National Aeronautics and Space Administration

Pleiades
Updates for 2012

Bob Ciotti

Supercomputing Systems Lead/System Architect

Open Fabrics Alliance - 2011

Facility/Mission

We are mostly users of infiniband

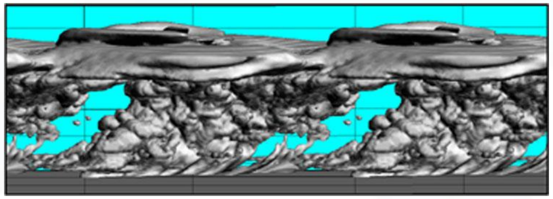
Some feature development

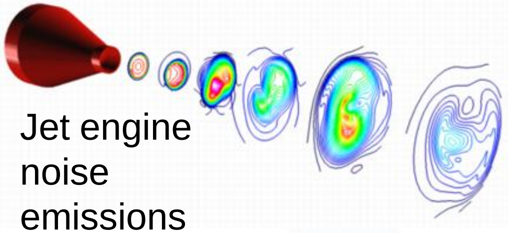
Supercomputing Support for NASA Missions

- 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

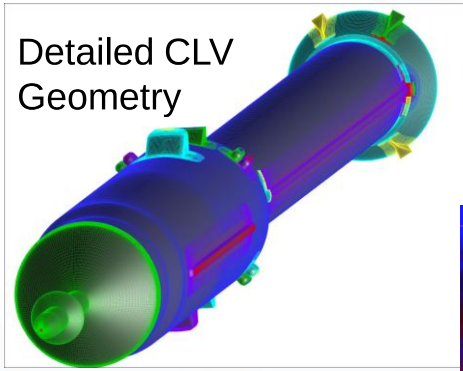
Jet aircraft wake vortices



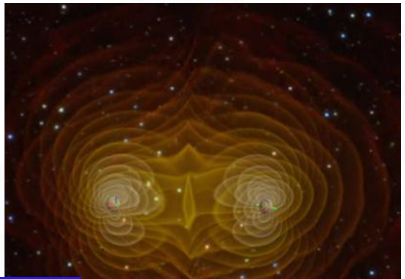


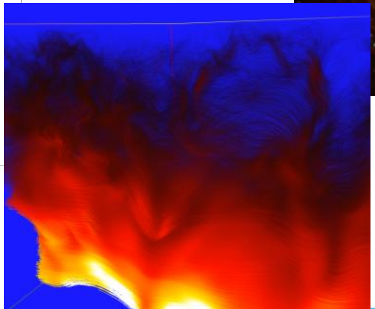
Jet engine noise emissions

Detailed CLV Geometry

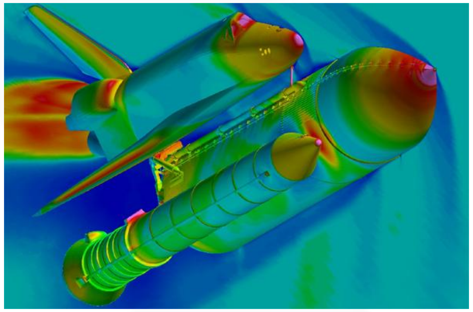


Orbiting, Spinning Black Holes



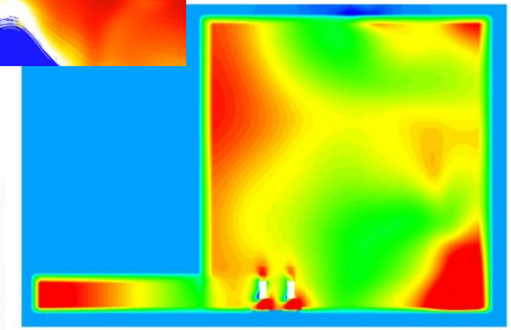


Solar surface convection

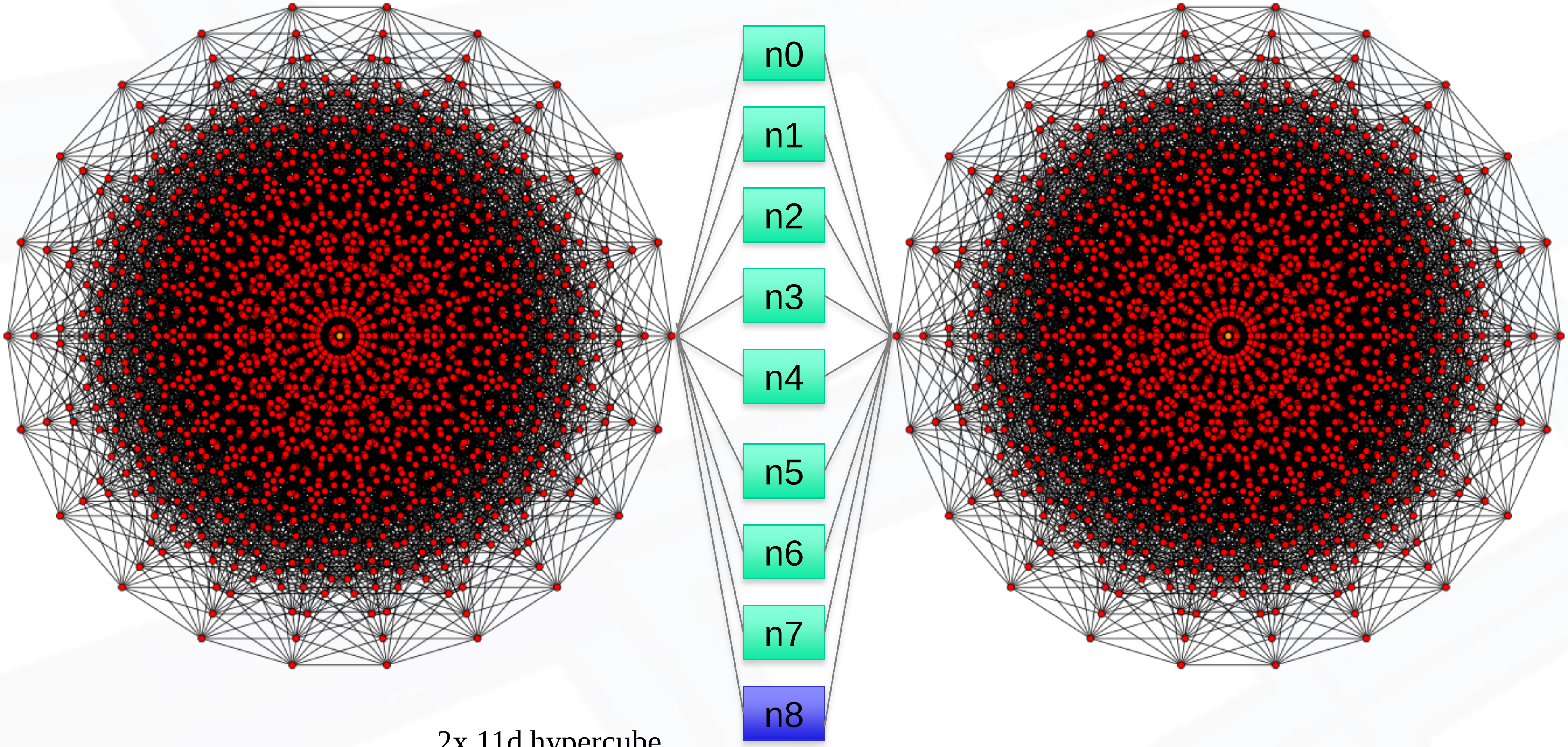


Shuttle Ascent Configuration

2-SRB Burn in VAB



SGI ICE Dual Plane – Topology



ib0

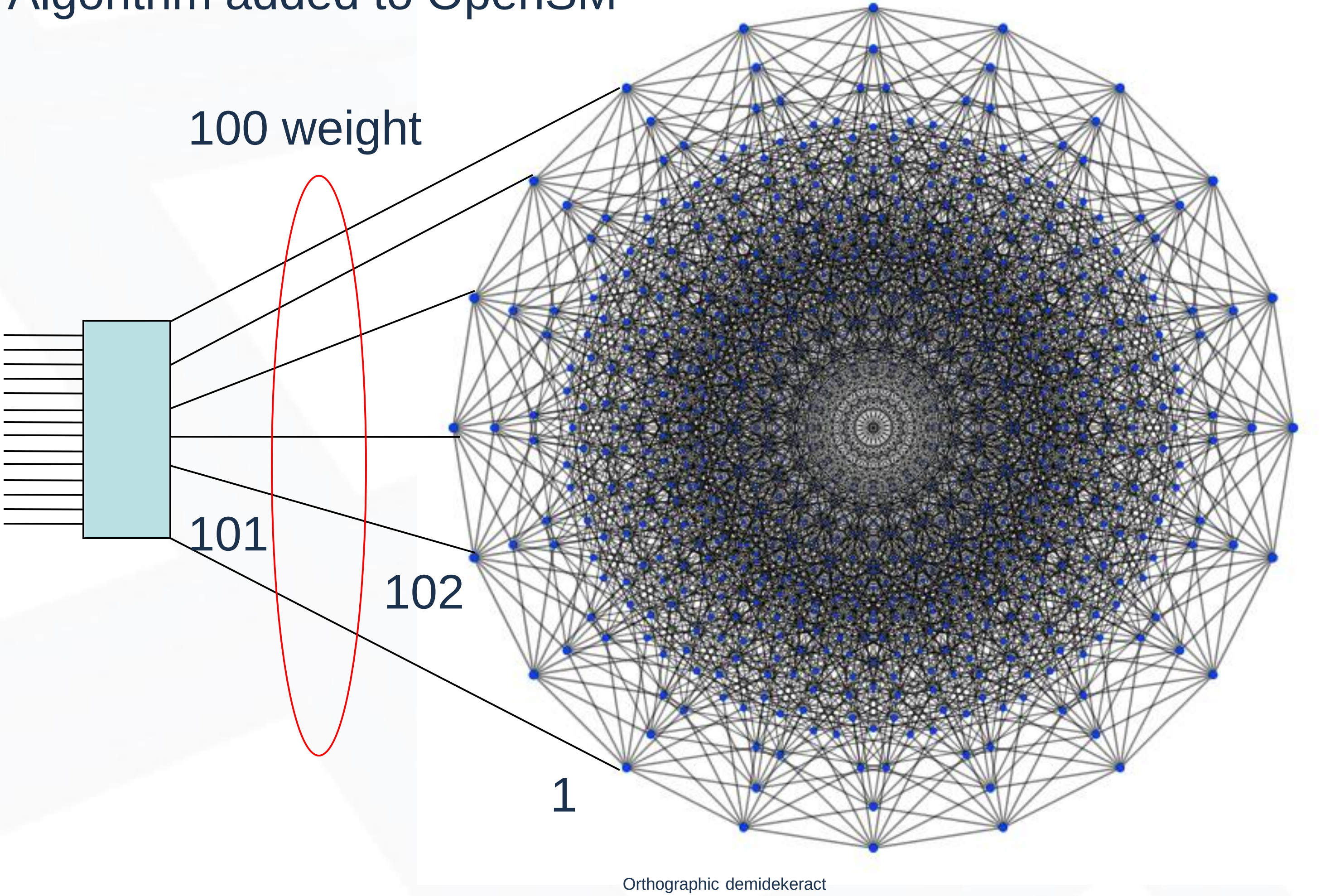
2x 11d hypercube
full 2048 vertices
Pleiades 1352/11d (2704 across both cubes)

ib1

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

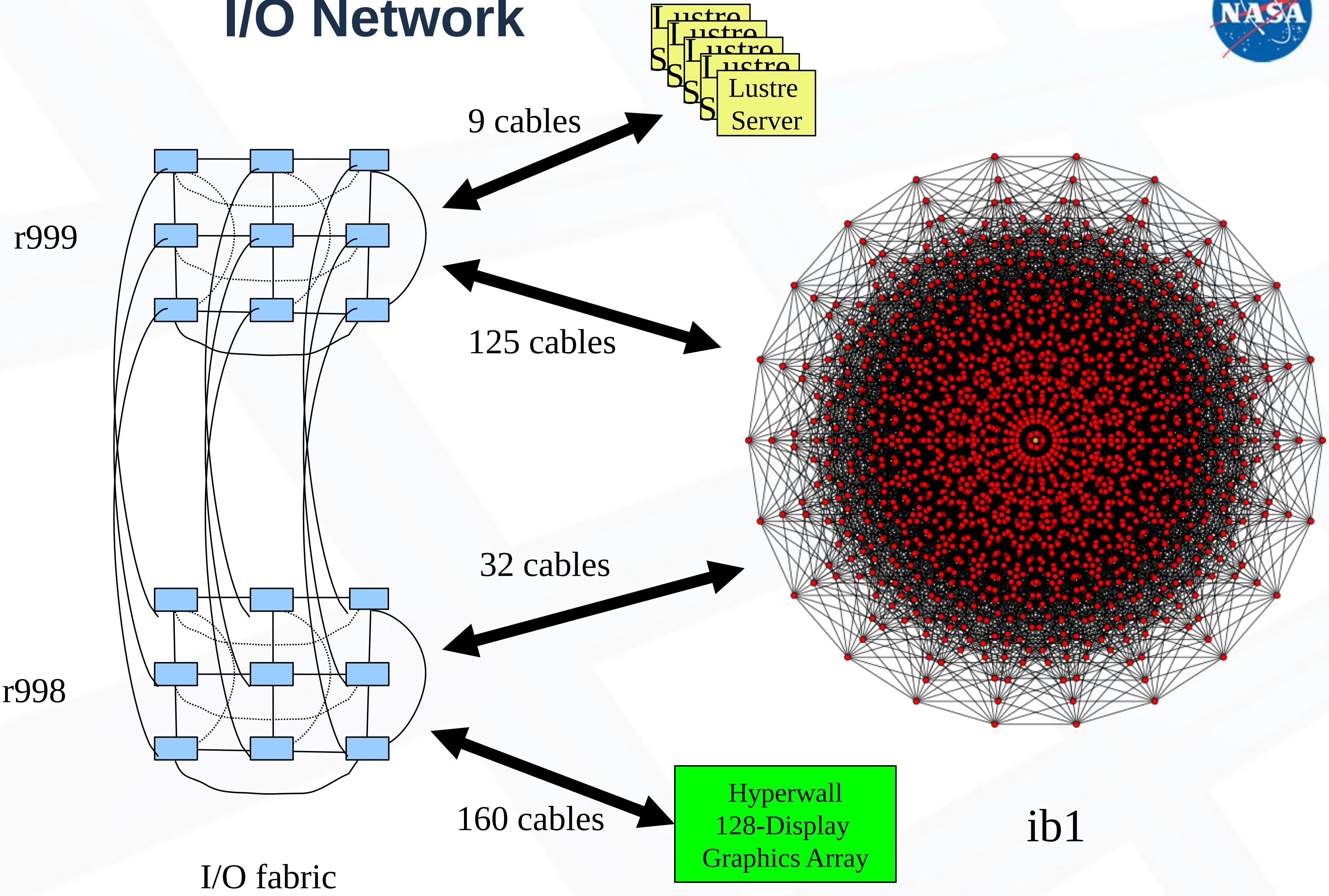
Infiniband – Subnet Discovery

- 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

I/O Network



Lustre Server

9 cables

r999

125 cables

32 cables

r998

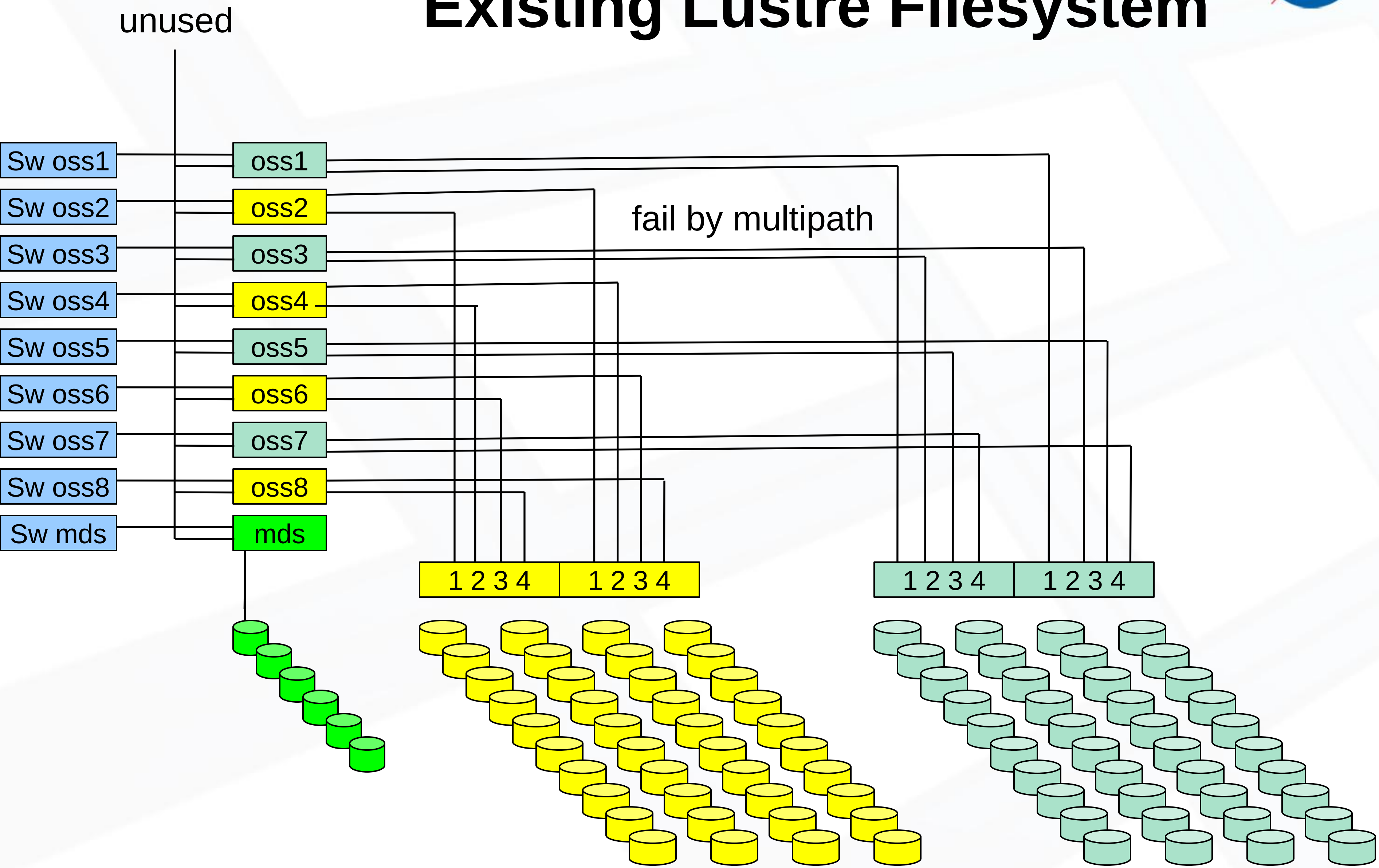
160 cables

Hyperwall
128-Display
Graphics Array

I/O fabric

ib1

Existing Lustre Filesystem



Real Time I/O Monitor

```
Every 1.0s: abracadabra -i 1
Mar 26 00:31:37 2012

    io_swx     nbp1        .     nbp2        .   nbp3/4        .     nbp5        .     nbp6        .      tot        .
         .     read    write     read    write     read    write     read    write     read    write     read    write

 r999i_mds      0.7      0.4      2.4      1.4     16.7     11.5      0.3      0.3      1.3      0.7     20.7     13.9
r999i_oss1      2.3      6.5     18.4    208.5      4.1     11.6      2.2      2.2      2.3      2.3     11.0     22.6
r999i_oss2      3.5    122.1      2.8     51.3      2.5      7.0      2.2      2.3      2.3      2.3     13.4    184.9
r999i_oss3      2.3      9.7     16.0     39.7      2.5      4.8      2.2      2.2      2.3      3.2     25.3     59.6
r999i_oss4      2.3      8.1     79.9     34.1      2.4      4.0      2.2      2.2      2.3      2.2     89.2     50.7
r999i_oss5      2.4      9.0      2.7     42.5      2.2     10.4      2.2      2.2      2.2      2.3     11.7     66.4
r999i_oss6      2.3     10.6      6.4     38.7      2.2      5.6      2.2      2.2      2.2      2.2     15.5     59.4
r999i_oss7      2.3     10.6      6.3     23.5      2.2     12.3      2.2      2.2      2.2      2.2     15.3     50.8
r999i_oss8      2.3     10.2    270.5     35.7      2.2      7.1      2.2      2.2      2.2      3.2    279.3     58.4
     Total     20.4    187.2    405.4    475.4     37.0     74.3     17.9     18.0     19.3     20.6    481.4    566.7
       Max   2809.2  16138.9   5943.9   5003.6   2310.6   4719.3     50.9    171.3  14930.3  15173.6  15127.3  16845.9

       Max  RcvData: 1514.8 8451.6 3319.8 1252.6 6261.4 7874.4 14207.8 3903.5 10441.4 8181.3 6720.7 5473.9   7.1   3.6   9.2   1.9   8.8   1.7  11.1   1.2   3.6 16847.1
       Max XmitData:   14.1 1393.7 6645.3 3405.3 1478.8 5506.1 13417.8 1675.2 2846.6 2498.5 1365.8 1210.5   8.8   2.0   6.9   3.8  10.4   1.2   8.9   2.1   4.7 15130.8

     Total  RcvData:    0.1   62.4    4.1    6.0    5.7   14.4   52.2   22.8  128.4   18.4  171.4  288.3   0.3   0.1   0.3   0.0   0.2   0.3   0.3   0.3   1.3   777.6
     Total XmitData:    0.1   17.7   11.2    6.4    6.3  105.0   15.0   15.0    8.9    9.8    2.8  301.8   0.3   0.1   0.3   0.1   0.3   0.3   0.2   0.4   1.3   502.7

 r999i_mds         .      .  r41i0  r49i1  r57i1  r17i0  r25i0 r129i0 r137i0 r145i0 r153i0      .   r9i0  oss1  oss1  oss2  oss2  oss3  oss3  oss6  oss6 hwsw0     tot
 r999i_mds  RcvData:    0.0    0.2    0.6    0.4    0.2    0.1    0.7    2.0    0.3    1.2    0.0    8.5   0.1   0.1   0.1   0.0   0.2   0.1   0.2   0.1   0.0    15.1
 r999i_mds XmitData:    0.0    1.9    1.3    0.9    0.2    0.1    1.2    2.2    0.3    2.1    0.0   11.2   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.8    22.2

r999i_oss1         .      .  r41i3  r49i3  r57i3  r17i3  r25i3 r129i3 r137i3 r145i3 r153i3   r1i3   r9i3  oss2  oss2   mds   mds  oss4  oss4  oss7  oss7 hwsw1     tot
r999i_oss1  RcvData:    0.0    5.2    0.5    0.3    0.8    2.9    4.9    2.0    1.9    5.6  170.4   37.2   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.1   231.9
r999i_oss1 XmitData:    0.0    3.3    1.3    0.5    0.8   12.0    1.8    1.7    1.1    0.9    1.9    4.2   0.0   0.0   0.1   0.0   0.0   0.0   0.0   0.0   0.0    29.7

r999i_oss2         .      .  r42i2  r50i2  r58i2  r18i2  r26i2 r130i2 r138i2 r146i2 r154i2   r2i2  r10i2   mds   mds  oss1  oss1  oss5  oss5  oss8  oss8 hwsw2     tot
r999i_oss2  RcvData:    0.0    7.3    0.5    0.3    0.7    2.8    7.6    2.0  115.3    1.8    0.2   46.3   0.0   0.0   0.0   0.0   0.0   0.1   0.0   0.1   0.3   185.1
r999i_oss2 XmitData:    0.0    1.8    1.3    0.5    0.7    0.9    1.8    1.7    2.2    0.9    0.2    1.3   0.1   0.0   0.0   0.0   0.1   0.1   0.0   0.0   0.2    13.6
```

Pleiades Infiniband Specifics

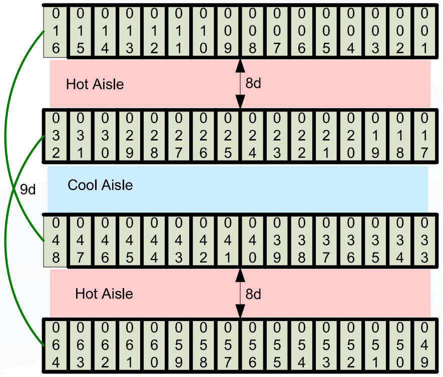
- Mix of infinihost III, Connect-X 1-2-3 [DDR,QDR,FDR] HCA
  - ~12,500 cables (over 50 miles - combination of optical/copper)
  - ~22,000 active host ports
- Mix of infiniscale III, infiniscale IV, switch X switches
  - 2,914 total switch chips
- Two Major subnets (~12,000 endpoints)
- 73,142 ports (21,704 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)

Pleiades April 2012 Target Configuration
SGI ICE System

- 11,712 nodes – 23,434 sockets - 126,720 x86 cores
  - 4,096 harpertown nodes x5670
  - 1,280 nehalen nodes x5472
  - 4,672 westmere x5570
  - 1,720 sandybridge x
  - +128 hw2 - vis (opteron 2354)

- Resite 1,752 harpertowns (n233)
  - how to go 1.8 KM
  - color chip transceivers
    - modified switch firmware - consolidates vls and port group buffers
      - achieves qdr line rate
      - 3 or 4 ports

NASA (Pleiades) Rack Layout



A (ICE DDR)

B (ICE DDR)

C (ICE DDR)

D (ICE DDR)

64 racks – 2008
393 teraflops