May 16, 2011

Report on the Third Kavli Futures Symposium

Growing High Performance Computing in a Green Environment

September 9-11, 2010, Tromø, Norway

Roger Blandford, Michael Roukes, Larry Abbott, Terry Sejnowski (Chair)

Participants:

Larry Abbott (Columbia)
Tom Abel (Stanford)
Blaise Aguera y Arcas (Microsoft)
Andreas Andreou (Johns Hopkins)
Roger Blandford (Stanford)
Kwabena Boahen (Stanford)
David Dean (DOE)
John Grunsfeld (NASA)
Tod Hylton (DARPA)
Steve Koonin (DOE)
Michael Roukes (Caltech)
Felix Schuermann (EPFL)
Terry Sejnowski (Salk)
Horst Simon (Berkeley)
Risa Wechsler (Stanford)

Observers

Miyoung Chun (Kavli)
Tom Everhart (Kavli)
James M. Gentile (Research Corporation for Science Advancement)
Ziba Mahdavi (Stanford)
Jim Omura (Moore Foundation)
Tryvge Solstad (Salk) - Rapporteur

Executive Summary

The Third Kavli Futures Symposium held in Tromsø, Norway on September 9-11, 2010, took a long-term look at what a green future for computing might hold, based on current computing architectures and new ones designed with energy limits in mind.

Although the cost of computing is decreasing, the current large-scale parallel architectures that are being planned to achieve exascale performance levels are projected to require vastly more power[1]. This path will lead to problems, as the power needed to follow Moore's law will soon become environmentally unacceptable. If we do not begin planning now, high performance computing may find itself at the end of a *cul de sac* based on unsustainable technologies in a world with finite energy constraints.

Astronomical data and astrophysical simulations have reached "astronomical" proportions. Processing power is no longer rate-limiting and for all practical purposes, "flops are free." Rather, processing power is outstripping the vast amount of fast memory needed to analyze astronomical data and the rate limiting bottleneck is the communication between the memory and the processors. An even more severe problem is on the horizon as the power needed to reach exascale computers is estimated to reach 50 megawatts, more than the power needed to run the New York subway.

Nature faced similar problems with the flood of sensory data that is continuously processed online and committed to long-term memory. Brains evolved many highly efficient special-purpose processors in contrast to the general purpose computers that dominate high performance computing; the bottleneck problem was solved by integrating the memory into the processors; finally, nature achieved at least petascale computing with 20 watts of power by exploiting the physical properties of matter at the molecular level.

Progress toward exascale computing may benefit from current efforts in large-scale parallel integration for building a silicon cortex and advances in nanoscience that will shrink memory and logic gates down to the molecular level. However, an exascale machine pushing forward the frontiers of astrophysics, neuroscience and nanoscience in a sustainable energy environment needs to have an architecture that integrates device physics, algorithm development, and data storage.

This report summarizes the discussions that took place in Tromsø that led to these conclusions and to a set of questions that could guide the future of exascale computing in science and engineering.

[1] "Kogge Report" ExaScale Computing Study: Technology Challenges in Achieving Exascale Systems" http://www.cse.nd.edu/Reports/2008/TR-2008-13.pdf

The energy problem

More of the same or no Moore?

Integrated circuits have roughly doubled their performance every 18 months since the first digital computers were available in the 1950s (Moore's law). Extrapolating this trend, we could expect exaflops (a billion billion floating point operations per second) computers in 2015.

However, continued miniaturization of the CMOS (Complementary Metal-Oxide Semiconductor) technology is reaching a physical performance limit at which further increase in speed leads to an exponential increase in energy consumption from leak currents. Fundamental physics (Boltzmann's law) does not allow the physical implementation of an MOS transistor that could behave as a perfect switch. Transistors must always conduct minute amount of current in their "off" state, called the "leak", or subthreshold current. With chips having billion of transistors on them, these minute currents add up to be a significant problem. Optimizing devices for higher speed has the unfortunate consequence of increasing the "leak" currents. Clock frequencies have stagnated since 2005, and today increases in computing power are achieved by increasing the number of processors per chip with the processors operating at slower speeds. High performance computing is dominated by massively parallelism.

The need for exploring alternative architectures is imminent and minimizing energy consumption is central to further progress.

Energy cost

The quest for exascale computing comes at an economic and environmental cost. With today's technology, a 1 petaflops machine requires approximately 3 MW of power, which costs approximately $3 M a year. Historically the number of computations per kWh has increased by a factor of 1.5 every year (Koomey's Law). According to this figure, a projected 1 exaflops computer operating at a more realistic date of 2020 would require 50 MW and would generate a yearly electricity bill of $50 M. In comparison, the NY subway runs on a mere 20 MW.

We have already passed the point at which the cost of running and maintaining a high performance computer over its lifetime surpasses its capital costs. The main bulk of these additional expenses go towards building space and cooling. From 2000 to 2005, the total electricity consumption associated with servers doubled, reaching 45 billion kWh/year in the US and 120 billion kWh/year worldwide. The corresponding electricity costs were $3.6B/year and $9.6B/year, respectively (at $0.08/kWh)[2]. Realizing the importance of energy cost, large data

[2] Koomey, Analytical Press, 2007: Estimating total power consumption by servers in the U.S. and the World; 2005: U.S. Energy Information Administration (DOE): http://www.eia.doe.gov/cneaf/electricity/epa/epat7p4.html

centers like those of Google and Microsoft are built in locations with easy access to cheap energy and water cooling. Microsoft's data center in Quincy, Washington is presently one of the world's largest, taking up 470,000 Sq Ft and running on 47 MW, and a new data center is now being built next to the existing one. The total supercomputer capacity available to US academia in 2010 was 3.5 Pflops, requiring 35 MW of power. Obtaining corresponding figures for commercial users in the USA and both academic and commercial users in the rest of the world would be highly valuable.

Environmental impact

The enormous power consumption of high performance computing not only constitutes several percent of the US budget, but also contributes to a large IT carbon footprint. The collective 2007 IT carbon footprint was comparable to that of the global aviation industry (830 Million tons CO_2) and is expected to increase to 1.4 Billion tons CO_2 by 2020, amounting to 4% of the total carbon emissions. Computing centers alone are responsible for 150 M tons of CO_2 per year (corresponding to 30 million cars).

In conclusion, the projected trend of increased computational performance is not sustainable from the point of view of energy consumption, energy cost, and environmental impact. To meet the scientific demand for exascale computing, developing new energy efficient or 'green' technologies is essential and time is of the essence.

The data deluge problem

The problems facing scientific computing in the near future have been apparent for some time in astrophysics[3], which is struggling with understanding the origin and evolution of the universe and the nature of its constituent galaxies, stars and planets. The greatest computational challenges come from analyzing large data streams and performing giant computer simulations.

Towards the end of the decade, the 8 m optical telescope Large Synoptic Survey Telescope (LSST) will begin its mission to survey over half the sky by taking 3 gigapixel images very 15 s. It will produce over 100 petabytes of archival data over 10 years. Increased computational capability will be necessary to allow professional astronomers and members of the public to mine this archive. Even greater challenges are presented by the proposed international Square Kilometer Array radio observatory. The main computational need is not so much faster processing speed, but much greater data storage and bandwidth. Data transfer, error correction, and fault tolerance are topics of intense interest to the astrophysics community. Experience from

[3] Szalay and Gray, ("Science in an exponential World" Nature vol. 440, March 23 2006; Science special issue February 14, 2011

astrophysics also reminds us that software development is often as important to increased performance as hardware development

The second computing grand challenge for astrophysics is to scale up simulations. Since the last Futures symposium, Tom Abel reports an increasing focus on programming emphasizing data localization. Modern frameworks for solving differential and integral equations combine message passing, thread parallelism and frequently utilize graphics processing units. For example, parallelizing the simulations of three-dimensional cosmological radiation hydrodynamics resulted in a 100x simulation speedup and scientific discoveries that could not have been made with lower spatial and temporal resolution. Similarly, the development of the adaptive mesh refinement technique has been instrumental to simulating cosmological data with a dynamic range spanning 10^{15} in length. (This is comparable to resolving scales from the earth to a single bacterium.) Finally, because computers are already extremely fast, computational efficiency is often limited by the human-machine interfaces. Functional ways of visualizing data can save enormous amounts of computation time.

These examples draw attention to the impact of software development and improving memory and bandwidth rather than processor speed in maximizing computational performance.

Hence the mantra: “Flops are free! free the flops!”

Can we learn from Nature how to solve the energy problem?

Only recently has energy consumption become the limiting factor for progress in high performance computing. In contrast, evolution has likely favored energy efficiency from the beginning, with the result that two grams of spaghetti is enough to power the brain for an hour. From this perspective, the brain might be a good model for low energy computing. Although no ‘standard model’ of brain function has yet been formulated, some basic design principles of the brain can be identified:

Specialization of function: the brain is highly compartmentalized into brain areas serving different functions and performing different computations. Energy is supplied to different areas on demand, a principle exploited to perform brain imaging, such as functional magnetic resonance imaging (fMRI). In some sense, the brain is a multipurpose computer composed of hundreds of interacting, special purpose machines.

Parallel processing: A typical neuron does not gain computational power with an impressive clock speed, maxing out at a few hundred spikes per second at best. Rather, the brain's computational power derives for its ability to compute with billions of neurons in parallel, reaching for over 10^{15} synaptic operations per second[4]. The balance between parallelization and clock speed probably reflects a tradeoff between energy consumption and performance, a trade that is reflected in microcircuit design with the advent of multicore processor architectures. Finding the optimal balance for particular architectures is an open challenge to future microchip design.

Local memory: A typical cortical neuron is connected to 10,000 other neurons through chemical synapses that form a 3-dimensional connectivity matrix. This synaptic matrix holds our long-term memory and all directly accessible memory is stored locally at each processing unit, with the redundancy required for fault tolerance and graceful degradation. With enormous amounts of data, component failure is a critical impediment for today's computer technology. Assuming a minimum of 1 byte of storage per synapse gives a tentative lower bound on the storage capacity of a human brain of approximately 1 petabyte. It is likely that there are other ways that information can be stored in the complex dynamical behavior of local neuronal micro-circuits, which could yield equivalent "storage" capacities that are orders of magnitude higher than what we estimate here, so this estimate should only be considered a provisional lower bound on the memory capacity of brains.

Analog data processing with digital data transfer: Neurons are principally analog processors but transmit information in the form of digital spikes along active cables. Digital signal transmission is highly energy efficient and amenable for operations like error correction. The absolute energy consumption of a spike is in the range of a picojoule, about a tenth of the power need to switch on a transistor. Most neural systems smaller than a millimeter do not use spikes but use graded signals and graded synaptic transmission (including direct connectivity through gap junctions). Larry Abbott presented evidence that due to the noisy nature of spikes, the spiking network equivalent of a neural system with analog communication using the average firing rate might require 1000 times the number of neurons. This emphasizes the role of the spike in performing energy efficient transmission of information over "long-distance" wires. Alternatively, and quite likely, nature may have taken advantage of spike timing in addition to the total number of spikes, and especially the relative timing of spikes in a neural population. Several projects in neuromorphic engineering are exploring the benefits of the spiking network architecture.

The Brain as an Inference Engine: Understanding the efficiency and effectiveness of the brain is clearly one of the greatest scientific challenges of all time. One of the major functions of the brain is inference from probabilistic data. Graphical probabilistic models that have been studied

[4] This is estimated from the average firing rate of 1 Hz for 10^{11} neurons each of which has 10^4 synapses. The computation performed by one synaptic operation is a few flops in a few ms initially, but could involve millions of flops if long-term biochemical changes are taken into account that extend over hours and are involved in memory storage. The nature of these changes at synapses is at the forefront of current research.

in machine learning often compute likelihood functions for maximum likelihood estimation. There is growing evidence that neurons integrate sensory information as log-likelihood sums and make decisions by setting thresholds.

Mapping the brain: Until recently the amount of neurophysiological data has been limited by the technology used to record neural responses. We are now in the midst of a revolution in computer-controlled optical imaging technology with the potential of monitoring the activity of hundreds of thousands of neurons simultaneously. Nanotechnology is also starting to have an impact on neurophysiology and the first nanoscale electrodes are being developed for intracellular recordings. Advances in nano-photonics and optogenetics are also providing a new way of mapping and controlling activity brain. Initially inspired by the success of the human genome project, another advance since the last Futures Symposium is the exploding interest in connectomics: mapping the full connectivity of the brain down to the level of individual synapses.

The simple roundworm C. elegans with its 302 neurons was the first "connectome" to be mapped, and the approximately 100,000 neurons of the fly brain are being mapped at Janelia Farm of the Howard Hughes Medical Institute and elsewhere. Mapping the full connectivity of the cerebral cortex is a grand challenge, but the task has already been undertaken by the Blue Brain project at the École Polytechniques Fédérale de Lausanne led by Henry Markram and Felix Schuermann.

Eventually, the question will be to find connectivity on the molecular level. Reconstructing a cube with 6x6x5 μm^3 volume of rat neuropil in the hippocampus to the level of molecules has taken 3 man-years of work and generated a terabyte of data in the Sejnowski lab. Reconstruction of an entire brain on this level of detail would generate several petabytes of data for a fly brain and 1 exabyte of data for a mouse brain.

The latest results from the Blue Brain Project indicates that to build a full cellular level neural network model of the human brain, 100 petabytes of memory and more than 1 exaflops are needed. Such a machine would consume more than 50 MW, a million times more than the system it is designed to simulate, the human brain operating at 20 W. How can we close this gap?

Nanoscience and computer architecture

The great hope is for nanoscience to come to the rescue. Nanotechnology has the potential of producing smaller components, more chips per die, denser, more fault tolerant memory, and 3-dimensional interconnections between cores. Nanoscience is expected to become a massive user of high performance computation itself, through the emergence of fields like computational material science, structural biology, and personalized medicine, allowing, for example, simulation of human interaction with microbes and designing drugs virtually. However, nanoscience is in its early days, and currently is at the level of exploring new materials and architectures. How can we best meet our computational needs both in the near and the long term?

Four main paths towards energy efficient exascale computing were discussed at the symposium:

I. Developing fundamentally new computer technology

II. Optimizing current microcircuit architecture

III. Building special purpose computers

IV. Decentralized computation

Developing fundamentally new computer technology

An important issue raised at the meeting was that the current CMOS technology was not designed with energy efficiency in mind and is reaching physical limits for miniaturization. Eventually, substantial increase in computational performance will depend on the development of fundamentally new technology for computation. What alternatives are there and what are the resources necessary to develop them?

By spelling 'IBM' with atoms in 1989, Kavli prize winner Donald M. Eigler gave an inspiring example of what can be achieved with nanoscience. In 2002, Thorsen et al presented a chip containing a high-density array of 1000 individually addressable chambers on the picoliter scale. In 2010 they are up to 40,000 valves per chip. Continuing this trend, we will have to wait another 40 years to reach the level of complexity that modern CMOS computers have already achieved, and even if nano solutions were ready today it would still take a decade to deploy.

Over the past 40 years, massive resources and major investments have gone into increasing the number of transistors on a chip from thousands to billions[5]. Similar resources will be needed if want to develop fundamentally new applications of nanocomputer technology with high-societal impact such as personalized medicine, rapid vaccine development, and national security.

Many different alternative logic gates based on atoms, spins, photonics, etc, are being explored today. However, the realization that flops are free invalidates the justification for alternative (non-silicon CMOS) logic gates. Rather than bringing about revolutionary technology, the near-term impact of nanoscience is likely to be the improvement of memory density and co-locality.

[5] "The semiconductor industry invested over $18 billion in research and development in 2006, the last year for which we have data, or over 15 percent of sales in that year .This constitutes about one out of every 12 dollars spent on R&D in the U.S.by private industry, and one out of every 8 dollars spent by U.S.manufacturing." E M. Ehrlich, Manufacturing, Competitiveness and Technological Leadership in the Semiconductor Industry.

Nanoscience can “free the flops” by producing dense, local atomic scale memory with a high degree of fault tolerance and less need for cooling. The symposium members were dazzled by the promise of someday achieving what nature has already achieved: computing with single molecules.

However, nanoscience is still in its infancy and energy efficient solutions to problems are not yet in hand. Even though Eigler’s work is truly remarkable in that it demonstrates computations at the atomic level, the energy per op (when normalized to the temperature) is comparable and even higher than the energy per op in CMOS. There is still a fundamental gap in our understanding on how to build efficient computational systems down to the nano/atomic scale.

Optimizing current microcircuit architecture

Looking at the latest development in microchips, the trend is towards more brain-like low power design. IBM’s Power5 processor was introduced in 2004 with two cores running at 1,900 MHz and 120 W. Lower clock frequencies enable the use of simpler cores that are easier to program, creating fewer bugs, more robust, and more concentrated allowing more cores to be deployed per processor. Processors developed especially for low power applications in handheld computers and cell phones have a flops-to-Watt ratio that is larger by factors 80 and 400 respectively than the IBM Power5 microprocessor [6] The ARM processor was also highlighted at the meeting for trending towards biological design principles in having a heterogeneous collection specialized analog and digital circuits solving separate tasks and shutting off components not in use at any given time.

Utilizing the high energy efficiency means dealing with highly parallel architectures. A thousand cores per chip is expected by 2018 Tilera has already announced that it can place 512 cores on a single chip, and NVIDIA has in production 1000 cores/chip, with 10,000 cores/chip on the horizon. It is essential that software tools and algorithms be able to keep up with this rapid parallelization in order to capitalize on the full potential of parallel computing that high performance computers now offer and depend on.

A further consequence of massive parallelization is that the cost of sensing, collecting, generating and calculating with data is declining much faster than the cost of accessing, managing and storing it[7]. In this sense, flops are free, but memory is dear and constitutes the main cost of computing.

[6] For example, Intel’s “Atom” microprocessor has a clock speed of 800MHz and consumes 0.6W while Tensilica’s “Xtensa” has a clock speed of 600 MHz and consumes 0.08W.

[7] For example, in the six years from 2002 to 2008, the processing cost declined by a factor 100 from $10 per megaflops to 10c per megaflops, while the cost of storage remained constant at 10c per megabyte.

Because of the limited area of a microchip, one challenge that comes with the massive parallelism is how to divide the available area between memory and processors in the most efficient way. Accepting this challenge from the previous Kavli Futures Symposium, Andreas Andreou presented a new framework for calculating the optimal on chip memory, processor and communication resources areas. This represents an important step towards being able to optimize the memory, processing and communications in microchip architecture for specific applications.

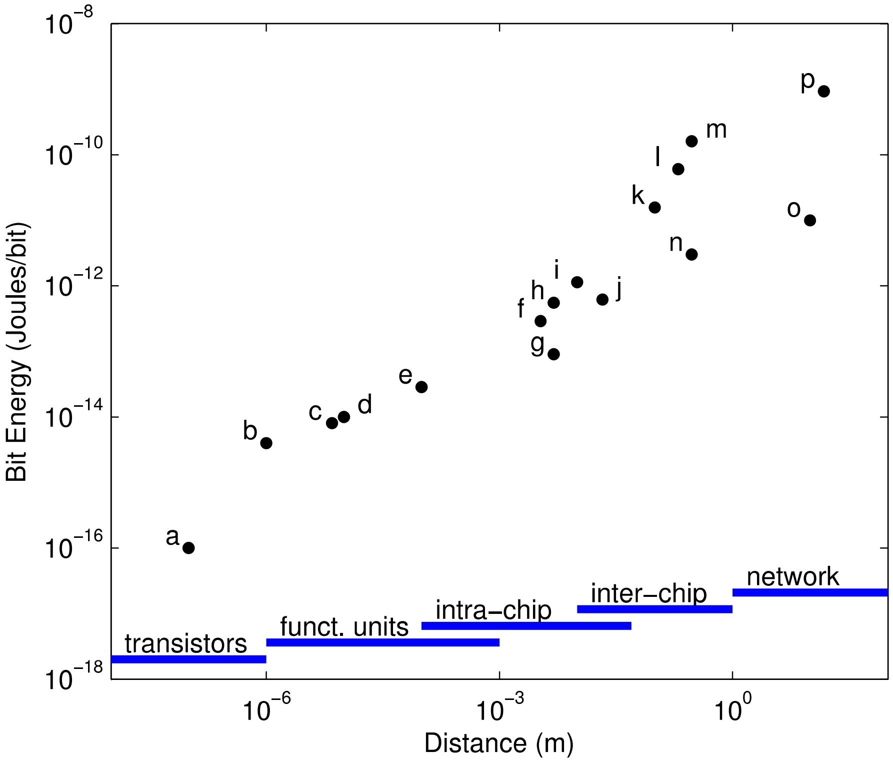

Figure 1. a. 100 nm CMOS inverter, **b**. 1000 nm CMOS inverter, **c.** SOI 3D via, **d.** 3D Through Silicon Via, **e.** MPU register file, **f.** L2 cache **g.** IC copper trace (future), **h.** IC copper trace (current), **i.** electrical switching across 1cm die, **j.** L3 cache **k.** electrical chip-to-chip link, **l.** DDR DRAM, **m.** SERDES (current), **n.** SERDES (future), **o.** optical, **p.** Firewire (IEEE 1394b)

Since energy to move data across a wire is proportional to the length of the wire, keeping memory and data movement local to the processor is key to minimizing energy cost. Moving data between chips is approximately 100 times more expensive than moving data within a chip. To maximize the amount of data that is locally accessible to each core, microchips with 3-dimensional interconnections are being built by Tezzaron and tested in Andreou's laboratory, who reported improvements in energy efficiency by two to three orders of magnitude.

Building special purpose computers

We have already seen that the von Neumann architecture of conventional digital computers that separates memory and processing is poorly matched to that of neural circuits and inefficient for performing neural network simulations. The human brain runs on the power of a 20W light bulb; in comparison, IBM's Blue Gene computer needs approximately 100 KW of power to simulate a 100,000 neuron cortical column in the Blue Brain project.

In answer to this shortcoming, neuromorphic chips, electronic chips with analog circuits that mimic the architecture of biological neural systems pioneered by Carver Mead in the 1980s, explore hardware implementations of computing with spikes. Kwabena Boahen presented a neuromorphic chip that implements 1 million Hodgkin-Huxley based neurons and 6 billion synapses in hardware and runs on 1 Watt of power. The potential of such brain-inspired technology for other purposes than simulating the brain is now being explored by DARPA's SyNAPSE project, led by Tod Hylton, aiming to control intelligent robots with energy efficient neuromorphic devices. Increasing the memory density to the levels found in biological systems (1 synapse/μm^3) and the energy consumption to 10 femtojoule per synaptic operation would be a game changer for data storage.

To solve certain classes of problems it might be more efficient to develop specialized computers rather than depend exclusively on general purpose computing. An essential step towards specialized computational solutions is to identify common problem categories across scientific disciplines. For example needs have already been identified for specialized differential equation solvers, Fast Fourier Transform chips and Monte Carlo processors. The Anton supercomputer developed by DE Shaw Research, a specialized machine aimed at molecular dynamics is an interesting example of emerging technology aimed at specialized solutions. Special purpose hardware can be effective in speeding up problems of fixed size that do not scale favorably as the number of general purpose processors increases. This regime is called "strong scaling."

It is still possible that multipurpose computers are required to solve multiphysics problems, and the need to assess hybrid computer architectures was discussed at some length. Hybrid systems could consist of several small and a few big computers, or several special purpose computers bridged by software that brings different computer architectures together in a hierarchical computer system. Cell phones have already adopted this approach, driven by the need to minimize energy consumption. This is the direction that the high performance computing is heading.

Decentralized computation

Referred to as one of the tyrannies of high performance computing, the low-end consumer market is a primary force behind the development of computer components, now driving the development of mobile devices. Slow writing, visual inferiority and low-feature gaming are issues of the past. With powerful, directly addressable GPUs and dual 1.5G Hz battery operated processors as the motto of 2011, there is now more combined compute power in the pockets of ordinary people than in supercomputers available to academia. With unlimited data plans, WiFi and 4G, these devices are anywhere and everywhere connected, and often idling. As we are entering an era of computing with mobile devices, the feasibility of harnessing this resource for scientific purposes should be evaluated.

Blaise Aguera y Arcas gave examples of where distributed computation outperforms big data centers in the mapping or making of a physical representation of the world. Through interactive web applications like Flickr and Bing maps, people take the computation to the source, using mobile devices with cameras and GPS coordinates.

On the metric of performance:

At the 2nd Kavli Futures Symposium discussed major theme was the issue of whether “bandwidth cost integral” should replace flops as the standard metric from computational performance. A clear conclusion from the 3rd Kavli Futures Symposium is that “it all boils down to bits or flops per Watt”. Measuring flops per joule or Joules/bit (bit-energy) might serve as appropriate measures of computational performance.

Conclusions from the 3rd Kavli Futures Symposium:

- The state of the Kavli disciplines was summarized as “Neuro insights, Nano play, and Astro desperation.”
- Developing algorithms and human-computer interfaces can improve computational performance by several orders of magnitude.
- New architectures for integrated circuits should focus more on improving memory and bandwidth rather than processing speed.
- Key principles to energy efficient computation are found in biological neural circuits and can now be implemented in integrated circuits using nanotechnology.
- In the short term, nanotechnology is best put to use in developing denser, more fault tolerant memory that can be stored locally (on-chip) to processors with 3-D connections.
- There is a need for ways to calculate the optimal fraction of memory to processor area for specific purpose chips. A framework for calculating such cost functions is being developed.

- For certain applications, specialized computers are favored over general purpose computers. Neuromorphic chips represent an implementation of specialized hardware for simulating neural networks.

- We need to think in a radically different way about parallelism; we need architectures that are fundamentally parallel and algorithms that make full use of this parallelism.

- A surprising outcome of the meeting was realization the fact that the bulk of computational power now is shifting from high performance computers to mobile devices. Due to limited battery capacity and the relatively high demand for resource intensive applications like video and WiFi these devices also tend to have very energy efficient hardware. This trend might be exploited for scientific purposes.

- Along with the rise of mobile computing, e-waste is becoming a large problem, and ways to design recyclable components are of increasing importance.

Plan of Action

- Write a paper reviewing the current state of computational resources available to academic and commercial agents, energy consumption and energy cost both in USA and the world as a whole.
- Derive a simple equation for relating computational power, energy consumption, price, and carbon emission. This could someday become another law of computing.
- Write a nuts and bolts paper on how far high performance computing can get in 20 years, exploring continuation of CMOS versus a shift to fundamentally new technology (incentive for funding).
- Find the estimated economic cost of building an exascale computer, and how to best divide money between software, network, maintenance, and labor.
- Assess the need for special purpose machines for strong scaling, like a special purpose differential equation solver, as part of hybrid systems of interacting special purpose computers, and software interfaces for bringing together different architectures.
- Stimulate design of energy efficient chips through incentives like the X-Prize.
- Stimulate development of software for parallel supercomputing.
- Investigate the possibility of utilizing the many idling mobile devices around the world, e.g. similar to the seti@home project.
- Understand what would be the impact of changing metric of computation from bits per second to bits per second per Joule?
- How should we prepare for zetascale (10^{21}) computing?

Working groups

Aiding the work towards a plan of action, the following questions were addressed at round table sessions:

1. *X-Prize for exascale computing:*

   Leader: Terry Sejnowski
   David Dean
   Tod Hylton
   Miyoung Chun
   Blaise Aguera y Arcas
   Horst Simon

   Competitions, such as the 2010 Student Cluster Competition for the high performance computing, spur creativity and innovation. The X Prize Foundation has successfully run a number of high-tech competitions, such as the Ansari X Prize for Space Flight. An X Prize for efficient computing could incentivize innovation to develop energy efficient computers. A competition could be held with the following problem: For a given amount of power (e.g. 20W), produce a set amount of compute power or solve a given problem (e.g. chess, or go). Because companies want to make money on products they already have, they are not interested in pushing for revolutionary technology. The X Prize could provide a visible target for startups and innovation.

   How can foundations and agencies help in funding projects?

   How can we get funding for developing algorithms and visualization tools that utilize the parallel nature of high performance computing?

2. *Exascale Computer vs. Exascale Computing: Are there natural breakpoints on the way to exascale computing?*

   Leader: Horst Simon
   David Dean
   Kwabena Boahen

   There are several daunting barriers to effective exascale computing. The evolution of semiconductor technology is dramatically changing the balance of future computer systems. Clock frequencies are expected to stay constant or even decrease to conserve power. One effect of this is that the number of processing units on a single chip will have to increase exponentially from this point onward. In addition, the energy costs of moving data both on-chip and off-chip will become much more important. Based on our discussions, two different machine architectures appear possible by the end of the decade. The first is a chip architecture where all of the processing units are similar low power units, similar to the current IBM Blue Gene architecture. The second design approach is heterogeneous systems where some of the processing units are much different than the others -- similar to IBM's Roadrunner and today's GPU enhanced clusters. Both of these designs pose a common set of challenges:

1. Total concurrency in the applications must rise by a factor of more than 1000;

2. Memory per processor decreases dramatically which makes current weak scaling approaches problematic;

3. Whereas the cost of computation used to be dominant in past systems, the energy cost of moving data over long distances will become more costly than computation, so the locality of data and computation will be increasingly important for future systems. Therefore models for parallelism, which assume all computing elements are equidistant, will perform poorly;

4. Manufacturing variability and practical approaches to handling manufacturing defects (building spare components into designs) make it unreasonable to assume the computer components will offer deterministic performance across large systems. Bulk-synchronous approaches to computation, which dominate current programming models, may need to be rethought and redesigned to handle more asynchronicity and non-uniformity.

5. Error rates for computing elements will affect the results of computations by 2018, and require new approaches to error recovery because current checkpoint/restart schemes have reached their limits for scalability. The increased error rates also may overcome current error detection methods, leading to silent errors (errors that evade detection, but corrupt scientific results).

6. Synchronization will be very expensive. As the number of elements rise, the probability that one member will be much slower is nearly 1 -- forcing a disruptive shift from managing parallelism using straightforward bulk-synchronous models to fully dynamic asynchronous approaches.

7. The I/O system at all levels – chip to memory, memory to I/O node, I/O node to disk and disk I/O – will be much harder to manage due to the relative speeds of the components. These challenges represent a disruptive change in the computing cost model, from expensive flops coupled with almost free data movement, to free flops coupled with expensive data movement reduced reliability and increased system and application complexity.

In either architectural design, it will be more complicated to manage energy-efficiency, concurrency, and resiliency in operating system software and programming models. Fundamental advances in data management and analysis, along with visualization, will be needed to support scientific discovery and understanding. Numerical algorithms, mathematical models, and scientific software codes must also be reformulated (perhaps radically) to take full advantage of these emerging computational platforms. Achieving these improvements will require an unprecedented cooperation between computer architects, domain scientists, application developers, applied mathematicians, and computer scientists.

3. *Green computing: Recycling e-waste.*

Leader: Blaise Aguera y Arcas
David Dean
Andreas Andreou

The basic premise for re-using old computing is that the environmental cost of manufacturing silicon chips outweighs the costs of using them through their lifetime. This may be true for most electronic devices that do not operate 7/24.

The environmental impact of the semiconductor industry was first explored in a controversial paper, in which the authors follow the lifetime of a DRAM microchip from its fabrication to its utilization by consumers by analyzing the material and energy that needs to be expended in order to create one chip[8]. The information in this paper came from several resources, including a United Nations Environment Program (UNEP) and United Nations Industrial Development Organizations (UNIDO) joint publication and a study published by the Microelectronics and Computer Technology Corporation (MCC).

Overall, several forms of chemicals including, deposition/dopant gases, etchants, acids and bases, and photolithographic chemicals are needed. It is estimated that 45.2 grams of chemicals are necessary for the production of 1 square centimeter of semiconductor. The largest composition comes from elemental gases and of which the largest contributor is nitrogen gas. Aggregate chemical usage information widely differs from source to source, anywhere of 1.2 grams of water per square centimeter of fabricated wafer (National level study conducted by the Toxic Release Inventory) to 610 g/per sq cm (UNEP/UNIDO), with a baseline of 45 g/ per sq. cm from the firm. Approximately 18-27 liters of water is necessary to manufacture 1 square centimeter of wafer.

Energy consumption is separated primarily into two divisions – purifying silicon material into silicon wafers and the processing of the wafer into a chip. The first process that takes quartz (sand) and turns it into a silicon wafers totals nearly 3000 kWh per kilogram of fabricated silicon wafer. It takes about 1.5 kWh to process 1 square centimeter of silicon. On a per-chip basis, including consumer use, in its lifetime, it will use up approximately 56 MJ of energy. Every chip manufactured today must be used for about 2 years to outweigh, the energy cost of manufacturing.

[8] *The 1.7 Kilogram Microchip: Energy and Material Use in the Production of Semiconductor Devices* paper by Williams, Ayres, and Heller (Environmental Science and Technology 2002) and follow-up editorial in Nature

Hence it would not be unreasonable to re-use old chips, perhaps in non-computing intense applications to improve the greenness of the semiconductor industry.

4. *What is the processing vs. memory point of cortex?*

Leader: Andreas Andreou
Terry Sejnowski
Larry Abbott
Kwabena Boahen

In modern computer VLSI processors, the physical constraint of a fixed die area imposes tradeoffs on how this space is utilized. More specifically, there is a fundamental tradeoff between the numbers of processors that can be accommodated on a single die and the amount of local memory, notable L2 and L3 cache that can co-exist on a single die (the physical space occupied by the computational resources). Local memory is important because of the high cost in time and energy to access the main memory. This is called locality of reference in computer architectures: For a given problem, the amount of local memory can be optimized so that so that data are used and re-used efficiently by the processors. Modern computing has been shaped by the locality of reference and the related concept of a memory hierarchy: How data are stored as a function of distance from the processor.

Discussions on locality of reference at the 2nd Kavli workshop in Costa Rica has led to a systematic exploration of optimal resource allocation (chip areas dedicated to computation, communication and memory). Given a task or a set of problems, such as inference using graphical models or cross-correlation of astronomical measurements, the overall system performance can be optimized by balancing the performance gains from parallelism, processor microarchitecture, and cache-local-memory with the energy-delay costs of computation and communication.

For example, in speech recognition speech, an optimum multiprocessor architecture for Hidden Markov Models using standard integrated circuit technology has approximately 25 processors. This number does not strongly depend on the absolute area of the die. The processor area is approximately one sixth of the area occupied by memory, which in this case was about 3MB. More generally, for tasks such as inference algorithms based on graphical models the optimal architecture has a "large" number of simple processors with enough local memory so that the processors are able to do their work effectively (low latency) and efficiently (low energy).

The cost function J_D for a given problem that must be optimized to determine the best architecture is shown below (Fig. 2) as a function of the complexity of the processors (A_p) and the amount of cache (A_{L2}). The surface resembles an L-shaped canyon, with a sharp (nearly vertical) back wall and a front wall that rises to the left and right. There is a minimum complexity of the processor and minimum size of the cache, represented by the steep walls. The best combination occurs when the processing complexity and the cache size both exceed a minimum and are roughly matched in capacity.

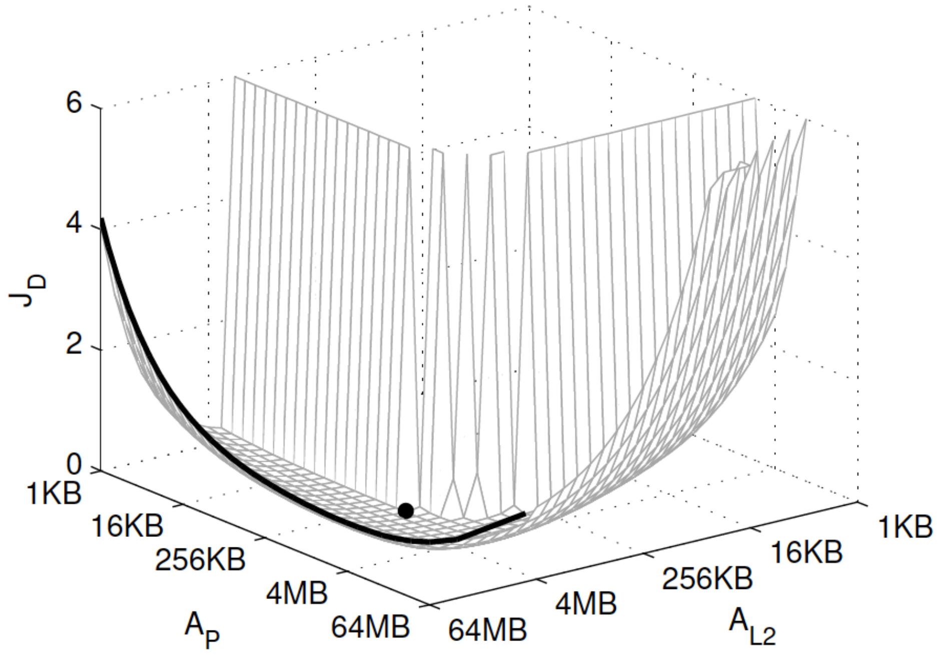

Figure 2. The cost function J_D on the z-axis as a function of the complexity of the processors on the y-axis (A_p) and the amount of cache on the x-axis (A_{L2}).

The brain has billions of neurons interconnected with thousands of synapses. Within the cortex, the density of interconnection is much higher within a cortical column than across long-distance connections. Can we understand complexity and size of computational structures in the brain such as a cortical column based on delay and energy costs? Could it be that in the cortical column neurons form a computational structure that can be considered a single die where locality of reference prevails? If this is the case can we calculate the optimal amount of memory needed in this volume of tissue? Can we identify neural computational structures that correspond to the communication resources found on modern chip multiprocessors? These questions may lead to a better understand of how to engineer modern computing systems that match the high degree of parallelism and complexity in the human brain.

5. *What is the currently best idea for specifying an exascale machine?*

Leader: Roger Blandford
Tom Abel
Risa Wechsler
Larry Abbott
Felix Schuermann

There is no single "best" machine to try to develop for general purpose usage. Conversely, any functioning machine would surely find impressive usage. Indeed, it is not even clear what the prefix "Exa-" qualifies as it could apply to processing speed, fast memory or storage, (though hopefully not dollars!). If one applies the metric of optimizing the science per outlay then we must be prepared to allow the design to evolve over a series of stages as generalized computing capability develops and usage and "proposal pressure" indicate the wisest future investment to make, allowing for the true cost of operations under "green" conditions. Indeed, it may well turn out that putting the resources into software to accommodate multiple hierarchies of memory with built in fault tolerance may produce the largest gains, a form of "hardware-software codesign". Alternatively, for a number of science applications it may be more cost efficient to put the resources into smaller installations that require less sophisticated network designs and I/O infra-structure. That having been said, the quest for exascale computing will inevitably bring down the costs of multi-petascale computing, extrapolating from past experience.

This is illustrated within the subfield of astrophysical simulation. The most complete numerical models of the formation of galaxies and large scale structure in the universe follow aspects of dark matter and gravity, plasma physics, radiation, relativistic particles, chemical processes and increasingly sophisticated subgrid models encapsulating the physics of stars, supernovae and accreting black holes. Today, there are roughly ten competitive codes in production or development. None of them scale efficiently to the petascale let alone the exascale. The inherently wide range of time scales and propagation speeds involved may preclude this. (Illusory, inefficient scaling may be achievable, through performing copious superfluous calculation.) The exascale architectures envisioned now will be a poor match to the software technology for most existing astrophysical simulation. Astrophysical calculations are typically very inhomogeneous requiring a lot of communication and a high dynamic range while machine progress will come most rapidly by developing the raw processing speed. Most of the gains we will see in the near future will come from comprehensive parameter and resolution studies. Thus, an exascale computer could still be used in an embarrassingly parallel fashion, utilizing smaller and smaller portions of the entire machine per workload to run medium-sized N-body volumes exploring parameter space with thousands of models and realizations.

6. *Can we do with less memory?*

Leader: Michael Roukes
Tod Hylton
Andreas Andreou

The traditional transistor scaling paradigm reduced all the dimensions of a transistor proportionately - gate width, length and oxide thickness all becoming smaller by a factor of ~0.7 per node resulting in a doubling of transistor density at each node. In this way of scaling, the transistor channel resistance does not change but the gate capacitance is reduced by a factor of 0.7 resulting in a corresponding decrease of the RC time constant by ~0.7 per node. As gate oxides became thinner, operating voltages also decreased to prevent tunneling and damage of the oxides resulting in roughly constant electric field stress on the materials. This voltage scaling was a welcome effect because it also reduced power consumption by a factor of V^2. Hence, in the traditional scaling paradigm everything was good - more devices per unit area, faster operation and lower power consumption. Also, because of the regularity between technology nodes, it was possible to port designs between nodes without redesigning the entire chip.

Today, however, the scaling paradigm has changed - device size scaling continues (current industry projections indicate this trend may continue for roughly another decade - to the 5 nm to 8 nm node), but operating voltage and frequency scaling have ended. The following is a list of reasons for this change along with some of the associated effects.

- Voltage scaling is limited by the transistor threshold voltage, which must be reduced as operating voltage is reduced. Reducing threshold voltage is problematic as it also reduces the on-off resistance ratio of the transistor. Low on-state resistance (low threshold) is needed for high speed switching, but high off-state resistance (high threshold) is needed to prevent thermally excited carriers from leaking through the switch (resulting in static power dissipation).
- Power consumption in computing systems is reduced by slowing clock rates - dynamic power is reduced approximately linearly with frequency and static power can be reduced by using larger threshold voltages - with an obvious linear reduction in performance. Because consumption scales more than linearly with frequency, multicore systems at lower frequency are an attractive option to improve power efficiency while maintaining performance.
- Applications drive the selection of the clock rate today, but the clock rate of high performance processors have saturated near 3GHz because of limitations in getting heat off the chips. Large supercomputing systems operate at significantly lower clock rates because of power distribution and heat removal constraints at the system level. Power distribution, heat removal and overall power consumption are the principal design constraint in most computer systems today.
- Continued size scaling also introduces many complexities in transistor design making the trades between voltage, frequency and power even more difficult.

- Porting between nodes requires new high level designs and chip layouts.

The net effect of the current CMOS scaling paradigm is that we are essentially out of headroom in power, voltage and frequency - only size scaling remains (but not even that for much longer). Device, circuit and system level designs are also more difficult. However, a new kind scaling has emerged - multiple processors on a chip (including specialized processors or "accelerators") driven by the abundance of transistors and the inability to increase clock rate. This brings with it additional challenges in software and memory access.

Computing systems employ a hierarchy of memory technologies in order to efficiently store and access large amounts of data. The closer the memory is to the processor, the smaller its size, the faster its access time and more power it uses. Most processors use several levels of cache memory (L1, L2, L3) on chip constructed from SRAM. The next level is typically off-chip DRAM followed variously by flash, hard disk, and tape backup. While cache memories are expected to scale with the number of processors (because they employ the same on-chip transistors), access to off-chip memory like DRAM becomes increasingly difficult as the number of processors increases. There simply is not enough bandwidth off the chip to the DRAM to handle the needed memory access as the number of cores on a chip increases (Hence, it is expected that future computing systems will have less memory available per core than has been the standard in the recent past.). This communications bottleneck is replicated also at larger scales in computing systems (between boards and between modules) as the number of processors increases. In general, the communication of information in a computer system is extremely power in intensive. Ongoing efforts to mitigate the communications bottleneck include

- Introduction of new memories integrated on-chip with the multi-core processors (e.g. PCM and other resistive memories)
- Integration in "3D" by chip stacking methods to bring memory chips or communication interfaces in closer proximity to the cores
- Introduction of optical communication technologies at various levels of the system hierarchy.

Ultimately there is no "cure" for the communications problem because while the number of processors scales as the area of the chip (or board), the amount of communication interface scales only as the perimeter. Although mitigated, it is also not "solved" by 3D architectures - here the number of processors will scale as the volume but the communication interface will scale only as the area. Also, today we effectively use the 3rd dimension to remove heat, which we will lose if the computing system is also 3D.

The memory access problem suggests some obvious near-term research priorities: 1) integrating extremely high density DRAM quality memory on multicore chips or 2) stacking memory chips on processor chips with high communication bandwidth connecting them. Many researchers are currently investigating resistive memories such as memristor, MRAM, or PCM integrated into the back-end wiring. Most of these memories suffer from numerous technical challenges, and development costs are extremely large even when candidate devices are identified. Chip stacking technologies face difficult issues of interchip vias, packaging and cooling. The industry clearly recognizes this problem and is investing heavily to address them.

In addition to the memory access problem, the multicore scaling problem brings with it the problem of "parallel" programming. How do we effectively instruct a machine of many cores what to do? There are some classes of problems or applications where we do know what to do - (1) problems that are highly compartmentalized where the processors are assignable to an individual task, (2) problems that are "embarrassingly" parallel with each node computing and communicating with other nodes in a uniform way and (3) problems with modest internode communications ("messages"). This presumably leaves unaddressed myriads of complex problems that the multicore systems might be able to solve if we could only tell them what to do. This limitation may not surmountable in the current practice because the potential state space of even modest computing machines is simply inconceivably large: Full-time coding by all of the humans alive today will not be able to use a tiny a fraction of any modern computing machine. This is not to say that humans cannot still create valuable algorithms, but our ability to create them is impeded by the scale and dwarfed by the capacity of the machines.

Other current trends in multicore scaling are (1) systems heterogeneous cores, thereby allowing specialization of cores for certain tasks, and (2) systems with reconfigurable cores (FPGAs), thereby allowing the portions of the hardware to be configured by the programmer as needed for certain tasks. Each of these approaches enables performance improvement (in power or speed), but generally at the expense of further increase in the complexity of software.

The trends and limits of the current scaling paradigm suggest that computing systems of the future must have the following features.

- The computer will be a "fabric" of processing units, embedded memory and communication network. At the extreme limit, it will be difficult to distinguish "processor" and "memory".
- Individual processing units must operate at low rates and high levels of coordination with other units that are active simultaneously. As mentioned above, slowing the speeds of the devices can yield substantial dynamic and static power savings. The static power savings could be realized by using higher threshold transistors (which are not the focus of current research activities).
- Activity must be driven by "events" rather than by a clock. Systems that are event-driven will be naturally more tolerant to variations in the (presumably much slower) transistor speeds.
- The system must be organized as a kind of heterarchy or "fractal" structure in order to address the surface-to-volume limits inherent in communication. This organization will be needed to maximize information throughput and minimize power.

- The system must be able to learn its own "programs" as it operates because no human will be able to program it effectively.
- The system must continuously and rapidly "heal" itself when components fail because there will be so many failures and so much integration that we will not be able to replace them. The key here is that the system must have a large redundancy in its ability to organize itself to solve a problem. If one component fails then the whole system spontaneously reorganizes around it. This redundancy will be a natural part of the system architecture, rather than component / module / coding specific redundancies used today.
- The system must innately "recognize" and "regulate" its energy consumption so that power is used only as it is needed. Even better, the energy consumption of the machine will be an inherent part of its design and function.

Brains and other natural systems have exactly these features already. The point is not that computers should work like brains; rather, that the trends and challenges facing computing today lead necessarily to that conclusion.

Here are some key research questions associated with the challenges detailed above that could change the way we think about memory and high performance computing:

1. Currently thermodynamics is viewed as a constraint in the realization of computing systems. How do we make thermodynamics a core part of the computational paradigm and the technology that implements it?
2. How do natural systems self-organize and develop algorithms to solve problems? Are overarching physical or thermodynamic optimization principles at work? If so, how do we apply them to computing systems?
3. How can we get computing machines to evolve their own algorithms in a scalable way? How do we describe a problem to be solved and appropriately "seed" the machine with the parts of the answer that we already know and have the machine evolve the rest?
4. Is intelligence due to the thermodynamic organizational principles operating in the brain and other natural systems?
5. What would the components and architecture of new computing machines based on these principals look like?

If these questions can be addressed then progress in computing will be continue to expand beyond our current extrapolation of the current paradigm; if not, then computing could stagnate on the way to exascale computing.

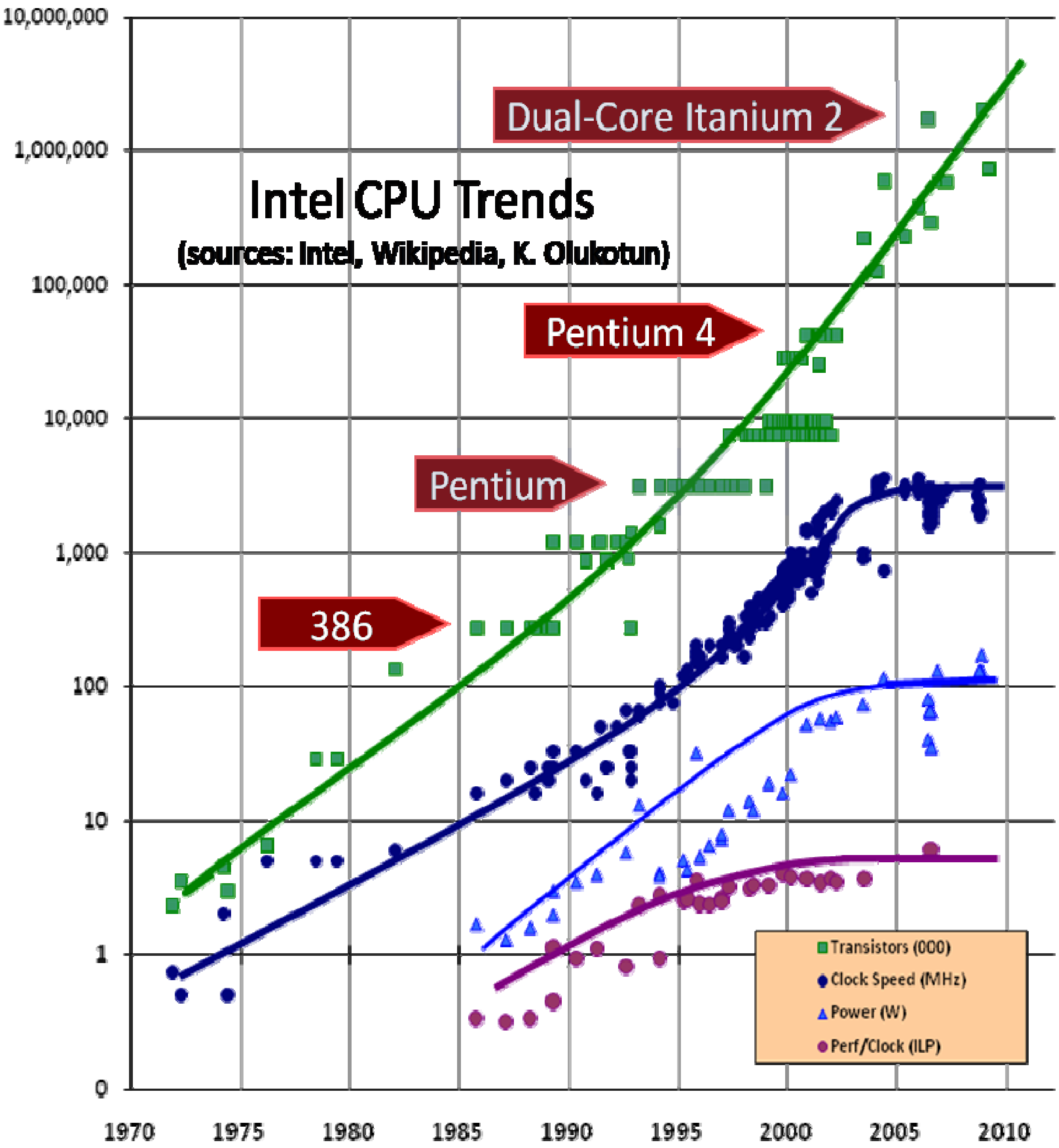

Figure 3. Log of the number of transistors (000) on VLSI processor chips (green squares), clock speed (black circles), power (blue triangles) and instruction level parallelism (red) as a function of time.

"Growing High Performance Computing in a Green Environment"

September 9-11, 2010 in Tromso, Norway

Schedule

Wed September 8 - reception after dinner.

Thursday September 9

9 AM - Setting the Stage

Astrophysics: Roger Blandford (Stanford)
Neuroscience: Terry Sejnowski (Salk)
Nanoscience: Michael Roukes (Caltech)

12:00 Lunch

1 PM: CS and Energy

Horst Simon (Berkeley)
"From Bits to Buildings: Energy Efficiency and the Path to Exaflops Computing"
Steve Koonin (DOE)
"Simulation for Energy (and conversely)"
John Grunsfeld (NASA)
"HPC at NASA"
Dinner

Friday September 10

9 AM – Nanoscience

Tod Hylton (DARPA)
"Perspectives on Intelligence and Computation"
Andreas Andreou (Johns Hopkins)
"Amdahl meets Moore in the multi-core arena"
Kwabena Boahen (Stanford)
"Simulating the brain with neuromorphic chips"

12:00 Lunch

1 PM – Hiking Excursion

Dinner

Saturday September 11

9 AM – Astrophysics

Risa Wechsler (Stanford)
"Computing the Universe in the Age of 10-Billion-Galaxy Surveys"
Tom Abel (Stanford)
"Efficient adaptive algorithms for astrophysics"
Blaise Aguera y Arcas (Microsoft)=20
"High performance computing at pocket scale"

12 Lunch

1 PM – Neuroscience

Larry Abbott (Columbia)
"The high price of spikes"
Felix Schuermann (EPFL)
"Next Generation Blue Brain Architecture"
Panel Discussion

Dinner