Advancements in Artificial Intelligence for Science

DE-FOA-0003264

Dr. Hal Finkel, Dr. Steven Lee, Dr. Margaret Lentz, Dr. Kalyan Perumalla, Dr. Robinson Pino, and Dr. William Spotz

Advanced Scientific Computing Research

February 26, 2024

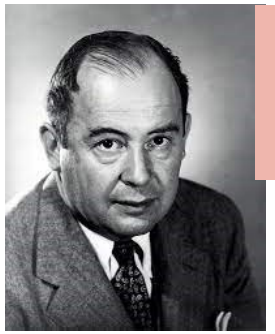
Disclaimer : This presentation summarizes the contents of the FOA. Nothing in this webinar is intended to add to, take away from, or contradict any of the requirements of the FOA. If there are any inconsistencies between the FOA and this presentation or statements from DOE personnel, the FOA is the controlling document.

FOA Issue Date:	**February 13, 2024**
Submission Deadline for Pre-Applications:	**March 19, 2024 at 5:00 PM Eastern Time** **A Pre-Application is required.**
Pre-Application Response Date:	**April 11, 2024 at 11:59 PM Eastern Time**
Submission Deadline for Applications:	**May 21, 2024 at 11:59 PM Eastern Time**

Outline

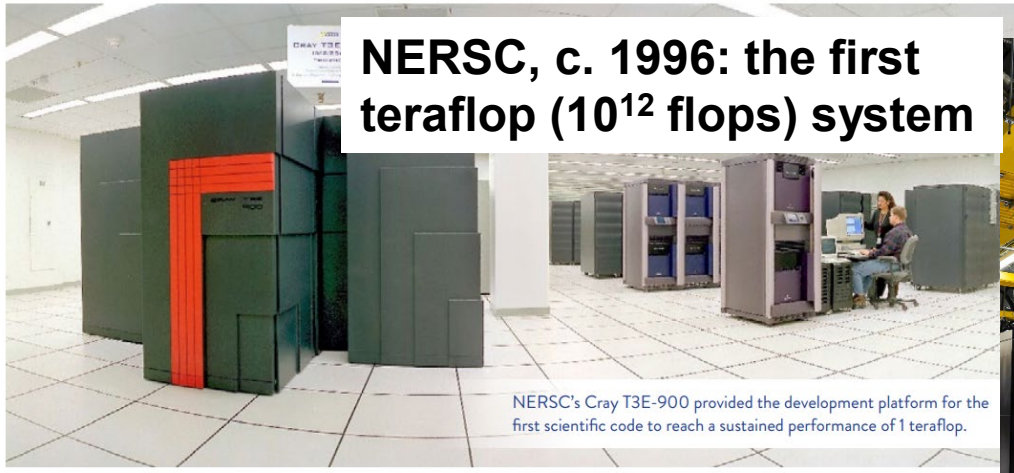
- Introduction to the Advanced Scientific Computing Research (ASCR) program.
- Overview of the Funding Opportunity Announcement
  - Research Areas
  - Eligibility
  - Procedure for Applying
- Q&A

ASCR – over 70 years of Advancing Computational Science



Beginnings: During the Manhattan Project, John Von Neumann advocated for the creation of a Mathematics program to support the continued development of applications of digital computing

ASCR has a rich history of investment in computational science and applied mathematics research, and revolutionary computational and network infrastructure.



NERSC, c. 1996: the first teraflop (10^{12} flops) system

NERSC's Cray T3E-900 provided the development platform for the first scientific code to reach a sustained performance of 1 teraflop.

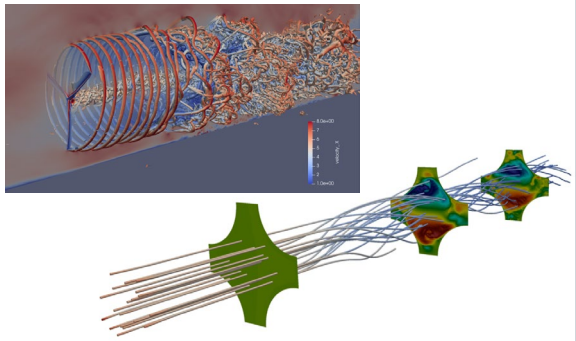
Today, Frontier at OLCF: first to exascale (10^{18} flops)

WHY COMPUTATIONAL SCIENCE?

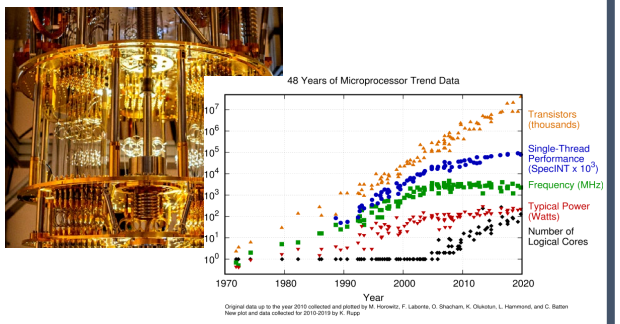
- Computational science added a third pillar to researcher's toolkit along side theory and experiments
- Valuable when experiments are too expensive, dangerous, time-consuming or impossible
- Facilitates idea-to-discovery that leads from equations to algorithms
- Virtually every discipline in science and engineering has benefited from DOE's sustained investments in computational science

Emerging Technology Trends for Scientific Computing

Advanced Modeling, Simulation, and Visualization

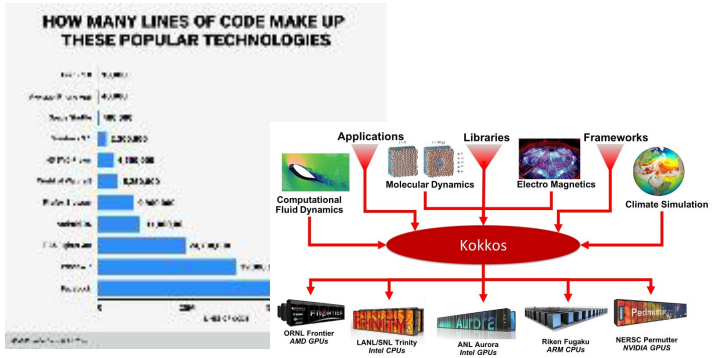


Trustworthy Artificial Intelligence and Data



Heterogeneous, Distributed, Co-Designed, Energy-Efficient Computing and Algorithms

Software Complexity for Increased Versatility



High-Performance Computing and Networking across Experiments, Exascale and the Edge

FOA: Advancements in Artificial Intelligence for Science

- Funding Opportunity Announcement (FOA) DE-FOA-0003264.
- Seeking basic computer science and applied mathematics research in the fundamentals of Artificial Intelligence (AI) for science. Specifically, advancements in this area are sought that can enable the development of:
  - Foundation models for computational science;
  - Automated scientific workflows and laboratories;
  - Scientific programming and scientific-knowledge-management systems;
  - Federated and privacy-preserving training for foundation and other AI models for
  - science; and
  - Energy-efficient AI algorithms and hardware for science.
- The development of new AI techniques applicable to multiple scientific domains can accelerate progress, increase transparency, and open new areas of exploration across the scientific enterprise.

Research Area 1: Extreme-Scale Foundation Models for Computational Science

- ASCR sees an opportunity to capitalize on the rapidly-advancing foundation-model techniques, combined with recently-deployed exascale computing resources, to expeditiously jumpstart the impact of foundation models on computational science in order to *significantly advance the state of the art in computational science*.
- A foundation model is trained on broad data, generally uses self-supervision for training, and is applicable across a wide range of contexts. While many well-known foundation models are Large Language Models (LLMs) processing textual data, foundation models can also be created for other kinds of data, or moreover, can be multimodal in nature (processing multiple kinds of data).
- For the purposes of this research area, a foundation model is expected to have at least tens of billions of parameters.
- It is expected that the foundation model, or foundation models, will explore innovative next-generation capabilities such as flexible multimodality, advanced external tool integration, sophisticated reasoning and planning, and robust memory.
- Given the importance of accuracy and quantified uncertainty in science, the research must address how these factors will be assessed.
- The proposal must address how the training process, or processes, will minimize risks associated with model misuse, unintentional biases or inaccuracies in model outputs, and other potential undesirable consequences.

Research Area 1: Extreme-Scale Foundation Models for Computational Science – Computing Resources

- ASCR's exascale supercomputers, Frontier and Aurora, each contain tens of thousands of state-of-the-art graphics processing units (GPUs) and rank among the world's most-powerful resources for AI model training.
- Proposed efforts are encouraged to make use of ASCR's exascale resources to create a foundation model, or foundation models, that take advantage of extreme-scale computing to advance the state of the art in foundation-model creation.
- *For this research area only*, a pre-application in response to this FOA may, at the applicant's discretion, also include an articulation of likely HPC resource requirements as might be requested from the ASCR Leadership Computing Challenge (ALCC) allocation program.
  - The quantity of physical node hours required (not multiplied by any site charging factor) and a detailed summary of the software to be used, the planned production runs, the number of nodes per run, the expected wall clock time required, and how benchmarking was determined.
  - ASCR is seeking projects that will be ready to use HPC resources starting in 2024 on Frontier, Polaris, Perlmutter-GPU, and/or Aurora.
- See the solicitation for important additional information.

Research Area 2: AI Innovations for Scientific Knowledge Synthesis and Software Development

- The state-of-the-art in knowledge synthesis and programming tools are changing rapidly, fueled by AI Large Language Models (LLMs) trained on text, source code, and other data sources.
- New AI-driven tools are currently not trustworthy; do not systematically understand mathematical and physical principles; cannot properly ingest and understand scientific literature and data; and do not produce consistent, verified, uncertainty-quantified, reproducible results.
- This research area seeks fundamental advancements in knowledge synthesis and programming tools for science. Moreover, realizing AI systems that can truly understand, and assist with, all aspects of the scientific process requires innovation in many areas, including multimodality, tool use, deeper reasoning and planning, memory, and external interaction.
- Additionally, investigations into AI-driven tools for science should be conceptualized accounting for the iterative and collaborative processes that define modern science and scientific-software development.
- Methods proposed for investigation should use any appropriate techniques that might be necessary to accomplish their goals, including, but not limited to, machine learning, natural-language processing, formal reasoning, instrumentation, data management, and compiler technology.
- See the solicitation for important additional information.

Research Area 3: AI Innovations for Computational Decision Support of Complex Systems

- Scientific machine learning is a core component of artificial intelligence and a computational technology that can be trained, with scientific data, to augment or automate human skills.
- The term “outer loop” is used increasingly to describe computational applications that form outer loops around a forward simulation. Examples of outer-loop applications include optimization, uncertainty quantification, inverse problems, data assimilation, and control.
- The principal focus of this research area is on the use of scientific AI/ML for intelligent automation and decision support for complex systems.
- This area is defined by PRD 6, “Intelligent Automation and Decision Support” of the Basic Research Needs for Scientific Machine Learning: Core Technologies for Artificial Intelligence report.
- Applicants must demonstrate that their proposed work is not specific to a particular complex system.
- It is expected that the proposed projects will significantly benefit from the exploration of innovative ideas or from the development of unconventional approaches.
- Proposed approaches may include innovative research with one or more key characteristics, such as asynchronous computations, mixed-precision arithmetic, automatic differentiation, compressed sensing, coupling frameworks, graph and network algorithms, randomization, Monte Carlo or Bayesian methods, probabilistic programming, or other relevant facets.
- See the solicitation for important additional information.

Research Area 4: Federated and Privacy-Preserving Machine Learning and Synthetic Data Creation

- Key challenges in training foundation and other AI models for science, are created by the reality that the data available for training is often sparse and distributed across different computing systems.
- Sometimes, when there is too little data for training, or the underlying data cannot be used or shared, more-voluminous synthetic data can be created for AI training. Alternatively, there may be sufficient data, but it might be too voluminous to be transferred to a common system for training, or it might be restricted from being aggregated to protect privacy-sensitive or proprietary data, thus requiring federated AI training techniques.
- Another major challenge is the need for improved energy efficiency, scalability, and performance in the federated training process.
- For this research area, the focus for federated learning is on the role and innovative use of federated approaches to train and fine-tune foundation and other AI models for science on massive amounts of multi-modal, distributed, and potentially privacy-sensitive datasets.
- Significant advances will be needed in asynchronous, randomized, mixed-precision, and other algorithmic developments, optimization, and numerical analyses that underpin the overall process of building and fine-tuning trustworthy foundation and other AI models for science.
- Additionally, challenges relevant to the creation and validation of foundation and other AI models for science might be addressed by innovative methods for the creation of synthetic data sets.
    - Moreover, research may investigate new tradeoffs between computation, communication, and storage, such as in terms of algorithmic entropy and overall usage costs, by exploiting the potential for deterministic regeneration of synthetic data via on-demand recomputation.
- See the solicitation for important additional information.

Research Area 5: The Co-Design of Energy-Efficient AI Algorithms and Hardware Architectures

- This research area seeks innovative approaches to energy-efficient scientific AI, including corresponding mathematical paradigms and modeling capabilities capable of predicting the resource efficiency of AI systems, potentially up to the largest scales.
- Energy-efficient AI technologies, covering both training and inference, and including current technologies as a point of comparison, should be addressed holistically, accounting for energy for parameter representation, model size, and other representational factors, plus computational energy for activations and other inherent operations in the neural networks.
- Modeling capabilities to address energy efficiency should include both the ability to estimate resource usage relatively cheaply and the ability to perform detailed simulations of next-generation AI systems at extreme scales.
- It seems useful to note that, in many respects, biological neurons are significantly more efficient than present-day computational approaches, and it has long been recognized that one factor that is likely key to this energy-efficiency gap is the inherent use of sparsity of connectivity and timing by biological neurons: neurons in the brain are "spiking" and transmit data asynchronously on an as-needed basis to a limited neighborhood.
- SNNs are not the only kind of AI approach promising significant increases in energy efficiency over modern AI systems. Hardware and algorithms, for example, neuromorphic, utilizing statistical, probabilistic, approximate, sub-threshold, photonic, cryogenic, and other hardware and software approaches remain the subject of important investigations.
- See the solicitation for important additional information.

Research Areas: Out of Scope

- Out of scope are pre-applications and applications that:
  - Fail to focus on one of the research areas described above.
  - Fail to focus on fundamental advances in AI for science;
  - Fail to describe the kinds of scientific-computing workloads targeted by the proposed investigations;
  - Provide discipline-specific and/or application-specific solutions that do not generalize to multiple applications; or
  - Focus on the development of applications or approaches for quantum computers.
- See the solicitation for important additional information.

Teaming and Award Size

- SC uses two different mechanisms to support teams of multiple institutions: collaborative applications and subawards.
  - DOE/National Nuclear Security Administration (NNSA) National Laboratories, other Federal agencies, and another Federal agency's FFRDCs, if participating in a team led by another institution, must use the collaborative application process described above and may not be proposed as subrecipients when requesting funding more than or equal to the applicable floor.
- A multi-institutional team, whether applied for as a prime applicant with subawards or as collaborative applications, is limited to a request of no more than $2,350,000 per year.
- Note that the ceiling and floor specified below apply to each institution's proposed budget inclusive of any proposed subawards.
- The ceiling and floor specified below are for total costs, both direct and indirect costs.
  - Ceiling
    - DOE National Laboratories: $2,000,000 per year
    - All other applicants: $350,000 per year
  - Floor
    - DOE National Laboratories: $250,000 per year
    - All other applicants: $100,000 per year
- Approximately six (6) to eight (8) awards are expected.
- See the solicitation for important additional information.

Pre-Applications and Applications

- Submission Deadline for Pre-Applications: March 19, 2024 at 5:00 PM Eastern
  - **A Pre-Application is required**.
- Pre-Application Response Date: April 11, 2024 at 11:59 PM Eastern
- Submission Deadline for Applications: May 21, 2024 at 11:59 PM Eastern

- For details on pre-applications and the submission procedure, see the FOA Section IV.B.2.
- The first page of the pre-application must specify at least one scientific hypothesis whose investigation motivates the proposed work, using no more than 100 words, in a box with a black border. For any hypothesis that is not itself innovative, the pre-application must describe at least one innovative insight into how the hypothesis can be investigated that may be exploited by the planned research.
- The pre-application must include a listing of senior/key personnel and a listing of individuals who should not serve as merit reviewers of a subsequent application. The list of individuals must be included as an "Additional Attachment" to your pre-application in PAMS.
  - The lists must be submitted in tabular format, preferably as Microsoft Excel (.xls or .xlsx) files.
  - For your convenience, a Collaborator Template is available at https://science.osti.gov/grants/Policy-and-Guidance/Agreement-Forms.
- Written feedback about pre-applications will be provided upon request after award selections have been announced.