Lustre on OCTOPUS

Susumu Date
Applied Information Systems Research Division,
Cybermedia Center, Osaka University, Japan
大阪大学　サイバーメディアセンター
応用情報システム研究部門　伊達　進

Cybermedia Center
Osaka University

Introduction

Osaka University selected as a Designated National University on Oct.23 2018

Under the motto "Live Locally, Grow Globally,"
Osaka University grapples with the challenges of education and research

http://www.osaka-u.ac.jp/en/index.html

Organization

- 11 Schools with 10 Corresponding Graduate Schools
- 6 Independent Graduate Schools
- 6 Research Institutes
- 2 National Joint-Use Facilities
- 14 Joint -Use Education and Research Facilities
- University Library
- 2 University Hospitals
- World Premier International Research Center (WPI) Initiative
- Institute for Academic Initiatives
- Center for Education in Liberal Arts and Sciences

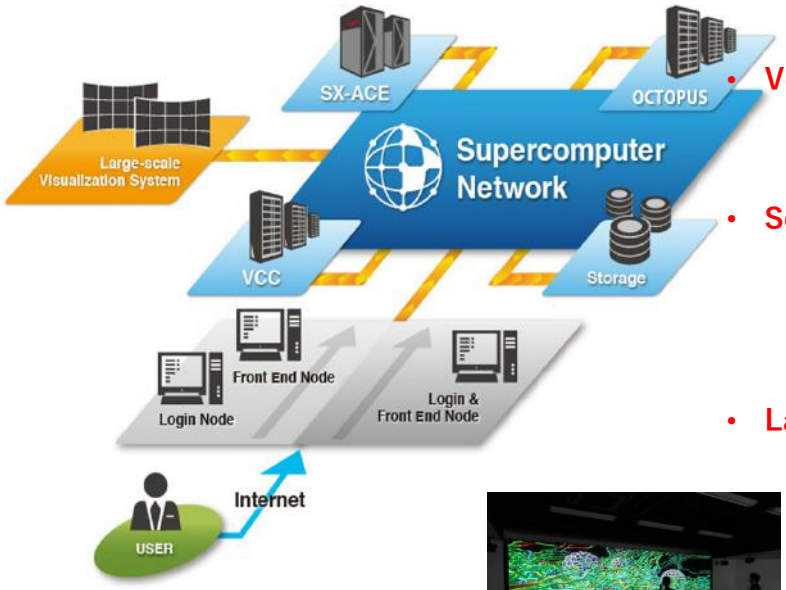
Faculty and staff 6,654

Students

- Undergraduate 15,250
- Graduate 8,054

Cybermedia Center

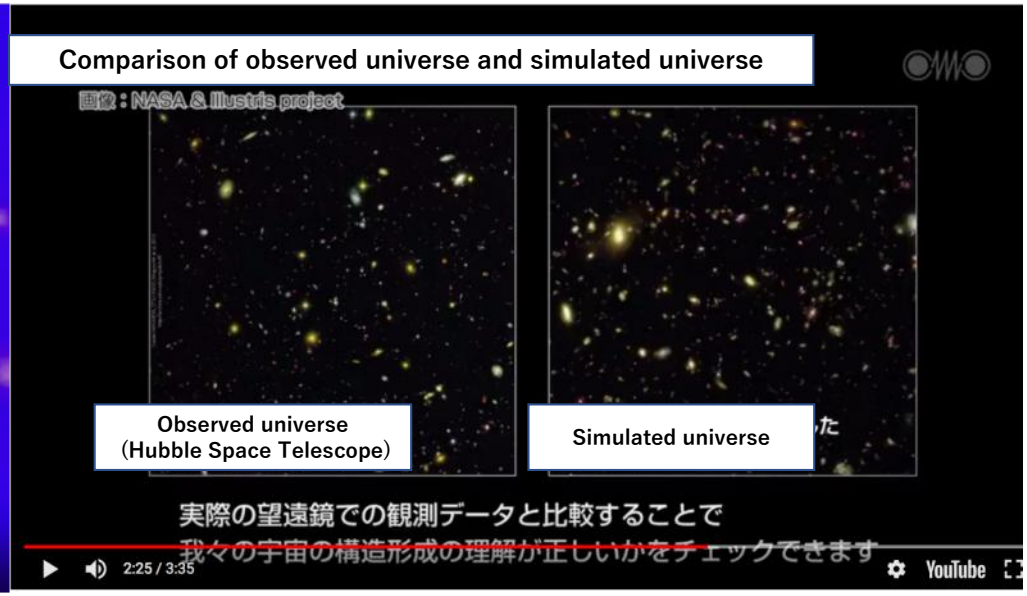
- a supercomputing center at Osaka University
  - *http://www.cmc.osaka-u.ac.jp/*
- has a responsibility of providing a powerful high-performance computing environment for university researchers across Japan as a national joint-use facilities.



- **Vector-typed Supercomputer**
  - SX-ACE
- **Scalar-typed Supercomputer**
  - PC cluster system for large-scale visualization (VCC)
  - Osaka university Cybermedia cenTer Over-Petascale Universal Supercomputer
- **Large-scale Visualization System**
  - 24-screen Flat Stereo Visualization System
  - 15-screen Cylindrical Stereo Visualization System

Application example from High-performance Scientific News

Cosmology with Numerical Simulations
Prof. Kentaro Nagamine
Graduate School of Science, Osaka Univ.

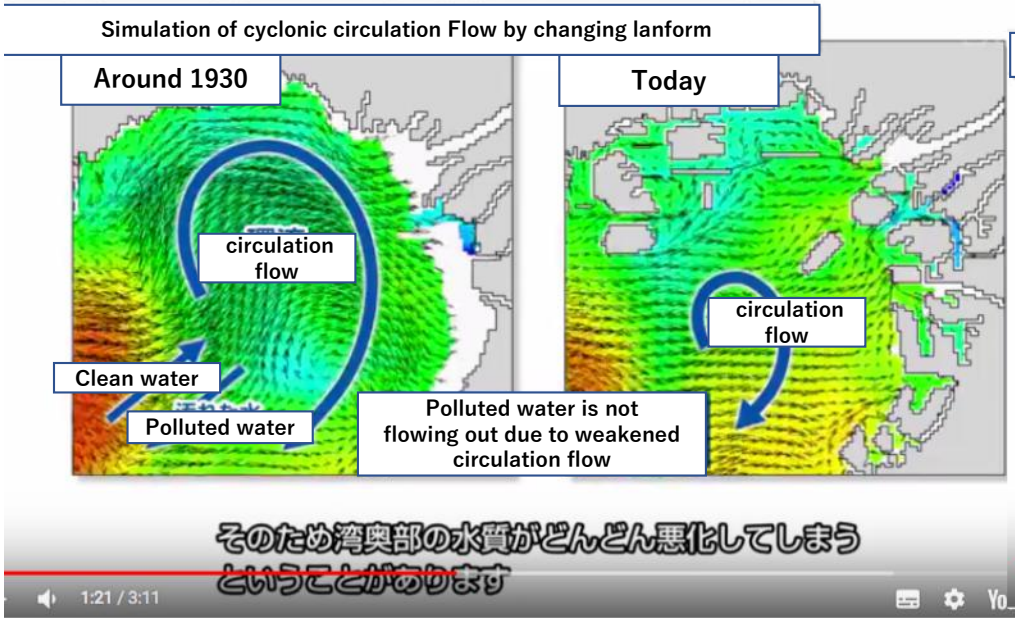


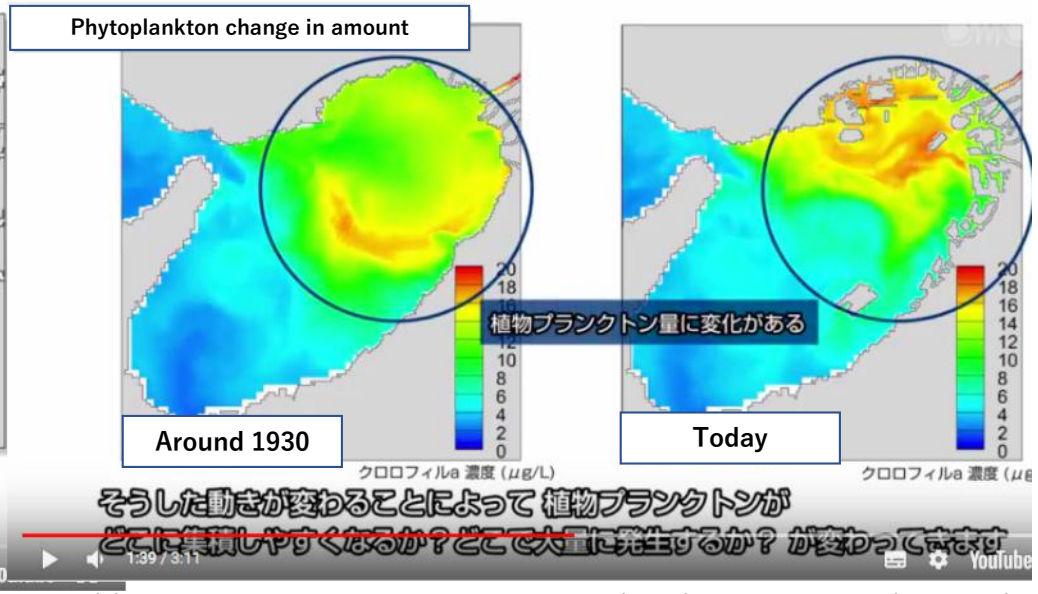
HPSC vol.2 http://www.hpc.cmc.osaka-u.ac.jp/en/hpsc-news/vol02/

Application example from High-performance Scientific News

Cybermedia Center
Osaka University

Numerical simulation of flow and water quality in Osaka Bay
Assit. Prof. Yusuke Nakatani
Graduate School of Engineering, Osaka Univ.





HPSC vol.3 http://www.hpc.cmc.osaka-u.ac.jp/en/hpsc-news/vol03/

OCTOPUS

ペタフロップス級ハイブリッド型スーパーコンピュータ

Cybermedia Center
Osaka University

1 全国の研究者が利用可能

2 多様な計算ニーズへの対応

3 ペタフロップス級大規模計算能力

4 安定した動作環境の提供

OCTOPUS since Dec. 2017

Cybermedia Center
Osaka University

- is short for Osaka university Cybermedia cenTer Over-Petascale Universal Supercomputer.
- We were really eager to obtain a Peta-Flops system.
- The CMC' system is difficult to remember. Not fun!
  - Vcc, Ace, Hcc.. , PCC.
- Many Japanese people associate TAKOYAKI with Osaka.

Takoyaki (たこ焼き *or* 蛸焼) is a ball-shaped Japanese snack made of a wheat flour-based batter and cooked in a special molded pan. It is typically filled with minced or diced octopus (*tako*), tempura scraps (*tenkasu*), pickled ginger, and green onion.[1][2] Takoyaki are brushed with takoyaki sauce (similar to Worcestershire sauce) and mayonnaise, and then sprinkled with green laver (*aonori*) and shavings of dried bonito. There are many variations to the takoyaki recipe, for example, ponzu (soy sauce with *dashi* and citrus vinegar), goma-dare (sesame-and-vinegar sauce) or vinegared dashi.
Yaki is derived from "*yaku*" (焼く) which is one of the cooking methods in Japanese cuisine, meaning "to fry or grill", and can be found in the names of other Japanese cuisine items such as *okonomiyaki* and *ikayaki* (other famous Osakan dishes) *from wikipedia*

OCTOPUS 概要

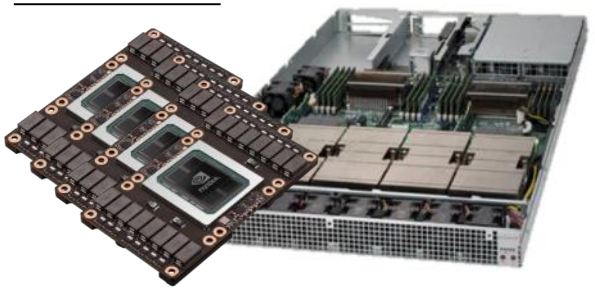
CPU nodes

NEC LX 2U Twin2 Server 406 Rh-2 (59: 236nodes)

Storage

DDN EXAScaler (3.1PB)

GPU nodes



NEC LX 1U 4GPU Server 102Rh-1G (37: 37 nodes)

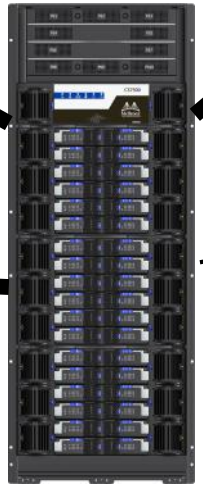
Many-core nodes

NEC Express5800/HR110c-M (11: 44nodes)

Large-scale shared memory nodes

NEC LX 116Rg-7 (2: 2nodes)

NEC LX 4U-GPU server 108Th-4G (4 node)



Mellanox CS7500
648 port EDR InfiniBand Director Switch

1.46 PFlops

OCTOPUS システム構成

汎用 CPUノード	合計：236ノード（471.24 TFlops）
CPU	Intel Xeon Gold 6126 (Skylake / 2.6 Ghz, 12 cores) x 2
Memory	192 GB

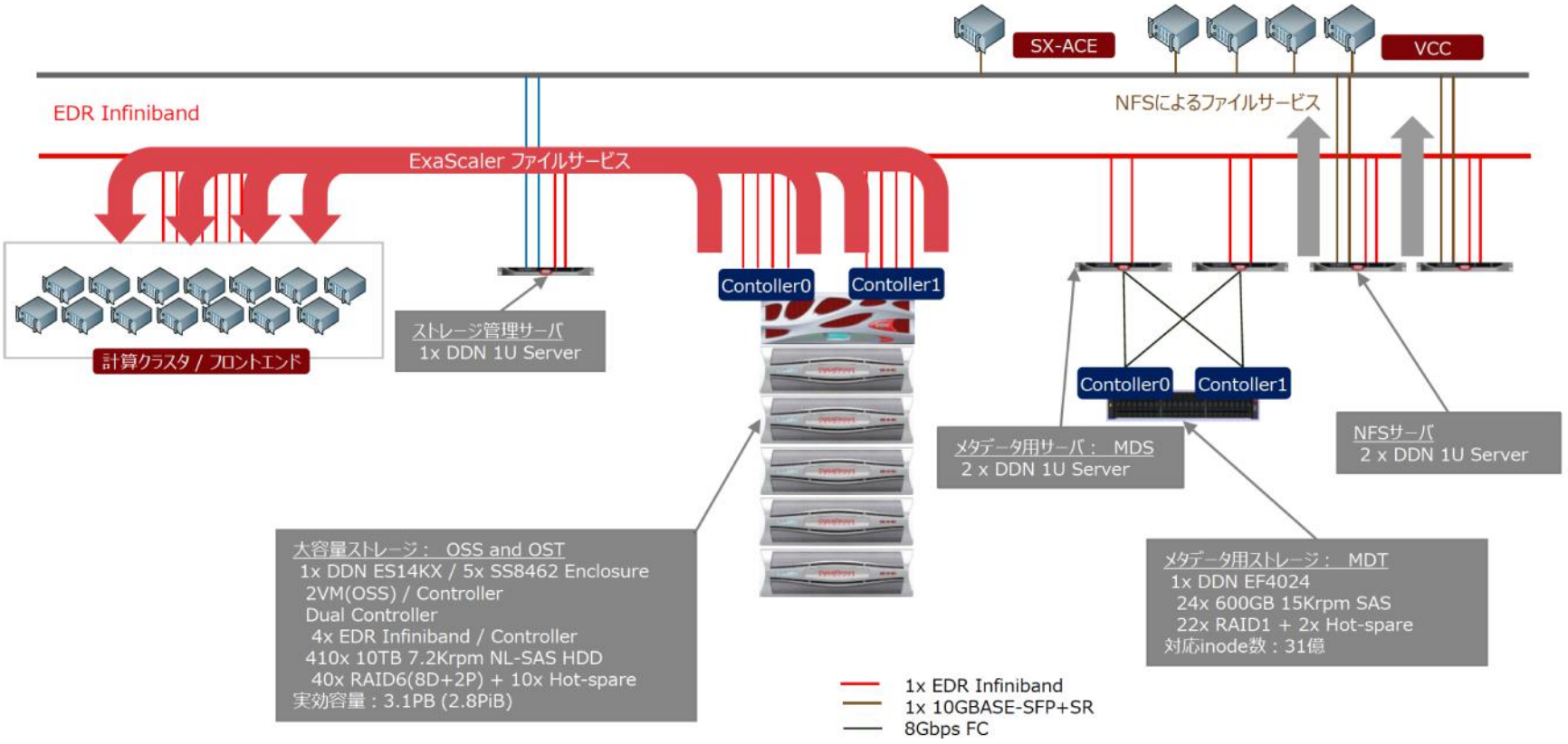
GPUノード	合計： 37ノード（858.28 TFlops）
CPU	Intel Xeon Gold 6126 (Skylake / 2.6 Ghz, 12 cores) x 2
Memory	192 GB
Accelerator	NVIDIA Tesla P100 (NVLINK) x 4

メニーコアノード	合計：44ノード（117.14 TFlops）
CPU	Intel Xeon Phi 7210 (KNL / 1.3 Ghz, 64 cores)
Memory	192 GB

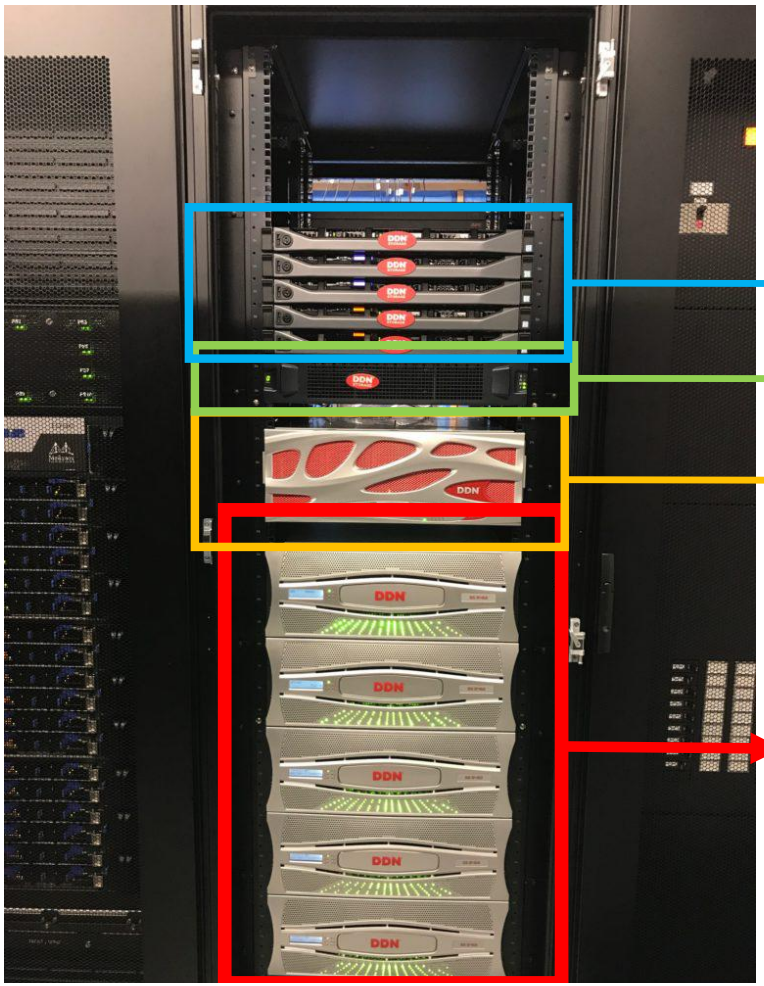
大容量主記憶搭載ノード	合計： 2ノード（15.38 TFlops）
CPU	Intel Xeon Platinum 8153 (Skylake / 2.0 Ghz, 16 cores) x8
Memory	6TB

大容量ストレージ	合計： 3.1PB
File System	DDN EXAScaler(Lustre)
Size	3.1 PB

OCTOPUS Storage/File server



OCTOPUS Storage/File server



DDN製ExaScaler

- Lustreファイルシステムベース
- 1ラックに3PB収容

管理サーバ(1台)
NFSサーバ(2台)
MDS (Meta Data Server) (2台)

DDN EF4024 (Meta Data Target)

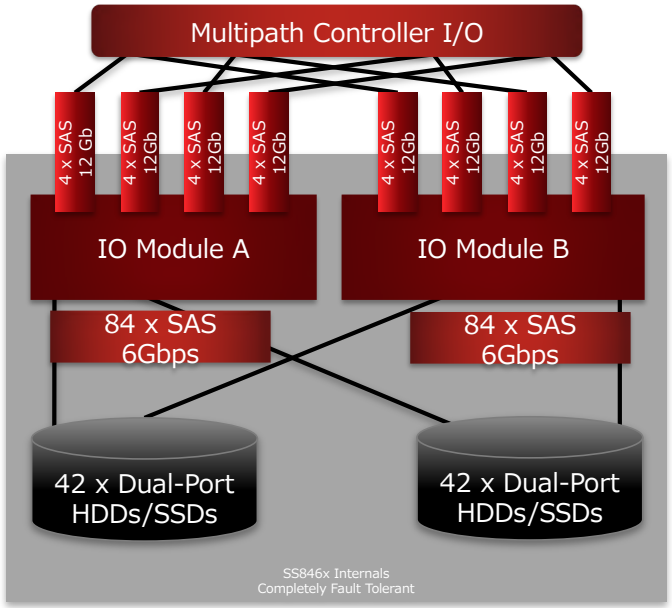
DDN ES14KX
RAIDコントローラ
+OSS (Object Server Storage)

DDN SS8462
ストレージ
エンクロージャ
(OST: Object Storage Target)

DDN SS8462: OST on OCTOPUS

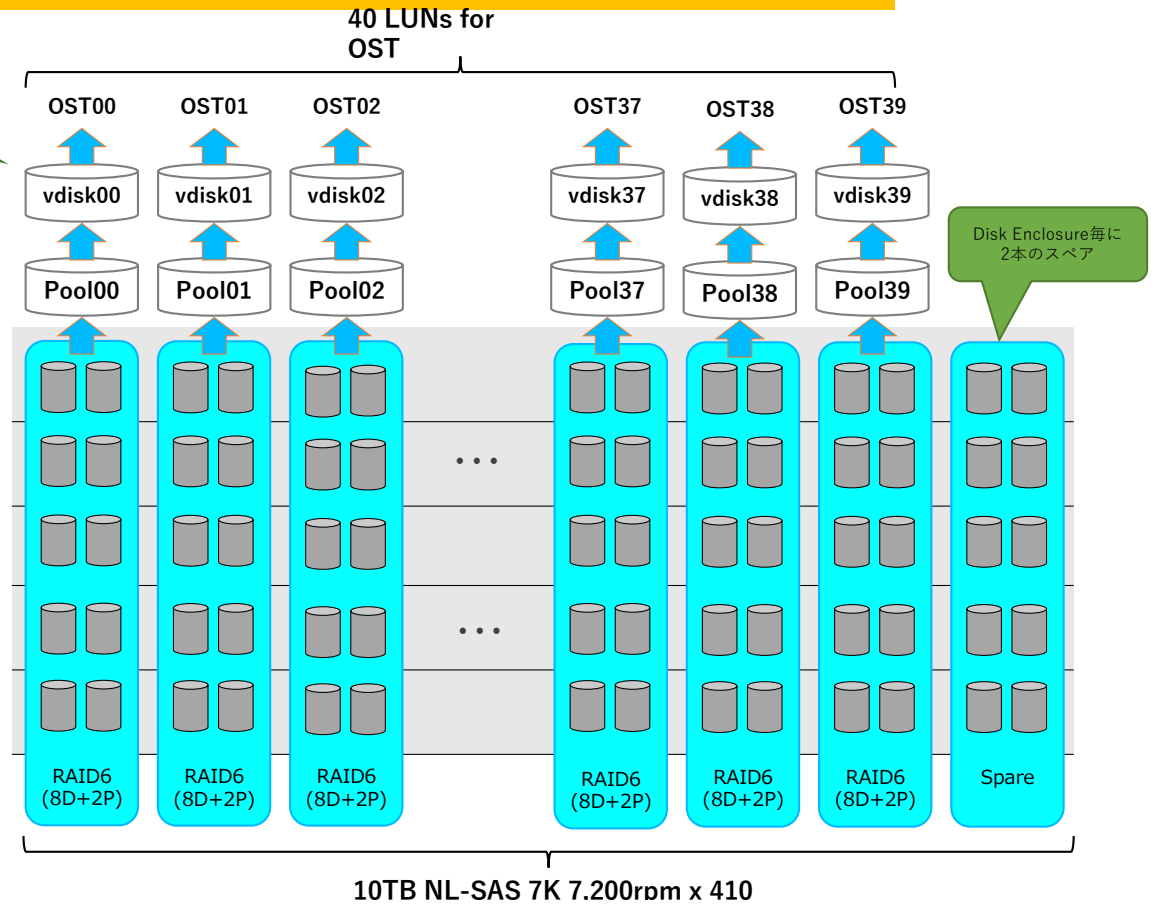
- 4U筐体に最大で84本のドライブを収容する高密度なストレージエンクロージャ
- 全てのコンポーネントが冗長化

OCTOPUSでは、10 TB NL-SAS(7.2krpm) HDD ✖ **82個** を１ラックに搭載

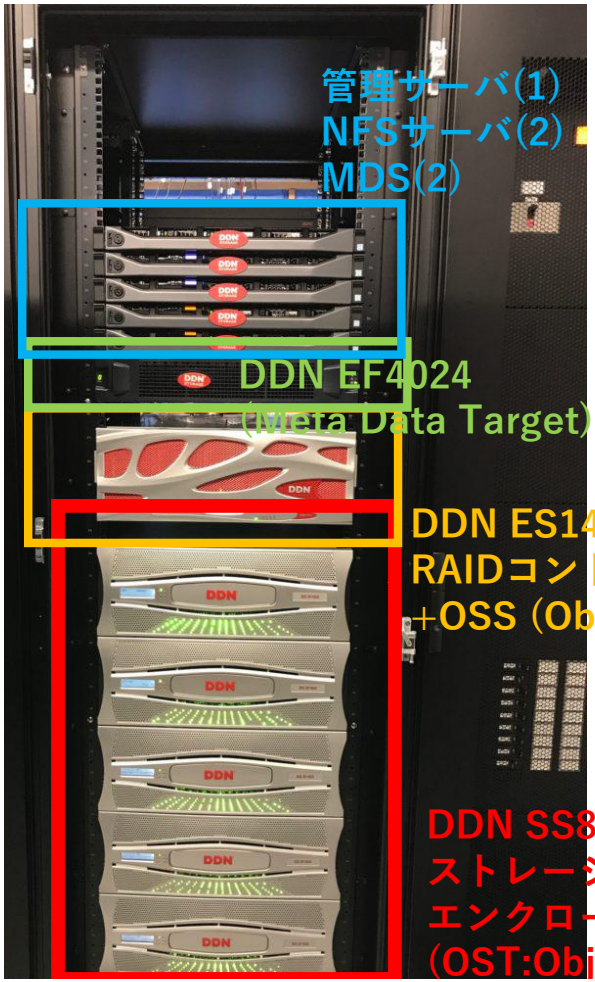


OCTOPUS　ストレージ構成

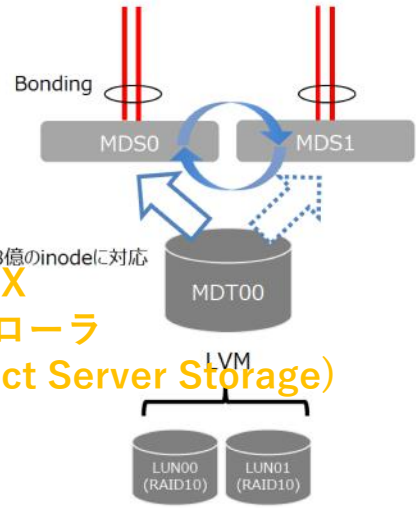
- 5エンクロージャにまたがり、10ドライブごとにLUNを構成
- LUN内でRAID6(8D+2P)を構成



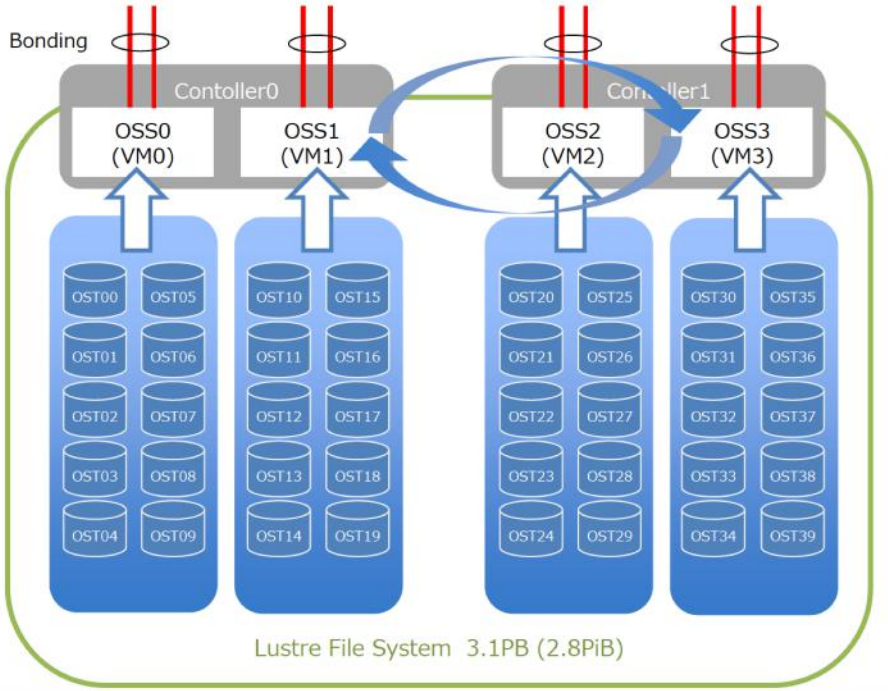
OCTOPUS: ファイルシステム構成



- MDSサーバ: 2サーバでActive/Stanby冗長構成



- ファイルサーバ：RAIDコントローラごとに2VMを配備し、合計4VMでActive/Active構成



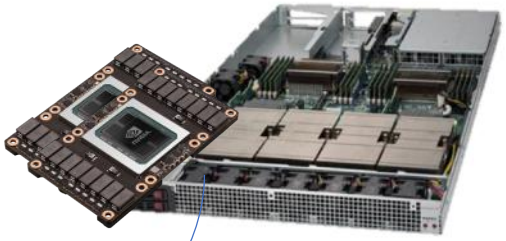
Cybermedia Center
Osaka University

Lustreの運用 on OCTOPUS

OCTOPUS file system (1)

- /octfs: mount point as a single file system
  - Lutre storage is accessed at /octfs/
    from all computing nodes and frontend nodes.

Large Memory nodes



GPU nodes

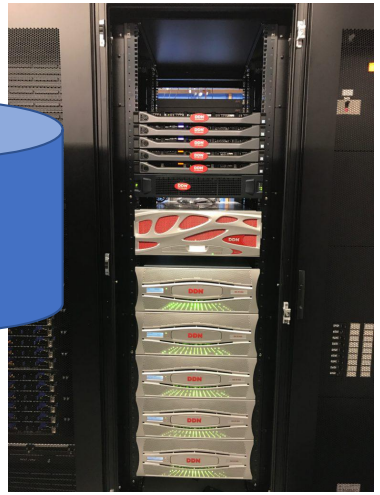
CPU nodes

KNL nodes

Frontend nodes

/octfs

3.1PB
Lustre



OCTOPUS file system (2)

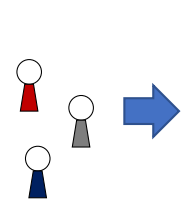
- Design goal: Quota design/configuration **well-suitable for our user management method and storage charging policy**.

Storage-Charging Policy
- 1 TB disk space is allocated to each application request (ディスク容量は1申請単位で1TBを割り当てる。)
- 1 TB disk space can be added per 3000 JPY/year. (年間3000円で1 TBの追加ストレージを割り当て可能。)

How can we use Lustre quota functions?
(どのようにクオータによってストレージ容量を制限すべきか？)

User management

- Application request is done by an individual or group. (申請は個人あるいはグループ単位)
- Currently, different user-ids are issued to a user in the case that the user belongs to multiple application requests. (現在は、同一ユーザであっても 申請グループごとにアカウント発行)
- In future, the same account for a user can be shared among different groups. (将来的に、1ユーザが複数のグループに所属できるようにしたい)



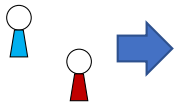
PI:
Mr. Incredible
Members:
Elastigirl
Violet

Mr. Incredible, Elastigirl, and Violet share 1 TB storage.

Application request form

PI:
Dash

Dash can use 1TB storage.

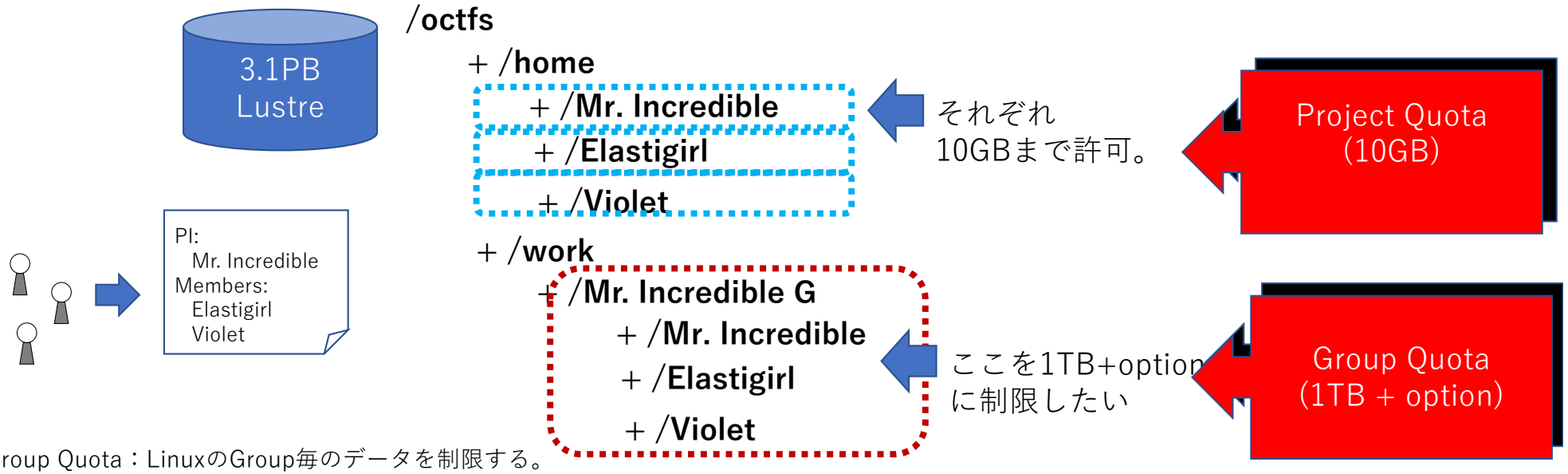


PI:
Syndrome
Members:
Mr. Incredible

Often, some users appear on the different application form. (時折、ユーザが複数の申請書に関係)

Partition and Quota Configuration on OCTOPUS

- Partitioning (パーティション設定)
  - ユーザhome領域: */octfs/home/(user id)/*
  - グループwork領域： */octfs/work/(group)/(user id) /*
- Quota Configuration(クオータ設定)



※Group Quota：LinuxのGroup毎のデータを制限する。
ファイルシステム単位で1つだけ設定可能。
※Project Quota：特定のディレクトリ内のデータ容量を制限する。
ファイルシステム内に、ディレクトリ単位で複数設定可能

Cybermedia Center
Osaka University

ファイルアクセス性能 on OCTOPUS

比較対象システム：VCC 2013年度導入

http://www.hpc.cmc.osaka-u.ac.jp/vcc-sys/

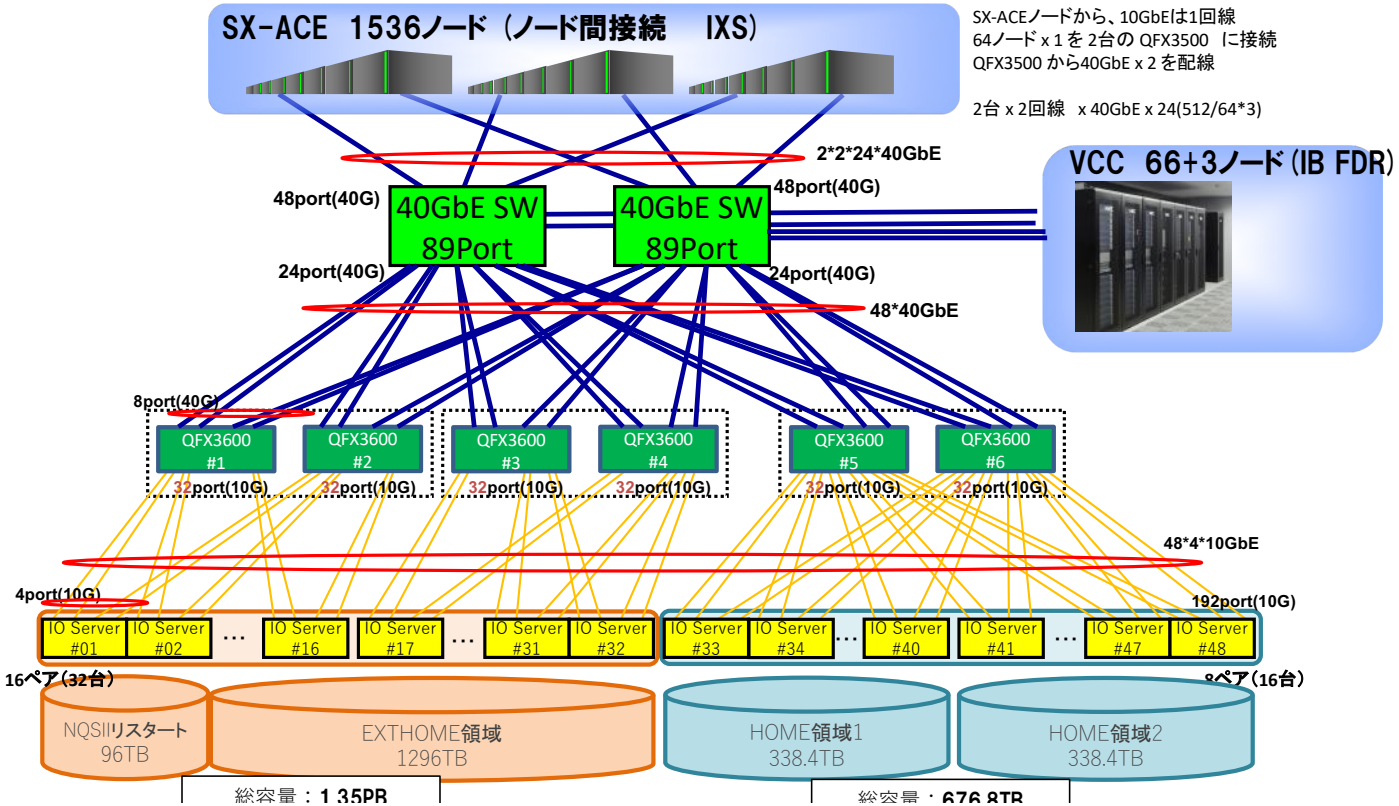
- Dynamically Reconfigurable Cluster System
  using ExpEther, system virtualization technology from NEC

CPUノード		**合計：66ノード**
CPU	Intel Xeon E5-2670v2(Ivy Bridge / 2.5 Ghz, 10 cores) x 2	
Memory	64GB	
増設ノード		**合計： 3ノード**
CPU	Intel Xeon E5-2690v4 (Broadwell / 2.5 Ghz, 14 cores) x 2	
Memory	64 GB	
再構成可能資源		
Accelerator	NVIDIA Tesla K20	59基
Flush drive	ioDrive2 (365 GB)	4個
Storage	PCIe SAS (36TB)	9個
大容量ストレージ		
File system	NEC ScaTeFS (Scalable Technology File System)	
Size	2 PB	

NEC ScaTeFS for SX-ACE & VCC

- 複数のI/Oサーバにメタデータおよびデータを均等に分散させることで、高速な並列I/Oと大量のファイル操作に強いファイルシステムを実現するだけでなく、単一障害点の排除と自動リカバリによる高い耐障害性とデータ保護も同時に実現した高信頼性を確立したファイルシステム

NEC ScateFSで2PBのストレージをSX-ACE、VCCに提供



設置図

Cybermedia Center
Osaka University

ScaTeFS (2PB) for SX-ACE and VCC

DDN Lustre (3PB) for OCTOPUS

IORによる性能評価

- 2018年10月18日(運用中:**中負荷時**)にスケジューラNQS II よりジョブ投入して計測
  - *ior –w -r –b 8g –t m –C –Q 21-e –vv –F –o output*

			OCTOPUS	VCC
IOR 2.10.3	32 nodes 640 processes	write	30,405 MiB/sec	6,959 MiB/sec
		read	30,392 MiB/sec	10,224 MiB/sec
	16 nodes, 320 processes	write	28,995 MiB/sec	9,067 MiB/sec
		read	29,631 MiB/sec	6,805 MiB/sec
	8 nodes, 150 processes	write	26,443 MiB/sec	7,138 MiB/sec
		read	20,772 MiB/sec	8,443 MiB/sec
	4 nodes, 80 processes	write	20,896 MiB/sec	4,134 MiB/sec
		read	13,861 MiB/sec	4,284 MiB/sec

運用中でも調達当初の性能(read/write合算60GB/sec以上)を確認

(注)全く違う環境での評価なので、VCCのデータはあくまで参考データ。

IORによる性能評価

- 2018年10月18日(運用中:**高負荷時**)にスケジューラNQS II よりジョブ投入して計測
  - *ior –w -r –b 8g –t m –C –Q 21-e –vv –F –o output*

			OCTOPUS
IOR 2.10.3	32 nodes 640 processes	write	25,037.06 MiB/sec
		read	15,908.67 MiB/sec
	16 nodes, 320 processes	write	19,870.49 MiB/sec
		read	10,768.68 MiB/sec
	8 nodes, 150 processes	write	14,859.89 MiB/sec
		read	17,179.97 MiB/sec
	4 nodes, 80 processes	write	12,588.92 MiB/sec
		read	4726.19 MiB/sec

ddによる性能評価

- 2018年10月18日の運用中にスケジューラNQS II よりジョブ投入して計測
- *ブロックサイズ＝1 Mに固定して、256, 512, 1024MBのファイルをwrite/read した際のスループット計測*
  - *Write: dd if=/dev/zero of=$LUSTERDIR/$FILE bs=1M count=??? conv=fsync*
  - *Read: dd if=$LUSTERDIR/$FILE of=/dev/null bs=1M count=???*

(注)全く違う環境での評価なので、VCCのデータはあくまで参考。

			OCTOPUS	VCC
dd	1 node, 1process	256MB write	1.2 GB/sec	0.33 GB/sec
		256MB read	1.6 GB/sec	0.98 GB/sec
		512MB write	1.2 GB/sec	0.43 GB/sec
		512MB read	1.5 GB/sec	1.00 GB/sec
		1024MB write	1.5 GB/sec	0.57 GB/sec
		1024MB read	1.5 GB/sec	1.00 GB/sec

MDTestによる性能評価

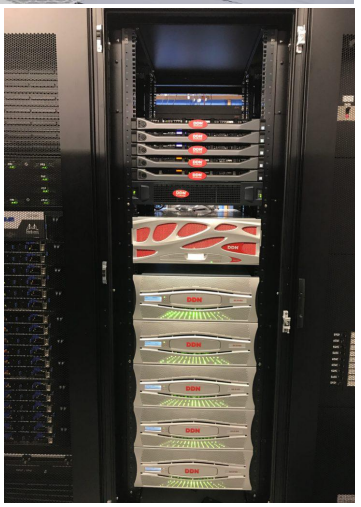
- 運用中にスケジューラNQS II よりジョブ投入して計測
  - Unique: 各プロセスが個別ファイルを処理
  - Shared: 各プロセスが同一ファイルを処理

(注)全く違う環境での評価なので、VCCのデータはあくまで参考。

,			OCTOPUS	VCC
MDTest 1.9.3	32 nodes 64 processes	unique file creation	75,963 op/sec	33,528 op/sec
		unique file stat	272,689 op/sec	151,143 op/sec
		unique file read	212,684 op/sec	271,262 op/sec
		unique file removal	106,531 op/sec	44,003 op/sec
		shared file creation	64,689 op/sec	2,703 op/sec
		shared file stat	139,334 op/sec	28,100 op/sec
		shared file read	131,328 op/sec	278,557 op/sec
		shared file removal	76,596 op/sec	10,613 op/sec

Summary (まとめ)

- OCTOPUSを支えるLustreストレージについて報告・紹介
  - **DDN ExaScalerを基軸としたストレージサービス**の実現
    - DDN ES14KX, SS8462を活用
    - 3PBを1ラックに高密度搭載
    - 10 TB NL-SAS(7.2krpm) HDD 410個
    - 高信頼
      - 2017年12月より障害がほとんどない
  - Lustre運用の実際面
    - Partition &Quota Configuration
  - 性能評価
    - 運用中でも高性能を提供



Expectation and Concern to Lustre (データ基盤?)

Cybermedia Center
Osaka University

- **Expansion of functionalities or familiarities with other storage service?(機能拡張 & 他ストレージサービスとの親和性強化?)**
  1. User-Friendly Data Aggregation (**データ集約機能**)
     - Large data are still *immovable like a mountain.* *"動かざること山の如し"*
  2. Security/Data-protection functionality from not only technological but also ethical and regulation aspects. (**セキュリティ/データ保護機能**)
     - Security-sensitive data still sits in the outside of our supercomputer environment.
  3. For hybrid use of Cloud and on-premises environment

- **Continuous, prompt, and versatile support and information sharing (継続的、迅速かつ臨機応変なサポートと情報共有)**
  - For prompt feedback: Users and administrators can benefit from new technologies and functionalities.
  - For stable system operation: "Never waste researchers' time and efforts."