Computing Community Consortium (CCC) Response to NITRD "RFI Information on Update to Strategic Computing Objectives"

**Ian Foster, Sujata Banerjee, Randal Bryant, Tom Conte,
Juliana Freire, Mark D. Hill, and Ann Schwartz Drobnis**

This response was prepared by the Computing Community Consortium (CCC). The mission of the CCC is to catalyze the computing research community and enable the pursuit of innovative, high-impact research. Our goal is to identify and call attention to major research opportunities for the computing community.

1) What are emerging and future scientific and technical challenges and opportunities that are central to ensuring American leadership in SC, and what are effective mechanisms for addressing these challenges?

As the science, engineering, and biomedical communities pursue innovative applications of machine learning, we see unprecedented new demands for data and computing. These demands differ in important ways from those encountered in traditional scientific computing (SC). Meanwhile, the slowing of Moore's Law is driving a Cambrian explosion of accelerator designs, and increased interest in non-Von Neumann (e.g., neuromorphic) hardware. Mastering this new landscape will require substantial R&D, at every level from chips to system software and applications. With the technological future uncertain and America less clearly dominant, it is more important than ever to engage academia and government laboratories on these questions.

We encourage and applaud present efforts designed to maintain the US semiconductor lead as Moore's Law slows, e.g., via DARPA's Electronic Resurgence Initiative.

We recommend expanded efforts, especially with respect to AI's and deep neural network's unsustainable demand for compute, memory/storage, and bandwidth, which can perhaps be aided by an increased focus on coordination (co-design) from algorithms to hardware .

Finally, computational reproducibility is another area that deserves expanded efforts. Computational reproducibility is a requirement for science to progress, and currently, we lack the necessary tools and infrastructure to make computations reproducible. The need for reproducibility and transparency goes beyond science: it is a requirement for one to trust the results of any computation. This is thus a pressing issue, as important decisions are increasingly being made based on data computation.

2) What are appropriate models for partnerships between government, academia and industry in SC, and how can these partnerships be effectively leveraged to advance the objectives of SC?

We encourage industrial involvement to inform and perhaps support pre-competitive efforts to complement industrial development to aid industry and American competitiveness in the long run.

Given the expected importance of new hardware designs, including accelerators, to both SC and AI, it will be desirable to establish one or more research facilities where industry can place prototype systems for evaluation, and academia and government can gain experience in the use of those systems. Academia and government can also evaluate and analyze the systems more deeply (e.g., to identify possible security issues).

3) How do we develop and nurture the capable workforce with the necessary skill and competencies to ensure American leadership in SC? What are effective nontraditional approaches to lowering the barriers to knowledge transfer?

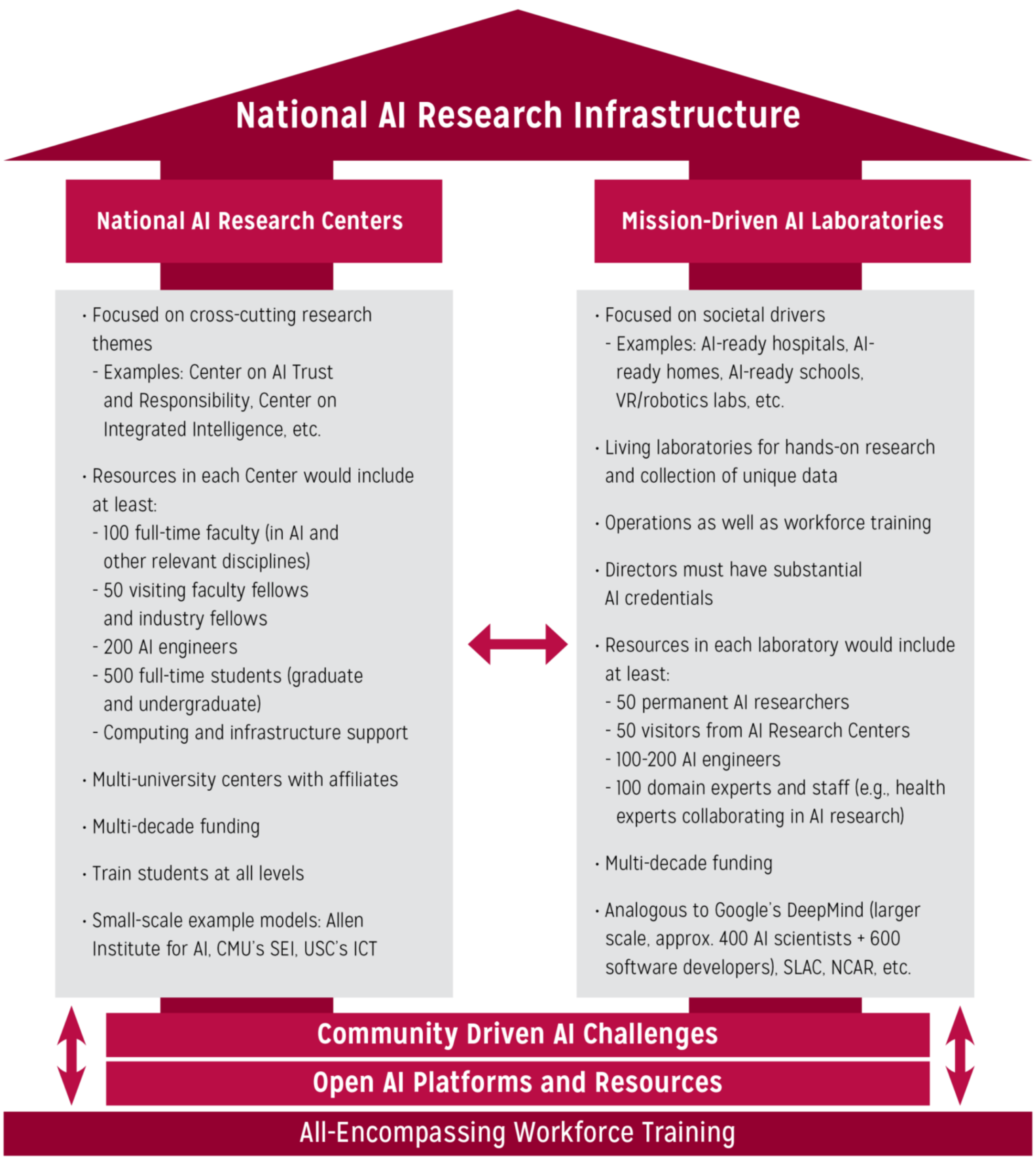
As AI methods (e.g., deep learning) become ever-more computationally demanding, and are increasingly applied to problems formerly addressed solely via numerical simulation, the workforce needs of SC and AI start to converge. New approaches to workforce development are required to meet growing needs for traditional SC, SC-aware AI, and AI-aware SC expertise. These new approaches must produce people with depth in both computer science and science/engineering disciplines.

Recommendations of the recently released CCC/AAAI 20-Year Roadmap for AI are relevant here[A 20-Year Community Roadmap for Artificial Intelligence Research in the US]. We point, in particular, to that report's recommendations to "Re-conceptualize and Train an All-Encompassing AI Workforce." There are eight major recommendations: 1) Developing AI curricula at all levels, 2) Recruitment and retention programs for advanced AI degrees, 3) Engaging underrepresented and underprivileged groups, 4) Incentivizing emerging interdisciplinary AI areas, 5) AI ethics and policy 6) AI and the future of work, 7) Training highly skilled AI engineers and technicians 8) Workforce retraining. Success in all of these areas will be key to providing the general public and decision makers with general knowledge about AI that will be fundamental for everyone, both as intentional consumers or unintentional recipients of AI technologies.

4) How can technical advances in SC and other large government and private initiatives, including infrastructure advances, provide new knowledge and mechanisms for executing next generation research?

The history of computing in the US shows that when powerful new tools are placed in the hands of talented scientists and students, remarkable things happen. The need now, therefore, is to engage today's scientists and engineers in working with early versions of the computing environments of tomorrow--environments in which interdisciplinary teams will work with powerful combinations of big computers, powerful AI, and massive data.

The CCC/AAAI 20-Year Roadmap for AI's recommendations to "Create and Operate a National AI Infrastructure" [A 20-Year Community Roadmap for Artificial Intelligence Research in the US] are directly relevant here. In building this national SC/AI infrastructure, we especially encourage creating arrangements that link commercial clouds and government laboratories's world-class supercomputers. The goal of the National AI Infrastructure is to reinvent the AI research enterprise through four interdependent capabilities: 1) open AI platforms and resources, 2) sustained community-driven challenges, 3) national AI research centers, and 4) national mission-driven AI centers. Figure 6 from the Roadmap gives an overview of these components and their interactions.



5) What are the future national-level use cases that will drive new computing paradigms, and how will new computing paradigms yield new use cases?

The research labs, design labs, manufacturing plants, communities, and hospitals of the future will all be deeply networked; connected to rich arrays of sensors producing large quantities of data; imbued with pervasive AI capabilities; and supported by powerful numerical simulations. These AI-, data-, and simulation-driven systems will provide a rich source of national-level use cases that will raise profound technical challenges at every level.

Pervasive AI poses new challenges for SC. To realize the full competitive potential of AI for the US, recent successes with deep neural networks (DNNs)—and industry's advances to come—need to be complemented with broader research in the three research areas outlined in CCC/AAAI 20-Year Roadmap for AI [A 20-Year Community Roadmap for Artificial Intelligence Research in the US], namely *integrated intelligence*, *meaningful interaction*, and *self-aware learning*. Only with a full complement of reasoning technique developed over one to two decades will the US move beyond the perception and categorization wonders of DNN to deeper intelligence with American values.

6) What areas of research or topics of the 2016 NSCI Strategic Plan should continue to be a priority for federally funded research and require continued Federal R&D investments? What areas of research or topics of the 2016 Strategic Plan no longer need to be prioritized for federally funded research?

All objectives remain important.

We observe that while progress towards Capable Exascale (Objective 1) is good, with deployment of first systems planned for 2021, work in this area needs to continue to address the needs of new workloads (e.g., AI) and to broaden access across industry, academia, and government. Growing demands from AI workloads will result in exascale capabilities being needed in every research laboratory and university. There are great challenges to addressing this growing demand without (much) traditional technology scaling, motivating a need to seek both new architectures and technologies.

7) What challenges or objectives not included in the 2016 NSCI Strategic Plan should be strategic priorities for the federally funded SC R&D? Discuss what new capabilities would be desired, what objectives should guide such research, and why those capabilities and objective should be strategic priorities.

Objective 2, Technology Coherence, should be expanded to encompass convergence with the rapidly expanding "AI" space.

Reduced precision is a new capability and also has a lot of potential. Faster is not the only consideration anymore. Making effective use of reduced precision and other more radical alternatives to 64-bit arithmetic (e.g., various accelerators, neuromorphic) will require potentially major changes to applications, algorithms, compilers, mathematical libraries, and other system components.

Many—if not most—of the benefits that information technology has provided to society have in turn depended on tremendous progress in technology (Moore's Law) and in hardware designs for compute, storage, and communication. Future information technology benefits cannot assume further transparent technology/hardware progress. Thus, we face three options: seek to provide benefits without technology/hardware change, work with hardware for synergies to affect change (e.g., reduced precision, specialized accelerators), or await new (post-CMOS) technologies to develop. Prudence suggests all three options should be exploited.