Capacity Systems:
Their Synergy with,and Relationship to, Capability Systems

by
Jim Ang
Sandia National Laboratories

Presented at: SOS10
March 8, 2006
Wailea, Maui, HI

National Nuclear Security Administration

Sandia National Laboratories

Panel Questions

- What are the principal factors that distinguish Capacity Systems from other forms of computing systems?
- What is the role and impact of capacity computing for current and future scientific problems?
- What technical challenges confront the continued growth of capacity computing performance?
- What will be the dominant directions for future generation capacity computing and system types?
- Will the current division between capacity and capability computing be retained over the next decade or will there emerge a different useful distinction in form and function?

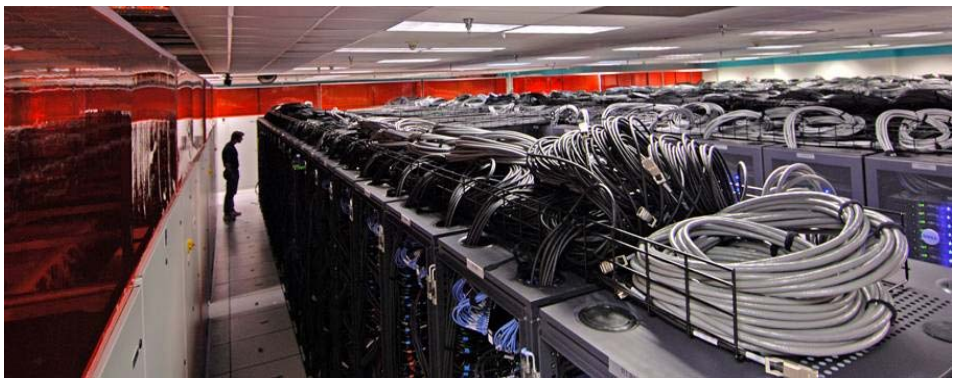
Capability and Capacity Computing

- The largest supercomputers are used for capability or turnaround computing where the maximum processing power is applied to a single problem. The goal is to solve a larger problem, or to solve a single problem in a shorter period of time. Capability computing also enables the solution of problems that cannot otherwise be solved in a reasonable period of time (for example, by moving from a two-dimensional to a three-dimensional simulation, using finer grids, or using more realistic models). **The main figure of merit is time to solution.**
- Smaller or cheaper systems are used for capacity computing, where smaller problems are solved. Capacity computing can be used to enable parametric studies or to explore design alternatives; it is often needed to prepare for more expensive runs on capability systems. Capacity systems will often run several jobs simultaneously. **The main figure of merit is sustained performance per unit cost.**
- There is often a trade-off between the two figures of merit, as further reduction in time to solution is achieved at the expense of increased cost per solution different platforms exhibit different trade-offs. Capability systems are designed to offer the best possible capability, even at the expense of increased cost per sustained performance, while capacity systems are designed to offer a less aggressive reduction in time to solution but at a lower cost per sustained performance.
  - *Susan L. Graham, Marc Snir, and Cynthia A. Patterson, editors, Getting Up to Speed: The Future of Supercomputing, National Research Council, Committee on the Future of Supercomputing, National Academies Press, page 24, 2005.*

Thunderbird vs. Red Storm

Thunderbird System Parameters

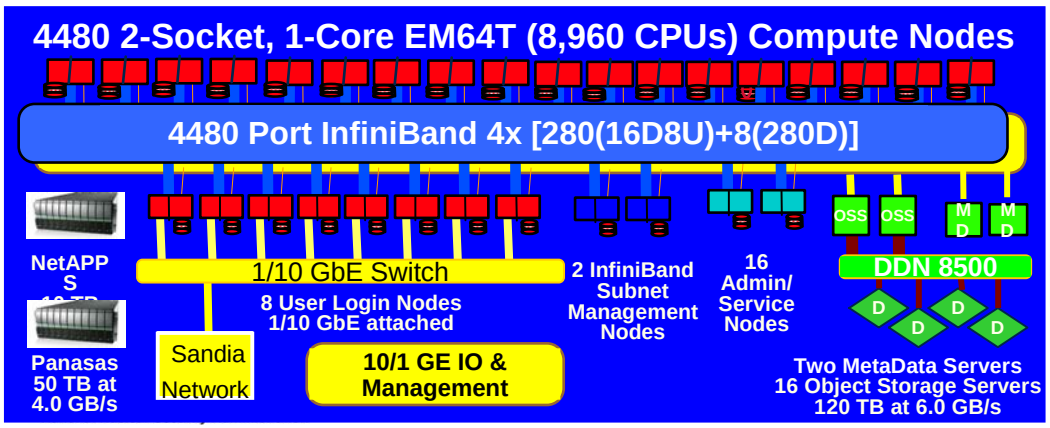
- 140 compute node cabinets and 4,480 compute nodes
- 14.4 GF/s dual socket 3.6 GHz single core Intel SMP nodes with 6GB DDR-2 400 SDRAM
- 10 IB Switch cabinets, ~300 InfiniBand switches to manage
- ~9,000 InfiniBand ports, 2:1 oversubscribed fat-tree topology
- ~33,600 meters (or 21 miles) of 4X InfiniBand copper cables
- ~10,000 meters (or 6 miles) of copper Ethernet cables
- 26,880 1 GB DDR-2 400 SDRAM modules
- 1.8 MW of power, 400 tons of cooling

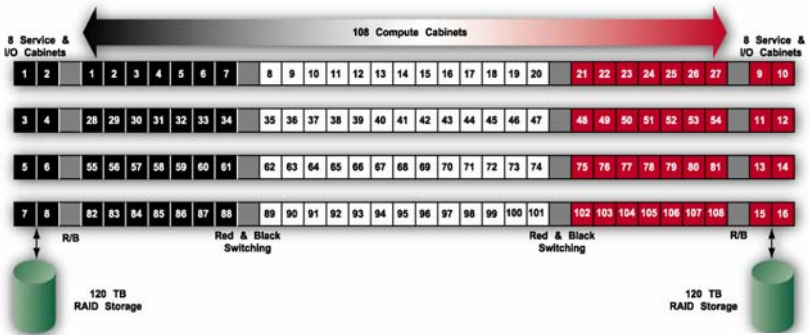


Red Storm System Parameters

- 108 compute node cabinets and 10,368 compute node processors (AMD *Opteron* @ 2.0 GHz)
- ~30 TB of DDR compute node memory
- Fully connected 3-D mesh interconnect.
- Link MPI Latency ~4μsec
- Peak Bi-directional MPI BW per link 5.2 GB/s
- Min Bisection BW 2.0 TB/s
- 8 Service and I/O cabinets on each end (256 processors for each color)
- ~400 TB of disk storage (~200 TB per color)
- Less than 2 MW total power and cooling
- Less than 3,000 ft^2 of floor space

QuickTime™ and a TIFF (LZW) decompressor are needed to see this picture.





Sandia National Laboratories

Capacity vs. Capability System Architecture Differences

- Capacity Systems are focused on Commodity Technologies, Capability Systems leverage Commodity Components make selective investments in non-commodity technologies to impact:
  - Performance at Scale
  - Reliability at Scale
- Interconnect Network Performance
  - Capacity: B/F < 0.1
  - Capability: B/F > 1
- Compute Node Operating System
  - Capacity: Linux
  - Capability: LWK
- Scalability of Runtime System Software
  - Scale of workload: Number of jobs vs. Size of jobs
- Integrated, Independent RAS Subsystem
  - Capacity: Area for future development?
  - Capability: Yes, required for scalability

ASC Capability Systems Governance Model

- Access to ASC capability systems will be similar to that of experimental facilities
  - Major programmatic computing efforts organized as computing work packages that are reviewed and prioritized for programmatic importance
  - Each proposed work package, called a *Capability Computing Campaign* (CCC), includes at least one major calculation needing a significant proportion of an ASC capability system and could also include related supporting jobs of smaller sizes
- The portfolio of CCCs will balance objectives:
  - To ensure resources address the highest programmatic needs
  - To efficiently use ASC capability systems for large capability mode jobs that cannot be run on other systems

ASC Program Performance Metrics
Capability Performance Indicator (CPI)

- One component in set of 4-5 ASC Program Performance Indicators under preparation for NNSA & OMB
- The metric of the capability mode usage on this class of systems is an indicator of the overall ASC program health in many respects
  - Capability systems reliably run the scale of workload for which they were purchased
  - ASC codes and their underlying algorithms perform at this scale and are being used to solve this class of problem
  - ASC has ensured that sufficient *capacity* resources are available to free the *capability* systems to run the jobs they were intended to run

$$\text{CPI} = \frac{\text{node-hours of capability usage}}{\text{total node-hours of usage}}$$

CPI Components

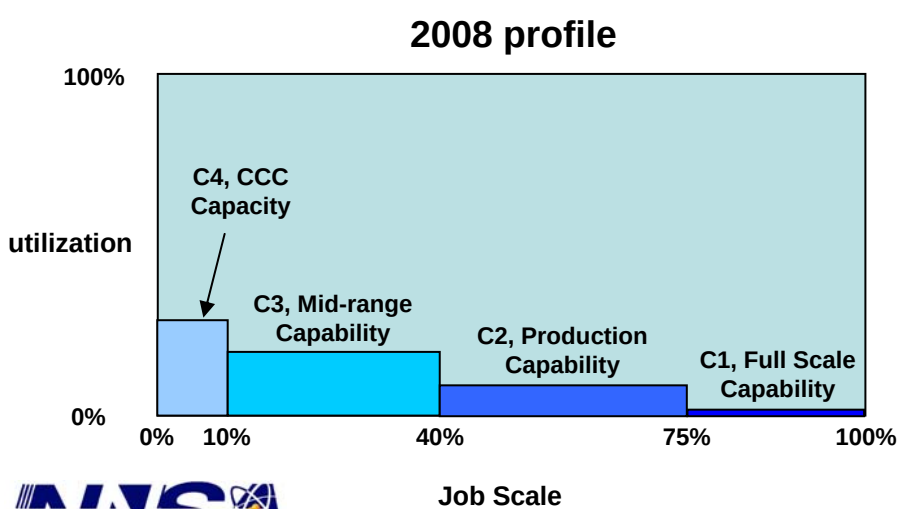
- **Four Regimes of Capability Mode usage:**
  - ***Category 1* (C1) jobs use 75% or more of the available nodes**
    - This class of job uses the full capability of the machine, and is typically used for scaling studies, such that the system will be effectively dedicated during the duration of the run.
  - ***Category 2* (C2) jobs use 40-74.9% of the available nodes**
    - This class of job will typically consist of large production weapons calculations and performance studies.
  - ***Category 3* (C3) jobs use 10-39.9% of the available nodes**
    - These are jobs similar to C2, but in a smaller size range, so that many such jobs could use the machine simultaneously. This class of job should include a tie to production weapons calculations and performance studies.
  - ***Category 4* (C4) jobs use less than 10% of the available nodes**
    - These smaller jobs are essential for carrying out a CCC work package. This class of job must have direct impact on programmatic drivers for production weapons calculations and performance studies.

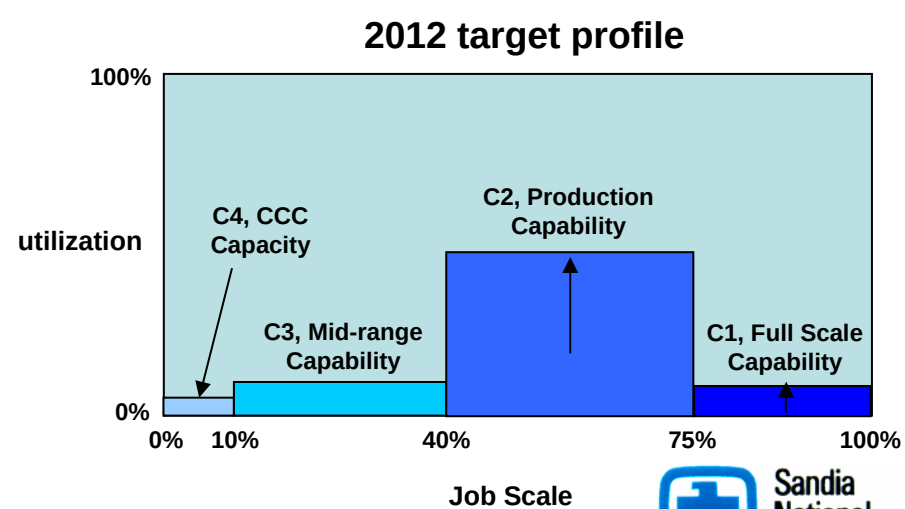
Strawman CPI Evolution

- Integrated Capability Performance Indicator

$$CPI_{Integrated} = CPI_{C1} + CPI_{C2} + CPI_{C3} + CPI_{C4}$$

CPI	FY08	FY09	FY10	FY11	FY12 Target
CPI_{C1}	2.0%	4.0%	6.0%	8.0%	10.0%
CPI_{C2}	10.0%	20.0%	30.0%	40.0%	50.0%
CPI_{C3}	20.0%	17.5%	15.0%	12.5%	10.0%
CPI_{C4}	25.0%	20.0%	15.0%	10.0%	5.0%
$CPI_{Integrated}$	57.0%	61.5%	66.0%	70.5%	75.0%





National Nuclear Security Administration

Sandia National Laboratories

Questions for the Panel
Capacity Machines

- What are the principal factors that distinguish Capacity Systems from other forms of computing systems (Capability Systems)?
  - Scale of workload: Number of jobs vs. Size of jobs
- What is the role and impact of capacity computing for current and future scientific problems?
  - Cost effective computing (for smaller job sizes than Capability Computing)
  - Perform simulations that prepare for & scale up to Capability Computing problems
- What technical challenges confront the continued growth of capacity computing performance?
  - Continue to increase scalability in job sizes - TBird is currently limited to 1,024 processor jobs
  - Expect to see continued technology transfer from Capability Systems to Capacity Systems in OS, RT, Communication library scalability
- What will be the dominant directions for future generation capacity computing and system types?
  - Multi-core processors, Integrated RAS capabilities, Virtualization
- Will the current division between capacity and capability computing be retained over the next decade or will there emerge a different useful distinction in form and function?
  - Current Division will continue due to differences in investments and roles