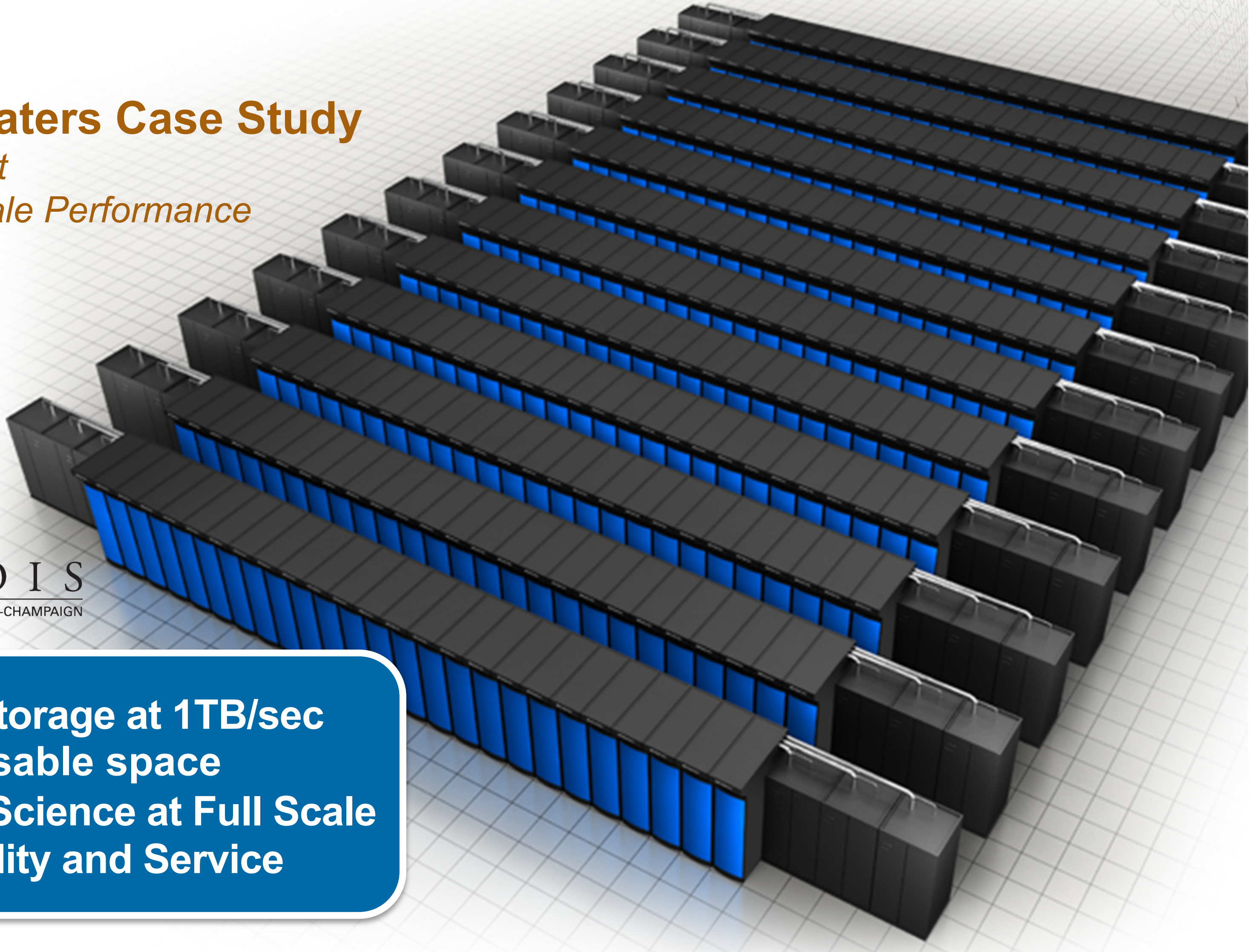
Enabling Scientific Breakthroughs at the Petascale



NCSA Blue Waters Case Study

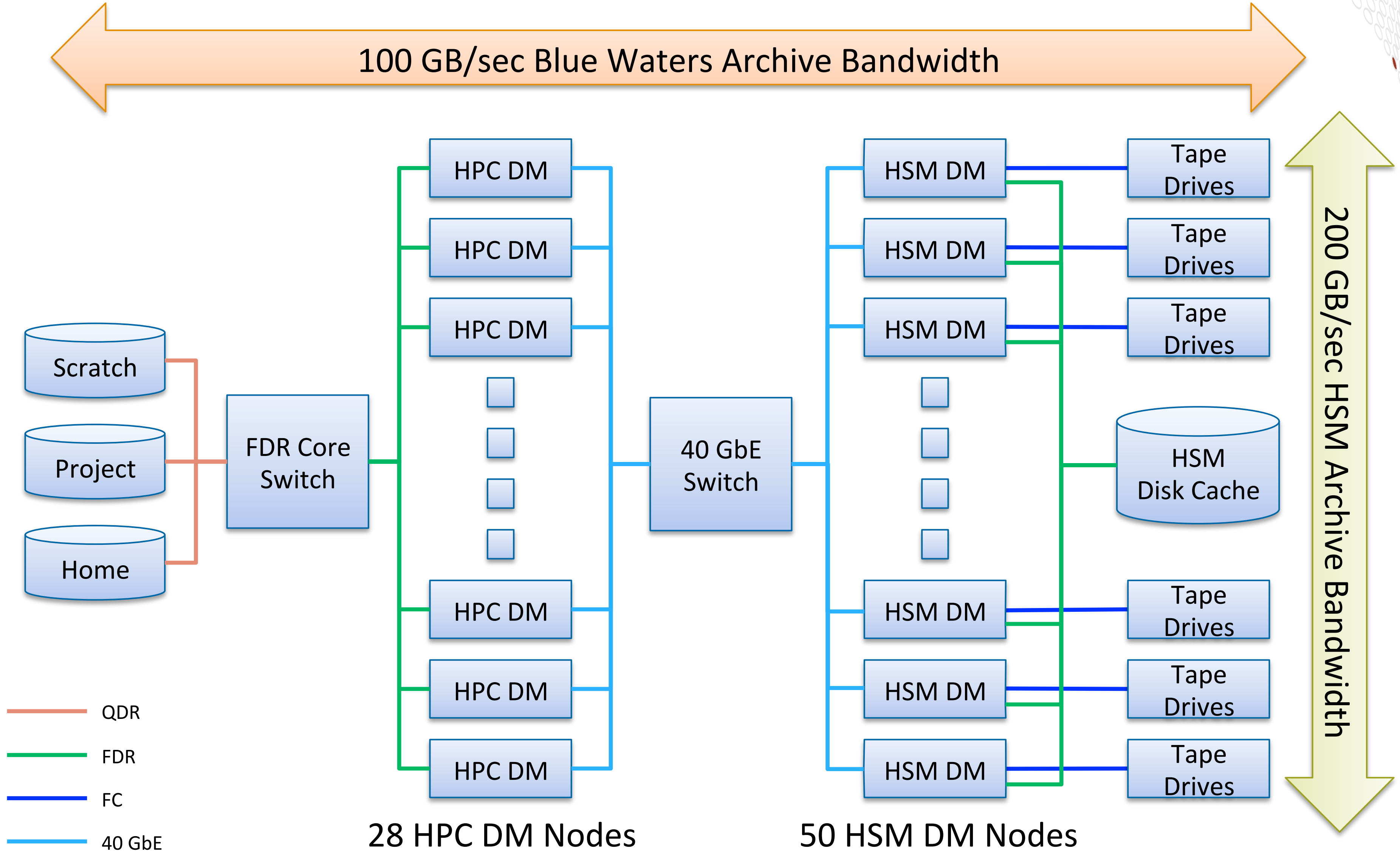
Supercomputing at
Sustained Petascale Performance

NCSA

ILLINOIS
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

- Integrated Storage at 1TB/sec
- 25+ PB of usable space
- Production Science at Full Scale
- Cray Reliability and Service

Blue Waters HSM Architecture (Traditional)



100 GB/sec Blue Waters Archive Bandwidth

200 GB/sec HSM Archive Bandwidth

Scratch

Project

Home

FDR Core Switch

HPC DM

HPC DM

HPC DM

HPC DM

HPC DM

HPC DM

40 GbE Switch

HSM DM

HSM DM

HSM DM

HSM DM

HSM DM

HSM DM

Tape Drives

Tape Drives

Tape Drives

HSM Disk Cache

Tape Drives

Tape Drives

Tape Drives

QDR

FDR

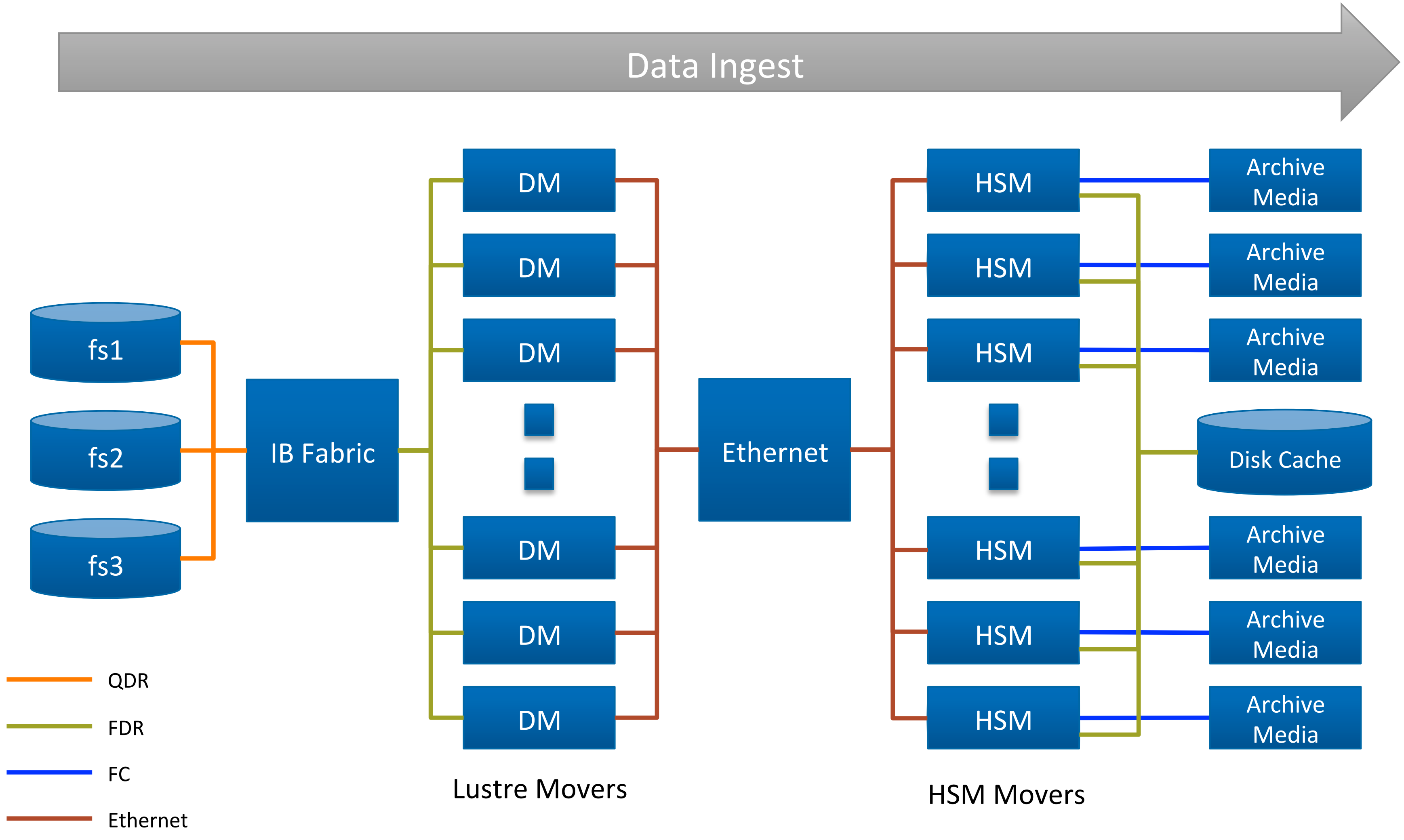
FC

40 GbE

28 HPC DM Nodes

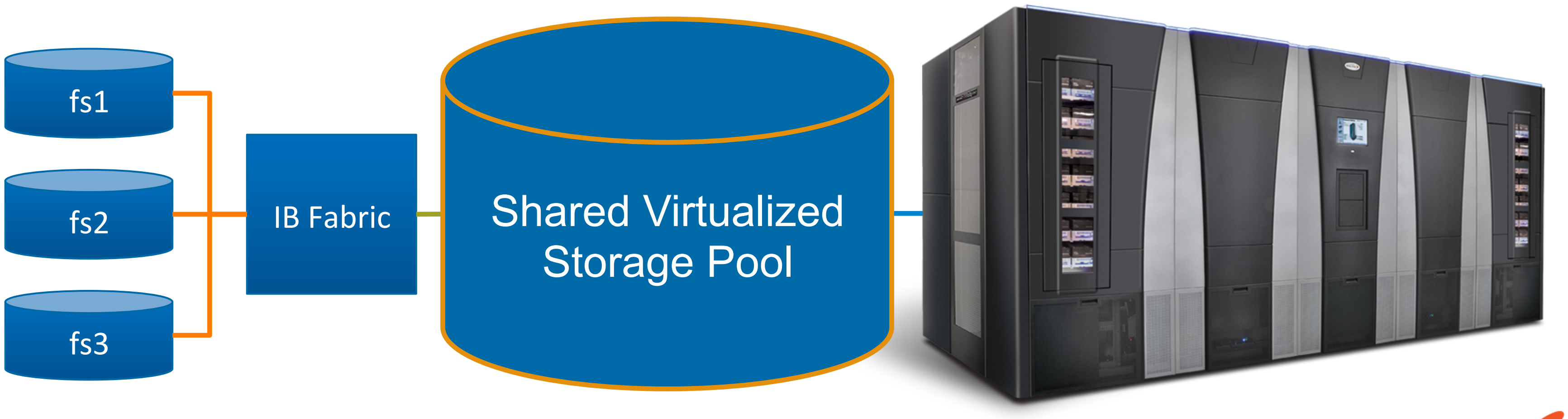
50 HSM DM Nodes

Traditional HSM Implementation



Cray Tiered Adaptive Storage

Data Movement and Transparent User Access

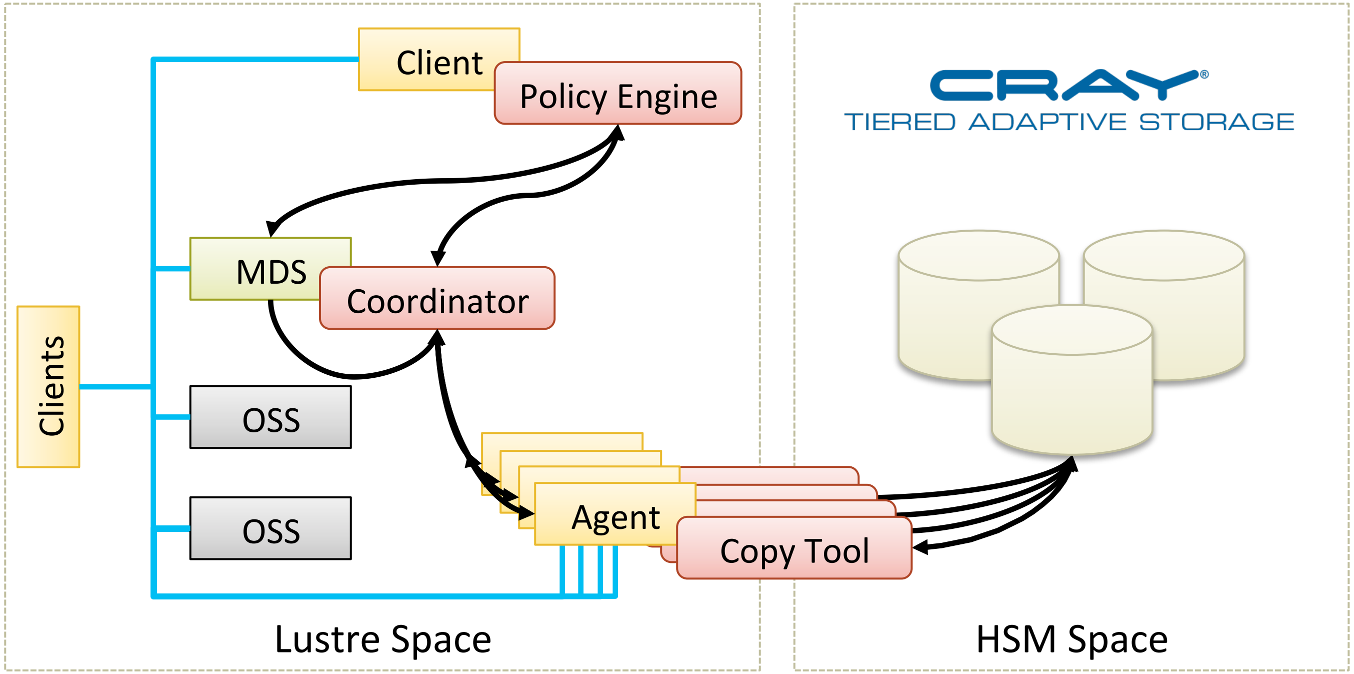


- QDR
- FDR
- FC
- Ethernet

CRAY® TIERED ADAPTIVE STORAGE

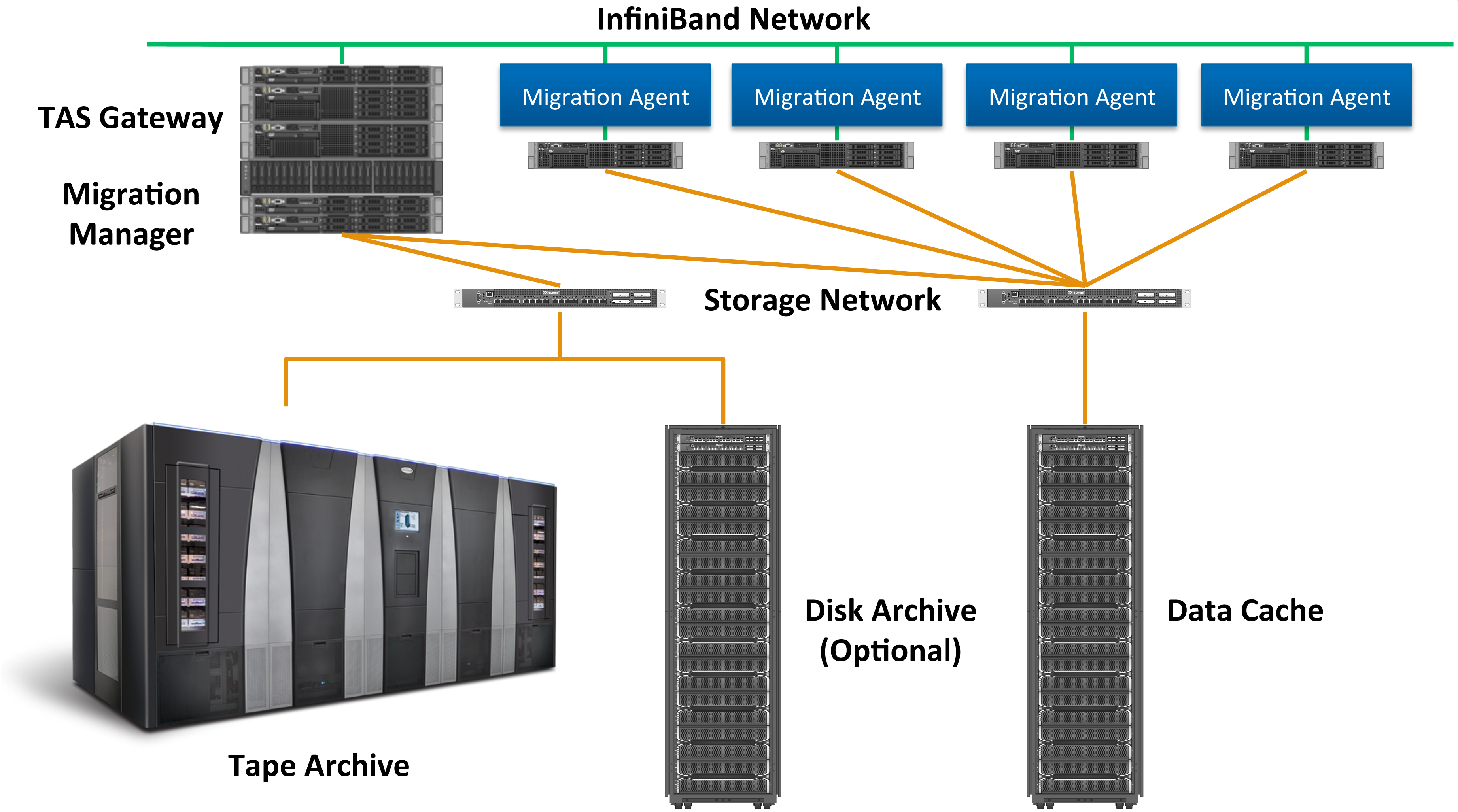
Versity

Cray TAS Connector for Lustre File System



- **Policy engine**
  - Robinhood policy engine to manage Lustre namespace activity
- **Coordinator**
  - Communicates with policy engine and agents to manage data movement
- **Agent and Copy Tool**
  - Lustre clients with copy tool software to migrate data between Lustre and TAS

Cray Tiered Adaptive Storage Lustre File System HSM Solution



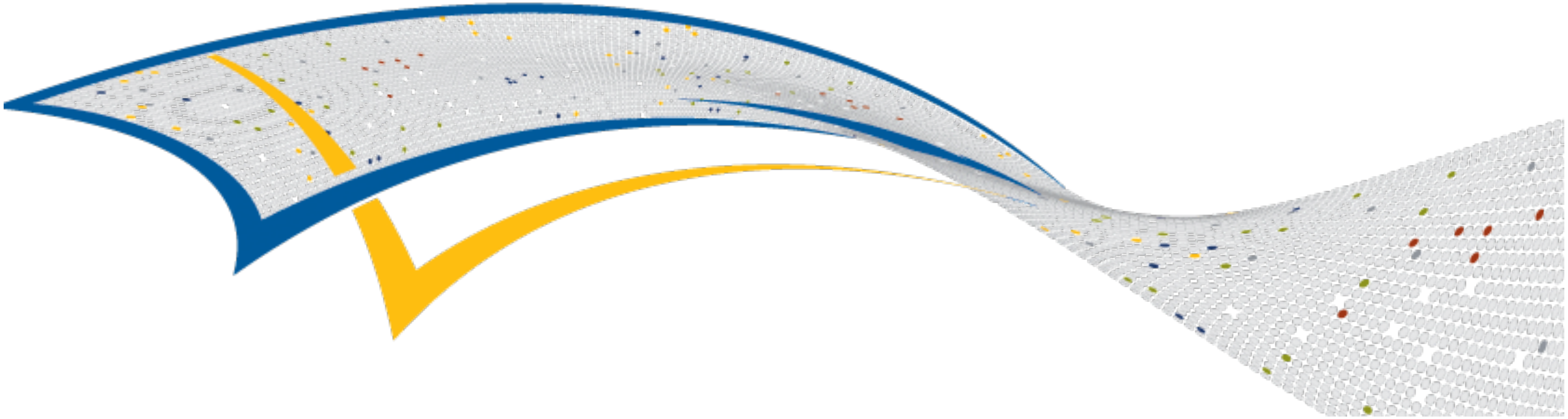
Cray Lustre Connector

- **User and application interface to all data is Lustre file system**
- **Complete solution for Lustre HSM capabilities**
  - Requires newly released Lustre 2.5 with HSM server-side extensions
    - Support for 3rd party Lustre 2.5 solutions 1Q2015
  - Robin Hood policy engine
    - Provided as part of Cray Lustre Connector solution
- **Scalable performance**
  - Optimized Lustre data movers
  - Parallel scaling of data movers to improve performance
  - Distributed namespace (DNE) compatible
- **Cray storage platform availability**
  - 1Q2015 - Lustre File System by Cray (CLFS)
  - 1H2015 - Cray Sonexion Data Storage System

Cray Tiered Adaptive Storage Complete Data Management Solutions

- **Cray Tiered Adaptive Storage**
  - Complete data management solutions
  - Architected and tuned by Cray
  - Simplifying complex architectures
  - Integration with Lustre HSM feature
- **Versity Storage Manager**
  - Built for Linux
  - Open archive data format
  - Mature technology based on open source SAM-QFS
- **World-class support from Cray**

CRAY®
TIERED ADAPTIVE STORAGE

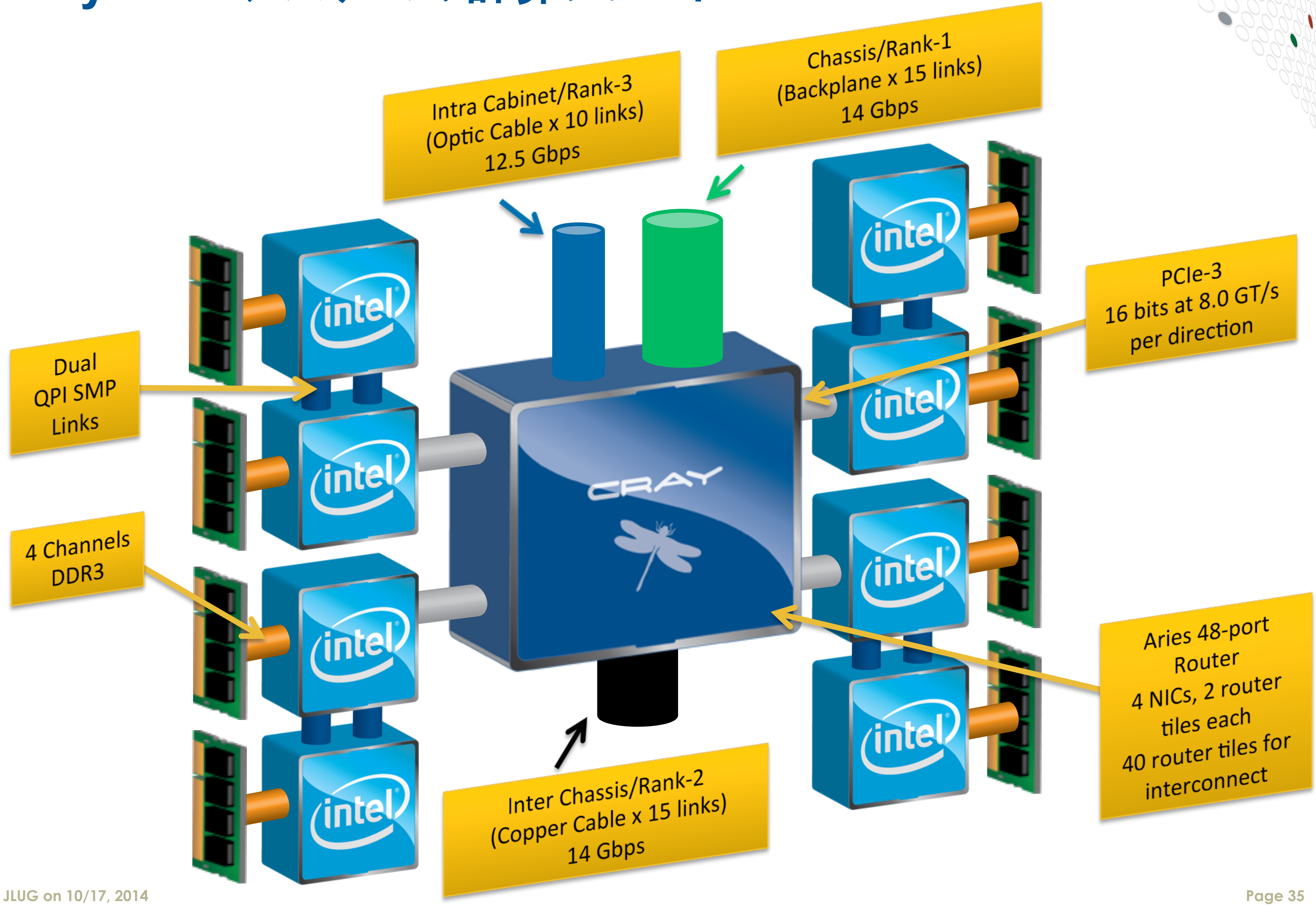


CRAY®

THE SUPERCOMPUTER COMPANY

バックアップ

Cray XC40システムの計算ブレード



Intra Cabinet/Rank-3
(Optic Cable x 10 links)
12.5 Gbps

Chassis/Rank-1
(Backplane x 15 links)
14 Gbps

PCIe-3
16 bits at 8.0 GT/s
per direction

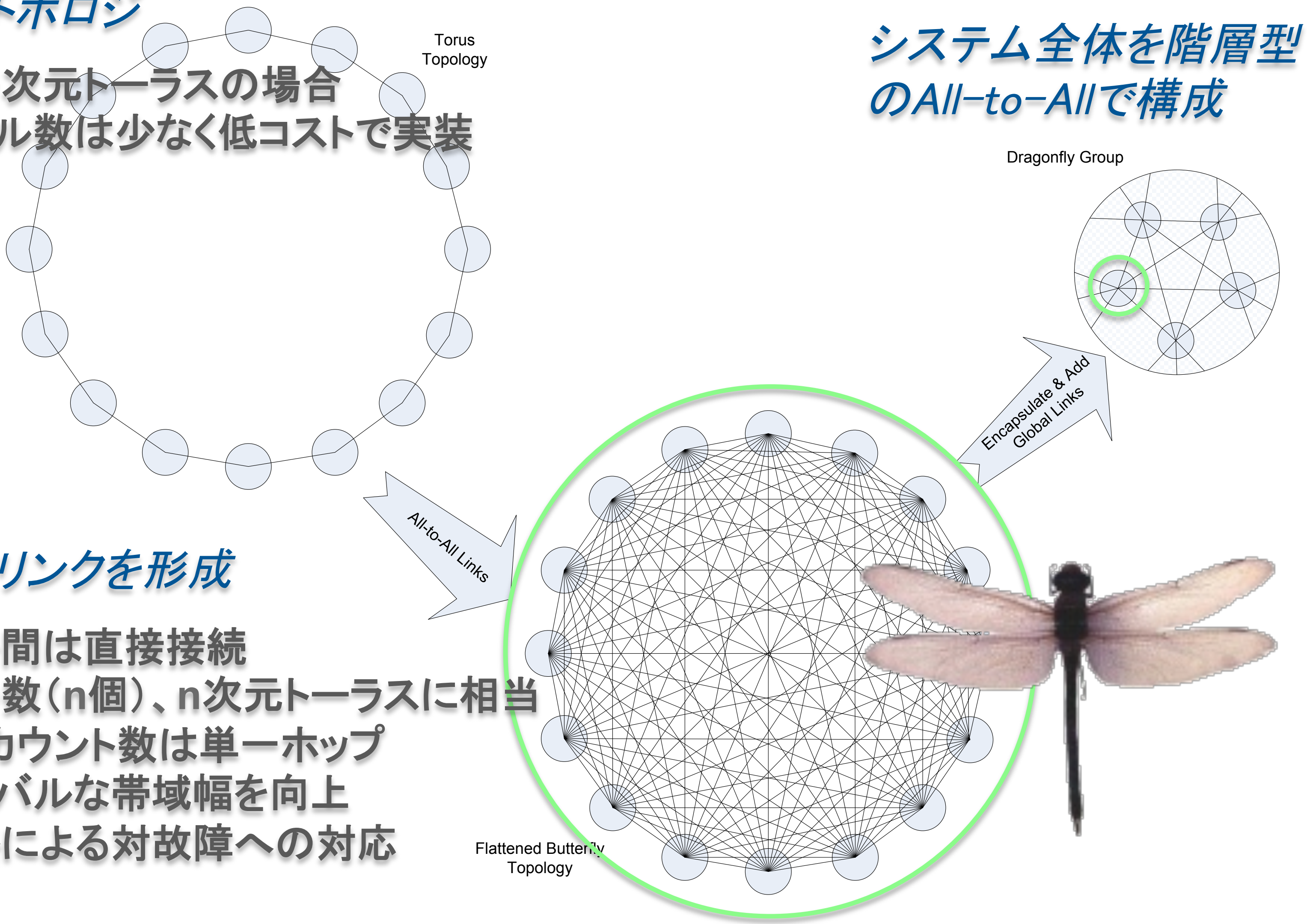
Dual
QPI SMP
Links

4 Channels
DDR3

Aries 48-port
Router
4 NICs, 2 router
tiles each
40 router tiles for
interconnect

Inter Chassis/Rank-2
(Copper Cable x 15 links)
14 Gbps

階層型All-to-AllのDragonflyネットワークトポロジー



トーラス・トポロジ

- 図は一次元トーラスの場合
- ケーブル数は少なく低コストで実装

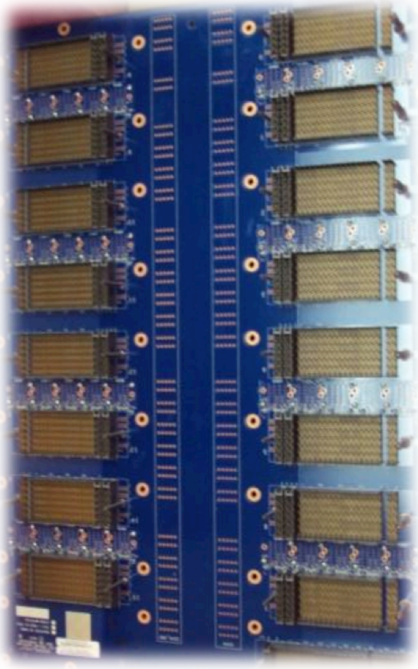
システム全体を階層型のAll−to−Allで構成

All−to−Allリンクを形成

- 各端子間は直接接続
- 端子の数（n個）、n次元トーラスに相当
- ホップカウント数は単一ホップ
- グローバルな帯域幅を向上
- 迂回路による対故障への対応

インターコネクトのケーブリング

- Cray XC40システムは、インターコネクトのバンド幅を十分に確保し、かつケーブルコストを考慮した設計になっています。

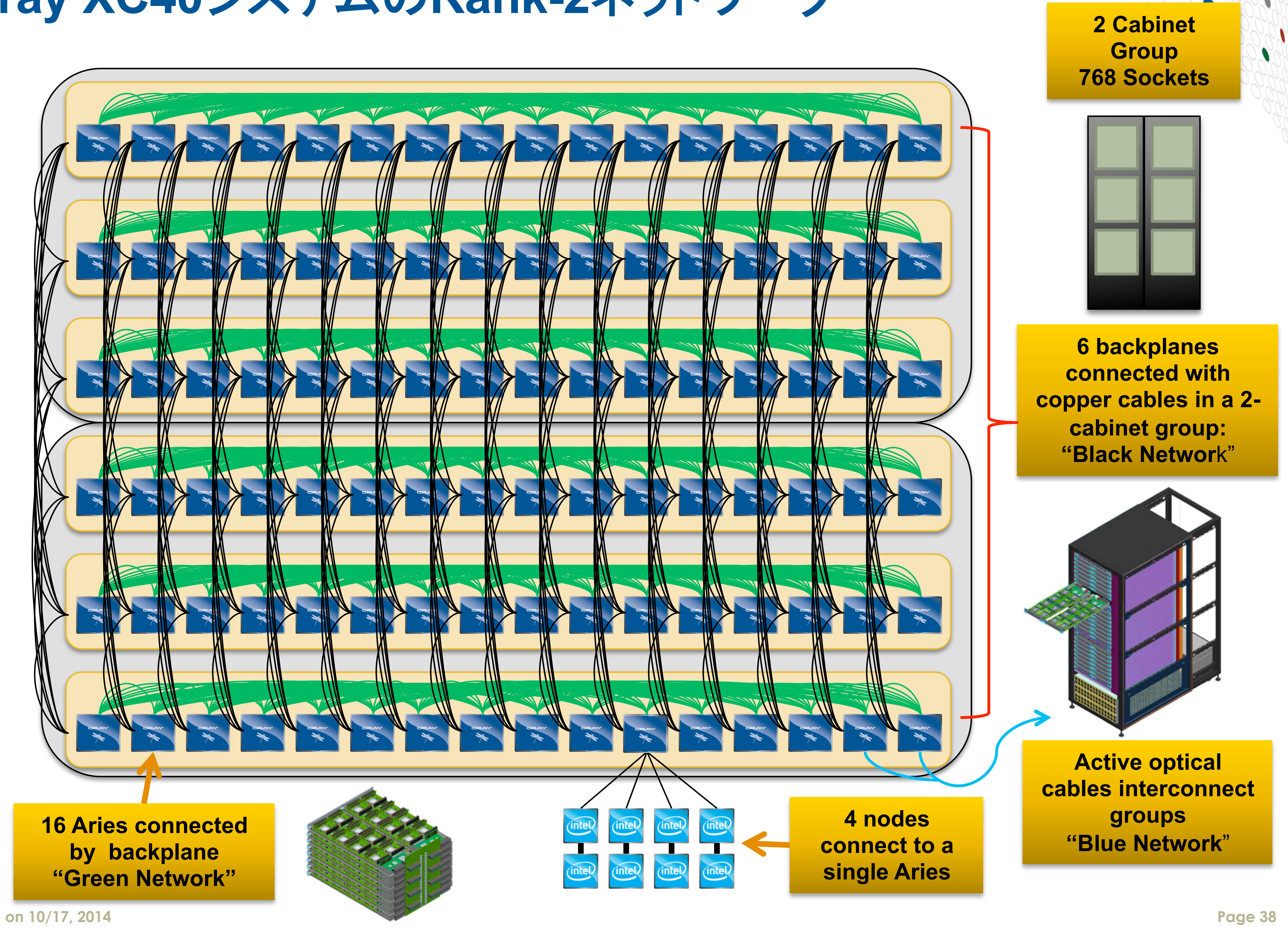


シャーシ内
（Rank-1）
バックプレーン
直接接続

グループ内
（Rank-2）
パッシブ・カッパー
ケーブル

グループ間
（Rank-3）
アクティブ・ファイバー
ケーブル

Cray XC40システムのRank-2ネットワーク



2 Cabinet Group
768 Sockets

6 backplanes connected with copper cables in a 2-cabinet group: “Black Network”

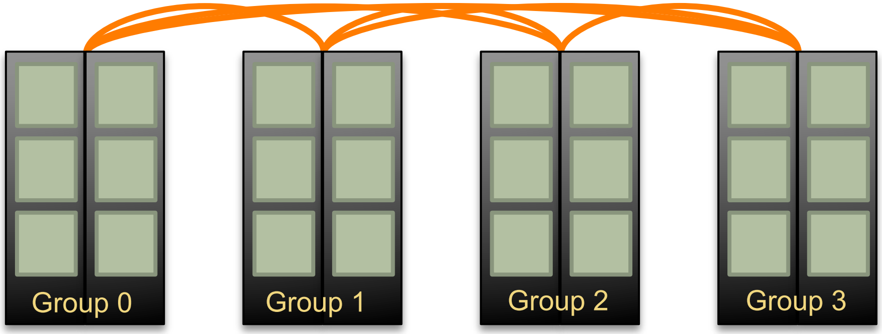
Active optical cables interconnect groups “Blue Network”

16 Aries connected by backplane “Green Network”

4 nodes connect to a single Aries

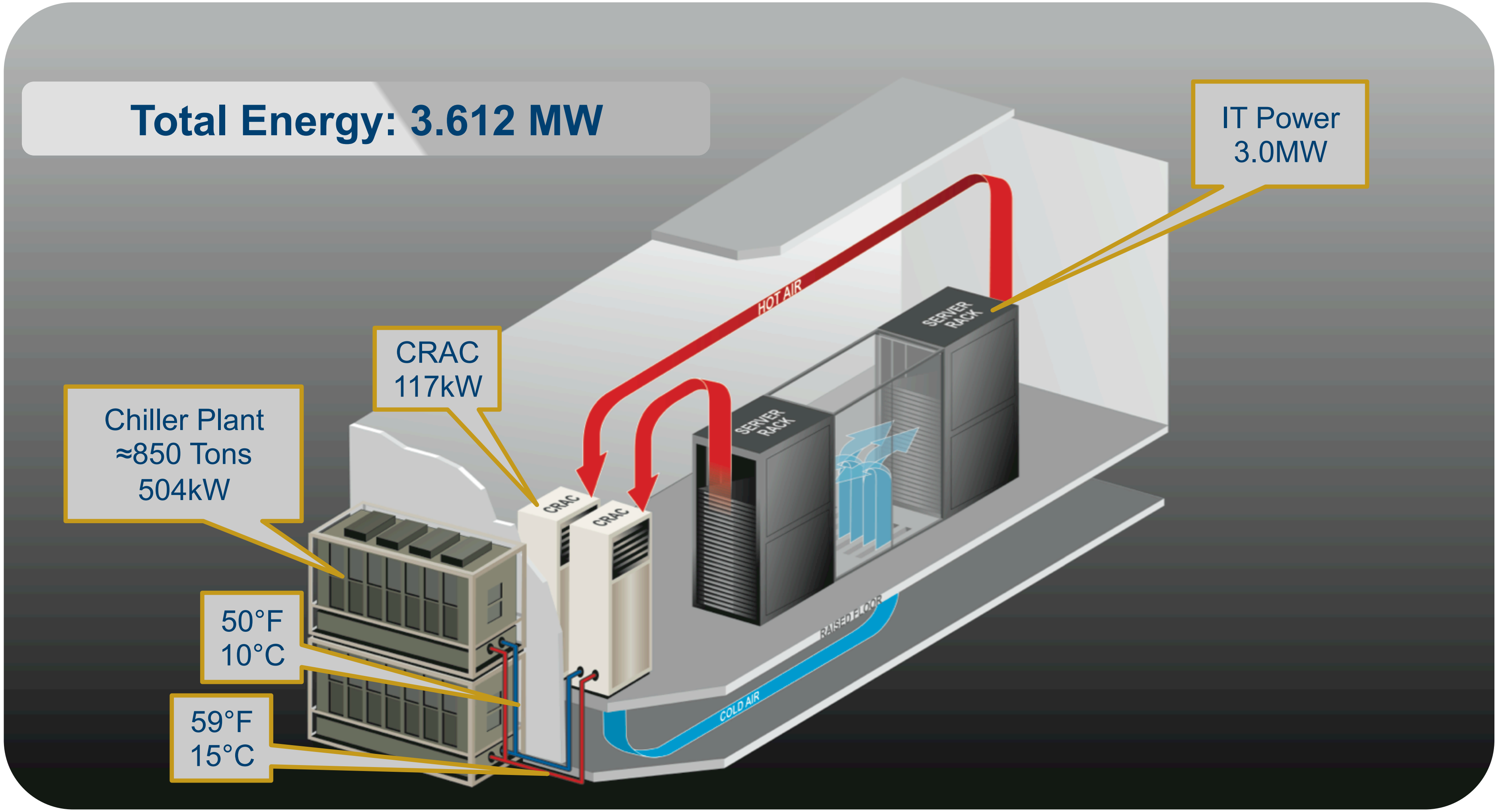
Cray XC40システムのRank-3ネットワーク

- An all-to-all pattern is wired between the groups using optical cables
- Up to 240 ports are available per 2-cabinet group
- The global bandwidth can be tuned by varying the number of optical cables in the group-to-group connections



Example: An 4-group system is interconnected with 6 optical "bundles". The "bundles" can be configured between 20 and 80 cables wide

Sample Air Cooled 3MW System



Total Energy: 3.612 MW

IT Power
3.0MW

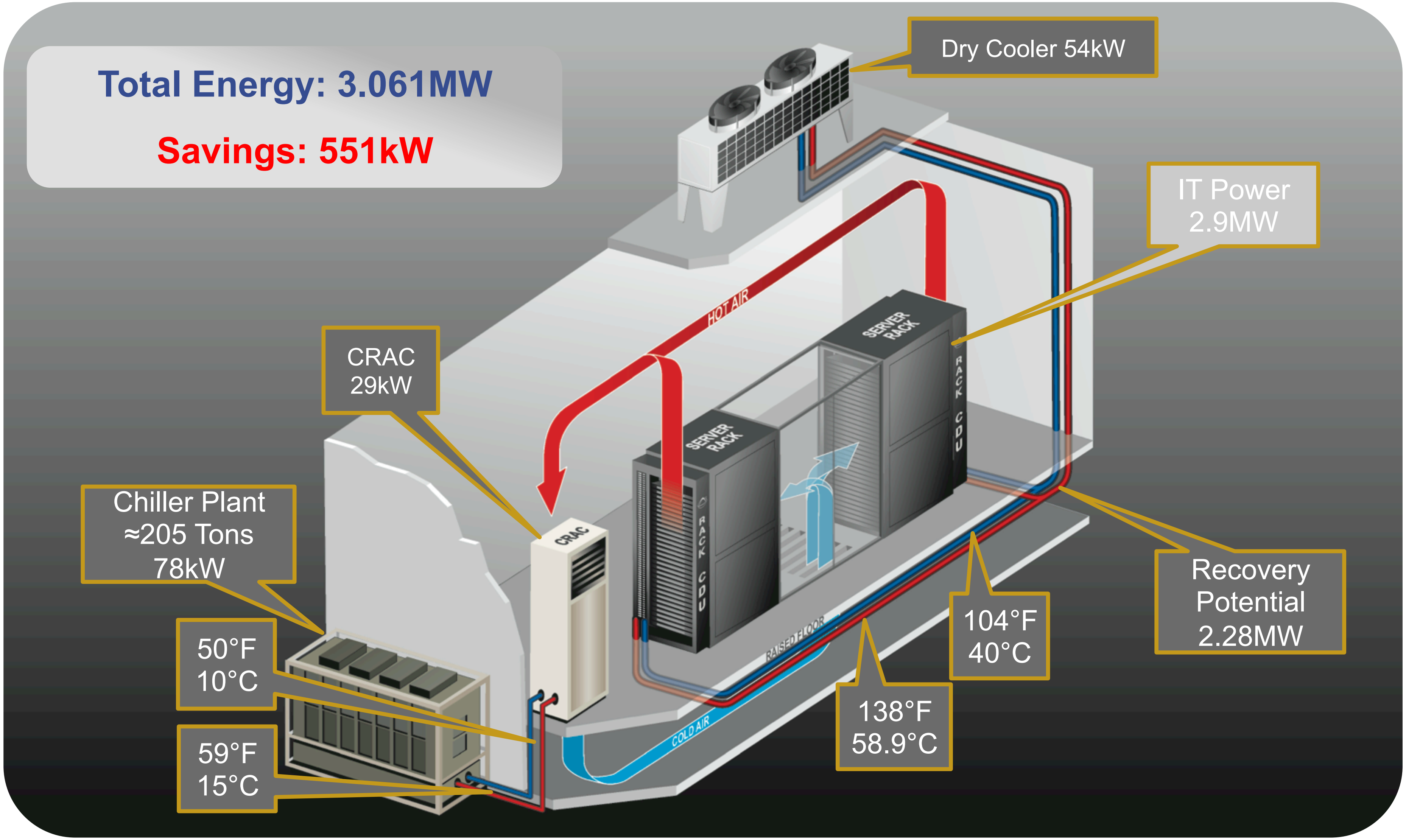
CRAC
117kW

Chiller Plant
≈850 Tons
504kW

50°F
10°C

59°F
15°C

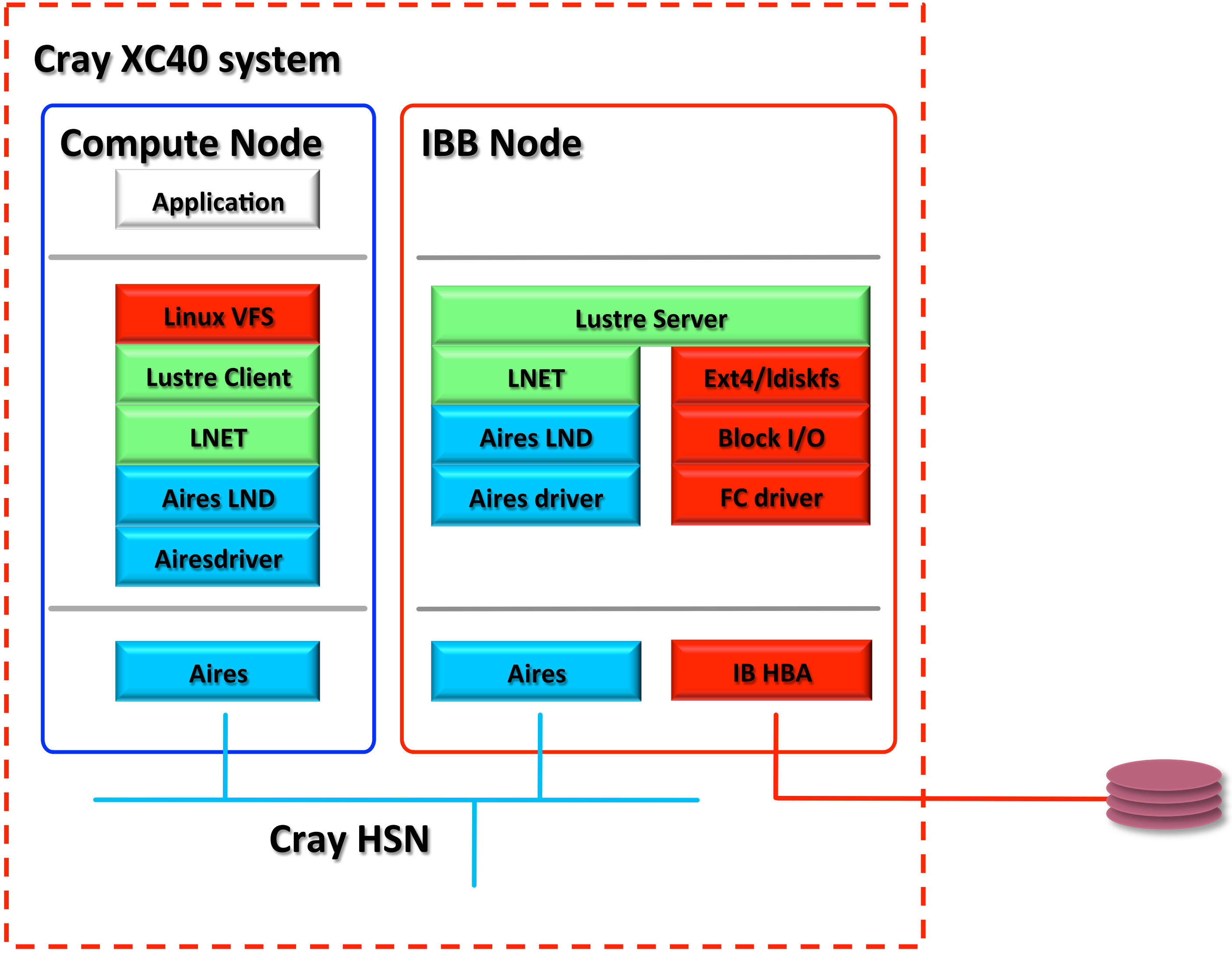
Example Savings for 3MW System



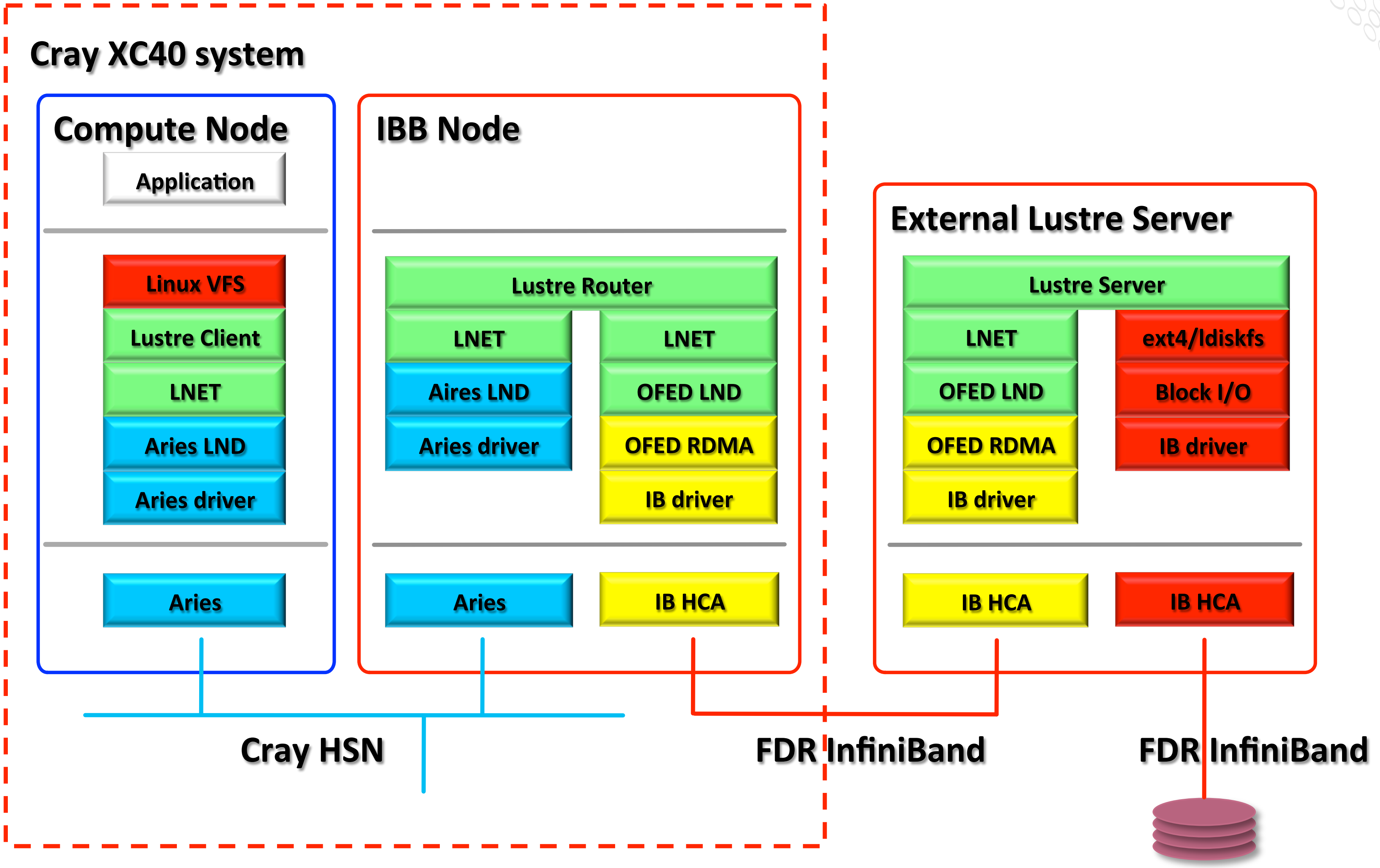
Total Energy: 3.061MW

Savings: 551kW

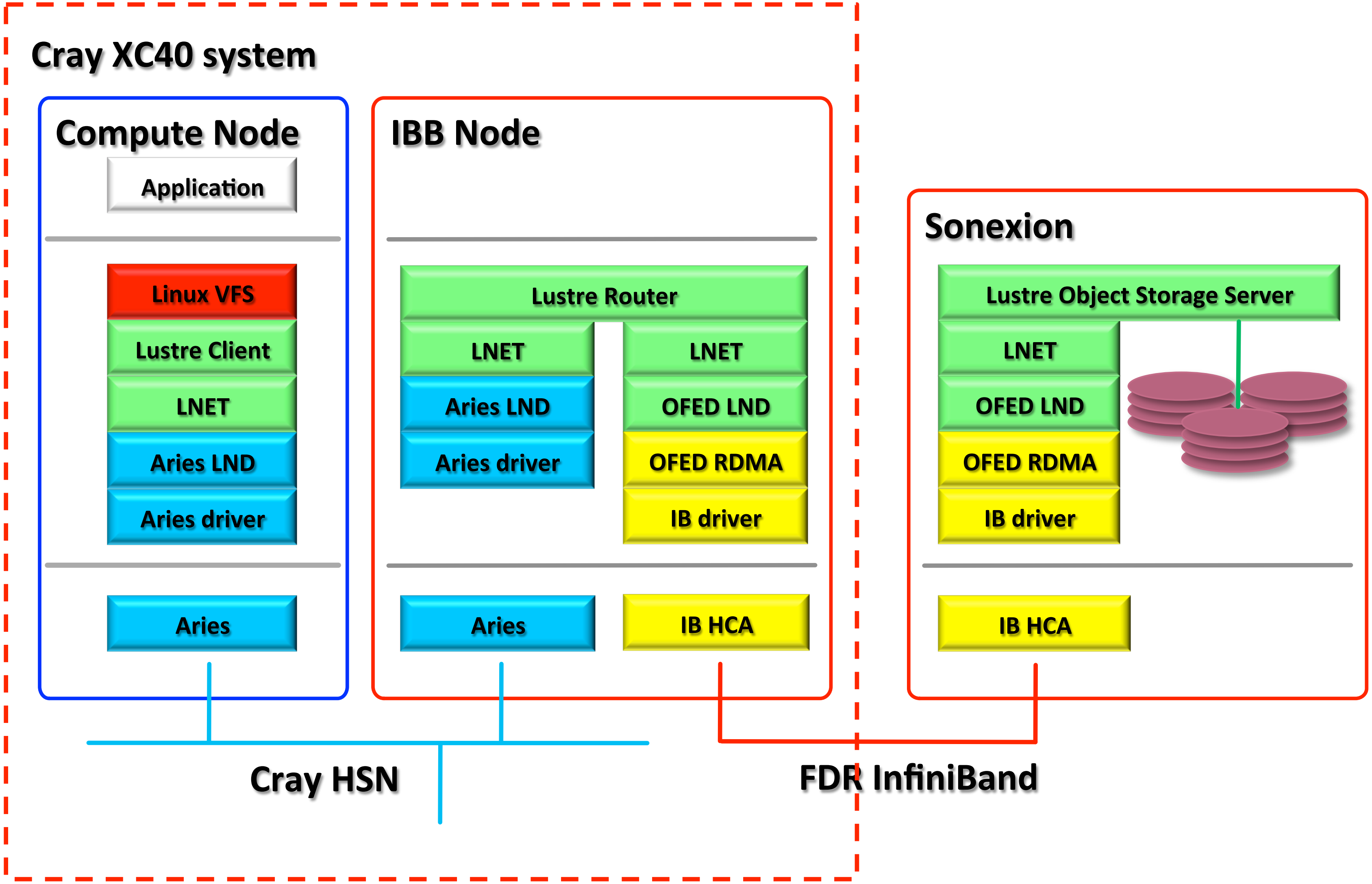
Cray Direct Attached Lustre



Cray CLFS


Cray XC40 system
Compute Node
Application
Linux VFS
Lustre Client
LNET
Aries LND
Aries driver
Aries
IBB Node
Lustre Router
LNET
Aires LND
Aries driver
LNET
OFED LND
OFED RDMA
IB driver
Aries
IB HCA
External Lustre Server
Lustre Server
LNET
OFED LND
OFED RDMA
IB driver
ext4/ldiskfs
Block I/O
IB driver
IB HCA
IB HCA
Cray HSN
FDR InfiniBand
FDR InfiniBand

Cray Sonexion



Spectra Logic

- **Midrange Libraries**
  - T200, T380 and T680
    - 500 TB to 1.7 TB native capacity
    - Up to 14 LTO-6 drives
- **Spectra T950**
  - 2.3 PB to 25.1 PB native capacity
  - Up to 120 LTO-6 drives
- **Spectra T-Finity**
  - Single library 2.3 PB to 125 PB
  - Library complex up to 1 EB
- **BlueScale library management**
  - Media Lifecycle Management (MLM)
  - Library Lifecycle Monitoring (LLM)
  - BlueScale encryption
  - Data Integrity Verification (DIV)

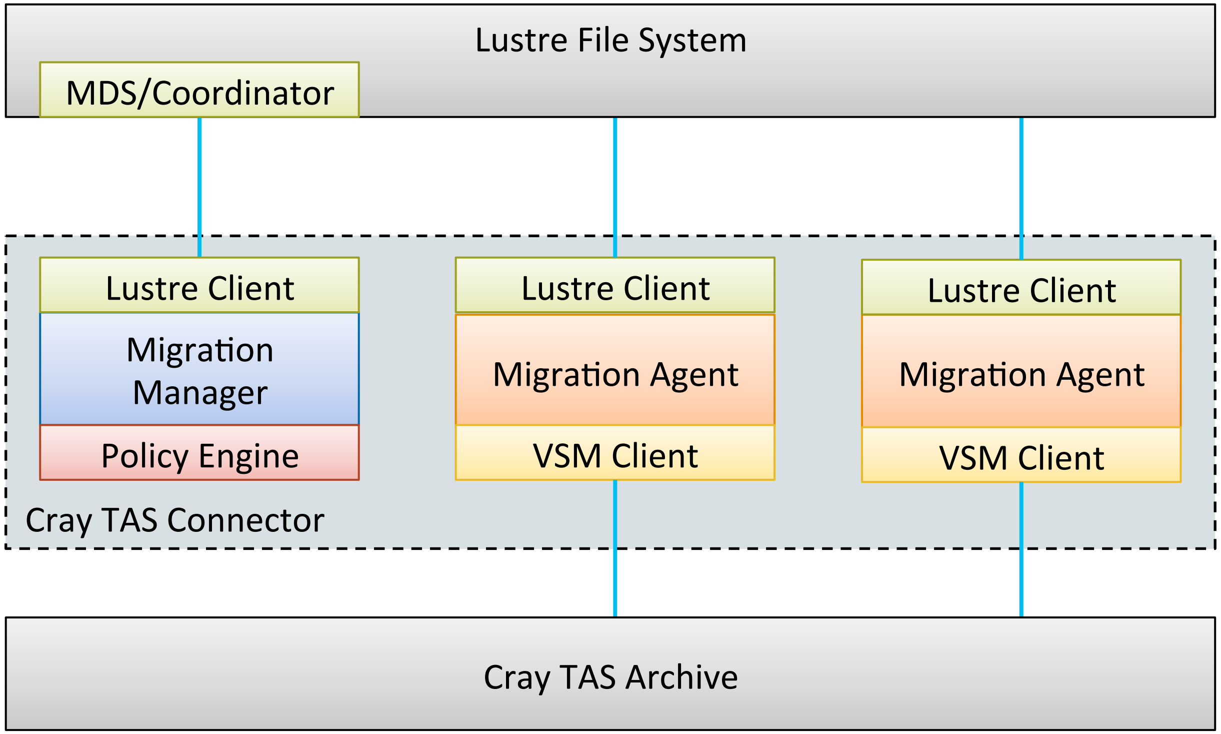
SPECTRA
SPECTRALOGIC.COM

Cray TAS Connector Components

- **Integrated Lustre HSM solution including**
  - Migration manager and agents for scale out performance
  - Currently backend storage agnostic by using VFS interface
- **Policy engine**
  - Consumes Lustre change log events
    - File and directory creation, modification, removal and attribute updates
  - Migration policies for file classes dictate when and where files are archived
- **Migration Manager**
  - Receives HSM actions from Lustre file system coordinator
  - Queues and dispatches requests to Migration Agents
  - Provides for better control and parallelism than multiple copy tools
- **Migration Agents**
  - Scale-out data movement from Lustre to backend HSM
  - Currently uses file system interface for HSM backend
  - Stores Lustre file layout and attributes in single file along with data

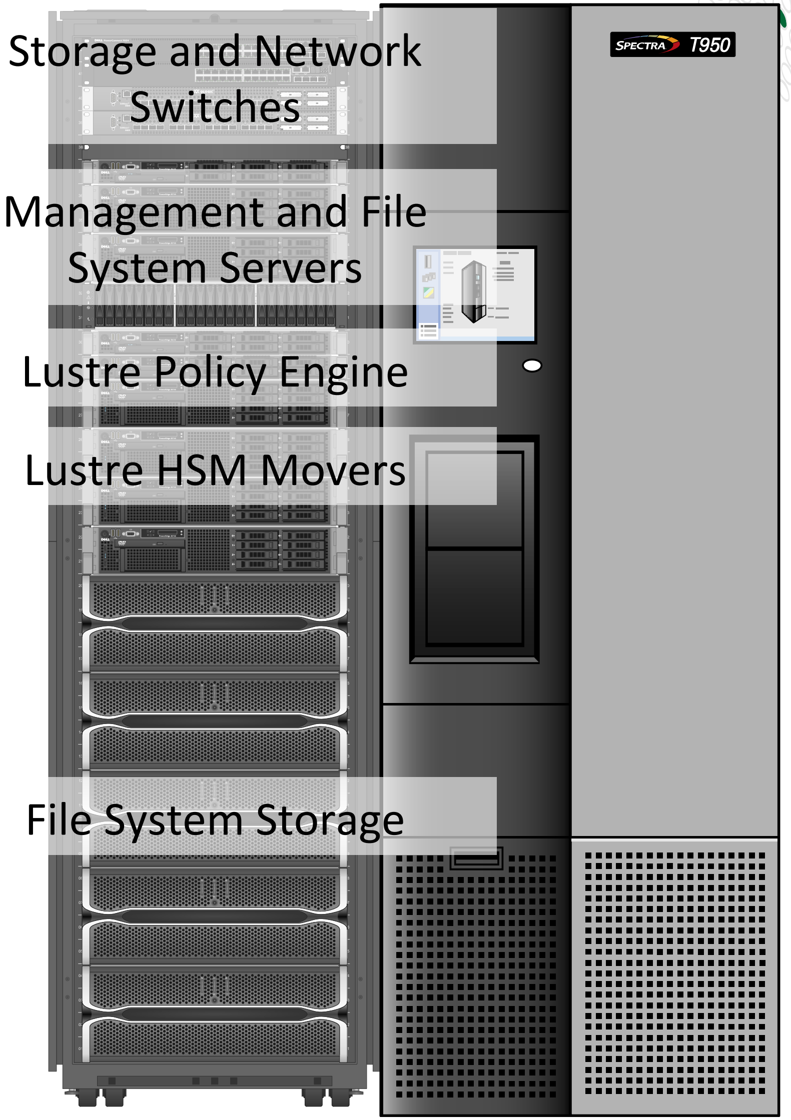
Cray TAS Connector Connections and Clients

- **Migration Manager**
  - Changelog consumer and policy engine
  - Communicates with Coordinator and updates status of requests
- **Migration Agents migrate data between Lustre and TAS**



Sample TAS Configuration

- **Cray TAS Gateway**
  - Two (2) file system servers
  - One (1) management server
  - VSM metadata storage
- **Lustre HSM servers**
  - Two (2) Lustre policy engine servers
  - Four (4) Lustre HSM data movers
- **Spectra T950 Library**
  - Single frame with 8 drives
  - Up to 920 LTO-6 cartridges
- **File system capabilities**
  - 1.1 PB of file system capacity
  - Up to 6 GB/sec of throughput
- **Tiering capabilities**
  - 2.0 PB of native LTO-6 capacity
    - Expandable to 25.1 PB
  - 1.2 GB/sec throughput to tape media



Storage and Network Switches
Management and File System Servers
Lustre Policy Engine
Lustre HSM Movers
File System Storage
SPECTRA T950