



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 7, 2013

**Subject: Question and Answer Set 1
Trinity and NERSC-8 Computing Platforms Project
LA-UR-13-24133**

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

Respondents to the RFI have expressed a concern regarding the sharing of proprietary and/or intellectual property (IP) information with Offerors of fully integrated systems.

Project Response 1

RFP language will require that each proposal be an offer for complete system integration for two separate systems, Trinity and NERSC-8. RFP language will allow the direct submission to the LANS Procurement Specialist of proprietary and/or IP information that is a supplement to an offer for complete system integration. The Instructions to Offerors will provide guidance for the submittal of such supplemental proprietary and/or IP information. Each Offeror for full system integration will be responsible for timely submission of the supplemental information and for each supplement's clear reference to the offer that is being supplemented. Submissions of supplemental information must comply with proposal preparation instructions (i.e. format and page limitation requirements) and LANS will assume no responsibility/liability for any failure to comply with the instructions.

Question/Issue 2

Processor technology is controlled by the processor suppliers. If projected performance isn't realized, how is the system vendor protected?

Project Response 2

Subcontracts will be executed with the system vendor, presumably with the processor technology suppliers as lowertier suppliers to that system vendor. It is the responsibility of the system vendor to determine reliable sustained system performance and to propose accordingly. Any failure to meet proposed performance by the successful offeror is subject to usual contractual provisions, which will be stated in the RFP.

Question/Issue 3

Please provide more information on NRE funding availability for Trinity. Primarily provide how much NRE funding will be available and what rules may apply for those seeking it.

Project Response 3

We are not specifying how NRE will be funded or how much NRE will be available. It depends on the proposals and their value add to the projects.

Question/Issue 4

Is there any information available on what the acceptance process would entail?

Project Response 4

Information on the acceptance process has been posted with the DRAFT Technical Requirements.

Question/Issue 5

Extrapolating benchmark numbers from a smaller system to a system of this magnitude has many unknowns. Also, the technology is some number of months out, further complicating the estimates.

Project Response 5

Providing benchmark estimates has always been a challenge. However, we have used this process for most of the past ASC systems in the last 15 years. We have asked for risk mitigation for specific components such as processors in case the estimated performance doesn't measure up.

Question/Issue 6

The recent benchmark instructions and application README's state that the 'Base case' shall be run with "no OpenMP" threads or "any existing APIs in the codes which exploit additional parallelism, e.g. OpenMP, may not be enabled". However, I have not been able to build the provided SNAP, or MILC codes without using the OpenMP enabling compiler flag. I can provide details of the code why this is the case, but just to be safe should the documents say something like "if OpenMP is enabled only 1 thread may be specified"?

Project Response 6

Yes, it is acceptable to run with the OpenMP enabled version and use OMP_NUM_THREADS=1 for the MPIonly base case.

Question/Issue 7

SNAP

In the May06 update to the README and web page there is a section that states : "Required Runs We require the weak scaling results, where the problem size per processor is constant regardless of the number of processors(sockets) with 1MPI rank per core, no OpenMP threads. Focus on 16x16x16 per processor on hopper. In *addition* to the above you can adjust the # MPI ranks and OpenMP threads that give the best performance for your architecture and report the configuration and results."

Does the customer expect the offerer to provide weakly scaling results i.e. modifying ny and nz proportional to npey and npez? Can the customer clarify?

Project Response 7

The description provided on the SNAP benchmarking web page describes how the baseline results were collected on NERSC's Hopper system, and how to modify parameters to best fit the vendors architecture. The vendor is allowed to run the required problem sizes at a different number of MPI ranks while preserving the total number of cells. The run rules for SNAP will be updated to provide more clarity.

Question/Issue 8

MILC

On the web site and in the documentation the problem size is described as follows. "The single node case is designed to have 8x8x8x8 sites per MPI task and is sized to fit on a current node of NERSC's hopper system (24 core/node Cray XE6). The large test case has the same number of sites per core but is sized to fit on 1024 nodes of hopper." Suggesting that the inputs should be scaled weakly based on a 8x8x8x8 site lattice per MPI task. This is consistent with all the sample input files and output files. In the "Capability Improvement Runs" section it says that "For the 24,576MPI rank large problem, the size of the fourdimensional spacetime lattice is controlled by the following parameters in the input deck:

```
nx 64
ny 64
nz 128
nt 192"
```

...

In general, to weak scale a 4x4x4x4 (nx x ny x ny x nt) problem, one begins by multiplying nt by 2.. " Suggesting that a 4x4x4x4 base lattice size should be used, but the parameters above is consistent with the 8x8x8x8 size lattice per MPI task. There seems to be some inconsistencies between the different statements and it would help if they could be reconciled or clarified.

Project Response 8

The 4x4x4x4 lattice size in the Capability Improvement description was meant to be an example. For the Capability Improvement metric, it is not a requirement to weak scale the "large" problem defined for the SSP metric. The vendor is allowed to choose the parameters that best fits their architecture. The run rules for MILC will be updated to provide more clarity.

Question/Issue 9

UMT (The following is also relevant to AMG)

With the instructions in the new README file we have been able to run UMT with different MPI rank counts while keeping the problem size constant. As for the variation in the iteration counts, we have recently discovered the same phenomenon is at work for AMG. The number of iterations changes with both MPI count and with the compiler used, with different compilers having the lowest iteration count at different processor counts. However, it does appear that the time per iteration is rather constant. It seems that it is important for the purposes of this benchmark and specifically the SSP calculation that the number of iterations be a constant.

The current SSP calculation does not take into account a change in convergence behavior due to a change in problem size; it is based on a reference time and a reference, fixed FLOP count. So while we agree it is possible and very useful for end users to examine the time per iteration, that is currently not part of the SSP calculation. Since we don't actually know how many iterations were completed during the reference runs, it is not clear to us how to project to a final time, which is dependent on the number of iterations. Furthermore it is not clear to us that we can guarantee that we will be executing the same number of iterations on a future machine with different processors and compilers.

Perhaps an alternative would be to base the SSP calculation off of a Tflops/iteration and a time/iteration. That way the program can be allowed to converge but timings and SSP calculations can account for variations in the number of iterations. Please advise on how we should handle this situation.

Project Response 9

The Benchmarking Run Rules document and README files have been updated to account for the differences in the number of iterations to convergence. Make sure to provide the number of iterations along with the benchmark time.

Question/Issue 10

We have observed that in the subroutine src/Teton/transport/Teton/snac/snflwxyz.F90 arrays are allocated by each MPI rank to be used in subsequent calls to calculate fluxes. Most of the computational work is in this section. Below are the allocate statements we are referring to:

```
allocate( omega(ndim,n_cpuL) )
allocate( abdym(nbelem,n_cpuL) )
allocate( siginv(npart,ncornr) )
allocate( sigvol(npart,ncornr) )
allocate( tphic(npart,ncornr,n_cpuL) )
allocate( tpsic(npart,ncornr,n_cpuL) )
allocate( qc(npart,ncornr,n_cpuL) )
allocate( psifp(npart,maxcf,ncornr,n_cpuL) )
allocate( source(npart,ncornr,n_cpuL) )
```

By our calculation, these allocates would total about 20 Mbytes of data per MPI rank on average, assuming 1 OpenMP thread, and 36 Mbytes of data per rank assuming 2 OMP threads. We have two questions regarding the size of these arrays:

- a) Is there an input parameter than can control and change the size of ncornr while keeping the global problem size and science problem fixed? If one exists that would allow one to impact the total size of those allocates per rank.
- b) Conversely, do the size of ncornr and npart change depending on the science problem? If yes, are they typically larger or smaller than the benchmark problem?

Project Response 10

- a) No, ncornr is not a controllable parameter and is not to be changed.
- b) Ncornr and npart are independent of the size of the problem, i.e. number of zones.

Question/Issue 11

Our benchmarking center isn't equipped with systems that large. Can you provide options for benchmarks on smaller subsets?

Project Response 11

The draft technical requirements we posted in December stated that performance benchmarks, application and microbenchmarks, can be "actual, predicted and/or extrapolated". In addition, the benchmarking run rules posted on the web site provide further guidance under the "Submission Guidelines" section for performance projections (predicted and/or extrapolated). With a 2015 delivery, we expect performance projections are necessary as many key technologies are not available at this time.

Question/Issue 12

The 1st draft of the RFP Technical Specifications released in December 2012 stated the system shall provide resource management functionality including checkpoint/restart, Moab Compatible.

- a) Does this mean application level checkpoint restart, not system C/R support?
- b) What does it mean for Moab compatible? Does it mean Moab as scheduler, vendor provided resource management and job launching system? Or Moab only plays as meta scheduler to forward job to the vendor scheduler. Please clarify this.

Project Response 12

This requirement has been revised in the second draft of the RFP Technical Specifications release June 3, 2013 and the language for checkpoint/restart has been removed. For Moab capability, the

requirement is to provide Moab as the scheduler, with a vendor provided resource manager and job launching system.

Question/Issue 13

The draft RFP Technical Specifications state that a job interrupt shall not require a complete resource re-allocation. Does this mean if one of hosts fails during job run time, system will allocate a new resource to replace the failure one for a running job? Please help clarify this.

Project Response 13

The requirement intent is for the job to not have to resubmit to the scheduler, and hence wait in the queue, to allocate resources to continue.

Question/Issue 14

The draft RFP technical specifications state a complete system initialization shall take no more than 30 minutes. How will this measured and please indicate what constitutes the start and stop time?

Project Response 14

The requirement is to describe the system's "full system initialize sequence and timings." How it is measured is system dependent and the method will be determined in the statement of work.

Question/Issue 15

What Globus toolkit version do you plan to use? GT4 supports our LSF product. However, GT5 changed the interface and now it does not work with LSF. As GT is a community program, historically it is has been the community who does integration work. There is not guarantee the community will take such action.

Project Response 15

There is not a requirement for a specific version of the Globus toolkit. Please describe what version of the you will support.

Question/Issue 16

The draft RFP technical specifications state job launch in 30 seconds. How will this measured and please indicate what constitutes the start and stop time?

Project Response 16

See Table 2 in the RFP Technical Requirements.

Question/Issue 17

Question: How will I/O performance be measured?

Project Response 17

See Table 2 in the technical requirements.

Question/Issue 18

The 1st draft of the RFP Technical Specifications released in December 2012 stated JMTTI/DELTA > 30 and also stated that the JMTTI > 30, from which it follows that DELTA must be less than $30/30 = 1$ (hour). However, Table 4 gives values of 20 and 35 minutes. And the Burst Buffer Use Scenarios gave significantly different values as well. Please clarify the IO bandwidth performance requirement.

Project Response 18

The PFS needs to be sized based on the performance requirements specified in Table 4. Delta for the

PFS and Burst Buffer are separate requirements. This has been clarified in a the second draft of the RFP Technical Specifications released June 3, 2013.

Question/Issue 19

Please clarify the External and Local client requirements in Table 4. For example, are these to be provided via nfs? Are these external and local clients on known networks that we need to bridge to? Are you requesting Gateway nodes?

Project Response 19

The external bandwidth requirements have been further clarified in the second draft of the RFP Technical Specifications release June 3, 2013. In particular, a facilities interface diagram has been provided in the appendix.

Darren Knox



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