Why Would You Need 63.000 Cores, 1.7PBytes storage and 57GB/s data transfer over Lustre

Philippe Trautmann
BDM High Performance Computing
philippe.trautmann@sun.com

Ranger

Texas Advanced Computing Center (TACC)

AGENDA:

- Overview
- History
- Hardware configuration
- Software stack
- Lustre and data management

Overview

Solid #2

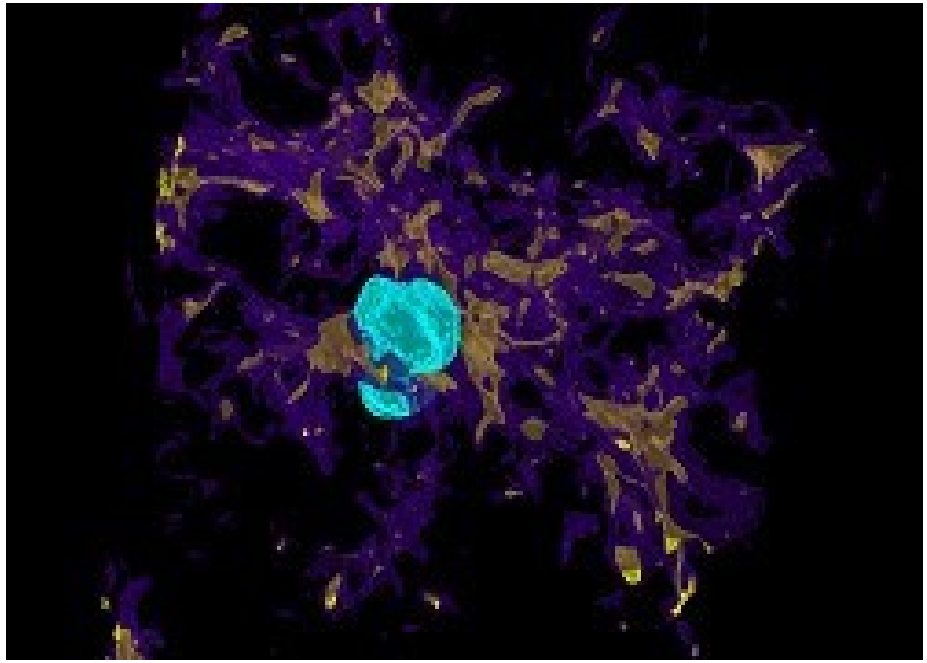
- TACC = University of Texas at Austin with collaboration from AZU, and Cornell.
- Founded by the National Science Foundation, NSF
- First track 2 award, for $59M of which $30M was for the computer system.
- Ranger entered production on 2/4/08
- Peak performance of 504 teraflops, measured is 326 Tflops.
- More Memory than any other system on Top 500
- 5 x more capable than any open-science computer available to the national science community

What's it for?

Computational power for the nation's research scientists and engineers.

Attaining the Unattainable

- Climate Research
- Astrophysics
- Complex Biology
- Biochemistry,
- Quantum physics
- Cosmology
- Geology



History

- 2001 – NSF launched the TeraGrid project to create a national network of Supercomputers.
  - To provide Computational power for the nation's research scientists
- 2003 – NSF expanded & included a system from TACC
- Feb. 2006 – NSF releases RFP for the 1st Track 2 supercomputer.
- Sept. '06 - NSF awards first Track 2 proposal to Sun and TACC
- Feb. '08 – Ranger goes into production

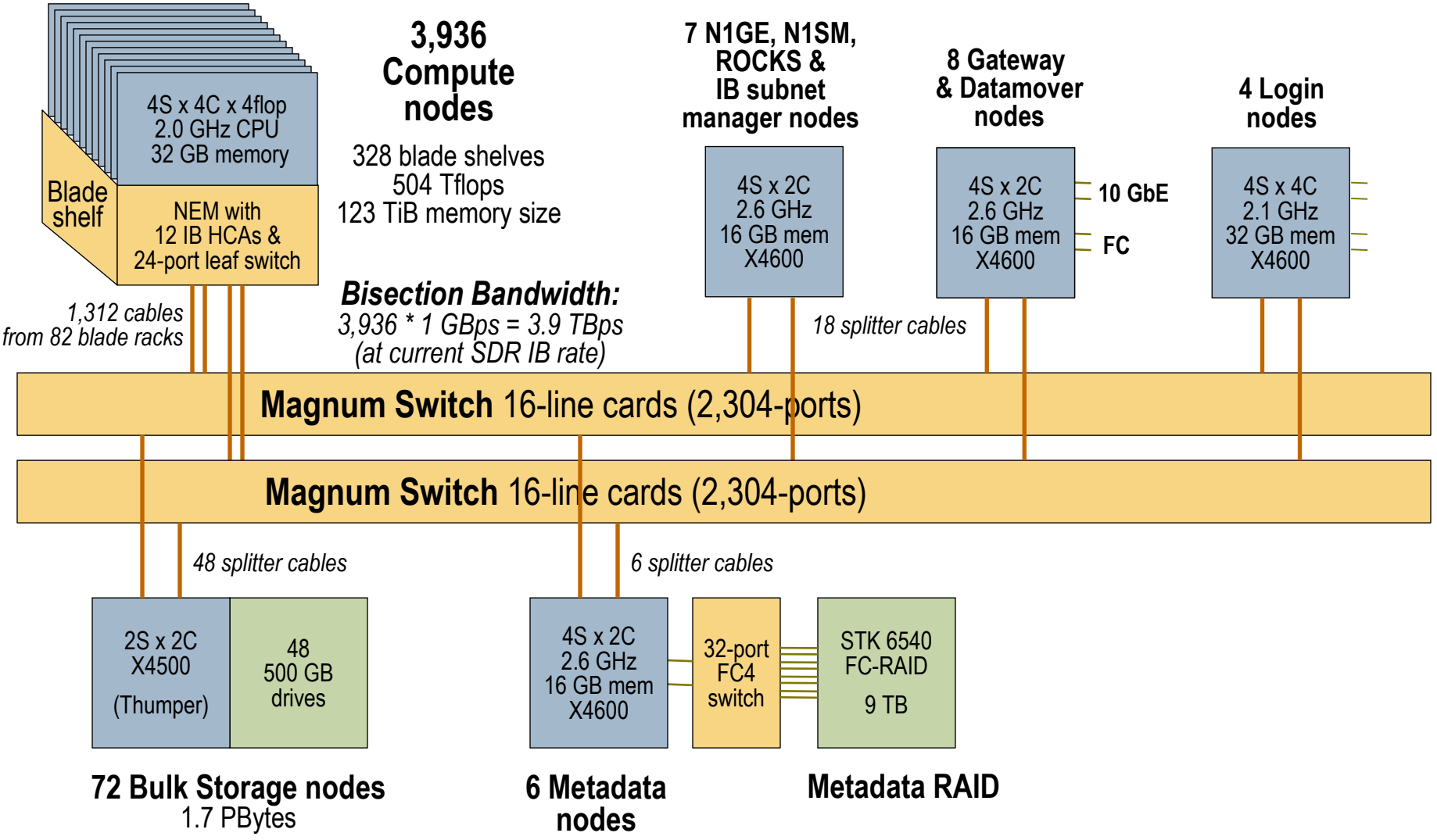
"*The vision is to make the grid a more seamless fabric, so a researcher has a virtual laboratory in which he has access to data, storage, simulations.*"

Configuration Summary

- **Two-tier InfiniBand topology** – SDR now, DDR later
  - A 24-port IB-NEM leaf switch on each 12-blade shelf
  - Two Magnum central switches: 16 line cards each
  - One 12x IB cable for every three nodes, 1,384 IB cables total
- **82 C48 Compute racks**, (Sun Blade 6048)
  - 3,936 Pegasus Blades w/ 2.0 GHz AMD BA, 4-socket, quad-core, 32 GB (15,744 chips, 62976 Cores and Dimms)
  - Each C48 rack has four 12-blade shelves
- **12 infrastructure node racks:**
  - 25 X4600 quad-socket nodes (Login, Data Mover, Metadata, etc)
  - 72 X4500 bulk storage nodes with 24-TB each
- **1 Metadata storage rack**: STK 6450 RAID, 9.3 TB

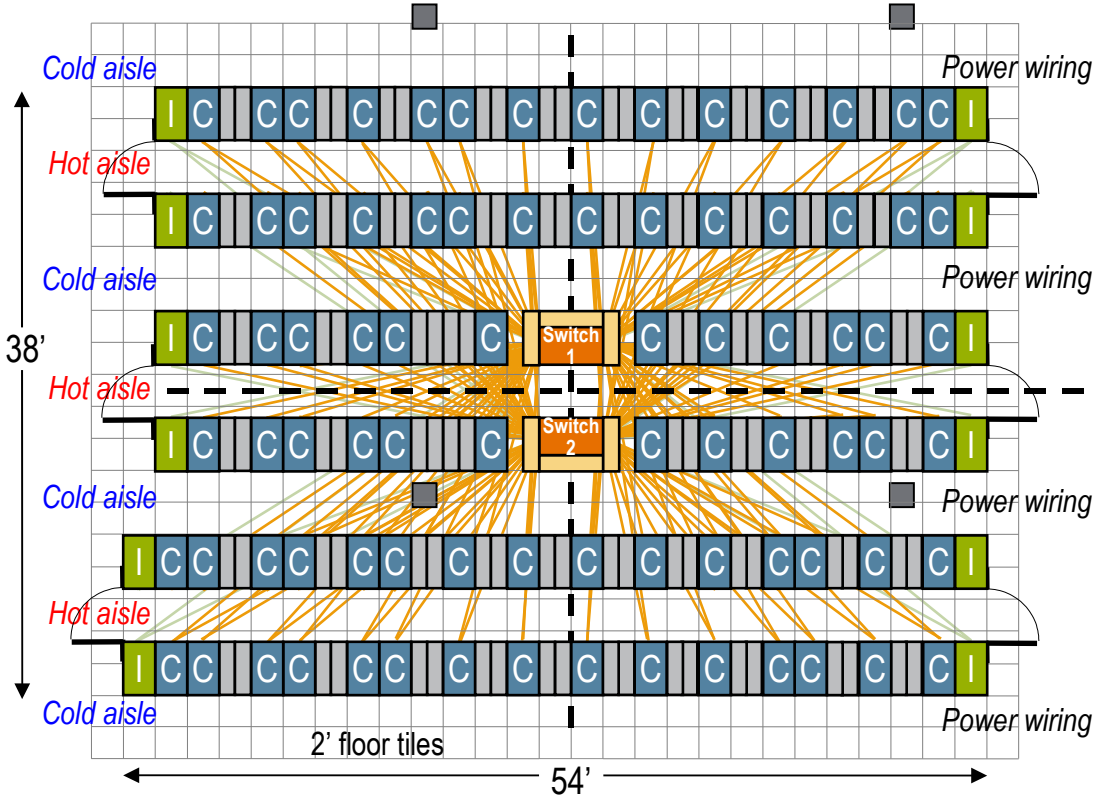

Sun

TACC Configuration



TACC Floorplan

Size: approximately half a basketball court



Switch — 2 Magnum switches
16 line cards each
(2,304 4x IB ports each)

C — 82 blade compute racks
(3,936 4S blades)

I — 12 IO racks
(25 X4600 4 RU
72 X4500 4 RU)

112 APC Row coolers

1,312 12x cables
(16 per rack)
16 km total length

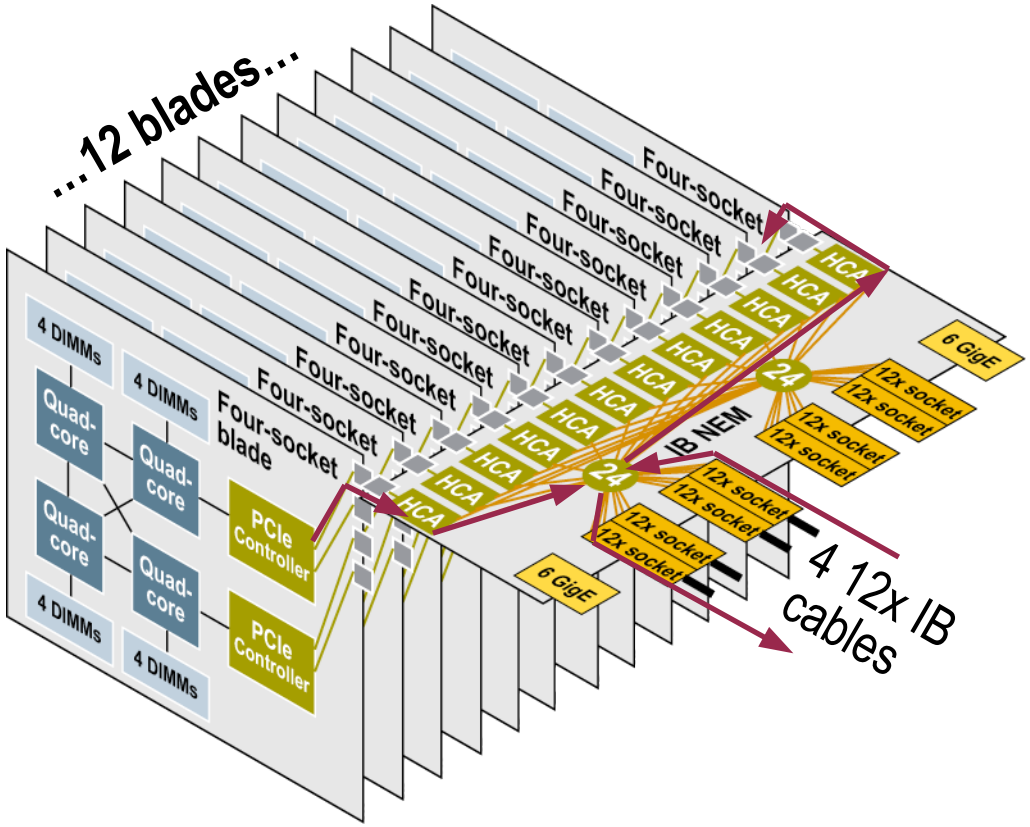
72 splitter cables
6 per IO rack

12x cable lengths: 171 9m, 679 11m, 406 13m, 56 15m

Splitter cable lengths: 54 14m, 18 16m

Blades and Shelf as Used at TACC

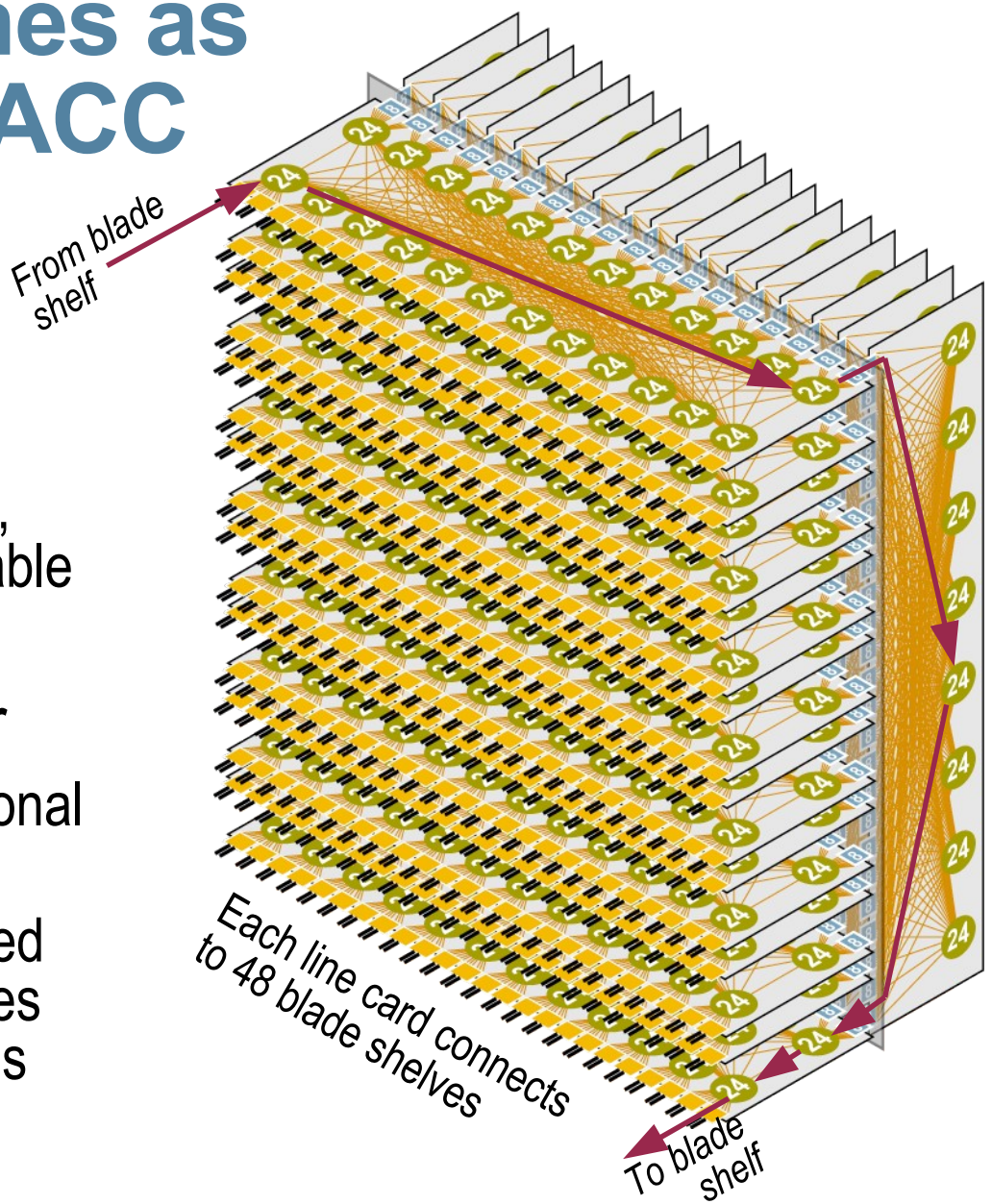
- Per blade:
  - 4 quad-core CPUs
  - 16 2 GB DIMMs
  - 8 GB flash for booting
  - One PCIe 8x connection
  - One Mellanox IB HCA (on NEM)
- Per shelf:
  - One 24-port IB leaf switch
  - Four 12x cables, each to a different line card
  - CMM connection to 100 MbE



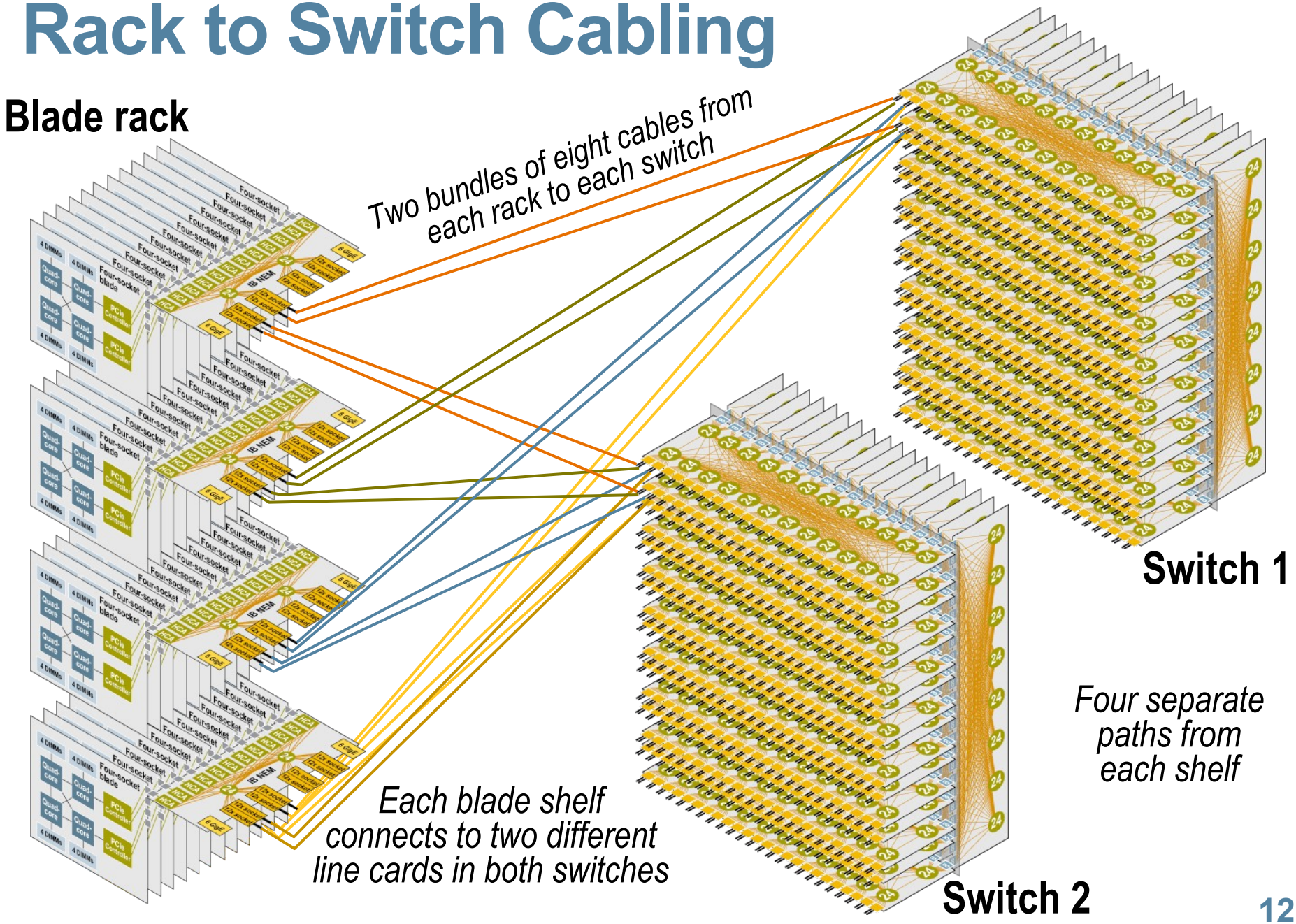
NEM GigE ports, second NEM switch chip, and PEM slots are not used

Magnum Switches as Configured at TACC

- Two Magnum switches
  - Each with 16 line cards
  - A total of 4,608 4x ports
  - Line cards cabled in pairs, with empty slots left for cable access
- Largest IB network so far
  - Equivalent to 42 conventional 288-port IB switches
  - “Only” 1,400 cables needed — conventional IB switches would require 8,000 cables

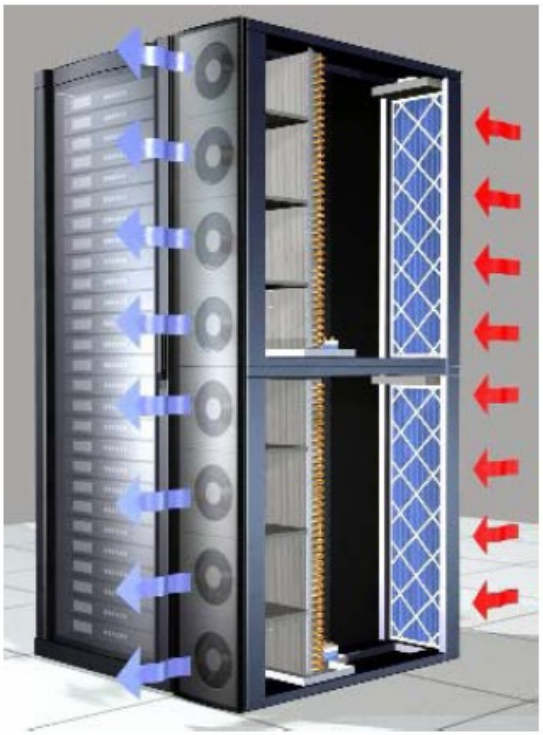


Rack to Switch Cabling

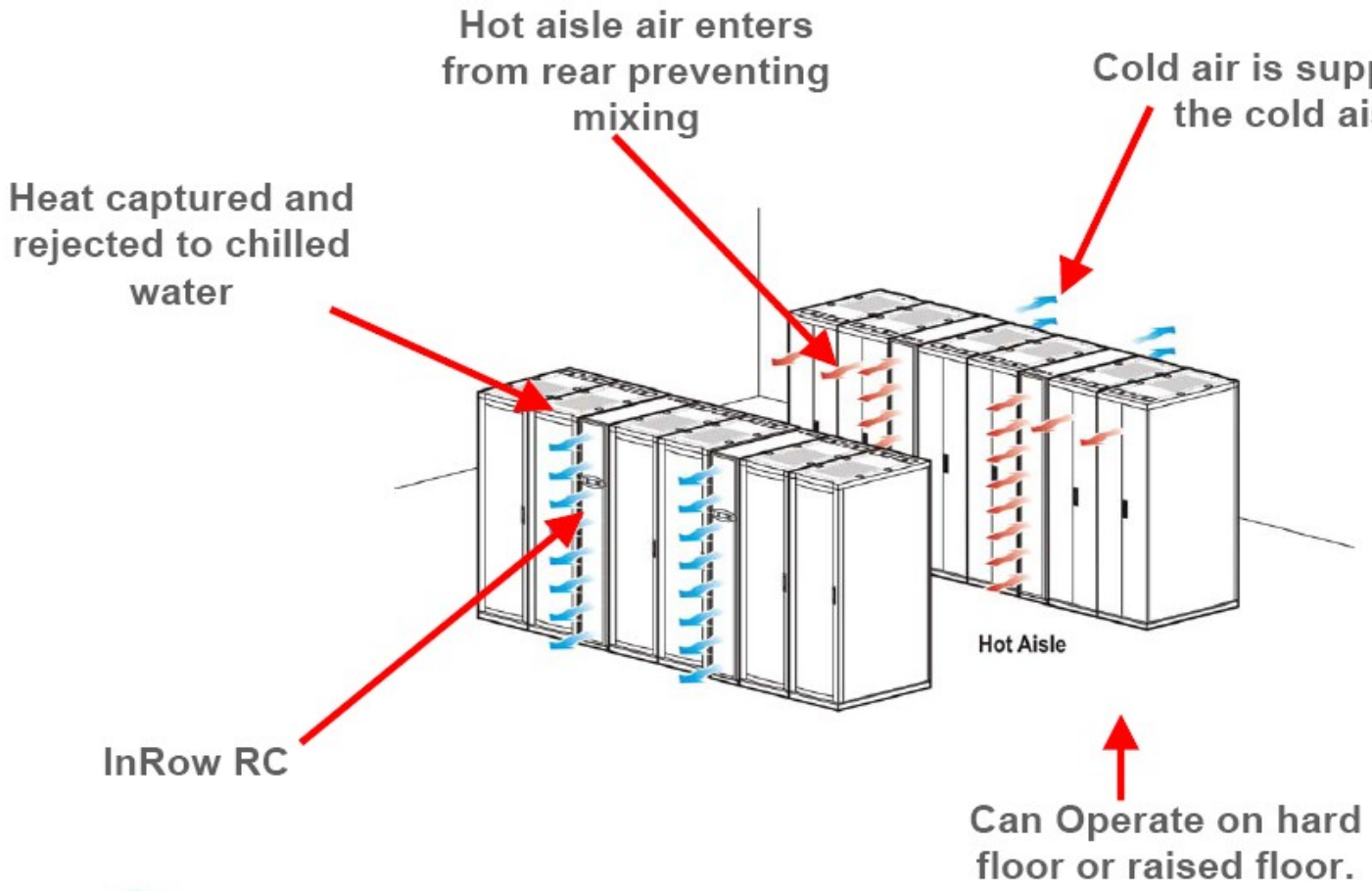


Sun Blade 6048 Cooling Option with APC (or Liebert)

- Chilled Water or Refrigerant
- Variable Capacity Control
- kW Metering
- Front & Rear Serviceable
- Network Manageable



In-Row Chillers with Sun Blade 6048



a company of Schneider Electric
InfraStruXure InRow RC

APC | MGE
Critical Power and Cooling Services

The Software Stack

TACC owns the stack

- **Base**
  - Linux – CentOS 4.4
  - Kernel 2.6.18.8
  - Kernel mods 2.6.26 prefcounter
- Mellanox Firmware
- UCSD – ROCKS 4.2.1
- OpenSM

- Grid Engine – latest
- Lustre 1.6.4.2
- OFED Stack 1.2.5.4
- Latest IBSRM - pending
- Scripts to manage the system.
- xVM OpsCenter - pending

For the applications:

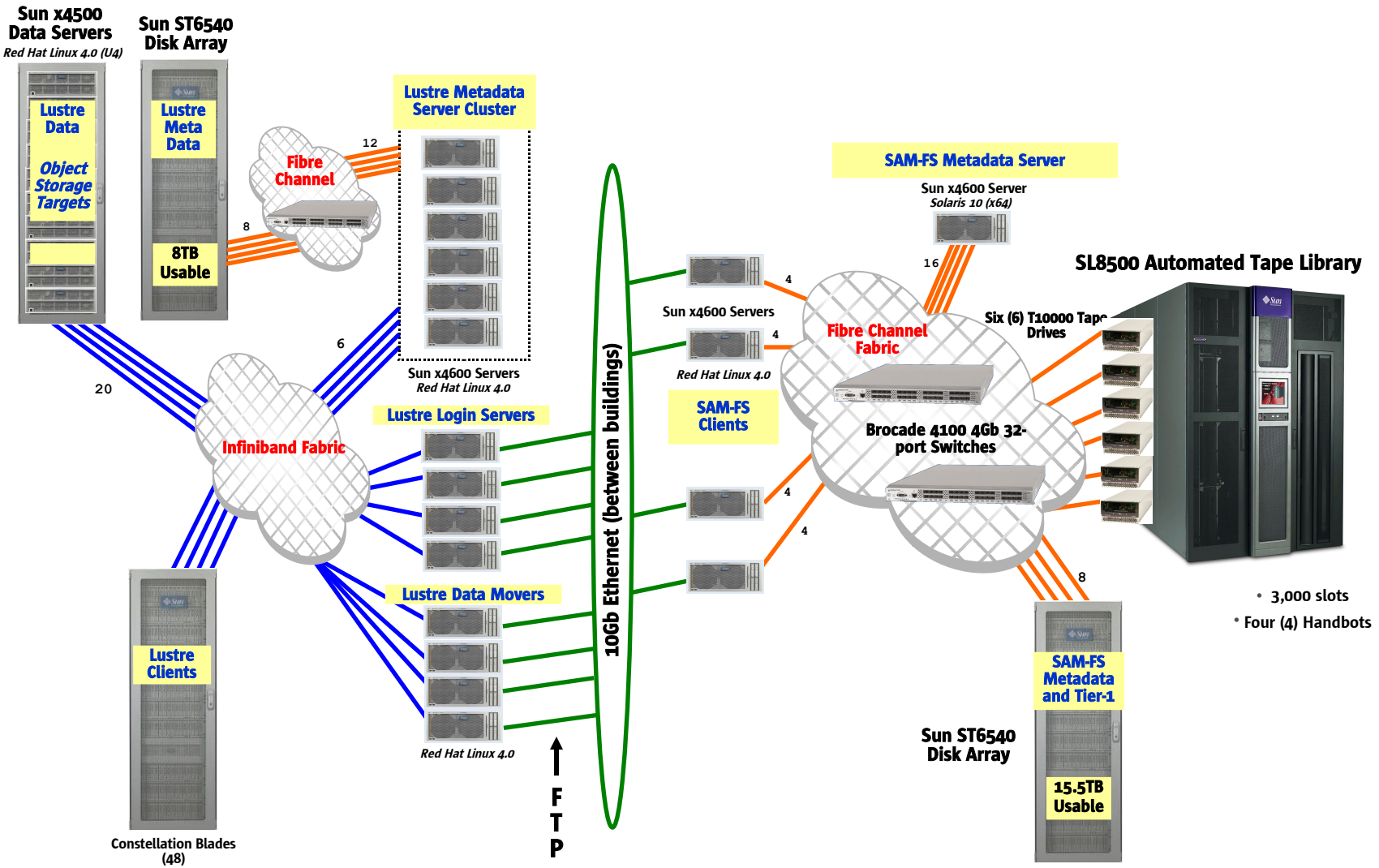
Compilers

- *Portland Group*
- *GCC*
- *Pathscale*
- *Intel*
- *Sun Studio*

Tools

- *MVAPICH and*
- *MVAPICH2*
  - *With Tuning help from OSU (Dr. Panda students)*
- *OpenMPI*

TACC Petascale Storage Design Overview
(Phase 1)



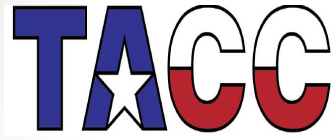
Challenges

- System Usability tools are desperately needed for extremely large scale clusters
  - TACC invested a huge amount of $$ (more than hardware) on tools and processes
- Data Access architecture requires extremely well-built systems
- Moore's Law applies for supercomputers, but still huge demands on Watt count and rack cooling
  - 2.5MW required for Ranger
- Do not try to anticipate too much on components availability, delivery issues DO happen

TACC Ranger

QUESTIONS??

Philippe Trautmann
philippe.trautmann@sun.com



Challenges

- System Usability tools are desperately needed for extremely large scale clusters
  - TACC invested a huge amount of $$ (more than hardware) on tools and processes
- Moore's Law applies for supercomputers, but still huge demands on Watt count and rack cooling
  - 2.5MW required for Ranger
-