HPCG Pleiades

Bob.Ciotti@nasa.gov
John Baron jbaron@sgi.com
NASA Ames Research Center

Pleiades HW Environment

- 11, 472 compute nodes 246,048 x86 cores
  - 1,968 Sandybridge
  - 5,400 Ivybridge
  - 2,088 Haswell
  - 2,016 Broadwell
- 938TB Memory
- FDR Infiniband – dual rail hypercube
- Additional task specific nodes
  - GPU
  - Xeon Phi (KNC+KNL)
  - 1024/512 cpu large shared memory
  - Large memory data analysis nodes
  - Front Ends
  - hyperwall viz/data analysis
- + a couple hundred administration/management nodes of various types.

Pleiades SW Environment

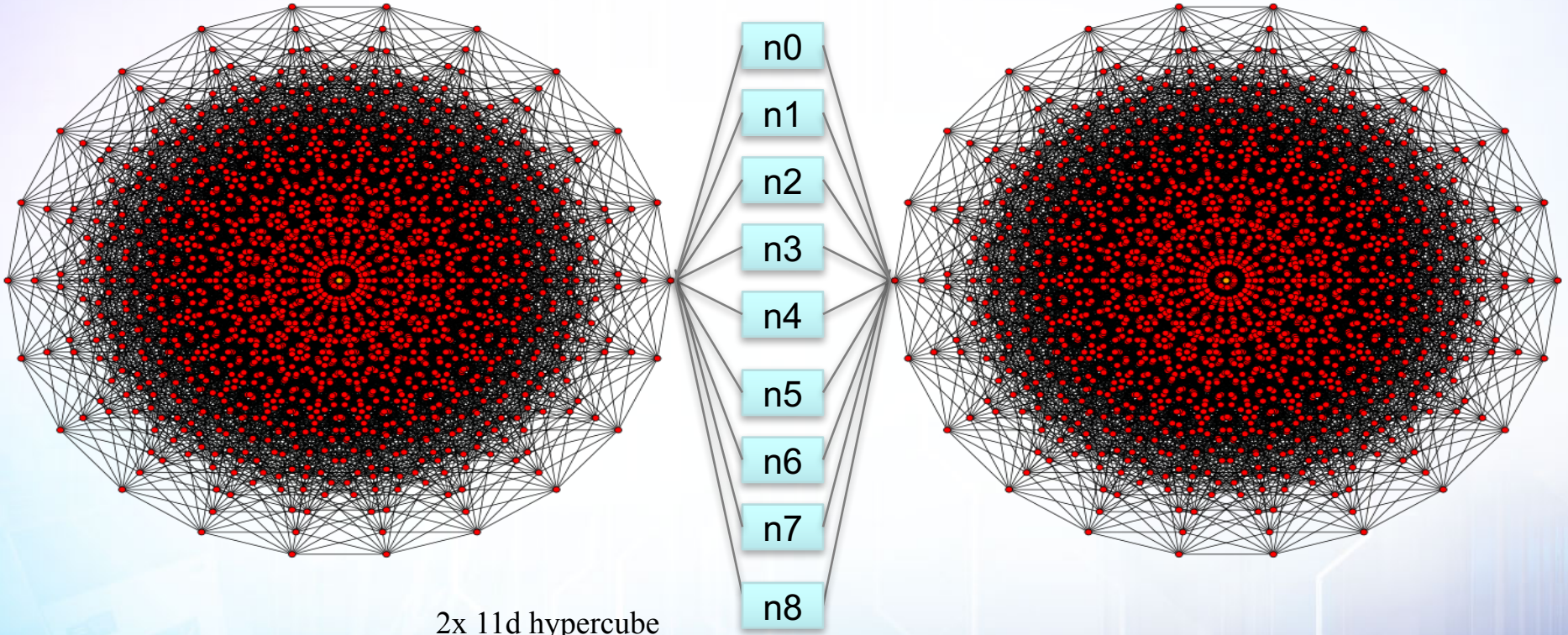
- LINUX
  - SLES11/12 (most user facing systems)
  - Red Hat/Centos (lustre servers)

- Lustre
- NFS

- Continuous Availability
- *All* software can be updated without full system dedicated outage
  - ‘rolllling updates for compute nodes
  - Suspend/Resume for service nodes (lustre/NFS servers, rack leaders)

- Compute nodes added/removed without dedicated system down

SGI ICE Dual Plane – Topology

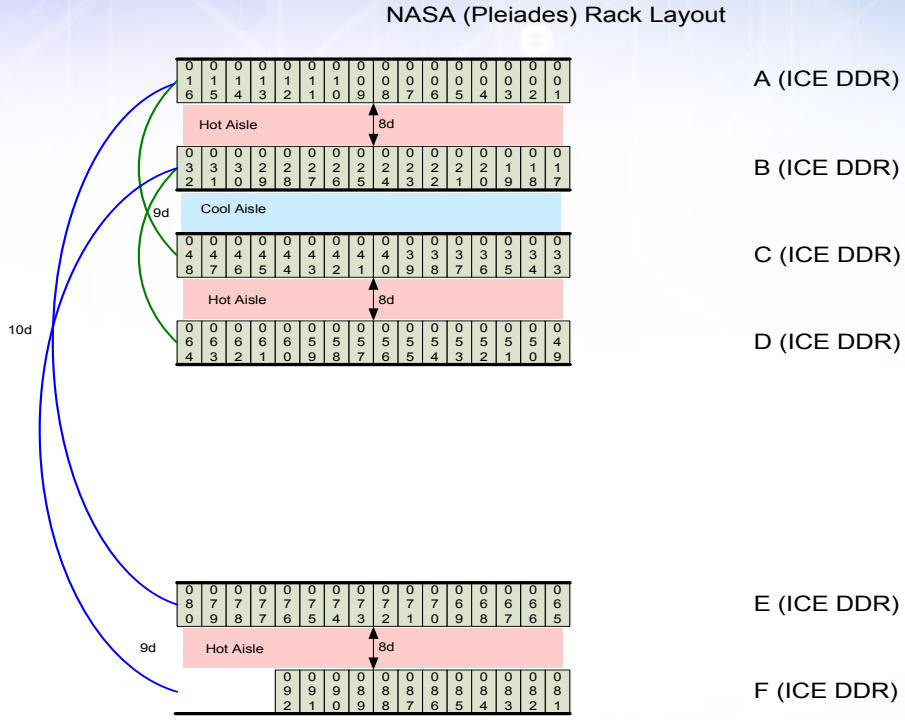


ib0

2x 11d hypercube
full 11d == 2048 vertices
Pleiades – partial 11d - 1296 vertices (2592 across both cubes)

ib1

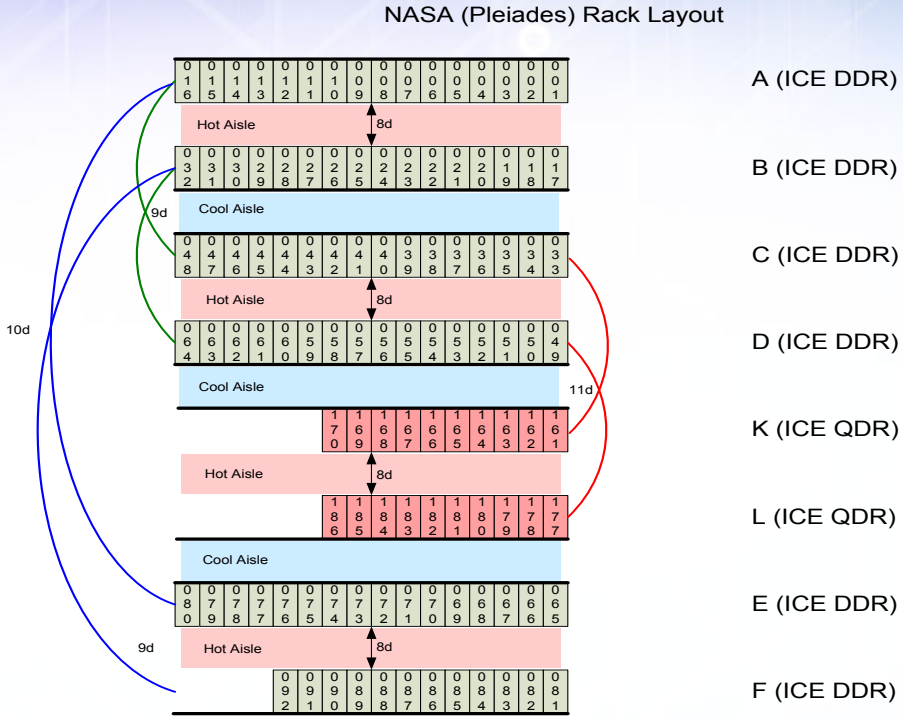
http://en.wikipedia.org/wiki/User:Qef/Orthographic_hypercube_diagrams



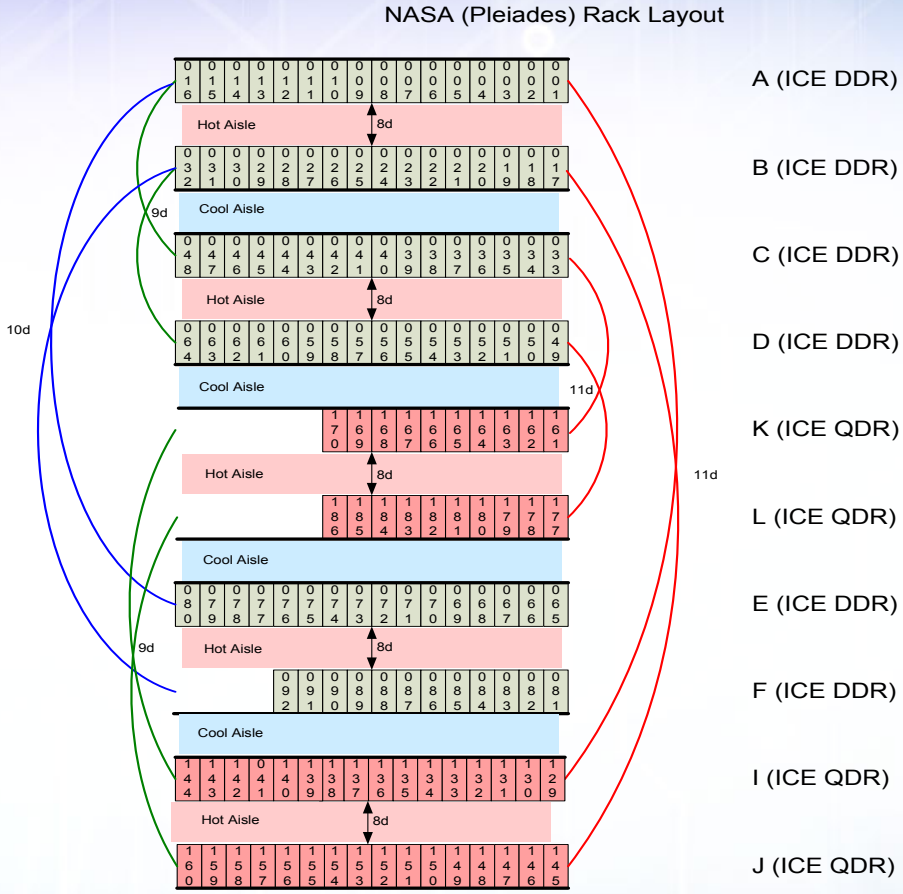
92 racks – 2008
565 teraflops

#3 Top500

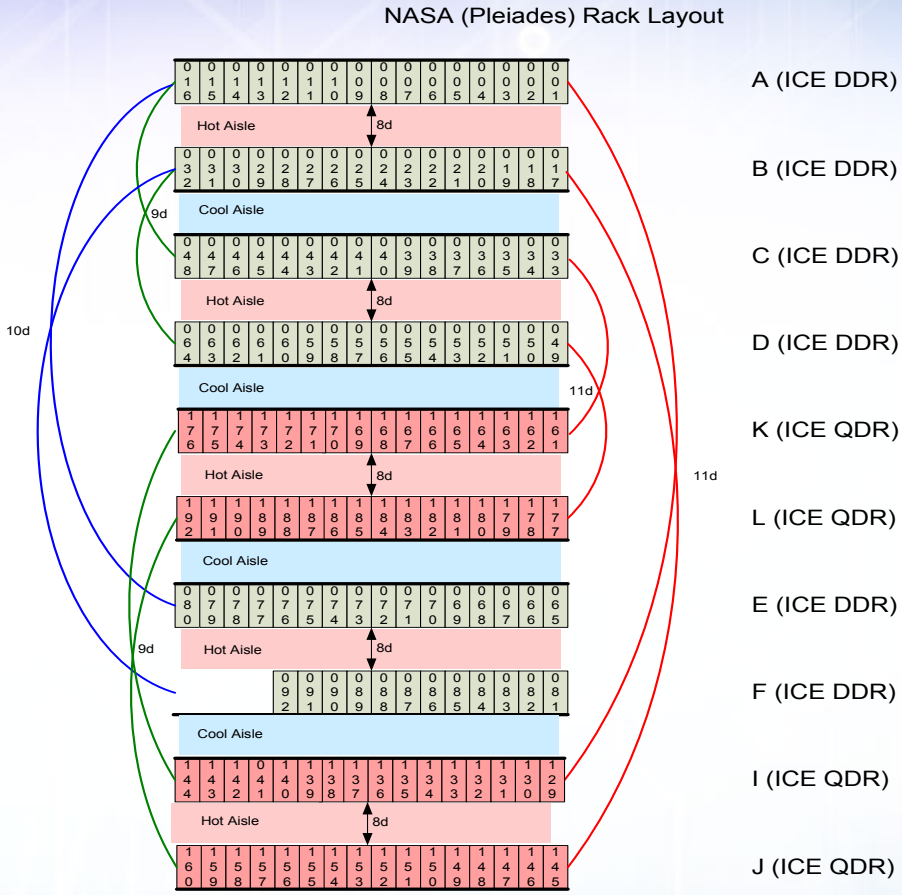
112 racks – 2009
683 teraflops



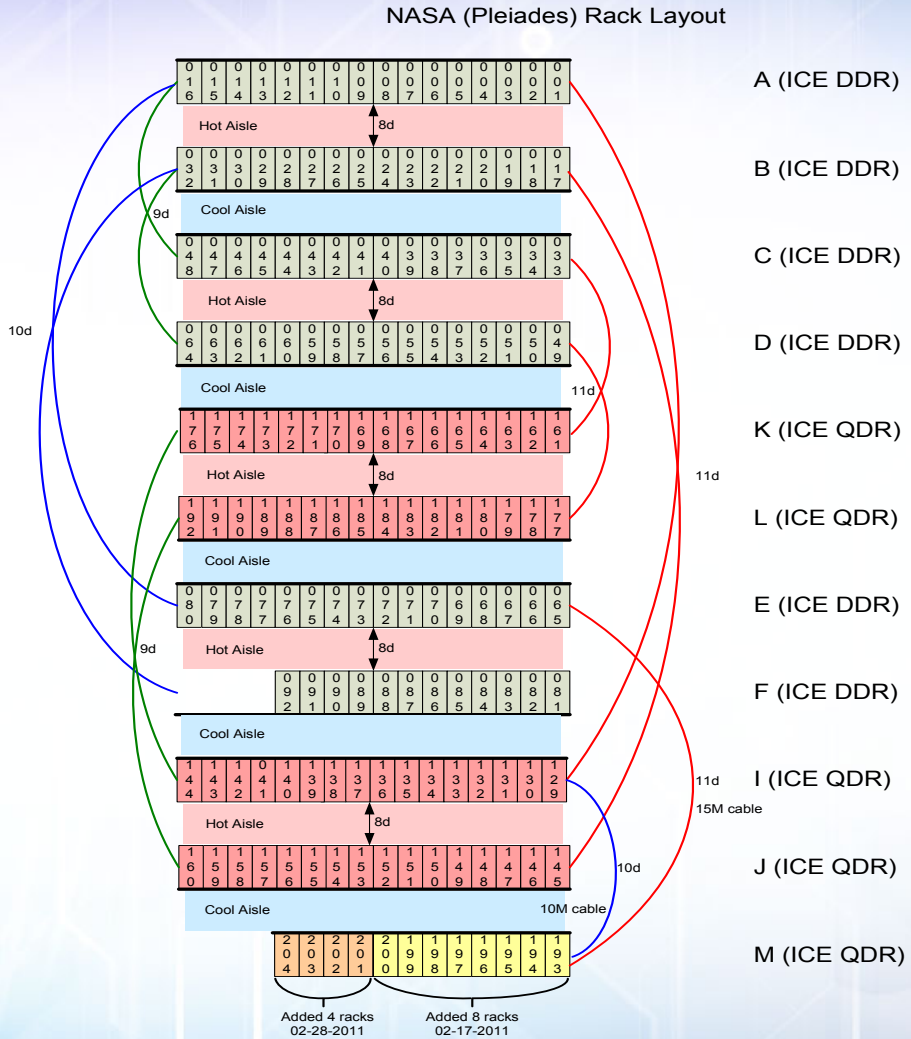
144 racks – 2010
969 teraflops



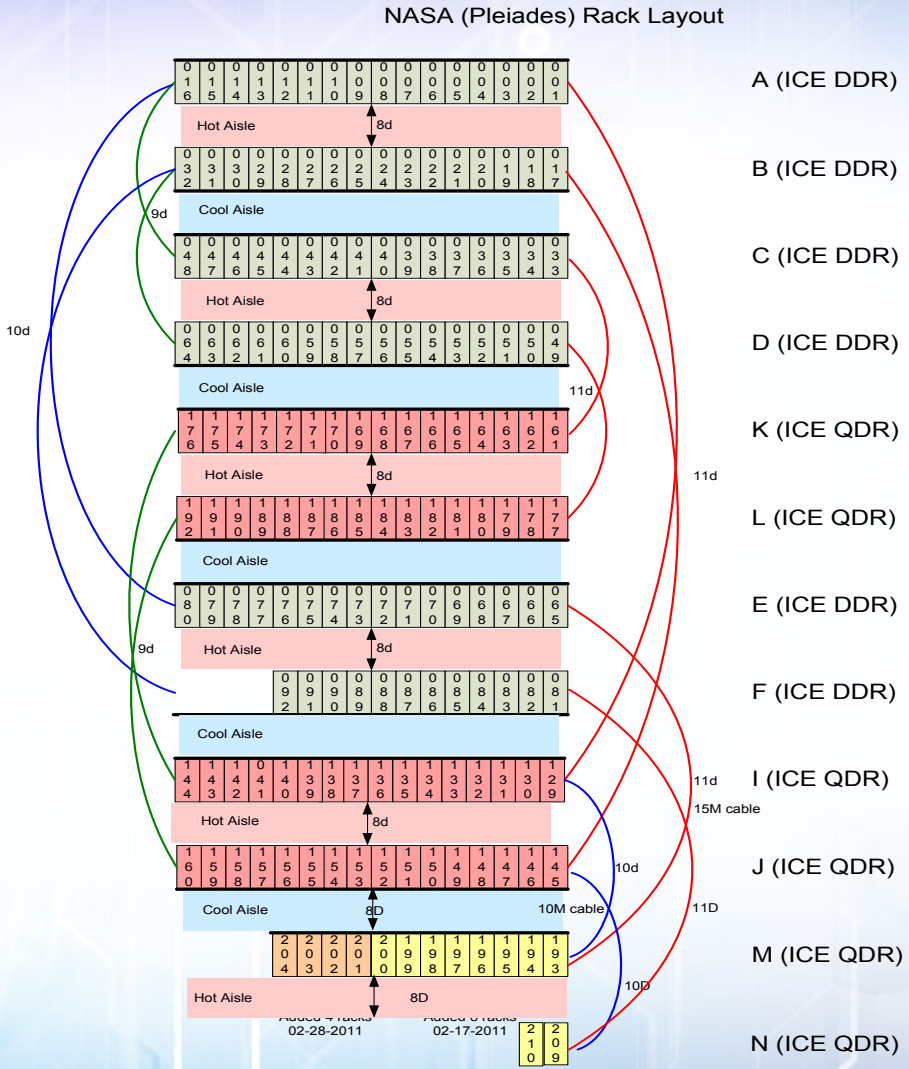
156 racks – 2010
1.08 petaflops



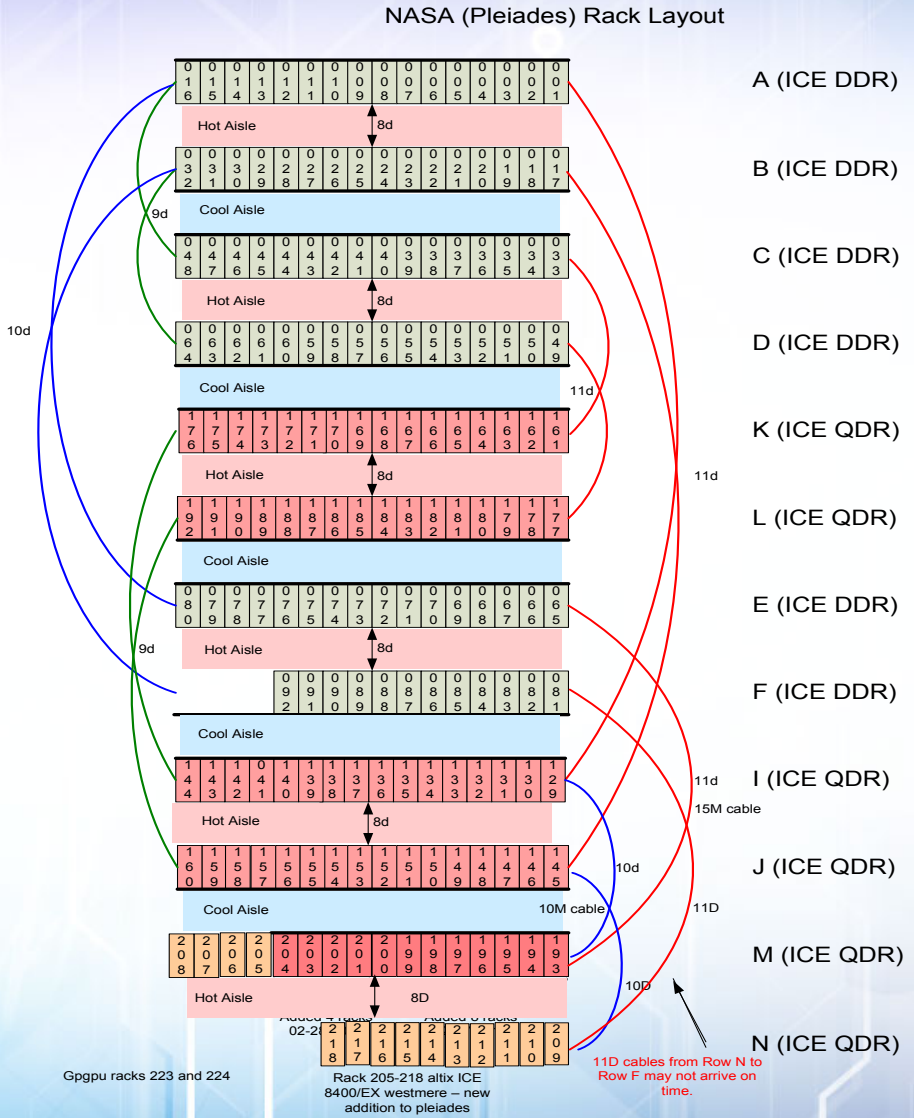
168 racks – 2011
1.18 petaflops



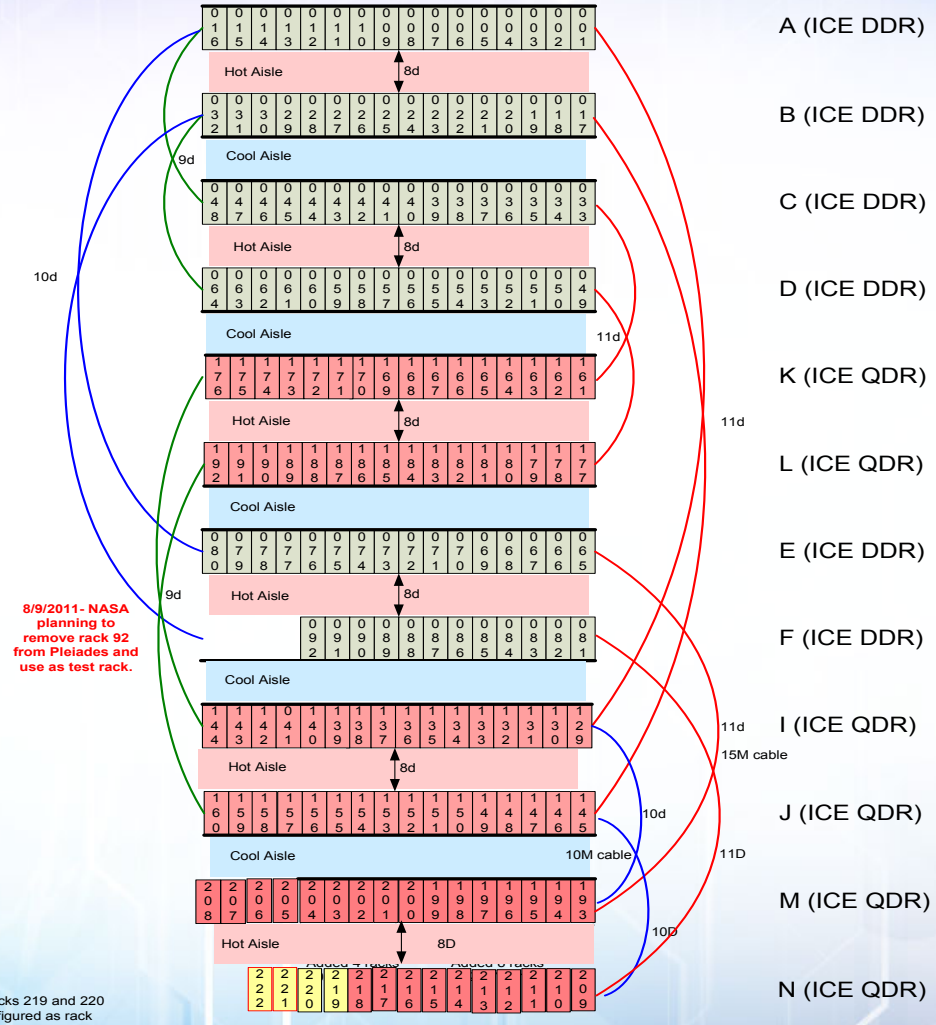
170 racks – 2011
1.20 petaflops



182 racks – 2011
1.31 petaflops



NASA (Pleiades) Rack Layout



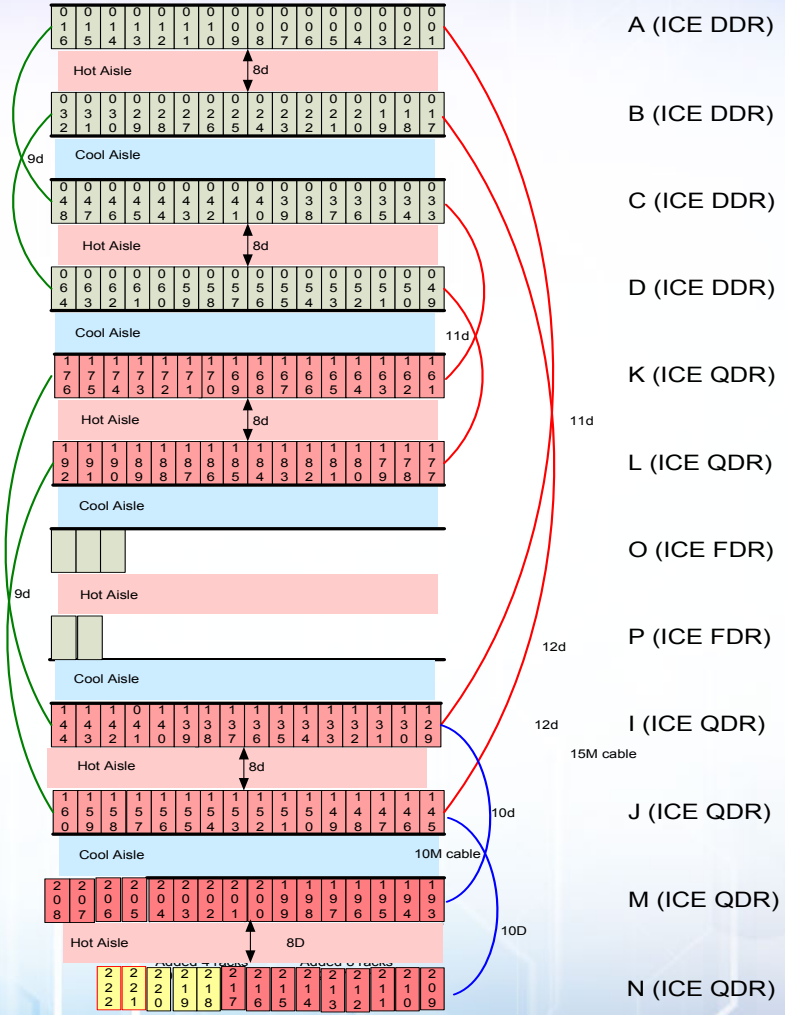
186 racks – 2011
1.33 petaflops

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

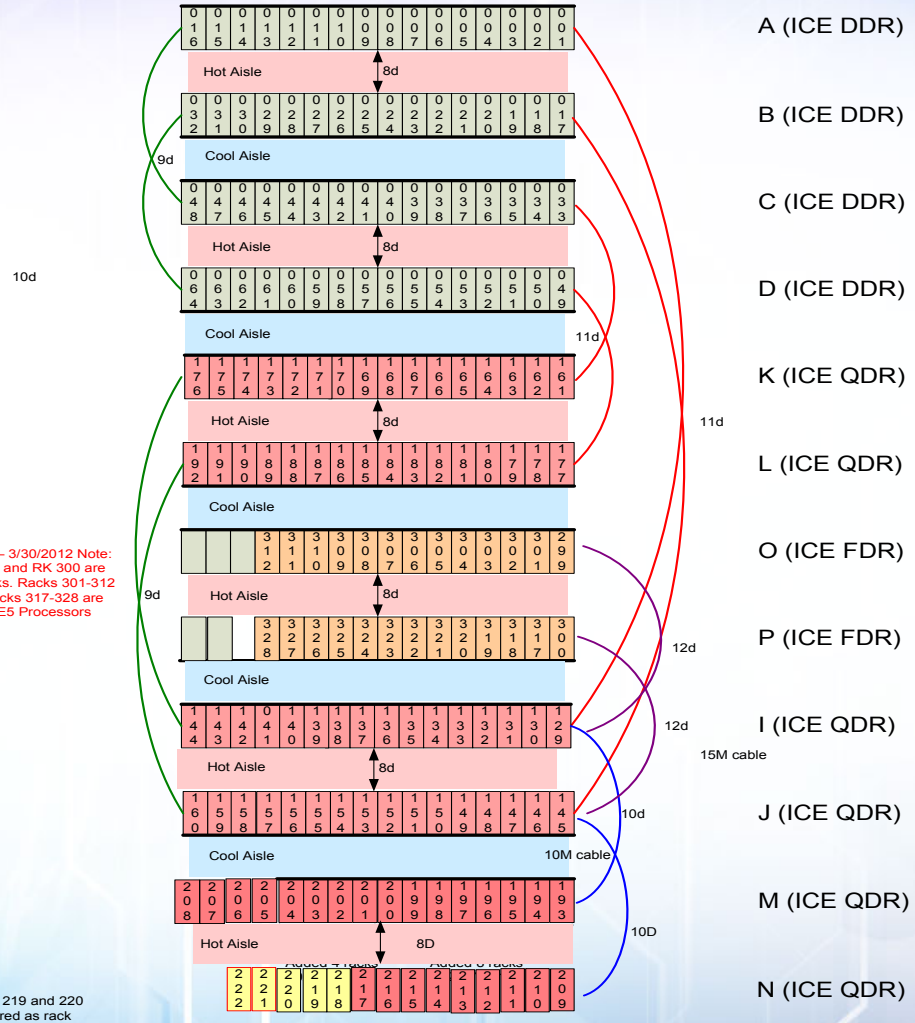


*Note: Harpertown Racks Removed 3/21/2012 in preparation for SGI ICE X Racks installation. I/O Racks remain

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.

NASA (Pleiades) Rack Layout



182 racks – 2012
1.7 petaflops

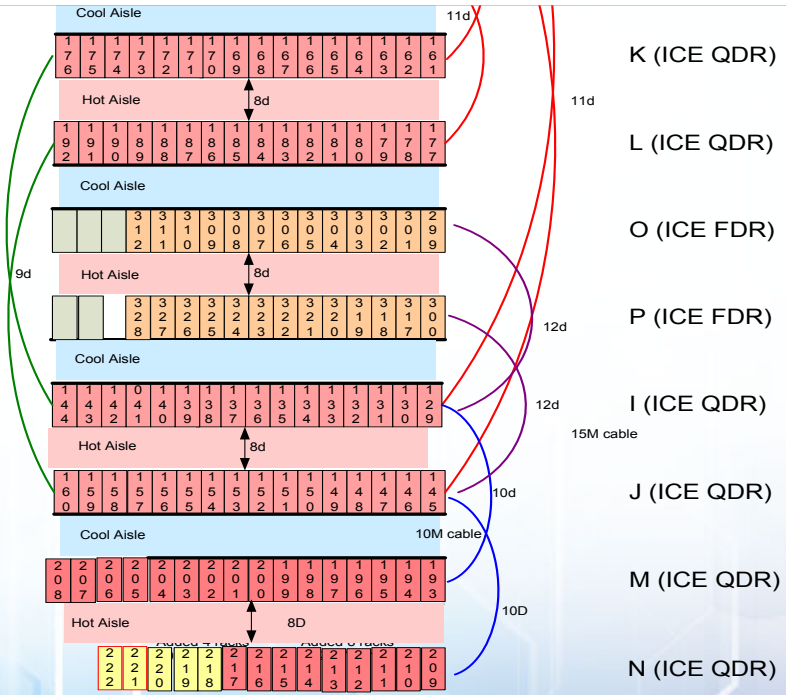
* Install – 3/30/2012 Note: RK 299 and RK 300 are RLC racks. Racks 301-312 and Racks 317-328 are Intel E5 Processors

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.

NASA (Pleiades) Rack Layout

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.

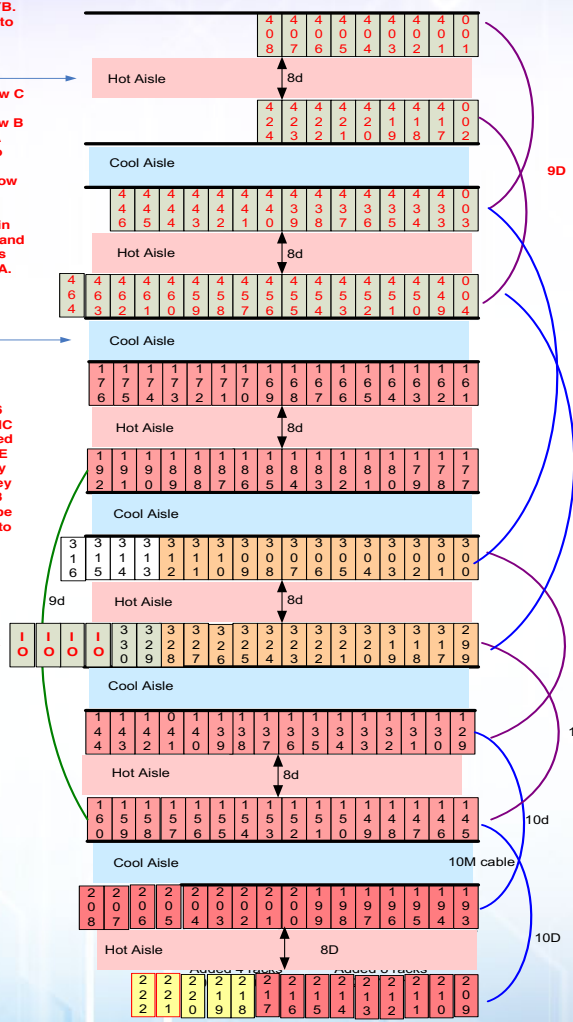
NASA (Pleiades) Rack Layout

Note: 06/21/2013 -Rack 001-004 are I/O racks for RLC and switches.. RowS A,B,C,D – 46 racks are proposed IVYB. They will connect via 10D to Row O and P. This will be partial 9D and partial 10D.

Note:
1st delivery : 8 racks in Row C
8 racks Row D
2nd delivery: 8 racks in Row B
8 racks in Row A
3rd delivery: add 8 racks to row D
add 6 racks to Row C

Rack 001-004 are the admin racks that house the RLC and ethernet switches. There is one being added for Row A.

167 racks – 2013
2.9 petaflops



* Install – Rks 313-316 are the pyramid with MIC racks and are configured as two racks of SGI ICE X except there are only 64 nodes per Rack. They are virtually racks 313 and 314. They will not be delivered till Nov 2012 to NASA

A (ICE FDR)
SGI Ice X – the 48 port cable version

B (ICE FDR) SGI ICE X – the 48 port version

C (ICE FDR)
SGI ICE X – the 48 port version

D (ICE FDR)
SGI ICE X – the 48 port version

K (ICE QDR)
RK 161-170 are Altix 8200 with QDR IB SW/but DDR compute blade. RKS 172-176 are Altix 8400EX with QDR

L (ICE QDR)
RK 177-1186 are Altix 8200 with QDR IB SW/but DDR compute blade. RKS 187—192 are Altix 8400EX with QDR SW and blades

O (ICE FDR)
SGI ICE X – the 48 port version – rk 301-312 are FDR/rk 313-316 uses 2U servers and they use Mellanox 6036 unmanaged switches. There are 32 nodes/rack but arrange in hypercube like two virtual ICE X racks

P (ICE FDR)
SGI ICE X – the 48 port version – Rks 317-330 are all FDR including the compute blade.

I (ICE QDR) rk 129-144 are Altix ICE 8400 EX – all QDR

J (ICE QDR)
rk 145-260 are Altix ICE 8400 EX – all QDR

M (ICE QDR)
rk 193-208 are Altix ICE 8400 EX – all QDR

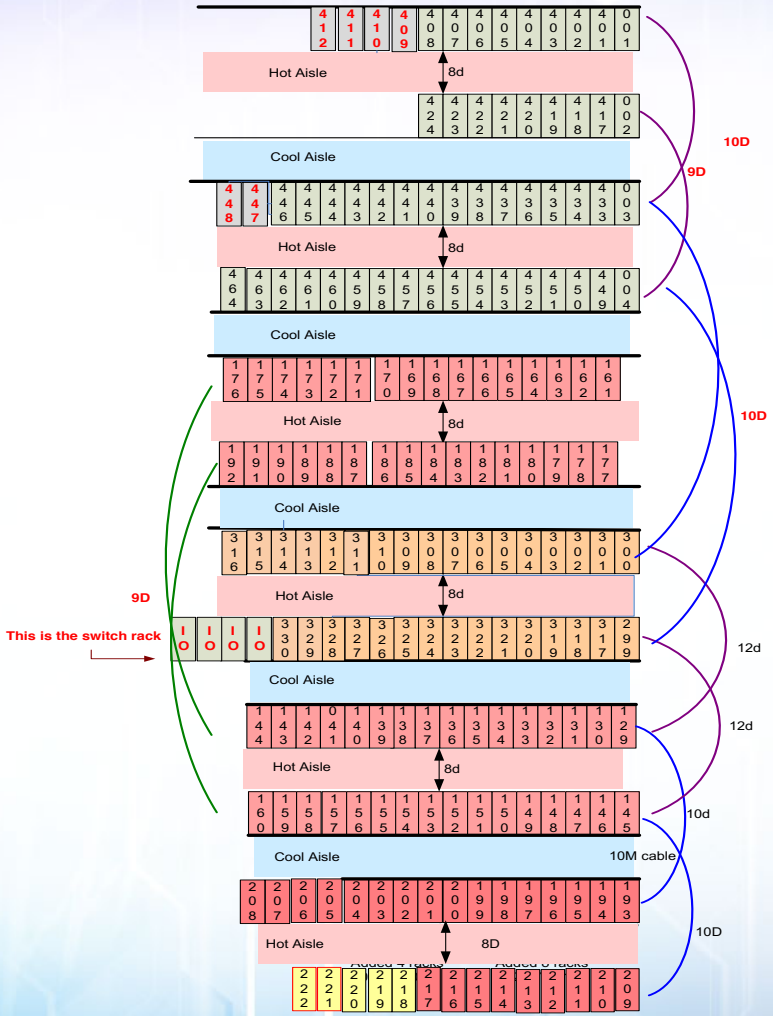
N (ICE QDR)
rk 209-217 are Altix ICE 8400 EX – all QDR / rk 219-220 is 2 racks with 32 servers each configured in hypercube using 5025 sw QDR./ rack 22-222 – Altix Ice 8400 EX all QDR SW and Blades.

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 12/30/2013

160 racks – 2013
3.1 petaflops



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

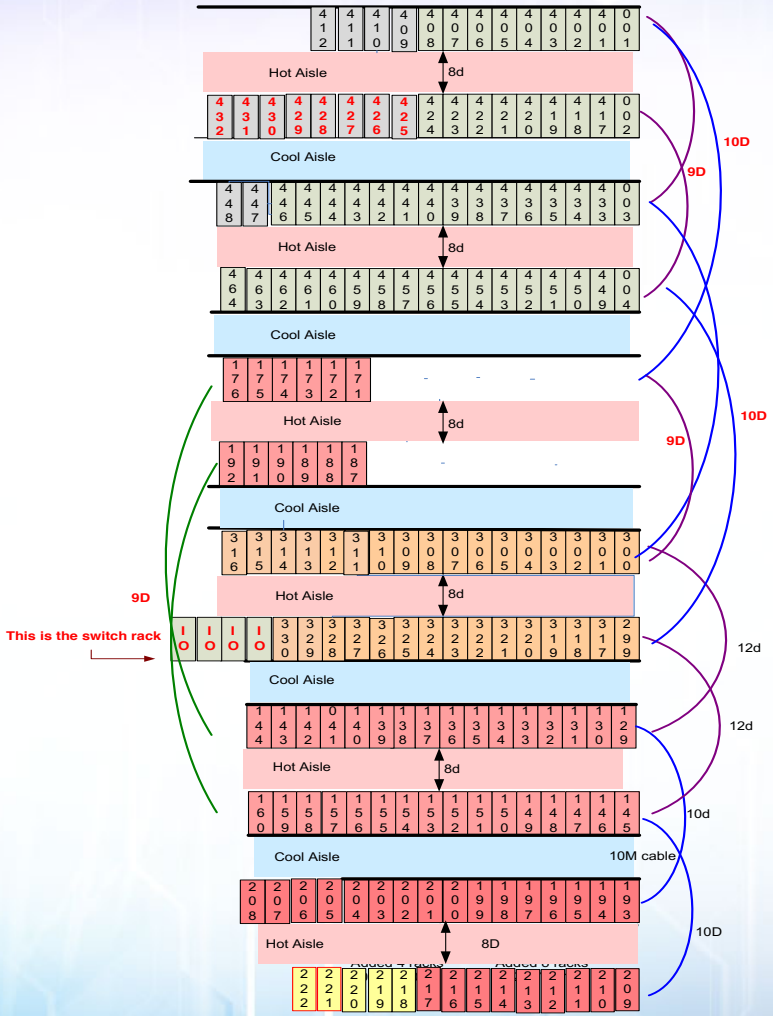
J (ICE QDR) RKS 145-160 ALTIX ICE 8400 EX

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

168 racks – 2013
3.2 petaflops



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

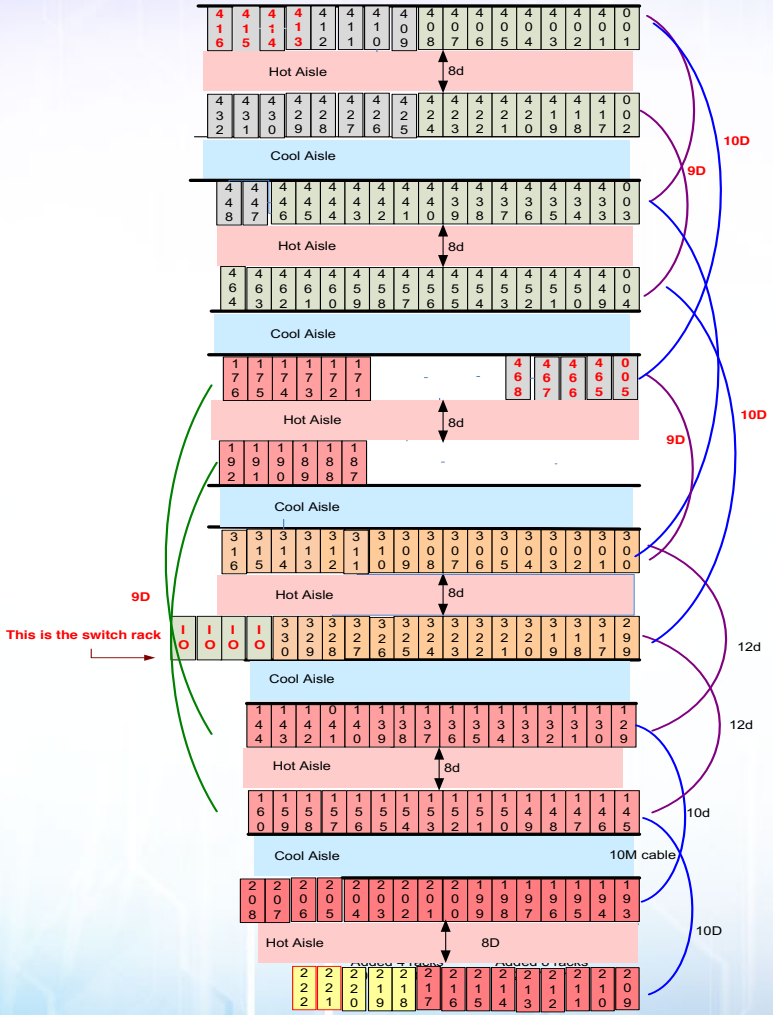
J (ICE QDR) RKS 145-160 ALTIX ICE 8400 EX

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

168 racks – 2014
3.3 petaflops



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

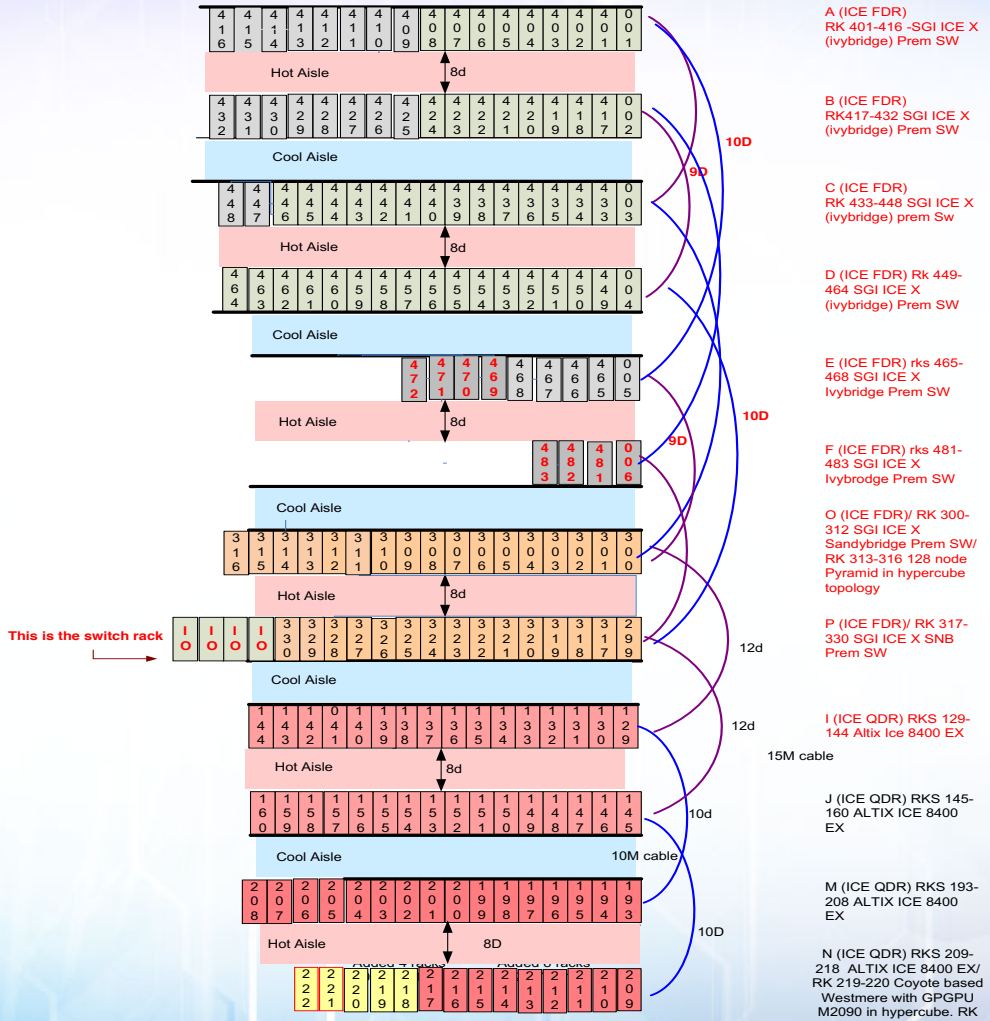
I (ICE QDR) RKS 129-144 Altix Ice 8400 EX

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

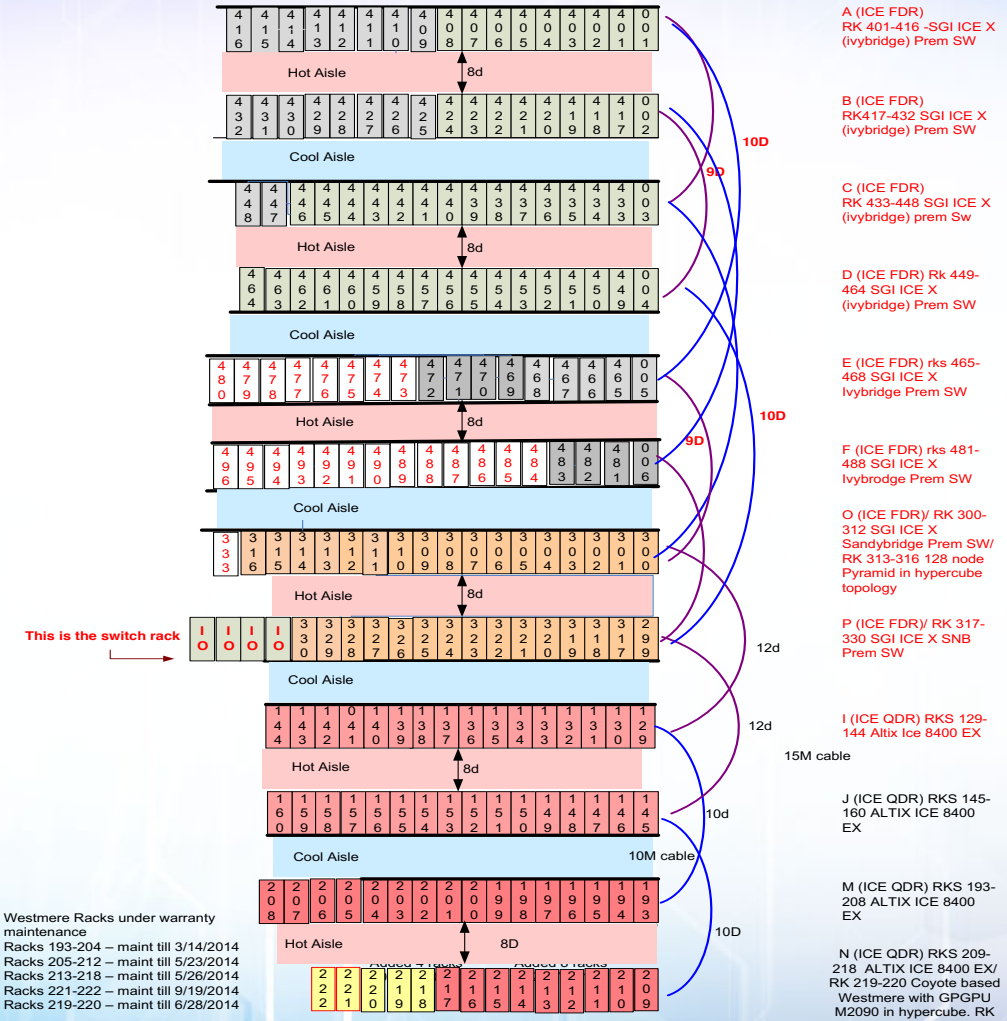
NASA (Pleiades) Rack Layout as of 2/25/2014



170 racks – 2014
3.5 petaflops

NASA (Pleiades) Rack Layout

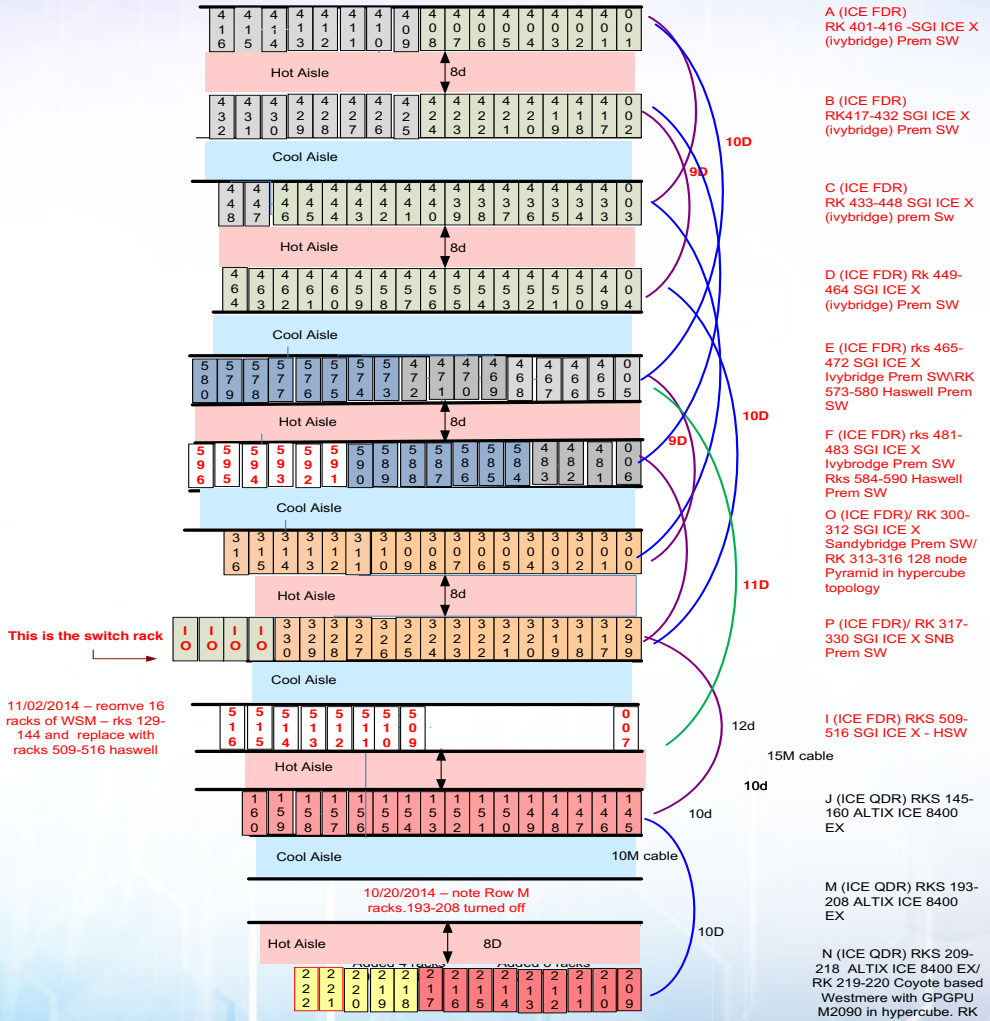
168 racks – 2014
4.5 petaflops



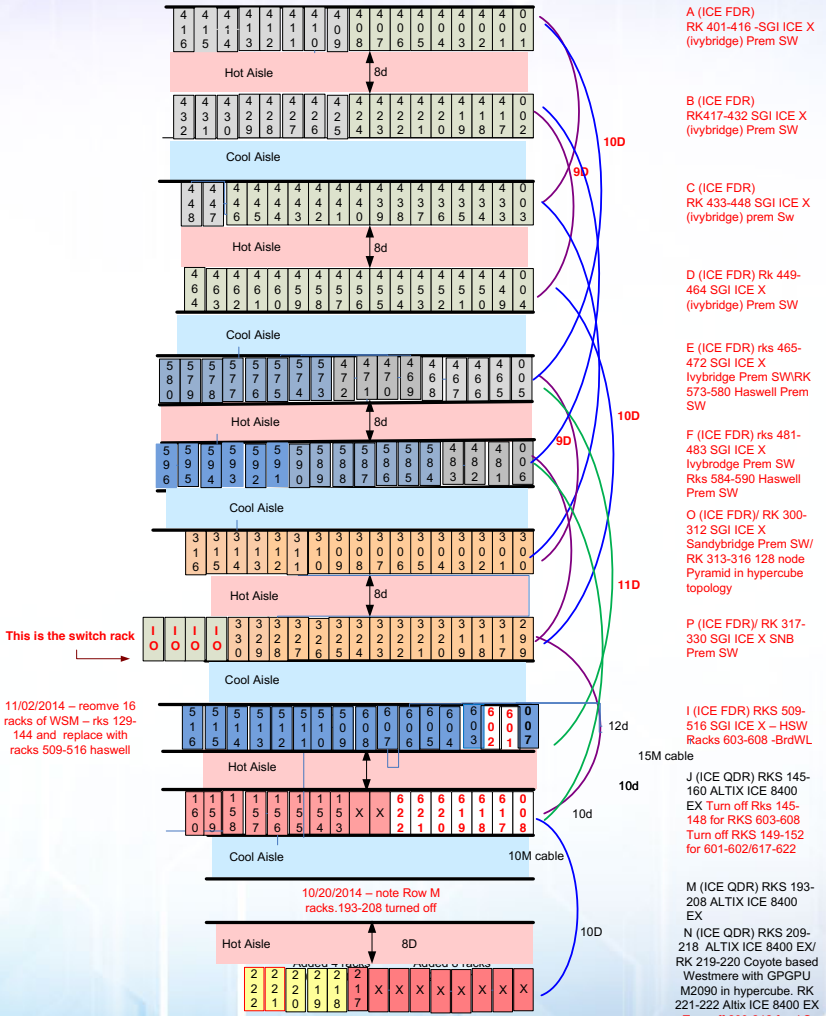
Westmere Racks under warranty maintenance
Racks 193-204 – maint till 3/14/2014
Racks 205-212 – maint till 5/23/2014
Racks 213-218 – maint till 5/26/2014
Racks 221-222 – maint till 9/19/2014
Racks 219-220 – maint till 6/28/2014

NASA (Pleiades) Rack Layout

168 racks – 2015
5.4 petaflops



NASA (Pleiades) Rack Layout



162 racks – 2016
7.1 petaflops

#13 Top500 Nov
#9 HPCG ISC 16

>20 major upgrades

Highlights of SGI Optimized HPCG Code

- Lexicographical ordering for maximum data locality
- Left and right data structures for full matrix representation
- A variant of CSR storage format
- Pure MPI
- No overlapping of computation and communication
- Additional tuning for contiguous memory, setup time and combined computations

Heterogeneous considerations

Load balancing via number of ranks per node

Broadwell E5-2680 v4 14-core 2.4 GHz

- 2015 nodes, 12 ranks/socket, 0.85 GF/rank

Haswell E5-2680 v3 12-core 2.5 GHz

- 2080 nodes, 10 ranks/socket, 0.91 GF/rank

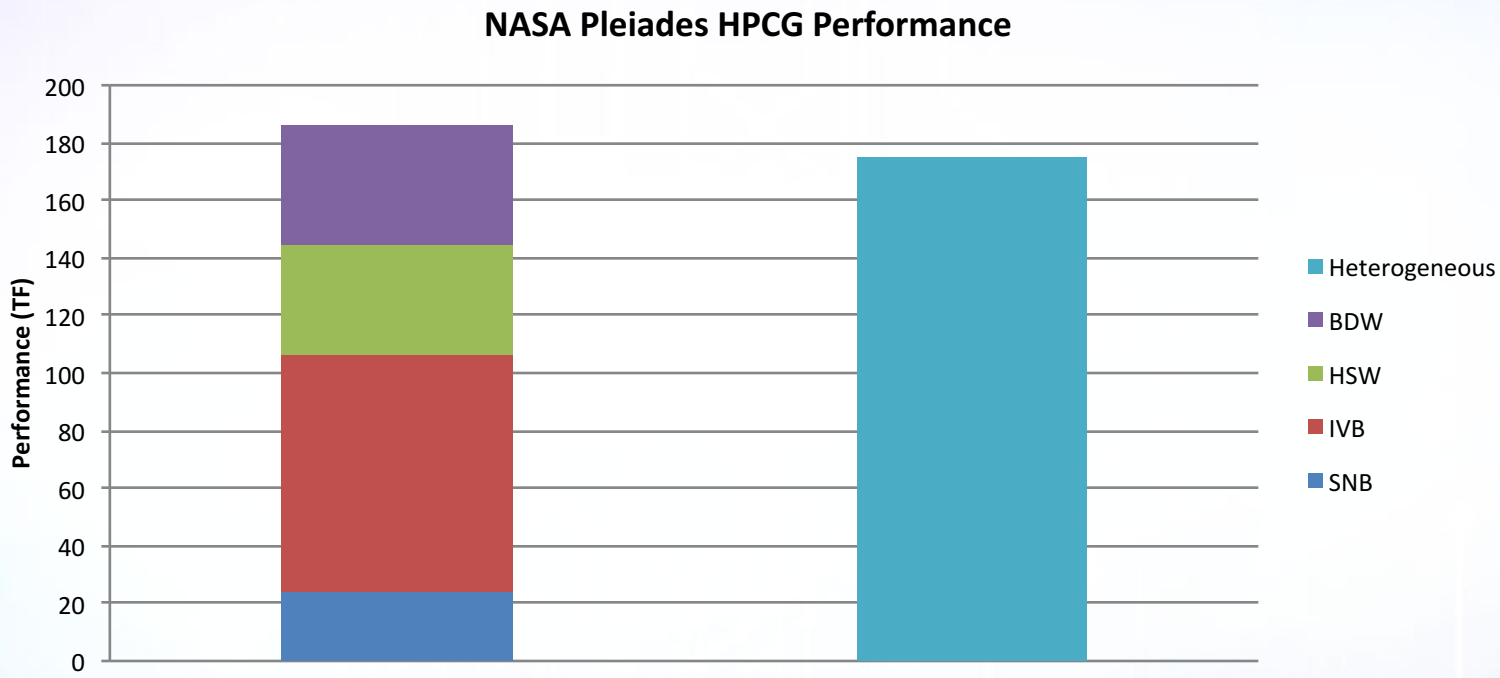
Ivy Bridge E5-2680 v2 10-core 2.8 GHz

- 5351 nodes, 9 ranks/socket, 0.86 GF/rank

Sandy Bridge E5-2670 8-core 2.6 GHz

- 1853 nodes, 7 ranks/socket, 0.92 GF/rank

Pleiades results



Aggregate performance is over 94% of the sum of the individual component results

Credits to the Team

John Baron SGI
Cheng Laio SGI
Michael Raymond SGI
Jay Lan SGI
Scott Emery SGI
Jennifer Fung SGI
Jose Rodriguez SGI
Matt Lepp SGI
Jason Inoue SGI
Rich Davila SGI
John Dugan SGI

Davin Chan CSRA
Dale Talcott CSRA
Jim Karella CSRA
Greg Matthews CSRA
Herbert Yeung CSRA
Mahmoud Hanafi CSRA
Mike Hartman CSRA
Jeff Becker CSRA
Nathan Dauchy CSRA
Bill Thigpen NASA
Mark Tangney NASA
Bob Ciotti NASA