INTERNATIONAL ROADMAP FOR DEVICES AND SYSTEMS

INTERNATIONAL ROADMAP FOR DEVICES AND SYSTEMS

2017 EDITION

EXECUTIVE SUMMARY

THE IRDS IS DEVISED AND INTENDED FOR TECHNOLOGY ASSESSMENT ONLY AND IS WITHOUT REGARD TO ANY COMMERCIAL CONSIDERATIONS PERTAINING TO INDIVIDUAL PRODUCTS OR EQUIPMENT.

Wi-Fi® and Wi-Fi Alliance® are registered trademarks of Wi-Fi Alliance.

LTE™ is a Trade Mark of ETSI registered for the benefit of its Members and of the 3GPP Organizational Partners.

The IEEE emblem is a trademark owned by the IEEE.

"IEEE", the IEEE logo, and other IEEE logos and titles (IEEE 802.11™, IEEE P1785™, IEEE P287™, IEEE P1770™, IEEE P149™, IEEE 1720™, etc.) are registered trademarks or service marks of The Institute of Electrical and Electronics Engineers, Incorporated. All other products, company names or other marks appearing on these sites are the trademarks of their respective owners. Nothing contained in these sites should be construed as granting, by implication, estoppel, or otherwise, any license or right to use any trademark displayed on these sites without prior written permission of IEEE or other trademark owners.

Table of Contents

List of Figures

List of Tables

ACKNOWLEDGMENTS

International Roadmap Committee

Europe—Francis Balestra, Mart Graef, Bert Huizing

Japan—Yoshihiro Hayashi, Hidemi Ishiuchi

U.S.A.—Tom Conte-vice-chair, Paolo Gargini-chair

IEEE

Rebooting Computing and Standards Association, with special thanks to Erik DeBenedictis, Terence Martinez, Rudi Schubert, William Tonti, and Elie Track

Communications Society—Chi-Ming Chen

Electron Devices Society—Fernando Guarin and Terence Hook

The outstanding work by the members of the International Focus Teams is acknowledged in each of their roadmap chapters.

The chairs and co-chairs of these teams are as follows:

Application Benchmarking—Tom Conte

Systems and Architectures—Marilyn Wolf

Outside Systems Connectivity—Michael Garner

More Moore—Mustafa Badaroglu

Lithography—Mark Neisser

Factory Integration—Supika Mashiro and James Moyne

Yield—Slava Libman and Ines Thurner

Beyond CMOS—An Chen, Shamik Das and Matt Marinella

Emerging Research Materials—Eric Vogel

Packaging Integration—Dev Gupta

Metrology—George Orji

Environment, Safety, Health, and Sustainability—Leo Kenny and Steve Moffat

IRDS Project Manager

Linda S. Wilson

Special Acknowledgment

Alan K. Allan

OVERVIEW

1. INTRODUCTION

IRDS MISSION

Identify the roadmap of electronic industry from devices to systems and from systems to devices.

IRDS STRUCTURE

This initiative focuses on an International Roadmap for Devices and Systems (IRDS) through the work of International Focus Teams (IFT) closely aligned with the advancement of the devices and systems industries. Led by an international roadmap committee (IRC), IFTs collaborated in the development of the 2017 IRDS roadmap, and engaged with other segments of the IEEE, such as Rebooting Computing Initiative (RCI), Electron Devices Society (EDS), Computer Society (CS), Communication Society (ComSoc), and also with related industry communities like the System and Device Roadmap Japan (SDRJ) and the European SINANO Institute (ESI) in complementary activities to help ensure alignment and consensus across a range of stakeholders, such as:

- Academia
- Consortia
- Industry
- National laboratories

IEEE, the world's largest technical professional organization dedicated to advancing technology for humanity, through the IEEE Standards Association (IEEE-SA) Industry Connections (IC) program, supports the IRDS to ensure alignment and consensus across a range of stakeholders to identify trends and develop the roadmap for all the related technologies in the computer industry.

IEEE SPONSORS

The IRDS is sponsored by the IEEE Rebooting Computing (IEEE RC) Initiative in consultation and support from many IEEE Operating Units and Partner organizations including:

CASS—Circuits and Systems Society
CEDA—Council on Electronic Design Automation
CPMT—Components, Packaging and Manufacturing Society
CS—Computer Society
CSC—Council on Superconductivity
EDS—Electron Devices Society
MAG—Magnetics Society
NTC—Nanotechnology Council
RS—Reliability Society
SSCS—Solid State Circuits Society
SRC—Semiconductor Research Corporation

IEEE Standards Association

INTERNATIONAL SPONSORS AND COOPERATION

There are several international roadmap efforts directly aligned with the IRDS:

- European SINANO Institute (ESI) and NanoElectronics Roadmap for Europe: Identification and Dissemination" (NEREID) https://www.nereid-h2020.eu

- The System Device Roadmap Committee of Japan (SDRJ) https://sdrj.jp/
- The International Electronics Manufacturing Initiative (iNEMI) www.inemi.org

1.1. THE NEW ECOSYSTEM OF THE ELECTRONICS' INDUSTRY

1.1.1. MOORE'S LAW

For over 50 years the semiconductor industry has marched at the pace of Moore's Law. Transistor scaling associated with doubling the number of transistors every two years on the average has been and continues to be the unique feature of the semiconductor industry. As a consequence, as transistors became smaller they could also be switched from the off to the on state at faster rates while simultaneously became cheaper to manufacture. System integrators assembled new products utilizing the building blocks provided by the semiconductor industry, but they were barely able to complete assembling a new system when a yet new more powerful IC was becoming available. Any new technology generation enabled multiple new products with better performance than the previous ones. Integrated device manufacturers (IDM) in conjunction with software companies providing operating systems and applications were in full control of the pace at which the whole electronics industry ecosystem was moving forward. Therefore, past editions of technology roadmaps (i.e., National Technology Roadmap for Semiconductors (NTRS) and the International Technology Roadmap for Semiconductors (ITRS)) concentrated on forecasting the rate of transistor scaling and how transistor density and performance affected the evolution of integrated circuits.

During the past 10 years the advent of fabless design houses and foundries has revolutionized the way in which business is done in the new semiconductor industry, and because of this change system integrators have regained full control of the business model. This implies that system requirements are set at the beginning of any new product design cycle and step-by-step-related requirements percolate down through the manufacturing production chain to the semiconductor manufacturers. No longer does a faster microprocessor trigger the design of a new PC but on the contrary the design of a new smart phone generates the requirements for new ICs and other related components. In addition, fast approaching fundamental 2D topological limits have been threatening the ability of the semiconductor industry to continue scaling at historical rates. New very creative 3D transistor, memory cell, and overall IC structures are revolutionizing the way ICs are designed and produced.

1.1.2. THE COMPUTER INDUSTRY

The first arithmetic mechanical machine was invented by Blaise Pascal as far back as 1642 but the first machine encapsulating most of the elements of modern computers was introduces by Charles Babbage in 1837. ENIAC was introduced in 1946 and was the first fully electronic computer powered by vacuum tubes. By the early 60's IBM had established itself as a leader in transistorized computers in four product lines aimed at multiple applications, but, to merge these different lines in a more productive way, in 1964 it introduced the first general-purpose machine. The Model 360 could perform up to 34,500 instructions per second, with memory from 8 to 64 KB. Bipolar transistors were by far faster and more reliable than any MOS device and for the subsequent 30 years it was bipolar technology that powered the evolution of computers.

The personal computer industry began as a "hobby past-time" in the mid 70's. Apple was the most significant company shipping personal computers until IBM decided to enter this business. The IBM PC was introduced in 1981 with the support of Intel and Microsoft. As time went by personal computers became more powerful while large computers pushed the performance of bipolar transistors beyond their power limits despite the use of very sophisticated cooling techniques. By the mid 90's both industries relied on CMOS technology for both logic and memory products. However, once again power limits were reached by the middle of the first decade of the new century imposing severe limitations on performance (see section 1.1.5 and 1.1.6). New ways of performing a variety of computing functions have been demonstrated and more yet are under development. Among them neuromorphic computing, approximate computing, and most of all, quantum computing appears as the most promising architectures for some special applications.

1.1.3. SOC AND SIP

In the past 15 years the advent of the Internet; the extensive deployment of Wi-Fi base stations; consumer acceptance of a broad variety of cell phones and wireless mobile appliances, plus the successful combination of fabless companies working in conjunction with foundries has completely changed the electronics industry. System integrators are nowadays able to conceive, design, and realize any integrated circuit they wish without having to refer to integrated device manufacturers. System integrators can nowadays integrate multiple functionalities in a single chip called System on Chip (SOC) or by means of integrating multiple dice in a single package called System in Package (SIP) as opposed to connecting multiple standard specialized ICs on a board. It is clear that these methods of integration are more efficient and less costly than

acquiring several separate ICs (e.g., microprocessor, graphic processor, multiple memory types, USB, etc.) and assembling them on a board. In addition, the limited space available in mobile products has further accelerated the integration of multiple capabilities in a very confined environment. System integrators are by and large setting the pace of innovation for the electronics industry. The IC industry has also contributed to provide valuable technology building blocks to other industries that either did not exist or were in their infancy before and by adopting well-established technologies and brand-new devices like micro-mechanical systems (MEMS), flat panel displays, multiple sensors and so on have been realized. All these somewhat dissimilar technologies have been readily included into mobile appliances by means of heterogeneous integration. These new technology families were forecasted as far back as 2006 by the ITRS under the name of More than Moore (MtM).

1.1.4. POWER CHALLENGE

Each new technology generation has continued to produce smaller and better transistors that could switch faster than those produced with any previous technology generation. In the past this electrical feature of transistors (e.g., intrinsic transistor delay) enabled microprocessors to operate in each new generation at higher frequencies and therefore computer performance, as measured by industry benchmarks, (e.g., measured by millions of instruction executed per second (MIPS)), continued to improve at a very fast rates (almost doubling with each new technology generation) without any major change in computer architecture. In fact, the basic computer architecture has not changed though the years much since Von Neumann introduced its concept on how to perform computing in 1945. However, power consumption of integrated circuits kept on increasing until the beginning of the past decade when fundamental thermal limits were finally reached by some ICs. It became clear then that it had become practically impossible to keep on concurrently increasing both the frequency of operation and the number of transistors, one of the two features (i.e., either the frequency or the number of transistors) had to level off in order to make the ICs capable to operating under practical thermal conditions. Frequency was selected as the sacrificial victim and indeed it has stalled in the few GHz since the middle of the previous decade. New transistor designs and new architectures have been aimed at alleviating this problem especially in the past 5-10 years.

1.1.5. CONSEQUENCES OF FREQUENCY LIMITATIONS

These limitations on maximum useable frequency have impacted the rate of progress of the computer industry that has been compelled to develop such methods as complex software algorithms and clever instruction management to improve performance to partially compensate for the aforementioned conditions. The architecture of the microprocessors has changed from single core to multi-core. With this partitioning of the process architecture (very easy from a technology and layout point of view) each core can run in the few GHz range while the output rate is increased multifold by combining the output of multiple cores to produce the output signal. Unfortunately, this parallel type of solution cannot be used in all computational cases since some problems, or part of them, can only be solved in a serial way. However, these performance limitations did not impact the development and expansion of the Internet or that of mobile appliances. Cell phone and mobile devices in general operate close or below 2GHz due to the way operating frequencies are allocated in each country by very specific rules and therefore all of them have not been affected by the frequency limitation so far. This situation may change to some degree with the advent of 5G. (Refer to section 1.1.8.) Consumers began accessing the Internet via desktop appliances and then progressively got used to access it via mobile multipurpose appliances. Reaching any source of information via the Internet takes tens of milliseconds due to the speed at which signals can travel on any interconnect lines so microprocessors operating in the few GHz frequency range are more than adequate to handle the communication traffic.

1.1.6. INTERNET OF THINGS, INTERNET OF EVERYTHING (IOT, IOE)

The US Department of Defense awarded contracts as early as the 1960s for packet network systems, including the development of the ARPANET (which would become the first network to use the Internet Protocol.)

Access to the ARPANET was expanded in 1981 when the National Science Foundation (NSF) funded the Computer Science Network (CSNET). Since the mid-1990s, the Internet has had a revolutionary impact on culture and commerce, including the rise of near-instant communication by electronic mail, instant messaging, voice over Internet Protocol (VoIP) telephone calls, two-way interactive video calls, and the World Wide Web. This worldwide connectivity has created new phenomena like social networking and online shopping and banking. Increasing amounts of data are transmitted at higher and higher speeds over fiber optic networks operating at 1-Gbit/s, 10-Gbit/s, and beyond 40-Gbit/s. The Internet's takeover of the global communication landscape occurred almost instantly in historical terms: it only communicated 1% of the information flowing through two-way telecommunications networks in 1993; increasing to 51% by 2000, and more than 97% of the telecommunicated information by 2007! Today the Internet of Things continues to grow, driven by ever-greater amounts of online information, commerce, entertainment, and social networking, just to mention a few. Access to the Internet was originally done via hardwired desktop computers but the introduction of wireless

technology (Wi-Fi), smart phones in 2007 and tablets in 2010 has revolutionized the way people interact via the Internet. The world of communications has truly become a wireless, ubiquitous, and continuously interconnected world.

With this all said and done it is important to remember that the ever-expanding Internet of Everything (IoE) could not have happened if semiconductors had not powered a variety of communication devices, data centers, routers, and sensors. The advent of foundries and fabless companies enabled the customization of semiconductor products that now can cover all the aspects of IoE and it would be a mistake to assume that the semiconductor industry is by now a mature industry and it has not much more to offer. The new fabless/foundry ecosystem has opened the door to an endless flow of innovation available at very reasonable and affordable costs. The advent of the third phase of device integration (i.e., 3D power scaling) plus the many new capabilities associated with the introduction of revolutionary materials in the semiconductor industry will revolutionize how computers are built. New computers built with revolutionary architectures enabled by new devices will be surrounded from top to bottom with a variety of new sensorial capabilities and communications capabilities that will offer new and exciting options to system designers (see the Rebooting Computing section for more details).

1.1.7. 5G AND BEYOND IS COMING!

Cell phones began operation in the 90's using frequencies in the 800-900 MHz ranges in accordance with specifications of the Global System for Mobile Communications (GSM). These operational frequencies utilized by cell phones have increased multiple times and have now reached the present revision named 4G and LTE that operate in the 2,500-2,700MHz ranges. The adoption of a more powerful communication infrastructure under the name of 5G has been under discussion for the past few years. In 5G the utilization of frequencies ranging from 3 to 28GHz is under consideration. Operation in these frequency ranges is still well within the capabilities of ICs. In the past 10 years, cell phones and mobile appliances in general have become a viable means of accessing the Internet. Cell phone power consumption is typically below 5 Watts, so this value is well within the thermal limits of ICs operation. Most recently, access to the Internet via LTE or Wi-Fi has been continuously increasing since the areas of coverage are continuously extending; mobile appliances have become the most convenient means of communication and access to any source of information anywhere at any time. However, even though Wi-Fi and 4G/LTE have been developed with completely different market models and applications (i.e., Internet and cell phone, respectively) it is quite clear that both technologies are now interchangeably used, and they are both contending for access to the same range of frequencies. Is this a recipe for some type of unification and/or consolidation?

1.1.8. DATA CENTERS

The insatiable demand for information has led to the creation of gigantic clusters of servers and memory banks named 'data centers'. In this environment performance is still the name of the game and using complex cooling systems can mitigate power issues. Power consumption of data centers is rapidly escalating into the hundreds of megawatts range. Communications within the data centers and for long distances is handled via fiber optics because of their stellar low rate of attenuation. Adoption of multicore processors has also found the perfect application in servers used in data centers. In the past a separate bank of processors and memory aimed at a specific application had to be isolated on a specific rack because the operating system required by the application was different from the operating system used for other applications. Under these conditions utilization of processors and memory devices was very inefficient. The advent of multicore processors, however, offered the opportunity to "host" different operating systems in each of the cores residing within the same microprocessor and therefore leading to a dramatic increase in efficiency. Since multiple applications were now residing within the same processor it followed that the rate at which inlet of data could be handled by a single server was drastically increased; this led to higher requirements of the optical networks operating within a data center. To satisfy this requirement single mode fibers are now being implemented in data centers.

1.1.9. PRODUCT CONFLUENCE AND TECHNOLOGY FUSION

In the past ten years the CMOS technology has evolved and continuously delivered, generation-to-generation, reduced transistor size while burning less power/transistor. CMOS technology derivatives can be found from cell phone limited to 5-6W power dissipation to data center and large computer where power dissipation of few hundred watts can be managed. Will CMOS remain the only technology of choice for the foreseeable future? Different types of microprocessors operating at few gigahertz can be found in cell phones, in Wi-Fi and in large computers. Will 5G become the wireless technology of choice for cell phones and Internet devices? Similarly, the basic Von Neumann architecture where bits are moved back and forth between logic and memory still represent the architecture of choice.

Will this architecture remain the architecture of choice for any product in the future? After more than 50 years the foundation of the electronics industry still relies on the same basic pillars. Some people may say that this is the way it was, and it will always be the only viable way in the future. It would be unwise not to prepare for real changes in the years to come.

For this reason, the research community has been studying new logic and memory devices operating on completely new physical principles since the middle of the past decade. Similarly, new architectures have been explored for the past 10 years. Will tunnel transistors (TFET) and neuromorphic computing be the way of the future? It is in any case clear that devices and systems can no longer be independently developed. New and different products are nowadays driving the growth of the electronics industry and therefore the ITRS (i.e., bottom up) had to evolve into the IRDS where top down and bottom up requirements are conjunctly harmonized.

All of the above systems specifications dictate the requirements for the semiconductor industry. [Figure ES1].

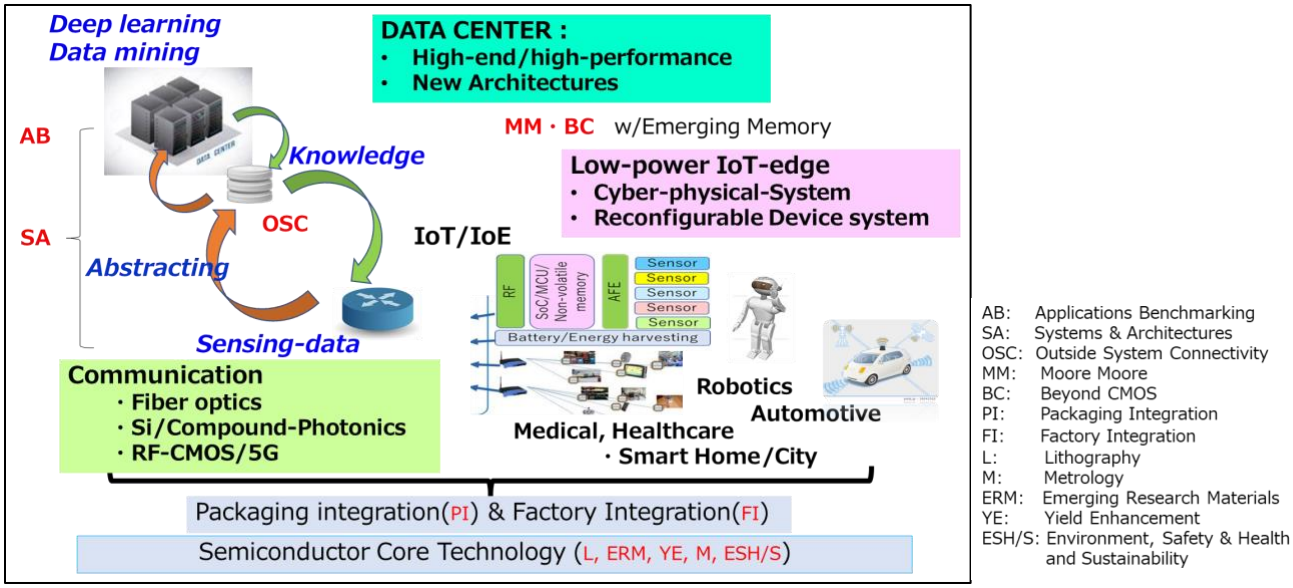

Figure ES1 The New Ecosystem of the Electronics' Industry based on Semiconductor Technologies

2. HISTORICAL EVOLUTION OF THE ROADMAP METHODOLOGY

from NTRS, to ITRS, and finally to IRDS

2.1. THE 3 ERAS OF SCALING

The foundations of the IC industry were laid out with the invention of the self-aligned silicon gate planar process in the late 1960's. Moore's predictions of the doubling of the number of transistors per die on an annual and then bi-annual pace formulated in 1965 and in 1975, respectively, in conjunction with Dennard's scaling guidelines led to the growth of the semiconductor industry until the beginning of the last decade.

Geometrical scaling characterized the 70's, 80's and 90's. This was the first generation of transistor scaling. The NTRS was initiated in the U.S. with a workshop held in 1991 and subsequent publications occurred in 1992, 1994, and 1997, respectively. The electronics industry was primarily "bottom up" driven during this period since any new technology generation provided transistors operating with continuously better performance that could power new memory and processors that easily fitted in the existing system architecture. System integrators could barely keep up with the rate at which new memory and new processors products were introduced since the industry changed in the 90's from a 3-4-year cycle to a 2-year technology cycle. However, major upcoming material and structural limitations were identified by the NTRS between 1994 and 1997. These problems were so fundamental that it was deemed necessary to engage the whole international semiconductor community to successfully identifying possible solutions in a timely fashion. Proposal to extend the NTRS to European, Japanese, Korean, and Taiwanese technical communities was presented in April 1998 to the World Semiconductor Council (WSC), the proposal was accepted and the ITRS was formed. The international research community met in San Francisco on July 1998 and at that meeting the research activities necessary to completely restructure the MOS transistor and the necessary methodology were approved and launched worldwide [Figure ES2].

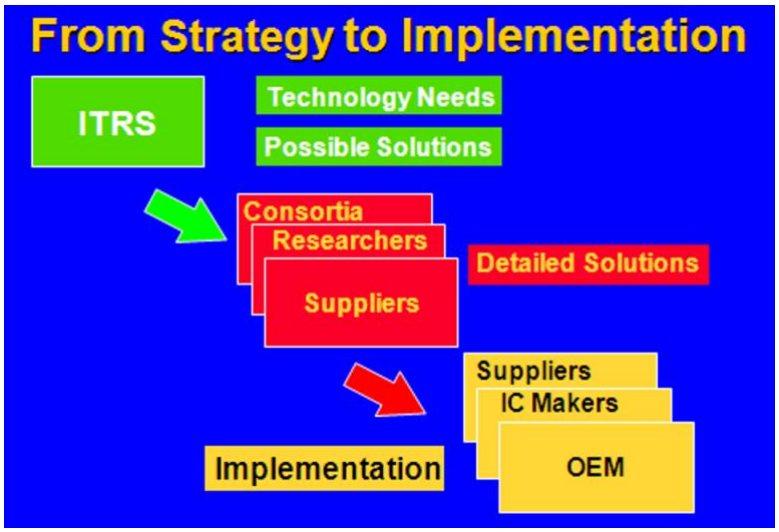

Figure ES2 1998 ITRS Program: From Strategy to Implementation

This new approach to restructuring the transistor was named 'equivalent scaling'. The goal of this program consisted in reducing the historical time of ~25 years between major transistor innovations to less than half in order to save the semiconductor industry from reaching a major crisis. Strained silicon, high-κ/metal gate, finFET, and use of other semiconductor materials (e.g., Germanium) represented the main features of this scaling approach [Figure ES3]. By 2011 all these new process modules were successfully introduced into high volume manufacturing [Figure ES4].

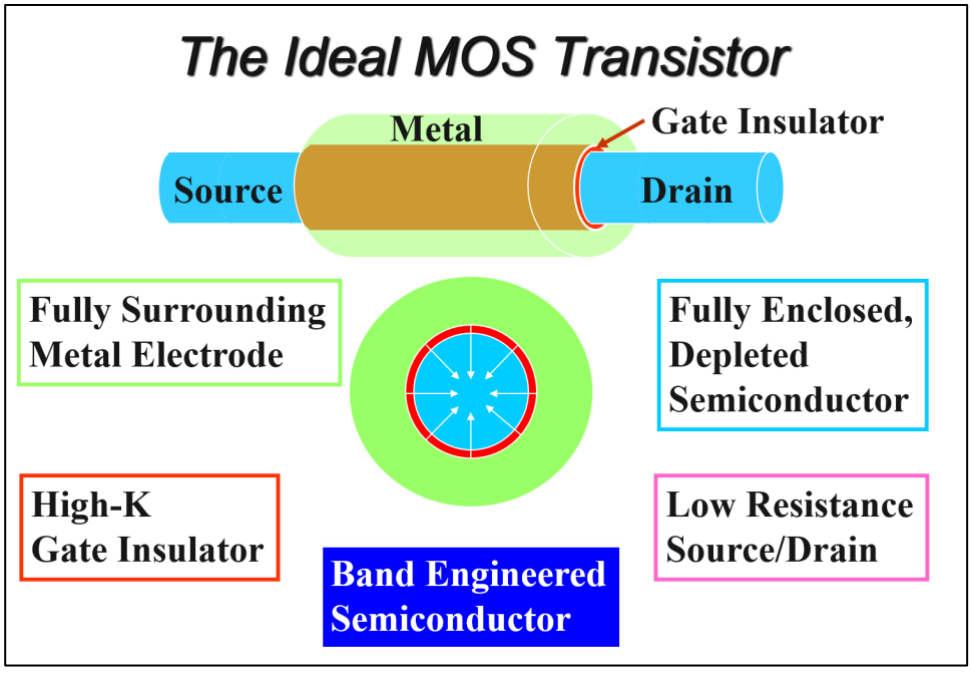

Figure ES3 Vision of the Completely Refurbished MOS Transistor

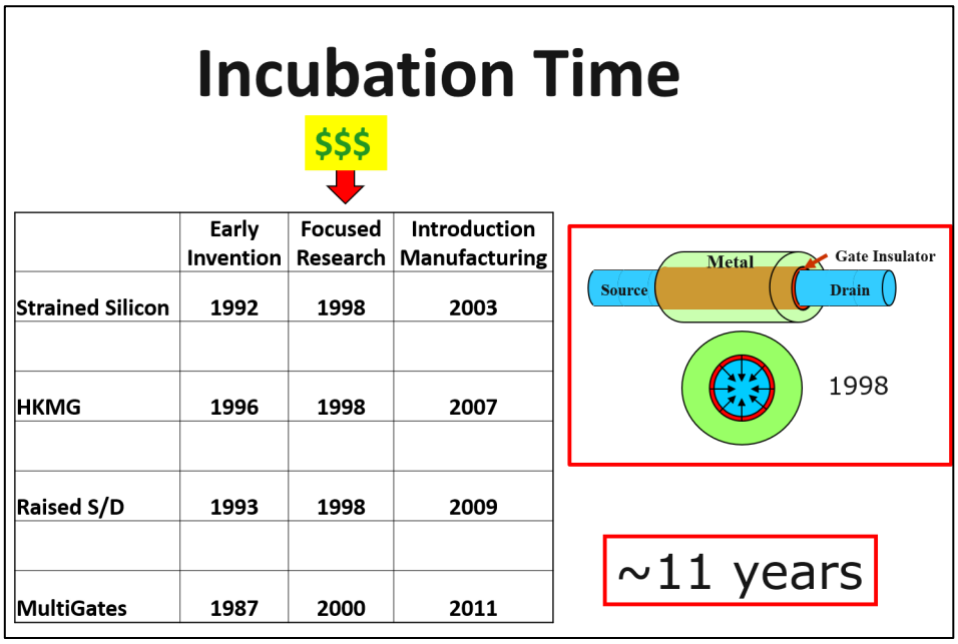

Figure ES4 From Strategy to Implementation in High-volume Manufacturing in Record Time

The advent and success of the combination of fabless design houses and foundries about 10 years ago revolutionized the way in which business was done and heralded the coming of the new semiconductor industry. Because of this environmental change system integrators finally regained full control of the business model. This implied that system requirements were going to be set at the beginning of any new product design cycle and step by step corresponding device requirements percolated down through the design/development/manufacturing production chain to the semiconductor manufactures. No longer was a faster microprocessor triggering the design of a new PC but on the contrary the design of a new smart phone generated the requirements for new ICs and other related components. Under these conditions it became clear in 2012 that the ITRS needed to adapt and morph to the new ecosystem [Figure ES5]. It was anticipated then that this transformation process would take some time and it was decided that the 2013 ITRS was going to be the last of its kind. Next, 2014 and 2015 were going to be dedicated to the construction of a new intermediate roadmap that was named ITRS 2.0.

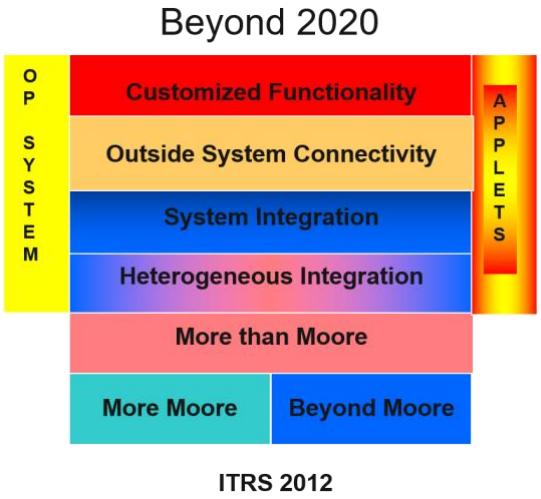

Figure ES5 The New Ecosystem of the Electronics Industry

During the preparation of the 2013 ITRS it was also assessed that horizontal (2D) features were going to be approaching the range of a few nanometers shortly beyond 2020 [Figure ES6] and so it became clear that the semiconductor industry was going to be running out of horizontal space by then! The question was: "Which products were going to be reaching

these 2D limits first?" Memory products have always been the leaders in transistor density (i.e., smallest feature pitch) and so it should have not been surprising to realize that the solution to this problem was to come first from companies producing Flash memories. In fact, multiple companies announced in 2014 that future products were going to fully utilize the vertical dimension [Figure ES7]. This is not too dissimilar from the approach taken in Manhattan, Tokyo, Hong Kong, or similarly highly crowded places to deal with space limitations: skyscrapers have become the standard approach to maximize "packing density". In addition, the rapid increase in the number of transistors (i.e., 2×/2-years) and the comparably rapid increase in operating frequency throughout the 80's and 90's drove the power dissipation of microprocessors way beyond the 100W by the 2003−2005 timeframe. This implied that number of transistors and frequency could no longer simultaneously increase. Under these conditions the electronics industry decided to convert to a multicore architecture, and continued to increase the number of transistors at historical rate but limited the operating frequency to few gigahertz. All the above considerations indicated that the structure of integrated circuits needed to evolve from 2D to 3D structures and transistor design needed to be aimed at reduced power consumption as opposed to be optimized for maximum operating frequency

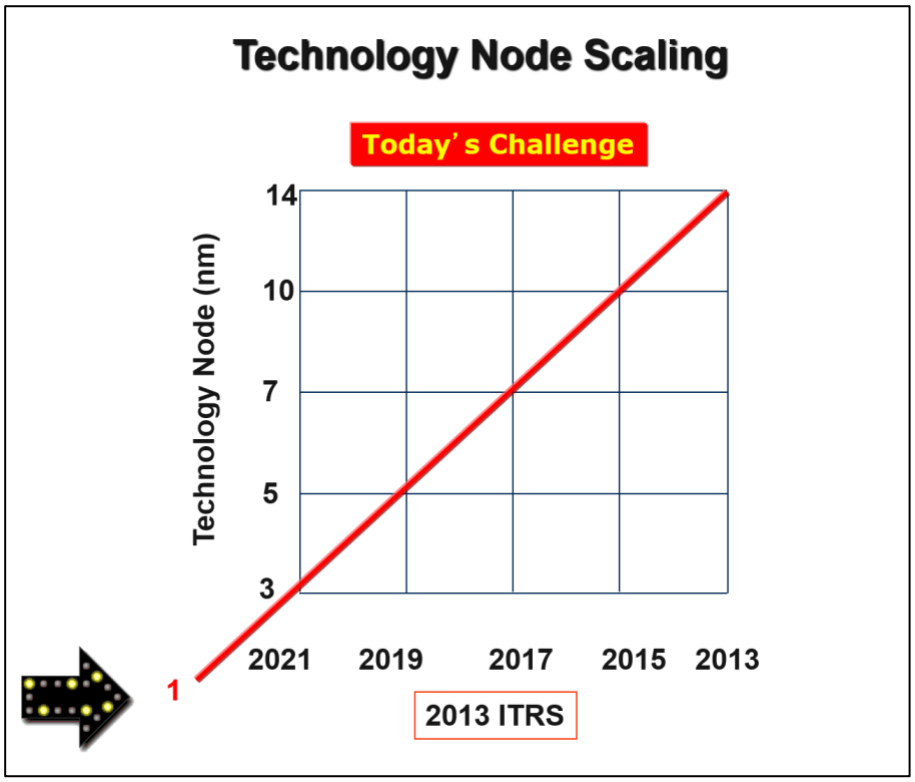

Figure ES6 2D scaling will reach fundamental limits beyond 2020

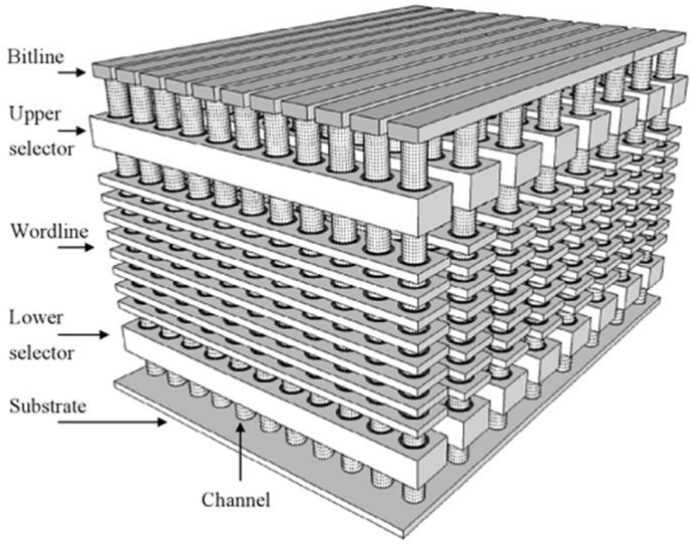

Figure ES7 Flash Memory Aggressively Adopts 3D Scaling in 2014

For the reasons described above, the new scaling method was named '3D Power Scaling' by the IRDS to symbolically include in a very succinct way all the challenges facing the semiconductor and electronics industries in the next 15 years [Figure ES8].

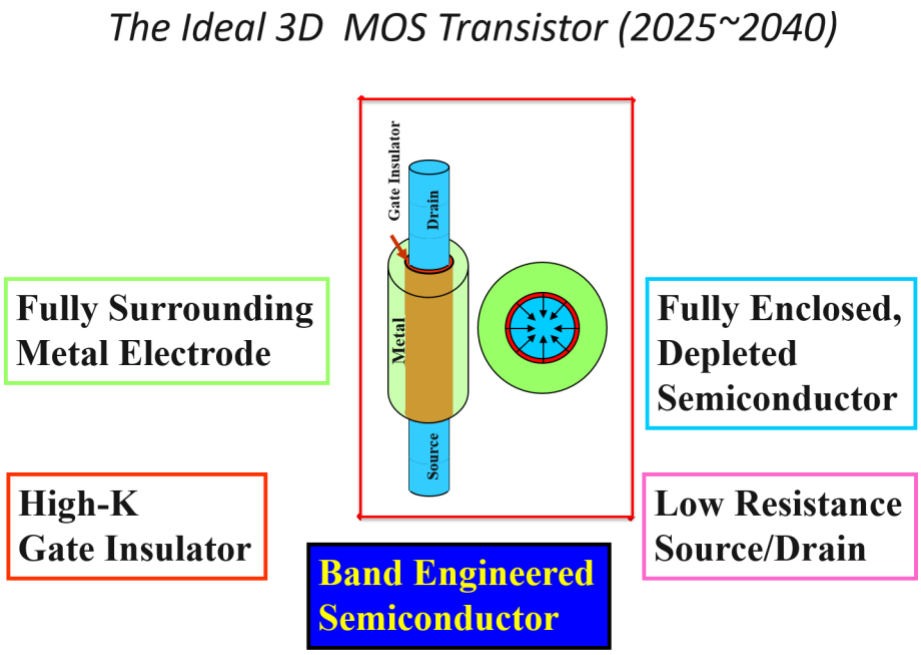

Figure ES8 The Ideal 3D Transistor

A summary of the 3 eras of transistor scaling is shown in Figure ES9.

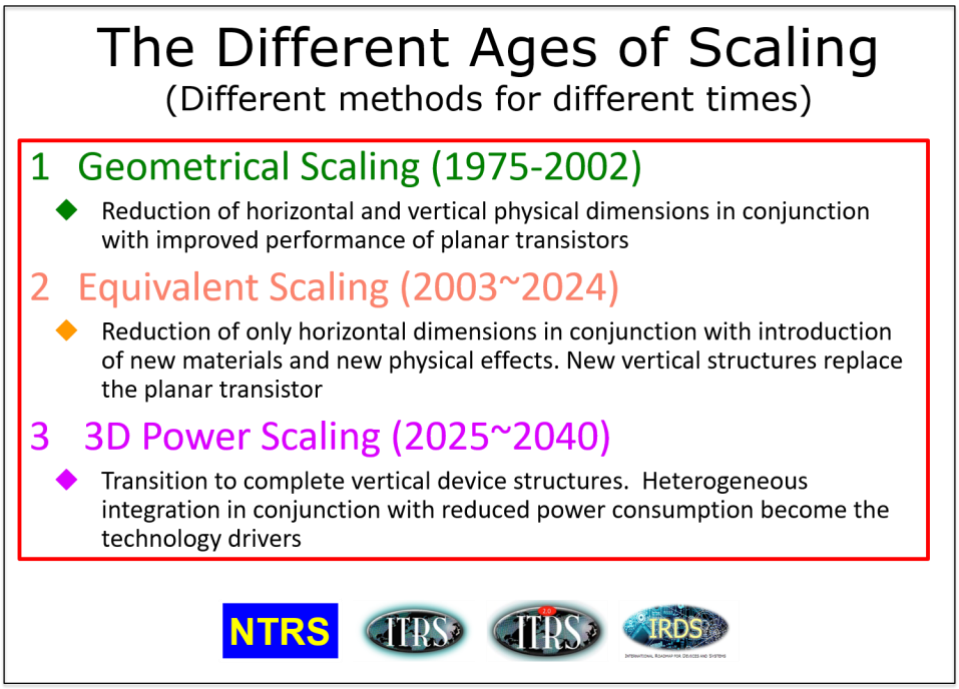

Figure ES9 The 3 Era of Scaling Heralded by NTRS, ITRS, ITRS 2.0, and IRDS

The considerations made it clear that it was no longer possible to design new transistors without keeping into account system requirements and so the International Roadmap for Devices and Systems (IRDS) came into being in May 2016 under the sponsorship of IEEE Rebooting Computing.

In the Beyond CMOS roadmap chapter the reader will find multiple new and exciting devices that, after 10 years of research, have already become or are soon becoming key players in the next decade. Integration of several memory circuits on top of logic circuits has been successfully demonstrated with consequent improvement in performance [Figure ES10]. This monolithic heterogeneous integration is made possible by the fact that these memory circuits can be fabricated at temperatures below 400°C. These temperatures are comparable to temperatures utilized nowadays to produce multiple interconnect lines on top of microprocessors.

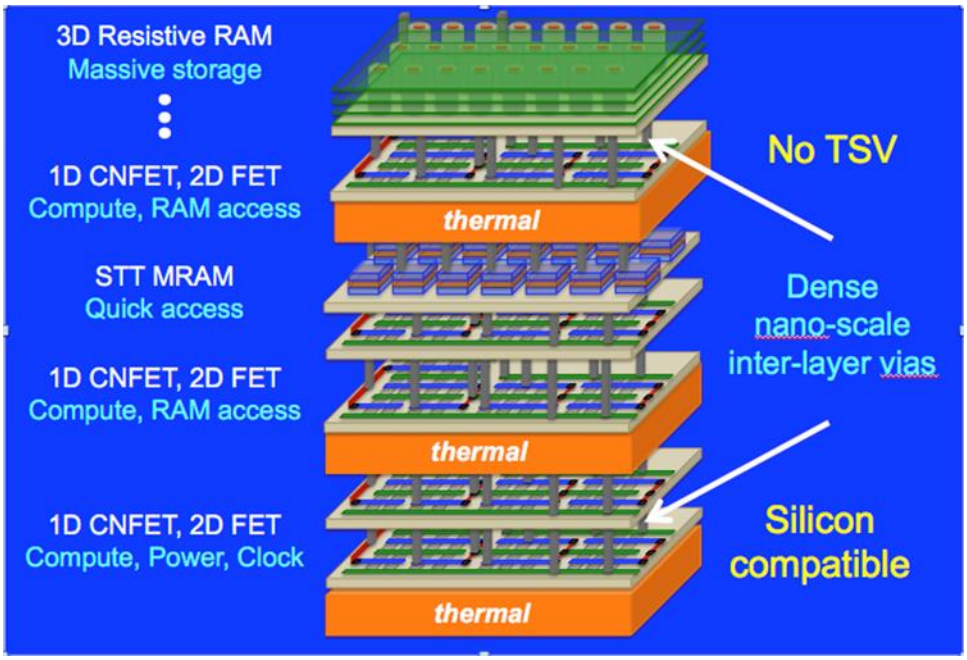

Figure ES10 Planning for the Advent of Monolithic Heterogeneous Integration

3. ROADMAP PROCESS AND STRUCTURE

3.1. ROADMAP PROCESS

The IRDS process is an evolution of the NTRS, ITRS, and ITRS 2.0 roadmapping process. The most relevant change consists in the fact that in the past device requirements were determined by the readily available technologies and system integrators were left with very few options on how to assemble their products. however, with the advent of the fabless/foundry ecosystem the system integrators regained the leadership position in establishing device requirements. To adjust to this new environment the IRDS process has been strengthen by a close association with the IEEE Rebooting Computing Initiative and by adding Applications Benchmarking and Systems and Architectures requirements to the 2017 IRDS roadmap process.

For the 2017 IRDS the Focus Teams are the following (The names link to each of the 2017 roadmap chapters.):

1. Application Benchmarking (AB)
2. Systems and Architectures (SA)
3. Outside System Connectivity (OSC)
4. More Moore (MM)
5. Beyond CMOS (BC)
6. Packaging Integration (PI)
7. Factory Integration (FI)
8. Lithography (L)
9. Emerging Research Materials (ERM)
10. Yield Enhancement (YE)
11. Metrology (M)
12. Environment, Safety, Health (ESH/S), and Sustainability

Additional roadmap documentation new for 2018 is a set on market drivers. The first to publish is the Medical Devices Drivers, which is provided to the document as part of the 2017 IRDS. Other market drivers' documents will be added as they become available to the IRDS report website. One can access this report at this link.

Furthermore, superconducting electronics and quantum computing are finally in the prototyping stage. A white paper on the subject is part of the 2017 IRDS and is found on the IRDS report website. A full chapter on this subject will be added as part of 2018 IRDS.

3.2. IRDS International Focus Teams (IFTs)

3.2.1. NEW! Application Benchmarking (AB)

The mission of the Applications Benchmarking (AB) IFT in the 2017 IRDS is to identify key application drivers, and to track and roadmap the performance of these applications for the next 15 years. The output of the AB market drivers in conjunction with the drivers of the Systems and Architectures (SA) IFT, generates a cross-matrix map showing which application(s) are important or critical (gating) for each market.

Historically, applications drive much of the nanoelectronics industry. For example, 10 years ago the PC industry put pressure on semiconductor manufacturers to advance to the next node in the roadmap. Today, as applications shift to the mobile market, it is again system manufacturers that are applying pressure for new technologies. The market for *Internet of Things edge devices* (IoT-e) has its own set of requirements and needs, including low cost and low energy consumption. Most of this is discussed in the Systems and Architectures (SA) 2017 roadmap chapter. It is the function of the AB chapter to step back from the current markets and their needs, and to consider the current and near-future application needs in each of these markets. For this reason, AB was created as new critical part of the 2017 IRDS.

3.2.2. Systems and Architectures (SA)

The mission of the System Architecture (SA) chapter in 2017 IRDS is to establish a top-down, system-driven roadmapping framework for key market drivers of the semiconductor industry in the 2017-2033 period. The SA chapter is proposing roadmaps of relevant system metrics for mobile applications, datacenter, IoT, and cyber-physical systems (CPS). The Systems and Architectures (SA) roadmap chapter of the 2017 IRDS roadmap constitutes a bridge between application benchmarks and component technologies. The systems analyzed in this chapter cover a broad range of applications of computing, electronics, and photonics. By studying each of these systems in detail, we can identify requirements for the semiconductor and photonics technologies that make these systems and applications possible.

The SA chapter considers four different types of systems: IoT edge (IoTe) devices provide sensing/actuation, computation, security, storage, and wireless communication. They are connected to physical systems and operate in wireless networks to gather, analyze, and react to events in the physical world. Cyber-physical systems provide real-time control for physical plants. Vehicles and industrial systems are examples of CPS. Mobile devices such as smartphones provide communication, interactive computation, storage, and security. For many people, smartphones provide their primary or only computing system. Cloud systems power data centers to perform transactions, provide multimedia, and analyze data. Cloud systems represent a trend towards common design principles and methodologies between traditional enterprise, high-performance scientific, and web-native computing.

3.2.3. Outside System Connectivity (OSC)

The mission of the OSC IFT in the 2017 IRDS consists in identifying and assessing capabilities needed to connect most elements of the Internet of Everything (IoE) and highlight technology needs and gaps. This includes supporting interconnection of a broad range of sensors, devices, and products to support information communication, processing and analysis for many applications including automobiles, aerospace, and a wide range of IoT applications for personal use, home, transportation, factory, and warehouse. Communication of data over fiber optic circuits to data centers and fiber optic communication within data centers is in scope for this chapter.

3.2.4. More Moore (MM)

The More Moore (MM) IFT of 2017 IRDS provides physical, electrical and reliability requirements for logic and memory technologies to sustain More Moore (Power, performance, area, cost (PPAC) scaling for big data, mobility, and cloud (IoT and server) applications and forecasts logic and memory technologies (15 years) in main-stream/high-volume manufacturing (HVM). The 2013 ITRS already anticipated that fundamental limits of 2D scaling were going to be reached for all product lines between 2015 and 2021. Flash products have been the technology leaders in pitch scaling since the mid-70's and they have already overcome the 2D limitations by aggressively implementing 3D memory cell strictures—already 72-96 layers of Flash memory cells have been demonstrated. It is anticipated that logic technologies will transition

to 3D approaches in the next few years. These technological solutions will assure continuation of Moore's Law for an additional 10−15 years

3.2.5. BEYOND CMOS (BC)

The goal of the Beyond CMOS (BC) IFT of the 2017 IRDS is to survey, assess, and catalog viable new information processing devices and system architectures due to their relevance on technological choices. It is also important to identify the scientific/technological challenges gating their acceptance by the semiconductor industry as having acceptable risk for further development. Another goal is to pursue long-term alternative solutions to technologies addressed in More-than-Moore (MtM) entries. This is accomplished by addressing two technology-defining domains: 1) extending the functionality of the CMOS platform via heterogeneous integration of new technologies, and 2) stimulating invention of new information processing paradigms. It is important to notice that many new memory devices identified by BC in past roadmaps have already been successfully demonstrated and are making their way into manufacturing by means of heterogeneous monolithic integration [Figure ES10].

3.2.6. PACKAGING INTEGRATION (PI)

The Packaging Integration (PI) focus of the 2017 IRDS is divided between the near-term assembly and packaging roadmap requirements and the introduction of many new requirements and potential solutions to meet market needs in the longer term. Packaging integration is the final manufacturing process transforming semiconductor devices into functional products for the end user. Packaging provides electrical connections for signal transmission, power input, and voltage control. It also provides for thermal dissipation and the physical protection required for reliability. Packaging is a limiting factor in both cost and performance for electronic systems. This is stimulating an acceleration of innovation. Heterogeneous integration of multiple technologies has become a dominant factor in the past 10 years enabling a variety of new products, especially in the mobile category. Design concepts, packaging architectures, materials, manufacturing processes, and systems integration technologies are all changing rapidly. This accelerated pace of innovation has resulted in development of several new technologies and the expansion and acceleration of others introduced in prior years. Wireless and mixed-signal devices, biochips, optoelectronics, and MEMS are placing new requirements on packaging and assembly as these diverse components are introduced as elements for System-in-Package (SiP) architectures.

3.2.7. FACTORY INTEGRATION (FI)

The Factory Integration (FI) focus area of 2017 IRDS is dedicated to ensuring that the semiconductor-manufacturing infrastructure contains the necessary components to produce items at affordable cost and high volume. Realizing the potential of Moore's Law requires taking full advantage of device feature size reductions, new materials, yield improvement to near 100%, wafer size increases, and other manufacturing productivity improvements. This in turn requires a factory system that can fully integrate additional factory components and utilize these components collectively to deliver items that meet specifications determined by other IRDS focus areas as well as cost, volume, and yield targets.

3.2.8. LITHOGRAPHY (L)

The Lithography (L) focus area of patterning technology has been high-performance logic chips, DRAM memory, and Flash memory. New capabilities to shrink dimensions enabled smaller devices that performed better than the previous generation. Once a new lithography technology became available, it was adopted for both memory and logic, perhaps with slightly different timing. But now devices are small enough that just shrinking them can give unacceptable electrical performance. In addition, the necessity to use as many as 4 mask exposures to pattern a single layer is driving manufacturing costs towards levels no longer affordable. These considerations have driven the industry towards major innovations in device design to avoid these effects and increasing the number of devices per unit of silicon area. These innovations in device and circuit design that have taken different forms for different types of devices as outlined in the 2017 IRDS. By 2014 it became clear that Flash memory products were becoming unreliable, as the physical size of the gate where the bits were stored was becoming too small (i.e., only few electrons could be stored). As a result, the IRDS forecasts that planar (or 2D) flash memory will stop shrinking critical dimensions (CDs), and the smallest CD devices will have a critical dimension of about 12 nm half-pitch. Several dimensions may actually undergo a process of reverse scaling and become actually larger from one technology generation to another. The industry is moving to three-dimensional (3D) flash memory to enable improved bit density on chips. It is expected that this trend will extend to logic chips beyond 2020.

3.2.9. EMERGING RESEARCH MATERIALS (ERM)

The mission of the Emerging Research Materials (ERM) IFT of the 2017 IRDS consists in aligning requirements for new materials with the needs of several IRDS working groups. The chapter emphasizes strategic difficult challenges and/or enabling of novel, breakthrough, and potentially disruptive opportunities for emerging material properties, synthetic methods, and metrology, organized in three areas: 1) *scaled technology materials needs for More Moore:* transistors,

memory, interconnects, lithography, heterogeneous integration, assembly and packaging; 2) *novel materials for beyond CMOS:* emerging logic and information processing devices, emerging memory and storage devices, and novel computational paradigms and architectures, and 3) *potentially disruptive material opportunities for functional scaling and convergent applications:* heterogeneous components, outside system connectivity, and high impact application areas such as energy, environment, agriculture, health, medical, etc.

3.2.10. YIELD ENHANCEMENT (YE)

The Yield Enhancement (YE) focus area is dedicated to activities ensuring that semiconductor manufacturing is optimized for production of the maximum number of functional units. Identifying, reducing, and avoiding relevant defects and contamination that can adversely affect and reduce overall product output are necessary to accomplish this goal. Yield in most industries has been defined as the number of functional and sealable products made divided by the number of products that can be potentially made. In the semiconductor industry, yield is represented by the functionality and reliability of integrated circuits produced on the wafer surfaces. During the manufacturing of ICs yield loss is caused, for example, by defects, faults, process variations, and design. The relationship of defects and yield, and an appropriate yield to defect correlation, is critical for yield enhancement. The Yield chapter of 2017 IRDS presents the current advanced and next generation future requirements for high-yielding manufacturing of More Moore as well as More than Moore products separated in "critical process groups" including MEMS, and back-end processes (e.g., packaging). Consequently, an inclusion of material specifications for Si, SiC, GaN, etc. is considered.

3.2.11. METROLOGY (M)

The Metrology Chapter (M) of the 2017 IRDS identifies emerging measurement challenges from devices, systems, and integration in the semiconductor industry and describes research and development pathways for overcoming them. This includes, but is not limited to, measurement needs for extending CMOS, beyond CMOS technologies, novel communication devices, sensors and transducers, materials characterization and structure/function relationships. This also includes metrology required in research and development, and techniques providing process control in manufacturing, yield, and failure analysis. With device feature sizes projected to decrease to less than 5 nanometers within the next 10 years, scaling as we know it is expected to soon reach its physical limits or get to a point where cost and reliability issues far outweigh the benefits. Already transistors for chips at the 7-5 nm nodes have already been demonstrated. The adoption of complex 3D structures fabricated using new materials and processes with ever decreasing dimensions are also projected to make their way into manufacturing within the next few years. The metrology roadmap addresses some of the measurement science challenges caused by these new developments and aims to provide a long-term view of the challenges, potentials solutions, technology, tools, and infrastructure needed to characterize new devices and materials for process control, and manufacturability.

3.2.12. ENVIRONMENT, SAFETY, HEALTH AND SUSTAINABILITY (ESH/S)

The Environmental, Safety, Health, and Sustainability (ESH/S) chapter of the 2017 IRDS serves to provide a long-range framework and process for all key stakeholders in the semiconductor and microelectronics industry, to develop proactive technical solutions to address critical ESH/S challenges up front, without gating industry R&D, mitigating cost, ensuring business continuity, and identifying key new markets and opportunities. This current version of the ESH/S Chapter reflects that *transitional nature of the technology roadmap itself,* from the previous ITRS to the new scope and vision of the IRDS. Reflecting this fundamental shift, this 2017 edition of the ESH/S chapter is primarily grounded on the work done by the transitional team in 2016, to form a basis for a substantial rewrite of the next roadmap update in 2018. However, there are several significant additions in this edition. First, we have formally added the area of 'sustainability,' given the increasing constraints posed by natural resources (water, energy), and materials usage. We will also include the topic of governance, in the context of how processes and systems are managed and reported. Note that broader sustainability topics typically included in standard reporting (such as fair labor practices and social responsibility that are not directly related to technology and operations), are considered out of the scope of this roadmap chapter.

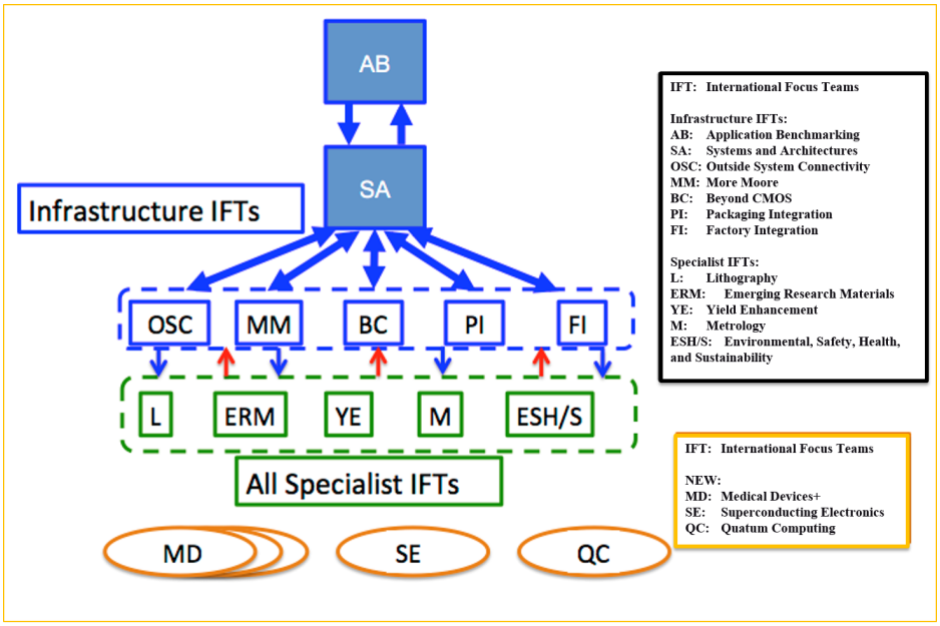

Figure ES11 IFT Structure of the IRDS

4. Overall Roadmap Drivers—ORSC and ORTC

4.1. System Performance Considerations

System performance is determined by the harmonious confluence of hardware, architecture, and software algorithms. In the PC era any hardware upgrades easily "fit in" the well-established system architecture and dramatically boosted up system performance. The NTRS and the ITRS followed the impact of any new technology generation "up the food chain" to report continuous system performance improvements. However, the imbalance between the speeds at which processor and memory devices operated compelled the migration of ever-larger amount of cash memory on the processor chip.

However, this solution was not good enough and further worsening the situation was that superscalar microarchitectures were enabling higher frequencies though deeper pipelines. This meant more instructions needed to be "in flight" than was possible by waiting for branch instructions to execute. This lead to *speculative execution*: predicting what path a program would take and then doing that work ahead of time, in parallel. Thus, higher frequencies meant deeper pipelines, which in turn required more and more speculatively executing instruction. But no prediction is 100% accurate. Invariably, these microprocessors did a lot of extra, wasted work by miss-speculation. The deeper the pipeline, the more power was wasted on these phantom instructions.

In the middle of the past decade, microprocessor's power dissipation reached fundamental limits and operational frequency could no longer be increased even though transistor performance could have easily allowed circuits to operate in the tens of GHz and above. These power limitations compelled a dramatic change in processor architecture to multicore in an attempt to partition data fetch and computation tasks among several cores that could then operate almost independently. Nowadays multiple new architectures are being explored and especially tailored for specific applications.

The restructuring of the roadmap process to ITRS 2.0 was initiated in 2014 and published early in 2016 (www.itrs2.net).

In the same year the roadmap was further restructured under IEEE and this process led to the 2017 IRDS. This edition of the IRDS is a first attempt to formulate and harmonize system and device requirements in a comprehensive and synergistic way. This explains why new and more powerful system indicators needed to be generate in parallel to logic and memory indicators.

4.1.1. Application Benchmarking and Systems and Architectures

The new requirements outlined above led to the formation of the Application Benchmarking (AB) IFT and the expansion of the System Integration IFT to also include architectural elements and thus became the Systems and Architectures (SA)

IFT. Much more information on these subjects can be found in the AB and SA chapters but it is worth including one example that epitomizes the new comprehensive approach of the IRDS.

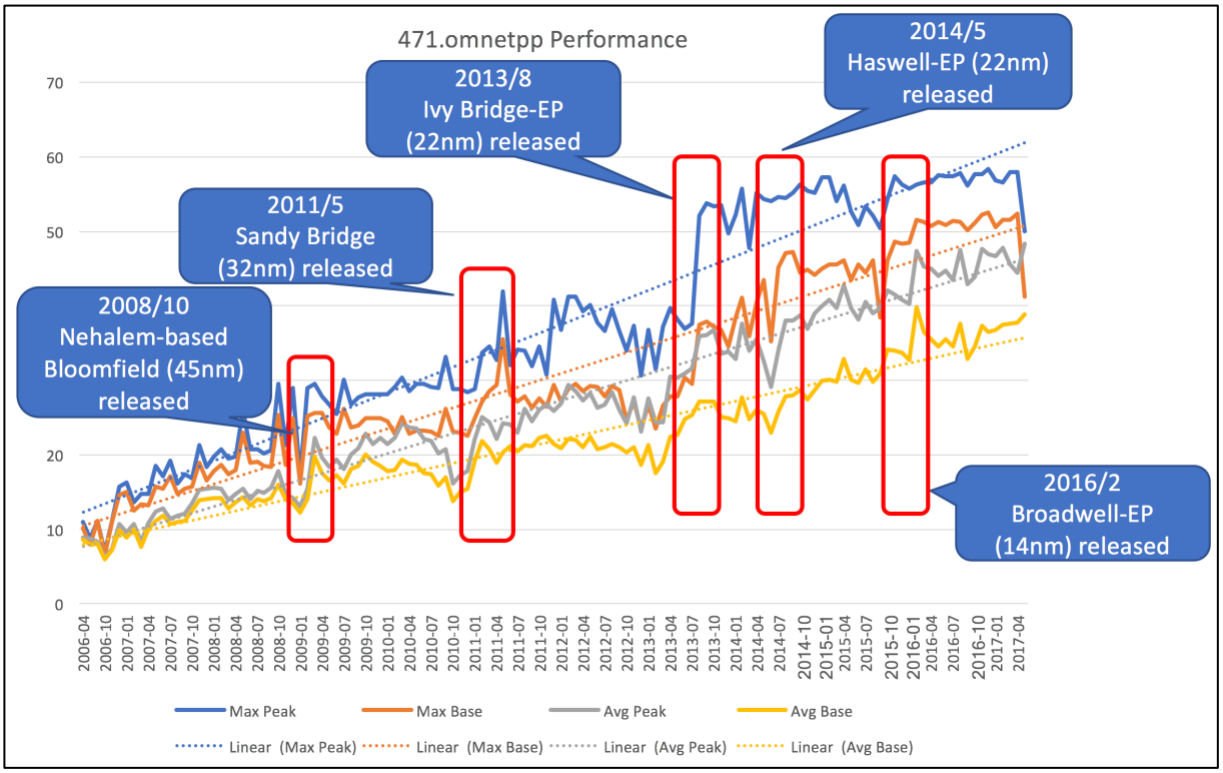

Figure AB7 Historical Performance Over Time of 471.omnetpp Benchmark

Figure AB-7 from the Application Benchmarking chapter shows the performance of 471.omnetpp over time. Discrete Event Simulation (DES) models the operation of a system as a discrete sequence of events in time. The plot indicates both base and peak performance with max and average scores per monthly bins, represented by the four solid lines. The linear regression of each metric is also represented by the four dotted lines. In general, performance appears to be improving over time, but it should be noted that there are occasional steep jumps in performance. These generally align well with major CPU releases from Intel as depicted in the figure. One thing to note is that the data uploaded to SPEC only has the test date associated with the data, not the system release date. While, in theory, it is possible to look up all the system release dates for all 8,600 data points, we assumed that, in general, newer systems would be more popular to benchmark at any given time, and, thus, the test dates would align well with the system release dates.

The Ivy Bridge-EP processor released around this timeframe is the first processor to have a 25MB last-level cache, which is enough to fit the entire working set of 471.omnetpp. This could also explain why the max peak performance of the benchmark has remained somewhat flat after that point. Of course, the whole domain of DES should not be restricted to this single benchmark with a limited size input set. However, being memory bound and cache sensitive would be a general characteristic to describe DES workloads.

4.2. Overall Roadmap Systems and Technology Characteristics (ORSC and ORTC)

Providing a simple top-down view of the new very complex ecosystem it is not an easy task and it is therefore necessary to have a simplified depiction of what is under study including some basic definitions. The 2017 IRDS has developed several summary tables to capture all these elements and the reader will be able to find detailed information in the chapters addressing these subjects.

The executive summary presents a succinct overview of the above elements progressing from system requirement all the way to device specifications.

Table ES1 Overall Roadmap System Characteristics

2017 IRDS Executive Summary - ORSC							
YEAR OF INTRODUCTION	2018	2019	2021	2024	2027	2030	2033
Cloud Computing (CC)							
# Cores per Socket [1]	32	38	46	58	70	70	70
Processor Base Frequency (for multiple cores together) [2]	2.75	3.00	3.20	3.50	3.80	4.10	4.40
L1 Data Cache Size (in KB) [3]	36	36	38	42	44	44	44
L1 Instruction Cache Size (in KB) [4]	48	48	64	128	160	160	160
HBM Bandwidth (TB/s) [5]	2	2.4	6	6.6	6.6	6.6	6.6
Socket TDP (Watts)	215	226	249	288	334	387	425
SA Mobile Table - Focus Drivers Line Items							
# CPU cores	8	10	12	18	25	28	30
# GPU cores	16	16	32	64	128	256	512
Max Freq (GHz)	2.5	2.8	3.3	4.4	5.9	7.8	10.4
Cellular Data rate (Mb/s)	13	22	1000	1000	1000	1000	1000
# Sensors	4	6	8	12	12	16	16
Board Power (mW)	4900	5096	5618	6504	7529	8716	10090
SA IoT Table - Focus Drivers Line Items							
CPUs per device	1	1	2	4	6	8	8
Max CPU Frequency (MHz)	255	277	305	320	335	352	369
Energy Source (B = battery, H = energy harvesting)	B+H	B+H	B+H	B+H	B+H	B+H	B+H
Sensors per device	4	4	8	12	16	16	16
SA CPS Table - Focus Drivers Line Items							
Number of Devices	64	64	64	128	256	512	512
CPUs per Device	4	4	8	8	12	16	16

Notes for Table ES1:

[1] Required for SpecINT-Rate scaling. See the AB roadmap chapter for more information.
[2] Frequency increase slowing down, but increases because of better cooling (allowing higher TDP))
[3] Load-to-use latency dictates constant or limited growth in L1 data cache size
[4] Instruction footprint for cloud apps going up (refer Google data warehouse paper)
[5] HBM: 128GB/s per port, in sockets 2015, HBM2: 256GB/s per port, can be in sockets 2017, HBM3: 512GB/s, can be in sockets 2019

Table ES1 summarizes some of the major system characteristics in cloud computing, mobile, IoT, and cyber-physical systems.

It can be noticed that in all cases the number of CPU or GPU cores continues to increase throughout the timer horizon. In all cases it is expected that the amount of data elaboration will continue to increase and actuality it does not seem that there is any limit on these requirements. Frequency of operation will continue to increase as well but only at a moderate rate except for mobile systems where the strong demand for wider bandwidth can only be satisfied if the operational frequency is increase. It may be observed however that this increase in frequency can only be tolerated if power dissipation is kept at very low levels by careful power management.

Table ES2 summarizes the major technology characteristics of logic and memory devices. For convenience the traditional industry "Node Range" Labeling is indicated even though in the past it only closely tracked the half-pitch of nonvolatile memory (NVM) products. The NTRS, ITRS and IRDS definition of technology naming based on half-pitch of gate (traditionally polysilicon based) and metal are reported [Figure ES12].

Table ES2 Overall Roadmap Technology Characteristics

2017 IRDS Executive Summary - ORTC							
YEAR OF PRODUCTION	2017	2019	2021	2024	2027	2030	2033
Logic device technology naming	P54M36	P48M28	P42M24	P36M21	P32M14	P32M14 T2	P32M14 T4
Logic industry "Node Range" Labeling (nm)	"10"	"7"	"5"	"3"	"2.1"	"1.5"	"1.0"
Logic device structure options	finFET FDSOI	finFET LGAA	LGAA finFET	LGAA VGAA	LGAA VGAA	VGAA, LGAA, 3DVLSI	VGAA, LGAA, 3DVLSI
LOGIC CELL AND FUNCTIONAL FABRIC TARGETS							
Average cell width scaling factor	1.00	0.90	0.90	0.90	0.90	0.90	0.90
LOGIC DEVICE GROUND RULES							
MPU/SoC Metalx ½ Pitch (nm) [1,2]	18	14	12	10.5	7.0	7.0	7.0
Physical gate length for HP Logic (nm) [3]	20	18	16	14	12	12	12
Lateral GAA (nanosheet) Minimum Width (nm)			7.0	7.0	6.0		
Minimum Device Width (fin, nanosheet) or Diameter (nm)	8	7.0	7.0	7.0	6.0	6.0	6.0
LOGIC DEVICE Electrical							
Vdd (V)	0.75	0.70	0.65	0.65	0.65	0.60	0.55
DRAM TECHNOLOGY							
DRAM ½ Pitch (nm) [1]	18	17.5	17	14	11	8.4	7.7
DRAM cell size factor: aF^2 [11]	6	6	4	4	4	4	4
DRAM bits/1chip target	8G	8G	16G	16G	32G	32G	32G
NAND Flash							
Flash ½ Pitch (nm) (un-contacted Poly)(F) (2D) [1]	15.0	15.0	15.0	15.0	15.0	15.0	15.0
Flash Product Highest Density (independent of 2D or 3D)	512G	1T	1T	1.5T	3T	4T	4T+
Flash 3D Maximum Number of Memory Layers [6]	64	96	128	192	384	512	>512

Notes for Table ES2:

ORTC: Logic Notes

[1] Based on 0.71x reduction per "Node Range" ("Node" = ~2x Mx).

[2] Based on 0.71x Mx reduction per "Generic Node", or .5x cell; 2x density; beginning 2013/"G1"/40nm.

[3] Defined as distance between metallurgical source/drain junctions

ORTC: DRAM Notes

[1] The definition of DRAM Half pitch has been changed from this edition. Because of 6F2 DRAM cell, BL pitch is no more critical dimension. pitch= (Cell Area/ Call size factor)^0.5."

Critical dimension for process development, the Minimum half pitch is also introduced. Currently Active area (long rectangle island shape) half pitch is the critical dimension of 6F2 DRAM.

Calculated half pitch is use the following equation "Calculated half pitch= (Cell Area/ Call size factor)^0.5." Critical dimension for process development, the Minimum half pitch is also introduced.

Currently Active area (long rectangle island shape) half pitch is the critical dimension of 6F2 DRAM.

[11] Cell size factor = a = (DRAM cell size/F2), where F is the DRAM ½ pitch. The current values of a are 6 from 2009. And a=4 will be predicted in 2021.

ORTC: NAND Flash Notes

[1] 2D NAND strings consist of closely packed polysilicon control gates (the Word Lines) that separate the source and drain of devices with no internal contact within the cell.

Up to now this uncontacted word line pitch is still the tightest in all technologies.

[6] The number of 3D layers is not a unique function, depending on the cell 1/2 pitch and 3D NAND technology architecture chosen. Lower number of 3D layers generally has lower bit cost,

but other factors such as decoding method, speed performance, easier or harder to get yield, also need to be considered.

The number of 3D layers spans a range since the same density product may be achieved by using smaller 1/2 pitch and fewer layers, or larger 1/2 pitch and more layers.

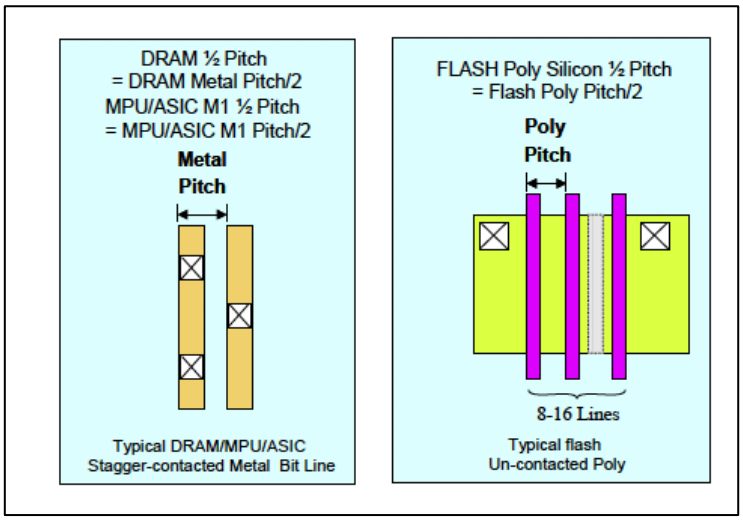

Figure ES12 Technology Node Definition

5. GRAND CHALLENGES

5.1. IN THE NEAR-TERM

5.1.1. APPLICATION BENCHMARKING

5.1.1.1. BIG DATA ANALYTICS

The gains in graph processing performance come from three main sources: 1) improvements in algorithms, 2) memory bandwidth, and 3) bandwidth. Processor performance does not have a first-order impact.

Some of the gains in traversed edges per second (TEPS) over the years have been due to improvements in the algorithm used. While we expect this factor to continue to have an impact, the improvements are likely to provide diminishing returns.

The most critical resources today are bandwidth and latency of the memory and the global network.

- Graph problems have a high ratio of communication to computation
- Very little locality

Memory bandwidth needs: the current top machines have an aggregate memory bandwidth of 5 petabytes per second.

- The next generation machines will attain about 20 petabytes per second, thanks to in-package DRAMs like high-bandwidth memory (HBM).

5.1.1.2. FEATURE RECOGNITION

For near-term digital hardware, the critical hardware needs are the design of systolic computing units that align well with the deep neural network (DNN) algorithms; and, ability to receive and send data to large amounts of memory at high-bandwidth yet reasonable power.

Further improvements can be obtained for distributed training by highly efficient use of network bandwidth, allowing the multiple "workers" to collaborate with a minimal amount of information exchange.

5.1.2. SYSTEMS AND ARCHITECTURES

5.1.2.1. IoT EDGE DEVICES

IoT edge devices must satisfy several stringent requirements. They must consume small amounts of energy for sensing, computation, security, and communication. They must be designed to operate with strong limits on their available bandwidth to the cloud.

Many IoT devices will include artificial intelligence (AI) capabilities; these capabilities may or may not include online supervision or unsupervised learning. These AI capabilities must be provided at very low-energy levels. A variety of AI-

enabled products have been introduced. Several AI technologies may contribute to the growth of AI in IoT edge devices: convolutional neural networks; neuromorphic learning, and stochastic computing.

IoT edge devices must be designed to be secure, safe, and provide privacy for their operations.

5.1.2.2. MOBILE SYSTEMS

Mobile systems present several challenges for system designers. Multimedia viewing, such as movies and live TV, have driven the specifications of mobile systems for many years. We have now reached many of the limits of human perception, so increases in requirements on display resolution and other parameters will be limited in the future based on multimedia needs. However, augmented reality will motivate the need for advanced specifications for both input and output (I/O) in mobile devices. Mobile device buyers demand frequent, yearly product refreshes. This fast refresh rate influences design methodologies to provide rapid silicon design cycles; it can also suggest the use of programmability to provide a broad range of models on a given platform. Financial transactions are now performed using mobile devices. We expect this trend to grow, particularly in developing nations, where financial technology will leapfrog. Security and privacy are key concerns for mobile devices, particularly for financial transactions.

5.1.2.3. CLOUD APPLICATIONS AND SYSTEMS

Cloud applications present several challenges for system designers. Data centers are starting to take advantage of heterogeneous core types, much as embedded systems have done for many years. System architects need to balance the performance improvements for chosen applications provided by specialized accelerators against the utilization of these specialized cores. The huge scale of problems in social networking and AI and other problems means that algorithms run at memory speed and that multiple processors are required to compute. The radius of useful locality—the distance over which programmers can use data as effectively local—is an important metric. We expect optical networking to greatly enlarge useful locality radius over the next few years. Memory bandwidth is a constraint on both core performance and number of cores per socket. Stacked memories, which are starting to come into commercial use, provide higher bandwidth memory connections. Thermal power dissipation continues to be an important limit.

Cloud systems present significant challenges. Heterogeneous architectures can provide more efficient computation of key functions. Novel memory systems, including stacked memories, offer high performance and lower power consumption. Advances in internal interconnect may create tipping points in system architecture. The term hyperconvergence is used to describe the point at which I/O speeds approach internal interconnect speeds.

5.1.3. OUTSIDE SYSTEM CONNECTIVITY

5.1.3.1. RF ANALOG TECHNOLOGY

The key challenges for radio frequency (RF) are to achieve high-performance, energy-efficient RF analog technology compatible with CMOS processing and delivering capabilities to support a broad range of applications for IoT devices. To achieve high-performance RF with high-energy efficiency, CMOS gate resistance must be reduced with technologies that are compatible with CMOS processing. Furthermore, SiGe and III-V performance needs increased ft and fmax while being integrated with CMOS. Passive devices need to be integrated on CMOS with higher performance.

To support systems and components to meet 5G performance requirements, we need devices to support <6 GHz massive multiple input multiple output (MIMO) and 28 GHz communication with low power and cost effectively, increasing energy efficiency of amplifiers while increasing operating frequency; systems to deliver high density communication without interference and with low noise; antennas to support multiple band communication in compact mobile devices, and low-cost high-efficiency directional antennas to support mmWave and massive MIMO 5G.

5.1.4. MORE MOORE

5.1.4.1. LOGIC DEVICE SCALING

Beyond 2019 a transition from finFET to gate-all-around (GAA) will start and potentially a transition to vertical nanowires devices will be needed when there will be no room left for the gate length scale down due to the limits of fin width scaling (saturating the Lgate scaling to sustain the electrostatics control) and contact width.

FinFET and lateral GAA devices enable a higher drive at unit footprint if fin pitch can be aggressively scaled while increasing the fin height. This increased drive at unit footprint by scaling the fin pitch comes at a trade-off between fringing capacitance between gate and contact and series resistance. This trend in reducing the number of fins while balancing the drive with increased fin height is defined as fin depopulation strategy, which also simulataneuously reducing the standard cell height, thefore the overall chip area.

The most difficult challenge for interconnects is the introduction of new materials that meet the wire conductivity requirements and reduce dielectric permittivity. As for the conductivity, the impact of size effects on interconnect structures must be mitigated. Future effective κ requirements preclude the use of a trench etch stop for dual damascene structures.

5.1.4.2. DRAM AND 3D NAND FLASH MEMORY

Since the DRAM storage capacitor gets physically smaller with scaling, the EOT must scale down sharply to maintain adequate storage capacitance. To scale the EOT, dielectric materials having high relative dielectric constant (κ) will be needed. Therefore metal-insulator-metal (MIM) capacitors have been adopted using high-κ ($ZrO_2/Al_2O/ZrO_2$) as the capacitor of 40−30's nm half-pitch DRAM. And this material evolution and improvement will continue until 20 nm high-performance (HP) and ultra-high-κ (perovskite κ > 50 ~ 100) materials are released. Also, the physical thickness of the high-κ insulator should be scaled down to fit the minimum feature size. Due to that, capacitor 3D structure will be changed from cylinder to pillar shape.

The economics of 3D NAND is further confounded by its complex and unique manufacturing needs. Although the larger cell size seems to relax the requirement for fine-line lithography, to achieve high data rate it is desirable to use large page size and this in turn translates to fine-pitched bit lines and metal lines. Therefore, even though the cell size is large, metal lines still require ~20 nm half-pitch that is only achievable by 193i lithography with double patterning. Etching of deep holes is difficult and slow and the etching throughput is generally very low. And depositing of many layers of dielectric and/or polysilicon, as well as metrology for multilayer films and deep holes all challenge unfamiliar territories. These all translate to large investment in new equipment and floor space and new challenges for wafer flow and yield.

5.1.5. EMERGING RESEARCH MATERIALS

5.1.5.1. MATERIALS FOR LOGIC DEVICE SCALING

Materials and processes that achieve performance and power scaling of lateral fin- and nanowire FETs (Si, SiGe, Ge, III-V) are as follows:

- Integrated high-κ dielectrics with equivalent oxide thickness (EOT) <0.5 nm and low leakage
- Integrated contact structures that have ultralow contact resistivity
- Achieving high-hole mobility in III-V materials in FET structures
- Achieving high electron mobility in Ge with low-contact resistivity in FET structures
- Processes for achieving low dislocations and anti-phase boundary generating interface between Ge/III-V channel materials and Si
- Dopant placement and activation, i.e., deterministic doping with desired number at precise location for Vth control and source/drain (S/D) formation in Si as well as alternate materials

5.1.5.2. MATERIALS FOR COPPER INTERCONNECT

Materials and processes that improve copper interconnect resistance and reliability are as follows:

- Mitigate impact of size effects in interconnect structures. Line and via sidewall roughness, intersection of porous low-κ voids with sidewall, barrier roughness, and copper surface roughness will all adversely affect electron scattering in copper lines and cause increases in resistivity.
- Patterning, cleaning, and filling at nano-dimensions. As features shrink, etching, cleaning, and filling high aspect ratio structures will be challenging, especially for low-κ dual damascene metal structures and DRAM at nano-dimensions.
- Cu wiring barrier materials must prevent Cu diffusion into the adjacent dielectric but also must form a suitable, high quality interface with Cu to limit vacancy diffusion and achieve acceptable electromigration lifetimes.
- Reduction of the κ value of intermetal dielectrics. Reduction of the interlevel dielectric (ILD) κ value is slowing down because of problems with manufacturability. The poor mechanical strength and adhesion properties of low-κ materials are obstructing their incorporation.

5.1.6. BEYOND CMOS

5.1.6.1. EMERGING MEMORIES AND LOGIC DEVICES

One of the grand challenges is to realize scaled volatile and nonvolatile memory technologies to replace SRAM and NAND flash memory in appropriate applications. The key components of such emerging memories are novel memory devices and selector devices.

Another grand challenge is to extend ultimately scaled CMOS as a platform technology into new domains of applications. The emerging logic and information processing devices will be extended CMOS devices and/or beyond CMOS devices.

5.1.7. LITHOGRAPHY

5.1.7.1. EUV LITHOGRAPHY FOR 7 NM NODE LOGIC AND BEYOND

Logic foundry producers have announced their commitment to producing products using extreme ultraviolet (EUV) with a target date of early 2019 and possible use late in 2018. There are three grand challenges: 1) the only manufacturing-level EUV exposure tool supplier has a target of 95% uptime by the time actual manufacturing commences; 2) the second challenge is the resolution, line edge roughness, and sensitivity (RLS) trade-off, that is, the need to have usable photospeed and resolution in combination with low enough stochastic effects, and 3) the third challenge is defectivity, particularly for masks. There is no actinic inspections system for patterned mask inspection. This makes eliminating mask defects a challenge.

Pellicles for EUV masks are under development, but not available yet. When pellicles are available, they will reduce exposure throughput by absorbing some EUV, leading to a different sort of cost impact. EUV tool uptime and the lack of pellicle technology are concerns that, once resolved, should stay resolved.

5.1.8. PACKAGING INTEGRATION

5.1.8.1. PACKAGING TECHNOLOGY

The challenges in 3D and 2.5D packaging are as follows:

- Materials and process for through silicon vias (TSVs) compatible with silicon
- Improve process for die stack compatible with future shrinks
- Head extraction form die stack
- Dense planer (2.5D) bridge to fill in the “interconnect gap”
- Establish hard (simulation and/or measurement based) universal performance metrics for package selection

5.1.9. METROLOGY

5.1.9.1. MEASUREMENT OF COMPLEX THREE-DIMENSIONAL (3D) STRUCTURES

The 3D structures, such as finFETs, place increased need for inline metrology for dimensional, compositional, and doping measurements. The materials’ properties of block co-polymers for directed self-assembly (DSA) result in new challenges for lithography metrology. The increased use of multi-patterning techniques introduces the need to independently solve a large set of metrics to fully characterize a multi-patterning process.

The 3D interconnect comprises several different approaches and new process control needs are not yet established. For example, 3D (critical dimension and depth) measurements will be required for trench structures including capacitors, devices, and contacts.

5.1.9.2. MEASUREMENT OF COMPLEX MATERIAL STACKS AND INTERFACIAL PROPERTIES

Reference materials and standard measurement methodology for new high-κ gate and capacitor dielectrics with engineered thin films and interface layers as well as interconnect barrier and low-dielectric layers, and other process needs are required.

Optical measurement of gate and capacitor dielectric averages over too large an area and needs to characterize interfacial layers. Carrier mobility characterization will be needed for stacks with strained silicon and silicon on insulator (SOI), III-V, GeOI, and other substrates, or for measurement of barrier layers. Metal gate work function characterization is another pressing need.

5.1.10. FACTORY INTEGRATION

5.1.10.1. RESPONDING TO BUSINESS REQUIREMENTS

In responding to rapidly changing and complex business requirements, the grand challenges are as follows:

- Increased expectations by customers for faster delivery of new and volume products
- Rapid and frequent factory plan changes driven by changing business needs
- Ability to load the fab within manageable range under changeable market demand, e.g., predicting planning and scheduling in real-time

- Enhancement in customer visibility for quality assurance of high reliability products: tie-in of supply chain and customer to factory information and control systems (FICS) operations
- Addressing the big data issues, thereby creating an opportunity to uncover patterns and situations that can help prevent or predict unforeseeable problems difficult to identify such as current equipment processing/health tracking and analytical tools
- To strengthen information security: maintaining data confidentiality (the restriction of access to data and services to specific machines/human users) and integrity (accuracy/completeness of data and correct operation of services), while improving availability (a means of measuring a system's ability to perform a function in a particular time) contradictive to needs of data availability.

5.1.10.2. RE-EMERGENCE OF 200MM PRODUCTION LINE

The increased heterogeneity and variety of devices combined with market pressures such as those associated with IoT solutions have given rise to 200 mm production as an important component of microelectronics ecosystem. While basic tenants of FI challenges and potential solutions associated with 300 mm translate well to 200 mm, there are specific FI challenges such as connectivity, variability, and availability of replacement components that must be addressed so that 200 mm can remain as a viable production capability in the ecosystem.

5.1.11. YIELD ENHANCEMENT

5.1.11.1. DETECTION OF MULTIPLE KILLER DEFECTS

One of the important key challenges will be the detection of multiple killer defects and the signal-to-noise ratio. It is a challenge to detect multiple killer defects and to differentiate them simultaneously at high capture rates, low cost of ownership, and high throughput. Furthermore, it is difficult to identify yield relevant defects under a vast amount of nuisance and false defects.

- Existing techniques trade off throughput for sensitivity, but at expected defect levels, both throughput and sensitivity are necessary for statistical validity.
- Reduction of inspection costs and increase of throughput is crucial in view of cost of ownership (CoO).
- Detection of line roughness due to process variation.
- Electrical and physical failure analysis for killer defects at high capture rate, high throughput and high precision.
- Reduction of background noise from detection units and samples to improve the sensitivity of systems.
- Improvement of signal to noise ratio to delineate defect from process variation.
- Where does process variation stop and defect start?

5.1.12. ENVIRONMENT, SAFETY, HEALTH, AND SUSTAINABILITY

5.1.12.1. MATERIAL CHALLENGES IN ENVIRONMENTAL, SAFETY, HEALTH, AND SUSTAINABILITY

Material challenges in Environmental, Safety, Health, and Sustainability are as follows:

- Emerging/novel materials, i.e., III-V (GaN, InP, InGaP, etc.), nano and energetic materials (assessing their ESH/S impacts, along with social responsibility implications)
- Utilization challenge (materials efficiency of incoming fab materials is <2%)
- Treatment and abatement solutions to meet current and future regulatory requirements can gate development ramp, and costs increasing
- Restrictions on recycling, repurposing, and reuse are significant due to technology and regulatory hurdles
- There is no universally accepted or applicable alternative assessment (frameworks, methods, and tools) strategy, nor are there clear guidelines or standards for how these should be applied for picking less ESH/S impactful materials.

5.2. IN THE LONG-TERM

5.2.1. APPLICATION BENCHMARKING

5.2.1.1. BIG DATA ANALYTICS

The gains in graph processing performance came mainly from improvements in algorithms, memory bandwidth, and bandwidth. Processor performance does not have a first-order impact.

In the long term, large memory bandwidth and lower latency and higher bandwidth of the global network, which could use optical links, will be needed.

5.2.1.2. FEATURE RECOGNITION

The main long-term challenges are DNN hardware based on either in-memory digital or analog computation, with the following critical technology need: analog memory devices, low-power, high-bandwidth, moderate-precision and extremely area-efficient A/D converters.

5.2.2. SYSTEMS AND ARCHITECTURES

5.2.2.1. IoT

The development of IoT will face significant long-term challenges. Low cost-efficient energy harvesting using multiple sources in order to develop autonomous systems, energy storage and management, low power sensing, computing and communication, automatic network configuration, and security will be needed.

5.2.2.2. MOBILE

Video will drive demand for both bandwidth and display, and augmented reality applications will require further increases in communication, computation, capture, and display. An important long-term challenge is also the substantial reduction of power consumption and increase of battery capacity to meet the demand of very active users.

5.2.2.3. CLOUD

Three-fourths of all data important to organizations will never be in the data center. High bandwidth memory and large socket thermal power dissipation using improved packaging and cooling will be needed.

5.2.2.4. CYBER-PHYSICAL SYSTEMS

Long-term challenges are associated with hardware and software reliability, security, exponential increase in the number of bytes generated and need of local analysis, and substantial advances in storage technologies.

5.2.3. OUTSIDE SYSTEMS CONNECTIVITY

Long-term grand challenges are the following: development of circuits for cancellation of 5G mmWave noise; reconfigurable and high-efficiency directional MIMO antenna with circuits to reconfigure and synchronize signals; agreement on optical technology standards; development of technologies for communication between systems with different wavelengths; polarization and modulations; reduction of latency of communication between CPUs and memory in data centers specifically due to routing, and conversion of electrical and photonic signals.

5.2.4. MORE MOORE

Power scaling is a major long-term challenge, which should use steep subthreshold slope devices but there is a lack of manufacturable candidates up to now. Diversification with novel architectures like vertical GAA (VGAA) devices, 3D stacking and possible co-integration of CMOS and beyond CMOS will be required for improving performance. These will need good management of thermal challenges, yield, and cost, together with introduction of alternatives to Cu-interconnects with low resistance and good reliability.

5.2.5. LITHOGRAPHY

Resolution may not be a challenge after 2025 due to the possible introduction of 3D devices. Potential patterning challenges will thus be related to cost, yield, defectivity, and optimization of complex 3D structures. Etch and deposition of sub 10 nm structures will also become major challenges. Another potential challenge might be implementing patterning on 450 mm wafers. However, if EUV is a mainstream patterning method in widespread use, this could limit the financial benefit of switching to 450 mm wafers. Little work is currently being done on extending any other patterning methodology (multiple patterning, nanoimprint, maskless, DSA) to larger wafer sizes than 300 mm.

5.2.6. FACTORY INTEGRATION

Important long-term challenges are the flexibility, extendibility, and scalability needs of a cost-effective, leading-edge factory, tackling environmental issues like material recycling and substitution (scarce, toxic) and future global regulations, and management of the uncertainty of novel device types replacing conventional CMOS and the impact of their manufacturing requirements on factory design.

5.2.7. YIELD ENHANCEMENT

The next generation inspection is a significant challenge. We need to explore new alternative technologies that can meet inspection requirements to discriminate defects of interest, like high speed scanning probe microscopy, near-field scanning

optical microscopy, interferometry, scanning capacitance microscopy, and e-beam. In-line defect characterization and analysis for smaller defect sizes and feature characterization will be required alternatively to optical systems and energy dispersive X-ray spectroscopy.

5.2.8. BEYOND CMOS

The beyond CMOS era is facing major research challenges. Nanoscale volatile and nonvolatile memory technologies to replace traditional SRAM, DRAM and FLASH in appropriate applications are needed, for instance by using resistive memories (phase-change RAM (PCRAM), Resistive RAM (ReRAM), magnetic RAM (MRAM)). The scaling of information processing technology substantially beyond that attainable by ultimately scaled CMOS will require new computing paradigms like neuromorphic or quantum computing, novel architectures, device technology breakthroughs using charges (e.g., small slope switches) or in the longer term alternative state/hybrid state variables (e.g., spin, magnon, phonon, photon, electron-phonon, photon-superconducting qubit, photon-magnon), the states being be digital, multilevel, analog, or entangled.

5.2.9. EMERGING RESEARCH MATERIALS

Many novel materials will be needed in the long term to satisfy the requirement of a lot of applications. We are facing substantial challenges for the development and integration of these alternative materials in the following fields: 3D monolithic and vertical integration of high mobility and steep subthreshold transistors (III-V, Ge, 2D, carbon nanotube (CNT); complex metal oxides, etc.) for extending or replacing CMOS; emerging non-charge-based memories and select devices (ferromagnetic, multiferroic, complex oxides, etc.) to replace DRAM/SRAM/NVM; interconnects with improved reliability and electromagnetic performance at the nanoscale (CNT novel interlayer dielectrics, e.g., metal organic framework and carbon organic framework) to replace copper, and integration on CMOS platforms, with flexible electronics and of biocompatible functional materials.

5.2.10. PACKAGING INTEGRATION

In the field of packaging, the significant long-term challenges are: reliable interconnects and substrates for wearable electronics (bendable, washable); bio compatible systems for miniaturized implants; efficient integration of electronic and optical components; and integration of cooling systems for quantum computing.

5.2.11. METROLOGY

Nondestructive wafer and mask-level metrology with better precision for novel device architectures and 3D structures are needed. Complementary and hybrid metrology combined with state of the art statistical analyses will be required to reduce the measurement uncertainty due to statistical limits of sub-7 nm process control. Materials characterization and metrology methods are also needed for control of interfacial layers, dopant positions, defects, size, location, alignment and atomic concentrations relative to device dimensions and for direct self-assembling processes.

5.2.12. ENVIRONMENT, SAFETY, HEALTH, AND SUSTAINABILITY

We are facing substantial challenges with the possible impact of health and environment of emerging materials (e.g., III-V materials, perfluorooctanoic acid (PFOA)) and potential biological interactions with e.g., mmWave (28-330 GHz). Driving green chemistry and engineering concepts will become a very important asset for future technologies, considering their impact on sustainability and future regulations in this domain.

6. APPENDIX

6.1. APPENDIX A—IFT CHAPTER FILES LINKS

- Application Benchmarking (AB)
- Systems and Architectures (SA)
- Outside System Connectivity (OSC)
- More Moore (MM)
- Beyond CMOS (BC)
- Packaging Integration (PI)
- Factory Integration (FI)
- Lithography (L)
- Emerging Research Materials (ERM)
- Yield Enhancement (YE)
- Metrology (M)
- Environment, Safety, Health (ESH/S), and Sustainability

6.2. APPENDIX B—OVERALL ROADMAP CHARACTERISTICS (ORSC AND ORTC) SOURCE INFORMATION LINKS

- Systems and Architectures Tables
- More Moore Tables