OPENFABRICS ALLIANCE

12th ANNUAL WORKSHOP 2016

Fabrics and Topologies for Directly Attached Parallel File Systems and Storage Networks

Susan Coulter, Kyle Lamb, Jesse Martinez, Dominic Manno, Parks Fields, Mike Mason, Brad Settlemyer

Los Alamos National Laboratory

[April 6th, 2016]

Los Alamos NATIONAL LABORATORY EST. 1943

LA-UR-16-21831

LANL transitions to Lustre

Fine Grain Routing (FGR) – To implement or not

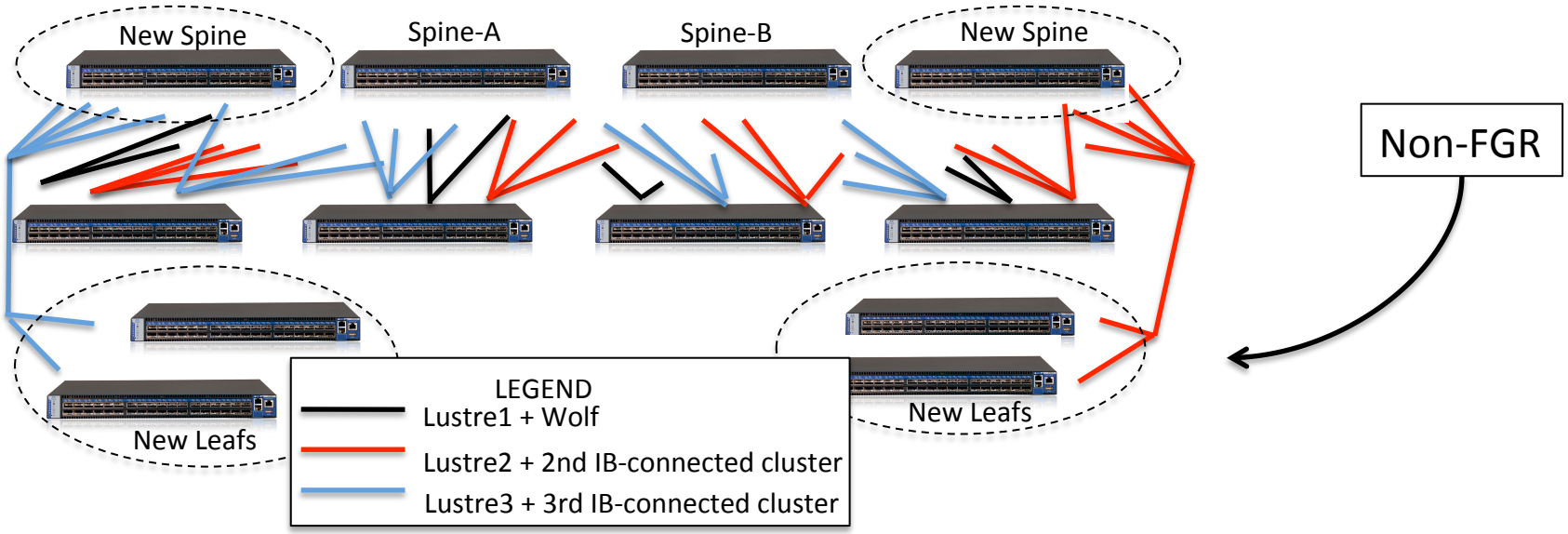
- Long term configuration
  - 3 Lustre file systems
  - 3 IB-connected clusters
- Non-FGR: Close impedance match between hosts and uplinks
- Hardware requirements (switches & cables)

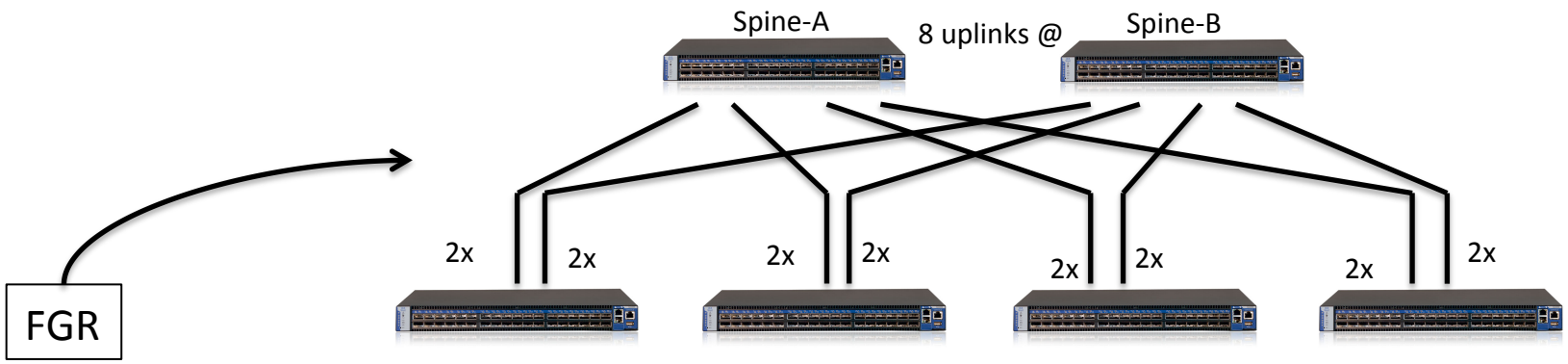
- Support model

References: I/O Congestion Avoidance via Routing and Object Placement
D. A. Dillow, G. M. Shipman, S. Oral, and Z. Zhang (CUG 2011)

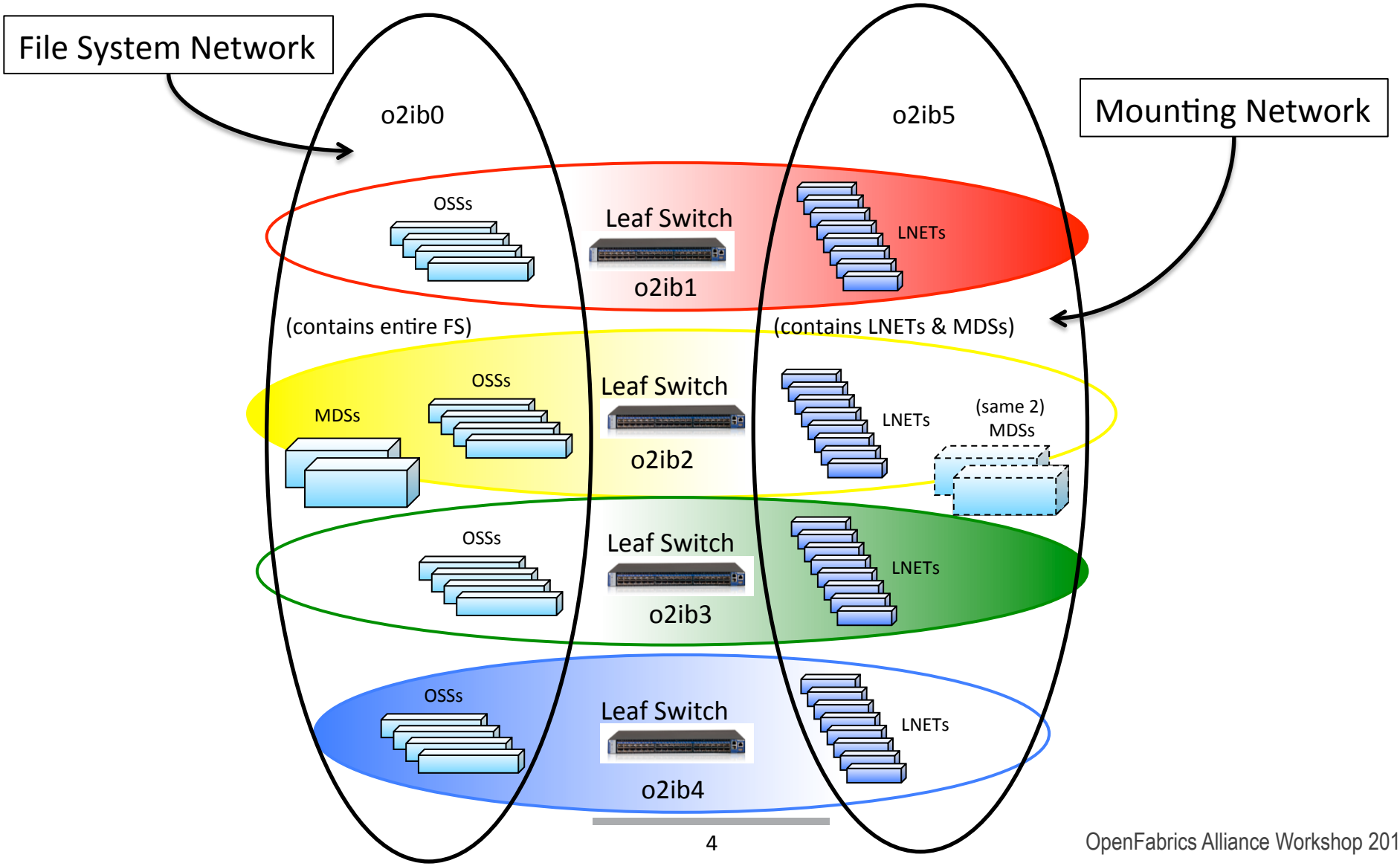
Acknowledgements: Bob Pearson, Dave McMillen (Cray), Oz Rentas (DDN), David Sherrill (LANL)

Non-FGR versus FGR

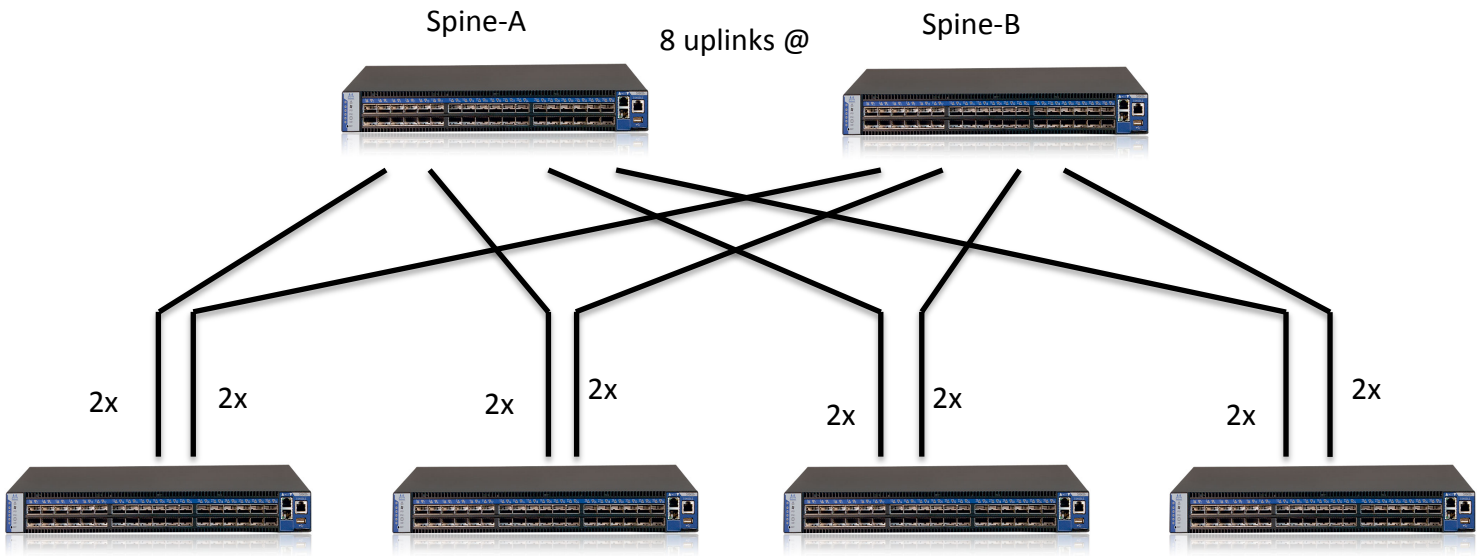




LANL implements FGR



Deployment after L1 and L2



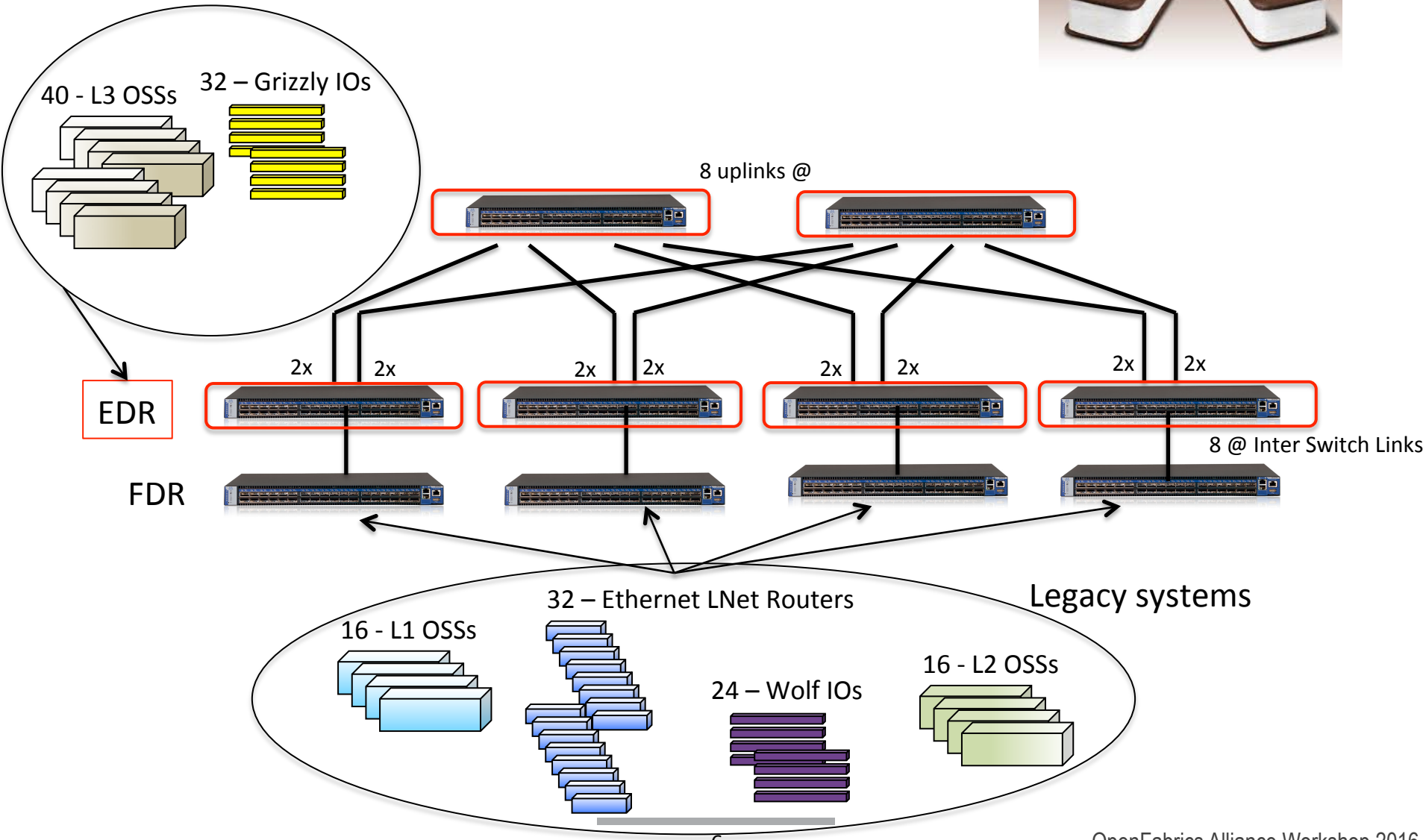
L1 – 16 OSSs

L2 – 16 OSSs

32 LNet routers convert from Ethernet to IB

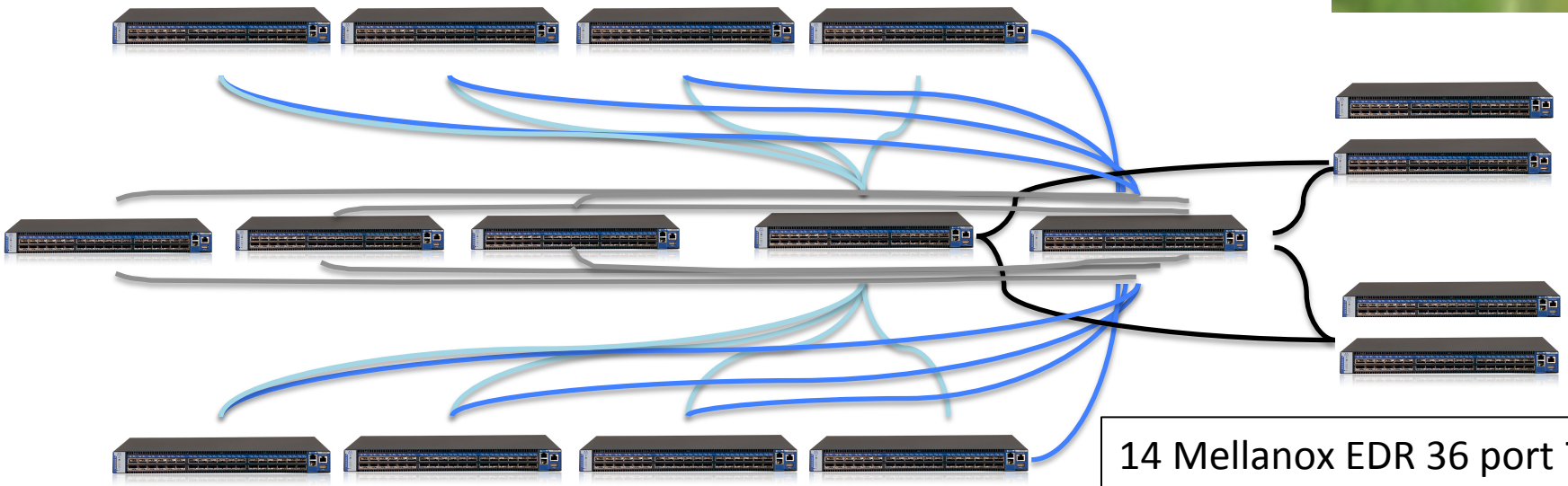
Turquoise L3 Expansion

(the ice cream sandwich)



Classified Network Storage Backbone

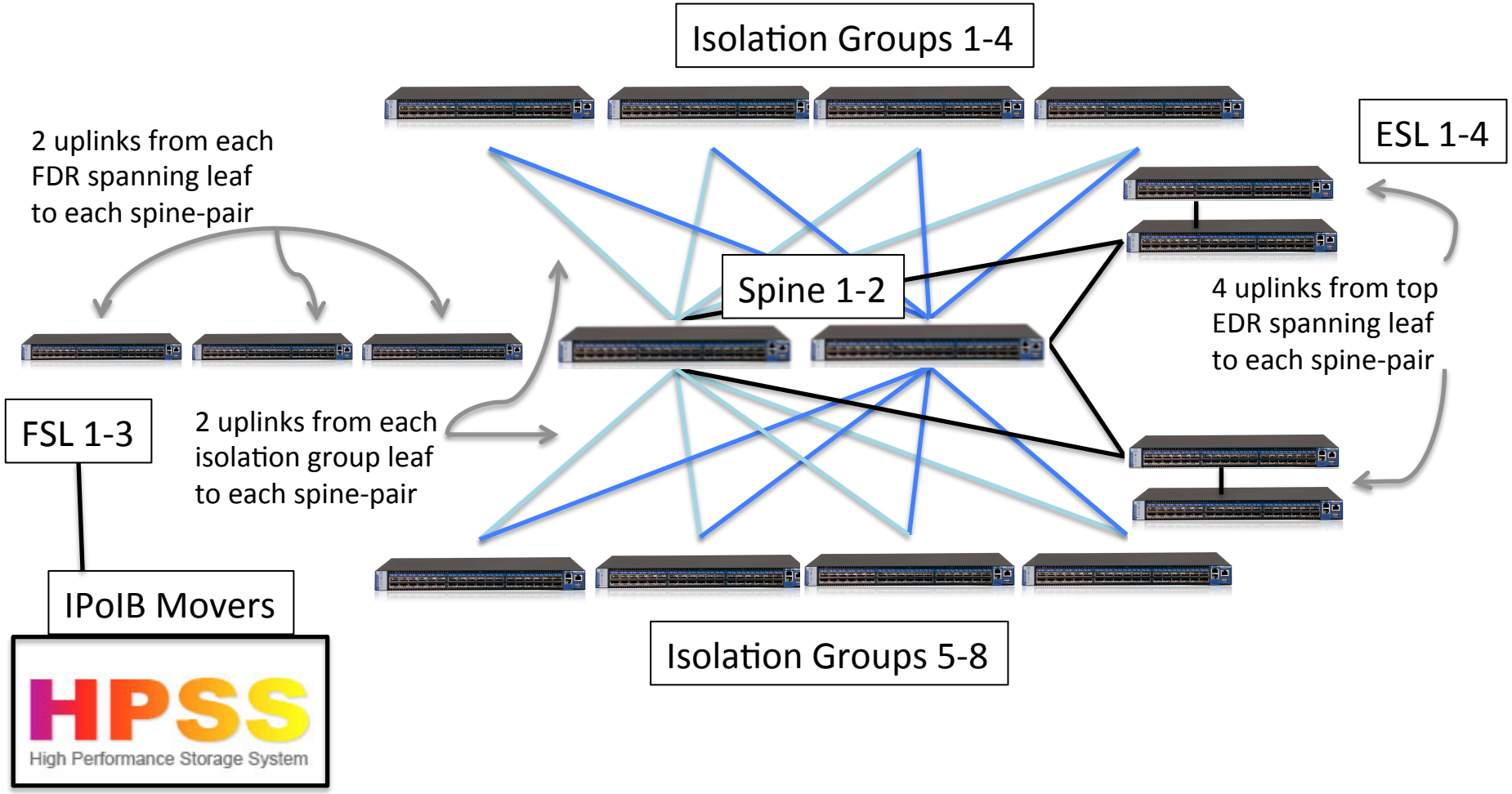
Damselfly Topology



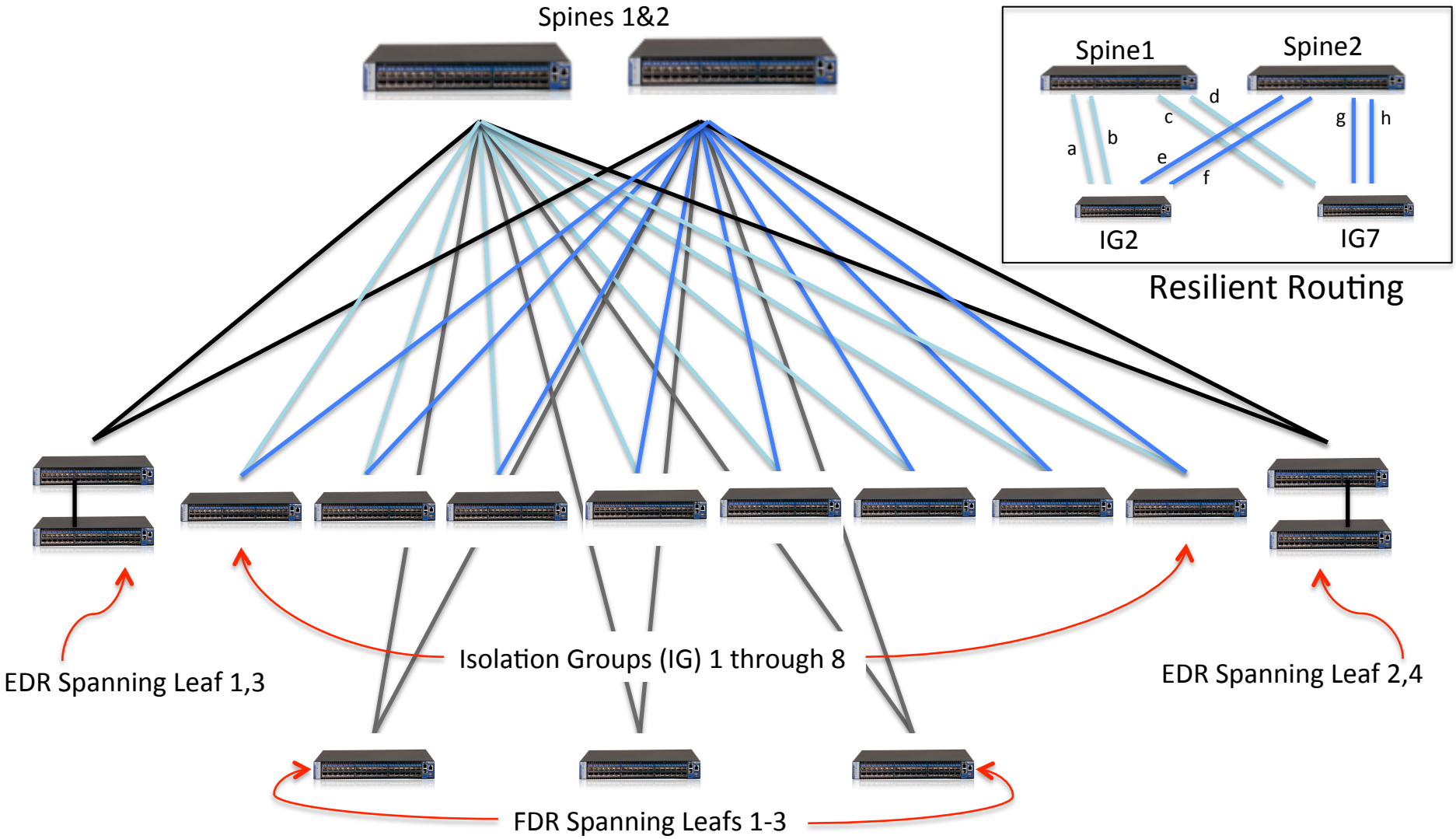
14 Mellanox EDR 36 port 7700s
3 Mellanox FDR 36 port 6036s

Damselflies (suborder Zygoptera) are insects in the order Odonata. Damselflies are similar to dragonflies, but the adults can be distinguished by the fact that the wings of most damselflies are held along, and parallel to, the body when at rest. Additionally, the eyes of a damselfly are separated; one on each side of it's head.

Damselfly


Isolation Groups 1-4
ESL 1-4
2 uplinks from each
FDR spanning leaf
to each spine-pair
Spine 1-2
4 uplinks from top
EDR spanning leaf
to each spine-pair
FSL 1-3
2 uplinks from each
isolation group leaf
to each spine-pair
IPoIB Movers
HPSS
High Performance Storage System
Isolation Groups 5-8

Damselfly
Routing: Up/Down or MinHop



Damselfly Connections and Port Counts

Spines: 16 IG uplinks
(@) 14 SL uplinks
2 MDSs
1 SM
3 spare ports

Isolation Groups - 8
4 uplinks, 32 hosts @
Total: 256 hosts

(40 OSSs, 48 CTS-1 LNet routers)
(48 Luna LNet routers, 8 VM LNet routers)
112 spare ports

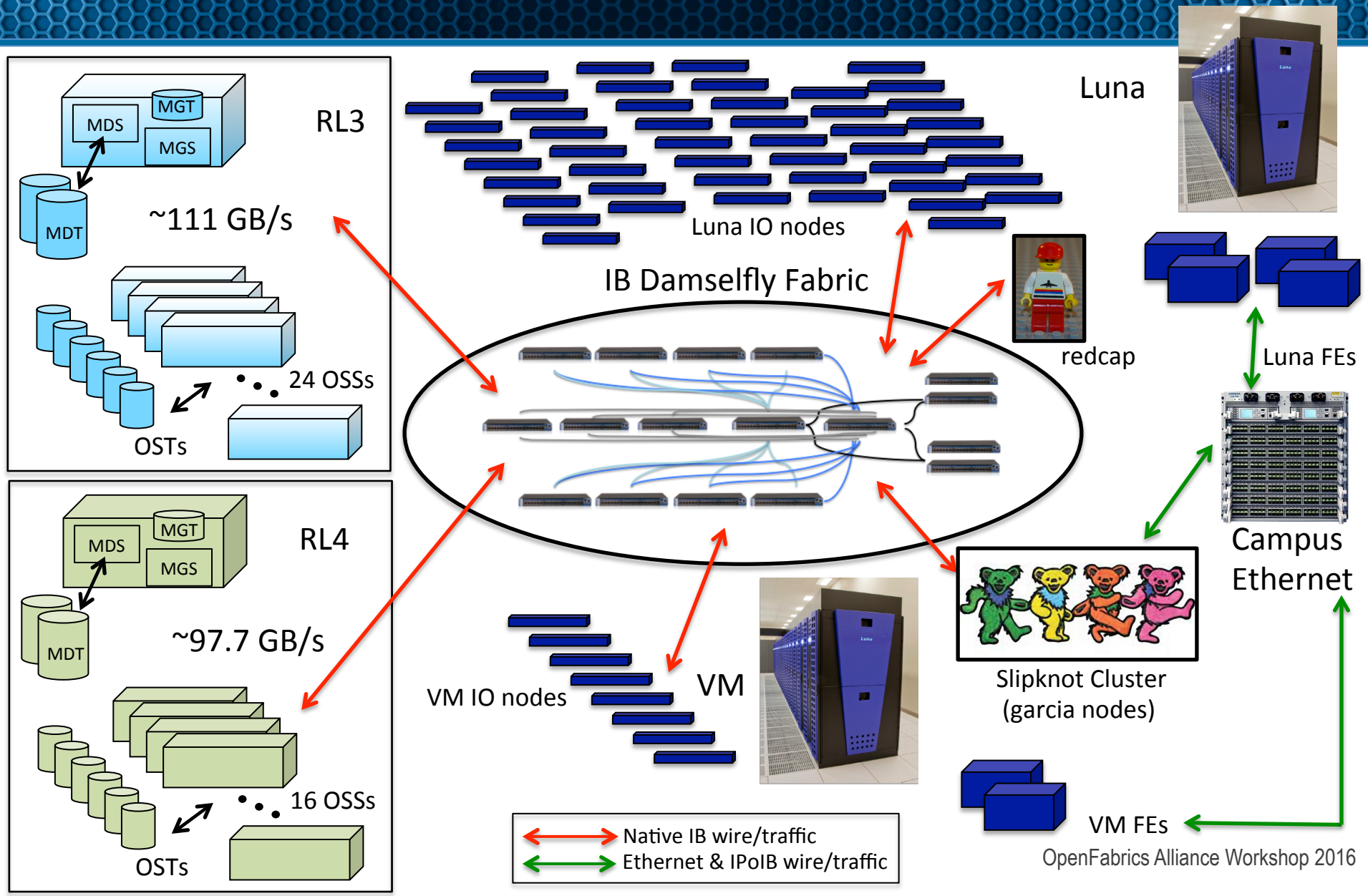
EDR Spanning Leafs (ESLs) - 4
Top EDR SLs: 8 uplinks, 8 downlinks, 20 hosts
Bottom EDR SLs: 8 uplinks, 28 hosts
Total: 96 hosts

(30 FTAs, 4 GW/garcia)
(22 Trinity LNet Routers, 6 CTS-1 FEs, 4 RDMS)
30 spare ports – 6/top, 24/bottom

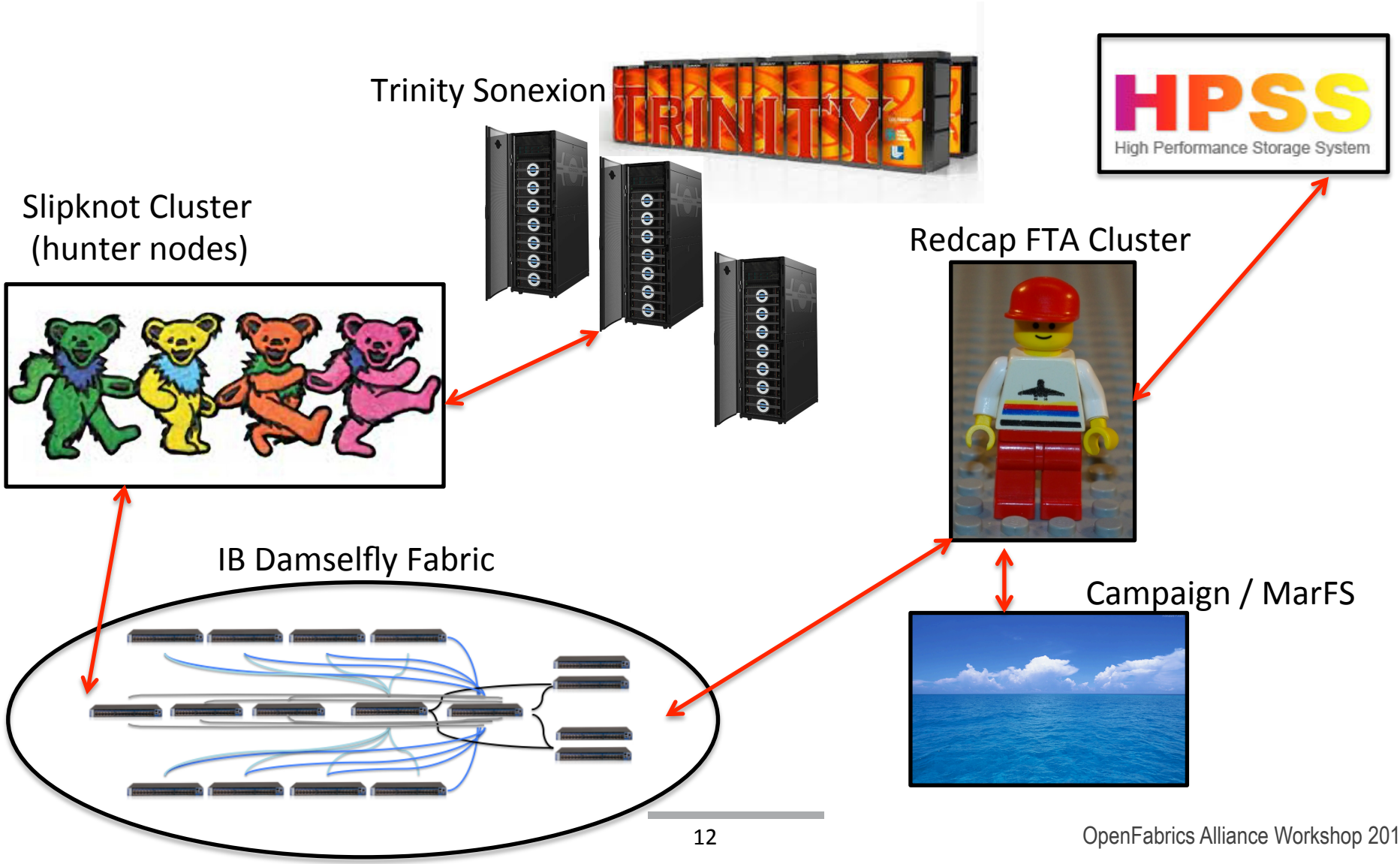
FDR Spanning Leafs (FSLs) – 3
FDR SLs: 4 uplinks, 32 hosts @
Total: 96 hosts

(70 HPSS Movers and 1 HPSS Master)
25 spare ports

Damselfly Deployment



Damselfly Deployment – Trinity View



Comparison

- Cost of EDR Chassis that could support this deployment =~ $575k
- Cost of this deployment, switches and cables =~ $290k

- Maintenance and deployment costs are higher
- Trouble shooting is more difficult
- Expansion options are painful and restrictive
- Responsibility for insuring the performance is much higher

- Fun-factor and/or panic-factor is much higher ☺

OPENFABRICS ALLIANCE

12th ANNUAL WORKSHOP 2016

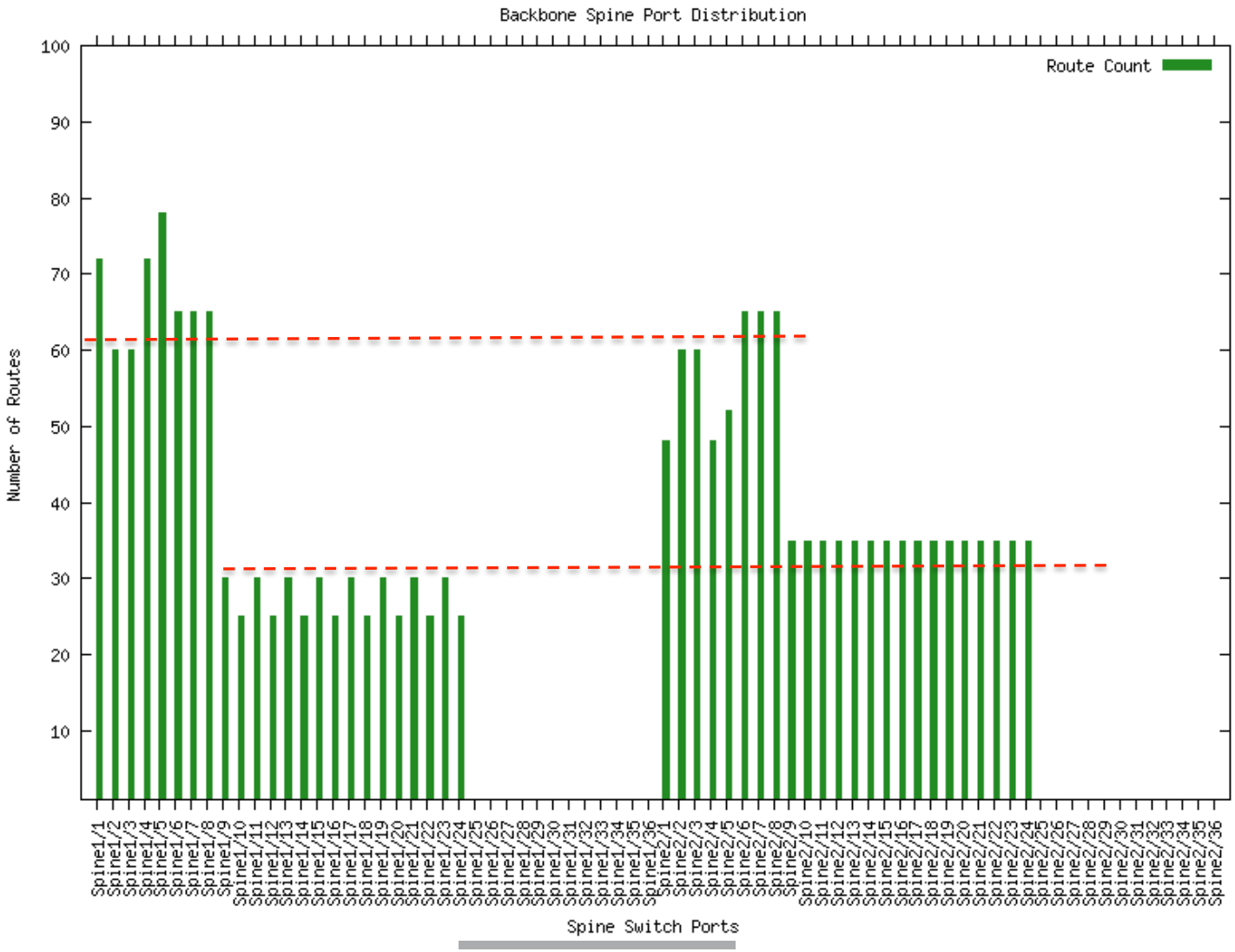
THANK YOU

Susan Coulter

Los Alamos National Laboratory

Los Alamos NATIONAL LABORATORY EST. 1943

Routing


Backbone Spine Port Distribution
Route Count
Number of Routes
100
90
80
70
60
50
40
30
20
10
Spine1/1
Spine1/2
Spine1/3
Spine1/4
Spine1/5
Spine1/6
Spine1/7
Spine1/8
Spine1/9
Spine1/10
Spine1/11
Spine1/12
Spine1/13
Spine1/14
Spine1/15
Spine1/16
Spine1/17
Spine1/18
Spine1/19
Spine1/20
Spine1/21
Spine1/22
Spine1/23
Spine1/24
Spine1/25
Spine1/26
Spine1/27
Spine1/28
Spine1/29
Spine1/30
Spine1/31
Spine1/32
Spine1/33
Spine1/34
Spine1/35
Spine1/36
Spine2/1
Spine2/2
Spine2/3
Spine2/4
Spine2/5
Spine2/6
Spine2/7
Spine2/8
Spine2/9
Spine2/10
Spine2/11
Spine2/12
Spine2/13
Spine2/14
Spine2/15
Spine2/16
Spine2/17
Spine2/18
Spine2/19
Spine2/20
Spine2/21
Spine2/22
Spine2/23
Spine2/24
Spine2/25
Spine2/26
Spine2/27
Spine2/28
Spine2/29
Spine2/30
Spine2/31
Spine2/32
Spine2/33
Spine2/34
Spine2/35
Spine2/36
Spine Switch Ports