U. S. DEPARTMENT OF ENERGY
FIELD WORK PROPOSAL

1. B&R No. KA2401022	2. Contractor No.: 2020-BNL-CC117-FUND	3. Date Prepared: 20191204	4. Task Term: Begin: 20191215 End: 20221214

5. Work Proposal No.: N/A	6. Work Authorization No.: KACH130

7. Title: HEP Center for Computational Excellence (HEP-CCE): Next-Generation HPC Systems and HEP Applications

8. Principal Investigator(s) : Kleese van Dam, Kerstin, (631) 344-6019

9. Headquarters/Operations Office Program Manager: Chatterjee, Lali (301) 903- 0435 10. Operations Office Work Proposal Reviewer: 11. Contractor Work Proposal Manager: Tribble, Robert (631) 344-3177	12. Headquarters Organization: Office of Science – Office of High-Energy Physics 13. Operations Office: CHICAGO, IL 14. Contractor Name: BROOKHAVEN SCIENCE ASSOCIATES BROOKHAVEN NATIONAL LABORATORY	15. HQ Organizational Code: SC 16. DOE Organizational Code: CH 17. Contractor Code: BN

18. Work Proposal Description (Approach, anticipated benefit in 200 words or less, suitable for public release) :

The HEP-CCE is a multi-Lab (Argonne, BNL, Fermilab, LBNL) activity aimed at bringing new computational capabilities to bear in support of HEP science goals. These include porting and optimizing HEP codes and frameworks for next-generation HPC systems, data-intensive computing tasks on HPC platforms, and addressing computation-related training issues for the HEP workforce. Many of these activities are carried out in partnership with DOE ASCR researchers.

HPC systems at DOE supercomputing facilities can be a powerful resource for HEP experiments provided the software can be suitably refactored, while satisfying two key requirements – efficiency and portability. The aim of this HPC-CCE project is to address this question via pilot sub-projects and collaborative research studies with the DOE ASCR HPC community in 1) portable parallelization strategies, 2) fine-grained I/O and related storage issues, 3) optimizing event generators, and 4) running complex workflows on HPC systems. Several of the broad questions are also being addressed within DOE's Exascale Computing Project (ECP), and there will be significant synergy with this work.

The HEP-CCE current BNL FWP is being submitted under a new FWP#, CC117. The prior HEP-CCE FWPs were submitted under BNL FWP# PO183.

19. Principal Investigator (s) :

Signature(s) Date: 12/04/2019

20. Contractor Work Proposal Manager: Signature Date 12/04/2019	21. Operations Office Review Official: Signature Date

22. Detail Attachments:

- (☒) a. Purpose
- (☒) b. Approach
- (☐) c. Technical progress
- (☐) d. Future accomplishments
- (☐) e. Relationships to other projects
- (☒) f. Explanation of milestones
- (☒) g. Other (Specify Topic) Data Management Plan

WORK PROPOSAL REQUIREMENTS FOR OPERATING/EQUIPMENT OBLIGATIONS AND COSTS

Contractor Name: BROOKHAVEN SCIENCE ASSOCIATES BROOKHAVEN NATIONAL LABORATORY	Contractor Number: 2020-BNL-CC117-FUND	Work Authorization No.: KACH130	Date Prepared: 20191204
TITLE: HEP Center for Computational Excellence (HEP-CCE): Next-Generation HPC Systems and HEP Applications			B&R Code: KA2401022

23. Staffing (in staff years)	Prior Years	FY2020	FY2021 President's	FY2021 Revised	FY2022 Request	FY2022 Authorized	FY2023	FY 2024	Total to Complete
a. Scientific		2.12		2.16	2.07				6.35
b. Other Direct		0.16		0.16	0.16				0.48
c. Total Direct		2.28		2.32	2.23				6.83
24. Operating Expense (in thousands)									
a. Total Obligations		$ 800		$ 800	$ 800				$ 2,400
b. Total Costs		$ 800		$ 800	$ 800				$ 2,400
25.1 Equipment (in thousands)									
a. Equip Obligations									
b. Equip Costs									
25.2 Accelerator Improvement Projects									
a. AIP Obligations									
b. AIP Costs									
25.3 General Plant Projects									
a. GPP Obligations									
b. GPP Costs									

26. Milestone Schedule (Tasks)	Proposed Schedule

Foot Notes:

22. Detail Attachments

a. Purpose

The current HEP-CCE proposal aims to connect the DOE ASCR and R&D programs to HEP computing across all three HEP science thrusts – the Cosmic, Energy, and Intensity Frontiers. The goal of the activity is to promote new capabilities that can aid HEP theorists and experimentalists in overcoming the challenges posed by next-generation computing platforms and to aid the HEP office in organizing and carrying out a number of cross-cutting topical evaluations, investigations, and studies.

Two key issues motivate the work proposed here. The first is a major increase in expected HEP experimental data volumes and the second is the potential lack of conventional HEP computing and storage resources to meet the associated demand. ASCR supercomputing facilities have the potential to play a significant role in addressing the computational challenge provided critical components of the HEP software stack can be refactored appropriately. Since such refactoring is a problem faced by many scientific communities today, the HEP effort can benefit from a strong connection to the approaches being taken by the HPC community. The bigger purpose of the HEP-CCE program described here is to bring these two communities together. The HEP program stands to benefit in several different ways – an easier and less risky path to evaluating and using next-generation computational systems, adoption of a future-looking approach to portable software, and the risk reduction in possessing an alternative mechanism for carrying out large-scale computing tasks.

Next-generation HEP experiments will have an order of magnitude increase in data rates and will face exponentially increasing data complexity starting in the mid-2020’s. New capabilities and methods for data processing, simulations, and data management are needed, as well as new analysis capabilities for exploring and understanding the recorded data and the corresponding simulations. ASCR computing resources (as well as ESnet) are expected to play a major role in addressing several of these issues. In this proposal we focus primarily on the needs of the HEP experiments (the theory/modeling component is in large part supported by the ECP) with a view to exploiting – in full production mode – the large-scale computational resources available at the ASCR-supported Leadership Computing Facilities (the two LCFs are at Argonne and Oak Ridge – ALCF and OLCF) and at NERSC.

Even though the available computational power will be very substantial, there is an important cautionary note: First, the increase in raw capability will be made possible by a new generation of computational hardware (GPU accelerators, vector units, specialized chips, ARM, FPGAs, etc.); second, high-performance computing system architectures are very different from the Grid, the archetype for HEP data processing. This poses a major challenge to the experiments’ software suites in adapting algorithms and making sure that they can be executed efficiently on multiple advanced hardware solutions, as the hardware and the HPC systems evolve. Managing the required software evolution is a very complex and difficult task and is currently being faced by not just HEP researchers, but the entire HPC community.

The work proposed here addresses the essential tasks that are needed to assess whether, and in what way, HPC systems can meet the needs of the experiments (in principle), and to study portable solutions that can be deployed across multiple platforms. This proposal involves a close collaboration between the domain scientists, the ASCR experts (including the ECP community), and the wider set of users who work with the experimental data, to ensure that the proposed solutions do in fact properly address the needs of the experiments. The timeline of the proposed activities is also connected with decisions that have to be made by the experimental teams.

b. Approach

Specific areas for HPC-focused HEP applications considered by this HEP-CCE proposal are 1) portable parallelization strategies (PPS) , 2) fine-grained I/O and related storage issues (IOS), 3) optimization of event generators (EG), and 4) the ability to run complex workflows (CW) on HPC systems at the LCFs and NERSC. The results of this work will also be useful for the ASCR facilities in planning for support of data-intensive computing campaigns. The proposal has been structured as a 3-year effort to yield deliverables of direct relevance to experiments on a relatively short timescale. Aside from the technical effort, there will be a significant multi-faceted outreach program to help ready the community for making best use of next-generation HPC systems. To summarize, the aim of the project is to explore and develop concrete, portable strategies for exploiting next-generation architectures by HEP experiments and to make the results widely available, so that they can be best exploited by the experimental community.

The proposal team consists of HPC experts with both ASCR and HEP backgrounds, and computational experts from the HEP experiments. The combined expertise covers deep knowledge of next-generation HPC architectures and systems, HPC software and performance optimization, the experiments' software base and HEP computing facilities (at BNL and FNAL), and the respective use cases and community needs. Because work at the ASCR facilities and coordinated data movement across them will be an essential component factoring into its overall success, we will have POCs to the project from the LCFs, NERSC, and ESnet.

The Management team for this project will be responsible for project execution and oversight of day-to-day tasks; it consists of the PI (Salman Habib, Argonne) and Deputy PI (Palo Calafiura, LBNL), and Technical leads for each main project R&D activity (PPS: Oli Gutsche, FNAL, deputies: Charles Leggett, LBNL and Meifeng Lin, BNL; IOS: Rob Ross and Peter van Gemmeren, Argonne; EG: Taylor Childers, Argonne, and Stefan Hoeche, FNAL; CW: Kyle Chard, Argonne, and Shantenu Jha, BNL). A Steering group (PI/Deputy PI, experiment POCs and the LCF/NERSC/ESnet POCs) will initially meet monthly and later on, at least quarterly, to assess roadmapping, scoping, and prioritization of project tasks, and to monitor and evaluate progress. It is important that the associated institutions work closely together; this requires a set of Institutional leads (Salman Habib, Argonne, Kerstin Kleese Van Dam, BNL, Liz Sexton-Kennedy, FNAL, and Peter Nugent, LBNL) who control local resources and availability of expert personnel and can help resolve problems connected with fluctuating project needs and any conflicts that might arise. As the set of HEP-CCE tasks evolves over time we will periodically reassess the organizational structure in order to respond to new developments.

f. <u>Explanation of Milestones</u>

Portable Parallelization Strategies (PPS): We will study the feasibility of using Kokkos, RAJA, OpenMP, SYCL, Numba, and other promising programming models in a few representative example codes (use cases) for single-node concurrency portability on both CPU and accelerator architectures. We will also explore the interplay of different concurrency paradigms, specifically multi-node parallelism implemented with MPI, and intra-node parallelism using the above-mentioned libraries. The emphasis for this project is to explore the use of high-level libraries in standalone programs, and to deliver recommendations and implementations for the experiments and their algorithm processing frameworks. We therefore favor technologies that mesh well with code that is primarily intended for CPUs. This work will go hand in hand with investigating the underlying event data models from the point of view of concurrency.

From project initiation, the PPS milestones are as follows:

- Q1, Preparation: Mapping of use cases to parallelization technologies and measurement platforms, deliverable – matrix of benchmarks of the (unaltered) use cases.
- Q4-7, First Implementation: Benchmark first implementations according to portability, performance, and usability.
- Q8, Consolidate Benchmarking Results: Write-up summarizing benchmarking results.
- Q12, Fully Functional Prototypes: Delivery of fully functional prototypes made available to the experiments.

Fine-Grained IO and Storage (IOS): The objective of this part of the project is to optimize HEP workflows on HPC systems from the point of view of data handling. We will explore how modifications to the event data models and the data representations that describe the physics objects can be optimized and handled both in runtime memory and in storage, to maximize the efficiency and throughput of HEP analysis on HPC machines. The investigations will focus on the viability of developing alternative data representations which are specifically tuned and persist-able on HPC installations, which can be used for archival storage at these sites, or reconstituted into the runtime environment to take advantage of high-speed storage buffer systems. This investigation will consider the need to account for the total cost of processing and conversion to the final archival format(s).

The IOS milestones are:

- Q1, Preparation: Document IO patterns and event data models (EDMs).
- Q2, Preparation: Benchmark performance of synthetic benchmarks on Grid resources.
- Q3, Preparation: First Implementation: Decide on storage technology optimization targets.
- Q9-10, Prototyping: Optimized persistent and in-memory EDM for synthetic benchmarks.
- Q11-12, Benchmarking and Recommendations to Experiments: Benchmark previously developed prototypes, provide experiment-specific recommendations.

Event Generators (EG): Event generators are the first step in the LHC simulation workflow, calculating the complex particle interactions using perturbative QCD to produce the particles measured by the detector. The precision of these calculations must be increased to enable scientific discovery at the HL-LHC, where new physics may manifest itself in form of small deviations from Standard Model predictions. Even at lowest order in perturbation theory, the computational complexity of the calculations underlying event generators grows exponentially with the number of experimentally observed objects in the final state. Developing accelerator-friendly versions of generators and integrators is an important step in reducing the computing demands for LHC experiments and will be critical to the success of the high-luminosity phase of the LHC.

The EG milestones are:

- Q1-3, Preparation: Implement initial set of test calculations; measure performance on GPUs, study and benchmark neural importance sampling.
- Q4-6, Initial Implementation: Implement generator for limited number of processes, cross-check with existing codes.
- Q7-12, Final Implementation: Design a generalized implementation covering the range of physics models of interest. Integrate with Sherpa/Pythia frameworks for distribution.

Complex Workflows (CW): HPC systems are not natively designed for handling complex, dynamic workflows and neither is the standard HPC software stack. As HPC systems begin to handle data-intensive tasks more routinely, there is a significant effort in making complex workflows run at the LCFs and at NERSC at scale. The efficient execution of the collective union of large numbers of heterogeneous tasks presents a difficult challenge of supporting computational campaigns that are comprised of millions of tasks from a diverse mix of size, duration, type (memory, I/O intensive), and obviously GPU and CPU mix. The imbalanced nature of HEP workloads, especially with granular tasks that have vastly different resource requirements and long-tail distributions, requires new workflow scheduling approaches and rich monitoring infrastructure to measure performance, limit resource usage, and learn characteristics for improved scheduling performance. We will explore methods for scaling online monitoring infrastructure for both production and user-focused workflows and explore the use of real-time information to inform job placement and resource usage. Given the likelihood of node and task failure, and the potential for significant inefficiency as a result of re-executing workflows, it is crucial that HEP workflows be able to quickly identify faults, integrate with external checkpointing models, implement workflow-level checkpointing to enable restart of unfinished work, and offer other resiliency measures. These methods will allow errors to be automatically resolved and for fault tolerant workflows to be executed efficiently without re-executing tasks that have completed successfully. The CW milestones are following:

- Q1-2, Preparation: Derive requirements for production and user-focused workflows, components, and execution environments.
- Q3-4, Initial Implementation: Develop and implement modular interfaces for execution and data management components.
- Q5-6, Containerization: Container-based models for translating requirements between systems, HPC orchestration support, and container-based scheduling models.
- Q7-8, Scheduling: HPC, container, and data scheduling models for performance improvement.
- Q9-10, Checkpointing: Checkpoint models for enabling workflow re-execution.
- Q11-12, Provenance: Provenance capture and export formats, analysis reproducibility.

Tasks for Brookhaven National Laboratory: The activities at Brookhaven National Laboratory will be coordinated by Kerstin Kleese van Dam, director of Computational Science Initiative (CSI), in consultation with the project management and experimental POCs. The BNL tasks described below will be carried out in close collaboration with the BNL Physics Department.

The technical focus areas at BNL will be Portable Parallelization Strategies (PPS) and Complex Workflows (CW). The PPS activity at BNL will be led by Meifeng Lin, who will be the PPS technical deputy co-lead (lead: Oliver Gutsche – FNAL, deputy co-lead: Charles Leggett – LBNL). Contributions to the PPS activity will mainly come from members of the CSI High Performance Computing group (e.g. Zhihua Dong, Hubertus van Dam, Kwangmin Yu) with support from relevant high-energy physics staff at BNL, including Brett Viren (DUNE) and Torre Wenaus (ATLAS). The BNL CW activity will be led by Shantenu Jha who is the CW technical co-lead (with Kyle Chard at ANL). Members in the C3D division of CSI will be brought in to work on this area as appropriate. These activities will be closely coordinated with the other labs working on complementary components and with the IOS and EG areas under the guidance of the project management.

g. Data Management Plan:

Data Sharing and Preservation: The work outlined in this project is primarily to develop portable accelerated/parallel computational techniques and associated software tools. We do not anticipate generating large quantities of data as an integral component of the proposed work. Nevertheless, all data with scientific value that is generated as part of this project will be archived at least for the duration of the project, and possibly longer. Parts of the archived data sets that can be used to check or validate results by comparing against other methods or implementations will be made publicly available.

Data in Publications: All data that appears in publications whether in the form of tables, figures, or images, will be simultaneously made publicly available from open websites and through other dissemination mechanisms. All publications will be available to the general public at arXiv.org.

Software Policy: Software tools generated under this project will be made publicly available through the project website and other resources such as GitHub and HEP-CCE portals.

Data Management Resources: As part of this project activity, we do not expect to commit data resources at any facility beyond what is considered nominal. (Archival data may be used for R&D in this area.)

Data Protection: We do not anticipate issues with confidentiality, personal privacy, Personally Identifiable Information, and U.S. national, homeland, and economic security. Activities and policies will be consistent in recognizing proprietary interests, business confidential information, and intellectual property rights; will avoid significant negative impact on innovation, and U.S. competitiveness; and will otherwise be consistent with all applicable laws, regulations, and DOE orders and policies.