AUG 2024

# ALCF-4 DESIGN REVIEW BENCHMARKS


**CHRISTOPHER KNIGHT**
ALCF Catalyst Team Lead

# AGENDA

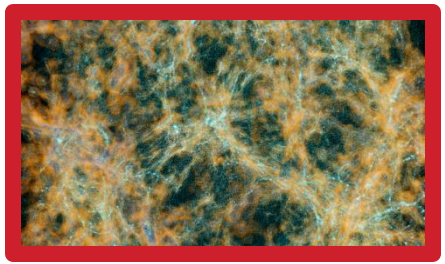
| Times | Item | Owner |
|---|---|---|
| 8:30 | Executive Session | Review Chair |
| 9:00 | Welcome | Mike Papka |
| 9:10 | Project Overview | Jini Ramprakash |
| 9:40 | Technical Overview and Early Science | Kevin Harms<br>Chris Knight |
| 10:15 | Break | |
| 10:30 | Technical Requirements | Taylor Childers |
| 11:30 | Benchmarks | Chris |
| 12:15 | (Working Lunch)<br>Discussion & Questions from the committee | ALCF-4 Team |
| 12:30 | (Working Lunch) Executive Session | Review Chair |
| 13:30 | Facilities | Jon Cisek |
| 14:15 | ALCF-4 Risks Review | Noah / Jini |
| 15:00 | Break | |
| 15:15 | Executive Committee Q&A with ALCF-4 team | Review Chair |
| 15:45 | Executive Writing Session | Review Chair |
| 17:00 | Adjourn / Tour of Aurora | Susan Coghlan |
| 18:00 | Dinner | |

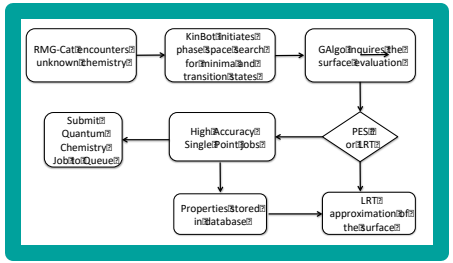
# CHARGE QUESTIONS

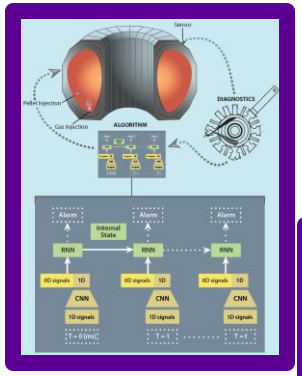
1. Is the technical approach appropriate to support the ALCF-4 Mission Need requirements?
2. **Are the RFP technical requirements reasonable, clear, and consistent with the goals and objectives for the ALCF-4 project?**
3. Does the ALCF facility upgrade plan support the system requirements specified in the RFP for the onsite options?
4. Have the major technical risks and appropriate mitigation strategies been correctly identified for this stage of the project?

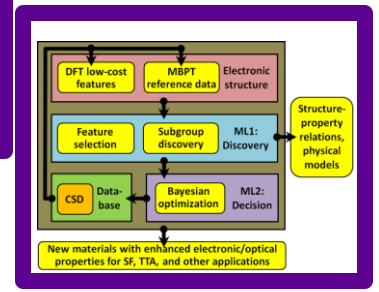
# ENABLING DIVERSE WORKLOADS AT-SCALE

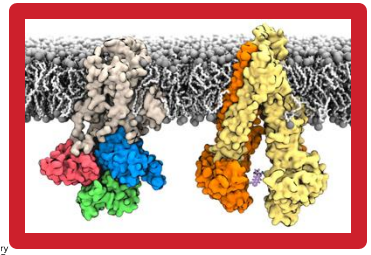
- As the complexity of scientific questions being asked continues to increase, so to does the diversity of computational workloads.
- ALCF projects (INCITE, ALCC, ESP, DD, …) have grown to include combinations of the 3 traditional Modeling & Simulation, Artificial Intelligence, and Data-Intensive workloads, often executed within Workflows.

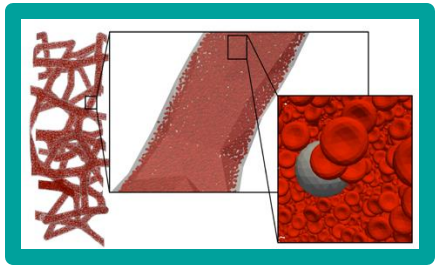






# ALCF-4 BENCHMARK VISION

- Identify a representative subset of ALCF workloads ensuring user community is able to explore new science and the ALCF-4 system is competitive.
- Support for “3-pillars” of computing
  - Modeling & Simulation (ModSim)
    - Traditional fp64/fp32 computing (w/ mixed precision)
    - Portability
  - AI/ML
    - Training and inference using industry standard frameworks
    - Integration with ModSim applications
  - Data Intensive
    - Processing and analysis of data (throughput focus)
    - Potentially more integer instructions

# ALCF-4 BENCHMARK VISION

- Diverse set of benchmarks
  - — Programming languages & models
  - — Run configurations (e.g. # MPI ranks per device)
  - — Stress different components of SW & HW

- Weak- and Strong-scale to full-machine
  - — Various combinations of workloads: ”Hero” runs & ensembles

- Intend to use Aurora to measure baseline Figure-of-Merits (FOM) and project performance
  - — Profile data from Aurora will be shared to help vendors with their estimates
  - — Work-in-progress…

# ALCF-4 BENCHMARK VISION

- Benchmarks are aligned with top priorities called out in (draft) RFP to support current and future ALCF users
  - C/C++, Fortran, and Python
  - OpenMP and SYCL
  - AI Frameworks
  - …

- Recognize growing importance of utilizing lower precision data types for performance (i.e. mixed precision algorithms)

- Workflows include multiple computational elements, often with different hardware and software requirements and complex inter-dependencies.

# ALCF-4 MODSIM BENCHMARKS

- HACC
  - Extreme-scale cosmological simulation code
- nekRS
  - Computational fluid dynamics (CFD) solver
- Thornado
  - Spectral neutrino transport in stellar astrophysics simulations
- QMCPACK
  - Quantum Monte Carlo simulations

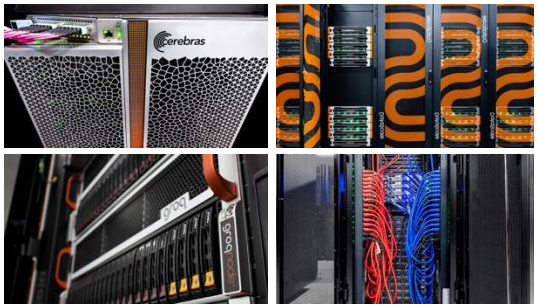
- Algorithmic Patterns
  - Dense Linear Algebra, Monte Carlo, FFTs, Particles, Structured Grids, pt2pt, all2all, …

# ALCF-4 AI/ML BENCHMARKS

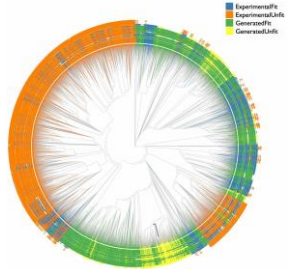
- Benchmarks
  - Dense and Sparse model pre-training (GPT & MOE)
  - 3D Vision Transformers
  - Distributed GNN (coupled with simulations)
  - Inference Suite
  - Clustering at Scale (traditional ML)
- ALCF making impactful contributions to developing AI/ML Benchmarks

MLCommons

HPC, Storage, Science, AI Safety




ALCF AI Testbed



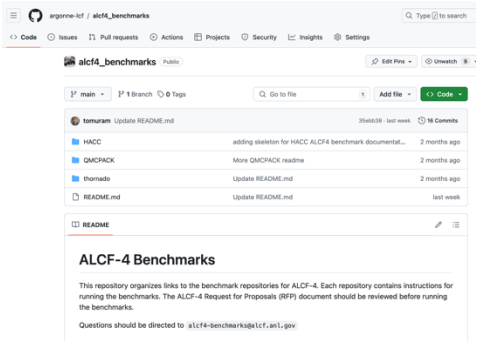
SC'24 Gordon Bell Finalist scaling on 5 diverse systems

# ALCF-4 WORKFLOW BENCHMARKS

- Static Worker
  - Single task launches worker tasks, which then manage execution of sub-tasks
  - Targets hardware with different granularity and spanning across nodes.
- Multi-size Ensemble
  - Execute set of single- or multi-rank MPI tasks running same application but with different sizes and/or input decks
  - Tests capabilities of “mpiexec” launcher
- Heterogeneous Workflows
  - Varied set of tasks requiring periodic data/meta-data/ML model synchronization across components
  - Involves frequent movement of large data across the interconnect

# ALCF-4 BENCHMARKS DRAFT

- Preparing superset of benchmarks spanning ModSim, AI/ML, and Workflows
  - — Standard benchmarks also included
    - • GEMM (multiple precisions), STREAM,




- GitHub Repo:
  - — Work-in-Progress
  - — Instructions for obtaining code, building, running, and validating results
  - — Codes expected to remain static once finalized

- Offeror's will be requested to provided baseline projections
  - — Also able to provide "optimized" projections

https://github.com/argonne-lcf/alcf4_benchmarks


ANY QUESTIONS?


Argonne
NATIONAL LABORATORY
Argonne Leadership
Computing Facility