Trinity Facilities and Operations Planning and Preparation: Early Experiences, Successes, and Lessons Learned

Ron Velarde, Carolyn Connor, Alynna Montoya-Wiuff, and Cindy Martin,
Los Alamos National Laboratory

Abstract

There is considerable interest in achieving a 1000 fold increase in supercomputing power in the next decade, but the challenges are formidable. The need to significantly decrease power usage and drastically increase energy efficiency has become pervasive in the high performance computing community, extending from chip design to data center design and operations. In this paper the authors present a short summary of early experience, successes, and lessons learned with respect to facilities, operations, and monitoring of the New Mexico Alliance for Computing at Extreme Scale (ACES), a collaboration between Los Alamos National Laboratory (LANL) and Sandia National Laboratories (SNL), Trinity Supercomputer during the facility preparation and pre-acceptance testing phases of the project. The Trinity Supercomputer, which is designed to exceed 40 Petaflops/s, is physically located at Los Alamos' Strategic Computing Center (SCC) and is a next step toward the goal of exascale computing (a million, trillion operations per second). Discussion topics include facilities infrastructure upgrades, Sanitary Effluent Reclamation Facility (SERF) water use, adaptive design and installation approaches, scalability and stability of monitoring systems, and early power-capping investigation results.

Keywords: High Performance Computing, SERF, HPC monitoring, Data Center, facilities infrastructure

1. Introduction

The supercomputers of the future will not only be extremely powerful, but will also need to be much more energy efficient and resilient than current designs. One of the most important obstacles to achieving the next three orders of magnitude performance increase in large-scale computing systems is power, and by extension, cooling. Gaining efficiencies in data center design and operation has become an ongoing focus of interest and investment. Current multi-petascale systems require several megawatts of power (e.g., Trinity's design specification was not to exceed 12 MW of power). The 2008 Exascale computing study projected that Exascale systems will consume one to two hundred megawatts if the power issues are not addressed. [1]

The installation of powerful supercomputers is no small task, and in fact planning and preparations for a new platform begin years in advance of its physical arrival. The Trinity supercomputer is the first of the NNSA's Advanced Simulation and Computing (ASC) Program's advanced technology systems. Once fully installed, Trinity will be the first platform large and fast enough to begin to accommodate finely resolved 3D calculations for full-scale, end-to-end weapons calculations. The Trinity system will reside in the SCC in the Nicholas C. Metropolis Center for Modeling and Simulation. The SCC is a 300,000-square-foot building. The vast floor of the supercomputing room is 43,500 square feet, almost an acre in size. [2]

In order to accommodate Trinity and its successors, cooling and electrical subsystems supporting supercomputing in the SCC had to undergo major upgrades. Recent SCC facility upgrades and projects supporting LANL's programmatic and institutional supercomputing mission have included a dramatic expansion of LANL's SERF, adding warm-water cooling capabilities (75°F water supplied to the racks, as compared to 45°F chilled water), increasing overall water cooling capacity to the computer room floor to 15 MW, and enhancing the electrical distribution to the computer room floor to 19.2 MW.

Additionally, because energy conservation is critical, ASC Program staff conducted field trips to observe water, power, and cooling operations at supercomputing facilities around the country, including ORNL, NREL, NCAR, LLNL, and NERSC. Staff from Los Alamos visited the largest hybrid cooling tower operation in the country in Eunice, New Mexico. The trip to URENCO in Eunice aided in the evaluation of hybrid versus evaporative cooling tower technologies. The site visits inspired design changes in the SCC cooling towers, for example: the addition of a strategically located valve will save money as a result of cooling without recirculation during months when the outside air temperature can provide adequate cooling temperatures. The ultimate goal is to maximize the availability of computing platforms to the end users with minimum expense and effort required of the computing center.

In this paper we present a short summary of early experience, successes, and lessons learned with respect to facilities, operations, and monitoring of the ACES Trinity supercomputer during the facility preparation and pre-acceptance testing phases of the project. Section 2 discusses the success and impact of LANL's SERF. Section 3 discusses three specific examples of adaptive design and installation approaches that resulted in either increased efficiency, significant cost savings, or both. Section 4 discusses the importance of monitoring and some early experiences with the vendor-supplied monitoring system for Trinity. Finally, Section 5 outlines future work and some key priorities moving forward.

2. SERF Supplied Water Cooling

The advantages of using water-cooling over air-cooling include water's higher specific heat capacity, density, and thermal conductivity, which allow water to transmit heat over greater distances with much less volumetric flow and reduced temperature difference. Because both water and energy conservation are priorities, the recent facility upgrades included a shift to economical warm-water cooling technology, as well as a dramatic expansion of LANL's Sanitary Effluent Reclamation Facility, which was completed in 2013, to supply water to the SCC. Figure 1 shows the resulting decrease (approximately 20% reduction in FY14 and continuing trend in FY15) in institutional water use.

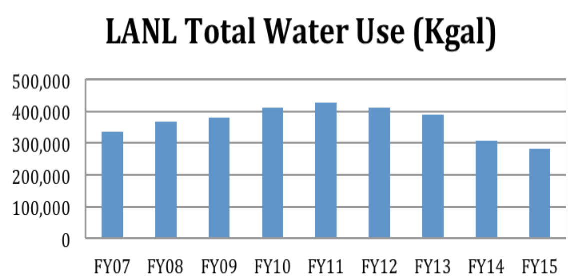

Figure 1 Graph showing LANL's total water use measured in Kilo-gallons by fiscal year. Operational upgrades to allow the SERF to supply the SCC were completed in 2013 allowing for significant institutional water savings.

LANL diverts treated sanitary wastewater and water flushed from cooling towers to the SERF [3], which can process up to 100 gallons per minute of sanitary effluent—or 120,000 gallons of water a day. It can produce up to 88 million gallons of water per year. In FY15, LANL set a record, exceeding 30 million gallons of SERF water used. Post upgrade, the SERF is supplying all water for the SCC facility, as shown in Figure 2, with a few noted exceptions that typically correspond to planned facility maintenance or short periods of low institutional water use that limit the quantity of water available for reclamation.

The immediate impact for the Laboratory of the use of SERF water in place of city/well water was an annual savings of tens of millions of gallons of well water per year. The longer-term impact is that LANL was poised to take delivery of the next large supercomputer, Trinity, and still stay below the site-wide annual limit of 51 million gallons of potable water. SERF's new robustness also means that future large supercomputers can be sited at Los Alamos with a sustainable water usage plan.

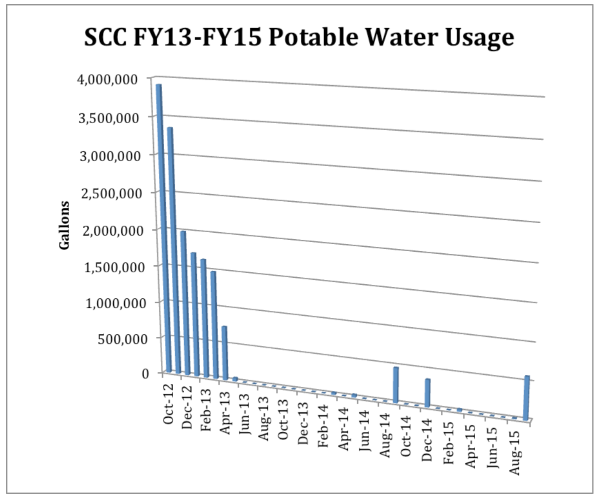

Figure 2 SCC fiscal years 2013-2015 Potable Water Usage graph illustrates the dramatic decrease in potable water use. The SERF is supplying all water for HPC in the SCC facility.

Currently LANL is considering additional options to increase capacity of the SERF, including expanding the water input beyond the wastewater treatment plant. The SERF is also looking into additional water storage at the facility and possibly more evaporation ponds. Finally, LANL is exploring ways to further increase cycles of concentration in the SCC cooling towers.

3. Adaptive Design and Installation Approaches

The LANL Facilities Team has adopted a philosophy of evaluating and incorporating adaptive design and installation approaches built upon historical best practices to realize efficiencies and cost savings when possible and advantageous. The power distribution approach is N+1 available power with rotary UPS to supply file systems, disk storage, and network switches; redundant utility power for mechanical systems; and raw utility power for compute racks and the building automation system

(BAS). In this section, we highlight three specific examples of adaptive facility design and installation approaches that resulted in significant cost savings, including a junction box design modification, installation of smart breakers, and the introduction of TC cable in our facility in preparation for Trinity

3.1. Junction Box (J-Box) Design Modification

As previously noted, preparation for Trinity included an emphasis on building relationships and sharing lessons learned and best practices among peer facilities. During a visit to NERSC, the LANL Facilities Team realized that the vendor design for supplying power to the Cray compute racks specified two 100-AMP feeds. Realizing the expense that would be associated with this approach, the LANL team conducted a detailed requirements review and proposed a modified single feed design for the Trinity installation. Both original and modified designs are illustrated in Figure 3. The modified (single 150-AMP feed) design was adopted with Cray's approval and resulted in a 50% overall cost reduction due to reduced material (i.e., fewer breakers, switchboards, bus way, and cabling to be installed) and associated labor savings, as well as an estimated 25% schedule gain. The estimated cost savings, due to the overall size of the Trinity project was $3M. One other peer facility has adopted this design and at least one additional facility is evaluating it for potential use.

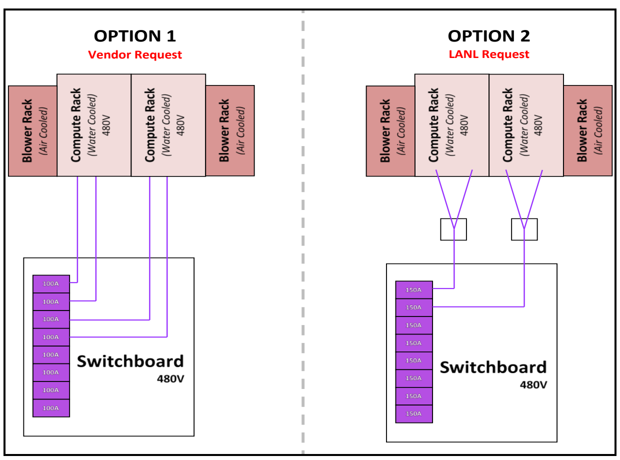

Figure 3 shows two facility designs for supplying power to the Trinity compute racks. Option 1 is the two 100-amp feed, vendor-proposed design. Option 2 is the modified J-Box design which uses a single 150 amp feed per rack.

3.2. Installation of Smart Breakers

A second example of an adaptive facility design and installation approach, the incorporation of smart breakers in the 480V Switchboard shown in Option 2, as illustrated in Figure 3, resulted in significant cost savings, and additionally important new insight. Smart breakers enable rack-level data collection, including validation of the integrity of the power supplied to each rack, as well as validation and trending of the actual per-rack power draw, both of which result in improved issue isolation and tracking. Exploitation of rack-level data collection at the smart breakers resulted in the collection of key information during Trinity pre-acceptance testing with respect to both power draw and power distribution, as we will now discuss.

As part of standard platform acceptance testing procedures, the computing system under test is exercised under a variety of conditions and workloads. The LINPACK benchmarks are a measure of a system's floating point computing power. They measure how fast a computer solves a dense n-by-n system of linear equations of the form $\boldsymbol{Ax} = \boldsymbol{b}$, which is a common task in science and engineering. Here A is an n-by-n matrix of coefficients, $\mathbf{x}$ is a column vector of unknowns with n entries, and $\mathbf{b}$ is a column vector of constants. The latest version of these LINPACK benchmarks is used to build the TOP500 list, which ranks the world's most powerful supercomputers. [4]

The first example of the impact of adding smart breakers to the facility design pertains to collection and monitoring of per-rack power draw data. During pre-acceptance testing, in the course of running the workload tests and ongoing monitoring and analysis of the data collected by means of the smart breakers, we observed incidences of Trinity compute racks exceeding the vendor's maximum power draw specification. In the most egregious instance, 92 kW of power was drawn during memory testing against a 74 kW maximum vendor per-rack specification.

Two benchmarks that were included in the workload package were HPL, the MPI implementation of the high performance LINPACK benchmark [5] and Cray's proprietary memory test diagnostic (HSW_MT). A highly regular dense LU factorization, HPL is computationally intensive and is recognized as an accepted LINPACK standard and thus the highest load one can put on a machine for purposes of exercising and characterizing supercomputers. Inquiry into the cause of the anomalous power-draw behavior revealed that the Cray diagnostic, HSW_MT, also known as "Memtest," has been designed to heavily load both the CPU and the memory. Thus, a higher power consumption was observed during HSW_MT than HPL, which is light on memory bandwidth use. This behavior (i.e., far exceeding the per-rack power draw specification) was neither predicted nor expected by the vendor and had obvious facility design implications, as well as potential operational impact, including tripped breakers, node shutdowns, etc. Fortunately, in this instance, these impacts were avoided due to the conservative power distribution design methodologies that were employed.

A second example of gained insight was a direct result of analysis of smart breaker data associated with power distribution. During pre-acceptance testing, LANL observed recurring, unexplained incidences of compute nodes powering down. Investigation efforts first focused on the water-cooling system and associated data (e.g., temperature, flow, etc.); however, nothing anomalous surfaced. Attention was then focused on the electrical system via analysis of the smart breaker data, paying specific attention to the measured supply voltage feeding each rack. Cray bench tested their power supplies against the LANL smart breaker data and the root cause of the behavior was determined to be power supply sensitivity on the high end, which resulted in node shutdown. Cray then experimentally determined a safe operating range for the chosen power supplies (-25% to +5%). LANL responded by adjusting the taps on the transformers to guarantee operation within the newly specified range and closely monitored system response. The availability of the smart breaker data allowed timely issue isolation and resolution with no further observed incidences. Cray was informed of a potentially serious issue of which they were previously unaware, and they worked collaboratively to resolve prior to the arrival of the full Trinity system. LANL was able to adapt its facility design to account for actual equipment behavior under LANL load conditions, thus avoiding potential operational implications.

3.3. Introduction of TC Cable

Historically, LANL has relied upon in-house assembly of electrical cable. In planning for the arrival of Trinity and to provide the required 24,000 linear feet of electrical cabling, the Facilities Team investigated alternatives that would gain efficiencies against tight construction schedules while fully meeting design specifications and requirements. Following analysis, the Team decided to pursue a preassembled alternative that would provide equivalent functionality and safety performance at a significantly reduced cost. Because the cabling must be Information Technology Equipment Room Approved per National Fire Protection Association standards (NFPA75), the Facilities Team worked through approval processes with the LANL Fire Marshall and other authorities having jurisdiction with the favorable outcomes of saved installation time (approximately a one month schedule gain for the team) and reduced labor costs resulting in an estimated overall project savings of an additional $1M.

Figure 4 The design choice to use pre-assembled cabling (TC Cable, as pictured) instead of in-house assembled cabling to provide the 24,000 linear feet of electrical cable required to connect Trinity resulted in an estimated $1M project cost savings.

Based upon LANL's analysis and experience, one other peer facility has adopted the use of TC Cabling in their facility and at least one additional facility is evaluating it for potential use.

4. Monitoring

As HPC platform scale continues to increase, systems are becoming concurrently more heterogeneous in computational, storage, and networking technologies. Furthermore, as the volume and complexity of critical operational information continues to increase, it will become impossible to efficiently manage platforms without tools that perform real-time run-time analysis continuously on all available data and take appropriate action with respect to problem resolution and power management. The smart breaker examples in the previous section illustrate the insight and essential information that system monitoring can provide in the design and operation of large and complex supercomputing platforms as well as the need for sophisticated and automated mechanisms for collecting operational data.

Operations management of the ACES Trinity supercomputer will rely on data from a variety of sources including System Environment Data Collections (SEDC); node level information, such as high speed network (HSN) performance counters and high fidelity energy measurements; scheduler/resource manager; and data center environmental data. The SEDC data provides information about voltages, currents, and temperatures of a variety of components at the cabinet, blade, and node level. This data also includes dew point, humidity and air velocity information. While the system utilizes many of these measurements to identify out of specification, and hence unhealthy, components, it relies on the crossing of fixed thresholds to trigger knowledge of an unhealthy situation. The node-level information provides high-fidelity energy measurements, OS (Operating System) level counters, and high speed network performance counters. Scheduler/ resource manager information provides time windows and components associated with user applications. Data center environmental data provides fine-grained power draw, information about noise on the power feeds, and water temperatures and flow rates. The increase in power density of HPC components has necessitated the use of water-based solutions for heat transport rather than traditional air-cooling solutions. This in turn requires feedback mechanisms to maintain proper water temperature, pressure, and flow rates as well as active fan control in the case of hybrid solutions. Thus, the water-cooled Cray XC platform requires a coordinated and comprehensive way to manage both the facility infrastructure and the platform due to several critical dependencies. As a first step to accomplish this, the facility employs multiple building automation systems, including a Tracer-Summit BAS to monitor mechanical systems, A Schneider Electric Power Management Expert (PME) system to monitor electrical systems, and an Environet BAS to record environmental information in the data center, including airflow (supply and return) room and rack temperatures, etc.

In February of 2015, LANL obtained an Application Readiness Testbed (ART) system known as Trinitite, a single cabinet Cray XC40, to prepare for Trinity with respect to the applications that will be run. Additionally, Trinitite provides a platform for validation of facilities–platform interaction and comparison.

Early experiences with the Cray monitoring network are as follows. During facility testing, we experienced a failure in SEDC data collection. We found that there is currently no mechanism for monitoring the SEDC data heartbeat, meaning that a data collection component may fail and that there is no knowledge that the data in question are no longer being collected, nor any mechanism for notification of failure. Secondly all data must pass through the System Management Workstation (SMW), noted in Figure 5. This impacts the amount of data that can be stored.

These observations have led ACES to develop a more scalable and cohesive infrastructure that includes both facilities and platform monitoring at Trinity scale. Thus, ACES is developing a mechanism to continuously remove data from the SMW for monitoring and analysis to mitigate the bottleneck and try to ensure collection of all relevant data.

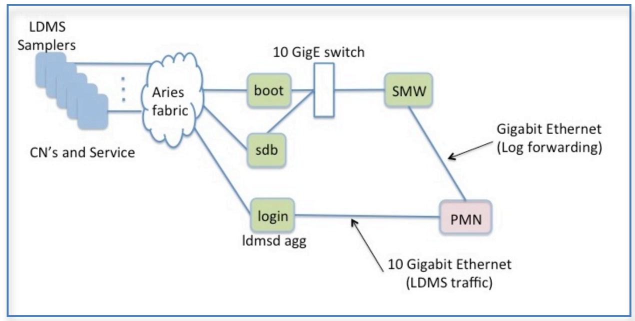

Figure 5 illustrates the Application Readiness Testbed (ART) system monitoring configuration

5. Future Work

Large-scale HPC platforms are continuing to push the limits of data center power and cooling infrastructure. Modern large-scale platforms with power draw requirements in the 20MW range can stress data center and site power infrastructure (e.g., power demands that change abruptly can cause power disruption in the data center and possibly including the local power grid). In addition, the power supplied to machine rooms tends to be over-provisioned because it is specified in practice not by workload demands but rather by high-energy LINPACK runs or nameplate power estimates. This results in a considerable amount of *trapped power capacity*—excess power infrastructure. Instead of being wasted, this trapped power capacity should be reclaimed to accommodate more compute nodes in the machine room and thereby increase system throughput. Thus, the ability to prioritize and manage platform power allocations is becoming essential and active management of a platform's average and peak power draw through processor frequency management (another parameter that affects performance) has become a high priority for both HPC data centers and vendors and is currently a high priority research topic.

Further, a study entitled the Power-Aware Data Center Project, released December 2012, [6] found the following.

- LANL supercomputers contain significant trapped capacity, not just on average but even in the worst case;
- Variability in power draw can be quite different across different architectures;
- There is a qualitative difference between power drawn while running benchmarks and when running a production workload;
- Power capping has the potential to free large amounts of power and cooling infrastructure with minimal impact on applications;
- An ability to co-schedule high-power-

consuming jobs with low-power consuming jobs would offer the potential to reduce peak power draw to further support penalty-free power capping;
- And that "race to halt" appears to be a more effective way to run a scientific workload than are various power-saving strategies.

LANL has already begun to investigate power capping. A detailed description of workload, test methods, and early results was presented at the Cray Users' Group (CUG 2015) [7] and these investigations will be continuing. Concurrently, there is research underway to investigate methods to enforce a system-wide power cap [8]. Additionally, there is ongoing interest around energy efficiency issues and research in the broader community. For instance, a recent paper noted that the Intel Xeon E5-1600 v3 and E5-2600 v3 series processors–codenamed Haswell-EP–implement major changes compared to their predecessors. Among these changes are integrated voltage regulators that enable individual voltages and frequencies for every core. The authors analyzed a number of consequences of this development that are of utmost importance for energy efficiency optimization strategies such as dynamic voltage and frequency scaling (DVFS) and dynamic concurrency throttling (DCT). This includes the enhanced RAPL (running average power limiting) implementation and its improved accuracy as it moves from modeling to actual measurement [9].

All of this prior work points to the critical importance of building a robust monitoring system that integrates data from a variety of sources to promote system understanding, to improve system performance, and to diagnose problems, as well as the need to evaluate how scheduling might impact power reductions.

The Trinity Haswell installation (Trinity Phase I) in late FY15 and then the KNL installation (Trinity Phase II) in late FY16 will provide for the first time per job power usage and some power management capabilities within the system. We will begin to look at how to use this information for management of power usage during FY16. Power interfaces have been defined. Trinity Haswell is installed and being accepted 12/2015. Integration of the Trinity Haswell partition into the power management systems is nearly complete. Power management for jobs and job scheduling will be tested in the spring of 2016 and utilization of power management of the Trinity Haswell partition will begin in the summer of 2016. Full utilization of Trinity power management for power savings/control will occur in 2017 after Trinity's KNL partition is installed and in production. A next step is dynamic facility management informed by platform data, including job scheduling and job profile. Separately, but concurrently, we will need to look at how constraints and cost efficiencies can be leveraged to drive scheduling (e.g., identifying a set of jobs/workloads that may be advantageous to run overnight when energy costs are reduced).

ACKNOWLEDGEMENTS

The authors would like to thank Monica Witt for access to LANL water usage data and Josip Loncaric for many fruitful technical discussions and shared insight.

REFERENCES

[1] Kogge PM, et al. (2008) "ExaScale computing study: Technology challenges in achieving exascale systems," Technical ReportTR-2008-13, University of Notre Dame CSE Department, Notre Dame, IN.

[2] "The Tip of the Iceberg: Say, what is "under the floor" of a supercomputer?" National Security Science, April 2013. Available online: http://www.lanl.gov/discover/publications/national-security-science/2013-april/_assets/docs/under-supercomputer.pdf

[3] See graphic available online: http://www.lanl.gov/environment/sustainability/water-reuse.php

[4] "The Top500 List" [online] Available: http://www.top500.org/lists/

[5] "HPL." [Online] Available: http://www.netlib.org/benchmark/hpl/

[6] Pakin, S., et al (2012), Power-Aware Data Center Project: Final Report. LA-UR-13-20710.

[7] "Enabling Advanced Operational Analysis Through Multi-Subsystem Data Integration on Trinity," J. Brands, D. DeBonis, A. Gentile, J. Lujan, C. Martin, D. Martinez, S. Olivier, K. Pedretti, N. Taerat, and R. Velarde, Cray Users' Group (CUG), 2015.

[8] Z. Zhang, M. Lang, S. Pakin, S. Fu, "TracSim: Simulating and Scheduling Trapped Power Capacity to Maximize Machine Room Throughput," Parallel Computing, available online 2 December 2015.

[9] D. Hackenberg, R. Schone, Thomas Ilsche, D. Molka, J. Schuchart, R. Geyer, "An Energy Efficiency Feature Survey of the Intel Haswell Processor," Proceedings of the 2015 IEEE International Parallel and Distributed Processing Symposium Workshop (IPDPSW), pp. 896-904.