Community Response to RFI on Incentives, Infrastructure, and Research and Development Needs To Support a Strong Domestic Semiconductor Industry

Written by: Tom Conte (Georgia Tech), Nadya Bliss (Arizona State University), Ian Foster (University of Chicago), William Gropp (University of Illinois), Brian LaMacchia (Microsoft Research), Vivek Sarkar (Georgia Tech), and Cliff Young (Google)

Introduction

This response is from the Computing Research Association (CRA)'s Computing Community Consortium (CCC), with input from CRA-Industry. CRA is an association of nearly 250 North American computing research departments - academic, industrial and professional societies. The mission of the CCC is to enable the pursuit of innovative, high-impact computing research that aligns with pressing national and global challenges. CCC is a responsive, respected, visionary organization that seeks diversity, equity, and inclusivity in all of its activities. The CCC brings together a diverse set of individuals representing the broad community to lead initiatives and activities, such as this response.

Historically, computing applications drive much of the semiconductor industry. For example, 15 years ago the PC industry put pressure on semiconductor manufacturers to advance to the next node in the roadmap. Semiconductors and computer performance remain closely tied. The popular interpretation of Moore's Law was not that semiconductors became cheaper exponentially but that computer performance increased exponentially. This led to a general philosophy among computer scientists that anything was possible and Moore's Law would make it so. "Andy [Grove] makes my computer faster. Bill [Gates] uses more of it" was the adage of the day [1].

After the end of Dennard scaling in the mid 2000's, there began a divergence between computer application needs and semiconductor device performance. Computing realized that continuously improving the cost per transistor– the overarching goal of the semiconductor industry– did not guarantee to improve system performance overall. Moore's Law and computer performance are not inextricably linked. System power efficiency is an equally important dimension. The comfortable ignorance between the semiconductor industry process designers and the computer system application designers was no longer possible [2].

Increasingly, there is an understanding that computing applications should be a driver that determines semiconductor industry decisions [3]. In order to continue to scale computing

performance, a holistic approach must consider changes at all levels of the computing stack. This includes changes not only in algorithms and software systems, but also in system architectures, circuit design, and semiconductors. There are potential solutions to application problems that are dismissed by the semiconductor industry as unimportant or impractical. For example, it's well known that devices designed for analog and those designed for digital are very different. But computing is "stuck" using digital devices to solve inherently analog problems such as machine learning, which is in essence brain-inspired computing. Analog and mixed-mode solutions are today still thought of as fringe technologies. What's more, the analog approaches must use CMOS devices in order to get economies of scale even though such devices have poor performance in analog.

We view that computing applications should have an equal part in influencing decisions about semiconductor technology and investment. This response focuses on the need to achieve this balance in implementation of the *Semiconductor Financial Assistance Program, (*Section 9902 of the *William M. (Mac) Thornberry National Defense Authorization Act for Fiscal Year 2021)*. Below we focus on driving applications in computing, broadly, as the areas of artificial intelligence, high-performance computing, experimental scientific computing, and security. We address this need in the context of the National Semiconductor Technology Center. In addition, we provide input in the RFP areas of fostering collaboration and dealing with intellectual property.

Importance of Computing Application Co-Design in NSTC

High performance computing

In high performance computing (HPC), many applications can make effective use of parallel computing. We can define $T_{A,X}(N,P)$ as the time to solution for application A on platform X, where the size of the input is N, and the number of "processors" allocated from X to the application is P [4]. We refer to the combination of A and X as the system configuration. A system configuration A,X is said to exhibit *strong scaling* if, when we hold the problem size N constant, the time to solution decreases as P is allowed to increase, that is $T(N,P1) > T(N,P2)$ if $P1 < P2$. We say that the system exhibits perfect strong scaling if $T(N,P) = T(N,1)/P \ \forall \ P$.

In many cases, applications are only capable of exhibiting *weak scaling* rather than strong scaling. In weak scaling, the amount of work performed by an application increases with the number of processors (instead of the execution time decreasing for a fixed amount of work, as in strong scaling). More formally, the problem size N that can be solved in constant time increases as the number of processors P increases, i.e., for any $P2 > P1$, there is some $N2 > N1$ such that $T(N2,P2) = T(N1,P1)$. The opportunity for weak scaling was stated in John Gustafson's classic CACM 1988 paper on "Reevaluating Amdahl's Law". Broadly speaking, weak scaling requires increasing bandwidth and total systems memory, which worked well in the early days of HPC and Massively Parallel Processing (MPP).

However, since the end of Dennard Scaling, there has been an increasing need for strong scaling so that hardware parallelism can be used to help reduce latency, since traditional

hardware latencies for inter-node communication, memory access, and intra-node data transfers (e.g., between CPUs and GPUs) have stayed nearly flat since the end of Dennard scaling. This is a clear illustration of the divergence between computer application needs and semiconductor device capabilities. It has been observed [5] that we are rapidly approaching a disruptive period of application redesign and reimplementation of applications due to this divergence, and that this disruption will surpass the significant disruption experienced by the HPC community when transitioning from vector to MPP platforms. Further, there is an increasing diversity in the hardware technologies that are emerging in the future, a phenomenon that has been labeled as “extreme heterogeneity”. Nevertheless, there are opportunities that can be explored across all levels of the computing stack to further improve performance and reduce latencies, despite these challenges. Some of these opportunities rely on converting synchronous operations to asynchronous operations, e.g., replacing synchronous accesses to remote data by asynchronous active messages as in actor models, and by performing collective scatter/gather operations asynchronously. In general, co-design of semiconductor innovations with advances in HPC applications is necessary to address the needs of future applications on future hardware platforms.

Artificial Intelligence

Artificial Intelligence has succeeded over the last decade in large part to the advance of Machine Learning (ML). ML works in two separate phases. In the first phase, labeled data is used to train a neural network. This is done via back propagation and stochastic gradient descent. Larger networks (more neurons and more weights) are more accurate, but they require more training. The success of AI in general and ML in particular has been due to the availability of powerful compute platforms. But there is a disconnect between the needs of machine learning and the goals of the semiconductor industry.

Today, machine learning accelerators for training share many of the attributes of both server-class CPUs and HPC: they are aggressive in all three of computation (hundreds of teraflops/s per chip), memory bandwidth (HBM enables TB/s per chip), and interconnect (dedicated links such as TPU ICI or NVLINK provide 100s of GB/s/chip). However, there are several areas where priorities for machine learning accelerators differ from server and HPC-class machines, summarized below.

Memory capacity for ML remains a concern. Only by aggregating thousands of nodes of high-bandwidth memory (HBM) can we reach the multiple terabytes of memory capacity required to train giant language models such as GPT-3. Coupled to this is that on-chip SRAM is running out of steam and scalability with each process node. While today transistor counts are faithfully tracking Moore’s Law, the ability to put more than about 100MB of SRAM on-chip is limited. Packaging and chiplets are not viable solutions. Compute die stacking doesn’t “power scale” for multiple hundred-watt machines—you can’t stack SRAM over a hot compute unit.

While time-to-solution remains important, micro-latency is not architecturally important in ML machines. That is, they do not need general-purpose CPUs with 4-5GHz clock rates and aggressive branch predictors, and they don’t need fine-grained (32 bit) access to memory at single-clock latencies served out of a tiny L1 cache. Instead, what is needed is bulk throughput: some way to move all of the terabytes/sec of available HBM bandwidth to the compute units.

The latency of the transfer is not a primary concern because the computations can be scheduled ahead of time: the computations are pipelinable and easily scheduled.

Of the three axes, interconnect is the hardest to scale. Electrical SERDES technologies advance at a certain rate, and the only way to improve bandwidth is to use more pins and wires. Optical alternatives are potentially interesting but still expensive for power, area, and integration complexity.

The above illustrates how far the semiconductor industry is out of sync with the needs of machine learning and AI. There is a critical need for coordination and collaboration between machine learning R&D and the NSTC.

Experimental Scientific Computing

Advanced pixel detectors within scientific instrumentation represent an exciting but also challenging source of requirements for custom VLSI [6]. Advanced pixel detectors have revolutionized numerous scientific disciplines, from astronomy to biochemistry and materials science–as well as transforming the photography industry. New instruments are now demanding advances not only in detector technology but also in VLSI. To give one example, nanoscale X-ray imaging is a crucial tool for a wide range of scientific explorations, from materials science and biology to mechanical and civil engineering. Next-generation light sources will increase X-ray beam brightness and coherent flux by 100 to 1,000 times, opening up the possibility of imaging macroscopic objects at nanometer resolution. Such a capability would make it possible, for example, to determine the synaptic connectivity of an entire mouse brain.

Imaging larger samples in this way requires that the continuous frame rate of pixel array detectors be increased to 1 MHz or even more. Such increases are technically feasible: indeed, European groups have demonstrated data collection at 4 MHz [7][8]. The critical bottleneck to effective sustained MHz+ imaging is the resulting data collection rate: with just a 256×256 array of 16-bit pixels, 1 MHz translates to a sustained data rate of 1,000 Gbps (i.e., 1 Tbps); with higher frame rates and larger pixel arrays, 10s of Tbps can easily be imagined. The 4 MHz detectors just mentioned overcome this problem by collecting data in brief bursts to pixel-adjacent buffers that are then drained, over a much longer period, to off-chip memory for analysis. Sustained MHz+ imaging requires instead the use of pixel-adjacent VLSI to perform data compression or AI-based feature extraction in a streaming manner so as to reduce off-chip data rates by several orders of magnitude. This capability will enable not only the observation of phenomena at faster timescales, but also smarter experiments that can focus quickly on important regions of interest and detect rare events.

Such applications require methods that can allow rapid design and fabrication of custom integrated sensor + embedded VSLI chips capable of processing multi-Tbps data streams.

Security

There are at least three dimensions of security that NSTC should consider. The first is the operational aspect of cybersecurity, information security, and security-related intellectual property that is addressed by other sections of this RFI and is out of scope for the CCC response. The second is ensuring that there is a significant, cross-cutting security element

within the NSTC research agenda. Most of this section outlines specific research directions relevant to NSTC. It additionally may be relevant to develop a similar set of “Bill of Materials" requirements for hardware manufacturing as are outlined in the 2021 Cybersecurity Executive Order [9]. Finally, the third dimension is the consideration of resilience and stability of semiconductor supply chains (especially in the context of critical minerals), as touched on below.

From a research perspective, it is vital that NSTC supports and leads efforts in hardware security and verification, including support for cryptography-related requirements. A growing research challenge is potential vulnerabilities that can be introduced into systems via hardware security design flaws [10]. Any semiconductor research effort should consider these potential vulnerabilities during the design process and thus it is vital to have cybersecurity experts embedded with semiconductor experts. Additionally, vulnerabilities can be introduced into the fabrication process so developing ways to mitigate those risks would be important.

Given the tight coupling between hardware and software (including the need for co-design and co-optimization of hardware and software components), it may be beneficial to consider a companion set of initiatives/guidelines to the Executive Order on Improving the Nation’s Cybersecurity [11]. Of particular relevance is Section 4 on Enhancing Software Supply Chain. It would similarly make sense to consider enhancing hardware supply chain and ensuring that interactions between hardware and software do not introduce new potential vulnerabilities.

Another important consideration for future silicon architectures is the degree to which they support cryptography-related requirements and use cases, in particular the design principle of cryptographic agility and secure and performant implementation of post-quantum cryptographic (PQC) algorithms. Cryptographic agility is a security design principle that allows computing systems to be easily reconfigured from using one cryptographic algorithm to another. This property is extremely important when designing secure systems to be robust against cryptographic attack or cryptanalytic weaknesses in underlying algorithms. Cryptographic algorithms can weaken and fail over time due to improvements in cryptanalytic techniques; when that happens, devices need to be reconfigured quickly to no longer use or depend upon the now-weakened algorithms.

Our ability to transition devices and ecosystems to new cryptographic algorithms is already being tested by the upcoming transition to post-quantum (a.k.a. quantum-resistant) public-key cryptographic algorithms [12][13]. NIST is currently in the process of selecting new public-key algorithms that are designed to be secure even against an adversary with access to an industrial-scale quantum computer (>= 1,000,000 physical qubits). New semiconductor architectures should consider functional elements designed for these new PQC algorithms; many of the candidate PQC algorithms will benefit from larger multiplier units and more parallel multiplier units on the die. Specialized instructions designed to accelerate standard cryptographic hash functions like SHA-3 would also be helpful to the cryptographic community and its customers.

In addition to potentially providing functional units designed to improve the runtime and side-channel security of widely deployed cryptographic algorithms, new architectures and manufacturing processes would also benefit by including security and cryptography features

aimed at ensuring the integrity of the semiconductor manufacturing process and overall supply chain.

CCC Views On Other Aspects of the RFI

Value of Community and Collaborative Research in the NSTC

The CCC believes that community access to results and work products are essential to the success of any large endeavor such as the NDAA. As discussed above, often today the design of hardware and software happens separately, leading to significant effort to optimize applications and efficient use of said hardware. Working from the application level and the device level in parallel and collaboratively is essential to the next generation of semiconductor technology. The NSTC must support R&D groups that operate in a collaborative environment to support co-design and optimization of materials, hardware, and software layers. It is the CCC's position that the NSTC must not only work with leading computing researchers but must also incorporate a division of these researchers inside its organizational structure. If such co-design of applications and devices happened in the earlier stages of semiconductor device development, it would not only contribute to increasing usability and efficiency but also to an increased likelihood of transition to and engagement of the private sector.

NSTC Intellectual Property and Openness

The CCC promotes the use of open source policies wherever possible. However we understand that open source is not a panacea to intellectual property issues. The importance of intellectual property protections for advanced technology cannot be obviated via open source.

In computing research, consortia established with non-exclusive, royalty-free licenses to consortium members has proven to be an effective model for industry/university collaborations. But there is an increasing realization that this is not enough. Research must be published for academic teams to succeed while delayed for property rights to be preserved. It is our opinion that, in the post *America Invents Act* era of "first to file," the NSTC must provide sufficient funds to support *agile patent filing:* rapid and well-funded filing of all potential patentable art with only minimal involvement of review committees. Review committees must not be forced to make decisions on filing based on scarce resources for filing costs. Only with sufficient funds can the conflict between the need for research publications and for intellectual property protections be resolved. The NSTC structure must take this need into account as it is not insignificant nor is it a secondary consideration. It must be worked into the structure of its intellectual property policy.

CCC Views on Roadmapping

The CCC believes that the NSTC should not reinvent the wheel. The IEEE has an active, broad, and comprehensive road mapping effort in the International Roadmap for Devices and Systems [14] which is the direct descendant of the International Technology Roadmap for Semiconductors. Indeed, the IEEE created the IRDS with the team from the ITRS when the Semiconductor Industry Association ceased supporting the ITRS in 2016 [15].

The IRDS roadmap is a consortium of constituent roadmap entities in the US, in Japan [16] and the European Union [17]. There is only a loosely organized presence for the US in the IRDS (composed of IRDS contributors who are also IEEE-USA members). NSTC can play that role of an US entity in the IRDS.

Bibliography

[1] Nicholas Negroponte. (2005) Negroponte: Laptop for Every Kid. Wired. https://www.wired.com/2005/11/negroponte-laptop-for-every-kid/

[2] T. M. Conte, E. P. DeBenedictis, P. A. Gargini & E. Track, "Rebooting Computing: The Road Ahead," Computer, January 2017, Vol. 50 No. 1, Pages 32-42

[3] (2020) IEEE International Roadmap for Devices and Systems. “Applications Benchmarking.” https://irds.ieee.org/images/files/pdf/2020/2020IRDS_AB.pdf

[4] (2009) Sarkar, V., et al. “DARPA Exascale Software Study: Software Challenges in Extreme Scale Systems.” https://cpn-us-w2.wpmucdn.com/sites.gatech.edu/dist/5/462/files/2016/08/ECSS-report-101909.pdf

[5] (2019) Sarkar, V., et al. “Future High Performance Computing Capabilities.” Summary Report of the Advanced Scientific Computing Advisory Committee (ASCAC) Subcommittee. https://science.osti.gov/-/media/ascr/ascac/pdf/meetings/201903/Future_High_Performance_Computing_Capabilities_ASCAC_20903.pdf?la=en&hash=629548777233D4B8043E0C07706DA519101367E3

[6] Garcia-Sciveres, M. & Wermes, N. “A Review of Advances in Pixel Detectors for Experiments with High Rate and Radiation,” Reports on Progress in Physics, 2018, Vol. 81 No. 6, Pages 066101

[7] Allahgholi, A. et al., “Front End Asic for AGIPD, a High Dynamic Range Fast Detector for the European XFEL,” Journal of Instrumentation, 2016, Vol. 11 No. 1, Pages C01057-C01057 https://doi.org/10.1088/1748-0221/11/01/C01057

[8] Hart, M., et al. “In Development of the LPD, a High Dynamic Range Pixel Detector for the European XFEL,” IEEE Nuclear Science Symposium and Medical Imaging Conference Record (NSS/MIC), October 2012, Pages 534-537 https://doi.org/10.1109/NSSMIC.2012.6551165

[9] Joseph R. Biden Jr. (2021) Executive Order on Improving the Nation’s Cybersecurity. https://www.whitehouse.gov/briefing-room/presidential-actions/2021/05/12/executive-order-on-improving-the-nations-cybersecurity/

[10] Mark Hill. (2018) A Primer on the Meltdown & Spectre Hardware Security Design Flaws and their Important Implications.

https://www.sigarch.org/a-primer-on-the-meltdown-spectre-hardware-security-design-flaws-and-their-important-implications/

[11] Joseph R. Biden Jr. (2021) Executive Order on Improving the Nation's Cybersecurity. https://www.whitehouse.gov/briefing-room/presidential-actions/2021/05/12/executive-order-on-improving-the-nations-cybersecurity/

[12] Brian LaMacchia, "The Long Road Ahead to Transition to Post-Quantum Cryptography," Communications of the ACM, January 2022, Vol. 65 No. 1, Pages 28-30

[13] Campagna M., LaMacchia B., & Ott D. (2020) Post Quantum Cryptography: Readiness Challenges and the Approaching Storm. https://cra.org/ccc/wp-content/uploads/sites/2/2020/10/Post-Quantum-Cryptography_-Readiness-Challenges-and-the-Approaching-Storm-1.pdf

[14] IEEE International Roadmap for Devices and Systems. https://irds.ieee.org

[15] (2017) IEEE International Roadmap for Devices and Systems. "Executive Summary." https://irds.ieee.org/images/files/pdf/2017/2017IRDS_ES.pdf

[16] Japan Society of Applied Physics. "System and Devices Roadmap of Japan" https://www.sdrj.jp

[17] European Academic and Scientific Association for Nanoelectronics. "SiNANO Institute." https://www.sinano.eu

****

This material is based upon work supported by the National Science Foundation under Grant No. 1734706. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.