Applied Math and Computer Science Research Portfolio for Next-Gen Integrated Scientific Experimental Facilities (NEXT-SCI)

Kerstin Kleese van Dam, Director, Computational Science Initiative (CSI)
Brookhaven National Laboratory
631-344-6019, kleese@bnl.gov
Announcement Number: LAB 24-3210

Senior/Key Personnel	
Institution	**Name**
Brookhaven National Laboratory (BNL)	**Kerstin Kleese van Dam**; Stuart Campbell; Adolfy Hoisie; Jin Huang; Shantenu Jha; Christopher Kelly; Meifeng Lin; Vanessa Lopez-Marrero; Xiaoning Qian; Nathan Urban; Shinjae Yoo; Byung-Jun Yoon
Thomas Jefferson National Accelerator Facility (JLab)	**Malachi Schram**
Princeton Plasma Physics Laboratory (PPPL)	**Ammar Hakim**; Randy Churchill
Oak Ridge National Laboratory (ORNL)	**Rafael Ferreira da Silva**; Ana Gainaru; Frederic Suter
New Jersey Institute of Technology (NJIT)	**Jing Li**
University of Delaware (UD)	**Rudolf Eigenmann**

Research Portfolio Vision

Brookhaven Lab is home to five of 28 distinct Department of Energy, Office of Science (DOE-SC) scientific research user facilities. In addition, it hosts the NASA Space Radiation Facility and an Isotope Production facility and supports the ATLAS and Belle II experiments. DOE's **large-scale experimental user facilities** are dynamic resources that propel scientific discovery. Their **operation is characterized by real-time situational analysis and steering in a high-consequence decision-making setting with extremely limited energy available to power such processes**. Perhaps unsurprisingly, many science user facilities currently operate as internally focused **islands of scientific discovery**, while **the typical procedures for engaging these sophisticated experimental sites remain decidedly "low tech."** However, artificial intelligence (AI)- and numerical-modeling-aided **generative design campaigns now are creating thousands** of possible options and optimal **objective-driven experimental protocols** for experimental test and validation. To remain at the forefront of scientific discovery, **both the inner loop facility optimization and outer loop experimental design need to be effectively connected and integrated**. Solutions will rely heavily on a connected compute and data environment of highly heterogenous computing hardware from high-performance computing (HPC) to edge and embedded technologies, while **energy-efficient hardware and software also will be paramount**.

Project Objectives

Through its applied math and computer science research, BNL's Advanced Scientific Computing Research portfolio will address several application challenges in the next four years, including:

- Optimized design of autonomous real-time operation and control of large-scale, complex experimental facilities that is adaptive and open to external objective-driven requests within a stringent energy envelope.
- Secure and safe decision making in high-consequence environments.
- Integration of an autonomous experimental user facility in external generative design loops.

Technical Approach

To support this portfolio vision, BNL has grouped its research thrusts, corresponding to DOE's assigned categories of *Applied Mathematics*, *Computer Science*, and *Advanced Computing Technologies and Testbeds*, under six distinct topics.

1. Near-facility, Composable, Experimental Science-tailored Testbeds

BNL hosts some of the largest experimental facilities and serves as a leading center for data storage and analysis that originate from a multitude of global experiments. That said, responding to the sheer complexity required for characterization, optimization, codesign, control, and optimal design of experiments for facilities is a tall order of highest scientific magnitude. Recognizing this challenge, BNL invested in a data-science-oriented testbed facility, the **Advanced Computing Lab** (ACL), to enable codesign of experimental facilities and workflows, as well as provide a nexus for developing state-of-the-art technologies via active integration of computer science and applied mathematics methods with tools, including AI. The ACL is unique in the DOE complex, affording the ability to ingest actual data from diverse experiments—at the Center for Functional Nanomaterials for microscopy, National Synchrotron Light Source II, and soon Relativistic Heavy Ion Collider—and exposing these data and computation challenges to advanced technologies that can accelerate science conducted in DOE experimental facilities. ACL offers a true collaborative space for national labs, academia, and industry, where we codesign essential, multi-use technologies and methods expressly for experimental science that otherwise may not emerge from commercial sources. We aim to build unique research and development teams and engage hardware innovations provided from industry (current and past ACL partners include NVIDIA, IBM, DDN, HPE, Lightmatter, and Tenstorrent). Some specific technologies under investigation relate to memory, storage, and data-intensive technologies. In ***Advanced Computing Technology and Testbed Thrust 1.1***, we propose using the ACL collaboratively to: 1) capture and standardize key workflows from experimental facilities (data from instruments); 2) explore "composability" in the codesign of streaming data sources and storage for testing in-network AI and HPC hardware from the edge to the extreme; 3) facilitate research to improve energy efficiency and increase reliability for scientific instruments and workflows; and 4) test, codesign, and deploy customized scientific hardware/firmware and novel industrial AI hardware. In addition to the computer scientists and applied mathematicians from BNL's CSI group, the proposed team features key experimental scientists and detector/facility designers. Moreover, the ACL testbed not only will support other BNL research areas, but it will function as a true facility open to the entire DOE research community and beyond, e.g., uses involving national security or applied energy.

2. Portable Middleware Systems that Support *Learning Everywhere and Every Scale*

The *IRI* and *AI for Science* missions will require scalable, composable, and portable middleware systems to support "learning everywhere and every scale" in connection with large-scale experimental facilities. Specifically, the middleware must support heterogeneous computing and data infrastructure and diverse data and computing requirements while being responsive to technology trends and evolving application requirements. This necessitates an ambitious and radical re-architecting and design of future middleware systems and capabilities, which will be the focus of ***Computer Science Thrust 2.1***. We will research novel, multi-level active learning algorithms to plan, map, and schedule heterogeneous workflows on heterogeneous computing platforms, accounting for diverse data and compute requirements. ***Computer Science Thrust 2.2*** will develop a workflow runtime system to coordinate the dynamic execution of workflow with data/compute services and tasks, dynamically controlling concurrency between services and tasks and ratio/partitioning/colocation of service/tasks at runtime. Performance metrics will include latency request/response, resource utilization, and start/stop overheads. ***Computer Science Thrust 2.3*** will use system design for an application-defined network fabric to enable arbitrary communication among workflow tasks and middleware components, moving the design of communication and coordination protocols from the task level to the workflow application level.

3. Safe, Secure, and Trustworthy Composable, Energy-efficient Real-time Data Analytics

To enable energy-efficient real-time data analytics, it is essential to introduce approximated computing on hardware acceleration with numerical library abstraction. However, these methods must be accompanied by trust-building measures. Our current approximation methods have delivered a factor of 1000 speedup for streaming data analysis, but new data rates at hundreds of TB/s+ require additional breakthroughs. In ***Computer Science Thrust 3.1***, we will research novel randomized algorithms (sketching, projection, sampling) and methods for uncertainty quantification (UQ) and optimal decision theory on hardware accelerators (in-memory/neuromorphic/quantum/accelerators beyond CPU and GPUs). This will be combined with approaches that exploit the scientific applications' physical constraints. Too, the ***Applied Math Thrust 3.2*** will cultivate methods to implement UQ propagation from the randomized algorithms to downstream tasks efficiently. This will focus on parametric uncertainties and address functional uncertainties about a system's underlying structure. We intend to research a combination of functional UQ methods and dimension reduction techniques and develop tractable and fast approximations to this infinite-dimensional UQ problem.

4. Foundations for Multiscale, Complex, Safe, and Secure Digital Twins to Steer Autonomous Facilities and Experiments

The success of predictive digital twins for autonomous facilities depends highly on numerical methods that enable modeling and simulation (ModSim) of quantities of interest along with their uncertainties. In ***Applied Math Thrust 4.1***, we propose to research numerical methods that enable predictive capabilities of digital twins. These will include expanding state-of-the-art research in numerical methods for (optimal) measure transport. We will investigate methods inclusive of (although not exclusively for) cases when data are limited as such scenarios are common. Research into the suitability of current or new numerical optimization techniques for the resulting measure transport optimization problems also will be required. Digital twins often need to simulate realistic system dynamics and respond to initial and boundary conditions. Detailed dynamical information can be difficult to emulate. In ***Applied Math Thrust 4.2***, we propose a highly (sample/energy) efficient formulation of digital twins that combines system identification, which learns and quantifies uncertainties in dynamics, with coarse-grained or reduced-order models that efficiently predict system responses. It also will permit exploration of "structural" system uncertainties that normally are too difficult within the paradigm of computer model calibration and experimental data tuning. Meanwhile, ***Multifaceted Thrust 4.3*** will create an integrated framework for digital twins capable of providing real-time feedback within a given energy envelop and time constraint. Thus, efficient algorithms and programming models exploiting novel architectures will be needed. This thrust will center on basic research of programmability and portability of dataflow architectures, computational storage, and memory-centric architectures in an *in situ* or near-facility processing environment. Algorithms to exploit the specific architecture features will be developed with an emphasis on applications to digital twins.

5. Optimal Experimental and Facility Design, Decision Making, and Operation Under Uncertainty for High-consequence Environments

Decision-theoretic, uncertainty-aware optimal experimental design (OED) provides a powerful means to prioritize the most valuable experiments while expending the fewest resources (time, cost, etc.). For high-consequence experiments and facilities, multiple operational objectives often must be considered. Decision making during the codesign procedure must have certified robustness to reduce hazardous risk and prevent incurring costs. In ***Applied Math Thrust 5.1***, we propose to accelerate OED calculations through both algorithmic improvements and HPC, greatly decreasing computational and energy requirements and enabling time-bound OED applications, e.g., real-time decision making. These improvements will be implemented in a new software library for efficient OED that includes integration with differentiable and probabilistic programming, potentially leading to a new domain-specific language for OED. HPC acceleration will exploit the often embarrassingly parallel nature of the problem via parallel/distributed and heterogeneous computing to demonstrate strong scaling on DOE exascale applications. ***Applied Math Thrust 5.2*** will develop decision-theoretic techniques to optimize objective-driven goals (energy efficiency, cost, throughput, scientific information gain) constrained by available resources (computational, experimental, human, or physical). This will include methods for sequential decision making over dynamic process graphs, e.g., choices of which computational or lab experiment to conduct with serial dependencies and opportunities for parallel execution. ***Multifaceted Thrust 5.3*** will develop a new **decision-oriented visualization** framework, **VISTA** (**VIS**ualization framework for **T**ask-aware decision-making in **A**utonomous experiment design). Based on modern digital twin settings for **autonomous experiment design**, VISTA aims to integrate computational models, AI/machine learning (ML) surrogates, and human-AI interactions to facilitate more effective and efficient data assimilation, data representation, model calibration, and ultimately adaptive/interactive decision making. ***Applied Math Thrust 5.4*** will aim to create multifidelity, multi-objective Bayesian optimization (MFMOBO) for intelligent data-acquisition with certified robustness as part of VISTA, specifically in cases where the multifidelity approximations vary over the input space. We will further explore adaptive MFMOBO to better comprehend the Pareto front with certified robust risk, considering high-consequence outcomes using the approximate evaluation under different fidelity.

6. AI-driven, Energy-efficient Integrated Research Platform for Experimental Facilities

Foundation Models (FMs) have demonstrated outstanding performance. Yet, bringing FMs to production research requires overcoming significant challenges, i.e., how we compute needs to be radically different. A leading large language model, GPT4, requires 1 GW for training and 260 MW per day, running a GPT4 inference service. This is impractical on DOE's exascale resources and impossible for real-time inference at the edge. ***Multifaceted Thrust 6.1*** will combine OED, few-shot model reduction, and efficient (energy/time) training and inference optimization. Meanwhile, ***Applied Math Thrust 6.2*** will seek energy-efficient UQ for FMs and hybrid physical-ML models.

Multi-institutional Team Investigators and Collaborators

Lead Principal Investigator (PI) and Co-PIs

Brookhaven National Laboratory: Kerstin Kleese van Dam (Lead PI)
Thomas Jefferson National Accelerator Facility (JLab): Malachi Schram (Co-PI)
Princeton Plasma Physics Laboratory (PPPL): Ammar Hakim (Co-PI)
Oak Ridge National Laboratory (ORNL): Rafael Ferreira da Silva (Co-PI)
New Jersey Institute of Technology (NJIT): Jing Li (Co-PI)
University of Delaware (UD): Rudolf Eigenmann (Co-PI)

Table 1. Summary Budget

	Name	**Institution**	**Year 1 Budget**	**Year 2 Budget**	**Year 3 Budget**	**Year 4 Budget**	**Total Budget**
Lead PI	Kleese van Dam, K.	BNL	$8,500	$8,500	$8,500	$8,500	$34,000
Co-PI	Schram, M.	JLab	$400	$400	$400	$400	$1,600
Co-PI	Hakim, A.	PPPL	$400	$400	$400	$400	$1,600
Co-PI	Ferreira da Silva, R.	ORNL	$400	$400	$400	$400	$1,600
Co-PI	Li, J.	NJIT	$150	$150	$150	$150	$600
Co-PI	Eigenmann, R.	UD	$100	$100	$100	$100	$400

($ in thousands)