# Rebooting Computing

## Developing a roadmap for the future of the Computer Industry

**Thomas M. Conte, Erik DeBenedictis, Paolo A. Gargini, Alan Kadin, Elie K. Track**

**Sommario**

*La Legge di Moore ha ormai regnato indiscussa per 50 anni ed in pieno accordo con la sua previsione il numero dei transistori è cresciuto in modo esponenziale creando di conseguenza la rivoluzione informatica. Ci si chiede, è possibile che una crescita (esponenziale!) possa durare per sempre? Ci stiamo forse avvicinando al momento in cui non sarà più possibile ridurre ulteriormente le dimensioni dei transistori? Siamo forse vicini alla fine della rivoluzione informatica? Il seguente articolo suggerisce che al contrario il prossimo decennio vedrà una rinascita (Rebooting) di tutta l'industria informatica attraverso una completa riprogettazione da cima a fondo di entrambi hardware e software. Questa riformulazione consentirà una continuazione della crescita (ancora esponenziale!) delle capacità di elaborazione dati mantenendo la rivoluzione industriale "viva e vegeta". Il fulcro di questa rinascita risiede nella cooperazione e convergenza di due iniziative complementari. Queste sono "IEEE Rebooting Computing Initiative" (RCI) e la "International Roadmap for Devices and Systems" (IRDS).*

*Buona lettura!*

## Abstract

*The 50-year reign of Moore's Law, with its exponential increase in integrated circuit density, has created the computer revolution. Is device scaling coming to an end, and will this lead to the end of the revolution? On the contrary, we suggest that the next decade will see a "rebooting" of the entire computing industry, by redesigning computer hardware and software from top to bottom. This will enable continued exponential growth of computing capabilities, keeping the computer revolution alive and well. We describe preliminary efforts along these lines, focusing on the "IEEE Rebooting Computing Initiative" (RCI) and the "International Roadmap for Devices and Systems" (IRDS).*


**Keywords:** Computers, Device scaling, Moore's Law, solid-state circuits, VLSI, CMOS, Nanotechnology, Neuromorphic computing, Approximate computing, Parallel computing, Power efficiency, Supercomputers, Computer networks

## 1. Introduction

Over the past 50 years, the Computer Industry has fueled the information revolution. Society has been completely changed by the introduction of personal computers, smartphones, tablets and many other devices that have become part of everyday life. In addition, progress in High Performance Computing has allowed solving of the most complicated problems in relatively short times. The foundations of this revolution were the von Neumann computer architecture and the invention of the integrated circuit. However, in the last 10 years the progress in computational performance has substantially slowed down due to limitations in operational performance imposed by limits on power dissipation of integrated circuits, increases in signal propagation delays, and intrinsic limitations imposed by the von Neumann architecture. We propose that this state of affairs can be changed only by a major effort of revisiting all of the conventional assumptions, and "rebooting" the entire computer industry.

The computer architecture proposed by von Neumann and the demonstration of the transistor occurred in the late 1940s, and laid the foundation of the modern computer industry. The enabling integrated circuit was commercialized in the early 1960s, and led to the exponential growth in transistor densities reflected in Moore's Law [1]. Progress in both computer microarchitecture and semiconductor technology allowed enhancing the performance of the original computer architecture beyond all expectations. Each new generation of scaled-down transistors enabled computers to operate at higher frequency, performing more operations per second than the previous generation. This in turn enabled such design techniques as deep pipelining, speculative execution and superscalar microarchitectures. As a result, computational performance continued to improve without programmers' being aware of any dramatic change in the von Neumann architecture.

However, fundamental power limits (due to excessive local heating) were reached by the middle of the last decade when microprocessors tried to operate well beyond the 100W power level. These physical limits prevented any substantial increase in pipeline depth and operating frequency from being realized to further enhance computing performance, even though transistors could operate at yet higher frequencies and larger numbers of transistors were available, in accordance with Moore's Law, with each new technology generation.

In response, the computer industry created multicore processors that required substantial rewriting of software in order to scale computer performance. But engineering of software for the von Neumann architecture was a difficult endeavor. The further need for explicit programming of parallelism made this largely untenable. As a result, computer performance stalled.

These problems have been widely recognized within the computer engineering community. In order to address these problems more coherently, the *Institute of Electrical and Electronics Engineers* (IEEE) created a new initiative in 2012, under the auspices of the IEEE Future Directions Program, called the *Rebooting Computing Initiative* (RCI). RCI is a multi-disciplinary effort of volunteers from 10 IEEE Societies and Councils, which leverages IEEE's pre-competitive community to explore ways to restore computer performance to its historic exponential growth path. RCI works from a holistic viewpoint, taking into account evolutionary and revolutionary approaches to rethink the computer "from soup to nuts" including all aspects from device, through circuit, architecture, software, algorithms, and applications. RCI has sponsored a series of Summits and co-sponsors workshops and conferences on related topics. For more information, see http://rebootingcomputing.ieee.org .

Over the same period, the limitations of device scaling were also recognized by the semiconductor industry planning consortium, the International Technology Roadmap for Semiconductors (ITRS). ITRS was founded in 1998, and has been sponsored by the five leading chip-manufacturing regions in the world: Europe, Japan, Korea, Taiwan, and the United States. While ITRS traditionally focused predominantly on device scaling, in 2012 it initiated a process of reorganization that recognized the need to go beyond device scaling to encompass a broader focus on computer systems and applications. ITRS 2.0 is organized into 7 Focus Teams: System Integration, Heterogeneous Integration, Heterogeneous Components, Outside System Connectivity, More Moore, Beyond CMOS and Factory Integration. For more information, see http://www.itrs2.net/.

In 2014, RCI and ITRS 2.0 initiated collaboration with joint workshops. In 2015 the two organizations agreed that the development of a new paradigm for future computing requires a synergistic integration of new computer architectures with new revolutionary devices. The goal of this cooperation is to create a new roadmap to successfully restart computer performance scaling. This new vision of the roadmap was well beyond the boundaries of the historical semiconductor industry trends and required a new and broader platform to properly operate.

In May 2016, IEEE announced the formation of the *International Roadmap for Devices and Systems* (IRDS). IRDS will synchronize and merge system requirements with present and future devices capabilities under the auspices of the Industrial Connections program of the *IEEE Standards Association* (IEEE-SA). For more information, see http://irds.ieee.org

These joint efforts will be expedited by assistance and coordination provided by various governmental initiatives, which will help align efforts in industry with those in academic research programs. Recently announced programs in the United States include the *National Strategic Computing Initiative* (NSCI) and the *Nanotechnology-Inspired Grand Challenge for Future Computing*. For further information see https://www.whitehouse.gov/the-press-office/2015/07/29/executive-order-creating-national-strategic-computing-initiative and http://www.nano.gov/node/1522.

## 2. The Problem and How We Got Here

In the beginning, von Neumann described his Computer Architecture in a report in 1945. It identified a processing unit containing an Arithmetic Logic Unit (ALU) and several registers, a control unit containing an instruction register and program counter as well as memory units for data and instructions. Access to a large external memory storage unit was also part of the overall structure. This simple view enabled complex software to be written and debugged in a relatively straightforward manner. In this way, the software industry was born. In 1965 Gordon Moore predicted that it would be possible to double the number of useable transistors every year by means of design evolution and technology improvements. Continuing this trend for 10 years would yield 65,000 transistors available to design a product. In essence, he asked the question, “How could a system designer take advantage of this abundance of transistors?” Between 1972 and 1974, Robert H. Dennard announced a new methodology that allowed one to reduce the size of a transistor and also predict all of its electrical properties. This methodology acquired the common name of “Geometrical Scaling”. By the time Gordon Moore made his second prediction in 1975 (the number of transistors will double every 2 years), more than 40 companies had been launched in Silicon Valley. In parallel with this, IBM launched Project Stretch in 1960 to study ways to enhance computing performance through changing the organization of the computer. Computer architecture was effectively made a discipline. IBM continued to dominate with Gene Amdahl’s idea of the IBM 360: separating the microarchitecture from the instruction-set architecture. Multiple models of the 360 could be made, all with different microarchitectures, all capable of running the same software without recompilation. This led to the invention of instruction-level parallelism and out-of-order execution by Robert Tomasulo for the IBM 360 model 90, and in parallel (pun intended) by Jim Thornton and Seymour Cray at Control Data for the CDC 6600.

These techniques enabled computer performance to grow while maintaining the illusion of the von Neumann architecture to the programmer. The modern computer’s performance growth is a result of the combination of these two

mega-trends that enabled computer performance to grow exponentially from generation-to-generation:

(1) the rapid increase in semiconductor technology, and

(2) the rapid increases in computer architecture enabling cross-generation binary code compatibility.

The Personal Computer (PC) market was born in the mid-1980s, and soon locked the semiconductor industry and consequently the computer industry into a very unusual situation: the system manufacturers were all using microprocessors and software mostly produced by only two companies, Intel and Microsoft. In a market where silicon technology and architecture as well as software architecture were defined and locked into a "backward/forward compatibility mode," only one main avenue remained opened for the whole ecosystem to make progress: the lessons of the IBM 360 needed to be re-learned. In order to keep microprocessors based on the Intel's x86 architecture improving at a pace of doubling every two years, the industry went back to the 1960s and re-implemented complex microarchitectures from Project Stretch, the IBM 360 model 91, the CDC 6600, etc. Combined, these "tricks" worked behind the scenes to enable instructions to be run in parallel. Over time, these tricks in general became known as "superscalar" microarchitectures. All along, pipelined superscalar microarchitectures enabled designers to increase frequency (f) each and every generation. This was an easy solution indeed, but it came at a price. There is a high-level relation that ties together the key electrical parameters of any technology: $P=CV^2f$, where C is the device capacitance. So, making any new microprocessor faster implied operating at a higher frequency at the expense of an increase in power. Of course, reducing the operating voltage could somewhat reduce the power increase. However, frequency of operation was increasing faster than the any decrease in voltage. In few words, any voltage reduction was only delaying the unavoidable power debacle. Further worsening the situation was that superscalar microarchitectures were enabling higher frequencies though deeper pipelines. This meant that more instructions needed to be "in flight" than was possible by waiting for branch instructions to execute. This lead to "speculative execution": predicting what path a program would take and then doing that work ahead of time, in parallel. Thus higher frequencies meant deeper pipelines, which in turn required more and more speculatively executing instructions. But no prediction is 100% accurate. Invariably, these microprocessors did a lot of extra, wasted work by mis-speculation. The deeper the pipeline, the more power was wasted on these phantom instructions.

But increased performance was the name of the game, and operating frequency kept on increasing through the 1990s until the processor exceeded the 100W operating level! Crossing this power threshold required a drastic change in cooling techniques, inconsistent with the PC hardware of the time. Increasing operating frequency as the main tool to increasing computing performance [see Fig. 1] was *No Longer Viable*! The consequence of reaching the power wall had implications beyond the PC industry. The PC and microprocessor ecosystem had driven the cost of mainstream processors to a very low cost, fostered by the

continually increasing volumes of logic ICs. As a result these types of microprocessors and also other main elements of the PC ecosystem had migrated upward, affecting systems operating at much higher level of complexity than PCs. Supercomputers were being built using microprocessors. The microprocessor crisis had infected the whole computer industry all the way to the High Performance Computing (HPC) level!

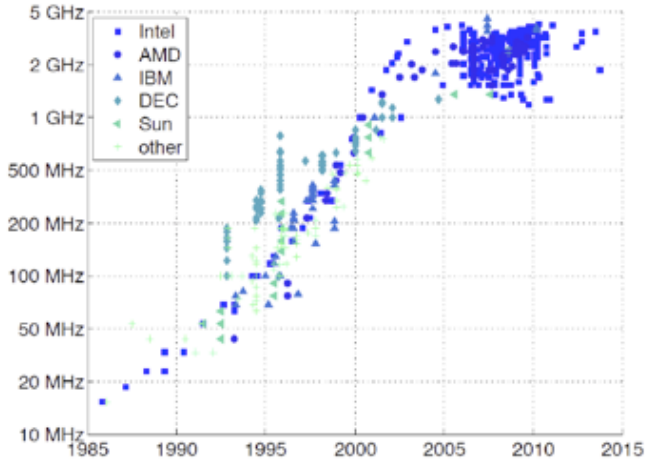


**Figura 1**
*Historical trends in clock speeds of microprocessors, showing saturation in speed by 2005.*

By 1995, it had actually became clear that a crisis of unprecedented dimensions was looming on the horizon; by 2005, at the latest, no further scaling of transistors according to Geometrical Scaling was possible. Furthermore, interconnect propagation was becoming larger than transistor delay. Analysis of the Pentium family showed also that 50% of the dynamic power was consumed in interconnections! Focus Centers Research Program (FCRP) was launched in 1997 as an alliance between IC companies, equipment suppliers and DARPA with the goal of promoting university research in the US on technology challenges for the next 10 years. Between 1999 and 2003, the semiconductor industry converted to copper interconnects and low-k inter-metal dielectrics. In 1998, Paolo Gargini (then Intel Director of Technology Strategy) realized that these were international problems. In order to address these, the *International Technology Roadmap for Semiconductors* (ITRS) was launched in association with organizations from U.S., Europe, Japan, Korea and Taiwan. The ITRS identified that the end of Dennard's scaling was imminent and outlined a set of revolutionary innovations aimed at the continuation of historical trends of the semiconductor industry; this new scaling paradigm was named "Equivalent Scaling". This approach has kept the semiconductor industry successful for the past decade, and will still remain the primary guide to the semiconductor industry until the end of this decade and beyond.

As for microprocessors, the transistors per chip kept on increasing, but the superscalar microarchitecture had stalled out. Frequency scaling was dead and along with it, "hiding" parallel execution using superscalar speculative execution while maintaining generation-to-generation software compatibility. What were the microprocessor vendors to do with these extra transistors? The solution was obvious: place more than one processor "core" on a chip and let programmers worry about how to keep them busy. The Multicore Era was born.

In parallel with the industry-led efforts, in 2000 the US government launched the National Nanotechnology Initiative (NNI, http://www.nano.gov/) under the guidance of Mike Roco of the US National Science Foundation (NSF). The NNI announcement triggered an escalation of investments in Nanotechnology across the world [Fig. 2]. Both Governments and IC industry were, as many times in the past, on a cooperative course and synchronization of efforts was essential. Paolo Gargini launched the Nanoelectronics Research Initiative (NRI) in 2005 with the cooperation of leading US semiconductor companies, NSF and NIST. The goal of NRI (https://www.src.org/program/nri/) consisted in identifying and developing new types of transistors operating under new physical phenomena. These devices presented some features different from CMOS. In the past 5 years several interesting candidates have emerged and are being intensively developed. However, it should not be expected that any of these new types of transistors could be a simple "plug in" replacement for CMOS. It is expected that Equivalent Scaling by itself will not be able to continue historical trends through the next decade, but fortunately a new approach is within reach.

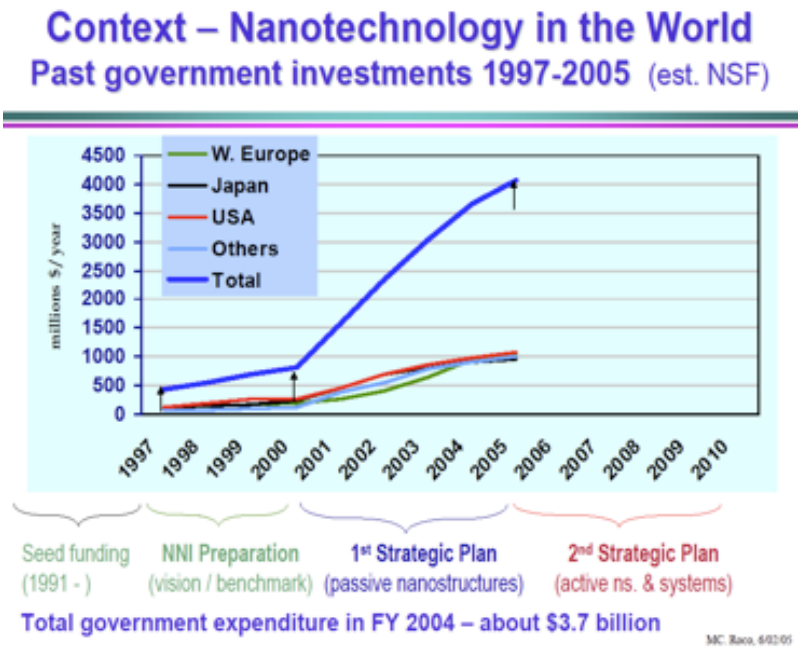


**Figura 2**
*Past trends in government research funding for nanotechnology across the world, showing sharp rise starting around 2000.*

The semiconductor industry will be approaching transistor features (around 5nm) in the next decade; these features are at the limit of the functionality of MOS transistors, the industry standard. However a new scaling paradigm is underway addressing the two major limitations upcoming in the next decade: available space for more transistors and power. In the 3D Power Scaling approach, the planar transistor is rotated along the source edge by 90 degrees; the transistor is standing up supported only by the outmost edge of the source. This methodology allows continuing packing transistors at Moore's Law pace, but in addition, multiple planes of verticals transistors can be stacked on top of each other. In fact, columns of transistors can be grown by multiple sequential depositions, and then connections between the transistors in plane and from plane to plane can be made [Fig. 3]. Memory IC makers have already announced Flash memory chips stacking as many as 48 layers of transistors built with this 3D approach, producing a staggering 128Gbit memory. Aggressive forecast of 1Terabit Flash memories have been presented [Fig. 4]. The number of transistors will continue to grow and even faster than Moore's Law!

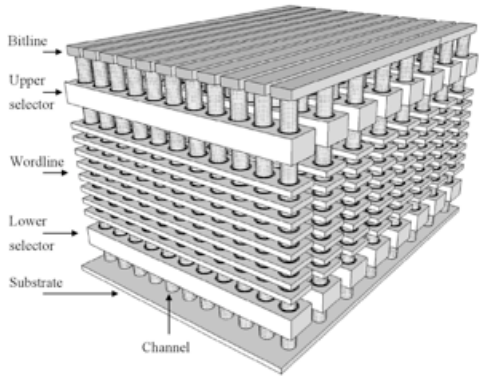


**Figura 3**
*Conceptual diagram of 3D Memory Chip.*

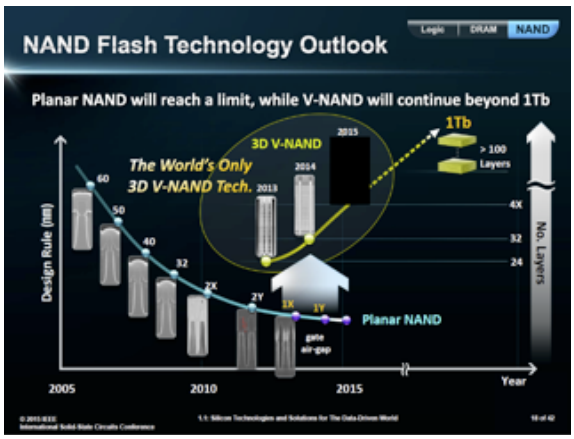


**Figura 4**
*Trends in Flash Memory capacities, focusing on impact of 3D integration*

On the power-reduction side, tunnel transistors (also known as TFETs) have shown the capability of running similar to MOS transistors, but with almost no leakage current. Would a drastic reduction in power consumption in the "off" condition be sufficient to extend the historical performance trends? These TFET transistors also present a very abrupt transition from the "on" condition to the "off" condition, and this feature could allow further reduction in power supply voltage. Some other types of devices operating at much lower frequency than MOS transistors but capable of storing information in a nonvolatile mode and using much less power have also been demonstrated [Fig. 5]. See Nikonov [2] for further information.

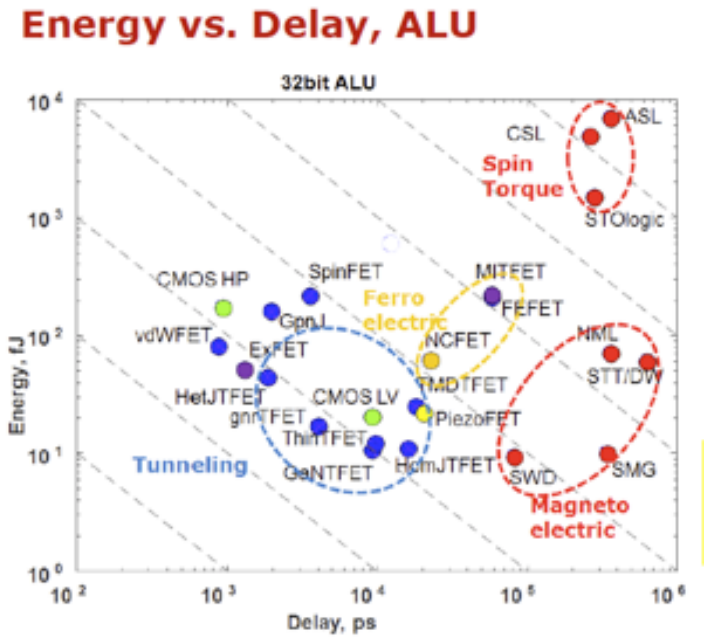


**Figura 5**
*Energy vs. delay for arithmetic logic units in conventional and exotics electronic technologies. See Nikonov [2] for further information.*

## 3. Where Do We Go From Here?

It should be clear from the previous paragraphs that computing performance increased from the mid-1980s until 2005 as a result of higher frequency of operation and superscalar's clever ways of keeping the processing unit busy all the time. By 2005 the MOS technology reached a fundamental power wall [Fig. 6] that prevented microprocessors' designers from further increasing operational frequency. Leading IC makers have continued to reduce the size of transistors and increase their number [Fig. 7] according to Moore's Law for the past 10 years just like they had done since 1975, suggesting that nothing has changed! Transistors could operate at higher frequency than in the past, but they would self-destruct due to overheating. Microprocessors have evolved over time. But

like biological evolution, sometimes there occurs a catastrophic event that brings about a new order. This is the theory of Punctuated Equilibrium as proposed by evolutionary biologist Stephen J. Gould. Computing has clearly reached a catastrophe: the Power Wall. But, what would the next era, the "new equilibrium" of computing look like?

With this understanding we can now correctly formulate the problem: The IC industry has produced and will continue to produce smaller, faster transistors at the rate predicted by Moore's Law. Smaller dimensions, higher switching frequencies and more transistors will remain possible in the future, but these transistors will not be operated at frequencies that would allow microprocessor power dissipation exceed a 100W limit because the circuit would self-destruct. This limitation has brought the rate of progress in Computing Performance to a snail's pace. A new way of computing is urgently needed.

3D Power Scaling gives us however a glimpse of how the process and the circuit architecture could positively affect the re-engineering of Computer Architecture. In the 3D Power scaling approach, a logic block could have memory, registers and other circuits stacked in the planes immediately above and below. This would greatly reduce the distance interconnect lines have to travel and also their cross section could be greatly increased with consequent reduction in signal propagation delay.

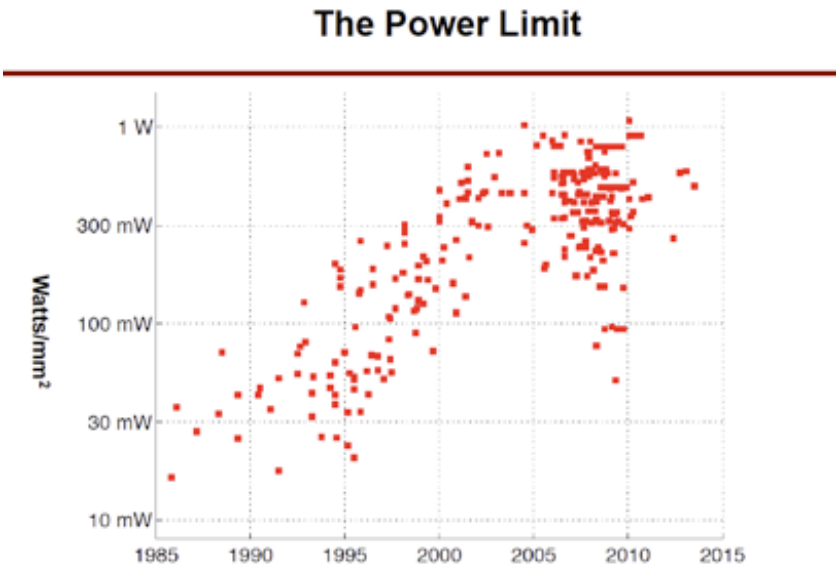


**Figura 6**
*Trends in power density in microprocessor chips, showing saturation starting around 2005.*

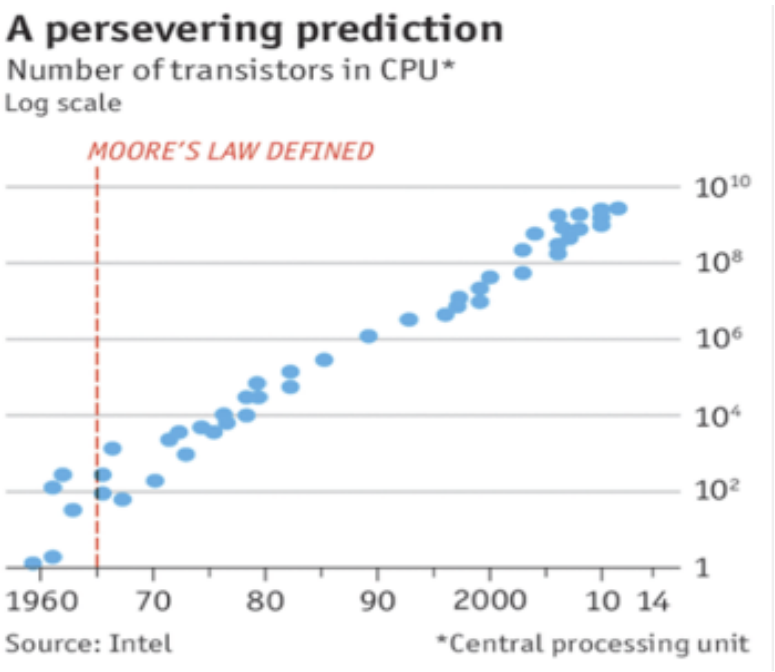


**Figura 7**
*Trends in number of transistors in Intel microprocessors, showing continuation of Moore's Law scaling, even while performance saturated.*

## 4. Rebooting Computing

In 2011, Elie Track, then-president of the IEEE Council on Superconductivity, and Tom Conte, professor of CS and ECE at Georgia Institute of Technology and himself a computer microarchitecture researcher, independently realized that computing itself needed to take a new direction. Because Conte was the then vice-president of the IEEE Computer Society, the two met and shared ideas at an IEEE event in January of 2012. They both decided that any change had to be fundamental and incorporate changes all the way up the "Computing Stack," from the device level, to circuits, to architecture, on up to algorithms and applications themselves. The von Neumann architecture itself could no longer be propped up. Everything in the computer needed to be re-thought, "from soup to nuts." The only place where experts in every level of the computing stack meet is in the IEEE; thus, they realized that IEEE itself would be the catalyst to enable this change. Conte coined the term "Rebooting Computing," and the IEEE Rebooting Computing initiative (http://rebootingcomputing.ieee.org) was born, under the umbrella of IEEE Future Directions. RCI is co-chaired by Track and Conte, with a committee composed of IEEE volunteers from 10 Societies and Councils: Circuits and Systems Society (CAS), Council on Electronic Design Automation (CEDA), Computer Society (CS), Council on Superconductivity (CSC), Electron Devices Society (EDS), Magnetics Society (MAG), Reliability Society RS Components, Packaging, and Manufacturing Technology Society

(CPTM), Council on Nanotechnology (NT), and Solid-State Circuits Society (SSC). This has created a unique interdisciplinary flavor in the RCI.

With funding from the IEEE Technical Activities Board (the body inside the IEEE that encompasses all technical societies and councils across the discipline), IEEE RCI began holding invitation-only summits to begin brainstorming a way forward. (Overviews of RC Summits are available from [3], http://rebootingcomputing.ieee.org/rc-summits.) The first IEEE RCI summit was held in Washington, DC in December 2013, and included thought leaders from major government agencies, the White House Office of Science and Technology Policy, industry giants and accomplished academics. The exercise produced the realization that there were three Pillars to Rebooting Computing [Fig. 8]: (1) New and Emerging Applications that drive the need for computer performance, (2) Power and Energy limits that brought about the demise of the Von Neumann architecture, and (3) Secure computing, because, as the group reasoned, if one were to re-invent the computer, it should be made implicitly secure from the start.

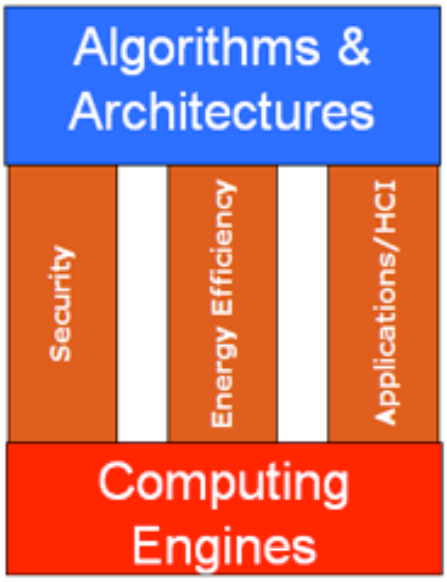


**Figura 8**
*The three Pillars of Future Computing as identified in the 1st Rebooting Computing Summit (2013). These sit atop the devices (the computing engines), and in turn support high-level algorithms and architectures.*

Over the course of 2014, IEEE RCI held two additional summits, both in the Silicon Valley area. The second summit (RCS 2) looked at "new engines of computing." Old and new ideas in how to compute were welcome to the table. Topics included Neuromorphic computing [4] (also called brain-inspired computing) and Approximate Computing [5], as well as Energy-Efficient Computing approaches [6]. These approaches are summarized in Table I below:

| Approach | Advantages | Research questions |
|---|---|---|
| Asynchronous circuits | Known potential for speedup | Design tools, complexity |
| Adiabatic/reversible computing | Could enable far lower power | All known approaches clock slower: requires more inherent parallelism to compensate |
| Neuromorphic | Proven for recognition problems | Programmability, repeatability/reliability of results |
| Computationally error tolerant | Enables < 1 *volt* operation | Codec could consume all potential power gains, proof-of-concept prototyping needed |
| Random, Stochastic and Approximate | Leverages current over-computing of accuracy/ precision | Programming languages, application space expansion required, potentially one-time speedup |
| Memory-centric, near memory processing | Many problems are memory bound, could build on 3D | Needs more prototyping, application space expansion required, potentially one-time speedup |

**Table I**
*Alternative approaches to Rebooting Computing – advantages and disadvantages.*

The third summit (RCS 3) looked at security and algorithms. The consensus of this summit was that there were select classes of problems, some old and some new, that would be the drivers in the years to come: the demands of Big Data, the need for ever-more-accurate yet fast recognition/machine learning, the need to improve the speed of solving optimization problems, the requirements of computational science and its simulation of physical systems, the requirements of simulation of engineering systems, the need for computationally strong encryption, acceptable yet efficient processing of multimedia data, and enabling truly immersive human-computer interaction. This is of course only a partial list, but it represents the key challenges to what and how we may compute in the future. Many of the "rebooted computer" ideas explored by IEEE RCI [Table I] take advantage of properties of semiconductor devices heretofore thought of as undesirable: unreliable switches, multi-valued (analog) properties, slower yet far

more power-efficient gates, devices that work as both logic and memory, but not optimally for either, etc.

The 4th and final RC Summit (RCS 4) focused on several alternative tracks, including Extending Moore's Law through novel devices (such as tunneling FETs, memristors, spintronic elements, and carbon nanotubes), Approximate/ Probabilistic Computing, Neuromorphic Computing, and Superconducting Computing.

Recognizing the needs for low-power approximate computing for widespread applications, RCI has sponsored an annual technology competition directed primarily at students: the *Low-Power Image Recognition Challenge* (LPIRC). This challenge, held in connection with the Design Automation Conference (DAC), aims to discover the best technology in both image recognition and energy conservation. Winners will be evaluated based on both high recognition accuracy and low power usage, with speed another consideration. For further information, see http://lpirc.net/ and [7].

In addition to the RC Summits, RCI hosted a special issue of *IEEE Computer Magazine* in December 2015 on *Rebooting Computing* [8], with special editors Tom Conte and Elie Track. Contributions addressed a variety of conventional and exotic technological approaches. An overview comparison of alternative technologies is provided by Shalf and Leland [9]. Other approaches include superconducting computing [10], memory-centric computing [11], and 3D memory-logic integration [12].

Following the success of the RC Summits, the 1st IEEE International Conference on Rebooting Computing (ICRC) was held in San Diego, California, October 17-19, 2016. The goal of ICRC 2016 was to discover and foster novel methodologies to reinvent computing technology, including new materials and physics devices and circuits, system and network architectures, and algorithms and software. This covered some exotic approaches to future computing such as superconducting computing, neuromorphic computing, nanocomputing, and even quantum computing [10, 13]. See for more information. A second ICRC 2017 is being planned for October 2017 in Washington, DC, USA. See http://icrc.ieee.org for more information.

## 5. Government-Coordinated Research & Development Programs

We spoke earlier about the key role of the US *National Nanotechnology Initiative* in promoting worldwide R&D into nanoscale semiconductor devices and circuits in the last decade. Similar large-scale government-industry coordinated programs will be necessary to stimulate and catalyze the next transition in the computer industry over the next decade and beyond.

An important new program in this regard is the US *National Strategic Computing Initiative* (NSCI), which was announced in July 2015, but is still being formulated (see, for example, https://en.wikipedia.org/wiki/National_Strategic_Computing_Initiative). While NSCI focuses on high-

performance computing (HPC, also known as supercomputing), it also features research and development of novel computing technologies that will impact computing systems from data centers to smartphones to the internet of things (IoT). This aligns strongly with the goals of the IEEE Rebooting Computing and the International Roadmap for Devices and Systems (IRDS). NSCI is a broad interagency multi-year initiative, with participation from the Department of Energy, Department of Defense (DoD), National Institutes of Health, National Science Foundation, National Aeronautics and Space Administration (NASA), National Oceanographic and Atmospheric Administration, and National Institute of Standards and Technology. The Intelligence Research Projects Agency (IARPA) within DoD will take the lead at promoting alternative technologies beyond CMOS, including superconducting and quantum computing.

Another new US government program is an offshoot of the current National Nanotechnology Initiative. The *Nanotechnology-Inspired Grand Challenge for Future Computing*, announced in October 2015 (http://www.nano.gov/futurecomputing), challenges researchers and industry to *"Create a new type of computer that can proactively interpret and learn from data, solve unfamiliar problems using what it has learned, and operate with the energy efficiency of the human brain. … This grand challenge will bring together scientists and engineers from many disciplines to look beyond the decades-old approach to computing based on the Von Neumann architecture as implemented with transistor-based processors, and chart a new path that will continue the rapid pace of innovation beyond the next decade. … Many of these approaches will require new kinds of nanoscale devices and materials integrated into three-dimensional systems. These nanotechnology innovations will need to be developed in close coordination with new computer architectures and informed by our growing understanding of the brain."*

A team led by RCI and ITRS (together with HP and Sandia) has submitted a proposal in response to the Grand Challenge, focusing on the development of "Sensible Machines" that would incorporate a neuromorphic computer based on nanoscale devices into a system smaller than a millimeter, capable of automated learning. For further information, see the White Paper at http://rebootingcomputing.ieee.org/images/files/pdf/SensibleMachines_v2.5_N_IEEE.pdf

The focus on Brain-Inspired Computing is also reflected in major research programs on the Brain in both US and Europe. These include the US BRAIN initiative (https://www.whitehouse.gov/BRAIN) and the European Human Brain Project (https://www.humanbrainproject.eu/). While these are primarily focused on neuroscience, both projects also include key components in computer science and engineering, and on development of computers that can compute more efficiently based on neuromorphic foundations.

There are also a series of innovative research programs in future computing technology sponsored by the US *Semiconductor Research Corporation* (SRC, https://www.src.org/), which is affiliated with the US-based *Semiconductor Industry Association* (SIA, http://www.semiconductors.org/). SRC coordinates with the US National Science Foundation (NSF) to provide grants to university

researchers in advanced device technology, and also has issued recent reports that align directly with what RCI and ITRS have been promoting, such as:

*Rebooting the IT Revolution: A Call to Action*

https://www.src.org/newsroom/press-release/2016/758/

*Energy-Efficient Computing*

https://www.nsf.gov/news/news_summ.jsp?cntn_id=136662

SRC also coordinates research programs in nanoelectronics and nanomanufacturing.

## 6. Extending the Roadmap: IRDS

In May 2016, IEEE announced the launch of the *International Roadmap for Devices and Systems* (IRDS), a new organization under the Industry Connections program of the *IEEE Standards Association* (IEEE-SA). The IRDS is sponsored by the *Rebooting Computing Initiative* in consultation with the *IEEE Computer Society*, and represents the next phase of work that began with the partnership between the IEEE RC Initiative and ITRS 2.0 initiated in early 2015. With the launch of the IRDS program, IEEE is taking the lead in building a comprehensive, end-to-end view of the computing ecosystem, including devices, components, systems, architecture, and software.. Plans are for IRDS to deliver a 15-year vision that encompasses systems and devices, setting a new direction for the future of the semiconductor, communications, networking, and computer industries. Bringing the IRDS under the IEEE umbrella will create a new 'Moore's law' of computer performance, and accelerate bringing to market new computing technologies.

## 7. Conclusions

We believe that a new direction for the semiconductor and computer industries is required, and must focus on solving two, inter-related problems:

1. Virtually all computers known today are designed in accordance with the architecture unveiled by Von Neumann in 1945. A New, efficient and yet less power-hungry Computer Architectures need to be invented.
2. A new less power-hungry "switch," operating differently from a MOS transistor, needs to be demonstrated.

Solving (2) does not reduce the need for (1). New switches will have properties that enable new, non-von-Neumann ways of computing. Any solution of these challenging problems requires the contributions of the global computing and semiconductor communities, but most of all, it is absolutely essential that New Architectures and New Devices are synergistically developed.

In order to achieve these ambitious goals in the next decade and beyond, coordination and catalysis among the major players (in industry, academia, and government) is critical. The IEEE, through the Rebooting Computing Initiative and the International Roadmap for Devices and Systems, will a key part of this effort.

## Bibliography

[1] "Special Report: 50 Years of Moore's Law", *IEEE Spectrum*, April 2015. http://spectrum.ieee.org/static/special-report-50-years-of-moores-law

[2] D. Nikonov and I. Young, "Benchmarking of Beyond-CMOS Exploratory Devices for Logic Integrated Circuits", *IEEE Journal on Exploratory Computational Devices and Circuits*, 2015. DOI: 10.1109/JXCDC.2015.2418033.

[3] *IEEE Rebooting Computing Summits*, http://rebootingcomputing.ieee.org/rc-summits .

[4] *Neuro-Inspired Computational Elements Workshop*, Feb. 2015, http://rebootingcomputing.ieee.org/images/files/pdf/NICE2015_Workshop_Report.pdf

[5] S. Venkataramani, et al., "Approximate Computing and the Quest for Computing Efficiency", *Proc. Design Automation Conference*, 2015. doi 10.1145/2744769.2751163

[6] *Berkeley Symposium on Energy Efficient Electronic Systems*, October 2015, https://www.e3s-center.org/events/e3s-symposium/e3s-sym2015-program.htm

[7] Y.H. Lu, et al., "Rebooting Computing and Low-Power Image Recognition Challenge", *Proc. Int. Conf. on Computer-Aided Design*, ICCAD'15, pp. 927-932, 2015. DOI: 10.1109/ICCAD.2015.7372672

[8] T. Conte, E. Track, and E. DeBenedictis, editors, "Rebooting Computing", Special Topical Issue of *IEEE Computer Magazine*, December 2015. DOI: 10.1109/MC.2015.363

[9] J. Shalf and R. Leland, "Computing Beyond Moore's Law", IEEE Computer Magazine, December 2015. DOI: 10.1109/MC.2015.374

[10] D. Holmes, et al., "Superconducting Computing in Large-Scale Hybrid Systems," IEEE Computer Magazine, December 2015. DOI: 10.1109/MC.2015.375

[11] K. Bresniker, et al., "Adapting to Thrive in a New Economy of Memory Abundance,"*IEEE Computer Magazine*, December 2015. DOI: 10.1109/MC.2015.368.

[12] M. Sabry Ali, et al., "Energy-efficient Abundant Data Computing: The N3XT 100x", *IEEE Computer Magazine*, December 2015. DOI: 10.1109/MC.2015.376

[13] S. Murugesan, "Radical Next-Gen Computing", *Computing Now*, June 2015. https://www.computer.org/web/computingnow/archive/radical-next-gen-computing-june-2015

## Biographies

**Thomas M. Conte** is a professor with joint appointments in the Schools of Computer Science and Electrical and Computer Engineering at Georgia Tech. His research interests include computer architecture and compiler code generation. Conte received a PhD in electrical engineering from the University of Illinois at Urbana–Champaign. He is the 2015 IEEE Computer Society president, co-chair of the IEEE Rebooting Computing Initiative, and a Fellow of IEEE.

Email conte@gatech.edu

**Paolo A. Gargini** is Chairman of the International Roadmap for Devices and Systems (IRDS), and was previously Chairman of International Technology Roadmap for Semiconductors (ITRS) since it started in 1998. He was also with Intel from 1978 until his retirement as Intel Fellow in 2012. He was born in Florence, Italy and received a doctorate in Electrical Engineering in 1970 and a doctorate in Physics in 1975 from the Università di Bologna, Italy. He is a Fellow of the IEEE.

Email paologargini1@gmail.com

**Erik Debenedictis** is a technical staff member in the Non-Conventional Computing Technologies Department at Sandia National Laboratories. His research interests include computing approaches across the entire technology stack, including further scaling of von Neumann architectures, brain-inspired computing approaches, and superconducting electronics. He received a PhD in computer science from Caltech. He is a key member of the IEEE Rebooting Computing Initiative, the IEEE Computer Society, ACM, and the American Physical Society.

Email epdeben@sandia.gov

**Alan M. Kadin** is a technical consultant in Princeton Junction, New Jersey, USA, and formerly tenured Professor of Electrical Engineering at the University of Rochester where he authored the textbook “Introduction to Superconducting Circuits.” His research interests include superconducting electronics, future computing and communication systems, and energy conversion. Dr. Kadin received a PhD in physics from Harvard University. He is a Senior Member of IEEE and an active participant in the IEEE Rebooting Computing Initiative.

Email amkadin@verizon.net

**Elie K. Track** is CEO of nVizix LLC, a startup developing novel photovoltaic technology for solar power based in Stamford, Connecticut. His research interests include innovative high-efficiency solar cells as well as superconducting electronics and its applications in high-performance communications and computing. He received a PhD in physics from Yale University. Track is co-chair of the IEEE Rebooting Computing initiative, past president of the IEEE Council on Superconductivity, and a Fellow of IEEE.

Email elie.track@nvizix.com