Capability Machines: Red Storm

Jim Tomkins

SOS10, Maui, Hawaii, March 6 - 9, 2006

Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company,
for the United States Department of Energy's National Nuclear Security Administration
under contract DE-AC04-94AL85000.

Red Storm System

- True MPP, designed to be a single system
- Distributed memory MIMD parallel supercomputer
- Fully connected 3-D mesh interconnect. Each node has its own high bandwidth, bi-directional connection to the primary communication network.
- 108 compute node cabinets and 10,368 compute node processors (AMD Opteron @ 2.0 GHz)
- ~30 TB of compute node memory
- Red/Black Switching - ~1/4, ~1/2, ~1/4
- 8 Service and I/O cabinets on each end (256 nodes for each color)
- 400+ TB of disk storage (200+ TB per color)

Red Storm System

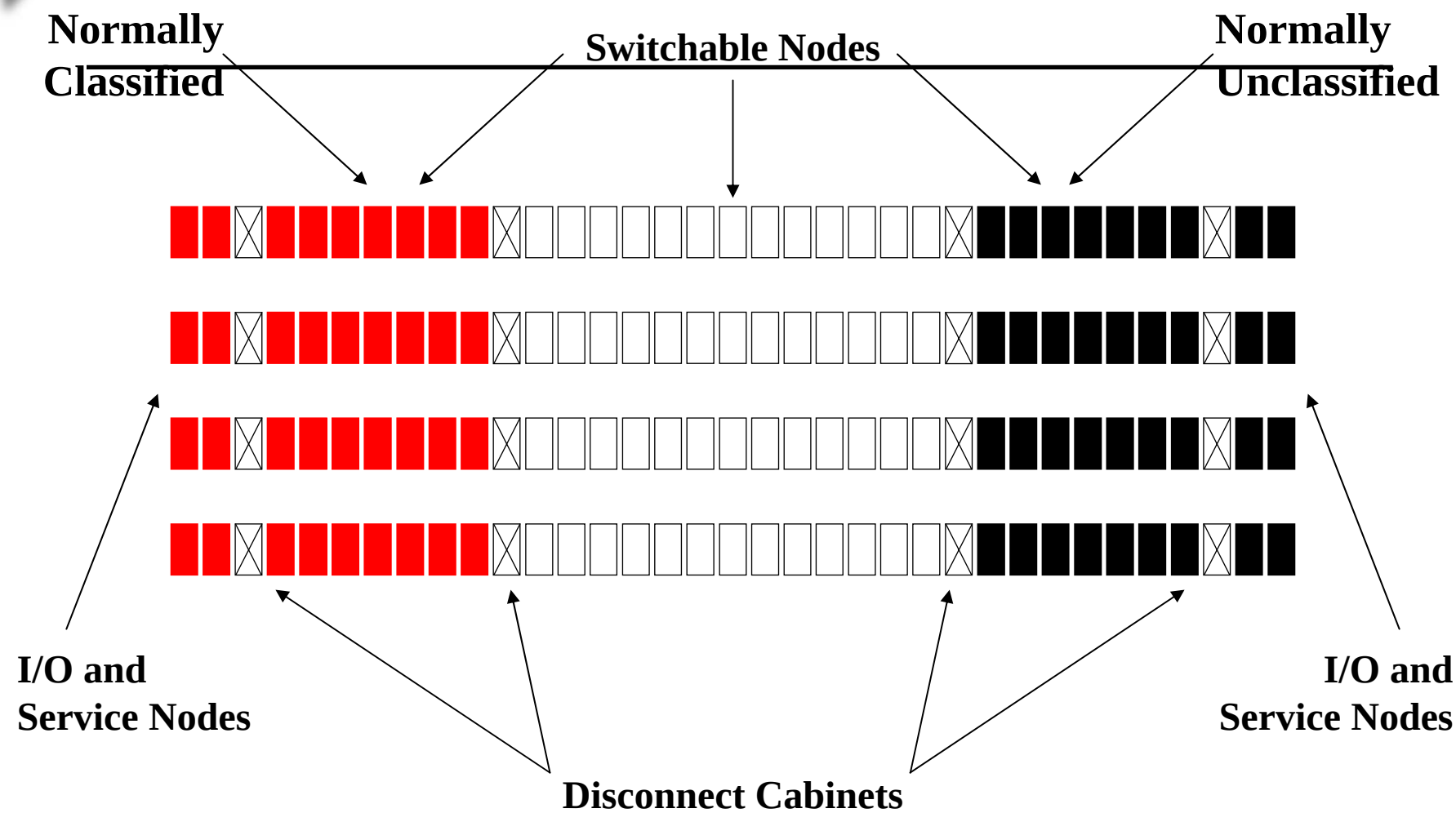
- Functional Hardware Partitioning - service and I/O nodes, compute nodes, RAS nodes
- Partitioned Operating System (OS) - LINUX on service and I/O nodes, LWK (Catamount) on compute nodes, stripped down LINUX on RAS nodes.
- Separate RAS and system management network (Ethernet).
- Less than 2 MW total power and cooling.
- Less than 3000 sq ft.

Red Storm Topology

- Compute Nodes
  - 27 x 16 x 24 (X, Y, Z) - Red/Black split 2688 - 4992 - 2688
- Service and I/O Nodes
  - 2 x 8 x 16 (X, Y, Z) nodes on each end
  - Mesh is 2 x 16 x 16 (X, Y, Z) on each end
  - Center card cage is empty

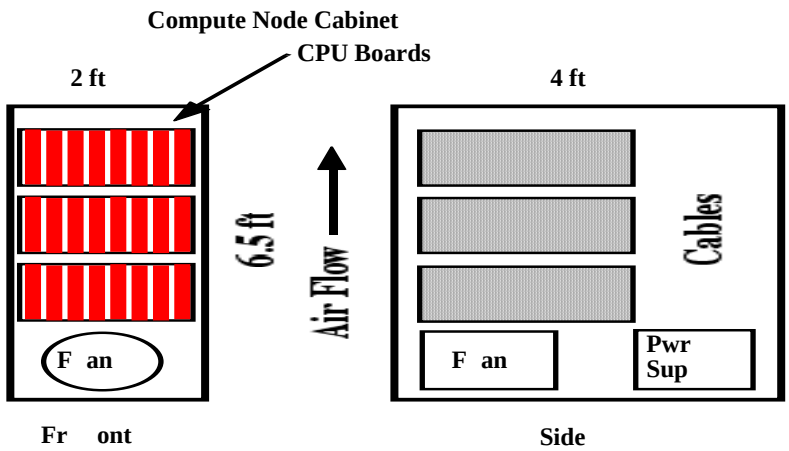
Red Storm Layout

(27 × 16 × 24 mesh)



Disk storage system not shown

Red Storm Cabinet Layout



Compute Node Cabinet

- 3 Card Cages per Cabinet
- 8 Boards per Card Cage
- 4 Processors per Board
- 4 NIC/Router Chips per Board
- N+2 Power Supplies
- Passive Backplane

Service and I/O Node Cabinet

- 2 Card Cages per Cabinet
- 8 Boards per Card Cage
- 2 Processors per Board
- 4 NIC/Router Chips per Board
- 2 PCI-X slots per Node
- N+2 Power Supplies
- Passive Backplane

Red Storm Performance

- HPL - 36.19 TF (6)
- HPCC baseline
  - G-PTRANS - 1813 GB/s (1)
  - G-Random Access (Gups) - 1.018/s (1)
- Interconnect performance
  - Latency <5 (3.6) μs (neighbor) <8 μs (full machine)
  - Peak Link bandwidth - 3.84 GB/s each direction
  - Peak HT bandwidth - 3.2 GB/s each direction
  - Bi-section bandwidth ~2.95 TB/s Y-Z, ~4.98 TB/s X-Z, ~6.64 TB/s X-Y

- I/O system performance
  - Sustained file system bandwidth of 50 GB/s for each color.
  - Sustained external network bandwidth of 25 GB/s for each color.

Red Storm Upgrade Plans

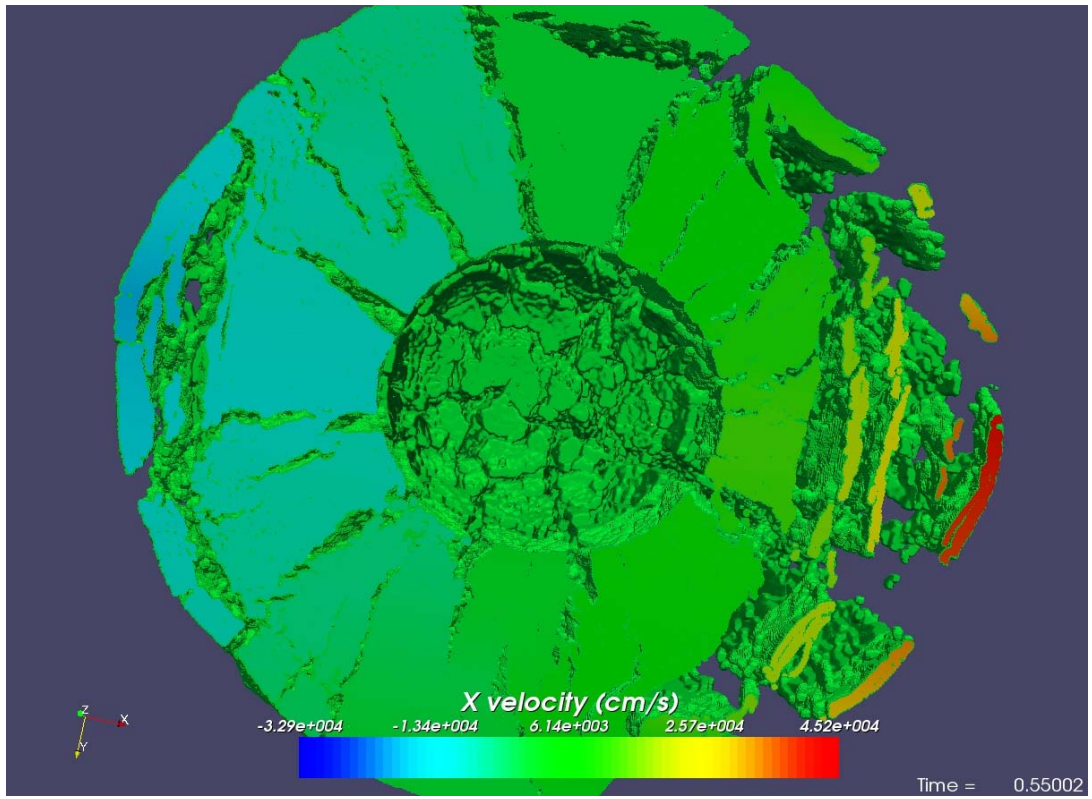
- Current Contract
  - 5th row of cabinets to be installed in July (mostly built) - system will have 12,960 compute nodes and 640 (320 on each end) service and I/O nodes
  - Replace existing Seastar v1.2 with v2.1 to improve HT bandwidth (approximately double delivered bandwidth) - Installation planned for July
- Additional Planned Upgrades
  - Replace all single processor 2.0 GHz Opterons with dual core 2.4 GHz chips
  - Increase memory on each compute node to 6 GB of DDR 400
  - System peak will increase to ~125 TF with ~75 TB of memory
  - Expect to complete upgrade in early GFY'07

Some Applications Running on Red Storm

- CTH (SNL) - Shock/Hydro
- Partisn (LANL) - Discrete Ordinates Radiation Transport
- ITS (SNL) - Monte-Carlo Radiation Transport
- Sage (LANL) - Hydro
- Salinas (SNL) - Structural Mechanics
- Alegra (SNL) - Hydro/ICF
- Presto (SNL) - Structural Dynamics
- Calore (SNL) - Heat Transfer
- Fuego (SNL) - Fire
- SEAM (NCAR) - Atmospheric Climate
- POP (LANL) - Ocean Model

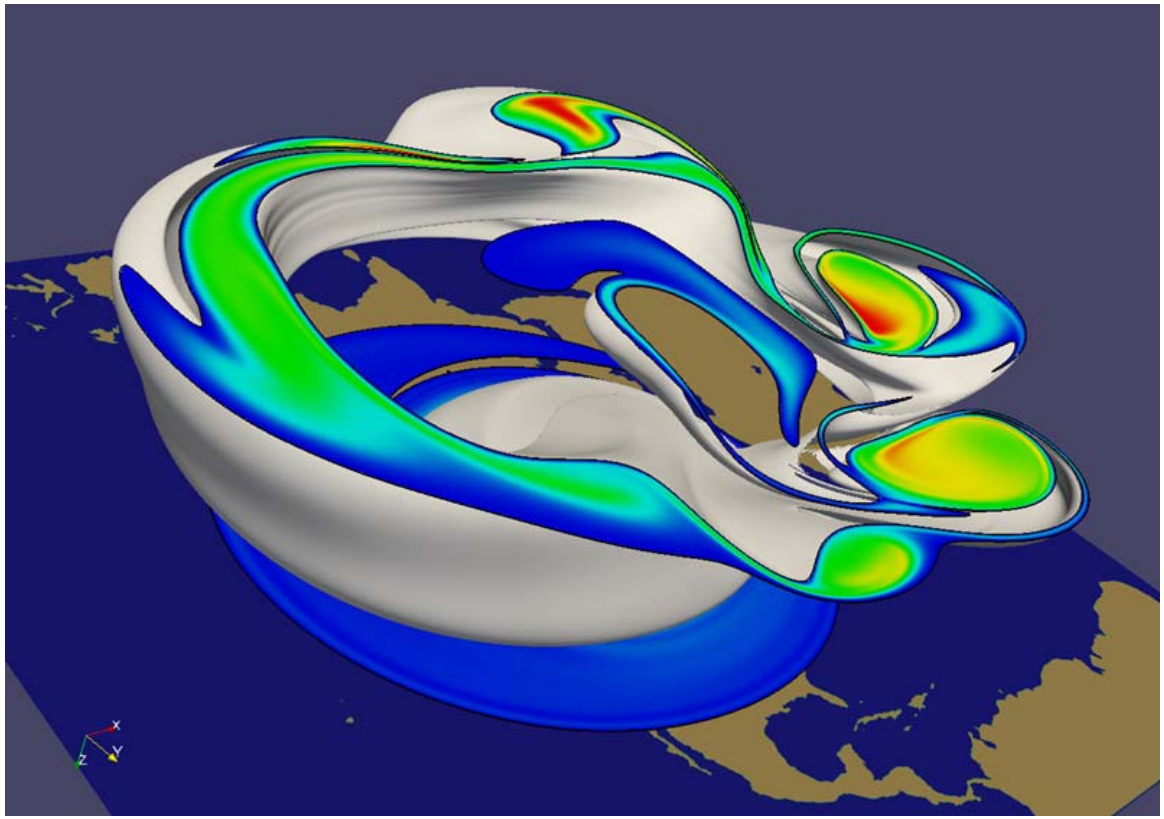
Calculation of Golevka Asteroid Explosion

(Mark Boslough)



- 10 MT Explosion at center of mass
- 0.5 s simulation time
- Billion cell simulation, ~0.5m cell size
- 7200 nodes of Red Storm
- >12 hr run

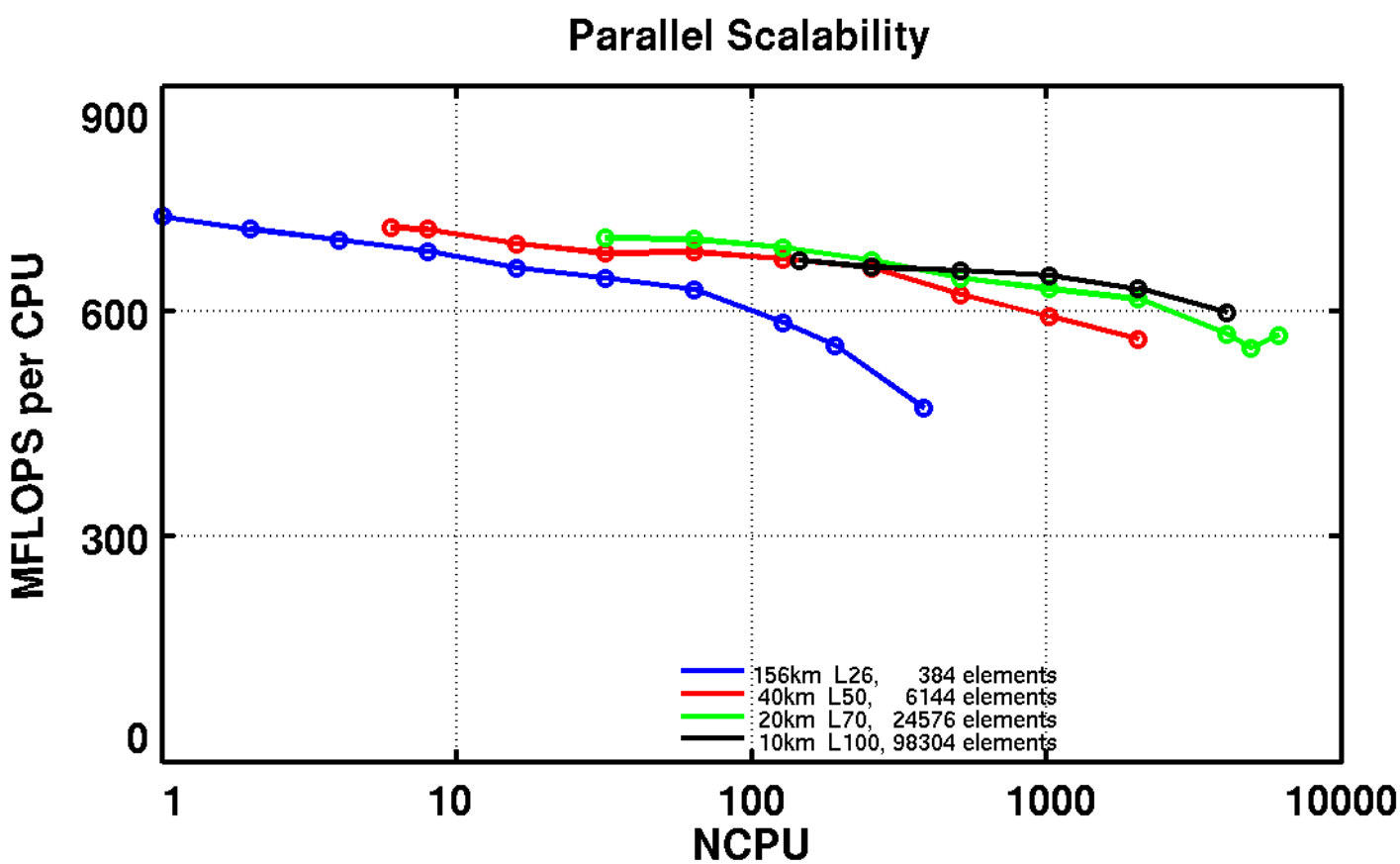
SEAM
(Mark Taylor)



- Billion grid point simulation of a polar vortex that has trapped air at the pole.
- Day 16 in a 20 day simulation.
- Run on 7200 nodes of Red Storm for 36 hrs.

SEAM Scalability on Red Storm

(Mark Taylor)



Capability Computing

- Capability computing is leading edge in terms of scale. Currently this means efficiently using thousands of processors for a single calculation.
  - It is more than increased resolution or reduced run-time
  - Over the next decade capability computing will probably mean using tens of thousands of processors.

Capability Versus Capacity

- **The distinction is important.**
  - **Capability systems are designed to perform capability calculations**
  - **Capability systems can also perform capacity calculations**
  - **Capacity systems don’t do well on capability calculations**

Proposed ASC Capability Metrics

- **Four Usage Categories:**
  - ***Category 1* (C1) Capability 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, or the largest jobs to be run within the complex, such that the system will be effectively dedicated during the duration of the run.
  - ***Category 2* (C2) Capability jobs use 40-74.9% of the available nodes**
    - This class of job will typically consist of large production calculations and performance studies, which are nevertheless sized such that two jobs could run on the system simultaneously.
  - ***Category 3* (C3) Capability 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 mid-range production calculations and performance studies.
  - ***Category 4* (C4) Capability jobs use less than 10% of the available nodes**
    - These are Capacity jobs that use less than 10% of the available nodes, but these smaller jobs are an essential step to running related capability calculations.

Proposed ASC Capability Metrics

- **Capability systems reliably run the scale of workload for which they were purchased**

- **Capability Performance Indicator (CPI)**

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

  - **C4 is really capacity computing**
  - **C3 represents the transition from capacity to capability computing**