Terascale to Petascale: A design for a reliable, scalable network for storage and clusters

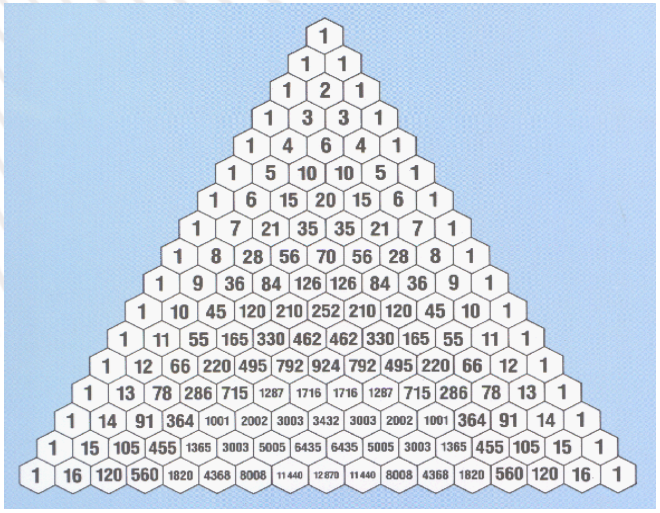
Hsing-bung (HB) Chen, Gary Grider, Parks Fields
HPC-5, Los Alamos National Lab
New Mexico, USA
Date:3/23/2009

www.openfabrics.org

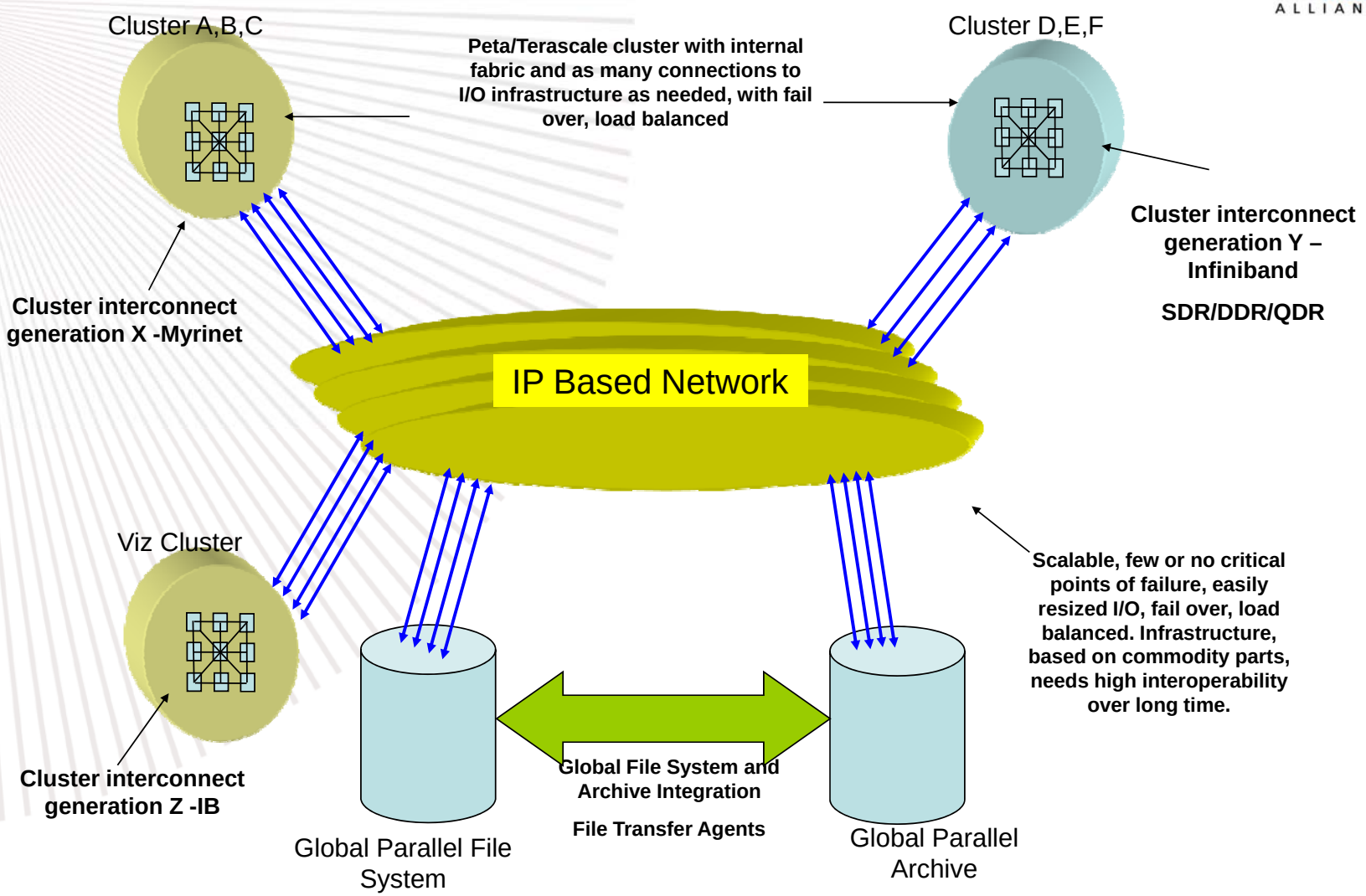
Abstract of Presentation

We present a cost-effective, high bandwidth server I/O network architecture, named PaScal (Parallel and Scalable). We use the PaScal server I/O network to support data-intensive scientific applications running on very large-scale Linux clusters. PaScal server I/O network architecture provides (1) Bi-level interconnection network by combining high speed interconnects for computing Inter-Process Communication (IPC) requirements and low-cost Gigabit Ethernet interconnect for global IP based storage/file access, (2) A bandwidth on demand I/O network architecture without re-wiring and reconfiguring the system, (3) load balancing and failover multi-path routing scheme, (4) Improving reliability through reducing large number of network components in server I/O network, and (5) Supporting global storage/file systems in heterogeneous multi-cluster and Grids environments. We have compared the PaScal server I/O network architecture with the Federated server I/O network architecture. Concurrent MPI-I/O performance testing results and deployment cost comparison demonstrate that the PaScal server I/O network architecture can outperform the Federated server I/O network architecture in many categories: cost-effective, and ease of growing and management very large-scale I/O network.

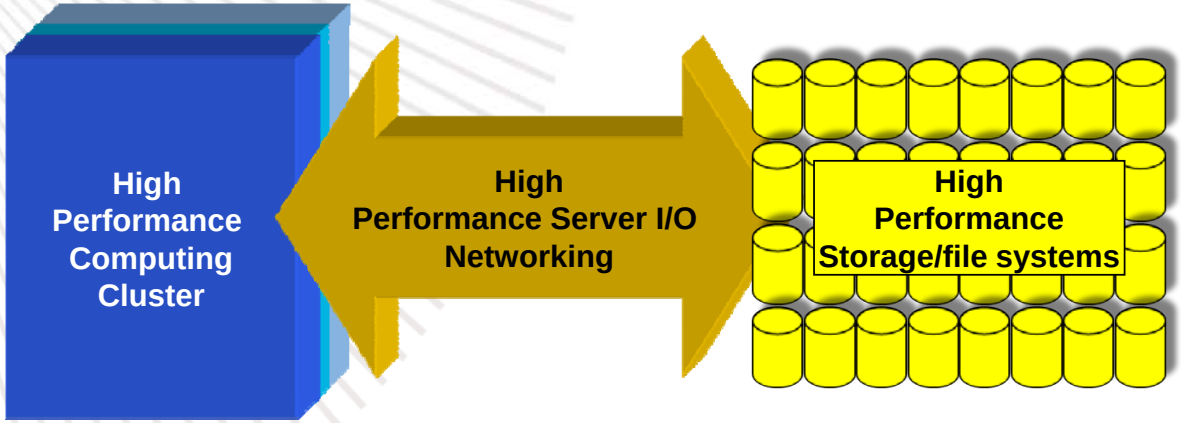
If we deploy this multi cluster shared global parallel file system, we will need a scalable and reliable network that connects the clusters to the file system



High Level Conceptual Cluster Server I/O network
Where we want to be!

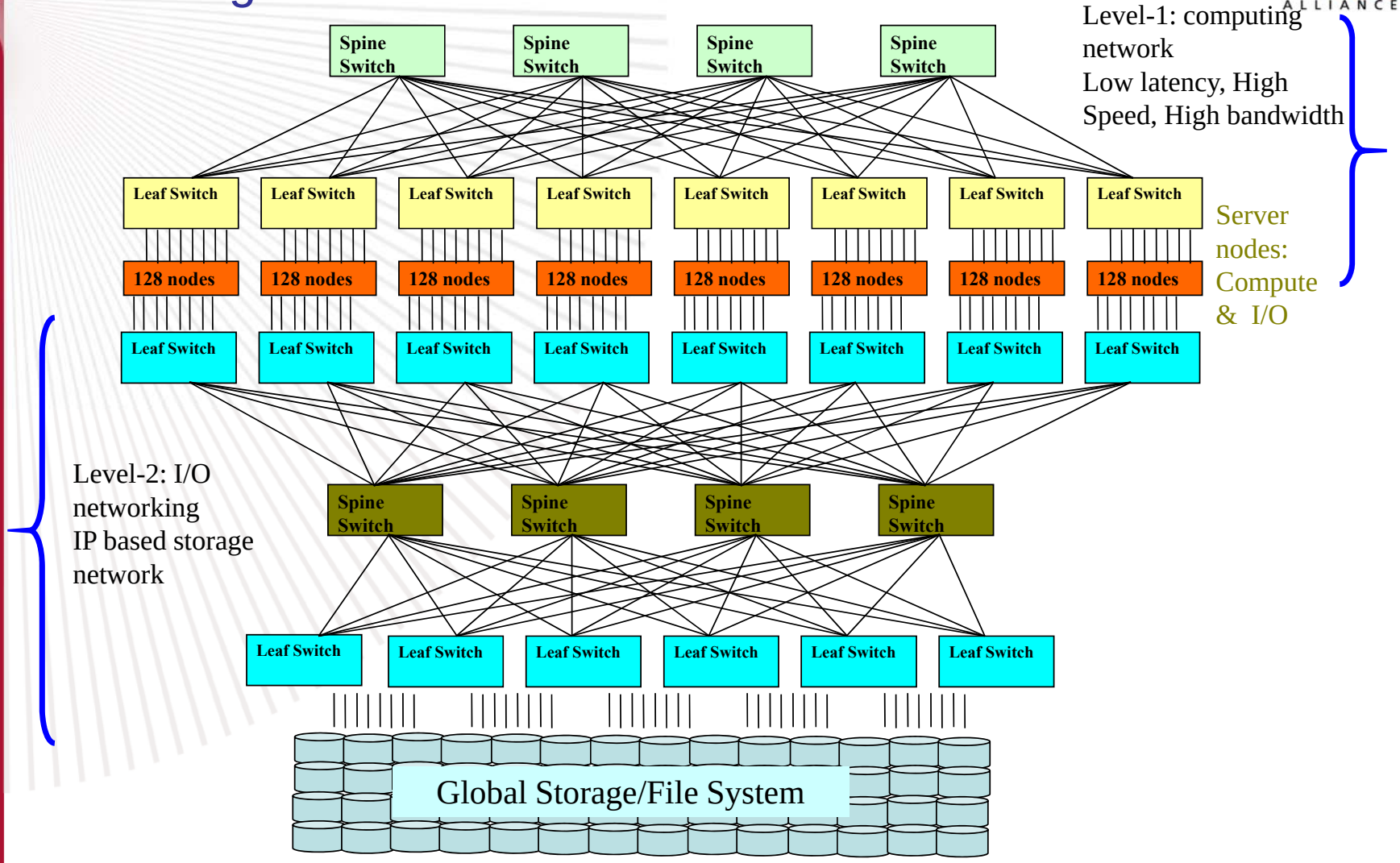


Expensive (all to all connection)
High bandwidth data Path
Parallel but not very scalable I/O networking

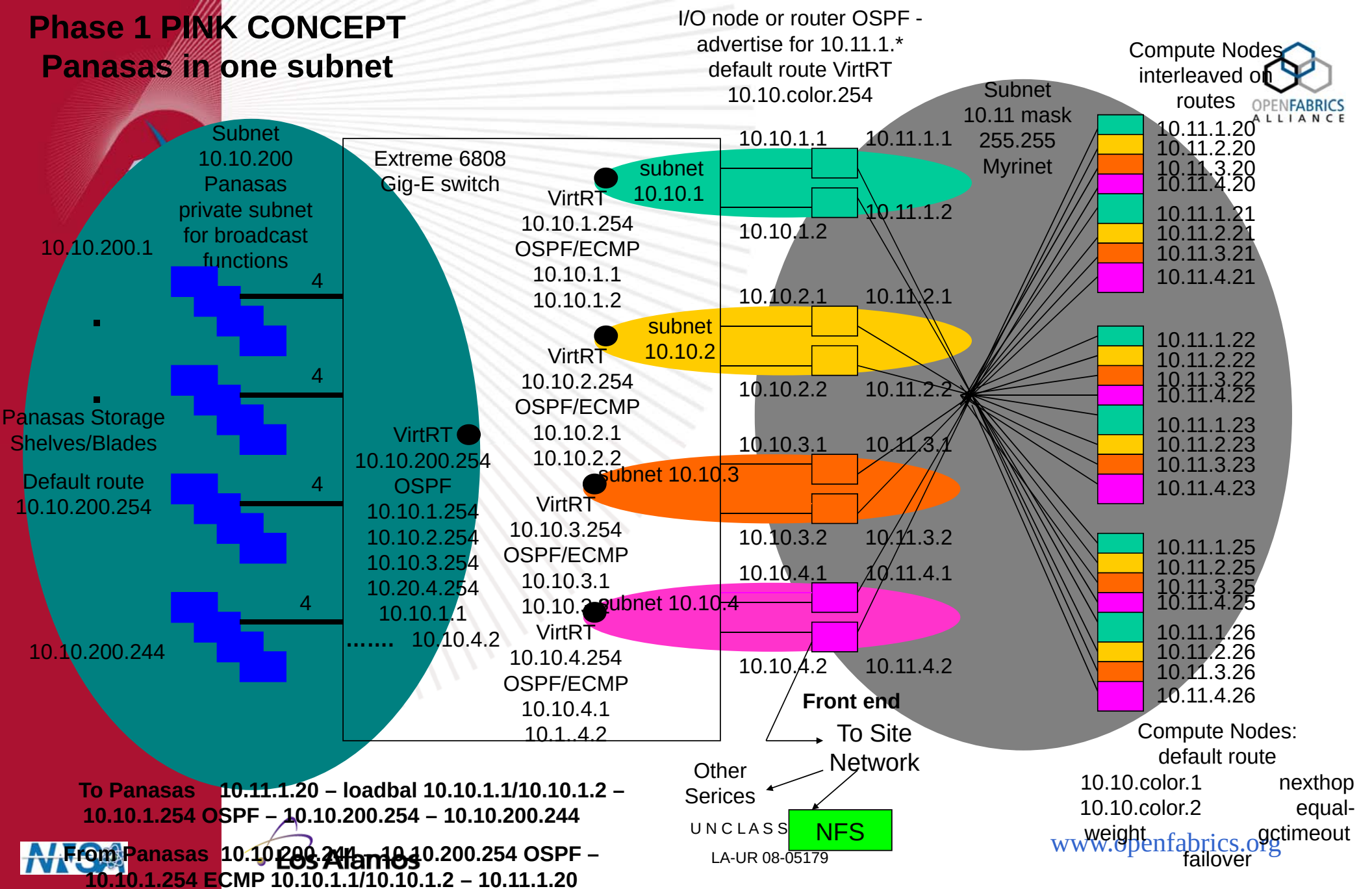


Storage/File system
Multiple mounts

Two Levels of Interconnect – Computing and I/O networking



Phase 1 PINK CONCEPT
Panasas in one subnet



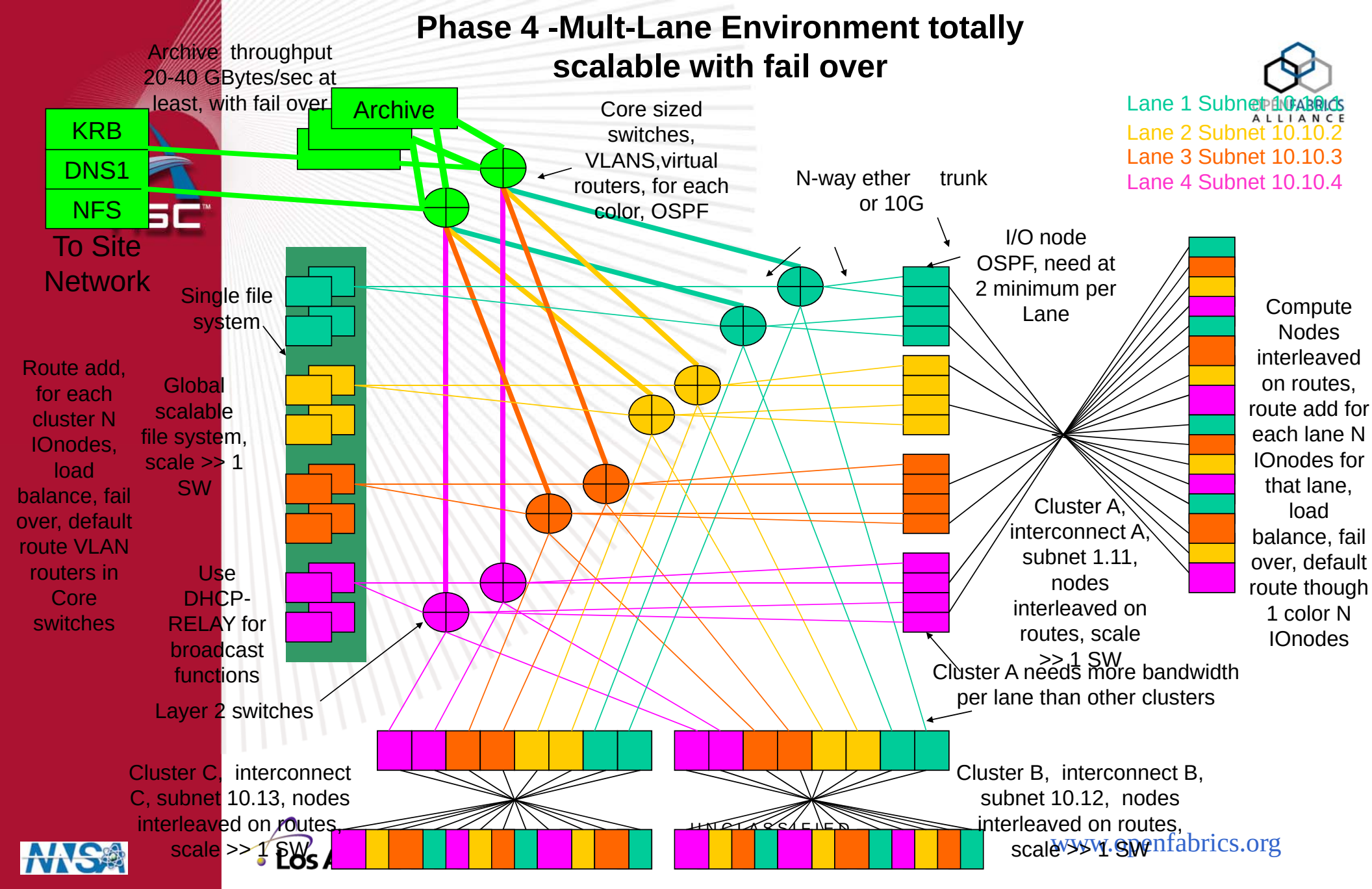
To Panasas 10.11.1.20 – loadbal 10.10.1.1/10.10.1.2 – 10.10.1.254 OSPF – 10.10.200.254 – 10.10.200.244

From Panasas 10.10.200.244 – 10.10.200.254 OSPF – 10.10.1.254 ECMP 10.10.1.1/10.10.1.2 – 10.11.1.20

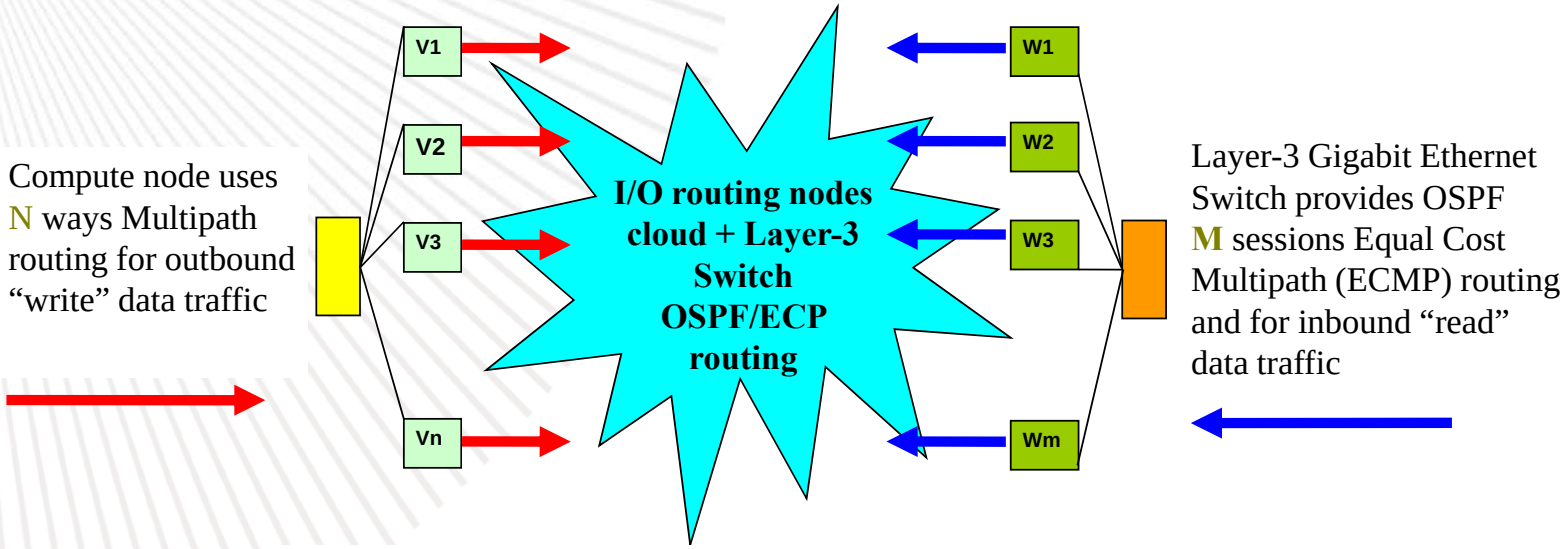
LA-UR 08-05179

www.openfabrics.org

Phase 4 -Mult-Lane Environment totally scalable with fail over



Bi-direction Equal Cost MultiPath routing – load balancing and fail-over



IO nodes cloud uses OSPF to route read and write traffic from the Level-1 and the Level-2 networks

Advantage of PaScalBB Server I/O architecture

- Bi-level switch-fabric interconnected systems by combining high speed interconnects for computing IPC requirement and low-cost Gigabit or 10GIGE Ethernet interconnect for IP based global storage access,
- A bandwidth on demand linear scaling I/O network architecture without re-wiring and reconfiguring the system *,
- load balancing and failover multi-path routing scheme,
- Improve reliability through reducing large number of network components in server I/O network, and
- Support for global storage/file systems in heterogeneous multi-cluster and Grids computing environment.
- Dead Gateway Detection (DGD)

Petascale Storage

- **Massive parallelism**

  O(1K-10K-100K) disk drives

  distributed data coordination and placement

  flexible data stripping

  distributed/parallel metadata servers

  parallel/scalable data rebuild capability

- **Design implications**

  Complexity

  capacity vs. bandwidth – It's about scaling

  adaptively distribute data – distributed control (metadata) and adaptation

DGD – Dead Gateway Detection

- Life still wasn't good enough, we needed more resiliency.
- Why did we do DGD, what was wrong?
  - With IB and 10Gige we start seeing interruptions
- What is DGD
  - Set of scripts that monitors interfaces to make sure data can be passed.
  - If a bad interface is found then adjust the routes on the compute nodes

DGD: Reliable, Highly available, and Manageable, Petascale High Performance Cluster Systems

- **Reliable –**

  24/7 - no service interruption (Computing, Networking, I/O, Storage)

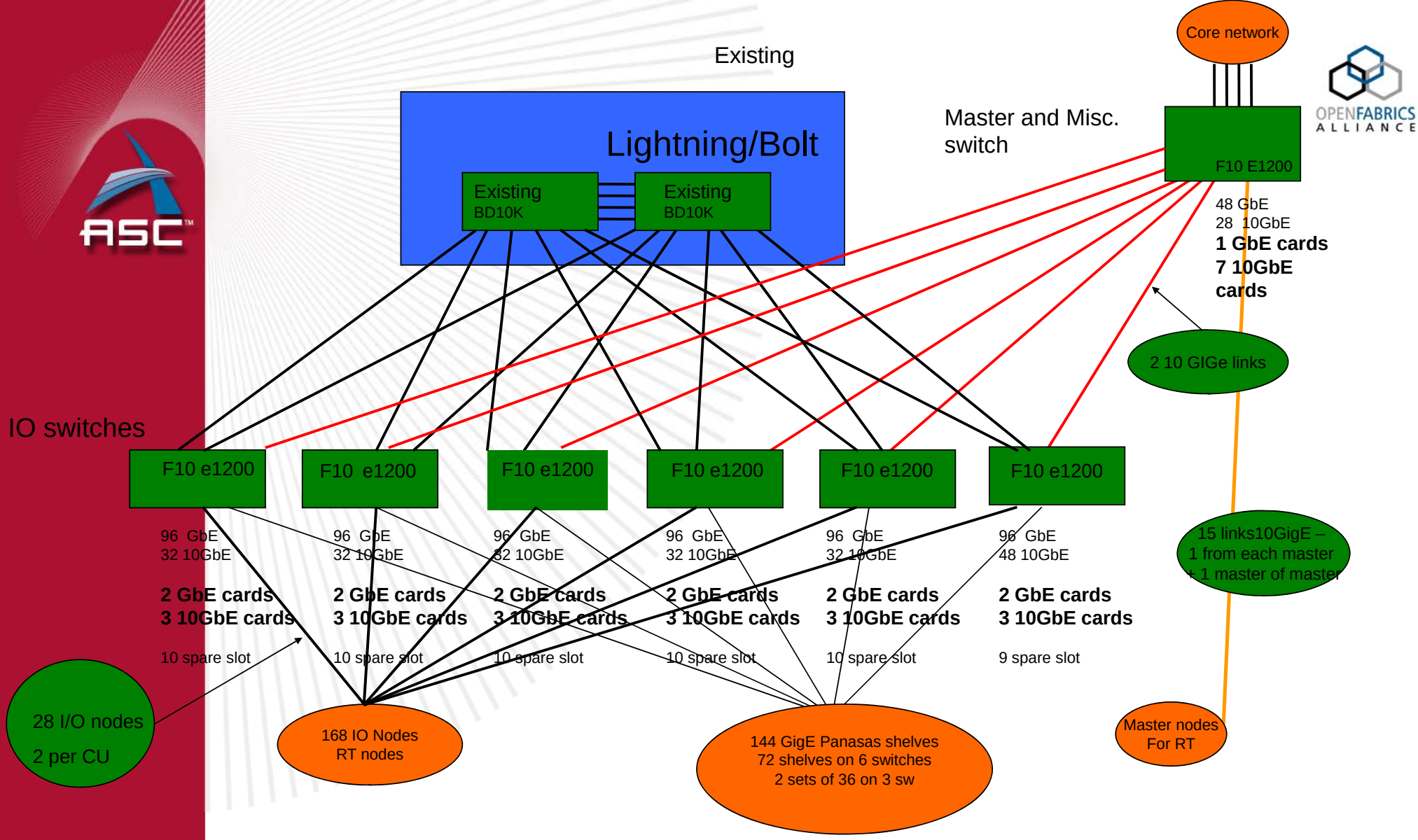
- **Highly Available –**

  Fault resistance, Fail-over, Fault tolerance, If hardware is designed

- **Manageable –**

  1K -10k – 100K Sever nodes (Multisocket / Multicore)

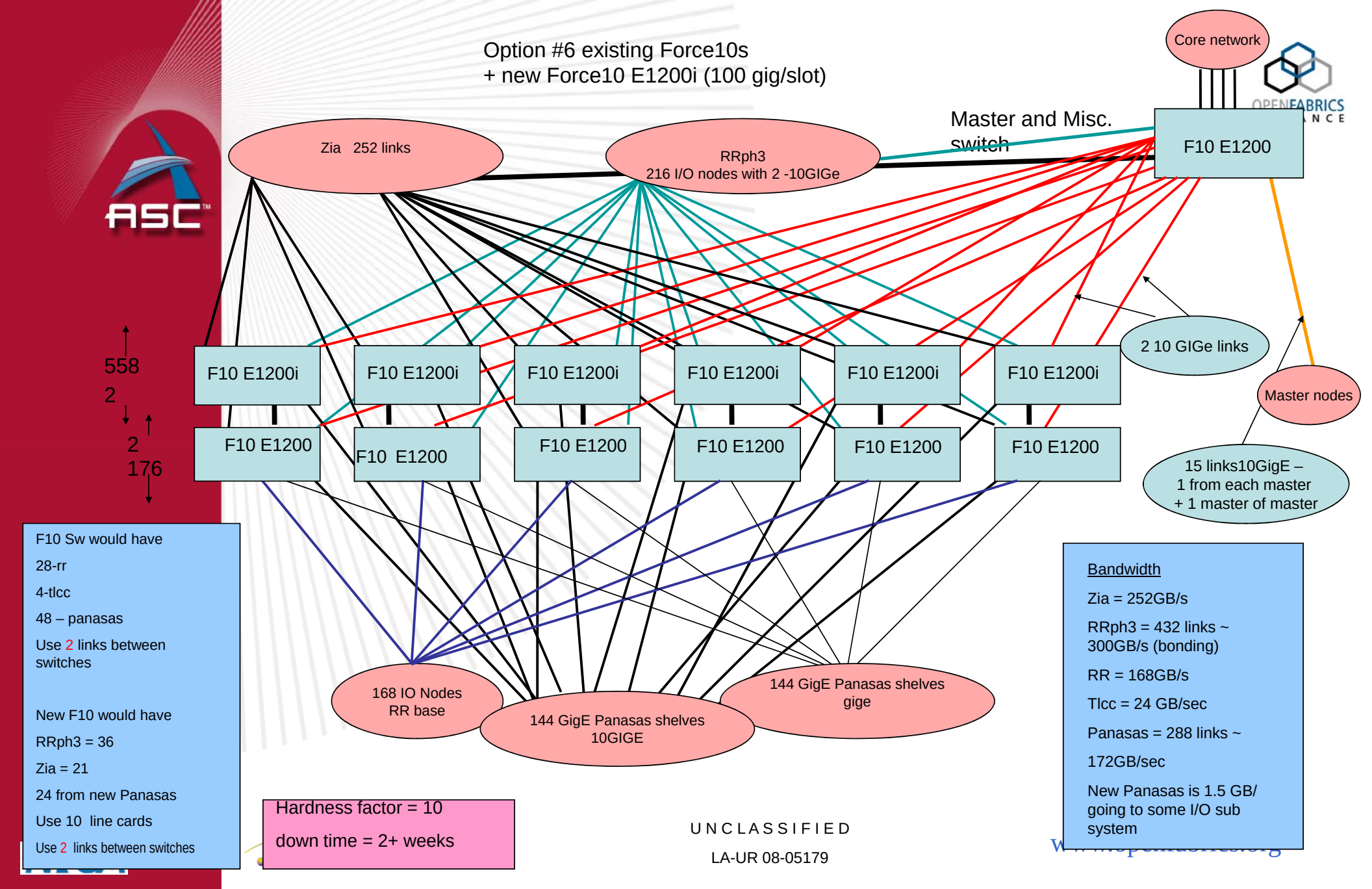
  Petascale Computing / Parallel IO / Storage /Global File systems


Existing
Lightning/Bolt
Existing
BD10K
Existing
BD10K
Master and Misc.
switch
Core network
F10 E1200
48 GbE
28 10GbE
1 GbE cards
7 10GbE
cards
2 10 GIGe links
IO switches
F10 e1200
F10 e1200
F10 e1200
F10 e1200
F10 e1200
F10 e1200
96 GbE
32 10GbE
2 GbE cards
3 10GbE cards
10 spare slot
96 GbE
32 10GbE
2 GbE cards
3 10GbE cards
10 spare slot
96 GbE
32 10GbE
2 GbE cards
3 10GbE cards
10 spare slot
96 GbE
32 10GbE
2 GbE cards
3 10GbE cards
10 spare slot
96 GbE
32 10GbE
2 GbE cards
3 10GbE cards
10 spare slot
96 GbE
48 10GbE
2 GbE cards
3 10GbE cards
9 spare slot
15 links10GigE –
1 from each master
+ 1 master of master
28 I/O nodes
2 per CU
168 IO Nodes
RT nodes
144 GigE Panasas shelves
72 shelves on 6 switches
2 sets of 36 on 3 sw
Master nodes
For RT

So how large can we scale?

- Capacity of F10 E1200 / 6 lanes today
  - 14 slots = ~84 GB/sec per switch * 6 switch= 504GB/sec
    - 56 gigabits to back plane per slot
    - 16 port 10gige cards (4to1) over subscribed
    - With 2 gige cards and the rest 10gige cards = 1152 10gige ports
- This summer adding 6 more lanes
    - These are new E1200i at a 100gigbits per slot
    - 1.4 Tbits per switch * 6 = 8.4 tb/sec or 1.05 TB/sec
    - 40 port 10gige cards for a total of 3360 ports (4to1)
- Total ~1.5 TB/sec BW to I/O nodes and Storage and 4512 ports of 10gige (need to buy some cards)

Option #6 existing Force10s + new Force10 E1200i (100 gig/slot)



F10 Sw would have

28-rr

4-tlcc

48 – panasas

Use 2 links between switches

New F10 would have

RRph3 = 36

Zia = 21

24 from new Panasas

Use 10 line cards

Use 2 links between switches

Hardness factor = 10

down time = 2+ weeks

Bandwidth

Zia = 252GB/s

RRph3 = 432 links ~ 300GB/s (bonding)

RR = 168GB/s

Tlcc = 24 GB/sec

Panasas = 288 links ~

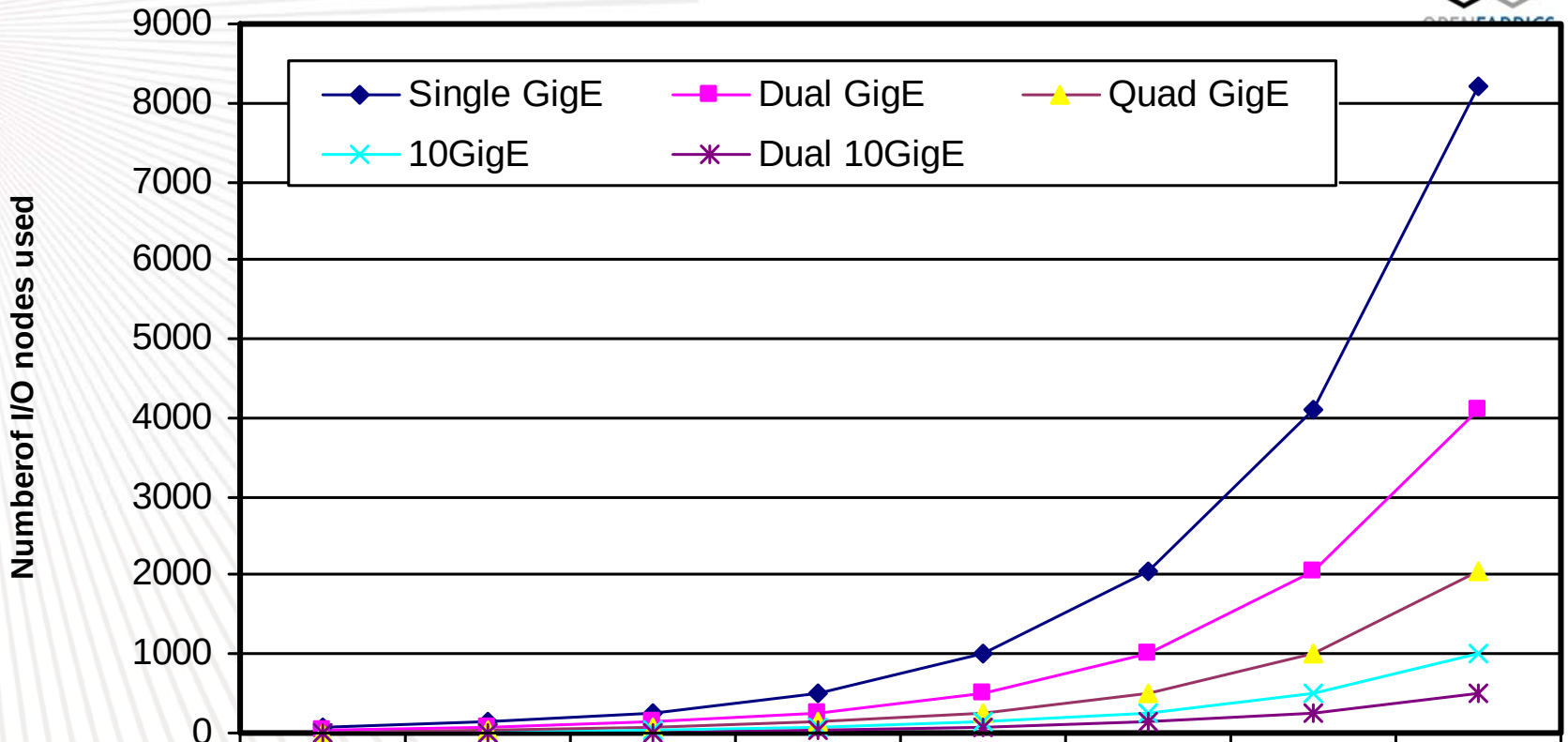
172GB/sec

New Panasas is 1.5 GB/ going to some I/O sub system

UNCLASSIFIED

LA-UR 08-05179

TeraScale to PetaScale



	8GB	16GB	32GB	64GB	128GB	256GB	512GB	1024GB
Single GigE	64	128	256	512	1024	2048	4096	8192
Dual GigE	32	64	128	256	512	1024	2048	4096
Quad GigE	16	32	64	128	256	512	1024	2048
10GigE	8	16	32	64	128	256	512	1024
Dual 10GigE	4	8	16	32	64	128	256	512

I/O Data Bandwidth **GB/sec**

Conclusions: PaScalBB I/O

1. It is very cost-effective and easy to grow Level-2 server I/O network,
2. It can scale very well as the system keeps growing,
3. It eliminates the I/O routing interferences on back-end Computer nodes and reduce significant amount of interactions between applications and operation system hence provides a “noiseless” operating system and allow applications to use as many cycles as possible,
4. Smart Bi-direction Equal-Cost MultiPath routing
5. It has no redundant network on compute nodes
6. It provides load balancing between the Level-1 and the Level-2 networks, and
7. It has much less NIC cable installation and complicated cable management overhead.

Future Works

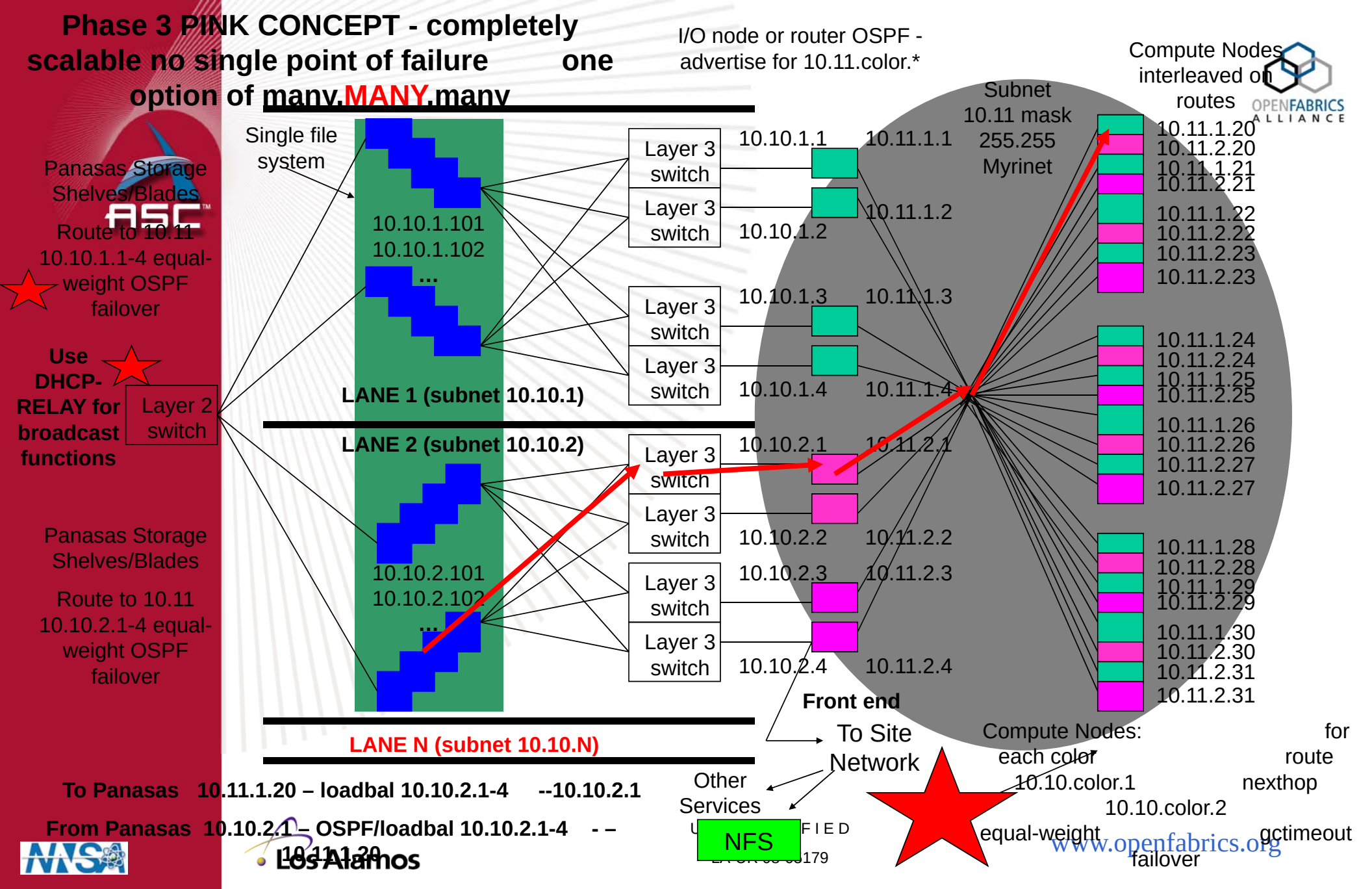
1. Passive Dead I/O node detect and react (FS/IO Networking)
2. Storage side fail over;
   Depends on Panasas implementation of shelf network fail over
3. Server side pull protocol exploration (ISER)
   Only exploratory work so far (depends on standards)
4. Possible need for function shipping if we end up with a machine that has no OS capable of file system client
   Sandia/ANL/IBM are working on this for BG/L and RS, will leverage if needed.

Bonus Slides AKA: backup slides

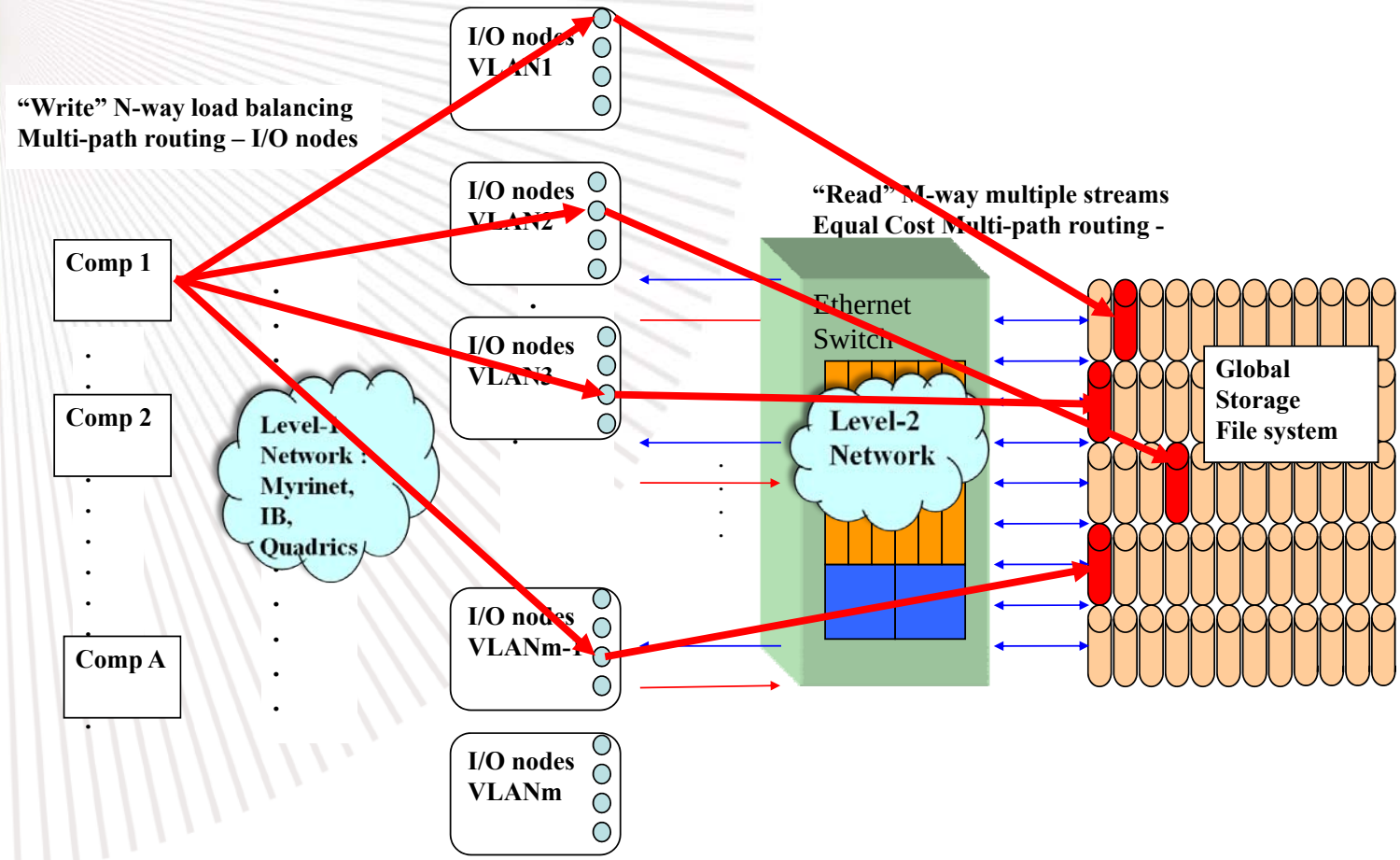
More on Failover and resiliency

- We connect the 12 switches together by 2 10gige links, saving ports for the system
- We set up routes on I/O nodes when clusters have on 12 I/O nodes or if they have only 6 I/O we use a 2 port 10gige nic. Then routes are set to use “sister” switches to get to Panasas.

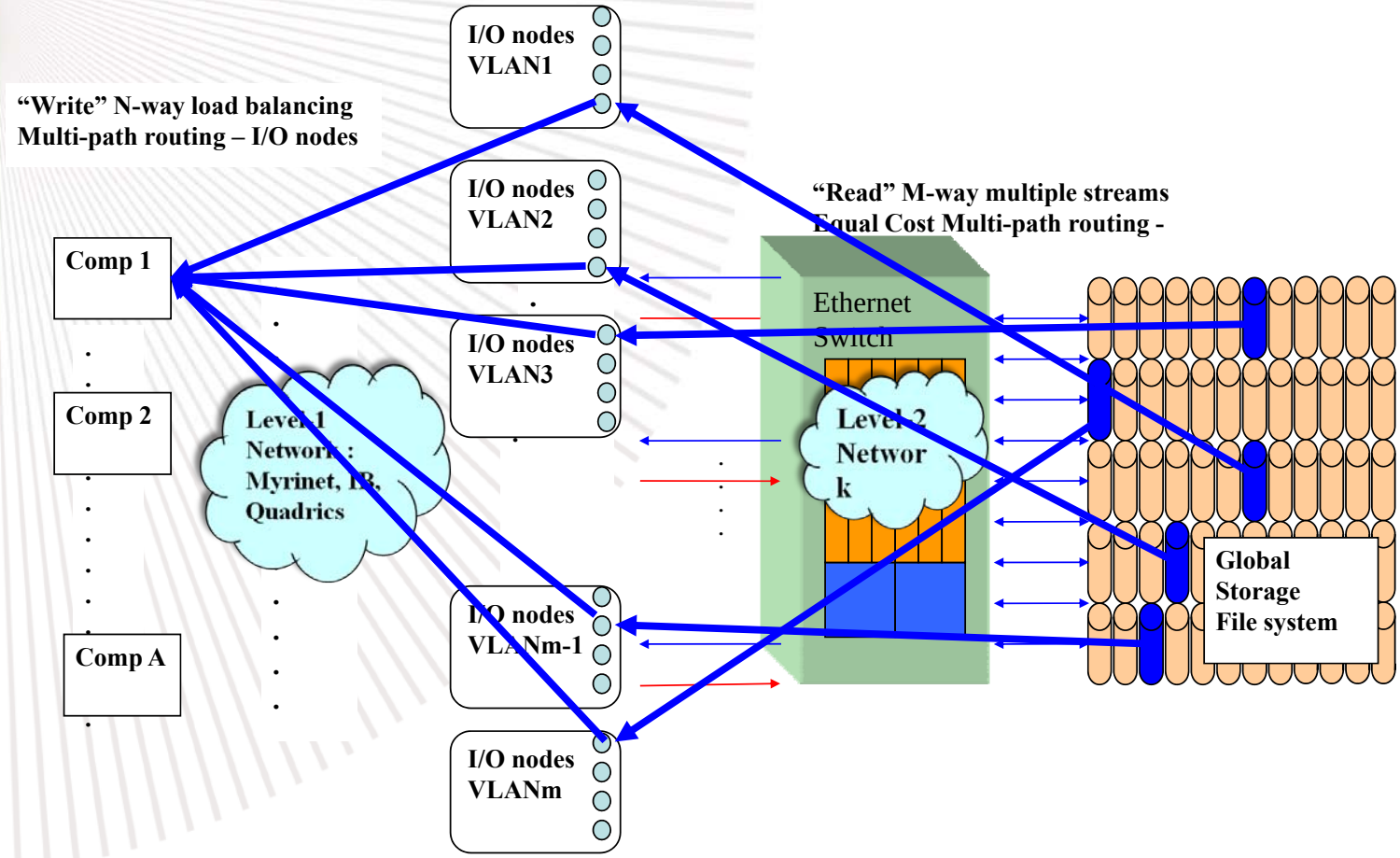
Phase 3 PINK CONCEPT - completely scalable no single point of failure one option of many.MANY.many



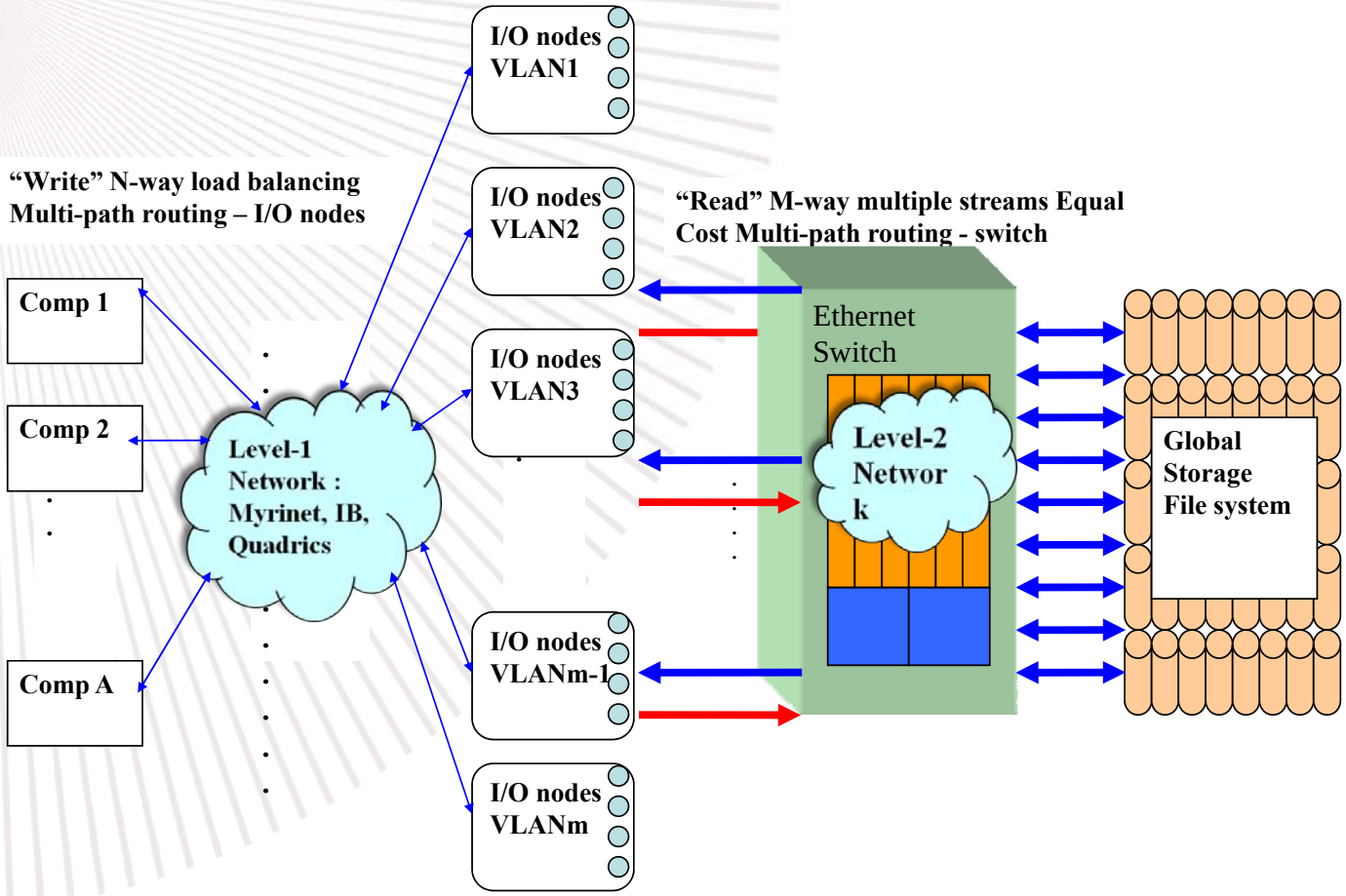
Write – N-way Multipath routes



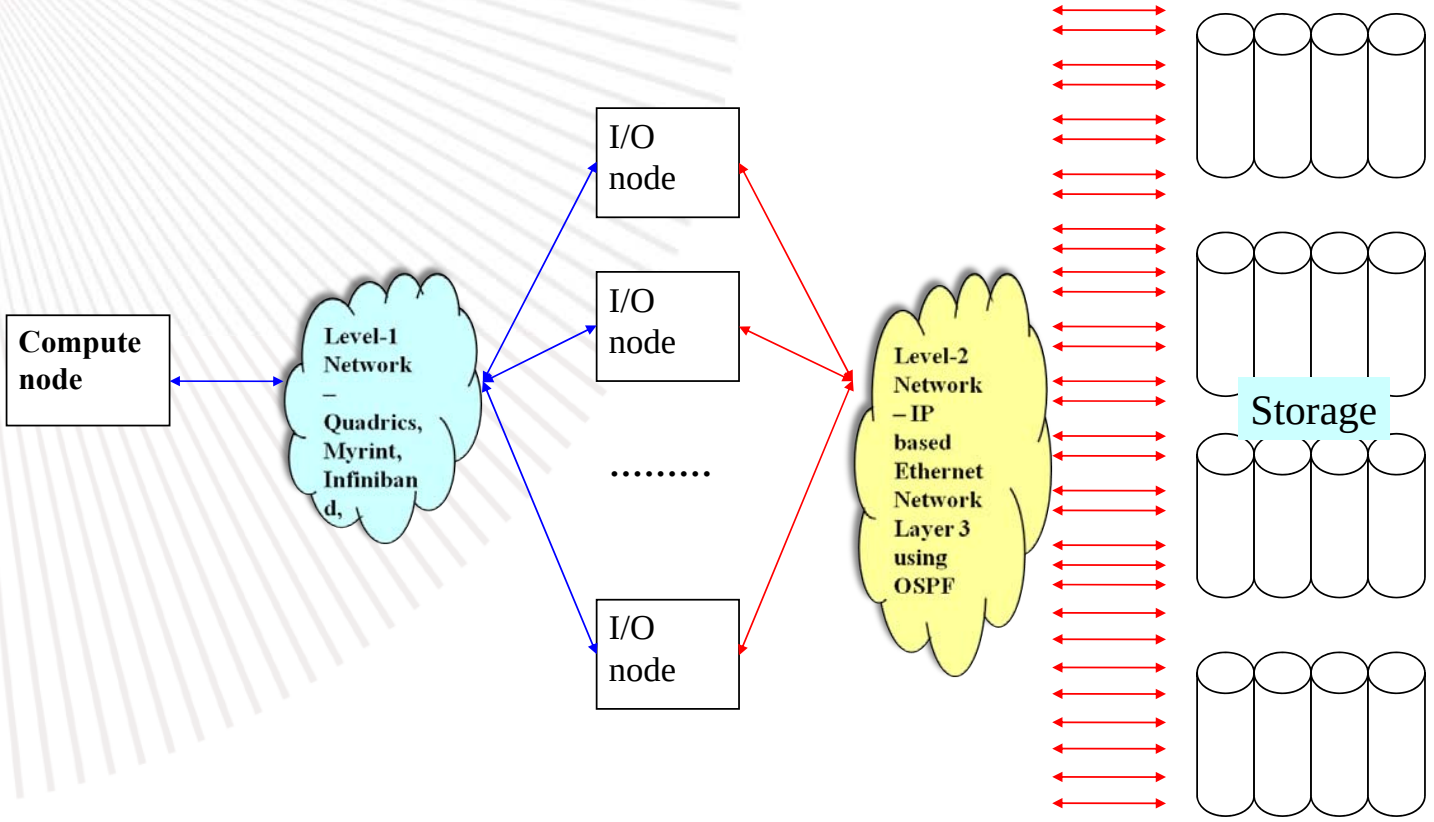
READ – M way Equal Cost MultiPath routes

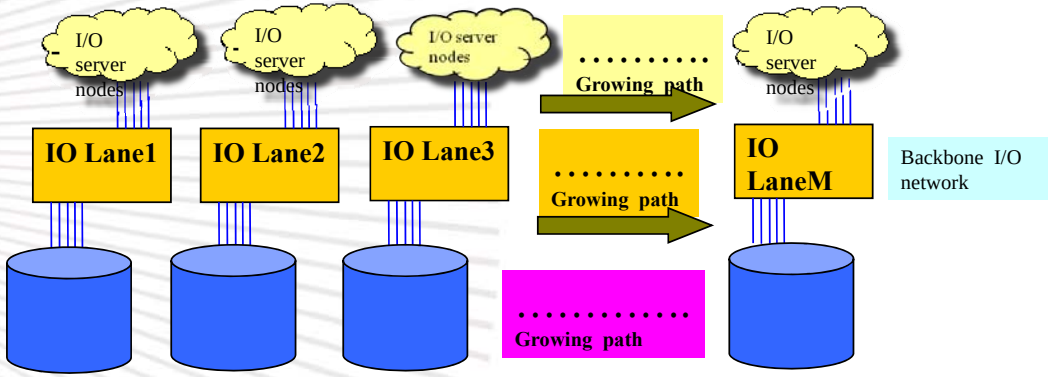


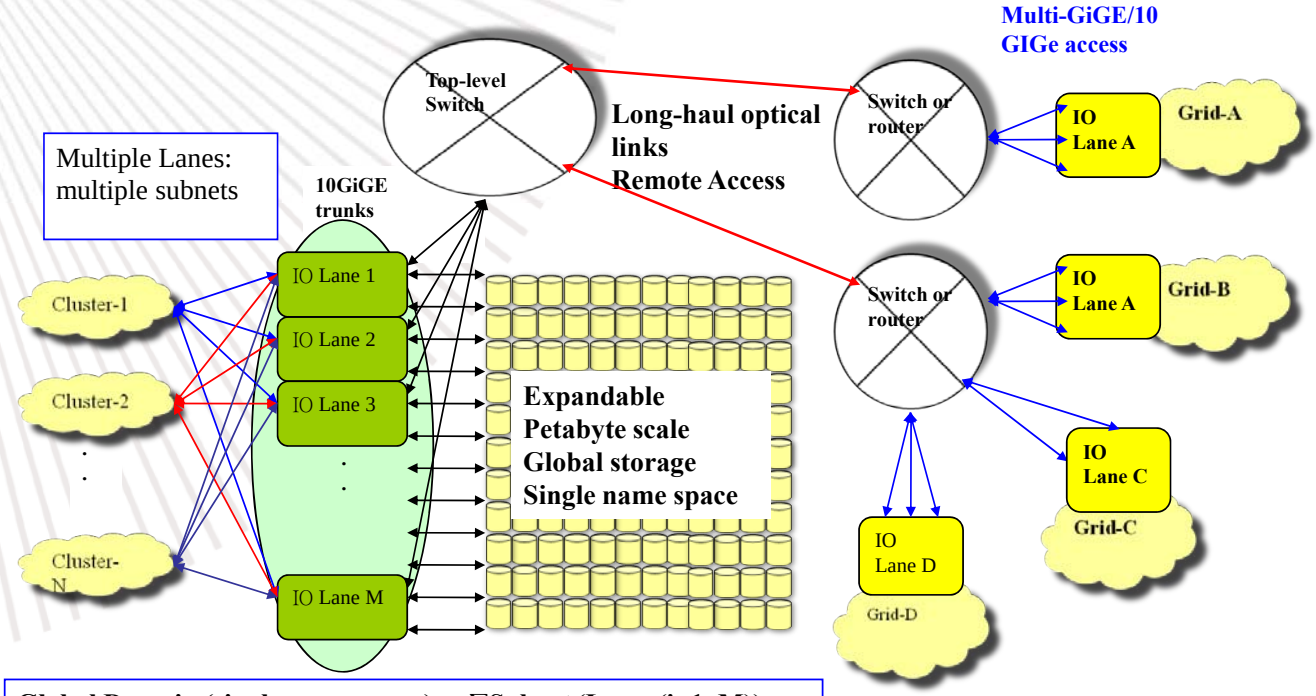
Bi-direction Extreme Load balancing



Compute node concurrent I/O using Equal Cost Multi-Path routing – load balancing and fail-over



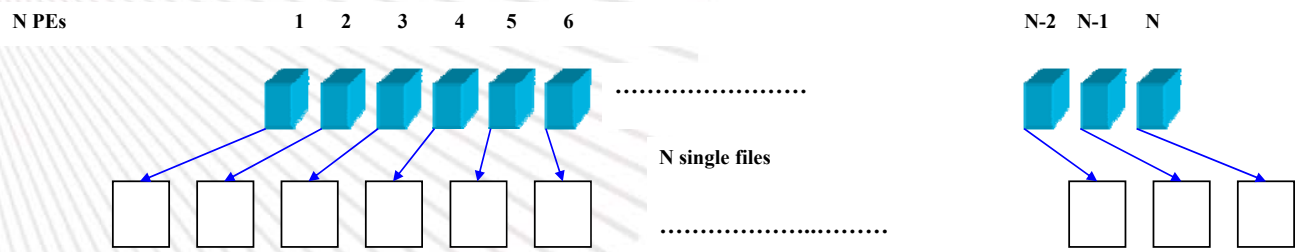




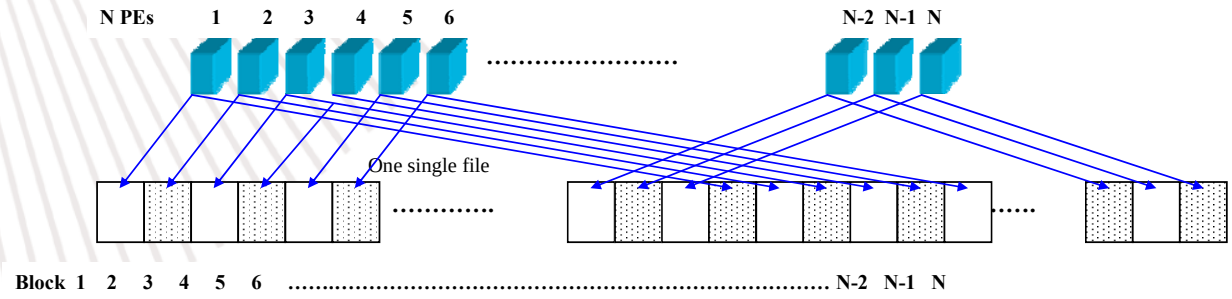
Global Domain (single name space) = ∑Subnet (Lane$_i$ (i=1..M)),
Level-2 Linear scaling IP routing network

Concurrent I/O access patterns

N-to-N



N-to-1



The future holds more capability!

	ZIA	TRINITY
Peak PF	> 2	> 50
Total memory	> 0.5 PB	> 5 PB
Aggregate[a] Memory BW	> 1 PB/sec	> 5 PB/sec
Aggregate Interconnect BW	> 1 PB/sec	> 7 PB/sec
Aggregate Bisection BW[b]	> 80 TB/sec	> 450 TB/sec
Aggregate Message Rate	> 10 GMsgs/sec	> 80 GMsgs/sec
Aggregate I/O BW	**> 1 TB/sec**	**> 10 TB/sec**
Disk Capacity	> 20 PB	> 200 PB
System Power (MW)	5 - 8	10 - 16
Floor Space (sq ft)	< 8,000	< 8,000
MTTI (Job) / MTBF (System) (Both @ Full Scale)	> 50 / > 200 Hrs.	> 50 / > 200 Hrs.