"There and back again" The Battle of Lustre at LANL

Susan Coulter, Kyle Lamb, Mike Mason

April13-15, 2015

Overview

- Background
- Current Lustre status at LANL
- Future developments at LANL
- Fine Grain Routing FGR at LANL

Background of LANL and Lustre

- LANL was primarily Panasas
- Lustre came about with Cielo
- Lustre taking presence as FS of choice for future systems
- A few changes in order to adjust to Lustre vs. Panasas
  - Purging
  - User load balancing
  - Monitoring

Cielo Lustre Deployment

- Lustre 1.8 *if it ain't broke don't fix it
- 3 File systems (2PB, 2PB, and 4PB)
- Aggregate of 160GB/s across all 3
- Fat tree topology for IB
- Only connected to Cielo

lustre®
8PB

HPSS
High Performance Storage System

Turquoise Lustre File Systems

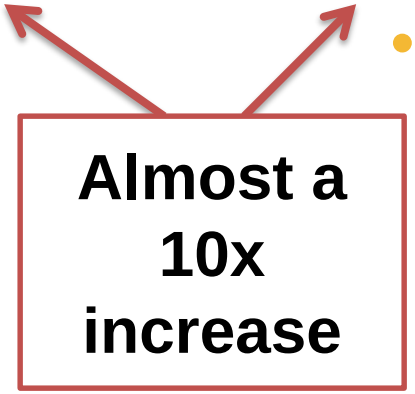
- L1
  - DDN SF12K system
  - Lustre Version 2.5.19
  - DDN OS and Stack (will likely change to TOSS early next year)
  - 3PB in aggregate
  - 35GB/s with direct IB connectivity

Turquoise Lustre File Systems

- L2
  - DDN SF12K
  - TOSS OS
  - Lustre Version 2.5.3
  - ZFS on OSTs LDISKFS on MDT
  - 1PB aggregate
  - Small deployment that we plan to increase over time
  - Current compression ratio of 1.5

The Future of Lustre at LANL

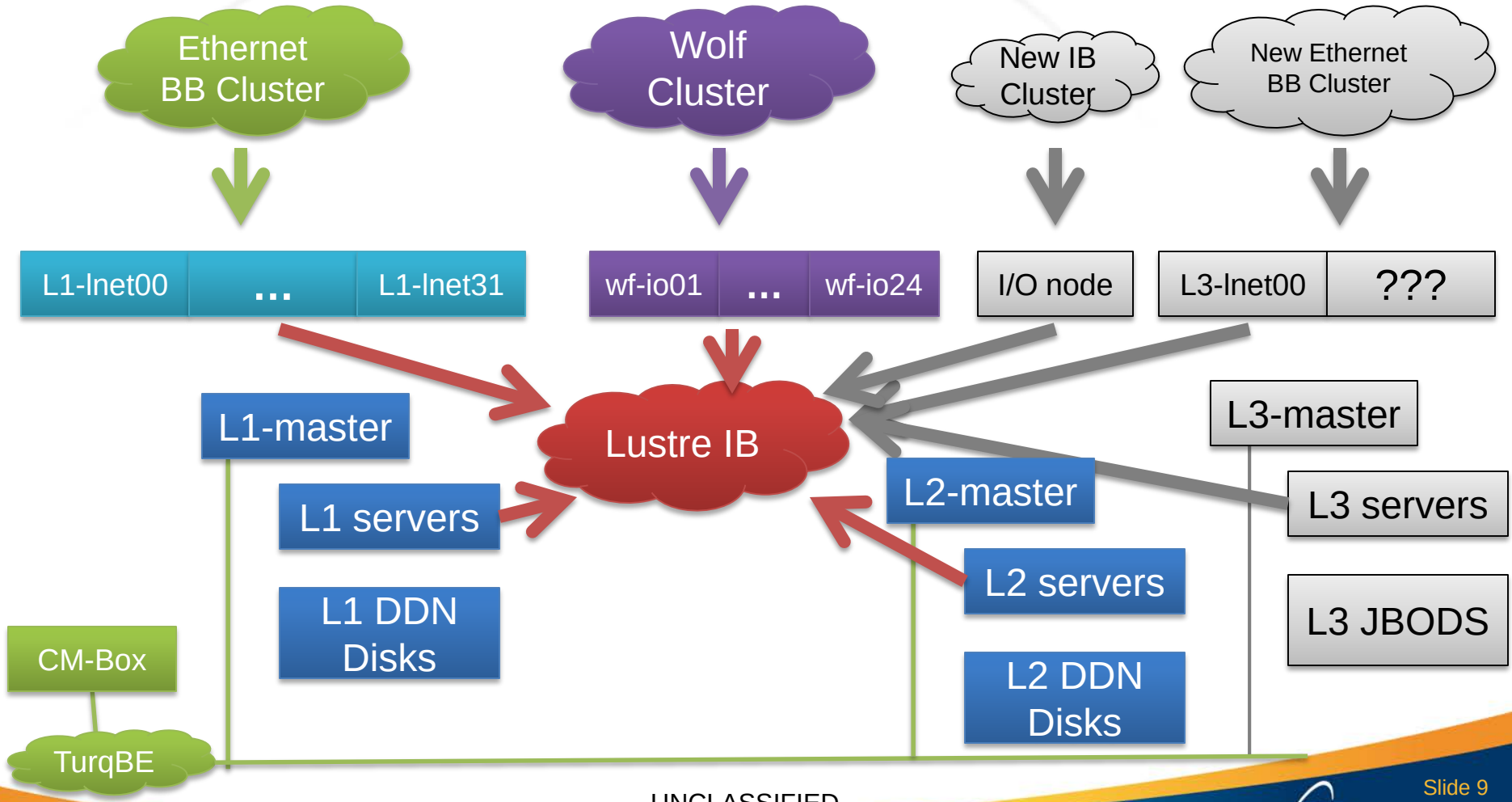
- Open
  - Current
    - Lustre 4.3 PB
    - Other 2 PB
  - End of this year
    - Lustre 15.1 PB
    - Other 2 PB
- Secure
  - Current
    - Lustre 8.3 PB
    - Other 3.7 PB
  - End of this year
    - Lustre 106.3 PB
    - Other 3.7 PB



LANL Open Infrastructure

- Current Open Infrastructure
  - L1 [3.5PB]
  - L2 [836TB]
    - One rack
    - At least double by FY16
  - /scratch (Panasas) [740TB]
  - /scratch3 (Panasas) [1.1PB]
    - Removal by early 2016
- L3 RFP
  - Minimum 5 PB (likely ~10PB)
  - 80-100GB/s @ 70% capacity
    - Sized for our FY16 Open System
  - Few Mandatory Requirement more Target Requirements
    - Allows us to see what vendors have to offer
    - Give us flexibility with our purchasing department
  - Target date Dec 2015

Open Lustre Infrastructure



LANL Secure Infrastructure

- Current Red Infrastructure
  - Cielo
    - /lscratch2,3,4 [2.1, 4.1, 2.1 PB]
  - /scratch6 (Panasas) [410TB]
  - /scratch8 (Panasas) [1.7PB]
  - /scratch9 (Panasas) [1.6PB]
    - Removal by early 2016

- L3 RFP
  - Piggyback
  - 2 new Red systems
  - Consolidate FS types
  - Target date Dec 2015
    - Build (test/debug) 3 FSes in 5 months
    - 3 admins + 4 others

ACES Supercomputer: Trinity

- Partial HPC system and compete file systems
  - Summer 2015
  - 2 PB memory
  - Burst buffer
    - 3.7 PB @ 3.3TB/s
  - LNET router 222
- Cray Hardware administration
- LANL/SNL Software administration

Trinity File System: Sonexion

- Two Cray Sonexion 2000 File Systems
  - 39PB each (78PB total usable)
  - 1.33 TB/s (80% memory in 20 minutes)
  - 19 Racks per file system (38 total)
  - 108 SSU per file system (216 total)
    - 216 OSSs with 1 OST each
  - 6 TB drives; GridRAID 41 drives
  - Lustre 2.7
  - DNE phase 2
    - 5 MDS

Trinity Test Environment: Trinitite and Gadget

- Application Regression Test: Trinitite
  - 2 racks, 200 nodes
  - 38TB Burst Buffer @ 34GB/s
  - 3 LNET Routers
  - 720TB Sonexion 2000 @ 15GB/s
- System Development Test: Gadget
  - 1 rack, 40 nodes
  - 13TB Burst Buffer @ 11GB/s
  - 2 LNET Routers
  - 360TB Sonexion 2000 @ 7.5GB/s
- Posters
  - Early Performance and Scaling of Sonexion 2000
  - Bottom-up Performance Estimation for a Cray Sonexion 2000

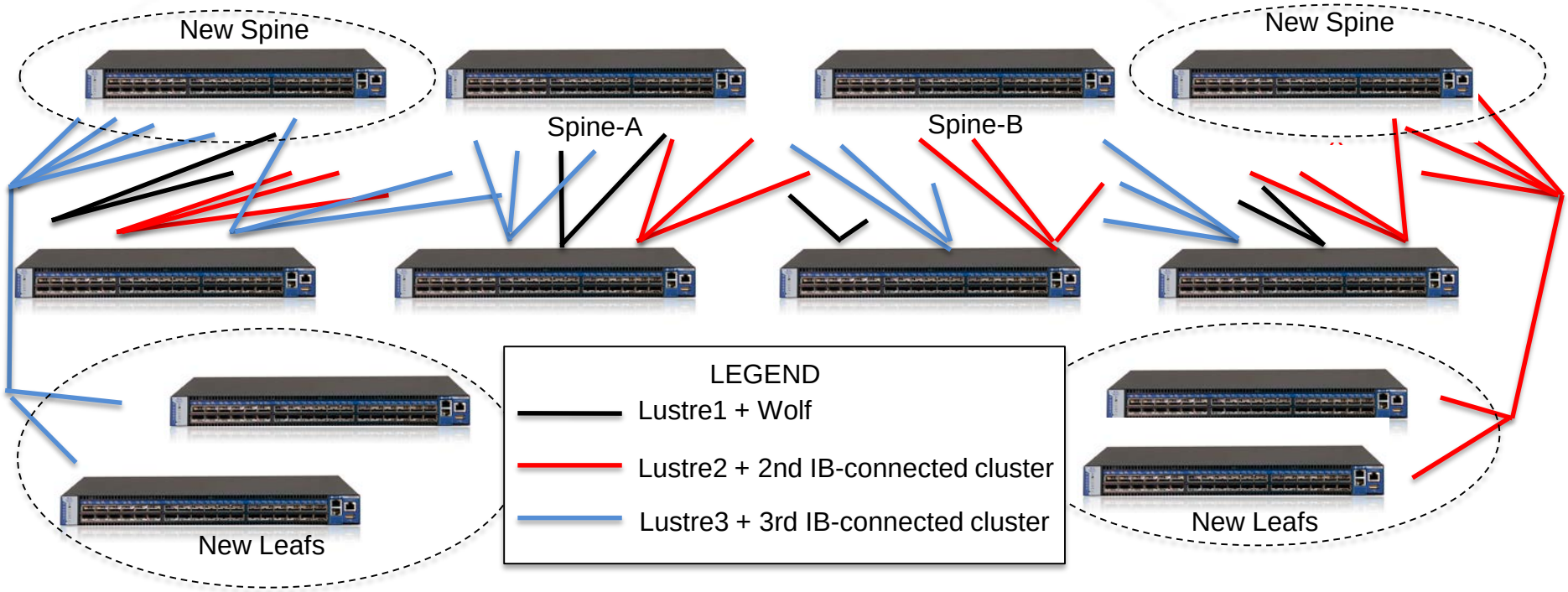
FGR – Decision

- Long term configuration
  - 3 Lustre file systems
  - 3 IB-connected clusters
- Close impedance match between hosts and uplinks for non-FGR option
- Hardware requirements
- Ease of expansion
- Support model

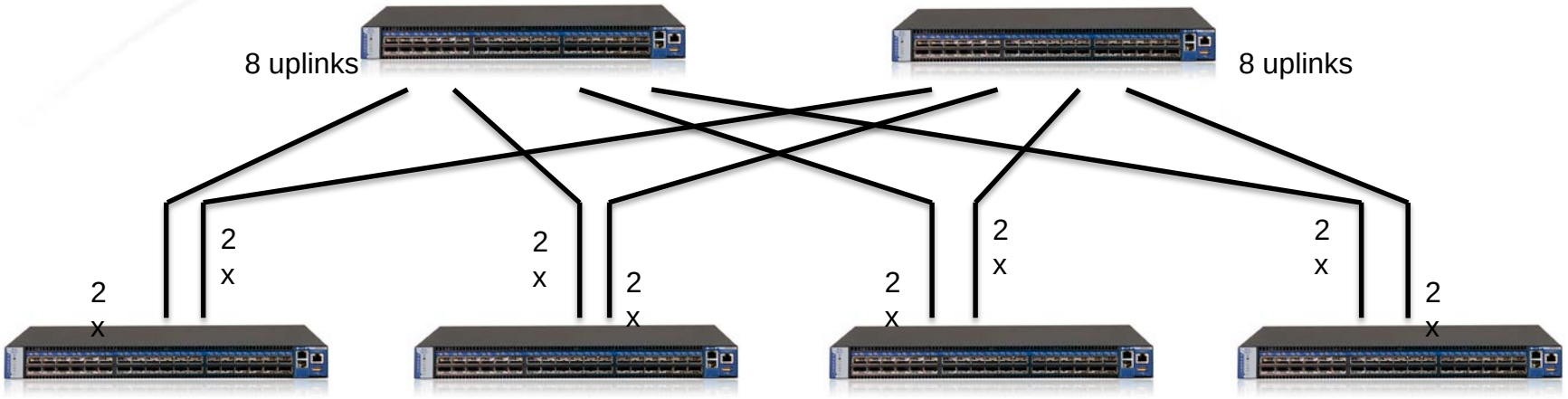
References: I/O Congestion Avoidance via Routing and Object Placement / D. A. Dillow, G. M. Shipman, S. Oral, and Z. Zhang (CUG 2011)

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

FGR – Decision / Configuration without FGR

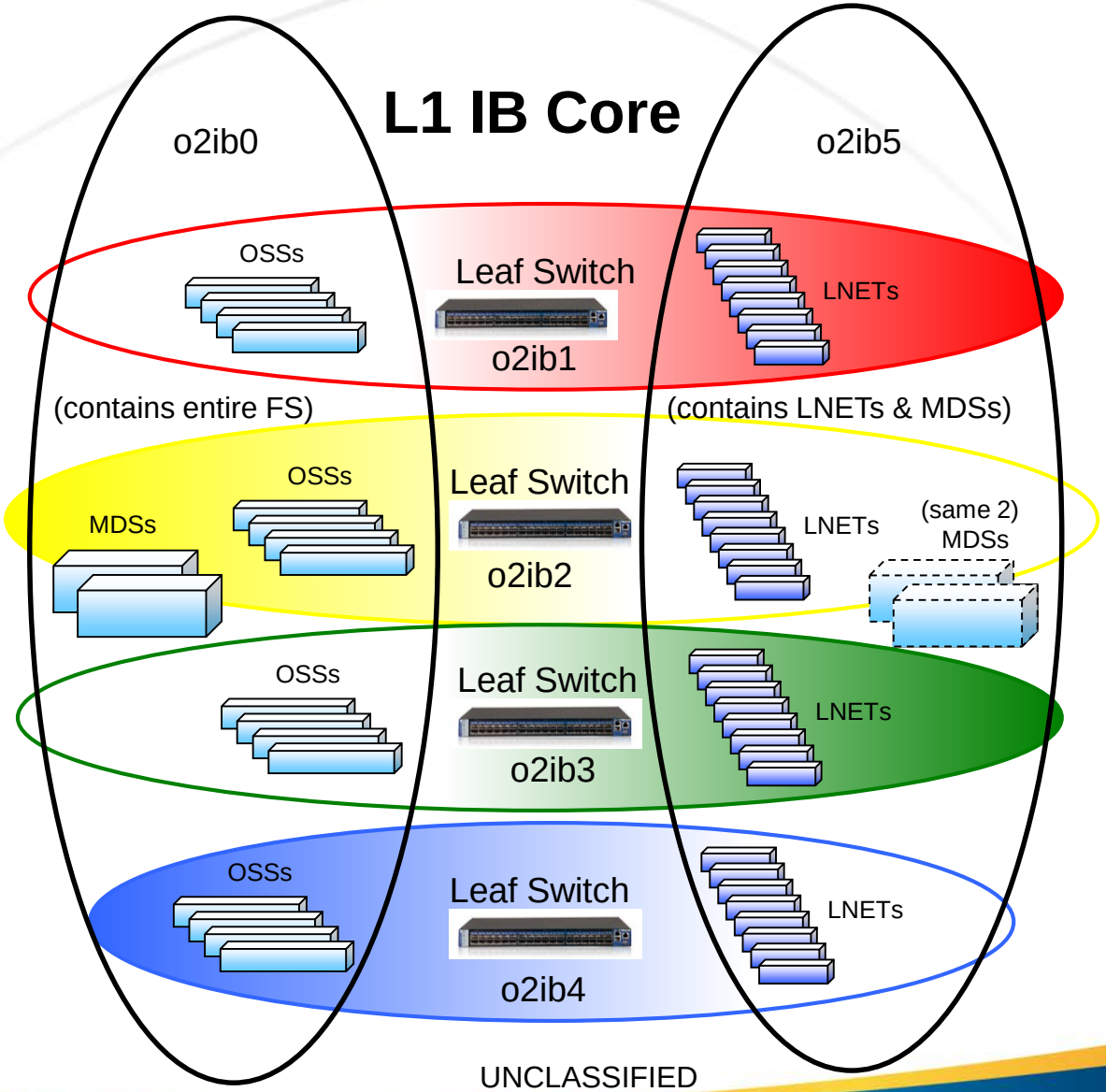


FGR – Decision / Configuration with FGR

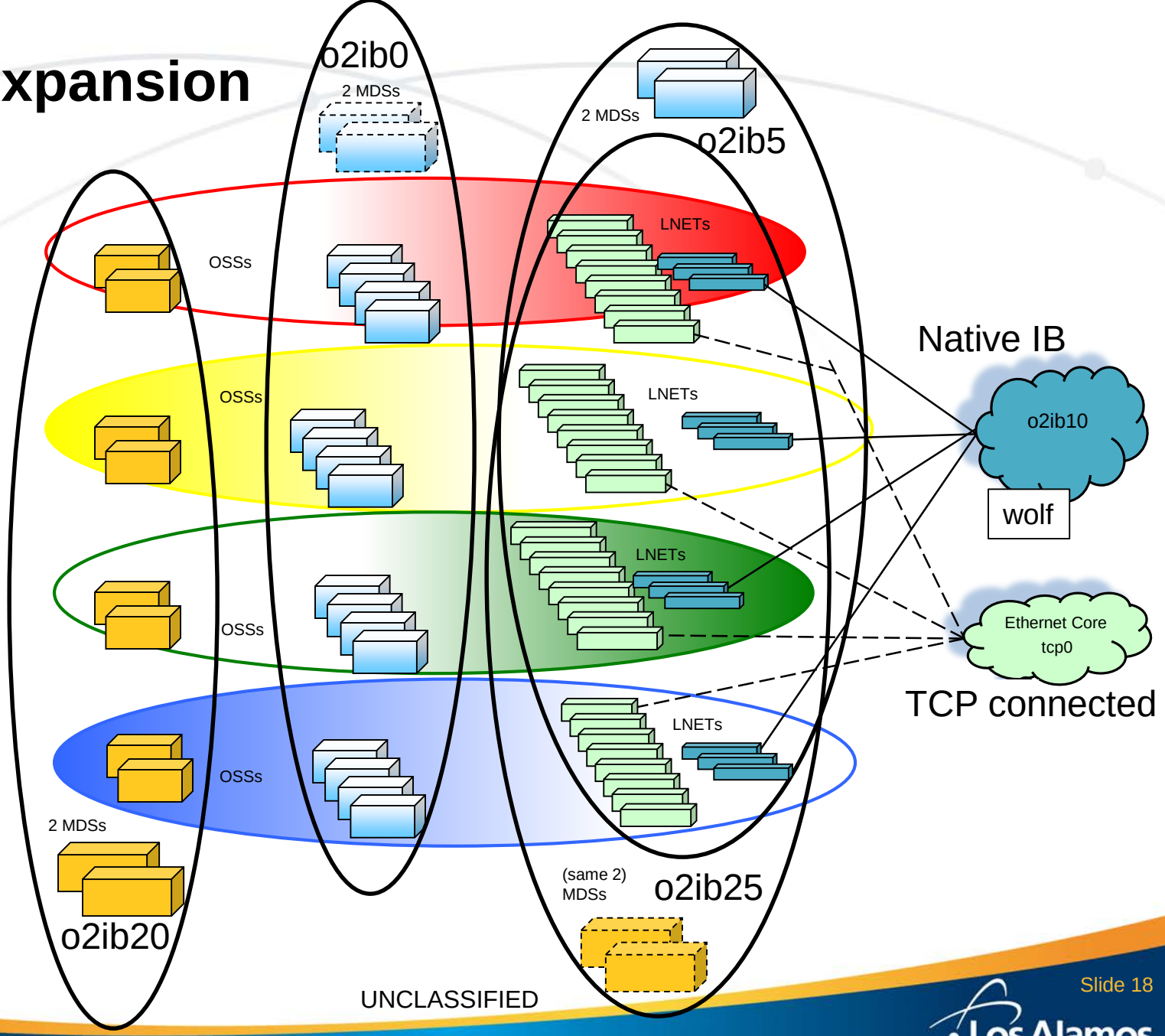


- PROs
  - Simpler IB fabric
  - Much easier to expand
  - Significantly lower hardware costs
  - Translatable to secure implementation
- CONs
  - FGR relatively new
  - LANL Lustre experience minimal
  - DDN support of FGR tentative
  - More complex LNET configuration

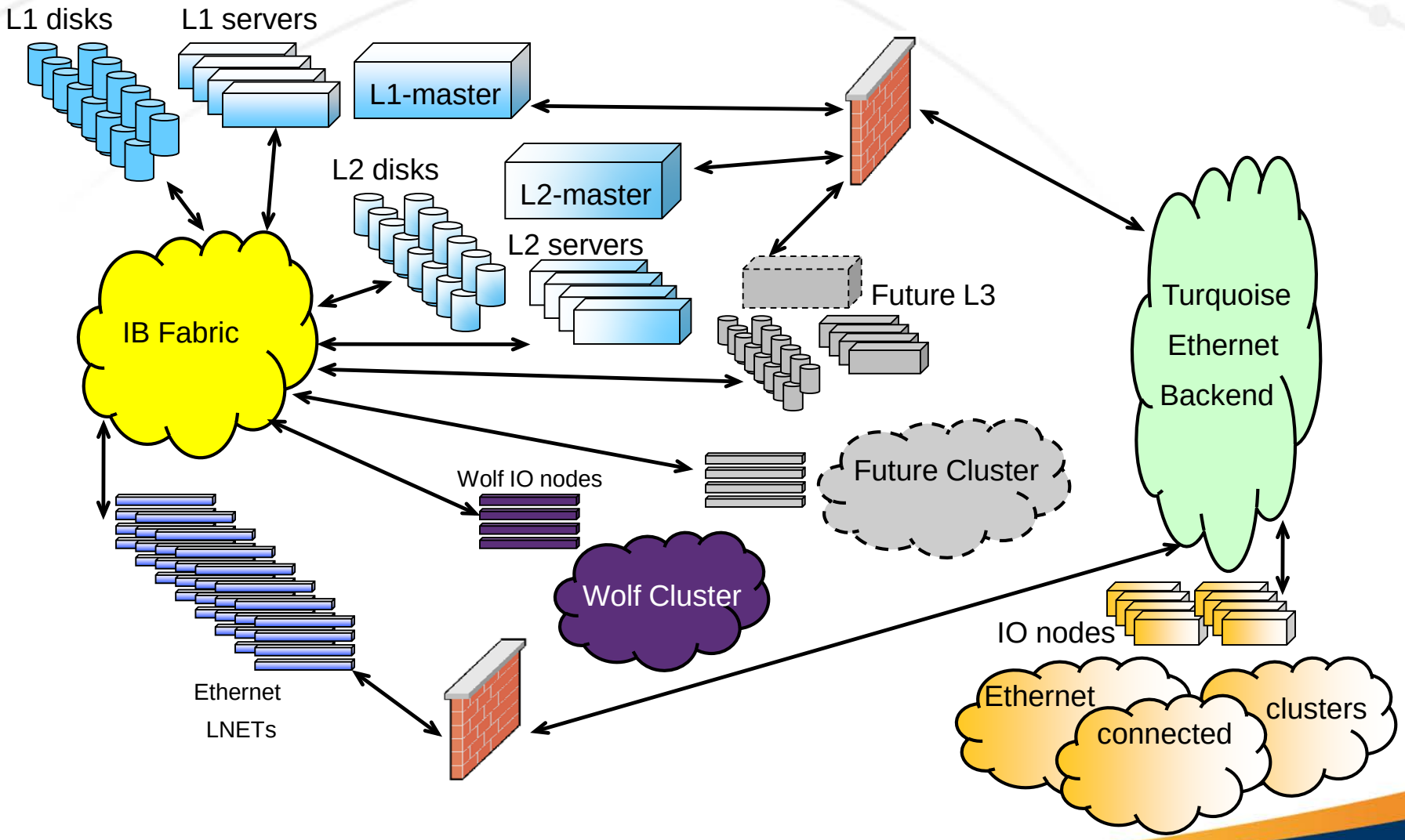
FGR – Implementation



FGR – Expansion

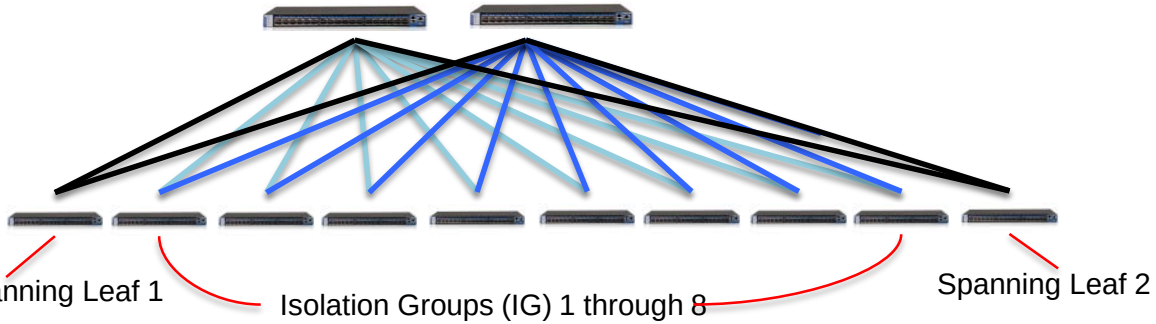


FGR - Future



FGR – Lessons

- Modprobe configuration file
  - Single common file vs custom files
  - Failover complexity
    - Limitation on number of characters
- Knowledge translation from turquoise to red
  - Use of FGR informed IB backbone Damselfly design



Thank You