Advanced Computing Ecosystem

Request for Information

Version: 1.6

1. Introduction

The US Department of Energy (DOE) has a long history of deploying leading-edge computing capabilities for science and national security. The acquisition plans of the large DOE compute facilities continue to march forward, with new systems being deployed on a regular basis. Traditionally, a new system has been deployed approximately every five years; however, some facilities might be interested in deploying smaller systems more frequently (e.g., every one to two years). This request for information (RFI) from computing hardware and software vendors, system integrators, and other entities will assist the DOE national laboratories (labs) to plan, design, commission, and acquire the next generation of supercomputing systems in the 2025 to 2030 timeframe.

Future DOE supercomputers will need to tackle scientific discovery challenges against a backdrop of emerging edge computing technology, data science, and machine learning advances, in addition to traditional modeling and simulation application requirements. DOE also is planning for and designing an Advanced Computing Ecosystem (ACE) for this timeframe that will enable integration with other DOE facilities, including light source, data, materials science, and advanced manufacturing. The next generation of supercomputers will need to be capable of being integrated into an ACE environment that supports automated workflows, combining one or more of these facilities to reduce the time from experiment and observation to scientific insight.

For this RFI, DOE is interested in the deployment of one or more supercomputers that can solve scientific problems 5 to 10 times faster—or solve more complex problems, such as those with more physics or requirements for higher fidelity—than the current state-of-the-art systems. These future systems will include associated networks and data hierarchies. A capable software stack will meet the requirements of a broad spectrum of applications and workloads, including large-scale computational science campaigns in modeling and simulation, machine intelligence, and integrated data analysis. We expect these systems to operate within a power envelope of 20–60 MW. These systems must be sufficiently resilient to hardware and software failures, in order to minimize requirements for user intervention. As the technologies evolve, we anticipate increased attention to resilience in other supercomputing system developments.

We also wish to explore the development of an approach that moves away from monolithic acquisitions toward a model for enabling more rapid upgrade cycles of deployed systems, to enable faster innovation on hardware and software. One possible strategy would include increased reuse of existing infrastructure so that the upgrades are modular. A goal would be to reimagine systems architecture and an efficient acquisition process that allows continuous injection of technological advances to a facility (e.g., every 12–24 months rather than every 4–5 years). Understanding the tradeoffs of these approaches is one goal of this RFI, and we invite responses to include perceived benefits and/or disadvantages of this modular upgrade approach.

The information supplied in responses to this RFI will inform how DOE and the labs update their long-term advanced computing roadmaps, and set up requests for proposals (RFPs) for their next-generation systems. As such, DOE will share and discuss responses to the RFI with national labs, including Oak Ridge National Laboratory, Lawrence Berkeley National Laboratory, Lawrence Livermore National Laboratory, Los Alamos National Laboratory, Sandia National Laboratory, and Argonne National Laboratory. Any response that includes information requiring a non-

disclosure agreement (NDA) before dissemination and discussion with these parties must include a written notice of such restriction and agreement to negotiate any required NDA in good faith and as expeditiously as possible. In response to this RFI, we also explicitly request information regarding critical R&D challenges faced by the vendor community in delivering computing resources, communication resources, data infrastructure, and usable software; these resources could include the programming environment, tools and libraries, and management or monitoring software. The responses to this RFI will inform the DOE Advanced Scientific Computing Research (ASCR) 10-year roadmap for strategic vendor engagement. In addition, responses to this RFI may be considered in setting up targeted RFPs as done in the earlier DOE[1] programs. DOE may support near-term targeted R&D investments, not tied to a specific procurement, to advance the state-of-the-art technology. Non-recurring engineering (NRE) may be part of the RFP and tied to a specific procurement. These NRE activities may be performed in conjunction with the acquisition to increase capabilities, offer higher performance, lower the total cost of ownership, and/or increase productivity.

2. Who May Respond?

Responders may include any vendor of hardware or software that would be deployed in a future advanced computing system. Examples include, but are not limited to, basic hardware component (e.g., memory, processors, interconnects, storage media) vendors, system (e.g., compute, storage) integrators, and software (e.g., compilers, system management, storage management) vendors. Any company that provides hardware or software technologies relevant to delivering multi-exaflop supercomputing may respond.

Companies that could respond as the prime contractor to any of the described platform RFPs, as well as providers of technologies that could potentially be part of platform RFP responses, are encouraged to respond to this RFI. In cases where a company could act as both the prime contractor and as a technology provider, it may provide information from either or both perspectives in the response. Your written response should present your answers to the questions in this attachment and any other information you deem helpful in developing an RFP.

3. Mission Need

High-performance computing (HPC) and data-driven modeling and simulation are used extensively in advancing DOE missions in science and engineering and in the stewardship of the nation's nuclear weapons stockpile. In order to maintain leadership and to address the future challenges in science, energy, health, and growing security threats, however, the United States must continue to push strategic advancements in HPC—bringing about a grand convergence of modeling and simulation, data analytics, deep learning, artificial intelligence (AI), quantum computing, and other emerging capabilities—across integrated infrastructures in computational ecosystems. New approaches to predictive analysis for scientific discovery and solutions to complex data-driven engineering problems will arise from this convergence.

The DOE exascale systems deployed this year and being deployed next year (i.e., Frontier, Aurora, and El Capitan) are designed to address this emerging convergence. They can run simulations that require the entire platform and take days to weeks to complete. The AI-driven approaches on these systems will be used to perform uncertainty quantification and to discover complex, non-linear relationships in the output of large multi-physics simulations and large science experiments. The

[1] https://www.osti.gov/biblio/1845203, de Supinski, B R, Atchley, S, Hughes, C, Goldstone, R, Finkel, H, Karlin, I, Pakin, S, and Daley, C. Non-Proprietary Companion to the Q1 CY2021 PathForward Final Assessment WBS 2.4.1, Milestone PM-HI-1040. United States: N. p., 2022. Web. doi:10.2172/1845203.

new capabilities of these systems will revolutionize scientific areas, such as energy production, materials design, chemistry, precision health care, advanced manufacturing, stockpile stewardship, and national security. For the past decade, the six science programs in the DOE Office of Science have formulated strategic plans for the disciplines that they steward. These plans rely on HPC in ever-increasing proportion, and, in recent years, the explicit call for HPC at exascale performance levels has been a common and defining theme.[2,3] Examples include discovery and characterization of next-generation materials; systematic understanding and improvement of chemical processes; analysis of the extremely large datasets resulting from the next generation of particle-physics experiments; and extraction of knowledge from systems-biology studies of the microbiome. Advances in applied energy technologies also are dependent on next-generation simulations, notably whole-device modeling in plasma-based fusion systems. The current Exascale Computing Project (ECP)[4] has developed a portfolio of applications and technologies at exascale that will use the current DOE exascale systems, while benefitting next-generation systems.

4. Planning for System Acquisition and NRE, and Tactical R&D Opportunities

The responses to this RFI will inform one or more DOE system acquisition RFPs, which will describe requirements for system deliveries in the 2025–2030 timeframe. These systems are expected to solve emerging data science, artificial intelligence, edge deployments at facilities, and science ecosystem problems, in addition to the traditional modeling and simulation applications.

Due to the need to understand the breadth of potential diversity for exascale systems in the 2025–2030 timeframe, your response to this RFI should include (at a high level) the different notional solutions that you could provide, along with more detailed information on your assumptions about underlying technologies upon which your strawman architectures depend.

One of the other goals of this RFI is to examine other acquisitions during this decade (2025–2030). If R&D targeting the end of this decade will enhance the production systems delivered in this time frame, your response to this RFI should describe the R&D in addition to system- and procurement-specific NRE.

5. Requested Information Categories

The labs are collecting information from potential vendors in all areas specified in this section. If the area does not apply to you, state in a short paragraph why it does not, and if it does, provide your response.

A. Hardware

While vendor roadmaps typically only extend 12–24 months into the future, far short of the targeted delivery timeframe, they are still useful in providing current and new product information. They also provide context for discussions about future directions and market dynamics. Technology trends will allow DOE to better understand the solution space and constraints on possible designs for the next-generation computing system.

---

[2] Pioneering the future Advanced Computing Ecosystem: A Strategic Plan, Committee on Technology, National Science and Technology Council, November, 2020, https://www.nitrd.gov/pubs/Future-Advanced-Computing-Ecosystem-Strategic-Plan-Nov-2020.pdf.

[3] DOE Exascale Requirement Review, http://exascaleage.org/

[4] Details of the individual ECP applications, software technologies, and hardware projects, can be found at the Exascale Computing Project website (https://www.exascaleproject.org/).

We request information in the following categories:

A.1 Underlying technology trends

Describe the fundamental technologies that impact your company's product designs. For processors, the drivers could be process node improvements, packaging of chiplets into a single package, interconnects (e.g., chip-to-chip, processor-to-processor, processor-to-memory), 2.5D and 3D stacking (whether memory-on-logic or logic-on-logic), or thermal and manufacturing constraints. For memory, drivers of change include new standards for DDR and HBM, process node improvements, new devices, capacity improvements, and persistent memories.

For interconnects, drivers can include PHYs (e.g., host and endpoint-to-endpoint), protocols (i.e., PCIe, CXL, Ethernet, proprietary), and user interfaces (e.g., libfabric, UCX, sockets, proprietary). Please provide a roadmap of capabilities for your key technologies and a vision for systems during the 2025–2030 timeframe. Information about current product offerings can be used to provide a needed baseline for comparison. We prefer quantitative capability targets and goals instead of qualitative descriptions. When in doubt, please base projections on assumptions of a constant power and/or cost baseline from today's leadership systems as this will enable us to understand your roadmap.

A.2 Process node/building blocks

- If your products are silicon-based, which process node(s) does your company use for current generation products?
- Which process node(s) will be available for your components during the targeted timeframe?
- What will be the range of performance improvement and performance per watt for each new process node, starting with the process node used in your current products to the targeted timeframe?
- Notional system architecture sketches with characteristics of the nodes, memory hierarchy or hierarchies, interconnect(s), and system(s); Include, if needed, high-level considerations of the balance between traditional HPC (FP64) needs and AI (BF16/FP16) needs; Include considerations, if needed, of architecture optimizations for large-scale AI training (100 trillion parameter models); domain specific architectures (e.g., for HPC+AI surrogates and hybrid classical–quantum deployments). Our rough estimate of targets includes traditional HPC (based upon past trends over the past 20 years) systems at the 10–20 FP64 exaflops level and beyond in the 2025+ timeframe and 100+ FP64 exaflops and beyond in the 2030+ timeframe through hardware and software acceleration mechanisms. This is roughly 8 times more than 2022 systems in 2026 and 64 times more in 2030. For AI applications, we would be interested in BF16/FP16 performance projections, based on current architectures, and would expect additional factors of 8 to 16 times or beyond the FP64 rates for lower precision. Please indicate if your performance projections are based on hardware capabilities such as tensor or matrix operations as opposed to vector operations and any assumptions about sparsity.
- List available options for heterogeneous and accelerated computing on a node, and at other components of the system, such as within racks and across interconnects; Include if feasible emerging accelerator paradigms, such as quantum and neuromorphic accelerators; Include considerations of disaggregation and data-centric designs in the system interfaces, if appropriate.

A.3 Packaging

The industry has seen growth in the use of homogenous, multi-chip modules, as well as heterogeneous, chiplet-based System-on-Chips (SoCs) to improve yield, to drive down cost, to provide the ability to customize SoC designs (e.g., different combinations of cores), and to build larger SoCs than traditional, monolithic designs that are reticle-limited. In addition to enabling larger SoCs, packaging innovations such as 2.5D and 3D stacking are becoming common.

- In the targeted timeframe, what, if any, packaging improvements will benefit your components?
- What impact might these changes provide? Some questions to consider are as follows:
    - What will be realistic SoC sizes during this timeframe?
    - How much compute and memory is possible in a single socket?
    - What would power per SoC be during this timeframe?
    - Will we see more 3D stacking? Logic-on-logic?
    - What solutions might exist (or need to be developed) to manage thermal challenges of 3D packaging?
    - For non-compute ASICs (e.g., NICs, DPUs, FPUs), what impacts might packaging have on features, performance, power efficiency, and thermal management on future designs?
    - What are the tradeoffs between manufacturability and/or cost and performance?

A.4 Memory

- What are your expectations for memory technologies, capacity, latencies, and bandwidths during that timeframe?
- What fundamentally new memory devices (i.e., alternatives to CMOS DRAM and Flash) do you expect to become feasible during that timeframe?
- The HBM roadmap or projections, including the following:
    - Will power simply scale linearly with these bandwidths or are there any opportunities for performance or power (i.e., wattage) improvements?
    - How many stacks per SoC could be reasonable?
    - Are there new technologies and how might they be used (e.g., processing-in-memory, silicon photonics connecting off-SoC memory modules, new memory devices)?

A.5 Interconnects

- What are your expectations for on-die (e.g., GMI, EMIB, UCIe) bandwidths and power during this timeframe?
- What are your expectations for intra-node (e.g., PCIe, CXL, NVLink, xGMI, OpenCAPI) bandwidth and power during this timeframe?
- What are your expectations for coherency or memory consistency models for intra-node components during this timeframe?
- What are your expectations for inter-node bandwidths? Port bandwidths (e.g., 800 Gb/s, 1Tb/s)? Injection ports per node?
- Electrical versus optical (i.e., is electrical still viable during this timeframe and at what distance)?

- What topologies are you considering for your future deployments?
- High performance and scalable interconnect fabrics that support HPC, AI/ML, and data analytics workloads: how will these solutions evolve and how might they be used (e.g., software defined networks, DPU/IPU, passive optical fabrics)
- Are there opportunities to improve power efficiency?

A.6 Processor options

- Given the previously listed SoC constraints, which options would you propose?
- APU/XPU, CPU+wide vectors+HBM, reconfigurable (e.g., FPGA, coarse-grained reconfigurable arrays), or any others?
- What are your projections for datatype/datapath mix (e.g., FP64, FP32, FP16 datatypes and FMA versus GEMM for datapath) in future processors?
- Will an emphasis on AI drive lower or mixed precision improvements faster than FP64 performance? If so, how?

A.7 Storage

- What are your expected Bandwidths/IOPS for HDD, SSD, PCM, or any others?
- What capacities will be available for each media type?
- Do you see flash drives narrowing the gap in cost per byte with HDD? If so, in what ways?
- Are there new technologies and how might they be used (e.g., persistent memory, headless storage, compute-in-storage, new interfaces)?
- Are there new software and storage options and how would they apply to HPC/AI?
- What metadata and data backplane solutions are on your roadmap to better support campaigns running on the future computing ecosystems?
- If you are an integrator, what data tiering and I/O subsystems do you anticipate including, such as a data hierarchy from nodes to center-wide file systems?

A.8 Potential node configurations

- Given the previously mentioned constraints, how might a node be configured during the requested time frame?
- What compute blade configurations might be anticipated?
- What might a collection of disaggregated nodes consist of?

A.9 Cabinets (Integrators only)

- Do you envision a new, higher-power cabinet design during this timeframe?
- What power envelope do you expect per node (i.e., or board or chassis)?
- What is a reasonable power density (e.g., kVA/ft^2)?
    - The power cabinets for today's exascale systems produce an approximate 12.5 kVA/ft^2 (400 kVA/32 ft^2).
- For cooling, will Direct Liquid Cooling (DLC) be sufficient, or do we need to look at alternatives (e.g., immersion)?

- Will W32 cooling provide enough cooling for your notional architectures? What design challenges must be managed to continue use of primary-side water cooling to 32 °C? Will warmer options (e.g., 40 °C) be possible?
- If DLC, what options might exist for cooling fluid (e.g., water, propylene glycol water)?

A.10 Overall System

- Given the processors, nodes, interconnects, and cabinets, describe future systems?
- What technology components of the system are upgradeable (i.e., added or swapped in a systematic fashion) on a 12–24 month cycle? This should include any processors, memory, nodes, interconnects, storage, or appliance-class accelerators.
- Given the same footprint as Frontier (i.e., an approximate 4,000 ft^2), what might performance and specifications look like?
- If the footprint were 50 percent larger than Frontier, what would the performance and specifications be?
- Do you intend to include on- or off-premises cloud resources as a potential part of the solution, and what are the main associated cost and performance tradeoffs that you anticipate?
- What might the system MTBF be? What are Reliability, Availability, and Serviceability (RAS) solutions for compute, network, and data you might include in your system?
- What specific actions are you taking to reduce risk from supply-chain issues throughout the design, test, validation, and manufacturing phases?
- For major subsystem or appliance offerors, what are the main concerns (or risks) you anticipate in providing solutions directly to DOE to integrate into their system as opposed to through a single-system integrator? What are the main benefits to you of such an approach?

B. Describe your possible software and application support infrastructure

B.1 Provide a high-level description of the software stack, including the following components:

- Software environment(s) (i.e., management, development, operating)
    - Simplified capabilities of end-to-end use
    - Structure of system management
- Scope of workflow support integrated with the underlying system architecture.
- Candidate deployment interfaces for ACE components at multiple facilities connecting to central supercomputers?
- How will MPI need to evolve to address the scalability and heterogeneity of future systems and the mixed workloads and workflows these systems enable?
- Have you identified any trends in programming models that we need to be aware of? Similarly, trends in languages? If so, please specify.
- Is AI/ML/DL pushing programming in new directions, and if so, what are they?
- Would you maintain a full stack or use more open source software? If yes, how and in what ways?

B.2 Advances needed for large-scale acceleration (including those with HPC + AI hybrid appliances, quantum computing, or domain-specific systems)

- For HPC modeling and simulation applications, describe the technological advances that are anticipated or needed to deliver realized gains—of approximately 8 to 64 times over 2022 HPC system (e.g., multiphysics, PDEs) applications —by 2030. You may use as examples themes from past AI for Science workshops and roundtables conducted by DOE[5].
- Ability to support portability, since application modules may have to run on multiple parts of the ACE.
- Ability to provide compilers and runtimes that offer supercomputing performance in the context of a heterogeneous ecosystem; include, if needed, fundamental R&D challenges in the field of compilers and runtimes.
- General-purpose supercomputing architectures in the 2025–2030 timeframe may be paired with specialized types of accelerators to support AI and/or quantum applications. Please describe how you might support building large-scale systems that incorporate new types of AI and/or quantum computing accelerators. Please indicate any existing or planned joint R&D or partnership agreements with component providers that might supply such capabilities.
- For accelerator vendors (e.g., AI, quantum, vector, domain-specific), describe the state of the software development and market differentiators.
- For accelerator vendors (e.g., AI, quantum, vector, domain-specific), describe the roadmaps and quantitative capability targets (e.g., sizes of large language models).

[5] https://web.cvent.com/event/fc3922f8-fc75-4041-a317-f13a1da44f7c/summary
https://ai4esp.org/
https://doi.org/10.2172/1604756

B.3 Cloud offerings

Our potential interest in offerings from cloud computing providers is twofold. First, we are interested in understanding the possibility of using cloud resources to augment or replace our traditional HPC computing resources. We are interested, therefore, in projected use-case types available for leadership-scale supercomputing in the cloud. Secondly, we are interested in the possibility of cloud providers working with us to build systems that incorporate new technologies and components into full systems, deployable both on our premises and in commercial cloud data centers that would be then by available for use by us and by the broader community.

- Summarize your current cloud offerings and your approach to providing large-scale virtualized or bare-metal computing resources to end users. Large-scale here means "computing clusters" with from 1,000 to 10,000 nodes each and 4 to 8 GPUs or with other accelerators and the ability to be scheduled for units of long time periods. We are also interested in the hybrid cloud deployment model with the possibility of some quantity, perhaps significant, of gear deployed on-premises and paired with large-scale quasi-dedicated capability in the cloud, both running similar software.
- We are interested in your expected roadmap of capabilities for your key technologies and vision for cloud systems during this timeframe. Information about current product offerings can be used to provide a baseline if needed for comparisons. We prefer quantitative capability targets and goals instead of qualitative descriptions. Based on your current business model and deployment strategies, what advances in power efficiency, packaging, memory capacity and bandwidth, interconnects, and storage are you assuming will be needed to reach these overall targets.
- We are interested in understanding the component technologies and systems building blocks that will enable the construction of world-leading systems. Include any specific strategies, such as modular building blocks or industry standard packaging, that could streamline integration at scale. Please describe any unique capabilities you might have to support integration work at scale and to enable rapid upgrades of components. In some scenarios, we may consider electing to specify certain components and technologies; therefore, we are interested in your "default" or business-as-usual assumptions and your views toward integrating alternatives that may be specified by us or jointly developed. If this is of interest, please describe how you imagine this might work. Of particular interest is applying this approach to a hybrid cloud deployment, where we would have on-premises systems and access to cloud resources based on these technologies.
- Discuss how your systems would be expected to perform on large-scale HPC and/or HPC + AI exemplar applications likely to be of interest to DOE and NNSA. You may choose to base your projections on any relevant ECP mini-apps or AI benchmarks. If possible, describe how your technology could be used to support any of these use cases at scale.
- Describe how you might support building large-scale systems that incorporate new types of AI accelerators and/or quantum computing accelerators. Please indicate any existing or planned joint R&D or partnership agreements with component providers that might supply such capabilities.

- Describe your software stack and identify which elements of your software stack would be developed and supported by your company and which elements would need to be sourced to support complete systems. Please describe your preferred software partners for those additional elements. Describe the software ecosystem and roadmap to support workflows spanning one or more geo-graphically distributed systems, including with edge systems and in ACE environments. How would rapid technology injections and upgrades be realized? How can the systems be integrated as part of ACE environments?
- Describe your plans to develop and execute secure software technologies and applications.

C. Address the potential impact of DOE R&D investment and NRE funding on your proposed system(s)

Highlight those innovations that could significantly contribute to accelerating the trajectory of computing capabilities. Responses to this question can include innovations in any or all of your proposed solutions, but particularly the 2026 solutions and solutions that would promote wide ecosystem diversity.

- What are the primary areas in which you feel this funding is needed to deliver a capable multi-exascale system?
- Indicate your priorities for these efforts. Describe how they complement existing internal strategies and timelines.
- How would NRE funding alter the system that might otherwise be bid (e.g., compressed schedule, improved technology, reduced cost)?
- How would the NRE work complement your related R&D funding (e.g., internal and external, such as from earlier DOE Path Forward-like programs)?
- How does NRE funding impact cost? Higher capability at same cost? Reduced cost for same capability? Faster delivery of specific capabilities (which ones)?

For example, your answer could include innovations in power consumption, performance, performance analysis, programmability, reliability, data science, machine learning, portability, languages, compilers, runtimes, data management, or any other areas that your system design provides, including:

- Capability to merge data science, machine learning, and simulation into a single system.
- Hardware innovations (e.g., node, board, interconnect design)
    - Diverse processor technologies to promote a rich ecosystem
    - Memory system technology, specifically to improve effective memory bandwidth, latency, and/or capacity limitations
    - Improved interconnect performance
    - Improving I/O system performance
    - Resilience and RAS system
- Software innovations (e.g., ease of monitoring and managing the system)
    - Data management

- System software innovations that simplify the user experience of using an end-to-end ecosystem capability
    - Performance portability
    - Programmability
    - Usability by a wide variety of scientists and engineers across a range of applications
    - Usability by a wide variety of scientists across subcomponents (from Edge-to-Exascale) in the ACE
- Innovations that reduce total cost of ownership.
    - Reduced power consumption
- Innovations that support complex scientific workflows and user models.
- Innovations that might enable dramatic performance increase on certain applications but might not be broadly applicable across all DOE science areas.
- Packaging for purposes such as density, cooling, energy reduction, maintainability, energy reuse (i.e., efficiently managing waste heat rejection to support reuse strategies).

C.1 Cost Estimates

Provide cost estimates for NRE and for the described system(s).

Glossary of Acronyms

APU - AMD's Accelerated Processing Unit, combines CPU and GPU cores
ASCR - DOE's Advanced Scientific Computing Research
ASIC - Application-Specific Integrated Circuit
BF16 - Brain Float 16 data format
CMOS - Complementary Metal-Oxide Semiconductor
CPU - Central Processing Unit
CXL - Compute Express Link
DDR - Double Data Rate, standard memory in DIMM form factors
DIMM - Dual Inline Memory Module
DLC - Direct Liquid Cooling
DRAM - Dynamic Random Access Memory, used in DDR DIMMs and HBM
DPU - Data Processing Unit
EMIB - Intel's Embedded Multi-die Interconnect Bridge
FP16 - 16-bit Floating Point data format
FP32 - 32-bit Floating Point data format
FP64 - 64-bit Floating Point data format
GEMM - General Matrix-Matrix multiplication
GMI - AMD's intra-socket Infinity Fabric
GPU - Graphics Processing Unit
HBM - High Bandwidth Memory
HDD - Hard Disk Drive
IOPS - I/O Operations Per Second
IPU - Infrastructure Processing Unit
kW - Kilowatt (also KW)
kVA - kilovolt-amps
MPI - Message Passing Interface
MW - Megawatt

MTBF - Mean Time Between Failures
NIC - Network Interface Card
NNSA - DOE' National Nuclear Security Administration
NVLink - NVIDIA's proprietary inter-socket interface
OpenCAPI - IBM's Open Coherent Accelerator Processor Interface
PCIe - Peripheral Component Interface Express
PCM - Phase Change Memory
PDE - Partial Differential Equation
PHY - Physical layer device
SSD - Solid State Drive
UCIe - Universal Chiplet Interconnect Express
UCX - Universal Communications X software
W32 - 32°C cooling water standard
xGMI - AMD's inter-socket Infinity Fabric
XPU - Intel's Accelerated Processing Unit