Toward five-dimensional scaling: How density improves efficiency in future computers

P. Ruch
T. Brunschwiler
W. Escher
S. Paredes
B. Michel

We address integration density in future computers based on packaging and architectural concepts of the human brain: a dense 3-D architecture for interconnects, fluid cooling, and power delivery of energetic chemical compounds transported in the same fluid with little power needed for pumping. Several efforts have demonstrated that by vertical integration, memory proximity and bandwidth are improved using efficient communication with low-complexity 2-D arrays. However, power delivery and cooling do not allow integration of multiple layers with dense logic elements. Interlayer cooled 3-D chip stacks solve the cooling bottlenecks, thereby allowing stacking of several such stacks, but are still limited by power delivery and communication. Electrochemical power delivery eliminates the electrical power supply network, freeing valuable space for communication, and allows scaling of chip stacks to larger systems beyond exascale device count and performance. We find that historical efficiency trends are related to density and that current transistors are small enough for zetascale systems once communication and supply networks are simultaneously optimized. We infer that biological efficiencies for information processing can be reached by 2060 with ultracompact space-filled systems that make use of brain-inspired packaging and allometric scaling laws.

Introduction

Computers have developed in an extraordinary fashion since the demonstration of the first room-sized electromechanical computers (Eniac and Zuse) in the 1940s; they shrank to the size of personal computers (PCs) (30 liters) around 1980 while improving their performance and efficiency by many orders of magnitude, respectively. Initially, the parallel shrinking of size and cycle time kept memory proximity and equal fractions for communication and computation power. Since then, the efficiency improved by five orders of magnitude, but the form factor and communication distances over the printed circuit board did not change. The wider data bus and faster data transfer rate did not match the processor development, creating a communication bottleneck. This was partially compensated by hierarchical caches, but during a cache miss, a long delay is created, and transport of hundreds of bytes is needed. We discuss that current architectures are communication dominated because the increased power densities lead to larger sizes of the chip package plus air cooler. Low-power smartphones and microservers continued to shrink: They are 100 times smaller (0.3 liters) and almost an order of magnitude more efficient. Thus, packaging initially developed parallel to performance but slowed for high-performance computers after the introduction of the PC geometry where the development focused on transistor scaling, opening a communication bandwidth and latency gap.

The dominance of communication in terms of latency and power consumption as a consequence of device-centric scaling has been anticipated [1–8]. With growing system sizes, communication requires larger fractions of the overall system resources. Even memory can be considered a form of temporal communication, routing data from one moment in time to another. For this reason, demand for memory and communication needs joint optimization because interconnect prediction evolves into system-level prediction.

Digital Object Identifier: 10.1147/JRD.2011.2165677

©**Copyright** 2011 by International Business Machines Corporation. Copying in printed form for private use is permitted without payment of royalty provided that (1) each reproduction is done without alteration and (2) the Journal reference and IBM copyright notice are included on the first page. The title and abstract, but no other portions, of this paper may be copied by any means or distributed royalty free without further permission by computer-based and other information-service systems. Permission to republish any other portion of this paper must be obtained from the Editor.

0018-8646/11/$5.00 © 2011 IBM

With exponential densification, heat dissipation becomes increasingly demanding [9, 10], and water cooling becomes a key for volume reduction by several orders of magnitude with minimal size and thermal resistance being achieved by embedding fluid channels directly in the silicon die [11]. While heat transfer is enhanced in microscopic channels, the pumping power increases with the inverse fourth power of hydraulic diameter such that optimized fluid distribution manifolds are needed to reduce pressure drop [12, 13].

The transition from 2-D scaling to 3-D integration offers an excellent opportunity to improve computer density and efficiency. Interlayer cooled chip stacks have been demonstrated as a scalable solution allowing the integration of several logic layers each with massive amounts of main memory [14]. These systems use 3-D communication and ultracompact cooling similar to a human brain. We are therefore on the verge of bionic volumetric computer scaling. This volumetric scaling is well exploited by biological organisms gaining a performance advantage by means of hierarchically branched supply and drain networks [15]. We explore how these brain-inspired packaging concepts can be transferred to future 3-D computers to eliminate several bottlenecks without fully deviating from current device technology.

The main motivation of this paper is to emphasize that the transition to 3-D stacked chips bears a huge potential to improve computer efficiency that can be fully exploited only with new architectural concepts. The topic is presented in the following sequence: First, the computation and communication energy demands of current computers are compared with the efficiency of a biological brain. Then, computer density trends are analyzed, and volumetric scaling theories are introduced. In a next section, interlayer cooling of 3-D chip stacks with combined electrochemical power delivery is described. Finally, the potentials of the new packaging architectures are evaluated, in particular the globally optimized hierarchical networks that endow systems with higher dimensional scaling.

Energy demands of information processing

To distinguish logical states in the presence of thermal fluctuations, irreversibly operating devices store many $k_B T$ of energy and dissipate it with every switching event. Reversible devices may conserve most of the energy, thus reducing the energy dissipation [16–18]. However, the performance of such adiabatic computing (in operations per second) is significantly less than that of present-day complementary metal-oxide semiconductor (CMOS) technology for a given device footprint. Assuming 10^5 fully dissipative primary operations for one useful computation (data transport, clock, and register operations), we get 10^{16} operations/J, or approximately ten-million-fold better than best current computer efficiencies of 10^9 operations/J. [Note that, in our comparisons, an operation is considered equivalent to a floating-point operation (FLOP).] Biology has developed energy-aware architectures with efficiencies of 10^{14-15} operations/J, five to six orders of magnitude better than those of current computers [19–21]. Most insights into brain efficiencies are derived from comparisons with robotic systems that perform like organisms or from estimates extrapolated from the retina [20]. This 1-cm^2 tissue of 500-μm thickness with 10^8 neurons performs edge and motion detection requiring 10^9 operations/s. Extrapolated to the 10^5 times larger brain, a performance of 10^{14} operations or 0.1 peta-operations is estimated. This coincides with a performance of one neuron equivalent to 1,000 operations/s [22]. A computer with this performance currently requires about 200 kW or 10^4 times more energy than the entire human brain (e.g., for *Jeopardy!*). This efficiency comparison is for tasks that are difficult for computers, but for simpler tasks (e.g., chess), the difference is smaller.

Direct comparisons between computers and brains in terms of computational efficiency are difficult due to the strong workload dependence [20]. Literature values to attain human-level intelligence vary widely, starting with 10^8 operations/s and extending to 10^{16} operations/s based on experiences with cars driven by artificial intelligence: For example, equipped with the capability of 10^8 operations/s, self-driving cars could handle desert terrain but miserably failed in city traffic. Moravec's paradox states that high-level reasoning requires little computation, but low-level sensorimotor skills require enormous resources: It is easy to make computers exhibit adult-level performance on intelligence tests but difficult to give them the skills of a one-year-old child when it comes to perception and mobility [23]. For a complete and meaningful comparison of the efficiencies of computers and brains, overall energy consumption is important. Similar to waste heat utilization in biology, the recovery of thermal energy in computers has been demonstrated, thus replacing conventional fossil fuel heating and lowering the carbon footprint of systems comprising data centers and heat consumers [23–28].

Energy is required for device switching and for transmitting signals along wires, and the overall energy fractions (switching versus transmitting signals) have changed with technology scaling. The physical basis for the improvement of device efficiency was described by Dennard et al. [29]. Since 2004, the rate of efficiency improvement has stalled due to the end of voltage scaling [30]. Computational energy efficiency is currently limited by the power required for communication despite the introduction of copper wiring and low-permittivity dielectrics [31]. Meanwhile, the number of back-end-of-line (BEOL) metal layers has increased to 12, with the total wire length exceeding several kilometers per square centimeter [32]. The wire latency was not significantly reduced for recent technology generations, which means that system efficiency

Table 1 Time and energy requirements for on- and off-chip operations for 130- and 45-nm node sizes [38]. (ALU: arithmetic logic unit.)

	Delay (ps)		*Energy* (pJ)	
Operation	130 nm	45 nm	130 nm	45 nm
32-byte ALU operation	650	250	5	0.3
Transfer 32 bytes across chip (10 mm)	1,400	2,300	100	17
Transfer 32 bytes off-chip	—	—	1,300	400

did not improve. In addition, latency-induced idling creates a communications-related energy consumption. With technology scaling, wiring and communication latencies grew drastically and have dominated transistor performance for the last two decades [5, 8]. Local wires became shorter but were slower because of increased electrical resistance in nanoscale structures [3]. Global wires with reduced cross sections lengthened, leading to much larger delays, requiring power- and area-hungry repeaters to keep delays manageable. Still, only 1% of the die area or $2 \cdot 10^7$ elements are accessible within one clock cycle [33]. Logic-centric scaling created a widening disparity, which is often referred to as the memory wall [34, 35]. Memory latency [36] now exceeds hundreds of clock cycles, with a painful impact on performance since memory access to noncached locations is frequently needed [35, 37].

The time for global on-chip communication compared with arithmetic operations has increased from 2:1 at the 130-nm node to 9:1 at the 45-nm technology node, respectively (see **Table 1**). The energy required for off-chip communication was 260 times greater than that for arithmetic operations at 130 nm, whereas this ratio increased to 1,300 for the 45-nm node (see Table 1). Both performance (latency) and energy efficiency are dominated by communication [7]. The efficiency of an algorithm depends on data movement in both time and space, and optimization efforts are needed here as well [4]. With scaling from 1 μm to 35 nm, the switching delay of transistors was reduced from 20 to 2.5 ps, whereas the *RC* response time of a 1-mm-long wire increased from 1 to 250 ps. From an energy point of view, the transistor switching energy reduced 3,000-fold, i.e., from 300 to 0.1 fJ, whereas the interconnect switching energy changed only 100-fold, i.e., from 300 to 3 fJ [6].

Communication efficiency depends on overall length and organization, where frequently communicating elements are arranged in proximity. The communication intensity and structure are characterized with Rent's parameters p and k being the slope and y-intercept of log–log plots of the number of input/output (I/O) pins as a function of the number of gates, respectively [39, 40]. More generalized, Rent's rule relates the number of connections to the number of logic elements in a modular circuit architecture [41–43]. Rent's rule may be expressed as $C = k \cdot N^p$, where C is the number of edge connections, and N is the number of logic elements. Rent coefficient k gives the average number of connections per logic element, and Rent exponent p is a measure for the network complexity, with larger values of p indicating more complex networks with larger terminal requirements. The maximum value of $p = 1$ indicates random placement and connectivity of logic elements. The importance of a Rent analysis grows with system size: Systems with less than one billion functional elements tend to be device dominated, whereas more complex systems are communication dominated. Rent's rule is not as well known as Moore's Law but is more fundamental: Neglecting its key messages had a detrimental effect on the performance of current computers [43]. Rent's rule can also be applied to a human brain, where a neuron with certain synaptic complexity is compared with logic gates and axons are compared with wires [44].

Shortening of communication paths in computers is enabled by stacking of dies [3]. Stacking also allows integration of main memory, which eliminates more than 50% of off-chip communication, thereby softening the package-level bottleneck. Two key challenges have to be tackled: 1) The impact of areal heat dissipation will multiply since both thermal power and thermal resistance rise; and 2) introducing sufficient electrical power to meet the demand of the entire stack. Approximately 75% of interconnects are allocated to power and ground in planar integrated circuit (IC) designs. With 3-D chip stacks, even more pins will be needed for power delivery. Current planar IC designs fully rely on one-sided heat removal and power delivery. As future chips begin to scale out-of-plane, the heat dissipation and power demand requires volumetrically scalable packaging solutions to provide cooling and power.

With technology scaling, transistors became cheap and abundant, but availability of chip and board-level wiring became even scarcer. Managing the demand for wires is overdue but has been delayed since it requires architectural changes that are more difficult than expanding the transistor count. Architectural changes are particularly important when the efficiency of systems is the main metric. For this, simpler cores are needed (e.g., accelerators or graphics processing units) that reduce fan-out, network complexity, and wire length. Current fan-out factors drive power consumption and need to be reduced. Another soft spot of exascale systems is the dynamic random access memory (DRAM): Multitasking cores have high demand for memory, which slows DRAM access and further increases latency and demand due to the larger number of address decoding steps and the serialization of the data. Core-level multithreading is a strong driver for memory demand and triggers a faster encounter with the memory wall. There are developments needed to reach higher performance with fewer components

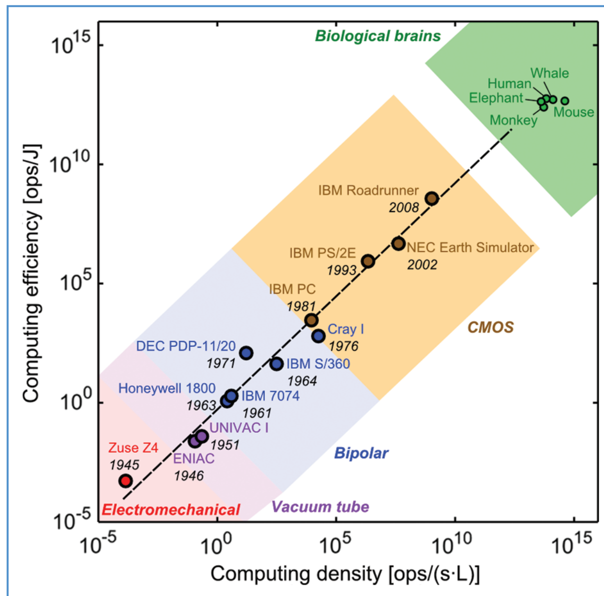

Figure 1

Computational efficiency and computational density of computers compared with mammalian brains.

since this addresses active energy and leakage. Today, it takes about 10^{15} elements for 1 peta-FLOPs (PFLOPs). With the elimination of complexity and tenfold smaller memory demand per CPU, more than ten times higher FLOPs per element, and therefore exascale performance, can be more easily reached than exascale device count.

Volume demands in information processing

Historically, the energy efficiency of computation doubled every 18–20 months [45, 46], with a 1,000-fold performance improvement every 11 years [45, 47], leading to a net 10-fold energy consumption increase, which is clearly not a sustainable trend. Improved computational efficiency was enabled due to new device technology and scaling: Ten orders of magnitude efficiency improvement have been achieved, but with the anticipated end of CMOS scaling [48], it is unclear how this will continue (see **Figure 1**) upward to biological efficiencies (green region). The computational performances of computers were compiled from [20, 46, 49] and normalized to the power consumption and volume of each computer without supporting infrastructure. The computational performance of mammalian brains was estimated from [20, 46, 49] and normalized to brain volume, taking into account the brain metabolic rate [50], power consumption, and volume fraction [51, 52]. It is striking that six different technologies (i.e., electromechanical switches, vacuum tubes, discrete transistor–transistor logic transistors, emitter-coupled logic ICs, very large scale integrated (VLSI) CMOS, and biological brains) all fall onto one line on the log–log plot of operations per joule versus operations per second per liter, and changes in device technology did not lead to discontinuities. For microprocessor-based computers, this historic evolution is attributable to device shrinkage, which was motivated by a reduced cost per transistor and resulted in improved performance and power use per transistor. The main driver for efficiency improvement, therefore, has been integration density. Another way of expressing this trend is by stating that more efficient systems can be built more compact. In all technologies, the volume fraction occupied by the devices was less than 0.1%. This ratio now has become extremely small with the current transistor generation just occupying 1 ppm of computer volume. Further CMOS device shrinkage is challenging due to the rising passive power fraction. We suggest that device technology needs to be efficient and small enough with low leakage currents to support integration density. Integration density wins over single-device performance when system efficiency is considered as observed during the transition from bipolar to CMOS technology and during other technology transitions.

Since current computers are mainly limited in efficiency and performance by their communication capability, we have to look for examples with better communication architecture. The human brain is such an example. It consists of 44% by volume of white matter with axons implementing long-range 3-D connections between cortical areas. The gray matter contains cell bodies, dendrites, and axons. The latter two occupy approximately 60% of the gray matter (or 33% overall), indicating a high degree of local communication. Overall, the human brain uses 77% of its volume for communication. As brain size increases, the volume of the communication infrastructures increases faster than the volume of functional devices according to a power law. In fact, brains show similar Rent coefficients as VLSI systems [44, 53] and are power intensive, dissipating about 20% of the total body energy in less than 3% of the tissue [54].

The size of organisms spans 27 orders of magnitude from the largest animals and plants to the smallest microbes and subcellular units. Despite this large diversity, many phenomena scale with size in a surprisingly simple fashion: Metabolic rate scales with 3/4 power of mass, whereas timescales and sizes (heart rate, life span, and aorta length) scale with 1/4 power of mass, whereas other values are constant such as the number of heartbeats during life and the energy used by all organisms of a size class. This is due to hierarchical branching networks that terminate in size-invariant capillaries and show maximized metabolic capacity by minimizing transport distances and times [55]. The distribution of energy, resources, and information plays a central role in constraining the structure and organization of life. Space-filled hierarchical fractal-like branching networks distribute energy and materials between reservoirs

and microscopic sites in animal circulatory, respiratory, renal, and neural systems, as well as plant vascular systems. In addition, these networks are seen in the systems that supply food, water, power, and information to human societies in cities.

Quarter-power scaling laws are as universal as metabolic pathways, the structure and function of the genetic code, and the process of natural selection. Fractal-like networks effectively endowed life with an additional fourth spatial dimension and a strong selective advantage [15, 55]. Neural cells, as opposed to other cells, resemble snowflakes, so that they can receive more information and maximize their mass and heat transfer. For this reason, the metabolic rate of the brain scales with a 4/5 power law as opposed to the 3/4 power law for other cells. This means that nature endows brains uniquely with a fifth dimension originating from the fractal structures of brain cells and the associated communication networks [56–58]. Functional elements in biology occupy a spatial fraction exceeding 80%. In current computers, this fraction is much lower: $\sim$96% is used for thermal transport (air and copper), 1% for electrical and information transport, 1% for structural stability, and 1 ppm for transistors and devices, whereas in the brain, 70% of the structure is used for communication, 20% for computation, and $\sim$10% for energy and thermal supply and drain, and structural stability. With 2-D communication, fewer communication lines are available, and much more energy is needed for communication. The increased functional density in brains is a major driver for the improved efficiency, as shown in Figure 1. Therefore, adopting packaging and other architectural features of the brain for future computers is a promising approach to building much more efficient and compact computers in the future.

Toward volumetric scalability in 3-D packages

VLSI architectures exhibit similar Rent exponents and network complexities as the human brain [44], but their computational efficiencies and functional densities lag by orders of magnitude (see Figure 1) due to the lack of scalability of current VLSI architecture in the third dimension. The stacking of multiple active silicon dies leads to enhanced interconnect densities of $10^6/\text{cm}^2$, as compared with $10^3/\text{cm}^2$ for the chip-to-board pins [59]. The critical components are the through-silicon vias (TSVs), which have been demonstrated with pitches of 3 μm [32, 60]. A significant increase in performance is achieved by incorporating memory and logic interconnected by TSVs [61]. Scalable cooling between active layers can handle heat fluxes of 180 W/cm^2 per layer [14, 62]. Such forced convection schemes maximize hot-spot cooling compatible with vertical interconnects at practical pressure drops [14]. While the stack height is dilated through microchannel incorporation by a factor of 1.3–2, this dilation can be reduced to less than 1.2 by offering cooling only in between stacks comprising one logic layer with ten memory layers. The second major constraint is the need for power delivery pins. In planar designs, about 75% of all microprocessor off-chip I/Os are allocated to power delivery. In stacked-chip configurations with multiple processor dies, the demand will multiply and quickly use up all available pins. Higher voltages for chip-to-board power interconnects in combination with on-chip voltage transformers may aid in delaying this cutoff point but only at the cost of silicon real estate and efficiency.

A big step toward volumetric scalability may be taken by unification of thermal and power I/O requirements, which is achieved via a fluidic network. The technological implementation of such a multifunctional network is an electrochemical power delivery scheme in which soluble redox couples are added to the coolant medium, representing an on-chip realization of microfluidic redox flow cells. First conceived as a form of secondary battery for long-lived large-scale energy storage [63], the miniaturization of redox flow cells [64] favors their integration into chip stacks with power demand at voltage levels on the order of 1 V. A unification of thermal and power fluidic I/O is attractive from the perspective of overlapping intensity distributions for power consumption and heat dissipation and the avoidance of ohmic losses when large currents are distributed at low voltage levels since charge carriers are transported by forced convection instead. This provides an alternative to area-hungry power TSVs, which hamper signaling I/O between layers in a 3-D stack. In addition, the inherent capacitance of the electrochemical double layer (about 10 μF/cm^2 with respect to the electrode area) can take over the role of decoupling capacitors to minimize supply voltage variations. To ensure voltage stability, voltage transformation and domains with switched supply islands for standby leakage power minimization and combination with on-chip voltage regulators is possible.

For the laminar flow conditions encountered in microchannels, the average current density j provided by the forced convection of spontaneously discharging electrolyte solutions depends on fluid velocity v, hydraulic diameter D_h, and channel length x in the same way as the heat and mass transfer coefficients, i.e., $j \sim h_\text{HT} \sim h_\text{MT} \sim (v/(D_\text{h} \cdot x))^{1/3}$. High current densities are favored by microchannel geometries with high fluid velocities and small hydraulic diameters. Hierarchical fluid manifolds allow for rapid energy delivery in channels with short lengths x while distributing the bulk electrolyte solutions via larger channels.

By pumping the electrochemically active species between logic plus memory stacks, the electrochemical energy conversion occurs near the highest BEOL metallization level and thereby bypasses chip-to-board pins and TSVs. Furthermore, the convective fluid transport may extend to the data-center infrastructure where an electrochemical recharging unit can regenerate the discharged electrolyte

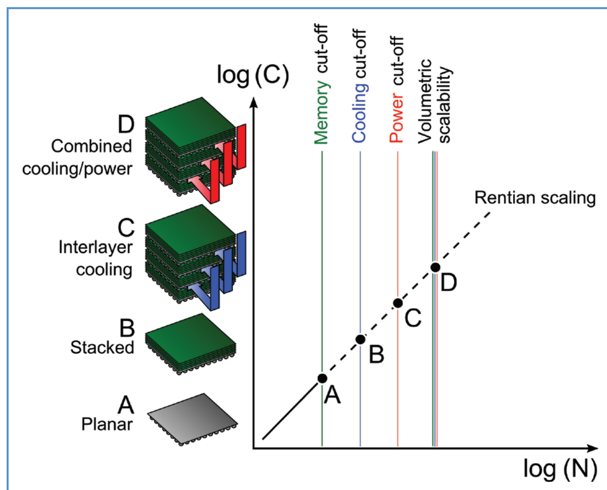

Figure 2

Scalability in a log–log plot of connection count $[\log(C)]$ against logic element count $[\log(N)]$. A: Conventional planar design (2-D); B: stacked processor and memory dies with one-sided power delivery and heat removal (3-D); C: stack of stacks with volumetric heat removal and one-sided power delivery; D: stack of stacks with combined volumetric heat removal and power delivery (3-D bionic packaging).

solutions. The complete electrochemical charge–discharge cycle has energy efficiencies of more than 80% per pass. Notably, the convective charge transport scheme corresponds to the power distribution at voltages on the order of 1 V: Duplicating this with conventional wiring would require prohibitively large amounts of copper. The parasitic pumping power favorably compares with the ohmic losses in electrical conductors for large cross sections. Therefore, in addition to the benefit of scaling with the number of stack interlayer gaps, convective electrochemical power delivery may improve system efficiency by reducing power conversion and distribution losses.

For a beneficial implementation of combined cooling and power delivery, the power per unit electrode area of microfluidic electrochemical systems must be increased by almost two orders of magnitude with respect to reported values [64]. The ability to deliver power independent of hard-wired connectivity enables the same degree of volumetric scalability as for the coolant distribution. As microfluidic redox flow cell performance is typically diffusion limited, the path toward power density improvement relies in part on enhanced mass transport and short diffusion lengths through electrode microstructuring and elevated temperature operation.

We can contemplate the impacts of high-performance fluidic power distribution networks on the volumetric scalability of 3-D ICs, as shown in **Figure 2**. Planar designs are limited in terminal and logic element counts due to the limited off-chip access to memory I/O (see A in Figure 2). This memory-related limitation can be alleviated by stacking of logic and memory dies as it is developed in the IT industry today (see B in Figure 2). Stacking of several of these entities is not possible due to limits of heat dissipation through the stack to a backside heat sink. Volumetric cooling such as the interlayer forced convection described above copes with the increased heat dissipation footprint of 3-D chip stacks (see C in Figure 2) and allows stacking of several such entities each having a logic layer and many layers of memory until the system is limited by the ability to deliver sufficient power to all tiers in a stack through the TSVs. Unified volumetric cooling and power delivery has the potential to surpass this power wall and push the scalability of stacked-chip architectures to the interlayer I/O limit (see D in Figure 2). The replacement of power TSVs by signal TSVs enables higher interlayer bandwidth for improved I/O access within the stack of stacks and to the outside through all faces of the cube.

True volumetric scalability implies direct proportionality between cooling or power delivery capacity and system volume. However, such systems are not readily realized in practice. For example, surface access to a cube scales as $V^{2/3}$, where V is the volume of the cube. Generally, the hypersurface of a D-dimensional hypervolume scales with the $(D-1)/D$-dimensional power of the hypervolume [65]. Space-filling networks in biology are observed to follow allometric scaling, in which supply networks subproportionally grow with system volume. Rent's rule relates the number of devices in a block with the need for communication. Allometric scaling is inverse since it defines the amount of elements with a given metabolic rate in a volume element by the availability of supply that comes through the surface of this volume element, which means that, for a 2-D system, the perimeter scales with a power of 1/2, and for a 3-D element, the surface scales with a power of 2/3. Hierarchical branched networks allow biology to scale in a 4-D space, which results in a scaling exponent of 3/4. If we combine this with nonspherical shapes of brain cells and optimized Rentian scaling of networks, the performance of a brain scales with the power of 4/5 of its volume as if it would use the surface in a 5-D hyperspace.

We now derive performance guidelines for space filling computers by combining allometric scaling with Rent's rule assuming that communication and supply compete for the same surface in a single element. For simplicity, we start the scaling based on the I/O supply infrastructure with different slopes given in **Figure 3**. For 2-D chips, we allow the number of available connections to be multiplied by the number of metal layers. This connection space is then compared with the need for connections of an individual cell. Information processing devices are all subject to a fundamental access optimization problem. A given unit cell of a computer optimally performs when it has low-loss access to the resources listed in **Table 2**.

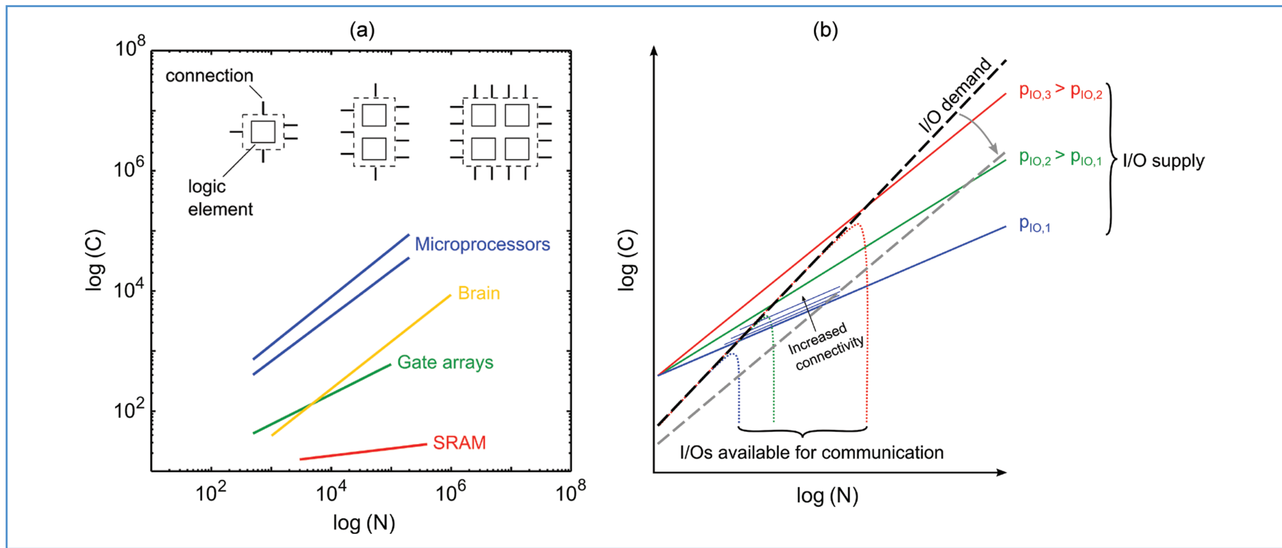

Figure 3

(a) Comparison of the Rentian scaling of different circuit types [40] with the human brain [44]. (b) Influence of the scalability of the I/O supply to meet the connectivity demand for growing system sizes. The black curve shows the total demand, whereas the gray curve shows the demand without power supply pins; the gray curve does not intersect with the maximal available I/O curve (red) and is therefore scalable to very large systems.

Table 2 Resource demand for space-filling computers.

Communication in/out (exponent <0.5)
Communication ground in/out (exponent <0.5)
Clock in/ground (exponent <0.5)
Power supply/ground (exponent >0.67)
Cooling supply/drain (exponent >0.67)

The I/O requirements, including fan-out, define a starting point on the vertical axis in Figure 3(b) that affects the onset of the memory and power walls. From this, it becomes clear that low fan-out architectures can be extended to larger block sizes. In the past, architectures with very high primary fan-out were advantageous because switches were slow and wires were fast. This is different now: The communication demand has to be limited from the very beginning by creating a low fan-out microarchitecture. The onset of the cooling, power, memory, and communication wall occurs as the I/O need of a given block exceeds the available surface. Stanley-Marbell and Cabezas showed that with current power density trends in processors, there will be eventually no pins left for communication [37]. Obviously, such a system does not perform, and a compromise on power density has to be made. Communication links in a brain [see Figure 3(a), yellow] are counted with respect to neurons. For a fair comparison with computers, neurons are set equal to 1,000 transistors. Under this normalization, the specific communication needs (fan-out) in a brain are lower than those in current logic technology, and the yellow curve is close to gate arrays (green) and memory (red).

The wiring complexity indicated by the value of p varies from 0.1 for memory with serial data readout to 0.8 for high-performance microprocessors [see Figure 3(a)]. Packaging constraints lead to a breakdown of Rent's rule for off-chip communication when the package pin count available for communication sets an upper bound for the wire count [see Figure 3(b)]. Such discontinuities are also observed in the corpus callosum of the human brain [65], but breaking Rent's rule between computational nodes and main memory has a much more severe impact on system performance. On-chip cache reduces the negative impact of this bottleneck at the expense of more than half of the chip area being allocated to memory [7, 66]. Serialization of the information stream enlarges overhead and latency. Optical and wireless interconnects [67] cannot reduce latency: The only approach that can reduce latency through reduction of overall distances and more parallel interconnects is 3-D chip integration. Improving the scalability of the number of I/O pins available for communication can extend Rentian scaling to larger system sizes with higher gate counts and improved computational performance without the drawbacks of data serialization [see Figure 3(b)]. Increasing connectivity without changing scalability, for example, by increasing the number of BEOL metallization layers, can only marginally improve system scaling.

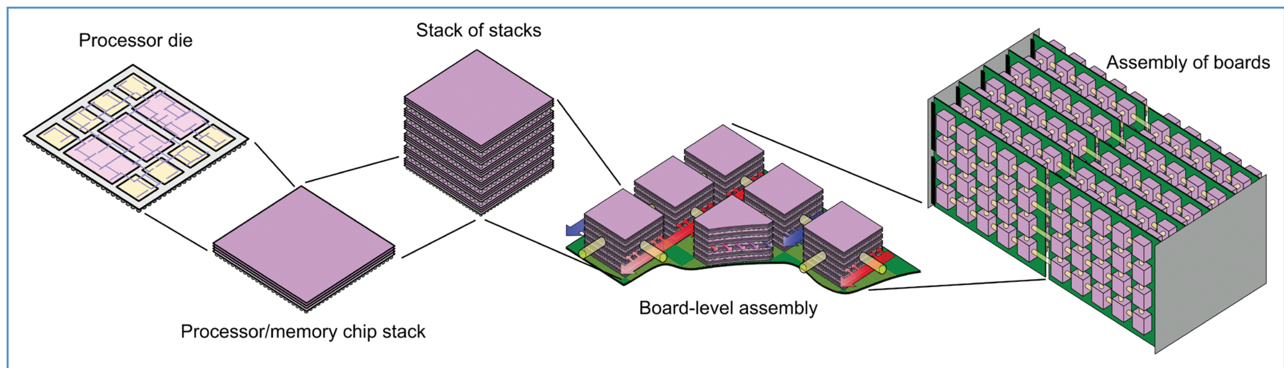

Figure 4

Volumetrically scalable stack of stacks. Note that the communication links between logic elements as well as the fluidic thermal and power I/O enable high-density space-filling supply and drain networks.

So far, access optimization has been only performed on individual power and signaling networks. Full optimization requires a parallel optimization on a universal connectome or supranet that has to satisfy all the needs of all devices within one block since all these resources compete for the same surface. There are different types of needs that scale with different exponents, leading to different volume requirements for each of the I/O networks. The combination of hierarchical fluid supply and Rentian information transport allows scaling of information processing systems in a higher-dimensional space, collapsing their volume demand to the same efficiency and functional density as biological systems.

A compact system is crucial: The shorter the communication paths, the lower both the latency and energy consumption. Since in the future overall efficiency will be a key metric, any tricks to hide a slow communication (such as caching and multithreading) cannot hide the flaws of the underlying system. With volumetrically cooled chip stacks, the overall volume is compressed by a factor of $10^5 - 10^6$ so that the following volume fractions result: transistors and devices 5%, interconnects 25%, power supply 30%, thermal transport 30%, and structural stability 10%.

The functional density in current computers is 10,000 transistors/mm^3 compared with 100,000 neurons/mm^3 in a brain. Since a neuron has a performance equivalent of 1,000 transistors, the density of the brain is 10,000 times higher. On an interconnect level, it is more than 1,000-fold denser albeit at a much smaller communication speed. Volumetrically cooled chip stacks are space-filled like biological brains with ten times higher density in "neuron equivalents." After 3/4-power allometric scaling from chip scale (1 cm^3) to brain scale (1,500 cm^3), the component density is similar. Interconnect density assuming 4/5-power allometric scaling is ten times higher, and with the much higher communication speed, the overall communication capacity is 100 times better. Since in a 3-D chip stack the communication network is not equivalent in all dimensions of space but lower in the vertical dimension, we estimate a small communication capacity advantage over the human brain.

Today, most of the communication performance in a computer is gained via the higher speed at the expense of a more than quadratically higher energy demand [68]. For improved efficiency, frequent communication should be constrained to proximal locations within the 0.1-cm^3 chip stacks, whereas long-distance communication crossing this boundary should be limited to a minimum. This can be realized with our brain-inspired packaging approach, which combines cooling and power supply networks, leading to similar functional and communication densities as in a biological brain. The exact approach to scale the chip stack to a more than 1,000 cm^3 system following an allometric scaling factor still remains to be established. If we partially compromise on density, we still may be able to build an electronic brain with similar overall performance and efficiency as the human brain but slightly lower density, as depicted in **Figure 4**.

The integration process in Figure 4 starts from a CPU die with ten added layers of memory that extend the memory wall and allow ten times higher performance than a current CPU (100 cores, 1 tera-FLOPs (TFLOPs), 10 GB memory, and 10^{11} logic elements). Next, ten of these systems are vertically stacked to a volume of 300 mm^3 (100 GB, 1,000 cores, 10 TFLOPs, and 10^{12} elements). Then, 100 of these elements are combined to a multichip module (MCM) (100,000 cores, 10^{14} elements, and 1 PFLOPs). Finally, ten of these MCMs are stacked to a cube of ten times the volume of a human brain and a performance of 100 times the performance of a human (one million cores, 10 PFLOPs, and 100 terabytes). Maximum communication distances in the system are shorter than 50 cm.

Table 3 Scaling exponents and functional density for increasing logic element count of a hypothetical computer unit cell with power, signal, and thermal I/O demands.

	Scaling exponents		*Volume demand scaling exponents*		*$N = 10^{10}$ elements*		*$N = 10^{15}$ elements*	
Scenario	Communication (Rent)	Power and cooling	Signal	Power and cooling	Power (W)	Volume (cm^3)	Power (kW)	Volume (m^3)
2-D, with data serialization	0.80	1.1	1.2	1.6	50	1	10,000	10,000
3-D, air-cooled	0.80	1.0	0.80	1.5	5	200	400	50
3-D, air-cooled with advanced devices	0.80	1.0	0.80	1.5	1	200	100	10
3-D, bionic packaging	0.65	0.96	0.65	1.3	<1	5	3	0.01

Discussion

A key characteristic of a biological brain is the close proximity of memory and computation, which allows recognizing images stored many years ago in a fraction of a second. Computers will continue to be poor at this unless architectures become more memory centric. Caches, pipelining, and multithreading improve performance but reduce efficiency. Multitasking violates the temporal Rent rule because it puts processes in spatial or temporal proximity that do not communicate. If the underlying hardware follows Rent's rule, a system optimization segregates the tasks. However, the breakdown of Rent's rule at the chip boundary leads to a serialization of the data stream and, accordingly, to penalties in performance and efficiency. The reason for this situation is readily appreciated: Assuming typical Rent coefficients of $k = 10$ and $p = 0.8$, the terminal count required for 10^9 logic elements is 10^8, exceeding current package pin counts by five orders of magnitude. The current industry-wide effort to introduce 3-D integration allows a considerable improvement of bandwidth and latency between the logic and memory units. With parallel introduction of new architectural concepts, the transition to 3-D stacked chips can unfold a much larger potential, in particular to increase the specific performance from 1 to more than 10 FLOPs per element.

A coarse assessment of the power and volume requirements for stacked systems in combination with our proposed bioinspired packaging scheme can be performed based on current device and wire characteristics, with wiring needs derived from Rent's rule (with $k = 10$) according to Donath [68] and its extension to 3-D systems [69]. Power consumption is estimated from the device count, including switching and leakage power, as well as from the switching power associated with wiring. The volume demand is derived from the I/O requirements for power, cooling, and communication (based on the terminal count according to Rent's rule). For simplicity, these I/O requirements are consolidated to a computer unit cell with the shape of a cube.

We now consider different reference cases based on varying I/O demands and supply scaling with increasing element count (see **Table 3**). Present-day scaling behavior (2-D scenario) based on a complex communication architecture with $p = 0.8$ is dominated by the overwhelming demand for wiring, which results in both an overproportional power scaling coefficient of 1.1 with a growing number of elements and excessive area requirements for the terminal count. The volume grows even more aggressively with the element count, following a 1.5-fold higher scaling exponent due to the surface-to-volume ratio of the unit cell cube. The extrapolation of power and volume demands for increasing element count is illustrated in **Figure 5**: Scaling to 10^{18} elements results in excessive power and volume demands of 30 GW and 1 km^3, even with data serialization, which considerably reduces the demand for package pin count and therefore volume.

Three-dimensional integration enables close proximity between logic and main memory, where ideally the communication supply scales volumetrically (see Table 3: scenario 3-D, air-cooled). This proximity eliminates the cache hierarchy with its power consumption overhead, eliminates the need for core-level multitasking, reduces memory demand per core, and allows higher FLOPs per device count numbers than today. Due to reduced wiring, the power demand is diminished [see Figure 5(a), red line] and scales roughly proportionally with the number of elements in the system with the volume demand following a scaling exponent of 1.5. Therefore, the system volume [see Figure 5(b), red line] becomes dominated by cooling for large element sizes. Overall, a petascale machine can be built

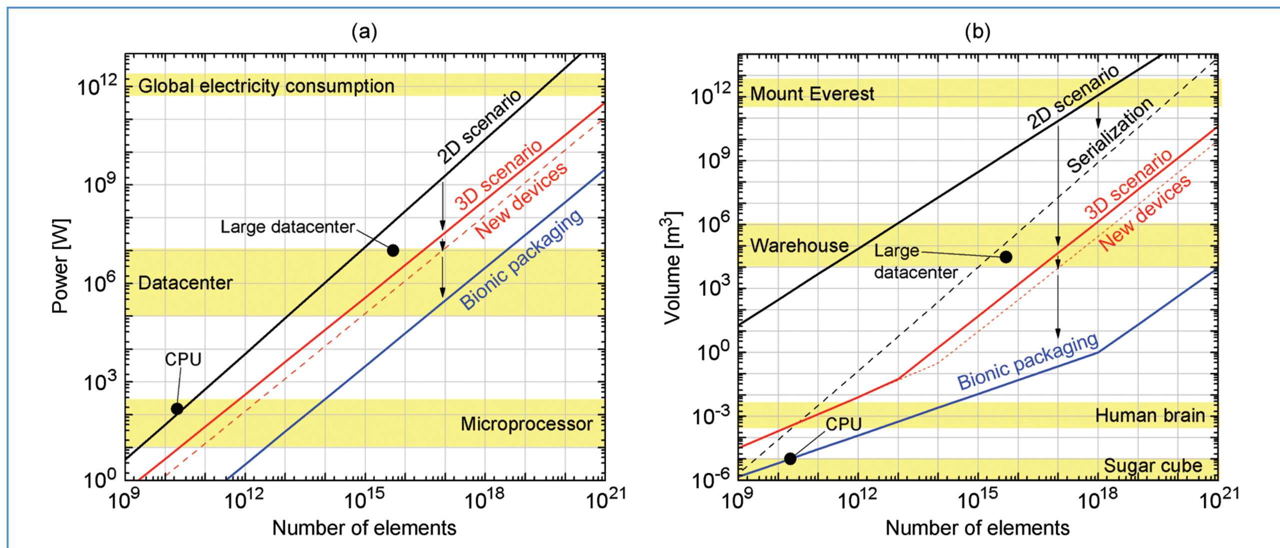

Figure 5

(a) Power and (b) volume of computers as functions of element count for different scaling scenarios.

within 50 m^3 and a power budget of 400 kW (see Table 3). For an exascale system 2 million m^3 or about twice the largest known data-center volume is needed with a steep power bill for the 0.2 GW consumed. Improved device technology cannot alleviate this: Shrinking the gate length to 10 nm, lowering the operating voltage to 0.3 V, and reducing leakage currents by a factor of 10 improves the absolute values of power and volume demands but does not fundamentally change the system scalability [see Figures 5(a) and 5(b), dashed red lines] [30, 70]. The scalability can only be improved by allowing for a change in the way in which power and cooling are supplied.

The biologically inspired packaging approach allows improved scaling: An ideal I/O supply would scale proportionally with system volume in order to meet the power demand. However, a practical implementation of the proposed combined fluidic cooling and power delivery is likely to scale underproportionally with system volume. Biological supply and drain networks imply that a scaling exponent of 3/4 is feasible, so that both cooling and power delivery scale $\sim V^{3/4}$ instead of $\sim V$. A simplification of wiring is assumed since power is delivered electrochemically, leading to a lower Rent exponent of 0.65 (see Table 3, scenario 3-D, bionic packaging). The improved volumetric scalability results in the most compact system size [Figure 5(b), blue line]. The system volume corresponds to the minimum volume to implement the wiring required by Rent's rule up to 10^{18} elements, until the faster growing volume demand for power and cooling dictates system sizes above this value. While performance and size differences between 2- and 3-D bionic systems are minute for small system sizes (current CPUs), they explosively grow with system size: While a petascale system built with 2-D technology fits in a few 1,000 m^3 using 10 MW, a zetascale system requires a volume larger than Mount Everest and consumes more than the current human power usage (20 TW). A 3-D bionic system scales to 10 L, 1 m^3, and 10,000 m^3 for peta-, exa-, and zetascale systems, respectively. Importantly, today's transistors ($100 \times 100 \times 100$ nm^3 cube) are already dense enough to sustain such high density systems, i.e., they occupy 1 cm^3, 1 dm^3, and 1 m^3 for peta-, exa-, and zetascale systems, respectively. The key toward volume reduction therefore lies in the volumetric packaging approach.

Figure 5 shows that the transition from the current 2-D scenario (black dashed) to a 3-D scenario can shrink the volume of a petascale computer by a factor of 300 and by a factor of 1,000 if a continued device development is assumed. The introduction of 3-D bionic packaging allows further shrinkage by a factor of 1,000. In terms of energy, 3-D integration can reduce energy consumption by a factor of 30 and, with improved devices, by a factor of 100. Three-dimensional bionic packaging allows extending this by another factor of close to 100, thereby reaching biological efficiencies. The implementation of such systems may be realized within a 50-year timeframe and will allow with moderate added investments to harvest the benefits of 3-D stacking and TSV technology much better and much longer. Then, the performance of the human brain will be matched with 10^{14} elements that consume about 100 W and occupy a volume of 2,500 cm^3 (excluding storage and external communication). Key is that the system core is

compressed one-million-fold volumetrically and 100-fold linearly: This is the same scaling factor that was reached between 1945 and 1985 but projected to happen in half the time to compensate for slower device technology developments.

Outlook

We have shown that integration density is crucial for biological information processing efficiency and that computers can—from a volume and power consumption point of view—develop along the same route up to zetascale systems. We conjecture that the information industry could reach the same efficiency as biological brains by 2060, 100 years after the introduction of electronic computers. We may recall that the first steam engine had less than 0.1% efficiency and it took 200 years (1780 to 1980) to develop steam engines that reached biological efficiency. The evolution described above would be twice as fast a development in the information industry as for the process that laid the foundation for the Industrial Age. Economically, it is important that the cost per transistor continues to shrink. The slower scaling speed at continued manufacturing efficiency improvements (e.g., 450-mm wafers) will allow longer depreciation times in chip factories and reduce the manufacturing cost per silicon area. What needs to shrink faster is the cost per interconnect, in particular the cost per long-distance interconnect. Three-dimensional stacking processes need to be well industrialized to make sure that interconnect and memory cost that dominate in future systems over (logic) transistor cost will show a similar Moore's trend in the next 50 years to make technological predictions of this paper also economically feasible.

Whether it makes sense to strive for a zeta-FLOPs capability system or whether a large number of smaller systems are able to meet information processing demands for IBM's 150-year anniversary remains to be seen. What we want to state here is that, while the key message for the last 50 years was that "there is a lot of room at the bottom"; the message for the next 50 years is "there is a lot of room between devices" or "there is a lot of slack in wires." Integration density will improve in the next 50 years to similar values as biological systems and similar efficiencies even for the most challenging comparisons. It remains to be seen whether a novel device technology could be mass produced by then that provides better base efficiency while not jeopardizing integration density. Availability of an improved device technology might help accelerate the rate of progress and allow reaching information processing efficiencies beyond biological brains. Predicting developments in the information industry after 2060 remain difficult, but there is a chance that, by IBM's 200-year anniversary, packaging and device technology may enable integrated electronic super brains with 100-fold human intelligence—whatever this means.

Acknowledgments

The authors would like to thank Ronald Luijten, Philip Stanley-Marbell, Wilfried Haensch, and Thomas Theis for discussions, as well as Walter Riess for support.

References

1. M. T. Bohr, "Interconnect scaling—The real limiter to high performance ULSI," in *IEDM Tech. Dig.*, 1995, pp. 241–244.
2. W. J. Dally, "Interconnect-limited VLSI architecture," in *Proc. IEEE Int. Interconnect Technol. Conf. (Cat. No.99EX247)*, 1999, pp. 15–17.
3. J. A. Davis, R. Venkatesan, A. Kaloyeros, M. Beylansky, S. J. Souri, K. Banerjee, K. C. Saraswat, A. Rahman, R. Reif, and J. D. Meindl, "Interconnect limits on gigascale integration (GSI) in the 21st century," *Proc. IEEE*, vol. 89, no. 3, pp. 305–324, Mar. 2001.
4. D. Greenfield and S. Moore, "Implications of electronics technology trends for algorithm design," *Comput. J.*, vol. 52, no. 6, pp. 690–698, Aug. 2009.
5. R. Ho, K. W. Mai, and M. A. Horowitz, "The future of wires," *Proc. IEEE*, vol. 89, no. 4, pp. 490–504, Apr. 2001.
6. J. D. Meindl, J. A. Davis, P. Zarkesh-Ha, C. S. Patel, K. P. Martin, and P. A. Kohl, "Interconnect opportunities for gigascale integration," *IBM J. Res. Develop.*, vol. 46, no. 2/3, pp. 245–263, Mar. 2002.
7. S. Moore and D. Greenfield, "The next resource war: Computation vs. communication," in *Proc. Int. Workshop Syst. Level Interconnect Prediction*, 2008, pp. 81–85.
8. K. C. Saraswat and F. Mohammadi, "Effect of scaling of interconnections on the time delay of VLSI circuits," *IEEE J. Solid-State Circuits*, vol. SSC-17, no. 2, pp. 275–280, Apr. 1982.
9. S. P. Gurrum, S. K. Suman, Y. K. Joshi, and A. G. Fedorov, "Thermal issues in next-generation integrated circuits," *IEEE Trans. Device Mater. Reliab.*, vol. 4, no. 4, pp. 709–714, Dec. 2004.
10. M. Saini and R. L. Webb, "Heat rejection limits of air cooled plane fin heat sinks for computer cooling," *IEEE Trans. Compon., Packag., Technol.*, vol. 26, no. 1, pp. 71–79, Mar. 2003.
11. D. B. Tuckerman and R. F. W. Pease, "High-performance heat sinking for VLSI," *IEEE Electron Device Lett.*, vol. EDL-2, no. 5, pp. 126–129, May 1981.
12. W. Escher, B. Michel, and D. Poulikakos, "A novel high performance, ultra thin heat sink for electronics," *Int. J. Heat Fluid Flow*, vol. 31, no. 4, pp. 586–598, Aug. 2010.
13. W. Escher, B. Michel, and D. Poulikakos, "Efficiency of optimized bifurcating tree-like and parallel microchannel networks in the cooling of electronics," *Int. J. Heat Mass Transf.*, vol. 52, no. 2–6, pp. 1421–1430, Feb. 2009.
14. T. Brunschwiler, B. Michel, H. Rothuizen, U. Kloter, B. Wunderle, H. Oppermann, and H. Reichl, "Interlayer cooling potential in vertically integrated packages," *Microsyst. Technol.*, vol. 15, no. 1, pp. 57–74, Oct. 2008.
15. G. B. West, J. H. Brown, and B. J. Enquist, "A general model for the origin of allometric scaling laws in biology," *Science*, vol. 276, no. 5309, pp. 122–126, Apr. 1997.
16. R. Keyes and R. Landauer, "Minimal energy dissipation in logic," *IBM J. Res. Dev.*, vol. 14, no. 2, pp. 152–157, Mar. 1970.
17. S. Lloyd, "Ultimate physical limits to computation," *Nature*, vol. 406, no. 6799, pp. 1047–1054, Aug. 2000.
18. L. B. Levitin, "Energy cost of information transmission (along the path to understanding)," *Phys. D, Nonlinear Phenom.*, vol. 120, no. 1/2, pp. 162–167, Sep. 1998.
19. R. C. Merkle. (1989, Aug.). Energy limits to the computational power of the human brain. *Foresight Update.* [Online]. *6,* pp. 1–4. Available: http://www.merkle.com/brainLimits.html
20. H. Moravec. (1998). When will computer hardware match the human brain? *J. Evol. Technol.* [Online]. *1*. Available: http://www.transhumanist.com/volume1/moravec.htm

21. R. Kurzweil, *The age of spiritual machines*. New York: Viking Penguin, 1999.
22. H. Moravec, *Robot: Mere Machine to Transcendent Mind*. London, U.K.: Oxford Univ. Press, 2000.
23. V. P. Carey and A. J. Shah, "The exergy cost of information processing: A comparison of computer-based technologies and biological systems," *J. Electron. Packag.*, vol. 128, no. 4, pp. 346–353, Dec. 2006.
24. M. Iyengar and R. Schmidt, "Analytical modeling for thermodynamic characterization of data center cooling systems," *J. Electron. Packag.*, vol. 131, no. 2, pp. 021009-1–021009-1, Jun. 2009.
25. T. Brunschwiler, B. Smith, E. Ruetsche, and B. Michel, "Toward zero-emission data centers through direct reuse of thermal energy," *IBM J. Res. Dev.*, vol. 53, no. 3, pp. 11:1–11:13, May 2009.
26. G. Meijer, "Cooling energy-hungry data centers," *Science*, vol. 328, no. 5976, pp. 318–319, Apr. 2010.
27. T. Brunschwiler, G. I. Meijer, S. Paredes, W. Escher, and B. Michel, "Direct waste heat utilization from liquid-cooled supercomputers," in *Proc. 14th IHTC*, 2010, pp. 429–440.
28. A. Haywood, J. Sherbeck, P. Phelan, G. Varsamopoulos, and S. K. S. Gupta, "A sustainable data center with heat-activated cooling," in *Proc. IEEE 12th Intersoc. Conf. Thermal Thermomech. Phenom. Electron. Syst. (ITherm)*, 2010, pp. 1–7.
29. R. H. Dennard, F. H. Gaensslen, H.-N. Yu, V. L. Rideout, E. Bassous, and A. R. LeBlanc. (1974, Oct.). Design of ion-implanted MOSFETs with very small physical dimensions. *IEEE J. Solid-State Circuits.* [Online]. *SSC-9(5)*, pp. 256–268. Available: http://www.ece.ucsb.edu/courses/ECE225/225_W07Banerjee/reference/Dennard.pdf
30. L. Chang, D. J. Frank, R. K. Montoye, S. J. Koester, B. L. Ji, P. W. Coteus, R. H. Dennard, and W. Haensch, "Practical strategies for power-efficient computing technologies," *Proc. IEEE*, vol. 98, no. 2, pp. 215–236, Feb. 2010.
31. K. Maex, M. R. Baklanov, D. Shamiryan, F. Lacopi, S. H. Brongersma, and Z. S. Yanovitskaya, "Low dielectric constant materials for microelectronics," *J. Appl. Phys.*, vol. 93, no. 11, pp. 8793–8841, Jun. 2003.
32. *International Technology Roadmap for Semiconductors: 2010 Update*. [Online]. Available: http://www.itrs.net
33. V. Agarwal, M. Hrishikesh, S. W. Keckler, and D. Burger, "Clock rate versus IPC: The end of the road for conventional microarchitectures," in *Proc. ACM SIGARCH Comput. Archit. News*, 2000, vol. 28, pp. 248–259.
34. Q. Harvard, R. J. Baker, and R. Drost, "Main memory with proximity communication: A wide I/O DRAM architecture," in *Proc. IEEE Workshop Microelectron. Electron Devices*, Apr. 2010, pp. 1–4.
35. W. A. Wulf and S. A. McKee, "Hitting the memory wall: Implications of the obvious," in *Proc. ACM SIGARCH Comput. Archit. News*, 1995, vol. 23, pp. 20–24.
36. D. A. Patterson, "Latency lags bandwidth," *Commun. ACM*, vol. 47, no. 10, pp. 71–75, Oct. 2004.
37. P. Stanley-Marbell and V. C. Cabezas, private communication, 2011.
38. W. J. Dally, "Computer architecture is all about interconnect," in *Proc. 8th Int. Symp. High-Performance Comput. Archit.*, 2002, pp. 1–11. [Online]. Available: http://www.hpcaconf.org/hpca8/sites/hpca8-panel/DallySlides.pdf
39. M. Y. Lanzerotti, G. Fiorenza, and R. A. Rand, "Interpretation of Rent's rule for ultralarge-scale integrated circuit designs, with an application to wirelength distribution models," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 12, no. 12, pp. 1330–1347, Dec. 2004.
40. M. Y. Lanzerotti, G. Fiorenza, and R. A. Rand, "Microminiature packaging and integrated circuitry: The work of E.F. Rent, with an application to on-chip interconnection requirements," *IBM J. Res. Dev.*, vol. 49, no. 4/5, pp. 777–803, Jul. 2005.
41. B. Landman and R. Russo, "On a pin versus block relationship for partitions of logic graphs," *IEEE Trans. Comput.*, vol. C-20, no. 12, pp. 1469–1479, Dec. 1971.
42. P. Christie and D. Stroobandt, "The interpretation and application of Rent's rule," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 8, no. 6, pp. 639–648, Dec. 2000.
43. D. Stroobandt, "Interconnect research influenced," *IEEE Solid-State Circuits Mag.*, vol. 2, no. 1, pp. 21–27, Winter 2010.
44. D. S. Bassett, D. L. Greenfield, A. Meyer-Lindenberg, D. R. Weinberger, S. W. Moore, and E. T. Bullmore. (2010, Apr.). Efficient physical embedding of topologically complex information processing networks in brains and computer circuits. *PLoS Comput. Biol.* [Online]. *6(4)*, p. e1 000 748. Available: http://www.ploscompbiol.org/article/info:doi/10.1371/journal.pcbi.1000748
45. Standard Performance Evaluation Corporation. [Online]. Available: http://www.spec.org/power_ssj2008/results/
46. J. Koomey, S. Berard, M. Sanchez, and H. Wong, "Implications of historical trends in the electrical efficiency of computing," *IEEE Ann. Hist. Comput.*, vol. 33, no. 3, pp. 46–54, Mar. 2011.
47. The Green 500. [Online]. Available: http://www.green500.org
48. E. Nowak, "Maintaining the benefits of CMOS scaling when scaling bogs down," *IBM J. Res. Dev.*, vol. 46, no. 2/3, pp. 169–180, Mar. 2002.
49. W. D. Nordhaus, "Two centuries of productivity growth in computing," *J. Econ. Hist.*, vol. 67, no. 1, pp. 128–159, Mar. 2007.
50. M. Kleiber, "Body size and metabolic rate," *Physiol. Rev.*, vol. 27, no. 4, pp. 511–541, Oct. 1947.
51. H. Jerison, "Brain to body ratios and the evolution of intelligence," *Science*, vol. 121, no. 3144, pp. 447–449, Apr. 1955.
52. M. A. Hofman, "Energy metabolism, brain size and longevity in mammals," *Quart. Rev. Biol.*, vol. 58, no. 4, pp. 495–512, Dec. 1983.
53. V. Beiu and W. Ibrahim, "Does the brain really outperform Rent's rule?" in *Proc. IEEE Int. Symp. Circuits Syst.*, 2008, pp. 640–643.
54. J. E. Niven and S. B. Laughlin, "Energy limitation as a selective pressure on the evolution of sensory systems," *J. Exp. Biol.*, vol. 211, no. 11, pp. 1792–1804, Jun. 2008.
55. G. B. West and J. H. Brown, "The origin of allometric scaling laws in biology from genomes to ecosystems: Towards a quantitative unifying theory of biological structure and organization," *J. Exp. Biol.*, vol. 208, no. 9, pp. 1575–1592, May 2005.
56. J. H. He, "A brief review on allometric scaling in biology," in *Lecture Notes in Computer Science*. Berlin, Germany: Springer-Verlag, 2004, pp. 652–658.
57. J.-H. He and Z. Huang, "A novel model for allometric scaling laws for different organs," *Chaos Solitons Fractals*, vol. 27, no. 4, pp. 1108–1114, Feb. 2006.
58. J. P. DeLong, J. G. Okie, M. E. Moses, R. M. Sibly, and J. H. Brown, "Shifts in metabolic scaling, production, and efficiency across major evolutionary transitions of life," *Proc. Nat. Acad. Sci. U. S. A.*, vol. 107, no. 29, pp. 12 941–12 945, Jul. 2010.
59. J. Knickerbocker, P. Andry, B. Dang, R. Horton, M. Interrante, C. Patel, R. Polastre, K. Sakuma, R. Sirdeshmukh, E. Sprogis, S. M. Sri-Jayantha, A. M. Stephens, A. W. Topol, C. K. Tsang, B. C. Webb, and S. L. Wright, "Three-dimensional silicon integration," *IBM J. Res. Dev.*, vol. 52, no. 6, pp. 553–569, Nov. 2008.
60. M. Motoyoshi, "Through-silicon via (TSV)," *Proc. IEEE*, vol. 97, no. 1, pp. 43–48, Jan. 2009.
61. A. Rahman and R. Reif, "System-level performance evaluation of three-dimensional integrated circuits," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 8, no. 6, pp. 671–678, Dec. 2000.
62. B. Dang, M. S. Bakir, D. C. Sekar, J. C. R. King, and J. D. Meindl, "Integrated microfluidic cooling and interconnects for 2D and 3D chips," *IEEE Trans. Adv. Packag.*, vol. 33, no. 1, pp. 79–87, Feb. 2010.
63. L. Thaller, "Electrically rechargeable redox flow cells," NASA, Washington, DC, Tech. Memorandum X-71 540, 1974.
64. E. Kjeang, N. Djilali, and D. Sinton, "Microfluidic fuel cells: A review," *J. Power Sources*, vol. 186, no. 2, pp. 353–369, Jan. 2009.
65. H. M. Ozaktas, "Paradigms of connectivity for computer circuits and networks," *Opt. Eng.*, vol. 31, no. 7, pp. 1563–1567, Jul. 1992, DOI:10.1117/12.57685.

66. A. Hartstein, V. Srinivasan, T. Puzak, and P. Emma, "Cache miss behavior: Is it sqrt (2)?" in *Proc. Conf. Comput. Frontiers*, 2006, pp. 313–321.
67. R. H. Havemann and J. A. Hutchby, "High-performance interconnects: An integration overview," *Proc. IEEE*, vol. 89, no. 5, pp. 586–601, May 2001.
68. W. Donath, "Placement and average interconnection lengths of computer logic," *IEEE Trans. Circuits Syst.*, vol. CAS-26, no. 4, pp. 272–277, Apr. 1979.
69. D. Stroobandt and J. Van Campenhout, "Estimating interconnection lengths in three-dimensional computer systems," *IEICE Trans. Inf. Syst.*, vol. E80-D, no. 10, pp. 1024–1031, Oct. 1997.
70. D. J. Frank, R. H. Dennard, E. Nowak, P. M. Solomon, Y. Taur, and H.-S. P. Wong, "Device scaling limits of Si MOSFETs and their application dependencies," *Proc. IEEE*, vol. 89, no. 3, pp. 259–288, Mar. 2001.

Received May 15, 2011; accepted for publication July 18, 2011

Patrick Ruch *IBM Research Division, Zurich Research Laboratory, Säumerstrasse 4, CH-8803 Rüschlikon, Switzerland (ruc@zurich.ibm.com).*

Dr. Ruch joined the Advanced Thermal Packaging Group at the IBM Zurich Research Laboratory in 2009 as a postdoctoral researcher. He had studied materials science at the Swiss Federal Institute of Technology (ETH) Zurich and received his Ph.D. degree from ETH in 2009 for his work performed at the Paul Scherrer Institut (PSI) on electrochemical capacitors. His main research interests are in energy conversion and storage with applications to efficient data centers. He is currently a Research Staff Member working on exploratory research regarding microfluidic electrochemical energy conversion and the development of adsorbent materials for solid sorption refrigeration.

Thomas Brunschwiler *IBM Research Division, Zurich Research Laboratory, Säumerstrasse 4, CH-8803 Rüschlikon, Switzerland (tbr@zurich.ibm.com).*

Mr. Brunschwiler is a member of the advanced thermal packaging team. His main interests are in micro- and nanoscale heat and mass transfer and its utilization in efficient data-center operation. He contributed with his work to the development and realization of the zero-emission data-center demonstrator. He is currently finishing his Ph.D. degree at the Technical University of Berlin, Germany, in electrical engineering. In the context of his thesis, he focuses on exploring liquid cooling solutions for 3D chip stacks, in particular interlayer convective heat removal. He joined IBM Research in 2001 to work on integrated optics in the silicon oxynitride project where he contributed to the design and processing of all optical, thermally tunable add/drop multiplexers. In 2002 he moved to the organic light emitting display technology project where he was responsible to develop carrier injection layers to increase system lifetime. Since 2003 he has been working on microscale heat transfer and improved high performance thermal interfaces with micro channel patterning and demonstrated nature-inspired jet-impingement concepts.

Werner Escher *IBM Research Division, Zurich Research Laboratory, Säumerstrasse 4, CH-8803 Rüschlikon, Switzerland (wes@zurich.ibm.com)*.

Dr. Escher joined the Advanced Packaging Group as a Ph.D. student in 2006. He received his doctorate in 2009 from Swiss Federal Institute of Technology (ETH) in the area of micro- to nanoscale heat and mass transfer. In the scope of his thesis he explored the potential of nanofluids for electronics cooling and developed a high efficiency micro-fluidic heat sink with an incorporated hierarchical fluid supply network. Currently he is a Research Staff Member at the IBM Zurich Research Laboratory. His research interests include single phase cooling, thermal interfaces, novel packaging techniques for high power thermal-fluidic microelectronic packages and high concentrating photovoltaic systems.

Stephan Paredes *IBM Research Division, Zurich Research Laboratory, Säumerstrasse 4, CH-8803 Rüschlikon, Switzerland (spa@zurich.ibm.com).*

Mr. Paredes received his bachelor's degree in systems engineering, major in physics engineering from the University of Applied Science NTB in Buchs, Switzerland, in 2000. From 2001 to 2004 he worked at the Institute for Microsystems, Buchs, Switzerland, in the field of MEMS, optical packaging and waveguide technology for optical interconnects. In 2004, he received his postdiploma degree in optical systems engineering. From 2004 to 2007, he worked in the development of optical interconnects and packaging technology for printed circuit boards at Varioprint AG in Heiden, Switzerland. In 2008 he joined the IBM Research Laboratory in Zurich, Switzerland, where he is currently working with the Advanced Thermal Packaging Group on microprocessor liquid cooling, improved thermal interfaces and concentrated photovoltaics.

Bruno Michel *IBM Research Division, Zurich Research Laboratory, Säumerstrasse 4, CH-8803 Rüschlikon, Switzerland (bmi@zurich.ibm.com).*

Dr. Michel received a Ph.D. degree in biochemistry/biophysics from the University of Zurich, Switzerland, in 1988 and subsequently joined the IBM Zurich Research Laboratory to work on scanning probe microscopy and its applications to molecules and thin organic films. He then introduced microcontact printing and led a worldwide IBM project for the development of accurate large-area soft lithography for the fabrication of LCD displays. He started the Advanced Thermal Packaging Group in Zurich in 2003 in response to the needs of the industry for improved thermal interfaces and miniaturized convective cooling. He has published more than 100 research articles with a Hershey index of 44, is an IBM master inventor holding more than 50 Granted patents. He received the IEEE Harvey Rosten award and several IBM outstanding technological achievement and IBM patent portfolio awards. The main current research topics of the Zurich Group are microtechnology/microfluidics for nature-inspired miniaturized tree-like hierarchical supply networks, hybrid liquid/air coolers, 3-D packaging, and thermophysics to understand heat transfer in nanomaterials and structures. He started the energy aware computing initiative at IBM Zurich and triggered the Aquasar project to promote improved data-center efficiency and energy reuse in future green data centers. He currently leads the smart energy effort at IBM Zurich and as part of that has begun research efforts on concentrated photovoltaics, thermal desalination, and adsorption heat pumps.