Roadmap towards Ultimately-Efficient Zeta-Scale Datacenters

Patrick Ruch, Thomas Brunschwiler, Stephan Paredes, Ingmar Meijer, and Bruno Michel
IBM Research – Zurich, Advanced Thermal Packaging, Säumerstrasse 4, CH-8803 Rüschlikon Switzerland
bmi@zurich.ibm.com

Abstract—**Chip microscale liquid-cooling reduces thermal resistance and improves datacenter efficiency with higher coolant temperatures by eliminating chillers and allowing thermal energy re-use in cold climates. Liquid cooling enables an unprecedented density in future computers to a level similar to a human brain. This is mediated by a dense 3D architecture for interconnects, fluid cooling, and power delivery of energetic chemical compounds transported in the same fluid. Vertical integration improves memory proximity and electrochemical power delivery creating valuable space for communication. This strongly improves large system efficiency thereby allowing computers to grow beyond exa-scale.**

Keywords—datacenter; energy; reuse; packaging; cooling; power-supply; stacking

I. Introduction

Over the first decades of their development the volume occupied by electronic computers reduced by many orders of magnitude while performance and efficiency improved by similar factors. Since the introduction of the personal computer (PC), performance and transistor count on processors improved by six orders of magnitude from 2500 to 2,500,000,000, but form factor and communication distance via the printed circuit board did not change, creating a communication bottleneck. This slowed the efficiency improvements for PCs and servers causing an increase in power consumption in the last few years to an extent that energy cost now dominates over hardware cost. Datacenters currently consume ~2% of the overall electricity in the USA. As with other large industries, the energy consumed and the carbon footprint are important issues when assessing the efficiency of producing goods or services.

Low-power microservers continued to shrink more than 10 fold and reached almost 10 times better efficiency. Thus, packaging density initially developed parallel to performance but slowed after the introduction of the PC where development focused on transistor scaling thus opening a communication bandwidth and latency gap. This created two major roadblocks to further progress: power density and communication delays between processors and memory or other processors. With growing system sizes, spatial and temporal communication requires larger fractions of the overall system resources. For this reason, demand for memory and communication needs joint optimization because interconnect prediction evolves into system-level prediction.

978-3-9815370-0-0/DATE13/©2013 EDAA

The transition from 2-D scaling to 3-D integration offers an excellent opportunity to improve computer density and efficiency. Interlayer cooled chip stacks are a scalable solution allowing the integration of several logic layers each with massive amounts of main memory [1]. These systems use 3-D communication and ultra-compact cooling similar to a human brain or a bionic volumetric computer scaling. We explore how brain-inspired bionic packaging concepts can be transferred to future 3-D computers to eliminate bottlenecks. A processor architecture that imitates the mammalian brain promises a revolution. Compared with the mammalian brain, today's computers are terribly inefficient and wasteful of energy, with the number of computing operations per unit energy of the best man-made machines being in the order of ~0.01% of the human brain depending on the workload. To finally reach human level performance computer design needs to undergo several major paradigm changes:

II. Paradigm Changes

A. From Cooling to Energy Reuse

Currently, computers are inefficient not only in the processors but also in the energy-hungry air conditioners that are needed to remove the heat: At least half the energy in a moderate to warm climate is needed for cooling the servers in a datacenter. Similar to waste heat reuse 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. Low-resistance, exergy-conserving, liquid cooled datacenters run with up to 3 times lower energy cost and are close to carbon-neutral during operation. Higher cooling water temperatures without increased transistor temperatures require lower thermal resistances from the nanoscale devices to the cooler. Equally important is an effort to reduce the required temperature levels for the consumers of the datacenter waste heat. Space heating has undergone an innovation cycle in the course of the past years and has reached a high level of exergy efficiency. For adsorption chiller reduced thermal resistance between the heat exchanger and the adsorber and improved adsorption isotherms matches them better to low grade datacenter heat sources. The concept of direct waste heat utilization in computers has been recently demonstrated [2]. The ‘Aquasar’ 86 processor prototype has been extensively characterized in terms of energy and exergy efficiency [3]. Using this hot water cooling concept a large 20’000 processor supercomputer has been built at the Leibnitz Rechenzentrum in Garching, Germany establishing the concept as a commercial product. Although the paradigm change from cold air cooling to hot water energy re-use may seem small it is quite fundamental and influential since it opens the door to space filling systems and volumetric scaling.

B. *From Performance to Efficiency*

Up to ten years ago the cost for purchasing a computer exceeded the cost for energy to run it by a large factor. In the last decade this has changed so that now energy cost is higher than hardware cost. This triggers a paradigm change with severe consequences: While in the past it was important to get a maximal **performance** per processor, in the future it is crucial to get the maximum amount of computations per energy unit which we term **efficiency** and computers have to become energy aware.

Biology demonstrates energy-aware architectures with efficiencies of $10^{14\text{-}15}$ operations/J, four to five orders of magnitude better than current computers. A computer with the equivalent performance of a human brain currently requires about 200 kW or 10^4 times more energy (e.g., for Jeopardy!). Energy is required for device switching and for transmitting signals along wires. Since 2004, processor efficiency improvements stalled due to the end of voltage scaling and the overall energy fraction of transmitting signals versus switching has drastically increased. Computational energy efficiency is currently limited by the power required for communication despite the introduction of copper wiring and low-permittivity dielectrics. Logic-centric scaling created a widening disparity, which is often referred to as the memory wall. Memory latency now exceeds hundreds of clock cycles with a painful impact on performance and a disastrous impact on efficiency. Both performance (latency) and energy efficiency are dominated by communication. The efficiency of an algorithm depends on data movement in both time and space, and optimization efforts are needed here as well [4].

C. *From Areal Device Size to Volumetric Density Scaling*

Historically, the energy efficiency of computation doubled every 18 months. Improved computational efficiency was enabled by device size scaling: Ten orders of magnitude efficiency improvements have been achieved, but with the anticipated end of CMOS scaling, it is unclear how this will continue (Fig. 1) upward to biological efficiencies (green region). 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 (Fig. 1). The main driver for efficiency improvement, therefore, has been integration density. In most 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 the computer volume. Further CMOS device shrinkage is challenging due to the rising passive power fraction. But overall system volume reduction is possible when communication density, cooling and power delivery can be managed. This will trigger a paradigm change away from device size and areal density scaling to a system-level volumetric density scaling.

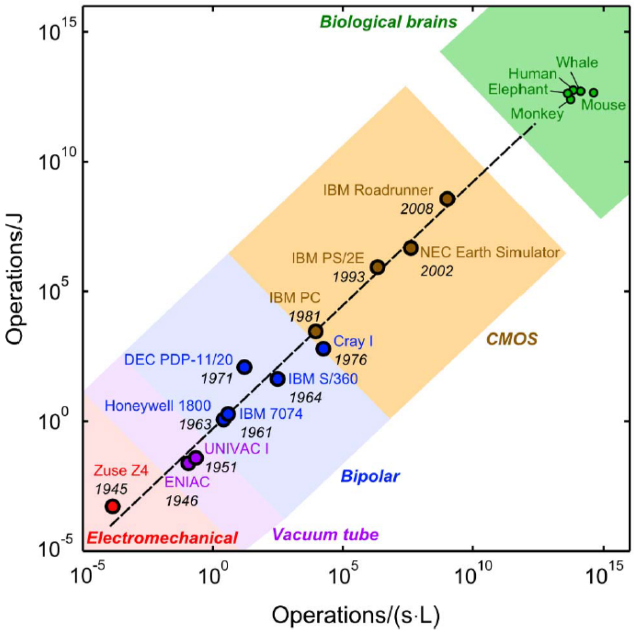

Fig 1. Computational efficiency and computational density of computers compared with mammalian brains [8].

III. VOLUMETRIC SCALING

Although heat re-use and efficiency paradigm changes are highly influential, the key paradigm change is the change from areal to volumetric scaling. This concept needs further explanation and a theoretical base. Since volumetric scaling systems are ubiquitous in biological organisms and brains we will explain the concept based on biological scaling concepts known as allometric scaling. Volumetric scaling also needs understanding of concepts like demand for communication, supply of surface, and power supply which are explained in paragraphs Rent’s rule and electrochemical power supply.

A. *Allometric Scaling*

Despite a large size range of 27 orders of magnitude in biological objects, many phenomena scale with size in a surprisingly simple fashion: Metabolic rate scales with 3/4 power of mass, timescales and sizes scale with 1/4 power of mass, and 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. Space-filled hierarchical fractal-like branching networks distribute energy and materials between reservoirs and microscopic sites.

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 [5]. In current computers ~96% of the volume 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 ~10% for energy and thermal supply and drain, and structural stability. The increased functional density in brains is a major driver for the improved efficiency (Fig. 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.

The allometric scaling rules allow a biological system to scale in size while keeping a similar basic blueprint and allow space-filling networks to grow sub proportionally with system volume. Since energy supply is delivered through the surface while energy demand is created in the volume, supply becomes more critical with large organisms because the surface grows less than the volume: The surface of a cube scales as $V^{2/3}$, where V is the volume. Generally, the hypersurface of a D-dimensional hypervolume scales with the (D-1)/D dimensional power of the hypervolume. This means large animals could not support high power densities. Hierarchical branched transport networks give access to four dimensional scaling: The 3/4 exponent allows higher power densities in large mammals. The 2/3 scaling of snails does not allow them to become large and severely limits their power density and performance.

B. Rent's Rule

Rent's rule relates the number of devices in a block with the need for communication. The efficiency depends on 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. Rent coefficient k gives the average number of connections per logic element, and Rent exponent p is a measure for the network complexity. The importance of a Rent analysis grows with system size: Systems with less than one billion functional elements are 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 [6]. VLSI architectures exhibit similar Rent exponents and network complexities as the human brain, but their computational efficiencies and functional densities lag by orders of magnitude (see Fig. 1) due to the lack of scalability in the third dimension.

Today, all microprocessors suffer from a break-down of Rent's Rule for high logic block counts because the number of interconnects does not scale beyond the chip edge. The limited number of package pins is one of the main reasons behind the performance limitation known as the memory wall. On-chip

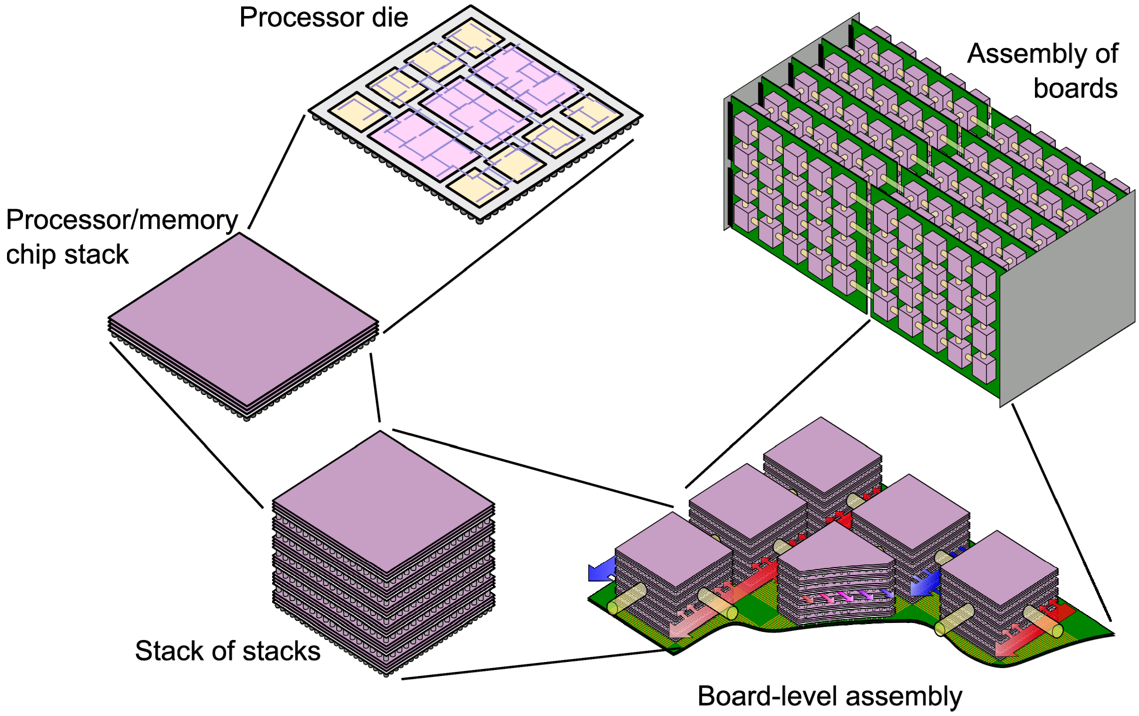

Fig 2. Volumetrically scalable stack of stacks. Note that the communication links between logic elements as well as the fluidic thermal and power I/O enable high-density space-filling supply and drain networks.

cache reduces the negative impact of this bottleneck at the expense of more than half of the chip area being allocated to memory. Serialization of the information stream enlarges overhead and latency. A dense, three-dimensional physical arrangement of semiconductor chips would allow much shorter and more parallel communication paths and a corresponding reduction in internal-delay roadblocks.

C. *Electrochemical Power Supply*

Going further in the new paradigm, another cause of energy inefficiency is the loss (in power and in space) in delivering the required electrical energy to the chips. The pins dedicated to power supply in a chip package easily outnumber the signal pins in high-performance microprocessors, and the number of power pins has been growing faster than the total number of pins. This power problem is essentially a wiring problem: The energy needed to switch wires is currently more than an order of magnitude larger than the energy needed to switch transistors. This disparate evolution of communication and computation is even more pronounced for latency because while transistors have become smaller the chip size has not followed suit, leading to substantially longer total wire lengths.

A new solution to this two-fold wiring problem, again designed following the mammalian brain, is to use the coolant fluid as the means of delivering energy to the chips. We use technological analogs for sugar as fuel: These analogs are inorganic redox couples which have been mainly studied for grid-scale energy storage in the form of redox flow batteries. These electrochemical systems offer superior power densities but would still be challenged to satisfying the power demand of a microprocessor. However, future computers which could be built around this fluidic power delivery scheme would be less power-intensive due to their reduced communication power demand. The fluidic means of removing heat allows huge increases in packaging density while the fluidic delivery of power with the same medium saves the space used by hardwired energy delivery. The savings in space allow much denser architectures, sharply reduced energy needs and latency as communication paths shrink.

On the path to building such systems the power density limits of model redox couples such as V(II)/V(III) and V(V)/V(IV) were explored using rotating disk electrode (RDE) measurements and fundamental data for mass transport and reaction kinetics were extracted. With suitable surface functionalization for carbonaceous electrodes, reaction kinetics were enhanced to the diffusion-limit. Implementation in analytical and numerical models of flow-through microfluidic cell geometries allowed an assessment of the power density as function of channel dimensions, flow rate and temperature. Experimental investigations of electrochemical conversion in single-channel redox flow systems are being carried out to explore approaches to overcome power density limitations.

IV. Discussion

The integration process in an ultra-dense computer starts from a processor die with many added layers of memory that extend the memory wall and allow ten times higher performance than a current processor. Next, ten of these systems are vertically stacked to a volume of 300 mm^3 and 100 elements are combined to a multichip module (MCM). Finally, ten 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. Maximum communication distances in the system are shorter than 50 cm and communication distances between processor and memory are shorter than 0.5 mm. The exact approach to scale the chip stack to a more than 1,000 cm^3 system following an allometric scaling factor still and hierarchical transport networks remains to be established. With a small density compromise, we still may be able to build an electronic brain with similar performance and efficiency as the human brain but slightly lower density (Fig. 2).

A. *Architectural Changes*

With technology scaling, transistors became cheap and abundant, but availability of chip and board-level wiring became scarce. Managing the demand for wires has been delayed since it requires architectural changes that are more difficult than expanding the transistor count. Stanley-Marbell showed that with current power density trends in processors, there will be eventually no pins left for communication [7]. 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. Core-level multithreading is a strong driver for memory demand and triggers a faster encounter with the memory wall. It is clear that low fan-out architectures can be extended to larger block sizes. The communication demand has to be limited from the very beginning by low fan-out micro-architectures. The onset of the cooling, power, memory, and communication wall occurs as the I/O need of a given block exceeds the available surface.

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. 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. This connection "space" or "supply" is then compared with the need for connections.

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 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 to more FLOPs per element.

B. Scaling to Exa- and Zeta-Scale Systems

A coarse assessment of the power and volume demand for stacked systems in combination with the proposed bio-inspired packaging scheme can be performed based on current device and wire characteristics, with wiring needs derived from Rent's rule and its volumetric extension. 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.

We consider different reference cases based on varying I/O demands and supply scaling with increasing element count. Present-day scaling behavior (2-D scenario) based on a complex communication architecture which is dominated by the overwhelming demand for wiring, which results in both an over proportional power scaling coefficient 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 Fig. 3: Scaling to 10^{18} elements results in 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. 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 (Fig. 3(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 (Fig. 3(b), red line) becomes dominated by cooling for large element sizes. Overall, a peta-scale machine can be built within 50 m^3 and a power budget of 400 kW.

For an exa-scale 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 can only partially 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 (Figs. 3(a) and 3(b), dashed red lines). The scalability can only be improved by allowing for a change in the way in which power and cooling are supplied.

C. Bionic Packaging

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 under proportionally 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 ~$V^{3/4}$ instead of ~V. A simplification of wiring is assumed since

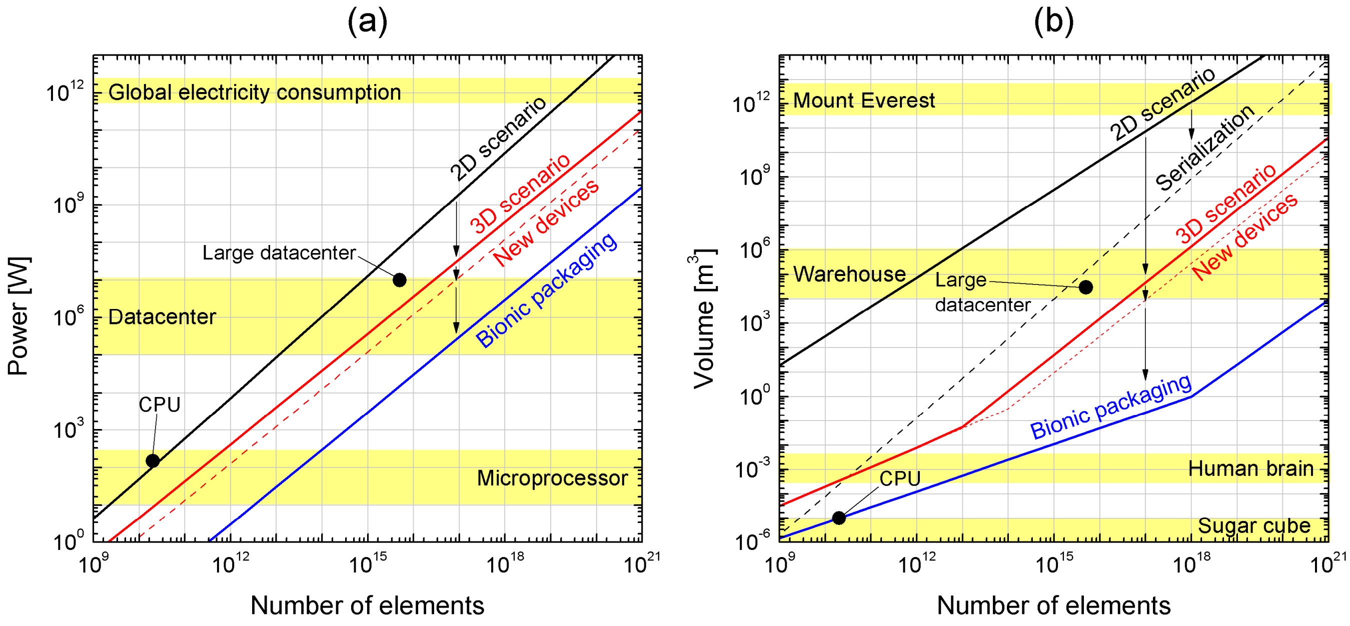

Fig 3. (a) Power and (b) volume of computers as functions of element count for different scaling scenarios [8].

power is delivered electrochemically. The improved volumetric scalability results in the most compact system size (Fig. 3(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 dominates 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 peta-scale system built with 2-D technology fits in a few 1,000 m^3 using 10 MW, a zeta-scale 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 zeta-scale systems, respectively.

Importantly, today's transistors are already dense enough to sustain such high density systems: They occupy 1 cm^3, 1 dm^3, and 1 m^3 for peta-, exa-, and zeta-scale systems, respectively. The key toward volume reduction therefore lies in the volumetric packaging. Figure 3 shows that the transition from the current 2-D scenario (black dashed line) to a 3-D scenario can shrink the volume of a peta-scale computer by a factor of 300, and by a factor of 1,000 for a continued device development. 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 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 at least one-million-fold volumetrically and 100-fold linearly.

V. Summary and Conclusion

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 to zeta-scale systems. We conjecture that the information industry could reach the efficiency of biological brains once a packaging approach similar to the one of human brains can be implemented. Economically, it is important that the cost per transistor continues to shrink. 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 ensure that interconnect and memory cost that dominate in future systems over (logic) transistor cost will show a similar Moore's trend in the next 20 years to make technological predictions of this paper economically feasible. It remains to be seen whether a novel device technology could be mass produced by then that provides better base switch efficiency while not jeopardizing integration density.

Based on the proposed paradigm changes integration density can improve in the next 20 years to similar values as biological systems allowing similar compute efficiencies. The bionic packaging copies the integration density, the power supply, and the cooling approaches of the human brain – things that we understand since more than a century since microscopy of brains was established. It does not distinguish whether the computing architecture still uses a von Neumann concept or switches to non von Neumann concepts. Neural networks could also be supported by bionic packaging but this computing approach will have to prove its benefits independent of packaging density. Finally, the potential computing power of DNA could be used in a more distant future on our way to copy the – Structure – Function – and finally material base of biological brains. Our understanding of bionic computing does not include an identical copy of the brain by using neurons, however. The goal is to reach the computational performance of the brain by a partial copy of concepts and to repeat the Jeopardy competition in two decades with machine and human using the same power.

Acknowledgment

The authors wish to thank Ronald Luijten, Philip Stanley-Marbell, Wilfried Haensch, Thomas Theis, Tobias Rapp, and Thomas J. Schmidt for discussions, as well as Walter Riess for support.

References

[1] T. Brunschwiler, B. Michel, H. Rothuizen, U. Kloter, B. Wunderle, H. Oppermann, and H. Reichl, „Interlayer cooling potential in vertically integrated packages," Microsyst. Technol., 15(1), 57–74 (2008).

[2] T. Brunschwiler, G.I. Meijer, S. Paredes, W. Escher, and B. Michel, "Direct Waste Heat Utilization from liquid-cooled supercomputers," Proc. 14th Intl. Heat Transfeer Conf. IHTC14, August 8-13, 2010, Washington DC, USA.

[3] S. Zimmermann, I. Meijer, M.K. Tiwari, S. Paredes, B. Michel, and D. Poulikakos, „Aquasar: A hot water cooled data center with direct energy reuse", Energy 43, 237-245 (2012).

[4] S. Moore and D. Greenfield, "The next resource war: Computation vs. communication," Proc. Int. Workshop Syst. Level Interconnect Prediction, 81–85 (2008).

[5] G. B. West, J. H. Brown, and B. J. Enquist, "A general model for the origin of allometric scaling laws in biology," Science, 276 (5309), 122–126 (1997).

[6] 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. 12 (12), 1330–1347 (2004).

[7] P. Stanley-Marbell, V.C. Cabezas, R.P. Luijten, "Pinned to the walls - Impact of packaging and application properties on the memory and power walls," Intl. Symp. Low Power Electronics and Design (ISLPED 2011) 1-3 Aug. 2011, Fukuoka, Japan.

[8] P. Ruch, T. Brunschwiler, W. Escher, S. Paredes, and B. Michel, „Toward five-dimensional scaling: How density improves efficiency in future computers," IBM J. Res. Develop. 55(5), 15:1-13 (2011).