LLNL Sequoia-25 ISC13 Green500 Submission

Robin Goldstone, Anna Maria Bailey

Livermore Computing HPC Facility, LLNL

Lawrence Livermore National Laboratory

LLNL-PRES-646353

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under contract DE-AC52-07NA27344. Lawrence Livermore National Security, LLC

Sequoia Overview

- IBM BlueGene/Q System
  - 20 petaFLOP/s peak - 16.32 PF achieved
  - Memory 1.5 PB, 4 PB/s bandwidth
  - 98K nodes, 1.5M cores, 96 racks
  - 3 PB/s interconnect bandwidth
  - 0.5−1.0 TB/s Lustre bandwidth
  - 50 PB disk

- Power – 9.6 MW in 4,000 ft^2
- #1 on June 2012 Top500 ***and*** Green500

2,100.88 MFLOPS/W L1 submission (compute only, no network)

Sequoia Power Distribution

Utility Transformers

Qty 6 @ 2000 KVA Transformers with Utility Metering.

Distribution Switchboards

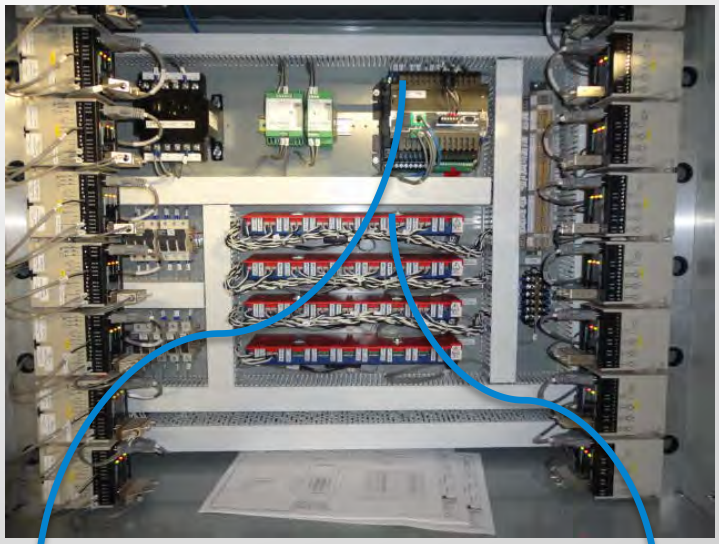
Qty 6 Distribution Switchboards, each feeding 16 underfloor PDUs.

Underfloor PDUs

Qty 96 underfloor 480V PDUs, each feeding one Sequoia rack.

Sequoia Metering Components

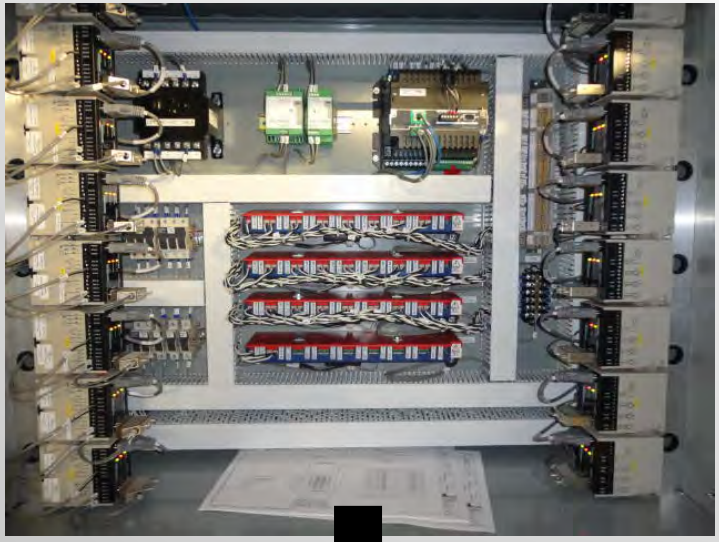
Distribution Switchboard (x6)



Main Breaker Meter:
Siemens 9510

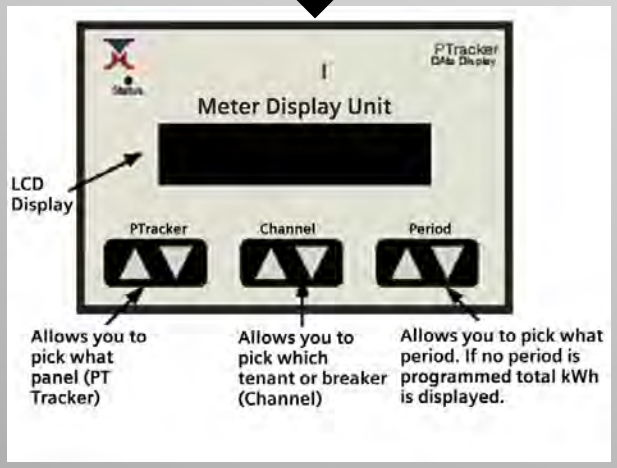
16 Feeder Breaker Meters:
Siemens MP636-EXTC

Sequoia Meter Data Collection



OSIsoft.

Data from Sequoia power meters, along with other LLNL HPC systems and facilities data, is collected and analyzed by the OSIsoft PI infrastructure management system.



The Siemens remote meter display unit (MDU) provides a local display of the energy readings. One MDU can support up to 256 meters. The MDU provides a quick way to view the energy information at an accessible location.

Sequoia-25

- Sequoia (20 PF peak) + Vulcan (5 PF peak)
- Systems were combined in early/mid 2013 to enable extreme-scale science calculations prior to transition of Sequoia to classified operations.
- Sequoia-25 achieved a Linpack result of 21.46 PF, however submission was rejected for inclusion in June 2013 Top500 list due to the fact that the system was subsequently de-coupled.

May 06, 2013

Sequoia Hits Warp Speed

Tiffany Trader

Chalk up another win for Sequoia and high-performance computing. The record-setting supercomputer is helping pave the way for future planetary-scale simulations.

Researchers from Lawrence Livermore National Laboratory (LLNL) and Renesselaer Polytechnic Institute (RPI) created a protocol called Time Warp that carried out 7.8 million MPI tasks on 1,966,080 cores of the Sequoia Blue Gene/Q supercomputer system.

Time Warp automatically exposes available parallelism in a model via its error detection and rollback recovery mechanism. The effect on performance is nothing short of remarkable: scaling from 32,768 cores to almost 2 million cores resulted in a 97x performance improvement. This long-sought-after linear performance gain is attributed to a sophisticated caching mechanism.

The team also used Sequoia to process 33 trillion events in 65 seconds, which comes out to over 504 billion events/second – that's 41 times better than the previous record of 12.2 billion events per second, which was set in 2009.

Top500
Green500

Despite not making the June 2013 Top500 list, Sequoia-25 still qualifies for Green500.

"In order to qualify for The Green500 List, a supercomputer must achieve performance at least as high as the 500th ranked system in the current Top500 List."

sequoia-25-hpl-out.txt

```
stdout[0]: ================================================================================
stdout[0]: T/V                N    NB     P     Q               Time                 Gflops
stdout[0]: --------------------------------------------------------------------------------
stdout[0]: WR16L2L8     14942207   256   512   960 103607.50158385312533937     2.14665301560507566e+07
stdout[0]: --VVV--VVV--VVV--VVV--VVV--VVV--VVV--VVV--VVV--VVV--VVV--VVV--VVV--VVV--VVV
stdout[0]: Max aggregated wall time rfact . . . :  7.70762757369811879
stdout[0]: + Max aggregated wall time pfact . . :  5.59436735105555272
stdout[0]: + Max aggregated wall time mxswp . . :  3.76747948728734627
stdout[0]: + Max aggregated wall time laswp . . : 3289.69090172152209561
stdout[0]: Max aggregated wall time up tr sv  . : 24.85606373313930817
stdout[0]: Max aggregated wall time dgemm . . . : 96135.22387718831305392
stdout[0]: Max aggregated wall time dtrsm . . . : 3681.57283625347190537
stdout[0]: --------------------------------------------------------------------------------
```

Would have been #2 on June 2013 Top500 List

Sequoia-25 Green500 Submission: Why L2, not L3?

B	C	D	
Time stamp	Sequoia-At-120Racks-KW-Total	Sequoia-At-120Racks-kWh-Total	
26-Mar-13 21:34:29	7694.647461	58198392	
26-Mar-13 21:34:30	7693.155273	58198396	
26-Mar-13 21:34:31	7691.663086	58198396	
26-Mar-13 21:34:32	7690.170898	58198400	
26-Mar-13 21:34:33	7688.678711	58198400	
26-Mar-13 21:34:34	7687.186523	58198404	
26-Mar-13 21:34:35	7685.694336	58198404	
26-Mar-13 21:34:36	7684.202148	58198408	
Total kW	1188413328	300988	Total kWh
Avg kW	11470.40	10457.02	Avg kW

△ 9.2%

Siemens meters report both instantaneous power (kW) and cumulative energy (kWh). When computing average power from the two data sets, we noted a discrepancy and subsequently discovered that 11 of our 120 meters were not correctly reporting their results to our data collection system.

All L3 Requirements were otherwise met

Aspect	L3 Requirement	Sequoia-25
Aspect 1a: granularity of power measurements	Continuously integrated total energy	Met*
Aspect 1b: timespan of power measurements	A time series of equally spaced integrated total energy values	Met*
Aspect 1c: reported measurements	Core phase avg power, >10 energy measurements within core phase, whole app avg power, idle power	Met
Aspect 2: machine fraction	Whole machine	Met
Aspect 3: subsystems included	All participating subsystems	Met
Aspect 4: power measurement point	Upstream of power conversion	Met

* Some meters failed to report due to misconfiguration.

Results

- Sequoia-25 Linpack: 21.41 PF
- L2 Average Power
  - Core phase: 11,501.98 MW
  - Full run: 11,470.29 MW
- Green500 Result: 1861.41 MFLOPS/W
  - ~12% less efficient than L1 measurement (with network excluded)

Takeaways

- In much the same way that preparing a Top500 submission can help shake out your HPC system hardware, preparing a Green500 submission can help shake out your facility metering infrastructure.
- … But it would have been better to have discovered misconfigured meters beforehand.
- LLNL's multiple levels of metering (per transformer, per main breaker panel and per PDU) have proven to be very useful, allowing for cross-correlation of measurements as well as isolation of faulty meters and anomalous power conditions.