



Acquisition Services Management (ASM) Division
Subcontracts, ASM-SUB
P.O. Box 1663, Mail Stop D447
Los Alamos, New Mexico 87545
505-665-3814 / Fax 505-665-9022
E-mail: dknox@lanl.gov

DATE: June 20, 2013

**Subject: Question and Answer Set 2
Trinity and NERSC-8 Computing Platforms Project
LA-UR-13-24524**

Greetings:

Interested parties are advised of the following questions or concerns that have been submitted to the Trinity and NERSC-8 Project team and to the accompanying Project responses below:

Question/Issue 1

Requirement 3.4.4 specifies that the Offeror provide the time to perform the metadata operations "enumerate and retrieve" on one million objects. However, the "Required Runs" section of the mdtest web page, <http://www.nersc.gov/systems/trinity-nersc-8-rfp/draft-nersc-8-trinity-benchmarks/mdtest/> specifies that only the times to perform creation and removal of objects be provided. Should we abide by the web page?

Project Response 1

The mdtest benchmark is not to be used for the "enumerate and retrieve" portion of 3.4.4. We anticipate using another method and/or benchmark to test this at acceptance. You do NOT need to address this aspect of the requirement at proposal time.

Question/Issue 2

The "Required Runs" section of the mdtest benchmark web page describes the syntax for an mdtest run that will cause N processes to operate on a single file:

```
aprun -n <#pes> N <#pes-per-node> ./mdtest -S -C -T -r -n 1 -d <path-to-pfs>/<n1_shared-dir>/ -F
```

However, when running with this command line, it appears to operate on 1 file per process, not a single file:

```
mdtest-1.8.3 was launched with 256 total task(s) on 8 nodes Command line used: ./mdtest -S -C -T -r -n 1 -d /lus/scratch/u3186/mdtest/mdtest-1.8.4/4121812.sdb/test4 -F  
Path: /lus/nid00030/u3186/mdtest/mdtest-1.8.4/4121812.sdb  
FS: 25.8 TiB Used FS: 20.8% Inodes: 87.2 Mi Used Inodes: 2.4%
```

256 tasks, 256 files

Project Response 2

Mdtest does report writing to N files on stdout, but if you check the actual directory for the resultant file you should find a single file. We confirmed this on our Cielo test bed.

Question/Issue 3

Note that the mdtest.c file in the 1.8.4 tar file contains the string:

```
#define RELEASE_VERS "1.8.3"
```

Can we get confirmation that the mdtest.c in the tar file is the correct version?

Project Response 3

The version of mdtest is 1.8.4, the release string was not updated. We'll correct that in the next distribution.

Question/Issue 4

Is the target SSP increase over Hopper (10x-30x for NERSC-8, 20x-60x for Trinity) using the optimized MPI+X SSP or base case (MPI-only) SSP?

Project Response 4

The Optimized MPI+X Case

Question/Issue 5

Is there an example calculation for the Capability Improvement metric?

Project Response 5

An example calculation for the Capability Improvement metric has been provided for the 3 NERSC-8 applications on the benchmarking web page.

Question/Issue 6

We are thinking of proposing a system with 2 different node types. In our response do we need to include benchmark times and an SSP calculation for both node types? How is the SSP calculated for a system with more than one type of node?

Project Response 6

Yes, your response should include benchmark times and an SSP calculation for both types of nodes. We have provided an example SSP calculation for a system with two different kinds of nodes on the benchmarking web page. This example shows that the SSP is calculated using all benchmarks and across all node types. The resulting SSP is the sum of the SSPs for each node type.

Question/Issue 7

The RFP Technical Specifications specifies some benchmarks for "RFP Response" and others for "Acceptance". What does this mean?

Project Response 7

Results of benchmarks labeled "RFP Response" should be submitted as part of the RFP response. Results of benchmarks labeled "Acceptance" must be provided by the selected vendor at the start of negotiations for inclusion in the Statement of Work.

Question/Issue 8

In reference to **Question and Answer Set 1, Project Response 12**, it appears that Moab is mandatory to be used as the only scheduler for all Trinity systems. Some of the more advanced capabilities from the RFP such as controlling the power use of the cluster, burst buffer operation, or system scale rely on functionality of the scheduler.

1. Please confirm if Moab is required as the scheduler for the clusters or can an alternate be proposed.

2. If Moab is mandatory, how will enhancements for the complex scheduling requirements needed for advanced power management and burst buffer data management be implemented?

Project Response 8

Since the requirement for Moab is a Target Requirement, it is not mandatory that Moab be the scheduler. Per the definition provided for a Target Requirement, failure to meet a Target Design requirement does NOT make the proposal non-responsive. However, if a requirement cannot be met the Offeror should provide the technical rationale and it is desirable that they propose an alternative solution.

Question/Issue 9

Please clarify in-situ visualization. What is the role of the visualization nodes with in-situ visualization? For example, are they running the simulation as well as the visualization/analysis? What is meant by "requires use of the main compute resources"?

Project Response 9

In-situ visualization and/or data analysis can take different forms, but in general, it requires that at least 2 jobs running on the system are able to communicate with each other over the high speed interconnect. It could be part of the visualization analysis is being done on the compute nodes and some reduced form of the data is being sent to the visualization nodes. It may also be the simulation and all of the in-situ visualization analysis is being done on the compute nodes, with viz and analysis dumps written to the file system as needed.

Question/Issue 10

How is does Section of the Draft Technical Requirements 4.1 workload #4 "Analysis of large ensembles of data" different from #1 and #3? Is this analysis that does not include visualization?

Project Response 10

To clarify, 4.1 #1, #3 and #4 could be non-viz data analysis. Number 4 is a special case of 1 and 3, called out because of the parallelism involved, and also because ensembles are a Sandia use case for simulation.

Question/Issue 11

In Section 4.1, what are the networking requirements for remote display? What bandwidth and latency is sufficient? What access is needed from the remote display, for example, do the visualization nodes need to have IP addresses on the same network as the remote displays? Figure A1 in Appendix A shows a 10 GB/s connection between the Visualization partition and DISCOM (Distance Computing WAN) at LANL. Does that imply that the visualization nodes share a 10 GB/s connection to the network outside the cluster?

Project Response 11

Yes, the 10 GB/s connection shown in Figure A1 is for the visualization nodes. We have not specified a latency requirement because it is dependent on the LANL external network configuration. The visualization nodes do need to be able to access an IP address off of the platform.

Question/Issue 12

Does the customer have information about the roadmaps of these visualization packages that they can share? Does the statement that ACES and NERSC will provide porting support indicate that there is ACES and NERSC will provide support to adapt these packages to the upcoming technologies?

Project Response 12

The ports are expected to be done by the respective visualization software vendor. We need to know that these packages will work on the given architecture. We expect the System vendor to provide porting

assistance to the software vendors. We also expect that the System vendor will have discussed these issues with the software vendors and will be able to provide some level of assurance that the software packages will work on the proposed architecture.

We recommend the following contacts:

ParaView: Berk Geveci, berk.geveci@kitware.com
VisIt: Dave Pugmire (ORNL), pugmire@ornl.gov
EnSight: Anders Grimsrud (CEI), ang@ensight.com

Question/Issue 13

If the visualization nodes are the same architecture as the compute nodes then this is straightforward. If they are different, then how do we figure 5%?

Project Response 13

If the visualization nodes are of a different architecture, the baseline proposal should be sized for 5% of the main compute partitions aggregate memory. E.g. if the aggregate memory of the main compute partition is 2 PB, the visualization partition shall provide 0.1 PB. In the requirements, we also ask for the option of doubling the memory capacity, this would also apply to the visualization partition.

Question/Issue 14

In Section 4.1, #1, a possible interpretation is GPUs are not essential for visualization nodes.

Project Response 14

GPUs can help rendering, but the visualization and data analysis needs are much greater than just rendering. These nodes require high performance, general-purpose capabilities.

Question/Issue 15

One of the requirements in the Draft RFP that has potential significant impact on system design (HW and SW) is the Max Power Rate of Change:

"The hourly average in platform power should not exceed the 2MW wide power band negotiated at least 2 hours in advance."

Would it be possible for you to elaborate on this requirement? How do you envision the negotiated parameter to be presented to the system? How often will the negotiated power band be changed? Is the hourly average a *rolling* hour or delineated by wall clock hours? Is the system really permitted to reach the peak power consumption (15 MW?) within the hourly window if the power band is significantly lower than that?

Project Response 15

Please recognize that advanced power management is one of the areas targeted for further development for DoE's Advanced Technology systems.

Industrial electricity isn't priced by kW*hr. Other components of the pricing formula include demand charge (basically, maximum power demand measured by some contractually specified formula), power factor, and other variables designed to estimate how hard is it for the utility to deliver the electricity. These factors are specified in the rate charts for industrial electricity and usually approved by relevant regulatory commissions. For example in NM, it may be that variable electricity costs about 65% more to deliver than steady electricity (a rough estimate based on a single data point for NM and not by a proper study of national averages). There are power fluctuation constraints arising from data center cooling infrastructure as well (e.g. sequencing of chillers/pumps/cooling towers).

More specific to the situation at Los Alamos is the pricing formula we currently have. We are supposed to stay within the power band negotiated in advance. Details of this may change when power contracts change, but basically they are designed to capture the power grid's costs of producing and delivering power. Currently, demand charge is computed based on hourly averages, i.e. not rolling averages. Those hourly averages should stay within the agreed upon power band, or else these things happen:

* Above the power band, must pay the electrical utility provider emergency power rates, or buy power on the spot market (either way, power is much more expensive)

* Below the power band, must pay for unused power or sell it on the spot market (thus dramatically reducing the cost benefit of energy savings)

* If peak hourly average power occurs in the wrong hour of the month, that extra MW could cost over 300 times more than normal; that normally means that mid-day peak hourly power should be watched carefully

We are looking to minimize operating **costs**, so power variability has to be considered. We hope that there is an opportunity to reduce the operating cost impact of power variability through advance warnings, which would allow site electrical utility people to trade energy on the spot market and thus avoid extra costs of emergency power rates or unused power. This normally implies at least an hour (preferably two) advance notice of major changes in power demand. The specific formula we have now has hourly averages, but this may change depending on terms in future power contracts. We'll be looking for advances in power management which can make large platforms behave as a good citizen on the power grid, where "good" is defined by cost incentives pushed to us by the power companies through rate structure we've got to pay. Our expectation is that for all data centers, power companies will find ways to pass on the cost of delivering variable power; that those costs may add about 65% to the cost of steady power (per the NM example rate formula), and that any energy saving technique which converts steady power to visibly variable power (meaning larger than something the power company can ignore) may have to achieve more than 65% energy efficiency improvement in order to be cost justified.

Grossly simplified, just think of variable power costing substantially more than steady power; even on hour-by-hour average basis.

Question/Issue 16

In the Draft benchmark run rules posted at <http://www.nersc.gov/assets/Trinity--NERSC-8-RFP/Documents/N8BmkInstructJune13.pdf>, 5 of the benchmarks have an extra-large dataset defined. Can you provide scaling data from the reference platform, Hopper, for these data sets?

Project Response 16

We are not providing scaling data for the applications specified in Table 2 of the document, only the reference values used for the SSP calculation.

Darren Knox



Acquisition Services Management
Los Alamos National Security, LLC
Los Alamos National Laboratory