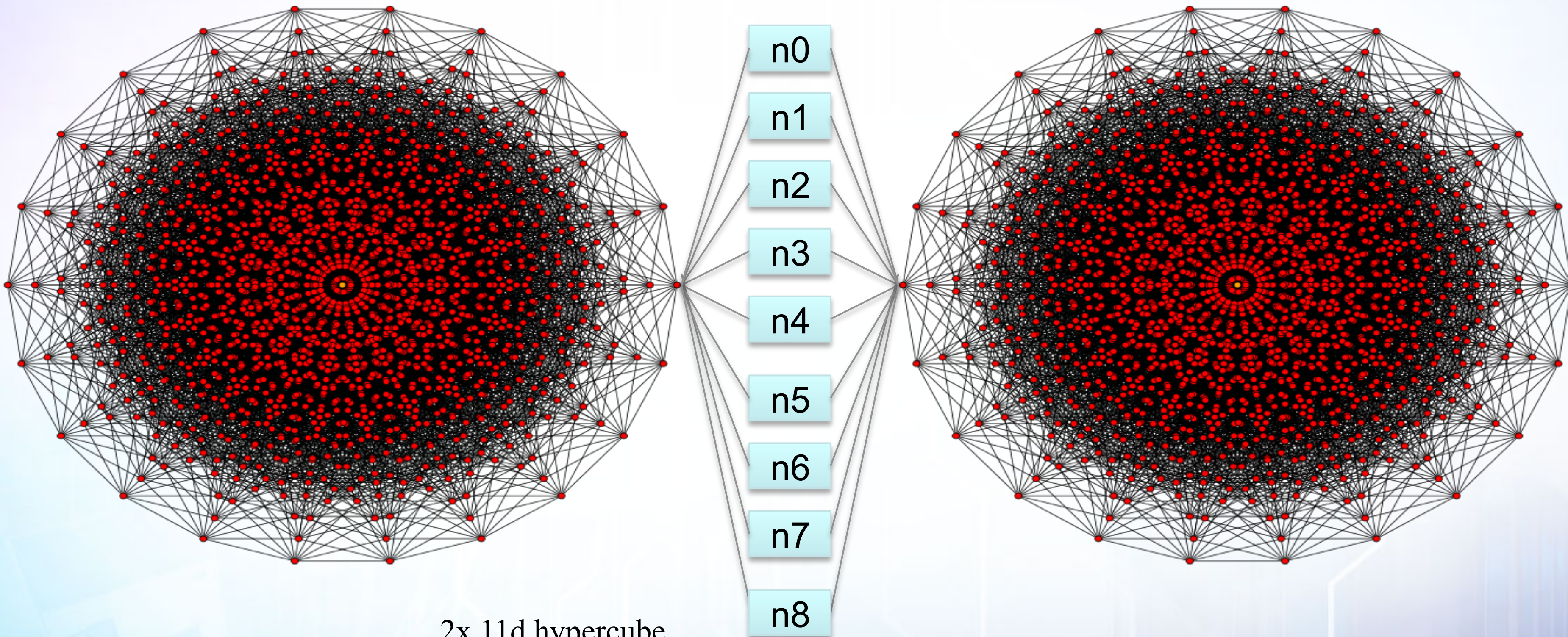
National Aeronautics and Space Administration

HPCG/HPL on Pleiades

Bob.Ciotti@nasa.gov
Chief Systems Architect
NASA Ames Research Center

www.nasa.gov

SGI ICE Dual Plane – Topology



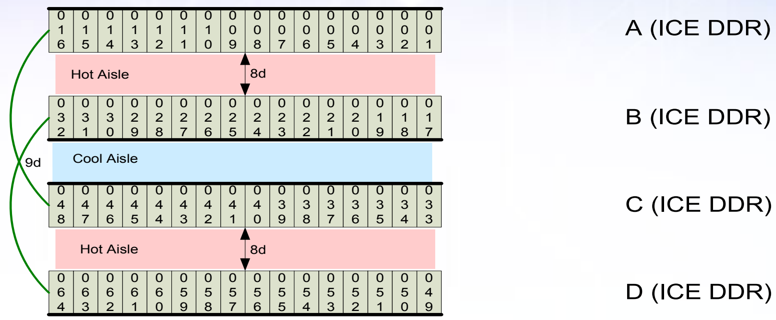
ib0

ib1

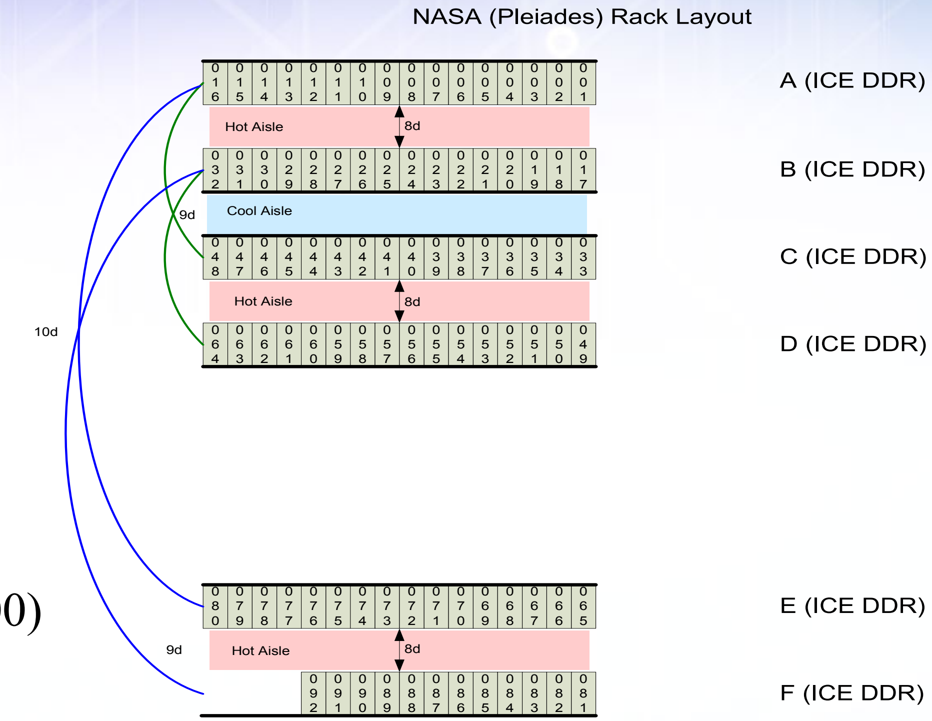
2x 11d hypercube
full 11d == 2048 vertices
Pleiades – partial 11d - 1336 vertices (2672 across both cubes)

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

NASA (Pleiades) Rack Layout

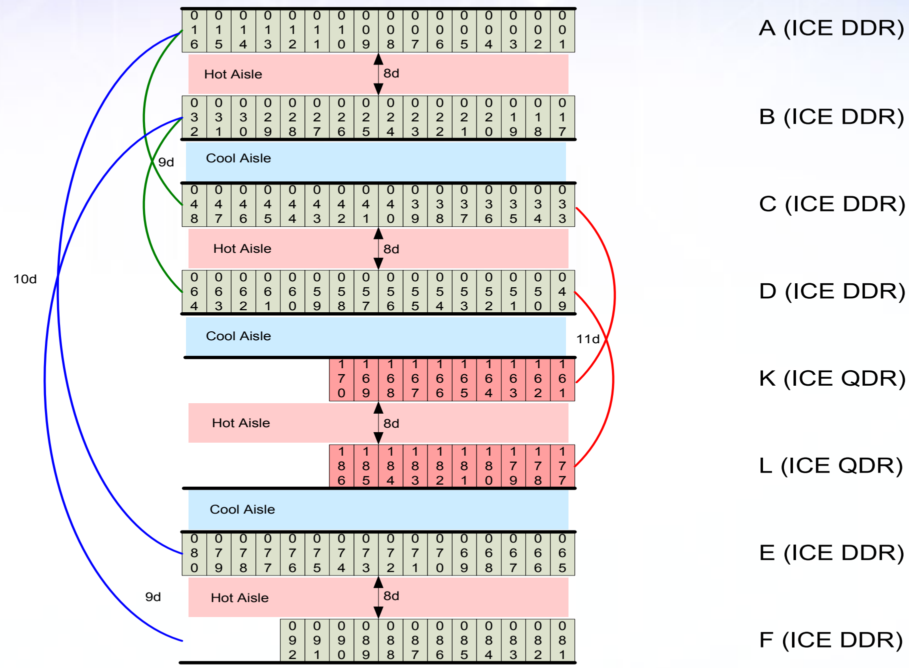


64 racks – 2008
393 teraflops



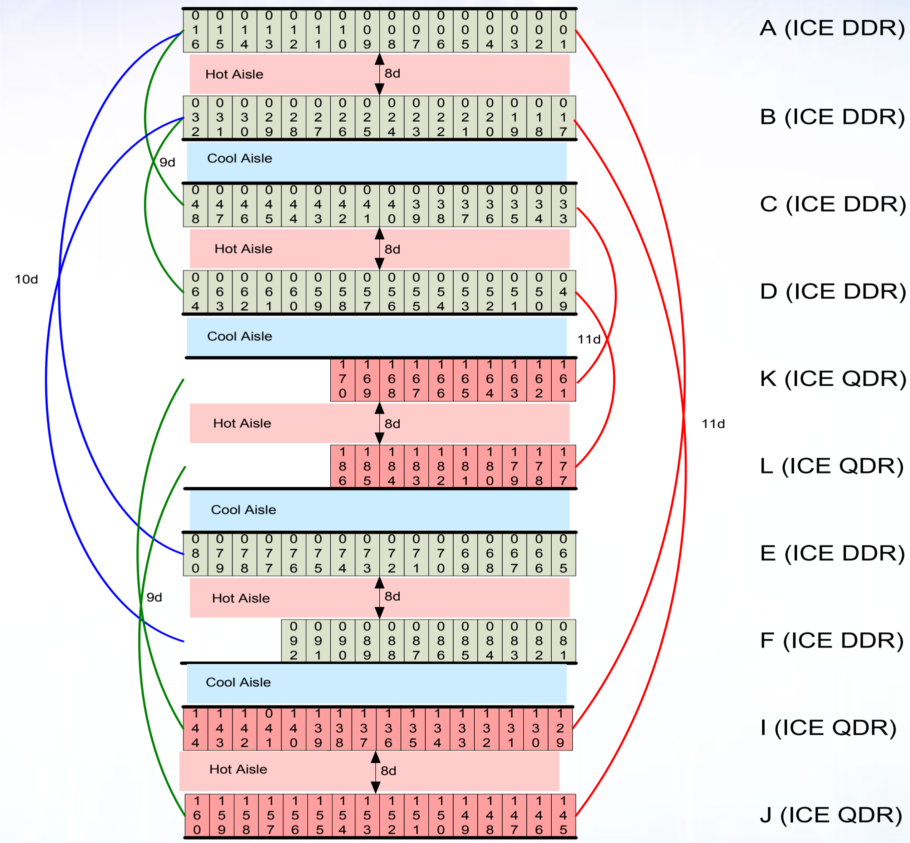
92 racks – 2008
565 teraflops (#3 t500)

NASA (Pleiades) Rack Layout

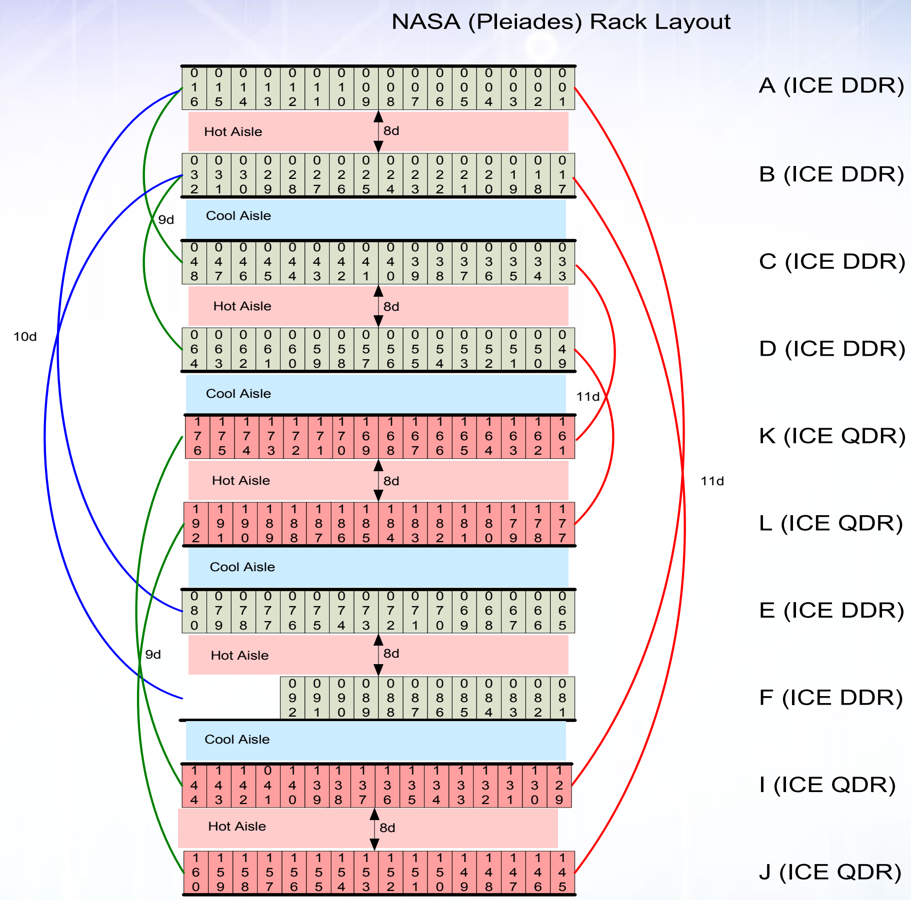


112 racks – 2009
683 teraflops

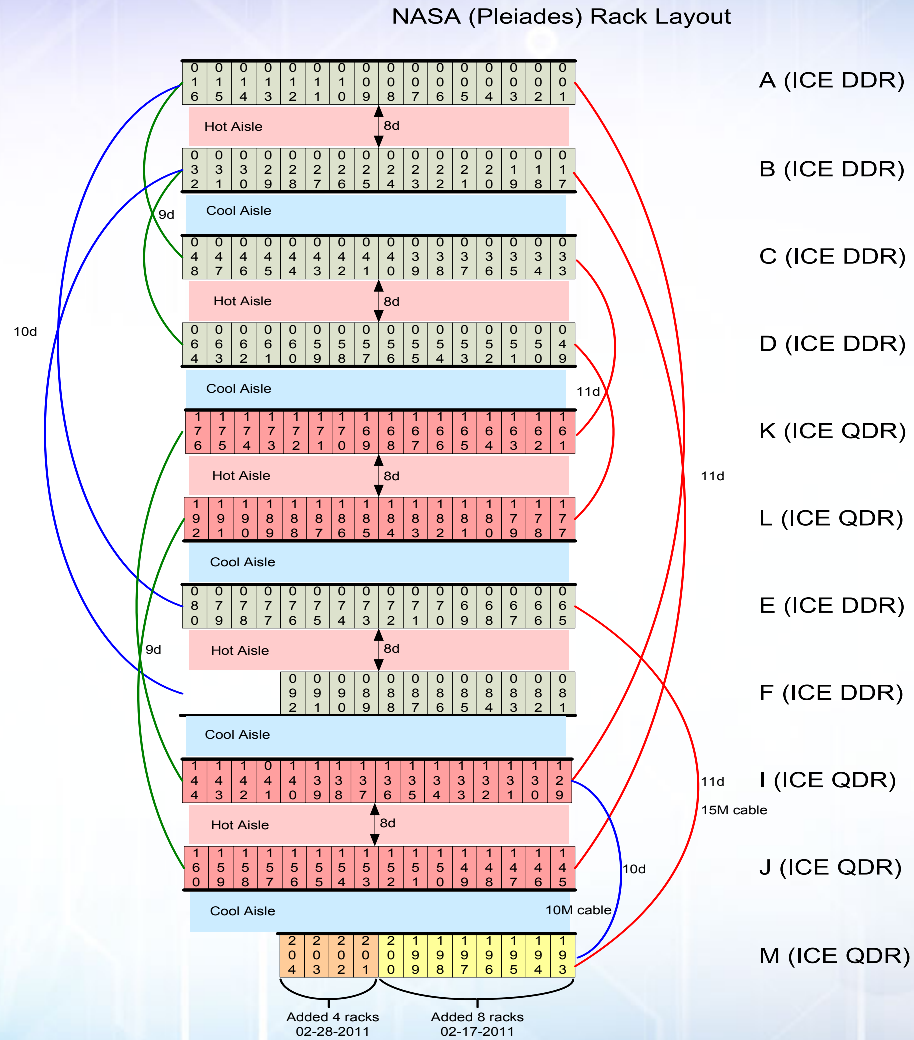
NASA (Pleiades) Rack Layout



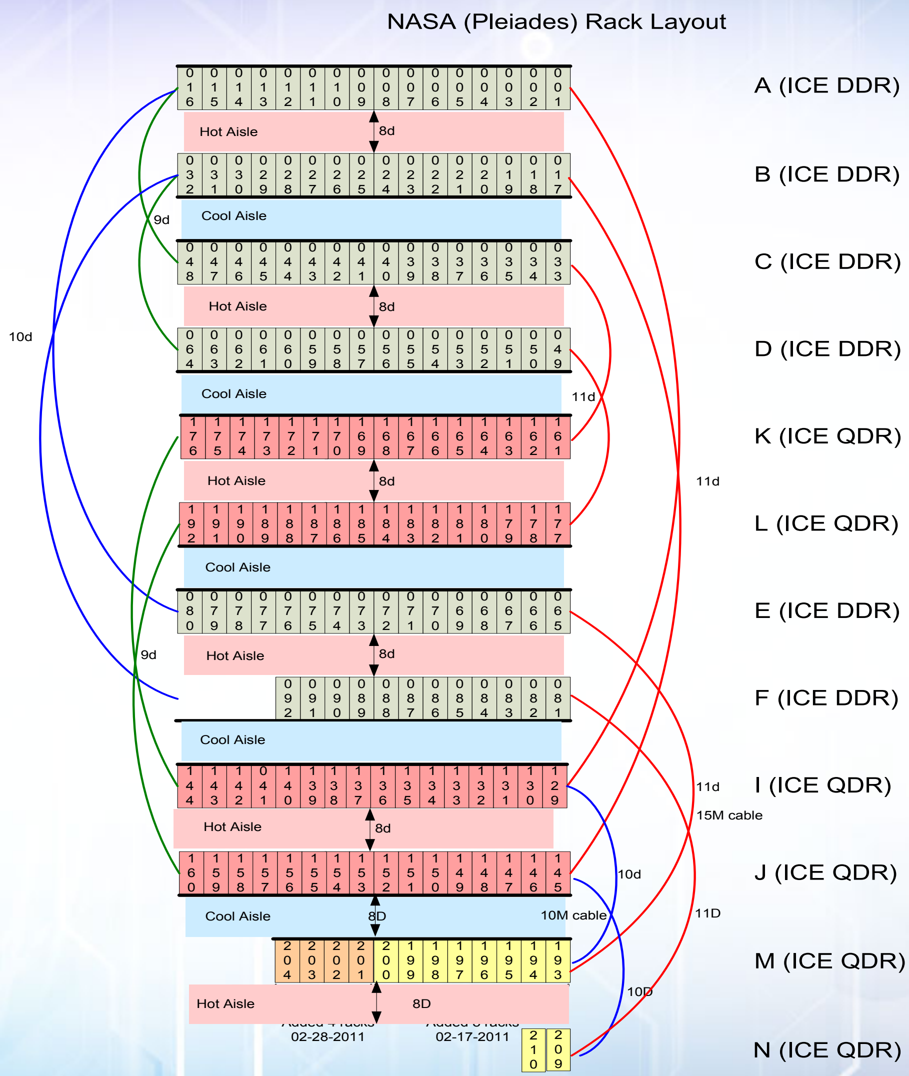
144 racks – 2010
969 teraflops



156 racks – 2010
1.08 petaflops

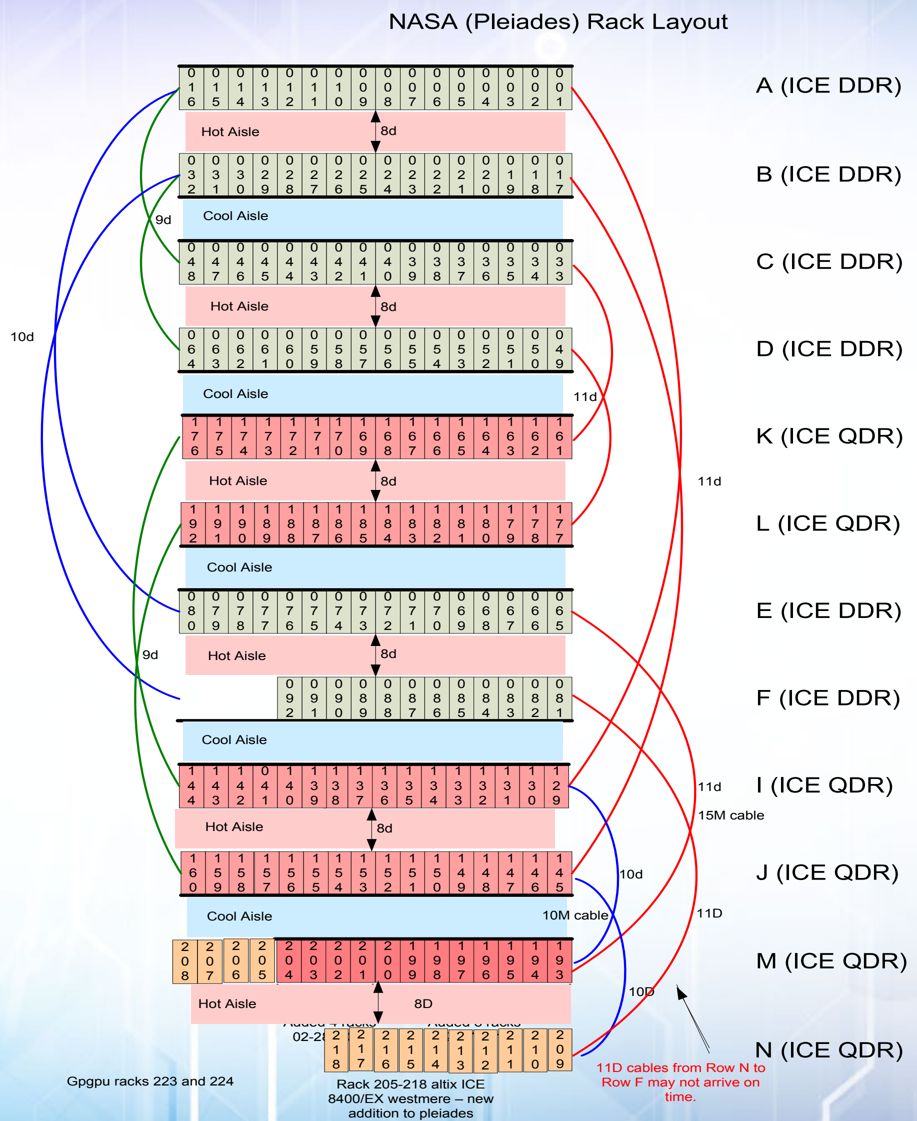


168 racks – 2011
1.18 petaflops

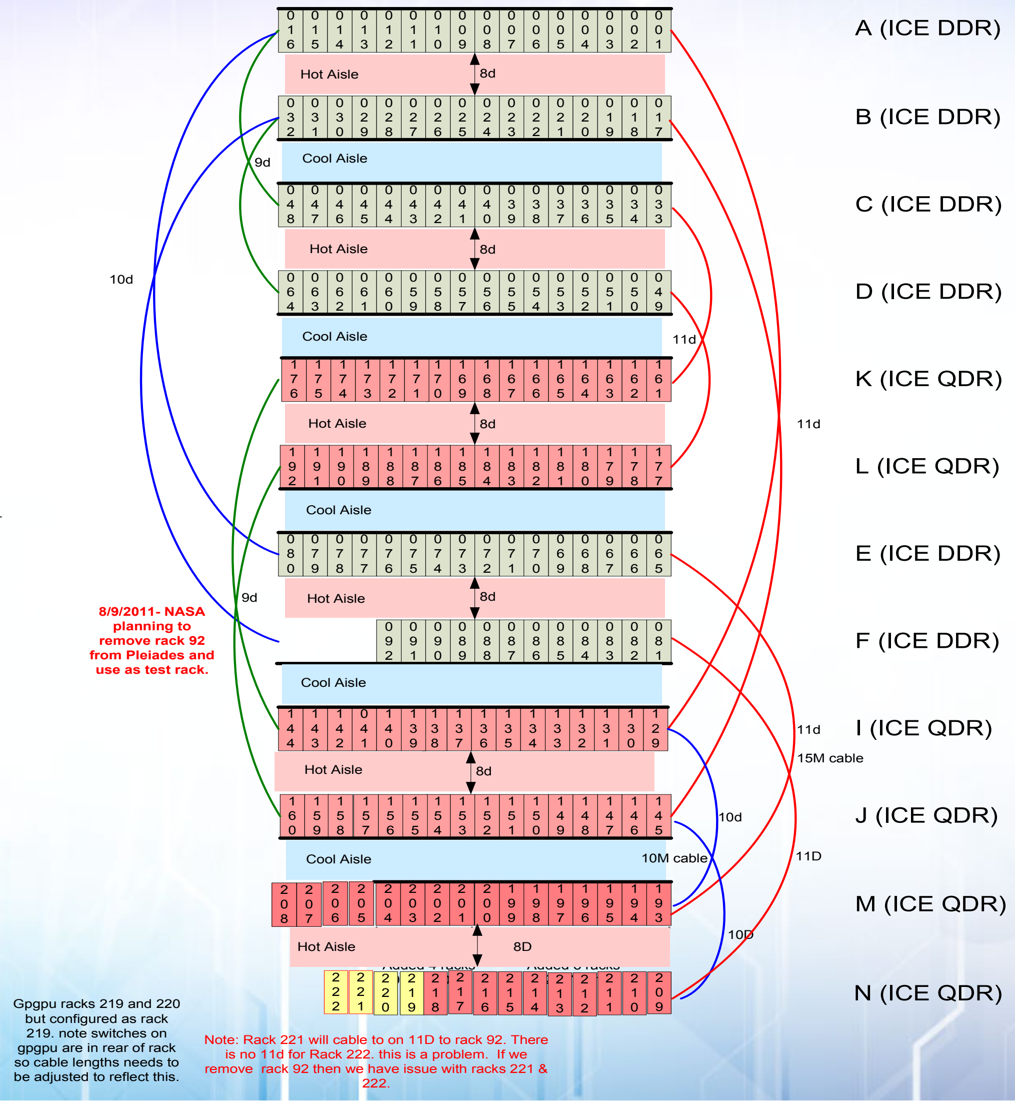


170 racks – 2011
1.20 petaflops

182 racks – 2011
1.31 petaflops



NASA (Pleiades) Rack Layout



186 racks – 2011
1.33 petaflops

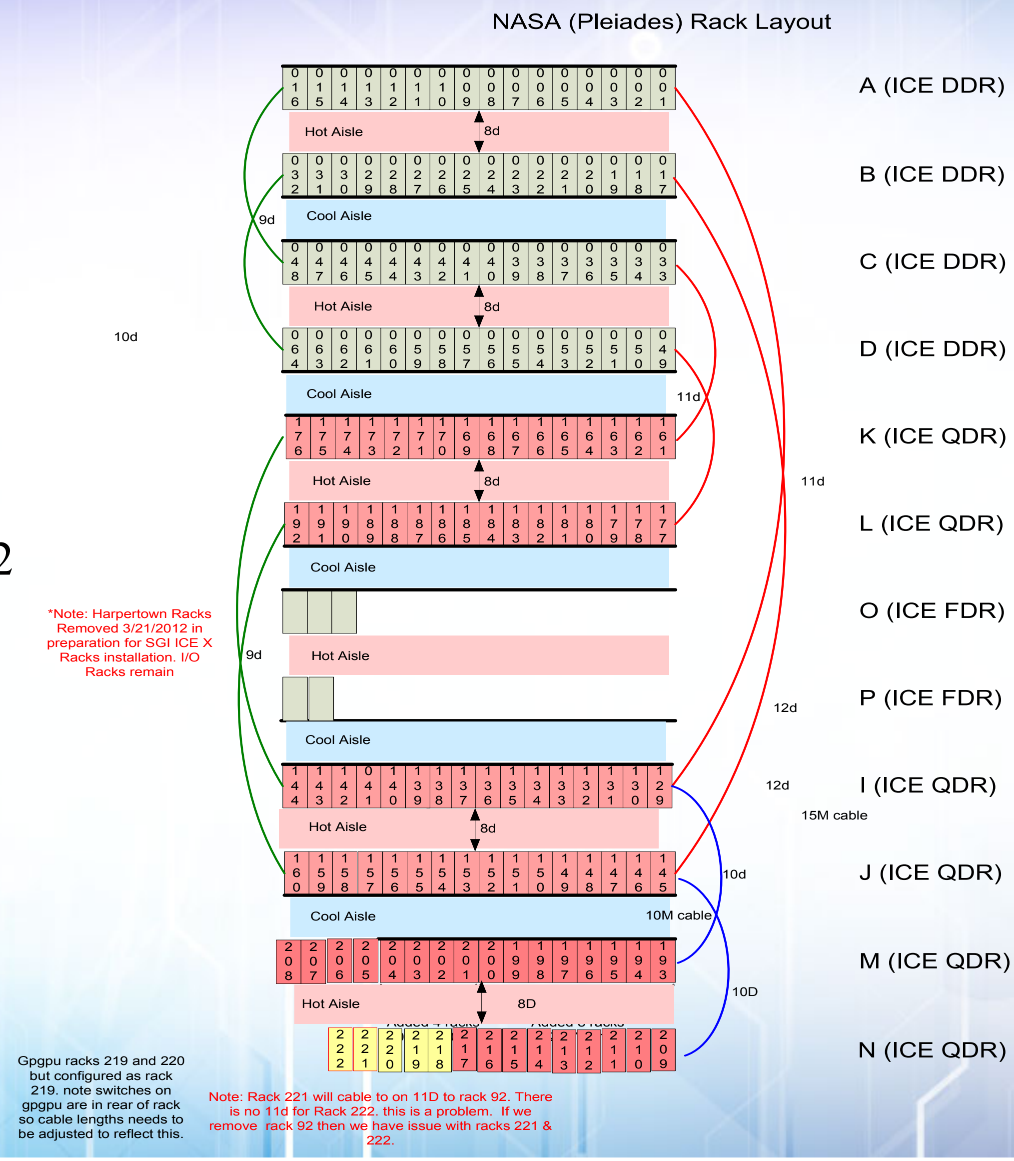
Pleiades - Sustained SpecFP rate base (2011 timeframe)

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

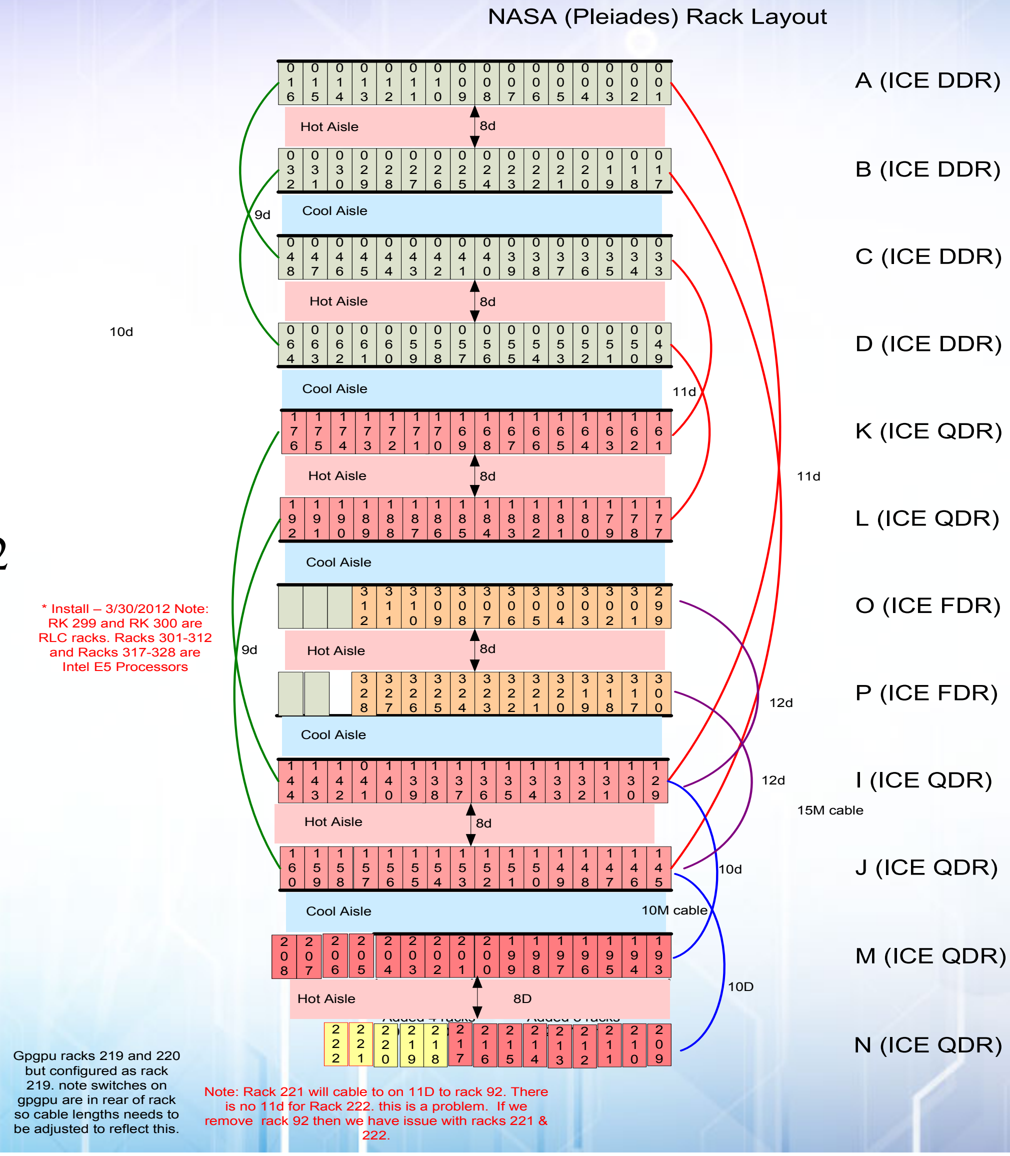
Spec	Top500	Machine	CPU	#Sockets	FPR/Socket	TSpec
1	2	Jaguar	AMD-2435	37,360	65.2	2,436,246
2	**6**	Tera-100	Intel-7560	17,296	133.4	2,307,805
3	**5**	Hopper	AMD-6176	12,784	149.8	1,800,115
4	**1**	Tianhe-1a	Intel-x5670	14,336	119.5	1,713,868
5	**11**	**Pleiades**	**Intel-x**	**21,632**	**72.2**	**1,562,510**
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

158 racks – 2012
1.15 petaflops
deinstall

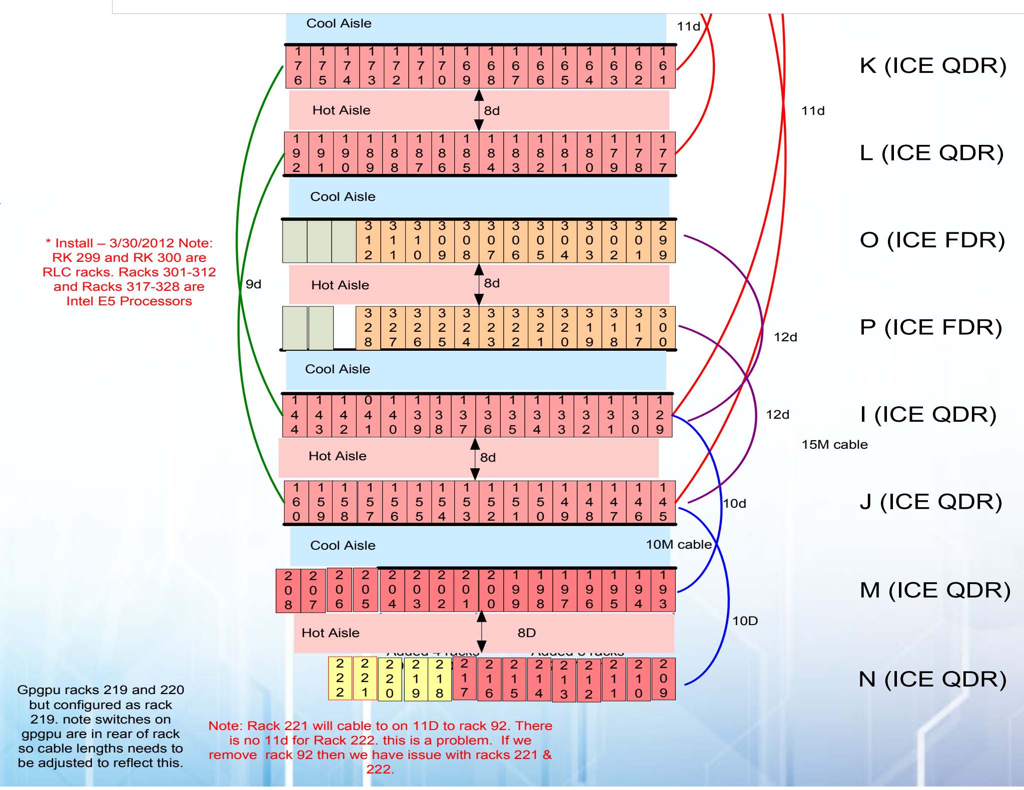


182 racks – 2012
1.7 petaflops



NASA (Pleiades) Rack Layout

64 rack deinstall 2013



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.

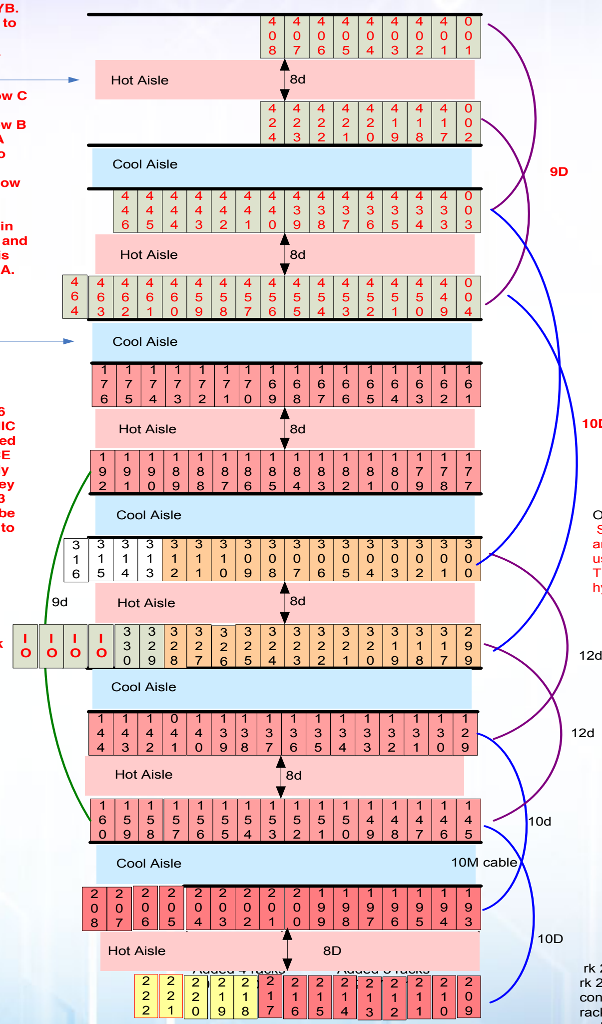
Note:
1st delivery : 8 racks in Row C
8 racks Row D
2nd delivery: 8 racks in Row B
8 racks in Row A
3rd delivery: add 8 racks to row D
add 6 racks to Row C

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

* 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

This is the switch rack

167 racks – 2013
2.9 petaflops



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

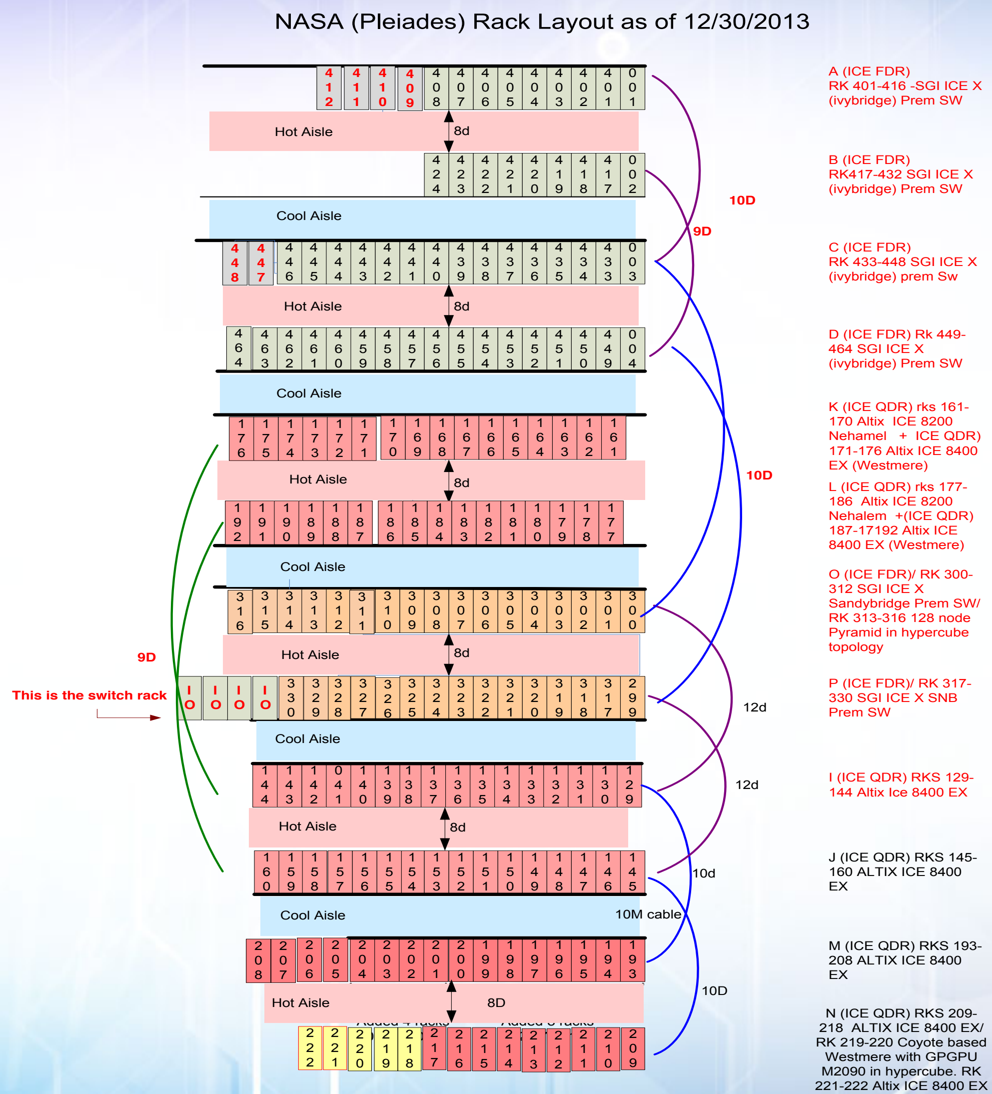
M (ICE QDR)
rk 193-208 are Altix ICE 8400 EX – all QDR

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.

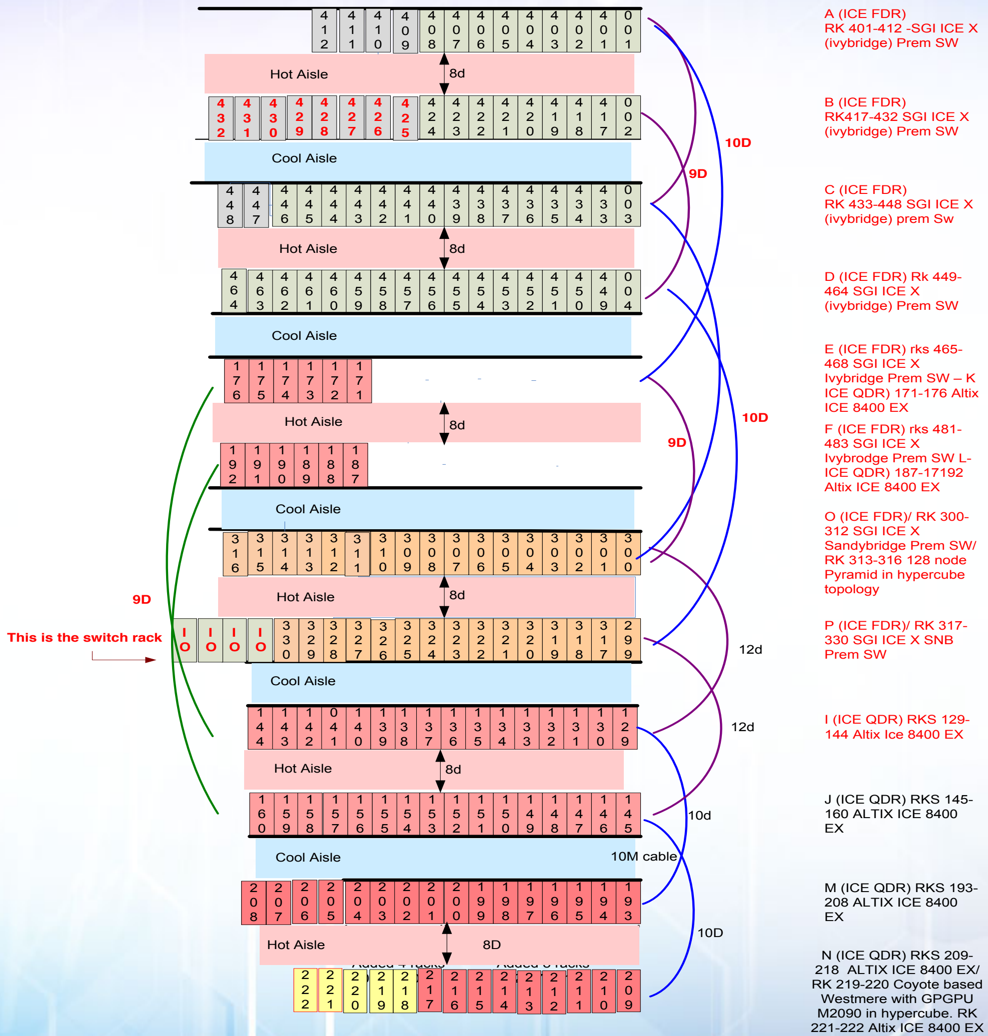
Note: Rack 221 will cable to on 11D to rack 92. There is no 11d for Rack 222. this is a problem. If we remove rack 92 then we have issue with racks 221 & 222.

160 racks – 2013
3.1 petaflops

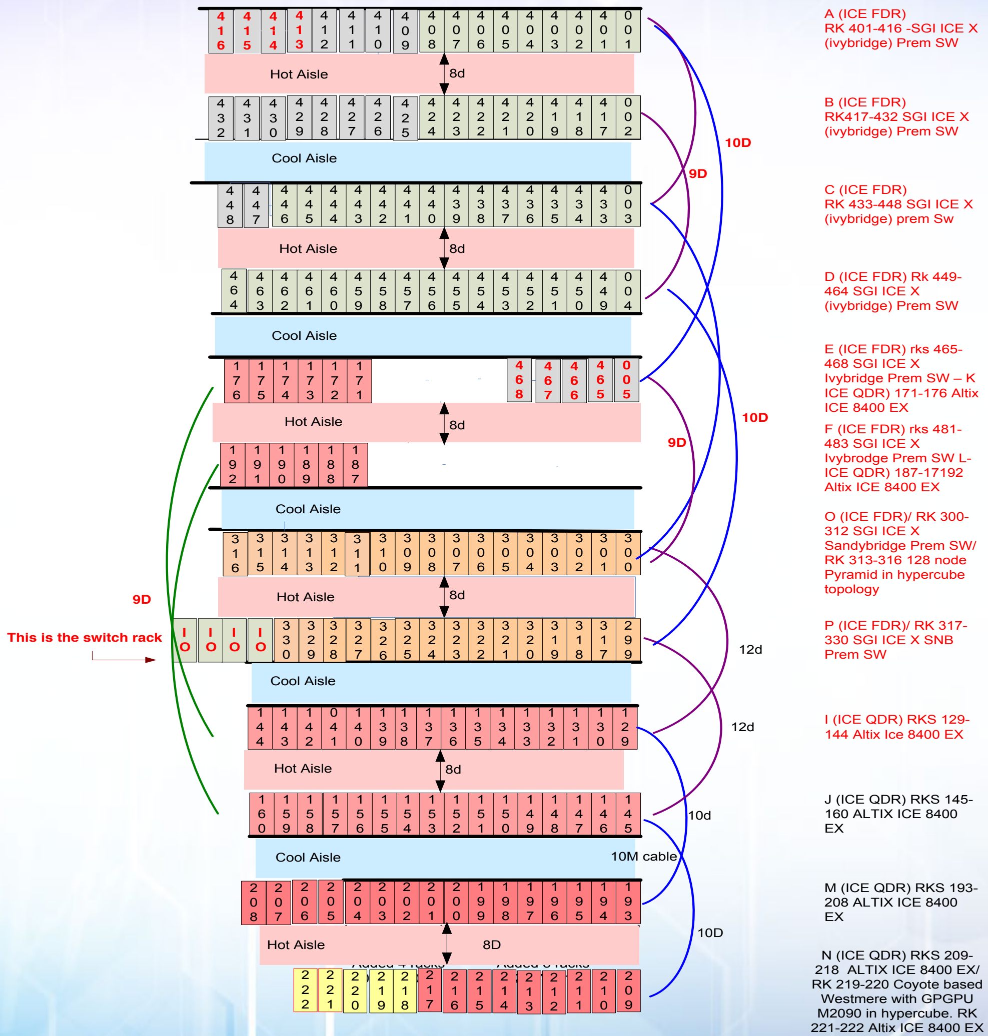


NASA (Pleiades) Rack Layout as of 1/30/2014

168 racks – 2013
3.2 petaflops

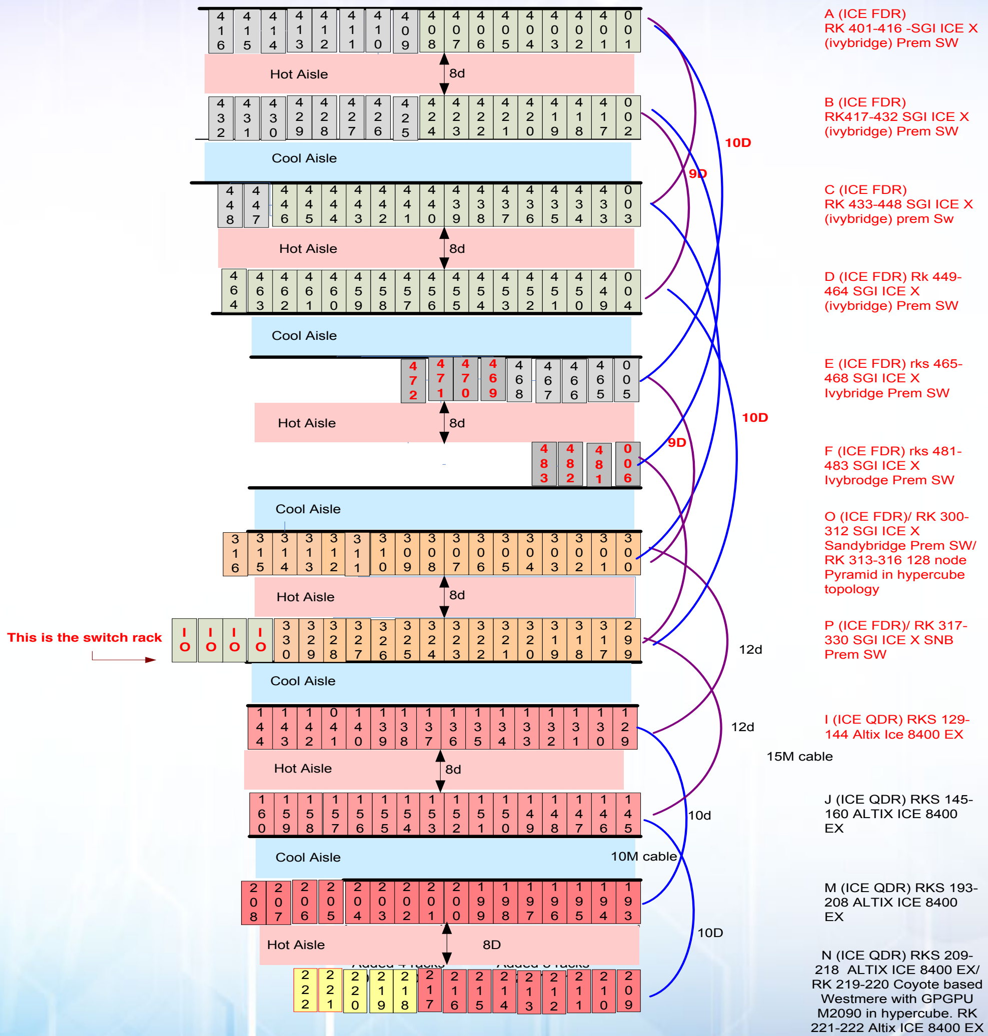


NASA (Pleiades) Rack Layout as of 2/18/2014



168 racks – 2014
3.3 petaflops

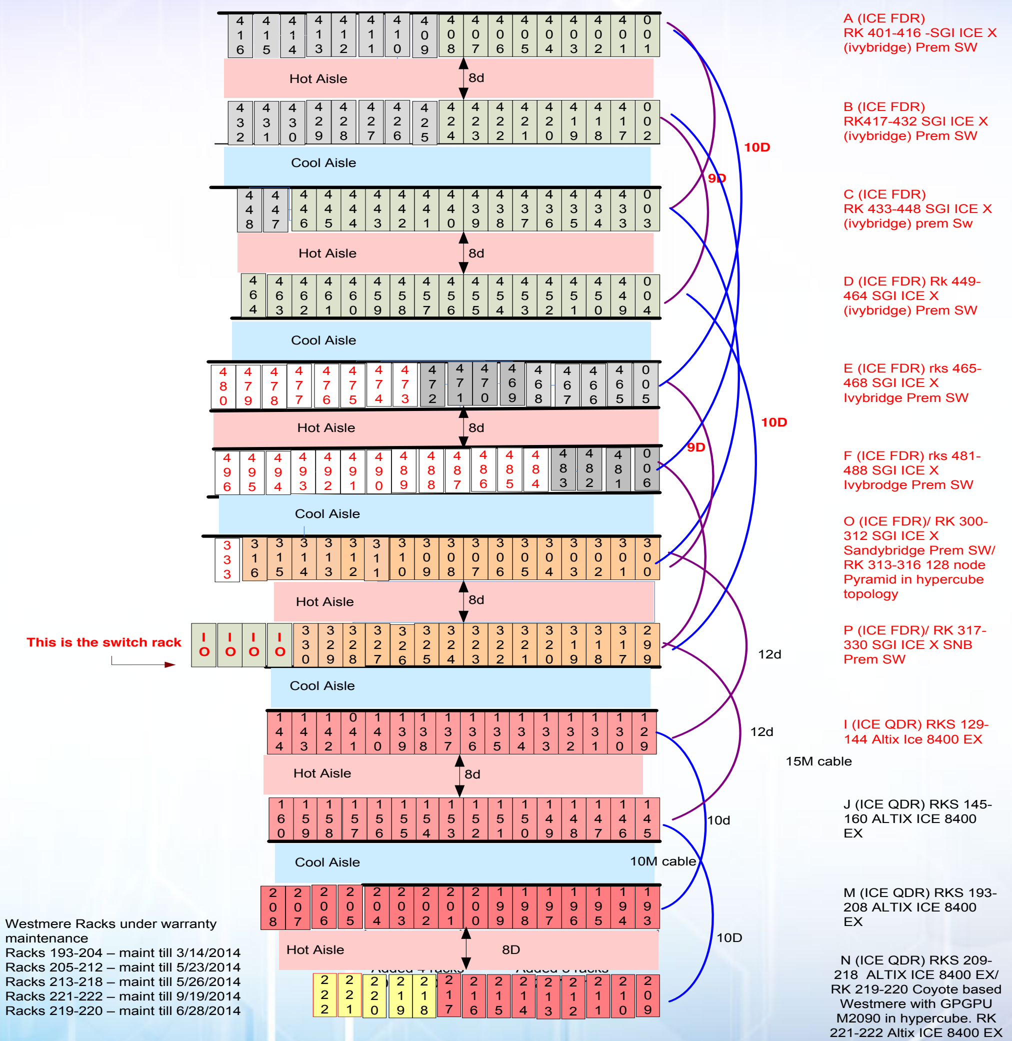
NASA (Pleiades) Rack Layout as of 2/25/2014



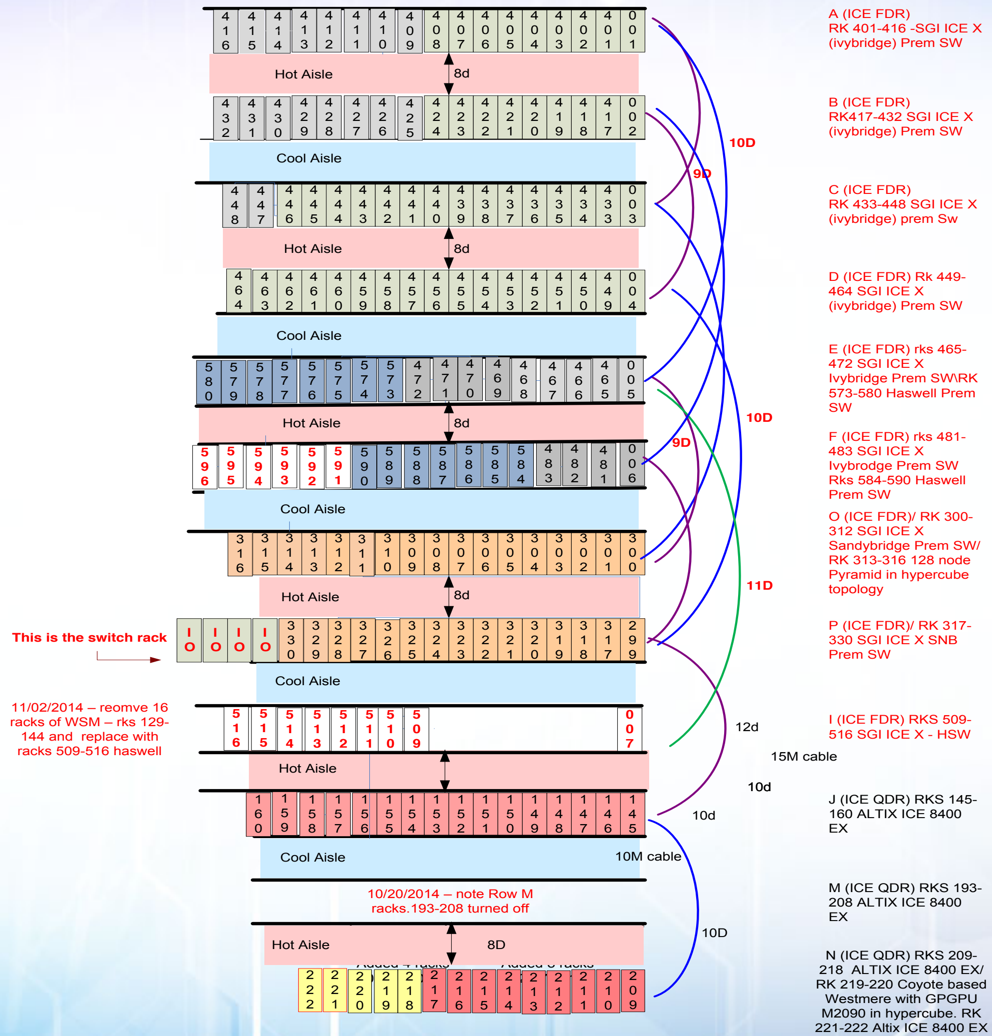
170 racks – 2014
3.5 petaflops

NASA (Pleiades) Rack Layout

168 racks – 2014
4.5 petaflops



NASA (Pleiades) Rack Layout



168 racks – 2015

5.4 petaflops

Pleiades 2015 – Based on MemoryBW (ignore GPU/PHI)

Machine	Type	11/14 T500	Sockets	Type	Mem BW Socket	Spec Socket	Mem BW (PB/Sec)	Mega Spec	Rmax	Rpeak	PctPeak
K computer	Sparc64	4	88,128	VIII fx	64.0	373.2	5,640	32.9	10,510	11,280	93.2%
Sequoia	BGQ/Power	3	98,304	BGQ-A2	42.7	144.3	4,198	14.2	17,173	20,132	85.3%
BlueWater	XK6/XK7		49,200	6276	51.2	176.0	2,519	8.7		71,378	
Mira	BGQ /Power	5	49,152	BGQ-A2	42.7	144.3	2,099	7.1	8,586	10,066	85.3%
Tianhe-2	Xeon/Xeon Phi	1	32,000	E5-2692v2	59.7	321.5	1,910	10.3	33,862	54,902	61.7%
Pleiades	**SGI/Xeon Mix**	**11**	**22,896**	**XeonMix**	**54.8**	**283.7**	**1,255**	**6.5**	**3,375**	**3,987**	**84.7%**
Juqueen	BGQ/Power	8	28,672	BGQ-A2	42.7	144.3	1,224	4.1	5,008	5,872	85.3%
Secret2	XC30/Xeon	13	18,832	E5-2697v2	59.7	341.0	1,124	6.4	3,143	4,881	64.4%
Vulcan	BGQ/Power	9	24,576	BGQ-A2	42.7	144.3	1,049	3.5	4,293	5,033	85.3%
Titan	XK7/Opteron/K20x	2	18,688	6274	51.2	173.0	957	3.2	17,590	27,112	64.9%
SuperMUC	iData/Xeon	14	18,432	E5-2680	51.2	244.5	944	4.5	2,897	3,185	91.0%
Pangea	SGI/Xeon	20	13,800	E5-2670	51.2	240.5	707	3.3	2,098	2,296	91.4%
Stampede	Dell/Xeon/Phi	7	12,800	E5-2680	51.2	244.5	655	3.1	5,168	8,520	60.7%
Hornet	XC40/Xeon	16	7,884	E5-2680v3	68.0	396.5	536	3.1	2,763	3,784	73.0%
Tianhe-1A	Xeon/Nvidia2050	17	14,336	X5670	32.0	132.0	459	1.9	2,566	4,701	54.6%
Secret1	CS/Xeon/K40	10	7,280	E5-2660v2	59.7	287.5	435	2.1	3,577	6,131	58.3%
HPC2	iData/Xeon/K20x	12	7,200	E5-2680v2	59.7	313.0	430	2.3	3,188	4,605	69.2%
Excalibur	XC40/Xeon	19	6,254	E5-2698v3	68.0	434.0	425	2.7	2,485	3,682	67.5%
Piz Daint	XC30/Xeon/K20x	6	5,272	E5-2670 snb	51.2	240.5	270	1.3	6,271	7,788	80.5%
Cascade	Xeon/Xeon Phi	18	1,880	E5-2670	51.2	240.5	96	0.5	2,539	3,388	74.9%
Tsubame	Nec/Xeon/K20x	15	2,816	X5670	32.0	132.0	90	0.4	2,785	5,735	48.6%

Numbers in Red are sWAG

Pleiades Environment

- 11,280 compute nodes – 22,560 sockets - 211,360 x86 cores
  - Westmere, Sandybridge, Ivybridge, Haswell
- 128 visualization nodes
- 192 GPU Nodes
- 192 Xeon Phi Nodes
- 10 Front End Nodes
- 4 “Bridge Nodes”
- 4 Archive Front Ends
- 8 Data Analysis Nodes
- 8 Archive Nodes
- 2 large memory nodes 2 TB + 4 TB

- + a couple hundred administration/management nodes of various types.

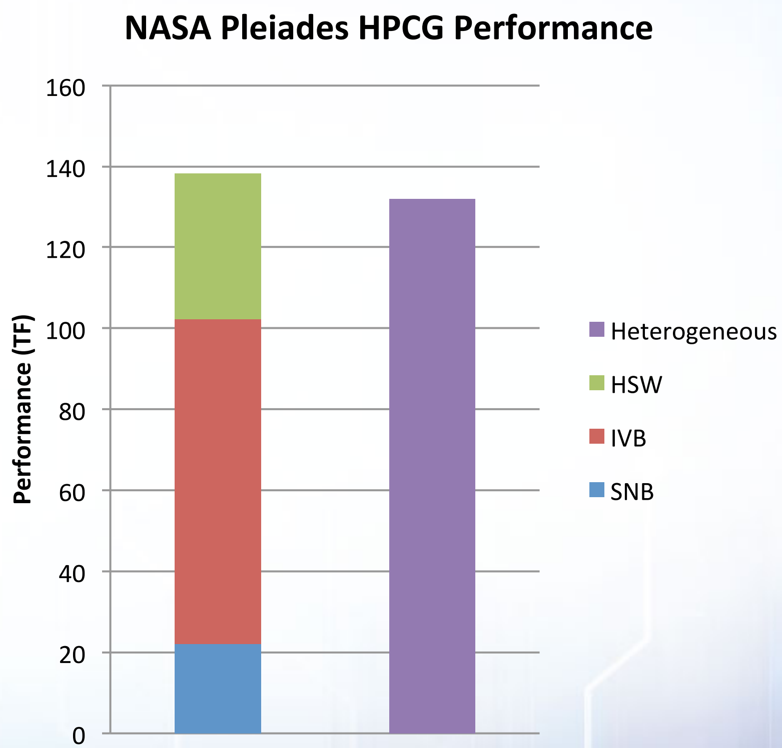
Pleiades Results

Load balancing with 1 MPI task/core

- SNB E5-2670 8c 2.6 GHz: STREAM Triad 75.9 GB/s/node, **4.74** GB/s/core
- IVB E5-2680v2 10c 2.8 GHz: STREAM Triad 95.7 GB/s/node, **4.79** GB/s/core
- HSW E5-2680v3 12c 2.5 GHz: STREAM Triad 117.2 GB/s/node, **4.89** GB/s/core

Performance measurements

- SNB: 99.3% scaling efficiency from 1 to 1868 nodes
- IVB: 97.4% scaling efficiency from 1 to 5347 nodes
- HSW: 95.9% scaling efficiency from 1 to 2073 nodes



Credit - Cheng Laio - SGI

SGI Optimized HPCG Code

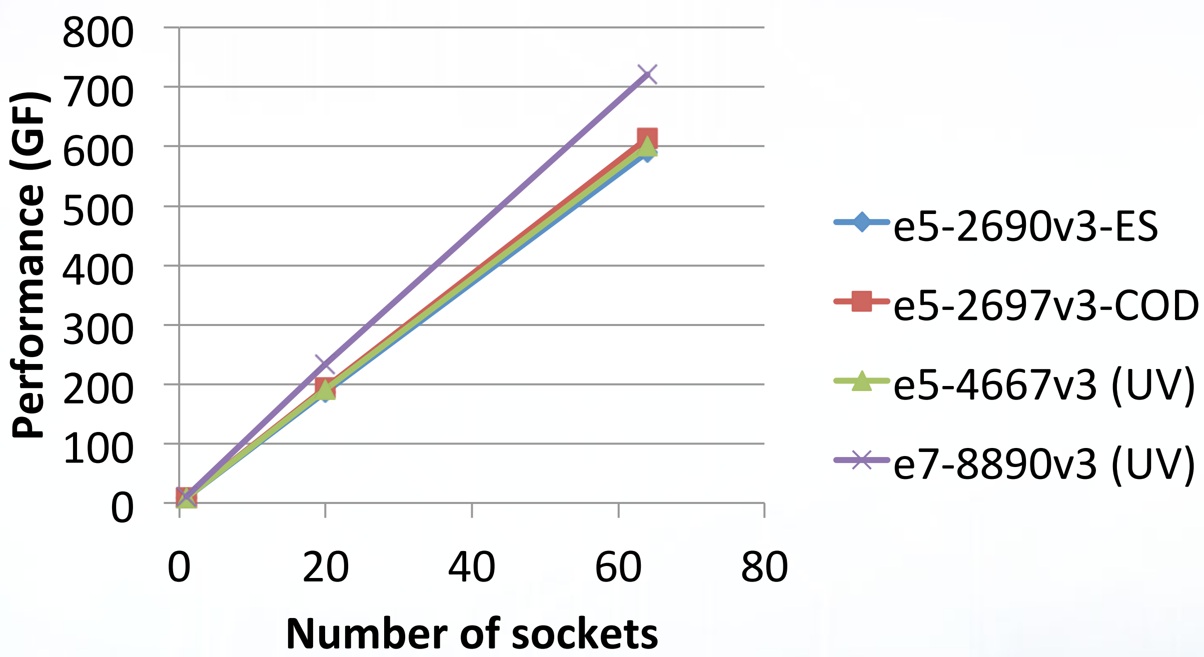
The SGI code is optimized using common techniques such as contiguous memory, storage format tuning, multi-color reordering and combined computations.

The code is pure MPI.

Improvements and Extensions are being planned.

Credit - Cheng Laio - SGI

Performance on Various Systems



Credit - Cheng Laio - SGI

HPCG June 2015

HPCG Rank (Jun 15)	#5	#1	#2	#3	#4
Top500 Rank (Nov 15)	#13	#1	#4	#2	#5
	Pleiades	Tianhe	K Computer	Titan	Mira
Cores	186,288	3,120,000	705,024	560,640	786,432
HPCG PF	0.131	0.580	0.461	0.322	0.167
HPL PF	4.089	33.863	10.51	17.59	8.567
Peak PF	4.970	54.902	11.280	27.112	10.066
HPCG MF/Core	703.21	185.90	653.59	574.88	212.35
HPL GF/Core	21.95	10.85	14.91	31.37	10.89
Peak GF/Core	26.68	17.60	16.00	48.36	12.80
HPCG %of HPL	3.20%	1.71%	4.38%	1.83%	1.95%
HPCG %of Peak	2.64%	1.06%	4.09%	1.19%	1.66%

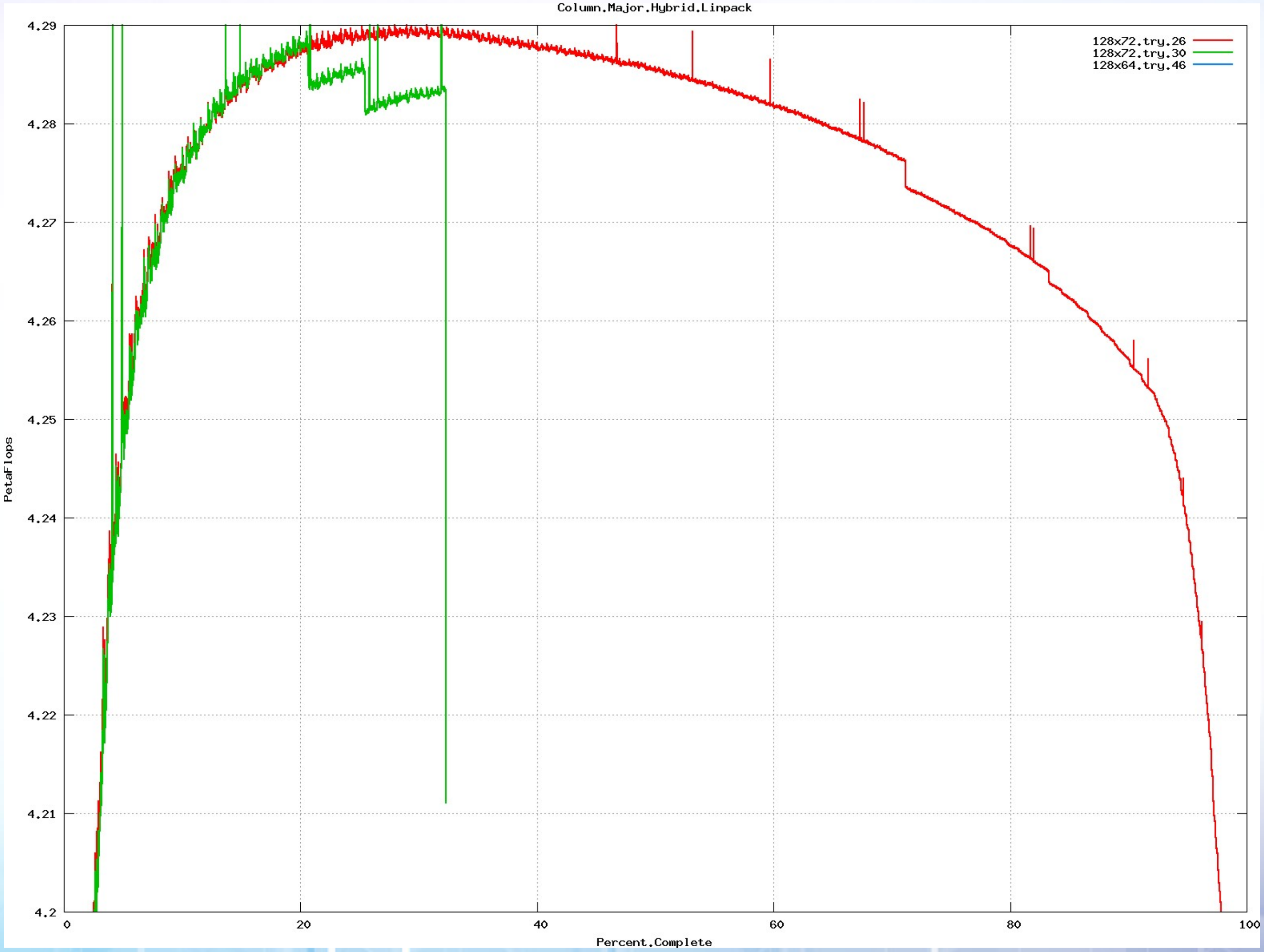
No one has built a 1 PetaFlop machine yet ☹

HPL Runtime Performance

Memory error burst or partial DIMM failure result in ~5TF performance drops or run failure on green try30.

<3TF lost on some memory errors or a single network transmission error in red try26.

Nice to modify HPCG to give similar real time performance metrics.



Fail

Node crash on DIMM issue in last seconds.

```
r509i0n0 0:
================================================================================
r509i0n0 0: T/V                N    NB     P     Q               Time                 Gflops
r509i0n0 0: --------------------------------------------------------------------------------
r509i0n0 0: WHC01L2L4     5459520   192   128    72           26568.92            4.08318e+06
r509i0n0 0: HPL_pdgesv() start time Thu Jun  4 23:07:09 2015
r509i0n0 0:
r509i0n0 0: HPL_pdgesv() end time   Fri Jun  5 06:29:58 2015
r509i0n0 0:
r509i0n0 0: --------------------------------------------------------------------------------
r509i0n0 0: ||Ax-b||_oo/(eps*(||A||_oo*||x||_oo+||b||_oo)*N)=    77939.0629011 ...... FAILED
r509i0n0 0: ||Ax-b||_oo  . . . . . . . . . . . . . . . . . =         224.874403
r509i0n0 0: ||A||_oo . . . . . . . . . . . . . . . . . . . =     1366608.843671
r509i0n0 0: ||A||_1  . . . . . . . . . . . . . . . . . . . =     1366651.513851
r509i0n0 0: ||x||_oo . . . . . . . . . . . . . . . . . . . =           3.483180
r509i0n0 0: ||x||_1  . . . . . . . . . . . . . . . . . . . =     2887333.826247
r509i0n0 0: ||b||_oo . . . . . . . . . . . . . . . . . . . =           0.500000
```

Success

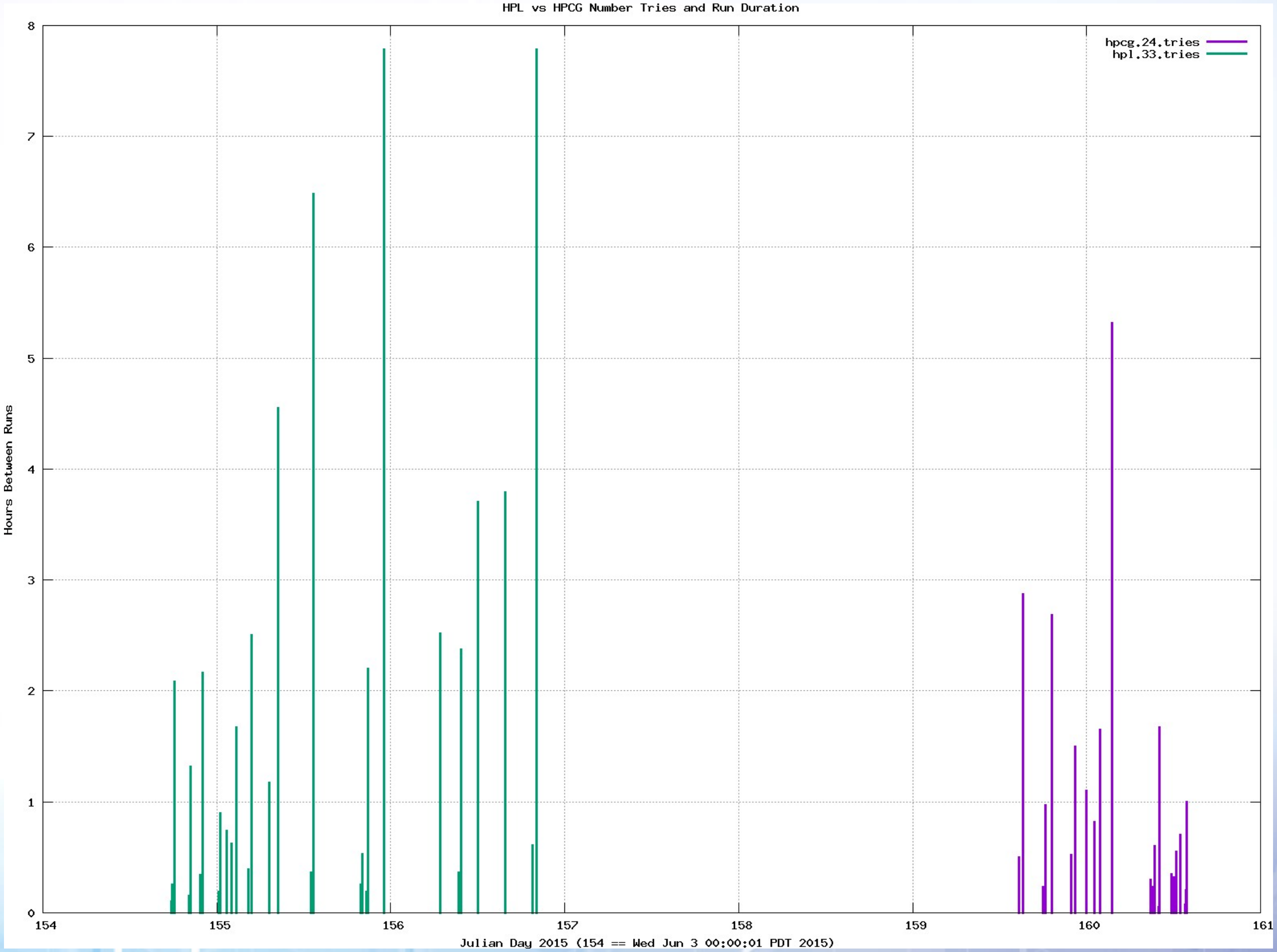
After many years of HPL, we observe that a successful run always begin late evening.

```
r509i0n0 0: ================================================================================
r509i0n0 0: T/V                N    NB     P     Q               Time                 Gflops
r509i0n0 0: --------------------------------------------------------------------------------
r509i0n0 0: WHC01L2L4      5459520   192   128    72           26528.30            4.08943e+06
r509i0n0 0: HPL_pdgesv() start time Fri Jun  5 20:14:31 2015
r509i0n0 0:
r509i0n0 0: HPL_pdgesv() end time   Sat Jun  6 03:36:40 2015
r509i0n0 0:
r509i0n0 0: --------------------------------------------------------------------------------
r509i0n0 0: ||Ax-b||_oo/(eps*(||A||_oo*||x||_oo+||b||_oo)*N)=        0.0026823 ...... PASSED
r509i0n0 0: ================================================================================
r509i0n0 0:
r509i0n0 0: Finished      1 tests with the following results:
r509i0n0 0:               1 tests completed and passed residual checks,
r509i0n0 0:               0 tests completed and failed residual checks,
r509i0n0 0:               0 tests skipped because of illegal input values.
r509i0n0 0: --------------------------------------------------------------------------------
r509i0n0 0:
r509i0n0 0: End of Tests.
r509i0n0 0: ================================================================================
Sat Jun  6 03:38:47 PDT 2015
```

HPL vs HPCG Runs

Memory DIMM errors dominate failures on HPL runs. These far exceed those seen in normal production likely due to larger memory footprint and higher CPU load (temperature). HPL then exposes these errors that were latent (dormant).

HPCG retries were mostly related to debugging a network layer/MPI issue that had been occurring in production without a reproducer – until this HPCG run. This allowed us to identify and correct the issues. Some HW/memory fallout did occur when bringing up system after two days powered off while system returns to nominal operating temperature.



Summary

HPL Strengths:

- Good for burn-in, clean-up.
- Useful in finding problems.
    - SW, Processors, memory, network, building power distribution, cooling.

HPL too time consuming, skipped runs on several major upgrades.

HPCG Strengths:

- Easy to map to system
- Configurable runtime
- Useful performance information on short (<1 hr) runs.
- Also found problems – SW, HW
- More Representative of performance seen on NASA codes

Most significant issue by far: Memory DIMMS

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 CSC
Dale Talcott CSC
Jim Karella CSC
Greg Matthews CSC
Herbert Yeung CSC
Mahmoud Hanafi CSC
Mike Hartman CSC
Jeff Becker CSC
Bill Thigpen NASA
Mark Tangney NASA
Bob Ciotti NASA