Overview Brochure

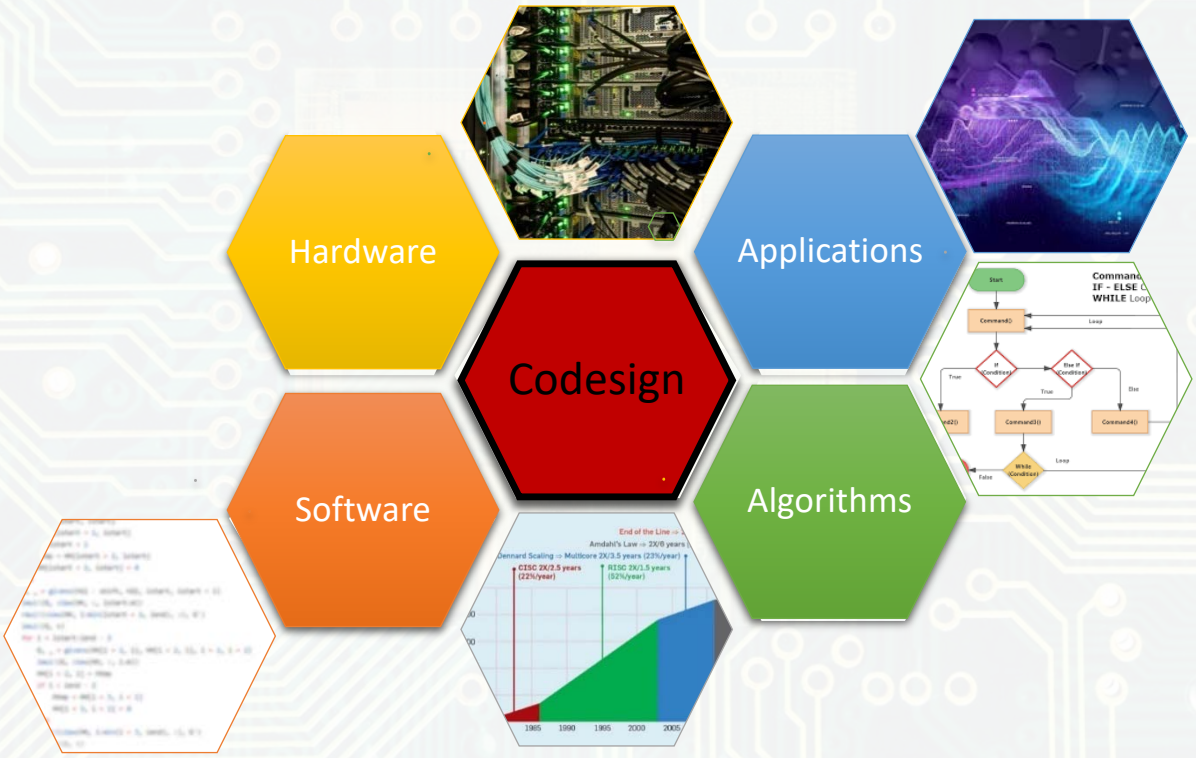
Basic Research Needs for

Reimagining Codesign for Advanced Scientific Computing

Unlocking Transformational Opportunities for Future Computing Systems for Science

16-18 March 2021

https://doi.org/10.2172/1822198



U.S. DEPARTMENT OF ENERGY | Office of Science

Reimagining Codesign in the face of extremely heterogeneous architectures and AI-enabled application workflows

In March 2021, DOE's Advanced Scientific Computing Research convened the Workshop on Reimagining Codesign. The workshop was organized around discussions on eight topic areas. From these, the workshop panelists identified four priority research directions, listed below. The full report for *Reimagining Codesign for Advanced Scientific Computing: Report for the ASCR Workshop on Reimaging Codesign* is available at https://doi.org/10.2172/1822199.

Motivation. Silicon-based transistors are nearing the limits of miniaturization, and the slowing pace of performance gains from smaller transistors is driving large-scale disruption of the entire computing ecosystem.[1] As a result, the high-performance-computing (HPC) ecosystem has transitioned from being dominated by a few relatively-similar general-purpose central-processing-unit (CPU) architectures to being dominated by a few relatively-similar increasingly general-purpose graphics-processing-unit (GPU) accelerator architectures. In alignment with the computing industry, it seems likely that these CPU/GPU architectures will rapidly be augmented by a diverse set of systems that comprise a broad portfolio of commodity plus customized modular components that include CPUs, GPUs, and artificial intelligence (AI) accelerators[2] where specialization provides benefits (as illustrated in Figure 1). This new ecosystem offers opportunities to significantly improve the performance, energy efficiency, productivity, reliability, and security of scientific applications, but exploiting these opportunities requires the development of innovative techniques and tools to rapidly codesign future software and hardware using verified, data-driven methodologies. Enabling transformative research in this new technological environment are five *key factors* distinguishing the present or near future from the past, as shown in Figure 2.

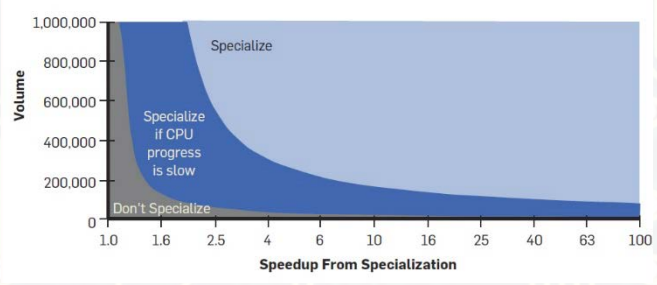


Figure 1: Benefits of Specialization. (Courtesy N.C. Thompson.)

Reimagining Codesign. Although the US Department of Energy (DOE) has successfully employed the codesign methodology to improve the software and hardware in several advanced HPC systems (see Figure 3), codesign was a distinct process, starting with workload analysis and ending with deployment and operation. Our workshop attendees concluded that a reimagined codesign process (as illustrated in Figure 4) that would be continuous, agile, and secure would better reflect the new reality of rapidly changing workloads and architectures. That is, future computing architectures will require substantially expanded scope to account for a broadened spectrum of applications that include AI/ML and graph methods for data-driven science; end-to-end processing for experimental instruments that aggregate and analyze real-time experimental data; and traditional numerically intensive HPC workloads. Each will

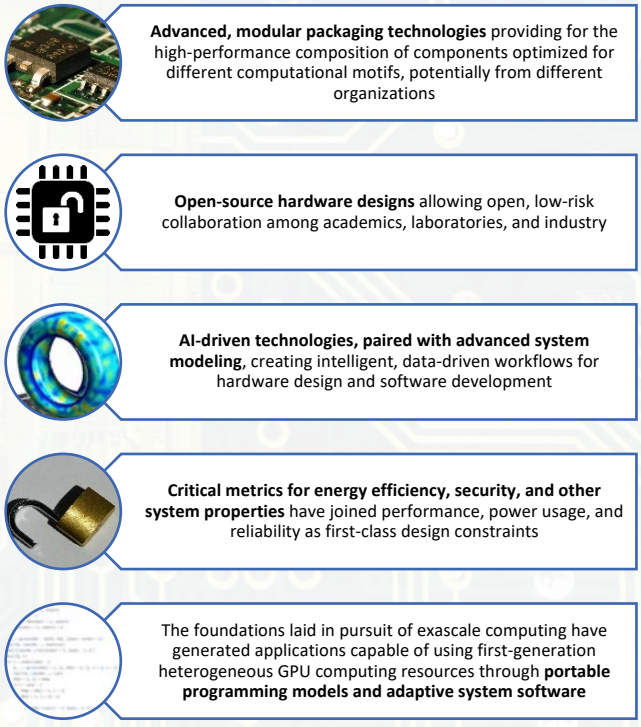


Figure 2: Enabling Key Technology Factors.

[1] "Chipmaking is being redesigned. Effects will be far reaching," *The Economist*, January 23, 2021. Accessed June 28, 2021.

[2] Vetter et al., *Extreme Heterogeneity 2018: Productive Computational Science in the Era of Extreme Heterogeneity: Report for DOE ASCR Workshop on Extreme Heterogeneity*, 2018, https://doi.org/10.2172/1473756.

require an end-to-end codesign approach to meet these new and more diverse mission requirements. In addition, the design of these computing architectures must account for new deployment scenarios at the edge and in the cloud, alongside traditional HPC data centers. In addition, the design of these computing architectures must account for new deployment scenarios at the edge and in the cloud, alongside traditional HPC data centers. Figure 5 summarizes the expected technology targets of this reimagined codesign process.

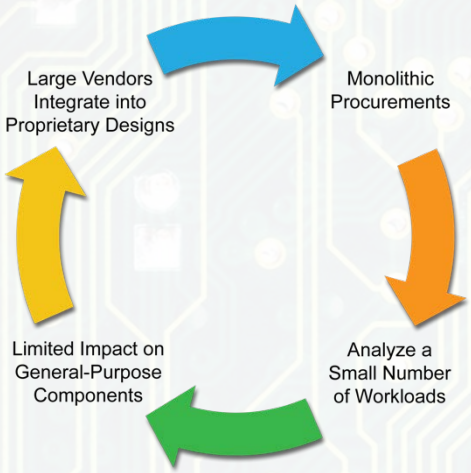


Figure 3: Past Codesign Activities

Priority Research Directions

1. Drive Breakthrough Computing Capabilities with Targeted Heterogeneity and Rapid Design

Key Questions: *What new methods and technologies are required to rapidly create breakthrough hardware designs? How can we ensure that they align to support increasingly diverse and demanding computing requirements?*

In today's rapidly changing technology landscape, designing radically heterogeneous systems requires new codesign methodologies that create an accurate understanding of workloads and use it to target and drive the creation of complementary sets of accelerators and heterogeneous structures that combine to create breakthrough systems. The flexible accelerated integration of such heterogeneous customized elements depends on advanced new physical integration technologies (e.g., chiplets and other advanced packaging), architectural integration (e.g., new memory interfaces, communication links, open standards/protocols), and accelerated hardware development (e.g., open-source designs or technology libraries and open-source tools). This research will yield innovative approaches to system design requiring only incremental enhancements to underlying microelectronics technologies while simultaneously complementing longer-term microelectronics research.

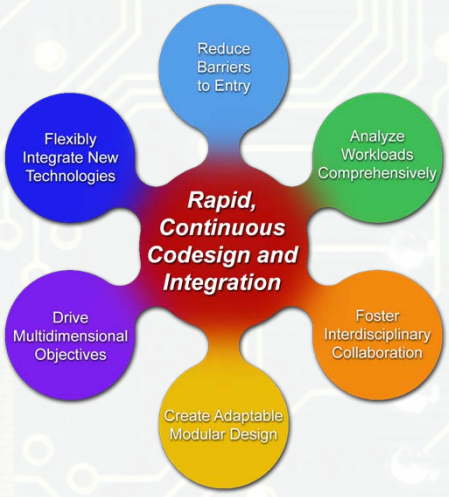


Figure 4: Re-imagined Codesign - Rapid, Continuous Codesign and Integration.

2. Software and Applications that Embrace Radical Architecture Diversity

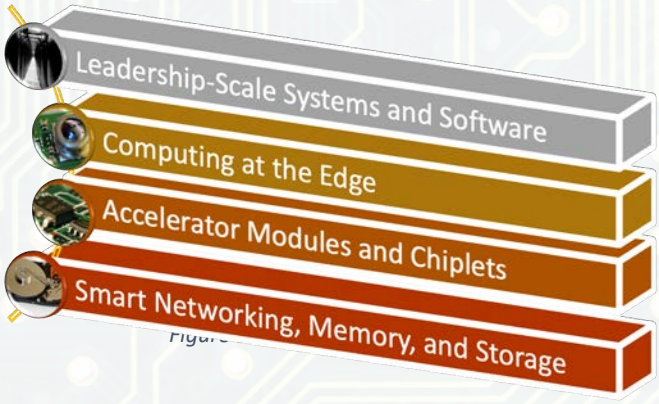
Key Question: *What novel approaches to software design and implementation can be developed to provide performance portability for applications across radically diverse computing architectures?*

Programming models, compilers, libraries, and run time systems must work with many types of compute engines and even new compute paradigms that differ dramatically from the traditional von Neumann architecture abstraction. New developments are needed in software abstractions to increase application portability; in dynamic run time systems to discover, schedule, monitor, and control highly varied resources; in tools for analyzing and predicting performance in the context of radical architectural diversity; and in metrics and benchmarks for quantifying progress beyond mere performance.

3. Engineered Security and Integrity from Transistors to Applications

Key Questions: *How does codesign consider needs for end-to-end scientific computing and scientific data security, provenance, integrity, and privacy? What computer security innovations from the commercial computing ecosystem (e.g., trusted execution environments) can be codesigned to provide security for DOE scientific discovery? How do we validate components with increasingly diverse supply chains and sources of development?*

Revolutionary advances are needed to extend roots of trust and other security capabilities to support DOE's scientific discovery continuum that leverages research networks, such as the Energy Sciences Network, to integrate supercomputing facilities with experimental user facilities, and this might extend to advanced wireless networking enabled Internet of Things sensors. This distributed scientific ecosystem provides increased exposure to threats, but it also offers the opportunity to leverage ubiquitous logic capabilities to enhance computer and data security. As a new area of concern, quantitative metrics assess how security codesign trade-offs can be made in conjunction with traditional metrics (e.g., power, performance, reliability). Moreover, given the proliferation of complex, modular components and community-developed open-source technology, our ability to validate new technology to protect against defects, including any intentional defects, must grow substantially.



4. Design with Data-Rich Processes

Key Questions: *What are the quantitative tools that are practical, accurate, and applicable to codesigning various layers of the hardware/software stack and of data-driven, dynamic, irregular workflows, such as those occurring in experimental science or AI/machine learning workloads?*

To be successful, quantitative tools—such as simulators, emulators, or profilers—must be applicable to design ahead (i.e., in advance of implementation) and follow through to assist optimizations during the execution of complex workflows on the target systems. These modeling and simulation capabilities must: (1) be sufficiently fast for repeated and potentially online use; (2) consider the triad of performance, power, and reliability in an integrated fashion; (3) be accurate over a broad range of hardware and software architectures; and (4) be scalable as the system complexity and size increases. Dominant workflows are increasingly data-driven, dynamic, and irregular and require new methods and tools of codesign that are dynamic and run time oriented.

Summary

Codesign in High Performance Computing and Artificial Intelligence has been critical to the design and implementation of contemporary computer architectures. As HPC applications evolve to include features for AI, connections to experimental facilities, and potentially mobile devices, the architectures and software will have to adapt much more quickly to serve these new emerging workloads efficiently. In this regard, the process of codesign must be reimagined to be continuous, agile, and secure to reflect the new reality of rapid change in both workloads and architectures. The four Priority Research Directions outlined above provide a sound foundation for a cohesive, long-term research and development strategy in reimaging codesign for advanced scientific computing. Over the last decade, DOE has invested heavily in codesign through the Exascale Computing Initiative; this effort created a baseline for codesign activities that will underpin key advances in these four PRDs. Such advances will build on this prior work from leading researchers in computer architecture, programming systems, simulation tools, workflow management; new research areas will emerge from the pursuit of next generation computing systems.

U.S. DEPARTMENT OF ENERGY | Office of Science