# *GAEA-C5*

# High Performance Computing System Supporting Critical Numerical Climate and Weather Prediction

# Statement of Work

November 19, 2020

**Attachment 1, RFP No. 6400016493**

## Change Control

| Document Version | Date | Affected Section(s) | Description |
|---|---|---|---|
| DRAFT (pre-release) | 09/11/2020 | All | Draft (pre-release) for comment. |
| 1.0.2 | 11/19/2020 | All | Initial Release |

# Table of Contents

1. Introduction ....1
2. Program Overview ....2
3. High-Level System Design ....3
   3.1 Key Design Elements ....4
   3.2 C5 High Level System Description (MR) ....6
   3.3 C5 High Level Software Model (MR) ....7
   3.4 C5 High Level Performance Objective (MR) ....7
   3.5 C5 High Level Project Management (MR) ....7
4. Benchmarks ....8
   4.1 Benchmark Availability ....8
   4.2 Benchmarking Procedures ....9
   4.3 Volume of Work Metric ....11
   4.4 Benchmark Runtimes ....12
   4.5 Volume of Work Results ....13
   4.6 Extrapolations ....13
   4.7 Scalability ....14
   4.8 Accuracy ....14
   4.9 Run-Time Variability ....14
5. C5 Compute Partition ....14
   5.1 C5 System Performance (TR-1) ....14
   5.2 C5 Memory Configuration (TR-1) ....15
   5.3 System Resilience (TR-1) ....15
   5.4 Early Access to C5 Hardware Technology (TR-1) ....15
   5.5 Early Access to C5 Software Technology (TR-1) ....15
   5.6 C5 Hardware Options ....15
   5.7 Security Controls (TR-1) ....16
   5.8 Options for Mid-Life Upgrades (TO-1) ....16
   5.9 Compute Partition Hardware Requirements ....16
   5.10 Compute Node Operation System (CNOS) Execution Model (TR-1) ....17
   5.11 Runtime Variability ....17
6. Input/Output Subsystem (IOS) (TR-1) ....17
   6.1 High-Level Requirements for Communicating with the C5 I/O Subsystem (IOS) ....18
7. High Performance Interconnect (TR-1) ....19
8. Base Operating System, Middleware and System Resource Management ....19
   8.1 Base Operating System Requirements (TR-1) ....19
   8.2 Distributed Computing Middleware ....21
   8.3 System Resource Management (SRM) (TR-1) ....21
9. Front-End Environment ....22
   9.1 Front-End Node (FEN) Hardware Requirements ....22
   9.2 Front-End Environment Software Requirements ....23
10. System Management and RAS Infrastructure ....26
   10.1 System Availability ....26
   10.2 Robust System Management Facility (TR-1) ....26
   10.3 Reliability, Availability and Serviceability (TR-1) ....28
   10.4 Company-provided Telemetry Database Support (TR-1) ....29
11. Local Area Networks and Services ....30
   11.1 Network Infrastructure ....30
   11.2 Company-Provided Services ....30

12. Maintenance and Support....30
12.1 Hardware Maintenance (TR-1)....30
12.2 Software Support (TR-1)....31
12.3 Problem Escalation (TR-1)....31
12.4 On-site Analyst Support (TR-1)....31
13. C5 Facilities Requirements....31
13.1 Laboratory Facilities Overview (TR-1)....32
13.2 Power & Cooling Requirements (TR-1)....36
13.3 Floor Space Requirements (TR-1)....38
13.4 Cable Management Requirements (TR-1)....38
13.5 Physical Access Requirements (TR-1)....38
13.6 Safety Requirements (TR-1)....38
13.7 Safety and Power Standards (TR-1)....38
13.8 System Installation and Integration (TR-1)....39
13.9 System Decommissioning (TR-1)....39
14. Project Management (TR-1)....39
14.1 Key Planning Deliverables....39
14.2 Project Meetings....40
14.3 Key Build Phase Milestone Dates (TR-1)....40
14.4 Key Elements of the Plan of Record....40
14.5 Key Elements of Factory Test Plan....41
14.6 Key Elements of Full-Term Hardware Development Plan....41
14.7 Key Elements of Site Preparation Plan....41
14.8 Key Elements of Installation Process Plan....42
14.9 Key Elements of Maintenance and Support Plan....42
15. Acceptance Requirements (TR-1)....42
Appendix A Technical Volume Instructions....1
General Instructions....1
1. Introduction....1
2. Program Overview....1
3. High Level System Overview....1
3.1 Key Design Elements....1
3.2 C5 High Level System Description....1
3.3 C5 High Level Software Model....1
3.4 C5 High Level Performance Objective....1
3.5 C5 High Level Project Management....2
4. Benchmarks....2
5. C5 Compute Partition....2
5.1 C5 System Performance....2
5.2 C5 Memory Configuration....2
5.3 System Resilience....2
5.4 Early Access to C5 Hardware Technology....2
5.5 Early Access to C5 Software Technology....2
5.6 C5 Hardware Options....2
5.7 Security Controls....2
5.8 Options for Mid-Life Upgrades....3
5.9 Compute Partition Hardware Requirements....3
5.10 Compute Node Operation System Execution Model....3
5.11 Runtime Variability....3
6. Input/Output System....3

7. High Performance Interconnect ....3
8. Base Operating System, Middleware and System Resource Management ....3
9. Front-End Environment ....4
10. System Management and RAS Infrastructure ....4
11. Local Area Networks and Fabrics ....4
12. Maintenance and Support ....4
13. C5 Facilities Requirements ....4
14. Project Management ....5
15. Acceptance Requirements ....5

Appendix B Glossary ....1
Hardware ....1
Software ....2

## List of Figures

Figure 3-1. Notional C5 Design ....3
Figure 13-1. Physical footprint available to the C5 system in DC K200 ....32

## List of Tables

Table 2-1 - HPC Systems Acquired Through The Strategic Partnership Between ORNL and NOAA ....2
Table 3-1 Existing Company-provided Storage/File System(s) ....5
Table 4-1. Benchmark Baselines and Weighting ....11
Table 4-2. Volume of Work Template ....13
Table 4-3. CM4, ESM4, SHiELD, SPEAR, UFS/GFS Scaling Results Template ....14
Table 13-1. Characteristics of the MTW Supply Serving C5 ....33
Table 13-2: Cooling Equipment Monitoring Interface ....34
Table 14-1: Project Meetings ....40

## Requirements Definitions

Specific sections of these technical requirements have priority designations, which are defined as follows:

- Mandatory Requirements designated as (MR)

Mandatory Requirements (designated MR) in this Statement of Work (SOW) are performance features that are essential to the Laboratory's requirements, and an Offeror must satisfactorily propose all Mandatory Requirements in order to have its proposal considered responsive.

- Mandatory Option Requirements designated as (MO)

Mandatory Option Requirements (designated MO) in this SOW are features, components, performance characteristics, or upgrades whose availability as options to the Laboratory are mandatory, and an Offeror must satisfactorily propose all Mandatory Option Requirements in order to have its proposal considered responsive. The Laboratory may or may not elect to include such options in any resulting subcontract. Each MO shall appear as a separately identifiable item in Offeror's proposal.

- Target Requirements designated as (TR-1 or TR-2).

Target Requirements (designated TR-1 or TR-2), identified throughout this SOW, are features, components, performance characteristics, or other properties that are important to the Laboratory, but that will not result in a nonresponsive determination if omitted from a proposal. Target Requirements are prioritized by dash number. TR-1 is more desirable than TR-2. Target Requirements add value to a proposal.

- Technical Option Requirements designated as (TO-1 or TO-2)

Technical Option Requirements (designated TO-1 or TO-2) in this SOW are features, components, performance characteristics, upgrades, or other properties that are important to the Laboratory, but that will not result in a nonresponsive determination if omitted from a proposal. Technical Options are expected to add value to a proposal. Technical Options are prioritized by dash number. TO-1 is more desirable than TO-2. Technical Options will be considered as part of the proposal evaluation process but the Laboratory may or may not elect to include Technical Options in any resulting subcontract. Each TO should appear as a separately identifiable item in Offeror's proposal.

- Information for Offeror (I)

Information for the Offeror, not prefaced with (MR), (MO), (TR-1), (TR-2), (TO-1), or (TO-2) in this SOW, describes the existing operating environment that supports the Strategic Partnership between Oak Ridge National Laboratory and NOAA; the anticipated operating environment that includes the new HPC resource C5; services provided by Company; each and any demarcation point (demarc) between the Offeror's solution and Company; and other information that is intended to clarify the responsibilities of Offeror and Company in any resulting subcontract. In general, Information for the Offeror does not include this specific designation (I). However, the SOW may designate Information for the Offeror as (I) where some clarity between a priority designation (MR), (MO), (TR-1), (TR-2), (TO-1), or (TO-2) and information is needed or warranted.

# 1. Introduction

This Statement of Work (SOW) describes the technical requirements for a new Department of Energy (DOE) high performance computing (HPC) system. This new HPC resource, the fifth HPC system acquired as part of the Strategic Partnership between Oak Ridge National Laboratory (ORNL) and National Oceanic and Atmospheric Administration (NOAA), and hereafter referenced as *C5*, is expected to dramatically increase the skill, resolution, complexity and the throughput of computer model-based projections of climate variability and change that will further enable sound decision-making by NOAA on issues of national importance in the time period 2021-2026. This SOW describes specific technical requirements related to the hardware and software capabilities of the C5 system as well as specific application requirements.

Oak Ridge National Laboratory, managed by UT-Battelle, LLC, hereafter referred to as Laboratory and/or Company, anticipates the following schedule for the C5 system acquisition:

| | |
|---|---|
| Q1CY 2021 | System Delivery |
| Q2CY 2021 | System Acceptance |
| Q3CY 2021 | Limited Production |
| September 2021 | Full Production |
| September 2026 | Decommissioning |

The Laboratory reserves the right to revise any or all of the dates reflected in the above schedule based upon Laboratory and/or NOAA needs.

## 2. Program Overview

Responding to climate change requires understanding, adaptation, and mitigation to achieve transition to a low carbon society and global sustainability objectives. The Strategic Partnership between Oak Ridge National Laboratory (ORNL) and the National Oceanic and Atmospheric Administration (NOAA) provides research collaboration and tools that help NOAA better understand and predict climate variability and change, as well as produce decision-support tools to facilitate understanding climate change, mitigation strategies, and adaptation options for the Nation.

In 2008, NOAA and the Department of Energy (DOE) first signed a Memorandum of Understanding (MOU) on collaborative research. Through the execution of this agreement, Oak Ridge National Laboratory (ORNL) provides NOAA with advanced high-performance computing for prototyping critical weather and climate applications in support of their mission. This partnership has contributed to the accelerated development and deployment of NOAA's unified modeling suite of four major configurations for weather and subseasonal-to-seasonal forecasting (SHiELD), seasonal-to-multidecadal forecasting (SPEAR), high-resolution-ocean-based climate modeling (CM4), and earth system modeling (ESM4). These models share many major components: the atmospheric models use the FV3 dynamical core; and SPEAR, CM4 and ESM4 use the OM4 ocean with the MOM6 dynamical core and physics, SIS2 for sea ice, and LM4 for land. These configurations will form the basis for further model development and specialized configurations for a diverse range of understanding and prediction objectives.

Since 2009, ORNL has procured and operated, on behalf of NOAA, a series of high performance computing (HPC) systems. The summary descriptions of those systems is shown in Table 2-1. C3 and C4 remain in operation as of this writing. C3 may be retired as part of the transition to production of C5.

**Table 2-1 - HPC Systems Acquired Through The Strategic Partnership Between ORNL and NOAA**

| System | Description, Final Configuration | Service Dates |
|---|---|---|
| C1MS | Cray XT6. 2,576-socket AMD Magny-Cours CPU, SeaStar HSN. Liquid-cooled. 260 TF | June 2010 – Sept 2012 |
| C1 | Cray XT6. 2,624-socket AMD Opteron/Interlagos CPU, Gemini HSN. Liquid-cooled. 386TF. | Sept 2012 – 2016 |
| C2 | Cray XE6. 4,896-socket AMD Opteron/Interlagos CPU, Gemini HSN. Liquid-cooled. 721TF. | Dec 2011 – 2016 |
| C3 | Cray XC40. 3,008-socket Intel Haswell CPU, Aries HSN, Liquid-cooled. 1.77 PF. | Jan 2016 - Dec 2021 (plan of record) |
| C4 | Cray XC40. 5,312-socket Intel Broadwell CPU, Aries HSN, Liquid-cooled. 3.52 PF. | Oct 2017 – present |
| C5 | | Sept 2021 – Oct 2026 (plan of record) |

NOAA regularly upgrades its numerical climate and weather prediction models to increase the accuracy and frequency of the output products. The current NOAA HPC systems require a complimentary refresh that is delivered and accepted no later than the Summer of 2021, with completion of the transition to production in Fall 2021 to meet its compute requirements.

As part of the continued Strategic Partnership between ORNL and NOAA, ORNL will manage the full life cycle (acquisition, installation, operation, and maintenance) of this new HPC capability. The programmatic objective is to install and operate C5 alongside C4 at Oak Ridge National Laboratory.

## 3. High-Level System Design

The overarching requirements for the C5 system are based on its ability to deliver highly available, timely, and accurate numerical climate and weather prediction output in support of the NOAA mission, and further, that this new capability is fully integrated in to the existing operational environment. To meet these requirements, some desired architectural design features are described. The successful Offeror may meet the requirements of these design features, either directly by implementing the architectural design features as described, or by proposing alternative architectural design features and explicitly describing how that solution meets or exceeds the desired feature.

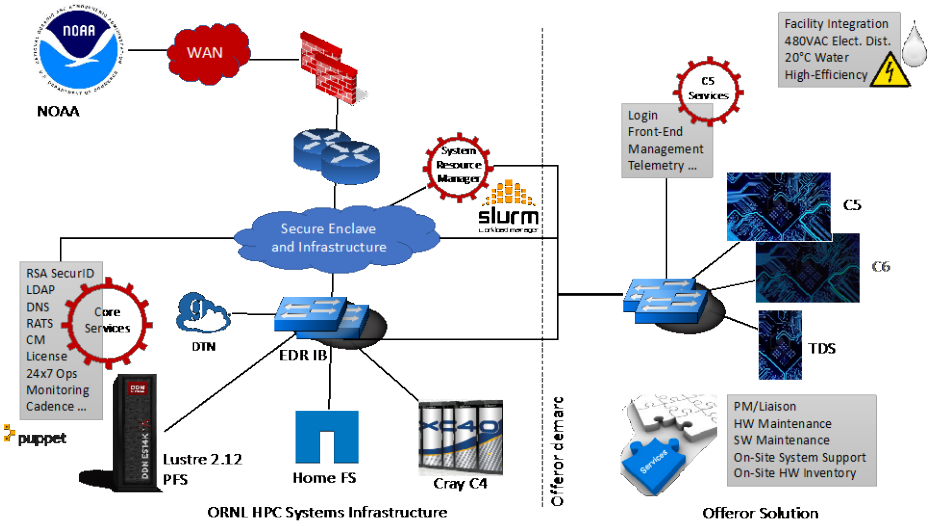



**Figure 3-1. Notional C5 Design**

A notional design is presented in Figure 3-1, describing the existing ORNL HPC Systems Infrastructure that supports the NOAA SPP, and the anticipated, fully integrated approach of the Offeror's solution for C5.

Offeror's solution comprises hardware and software services, including any and all additional login nodes, front end nodes, management servers, telemetry, and supporting network equipment, beyond what currently exists within the ORNL HPC Systems Infrastructure.

- Offeror's solution includes the production compute capability, notionally or potentially represented here as two distinct and independent compute partitions, labeled here as C5 and C6. Whether the total compute capability is proposed as a single C5 system, or as a pair of compute systems is the defended decision of the Offeror.
- Offeror's solution includes a separate test and development system (TDS) that allows independent software and firmware update/regression testing prior to introduction on the production platforms.
- Offeror's solution includes all facility design/integration services that allows the hardware associated with C5 to operate within the existing facility.
- Offeror's solution includes a full suite of maintenance and support services, including on-site system administration that can be integrated with Company's existing system support team.
- Offeror's solution names both a Technical Project Manager and Executive Liaison that will work with Company to ensure that the relationship is a partnership.

Note: it is not required to propose two separate systems, i.e. a C5/C6 configuration. This is suggested due to some of the potential benefits that surround the software upgrade management requirements throughout its lifetime. This diagram is purely notional. Detailed design, including how physical systems may be incorporated in logical abstractions, availability zones, or similar construct(s) is the responsibility of the Offeror to describe.

C5 will be integrated in to an existing Federal Information Protection Standard (FIPS) 199 Moderate (Confidentiality, Integrity) security enclave that is managed by Laboratory. That enclave provides secure and

production-hardened services including RSA authentication, LDAP, DNS, license servers, monitoring capabilities, telemetry and other data and system log capture and analysis, and home directories.

Company provides a 38 PB scratch Lustre namespace that is mounted by all compute partitions. As of the release of this request for proposals, that file system is based on Lustre 2.12 and may be referenced as *F2* throughout this document. Reference § 6.

Company provides a Front End Environment (FEE) for C3 and C4 that includes eight redundant login nodes. Each login node is identical, providing programming environments based on the Intel (19.x) and GNU C, C++ and FORTRAN compilers. All existing login nodes mount F2 and $HOME. These login nodes are specifically configured to support existing C3 and C4 systems. However, Offeror shall make no assumptions regarding the hardware and software that comprise the existing FEE. Reference § 9.

Company provides 16 highly optimized data transfer nodes (DTNs) that serve as the end point for workload components moving between Laboratory and NOAA. These DTNs mount the existing Lustre F2 file system via Enhanced Data Rate (EDR) InfiniBand and expose Globus[1] endpoints via 10 Gigabit Ethernet.

Company provides diverse and hardened 10/40/100 Gigabit Ethernet routing and switching capacity within the enclave. Company also provides a 100Gb/s InfiniBand network fabric that allows the F2 filesystem to be mounted by the compute partitions through LNET routers. These compute partitions include C4 (and potentially C5). Those physical ports in the Layer 2 devices and EDR IB devices that service the existing NOAA systems shall define the demark between the Laboratory and the Offeror equipment. Company will ensure that adequate IB ports are provided to connect elements of C5 to the existing IB fabric. Company will ensure that adequate ports are provided to uplink and connect the C5 Ethernet fabric(s).

## 3.1 **Key Design Elements**

### 3.1.1 Compute Partitions (TR-1)

Two discrete and identical compute partitions are desired. They shall be configured so that they are concurrently active, concurrently scheduled, but can be independently removed from service without affecting the other system. This concept is referenced hereafter as C5/C6.

A smaller Test and Development System (TDS) is desired. This system should have all of the attributes of the production computing capability, i.e. be able to operate as an independent computing resource. Company will primary use this TDS for regression testing and software stack/feature testing. The TDS may also be used by Offeror as part of their maintenance strategy. There is no requirement for the TDS to demonstrate the concurrently active, concurrently scheduled attributes described by C5/C6, i.e., this is expected to be a single compute system.

Company maintains an existing system resource manager (SRM), SchedMD's *Slurm*. Company leverages specific features of this SRM including management of federated clusters, multiple partitions, in-depth accounting, advanced reservations, backfill scheduling, topology optimized resource selection, and resource limits, among others. Slurm is currently highly integrated into the daily user workflow. Offeror is encouraged to integrate their solution with Slurm, although an alternate solution is acceptable. Company expects to continue an existing support contract with SchedMD for Slurm, independent of the contract associated with C5. As such, Offerors need only ensure that their proposed architecture is supported by SchedMD. No contractual elements are necessary. Offerors may propose an alternative SRM, including appropriate description of the licensing and

[1] Globus is a trademark of the University of Chicago.

support costs for that SRM. Offerors that do not propose Slurm are responsible for these licensing and support costs. Do not provide cost data in the Technical Volume but complete this in the Business Volume.

C5 (and potentially C6) and the TDS will be collocated in a Company-supplied facility. It is also assumed that any additional hardware necessary to deliver the front-end/supporting services will also be collocated in the same immediate physical area of the facility. As described in § 13, Company has identified available power, space and cooling that are available to the C5 compute partition(s).

### 3.1.2 Input/Output (I)

Company currently provides a NFS storage appliance for non-scratch data and a separate high-performance single-namespace Lustre filesystem as a scratch space. This is commonly referenced as $HOME. The configurations of the $HOME and F2 filesystems are summarized in Table 3-1.

**Table 3-1 Existing Company-provided Storage/File System(s)**

| System | Description, Final Configuration |
|---|---|
| $HOME | NetApp |
| F2 | DDN SFA14KX block storage system, with 6 controller pairs, 6 Dell r640 OSS nodes per controller pair, 38PB scratch capacity, and can realize sequential read/write performance in excess of 240GB/s. The F2 metadata subsystem is built on a single NetApp EF570 with 4 Dell r640 metadata servers as MDSs, and can realize a sustained IOP rate in excess of 80k. This storage hardware is formatted into a single Lustre 2.12 namespace. |

Sixteen Company-provided Data Transfer Nodes mount both $HOME and the F2 Input / Output System (IOS).

The I/O patterns on F2 are generally metadata intensive and composed of small unaligned reads/writes, which is typical of weather and climate simulation codes. This is interspersed with large sequential I/O activity making for a diverse I/O workload.

### 3.1.3 Test and Development Systems (TR-1)

To support early testing and to reduce operational risk to the Program, Company seeks a Test and Development Compute System (TDS), which architecturally mimics the C5 system at a smaller scale. It may leverage the Offeror's separately provided login, management, telemetry and other supporting services. It shall mount both the existing production F2 IOS and NFS non-scratch file system. The TDS will explicitly be part of the subcontract.

Company may seek additional testbed systems that are architecturally diverse from that which is represented in C5, and provide a hardware and software environment suitable for evaluating the performance of global and regional climate/weather applications on new and emerging technologies. Options for integrating these new and emerging technologies that can be physically incorporated in to the TDS and/or C5 are preferred, but not required. Additional testbed systems will be funded separately.

### 3.1.4 Benchmark Performance (TR-1)

Laboratory will assess the performance of five benchmarks part of the selection process. These benchmarks are highly representative of the current workload on C3 and C4 and the anticipated emerging workload on C4 and C5.

- NOAA's Coupled Model (CM4) is a coupled atmosphere-ocean general circulation model (AOGCM) that includes AM4 (atmosphere), OM4 (ocean), SIS2 (sea ice), and LM4 (land model)

- NOAA's Earth System Model (ESM4) includes AM4 (atmosphere), OM4 (ocean), SIS2 (sea ice), LM4 (land model), COBALTv2 (ocean biogeochemical component) and dust/iron cycling;
- NOAA's System for High-resolution modeling for Earth-to-Local Domains (SHiELD) is a next-generation modeling system for weather and subseasonal-to-seasonal forecasting;
- NOAA's SPEAR is a next-generation modeling system for seasonal to multidecadal prediction and projection;
- NOAA's UFS/GFSv16 prototype is a next-generation modeling system for weather prediction that remains a candidate operational configuration, under test, as part of the next scheduled upgrade of the Global Forecast System used by the NWS.

The problem sizes for each benchmark have been specifically chosen to accommodate benchmarking of smaller, but representative systems. Test cases are distributed across these five benchmarks to ensure that the production workload is adequately and appropriately represented.

Laboratory will assess the performance of each of these benchmarks, applying *weight*, i.e. importance of the application to the production target, to each.

### 3.1.5 C5 Integration Sequence (TR-1)

C5 is being added to an existing ecosystem that includes data center facilities, hardware, software, local- and wide-area networking, cyber security and other supporting infrastructure. That environment will be prepared for receipt of C5 elements prior to the end of Q1CY21. Laboratory anticipates a series of two explicit delivery opportunities as part of the integration of C5 into this existing infrastructure. These Key Build Phase Milestones are:

- Delivery, installation, integration and acceptance of the Front End Nodes, i.e. login, management, telemetry, and other similar items (FENs) and the Test and Development System (TDS);
- Delivery, installation, integration and acceptance of C5. Offerors should provide adequate detail of all components that may comprise C5, especially if a C5/C6 approach is offered.

The FENs are expected to be a prerequisite for the TDS. The availability of the FENs and TDS allows Laboratory to extend the end to end workflow for the existing NOAA user community to a smaller but distinct compute partition. Company acknowledges that delaying the timing for the C5/C6 system(s) may provide greater flexibility to the Offeror for delivering valuable, viable and appropriate product offerings. However, the overall schedule for delivery, acceptance and transition to production in Summer 2021 remains firm.

Offeror may combine these events in a single delivery. Offeror may also choose to defend the separate delivery of C5 and C6 for specific technical or other reasons, subject to the transition to production in Summer 2021.

## 3.2 **C5 High Level System Description (MR)**

The Offeror shall provide a concise description of its proposed system architecture, including all major system components plus any unique features that should be considered. The description shall include:

- An overall system architectural diagram that shows all node types and their quantity, the interconnect(s), including to the existing IOS (via Infiniband) and Layer 2 Ethernet fabric(s) as well as the latencies and bandwidths of data pathways between components. Detail must be sufficient to

ensure that Company can appropriately plan for connections from the existing infrastructure to the Offeror's solution;
- An architectural diagram of each node type showing all elements of the node along with the latencies and bandwidths to move data among node elements.

The response shall identify features of the architectural design that improve, enhance, or contribute to component or system reliability, availability, serviceability, repeatability, reproducibility, reduced variability, reduced jitter, and similar.

The response shall concisely describe the complete system architecture that the Offeror considers be the preferred technical choice. In addition, the response can describe and propose alternative technologies (e.g. consideration of different processor technologies) and the date when a change from the original proposal to an alternative technology would be appropriate. Regardless of the choices for these alternative technologies, the resulting system should meet all C5 high-level system requirements.

The Offeror shall describe how the proposed system fits into their long-term (no less than 36 months after delivery) product roadmap.

## 3.3 C5 High Level Software Model (MR)

The Offeror shall provide a high-level software architecture diagram. This diagram shall show all major software elements. The Offeror shall also describe the expected licensing strategy for each. The C5 software model and resulting requirements shall be described from the perspective of a scalable system that consists of compute nodes (CNs) that number in the range of tens to thousands, and a Front End Environment (FEE) of sufficient capability to provide access, system management, and programming environment to the system.

The response shall identify features of the software design that improve, enhance, or contribute to component or system reliability, availability, serviceability, repeatability, reproducibility, reduced variability, reduced jitter, and similar.

The response shall describe how the system software tracks early signs of system faults; manages power dynamically; collects power and energy statistics; and reports accurate and timely information about the hardware, software, and applications from all components in the system.

The Offeror is only required to supply one OS per node type, but the architecture shall not preclude the booting of different node OS's in support of containerized models.

## 3.4 C5 High Level Performance Objective (MR)

For the purposes of the RFP, Offeror will execute multiple versions and copies of CM4, ESM4, SHiELD, SPEAR and UFS/GFSv16, introduced in § 3.1.4, in parallel. NOAA provides initial run time scripts that are based on the C4 system configuration and that take in to account subtleties in the domain decomposition for each code. Offeror may revise that domain decomposition if they so choose, subject to the rules prescribed in § 4.2.1. Offeror will identify the best combination of compute resources, domain decomposition and runtime for each of the codes, in accordance with the detailed instructions found in § 4.

## 3.5 C5 High Level Project Management (MR)

The Offeror's proposal shall include the following:

- An overview of a collaboration plan and discussion of any requirements that the Offeror has for Laboratory in the management of the project;
- A preliminary Risk Management Plan that describes any aspects or issues that the Offeror considers to be significant risks for the system, including technical milestones, management of secondary subcontractors, and planned or proposed management and mitigations for those risks;
- Discussion of the delivery schedule and how Offeror proposes to manage the system delivery and deployment, e.g., personnel and communications, factory staging, onsite staging, installation, integration, testing, and bring-up. Any schedule risks should be clearly identified, as well as potential methods to mitigate those risks;
- Discussion of Offeror's general approach for software licensing, e.g., range of licenses and criteria for selection from that range of licenses for a particular package;
- Discussion of Offeror's quality assurance and factory test plans.

# 4. Benchmarks

## 4.1 Benchmark Availability

The benchmarks used as part of this Offer are developed, maintained, and licensed by NOAA. To provide appropriate access to the benchmarks for this specific acquisition activity, NOAA has agreed to distribute the materials via secure transfer methods. Company provides and maintains information related to requesting the benchmarks at https://noaa-c5.ornl.gov/. This site will be maintained with updated information throughout the proposal response period, including updates to instructions for build and execution.

The specific elements of the benchmarks used in this acquisition are:

- CM4. Includes AM4.0: atmosphere at approximately ½ -degree resolution with 33 levels and sufficient chemistry to simulate aerosols (including aerosol indirect effect) from precursor emissions; OM4: MOM6-based ocean at 1/8 -degree resolution with 75 levels using hybrid pressure/isopycnal vertical coordinates; SIS2: sea ice with radiative transfer and C-grid dynamics for compatibility with MOM6; and LM4: land model with dynamic vegetation. There is a single CM4 benchmark test case.
- ESM4: Includes AM4.0 atmosphere at approximately 1-degree resolution with 49 levels of comprehensive, interactive chemistry and aerosols (including aerosol indirect effect) from precursor emissions; OM4: MOM6-based ocean at ¼ -degree resolution with 75 levels using hybrid pressure/isopycnal vertical coordinates; SIS2: sea ice with radiative transfer and C-grid dynamics for compatibility with MOM6; LM4.1 land model with a new vegetation dynamics model with explicit treatment of plant age and height structure and soil microbes, with daily fire, crops, pasture, and grazing tiles; COBALTv2 ocean biogeochemical component representing ocean ecological and biogeochemical interactions; and fully interactive dust and iron cycling between land-atmosphere and ocean. There are two ESM4 test cases, "small" and "large".
- SHiELD: Includes the FV3 dynamical core with the non-hydro option, mixed layer ocean model, NOAH land model and GFS Physics w/ GFDL MP modifications from AM2. There is a single SHiELD benchmark test case. SHiELD leverages the inherent hardware threading capability of the currently installed processors.
- SPEAR: Includes AM4.0: atmosphere at 1-degree to 0.5-degree resolution with 33 levels and sufficient chemistry to simulate aerosols from precursor emissions; OM4: MOM6-based ocean at 1-degree resolution with tropical refinement to 0.3-degree and 75 levels using hybrid vertical coordinates, SIS2: sea ice with radiative transfer and C-grid dynamics for compatibility with MOM6; and LM4: land model with dynamic vegetation. There is a single SPEAR benchmark test case.
- UFS/GFSv16: Includes the GFSv16 prototype atmosphere model, the Noah and Noah-MP land models and the WAVEWATCH III wave model. Licensing for some of the components of

UFS/GFSv16 are specific to the owners of the components. More information is available at https://ufscommunity.org. There is a single UFS/GFS benchmark test case.

All five applications are written to the Fortran 2003 standard and use a mixed mode message-passing/threading paradigm. Message passing is achieved by using calls to MPI libraries, mostly via a GCOM interface layer which is provided with the benchmarks. Threading is via OpenMP.

These benchmarks have been successfully executed on existing systems at both NOAA and Laboratory, on multiple architectures and processor technologies.

## 4.2 **Benchmarking Procedures**

Each benchmark includes a brief documentation file, and a tar file. Tar files contain source code and test problems. The documentation files contain instructions for building the codes, the RFP problems to run, and steps to ensure the codes were run correctly. RFP problems are usually a set of command line arguments that specify a problem setup and parameterization, and/or input files.

### 4.2.1 Allowable Modifications for the Benchmarks

The source code and compile scripts downloaded from NOAA may be modified as necessary to compile and to run the benchmarks on the Offeror's system. Other allowable changes to the compile scripts include optimizations obtained from compiler flags that do not require modifications of the source code.

Benchmarks can be optimized as desired by the Offeror. Performance improvements from pragma-style guidance in C, C++, and Fortran source files are preferred. Modifications must be documented and provided back to Company. Further, source code modifications will be provided back to NOAA under the same license agreement as the original source code, including the transfer of these changes to NOAA as intellectual property (IP).

Optimizations are permitted for the benchmark tests, although limited in size and content. Optimizations shall be limited in size to:

- 1,000 lines of changes for CM4.
- 1,000 lines of changes for ESM4.
- 1,000 lines of changes for SHiELD.
- 1,000 lines of changes for SPEAR.
- 1,000 lines of change for UFS/GFSv16.

These limits do not apply to:

- Fixes for correctness or essential porting changes.
- The runtime scripts or compiler option configuration files.
- Changes made to code necessary for optimal running on a specific platform.

These limits *do* apply for the introduction or change of compiler directives, including those for OpenMP.

The number of lines changed is the sum of added, deleted, or modified lines of code.

The following types of source code changes are permitted for the benchmark tests:

- Compiler directives may be inserted into the code to instruct the compiler to perform some function it otherwise would not, such as "ignore vector dependencies", "unroll a DO loop" or "align arrays on different cache lines";
- Open-MP directives may be added together with moderate code changes directly related to an Open-MP implementation;
- The compiler or pre-processor is permitted to inline routines;
- The code comprising the GCOM library may be modified;
- Code that is dominated by communications or I/O may be modified.

In addition to this, limited changes to the source code will be permitted provided the principles below are followed. Laboratory reserves the right to disallow any change where Laboratory considers that change difficult to implement or otherwise not supported by NOAA. For benchmark tests, any changes to source code shall obey the following principles:

- Changes shall comply with the Fortran 2003 and C99 standards.
- Changes shall ensure array bounds *and* conformance are maintained.
- Changes shall not affect numerical results in a meteorologically significant way.
- Changes shall not alter the precision of any variables used.
- Changes shall not affect the generality of the code. This refers to the ability to run in other configurations, at different resolutions, etc without recompilation.
- The NOAA codes may be used on a variety of HPC platforms. Changes shall therefore not adversely affect the performance that is seen on other platforms. Pre-processor definitions may be used to protect small amounts of code for different platforms.
- The modified code shall be reproducible. That is if a job is re-run with no change to input data or Fortran namelists on the same system, the results shall be bit-identical to the first run.
- Each benchmark is designed to give bit-identical results, when using appropriate global summation methods, for different processor numbers and decompositions on the same system. This shall be maintained for all the benchmark configurations. In particular, neither code changes nor optimization flags should affect this property.
- Changes shall not affect the readability or portability of code.

Any changes made shall be documented and easily identifiable. The Offeror shall provide all changes made to the benchmarks to include documenting changed lines of code and the final count of changed lines, including directives. Note that code changes may attract risk factors that affect evaluation of committed performance. These may be based on (but are not limited to) evaluation of the generality, ease of implementation, size and complexity of the code changes proposed. Note that Offeror shall not include performance improvements based on as yet unidentified optimizations.

### 4.2.2 Running the Benchmarks

The individual benchmarks provide documentation that includes information on:

- Building the executables
- Running the jobs
- Validation
- Example output
- Environment and Fortran namelist variables that may be altered/tuned.

The Offeror may choose the ratio of MPI vs. threading that will produce the best results on their system. The x86 CPU on the current platforms have a modest number of cores (no more than 18) per socket. Offerors may revise the mapping of cores or apply specific hyper-threading or other techniques to produce the best results on their system. Offerors may choose to disable or modify the existing thread and process affinity used with Slurm and prescribed in the input.nml files.

The following rules shall be adhered to when running the benchmarks:

- All input data shall reside on the production file servers.
- All data output shall be written to the production file servers.
- The use of memory resident file systems for any files required for the run or output by the run is prohibited.
- The runs shall be in an environment controlled by the proposed scheduler.
- The runs shall be from a randomized initial workload (to provide realistic job placement by the scheduler and to avoid problems with simultaneous job start-up).
- Job placement shall be via the standard scheduler algorithms and shall not rely on (for example) lists of specific nodes for each job.
- The runtimes used in calculating performance shall be those provided in the scripts, which are the longest run of the ensemble of runs and include scheduling and MPI start-up costs.

Tests shall be run as a user/group that will be running the three applications in production. In particular, benchmarks shall not be run with elevated privileges.

## 4.3 Volume of Work Metric

Each of the NOAA benchmarks is measured by the increase in the volume of work that it can achieve relative to NOAA's current Intel Broadwell-based HPC system C4. Contributions related to the Intel Haswell-based C3 are not considered. The increase in the volume of work described for C5, across all of the benchmarks, are then combined to give an overall performance improvement figure $V$.

Offeror is to maximize $V$ subject to both the funding constraints and to the Offeror's solution to the mandatory and target technical requirements. The funding constraints include both system maintenance and system analyst support under the terms of the anticipated firm fixed price contract.

Offeror is providing the C5 system as a supplement to an existing ecosystem. Therefore, the main loop time for the benchmarks, not the total runtime, is used to calculate V. The total runtime, which reflects initialization and other factors that may be affected by I/O patterns, must be submitted but does not affect the calculation of V. The prototype code UFS/GFS is not structured in this manner. The total runtime will be used for its contribution to KPP V. Offeror has no obligation to run the CM4 (Large), SHiELD (Medium) or SHiELD (Large) test cases, but may report them to demonstrate system performance and scaling features of the offered solution. The runtime information from C4 is provided as supporting information only.

**Table 4-1. Benchmark Baselines and Weighting**

| Benchmark | Baseline Time($T_b^i$) (Main Loop) (secs) | Baseline Time (Total Runtime) (secs) | Baseline Resources (CPU cores) | Baseline Copies ($C_b^i$) | Weighting ($W^i$) |
|---|---|---|---|---|---|
| **CM4 (Small** | 4089 | 4335 | 6120 | 1 | 0.20 |
| **CM4 (Large)** | 2610 | 2839 | 9720 | 0 | 0.00 |

| | | | | | |
|---|---|---|---|---|---|
| **ESM4 (Small)** | 7632 | 7739 | 2232 | 1 | 0.20 |
| **ESM4 (Large)** | 5884 | 5978 | 3168 | 1 | 0.15 |
| **SHiELD (Small)** | TBD | TBD | TBD | 1 | 0.20 |
| **SHiELD (Medium)** | 6337 | 6479 | 9216 | 0 | 0.00 |
| **SHiELD (Large)** | 3272 | 3432 | 18432 | 0 | 0.00 |
| **SPEAR** | 7687 | 7859 | 2808 | 1 | 0.20 |
| **UFS/GFSv16** | 6975 | 6975 | 6084 | 1 | 0.05 |

For each benchmark, *i*, the baseline time $T_b^i$ and baseline number of copies $C_b^i$ are given above. The baseline time $T_b^i$ for each benchmark correlates with the provided forecast model. Offeror shall provide benchmark results with a runtime $T^i$ and number of copies $C^i$ satisfying the conditions:

a) $T^i \leq T_b^i$ for all benchmarks

b) $C^i$ shall be chosen to fill the whole of each cluster being considered as fully as possible for each benchmark. Note that this means $C^i$ may be fractional. Fractional values of $C^i$ are determined by dividing the unused compute nodes by the number of nodes per individual run of the baseline code being tested.

c) $C^i \geq 1$ for each cluster being offered.

d) Offeror may choose the number of nodes assigned to a benchmark from the proposed solution at their discretion.

The increase in volume of work, $V^i$, is then given by the formula

$$V^i = \frac{T_b^i C^i}{T^i C_b^i}$$

The $V^i$ for each individual component will then be combined via a weighted harmonic mean, with weights, $W^i$ $W^i$ as given in the table above to create the final performance improvement factor $V$ as follows,

$$V = \frac{1}{\sum \frac{W^i}{V^i}}$$

## 4.4 **Benchmark Runtimes**

For all (each) benchmark components, the submitted runtimes shall be less than or equal to the baseline runtimes.

The set of concurrent runs for each benchmark shall be less than $T_b^i$ as specified above and have run-time variability, from minimum to maximum wall-clock time, of not more than +/- 5% from each other or from single runs using the same number of nodes on a dedicated system.

## 4.5 Volume of Work Results

Offeror will state the total benchmark performance $V$ as computed with the formula in Section 4.4 that can be achieved on the total delivered system. In addition to $V$, the Offeror will:

- Provide the $\boldsymbol{T^i}$, $\boldsymbol{C^i}$, and $\boldsymbol{V^i}$ used to compute $V$ for each benchmark in the report.
- Demonstrate that concurrent runs for each benchmark shall be less than $T_b^i$ as specified in Section 4.4.

The following template may be used. The Offeror will describe the ratio of MPI vs. threading that was used. Based on benchmark results or estimates, Offeror will identify the number of nodes to be used on the proposed system to attain the performance $V^i$. Offeror shall provide actual benchmark results, with a detailed description of the hardware and software environment used to produce those results, and may then use those results to validate assumptions made for the proposed or extrapolated results for V.

**Table 4-2. Volume of Work Template**

| | | Main Loop | | | Total Runtime | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Nodes | Ave $T^i$ (secs) | Min $T^i$ (secs) | Max $T^i$ (secs) | Ave $T^i$ (secs) | Min $T^i$ (secs) | Max $T^i$ (secs) | $C^i$ | $V^i$ | V |
| **Actual Results** | | | | | | | | | | |
| **CM4/small** | | | | | | | | | | |
| **CM4/large** | | | | | | | | 0 | 0 | |
| **ESM4/small** | | | | | | | | | | |
| **ESM4/large** | | | | | | | | 0 | 0 | |
| **SHiELD/sm** | | | | | | | | | | |
| **SHiELD/med** | | | | | | | | 0 | 0 | |
| **SHiELD/lg** | | | | | | | | 0 | 0 | |
| **SPEAR** | | | | | | | | | | |
| **UFS/GFS** | | | | | | | | | | |
| **Extrapolated Results** | | | | | | | | | | |
| **CM4/small** | | | | | | | | | | |
| **CM4/large** | | | | | | | | 0 | 0 | |
| **ESM4/small** | | | | | | | | | | |
| **ESM4/large** | | | | | | | | 0 | 0 | |
| **SHiELD/sm** | | | | | | | | | | |
| **SHiELD/med** | | | | | | | | 0 | 0 | |
| **SHiELD/lg** | | | | | | | | 0 | 0 | |
| **SPEAR** | | | | | | | | | | |
| **UFS/GFS** | | | | | | | | | | |
| **Overall** | | | | | | | | | | |

## 4.6 Extrapolations

The Offeror may use benchmark results from existing systems to extrapolate and/or to estimate the benchmark performance on proposed systems. This may include extrapolations based on job size or to account for architectural differences between the tested and proposed machines. Note that extrapolations in performance may

result in risk factors being applied to committed performance. Detailed justifications for the extrapolations will mitigate this risk

Each reported run must explicitly identify:

- The hardware and system software configuration used;
- The build and execution environment configuration used;
- The source change configuration used; and
- Any extrapolation and/or estimation procedures used.

### 4.7 Scalability

In production, NOAA will run a series of parameterized tests to identify optimal run time configurations based on the system size, run time constraints, and other factors. Offeror may provide the results of their scaling runs, where they may choose to make changes to the domain decomposition of one or more benchmarks or where they apply specific optimizations. Offeror may use a Table such as that shown in Table 4-3 to delineate specific notes about those particular runs. Company is interested in the impact of optimizations for node sizes at and near the actual node count specified for the calculation of V and how specific conditions or optimizations may influence the results.

**Table 4-3. CM4, ESM4, SHiELD, SPEAR, UFS/GFS Scaling Results Template**

| **Node Count** | **Time** (secs) | Notes (e.g. Actual/Extrapolated, Quiesced System/Concurrent, Compiler Optimizations, Accuracy) |
|---|---|---|
| | | |
| | | |
| | | |

### 4.8 Accuracy

Offeror shall provide evidence demonstrating the accuracy of solutions from the runs used to compute V. This shall be done following the procedures outlined in the documentation provided with the benchmark codes.

### 4.9 Run-Time Variability

The Offeror shall describe the attributes of their solution that will minimize run time variability, commit to a maximum (worst-case) run time variability for the benchmarks, and describe the rationale for that calculation. This description shall include the measured or anticipated MPI communication characteristics between nodes of the compute system(s) including both average (predictable) latency and bandwidth and tail latency and bandwidth, and the statistical description (mean, median, mode, skew) of these assumptions or measurements.

## 5. C5 Compute Partition

### 5.1 C5 System Performance (TR-1)

Offeror will describe system performance by the sustained performance of the benchmarks and volume, V, as described in Section 4.4.

## 5.2 C5 Memory Configuration (TR-1)

Company anticipates that the minimum main memory capacity per processor socket will be at least 2 GB per core, regardless of the core-count on a specific node, for solutions that use traditional DDR4/5 SDRAM. Company will consider alternative memory configurations that may include stacked memory technologies including high bandwidth memory (HBM). Those HBM configurations may not meet the 2GB/core TR-1 prescribed for SDRAM, but must meet the performance characteristics of § 4.

Offeror will describe any SDRAM memory configuration in terms of ranks, presence of chipkill and similar parameters. Offeror will describe any HBM configuration in terms of dies per stack, density and power per die, and similar parameters.

The System will provide a uniform quantity of memory across each compute node.

Offeror will use a consistent memory configuration between benchmarks and the delivered system, i.e. production memory configuration cannot substantively differ from the benchmark configuration.

## 5.3 System Resilience (TR-1)

Offeror will describe the features of C5 that contribute to system resiliency.

Applications that fail must be restarted by the System Resource Manager (SRM). If the application uses services that have been affected by a system fault, these services must be automatically restored so that the application can continue to progress.

## 5.4 Early Access to C5 Hardware Technology (TR-1)

The Offeror will propose mechanisms to provide the Company with early access to hardware technology for hardware and software testing prior to inserting the technology into the C5 system. Small early access systems are encouraged, particularly if they are sited at the Laboratory. It is expected that Offeror will meet this expectation through delivery, integration and operation of a Test and Development System (TDS) that is architecturally similar or equivalent to the larger compute system.

## 5.5 Early Access to C5 Software Technology (TR-1)

The Offeror will propose mechanisms to provide the Company with early access to software technology and to test software releases and patches before installation on the C5 system. It is expected that Offeror will meet this expectation through delivery, integration and operation of a TDS that is architecturally similar or equivalent to the larger compute partition.

## 5.6 C5 Hardware Options

The Offeror shall propose scalable options using whatever is the natural unit for the proposed architecture design as determined by the Offeror. For example, for system size, the unit may be an additional compute rack, or some other unit appropriate for the system architecture. Inflection points appropriate to the scaling options should be identified, i.e. cooling, infrastructure, or similar changes that are necessary to support a scaling options. If the proposed design has no option to scale one or more of these features, the Offeror should simply state this in the proposal response. Cost data shall not be reflected in the Technical Volume.

### 5.6.1 Scale the System Size (TO-1)

The Offeror will describe options for scaling the overall C5 system (size) beyond the baseline proposal using the appropriate natural unit for the proposed architectural design. These options must include and describe all supporting components that may be affected, including electrical and cooling distribution systems, network and switch fabrics.

### 5.6.2 C5 Test and Development System (TR-1)

The Offeror will describe a system configuration that consists of a minimally deployable system that includes the TDS compute partition as well as its supporting login, management, telemetry and other key services.

The Offeror will describe options for scaling the capacity of the TDS.

## 5.7 **Security Controls (TR-1)**

C5 will be operated as Unclassified, with both Offeror and Company providing required safeguarding or dissemination controls, pursuant to and consistent with any applicable laws, regulations, and government-wide policies.

## 5.8 **Options for Mid-Life Upgrades (TO-1)**

Although not currently funded, Company wishes to understand the options for upgrading the proposed system (mid-life upgrade). Options may be concisely framed as simply as additional compute capability (addressed through existing scaling target requirements) or more complex upgrades that involve modification or addition of components, subsystems, blades, processors, or other technology. Cost data shall not be reflected in the Technical Volume.

## 5.9 **Compute Partition Hardware Requirements**

### 5.9.1 Hardware Performance Monitors (TR-1)

The compute nodes (CNs) will have hardware support for monitoring system performance. This is expected to provide a rich set of events pertaining to the motion of data as it flows across the full memory hierarchy between all proposed computational units. Further, it will include hardware support for monitoring message passing performance and congestion on all node interconnect interfaces of all proposed networks.

### 5.9.2 Hardware Power and Energy Monitors and Control (TR-2)

CN hardware will support monitoring and control of system power and energy. The Offeror will document a Hardware Power Monitor and Control Interface (HPMCI) that will use this hardware support to measure and to log the total power of a node and to control its power consumption. HPMCI will support monitoring and control during idle periods as well as during active execution of user applications.

### 5.9.3 Hardware Debugging Support (TR-1)

CN cores will have hardware support for debugging of user applications, and in particular, hardware that enables setting regular data watchpoints and breakpoints. These hardware features will be made available directly to applications programmers in a documented API and utilized by the code development tools including the debugger.

### 5.10 **Compute Node Operation System (CNOS) Execution Model (TR-1)**

The Compute Node Operation System (CNOS) will support the following application runtime/job launching requirements:

- Processes may be threaded and can dynamically load libraries via dlopen() and related library functions, such as dlmopen.
- All tasks on a single CN will be able to allocate memory regions dynamically that are addressable by all processes on that node. Allocation and de-allocation may be a collective operation among a subset of the processes on a CN.
- The MPI API will be supported for process-to-process communication within a job. Additional native packet transport libraries will be exposed.
- The Pthread interface will be supported and will allow pinning of threads to hardware. MPI calls will be permitted from each Pthread.
- OpenMP threading will be supported. MPI calls will be permitted in the serial regions between parallel regions. MPI calls will be supported in OpenMP parallel loops and regions, with possible restrictions necessitated by the semantics of MPI and OpenMP.
- The Offeror is only required to supply and to support one OS per node type, but the architecture should not preclude the booting of different node OSs. So a job may specify the CNOS kernel, or kernel version, to boot and to run the job on the CN.

### 5.11 **Runtime Variability**

#### 5.11.1 Individual Application Runtime Variability (TR-1)

The main loop runtime of each application benchmark on a dedicated system will not vary by more than 5% across executions.

#### 5.11.2 Production Workload Runtime Variability (TR-1)

Job run-time variability negatively affects a variety of factors including system utilization, throughput, and scheduling. Given a situation where the C5 resources are heavily utilized (90% or more of compute nodes are scheduled), the Offeror shall describe the attributes of their solution that will minimize run time variability, commit to a maximum main loop run time variability for the class of applications described in the benchmarks, and describe the rationale for that calculation. The potential impact of I/O on runtime variability that is associated with F2 being a shared asset is eliminated by using main loop timings only.

## 6. Input/Output Subsystem (IOS) (TR-1)

The F2 IOS is an existing production resource. The requirements for C5 that are related to the IOS involve the ability to read and write from F2 across an EDR fabric. Offeror is responsible for all connections from C5 to the EDR fabric, including router nodes (if applicable) and cables. In no case should an individual cable exceed 50m. i.e. the facility layout as described in § 13 ensures that the worst-case distance from C5 to the IB switches that serve F2 is less than 50m. Company will provide corresponding EDR ports on a Mellanox Switch IB-2 director-class EDR switch.

The F2 IOS provides a single POSIX namespace. The filesystem is Lustre LTS, using version 2.12 or later (regularly updating the minor release level below 2.12 as appropriate). The namespace design efficiently supports a mix of concurrent small/large block read/write operations, and has been tuned to accommodate NOAA application I/O characteristics. This namespace is backed by compression-enabled ZFS for both block and

metadata storage. F2 takes advantage of Lustre's DNE feature for distributed metadata. The Offeror's Lustre client NID configuration shall be Company defined.

The IOS is currently significantly over-provisioned in terms of performance. The DDN SFA14KX block storage system, with 6 controller pairs, and six Dell r640 OSS nodes per controller pair has demonstrated sequential read/write performance in excess of 240GB/s. The F2 metadata subsystem, a NetApp EF570 with 4 Dell r640 metadata servers as MDSs, can realize a sustained IOP rate in excess of 80k.

The Offeror will defend their proposed configuration from C5 to the EDR fabric without consideration for whether the IOS may impact the performance results on C5. Similarly, the IOS is not expected to influence run time variability results obtained during performance and stability testing within the Acceptance Test Framework. As the main loop time for the benchmarks is key, not the end to end runtime, I/O-heavy startup and other similar concerns are mitigated.

The Offeror must commit to and defend a runtime varability metric (§ 4.9). In the event that Company identifies that the runtime variability metric is not met, Company agrees that analysis and migitation will exclude effects associated with the storage/file system.

## 6.1 **High-Level Requirements for Communicating with the C5 I/O Subsystem (IOS)**

All IOS tests listed in Section 6 will be performed on a quiesced storage and compute environment to demonstrate that the compute system has adequate network connectivity and client tuning to drive F2 to meet requirements. Clients will be mounted with noatime and with Lustre checksums disabled (assuming the compute HSN accommodates sufficient data integrity features). All clients will be tuned consistently and statically, and all servers will be tuned consistently and statically across all tests. If the Offeror proposes an LNET routed configuration to access F2 from C5, the Offeror shall provide the required number and configuration of the LNET routers to meet the requirements.

### 6.1.1 Lustre client software

C5 shall support native Lustre clients on the compute nodes that provide IOS access. Offeror will provide a Lustre client and applicable HSN Lustre Network Driver (LND) including build source for both. The compute system that supports Lustre features that include at a minimum: DNE, PFL, and DoM. The Offeror-provided Lustre client will follow the latest stable Lustre release (LTS) by no more than 4 months (available within 4 months of release at https://www.lustre.org/.

Support for the Lustre *client* and applicable HSN LND is the responsibility of the Offeror.

Support for the Lustre *server* is the responsibility of the Company.

### 6.1.2 Single Directory File Create Performance

The IOS namespace will provide a sustained file and directory create rate of 5,000 per second in a single directory. This performance will be observed using a single FEN and a single node of the compute partition separately.

### 6.1.3 Small File Size Transactions Performance

The IOS namespace will sustain an aggregate of 40,000 transactions per second for 5 minutes for parallel file operations where 32 KiB is written into each file from the CN partition in parallel using a sufficient number of CNs (create, open, write, close). The Offeror will describe the assumptions and required number of CNs used to achieve this performance.

### 6.1.4 Aggregate Client File I/O Read/Write Performance

The IOS namespace will provide a sustained aggregate parallel read and write I/O performance of 200 GB/s using 2 MB I/O sizes for no less than 10 minutes using a sufficient number of compute clients. This performance will be observed from the CN partition using sufficient CNs. The Offeror will describe the assumptions and required number of CNs.

### 6.1.5 Single Client File I/O Read/Write Performance

The IOS namespace will provide a sustained parallel read and write I/O performance of 10 GB/s using 2 MB I/O sizes for no less than 10 minutes from a single compute client. This performance will be observed from the CN partition using sufficient CNs. The Offeror will describe the assumptions and required number of CNs.

# 7. High Performance Interconnect (TR-1)

The Offeror will provide a physical network or networks for high-performance intra-application communication within the system. The Offeror will configure each node in the system with one or more high speed, high messaging rate interconnect interfaces. This (these) interface(s) will allow all compute elements in the system simultaneously to communicate synchronously or asynchronously with the high speed interconnect. The interconnect will enable low-latency communication for one- and two-sided paradigms.

The Offeror will provide a high-level description of the system interconnect's topology, latencies, bandwidths, bi-section bandwidth, routing algorithm, and congestion mitigation techniques. Offeror will describe anticipated bit error rates and recovery mechanisms. Offeror will describe connectivity to the F2 IOS including the required number of file system routers (if applicable).

# 8. Base Operating System, Middleware and System Resource Management

Key contributing factors include a stable and resilient operating system and software stack, a capable programming environment, including high-performing compilers, tuned libraries, and capable debugging, functional I/O that meets the bandwidth and latency requirements of a time-critical workflow.

## 8.1 **Base Operating System Requirements (TR-1)**

The Offeror will provide on Front End Environment (FEE) and System Management Nodes (SMN) a standard multiuser Linux Standards Base specification V4.1 or then current (http://www.linux-foundation.org/collaborate/workgroups/lsb ) compliant interactive base operating system (BOS). The BOS will:

- Include the full feature set and packages available in a then-current standard Linux distribution;
- Utilize standard Linux packaging methods; all files in the distribution will be owned by a package and dependencies between packages will be enforced;

- Include source code for all base Linux packages from which the corresponding binary packages can be built in a reproducible fashion;
- Provide consistent APIs and ABIs within a major release of distribution (meaning a binary built on version X.1 will run on any other version X.n).

The BOS will trail the official distribution release by no more than eight months. Updates will continue throughout the life of the system, including both major and minor versions and be buildable from source by the Company.

### 8.1.1 Kernel Debugging (TR-2)

The Linux kernel in the Offeror's BOS will function correctly when all common debugging options are enabled including those features that are enabled at compile time. Kdump (or equivalent) will work reliably and dumps will work over a network (preferred) or to local non-volatile storage (what that exists). Crash (or other online and offline kernel debugger) will work reliably.

### 8.1.2 Networking Protocols (TR-1)

The Offeror's BOS will include IETF standards-compliant versions of the following protocols: IPv4, IPv6, TCP/IP, UDP, NFSv3, NFSv4, LDAPv3, OSPF v2/v3, LACP 802.1ax, Multi chassis link aggregation (MLAG), VxLAN, Protocol Independent Multicast routing (SM, DM, Bidir, SSM), and BGP v4.

Provided switches will include a method of obtaining MAC addresses of connected hosts via SNMP, or API for management and automation purposes. Provided switches will be configured to ensure adequate buffering; Packet loss shall not exceed 0.1% based on flows between management/support devices (NFS/LDAP) and computes nodes.

### 8.1.3 Reliable System Logging (TR-1)

The Offeror's BOS will include standards-based system logging. The BOS will have the ability to log to local disk (where those exist) as well as to send log messages reliably to multiple remote systems. Company leverages standards-based centralized (remote) log management and monitoring systems, including or integrating Elasticsearch, Logstash, and Kibana (ELK), Kafka and Nagios. In case of network outages, the logging daemon should queue messages locally (where possible) and deliver them remotely when network connectivity is restored.

### 8.1.4 Operating System Security

#### 8.1.4.1 Authentication and Access Control (TR-1)

The Offeror's BOS will implement basic Linux authentication and authorization functions. All authentication-related actions will be logged including: logon and logoff; password changes; unsuccessful logon attempts; and blocking of a user along with the reason for blocking. User access will be denied after an administrator-configured number of unsuccessful logon attempts. All Offeror-supplied login utilities and authentication APIs will allow for replacement of the standard authentication mechanism with a site-specific pluggable authentication module (PAM).

#### 8.1.4.2 Software Security Compliance (TR-2)

The BOS will be configurable to comply with industry standard best security configuration guidelines such as those from the Center for Internet Security (http://benchmarks.cisecurity.org).

## 8.2 **Distributed Computing Middleware**

### 8.2.1 LDAP Client (TR-1)

The Offeror will provide LDAP version 3, or then current, client software, including support for SSL/TLS. The supplied LDAP command-line utilities and client/NSS libraries will be fully interoperable with an OpenLDAP Release 2.4 or later LDAP server.

### 8.2.2 Cluster Wide Service Security (TR-1)

All system services including debugging, performance monitoring, event tracing, resource management and control will support interfacing with the BOS PAM (Section 8.1.4.1) function. This protocol will be efficient and scalable so that the authentication and authorization step for any size job launch is less than 5% of the total job launch time.

## 8.3 **System Resource Management (SRM) (TR-1)**

The Offeror will provide SRM in an integrated system software design that seamlessly leverages the proposed design principles, potentially including a C5 / C6 two-system design. Offeror will harden the SRM services.

### 8.3.1 Batch Compatibility (TR-1)

#### 8.3.1.1 Slurm Compatibility (TR-1)

The SRM system will support all of the features of Slurm 20.02, including management of federated clusters, multiple partitions, in-depth accounting, advanced reservations, backfill scheduling, topology optimized resource selection, resource limits, and MPMD support.

#### 8.3.1.2 MPI Compatibility (TR-1)

Offeror shall describe the interface between the SRM application launcher and MPI, indicating if special compilation options are necessary and if user-compiled codes will need to be rebuilt when the SRM version changes.

### 8.3.2 Performance and Scalability

#### 8.3.2.1 Support for Hundreds of Jobs (TR-1)

The SRM system will support hundreds of simultaneous running jobs and thousands of simultaneous idle/blocked jobs.

#### 8.3.2.2 Job Start Performance (TR-1)

The SRM will be able to start a full-system job within three minutes. The SRM will be able to clean up after a successful full-system job within three minutes. The SRM will be able to clean up after an unsuccessful full-system job within ten minutes.

#### 8.3.2.3 Application Launch Performance (TR-1)

After a job has been fully-initialized, a full-system application launch must pass the MPI_Init() barrier in less than thirty (30) seconds.

#### 8.3.2.4 System State Responsiveness (TR-1)

The SRM and SRM-API will provide the overall system status in a scalable fashion, providing information needed in less than five (5) seconds.

# 9. Front-End Environment

The FEE includes Front-End Node (FEN) hardware as well as software necessary to support end-users of the system. Offeror shall make no assumptions regarding the hardware and software that comprise the existing FEE. Offeror shall deliver the system with all software and libraries necessary to build and execute the application software described in Section 4. The licenses for all software provided shall allow for use by any authorized user of the system.

A single FEN *type* is anticipated to support interactive use, code compilation, and job execution in a load-balanced manner as well as pre- and post- processing of data from the CNs. Due to the large number of use cases for the FEN, the ability to provide container-based implementations supporting different use cases on the FENs is strongly desired. The Front End Software Programming Environment should be able to run inside this container.

## 9.1 **Front-End Node (FEN) Hardware Requirements**

The following requirements are specific to the FEN hardware.

### 9.1.1 FEN Count (TR-1)

The Offeror will propose sufficient FENs to support 5 simultaneous parallel compiles and an anticipated 25 concurrent users at any one time.

Offeror will provide an option for additional FENs in appropriate scalable units to the underlying node architecture. This option must include and describe all supporting components that may be affected, including electrical and cooling distribution systems, network and switch fabrics. No cost data shall be included in the Technical Volume.

FENs will be compatible or similar to the compute node architecture to eliminate the need to cross-compile.

### 9.1.2 FEN Non-Volatile Resources (TR-1)

Each FEN will have sufficient NVMe-based SSD resources to store no less than twice the aggregate memory of the node.

The FEN will be able to boot over a network and mount a shared root file system.

### 9.1.3 FEN High-Availability (TR-1)

The FEN will have high availability features including but not limited to redundant and hot swappable power supplies, hot swappable fans, and ECC memory.

The FEN will operate completely independently from the management system that provisions and configures them. The management system can be down for maintenance and the FEN should remain operational and support user's code compilation, file manipulations on Lustre, and SRM client-server interactions.

#### 9.1.4 FEN I/O Configuration (TR-2)

All FENs will have sufficient network interfaces to access a local site network, the Offeror-supplied management network and the file system network. File system access from the FENs to the IOS shall be consistent with the compute partition access method. The FEN will have at least two free I/O slots such that the sites can configure additional network or storage interfaces as needed.

### 9.2 **Front-End Environment Software Requirements**

#### 9.2.1 Parallelizing Compilers/Translators

Offeror will provide at least two different compilers that support the underlying chipset proposed for the compute nodes. NOAA currently uses the GNU compiler and the Intel compiler for their Intel-based C3 and C4 systems.

Offeror will provide appropriate licensing to support five (5) simultaneous compilations of each baseline language.

##### 9.2.1.1 Baseline Languages (TR-1)

The Offeror will provide fully supported implementations of Fortran 2008 (ISO/IEC 1539-1:2010, ISO/IEC TR 19767:2005(E), ISO/IEC TR 29113 - https://www.iso.org/standard/50459.html), C (ANSI/ISO/IEC 9899:2011; ISO/IEC 9899:2011 Cor. 1:2012(E) - http://www.open-std.org/jtc1/sc22/wg14/www/standards), and C++ (ANSI/ISO/IEC 14882:2014 - http://www.open-std.org/jtc1/sc22/wg21/docs/standards) or then current versions. Fortran, C, and C++ are referred to as the baseline languages. An assembler will be provided. The Offeror will provide the fully supported capability to build programs from a mixture of the baseline languages (i.e., inter-language sub-procedure invocation will be supported).

##### 9.2.1.2 Baseline Language Optimizations (TR-1)

The Offeror will provide baseline language compilers that perform high levels of optimization that allow the application programmer to use all CN supported hardware features. Baseline language compilers will support directives to provide information (e.g., aliasing information beyond the restrict keyword) required for or to direct additional optimizations.

##### 9.2.1.3 Baseline Language Support for OpenMP Parallelism (TR-1)

The Fortran, C, and C++ compilers will support OpenMP Version 5.0 (or then current) (http://www.openmp.org). All baseline language compilers will include the ability to perform automatic parallelization. The baseline language compilers will produce symbol tables and any other information required to enable debugging of OpenMP parallelized applications.

The baseline languages and runtime library support for the compute node will include optimizations that minimize the overhead of locks, critical regions, barriers, atomic operations, tasks and self-scheduling "do-loops" by using special compute node hardware features.

#### 9.2.1.4 Python Support (TR-1)

The FEE will support the launching and running of applications based on multiple languages, including Python 3.6 (or then current versions) as released by http://www.python.org. Python applications may use dynamically linked libraries and SWIG (www.swig.org) and f2py generated wrappers for the Python defined API for the ability to call C, C++ and Fortran library routines. The Offeror will provide optimized versions of the Python modules numpy, scipy, and mpi4py, and will track future releases.

#### 9.2.1.5 Perl Support (TR-1)

The FEE will provide Perl 5.x or higher, including libraries and modules to run the Unified Model specifically including POSIX:calc, Date:manip, File:Copy.

#### 9.2.1.6 JAVA Support. (TR-1)

The FEE will provide JAVA 1.7 or higher.

### 9.2.2 Debugging and Tuning Tools (TR-1)

#### 9.2.2.1 Parallel Application Debugger (TR-1)

Offeror will provide Eclipse Integrated Development Environment (IDE) (http://elipse.org) and the GNU Debugger GDB.

#### 9.2.2.2 Stack Traceback (TR-2)

The Offeror will propose runtime support for stack traceback error reporting. Critical information will be generated to STDERR upon interruption of a process or thread involving any trap for which the user program has not defined a handler. The information will include a source-level stack traceback (indicating the approximate location of the process or thread in terms of source routine and line number) and an indication of the interrupt type.

Default behavior when an application encounters an exception for which the user has not defined a handler is that the application dumps a core file. By linking in an Offeror-provided system library the application may instead dump a stack traceback. The stack traceback indicates the stack contents and call chain as well as the type of interrupt that occurred.

#### 9.2.2.3 User Access to a Scalable Stack Trace Analysis Tool (TR-2)

The Offeror will supply a scalable stack trace analysis and display GUI-based tool that will allow non-privileged users to obtain a merged stack traceback securely and interactively from a running job.

#### 9.2.2.4 Lightweight Corefile API (TR-2)

The Offeror will provide the standard lightweight corefile API, defined by the Parallel Tools Consortium or a mutually agreed-upon equivalent format, to trigger generation of aggregate traceback data for all running threads. The Parallel Tools Consortium defines a specific format for lightweight core files (see http://web.engr.oregonstate.edu/~pancake/ptools/lcb/). The Offeror will provide an environment variable (or an associated command-line flag), with which users can specify that the provided runtime will generate lightweight corefiles instead of standard Linux/Unix corefiles. User control will be provided to generate either lightweight or

standard Linux/Unix corefiles from a selected subset of the MPI rank processes as well as to select the file-system locations into which to store them.

#### 9.2.2.5 Profiling Tools for Applications (TR-1)

The Offeror will provide a range of application profiling tools including Open|SpeedShop (https://openspeedshop.org/), HPCToolkit (http://hpctoolkit.org/), and gprof (https://www.gnu.org/software/binutils).

#### 9.2.2.6 Event Tracing Tools for Applications (TR-1)

The Offeror will provide the Score-P measurement infrastructure (http://www.vi-hps.org/projects/score-p) for tracing programming model events as well as collecting hardware performance counters. The Open|SpeedShop POSIX I/O tracer will also be provided through the Open|SpeedShop toolset (https://openspeedshop.org).

#### 9.2.2.7 Performance Monitor APIs and Tools for Applications (TR-1)

The Offeror will provide performance monitor APIs and tools, whereby performance measures from hardware performance monitors are obtained for individual threads or processes are reported and summarized for applications. The Offeror will provide a native API that allows full access to the performance monitor hardware. The Offeror will deliver PAPI, Version 5 (or then current), that gives user applications access to the 64b hardware performance monitors (Section 5) and exposes all HPM functionality to user applications.

#### 9.2.2.8 Timer API (TR-2)

The Offeror will provide an API for interval wall clock and for interval timers local to a thread/process. The system and user timers will have a global resolution of no worse than 3 microseconds (i.e., this wall clock is a system wide clock and is accurate across the system to 3 microseconds).

### 9.2.3 Application Building

#### 9.2.3.1 FEN Compilation Environment for CN (TR-1)

The Offeror will provide a complete compilation environment that allows the sites to compile and to link applications and daemons on the FEN for execution on the CN. The FEN environment will support the use of current versions of the standard GNU Autotools (Autoconf, Automake, Libtool) for automatic configuration and building of libraries and applications, OS, and runtime libraries.

#### 9.2.3.2 GNU Make Utility (TR-1)

The Offeror will provide the GNU make utility version 4.2 (or then current) with the ability to utilize parallelism in performing the tasks in a makefile.

#### 9.2.3.3 CMake (TR-1)

The Offeror will provide the CMake build system, version 3.8.2 (or then current). Offeror will also provide CMake platform files that enable cross-compiling. Platform files will correspond to tool chains available on the FENs and CNs, and all compiler tool chains available on these nodes.

#### 9.2.3.4 Linker and Library Building Utility (TR-1)

The Offeror will provide an application linker with the capability to link object and library modules into dynamic and static executable binaries.

### 9.2.4 Application Programming Interfaces (TR-1)

#### 9.2.4.1 Optimized Message-Passing Interface (MPI) Library (TR-1)

The Offeror will provide a fully supported, highly optimized implementation of the MPI-3.1 standard as defined by http://www.mpi-forum.org/docs/mpi-3.1/mpi31-report.pdf (or then current).

As the MPI standard evolves, the Offeror will provide an MPI implementation that is standards compliant of that new standard within nine months after MPI Forum standardization.

#### 9.2.4.2 Math Libraries (TR-2)

The Offeror will provide optimized single-node and parallel mathematics libraries.

# 10. System Management and RAS Infrastructure

NOAA has explicit requirements to deliver products on a regular and recurring basis. Company seeks a C5 system design that can deliver 100% availability, given the inherent ability of the system design to detect and remediate error conditions and manage unscheduled and scheduled maintenance.

## 10.1 **System Availability**

System availability is defined as:

$$A_o = \frac{Operational\ Uptime - Operational\ Downtime}{Operational\ Uptime}$$

In the situation where the Offeror proposes independent systems (C5/C6), the larger ecosystem is considered to be UP when at least one of the two individual systems is up and running the full application suite of jobs. the larger ecosystem is considered to be DOWN if the loss of any Offeror resource precludes the generation of any specific product.

## 10.2 **Robust System Management Facility (TR-1)**

The Offeror will provide a full-function, robust, scalable facility that enables efficient management of the system. This system management capability will run on one or more System Management Nodes (SMNs) and will control all aspects of system administration in aggregate, including modifying configuration files, software upgrades, file system manipulation (as appropriate), reboots, user account management and system monitoring. The system management capability should run natively on the SMN operating systems, i.e. not be abstracted into virtual machines or containerized solutions. A GUI should not be the primary or preferred method for controlling the system management software.

### 10.2.1 System Management Architecture (TR-1)

The Offeror will describe the hardware and software architecture and major components of the system management facility, highlighting features that provide ease of management, operational efficiency, scalability, state consistency (software and hardware) and effective fault detection/isolation/recovery.

### 10.2.2 Fast, Reliable System Initialization (TR-1)

The major components of the system will boot in less than one (1) hour. The boot process should include all infrastructure, hardware, software and any file systems required for the system to operate as designed. This excludes the Company-supplied IOS. Mounting of the IOS global namespace on all applicable nodes will add no more than an additional five (5) minutes to the boot process. The system boot will progress without human intervention with the exception of a final release to start batch jobs when all hardware and software is ready.

### 10.2.3 System Software Packaging (TR-1)

The Offeror will provide all software components of the system via a single Software Package Management System (SPMS). The SPMS will provide tools to install, to uninstall, to update, to remove, and to query all software components and versions. The SPMS will allow multiple versions of packaged software to be installed in Company-specified locations and used on the system at the same time, and will provide the ability to roll back to a previous software version.

### 10.2.4 System Logging and Formatting (TR-1)

The Offeror will provide a scalable logging solution that aggregates all system messages into a central location. All log messages shall have timestamps and those timestamps shall conform to the RFC3339 standard.

### 10.2.5 Remote Manageability (TR-1)

All nodes of the system will be 100% remotely manageable, and all routine administration tasks automatable in a manner that scales up to the full system size.

#### 10.2.5.1 Out of Band Management Interface (TR-1)

The system nodes will provide an Out of Band (OOB) management interface. This interface will be accessible over the system management network. This interface will allow system RAS and system administration functions to be performed without impact to or dependence on the high performance interconnect. It is preferred that Offeror solutions provide a separate network for system management from telemetry collection. Company will supply the fiber-optic uplink(s) from the OOB network(s) to the Laboratory management network. Company will ensure that appropriate ports are available for these connection(s).

#### 10.2.5.2 Remote Console and Power Management (TR-1)

The Offeror will provide a polled console input/output device for each instance of the operating system kernel that is available via a system-wide console network that scales to permit simultaneous and continuous access to all consoles. The Offeror will provide the capability to log all console output to logfiles. Rack PDUs that provide remote on/off switching control of individual outlets via a well-known API are desired.

### 10.2.6 System-wide Authentication/Authorization Framework (TR-1)

The Offeror's proposed system will provide a common authentication/authorization framework including some means of integrating with external directory services. A user's credentials, once validated, will be honored by all system components (e.g., FEE, batch system, and CNOS). Similarly, a user's privileges, once established, will be enforced by all subsystems. This framework will integrate seamlessly with the PAM provided in Section 8.1.4.1.

## 10.3 **Reliability, Availability and Serviceability (TR-1)**

The Offeror's proposed system will be designed with Reliability, Availability and Serviceability (RAS) in mind. The Offeror will provide a scalable infrastructure that monitors and logs the system health and facilitates fault detection and isolation.

### 10.3.1 Mean Time Between Failure Calculation (TR-1)

The Offeror will calculate the mean time to system failure for each major item, to include individual Field-Replaceable Units (FRU)s, compute nodes, compute partitions or systems, the compute system(s), and individual elements of the Front End systems (scheduler, login nodes, management nodes). Describe any formulas used, and assumptions made in completing these calculations.

### 10.3.2 Hardware RAS characteristics (TR-1)

The Offeror will describe component level RAS characteristics that are exploited to achieve a high level of system resilience and data integrity. This description should include methods of error detection, correction and containment across all major components and communication pathways. The Offeror will describe RAS features of the memory subsystem, including advanced error correction capabilities of DRAM in the proposed solution.

### 10.3.3 Failure Detection, Reporting and Analysis (TR-1)

The Offeror will provide a mechanism for detecting and reporting failures of critical resources, including processors, network paths, and disks. The diagnostic routines will be capable of isolating hardware problems down to the FRU level.

### 10.3.4 Scalable System Diagnostics (TR-2)

The Offeror will provide a scalable diagnostic code suite that checks processor, cache, memory, network and I/O interface functionality for the full system in under thirty (30) minutes. The supplied diagnostics will accurately isolate failures down to the FRU level.

### 10.3.5 Modular Serviceability (TR-1)

The service of system components, including nodes, network, power, cooling, and storage, will be possible with minimal impact and avoiding full-system outage. Hot swapping of failed FRUs will not require power cycling the cabinet in which the FRU is located.

### 10.3.6 Graceful Service Degradation (TR-2)

The Offeror's RAS facility will detect, isolate and mediate hardware and software faults in a way that minimizes the impact on overall system availability. Failure of hardware or software components will result in no worse than proportional degradation of system availability.

### 10.3.7 Comprehensive Error Reporting (TR-1)

All bit errors in the system (e.g., memory errors, data transmission errors, local disk read/write errors, and SAN interface data corruption), over-temperature conditions, voltage irregularities, fan speed fluctuations, and disk speed variations will be logged by the RAS facility. Recoverable and non-recoverable errors will be differentiated. The RAS facility will also identify irregularities in the functionality of software subsystems.

### 10.3.8 System Environmental Monitoring (TR-1)

The Offeror will provide the appropriate hardware sensors and software interface for the collection of system environmental data. This data will include power (voltage and current), temperature, humidity, fan speeds, and coolant flow rates collected at the component, node and rack level as appropriate. System environmental data will be collected in a scalable fashion, either on demand or on a continuous basis as configured by the system administrator.

### 10.3.9 Hardware Configuration Database (TR-2)

The RAS system will include a hardware database or equivalent that provides an authoritative representation of the configuration of system hardware. At minimum this will contain:

- Machine topology (compute nodes and I/O nodes);
- Network IP address of each hardware component's management interface;
- Status (measured and/or assumed) of each hardware component;
- Hardware history including FRU serial numbers and dates of installation and removal;
- Method for securely querying and updating the hardware database from system hosts other than the SMNs.

## 10.4 **Company-provided Telemetry Database Support (TR-1)**

The C5 compute system and supporting infrastructure will support sending monitoring and metrics telemetry to an external to C5 and Company-provided telemetry data store.

### 10.4.1 Compute telemetry (TR-1)

Offeror will support the collection and shipment of Linux client statistics to Company-provided telemetry data store.

### 10.4.2 IOS Client Telemetry (TR-1)

Offeror will support the collection and shipment of Lustre client statistics to Company-provided telemetry data store.

#### 10.4.3 HSN and Management Network Telemetry (TR-1)

Offeror will support the collection and shipment of HSN and Management Network statistics to Company-provided telemetry data store.

# 11. Local Area Networks and Services

## 11.1 Network Infrastructure

Company will provide network connections from Offeror-provided equipment to external services via 10– and/or 40- Gigabit Ethernet. Company will provide both the uplink port and fiber optic connection to the C5 system(s). Offeror should anticipate that uplinks will be short-range, fiber optic and configure their network equipment accordingly.

Company will enable direct connectivity to the F2 IOS via the existing EDR IB fabric by providing adequate ports on the Mellanox Switch IB-2 director. Offeror is responsible for the network connections to this switch fabric.

## 11.2 Company-Provided Services

Company will provide RSA, LDAP, DNS, license management (not the specific licenses), separate home directories and the parallel scratch file system F2 via an existing and separate file system. Offeror is not responsible for providing any of these services or functions.

# 12. Maintenance and Support

## 12.1 Hardware Maintenance (TR-1)

### 12.1.1 Hardware Maintenance Offerings (TO-1)

The Offeror will supply hardware maintenance for the system for a five-year period starting with system acceptance. The Offeror will propose the hardware maintenance offering that is appropriate to meeting the availability metrics of different system components. Offeror will describe how the offered maintenance solution manages both unscheduled and scheduled maintenance outages.

### 12.1.2 On-site Parts Cache (TR-1)

The Offeror will provide and refresh an on-site parts cache of FRUs and hot spare nodes of each type proposed for the system. The size of the parts cache, based on Offeror's MTBF estimates for each component, will be sufficient to sustain necessary repair actions on all proposed hardware and keep them in fully operational status for at least two weeks without parts cache refresh.

Company will provide power, space, and cooling for any diagnostic test equipment or cabinet that can be used for burn-in or causal analysis.

All spare parts and diagnostic equipment remain property of Offeror until such time that they are introduced in to the system.

### 12.1.3 FRU with Non-Volatile Memory Destroyed (TR-1)

The Offeror will identify any FRU in the proposed system that can persistently hold data in non-volatile memory or storage. FRU with non-volatile memory or storage that potentially contains user data shall not be returned to the Offeror. Instead, the Company will certify to Offeror that the FRU with non-volatile memory or storage that could potentially contain user data has been destroyed as part of the Offeror's RMA replacement procedure.

## 12.2 **Software Support (TR-1)**

The Offeror will supply software maintenance for each Offeror-supplied software component starting with the successful completion of system acceptance and ending five years after the system acceptance. The Offeror will propose the software maintenance offering that is appropriate to meeting the availability metrics of different system components. Because of Offeror's design, Offeror may describe separate software maintenance offerings for critical components (e.g. system boot, job launch, I/O and RAS) versus non-critical items. Offeror will describe how the offered maintenance solution manages both unscheduled and scheduled maintenance outages.

Offeror provided software maintenance will include an electronic trouble reporting and tracking mechanism and periodic software updates.

### 12.2.1 Software Feature Evolution (TR-1)

The Offeror will support new software features on the delivered hardware for the full term of the warranty or maintenance period covered by an exercised option, with the exception of when new software features are specific to a different hardware platform. For software produced by the Offeror, new features will appear on the delivered hardware at the same time as provided on the Offeror's other commercial platforms. For software not produced by the Offeror, new features will appear on the delivered hardware within 6 months of general availability.

## 12.3 **Problem Escalation (TR-1)**

The Offeror will describe their technical problem escalation mechanism in the event that hardware or software issues are not being addressed to the Company' satisfaction.

## 12.4 **On-site Analyst Support (TR-1)**

The Offeror will supply two full-time on-site analysts to Laboratory. These analysts will provide advanced, senior Linux systems administration skillsets for every aspect of the Offeror's infrastructure, hardware, software, and software stack. Due to Laboratory access controls and the requirements for privileged access to C5 systems, Analysts must be either U.S. Citizens or Legal Permanent Residents (LPR).

The base term for the on-site analysts will be 60-months, beginning with completion of the Entrance Criteria for the Hardware Acceptance Test.

The Company may request additional on-site analysts, which will be described separately in the Business Volume.

# 13. C5 Facilities Requirements

The requirements described in this section apply to the complete C5 system.

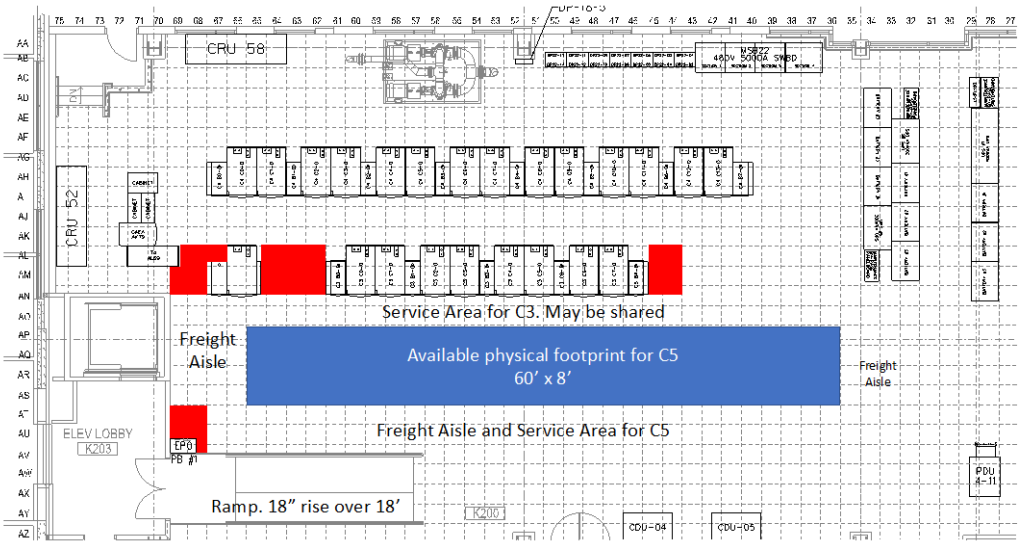
## 13.1 **Laboratory Facilities Overview (TR-1)**

### 13.1.1 Laboratory Facility - Space

The C5 system will be installed at physical address 1 Bethel Valley Road, Building 5600, Data Center (DC) K200 on the ORNL campus in Oak Ridge, TN 37830. K200 is an existing multi-tenant DC that contains C3, C4, F2, and existing FENs; C5 will occupy additional footprint in this same DC. The available physical footprint is shown in Figure 13-1.

The available gross square footage in this DC is approximately 660 $ft^2$. This takes in to account existing requirements to maintain appropriate service and freight areas. Offeror may leverage these service/freight areas for their needs, and to overlap service areas currently in use by C3. i.e. there is no need for the Offeror to use the 60'x8' physical space for their own service areas or clearance requirements. Within K200, the areas marked Freight Aisle are typically a minimum of 48" wide.

For the purposes of this request for proposal, Offeror should assume that C3 is not retired until after C5 is introduced in to production.




The designated area is clear-span, on-plenum (raised floor). There are no columns affecting the area. The raised floor is a pedestal and stringer system with a 250 pounds/$ft^2$ rating. Pedestal height is 18".

Total deck to deck height is approximately 14'. Total height from the raised floor to the drop-ceiling is no less than 104".

**Figure 13-1. Physical footprint available to the C5 system in DC K200**

Electrical and mechanical distribution is existing and can generally be modified as necessary to suit.

The Offeror should plan for all technical load to be included in this footprint, including infrastructure or management systems, network equipment associated with the system interconnect, and any supporting mechanical or electrical distribution items that are part of the solution. This may include CDUs, step-down transformers, or similar items. Some flexibility does exist within the freight and service aisle area defined for this effort, especially the area East (to the right) of the C5 footprint. Company will work with the successful Offeror to define any additional needs for floor space as needed.

Offeror will ensure that minimum service clearances to electrical and safety systems and existing tenant equipment are maintained. These minimum service clearance areas are shown above in red. Service areas/clearances for Offeror equipment may overlap with service areas/clearances for existing systems or electrical systems.

The Company will maintain the operating environment within the data center in accordance with Class A1, as defined in the ASHRAE Thermal Guidelines for Data Processing Environments. Room air is continuously filtered with MERV 8 filters. Air entering the data center is filtered with no less than MERV 11 filters.

Physical access (supporting delivery) to the K200 facility is via a commercial/industrial loading dock, with leveler, and a freight elevator. Laboratory maintains a freight aisle to this K200 room location that is no less than 52" wide.

Delivery and laydown space at the Laboratory is very limited. The Laboratory has minimal conditioned space to store hardware prior to installation activities. Coordinated efforts between Company and Offeror are required to support an effective delivery and installation effort.

### 13.1.2 Laboratory Facility – Mechanical Distribution

Laboratory has facility water supply distribution in K200 that includes supply temperatures described by ASHRAE Technical Committee 9.9 Liquid Cooling Guidelines as Liquid Cooling Class W2. The water temperatures for the W2 water supply to C5, based on the energy plant design and the annual range of environmental conditions in Oak Ridge, TN, will always range between 69F and 71F. The Offeror is expected to use this W2-distribution (medium-temperature water (MTW)) as part of their solution.

Company currently maintains the supply setpoint for the MTW loop at 21.5C/ 71F, and guarantees that the MTW supply will never exceed 24C/75.2F. Company further ensures that the supply temperature will always exceed dewpoint by no less than 4 degrees. Pressure for the W2 system is governed by both the MTW pumping differential pressure (dP) setpoint (part of the larger mechanical facility, and a primary setpoint for the Laboratory's Summit supercomputer) and the K200 booster pump dP setpoint (ensuring adequate pressure for the NOAA C3, C4 and C5 systems). In the current configuration, the maximum MTW inlet pressure is approximately 37 pounds per square inch gauge pressure (psig). The maximum MTW system's dP available is 20psi. C5 may absorb as much as 1000GPM at 20 pounds per square inch differential (psid) at the supply setpoint of 21.5C/ 71F. Company prefers solutions that minimize flow (GPM) and provide a corresponding high return temperature to the MTW system.

The MTW is treated with both biocides and corrosion inhibitors. The characteristics of the MTW supply are shown in Table 13-1.

**Table 13-1. Characteristics of the MTW Supply Serving C5**

| | | |
|---|---|---|
| All Metals < 0.10 ppm | Calcium < 1.0 ppm | Magnesium < 1.0 ppm |
| Manganese < 0.10 ppm | Phosphorus < 0.50 ppm | Silica < 1.0 ppm |
| Sodium < 0.10 ppm | Bromide < 0.10 ppm | Nitrite < 0.50 ppm |
| Chloride < 0.50 ppm | Nitrate < 0.50 ppm | Sulfate < 0.50 ppm |
| pH 6.5 – 8.5 | Turbidity (NTU) < 1 | Conductivity < 500.0 µS/cm at 21.5C |

Company will provide chilled water as part of fresh air pressurization, air conditioning, and relative humidity control of the room. However, chilled water is not available as a direct component for the Offeror's solution.

Inlet air may be assumed to be no higher than 75F. Laboratory maintains an RH setpoint of 50%.

Company seeks mechanical solutions that are as close to room-neutral as is practical. Based on the Offeror's engineering solution, and based on a (assumed/fixed) 71F inlet water temperature, a (assumed/fixed) 75F inlet air temperature, a (assumed/fixed) 50% RH setpoint, Offerors must calculate the ejection of waste heat to the room ambient air for both the compute and support rack *types* across the range of anticipated heat loads generated by

the compute solution (idle, typical, HPL, TDP). Specifically, Offeror shall provide a series of charts showing curves of these various load levels, where the solution is room neutral, with inlet water temperatures on the x-axis and required flow rates on the y-axis.

For support cabinets, the Offeror's solution shall use MTW water through Offeror-supplied rear door heat exchangers (RDHX) that contains a fail-closed, two way modulating valve. Pressure independent flow control and balancing is preferred.

For elements of the Offeror's solution that do not exchange air with the DC, Offeror will describe the anticipated parasitic (convective and radiant) heat load from those elements, across a range of water supply temperatures from 69.0F to 71F.

Delivery of mechanical services (MTW) to the Offeror's solution will be from below the raised floor. If necessary, cooling distribution units (CDUs) for rack-level flow control of the cooling water are the responsibility of the Offeror, who can provide operating envelopes for supply temperatures, pressure, and flow for the primary Facilities cooling system as well as wetted material list and makeup water quality requirements. Company will provide suitable delivery mechanisms, with final design for those based on the Offeror's solution. Company will provide appropriate tile cuts and seals/bushings. The Offeror's CDU may provide variable secondary water flow based on changing primary temperature and flow as well as changes in cooling demand by the connected IT equipment. Because the MTW serves multiple systems, including Summit and NOAA's C3 and C4, the CDU's primary flow control valve is to be a fail-closed, two way modulating valve. Pressure independent flow control and balancing is preferred.

Should the Offeror's solution connect directly to the MTW supply/return loop, Offeror will include rack-level flow control using a fail-closed, two way modulating valve. Pressure independent flow control and balancing is preferred.

If Facility MTW is to be used at the rack level with no heat exchanger, the water quality requirements are to be no tighter than that of the closed loop hydronic system serving 5600 K200. Offeror shall provide both a wetted material list and water chemistry requirements for the proposed system.

The use of containment systems by the Offeror is allowed, subject to personnel safety concerns related to high-temperature or "hot" aisles, and subject to Laboratory fire-protection regulations.

The Offeror's cooling equipment is to have a communications interface between it and the Facility. Table 13-1 reflects the types of data expected of this communications link. The Offeror will describe the actual data available on this communications link based on the proposed system.

**Table 13-2: Cooling Equipment Monitoring Interface**

| Name | Where | Typical Provider | Frequency of measurement | Accuracy/Units |
|---|---|---|---|---|
| **Water flow** | System | Facility | Once every 30 sec | +/- 5% Liter/min (gal/min) |
| | CDU | Negotiated | Once every 30 sec | +/- 10% Liter/min (gal/min) |
| **Thermal data** | System or branch | Facility | Once every 60 sec | +/-1° C (1.8 °F) |
| | CDU | Negotiated | Once every 60 sec | +/-1° C (1.8 °F) |

| Name | Where | Typical Provider | Frequency of measurement | Accuracy/Units |
|---|---|---|---|---|
| | Rack | HPC system | Once every 60 sec | +/-1° C (1.8 °F) |
| | Node | HPC system | Once per sec | +/-2° C (3.6 °F) |
| | Component | HPC system | Once per sec | +/-2° C (3.6 °F) |
| **Power** | System | Facility | Once per sec | +/- 5% Watts |
| | CDU | Negotiated | Once per sec | +/- 5% Watts |
| | Rack | HPC system | Once per sec | +/- 5% Watts |
| **Dew Point Temperature** | System | Facility | Once per 60 sec | +/-2° C (3.6 °F) |
| | Branch, rack or cabinet | Negotiated | Once per 60 sec | +/-2° C (3.6 °F) |
| **Pump Speed** | System | Facility | Once per 60 sec | +/-3 % of full speed |
| | CDU | Negotiated | Once per sec | +/-3 % of full speed |
| **Pressure Differential** | System | Facility | Once per 5 sec | +/-10 kpa (1.5 PSI) |
| | Branch, rack or cabinet | Negotiated | Once per 5 sec | +/-10 kpa (1.5 PSI) |
| **Valve Position** | System | Facility | Once per 60 sec | +/- 5% (% Open) |
| | Branch | Negotiated | Once per 60 sec | +/- 5% (% Open) |
| | CDU | Negotiated | Once per 60 sec | +/- 5% (% Open) |
| | Rack | HPC system | Once per 3 sec | +/- 2% (% Open) |

### 13.1.3 Laboratory Facility - Electrical Distribution

Laboratory will provide 3-phase 480VAC supply voltage that is derived from a wye connected transformer. Compute load can be either wye or delta type. The maximum size of any individual circuit supplying a compute rack will be 125A. While Laboratory expects to deliver 3-phase 480VAC supply voltage to the Offeror's equipment (i.e. the rack is the demarc), this does not constrain or preclude the Offeror from using rack-level step-down transformers or other methods to deliver lower voltage to individual power supplies within their solution. Note that this does not eliminate the overall requirement that equipment that is supplied by 480VAC systems shall be rated for 106% nominal voltage.

For support systems with significantly smaller loads, Laboratory will provide 208Y/120V supply voltage. Critical management systems that use the 208Y/120V supply voltage can be protected by diverse supply voltage sources, including generator/UPS-backed facilities. The maximum size of any circuit supplying a support rack using 208Y/120V supply voltage will be 60A.

Laboratory will continue to operate C3 and C4 as production resources during the installation and acceptance of C5. Laboratory maintains adequate electrical distribution capacity for all three compute systems. Offeror may assume a sustained electrical budget of 1.2MW, running the benchmarks defined in Section 4. Offeror need not constrain the solution to 1.2MW for short term benchmarks such as HPL. Laboratory can increase the sustained electrical budget by 1.0MW to 2.2MW in aggregate with some difficulty, including risk to the March substantial-

completion date for facility modifications. Company prefers solutions that can remain within the 1.2MW budget as well as the mechanical distribution constraint of 1000gpm and 20psid for the MTW supply.

Offeror's 480V and 208V equipment shall have a fault current withstand rating of 65kA and 10kA respectively.

Due to the limited plenum space below the system, it is required that delivery of electrical services is from overhead. Offeror may assume that the Company will provide overhead cable tray from the 480VAC distribution point (switchboard) to the C5 equipment, and from the 208V PDU, if necessary, to the C5 support rack(s). Offeror is responsible for cable management from the electrical distribution system (a Laboratory-provided terminal block) to their equipment. Offeror is responsible for cable management associated with any cabinet to cabinet network connections, including cable tray.

The connections from the Offeror's equipment to the facility's power distribution system will use liquid tight flexible metal conduit, flexible metal conduit, MC cable, or other permanent wiring method. It is preferred that a connection box be provided on the top of racks where Company can connect building supply wiring. Pin and sleeve connectors for 480VAC connections will not be used.

The Company will make direct connection of electrical service to the Offeror's equipment at Laboratory-provided terminal blocks. To avoid requiring multiple lug connections for each phase at the upstream overcurrent device, paralleling of conductors in raceway that supply compute racks is not allowed.

All equipment must be NRTL-certified.

## 13.2 **Power & Cooling Requirements (TR-1)**

Without compromising performance objectives, the Offeror will minimize the power and cooling required by the proposed systems. Further, Offeror will minimize the use of UPS/generator-backed power.

The Offeror will describe how its proposed system fits within the space, mechanical and electrical constraints. The Offeror will provide a detailed estimate of the total amount of power in kW (kilowatts) and kVA, and cooling in either refrigeration tons or BTU (British Thermal Units), required by the complete system. The estimate will describe the power and cooling loads for the individual racks (by rack type) and for each substantive component of the system. At a minimum, the Offeror will describe power and cooling estimates anticipated during special-purpose / diagnostic / benchmarking efforts such as HPL that may best approach the design TDP, and maximum (sustained execution of the class of applications represented by the benchmarks described in Section 4 acrross the entire solution), and idle (OS-only, minimum) operation.

The Offeror will describe the power factor, phase balancing, and harmonics management for their AC power systems.

### 13.2.1 Power and cooling utilization data collection (TR-2)

At every electrical or mechanical connection from the facility infrastructure to the Offeror's solution, the Offeror will provide a mechanism for providing power and cooling utilization or consumption data for that connection.

### 13.2.2 Rack-Integrated In-line PDU (TR-2)

The Offeror may propose a power distribution solution that supports more than one compute cabinet, using an in-line power distribution unit. If the Offeror does not provide an in-line PDU, the Offeror will provide a solution

that minimizes the number of connections to each rack. The design of any PDU will provide protection of each branch circuit in that PDU and remote management of each branch circuit in that PDU. Regardless of the PDU design (in-line serving more than one rack, or in-rack), the Offeror will provide a connection point for each device that the facility can permanently terminate to the appropriate building circuit.

### 13.2.3 Tolerance of Power Quality Variation (TR-1)

The design of the power system for the entire C5 ecosystem will be tolerant of power quality events.

All power supplies must be tolerant to voltage surge and sag in accordance with the latest version of SEMI F47. Computer power at the Laboratory is reliable and clean, but not conditioned. There is no uninterruptible power available for the compute system. 208-240VAC components of the Offeror's solution, which might include the network components, and other infrastructure, may be supported by UPS systems in dual-fed configurations that can provide tolerance (short-term ride-through) to power quality events that exceed the SEMI F47 specification.

The Offeror will describe the tolerance of their power system to power quality events, in terms of both voltage surge and sag, and in duration.

### 13.2.4 Power Factor and Harmonic Current Requirements (TR-1)

The power factor of the computer racks when operating at benchmark levels shall be $\geq 0.98$. The power factor of 208-240VAC components of the Offeror's solution, which includes the network components and other infrastructure, will be $\geq 0.95$.

At benchmark power levels, the maximum total harmonic current and the maximum individual harmonic current levels will meet recommendations provided in Table 2 of IEEE STD 519-2014 for $I_{sc}/I_l \leq 20$. Total harmonic current levels shall be at or less than a transformer with a K-4 rating can carry with no derating.

### 13.2.5 Cooling Requirements (TR-1)

All air and liquid cooling required for each system will be listed separately, in relation to the heat load created by the operation of each system. The Offeror should understand that both air and liquid cooled systems will reside in the same room.

For any portion of the Offeror solution that requires air cooling, the Offeror will specify the environmental conditions required in CFM, temperature, and humidity. The Offeror will specify the cooling envelopes required by each rack type, the allowable flow and temperature excursion magnitudes and durations, thresholds for flow and temperature at any warning, alarm, and critical/shutdown/throttling points in the envelope, and how any of the previously mentioned items may change at the idle, typical, and maximum operating points.

### 13.2.6 Liquid Cooling Solution Description (TR-1)

The Offeror will fully describe the liquid cooling apparatus and all implications for siting and facilities modifications (e.g., water connections, flow rates, temperature, humidity, pressure, quality, chemistry, and particulates). The solution will not preclude use of existing piping systems, which may include plastic, polypropylene, ductile steel, stainless steel, copper, and epoxy coated materials.

### 13.2.7 Liquid Cooling Temperature (TR-1)

The liquid temperature will be supplied at (nominally) 71F.

The Offeror will fully describe the range of operating conditions for the proposed solution.

## 13.3 **Floor Space Requirements (TR-1)**

The Offeror will provide a proposed floor plan for the C5 system that fits into Laboratory's multi-tenant DC K200, as described in Section 13.1.The floor plan will show the placement of all system components as provided by the Offeror.

## 13.4 **Cable Management Requirements (TR-1)**

The cable management system external to the cabinets will be accommodated above cabinets. Cable management configuration will assume no more than 40% fill, non-conductive materials, comprise solid-bottom tray and modesty panels, and will ensure that all minimum bend radius requirements are met. All cables will be contained in cable trays supplied by the Offeror.

## 13.5 **Physical Access Requirements (TR-1)**

The C5 system will be installed and physically located inside a controlled access area. The Company will only provide access to these areas for authorized personnel. All on-site personnel will be required to submit applications for access and to be approved by standard Laboratory procedures prior to entry into the facilities. Offeror personnel that are not U.S. citizens or LPRs may be further restricted, from both physical and system access, in accordance with the specific requirements of the Laboratory facility.

On-site space will be provided for personnel and equipment storage. The Offeror will describe the anticipated volume of equipment and supplies that must be accommodated as part of their maintenance schedule and plan.

## 13.6 **Safety Requirements (TR-1)**

Offeror personnel will practice safe work habits, and comply with all associated Laboratory Environment, Safety and Health (ES&H) requirements.

The Laboratory will allow Offeror to service, repair or replace any de-energized component(s) of a system, including components that are designed to be warm-swapped or hot-swapped, subject to restrictions regarding the overall availability of the C5 system. Any de-energized component will completely isolate all subsidiary components, through hardware and not software (i.e., on-off switches or switch-rated circuit breakers), without any potential for re-energization.

## 13.7 **Safety and Power Standards (TR-1)**

All equipment proposed by the Offeror will meet industry safety, and appropriate power quality and power supply standards.

Equipment that is supplied by 480VAC systems shall be rated for 106% nominal voltage.

### 13.8 System Installation and Integration (TR-1)

The Offeror will deliver, install, and integrate the elements of the C5 system in the ORNL facility. Offeror is responsible for appropriate handling and management of all materials including appropriate disposal of packaging waste.

### 13.9 System Decommissioning (TR-1)

The Offeror will decommission and remove all elements of the C5 system at the end of its service life. Offeror is responsible for appropriate handling and management of all materials. Equipment title will revert to Offeror at this time. Costs associated with the decommissioning and removal of the system are the responsibility of the Offeror. System decommissioning is notionally scheduled for October 1, 2026.

## 14. Project Management (TR-1)

As part of the RFP response, the Offeror will identify the Project Liaison Assignments for both the Executive Liaison and Technical Project Manager.

Documents described in this section are not required in the RFP response; however, the Offeror will confirm its commitment: 1) to include the following project management approaches and elements in its execution of any subcontract awarded; and 2) to provide the associated documentation by the required times and through a reliable and easily accessible mechanism that supports change control.

The Offeror will provide in its RFP response a set of milestones for the deliverables in this section as described in the **Key Build Phase Milestone Dates.**

This project management approach is designed to help the Offeror successfully meet its commitment, to help the Company to track the project, and to help Company and Offeror to understand and to mitigate risks successfully.

The specific detailed planning, effort tracking, and documentation requirements for the development, manufacturing, installation and support efforts that will be delivered as part of the subcontracts are delineated in the following sections.

### 14.1 Key Planning Deliverables

The Offeror will develop, deliver, submit for approval and maintain the following Key Planning Deliverables.

- Project Liaison Assignments: Offeror Executive Liaison and Technical Project Manager;
- Plan of Record, including Hardware and Software Schedule, and Project Milestones including late binding dates, assumptions, risks and opportunities;
- Factory Test Plan;
- Full-Term Hardware Development Plan;
- Site Preparation Plan;
- Installation Process Plan;
- Maintenance and Support Plan.

## 14.2 Project Meetings

Upon subcontract award, the project meetings and performance reviews described below shall commence.

Table 14-1: Project Meetings

| Purpose | Subcontractor Deliverables | End Date |
|---|---|---|
| Kickoff | • Within two weeks of subcontract award, face to face meeting of Offeror Executive and Technical Staff to review Key Planning Deliverables. | N/A |
| Monthly Project Teleconference | • Project status and issues updates<br>• Updated project action item list and assignments<br>• Updated schedule and critical path | Final acceptance |
| Site Preparation and Operations Planning, as needed. | • Site preparation, status, issues, action items, and assignments<br>• Updated Installation Plan and/or Installation Guide | Final acceptance |

## 14.3 Key Build Phase Milestone Dates (TR-1)

Offeror will provide the Company, in its proposal response, a proposed set of milestones for this section and, for each milestone, proposed associated payment that is applicable to Offeror's proposed deployment timeline and methodology. Offeror is encouraged to identify milestones for each year of the project that merit revenue that Offeror can legally recognize in that year.

Prior to award, Company and Offeror will finalize the list of Key Build Phase Milestone Dates. Following is a list of example key tasks/dates of importance to Laboratory.

- Project Liaisons assigned (identified in Offeror response);
- Plan of Record complete;
- Early system access begins;
- On Site System Administrators arrive on site;
- Begin delivery and installation of system;
- System Installation and Integration complete;
- System Acceptance Complete.

## 14.4 Key Elements of the Plan of Record

Within 60 days of subcontract award, the Offeror will provide a detailed Plan of Record (POR), which will include:

- **Project management plan** with management teams and organizational breakdown structure (OBS) identified.
- **Points of contact:** table of key personnel for contributing organizations within the company and its major subcontractors, and a description of their responsibilities and how these areas will be coordinated by the management team.
- **Full term project schedule** and Gantt chart for the duration of the contract will be kept under configuration control with an audit trail of changes. The schedule will be developed using the Critical Path Method (CPM) scheduling technique and will utilize the same numbering scheme as the WBS. The Company must concur with changes to capabilities, delivery/installation dates, and acceptance processes/schedules.

- **Project Plan Detail.** Using the same structure and sequence as this document, the POR will describe the planned tasks and their milestones in sufficient detail that Company can assess and track progress. The plan should cover the duration of this contract and reflect a level of detail that covers the major subsections of this document. The Project Plan Detail will be kept under configuration control with an audit trail of changes. The Company must concur with changes to capabilities, configurations, delivery/installation dates, and testing processes/schedules.

## 14.5 Key Elements of Factory Test Plan

Within 60 days of subcontract award, the Offeror will provide a Factory Test Plan, which will include the following elements.

- Process for qualifying vendors;
- Factory burn in and validation test plan;
- ASIC and system level margin testing;
- Pre-ship test plan for equipment.

The Factory Test Plan will be updated, as necessary, to reflect the Offeror's latest processes and plan throughout the life of the project.

## 14.6 Key Elements of Full-Term Hardware Development Plan

Within 60 days of subcontract award, the Offeror will provide a detailed Full-Term Hardware Development Plan, which will include the following elements. The Company must concur with changes to capabilities and delivery/installation dates.

- **Processor Technology**. Identify the planned milestones for processor development that lead to those to be deployed in the C5 system. In particular, provide milestones for silicon process development, sampling, engineering quantities, and production quantities for each processor generation leading to the C5 system.
- **Node Development**. Provide the planned tasks and milestones for product development for all node types covered by this contract. Include tasks and milestones for: memory architecture; cache coherency protocols; ASIC development; performance modeling efforts; applications analysis; functional verification test; and system test. Indicate how and when this technology will be inserted to meet subcontract milestones.
- **System Scalability and Performance Testing**. Provide the planned tasks and milestones for the scalability testing of system components. Include development of hardware for reliability, availability and serviceability (RAS).

## 14.7 Key Elements of Site Preparation Plan

Within 60 days of subcontract award, the Offeror will provide a detailed Site Preparation Plan that will include at a minimum the following items:

- Cabinet dimensions, packaging diagrams, and weights (in all configurations – in packaging, dry, with any liquid coolant);
- Electrical requirements for everything provided by selected Offeror;
- System layout and cabling requirements, including expansion options;
- Cable tray requirements;

- Environmental requirements;
- Expected power and cooling requirements;
- Cooling water quality requirements;
- Safety requirements.

### 14.8 **Key Elements of Installation Process Plan**

At least 90 days before the first equipment delivery, the Offeror will provide a detailed Installation Process Plan. The Installation Process plan will include the following, in the minimum:

- Core installation team and staffing plan;
- Equipment delivery and testing schedule;
- Staging and temporary storage area needs;

### 14.9 **Key Elements of Maintenance and Support Plan**

At least 60 days before the first equipment delivery, the Offeror will provide a detailed Maintenance and Support Plan. The Maintenance and Support Plan will include the following:

- Obtaining hardware and software support from Offeror;
- Reporting and tracking system problems;
- Trouble report escalation process;
- Preventative maintenance requirements;
- Cycling power;
- Parts replacement;
- On site spare storage volume and access controls.

## 15. Acceptance Requirements (TR-1)

Upon delivery and installation, a series of performance, functionality, and availability tests will be performed prior to acceptance. Acceptance testing will comprise multiple components where the overall goal is to ensure that the system as a whole is high-performance, scalable, resilient and reliable. Acceptance testing will exercise the system infrastructure with a combination of benchmarks, injected failures, and stability tests. Any requirement described in the Technical Specification may generate a corresponding acceptance test. Company may identify other system aspects that merit testing. In general, the successful Offeror should expect a hardware acceptance phase that includes (Offeror) demonstration that all hardware and software is present, installed, functional, and that the benchmarks can be executed from a non-privileged account, meeting or exceeding the offered performance. Offeror should expect a detailed functionality test that systematically walks through the Technical Specification. Offeror should expect a Performance Test that provides appropriate conditions for execution and verification of the benchmarks described in Section 4. Offerors should expect a concise Stability test, controlled through a test harness, that keeps the C5 system full subscribed. The specifics of the acceptance test plan will be determined during subcontract negotiation.

# Appendix A Technical Volume Instructions

### General Instructions

For each section of the Statement of Work (SOW), provide a point by point response to the technical requirements of that Section. Preface each response narrative with the corresponding numbering scheme so that the Offeror response to a particular item uses the same number as the SOW requirement, target, or option. SOW text may be included or paraphrased as appropriate and may be reflected in a smaller font (but no smaller than 8-point).

Each response should include a detailed discussion of how all mandatory requirements (MRs), mandatory option requirements (MOs), proposed technical options (TO-1and TO-2) and proposed target requirements (TR-1 and TR-2) are met or exceeded, as well as a discussion of any Offeror-identified additional performance features that are included in the technical solution.

For any target requirement (TR-1 or TR-2) that is not proposed or not met, the Offeror will include an explicit statement to that effect, an explanation of the choice not to meet the technical option or target requirement, and any proposed remediation.

## 1. Introduction

No response necessary.

## 2. Program Overview

No response necessary.

## 3. High Level System Overview

### 3.1 Key Design Elements

Provide an executive summary that describes how the proposed system meets or exceeds the key design criteria for all major hardware and software system components, including the compute partitions, input/output subsystem, strategy for test and development system(s) and proposed benchmark performance.

### 3.2 C5 High Level System Description

Complete Section 3.2 per the instructions found in the SOW.

### 3.3 C5 High Level Software Model

Complete Section 3.3 per the instructions found in the SOW.

### 3.4 C5 High Level Performance Objective

Describe the increase in the volume of work, $V^i$, and the proposed performance improvement figure $V$, as defined in Section 4.5 of the SOW. The successful Offeror will demonstrate the ability to balance both time to solution and volume of computation (number of concurrent models) across their proposed solution within a time-critical window.

### 3.5 C5 High Level Project Management

Complete Section 3.5 per the instructions found in the SOW.

Provide the Key Tasks and Build Phase Milestone Dates as described in Section 14. Describe each, where applicable, and identify pertinent cost, schedule, or technical risk associated with each task. Identify technical, schedule, cost or other triggers that may influence or impact the delivered system.

## 4. Benchmarks

Provide documentation on system performance for the proposed system by running the CM4, ESM4, SHiELD, SPEAR, and UFS/GFS benchmarks and evaluated using the Volume of Work metric, $V$, and procedures defined in Section 4 of the SOW. Note that these rules (including those for code changes) will form the basis for the C5 performance acceptance tests.

## 5. C5 Compute Partition

### 5.1 C5 System Performance

Summarize the total benchmark performance, $V$ as computed with the formula in Section 4 of the SOW. Explain the methodology used to obtain any projected results.

### 5.2 C5 Memory Configuration

Describe the memory architecture and configuration of the C5 system in accordance with the requirements described in Section 5.2 of the SOW.

### 5.3 System Resilience

Describe the features of C5 that contribute to system resiliency. Describe how the SRM manages application failure. Describe how specific system services may be restored in the event of a system fault.

### 5.4 Early Access to C5 Hardware Technology

Complete Section 5.4 per the instructions found in the SOW.

### 5.5 Early Access to C5 Software Technology

Complete Section 5.5 per the instructions found in the SOW.

### 5.6 C5 Hardware Options

Complete Section 5.6 per the instructions found in the SOW. Do not provide cost data in the Technical Volume but complete this in the Business Volume.

### 5.7 Security Controls

Complete Section 5.7 per the instructions found in the SOW.

### 5.8 Options for Mid-Life Upgrades

Describe the plan of record for system upgrades (other than scaling using scalable units proposed elsewhere) that can extend the computing performance or capability of the C5 system(s), extend the productive life of the C5 system, increase system performance, reliability, availability, or serviceability. Identify anticipated schedules for each distinct upgrade.

Do not provide cost data in the Technical Volume but complete this in the Business Volume.

### 5.9 Compute Partition Hardware Requirements

In accordance with the requirements described in Section 5.9 of the SOW, describe the availability of hardware performance monitors, power and energy monitors, and hardware debugging support, and how that information is made available to system administrators, system monitoring tools, application programmers and code development tools.

### 5.10 Compute Node Operation System Execution Model

In accordance with the target requirements described in Section 5.10 of the SOW, describe the Compute Node Operation System (CNOS) Execution Model. Describe the proposed functionality, features, remediations, risks, or additional benefits that address the target requirement.

### 5.11 Runtime Variability

Describe the anticipated production workload runtime variability, and any assumptions that are made regarding that metric. Describe the proposed individual application runtime variability, and any assumptions that are made regarding that metric for each of the applications.

Describe the architectural, hardware, software, interconnect, or similar features of the proposed system that contribute to the production workload runtime variability and individual application runtime variability.

## 6. Input/Output System

In accordance with the target requirements described in Section 6 of the SOW, describe the Offeror's solution to the performance requirements of the system with respect to the I/O requirements.

## 7. High Performance Interconnect

Complete Section 7 per the instructions found in the SOW.

## 8. Base Operating System, Middleware and System Resource Management

In accordance with the target requirements described in Section 8 of the SOW, describe the proposed functionality, features, remediations, risks, or additional benefits that address the target requirements.

## 9. Front-End Environment

For each item in Section 9 of the SOW, describe the proposed functionality, features, remediations, risks, or additional benefits that address the target requirements.

## 10. System Management and RAS Infrastructure

For each applicable item in Section 10 of the SOW, describe the proposed functionality, features, remediations, risks, or additional benefits that address the target requirements.

## 11. Local Area Networks and Fabrics

No response necessary.

## 12. Maintenance and Support

Describe the proposed hardware and software maintenance strategies throughout the life of the C5 subcontract. Include the level of service the Offeror intends to provide at various points during the subcontract period (i.e., system build, system installation, acceptance testing, capability period, and general availability period).

Delineate specific roles and responsibilities for the Company, Offeror, and any lower-tier subcontractor personnel.

Identify the number of full-time maintenance personnel dedicated to servicing the systems as well as their level of experience on the equipment and software being provided, their training, and other relevant qualifications.

Describe problem escalation procedures and the process for generating, tracking, and closing trouble tickets.

## 13. C5 Facilities Requirements

Complete Section 13 per the instructions found in the SOW. Company notes that the options for system packaging are widely varied. Company seeks mechanical solutions that are as close to room-neutral as possible, during typical/notional production, given the constraints, where those solutions are also sympathetic to the demands of other systems that share the MTW system.

Describe how the proposed solution fits within the power budget. Provide a detailed estimate of the total amount of power in kW (kilowatts) and kVA, and cooling in either refrigeration tons or BTU (British Thermal Units), required by the complete system. Describe the power and cooling loads for the individual racks (by rack type) and for each substantive component of the system. Describe power and cooling estimates anticipated during special-purpose / diagnostic / benchmarking efforts that may best approach the design TDP, and maximum, typical / notional, and idle (minimum) operation. Describe the basis for those estimates.

Describe the power factor, phase balancing, and harmonics management for AC power systems.

Provide a detailed floor plan including any subsystems.

List any other facilities requirements.

## 14. Project Management

Complete Section 14 per the instructions found in the SOW.

Identify the Project Liaison assignments for both Technical Project Manager and Executive Liaison.

Confirm the commitment to execute the elements of Section 14: Project Management.

Provide the anticipated milestones for deliverables of key elements of C5, including at a minimum the FEN, TDS and compute system(s).

## 15. Acceptance Requirements

No response necessary.

# Appendix B Glossary

## Hardware

| | |
|---|---|
| CN | System compute nodes. Compute Nodes (CN) are nodes in the system on which user jobs execute. |
| FPE | Floating Point Exception. |
| GB | gigaByte. gigaByte is a billion base 10 bytes. This is typically used in every context except for Random Access Memory size and is $10^9$ (or 1,000,000,000) bytes. |
| GiB | gibiByte. gibiByte is a 1,073,741,824 base 10 bytes (i.e., $1024^3$ bytes). |
| ISA | Instruction Set Architecture. |
| FEN | Front End Nodes. Front End Nodes are nodes where users and administrators can login in and interact with the system. |
| MB | megaByte. megaByte is a million base 10 bytes. This is typically used in every context except for Random Access Memory size and is $10^6$ (or 1,000,000) bytes. |
| MTBF | Mean Time Between (Hardware) Failure. A measurement of the expected hardware reliability of the system or component. The MTBF figure can be developed as the result of intensive testing, based on actual product experience, or predicted by analyzing known factors. See URL: http://www.t-cubed.com/faq_mtbf.htm |
| Node | A set of CPUs sharing random access memory within the same coherent memory address space. From the SRM perspective, is the indivisible resource that can be allocated to a job. |
| Non-Volatile | Non-volatile memory, nonvolatile memory, NVM or non-volatile storage, is computer memory that can retain the stored information even when not powered. |
| NVRAM | Non-Volatile Random Access Memory |
| PB | petaByte. petaByte is a quadrillion base 10 bytes. This is typically used in every context except for Random Access Memory size and is $10^{15}$ (or 1,000,000,000,000) bytes. |
| Processor | The computer ASIC die and package. |
| Scalable | A system attribute that increases in performance or size as some function of the peak rating of the system. |
| SECDED | Single Error Correction Double Error Detection. Storage and data transfer protection mechanism that can detect parity errors (single bit errors) and detect storage or data transfer errors with multiple bits in them. |
| SNMP | Simple Network Management Protocol is a popular protocol for network management. It is used for collecting information from, and configuring, network devices, such as servers, printers, hubs, switches, and routers on an Internet Protocol (IP) network. |
| Thread | Hardware threads are typically exposed through the operating system as independently schedulable sequences of instructions. A hardware thread executes a software thread within a Linux (or other) OS process. |
| TB | TeraByte. TeraByte is a trillion base 10 bytes. This is typically used in every context except for Random Access Memory size and is $10^{12}$ (or 1,000,000,000,000) bytes. |

## Software

| | |
|---|---|
| API | Application Programming Interface: Syntax and semantics for invoking services from within an executing application. |
| Baseline Languages | The Baseline Languages are Fortran, C, and C++. |
| BOS | Base Operating System (BOS). Linux (LSB 3.1) compliant Operating System run on the FEN. |
| Fully supported | A software product-quality implementation, documented and maintained by the HPC machine supplier or an affiliated software supplier. |
| Job | An allocation of resources to a user for a specified period of time. The user should be given control over which resources can be allocated to a job. |
| CNOS | Light-Weight Kernel providing operating system functions to user applications running on CN. |
| OS | Operating System |
| Task | A process launched as a job step component, typically an MPI process. |