Attachment or Appendix A
Definition: Climate Modeling and Research System Technical Specification
Solicitation Number RESERVED

Table of Contents

1. Program Background **5**
1.1. Memorandum of Understanding Between DOE and NOAA [I] *5*
1.2. Work for Others Agreement Between DOE and NOAA [I] *5*
1.3. Conventions [I] *5*
1.4. Purpose [I] *6*
1.5. NOAA's Existing Research and Development HPC Computing Centers [I] *6*
1.6. NOAA's Existing Long-term Storage [I] *6*
2. Procurement Goals and Scope [I] **6**
2.1. Computational Resources [I] *6*
2.2. Storage Resources [I] *7*
2.3. Infrastructure Services [I] *7*
2.4. Network Services [I] *7*
2.5. Security Support [I] *7*
3. Responsibilities of the Selected Offeror and Interface with ORNL **7**
4. Climate Modeling and Research System Requirements **7**
4.1. Composition and Expected Lifetime [C] *7*
4.1.1. Single Phase System Delivery 8
4.1.2. Multi-Phase System Delivery 8
4.1.3. System and Subsystem Upgrades 9
4.2. System Computational Capacity [C] *9*
4.3. System Computational Capability [C] *9*
4.4. Floating Point Standards *10*
4.5. Node Requirements *10*
4.5.1. General Requirements 10
4.5.2. Service Nodes 10
4.5.3. Login Nodes 10
4.6. Node Memory Subsystem Requirements [C] *11*
4.7. High Speed Interconnect Requirements [C] *11*
4.7.1. Interconnect Configuration 11
4.7.2. Interconnect Bandwidth 11
4.7.3. Interconnect Topology Symmetry 12
4.7.4. Link Bandwidth 12
4.7.5. MPI Performance Requirements [S] 12
4.7.6. Bit Error Rate [S] 12
4.8. Test and Development System [C] *12*

DRAFT

5. I/O Systems and Data Management 13
5.1. Fast Scratch (FS) File System 13
5.2. Long-term Fast Scratch (LTFS) File System 13
5.3. FS and LTFS File System Bandwidth and Capacity Requirements 14
5.4. Common Data Management Requirements 18

6. Software 19
6.1. Site Support Advocate 20
6.2. Workload and Resource Management 20
6.2.1. Workload Management [I] 20
6.2.2. Resource Management 20
6.3. Software Suite 20
6.4. Checkpoint/Restart 22

7. Resiliency, Reliability, Availability, Serviceability 22
7.1. RAS System Initialization and Reboot 22
7.2. RAS System Diagnostics and Maintenance 22
7.3. RAS System Event Monitoring and Archival 22
7.4. System Mean Time To Interrupt and System Effectiveness Level Requirements 23
7.4.1. System Mean Time Between Interrupt 23
7.4.2. System Mean Time Between Failure Requirements 23
7.5. Resiliency 23
7.6. Reliability 24
7.7. Availability 25
7.7.1. System Effectiveness Level 25
7.8. Serviceability 25

8. Infrastructure and Networking Services 25
8.1. Infrastructure 25
8.2. Wide Area Network Connectivity [I] 26
8.3. Local Area Network Connectivity 26

9. Security 27

10. Home File System 28

11. Warranty, Maintenance, and Support Services 28
11.1. Extended Warranty 28
11.1.1. Warranty for Single System Installation 29
11.1.2. Warranty for Multi-Phase System Installation 29
11.1.3. Warranty for TDS 29
11.2. Hardware 29
11.3. Software 29
11.4. System Support 30
11.5. System Administration 30

DRAFT

11.6. Application Support *30*
11.7. Facilities Engineering Support *30*
11.8. Training *31*
11.9. Documentation [S] *31*
11.10. Warranty, Maintenance, and Support Termination *31*

12. Facilities **31**
12.1. Basis [I] *31*
12.2. Facilities Description [I] *32*
12.2.1. General Facilities Description [I] 32
12.2.2. Electrical Infrastructure and Electrical Conditions Affecting the Offeror [S] 32
12.2.3. Mechanical Infrastructure and Mechanical Conditions Affecting the Offeror [S] 34

DRAFT

Attachment or Appendix A
Definition: Climate Modeling and Research System Technical Specification
Solicitation Number RESERVED

List of Figures

Figure 1. Conventions 6

Figure 2. Example lifetime of a two-phase CMRS Installation 9

Figure 3. CM2-HR File Size Distribution 14

Figure 4. (Example) Calculating R_E and D_E using 2008 Baseline Data 15

Figure 5. Data production rate for a 50K core system with S_c=2.0 16

Figure 6. Bandwidth assuming data is written during 5% of total program execution time 16

Figure 7. Total modified bandwidth required by the hypothetical file system 17

Figure 8. Estimated size of FS and LTFS based on hypothetical 50K cores 17

Figure 9. Conceptual CMRS Network Infrastructure 27

Figure 10. Peak power consumption and Sustained (maximum) power consumption 33

Figure 11. Available Floor Space for the Climate Modeling and Research System 35

DRAFT

1. Program Background

1.1. Memorandum of Understanding Between DOE and NOAA [I]

In September 2008, the National Oceanic and Atmospheric Administration (NOAA) and the Department of Energy (DOE) signed a Memorandum of Understanding (MOU) on collaborative research. Through this MOU, Oak Ridge National Laboratory (ORNL) provides NOAA with advanced high-performance computing resources for prototyping critical weather and climate applications in support of NOAA's mission. Computing allocations are fulfilled at ORNL through the Innovative and Novel Computational Impact on Theory and Experiment (INCITE) program.

1.2. Work for Others Agreement Between DOE and NOAA [I]

In July 2009, NOAA and DOE completed a Work for Others (WFO) Agreement to enter into more sustained projects for the challenges that require a closer cooperation and understanding between DOE/ORNL and NOAA. This Agreement will help the Nation prepare for the challenges and risks posed by climate change by improving predictive and adaptive capacities at global to local levels and supporting the developing world in carrying out vulnerability analyses and addressing their findings.

Under the terms of this Agreement, ORNL shall host high-performance computer and data systems that will deliver improved state-of-the-art climate data and model experiments. These models will be used to understand and predict climate variability and change, as well as to produce decision-support tools to facilitate understanding climate change, mitigation strategies, and adaptation options for the Nation.

To this end, ORNL is issuing a Request for Proposal (RFP) for a computing and storage system that is tailored to meet climate modeling and research mission requirements. This Climate Modeling and Research System (CMRS) shall provide the correct balance of reliability, sustained performance, run time variability, reproducibility, and architectural resiliency to meet the needs of this climate research community.

1.3. Conventions [I]

Attributes of the CMRS are ranked according to their relative contribution to functionality, productivity, and usability. The following conventions are used:

Attribute Priority	Description
Critical	Items considered the most important, or critical to the CMRS. Offeror shall describe any exceptions to items considered critical, and describe proposed methods for mitigating the impact of an offer that does not provide this item. Sections that include these items are marked with [C].
Significant	Items that are of sufficient significance that the effectiveness of the CMRS would suffer if not provided. Offeror is encouraged to provide as many such items as possible. Sections that include these items are marked with [S].
Enhancing	Items that enhance the utility of, and are desirable for, the CMRS. Offeror is encouraged to provide as many such items as possible. Sections that include these items are marked with [E].
Information	Items that are provided as additional information to an Offeror regarding information about the NCCS or NOAA that may aid an Offeror in determining

DRAFT

the most suitable system configuration. Sections that include these items are marked with [I].

Figure 1. Conventions

1.4. Purpose [I]

The CMRS acquisition will support ORNL efforts to provide NOAA with computational resources to achieve their mission goals. ORNL and the Offeror will provide technical support that contributes directly to operating high performance computer and data systems for NOAA so as to deliver improved climate and weather data and model experiments. These models will be used to understand and predict climate and weather variability and change, as well as to produce decision-support tools to facilitate understanding climate change, mitigation strategies, and adaptation options for the Nation. The proposed effort leverages the significant specialized expertise and unique ORNL capabilities.

1.5. NOAA's Existing Research and Development HPC Computing Centers [I]

NOAA currently operates three research and development (R&D) HPC computing centers, the Geophysical Fluid Dynamics Laboratory (GFDL) in Princeton, NJ, the National Center for Environmental Prediction (NCEP) in Gaithersburg, MD, and the Earth System Research Laboratory (ESRL) in Boulder, CO. The GFDL workload focuses on natural climate variability and anthropogenic changes and in the development of the required earth system models. In support of these activities, GFDL is also a participant in the Intergovernmental Panel on Climate Change (IPCC). NCEP is focused on operational weather and seasonal/interannual climate forecasting. The ESRL Global Services Division (ESRL/GSD) focuses on weather and other environmental applications.

CMRS computing resources provided by ORNL shall support the research and development activities of all three Centers. It is anticipated that the demand for system resources will exhibit a specific bias towards long-term climate modeling, followed by seasonal/interannual climate modeling and weather modeling.

1.6. NOAA's Existing Long-term Storage [I]

NOAA possesses significant long-term storage systems at both GFDL (currently storing 22 petabytes (PB)) and NCEP (currently storing 10PB), with a smaller system at ESRL/GSD (storing approximately 1PB). These systems are expected to remain at those NOAA locations. The existing workflow for all three NOAA R&D Centers allows for computational products (model output) to be moved from large computational resources, regardless of location, and including ORNL, across high bandwidth, low latency national networks to the respective local archive for post-processing activities. This workflow model will expand to include the CMRS computational resource. There is no archive requirement in this solicitation. [I]

2. Procurement Goals and Scope [I]

2.1. Computational Resources [I]

The selected Offeror shall provide computational resources such that maximum performance is achieved within a fixed-price contract and best meeting all requirements. A complete, fully functional system is required. ORNL envisions a large system providing a very high level of dependability. Furthermore, weather and climate codes are typically data intensive; therefore balanced system performance will be imperative.

DRAFT

2.2. Storage Resources [I]

The selected Offeror shall provide file systems and data storage that provides balanced performance for the proposed computational resources and system software. Reference Section 5 for additional descriptions of the file system and data storage requirements.

2.3. Infrastructure Services [I]

ORNL shall provide a broad range of infrastructure services including Lightweight Directory Access Protocol (LDAP), Domain Name Service (DNS), Dynamic Host Configuration Protocol (DHCP), Network Time Protocol (NTP) and syslog aggregation. The selected Offeror shall work with ORNL to integrate the elements of the CMRS into this existing Infrastructure, leveraging the availability of the prescribed services. Reference Section 8 for additional information about the provided infrastructure.

2.4. Network Services [I]

This acquisition provides for connectivity to network devices and for services to support efficient use of system and network resources. Wide area networking is provided under a separate vehicle and is not the responsibility of the Offeror. Reference Section 8 for additional information.

2.5. Security Support [I]

This acquisition supports all standard Government security standards for an unclassified system as well as any additional standards imposed by ORNL and/or NOAA. Reference Section 9 for additional information.

3. Responsibilities of the Selected Offeror and Interface with ORNL

This section reserved.

4. Climate Modeling and Research System Requirements

4.1. Composition and Expected Lifetime [C]

The Climate Modeling and Research System (CMRS) shall comprise one or more supercomputer system(s) delivered to and installed at ORNL in Oak Ridge, TN,

The overall design of the CMRS shall balance the significant operational requirement for highly reliable, available and resilient system(s) with the need for a substantial increase in computing capability and capacity over current systems. The Offeror is encouraged to propose novel approaches to system design, system installation, system upgrades, and system administration and maintenance activities that will best meet all of these goals. [I]

An Offeror proposal that includes more than one physically separate supercomputer system shall address each of those as *subsystems*. Where necessary, this Technical Specification shall distinguish between the requirements for an individual subsystem and the collection of subsystems that comprise the CMRS. [I]

An individual subsystem can be maintained as more than one logical partition to assist in meeting the reliability, availability, and serviceability (RAS) requirements. The Offeror may alternatively choose to deliver distinct subsystems to address these same requirements. [I]

DRAFT

The Offeror shall describe their rationale for proposing physically distinct subsystems, or partitioned subsystems. [I]

In no case shall a subsystem or any partition of a subsystem contain fewer than 4,000 processing cores. [C]

Each CMRS subsystem (or the subsystem partition) shall be capable of running complex technical applications at the full size of that subsystem or partition. [S]

In all cases, an initial system must be delivered in Summer 2010, and entered into production before 1 October 2010. ORNL will evaluate the delivery schedule for risk. [C]

4.1.1. Single Phase System Delivery

The CMRS may be delivered as a single, integrated computer system, as a single system that is partitioned, or as two physically distinct subsystems. [I]

In this design, the Offeror shall describe a single installation and integration schedule such that the entire system shall enter production prior to 1 October 2010. [C]

In the event that the Offeror proposes a single phase system delivery, the minimum total peak double-precision (64-bit) floating-point performance of all compute partition(s) in the system available on 1 October 2010 shall be at least 750 TF. This is a performance target intended to provide a sense of scale; the evaluation criteria shall examine the benchmark performance (scalability and throughput) as well as the delivered throughput over the life of the CMRS. [S]

4.1.2. Multi-Phase System Delivery

The CMRS may be delivered as two separate computer systems. Each of these two deliveries may consist of either logically partitioned systems, or physically distinct subsystems. [I]

In this design, the Offeror shall make the first delivery such that the system (or subsystems) in the first delivery shall enter production no later than 1 October 2010. The second delivery shall be made such that the system (or subsystems) shall enter production no later than 1 October 2011. [C

In the event that the Offeror proposes a two-phase delivery, the minimum total peak double-precision (64-bit) floating-point performance of the compute partition(s) in the initial delivery available on 1 October 2010 shall be at least 250 TF and the minimum total peak double-precision (64-bit) floating-point performance of the compute partition(s) of the second delivery available on 1 October 2011 shall be at least 700 TF. This is a performance target intended to provide a sense of scale; the evaluation criteria shall examine the benchmark performance (scalability and throughput) as well as the delivered throughput over the life of the CMRS. [S]

The minimum length of production time that the two systems (or collection of subsystems) must overlap is to be no less than (NLT) 6 months to allow for a transition from the initial subsystem to the second subsystem. The decommissioning schedule for the initial system is otherwise left to the Offeror. System installations must meet all warranty and support requirements described in Section 11. The Offeror shall describe the anticipated impact to both users and operations due to the architectural differences between the two subsystems and how that relates to the appropriateness of the proposed transition period. The level of effort required to transition applications to the Phase 2 system will be evaluated. Systems that are architecturally compatible across the delivery Phases (same ISA and compilers) are strongly preferred. [S]

The delivery strategy for a multi-phase installation is shown in Figure 2. The blocks designated System 1 and System 2 may each individually consist of a single system, a logically partitioned system, or two

DRAFT

physically distinct subsystems. The best strategy for meeting the capacity and RAS requirements is left to the Offeror.

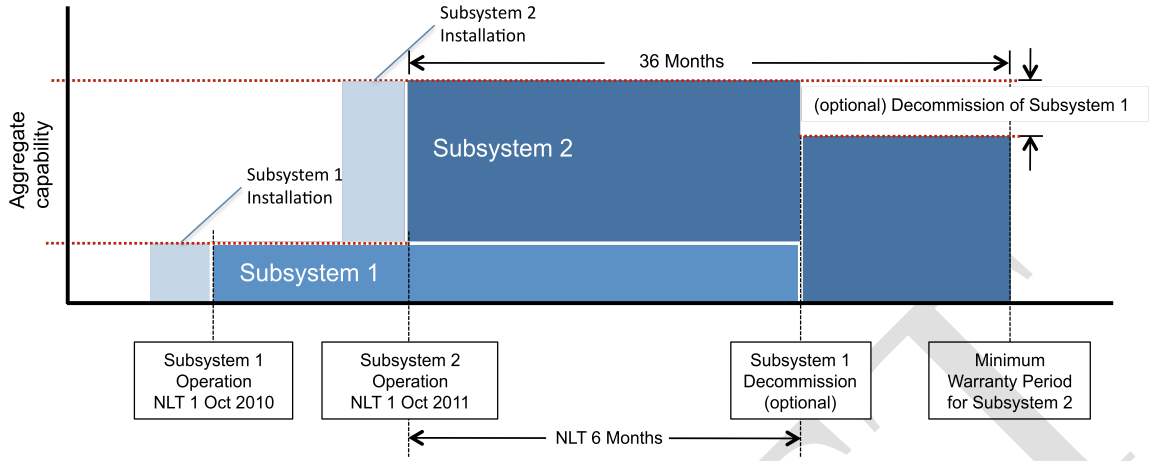

Figure 2. Example lifetime of a two-phase CMRS Installation

4.1.3. System and Subsystem Upgrades

The Offeror may propose a specific upgrade to any subsystem, regardless of whether the proposed solution is for a single- or multi-phased configuration. This upgrade may coincide with a processor, interconnect, or packaging option that will provide better performance, capacity, capability, or any other factor that will improve the value of the delivered subsystem. The Offeror shall describe the expected impact on application validation that may be associated with any proposed upgrades. [E]

An upgrade to a subsystem shall include a similar upgrade to the associated test and development system (TDS) (Reference Section 4.8).

4.2. System Computational Capacity [C]

The system computational capacity is based on the results of the workflow throughput benchmark. That benchmark is based on the GFDL high-resolution coupled climate model (CM2-HR). ORNL will assess the potential delivered capacity over the life of the delivered system(s) using the Offeror's delivery and warranty schedule and the Offeror's benchmark results. Reference the Benchmark Instructions for additional information.

4.3. System Computational Capability [C]

Each subsystem of the CMRS shall be a complete system as delivered, able to execute climate and weather modeling applications. Each CMRS subsystem (or the subsystem partition) shall be capable of running complex technical applications at the full size of that subsystem or partition. [S]

Benchmark codes will be employed to assist ORNL in assessing system performance. These benchmark codes will include climate and weather modeling codes. ORNL will also employ synthetic benchmark codes including both the C and Fortran 90 versions of the EPCC OpenMP Microbenchmarks (http://www2.epcc.ed.ac.uk/computing/research_activities/openmpbench/openmp_index.html), IOR, and mdtest. [I]

In the event that the Offeror proposes a single delivery as defined in Section 4.1.1, the minimum total peak double-precision (64-bit) floating-point performance of all compute partition(s) in the system

DRAFT

available on 1 October 2010 shall be at least 750 TF. This is a performance target intended to provide a sense of scale; the evaluation criteria shall examine the benchmark performance (scalability and throughput) as well as the delivered throughput over the life of the CMRS. [S]

In the event that the Offeror proposes a two-phase delivery as defined in Section 0, the minimum total peak double-precision (64-bit) floating-point performance of the compute partition(s) in the initial delivery available on 1 October 2010 shall be at least 250 TF and the minimum total peak double-precision (64-bit) floating-point performance of the compute partition(s) of the second delivery available on 1 October 2011 shall be at least 700 TF. This is a performance target intended to provide a sense of scale; the evaluation criteria shall examine the benchmark performance (scalability and throughput) as well as the delivered throughput over the life of the CMRS. [S]

4.4. Floating Point Standards

The Offeror's solution shall be compliant with 64-bit IEEE 754-2008 floating-point arithmetic. [S]

4.5. Node Requirements

4.5.1. General Requirements

For the purposes of this solicitation, a node is defined as the hardware on which a single instance of an operating system is running. The Offeror may further clarify this definition as needed, specific to their proposed architecture. [I]

All critical data paths internal to a processor shall be protected by an appropriate error correction code scheme such as single bit error correction, double bit error detection (SECDED). [S]

4.5.2. Service Nodes

For any service partition, any disk attached to a service node shall be configured for fault tolerance, e.g. a fault tolerant RAID type, and used to provide a service such as file I/O serving or booting. [S]

4.5.3. Login Nodes

For any service node that is designated as a login node, and will handle interactive user logins, compilations, and other typical login node services, these login nodes have additional requirements above and beyond the service and compute nodes. [I]

The login nodes are an important part of the development and production environment of the overall system. From a design perspective, they should be considered high-end servers. Login nodes must be able to see all secondary storage file systems available to the compute partition of the overall system, including the FS and LTFS. This is necessary to handle staging, production run monitoring, and code development processes. The login node(s) shall support 5 simultaneous parallel compiles using at least 8 compilation threads without degradation. ORNL will test this during acceptance testing by executing one compile of the CM2-HR application and then five simultaneous compiles of the CM2-HR application. The elapsed time for the completion of the five compiles can be no more than 10% longer than the time for the single compile. [S]

The login nodes shall have enough resources (memory, locally attached disk, processors) to facilitate this compilation mode. The network connectivity must be configured to facilitate visibility to the relevant file systems. [S]

The number of login nodes shall be sufficient to handle 25 simultaneous users, executing a mixture of compilation, job monitoring and file transfers. [S]

DRAFT

The minimum main memory capacity per processor socket for login nodes shall be at least 4 GB per core, regardless of the core-count on a specific node.

Login nodes should use the same type of processor or same family of processor as the compute nodes. If the login nodes are not binary compatible with the compute nodes, the Offeror shall describe the programming environment on the login nodes, how it differs from the programming environment on the compute nodes, the anticipated impact to such a configuration, and the mitigation strategy to be supplied by the Offeror. [S]

4.6. Node Memory Subsystem Requirements [C]

The minimum main memory capacity per processor socket shall be at least 2 GB per core, regardless of the core-count on a specific node. [C]

All paths from the main memory subsystem to the processor's integrated memory controller shall be protected by an appropriate error correction code scheme such as single bit error correction, double bit error detection (SECDED) logic. [S]

The memory subsystem shall provide support to protect against a fatal memory error due to the failure of any single memory chip in the subsystem. [S]

4.7. High Speed Interconnect Requirements [C]

The primary communication network (high speed interconnect) shall provide highly scalable performance. [I]

Sustained rates refer to the measured rate of data payload only and shall not include any hardware, protocol, and/or communication library overhead or any bandwidth needed for error correction or recovery. [I]

Injection bandwidth is defined as the bandwidth between a node's network interface card(s) and a high-speed network router. [I]

The Offeror shall propose an interconnect that meets the following requirements.

4.7.1. Interconnect Configuration

Each node shall have at least one dedicated network interface card (NIC) and connection to the high speed interconnect. [C]

4.7.2. Interconnect Bandwidth

For a mesh, torus or hypercube, the bisection bandwidth calculation shall be defined as the sum of the bisection bandwidths for each dimension. For all other topologies, including fat trees, the traditional minimum bisection definition shall be used. The bisection bandwidth of hybrid topologies is the bisection bandwidth of the highest level of the topology. This bandwidth shall be measured by pairing up all of the nodes such that all communication shall pass through the bisections of the platform with all nodes participating in the measurement. The measured bandwidth shall be the data payload only and shall not include any hardware, protocol overhead, and/or communication library overhead or any bandwidth needed for error correction or recovery. [I]

The Offeror shall describe the sustained bisection bandwidth and how that supports system performance requirements. [S]

DRAFT

4.7.3. Interconnect Topology Symmetry

If the topology is a mesh or hypercube, the topology shall be as symmetric as possible. The worst-case ratio of the bi-section bandwidth in any dimension should not be greater than 2:1. [S]

4.7.4. Link Bandwidth

The Offeror shall describe the sustained link bandwidth in each direction and bidirectional and how that supports system performance requirements. [S]

4.7.5. MPI Performance Requirements [S]

The Offeror shall specify the MPI ping-pong, zero byte latency between nearest neighbors (not in the same node). The Offeror shall specify the MPI ping-pong, zero byte latency between any pair of nodes.

The Offeror shall specify the MPI unidirectional bandwidth and bidirectional bandwidth as measured between two nodes.

The Offeror shall specify the maximum time measured or projected from all nodes for a full-scale floating-point (32-bit & 64-bit) MPI_ALLREDUCE (maximum and sum operations).

The Offeror shall specify the maximum time measured or projected from all nodes for a full-scale MPI_BARRIER (maximum and sum operations).

4.7.6. Bit Error Rate [S]

The uncorrected bit error rate (BER) of the interconnect shall be less than 1 bit in 10^{22} bits. The BER is for end-to-end data transfers within the high speed interconnect. At a minimum, all messages shall have an end-to-end 32-bit or greater CRC and all message headers shall have a separate 16-bit or greater CRC.

4.8. Test and Development System [C]

The Offeror shall include a Test and Development System (TDS) of about 2% the size of the CMRS. The size of the TDS shall be sufficiently large that all architectural features of the larger system are replicated. This specifically includes interconnect mechanisms that are deployed within the larger system. [C]

The TDS should physically span a minimum of two racks although each rack need not be fully populated. [E]

The TDS shall be delivered prior to the delivery of the associated CMRS system. The TDS shall remain operational at least through the same service term as the associated CMRS system.

A separate test system is required for each CMRS subsystem if they are architecturally distinct. The TDS shall be comprised of identical components (hardware and software), software tools, operation and support to the associated CMRS including a scaled version of the fast scratch (FS) file system as described in Section 5.

The TDS shall be independent of the CMRS except for the following elements:

- connectivity to the LTFS (as described in Section 5) [S]
- connectivity to the FS (in the case of a globally accessible FS) [S]

The batch and scheduler environment of the TDS and CMRS can be unified [E].

DRAFT

5. I/O Systems and Data Management

5.1. Fast Scratch (FS) File System

The Fast-Scratch (FS) file system is the high-bandwidth parallel file system that is visible and accessible by all processors within a given batch job. [I]

The FS will be regularly purged based on multiple factors including percentage of free space remaining, age of individual files, patterns of use by specific projects or users, and other considerations. The FS will not be backed up. [I]

The FS shall support no less than 100 million files and individual files that are up to 10 TB in size. [S]

It is strongly preferred that the FS is available as a single parallel file system to all CMRS and TDS subsystems. [S]

FS availability is incorporated into CMRS availability as a whole and is detailed in Section 7.7 [S]

FS mean time to interrupt and mean time to failure is incorporated into CMRS availability as a whole and is detailed in Section 7.4.1 and Section 7.4.2 respectively. [S]

The FS shall support quotas for both reporting and enforcement purposes. Quotas are defined as a mechanism to track total file system usage by user or group (reporting) and to enforce a configurable maximum capacity per user or group (enforcement). Enforcement of quotas shall be configurable separately from quotas for reporting purposes. [S]

5.2. Long-term Fast Scratch (LTFS) File System

The Long-term Fast Scratch (LTFS) file system is the staging area for users to temporarily place files that are frequently manipulated. The LTFS will be regularly purged based on multiple factors including percentage of free space remaining, age of individual files, patterns of use by specific projects or users, and other considerations. The LTFS will not be backed up. [I]

The LTFS file system shall support no less than 100 million files and individual files that are up to 10 TB in size. [S]

Scientific model data for the CMRS will be primarily accessible through the LTFS. The LTFS shall remain accessible in the absence of the computational platform. The LTFS shall be a separate file system constructed of completely independent components from the FS. The LTFS shall be designed with fault tolerant features including I/O server failover and path failover (Device-Mapper (DM) Multipath, Redundant Disk Array Controller (RDAC), or similar technologies). Additional requirements for fault tolerance are detailed in Section 5.3. [S]

The LTFS shall be highly available. Availability of the LTFS is defined as the ability to access the LTFS for reading, writing, creating, and deleting any or all files on the LTFS at the bandwidth and metadata rates specified for the LTFS. Additional requirements for availability are described in Section 7.7. [I]

The LTFS mean time to failure is described in section 7.4.2. [S]

The LTFS shall support quotas for both reporting and enforcement purposes. Quotas are defined as a mechanism to track total file system usage by user or group (reporting) and to enforce a configurable maximum capacity per user or group (enforcement). Enforcement of quotas shall be configurable separately from quotas for reporting purposes. [S]

The LTFS shall be configured with local data transfer nodes (LDTNs). The LDTNs are dedicated nodes that shall transfer data between the LTFS and the FS systems. LDTNs shall mount both the FS and LTFS systems to support these data transfer operations. LDTNs shall be shared between the FS and LTFS

DRAFT

systems. The Offeror shall provide sufficient LDTNs to support the minimum bandwidth requirements as described in section 5.3. [S]

It is anticipated that no fewer than 16 LDTNs shall be required. [I]

The LTFS shall provide Quad Data Rate (QDR) InfiniBand (IB) or 10 Gigabit Ethernet connectivity for no less than eight remote data transfer nodes (RDTN). RDTNs are data transfer nodes that mount the LTFS file system to execute remote data transfers via GridFTP, BBCP, or other data transfer protocols. The RDTNs are not the responsibility of the Offeror. RDTNs will be provided, configured, administered, and maintained by ORNL. [S]

The LTFS shall be mounted on all login and batch nodes on the CMRS. The LTFS need not be mounted on the compute nodes of the CMRS. [S]

A primary function of the LTFS is to provide a staging area for remote data transfers. This workload includes large data movement from the FS to the LTFS of computational products (model output) generated from application runs on the CMRS. The Offeror shall provide performance targets for data transfers from the FS to the LTFS using a parallel copy mechanism of their choosing using a data set with files whose size conforms to the distribution shown in Figure 3.

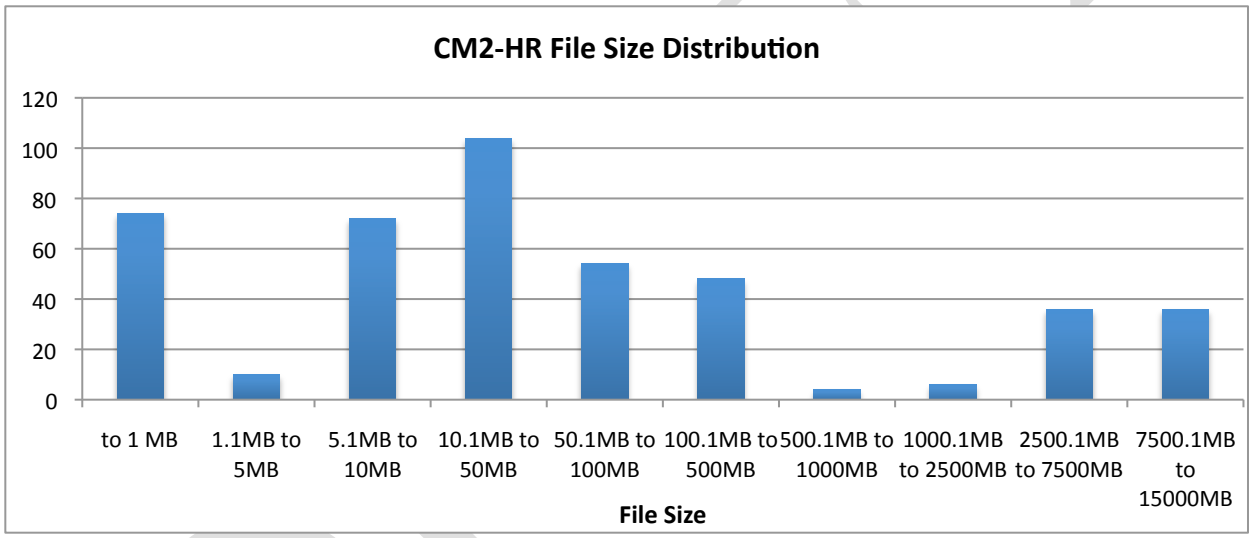

Figure 3. CM2-HR File Size Distribution

5.3. FS and LTFS File System Bandwidth and Capacity Requirements

The performance characteristics and capacity of both the FS and LTFS systems are a function of the total compute resource of the CMRS (relative to core count). [I]

For multi-phase CMRS deliveries as defined in Section 4.1.2, the LTFS and FS must be revised to ensure that they meet the requirements of each phase. [S]

These required performance characteristics are based on historical data volumes produced as a function of core counts on existing systems projected to the CMRS. In 2008 1.5 PB of data was generated on existing NOAA systems and will be used as the baseline for this formulation. [I]

The following definitions are used in deriving the required data rates for the FS and LTFS systems. [I]

V_E = Volume of data generated on existing systems in 1 year (GB/year)
N_E = Number of cores on existing systems
N_{CMRS} = Number of cores on the CMRS

D_E = Production data rate observed on existing systems in GB/ CPU-Hour (CPU=core)
D_{CMRS} = Production data rate projected on the CMRS in GB/ CPU-Hour (CPU=core)
R_E = Production data rate observed on existing systems in GB/hour
R_{CMRS} = Production data rate projected on the CMRS in GB/hour

Equations 1 and 2 describe the algorithms for converting the 2008 baseline system data production to average continuous total and per-CPU data generation rates, per hour, on the baseline system. The results of that calculation are shown in Figure 4.

$$R_E = \frac{V_E}{365 * 24} \tag{1}$$

$$D_E = \frac{R_E}{N_E} \tag{2}$$

N_E (cores)	V_E (GB/Year)	R_E (GB/hour)	D_E (GB/CPU-Hour)	t_E (sec)
8000	1,500,000	171.23	0.0214	11280

Figure 4. (Example) Calculating R_E and D_E using 2008 Baseline Data

The following additional definitions are used in deriving the required data rates for the FS and LTFS systems. [I]

TTPR = Total throughput ratio
t_{CMRS} = benchmark job work stream run time on proposed CMRS system as projected by the Offeror (see benchmark instructions).
t_E = benchmark job work stream run time on existing system
F_{sw} = Scale factor to account for potential software improvements expected in 2 years, The Offeror shall assume that F_{sw} = 2.0.
S_c = Capability scale factor

$$S_c = F_{SW} * \frac{t_E}{t_{CMRS}} \tag{3}$$

$$TTPR = \frac{R_{CMRS}}{R_E} = F_{SW} * \frac{t_E}{t_{CMRS}} * \frac{N_{CMRS}}{N_E} = S_C * \frac{N_{CMRS}}{N_E} \tag{4}$$

To compute the factor by which storage requirements scale up from the baseline, calculate $(TTPR)^{0.7}$. To calculate scaling on a per-core basis, divide $(TTPR)^{0.7}$ by (N_{CMRS} / N_E). To calculate the projected per-core data generation rate on the CMRS, multiply that result by the current data generation rate per core, D_E. The result is a per-core data generation rate which scales as the inverse 0.3 power of the core count:

$$D_{CMRS} = D_E * \left(F_{SW} * \frac{t_E}{t_{CMRS}} \right)^{0.7} * \left(\frac{N_E}{N_{CMRS}} \right)^{0.3} \tag{5}$$

DRAFT

As an example, the FS and LTFS bandwidth requirements using a hypothetical system with 50,000 cores that has a workstream runtime t_{CMRS} that is the same as that on the existing system so that S_{eq}=2.0. The data production rate D_{CMRS} in GB/cpu-hour for this example as shown by Equation 6 and 7, can be seen in Figure 5.

$$D_{CMRS} = D_E * (S_C)^{0.7} * \left(\frac{N_E}{N_{CMRS}}\right)^{0.3} \tag{6}$$

$$R_{CMRS} = D_{CMRS} * N_{CMRS} \tag{7}$$

N_E (core count)	N_{CMRS} (cores on CMRS)	S_c (capability scale factor)	R_{CMRS} (GB/hour)	D_{CMRS} (GB/cpu-hour)
8,000	50,000	2.0	1003	0.02

Figure 5. Data production rate for a 50K core system with S_c=2.0

Data is written to the file storage system during a fraction of the total execution time of the job. The bandwidth required by the storage system to absorb this data is then derived with the following definitions in Equation 8 and Figure 6.

D_{IO} = I/O Duty cycle presented as a fraction of the program execution time spent in I/O (5% or 0.05)

D_{FS} = The I/O bandwidth in GB/sec of data from the computer to the file system.

$$D_{FS} = \frac{R_{CMRS}}{D_{IO} * 3600} \tag{8}$$

R_{CMRS} (GB/hour)	D_{IO}	D_{FS} (GB/s)
1003	0.05	5.57

Figure 6. Bandwidth assuming data is written during 5% of total program execution time

The file system sees multiple accesses to this data from the executing program and the data movers. Hence the cumulative bandwidth of the file system (BW) to sustain this access load is expressed as shown below in Equation 9 and corresponding Figure 7 with the following definitions: [I]

F_{IO} = The scale factor representing the number of concurrent data movements on the appropriate file system

BW = Bandwidth of the file system in GB/s

$$BW = D_{FS} * F_{IO} \tag{9}$$

File System	D_{FS} (GB/s)	F_{IO}	BW (GB/s) of the file system
FS	5.57	3	16.72
LTFS	5.57	1.5	8.36

DRAFT

Figure 7. Total modified bandwidth required by the hypothetical file system

The data writes to the LTFS and both read and write operations to/from the FS will occur in bursts. The data will stream from the LTFS to the originating centers, GFDL, ESRL or NCEP. The estimated bandwidth required for the LTFS is reduced since the streaming reads associated with the WAN transfers will be less demanding. This is equivalent to adjusting the duty cycle upwards for the WAN transfers. This is reflected by weighting that transfer by 0.5 in the factor F_{IO} (concurrent data movements) for LTFS.

The FS is required to retain the computational data (model output) for a period of four days. The LTFS is required to retain the computational data (model output) for a period of 18 days. [S]

The estimated storage size for the file systems is given below in Figure 8 using Equations 10 and 11 and the following definitions:[I]

T_l	=	Retention time for data on Long Term File System (LTSFS) in days
T_a	=	Data aging on Long term file system in days
T_s	=	Data retention time on Fast Scratch File System (FSFS) in days
C_{FS}	=	Size of FS using example of 50,000 cores on the CMRS influenced by T_s
C_{LTFS}	=	Size of LTFS using the example of 50,000 cores on the CMRS influenced by T_a and T_l

$$C_{FS} = R_{CMRS} * 24 * T_S * 3 \tag{10}$$

$$C_{LTFS} = R_{CMRS} * 24 * (T_l * T_a) * 3 \tag{11}$$

T_l (days)	T_a (days)	T_s (days)	R_{CMRS} (GB/hour)	C_{FS} (GB)	C_{LTFS} (GB)
4	14	4	1003	288,945	1,300,257

Figure 8. Estimated size of FS and LTFS based on hypothetical 50K cores

In this example the space requirements are those anticipated for actual files. The Offeror is required to provide the file systems with adequate excess capacity to allow the file systems to operate at no more than 75% of total available storage capacity. [S]

The LTFS must conform to the following performance and capacity specifications: [S]

1. 3,000 metadata (file create) operations per second as measured by the mdtest benchmark. The Offeror should reference the benchmark instructions for information about how to run the mdtest benchmark.
2. Usable (formatted) capacity as specified in Equation 11
3. File system bandwidth as specified in Equation 9. File system bandwidth shall be measured from the LDTNs via the Long Term Fast Scratch Benchmark (LTFSB). This bandwidth must be achieved concurrently on both the LTFS and FS systems from the LDTNs. This is an application benchmark that measures sustained, delivered bandwidth. This is expected to differ considerably from the LTFS peak bandwidth. The Offeror shall specify the offered LTFS peak bandwidth and how it differs from that delivered to the LTFSB.

The FS must conform to the following minimum performance and capacity specifications: [S]

1. 3,000 metadata (file create) operations per second as measured by the mdtest benchmark
2. Usable (formatted) capacity as specified in Equation 10

DRAFT

3. File system bandwidth as specified in Equation 9. File system bandwidth will be measured from the computational nodes via the FSB (Fast Scratch Benchmark). This is an application benchmark that measures sustained, delivered bandwidth. This is expected to differ considerably from the LTFS peak bandwidth. The Offeror shall specify the offered FS peak bandwidth and how it differs from that delivered to the FSB.
4. File system bandwidth as specified in Equation 9. File system bandwidth will be measured from the LDTNs via the LTFSB.

The minimum file system bandwidth for both the FS and LTFS must be maintained throughout the life of the file system regardless of aging issues such as fragmentation. [S]

5.4. Common Data Management Requirements

References to RAID 6 indicate a requirement for either RAID 6 or an equivalent data protection and recovery mechanism[1]. [I]

The storage systems for both FS and LTFS shall conform to the following requirements:

The storage system shall support and be configured with tiers that are protected by RAID 6. The storage system shall support, and be configured with, parity check on read. Parity check on read is defined as all components of a RAID stripe are read and the parity data is confirmed to be consistent. Mismatches must result in a failed read operation and must be reported via a logging mechanism. Timeouts for slow disk may still be employed in conjunction with parity check on read but must also be reported via a logging mechanism. [S]

The storage system shall not possess any single points of failure including controllers, enclosure bays, power distribution units, and disks. [S]

The storage system shall support "hot-swapping" of all components including power supplies, fans, controllers, disks, and drive enclosure bays. [S]

The storage system shall support dynamic sector repair and background RAID verification and must be configured with these features enabled. [S]

The entire storage system shall have a mean-time between unscheduled interrupt of greater than 1 month when measured over a three month period. [S]

The storage system shall be configured to maintain the required file system bandwidth in the presence of concurrent rebuilds across up to 10% of the available storage tiers. In no case shall a rebuild require greater than 24 hours to complete (even in the presence of active I/O operations from the file system). It is desired to have a mechanism to prevent new data from being written to degraded logical unit numbers (LUNs). [S]

The storage system shall be configured with a minimum of 1 hot spare drive for every 4 RAID 6 tiers. [S]

A storage system which supports and that is configured with InfiniBand connectivity as well as the SCSI remote direct memory access (RDMA) protocol (SRP) is preferred. [S]

The storage system controllers shall support both automatic and manual failover. Controller level caching must either support mirroring or be disabled such that automatic failover via DM multipath or RDAC shall not result in data corruption. It is desired to have a mechanism to prevent new data from being written to LUNs with degraded performance due to controller failure. [S]

[1] Recovery information, such as parity for a tier, must be replicated on a minimum of two independent disk drives.

DRAFT

The Offeror shall provide a uniform power distribution design for all storage systems. Power distribution design shall be based on two independent inputs. [S]

Features including parity-declustering for fast rebuilds and T10-DIF (Data Integrity Field) or other end-to-end data integrity mechanisms for enhanced reliability are desired. [S]

The storage system for the FS shall be configured with either serial-attached SCSI (SAS) or fibre-channel (FC) enterprise class drives (10K RPM minimum). [S]

The storage system for the LTFS may utilize SAS, FC, or serial-ATA (SATA) drives. [S]

6. Software

The Offeror shall provide a complete software suite for the TDS and CMRS that includes, at a minimum, the following elements: [S]

- The CMRS shall be provided with a UNIX or Linux native 64-bit operating system, system interfaces, shell and utilities that comply with POSIX standards. The Offeror's proposal shall describe what system software is provided and describe compliance and non-compliance with POSIX standards.
- Resource and workload management software - reference Section 6.2
- Accounting software
- System monitoring software - Reference Section 7
- File system(s)
  The Offeror's proposal shall describe the filesystems to be provided with and supported on the proposed CMRS.

The CMRS shall be provided with a full-featured user and program development environment to support MPI, OpenMP and hybrid programming models. This shall include

- ANSI Standard Fortran 90/95, C and C++ compilers with OpenMP support
  - Compiler licenses to enable no less than 200 concurrent compiler invocations distributed among 25 unique user ids
  - Macro-preprocessors
  - MPI 1.2, MPI-2 I/O and MPI-2 one-sided communication libraries[S]. Full MPI-2 support is preferred [E].
  - Optimized serial and parallel numerical and scientific libraries
- Tools for placement and pinning of threads and processes to specific cpu cores.
- Libraries
- Serial and Parallel source-level debugger(s)
- Performance analysis tools including
  - Profiling
  - Tracing
  - Access to processor hardware counters
  - I/O
- Other programming environment software

The Offeror shall provide a detailed description, including compliance with relevant standards, of the user and program development environments to be provided with the CMRS. The Offeror should note particular attributes that may aid the NOAA users community in optimizing their use of the proposed CMRS. [S]

DRAFT

6.1. Site Support Advocate

The Offeror shall designate both a primary and alternate ORNL Site Support Advocate from within the Offeror's software support organization. The ORNL Site Support Advocate shall be responsible for providing fast-path escalation of CRMS software problem issues, maximizing quality of service, and minimizing time to resolution. [S]

6.2. Workload and Resource Management

6.2.1. Workload Management [I]

ORNL owns a site license for the Moab Cluster Suite® (http://www.clusterresources.com/products/moab-cluster-suite.php) for the CRMS. The Cluster Suite includes:

- Workload Manager® - a policy-based job scheduler.
- Cluster Manager ® - a cluster management interface, monitor, and reporting tool for Workload Manager.
- Access Portal – an end-user job submission and management tool that works with Workload Manager.

ORNL will provide the Offeror access to the Moab Cluster Suite license through FY2011 at no cost. The Offeror is responsible for costs for the workload manager beyond this date, or may propose their own solution for the entire period of the contract.

ORNL has significant operational and administrative experience managing scheduling policy for computer systems using the Moab Cluster Suite. ORNL will determine and define scheduling policy for the CMRS. If the Offeror chooses to manage the CMRS with the Moab Cluster Suite, ORNL will work with the Offeror to effect scheduling policy using Moab.

6.2.2. Resource Management

The Offeror shall provide, optimize, and maintain a resource manager for the CMRS that is compatible with the Moab Workload Manager (or their alternate Workload Manager). That resource manager may be one of the following products: TORQUE Resource Manager, OpenPBS, PBSPro, Sun Gridengine (SGE), SGE Enterprise Edition (SGEE), LoadLeveler, LSF, BProc/Scyld, Scalable System Software (SSS-RM), or Quadrics RMS. [S]

6.3. Software Suite

The Offeror shall provide the following software suite on the CMRS to support the scientific goals of the anticipated user community. This software consists of both free and commercially available software that includes parallelized and optimized numeric libraries. [E]

The selected Offeror shall work with ORNL to compile and distribute new versions of the following software packages and programs on the CMRS as they become available. [E]

- ANTLR (http://www.antlr.org)
- Bbcp (http://www.slac.stanford.edu/~abh/bbcp/)
- Bbftp (http://doc.in2p3.fr/bbftp/)
- BLAS (http://www.netlib.org/blas)
- Earth Systems Modeling Framework (http://www.esmf.ucar.edu)
- Environment Modules (http://modules.sourceforge.net/)
- Ghostview (http://pages.cs.wisc.edu/~ghost/)

DRAFT

- Globus toolkit (http://www.globus.org)
- GNU binutils (http://www.gnu.org/software/binutils)
- GNU coreutils (www.gnu.org/software/coreutils)
- GNU make (http://www.gnu.org/software/make/make.html)
- GNU tar (http://www.gnu.org/software/tar/)
- GrADS (http://www.iges.org/grads/)
- GRIB libraries and utilities (http://www.cpc.ncep.noaa.gov/products/wesley/wgrib.html)
- HDF5 (http://www.hdfgroup.org)
- Heirloom cpio (http://heirloom.sourceforge.net)
- ImageMagick (http://www.imagemagick.org)
- Lapack (http://www.netlib.org/lapack)
- libxml (http://xmlsoft.org)
- Ncview (http://meteora.ucsd.edu/~pierce/ncview_home_page.html)
- netCDF and udunits libraries (http://www.unidata.ucar.edu/software/netcdf and http://www.unidata.ucar.edu/software/udunits)
- netCDF Operators (http://nco.sourceforge.net)
- Octave (http://www.octave.org)
- Perl (http://www.perl.com)
- Petsc (http://mcs.anl.gov/petsc)
- Python (http://www.python.org, http://numpy.scipy.org, http://dirac.cnrs-orleans.fr/plone/software/scientificpython)
- R (www.r-project.org)
- Ruby (http://www.ruby-lang.com)
- Tau (http://www.cs.uoregon.edu/research/tau/home.php)
- Zlib (http://www.zlib.net)

The Offeror shall provide licenses for the following tools or functional equivalents and shall support the use of floating or enterprise license managers as appropriate. [S]

- Allinea DDT (http://www.allinea.com/) for up to 1,000 cores of the CMRS and TDS.
- IDL, version 7.1 or later (http://www.ittvis.com/ProductServices/IDL.aspx) for a single node (not subsystem or partition) of the CMRS or TDS.
- Matlab, release R2009B or later (http://www.mathworks.com/) for a single node (not subsystem or partition) of the CMRS or TDS.
- NAG, including at a minimum the Numerical Components C, Fortran, Fortran 90, Toolbox for Matlab (http://www.nag.co.uk) for a single node (not subsystem or partition) of the CMRS or TDS.
- TotalView (http://www.totalviewtech.com) for up to 500 cores of the CMRS and TDS.

In the event that separate architectures (not two systems of the same architecture) are proposed, as defined in Section 0, the Offeror shall ensure that the provided license structure supports both architectures for simultaneous use up to the minimum specified core count. The intent of this requirement is to ensure that separate systems that are based on different processor families each provide a suitably robust software tools license structure for disparate systems. [S]

All updates of software installed or used on the system shall be managed through a rigorous configuration management (CM) process for the lifetime of the system. Reference Section 9 for additional information on the CM process. [S]

DRAFT

6.4. Checkpoint/Restart

The NOAA R&D user community very frequently employs user-initiated checkpoint/restart. This practice is expected to continue. System initiated checkpoint-restart is not a requirement. As applicable, the Offeror shall describe system support for and user compliance requirements for checkpointing of jobs. [E]

7. Resiliency, Reliability, Availability, Serviceability

7.1. RAS System Initialization and Reboot

A full system initialization that includes 100% of the service partition components and no less than 99.9% of the compute partition components shall take no more than 60 minutes. Depending on the Offeror's solution, initialization may include the system power-on sequence. Initialization does not include file-system check or full memory check, nor a power-on sequence for either the FS or LTFS storage systems. [S]

A single node or any subset of nodes shall be able to be rebooted without affecting the balance of the running system. If the Offeror's solution has specific conditions that may preclude this, those conditions shall be described in the Offeror's proposal. [S]

A full memory check shall be available and be a configurable option on a per node basis. [S]

7.2. RAS System Diagnostics and Maintenance

The RAS system shall provide diagnostic capability sufficient to isolate problems down to, at a minimum, a Field Replaceable Unit (FRU) or units. [S]

All FRUs whose replacement requires shutdown of more than a single cabinet shall be identified. [S]

FRUs shall be replaceable during production uptime and be able to be returned to service after replacement. [S]

The RAS system shall track major hardware components automatically, preferably by serial number or some similar automatically identifiable characteristic. Detected hardware configuration changes shall be logged. Operators shall be able to log their observations and service actions by identifiable hardware component. [S]

7.3. RAS System Event Monitoring and Archival

The RAS system shall provide both real-time monitoring and archiving of historic events at the granularity of the individual component. [S]

Real-time monitoring of system components shall be operator initiated or configurable automatic monitoring. [S]

The historic archival of RAS events shall be accessible for a site configurable period of time. [S]

The historic data shall be separated into recent history and archived data, in order to facilitate real time analysis of recent history. All data shall be exportable in a fully documented format to local site systems for archival storage. [S]

The RAS system shall provide the capability to configure the level and content of the output of system monitoring, including output from sensors monitoring temperatures, voltages, current draws and fan speeds. All data shall be reported at the finest possible resolution. The physical meaning, accuracy and precision of sensor data shall be fully documented. [S]

DRAFT

All operations status transitions of all nodes shall be recorded. [S]

Operations status is one of the following three mutually exclusive values: [I]

- Production Uptime (ready to perform computations for production users),
- Scheduled Downtime (not in Production time for scheduled reasons approved through the CM process),
- Unscheduled Downtime (not in Production time for unscheduled reasons). Unscheduled downtime that is directly attributable to events that are out of the control of the Offeror is considered NULL time, and does not materially impact the calculation of system metrics.

Recording shall be as automatic (e.g. system-determined transitions) and convenient (e.g. human-determined transitions) as possible. [S]

7.4. System Mean Time To Interrupt and System Effectiveness Level Requirements

7.4.1. System Mean Time Between Interrupt

The Offeror shall calculate the System Mean Time Between Interrupt (SMTBI) for each subsystem. [S]

A system interrupt is defined as any hardware or system software error or failure, or cumulative errors or failures over time, that result in more than 1% of the system being unavailable at any given time during Production Uptime. This includes loss of access to any of the file systems. [S]

The minimum System Mean Time Between Interrupt (SMTBI) shall be at least 200 hours. [S]

The Offeror shall describe the process for estimating the SMBTI of each proposed subsystem. [I]

7.4.2. System Mean Time Between Failure Requirements

The Offeror shall calculate the hardware System Mean Time Between Failure (SMTBF) for each subsystem. [S]

This calculation focuses on system wide reliability, or the reliability of the system when functioning as a single unit. Individual components configured to operate collectively in a redundant manner may be considered an individual component with the appropriate error rate for the purposes of this calculation. (Example: a set of redundant power supplies) [I]

The minimum SMTBF shall be at least 30 hours. The SMTBF shall include the FS file system. [S]

It is desired that the LTFS file system demonstrate an MTTF of not less 30 days. The Offeror shall state the MTTF of the LTFS file system. [S]

The Offeror shall describe the process used for estimating the SMBTF of each proposed subsystem. [S]

7.5. Resiliency

The CMRS shall possess resiliency features such that user applications can continue to run to successful completion despite system faults. The underlying hardware/software/infrastructure shall be able to recover gracefully from faults and fails by rapid re-provisioning of resources or similar mechanism(s). [S]

The Offeror shall describe features of their hardware monitoring system(s). The Offeror shall describe their methods for executing predictive failure analysis. The Offeror shall describe their anticipated preventative maintenance (PM) schedule, and how predictive failure analysis (PFA) is integrated into or affects that schedule. The PM schedule shall consider the impact to users, including the executing of PM

DRAFT

during normal business hours. The Offeror shall describe features of the proposed system that support an ability to migrate existing jobs. [S]

The Offeror shall describe features of the LTFS and FS that improve its resiliency and support rapid recovery from hardware failures. The Offeror shall describe features of the proposed system that support an ability to migrate data from failing hardware components within the LTFS or FS, or to migrate data from one storage device to another to support planned maintenance operations. [S]

The Offeror shall provide estimates of the time required to recover from various file system faults due to hardware and/or software faults. These calculations shall include, at a minimum, analysis of the time to recover from the loss of individual disks, disk arrays, storage servers and the fabric components. [S]

The Offeror shall describe the resiliency features of the proposed LTFS and FS that ensure the accuracy of read/write operations, guard against single point of failure, and protect file system integrity. [S]

The CMRS shall be capable of scheduling/queuing jobs despite a loss of a portion of system resources, of rerunning batch jobs without user intervention and preventing job "timeout" due to long running jobs when the system is in a degraded mode. [S]

The Offeror shall identify any single point of failure in the CMRS. The CMRS shall support continued interactive login during degraded mode. Offeror shall describe the degree to which interactive nodes are decoupled from compute nodes. [S]

The CMRS should provide failover to binary-compatible processor architectures that are running the same operating system and level whenever a set of resources fails. [E]

The Offeror shall describe the features of the proposed solution that enhance resiliency. [S]

7.6. Reliability

The CMRS shall be comprised of components, hardware and software, and of a design selected to maximize reliability. [S]

System resources shall support continuous monitoring and a mechanism whereby errors are reported for analysis and subsequent action. Reliability shall be examined during Acceptance via tests of numerical reproducibility, numerical accuracy and run-time variability. At ORNL's discretion, similar tests, typically of a shorter duration, will follow system outages involving hardware/software upgrades. ORNL reserves the right to verify proper function at any time. [S]

Data integrity shall be demonstrated in the ability to reproduce results on the same hardware, data movement and data storage. The CMRS shall have the integrity that no more than one bit in every 10^15 bits shall be impacted by silent data corruption. [S]

The CMRS requires bit-wise reproducibility of answers by two identical jobs running on the same system. In the event that a bit-wise reproducibility problem is discovered, either before or after acceptance of the system, the Offeror shall promptly notify ORNL. Resolution of a reproducibility problem requires that the root cause be provided and appropriate mitigation of the problem shall occur. The Offeror shall keep ORNL informed as it works to identify the root cause. At ORNL's discretion, ORNL will dual-run its work to assure bit-wise reproducibility. If a bit-wise reproducibility problem is not resolved within six (6) weeks of the initial discovery, the second run of the dual-run shall be marked as downtime. [C]

The CMRS shall deliver job run-time variability of not more than +/- 5% for its benchmark codes under the following condition: at least 90% of CMRS resources are utilized by running multiple copies of the benchmark codes in a random manner (so as to engage a varied set of system resources for the codes).

DRAFT

Run-time variability is influenced by input data. Benchmark codes will employ fixed input datasets and comparisons will be made among benchmarks running identical datasets during a given test period. [S]

The Offeror shall describe the features of the proposed solution that enhance reliability. [S]

7.7. Availability

7.7.1. System Effectiveness Level

The system Effectiveness Level (EL) is defined in Equation 12 as:

$$EL = \left(\frac{PU - SD - NULL}{PU + UD - SD - NULL} \right) \tag{12}$$

where, as described in Section 7.3, PU is production uptime, SD is scheduled downtime, UD is unscheduled downtime, and NULL is time that is lost due to factors that are out of the Offeror's control, such as a utility outage affecting the facility. NULL time will be determined and reported by ORNL. This same formula is to be used for the calculation of the availability of CMRS systems, and LTFS and FS elements. [I]

The CMRS shall exhibit an effectiveness level of not less than 96% calculated on a monthly basis. The effectiveness level of the CMRS shall include the FS file system. The system shall provide for continuous system resource monitoring and logging to substantiate availability. These logs shall be made available to ORNL upon request. The Offeror shall provide monthly summary reports describing monthly availability. [S]

The LTFS file system shall demonstrate an effectiveness level of not less than 98 %. The Offeror shall state the proposed effectiveness level of the LTFS file system recognizing that an effectiveness level of over 98% is desired. [S]

The Offeror shall describe features of the proposed solution that enhance availability. The Offeror shall describe a preferred schedule of downtime for scheduled maintenance, security management, and preventative maintenance. [S]

7.8. Serviceability

A high level of serviceability is required to meet ORNL's availability needs. The CMRS design shall incorporate features supporting individual component service leaving the remainder of the resource available to perform work. Design features that permit work to continue in a degraded mode while service is underway are significantly more preferable to outages. Features that support scheduled maintenance rather than time-critical repairs are desired. [I]

The CMRS shall support hot swap of field replaceable components without impacting the operating state of the system. The Offeror shall describe the features of the proposed solution that enhance serviceability. [S]

The use of serviceable replacement parts that are warranted as new is acceptable. [I]

8. Infrastructure and Networking Services

8.1. Infrastructure

The CMRS shall support LDAP as a method for authorization and authentication. The CMRS shall also support PAM as an additional authentication method. The CMRS shall support configuring syslog to log in to multiple destinations outside of the CMRS. [S]

DRAFT

ORNL will use Nagios or equivalent as a monitoring solution. The Offeror shall work with ORNL to integrate the elements of the CMRS into the existing monitoring implementation. This requires SNMP queries. [S]

8.2. Wide Area Network Connectivity [I]

The CMRS will connect to the NOAA National R&D network footprint, and other necessary peerings via diverse and redundant high bandwidth, low latency connections. At the time of award, the anticipated external connectivity from ORNL to the research and education community will include, at a minimum, multiple connections to ESnet at 10 GigE. These connections will be to both Layer 3 services and to the Layer 2 ESnet Science Data Network (SDN) services. Other high bandwidth connections, for example, to Internet2, are available, but paths between ORNL and other NOAA facilities are expected to transit ESnet services. While these services will initially be based on 10 GigE technologies, it is expected that these services will be upgraded to 100 GigE services during the life of this contract. Wide Area Network connectivity is not a responsibility of the Offeror. [I]

The Offeror is responsible for providing the necessary interfaces within the LTFS to support connections to the RDTNs as described in Section 5.2. [S]

8.3. Local Area Network Connectivity

The demarcation point (demarc) between the CMRS and the broader ORNL network is defined by an Offeror-specified number of network interfaces using IEEE 802.3ae-compliant 10 Gigabit Ethernet interfaces. ORNL will provide the necessary number of physical 10 Gigabit Ethernet ports as specified by the Offeror to facilitate the movement of data among the elements of the CMRS and from the CMRS to an external location. ORNL will provide the network equipment described herein. ORNL will provide the fiber optic or copper cable connections from this network equipment to the Offeror's equipment. [I]

The Offeror is responsible for any interconnect fabric (switches, NICS, and fiber optic or copper cable) between the LTFS, FS, and CMRS that do not use 1- or 10 Gigabit Ethernet. [S]

The network infrastructure provided by ORNL/NCCS is a hierarchical design that provides for expansion, performance and redundancy. The core will consist of two Layer 2/3 network devices capable of VRRP, filtering, 10GigE connections with an upgrade path to 40/100GigE, and OSPF v2/3. The core devices will consist of at least 32 10GigE ports and be capable of expanding to 256 or more 10GigE ports. The distribution devices will consist of at least 50 10GigE ports per chassis. Top of rack switches will be used for Layer 2 connectivity for infrastructure/services. These switches will have at least 2 10GigE uplinks, and support 48 copper interfaces (non POE). If a fabric extension module or similar technology is used that only supports 1 Gigabit Ethernet, 2 top of the rack management switches will be used having at least 48 ports (non POE) that will support 10/100/1000 copper connections. [I]

The Offeror's equipment shall attach to the ORNL-provided distribution switches for TCP/IP based connections to the NOAA enclave. [S]

Above the core, two firewalls with 10GigE connections will provide Layer 2 and Layer 3 protection in a high availability configuration. Each firewall will connect to diverse upstream devices which announce CMRS-related IP space via BGP. NOAA will provide the IP address space for the enclave. [I]

Below the core, two Layer 2 network devices will function as distribution and access switches. [I]

The CMRS will be integrated into the existing security enclave and interconnected with other resources within the facility via existing switch infrastructure based on 10Gigabit Ethernet (IEEE 802.3ae). All applicable components of the CMRS shall be compliant with this standard. [C]

All devices shall support IEEE 802.3ad Link Aggregation Control Protocol. [S]

DRAFT

All uplinks and interconnections between network devices should not be oversubscribed. Network devices that have no oversubscription are strongly desired. [S]

Reference Figure 9 for an overview of the local area network architecture. [I]

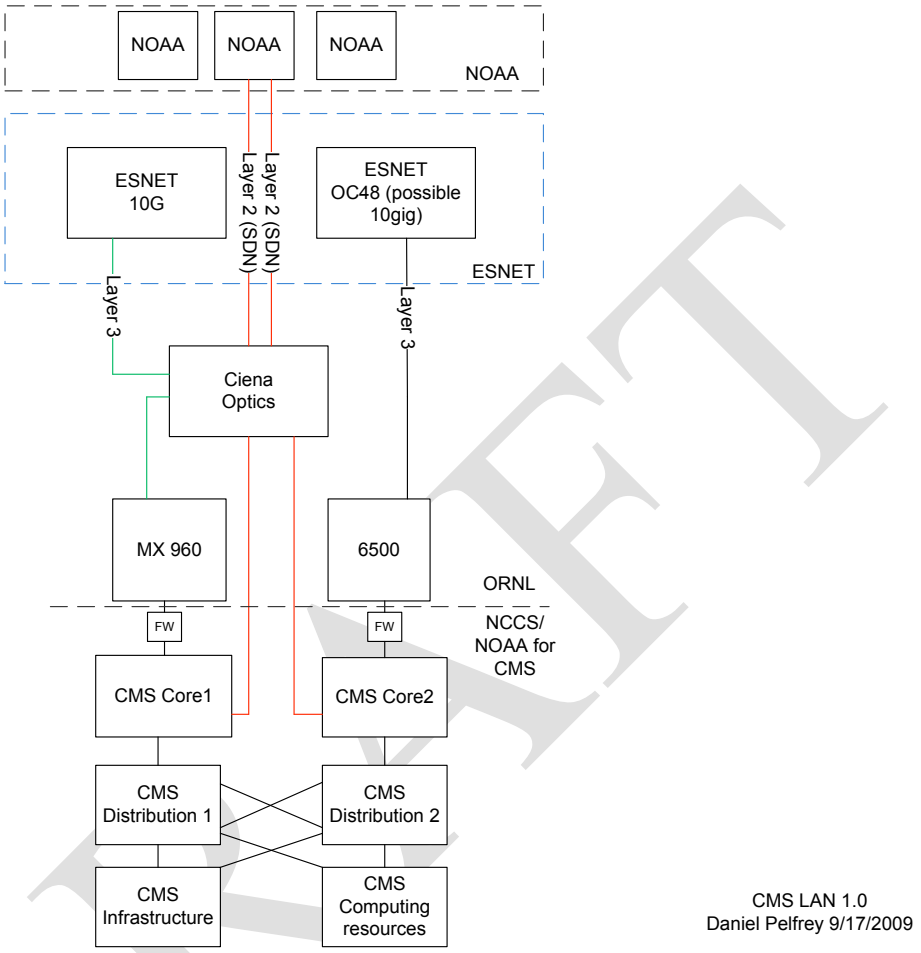

Figure 9. Conceptual CMRS Network Infrastructure

All applicable components of the CMRS shall include support for Ethernet frames with a MTU greater than 1,500 bytes. Support for jumbo frames of no less than 9,000 bytes is required. Support for super-jumbo frames of no less than 16,000 bytes is strongly desired. [S]

All applicable components of the CMRS shall include support for SNMPv2c (RFC 1901) and SNMPv3 (RFC 2570). All applicable components of the CMRS shall support NTP and syslog. All applicable components of the CMRS shall support 802.1q VLANs. [S]

9. Security

The CMRS shall reside in a security enclave that is maintained by ORNL. Security controls and assessments are used within this enclave to deliver secure and reliable services in accordance with a Confidentiality/Integrity/Availability risk assessment of Low/Moderate/Low, as defined in Federal Information Processing Standard 199, Standards for Security Categorization of Federal Information and Information Systems. The CMRS must operate within this security enclave. [S]

DRAFT

The Offerer shall comply with the requirements of the NOAA and ORNL Cyber Security Program Plans. [C]

Offeror staff that access the CMRS systems shall have the equivalent of an active DOE L or DoD SECRET or higher security clearance..

The Offeror shall comply with a rigorous Configuration Management (CM) process that is directed by ORNL. All changes to the system shall be approved by ORNL through the CM process. CM compliance and unauthorized change detection will be enforced using host-based intrusion detection facilities provided by ORNL. [C]

The Offerer shall comply with user/session behavior auditing practices provided by ORNL that allow for the inspection of user activity on the CMRS. [C]

The Offeror shall comply with network monitoring strategies provided by ORNL, including deep packet inspection and behavior analysis. [C]

The Offeror shall comply with existing requirements for ongoing security audits and personnel training. [C]

The Offeror shall ensure that system media and output are properly marked, controlled, stored, sanitized, and destroyed as appropriate. [C]

The Offeror shall comply with provided instructions on reporting potential security incidents. The Offeror shall ensure that authentication mechanisms issued for the control of their access to the CMRS are protected in accordance with NOAA and ORNL requirements. [C]

The Offeror shall protect CMRS administrative access end-points (e.g. terminals and workstations) from unauthorized access. [C]

Any worker who requires administrative access to the CMRS may be required to first undergo a background screening. [C]

The Offeror shall adhere to the least privilege principle in granting access to the CMRS, related resources, and data. [C]

The Offeror shall comply with authentication infrastructures provided by NOAA and ORNL (e.g. public key infrastructure, one-time passwords). [C]

10. Home File System

The Home File System (HFS) is a NetApp fabric-attached storage (FAS) NFS appliance. The HFS will interface to the CMRS via NFSv3 over TCP/IP. [I]

The Offeror shall provide Ethernet connectivity of at least 1 Gb/s per service node to this filesystem. An upgrade path to NFSv4 is desired. [S]

11. Warranty, Maintenance, and Support Services

11.1. Extended Warranty

The Offeror shall provide an extended warranty for each proposed system. The terms of the extended warranty are based on whether the Offeror proposes one system, or multiple subsystems. [I]

DRAFT

11.1.1. Warranty for Single System Installation

For an offer that describes a single system configuration, the Offeror shall provide an extended hardware warranty, or combination of warranty and maintenance, for a period of no less than three years from the date of system acceptance. The Offeror shall provide an extended software warranty, or combination of warranty and maintenance, that includes support, maintenance, and upgrades, for a period of no less than three years from the date of system acceptance. [C]

11.1.2. Warranty for Multi-Phase System Installation

For an offer that describes a multi-phase configuration, the Offeror shall provide an extended hardware warranty, or combination or warranty and maintenance for the second subsystem for a period of no less than three years from the date of system acceptance for the second subsystem. The Offeror shall provide an extended hardware warranty for the first subsystem such that it meets the restrictions for system retention as described in Section 0. [C]

For an offer that describes a multi-phase configuration, the Offeror shall provide an extended software warranty, or combination or warranty and maintenance, that includes support, maintenance, and upgrades, for the second subsystem for a period of no less than three years from the date of system acceptance of the second subsytem. The Offeror shall provide an extended software warranty for the first subsystem such that it meets the restrictions for system retention as described in Section 0. [C]

11.1.3. Warranty for TDS

The extended hardware and software warranty for the TDS shall match the warranty terms and conditions of the associated CMRS subsystem.

11.2. Hardware

The Offeror shall provide (at a minimum) 24x7, 4-hour response, on-site hardware maintenance for critical system components. [C]

The Offeror shall provide (at a minimum) 8x5, next business day, on-site hardware maintenance for non-critical system components. [S]

The Offeror shall provide an on-site spare part inventory for commonly failing components, with expedited replacement of parts removed from inventory. [S]

11.3. Software

The Offeror shall provide (at a minimum) 24x7, 4-hour response, remote support for critical system events. [S]

The Offeror shall provide online and telephone problem reporting, tracking, and support for non-critical system problems. [S]

The Offeror shall retain support for new software features on the delivered hardware for the full term of the extended warranty. New software features that are not hardware- or architecturally-specific to a different hardware platform shall be made available for the delivered hardware for the full term of the extended warranty. [S]

DRAFT

11.4. System Support

The Offeror shall provide all system support services necessary to meet requirements set forth in this solicitation. These services may include, but are not limited to, system administration, hardware maintenance, software maintenance and continuous system monitoring for security and availability. [C]

11.5. System Administration

The Offeror shall provide sufficient support staff to administer the CMRS and all peripheral devices and to work with ORNL staff members who help support operations and transition-to-operations type development work. The Offeror shall provide at least two on-site system administrators at the ORNL facility during normal business hours. The Offeror shall provide substantive access to a storage system administrator. ORNL will augment the Offeror-supplied system administration staff with two UT-Battelle system administrators. The Offeror is expected to work collaboratively with the ORNL system administrators to support a wide range of activities but the Offeror remains responsible for meeting system reliability, availability, run-time variability and other performance requirements. [C]

System administration support is required 7x24x365. The Offeror is responsible for organizing a system administration support structure sufficient to meet all requirements and for ensuring competent staffing at all hours. On-call system administrators are required to provide expert assistance to ORNL personnel engaged in supporting product generation and dissemination. On-call and emergency staffing arrangements must be sufficient to meet requirements set forth in this solicitation. The Offeror and ORNL will jointly execute system administration activities. The Offeror is fully responsible for meeting ORNL's requirements for system RRAS and for cooperating with ORNL staff engaged in system support. [C]

The Offeror shall support a training program for ORNL personnel assigned to assist the Offeror-provided system administrators. [S]

11.6. Application Support

The Offeror shall provide sufficient support staff to assist ORNL staff with code optimization, data migration, training, and code conversion. The Offeror shall describe their strategy for meeting this requirement. The Offeror shall consider the impact on application support needs for heterogeneous CMRS solutions. The Offeror may use Application Support staff to augment System Administration support during normal business hours. However, application support staff shall not replace system administrators and they are not expected to participate in on-call activities except during normal business hours. [C]

The Offeror shall provide opportunities for collaborative work to develop innovative software and/or support functions for the CMRS. Opportunities to examine/test new features or new system architectures are sought. [S]

11.7. Facilities Engineering Support

The Offeror shall provide facilities engineering support to ORNL relative to the power, cooling and space requirements of the proposed systems. The Offeror shall install the proposed systems. ORNL will provide the modifications to the facilities that are necessary to support the proposed systems, including installation and termination of branch circuits, installation and termination of chilled water supply and return lines, and minor modifications to the supplemental air systems that accommodate waste heat ejected to air. [C]

DRAFT

11.8. Training

The Offeror shall deliver a comprehensive training program for CMRS administrators and end users. The program shall include periodic refresher training for new releases. Classes for end users should be on-site at ORNL, GFDL, NCEP and ESRL/GSD. Classes shall be tailored to the needs of scientific staff and shall include classes on system basics, compiler features, performance tuning and optimization features. Application Support staff shall identify systematic user issues and structure on-site training classes to address them. Meteorological and Oceanographic applications and associated libraries should be highlighted, if applicable. Training for system administrators shall be a combination of standard classes taught at the Offeror's facilities and on-site custom classes. Pertinent topics including but not limited to advanced system administration, file systems, resource management, performance tuning, system troubleshooting and failure analysis should be offered. [S]

11.9. Documentation [S]

ORNL will provide both internal and external communication infrastructure that includes, at a minimum, secure and non-secure web access, wiki, and sharepoint resources. Electronic versions of documentation may be stored on this infrastructure. [I]

The Offeror shall make electronic versions of documentation available to CMRS system administrators and users. ORNL and Offeror will identify items that may require authorization and authentication for access, and the method for that authentication. Documentation shall include but should not be limited to:

- Site preparation and installation guides
- System hardware manuals
- System operations manuals
- User's manuals (system, commands, compilers, assembler, debugger(s), libraries, performance utilities, and other similar materials)
- System programmer's manuals
- System administrator's manuals
- UNIX-style "man pages,"

A low level user's guide and a problem reporting procedure are required. Advanced level documentation, in support of training activities is required.

11.10. Warranty, Maintenance, and Support Termination

The Offeror shall describe all warranty, maintenance, and support activities such that no activities extend past 30 September 2014.

12. Facilities

12.1. Basis [I]

In accordance with the Inter-Agency Agreement described in Section 1.2, the ORNL National Center for Computational Sciences (NCCS) will provide the base facilities and infrastructure associated with the Climate Modeling and Research System. This agreement allows NOAA to reduce its investments in computing facilities and reallocate its investments to focus on its core environmental mission. GFDL, NCEP and ESRL/GSD will continue to support pre and post processing applications at their facilities.

DRAFT

12.2. Facilities Description [I]

12.2.1. General Facilities Description [I]

The NOAA CMRS and all supporting equipment will be installed within Building 5600 at ORNL. This building was constructed in 2003, and contains a total of 135,670 ft^2. It is part of a series of three buildings that total more than 365,000 ft^2 of conditioned space.

Within Building 5600, there are two separate computer rooms, each with approximately 20,000 ft^2 of raised floor. These computer rooms include substantial utility, physical security, building automation systems, and fire protection systems. In addition, the facility is staffed 24x7x365 by operations, security, electrical systems, and mechanical systems personnel.

Physical access to the computer room is provided via a loading dock facility, 48" above grade, and a freight aisle with minimum dimensions of 78.75"(w) and 106.5" (h) including doors. Floor loading is limited to 250lbs/ft^2. Any equipment placement that exceeds this floor load rating must include an Offeror-provided and ORNL-approved solution.

12.2.2. Electrical Infrastructure and Electrical Conditions Affecting the Offeror [S]

Electrical power is supplied to the NCCS facility via two 13.8kV feeders that are a combination of overhead & underground circuits. These circuits are routed through switchgear at the HPCC to unit substations located in and around the HPCC facility. These unit substations are typically rated for 2.5/3.3 MVA and reduce voltage to 480Y/277V. For large computer systems, 480V 3 phase branch circuits are supplied from switchboards to both the computer cabinets and to the cooling equipment associated with the cabinets. These switchboards supply additional transformers that reduce voltage to 208Y/120V to supply disk drives, network equipment, and other supporting equipment. Uninterruptable power is provided via double conversion type UPS's that are backed by diesel generators.

The CMRS electrical infrastructure is based on delivery of 480v, 3-phase electrical power to the computer cabinets. In addition, a preference is given to disk/data storage equipment solutions that directly use a 480v, 3-phase service. The CMRS will be powered from main switchboards MSB-12 and MSB-13 that contain 480V breakers. The size, quantity, and number of poles of these individual breakers and the number of vertical sections will be determined by the selected CMRS. Each computer cabinet shall be powered by a single 480v 3-phase circuit. Delta-connected power supplies are preferred.

Subject to the total program facility restrictions for this system, ORNL will provide the electrical distribution system for the CMRS including transformers, switchboards, vertical sections, power distribution units, remote distribution units, and breakers. ORNL will provide the branch circuits from the switchboards, power distribution units, or remote distribution units to the individual cabinets of the CMRS. Connection of supply circuits to CMRS equipment shall via direct connection or via plug and receptacle, in which case Offeror shall provide the receptacles to be installed by ORNL.

In addition to the prescribed 480Y/277V 3-phase power available for powering CMRS equipment, there will be limited 208Y/120V, 3-phase electrical circuits available for network infrastructure, critical infrastructure, and perhaps storage systems. This same infrastructure shall support service workstations, management consoles, and other similar devices. ORNL will provide the branch circuits from the 208Y/120V power distribution panels to the individual cabinets of the support equipment.

The total peak power consumption for the CMRS is facility-limited to 5.0MW. The total peak power calculation shall be based on the realistic upper bound for worst-case power consumption for short durations. The sustained maximum consumption of power for the CMRS is limited to 4.0MW. The sustained maximum consumption shall be based on an execution of a typical workload, with very heavy (95% or greater) utilization. Occasional, modest consumption above the sustained maximum consumption but less than the total peak power consumption can be absorbed by the underlying electrical and cooling

DRAFT

infrastructure. Reference a comparison of total peak power consumption and sustained maximum power consumption in Figure 10.

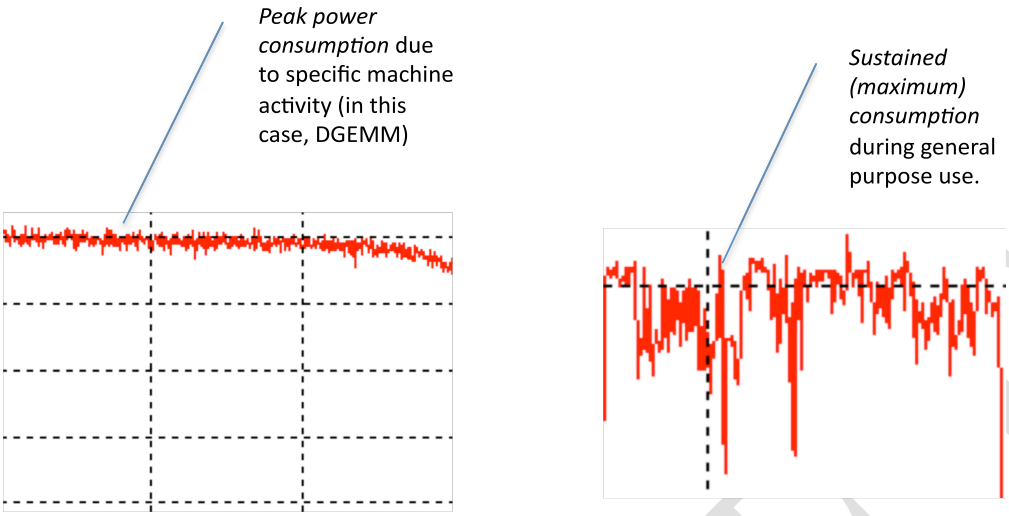

Figure 10. Peak power consumption and Sustained (maximum) power consumption

The 5.0 peak power consumption and 4.0MW sustained maximum power consumption conditions are independent of equipment that uses 208Y/120V electrical power.

The total sustained (maximum) power consumption of 208Y/120V, 3-phase electrical power for the storage, network, and other systems that support the CMRS computer is limited to 225kW for UPS-protected systems. This is the maximum amount of power that can be supplied from either side of redundant power sources or the total power that can be supplied from both sides of redundant sources. For example a piece of dual-corded equipment can present a 112.5kW load on each cord or 225kW on one cord in the event the other side of the redundant source is not available.

The total sustained (maximum) power consumption of 208Y/120V, 3-phase electrical power for devices that do not require UPS protection is limited to 400kW. The Offeror shall describe systems that are not protected by UPS, and the rationale for this configuration.

Power supplies for computer, network, and storage equipment; cooling systems, and other equipment directly associated with the CMRS shall comply with IEEE 519 requirements for harmonic currents. The Offeror shall provide ORNL with harmonic current data produced by CMRS computer loads. This data will be used by ORNL to model the electrical system to determine impact on other systems and to provide an estimation of total harmonic distortion for voltage at switchboards.

Power supplies for all equipment associated with the CMRS shall be designed to tolerate power quality events bounded by the SEMI F47 curve for voltage anomalies below nominal voltage and by the ITI (CBEMA) curve for voltage anomalies above nominal voltages.

Power configuration to storage systems shall be configured so that the system components are dual-fed, and load-balancing. This configuration must support utility feeds from separate sources such that the power supplies shall manage the failure of one source automatically, with total load from the remaining power supplies keeping the system stable.

The power factor (pf) for all equipment shall be greater than 98%.

DRAFT

12.2.3. Mechanical Infrastructure and Mechanical Conditions Affecting the Offeror [S]

Chilled water is available from a local Central Energy Plant (CEP) within 5600 that is interconnected as needed to two other independent CEPs. Total chilled water capacity within the CEP is 6,600 tons, or the equivalent of more than 23MW of heat load. This is provided by a 10”-12” chilled water loop with crossover and splitter lines located under the raised access floor. The loop is supplied from four 12” feeders connected to 16” and 18” main lines. The chilled water loop services both liquid cooled computer heat exchangers directly and precision Computer Room air-conditioning Units (CRUs). Conditioned air from CRUs is distributed thru access floor tiles with perforations. Air is returned under the ceiling tiles to the top of the CRUs. A leak detection system is provided on the concrete subfloor.

Automatic flow control valves limit the maximum chilled water flow through all connected components.. A condensate removal system employing gravity drains is provided. The gravity drains are located in the concrete floor supporting the access floor.

Total chilled water capacity available to the CMRS is about 1400 tons, the rough equivalent of the 5.0MW electrical power consumption upper bound for the CMRS.

The Offeror shall design its computer cooling systems such that the significant majority of the heat removal is accomplished through direct connections to the insulated chilled water loop located under the access floor. A very modest load shed directly to air is acceptable. No more than 70 tons is available for air cooling computer components.

Chilled water is provided with the following characteristics:

- Total maximum cooling capacity available for this program: 1400 tons.
- Maximum cooling capacity available for liquid cooled equipment: 1330 tons (95% of total chilled water load).
- Maximum cooling capacity available for air-cooled equipment: 70 tons (5% of total chilled water load).
- Typical supply temperature of the chilled water: 41.5F to 43.0F.
- Required differential temperature thru all connected components is 14F minimum.
- Chilled water chemistry and quality is typical for hydronic comfort cooling chilled water systems.
- Total maximum differential pressure available for all connected computing system components, including piping, fittings, strainers, isolation valves, control valves, and filters, is 25 psi.
- Maximum system pressure is 150 psig.
- Chilled water is strained thru 840 micron strainers.
- All taps for computer connections will be provided with:
    - Isolation valves on the supply and return
    - A 840 micron strainer on the supply
    - An automatic flow control valve sized for the maximum flow of the connected device
- Pipe insulation required is fiberglass with an all service jacket, National Fire Protection Association (NFPA) rated
- Hose insulation required is non-halogenated elastomeric, NFPA rated

Air is supplied from CRUs with the following characteristics:

- Air supply dry temperature is 57F at access floor level
- Air supply dew point temperature is 46F at access floor level
- Air is filtered with ASHRAE 30% filters

The ORNL Facilities allocated to this program are shown in Figure 11.

DRAFT

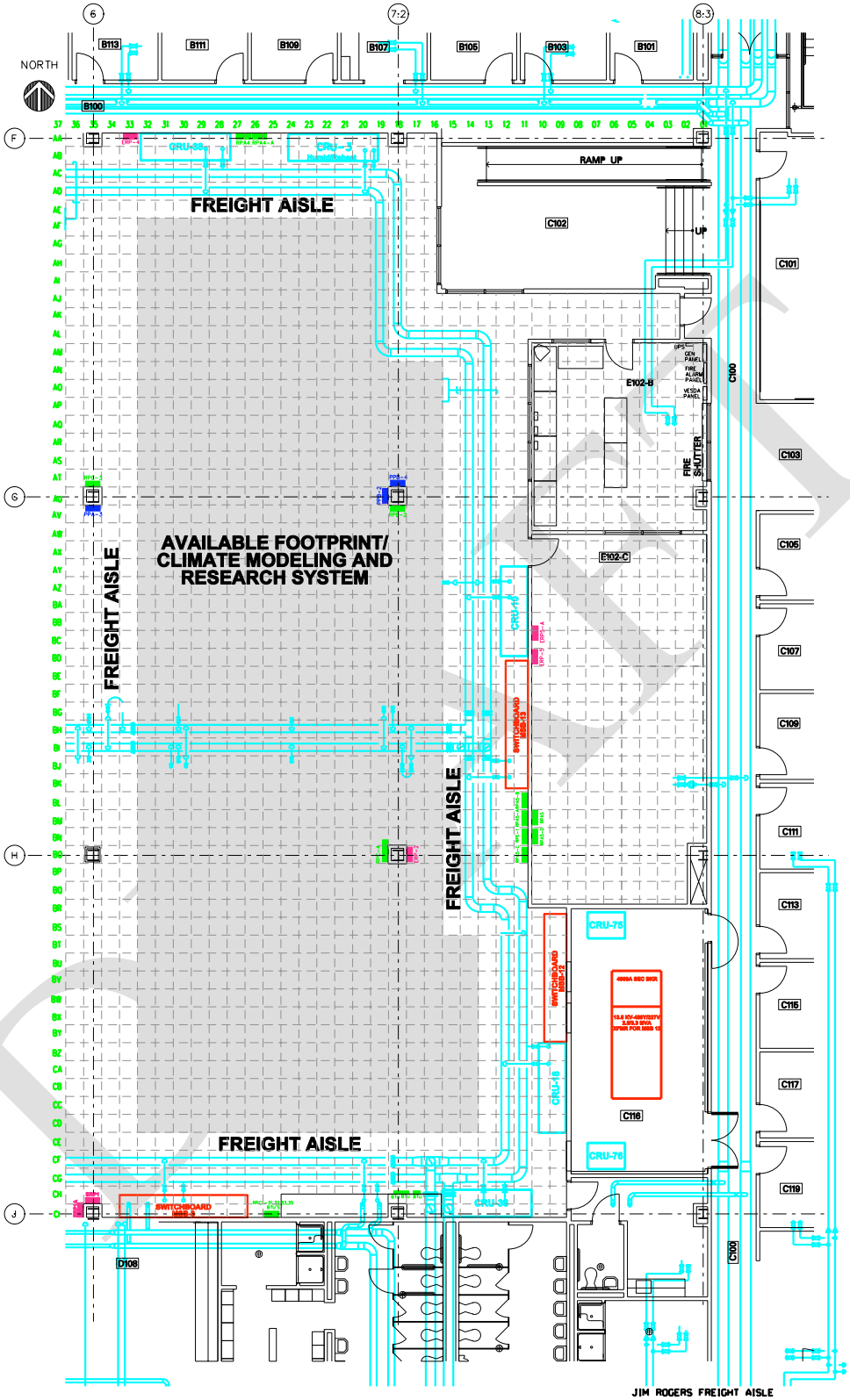

Figure 11. Available Floor Space for the Climate Modeling and Research System