REPORT FROM ASCAC SUBCOMMITTEE ON FUTURE HIGH PERFORMANCE COMPUTING CAPABILITIES

Vivek Sarkar, Ph.D.
Georgia Institute of Technology
Subcommittee Chair

ASCAC Meeting, December 20, 2017

Subcommittee Members

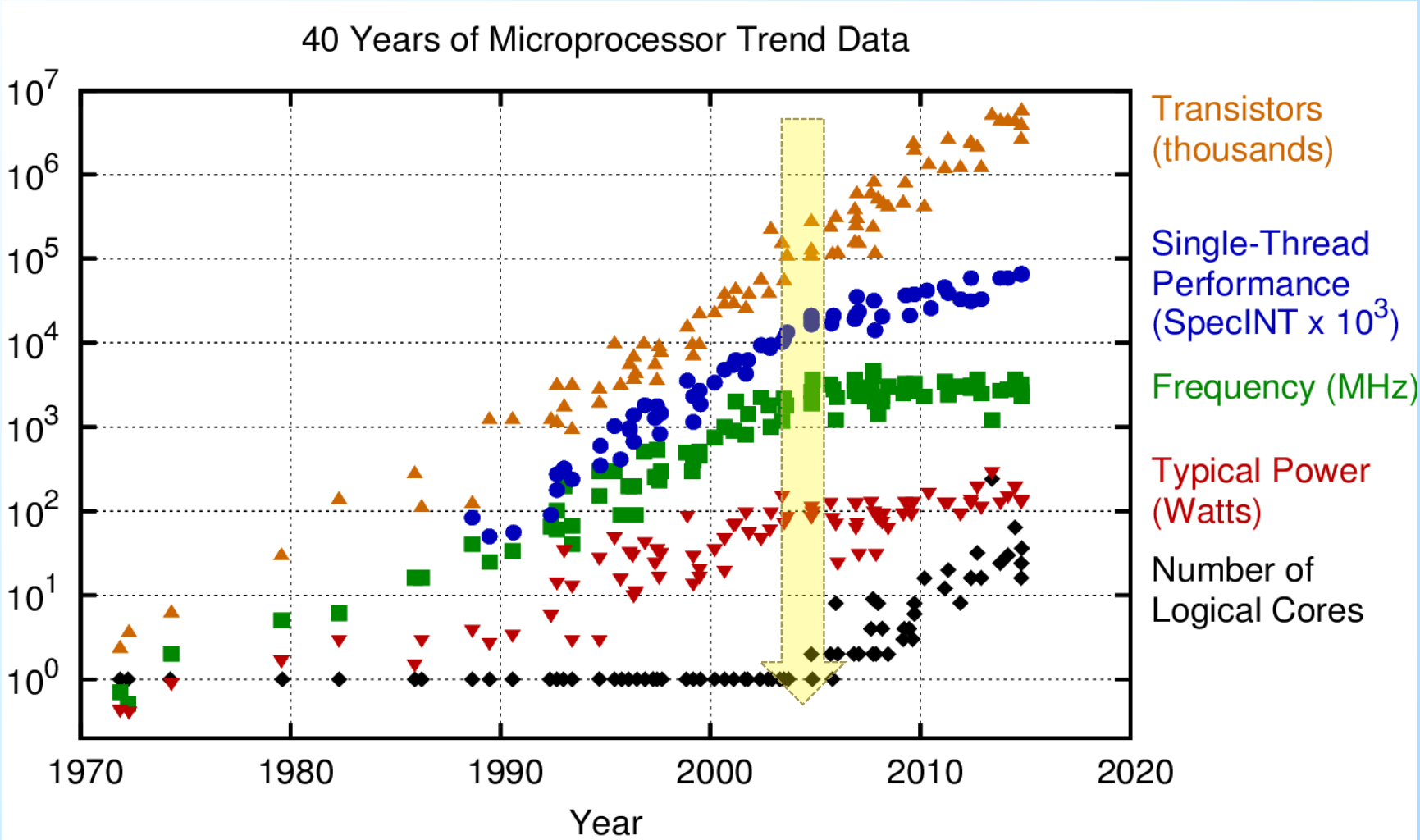
Last name	First name	Affiliation
Bergman[1]	Keren	Columbia U.
Conte	Tom	Georgia Tech
Gara	Al	Intel
Gokhale	Maya	LLNL
Heroux	Mike	Sandia
Kogge	Peter	Notre Dame
Lucas	Bob	ISI
Matsuoka[1]	Satoshi	Tokyo Tech
Sarkar[1,2]	Vivek	Georgia Tech
Temam	Olivier	Google

(1) ASCAC member, (2) Subcommittee chair

Outline

1. Background & Interpretation of Charge
2. Application lessons learned from past HPC Technology Transitions
3. Future HPC Technologies
4. Findings
5. Recommendations

Dennard scaling ended in 2005

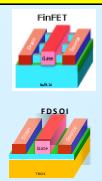
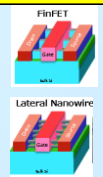
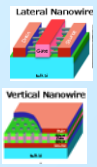
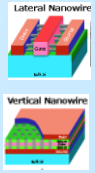
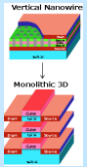
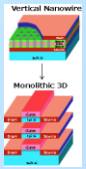
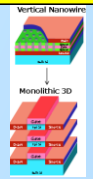


Original data up to the year 2010 collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, and C. Batten
New plot and data collected for 2010-2015 by K. Rupp

End of Moore's Law is approaching

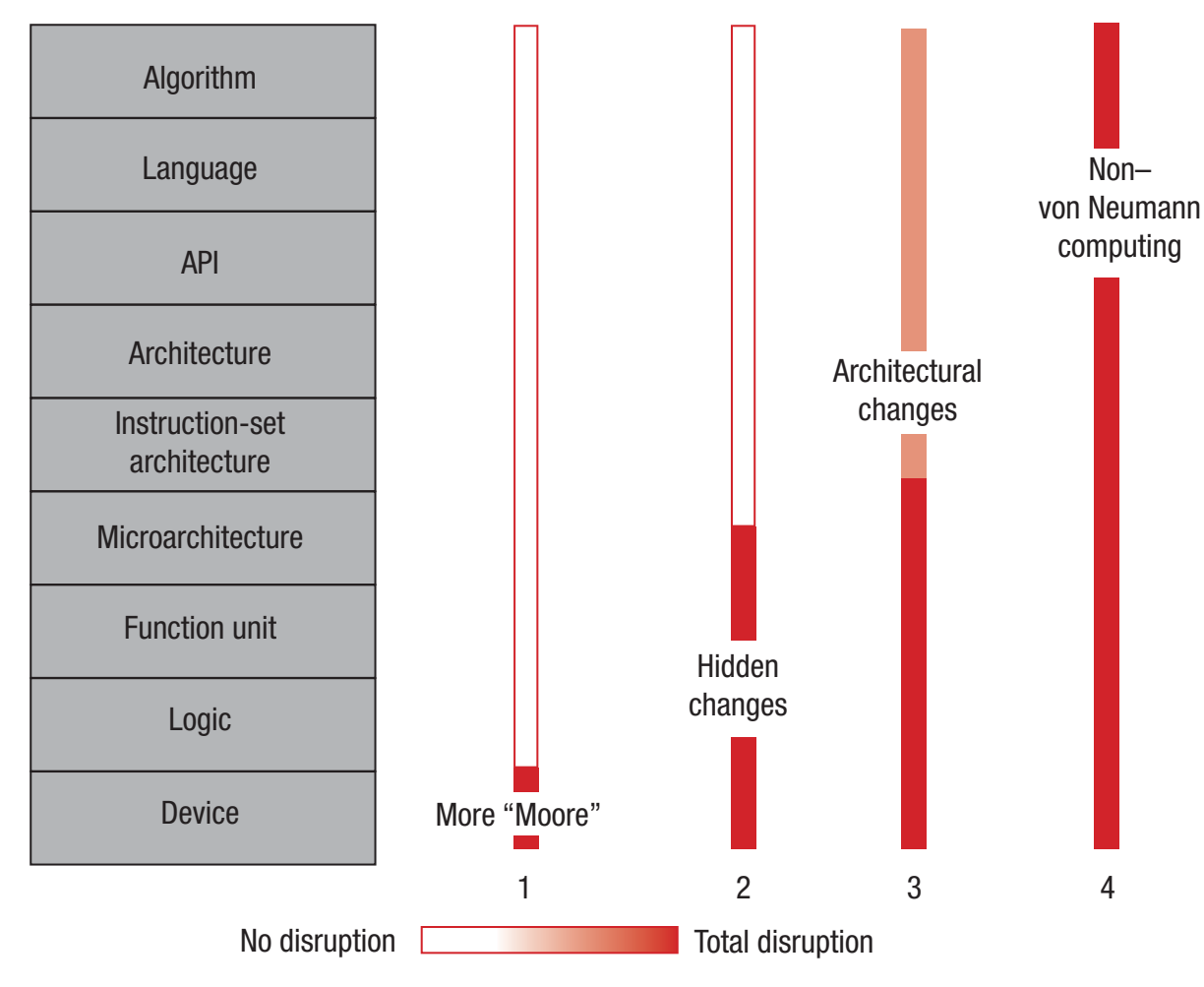
A slow tapering off --- feature sizes will continue to diminish until 1nm in 2033, *with monolithic 3D transistors expected from 2024 onwards*

Table MM01 - More Moore - Logic Core Device Technology Roadmap

YEAR OF PRODUCTION	***2017***	***2019***	***2021***	***2024***	***2027***	***2030***	***2033***
	P54M36	**P48M28**	**P42M24**	**P36M21**	**P28M14G1**	**P26M14G2**	**P24M14G3**
Logic industry "Node Range" Labeling (nm)	**"10"**	**"7"**	**"5"**	**"3"**	**"2.1"**	**"1.5"**	**"1.0"**
IDM-Foundry node labeling	**i10-f7**	**i7-f5**	**i5-f3**	**i3-f2.1**	**i2.1-f1.5**	**i1.5-f1.0**	**i1.0-f0.7**
Logic device structure options	**finFET FDSOI**	**finFET LGAA**	**LGAA VGAA**	**LGAA VGAA**	**VGAA M3D**	**VGAA M3D**	**VGAA M3D**
Logic device mainstream device	**finFET**	**finFET**	**LGAA**	**LGAA**	**VGAA**	**VGAA**	**VGAA**
Logic device technology naming							
Patterning technology inflection for Mx interconnect	**193i**	**193i, EUV**	**193i, EUV**	**193i, EUV**	**193i, EUV**	**193i, EUV**	**193i, EUV**
Channel material technology inflection	**Si**	**SiGe25%**	**SiGe50%**	**Ge, IIIV (TFET)**	**Ge, IIIV (TFET)**	**Ge, IIIV (TFET)**	**Ge, IIIV (TFET)**
Process technogy inflection	**Conformal deposition**	**Conformal Doping, Contact**	**Channel, RMG**	**CFET**	**Seq. 3D**	**Seq. 3D**	**Seq. 3D**
Stacking generation	**2D**	**2D**	**2D 3D: W2W or D2W**	**3D: P-over-N**	**3D: SRAM-on-Logic**	**3D: Logic-on-Logic, Hetero**	**3D: Logic-on-Logic, Hetero**
Design-technology scaling factor for standard cell	**-**	**1.11**	**2.00**	**1.13**	**0.53**	**1.00**	**1.00**
Design-technology scaling factor for SRAM (111) bitcell	**1.00**	**1.00**	**1.00**	**1.00**	**1.25**	**1.00**	**1.00**
Number of stacked devices in one tier	**1**	**1**	**3**	**4**	**1**	**1**	**1**
Tier stacking scaling factor for SoC	**1.00**	**1.00**	**1.00**	**1.00**	**1.80**	**1.80**	**1.80**
Vdd (V)	**0.75**	**0.70**	**0.65**	**0.60**	**0.50**	**0.45**	**0.40**
Physical gate length for HP Logic (nm)	**20.00**	**18.00**	**14.00**	**12.00**	**10.00**	**10.00**	**10.00**
SoC footprint scaling node-to-node - 50% digital, 35% SRAM, 15% analog+IO	**-**	**64.9%**	**51.3%**	**64.3%**	**64.2%**	**50.9%**	**50.7%**

Source: ***IEEE IRDS 2017 Edition***

Levels of Disruption in Post-Exascale and Post-Moore eras



At the far right (level 4) are non−von Neumann architectures, which completely disrupt all stack levels, from device to algorithm.

At the least disruptive end (level 1) are more "Moore" approaches, such as new transistor technology and 3D circuits, which affect only the device and logic levels.

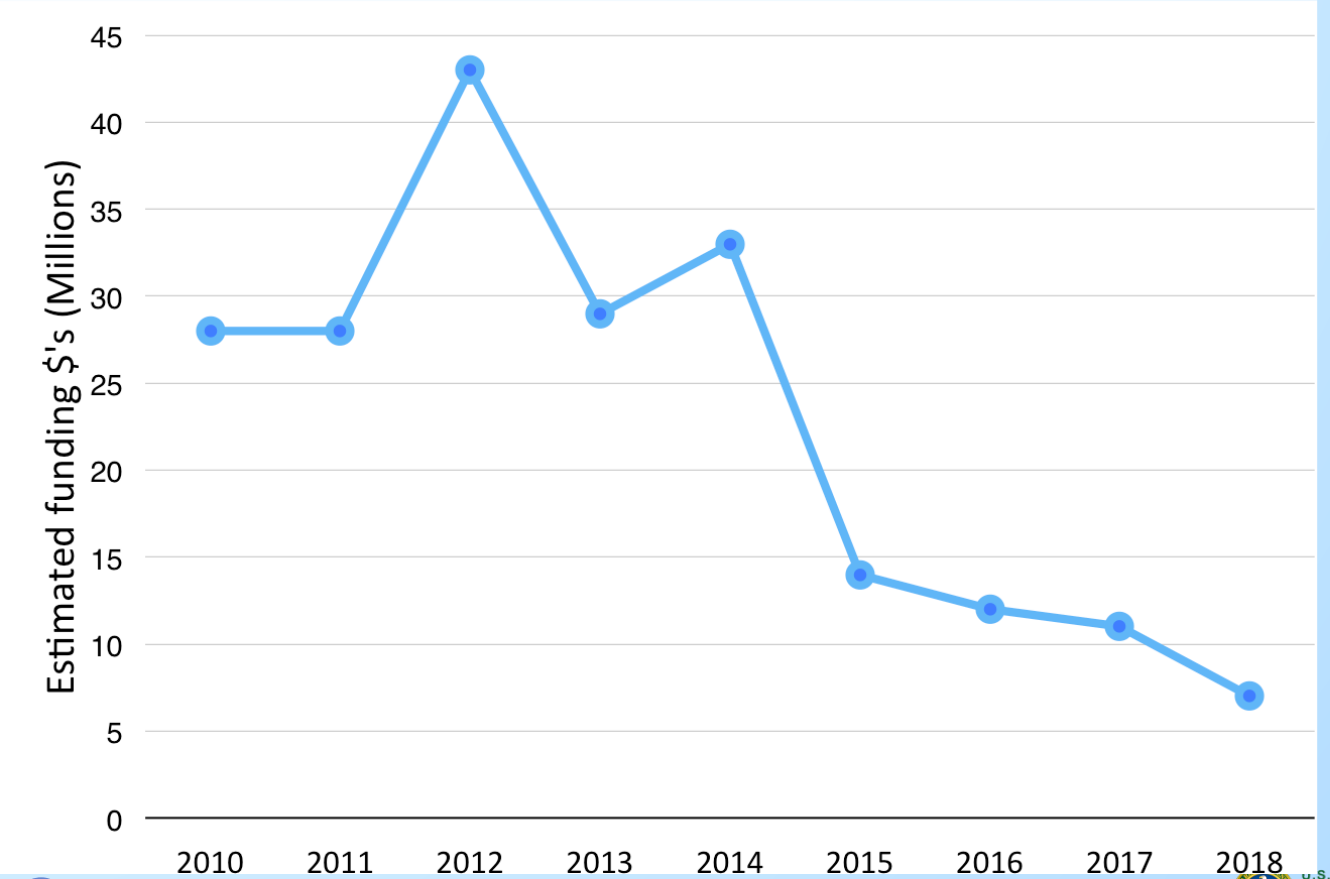
Hidden changes are those of which the programmer is unaware.

Our subcommittee is focusing on level 3 & 4 approaches.

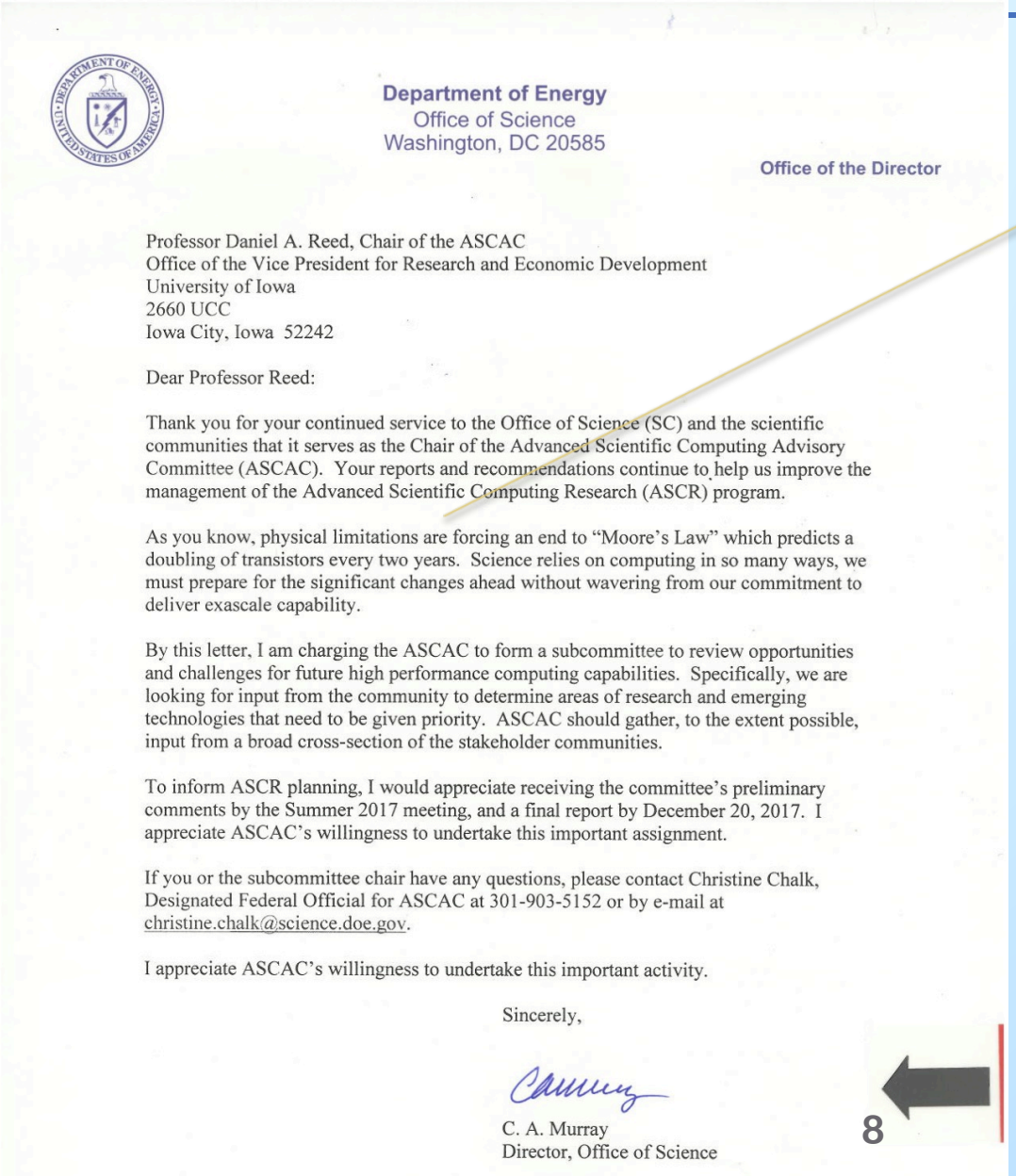
Source: "Rebooting Computing: The Road Ahead", T.M.Conte, E.P.DeBenedictis, P.A.Gargini, E.Track, IEEE Computer, 2017.

Research investments needed to prepare for disruptions, but there have been recent challenges in funding research

CS research programs related to Future Computing (estimates based on target funding $'s in solicitations, source: ASCAC presentation on X-Stack program, Sep'16)



Our Charge



Department of Energy
Office of Science
Washington, DC 20585

Office of the Director

Professor Daniel A. Reed, Chair of the ASCAC
Office of the Vice President for Research and Economic Development
University of Iowa
2660 UCC
Iowa City, Iowa 52242

Dear Professor Reed:

Thank you for your continued service to the Office of Science (SC) and the scientific communities that it serves as the Chair of the Advanced Scientific Computing Advisory Committee (ASCAC). Your reports and recommendations continue to help us improve the management of the Advanced Scientific Computing Research (ASCR) program.

As you know, physical limitations are forcing an end to "Moore's Law" which predicts a doubling of transistors every two years. Science relies on computing in so many ways, we must prepare for the significant changes ahead without wavering from our commitment to deliver exascale capability.

By this letter, I am charging the ASCAC to form a subcommittee to review opportunities and challenges for future high performance computing capabilities. Specifically, we are looking for input from the community to determine areas of research and emerging technologies that need to be given priority. ASCAC should gather, to the extent possible, input from a broad cross-section of the stakeholder communities.

To inform ASCR planning, I would appreciate receiving the committee's preliminary comments by the Summer 2017 meeting, and a final report by December 20, 2017. I appreciate ASCAC's willingness to undertake this important assignment.

If you or the subcommittee chair have any questions, please contact Christine Chalk, Designated Federal Official for ASCAC at 301-903-5152 or by e-mail at christine.chalk@science.doe.gov.

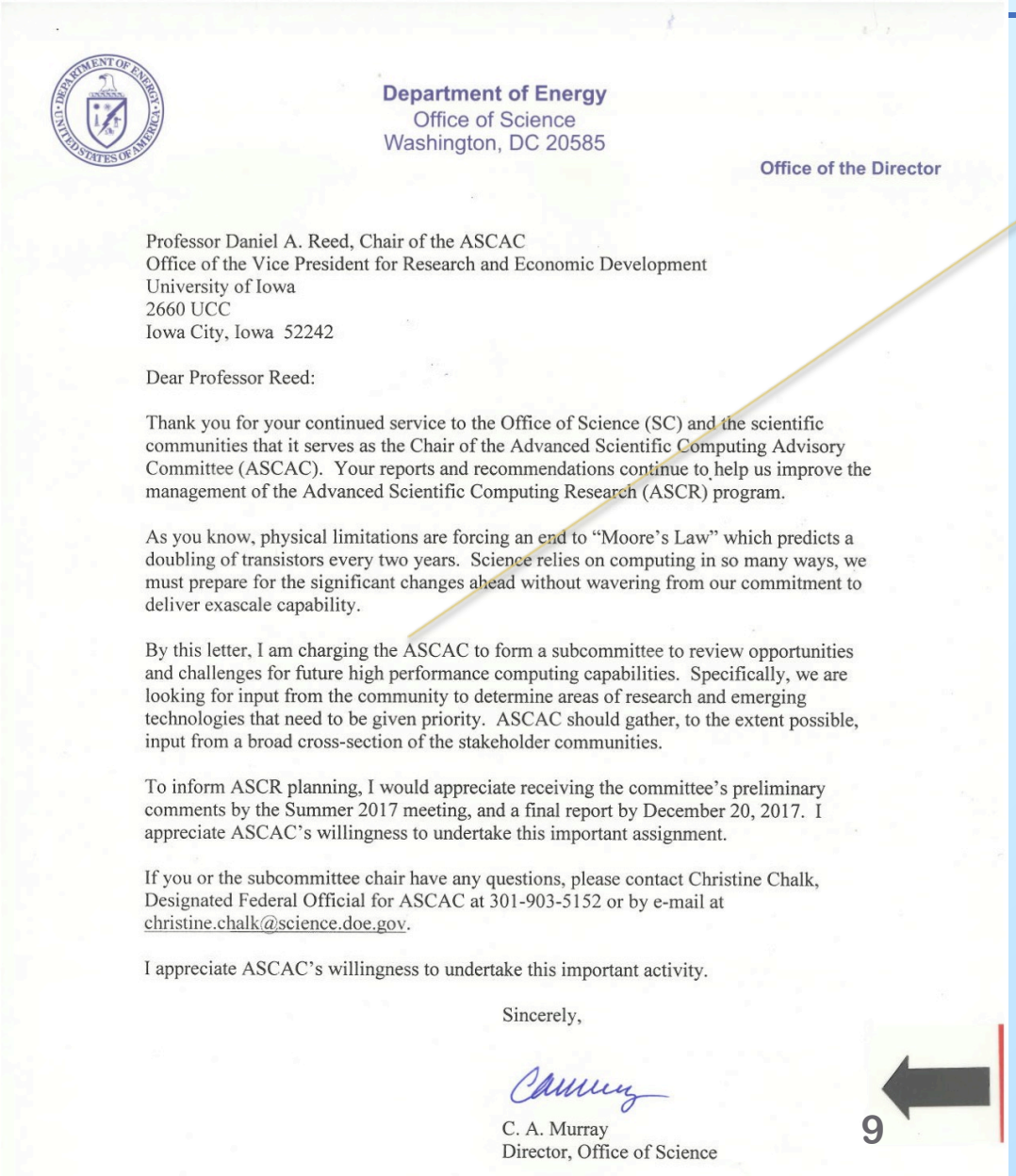
I appreciate ASCAC's willingness to undertake this important activity.

Sincerely,

C. A. Murray
Director, Office of Science

As you know, physical limitations are forcing an end to "Moore's Law" … we must prepare for the significant changes ahead without wavering from our commitment to deliver exascale capability.

Our Charge (contd.)



Department of Energy
Office of Science
Washington, DC 20585

Office of the Director

Professor Daniel A. Reed, Chair of the ASCAC
Office of the Vice President for Research and Economic Development
University of Iowa
2660 UCC
Iowa City, Iowa 52242

Dear Professor Reed:

Thank you for your continued service to the Office of Science (SC) and the scientific communities that it serves as the Chair of the Advanced Scientific Computing Advisory Committee (ASCAC). Your reports and recommendations continue to help us improve the management of the Advanced Scientific Computing Research (ASCR) program.

As you know, physical limitations are forcing an end to "Moore's Law" which predicts a doubling of transistors every two years. Science relies on computing in so many ways, we must prepare for the significant changes ahead without wavering from our commitment to deliver exascale capability.

By this letter, I am charging the ASCAC to form a subcommittee to review opportunities and challenges for future high performance computing capabilities. Specifically, we are looking for input from the community to determine areas of research and emerging technologies that need to be given priority. ASCAC should gather, to the extent possible, input from a broad cross-section of the stakeholder communities.

To inform ASCR planning, I would appreciate receiving the committee's preliminary comments by the Summer 2017 meeting, and a final report by December 20, 2017. I appreciate ASCAC's willingness to undertake this important assignment.

If you or the subcommittee chair have any questions, please contact Christine Chalk, Designated Federal Official for ASCAC at 301-903-5152 or by e-mail at christine.chalk@science.doe.gov.

I appreciate ASCAC's willingness to undertake this important activity.

Sincerely,

C. A. Murray
Director, Office of Science

By this letter, I am charging the ASCAC to form a subcommittee to ***review opportunities and challenges for future high performance computing capabilities.*** *Specifically, we are looking for input from the community to* ***determine areas of research and emerging technologies that need to be given priority.***

Our Charge (contd.)

Department of Energy
Office of Science
Washington, DC 20585

Office of the Director

Professor Daniel A. Reed, Chair of the ASCAC
Office of the Vice President for Research and Economic Development
University of Iowa
2660 UCC
Iowa City, Iowa 52242

Dear Professor Reed:

Thank you for your continued service to the Office of Science (SC) and the scientific communities that it serves as the Chair of the Advanced Scientific Computing Advisory Committee (ASCAC). Your reports and recommendations continue to help us improve the management of the Advanced Scientific Computing Research (ASCR) program.

As you know, physical limitations are forcing an end to "Moore's Law" which predicts a doubling of transistors every two years. Science relies on computing in so many ways, we must prepare for the significant changes ahead without wavering from our commitment to deliver exascale capability.

By this letter, I am charging the ASCAC to form a subcommittee to review opportunities and challenges for future high performance computing capabilities. Specifically, we are looking for input from the community to determine areas of research and emerging technologies that need to be given priority. ASCAC should gather, to the extent possible, input from a broad cross-section of the stakeholder communities.

To inform ASCR planning, I would appreciate receiving the committee's preliminary comments by the Summer 2017 meeting, and a final report by December 20, 2017. I appreciate ASCAC's willingness to undertake this important assignment.

If you or the subcommittee chair have any questions, please contact Christine Chalk, Designated Federal Official for ASCAC at 301-903-5152 or by e-mail at christine.chalk@science.doe.gov.

I appreciate ASCAC's willingness to undertake this important activity.

Sincerely,

C. A. Murray
Director, Office of Science

To inform ASCR planning, I would appreciate receiving the ***committee's preliminary comments by the Summer 2017 meeting****, and a* ***final report by December 20, 2017.***

Interpreting the Charge: Timeframe

- The charge did not specify a timeframe for the subcommittee to focus on ...
- ... however, it is clear that the charge refers to the **post-exascale** (2020's) and **post-Moore** (2030's and beyond) timeframes
- The subcommittee concluded that it was appropriate to focus on *different timeframes for different technologies*, when identifying potential areas of research needed to support the Science mission.

Caveat from subcommittee

"While the subcommittee appreciated the timeliness of the charge, we acknowledge that a single study cannot provide a comprehensive answer to identifying research opportunities and challenges for future HPC capabilities in the post-exascale and post-Moore timeframes, which span multiple decades, and trust that there will be follow-on studies to elaborate further on these challenges and opportunities as details of emerging HPC technologies become clearer in the coming years."

Outline

1. Background & Interpretation of Charge
2. <u>Application lessons learned from past HPC Technology Transitions</u>
3. Future HPC Technologies
4. Findings
5. Recommendations

Vector → Massively Parallel Processing (MPP)

- New computing paradigm with no incremental transition path
- Successful transitions enabled by creating new application frameworks with support for domain decomposition, halo exchanges & global reductions leveraging concepts from prior Applied Math and CS research
- Attrition of vectorization features as focus of on-node performance moved to cache locality
- Challenges in maintaining production vector version while developing new MPP version; development team had to be split across both versions

Terascale → Petascale

- The path to Petascale required attention to intra-node parallelism with OpenMP threading, use of accelerators, and exposing vectorizable code to compilers
- For many applications, this transition was incremental due to reuse of MPP frameworks for inter-node parallelism
- Initial ports of MPP codes were straightforward, but substantial data structure and execution strategy modifications were required to optimize on-node parallelism and locality, leveraging concepts from prior Applied Math and CS research

Petascale → Exascale

- New transition: significant growth in on-node parallelism, locality and heterogeneity, and increasing penalty for any sequential regions of code
- Performance portability becomes a significant challenge; many applications cannot deliver uniformly high performance across different platforms and problem formulations that they are designed to support
- Need for new algorithms, new control layers, new system software support to better handle simultaneous heterogeneous execution, and support task-enabled parallelism, asynchrony, and resilience
- Smaller body of prior research available in support of this transition than in past transitions

Lessons learned

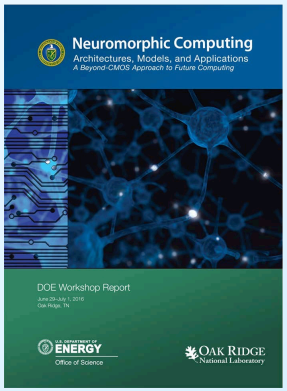
- Vector→MPP: Investing in new application frameworks was critical for success in this transition
- Terascale→Petascale: Leveraging incremental approaches to application migration can be extremely valuable, whenever possible to do so
- Petascale→Exascale: Investing in new control layers and system software support will be helpful for addressing the disruption of large on-node heterogeneous parallelism
- All above transitions were aided by prior research in Applied Math and Computer Science
- Continued opportunities to adopt best practices in software design to reduce application transition costs

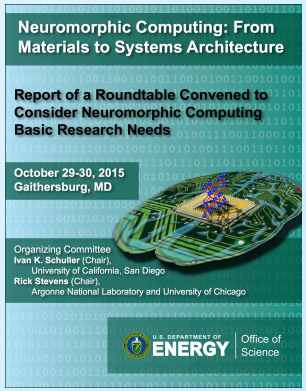
Outline

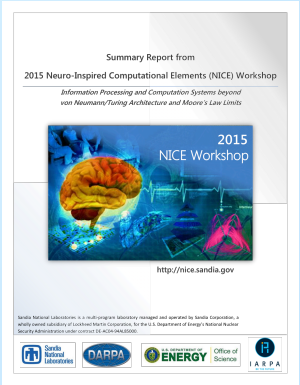
1. Background & Interpretation of Charge
2. Application lessons learned from past HPC Technology Transitions
3. <u>Future HPC Technologies</u>
4. Findings
5. Recommendations

Community investigation of future technologies

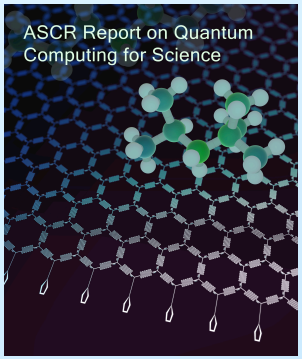
- Several recent DOE workshops and reports have focused on future HPC technologies

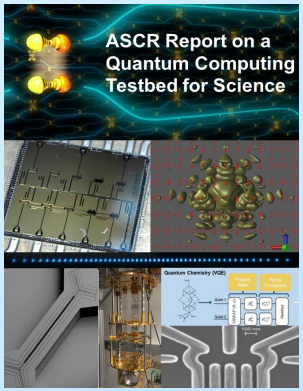






. . .





Future HPC technologies considered by our subcommittee

- Post-Exascale (2020's)
  - Reconfigurable logic
  - Memory-centric processing
  - Silicon photonics
- Post-Moore (2030's)
  - Neuromorphic computing
  - Quantum computing
  - Analog computing
- Common theme: extreme heterogeneity with continued use of digital computing as foundation

Reconfigurable Logic

Approach:

- For best performance, FPGA kernels are written in Hardware Description Languages (HDLs), which requires significant hardware expertise and development effort
- High Level Synthesis (HLS) of C, C++, or OpenCL continues to improve, but, unlike the use of HDL, HLS performance gain is often comparable to that of GPUs

Current & Future Promise:

- Improved energy efficiency & memory bandwidth utilization relative to CPUs/GPUs

Motivating Applications:

- Bioinformatics, signal processing, image processing, network packet processing
- Early adoption in data analysis and in-transit processing areas: use of FPGAs to compress, clean, filter data streams generated by scientific instruments

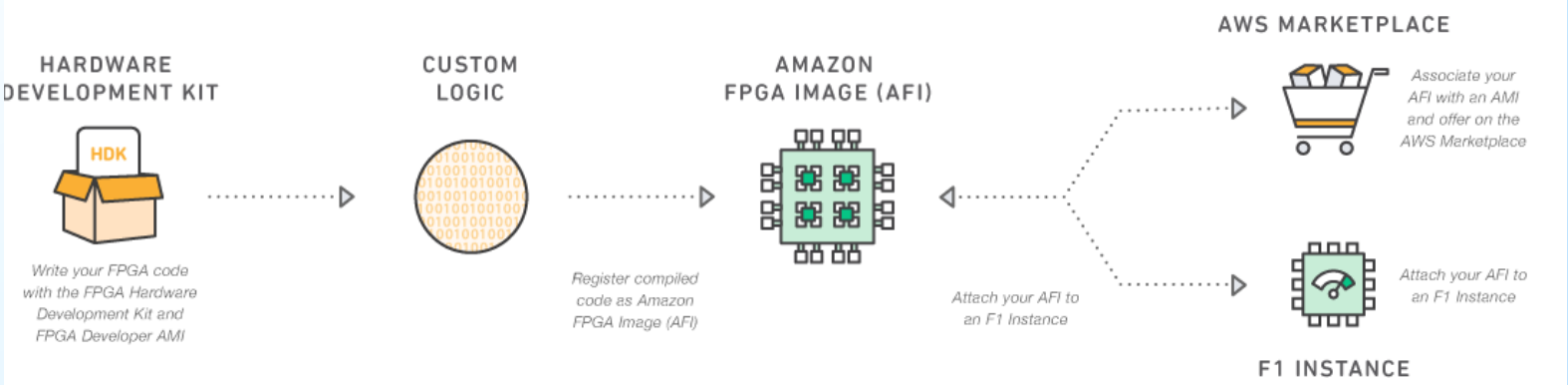
Timeframe:

- FPGA accelerators are already available now (even as cloud services!), and closer integration of CPU with reconfigurable logic is expected in 2-5 years

Research challenges:

- Lack of design tools that simplify application development remains a major obstacle, as does compile cycles (synthesis, map, place, route) that can take hours to days

FPGAs now available as Amazon EC2 F1 instances



DEVELOP

Develop custom Amazon FPGA Images (AFI) using the Hardware Development Kit (HDK) and full set of design tools and simulators.

DEPLOY

Deploy your AFI directly on F1 instances and take advantage of all the scalability, agility, and security benefits of EC2.

OFFER

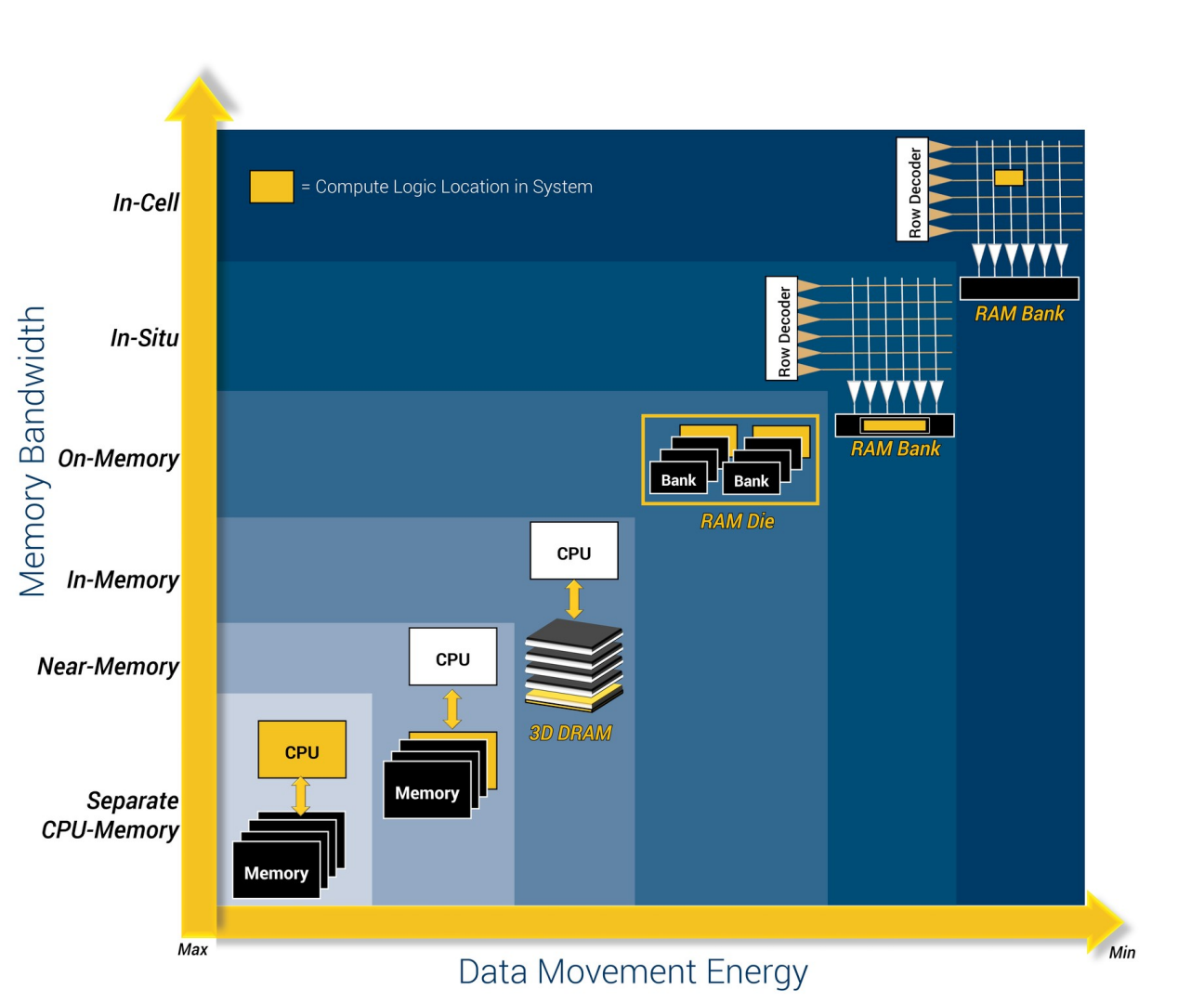
Offer AFIs you design on the AWS Marketplace for other customers.

PURCHASE

Purchase AFIs built and listed on AWS Marketplace to quickly implement common hardware accelerations.

Source: https://aws.amazon.com/ec2/instance-types/f1/

Range of Approaches for Memory-Centric Processing



Memory-Centric Processing

Approach:

- Memory-Centric Processing places computation closer to memory than conventional cores. These approaches are being explored at the *in situ*, *sense amps*, *memory bank*, *on-memory*, and *near-memory* levels.

Current & Future Promise:

- Reduce memory bandwidth bottlenecks by performing lightweight specialized operations close to memory. Additional benefits include reduced latency, reduced energy of transport, faster atomic operations, and higher levels of concurrency.

Motivating applications:

- Applications with memory–centric streaming operations, e.g., encryption/decryption, search, big data, big graphs, deep learning

Timeframe:

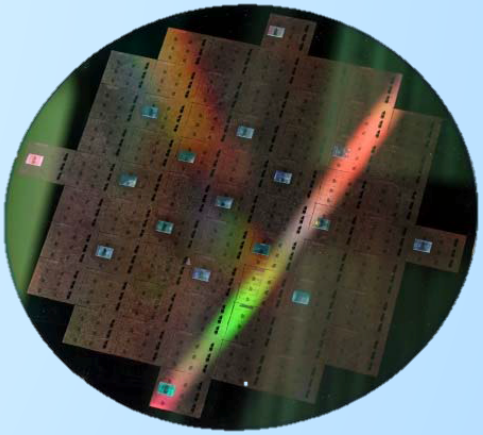
- Above approaches demonstrated at the research level. Near-Memory Processing appears to be the most viable for the next level, due to its synergy with 3D stacking.

Research challenges:

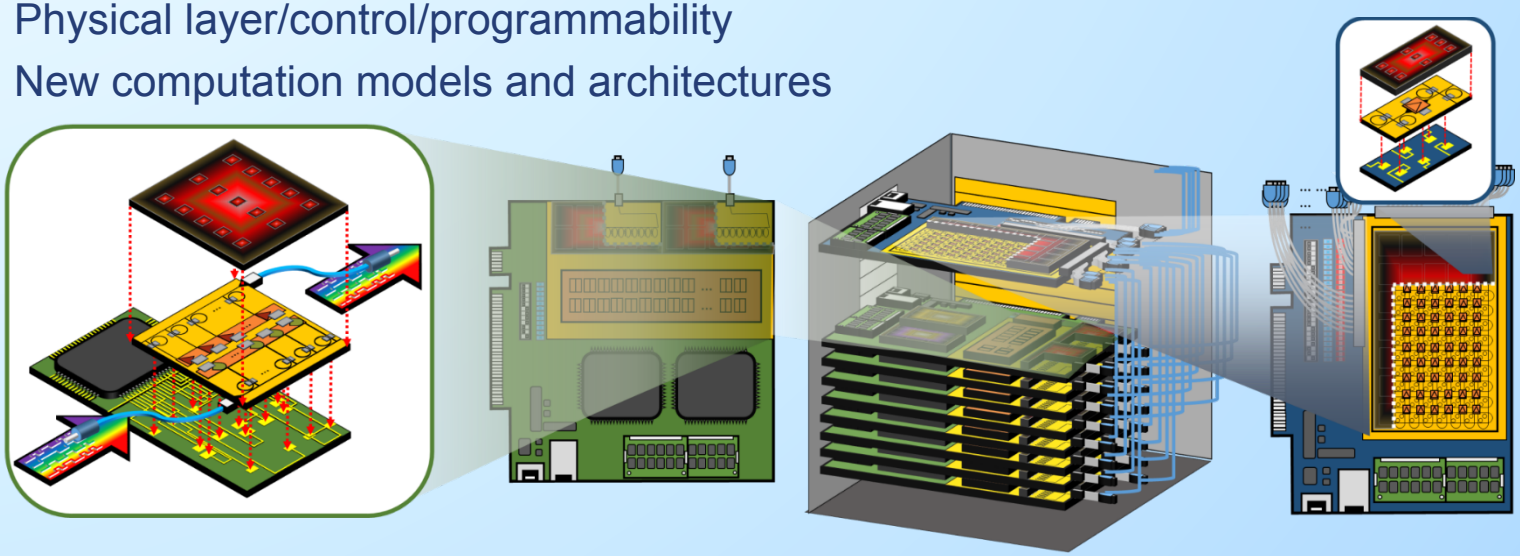
- How to maintain some level of coherence/consistency across data copies, how to support remote computations and a global address space, how to recognize completion of asynchronous operations, how to handle cases where data from separate memories need to be combined.

Silicon Photonics

- Silicon Photonics has emerged as platform for large scale integration of complex electronic-photonic ICs
- Enabling system scale CMOS-photonics
- AIM Photonics - Integrated Photonics Manufacturing Institute – state-of-art US facility (Albany) with 300mm tools for fabrication, 3D stacking with CMOS
- Research challenges:
  - Bridging photonics with computing systems
  - Physical layer/control/programmability
  - New computation models and architectures



300mm SiP wafer



Example future direction for Photonics: Optical Neural Networks

nature photonics

ARTICLES

PUBLISHED ONLINE: 12 JUNE 2017 | DOI: 10.1038/NPHOTON.2017.93

Deep learning with coherent nanophotonic circuits

Yichen Shen[1*†], Nicholas C. Harris[1*†], Scott Skirlo[1], Mihika Prabhu[1], Tom Baehr-Jones[2], Michael Hochberg[2], Xin Sun[3], Shijie Zhao[4], Hugo Larochelle[5], Dirk Englund[1] and Marin Soljačić[1]

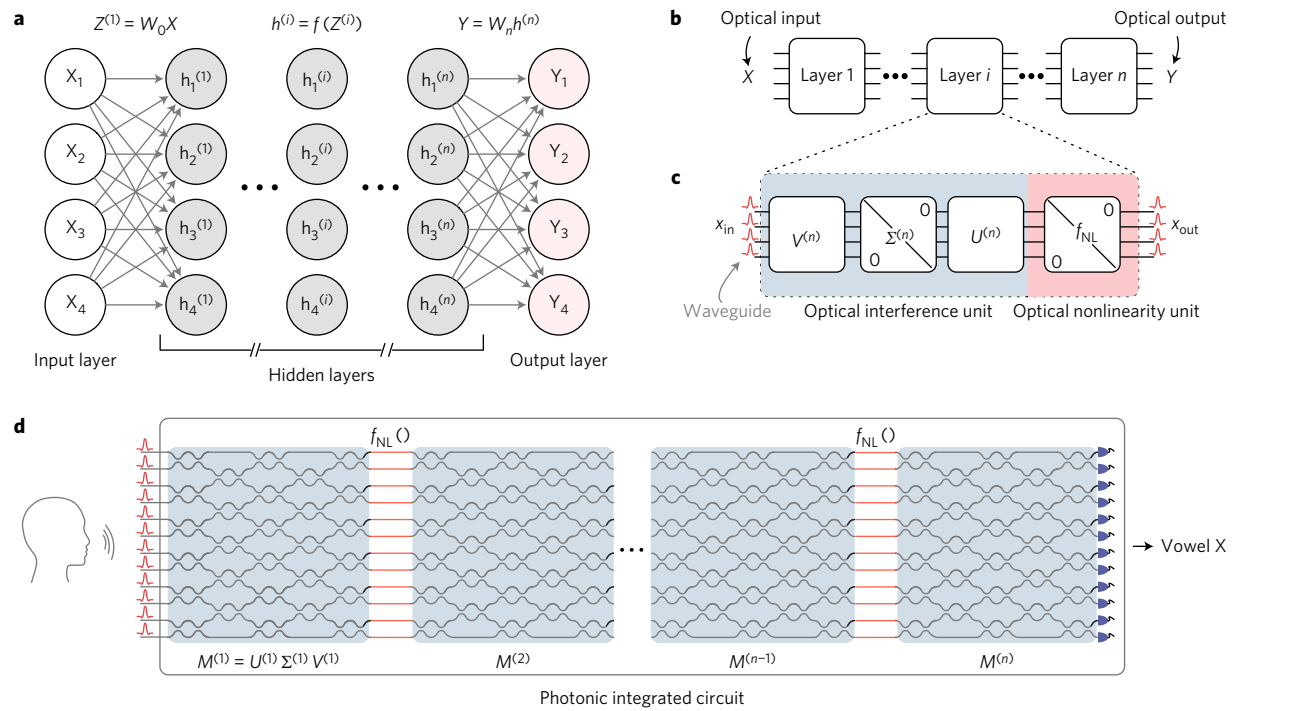


Figure 1 | General architecture of the ONN. a, General artificial neural network architecture composed of an input layer, a number of hidden layers and an output layer. **b**, Decomposition of the general neural network into individual layers. **c**, Optical interference and nonlinearity units that compose each layer of the artificial neural network. **d**, Proposal for an all-optical, fully integrated neural network.

Neuromorphic Computing

Approach:

- Emulate the behavior of a subset of the brain, e.g., via algorithms that simulate spiking neurons and can be used as modeling tools by neuroscientists
- Use artificial neural networks to achieve brain-like functionality, such as object or speech recognition e.g., via deep neural networks.

Current & future promise:

- Initial excitement in the 1950s with the Perceptron, followed by Multi-Layer Perceptrons in the 1980s/1990s. However, these were outperformed by running algorithms such as Support Vector Machines (SVMs) on stock hardware from those periods.
- Current hardware (notably GPUs) has made it possible for Deep Neural Networks to achieve human-level performance for non-trivial tasks such as object recognition & speech recognition.

Motivating applications:

- Modeling tools for neuroscientists, deep learning for science, numerous commercial applications

Timeframe:

- Current implementations include Google's TPUs and IBM's True North hardware, as well as efficient implementations of DNNs in GPUs and FPGAs
- Many companies are expected to propose and develop ASICs with efficient support for neuromorphic computing for use in data centers and embedded platforms (e.g., self-driving cars).

Research challenges:

- Modeling the human brain, expand use of neuromorphic computing in new applications

Neuromorphic Computing is already receiving a lot of attention in DOE activities

von Neumann Architecture | **Neuromorphic Architecture**

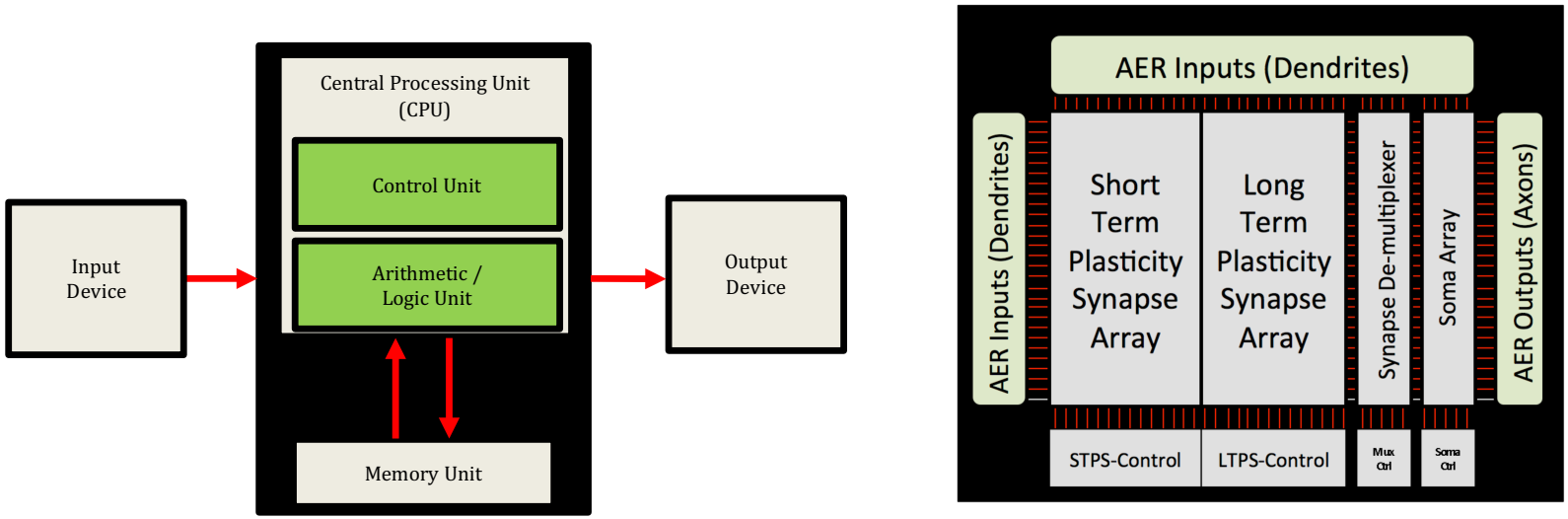


Figure 1. Comparison of high-level conventional and neuromorphic computer architectures. The so-called "von Neumann bottleneck" is the data path between the CPU and the memory unit. In contrast, a neural network based architecture combines synapses and neurons into a fine grain distributed structure that scales both memory (synapse) and compute (soma) elements as the systems increase in scale and capability, thus avoiding the bottleneck between computing and memory.

Figure source: "Report of a Roundtable Convened to Consider Neuromorphic Computing Basic Research Needs", October 2015, Gaithersburg, MD

Quantum Computing is also receiving a lot of attention in DOE activities

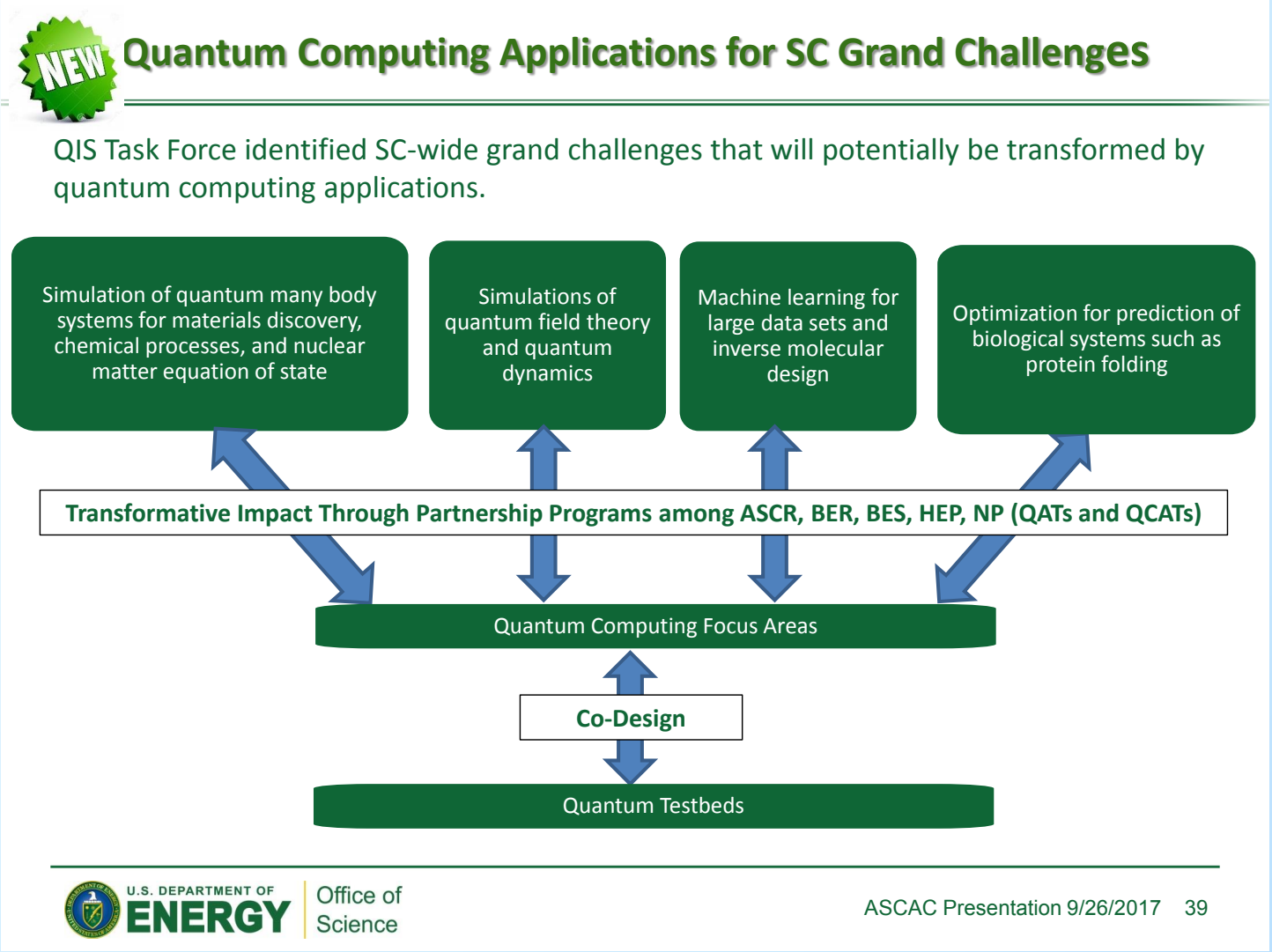


Figure source: presentation on "Advanced Scientific Computing Research", Barbara Helland, ASCAC meeting, Sep 2017. Also included updates on "Quantum Algorithm Teams (QATs)" and "Quantum Testbed Pathfinder" programs.

Quantum Computing

Approach:

- Exploit quantum-mechanical nature of specific physical phenomena to provide advantages relative to classical computing. Whereas N digital bits encode one N-bit state, N entangled quantum bits (qubits) can encode 2^N possible N-bit states states upon which operations can be simultaneously applied.

Current & future promise:

- Theoretical quantum algorithms have been discovered for multiple scientific problems of interest to DOE. These range from problems in chemistry and physics, to data analysis and machine learning, and to fundamental mathematical operations. However, without the existence of suitable quantum computers, they cannot yet be exploited to accelerate time to scientific discovery.
- Prototypes of small quantum systems, be they specialized annealing devices, or even general purpose computers, are beginning to appear (D-Wave, IBM, etc.).

Motivating applications:

- Quantum computing was originally conceived of as a way to use quantum mechanical phenomenon to solve problems in modeling other quantum mechanical properties of materials. The range of potential applications for which quantum computing offers advantages relative to classical computing has since expanded, including factoring composite integers (Shor), search (Grover), and optimization (quantum annealing).

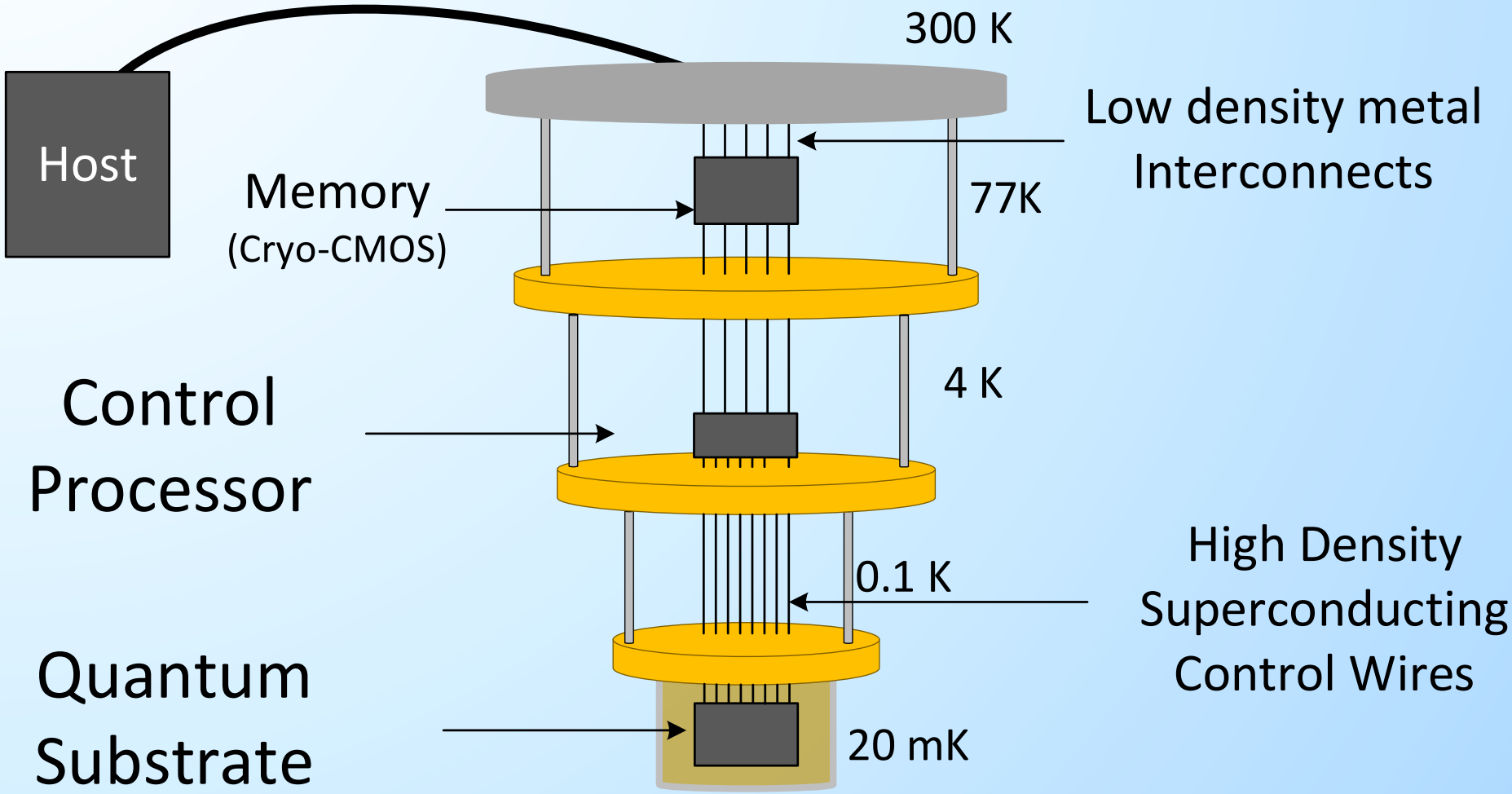
Timeframe:

- Quantum computing today is still itself an object of research, and not yet a tool that is ready to be applied for broader scientific discovery. Since the advent of Shor's algorithm, there has been substantial investment in quantum computing worldwide, first by governments, and more recently, commercial interests.

Research challenges:

- Development of quantum computing at larger scales where they will offer true computational advantage relative to classical machines.
- Development of programming approaches to make use of quantum computing more broadly accessible.

Integrating Quantum Computing with Digital host/control processors



Thermal hierarchy for host and control processors connected to a quantum substrate

Analog Computing

Approach:

- Mapping dynamical systems to analogous systems, where the latter is typically electronic, optical or electro-chemical systems.
- Exploit dynamical systems that have similar physics relationships to the system being simulated/modeled.

Current & future promise:

- Improved computational efficiency vs. traditional digital simulation/search. In some cases, orders of magnitude lower power than digital approaches.

Motivating applications:

- Physical system simulation, solving differential equations, near-optimal search (annealing).

Timeframe:

- Analog computing has a long history, but the success of digital computing has pushed it to the sidelines. New investments coupled with device/dynamical-process modeling has strong potential in a 10 year timeframe.

Research challenges:

- Increased bit precision of computation as a function of SNR, algorithm design for limited precision, software foundations for hybrid digital-analog computing

Common themes: extreme heterogeneity, specialization, hybrid digital-analog systems

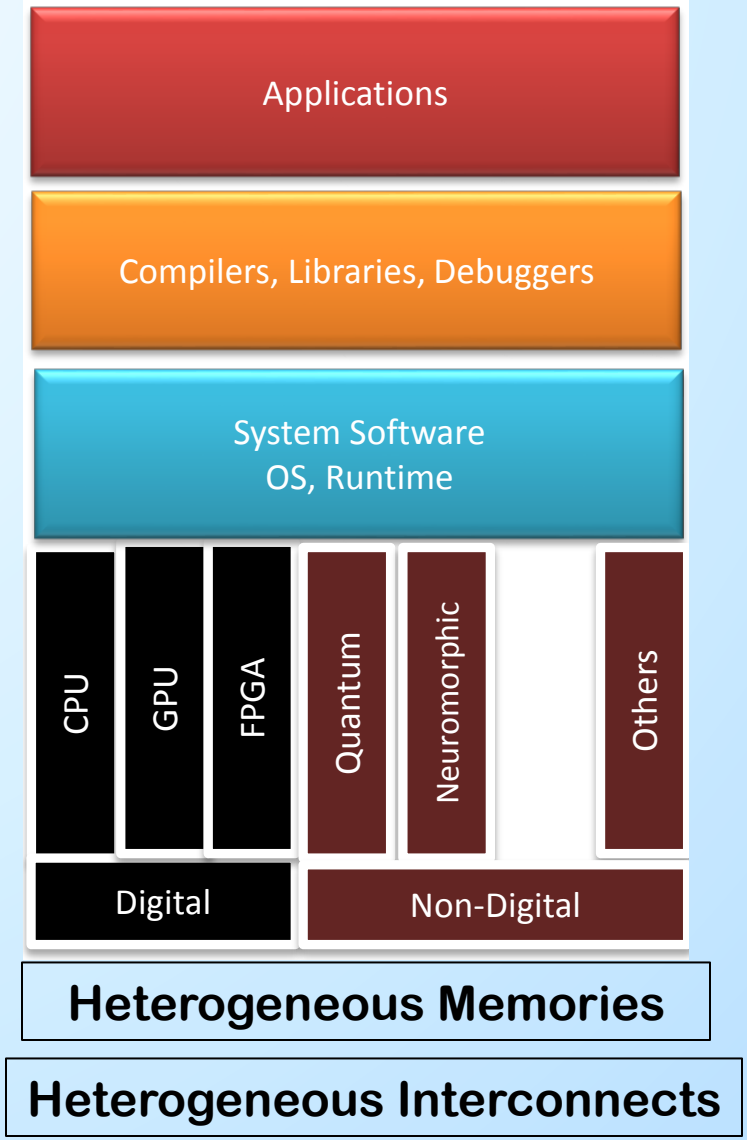


Figure source: presentation on "Advanced Scientific Computing Research", Barbara Helland, ASCAC meeting, Sep 2017.

Outline

1. Background & Interpretation of Charge
2. Application lessons learned from past HPC Technology Transitions
3. Future HPC Technologies
4. <u>Findings</u>
5. Recommendations

Findings

Finding 1: Need for clarity in future HPC roadmap → Science will need to prepare for a period of uncertainly in future HPC technologies and computing paradigms

- Significant attention on post-Moore computing from multiple agencies, but lack of clarity as to what the future HPC roadmap should be for Science
- Science will need to prepare for a period of uncertainty in future HPC technologies and computing paradigms, which is likely to be more disruptive than the Vector→MPP transition
- Due to this uncertainty, there is a need to adopt agile strategy and planning processes so as to better adapt to future HPC technology transitions

Findings (contd)

Finding 2: Extreme heterogeneity with new computing paradigms will be a common theme in future HPC technologies

- There is a great diversity in the technologies that are expected in the post-exascale and post-Moore eras, appropriately termed “extreme heterogeneity” in an upcoming ASCR workshop and related discussions
- Value in focusing on extreme heterogeneity with digital computing foundations as a common theme in future HPC technologies
- Within this theme, there are compelling research challenges in moving point solutions forward (e.g., neuromorphic computing, quantum computing) so that they can be integrated in future platforms with extreme heterogeneity

Findings (contd)

Finding 3: Need to prepare applications and system software for extreme heterogeneity

- We are rapidly approaching a period of significant redesign and reimplementation of applications that is expected to surpass the Vector→MPP transition
- Scientific teams will need to prepare for a phase when they are both using their old codes to obtain science results while also developing new application frameworks based on the new applied math and computer science research.

Findings (contd)

Finding 4: Need for early testbeds for future HPC technologies

- There is a need for building and supporting early testbeds for future HPC technologies that are broadly accessible to the DOE community, so as to enable exploration of these technologies through new implementations of science applications (proxy and full)
- There are multiple instances of individual research groups at DOE laboratories creating early testbeds, but administration of testbeds by research groups is necessarily ad hoc and lacks the support for broad accessibility that is provided by DOE computing facilities

Findings (contd)

Finding 5: Open hardware promises to be a major trend in future platforms

- With extreme heterogeneity, there is a growing trend towards building hardware with *open interfaces* so as to integrate components from different hardware providers
- There is also a growing interest in building *open source* hardware components through recent movements such as the RISC-V foundation
- For the purpose of this report, the term “open hardware” encompasses both open interfaces for proprietary components as well as open source hardware

Findings (contd)

Finding 6: Synergies between HPC and mainstream computing

- There are notable synergies between future HPC and mainstream computing requirements, e.g., there is already a growing commercial use of reconfigurable logic in mainstream platforms
- In addition, synergies will be leveraged in the area of data-intensive applications and data analytics. e.g., use of neuromorphic computing and accelerators for deep learning
- As observed in a past ASCAC study, there are also notable synergies between the data-intensive computing and high-performance computing capabilities needed for science applications

Outline

1. Background & Interpretation of Charge
2. Application lessons learned from past HPC Technology Transitions
3. Future HPC Technologies
4. Findings
5. <u>Recommendations</u>

Recommendations

Recommendation 1: The DOE Office of Science should play a leadership role in developing a post-exascale and post-Moore strategy/roadmap/plan, at both the national and international levels, for HPC as a continued enabler for advancing Science.

- Focus on the needs of science applications (some may be synergistic with vendor priorities, and some may not)
- Raise public awareness of upcoming post-Moore challenges (as we did for exascale)
- Engagement with existing technology roadmap efforts (e.g., IRDS) can play a key role in defining DOE's HPC roadmap
- International competitiveness dictates that DOE Office of Science continue its focus on ensuring USA's continued worldwide leadership in high performance computing.

Recommendations (contd)

Recommendation 2: DOE should invest in preparing for readiness of science applications for new computing paradigms in the post-exascale and post-Moore eras

- In partnership with other science programs (as in SciDAC), to ensure that sufficient investment is made with adequate lead time to prepare science applications for the post-exascale and post-Moore eras
- With clear methodology for making migration vs. rewrite decisions for different applications in different timeframes, as new technologies become ready for production use
- While balancing the criticality of both delivering exascale capability and exploring new computing paradigms for the future.
- Including investment in applied math and algorithms research (e.g., exploring new models of computer arithmetic) that is tightly coupled with application development for new computation and data models

Recommendations (contd)

Recommendation 3: DOE should invest in research to help foster an open hardware ecosystem as part of the future HPC technology roadmap

- Future hardware will require more innovation and agility in hardware design than in past decades, and an open platform approach will help foster this innovation while also mitigating risks associated with selecting a single vendor for hardware acquisition.
- Trend towards extreme heterogeneity in post-exascale and post-Moore computing reinforces the importance of integrating hardware components developed by different hardware providers.
- Research investment is necessary new approaches are needed to ensure that leadership-class HPC hardware can be built for future science applications by tightly integrating the best technologies from different hardware providers (proprietary or open source).

Recommendations (contd)

Recommendation 4: DOE should invest in research to advance system software technologies for post-exascale and post-Moore computing

- Past DOE investments have helped ensure a successful history of using advances in system software to reduce time and cost for developing and deploying production applications on leadership HPC systems
- Current system software stack is built on technology foundations that are more than two decades old, and are ill-prepared for new computing paradigms anticipated in post-exascale and post-Moore computing
- Combination of open hardware research and system software research will enable software/hardware co-design to occur with the agility needed for post-exascale and post-Moore computing
- System software has a long history of reducing the impact of hardware disruptions on application software, and this role will be even more important in the future

Recommendations (contd)

Recommendation 5: DOE computing facilities should prepare users for post-Moore computing by providing and supporting early access to testbeds and small-scale systems

- Includes acquiring testbeds and small-scale systems that are exemplars of future HPC systems, and investing in personnel who are qualified to provide support and training
- Will require building relationships with new hardware providers who are exploring new post-Moore technologies
- Will need to extend beyond system support, and also include training, workshops, and fostering of user groups for different systems.
- Without distracting from exascale commitments!

Recommendations (contd)

Recommendation 6: DOE labs should recruit and grow workforce members who can innovate in all aspects of mapping applications onto emerging post-exascale and post-Moore hardware

- Recruiting and retention challenges in computing-related areas have been documented in past studies
- New opportunities to recruit talent who are passionate about working with cutting-edge technologies
- Prioritization of future HPC in all avenues related to recruiting, growth and retention of top talent, including CSGF fellowships, postdoctoral appointments, LDRD-funded projects, awards, and other forms of recognition
- Engage with interested and qualified faculty in academia through sabbaticals and other channels

Leadership beyond exascale

- While DOE's commitment to deliver exascale capabilities is of paramount importance, we believe that it is essential for DOE ASCR to fund research and development that looks beyond the Exascale Computing Project (ECP) time horizon
- ECP focus has dampened recent efforts to explore new paradigms for post-exascale and post-Moore computing, and this dampening is in danger of intensifying due to reductions in the ECP delivery timeline
- Balancing the criticality of delivering production applications with research that explores new computing paradigms has been a successful strategy for past technology transitions (e.g., Vector → MPP); continuing such a strategy for post-exascale and post-Moore computing will ensure our nation's continued leadership in future HPC

Summary

- Wide range of technologies for future high performance computing capabilities in different timeframes.
- Extreme heterogeneity with digital computing foundations will be a common theme in future HPC
- There has been a loss in momentum in funding and sustaining a research pipeline in the applied math and computer science areas for future HPC, which should be corrected as soon as possible
- Applications will need to be agile in evaluating and adopting technologies that are most promising for their domain, as well as in making "migrate vs. rewrite" decisions
- Office of Science can play a leadership role in developing a post-exascale and post-Moore roadmap for Science on HPC, without distracting from exascale commitments