OPENFABRICS ALLIANCE

# New Storage Architectures

## Replacing LNET routers with IB routers

OpenFabrics Software User Group Workshop

#OFSUserGroup

# Lustre Basics

- Lustre is a clustered file-system for supercomputing
- Architecture consists of clients and three types of servers
- Basic configuration connects servers and clients with a single network.

# LNET Routers

- Separate storage and compute IB fabrics
- Different subnet managers, different topologies, fault isolation
- Limited to LNET traffic

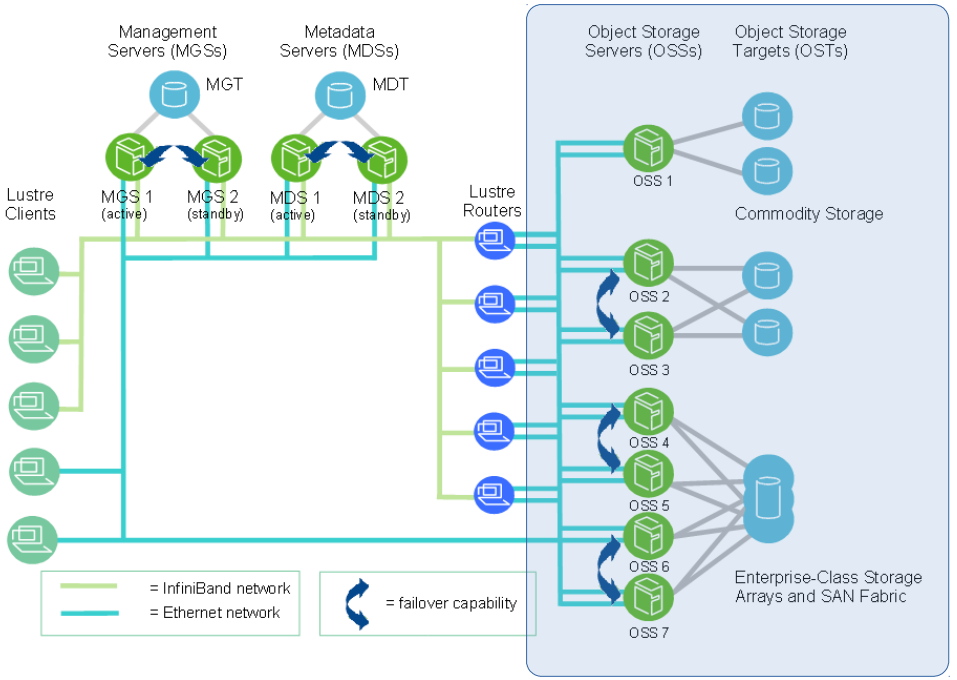



Image Credit: wiki.lustre.org

# Typical LNET Router

- Basic Compute Node
- Xeon Processor(s), lots of Memory, one or two HCAs
- Runs Linux and the LNET kernel stack
- Software involvement in packet routing

# The First Crossbow Device

**Crossbow R400-6**

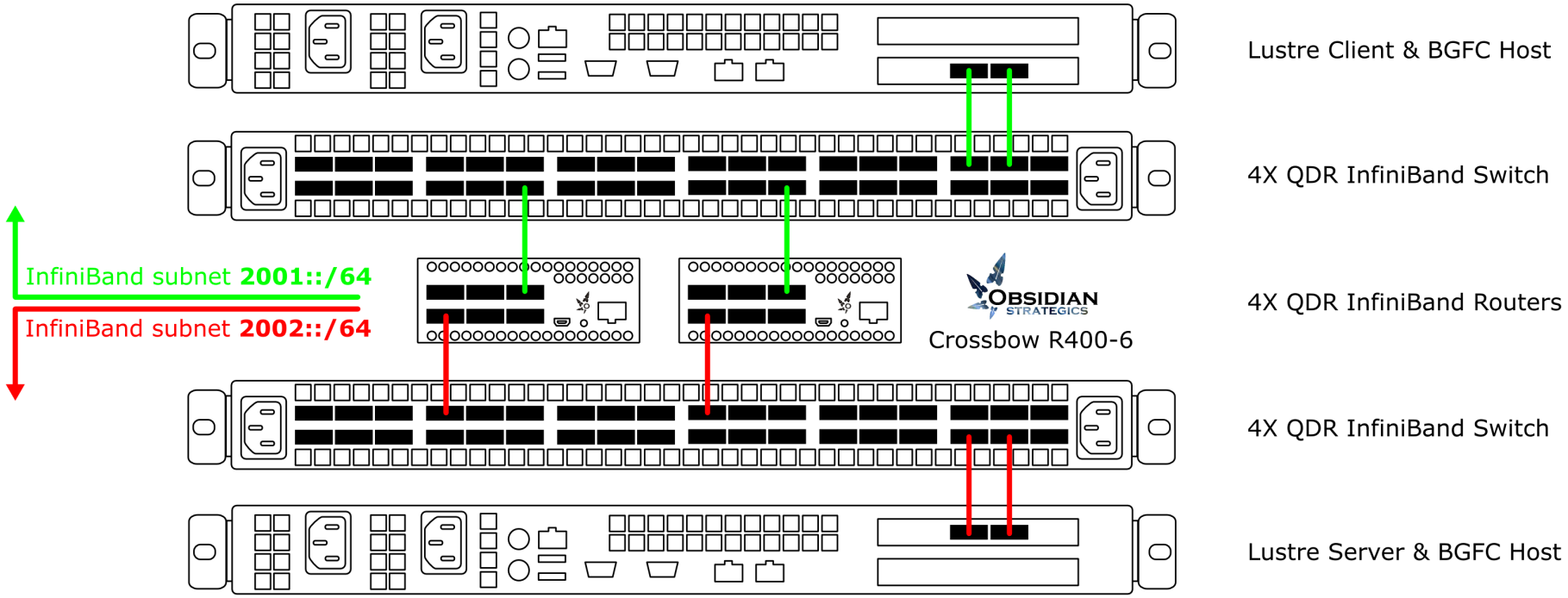
40G 6-port 4X QDR IB router

LAN or mixed LAN / WAN environments

# R400

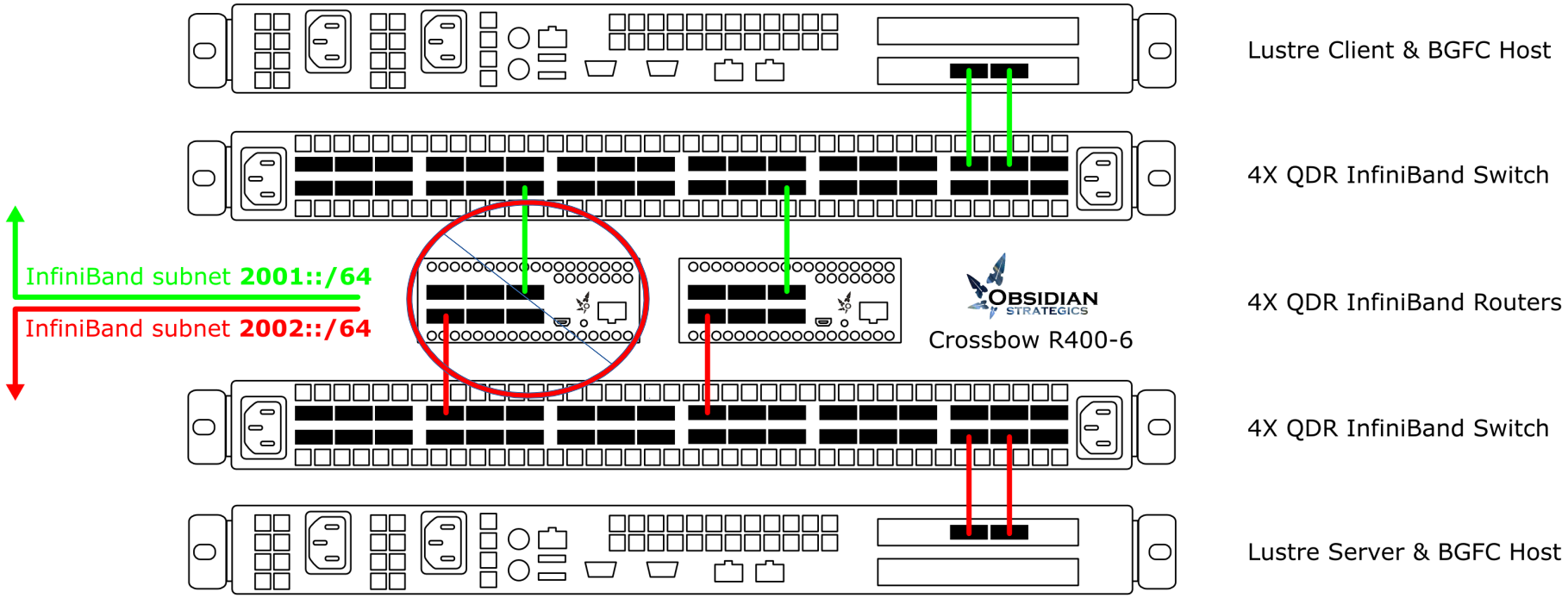
- Six port IB router
- Shared memory hardware switching architecture
- ~30 watts operating
- ~350ns port to port latency
- Wire speed
- Stripe with multipath many R400s for capacity

# Demo Configuration




- Ran for 3 days on the show floor during SC|14

# Demo Configuration



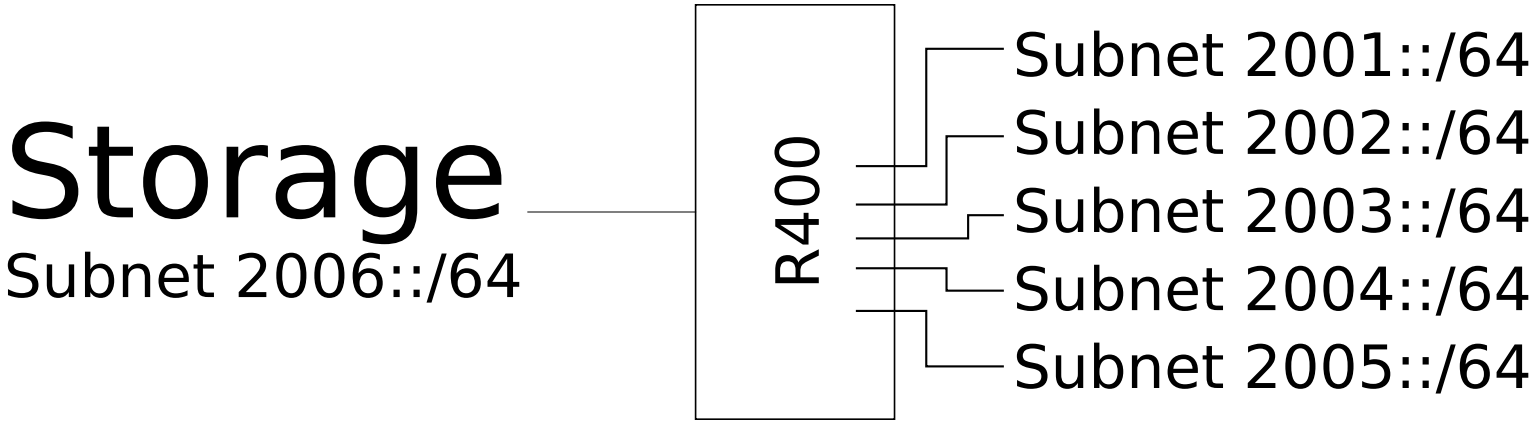

- Full bandwidth while deliberately faulting equipment

# Replacing LNET Routers

- Each Crossbow R400 replaces three LNET router nodes
- Automatic Path Migration is supported at the IB layer
- Routed IB can be natively extended over distance using Longbow

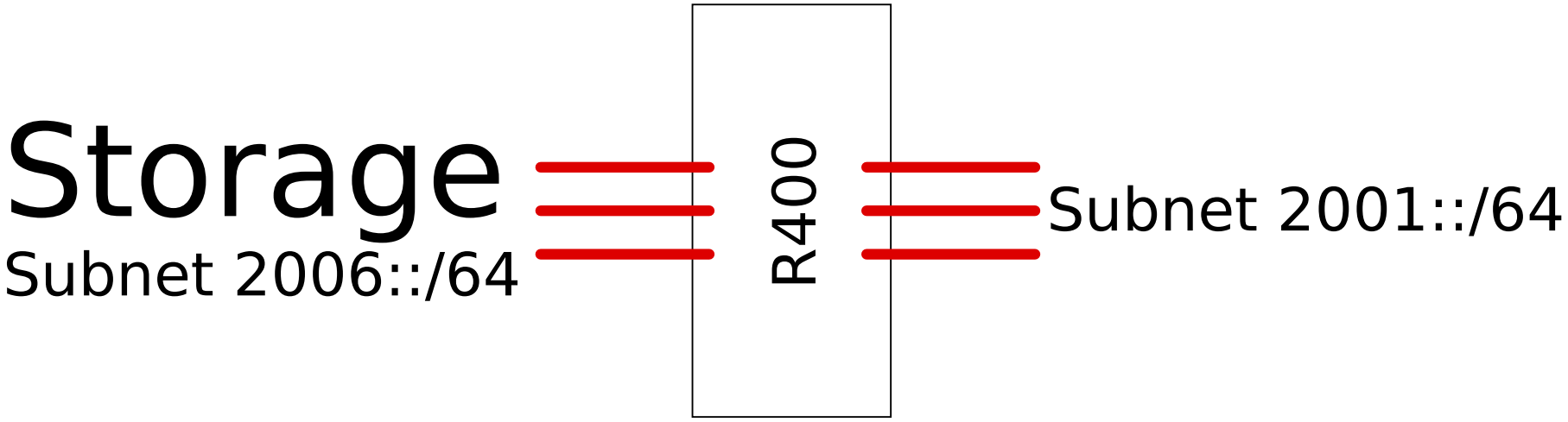
# Replacing LNET Routers

- Cross connect multiple subnets

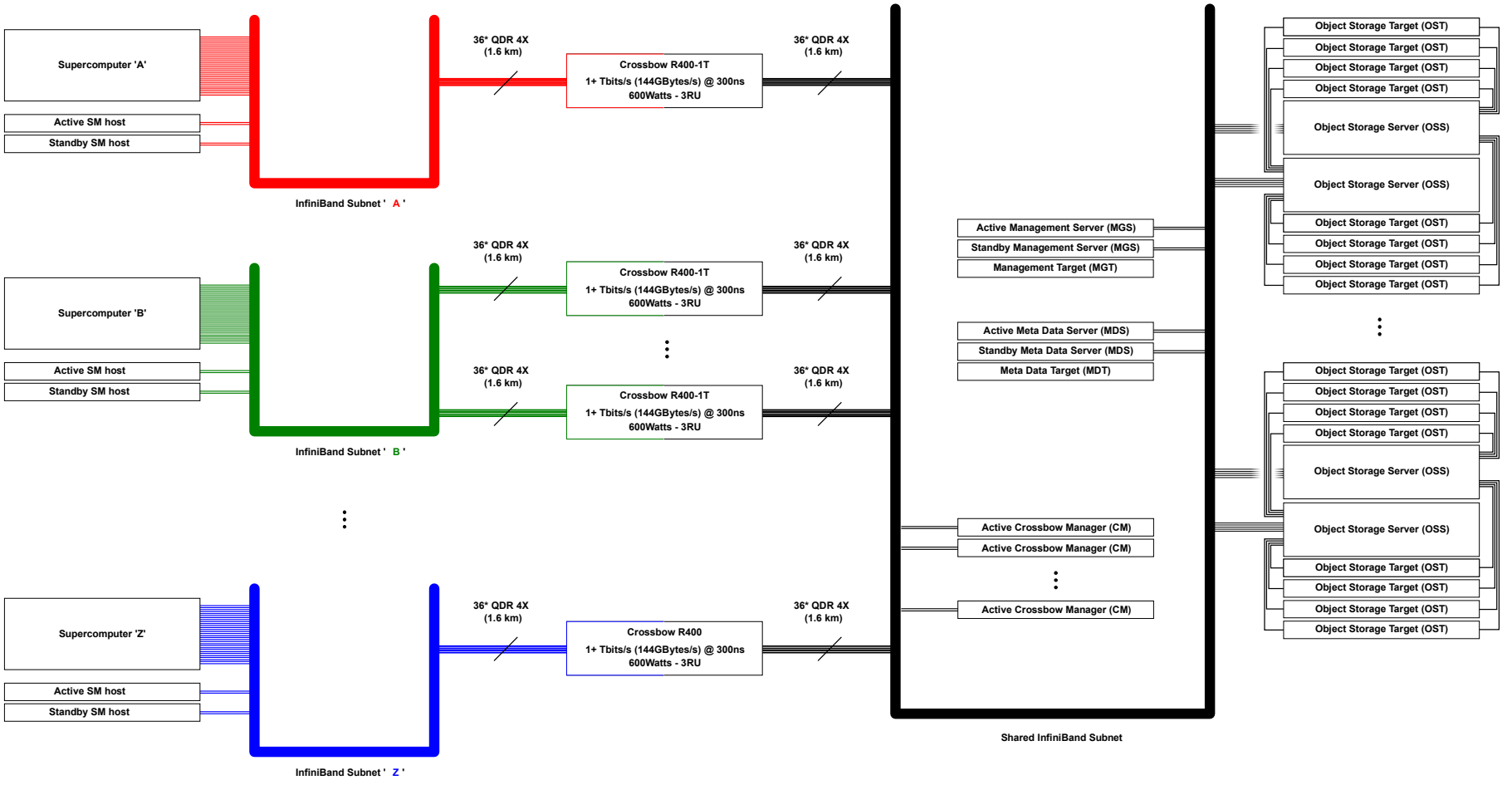


# Replacing LNET Routers

- Parallel connect two subnets for bandwidth



# Scale Up


Supercomputer 'A'
Active SM host
Standby SM host
InfiniBand Subnet ' A '
36* QDR 4X (1.6 km)
Crossbow R400-1T
1+ Tbits/s (144GBytes/s) @ 300ns
600Watts - 3RU
36* QDR 4X (1.6 km)
Supercomputer 'B'
Active SM host
Standby SM host
InfiniBand Subnet ' B '
36* QDR 4X (1.6 km)
Crossbow R400-1T
1+ Tbits/s (144GBytes/s) @ 300ns
600Watts - 3RU
36* QDR 4X (1.6 km)
Supercomputer 'Z'
Active SM host
Standby SM host
InfiniBand Subnet ' Z '
36* QDR 4X (1.6 km)
Crossbow R400
1+ Tbits/s (144GBytes/s) @ 300ns
600Watts - 3RU
36* QDR 4X (1.6 km)
Active Management Server (MGS)
Standby Management Server (MGS)
Management Target (MGT)
Active Meta Data Server (MDS)
Standby Meta Data Server (MDS)
Meta Data Target (MDT)
Active Crossbow Manager (CM)
Active Crossbow Manager (CM)
Active Crossbow Manager (CM)
Shared InfiniBand Subnet
Object Storage Target (OST)
Object Storage Target (OST)
Object Storage Target (OST)
Object Storage Target (OST)
Object Storage Server (OSS)
Object Storage Server (OSS)
Object Storage Target (OST)
Object Storage Target (OST)
Object Storage Target (OST)
Object Storage Target (OST)
Object Storage Target (OST)
Object Storage Target (OST)
Object Storage Target (OST)
Object Storage Target (OST)
Object Storage Server (OSS)
Object Storage Server (OSS)
Object Storage Target (OST)
Object Storage Target (OST)
Object Storage Target (OST)
Object Storage Target (OST)

# BGFC Clustered SM

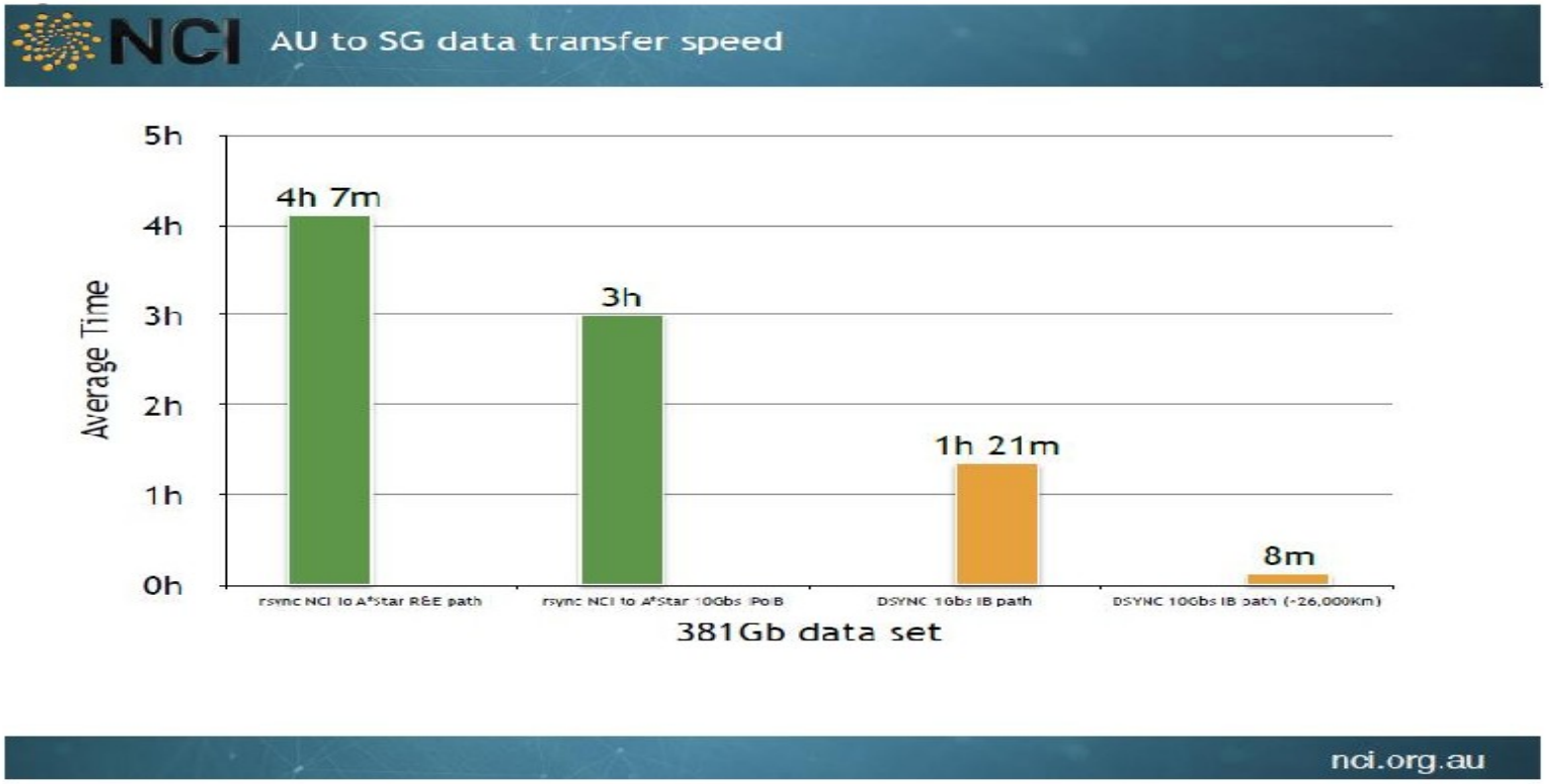
- New multi-subnet, router enabled SM
- Clustered per subnet, and clustered globally
- Sophisticated multi-subnet routing engine allows complex multi-subnet topologies
- Provably safe loss-less cross subnet routing
- Full support for multipathing
- Cross subnet path record queries, with multipathing and Alternate Path Migration

# File Copy over WAN

- Extend clustered file system data beyond a Campus
- Obsidian's dsync+ tool gives high performance file copy over RDMA
- Supports Longbow and Crossbow devices
- Very tolerant to latency

# DSYNC+

- APAN results: 300ms RTT, 26000 Km.



# Thank You

OPENFABRICS ALLIANCE

OpenFabrics Software
User Group Workshop