Exascale Research: Preparing for the Post-Moore Era

Marc Snir[1], William Gropp[2] and Peter Kogge[3]
6/19/11

Executive Summary

- Achieving exascale performance at the end of this decade or the beginning of next decade is essential for progress in science – including progress on problems of major societal impact (such as weather or environmental impact); essential for the continued certification of the nuclear stockpile; and essential to our national security.
- The rate of advance in the performance of CMOS technology is slowing down and is likely to plateau mid next decade. No alternative technology is ready for deployment.
- Therefore, achieving exascale performance in 20 years may not be significantly cheaper than achieving it in 10 years – even if we could afford the wait.
- It is essential (for continued progress in solving major societal problems, nuclear stockpile, security) to have a sustained growth in supercomputer performance and sustained advantage over competitors and potential enemies.
- To achieve this continued growth, we need research on (a) using CMOS more efficiently and (b) accelerating the development and deployment of a CMOS replacement.
- (a) is (or should be) the focus of exascale research: How to get significantly higher compute efficiencies from a fixed transistor or energy budget. (b) is essential to explore, even if not for exascale in 10 years, as it will be necessary to continue beyond exascale.

[1] Department of Computer Science, University of Illinois at Urbana Champaign, snir@illinois.edu
[2] Department of Computer Science, University of Illinois at Urbana Champaign, wgropp@illinois.edu
[3] Department of Computer science and Engineering, University of Notre Dame, kogge@cse.nd.edu

The Race to Exascale

As petascale systems begin to be deployed in national labs and academia, the race has started for the next level of performance: exascale systems. DOE has announced the goal of an exascale system consuming 20MW in 2018 and has proposed a major initiative toward this goal; various research projects have been started as part of this initiative. DARPA has launched the UHPC research program, with the aim of achieving petascale performance in a single rack system consuming no more than 57KW. We even have a digital "Exascale Report" periodical – clearly showing that exascale is "hot".

On the other hand, there also is some skepticism about exascale. Exascale may be very hard to achieve this decade. DARPA commissioned a study on exascale computing that was published in 2008 [8]. This study developed a strawman for an exascale system built from technology forecast to be available in 2015. The system would consume close to 70MW and would have many drawbacks: in particular, a small amount of memory (3.6 PB), very limited memory bandwidth (~0.002 Word/Flop), very limited network bandwidth, and very low MTTI (40 minutes). The study ignored many practical aspects of a modern architecture: no error detection or correction in memory, no caches, no virtual memory, etc. A subsequent, more detailed analysis by Kogge, showed that, when all the details are taken into account, power consumption is likely to be in the 400+MW range [9]. Therefore, exascale this decade at reasonable power consumption will require major technology advances, and may result in a system that is very different than current supercomputing systems in the balance between the various resources and the execution model it provides to the users. The research and development costs for producing such a system and developing the software needed to use it will be very significant – likely measured in billions of dollars; the yearly maintenance cost (e.g., power and cooling) could be measured in hundreds of millions.

The Need for Exascale

Faced with such costs and technical challenges, it is reasonable to ask hard questions: How important it is to stay on the same performance improvement curve that supercomputing has followed in the last decades? Should we perhaps wait longer, when this level of performance can be achieved at lower cost?

We believe that the answer to these questions is no, for two main reasons: (a) The lost opportunity cost of not achieving exascale performance as soon as technically feasible are very significant; and (2) waiting longer may not help much.

The first point has been documented in a series of reports produced by recent DOE workshops [1]. Exascale performance is essential for making progress on a variety of science problems of great societal impact: For example large-scale simulations play an essential role in the development of alternative energy sources and for increasing the efficiency of current energy consuming. While global warming is an irrefutable trend, assessments of how soon and how large the damage caused by global warming will be still have significant error margins. Both inaction and inadequate action for mitigating global warming or adapting to it can cost millions of lives and trillions of dollars. Exascale climate simulations can significantly reduce the error margins. While not all these applications will require scaling one simulation to billions of cores, they will all require dedicated exascale systems with billions of cores for months and years. Thus, it is essential that such systems be available at an acceptable cost and power consumption.

The need to certify the nuclear stockpile without performing nuclear experiments has been a main motivator for a rapid increase in DOE's ability to perform large-scale simulations. This motivation has not disappeared. As the expert knowledge of people that designed and tested nuclear weapons disappears, and as the design of current nuclear weapons increasingly diverges from designs that were tested, there is an increasing need for more accurate simulations with a broader range of scenarios and models and for better uncertainty quantification. A reduction in the size of the nuclear stockpile makes this problem more acute. This mission will continue to require fast growth in the performance of supercomputers.

Continuing the march toward higher performance is essential for our national security. In an era where information becomes the main weapon of war, the US cannot afford to be outcomputed anymore that it can afford to be outgunned. While the use of supercomputing in national security is, for obvious reasons, less well documented in the open literature than its use in science and engineering, it is no secret than national security agencies are major customers of leading supercomputing systems and will continue to be so in the foreseeable future.

The End of Moore's Law

The second point has attracted less attention. We have become accustomed to the relentless improvement in the density of silicon chips, leading to a doubling of the number of transistors per chip every 18 months, as predicted by "Moore's Law". In

the process, we have forgotten “Stein’s Law”: "If something cannot go on forever, it will stop." The continued miniaturization of transistors cannot go on forever. The International Technology Roadmap for Semiconductors [11] forecasts a feature size of 7.5 nm by 2024. Such a size corresponds to about 30 atoms in a silicon crystal. Clearly, we are approaching the limits of silicon technology. In fact, 7 nm seems to be the limit of CMOS technology, for a variety of reasons [7].

Classic Dennard scaling, where feature size decreases and chip speed increases in proportion, with no increase in power density, has ended at 130 nm, a decade ago [10]. The continuous increase in power density, mostly due to a plateauing of voltage levels and to increasing leakage current, has stopped the progress toward faster circuits. Since then, we have needed new “tricks” for each new silicon generation. ITRS predicts that new materials (e.g., III-V, germanium thin channels, nanowires, nanotubes or graphene) and new structures (e.g., 3D transistor structures) will be needed to continue progress in the coming decade. Even with all these advances, we shall continue increasing the number of transistors per chip, but will need to run them far slower than their potential in order to avoid melting them.

The supercomputers of 70’es and 80’es, in the heyday of bipolar technology, were a miracle of packaging and cooling and a pleasure to see. As we moved to CMOS, packaging became pedestrian and architectures boringly repetitious; however, this “boring” CMOS evolution gave us more than two decades of improving cost-performance in a stable software environment. Supercomputers are beginning to use liquid cooling again. While the technology is impressive and again a pleasure to see, it also indicates that we are approaching the end of the era of fast CMOS scaling. To quote the ITRS: “While power consumption is an urgent challenge, its leakage or static component will become a major industry crisis in the long term, threatening the survival of CMOS technology itself, just as bipolar technology was threatened and eventually disposed of decades ago.” The ITRS “long term” is the 2017-2024 timeframe.

While the similarity to the end of the bipolar era is compelling, there is one major difference: When bipolar technology was coming to its end, Intel had been using MOFSETs to manufacture microprocessors for almost two decades. Today, we have no alternative technology in use.

CMOS technology will not be replaced overnight and will continue to be the foundation for the microelectronics industry for the next couple of decades. However, each additional technological “trick” will raise manufacturing costs and, in particular, will raise non-recurring costs. New lithography techniques will be more expensive: “achieving constant/improved ratio of exposure-related tool cost to

throughput might be an insoluble dilemma" [11]; more materials and more elaborate transistor structures require more manufacturing steps; a tighter control of variance is required as feature size shrinks; 3D structures will be more expensive; and so on. It will become necessary to amortize investments over longer runs, hence to slow-down the rate of miniaturization. Economics (a shift toward lower cost devices, slow adoption of multicore technology, etc.) may further slow down progress. As we get closer to the intrinsic limits of CMOS, the rate of progress is bound to decelerate near the end of this decade.

Alternative technologies, such as spintronics [2], may eventually renew the growth in component performance. However, since we are still at the stage of exploring the basic circuit components, it is very unlikely that any alternative will be ready by the time CMOS plateaus, Furthermore, when MOSFETs replaced bipolar components, it was a case of a cheaper but "good enough" technology that replaced a better but more expensive technology [3]. Spintronics, or other alternative technologies, may not offer any cost savings as compared to CMOS. This would lead to a bifurcation where much of the mass market continues, for the foreseeable future, to use CMOS, with exotic and more expensive technologies used to satisfy higher performance requirements. We already see this happening: Rapid Single-Flux Quantum (RSFQ) Logic has been studied over the last five decades in academia and industrial labs; it has had limited use in products for over two decades [5]. While this technology can offer two orders of magnitude increase in logic circuit speed, and five orders of magnitude decrease in logic power consumption, it is highly unlikely to replace CMOS in mass products in the foreseeable future – as it requires cryogenic cooling and is incomplete in that there is no matching memory technology.

What's Needed

We can react in three ways to this deceleration in intrinsic technology growth:

1. Accept that the performance of supercomputers will plateau mid next decade.

2. Push the performance of CMOS-based supercomputers faster than the rate of increase in CMOS performance

3. Accelerate the rate of introduction of alternative technologies.

We believe that the first alternative is not acceptable – for the reasons previously discussed. As CMOS technology will plateau, deciding against building an exascale

computer late this decade or early next decade may be tantamount to deciding that exascale performance will not be reached for several decades.

Historically, the peak performance of supercomputers has progressed faster than the progress in CMOS performance (physical gate lengths – and thus approximately capacitance and delay – have decreased at a CAGR of 1/1.08, while Rpeak has grown by a CAGR of 1.8), and the cost of the top supercomputers has increased over the years. We have paid this price because of the great societal value of improved simulation capabilities, the importance of increased compute power for nuclear stockpile certification, and for security needs. An exascale supercomputer will be a scientific instrument of no less importance than a leading telescope or particle accelerator – and still cheaper than these. As long as nuclear weapons are still around, we shall need to ensure they are functioning – hence will need exascale performance. Finally, keeping ahead of the curve in supercomputer performance is essential for our national security. But if the curve is flat, there is little advantage in being ahead.

As we expect a hiatus between current, CMOS-based systems and systems using new device technologies, it is essential to pursue both the second and third alternative: Get more out of CMOS and prepare for post-CMOS.

We could, conceivably, push supercomputing performance by brute force: Build systems that consume 500MW or 1 GW and require machine rooms (or machine fields) the size of tens of football fields. Even if the cost of such an approach is acceptable, it is not clear that brute force can lead us to exascale: large physical size is an impediment to scalability and a source of frequent failures (failure rates today for Blue Gene systems have been reported on the order of 0.001 failures per year per socket, and projecting forward to a million sockets yields socket failure rates of minutes). Furthermore, with such an approach, the US will forfeit any advantage over competitors and possible enemies that are capable and willing to outspend us.

The alternative must be a smarter use of CMOS circuitry. The main constraint on large system performance is energy consumption. The main source of energy consumption in a large system is communication: on-chip communication to caches, off-chip communication to memory and inter-node communication. In the follow-on study of Linpack on a 2018 processor as outlined in the Exascale report, the amortized memory access energy per flop was 475pJ, versus an FPU that took only 10pJ; the bulk of this energy was in maintaining the cache hierarchy and in communicating off-chip and off-board to other memories. To improve performance, we need circuits and communication protocols that consume less power, and denser packages that reduce physical distances. Some of the required technology (e.g., low

power circuits) has broad applicability and industry is likely to invest sufficiently to drive its fast progress. Other technologies (e.g., advanced cooling for very dense packaging) may have a much more limited market and will require government funding of industry-research partnerships.

Most importantly, we need to reduce the amount of communication used by computations. One can expect improved algorithms to be the major source of communication reduction. There has been so far little work on communication-efficient algorithms (see [4,6], for recent exceptions): Reducing communication is often seen as a tuning issue, to be handled when the algorithm is coded, but not part of the algorithm design. This must change. We need to better understand the inherent communication requirements of various computations and trade-offs between computation and communication in time (storing to and loading from memory) and in space (communication across cores and nodes). Languages must enable locality control; compilers, runtimes and architecture must reduce the gap between the minimum amount of energy needed to move a chunk of data and the actual amount of energy spent by current memory and communication systems to move such a chunk.

In addition to this "CMOS acceleration" work, we need to prepare as soon as possible for the post-CMOS era. A first step should be an inventory of new device technologies that might be deployed in 10-15 years from now and an analysis of gaps and roadblocks to their deployment. We need to identify where commodity technologies are most likely to diverge from the technologies needed to continue the fast progress in the performance of high-end platforms; and we need government funding in order to accelerate the research and development of those technologies that are essential for high-end computing but are unlikely to have broad markets.

• *While power consumption is an urgent challenge, its leakage or static component will become a major industry crisis in the long term, threatening the survival of CMOS technology itself, just as bipolar technology was threatened and eventually disposed of decades ago.*

References

1. ASCR Program Documents, http://science.energy.gov/ascr/news-and-resources/program-documents/
2. David D. Awschalom, Michael E. Flatté and Nitin Samarth. Spintronics. Scientific American, May 13 2002.
3. Clayton M. Christensen. The Innovator's Dilemma. HarperCollins, 2000.
4. Demmel, J., Grigori, L., Hoemmen, M. and Langou, J. Communication-optimal parallel and sequential QR and LU factorizations, Arxiv preprint arXiv:0808.2664, 2008
5. Dorojevets, M., Current Status and Recent Developments in RSFQ Processor Design. In Serge Luryi, Jimmy Xu, Alex Zaslavsky (eds.) Future Trends in

Microelectronics: From Nanophotonics to Sensors and Energy, pp 229-239, Wiley, 2010.
6. Grigori, L., Demmel, J.W. and Xiang, H., Communication avoiding Gaussian elimination, Proceedings of the 2008 ACM/IEEE conference on Supercomputing, 2008
7. Haensch W.; Nowak, E. J.; Dennard, R. H.; Solomon, P. M.; Bryant, A.; Dokumaci, O. H.; Kumar, A.; Wang, X.; Johnson, J. B.; Fischetti, M. V.; Silicon CMOS devices beyond scaling. IBM Journal of Research and Development 50(4.5) July 2006, pp. 339-361.
8. Peter Kogge (Editor & Study Lead) ExaScale Computing Study: Technology Challenges in Achieving Exascale Systems, DARPA IPTO Technical Report, 2009.
9. Kogge, P.; La Fratta, P.; and Vance, M.. Facing the Exascale Energy Wall, Int. Workshop on Innovative Architectures, 2010.
10. Kung, K.J., CMOS scaling beyond 32nm: challenges and opportunities. 46th ACM/IEEE Design Automation, pp. 310--313, 2009.
11. International Technology Roadmap for Semiconductors.