LUG JAPAN 2022

OISTにおけるLustre活用のこれまでとこれから

田仲 康司
HPCシステムエンジニア@OIST

概要

1. OISTの紹介

2. OISTにおけるLustreの歴史（2008～）

3. 今後の課題と展望

OISTのHPCシステムエンジニアについて

- HPCシステムエンジニア@OIST
  1) 導入されたハードウェアに合わせてクラスタを構築する
  2) HPCだけでなくストレージやその他のシステムの維持管理も行う
  3) 新しいニーズやパッチに応じてシステムをアップデートする
  4) システム管理者の存在をなるべく感じさせないように心がける
  5) システム障害 → 研究が止まる → 「早く直してくれ！」と言われる

HPCシステムエンジニアを募集中!

https://unsplash.com/photos/5jz3T5LwuPA

OISTキャンパス

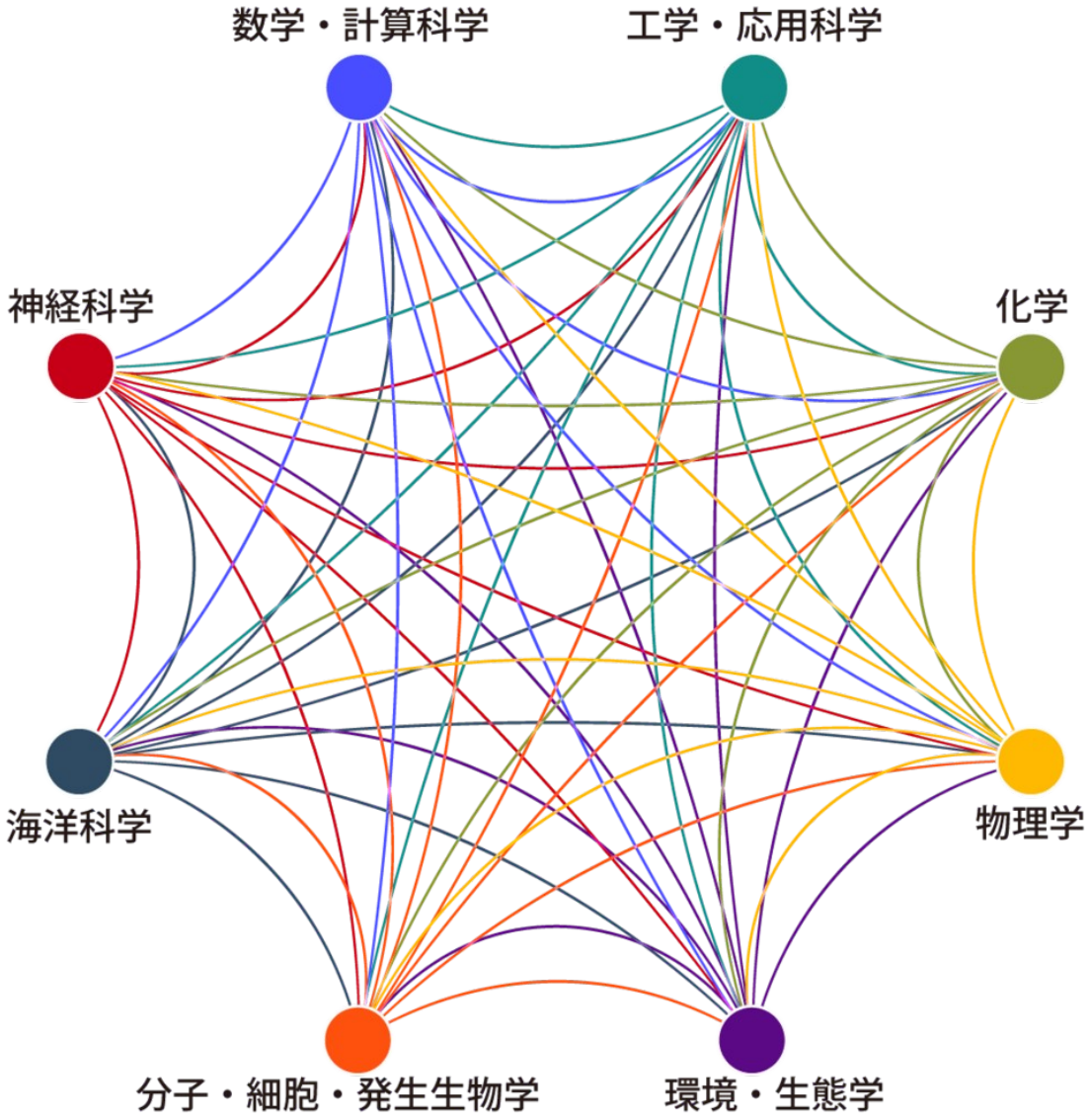
研究分野について

研究分野

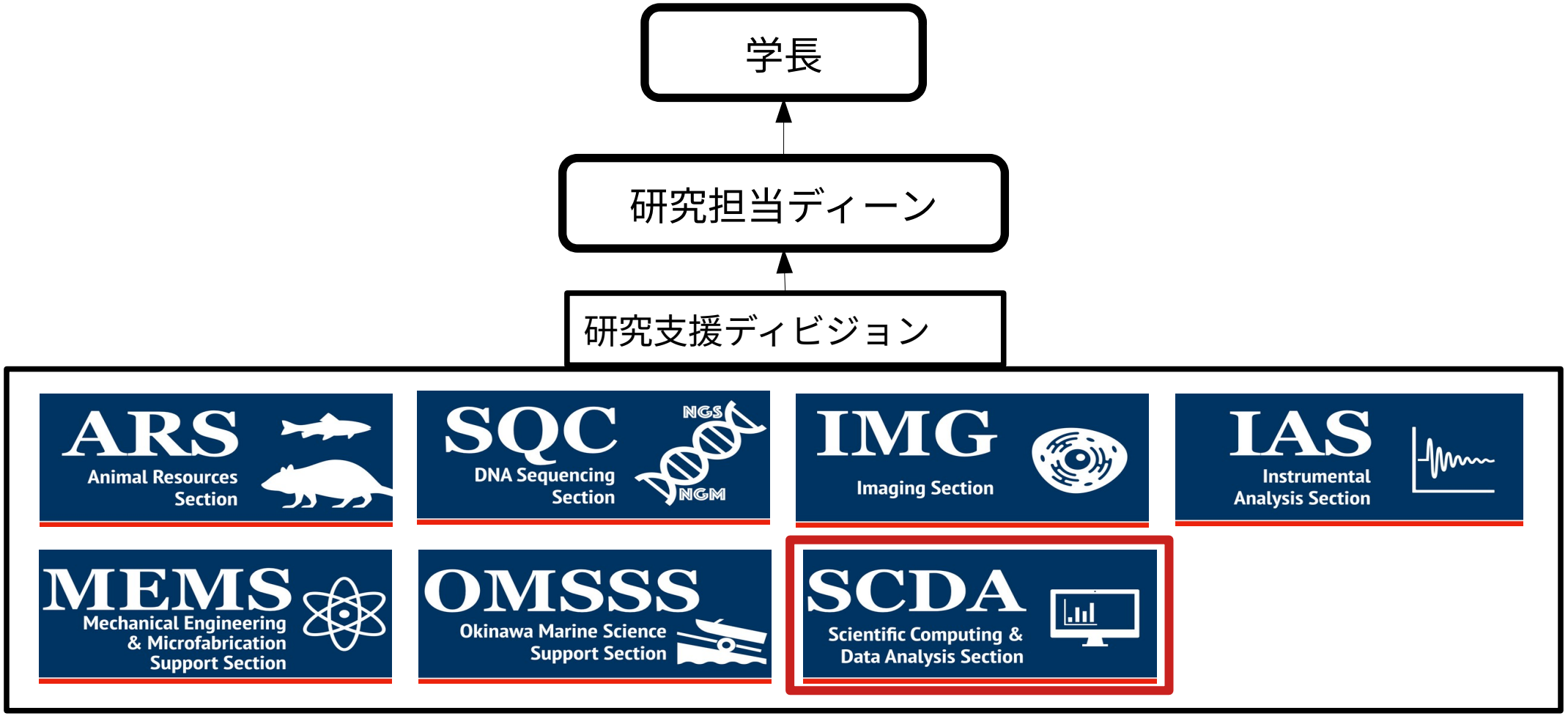
- 物理学
- 化学
- 神経科学
- 海洋科学
- 環境・生態学
- 数学・計算科学
- 分子・細胞・発生生物学
- 工学・応用科学

学際的な研究と教育

- 学部が存在しない
- 博士課程コースの分野が限定されない
- 研究支援ディビジョンが存在する



研究支援ディビジョン



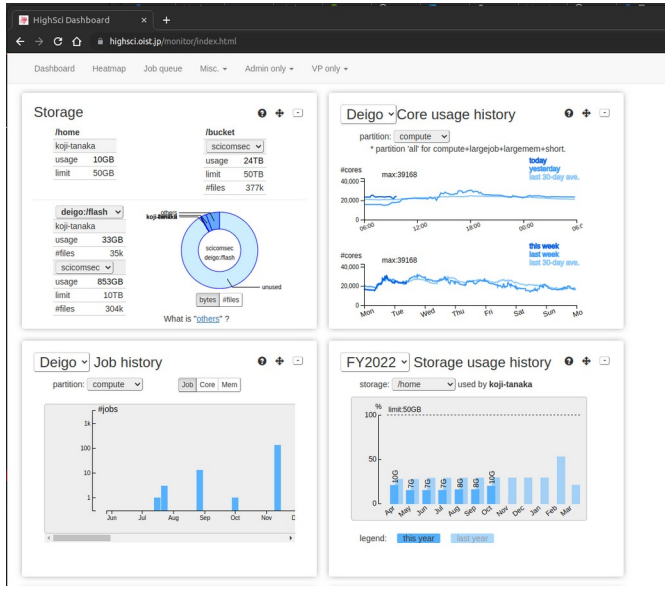
https://groups.oist.jp/rsd

SCDAセクションについて

- HPCとストレージの管理・運用
- HPC利用者へのトレーニングとサポート
  - 定期的にハンズオンセミナーを開催
  - 平日15:30~17:30はオープンアワー
- ユーザ向けウェブサービスHighSci(ハイサイ)の運用
  - HPCやストレージのステータス、グループ管理など
- Cloudサービスおよび VMの管理・運用
  - オンプレ、パブリッククラウド
- データ可視化のトレーニングとサポート
  - 研究データ → 可視化プリプロセス → 画像/動画
- データ取得と解析のサポート
  - 研究機器 → ストレージ → HPC
- 商用ソフトウェアライセンスの管理と運用
  - サイトライセンスとコンカレントライセンス活用を推進
  - 定期的にソフトウェア・ウィークを開催

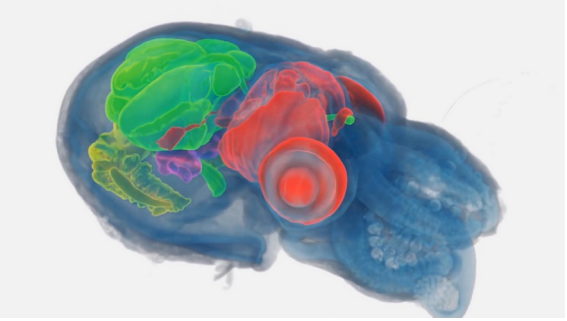
SCDA Open Hours

HighSci



https://axies.jp/_files/conf/conf2020/FA2-2.pdf

BlenderによるMicro-CT scanデータの
ボリューム・レンダリング



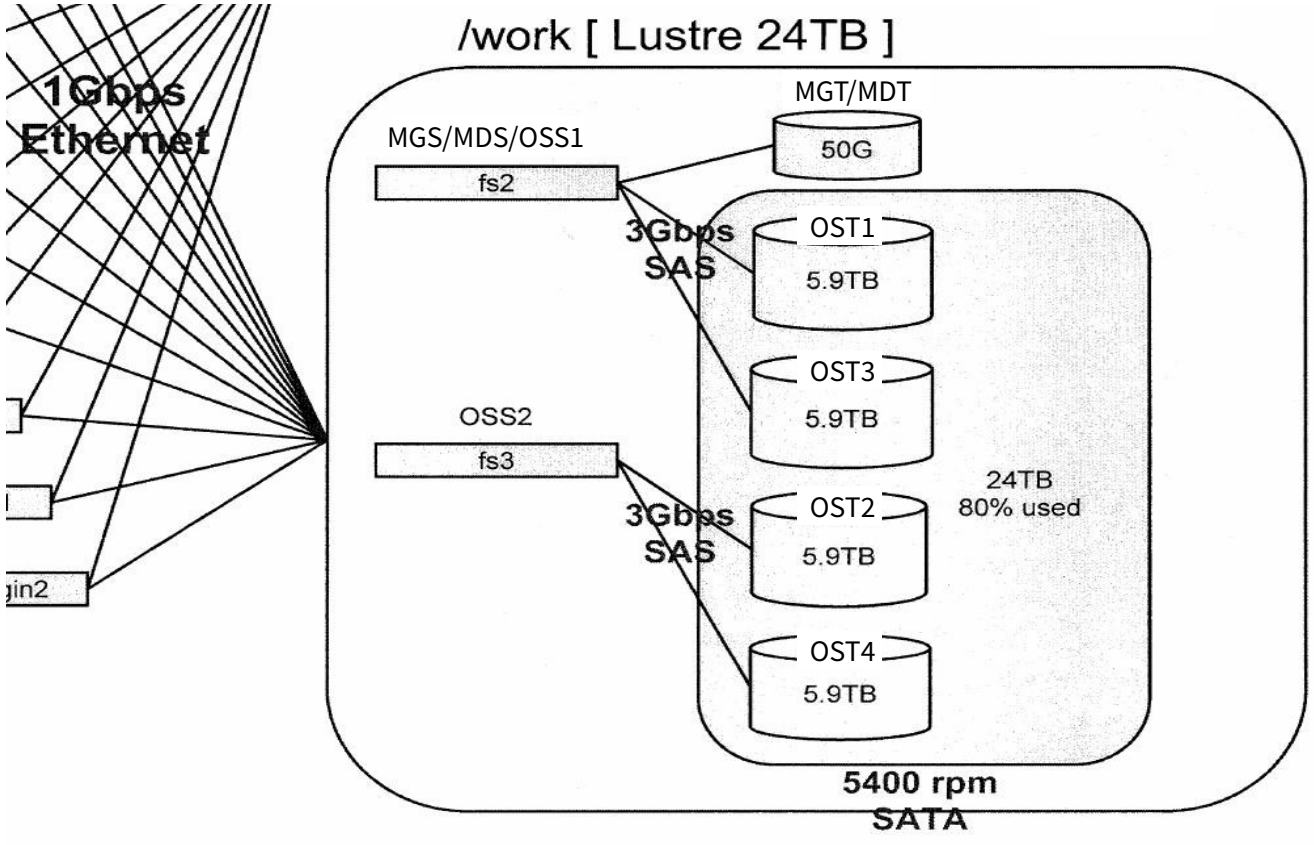
概要

1. イントロダクション

2. <u>OISTにおけるLustreの歴史（2008～）</u>

3. 今後の課題と展望

OISTにおけるLustreの歴史 – cstone (2008~2011)



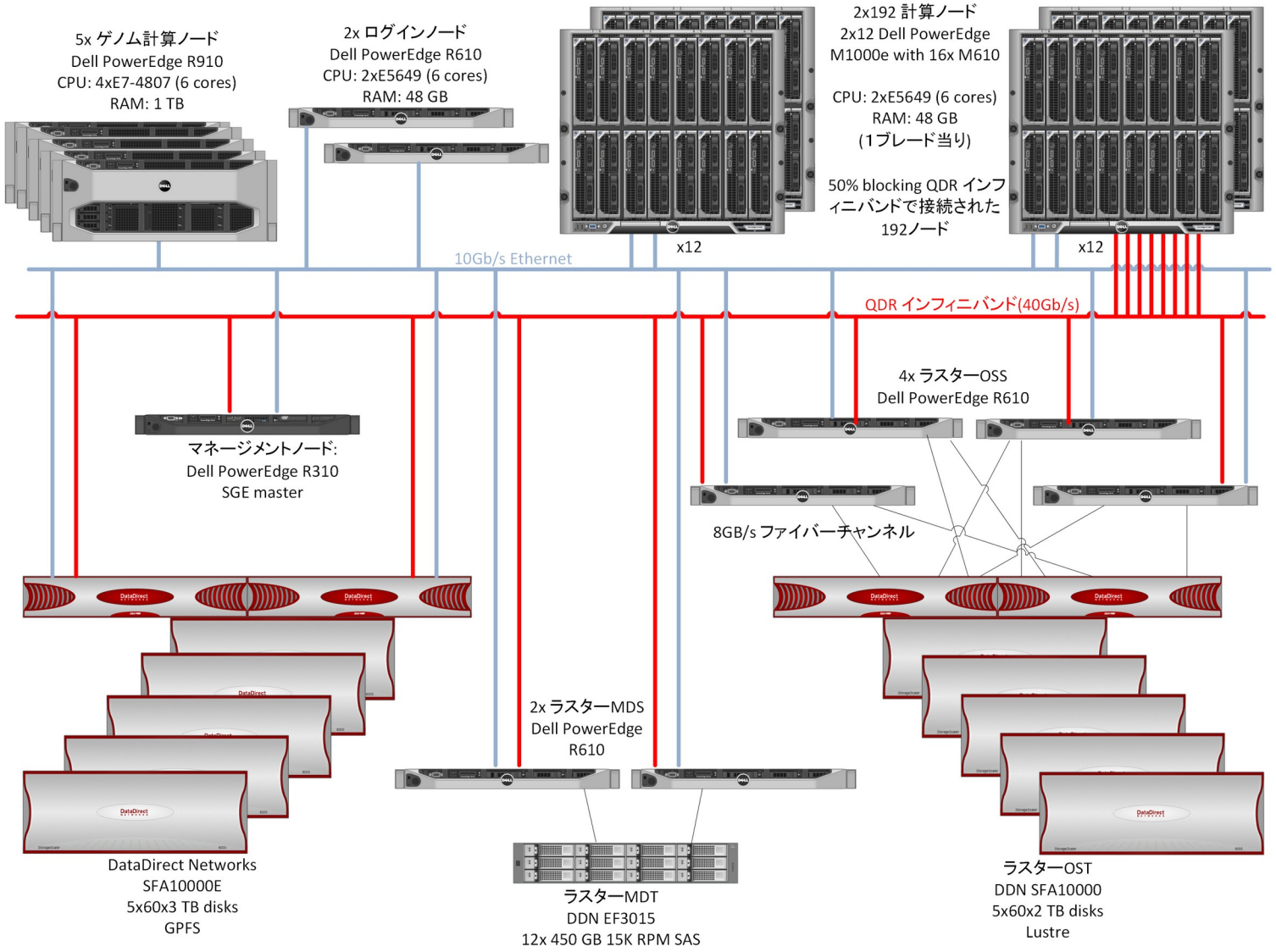
- Lustreクライアント数：144 nodes
- キャパシティ：24TB
- 1 x MDS, 2 x OSS

MGS/MDS/OSS: PowerEdge 2950 III

MGT/MDT/OST: PowerVault MD1000

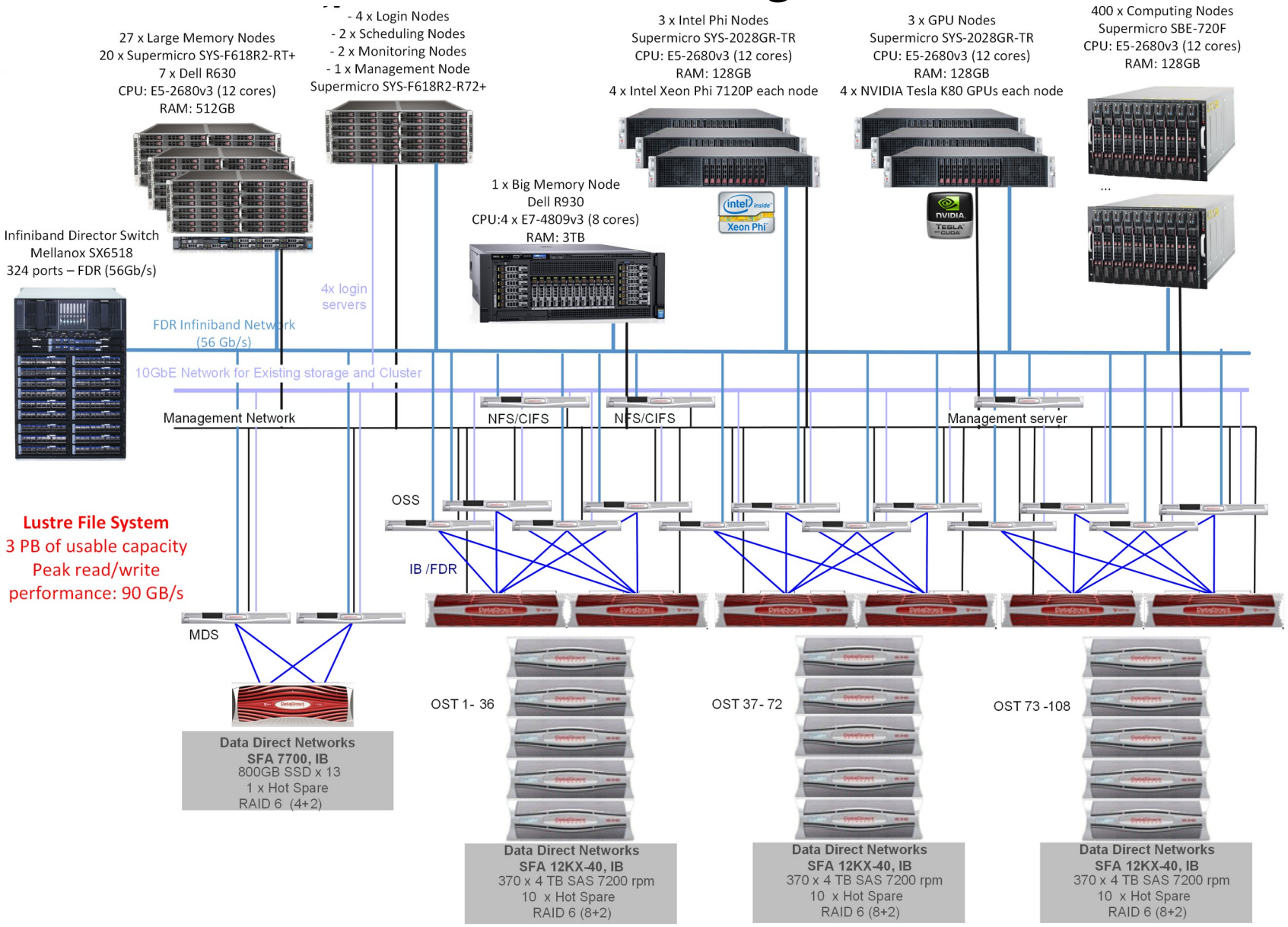
Compute nodes: PowerEdge M1000e with M600

OISTにおけるLustreの歴史 – tombo (2011~2015)



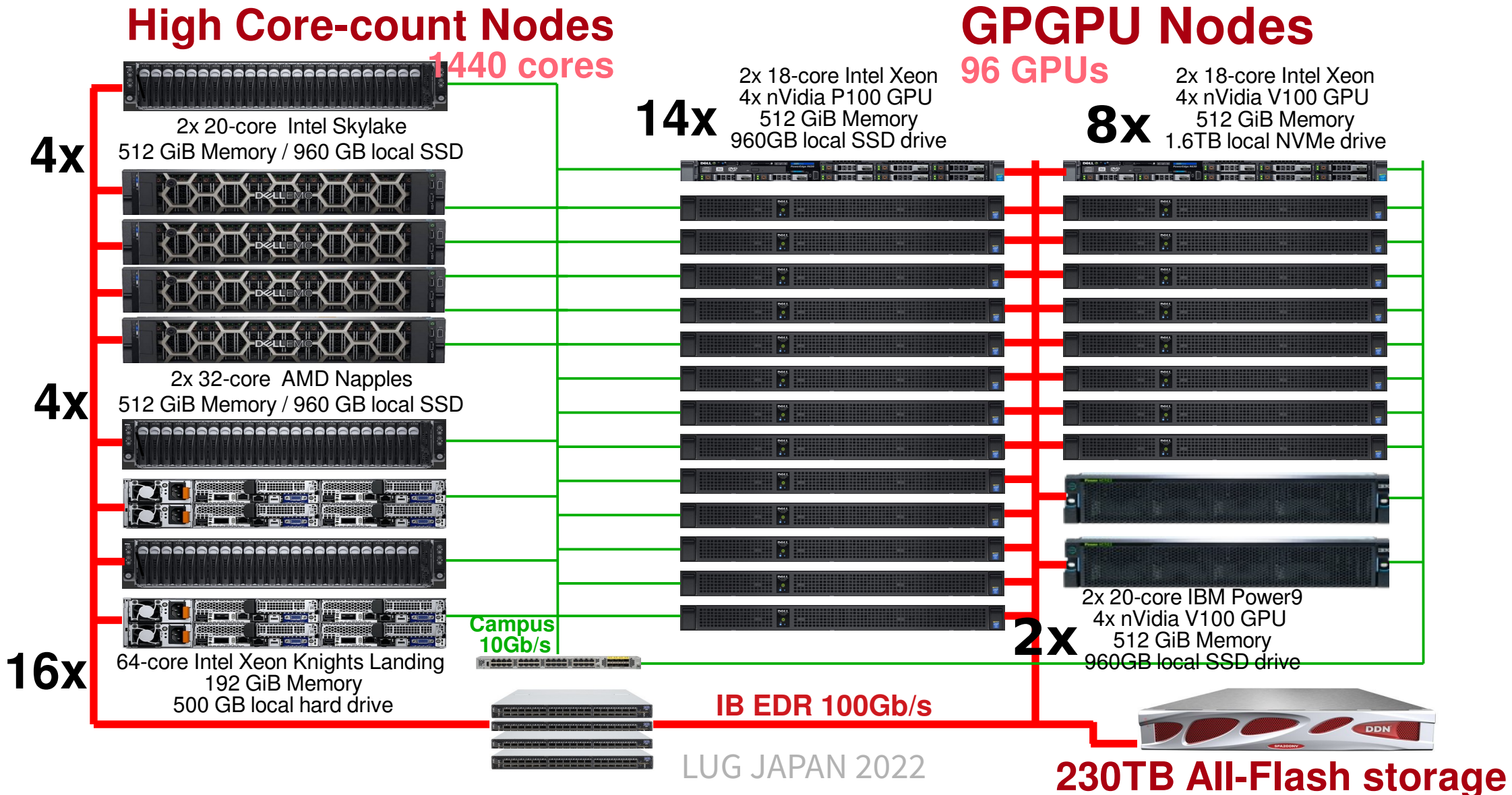
- Lustreクライアント数：391 nodes
- Lustreキャパシティ：500TB
- 2 x MDS(Active, Stand-by), 4 x OSS

OISTにおけるLustreの歴史 – Sango (2015~2020)



- Lustreクライアント数：400+
- キャパシティ：3PB
- 2 xMDS(Active, Stand-by), 12 x OSS
- 2 x NFS/CIFS

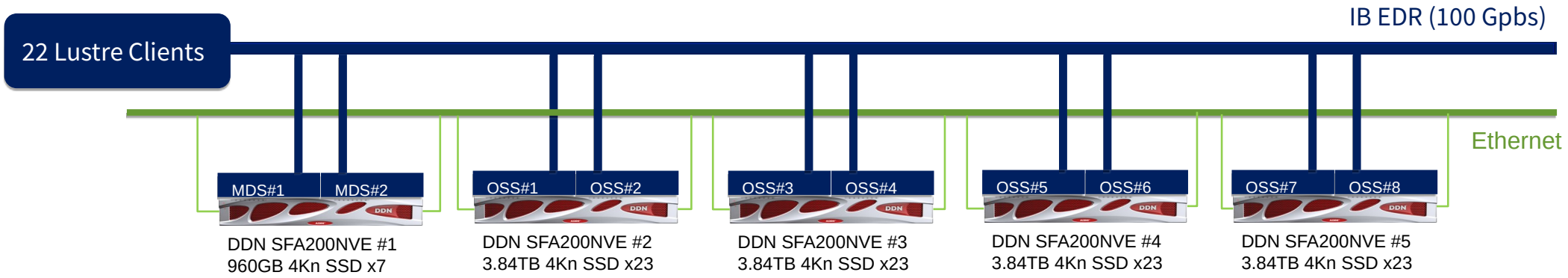
OISTにおけるLustreの歴史 – Saion (2017~現在) (1) 構成



OISTにおけるLustreの歴史 – Saion (2017~現在) (2) ベンチマーク - LUG 19

Benchmark Setup

- Lustre 2.10.5-ddn7
- Capacity: 229.26 TiB
- Number of inodes: ~1.5 billion

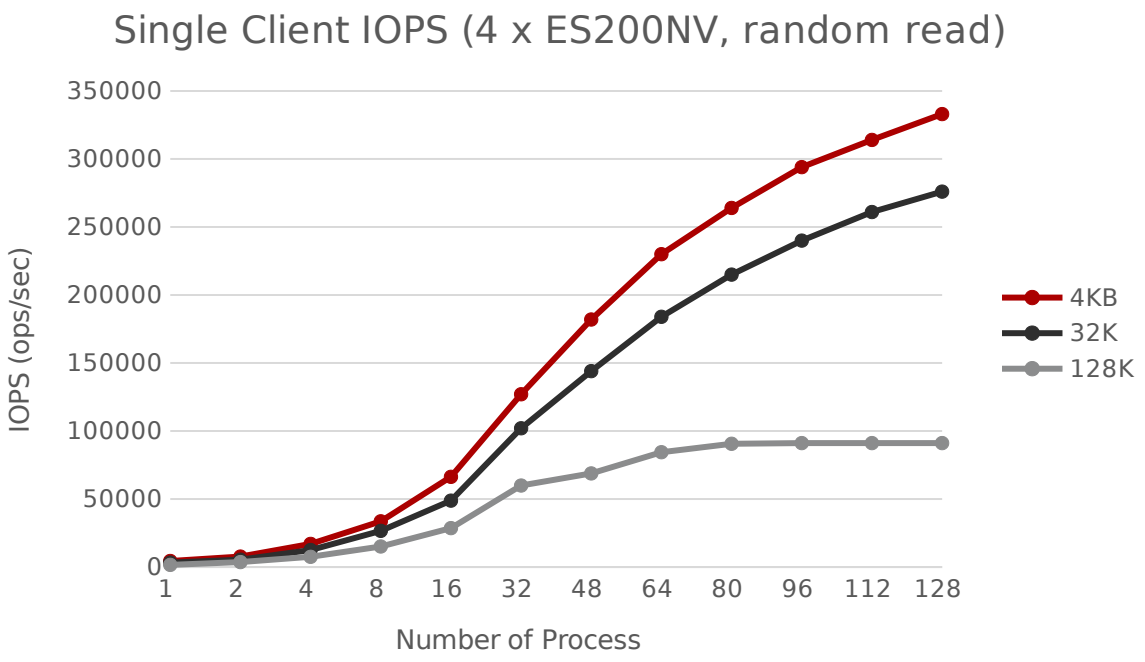


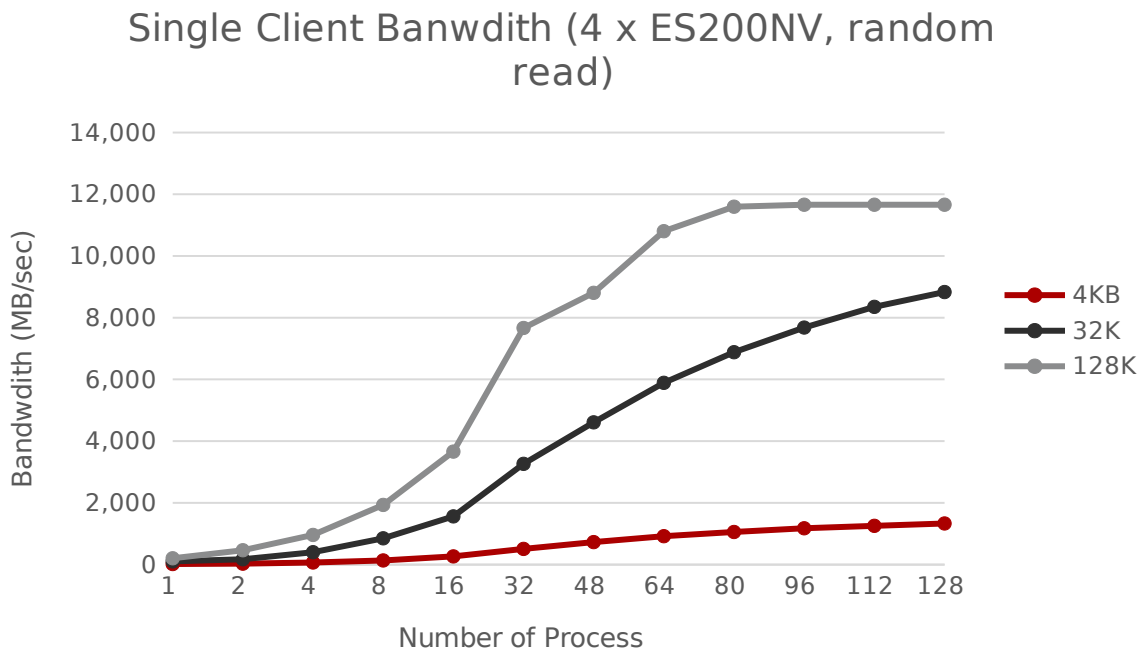
Method

- Use fio with server/client mode
- Create test files(4GB file x processes) first, then random read (4KB, 32KB, and 128KB)
- Minimize I/O cache effects (“sync” and O_DIRECT mode)
- 10 sec run time

OISTにおけるLustreの歴史 – Saion (2017~現在) (3) シングルクライアント性能

Single Client IOPS and Bandwidth (1 x client and 4 x ES200NV)

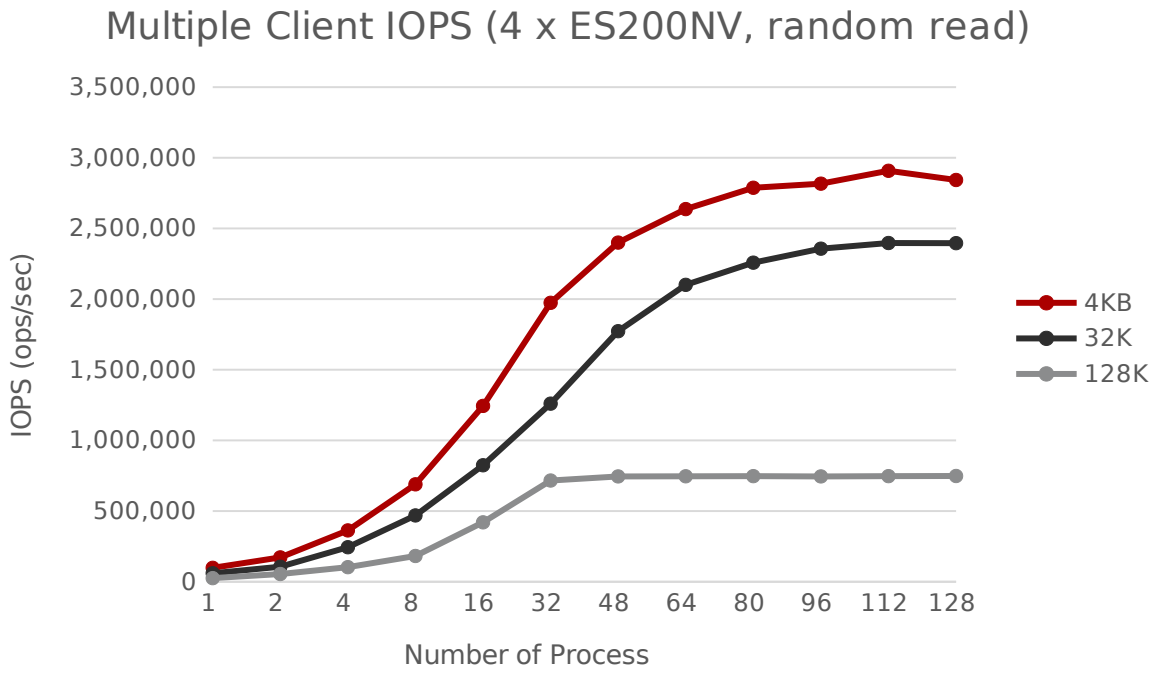


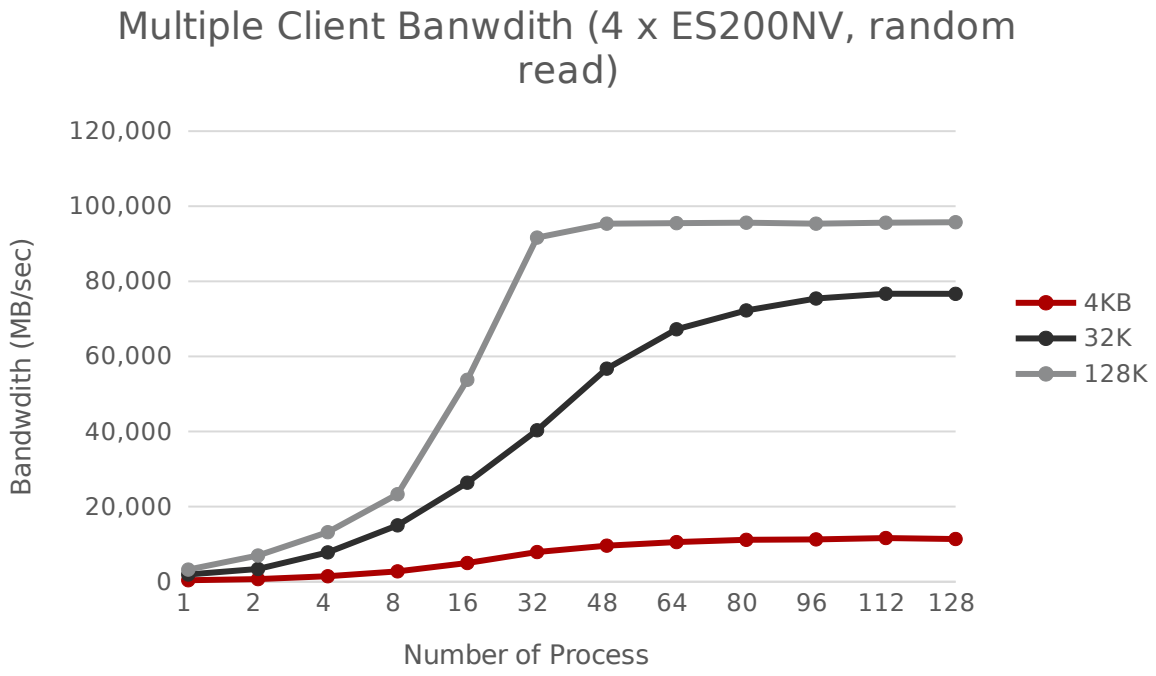


- Achieved 333K IOPS at 4K Random Read and 11.5GB/sec at 128KB Random Read
- 128K Random Read is limited by EDR Bandwidth at Client

OISTにおけるLustreの歴史 – Saion (2017~現在) (4) マルチクライアント性能

Multiple Client IOPS and Bandwidth (22 x client and 4 x ES200NV)





- Achieved 2.84M IOPS at 4K Random Read and 95.7GB/sec at 128KB Random Read
- Delivered balanced IOPS(2.39M IOPS) and Bandwidth(76.9GB/sec) for 32KB Random Read

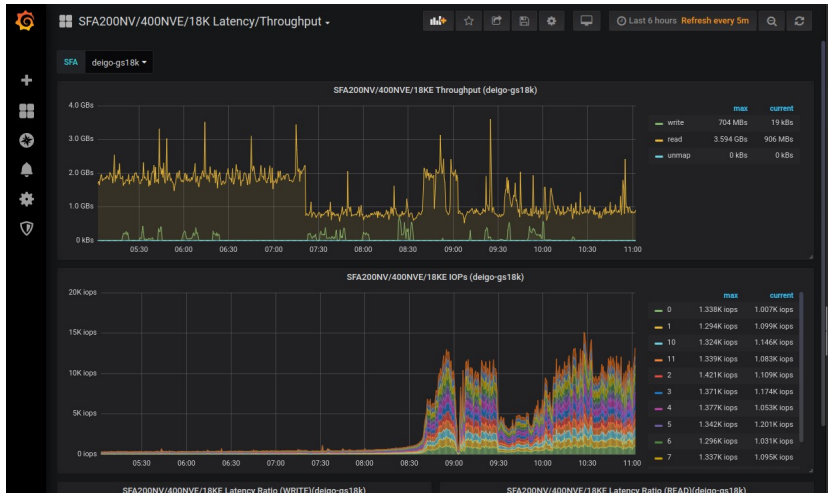
OISTにおけるLustreの歴史 – Deigo (2020~現在) (1)構成

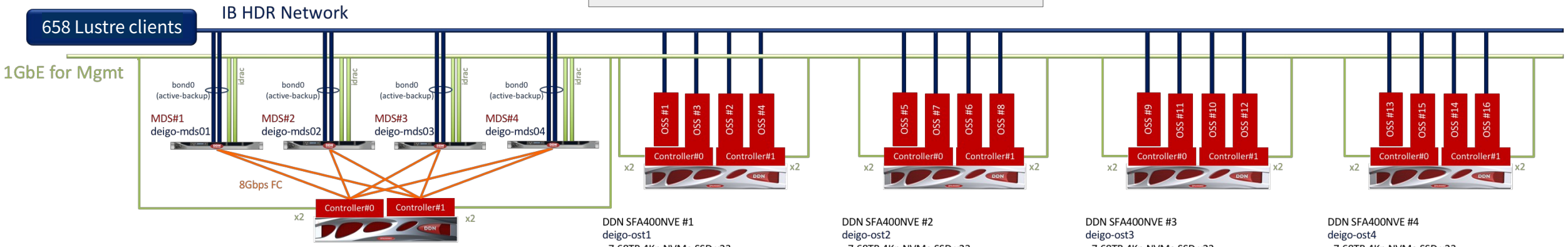
- Lustre 2.12.3_ddn13
- クライアント数：658+ノード
- キャパシティ：500TB
- 4 x MDS, 16 x OSSs
- 8 x LNET routers

Dell EMC PowerEdge C6525/C6420

- 456 x PowerEdge C6525
  - CPU: 2 x AMD EPYC 7702
  - Mem: 512GB (DDR4-3200)
  - Local storage: 894GB (SSD)
- 192 x PowerEdge C6420
  - CPU: 2 x Intel(R) Xeon(R) Gold 6230
  - Mem: 512GB (DDR4-2933)
  - Local storage: 894GB (SSD)

DDN LustrePerfMon





DDN SFA200NV
deigo-mdt
- 1.96TB NVMe SSD x13 (including spare x1)
- 2 xDCR RAID6(4d+2p)
- 2VDs from 1 DCR, Total: 4VDs

DDN SFA400NVE #1
deigo-ost1
- 7.68TB 4Kn NVMe SSD x23 (including spare x1)
- DCR RAID6(8d+2p, MinRebuild 0 w/ 10 disks) x2
- 4VD from 1 DCR RAID, Total: 8VDs.
- 4VM per SFA400NVE as OSS

DDN SFA400NVE #2
deigo-ost2
- 7.68TB 4Kn NVMe SSD x23 (including spare x1)
- DCR RAID6(8d+2p, MinRebuild 0 w/ 10 disks) x2
- 4VD from 1 DCR RAID, Total: 8VDs.
- 4VM per SFA400NVE as OSS

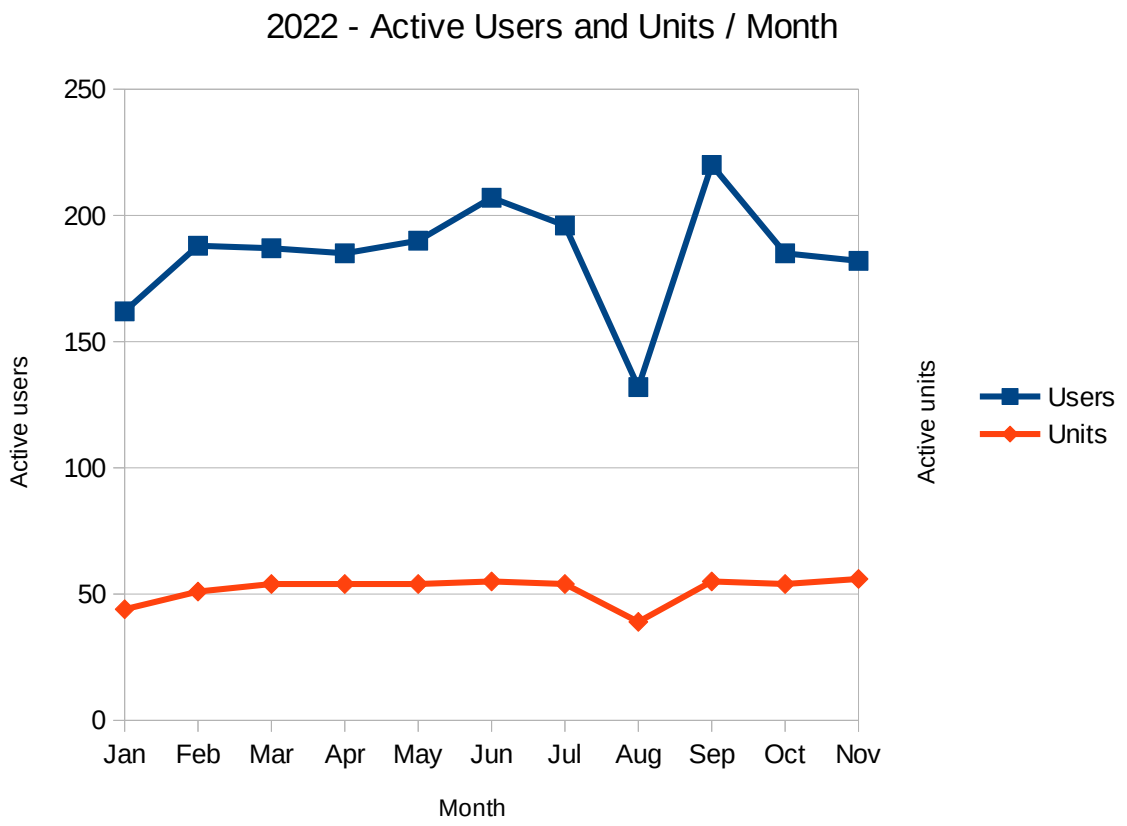
DDN SFA400NVE #3
deigo-ost3
- 7.68TB 4Kn NVMe SSD x23 (including spare x1)
- DCR RAID6(8d+2p, MinRebuild 0 w/ 10 disks) x2
- 4VD from 1 DCR RAID, Total: 8VDs.
- 4VM per SFA400NVE as OSS

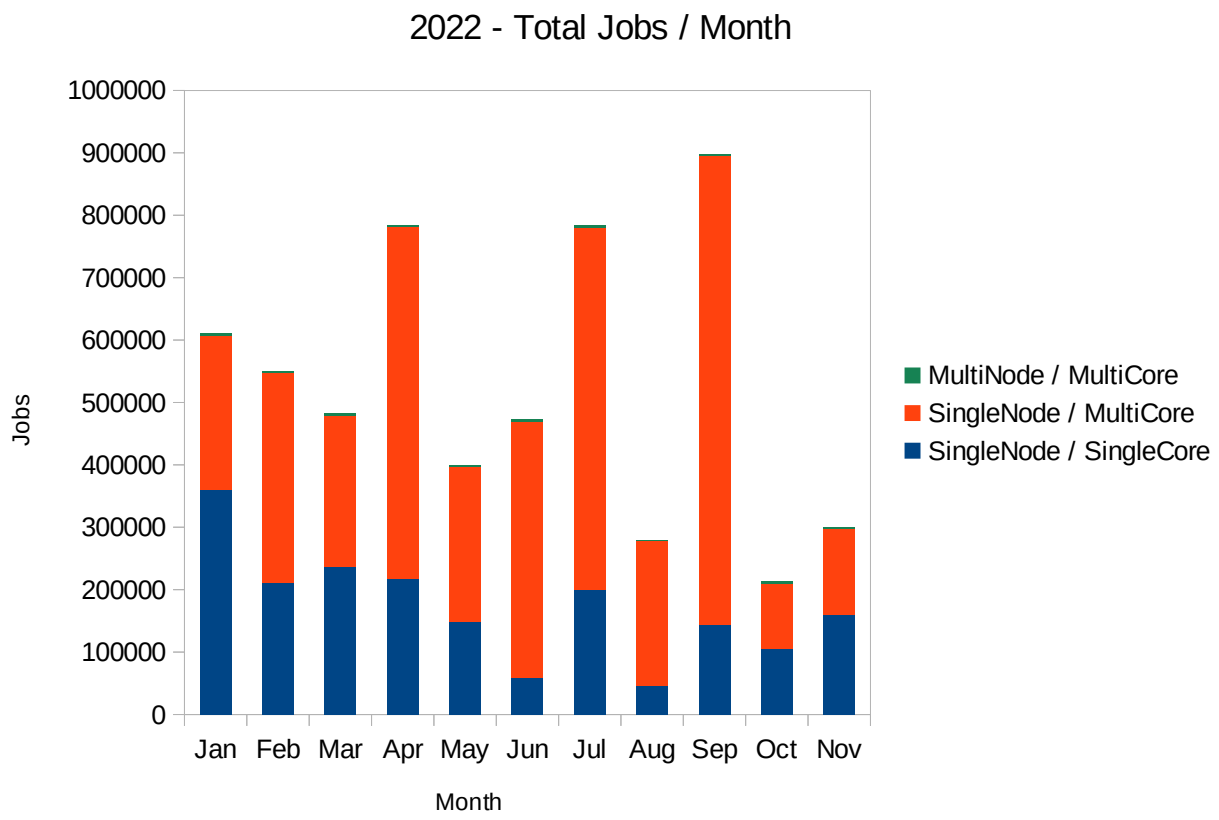
DDN SFA400NVE #4
deigo-ost4
- 7.68TB 4Kn NVMe SSD x23 (including spare x1)
- DCR RAID6(8d+2p, MinRebuild 0 w/ 10 disks) x2
- 4VD from 1 DCR RAID, Total: 8VDs.
- 4VM per SFA400NVE as OSS

OISTにおけるLustreの歴史 – Deigo (2020~現在) (2)ベンチマーク結果

- スループット(IORにて計測)
  - データのバッファサイズ: 1MB
    - Max Write: 152 GB/s
    - Max Read: 189 GB/s
  - データのバッファサイズ: 4KB
    - Max Write: 139 GB/s
    - Max Read: 189 GB/s
- 4KBランダムリード(fioにて計測)
  - 7M IOPs
  - 849MiB/s
- Metadataパフォーマンス(MDTestにて計測)
  - Unique Directory to Four MDSs/MDTs
    - 617,199.696
  - Shared Directory to Single MDS/MDT
    - 111,530.378

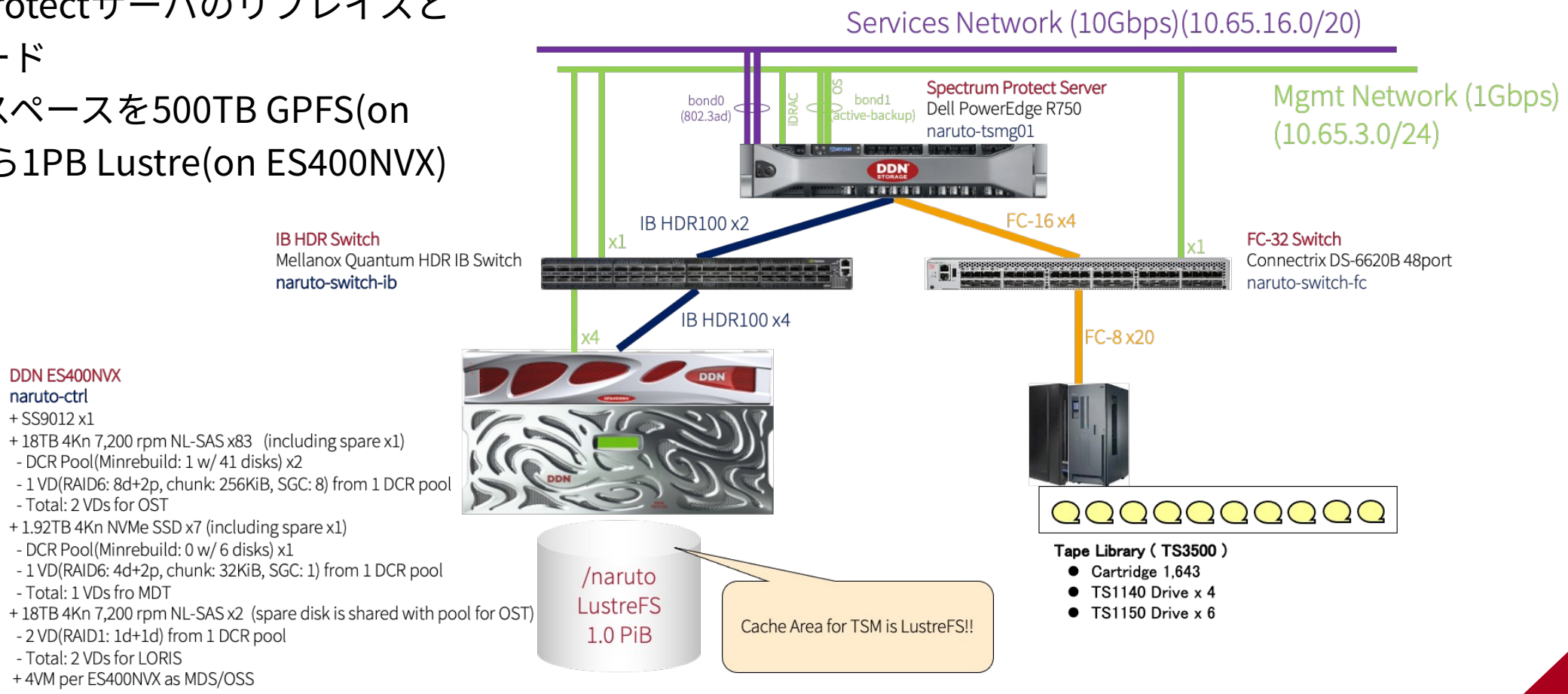
2022年度Deigoの月間アクティブなユーザー数、ユニット数、ジョブ数



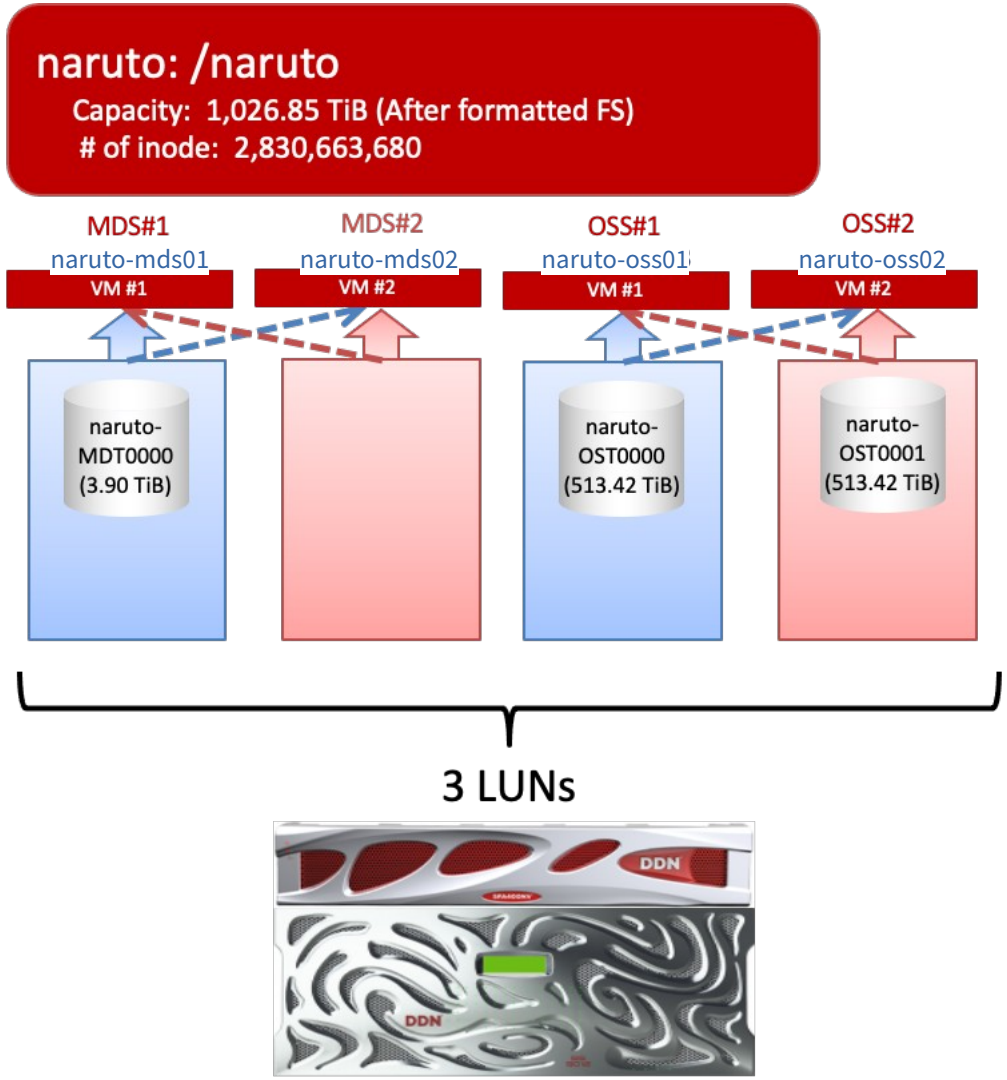


OISTにおけるLustreの歴史 – Naruto (2021~) (1)構成

- Spectrum Protectサーバのリプレイスとアップグレード
- キャッシュスペースを500TB GPFS(on SFA7700)から1PB Lustre(on ES400NVX)へ



OISTにおけるLustreの歴史 – Naruto (2021~) (2)LustreFSの構成



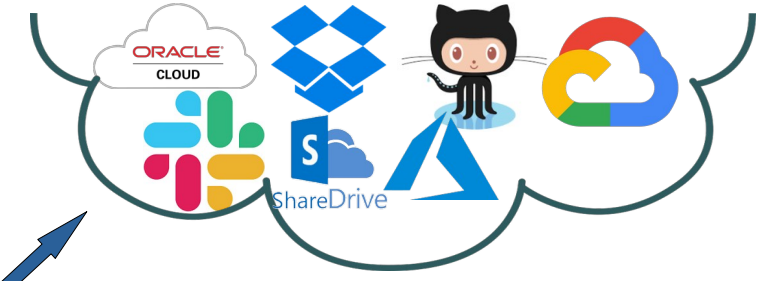
概要

1. イントロダクション

2. OISTにおけるLustreの歴史（2008～）

3. <u>今後の課題と展望</u>

OISTの計算環境



User

HPC

Deigo:/flash

500TB Lustre
10TB/Unit

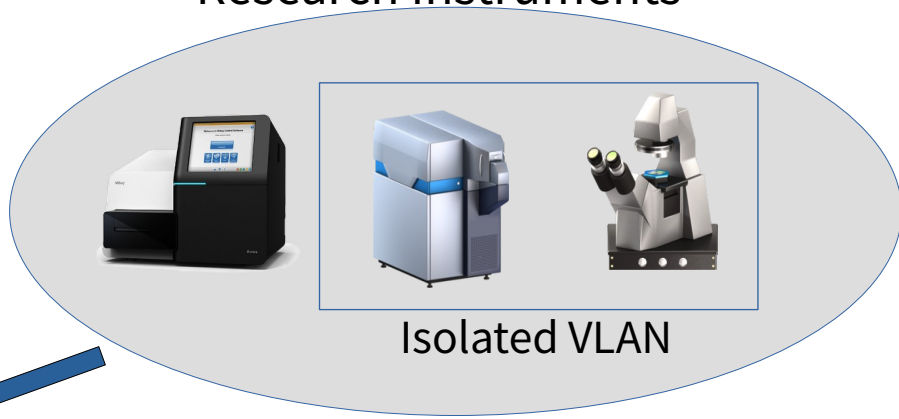
/home

20TB+ OneFS
50GB/User

Saion:/work

230TB Lustre
10TB/Unit

Research Instruments



/hpacquire

Highspeed data transfer
72TB

/bucket

12PB GPFS (SFA7700 + GS18K)
50TB+ / Unit

Naruto:/naruto

13PB tape archive
Incremental backups
Data archiving

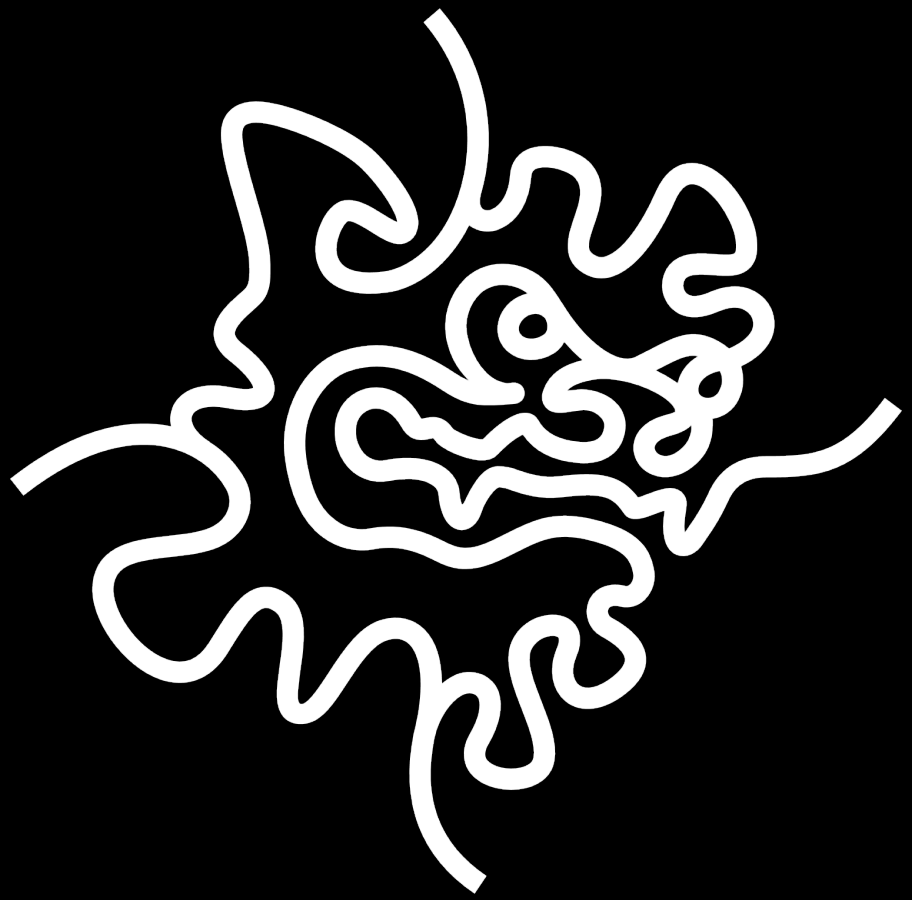
Off-campus (Ginoza IDC)

今後の課題と展望

- Deigo:/flashの拡張とノードの増設
- ポストDeigoに向けたデータセンター
- ポストBucket
  - Lustre, GPFS, or something else?
- HPCのハイブリッド化
  - オンプレ＆クラウド
- データ・マネジメント
  - データ・グラビティ
  - ラストマイル問題
  - ダーク・データ

注目しているLustreの機能

- シングルスレッド性能、シングルクライアント性能
  - Persistent Client Cache (2.13)
  - シングルストリームDirect I/O性能の向上 (2.15)
  - LustreクライアントWBC (2.16+)
- メタデータ性能
  - DNE auto-striping (2.14)
  - DNE MDT Space Balance (2.15+)
  - Lustre Metadata Redundancy (2.17+)



THANK YOU!