

CORAL-2 Questions & Answers

January 15, 2018

1. Please make sure the RFP is very clear about OS licensing requirements for the compute nodes. Ex. LLNL has a volume license price for RHEL. ORNL does not.

Thanks for feedback to clarify OS licensing requirements in the RFP, which we will do.

2. Section 3.3.1 Open Source Software (TR-1): Open Source. What specific resources are DOE / Labs intending to put into an open source framework? How can we specifically integrate DOE/Labs plans with our plans?

The 3.3.1 TR states, "The Laboratories strongly prefer that all offered software components are Open Source." The reason for this preference is because, in the past, Labs have acquired systems containing closed source software that the vendor was unwilling or unable to fix bugs in or to modify to meet the lab's needs. Open source reduces the chance that Labs would be stuck in a similar situation in the future. The TR is generic and does not refer to specific software components. We understand that some components will not be provided as open source. Our preference is that those components be minimized.

The Labs are not dependent on any specific software being developed by the DOE Exascale Computing Project (ECP). Vendors are expected to supply a complete software stack for their proposed system. If the vendor feels their stack can be enhanced (or gaps filled) by incorporating software being developed by the ECP, then vendors are welcome to work with and leverage the ECP efforts.

3. Section 3.6 Early Access to CORAL Software Technology (TR-1): Is this referring to a Development System?

Yes. Both 3.5 (HW) and 3.6 (SW) asks vendors to propose mechanisms for Labs to have early access to the technologies that will show up in the final system. This is typically done through a small early access system available a few months before the final system.

4. Section 3.7.6 CORAL-Scalable Unit System (MO): Who is referenced by "CORAL partners"? Does this imply anyone other than the Laboratories?

CORAL partners refer to just the three CORAL Laboratories. It does not imply anyone else.

5. Section 4.1 Benchmark Categories: Twenty-four applications is a LOT. Specifically, the effort required for the Scalable Science, Throughput and Data Science/Deep Learning suites will be significant. Need some discussion re: priority, emphasis and methodology. Can we incorporate benchmarking and performance validation as part of post-award effort? To allow more in-depth performance assessment lead to late-binding decisions.

The Benchmark Section has been revised and the number of applications that have to be analyzed has been significantly reduced.

6. Section 5.2.15 Signals (TR-2): Does “POSIX compliance” infer ‘partial compliance’ or perhaps ‘compatibility’? That is, is partial compliance expected?

Section 5.2.15 specifies that the compute node OS provide POSIX compliance for signals and threads including: nested signals; and proper saving and restoring of signal mask. It does not imply POSIX compliance in other respects. For example, 6.1.1.1 states the Laboratory will accept relaxed POSIX semantics. Specifically, the Laboratory does not require support for Read After Write (RAW) between different processes on different nodes on the same open file.

7. Section 6.1.2.1 Minimum I/O Subsystem Aggregate Capacity (TR-1): 100 times aggregate CN memory for the capacity tier works out to be 1 EiB raw storage, assuming RAID6(8+2) data encoding. Delivering this amount of storage will consume a very large portion of the system budget. Is there a smaller minimum capacity that would be acceptable?

The vendor should propose the configuration that it feels provides the best system for the Laboratories. Per section 3.7.4 the vendor should describe and separately price options for scaling the capacity and performance of the CORAL I/O subsystem. If 6.1.2.1 is not met then 3.7.4 should describe an option that does.

Note that the capacity in 6.1.2.1 applies to the 2021 system and not to the 2022 system.

8. Section 6.1.3.2.2 File Per Process Checkpoint Performance (TR-1): Checkpoints to flash could be bound by the endurance of the flash devices. Higher drive writes per day necessitates more drives or higher capability (and cost) drives to satisfy the requirements for the life of the machine. The draft SOW calls for 3 minute checkpoints executed every hour. Is that an expected sustained system average checkpoint interval, or is the average checkpoint interval longer (while still retaining the 3 minute checkpoint duration)?

The expected sustained rate is a checkpoint per hour (for the 2021 and 2022 I/O subsystems). The 2021 system’s 6.1.3.2 requirement specifies that the checkpoint size is only 1/5 the memory size every hour. This corresponds to less than 5 drive writes per day. Similarly, the requirements in 6.2 for the 2022 system corresponds to 6 drive writes per day.

Vendors should not base their projection of flash capacity requirements on overly conservative endurance estimations. Manufacturers’ flash endurance estimates are based on write amplification (WAF) levels that are higher than we have measured for checkpoint workloads, and also on environmental conditions that are not representative of how these devices will operate in our HPC systems.

9. Section 10.1.5 System Performance Analysis and Tuning (TR-1): This seems like a difficult request to satisfy. Can this requirement be further clarified and refined?

After further consideration the Section 10.1.5 requirement has been removed from the SOW.

10. Sections 12.1 through 12.4 – The attached CORAL-2_SOW_Feedback_Sections_12.1-12.4.pdf document attempts to tabularize all the facilities attributes pertinent to the individual sites. Ideally, each cell in the table would contain a value or guidance for the contractor.

Thanks for feedback to CORAL about missing facility attributes

11. Section 12 CORAL Facilities Requirements: What is the minimum acceptable service aisle width between adjacent rows of Offeror provided equipment in the CORAL-2 system?

[ORNL] There is no explicit service aisle width minimum between adjacent rows of Offeror-provided equipment. Offeror should stipulate the minimum service aisle requirement for each distinct system type, cabinet, or similar, as part of a description of their potential system layout. For example, compute cabinets, storage cabinets, network cabinets, and CDU may all have different dimensions and service/accessibility requirements.

ORNL maintains a 40" service minimum in front of all switchboards, electrical panels, PDUs, and air-conditioning units. Offeror should consider the impact of these 40" minimum clearance requirements. Freight aisles have a preferred width of 72".

The raised floor grid uses 24"x24" tiles. Minimizing overlap of equipment and tiles will enhance serviceability beneath the floor.

[LLNL] The minimum service aisle width between adjacent rows of Offeror-provided equipment is 4'. This width may need to be increased to adequately install all power and cooling solutions; however, the raised floor grid uses 24"x24" tiles so minimizing overlap of equipment and tiles will enhance serviceability beneath the floor.

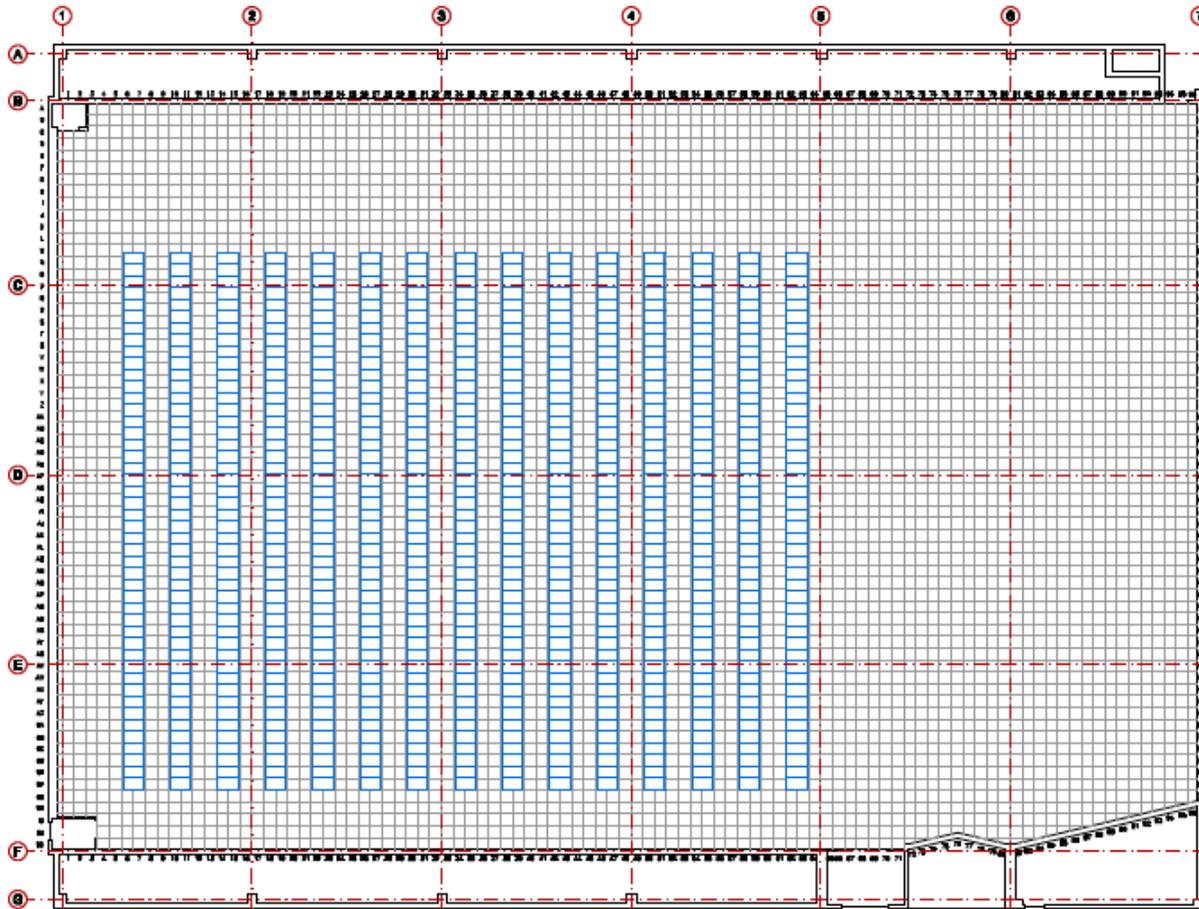
12. Section 12.1 ANL Facilities Overview: Fig 12.1 – (ANL siting area) Is the grey area in the figure 122 ft x 122 ft (meaning the yellow access areas indicated in the figure do not detract from the 122 ft x 122 ft siting area)?

Correct.

13. Section 12.2 LLNL Facilities Overview:

a. Fig 12.2 – (LLNL siting area) what are the length x width dimensions of the B453 siting area indicated by the purple area for CORAL-2? Is there any significance to the white lines that cut through the purple area in Fig 12.2?

Attached is an updated potential layout for CORAL-2 in B-453 depicting 4' aisles and standard size racks.



b. Fig 12.2 – (LLNL siting area) what are the length x width dimensions of the B654 siting area for the potential smaller CORAL-2 system?

A CORAL-2 system installed at B654 would be installed on the first level of the computer room not on the raised floor. It would be installed on the slab on grade. The area is about 3000 SF with columns every 16' on center so 3' aisles would be used for this installation.

14. Section 12.5.1 Minimal electrical and mechanical connections (TR-2): This requirement excludes the use of threaded cooling water connections. Can a preferred connection method to the facility piping be described (e.g. 150# ANSI flanged connections, brazed copper tubing, grooved mechanical joining, etc.)?

It is preferable that the Offeror provide a quick connect on the machine side via a hose with a NPTM barb and clamp where LLNL can supply a FNPT fitting to match the water supply end of the hose provided by the Offeror.

15. Section 12.5.3 Rack-Integrated In-line PDU (TR-2): If no in-line PDU is offered, the Offeror will provide a solution with one point of connection to each rack. Combining this requirement with ORNL's maximum 200 Amp circuit ampacity would limit that available kVA to single compute rack to be 133kVA. Is that the intent of this requirement?

No. The intent of the requirement is simply to minimize the number of connections to the rack. In this explicit example, Offeror may describe, for example, a pair of circuits per compute rack given the very high power-density of the offered solution.

Thanks for this observation. Section 12.5.3 has been modified to replace "one point of connection" with "minimize the number of connections."

16. Section 12.5.5 Tolerance of Power Quality Variation (TR-1): States there is no UPS available for the CORAL compute system, but then states the 208VAC to 240VAC components of the Offeror's solution which might include I/O, networking, or other infrastructure, may be supported by UPS systems in dual-fed-configurations. Please verify that customer facilities would provide the UPS systems should they be utilized. Additionally, please verify that 208VAC to 240VAC I/O, networking, and other infrastructure racks satisfy the overall requirements for the CORAL-2 systems. Would 208VAC to 240VAC racks provided with IEC 60309 pin-and-sleeve power connections be acceptable?

ORNL will provide a single 1200kVA battery-based uninterruptible power supply, with automatic bypass switch, and 1750kVA (1500kW) generator to support Offeror's solution. In their proposed solution, Offeror should demonstrate how this UPS might be incorporated into their design, including a maintenance scenario where the line-side input to the I/O, networking, or other infrastructure is unavailable for an extended time.

Offeror may choose to use 208VAC to 240VAC for these I/O, networking, or other infrastructure racks. It is expected that compute racks will use 480VAC. IEC 60309 pin-and-sleeve power connections for 208VAC to 240VAC connections are acceptable.

LLNL will also provide a single battery-based UPS but without a generator. LLNL's coverage is to cover transitions of power between the two utilities in the event of a short-term outage. We provide only 30 minutes of short term coverage of these ancillary loads. LLNL concurs with ORNL's answer on the IEC60309 portion of the question.

17. Section 12.8 Cable Management Requirements (TR-1): States non-conductive materials be used for cable management. Could sheet metal plenums also be accepted?

Non-conductive materials should be used for cable management that is external to the compute, I/O, network, or other infrastructure cabinets. The use of sheet metal plenums within the framework of a cabinet is acceptable.

Section 12.8 has been modified to clarify that this requirement is about cable management external to the cabinets.

Added January 18, 2018

18. Does CORAL prefer a single address space between CPUs and accelerators?

Yes, CORAL prefers a single address space, which is a programming convenience that enables faster and wider adoption of accelerators. Ideally, the programming environment will provide methods for using accelerator memory as a scratchpad as well as accepting malloc'd pointers. Section 5.1 of Draft SOW has been revised.

19. Does CORAL prefer high bandwidth connectivity for accelerators?

Yes, CORAL prefers that the connectivity bandwidth between CPUs and accelerators matches or exceeds the bandwidth of the system memory. If the connectivity bandwidth is less than system memory, the connectivity link reduces the effective bandwidth of the system memory. Section 5.1 of Draft SOW has been revised.

Added January 30, 2018

20. What is the largest problem size that the laboratories expect to run? A single 5 PB job? Multiple 2 PB jobs?

The laboratories have a mission need to run both single large jobs as well as multiple concurrent smaller jobs. The DRAFT SOW (section 3.2.3) indicates that either of these run modes will require up to 8 PiB of memory, with our application teams preferring 4-6 PiB (of the 8 PiB) being "fast and close" memory.

21. Can the laboratories provide file size distributions?

- The file size data for ORNL is at https://github.com/ORNL-TechInt/Atlas_File_Size_Data
- The file size data for LLNL is in the following image:

LLNL - file size data

fprof data from xxxxxxxx, May 2017

```
Directory count:      45,265,268
Sym links count:     10,430,723
Hard linked files:   309,219
File count:          1,342,586,738
Sparse files:        824,561,829
Skipped count:       2,977
Total file size:     3621.73 TiB
Avg file size:       2.83 MiB
Max files within dir: 26,546,573
Tree walk time:      16h 23m
Scanning rate:       23688/s
```

Fileset Histogram

Buckets	Num of Files	Size	%(Files)	%(Size)
<= 4.00 KiB	405,101,078	478.02 GiB	30.17%	0.01%
<= 8.00 KiB	125,265,283	711.78 GiB	9.33%	0.02%
<= 16.00 KiB	64,625,194	699.03 GiB	4.81%	0.02%
<= 32.00 KiB	60,576,954	1.33 TiB	4.51%	0.04%
<= 64.00 KiB	95,345,635	4.36 TiB	7.10%	0.12%
<= 128.00 KiB	132,993,524	10.06 TiB	9.91%	0.28%
<= 256.00 KiB	47,006,551	8.13 TiB	3.50%	0.22%
<= 512.00 KiB	113,780,368	42.45 TiB	8.47%	1.17%
<= 1.00 MiB	134,780,790	87.91 TiB	10.04%	2.43%
<= 2.00 MiB	72,754,593	95.37 TiB	5.42%	2.63%
<= 4.00 MiB	24,239,190	64.34 TiB	1.81%	1.78%
<= 16.00 MiB	40,482,662	316.71 TiB	3.02%	8.74%
<= 32.00 MiB	9,258,891	199.41 TiB	0.69%	5.51%
<= 64.00 MiB	9,478,784	386.65 TiB	0.71%	10.68%
<= 128.00 MiB	3,659,133	304.91 TiB	0.27%	8.42%
<= 256.00 MiB	1,654,034	289.96 TiB	0.12%	8.01%
<= 512.00 MiB	803,875	270.47 TiB	0.06%	7.47%
<= 1.00 GiB	381,657	259.78 TiB	0.03%	7.17%
<= 4.00 GiB	340,887	619.90 TiB	0.03%	17.12%
<= 64.00 GiB	57,012	449.67 TiB	0.00%	12.42%
<= 128.00 GiB	381	33.33 TiB	0.00%	0.92%
<= 256.00 GiB	105	18.75 TiB	0.00%	0.52%
<= 512.00 GiB	66	23.15 TiB	0.00%	0.64%
<= 1.00 TiB	65	46.65 TiB	0.00%	1.29%
<= 4.00 TiB	21	35.73 TiB	0.00%	0.99%
> 4.00 TiB	5	50.87 TiB	0.00%	1.40%

22. Re: Section 6.1.3.1.2, Small File Size Transactions Performance

How will this be measured — using mdtest and reporting the “File creation” rate with the 32KiB file size?

Answer: Yes

23. Re: Section 6.1.3.1.3, Single FEN Aggregate Transactions Performance

Can we confirm that with the single FEN generating 32KiB-files at 1 million/sec (32 GB/s) for 30 minutes, that would be 57TB of data from a single node?

Answer: Yes.

23. Section 6.1.3.1.4, Metadata Performance Scalability

For the 10 concurrent instantiations of the single FEN aggregate transaction performance, would there be a means of synchronization between the instantiations? For example, would the 10 instantiations only perform the file creation phase simultaneously before beginning the subsequent phases of mdtest?

The target is to measure the FEN performance under concurrent loads. synchronization of the overall mdtest instances is sufficient.

25. Section 6.1.3.1.5, Complete POSIX Namespace Tree Walk Performance

How is this to be measured? Is there a benchmark available for this?

ORNL has developed a benchmark for efficient parallel tree walk and it is called fprof and can be obtained from <https://github.com/olcf/pcircle/blob/master/man/fprof.md>.

26. Section 6.1.3.2[7|8], Shared File Non-Overlapping Write/Read Performance

Can we confirm that the parameters will be provided by the Offeror?

Yes.

27. Section 6.1.3.2[9|10], Shared File Overlapping Write/Read Performance

Can we confirm that IOR will be used for this test, and which IOR parameters will be used for testing overlapping I/O to a shared-file? Versions of IOR through IOR-2.x do not have that capability to our knowledge.

Sections 6.1.3.2.9 and 6.1.3.2.10 will NOT be included in the latest version of the SOW.

28. Section 6.2.4.1.2, CN Partition Aggregate Transactions Performance

How is this to be measured with mdtest? Mdtest will report the individual rates of open (create), stat, write, read, close. Can we confirm that the 15 million 32KiB-file create transactions is the reported mdtest rate of the "File creation" phase with a 32KiB-file?

Yes, the 15 million 32KiB file create is expected to be reported number by the mdtest for the file creation phase using 32KiB I/O.

29. Section 6.2.4.1.3, FEN Aggregate Transactions Performance

Can we confirm that the 1 million 32KiB-file create transactions per second is reporting the "File creation" rate from mdtest sustained for at least 20 seconds?

Yes.

30. Section 6.2.4.1.4, Metadata Performance Scalability

For the 10 concurrent instantiations, would there be a means of synchronization between the instantiations? For example, would the 10 instantiations only perform the file creation phase simultaneously before beginning the subsequent phases of mdtest?

A global synchronization between mdtest instantiations is sufficient.

31. Can we confirm that the performance specified in Section 6.2.4.1.2 and to be met with multiple CNs (X) will be met with 10 * X for this requirement?

Requirement 6.2.4.1.4 tests the scalability of metadata performance on the FENs by multiple (10) simultaneous users. It is an extrapolation of requirement 6.2.4.1.3, not 6.2.4.1.2. This was an error in the draft SOW and has been corrected.

**32. Section 6.2.4.1.5, Complete POSIX Namespace Tree Walk Performance
How is this to be measured? Is there a benchmark available for this?**

The fprof tool is recommended, see <https://github.com/olcf/pcircle/blob/master/man/fprof.md>.

**33. Section 6.2.4.2.[3|4], Shared File Write/Read Performance
Can we confirm that the parameters will be provided by the Offeror?**

Yes

Added February 8, 2018

34. Q. Can the Oak Ridge Leadership Computing Facility share any job size distribution information?

Yes, the attached file includes 2017 data from January, April, June, October, and December. Each record has SUBMITTIME, DISPATCHTIME, COMPLETETIME, and NUM_NODES. One can infer the length of time jobs sit in queue, the workload distribution, and how long certain sizes of jobs run for. This data is taken at the head of a month on a warm machine. If someone tries to replay the data, the first several days of data will be garbage because of the unknown state of the machine prior to the first record.

Please see the Titan Scheduling Policy to understand how its queue is managed. OLCF gives priority to jobs to using 20% or more of the machine, which OLCF considers Capability jobs.

<https://www.olcf.ornl.gov/support/system-user-guides/titan-user-guide/#358>

The job distribution information is in the file "Workload.tar.gz", attached on the CORAL-2 RFP web site.

35. Will the CORAL Labs make changes to the SOW MRs/MOs to allow a vendor to propose a storage solution only?

No. The Laboratories require a proposal for and intend to acquire an integrated solution.

36. What are the allowed source modifications to the baseline codes?

We are planning to update the first section of 4.4.5 in the SOW so that it reads as follows: The source code and compile scripts downloaded from the CORAL benchmark web site may be modified as necessary to get the benchmarks to compile and to run on the Offeror's system. Other allowable changes include optimizations obtained from compiler flags, and changes in the system software such as expected improvements to compilers, threading runtimes, and MPI implementations that do not require modifications of the source code. Source code changes to the baseline configuration are allowed to add or modify portable directives (e.g. OpenMP or OpenACC) or replace source code with standard library calls (e.g. LAPACK or BLAS). Once this baseline configuration is accomplished, a full set of benchmark runs must be reported with this unmodified or lightly modified source code. The extent of any modifications will factor into the evaluation of the projections with fewer changes preferred. If any source code modifications are made they must be documented and provided back to CORAL under the same license as the baseline.

37. What functionality is expected of the memory interface library (5.2.12)?

The functionality of the memory interface library to NVRAM should address performance, consistency and reliability of the data.

(i) The performance requirement has been explicitly called out in 5.2.12.

(ii) For data consistency, the APIs and semantics should ensure the safety of the variables in the face of application and system failures by avoiding bugs such as dangling pointers and locking errors.

(iii) For reliability, it is desirable to have automated methods to transfer the data to stable storage.

For more information on these issues, please refer to software libraries in the community such as NVHeaps

(<https://cseweb.ucsd.edu/~swanson/papers/Asplos2011NVHeaps.pdf>) and Mnemosyne

(<http://research.cs.wisc.edu/sonar/papers/mnemosyne-asplos2011.pdf>).

38. Would ORNL allow access to Titan in support of the work being done for benchmark responses? Is access to Summit available?

Titan and Summit-dev are available for use. A link with specific instructions and requirements has been added to the CORAL-2 RFP web site. Access to Titan or Summit-Dev are subject to the posted instructions and requirements. The Summit system is not available for use.

Added February 21, 2018

39. Is the Summit system available?

Summit is going through acceptance and the system is not available.

Added March 14, 2018

40. The draft SOW allows ORNL to choose from the set of proposals for systems delivered in 2022. Can we put in a proposal for a 2021 system and a separate proposal for a 2022 system that could also be delivered in 2021?

Yes, if the vendor wants to include an option to deliver their proposed 2022 system in 2021, then the proposal needs to clearly describe if there are any technology changes due to the early delivery and the pricing spreadsheet would need to include the pricing for the early delivery option. (If the early delivery option includes any additional NRE, this needs to be clearly spelled out and separated from the proposed 2022 system NRE). No additional pages are allowed. The option must be incorporated within the PEPPI specified page limits.

41. The draft SOW states that we can submit an individual proposal for the system delivered in 2021 and an individual proposal for the system delivered in 2022. It also states that we can submit a single proposal to cover both systems. If we choose to submit a single proposal, how should it be structured?

A single proposal will need to be structured per the instructions in the Proposal Evaluation and Proposal Preparations Instructions (PEPPI). The PEPPI allows higher page limits for such a single proposal and associated NRE. The PEPPI will become available once the RFP has received the required approvals. A single proposal must allow the Laboratories to easily distinguish the parts of the system that apply to the system delivered in 2021 from that delivered in 2022. The Laboratories will not make assumptions as to what does and does not apply to the different systems.

42. How many NRE proposals are allowed?

There should be one NRE proposal for each system proposal. If the vendor submits a single proposal to cover both systems, then there should be one NRE proposal that goes with this single proposal. If the vendor submits a separate 2021 system proposal and a separate 2022 system proposal, then each of these proposals should have a separate NRE proposal specific to it.

43. Is it acceptable for a proposal to including URLs to publicly available documents? An example might be a link to a complete API specification for a device or application.

The proposal should stand alone. URLs may be included to provide additional detail beyond what is required to address how requirements are met but the proposal text should have sufficient detail for a knowledgeable reviewer to understand the proposed solution.

44. The CORAL-2 SOW v20 states ORNL has 10 feet of height between the top of the raised floor and the bottom of the dropped ceiling. Are there any protrusions from the dropped-ceiling we need to be aware of (e.g. wet sprinklers, NOVEC nozzles, sensors, etc.). We are trying to understand how tall the racks can get before so we do not “bump” into any issues.

The Plan of Record (POR) space is Bldg. 5600 E102. ORNL has 138” from deck to deck in that space or 11’6” for Americans or 350cm for Europeans. Within the footprint that you would consider for a system, i.e. ignoring things like exit signs that are in the freight aisles, there are a number of objects that protrude from the deck, including vesda sensors, sprinkler heads, emergency lighting systems etc. Basically, all the normal stuff that you would normally expect to be in that space. The physical projection of these items never exceeds 3”, but of course a sprinkler head as an example will have a NFPA et al requirement of no less than 18” from sprinkler to object. This gives you 10’ of real working space (and an easy 11’ of delivery and installation space) without conflicting with the fire protection requirements.

Added March 27, 2018

45. Can you provide a better image of the space at ORNL where Frontier will located?

Yes, the image is now available on the website.

46. Can you provide more detail about section 5.2.3.1 including Performance Testing (PT) and Stability Testing (ST)?

The principal purpose of performance and scalability testing is to identify and fix any performance and scalability problems that may exist in the CORAL-2 systems that would interfere with the successful running of the ST for multiple days. Those tests provide an understanding of the behavior of applications over the full range of problem types and sizes on the system in order for the CORAL-2 laboratories to anticipate acceptable behavior for each application during the multiple days of ST.

Another purpose of the PT is to verify that the CORAL-2 systems can meet specific performance measures explicitly called for in the CORAL-2 SOW.

These tests may also be run for additional iterations in a scripted/batch mode during an extended period for additional stress testing and to verify consistent results.

47. Does the stability testing phase fill the system with application instances or just 1 application instance and the remainder of the system idle?

Stability testing involves a variety of job sizes that reflects the intended production usage of the system.

48. There are many potential sources of variability such as slow hardware, slow network links/switches, poor placement by the scheduler, I/O subsystem variability, as well as processor power management when not using some form of “QOS mode” which forces the processors to perform the same. Do you expect “QOS mode” to be the predominate policy for CORAL2 or would other power/performance modes be used?

The Laboratories expect an Offeror's proposal to reflect the best system possible. The Offeror should propose policies that will allow the Laboratories to balance job performance, system throughput, and power consumption/operating costs. If reducing variability complements this goal or is at odds with this goal, explain why.

49. Section 6.1.2.4 Metadata Capacity (TR-1). "The I/O subsystem will support the storage of at least 500 Billion files, at least 500 Billion directories, and at least 10 million files in a single directory." What is the number of inodes (is it 1000 Billion or 500 Billion), the number of files, and the number directories?

1,000 Billion inodes, 500 Billion files, 500 Billion directories.

49.a. The ORNL CORAL SOW states 30 Billion files, CORAL-2 states either 500 Billion or 1000 Billion files (depending on previous question). Is the requirement for 500 Billion (or 1000 Billion) files correct given that it is as much as 30X greater than ORNL CORAL SOW requirement? If so, why did it scale up 30X when compared to CORAL?

1,000 Billion is correct.

50. Section 6.1.3.1.2. Small File Size Transactions Performance (TR-1). "The I/O subsystem will provide an aggregate of 15 million transactions per second for parallel file create operations where 32 KiB is written into each file from the CN partition in parallel using a sufficient number of CNs. The Offeror will describe the assumptions and required number of CNs used to achieve this performance" What is the duration of the test and number of iterations?

The test duration is 20 minutes and 3 iterations are required.

51. Section 6.1.3.1.3. Single FEN Aggregate Transactions Performance (TR-1). "The I/O subsystem will sustain 1 million transactions per second, for parallel file create operations where 32 KiB is written into each file from each FEN for a duration of 30 minutes. The Offeror will report the performance of open, create, stat, write, read, and close operations separately."

This TR has been modified. The new language clarifies the requirement. The "read" operation is no longer required as part of this TR. The transaction is defined as a sequence of "create, open, stat, write, and close" operations on a given file.

51.A. Does one "transaction" consist of open, create, stat, write, read, and close?

See above answer.

51.B. Is the POSIX unlink operation optional?

Yes.

51.C. Is the POSIX read operation required?

See above answer.

51.D. Is the requirement for only one iteration of duration 30 minutes?

The 30 minutes is the whole duration of the test, including all elements.

52. Section 6.1.3.1.4. Metadata Performance Scalability (TR-2). "The I/O subsystem will support 10 concurrent instantiations of the metadata performance requirements specified in Section 6.1.3.1.3 without any degradation in performance. The Offeror will describe any assumptions made, e.g., how far apart in the directory tree the processes

must be to meet this requirement.” Is the performance requirement for the aggregate 10 concurrent instantiations an aggregate 10 Million transactions per second?

Yes. The target is to measure the FEN performance under concurrent loads. synchronization of the overall mdtest instances is sufficient.

53. Section 6.1.3.1.5. Complete POSIX Namespace Tree Walk Performance (TR-1). “The I/O subsystem will walk the entire global namespace with 500 Billion directories and 500 Billion files in not more than 18 hours. The Offeror will describe all assumptions required to achieve this performance.” What is the number of inodes (is it 1000 Billion or 500 Billion), the number of files, and the number directories?

1,000 Billion inodes, 500 Billion files, and 500 Billion directories.

53.A. Define “walk the entire global namespace” – is this stat, policy engine? What tool should be used?

ORNL has developed a benchmark for efficient parallel tree walk and it is called fprof and can be obtained from <https://github.com/olcf/pcircle/blob/master/man/fprof.md>. Fprof will be used as a benchmark for this test.

54. Section 6.1.3.1.6. Sustained unlink () Performance (TR-1). “The I/O subsystem will perform a parallel unlink () operation of 100 million files in not more than 2,000 seconds (approx. ½ hour). The Offeror will describe all assumptions required to achieve this performance.” What is the file size distribution and the distribution of files in directories?

The labs’ file systems have many empty directories due to purging and most directories have many files (number of ranks * N where N is 1-16).

55. Section 7.3.1 Low-level Network Communication API (TR-1). “Can Verbs be added as an example of a LLCA in 7.3.1? Would Verbs be acceptable as a LLCA?”

The list of example LLCA in 7.3.1 is not exhaustive. Yes, Verbs is an acceptable LLCA.

56. Section 7.3.1.4 Accelerator-Initiated/Targeted Operations (TR-2). Would CORAL consider rewording the requirement to: If the system has multiple processor types (i.e., accelerators or coprocessors), each processor type will be able to initiate LLCA communication operations without explicit host activity and the LLCA will support RMA communication to each processor type’s memory without explicit activity by the remote node host. Accelerator initiated communication shall be compatible with the LLCA.

No, the intent of this TR-2 is to have accelerators as first-class citizens when communicating with peers. An ideal solution would allow the accelerator to access the network interface directly and at a lower latency than having the accelerator coordinate with the CPU to schedule communication.

57. Section Item 8.3.3.4 covers container support. Is the intent of this requirement for the application launcher to be able to launch a fully external, Government provided container runtime service, or for the application launcher to provide the container runtime service?

Neither, the intent is to allow a job launcher to launch a container runtime (i.e. the program that images creates/starts/stops/execs container images) that will then start container images.

58. Section 9.2.1.8 Baseline Language Support for OpenMP Parallelism (TR-1). "All baseline language compilers will include the ability to perform automatic parallelization." OpenMP parallelism is directive based. To reflect that, can this requirement be changed from "automatic parallelization" to "parallelization"?

You can read that as "OpenMP parallelization".

59. Section 9.2.1.9 Baseline Language Support for OpenACC Parallelism (TR-2). "The Offeror will support interoperability of OpenACC with OpenMP 5.0 (or then current directives)." If multiple compiler suites are included in the offer, does this require interoperability of OpenACC and OpenMP across different compiler suites?

No.

60. How many applications depend on Co-array Fortran support (in the context of Fortran 2008)?

Currently, Co-array Fortran is not heavily used in DOE applications. However, many DOE application code teams would like to use Co-array Fortran. Its wider-spread adoption has been limited by the lack of software support and optimization on modern platforms. We want modern standards supported and optimal implementations provided to allow our codes to use these features for a variety of potential use cases.

61. In Section 6.1.3.2.2, ORNL states that the expected check point interval is hourly. What is the corresponding expected frequency of checkpoints at LLNL?"

The answer is "Hourly also" for LLNL.

Added March 28, 2018

62. Can you provide some details about LLNL's interest in cognitive simulation as a possible activity within the CoE?

Yes, please see https://asc.llnl.gov/content/assets/docs/EICap_IntelligentSimCOE_whitepaper.pdf

Added April 5, 2018

63. Can the pools of NRE funds and system funds be manipulated so that funds from one pool could be moved to the other pool? For instance, could funds for the system build pool be moved to the NRE pool or vice versa?

Response:

The Offeror should understand that the reallocation of funds between NRE pool and system build pool would be agreed upon during pre-award negotiations. The Offerors should be cautious about the Laboratories willingness or ability to agree to significant allocation changes among the pools. The availability of funds in a given year at a given laboratory also affects the ability to agree to reallocate these funds.

ANL or ORNL acquisitions: It is possible to move funds from the NRE pool to the system build pool. It is much more difficult to move funds from the system build pool to the NRE pool. Changing the split after contract signing may be possible but would be much more difficult.

LLNL acquisition: It is possible to move funds from the NRE pool to the system build pool and vice versa. Changing the split after contract signing may be possible but would be much more difficult.

Added April 13, 2018

64. Is the DOE accepting responses from vendors that can address a portion of the RFP, in this case, a workload manager or does it have to be a complete solution: hardware, storage, software etc.?

To clarify things, UT-Battelle LLC issued the RFP, not DOE.

We are not accepting responses for a portion(s) of the solution. We require a proposal for the complete solution.

65. a) We reviewed the documents associated with the CORAL 2 solicitation and don't see a Sample CORAL-2 Build PO for ORNL document.

b) This document is referenced in the NRE volume: "For more information on the CORAL-2 decision process, see article IV of the Sample CORAL-2 Build PO for ORNL or article 13 of B626589 Sample Build Subcontract for LLNL."

a) The sample CORAL-2 Build PO is the Draft Lease Agreement. The title will be corrected.

b) The process is actually addressed in LLNL's sample build subcontract Article 14, not Article 13.

66. Questions on Coral-2 SOW, Section 12

General Power Questions

1. Is the 100 kAIC series rated fuse protection preferred inside the PDU per branch circuit or upstream from the PDU?

Inside the PDU is preferred and provides the best protection.

2. Is it ok to have no internal PDU power measurements since these are redundant to the facility collection system? It is expected that the power consumption per node can be obtained and summed by the Offeror's tool (e.g., xCAT).

Internal PDU power measurements are preferred.

3. Is it acceptable to have 3x 480/277 Vac connections to each rack?

One is better. If it is hardwired, this is easier to achieve. What is the projected power ampacity of each rack? Having this information will help answer the question. The requested information should be part of the proposal.

12.2 LLNL Facilities Overview

1. What is the expected air temperature entering the racks?

70F to 75F

2. Will LLNL provide a facility CDU to distribute water to the individual racks?

Not planning to but can accommodate it if CDUs are vendor provided.

1. If so, what secondary temperature will be provided to the servers?

N/A unless vendor is providing CDUs. ~71-100F for in row CDUs.

2. If so, can LLNL provide the water quality to the racks to be equivalent to that provided currently on Coral?

Water quality can be achieved for any solution at LLNL, CDUs or no CDUs.

3. If LLNL will not provide a facility CDU, what is the primary water temperature that is provided?

60F to 85F

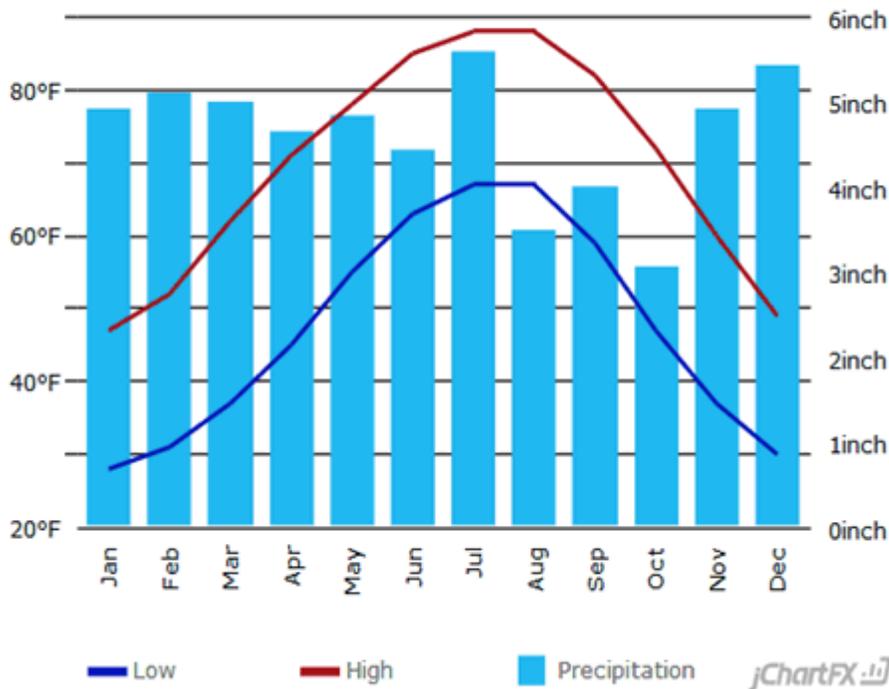
4. In either case, how will the water temperature vary throughout the year?

Across that range above. Colder in the winter and warmer in the summer.

12.3.3 ORNL Facilities - Mechanical Distribution

1. Outside temperature at ORNL can be summarized per

Oak Ridge Climate Graph - Tennessee Climate Chart



Does ORNL expect to provide facility supply water that varies throughout the year, within a few degrees of the chart above?

ORNL – Facility High Temperature Water (HTW; no chillers are involved in the production of HTW) water is expected to be 60-85F, ranging through the year, colder in the winter and warmer in the summer. Facility upgrades will be required to be able to provide the HTW. HTW is intended for compute racks and for RDHX when in the cooler temperature ranges. Facility chilled water (CHW) is 42F and this will be used for

dehumidification and miscellaneous air-cooled loads within the data center and used in conjunction with a CDU for racks that exchange air with the data center to cool HTW down to the point RDHXs become room neutral.

2. For support racks, the SOW states:

For support cabinets, the Offeror's solution is to utilize the W3 water when possible, but provide means to return rack discharge air to acceptable conditions using Facility chilled water when the W3 water is too warm
What is an acceptable return rack discharge air temperature?

ORNL – discharge air temperatures ideally match the rack inlet air temperatures.

What temperature is the facility chiller water maintained at?

ORNL – Facility chilled water is at 42F.

Is the use of Rear Door Heat Exchangers (RDHX) acceptable per this requirement? Here, W3 water would be used for much of the year when the W3 water temperature is low enough to meet the "acceptable return rack discharge temperature", and either chilled water or a mixture of W3 and chilled water would be used to cool the RDHX when the W3 water is not cool enough. Can the W3 loop be mixed with the Facility chilled water loop via ORNL controls?

ORNL – We will not mix HTW and CHW. An Offeror provided CDU piped in an arrangement to serve Offeror provided RDHXs and controlled in a way to monitor the supply temperature of the RDHXs it serves is required. If the HTW gets too warm for the RDHX to be room neutral, the CDU will engage and bring the supply temperature down. These CDUs preferably include a pump otherwise the pressure drop through this additional heat exchanger must be acceptably low.

3. For compute racks, the proposal states:

During normal operation, no air exchange will be allowed between the interior and exterior of the compute rack unless the specific site can accommodate this arrangement. During short-term maintenance, less than 2% of the system load may escape as parasitic heat to the data center ambient.

To be 100% certain, the requirement is "no air exchange" and not "no heat exchange". Is that correct?

ORNL – the requirement is no air exchange. As soon as air is exchanged with the data center and holding on to the room neutral requirement, the water supply temperatures must drop to accommodate the components with the most stringent requirements, those that are air cooled. This assumes a single water supply and return. The facility providing two supply temperatures via two piping systems is not an option.

If so, that implies water cooling the vast majority of the heat by water and then removing the rest of the heat using fans and a RDHX is not an acceptable solution. Do you agree?

ORNL – An RDHX on compute racks is not an acceptable solution assuming the majority of the air going through the RDHX has also gone through the compute nodes.

Does this apply to all of the components that the Offeror places in a compute rack (ie, Ethernet switches that could be placed in the compute rack to reduce the cable lengths in the total solution.)

ORNL – The amount of heat load not captured by the HTW must be quantified and located on a plan view of the data center floor to be able to answer this question.

What does the phrase "unless the specific site can accommodate this arrangement". Is interior to exterior air exchange acceptable at ORNL?

ORNL and LLNL facility designs differ when it comes to the ability and method of cooling air-side loads within the data center space. LLNL's air handling is done outside the whitespace, while ORNL's must be done within the whitespace.

During short-term maintenance, less than 2% of the system load may escape. Does the "system" refer to the entire installation and not the power of a single rack?

ORNL – 2% of the overall system load.

The proposal states:

The Offeror's CDU is to provide variable secondary water flow based on changing primary temperature and flow as well as changes in cooling demand by the connected IT equipment.

Providing variable secondary water flow based on changing primary temperature and flow is understood. As the primary temperature decreases, so can the secondary temperature, which can be accompanied by a decreased secondary flowrate. This would lead to reduced pump power and warmer return water temperatures

Providing variable secondary water flow based on changes in cooling demand requires a more extensive discussion. Within the scope of this initial analysis, the Offeror does not believe that providing variation in secondary flow for every individual server (including a valve for every node and the intelligence to control it) is economically justified based on the associated costs and complexity, while resulting in limited benefits. If ORNL disagrees, please elaborate on the potential benefit you hope to achieve.

ORNL - The method for variable secondary water flow is the responsibility of the Offeror.

For the Offeror water manifolds, we can control the water flowrate to the manifold, but that flow is evenly divided upon the N nodes in the rack. Hence the node with the higher power dissipation defines the flowrate/node and the flowrate/rack which is $N \times \text{flowrate/node}$ (for the node with the most power). Hence it only takes one node at max power to require max flowrate to the rack (for that given primary water temperature) if that one hot node is to receive sufficient cooling

4. Can ORNL comment on the expected uniformity of power from node-to-node within a rack.

- For instance, would ORNL exercise every node in the rack to the approximate same power level? That is, as a general rule, all of the nodes will have similar power, whether it is max power or idle conditions, or
- Is there no general trend for the rack, where as a general rule some nodes would be at idle while others will be exercised at max power?

If the power from node-to-node was expected to always be similar, and flowrate to the rack was based on a rack level power consumption or a rack level bulk heating of the water, would it be acceptable to deliver sufficient flowrate assuming a uniform power dissipation from nodes to node, which could imply insufficient flow and hence throttling for a node that had dramatically higher power dissipation than the average node power dissipation?

ORNL – No expectation of the usage model can be assumed.

5. The SOW states:

The CDU's primary flow control valve is to be a fail closed, two-way modulating valve with pressure independent flow control and balancing.

If the primary flow control valve fails in the closed position this is to the benefit of minimizing facility water, but it implies a loss of cooling to all of the servers and every server will shut down. Is this what ORNL desires?

ORNL must be able to isolate the HTW CEP from any CDU that is out of service.

6. Table 12-1:

What is expected from a thermal data for the rack? Is the collection of data from the components of the nodes within the rack sufficient?

ORNL – Thermal data is desired at whatever level the Facility interfaces with the Offeror’s cooling solution. In the case CDUs are provided for the compute racks, ORNL prefers to see the CDU’s primary flow control valve position. Depending on how the Offeror’s overall system interacts with the Facility, other data may be desired.

67. Question about the CORAL 2 Cover Letter.

In the RFP cover letter and sample Argonne Equipment PO, there are indications a LTO offer would be subject to a Master Lease to Ownership Agreement and a Lease to Ownership Order. We're not finding these documents in the solicitation attachments, however. Are they to be provided for consideration with our proposal?

Yes. The website has been updated. The documents have been added to the list of documents for ANL’s Build Specific Documents. The RFP cover letter has been updated to reflect the change to the list of documents associated with the RFP. RFP Cover letter, Rev 7 and Rev 1 to the summary matrices have been posted.

Added April 19, 2018

68. About the input data of the BDAS benchmarks, can the size of the feature column be changed from 250 to 256?

A: No. The size of the feature column was chosen to be representative in data analytic problems we’ve seen. While we understand the optimization perspective of have a binary representation, we want to see how the codes perform in a more realistic situation. Code optimization without altering the problem is encouraged.

69. At ORNL, what is the smallest door opening between the loading dock and the final room?

The smallest opening is 106.5" x 60" (H x W).

70. At LLNL, can we assume that power is delivered to the racks from overhead or below the raised floor?

Below the raised floor only.

71. SOW Section 11.1.1 Hardware Maintenance Offerings (TO-1), Page 83 and Price Schedule

Is it a correct assumption that the Government is asking for 9x5xNBD (Next Business Day) maintenance even though the price schedule shows 1 hour response on the 9x5 HW Maintenance Tab?

Yes. The SOW text has the intent which is one-hour response within the 9x5 window. Outside the window, the response time is one-hour within the Next Business Day.

72. PEPPI Section 5.3.8 Additional Licenses, Page 14

Is there a typo in PEPPI section 5.3.8?

5.3.8 Additional Licenses

Offeror must provide option pricing additional I/O client software licenses for 2,500 nodes of x86, ARM, and Power architectures as described in Volume 1, Section 6.1.8.2.1 and for increments of 1,000 additional I/O client software licenses as described in Volume 1, Section 6.2.8.3.

73. The CORAL-2 pricing worksheet has software maintenance as a line item in the Base System worksheet and hardware maintenance in separate worksheets. (a) Can the maintenance pricing for both the software and hardware bundle be put in one place or the other? (b) If the vendor has a 'recommended' maintenance plan, can the hardware and software support pricing be bundled together with the vendor describing how it meets the CORAL-2 SLA requirements in that bundle? Using the CORAL-2 pricing format as presented in the solicitation, it will be hard to separate out the pricing for the on-site personnel between hardware and software.

- (a) No.
- (b) No.

74. The RFP cover letter states that in taking advantage of the patent waiver a large business must fund at least 40% of the total price of performance. Section 5.1 of the PEPPI references identifying cost share for each milestone. (a) Is UT Battelle open to an offeror determining cost share on a per milestone basis, i.e. some milestones have cost share and some do not, provided the overall cost share is at least 40% of the total price of performance? (b) Alternatively, may an offeror apply the patent waiver only to select milestones?

- (a) No
- (b) No

75. The Phloem_MPI_Benchmarks_Summary_v1.0.pdf document specifies examples of interest. We have several questions below.

The summary file above is months out of date. Please use the latest summary files, benchmark source code, and benchmark results spreadsheet. This will address many of the questions.

75.a. Would it be acceptable to provide projections for the mpiBench benchmark run without '-d 2 -p 2' arguments? I.e. like mpirun -n <N x the number of cores per node> ./mpiBench
This would reduce the number of projections required.

No. Removing those arguments eliminates subcommunicator results. The use of subcommunicators is critical to many applications. We are trying to understand the MPI performance with many subcommunicators. The Offeror may have a general performance target that covers more than 1 specific case.

75.b: For SQMR benchmark, would it be sufficient to provide results for two and four nodes with one process per core?

The latest summary file and benchmark results spreadsheet contain updated messaging rate requirements.

75.c: Please confirm that no projections for the mpiGraph benchmark is expected.

We are not requesting mpiGraph results.

76. Reference: Attachment 2, RFP 6400015092, 4.5, Section 5. Subcontracting, page 11.

The RFP states that "The Offeror should describe any previous experience with the proposed third-party subcontractors and the experience that the proposed third-party subcontractors had on projects for similar equipment or services as being provided under the anticipated CORAL-2 Build subcontract. The Offeror must include proof of demonstrated experience and past performance for all proposed subcontractors and a commitment from integrated subcontractors to participate in the work."

Given that the information requested in this section is similar to the information requested in the "Supplier Attributes" (Volume 2, Section 1) section, can the offeror reference the information submitted in response to "Supplier Attributes" in this section?

The requirements in each section are different. It is up to the Offeror as to how they want to respond to each.

77. Reference: Attachment 2, RFP 6400015092, 4.5, Section 5. Subcontracting, page 11. The RFP states that "This section should describe any use of subcontracting or third parties for major software, hardware components, or services and associated areas of risk and risk mitigation."

Can the Government clarify what constitutes a major software, hardware component, or service?

- a. No. It is up to the Offeror to make this determination.
- b. NOTE: UT-Battelle LLC and the other managing and operating (M&O) contractors of LLNL and ANL are not the Government or representatives of the Government. This statement also applies to questions 78, 79, 80, and 81.

78. Reference: Attachment 2, RFP 6400015092, 6.8, Section 8. Workplace Substance Abuse Program Plan, page 17. The RFP states that "Before the work can begin, the Offeror selected for award must submit a written Workplace Substance Abuse Program Plan (WSAPP) or WSAPP Certification consistent with 10 CFR 707 for LLNS approval. Any lower-tier subcontractor's WSAPP must be approved before the lower-tier subcontractor is allowed to perform the work."

Can the Government please confirm that a WSAPP should not be included in the LLNS proposal and is not required until time of award?

As stated in RFP Attachment 2 (PEPPI), LLNS requires a WSAPP(s) from the selected offeror prior to award. It is up to the Offeror(s) how to address this subject within a proposal.

79. Can the Government confirm that the Draft SOW version 21 dated March 30, 2018 is considered the Final SOW?

Please see PEPPI Section 9.3, which includes this paragraph:

Target Requirements (designated TR-1, TR-2, or TR-3), identified throughout the SOW, are features, components, performance characteristics, or other properties that are important to the Laboratories but that will not result in a nonresponsive determination if omitted from a proposal. Target Requirements add value to a proposal and are prioritized by dash number. TR-1 is most desirable to the Laboratories, while TR-2 is more desirable than TR-3. MRs, MOs, TOs, TRs, and additional features proposed by the selected Offeror(s), and of value to the Laboratories, will be included in a final negotiated SOW(s) and incorporated within the resulting subcontract(s).

Note: Version 22 of the Draft SOW applies as a result of Amendment 1 to the RFP.

80. Reference: Attachment 2, RFP 6400015092, Section 1 Proposal Format, page 1.

The RFP states that "All proposals should be prepared using an 8-1/2 by 11 in. paper format and a minimum font size of 11 points."

Will the Government allow offerors to utilize 8pt font for proposal graphics?

Any text in graphics should be easily read when the proposal is printed

81. Reference: Attachment 2, RFP 6400015092, Section 2 CORAL-2 Build Technical Proposal (Volume 1), page 3. The RFP states that "SOW text should be included but may be formatted with a smaller font." Will the Government clarify if the SOW text counts against the page count in Volume 1?

Yes.

Added April 24, 2018

82. In accordance with the Proposal Evaluation and Proposal Preparation Instructions section 6.5, representations and certifications are being requested for all 3 labs. We're not finding a representations and certifications form for ANL in the solicitation materials. If required with proposal submission, please provide.

The representations & certifications for ANL's that applies to ANL's Build Specific Documents is now included, see the form 70B Reps & Certs. The document was left out in error. The Summary Matrices – Laboratory Specific Documents Rev 2 has also been updated.

Added April 26, 2018

83. In the proposal instructions under Other Documents section 6.6 we are requested to provide a completed EEO pre-award clearance request form applicable to ORNL awards. In the section Documents applicable to the Coral2 RFP General RFP documents, this form is carried under the ORNL specific NRE contract documents section only. Is this form to be completed for the build proposals as well?

Yes

84. SOW Page 88, Section 12.3.3 ORNL Facility Mechanical Distribution, pertaining to this statement: *During normal operation, no air exchange will be allowed between the interior and exterior of the compute rack unless the specific site can accommodate this arrangement.*

Question: *May the vendor put one air-cooled, off-the-shelf control network switch at the top of each compute rack? The 95 Watt heat load generated by the switch will be dissipated into the machine room air. Except for this control network switch the compute rack will otherwise be sealed, having no air exchange with the room.*

ORNL can allow this load to transfer from the interior to the exterior of a cabinet so long as the cumulative load transferred does not exceed what is proposed in the question (ie. 95W per rack).

85. PEPPI Section 2.1 and Section 8.4 (Summary Matrices Spreadsheet)

- Section 2.1 of the PEPPI states: that the summary matrices should be included at the beginning of the executive summary in *Volume 1*, Section 3 and as part of *Volume 7*.
- Section 8.4 of the PEPPI states that the summary matrices spreadsheet be included in *Volume 3* and as a separate attachment for *Volume 7*.
- Should Section 8.4 state "Volume 1 Section 3" instead of "Volume 3"?

Yes

86. SOW Section 9.2.1.10 – Sanitizer Instrumentations. The SOW does not include a TR level. Please enumerate.

See Amendment 1 to the RFP. The TR level is "2". Rev 22 of the Draft SOW posted 4/23/2018 now includes the TR level.

87. PEPPI Section 4.6 Other Research & Development

Section 4.6 describes a section 6, Other Research & Development, to be included in Volume 3 CORAL-2 NRE Technical Proposal that is not listed in the PEPPI Table 1, Proposal Format. Can you please confirm that this section should be included in Volume 3?

Yes. Amendment 2 to the RFP has been issued and Rev 13 of the PEPPI has been posted.

88. General

Are covers, front matter, tables of contents, glossary, and spreadsheets included in the maximum page count limit? Given the requirement to include the SOW requirements in the document, considerable page count is already consumed by the requirements, even at a reduced font size.

No.

89. General

Can the required spreadsheets be submitted as attachments instead of being embedded in a document?

As stated in the answer to #1, the Summary spreadsheets need to be included at the beginning of the executive summary of Volume 1 Sections 3.

90. PEPPI Section 5.3 Build – Mandatory Option and Technical Option Fixed Prices & CORAL-2 Price Schedule
Section 5.3 states that Offerors must complete the Optional Pricing tabs contained in the CORAL-2 Price Schedule spreadsheet. We do not see any Optional Pricing tabs within the CORAL-2 Price Schedule spreadsheet. Can you please clarify this requirement?

All tabs after the Risk Sharing tab are Optional Pricing tabs.

91. PEPPI Section 5.2 Build – CORAL-2 System Fixed Prices & CORAL-2 Price Schedule

Section 5.2 requests "maintenance options should include at least the 24x7 and 12x7 models. The CORAL-2 Price Schedule only has two tabs for hardware maintenance (9x5 and 12x7). Please clarify what hardware maintenance pricing is required.

PEPPI Section 5.2 should read, "maintenance options should include at least the 9x5 and 12x7 models". Revision 13 of the PEPPI has been incorporated correcting this error, see Amendment 2 to the RFP.

92. In preparing our response to the CORAL-2 RFP, we believe there is an error in the Benchmark Summary Spreadsheet:

http://procurement.ornl.gov/rfp/CORAL2/05_CORAL2_Benchmark_Results_v16.xlsx

Summary Tab

Cells B55, B56 $= (5 * B53 + 1 * D53 + 2 * H53 + 5 * I53) / 18$

It is our understanding that this formula is intended to generate a weighted mean of 4 benchmarks, in agreement with the description given in the deep learning suite summary as:

The overall FOM is computed separately for base and optimized run as weighted mean of the four FOM values from candle, convnet, RNN, and ResNet-50 for Imagenet, where the weights are 5.0, 1.0, 2.0, and 5.0, respectively. The weights for the coefficients match, but instead of dividing by 13 (sum of weights) the denominator is 18. We believe the denominator should be 13, not 18 as used in the official spreadsheet.

The error in the Benchmark Results spreadsheet has been corrected. Revision 17 of the CORAL2 Benchmark Results spreadsheet has been posted. Amendment 2 to the RFP has also been issued.

93. The draft SOW on page 88, Section 12.3.3 ORNL Facility Mechanical Distribution, has this statement: "During normal operation, no air exchange will be allowed between the interior and exterior of the compute rack unless the specific site can accommodate this arrangement."

May the vendor put one air-cooled, off-the-shelf control network switch at the top of each compute rack? The 95-Watt heat load generated by the switch will be dissipated into the machine room air. Except for this control network switch the compute rack will otherwise be sealed, having no air exchange with the room.

Offeror must describe the circumstances under which there is any deviation from the statement "During normal operation, no air exchange will be allowed between the interior and exterior of the compute rack".

Added May 3, 2018

94. Attachment 2, Volume 1 Section 13, page 8.

The proposal format requires the Offeror to address Volume 1, Section 13 Project Management. On page 8, there are no instructions listed for the Project Management content between Section 12 and Section 14.

Will instructions for Volume 1 Section 13 be provided?

The first paragraph of Section 13.0 provides the instructions for completing the section. Specifically, it states:

"Documents described in this section are not required in the RFP response; however, the Offeror will confirm its commitment: 1) to include the following project management approaches and elements in its execution of any CORAL subcontract awarded; and 2) to provide the associated documentation by the required times and through a reliable and easily accessible mechanism that supports change control. The Offeror will provide in its RFP response a set of milestones for the deliverables in this section as described in the Key Build Phase Milestone Dates."

Offeror may also describe any additions or extensions to the required documentation as well as any specifics of the mechanism that through which they will implement the required documentation.

95. Attachment 2, 2.5 Facilities Requirements (Section 12 SOW), page 8.

For the Offeror's response to Section 12 CORAL Facilities, can the Offeror omit sections related to labs that they are not proposing?

Yes, however, ORNL has the right to choose from the 2022 proposals and LLNL has the right to choose from 2021 proposals. By not providing a facilities plan for a laboratory, Offeror would significantly impede that laboratory from selecting their proposal.

96. SOW, 6.1.6.3 Reliability (TR-1), page 38

Can LLNL and ANL characterize the expected average lifespan of files in the global namespace? What is the expected global namespace daily/weekly/annual change rate over the course of 5 years?

CORAL does not have data upon which to forecast an expected average lifespan or expected change rates. Offeror should provide a solution that can sustain the I/O rates required in Section 6.2.4.2 for the contracted lifespan of the CORAL system.

97. Attachment 2, 5.1, Section 1 NRE Fixed Price

Because the COE component of the NRE is not related to intellectual property for new technology, we believe that this component would not be included in the cost sharing approach that applies to the remainder of the NRE components, to which the waiver on intellectual property would apply. Can Battelle confirm that this is the case? We understand that the answer to Question #74 clarified that the cost sharing must be on the entire set of NRE milestones [please confirm the "NRE milestones" that I added], however, the COE is a required Task in the Offeror's NRE proposal. Because the COE goal is to support porting and optimizing the Laboratories' applications to the CORAL-2 architecture, there will be no intellectual property to be waived. Given this, we don't believe that the costs of the COE component of the NRE proposal are intended to be included in the total price for cost sharing

As stated in the PEPPI, COE activities are a mandatory requirement of the NRE proposal and "This activity must reflect all terms and conditions of NRE activities including cost sharing." This is a requirement and is not open to negotiations.

98. SOW, 12.2, page 86.

The SOW notes: "Both B-453 and B-654 are unique facilities in that their construction consists of single story two level computer rooms. This design affords the capability of siting a machine with higher weight capacities on the slab on grade of the first level of the computer room"

Q&A 13(b) notes: "A CORAL-2 system installed at B654 would be installed on the first level of the computer room not on the raised floor. It would be installed on the slab on grade"

Is the intention of the customer to allow the offeror to recommend a siting position, or does LLNL have a defined configuration for each building?

How does air cooling capacity for both LLNL facilities differ in these configurations?

There is allocated space in each facility for the location of the proposed system by LLNL.

B-453 has adequate room air flow available from existing air handlers. B-654 would require rear door heat exchangers for air cooling.

99. SOW 12.2, page 87

The SOW describes the LLNL B453 siting area in terms of square feet, and also provides a sample layout in the Q&A 13(a).

It is still unclear what the specific dimensions of the siting space in B453 is intended to be. Can LLNL provide engineering documents related to the space, and/or specific measurements including any columns?

What is the maximum height available in B453?

B-453 has no columns or above floor impediments to site the system. The sample layout is indicative of the square footage available to site the machine.

The maximum height of B-453 is 10'6" in the computer room from the top of the floor to the bottom of the ceiling tiles. A minimum of 18" of clearance is required to the top of the racks from the ceiling.

100. SOW 12.2, page 87

The SOW describes the LLNL B654 siting area in terms of square feet, and also provides details on columns arrangement.

It is still unclear what the specific dimensions of the siting space in B654 is intended to be. Can LLNL provide engineering documents related to the space, and/or specific measurements including any columns? If there are not defined limits to the length and width, does LLNL prefer the offeror to provide suggested dimensions to fit within 3,000 SF? If so, what are the maximum dimensions?

B-654 has columns as indicated. The sample layout is indicative of the square footage available to site the machine. Please provide a suggested layout to fit within 3,000 SF.

101. SOW 12.2, page 87

The SOW does not include information on the facility electrical distribution details for the LLNL siting locations. Should Offerors assume 3-phase Y 277/480 can be made available?

480/277V is available in all facilities at LLNL.

102. Statement of Work Section 11.6, Clearance Requirements for CORAL Support Personnel at LLNL

Section references the need to obtain "DOE P" clearances for repair actions. Will the Government please clarify the level of a 'DOE P' clearance in relation to Q and L level clearances?

The RFP will be modified to refer to a DOE P "approval" rather than a DOE P clearance. A P approval is not a clearance. Rather, it is the result of a background investigation that is conducted to provide information to the Department of Energy as part of the initial Q/L clearance process or in this case to provide long-term site access to Offeror's support personnel. In most cases the scope of a preliminary background investigation is three years.

103. The price volume references 480VAC and 600VAC options. However, the SOW specifies only a 480VAC requirement. Please delete the 480VAC and 600VAC options from the price table or clarify if something beyond the base requirement is being requested.

Revision 4 of the Price Schedule spreadsheet resolves this issue.

104. The price table contains a typo. Please correct the section number for the Alternative Integrated Cooling Solution to be 12.5.12.

Revision 4 of the Price Schedule spreadsheet resolves this issue.

105. Will you grant a two-week extension to the proposal due date of May 24, 2018.

No.

Added May 9, 2018

106. Has the anticipated budget range for the CORAL-2 systems changed?

No. The Funding and Proposed Financing section of the cover letter includes the following: The RFP Cover Letter includes the following: "The anticipated budget range for each system, including any associated NRE, is \$400M to \$600M. The budget range for NRE alone is expected to be \$75M to \$150M. However, the actual NRE or system award amount may be more or less than the anticipated budget, depending on the Laboratories' perceived value of the proposal(s), resulting negotiations, and annual appropriated funding from Congress."

Added May 10, 2018

107. Can you post a draft of the advance patent waiver.

The advance patent waiver for the CORAL-2 RFP is in process at DOE. DOE has advised that the CORAL-2 advance patent waiver provisions will be the same as there were in the initial CORAL acquisition issued by LLNL for the CORAL Procurement effort in 2014. A copy of the CORAL Advance Patent Waiver is now provided.

108. In the Summary Matrices spreadsheet, the Socket and Node tabs include rows for Storage Memory, if included. What does Durability (in P/E cycles) mean?

This assumes a NAND Flash device or storage memory with similar properties. We believe Program/Erase (P/E) cycles for the NAND blocks is a more meaningful statistic than Drive-Writes/Day (DWD), which is what flash drive vendors typically report. Vendor DWD calculations typically include over-inflated write amplification factors reflecting usage patterns that do necessarily not match our workloads. Providing us with total raw P/E cycles (P/E cycles/block times the number of blocks) and the block size will enable us to better assess the device durability over the system lifetime by applying our own anticipated write amplification factors.

109. Reference: QMCPACK

We believe there is a possible source of misinterpretation in the SOW and we would like to confirm with you if we are understanding it correctly.

Regarding the following paragraph of section 4.4.1:

"The goal of scalable science is to push the boundaries of computing by demonstrating the calculation of problems that could not previously be performed. As such, the preference for the Scalable Science Benchmarks is that the Offeror achieve the designed increase in FOM by demonstrating the ability to run problems that are at least 5-10x larger than the baseline problems. The individual descriptions of each of the Scalable Science Benchmarks found on the benchmark website provide information regarding how this increase in problem size can be performed"

From this paragraph, we understand that we are allowed to select problem sizes as we wish, provided that they are at least 5 to 10x larger than the reference problem. In other words, any problem size larger than 5X is allowed, and the individual descriptions of the 4 problems should explain how to increase the problem size. However, the summary for QMCPACK mentions the problem size to be 5X the reference problem:

"The figure of merit is based on the time taken to obtain $819200000=5*163840000$ total samples (steps)"

We would like to validate that such text is just an example, and a larger problem can also be used for QMCPACK (larger than 5x) in order to achieve the designed increase in FOM.

The short answer is no, not until the Offeror has completed the required projections.

The text in the SOW is general to cover all scalable science (SS) benchmarks, but each SS benchmark may have a particular target problem size as defined in the summary file. All the scalable science targets fall within the 5-10x larger problem range, but the specific problem size selected for a given SS benchmark was designed to meet DOE's scientific and national security workloads in the CORAL-2 timeframe. If the SS summary file mentions a specific benchmark problem size then the Offeror is required to provide projections for that problem size. For example, if a particular SS benchmark sets the problem size at 8x (run 6.25x faster to achieve 50x) then the Offeror is required to provide projections for that problem size. So, the above mentioned QMCPACK problem size must be projected to the scale in the summary file, 5x larger, and not larger or smaller.

After completing the REQUIRED problem size projections for all baseline **and** optimized results, the Offeror may provide additional problem sizes that are within the 5-10x larger range and that demonstrate the performance efficiency of the proposed architecture.

110. What number should be used to report the Quicksilver FOM?

The number in the version 17 of spreadsheet posted on the RFP site is correct. It can be found on the last page of the baseline FOM tab. The additional values in the document about Quicksilver: https://asc.llnl.gov/coral-2-benchmarks/downloads/Quicksilver_CORAL2-V3.pdf are for reference purposes.

111. PEPPI requirements that "Each section should start on a new page."

Since Sections 1 and 2 easily fit on a single page, may offerors be allowed to remove the page break between Sections 1 and 2 to save a page for response content?

Yes

112. RFP Letter, Proposal Due Date

Due to the extensive level of detail that the Pricing Table requires and the large volume of contractual documents to be reviewed in depth, this Offeror requests that an extension of two weeks (to June 7 2018) be granted in order to provide a complete response that allows for the most competitively priced solution.

No. We do not intend to grant an extension.

112. PEPPI Proposal Format

If a separate proposal is being submitted for each site, Volume 1 responses are limited to 200-pages. When the requirements are shrunk down to a small font and included within the response document, as required by the RFP, roughly 100 pages of the volume are already consumed by the requirements themselves. This leaves an Offeror with only 100 pages to explain their vision and commitments for a future exascale solution. This Offeror does not feel that 100 pages of available writing is sufficient to provide a complete response. We are requesting that the page count for separate responses be increased by an additional 100 pages to a maximum of 300 pages per site.

No. The draft SOW is 109 pages total. The page count starts with Section 1 and does not include the glossary or the appendices. The draft in the original font has 94 pages. The Offeror may reduce the SOW text to 6 pt, which reduces the page count to 55 starting with Section 1 and excluding the appendices.

113. PEPPI Section 2.2 and Section 8.

- Section 2.2 of the PEPPI states that for every benchmark, the Offeror should describe all modifications to source code, makefile or scripts written to run the benchmarks and why the modifications were made and to explain the methodology used to obtain the projected results. This data is to be included in Volume 1.
- Section 8 of the PEPPI (on page 18 and 19) requests the same information to be included in Volume 7.
- In order to fully comply with the detail necessary to accurately explain the modifications and methodologies used, we will need to utilize a minimum of 30 pages, which will negatively impact the detail we can provide on the rest of our technical solution. In order to maximize the technical details of our solution, this vendor requests that the above requirements only be required to be addressed in the Performance of the System (Volume 7).

The Offeror can mention it in Vol. 1, but give the details in Vol. 7. It's not our intention to negatively impact their page limit on the main proposal.

114. May the Offeror include NRE proposals which do not directly impact 50X and thus are not assumed in the build proposal but may be of interest to the labs? If so, how should these be presented so they are not included in the build proposal price?

Yes. PEPPI section 5.3 states, "The Offeror may include additional options that it thinks would be of interest to the Laboratories. Offeror-defined options must include relevant technical, business, and price information in the appropriate proposal volume." The Offeror should include a separate tab for the option in the Price Schedule as well as describe the option in an appropriate section in the draft SOW or NRE volumes. PEPPI section 4.2 states, "The Laboratories do not anticipate delivery of hardware or software resulting from NRE activities." If the option includes delivery of hardware and/or software, it should not be in the NRE volume.

115. If there is no Offeror IP to be protected as part of an NRE project, can that specific project be submitted for 100% funding? This could apply to open source software that needs to be developed to provide needed functionality.

No, see question 74.

Added May 17, 2018

116. Can we get clarification on the following requirement from the "Proposal Evaluation Instructions":

8.1 SECTION 1: BENCHMARKS, MAKEFILES, SCRIPTS, AND OUTPUT FILES

The Offeror may return all benchmark source files, makefiles, modifications, scripts written to run the benchmarks, and actual output files. The output of each code build, each run reported, and all run scripts used must be provided in electronic form, organized in a manner that reflects a one-to-one correspondence with the benchmark results spreadsheet.

Please provide clarification on the following:

In reading this requirement, it is understood that the benchmark output **must be provided**. Can you provide an FTP site for the results?

The first sentence states "The Offeror **may** return all benchmark source files, makefiles,".

If the Offeror chooses to provide these benchmark files, the Offeror should use ORNL's File Upload System:

<https://ftp.ornl.gov/fileupload.html>

When using the ORNL File Upload System, Offerors will be asked to provide an email address of the ORNL recipient.

Please use coral2-team@ornl.gov

Please contact me if there are any questions.

Sincerely,

William (Willy) Besancenez
UT-Battelle, LLC
Procurement Officer