on}/I_{sub} ratio of four orders of magnitude is still achieved at a slightly higher supply voltage of 500 mV.

Fig. 4.23a shows the worst-case time dependent data deterioration after writing into a 2T gain-cell with a PMOS I/O write transistor under global and local variations. The blue (bottom) curves show the deterioration of a logic '0' level with WBL tied to V_{DD}, and the red (top) curves show the deterioration of a logic '1' level with WBL tied to ground. The plot was simulated with a sub-V_T 400 mV V_{DD} assuming a storage node capacitance of 2.5 fF. A worst-case retention time of 40 ms can be estimated from this figure, corresponding to the minimum time at

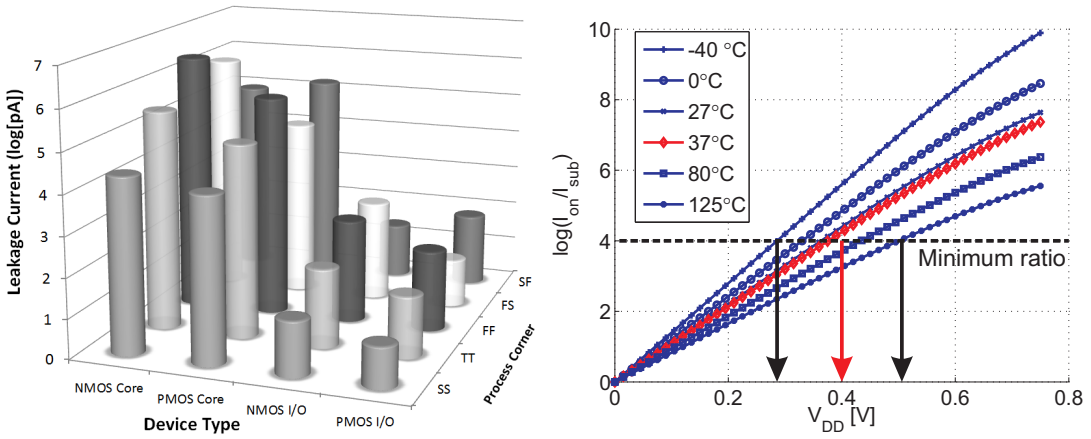


Figure 4.22: (a) Subthreshold conduction of different transistor types in an 0.18 µm node, and (b) I/O PMOS I_{on}/I_{sub} current ratio as a function of V_{DD} for the typical-typical (TT) process corner at different temperatures.

which the '0' and '1' levels intersect. It is clear that a logic '0' level decays much faster than a logic '1' level, corresponding with previous reports for the above-V_T domain [130, 125]. In fact, the decay of a '1' level is self-limited due to the steady increase of the reverse gate overdrive ($V_{GS,MW} = V_{DD} - V_{SN}$) and the increasing body effect ($V_{BS,MW} = V_{DD} - V_{SN}$) of MW with progressing decay. Both of these effects suppress the device's leakage. Furthermore, the charge injection (CI) and clock feedthrough (CF) that occur at the end of a write access (when MW is turned off), cause the SN voltage level to rise, strengthening a '1' and weakening a '0' level [29, 114]. Therefore, careful consideration must be given to the initial state of the '0' level following a write access, as will be discussed in Section 4.4.2.

Scaled 40 nm CMOS Node While choosing the best device option for MW, subthreshold conduction must again be kept as small as possible, as it affects both a '1' and a '0' level. The diffusion leakage, the GIDL current, and the edge-direct tunneling current weaken one logic level, while they strengthen the other. However, all three leakage components work against the logic level which has already been weakened through CI and CF at the end of a write pulse. For example, with a PMOS MW, the logic '0' level is weakened through a positive SN voltage step when closing MW, while I_{GIDL}, I_{diff}, and I_{EDT} further pull up SN, deteriorating the stored '0'. Therefore, in order to protect the already weaker level, the optimum device selection aims at minimizing all of these leakage components. Fig. 4.24a shows the leakage components of minimum sized devices provided in the 40 nm process[5] at a near-V_T supply voltage of 600 mV. This figure clearly shows that despite the increasing significance of other leakage currents with

[5]The LVT devices were left out of the figure for display purposes, as their leakage is significantly higher than the leakage of other devices.

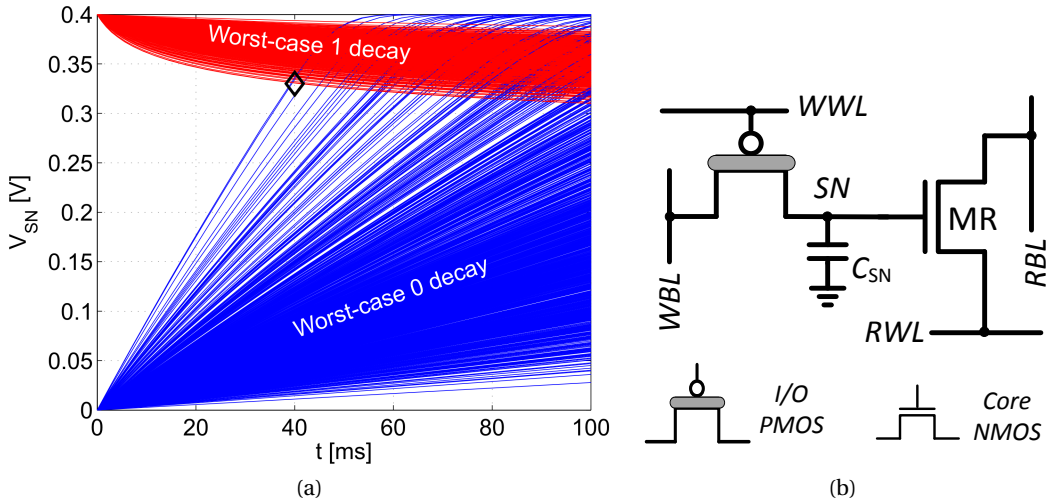


Figure 4.23: (a) Worst-case retention time estimation of 0.18 µm sub-V_T gain-cell with V_{DD} = 400 mV. (b) Best-practice gain-cell for sub-V_T operation in 0.18 µm CMOS.

technology scaling, I_{sub} is still dominant at this node[6]. However, the advantage of using an I/O device is lost, and a more compact HVT PMOS device provides the lowest total leakage. This trend is confirmed when evaluating the leakage components of intermediate process nodes, as well, showing that the leakage benefits of using an I/O device deteriorate to the point where the area versus leakage trade-off favors the use of an HVT device at around the 65 nm node.

Best-Practice Read Transistor Implementation

Mature 0.18 µm CMOS Node At the onset of a read operation, capacitive coupling from RWL to SN causes a voltage step on SN [114]. Our analysis from the previous section showed that MW should be implemented with a PMOS device, resulting in a strong logic '1' and a weaker logic '0'. Therefore, it is preferable to implement MR with an NMOS transistor that employs a negative RWL transition for read assertion. The resulting temporary[7] decrease in voltage on SN counteracts the previous effects of CI and CF, thus improving the '0' state during a read operation. As a side effect, this negative SN voltage step also lowers the '1' level and therefore slightly slows down the read operation; however, this level is already initially boosted due to deassertion of the WWL. An additional, and perhaps more significant reason to choose an NMOS device for readout is that NMOS devices are approximately an order-of-magnitude stronger than their PMOS counterparts at sub-V_T voltages. Therefore, implementing MR with an NMOS device provides a fast read access, which not only results in better performance, but is essential for ensuring high array availability. As mentioned, the considered 0.18 µm process

[6] Some of the leakage components are not modeled for the I/O devices; however, this does not impact our analysis, as the PMOS HVT already provides the lowest total leakage.

[7] The effect is reversed upon deassertion of the RWL.

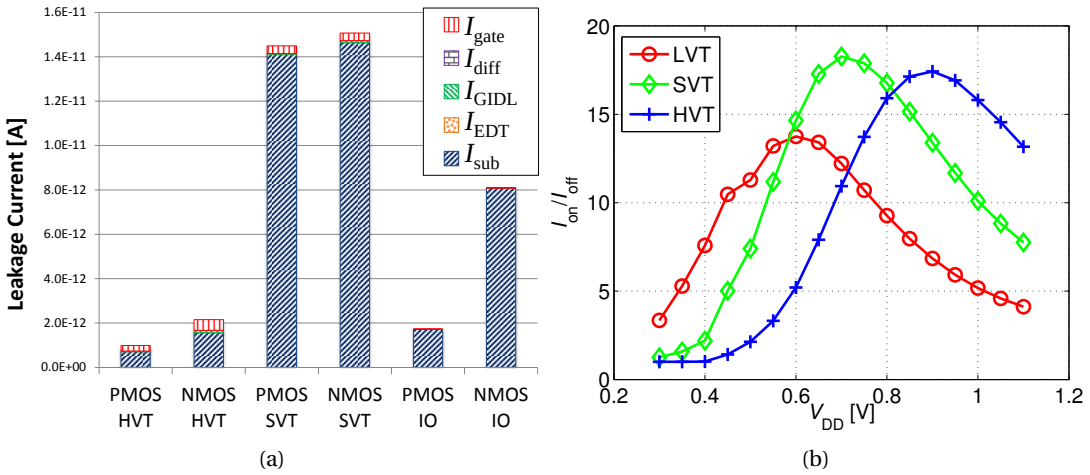


Figure 4.24: (a) Leakage components of various devices in the considered 40 nm node at a near-V_T supply voltage of 600 mV. (b) Worst-case I_{on}(weak'1')/I_{off}(weak'0') of MR, implemented with LVT, SVT, and HVT devices. Both plots were simulated under typical conditions.

provides core and I/O devices, and considering the three-orders-of-magnitude higher on-current for core devices at sub-V_T voltages, the choice of an NMOS core MR is straightforward.

To summarize, the most appropriate 2T gain-cell for sub-V_T operation in an above-100 nm CMOS node comprises an I/O PMOS write transistor and a core NMOS read transistor, as illustrated in Fig. 4.23b. The resulting hybrid NMOS/PMOS gain-cell shares the n-well on three sides between neighboring cells [63] to keep the area cost low, as discussed in Section 4.4.3.

Scaled 40 nm CMOS Node When considering the best device type for scaled nodes, the large number of options presents some interesting trade-offs for the implementation of MR. The increasing gate leakage currents (I_{gate} and I_{EDT}) at scaled nodes could potentially present an advantage for a thick oxide I/O device due to its reduced gate tunneling. However, at low voltages, the tunneling currents are small in comparison with the subthreshold conduction through MW, as shown in Fig. 4.24a. In addition, I_{gate} and I_{EDT} actually appear in opposite directions, as the stored '0' level rises, further reducing their impact. On the other hand, the two primary considerations for the above-100 nm nodes are even more relevant at scaled nodes. The achievable retention time in the 40 nm process turns out to be approximately three orders-of-magnitude lower than that of the 0.18 µm node. Therefore, the negative step caused by RWL coupling to SN is even more important, and fast reads are essential to provide sufficient array availability, despite the high refresh rates. To further enhance the read step, layout techniques can be implemented to increase the capacitive coupling between RWL and SN. However, when considering read access times, additional trade-offs arise. For maximum read performance, MR could be implemented with an LVT device. At the 40 nm node, an LVT NMOS provides an 8× increase in on-current at 400 mV compared to an SVT NMOS. However, as the supply voltage is increased, this benefit reduces to 3× at 600 mV. The superior

on-currents of LVT devices, as compared to SVT or HVT options, come at the expense of much higher off-currents, as well as increased process variations. When choosing the read device, this trade-off must be taken into consideration, as it is mandatory to correctly differentiate between the discharged level of RBL due to a stored '1' and the depleted level due to a weak stored '0'. Furthermore, the unselected cells on the same column of a selected cell storing a '1' will start to counteract the discharge of RBL during a read, as $V_{GS,MR}^{unselected} = V_{DD} - V_{RBL}$. In effect, this limits the speed and minimum discharge level of RBL, according to the drive strength of the unselected MR devices. When considering sub-V_T operation in the 40 nm node, the relatively low subthreshold conduction of the SVT, HVT, and I/O devices, render the LVT the only feasible option for MR to achieve a reasonable RBL discharge time. However, as V_{DD} is increased into the near-V_T region, an SVT device provides sufficient on-current, while the higher V_T and lower leakage enable better reliability under process variations, as well as improved array availability.

Fig. 4.24b shows the worst case current ratio I_{on}/I_{off} of the NMOS read transistor MR, implemented with different device types as a function of V_{DD}. I_{on} is given for a weak '1' level, estimated as the steady state high voltage of SN when tying WBL to V_{DD} ($V_{SN} = 0.85V_{DD}$). I_{off} is given for a weak '0' level, estimated at $V_{SN} = 0.4V_{DD}$, which would provide a sufficient margin to differentiate between the two levels[8]. For supply voltages below 600 mV, the LVT device has the highest current ratio and is therefore preferred, as it provides the best achievable array availability. Likewise, the SVT device is preferred for V_{DD} between 600 and 800 mV, while the HVT device is the best option for even higher V_{DD}.

Storage Node Capacitance and WWL Underdrive Voltage

Mature 0.18 µm CMOS Node To close the design of the 2T bitcell, two important design parameters must be taken into consideration. First, the storage node capacitance (C_{SN}), primarily made up of the diffusion capacitance of MW and the gate capacitance of MR, is typically around 1 fF for minimum device sizes. However, we find that by applying layout techniques, such as metal stacking, this value can be extended by over 5×, providing a configurable design parameter. Second, to address the V_T drop across MW especially affecting the write '0' operation (but also the write '1' operation in the sub-V_T regime), an underdrive voltage (V_{NWL}) needs to be applied to WWL, the magnitude of which affects the write access time and the SN voltage.

Fig. 4.25a shows the storage node voltage (V_{SN}) after a write '0' access as a function of C_{SN} and V_{NWL}, before and after closing MW. Fig. 4.25b emphasizes the impact of CI and CF by showing the voltage step ΔV that occurs while closing MW. It is clear that any V_{NWL} above −650 mV already results in a degraded logic '0' transfer prior to turning off MW. ΔV can be reduced by increasing C_{SN} and by decreasing the magnitude of V_{NWL}. Therefore, on the one hand, V_{NWL} must be low enough to ensure a proper logic '0' transfer, while, on the other hand, it should be

[8] This is verified for the chosen implementation at the minimum feasible bias in Section 4.4.4.

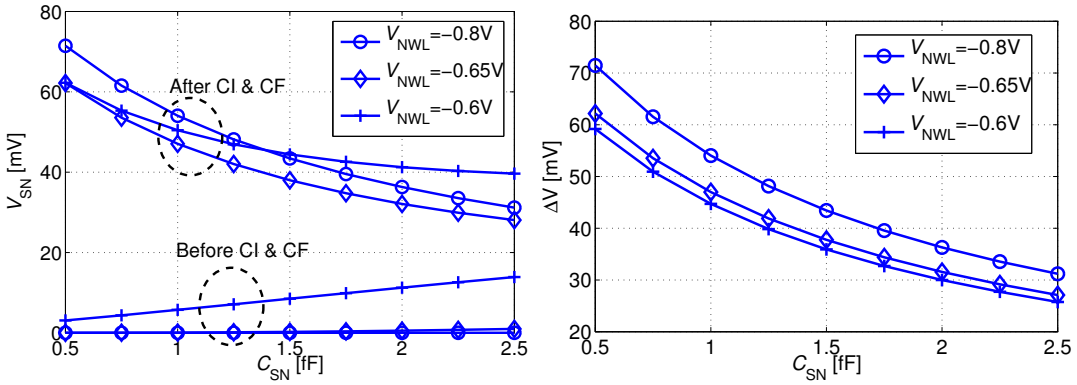


Figure 4.25: Following a write '0' operation: (a) V_{SN} before and after closing MW, as a function of C_{SN} and V_{NWL}. (b) ΔV due to charge injection from MW and due to capacitive coupling from WWL to SN.

as high as possible to minimize ΔV. The optimum value for V_{NWL} leading to the strongest '0' state after a completed write operation is found to be −650 mV, as shown in Fig. 4.25a. The optimum value for C_{SN} is clearly the maximum displayed value of 2.5 fF.

Scaled 40 nm CMOS Node It is clear that the storage node capacitance should always be as big as possible, regardless of the technology node. This not only results in an improved initial '0' level, as shown above, but also provides more stored charge and thus extends the retention time. A general characteristic of scaled CMOS nodes is the increased number of routing layers which, in the case of gain-cell design, can be used to build up the storage node capacitor. Here, we assume that all available metal layers can be used at no additional cost, as the memory is going to be embedded in a system-on-chip which already uses all the metal layers. Moreover, with technology scaling, the aspect ratio of metal wires changes to narrower but higher, and wires can be placed closer to each other, which is beneficial in terms of side-wall parasitic capacitance. However, much of this benefit is offset by the lower dielectric constants of the insulating materials (*low-k*) integrated into digital processes with technology scaling. In addition, the absolute footprint of the bitcell shrinks with technology, making it more challenging to allocate many inter-digit fingers for a high capacitance. In fact, in the considered 40 nm node, the footprint of a gain-cell containing only two core devices is so small that the minimum width and spacing rules for medium and thick metals are too large to exploit these metals for increasing the capacitance of the SN. Therefore, our layout of the 40 nm cell is limited to 5 routing layers, and the overall SN capacitance is much lower than that achieved in the 0.18 µm node. Fig. 4.26a summarizes the achievable storage node capacitance according to the number of thin metal layers provided by the two considered technology nodes.

Fig. 4.26b shows the SN voltage step ΔV of the 40 nm CMOS gain-cell that occurs during

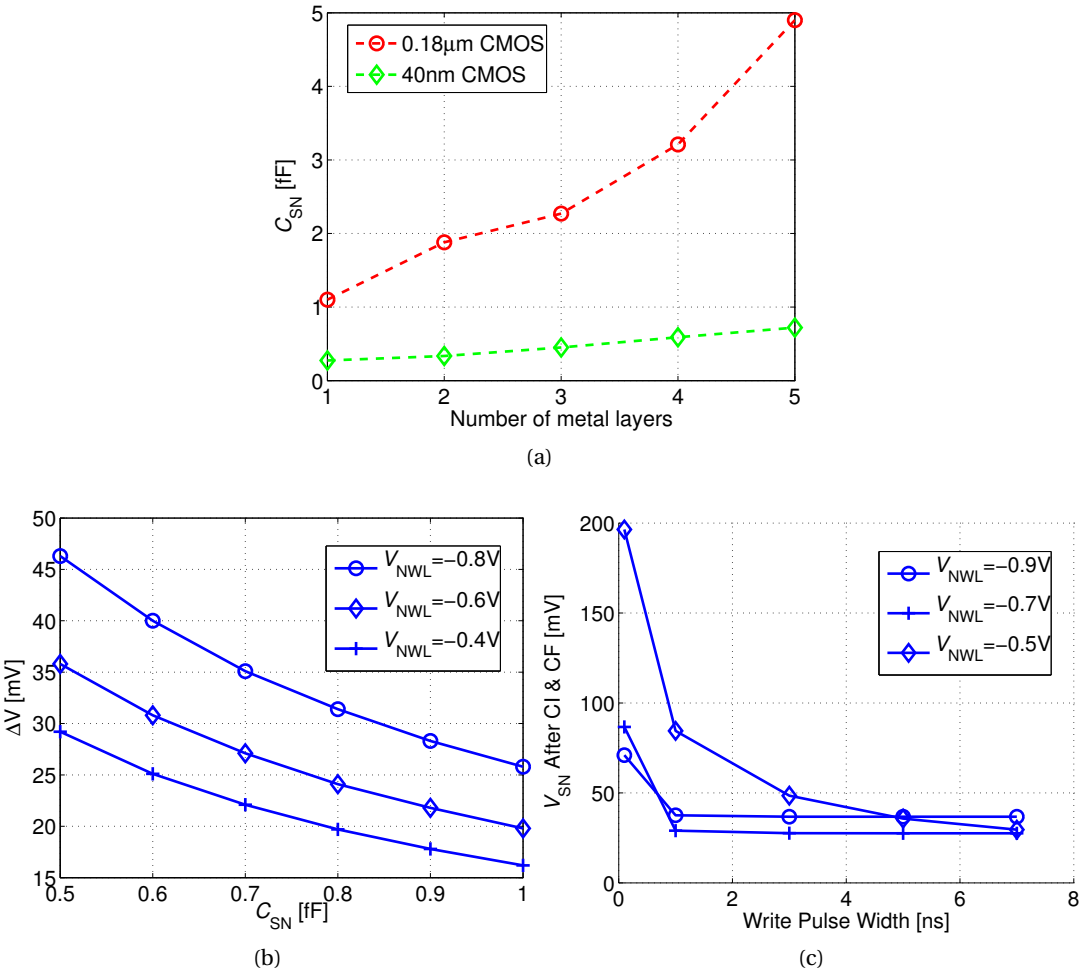


Figure 4.26: (a) Storage node capacitance versus number of employed metal layers. (b) ΔV due to CI and CF, as a function of C_{SN} and V_{NWL}, for V_{DD} = 700 mV. (c) V_{SN} after CI and CF versus write pulse width.

the positive edge of WWL for a logic '0' transfer. As already observed for the 0.18 µm node, ΔV decreases with increasing SN capacitance and with decreasing WWL step size (i.e., with decreasing absolute value of the underdrive voltage, V_{NWL}). While the charge injected from the large channel area of the selected I/O PMOS write transistor in the mature technology node results in a large voltage step severely threatening data integrity, the problem is slightly alleviated in more advanced nodes where small core transistors are preferred. The resulting voltage steps of 10 to 45 mV are rather small compared to the minimum V_{DD} where high array availability is achieved (as will be shown in Section 4.4.4). Moreover, it is worth mentioning that strong '0' levels are transferred to SN even with the least aggressive underdrive voltage of −0.4V (however, at the expense of write access time). Therefore, the ΔV values in Fig. 4.26b also correspond to the final SN voltage right after the write access. The final choice of V_{NWL} for the 40 nm node needs to account for the write access time, which must remain short to guarantee high array availability in a node with high leakage and short retention time (see Section 4.4.4).

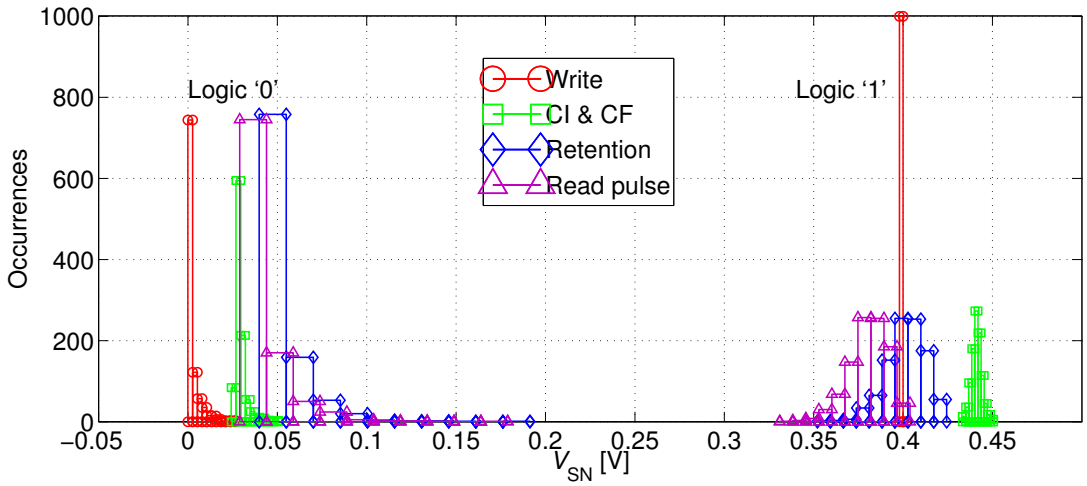


Figure 4.27: Distribution of the SN voltage of a logic '0' and a logic '1' at critical time points: 1) [circles] directly after a 1 µs write access (before turning off MW); 2) [squares] after turning off MW; 3) [diamonds] after a 40 ms retention period under worst-case WBL conditions; and 4) [triangles] during a read operation.

Therefore, Fig. 4.26c shows the final V_{SN} after CI and CF, as a function of the write pulse width. Over a large range of pulse widths as short as several ns, an underdrive voltage of −700 mV results in the strongest '0' levels, and is therefore preferred. Less underdrive, e.g., −500 mV, would result in weak '0' levels for pulse widths which are shorter than 3 ns.

4.4.3 Macrocell Implementation in 0.18 µm CMOS

This Section presents a 64×32 bit (2 kb) memory macro based on the previously elaborated 2T gain-cell configuration (Fig. 4.23b), implemented in a bulk CMOS 0.18 µm technology. The considered V_{DD} of 400 mV is clearly in the sub-V_T regime, as V_T of MW and MR are −720 mV and 430 mV, respectively. Special emphasis is put on the analysis of the reliability of sub-V_T operation under parametric variations. While the address decoders and the sense buffers are built from combinational CMOS gates and operate reliably in the sub-V_T domain [73], the analysis focuses on the write-ability, data retention, and read-ability of the gain-cell. All simulations assume a 1 µs write and read access time (1 MHz operation); a 3-metal SN capacitance of 2.5 fF, providing a retention time of 40 ms (according to previously presented worst-case estimation); a temperature of 37 °C and account for global and local parametric variations (1k-point Monte Carlo sampling).

Fig. 4.27 plots the distribution of the bitcell's SN voltage at critical time points for the '0' and the '1' states. As expected, nominal 0 V and 400 mV levels are passed to SN just before the positive edge of the write pulse. CI and CF cause the internal levels to rise by 20–50 mV, resulting in a slightly degraded '0' level and an enhanced '1' level, while the distributions remain sharp. After a 40 ms retention period with a worst-case opposite WBL voltage, the distributions are

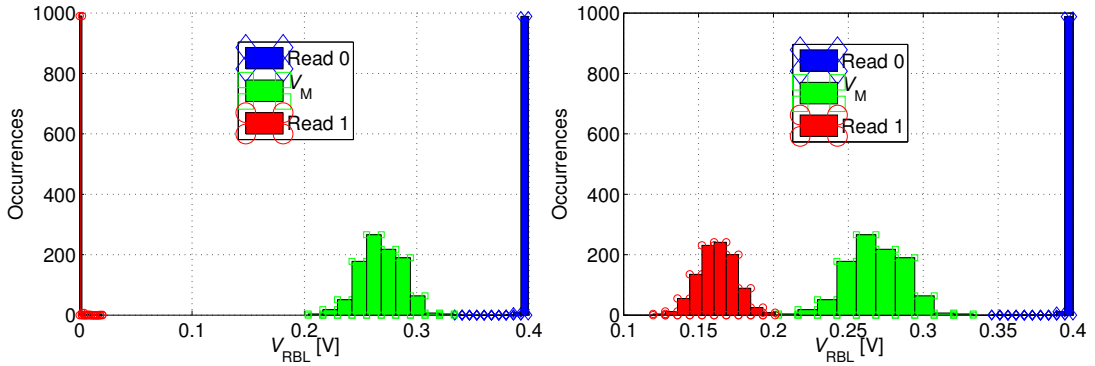


Figure 4.28: Distribution of RBL voltage (V_{RBL}) after read '1' [circles] and read '0' [diamonds] operations and distribution of the trip-point V_M of the read buffer [squares], for (a) favorable and (b) unfavorable read '1' conditions.

spread out, but the '1' levels are still strong, while the extreme cases of the '0' levels have severely depleted, approaching 200 mV. However, the '0' and '1' levels are still well separated, and moreover, the '0' levels are improved following the falling RWL transition, resulting in a 10–20 mV decrease.

To verify the read-ability of the bitcell, Fig. 4.28 shows the distribution of the RBL voltage (V_{RBL}) following read '0' and read '1' operations after the 40 ms retention period. In addition, the figure plots the distribution of the trip-point (V_M) of the sense buffer. While read '0' is robust in any case (RBL stays precharged), read '1' is most robust if all unselected cells on the same RBL as the selected cell store '0' (see Fig. 4.28a), while it becomes more critical if all unselected cells store '1' (see Fig. 4.28b), thereby inhibiting the discharge of RBL through the selected cell. This worst-case scenario for a read '1' operation is illustrated in Fig. 4.29a. In order to make the read operation more robust, V_M is shifted to a value higher than $V_{DD}/2$ by appropriate transistor sizing in the sense inverter. Ultimately, the V_{RBL} distributions for read '0' and read '1' are clearly separated, and the distribution of V_M is shown to comfortably fit between them, as shown in Fig. 4.28.

The layout of the 0.18 µm 2T gain-cell, comprising a PMOS I/O MW and an NMOS core MR is shown in Fig. 4.29b. The figure presents a zoomed-in view of one bitcell (surrounded by a dashed line) as part of an array. The chosen technology requires rather large design rules for the implementation of I/O devices; however, by sharing the n-well on three sides and stacking the bitlines, a reasonable area of 4.35 µm^2 per bitcell is achieved. In the same node, a single-ported 6T SRAM bitcell for above-V_T operation has a comparable area cost of 4.1 µm^2 (cell violates standard DRC rules), whereas SRAM bitcells optimized for robust operation at low voltages are clearly larger (e.g., the 14T SRAM bitcell in [89] has an area cost of 40 µm^2). The depicted layout also enables metal stacking above the storage node to provide an increased SN capacitance of up to 5 fF (see Fig. 4.26a).

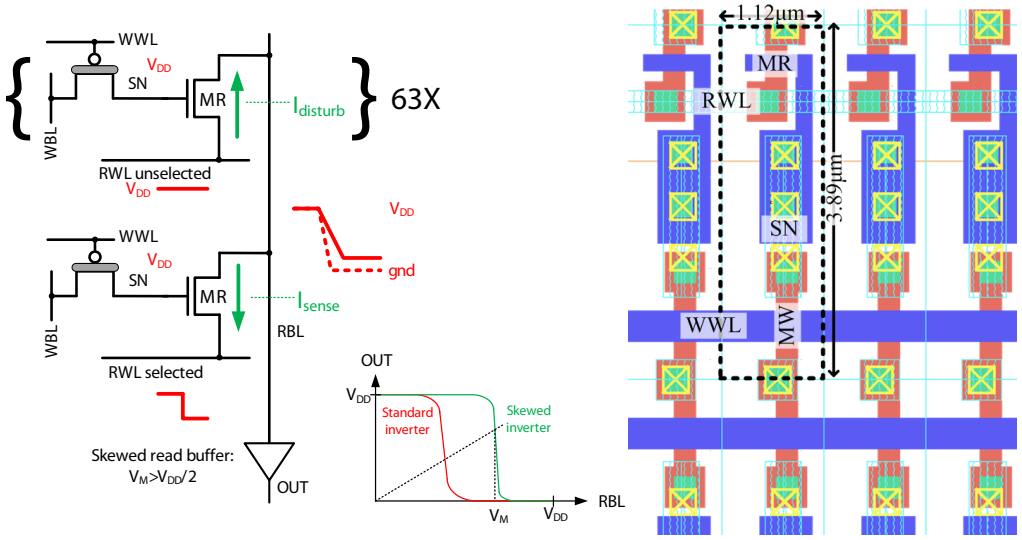


Figure 4.29: 180 nm gain-cell array: (a) Worst-case for read '1' operation: all cells in the same column store data '1'. To make the '1' operation more robust, the sense inverter is skewed, with a trip-point $V_M > V_{DD}/2$. (b) Zoomed-in layout.

At an operating frequency of 1 MHz, a full refresh cycle of 64 rows takes approximately 128 µs. With a worst-case 40 ms retention time, the resulting availability for write and read is 99.7%. As summarized in Table 4.4, the average leakage power of the 2kb array at room temperature (27 °C) is 1.95 nW, while the active refresh power of 1.68 nW is comparable, amounting to a total data retention power of 3.63 nW (or 1.7 pW/bit). This total data retention power is comparable to previous reports on low-voltage gain-cell arrays [130], given for room temperature as well.

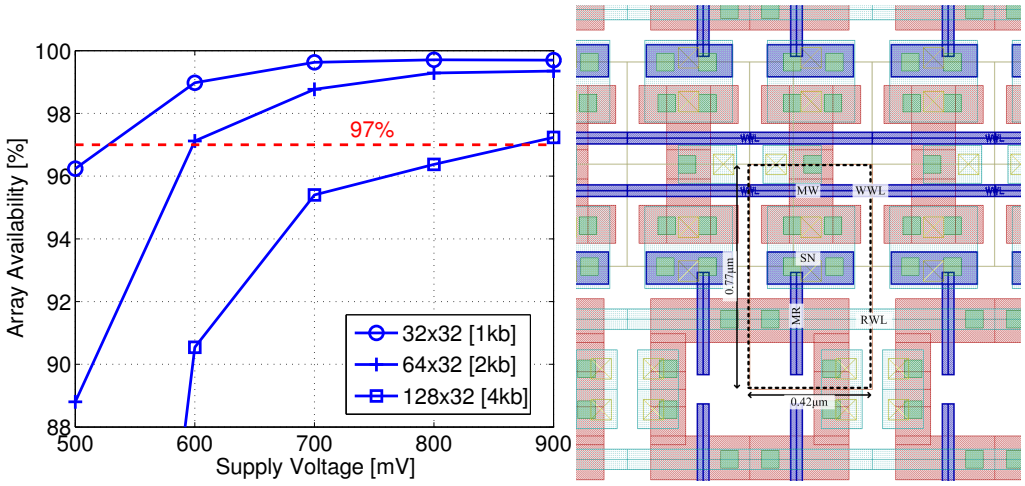
4.4.4 Macrocell Implementation in 40 nm CMOS

Whereas gain-cell implementations in mature technologies have been frequently demonstrated in the recent past, 65 nm CMOS is the most scaled technology in which gain-cells have been reported to date [29], as discussed in detail in Section 4.1. In this Section, for the first time, we present an 40 nm gain-cell implementation, and explore array sizes and the corresponding minimum operating voltages that result in sufficient array availability.

As previously described, core HVT devices are more efficient than I/O devices for write transistor implementation at scaled nodes, providing similar retention times with relaxed design rules (i.e., reduced area). In addition, the multiple threshold-voltage options for core transistors provide an interesting design space for the read transistor selection, trading off on- and off-currents, depending on the supply voltage. Two additional factors that significantly impact the design at scaled nodes are the reduced storage node capacitance, due to smaller cell area and low-k insulation materials, and severely impeded retention times, due to lower storage

Table 4.4: Figures of merit for 0.18 µCMOS and 40 nm CMOS ultra-low voltage GC-eDRAM macrocells.

Technology Node	**180 nm CMOS**	**40 nm LP CMOS**
Number of thin metal layers	5	5
Write Transistor	PMOS I/O	PMOS HVT
Read Transistor	NMOS Core	NMOS SVT
V_{DDmin}	400 mV	600 mV
Storage Node Capacitance	1.1 fF–4.9 fF	0.27 fF–0.72 fF
Bitcell Size	1.12 µm x 3.89 µm (4.35 µm^2)	0.77 µm x 0.42 µm (0.32 µm^2)
Array Size	64x32 (2 kb)	64 x 32 (2 kb)
Write Access Time	1 µs	3 ns
Read Access Time	1 µs	17 ns
Worst-Case Retention Time	40 ms	44 µs
Leakage Power	1.95 nW (952 fW/bit)	68.3 nW (33.4 pW/bit)
Average Active Refresh Energy	67 pJ	21.2 pJ
Average Active Refresh Power	1.68 nW (818 fW/bit)	482 nW (235.5 pW/bit)
Average Retention Power	3.63 nW (1.7 pW/bit)	551 nW (268.9 pW/bit)
Array Availability	99.7%	97.1%



(a) (b)

Figure 4.30: 40 nm gain-cell array: (a) array availability as a function of supply voltage and array size; and (b) zoomed-in layout.

capacitance and increasing leakage currents. Therefore, array availability becomes a major factor in gain-cell design and supply voltage selection. For this implementation, a minimum array availability of 97% was defined.

Considering a minimum array size of 1 kb (32x32), sufficient array availability is unattainable with the LVT MR implementation for a supply voltage lower than 500 mV, suitable for this device according to Fig. 4.24b. Therefore, an SVT device was considered with near-threshold

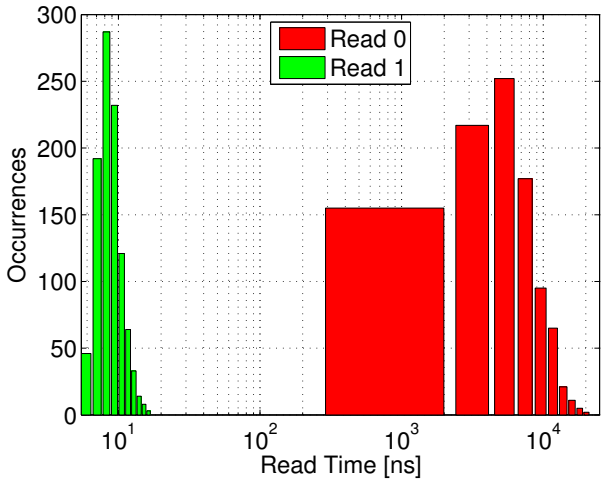


Figure 4.31: Read access time distribution for the GC-eDRAM implementation in 40 nm CMOS: RBL discharge time for correct data '1' sensing, and undesired RBL discharge time till sensing threshold through leakage for data '0'.

supply voltages above 500 mV. Fig. 4.30a shows the array availability achieved under varying supply voltages, considering array sizes from 1 kb to 4 kb. The red dashed line indicates the target availability of 97%, showing that this benchmark can be achieved with a 2 kb array at 600 mV. At this supply voltage, with a −700 mV underdrive write voltage, the write access time is 3 ns, and the worst-case read access time is 17 ns, while the worst-case retention time is 44 µs (see Table 4.4). Fig. 4.31 shows the distribution of the time required to sense the discharged voltage of RBL during a read '1' operation following a full retention period (green bars). The red bars (read '0') represent an incorrect readout, caused by a slow RBL discharge through leakage, such that the read access time must be shorter than the first occurrence of an incorrect read '0'. The clear separation between the two distributions shows that by setting the read access time to 17 ns, the system will be able to robustly differentiate between the two stored states.

A zoomed-in layout of the 40 nm gain-cell array is shown in Fig. 4.30b, with a bitcell area of 0.32 µm^2 (surrounded by the dashed line). For comparison, a single-ported 6T SRAM bitcell in the same node has a slightly larger silicon area of 0.572 µm^2, while robust low-voltage SRAM cells are considerably larger (e.g., the 9T SRAM bitcell in [4] has an area cost of 1.058 µm^2). As shown in Table 4.4, the implemented 40 nm array exhibits a leakage power of 68.3 nW, which is clearly higher than for the array in 0.18 µm CMOS technology. Even though the active energy for refreshing the entire array is only 21.2 pJ, the required refresh power of 482 nW is again higher than for the 0.18 µm node, due to the three orders-of-magnitude lower retention time. Consequently, the total data retention power is around 150× higher in 40 nm CMOS, compared to 0.18 µm CMOS.

4.4.5 Conclusions

This Section investigated 2-transistor (2T) sub-V_T and near-V_T gain-cell memories for use in ultra-low power systems, implemented in two very different technology generations. For mature, above-100 nm CMOS nodes, the main design goals of the bitcell are long retention time and high data integrity. In the considered 0.18 µm CMOS node, a low-leakage I/O PMOS write transistor and an extended storage node capacitance ensure a retention time of at least 40 ms. At low voltages, data integrity is severely threatened by charge injection and capacitive coupling from read and write word-lines. Therefore, the positive storage node (SN) voltage disturb at the culmination of a write operation is counteracted by a negative disturb at the onset of a read operation, which is only possible with an NMOS read transistor. Moreover, the write word-line underdrive voltage must be carefully engineered for proper level transfer at minimum voltage disturb during de-assertion. Monte Carlo simulations of an entire 2 kb memory array, operated at 1 MHz with a 400 mV sub-V_T supply voltage, confirm robust write and read operations under global and local variations, as well as a minimum retention time of 40 ms leading to 99.7% availability for read and write. The total data retention power is estimated as 3.63 nW/2kb, the leakage power and the active refresh power being comparable. The mixed gain-cell with a large I/O PMOS device has a large area cost of 4.35 µm^2, compared to an all-PMOS or all-NMOS solution with core devices only.

In more deeply scaled technologies, such as the considered 40 nm CMOS node, subthreshold conduction is still dominant at reduced supply voltages. Gate tunneling and GIDL currents are still small, but of increasing importance, while reverse-biased pn-junction leakage and edge-direct tunneling currents are negligible. In the 40 nm node, the write transistor is best implemented with an HVT core PMOS device, which provides the lowest aggregated leakage current from the storage node, even compared to the I/O PMOS device. A write word-line underdrive voltage of −700mV is employed to ensure strong '0' levels with a short write access time. Among various NMOS read transistor options, an SVT core device maximizes the sense current ratio between a weak '1' and a weak '0' for near-V_T supply voltages (600–800 mV) where 97% array availibility is achieved. Both the access times and the retention time are roughly three orders-of-magnitude shorter than in the 0.18 µm CMOS node, due to the increased leakage currents and smaller storage node capacitance. While the active refresh energy is low (21 pJ), the high refresh frequency results in high refresh power (482 nW), dominating the total data retention power (551 nW). As compared to the 0.18 µimplementation, the scaled down design provides better performance (17 ns read access and 3 ns write access), and a compact bitcell size of 0.32 µm^2.

To conclude, this analysis shows the feasibility of sub-V_T GC-eDRAM operation for mature process technologies and near-V_T operation for a deeply scaled 40 nm process, providing a design methodology for achieving minimum V_{DD} at these two very different nodes.

4.5 Multilevel GC-eDRAM (MLGC-eDRAM)

As discussed in detail in Section 2.3.1, there is a large number of VLSI systems which 1) require only short data retention times; and/or 2) are resilient to a small number of hardware defects, such as broken memory cells. Application fields for such VLSI systems include multimedia [51], wireless communications [2, 52, 21, 53], and data mining [54]. Note that beside these typical examples of systems which can tolerate some hardware defects, there is a general trend to error-resilient (or fault-tolerant) VLSI systems [55, 56] due to increased parametric variations and high defect levels in nanometric CMOS technologies. Unfortunately, random within-die process variations such as line edge roughness (LER) and random dopant fluctuations (RDFs) affect memory cells more than logic since the transistors are typically of minimum size in memory cells to satisfy high density requirements [56]. In order to strongly motivate and position our work on a multilevel GC-eDRAM array, presented in the following, we mention again the simulation-based analysis of a complete high speed-packet access (HSPA+) systems [21] (see Section 2.3.1 for some more details). In fact, this study [21] shows that the hybrid automatic repeat request (HARQ) memory, a major part of the entire system in terms of silicon area, can exhibit a bitcell failure rate of up to 1% while the HSPA+ system still achieves the required throughput. Furthermore, under preferential protection of the four most significant bits (MSBs) of the log-likelihood ratios in robust 8T SRAM bitcells, all remaining, less significant bits can be stored in highly unreliable bitcells exhibiting failure rates of up to 10% for a system-level throughput which is only slightly degraded compared to error-free hardware [21]. As a further example for the use of unreliable memories in fault-tolerant VLSI systems, we mention the work in [52], where the effect of unreliable storage of log-likelihood ratios on the performance of wireless communication transceivers is investigated. The system under consideration in [52] requires retention times below 10 µs and it is shown that error rates up to a few percent can be tolerated. These results encourage us to exploit innovative ways to compromise the reliability and the retention time of dynamic memories in general and of GC-eDRAM in particular for the benefit of increased storage densities.

While multilevel cells (MLC) [136] are extensively and industrially used in non-volatile Flash memory technology since several decades, only a small number of research works [137, 138, 139, 140] consider the possibility of storing more than one bit per cell in conventional 1T-1C eDRAM technology for increased storage density at the cost of compromised reliability and reduced retention times. The noise margin in an n-level multilevel DRAM (MLDRAM) is reduced by a factor of $1/(n-1)$ compared to the noise margin in a conventional single-bit-per-cell (two-level) DRAM [141] which implies that MLDRAMs are less reliable. Furthermore, the destructive read access of the conventional 1T-1C storage cell renders multilevel sensing a complex endeavor, particularly if sensing is to be done in a sequential manner to reduce the area overhead of the readout circuitry. Also the multilevel write and restore operations are rather complex; most MLDRAMs use charge sharing among ratioed or equal-sized capacitors, which typically are divisions of bitlines, to generate storage and reference levels [141].

For the first time, we apply the concept of storing many bits per memory cell to fully logic-

compatible GC-eDRAM technology. Besides the main advantage of logic compatibility, the non-destructive read access of gain-cells avoids the power-consuming restore operation and significantly simplifies the multilevel sense operation compared to conventional 1T-1C eDRAM bitcells: a stored data level can now sequentially be compared to several reference voltage levels. In the following, an 8-kbit multilevel GC-eDRAM macrocell, storing 2 bits per gain-cell, in 90 nm CMOS technology, including multilevel write and read circuits is proposed and analyzed with respect to its read failure probability due to within-die (WID) process variations by means of Monte Carlo simulations. With a view toward fault-tolerant VLSI signal processing systems, we investigate the dependency of the read failure probability on the *time upon write*, i.e., the time that passes between writing and reading back the data from the storage array. The results serve as a link to the area of fault-tolerant system performance analysis and design where the knowledge about the degree of data integrity for a given retention time can be taken into account.

Section 4.5.1 discusses the design of the multilevel GC as well as the corresponding multilevel read and write circuits. Section 4.5.2 discusses the failure mechanisms and studies the read failure probabilities under different operating conditions. Section 4.5.4 compares the area of the proposed memory macro with the one of an SRAM macrocell, and briefly presents a multilevel GC-eDRAM test chip. Section 4.5.3 proposes the use of a replica column in multilevel GC-eDRAM for fast memory access under varying PVT conditions. Finally, Section 4.5.5 concludes the work on multilevel GC-eDRAMs and presents an outlook.

4.5.1 Multilevel GC-eDRAM Design

As already discussed in Section 4.1, the basic idea behind GC-based memories is to store data in form of charge on a capacitive storage node (SN) formed by a MOSCAP (dedicated storage transistor MS), junction capacitance, as well as interconnect capacitance. In multilevel GC-based memories, many different voltage levels must be generated and transferred to the SN during the write operation. During the read operation, the transconductance gain of the ST is exploited to yield different sensing currents which can be compared to reference currents to yield a decision on the information stored in the cell. In summary, a multilevel GC-based memory comprises the following key components: an array of storage cells, a circuit for the generation of storage and reference levels, and a read circuit.

Multilevel gain-cell

In *single-bit-per-cell* storage arrays only an on- and an off-state of the storage transistor (MS), corresponding to two intervals of the SN voltage, must be distinguished. In our proposed *multilevel* GC, the drain current of MS is modulated by means of its gate voltage to distinguish between multiple levels during the read operation. To this end, the dynamic range of the voltage on the SN is partitioned into multiple non-overlapping regions corresponding to the individual symbols stored in a cell. This more fine-grained partitioning of the available

dynamic range of the SN voltage increases the sensitivity of the GC to leakage, which causes the SN voltage to drift, and therefore limits the retention time of the circuit. Furthermore, for multilevel-sensing smaller differences in sensing current must remain distinguishable compared to single-bit-per-cell storage arrays.

As a starting point for our multilevel GC implementation, we choose a conventional 3-transistor (3T) gain-cell [30] for reasons of its area efficiency compared to 4T GC topologies (using a "gated diode" or MOSCAP for increased SN capacitance and capacitive coupling from RWL to SN during read). The additional, separate read transistor (MR) compared to the more area-efficient 2T GC was chosen to avoid the masking issue[9] during read operation that was already critical in previous single-bit-per-cell implementations [28]. In order to simplify the multilevel read operation, the chosen 3T GC topology does not contain capacitive coupling from the read word-line (RWL) to the SN, a technique often being used in many single-bit-per-cell GC topologies to boost the SN voltage during read for larger sensing current and faster read.

The 3T multilevel GC topology can be implemented using different combinations of PMOS and NMOS devices. Clearly, an all-PMOS or an all-NMOS configuration yields the most compact cell layout. Unfortunately, the drawback of such a configuration is that the gate voltage of the write transistor (MW) must be boosted to be able to transmit the maximum available dynamic range for which the storage transistor MS is turned on to the SN during write operation in order to maximize the available margin between different levels. This implies the use of level shifters and a second power supply (or an embedded charge pump [88]) to generate the boosted write word-line (WWL) voltage. Furthermore, the correct functioning of the memory might be difficult to guarantee due to excessive gate tunneling and the long-term reliability might be compromised without a proper power-up sequence which ensures that the maximum voltage between the terminals of MW does never exceed the specifications of the technology.

To avoid the above described problems, we chose a configuration in which MW is implemented as PMOS transistor while MS and MR are implemented as NMOS transistors (vice versa would also be possible), as shown in the gray box in Fig. 4.32. The drawback of this solution is the area overhead required for the spacing between NMOS and PMOS devices. In our *mixed* GC configuration, this overhead is minimized by sharing the n-well on 3 sides between neighboring cells. Since the cell area is mostly limited by the contacts, the overall cell area increases only by a very small amount. As for the entire memory macro, requiring neither level shifters nor the generation of an additional boosted supply voltage, our mixed GC configuration results in much smaller overall area than the NMOS- or PMOS-only configuration.

[9] In area-efficient 2T gain-cells [28], the number of words which can be connected to the same read bit-line (RBL) is seriously limited, as the sum of the leakage currents drawn from the RBL by unselected cells quickly masks the sensing current of the selected cell to such an extent that the read operation fails. This problem is mitigated in 3T gain-cells (such as [121]) by adding a separate read transistor (MR) to the cell, at the price of a larger silicon area.

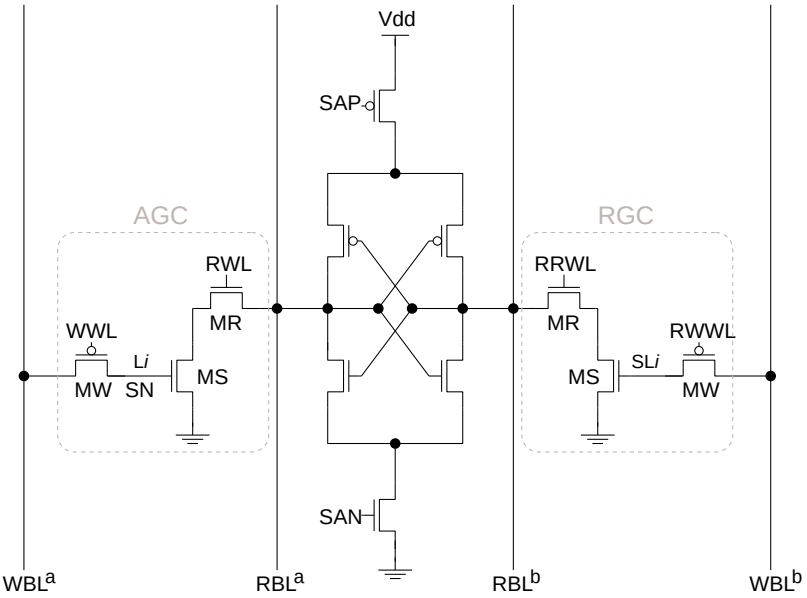


Figure 4.32: Sense amplifier connected to the gain cell being read and to the reference gain cell; the multilevel gain-cell topology is shown in the gray box.

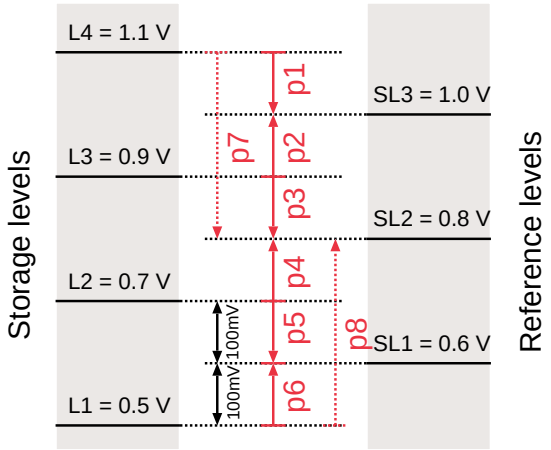


Figure 4.33: Allocation of storage and reference levels.

Level generation

Fig. 4.33 shows the 4 storage levels on the left-hand side and the 3 reference levels on the right-hand side which must be generated for storing and reading back 2 bits per cell. Note that at the end of the write operation, these voltage levels are slightly shifted (reduced by around 20 mV) due to charge injection and clock feedthrough from the PMOS write transistor.

In order to locally generate these levels within the macrocell, we follow the area-efficient approach proposed in [141, 140] by using charge sharing between bitline segments (sub-bitlines) which are precharged to either 0 V or to the supply voltage V_{DD} and then shorted

together. Fig. 4.34 shows one column of the memory macro and highlights the switches connecting two sub-bitlines. The resolution of this level generation technique is V_{DD}/W, where W denotes the number of words per WBL. One WBL cut must be performed for each different level to be generated, which results in $N+1$ sub-bitlines connected by N sub-bitline connectors (see Fig. 4.34) for N different levels.

Multilevel Sensing

As shown in Fig. 4.34, each column of the macro memory contains not only the actual GCs, but also a reference GC (RGC). The sense operation starts by writing a reference level to such a RGC in an unselected column of the storage array. Subsequently, the current drawn by the active GC (AGC), i.e., the GC being read, is compared to the current drawn by the RGC. To distinguish between multiple levels, one storage level must be compared to several reference levels. These comparisons can be done either sequentially [140] or in parallel [138]. For sequential 4-level sensing implementing a successive approximation, one storage level must be compared to two reference levels. As opposed to DRAMs based on the conventional 1T-1C cell, a storage level can easily be sensed multiple times in GC-based memories due to the non-destructive read access to the GCs. Using sequential rather than parallel multilevel sensing leverages this advantage to keep the area of the readout circuits small.

Fig. 4.32 shows the sense amplifier (SA) together with the AGC and the RGC. After storing the mid-range reference level (SL2 in Fig. 4.33) to the RGC, the RBL of the active and the reference column are precharged to V_{DD} and equalized by the bit line equalizer shown on the right-hand side of Fig. 4.34. The RWLs associated with the AGC and the RGC are then enabled at the same time which causes the RBLs to be discharged. Since the voltage levels stored in the GCs are different, the two RBLs are discharged unequally fast. The SA is triggered by the control logic after a short delay that is chosen long enough to allow for the development of a sufficient voltage difference between the two RBLs. The sense operation is then repeated with a second reference level that is chosen depending on the outcome of the first comparison.

4.5.2 Reliability/Failure Analysis

The dynamic storage mechanism combined with the reduced margin between the levels representing different symbols for the multilevel storage capability compromise the integrity of the data stored in the memory array. In the following, we presume a fault-tolerant application that can tolerate unreliable, but still mostly functional circuit behavior and we analyze the reliability of the proposed storage array for different operating conditions and process corners.

Read Failure Analysis

The two main reasons for not being able to read back the content of a memory cell correctly in the described storage array are:

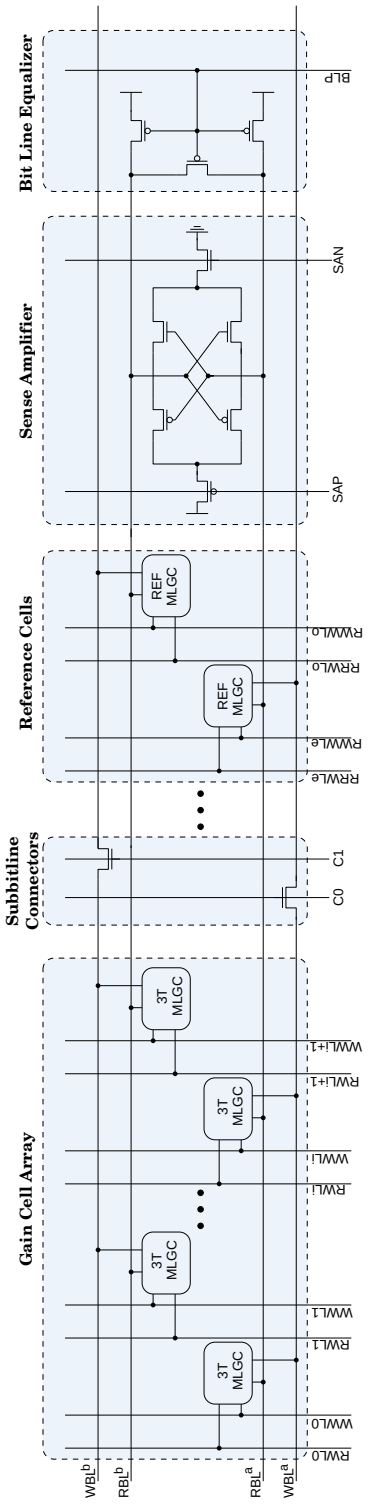


Figure 4.34: Multilevel GC-eDRAM macrocell architecture.

1. within-die (WID) process parameter variations that give rise to mismatch between the transistors on the active branch and the reference branch of the readout circuit (including the read port of the GC), and

2. the sum of leakage components from and to the SN which alters the voltage on the SN.

The second effect causes a shift of the SN voltage in the direction of one of the neighboring levels which reduces the sense margin that is available to compensate for process parameter variations. Hence, the percentage of errors due to process parameter variations depends on the time upon write which defines the time between the read operation and the last write operation to the corresponding multilevel GC.

Impact of Within-Die Process Variations We shall first investigate the impact of process parameter variations alone, without also explicitly considering the dependency of the error rate on the time upon write. To this end, we consider the voltage difference ΔV between the SNs of the AGC and the RGC as a parameter that we can set to emulate the voltage drift of the SN. A read failure can occur due to mismatch between the corresponding transistors in the active and the reference branches of the GCs and of the SA. The smaller ΔV, the higher the sensitivity of the sensing scheme to mismatch. For the GCs, the corresponding storage transistors (MS) as well as the corresponding read transistors (MR) should match, while in the SAs the NMOS (PMOS) transistors in the cross-coupled inverter pair should match (see Fig. 4.32). Transistors in the GCs are of minimum size and can be far apart. They can therefore hardly be matched and process parameters must be considered to be independent and identically distributed (i.i.d.). The opposite is true for transistors in the SA which can be placed in close proximity to each other and can be sized generously to improve matching. Nevertheless, i.i.d. process variations between the AGCs and the RGC as well as within the SAs are considered for all following analyses.

We evaluate the failure probabilities using Monte Carlo circuit simulations with back-annotation of all relevant layout parasitics in a 90 nm CMOS node. Depending on the level being stored in the AGC and depending on the state of the successive approximation algorithm (first or second comparison), 8 sense operations, labeled p1 … p8, are distinguished as shown in Fig. 4.33. The sense operations p7 and p8 have a much greater margin than the other sense operations (p1 to p6). We can therefore limit the analysis of the read failure probability to the sense operations p1 to p6. Fig. 4.35 shows the corresponding empirical failure probabilities p_{fail} for 1000 within-die process parameter realizations under *worst*-case conditions, corresponding to the fast-fast process corner at 85 °C. As expected, the read failure probabilities increase as the margin ΔV decreases and reach 50 % for $\Delta V = 0\,\text{V}$. We also observe that the failure probabilities depend mostly on ΔV and not much on the absolute SN voltage levels and are thus very similar for the six relevant sense operations.

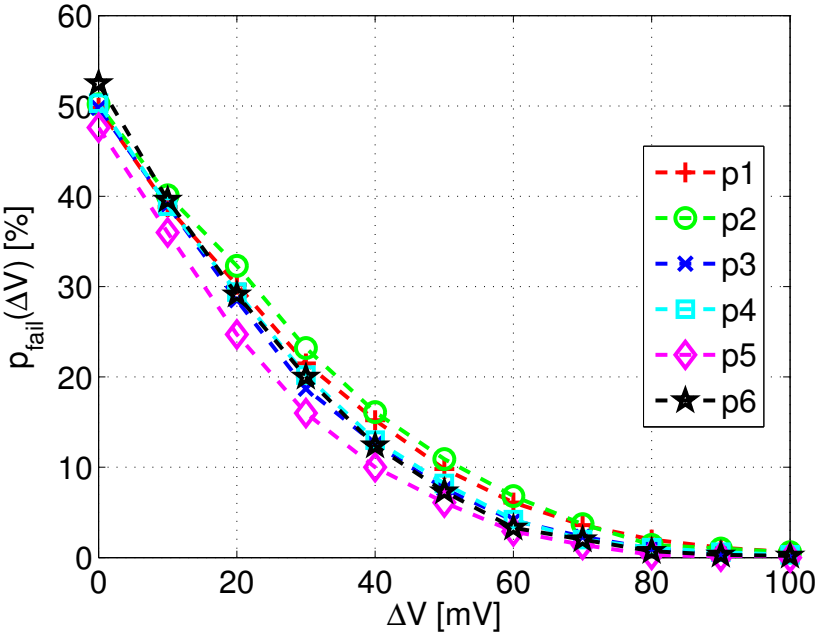


Figure 4.35: Read failure probability p_{fail} as a function of ΔV under *worst*-case conditions (defined in Table 4.5).

Table 4.5: Definition of operating conditions.

Name	Corner	Temp.	WBL state
Worst	ff	85 °C	Opposite
Bad	ff	85 °C	Middle
Typical	tt	25 °C	Middle

Impact of Time Upon Write t_{w} As discussed previously, ΔV for a particular sense operation can change over time due to leakage from and to the SN. This effect is negligible for the RGC which is set immediately before the read operation, but the time upon write t_{w} needs to be taken into account to determine the SN voltage of the AGC during the read operation.

Fig. 4.36 shows the sensing failure probabilities $p_{\text{fail}}(t_{\text{w}})$ as a function of t_{w}, again obtained through Monte Carlo simulations, for the fast-fast process corner at 85 °C. For each sense operation we have constructed a worst-case scenario that keeps the WBL constantly at a level that maximizes the subthreshold current of the MW pulling the SN voltage of the AGC toward the reference level of the respective sense operation. We observe from Fig. 4.36 that the sense operations p1, p3, and p5 are less likely to fail than p2, p4, and p6. The reason for this difference is that for the more reliable sense operations (p1, p3, and p5), the gate-induced drain leakage (GIDL) current of MW charges the SN, while the subthreshold current of MW discharges the SN. For the less reliable sense operations (p2, p4, and p6) both the GIDL current and the subthreshold current of MW charge the SN. The worst situation occurs for p6 due to the largest drain-to-source voltage of MW.

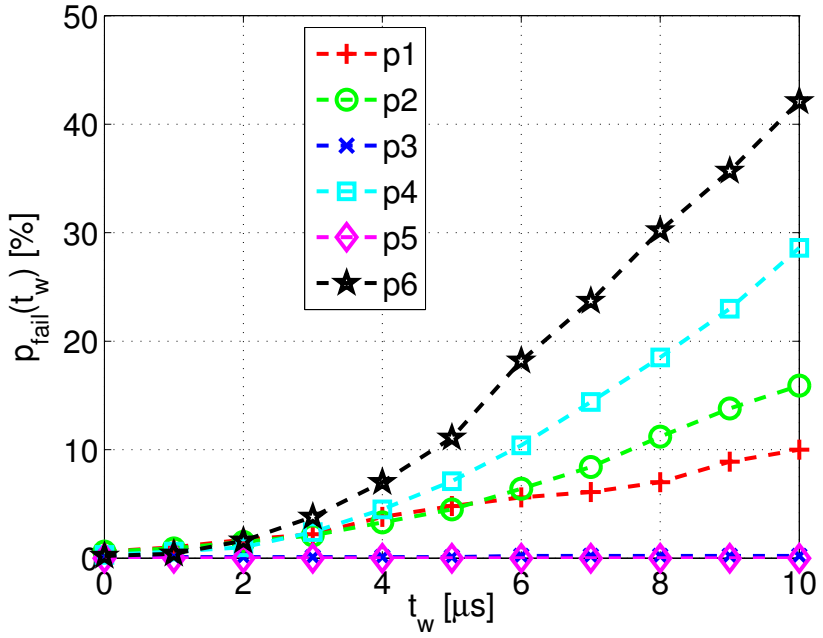


Figure 4.36: Read failure probability $p_{\text{fail}}(t_{\text{w}})$ as a function of the time upon write t_w under *worst*-case conditions (defined in Table 4.5).

For practical systems, the above worst-case assumption on the state of the WBL is highly unrealistic. In fact, during the idle state of the memory, the voltage on the WBL can be controlled and can be kept in the middle of its dynamic range. As can be seen in Fig. 4.37, $p_{\text{fail}}(t_{\text{w}})$ decreases significantly under this new assumption, as the subthreshold conduction of MW is smaller. Fig. 4.37 also shows that now the highest failure probabilities occur for the sense operations p1 and p6 due to the largest drain-to-source voltage values of MW. The sense operation p6 has a smaller failure probability than p1 as the PMOS MW has higher gate-to-source and gate-to-drain voltages and thus a smaller subthreshold current.

Keeping the same assumption on the WBL state, and for the typical-typical process corner at 25 °C, the maximum read failure probability among all possible sense operations 10 µs (50 µs) after writting is 1.7 % (7.9 %), as shown in Fig. 4.38.

So far, the read failure probabilities of single sense operations has been analyzed. Thereof, the failure probabilities of two successive sense operations corresponding to the detection of a storage level can be deduced. For *typical* operating conditions, the high failure probability of the sense operation p6 suggests using only 3 levels per cell. Thus, in order to reach higher reliability at the price of larger area, coding over two 3-level cells could be used, so that $3^2 = 9 > 8$ symbols are available using both cells, which corresponds to $\frac{1}{2}\log_2 8 = 1.5$ bits per cell if only 8 out of 9 symbols are used [142].

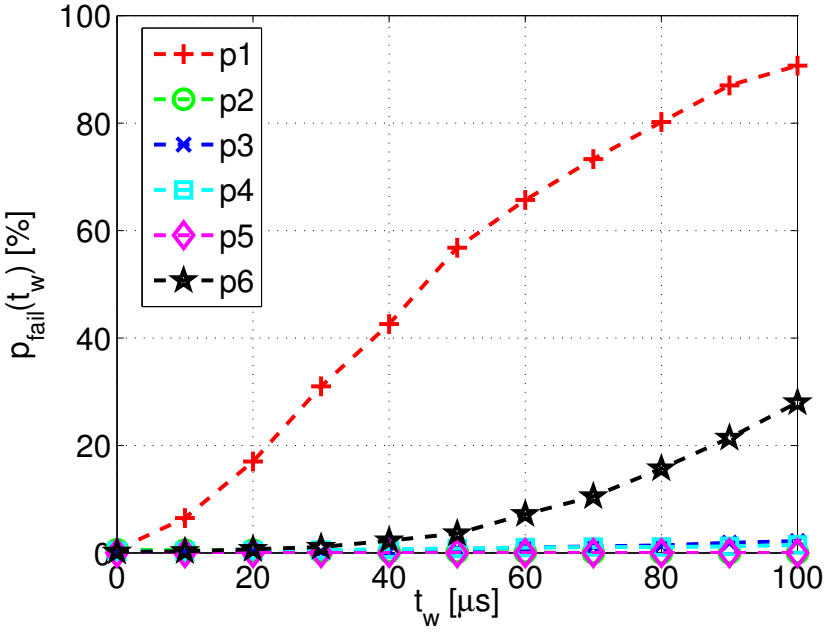


Figure 4.37: Read failure probability $p_{\text{fail}}(t_{\text{w}})$ as a function of the time upon write t_w under *bad* conditions (defined in Table 4.5).

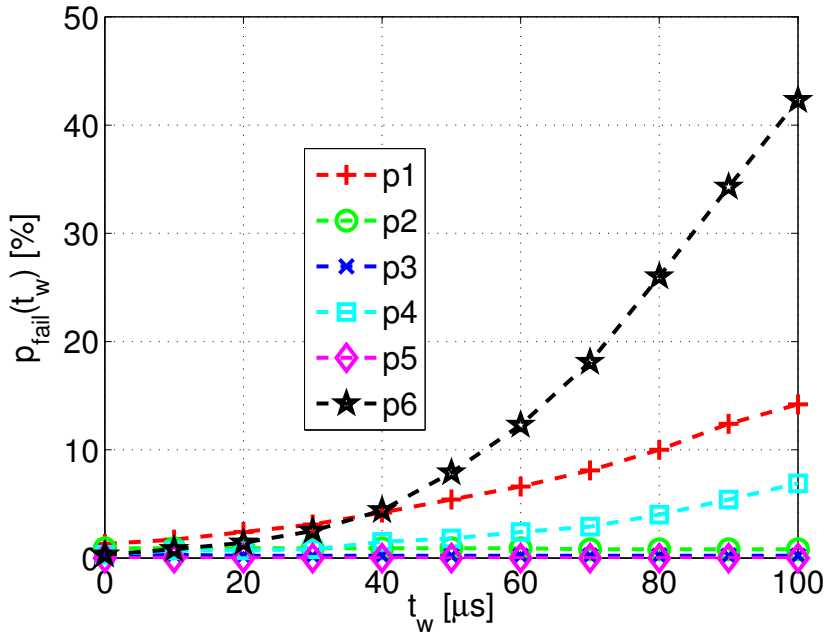


Figure 4.38: Read failure probability $p_{\text{fail}}(t_{\text{w}})$ as a function of the time upon write t_w under *typical* conditions (defined in Table 4.5).

4.5.3 Replica Techniques for Frequency Guardband Reduction

As seen before, among different multilevel write schemes summarized in [143], charge sharing between bit-line (BL) segments to locally generate many data levels has a small area cost.

Moreover, remember that the multilevel read operation is best performed in a sequential fashion to avoid an area-increase due to the need for many parallel sense amplifiers (SAs). However, these area-efficient multilevel write and read schemes result in long access times. This problem is aggravated in deep-submicron (DSM) and nanometric CMOS technologies where large timing margins are required due to increasing process variations if reliability is not to be further compromised.

In order to guarantee reliable sense operation and to yet trigger the SA at the earliest possible instant, even in the occurrence of large die-to-die (D2D) process, voltage, and temperature (PVT) variations, different flavors of replica BL techniques have been developed for SRAMs [144, 145, 146]. Some of these techniques do also address WID process parameter variation [145, 146]. The basic replica BL technique consists of a delay generator (the replica BL) which tracks the delay of the actual BLs across PVT corners [144]. To our knowledge, the replica BL technique has not been exploited yet to improve the access times of multilevel GC-eDRAMs.

In our simulation-based study [112], the replica BL technique is applied to the previously presented multilevel GC-eDRAM to maintain optimum read access times under PVT variations with a minimum area-overhead. In addition to generating read control signals, the same replica column is also used to generate write control signals with optimum delay. As for the multilevel write operation, the delay to pre-(dis)charge the capacitive WBL segments is the most significant contribution to the write access time. For the generation of the highest storage level of 1.1 V, 11 WBL segments need to be pre-charged to V_{DD}, which amounts for the longest possible pre-(dis)charge delay. The replica column which is added to the storage array is designed to track exactly this pre-(dis)charge delay, in order to optimally time the initiation of the charge sharing process and the assertion of the write word-line (WWL) for a successful write completion.

As for the multilevel read operation, the different voltage levels on the SN of the AGC and the RGC result in unequally strong RBL discharging currents, which eventually develops a voltage difference between the terminals of the SA, as expatiated on before. The SA is triggered as soon as this voltage difference is big enough to overcome its offset voltage. It is crucial to trigger the SA at the right time: triggered too early, the voltage difference might be too small to be resolved correctly; triggered too late, both RBLs might already have been discharged completely to ground. Finding a suitable trigger instant is especially difficult since there are many different voltage levels resulting in stronger or weaker discharging currents. The problem is further aggravated by PVT variations. Implemented in a 90-nm CMOS technology, the SA shown in Fig. 4.32 has an offset voltage of up to 30 mV. The RBL-discharge delay to be tracked by the replica column is defined as the required time to discharge a RBL from V_{DD} to $0.45 \times V_{DD}$ through the read path of a gain-cell storing the highest data level. For the highest data and reference levels, a voltage difference between the RBLs of 107 mV is developed within this delay, whereas the voltage difference that develops for the lowest data and reference levels is 71 mV and still high enough for reliable sensing. Consequently, the replica column, tracking

Table 4.6: Total access times for different PVT conditions.

PVT condition	t_{write} [ns]	t_{read} [ns]
Fast	2.3	9.2
Typical	3.0	12.0
Slow	5.0	20.0

the RBL discharge delay, allows for triggering the SAs at an early yet safe instant for any PVT condition. More details on the replica column design and the control signal generation can be found in [112].

Extensive simulation results presented in [112] demonstrate the effectiveness of the proposed replica column in firmly tracking the WBL segment pre-(dis)charge delay as well as the RBL discharge delay over large P(D2D)VT variations and that the small remaining timing margins of 50−100 ps are sufficient to cope with WID variations. Table 4.6 shows the total write access time t_{write} and the total read access time t_{read}, including the time required for address decoding, for fast (FF, 1.32 V, 10 °C), typical (TT, 1.20 V, 27 °C), and slow (SS, 1.08 V, 80 °C) PVT conditions. Note that $t_{read} = 4 \times t_{write}$, as a read access consists of two write accesses to a reference gain-cell (1 clock cycle each), each write access being followed by a sense operation (1 clock cycle each). The implemented replica BL technique provides savings in the write access time of 2.7 ns for fast PVT conditions, and 2.0 ns for typical PVT conditions, compared to a design with fixed timing margins which guarantee accurate level generation even for slow PVT conditions. Similarly, the savings in read access time are 10.8 ns and 8.0 ns for fast and typical PVT conditions, respectively. More importantly, the replica BL technique is much safer than using fixed timing margins, as it finds an appropriate SA trigger instant for each PVT condition.

4.5.4 Implementation Results

The implemented multilevel GC-eDRAM macrocell shown on the right-hand side of Fig. 4.39 has a storage capacity of 8 192 bits. With an area of $86 \times 138.1\,\mu m^2 = 11\,877\,\mu m^2$, the proposed 4-level GC-based macro memory is only 54.8 % the size of a corresponding commercially available single-port SRAM macrocell ($152.6 \times 141.9\,\mu m^2 = 21\,654\,\mu m^2$) with the same storage capacity (see Fig. 4.39 left-hand side), even though the SRAM macrocell has pushed DRC rules, i.e., it contains smaller than minimum-size features (e.g, narrower contacts) and also violates other design rules (e.g., minimum diffusion enclosure of contact, minimum poly to diffusion spacing) for higher density.

The presented multilevel GC-eDRAM macrocell was manufactured in a 90 nm CMOS node. The layout picture and the chip microphotograph are shown in Fig. 4.40. The test chip also contains a further multilevel GC-eDRAM macrocell based on an all-NMOS 3T gain-cell [147],

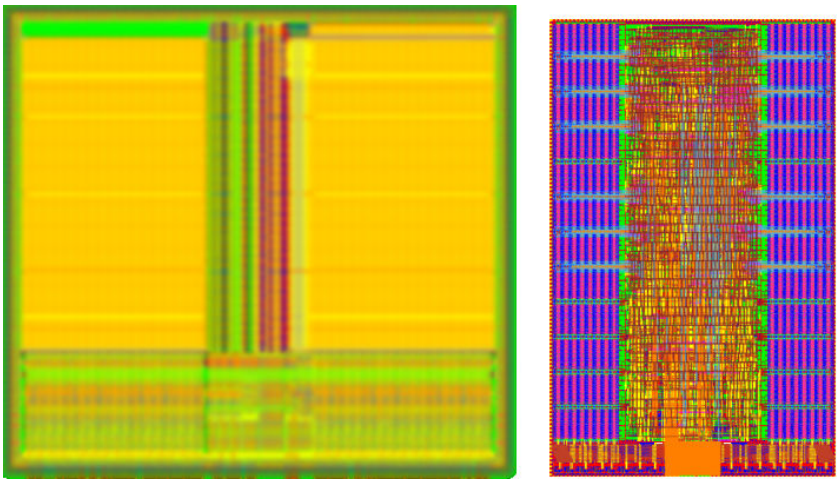


Figure 4.39: Commercially available SRAM macrocell (left) and proposed multilevel GC-eDRAM macrocell (right).

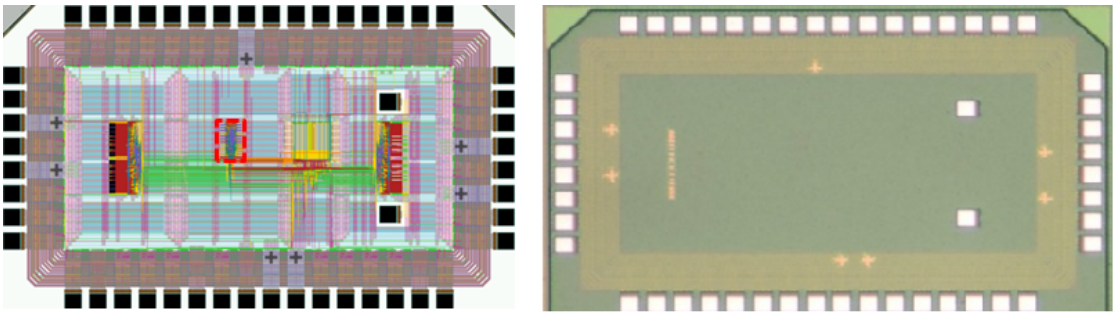


Figure 4.40: Layout picture (left) and microphotograph (right) of multilevel GC-eDRAM test chip; the multilevel GC-eDRAM macrocell described in this section is highlighted by a dashed red line in the layout picture.

a test array where the storage and reference voltages can be controlled externally, and a single-port SRAM macrocell for reference measurements. Unfortunately, while the above presented multilevel GC-eDRAM macrocell was designed to store 4 levels per bitcell, silicon measurements unveil that only 3 distinct levels can be read, even for a short time upon write t_w. This degraded storage density compared to pre-silicon expectations can be attributed to a combination of the following factors which could hardly be addressed by simulation: 1) variations in the bit-line segment capacitances leading to inaccurate storage and reference levels; 2) noise on the supply rail (V_{DD}); 3) cross-coupling effects between the bit-lines (BLs); and 4) more mismatch than in the Monte Carlo (MC) models. By using charge sharing among already existing BL segments for level generation and using minimum-size reference gain-cells for sensing, our multilevel GC-eDRAM design aimed at small overhead for the design of peripheral circuits. We conclude that in order to demonstrate storage of 4 (instead of 3) voltage levels per GC in a 90 nm CMOS node, it is necessary to employ a more robust level generation technique (accepting a larger overhead) and avoid variations in the reference currents used

for sensing (again, accepting a more expensive circuit).

4.5.5 Conclusions and Outlook

The concept of storing many bits per basic memory cell has been applied to fully logic-compatible gain cells, in order to trade reliability and retention time for higher storage density in future error-resilient (fault-tolerant) VLSI systems. An 8-kbit macro memory including multilevel write and read circuits was presented and analyzed regarding its failure mechanisms. The read failure probability at a given time upon write was shown to depend quite heavily on the state of the write bit-lines (WBLs) and is significantly decreased if the WBLs are kept in the middle rather than on either side of their dynamic range during the idle state of the memory. Under typical operating conditions, the maximum, simulated failure probability among all possible sense operations 10 µs (50 µs) after writting is less than 2 % (8 %), which can be tolerated by some fault-tolerant VLSI systems. The area of the proposed macro memory is only 54 % of the area of a commercially available single-port SRAM macrocell of equal storage capacity. The use of a replica column is encouraged for the generation of write and read control signals with optimum timing under varying process-voltage-temperature (PVT) conditions. The proposed replica technique significantly improves both the write and read access times of the otherwise rather slow and complex multilevel write and read operations.

Our analyses (including many post-layout circuit simulations and some silicon measurements) showed that multilevel GC-eDRAM is an interesting concept to enhance the storage density in error-resilient systems implemented in rather mature CMOS nodes (above 100 nm CMOS). Also, the concept is especially interesting for large storage arrays in order to justify the overhead for large and accurate voltage and reference current generation circuits; the low-overhead circuits used in our 90 nm CMOS test chip were only accurate enough to store and retrieve 3 (instead of 4) levels per cell. For sub-90 nm CMOS nodes with increased parametric variations and leakage currents, multilevel GC-eDRAMs are not a viable option to ensure read failure rates below 10 % after “retention” times of 50 µs, and the use of our more robust single-bit-per-cell GC-eDRAMs presented in the previous Sections is recommended instead for such deeply scaled nodes.

5 Conclusions

This PhD thesis has proposed various alternatives to SRAM macrocells for the implementation of embedded memories in VLSI SoCs, namely several novel and innovative implementations of standard-cell based memories (SCMs) and gain-cell based embedded DRAMs (GC-eDRAMs). Many prototype chips have been designed and manufactured, in a variety of CMOS nodes (mostly 180 nm, 90 nm, and 65 nm), in order to verify the various novel area- and energy-efficient memory technologies and circuit techniques by means of silicon measurements. Our innovative types of embedded memories have been specifically optimized to meet the requirements of a large range of completely diverse VLSI SoCs, from ultra-low power systems operated at subthreshold (sub-V_T) voltages all the way to high-performance, power-aware, potentially error-resilient systems. In addition, the proposed memory designs often exploit properties of the target system (such as frequent write updates) in order to improve a given metric which is of particular interest and concern at the system level (such as speed or storage density). For example, on the one hand, highly robust circuit operation and an extremely low power budget are the main challenges to be addressed for the design of ultra-low power systems which often find applications in the biomedical domain (biomedical implants); in this thesis, a straightforward, ultra-low power, sub-V_T SCM compilation flow has been proposed, and the feasibility of sub-V_T GC-eDRAM has been demonstrated, as well. The sub-V_T SCM designs achieve extremely low leakage power by relying on a custom-designed ultra-low leakage standard-cell whose design exploits low speed requirements imposed by the system. On the other hand, high-performance, power-aware VLSI systems, some of which require only short data retention times and can tolerate a few failing memory bitcells, with applications in domains such as wireless communications or multimedia, require high speed and preferably small silicon area (for low cost) as primary design goals; in this thesis, we have proposed to integrate high-density, fast, dynamic latches into SCMs, and have investigated for the first time the feasibility of multilevel GC-eDRAM.

5.1 Standard-Cell Based Memories (SCMs)

The studies in this thesis show that SCMs have several advantages compared to SRAM macrocells, including straightforward implementation and robust operation in any system and at any supply voltage, simple portability among technology nodes, modifications at design time, automatic placement and ability to merge storage with logic (where appropriate) for less routing and less switching power, lack of separate voltage supply rings, and high flexibility for fine-granular memory organizations with clock-gating (and power-gating) for reconfigurable VLSI systems. If avoiding the burden of custom-designed standard-cells for short design times and maximum portability, irrespective of the supply voltage, technology node, fab, and standard-cell library provider, the best-practice SCM architecture uses latches as storage cells (rather than flip-flops), clock-gates for the generation of write select pulses (rather than enable flip-flops/latches), and multiplexers on the readout path (rather than tri-state inverters). If custom design is affordable, targeting ultra-low power (ULP) VLSI systems, a latch topology based on tri-state inverters (instead of inverter and transmission gate), stack forcing in all inverters, channel length stretching, and a tri-state output inverter (for a tri-state read logic implementation) leads to ultra-low leakage power and access energy (primary concerns for ULP system design) at the cost of a degraded read access time (secondary concern for most ULP systems). Targeting high-performance VLSI systems with high density and short retention time requirements, the use of dynamic SCMs (D-SCMs) dramatically reduces the area cost compared to static SCMs; a 3-transistor dynamic latch is seamlessly merged with the first stage of the read multiplexer (a NAND gate) in a single standard-cell for maximum area-efficiency. In addition, the custom-designed, dynamic standard-cell is optimized to avoid short-circuit currents, as well.

SCMs synthesized using exclusively commercially available standard-cell libraries, i.e., only static latches and flip-flops, can be smaller than corresponding SRAM macrocells for storage capacities up to 1 kbit. If employing a custom-designed, robust, 8-transistor (8T) dynamic latch topology instead of a static latch as basic storage cell, this border for which SCMs are still smaller than SRAM macrocells moves up to around 2 kbit. A 3-transistor (3T) dynamic latch can clearly be smaller than a 6-transistor (6T) SRAM bitcell, enabling smaller SCMs as compared to SRAM macrocells irrespective of the storage capacity; however, it is challenging to integrate such a 3T dynamic latch into a standard digital design flow, and the reliability and retention time are degrated compared to the 8T dynamic latch topology.

A low-power low-density parity-check (LDPC) decoder was used as a case study to demonstrate the advantages and potential drawbacks of SCMs. In fact, replacing all SRAM macrocells in a baseline LDPC decoder design with static SCMs was shown to reduce the decoder's power consumption by 37% while the area cost increased by 50%. Since all internal memories of this LDPC decoder architecture are updated with new data periodically and frequently, it is possible to use refresh-free, dynamic SCMs (D-SCMs) for high storage density. In fact, a custom-designed, multi-functional dynamic storage and NAND gate entails a 70% reduction in silicon area compared to an implementation based on commercial standard-cells, which

results in a 44% area reduction at the decoder level. Both leakage and active power are also reduced thanks to the D-SCMs, as compared to static SCMs; in fact, short-circuit currents are systematically avoided at all time by circuit optimizations.

Unfortunately, 6T-bitcell SRAM macrocells typically obtained from commercial memory compilers do not work reliably at scaled voltages. Alternative 8T, 10T, ..., 14T SRAM bitcells, sometimes in conjunction with low-voltage write and read assist techniques are required to guarantee reliable circuit operation at aggressively scaled supply voltages, sometimes residing in the subthreshold (sub-V_T) domain. Due to the lack of good compilers for such robust sub-V_T SRAMs, the use of the proposed sub-V_T SCM compilation flow is highly interesting, especially for ultra-low power (ULP)/ultra-low voltage (ULV) systems requiring only a small storage capacity (per memory block) of several kb. In fact, sub-V_T SCMs synthesized exclusively from commercial standard-cell libraries operate reliably at sub-V_T voltages, and have short access times and good energy efficiency compared to corresponding full-custom sub-V_T SRAM macrocells. Unlike modified bitcells and low-voltage assist circuits for sub-V_T SRAM macrocells which often impede the performance at nominal voltage, the best-practice sub-V_T SCM topology is also the preferable choice for above-V_T operation, thereby supporting dynamic voltage and frequency scaling (DVFS) while keeping the optimum circuit topology.

Most of the energy in sub-V_T SCMs is consumed due to leakage currents while active energy plays only a minor role, especially for large configurations. Therefore, in order to improve the access energy and the leakage power, the design of custom standard-cells focuses on leakage reduction. In fact, opting for a tri-state read logic instead of the otherwise preferred CMOS multiplexers, all major leakage contributors of sub-V_T SCMs can be addressed by designing a single low-leakage standard-cell, namely a D latch with tri-state inverters, transistor stacks, channel length stretching till the point of diminishing returns, and a tri-state output inverter. Silicon measurements of a 4 kb sub-V_T SCM manufactured in 65 nm CMOS show that the leakage power and the access energy are cut into half compared to SCMs synthesized from commercial libraries only. Moreover, we reported the lowest access energy and leakage power per bit to date among all silicon-proven sub-V_T SRAMs in 65 nm CMOS technology.

For the first time, we have proposed a ReRAM-based non-volatile flip-flop (NVFF) which reliably operates in the sub-V_T domain (except for the ReRAM write operation which requires a CMOS-compatible voltage). The proposed NVFF circuit operates reliably in the sub-V_T domain even in the occurrence of parametric variations in the ReRAM device and MOS transistors. The energy for an active-to-sleep transition (i.e., for a ReRAM write operation) is relatively high due to the high voltage, while the energy for a sleep-to-active transition (i.e., for a ReRAM read operation) is successfully reduced thanks to the wake-up at a sub-V_T voltage. With the currently used oxide stack (OxRAM device) resulting in a large write energy, the break-even time for net energy savings compared to the retentive, low-leakage, 500 fW latch is relatively long (1.47 s).

5.2 Gain-Cell Based eDRAMs (GC-eDRAMs)

Gain-cell based eDRAM (GC-eDRAM) combines most of the advantages of SRAM and conventional 1T-1C eDRAM and avoids most of their respective drawbacks, making it an attractive option for the implementation of embedded memories. In fact, gain-cells are much smaller than SRAM bitcells (typically by 50%), they exhibit a much lower bitcell leakage current than SRAM bitcells, they are fully compatible with standard digital CMOS technologies (like SRAM, and unlike 1T-1C eDRAM requiring extra process steps and additional costs to build high-density 3D capacitors), they allow for non-destructive read access and can avoid power-hungry restore (write-back) operations (as opposed to 1T-1C eDRAM), and they have a separate read and write port (unlike conventional 6T SRAM and 1T-1C eDRAM) which allows to simultaneously and independently optimize the bitcell for high read and write robustness and allows for low-overhead two-port memory macrocell implementations with high access bandwidth. The main drawback of GC-eDRAM compared to 1T-1C eDRAM is the lower in-cell storage capacitor which can be built using exclusively MOSCAPs, junction capacitances, and interconnect capacitances available in a digital CMOS process, as compared to the dedicated trench or stacked DRAM capacitors; this typically results in lower data retention times and more frequent, power-consuming refresh operations.

While almost all previous works on GC-eDRAM were targeting large cache memories for high-end microprocessors, this thesis extends the application range of GC-eDRAM to low-voltage/low-power VLSI SoCs (such as biomedical implants or sensor networks) and to error-resilient VLSI systems (such as many wireless communications systems). In particular, we have pioneered the field of low-voltage operation for GC-eDRAMs, exploiting near-threshold (near-V_T) and even subthreshold (sub-V_T) circuit operation for low leakage power and low access energy, as well as voltage-compatibility with and integration into ultra-low voltage (ULV)/ultra-low power (ULP) VLSI systems. Specifically, the predominant drawback of GC-eDRAM, i.e., the rather low retention times, has been alleviated within this thesis by a number of innovative techniques, all applied to near-V_T GC-eDRAM arrays: 1) first of all, counter to intuition, it has been shown that the retention time can be improved by means of voltage scaling if the write bit-lines (WBLs) can be freely controlled to a desired voltage in case of unfrequent write accesses; 2) second, silicon measurements have shown, for the first time, the high impact which reverse body biasing can have to improve the retention time of GC-eDRAM; and 3) further silicon measurements verified the effectiveness of our proposed replica technique to find the optimum refresh timing, avoiding unnecessary power consumption due to early refresh triggering, across varying process-voltage-temperature (PVT) conditions and for varying write-access disturb frequencies (which degrade the retention time). Furthermore, aggressive voltage scaling down to the sub-V_T regime, as well as aggressive technology scaling (down to a 40 nm CMOS node) have been investigated. Our analyses show that GC-eDRAM implementations in mature CMOS nodes (such as a 0.18 µm node) can be safely operated in the sub-V_T regime where, despite heavily degraded on/off current ratios and the strong impact of parametric variations, high array availability for write and read access can still be

ensured (i.e., the retention time is still long enough compared to the access time). However, for aggressively scaled CMOS nodes (such as a 40 nm node) characterized by high leakage currents, increased parametric variations, and lower achievable in-cell storage capacitance, voltage scaling should be limited to the near-V_T domain in order to guarantee reasonably high array availability. Finally, we have proposed a multilevel GC-eDRAM storing up to 2 bits per basic memory cell in order to achieve high storage densities at the cost of a small number of read failures that can be tolerated by the target application system.

In a 2-PMOS GC-eDRAM implemented in a mature 0.18 µm node, voltage scaling from the nominal voltage (1.8 V) to a near-V_T voltage (0.7 V) enhances the data retention time by 4× provided that write access is unlikely and that the write bit-line (WBL) can be controlled to ground during standby and read. The retention time can be further improved by 3.3× if the WBL is set to a voltage between the supply rails, which, however, comes at the cost of voltage generation circuits and is particularly interesting only for large GC-eDRAM arrays. Even with this total 13.2× improvement in retention time, the data retention power is still dominated by the active refresh power, while leakage power in the GC-eDRAM array plays only a minor role. Therefore, several techniques to further improve the retention time and reduce the active refresh power (thus significantly reducing the data retention power) of near-V_T GC-eDRAM have been proposed in this thesis. First of all, silicon measurements of a 2 kb GC-eDRAM macrocell implemented in a 0.18 µm CMOS process show that the retention time can be improved by 2.3× (from 23 to 53 ms) by applying a reverse body bias (RBB) of only 100 mV. This is the first demonstration of successfully applying reverse body biasing to GC-eDRAM arrays, which has only been used in conventional 1T-1C eDRAM thus far. Moreover, silicon measurements show that 100 mV forward body biasing (FBB), which can be selectively applied for fast memory access, leads to a 2.9× retention time penalty. Sweeping the body voltage over a range of 375 mV spans a retention time range of almost 2 orders of magnitude, providing and interesting trade-off between access time and retention time. Second of all, a replica bitcell technique, also implemented on a 2 kb all-PMOS 2T GC-eDRAM array in 0.18 µm CMOS, successfully tracks the retention time of the GC-eDRAM array across process-voltage-temperature (PVT) variations and varying write-access disturb frequencies. Silicon measurements show that the implemented replica technique allows to trigger refresh cycles up to 5× less frequently compared to conventional worst-case design, which significantly reduces the refresh power.

The possibility of operating GC-eDRAM at subthreshold (sub-V_T) voltages, for use in ultra-low power systems, and of implementing GC-eDRAM in deeply scaled CMOS nodes, for use in future high-performance VLSI systems, has been investigated in this thesis. In order to enable sub-V_T operation in mature, above-100 nm CMOS nodes, the main design goals of the bitcell are long retention time and high data integrity. In the considered 0.18 µm CMOS node, a low-leakage I/O PMOS write transistor and an extended storage node capacitance ensure a retention time of at least 40 ms. Since at ultra-low voltages the data integrity is severely threatened by charge injection and clock feedthrough (capacitive coupling from read and write word-lines), a core NMOS transistor is used as read transistor to balance the storage

node (SN) voltage disturbs (positive for write, negative for read); in addition, the core NMOS device is the strongest among all possible device options, ensuring a fast read operation and high array availability (i.e., fast read compared to retention time). Monte Carlo simulations of an entire 2 kb memory array, based on this mixed sub-V_T gain-cell design, operated at 1 MHz with a 400 mV sub-V_T supply voltage, confirm robust write and read operations under global and local parametric variations, as well as a minimum retention time of 40 ms leading to 99.7% availability for read and write. In deeply scaled CMOS technologies, such as the considered 40 nm CMOS node, subthreshold conduction is still dominant at ultra-low supply voltages. Gate tunneling and GIDL currents are still small, but of increasing importance, while reverse-biased pn-junction leakage and edge-direct tunneling currents are negligible. In the 40 nm node, the write transistor is best implemented with an HVT core PMOS device, which provides the lowest aggregated leakage current from the storage node (SN), even compared to the I/O PMOS device. Among various NMOS read transistor options, a standard-V_T core device maximizes the sense current ratio between a weak '1' and a weak '0' for near-V_T supply voltages (600–800 mV) where 97% array availibility is achieved. Both the access times and the retention time are roughly three orders-of-magnitude shorter than in the 0.18 µm CMOS node, due to the increased leakage currents and smaller storage node capacitance. Briefly, we showed the feasibility of sub-V_T GC-eDRAM operation for mature process technologies and near-V_T operation for a deeply scaled 40 nm process, and provided best-practice bitcell designs for achieving minimum V_{DD} at these two very different nodes.

Finally, the idea of storing many bits per gain-cell was investigated for the first time in this thesis. Our analyses of an 8 kb multilevel GC-eDRAM macrocell unveiled that the read failure probability at a given time after writing the storage array depends strongly on the state of the write bit-lines (WBLs) and can be reduced if the WBLs are controlled to the middle of their dynamic range during non-write operations. According to post-layout simulation results in a 90 nm CMOS technology, in case of storing 4 data levels (equivalent to 2 bits) per gain-cell, the maximum failure probability among all possible sense operations 10 µs (50 µs) after writting is less than 2% (8%), which can be tolerated by some error-resilient VLSI systems (for example a HSPA+ system). In addition, it was shown that the use of a replica column for optimum write and read control signal timing significantly improves both the write and read access speed of the otherwise rather slow and complex multilevel write and read operations. The use of multilevel GC-eDRAM is mostly interesting for large storage arrays (to offset the overhead of robust voltage generation circuits) implemented in mature CMOS nodes (above 100 nm); however, for aggressively scaled CMOS nodes, retention time requirements up to 50 µs, and system-level bitcell failure tolerances below 10%, the use of multilevel GC-eDRAM is not considered to be a viable option, favoring the use of our more conventional single-bit-per-cell GC-eDRAM implementations.

5.3 Outlook: SCMs and GC-eDRAMs in Future Applications

The implementation of conventional 6-transistor (6T)-bitcell SRAM for operation at aggressively scaled supply voltages often residing in the sub-V_T domain (see Section 3.1) or in aggressively scaled CMOS nodes (such as 28 nm CMOS or below) is extremely challenging, especially if reliable circuit operation and a high manufacturing yield need to be guaranteed in a high volume manufacturing (HVM) context for cost effectiveness. Ensuring robust operation of 6T SRAM is even more challenging for simultaneous voltage and technology scaling (such as sub-V_T or near-V_T circuit operation in a 28 nm CMOS node). While is is generally difficult to provide good SRAM memory compilers for these extreme conditions, standard-cell based memories (SCMs) are straightforward to implement and will work reliably even at ultra-low voltages and in deeply scaled, nanometric CMOS nodes, provided that a standard-cell library (SCL) is available. Note that SCLs, containing both combinational and sequential cells, required to synthesize SCMs, are typically the first development for each new technology node and are typically released before any SRAM memory compilers. In addition, we believe that non-volatile flip-flops and latches based on ReRAM device technology, integrated in form of distributed, synthesized storage arrays and/or state registers, will enable future low-power VLSI SoCs with zero-leakage standby states. However, for break-even sleep times below 1 s, compared to ultra-low leakage, retentive CMOS memories, the ReRAM technology should evolve to enable energy-efficient write at low voltages; moreover, for a large adoption of such emerging memory devices, the manufacturing processes have to mature in order to ensure high repeatability and yield.

Beyond the material covered in this thesis, there have been already further developments in the field of sub-V_T SCMs. While the herein presented SCM with ultra-low leakage latches and tri-state read logic (see Section 3.3) continues to exhibit the lowest leakage power and access energy per bit, the read access time was significantly improved by using segmented read bit-lines (RBLs), i.e., by limiting the number of tri-state drivers per RBL segment and using a small number of conventional CMOS circuits to complete the read multiplexer [83]. Further silicon measurements showed that at the same sub-V_T voltage, an even faster read access time was achieved by reverting to a pure CMOS read multiplexer, integrating the first stage (a NAND gate) as output buffer of the custom-designed latch [83]. Moreover, a sub-V_T 10-transistor (10T) latch circuit with output NAND buffer, avoiding write contention and read failures that would be encountered in a 6T SRAM bitcell, properly characterized as standard-cell and integrated into the SCM compilation flow, does not only preserve all the advantages of SCMs, but can also compete with 8-14T sub-V_T SRAM in terms of silicon area. Finally, we have also proposed further ReRAM based non-volatile flip-flop (NVFF) topologies which operate at low voltages. In particular, a NVFF topology using a single ReRAM device instead of two complementary programmed ReRAM devices dissipates less write energy but requires a higher voltage for reliable read operations.

Beside LDPC decoders with frequent and periodic write updates (see Section 2.3.2), we believe that a large number of VLSI SoCs can benefit from dynamic SCMs (D-SCMs) in the future.

On the one hand, the ongoing paradigm shift from 100% correct circuit operation to error-resilient VLSI systems (with error detection and correction mechanism), due to increasing parametric variations and high defect levels in nanometric CMOS nodes, favors the use of D-SCMs as distributed storage arrays. On the other hand, a large number of systems, in the field of wireless communications, multimedia (video, image, and audio processing), and data mining, can tolerate a small number of failing memory cells without the need for a correction mechanism; the adoption of D-SCMs for high storage density is certainly an interesting option for future VLSI implementations of such systems. In addition to the material presented in this thesis, we have carried out an extensive comparative analysis of a large number of dynamic latch topologies, in order to determine the most energy-efficient and smallest dynamic latch topology for a given system-level retention time requirement and failure resilience.

Compared to conventional 6T SRAM and 1T-1C eDRAM, gain-cell based eDRAM (GC-eDRAM) has a crucial advantage which can make it appealing for the implementation of embedded memories in advanced CMOS nodes or for operation at scaled voltages. In fact, as explained in detail in Section 4.1, gain-cells have a separate read and write port, which allows the simultaneous and independent optimization of a gain-cell for both robust read and write operations. Unfortunately, beside the possibility of achieving both robust read and write, the large spread of per-cell retention time and the small in-cell storage capacitor, coupled with conventional refresh time guardbanding leads to power-hungry refresh cycles. Therefore, especially for large cache memories, where an extremely unlikely worst-case cell dictates the refresh period, the adoption of GC-eDRAM is not an attractive option for the major semiconductor companies, which, in turn, focus most of their research on innovative ways of obtaining large, dedicated DRAM capacitors in below 28 nm CMOS nodes. For example, recent patents of Intel propose to use the readily available fin structure, used to build FinFETs (tri-gate transistors) to build large and high-density capacitors. We believe that combining the advantages of the gain-cell read and write ports with large, emerging, dedicated DRAM capacitors would lead to a wining new type of memory bitcell for future VLSI applications. However, if the use of dedicated DRAM capacitors is not economic, GC-eDRAM is still an interesting option for many VLSI SoCs requiring medium-size memory arrays and rather short data retention times. There are certainly many applications similar to the LDPC decoder presented in Section 2.3.2 which can benefit from GC-eDRAM, either operated with periodic refresh cycles, or in a refresh-free way due to frequent write updates [50]. Besides such high-performance VLSI DSP systems, GC-eDRAMs are also an interesting memory option for the niche of future ultra-low power (ULP) VLSI systems operated at ultra-low voltages (ULV) and implemented in mature, low-leakage, cheap CMOS processes (such as 0.18 µm CMOS). In fact, as seen in Section 4.4, it is possible to operate GC-eDRAM at sub-V_T voltages in mature CMOS nodes, and such sub-V_T 2T-bitcell GC-eDRAM is an extremely high-density alternative to the currently used 8-14T-bitcell sub-V_T SRAM macrocells.

Beyond the material covered in this thesis, there has been further innovative work in the field of GC-eDRAM. In fact, a 3-transistor (3T) gain-cell exhibits a full transmission-gate instead of a single write transistor, and a conventional merged storage and read transistor. This gain-

cell topology ensures fast write access at low voltages and avoids the use of any overdrive or underdrive voltage, thereby facilitating its integration into a digital SoC and avoiding the need for a costly voltage regulator. Moreover, a 4-transistor (4T) gain-cell contains an internal feedback transistor to strengthen the weaker data level, while hardly affecting the stronger data level. This leads to a much more symmetric decay of data '0' and '1' and a significantly extended retention time compared to a baseline 2T gain-cell. This gain-cell is an important step toward the realization of GC-eDRAM in deeply scaled CMOS nodes characterized by high leakage currents. Finally, another research direction aims at modeling the retention time of GC-eDRAM; in fact, an analytical model for the distribution of the retention time is derived based on statistics on primary circuit parameters, such as the threshold voltage and other transistor's parameters as well as the gate, junction, and interconnect capacitances. This eventually allows to carry out a sensitivity analysis and identify the main contributors to the typically large retention time spreads, for different operating regimes (sub-V_T and above-V_T domain) and for implementation in different CMOS nodes. Ultimately, this analysis not only allows to improve the statistical distributions of the dominant circuit parameters to narrow down the retention time distribution and reduce the refresh rate, but also enables to model and exploit the trade-off between read failure probability and refresh power in future error-resilient VLSI systems. In such future VLSI systems, the refresh rate could even be set dynamically in order to selectively change between an accurate, power-hungry computing mode and a less accurate, low-power computing mode. Research at the system level and in particular from a fault tolerance perspective could take significant advantage of this large per-cell retention time spread coupled with a dynamically set refresh rate.

A Analytical Sub-V_T Model

In Chapter 3, to exhaustively compare the energy dissipation and the critical path delay of a large number of standard-cell based memory (SCM) architectures, the following analytical sub-V_T characterization model, based on [96], was used. According to Fig. 3.2a in Section 3.2.1, this analytical model is applied to SCMs which have previously been synthesized, placed, and routed at nominal supply voltage (V_{DD}) using only commercially available standard-cell libraries and commercial digital design tools. Furthermore, the analytical model relies on the results from the post-layout static timing analysis (STA) and voltage-change dump (VCD)-based power analysis, both performed at nominal V_{DD}, in order to eventually predict the behavior of the SCMs in the entire sub-V_T regime.

The total energy dissipation E_T of static CMOS circuits operated in the sub-V_T regime is modelled as

$$E_T = \underbrace{\alpha C_{tot} {V_{DD}}^2}_{E_{dyn}} + \underbrace{I_{leak} V_{DD} T_{clk}}_{E_{leak}} + \underbrace{I_{peak} t_{sc} V_{DD}}_{E_{sc}}, \tag{A.1}$$

where E_{dyn}, E_{leak}, and E_{sc} are the average energy dissipation due to switching activity, the energy dissipation resulting form integrating the leakage power over one clock cycle T_{clk}, and the energy dissipation due to short circuit currents, respectively. The energy dissipation E_{sc} has been shown to be negligible in the sub-V_T regime [148]. The switching current causing the energy dissipation E_{dyn} results from subthreshold currents [149], i.e., from the drain currents of MOS transistors whose gate-to-source voltage V_{GS} is equal to or lower than the threshold voltage V_T ($V_{GS} \leq V_T$). Whenever the subthreshold current is not used to switch a circuit node, it contributes to E_{leak} together with all other types of leakage currents.

For a given clock period T_{clk}, (A.1) may be rewritten as

$$E_T = \mu_e C_{inv} k_{cap} {V_{DD}}^2 + k_{leak} I_0 V_{DD} T_{clk}, \tag{A.2}$$

where I_0 and C_{inv} are the average leakage current and the input capacitance of a single inverter, respectively. Furthermore, k_{leak} and k_{cap} are the average leakage and the capacitance of the circuit, respectively, both normalized to a single inverter. Moreover, μ_e is the circuit's average switching activity.

In the sub-V_T domain, it is beneficial to operate at the maximum achievable frequency to reach minimum energy dissipation per operation. In the following, (A.2) is therefore written again for the case where the clock period T_{clk} is equal to the critical path delay (T_{clk} denotes the critical path delay in the remainder of this section). The critical path delay itself may be written as

$$T_{clk} = k_{crit} T_{sw\_inv}, \tag{A.3}$$

where k_{crit} is the critical path delay of the circuit normalized to the inverter delay T_{sw_inv}. In [148], the delay T_{sw_inv} of an inverter operating in the sub-V_T regime is given by

$$T_{sw\_inv} = \frac{C_{inv} V_{DD}}{I_0 e^{V_{DD}/(nU_t)}}, \tag{A.4}$$

where n and U_t denote the slope factor and the thermal voltage, respectively. By introducing (A.4) into (A.3), the the critical path delay is now given by

$$T_{clk} = k_{crit} \frac{C_{inv} V_{DD}}{I_0 e^{V_{DD}/(nU_t)}}, \tag{A.5}$$

and the reciprocal of (A.5) defines the maximum frequency at which the circuit may be operated for a given supply voltage V_{DD}.

Finally, the total energy dissipation E_T assuming operation at the maximum frequency is found by introducing (A.5) into (A.2), which yields

$$E_T = C_{inv} {V_{DD}}^2 \left[\mu_e k_{cap} + k_{crit} k_{leak} e^{-V_{DD}/(nU_t)} \right]. \tag{A.6}$$

For the architectural analysis presented in Section 3.2.2, (A.6) has been used.

B Glossary

α_{disturb}	Write-'1' disturb activity factor
C_{SN}	Storage node capacitor
p_{fail}	Read failure probability
t_{ret}	Retention time
t_{up}	Update rate
t_{w}	Time upon write
V_{B}	Body voltage
V_{DD}	Supply voltage
V_{T}	Threshold voltage
1T-1C	1-transistor-1-capacitor
6T	6-transistor
ABB	Adaptive body biasing
Above-V_{T}	Above-threshold
AGC	Active gain-cell
ASIC	Application-specific integrated circuit
BB	Body bias
BL	Bit-line
BIST	Built-in self test
CF	Clock feedthrough
CG	Clock-gate
CI	Charge injection
CMOS	Complementary metal-oxide-semiconductor
DRAM	Dynamic random-access memory
DVFS	Dynamic voltage and frequency scaling

eDRAM	Embedded dynamic random-access memory
EMV	Energy minimum voltage
FBB	Forward body bias (or biasing)
FFE	Flip-flop with enable feature
FSM	Finite-state machine
GC	Gain-cell
GC-eDRAM	Gain-cell based embedded dynamic random-access memory
GIDL	Gate-induced drain leakage
HRS	High resistance (or resistive) state
HVM	High volume manufacturing
LRS	Low resistance (or resistive) state
LSB	Least significant bit
MEP	Minimum-energy point
MLDRAM	Multilevel dynamic random-access memory
MLGC	Multilevel gain-cell
MOSFET	Metal-oxide-semiconductor field-effect transistor
MOSCAP	Metal-oxide-semiconductor capacitor
MR	Read transistor
MS	Storage transistor
MSB	Most significant bit
MUT	Memory under test
MW	Write transistor
Near-V_T	Near-threshold
NMOS	N-channel MOSFET
PMOS	P-channel MOSFET
PVT	Process, voltage, temperature
RAM	Random-access memory
RBB	Reverse body bias (or biasing)
RBL	Read bit-line
RGC	Reference gain-cell
RWL	Read word-line
SA	Sense amplifier

SBB	Standard body bias (or biasing)
SCL	Standard-cell library
SCM	Standard-cell based memory
SN	Storage node
SNM	Static noise margin
SRAM	Static random-access memory
Sub-V_T	Subthreshold
ULP	Ultra-low power
ULV	Ultra-low voltage
VTC	Voltage transfer curve (or characteristic)
WL	Word-line
WBL	Write bit-line
WWL	Write word-line

Bibliography

[1] "International technology roadmap for semiconductors," 2011. [Online]. Available: http://www.itrs.net/Links/2011ITRS/Home2011.htm

[2] C. Roth, P. Meinerzhagen, C. Studer, and A. Burg, "A 15.8 pJ/bit/iter quasi-cyclic LDPC decoder for IEEE 802.11n in 90 nm CMOS," in *Proc. IEEE Asian Solid State Circuits Conference (A-SSCC)*, 2010, pp. 1–4.

[3] L. Chang, D. Fried, J. Hergenrother, J. Sleight, R. Dennard, R. Montoye, L. Sekaric, S. McNab, A. Topol, C. Adams, K. Guarini, and W. Haensch, "Stable SRAM cell design for the 32 nm node and beyond," in *Proc. IEEE Symposium on VLSI Technology (VLSIT)*, 2005, pp. 128–129.

[4] A. Teman, L. Pergament, O. Cohen, and A. Fish, "A 250 mV 8 kb 40 nm ultra-low power 9T supply feedback SRAM (SF-SRAM)," *IEEE Journal of Solid-State Circuits (JSSC)*, vol. 46, no. 11, pp. 2713–2726, 2011.

[5] S. Jain, S. Khare, S. Yada, V. Ambili, P. Salihundam, S. Ramani, S. Muthukumar, M. Srinivasan, A. Kumar, S. Gb, R. Ramanarayanan, V. Erraguntla, J. Howard, S. Vangal, S. Dighe, G. Ruhl, P. Aseron, H. Wilson, N. Borkar, V. De, and S. Borkar, "A 280mV-to-1.2V wide-operating-range IA-32 processor in 32nm CMOS," in *Proc. IEEE International Solid-State Circuits Conference (ISSCC)*, 2012, pp. 66–68.

[6] B. Calhoun and A. Chandrakasan, "A 256-kb 65-nm sub-threshold SRAM design for ultra-low-voltage operation," *IEEE Journal of Solid-State Circuits (JSSC)*, vol. 42, no. 3, pp. 680–688, 2007.

[7] I. Kazi, P. Meinerzhagen, P.-E. Gaillardon, D. Sacchetto, A. Burg, and G. De Micheli, "ReRAM-based non-volatile flip-flop with sub-VT read and CMOS voltage-compatible write," in *Proc. IEEE International NEWCAS Conference*, 2013.

[8] H. Kaeslin, *Digital Integrated Circuit Design: From VLSI Architectures to CMOS Fabrication*, 1st ed. Cambridge University Press, 2008.

[9] "Nehalem part 3: The cache debate, LGA-1156 and the 32nm future," November 2008. [Online]. Available: http://www.anandtech.com/show/2671

[10] M. Bohr, "The new era of scaling in an SoC world," in *IEEE International Solid-State Circuits Conference (ISSCC)*, 2009, pp. 23–28.

[11] S. Damaraju, V. George, S. Jahagirdar, T. Khondker, R. Milstrey, S. Sarkar, S. Siers, I. Stolero, and A. Subbiah, "A 22nm IA multi-CPU and GPU system-on-chip," in *IEEE International Solid-State Circuits Conference (ISSCC)*, 2012, pp. 56–57.

[12] A. Burg, S. Haene, M. Borgmann, D. Baum, T. Thaler, F. Carbognani, S. Zwicky, L. Barbero, C. Senning, P. Greisen, T. Peter, C. Foelmli, U. Schuster, P. Tejera, and A. Staudacher, "A 4-stream 802.11n baseband transceiver in 0.13 μm CMOS," in *IEEE Symposium on VLSI Circuits*, 2009, pp. 282–283.

[13] J. Constantin, A. Dogan, O. Andersson, P. Meinerzhagen, J. Rodrigues, D. Atienza, and A. Burg, "TamaRISC-CS: An ultra-low-power application-specific processor for compressed sensing," in *Proc. IEEE/IFIP International Conference on VLSI System-on-Chip (VLSI-SoC)*, 2012, pp. 159–164.

[14] C. Studer, N. Preyss, C. Roth, and A. Burg, "Configurable high-throughput decoder architecture for quasi-cyclic LDPC codes," in *Proc. IEEE Asilomar Conference on Signals, Systems and Computers*, Oct. 2008, pp. 1137–1142.

[15] R. Dreslinski, M. Wieckowski, D. Blaauw, D. Sylvester, and T. Mudge, "Near-threshold computing: Reclaiming moore's law through energy efficient integrated circuits," *Proceedings of the IEEE*, vol. 98, no. 2, pp. 253–266, 2010.

[16] J. Yoo, L. Yan, D. El-Damak, M. Bin Altaf, A. Shoeb, H.-J. Yoo, and A. Chandrakasan, "An 8-channel scalable EEG acquisition SoC with fully integrated patient-specific seizure classification and recording processor," in *Proc. IEEE International Solid-State Circuits Conference (ISSCC)*, 2012, pp. 292–294.

[17] F. Zhang, Y. Zhang, J. Silver, Y. Shakhsheer, M. Nagaraju, A. Klinefelter, J. Pandey, J. Boley, E. Carlson, A. Shrivastava, B. Otis, and B. Calhoun, "A batteryless 19 μW MICS/ISM-band energy harvesting body area sensor node SoC," in *Proc. IEEE International Solid-State Circuits Conference (ISSCC)*, 2012, pp. 298–300.

[18] M. Bushnell and V. Agrawal, *Essentials of Electronic Testing for Digital, Memory and Mixed-Signal VLSI Circuits*. Springer Verlag, 2000, ch. 9.1.

[19] S. Jahinuzzaman, J. Shah, D. Rennie, and M. Sachdev, "Design and analysis of a 5.3-pJ 64-kb gated ground SRAM with multiword ECC," *IEEE Journal of Solid-State Circuits (JSSC)*, vol. 44, no. 9, pp. 2543–2553, 2009.

[20] A. Kumar, J. Rabaey, and K. Ramchandran, "SRAM supply voltage scaling: A reliability perspective," in *Proc. IEEE International Symposium on Quality Electronic Design (ISQED)*, 2009, pp. 782–787.

[21] G. Karakonstantis, C. Roth, C. Benkeser, and A. Burg, "On the exploitation of the inherent error resilience of wireless systems under unreliable silicon," in *Proc. ACM/EDAC/IEEE Design Automation Conference (DAC)*, June 2012, pp. 510–515.

[22] M. M. Sabry, G. Karakonstantis, D. Atienza, and A. Burg, "Design of energy efficient and dependable health monitoring systems under unreliable nanometer technologies," in *Proc. ACM International Conference on Body Area Networks*, 2012, pp. 52–58.

[23] D. Ernst, N. S. Kim, S. Das, S. Pant, R. Rao, T. Pham, C. Ziesler, D. Blaauw, T. Austin, K. Flautner, and T. Mudge, "Razor: a low-power pipeline based on circuit-level timing speculation," in *Proc. IEEE/ACM International Symposium on Microarchitecture (MICRO)*, 2003, pp. 7–18.

[24] K. Bowman, J. Tschanz, S. Lu, P. Aseron, M. Khellah, A. Raychowdhury, B. Geuskens, C. Tokunaga, C. Wilkerson, T. Karnik, and V. De, "A 45 nm resilient microprocessor core for dynamic variation tolerance," *IEEE Journal of Solid-State Circuits*, vol. 46, no. 1, pp. 194–208, 2011.

[25] P. P. Pande, A. Ganguly, and K. Chakrabarty, *Design Technologies for Green and Sustainable Computing Systems*. Springer, 2013, Chapter 9: Claremont: A Solar-Powered Near-Threshold Voltage IA-32 Processor, by Sriram Vangal and Shailendra Jain.

[26] "A solar powered IA core? no way!" Intel Developer Forum, Sept. 2011. [Online]. Available: http://blogs.intel.com/research/2011/09/ntvp

[27] S. Seo, R. Dreslinski, M. Woh, C. Chakrabarti, S. Mahlke, and T. Mudge, "Diet SODA: A power-efficient processor for digital cameras," in *Proc. ACM/IEEE International Symposium on Low-Power Electronics and Design (ISLPED)*, 2010, pp. 79–84.

[28] D. Somasekhar, Y. Ye, P. Aseron, S.-L. Lu, M. Khellah, J. Howard, G. Ruhl, T. Karnik, S. Borkar, V. De, and A. Keshavarzi, "2GHz 2Mb 2T gain-cell memory macro with 128GB/s bandwidth in a 65nm logic process," in *Proc. IEEE International Solid-State Circuits Conference (ISSCC)*, 2008, pp. 274–613.

[29] A. Teman, P. Meinerzhagen, A. Burg, and A. Fish, "Review and classification of gain cell eDRAM implementations," in *Proc. IEEE Convention of Electrical and Electronics Engineers in Israel (IEEEI)*, 2012, pp. 1–5.

[30] S. Kang and Y. Leblebici, *CMOS Digital Integrated Circuits: Analysis and Design*, 3rd ed. McGraw-Hill, 2003.

[31] M. Qazi, M. Sinangil, and A. Chandrakasan, "Challenges and directions for low-voltage SRAM," *IEEE Design and Test of Computers*, vol. 28, no. 1, pp. 32–43, 2011.

[32] P. Meinerzhagen, C. Roth, and A. Burg, "Towards generic low-power area-efficient standard cell based memory architectures," in *Proc. IEEE International Midwest Symposium on Circuits and Systems (MWSCAS)*, 2010, pp. 129–132.

[33] P. Meinerzhagen, S. Sherazi, A. Burg, and J. Rodrigues, "Benchmarking of standard-cell based memories in the sub-VT domain in 65-nm CMOS technology," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems (JETCAS)*, vol. 1, no. 2, pp. 173–182, 2011.

[34] K.-S. Yeo and K. Roy, *Low-Voltage, Low-Power VLSI Subsystems*. McGraw-Hill, 2005.

[35] B. Zhai, S. Pant, L. Nazhandali, S. Hanson, J. Olson, A. Reeves, M. Minuth, R. Helfand, T. Austin, D. Sylvester, and D. Blaauw, "Energy-efficient subthreshold processor design," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems (TVLSI)*, vol. 17, no. 8, pp. 1127–1137, Aug. 2009.

[36] A. Wang and A. Chandrakasan, "A 180-mV FFT processor using subthreshold circuit techniques," in *Proc. IEEE International Solid-State Circuits Conference (ISSCC)*, vol. 1, Feb. 2004, pp. 292–529.

[37] C. Roth, A. Cevrero, C. Studer, Y. Leblebici, and A. Burg, "Area, throughput, and energy-efficiency trade-offs in the VLSI implementation of LDPC decoders," in *Proc. IEEE International Symposium on Circuits and Systems (ISCAS)*, 2011, pp. 1772–1775.

[38] T.-C. Kuo and A. Willson, "A flexible decoder IC for WiMAX QC-LDPC codes," in *Proc. IEEE Custom Integrated Circuits Conference (CICC)*, Sept. 2008, pp. 527–530.

[39] P. Urard, L. Paumier, V. Heinrich, N. Raina, and N. Chawla, "A 360mW 105Mb/s DVB-S2 compliant codec based on 64800b LDPC and BCH codes enabling satellite-transmission portable devices," in *Proc. IEEE International Solid-State Circuits Conference (ISSCC)*, 2008, pp. 310–311.

[40] P. Urard, E. Yeo, L. Paumier, P. Georgelin, T. Michel, V. Lebars, E. Lantreibecq, and B. Gupta, "A 135Mb/s DVB-S2 compliant codec based on 64800b LDPC and BCH codes," in *Proc. IEEE International Solid-State Circuits Conference (ISSCC)*, 2005, pp. 446–609.

[41] M. M. Mansour and N. R. Shanbhag, "A 640-Mb/s 2048-bit programmable LDPC decoder chip," *IEEE Journal of Solid-State Circuits (JSSC)*, vol. 41, no. 3, pp. 684–698, 2006.

[42] C.-H. Liu, S.-W. Yen, C.-L. Chen, H.-C. Chang, C.-Y. Lee, Y.-S. Hsu, and S.-J. Jou, "An LDPC decoder chip based on self-routing network for IEEE 802.16e applications," *IEEE Journal of Solid-State Circuits (JSSC)*, vol. 43, no. 3, pp. 684–694, 2008.

[43] Y. Sun and J. R. Cavallaro, "A low-power 1-Gbps reconfigurable LDPC decoder design for multiple 4G wireless standards," in *Proc. IEEE International ScC Conference*, Sept. 2008, pp. 367–370.

[44] Q. Xie, Q. He, X. Peng, Y. Cui, Z. Chen, D. Zhou, and S. Goto, "A high parallel macro block level layered LDPC decoding architecture based on dedicated matrix reordering," in *Proc. IEEE Workshop on Signal Processing Systems (SiPS)*, 2011, pp. 122–127.

[45] Y. Cui, X. Peng, Z. Chen, X. Zhao, Y. Lu, D. Zhou, and S. Goto, "Ultra low power QC-LDPC decoder with high parallelism," in *Proc. IEEE International SOC Conference (SOCC)*, 2011, pp. 142–145.

[46] Y. Sun, G. Wang, and J. Cavallaro, "Multi-layer parallel decoding algorithm and VLSI architecture for quasi-cyclic LDPC codes," in *Proc. IEEE International Symposium on Circuits and Systems (ISCAS)*, 2011, pp. 1776–1779.

[47] J. Lillis and C.-K. Cheng, "Timing optimization for multisource nets: characterization and optimal repeater insertion," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 18, no. 3, pp. 322–331, March 1999.

[48] *IEEE Unapproved Draft Std P802.11n/D11.0*, June 2009.

[49] X.-Y. Shih, C.-Z. Zhan, and A.-Y. Wu, "A real-time programmable LDPC decoder chip for arbitrary QC-LDPC parity check matrices," in *Proc. IEEE Asian Solid-State Circuits Conference (A-SSCC)*, Nov. 2009, pp. 369–372.

[50] Y. S. Park, D. Blaauw, D. Sylvester, and Z. Zhang, "A 1.6-mm^2 38-mW 1.5-Gb/s LDPC decoder enabled by refresh-free embedded DRAM," in *Proc. IEEE Symposium on VLSI Circuits (VLSIC)*, 2012, pp. 114–115.

[51] A. Lingamneni, C. Enz, J.-L. Nagel, K. Palem, and C. Piguet, "Energy parsimonious circuit design through probabilistic pruning," in *Proc. IEEE Design, Automation, and Test in Europe Conference & Exhibition (DATE)*, March 2011, pp. 1–6.

[52] C. Novak, C. Studer, A. Burg, and G. Matz, "The effect of unreliable LLR storage on the performance of MIMO-BICM," in *Proc. IEEE Asilomar Conference on Signals, Systems, and Computers*, November 2010.

[53] C. Roth, C. Benkeser, C. Studer, G. Karakonstantis, and A. Burg, "Data mapping for unreliable memories," in *Proc. IEEE Allerton Conference on Communication, Control, and Computing*, October 2012.

[54] V. Chippa, A. Raghunathan, K. Roy, and S. Chakradhar, "Dynamic effort scaling: Managing the quality-efficiency tradeoff," in *Proc. ACM/EDAC/IEEE Design Automation Conference (DAC)*, June 2011, pp. 603–608.

[55] M. A. Breuer, "Let's think analog," in *Proc. IEEE Computer Society Annual Symposium on VLSI (CSAS on VLSI): New Frontiers in VLSI Design*, May 2005, pp. 2–5.

[56] S. Ghosh and K. Roy, "Parameter variation tolerance and error resiliency: New design paradigm for the nanoscale era," *Proceedings of the IEEE*, vol. 98, no. 10, pp. 1718–1751, Oct. 2010.

[57] Q. Xie, Q. He, X. Peng, Y. Cui, Z. Chen, D. Zhou, and S. Goto, "A high parallel macro block level layered LDPC decoding architecture based on dedicated matrix reordering," in *Proc. IEEE Workshop on Signal Processing Systems (SiPS)*, 2011, pp. 122–127.

[58] H. Zhong and T. Zhang, "Block-LDPC: a practical LDPC coding system design approach," *IEEE Transactions on Circuits and Systems I (TCAS-I)*, vol. 52, no. 4, pp. 766–775, 2005.

[59] K. Gunnam, G. Choi, W. Wang, and M. Yeary, "Multi-rate layered decoder architecture for block LDPC codes of the IEEE 802.11n wireless standard," in *Proc. IEEE International Symposium on Circuits and Systems (ISCAS)*, 2007, pp. 1645–1648.

[60] *Digital Video Broadcasting (DVB) User guidelines for the second generation system for Broadcasting, Interactive Services, News Gathering and other broadband satellite applications (DVB-S2)*, Feb. 2005.

[61] "Wireless LAN medium access control (MAC) and physical layer (PHY) specifications: Enhancements for higher throughput," *IEEE P802.11n/D5.02, Part 11*, July 2008.

[62] V. Kursun and E. Friedman, "Domino logic with variable threshold voltage keeper," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems (TVLSI)*, vol. 11, no. 6, pp. 1080–1093, 2003.

[63] P. A. Meinerzhagen, O. Andiç, J. Treichler, and A. P. Burg, "Design and failure analysis of logic-compatible multilevel gain-cell-based DRAM for fault-tolerant VLSI systems," in *Proc. IEEE/ACM Great Lakes Symposium on VLSI (GLSVLSI)*, 2011, pp. 343–346.

[64] K. C. Chun, W. Zhang, P. Jain, and C. Kim, "A 2T1C embedded DRAM macro with no boosted supplies featuring a 7T SRAM based repair and a cell storage monitor," *IEEE Journal of Solid-State Circuits (JSSC)*, vol. 47, no. 10, pp. 2517–2526, 2012.

[65] R. Iqbal, P. Meinerzhagen, and A. Burg, "Two-port low-power gain-cell storage array: Voltage scaling and retention time," in *Proc. IEEE International Symposium on Circuits and Systems (ISCAS)*, 2012, pp. 2469–2472.

[66] X. Peng, Z. Chen, X. Zhao, D. Zhou, and S. Goto, "A 115mW 1Gbps QC-LDPC decoder ASIC for WiMAX in 65nm CMOS," in *IEEE Asian Solid State Circuits Conference (A-SSCC)*, 2011, pp. 317–320.

[67] B. Xiang, D. Bao, S. Huang, and X. Zeng, "An 847-955 Mb/s 342-397 mw dual-path fully-overlapped QC-LDPC decoder for WiMAX system in 0.13um CMOS," *IEEE Journal of Solid-State Circuits (JSSC)*, vol. 46, no. 6, pp. 1416–1432, 2011.

[68] R. Sarpeshkar, "Ultra low power electronics for medicine," in *Proc. IEEE International Workshop on Wearable and Implantable Body Sensor Networks (BSN)*, April 2006.

[69] F. Kurdahi, A. Eltawil, K. Yi, S. Cheng, and A. Khajeh, "Low-power multimedia system design by aggressive voltage scaling," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems (TVLSI)*, vol. 18, no. 5, pp. 852–856, 2010.

[70] J.-J. Kim and K. Roy, "Double gate-MOSFET subthreshold circuit for ultra low power applications," *IEEE Transactions on Electron Devices (TED)*, vol. 51, no. 9, pp. 1468–1474, Sept. 2004.

[71] M. Sinangil, N. Verma, and A. Chandrakasan, "A reconfigurable 65nm SRAM achieving voltage scalability from 0.25-1.2V and performance scalability from 20kHz-200MHz," in *Proc. IEEE European Solid-State Circuits Conference (ESSCIRC)*, Sept. 2008, pp. 282–285.

[72] B. Calhoun, A. Wang, and A. Chandrakasan, "Device sizing for minimum energy operation in subthreshold circuits," in *Proc. IEEE Custom Integrated Circuits Conference (CICC)*, Oct. 2004, pp. 95–98.

[73] B. H. Calhoun, A. Wang, and A. Chandrakasan, "Modeling and sizing for minimum energy operation in subthreshold circuits," *IEEE Journal of Solid-State Circuits (JSSC)*, vol. 40, no. 9, pp. 1778–1786, Sept. 2005.

[74] J. Rodrigues, O. Akgun, and V. Owall, "A <1 pJ sub-VT cardiac event detector in 65 nm LL-HVT CMOS," in *Proc. IEEE/IFIP VLSI System on Chip Conference (VLSI-SoC)*, June 2010, pp. 253–258.

[75] J. Chen, L. Clark, and T.-H. Chen, "An ultra-low-power memory with a subthreshold power supply voltage," *IEEE Journal of Solid-State Circuits (JSSC)*, vol. 41, no. 10, pp. 2344–2353, Oct. 2006.

[76] N. Verma and A. Chandrakasan, "A 65nm 8T sub-Vt SRAM employing sense-amplifier redundancy," in *Proc. IEEE International Solid-State Circuits Conference (ISSCC)*, Feb. 2007, pp. 328–606.

[77] M. Sinangil, N. Verma, and A. Chandrakasan, "A reconfigurable 8T ultra-dynamic voltage scalable (U-DVS) SRAM in 65 nm CMOS," *IEEE Journal of Solid-State Circuits (JSSC)*, vol. 44, no. 11, pp. 3163–3173, Nov. 2009.

[78] S.-C. Luo and L.-Y. Chiou, "A sub-200-mV voltage-scalable SRAM with tolerance of access failure by self-activated bitline sensing," *IEEE Transactions on Circuits and Systems II (TCAS-II)*, vol. 57, no. 6, pp. 440–445, June 2010.

[79] M.-F. Chang, J.-J. Wu, K.-T. Chen, Y.-C. Chen, Y.-H. Chen, R. Lee, H.-J. Liao, and H. Yamauchi, "A differential data-aware power-supplied (D2AP) 8T SRAM cell with expanded write/read stabilities for lower VDDmin applications," *IEEE Journal of Solid-State Circuits (JSSC)*, vol. 45, no. 6, pp. 1234–1245, June 2010.

[80] T.-H. Kim, J. Liu, J. Keane, and C. Kim, "A high-density subthreshold SRAM with data-independent bitline leakage and virtual ground replica scheme," in *Proc. IEEE International Solid-State Circuits Conference (ISSCC)*, Feb. 2007, pp. 330–606.

[81] P. Meinerzhagen, O. Andersson, Y. Sherazi, A. Burg, and J. Rodrigues, "Synthesis strategies for sub-VT systems," in *Proc. IEEE European Conference on Circuit Theory and Design (ECCTD)*, 2011, pp. 552–555.

[82] P. Meinerzhagen, O. Andersson, B. Mohammadi, Y. Sherazi, A. Burg, and J. Rodrigues, "A 500 fW/bit 14 fJ/bit-access 4kb standard-cell based sub-VT memory in 65nm CMOS," in *Proc. IEEE European Solid-State Circuits Conference (ESSCIRC)*, 2012, pp. 321–324.

[83] O. Andersson, B. Mohammadi, P. Meinerzhagen, A. Burg, and J. Rodrigues, "Dual-VT 4kb sub-VT memories with <1 pW/bit leakage in 65nm CMOS," in *Proc. IEEE European Solid-State Circuits Conference (ESSCIRC)*, 2013.

[84] S. Mukhopadhyay, H. Mahmoodi, and K. Roy, "Modeling of failure probability and statistical design of SRAM array for yield enhancement in nanoscaled cmos," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 24, no. 12, pp. 1859–1880, 2005.

[85] S. Mukhopadhyay, K. Kang, H. Mahmoodi, and K. Roy, "Reliable and self-repairing SRAM in nano-scale technologies using leakage and delay monitoring," in *Proc. IEEE International Test Conference (ITC)*, 2005, pp. 1–10.

[86] E. Karl, Y. Wang, Y.-G. Ng, Z. Guo, F. Hamzaoglu, U. Bhattacharya, K. Zhang, K. Mistry, and M. Bohr, "A 4.6GHz 162Mb SRAM design in 22nm tri-gate CMOS technology with integrated active VMIN-enhancing assist circuitry," in *Proc. IEEE International Solid-State Circuits Conference (ISSCC)*, 2012, pp. 230–232.

[87] K. Zhang, U. Bhattacharya, Z. Chen, F. Hamzaoglu, D. Murray, N. Vallepalli, Y. Wang, B. Zheng, and M. Bohr, "A 3-GHz 70-Mb SRAM in 65-nm CMOS technology with integrated column-based dynamic power supply," *IEEE Journal of Solid-State Circuits (JSSC)*, vol. 41, no. 1, pp. 146–151, 2006.

[88] A. Raychowdhury, B. Geuskens, J. Kulkarni, J. Tschanz, K. Bowman, T. Karnik, S.-L. Lu, V. De, and M. Khellah, "PVT-and-aging adaptive wordline boosting for 8T SRAM power reduction," in *Proc. IEEE International Solid-State Circuits Conference (ISSCC)*, 2010, pp. 352–353.

[89] S. Hanson, M. Seok, Y.-S. Lin, Z. Foo, D. Kim, Y. Lee, N. Liu, D. Sylvester, and D. Blaauw, "A low-voltage processor for sensing applications with picowatt standby mode," *IEEE Journal of Solid-State Circuits (JSSC)*, vol. 44, no. 4, pp. 1145–1155, April 2009.

[90] Y.-W. Chiu, J.-Y. Lin, M.-H. Tu, S.-J. Jou, and C.-T. Chuang, "8T single-ended sub-threshold SRAM with cross-point data-aware write operation," in *Proc. IEEE/ACM International Symposium on Low-Power Electronics and Design (ISLPED)*, 2011, pp. 169–174.

[91] A. C. Cabe and M. R. Stan, "Experimental demonstration of standby power reduction using voltage stacking in an 8Kb embedded FDSOI SRAM," in *Proc. ACM/IEEE Great Lakes Symposium on VLSI (GLSVLSI)*, 2011, pp. 399–402.

[92] A. Teman, O. Yadid-Pecht, and A. Fish, "Leakage reduction in advanced image sensors using an improved ab2c scheme," *IEEE Sensors Journal*, vol. 12, no. 4, pp. 773–784, 2012.

[93] A. Fish, T. Rothschild, A. Hodes, Y. Shoshan, and O. Yadid-Pecht, "Low power CMOS image sensors employing adaptive bulk biasing control (AB2C) approach," in *Proc. IEEE International Symposium on Circuits and Systems (ISCAS)*, 2007, pp. 2834–2837.

[94] M. Alioto, "Impact of NMOS/PMOS imbalance in ultra-low voltage CMOS standard cells," in *Proc. IEEE European Conference on Circuit Theory and Design (ECCTD)*, Aug. 2011, pp. 536–539.

[95] S. Amarchinta, H. Kanitkar, and D. Kudithipudi, "Robust and high performance sub-threshold standard cell design," in *Proc. IEEE International Midwest Symposium on Circuits and Systems (MWSCAS)*, Aug. 2009, pp. 1183–1186.

[96] O. C. Akgun and Y. Leblebici, "Energy efficiency comparison of asynchronous and synchronous circuits operating in the sub-threshold regime," *Journal of Low Power Electronics*, vol. 4, Oct. 2008.

[97] O. Akgun, J. Rodrigues, Y. Leblebici, and V. Owall, "High-level energy estimation in the sub-VT domain: Simulation and measurement of a cardiac event detector," *IEEE Transactions on Biomedical Circuits and Systems (TBCAS*, vol. 6, no. 1, pp. 15–27, 2012.

[98] S. Hanson, B. Zhai, K. Bernstein, D. Blaauw, A. Bryant, L. Chang, K. Das, W. Haensch, E. Nowak, and D. Sylvester, "Ultralow-voltage, minimum-energy CMOS," *IBM Journal of Research and Development*, vol. 50, no. 4.5, pp. 469–490, 2006.

[99] A. Agarwal, B. Paul, S. Mukhopadhyay, and K. Roy, "Process variation in embedded memories: failure analysis and variation aware architecture," *IEEE Journal of Solid-State Circuits (JSSC)*, vol. 40, no. 9, pp. 1804–1814, Sept. 2005.

[100] B. Calhoun and A. Chandrakasan, "Static noise margin variation for sub-threshold SRAM in 65-nm CMOS," *IEEE Journal of Solid-State Circuits (JSSC)*, vol. 41, no. 7, pp. 1673–1679, July 2006.

[101] Y. Wang, H. J. Ahn, U. Bhattacharya, Z. Chen, T. Coan, F. Hamzaoglu, W. Hafez, C.-H. Jan, P. Kolar, S. Kulkarni, J.-F. Lin, Y.-G. Ng, I. Post, L. Wei, Y. Zhang, K. Zhang, and M. Bohr, "A 1.1 GHz 12 uA/Mb-leakage SRAM design in 65 nm ultra-low-power CMOS technology with integrated leakage reduction for mobile applications," *IEEE Journal of Solid-State Circuits (JSSC)*, vol. 43, no. 1, pp. 172–179, 2008.

[102] C.-M. Jung, K.-H. Jo, E.-S. Lee, H. M. Vo, and K.-S. Min, "Zero-sleep-leakage flip-flop circuit with conditional-storing memristor retention latch," *IEEE Transactions on Nanotechnology (TNANO)*, vol. 11, no. 2, pp. 360–366, 2012.

[103] G. Burr, B. Kurdi, J. Scott, C. Lam, K. Gopalakrishnan, and R. Shenoy, "Overview of candidate device technologies for storage-class memory," *IBM Journal of Research and Development*, vol. 52, no. 4.5, pp. 449–464, July 2008.

[104] D. B. Strukov, G. S. Snider, D. R. Stewart, and R. S. Williams, "The missing memristor found," *Nature*, vol. 453, no. 7191, pp. 80–83, 2008.

[105] S. Onkaraiah, M. Reyboz, F. Clermidy, J. Portal, M. Bocquet, C. Muller, H. Hraziia, C. Anghel, and A. Amara, "Bipolar ReRAM based non-volatile flip-flops for low-power architectures," in *Proc. IEEE International NEWCAS Conference*, June 2012, pp. 417–420.

[106] Y. Jung, J. Kim, K. Ryu, J. P. Kim, S. H. Kang, and S.-O. Jung, "An MTJ-based non-volatile flip-flop for high-performance SoC," *International Journal of Circuit Theory and Applications*, 2012.

[107] W. Zhao, E. Belhaire, and C. Chappert, "Spin-MTJ based non-volatile flip-flop," in *Proc. IEEE Conference on Nanotechnology (NANO)*, 2007, pp. 399–402.

[108] Y. Jung, J. Kim, K. Ryu, S.-O. Jung, J. Kim, and S. Kang, "MTJ based non-volatile flip-flop in deep submicron technology," in *Proc. IEEE International SoC Design Conference (ISOCC)*, 2011, pp. 424–427.

[109] Y. S. Chen, H. Lee, P. Chen, C. Tsai, P. Gu, T. Y. Wu, K. Tsai, S. S. Sheu, W. Lin, C. H. Lin, P. Chiu, W.-S. Chen, F. Chen, C. Lien, and M.-J. Tsai, "Challenges and opportunities for HfOX based resistive random access memory," in *Proc. IEEE International Electron Devices Meeting (IEDM)*, 2011, pp. 31.3.1–31.3.4.

[110] N. Sakimura, T. Sugibayashi, R. Nebashi, and N. Kasai, "Nonvolatile magnetic flip-flop for standby-power-free SoCs," in *Proc. IEEE Custom Integrated Circuits Conference (CICC)*, 2008, pp. 355–358.

[111] ——, "Nonvolatile magnetic flip-flop for standby-power-free SoCs," *IEEE Journal of Solid-State Circuits (JSSC)*, vol. 44, no. 8, pp. 2244–2250, 2009.

[112] M. Khalid, P. Meinerzhagen, and A. Burg, "Replica bit-line technique for embedded multilevel gain-cell DRAM," in *Proc. IEEE International NEWCAS Conference*, June 2012, pp. 77–80.

[113] P. Meinerzhagen, A. Teman, A. Burg, and A. Fish, "On the impact of body biasing on the retention time of gain-cell memories," *IET Journal on Engineering (JoE)*, vol. 1, August 2013.

[114] P. Meinerzhagen, A. Teman, A. Mordakhay, A. Burg, and A. Fish, "A sub-VT 2T gain-cell memory for biomedical applications," in *Proc. IEEE Subthreshold Microelectronics Conference (SubVT)*, 2012, pp. 1–3.

[115] P. Meinerzhagen, A. Teman, R. Giterman, A. Burg, and A. Fish, "Exploration of sub-VT and near-VT 2T gain-cell memories for ultra-low power applications under technology scaling," *Journal of Low Power Electronics and Applications (JLPEA)*, vol. 3, no. 2, pp. 54–72, April 2013.

[116] W. Luk and R. Dennard, "A novel dynamic memory cell with internal voltage gain," *IEEE Journal of Solid-State Circuits (JSSC)*, vol. 40, no. 4, pp. 884–894, April 2005.

[117] K. C. Chun, P. Jain, T.-H. Kim, and C. Kim, "A 667 MHz logic-compatible embedded DRAM featuring an asymmetric 2T gain cell for high speed on-die caches," *IEEE Journal of Solid-State Circuits (JSSC)*, vol. 47, no. 2, pp. 547–559, 2012.

[118] M. Kaku, H. Iwai, T. Nagai, M. Wada, A. Suzuki, T. Takai, N. Itoga, T. Miyazaki, T. Iwai, H. Takenaka, T. Hojo, S. Miyano, and N. Otsuka, "An 833MHz pseudo-two-port embedded DRAM for graphics applications," in *Proc. IEEE International Solid-State Circuits Conference (ISSCC)*, 2008, pp. 276–613.

[119] W. Luk and R. Dennard, "2T1D memory cell with voltage gain," in *Proc. IEEE Symposium on VLSI Circuits (VLSIC)*, 2004, pp. 184–187.

[120] D. Somasekhar, S.-L. Lu, B. Bloechel, G. Dermer, K. Lai, S. Borkar, and V. De, "A 10Mbit, 15GBytes/sec bandwidth 1T DRAM chip with planar MOS storage capacitor in an unmodified 150nm logic process for high-density on-chip memory applications," in *Proc. IEEE European Solid-State Circuits Conference (ESSCIRC)*, 2005, pp. 355–358.

[121] W. Luk, J. Cai, R. Dennard, M. Immediato, and S. Kosonocky, "A 3-transistor DRAM cell with gated diode for enhanced speed and retention time," in *Proc. IEEE Symposium on VLSI Circuits (VLSIC)*, 2006, pp. 184–185.

[122] D. Somasekhar, Y. Ye, P. Aseron, S.-L. Lu, M. Khellah, J. Howard, G. Ruhl, T. Karnik, S. Borkar, V. De, and A. Keshavarzi, "2 GHz 2 Mb 2T gain cell memory macro with 128 GBytes/sec bandwidth in a 65 nm logic process technology," *IEEE Journal of Solid-State Circuits (JSSC)*, vol. 44, no. 1, pp. 174–185, 2009.

[123] W. Zhang, K. C. Chun, and C. H. Kim, "Variation aware performance analysis of gain cell embedded DRAMs," in *Proc. ACM/IEEE International Symposium on Low Power Electronics and Design (ISLPED)*, 2010, pp. 19–24.

[124] K. C. Chun, P. Jain, J.-H. Lee, and C. Kim, "A sub-0.9V logic-compatible embedded DRAM with boosted 3T gain cell, regulated bit-line write scheme and PVT-tracking read reference bias," in *Proc. IEEE Symposium on VLSI Circuits (VLSIC)*, 2009, pp. 134–135.

[125] ——, "A 3T gain cell embedded DRAM utilizing preferential boosting for high density and low power on-die caches," *IEEE Journal of Solid-State Circuits (JSSC)*, vol. 46, no. 6, pp. 1495–1505, 2011.

[126] K. C. Chun, P. Jain, T.-H. Kim, and C. Kim, "A 1.1V, 667MHz random cycle, asymmetric 2T gain cell embedded DRAM with a 99.9 percentile retention time of 110 µsec," in *Proc. IEEE Symposium on VLSI Circuits (VLSIC)*, June 2010, pp. 191 –192.

[127] N. Ikeda, T. Terano, H. Moriya, T. Emori, and T. Kobayashi, "A novel logic compatible gain cell with two transistors and one capacitor," in *Proc. IEEE Symposium on VLSI Technology (VLSIT)*, 2000, pp. 168–169.

[128] M. Ichihashi, H. Toda, Y. Itoh, and K. Ishibashi, "0.5 V asymmetric three-tr. cell (ATC) DRAM using 90nm generic CMOS logic process," in *Proc. IEEE Symposium on VLSI Circuits (VLSIC)*, 2005, pp. 366–369.

[129] M.-T. Chang, P.-T. Huang, and W. Hwang, "A 65nm low power 2T1D embedded DRAM with leakage current reduction," in *Proc. IEEE SOC Conference (SOCC)*, 2007, pp. 207–210.

[130] Y. Lee, M.-T. Chen, J. Park, D. Sylvester, and D. Blaauw, "A 5.42nW/kB retention power logic-compatible embedded DRAM with 2T dual-VT gain cell for low power sensing applications," in *Proc. IEEE Asian Solid State Circuits Conference (A-SSCC)*, 2010, pp. 1–4.

[131] S. Hong, S. Kim, J.-K. Wee, and S. Lee, "Low-votage DRAM sensing scheme with offset-cancellation sense amplifier," *IEEE Journal of Solid-State Circuits (JSSC)*, vol. 37, no. 10, pp. 1356–1360, 2002.

[132] K. C. Chun, P. Jain, and C. Kim, "Logic-compatible embedded DRAM design for memory intensive low power systems," in *Proc. IEEE International Symposium on Circuits and Systems (ISCAS)*, 2010, pp. 277–280.

[133] M. Seok, D. Sylvester, and D. Blaauw, "Optimal technology selection for minimizing energy and variability in low voltage applications," in *Proc. ACM/IEEE International Symposium on Low Power Electronics and Design (ISLPED)*, 2008, pp. 9–14.

[134] A. Teman, O. Yadid-Pecht, and A. Fish, "Leakage reduction in advanced image sensors using an improved AB2C scheme," *IEEE Sensors Journal (JSEN)*, vol. 12, no. 4, pp. 773–784, 2012.

[135] A. Teman, A. Mordakhay, and A. Fish, "Functionality and stability analysis of a 400 mV quasi-static RAM (QSRAM) bitcell," *ELSEVIER Microelectronics Journal*, vol. 44, no. 3, pp. 236–247, 2013.

[136] M. Bauer, R. Alexis, G. Atwood, B. Baltar, A. Fazio, K. Frary, M. Hensel, M. Ishac, J. Javanifard, M. Landgraf, D. Leak, K. Loe, D. Mills, P. Ruby, R. Rozman, S. Sweha, S. Talreja, and K. Wojciechowski, "A multilevel-cell 32 Mb flash memory," in *Proc. IEEE International Solid-State Circuits Conference (ISSCC)*, 1995, pp. 132–133.

[137] M. Aoki, Y. Nakagome, M. Horiguchi, S. Ikenaga, and K. Shimohigashi, "A 16-level/cell dynamic memory," *IEEE Journal of Solid-State Circuits (JSSC)*, vol. 22, no. 2, pp. 297–299, 1987.

[138] T. Furuyama, T. Ohsawa, Y. Nagahama, H. Tanaka, Y. Watanabe, T. Kimura, K. Muraoka, and K. Natori, "An experimental 2-bit/cell storage DRAM for macrocell or memory-on-logic application," *IEEE Journal of Solid-State Circuits (JSSC)*, vol. 24, no. 2, pp. 388–393, 1989.

[139] T. Okuda and T. Murotani, "A four-level storage 4-Gb DRAM," *IEEE Journal of Solid-State Circuits (JSSC)*, vol. 32, no. 11, pp. 1743–1747, 1997.

[140] P. Gillingham, M. Incorp, and O. Kanata, "A sense and restore technique for multilevel DRAM," *IEEE Transactions on Circuits and Systems II (TCAS-II): Analog and Digital Signal Processing*, vol. 43, no. 7, pp. 483–486, 1996.

[141] B. Cockburn, J. Tapia, and D. Elliott, "A multilevel DRAM with hierarchical bitlines and serial sensing," in *Proc. IEEE International Workshop on Memory Technology, Design and Testing (MTDT)*, 2003, pp. 14–19.

[142] J. Koob, S. Ung, A. Rao, D. Leder, C. Joly, K. Breen, T. Brandon, M. Hume, B. Cockburn, and D. Elliott, "Test and characterization of a variable-capacity multilevel DRAM," in *Proc. IEEE VLSI Test Symposium (VTS)*, May 2005, pp. 189–197.

[143] J. Koob, S. Ung, B. Cockburn, and D. Elliott, "Design and characterization of a multilevel DRAM," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems (TVLSI)*, vol. 19, no. 9, pp. 1583–1596, Sept. 2011.

[144] B. Amrutur and M. Horowitz, "A replica technique for wordline and sense control in low-power SRAM's," *IEEE Journal of Solid-State Circuits (JSSC)*, vol. 33, no. 8, pp. 1208–1219, Aug. 1998.

[145] U. Arslan, M. McCartney, M. Bhargava, X. Li, K. Mai, and L. Pileggi, "Variation-tolerant SRAM sense-amplifier timing using configurable replica bitlines," in *Proc. IEEE Custom Integrated Circuits Conference (CICC)*, Sept. 2008, pp. 415–418.

[146] S. Komatsu, M. Yamaoka, M. Morimoto, N. Maeda, Y. Shimazaki, and K. Osada, "A 40-nm low-power SRAM with multi-stage replica-bitline technique for reducing timing variation," in *Proc. IEEE Custom Integrated Circuits Conference (CICC)*, Sept. 2009, pp. 701–704.

[147] P. Meinerzhagen, O. Andic, J. Treichler, and A. Burg, "Logic-compatible multilevel gain-cell-based DRAM for VLSI-SoCs," in *Proc. IEEE/IFIP International Conference on VLSI System-on-Chip (VLSI-SoC), PhD Forum*, Sept. 2010.

[148] E. Vittoz, *Low-Power Electronics Design*. CRC Press, 2004, ch. 16.

[149] H. Soeleman, K. Roy, and B. Paul, "Robust subthreshold logic for ultra-low power operation," *IEEE Transactions on VLSI Systems (TVLSI)*, vol. 9, no. 1, pp. 90–99, Feb 2001.

Pascal Meinerzhagen
PhD Student / Research Assistant at EPFL, Electrical Engineering
Intel Ph.D. fellowship student
(Successfully passed private Ph.D. defense on 12/16/2013)

EPFL STI IEL TCL
ELG 031 (Building ELG)
Station 11
CH-1015 Lausanne
Switzerland
pascal.meinerzhagen@epfl.ch
Phone: +41 21 69 31027
Mobile: +41 76 2455875
http://people.epfl.ch/pascal.meinerzhagen

Personal Details

Nationality: Swiss
Birth date: 18.11.1984
Fluent in English, German, French, and Spanish
Basics of Italian and Portuguese

Mission

- Enable **longer runtimes of biomedical implants and sensor nodes, as well as mobile computing devices**, by lowering the leakage power and the active energy of embedded memories beyond prior art and by employing advanced power management techniques
- Enable **higher integration densities and speed of wireless communications systems**, by exploiting their inherent resilience to hardware defects

Fields of Expertise

- Digital VLSI design flow
- Full-custom digital IC design
- PCB design and ASIC measurements
- Ultra-low voltage (ULV)/ultra-low power (ULP) systems
- Fault-tolerant/error-resilient VLSI systems
- ULV/ULP digital circuit design, logic synthesis, and backend design
- Design and characterization of custom standard-cell libraries (SCLs)
- Subthreshold/near-threshold memory design
- Gain-cell memory design

- Hybrid ReRAM-CMOS circuit design

Future Work/Research Interests

- Dynamically voltage scalable, error-resilient, embedded processors for ultra-low power consumption in the subthreshold / near-threshold domain and high-speed operation on demand
- Fine-grained, low-overhead power management techniques enabling fast, energy-efficient active-sleep and sleep-active transitions as well as charge recycling
- Demonstrating embedded memories with superior robustness, lower power, and higher density than SRAM in advanced CMOS nodes (28nm and below) and at low voltages
- Combining emerging technologies (e.g., memristors) with standard CMOS for smaller biomedical systems with higher energy efficiency

Education and Academic Positions

01/2009 – 12/2013 (successfully passed private Ph.D. defense on 12/16/2013)

PhD in Electrical Engineering and research assistant at EPFL, Lausanne, Switzerland (first 2 years at ETHZ, Zurich, Switzerland)

- Advisors: Prof. **Andreas Burg**; and Prof. **Yusuf Leblebici**
- Collaborations: Prof. **Joachim Rodrigues**, Lund University, Sweden; Prof. **Alexander Fish**, Bar-Ilan University, Israel; and (about to start) Prof. **Wayne Burleson**, University of Massachusetts, USA
- Within EPFL, involved in projects with Prof. **Yusuf Leblebici**'s group; Prof. **David Atienza**'s group; and Dr. **Pierre-Emmanuel Gaillardon** (from Prof. **Giovanni De Micheli**'s lab)
- Title of PhD dissertation: "*Novel Approaches toward Area- and Energy-Efficient Embedded Memories*"
- Achievements at a glance: **1** invited book chapter (under review); **25** journal and conference papers (out of which 4 under review); **2** patent applications; **8** keynotes, invited talks, and seminars; main advisor of **18** MSc/BSc students and interns; involved as designer or advisor in tape-out/measurement of **11** ASICs; reviewer for **11** international journals and conferences; Intel Ph.D. fellowship award ($35 000); and **2** best paper nominations

09/2006 – 09/2008

Master of Science (MSc) in Electrical Engineering from EPFL, and joint MSc degree in "Micro- and Nanotechnologies for Integrated Systems" from EPFL; Grenoble INP, France; and Politecnico di Torino, Italy

Master's Thesis: "*Design of a 12-bit low-power SAR A/D Converter for a Neurochip*", carried out jointly in Prof. **Sung-Mo "Steve" Kang**'s group at UC Merced as visiting researcher, under the guidance of Prof. **Shin-Il Lim**, and in Prof. **Yusuf Leblebici**'s laboratory

10/2003 – 09/2006

Bachelor of Science (BSc) in Electrical Engineering

Teaching Assistantships

Main advisor of 18 Master projects, semester projects, and internships

"*Full-Custom Digital, Semi-Custom Digital, and Full-Custom Analog Design Labs*", 2011-2013, EPFL

"*VLSI II: Design of Very Large Scale Integration Circuits*", 2009-2010, ETHZ

"*VLSI I: From Architectures to VLSI Circuits and FPGAs*", 2009-2010, ETHZ

"*C/C++ Programming*", 2005, EPFL

Research Assistantships

01/2011 – current

Telecommunications Circuits Lab (TCL), Institute of Electrical Engineering, EPFL, Lausanne, Switzerland, supported by the Swiss National Science Foundation (SNSF)

01/2009 – 12/2010

Integrated Systems Laboratory (IIS), Department Information Technology and Electrical Engineering, ETHZ, Zurich, Switzerland, supported by the Swiss National Science Foundation (SNSF)

Industry Experience

05/2013 – 07/2013, 3 months, full-time

Intern at Intel Corporation, **Intel Labs, Circuit Research Lab (CRL)** – head by Vivek De and Richard (Rick) Forand, Low Power Circuit Technology (LPCT) group – head by James (Jim) Tschanz, Hillsboro, OR, USA. Currently ongoing research projects:

- Energy-efficient power-gating schemes for microprocessor VLSI systems-on-chip (SoCs) based on charge recycling
- Modeling and simulation of switched-capacitor (SC) charge recycling circuits / charge pumps.

2007, 10 weeks, full-time

Internship at Asetronics AG, Berne, Switzerland, in the domain of PCB test engineering. Responsibilities:

- Reviewing state-of-the-art PCB testing technologies
- Developing models to calculate test cost

- Optimizing and enhancing existing PCB test strategies for reduced time and cost

2003, 5 weeks, full-time

Construction worker at WIRZ AG, Berne, Switzerland

Publications

Book Chapters

"*An Ultra-Low-Power Application-Specific Processor for Compressed Sensing*," Jeremy Constantin, Ahmed Dogan, Oskar Andersson, Pascal Meinerzhagen, Joachim Neves Rodrigues, David Atienza, and Andreas Burg, VLSI-SoC'12 book, IFIP Advances in Information and Communication Technology, edited by Ayse Coskun, Andreas Burg, Ricardo Reis, and Matthew Guthaus, published by Springer, **invited book chapter**

Journals (Peer-Reviewed)

"*Comparative Analysis of ReRAM-Based Non-Volatile Flip-Flop Topologies with Sub-VT Read and CMOS Voltage-Compatible Write*," I. Kazi, P. Meinerzhagen, P.-E. Gaillardon, D. Sacchetto, A. Burg, and G. De Micheli, IEEE Transactions on Circuits and Systems I (T-CASI), *in preparation*

"*Comparative Analysis of Energy, Area, and Failure Probability of Dynamic Latch Topologies*," P. Meinerzhagen, A. Bonetti, G. Karakonstantis, and A. Burg, IEEE Transactions on Very Large Scale Integration Systems (T-VLSI), *under internal review*

"*Area-Efficient Low-Density Parity Check (LDPC) Decoder with Refresh-Free eDRAMs*," P. Meinerzhagen, A. Bonetti, G. Karakonstantis, C. Roth, F. Gürkaynak, and A. Burg, IEEE Transactions on Circuits and Systems II (TCAS-II), *under review*

"*Replica Technique for Adaptive Refresh Timing of Gain Cell embedded DRAM*," A. Teman, P. Meinerzhagen, R. Giterman, A. Fish, and A. Burg, IEEE Transactions on Circuits and Systems II (TCAS-II), *accepted*

"*On the Impact of Body Biasing on the Retention Time of Gain-Cell Memories*," P. Meinerzhagen, A. Teman, A. Burg, and A. Fish, Journal of Engineering (JoE)

"*Exploration of Sub-VT and Near-VT 2T Gain-Cell Memories for Ultra-Low Power Applications under Technology Scaling*," P. Meinerzhagen, A. Teman, R. Giterman, A. Burg, and A. Fish, Journal of Low Power Electronics and Applications (JLPEA), **invited article**

"*Benchmarking of Standard-Cell Based Memories in the Sub-VT Domain in 65-nm CMOS Technology*," P. Meinerzhagen, Y. Sherazi, A. Burg, J. Rodrigues, in IEEE Journal on Emerging and Selected Topics in Circuits and Systems (JETCAS), June 2011

Conference Proceedings (Peer-Reviewed, Full ≥ 4-Pages Papers)

"Single-Supply High-Speed 3T Gain-Cell for Low-Voltage Low-Power Applications," R. Giterman, A. Teman, P. Meinerzhagen, A. Burg, and A. Fish, IEEE FTFC, *under review*

"An Area-Optimized Sub-VT Memory Using Full-Custom Storage Elements in 65 nm CMOS," O. Andersson, B. Mohammadi, P. Meinerzhagen, A. Burg, and J. N. Rodrigues, IEEE International Midwest Symposium on Circuits & Systems (MWSCAS), *under review*

"4T Gain-Cell with Internal-Feedback for Ultra-Low Retention Power at Scaled CMOS Nodes," R. Giterman, A. Teman, P. Meinerzhagen, A. Burg, and A. Fish, IEEE International Symposium on Circuits and Systems (ISCAS), *accepted*

"*FireBird: PowerPC e200 Based SoC for High Temperature Operation*," R. Cojbasic, Ö. Cogal, P. Meinerzhagen, C. Senning, C. Slater, T. Maeder, A. Burg, and Y. Leblebici, in Proc. IEEE Custom Integrated Circuits Conference (CICC), September 2013

"*Dual-VT 4kb Sub-VT Memories with <1 pW/bit Leakage in 65 nm CMOS,*" O. Andersson, B. Mohammadi, P. Meinerzhagen, A. Burg, and J. N. Rodrigues, in Proc. IEEE European Solid-State Circuits Conference (ESSCIRC), September 2013

"*ReRAM-Based Non-Volatile Flip-Flop with Sub-VT Read and CMOS Voltage-Compatible Write*," I. Kazi, P. Meinerzhagen, P.-E. Gaillardon, D. Sacchetto, A. Burg, and G. De Micheli, in Proc. IEEE International NEWCAS Conference, June 2013

"*Review and Classification of Gain Cell eDRAM Implementations*", Adam Teman, Pascal Meinerzhagen, Andreas Burg, and Alexander Fish, in Proc. IEEE Convention of Electrical and Electronics Engineering in Israel (IEEEI), November 2012, **invited paper**

"*A Successive Cancellation Decoder ASIC for a 1024-bit Polar Code in 180nm CMOS*," A. Mishra , A. J. Raymond, L. G. Amaru, G. Sarkis, C. Leroux, P. Meinerzhagen, A. Burg, and W. J. Gross, in Proc. IEEE Asian Solid-State Circuits Conference (A-SSCC), November 2012

"*A Sub-VT 2T Gain-Cell Memory for Biomedical Applications*," Pascal Meinerzhagen, Adam Teman, Anatoli Mordakhay, Andreas Burg, and Alexander Fish, in Proc. IEEE Subthreshold Microelectronics Conference, October 2012

"TamaRISC-CS: *An Ultra-Low-Power Application-Specific Processor for Compressed Sensing*," Jeremy Constantin, Ahmed Dogan, Oskar Andersson, Pascal Meinerzhagen, Joachim Neves Rodrigues, David Atienza, and Andreas Burg, in Proc. IFIP/IEEE International Conference on Very Large Scale Integration (VLSI-SoC), October 2012, **nomination best paper award**

"*A 500fW/bit 14fJ/bit-access 4kb Standard-Cell Based Sub-VT Memory in 65nm CMOS*," Pascal Meinerzhagen, Oskar Andersson, Babak Mohammadi, Yasser Sherazi, Andreas Burg, and Joachim Neves Rodrigues, in Proc. IEEE European Solid-State Circuits Conference (ESSCIRC), September 2012

"*Replica Bit-Line Technique for Embedded Multilevel Gain-Cell DRAM*," U. Khalid, P. Meinerzhagen, A. Burg, in Proc. IEEE International NEWCAS Conference, June 2012

"*Two-Port Low-Power Gain-Cell Storage Array: Voltage Scaling and Retention Time*," R. Iqbal, P. Meinerzhagen, A. Burg, in Proc. IEEE International Symposium on Circuits and Systems (ISCAS), May 2012

"*Synthesis Strategies for Sub-VT Systems*," P. Meinerzhagen, O. Andersson, Y. Sherazi, A. Burg, J. Rodrigues, in Proc. IEEE European Conference on Circuit Theory and Design (ECCTD), August 2011, **invited paper**

"*Design and Failure Analysis of Logic-Compatible Multilevel Gain-Cell-Based DRAM for Fault-Tolerant VLSI Systems*," P. Meinerzhagen, O. Andic, J. Treichler, A. Burg, in Proc. ACM/IEEE GLSVLSI, May 2011

"*A 15.8 pJ/bit/iter Quasi-Cyclic LDPC Decoder for IEEE 802.11n in 90 nm CMOS*," C. Roth, P. Meinerzhagen, C. Studer, A. Burg, in Proc. IEEE Asian Solid-State Circuits Conference (A-SSCC), November 2010

"*Towards generic low-power area-efficient standard cell based memory architectures*," P. Meinerzhagen, C. Roth, A. Burg, in Proc. IEEE International Midwest Symposium on Circuits & Systems (MWSCAS), August 2010, **nomination student paper contest**

Contributions to PhD Forums (Peer-Reviewed)

"*Standard-Cell Based Memories (SCMs): from Sub-VT to Error-Resilient Systems*," P. Meinerzhagen, **International Solid-State Circuits Conference (ISSCC)**, Student Research Preview, February 2012

"*Logic-Compatible Multilevel Gain-Cell-Based DRAM for VLSI-SoCs*," P. Meinerzhagen, O. Andic, J. Treichler, A. Burg, IEEE/IFIP VLSI-SoC, PhD Forum, September 2010

Theses

"Novel Approaches toward Area- and Energy-Efficient Embedded Memories," P. Meinerzhagen, Ph.D. Thesis, EPFL

"*Design of a 12-bit low-power SAR A/D Converter for a Neurochip*," P. Meinerzhagen, Master's Thesis, EPFL, August 2008

Patents

Pascal Meinerzhagen, Jaydeep Kulkarni, Muhammad Khellah, Jim Tschanz, Dinesh Somasekhar, and Vivek De, "*Apparatus for Dual Purpose Charge Pump*," *filed for patent application*

Jaydeep Kulkarni, Pascal Meinerzhagen, Dinesh Somasekhar, James Tschanz, and Vivek De, "*Apparatus for Charge Recovery during Low Power Mode*," *filed for patent application*

Keynotes, Invited Talks, and Seminars

P. Meinerzhagen, "*Charge Recovery Circuits for Energy-Efficient Active/Sleep Mode Transitions in Power-Gated Domains*," talk at Intel GmbH Braunschweig, Germany, September 6th, 2013

P. Meinerzhagen, and A. Burg, "*Challenges, Solutions, and Alternatives to SRAM for the Design of Embedded Memories*," **Keynote** speech at Swedish SoC Conference (SSoCC), May 7, 2013

P. Meinerzhagen, "*Embedded Memories Tailored for Ultra-Low Power and Error-Resilient VLSI Systems*," invited talk, Bar-Ilan University, Ramat Gan, Israel, March 6, 2013

P. Meinerzhagen, "*Odds of Gain-Cell based eDRAM for Future VLSI SoC Applications*," invited talk, SanDisk, Omer, Israel, March 5, 2013

A. Burg, P. Meinerzhagen, A. Dogan, J. Constantin, M. M. Sabry Ali, G. Karakonstantis, D. Atienza, L. Benini, "*Near- and Sub-Threshold Design for Ultra-Low-Power Embedded Systems*," keynote speech at Winter School on Design Technologies for Heterogeneous Embedded Systems (FETCH), Leysin, Switzerland, January 2013

P. Meinerzhagen, and A. Burg, "*A Standard-Cell Approach toward Simple and Robust Low-Power Embedded Memories,*" invited presentation at Workshop on Energy Efficient Electronics and Applications, Bar-Ilan University, Ramat Gan, Israel, November 2012

P. Meinerzhagen, "*Robust Low-Voltage/Low-Power Embedded Memories for Biomedical Systems*," talk, invited by Prof. Chulwoo Kim and Prof. Jongsun Park, Department of Electronics Engineering, Korea University, South Korea, May 2012

P. Meinerzhagen, "*Embedded Memories: from Sub-VT to Error-Resilient Systems*," lecture, invited by Prof. Kyeong-Sik Min, School of Electrical Engineering, Kookmin University, South Korea, May 2012

Awards

Oct. 2012: Received a prestigious, highly competitive "**Intel Doctoral Student Honor award**" for work on embedded memories; price value $35 000

Oct. 2012: Received a nomination for the "2012 IFIP/IEEE International Conference on Very Large Scale Integration (VLSI-SoC) best paper award"

Aug. 2010: Received a nomination for the "2010 IEEE International Midwest Symposium on Circuits and Systems (MWSCAS) student paper contest"

Professional Activities

Reviewer for the following journals and conferences:

- IEEE Transactions on Circuits and Systems I (TCAS-I)
- IEEE TCAS-I Special Issue on CICC 2013
- IEEE Transactions on Circuits and Systems II (TCAS-II)
- ELSEVIER Microelectronics Journal
- ELSEVIER Integration, the VLSI Journal – 2013
- IEEE Access
- IEEE International Symposium on Circuits and Systems (ISCAS)
- IFIP/IEEE International Conference on Very Large Scale Integration (VLSI-SoC)
- ACM/ECRC/IEEE GLS Very Large Scale Integration Conference (GLSVLSI)
- IEEE Norchip Conference
- Asia Symposium on Quality Electronic Design (ASQED) – 2013
- IEEE International Symposium on Bioelectronics and Bioinformatics (ISBB)

Member of TPC for ASQED 2013, Circuit and System Design track

Received invitation to serve as session chair at ASQED 2013