PRELIMINARY IEEE RC-ITRS REPORT
SEPTEMBER 2015

On The Foundation Of The New Computing Industry Beyond 2020

Thomas M. Conte and Paolo A. Gargini
IEEE Fellows

EXECUTIVE SUMMARY

IEEE rebooting COMPUTING

ITRS

The Organizations

IEEE RC.

The IEEE Rebooting Computing Initiative leverages IEEEs multi-disciplinary, pre-competitive community to explore ways to restore computer performance to its historic exponential growth path. IEEE RC 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. It sponsors RC Summits and co-sponsors workshops and conferences on related topics. For more information, see http://rebootingcomputing.ieee.org.

ITRS 2.0

The International Technology Roadmap for Semiconductors is sponsored by the five leading chip manufacturing regions in the world: Europe, Japan, Korea, Taiwan, and the United States.

The objective of the ITRS is to ensure cost-effective advancements in the performance of the integrated circuit and the advanced products and applications that employ such devices, thereby continuing the health and success of this industry.

The ITRS initiated a process of reorganized in 2012 to realign itself to electronic industry ecosystem.

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

http://www.itrs2.net/

OVERVIEW

The Computer Industry fueled the information revolution over the last 50 years. 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.

Von Neumann Computer Architecture and the invention of the Integrated Circuit represent the building blocks of this revolution. 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 Von Neumann architecture.

IEEE Rebooting Computing Initiative was launched in 2012 to revolutionize the way we approach computing in order to return and exceed the rate of historical progress in computational performance by leveraging the IEEE's multi-disciplinary, pre-competitive community

ITRS 2.0 concept was launched in 2012 to readjust the way integrated circuits are developed based on the new electronic industry ecosystem. The semiconductor industry grew with Geometrical Scaling as the main method to improve transistor cost, performance and number from the mid-70s to the end of the 90s.

Equivalent Scaling concept was introduced in the late 90s and it has supported the growth of the semiconductor industry since the past decay. 3D Power Scaling will become the dominant method to continue and exceed historical trends in the next decade.

IEEE RC and ITRS 2.0 organizations initiated an exchange of information in 2014 and initiated cooperation and joint workshops in 2015.

The two organizations believe that the development of a new computer paradigm needs the 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

INTRODUCTION

The Computer Architecture proposed by Von Neumann and the demonstration of the transistor that occurred in the late 1940s laid the foundation of the modern computer industry. The enabling integrated circuit was commercialized in the early 1960s. Progress in both computer microarchitecture and semiconductor technology allowed enhancing the performance of the original computer architecture beyond any expectations. Any 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 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 were reached by the middle of the previous decade when MPU tried to operate beyond the 130W power level. This physical limit 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.

The micorarchitectural tricks used to hide speculation and parallelism from the programmer began to break down. In response, industry created multicores that required substantial rewriting of software in order to scale computer performance. But, engineering of software for the Von Neumann architecture was itself a difficult endeavor. Adding the need for explicit programming of parallelism was largely untenable. Computer performance stalled.

Both the IEEE and the Semiconductor Community have been looking for new revolutionary solutions for sometime to solve this fundamental problem. New Architectures and New Devices need to harmoniously work together to assure success. Last year experts from the two fields of research decided to join forces to identify New Computer Architectures and New Revolutionary Switches. The first objective of the scientific alliance between **IEEE Rebooting Computing (RC)** Initiative and the **International Technology Roadmap for Semiconductors 2.0 (ITRS 2.0)** was to identifying the challenges posed by the above computational problem, and to establish a ***roadmap to successfully restart computer performance scaling.***

All of these efforts would be meaningless without adequate support from research and funding organizations. The cooperation between Government, Industry and Academia has repeatedly been proven successful in solving the most complex problems. The device roadmap and challenges outlined by ITRS in 1998 were shortly after supported by a broad set of investments made by the National Nanotechnology Initiative (NNI) in the year 2000, and this combination led to a complete and successful revolution in the way transistors are built. Similar efforts were also launched in many other regions of the world following the NNI announcement. There would have been no advancement in the semiconductor industry and consequently in the computer industry without the success of this global effort.

The recent announcement of NSCI effort adds all the necessary elements for another successful association between Government, Industry and Academia.

I. 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 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 strait forward manner. 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 and by so doing in 10 years there will be as many as 65,000 transistors available to design a product, he stated; in essence he formulated this implied question: ***"How could a system designer take advantage of this abundance of transistors?"***

Between 1972 and 1974, Robert H. Dennard announced to the world a new methodology that allowed to quickly reducing the size of a transistor and also predicting 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 lead 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.

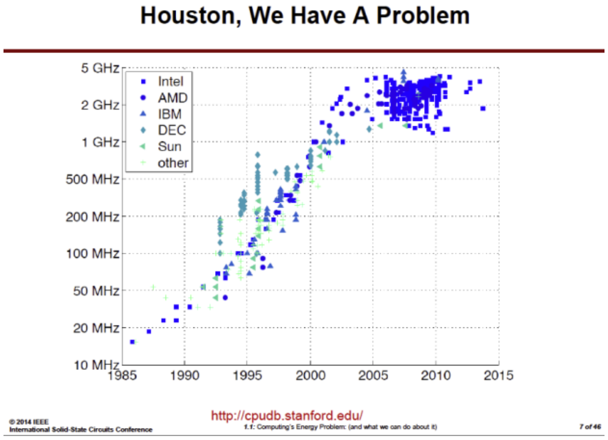
II. THE PC REVOLUTION

The PC market was born in the mid-80s and soon locked the semiconductor industry and consequently the computer industry in a very unusual situation since the system manufactures were all using microprocessors and software mostly produced by only two companies: Intel and Microsoft (resp.). In a market were silicon technology and architecture as well as software architecture were defined and locked in 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 every two years, the industry went back to the 1960s and re-implemented complex microarchitectures form Project Stretch, the IBM 360 model 91, the CDC 6600, etc. Combined, these "tricks" worked behind the scenes to figure out how to run instructions 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 ways 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^2*f*. 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 more instructions needed to be "in flight" than was possible by waiting for branch instructions to execute. This lead to *speculative execution*: predicting what path a program would take and then doing that work ahead of time, in parallel. Thus higher frequencies meant deeper pipelines, which in turn required more and more speculatively executing instruction. But no prediction is 100% accurate. Invariably, these microprocessors did a lot of extra, wasted work by miss-speculation. The deeper the pipeline, the more power was wasted on these phantom instructions.

But performance was the name of the game and operating frequency kept on increasing through the 90s until it eventually happened, the processor way 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 [Fig. 3] was ***No Longer Viable***!


Houston, We Have A Problem

http://cpudb.stanford.edu/

© 2014 IEEE
International Solid-State Circuits Conference

1.1: Computing's Energy Problem: (and what we can do about it)

7 of 46

Fig. 3. Actual operating frequency limitations

The consequence of reaching the power wall had actually much further reaching implications that just affecting the PC industry. The PC and microprocessor ecosystem had driven the cost of mainstream processors to a very attractive economical (low cost) level 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!

III. AN EARLY CALL TO ACTION

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 the year 1999 and 2003 the semiconductor industry converted to copper interconnects and low-k intermetal dielectrics.

Paolo A. Gargini (Director of technology strategy and Intel Fellow) realizing that the problem of re-engineering the MOS transistor (process and structure) and interconnect lines was not only impacting the US semiconductor industry but the global semiconductor industry as well was able to promote and launch the International Technology Roadmap for Semiconductors (ITRS) in association with organizations from Europe, Japan, Korea and Taiwan in 1998. 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 since the beginning of the past decade and Equivalent Scaling will still remain the undisputed guidance 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

The US government launched the National Nanotechnology Initiative (NNI) under the guidance of Mike Roco (NSF) in the year 2000. The NNI announcement triggered an escalation of investments in Nanotechnology across the world [Fig. 4].

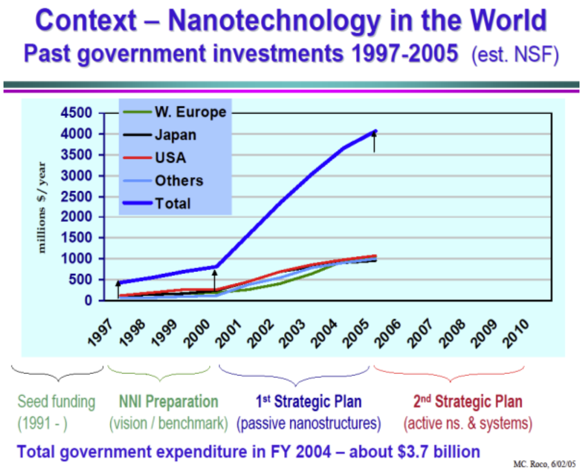


Fig. 4. Nanotechnology investments

Both Governments and IC industry were, as many times in the past, on a cooperative course and synchronization of efforts was a must. Paolo A. Gargini launched the Nanoelectronics Research Initiative (NRI) in 2005 with the cooperation of leading US semiconductor companies, NSF and NIST.

The goal of NRI consists in identifying and developing new types of transistors operating under new physical phenomena. These devices present some features that are different from CMOS. In the past 5 years several interesting candidates have emerged and are intensively developed but it should ***not be expecte***d 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 maintaining historical trends into most of the next decade but fortunately a new approach is within reach!

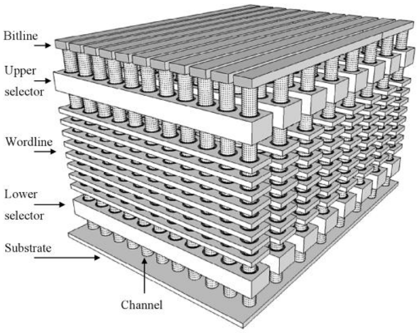
IV. 3D POWER SCALING

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 foreseeably 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 degree; 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 there is much more. Nothing prevents from stacking multiple planes of verticals transistors on top of each other. In fact, columns of transistors can be grown by multiple sequential depositions and then connections the transistors in plane and from plane to plane can be made [Fig. 5].

Memory IC makers have already announced Flash memory 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. 6]. ***The number of transistors will continue to grow and even faster than Moore's Law!***



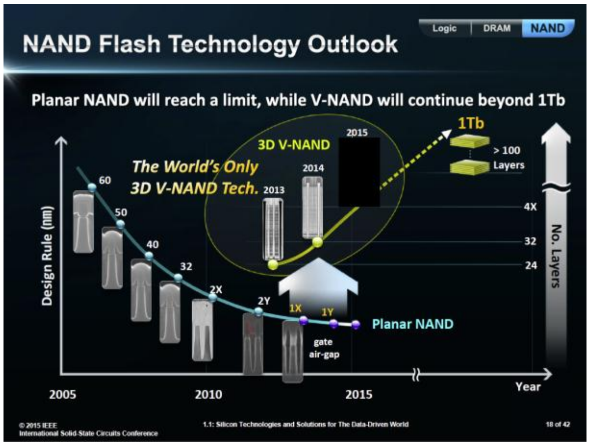


Fig. 5. Packing planes of transistors **Fig. 6. 1Tbit Flash memory forecasted**

On the power reduction side, tunnel transistors have shown the capability of running similar to an MOS transistor 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 present also 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 non-volatile mode and using much less power have been demonstrated also [Fig. 7]. This type of behavior is beginning to be rather similar to the way neurons and synapses operate in a human brain, enabling the Neuromorphic approach identified by the IEEE RC.

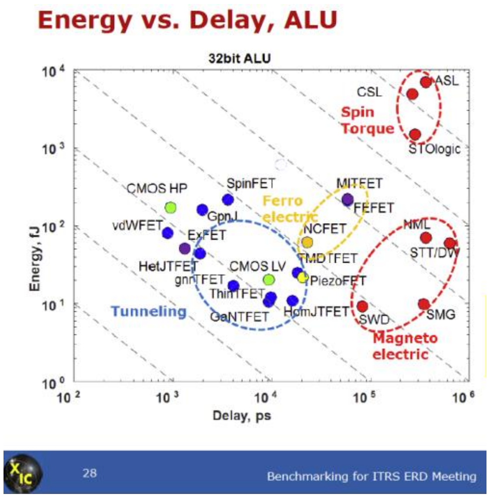


Fig. 7. Energy-Delay transistor families

V. WHERE ARE WE GOING NEXT?

It should be plainly clear from the previous paragraphs that computing performance increased from the mid 80s until the beginning of the previous decade as a result of higher frequency of operation and superscalar's clever ways of keeping the processing unit busy all the time. It should be noticed also that a continually higher operating frequency was responsible for the majority of the improvement in computing performance.

In the beginning of the previous decade the MOS technology reached a fundamental power wall [Fig. 8] that prevented microprocessors' designers from further increasing operational frequency.

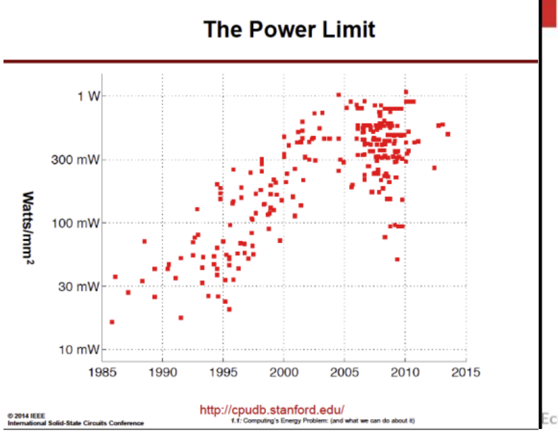


Fig. 8. Power limit wall

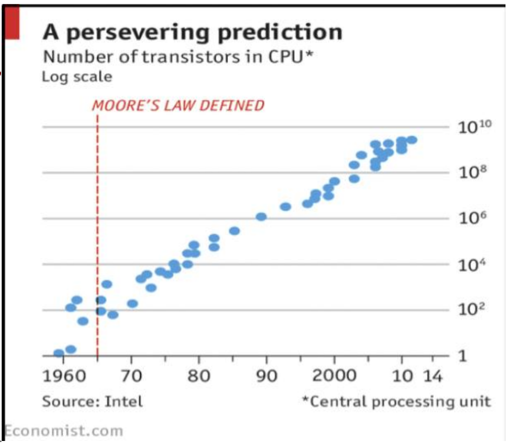


Fig. 9. Transistors growth trend

Leading IC makers have continued to reduce the size of transistors and increase their number [Fig. 9] according to Moore's Law for the past 10 years just like they had done since 1975, 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 evolution, sometimes there occurs a catastrophic event that brings about a new order. This is the theory of *Punctuated Equilibrium* as discovered 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 on 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 related 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 increase with consequent reduction in signal propagation delay.

VI. REBOOTING COMPUTING

In 2011, Elie Track, then-president of the IEEE Superconductivity Council, and Tom Conte, professor of CS and ECE at Georgia Institute of Technology and himself a computer microarchitecture researcher, independently realized 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 was the catalyst to enable this change. Conte coined the term "Rebooting Computing[1]," and the IEEE Rebooting Computing initiative was born. With funding from the Technical Activities Board of the IEEE (the body inside the IEEE that encompasses all technical societies and councils across the discipline), IEEE RC began holding invitation-only summits to begin brainstorming a way forward.

The first IEEE RC summit was held in Washington, DC 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: (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.

Over the course of 2014, IEEE RC held two additional summits, both in the Silicon Valley area. The second summit looked at "new engines of computing." Old and new ideas in how to compute were welcome to the table. These approaches are summarized 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

[1] It was later learned that Peter Denning had previously used the term for a prior effort. IEEE requested and received his permission to use the name.

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

Fig. 10. Example approaches to Rebooting Computing and their research challenges.

The third summit 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 driver 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 RC [Fig. 10] take advantage of properties of semiconductor devices heretofore though 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.

VII. A NEW DIRECTION FOR THE SEMICONDUCTOR & COMPUTER INDUSTRIES

We believe a new direction for the semiconductor and computer industries 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): instead, the two synergize. 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 important

that New Architectures and New Devices are synergistically developed. This consideration led to the cooperative effort between IEEE Rebooting Computing and the ITRS 2.0 towards producing a joint roadmap.

Once promising solutions are identified, it is up to any region to establish international and domestic programs leading to societal benefits and economic progress.

We believe that National Strategic Computing Initiative represents the first major step towards supporting the path of a successful solution to the computational challenge!