Georgios Michelogiannakis Research Statement

Moore's Law is a techno-economic model that has enabled the Information Technology (IT) industry to nearly double the performance and functionality of digital electronics roughly every two years within a fixed cost, power and area budget. Within a decade, it will be feasible to manufacture traditional MOSFET devices with characteristic dimensions in the 3nm-5nm range, which corresponds to a dozen or fewer Si atoms across critical device features and will therefore be a practical limit for controlling charge in a classical sense. This challenges multiple computing domains which have come to depend on the rapid, predictable, and cheap scaling of computing performance to meet mission needs such as for scientific theory, national security, large-scale experiments. Society more generally has come to expect the benefits provided by Moore's Law for consumer electronics and data centers. The deeper issue in these changes is the threat to the future economic growth of the U.S. computing industry and to society as a whole.

This approaching end of lithographic scaling does not mean the end of performance scaling for digital computing, but an invitation to identify novel methods to continue this scaling. Investments in digital computing need to co-exist alongside new forms of computing such as neuromorphic and quantum. However, both neuromorphic and quantum computing models apply to specific domains of problems that do not include a wide variety of critical applications for multiple computation domains [35]. As no single, general-purpose computational alternative has yet risen, combined with the significant investments already made to digital computing, a recent IDA-DARPA report expects that CMOS will remain dominant well after Moore's law ends [33].

In this statement, I describe my strategy to preserve digital computing performance scaling. This includes emerging devices, emerging memories, deep 3D integration, and specialization. These directions fit into the big picture of Figure 1. I then provide an overview of my other research interests, as well as relevant recent projects.

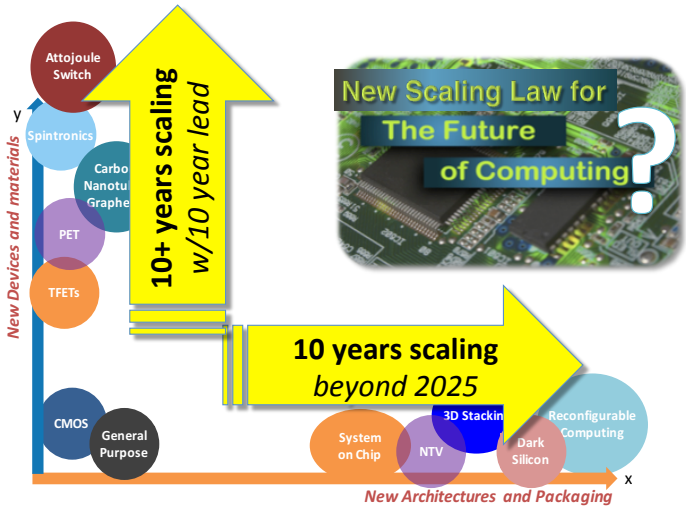

Figure 1: The roadmap for continuing performance scaling of digital computing includes new technologies and architectural specialization.

Architectural Specialization

Specialized architectures, typically in the form of accelerators, have been shown to provide orders of magnitude speedup for constant energy consumption on certain applications. However, typical accelerators today do not explore the spectrum of specialization because they retain a level of programmability (e.g., GPUs). As part of this thrust, I will quantitatively demonstrate the impact of architectural specialization for applications that are critical to society and depend on digital computing, such as climate modeling and density functional theory. Preliminary projections indicate we can pack the equivalent performance of a petaflop-scale system in a single 300mm2 100W package for a specific HPC application. An important aspect of this thrust is exploring the spectrum of specialization: from partly programmable and flexible (thus with higher overhead), to extreme specialization (e.g. one architecture per computational motif) through a family of novel non-VonNeumann architectures that conform to the structure of the problems they are solving.

This thrust includes different kinds of accelerators, not just computational accelerators. As an example, in my previous work I developed prefetchers and data transfer engines for data-parallel applications that accelerate memory access [27,34]. Future additions to this work may include cache access accelerators or data movement accelerators. In addition, this architectural exploration will include inexact computing because numerous of the applications that depend on digital computing can in fact tolerate some error, and inexact computing is a promising avenue to tolerate increased error rates from emerging devices and

other technologies without adding the software or hardware overhead to correct those errors. Finally, another focus of this thrust is how the software (such as compilers, run-time systems or the algorithms themselves) will need to adapt to best make use of different types of specialized architectures.

Emerging Devices

There are numerous new CMOS replacement device technologies proposed, but there is little understanding of their implications for system-scale architecture, application performance, and software programming environments. Numerous alternative transistor technologies to replace CMOS MOSFETS are being evaluated. In particular, tunnel FETs (TFETs), negative capacitance transistors (NFETs) and carbon nanotube FETs (CFETs) have all been demonstrated in fabricated chips. As such options keep emerging, it will be crucial to evaluate their impact at a circuit and systems level. For example, a number of new devices offer their best energy efficiency at extraordinarily low clock frequencies (1Mhz or even 100Hz), which requires many orders of magnitude more concurrency to match the performance of today's systems, but their energy efficiency would enable many orders of magnitude more growth in overall functionality per watt and per unit area.

As part of this thrust, I will first develop compact models of mature and promising new devices and use them in architectural-level simulators or synthesis flows. With this, I will use those models to first determine how to best make use of each device (i.e., what architectural or software changes it requires), what drawbacks each device introduces, and more importantly what is the potential for each device. This is a critical step for constructing the roadmap for the future of digital computing.

Emerging Memories and Deep 3D Integration

While today's 3D integration mostly manifests in 3D memory, future technology promises deep 3D stacks with hundreds of layers including multiple logic layers in different parts of the stack. This is a particularly promising direction since multiple levels of logic as well as multiple layers of memory (each of which is closer to logic) can vastly reduce the cost of data movement. In addition, this is combined with emerging memory technologies such as resistive and magnetic RAMs. Such new technologies offer attractive performance and cost tradeoffs compared to traditional memory technologies, and are non-volatile. Having non-volatile memories close to computation cores create opportunities such as attractive options for power management. However, benefits also come with drawbacks such as error rates as well as the bandwidth and energy challenge of inter-layer communication that stems from VIA technology. Future VIAs as well as alternative inter-VIA communication promise denser and cheaper connections.

As part of this thrust, I will develop models of a 3D integrated chip with emerging memory technologies. Then, I will explore the space of possibilities of different memory technologies and their location in relation to logic layers, as well as explore other challenges such as inter-layer technology and heat density. The end goal is an understanding of how far 3D integration and emerging memory technologies can take digital computing, and what is the most efficient way to do so.

Other Research Directions

My research interests extend to other microarchitectural challenges [13,14] relevant to large data [17,25], memory hierarchy [20], approximate computing, GPUs, impact to the software layer [30], and other adjacent areas. Many of these areas are pathways for further performance over power improvements, but introduce challenge such as at the very least heterogeneity.

In addition to the above, my research interests and past experience extend to networking. Numerous networking challenges today are fundamental to our current messaging models and network architectures. For instance, task placement is a complex problem that is constrained by not being able to change how data maps to tasks. This also complicates programming by forcing the programmer to decide data placement in tasks, often agnostically from the network and other runtime information. In addition, current networks often exhibit performance that is hard to explain or predict. Resiliency is also made harder by the lack of easy data replication and relocation, especially given predictions for failures to the order of minutes in future extreme-scale networks. To tackle these challenges, a name-centric networking approach in modern datacenter and HPC networks holds promise because the burden of deciding data placement is transferred to the network and system software. This enables easy and abstract data replication, security and isolation, but also enables task and data placement to be performed with full

knowledge of the network configuration and dynamic state. Also, programming becomes easier because programmers address data objects without having to know their location, and data objects can be detached from tasks. Building on name-based addressing, I will develop powerful placement algorithms that can place data independently from tasks, and can duplicate data to eliminate congestion or to respond to a failure, as well as cache data in network switches instead of just endpoints in order to reduce communication distance. Furthermore, I will develop performance analysis methodologies and theory that learn traffic and report to the programmer or user the source of performance degradation and provide recommendations for more suitable topologies or communication patterns. This couples with performance prediction models I will develop to be used in advance of application execution. This messaging programming interface will be both simpler and more powerful to the programmer, because the application can now provide hints to the network for upcoming significant communication, as well as predict the cost of communication at message granularity. This work creates avenues for improved optimization methodologies such as communication-avoiding algorithms because these optimization methods can now gain direct feedback on the actual of predicted cost of each communication event.

Recent Relevant Research

My past work has tackled several microarchitectural challenges relevant to HPC. In particular, I have designed a hardware prefetcher that detects data-parallel application access patterns to both make better predictions and also improve the access pattern that the memory experiences, which affects its performance and energy characteristics [34]. Complementary to this work, I have designed a software and hardware collective data transfer engine for the same purpose of improving memory performance [27]. Both of these mechanisms are critical to memory bandwidth-bound applications that are prevalent in numerous computing domains, and demonstrated an 8% and 20% average application performance improvement respectively, with an up to 2.2x reduction in DRAM read power. Another microarchitectural solution is on-going work on lock acquisition prediction for the purpose of eliminating the lock acquisition latency for non-contended locks in chip multiprocessors.

In addition, I have proposed hardware support for fixed-point representations of double-precision variables such that system-wide reduction operations, such as summations and dot products, can be performed with billions of operands with reproducibility and no precision loss [2]. This is feasible with performance comparable to that of double-precision summation, and often times with capabilities existing network interface cards have. I am also currently looking at task placement in a large-scale system by designing an efficient heuristic algorithm that is both scalable and improves on past work using simple key insights to improve heuristic algorithm performance [28]. Related to optimal use of networks, one of my previous studies was one of the first to explore hierarchical cache hierarchies as well as how choices of the on-chip network affect the cache hierarchy in a CMP, illustrating the benefits of co-design [12]. Among other conclusions, that work demonstrated that network topology affects whether private or shared level 2 caches were optimal, because each topology favors different communication patterns.

Another on-going project is “OpenHPC”, which strives to provide open-source software and hardware simulation capability and include all necessary components for a chip multiprocessor (such as the network, caches, cores). This direction is inspired by the need to reliably model future technologies [32].

I have also extensively worked on network design, both on and off chip. For off-chip HPC-style networks, I proposed the channel reservation protocol (CRP) [1] to completely prevent in-network and endpoint congestion with lossless flow control. CRP achieves this by enabling sources to reserve multiple network resources with a single request, but without idling resources like circuit switching. CRP also prevents congestion with many short-lived adversary flows, where other techniques such as ECN are ineffective.

For on-chip networks, I have developed the elastic buffer flow (EB) flow control which removes input buffers from routers and uses channel and router pipeline master-slave flip-flops as two-slot FIFOs [8,9,23,24]. I have also led a study that highlighted the drawbacks of deflection-based bufferless flow control for on-chip networks [7]. Furthermore, I have developed packet chaining that addresses the traditional tradeoff between allocator performance and complexity [6,22] and can outperform significantly more complex allocators. Moreover, I have designed an eager routing speculation scheme where flits bypass routers to follow pre-configured routes [10]. Finally, I have proposed adaptive bandwidth networks that address the growing concern of on-chip network leakage power [26]. ABNs divide channels and router buffers into lanes, and only activate the necessary resources. I have also participated in numerous other adjacent projects [3,4].

References

[1] Michelogiannakis, G.; Jiang, N.; Becker, D.; Dally, W.J.; "Channel Reservation Protocol for Over-Subscribed Channels and Destinations" in *SC 2013*.
[2] Michelogiannakis, G.; Li, X.S.; Bailey, D.H.; Shalf, J.; "Extending Summation Precision for Network Reduction Operations" in *SBAC-PAD 2013*.
[3] Jiang, N.; Becker, D.; Michelogiannakis, G.; Balfour, J.; Towles, B.; Kim, J.; Dally, W.J.; "A Detailed and Flexible Cycle-Accurate Network-on-Chip Simulator" in *ISPASS 2013*.
[4] Becker, D; Jiang, N; Michelogiannakis, G; Dally, W.J.; "Adaptive Backpressure: Efficient Buffer Management for On-Chip Networks" in *ICCD 2012*.
[5] Jiang, N.; Becker, D.; Michelogiannakis, G.; Dally, W.J.; "Network Congestion Avoidance through Speculative Reservation" in *HPCA 2012*.
[6] Michelogiannakis, G.; Jiang, N.; Becker, D.; Dally, W.J.; "Packet Chaining: Efficient Single-Cycle Allocation for On-Chip Networks" in *MICRO 2011*.
[7] Michelogiannakis, G.; Sanchez, D.; Dally, W.J.; Kozyrakis, C.; "Evaluating Bufferless Flow Control for On-chip Networks" in *NOCS 2010*.
[8] Michelogiannakis, G.; Dally, W.J.; "Router Designs for Elastic Buffer On-Chip Networks" in *SC 2009*.
[9] Michelogiannakis, G.; Balfour, J.; Dally, W.J.; "Elastic-Buffer Flow Control for On-Chip Networks" in *HPCA 2009*.
[10] Michelogiannakis, G.; Pnevmatikatos, D.; Katevenis, M.; "Approaching Ideal NoC Latency with Pre-Configured Routes" in *NOCS 2007*.
[11] Michelogiannakis, G.; Dally, W.; "Elastic-buffer Flow Control for On-chip Networks", *IEEE Transactions on Computers*, vol. 62 no. 2, pp. 295-309, February 2013.
[12] Sanchez, D.; Michelogiannakis, G.; Kozyrakis, C.; "An Analysis of On-Chip Interconnection Networks for Large-Scale Chip Multiprocessors", *ACM Transactions on Architecture and Code Optimization*. May 2010.
[13] Borkar, S.; "How to Stop Interconnects from Hindering the Future of Computing" in *OSI 2013*.
[14] Borkar, S.; "Thousand Core Chips: A Technology Perspective" in *DAC 2007*.
[15] Esmaeilzadeh, H.; Blem, E.; St. Amant, R.; Sankaralingam, K.; Burger, D.; "Dark silicon and the end of multicore scaling" in *ISCA 2011*.
[16] Shalf, J.; Dosanjh, S.; Morrison, J.; "Exascale Computing Technology Challenges" in *VECPAR 2010*.
[17] TechNavio Report: "Global Data Center Market 2012-2016".
[18] Greenberg, A.; Hamilton, J.; Matlz, D. A.; Patel, P.; "The Cost of a Cloud: Research Problems in Data Center Networks" in *ACM SIGCOMM Computer Communication Review*, vol. 39 no. 1, pp. 68-73, January 2009.
[19] Rogers B. M.; Krishna, A.; Bell, G. B.; Vu, K.; Jiang, X.; Solihin, Y.; "Scaling the bandwidth wall: challenges in and avenues for CMP scaling," in *ISCA 2009*.
[20] Udipi, A. N.; Muralimanohar, N.; Chatterjee, N.; Balasubramonian, R.; Davis, A.; Jouppi, N. P.; "Rethinking DRAM design and organization for energy-constrained multi-cores" in *ISCA 2010*.
[21] Bachrach, J.; Vo, H.; Richards, B.; Lee, Y.; Waterman, A.; Avizienis, R.; Wawrzynek, J.; Asanovic, K.; "Chisel: Constructing Hardware in a Scala Embedded Language" in *DAC 2012*.
[22] Michelogiannakis, G.; Jiang, N.; Becker, D.; Dally, W.; "Packet Chaining: Efficient Single-Cycle Allocation for On-Chip Networks", *IEEE Computer Architecture Letters*. June 2011.
[23] Michelogiannakis, G.; Becker, D.U.; Dally, W.J.; "Evaluating Elastic Buffer and Wormhole Flow Control", IEEE Transactions on Computers, vol. 60, no. 6, pp. 896-903, June 2011.
[24] Michelogiannakis, G.; Dally, W.; "Elastic-buffer Flow Control for On-chip Networks", *IEEE Transactions on Computers*, vol. 62 no. 2, pp. 295-309, February 2013.
[25] Pakbaznia, E.; Pedram, M.; "Minimizing Data Center Cooling and Server Power Costs", in *ISLPED 2009*.
[26] Michelogiannakis, G.; Shalf, J.; "Variable-Width Datapath for On-Chip Network Static Power Reduction" in NOCS 2014.
[27] Michelogiannakis, G.; Williams, A.; Shalf, J.; "Collective Memory Transfers for Multi-Core Chips" in ICS 2014.
[28] Michelogiannakis, G.; Wilke, J.; Kenny, J.; Ibrahim, K.; Shalf, J.; "Network Bandwidth Tapering Aware Task Placement". *Under review / development.*
[29] Michelogiannakis, G.; Williams, S.; Shalf, J.; "Collective-Aware Prefetcher for Optimizing Memory Bandwidth in Cache-Coherent Chip Multiptocessors". *Under review / development*.
[30] Unat, D.; Chan, C.; Zhang, W.; Michelogiannakis, G.; Bell, J.; Shalf, J.; "TiDA: Tiling as a Durable Abstraction". *International Supermputing Conference (ISC)*, June 2016.
[31] Mohammadiyaghi, P.; Michelogiannakis, G.; Gratz, P.; "Speculative Lock Allocation for Chip Multiprocessors". *Under review / development*.
[32] Fatollahi-Fard, F.; Donofrio, D; Michelogiannakis, G.; Shalf, J.; "OpenSoC Fabric: On-Chip Network Generator". *IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS)*, April 2016.
[33] Joneckis, L.; Koester, D.; Alspector, J.; "An Initial Look at Alternative Computing Technologies for the Intelligence Community". IDA-DARPA report, January 2014.
[34] Michelogiannakis, G.; Shalf, J.; "LLCP: Last Level Cache Prefetching for Data-Parallel Application". *Under review / development.*
[35] Aaronson, S.; "The Limits of Quantum". *Scientific American*, 2008