Summary Findings of the Workshop on the Frontiers of Extreme Computing 2007

Erik DeBenedictis (Sandia National Laboratory)
Thomas Sterling (Louisiana State University)
David H. Bailey (Lawrence Berkeley National Laboratory)
December 10, 2007

Abstract

Computers have made substantial improvements in performance year-to-year over the last few decades and have now reached petaflops level. The increased availability of computation has powered growth in science and the economy to the considerable benefit of society. This workshop seeks to reconcile user demands for Exaflops and then Zettaflops computers with the roadmaps and research possibilities of the computer builders.

This workshop is unusual in that it includes participants across a broad spectrum of talents. The users understand that only a subset of computer applications make sense at Exaflops or Zettaflops levels, and those applications have specific requirements. While the computer builders say general purpose computers at the Exaflops and Zettaflops levels will be extraordinarily challenging, restricting future capabilities to just those needed by relevant applications will tend to focus the R&D task and yield a simpler, less expensive solution.

Some workshop participants analyzed computer applications with high societal impact requiring Exaflops and Zettaflops levels of computation. Other participants identified ways computers could be built and programmed to the Exaflops level, which ended up requiring heroic extensions of today's methods.

The entire workshop concluded that building an Exaflops computer was a reasonable ten-year goal and the computer would be immediately used to considerable societal benefit. While applications demanding Zettaflops will exist, a computer of that performance will only be possible with new technologies that reach far a field of current methods.

The workshop and this report identify a series of issues that cut across the applications, software, architecture, and technology domains and which will collectively determine the path to reaching Exaflops and Zettaflops computing levels.

Crosscutting issue 1: Legacy and New Applications

The applications group identified supercomputing applications that could justify Exaflops through Zettaflops performance levels based on societal value. For many of the applications, the group went on to devise the requirements for an Exaflops supercomputer.

The Zettascale applications include:

- Climate modeling for mitigating the effects of global warming
- Personalized medicine and phylogenomics for improving health of citizens
- Fusion and astrophysics for advancing science
- Decision making, specifically for policy modeling
- Nano/material science and chemistry for international competitiveness

The requirements set forth by the applications group will be useful in discussions of the architecture and technology:

Performance	1 Exaflops
Main memory	~400 petabytes
Working storage	4 – 40 exabytes
Archival storage	40-1000 exabytes
Bisection bandwidth	0.5 to 1 exabytes/second
Memory performance	Balanced random access and High Performance Linpack (HPL) performance
Software compatibility	Support DOE and NSF's $2.5B MPI software base
Table I: Application demands	

Crosscutting issue 2: Parallelism

The need to accommodate growing parallelism is one of the two profound technical findings of the workshop. There is consensus that all paths to greater performance require more parallelism in the hardware, which software is not prepared to exploit. It has become universally known across the industry that microprocessor clock rates have leveled off at about 2-3 GHz and that vendors are putting additional resources into parallelism in the form of multiple cores. A simple division shows that Exaflops and Zettaflops systems will require a billion and a trillion floating point units operating in parallel.

The various groups illuminated the problem, with a consensus that the issue requires more work.

- The applications group acknowledges difficulty in programming even the 100,000 threads in the Blue Gene systems. This is of course far short of the billion and trillion thread targets.
- The systems software group offers that Thread Level Parallelism (TLP), Parallel Genetic Algorithms (PGAS), and Data Parallel programming are good directions to explore for the representation of parallelism.
- The architecture group accepts the clock rate limit and the resulting billion to trillion way parallelism. Table 2 offers two design points: the first based solely on microprocessors and the second including vector coprocessors.

FLOPS/module	Bytes per	Modules	Total	Power	

	module		Cores		
4 teraflops	16TB	250K	500M	250-500MW	Steve Poole, ORNL
5-10 teraflops	50-100GB	100K-200K	13-26M vector	40-50MW	Steve Scott, Cray
Table 2: Two Projections of 1 Exaflops Supercomputers					

- The technology group confirms that the clock rate flat lining has a bona fide physical basis in power consumption for logic and interconnect. Tom Theis of IBM speaker showed a viewgraph of commercial microprocessor clock rates where the rate overshot the asymptote of about 2 GHz, giving the explanation that the industry had slightly misjudged market demand for the speed power tradeoff.

The workshop-wide consensus is that the future increases in end-user performance needed to meet applications demands will only be possible if software, algorithms, and languages become better able to exploit parallelism in the underlying hardware.

Crosscutting issue 3: Memory and Storage

Memory and storage offer challenges, but there are many promising leads.

The applications group identifies total memory requirements for an Exaflops supercomputer at 400 petabytes and 4-40 exabytes per table 1. While the main memory requirement is large as an absolute number, at .4 bytes/flops it represents continuation of a very slow decline of this ratio that has been occurring over decades.

The software group is eager to manage this expanded memory resource through systems software enhancements. Their observation is that the memory should be accessible as a global address space (although there could be other options) and that this implies a deeper hierarchy that needs to be managed with more operating systems abstractions.

The architects identify the traditional “memory wall” as an important issue potentially mitigated by 3D packaging. The scheme is to explore 3D stacking of CMOS and/or Processor-In-Memory (PIM) options to get memory physically and temporally closer to the processing resource. The advantages of these approaches are to minimize latency, maximize bandwidth, and to introduce added memory functions such as scatter/gather.

The technology working group identified a sizeable range of technology options in the area of memory and storage, all compatible with the conclusions of other groups. There are several dozen companies exploring nano memory options. Many nano memory options are denser or lower in power (even static) compared to DRAM. Many of these are compatible, embeddable, or stackable with CMOS processes, typically offering higher bandwidth to the logic components and breaking down the “memory wall.” There do not appear to be any options that are as fast as DRAM and SRAM, suggesting that the new nano memory options will augment rather than replace current DRAM and SRAM.

There is a consensus of enthusiasm across the groups for nano memory closely integrated with processors to occupy a new spot in the memory hierarchy slightly below DRAM. For example, 1 Terabyte of 100 ns memory layered onto a processor chip. There is concern that commercial developments may not be sufficient to reach the goal on the necessary timescale.

Crosscutting issue 4: Interconnect

The trend in interconnect has been the infiltration of optics from long distance links to progressively shorter ones. This trend held true at the workshop and we now predict the need for optics down to the 100μm connection length.

The applications group identifies similar bandwidth/flops ratio for Exaflops as has been traditional for supercomputers over many years, although this of course implies much larger absolute bandwidths. The expanded bandwidth will require reengineering but does not cause concern. The new developments are in short distance interconnect.

There are technology opportunities for optical interconnect within a chip. The technology group finds that it may be necessary to replace bus-level electrical interconnect within a chip with optics to reach the Exaflops level. Electrical interconnect on a chip obey the diffusion equation (speed drops with distance) unless power-hungry repeaters are used. This makes optics a candidate for moderate length intra-chip communications, due to the properties of optical links that transmission speed and power are independent of distance. There are R&D activities in this area, such as reported by Jag Shah of DARPA in his presentation and many startup companies.

Silicon is not a good material for building lasers, so the chief direction is for externally generated laser light to be piped onto a chip and modulated.

There is a concern that the demands of the commercial marketplace will be insufficient to drive development of on-chip optical interconnect at a sufficient rate to be ready for Exaflops supercomputers when they are needed, thus DARPA and other technology prototyping efforts are endorsed.

Crosscutting issue 5: Reliability

The workshop projects major engineering efforts to meet evolving reliability requirements, but no scientific breakthroughs are needed.

The applications working group reports consistent requirements in terms of Mean Time Between Failures (MTBF) for an Exaflops or Zettaflops system, but these requirements become more difficult to achieve as the systems become larger. The consistent requirement is that jobs complete most of the time (more than about half the time) without crashing due to a hardware failure. This is because jobs can be rerun if failures are infrequent enough. With the flat lining of clock rate, the number of logic components must rise thousand- to million-fold to reach Exaflops or Zettaflops. To maintain constant

MTBF of the overall system, the MTBF of each subsystem must rise by the same factors. This is expected to be difficult to achieve, but possible with cooperation among all the stakeholders.

The applications working group is expecting that users will share the burden of reliability. They request error checking, notification, and checkpoint/restart features in future systems, which users would be expected to use.

The software community proposes to add new features to cope with large systems, specifically facilities for fault-oblivious applications and dynamic reconfiguration.

The architecture working group proposes a stronger set of error tolerance features to be built into the hardware. Hardware is expected to require full Error Detection and Correction (EDAC) for faults as well as design that avoids all single points of failure (or interrupt). The architects predict that system-scale fault modeling will be required.

Crosscutting issue 6: Programming Model

There is consensus across all working groups that current programming models are poorly equipped to deal with exponentially rising parallelism. Rectifying this deficiency promises to be remarkably difficult.

The applications working group points out that there is a 20 year lead time on the emergence of a new programming language, yet the flat lining of clock rate and consequent increase in parallelism have already happened.

The software working group is prepared to embrace new programming models for billion way parallelism, expecting that the emerging model should have a global address space and exploit new architectural features. Transactional memory could be of assistance. There does not seem to be a candidate programming language.

The architecture group is concerned about data locality, or the problem that moving data to the locus of a computation predominates over the cost of the computation. The architects are predicting that Exaflops and Zettaflops systems will trade easy programming (i. e. they will be harder to program) for the necessary features that will permit programmers to control data locality. The architects are expecting to put features into systems that permit data movement control by higher levels of the software stack. The architects expect that compilers will assume some of the burden of data placement.

It seems likely that the hardware community will produce computers of varying designs over the next few decades, all highly parallel but offering innovative features in an attempt to help the programmer cope with parallelism. Highly skilled and motivated programmers will harness some of the resulting machine, distilling a programming model and language over the next 20 years – much like Message Passing Interface (MPI) emerged as a programming model over the last 20 years.

Crosscutting issue 7: Power

The trend to excessive power consumption is a top finding of the workshop. With reference to table 2, the architects' power predictions cover the range of 40-250 MW for a 1 Exaflops supercomputer of a fairly conventional architecture. These power levels and the costs they imply are so high that there will be tremendous motivation to develop different, non-conventional architectures to reduce the costs. In fact, the options discussed in this report could collectively contribute to reducing power by 1-2 orders of magnitude to manageable levels. However, Zettaflops would require an additional 3 orders of magnitude, totaling 4-5 orders of magnitude improvement. Zettaflops seems unlikely with incremental improvements to current technology.

The systems software group is eager to manage power as a precious resource – just as today's systems software manages CPU time, memory, and disk space. One expectation is that hardware will provide information about and control of power consumption in real time for the operating system to manage. In addition, compilers are expected to create code to minimize power consumption based on an implicit understanding of the architecture.

The systems software group expects a power model to be developed, exposing controls on power consumption through cache management, bandwidth management, etc. The systems software group expects that data movement is a greater contributor to power consumption than arithmetic.

The architects understand that they have some control of power consumption. There is a general principle of engineering that lowering the clock rate decreases power consumption disproportionately (at least for dynamic power consumption). This offers a strategy of lowering clock rate while increasing the amount of computational hardware by equal proportions. While this would result in lower power for the same aggregate throughput, it would increase the amount of parallelism and make the computer more difficult to use.

Similar reasoning applies to processor core complexity. A traditional measure of processor architecture quality is the average number of Instructions executed Per Clock (IPC), as higher IPC translates to proportionately higher throughput for a given clock rate. It is possible to raise IPC by increasing cache size and having the processor execute instructions speculatively (i. e. executing instructions before it is known if they will be required based on branch instructions, throwing out the results of instructions that end up not being needed). Unfortunately, these methods of raising IPC disproportionately raise power consumption. The obvious mitigation would be for the architects to use larger numbers of simpler and thereby more power-efficient processor cores. This would again improve overall power efficiency while increasing the amount of parallelism and making the computer more difficult to use.

More locality in data and function placement helps. It takes a lot of energy to move data between memories and functional units. This energy can be reduced by smarter arrangements. Good placement is a universally good idea, yet finding a near-optimal organiza-

tion of a computer's components is a difficult task particularly since it may be different for each application program. The way forward is for the architects to do the best they can while providing information and control to the systems software.

The technologists are eager to confront power issues, although a key power issue is grounded in the laws of physics and subject to circumvention only by highly disruptive changes.

All projections for the basic logic operations that underlie computation (Boolean logic) show substantial reductions in energy per operation over 1-2 decades and then a flat lining at a technology-independent physical limit. The limit is due to the fact that the signals representing 0's and 1's must be sufficiently stronger than thermal noise to be detected reliably. The AND and OR gates of Boolean logic destroy one of their inputs, and so this energy must be dissipated by every logic gate. There will be a discussion of new devices that can circumvent this limit, but the technologists recommend following the industry-defined CMOS roadmap as it approaches this limit over the next couple decades.

Excessive power consumption in interconnect would become a severe problem, but short-distance optical interconnect are a viable mitigation if developed further. This has been discussed as cross-cutting issues 4.

Crosscutting issue 8: Heterogeneous Architectures and Accelerators

There is a workshop-wide consensus that simply replicating microprocessors to the Exaflops level will be at the limit of feasibility and doing so to the Zettaflops level will be imposisble. The systems software group recognizes an immediate need to support heterogeneous systems, which the other groups confirm is the emergence of a long-term trend.

The systems software community is confronted today by high-profile heterogeneous systems comprised of the pairing of a microprocessor with a coprocessor accelerator, such as IBM Cell or ClearSpeed. The existence of these systems highlights the need for systems software where a single, integrated source code can run on the processor-coprocessor combination, with tools and operating systems support.

The architects are aware that lowering control overhead relative to computation will be necessary even with all contemplated device improvements. The architects see the need to reduce control overhead in microprocessors through better design without changing the overall programming model. However, the architects also anticipate acceleration technologies as they tend to be substantially more power efficient. The architects intend to consider bio-inspired architectures as well.

The technologists reinforce the opportunity for architectural change. The technologists confirm that Exaflops are at the limits of conventional architectures with Boolean logic, with Zettaflops being clearly beyond the limit. However, it is possible to deploy transistors in different circuits and architectures that could possibly reach Zettaflops. Alternative

transistor circuits could process information in “analog” forms with much greater power efficiency, such as in biology. In addition, the coprocessors and special function units (e. g. molecular dynamics engines) are perfectly compatible with end-of-the-roadmap Boolean logic and their power efficiency improvement could have the same effect as continued device physics improvements.

Crosscutting issue 9: A New Device

After considering a broad range of options in software, the consensus of the workshop is that Exaflops hardware should be feasible with future semiconductor technology but that Zettaflops will require new technology at the device level. It is theoretically possible to circumvent power limits that apply to Boolean logic, although a new physical device would be needed for which there have been no experimental demonstrations to date. There are several alternatives:

- Reversible logic and logic “out of thermal equilibrium” with the environment recycle or otherwise preserve the energy of 0’s and 1’s through many logic levels instead of turning the energy into heat at each level. The upside potential of these approaches is a substantial reduction in overall heat generation while maintaining a programming model similar to today’s computers.
- The gates in a hypothetical quantum computer would operate on qubits instead of bits. Qubit algorithms can reduce the number of gate operations compared to a classical bit-based algorithm. Certain quantum algorithms reduce the number of gate operations to the square root or logarithm of the number required for a classical algorithm, thus reducing the effective power consumption dramatically.

The technologists consider these options to be viable research directions worthy of support, but they will not displace the need to follow conventional Boolean logic to its end game.

Community Process and Conclusions

The idea of an Exaflops systems simulation came up across several of the groups. The International Technology Roadmap for Semiconductors (ITRS) represents a thorough effort to predict the future of the mainstream CMOS technology that is inevitable for the Exaflops level. In addition, the ITRS roadmap is beginning to show what CMOS will look like at full maturity in 1-2 decades where it would be just barely able to support a 1 Exaflops supercomputer. There is cross-group interest in doing a point design study of an Exaflops supercomputer with fully mature CMOS. This study would give ballpark figures for power consumption and cost, but would additionally give insight into the programming model. Knowing the programming model for an Exaflops supercomputer would give the software community a target.

The systems software group a specific need for better tools, languages, and algorithms for coping with increasingly parallel nature of computers.

There is considerable interest in what comes beyond Exaflops. There is consensus among all groups that Zettaflops will not be achievable with incremental improvements in devices and architecture. However, the technologists encourage research into a new information processing device (i. e. transistor replacement) that could raise performance through low-level physical means. In addition, Zettaflops should be achievable through profound changes in architecture.

The technologists recommend careful monitoring of industry's direction. Industry's investment will dwarf anything the technical will be able to do, but it may be appropriate for the technical sector to augment industrial investment in certain specialized areas. Intra-chip optical interconnect and memory technology are immediate suggestions, but a long term monitoring program would be wise.