Technical Specifications

NCAR’s Next Generation HPC System - NWSC-2

NWSC-2 Technical Specifications

1 Introduction 3

1.1 NWSC-2 Procurement Objectives 3

2 Mandatory Elements of Offeror Response 3

3 Target Design Specifications 4

3.1 Scalability 5

3.2 System Software and Runtime 6

3.3 Software Tools and Programming Environment 8

3.4 Parallel File System 9

3.5 Integration with Existing NWSC Data Services 11

3.6 Application Performance Specifications and Benchmarks 12

3.7 Reliability, Availability, and Serviceability 15

3.8 System Management and Operations 16

3.9 Buildable Source Code 18

3.10 Test Systems 18

3.11 Facilities and Site Integration 19

3.12 Maintenance, Support, and Technical Services 21

4 Technical Options 23

4.1 Many-core Compute Partition 24

4.2 General Purpose GPU Compute Partition 25

4.3 Data Analysis, Visualization and Post-processing 26

4.4 Innovative Storage and Memory Technologies 27

4.5 Early Access System 28

4.6 Software Tools and Programming Environment 29

4.7 Upgrades and Expansions to the NWSC-2 PFS System 30

4.8 Maintenance, Support, and Technical Services 31

4.9 Upgrades and Expansions to the NWSC-2 HPC System 32

4.10 AMPS System 32

5 Delivery and Acceptance Specifications 33

5.1 Pre-delivery Testing 34

5.2 Site Integration and Post-delivery Testing 34

5.3 Acceptance Testing 34

6 Risk Management and Project Management 34

7 Documentation and Training 35

7.1 Documentation 36

7.2 Training 36

8 References 37

9 Facilities Interfaces 38

# Introduction

The University Corporation for Atmospheric Research (UCAR), on behalf of the Computational Information Systems Laboratory (CISL) at the National Center for Atmospheric Research (NCAR) has released a Request for Proposal (RFP) for the next-generation high-performance computing (HPC) system to be installed at the NCAR-Wyoming Supercomputer Center (NWSC). This new system, herein referred to as NWSC-2, is expected to be delivered in the second half of CY2016 for production use in January 2017. This document provides the technical specifications for the NWSC-2 system.

The NWSC-2 system is expected to run for four years, with options to extend beyond that. The system must support interoperability with the existing GPFS-based NCAR Globally Accessible Data Environment (GLADE) and HPSS-based NCAR data archive.

## NWSC-2 Procurement Objectives

The primary objective of the NWSC-2 procurement is to provide a computational system that will support the demands of the atmospheric sciences community that computes on NCAR’s Yellowstone and related systems. To this end, the predominant characteristic of NWSC-2 is its ability to run NCAR’s existing applications. An overview of NCAR’s computational workload [1] summarizes key aspects of the application and job mix, and provides a quantitative assessment of how the current Yellowstone system is being used.

While rooted in production computing, the NWSC-2 procurement is being conducted with awareness of the performance limitations of today’s climate and weather models and thus there is an interest in novel technologies that can improve the performance of these applications on current computing architectures.

Beyond this production computing capability, NWSC-2 acknowledges the need to move towards new HPC architectures, such as many-core processors and GPGPU accelerators, which is being driven by limitations of traditional processor design, the need for finer resolutions in the simulations, and power constraints of large-scale HPC systems.

This RFP is structured to give vendors the flexibility to propose well-balanced solutions that meet NCAR’s production computing needs, while providing technical options that allow UCAR to understand cost and performance trade-offs among possible system choices within the available funding.

Definitions of terms used in this document are contained in Article 1 of the RFP’s NWSC-2 Sample Subcontract Terms and Conditions.

# Mandatory Elements of Offeror Response

An Offeror shall address all mandatory elements and its proposal will be deemed non-responsive and will receive no further consideration if any of the following mandatory requirements is not met.

Offerors may propose architectural choices in their designs that may be advantageous for UCAR to consider (e.g. choice of high speed interconnect topology or connectivity options, file system software, etc.). However, proposals must include all of the mandatory elements described below. The Offeror should submit a single proposal that covers all design options, with the differences presented in side-by-side comparisons in both the Technical and Business/Price volumes.

* + 1. The Offeror shall provide a detailed architectural description of the proposed NWSC-2 computational and/or storage production and test systems. The description shall include: a high-level architectural diagram that includes all major components and subsystems; detailed descriptions of all the major architectural hardware components in the system to include: node, cabinet, rack, multi-rack or larger scalable units (if applicable), up to the total system, including the high-speed interconnect and network topology; detailed descriptions of the system software components; the storage subsystem and all I/O and file system components; electrical and cooling requirements; and a proposed floor plan.  
         
       An Offeror proposing only a computational or storage solution may be silent on specifications inapplicable to their solution, but must address the integration and interoperability of the computational and storage systems, and the Offeror’s ability to work with UCAR and other resource providers to successfully deploy, test, maintain and support a total NWSC-2 solution. UCAR intends to address the details of how NCAR and the selected Offeror(s) will work together during subcontract negotiations.
    2. The Offeror shall provide benchmark results in accordance with the proposed computational and storage solutions.
    3. The Offeror shall provide a detailed plan for delivery, installation, maintenance and support services necessary to meet the NWSC-2 target reliability through the proposed system lifetime, and a proposed delivery, installation and acceptance testing schedule for the NWSC-2 system, including the number and roles of any temporary or long-term on-site Offeror personnel.
    4. The Offeror shall describe how the proposed system fits into their HPC roadmap for the period of deployment as well as a potential NWSC-3 acquisition for the 2020-2023 timeframe.

# Target Design Specifications

This section contains detailed system design targets and performance features. It is desirable that the Offeror’s design meets or exceeds all the features and performance metrics outlined in this section. Failure to meet a given Target Design Specification will not make the proposal non-responsive. However, if a Target Design Specification cannot be met, it is highly desirable that the Offeror either provide a development and/or deployment plan and schedule to satisfy the specification or describe trade-offs the Offeror’s solution provides in lieu of the specification.

The Offeror shall address all Target Design Specifications and describe how the proposed system meets, exceeds, or does not meet the Target Design Specifications. Offerors proposing only a computational or storage solution should respond ‘N/A’ to those specifications that do not apply to their proposal.

The Offeror shall also propose any hardware and/or software architectural features that will provide improvements for any aspect of the system. Areas of interest include, but are not limited to: application performance, storage and memory technology, power usage, overall productivity of the system, and systems management.

## Scalability

The NWSC-2 production system workload will include large-scale jobs, up to and including full-system size; therefore, the system must scale well to ensure efficient usage. However, the anticipated job mix is likely to be dominated by jobs at smaller scales (see the Yellowstone Workload Study [1]), and thus CISL anticipates the Offeror’s system will reflect scalability trade-offs in order to serve the overall production workload.

* + 1. The system shall support hundreds of concurrent users and thousands of concurrent batch jobs. The Offeror shall describe and provide details on how the system supports this, including, for example, the number and design of login nodes required to support the routine workload associated with interactive use of the system.
    2. The system shall support running a single application at full scale.
    3. The system shall provide reproducible numerical results and consistent run times. The Offeror shall describe strategies available to system administrators and to user applications for minimizing runtime variability. An application’s runtime (i.e., wall clock time) shall not change by more than 3% from run-to-run in dedicated mode and 5% in production mode. Variability will be measured by using the Coefficient of Variation.
    4. If the system has heterogeneous node architectures, the Offeror shall describe any associated scalability limitations, impacts to the high-speed interconnect, and the scalability of applications running on each set of homogeneous nodes.
    5. The system’s high-speed interconnect shall support high bandwidth, low latency, high throughput, and minimal inter-job interference. The Offeror shall describe the high-speed interconnect in detail, including topology, all performance characteristics, mechanisms for adapting to inoperable links and heavy loads (including mixes of I/O and inter-process communication), and dynamic responses to failure and repair of links, nodes, and other systems components.
    6. The Offeror shall describe how the high-speed interconnect design and related software represents an appropriate balance between supporting single full-system jobs at full-scale and supporting the workload of smaller-scale jobs that are representative of the Yellowstone workload (see workload study [1]).
    7. The Offeror shall provide details on the flexibility in the proposed interconnect design and topology that could be, if it is deemed in the best interest of the overall system performance and capability for UCAR’s workload, exercised during subcontract negotiations. The Offeror should therefore provide design, performance, and pricing information that can be used by UCAR in assessing said flexibility and how it impacts the overall system and applications performance.

## System Software and Runtime

The Offeror shall propose a well-integrated and supported system software environment that is high-performing and reliable.

* + 1. The Offeror shall provide a system that includes a full-featured, POSIX compliant, Unix-like operating system (OS) environment on all user-visible OS instances including compute nodes, login nodes, service nodes, and management servers.
    2. The Offeror shall describe any system software optimizations or support for a low-jitter environment for applications and shall provide an estimate of a compute node OS’s noise profile, both while idle and while running a non-trivial MPI application (e.g., one of the benchmarks described in §3.6), including jitter-induced application runtime variability. If core specialization is used, describe the system software activity that remains on the application cores.
    3. The Offeror shall describe the security capabilities of the full-featured, Unix-like OS. The OS for the login nodes, service nodes and system management nodes shall provide at a minimum the following security features: ssh version 2, POSIX user and group permissions, kernel-level firewall and routing capabilities, centralized logging, and auditing.
    4. The compute partition OS shall provide a trusted, hardware-protected supervisory mode to implement security features. The Offeror shall describe how the supervisor/kernel provides authoritative user identification, ensures that user access controls are in place, employs the principle of least privilege, and interoperates with the same features on the login, service and management nodes. Logging and auditing features supported by the compute node OS shall have the capability to be enabled, disabled and custom-configured to site preferences.
    5. The Offeror shall describe how the system provides support for static libraries and objects and/or dynamic loading of shared objects. The Offeror should describe how the system will support applications using these techniques at the full scale of the system.
    6. The Offeror shall describe how the system provides efficient, secure, inter-process communication that allows cooperating applications running anywhere on the high-speed network to inter-communicate (e.g., the compute partition, the service partition, or both). The provided mechanism shall be as close to the underlying network stack as possible. The security model shall allow applications and users to set access controls based on authenticated or trusted values for process and user identifiers.
    7. The Offeror shall propose a job scheduler and resource management subsystem, and all necessary licensing, capable of simultaneously scheduling both batch and interactive workload.

#### The Offeror shall describe the features and capabilities available to administrators and users, including: hierarchical fair-share, backfill, targeting of specified resources, advance and persistent reservations, job preemption, monitoring of running and pending jobs, job reporting and accounting, architecture-aware job placement, and any unique features relevant to the Offeror’s proposed system.

#### The proposed job scheduler and resource management subsystem shall support an efficient mechanism to launch applications at sizes up to full scale. The Offeror shall describe the factors (such as executable size, number of jobs currently running or queued, and so on) that affect application launch time. The Offeror shall describe expected application launch times and how the factors noted increase or decrease the launch time.

#### If required to meet the §3.1.3 run time specifications, the proposed job scheduler and resource management subsystem shall utilize an optimized job-placement algorithm to reduce job runtime, lower variability, minimize latency, etc. The Offeror shall either explain why such an algorithm is unnecessary or describe in detail how the algorithm is optimized to the system architecture, and how it affects application efficiency and overall system efficiency and utilization.

* + 1. The Offeror shall describe its software development and release plan, regression testing and validation processes, for all system software, including security and vulnerability updates.

## Software Tools and Programming Environment

The primary parallel programming model currently used on existing NCAR systems is hybrid Message Passing Interface (MPI) with OpenMP. To support current climate and weather models that form the large majority of NCAR’s production workload, the Offeror’s proposed system shall support the hybrid MPI/OpenMP programming model for its primary production workload.

For the following software categories, the Offeror should describe the proposed set of the Offeror’s own optimized, integrated and/or recommended software tools and programming environment components, including any third-party software that CISL may have a compelling reason to make part of the NWSC-2 procurement.

* + 1. The production system shall support the Message Passing Interface 3.0 (MPI-3) standard specification. The Offeror shall describe the proposed MPI implementation, including version, optimizations for collective operations, support for features such as hardware-accelerated collectives, and the ability for applications to access the physical-to-logical mapping of the job’s node allocation, and describe any limitations relative to the MPI-3 standard.
    2. The Offeror shall describe and provide licenses for a minimum of 50 seats for a proposed set of high-performance, optimizing compilers capable of creating executables for the compute partition of the HPC system. These compilers shall support the latest International Standards for C, C++, and Fortran.
    3. The Offeror shall describe all programming APIs, languages, compiler extensions, etc., other than MPI (e.g., OpenMP, OpenACC, CUDA, OpenCL, and so on) by the compilers proposed in §3.3.2.
    4. For the compilers proposed in §3.3.2 Offeror shall describe support for mixed-language programs. In particular, use of parallel features of any one of C, C++ or Fortran shall not interfere with the use of parallel features in the others.
    5. For the compilers proposed in §3.3.2, the Offeror shall describe the support for partitioned global address space (PGAS) features, if any, such as coarray Fortran or Unified Parallel C (UPC). The Offeror shall describe production system hardware and programming environment software for exploiting PGAS capabilities.
    6. The Offeror shall list and describe all optimized mathematical and I/O libraries proposed for the production system.
    7. The Offeror shall describe any proposed software components that provide control of task and memory placement within a node for efficient performance of applications, the controls provided, and any limitations that may exist.
    8. The Offeror shall describe any other optimized, integrated or recommended programming environment components and software tools that are proposed as part of the programming environment and/or optimized compiler tool suite. These may include debuggers, performance analysis tools including those accessing node and fabric hardware counters, event-tracing tools, and stack-tracing tools. Such tools should support at least 20 simultaneous users and a single job up to at least 2000 cores of the production system.

## Parallel File System

The Offeror shall propose hardware and software to support a high-performance parallel file system (PFS) that presents one or more globally-consistent name space(s) to the HPC platform. The Offeror shall explain how the proposed PFS solution satisfies each of the following specifications.

* + 1. The Offeror shall provide a detailed description of the PFS implementation including:
* File system architecture and proposed implementation
* Expected scaling characteristics
* Management, diagnostic, deployment, security and configuration tools
* Externalized error and diagnostic information
* Support for user and group quotas at filesystem and finer granularity (e.g. directory/container/fileset) levels
* Real-time monitoring tools
  + 1. The PFS shall have a minimum usable capacity given in Table 1.
    2. The PFS shall achieve the target aggregate bandwidth given in Table 1.
    3. The PFS shall be externally accessible with aggregate external bandwidths as specified in Table 1.
    4. The peak bandwidth shall be achievable by the Interleaved Or Random I/O benchmark (IOR) running on arbitrary sets of compute nodes. The Offeror shall demonstrate that this is achieved when running IOR on 10%, 25%, 50%, 75% and the full-scale system, and that the aggregate bandwidth does not degrade as the number of nodes increases. The Offeror shall also demonstrate this is achieved when the file system is 70% full (data space) and contains 500,000,000 objects (metadata entries).
    5. The PFS shall be able to operate independently of the production NWSC-2 system and be capable of serving data to external clients when the production NWSC-2 system is down.
    6. The PFS shall be resilient and have no single point of hardware failure or failure due to loss of one of the redundant facility power feeds.
    7. The parallel file system shall be designed to minimize the risk of data corruption and data loss. Reliability of the PFS will be assessed as part of the overall system reliability metrics. For the proposed PFS implementation, the Offeror shall provide a detailed description of the following:
* How the I/O solution strives to achieve zero corruption and zero data loss over the life of the system
* How silent data corruption errors (e.g., high-fly writes, short writes, misdirected I/Os) are addressed
* How the I/O solution will function when a failure is exposed beyond the RAID level
* Ability to run in a partial or degraded mode
* Estimates of failure rates of the various components comprising the PFS implementation.
  + 1. The PFS shall provide a robust, interactive environment for users that is prompt and not substantially impacted by unrelated applications running on the rest of the HPC system. The Offeror shall describe the time required to insert, delete, enumerate, and retrieve one million file system objects within a single directory on login, compute, or file transfer service nodes.
    2. The Offeror shall describe how the PFS directories are structured to prevent a linear search for a random child object within a directory containing one million objects.
    3. The NWSC-2 systems shall provide POSIX I/O and MPI I/O functionality that is tightly integrated with file system software to provide high-performance small- and large-block I/O. The Offeror shall discuss the relative performance of MPI I/O shared file versus POSIX I/O, including the case of a single file per processor at full system bandwidth.
    4. The PFS system shall support Ethernet based TCP/IP client access.
    5. The Offeror shall describe expected PFS maintenance procedures and their impacts on the PFS performance under normal load and other routine operations including: file system start-up, shut-down, purging, health monitoring, performance statistics, problem alerts, diagnosis and repair, and reconstruction after a drive replacement, including the time required for each of these activities.
    6. The Offeror shall provide the amount of time required to perform a file system consistency check when a 20 PB file system is at 70% full with 500,000,000 objects.
    7. The Offeror shall provide low level test procedures and code specific to the storage hardware. These tests shall be used to show that hardware performance of all components of the system are within vendor specifications. At least a subset of these tests must be able to be run with a file system present and in read-only mode. It is NCAR's intent to use these tests as hardware verification after hardware repair or upgrade.
    8. The Offeror shall describe how the PFS design allows for upgrades in performance, capacity, or both. Specific options for upgrading are requested in §4.7.

## Integration with Existing NWSC Data Services

NCAR currently operates the GPFS-based GLADE filesystem services, the HPSS-based NCAR Archive, and an Ethernet-based local area network at the NWSC. The NWSC-2 systems must integrate into the overall NWSC environment and interoperate with these existing resources. The Offeror shall work with UCAR to identify appropriate technologies for accomplishing that integration.

* + 1. The NWSC-2 HPC system shall interoperate with NCAR’s existing GPFS-based GLADE resource, providing an aggregate, sustainable bandwidth specified in Table 1. This bandwidth is in addition to the bandwidth provided for access to the NWSC-2 production system’s PFS. UCAR and IBM have executed a GPFS Multi-System Statement of Work (SOW) which provides for GPFS licensing at UCAR[[1]](#footnote-1). The Offeror shall provide a technical solution for integration of the NWSC-2 HPC system with the existing GLADE resource; UCAR will be responsible for any requisite GPFS licensing. Figure 1 in §9 provides a high-level diagram of the network architecture of the current GPFS deployment and integration with Yellowstone and the *NCAR GLADE Environment Overview* at [1] provides additional details of the existing GLADE environment.
    2. The NWSC-2 PFS system shall provide connectivity to NCAR’s TCP/IP network, including NCAR’s HPSS-based data archive, with an aggregate, sustainable bandwidth in excess of that specified in Table 1. This bandwidth is in addition to the bandwidth provided for access to the NWSC-2 production system’s PFS and that specified in §3.5.3.
    3. The NWSC-2 PFS system shall support other NCAR clients, providing an aggregate, sustainable bandwidth in excess of that specified in Table 1. This bandwidth is in addition to the bandwidth provided for access to the NWSC-2 production system’s PFS and that specified in §3.5.2.

## Application Performance Specifications and Benchmarks

Assuring that real applications perform well on the NWSC-2 platform is critical to the success of the system. Because the full applications are large, often with millions of lines of code, NCAR has put together a suite of applications that will be used to assess NWSC-2’s sustained performance, scalability, and performance as part of proposal evaluation and during system acceptance. These consist of

* a suite of application kernels and full applications with reduced or simplified input data sets, referred to as the NCAR Benchmarking Suite (NBS),
* micro-benchmarks designed to test memory, GPU, and interconnect subsystems, and
* I/O benchmarks designed to test the parallel file system.

These applications, which are listed in Table 2 and Table 3, are representative of real applications, but are smaller in terms of the effort and compute resources required to run them. Refer to Attachment 2, NWSC-2 Benchmark Instructions, available from the NWSC-2 procurement website [1], for information on obtaining, building and running the benchmarks and for guidance on the benchmark-specific information to return to UCAR.

* + 1. The Offeror shall provide performance results (actual, predicted and/or extrapolated) for the proposed systems for the benchmarks listed in Table 2 and Table 3. The Offeror shall report results of selected benchmarks in the NWSC-2 Benchmark Results spreadsheet available on the NWSC-2 Procurement website [1].
    2. The Offeror shall provide licenses for the delivered system for all compilers and tools used to achieve benchmark performance. Licenses may be provided on a temporary basis if the associated software products are not part of the delivered system.
    3. The Offeror’s proposal shall state a minimum sustained performance as measured by a weighted average of several of the NBS applications. The NWSC-2 system must meet or exceed the target minimum Yellowstone Equivalent Sustained Performance (YSEP) at acceptance. The target is a minimum of a 2x increase over the Yellowstone platform [2]. Refer to the NWSC-2 Benchmark Results spreadsheet for Yellowstone’s baseline YSEP and the constituent applications used to compute it.
    4. Following award, and prior to delivery of the production NWSC-2 system, Offeror will be required to validate that the production HPC system generates correct solutions for all of the applications in the NBS suite. These runs will be made on the same processor and interconnect technology that will be delivered on the production HPC system, but run at smaller scale. The validation runs will be performed jointly by the Offeror and NCAR staff. The optional Early Access System (§4.5) may be used for this testing, provided it meets the criteria above.
    5. As part of the acceptance testing of the final system, Offeror must demonstrate that application scalability is as good as or better than the Yellowstone system. This test will use the MPAS application, noted in Table 2, with a high-resolution input case.
    6. All performance tests must continue to meet acceptance performance and reproducibility criteria throughout the lifetime of the system.

1. **Target Configuration and Performance Specifications**

| **Attribute** | **NWSC-2** |
| --- | --- |
| Minimum sustained HPC system performance increase over Yellowstone [2] (§3.6.3) | 2 |
| Minimum memory on an HPC system compute node. This metric is for main memory capacity only, e.g. DDR4. It does NOT include memory associated with caches, accelerators, scratch pads, etc. | 64 GB |
| Aggregate external bandwidth on/off the HPC system for accessing NCAR’s GLADE file system (§3.5.1) | 100 GB/s |
| Minimum usable disk capacity of the PFS system (§3.4.2) | 20 PB |
| Minimum bandwidth of the PFS system (§3.4.3) | 400 GB/s |
| Aggregate external bandwidth on/off the PFS system for general TCP/IP connectivity, including NCAR HPSS archive and other data services (§3.5.2) | 25 GB/s |
| Aggregate external bandwidth on/off the PFS system for access by other NWSC client systems. This could include, for example, the existing Caldera, Geyser, and/or Yellowstone systems (§3.5.3) | 100 GB/s |

1. **NCAR Benchmarking Suite (NBS)**

| **App Kernel Name** | **Description** | **Availability** | **RFP**  **Response** | **Acceptance** |
| --- | --- | --- | --- | --- |
| HOMME  HOMME\_COMM | HOMME and its communication kernel | Code and input cases available from the NWSC-2 procurement site. | X | X |
| HPCG | High Performance Conjugate Gradient Solver | Code and input cases available from the NWSC-2 procurement site. | X | X |
| LES | Large Eddy Simulation kernel | Code and input cases available from the NWSC-2 procurement site. | X | X |
| MG2 | Atmosphere microphysics kernel | Code and input cases available from the NWSC-2 procurement site. | X | X |
| MPAS-A | Model for Prediction Across Scales | Code and input cases available from the NWSC-2 procurement site. | X | X |
| POPperf | Ocean circulation model | Code and input cases available from the NWSC-2 procurement site. | X | X |
| WRF | Weather Research and Forecasting model | Code and input cases available from the NWSC-2 procurement site. | X | X |
| Community Earth System Model | CESM | Not to be included in proposals. Will be part of numerical correctness and system acceptance testing. |  | X |

1. **I/O and Micro-benchmarks**

| **Benchmark** | **Description** | **Availability** | **RFP Response** | **Acceptance** |
| --- | --- | --- | --- | --- |
| **Micro-Benchmarks** | | | | |
| OSU MPI Benchmarks | Interconnect performance | Code and input cases available from the NWSC-2 procurement site. | X | X |
| STREAM | Memory bandwidth | Code and input cases available from the NWSC-2 procurement site. | X | X |
| SHOC | Scalable HeterOgeneous Computing benchmark | Code and input cases available from the NWSC-2 procurement site. | X | X |
| **I/O Benchmarks** | | | | |
| IOR | I/O latency and bandwidth | Code and input cases available from the NWSC-2 procurement site. | X | X |
| mdtest | Metadata performance | Code and input cases available from the NWSC-2 procurement site. | X | X |
| Pyreshaper | Application I/O kernel | Code and input cases available from the NWSC-2 procurement site. | X | X |

## Reliability, Availability, and Serviceability

For each metric specified below, the Offeror must describe how they arrived at their estimates. Terms used in this section can be found in Article 1 “Definitions” of Attachment 5, NWSC-2 Sample Subcontract Terms and Conditions.

* + 1. The NWSC-2 production HPC system shall meet or exceed 98% System Availability.
    2. The NWSC-2 production PFS shall meet or exceed 99% File System Availability.
    3. The NWSC-2 production system’s System Mean Time Between Interrupt (SMTBI) shall exceed 14 days.
    4. Any component of the NWSC-2 production systems that fails shall be expeditiously repaired (cf, §3.12). No component may fail and remain out of service longer than seven calendar days unless mutually agreed to by the Offeror and UCAR.
    5. Failure of the system management and/or RAS (Reliability, Availability, and Serviceability) system(s) shall not cause a system or job interrupt. This does not apply to a RAS system feature that automatically shuts down the system for safety reasons, such as an overheating condition.
    6. The Offeror shall describe the full system initialization sequence and timings, including the time to complete the HPC system initialization. HPC system initialization is defined to be the time to power up, and initialize 98% of the installed compute resource including all peripherals and 100% of any service resource to the point where a job can be successfully launched.
    7. The Offeror shall describe the full system initialization sequence and timing, including the time to complete the PFS system initialization. PFS system initialization is defined to be the time to power up and initialize all file system servers, controllers and subsystems to the point where all hosted filesystems can be accessed from applications running on external clients.
    8. The Offeror shall discuss the RAS mechanisms and capabilities of the proposed HPC and PFS systems including, but not limited to:
* Any condition or event that can potentially cause a job interrupt or loss of file system accessibility
* Resiliency features to achieve the availability targets
* Single points of failure, hardware or software, and the potential effect on running applications and system availability
* How a job maintains its resource allocation and is able to relaunch an application after an interrupt
* How a file system remains available when a PFS component (server, controller, drive) fails
* How efficiency and availability is monitored and calculated when the system or file system hangs

## System Management and Operations

The Offeror shall provide a centralized ability to manage and operate the HPC and PFS systems independently as specified in the following subsections.

* + 1. The Offeror shall describe the proposed HPC and/or PFS system management and operation infrastructure, including its networking, number of management and boot nodes, boot filesystems, and other requisite infrastructure used for system management and operation.
    2. The HPC and PFS system management and monitoring capabilities shall be centralized, integrated, and scalable and provide: human interfaces and APIs for system configuration and its ability to be automated; software management; change management; and local site integration. The description shall include the effects and overhead of software management tool components on CPU, network and/or memory of system nodes.
    3. The Offeror shall describe and provide a means for independent, centralized HPC and/or PFS system management, tracking and analyzing of all software and firmware updates, software and hardware changes and failures, and hardware replacements over the lifetime of the system. The Offeror shall also include the estimated time and effort required to install both a major and a minor system software update. Centralized management support shall include:
* Updates that do not require full-system outages must allow previous version and next version to be simultaneously running on independent nodes, or partitions, the production systems
* Support for multiple simultaneous or alternative system software configurations
* Notwithstanding upgrades required for reasons such as security, it is highly desirable that all system software must be upgradable without need for complete reinstall or major service interruptions.
  + 1. The Offeror shall describe the processes and procedures in place for generating and providing software updates for the proposed HPC and/or PFS OS. Processes should allow expeditious updating of kernel and non-kernel packages to address issues that impact user application and system performance, including security vulnerabilities, in the suite of software.
    2. The HPC and PFS system management capabilities shall provide
* a single, scalable log analysis capability for all logs originating from any component of the proposed system
* user activity tracking, such as audit logging and process accounting
* logs for the PFS and HPC systems shall be separately stored to facilitate independent analysis
  + 1. The Offeror shall describe how the HPC and/or PFS system can be operated, power managed, and administered from a remote location via lights-out management. This may be in the form of an embedded controller, or an equivalent capability, that allows the system to be managed remotely.
    2. CISL system administrators shall have unrestricted privileged access to all hardware and software components delivered with the system.
    3. All critical components must provide hot-swap redundancy for all services that are required for normal production operations.
    4. The Offeror shall provide CISL with a complete plan for support of offline backups and bare metal recovery of all components and systems. This should include any procedures required to configure components for remote administration, including but not limited to, UEFI/firmware parameters, and BMC/IMM parameters.

## Buildable Source Code

The Offeror shall describe the software components that will be provided as buildable source code.

* + 1. Source code, and necessary build environment, shall be provided for all software except for firmware, compilers and third-party products. Exceptions will be granted for vendor-supplied software that is proprietary in nature and/or where copyrights do not permit distribution of source.
    2. Updates of source code, and any necessary build environment, for all software shall be provided over the life of the subcontract.

## Test Systems

The Offeror shall propose test HPC and PFS systems. The test systems will be used for testing upgrades to the production NWSC-2 systems, therefore they shall contain all the same hardware, software and functionality as the production systems, but scaled down to an appropriate configuration.

* + 1. The test system shall be comprised of an HPC and PFS resource with architecturally appropriate sizing of: the greater of 60 batch computation nodes or a suitable number of nodes conforming with either the Offeror’s scalable unit or a comparable architectural unit (e.g., a single rack); two login nodes; minimum of 250 terabytes of usable, dedicated filesystem storage; and all requisite administrative resources to be independently administrable and operable. Components that are redundant in the production systems must also be present in the test systems as redundant components.
    2. The test system shall not share any internal or external communications paths, peripherals, storage, power supplies or subsystems required for normal operation with the NWSC-2 production HPC and PFS systems.
    3. All procedures for the production system shall be testable on the test system, including upgrades, patches, operation, monitoring, etc. Similarly, the test system PFS shall be externally mountable to other test platforms just as the production PFS is externally mountable to Yellowstone.
    4. The test system shall be delivered to and installed at the designated NCAR facility a minimum of 30 days prior to initial production system equipment delivery. It is highly desirable to have the system delivered 60 days prior to initial production equipment delivery.

## Facilities and Site Integration

NWSC facility attributes are provided in Table 4 and target NWSC-2 HPC and PFS locations within the NWSC facility along with notional diagrams of site network connectivity can be located in §9. These are intended to assist Offerors in understanding the integration requirements for NWSC-2 and related systems. These are subject to change prior to the NWSC-2 deployment.

* + 1. The computational system shall use 3-phase 480V AC. Both four and five-wire cabling can be accommodated. If line-to-neutral power supplies are used, phase balancing is necessary and must be verified. Other power sources (208V, 110V) are available to support the system's infrastructure such as disks, switches and consoles.
    2. All equipment and power control hardware shall be Nationally Recognized Testing Laboratories (NRTL) certified. All equipment shall bear appropriate NRTL labels. All equipment shall comply with the IEEE, NEC and NFPA environmental standards and codes as referenced in Section 7, particularly for Performance Level 2 systems as defined in IEEE 1156.2-1996. All proposed equipment shall comply with the new Class 1 & 2 recommended operating environment range as specified in the 2008 ASHRAE Environmental Guidelines for Datacom Equipment.
    3. Every rack, network switch, interconnect switch, node, and disk enclosure shall be clearly labeled with a unique identifier visible from the front and/or rear of the rack, or on pull-out tabs, as appropriate, when the rack door is open. These labels will be high quality to be usable and readable throughout the lifetime of the system. Nodes shall be labeled on the service-access side with a designator matching that assigned via administrative software and a unique serial number for inventory tracking.
    4. All NWSC-2 system power cabling and water connections shall be below the access floor. All other cabling (e.g. system interconnect, administrative networking) should be above floor. The systems shall be provided with cable trays or cable containment integrated with and spanning between the system cabinetry for the system interconnect and networking cables. All cables shall be plenum rated. All communications, power and other cables, wherever installed, shall be labeled with source/destination and a unique serial number at both ends.
    5. The Offeror shall describe the features of the system related to facilities and site integration including:

#### Remote environmental monitoring capabilities of the system and the proposed integration into facility monitoring.

#### Detailed descriptions of power and cooling distributions throughout the system including power consumption and cooling requirements for all subsystems, at idle, observed maximum (e.g., HPL), and design limit states.

#### Detailed descriptions, quantities and types of all electrical, mechanical connections made to facility infrastructure.

#### OS distributions or other client requirements to support off-platform access. This would include, for example, remote management and troubleshooting of hardware and software components

* + 1. The Offeror shall provide a description of facility and installation planning services with their proposal. The description shall include the facility preparation and planning processes to be conducted with UCAR, including shipping, receiving and staging of NWSC-2 equipment and all on-site assembly thereof, as well as all logistics information, including crated and uncrated sizes, weights and floor loading.
    2. The Offeror shall provide transportation, delivery and installation of all NWSC-2 equipment as well as replacement and spare parts. The Offeror shall provide unpacking, uncrating, assembly and interconnection of the NWSC-2 system components at the NWSC facility in Cheyenne, WY, (and the NCAR Mesa Laboratory facility in Boulder, CO, if necessary). The Offeror shall remove all packing materials and trash associated with delivery and installation.

1. **NWSC-2 Facility Specifications**

| **NWSC-2** | |
| --- | --- |
| Location | NCAR-Wyoming Supercomputer Center, Cheyenne, Wyoming.  The production system will be housed in NWSC Module A, which will be built out to support the system. |
| Altitude | 6,260 feet |
| Seismic | N/A |
| Water Cooling | The system must operate within ASHRAE TC 9.9 temperature ranges. NWSC provides 65F chilled water, currently has 75F return water, and can accommodate higher return temperatures up to 80F without coordination with or modifications to the NWSC facility. The NWSC chilled water system can accommodate a large range of flow rates. Offerors shall provide requirements for the proposed equipment. De-ionized water is available. |
| Air Cooling | The system must operate within ASHRAE TC 9.9 temperature ranges. |
| Maximum Power | 3 MW |
| Maximum Power Rate of Change | No restrictions. |
| Floor | 10’ raised floor |
| Ceiling | 12’ ceiling; maximum cabinet height is 9’ 9” |
| Maximum Footprint | 12,000 square feet (inclusive of compute, storage and service aisles) |
| Shipment Dimensions and Weight | For delivery, system components shall weigh less than 7,000 pounds. All doors and pathways are 6’ 0” in width and 9’ 9” in height, or larger. |
| Floor Loading | The floor loading shall not exceed a uniform live load of 250 pounds per square foot with a deflection of not more than 0.04 inch, and a concentrated load of 2,500 pounds on one square inch. |
| Cabling | All power cabling and water connections are below the access floor. All other cabling (e.g. system interconnect, administrative networking) are above floor. |
| External network interfaces supported by the site for connectivity requirements specified below | 1GbE, 10GbE, 40GbE, 100GbE |

## Maintenance, Support, and Technical Services

The Offeror shall propose maintenance and support with the following minimum features.

* + 1. Pricing and the Maintenance Period

The Offeror shall propose technical services, warranty, maintenance and support for a period four (4) years subsequent to the date of Acceptance of the NWSC-2 system by UCAR. The maintenance and support pricing shall be for each year of the above period after the warranty expires. The warranty period begins at the date of Acceptance of the system.

* + 1. Maintenance and Support

The Offeror shall describe proposed maintenance and support services which, at a minimum: (1) provide replacement hardware for all failed components and return shipping of failed components to the Offeror, (2) train and certify NCAR staff to perform hardware failure diagnosis, isolation and repair activities for all field-replaceable unit (FRU) components, provide Offeror employees or contractors to perform those activities, or a combination thereof, and (3) supply hardware maintenance procedural documentation, training, and manuals.

* + 1. On-site Parts Cache

The Offeror shall maintain a parts cache on-site at the NWSC facility. The parts cache shall be sized and provisioned sufficiently to support all normal repair actions for two weeks without the need for parts refresh. The initial sizing and provisioning of the cache shall be based on Offeror’s MTBF estimates for each Field Replaceable Unit (FRU) and each rack, and scaled based on the number of FRU’s and racks delivered. The parts cache configuration will be periodically reviewed for quantities needed to satisfy this requirement, and adjusted if necessary, based on observed FRU, component or node failure rates. The parts cache will be resized, at the Offeror’s expense, should the on-site parts cache prove to be insufficient to sustain the actually observed FRU or node failure rates.

* + 1. Hardware and Software Protection Plan

The Offeror shall support and maintain the hardware comprising the NWSC-2 systems and the software stack delivered with the NWSC-2 systems for the duration of the base subcontract plus any optional extensions.

* + 1. Software and Firmware Update Service

The Offeror shall provide an ability for UCAR to obtain updates for all software and firmware provided with the NWSC-2 systems and guidance on Offeror tested/recommended software stack versioning and best practices. This shall include new releases of software/firmware and software/firmware features, bug fixes and security patches as required.

* + 1. Problem Reporting and Resolution Service

The Offeror shall provide 24x7 telephone and web-based problem reporting, ticketing, and resolution services. The Offeror’s ticketing system shall be full-functioned and provide participant roles, keyword searches, a unified view of hardware and software problem reports/resolutions, and the ability to transition and associate tickets between hardware and software.

* + 1. Production Transition Support

The Offeror shall provide remote and/or on-site assistance to UCAR to transition the production workload and key production applications within NCAR’s production workload to the NWSC-2 system from its predecessor system. This shall include system and file system configuration and tuning, software problem isolation and resolution, and application tuning and optimization. This support shall be provided from subcontract execution through six months after system Acceptance.

# Technical Options

This section contains options to the base system described in §3. Some options may represent collaborations between NCAR, with the Offeror providing functionality that doesn’t currently exist or is inadequate in the current marketplace, but could be delivered during the term of the subcontract. Table 5 provides Offerors with guidance to assist them in developing their responses to each of the specific technical options.

1. **Technical Options Overview**

| **Technical Option** | **Guidance** |
| --- | --- |
| **Technological enhancements to the base systems** | |
| §4.1 Many-core partition | Development work may be required by both the Offeror and UCAR. In proposing a delivery date for the many-core component, Offeror should consider market maturity, reliability, and programming model, among other criteria. |
| §4.2 General Purpose GPU | General market availability of the GPGPU is assumed. If exercised, some quantity is expected at initial system delivery. Subsequent quantities are possible. |
| §4.3 Data Analysis, Visualization and Post Processing | General market availability assumed. If exercised, some quantity is expected at initial system delivery. Subsequent quantities are possible. |
| §4.4 Innovative Storage and Memory Technologies | Development work may be required by both the Offeror and UCAR. In proposing a delivery date for the technology, the Offeror should consider market maturity, reliability and programming model, among other criteria. |
| §4.5 Early Access System | Minimum configuration due 6-months prior to initial system delivery. Other configurations will be negotiated as part of specific technical options that are exercised. |
| **Upgrades or enhancements to the base systems or services** | |
| §4.6 Software Tools and Programming Environment | General market availability assumed. Tools that are part of Technical Options that require development may be deployed with corresponding technology. |
| §4.7 Upgrades and Expansions to the NWSC-2 PFS System | Upgrade to performance and/or capacity of PFS. Specific quantities and delivery date will be negotiated at award. Subsequent quantities are possible. |
| §4.8 Maintenance, Support, and Technical Services | Enhanced services beyond the base §3.12 services. Offeror should propose incremental enhancements, if applicable, as described. |
| §4.9 Upgrades and Expansions to the NWSC-2 HPC System | Based on the delivered system configurations. Options could be negotiated at any time during life of the contract. |
| §4.10 AMPS System | Based on the delivered NWSC-2 system technologies and software, a real-time operational forecast system for the Antarctic Mesoscale Prediction System |

The Offeror shall provide all relevant technical, business and price information (as described in the NWSC-2 RFP) for all options listed below. Options shall be priced separately and shall include delivery, installation, warranty, maintenance, and support for the life of the system. Pricing shall be firm fixed prices. The technical, business and price information for vendor-proposed options will be evaluated during the selection process and represent a critical element of the overall Offeror proposal.

Beyond the options described below, the Offeror is encouraged to propose additional areas of collaboration that they feel provide substantial value to the NWSC-2 system and its user community.

## Many-core Compute Partition

An emerging class of processors, so-called many-core processors, have characteristics that may include: a slightly higher number of cores than is typical for contemporary microprocessors (i.e. a few dozen processing elements); being self-hosted, i.e., running a native instance of the OS; binary compatibility with traditional processors; allowing developers to maintain a single application source tree; and an additional level of high-speed local memory.  
  
In the NWSC-2 time frame, CISL does not expect that most of its production workload, which is composed primarily of climate and weather models, will be able to fully leverage these emerging architectures. However significant efforts are underway by CISL and the model development teams to transition the models to, and evaluate model performance on, these future processors. Therefore, while the production NWSC-2 system should support the anticipated production workload, CISL anticipates investing in sufficient capabilities through technical options to expand the system in support of these model transition and evaluation efforts.

* + 1. The Offeror shall describe the node-level architecture for the many-core partition. It is highly desirable that the many-core partition be based on the same node-level architecture as the main compute nodes. If the Offeror determines that the proposed compute node architecture is not consistent with the roadmap of the many-core processor, the Offeror shall describe the alternative architecture(s).
    2. The Offeror shall describe how the many-core components will be integrated into the system. It is highly desirable that these components be integrated on the same high-speed network as the main compute resources and have equal access to all other system resources, e.g., file systems and storage. If this is not possible, Offeror should explain why and what the performance implications are of having them on an independent network.
    3. The Offeror shall describe how the many-core partition affects the scalability of the main system partition and impact the high-speed interconnect of the cluster. The Offeror should address the scalability of the different partitions as well as the combined scalability of the integrated system.
    4. Offeror shall provide options for the following sizes (in terms of node count) of a many-core partition. Sizes are approximate, and the Offeror may propose architecturally appropriate quantities.

#### ~5% of the total system is comprised of many-core nodes.

#### ~10% of the total system is comprised of many-core nodes.

#### ~20% of the total system is comprised of many-core nodes.

* + 1. All of the application performance specifications in §3.6 shall apply to the many-core partition. If the many-core partition is delivered at a date later than the primary NWSC-2 production system, its contribution to the Sustained Performance (NBS, §3.6.3) shall be applied at the time of acceptance of the many-core partition.

## General Purpose GPU Compute Partition

General Purpose Graphics Processing Units (GPGPU) have characteristics that may include: a much higher number of processing elements than is typical for contemporary microprocessors (i.e. thousands processing elements); not being self-hosted, i.e., relying on a host processor to receive instructions and data on which to operate; and typically having high-speed local memory. CISL already provides support for a limited number of applications that make use of GPGPUs.

The Offeror shall provide options for a GPGPU compute partition as follows:

* + 1. The Offeror shall describe the node-level architecture for each GPGPU partition. It is highly desirable that the partition be based on the same node-level architecture as the main compute nodes. If the Offeror determines that the proposed compute node architecture is not consistent with the roadmaps of the GPGPU, the Offeror shall describe the alternative architecture(s).
    2. The Offeror shall describe how the GPGPU components will be integrated into the system. It is highly desirable that these components be integrated on the same high-speed network as the main compute resources and have equal access to all other system resources, e.g., file systems and storage. If this is not possible, Offeror should explain why and what the performance implications are of having them on an independent network.
    3. The Offeror shall describe how the GPGPU partition affects the scalability of the main system partition and impacts the high-speed interconnect of the cluster. The Offeror should address the scalability of the different partitions as well as the combined scalability the integrated system.
    4. Offeror shall provide options for the following sizes (in terms of node count) a of GPGPU partition. Sizes are approximate, and the Offeror may propose architecturally appropriate quantities.

#### ~5% of the total system is comprised of GPGPU nodes

#### ~10% of the total system is comprised of GPGPU nodes

#### ~20% of the total system is comprised of GPGPU nodes

* + 1. All of the application performance specifications in §3.6 shall apply to the GPGPU partition. The SHOC application in Table 3 identified as GPGPU-specific is required for the GPGPU itself, while the other benchmarks are required for the host processor within the GPGPU partition. To the extent that the GPGPU nodes satisfy the criteria in §4.2.1, their compute performance may be counted towards the Sustained Performance requirement given in §3.6.3.

## Data Analysis, Visualization and Post-processing

The Data Analysis, Visualization and Post-Processing (DAVP) option provides for a portion of the system capable of supporting post-processing, visualization and analysis workloads. Data analysis and visualization workloads require substantial compute, memory, GPU resources, and especially good bandwidth into the parallel file system.

* + 1. The Offeror shall describe how or confirm that the visualization resources shall support the following packages:
* VAPOR - https://www.vapor.ucar.edu/
* VisIt - https://wci.llnl.gov/simulation/computer-codes/visit/
* Kitware’s ParaView (open source).
* VirtualGL
* NCL - http://www.ncl.ucar.edu/
* Matlab
* IDL

These packages use some or all of the following system software capabilities: full support of sockets, dynamic linked libraries, POSIX threads, Python scripting, MPI, OpenGL, NumPy, NetCDF, HDF5 and MPI I/O. NCAR will assume responsibility for porting of NCAR packages and separately provide any required licensing for third-party visualization applications.

* + 1. The Offeror shall describe the node-level architecture for each DAVP partition. It is highly desirable that the visualization/data analysis partition(s) have a similar node-level architecture (e.g., CPU, motherboard) as the main compute nodes. If the Offeror determines that the proposed compute node architecture is not consistent with the roadmaps of all the visualization packages listed above, the Offeror shall propose an alternative architecture that is consistent.
    2. The visualization/data analysis resource shall be integrated into the system, sharing the same high-speed network as the main compute resources and have equal access to all other system resources, e.g., file systems and storage.
    3. The Offeror shall describe and provide an option for a visualization/data analysis capability that consists of 8 nodes as follows:

#### Four large-memory nodes of at least 1 TB of memory per node.

#### Four GPU visualization nodes with a minimum of 512 GB of memory per node. If the large-memory node architecture in 4.3.4.1 can be used for these, Offeror should provide an option that combines the large memory nodes with the GPU nodes, and thus results in eight, 1 TB memory nodes with GPUs.

## Innovative Storage and Memory Technologies

NCAR is interested in innovative storage and memory technologies that have the potential to dramatically improve the performance of NCAR’s user applications. These may include, for example: stacked memory, non-volatile memory, burst buffers, hybrid SSD/HDD storage systems, processor-in-memory, or software enhancements that leverage these new technologies. These innovative technologies may be deployed subsequent to the initial system deployment, but no later than one year following Acceptance of the NWSC-2 production systems. It is understood that long-term projections of the performance of these technologies is difficult given the level of development work that may exist between now and the expected deployment of them. Therefore, NCAR will negotiate these acceptance criteria at the time a decision is made to exercise these options. The Offeror shall propose separately priced options for innovative storage and memory technologies as follows:

* + 1. Description

The Offeror shall fully describe the technology, including all relevant architectural details, how it will be integrated into the production HPC system, many-core, or GPGPU partitions, and the expected performance characteristics, software programming model, and reliability. Where non-volatile memory (NVM) technologies are proposed, Offeror shall address endurance or any other attribute where performance may degrade or substantially change with time or variations in I/O patterns.

* + 1. Cost/benefit

Offeror shall provide a cost/benefit analysis of deploying the proposed technology. For example, if a burst buffer is proposed, Offeror shall provide the rationale for the amount, and to what extent it will reduce the amount of storage and/or bandwidth required in other parts of the system.

* + 1. Market Maturity Information

Offeror shall provide information that can be used to assess the state of the technology development and adoption at the expected time of contract award, and at the time it is proposed to be deployed in the NWSC-2 system. This can include, for example: current or planned installations of the technology; committed market release dates; data or other results that indicate the current state of development and the risks that remain in integrating it into a production HPC system.

* + 1. Performance of Storage and Memory Technology

Offeror shall estimate the expected performance improvements to NCAR’s application workload and propose a set of benchmarks that could be used to demonstrate these performance benefits.

## Early Access System

To allow for early and/or accelerated development of applications, development of functionality required as a part of the subcontract’s statement of work, or for the correctness validation of the NCAR Benchmark Suite results (§3.6.4), the Offeror shall propose options for early access development systems (EAS). These can be systems either delivered to NCAR or provided via dedicated access at the Offeror’s site. The systems can be in support of the baseline system specifications or any proposed options. EAS system(s) shall be delivered or made available six months prior to the delivery of any hardware or software that is planned for production.

* + 1. The Offeror shall propose Early Access System(s). The primary purpose is to expose the application to the same programming, runtime, and system software environment as will be found on the final system, and to provide access to hardware or software that may be proposed as part of the base system and/or technical options. Except for tests described in §3.6.4, it is acceptable for the early access system to not use the final processor, node, or high-speed interconnect architectures. However, the programming, runtime, and operating system environment must be sufficiently similar such that a port to the final system is trivial. The early access system shall contain similar functionality of the final system, including file systems, but scaled down to the appropriate configuration.
    2. The specific size of the EAS shall be negotiated with Offeror and will be based on details of the specific NWSC-2 solution and options proposed.

## Software Tools and Programming Environment

The Offeror shall provide information about optional software tools and programming environment software in the following product categories. The Offeror may respond either by simply identifying compatible or available third-party offerings (for CISL to evaluate separately) or by describing and providing full technical options for software that CISL may elect to execute as part of the NWSC-2 procurement.

Descriptions of the software tools should reflect that the primary programming model used by application scientists running in production on the NWSC-2 systems will be the Message Passing Interface (MPI) with hybrid MPI/OpenMP parallelism used by key applications. However, significant development activity for NCAR’s climate and weather models is expected to target the transition to programming models on the many-core partitions proposed by the Offeror in §4.1.

* + 1. The Offeror shall identify or describe any compatible alternate vendor-supported job-schedulers/resource managers, consistent with the details requested in §3.2. The Offeror shall also identify and describe any advanced or distinguishing features of optional job schedulers/resource managers, such as: checkpoint-restart, job migration, concurrent scheduling of a single application across more than one resource (e.g. production and many-core or GPGPU partitions), and/or job interrupt/resume.
    2. The Offeror shall identify or describe additional available or supported MPI 3.0 implementations, consistent with the details requested in §3.3.
    3. It is highly desirable for the production NWSC-2 system to support multiple compilation environments. The Offeror shall identify or describe all compilation suites and languages, including version numbers and integrated components (such as debuggers or profiling tools), that are compatible with the proposed system.
    4. The Offeror shall identify or describe the interactive debuggers or debugging suites (available separately from a previously described compiler or tool suite) with a graphical user interface that are compatible with the proposed system and provide a single-point of control for debugging applications using all granularities of parallelism and programming environments provided by the system.
    5. The Offeror shall identify or describe the available tools for the proposed system that provide detailed performance analysis and profiling of user applications (available separately from a previously described compiler or tool suite). The tools shall support all granularities of parallelism and the programming environment of the system. The Offeror shall describe all tools supported and any limitations, e.g. limits on scalability.
    6. The Offeror shall identify or describe the tools available for the proposed system (separate from a previously described compiler or tool suite) that provide event-tracing capabilities and any limitations. Events of interest include Message-Passing Event Tracing, I/O Event Tracing, Floating Point Exception Tracing, and Message-Passing Profiling.
    7. The Offeror shall identify or describe the stack-tracing tools available for the proposed system (separate from a previously described compiler or tool suite). The tools identified should include a source-level stack trace back, including an API that allows a running process or thread to query its current stack trace.

## Upgrades and Expansions to the NWSC-2 PFS System

The initial PFS storage deployment specified in Table 1 is considered sufficient in capacity and performance to accommodate the major application use cases during the first two years of production. However, it is likely thereafter that additional storage capacity and performance will be required and thus, Offeror shall provide options for upgrading storage in accordance with the design described in §3.4.

* + 1. Options will be described as scalable units (SU), which UCAR may purchase in variable quantities. Any additional infrastructure required to support the SU will either be included in the SU or identified as additional requirements to support the SU.
    2. Offeror shall describe and provide options that increase the capacity of the initial storage deployment by 50% and 100% without necessarily increasing performance. Describe any impact on performance of adding capacity only.
    3. Offeror shall describe and provide options that increase both bandwidth performance and capacity of the initial storage deployment by 50% and 100% over the initial deployment; i.e., one option where performance and capacity increases by 50% each; and one option where performance and capacity increases by 100% each.

## Maintenance, Support, and Technical Services

The Offeror shall propose separately priced maintenance and support options with the following enhanced features beyond the basic set proposed in response to §3.12:

* + 1. Maintenance and Support Options

The Offeror shall, at a minimum, propose the following maintenance and support options. The Offeror is encouraged to propose additional options considered to be advantageous to UCAR, such as different maintenance levels for the test and production systems, or a combination of options covering specific components or optional partitions of the NWSC-2 systems.

#### Option 1 – 24x7

The Offeror shall price Option 1 as full Offeror-supplied hardware maintenance for Offeror-provided hardware components. The principal period of maintenance (PPM) shall be 24 hours by 7 days a week with a four hour response to any request for service.

#### Option 2 – 9x5

The Offeror shall price Option 2 as full Offeror-supplied hardware maintenance for Offeror-provided hardware components. The PPM shall be 9 hours by 5 days a week (exclusive of holidays observed by NCAR).

#### Option 3 – Mixed

The Offeror shall price Option 3 as full Offeror-supplied hardware maintenance for Offeror-provided hardware components. The PPM for critical components shall be 24 hours by 7 days a week with a four hour response and the PPM for non-critical components shall be 9 hours by 5 days a week (exclusive of holidays observed by NCAR).

* + 1. Software and Firmware Update Service

The Offeror shall provide an update service for software and firmware delivered with the NWSC-2 system. Offeror personnel, supervised by UCAR staff, shall perform this update service. This shall include new releases of software/firmware and software/firmware patches as required.

* + 1. Extended Call Service

If not already included by the Offeror in the support services proposed in response to §3.12, the Offeror shall propose an option which would make key technical personnel with knowledge of the proposed equipment and software available to NCAR. These personnel shall be available for consultation by telephone and email with NCAR personnel. In the case of degraded system, file system or application performance, the Offeror’s services shall be made readily available to develop strategies for improving performance, e.g., patches and workarounds.

* + 1. Extended Services, Warranty, Maintenance, and Support

The Offeror shall propose separately priced options to extend all technical services, warranty, and Maintenance and Support for each of two additional years beyond the end of the initial 4-year maintenance period.

## Upgrades and Expansions to the NWSC-2 HPC System

It is possible that NCAR will have future requirements for HPC system upgrades and/or additional quantities based on the production system configuration(s) proposed for NWSC-2. To address these potential requirements, the Offeror shall propose separately priced options for HPC system upgrades and expansions as indicated below. Pricing should include all hardware, software and maintenance costs associated with the upgrade, e.g., adding compute nodes requires additional switching infrastructure. The Offeror shall address any technical challenges foreseen with regard to scaling and any other production issues.

* + 1. Upgrade/expand the NWSC-2 HPC configuration by the following fractions of the proposed systems as measured by the NCAR Benchmarking Suite. Sizes are approximate, and the Offeror may propose architecturally appropriate quantities.

#### ~10%

#### ~25%

#### ~50%

## AMPS System

The AMPS System option shall provide a stand-alone replacement for the current system, named Erebus (<https://www2.cisl.ucar.edu/resources/erebus>), which supports the Antarctic Mesoscale Prediction system (AMPS, <http://www2.mmm.ucar.edu/rt/amps/>). The Offeror’s proposal shall describe the proposed AMPS System option.

* + 1. The AMPS System shall be comprised of the identical technology as the proposed NWSC-2 production systems: node and high-performance interconnect, disk storage, management and administrative infrastructure. It shall have a software stack identical to that of the production HPC and PFS resources. Thus, the AMPS System and its user environment shall be compatible with the NWSC-2 production and test resources so that they may be used for AMPS runs, or as a backup when, for example, the AMPS System is unavailable due to system upgrades or maintenance.
    2. The AMPS system shall be comprised of an HPC and PFS resource with sizes identical to the test system described in §3.10 except that it must have a minimum of 100 batch computation nodes.
    3. The AMPS system shall not share any internal or external communications paths, peripherals, storage, power supplies or subsystems required for normal operation with either the NWSC-2 production or test systems.
    4. The AMPS HPC and PFS systems shall be subject to the same lifetime reliability specifications as the NWSC-2 production systems and shall be supported and maintained for 4 years at the same levels as that proposed for the NWSC-2 production resources. The Offeror shall provide options for 5th and 6th years of AMPS System support and maintenance.
    5. The AMPS HPC and PFS systems delivery schedule shall be determined during subcontract negotiations.
    6. The AMPS HPC and PFS systems shall be subject to acceptance testing that will include the production of AMPS operational forecasts. WRF is the primary application used for these forecasts.

For more information on the Antarctic Mesoscale Prediction System, its science and support of the United States Antarctic Program, see Powers, J.G., K.W. Manning, D.H. Bromwich, J.J. Cassano, and A.M. Cayette, 2012: *A decade of Antarctic science support through AMPS*. Bulletin of the American Meteorological Society, 93, 1699-1712. DOI: <http://dx.doi.org/10.1175/BAMS-D-11-00186.1>

# Delivery and Acceptance Specifications

Testing of the system shall proceed in three steps: pre-delivery, post-delivery and acceptance. Each step is intended to validate the system and support subsequent activities. A sample acceptance test plan is include in the Sample Subcontract, Schedule F. A detailed acceptance test plan will be developed during subcontract negotiations.

Some limitations may exist, post-delivery, for vendor access to the system (both on-site and remote). All access to UCAR facilities and systems shall be coordinated through the NWSC facility’s security staff and CISL’s IT staff, respectively.

## Pre-delivery Testing

NCAR and vendor staff shall perform pre-delivery testing at the factory on the hardware to be delivered. Any limitations for performing the pre-delivery testing need to be identified including scale and licensing limitations. During pre-delivery testing, the successful Offeror shall:

* Demonstrate RAS capabilities and robustness, using simple fault injection techniques such as disconnecting cables, powering down subsystems, or installing known bad parts.
* Demonstrate functional capabilities on each segment of the system built, including the capability to build applications, schedule jobs, and run them using the customer-provided testing framework. The root cause of any application failure must be identified.
* Provide a file system sufficiently provisioned to support the suite of acceptance tests.
* Provide onsite and remote access for NCAR staff to monitor testing and analyze results.
* Instill confidence in the ability to conform to the statement of work.

## Site Integration and Post-delivery Testing

NCAR and vendor staff shall perform site integration and post-delivery testing on the fully delivered system.

* During post-delivery testing, the pre-delivery tests shall be run on the full system installation.
* Where applicable, tests shall be run at full scale.

## Acceptance Testing

NCAR and vendor staff shall perform onsite acceptance testing on the fully installed system.

* The Offeror shall provide any suggested modifications to the sample acceptance test plan provided in the Subcontract Schedule F.
* The systems shall be subject to functionality, resilience, performance and availability testing and meet the criteria specified in the Subcontract Schedule F.
* The Subcontractor shall demonstrate that the delivered systems conform to the subcontract’s Statement of Work.

# Risk Management and Project Management

The Offeror’s proposal shall provide a project management plan, based upon the sample project management plan included in the RFP’s Sample Subcontract, Schedule G, that addresses:

* + 1. Provide a risk management strategy for the proposed system in case of technology problems or scheduling delays that affect availability or achievement of performance targets in the proposed timeframe. Describe the impact of substitute technologies on the overall architecture and performance of the system as described in §2.1.1. In particular, the Offeror shall address the technology areas listed below.
* Processor
* Memory
* High-speed interconnect, both hardware and software
* Storage subsystem, both hardware and software
* Technical Options
  + 1. Identify any other high-risk areas and accompanying mitigation strategies for the proposed system.
    2. Provide a clear plan for effectively responding to software and hardware defects and system outages and document how problems or defects will be escalated.
    3. Discuss additional capabilities including the Offeror’s:
* Ability to produce the proposed system and maintain it for the life of the platform
* Ability to achieve the target quality assurance, reliability and availability goals
* In-house testing and problem diagnosis capability, including hardware resources at appropriate scale

# Documentation and Training

The Offeror shall provide documentation and training for the proposed solution to the operators, system administrators, and users of the NWSC-2 system in order to effectively operate, configure, monitor and use the platform. NCAR may, at its option, make audio and video recordings of presentations from Offeror’s speakers at public events targeted at the NCAR user community. Offeror grants NCAR use and distribution rights of vendor-provided documentation, session materials and recorded media to be shared with NCAR staff and all authorized users and support staff for NWSC-2.

## Documentation

* + 1. The Offeror shall provide system level documentation for each delivered system describing the configuration, interconnect topology, labeling schema, hardware layout, etc., of the system as deployed before the commencement of system acceptance testing. For storage hardware, such as RAIDs, the Offeror shall additionally provide details of the logical and physical interconnections within the storage hardware to facilitate failure analysis and fault isolation by CISL.
    2. The Offeror shall supply and support system-level documentation necessary for operation and maintenance of the system for all components of the system. Documentation shall include available monitoring APIs (e.g. accessible by SNMP or other interfaces) available for use by CISL for enhanced monitoring capabilities.
    3. The Offeror shall supply user-level documentation that can be shared with all authorized users of NWSC-2 for all user-accessible software tools and programming environment components. The Offeror shall describe any limitations on the distribution of these materials.
    4. The Offeror shall describe how system-level and user-level documentation will be distributed and updated electronically (e.g., as electronic documents for inclusion on the CISL web site, or as electronic documents accessed via the Offeror’s support site).
    5. All documentation shall be updated in a timely manner. Changes to the system shall be accompanied by relevant documentation updates. Documentation of changes and fixes may be distributed as release notes. Full reference manuals may be updated later, but should be no older than two minor releases behind the installed version.

## Training

* + 1. The Offeror shall provide the following types of training at facilities specified by UCAR.

|  |  |  |
| --- | --- | --- |
| **Class Type** | **# Classes annually** | |
|  | 1st Year | Subsequent Years |
| Customer-replaceable hardware (FRU) maintenance and service | 3 | 1 |
| System Operations and Advanced Administration | 2 | 1 |
| Application Programming and Runtime Optimization | 4 | 1 |

* + 1. The Offeror shall describe how all proposed training relevant to the systems will be delivered (offsite classroom training, onsite training, online training, etc.).

# References

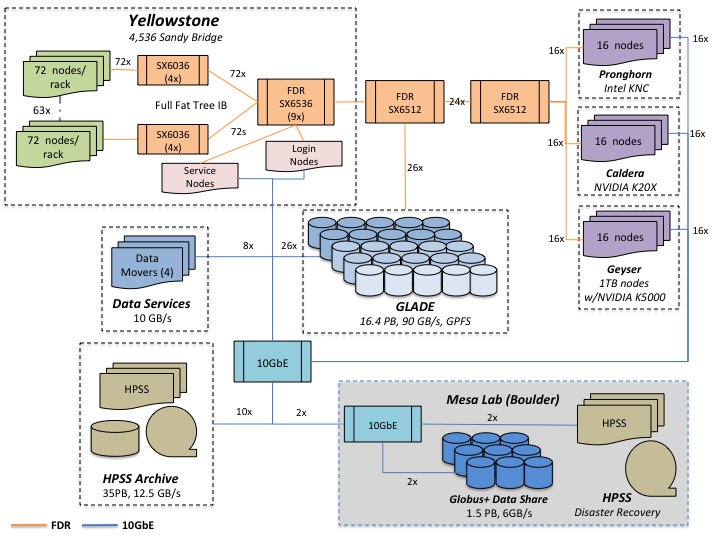
[1] https://www2.cisl.ucar.edu/NWSC-2

[2] https://www2.cisl.ucar.edu/resources/yellowstone

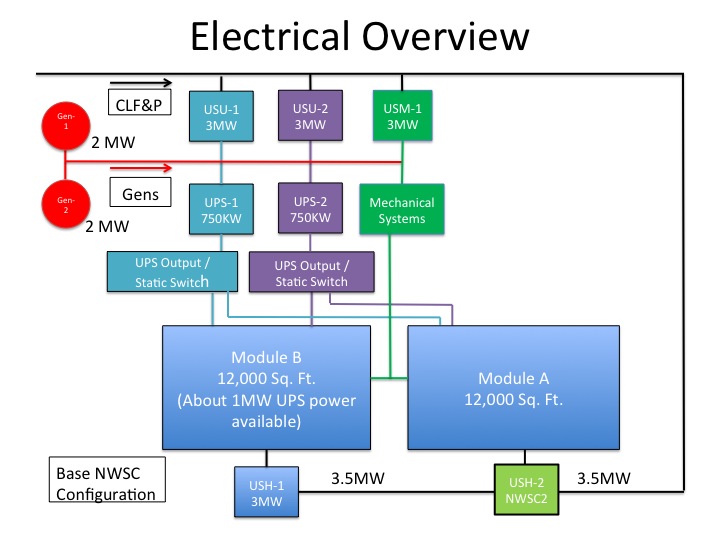
# Facilities Interfaces

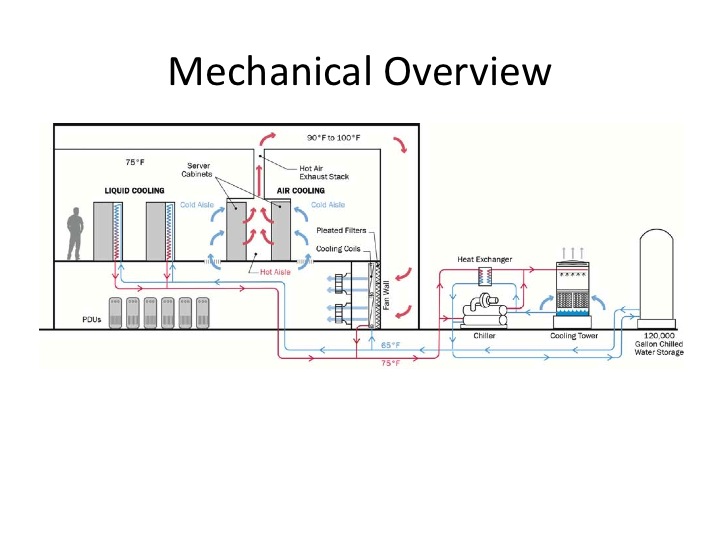
The following figures provide notional diagrams of the existing NCAR HPC/GLADE and NWSC facility environment.

**Figure 1: Current NCAR HPC Environment**

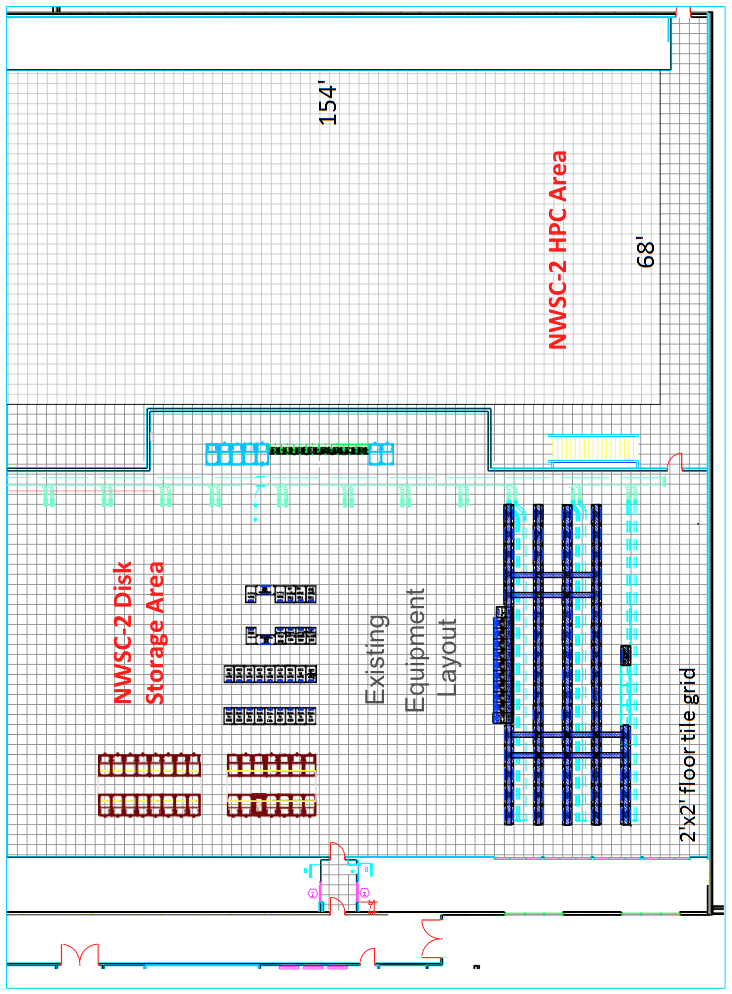


**Figure 2: Current NWSC Module B Electrical and Mechanical Environment**





**Figure 3. NWSC Modules A and B floor plan showing existing equipment and proposed location of NWSC-2 systems.**



1. The UCAR/IBM GPFS Multi-System SOW prices GPFS based upon the number of servers; clients are free. A GPFS server is any node which performs GPFS management functions such as cluster configuration manager, quorum or manager nodes, or NSD servers; additionally, nodes providing NFS, FTP, CIFS, or HTTP are considered GPFS servers. Client nodes are not allowed to share GPFS data through any application, service, protocol or method. [↑](#footnote-ref-1)