



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

Table of Contents

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



5.2.	<i>FS and LTFS File System Bandwidth and Capacity Requirements</i>	12
5.3.	<i>Common Data Management Requirements</i>	15
6.	Software	16
6.1.	<i>Site Support Advocate</i>	16
6.2.	<i>Workload and Resource Management</i>	17
6.2.1.	<i>Workload Management [I]</i>	17
6.2.2.	<i>Resource Management</i>	17
6.3.	<i>Software Suite</i>	17
6.4.	<i>Checkpoint/Restart</i>	18
7.	Resiliency, Reliability, Availability, Serviceability	18
7.1.	<i>RAS System Initialization and Reboot</i>	18
7.2.	<i>RAS System Diagnostics and Maintenance</i>	18
7.3.	<i>RAS System Event Monitoring and Archival</i>	19
7.4.	<i>System Mean Time To Interrupt and System Effectiveness Level Requirements</i>	19
7.4.1.	<i>System Mean Time Between Interrupt</i>	19
7.4.2.	<i>System Mean Time Between Failure Requirements</i>	19
7.4.3.	<i>System Effectiveness Level [S]</i>	20
7.5.	<i>Resiliency</i>	20
7.6.	<i>Reliability</i>	20
7.7.	<i>Availability</i>	21
7.8.	<i>Serviceability</i>	21
8.	Infrastructure and Networking Services	22
8.1.	<i>Infrastructure</i>	22
8.2.	<i>Wide Area Network Connectivity [I]</i>	22
8.3.	<i>Local Area Network Connectivity</i>	22
9.	Security	23
10.	Home File System	24
11.	Warranty, Maintenance, and Support Services	24
11.1.	<i>Extended Warranty</i>	24
11.1.1.	<i>Warranty for Single System Configurations</i>	24
11.1.2.	<i>Warranty for Multi-Phase System Configurations</i>	25
11.2.	<i>Hardware</i>	25
11.3.	<i>Software</i>	25
11.4.	<i>System Support</i>	25
11.5.	<i>System Administration</i>	25
11.6.	<i>Application Support</i>	26
11.7.	<i>Facilities Engineering Support</i>	26
11.8.	<i>Training</i>	26
11.9.	<i>Documentation [S]</i>	27



12. Facilities	27
12.1. <i>Basis [I]</i>	27
12.2. <i>Facilities Description [I]</i>	27
12.2.1. General Facilities Description [I]	27
12.2.2. Electrical Infrastructure and Electrical Conditions Affecting Offerors [C]	27
12.2.3. Mechanical Infrastructure and Mechanical Conditions Affecting Offerors [C]	29

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 for prototyping critical weather and climate applications in support of NOAA’s mission. Computing allocations are fulfilled at ORNL through the DOE Advanced Scientific Computing Research (ASCR) 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 which 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 provide research collaboration and technical support for high performance computer and data systems that will deliver improved 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.

A primary research goal is to develop, test, and apply state-of-the-science computer-based global climate simulation models that are based upon a strong scientific foundation and leverage leading-edge high-performance computing and information technologies. The objective is to increase dramatically the skill, resolution, complexity, and throughput of computer model-based projections of climate variability and change to enable sound decision-making on issues of national importance, such as future energy use and technology options.

To further these research goals, ORNL is issuing a Request for Proposal 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 collaborative climate research community.

1.3. Conventions [I]

Attributes of the Climate Modeling and Research System 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 the most suitable system configuration. Sections that include these items are marked with [I].

Figure 1. Conventions

1.4. Purpose [I]

The Climate Modeling and Research System (CMRS) acquisition will support ORNL efforts to provide NOAA with computational resources to achieve their mission goals. ORNL 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 R&D 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). 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 22 petabytes (PB)) and NCEP (currently 10PB), with a smaller system at ESRL/GSD (approximately 1PB). These systems are expected to remain at those NOAA locations. The growth rate in each of these archives is exponential. The existing workflow for all three Centers allows for computational products to be moved from large computational resources, regardless of location, and including ORNL, to the local archive for post-processing activities.



2. Procurement Goals and Scope [I]

2.1. Computational Resources [I]

The selected Offeror shall provide computational resources such that maximum performance, as determined by benchmark performance, is achieved within a fixed-price contract and 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.

2.2. Storage Resources [I]

The selected Offeror shall provide data storage that is designed to provide balanced performance when considered with the computational resources and system software (especially the file system) and meeting all requirements.

2.3. Infrastructure Services [I]

ORNL shall provide a broad range of infrastructure services including LDAP, DNS, DHCP, 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.

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 contract. 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 in the NCCS computing facility in Oak Ridge, TN, operated, administered, maintained and supported by the selected Offeror, and operated and administered collectively by the Offeror and ORNL. [C]

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]



The overall design of the CMRS shall balance the significant desire for a 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 proposed novel approaches to system design, system installation, system upgrades, and system administration and maintenance activities that will best meet all of these goals. [I]

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

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

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

4.1.1. Single System Configurations

The CMRS may be proposed as a single, integrated computer system. [I]

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

This system may be logically separated into more than one partition, or may be delivered as two physically distinct subsystems. [I]

4.1.2. Multi-Phase System Configurations

The CMRS may be proposed as two separate computer system deliveries. 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]

Each system (or subsystem) should be able to be used in its entirety by a single application. For a system that is logically partitioned, that partition should be able to be used in its entirety by a single application. [S]

The minimum length of production time that the two systems (or collection of subsystems) must overlap is to be no less than 6 months to allow for a transition from the initial subsystem to the second subsystem. System installations must meet all warranty and support requirements described in Section 11. Offerors 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. [S]

The delivery strategy for a multi-phase installation is shown in Figure 2. The blocks designated System 1 and System 2 could each individually consist of a single system, a logically partitioned system, or two physically distinct subsystems. The best strategy for meeting the capacity and RAS requirements is left to the Offeror.

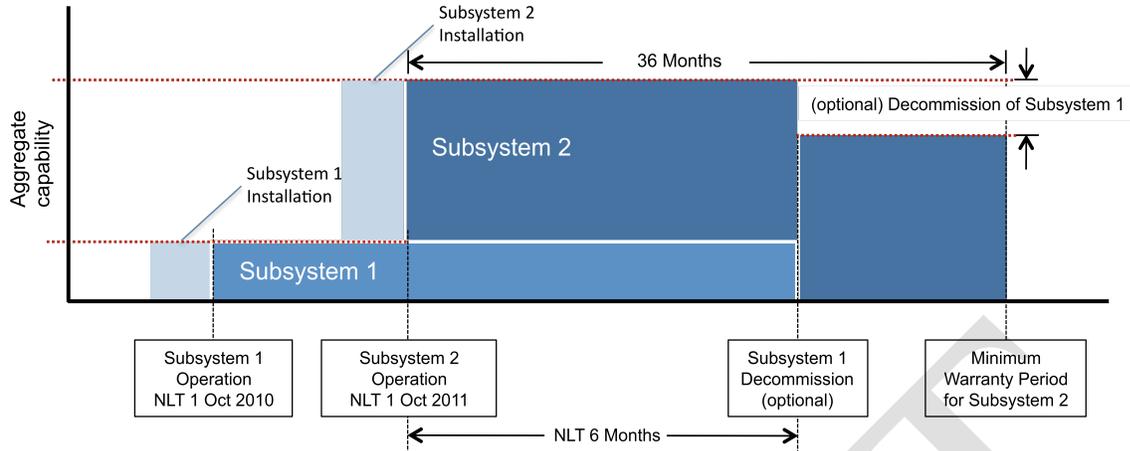


Figure 2. Example lifetime of a two-phase CMRS Installation

4.1.3. System and Subsystem Upgrades

Offerors 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.

4.2. System Computational Capacity [C]

The system computational capacity is based on the results of the workflow throughput benchmark, based on the GFDL high-resolution coupled climate model (CM2-HR). We will extrapolate the potential delivered capacity over the life of the delivered system(s) using the Offeror's delivery and warranty schedule. Reference 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 system (or subsystem) should be able to be used in its entirety by a single application. For a system that is logically partitioned, that partition should be able to be used in its entirety by a single application. [S]

Benchmark codes will be employed to assist ORNL in assessing system performance. These benchmark codes will include climate and weather modeling codes. ORNL may also elect to employ synthetic benchmark codes and/or to analyze benchmark code execution in detail and not rely upon simple wall-clock run-time assessment techniques. [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 available on 1 October 2010 shall be at least 750 TF. [S]

In the event that the Offeror proposes a two-phase delivery as defined in Section 4.1.2, 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. [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

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]

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]

4.6. 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]

Each CMRS subsystem shall work as a single system capable of running complex technical applications at the full size of the platform. To achieve this, the primary communication network (high speed interconnect) shall provide highly scalable performance. [C]

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. [S]

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 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]

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 should be sufficiently large that all architectural features of the larger system are replicated. This specifically includes interconnect mechanisms that will be deployed within the larger system. The TDS should physically span a minimum of two racks although each rack need not be fully populated. A separate test system is required for each CMRS subsystem if they are architecturally distinct. Test system(s) shall be comprised of identical components (hardware and software), operation and support to the associated CMRS including a scaled version of the FS file system (as described in Section 5).

Test system(s) shall be independent of the CMRS except for the following elements:

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

The batch and scheduler environment and the CMRS can be unified (optional).



5. I/O Systems and Data Management

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 Fast-Scratch file system 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 Fast-Scratch file system will not be backed up. [I]

Initially the Fast-Scratch file system shall be able to support 100 million files and individual files that are up to 10 TB in size. It is expected that these requirements will grow over the life of the contract. [S]

Should the CMRS be partitioned into multiple systems (as described in Section 4.1.2), a single parallel file system accessible from all computation partitions is preferred for the FS. Should a single parallel file system be offered and should the Offeror choose a phased deployment, the single parallel file system should be expanded to meet the requirements of the second phase system at the time of the second phase delivery. [S]

The FS must support quotas for reporting and enforcement. Enforcement of quotas shall be configurable separately from quotas for reporting purposes. [S]

5.1. Long-term Fast Scratch (LTFS)

Long-term Fast Scratch is a staging area for users to temporarily place files that will frequently be manipulated. This file system does not need to be backed up and can be purged of aged data. [I]

The LTFS file system shall be able to support 100 million files and individual files that are up to 10 TB in size. It is expected that these requirements will grow over the life of the contract. [S]

It is anticipated that the scientific data for the CMRS will be primarily accessible through the LTFS. The LTFS shall remain accessible even 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 such as I/O server failover and path failover (DM Multipath, RDAC, or similar technologies); other requirements for fault tolerance are detailed in Section 5.2. ORNL expects a high availability for the LTFS with a target of 98% availability over the life of the contract. The Offeror shall provide a availability commitment for the LTFS over the life of the contract. Availability is defined as the ability to access the LTFS for reading, writing, creating, and deleting any/all files on the LTFS at the bandwidth and metadata rates specified for the LTFS. The LTFS shall support quotas for reporting and enforcement. Enforcement of quotas shall be configurable separately from quotas for reporting purposes.

The LTFS shall be configured with local data transfer nodes (LDTNs) that are defined as dedicated nodes that shall be used to transfer data between the LTFS and the FS systems. LDTNs shall mount both the FS and LTFS systems to support data transfer operations. LDTNs shall be shared between the FS and LTFS systems. The Offeror shall provide enough LDTNs to support the minimum bandwidth requirements as outlined in section 5.2, a minimum of 16 LDTNs shall be provided. The LTFS shall provide QDR InfiniBand or 10 Gbe connectivity for an additional 8 remote data transfer nodes RDTNs. RDTNs are defined as data transfer nodes that mount the LTFS file system for remote data transfers via GridFTP, BSCP, and other data transfer protocols. The RDTNs will be provided, configured, and maintained by ORNL staff. The LTFS shall be mounted on all login and batch nodes on the CMRS but need not be mounted on the compute nodes of the CMRS.

The Offeror must provide an itemization of the cost of the LTFS, LDTNs, and all related components including support and maintenance separate from the broader CMRS.

One of the main workloads of the LTFS is to act as a staging area for remote data transfers. This workload requires large data movement from the FS to the LTFS of datasets generated from application runs on the CMRS. The Offeror must 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.

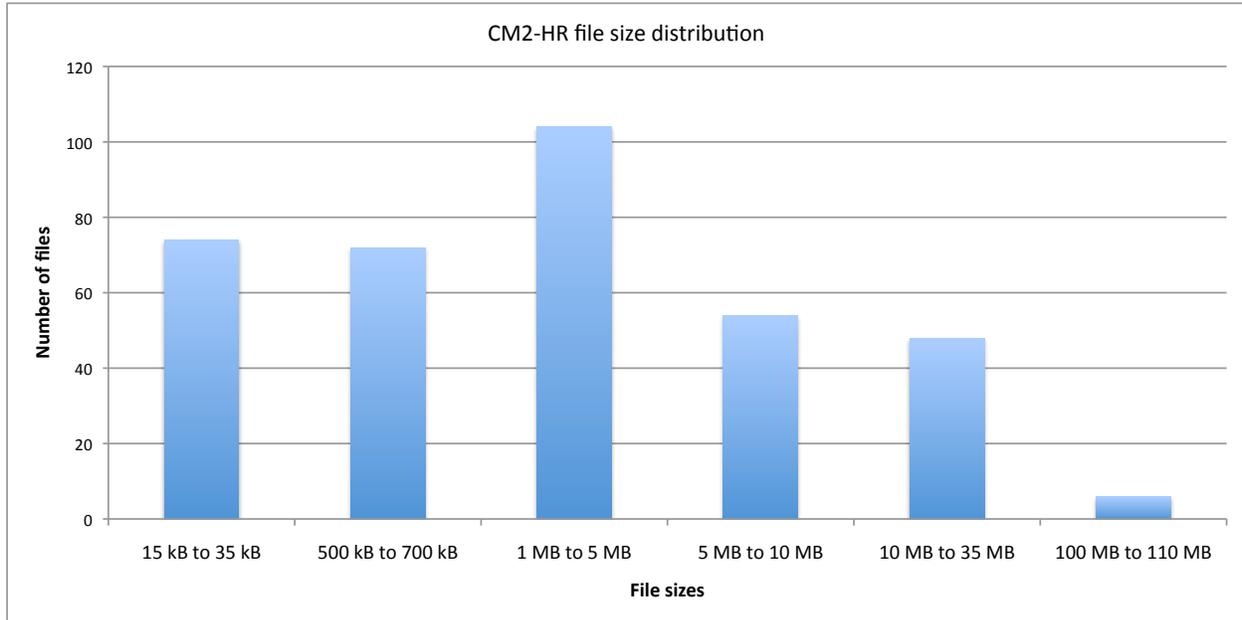


Figure 3. File Size Distribution

5.2. FS and LTFS File System Bandwidth and Capacity Requirements

Required data rates and capacity for both the FS and LTFS systems are a function of the total compute resource of the CMRS in terms of core count. Should a phased deployment occur each phase of the deployment must conform to the required data rates and capacity. These required data rates are based on historical data volumes produced as a function of core counts on existing systems projected to the CMRS. In 2008 1.5 Petabytes of data was generated on existing systems and will be used as the baseline for this formulation. The following definitions will be used in deriving the required data rates for the FS and LTFS systems.

- 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

Using the the existing system as a baseline, it is expected that the data rate will scale with computing capacity at 0.0214 GB/CPU-HOUR measured for 2008 workloads. This is illustrated below by Equation 1, 2 and Figure 4.

$$R_E = \frac{V_E}{365 * 24} \quad (1)$$



$$D_E = \frac{R_E}{N_E} \quad (2)$$

N_E (cores on existing systems)	V_E (GB/Year)	R_E (GB/hour)	D_E (GB/CPU-Hour)
8000	1,500,000	171.23	0.0214

Figure 4. Averaged sustained data production rate derived from total archived data volume (2008)

The following additional definitions will be used in deriving the required data rates for the FS and LTFS systems.

- TTPR = Total throughput ratio
- t_{CMRS} = benchmark job work stream run time on proposed CMRS system as projected by the Offeror
- t_E = benchmark job work stream run time on existing system
- F_{sw} = Scale factor from potential software improvements expected in 2 years

$$S_{eq} = F_{sw} * \frac{t_E}{t_{CMRS}} \quad (3)$$

$$TTPR = \frac{R_{CMRS}}{R_E} = F_{sw} * \frac{t_E}{t_{CMRS}} * \frac{N_{CMRS}}{N_E} = S_{eq} * \frac{N_E}{N_{CMRS}} \quad (4)$$

t_{CMRS}	t_E	F_{sw}	S_{eq}
1	1	2.0	2.0

Figure 5. Capability scale factor S

To get the factor by which storage requirements scale up from the baseline, take TTPR to the 0.7 power. To get the scaling on a per-core basis, divide $(TTPR)^{0.7}$ by (N_{CMRS} / N_E) . To get 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 .3 power of the core count:

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

As an example we will now estimate the FS and LTFS bandwidth requirements using an example CMRS system with 50,000 cores. 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 6:

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

$$R_{CMRS} = D_{CMRS} * N_{CMRS} \quad (7)$$

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



Figure 6. Data production rate using an example of 50K core system with a 2x capability increase

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 7.

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.

$$D_{FS} = \frac{R_{CMRS}}{d_{IO} * 3600} \quad (8)$$

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

Figure 7. The modified bandwidth from the assumption that 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 can be expressed as shown below in Equation 9 and corresponding Figure 8 with the following definitions:

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

$$BW = D_{FS} * F_{IO} \quad (9)$$

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

Figure 8. Total modified bandwidth required by the file system

The data writes to the LTFS and both reads from and write operations to 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 .5 in the factor F_{IO} (concurrent data movements) for LTFS.

The FS is required to hold the data output for a period of only 4 days. The LTFS is required to retain the data for a period of 4 days with an additional 14 days for data aging. The estimated storage size for the file systems is given below in Figure 9 using equations 10 and 11 and the following definitions:

T_1 = 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 9. Estimated size of FS and LTFS in GigaBytes based on the example using 50K cores

It should be noted that in this example these are actual file system sizes. The Offeror would be required to provide the file systems with adequate head room to allow efficient operation. Typically best practice operations allow the file systems operate at no more than 75% of total available storage capacity.

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

1. 3,000 metadata operations per second as measured by the mdtest benchmark
2. Usable (formatted) capacity as specified in equation 11
3. File system bandwidth as specified in equation 9. File system bandwidth will be measured from the LDTNs via the LTFSB (Long Term Fast Scratch Benchmark). This bandwidth must be achieved concurrently on both the LTFS and FS systems from the LDTNs.

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

1. 3,000 metadata operations per second as measured by the mdtest benchmark
2. Usable (formatted) capacity as specified in equation 10
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).
4. File system bandwidth as specified in equation 9. File system bandwidth will be measured from the LDTNs via the LTFSB.

The specified 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.

5.3. Common Data Management Requirements

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 or an equivalent data protection and recovery mechanism¹. 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]

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



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.

The storage system shall be configured to maintain the required file system bandwidth in the presence of concurrent rebuilds across 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 LUNs.

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

A storage system which supports and that is configured with InfiniBand connectivity as well as the SCSI RDMA protocol (SRP) is preferred.

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.

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

Features including parity-declustering for fast rebuilds and T10-DIF (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 SAS or FC enterprise class drives (10K RPM minimum). [S]

The storage system for the LTFS may utilize SAS, FC, or 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]

- Operating system
- Resource and workload management
- File system(s)
- Compilers
- Libraries
- Parallel computing environment
- Debugger(s)
- Performance tool(s)
- Other programming environment software

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 shall provide 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.

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. 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/>)
- 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.wmo.ch/web/www/WDM/Guides/Guide-binary-2.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>)
- 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>)
- R (www.r-project.org)
- Rudy (<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 Enterprise Managers as appropriate. [S]

- Allinea DDT (<http://www.allinea.com/>) for up to 1,000 cores
- IDL (<http://www.itvis.com/ProductServices/IDL.aspx>) for a single node (not subsystem or partition) of the CMRS
- Matlab (<http://www.mathworks.com/>) for a single node (not subsystem or partition) of the CMRS
- NAG (<http://www.nag.co.uk>) for a single node (not subsystem or partition) of the CMRS
- TotalView (<http://www.totalviewtech.com>) for up to 500 cores

All updates of software installed or used on the system shall be managed through a formal configuration management process. [S]

6.4. Checkpoint/Restart

The CMRS shall support checkpoint-restart of running jobs. The Offeror shall describe system support for and user compliance requirements for checkpointing jobs. System initiated checkpoint-restart is not a requirement. [S]

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 should 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.

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 should 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]

All operations status transitions of all nodes shall be recorded. Operations status is one of the following three mutually exclusive values:

- 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 are considered NULL time, and do not materially impact 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 minimum System Mean Time Between Interrupt (SMTBI) shall be at least 200 hours. A system interrupt is defined as any hardware or system software error or failure, or cumulative errors or failures over time, resulting 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]

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

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. [S]

7.4.3. System Effectiveness Level [S]

The system Effectiveness Level (EL) shall be at least 90%. The EL is defined as:

$$EL = (PU - SD - NULL) / (PU + UD - SD - NULL)$$

Where PU is production uptime, SD is scheduled downtime and UD is unscheduled downtime. The period of measurement for EL shall be at least 336 hours (14 consecutive days).

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

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.

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.

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.

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 should 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]

7.6. Reliability

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



System resources should support continuous monitoring and a mechanism whereby errors are reported for analysis and subsequent action. Reliability will be examined during Acceptance via tests of numerical reproducibility, numerical accuracy and run-time variability. Similar tests, although of a shorter duration, will follow every system outage involving hardware/software upgrades, including quarterly upgrades. ORNL reserves the right to verify proper function at any time. [S]

Data integrity should 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 will 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, the Offeror shall 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 will be marked as downtime. [C]

The CMRS should 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). Run-time variability is influenced by input data. Benchmark codes will employ fixed input datasets. [S]

7.7. Availability

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

The LTFS and FS file systems shall demonstrate an availability of 99.9%. [C]

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.

7.8. Serviceability

A high level of serviceability is required to meet ORNL's availability needs. The CMRS design should 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. [S]

The CMRS shall support hot swap of field replaceable components (such as disk drives and cooling fans) without impacting the operating state of the system. [S]

The used of serviceable replacement parts that are warranted as new is acceptable.



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]

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 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]

8.3. Local Area Network Connectivity

The 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 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. [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]

Above the core will consist of two firewalls with 10GigE connections in/out, that can do both Layer2 and Layer3 functionality and function in active/standby mode for high availability. Each firewall will be connected to diverse upstream devices which will announce CMRS related IP space via BGP. [I]

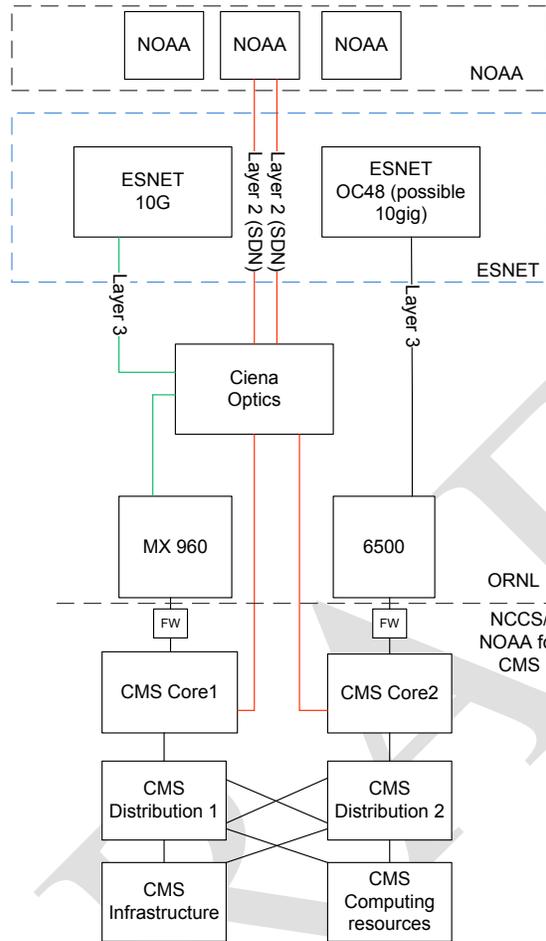
Below the core will consist of two Layer 2 network devices that will function as distribution and access. [I]

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 should support IEEE 802.3ad Link Aggregation Control Protocol. [S]

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

Reference Figure 10 for an overview of the local area network architecture. [I]



CMS LAN 1.0
 Daniel Pelfrey 9/17/2009

Figure 10. Conceptual CMRS Network Infrastructure

All applicable components of the CMRS shall include support for Ethernet frames with a MTU greater than 1500 bytes. Support for jumbo frames of no less than 9000 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 should include support for SNMPv2c (RFC 1901) and SNMPv3 (RFC 2570). All applicable components of the CMRS should support NTP and syslog. All applicable components of the CMRS should support 802.1q VLANs. [S]

9. Security

The CMRS will 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]



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

The Offerer shall comply with a rigorous Configuration Management (CM) process that will be provided 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 Offerer shall comply with network monitoring strategies provided by ORNL, including deep packet inspection and behavior analysis. [C]

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

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

The Offerer shall comply with provided instructions on reporting potential security incidents. 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 Offerer shall protect CMRS administrative access end-points (e.g. terminals and workstations) from unauthorized access. [C]

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

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

The Offerer 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 FAS NFS appliance. The HFS will interface to the CMRS via NFSv3 over TCP/IP. [I]

The Offerer 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

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

11.1.1. Warranty for Single System Configurations

For an offer that describes a single system configuration, the Offerer 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 Offerer shall provide an extended software warranty, or combination or



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 Configurations

For an offer that describes a multi-phase configuration, the Offeror shall provide an extended hardware warranty, or combination of 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 4.1.2. [C]

For an offer that describes a multi-phase configuration, the Offeror shall provide an extended software warranty, or combination of 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 subsystem. 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 4.1.2. [C]

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]

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

11.6. Application Support

The CMRS 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. Offerors 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]

Offerors should 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]

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 should be tailored to the needs of scientific staff and should 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 should be a combination of standard classes taught at 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]

Offerors are expected to deliver extensive, on-line system documentation available to all CMRS system users. The Offeror shall provide the host (and host maintenance/support) for all system related documents. On-line document availability shall be 99% (measured monthly). CDs of documentation shall be available upon request. Documentation should 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 others)
- System programmer's manuals
- System administrator's manuals
- Standard UNIX "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.

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.

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 ft². It is part of a series of three buildings that total more than 365,000 ft² of conditioned space.

Within Building 5600, there are two separate computer rooms, each with approximately 20,000 ft² 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/ft². 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 Offerors [C]

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 can 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 can 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 infrastructure. Reference a comparison of total peak power consumption and sustained maximum power consumption in Figure 11.

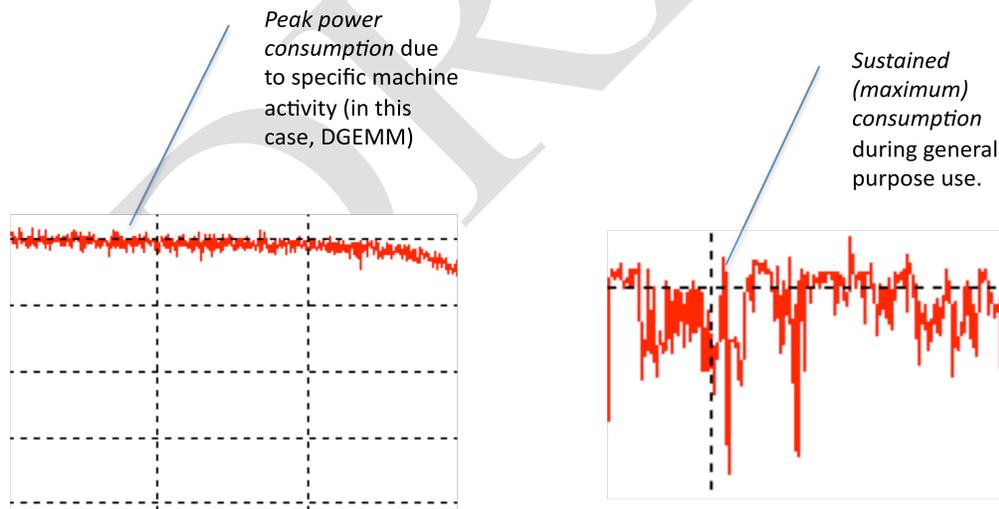


Figure 11. Peak power consumption and Sustained (maximum) power consumption

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 112kW 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. Offerors 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. 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%.

12.2.3. Mechanical Infrastructure and Mechanical Conditions Affecting Offerors [C]

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.

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: 1140 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 20 mesh (840 microns) strainers.
- All taps for computer connections will be provided with:
 - Isolation valves on the supply and return
 - A 20 mesh (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, 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 12.

DRAFT

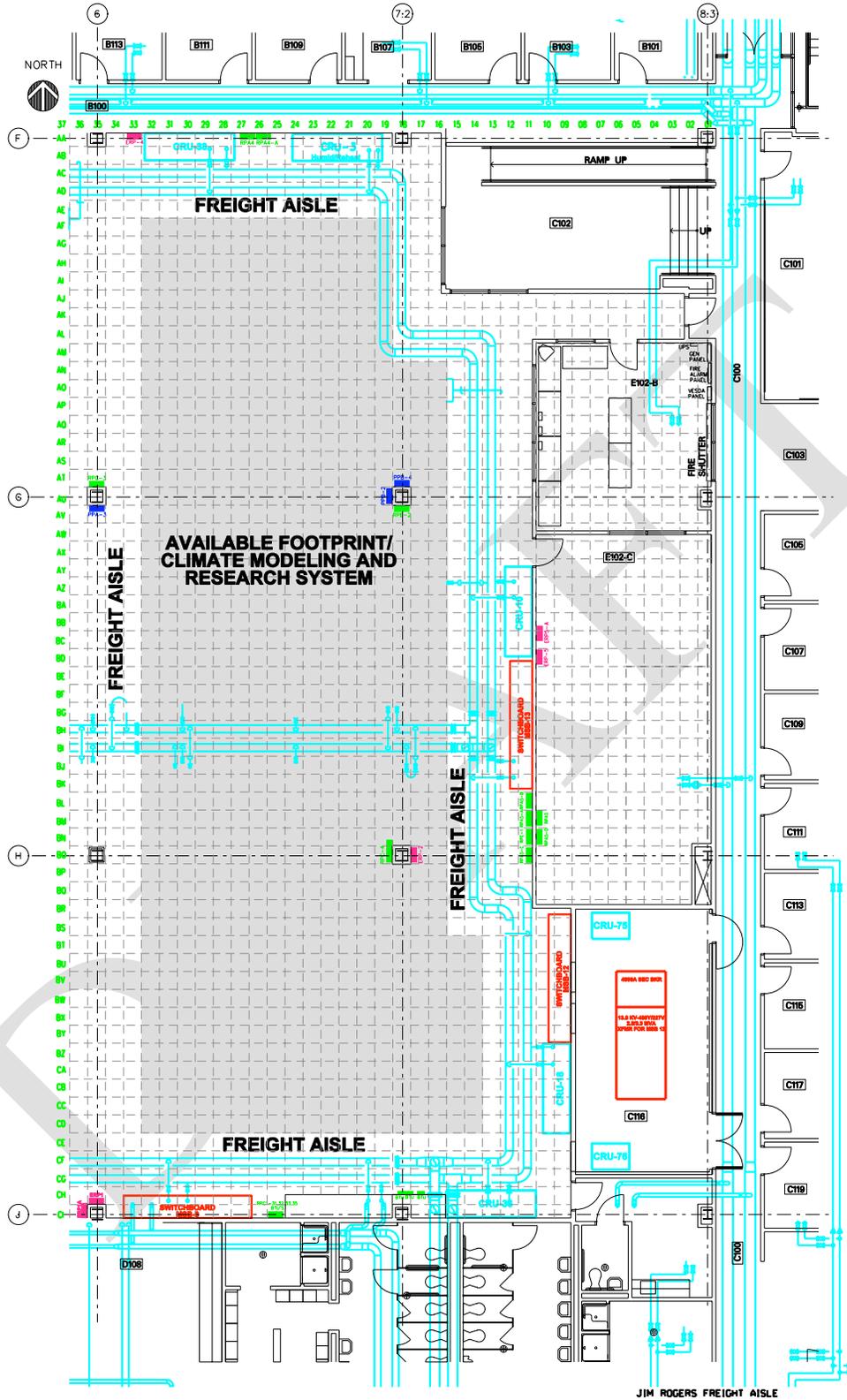


Figure 12. Available Floor Space for Climate Modeling and Research System