Communications regarding UCAR RFP000074 (NWSC-3)

Revision:
Version 1.2, 20 April 2020

Table of Contents

1 Overview 2
2 Conventions 2
2.1 Example brief description of question 2
3 RFP Questions and Answers, issued 13 April 2020 2
3.1 Attachment 1, Technical Specifications, Section 1, Software 2
3.2 Attachment 1, Technical Specifications, Section 3.3.4, Production PFS 2
3.3 Attachment 2, Benchmark Rules, Sections 5.1.3 and 5.1.4, and Benchmark Website Instructions 3
3.4 Attachment 2, Benchmark Rules, Section 5.1.1, and Benchmark Website Instructions 4
3.5 Attachment 2, Benchmark Rules, Section 5.3 4
4 RFP Questions and Answers, issued 20 April 2020 5
4.1 Attachment 2, Benchmark Rules, Section 4.4 5
4.2 Attachment 1, Technical Specifications, Section 3.4.3, Production PFS 5
4.3 Attachment 1, Technical Specifications, Section 3.12.2, Facilities & Site Integration 6
4.4 Attachment 1, Technical Specifications, Section 3.12.7, Facilities & Site Integration 6
4.5 Attachment 2A, NWSC-3 Benchmark Results Spreadsheet 6
4.6 Attachment 2A, NWSC-3 Benchmark Results Spreadsheet 6
4.7 Attachment 2A, Benchmark Results Spreadsheet 7
4.8 Attachment 1, Technical Specifications, Section 3.13.2 and 3.13.5, Maintenance, Support, and Technical Services 7
4.9 Attachment 1, Technical Specifications, Section 3.3.6, Production PFS 7

1 Overview

This document contains prospective Offeror questions related to UCAR RFP000074 (NWSC-3) and UCAR's responses to those questions.

2 Conventions

Each question and its corresponding response is formatted as shown below, providing a unique question identifier and a brief title for the question, the question itself, and UCAR's response to the question.

Example:

2.1 Example brief description of question

Question — The text of the Respondent's question will appear here. It may be stated verbatim or modified slightly to remove any irrelevant attributes of the question or any indication of the Offeror's identity.

UCAR's response to the question immediately follows.

3 RFP Questions and Answers, issued 13 April 2020

The following questions were received by UCAR between the release of the RFP, on 02 April 2020, and 13 April 2020.

3.1 Attachment 1, Technical Specifications, Section 1, Software

Question — Prior to submitting our "Registration of Interest," we are seeking confirmation on the response requirement. Will NCAR accept a proposal for a software portion only, or does the response need to include all components, i.e. software, hardware, and storage, to be accepted?

An Offeror proposal in response to UCAR RFP000074 must include a complete NWSC-3 solution, comprising all hardware, software, infrastructure, networking, delivery, installation, and five (5) years of software licenses and hardware/software maintenance, support, and other services. An exception, as described in §2 of Attachment 1 of the RFP, is provided for an Offeror who chooses to propose only an HPC or PFS solution. If an Offeror wishes to submit a quotation for a specific hardware or software component of NWSC-3, the Offeror may do so, but it will not be considered a response to UCAR RFP000074.

3.2 Attachment 1, Technical Specifications, Section 3.3.4, Production PFS

Question — As stated in Section 3.3.4, "*The PFS solution shall have an initial usable file system capacity of 60 PB (petabytes) and a rack infrastructure that allows the usable capacity to be doubled by the simple addition of data storage devices.*" Does this mean it is required that all of the needed additional infrastructure,

such as drive enclosures, controllers, cables, racks, and power be in place at the initial installation, so that doubling the capacity is done by merely adding HDDs (and SSDs as specified in 3.3.5)?

UCAR's requirement stipulates that the proposed solution has the ability to increase capacity simply by adding additional HDD/SSD drives. The Offeror's proposed solution should include all of the needed storage infrastructure components, such as drive enclosures, controllers, cables, and rack power in place at the initial installation. If the architecture allows for additional drive enclosures and cabling to easily be added within the rack/controller infrastructure, that is an acceptable alternative, as long as it can be done in a manner that is non-disruptive to the services provided by the initially installed storage.

3.3 Attachment 2, Benchmark Rules, Sections 5.1.3 and 5.1.4, and Benchmark Website Instructions

Question

For the CESM2_MG2 kernel benchmark, the last sentence of the first paragraph of page 2 of the PDF containing instructions on the benchmarks website requests: *"Please provide output files for a number of MPI ranks that both fully-subscribed and over-subscribed hardware cores,"* but it is stated on page 10 of the UCAR_RFP000074_Attachment_2_NWSC-3_Benchmark_Rules_v1.docx in Section 5.1.3 MG2 that *"MG2 should be run on a single node, using all available cores, and using one MPI rank for each of the available cores."*

Analogous to CESM2_MG2, for the WACCM_imp_sol_vector kernel benchmark, the last sentence of the second paragraph on page 2 requests: *"Please provide output files for a number of MPI ranks that both fully-subscribed and over-subscribed hardware cores,"* but it is stated on page 10 of the UCAR_RFP000074_Attachment_2_NWSC-3_Benchmark_Rules_v1.docx in Section 5.1.4 WACCM that *"WACCM should be run on a single node, using all available cores, and using one MPI rank for each of the available cores."*

Do the benchmark rules override the PDF so that oversubscribed runs are no longer required? Conversely, if oversubscribed runs are still required or desired, then which achieved figure of merit (FOM) must be entered into the UCAR_RFP000074_Attachment_2A_Benchmark_Results_Spreadsheet_v1.xlsx; i.e., the best FOM or always the FOM from the fully subscribed (but not over-subscribed) run, even if the oversubscribed run yielded a higher FOM?

UCAR would like the benchmark results to be returned for both the fully subscribed and oversubscribed cases, as requested in the instructions provided on the NCAR HPC Benchmarks website[1]. The result for the fully subscribed case (i.e., one MPI rank for each available core) should be used as the figure of merit (FOM) to enter in the Benchmark Results spreadsheet[2].

3.4 Attachment 2, Benchmark Rules, Section 5.1.1, and Benchmark Website Instructions

Question

Based on the following language found in Section 5.1.1 of Attachment 2: *"5.1.1 CLUBB: 'CLUBB should be run on a single node, using all available cores, and using one MPI rank for each of the available cores,'"* the results for this benchmark will be for runs on a node of the proposed system which is fully subscribed with MPI tasks but NOT oversubscribed (that is, with hyper-threads) as requested in previous documentation, correct? The CLUBB benchmark data only provide reference files for pcols=16 and pcols=192. The README and PDF state that results for any value between 16 and 192 would be accepted. Without the reference files, there is no way to verify the results of a different value of pcols between 16 and 192. Is it correct then to assume we can only test with pcols=16 and pcols=192 for CLUBB?

For CLUBB, the fully subscribed result (one MPI rank per core) is required to be returned and should be entered into the Benchmark Results spreadsheet[2] as the figure of merit (FOM). An oversubscribed result may optionally be returned, in addition to the fully subscribed result, if it showcases interesting performance.

The CLUBB benchmark is used outside of the NWSC-3 benchmark suite with other values for pcols, hence the language in the README and instructions. However, for the NWSC-3 procurement, you are correct: we are only requesting results for pcols=16 and/or pcols=192. For CLUBB, the fully subscribed result (one MPI rank per core) is required to be returned and should be entered into the Benchmark Results spreadsheet[2] as the FOM. An oversubscribed result may optionally be returned, in addition to the fully subscribed result, if it showcases interesting performance.

3.5 Attachment 2, Benchmark Rules, Section 5.3

Question

The benchmark rules document mentions two Microbenchmarks, STREAM and OSU MPI, that vendors need to complete as part of the RFP requirements. However, the results spreadsheet supplied doesn't have provision to include results from these two micro benchmarks. Please clarify.

The primary purpose of the Benchmark Results spreadsheet[2] is to calculate the aggregate Cheyenne Sustained Equivalent Performance (CSEP) value. Since CSEP is intended to be a comparative measure of a system's capacity based upon the relative performance of NCAR applications, the synthetic STREAM and MPI benchmark results are not expected to be entered into the spreadsheet. Nevertheless, the STREAM and MPI benchmark results are important to UCAR's assessment; thus, they should be returned as files capturing STDERR and STDOUT. The STREAM and MPI benchmarks are required to be run, and their results are required to be returned with the Offeror's proposal.

4 RFP Questions and Answers, issued 20 April 2020

The following questions were received by UCAR between the release of Version 1.1 of this document, on 13 April 2020, and 20 April 2020.

4.1 Attachment 2, Benchmark Rules, Section 4.4

Question — Context: It is stated in UCAR_RFP000074_Attachment_2_NWSC-3_Benchmark_Rules_v1.docx in paragraph "*4.4 As-is and Optimized Benchmark Results*" for the As-is results at the top of page 7 that *"No application source code modifications are allowed."* Does this extend to/include also: a) No compiler directives for optimization purposes are allowed for the as-is runs? and b) No porting changes are allowed? E.g., we could write a C-Language wrapper for getpid, or otherwise, would compile with -D_NOGETPID.

a) For the "**as-is**" runs, additional compiler directives may not be added to the source code for purposes of improving performance. Directives that already exist in the source code may be used, e.g. by compiling with -qopenmp, etc.

b) For the "**as-is**" runs, only source code modifications that are required in order to make a code execute correctly and/or pass validation criteria are permissible. Any such changes should be placed inside of conditional compilation blocks such that the original source code can still be compiled. The blocks should clearly identify the vendor making the changes, for example:

```
#ifdef NWSC3_Offeror
<source code modifications>
#else

<original source code>
#endif
```

It should be noted, though, that the Benchmark Rules §4.4 does allow compiler directive and source code changes to be made and submitted as "**optimized**" results, so long as those changes adhere to Benchmark Rules §4.6, and the benchmark continues to pass its validation criteria.

4.2 Attachment 1, Technical Specifications, Section 3.4.3, Production PFS

Question — Section 3.4.3 states *"The NWSC-3 PFS solution shall support connectivity with NCAR client systems other than the NWSC-3 HPC system and provide an aggregate, sustainable bandwidth in excess of 200 Gb/s."* Does the 200 Gb/s in the requirement mean 200 Gigabits per second or 200 Gigabytes per second?

Section §3.4.3 of the Technical Specifications is correct. In addition to the bandwidth to the NWSC-3 HPC system, the NWSC-3 PFS must have, at a minimum, an additional 200

Gigabits per second (Gb/s) aggregate, sustainable bandwidth for connection to other NCAR client systems.

4.3 Attachment 1, Technical Specifications, Section 3.12.2, Facilities & Site Integration

Question Please clarify the statement *"Other power sources (208V, 110V) are available to support a system's infrastructure such as storage, switches, and consoles."* Is 3-phase 208 Vac available?

Yes, 3-phase 208V is available. However, UCAR wishes to reiterate, as stated in the preceding sentences of §3.12.2, that the high density compute nodes should be powered at 480V, so that the NWSC can maintain its electrical efficiencies.

4.4 Attachment 1, Technical Specifications, Section 3.12.7, Facilities & Site Integration

Question Please clarify the statement *"All cables shall be plenum rated…"* Is this just limited to the networking and communications cables? There are no plenum requirements in the National Electrical Code or ITE product safety standards for power-supply cords.

This is acknowledged and understood. The requirement is limited to network and system interconnect cabling.

4.5 Attachment 2A, NWSC-3 Benchmark Results Spreadsheet

Question For the heterogeneous node benchmarks, the comparison points for accelerator performance relative to Cheyenne cores are not consistent.

This observation is correct and the difference is intentional. The two heterogeneous node benchmarks are being compared to Cheyenne using different methods. The MPAS 15 km benchmark compares a fixed number of Cheyenne cores (or nodes) to a fixed number of proposed GPUs/Accelerator devices, without fixing the number of proposed nodes (i.e. the number of devices per proposed node is not specified by the benchmark rules). In contrast, the GOES benchmark compares a fixed number of Cheyenne nodes, to a fixed number or proposed nodes (one in both cases) again without specifying the number of proposed devices per node. Because of this difference in comparison methodology, the formulas in the benchmark results spreadsheet[2] use different normalizations to calculate speedups relative to Cheyenne.

4.6 Attachment 2A, NWSC-3 Benchmark Results Spreadsheet

Question For the GOES benchmark the comparison is one "heterogeneous node" vs. 36 cores of Cheyenne, while for MPAS-A at 15 km the comparison is "one accelerator" vs. 118.5 Cheyenne cores (2844/24 = 118.5). As a result, speed-ups in the spreadsheet come from ratios as diverse as a minimum of 4

accelerators vs. 1 Cheyenne node, to one accelerator vs. ~3.3 Cheyenne nodes.

This observation is correct and the difference is intentional. Please refer to §4.5, which also covers this question.

4.7 Attachment 2A, Benchmark Results Spreadsheet

Question For MPAS-A at 30 km there are two very different comparison points: one "heterogeneous node" vs. 36 cores of Cheyenne, and one "two accelerators" vs. 150 cores of Cheyenne. The RFP document requests heterogeneous nodes with four to eight accelerators, so the differences between these comparison methods is very large.

This observation is correct and the difference is intentional. Please refer to §4.5, which also covers this question. Similar to the response to §4.5, there are two comparison methods being employed—either Cheyenne nodes versus proposed nodes, or Cheyenne nodes versus proposed GPU/Accelerator devices, without specifying how many GPUs, or devices, should be within a proposed node. Again, the speedups are calculated differently depending on which comparison method is being used.

4.8 Attachment 1, Technical Specifications, Section 3.13.2 and 3.13.5, Maintenance, Support, and Technical Services

Question Please clarify the statements in Section 3.13.2 *"UCAR's target for on-site Offeror responsiveness is 9x5-NBD (Next Business Day)"* and in Section 3.13.5 *"The Offeror shall provide technical support services with 24x7 telephone and web-based technical support, problem reporting, ticketing, diagnosis and resolution services."*

Section §3.13.2 specifically relates to all Field-Replaceable Unit (FRU) work or any other work that implicitly requires the physical presence of an Offeror representative at the NWSC. This on-site work requires a responsiveness of 9x5-NBD (Next Business Day), with the caveat stated in §3.13.2, that "*…a more immediate response should be available for critical downtime situations*."

Section §3.13.5 is for any other support services and assistance that can be handled remotely, such as software support, problem reporting and escalation.

4.9 Attachment 1, Technical Specifications, Section 3.3.6, Production PFS

Question Does 3.3.6 require the 100/200 Gb Ethernet switch infrastructure to be provided by the HPC cluster, by the PFS, or part of the NWSC infrastructure?

An intent of the §3.3.6 specification for the PFS, and its counterpart §3.2.10 specification for the HPC system, is for the PFS and HPC systems to be independently operable, particularly if they might be supplied by independent Offerors. However, an Offeror may propose a complete solution with integrated PFS and HPC networking infrastructure.

Any NWSC-3 PFS solution provided must be able to integrate into the 100/200GbE HPC network and provide full, non-blocking communications to systems within the NWSC-3 HPC network. In such a case, the Offeror should provide all switches, cabling, optics, and/or gateways for connectivity with the NWSC-3 HPC network. Likewise, the Offeror may choose to rely on the HPC network infrastructure for PFS connectivity, providing all cabling and optics necessary for connection to the HPC network switches.

It should be noted that, per §3.4 of the Technical Specifications, the solution will also need to integrate the provided PFS and HPC networks into the NWSC's TCP/IP network. The vendor shall supply suitable cabling, optics, and/or gateways needed for connectivity with the NWSC TCP/IP network.

[1] NCAR HPC Benchmarks Website: https://www2.cisl.ucar.edu/hpc_benchmarking

[2] UCAR_RFP000074_Attachment_2A_Benchmark_Results_Spreadsheet_v1.xlsx