National Aeronautics and Space Administration

Pleiades
Field Experience on a 10,000+ Node Subnet

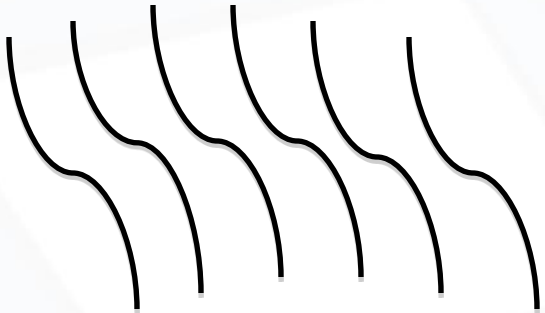
Bob Ciotti
Supercomputing Systems Lead/System Architect
Open Fabrics Alliance - 2011

Agenda

- Facility/Mission
- Pleiades Systems and Infiniband Subnet
- Issues with Scale

Facility/Mission

$$, 6 MW



Mission
Science/Engineering

Facility/Mission

We are mostly users of infiniband

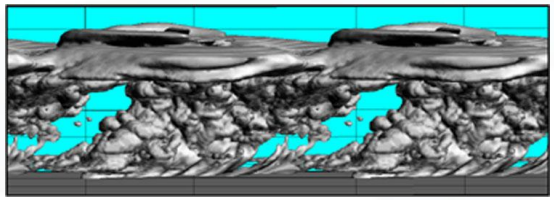
Some feature development

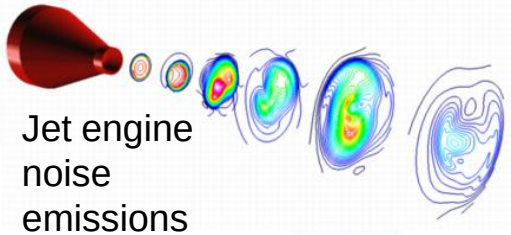
Supercomputing Support for NASA Missions

- Agency wide resource
- Production Supercomputing
  - Focus on availability
- Machines mostly run large ensembles
- Some very large calculations (50k)
  - Typically o500 jobs running

- Example applications
- ARMD
  - LaRC: Jet wake vortex simulations, to increase airport capacity and safety
  - GRC: Understanding jet noise simulations, to decrease airport noise
- ESMD
  - ARC: Launch pad flame trench simulations for Ares vehicle safety analysis
  - MSFC: Correlating wind tunnel tests and simulations of Ares I-X test vehicle
  - ARC/LaRC: High-fidelity CLV flight simulation with detailed protuberances
- SMD
  - Michigan State: Ultra-high-resolution solar surface convection simulation
  - GSFC: Gravity waves from the merger of orbiting, spinning black holes
- SOMD
  - JSC/ARC: Ultra-high-resolution Shuttle ascent analysis
- NESC
  - KSC/ARC: Initial analysis of SRB burn risk in Vehicle Assembly Building

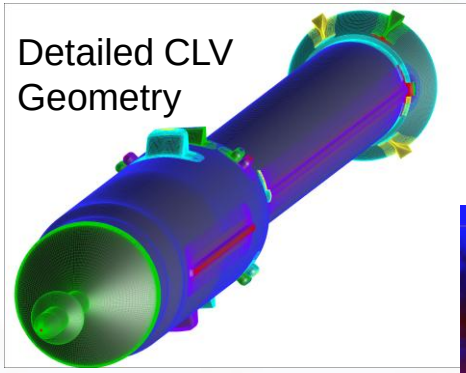
Jet aircraft wake vortices



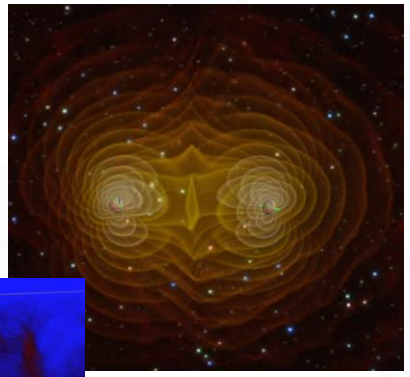


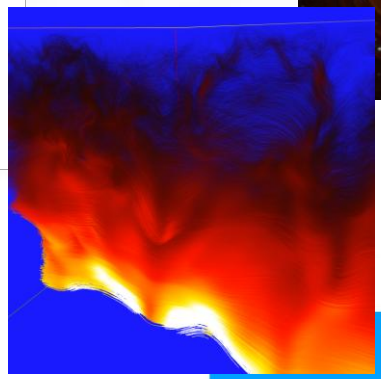
Jet engine noise emissions

Detailed CLV Geometry

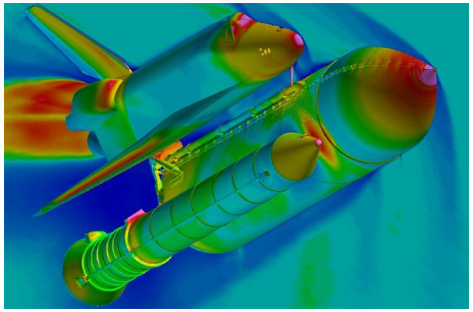


Orbiting, Spinning Black Holes

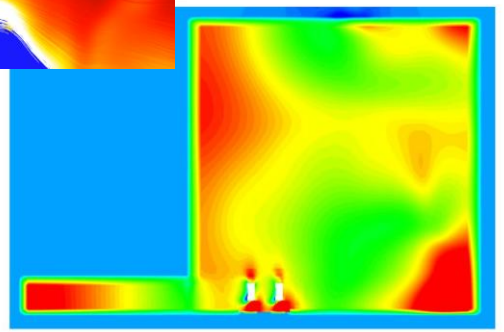




Solar surface convection



Shuttle Ascent Configuration



2-SRB Burn in VAB

Major Systems

Pleiades

Columbia

hyperwall

Systems Connectivity

Columbia

Bridge Node

<-fc lb1->

Archive Server

Pleiades

lb0 288sdr | lb0 96ddr | lb0 24ddr

lb0

hyperwall

lb1

lustre — Oss Oss Oss Oss Mds Mds — lb0 288ddr

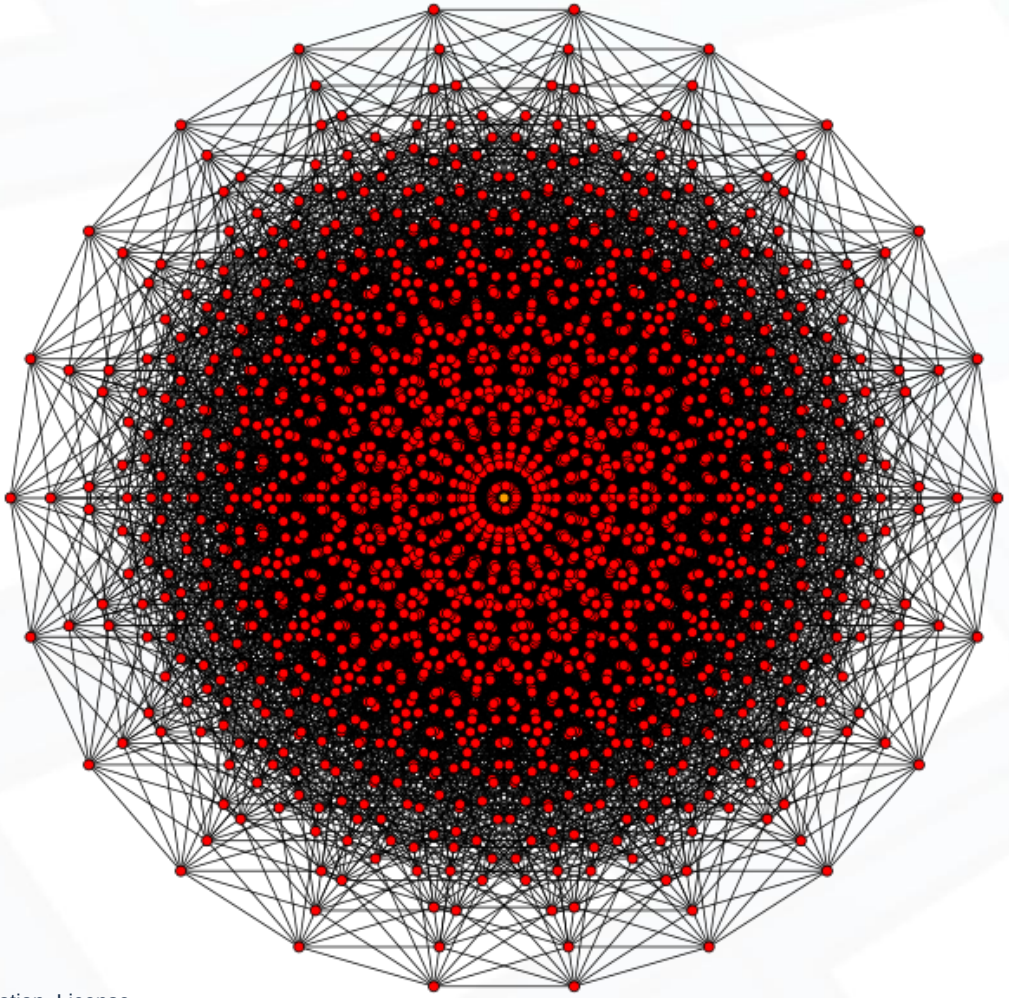
Oss Oss Oss Oss Mds Mds — lustre — Mds Mds Mds Mds

Infiniband – Topology

Partially Populated 11d Hypercube

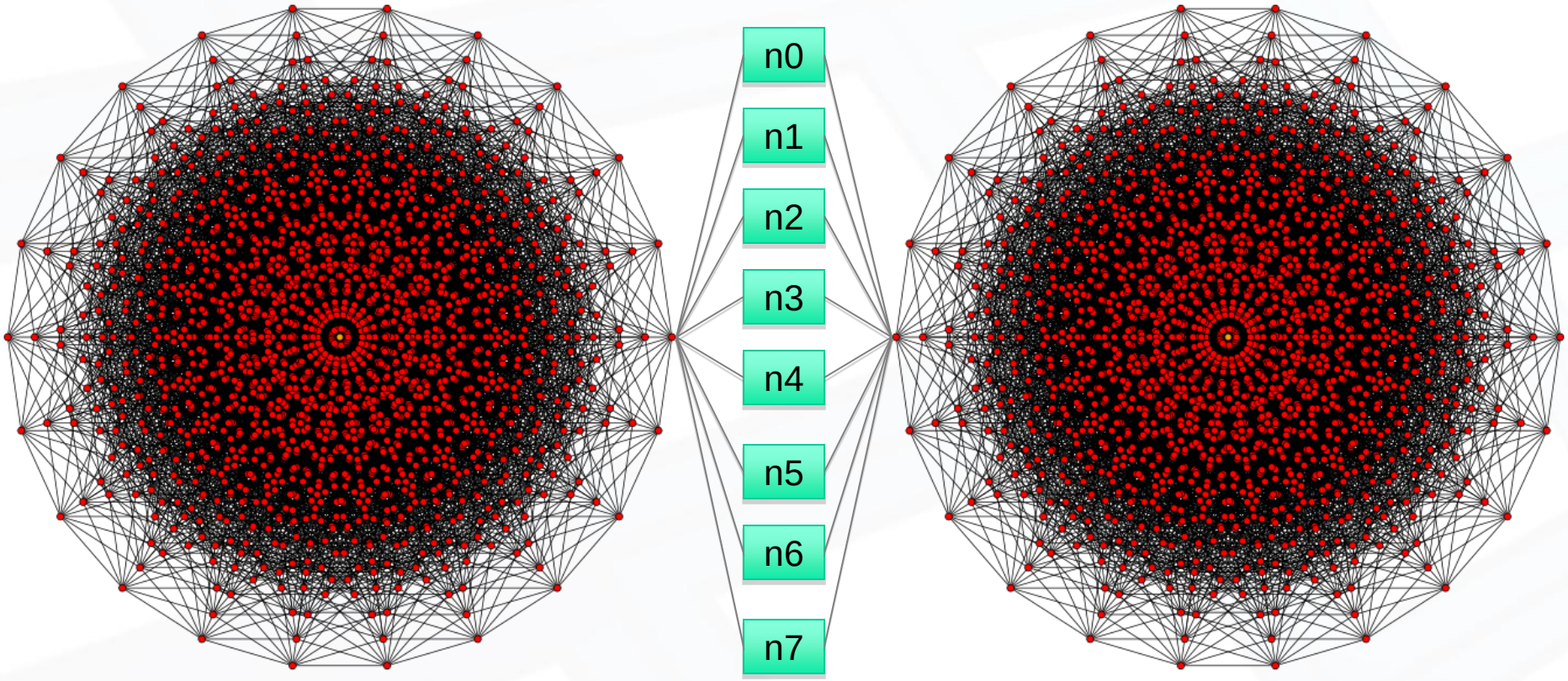
- Subnet manager algorithm
  - Minimum Hop Count
  - Break Ties by Port Number

→Dimension Ordered Routing (DOR)



(Orthographic demidekeract
by Claudio Rocchini, wikipedia
Copyright GNU http://en.wikipedia.org/wiki/GNU_Free_Documentation_License
Creative Commons 3.0 http://creativecommons.org/licenses/by/3.0/)

SGI ICE Dual Plane – Topology



ib0

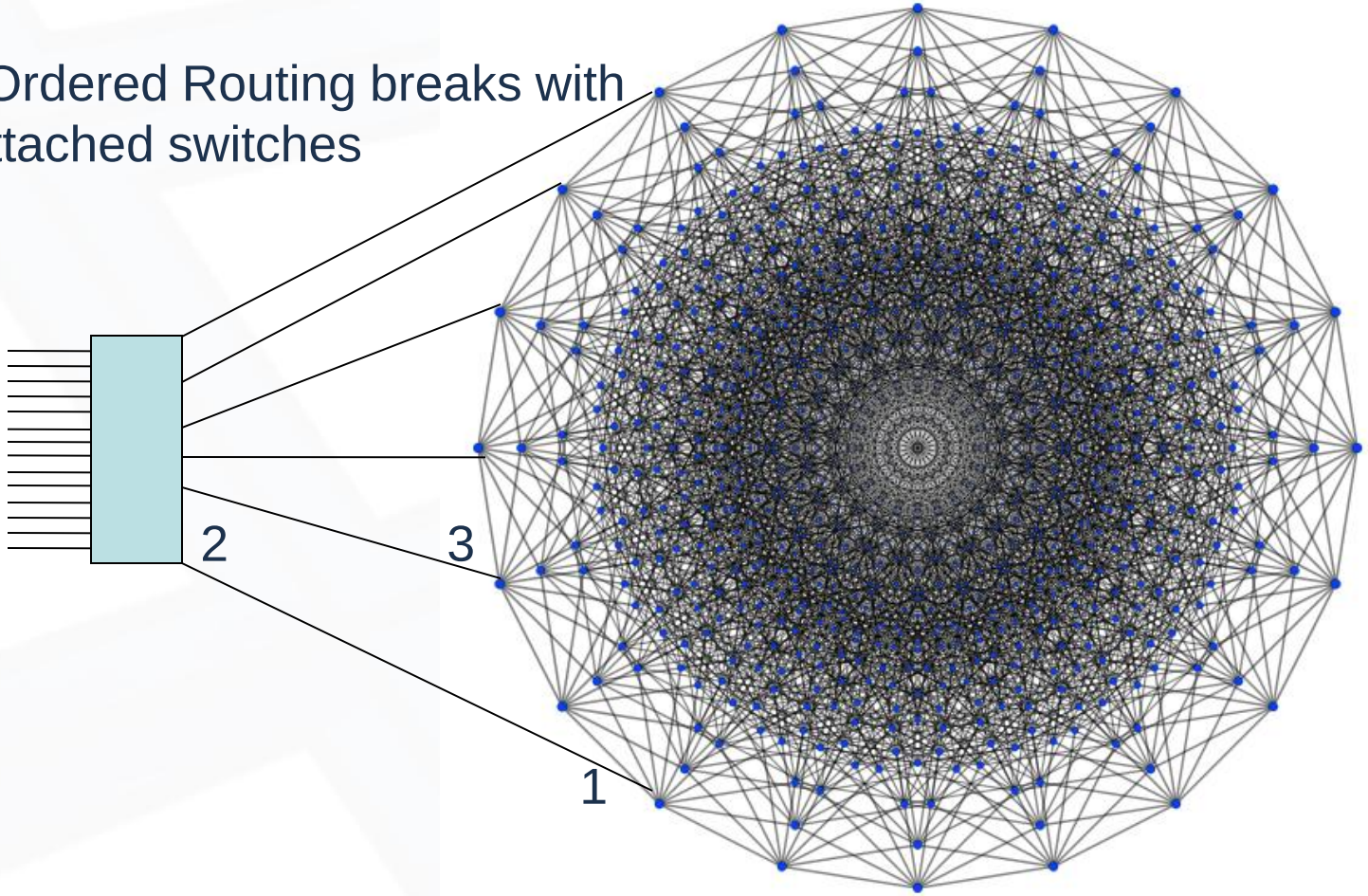
2x 11d hypercube
full 2048 vertices
Pleiades 1344/11d

ib1

http://en.wikipedia.org/wiki/User:Qef/Orthographic_hypercube_diagram

Infiniband – Subnet Discovery

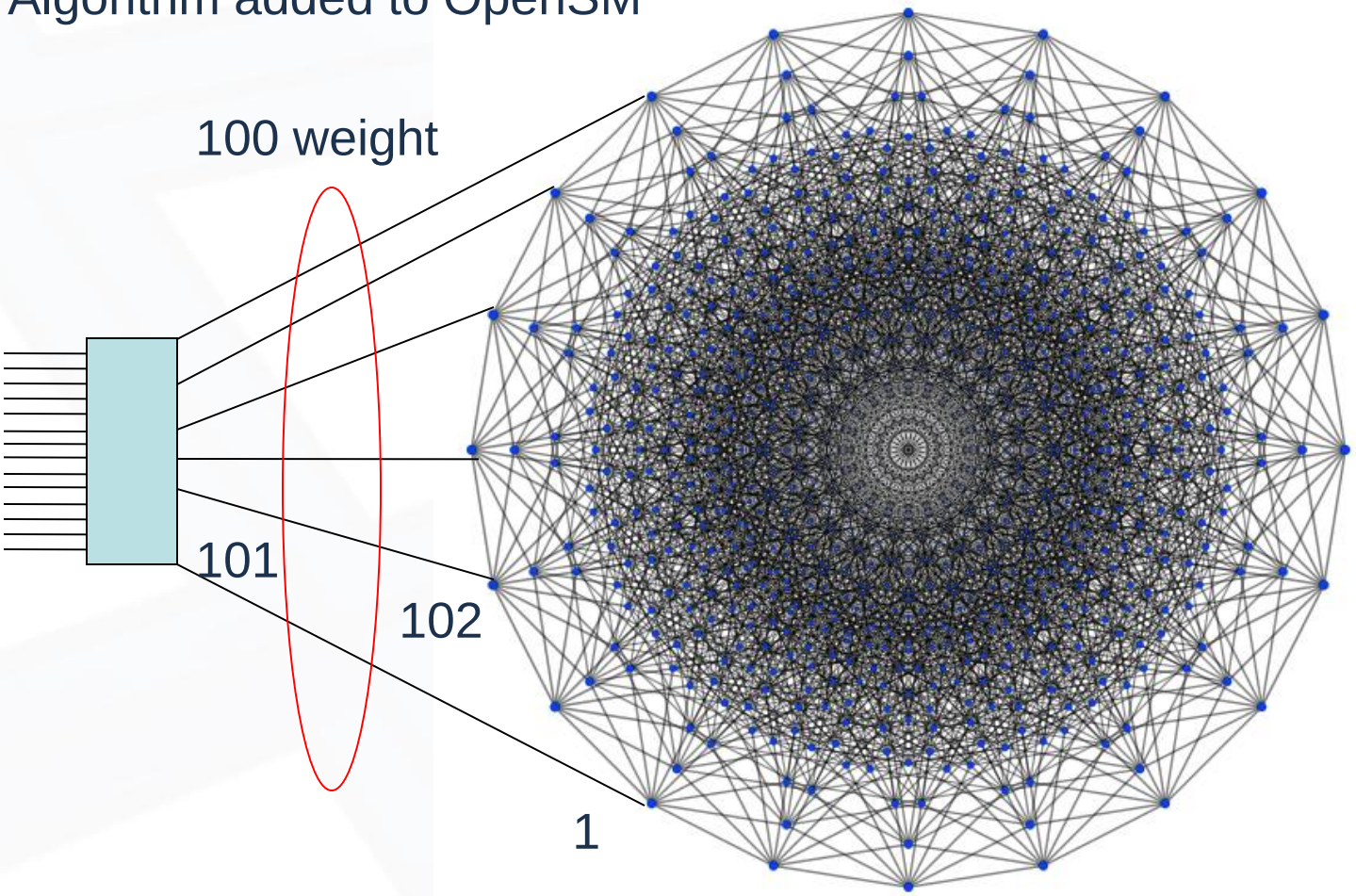
Dimension Ordered Routing breaks with externally attached switches



Orthographic demidekeract
by Claudio Rocchini, wikipedia

Infiniband – Subnet Discovery

- Weighting Algorithm added to OpenSM



Orthographic demidekeract
by Claudio Rocchini, wikipedia
Copyright GNU http://en.wikipedia.org/wiki/GNU_Free_Documentation_License
Creative Commons 3.0 http://creativecommons.org/licenses/by/3.0

Infiniband Subnet LAN

LAN Implemented with out board IB switches

Archive Servers

NFS File Servers

Hyperwall Graphics System

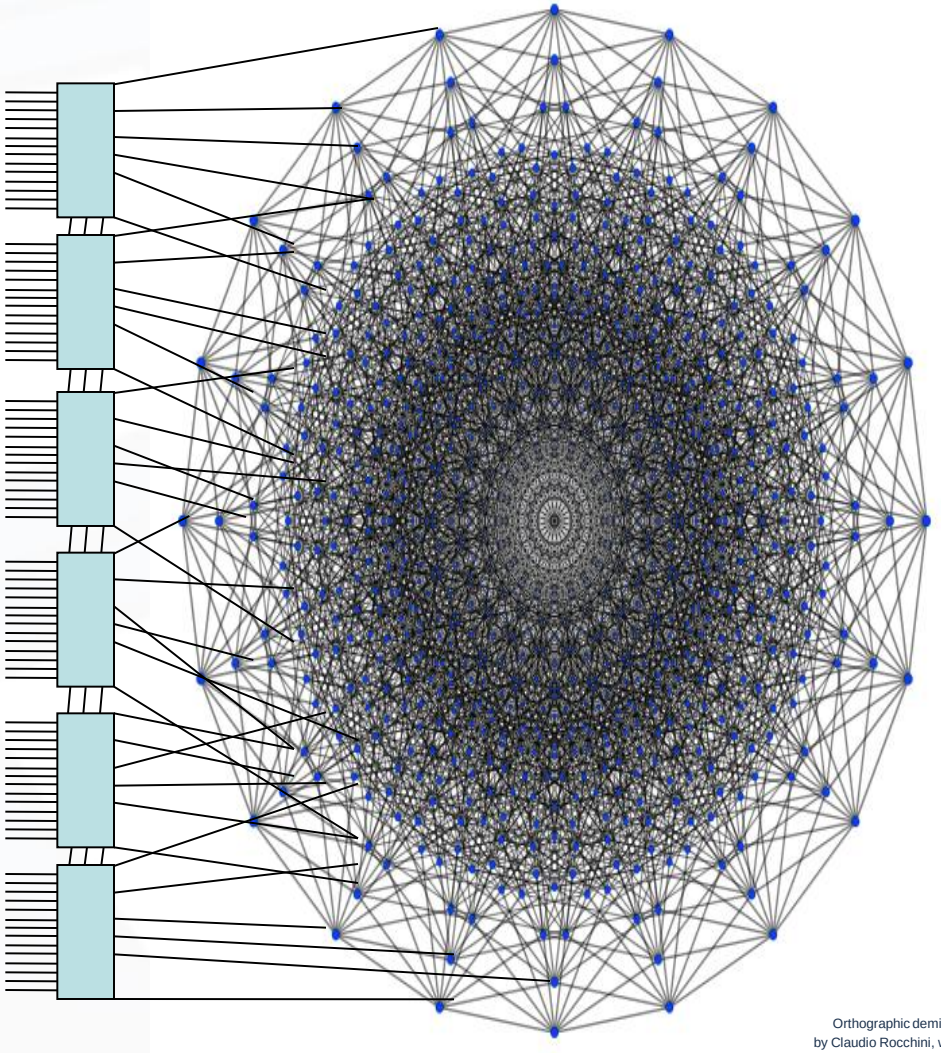
Columbia Nodes

Columbia Bridge Nodes

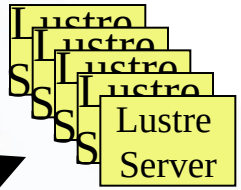
Front End Nodes

Data Analysis Nodes

Lustre Filesystems

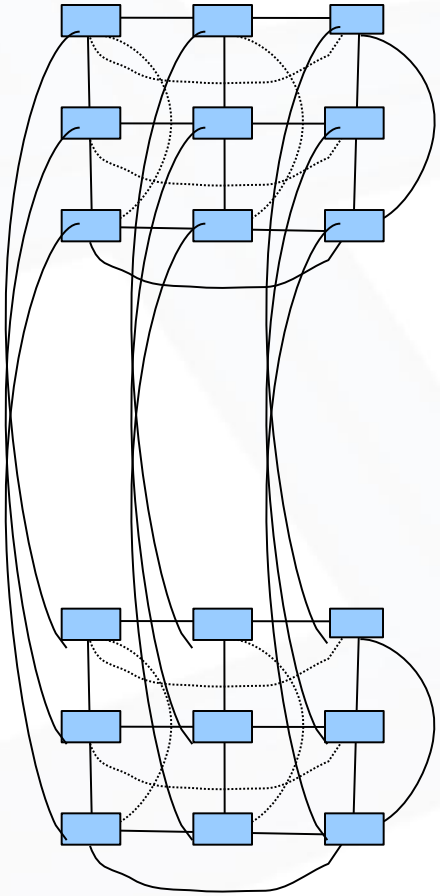


I/O Network

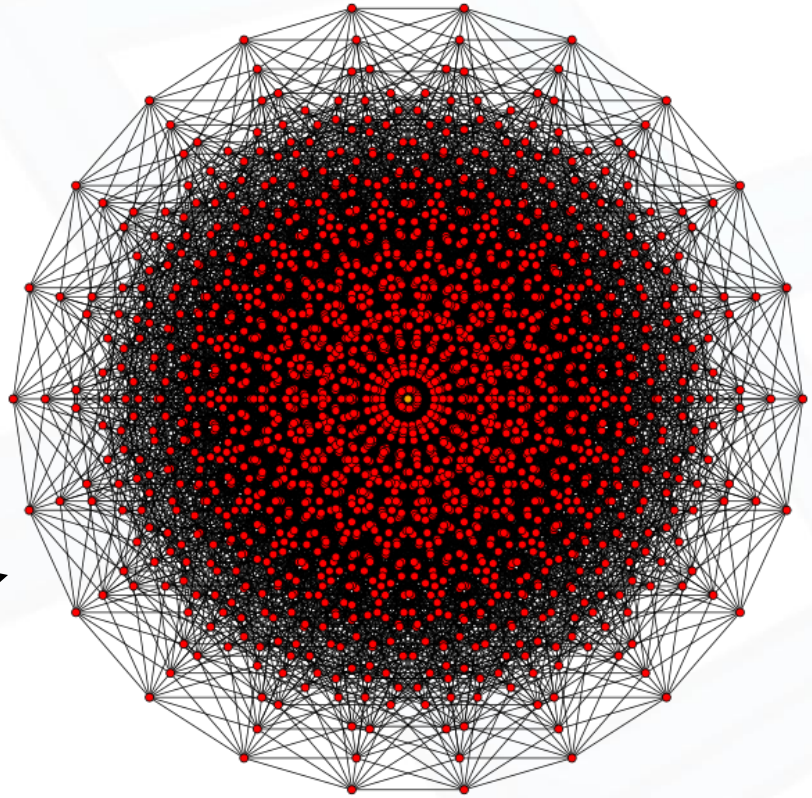


9 cables

r999



125 cables



32 cables

r998

160 cables

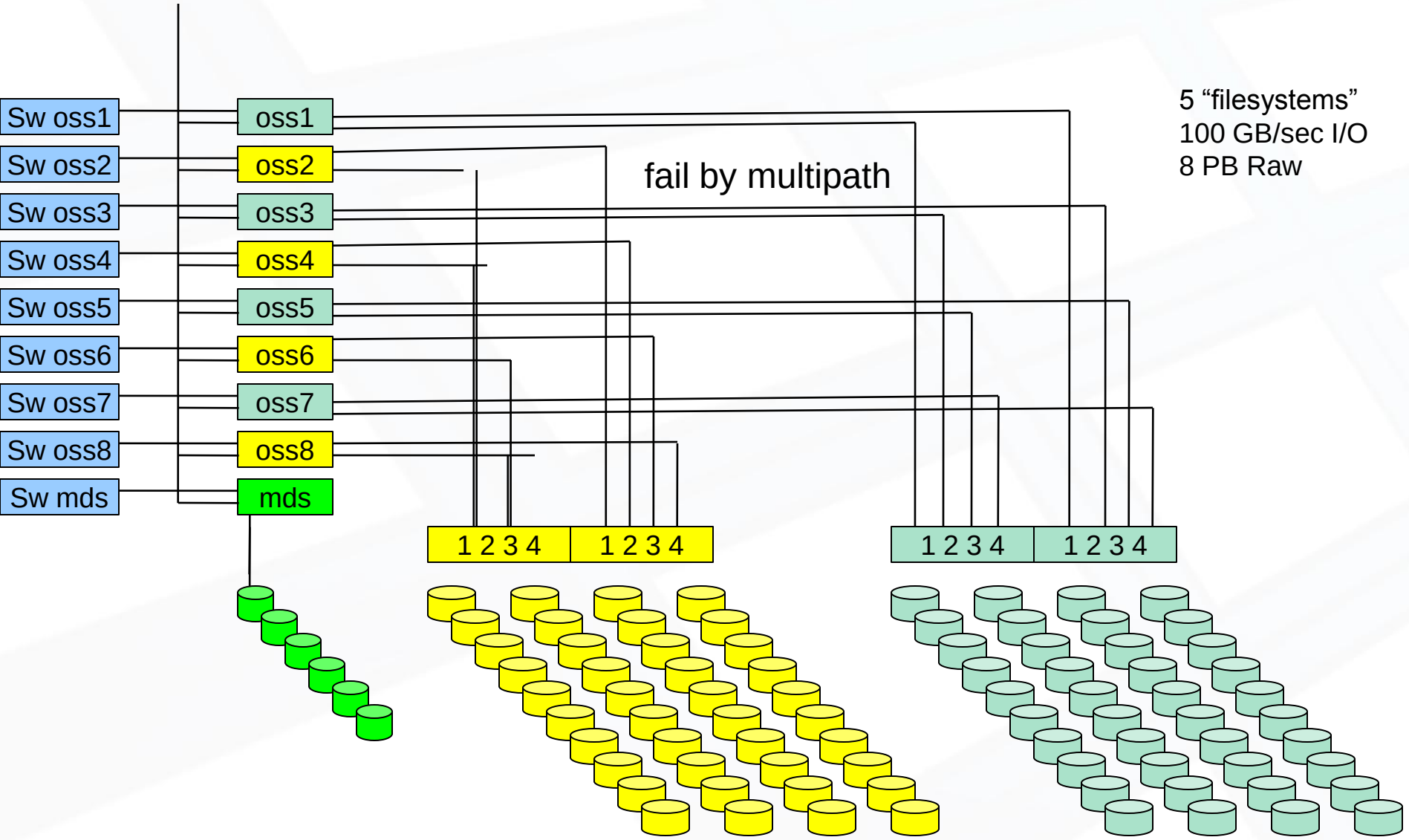
Hyperwall
128-Display
Graphics Array

I/O fabric

ib1

Target Lustre Filesystem

Gen2 I/O Fab



5 “filesystems”
100 GB/sec I/O
8 PB Raw

Pleiades Current Configuration
SGI ICE System

- 10,880 nodes – 21,760 sockets - 101,376 x86 cores
  - 10,816 compute/128 Vis (opteron)
  - X5670/X5472/X5570 Xeon @3.0/2.93ghz
    - (harpertown, nehalem, westmere)
- Infiniband Back End disk subsystem for lustre

- # mount on most nodes (5 nfs servers/7 lustre filesystems)
  - delta-ib1-0 /mnt/home1 nfs
  - galileo-ib1-0 /mnt/home3 nfs
  - pioneer-ib1-0 /mnt/home4 nfs
  - delta-ib1-0 /mnt/nasa nfs
  - saturn-ib1-0 /mnt/nobackup nfs

  - service110-ib1@o2ib /nbp10 lustre
  - service150-ib1@o2ib /nbp20 lustre
  - service200-ib1@o2ib /nbp30 lustre
  - service159-ib1@o2ib /nbp50 lustre
  - service100-ib1@o2ib /nbp40 lustre
  - service100-ib1@o2ib /nbp60 lustre
  - service160-ib1@o2ib /nbp70 lustre
  - service140-ib1@o2ib /rtj-home lustre

Pleiades - Sustained SpecFP rate base

- **SpecFP rate base estimates** (eliminates cell/GPU/blue-gene/SX vec)

Spec	Top500	Machine	CPU	#Sockets	FPR/Socket	TSpec
1	2	Jaguar	AMD-2435	37,360	65.2	2,436,246
2	**6**	Tera-100	Intel-7560	17,296	133.4	2,307,805
3	**5**	Hopper	AMD-6176	12,784	149.8	1,800,115
4	**1**	Tianhe-1a	Intel-x5670	14,336	119.5	1,713,868
5	**11**	**Pleiades**	**Intel-x**	**21,632**	**72.2**	**1,562,510**
6	**10**	Cielo	AMD-6136	13,394	115.5	1,547,408
7	8	Kraken	AMD-2435	16,448	65.2	1,075,182
8	14	RedSky	Intel-x5570	10,610	90.3	958,401
9	17	Lomonosov	Intel-x5570	8,840	90.3	798,517
10	15	Ranger	AMD-2356	15,744	37.3	588,196

- Tspec == number of 2-core 296mhz UltraSPARC II

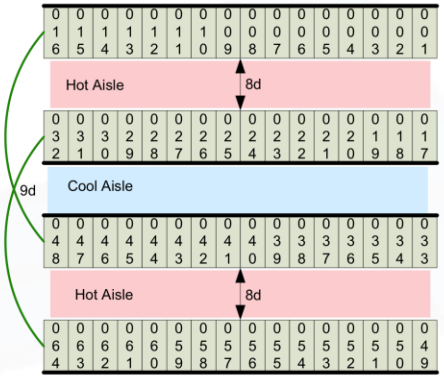
Pleiades Infiniband Specifics

- Mix of infinihost III, Connect-X DDR, Connect-X QDR HCA
  - ~12,379 cables (over 50 miles - combination of optical/copper)
  - 21,704 active host ports
- Mix of infiniscale III and infiniscale IV switches
  - 2,706 total switch chips
  - 51,438 active switch ports
- Two Major subnets (>10,000 endpoints)

- 73,142 ports (21,704 hca, 51,438 switch == ~7 ports/node)
  - 36,571 port-port links
  - 24,192 backplane
  - 12,379 cables (>50 miles, average length 7m)

- 1.6 million base counters (+extended/mellanox specific)

NASA (Pleiades) Rack Layout



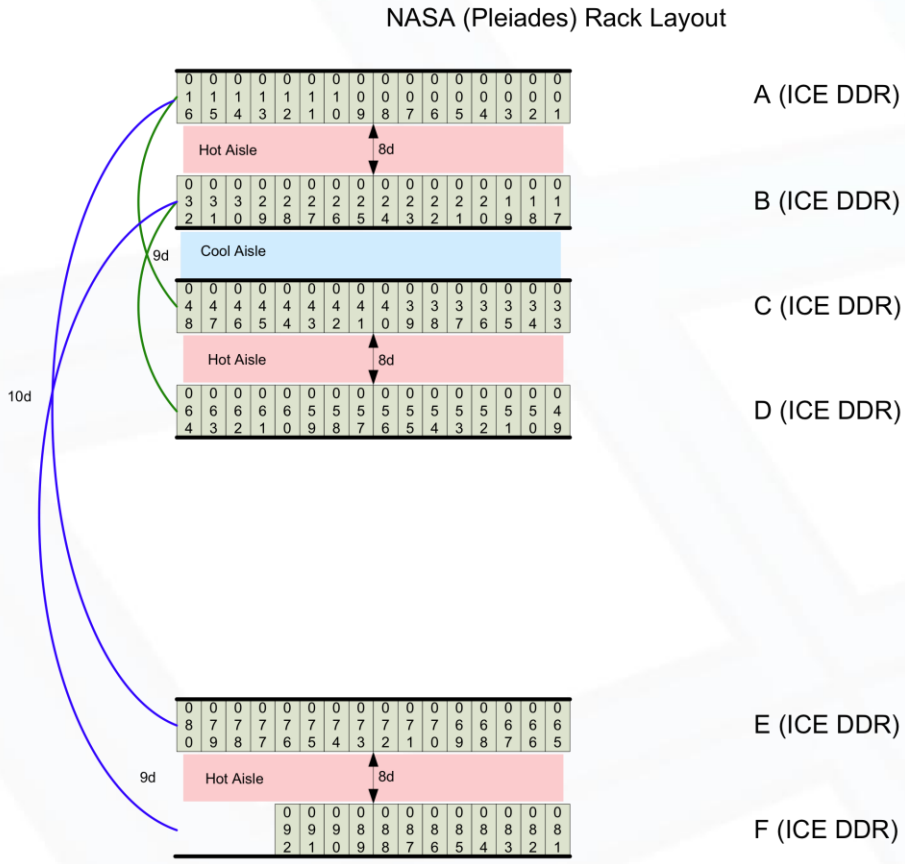
A (ICE DDR)

B (ICE DDR)

C (ICE DDR)

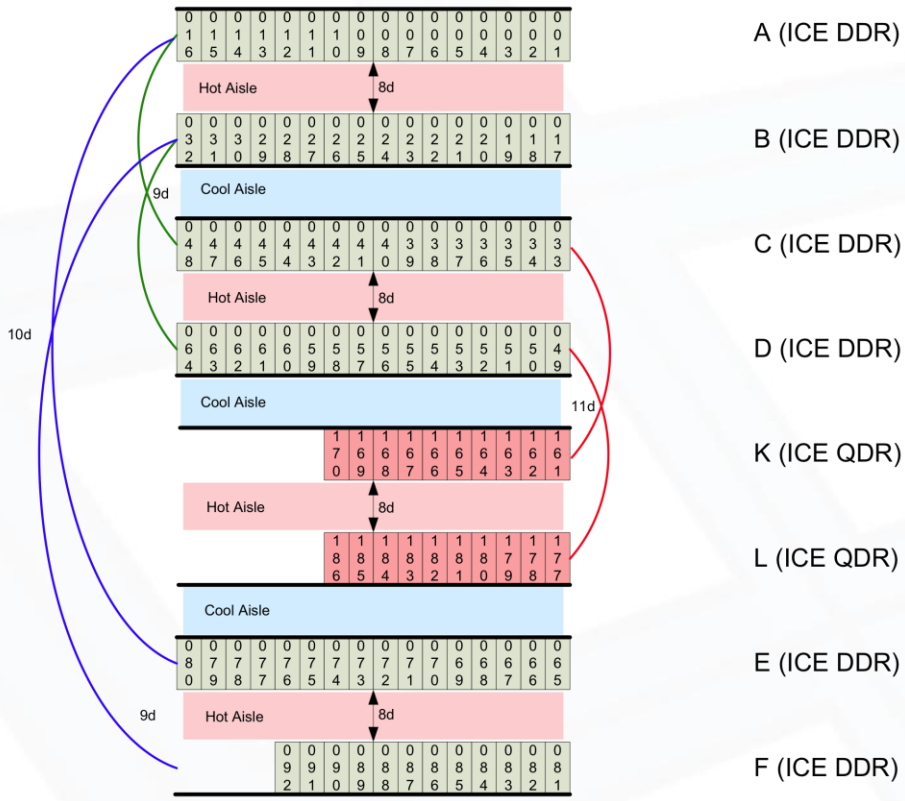
D (ICE DDR)

64 racks - 2008



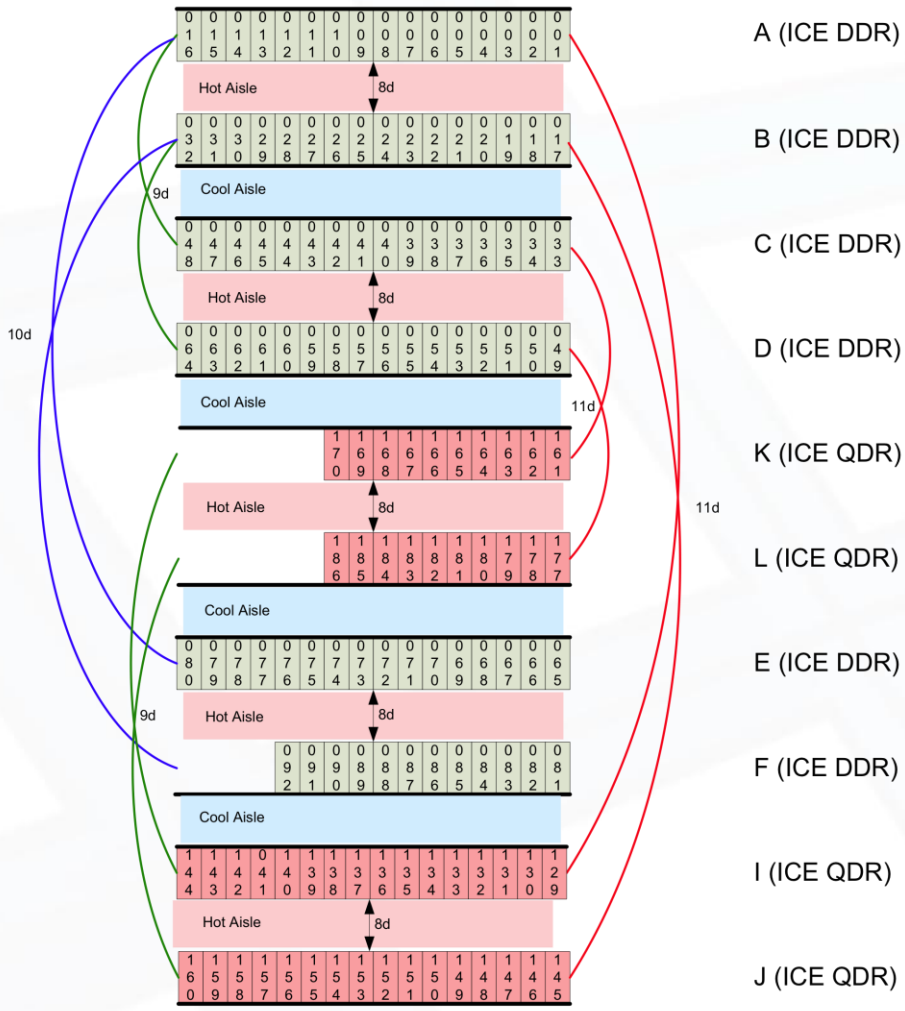
92 racks - 2008

NASA (Pleiades) Rack Layout



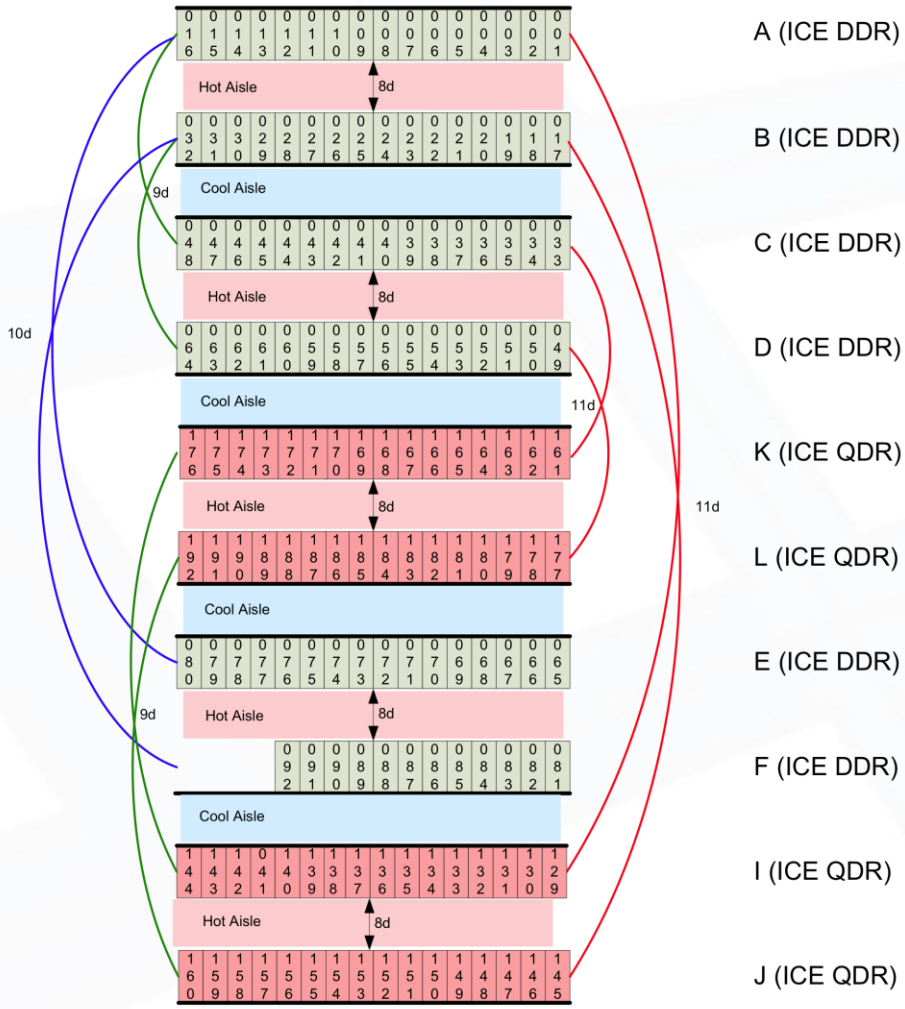
112 racks - 2009

NASA (Pleiades) Rack Layout



144 racks - 2010

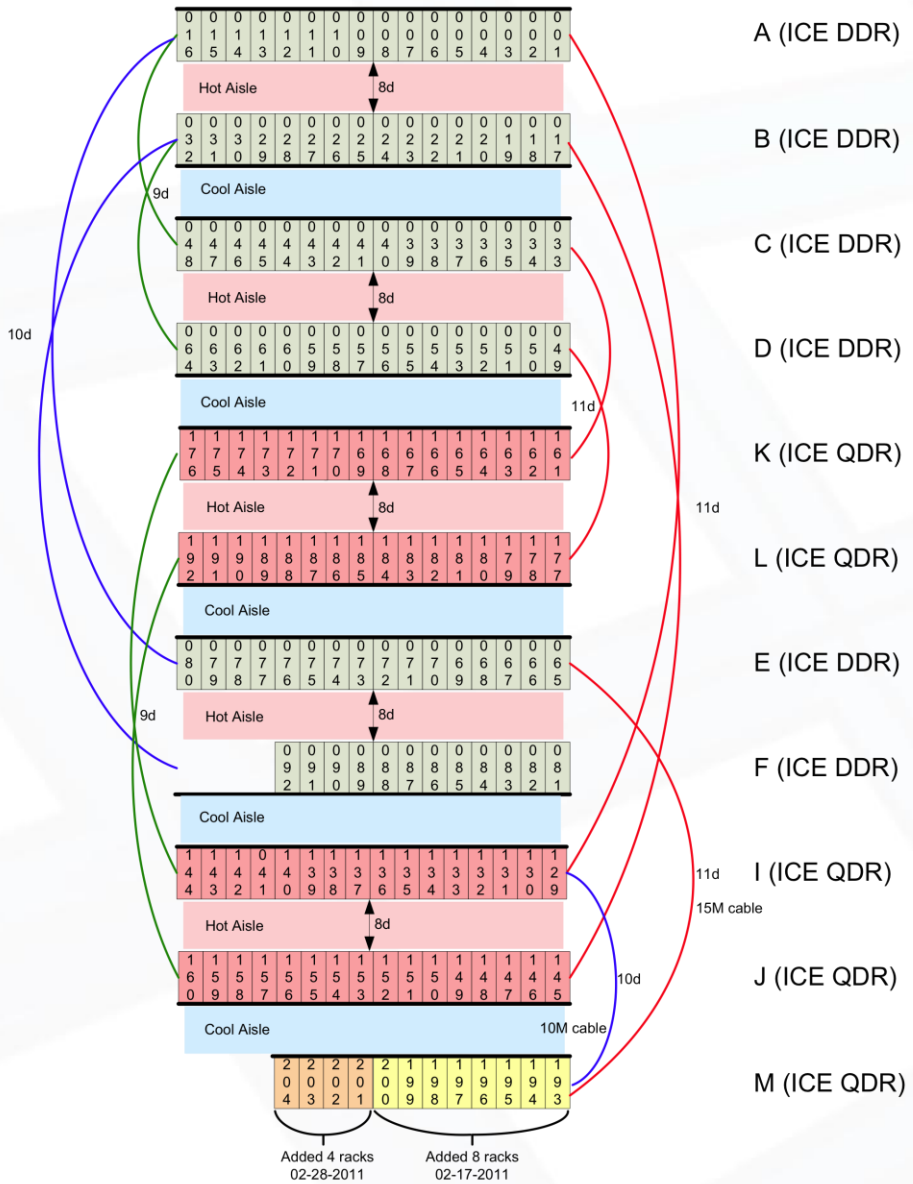
NASA (Pleiades) Rack Layout



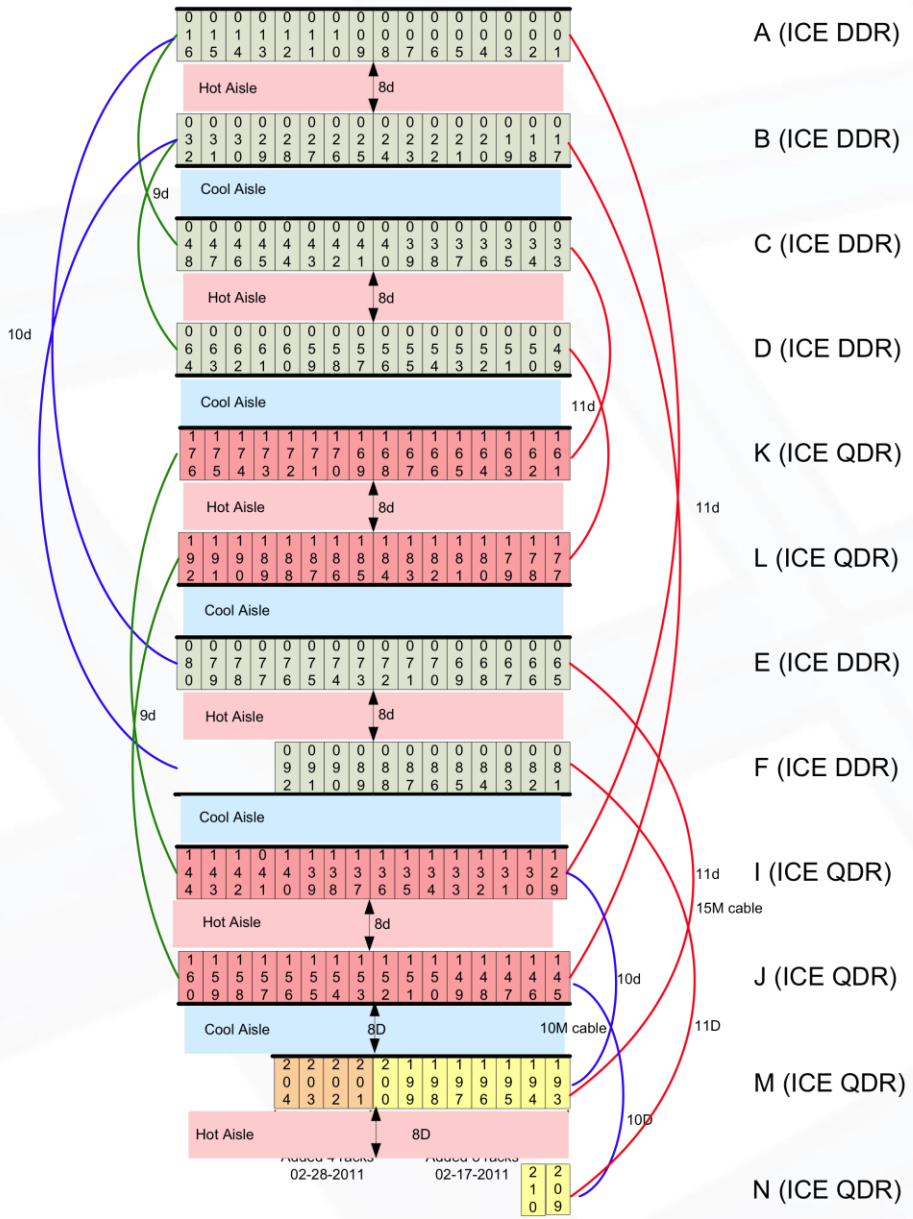
156 racks - 2010

NASA (Pleiades) Rack Layout

168 racks - 2011



NASA (Pleiades) Rack Layout



170 racks - 2011

Continuous Availability
24/7 Operations

- Goal – never take the whole system down
  - Outages are very disruptive (except NFS)
  - Dedicated time very costly
  - Not even possible to update entire system in one dedicated session.
  - Things go wrong

- Components
  - Lustre, NFS, CXFS, OFED, OpenSM, Linux Distro patches, cluster management software,
  - Firmware
    - its in everything – including the cables.

Continuous Availability
24/7 Operations

- Rolling updates of various components
  - Lustre/NFS clients/compute node images
    - Easy – simply done at end of user job
  - NFS, Lustre servers
    - Hot swap
      - Nfs hard mounts
      - Lustre recovery
  - Schedule filesystems as a resource in addition to nodes
    - Allow us to use all compute nodes and figure out share later
- Various admin, front ends, bridge nodes are easier or less urgent.

Continuous Availability
24/7 Operations

- Hot Plug - Grow system while in operation
  - Cable up new components powered off
    - Check cabling
  - Signal OpenSM to turn off sweep
  - Power on equipment
  - Run ibnetdiscover to verify cabling
  - Signal OpenSM to sweep

Warning!

- Some issues on the following slides may have been solved or may be solved by work in progress.

  Searching Amazon for “infiniband for dummies” returned:

  How to fix Everything – for dummies

- Welcome any input or questions regarding anything we do

Management
experience

- Adding/removing switches is disruptive.
  - a) recovery often forced up to library layer
  - b) takes too long to reprogram switches
  - - how long do we have before Error returned to QP?
  - Should be able to leverage facts about the system:
    - E.g. This port on switch GUID x will always have an HCA on it
- Need robust mechanism for switch discovery and routing when since nodes may not always be responsive to SMP traffic. Objective is to sweep large fabrics in under 30 seconds. To better provide for planned maintenance, consider rerouting prior to bringing down a switch.

Dueling SMs
experience

- Default installs that include or require SM
  - Bad
  - need diags/utils everywhere
- SM failover problematic
  - Network connectivity will come and go
  - All IB SW needs to ride through instability
  - large IB fabric go unstable
  - Network instability can be localized
    - Dedicated network between two servers
    - Deliberate action to enable SM

Cable Issues
experience

- Physical layer cable stability
  - Over the years MANY different odd cable problems.
    - Too long copper
    - Too weak laser
    - Temperature – too cold
    - Broken fibers in packaging
    - Reseats often required
  - Enet and FC haven't had these issues
    - OK - they don't go as fast
  - QSFP/QDR in the field has NOT yet improved reliability over CX4/DDR.
  - Will FDR improve/worsen problem

Recovering from Credit Loop
experience

- Fairly common for us to have a credit loop
  - (try to avoid system outages)
  - Only locked up everything once!
  - Had to power down all nodes to bring things back
  - Related to ARP Storm/IPoIB MC subscription
    - TIP:
      - Some switches still subscribe to IPoIB MC group even when SM is disabled.
      - SM (whether switch or server based) does not require IPoIB
      - ARP/IPoIB MC causes catastrophic VL15 drops
      - Ifdown ib*
- Looking at possible solutions (lash, taking other ports away, …)

VL15 limitations
experience

- DDR switches much more prone to drop VL15 traffic
- limits outstanding SMP SM settings in SM
- ARP storms causes interference on anything subscribed to IPoIB mc
- "smart" switches
- CA - most importantly SM node

Human Interface
experience

- Human Interface issues
  - GUID/LIDS/Ports/Subnets
  - Easy to get confused
  - Node name maps – right direction

Management Issues
feature

- Parallelism
  - As systems grow, network complexity multiplies
- SA queries
- SM scans
  - Prune work in heavy sweep
    - timeout/light sample helpful but insufficient
  - More SMP not option in older switches
  - Parallel discovery/route
  - target a large system discover/route in seconds even some number of seconds*)
- Performance and error collection
  - sub second target

dasw

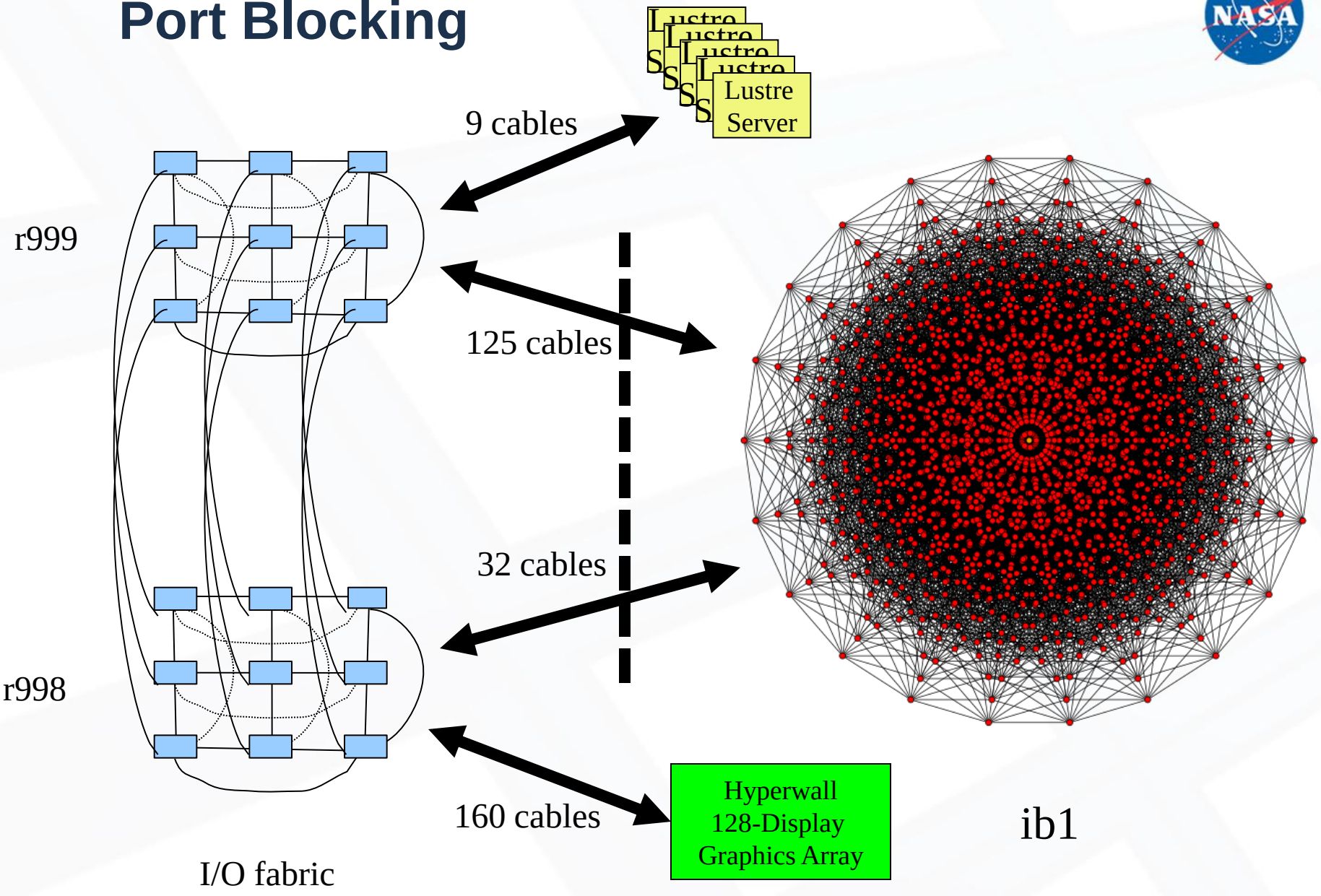
```
# time /usr/local/ib/dasw -ap1
Fri Apr  1 22:42:26 PDT 2011
/usr/local/ib/dasw -ap1
last lid/guid  map on Sat Feb 12 02:10:53 PST 2011 license 1
last counter reset at Mon Feb 28 11:50:33 PST 2011
last counter  read at Fri Apr 1 22:41:47 PDT 2011

Checking 1344 ib0 switches (448 sw0 224 sw1 672 sw3) switches

   -- cb1 ib0 sw0    .   .    .              .   n2 n3  n0  n1  n5  n4  d1  n7  x n6 d2 d3 d4 d5 d6 d7 d8 d9 d10 d11 d12  io
   == cb1 ib0 sw0    .   .    .              .    1  2   4   6   7   8   9  10 11 12 13 14 15 16 17 18 19 20  21  22  23  24
 r72i3 cb1 ib0 sw0 SwLid  6535 port-1       SymbolErrors:    .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  5  .  .

   -- cb3 ib0 sw1    .   .    .              .   n0 n1  n2  n3  n4  n5  n6  n7 d1 d1 d1 d1 d2 d3 d4 d5 d6 d7  d8  d9 d10 d11 d12 d13 d14 d15 d16 d17 io
   == cb3 ib0 sw1    .   .    .              .    1  2   3   4   5   6   7   8  9 10 11 12 18 19 20 21 22 23  24  25  26  27  28  29  30  31  32  33 34
r135i1 cb3 ib0 sw1 SwLid  7333 port-1       SymbolErrors: 21066 .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .
r135i1 cb3 ib0 sw1 SwLid  7333 port-1         RcvErrors:   132 .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .
r149i0 cb3 ib0 sw1 SwLid  9227 port-1 RcvRemotePhysErrors:    .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  1  .  .  .  .  .  .  .  .  .  .  .
r153i0 cb3 ib0 sw1 SwLid  9239 port-1 RcvRemotePhysErrors:    .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  1  .  .  .  .  .  .  .  .  .  .  .  .
r155i0 cb3 ib0 sw1 SwLid  8886 port-1 RcvRemotePhysErrors:    .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  1  .  .  .  .  .  .  .  .  .  .  .  .  .
r155i2 cb3 ib0 sw1 SwLid  8875 port-1 RcvRemotePhysErrors:    .  .  202  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .
r157i0 cb3 ib0 sw1 SwLid  8989 port-1 RcvRemotePhysErrors:    .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  1  .  .  .  .  .  .  .  .  .  .  .
real    0m1.183s
user    0m0.064s
sys     0m0.024s
```

Port Blocking



Port Blocking
feature

- Port Blocking
- use ibportstate to disable port
- can be used to disable a failing port/cable
- will not survive a power cycle on the switch
  - file that describes ports to never bring up/route across
  - needs to fairly dynamic

- SM feature to take list of GUIDs to ignore
  - Or guid/port to keep down

Other Features

- IB failures to appear in syslog
  - Give me some useful log information.
- Need an OFED testing matrix:
  - - what versions of OFED work with what hardware/software?
  - - what software works with what library versions / Linux kernel,
  - e.g., want to use version X of opensm – what versions of Libibmad, libibumad and linux kernel will work?
- Knowledge base/Wiki

Barriers

- Resistance in the admin community to use infiniband
  - General purpose networking.
    - Comfortable with ethernet and ethernet knowledge base
      - E.g. Image management, booting, logging.
    - Routing between IB/ethernet
    - Big ib networks do have limitations/“features”
  - Replace FC
    - Reuse ib equipment and knowledge
      - More risk/familiarity in storage community
  - Infiniband still ARCANE – steep learning curve
    - Need more centralized knowledge base

2x Scale Blockers (let alone 1000x)

- Link Fail/Rescans
  - Pretty close to breaking NOW
- ARP
  - causing vl15 drops
- SAQ Paths
- Congestion – will want to make better use of whats there.
  - May require per job routing
- Whole fabric may NEVER be up
  - External database of GUIDS, lids, locations, movement, etc
  - 10,000's of nodes down
    - link active/no sa response
  - links may (will) constantly be going up and down

Future Directions

- Improve participation in OFA
- User Space Path Queries (ib_acm)
- Hierarchical ARP extension
- Performance/Error monitoring
- Cycle Prevention for degraded DOR system
- Topologically aware scheduling
- Upgrade campus extensions

Questions?