Brookhaven National Laboratory

U.S. DEPARTMENT OF ENERGY

Architecting an AI-Centric Workload Management System: Evolving from Expert-Driven Complexity to an AI-Empowered Approach

PI: Tadashi Maeno (NPP)

Co-PIs: Sara Mason (CFN), Alexei Klimentov (CDF), Paul Nilsson (NPP)

Date: 9th January 2026

@BrookhavenLab

FY27 NPP LDRD Type B Pre-Proposal

Title: Architecting an AI-Centric Workload Management System: from Expert-Driven Complexity to an AI-Empowered Approach

PI: Tadashi Maeno[1]

Co-PIs: Sara Mason[2], Alexei Klimentov[3], Paul Nilsson[1]

Senior personnel: Torre Wenaus[1,3], Johannes Elmsheuser[1], Ofer Rind[3]

Cross-directorate: Yes ✓ No___

Directorates/organizations: [1]NPP, [2]CFN, [3]CDF

Proposal Term: From: 10/01/2026 To: 09/30/2028

Abstract: Today's scientific Workload Management Systems (WMSs) face critical scalability and accessibility challenges due to their heavy reliance on human expertise for configuration, optimization, and troubleshooting. To address this, we propose a shift to an **AI-centric WMS architecture** that integrates Artificial Intelligence (AI) agents, Machine Learning (ML) models, and Large Language Models (LLMs) as core components. By utilizing the Model Context Protocol (MCP), the system enables autonomous workflow orchestration, predictive resource forecasting, and natural-language interfaces for enhanced reasoning and diagnostics. This research aims to provide a generalizable and practical foundation for next-generation scientific computing demands.

Program: NP, HEP, EIC, Multiprogram.

Return on Investment: The proposed project aligns closely with DOE/ASCR AI-focused initiatives, such as AmSC, and BNL strategic goals for EIC, FCC, and HL-LHC, providing a foundation for follow-on funding within these emerging programs. Furthermore, standardizing on MCP and AI agents ensures interoperability with DOE–industry AI partnerships, positioning this work for future IRI funding focused on autonomous facility management.

Broader impact on the activities at the laboratory: Delivering validated research artifacts to support high-priority programs, including EIC, FCC, and HL-LHC, while streamlining operations for CFN-hosted research programs and BNL scientists (beyond NPP) needs.

Total planned funding per year in FY27 and FY28: $250k

Proposal Description: Motivation

The Challenge

- Today's WMSs have evolved into mission-critical platforms that must coordinate geographically-distributed, heterogeneous resources (CPU, GPU, HPC, Cloud, Storage, Network) with conflicting optimization strategies.
- Modern scientific analysis workflows involve multi-step dependencies with diverse compute, memory, I/O, and storage requirements, observed not only in NP/HEP experiments such as ATLAS and sPHENIX, but also in the growing diversity and complexity of batch workloads at the CFN compute facility.
- Current systems rely heavily on human expertise for tuning and troubleshooting, creating a scalability limit, slow adaptation to emerging scientific demands, and a significant barrier to entry for the scientific computing workforce.

Foundational Successes

- **PanDA system:** A BNL-originated, multi-experiment WMS that supports ATLAS, the Vera C. Rubin Observatory, DarkSide-20k, and other projects, managing large-scale data processing across globally distributed resources.
- **AskPanDA:** Successfully integrated LLMs and Retrieval-Augmented Generation (RAG) using MCP to automate diagnostics for the PanDA system.
- **Predictive ML Pipelines:** Demonstrated ability in REDWOOD to eliminate "two-stage" execution overhead by predicting memory, CPU, and walltime requirements using deep neural networks and gradient boosting.

→ The opportunity to shift from a static collection of heuristics to a learning-based cognitive control system.

Proposal Description: Research Objective

- Design and study a next-generation WMS that transforms an "expert-only" black box to adaptive, learning-based control, integrating AI agents, ML models, and LLMs as core components.
- Replace static heuristics and manual coordination with AI agents that learn from telemetry, historical trends, and system feedback, and interact with the system through MCP.
- Enable advanced automation for workload diagnostics, resource prediction, anomaly detection, and proactive error mitigation.
- Support the real-time coordination of complex, multi-step workflows across heterogeneous computing and storage resources.
- Lower knowledge barriers through LLM/RAG-based natural-language interfaces that allow operators to query system state and receive guided troubleshooting support.
- Leverage LLMs to automate the database lifecycle, including schema design and dynamic query optimization based on historical access patterns.

Proposal Description: Vision and Deliverables

The Vision

- Conducting a deep dive into MCP, MLOps platforms, and candidate LLMs to build a robust, modern foundation for AI integration.
- Designing a comprehensive WMS reference architecture that defines the specific roles and boundaries of AI agents.
- Delivering validated research artifacts to support high-priority NP/HEP programs, including EIC, FCC and HL-LHC, while streamlining operations for CFN-hosted programs.
- Positioning BNL to meet the extreme data demands of future experiments through a pivotal shift toward intelligent, self-evolving systems.

Key Deliverables

- **Reference Architecture:** A validated design framework for AI-centric WMSs, applicable across the next generation of large-scale scientific computing environments.
- **Research Artifacts:** Reusable AI agent structures, ML pipelines, and LLM-driven tools to accelerate BNL-wide R&D.
- **Strategic Impact:** Validated proof-of-concepts that demonstrate to reduce manual overhead and improve system transparency.

Summary

- Multi-directorate collaboration converging from NPP, CFN, and CDF, and uniting stakeholders from EIC, FCC, sPHENIX, and HL-LHC to build a universal solution.
- Transformation of WMSs from "expert-only" black boxes to AI-empowered cognitive control systems, ensuring infrastructure can scale without a linear increase in human operational costs.
- While rooted in NP and HEP, the framework provides immediate and broader benefits across various BNL activities including CFN compute facility, addressing the issues with the growing diversity of cross-disciplinary workloads.
- A strategic investment in a coordinated, AI-centric approach,leveraging in-house extensive expertise in WMSs, positions BNL to deliver a reusable WMS architecture and capabilities needed for EIC, FCC, HL-LHC, and future data-intensive programs.