CORAL
COLLABORATION
OAK RIDGE ARGONNE LIVERMORE

Briefing on CORAL-2 RFP and Draft Technical Requirements

Vendor Webinar Meeting
December 6, 2017

http://procurement.ornl.gov/rfp/CORAL2/
What is CORAL?

- CORAL is a Collaboration of Oak Ridge, Argonne, and Lawrence Livermore Labs

- Began in 2012 to acquire three systems for delivery in 2017 with a single RFP

- This CORAL partnership and process was considered a big success in the first acquisition

- So the partnership is being continued for the second round of system acquisitions (called CORAL-2) with a single RFP for delivery of a system in late 2021 and one or two systems in late 2022.
# DOE HPC Facilities Systems

## Pre-Exascale Systems

<table>
<thead>
<tr>
<th>Year</th>
<th>System</th>
<th>Location</th>
<th>Architecture</th>
<th>Classification</th>
</tr>
</thead>
<tbody>
<tr>
<td>2013</td>
<td>Mira</td>
<td>Argonne</td>
<td>IBM BG/Q</td>
<td>Unclassified</td>
</tr>
<tr>
<td>2016</td>
<td>Theta</td>
<td>Argonne</td>
<td>Intel/Cray KNL</td>
<td>Unclassified</td>
</tr>
<tr>
<td>2018</td>
<td>C0RE</td>
<td>ORNL</td>
<td>Cray/Intel Xeon/KNL</td>
<td>Unclassified</td>
</tr>
<tr>
<td>2020</td>
<td>Titan</td>
<td>ORNL</td>
<td>Cray/NVidia K20</td>
<td>Unclassified</td>
</tr>
<tr>
<td></td>
<td>Sequoia</td>
<td>LLNL</td>
<td>IBM BG/Q</td>
<td>Classified</td>
</tr>
<tr>
<td></td>
<td>Trinity</td>
<td>LANL/SNL</td>
<td>Cray/Intel Xeon/KNL</td>
<td>Classified</td>
</tr>
</tbody>
</table>

## Exascale Systems

<table>
<thead>
<tr>
<th>Year</th>
<th>System</th>
<th>Location</th>
<th>Architecture</th>
<th>Classification</th>
</tr>
</thead>
<tbody>
<tr>
<td>2021-2022</td>
<td>Summit</td>
<td>ORNL</td>
<td>IBM/NVidia P9/Volta</td>
<td>Unclassified</td>
</tr>
<tr>
<td></td>
<td>ORNL</td>
<td>ORNL</td>
<td>TBD</td>
<td>TBD</td>
</tr>
<tr>
<td></td>
<td>LLNL</td>
<td>LLNL</td>
<td>TBD</td>
<td>Classified</td>
</tr>
<tr>
<td></td>
<td>LANL</td>
<td>LANL/SNL</td>
<td>TBD</td>
<td>TBD</td>
</tr>
</tbody>
</table>

### NERSC-9

- LBNL
- LBNL
- TBD
- Unclassified

### Crossroads

- LLNL
- LLNL
- TBD
- Classified

### Frontier

- ORNL
- ORNL
- TBD
- Unclassified

### El Capitan

- LLNL
- LLNL
- TBD
- Classified
RFP Will Request Two Proposals

• ANL is getting a 2021 Intel system (Aurora) outside this RFP.
• The RFP requests proposals for an ORNL system delivered in late 2021 (accepted in 2022) with the requirement that the system must be diverse from the ALCF 2021 system.
• The RFP requests proposals for a LLNL system delivered in late 2022 (accepted in 2023). There is no diversity requirement.
• ANL, as CORAL partner, has the potential to choose a system from the same pool as LLNL.

• The reasons for just one RFP:
  – Same RFP release time
  – Same system requirements (based on ECP exascale definition)
  – Only differences are facility descriptions and I/O subsystems
  – Potential opportunity to share NRE costs
CORAL-2 RFP Time Line

• Vendors are free to talk with the DOE labs about the RFP until the RFP is officially released (e.g. unachievable requirements)

• RFP Release goal is February 2018

• Responses due eight business weeks later (early April)

• Evaluations by Lab teams

• Selections made in May 2018

• Negotiations for NRE contracts begin (50 pg proposal)
  – with goal of awards before October 2018
  – one for each build contract (less if shared NRE)

• Negotiations for System Build Contracts begin
  – with goal of awards by start of 2019
  – Set Go/No Go date to change targets to firm requirements
LLNL El Capitan System goals and objectives

- Support the ongoing and future workload for the NNSA Stockpile Stewardship Program.
  - Life Extension Programs
  - Demands from the Predictive Capability Framework and the experimental campaigns

- Replace Sierra with a system to support:
  - Hostile environment assessment
  - Uncertainty quantification for certification
  - Very large weapons’ science calculations

- Siting in 2022, acceptance in 2023

- Available to support growing mission needs through 2027
ORNL Frontier System goals and objectives

- Provide the Leadership computing capabilities needed for the DOE Office of Science mission thru 2026
  - Capabilities for INCITE, ECP, and ALCC science projects
  - Capabilities assessed from gathered user requirements
  - Increased need for data science and deep learning
- Siting in 2021. Run in parallel with Summit until Summit end-of-life (2023)
Technical requirements contents

Started with the successful RFP developed last time in CORAL
As before, chapter teams with experts from all three labs worked on each chapter

Chapter 3 High Level System Requirements
Chapter 4 Application Benchmarks
Chapter 5 Compute Partition
Chapter 6 I/O Subsystem
Chapter 7 Interconnect
Chapter 8 OS, Middleware, System Resource Management
Chapter 9 Front-end Environment
Chapter 10 System Management and RAS
Chapter 11 Maintenance and Support
Chapter 12 Facility Requirements
Chapter 13 Project Management
Basic RFP Terminology Definitions

The technical requirements have these designations:

• **Mandatory Requirement (MR):** *(only 3)*
  – Requirements that are essential to CORAL RFP
  – Required to be *responsive*

• **Mandatory Option (MO):**
  – Separately priced system options that may be exercised in individual build contracts
  – Required to be *responsive*

• **Technical Option (TO-1):**
  – System upgrade option that each laboratory may choose
  – *Not* required to be *responsive*

• **Target Requirement (TR):**
  – Performance features that are important to CORAL requirements
  – Graded (1, 2 or 3) based on overall importance to CORAL
    – 1 is essential; 2 is important 3 is desirable
  – *Not* required to be *responsive*
Technical requirements guiding principles

• Minimize number of mandatory requirements and allow consideration of widest range of architectural solutions.

• Focus on requiring science and throughput performance. Avoid overly prescriptive explicit speeds and feeds.

• Agree on common technical requirements across all three Laboratories – not three separate sets of requirements.

• Require vendors to describe available options to adjust system size and configuration to meet individual site needs and/or budgetary constraints.
SOW has three mandatory requirements

1. Description of the CORAL system
   – Overall system architecture and details of the interconnect
   – Detailed node architecture diagram with all data movement paths

2. Description of the software stack
   – Describe all software components and license strategy for each
   – Entire system software stack including system management and program development

3. Description of high level project management
   – Proposed multi-year collaboration plan
   – Plan to deliver systems
   – Perceived risks, risk management plan, and potential mitigations
   – Prime’s ability to ensure the responsiveness of its partners to the performance requirements
High level system requirements (1 of 4)

It is desirable for overall system to meet or exceed the following:

**CORAL system peak (TR-1)**
- Peak of at least 1300 PF (double precision floating point)

**Reasons**

- Why a peak requirement? Provides vendors a **target** to shoot for.
  - Vendor may have to propose more to meet 50X benchmarks FOMs
  - There is a technical option requirement to scale peak up or down.
- Providing 50X science and throughput over today's largest DOE systems is a mission need for ECP
  - Given Titan and Sequoia performance today, 1300 PF was considered a minimum size to achieve 50X science. 1 EF peak not expected to be enough
High level system requirements (2 of 4)

It is desirable for overall system to meet or exceed the following:

**Total Memory (TR-1)**
- Aggregate of at least 8 PB total system memory
- Must propose the same memory capacity and configuration as is used to obtain the reported benchmark results
- Place a high value on solutions that provide mostly fast memory (e.g., HBM)

**Reasons**

- Science teams wanted to be sure system has enough memory to do the science without prescribing a particular memory architecture or configuration.

- Study of exascale apps showed that at least 4-6 PB of HBM needed. Benchmarks set up to require 5 PB.

- Data Science and Deep Learning apps often desire large memory per node.

- The vendors can propose a wide range of memory architectures. RFP does not prescribe a particular architecture.
High level system requirements (3 of 4)

It is desirable for overall system to not exceed the following:

**Maximum Power Consumption (TR-1)**

- Cannot exceed 40 MW for the 2021 system or the 2022 system including all peripherals and the file system.
- Power consumption between 20-30 MW preferred

**Reasons**

- All three labs agreed that 40 MW would be the maximum requirement based on
  - Available power at each facility for the new system
  - Annual cost of power
  - Evaluation of the RFI responses for expected power requirements for systems in the 2021-2023 timeframe
- The 2022 proposals are expected to have more capability compared to the 2021 proposals for a given power consumption.
Require vendors to describe any available options to adjust system

- Scale the system size (TO-1)
- Scale the system memory (TO-1)
- Scale the system interconnect (TO-1)
- Scale the system I/O (TO-1)
- Scale the integrated telemetry DB (TO-1)

Because the different CORAL sites may want to procure different system configurations

- CORAL-scalable unit (MO)

CORAL sites want to procure smaller versions of the system

- Options for mid-life upgrades (TO-1)

System does not have to be upgradeable, but if it is, what are the options and time-line
Other notable requirements

Early access to CORAL hardware/software (TR-1)

Throughout the long-term collaboration, Offeror will propose mechanisms to provide the Laboratories with early access to hardware and software technology for evaluation, feedback, and application testing

Reason To enhance the partnership nature of the vendor relationship

Two runtime variability requirements (TR-1)

Runtime variability $\leq 5\%$ on a dedicated system running the benchmarks

Runtime variability $\leq 30\%$ on a busy system with representative workload

Reason Reproducible performance from run to run is a highly desired property by science teams. Values of 5% and 30% come from experience with original CORAL RFP
Integrated Telemetry Database and Analytics Infrastructure (TR-1)

New requirement that has arisen since the original CORAL RFP. Sys admins, operators, tool developers, and end-users all had input into this requirement.

Offeror will provide an integrated telemetry database (ITDB) that facilitates collection and analysis of system management and monitoring data to help understand utilization of all system resources, failure modes and trends, and to correlate the various data sources. The database will be scalable and accessible through a web interface and command-line interface.

• The database will provide rich search capabilities and advanced analytics that enable authorized users to perform complex event correlation and problem determination as well as to avoid outages proactively and to optimize system operation and performance.

• Authentication and authorization mechanisms will enable subsets of data to be made available to different groups of people (e.g., system administrators, operators, end-users, and researchers).
The evolution of node local and system local storage has led to the replacement of an *optional* storage system requirement in the original CORAL RFP with a TR-1 requirement to include the entire I/O subsystem in the RFP response.

ORNL and LLNL use their file systems in different ways that require different I/O subsystem requirements:

- The I/O subsystem for the 2021 system at ORNL will be the center-wide storage system, providing a global POSIX namespace that is accessible by both the CN partition as well as external resources in the facility;
- The I/O subsystem for the 2022 system at LLNL will not be a center-wide system but will support two types of namespaces: a *global namespace* and *transient namespaces* and will support two logical storage tiers, each optimized based on their performance and data residency requirements.

The 2021 System’s I/O subsystem and the 2022 System’s I/O subsystem requirements are described in separate sections to make it clear to the vendors what to respond to in each of the two proposals.
Application performance requirements are the highest priority to CORAL

An average “figure of merit” (FOM) improvement of at least 50X for scalable science, throughput and data science apps over today’s DOE systems.

- Reference machines are Sequoia and Titan
- The Offeror will provide actual, predicted and/or extrapolated performance results for the proposed system for the following:

• **CORAL-2 system performance (TR-1)**
  - Average FOM over five scalable science apps >= 50
  - Average FOM over six throughput apps >= 50
  - Average FOM over four data science and machine learning apps >= 50
  - Raw results for nine skeleton apps
CORAL-2 benchmarking strategy

• Assuring that actual DOE applications perform well on CORAL platforms is key to success
  – Develop a benchmark suite that is representative of the DOE workload
  – Develop a benchmark suite that can evaluate future technology options
  – Provide flexibility to vendors for diverse architectures

• Full applications not ideal to achieve all these goals
  – Millions of lines of code, may be MPI only, & perhaps export controlled
  – Limit benchmark suite to smaller applications, skeleton benchmarks, and micro-benchmarks
Coral-2 benchmark categories represent DOE workloads and technical requirements

- **Scalable Science Benchmarks**
  - Expected to run at full scale of the CORAL-2 systems

- **Throughput Benchmarks**
  - Represent large ensemble runs; may be subsets of full applications; run many small to medium sized jobs across full system

- **Data Science & Deep Learning Benchmarks**
  - Represent emerging data intensive & deep learning workloads
  - Integer operations, instruction throughput, indirect addressing, graph analysis
  - ML algorithms, regression analysis, training

- **Skeleton Benchmarks**
  - Investigate various platform characteristics including network performance, threading overheads, I/O, memory, memory hierarchies, system software, and programming models
**CORAL-2 benchmarking suite uses proxy-apps and a few larger applications**

<table>
<thead>
<tr>
<th>Scalable Science</th>
<th>Throughput</th>
<th>Data Science Deep Learning</th>
<th>Skeleton</th>
</tr>
</thead>
<tbody>
<tr>
<td>QMCPACK, HACC, NEKbone, ACME, VPIC</td>
<td>Quicksilver, Kripke, LAMMPS, AMG, PENNANT, Laghos</td>
<td>Integer sort, Havoq, GEMM, Pynamic, CLOMP, Deep Learning Suite, Big Data Suite, ML suite, I/O suite, RAJA suite, Hash</td>
<td>MPI suite, Memory suite, GEMM, Pynamic, CLOMP, ML suite, I/O suite, RAJA suite, Hash</td>
</tr>
</tbody>
</table>

All Benchmarks are Target Requirements (TR-1) in the RFP

CORAL-2 Benchmarks code and procedures are available at https://asc.llnl.gov/CORAL-2-benchmarks
CORAL-2 performance targets for scalable science and throughput codes address key application/workload requirements

50X Improvement on Scalable Science
- Geometric Mean
- QMCPACK
- HACC
- NEKbone
- ACME
- VPIC

Each run/projected at full machine scale
Recommend 5X larger problem and 10X faster

50X Improvement on Throughput
- Geometric Mean
- Quicksilver
- Kripke
- LAMMPS
- AMG
- PENNANT
- Laghos

Each run/projected at 1/24\textsuperscript{th} and 1/4\textsuperscript{th} machine scale
Multiple jobs to fill machine
Recommend 2-10X larger problem and 5-25X faster
CORAL-2 system performance targets will be projected for each benchmark category.

Box represent full scale

<table>
<thead>
<tr>
<th>QMCPACK</th>
<th>HACC</th>
<th>NEKbone</th>
<th>ACME</th>
<th>VPIC</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Throughput 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Throughput 2</td>
</tr>
<tr>
<td>Throughput 3</td>
</tr>
<tr>
<td>Throughput 4</td>
</tr>
</tbody>
</table>

Quicksilver | Quicksilver | Quicksilver | Quicksilver | Quicksilver | Laghos | Laghos | Laghos | Laghos | Kripke | Kripke | Kripke | Kripke | LAMMPS | LAMMPS | LAMMPS | LAMMPS | AMG | AMG | AMG | AMG | PENNANT | PENNANT | PENNANT | PENNANT |

\[ S_i = \frac{\text{projected FOM}_i}{\text{baseline FOM}_i} \]

**Scalable Science Benchmarks**
(each run/projected at full machine scale)

\[ S_{\text{scalable}} = \left( \prod_{i=1}^{5} S_i \right)^{\frac{1}{5}} \]

**Throughput Benchmarks**

**Run mode#1: Capability @ 1/4th machine scale**

\[ S_{4 \text{throughput}} = \left( \prod_{i=1}^{6} S_i \right)^{\frac{1}{6}} \]

**Run mode#2: Ensembles @ 1/24th machine scale**

\[ S_{24 \text{throughput}} = \left( \prod_{i=1}^{24} S_i \right)^{\frac{1}{24}} \]
CORAL-2 addresses emerging data centric and deep learning workloads

- Both full machine and single node benchmarks
- Parallel runs across entire target platform

**Exercised capabilities**
- Interconnect
- Entire memory hierarchy
- Irregular access patterns

**Stressed features**
- Integer operations
- Instruction throughput
- Indirect addressing

50X Improvement on Data Science & Deep Learning

Geometric Mean

Integer Sort
Havoq
Big Data Suite
Deep Learning Suite
CORAL-2 benchmarking includes emerging Data Science, Deep Learning, & Big Data Analytic workloads

- **Data Science**
  - Parallel Integer Sort + Havoq (Graph Analytics) Triangle counting

- **Deep Learning Benchmark suite contains:**
  - Convolutional Neural Networks (CNNs) that comprise convolutional layers followed by fully connected layers;
  - Long short-term memory (LSTM) recurrent neural network (RNN) architecture that remembers values over arbitrary intervals to deal with temporal and time-series prediction;
  - Distributed training code for classification in ImageNet data set at scale

- **Big Data Benchmark suite contains:**
  - Subset of CANDLE benchmark codes that implement deep learning architectures that are relevant to problems in cancer.
  - Address problems at different biological scales, specifically problems at the molecular, cellular and population scales
  - Two diverse benchmark problems, namely
    - P1B1, a sparse autoencoder to compress the expression profile into a low-dimensional vector
    - P3B1 a multi-task deep neural net for data extraction from clinical reports.
CORAL-2 system performance targets will be projected for each benchmark category

Data/Graph Analytics

$S_i = \frac{\text{projected FOM}_i}{\text{baseline FOM}_i}$

Run mode: Full machine scale

$S_{\text{data}} = \left( \prod_{i=1}^{4} S_i \right)^{\frac{1}{4}}$

Big Data and Deep Learning

Run mode: Small Ensembles
Rules for CORAL-2 benchmark procedures

- Define a variety of problem sizes:
  - **Small** is single node
  - **Medium** is ~1K nodes
  - **Large** is reference problem on current platforms
  - **CORAL-2 class problems** are:
    - SS: 5X or larger + 10X faster
    - TP: 2-10X larger + 5-25X faster
    - Require 5 PB of memory

- Follow CORAL-2 code modification guidelines
  - Baseline and optimized results allowed

- Offeror estimates sustained performance for science, throughput, & data science deep learning benchmarks
  - **CORAL-2 class problem sizes**
Allowed code modification to CORAL-2 benchmarks

- Benchmarks may be modified as necessary to get them to compile and run
  - Portability changes for programming models are allowed

- A full set of benchmark runs must be reported with this “baseline” source code
  - Can include non-intrusive and/or portable optimizations
    - E.g., compiler flags and standard pragma-style guidance
  - Can include anticipated changes to system software
    - E.g., MPI and OpenMP runtime improvements

- At least 1 GB per MPI task and use threading within each task if necessary to utilize all compute resources
  - Requirement tied directly to current CORAL codes and production usage
Allowed code modifications for optimized results

• Offeror may also report optimized results
  – Any and all code modifications are allowed
  – However, wholesale algorithmic changes that are strongly architecture specific have less value
  – All benchmark code modification will be documented and provided to CORAL-2

• CORAL-2 and Offeror will continue to improve the efficiency and scalability of all benchmarks between award of the contracts and delivery of the systems
  – Emphasis on higher level optimizations as well as compiler optimization technology improvements while maintaining readable and maintainable code
Questions?

ORNL Contract Administrator Contact Information:
William (Willy) Besancenez
Procurement Officer
Acquisition Management Services Division
Oak Ridge National Laboratory
Phone - (865) 576-1538
E-mail - besancenezwr@ornl.gov

http://procurement.ornl.gov/rfp/CORAL2/