Expanding a Supercomputer Facility Using Modular Data Center Technology

ABSTRACT

With the expansion of high-end computing resources needed to support NASA's increasing demands for physics-based simulations, the facility housing Pleiades—the agency's largest supercomputer—recently reached its power and cooling capacity. In response, the NASA Advanced Supercomputing Division at Ames Research Center undertook a prototype project that resulted in a new facility based on modular data center technology. The facility, a ~1000 square-foot module on a concrete pad with room for 16–18 compute racks, was completed in fall 2016 and an SGI computer system, named Electra, was deployed there in early 2017. Cooling is performed via an evaporative system built into the module, and preliminary experience shows a Power Usage Effectiveness (PUE) of ~1.03. Electra achieved over a petaflop on the LINPACK benchmark, sufficient to rank number 96 on the November 2016 TOP500 list. The system consists of 1,152 InfiniBand-connected Intel Xeon Broadwell-based nodes. Its users access their files on a facility-wide file system shared by all compute assets via Mellanox MetroX InfiniBand extenders, which connect the Electra fabric to Lustre routers InfiniBand fabric over fiber-optic links about 300 meters long. The prototype has exceeded expectations and is serving as a blueprint for future expansions.

KEYWORDS

High Performance Computing, Modular Data Center, Supercomputer

1 INTRODUCTION

For the last two years, growth of the NASA Advanced Supercomputing (NAS) Division's compute resources was hampered by limited power and cooling capability in its facility space at NASA's Ames Research Center (ARC), Moffett Field, California. For nearly a decade, NASA scientists and engineers used the Pleiades supercomputer at NAS to perform physics-based simulations. The system evolved over time, integrating new processor types and interconnect fabric technologies as they became available. Today, Pleiades consists of over 11,000 compute nodes based on four generations of Intel Xeon processors: Sandy Bridge, Ivy Bridge, Haswell, and Broadwell. Its interconnect is Fourteen Data Rate (FDR) InfiniBand in a dual-plane, partial 11-D hypercube topology. Pleiades is currently ranked #13 on the November 2016 TOP500 list at 5.952 petaflops (LINPACK) and 7.107 petaflops peak performance [5]; it is #9 on the November 2016 HPCG Performance List [1].

Starting with the integration of new Haswell-based nodes in 2015, the facility housing Pleiades could not support the power and cooling requirements of new compute nodes without the removal of older-generation nodes that were still cost-effective to run. The early retirement of productive nodes effectively drove up the cost of providing additional compute resources.

NASA's High End Computing Capability (HECC) Project, which funds the facility and resources at NAS, predicts that supercomputing requirements will to continue to grow exponentially over time, as the agency leverages high-end computing to pursue its challenging missions. To meet its rapidly increasing computational requirements, HECC provides regular funding for NAS to significantly upgrade and replace the supercomputing resources that it hosts for the agency. It is critical, therefore, for NAS to realize the full value of adding new compute resources by overcoming the limitations of its current facility space.

In this paper, we detail how NAS accomplishes this through the use of modular data center technology. In the next section, we examine NASA's requirements and goals for the new facility space. In Section 3 we detail the facility design and construction process. In Section 4, we describe the design of the supercomputer system, named Electra, and explain how the system has been integrated with existing NAS resources. Section 5 details our experiences with the performance of the facility and system. We conclude with lessons learned and projections about where we expect NAS to be two years from now.

2 PROJECT APPROACH

2.1 PROJECT GOALS

The NAS Division at Ames Research Center (ARC) operates NASA's primary high-performance computing (HPC) facility under the High-End Computing Capability (HECC) project, with users broadly distributed across all NASA mission areas, most NASA centers, and numerous partner organizations (universities, corporations, and other organizations performing work for or with NASA). With more demand for computing services than supply, HECC needed more computing forcing NAS to physically expand.

At the time the need for expansion was realized, in July 2015, the NAS Division had two different facilities for housing computer equipment. The largest of these is in Building N258 at ARC. The computing floor area is approximately 14,000 sq. ft. and houses the Pleiades supercomputer [3], multiple filesystems, and archive storage. The primary cooling system is provided via a chilled water plant. The estimated Power Usage Effectiveness (PUE) [4] rating is ~1.3. The second facility is in Building N233A

[*] An employee of ....

[†] Responsible government official for the work. Please direct any correspondence about the project to ....

and houses the Merope supercomputer in 6600 sq. ft. of floor space. Also cooled by a chilled water plant with an estimated PUE of 1.5 [6].

Once it was determined that HECC had to expand its computing system services, NASA commissioned a trade-off study to evaluate its options: restructuring the existing N258 facility; constructing a new data center building; moving compute resources off-site to a larger, existing data center; utilizing commercial HPC Cloud resources; or installing a quickly deployable modular data center. The study indicated that the fastest, most cost-effective approach would be to deploy a modular data center. However, a large-scale HPC system in a modular building is not a common facility methodology in the HPC world, so NASA management decided that developing a prototype, or "proof-of-concept," facility would be a lower-risk approach before committing to a large-scale facility. NAS then undertook the Modular Supercomputing Facility (MSF) prototype project, which had the following goals:

- The new compute resources should represent the most cost-effective way of delivering compute resources that enable science and engineering returns to NASA.
- The new compute resources should be capable of being operated in an energy-efficient and environmentally friendly manner.
- The new system should interoperate with existing InfiniBand-connected Lustre and NFS filesystems, located about 300 meters away.

An important selection criterion for the MSF prototype was to be environmentally friendly. The existing N258 supercomputer facility is a traditional data center that uses a chilled water loop for heat transfer from the computer floor. With a constant compute power load of 4.0 megawatts (MW), N258 uses approximately 1.0 MW to power the chillers and cooling tower that cool the computer systems. In addition to its power usage, the N258 cooling system consumes an average of 50,000 gallons of water per day.

In the highly urban environment of Silicon Valley, where local cities have mandated water restrictions to their residents as a result of the California drought, good water management was a major requirement. Fortunately, the weather in the San Francisco Bay Area is very temperate and for most hours of the day, the outdoor air temperature is sufficiently cool for computer operation. Figure 1 shows psychrometric diagrams for the Ames Research Center and for the Kennedy Space Center to contrast how typical weather conditions affect data center operations in different locations. Finding a design that could exploit the natural San Francisco Bay Area weather was one of the primary goals.

2.2 Acquisition

The NAS team chose to site the MSF on an approximately 4000-square-foot lot across the street from N258. The team then requested potential vendors to propose a strategy to achieve a flexible, energy-efficient approach for deploying supercomputers on the site, and ultimately partnered with SGI on the project.

The MSF tasks were divided between two contractor teams: the infrastructure team, Cyber Professional Solutions and AECOM, and the computer and Modular Data Center (MDC) vendor team, SGI and CommScope. The infrastructure contractors were responsible for concrete slab construction and installation of site utilities, power, water, and drainage. The SGI/CommScope team was responsible for module and computer installation, as well as post-installation maintenance.

Modular data centers are noted for their rapid deployment and our experience is that the reputation is warranted. The schedule in Figure 2 describes the project's high-level tasks and durations. While this chart is not the actual schedule (the project had an unplanned 5-month stop-work unrelated to project performance) it does represent the actual durations, showing that modular data centers can be deployed in less than 6 months—even at government facilities.

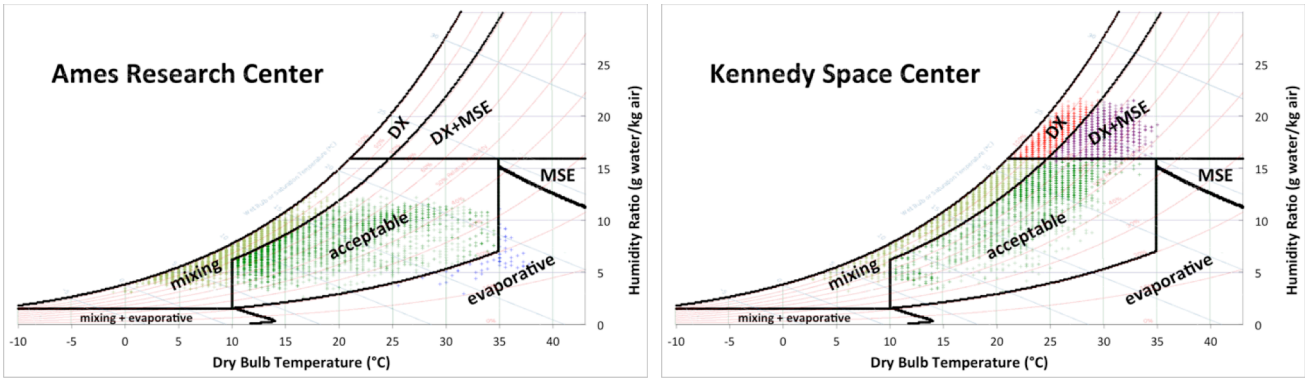


Figure 1: Psychrometric plots of approximately 1 year of hourly weather data from National Weather Service websites for two NASA centers. Ames shows nearly all points requiring no cooling whatsoever, although many require "mixing" of return air with outside air to dry it. A small number of points require adiabatic ("evaporative") cooling. Kennedy has a substantial number of points requiring direct expansion (DX) air conditioning, potentially in conjunction with multi-stage evaporative (MSE) coolers ("DX" and "DX+MSE").

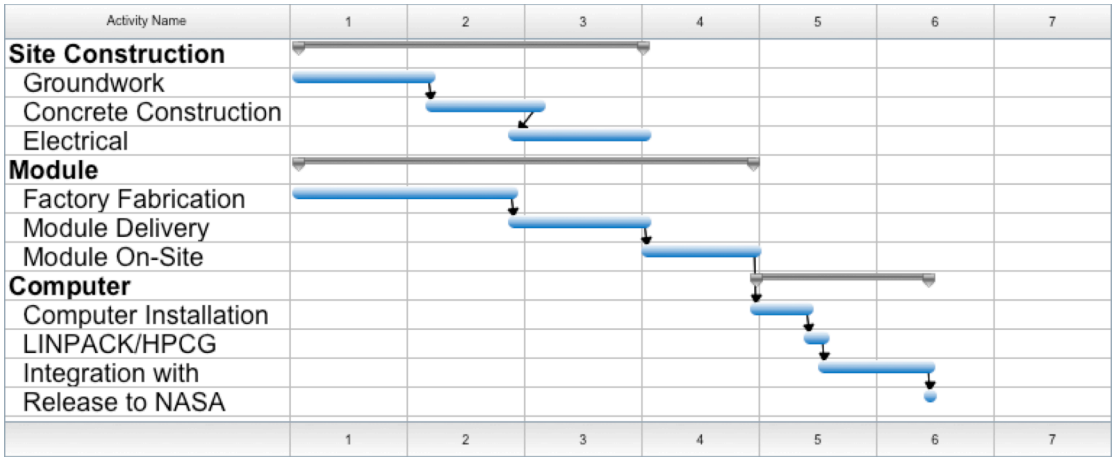


Figure 2: The high-level tasks and durations of the MSF project.

3 FACILITY DESIGN AND CONSTRUCTION

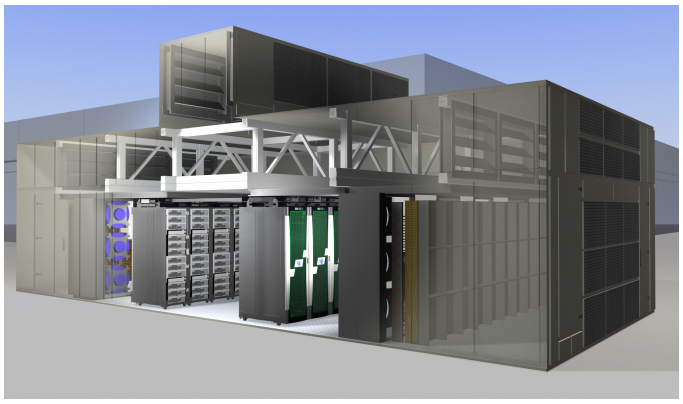
3.1 DESIGN CONSIDERATIONS

The MDC team was a partnership between SGI, the computer vendor; CommScope, the MDC integrator; and Saiver, the module manufacturer. The MDC design (see Figure 3) draws outside air into each end of the module via two fan banks of 12 centrifugal fans each. The air is filtered (and conditioned if required) before it travels into two separate cold aisles and through the racks, and then exhausted into a common hot aisle in the middle of the module for release back to the outdoors.

A programmable logic controller (PLC) controls the operating environment, adjusting dampers for recirculation of hot air, solenoid valves to run water through the evaporative cooler, and fan speeds based on temperature and pressure sensors located inside and outside of the module. Power meters measure total power draw for the module and each of the four Starline busways that feed the compute racks.

The MDC is capable of holding twenty 24-inch wide racks, in two rows of ten each. For our installation, we installed only eighteen racks (sixteen compute racks and two I/O racks) with two blanking panels (to control air circulation) at the end of each cold aisle. The module's four busways each feed power to four compute racks. Each compute rack is powered by two 415VAC, 3-phase feeds through IEC309 32A, 5-pin connectors. The power distribution units (PDUs) within the rack distribute the 415V, 3-phase input as 240VAC single-phase to the 16 power supplies in the rack that output 12 VDC to the compute and fans. Powering the module at 415V, 3-phase eliminates the electrical losses associated with the traditional 480–208V step-down that occur in our N258 data center, saving 12 kW.

While the San Francisco Bay Area's weather is mild, there are days when the outside temperature cannot meet the 15–27°C (59–81°F) desired cold aisle settings. Fortunately, hot days are almost always complemented with low relative humidity and cooling the supply air can easily be accomplished with an evaporative cooler. When the outside air temperature exceeds 27°C, air is drawn through an evaporative media (an impregnated glass fiber, honeycomb-like material) that has been saturated with water. The heat in the air evaporates the water as it passes through the media, raising the humidity of the air stream and lowering the temperature of the air. We estimate that evaporative cooling will be required about 30 hours per year,



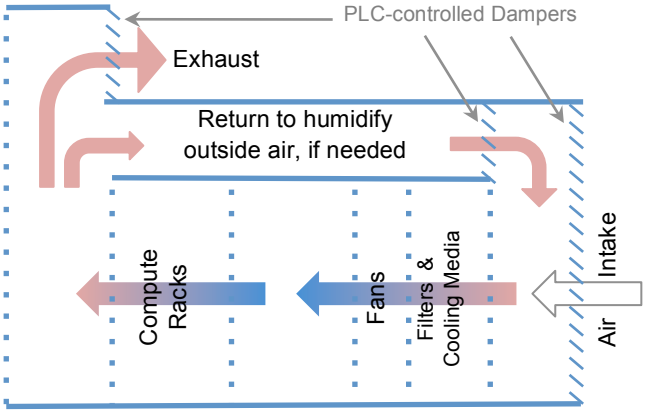


Figure 3: Design of the Modular Data Center; the line drawing shows an elevation view of the airflow for the right half of the rendering.

requiring about 9 gallons per minute of water, for a total yearly water consumption of 16,000 gallons.

Solenoid valves control the wetting of the evaporative media by pulsing on and off to limit waste. The evaporative cooler has four columns of media, and the PLC controls which columns to wet to provide the proper amount of cooling. The water is run through the evaporative media only once without the use of any water treatment, which allows it to be drained into the center storm drain system. Measurements show that the amount of water lost due to evaporation or drainage in the module is 1% of the amount that would be needed if the Electra racks were installed in the conventionally cooled computer facility in N258.

On cold days below 15°C (59°F), the air in the hot aisle can recirculate back to mix with the incoming supply air to increase the air temperature and/or lower the relative humidity (RH) below the 80% maximum setting. In the San Francisco Bay Area, the recirculating configuration is quite common, as the outside supply is at or above 15°C/80% RH or less for half of the yearly hours.

3.2 SITE PREPARATION

The site selected for the MSF prototype was bare dirt. NASA's Ames Research Center sits on the southern end of the San Francisco Bay, and the site is about a mile from the bay, so flood mitigation was required. Groundwork included raising the level of the site with compacted engineered fill to an elevation above a 50-year flood level—13.5 feet above sea level. In addition, existing storm drainage was relocated away from the MSF site.

While utilities were available close to the site, there were none running directly to it. After the site groundwork was complete, utility duct banks were dug for the power conduits, communication conduits, and water supply and drain. Over 600 feet of high-voltage, underground, concrete-encased duct bank was installed to carry power from one of ARC's substations to the MSF site. Four hundred feet of underground, concrete-encased duct bank was installed to house a 144-strand bundle of single-mode fiber cable for network communications. A 2-inch water line was hot-tapped into an existing underground 10-inch water main to provide domestic cold water to the site and a 4-inch storm drain line was installed to carry away domestic clean water as well as rain water. The final water connection at the module is a ½-inch line at each end of the module.

After the site groundwork was complete, the concrete slab foundation for the module was poured. The 45×50 foot slab is 16 inches thick with two courses of #6 rebar on 12-inch centers, and was designed to hold two modular data centers weighing up to 135,000 lbs. each. The slab was poured in one day as a monolithic slab and required 20 concrete truckloads to complete. Embedded into the slab are twelve 8×10×0.75-inch thick steel weld pads that the module is welded to for seismic restraint due to the site's earthquake prone location.

Once the concrete slab cured, the transformer and switchgear were set onto the slab over conduits for power conductors, which were run under and up through the slab. Power is delivered to the MSF site via 15 kV-rated conductors directly fed from a dedicated breaker in the substation to a 2800 kVA transformer at the site. The transformer steps down from 13.8 kV to 415V, 3-phase power, which is then distributed through onsite switchgear with each module having its own dedicated breaker. Ten sets of 750-kcmil conductors run from the transformer secondary to the switchgear main breaker. Four sets of 500-kcmil conductors run from the switchgear into the module's electrical panel.

3.3 MDC ASSEMBLY

The MDC was manufactured in Monza, Italy and shipped to California in ISO containers, arriving at NASA Ames in February 2016. Because of the previously described delay, assembly did not commence until September 9th. At that time, the 15 separate sections that comprise the MDC were moved into position via a forklift and fastened together on site to construct one 21×46 foot module. Assembly of the module with roofing and interior electrical was completed on September 19th, just in time for the installation of 18 racks into the module. Late September's tasks included installation of the fire suppression system and cabling of the compute racks. The module earned its ETL Mark electrical certification after an October 11th inspection and the Ames Fire Department provided occupancy approval the next day, after the successful completion of a fan integrity test to verify the module seal for the Novec 1230 fire suppression system and functional testing of the fire detection and alarming system.

4 SYSTEM DESIGN AND INTEGRATION

The computational system installed in the MDC is known as Electra. It consists of 1,152 nodes, installed in 16 racks, and has a theoretical peak performance of 1.23 Pflop/s. Each node has two Intel E5-2680v4 (Broadwell) processors (28 cores per processor) with 128 GB of memory. The nodes are interconnected via two independent FDR InfiniBand fabrics in a hypercube topology. I/O traffic is isolated to the ib1 fabric and MPI/system communication is primarily on the ib0 fabric.

The compute system is mostly self-contained within the module, but from a user perspective, it has been highly integrated with the Pleiades system in N258. Electra users log into Pleiades front ends, and batch jobs have full access to the shared NFS and Lustre filesystems serving all of HECC, which are located in the primary compute facility in N258.

A single Portable Batch System (PBS) server manages all of the jobs for Pleiades as well as Electra. User jobs are routed to the appropriate system by specifying the hardware model types associated with each system.

Access to the filesystems is facilitated through Lustre and IP routers connected via Mellanox MetroX long-haul InfiniBand switches and Obsidian Longbow range extenders (see Figure 4). The Lustre routers were proven in another deployment. But with Electra, IP routers were utilized to minimize the InfiniBand connectivity required on the NFS servers. In the other (previous) deployment, additional InfiniBand host channel adapters were added to multi-home the NFS servers on all of the required IB fabrics.

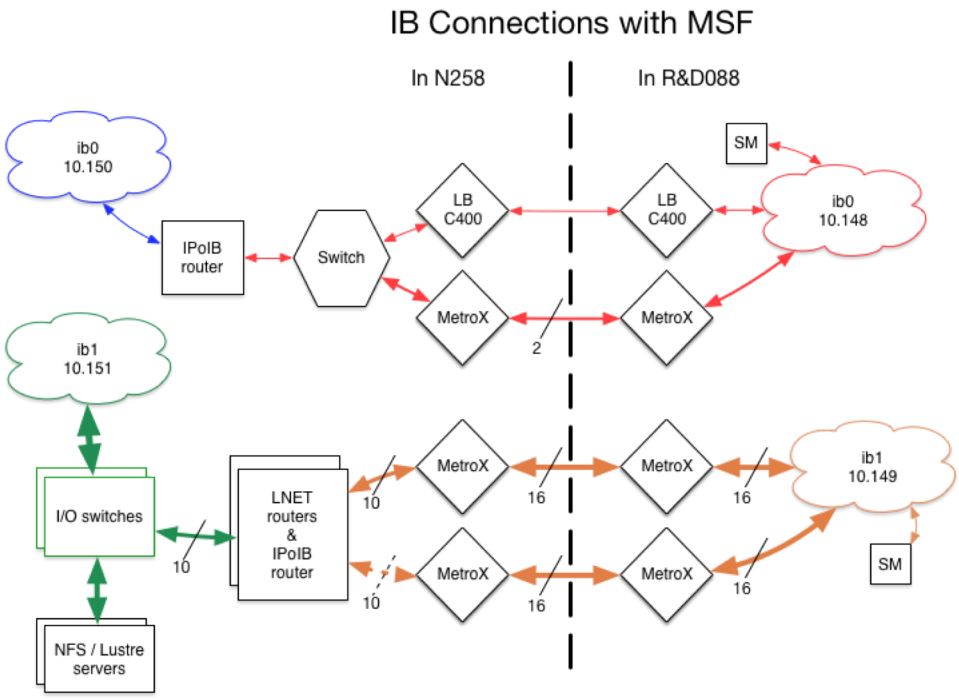


Figure 4: Components connecting Electra's compute nodes to user filesystems on Pleiades 300m away in N258.

One of the team's design decisions was to integrate the management of Electra with the Pleiades system to minimize the system management overhead. Even though Electra and Pleiades are on independent InfiniBand fabrics, they share common access to the filesystems, batch scheduler, and cluster management infrastructure.

The cluster management infrastructure used to provision and manage the Pleiades and Electra cluster utilizes the same software and hardware components, so the cluster can be managed as a single instance rather than two independent systems. This reduces the amount of required hardware and labor, and maintains consistency between the systems. Although they are managed as a single system and highly integrated, the systems are distinct from the user's perspective: jobs are not allowed to span the two systems due to bandwidth limitations.

5 OPERATIONAL EXPERIENCE

5.1 MODULE PARAMETERS

The MDC is controlled by the PLC, which queries sensors throughout the module and makes adjustments to maintain the internal environment within the set operating parameters. Access to the operational control is provided through a user interface (UI) that displays the current module conditions (see Figure 5) on a monitor in the module as well as in the N258 main control room and the N258 facility engineering office. Each heading can be clicked to review detailed measurements for temperature sensors, fan performance, and power/voltage/current draw. The data that is displayed on the UI is sampled every 10 seconds and stored in a log file for trending and evaluation.

The computer system's power draw, in kW, is typically in the mid-300s, averaging 20–22 kW per rack. (For comparison, LINPACK testing was conducted with an average power of 439 kW). At 22 kW per rack, the fans on the back of the SGI rack are moving 3000 CFM of air per rack. Originally, the SGI racks were not sealed for a tight cold-aisle/hot-aisle configuration, and the module's supply fans had to be overdriven to over-pressurize the cold aisle and to keep air from the hot aisle from being drawn back into the cold aisle by the rack fans. When the racks were first powered on to run diagnostic tests, there were several instances where nodes powered off due to overheating because

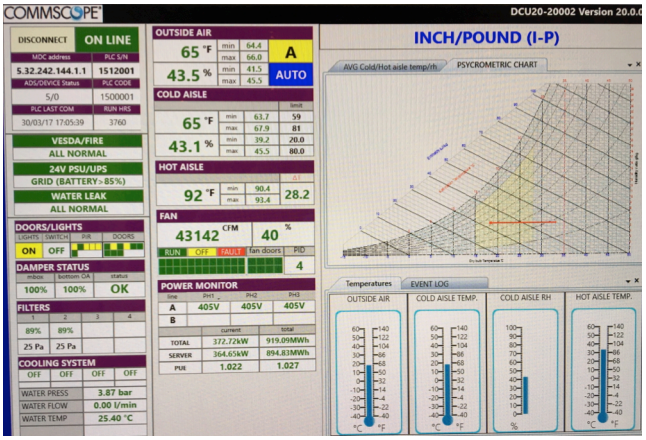


Figure 5: User interface for the module's PLC, which controls the climate inside the module.

the module's fans were set too low. While overpressurizing the cold aisle was effective in cooling the nodes, it was to the detriment of total power consumption, and consequently the PUE. For the initial setup, depending on how conservative the control parameters were, fan power ranged from 15–45 kW.

Subtle changes were made to the module that improved air management and lowered the fan power. Large metal mesh debris filters in the exhaust airflow of the module were removed to reduce pressure in the hot aisle. Open spaces between the I/O racks and the adjacent compute racks were sealed with large blanking panels. The decorative openings in the top sheet metal enclosure above each rack to contain cables were sealed. The gap between the floor and the rack bottom as well as the gap between the rack sides and module's wall, were sealed. Finally, the open area above the top nodes was blanked off, basically extending the top of the rack. After these changes were made, processor temperature testing was conducted to determine the best operating settings.

The target processor temperature is 70°C, which is a typical high temperature on a 22 kW Broadwell rack in the 20°C (68°F) N258 computer room floor. For temperature testing, diagnostic software was run to provide a stable power load on each processor. The total power load was 400 kW (25 kW per rack), about 15% higher than a typical workload. SGI rack management software recorded the processor and air intake temperatures while the module's PLC recorded airflow. The test results from Rack 16, shown in Figure 6, are representative of the racks in the module. The nodes at the top of the rack are most affected by poor air separation, with hot air spilling over the top of the rack. By setting the module's fans to supply 43–45,000 CFM (air flow total to 2 cold aisles), processor temperatures drop below the 70°C target. When module supply airflow drops below 40,000 CFM, the compute rack fans pull air from the hot aisle to meet their 48,000 CFM requirement (16 racks @ 3000 CFM/rack). As shown in Figure 6, increasing airflow continues to reduce processor temperature, but at the cost of additional fan power. To keep fan power at a minimum, the module is operated at 43,000 CFM, which requires 8 kW of power.

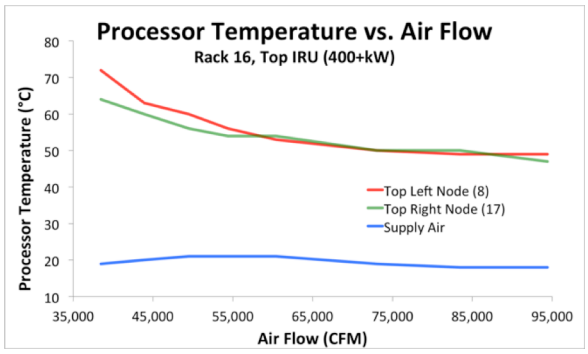


Figure 6: Effect of airflow on processor temperature.

5.2 FACILITY PERFORMANCE

In this section we present more detailed information about recent operating trends. The time period presented in each graph is from data recorded March 1st through March 14th of this year. The data is representative of the facility's performance since it went in production in January 2017.

The MSF has only been in operation for four months and has not operated in summer weather. While the cold aisle temperature will increase, we do not expect to change our operating parameters as summer progresses other than to see an increase in the running of water. Testing of the evaporative system has shown that passing water over the evaporative media does not require an increase in fan power to improve airflow.

Current operating settings for the cold aisle are a temperature range of 15–27°C (59–81°F) with relative humidity at 20–80%. The lowest airflow setting is 43,000 CFM, which is a setting of 40% fan speed in the PLC drawing only 350 watts per fan. The airflow will never drop below this setting and will only increase when the hot aisle exceeds 44.5°C (~112°F). A 44.5°C hot aisle means that the difference between the cold aisle and hot aisle temperatures, Delta T, is 17.5°C (~31°F), at the maximum cold aisle of 27°C.

PUE: Figure 7 presents a rolling 15-minute average of the PUE over our two-week data period. Electra's power draw varies from a low of 250 kW to a high of 360 kW. The PUE is

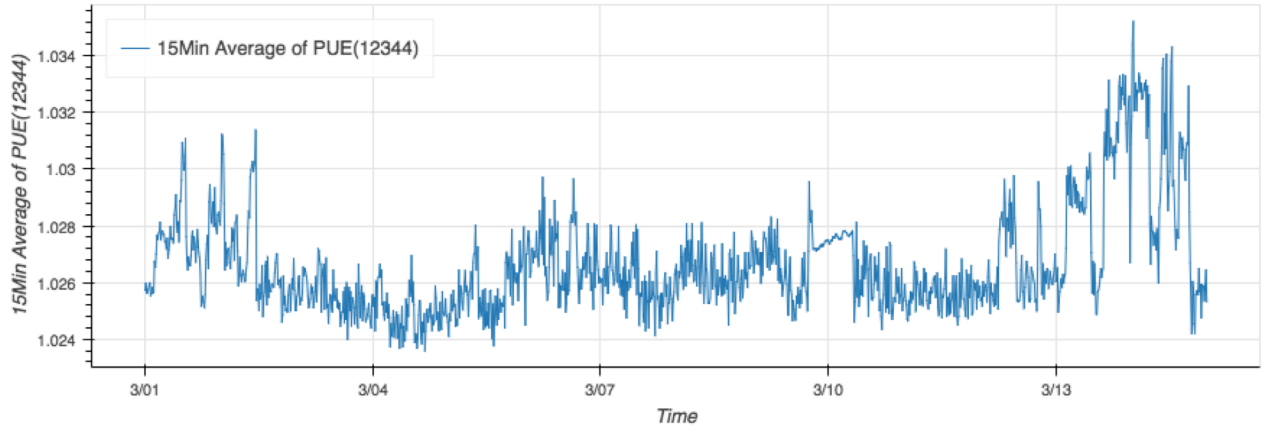


Figure 7: PUE for the MSF over a two-week period in March 2017.

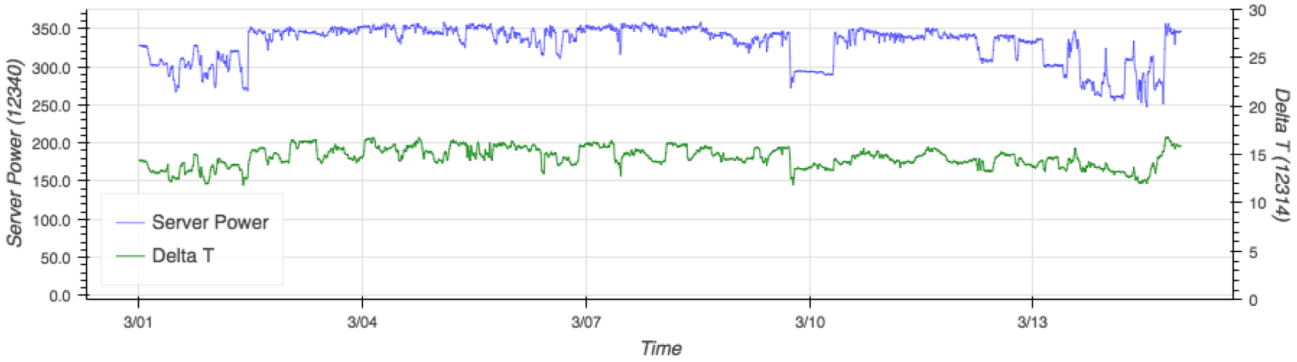


Figure 8: Power consumption and Delta-T for the MSF over a two-week period in March 2017.

consistently under 1.03 when Electra is performing a typical workload of 300 kW or higher. Since the module's power consumption is constant at 8 kW, the PUE only rises above 1.03 when the batch processing system is collecting nodes to run a large job or nodes are taken out of service for system testing.

Electra's Power Effect on Delta T: With the module's fans set to a constant airflow of 43,000 CFM, the hot aisle/code aisle Delta T is dependent on Electra's power consumption. As shown in Figure 8, the Delta T trends at 15-16°C for normal workloads of 350 kW. For the module's control logic, Delta T is not limited until the hot aisle temperature exceeds 44.5°C, at which time the airflow will increase to maintain the hot aisle maximum setting. It should be noted that hot aisle control is just one of five fan strategy programs that can be used to control the module's environment. Each program controls airflow and Delta T in slightly different ways. Because of the significant airflow on the SGI rack fans, hot aisle control has been identified as the best fit for our application, but other strategies may be considered as experience is gained from more operating time with the module.

Damper Settings Dependent on Outside Temperature: Figure 9 shows how the module's dampers adjust to maintain a constant cold aisle temperature over a fluctuating outside temperature. When the outside temperature falls, the outside air damper changes from 100% to 20% open while the recirculating air damper acts in a complementary fashion and changes from 95% to 65% closed (35% open). As shown in Figure 9, the adjustable dampers allow for a 16°C cold aisle temperature while the outside air varies from 5° to 20°C.

Although not shown in the figure, the dampers also adjust to maintain the cold aisle humidity even if it means raising the cold aisle temperature. In cases where the outside air is within the cold aisle range 15-27°C, but the outside air relative humidity is over 80%, such as on a rainy day, the recirculating air damper will open to mix hotter air with the incoming air to lower the relative humidity below the set point. In the San Francisco Bay Area, 27°C dry-bulb temperature days with 80% humidity do not occur, so there is little concern about inability to control the humidity in the module.

Cold Aisle Temperature Compared to Outside Temperature: Similar to Figure 9, Figure 10 compares the cold aisle temperature with the outside temperature, but replaces the damper positions with the quantity of water used and the wet-bulb temperature. The wet-bulb temperature is slightly lower than the lowest temperature achievable by the evaporative cooler, so it is shown to provide a reference point.

For this data set, the maximum cold aisle temperature was set at 23.9°C (75°F) to take advantage of a March warm spell to test the evaporative cooling system. As shown in Figure 10, the cold aisle temperature matches the outside temperature peaks until March 12, when the outside temperature exceeded 23.9°C,

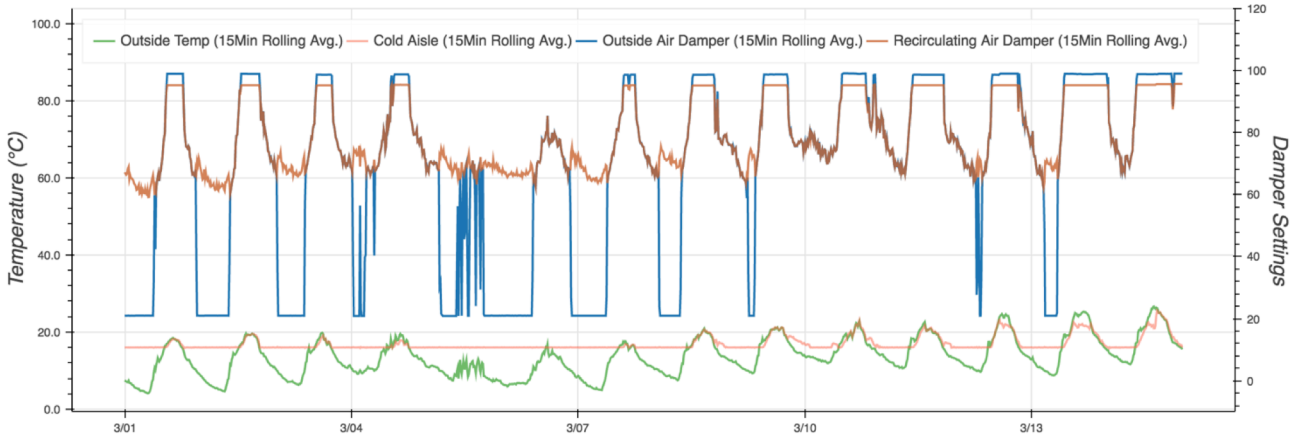


Figure 9: Damper settings and outside temperature over a two-week period in March 2017.

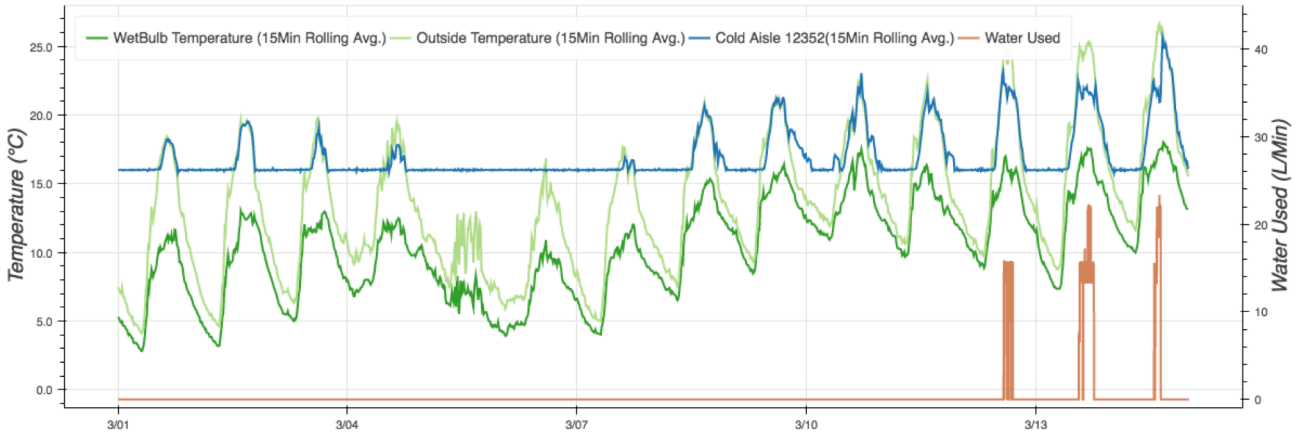


Figure 10: Power consumption and Delta-T in the MSF over a two-week period in March 2017.

causing water to flow through the evaporative cooler. Figure 11 enlarges a 5½-hour window on March 12th, presenting the cold aisle temperature reduction as a result of water running through the evaporative media with the outside temperature exceeding the 23.9°C set point. Two of the four evaporative media columns were used to provide a 3°C drop on March 12 and 4°C drop on March 14. This increases the confidence of the NAS facility team in the MDC's ability to maintain a 27°C cold aisle maximum during the summer months.

Of additional interest is the sensitivity of the control sensors to stop the flow of water each time the outside air falls below 23.9°C. For the March 12 afternoon, 1969 liters (520 gallons) were used over 2½ hours, averaging to 13.5 liters per minute (3.6 gpm) to cool Electra while operating at 340 kW.

5.3 SYSTEM PERFORMANCE

After the assembly of the MDC and the installation and cabling of Electra's computer racks were completed in October 2016, SGI ran diagnostics on Electra to test it and to provide a heat load for testing the module's environmental systems. As part of that testing, they ran LINPACK to get a number for the TOP500 list. On October 20th, 6 weeks after the start of module assembly and on its very first attempt, Electra achieved an Rmax of 1.096 Pflops/second, which was sufficient to place it in the top

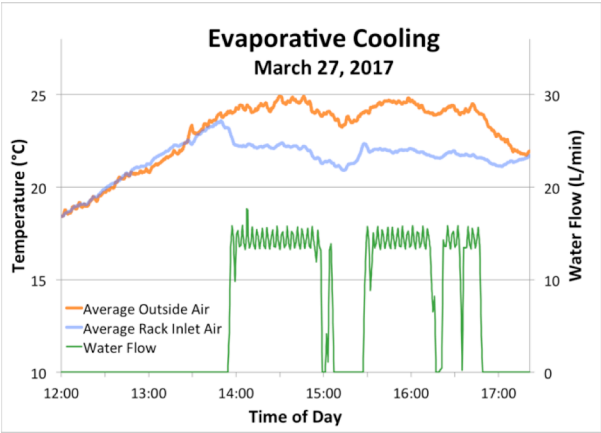


Figure 11: Evaporative cooling over a 5½-hour period.

100 when the list came out in November 2016. While facility information is not available for some systems in the TOP500 list, we believe Electra to be to top system in the list that is module-based. SGI also ran the HPCG benchmark, which measured 25.2 Tflops/second, sufficient to place Electra at #46 in the world on the November HPCG list.

After SGI finished their installation, the NAS systems team configured the system software, and then the application performance team conducted its tests. With the exception of the I/O subsystem, Electra is very similar to the Broadwell subsystem of Pleiades, which they had previously tested. Given that fact and the 300m distance to the user filesystems, the team focused on the performance of the I/O infrastructure described in Section 4. In particular, they wanted to determine how many Lustre routers were required to achieve similar I/O performance to Pleiades.

The systems team planned to use up to 10 Lustre routers in the I/O infrastructure (these are called "LNET routers" in Figure 4). In order to test the sensitivity of I/O performance to the number of routers, they varied the I/O configuration to have 2, 4, and then 10. For each setup, the applications team used a variety of applications from the standard workload on Pleiades to measure I/O performance. Some applications ran stand-alone, filling the system, while others were run with a variety of user jobs sharing the system. The team found that application performance suffered significantly when only 2 or 4 Lustre routers were available, and consequently the full complement of 10 was put into production.

6 LESSONS LEARNED AND FUTURE WORK

Facing a situation where its primary supercomputer facility could not economically support the addition of new equipment, the NAS Division elected to conduct an experiment to study whether high-performance computing equipment was compatible with modular data center technology. Specifically, they undertook the design, installation, and deployment of the MSF prototype, which resulted in the Electra supercomputer being housed in a ~1,000 square foot module about 300m from their primary facility. In this section we summarize the lessons

learned from the experiment and also describe how this will impact future expansion at NAS.

Overall, the prototype has been a huge success. With the exception of a 5-month unrelated delay, the facility went from an unimproved lot to a production supercomputer in less than 6 months. The additional compute resources increased the capacity available for NASA's physics-based simulations by ~16%.

Operationally, the biggest lesson learned is that cold-aisle/hot-aisle containment is required to achieve a low PUE. Poor containment can be overcome by increasing airflow, but it increases facility power consumption and PUE.

Installing a 27 kW-per-rack supercomputer is not a common use case for a modular data center. Installing power into the module had its difficulties. The four sets of THHN 500-kcmil conductors could not make the tight radius required to land the conductors in the module's electrical switchgear panel. Clever onsite field engineering modified the module design with an oversized junction box and a conductor splice from THHN to highly flexible 535-kcmil DLO cable from the box into the panel.

A low-restriction exhaust is key at power densities of 1000 W/sq.ft. Any component that reduces the free air opening of the exhaust needs to be selected carefully to mitigate the risk of inadvertent heating of the hot aisle.

It is important to keep the core group of partners as small as possible, ideally, all within one company. Multiple designers, builders, and vendors lead to interface mismatches that need to be corrected on site. One case in point was that the pad designer's flatness callout for the concrete pad was misunderstood by the contractor, who added a slope for drainage, necessitating thick shims to be fabricated on site to bring the module back to level.

Data collection proved to be difficult to implement. The NAS Division's standardization on Mac and Linux was an issue because the user interface/PLC operates on Windows. This issue was further complicated by originally requesting compatibility with the server-monitoring tool Nagios. That proved to be too labor intensive for NAS to implement and support in a Windows environment, so the requirement was dropped. It is important to have a thorough understanding early in the process of how module instrumentation data can flow to where it will be analyzed.

Building on the success of the prototype, NAS plans to use modular data center technology for two future expansions. One, which is already underway, will add a second module on the pad next to the first. Rather than using outside air for cooling, NAS will aim to gain experience with warm-water cooling in the new module. While the Electra system uses a traditional hot/cold-aisle setup, the next iteration will utilize the HPE[‡] water-cooled E-Cell technology, which allows for a higher density configuration. Although the PUE is expected to go up, the Total-power Usage Effectiveness (TUE) [2] is expected to improve. This is due to the integrated fans in Electra being counted as compute load for its PUE calculation and as cooling load in its TUE calculation.

In the longer term, NAS plans to significantly expand the facility space available to it. A procurement is underway that is requesting vendor proposals for computer systems and facility space that would support a growth path to ~10MW of equipment over the next 5 years. Because of the success of the MSF, it is expected that many vendors will propose solutions based on modular data center technology along the lines of the MSF project.

[‡] SGI was acquired by HPE in November 2016.

ACKNOWLEDGMENTS

The authors wish to acknowledge….

REFERENCES

[1] HPCG Benchmark. http://www.hpcg-benchmark.org
[2] M. Patterson, S. Poole, C-H. Hsu, D. Maxwell, W. Tshudi, H. Coles, D. Martinez, N. Bates. TUE, a new energy-efficiency metric applied at ORNL's Jaguar. https://pdfs.semanticscholar.org/2eb5/b7171842f85899faa9cf4146a97d9084e76b.pdf
[3] Pleiades Supercomputer. https://www.nas.nasa.gov/hecc/resources/pleiades.html
[4] PUE definition. https://en.wikipedia.org/wiki/Power_usage_effectiveness
[5] TOP500. https://www.top500.org
[6] Merope Supercomputer. https://www.nas.nasa.gov/hecc/resources/merope.html