DISCLAIMER

HTTPS://DOI.ORG/10.2172/2476958

BASIC RESEARCH NEEDS IN

Energy-Efficient Computing for Science

September 9-11, 2024

U.S. DEPARTMENT of ENERGY | Office of Science

PREPARED FOR THE U.S. DEPARTMENT OF ENERGY ADVANCED SCIENTIFIC COMPUTING RESEARCH

ENABLING ENERGY EFFICIENCY FOR FUTURE GENERATIONS OF SCIENTIFIC COMPUTING

Large-scale computing has enabled numerous scientific discoveries, including ground-breaking achievements facilitated by the US Department of Energy (DOE) supercomputers and advances in applied mathematics and computer science. While important advances were made in energy efficiency to enable exascale computing, continued efforts are needed to dramatically improve the energy efficiency of the next generation of high-performance computing (HPC) systems and, more broadly, AI data centers. Without substantial improvements in energy efficiency, the energy consumption associated with computing could become a limiting factor for future scientific discovery, national security, and technological advancement. Therefore, realizing energy-efficient scientific computing remains a key challenge, and significant questions remain. In particular, how will next generation energy-efficient systems be designed, what new algorithms are needed to solve problems of interest on these platforms, how will they be programmed, and how will users manage the data generated by these tremendous computational capabilities in an energy-efficient way?

In September 2024, DOE's Advanced Scientific Computing Research (ASCR) program in the Office of Science convened the Workshop on Energy Efficient Computing for Science (alongside workshops on analog and neuromorphic computing) to identify research opportunities and grand challenges on this topic. Over three days, the participants identified five synergistic priority research directions (PRDs).

PRIORITY RESEARCH DIRECTIONS

1. Co-design energy-efficient hardware devices and architectures for important workloads

Key Questions: *How can energy-efficient devices using innovative compute methods be designed for scientific applications at scale? How can simulation and modeling tools help evaluate and develop new devices, circuits, and architectures? How can architects identify key HPC application kernels for specialized hardware? What methods can best support heterogeneous integration while minimizing energy loss at hardware component interfaces?*

Identifying promising candidates for energy-efficient devices and bridging device-level and architectural research are essential for the future of scalable scientific computing. These tasks involve taking a broad view of innovative computing paradigms, including analog, stochastic, optical, cryogenic, neuromorphic, quantum, and biological systems. Research on these paradigms must target large scale systems and include end-to-end evaluation of performance, energy use, manufacturing readiness, and yield. To maximize the benefit of specialization, innovation should focus on important computing kernels that are most likely to benefit DOE workloads and then develop synergistic technologies that ease the heterogeneous integration of specialized hardware for science into future scalable heterogeneous production systems. Advancements in synergistic technologies, such as chiplets and chip design software, will lower integration effort and costs, and access to fabrication facilities and toolchains will aid in accurate estimation and frequent prototyping. By co-designing these components, the DOE community can maximize the scientific impact while achieving unprecedented energy efficiency across diverse computational workloads.

2. Define the algorithmic foundations of energy-efficient scientific computing

***Key Questions**: Can energy complexity measures for algorithms be used to evaluate algorithm-hardware combinations accurately? How can telemetry of the current execution environment inform algorithmic choices in real time? How can knowledge of the algorithm inform execution to improve energy efficiency? What gains in energy efficiency are possible when focusing on scientific campaigns comprehensively rather than only individual tasks or operations? How can data-driven algorithms or models improve over conventional simulation?*

The study of energy efficiency in algorithms requires a mathematical notation that includes the energy costs of data movement and storage. With this new notation, it becomes easier to understand and optimize the energy complexity of different algorithm-hardware combinations and make informed algorithm choices based on the execution environment. By understanding how various aspects of computational workflows influence energy demands, users can optimize not just individual tasks but entire scientific campaigns. Strategies must be developed to adjust algorithmic parameters dynamically, responding to the specific hardware and resource availability in each environment. This approach balances energy costs across the entire lifecycle of scientific computation, including AI training and inference phases in tasks, such as training neural network surrogates. Integration of the algorithmic model with the dynamic heterogeneous hardware environment will provide substantial opportunities for reducing energy consumption across entire scientific campaigns.

3. Reconceptualize software ecosystems for energy efficiency

***Key Questions**: How can software ecosystems adapt for energy efficiency in innovative heterogeneous architectures? How should the foundations of software development shift to enable energy efficient computing technologies? How can developers program new energy-efficient hardware productively for complex scientific workflows?*

Software systems must evolve to achieve energy efficiency across existing and emerging architectures. Effective languages and compilers should produce energy-efficient code, while specialized libraries need to encapsulate hardware complexities without sacrificing efficiency. Runtime systems should dynamically optimize workload distribution with energy use in mind, while operating systems must intelligently balance energy consumption against the needs of performance, security, and reliability. Automating these capabilities will enable the integration of energy-efficient practices across diverse hardware technologies and facilitate development of fresh techniques that streamline abstractions, protocols, and interfaces to minimize energy use across deep software stacks.

4. Enable energy-efficient data management for data centers, instruments, and users

***Key Questions**: What strategies can minimize energy use in data movement across applications, systems, and distributed resources? How can new storage devices be designed and leveraged to lower energy consumption of I/O and data management? How can system architects balance I/O performance with energy-efficient data movement?*

Data movement constitutes a major energy expenditure in scientific workflows, but current storage systems are largely energy-agnostic and homogeneous, creating significant barriers to energy-efficient scientific computing. To address this, a paradigm shift in data management and storage strategies is imperative for sustainable HPC across distributed environments while preserving performance and dependability. Research efforts must span the entire storage infrastructure, including hardware and software stacks. Additionally, integrating innovative energy-efficient storage technologies will be essential to constructing robust data systems that operate within a limited energy budget while maintaining performance targets. Through these efforts, the scientific community can continue to derive vital insights as data volumes grow and experimental and computational infrastructures become more complex and geographically dispersed.

5. Develop integrated, scalable, and sustainable energy measurement and modeling for next-generation computing systems

Key Questions: *How should simulation and modeling infrastructures enable accurate multilevel energy modeling across devices, architectures, and systems? How can multiresolution measurement frameworks attribute energy usage across end-to-end hardware and software systems at scale? Which approaches balance fidelity and simulation efficiency to support early-stage design exploration? What strategies ensure reproducibility and sustainability as technologies evolve?*

Accurate, scalable modeling tools must quantify and predict energy use across all system levels—from devices to facilities. Current methods remain fragmented, rely on proprietary software, and ineffectively model emerging heterogeneous systems and technologies as parts of larger systems. Innovative simulation and modeling strategies should integrate energy models for diverse hardware, share open interfaces, support co-design of energy and performance, and accelerate innovation in novel computing paradigms. To ensure high confidence, these tools should enable reproducible and sustainable results, which can be replicated by industry, government, and academia. Similarly, energy measurement methodologies for operational systems and prototypes need to enable real-time, energy-aware optimizations and validate modeling predictions. Partnerships with prototyping facilities will create hardware and software testbeds that validate energy models for real applications, refining designs for real-world applicability.

SUMMARY

Energy-efficient computing is vital to the continued advancement of scientific research and the sustainable development of cutting-edge computing technologies. As we advance beyond the exascale era, the need to balance computational capabilities with energy consumption becomes increasingly crucial. The Workshop on Energy-Efficient Computing for Science, organized by the DOE's ASCR program, identified research opportunities and grand challenges for this topic. The workshop identified five priority research directions that promise to drive significant progress toward energy-efficient computing. Collectively, solutions to these PRDs will ensure that scientific advancements can be sustained.

U.S. DEPARTMENT of ENERGY | Office of Science