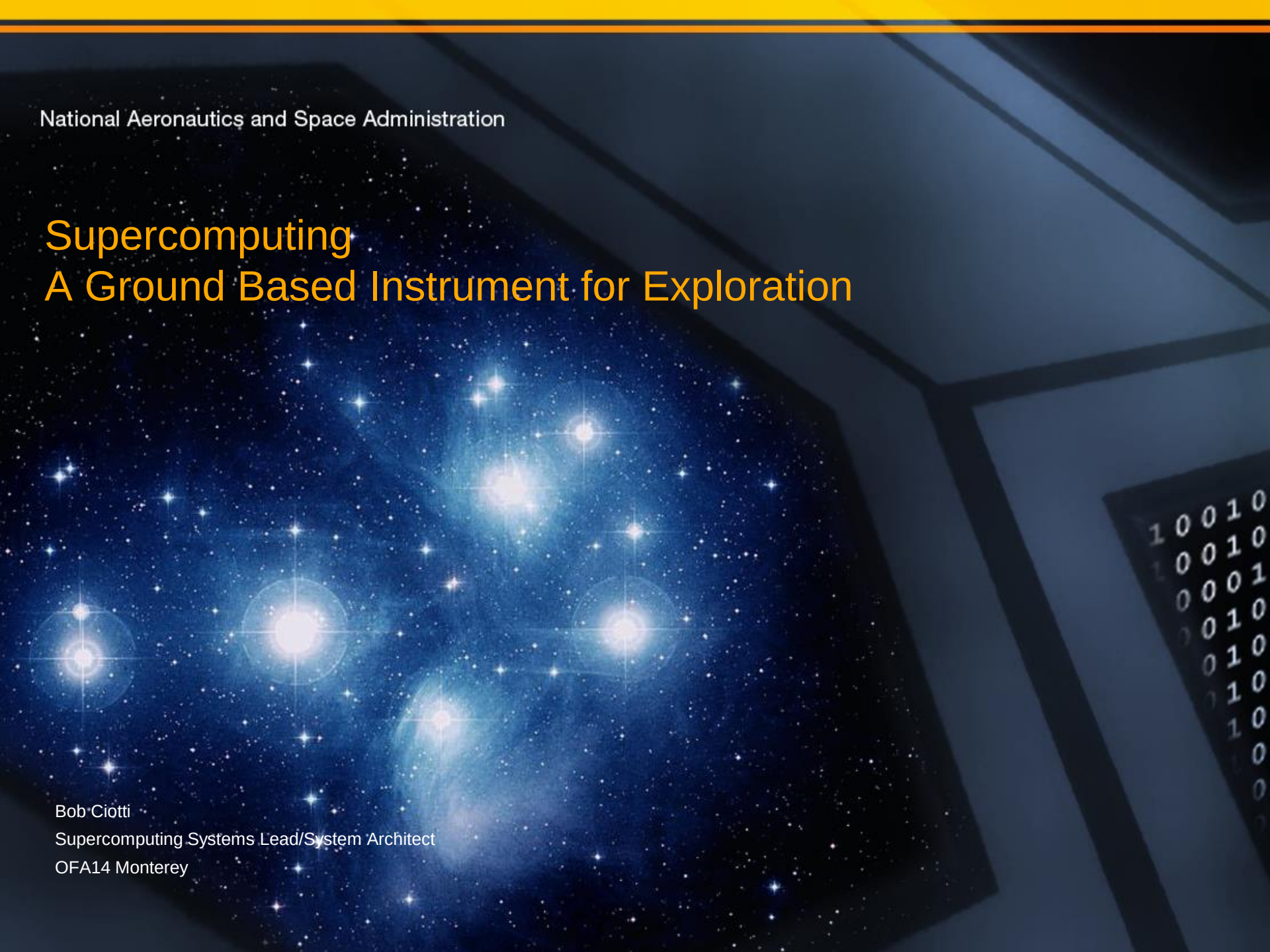
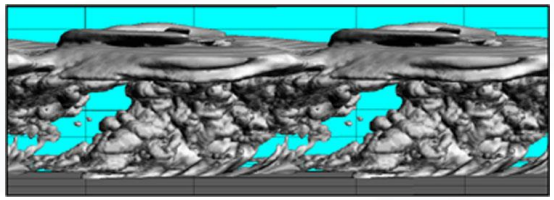

National Aeronautics and Space Administration
Supercomputing
A Ground Based Instrument for Exploration
Bob Ciotti
Supercomputing Systems Lead/System Architect
OFA14 Monterey

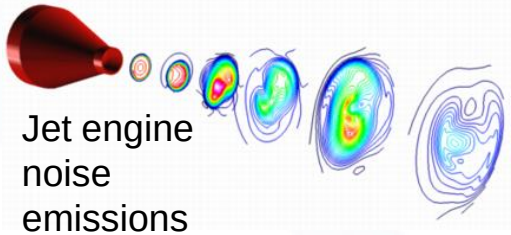
Supercomputing Support for NASA Missions

- 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

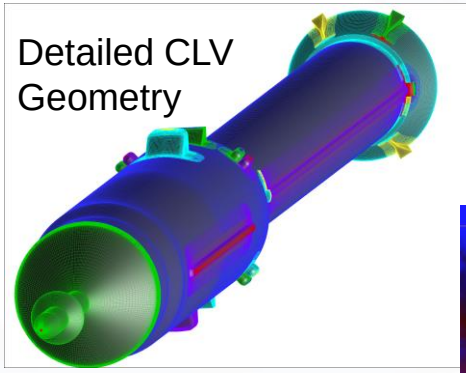
Jet aircraft wake vortices



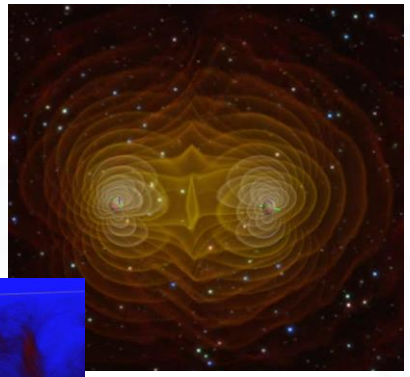


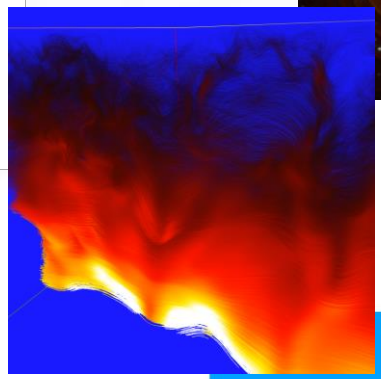
Jet engine noise emissions

Detailed CLV Geometry

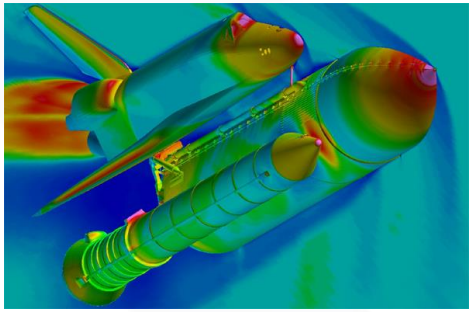


Orbiting, Spinning Black Holes



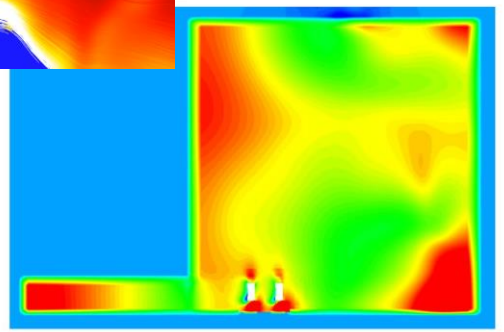


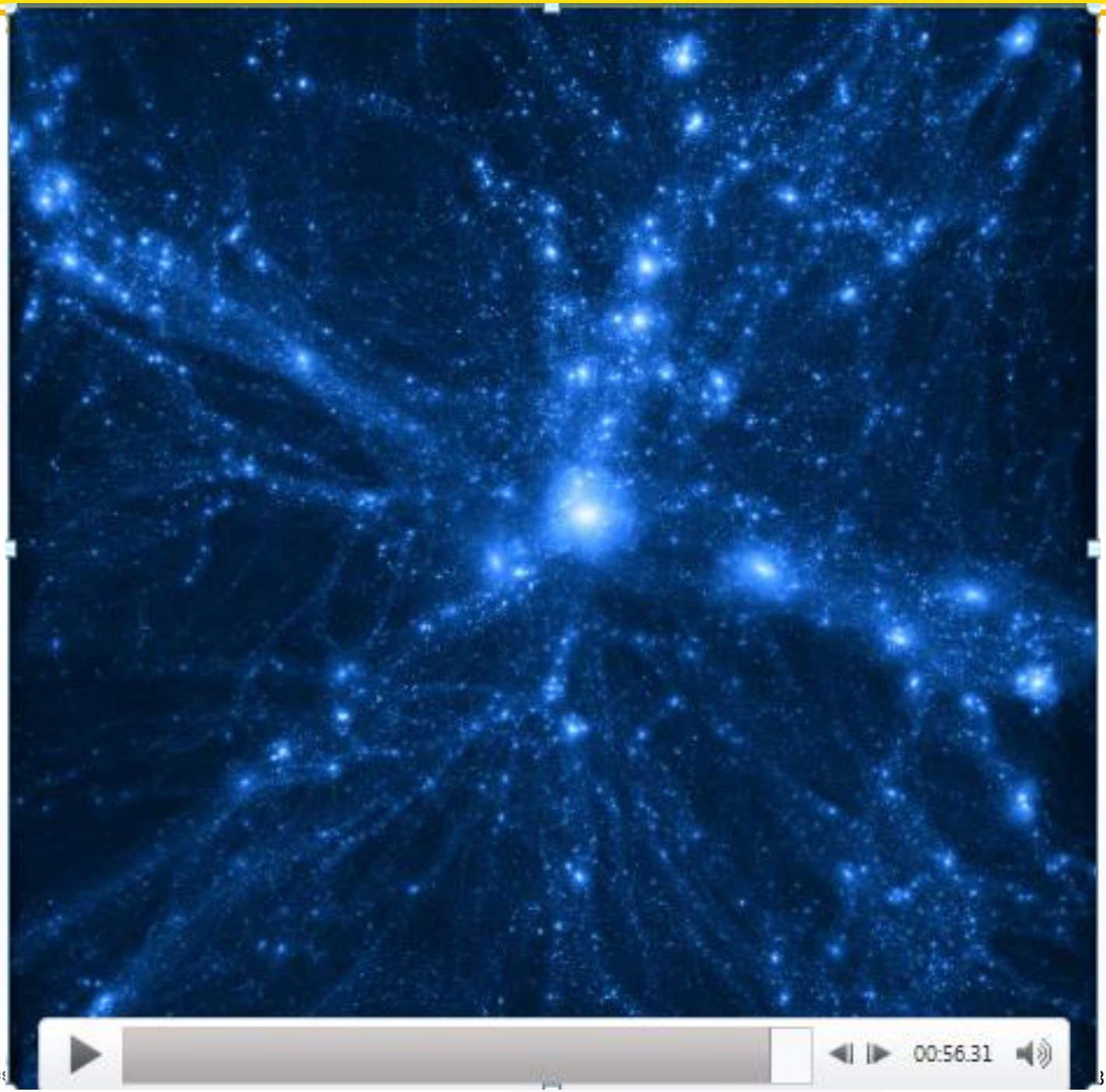
Solar surface convection

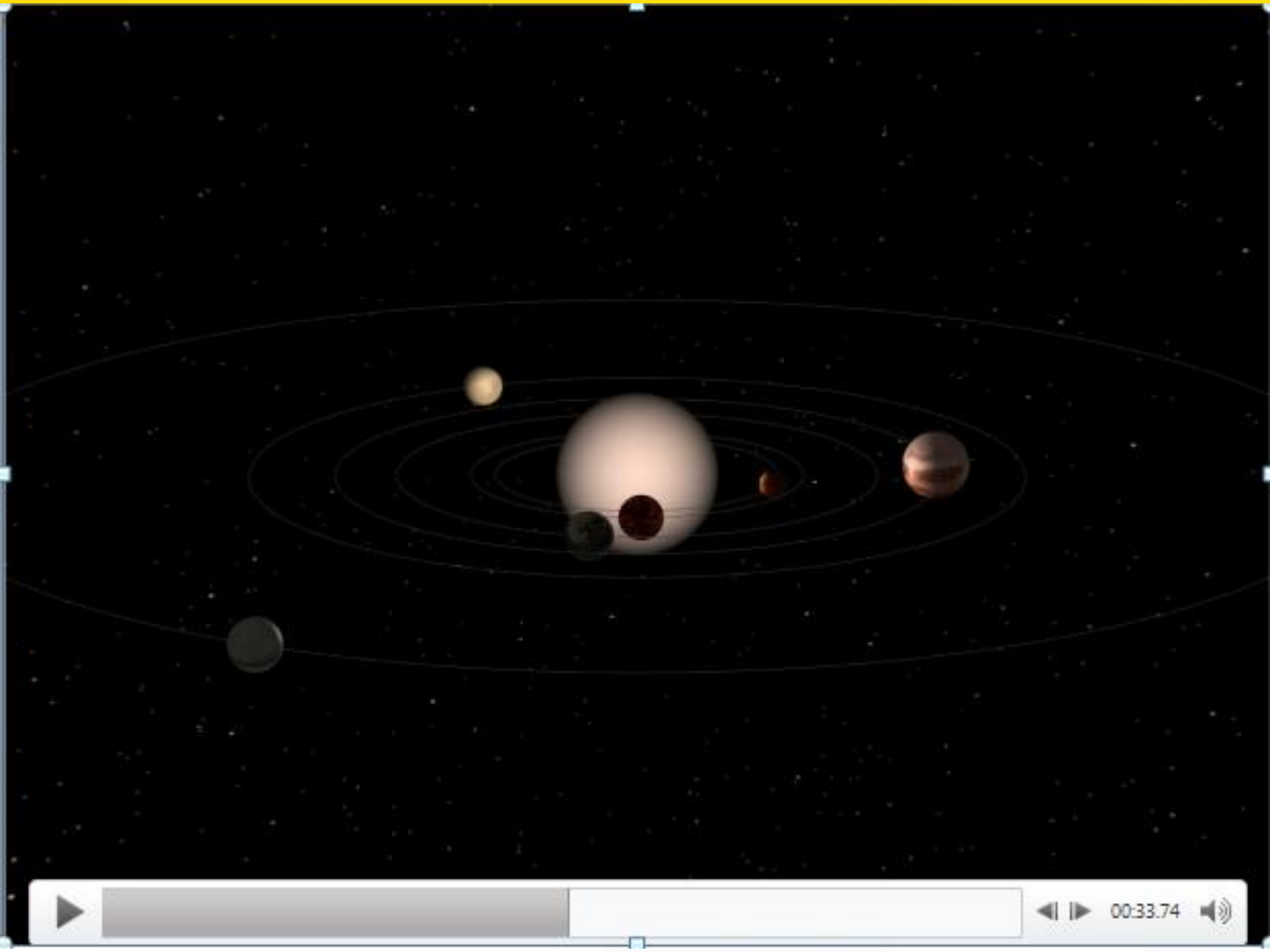


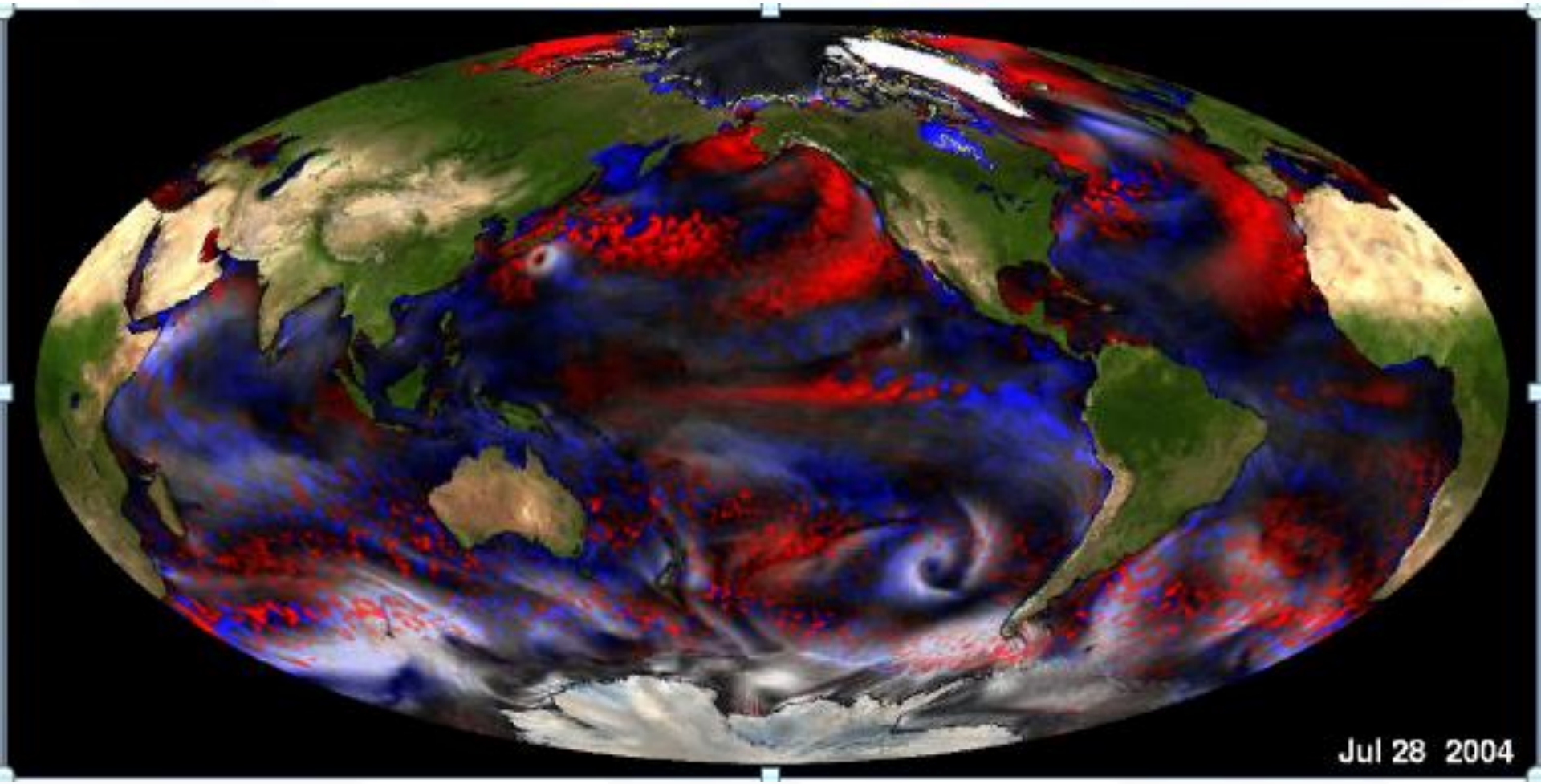
Shuttle Ascent Configuration

2-SRB Burn in VAB

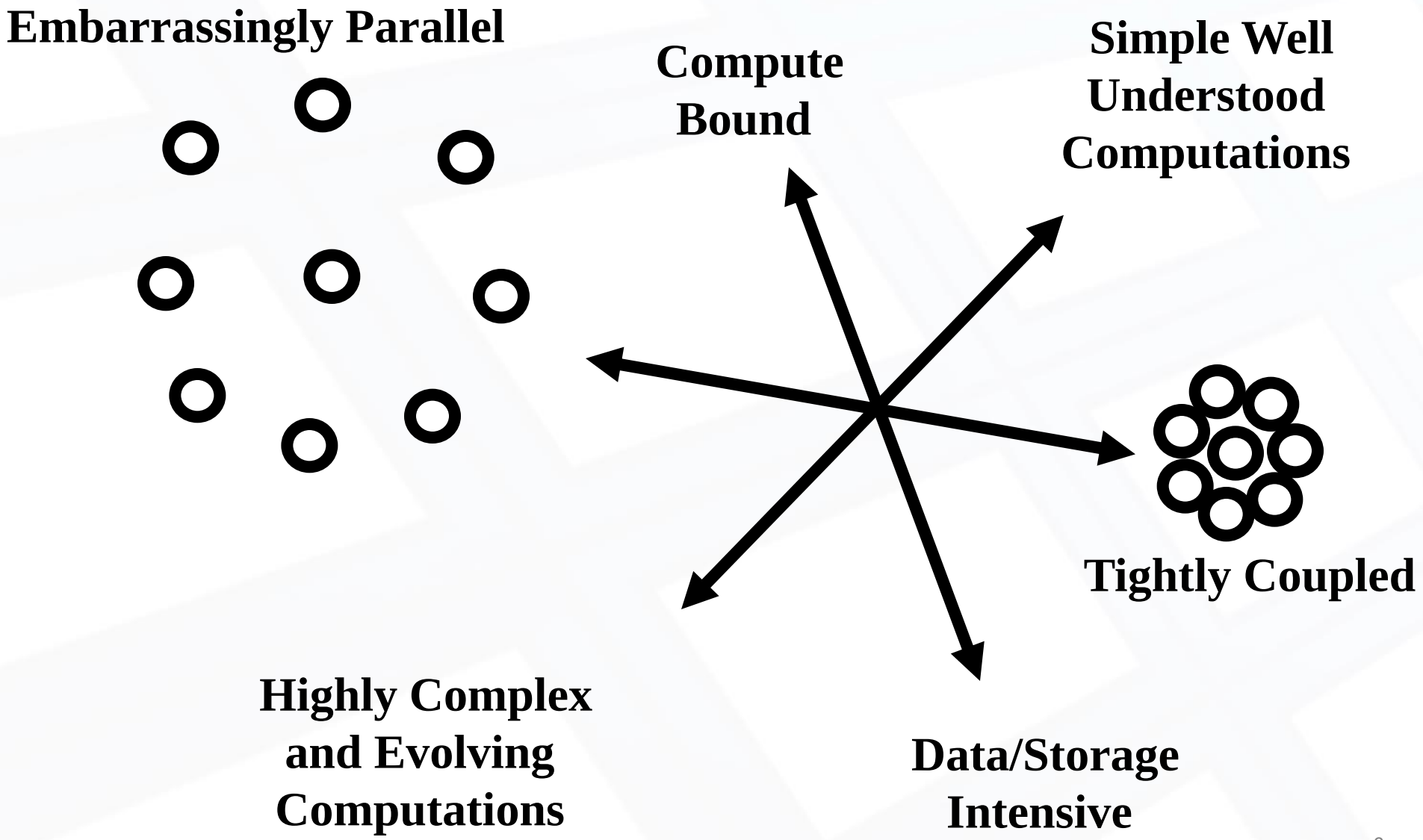



00:56.31


00:33.74

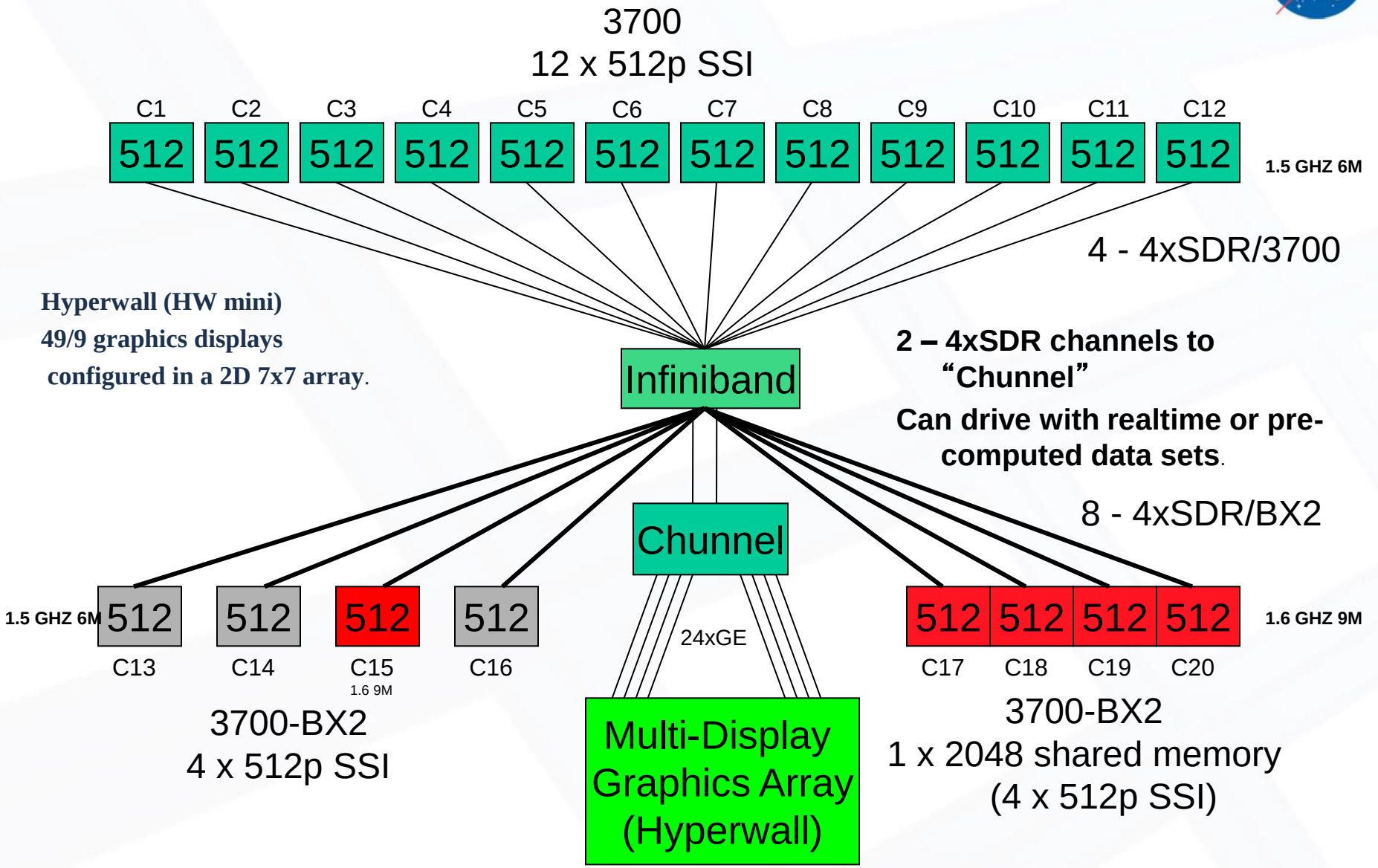

Jul 28 2004

NASA's Computational Landscape



Columbia System - October 2004

Columbia


3700
12 x 512p SSI
C1 C2 C3 C4 C5 C6 C7 C8 C9 C10 C11 C12
512 512 512 512 512 512 512 512 512 512 512 512
1.5 GHZ 6M
4 - 4xSDR/3700
Hyperwall (HW mini)
49/9 graphics displays
configured in a 2D 7x7 array.
Infiniband
2 – 4xSDR channels to "Chunnel"
Can drive with realtime or pre-computed data sets.
8 - 4xSDR/BX2
Chunnel
1.5 GHZ 6M
512 512 512 512
C13 C14 C15 C16
1.6 9M
24xGE
512 512 512 512
C17 C18 C19 C20
1.6 GHZ 9M
3700-BX2
4 x 512p SSI
Multi-Display Graphics Array (Hyperwall)
3700-BX2
1 x 2048 shared memory
(4 x 512p SSI)

OFA/Infiniband Connectivity

Columbia

Pleiades

Ib0
24sdr

Ib0
ddr/qdr

Bridge
Node

<-fc Ib1->

Ib0
288sdr

Ib0
96ddr

Ib0
96ddr

Archive
Server

Ib1
24ddr

longbow

hyperwall

lustre

Oss
Oss
Oss
Oss
Mds
Mds

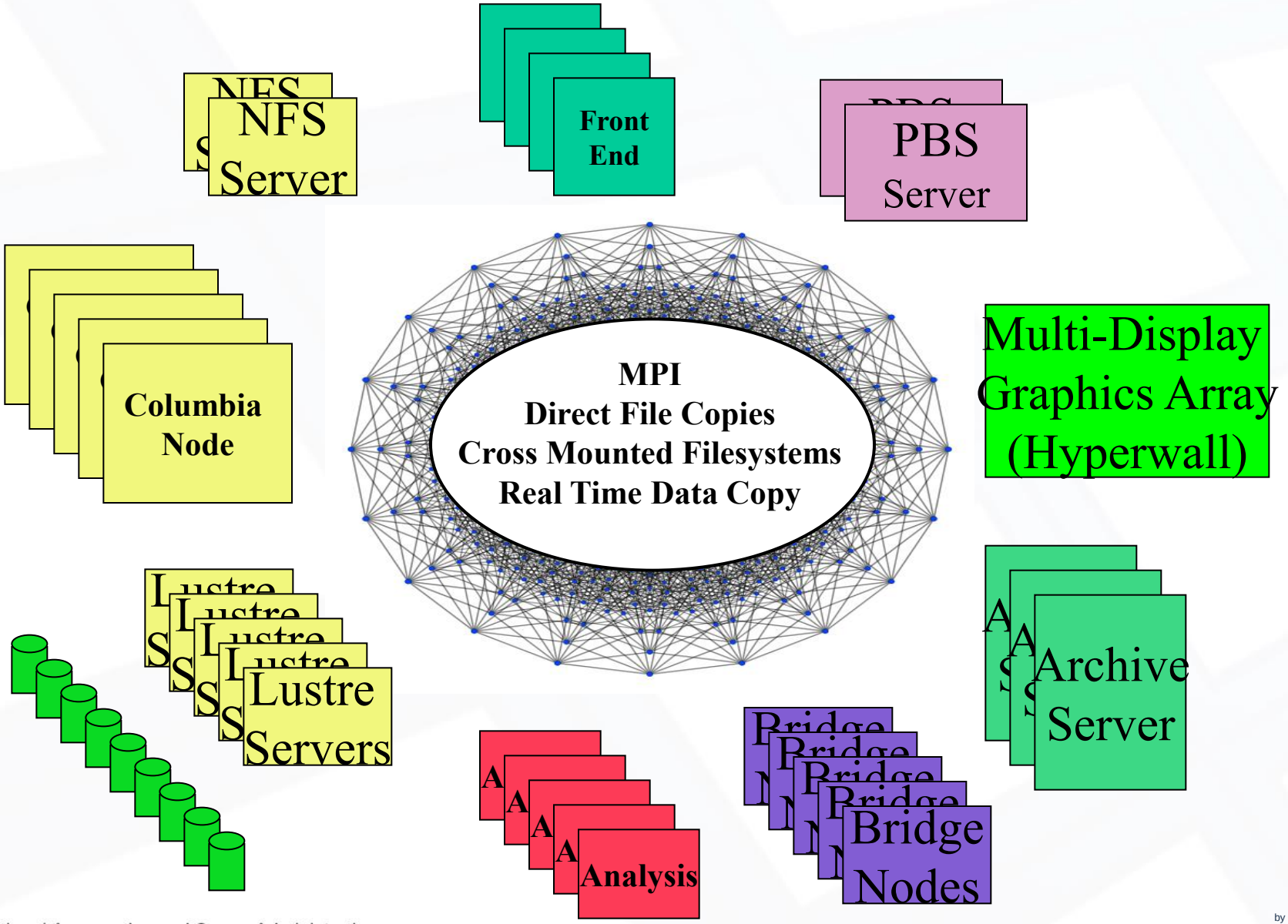
Ib0
288ddr

Ib1
36qdr

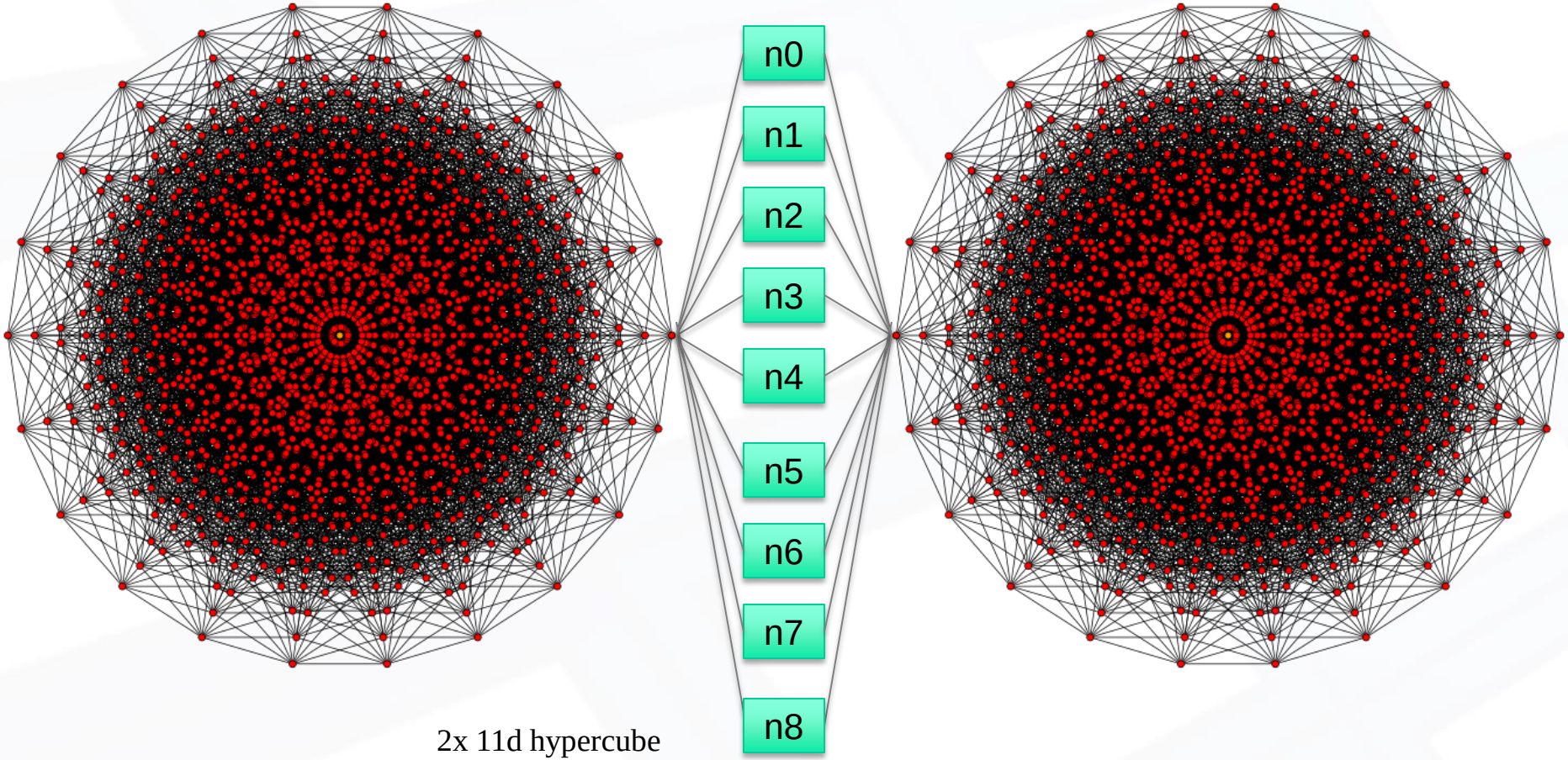
Oss
Oss
Oss
Oss
Mds
Mds

lustre

Architecture Target



SGI ICE Dual Plane – Topology



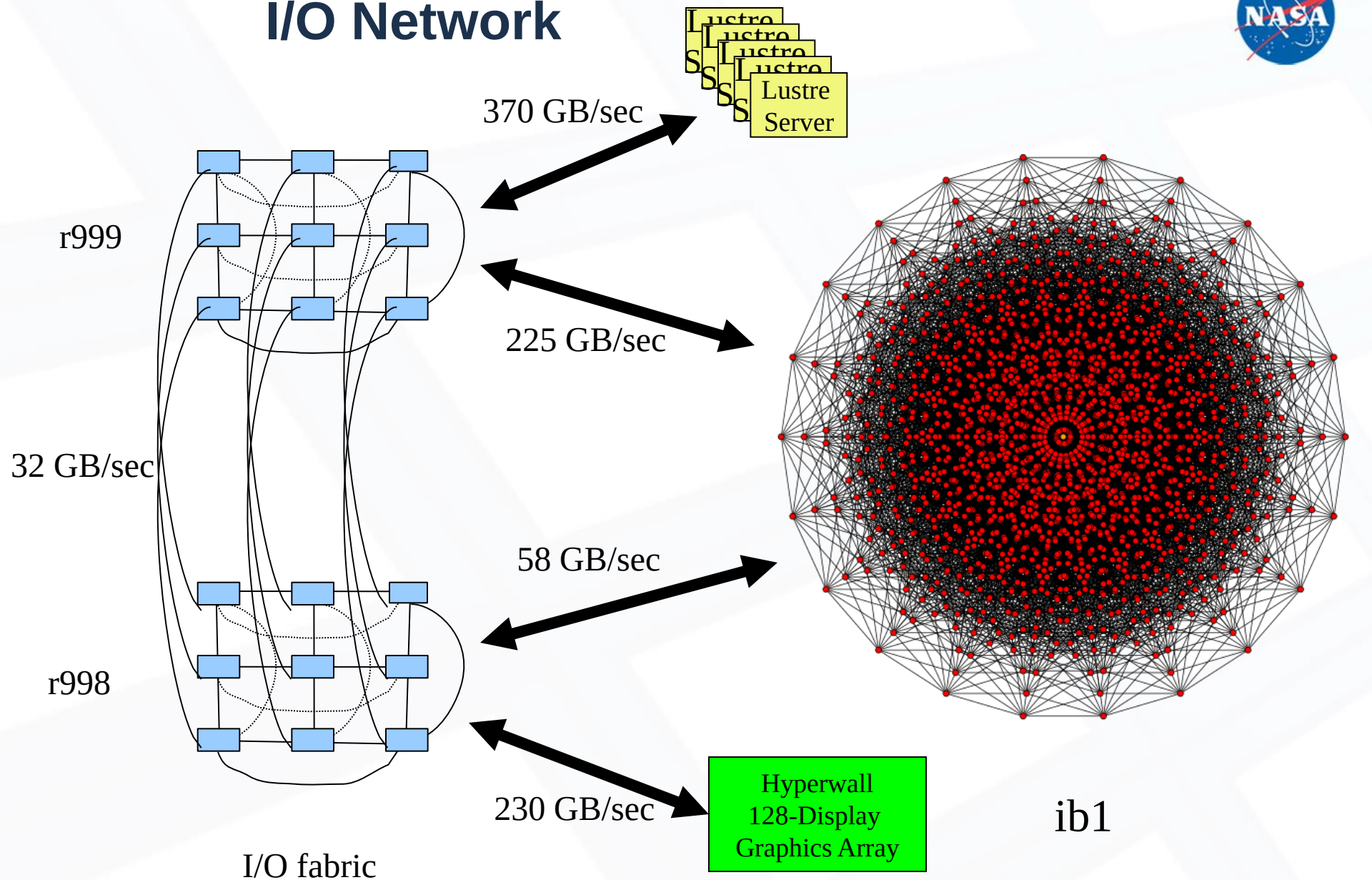
ib0

2x 11d hypercube
full 2048 vertices
Pleiades 1336/11d (2672 across both cubes)

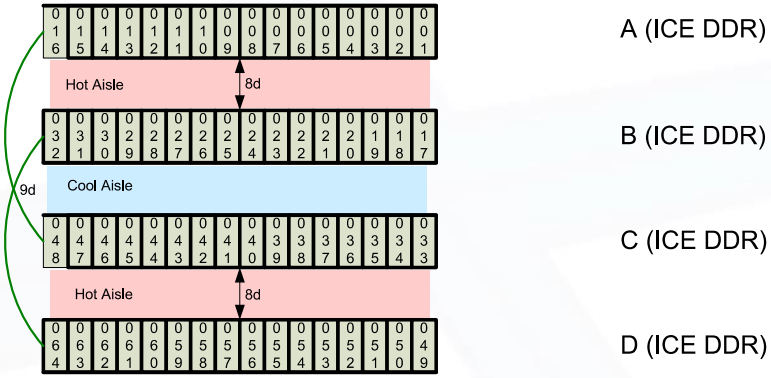
ib1

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

I/O Network

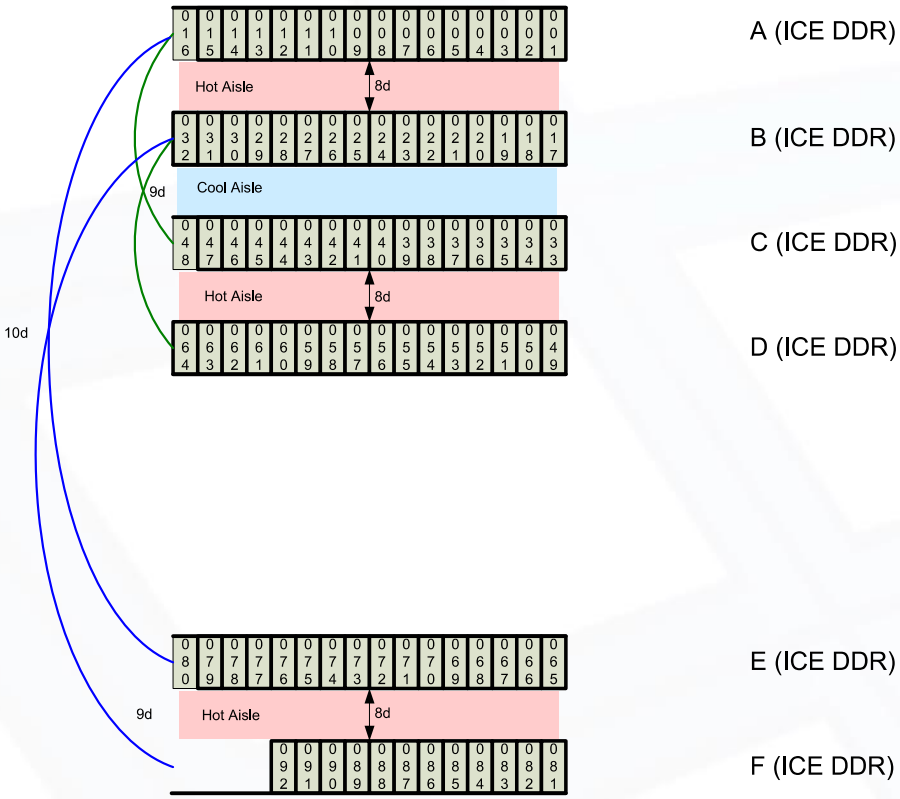

Lustre
Server
370 GB/sec
r999
225 GB/sec
32 GB/sec
58 GB/sec
r998
230 GB/sec
Hyperwall
128-Display
Graphics Array
ib1
I/O fabric

NASA (Pleiades) Rack Layout



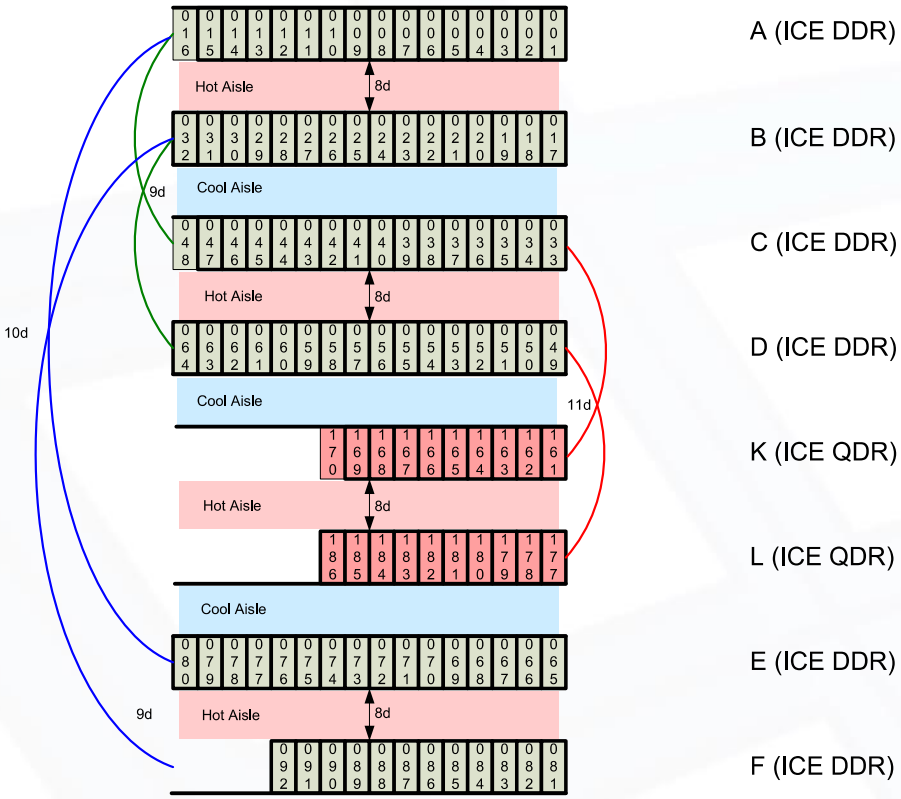
64 racks – 2008
393 teraflops

NASA (Pleiades) Rack Layout



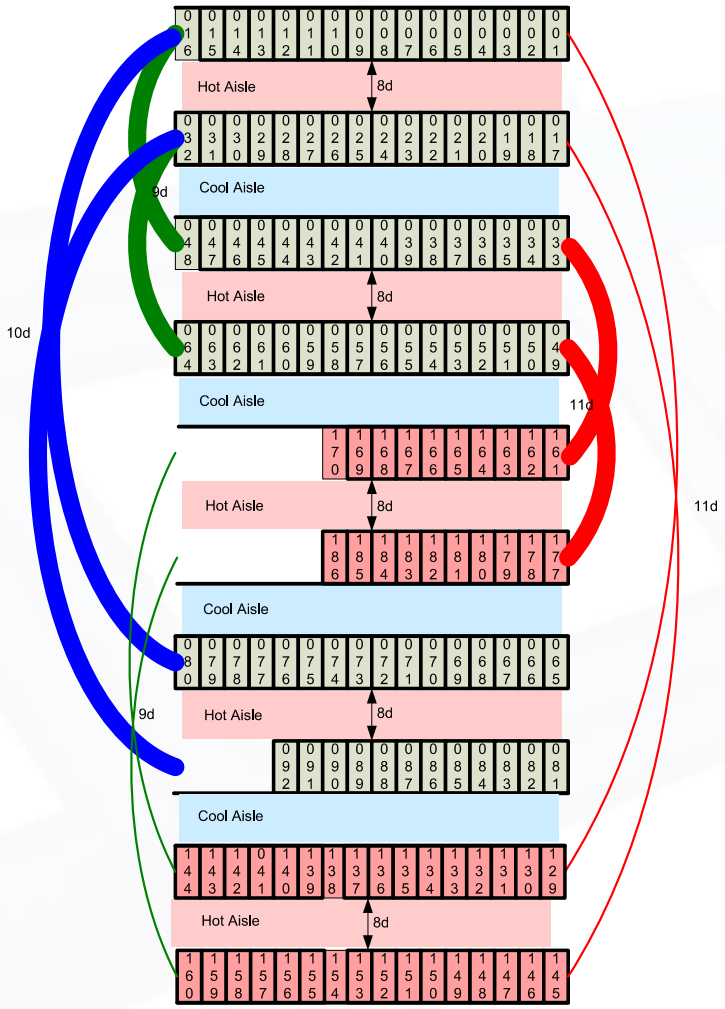
92 racks – 2008
565 teraflops

NASA (Pleiades) Rack Layout



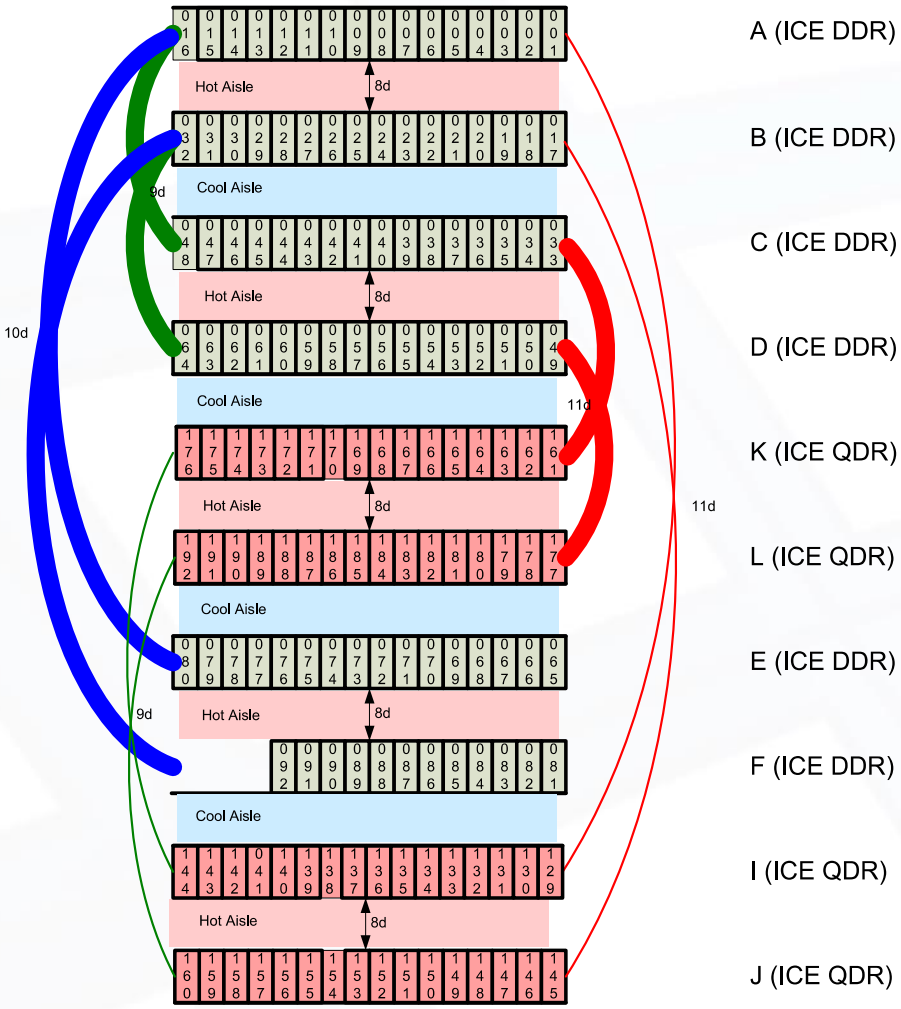
112 racks – 2009
683 teraflops

NASA (Pleiades) Rack Layout

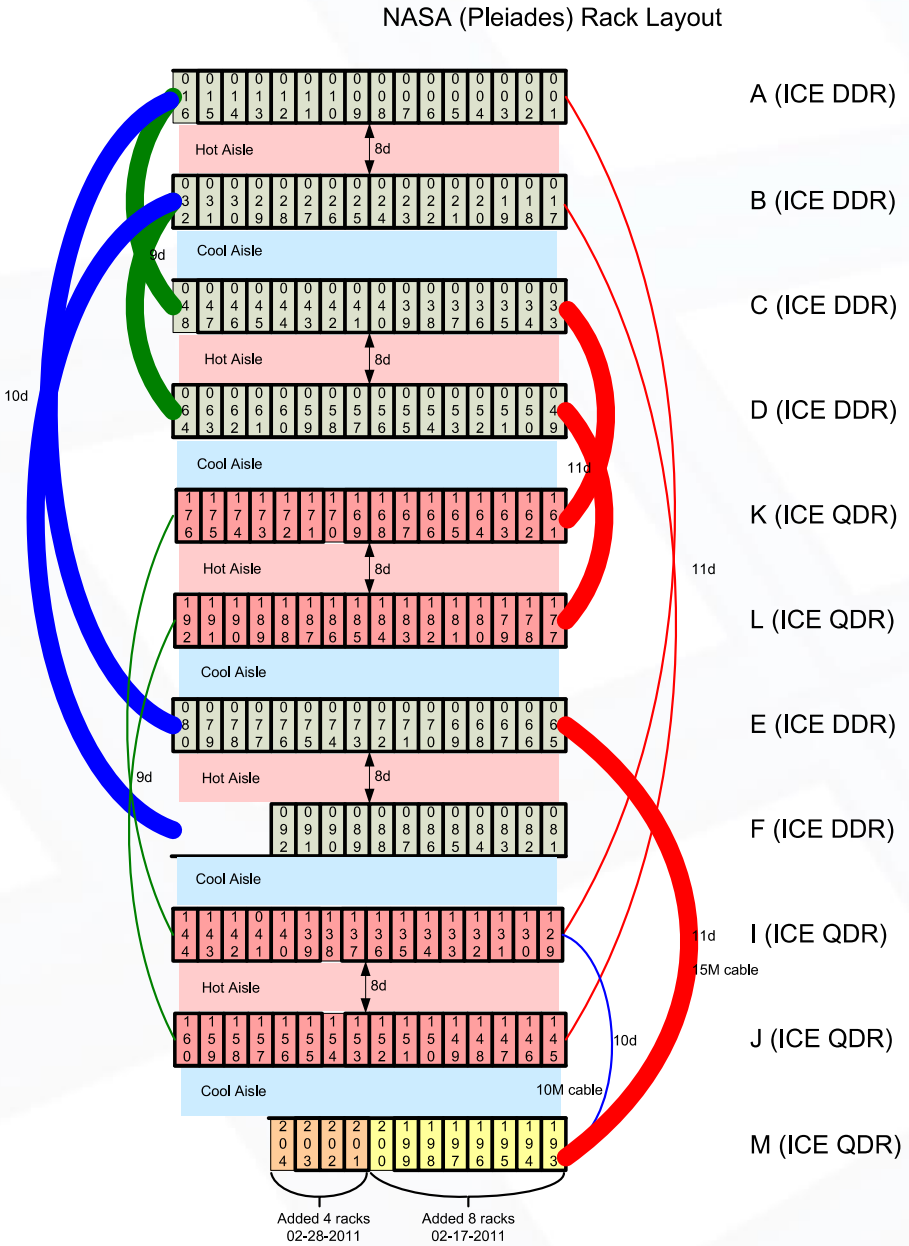


144 racks – 2010
969 teraflops

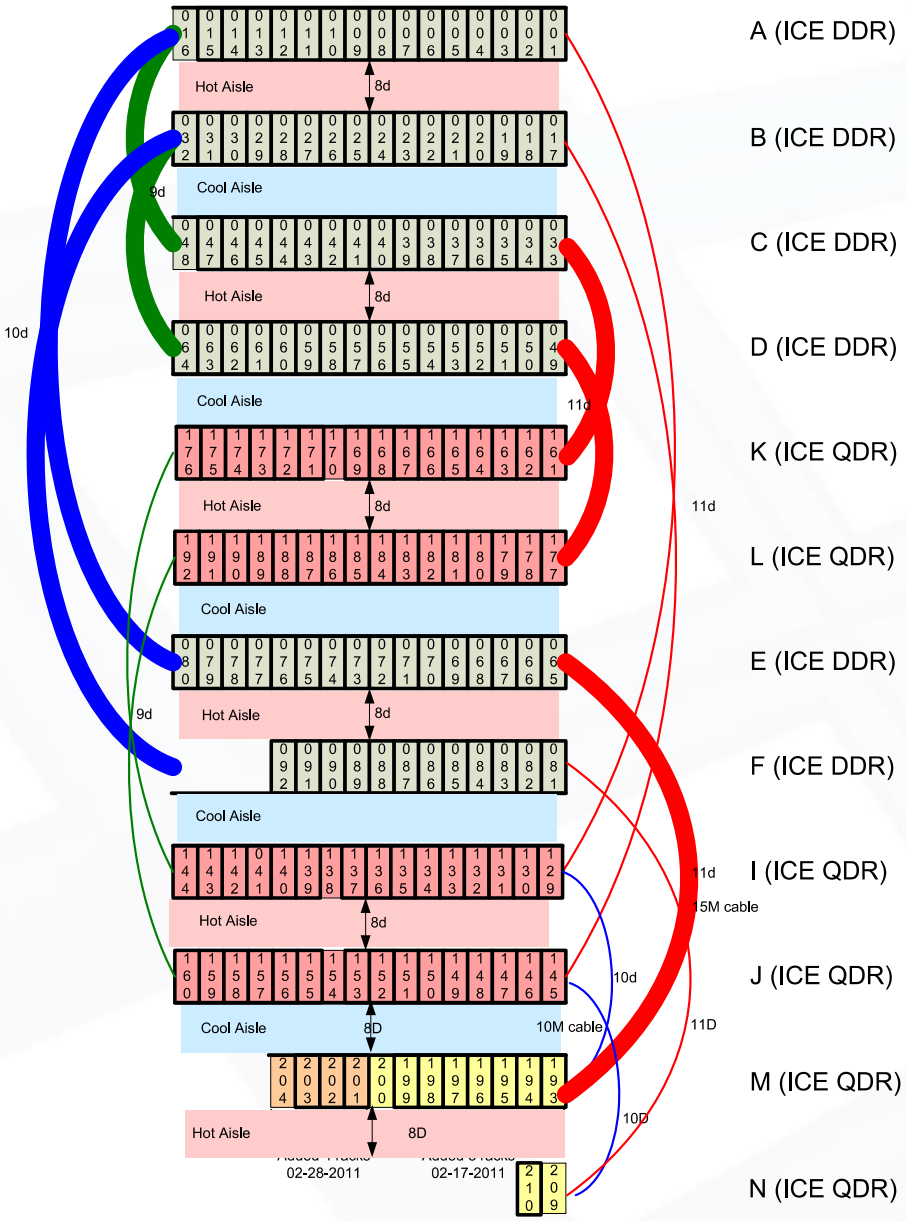
NASA (Pleiades) Rack Layout



156 racks – 2010
1.08 petaflops



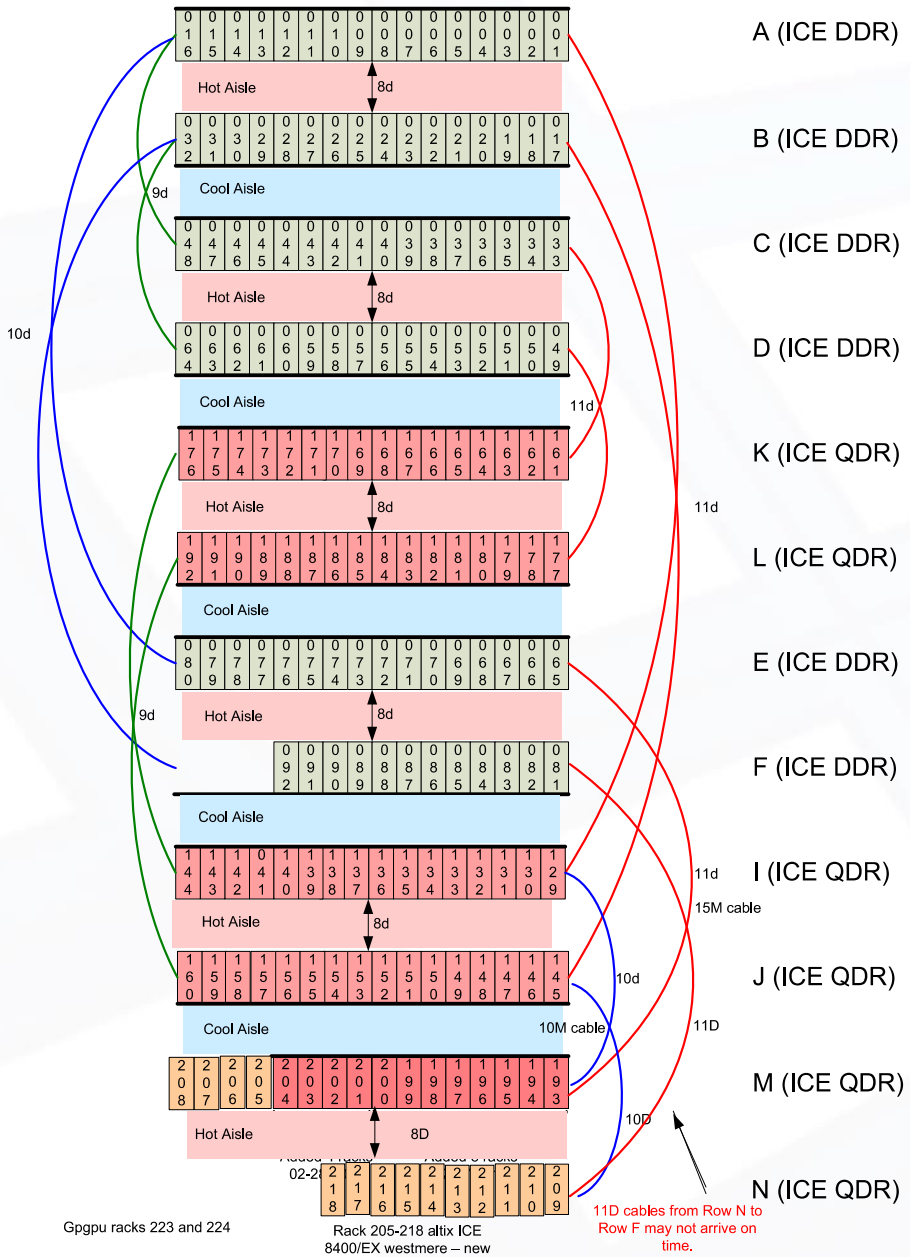
168 racks – 2011
1.18 petaflops



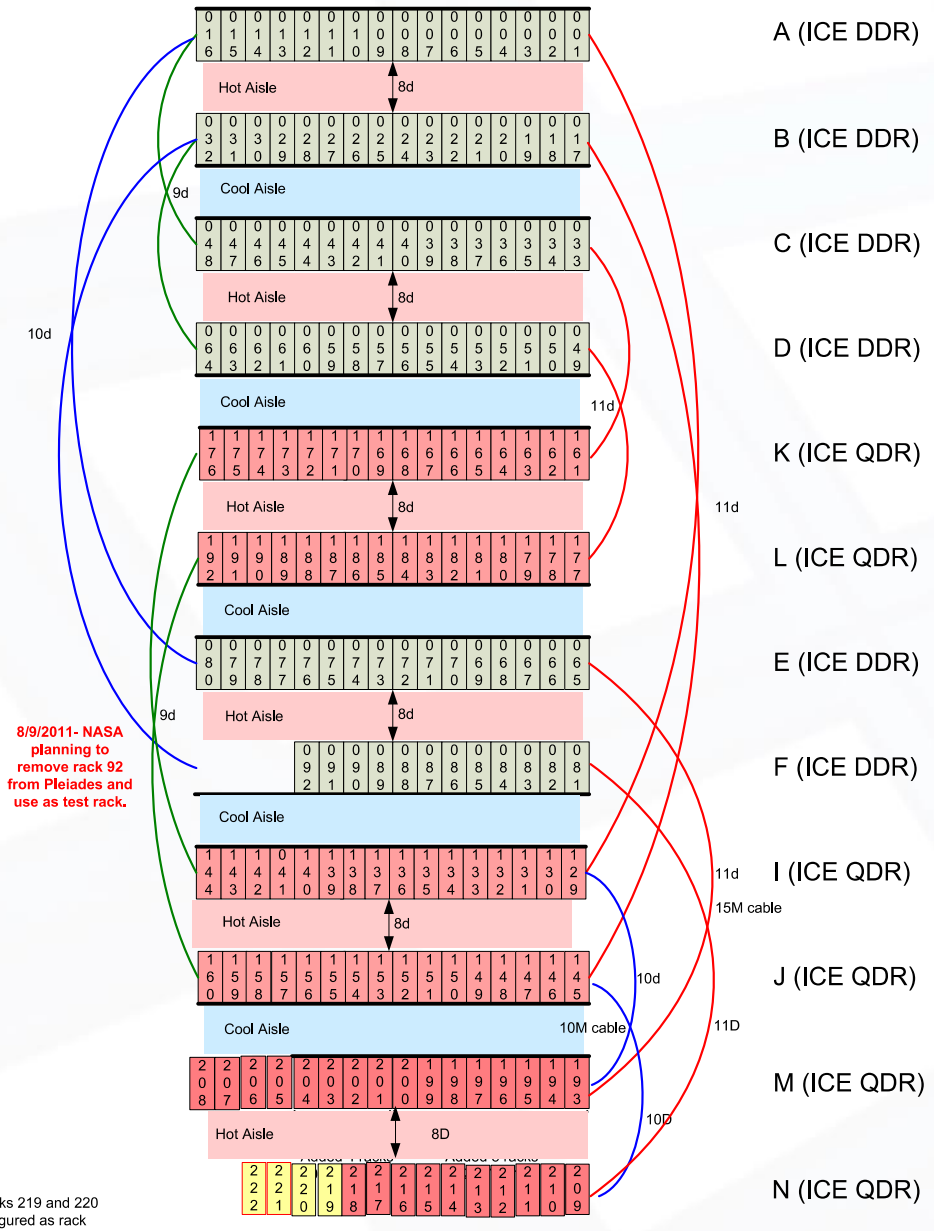
170 racks – 2011
1.20 petaflops

NASA (Pleiades) Rack Layout

182 racks – 2011
1.31 petaflops



NASA (Pleiades) Rack Layout

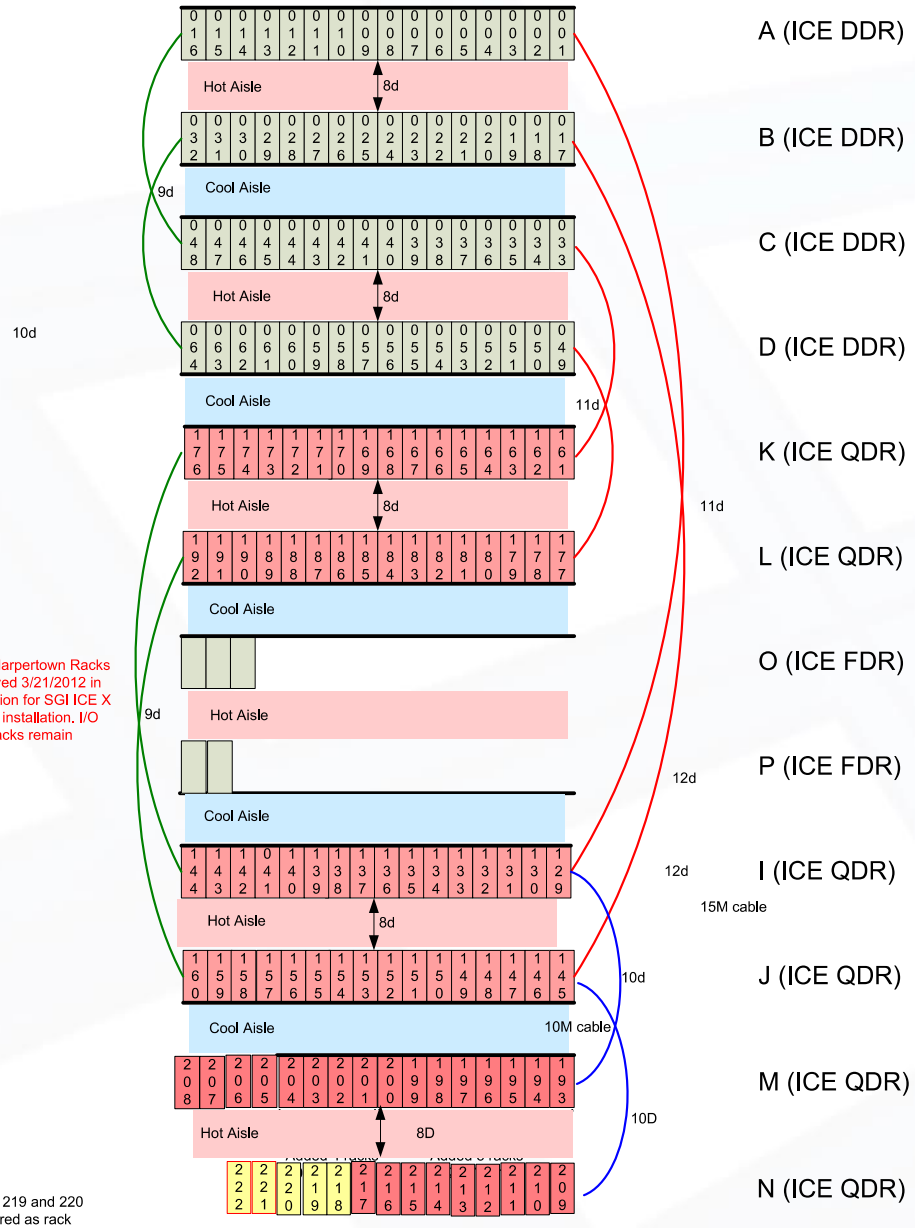


186 racks – 2011
1.33 petaflops

Gpgpu racks 219 and 220 but configured as rack 219. note switches on gpgpu are in rear of rack so cable lengths needs to be adjusted to reflect this.

Note: Rack 221 will cable to on 11D to rack 92. There is no 11d for Rack 222. this is a problem. If we remove rack 92 then we have issue with racks 221 & 222.

NASA (Pleiades) Rack Layout

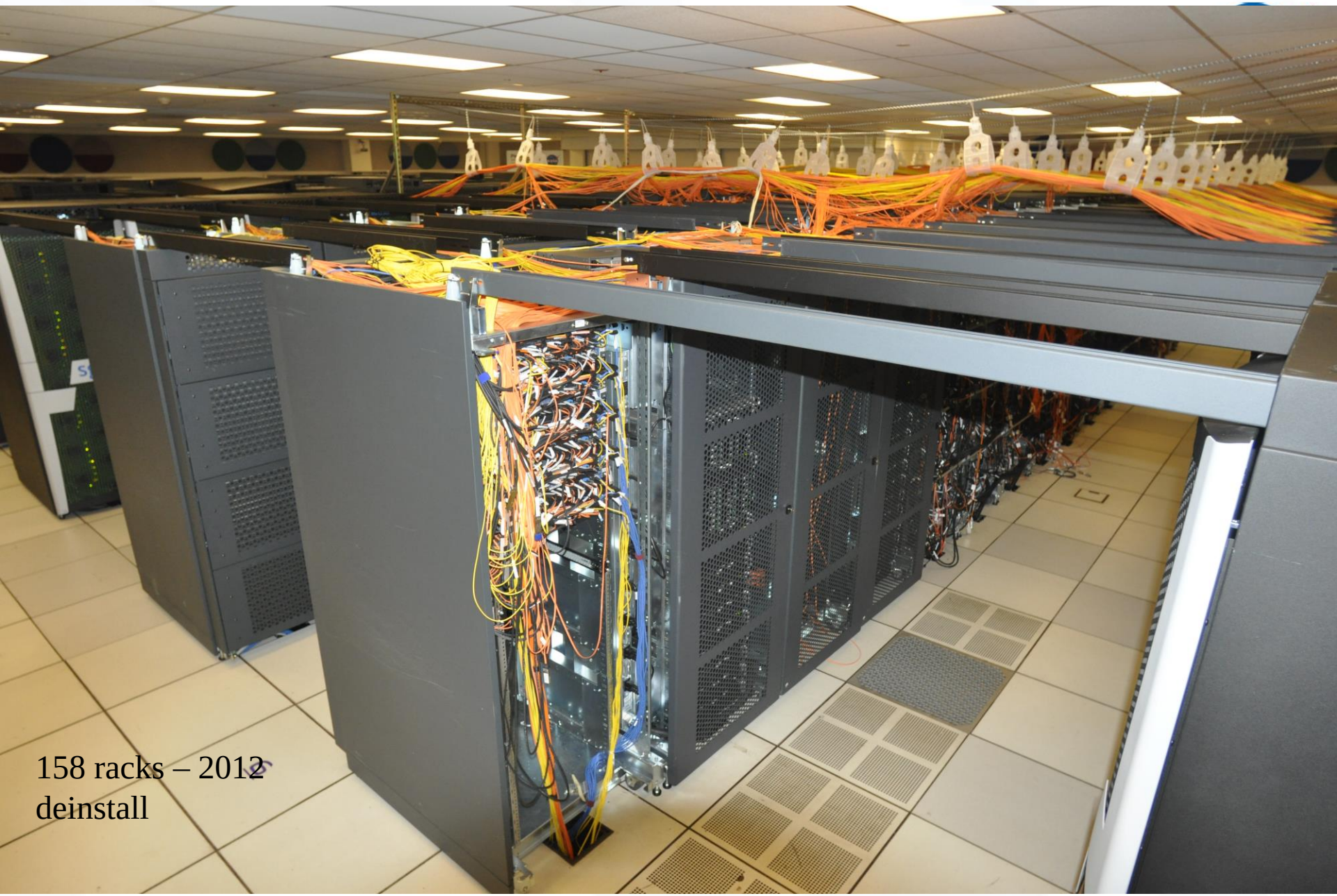


158 racks – 2012
1.15 petaflops
deinstall

*Note: Harpertown Racks Removed 3/21/2012 in preparation for SGI ICE X Racks installation. I/O Racks remain

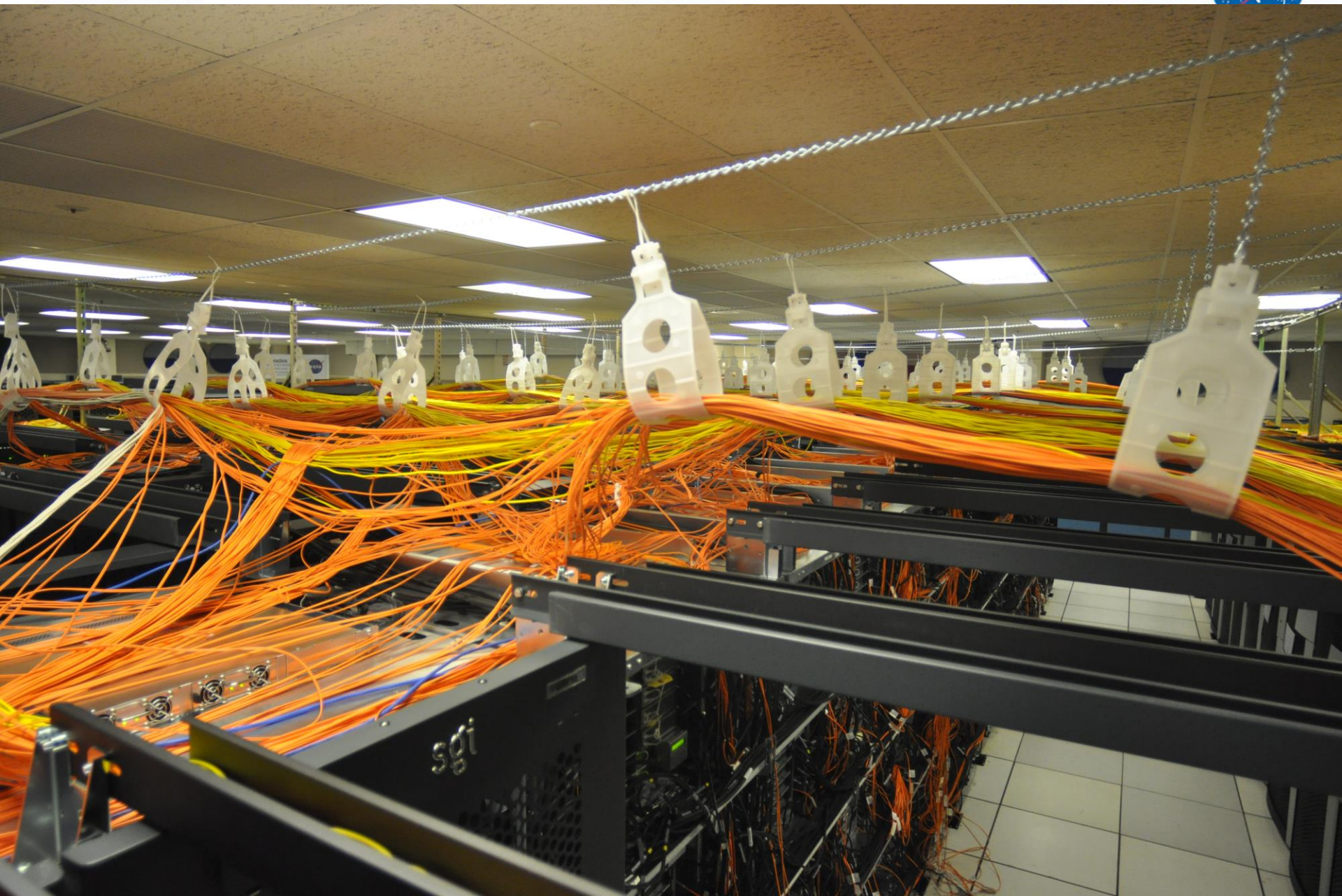
Gpgpu racks 219 and 220 but configured as rack 219. note switches on gpgpu are in rear of rack so cable lengths needs to be adjusted to reflect this.

Note: Rack 221 will cable to on 11D to rack 92. There is no 11d for Rack 222. this is a problem. If we remove rack 92 then we have issue with racks 221 & 222.

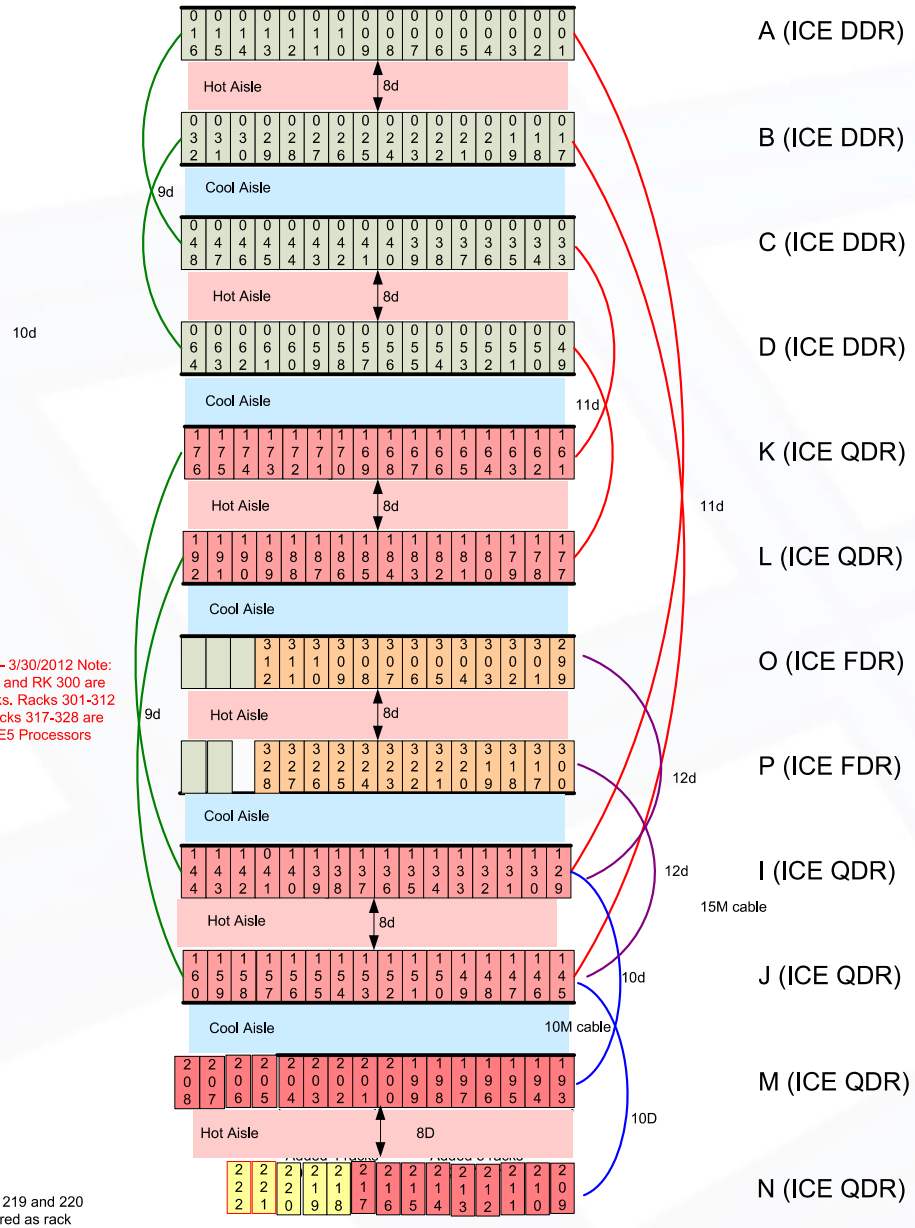

158 racks – 2012
deinstall


158 racks – 2012
deinstall

158 racks – 2012
deinstall



NASA (Pleiades) Rack Layout



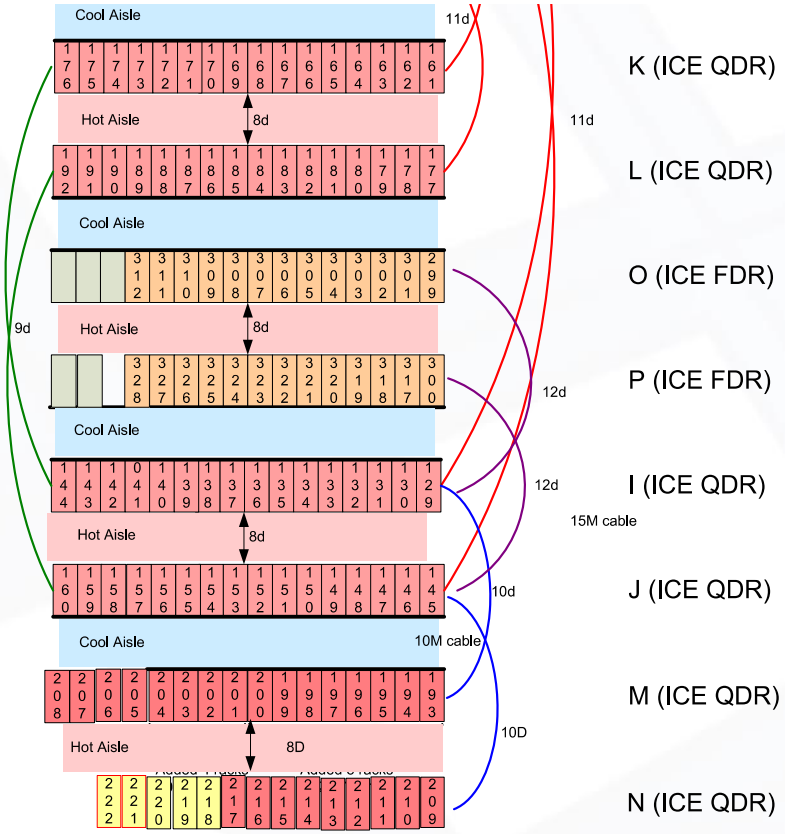
182 racks – 2012
1.7 petaflops

* Install – 3/30/2012 Note: RK 299 and RK 300 are RLC racks. Racks 301-312 and Racks 317-328 are Intel E5 Processors

Gpgpu racks 219 and 220 but configured as rack 219. note switches on gpgpu are in rear of rack so cable lengths needs to be adjusted to reflect this.

Note: Rack 221 will cable to on 11D to rack 92. There is no 11d for Rack 222. this is a problem. If we remove rack 92 then we have issue with racks 221 & 222.

64 rack deinstall 2013

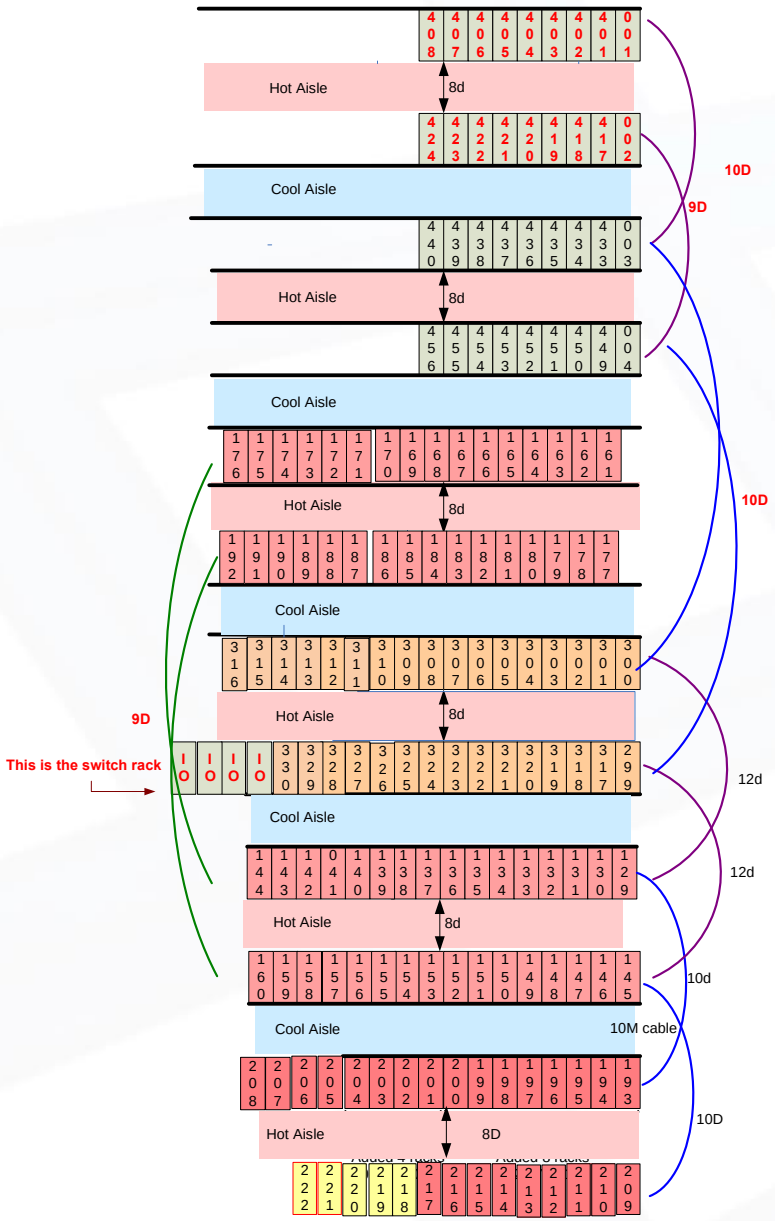


* Install – 3/30/2012 Note: RK 299 and RK 300 are RLC racks. Racks 301-312 and Racks 317-328 are Intel E5 Processors

Gpgpu racks 219 and 220 but configured as rack 219. note switches on gpgpu are in rear of rack so cable lengths needs to be adjusted to reflect this.

Note: Rack 221 will cable to on 11D to rack 92. There is no 11d for Rack 222. this is a problem. If we remove rack 92 then we have issue with racks 221 & 222.

NASA (Pleiades) Rack Layout as of 08/05/2013



A (ICE FDR)
RK 401-408 -SGI ICE X (ivybridge) Prem SW

B (ICE FDR)
RK417-424 SGI ICE X (ivybridge) Prem SW

C (ICE FDR)
RK 433-440 SGI ICE X (ivybridge) prem Sw

D (ICE FDR) Rk 449-464 SGI ICE X (ivybridge) Prem SW

K (ICE QDR) rks 161-170 Altix ICE 8200 Nehamel + ICE QDR) 171-176 Altix ICE 8400 EX (Westmere)

L (ICE QDR) rks 177-186 Altix ICE 8200 Nehalem +(ICE QDR) 187-17192 Altix ICE 8400 EX (Westmere)

O (ICE FDR)/ RK 300-312 SGI ICE X Sandybridge Prem SW/ RK 313-316 128 node Pyramid in hypercube topology

P (ICE FDR)/ RK 317-330 SGI ICE X SNB Prem SW

I (ICE QDR) RKS 129-144 Altix Ice 8400 EX

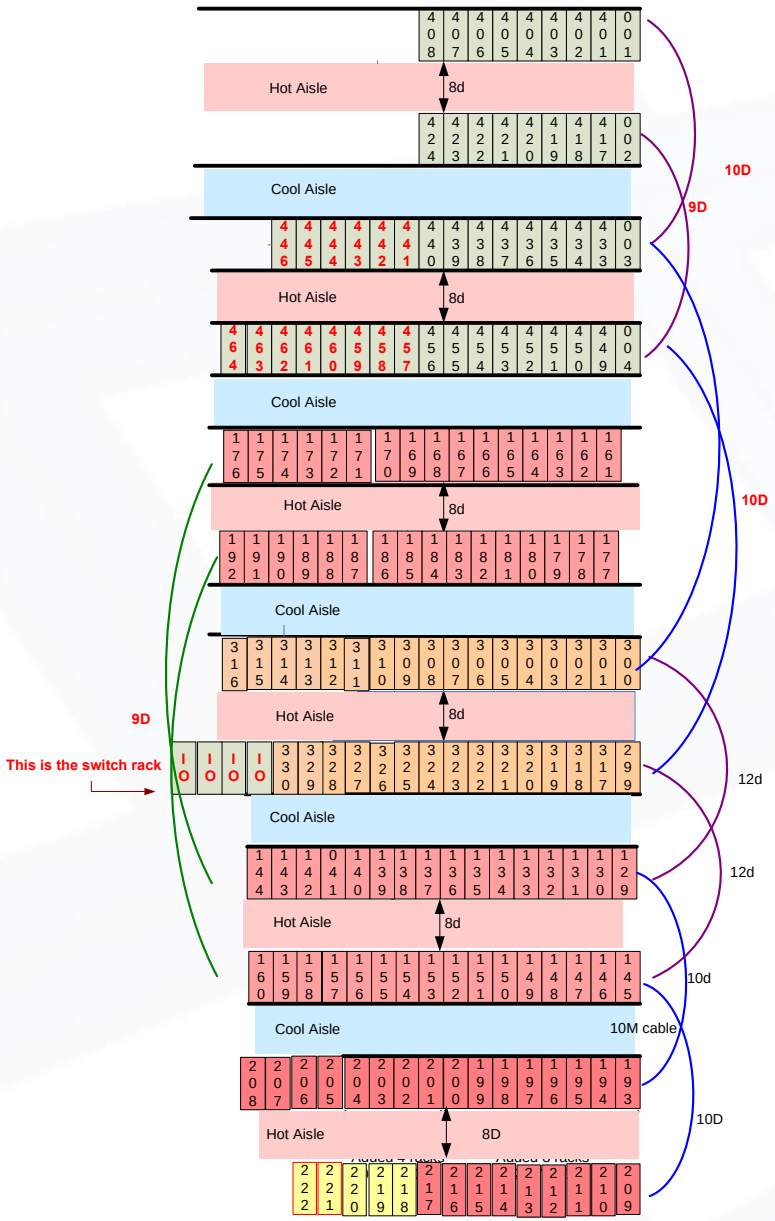
J (ICE QDR) RKS 145-160 ALTIX ICE 8400 EX

M (ICE QDR) RKS 193-208 ALTIX ICE 8400 EX

N (ICE QDR) RKS 209-218 ALTIX ICE 8400 EX/ RK 219-220 Coyote based Westmere with GPGPU M2090 in hypercube. RK 221-222 Altix ICE 8400 EX

NASA (Pleiades) Rack Layout as of 08/12/2013

167 racks – 2013
2.9 petaflops



A (ICE FDR)
RK 401-408 -SGI ICE X (ivybridge) Prem SW

B (ICE FDR)
RK417-424 SGI ICE X (ivybridge) Prem SW

C (ICE FDR)
RK 433-446 SGI ICE X (ivybridge) prem Sw

D (ICE FDR) Rk 449-464 SGI ICE X (ivybridge) Prem SW

K (ICE QDR) rks 161-170 Altix ICE 8200 Nehamel + ICE QDR) 171-176 Altix ICE 8400 EX (Westmere)

L (ICE QDR) rks 177-186 Altix ICE 8200 Nehalem +(ICE QDR) 187-17192 Altix ICE 8400 EX (Westmere)

O (ICE FDR)/ RK 300-312 SGI ICE X Sandybridge Prem SW/ RK 313-316 128 node Pyramid in hypercube topology

P (ICE FDR)/ RK 317-330 SGI ICE X SNB Prem SW

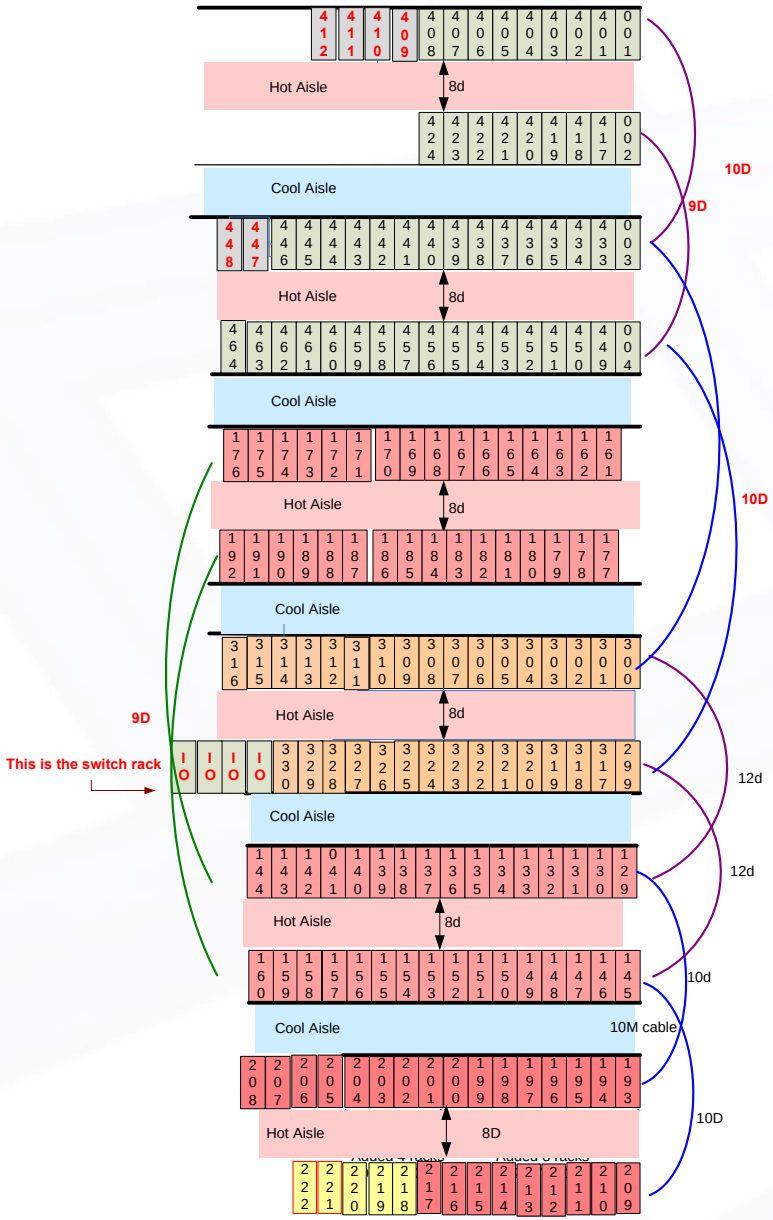
I (ICE QDR) RKS 129-144 Altix Ice 8400 EX

J (ICE QDR) RKS 145-160 ALTIX ICE 8400 EX

M (ICE QDR) RKS 193-208 ALTIX ICE 8400 EX

N (ICE QDR) RKS 209-218 ALTIX ICE 8400 EX/ RK 219-220 Coyote based Westmere with GPGPU M2090 in hypercube. RK 221-222 Altix ICE 8400 EX

NASA (Pleiades) Rack Layout as of 12/30/2013



A (ICE FDR)
RK 401-416 -SGI ICE X (ivybridge) Prem SW

B (ICE FDR)
RK417-432 SGI ICE X (ivybridge) Prem SW

C (ICE FDR)
RK 433-448 SGI ICE X (ivybridge) prem Sw

D (ICE FDR) Rk 449-464 SGI ICE X (ivybridge) Prem SW

K (ICE QDR) rks 161-170 Altix ICE 8200 Nehamel + ICE QDR) 171-176 Altix ICE 8400 EX (Westmere)

L (ICE QDR) rks 177-186 Altix ICE 8200 Nehalem +(ICE QDR) 187-17192 Altix ICE 8400 EX (Westmere)

O (ICE FDR)/ RK 300-312 SGI ICE X Sandybridge Prem SW/ RK 313-316 128 node Pyramid in hypercube topology

P (ICE FDR)/ RK 317-330 SGI ICE X SNB Prem SW

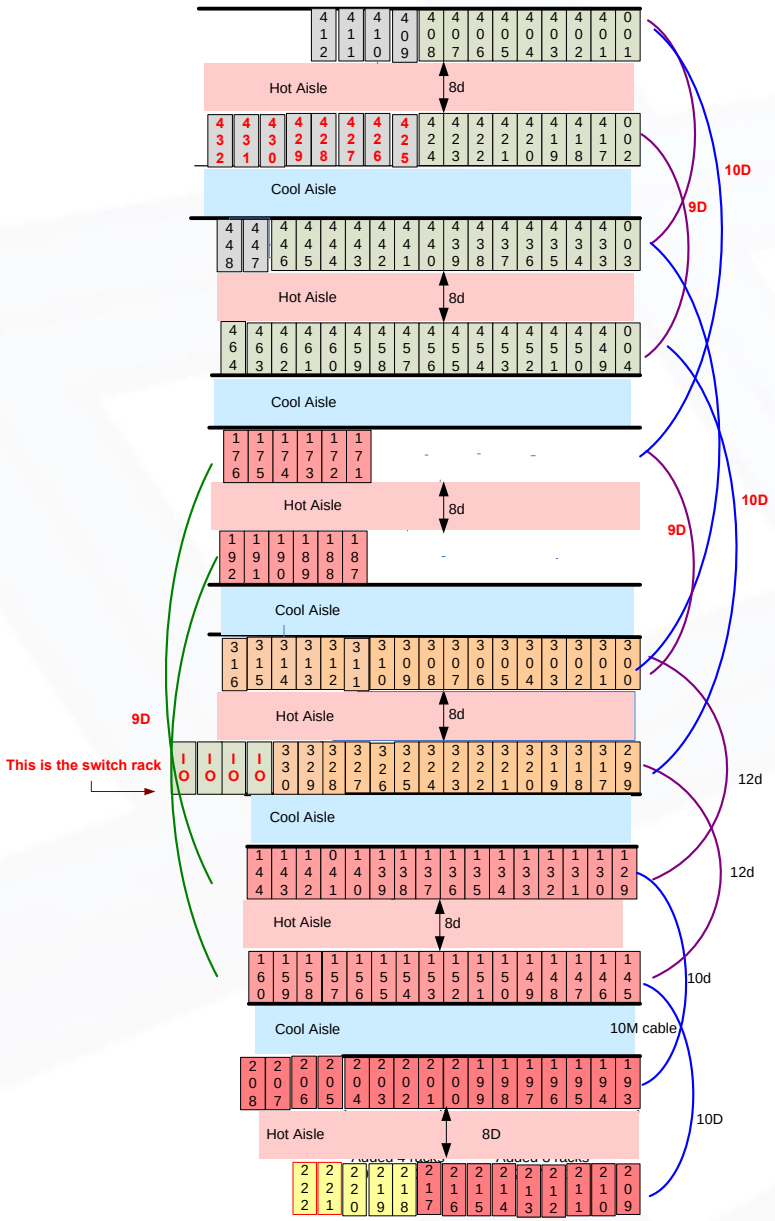
I (ICE QDR) RKS 129-144 Altix Ice 8400 EX

J (ICE QDR) RKS 145-160 ALTIX ICE 8400 EX

M (ICE QDR) RKS 193-208 ALTIX ICE 8400 EX

N (ICE QDR) RKS 209-218 ALTIX ICE 8400 EX/ RK 219-220 Coyote based Westmere with GPGPU M2090 in hypercube. RK 221-222 Altix ICE 8400 EX

NASA (Pleiades) Rack Layout as of 1/30/2014



A (ICE FDR)
RK 401-412 -SGI ICE X (ivybridge) Prem SW

B (ICE FDR)
RK417-432 SGI ICE X (ivybridge) Prem SW

C (ICE FDR)
RK 433-448 SGI ICE X (ivybridge) prem Sw

D (ICE FDR) Rk 449-464 SGI ICE X (ivybridge) Prem SW

E (ICE FDR) rks 465-468 SGI ICE X Ivybridge Prem SW – K ICE QDR) 171-176 Altix ICE 8400 EX

F (ICE FDR) rks 481-483 SGI ICE X Ivybrodge Prem SW L-ICE QDR) 187-17192 Altix ICE 8400 EX

O (ICE FDR)/ RK 300-312 SGI ICE X Sandybridge Prem SW/ RK 313-316 128 node Pyramid in hypercube topology

P (ICE FDR)/ RK 317-330 SGI ICE X SNB Prem SW

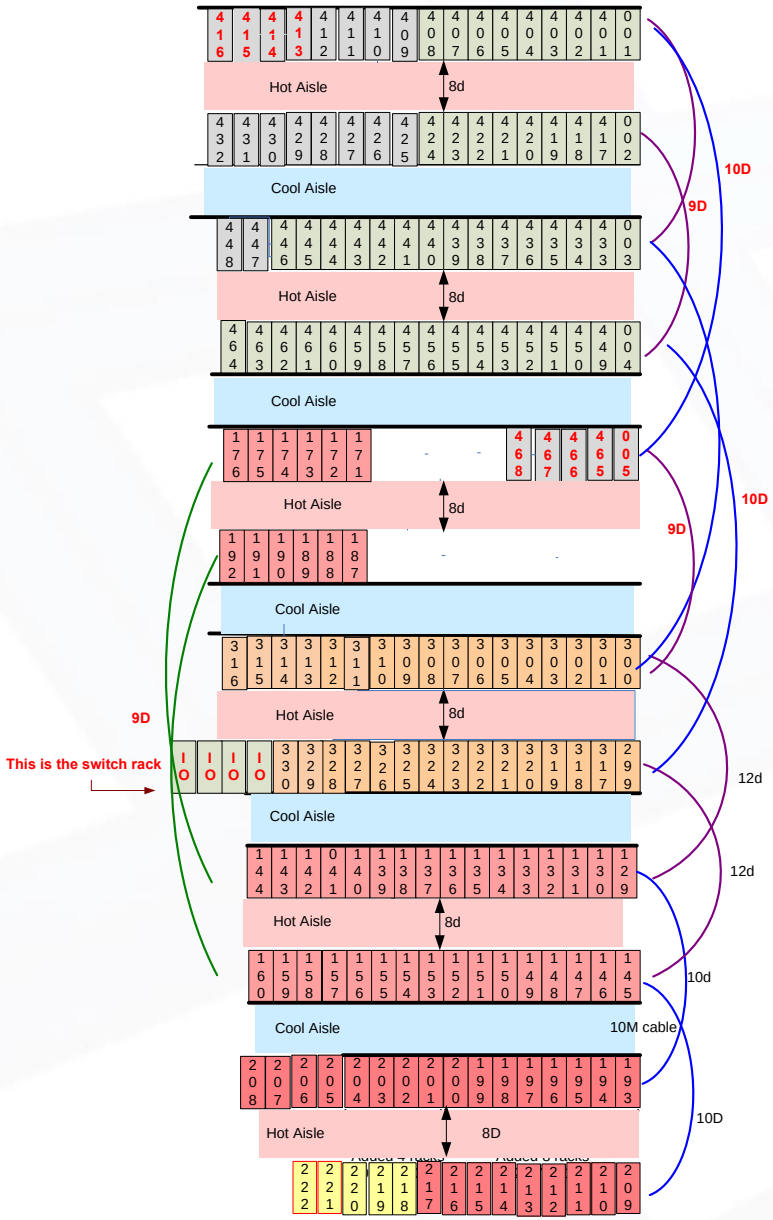
I (ICE QDR) RKS 129-144 Altix Ice 8400 EX

J (ICE QDR) RKS 145-160 ALTIX ICE 8400 EX

M (ICE QDR) RKS 193-208 ALTIX ICE 8400 EX

N (ICE QDR) RKS 209-218 ALTIX ICE 8400 EX/ RK 219-220 Coyote based Westmere with GPGPU M2090 in hypercube. RK 221-222 Altix ICE 8400 EX

NASA (Pleiades) Rack Layout as of 2/18/2014



A (ICE FDR)
RK 401-416 -SGI ICE X (ivybridge) Prem SW

B (ICE FDR)
RK417-432 SGI ICE X (ivybridge) Prem SW

C (ICE FDR)
RK 433-448 SGI ICE X (ivybridge) prem Sw

D (ICE FDR) Rk 449-464 SGI ICE X (ivybridge) Prem SW

E (ICE FDR) rks 465-468 SGI ICE X Ivybridge Prem SW – K ICE QDR) 171-176 Altix ICE 8400 EX

F (ICE FDR) rks 481-483 SGI ICE X Ivybrodge Prem SW L-ICE QDR) 187-17192 Altix ICE 8400 EX

O (ICE FDR)/ RK 300-312 SGI ICE X Sandybridge Prem SW/ RK 313-316 128 node Pyramid in hypercube topology

P (ICE FDR)/ RK 317-330 SGI ICE X SNB Prem SW

I (ICE QDR) RKS 129-144 Altix Ice 8400 EX

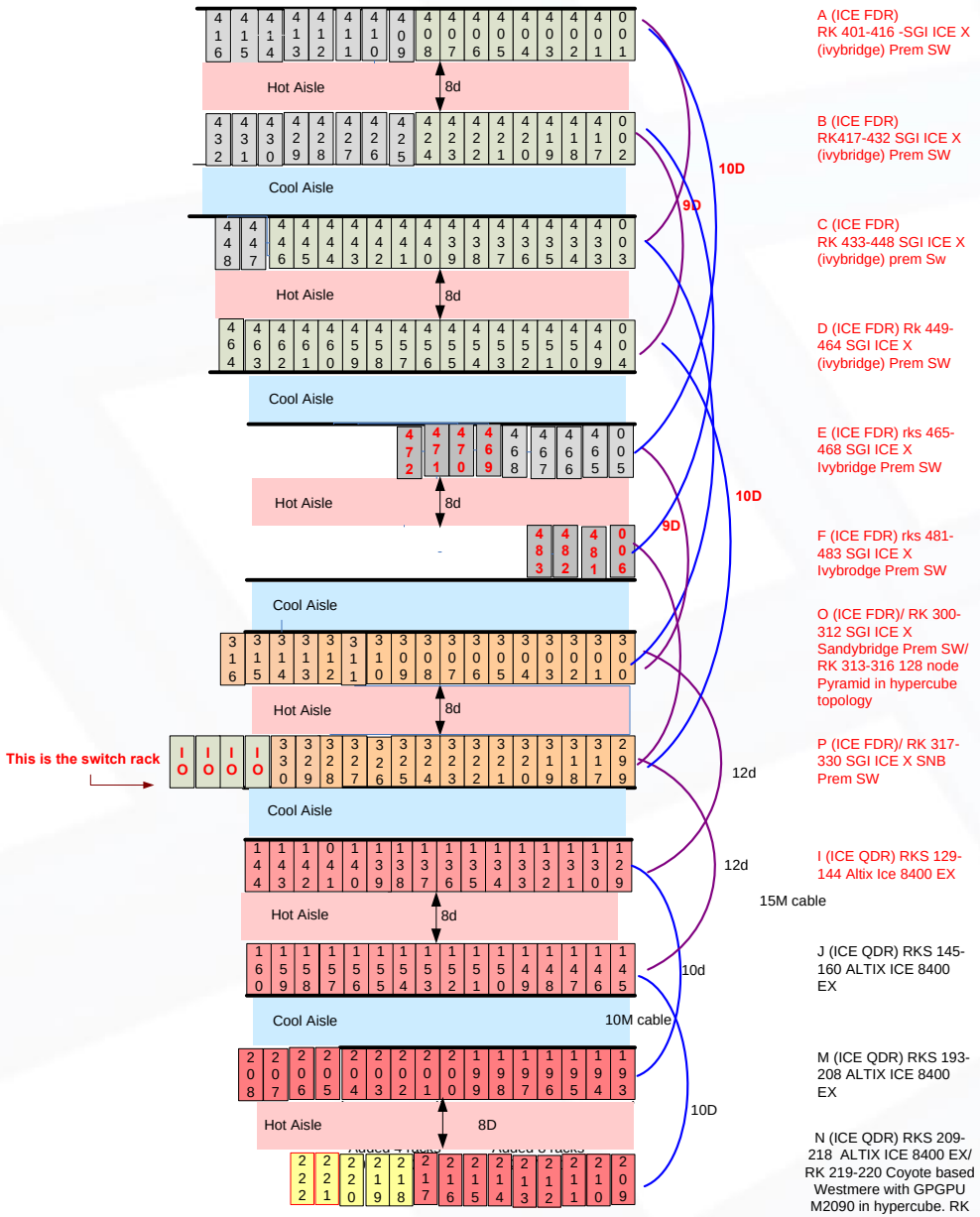
J (ICE QDR) RKS 145-160 ALTIX ICE 8400 EX

M (ICE QDR) RKS 193-208 ALTIX ICE 8400 EX

N (ICE QDR) RKS 209-218 ALTIX ICE 8400 EX/ RK 219-220 Coyote based Westmere with GPGPU M2090 in hypercube. RK 221-222 Altix ICE 8400 EX

NASA (Pleiades) Rack Layout as of 2/25/2014

170 racks – 2014
3.5 petaflops



Incremental Expansion – Driving Factors

- Annual Funding/Budget Uncertainty
- Synthetic Leases/Sarbanes-Oxley cost
- Risk Mitigation for Fast moving technology
- Supports Short Lead/Opportunistic Strategy
- Timed adoption based on technology readiness
- Decouples technologies on different development cycles
- Dynamic project support

Maintains leading edge components throughout our

"Ground Based Instrument"

Production Filesystems

- 16 different filesystems
  - 8 lustre scratch filesystems spread over about 12PB
    - Speed Optimized
    - IOP Optimized
    - Dedicated
  - 8 nfs filesystems
    - Home, scratch, system images

Production Software Environment

- 4 different production selectable operating systems
  - AOE: 3 sles, centos
  - Additional test images
- 251 different loadable modules
  - 58 different compilers (32 intel, 8 PGI, 4 gcc, 3 cuda, 3 matlab…)
  - 26 different MPIs (10 SGI MPT, 12 Intel MPI, 8 MVAPICH)
  - 23 libraries (13 hdf, 6 netcdf, 4 mkl)
- Various debuggers, performance analyzers, plotting/graphing, editors
- Driven by user requests/requirements

This is an HPC Cloud

What is Todays General Purpose Supercomputer

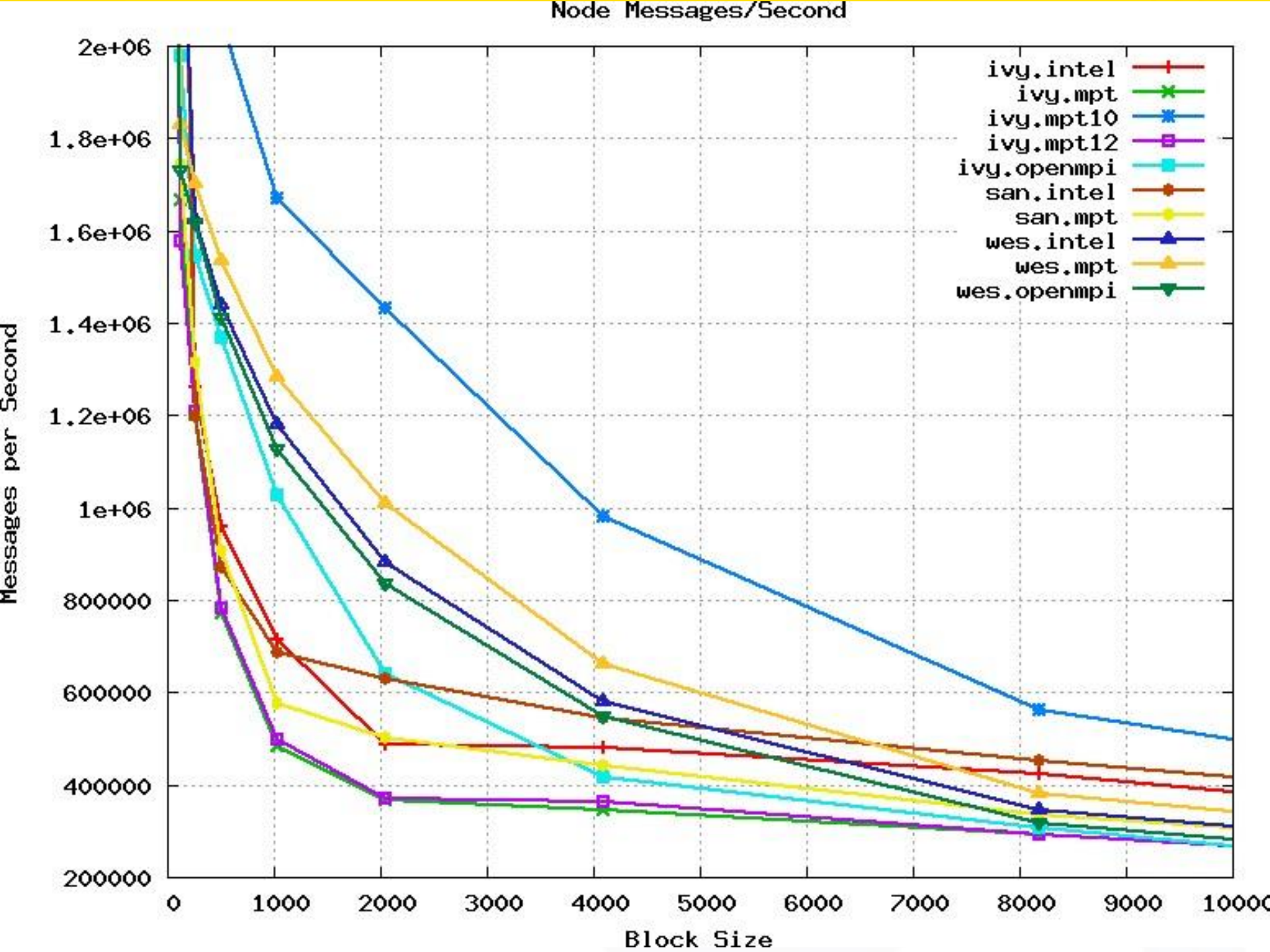
- 1980s/1990s – a monolithic system with limited access
  - Typically served smaller communities
  - Local dedicated disk with limited network connectivity
- Today – its a collection of heterogeneous elements both SW & HW
  - Supports a wide variety and types of computation
  - Tuned for user productivity
- General Purpose - a compromise in some ways
  - MAY not be the #1 top 500 machine
  - But should be the most productive for highly varied requirements in multiple science and engineering domains.

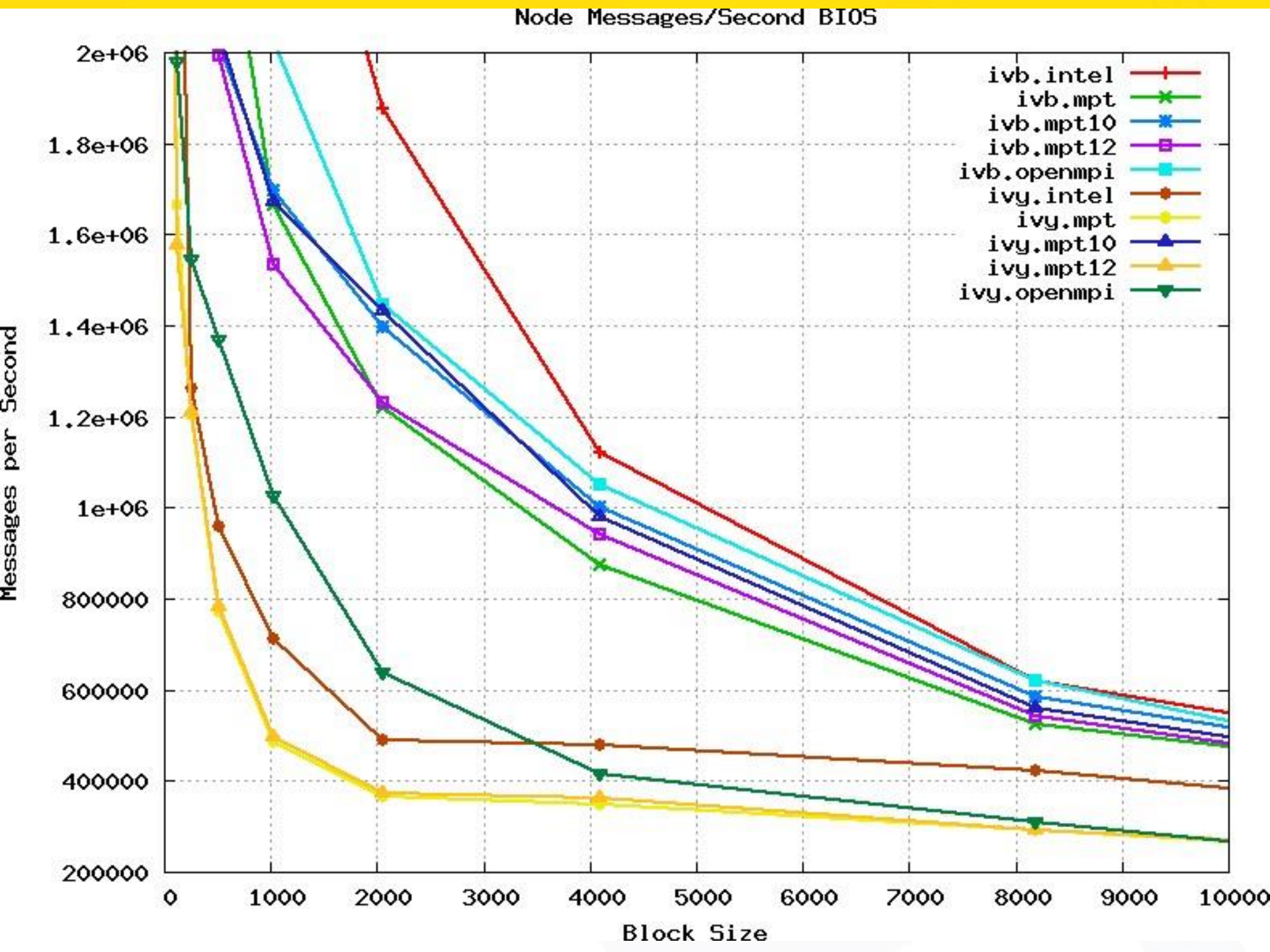
How does this relate to Exascale

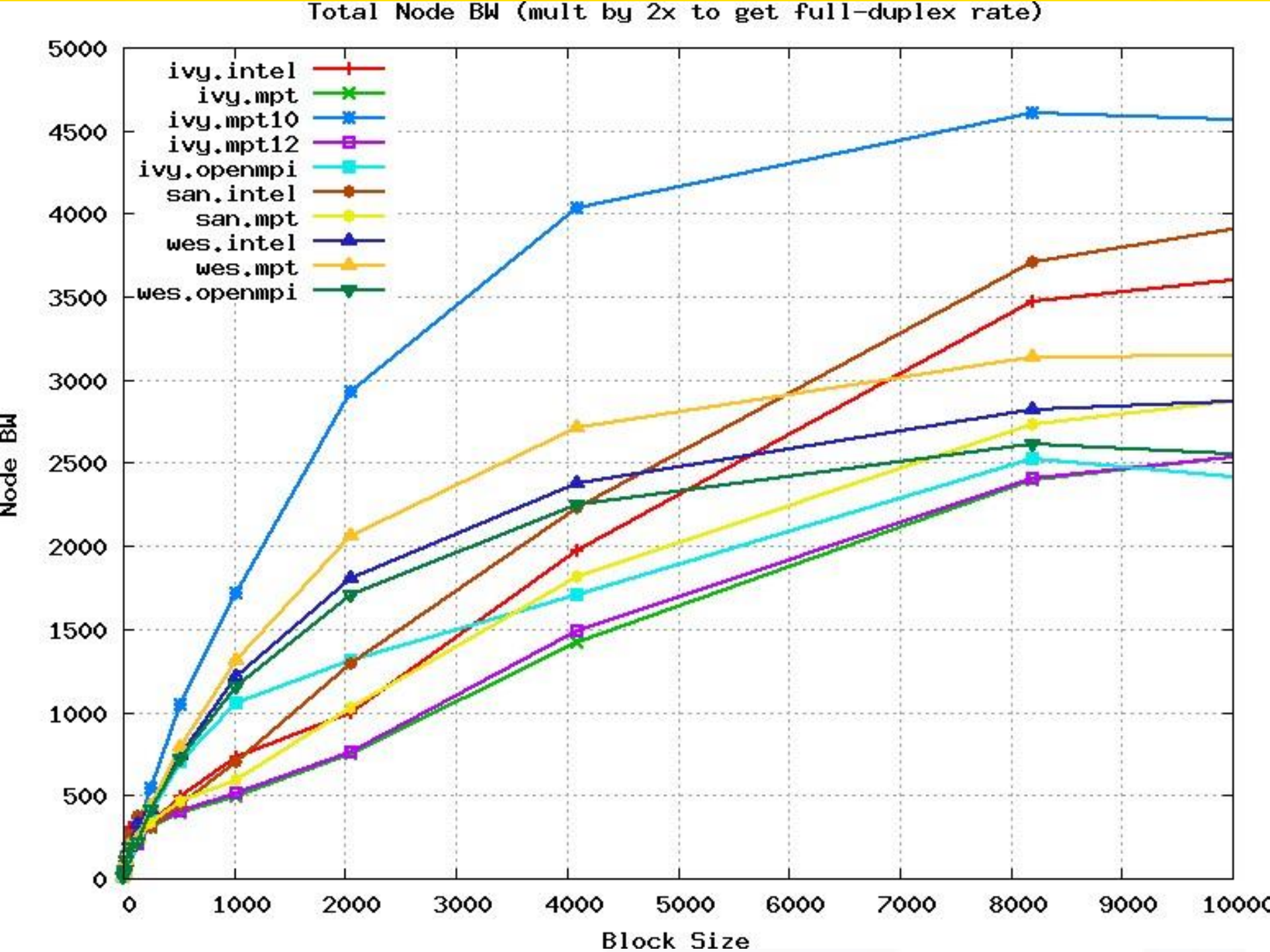
- NASA will get there – Incrementally
  - As long as doing so enables engineering and scientific discovery
- Exascale systems are likely to be more heterogeneous - diverse
  - More specialized components
    - Increased relevance around GPU/ARM/Phi
  - Cost saving realized from leveraging existing infrastructure
- Exascale systems will have significantly more parallelism
  - More things to break
- **Grand'ist challenge – dealing with component failure of HW/SW**

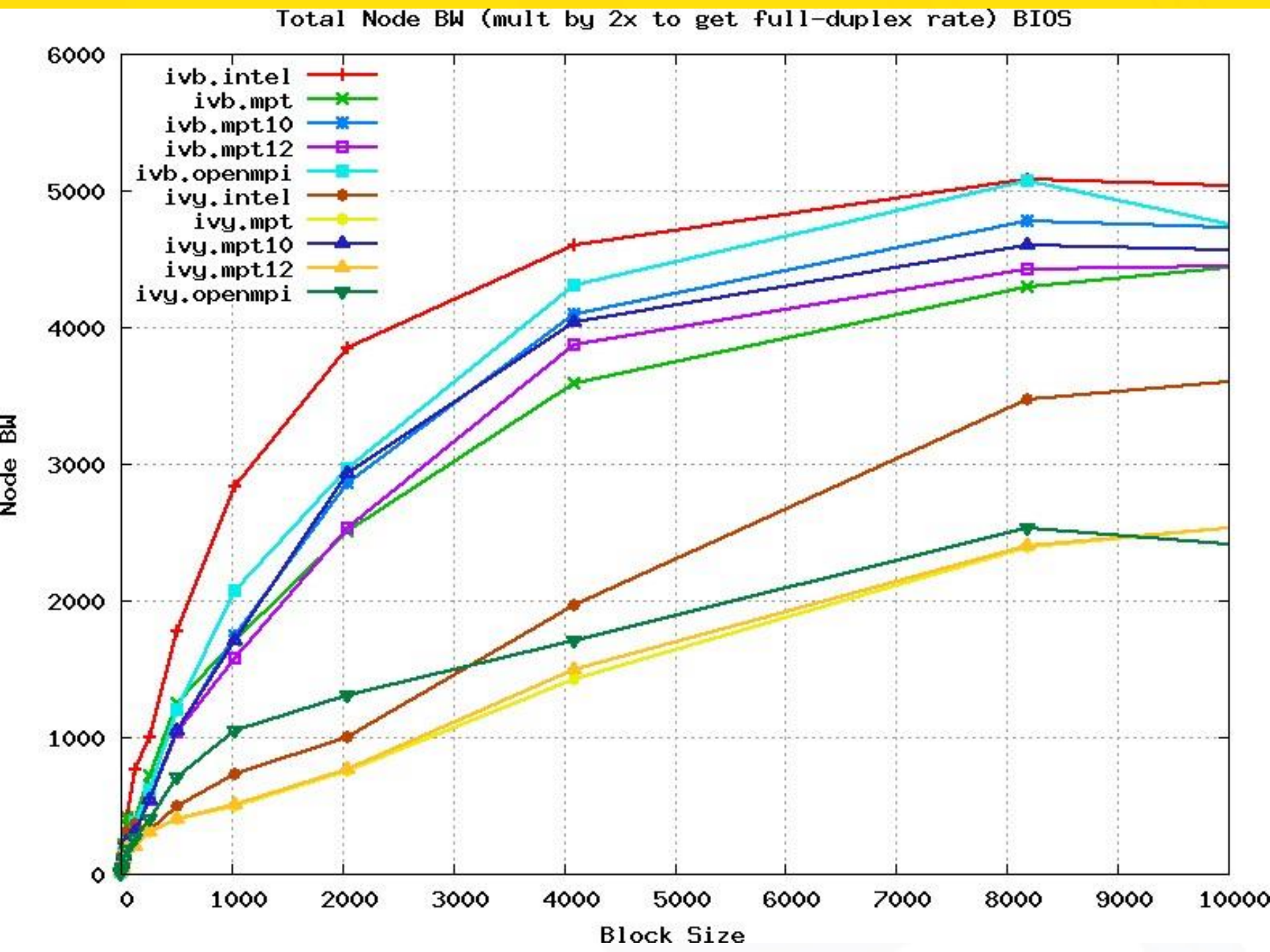
Infiniband Limitations (10x)

- Infiniband Cables
  - Physical layer issues – diagnostics – light queues – SM support
- Subnet Management
  - Does not scale - Scans, SA, multi-cast, restarts, continuous operation
- Packet Loss/RC not reliable
- Bonding Performance/Failover


Node Messages/Second
Messages per Second
Block Size
2e+06
1.8e+06
1.6e+06
1.4e+06
1.2e+06
1e+06
800000
600000
400000
200000
0
1000
2000
3000
4000
5000
6000
7000
8000
9000
10000
ivy.intel
ivy.mpt
ivy.mpt10
ivy.mpt12
ivy.openmpi
san.intel
san.mpt
wes.intel
wes.mpt
wes.openmpi


Node Messages/Second BIOS
Messages per Second
Block Size
2e+06
1.8e+06
1.6e+06
1.4e+06
1.2e+06
1e+06
800000
600000
400000
200000
0
1000
2000
3000
4000
5000
6000
7000
8000
9000
10000
ivb.intel
ivb.mpt
ivb.mpt10
ivb.mpt12
ivb.openmpi
ivy.intel
ivy.mpt
ivy.mpt10
ivy.mpt12
ivy.openmpi


Total Node BW (mult by 2x to get full-duplex rate)
ivy.intel
ivy.mpt
ivy.mpt10
ivy.mpt12
ivy.openmpi
san.intel
san.mpt
wes.intel
wes.mpt
wes.openmpi
Node BW
Block Size
0
500
1000
1500
2000
2500
3000
3500
4000
4500
5000
0
1000
2000
3000
4000
5000
6000
7000
8000
9000
10000


Total Node BW (mult by 2x to get full-duplex rate) BIOS
ivb.intel
ivb.mpt
ivb.mpt10
ivb.mpt12
ivb.openmpi
ivy.intel
ivy.mpt
ivy.mpt10
ivy.mpt12
ivy.openmpi
Node BW
Block Size
0
1000
2000
3000
4000
5000
6000
7000
8000
9000
10000

Infiniband Limitations (10x)

- Congestion Control
  - Getting worse for us
  - Nodes more easily overwhelm fabric
  - Discards more common
    - Particularly problematic in one code
  - Retry mechanisms too simply
  - Who is causing the problem
    - Slow node
    - Improper Software
    - Link Recovers
    - Slow Management

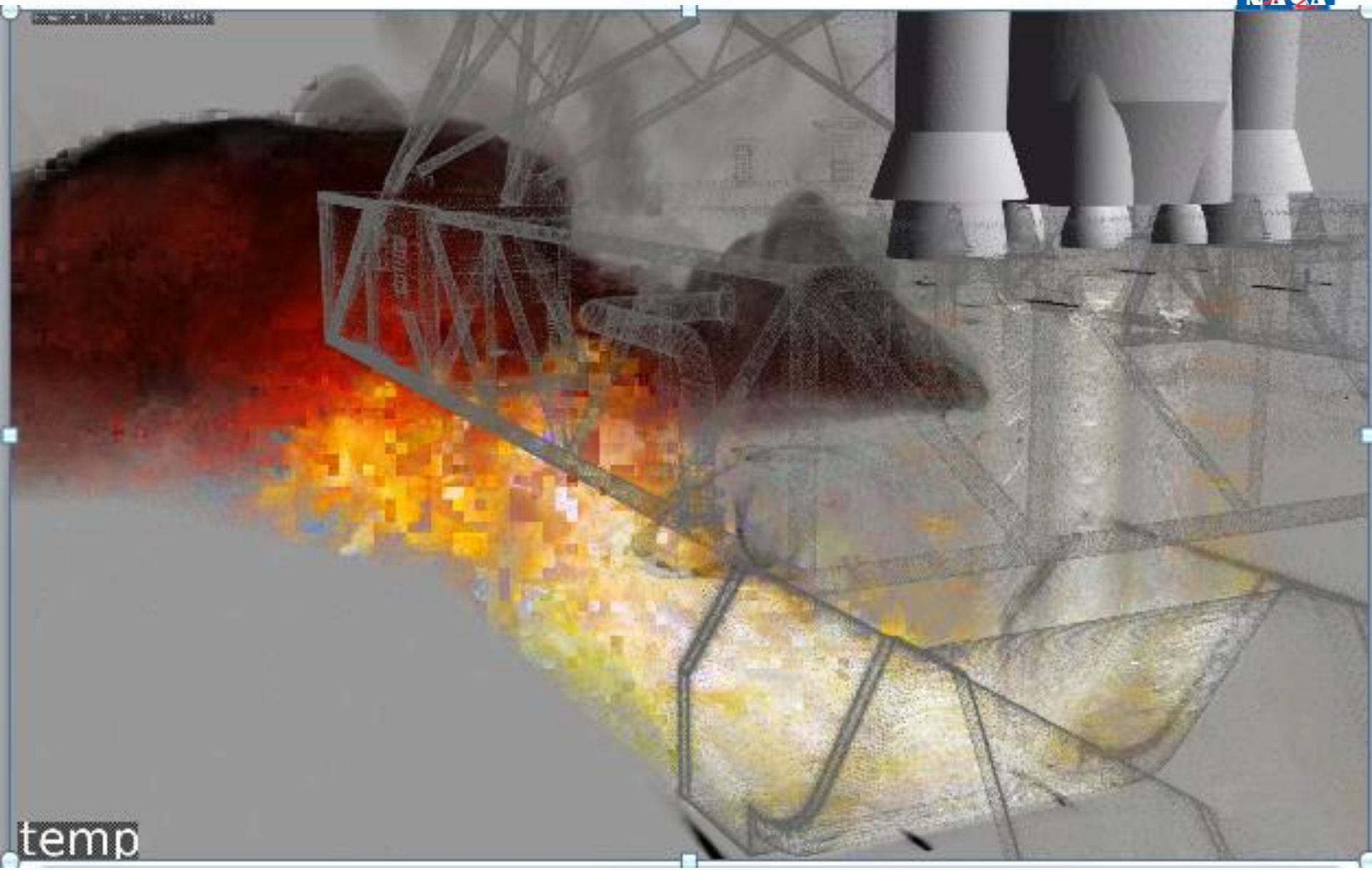
SM/SA Requirements

- Does not scale
  - Scans, SA, multi-cast, restarts, continuous operation
- Must handle 1000's of nodes:
  - Crashing
  - Rebooting
  - Being slow or non-responsive or broken
- System Maintenance and Expansion
  - Active SW GUIDS, Active Ports (cable maint *)
- Large Scale systems have regularity to leverage static information
  - SW Ports known to have HCA (*)
  - Static or locally discoverable) information (LIDs, QP services)
  - Decentralized/Distributed services (H-ARP *),
- Separate Switch and HCA management
- Intelligent Switches (am I connected to the right components)

(* working these)

Conclusions

- Backward and forward compatible high speed networks have big advantages in leveraging existing infrastructure.
- Future networks need to address interoperability at a reasonable cost.
  - Protocol translators cannot require significant HW (e.g. server)
- Converged/Heterogeneous Super Scale environments benefit
  - Agile system implementations
  - Cost saving by leveraging existing infrastructure
  - Quick deployments
- Infiniband has been the key to implementing this strategy
- **Most Important Factor – performance/reliability/cost of compute**


temp