U.S. DEPARTMENT OF ENERGY | Office of Science

ASCAC SUBCOMMITTEE ON FUTURE HIGH PERFORMANCE COMPUTING CAPABILITIES: PRELIMINARY COMMENTS

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

ASCAC Meeting, Arlington, VA, September 27, 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

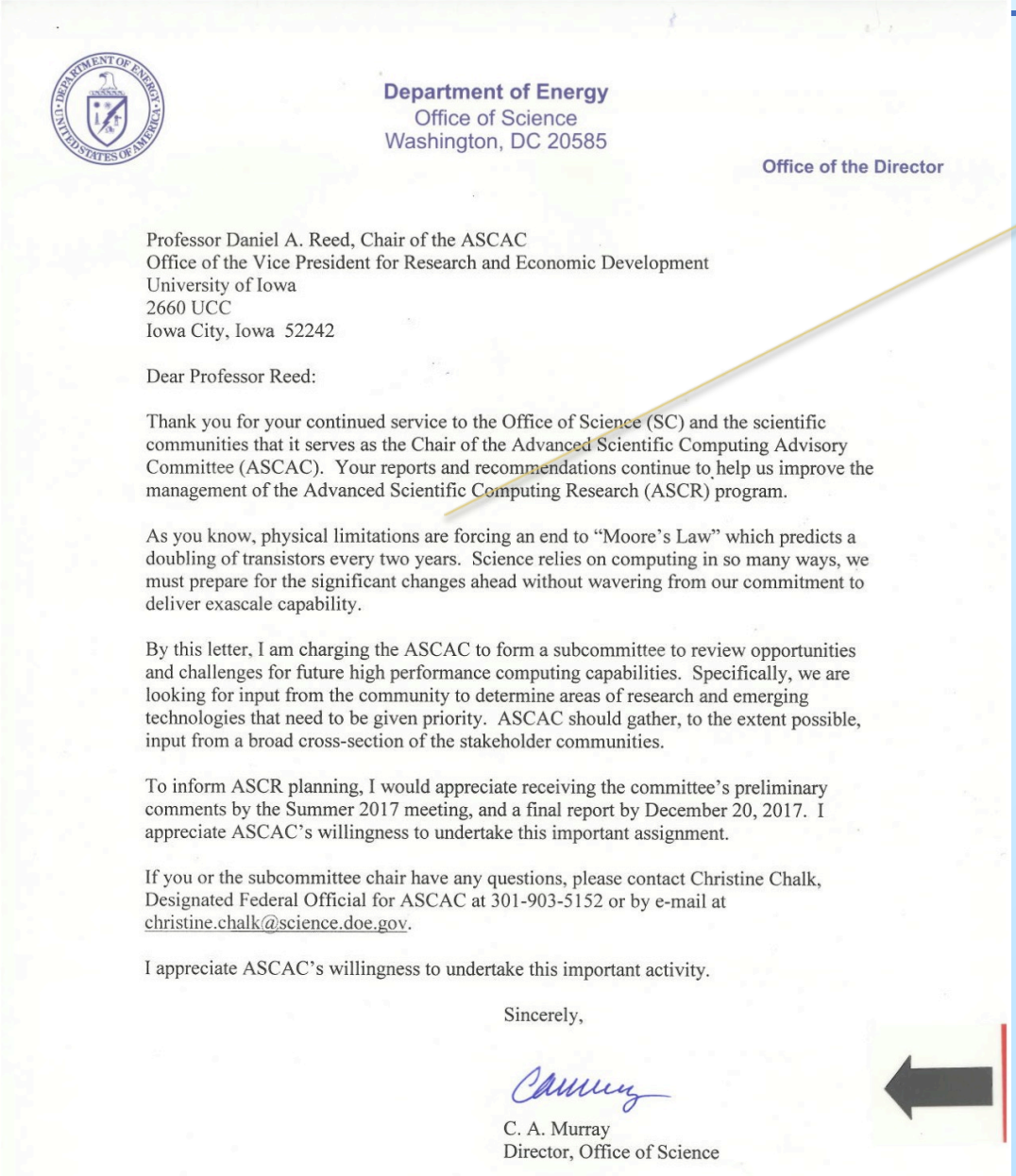
Outline

1. **Our charge**

2. Post-Moore opportunities and challenges in Office of Science's mission

3. Preliminary Recommendations

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.

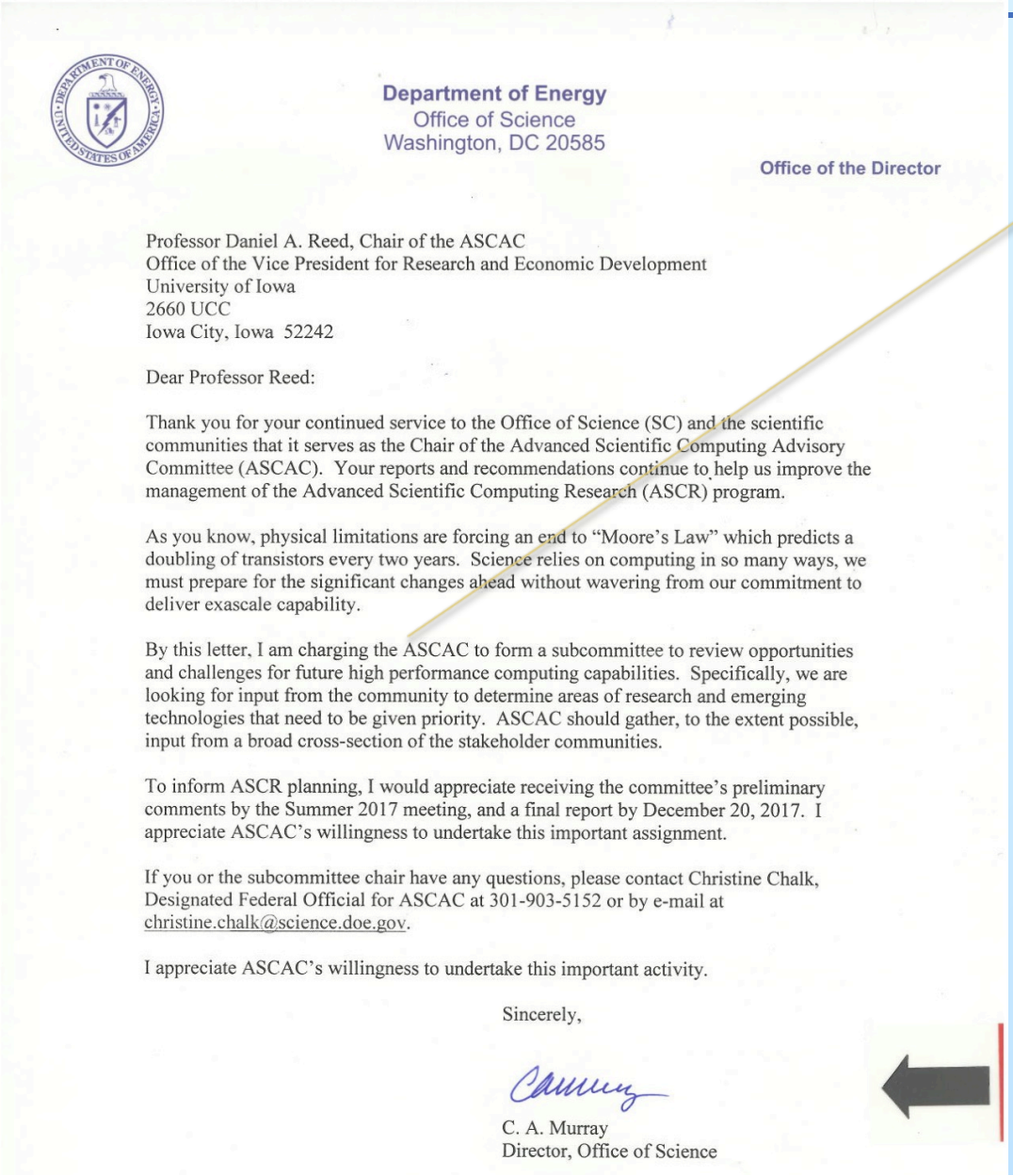
Sincerely,

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.

U.S. DEPARTMENT OF ENERGY | Office of Science

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 a **post-exascale timeframe**.
- 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.

Methodology

- Findings related to future HPC technologies
  - Identify potential technology areas for future HPC systems
  - Identify synergistic community activities in these technology areas (workshops, studies, white papers)
  - Estimate timeframes for different levels of technology readiness for these technologies
  - Create a framework for assessing the ability of applications to exploit different technologies
- Regular conference calls among subcommittee members to discuss findings and potential recommendations
  - Conducted 13 conference calls thus far
- Preparation of preliminary comments (this presentation)
- Discussion with external experts beyond subcommittee
- Preparation of final report

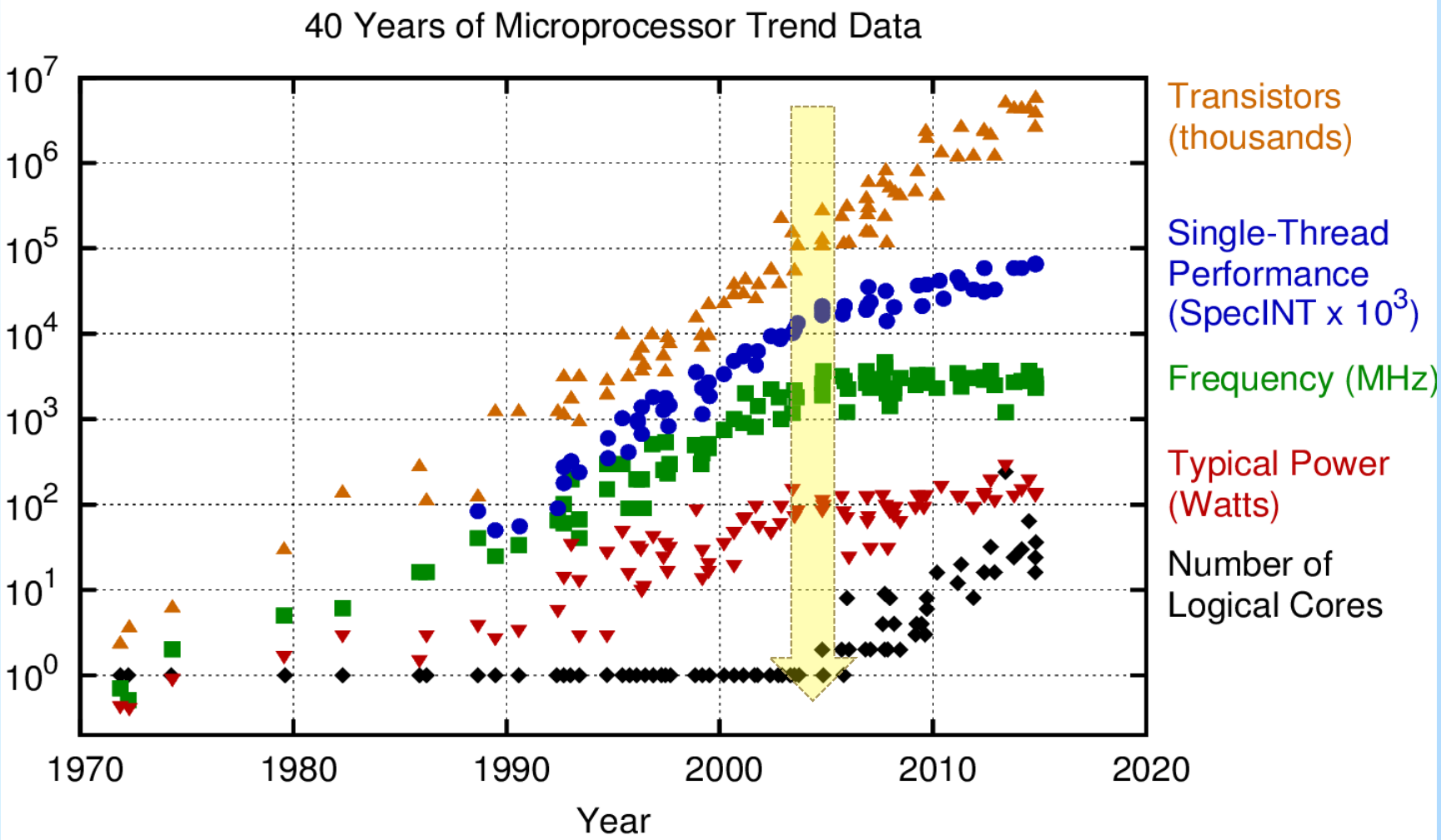
Outline

1. Our charge

2. **Post-Moore opportunities and challenges in Office of Science's mission**

3. Preliminary Recommendations

Dennard scaling ended in 2005

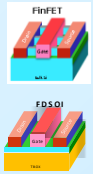
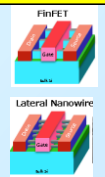
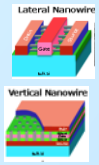
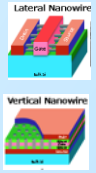
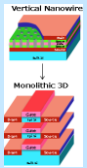
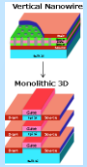
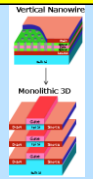


Original data up to the year 2010 collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, and C. Batten
New plot and data collected for 2010-2015 by K. Rupp

End of Moore's Law

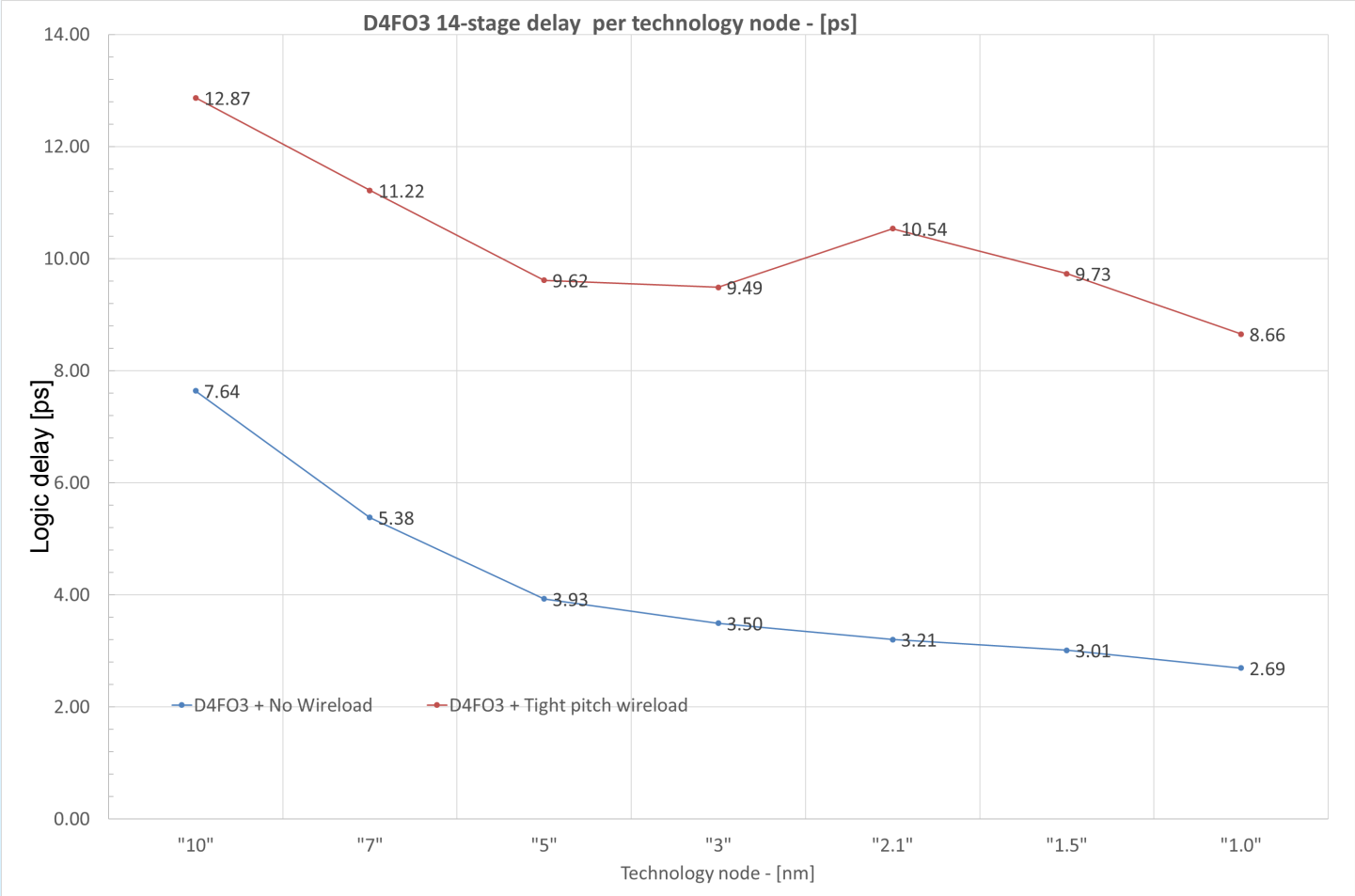
A slow tapering off --- feature sizes will continue to diminish until 1nm in 2033, *with monolithic 3D transistors expected from 2024 onwards*

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%**

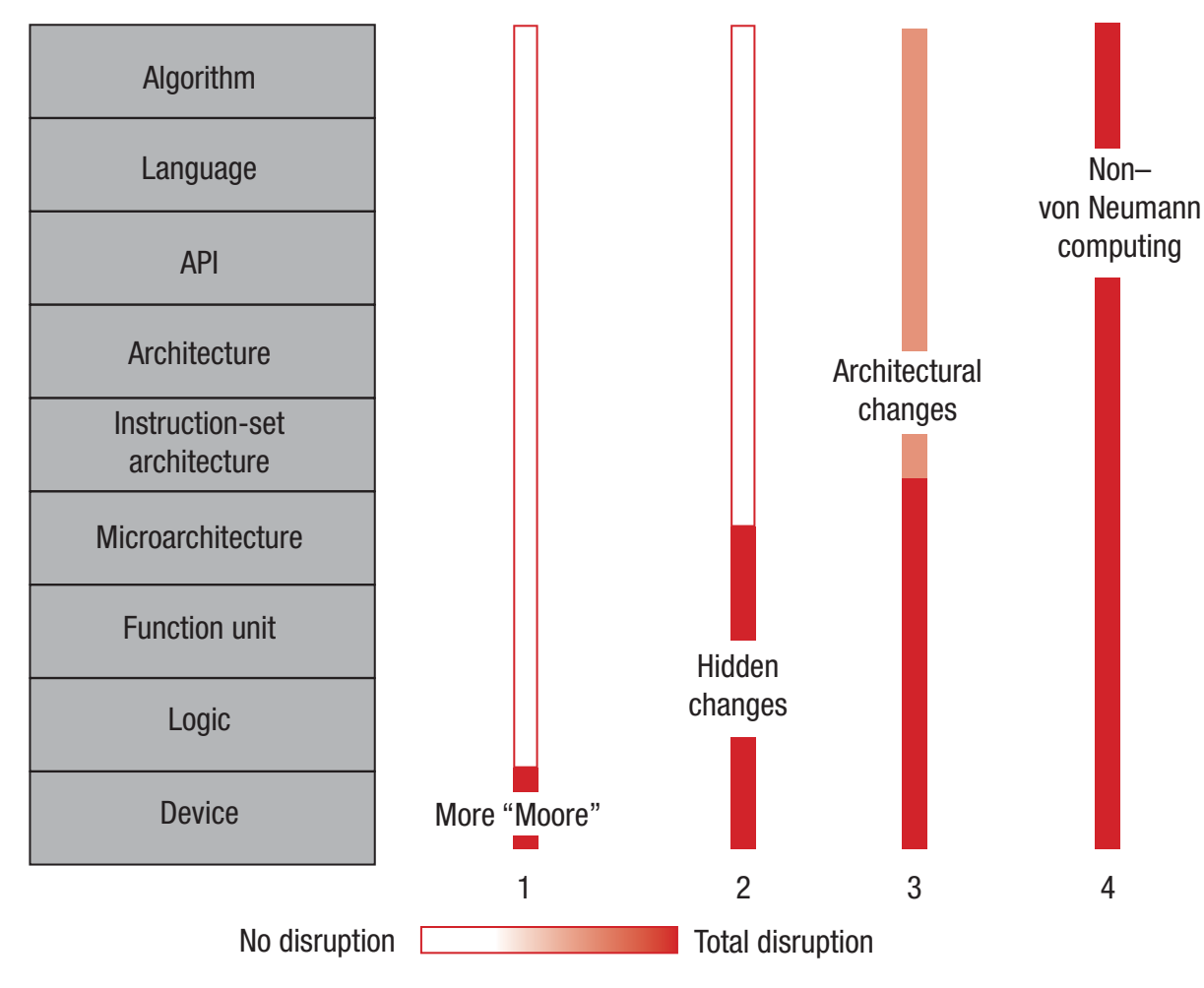
Source: ***IEEE IRDS 2017 Edition***

Minimal performance improvement past node 5



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

Levels of Disruption in Moore's Law End-Game 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.

Taxonomy of Future HPC technologies being considered by our subcommittee

(In order of increasing levels of disruption)

- Von Neumann approaches with specialized computing
  - GPU accelerators
  - *Reconfigurable logic*
  - CPU-integrated accelerators
- *Memory-centric computing*
- *Photonics*
- Non-Von Neumann approaches
  - *Neuromorphic computing*
  - *Analog computing*
  - *Quantum computing*

Common themes: extreme heterogeneity, specialization, hybrid systems

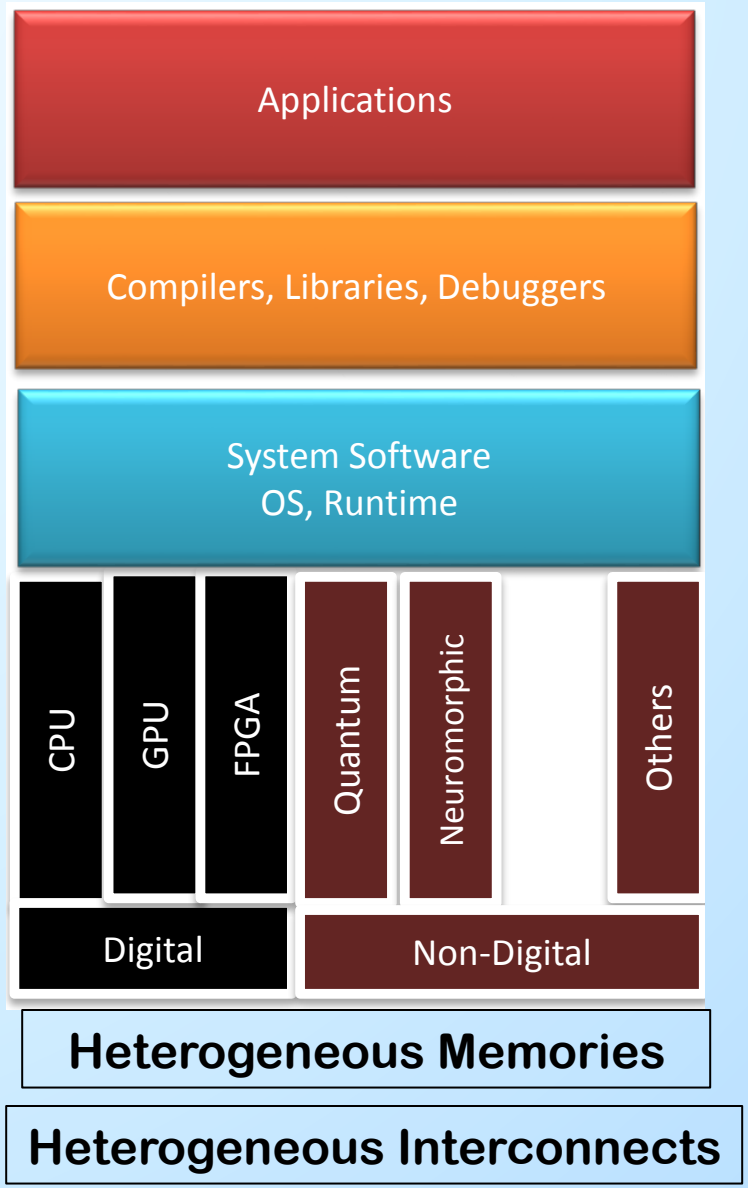
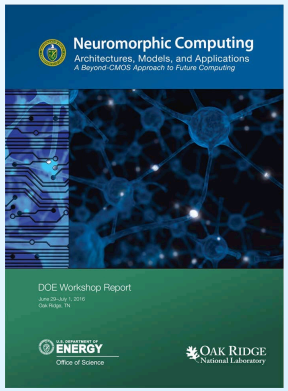
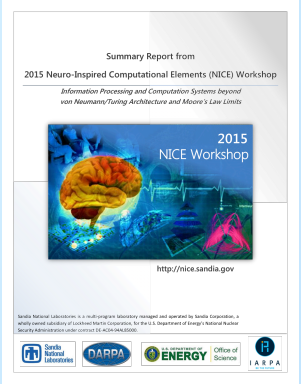


Figure source: presentation on "Advanced Scientific Computing Research", Barbara Helland, ASCAC meeting, Sep 2017.

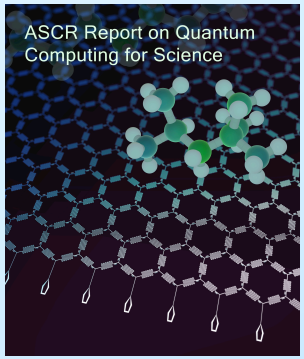
Community investigation of future technologies

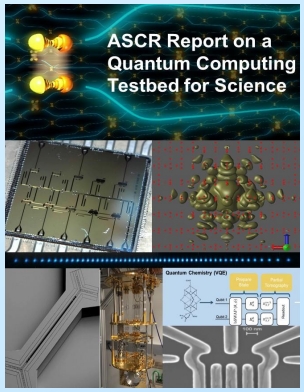
- Several recent DOE workshops and reports have focused on future HPC technologies

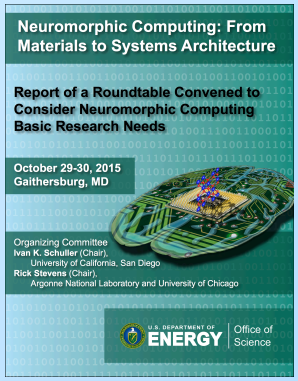




. . .







Technology Readiness Levels (TRLs)

Relative Level of Technology Development	Technology Readiness Level	TRL Definition	Description
	TRL 4	Component and/or system validation in laboratory environment	The basic technological components are integrated to establish that the pieces will work together. This is relatively "low fidelity" compared with the eventual system. Examples include integration of ad hoc hardware in a laboratory and testing with a range of simulants and small scale tests on actual waste[2]. Supporting information includes the results of the integrated experiments and estimates of how the experimental components and experimental test results differ from the expected system performance goals. TRL 4-6 represent the bridge from scientific research to engineering. TRL 4 is the first step in determining whether the individual components will work together as a system. The laboratory system will probably be a mix of on hand equipment and a few special purpose components that may require special handling, calibration, or alignment to get them to function.
Research to Prove Feasibility	**TRL 3**	Analytical and experimental critical function and/or characteristic proof of concept	Active research and development (R&D) is initiated. This includes analytical studies and laboratory-scale studies to physically validate the analytical predictions of separate elements of the technology. Examples include components that are not yet integrated or representative tested with simulants.[1] Supporting information includes results of laboratory tests performed to measure parameters of interest and comparison to analytical predictions for critical subsystems. At TRL 3 the work has moved beyond the paper phase to experimental work that verifies that the concept works as expected on simulants. Components of the technology are validated, but there is no attempt to integrate the components into a complete system. Modeling and simulation may be used to complement physical experiments.
Basic Technology Research	**TRL 2**	Technology concept and/or application formulated	Once basic principles are observed, practical applications can be invented. Applications are speculative, and there may be no proof or detailed analysis to support the assumptions. Examples are still limited to analytic studies. Supporting information includes publications or other references that outline the application being considered and that provide analysis to support the concept. The step up from TRL 1 to TRL 2 moves the ideas from pure to applied research. Most of the work is analytical or paper studies with the emphasis on understanding the science better. Experimental work is designed to corroborate the basic scientific observations made during TRL 1 work.
	TRL 1	Basic principles observed and reported	This is the lowest level of technology readiness. Scientific research begins to be translated into applied R&D. Examples might include paper studies of a technology's basic properties or experimental work that consists mainly of observations of the physical world. Supporting Information includes published research or other references that identify the principles that underlie the technology.

Relative Level of Technology Development	Technology Readiness Level	TRL Definition	Description
System Operations	**TRL 9**	Actual system operated over the full range of expected conditions.	The technology is in its final form and operated under the full range of operating conditions. Examples include using the actual system with the full range of wastes in hot operations.
System Commissioning	**TRL 8**	Actual system completed and qualified through test and demonstration.	The technology has been proven to work in its final form and under expected conditions. In almost all cases, this TRL represents the end of true system development. Examples include developmental testing and evaluation of the system with actual waste in hot commissioning. Supporting information includes operational procedures that are virtually complete. An Operational Readiness Review (ORR) has been successfully completed prior to the start of hot testing.
	TRL 7	Full-scale, similar (prototypical) system demonstrated in relevant environment	This represents a major step up from TRL 6, requiring demonstration of an actual system prototype in a relevant environment. Examples include testing full-scale prototype in the field with a range of simulants in cold commissioning[1]. Supporting information includes results from the full-scale testing and analysis of the differences between the test environment, and analysis of what the experimental results mean for the eventual operating system/environment. Final design is virtually complete.
Technology Demonstration	**TRL 6**	Engineering/pilot-scale, similar (prototypical) system validation in relevant environment	Engineering-scale models or prototypes are tested in a relevant environment. This represents a major step up in a technology's demonstrated readiness. Examples include testing an engineering scale prototypical system with a range of simulants.[1] Supporting information includes results from the engineering scale testing and analysis of the differences between the engineering scale, prototypical system/environment, and analysis of what the experimental results mean for the eventual operating system/environment. TRL 6 begins true engineering development of the technology as an operational system. The major difference between TRL 5 and 6 is the step up from laboratory scale to engineering scale and the determination of scaling factors that will enable design of the operating system. The prototype should be capable of performing all the functions that will be required of the operational system. The operating environment for the testing should closely represent the actual operating environment.
Technology Development	**TRL 5**	Laboratory scale, similar system validation in relevant environment	The basic technological components are integrated so that the system configuration is similar to (matches) the final application in almost all respects. Examples include testing a high-fidelity, laboratory scale system in a simulated environment with a range of simulants[1] and actual waste[2]. Supporting information includes results from the laboratory scale testing, analysis of the differences between the laboratory and eventual operating system/environment, and analysis of what the experimental results mean for the eventual operating system/environment. The major difference between TRL 4 and 5 is the increase in the fidelity of the system and environment to the actual application. The system tested is almost prototypical.

Source: U.S. DEPARTMENT OF ENERGY Technology Readiness Assessment Guide, DOE G 413.3-4 10-12-09

ASCAC

U.S. DEPARTMENT OF ENERGY | Office of Science

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 still only 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

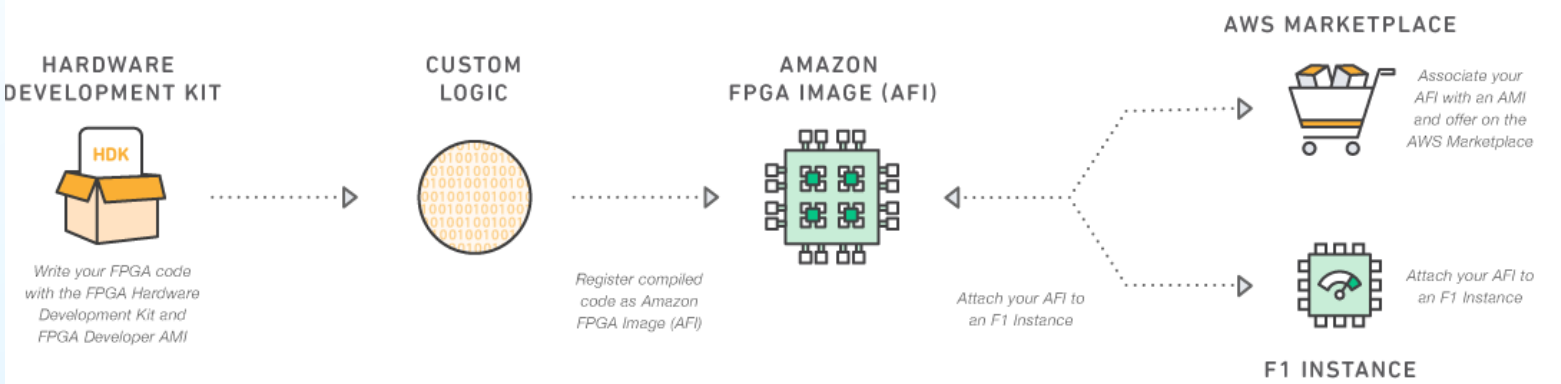
Timeframe:

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

Key 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

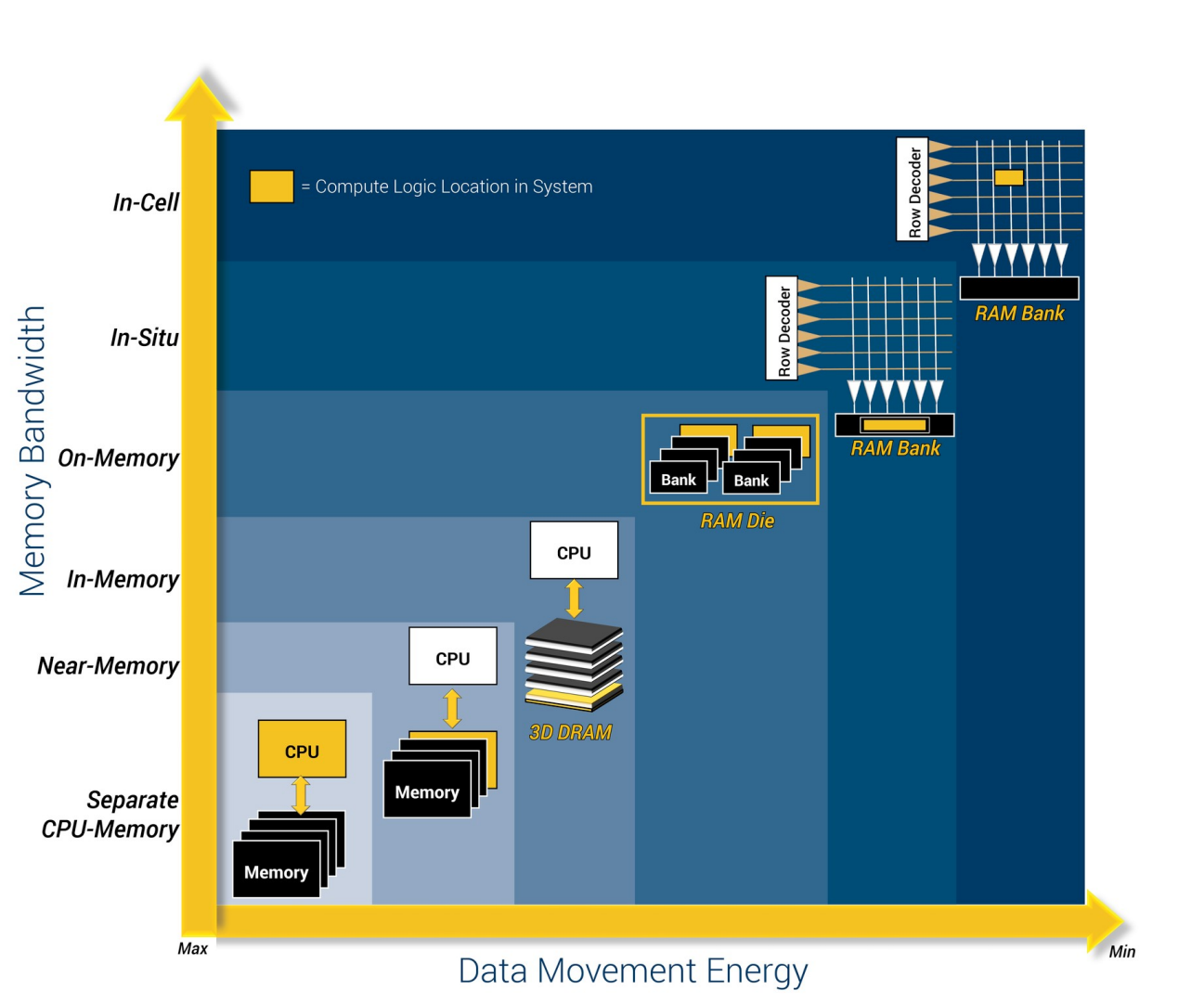
Offer AFIs you design on the AWS Marketplace for other customers.

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:

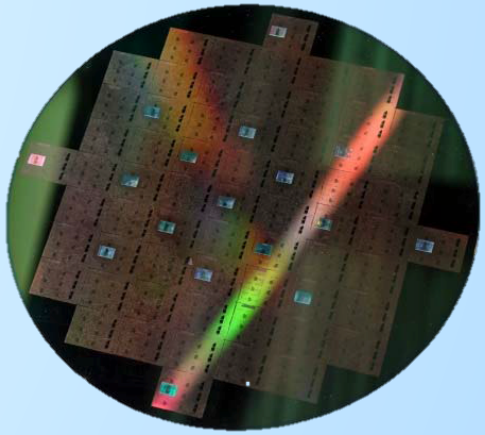
- Above approaches demonstrated at the research level. Near-Memory Processing appears to be the most viable for the next level, due to its synergy with 3D stacking.

Key challenges:

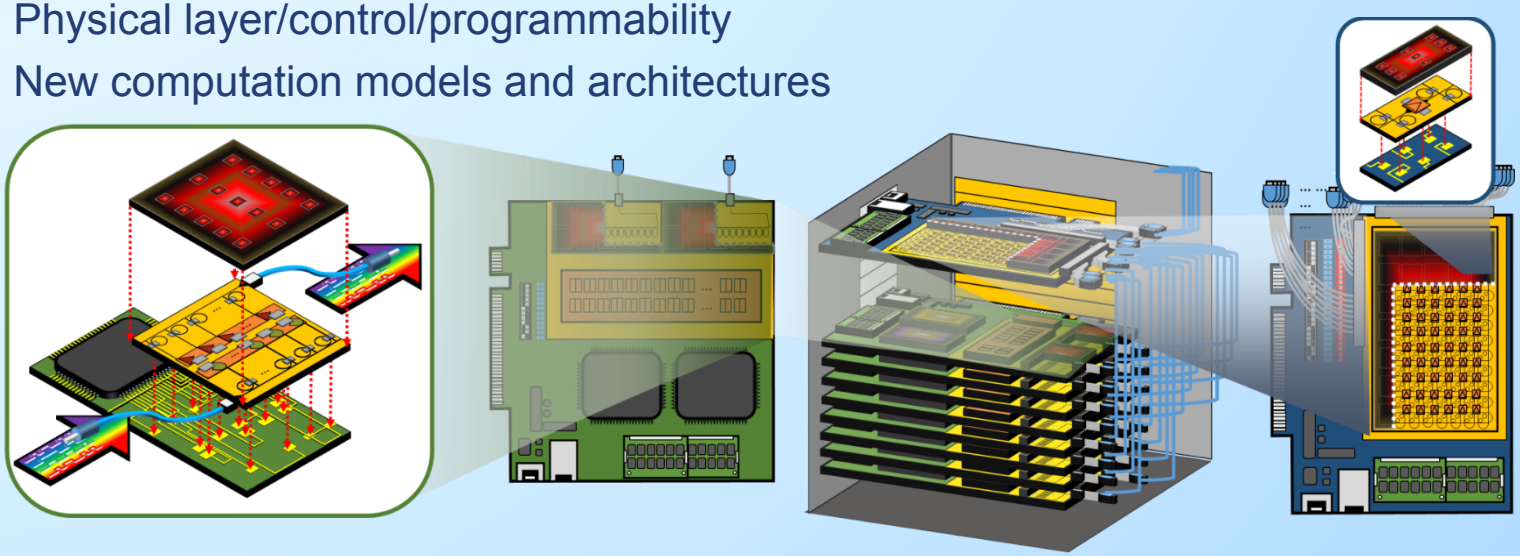
- How to maintain some level of coherence/consistency across data copies, how to support remote computations and a global address space, how to recognize completion of asynchronous operations, how to handle cases where data from separate memories need to be combined.

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
- Challenges:
  - Bridging photonics with computing systems
  - Physical layer/control/programmability
  - New computation models and architectures



300mm SiP wafer



Future directions for Photonics (example)

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]

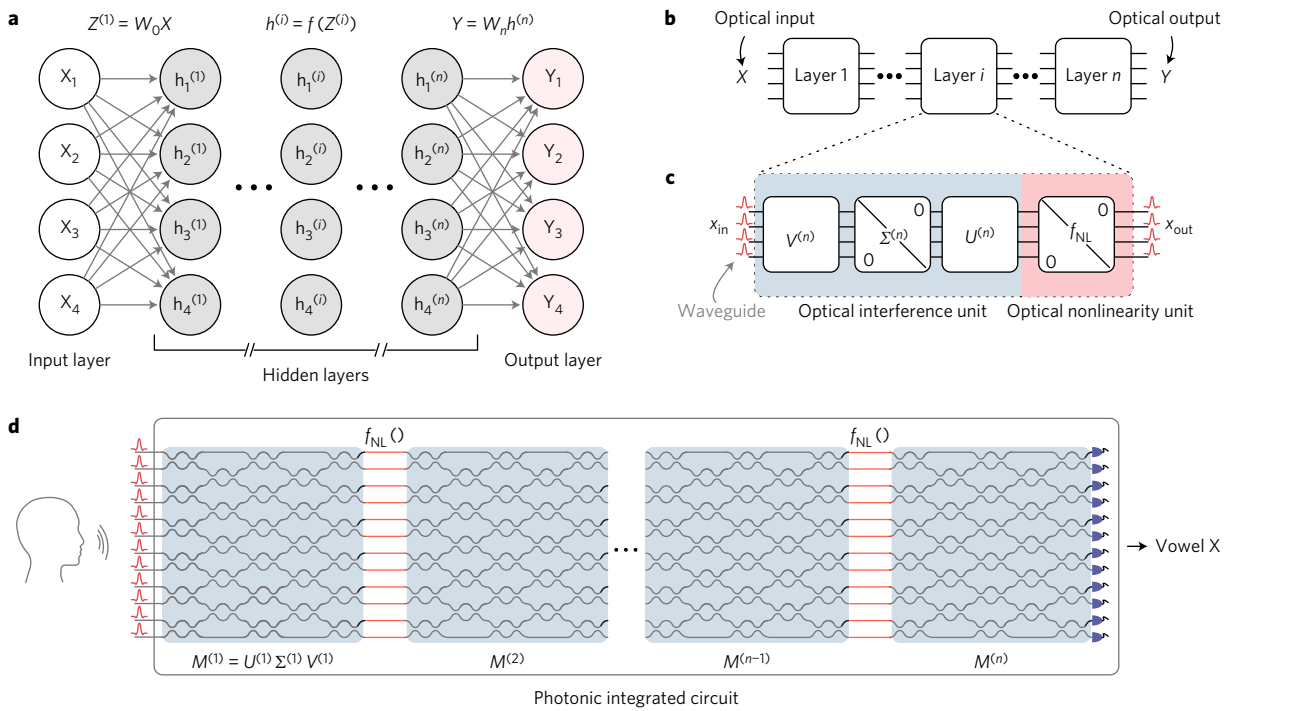


Figure 1 | General architecture of the ONN. a, General artificial neural network architecture composed of an input layer, a number of hidden layers and an output layer. **b**, Decomposition of the general neural network into individual layers. **c**, Optical interference and nonlinearity units that compose 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.
- Learning now emerging as third pillar of computational science (in addition to simulation & data)

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
- Going forward, we can expect neuromorphic computing to be used broadly, across data centers and embedded platforms (e.g., self-driving cars). Many companies are expected to propose and develop ASICs with efficient support for neuromorphic computing.

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

von Neumann Architecture | **Neuromorphic Architecture**

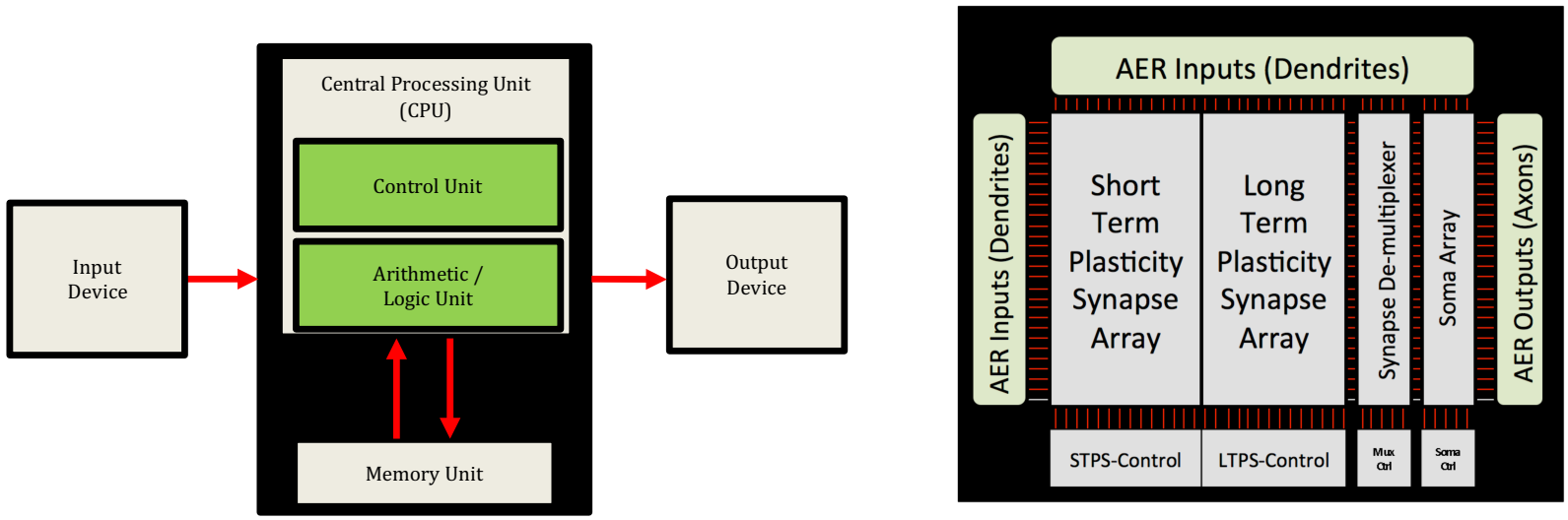


Figure 1. Comparison of high-level conventional and neuromorphic computer architectures. The so-called "von Neumann bottleneck" is the data path between the CPU and the memory unit. In contrast, a neural network based architecture combines synapses and neurons into a fine grain distributed structure 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

Overview of Electronic Analog Computing from the past

- Analog computers are especially well-suited to representing situations described by differential equations. Occasionally, they were used when a differential equation proved very difficult to solve by traditional means.
- The similarity between linear mechanical components, such as springs and dashpots (viscous-fluid dampers), and electrical components, such as capacitors, inductors, and resistors is striking in terms of mathematics. They can be modeled using equations of the same form.
- The electrical system is an analogy to the physical system, hence the name, but it is less expensive to construct, generally safer, and typically much easier to modify. As well, an electronic circuit can typically operate at higher frequencies than the system being simulated. This allows the simulation to run faster than real time (which could, in some instances, be hours, weeks, or longer).
- Electronics are limited by the range over which the variables may vary. Floating-point digital calculations have a comparatively huge dynamic range.

Source: https://en.wikipedia.org/wiki/Analog_computer#Electronic_analog_computers

EAI PACE TR-48 analog computer (1962)

EAI was the largest supplier of general-purpose analog computers. Transistorized models like the TR-48 were used for satellite design, chemotherapy studies, chemical reactor simulation, and more.

Source: http://www.computerhistory.org/revolution/analog-computers/3/152/430

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.

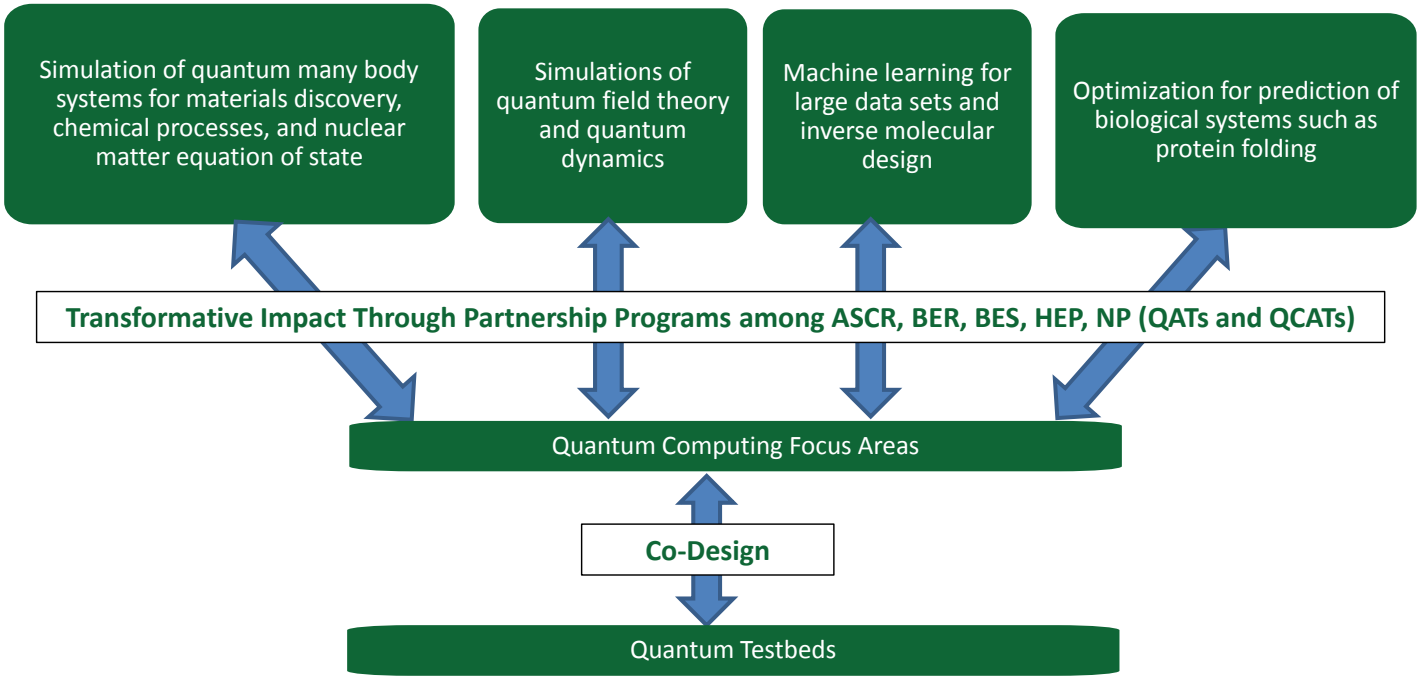
Key challenges:

- Effective bit precision of computation as a function of SNR is limited today, software support for (re-)configuration largely absent, and manufacturing of devices with useful dynamical behaviors is currently not an industry priority.

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

Quantum Computing Applications for SC Grand Challenges

QIS Task Force identified SC-wide grand challenges that will potentially be transformed by quantum computing applications.



U.S. DEPARTMENT OF ENERGY | Office of Science

ASCAC Presentation 9/26/2017 39

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.

Key challenges:

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

Framework for assessing application readiness for adopting new architectures

- **Application:** Scientific problem or subproblem with demand for extreme-scale computing.
- **Potential:** Evidence that one or more novel architectures could be suitable.
- **Readiness**: Suitability of current algorithms to novel architectures
- **Novelty:** Possibilities for new algorithmic approaches to addressing the same problem
- **Demand:** Urgency and demand for novel approaches
- **Agility**: Ability to quickly adapt to novel approaches
- **Total ranking:** The overall possibility that this is a driving application for one or more novel approaches. Will also feed into migrate vs. rewrite assessment.

Outline

1. Our charge

2. Post-Moore opportunities and challenges in Office of Science's mission

3. **Preliminary Recommendations**

Preliminary Recommendations

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

- 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)
- Longer & different time horizons for different technologies
- Need for agile and adaptive methodology/planning
- Play a leadership role in national and international collaborations

Preliminary Recommendations (contd)

Recommendation 2: DOE should prepare to invest in preparing for readiness of science applications for post-Moore

- In partnership with other science programs (as in SciDAC programs)
- Assess applications that will be prepared and need to be prepared for post-Moore
  - Which application areas are better positioned for post-Moore?
  - What are next game-changers for Science? What computation models do they need? Which post-Moore technologies can have the biggest game-changer impact on a science domain? New metrics, e.g., energy to knowledge?
- Workshops on post-Moore readiness, as was done for exascale readiness

Preliminary Recommendations (contd)

Recommendation 3: Identify and grow talent/staffing who can innovate in mapping applications onto emerging hardware (includes recognition of top talent in this regard)

- Also build pipeline e.g., CSGF is a foot in the door
- Encourage increase of named postdoc programs and LDRDs related to post-Moore
- Engage with interested & qualified faculty in academia through sabbaticals and other continuing engagements (e.g., joint faculty appointments)

Preliminary Recommendations (contd)

Recommendation 4: Facilities should prepare users for early access to testbeds and small-scale systems

- Includes training, workshops, support
- Build relationships with new classes of system/chip/device vendors)
- Without distracting from exascale commitments!

Summary

- Wide range of technologies for future high performance computing capabilities in different timeframes.
- Subcommittee is studying areas of research and emerging technologies that need to be given priority, but further investigation of technologies (requirements, workshops, etc.) will be needed beyond our study.
- Heterogeneity and hybridization are common themes in future HPC. No single technology will be the answer.
- Applications will need to be agile in evaluating and adopting technologies that are most promising for their domain, as well as “migrate vs. rewrite” decisions.
- Office of Science should play a leadership role in developing a post-Moore strategy/roadmap/plan for Science on HPC, without distracting from exascale commitments.