Communications regarding UCAR RFP000074 (NWSC-3)

Revision:
Version 1.5, 08 May 2020

Table of Contents

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.

5 RFP Questions and Answers, issued 27 April 2020

The following questions were received by UCAR between the release of Version 1.2 of this document, on 20 April 2020, and 27 April 2020.

5.1 Attachment 2A, Benchmark Results Spreadsheet

Question — The requirement to have an 80% - 20% split in the homogeneous vs. heterogeneous CSEP contribution is leading us to a very large heterogeneous node count and subsequently large total GPU count. Can NCAR verify the formulae used to compute speed-ups for the heterogeneous node benchmarks, or perhaps update the spreadsheet?

UCAR and NCAR subject matter experts have reviewed the formulae used for the 80/20 ratio, along with other formulae and metrics, in the NWSC-3 Benchmark Results Spreadsheet and have noted both errors and deficiencies in the CSEP calculations. UCAR apologizes for these problems in the original spreadsheet. UCAR is releasing, as part of Amendment #1 to UCAR RFP000074, a revised spreadsheet. Prospective Offerors should discard the originally-released Benchmark Results Spreadsheet and use the **'v1.1'** version, released with Amendment #1, for calculating CSEPs.

5.2 Follow-up Question to UCAR's Response to Question 4.9

Question — The statement in the response to question 4.9 in the latest Q&A says: "However, an Offeror may propose a complete solution with integrated PFS and HPC networking infrastructure." Does this mean that an HPC provider can provide an integrated PFS and not allow access from a third-party PFS?

The RFP's Attachment 1, NWSC-3 Technical Specifications, §3.2.10 and §3.3.6, describe how UCAR desires the HPC system and PFS to be interconnected. This design is meant to allow the NWSC-3 resources to fit into a broader HPC environment in which computational and storage resources (outside of those acquired via the NWSC-3 RFP) reside. In addition, the PFS solution must be capable of providing file system services to clusters other than the NWSC-3 HPC solution, and the HPC solution must be capable of mounting additional storage resources such as the current Campaign Storage solution.

While an HPC provider can propose an integrated PFS, UCAR also encourages standalone PFS-only proposals. UCAR discourages an HPC solution that disallows third-party PFS solutions as this will not meet the requirement of integration with current resources. The PFS solution must be capable of connecting to the NWSC 100/200 GbE (gigabits/second)

network and provide a minimum of 300GB/s (gigabytes/second) bandwidth between the PFS and the NWSC-3 HPC resource.

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

Question: Please clarify the statement "The systems shall be provided with cable containment integrated with and spanning between the system racks to accommodate the system interconnect and networking cables." We assume containment means cable tray. Does this say, 'NCAR shall provide'? Or, 'Vendor shall provide'? Does this also apply to Module B?

The Offeror is responsible for the cable tray and cable management between the HPC racks and cabinets for the high-speed interconnect and network cables to make the system functional. NCAR will provide the cable tray paths to connect the HPC system to other systems like the PFS and the main networking hubs within the NWSC computer rooms.

5.4 Benchmark Website Instructions

Question: The output of GOES shows all 3 numbers (elapsed, epoch, and elapsed/epoch), but which is the metric to use? I understand from the benchmark document for GOES (GOES16_2020_03_18.pdf) that it mentions the epoch time is the most parallel section of the application. I would like to confirm that “epoch” is the metric for our measurement that we should focus on. Is it possible to change the conv net parameters in the yml file? We would like to experiment with certain parameters such as the precision from the default float32 to bfloat16, batch_size to change from the default 1024. I am using our own GPU which has support for tensorflow and other ML libraries, but some of the software stack would be somewhat different from the ones listed in the build procedure. Hope it’s ok.

The ‘Epoch’ timer is the official benchmark metric and is the value that should be entered in the spreadsheet. The output file containing all timing values should be returned as part of the response.

The benchmark_config_default.yml file should not be altered for the "as-is" benchmark runs. The parameters in the file can be altered for use in optimized runs. The Offeror must document any changes made for the optimized runs, and the requested output files should be returned in addition to those for the "as-is" runs. UCAR has also provided updated benchmark instructions, and an updated benchmark spreadsheet, that will address this question as well.

5.5 Benchmark Website Instructions

Question: Changes were needed in order to get the benchmark to compile and run on a certain accelerator device. When PAPI is enabled, it appears to trigger some errors. If PAPI is not enabled, it complains about "current_timer" and

"check_flag" and there are minor errors when using PGI fortran-compiled mpif90 wrapper script.

These errors occur when attempting to build an older version of MPAS than is provided in the NWSC3_benchmarking branch of the restricted access MPAS repository.

There are many versions of the code on the MPAS repository, so it's important that you check out the NWSC3_Benchmarking branch. From the directory where you'd like to download the code, you can use the commands:

```
yourDirectory> git clone https://github.com/cenamiller/MPAS
yourDirectory> cd MPAS
yourDirectory/MPAS> git checkout NWSC3_Benchmarking
```

You can confirm the correct branch using the command:

```
yourDirectory/MPAS> git branch
```

and the result should have an asterisk next to the NWSC3_Benchmarking branch.

5.6 Benchmark Rules and Instructions

Question — Is it permissible, for the GOES benchmark, to increase the number of epochs to a larger number that better represents a full training, or could NCAR provide a formula that can convert the existing benchmark metrics to mimic a more realistic training (for example, 100 epochs)? If so, we would request a new baseline to compare against.

The Epoch timer, which is the timer to be recorded for the benchmark performance result, only measures the time of the 2nd epoch, and does not include the first epoch which contains setup overhead. The epoch timer is representative of the timing for subsequent training epochs, and so longer training times can be reliably estimated from it.

5.7 Benchmark Rules and Instructions

Question — For the GOES benchmark, on a CPU, epochs 1 and 2 take about the same time. On a GPU, the first epoch can be 10-20x longer than epoch 2. However, the 2nd epoch time (and all subsequent epochs) is very short (on the order of a few seconds). Most of the GPU runtime is set up during the first epoch. Only considering two epochs doesn't represent how a typical training (for example 100 epochs) would perform in order to achieve appropriate accuracy for the model.

Q1: The GOES PDF instructions list a set of libraries required for the goes16ci including Cuda, cuDNN, NCCL, and Python. Are the exact versions of these libraries required, or can later library versions be used?

Q2: In the Run Procedure section of the GOES PDF, there is a sentence, “We are interested in CPU node performance on your system, so please increase the number of “. What was the rest of the sentence supposed to say?

Q3: In the GOES benchmark, the code to find the cudnn and nccl libraries make an assumption on how the software is installed. May we make modifications to monitor.py to solve this issue without violating the “As-is” benchmark rules in section 4.4?

Regarding Q1: Later library versions may be used. Please document the version of the libraries used in the response.

Regarding Q2: The offending partial sentence has been removed. This is reflected in an updated benchmark instructions document for GOES in conjunction with Amendment #1.

Regarding Q3: Yes, the rules for the as-is benchmarks allow modifications necessary to make the benchmark run. Please document any modifications that were necessary in the response.

5.8 MPAS-A Benchmark Question

Question — For MPAS, is it permissible to increase the number of time steps that are used to compute the average performance per time step (for example, move the number of time steps from 100 to 1000 or 10,000 to compute a better average value)? If so we will NOT need a new baseline for performance/timestep.
Rationale: The first time step includes a significant amount of overhead for memory allocations.

The MPAS-A benchmark, both the 15 km and 30 km cases, should be simulating 200 time steps. We have realized that the 30 km case was mistakenly running for only 100 time steps. A new version of the MPAS-A input data (MPAS_2020-04-27_data.tar.gz) has been placed on the NWSC-3 Globus Benchmark endpoint to correct this error. We feel that 200 time steps is adequate for this test, and the number of time steps should not be changed. Additionally, please note that the value to be entered into the results spreadsheet should be entered in the "total" column for the "time integration" timer, not in the "avg" column, as may have been requested in an earlier version of the benchmark instructions.

5.9 NWSC Virtual Site Visit Question

Question — Will the slides/charts and other material be available after the meeting?

Yes. The slides and videos are available on the “Resources” tab of the NWSC-3 RFP Website at https://www2.cisl.ucar.edu/resources/nwsc-3. Offerors are advised to periodically check the “Updates” tab on the NWSC-3 RFP website for more details and regular updates regarding the NWSC-3 RFP.

5.10 NWSC Virtual Site Visit Question

Question — Regarding redundancy: With N+1, what is the typical N? N=1, N=2, etc.? Can you explain “dual tailed”?

At the server level, dual tailed means that there are two (2) physical power supplies and only one (1) is needed to operate the server. There are at least two (2) places a power cord plugs in and if one (1) power source is lost, the equipment still operates. A single device could have more than two (2) power supplies, but half of them need to be able to be powered off with no effect to the server. The same is true for an entire rack PDU; if a rack only needs one (1) PDU for all loads, there needs to be a redundant PDU installed in all critical equipment racks so that each PDU can be independently shut off and all equipment in the rack continues to operate with no issues. If racks need more than two (2) PDUs to power them, that is fine; but again, for critical equipment racks, there needs to be redundant PDUs so that the servers are not affected by a loss of a PDU or a loss of power to a PDU.

5.11 NWSC Virtual Site Visit Question

Question Will 415/240v power be available?

240v is more of a standard voltage for a single phase, or high leg three-phase system in the US. Normally anything that operates on 240v will also operate on 208v. While 415v seems to be more of a European voltage, this can be provided with the use of transformers. If 240v has to be provided, it can also be with the use of transformers. Please keep in mind that the NWSC operates at 60Hz; if any equipment operates at different frequencies (50Hz or other), this will need to be clearly stated in the Offeror's RFP proposal as this will require specialized electrical gear to change the frequencies to the proposed NWSC-3 systems.

5.12 MPAS-A Benchmark Question

Question For MPAS-A at 15 km resolution, the benchmark spreadsheet computes a speedup factor, column E, by comparing 24 accelerators with 2844 Cheyenne cores. In order to achieve a speedup of 1.0, each accelerator must provide performance equivalent to 3.29 Cheyenne nodes. By placing such a low value on accelerator performance relative to a Cheyenne node, a quite large number of accelerators is likely required to ensure that 20% of the CSEP metric is from heterogeneous nodes. This will result in a system balance tilted toward more heterogeneous nodes than one might expect, given NCAR's current workload as described in Attachment 5. Can NCAR confirm the spreadsheet computation of the CSEP fraction for MPAS-A?

The v1 version of the spreadsheet contained errors and deficiencies. UCAR is releasing, as part of Amendment #1 to UCAR RFP000074, a revised spreadsheet. Prospective Offerors should discard the originally-released Benchmark Results Spreadsheet and use the **'v1.1'** version released with Amendment #1 for calculating CSEPs.

5.13 Attachment 2A, Benchmark Results Spreadsheet

Question In the calculation for CSEP contributions for the homogenous codes, the formulae in column I divide by cell L27 (= 80%). Can NCAR please explain the

rationale for that and why there is not a corresponding term (divide by N27) for the heterogenous codes?

This was an error in the spreadsheet. A new v1.1 version of the spreadsheet, **UCAR_RFO000074_Attachment_2A_Benchmark_Results_Spreadsheet_v1.1.xlsx,** has been produced which corrects this error, and also contains additional updates. The revised spreadsheet is available from the NWSC-3 website at: https://www2.cisl.ucar.edu/resources/nwsc-3

5.14 NWSC-3 RFP, Section 2.8, Proposal Content and Format

Question

RFP Paragraph 2.8, Proposal Content and Format, Subparagraph (f). This subparagraph states that the combined 400 page limit includes all attachments, appendices, and all supplementary documentation. However, in requiring the submission of the Offeror's annual report, which in this Offeror's case is more than 150 pages, this requirement places an unnecessary limitation on the information offerors may submit for evaluation of the proposed solution. Would UCAR consider excluding 'supplementary documentation' from the page count? Such excluded supplementary documentation might be defined as "Annual Reports (2.8.1.2), Price Support Documentation such as GSA Schedules and Commercial Price Lists (2.8.1.10), Modified Copies of Attachment 4B Statement of Work, Attachment 4C Contact Information, and/or Attachment 4E Deliverable Requirements (2.8.1.13), Modified Copies of Attachment 4B Schedule B Statement of Work, Attachment 4C Contact Information, and/or Attachment 4E Deliverable Requirements (2.8.2.6)" and any other anticipated lengthy documentation that is supplementary in nature?

We have changed §2.8 of the RFP to remove item (f) from the 400 page limit (i.e.: attachments, appendices, and all supplementary documentation). Please see the new v1.1 version of the NWSC-3 RFP document, **UCAR_RFP000074_NWSC-3_RFP_v1.1.docx,** which removes the restriction, and also contains additional updates. The revised RFP document is available from the NWSC-3 website: https://www2.cisl.ucar.edu/resources/nwsc-3

5.15 NWSC-3 RFP, Section 2.16, Period of Acceptance of Proposals

Question

This paragraph requires that offerors agree to furnish any or all items at the price set forth in their proposal for up to 360 calendar days. With the volatility of many component prices brought about by the current global health emergency, this requirement may place an undue burden of risk on the offeror. While this offeror acknowledges that one or more interim submissions and final subcontract negotiations may provide opportunities to mitigate that risk through re-pricing, RFP Paragraph 3 reserves to UCAR the right to award on initial submissions without specifying a timeframe. As a result, this offeror respectfully requests this validity period be shortened to as little as 30 or 60 days. It may prove to be in UCAR's best interests as well by reducing the need for "risk mitigation" pricing by Offerors.

UCAR acknowledges the uncertainties that have been induced by the COVID-19 pandemic. Unfortunately, UCAR's proposal evaluation and selection process, along with requisite NSF sponsor approval, is sufficiently time consuming that sixty (60) days is untenable. However, in recognition of the pandemic uncertainties, UCAR is changing the proposal acceptance period to one hundred eighty (180) days. Please see the new v1.1 version of the NWSC-3 RFP document, **UCAR_RFP000074_NWSC-3_RFP_v1.1.docx**, for this and other updates. The revised RFP document is available from the NWSC-3 website at: https://www2.cisl.ucar.edu/resources/nwsc-3

6 RFP Questions and Answers, issued 04 May 2020

The following questions were received by UCAR between the release of Version 1.3 of this document, on 27 April 2020, and 04 May 2020.

6.1 Attachment 1, Technical Specifications, Section 3.4.3, Existing Ethernet

Question

Section 3.4.3: Can you elaborate on the existing Ethernet network that must connect to the NWSC-3 system? Whether through gateways or bridges, we understand that the RoCE-capable 100/200Gb Ethernet network for the PFS must be connected in some way to the facility Ethernet network. Since some network vendors don't necessarily coexist well with other network vendors, it would be helpful to know what exactly we need to connect to [network vendor and switch model number(s)], and how many ports are available to enable us to achieve the requested 200Gb/s throughput.

NCAR will expand the 100/200GbE network to accommodate connections to the PFS system or any gateways/bridges proposed. The current standard host interface within the NWSC is Mellanox ConnectX VPI HBAs and the current Ethernet switches are Juniper QFX10000-series switches. This said, we must caution Offerors that UCAR is currently in the process of upgrading the NWSC network infrastructure, and therefore cannot currently comment on the network equipment or its supplier. This information will be available prior to final subcontract negotiations.

6.2 Attachment 1, Technical Specifications, Section 3.12.12, PFS Expansion Footprint

Question

Section 3.12.12: Can we assume that should all or part of the second 60PB of PFS capacity be purchased that it could occupy additional space beyond what is shown in the illustration on Page 20 of the Technical Specification? Or must all 120PB fit within the shaded area shown?

The NWSC can accommodate a PFS footprint larger than noted on Page 20 of the Technical Specification. However, the expansion should be accomplished within the initial proposed infrastructure; therefore, NCAR would not expect an increase in floor space associated with the 60PB expansion of the initial PFS capacity. Please refer to Question 3.2 in this document for further guidance.

6.3 Attachment 1, Technical Specifications, Section 3.9.3, File System Lifetime Availability

Question: Section 3.9.3: When measuring the 99% uptime, is that inclusive of or independent of any scheduled maintenance? If inclusive, can you describe your expectations of scheduled downtime?

As stated in the **UCAR_RFP000074_Attachment_4_Terms_and_Conditions_v1.docx** document, File System Availability is defined as:

File System Availability = (∑_i^N (w_i*S_i - w_i*D_i))/(∑_i^N (w_i*S_i))

where:

N is the number of file systems in the PFS
S_i is the number of Scheduled Hours for file system i
D_i is the number of hours of downtime for file system i
w_i is the weight factor for file system i, determined by the ratio of the file system's size to the total size of all file systems

Scheduled Hours is wall-clock time minus Null Time minus any Downtime scheduled by UCAR.

File System Downtime is defined as "any period of time during which data residing in a file system is inaccessible from client systems, client systems cannot write data to the file system, or the file system's metadata is inaccessible. In calculating File System Availability, only downtime due to the failure of Subcontractor-supplied hardware or software applies."

The Offeror should refer to Article 1, Definitions, of Attachment 4 Terms and Conditions for all system reliability and other definitions, which is available from the NWSC-3 RFP website: https://www2.cisl.ucar.edu/resources/nwsc-3

6.4 Attachment 1, Technical Specifications, Sections 3.10.4 and 3.12.10, Environmental Monitoring

Question: Section 3.10.4 and 3.12.10: For environmental monitoring, are rack-level temperature sensors required? Are "smart" PDUs required? Can you provide more details of the available or planned facility monitoring capabilities we should integrate with?

There are no requirements of additional sensors or metering outside of what is contained in the RFP documents. The NWSC Facility Team is interested in what types of information are available from the proposed NWSC-3 HPC system and PFS solutions to see if there is a cost benefit for integration during the deployment. The NWSC Facility Team is interested in integrating any type of environmental information collected by either smart PDUs, sensors, or from the equipment inside the rack. A detailed list of currently used protocols

by the NWSC Facility Team was provided in the presentation from the NWSC Virtual Site Visit. The NWSC Facility Team is interested in understanding if these types of protocols are available for integration in any of the equipment in the proposed NWSC-3 solution, or if other standard protocols are available for integration with the facility. The NWSC EPMS system is extremely versatile and can integrate with many different protocols, but these protocols need to be identified so that licensing options can be discussed. Most equipment specification sheets have a list of integration points and what languages the equipment 'speaks.' Regarding smart PDUs, CISL's system administrators prefer switchable units with metering at the receptacle level, if possible.

6.5 Attachment 1, Technical Specifications, Section 3.12.4, Power Redundancy

Question: Section 3.12.4: Please elaborate on the available separate power sources. If we provide redundant power in each component, and the redundancy is divided across two rack PDUs, will it suffice to connect each of the two PDUs to a different facility power source to be compliant with this requirement?

All critical equipment must have N+1 capability, where half of the needed power supplies for each piece of equipment are plugged into a different PDU. Each PDU is sized to handle all of the load inside of the rack, and then each of these PDUs are plugged into a different NWSC facility power source. The design is to protect production use of the system from interrupt should a PDU fail. Additionally, this design allows the NWSC to shut down and perform maintenance on one power source to a rack while keeping the rack functional. Offerors should note that the system resilience testing (see the RFP's Attachment 4F, Acceptance Criteria and Testing) will test these capabilities.

6.6 Follow-up to NWSC-3 Communications Question 3.2

Question: The answer provided in response to the vendor question in Section 3.2 of version 1.3 (and 1.2 and 1.1) of the Communications document is a bit confusing. It seems to describe a preference for pre-installing all of the components for the second 60PB at the initial install, devoid of the HDDs and SSDs. This would seem to increase costs up front for the potential of expansion in the future. Our preference would be to provide only the components needed for the initial 60PB deployment at the initial installation time and, when additional capacity or performance is required, provide the appropriate number of "Scalable Storage Units" (Section 3.3.7) as add-on performance and capacity. This addition would be non-disruptive to the running PFS system, and we believe would make better use of the available funds. We believe the answer provided will allow this approach, but want to confirm our understanding.

UCAR prefers to purchase all the infrastructure for capacity expansion up front. This allows for the addition of HDDs/SSDs at a later date as budgets allow. There is no requirement for expansion of performance; therefore, the initial infrastructure should support the 300GB/s (gigabytes/second) minimum performance specification. All controllers and racks should be included to support the capacity expansion. If the

architecture allows for additional drive enclosures within existing racks and cabled to existing controllers, that is an acceptable alternative, but please provide the appropriate pricing for this solution.

6.7 Attachment 1, Technical Specifications, Section 3.3, PFS Encryption

Question — Section 3.3: For the PFS, are there any requirements for encryption-at-rest, and/or will there be any classified data stored that will preclude the return of failed storage devices to the manufacturer upon failure? This would impact the service and support costs for the PFS.

UCAR does not currently run encryption or support restricted/classified data on its HPC storage resources. At this time, failed hardware can be returned to the Offeror. If there are additional costs to support a restricted environment, please provide those as an option for future reference.

6.8 Attachment 1, Technical Specifications, Section 3.3.2, Spectrum Scale Licensing

Question — Section 3.3.2: With respect to IBM Spectrum Scale, does NWSC have a site license for any of the Spectrum Scale products, and/or will a new Spectrum Scale system require all new software licenses for the storage and cluster nodes?

UCAR currently runs the Data Management Edition of Spectrum Scale. As this is capacity-based licensing, the Offeror should include licensing for the initial 60PB of PFS capacity for Spectrum Scale Data Management Edition. UCAR may decide to request that IBM add this capacity to UCAR's current license agreement.

6.9 Attachment 1, Technical Specifications, Section 3.12.2, PFS Power Requirements

Question — Section 3.12.2: This section addresses the HPC system with a loose reference to "storage." Does the PFS also need to use "3-phase 400V to 480V AC power," or can it utilize the 208V power indicated here as "available"? If so, is there a limit to the number of 208V power drops? Also related, is there a make/model of rack PDU that is preferred by NWSC?

UCAR understands that most PFS solutions will be supplied power at 208v. There is no (reasonable) limit to the number of power drops, but the needed power drops for the PFS solution need to be plainly identified in the Technical Volume of the Offeror's proposal. The NWSC prefers to use Vertiv/Geist PDUs. The 60 Amp 208v 4 wire PDU used at the NWSC is: https://www.geistglobal.com/power-distribution-unit/NU30055L

While this exact PDU does not need to be used, these options are preferred by the system administrators and the preferred manufacturer at the NWSC.

6.10 Attachment 1, Technical Specifications, Section 3.4.3, NWSC LAN Connectivity

Question

Section 3.4.3: We understand that connectivity from the HPC system to the PFS will be over a 100/200GbE Ethernet fabric which must be included in the proposal at a data rate of at least 300GB/s, and that the existing GLADE system can also be accessed through the existing Ethernet fabric. Does this paragraph imply that the access from the HPC system to the GLADE system is subject to the 200Gb/s requirement of this paragraph? If not, what is the connectivity requirement from the HPC system to the GLADE storage? Also, the provided illustrations show that the HPSS Archive system is currently connected to an InfiniBand fabric. Will this system also be accessed through the existing Ethernet and be subject to the 200Gb/s requirement? If not, what connectivity is required to the HPSS Archive system?

The Offeror shall decide what network the PFS and HPC system will communicate over at a minimum of 300GB/s (gigabytes/second). There is an additional requirement that the PFS system be able to communicate on the NWSC 100/200GbE (gigabits/second) network to support access by additional clusters. The expectation is that storage servers or the HPC network will be connected to the 100/200GbE network providing a significant portion of the 300GB/s aggregate bandwidth to Ethernet-connected clients. GLADE currently supports 300GB/s aggregate bandwidth and is currently available on the NWSC network.

The HPSS Archive system is an Ethernet-based solution connected to the NWSC network. The requirement for the PFS to communicate on the NWSC network provides the connectivity necessary for UCAR Archival services.

6.11 Follow-up to NWSC-3 Communications Question 5.3

Question

The answer provided in the Communications v1.3 document in response to question 5.3 does not address the final point of the question, i.e. "Does this apply to Module B?" Since the PFS will be in Module B, does the PFS vendor need to provide the cable trays for just the inter-rack cables used in the PFS? Then will NWCS provide other cable trays to connect between the Module A and Module B and the main networking hubs? Is there a model/brand of cable tray currently in use that would be easier for the Vendor cable trays to integrate with?

The NWSC facility will provide cable tray paths between the modules and, if necessary, cable tray(s) will be installed on top of the proposed PFS solution where applicable. The Offeror shall provide the inter-rack cables and, unless a proprietary cable management solution for rack top needs to be provided, the NWSC will provide and install CommScope ADC 12" x 4" Yellow Fiber Tray and 12" CommScope Ladder Tray for Copper Connections.

6.12 MG2 Benchmark Question

Question

We found that the MG2 benchmark has incomplete input datasets for pcol=32 case. Instead of 20 input files, the benchmark directory provides only 9 files.

According to kgen_statefile.lst, it should be 20 files (11 files are missing). All other cases (pcols16/ pcols192/ pcols48/ pcols64/ pcols96) have a complete set of 20 files per each case. Would it be possible to provide missing input files for pcol=32?

UCAR recommends continuing as is with the existing data. UCAR requests that the Offeror return "Average columns per sec" for the best configuration of pcols. It is optional to provide results for all values of pcols. This allows the benchmark to adapt to hardware that may prefer a longer vector length.

6.13 Attachment 4F, Section 2.1, Pre-Delivery Assembly and Testing

Question: Do the requirements outlined in Attachment 4F, Section 2.1, apply to the storage component as well?

Yes, this requirement applies to storage (i.e. PFS) components.

6.14 Attachment 1, Technical Specifications, Section 3.3.2, File System Feature Matrix

Question: Section 3.3.2: Regarding the request for a matrix comparing the technical features of the PFS with those of IBM Spectrum Scale, we note that Spectrum Scale has an extensive list of features and options. Would it be possible to get a list of the specific features of Spectrum Scale that are currently being used at NWSC or that are critical to the functioning of the workload at NWSC? This would allow us to create a more focused comparison matrix.

Yes, the requested matrix will be provided by 08 May 2020 with the release of v1.5 of the Communications document.

7 RFP Questions and Answers, issued 08 May 2020

The following contains UCAR's response to the remaining question between the release of Version 1.4 of this document, on 04 May 2020, and 08 May 2020.

7.1 Attachment 1, Technical Specifications, Section 3.3.2, File System Feature Matrix

Question: Section 3.3.2: Regarding the request for a matrix comparing the technical features of the PFS with those of IBM Spectrum Scale, we note that Spectrum Scale has an extensive list of features and options. Would it be possible to get a list of the specific features of Spectrum Scale that are currently being used at NWSC or that are critical to the functioning of the workload at NWSC? This would allow us to create a more focused comparison matrix.

The list of the specific features of Spectrum Scale that are currently being used by NCAR, and other features which are desired to be used with NWSC-3, are listed in the following table.

Feature	Description	Used Today	Desired
Multi-Cluster	Sharing of resources across multiple SS clusters	X	
Snapshots	File-level snapshots, per fileset	X	
Filesets	Directory containers	X	
User Quotas	User level quotas	X	
Group Quotas	Group level quotas	X	
Fileset Quotas	Directory level quotas	X	
Protocols	NFS protocol service		X
cNFS	Clustered NFS	X	
TCT	Object Storage Interface		X
ILM	Policy Engine (used for purging, migrations, dumping file attributes)	X	
API	Programmatic access to GPFS metadata (DMAPI)	X	
SNMP	Feed events to central logging		X
Compression	File-level compression, by user choice		X
Syslog	Centralized syslog ability	X	
Storage Pools	Ability to tier storage pools, separate metadata	X	
LUN mgt	Ability to add LUNs to a file system Ability to delete LUNs from a file system Ability to rebalance LUNs in a pool/file system	X	
Multiple Data/MD copies	Maintain file system/fileset availability during partial storage system outages	X	
VERBS RDMA	Uses verbs over IB when possible	X	
ACL support	Both POSIX and NFSv4 ACLs	X	

Multiple Networks	Ability to serve a file system over multiple networks simultaneously (e.g. IB + IP)	X	

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

[2] UCAR_RFP000074_Attachment_2A_Benchmark_Results_Spreadsheet_v1.1.xlsx