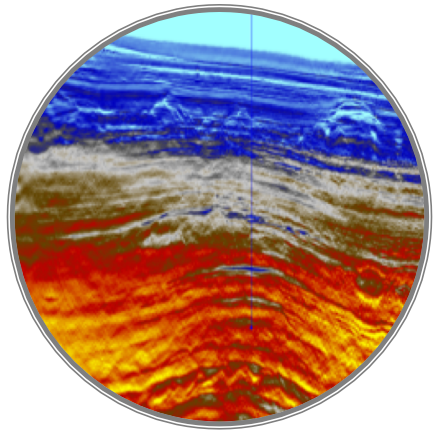
ExxonMobil

May 2019

Long distance Lustre Communication

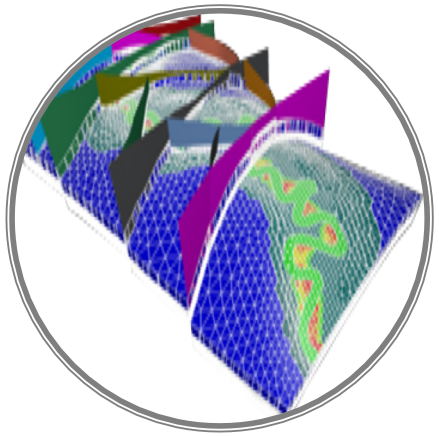
RAJ GAUTAM

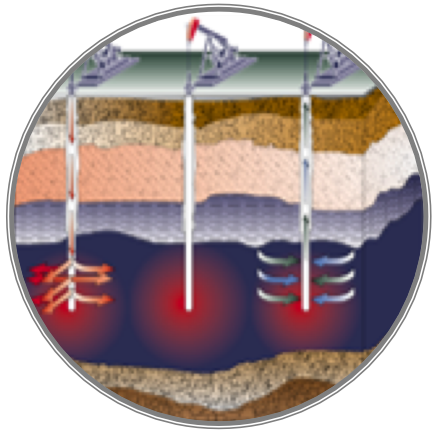
HPC is used across ExxonMobil



Geophysics

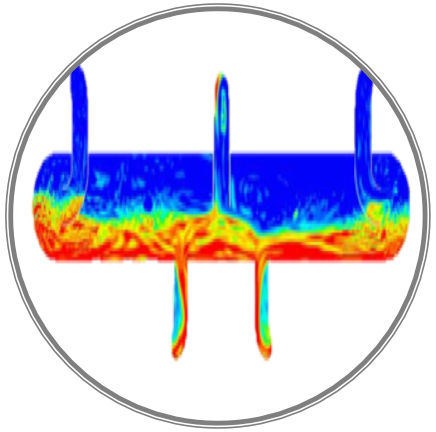
Reservoir

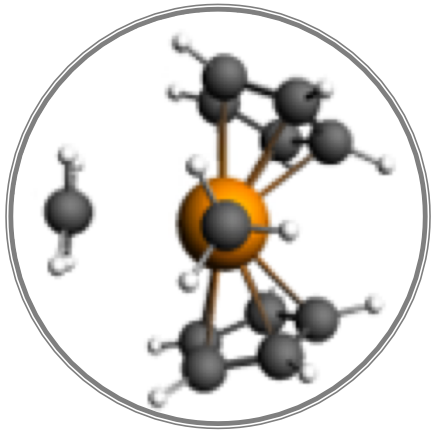




Drilling and Subsurface

Facilities





Molecular and Quantum Modeling

Capability, cycle time, and cost enable technology progression and business value

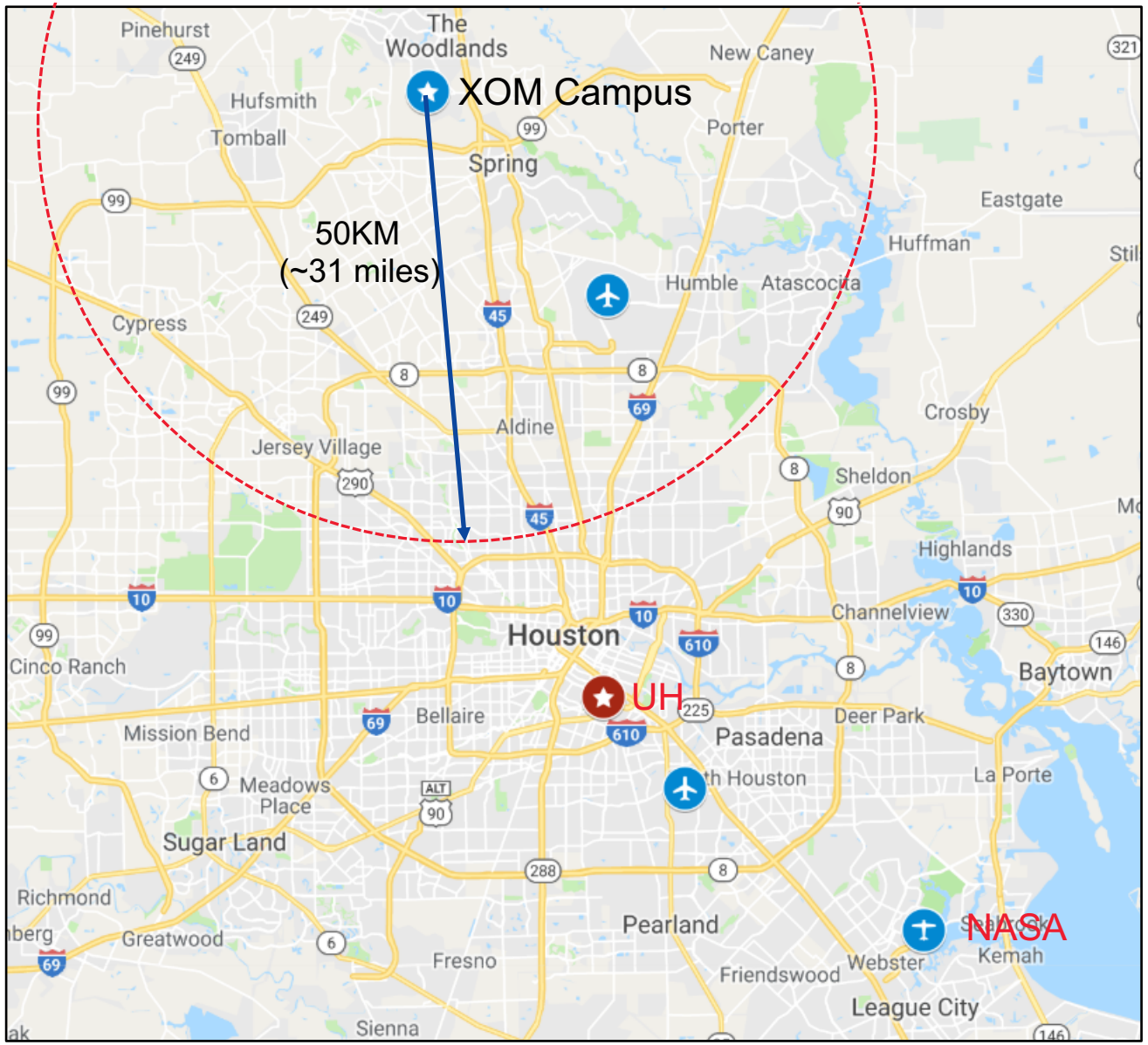
ExxonMobil

Lustre in ExxonMobil

2008: Lustre
2009: GPFS replaced Lustre
2012: Lustre v2.1 then v2.3
2013: v2.5.1 with distributed parity RAID
2018: v2.7 with DNE phase 1
2019: v2.11.0.201 with PFL

Two Datacenters

- Offsite HPC datacenter
- 50+ Petabytes of data
- New campus datacenter at ExxonMobil Campus
- 50KM (~31Miles) apart
- How to migrate Petabytes of data from offsite datacenter to new campus datacenter across Houston?

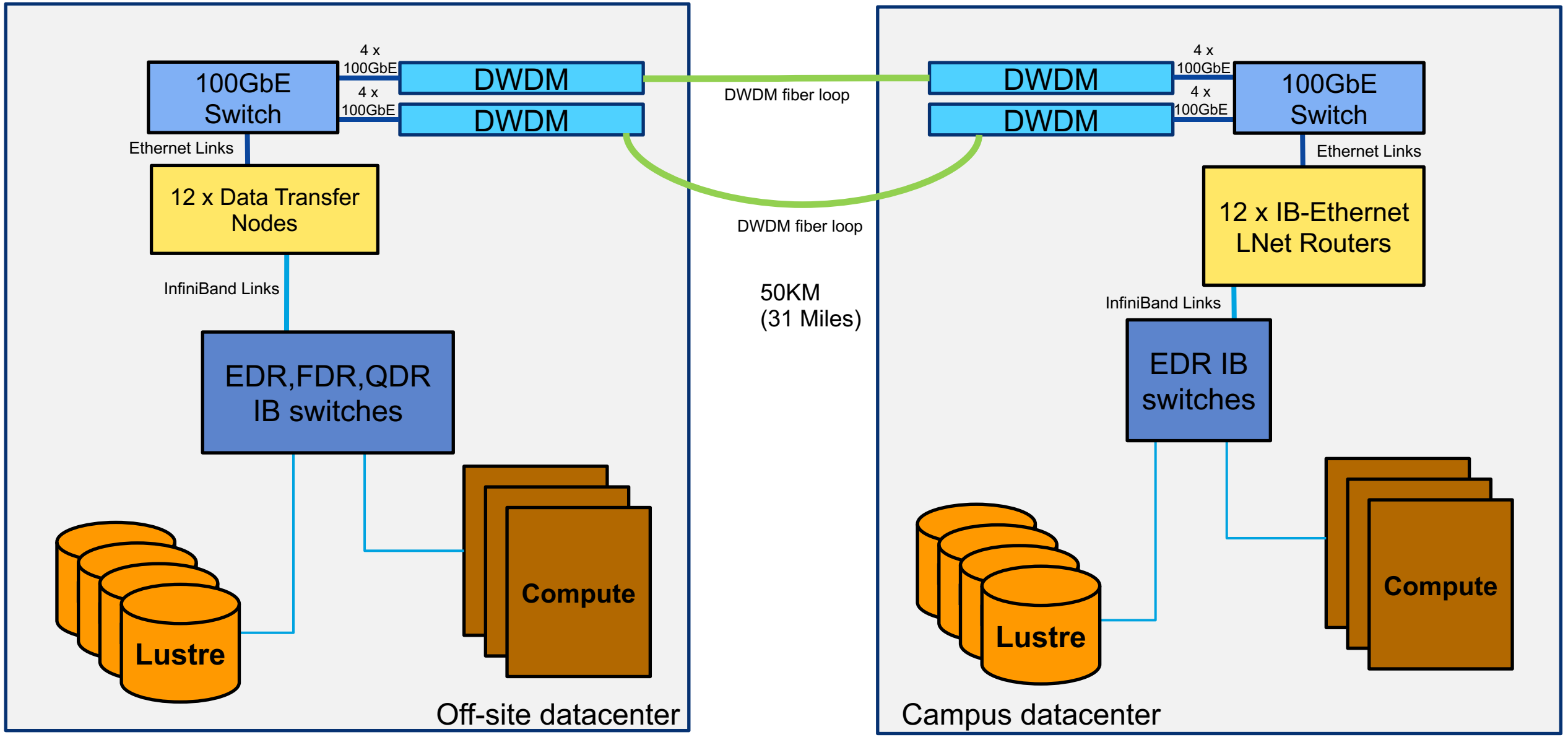


ExxonMobil

How to migrate Petabytes of data across Houston?

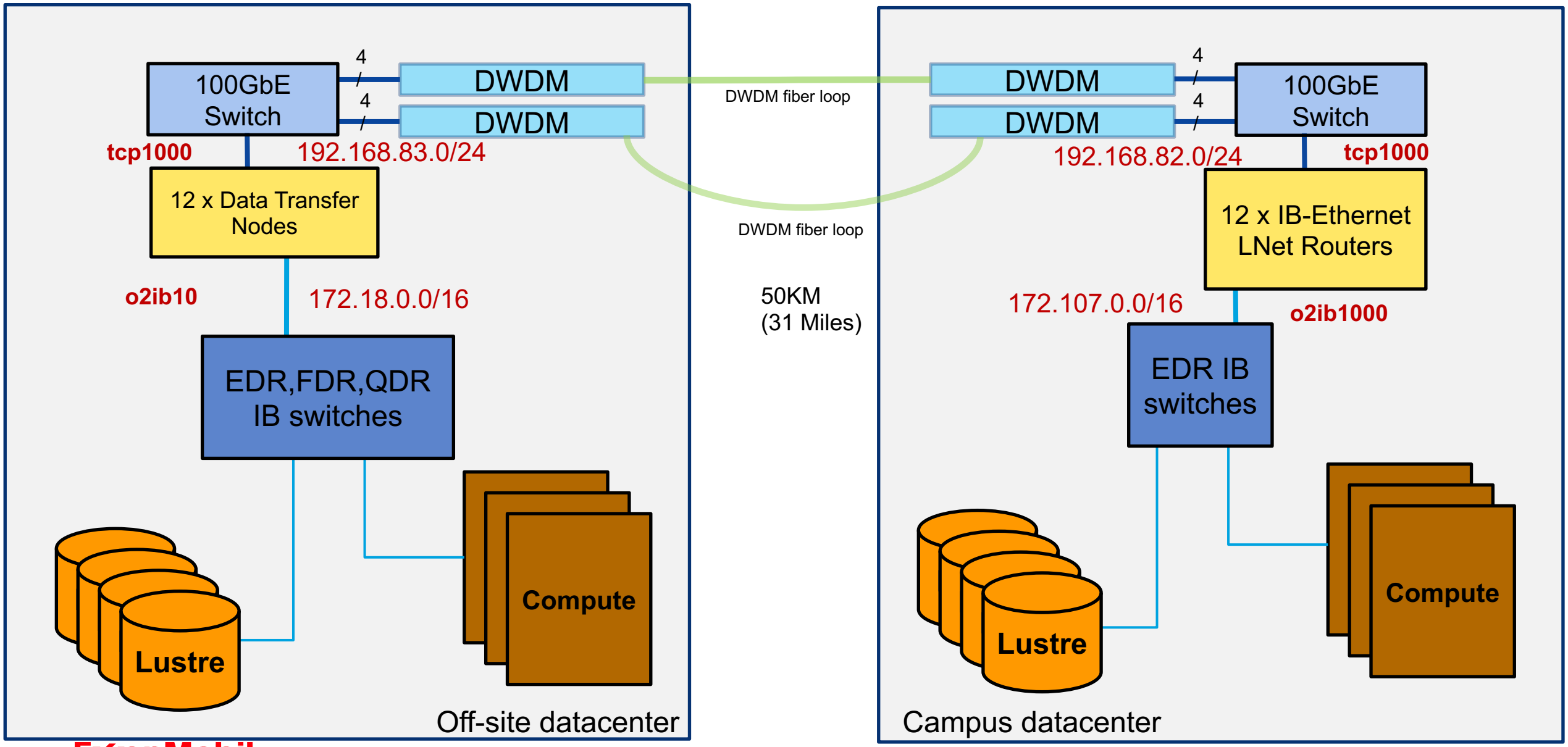
- More than 14K spinning disks
- No backups
- Infiniband network
- MetroX solution was very expensive with limited bandwidth
- Build an 8 x 100Gb/s network between data centers using DWDM technology
- Use LNet routers
- Use existing data copy tools to migrate projects between two centers

Design using LNet Routers and Data Transfer Nodes



ExxonMobil

Configuration

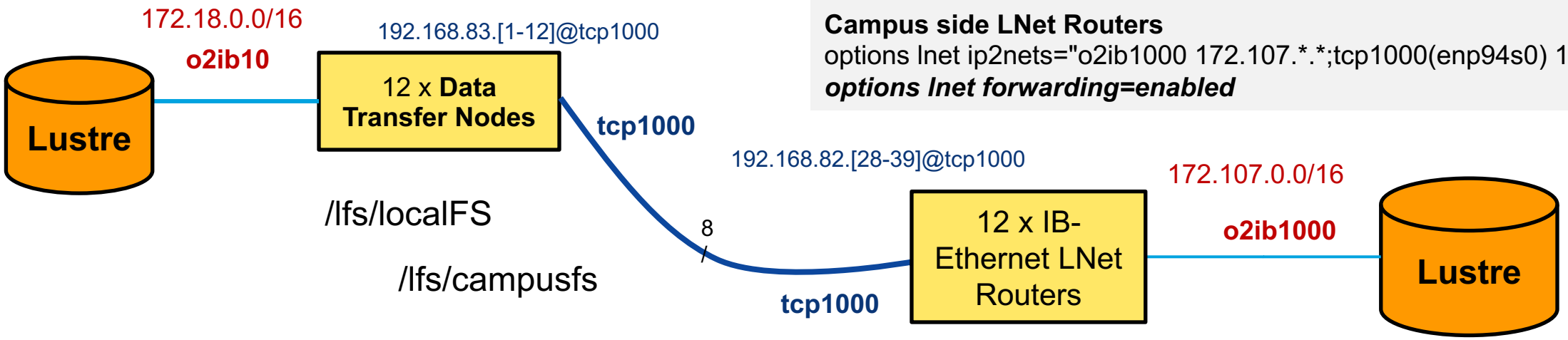


ExxonMobil

Lustre Configuration

LNet Router conf
CentOS 7.5
Lustre version 2.10.5
MOFED 4.4-2.0.7.0
100Gb Ethernet tuning from ES.net

Off-site datacenter side DTN nodes
options lnet ip2nets="tcp1000(enp94s0) 192.168.*.*; o2ib10 172.18.*.*"
options lnet routes="o2ib1000 1 192.168.82.[28-39]@tcp1000"

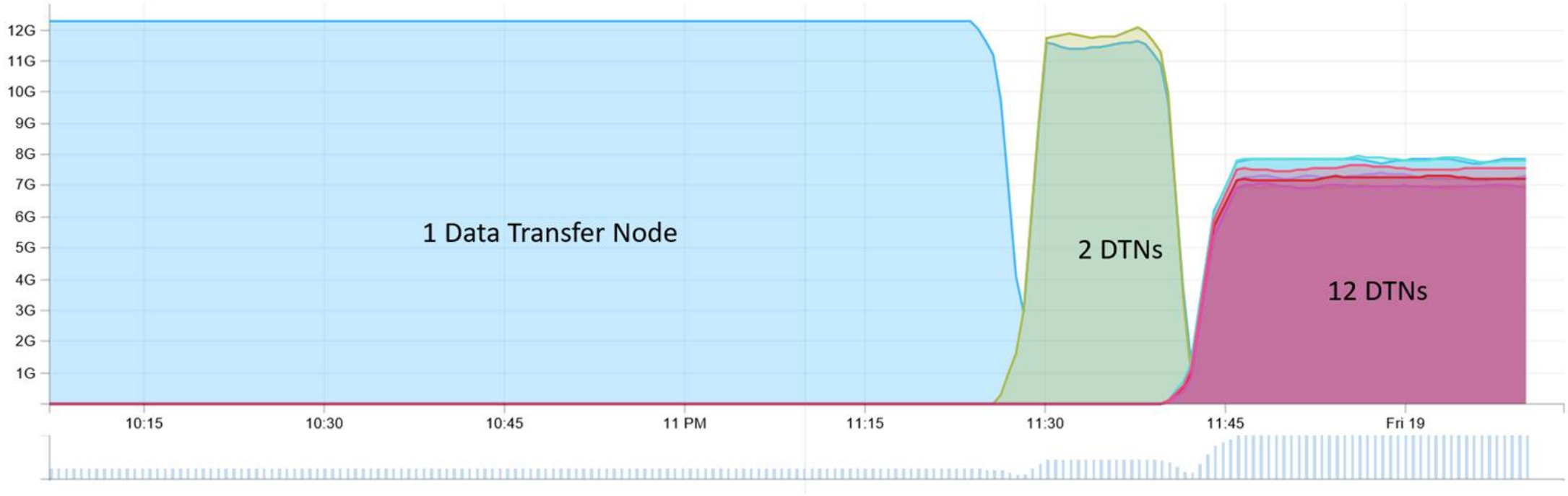


Campus side LNet Routers
options lnet ip2nets="o2ib1000 172.107.*.*;tcp1000(enp94s0) 192.168.*.*"
options lnet forwarding=enabled

Campus side Lustre FS
options lnet ip2nets="o2ib1000 172.107.28.*"
options lnet routes="tcp1000 1 172.107.0.[28-39]@o2ib1000"

ExxonMobil

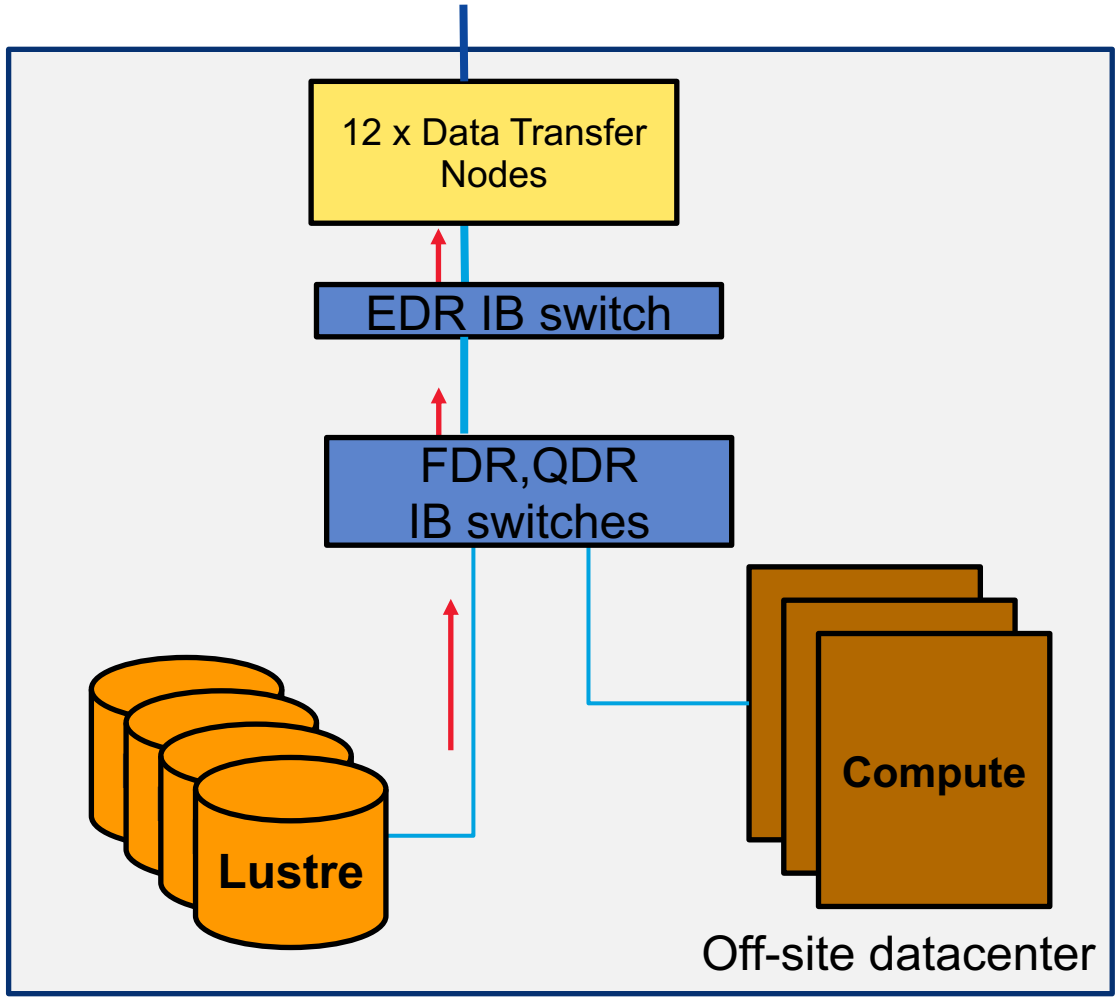
Throughput measured using LNet Self Test



- Throughput measured on Ethernet interfaces on DTN in colocation datacenter and using all 12 LNet routers in Campus.
- Since the pipe between centers is 8x100Gbps, they(12 DTNs) had to throttle down to an average of 6.8GBps using all DTNs. (6.8 x 12 = 79.2GBps)

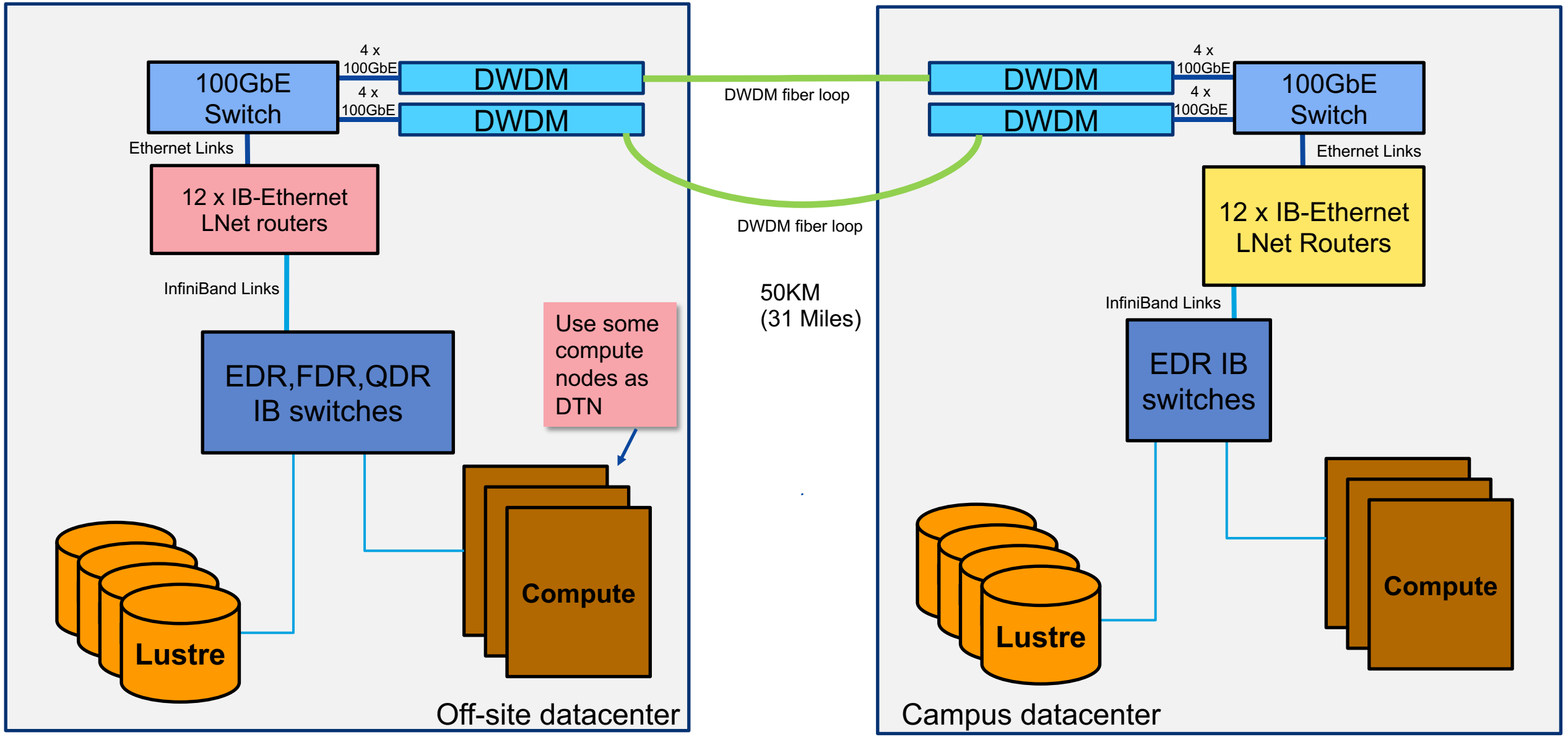
ExxonMobil

Data transfer test



- End to end data copy throughput was not as expected
- Transfer rate saturated around 30GBps
- This was due to slow read from local Lustre FS
- QDR core switch with up/down routing algorithm
- IB routes to DTN nodes' EDR switch from Lustre OSS nodes were not using all the uplinks ports
- Spreading DTN nodes switch uplinks to different ports on the core switches did not help much
- Use existing IB compute nodes to do the transfer
- Using compute nodes, we can scale the number of nodes easily

Convert DTN nodes to LNet Routers

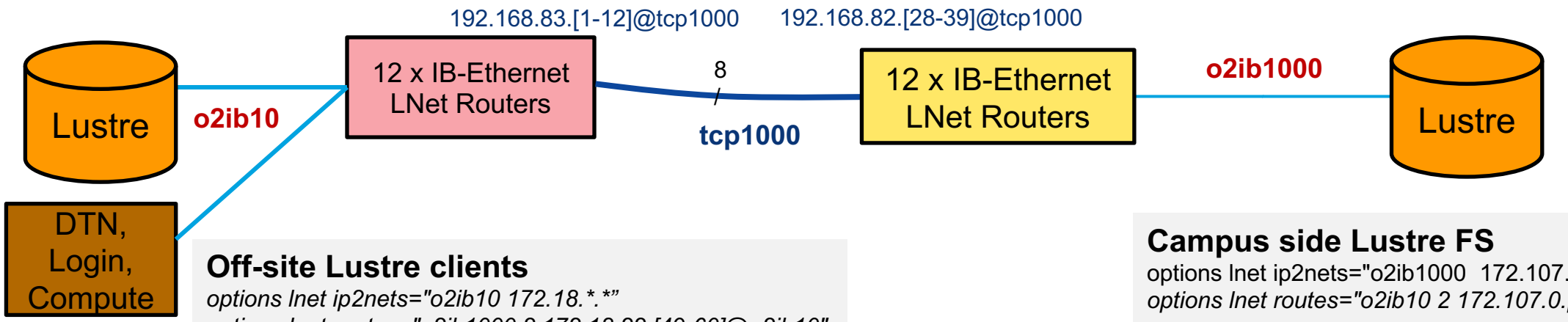


ExxonMobil

Lustre Configuration

Off-site datacenter side LNet Routers
options lnet ip2nets="tcp1000(enp94s0) 192.168.*.*; o2ib10 172.18.*.*"
options lnet routes="o2ib1000 1 192.168.82.[28-39]@tcp1000“
options lnet forwarding=enabled

Campus side LNet Routers
options lnet ip2nets="o2ib1000 172.107.*.*;tcp1000(enp94s0) 192.168.*.*"
options lnet routes="o2ib10 1 192.168.83.[1-12]@tcp1000 “
options lnet forwarding=enabled

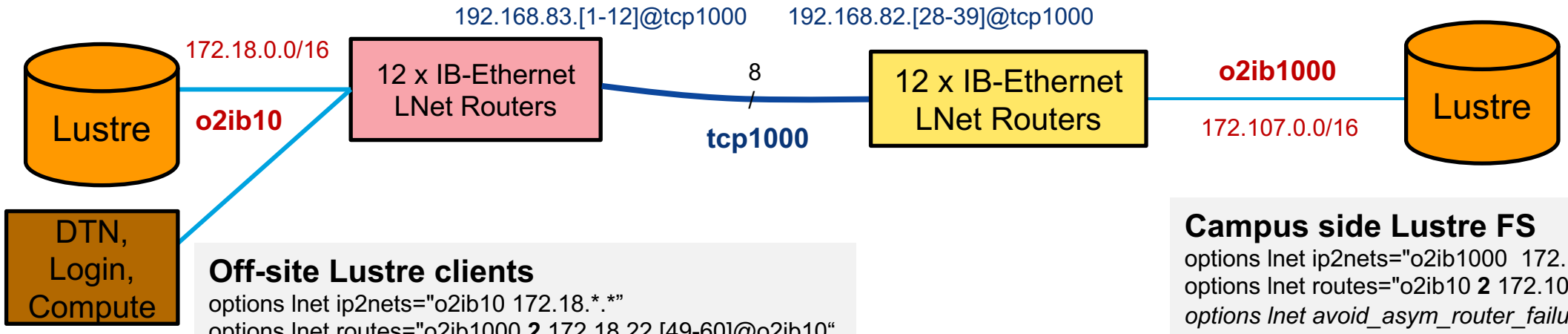


Off-site Lustre clients
options lnet ip2nets="o2ib10 172.18..*”*
options lnet routes="o2ib1000 2 172.18.22.[49-60]@o2ib10"

Campus side Lustre FS
options lnet ip2nets="o2ib1000 172.107.28.*"
options lnet routes="o2ib10 2 172.107.0.[28-39]@o2ib1000"

Client side issue

- lctl ping from repurposed compute node to OSS node on campus datacenter was failing after ~60 seconds after lnet was up
- Turn off Asymmetric Router Failure (enabled by default)



Off-site Lustre clients

options lnet ip2nets="o2ib10 172.18.*.*"
options lnet routes="o2ib1000 **2** 172.18.22.[49-60]@o2ib10"
options lnet avoid_asym_router_failure=0

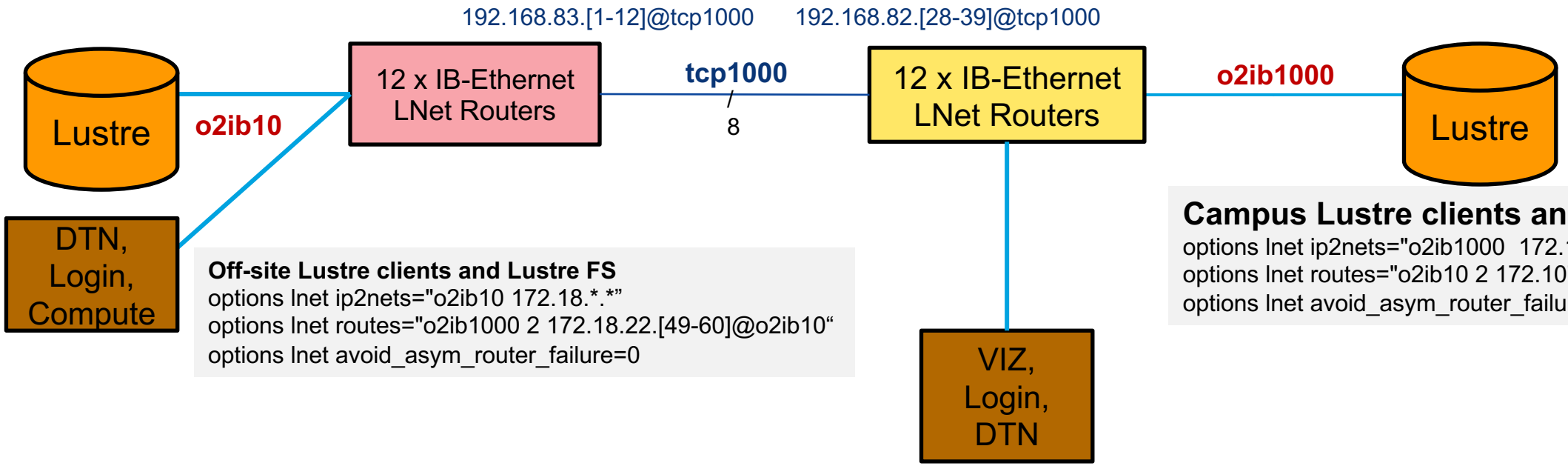
Campus side Lustre FS

options lnet ip2nets="o2ib1000 172.107.28.*"
options lnet routes="o2ib10 **2** 172.107.0.[28-39]@o2ib1000"
options lnet avoid_asym_router_failure=0

Campus clients to offsite Lustre FS

Off-site datacenter side LNet Routers
options lnet ip2nets="tcp1000(enp94s0) 192.168.*.*; o2ib10 172.18.*.*"
options lnet routes="o2ib1000 1 192.168.82.[28-39]@tcp1000“
options lnet forwarding=enabled

Campus side LNet Routers
options lnet ip2nets="o2ib1000 172.107.*.*;tcp1000(enp94s0) 192.168.*.*"
options lnet routes="o2ib10 1 192.168.83.[1-12]@tcp1000 “
options lnet forwarding=enabled



Off-site Lustre clients and Lustre FS
options lnet ip2nets="o2ib10 172.18.*.*”
options lnet routes="o2ib1000 2 172.18.22.[49-60]@o2ib10“
options lnet avoid_asym_router_failure=0

Campus Lustre clients and Lustre FS
options lnet ip2nets="o2ib1000 172.107.28.*"
options lnet routes="o2ib10 2 172.107.0.[28-39]@o2ib1000“
options lnet avoid_asym_router_failure=0

ExxonMobil

Conclusion

- Built an 8 x 100Gbps links between two datacenters which are ~50KM apart
- Used IB-Ethernet LNET routers on both datacenters
- Used multi hops LNET routing
- Enabled us to use existing data copy tools to migrate projects between two centers
- Enabled interactive users to access data across both centers

Acknowledgements

- HPC Systems Team. Bryan White for running IOR countless number of times.
- Networking Team for pointing us towards DWDM technology

Thank you