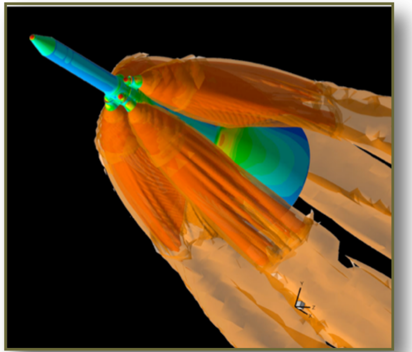
NASA Mult-Rail Router Deployment

Mahmoud Hanafi
Bob Ciotti
Dale Talcott
Mike Hartman

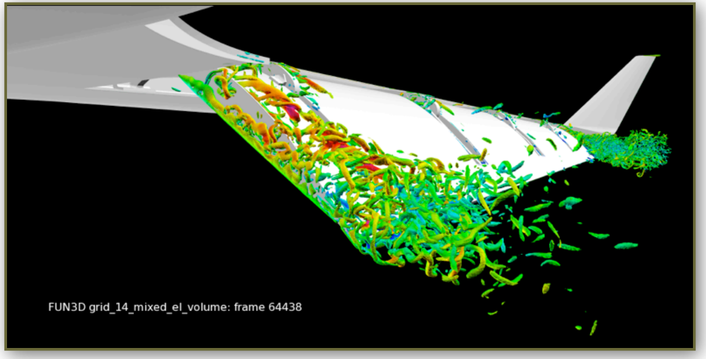
NASA's HEC Requirements: Capacity

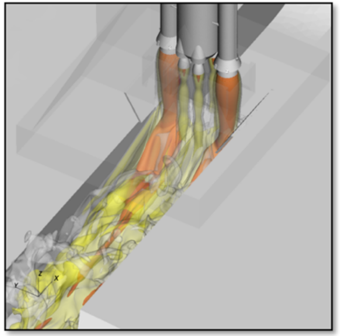
HEOMD (engineering-related work) require HEC resources that can handle large numbers of relatively-low CPU-count jobs with quick turnaround times.



Over 1500 simulations utilized ~ 2 million processor hours to study launch abort systems on the next generation crew transport vehicle

The formation of vortex filaments and their roll-up into a single, prominent vortex at each tip on a Gulfstream aircraft

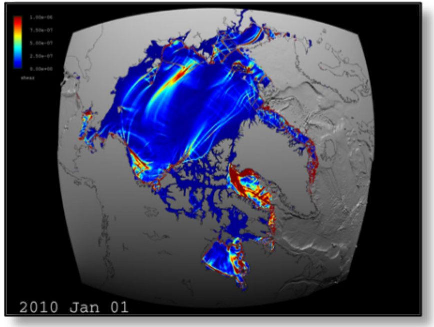




Over 4 million hours were used over a 4 month project to evaluate future designed of the next generation launch complex at the Kennedy Space Center

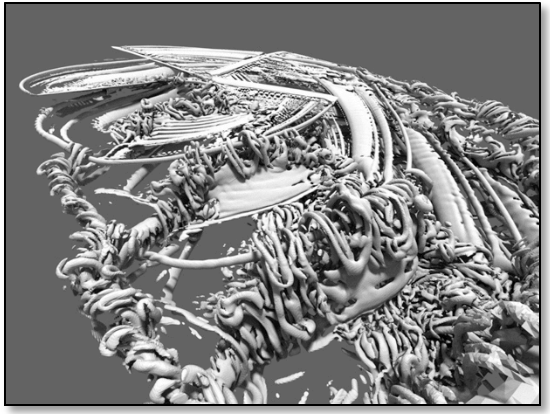
NASA’s HEC Requirements: Capability

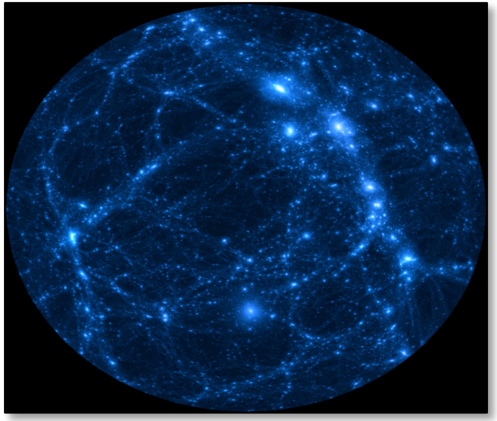
ARMD and SMD (aeronautics and science related work) require HEC resources that can handle high fidelity relatively-large CPU-count jobs with minimal time-to-solution. Capability enables work that wasn’t possible on previous architectures.



NASA is looking at the oceans, running 100’s of jobs on Pleiades using up to 10,000 processors. Looking at the role of the oceans in the global carbon cycle is enabled by access to large processing and storage assets

For the first time, the Figure-of-Merit has been predicted within experimental error for the V22 Osprey and Black Hawk helicopter rotors in hover, over a wide range of flow conditions





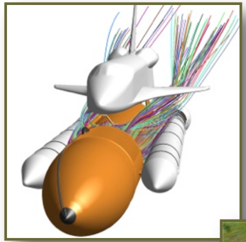
To complete the Bolshoi simulation, which traces how the largest galaxies and galaxy structures in the universe were formed billions of years ago, astrophysicists ran their code for 18 straight days, consuming millions of hours of computer time, and generating massive amounts of data

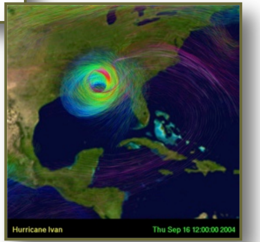
NASA's HEC Requirements: Time Critical

NASA also has need for HEC resources that can handle time-sensitive mission-critical applications on demand (maintain readiness)

HECC enables the enormous planetary transit searches to be completed in less than a day, as opposed to more than a month on the Kepler SOC systems, with significantly improved accuracy and effectiveness of the software pipeline

REENTRY





STORM PREDICTION

UAVSAR produces polarimetric (PolSAR) and interferometric (repeat-pass InSAR) data that highlight different features and show changes in the Earth over time

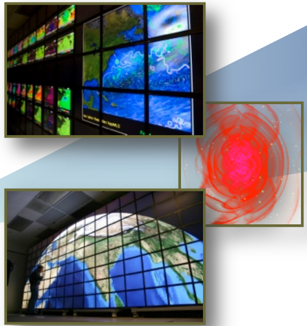
HECC Conducts Work in Four Major Technical Areas

Supercomputing Systems

Provide computational power, mass storage, and user-friendly runtime environment through continuous development of management tools, IT security, systems engineering

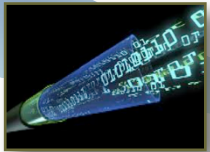
Application Performance and User Productivity

Facilitate advances in science and engineering for NASA programs by enhancing user productivity and code performance of high-end computing applications of interest



Data Analysis and Visualization

Create functional data analysis and visualization software to enhance engineering decision support and scientific discovery by incorporating advanced visualization technologies



Networking

Provide end-to-end high-performance networking analysis and support to meet massive modeling and simulation distribution and access requirements of geographically dispersed users

Supporting Tasks

- **Facility, Plant Engineering, and Operations:** Necessary engineering and facility support to ensure the safety of HECC assets and staff
- **Information Technology Security:** Provide management, operation, monitoring, and safeguards to protect information and IT assets
- **User Services:** Account management and reporting, system monitoring and operations, first-tier 24x7 support
- **Internal Operations:** NASA Division activities hat support and enhance the HECC Project areas

HECC Platforms

Major HECC Systems

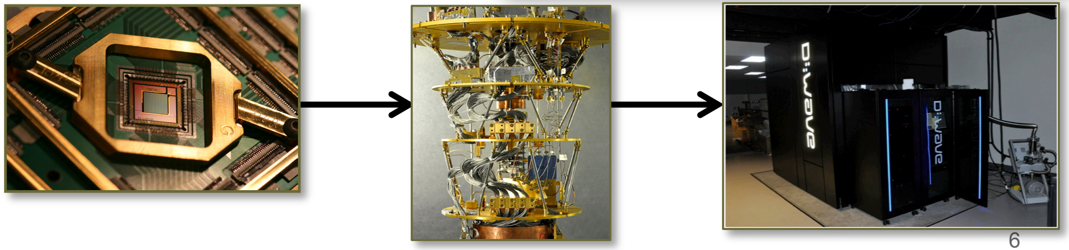
4 Compute Clusters

- Pleiades — 161 Racks / 11,340 nodes / 7.57 PF / 32,230 SBU/hr
- Electra — 20 Racks / 2,304 nodes / 4.78 PF / 11,566 SBU/hr
- Merope — 56 ½ Racks / 1,792 nodes / 252 TF / 1,792 SBU/hr
- Endeavour — 3 Racks / 2 nodes / 32 TF / 140 SBU/hr

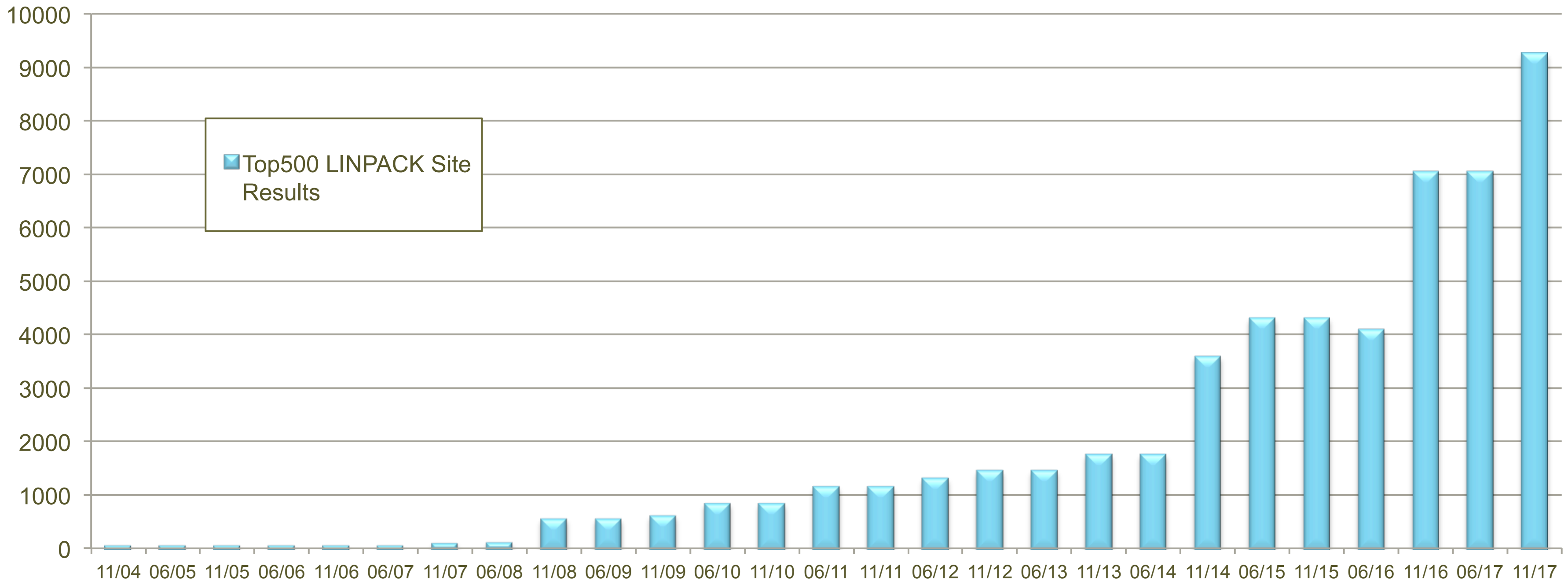
1 Visualization Cluster	245 million pixel display / 128 node / 703 TF
7 Lustre File Systems	39.6 PB
6 NFS File Systems	1.5 PB
1 BGFS Converged nVME	250 TB
Archive System	490 PB

Experimental Quantum D-Wave 2

- System with 1,097 qubits

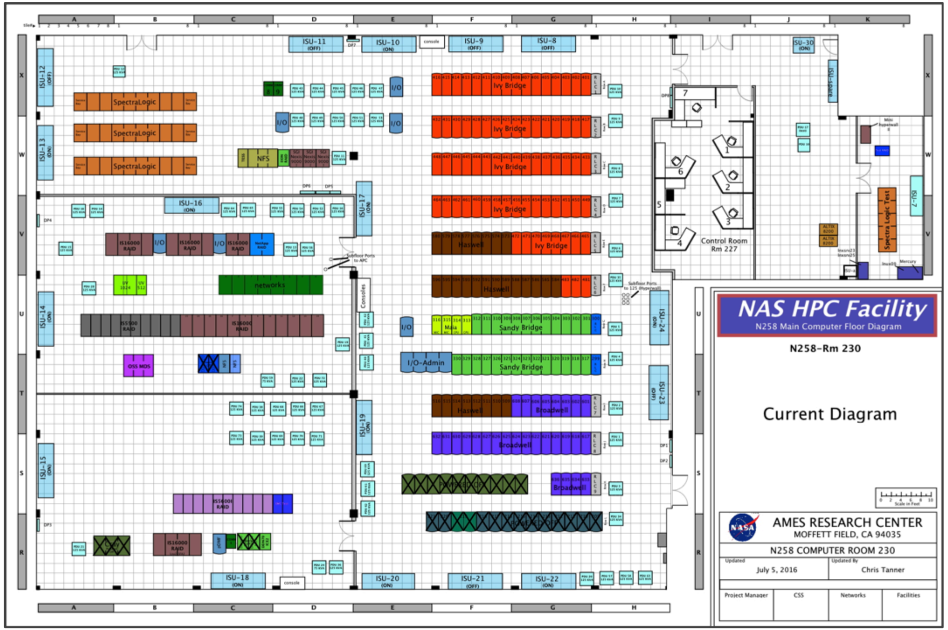


HECC Growth

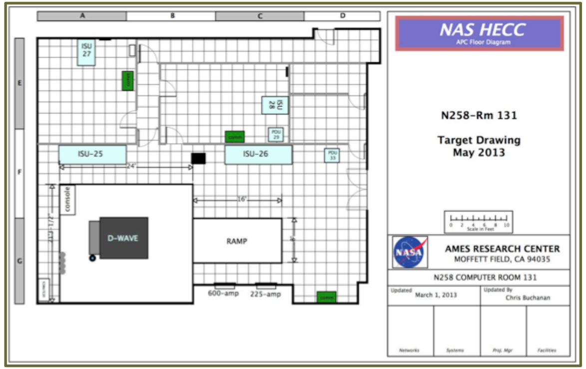


10000
9000
8000
7000
6000
5000
4000
3000
2000
1000
0
Top500 LINPACK Site Results
11/04 06/05 11/05 06/06 11/06 06/07 11/07 06/08 11/08 06/09 11/09 06/10 11/10 06/11 11/11 06/12 11/12 06/13 11/13 06/14 11/14 06/15 11/15 06/16 11/16 06/17 11/17

HECC Traditional Computer Facilities



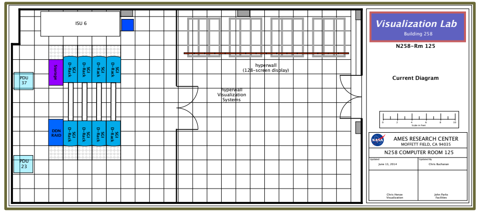
230
1,560 m²



131
193 m²

190
640 m²

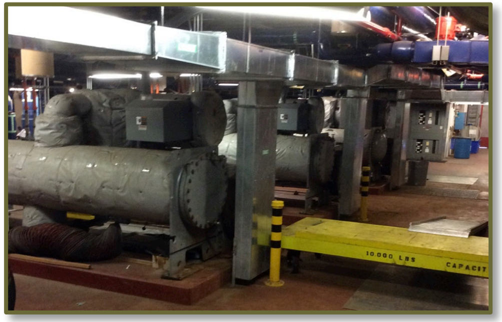
189
250 m²



125
118 m²

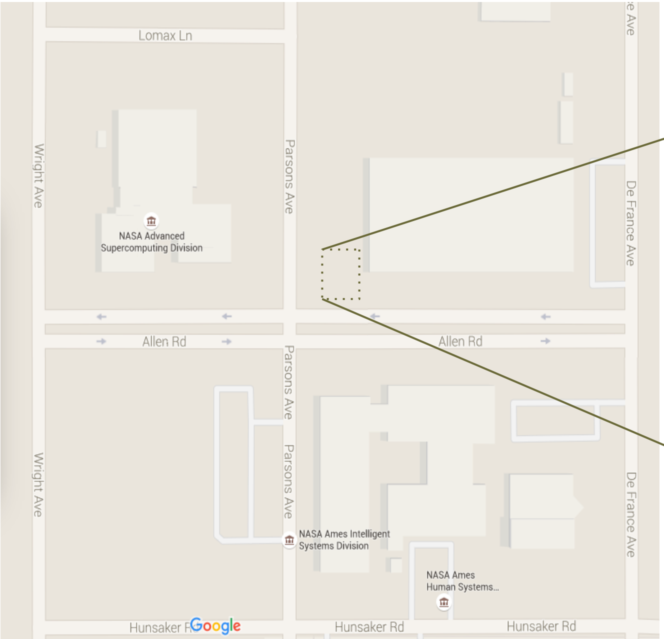
Current Primary Installation

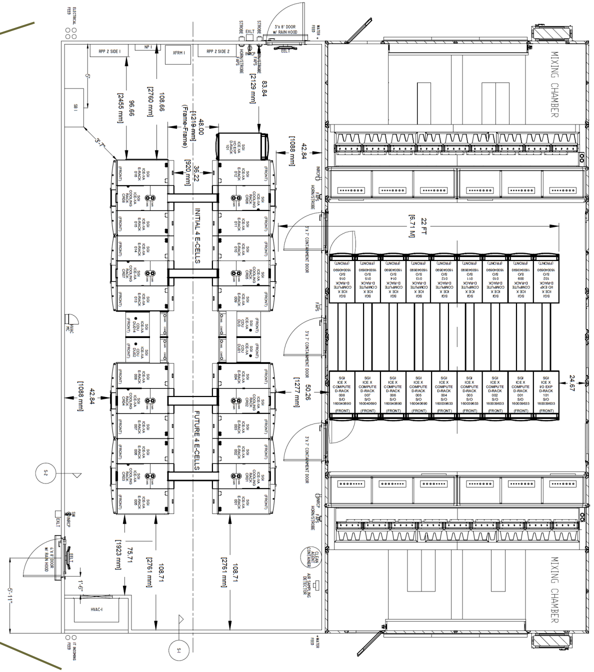
- **Limited by power and cooling**
- **Our Current Cooling System**
  - Open Air Cooling Tower with 4 50HP pumps
  - 4 450 Ton Chillers
  - 7 pumps for outbound chilled water
  - 4 pumps for inbound warm water
- **Our Electrical System**
  - Nominally the facility is limited to 6MW
  - 20% - 30% is used for cooling
  - 4MW – 5MW for computing



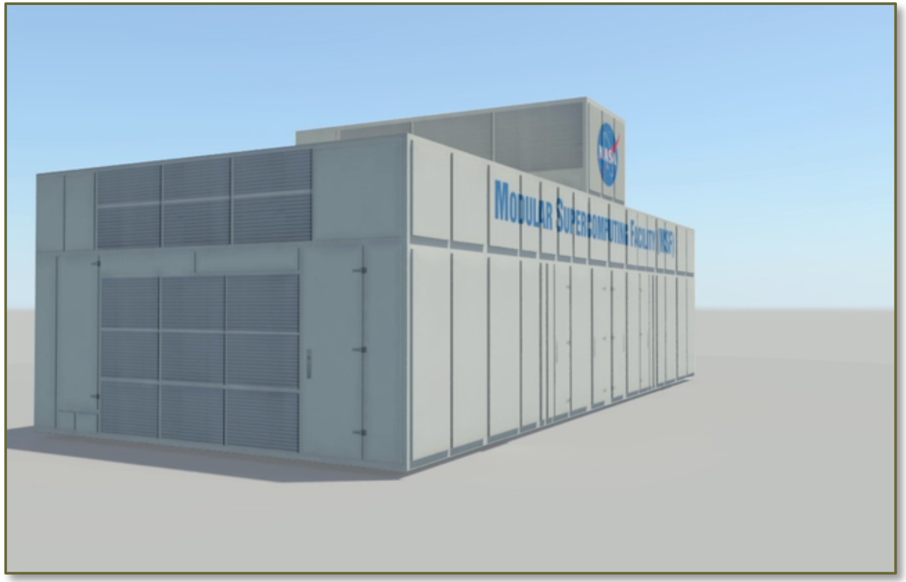
HECC Modular Facility

- **HECC Prototype Facilities**
- **Modular Supercomputer Facility**
- **Concrete Pad** — **362 square meters / 2.4 MW / will hold 2 adjacent DCoD-20 modules**
  - DCoD-20 Module 1 — 90 square meters / 40 square meters computer floor / 500 KW
  - Custom Module 2 — 90 square meters / 86 square meters computer floor / 1,200 KW





DCU-20

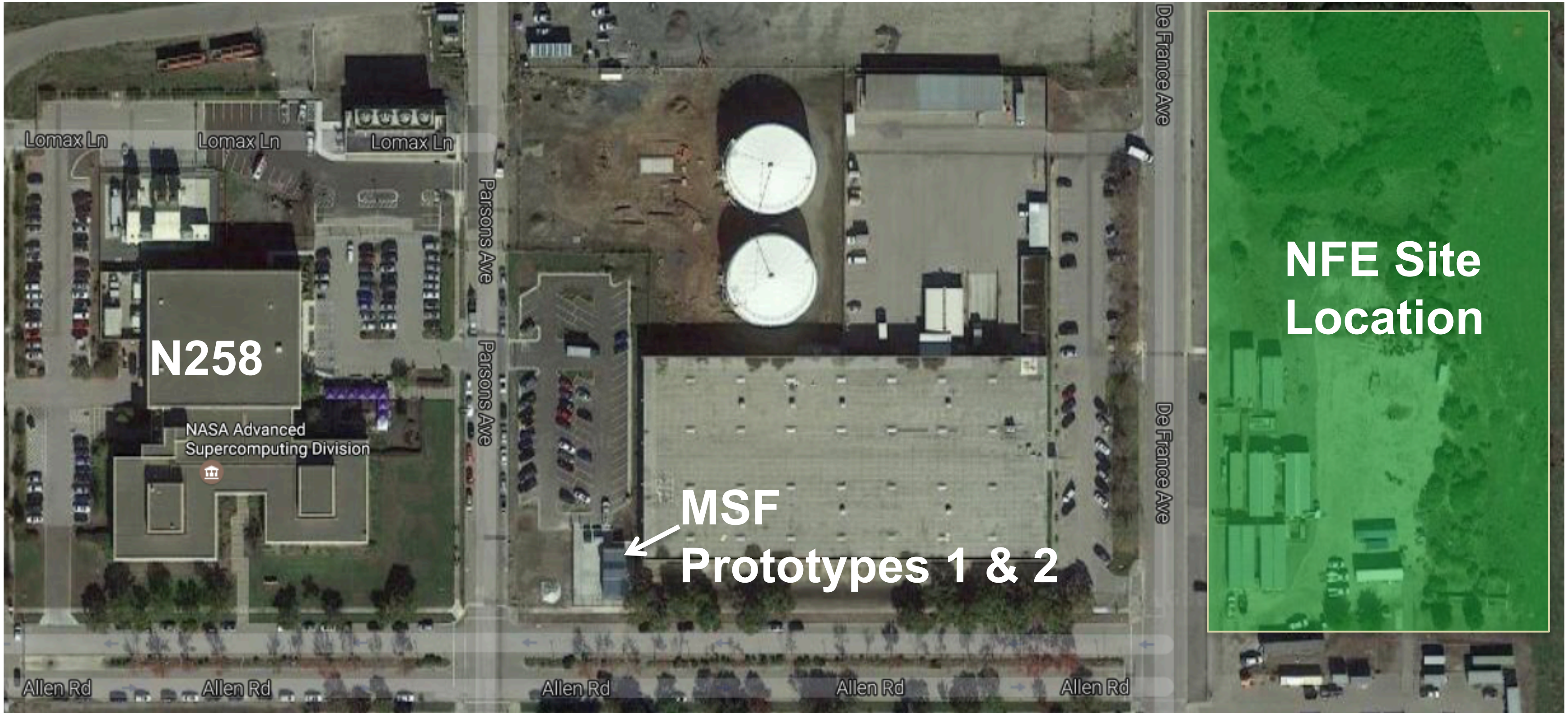


Module 2 Assembly

Full Site Deployment Concept

- 8 Compute modules house 96 tightly coupled E-Cells providing 84.9 PF
- 5 Data modules house 420 PB of formatted storage protected by dual generators and battery UPS
- System joins existing HECC assets with shared file systems and data archive
- Project deployed on site currently being constructed and available in early FY19
- Project fully operational in FY19

Site Location



Motivation

- **NASA HECC computer facilities physically distributed**
  - Leverage Exiting/Historical facility space
    - N258 – Primary Facility – 6 MW
    - N233 – Aux Facility – 1 MW
    - Electra – 1 MW
    - NFE – up to 30MW

- Distances from 2KM to 200 Meters

- Collection of various Infiniband/Ethernet Campus Area networking Equipment

  Metro-X 10 km – 160 gb (4 x 40gb QDR)
  Metro-X 1 km – 640 gb (16 x 40gb FDR10)
  Obsidian encrypted long haul 10gb – 1
  Obsidian 10 km 40gb (1 x QDR)
  Experimented with Luxtera cables, Voltaire switch firmware to extend infiniband layer-2

» Bottom Line: on-going need for HPC level Campus Area Networks (scalable to 100's GB/sec)

Procurement Strategy

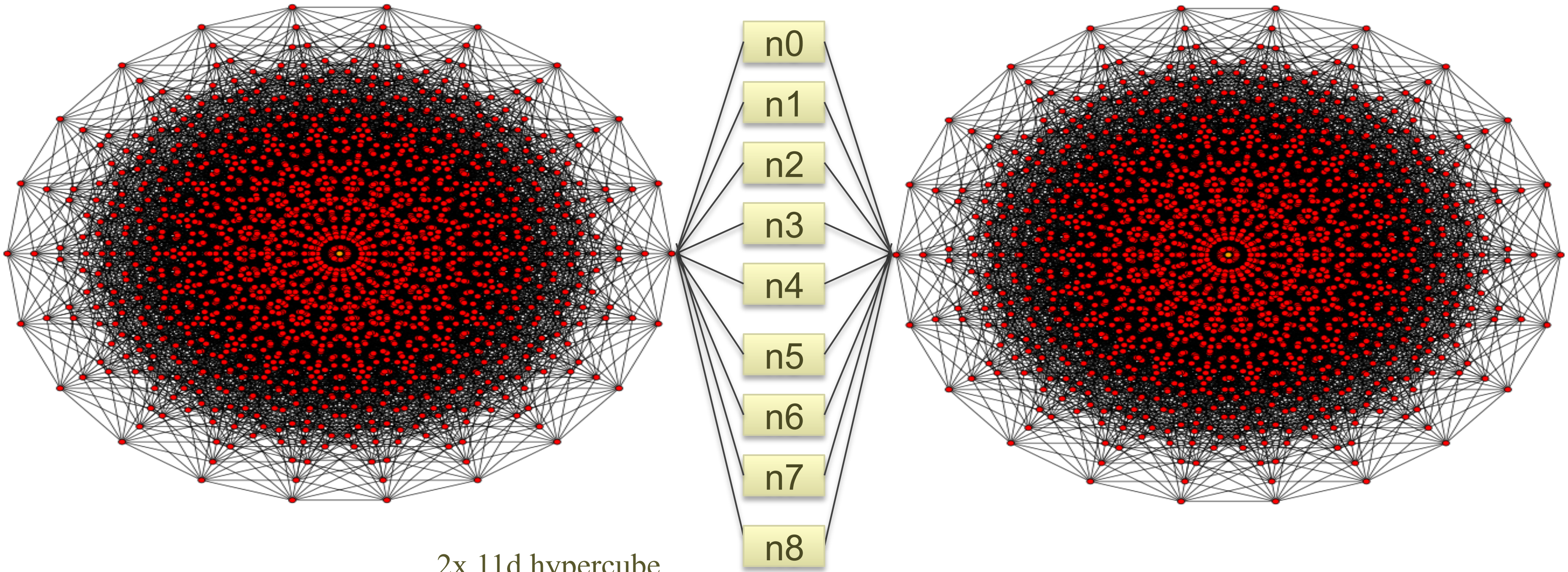
- Buy as yearly funds become available
- Some timing involved for product releases (e.g. new processors/networks)
- Results in several incremental smaller builds
- Thin provisioning for most non-compute elements (e.g. storage)

Future System Design

- **Building a campus area distributed, large scale HPC infrastructure**
  - Multiple phases
    - 2 phase Prototype
    - Multiple incremental deployments
    - Maximum scale at ~ 30MW

- **Networking Design Requirements - Must Be:**
  - High performance, cost effective and scalable
  - Redundant HW, Ideally Active/Active

Pleiades
SGI/HPE ICE Dual Plane – Topology



ib0

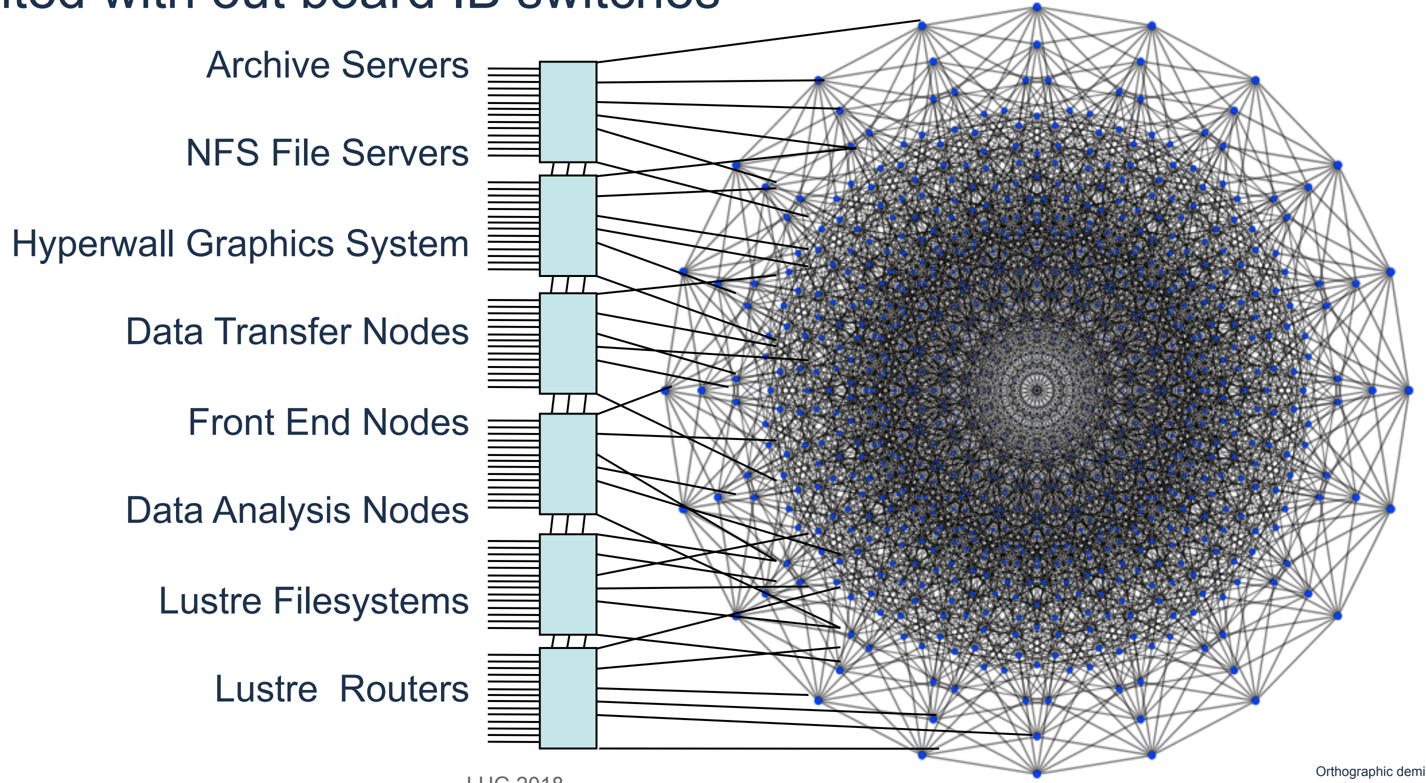
2x 11d hypercube
full 2048 vertices
Pleiades 1336/11d (2672 across both cubes)

ib1

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

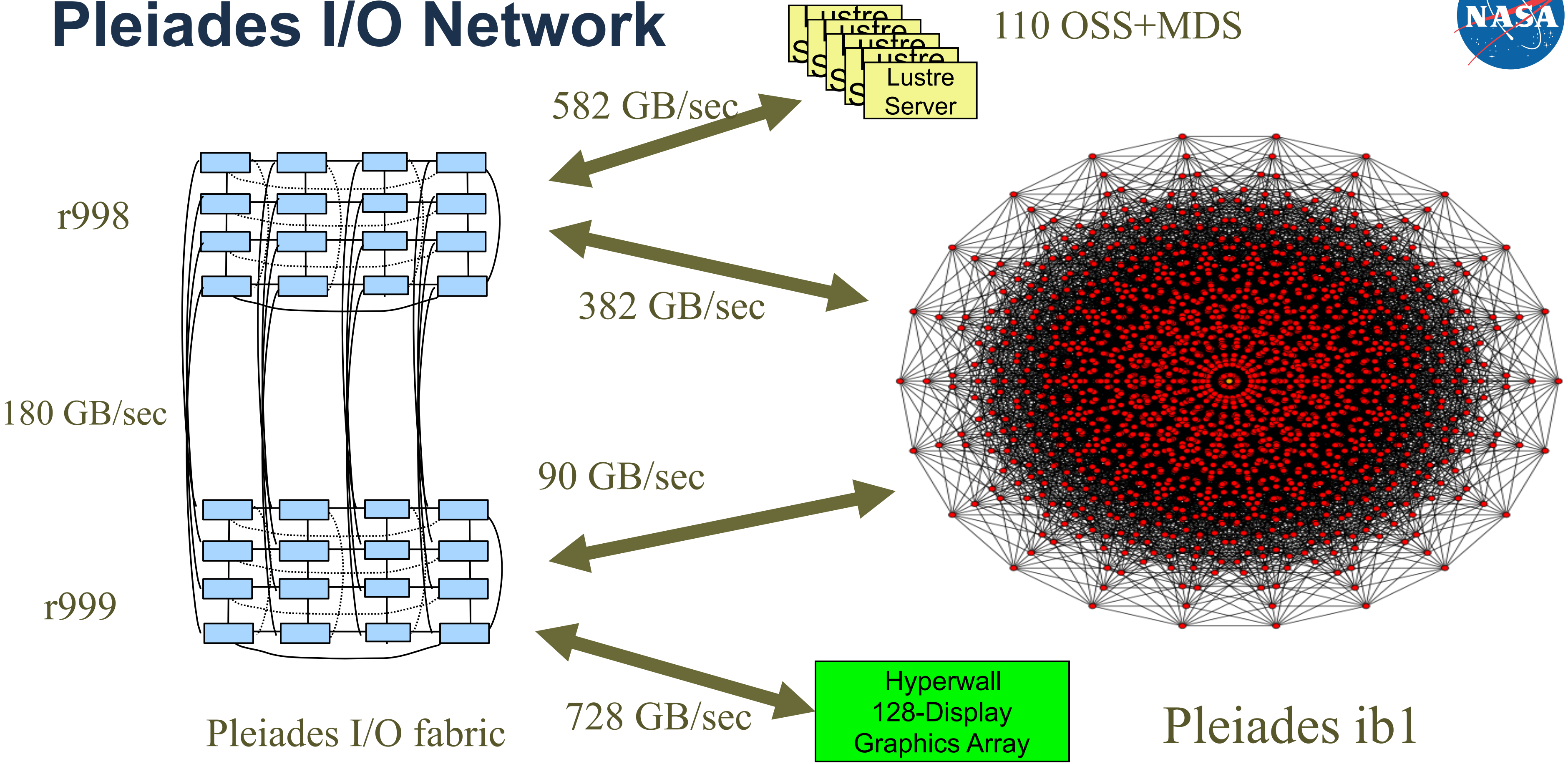
Pleiades Infiniband Subnet LAN

LAN Implemented with out board IB switches

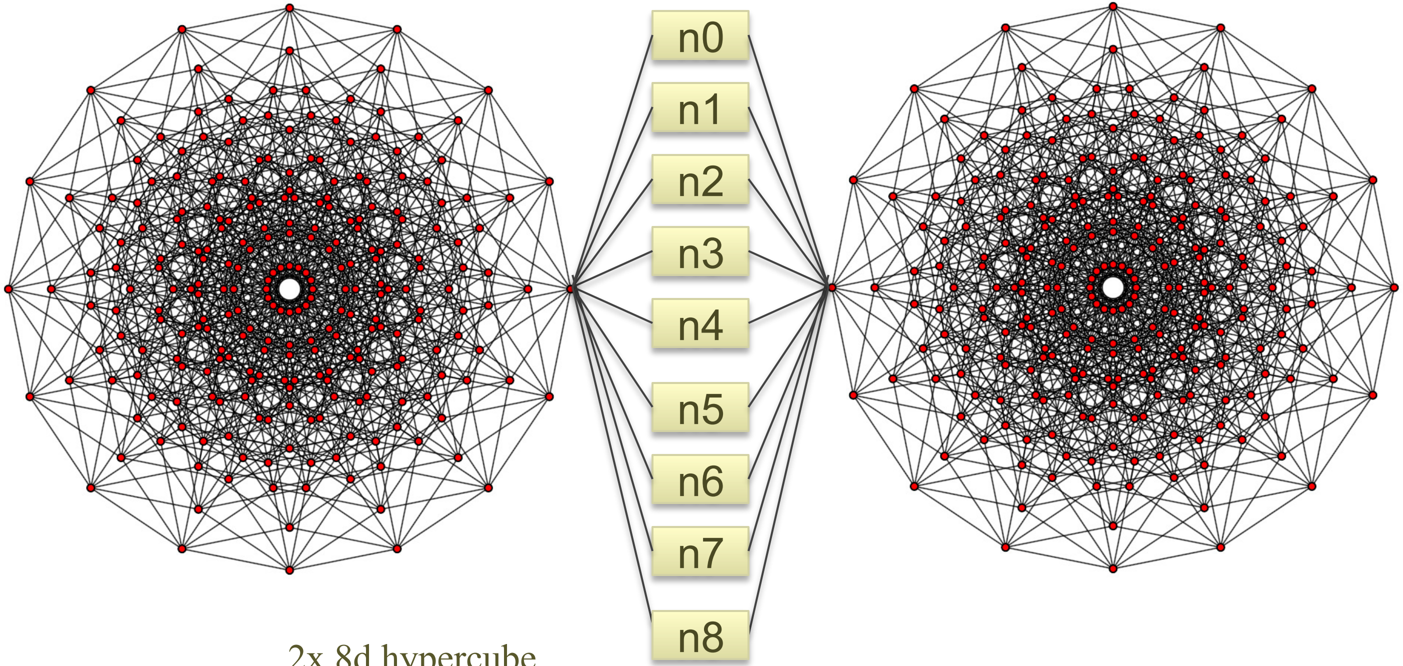


Orthographic demidekeract
by Claudio Rocchini, wikipedia

Pleiades I/O Network



Electra V1
SGI/HPE ICE Dual Plane – Topology



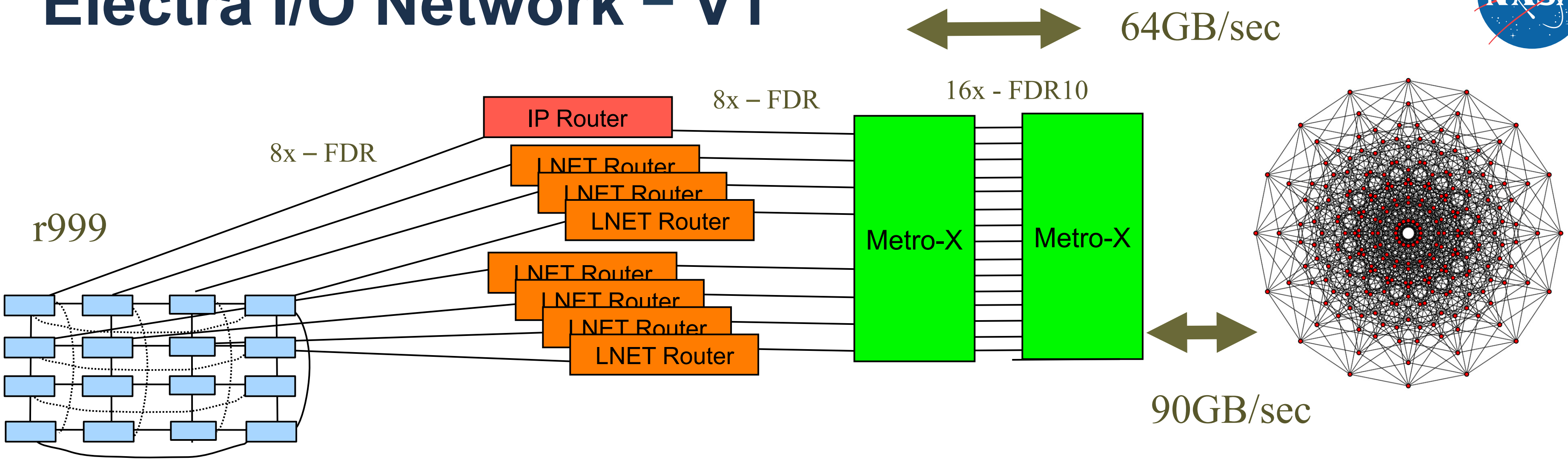
ib0

2x 8d hypercube
full 256 vertices
Electra 256/8d (512across both cubes)

ib1

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

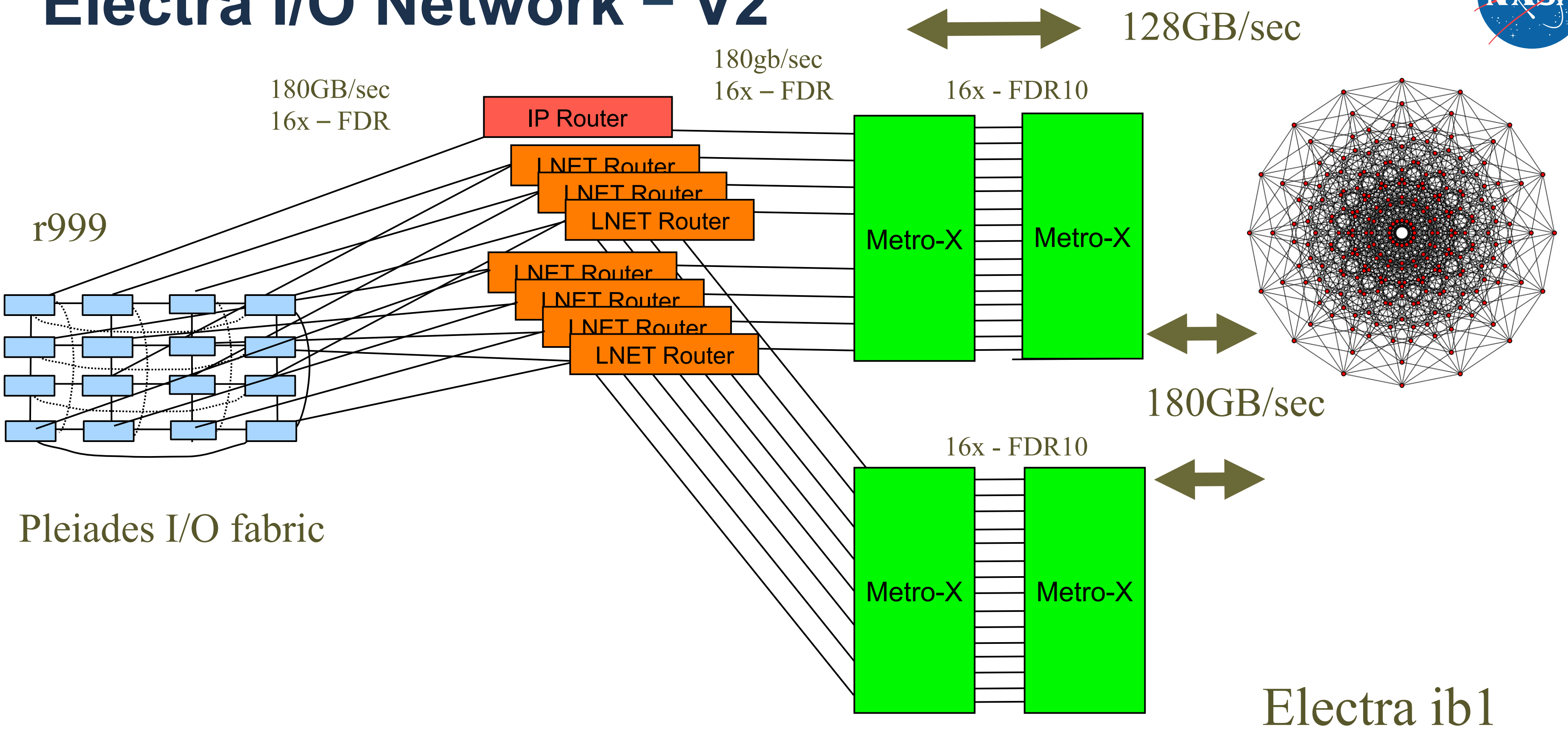
Electra I/O Network – V1



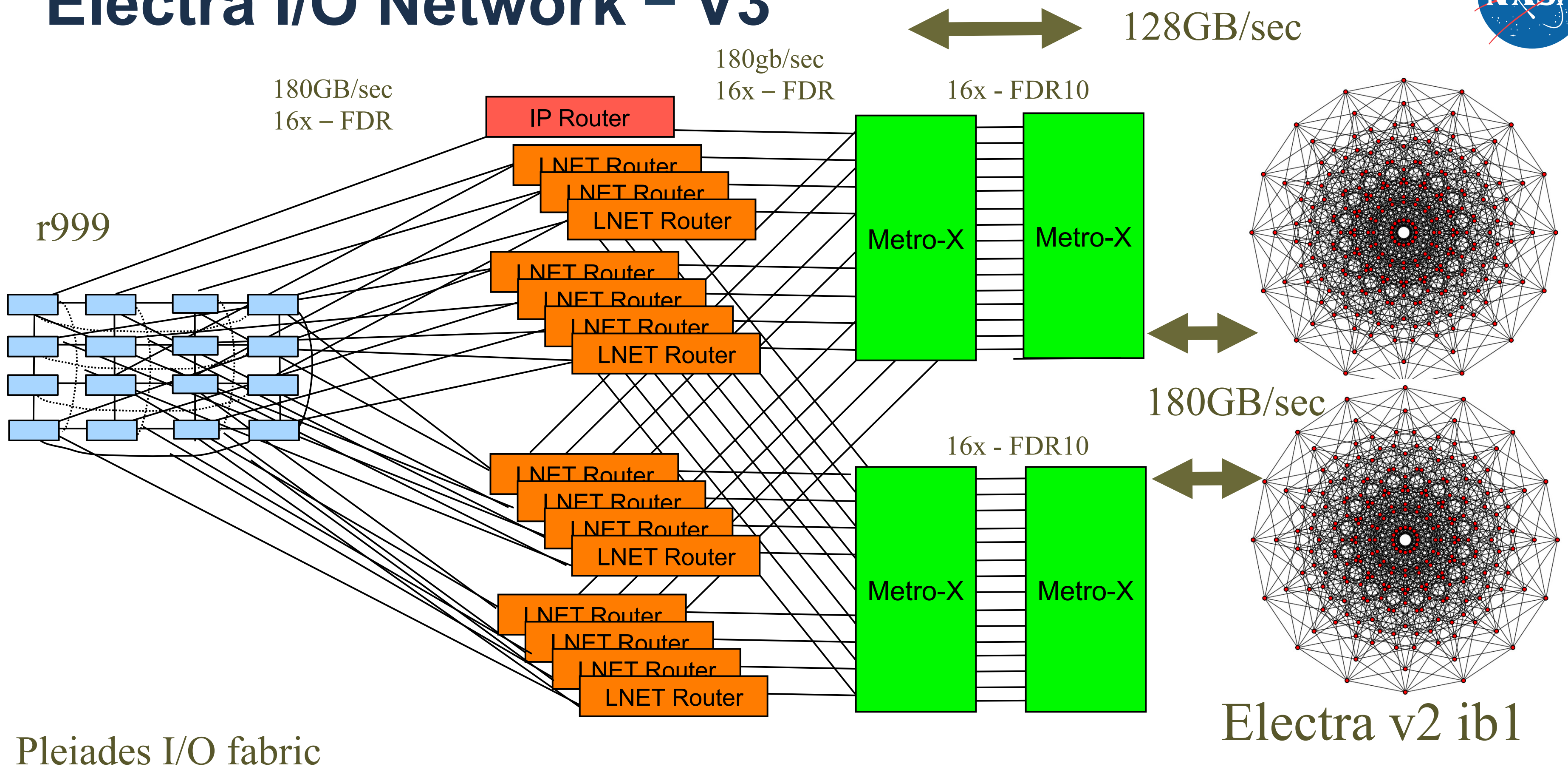
Pleiades I/O fabric

Electra ib1

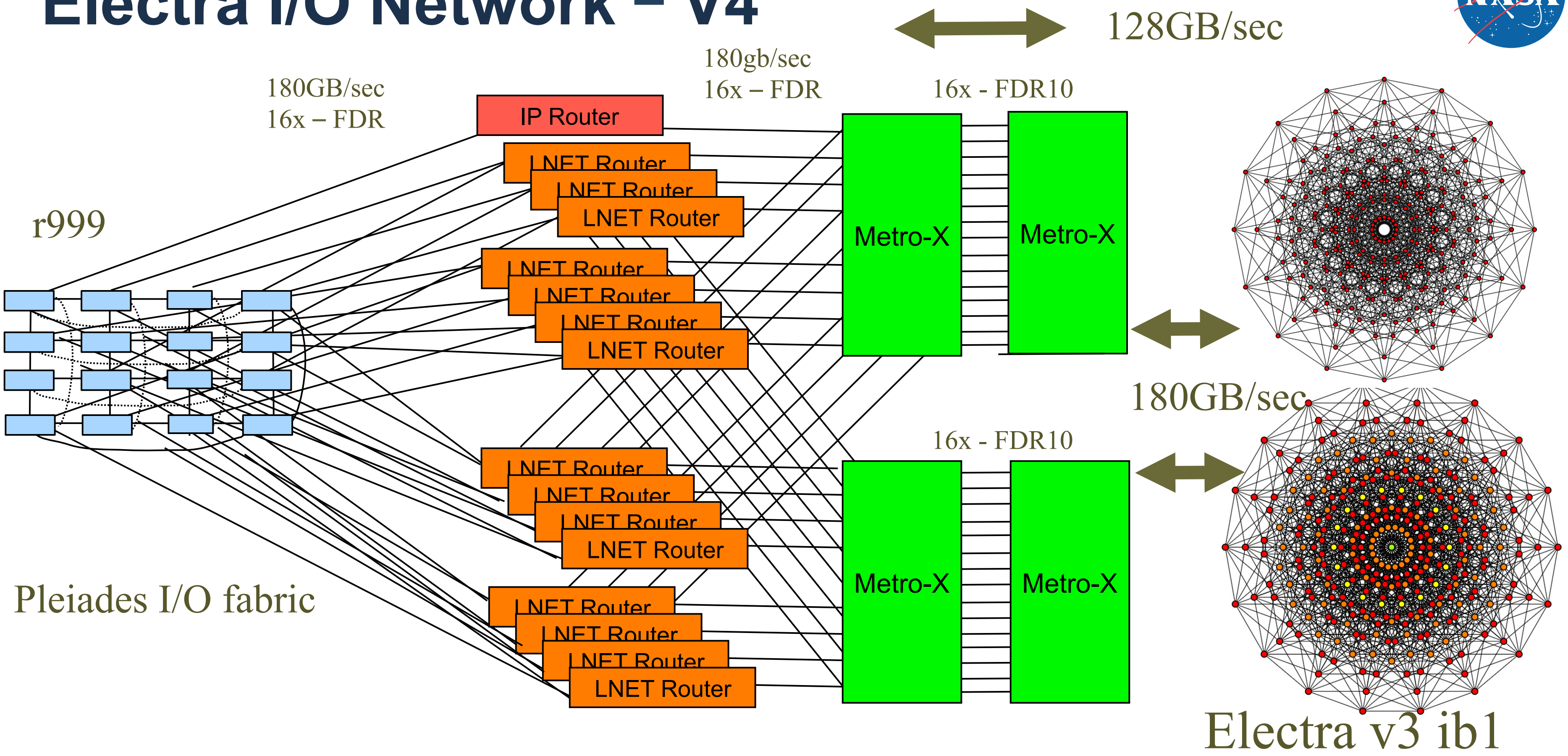
Electra I/O Network – V2



Electra I/O Network – V3



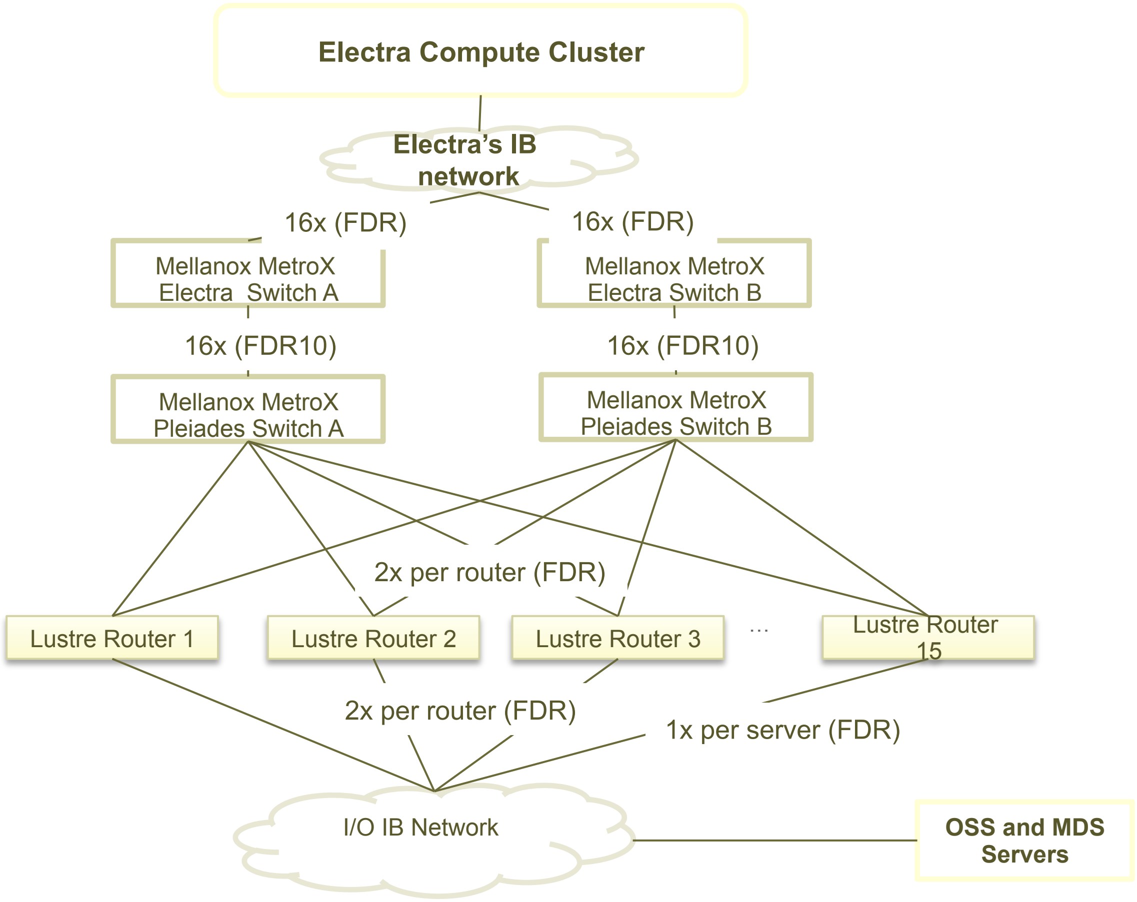
Electra I/O Network – V4



Configuration

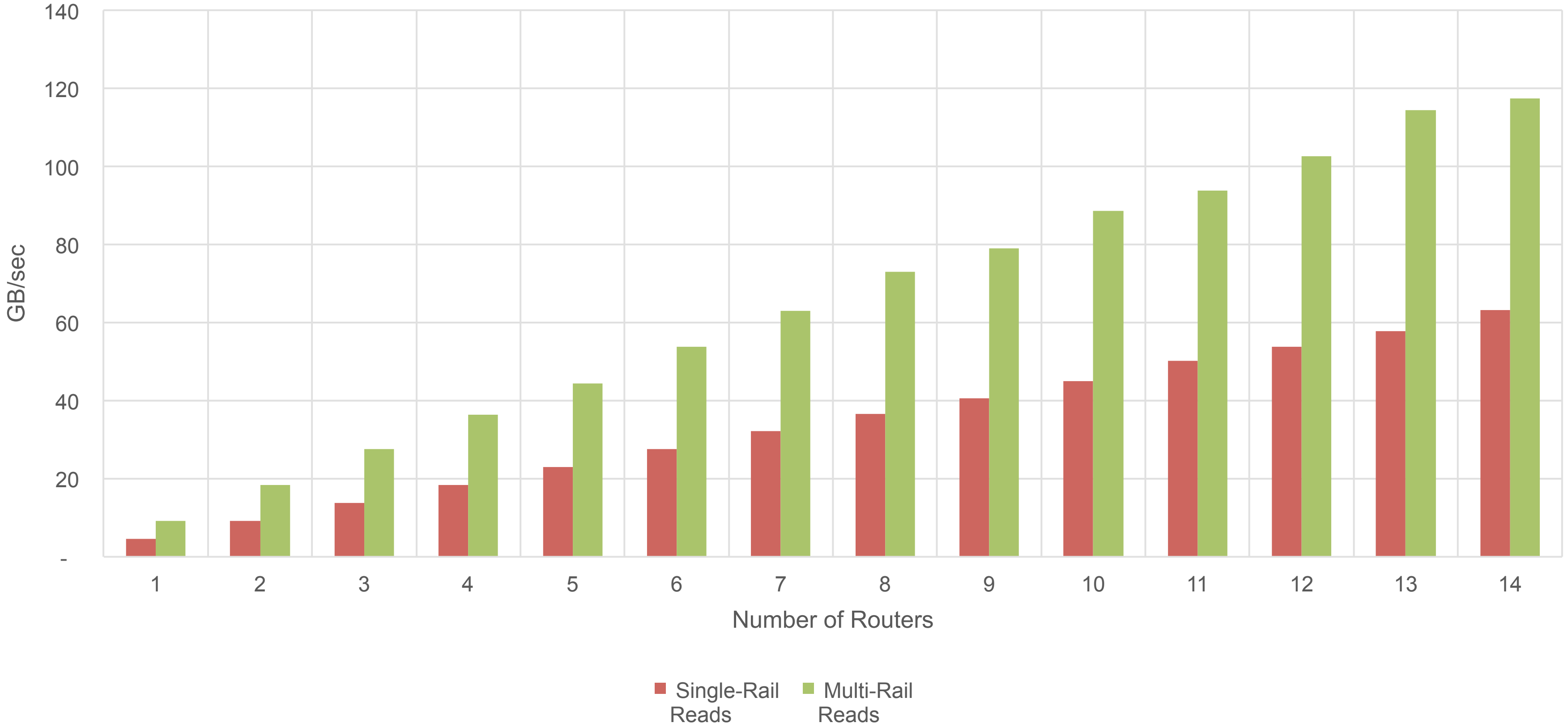
- **15 Routers Lustre 2.10.2 Multi-Rail**
  - CPU E5-2680 v4 @ 2.40GHz
  - 128 GB memory
  - 2x Mellanox Technologies MT27600
    - » Connect-IB
    - » 2 Ports each
  - Dual Interface single subnet ARP configuration
    - » Requires Policy Routing

      https://access.redhat.com/solutions/30564
- **Clients 2.10 and 2.9 Single-Rail config**
- **Servers 2.7 Single-Rail**
- **Client and Servers sees multi-interfaces as individual routers.**
  - No multi-rail config required
  - Each interface is listed in lnet routes


Electra Compute Cluster
Electra's IB network
16x (FDR)
16x (FDR)
Mellanox MetroX Electra Switch A
Mellanox MetroX Electra Switch B
16x (FDR10)
16x (FDR10)
Mellanox MetroX Pleiades Switch A
Mellanox MetroX Pleiades Switch B
2x per router (FDR)
Lustre Router 1
Lustre Router 2
Lustre Router 3
…
Lustre Router 15
2x per router (FDR)
1x per server (FDR)
I/O IB Network
OSS and MDS Servers

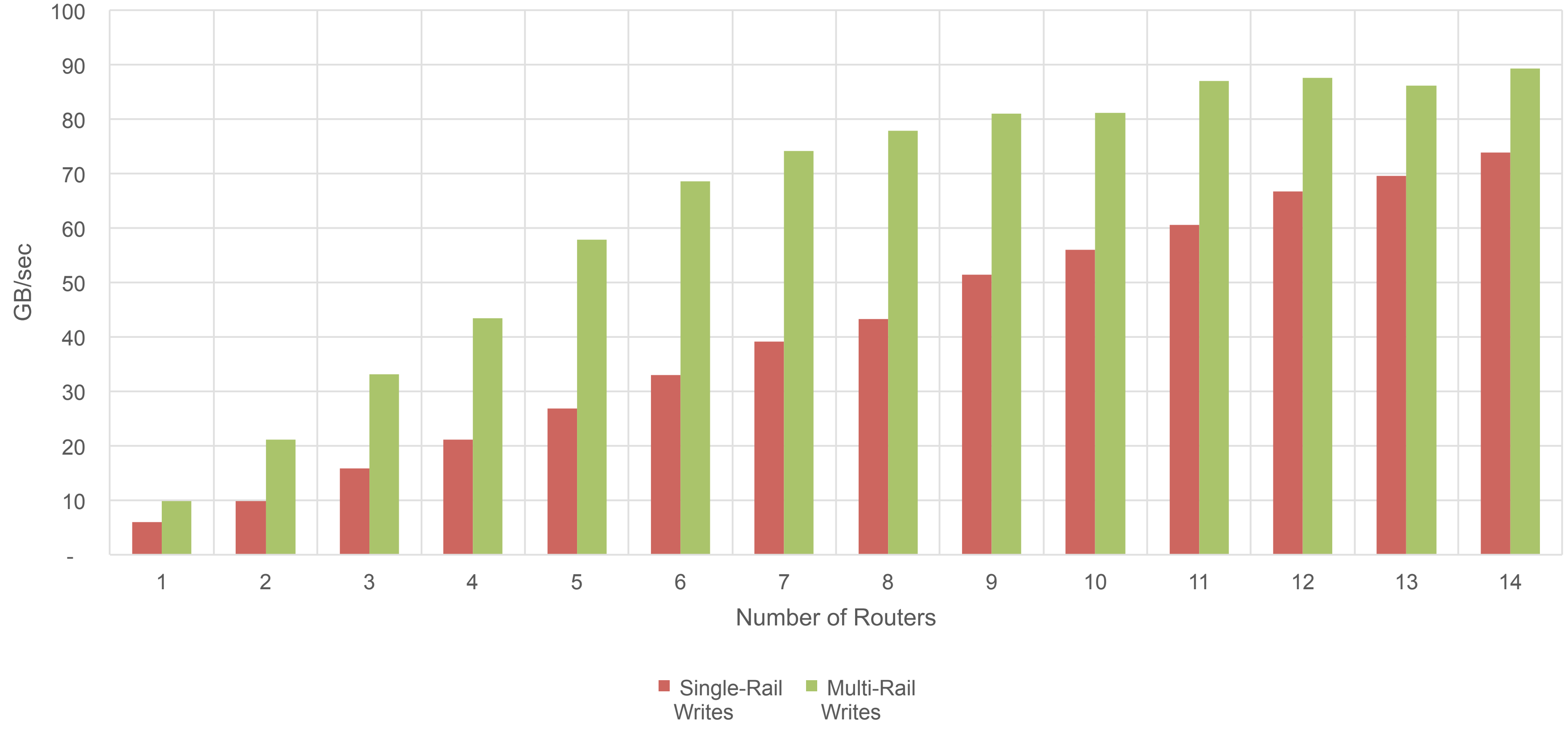
Electra Routers lnet-selftest Read

54 clients 6 lnet Threads per client



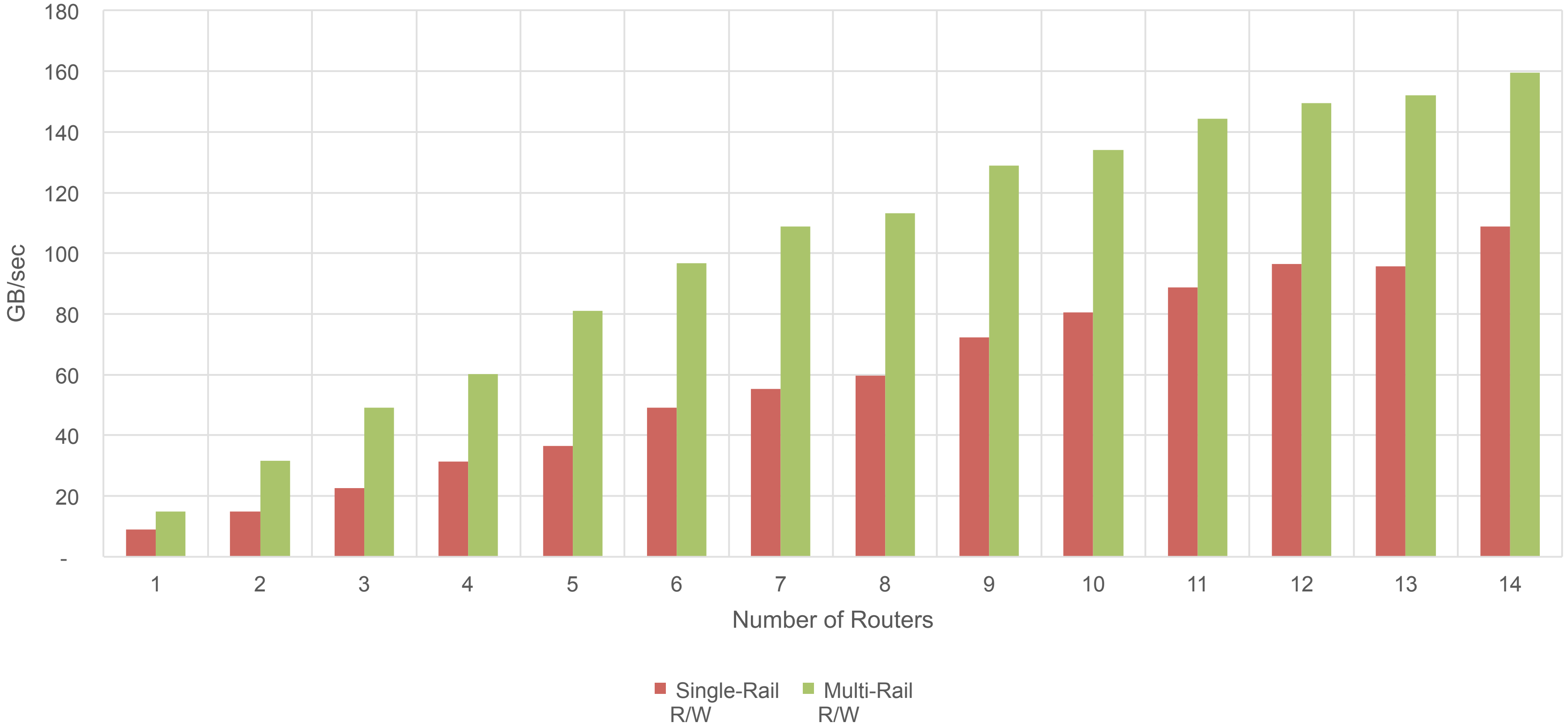
Electra Routers lnet-selftest Write

54 clients 6 lnet Threads per client



Electra Routers lnet-selftest Read/Write

54 clients 6 lnet Threads per client



Typical Server Module File

```
options ko2iblnd require_privileged_port=0 use_privileged_port=0
options ko2iblnd ntx=125536 credits=62768 fmr_pool_size=31385
options ko2iblnd timeout=150 retry_count=7 peer_timeout=0 map_on_demand=32 peer_credits=63 concurrent_sends=63


#lnet
options lnet networks=o2ib(ib1)
options lnet routes="o2ib233 10.151.26.[80-94]@o2ib; o2ib313 10.151.25.[167-170,195-197,202-205,222]@o2ib 10.151.26.[60,127,140-144,146-154]@o2ib"
options lnet dead_router_check_interval=60 live_router_check_interval=30
options lnet avoid_asym_router_failure=1 check_routers_before_use=1 small_router_buffers=65536 large_router_buffers=8192
```

Kudos

Intel Lustre Team

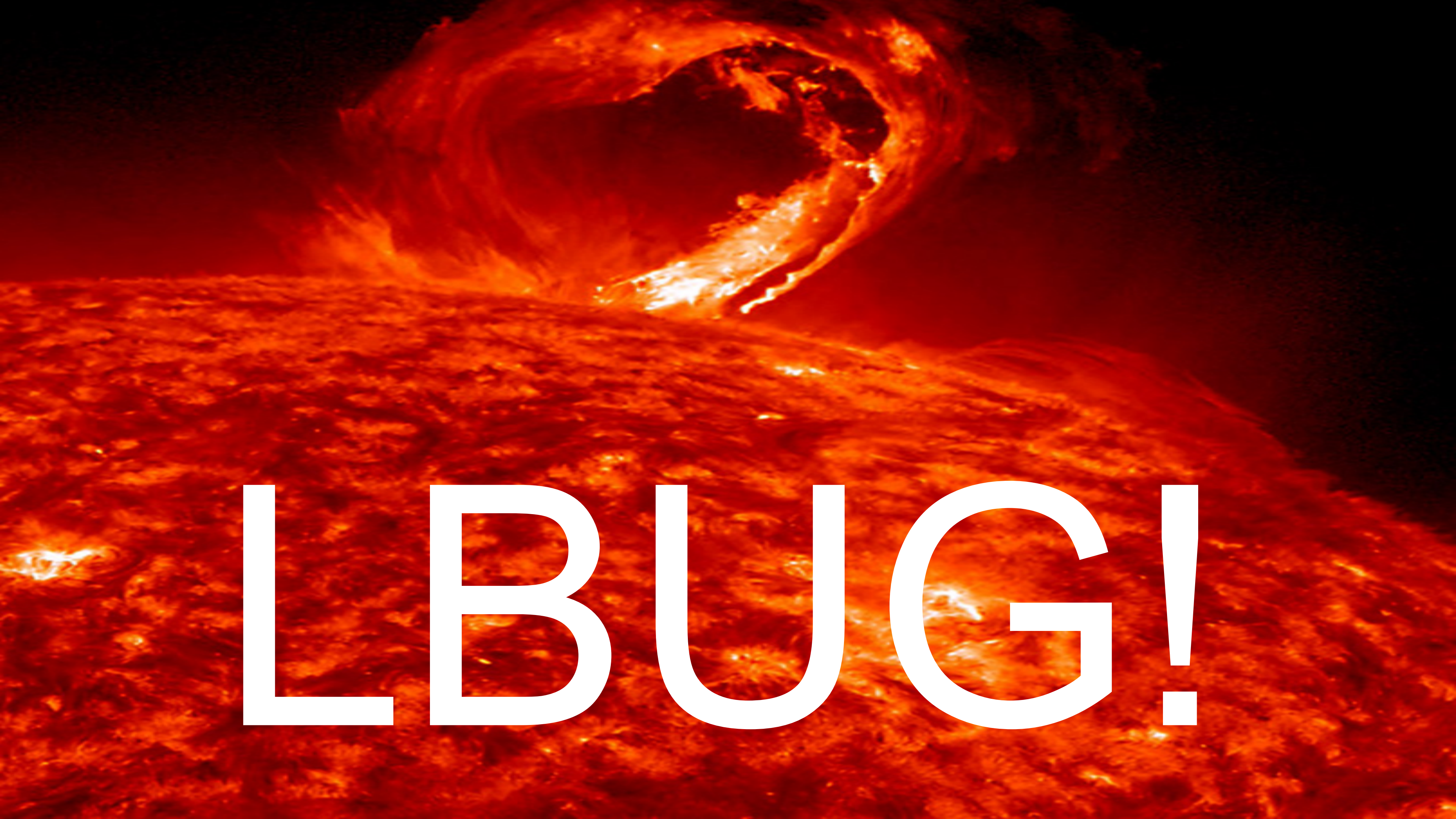
Specific Thanks to Amir Shehata

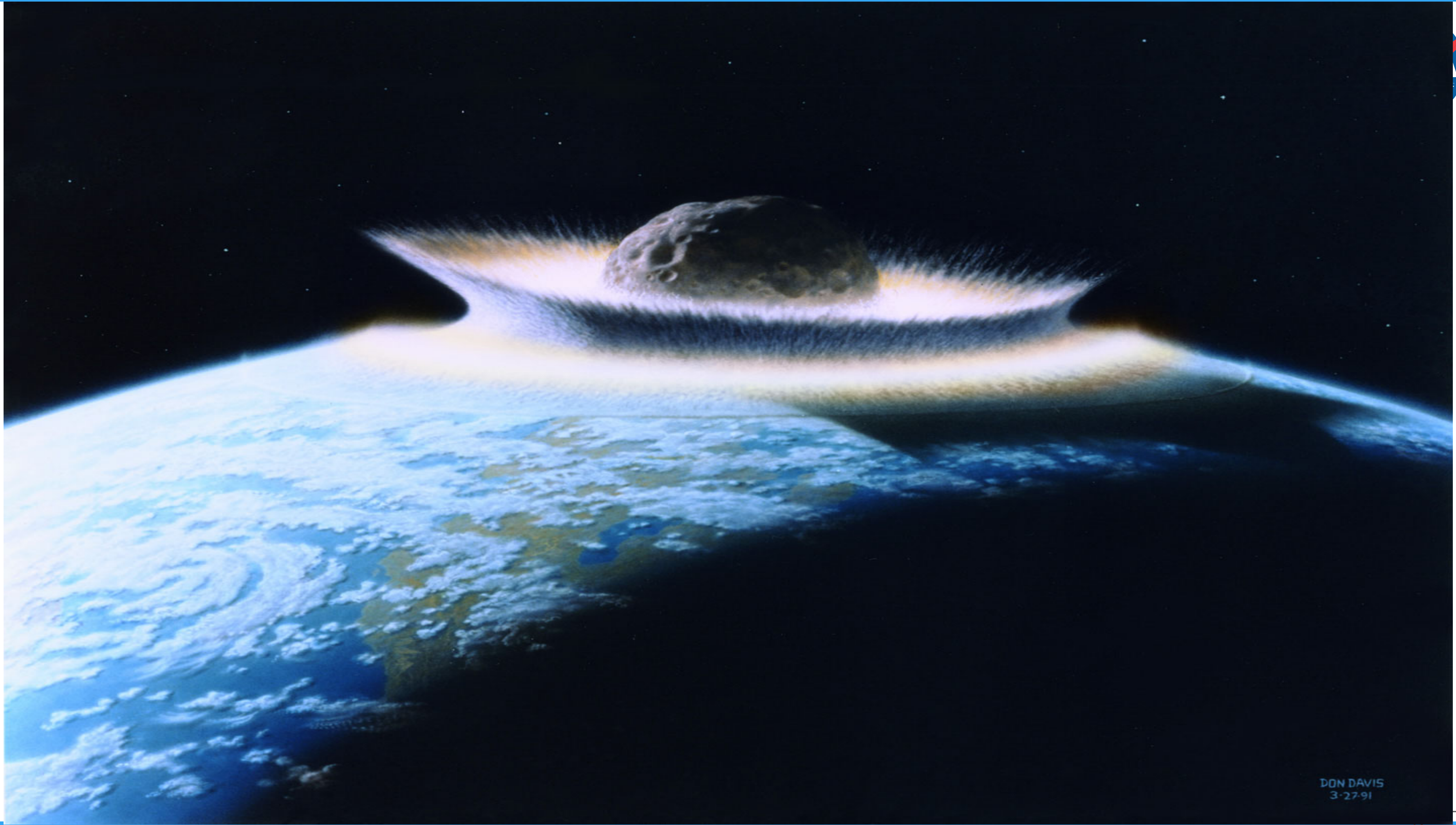
NASA Team:

Mahmoud Hanafi
Dale Talcott
Mike Hartman
Bob Ciotti

Questions

http://www.nas.nasa.gov/hecc


LBUG!


DON DAVIS
3-27-91