ASCR Research Priorities

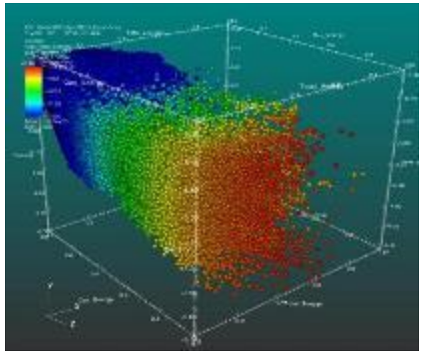
Dr. Hal Finkel

Acting Director, Computational Research and Partnerships Division,
Advanced Scientific Computing Research

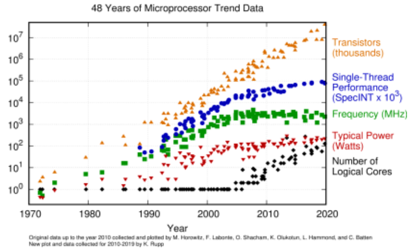
September 28, 2023

Critical Technology Trends Motivating ASCR Today

Data, Privacy, and Scientific Integrity

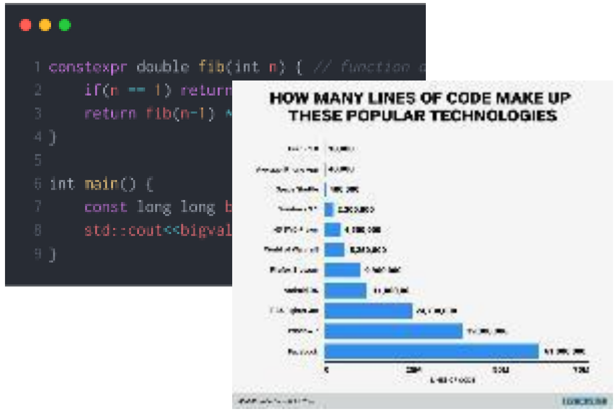


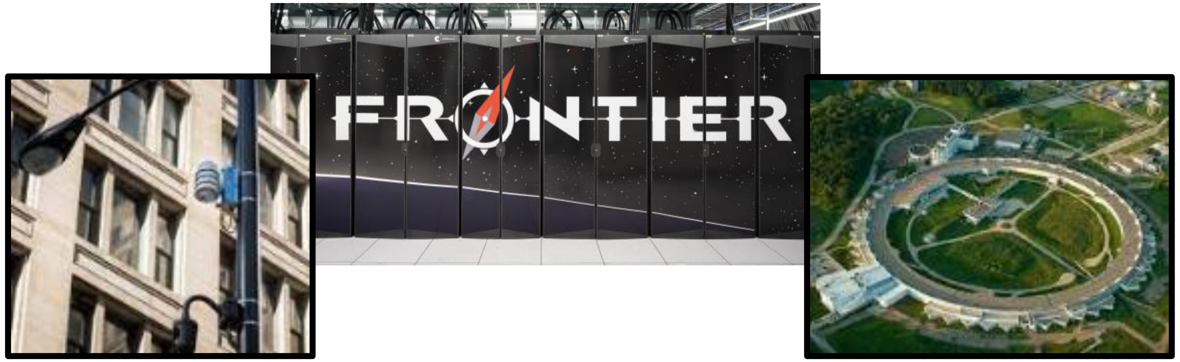
Artificial Intelligence



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

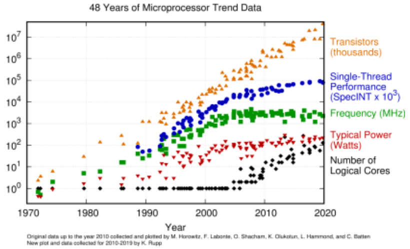
Exploding Software Complexity



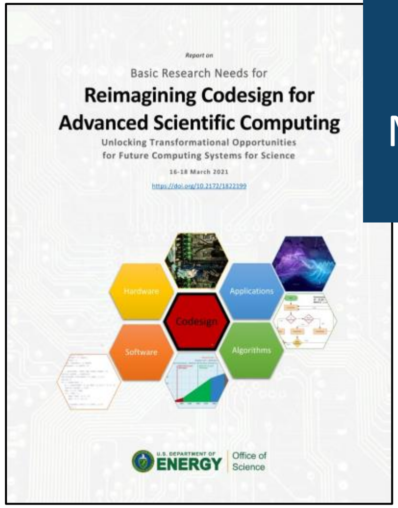


Scientific Computing and Networking: from Exascale to the Edge

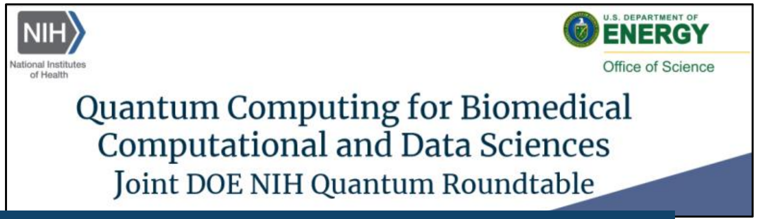
Transforming the Fundamentals of Computing



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

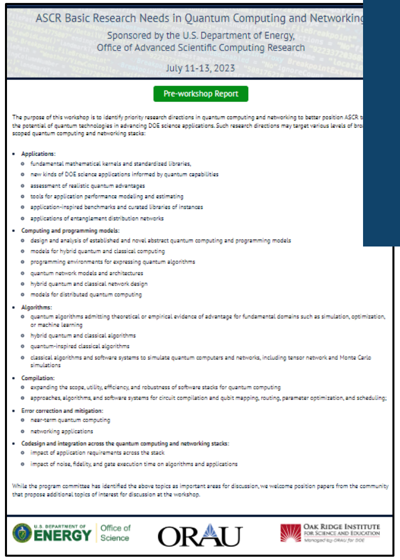


ASCR Workshop on Reimagining Codesign, March 2021: https://doi.org/10.2172/1822199



Quantum Computing for Biomedical Computational and Data Sciences Joint DOE NIH Quantum Roundtable March 2023

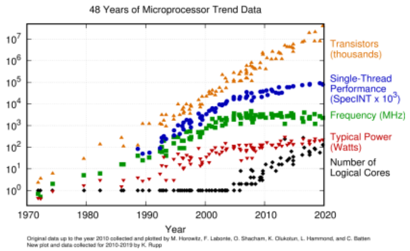
FY 2023



FY 2023

ASCR Basic Research Needs in Quantum Computing and Networking, July 2023: https://www.orau.gov/ASCR-BRN-Quantum (report forthcoming)

Transforming the Fundamentals of Computing



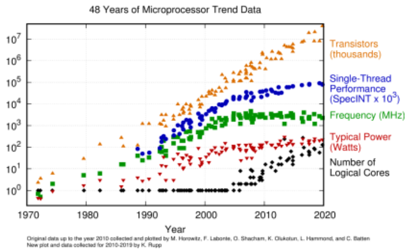
Heterogeneous, Distributed, Co-Designed, Energy-Efficient Computing

Past Solicitations (FY 2021 – 2022):

- ▲Entanglement Management and Control in Transparent Optical Quantum Networks, 2021.
- ▲Microelectronics Co-Design Research, 2021.
- ▲Quantum Internet to Accelerate Scientific Discovery, 2021.
- ▲Quantum Algorithms and Mathematical Methods, in Exploratory Research for Extreme-Scale Science (EXPRESS), 2022.
- ▲Quantum Computing at the Edge, in Exploratory Research for Extreme-Scale Science (EXPRESS), 2022.

Transforming the Fundamentals of Computing

FY 2023



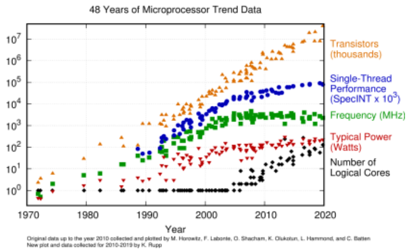
Heterogeneous, Distributed, Co-Designed, Energy-Efficient Computing

Solicitations (FY 2023):

- Accelerate Innovations in Emerging Technologies, 2023.
  - Two ASCR-funded projects, including: Thomas Jefferson National Accelerator Facility will lead a project to develop concepts for superconducting microelectronics to achieve ultra-energy-efficient computing (co-funding with NP).
- Modeling Future Supercomputing Systems, in Exploratory Research for Extreme-Scale Science (EXPRESS), 2023.
  - Four projects: projects include modeling cryogenic and photonic beyond-exascale supercomputing systems.
- Programming Techniques for Computational Physical Systems, in Exploratory Research for Extreme-Scale Science (EXPRESS), 2023.
  - Four projects: projects explore the fundamentals of analog computing in microelectronic and chemical systems toward next-generation computing and storage technologies.

Transforming the Fundamentals of Computing

FY 2023



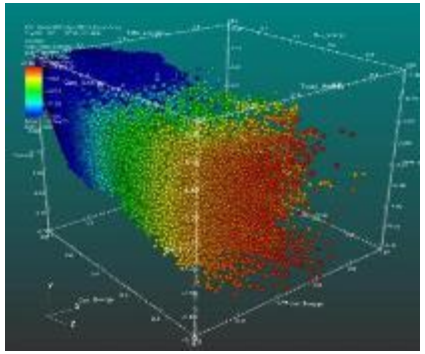
Heterogeneous, Distributed, Co-Designed, Energy-Efficient Computing

Solicitations (FY 2023):

- Quantum Algorithms across Models, in Exploratory Research for Extreme-Scale Science (EXPRESS), 2023.
  - Seven projects: projects include innovative techniques for converting quantum circuits into dynamic quantum walks and map the converted circuits onto different kinds of quantum computers.
- Quantum Testbed Pathfinder, 2023.
  - Six projects: Projects explore the limitations of the noisy, intermediate-scale quantum processors available today and aim to develop tools for assessing whether a particular quantum processor may be able to advance the frontiers of computational science even in the absence of formal error correction on the device.
- Scientific Enablers of Scalable Quantum Communications, 2023.
  - A collaborative research effort led by Argonne National Laboratory following a heterogeneous, full-stack approach in codesigning scalable quantum networks.
  - A collaborative research effort led by Oak Ridge National Laboratory developing the architecture and protocols for a performance-integrated scalable quantum internet.
  - A collaborative research effort led by Fermi National Accelerator Laboratory developing hyper-entanglement-based networking and error noise-robust correction techniques for developing advanced quantum networks for science discovery.

Empowering Science Through Data Innovations

Data, Privacy, and
Scientific Integrity

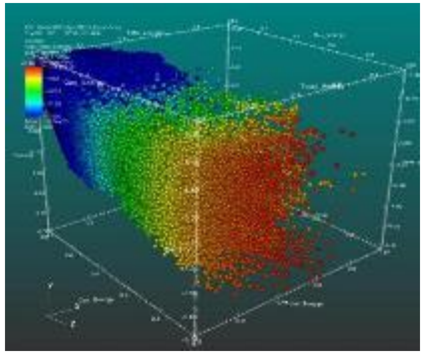


ASCR Workshop on Basic Research Needs for Management and Storage of Scientific Data, January 2022: https://doi.org/10.2172/1845707

ASCR Basic Research Needs Visualization for Scientific Discovery, Decision-Making, and Communication, January 2022: https://doi.org/10.2172/1845708 (brochure; report forthcoming)

Empowering Science Through Data Innovations

Data, Privacy, and Scientific Integrity



Past Solicitations (FY 2021 – 2022):

- Data Reduction for Science, 2021.
- Management and Storage of Scientific Data, 2022.
- Data Visualization for Scientific Discovery, Decision-Making, and Communication, 2022.

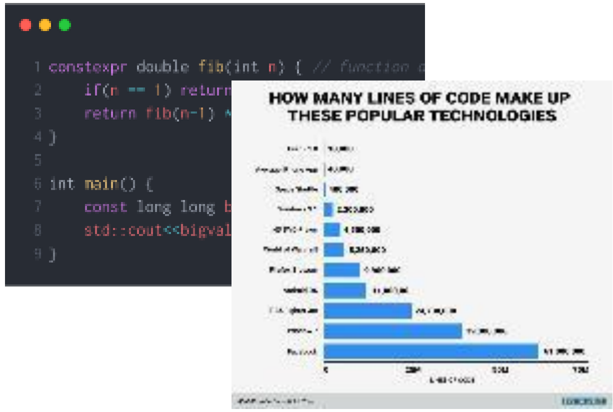
"Over the next decade, AI can help unlock world-leading simulation capabilities that can be augmented seamlessly with scalable, trusted, and efficient data-driven tools, including trusted and validated machine learning methods… but advantages in AI and machine learning can only be unlocked through powerful computing capability and commensurate amounts of good data."
- Deputy Secretary David Turk, testimony before the U.S. Senate Committee on Energy and Natural Resources

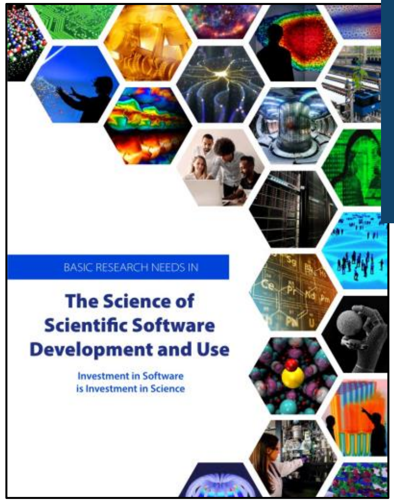
ECP EXASCALE COMPUTING PROJECT

Many projects are building on ECP investments!

Enhancing Scientific Programming

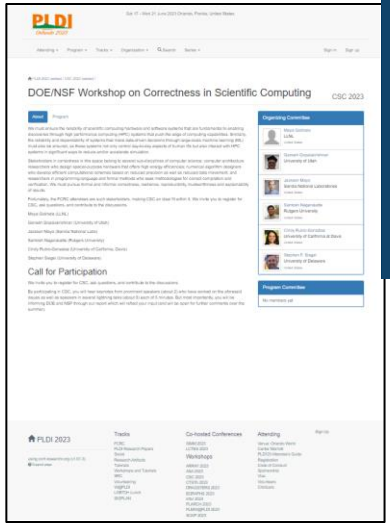
Exploding Software Complexity





ASCR Workshop on Basic Research Needs in The Science of Scientific Software Development and Use, December 2021: https://doi.org/10.2172/1846009

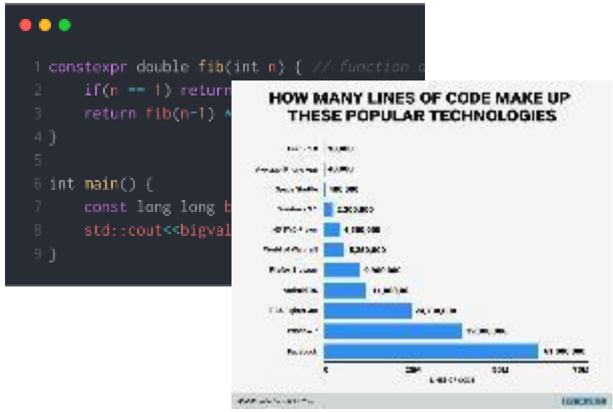
FY 2023



DOE/NSF Workshop on Correctness in Scientific Computing, June 2023: https://pldi23.sigplan.org/home/csc-2023 (report forthcoming)

Enhancing Scientific Programming

Exploding Software Complexity



Past Solicitations (FY 2021 – 2022):

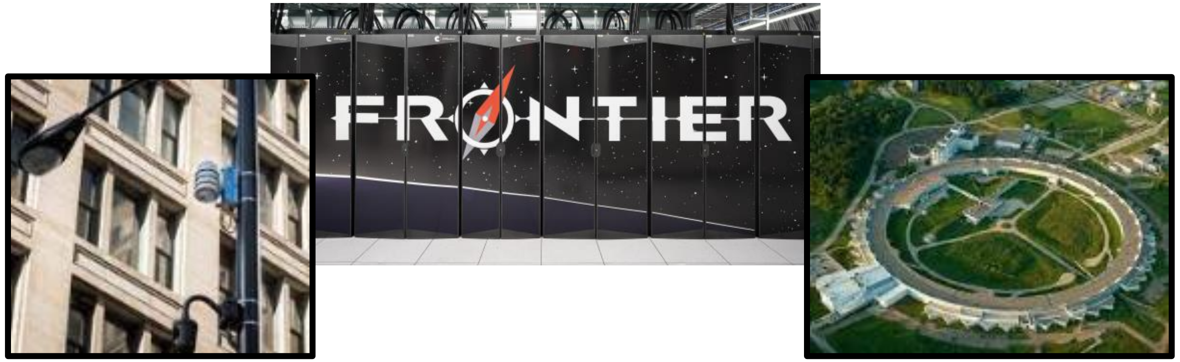
- X-STACK: Programming Environments for Scientific Computing, 2021
- Differentiable Programming, in Exploratory Research for Extreme-Scale Science (EXPRESS), 2022

Projects in this portfolio enhance the productivity of scientific programming, portable across current and future high-performance computers, and focus on challenges that will be critical as we integrate AI: Differentiable programming enables integrating modeling and simulation applications with AI training, and verification and testing methods seem essential to taking advantage of AI-generated source code.

ECP EXASCALE COMPUTING PROJECT

Many projects are building on ECP investments!

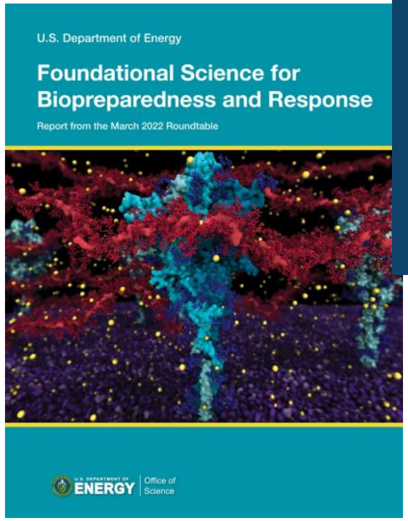
Accelerating Science from Exascale to the Edge



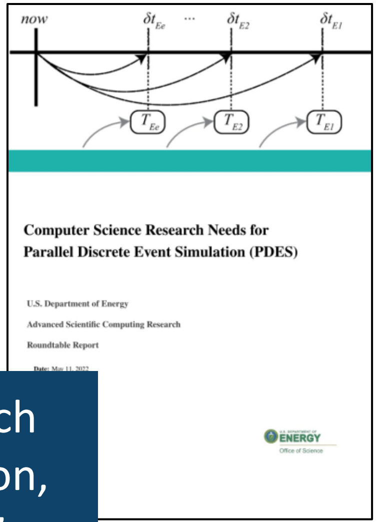
Scientific Computing and Networking: from Exascale to the Edge

Integrated Research Infrastructure Architecture Blueprint Activity, 2023: https://doi.org/10.2172/1984466

FY 2023

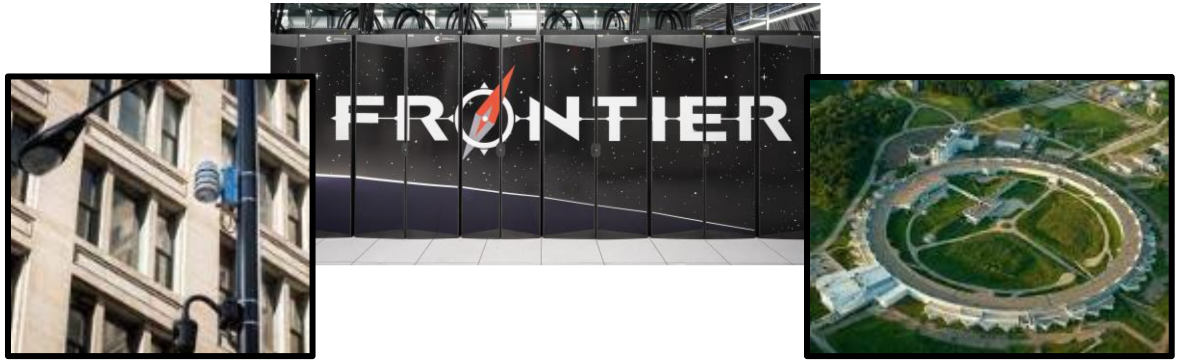


Roundtable on Foundational Science for Biopreparedness and Response, March 2022: Report available from https://science.osti.gov/ascr/Community-Resources/Program-Documents



Roundtable on Computer Science Research Needs for Parallel Discrete Event Simulation, 2022: https://doi.org/10.2172/1855247

Accelerating Science from Exascale to the Edge



Scientific Computing and Networking: from Exascale to the Edge

Past Solicitations (FY 2021 – 2022):

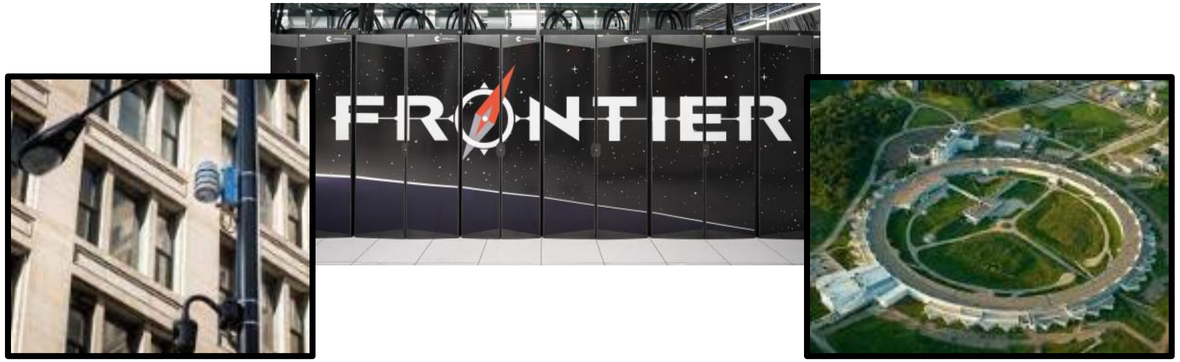
- 5G Enabled Energy Innovation Advanced Wireless Networks for Science, 2021.
- SciDAC: Partnerships in Basic Energy Sciences, 2021.

- Integrated Computational and Data Infrastructure for Scientific Discovery, 2021.
- EXPRESS: Randomized Algorithms for Extreme-Scale Science, 2021.
- SciDAC: Partnerships in Earth System Model Development, 2022.
- SciDAC: Partnership in Nuclear Energy, 2022.

ECP EXASCALE COMPUTING PROJECT

Many projects are building on ECP investments!

Accelerating Science from Exascale to the Edge



Scientific Computing and Networking: from Exascale to the Edge

Past Solicitations (FY 2021 – 2022):

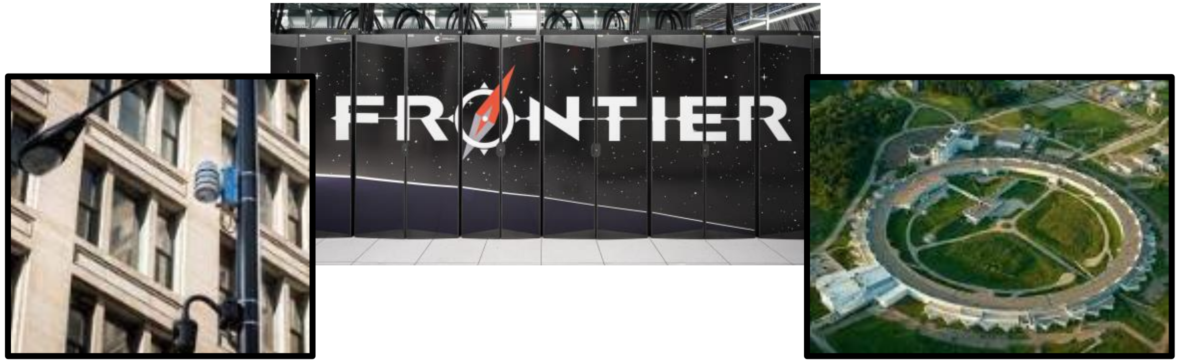
- SciDAC: High Energy Physics, 2022.
- Advancing Computer Modeling and Epidemiology for Biopreparedness and Response, 2022.
- SciDAC: Partnership in Nuclear Physics, 2022.
- Mathematical Multifaceted Integrated Capability Centers (MMICCS), 2022.
- Randomized Algorithms for Combinatorial Scientific Computing, 2022.
- Parallel Discrete-Event Simulation, in Exploratory Research for Extreme-Scale Science (EXPRESS), 2022

ECP EXASCALE COMPUTING PROJECT

Many projects are building on ECP investments!

Accelerating Science from Exascale to the Edge

FY 2023



Scientific Computing and Networking: from Exascale to the Edge

Solicitations (FY 2023):

- ▲ Energy Earthshot Research Centers, 2023.
  - [Award announcement forthcoming.]
- ▲ Science Foundations for Energy Earthshots, 2023.
  - [Award announcement forthcoming.]

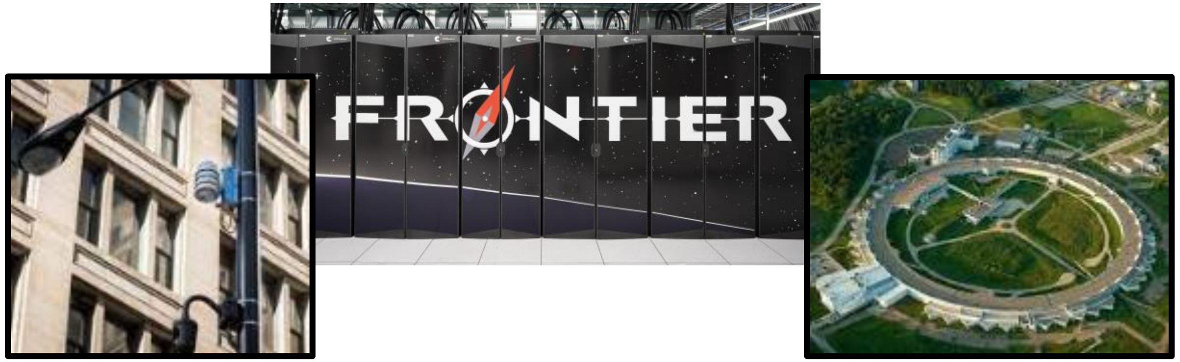
- ▲ Biopreparedness Research Virtual Environment (BRaVE), 2023.
  - Four ASCR-funded projects: Includes awards to Lawrence Berkeley National Laboratory and Oak Ridge National Laboratory will examine diverse data streams using DOE computational capabilities and AI-related tools to accurately detect the onset of emerging pandemics and help enable decision making for both healthcare and policy makers. Portfolio includes co-funding from BES and BER.

ECP EXASCALE COMPUTING PROJECT

Many projects will build on ECP investments!

Accelerating Science from Exascale to the Edge

FY 2023



Scientific Computing and Networking: from Exascale to the Edge

▲ Accelerate Innovations in Emerging Technologies, 2023.

- Two ASCR-funded projects, including: Argonne National Laboratory will lead a project to develop innovations that combine robotics, human interfaces, digital twins, and artificial intelligence to replace 80-year-old technologies currently used to produce isotopes used in medical diagnostics and treatments, research, and industrial applications (co-funding with IP).

Solicitations (FY 2023):

▲ Advanced Scientific Computing Research for DOE User Facilities, 2023.

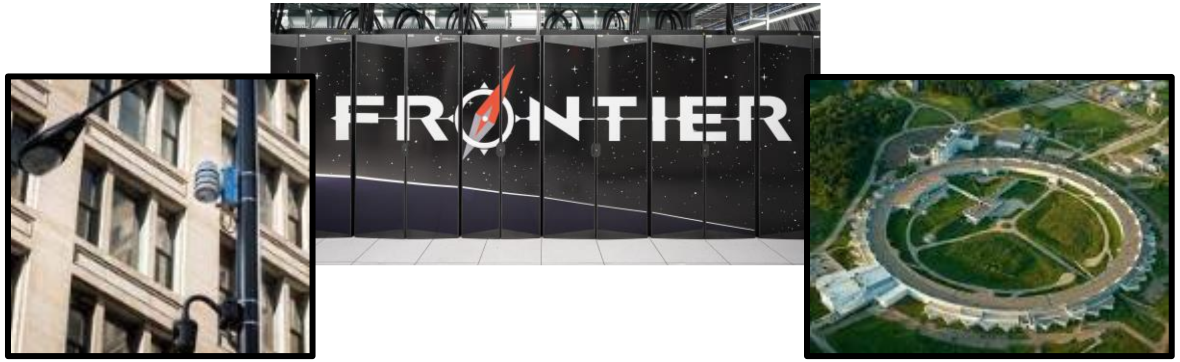
- Two ASCR-funded projects:
- A center for advanced mathematics for energy research applications led by Lawrence Berkeley National Laboratory to develop the algorithms, software, autonomous workflows, and real-time analysis at the edge for next-generation scientific user facilities.
- A research project led by Argonne National Laboratory that addresses the technical challenges and tools needed to enhance the optimization, prediction, and experimentation capabilities at science facilities.

ECP EXASCALE COMPUTING PROJECT

Many projects will build on ECP investments!

Accelerating Science from Exascale to the Edge

FY 2023



Scientific Computing and Networking: from Exascale to the Edge

- SciDAC- FES Partnerships, 2023.
  - 12 projects (co-funded by FES): Projects funded through this program will use computing resources to model plasmas, study turbulence, and use artificial intelligence to predict and solve problems like energy losses.

Solicitations (FY 2023):

- Distributed Resilient Systems, 2023.
  - Five projects, including:
  - A collaboration led by the University of Southern California is developing novel methods in swarm intelligence to distributed resource allocation.
  - A collaboration led by the University of California, Merced is improving our understanding of scalable, federated, privacy-preserving machine learning.

ECP EXASCALE COMPUTING PROJECT

Many projects will build on ECP investments!

Creating Trustworthy and Efficient AI For Science

Artificial Intelligence

Past Solicitations (FY 2021 – 2022):

- ▲ Bridge2AI And Privacy-Preserving Artificial Intelligence Research, 2021.
- ▲ Data-Intensive Scientific Machine Learning and Analysis, 2021.
- ▲ Federated Scientific Machine Learning, in Exploratory Research for Extreme-Scale Science (EXPRESS), 2022.
- ▲ Explainable Artificial Intelligence, in Exploratory Research for Extreme-Scale Science (EXPRESS), 2022.

Creating Trustworthy and Efficient AI For Science

FY 2023

Artificial Intelligence

Solicitations (FY 2023):

- Scientific Machine Learning for Complex Systems, 2023.
  - Four projects, including:
  - A collaboration led by Pacific Northwest National Laboratory, partnering with Spelman College, for quantifying uncertainties and improving predictions in atmospheric simulations and measurements.
  - A project led by Johns Hopkins University for research on the properties and behavior of additively manufactured composites in materials science and turbulent, high-speed fluid flow in aerospace engineering applications.
  - A project, led by Lawrence Berkeley National Laboratory, to develop a new generation of uncertainty quantification (UQ) tools to ensure optimal operation or accurate predictions for accelerators, reactors, climate science, and other complex systems or processes.

EXASCALE COMPUTING PROJECT

Building on ECP investments!

Growing and Diversifying Our Research Community

Solicitations (FY 2023):

FY 2023

- Early Career Research Program, 2023.
  - Ten projects, focusing on reconfigurable computing; a programming framework for graph algorithms; intelligent scheduling for heterogeneous computing; data-driven discovery of dynamic models; model reduction using deep learning; Quasi-Trefftz methods for problems governed by vector-valued partial differential equation; randomized optimization; multi-linear representations for quantum characterization, control, and computation; Markov random fields for scientific machine learning, and statistical modeling of extreme events in complex systems.
- FY 2023 Funding for Accelerated, Inclusive Research (FAIR), 2023.
  - Five projects focusing on machine-learning-based surrogate modeling for stochastic multiscale simulations (University of California, Merced and Lawrence Berkeley National Laboratory), the performance and scalability of distributed deep learning (University of Texas at El Paso and Pacific Northwest National Laboratory), storage-driven machine-learning performance models (University of South Dakota and Lawrence Berkeley National Laboratory), computational storage using data-tasks and asynchronous I/O (Illinois Institute of Technology and Argonne National Laboratory), and entanglement estimation for quantum computing (Texas Tech University and Argonne National Laboratory).
- Reaching A New Energy Sciences Workforce (RENEW), 2023.
  - Two many-institution projects focusing on innovative workforce development in quantum information science and artificial intelligence.

Small Business Innovation Research (SBIR) and Small Business Technology Transfer (STTR)

FY 2023

Many projects will build on ECP investments!

FY 2023 ASCR SBIR/STTR Phase I Awards:

Topic	Company	Title	State
Accelerating the deployment of advanced software technologies	Exabyte Inc.	Wide-Scale Adoption Through Visually-Enhanced HPC Codes (WAVE-HPC) for Semiconductor R&D	CA
Accelerating the deployment of advanced software technologies	Lifeboat LLC	Supporting Sparse Data in HDF5	IL
Accelerating the deployment of advanced software technologies	Veracity Nuclear, LLC	Accelerating the Deployment of Advanced Software Technologies: Deployment of ASCR-Funded Software	TN
Accelerating the deployment of advanced software technologies	Advanced Cooling Technologies, Inc.	Semiconductor Device Simulation Software with Monte Carlo Based Thermal Transport Modeling	PA
Accelerating the deployment of advanced software technologies	Coreform LLC	Advanced Tire Tread Simulation Through Adaptive Isogeometric Analysis	VT
Accelerating the deployment of advanced software technologies	Equity Engineering Group, Inc., The	SimPulse: Scalable Hydraulic Transients in 21st Century Piping Systems	OH
Accelerating the deployment of advanced software technologies	IERUS Technologies, Inc.	Multiphysics Motor Design Software Package Using MFEM	AL
Accelerating the deployment of advanced software technologies	Osazda Energy	An Optimization-Based Design Ecosystem Targeting Performance, Reliability, and Stability of Photovoltaic Modules in Solar Energy Market	NM
Accelerating the deployment of advanced software technologies	Protection Engineering Consultants	Network Risk Assessment Toolkit (NetRAT)	TX
Technology to facilitate the use of near-term quantum computing hardware	Atlantic Quantum Corp	Software for Automatic Control, Calibration and Validation of Quantum Processors	MA
Technology to facilitate the use of near-term quantum computing hardware	HighRI Optics, Inc	Highly Efficient Low Loss Fiber-Chip Light Coupling for Quantum Networks	CA
Technology to facilitate the use of near-term quantum computing hardware	memQ Inc.	Low-Loss Thermomechanically Stable Packaging for Cryogenic Quantum Photonic Network Devices	IL
Technology to facilitate the use of near-term quantum computing hardware	Physical Sciences Inc.	Volumetric Fiber-to-Chip Mode Converters	MA

Small Business Innovation Research (SBIR) and Small Business Technology Transfer (STTR)

FY 2023

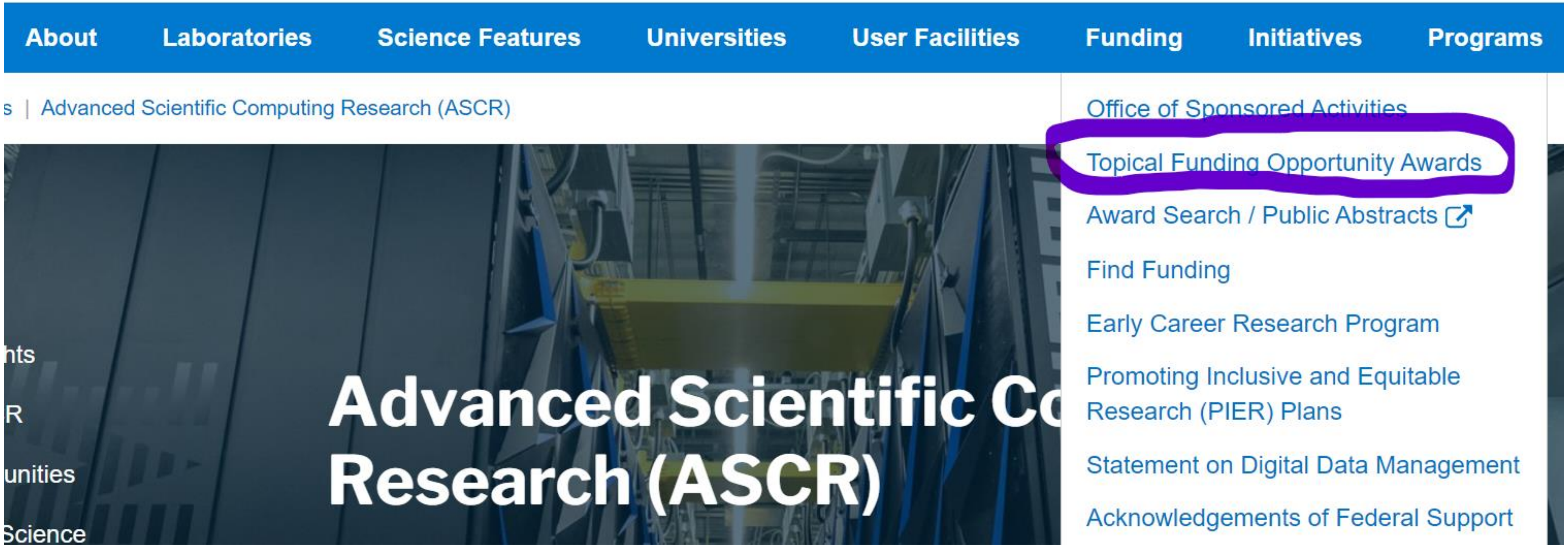
FY 2023 ASCR SBIR/STTR Phase II Awards:

Topic	Company	Title	State
Accelerating the deployment of advanced software technologies	Lifeboat LLC	Toward multi-threaded concurrency in HDF5.	IL
Accelerating the deployment of advanced software technologies	RadiaSoft LLC	Efficient Monte Carlo Simulations in the Cloud	CO
Accelerating the deployment of advanced software technologies	X-ScaleSolutions	SMART-PETSc: Smart Middleware for Accelerating PETSc	OH
Accelerating the deployment of advanced software technologies	Coreform LLC	Integrating MFEM for Commercial IGA Simulation Acceleration	UT
Accelerating the deployment of advanced software technologies	ParaTools, Inc	E4S: Extreme-Scale Scientific Software Stack for Commercial Clouds	OR
Transparent optical quantum network technologies	Physical Sciences Inc.	Optical Quantum Network Time-Frequency Multiplexer	MA
Technology to facilitate the use of near-term quantum computing hardware	Error Corp.	Functional Gradient-Based Geometric Curve Synthesis for Dynamic Quantum Error Suppression	MD

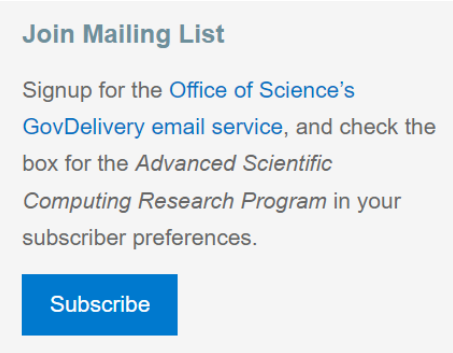
ECP EXASCALE COMPUTING PROJECT

Many projects are building on ECP investments!

Award Lists – A New Website Location



Award lists are now posted to https://science.osti.gov/Funding-Opportunities/Award along with other awards from the Office of Science. To receive award and solicitation announcements, and other ASCR-related news, signup for the Office of Science's GovDelivery email service, and check the box for the Advanced Scientific Computing Research Program in your subscriber preferences:



Anticipated Solicitations in FY 2024

- Compared to FY 2023, expect a smaller number of larger, more-broadly-scoped solicitations driving innovation across ASCR's research community.
- In appropriate areas, ASCR will expand its strategy of solicitating longer-term projects and, in most areas, encouraging partnerships between DOE National Laboratories, academic institutions, and industry.
- ASCR will continue to seek opportunities to expand the set of institutions represented in our portfolio and encourages our entire community to assist in this process by actively exploring potential collaborations with a diverse set of potential partners.
- Areas of interest include, but are not limited to:
  - Applied mathematics and computer science targeting quantum computing across the full software stack.
  - Applied mathematics and computer science focused on key topics in AI for Science, including scientific foundation models, decision support for complex systems, privacy-preserving federated AI systems, AI for digital twins, and AI for scientific programming.
  - Microelectronics co-design combining innovation in materials, devices, systems, architectures, algorithms, and software (including through Microelectronics Research Centers).
  - Correctness for scientific computing, data reduction, new visualization and collaboration paradigms, parallel discrete-event simulation, neuromorphic computing, and advanced wireless for science.
  - Continued evolution of the scientific software ecosystem enabling community participation in exascale innovation, adoption of AI techniques, and accelerated research productivity.

ASCR Welcomes Todd Munson

ASCR welcomes Todd Munson, now on part-time detail from Argonne National Laboratory. Todd received his Ph.D. from the University of Wisconsin-Madison in 2000. He is a Senior Computational Scientist in the Mathematics and Computer Science Division at Argonne National Laboratory and a Senior Scientist for the Consortium for Advanced Science and Engineering at the University of Chicago. Todd is Deputy Director for the ECP Co-Design Center for Online Data Analysis and Reduction (CODAR) and lead developer for the Toolkit for Advanced Optimization (TAO).
https://www.anl.gov/profile/todd-munson

Highlights

(The focus for this meeting will be trustworthy and efficient AI)

Scalable Transformers on Frontier for Real-Time Experiment Steering

Scientific Achievement

ORNL developed a scalable transformer on OLCF Frontier for real-time decision-making in neutron diffraction experiments at the TOPAZ beamline of SNS. This work:

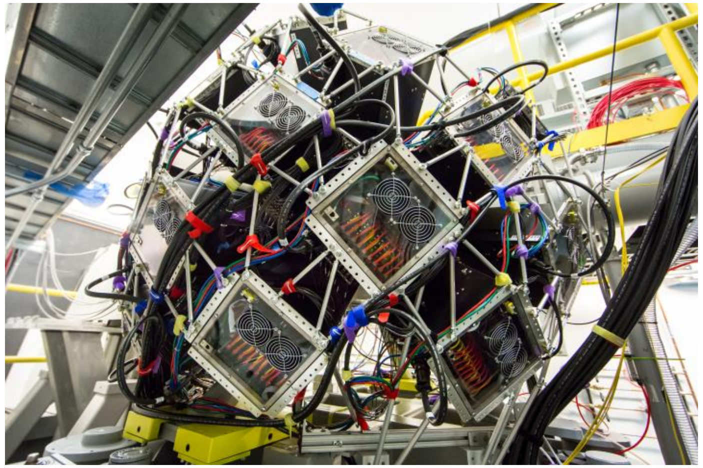
- Develops a stochastic process model for the time-of-flight neutron scattering data and exploits a temporal fusion transformer to **help reduce the experiment time**.
- Demonstrates outstanding scalability of the ML model on Frontier, which is necessary to synchronize neutron diffraction experiments, data analysis, and decision making.

Significance and Impact

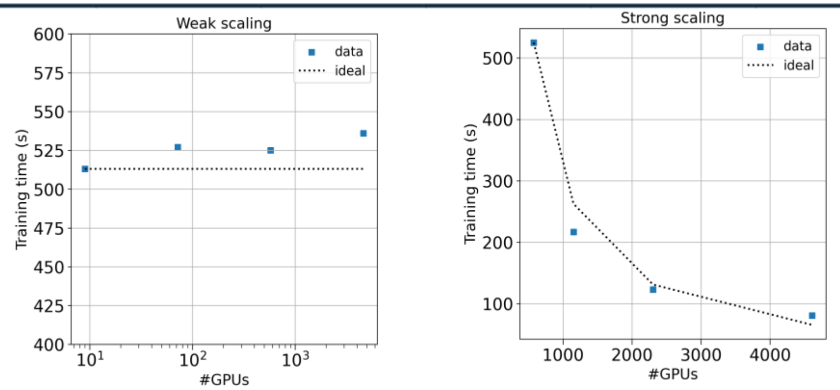
- The ML algorithm could help neutron scientists to **reduce the over-counting beamtime by around 30%** at TOPAZ, while achieving the similar data quality.
- This effort proves the concept of connecting BES's neutron facilities and ASCR's HPC facilities through AI/ML, **forming an integrated research infrastructure**.

Technical Approach

- The developed stochastic process model provides a novel and effective approach to describe the time-of-flight neutron scattering data.
- The hierarchical parallelization approach effectively uses ~60% of Frontier's computing power to keep up with the neutron experiment speed.



A single-crystal diffractometer on the TOPAZ beamline at SNS



Outstanding weak and strong scalability on Frontier with up to 4608 GPUs.

PI : Guannan Zhang (ORNL); ASCR Program: Data-Intensive Scientific Machine Learning and Analysis; ASCR PM: Steve Lee
Publication: J. Yin, S. Liu, V. Reshniak, X. Wang, and G. Zhang, *A scalable transformer model for real-time decision making in neutron scattering experiments*, ***Journal of Machine Learning for Modelling and Computing***, Vol 4 (1), pp. 95-107, 2023

SuperNeuro: An Accelerated Neuromorphic Computing Simulator

Scientific Achievement

ORNL scientists have developed SuperNeuro, the world's fastest simulator for neuromorphic computing. It was designed for speed and scalability, and is capable of running **300 times faster** than its competitors, garnering the team the **2023 R&D 100 Award in the Software/Services Category**.

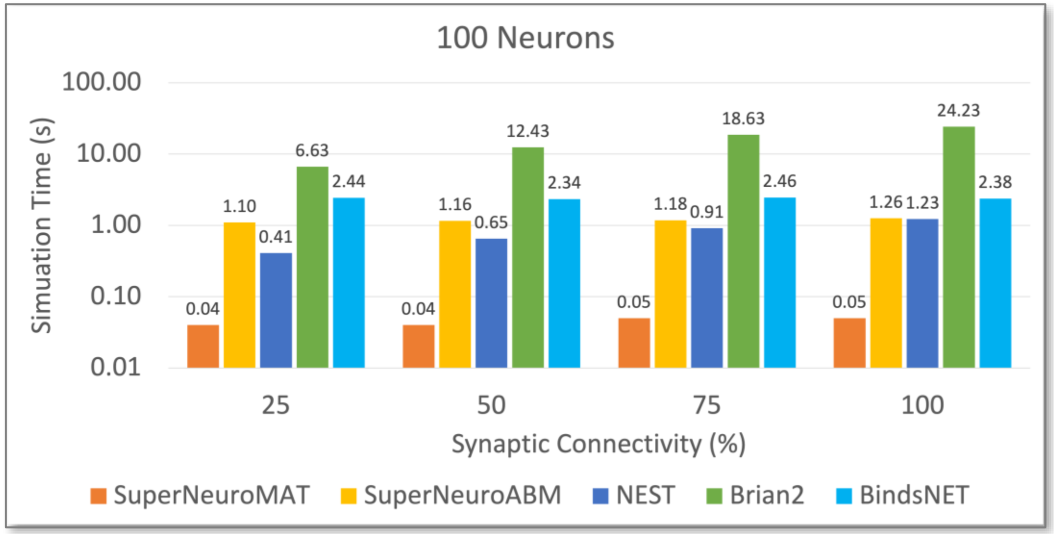
Significance and Impact

Neuromorphic architectures have the potential to increase computing power and efficiency, as well as advance AI applications. SuperNeuro provides an indispensable capability for this effort via the leveraging of GPU computing to provide superior performance for neuroscience, increased adaptability, spiking neural networks (SNNs), and general-purpose computing workloads.

Technical Approach

Two novel approaches used: matrix computation (MAT) and agent-based modeling (ABM).

- MAT Mode: Homogeneous simulations, built-in learning, CPU execution
- ABM Mode: Heterogeneous simulations, GPU acceleration



Simulating 100 neurons on 5 neuromorphic simulators with 4 different synaptic connectivites. SuperNeuroMAT performs 300 times faster than other neuromorphic simulators.

https://github.com/ORNL/superneuromat

OAK RIDGE National Laboratory

PI(s): Prasanna Date, Chathika Gunaratne, Shruti Kulkarni, Robert Patton, Mark Coletti, and Thomas Potok
Collaborating Institutions: Oak Ridge National Laboratory
ASCR Program: Neuromorphic Computing for Accelerating Scientific Discovery
ASCR PM: Robinson Pino
Publication(s) for this work: Date, Prasanna, Chathika Gunaratne, Shruti R. Kulkarni, Robert Patton, Mark Coletti, and Thomas Potok. "SuperNeuro: A Fast and Scalable Simulator for Neuromorphic Computing." In Proceedings of the 2023 International Conference on Neuromorphic Systems, pp. 1-4. 2023.

2023 R&D 100 WINNER

SUPERNEURO
ACCELERATING NEUROMORPHIC COMPUTING

Privacy-Preserving Federated Learning as a Service using APPFL

Scientific Achievement

APPFL as a Service (APPFLaaS) enables end-to-end secure and privacy-preserving federated learning. Using APPFL and Globus services, the service provides supervised learning of a model on distributed sensitive datasets while preserving data privacy.

Significance and Impact

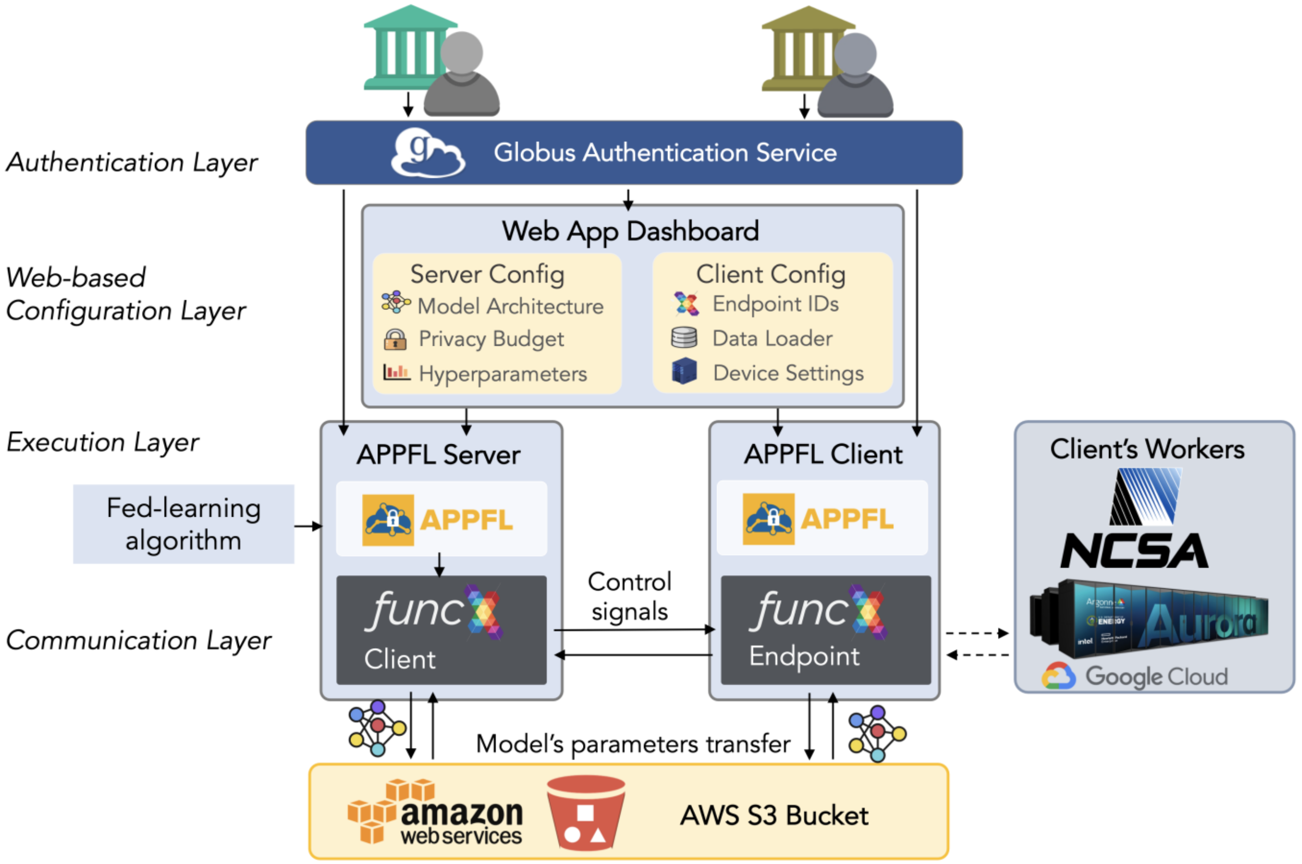
APPFLaaS will enable secure collaborations across countries and institutions while addressing the privacy and data shift challenges in many DOE applications (e.g., scientific machine learning, critical infrastructure) leading to fair and trust-worthy AI models

Research Details

- Integration with Globus Auth and Compute enables secure access controls and integration with heterogenous compute resources
- Novel distributed optimization algorithms with differential privacy result in better convergence and learning performance
- In collaboration with medical institutions, APPFL is used to train various ML models for disease prognosis, diagnosis and treatment planning
- APPFL used for federated control of power system operations maintaining data privacy against an adversary
- APPFLaaS provides comprehensive report for each federation learning experiment including training logs, hyperparameters, validation results, training metrics and Tensorboard visualization

https://github.com/APPFL/APPFL

Argonne NATIONAL LABORATORY



PI(s): Ravi Madduri and Kibaek Kim; Argonne National Laboratory
ASCR Program: Bridge2AI And Privacy-Preserving Artificial Intelligence Research
ASCR PM: Steven Lee
Publication(s) for this work: Ryu, Kim, Kim, Madduri. "APPFL: Open-Source Software Framework for Privacy-Preserving Federated Learning" 2022 IEEE IPDPS Workshop

Correctness of Autodiff on Machine-Representable Inputs

Scientific Achievement

Automatic Differentiation (AD), “Autodiff”, is a family of algorithms for computing derivatives of programs. Recent work has shown that AD over **real-valued** inputs is almost always correct in a mathematically precise sense. However, programs work with **machine-representable** floating-point numbers, not reals. We show that AD is almost always correct for neural networks with floating point arguments, and that AD is correct more often if network layers use bias parameters.

Significance and Impact

Our work provides the first theoretical understanding of when AD can be incorrect on machine-representable inputs. Our results also show that not all neural networks are the same with respect to AD and that certain network designs will give more reliable results from AD than others.

Technical Approach

- We study the set of machine-representable inputs where AD can be incorrect.
- We measure the density of this set over all machine-representable inputs.
- We derive simple conditions that decide whether a given input is in the set or not.

program P, input x

↓

TensorFlow PyTorch …

↓

computed derivative $\nabla^{\mathrm{AD}} P(x)$

Our main result:

If P represents a practical neural network, then

$$\nabla^{\mathrm{AD}} P(x) = \nabla P(x)$$

for **most machine-representable** inputs x.

A main result of our work: AD is correct for most machine-representable inputs, when applied to programs denoting practical neural networks.

PI(s): Alex Aiken, SLAC
Collaborating Institutions: Stanford University
ASCR Program: EXPRESS, Differentiable Programming
ASCR PM: Hal Finkel
Publication(s) for this work: Lee, et al., “On the correctness of automatic differentiation for neural networks with machine-representable parameters,” *International Conference on Machine Learning (ICML)*, 2023.

SyReNN: A Tool for Analyzing Deep Neural Networks

The Science

Deep Neural Networks (DNNs) are used in safety-critical applications such as image recognition and autonomous vehicle controllers. However, DNNs have been shown to be vulnerable to attacks and buggy behavior.

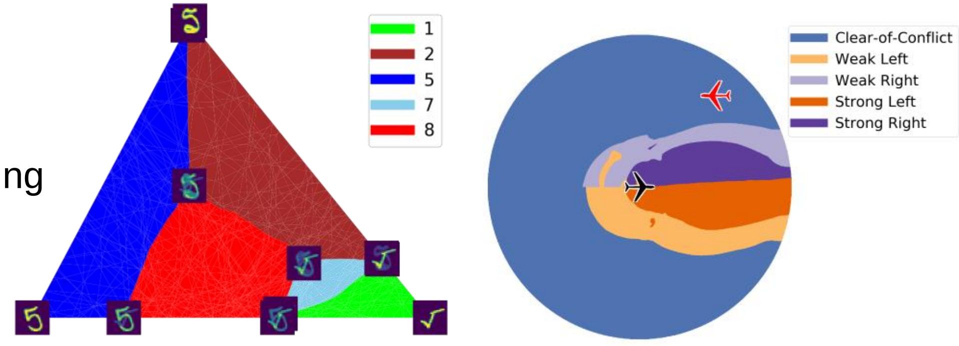
This paper introduces **SyReNN**, a tool for understanding and analyzing a DNN by computing its *symbolic representation*, which decomposes the DNN into linear functions.

The key insight to obtain scalability and precision is to restrict the analysis to *low-dimensional subsets* of the input space and design a novel algorithm that exploits *GPU parallelism*. (GPU: Graphics Processing Unit)

The Impact

SyReNN has been used in a variety of applications:

- Visualizing the *precision decision boundaries* of a DNN, enabling a human to understand its behavior. Prior to SyReNN one could only approximate decision boundaries via sampling.
- Enable *provable repair* of DNNs to precisely correct its behavior on an infinite set of points, which was not possible prior to SyReNN.
- Precise computation of Integrated Gradients (IG), a state-of-the-art measure to understand DNNs by determining which input dimensions (e.g., pixels in an image) were the most important in the final classification produced by the network. Prior to SyReNN, only imprecise approximations of IG were possible. This improvement is been incorporated by Google, the original inventors of IG.



Precise visualization of decision boundaries computed using **SyReNN** for the *(a) MNIST digit recognition network* and *(b) ACAS Xu network*. This is not a plot interpolating between finitely-many sampled points, instead **SyReNN** was used to quickly and *precisely* compute the exact decision boundaries.

https://github.com/95616ARG/SyReNN

PI(s) : Aditya V. Thakur (University of California, Davis)
ASCR Program: Early Career Research Program
ASCR PM: Hal Finkel
Publication(s) for this work: Sotoudeh, M., Tao, Z. & Thakur, A.V. SyReNN: A tool for analyzing deep neural networks. *Int J Software Tools Technol Transfer (STTT)* **25**, 145–165 (2023).

Dehallucination of LLMs for High-Level Planning

Scientific Achievement

- Large language models can generate plans for solving high-level planning problems, such as the operation of robots in DOE national laboratories.
- While the plans may appear to be of high quality, it is not uncommon for the produced plans to contain actions that cannot be executed in reality.
- We have developed a framework that mitigates hallucinations (generated errors) in LLM generated plans.

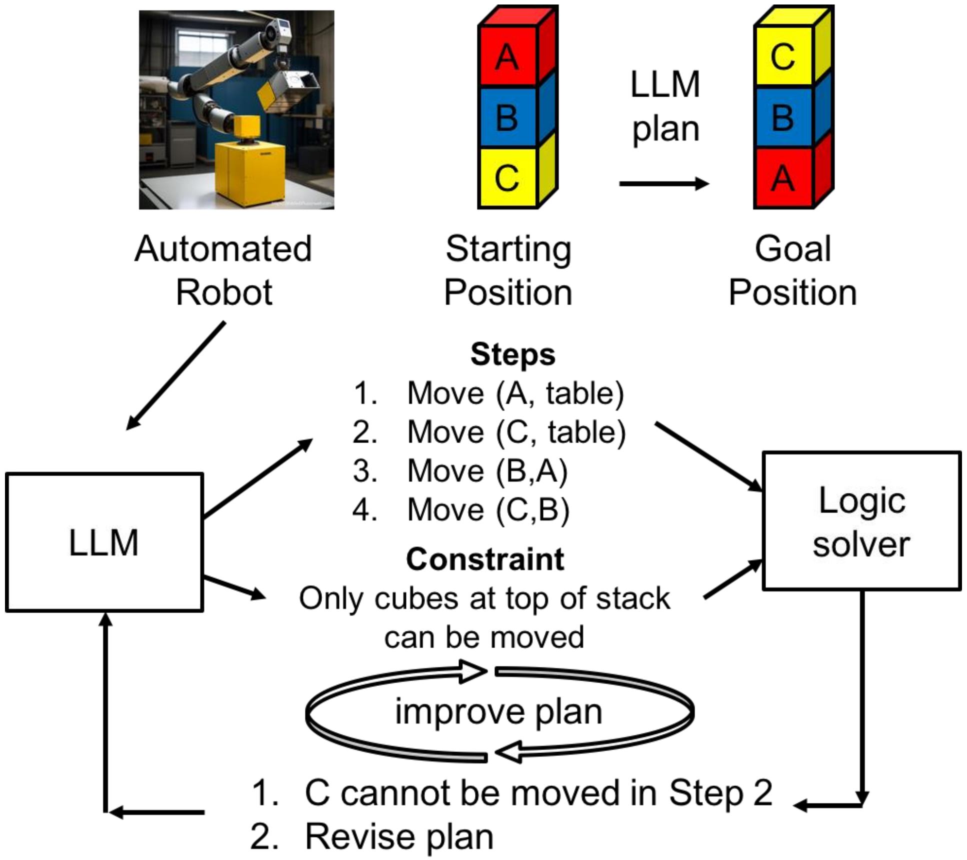
Significance and Impact

The project provides a solution to specifying scientific problems in natural languages (or text) while solving them using neuro symbolic methods. This is a step towards lowering technical barriers for future engineers and scientists.

Technical Approach

- The code generation capabilities of the LLM is used to specify logical constraints that every generated plan must satisfy.
- A solver is used to automatically check the adherence to the constraints and provide feedback to the AI model regarding unsatisfied constraints.
- The feedback allows the LLM to generate a new provably correct plan.

PI(s)/Facility Lead(s): Rickard Ewetz, Sumit Kumar Jha; University of Central Florida
Collaborating Institutions: Florida International University
ASCR Program: EXPRESS, Explainable AI
ASCR PM: Margaret Lentz
Publication(s) for this work: S. Jha, et al., "Counterexample Guided Inductive Synthesis Using Large Language Models and Satisfiability Solving," MILCOM, November, (2023). (to appear).



The LLM generates a high-level plan for moving the starting position to the goal position. The LLM also generates mathematical constrains describing how cubes are allowed to be moved. The plan and the constraints are fed into a logic solver, which determines that a constraint is violated in step 2. The C cube is attempted to be moved while cube B is on top. The solver provides feedback to the LLM why the plan is infeasible such that a new legal plan can be generated.