

地球シミュレータ(ES4)とストレージシステムの紹介

国立研究開発法人海洋研究開発機構
付加価値情報創生部門 地球情報基盤センター
計算機システム技術運用グループ
中川 剛史



Sailing for the Earth,
Diving for Science & Technology

Dec.10 2021 JLUG 2021

JAMSTEC
<http://www.jamstec.go.jp/>



Agenda

1. JAMSTECの紹介
2. ES4計算ノードとストレージの紹介
3. ストレージの性能と運用状況の紹介



Dec.10 2021 JLUG 2021

海洋研究開発機構(JAMSTEC)とは？



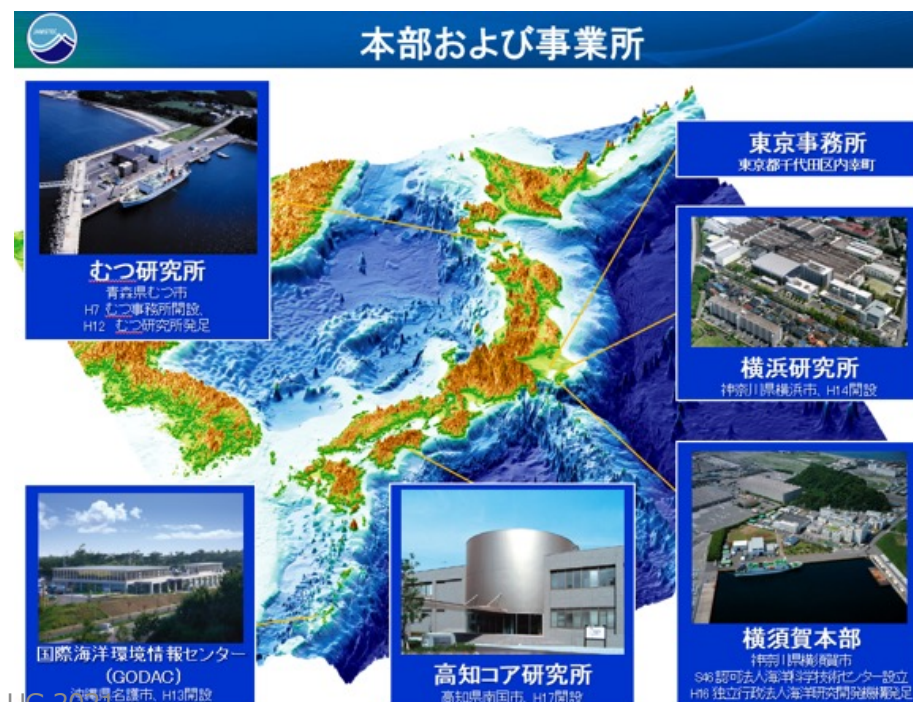
海洋研究開発機構 (Japan Agency for Marine-Earth Science and Technology: JAMSTEC ジャムステック) は、平和と福祉の理念に基づき、海洋に関する基盤的研究開発、海洋に関する学術研究に関する協力等の業務を総合的に行うことにより海洋科学技術の水準の向上を図るとともに、学術研究の発展に資することを目的とした組織

JAMSTECは、海洋・地球・生命の統合的理解のため、

- (1) 地球環境変動の統合的理解とその予測
- (2) 地球内部ダイナミクスの統一像の構築と
地震・津波の防災研究
- (3) 生命の進化と海洋地球生命史
- (4) 資源研究・海洋地球生命工学の新たな展開

が主な研究開発テーマ。地球深部探査船「ちきゅう」や
有人潜水調査船「しんかい6500」、スーパーコンピュータ
「地球シミュレータ」を運用

地球惑星分野における幅広い研究領域がターゲット



Dec.10 2021 JLUG 2021

JAMSTEC Super Computer & Storage



■ 大型計算機システム (Data Analyzer システム)



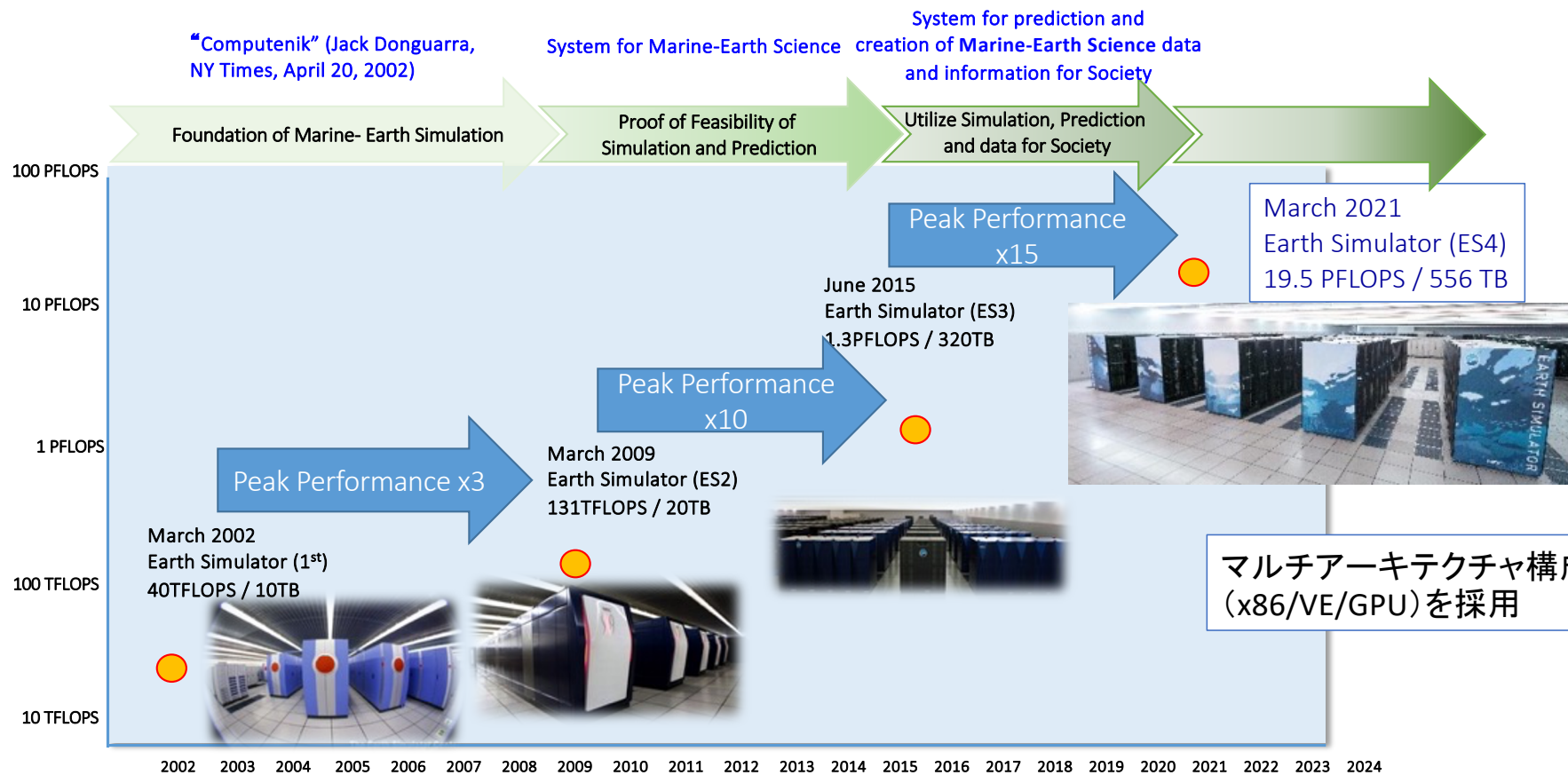
Feb 2018 -
HPE Apollo6000 & 2000
380 Nodes; 760 CPU (30400 cores);
1.16 PFLOPS
76.3 TB memory
5 PB storage (DDN EXAScaler(Lustre))

■ 地球シミュレータ

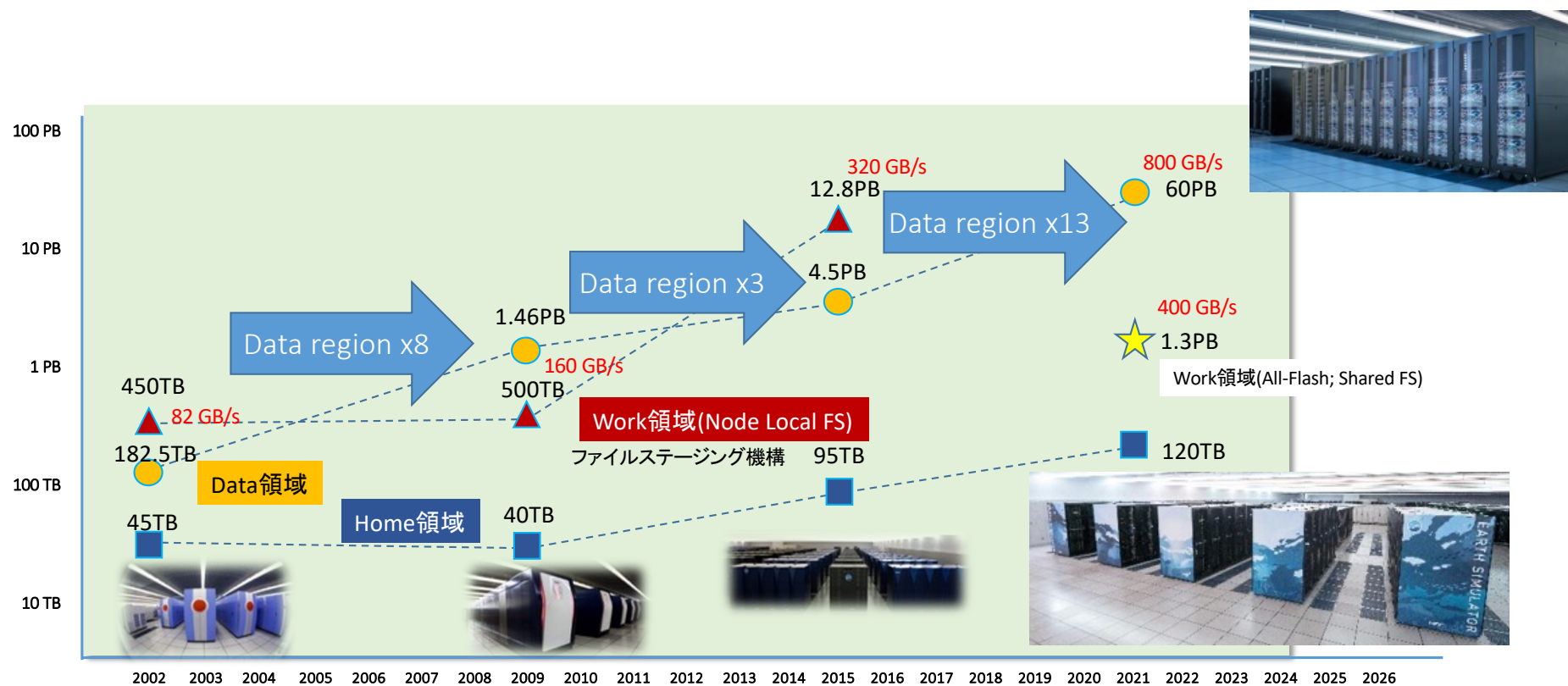


Mar 2021 -
NEC SX-Aurora TSUBASA & HPE Apollo 2000 & 6500
1412 Nodes; 2104 CPU (134656 cores);
5472 VE; 64 GPU; Total 19.5 PFLOPS
556.5 TB memory
60 PB + 1.3 PB storage (DDN EXAScaler(Lustre))

Earth Simulator since 2002

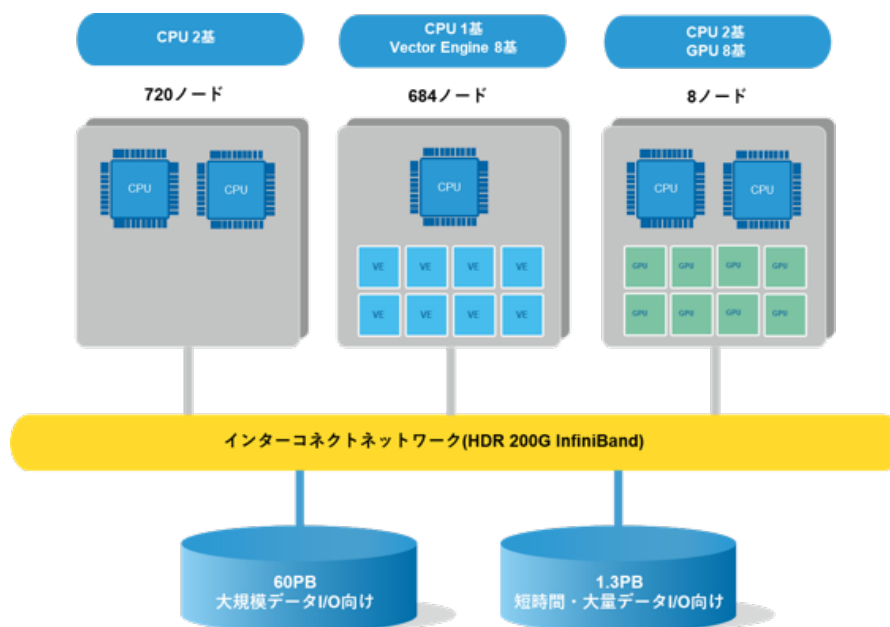


History of ES Storage system



Dec.10 2021 JLUG 2021

システム構成



インターコネクト

右記に記載の全ての機器は、HDR 200GのInfinibandファブリック(fat-tree構成)に接続されます。さらにVE搭載ノードは演算通信のみで利用するインターコネクトとして、前述とは別のHDR 200G InfiniBandファブリック(3つのfat-treeで構成されるDragonfly+)に接続されています。

VE搭載ノード

各VE搭載ノードは、AMD EPYC 7742 と 128 GiB DDR4 および 8基の NEC SX-Aurora TSUBASA Vector Engine (VE) を搭載しています。VEは 2.45 TFLOPS の演算性能と 48 GiB のHBM2を搭載しています。各ノードの性能は 21.9 TFLOPS で、全 648 ノードで 14.9 PFLOPS になります。

CPUノード

CPUノードは2基の AMD EPYC 7742 と 256 GiB DDR4 を搭載しています。各ノードの性能は4.6 TFLOPSです。CPUノードは720ノードあり、ピーク性能は3.3 PFLOPS、メモリ容量は180 TiBになります。

GPU搭載ノード

各GPUノードは2基のAMD EPYC 7742 と 4 TiB DDR4 、8基の NVIDIA A100 Tensor Core GPUを搭載しています。GPUの性能は 19.5 TFLOPS (DP, Tensor Core) です。さらにこのノードはアプリケーションの一時的なストレージとして7つの 3.2TB NVMe SSDを搭載しています。

ストレージ

Data Direct Networks社製の2つのストレージがあり、1つは61 PBの Lustre ファイルシステムで、合計800 GB/s の帯域があります。もう1つは1.3 PB のオールフラッシュLustreファイルシステムで、30M IOPSの性能があります。

計算ノードの仕様およびアーキテクチャ

VE搭載ノード 684ノード



684台あるVE搭載ノードは5472基の NEC SX-Aurora TSUBASA Vector Engineにより、新システムの中で最も強力な演算性能、最も大きなメモリ容量、最も広いメモリ帯域を提供します。VE搭載ノード部は、現行地球シミュレータの代表的なワークロードに対応します。

CPU	AMD EPYC 7742 @ 2.25 GHz
メモリ	128 GiB DDR4
アクセラレータ	8 SX-Aurora TSUBASA Type 20B (2.45 TFLOPS, 48 GiB per VE)

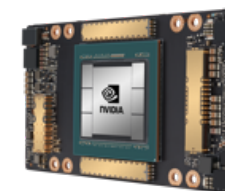
CPUノード 720ノード



720台のHPE Apollo 2000は1440基のCPUと180 TiBのメモリを搭載し、3.3 PFLOPSのピーク性能を提供します。このCPUノード部は、x86アーキテクチャに最適化されたアプリケーションや商用アプリケーションなど様々なアプリケーションに対応します。

CPU	Dual AMD EPYC 7742 @ 2.25 GHz
メモリ	256 GiB DDR4

GPU搭載ノード 8ノード



各GPU搭載ノードは4 TiBのメモリと8基のNVIDIA A100及び7基の3.2 TB NVMe SSDを搭載します。このGPU搭載ノード部は、GPUアプリケーションやML/DLアプリケーションのワークロードに対応します。さらに、大容量メモリや高速なI/O性能が必要なアプリケーションにも対応します。

CPU	Dual AMD EPYC 7742 @ 2.25 GHz
メモリ	4 TiB DDR4
アクセラレータ	8 NVIDIA A100 Tensor Core (19.5 TFLOPS, 40 GiB HBM2 per GPU)

Dec.10 2021 JLUG 2021

ストレージ構成

Home領域 120 TB



- ホームディレクトリが置かれ、ユーザ利用環境（ソースコードやツールなど）の保存領域
- 総容量: 120 TB(=60 TB × 2パーティション)
- 定期バックアップあり

Data領域 60 PB



- 計算ノードからの大規模I/O用途
- ES3におけるデータ領域とワーク領域の役割を兼ねる
- 総容量：60 PB（=30 PB × 2パーティション）
- IO性能：400 GB/s, Meta 630K IOPS
※パーティション毎の性能

Work領域 1.3 PB



- 短時間における大量データへのI/O用途
- シミュレーションとデータ解析の連携
- NVMe SSDで構成
- 総容量：1.3 PB、1パーティション
- IO性能：400 GB/s, Meta 315K IOPS, 4K rand 30M IOPS

ファイルシステム: DDN EXAscaler(Lustre)

計算ノードから透過的にアクセス可能 ※ファイルステージングシステムは未採用

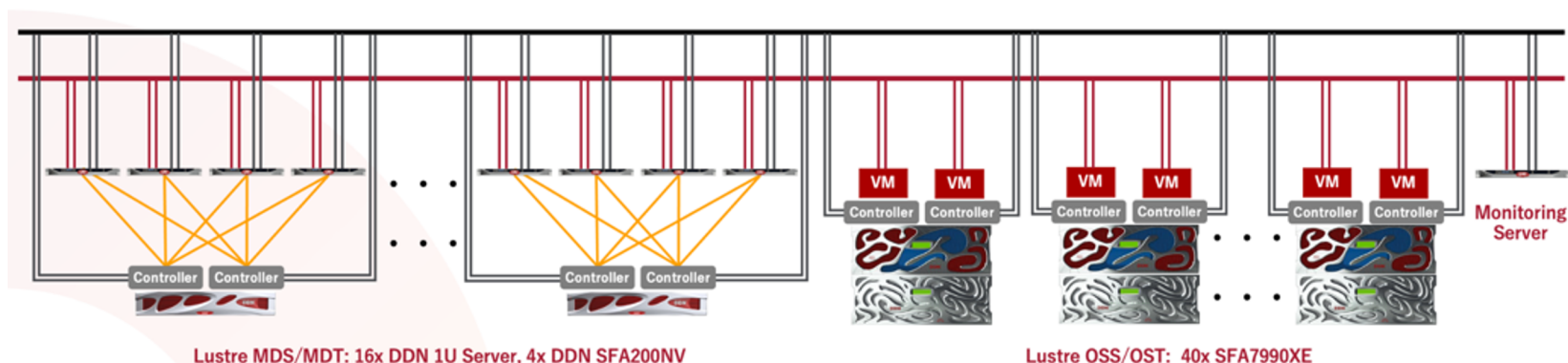
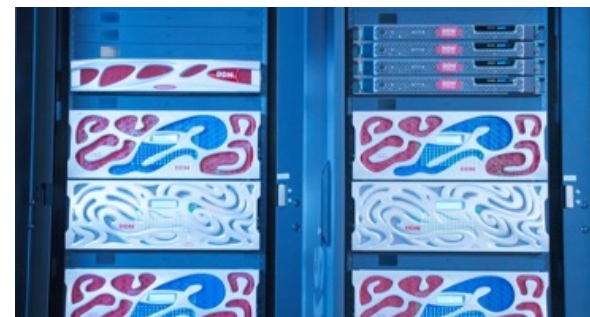
ES4ストレージの詳細 データ領域(data)

File System	DDN EXAScaler (Lustre 2.12.6)
Lustre MDS	16x DDN 1U Sever
Lustre MDT	4x DDN SFA200NV, 44x 1.92TB NVMe SSD
Lustre OSS/OST	40x DDN SFA200NVXE, 41x DDN SS9012, 7021x 12TB HDD
Monitoring Server	1x DDN 1U Server
Usable Capacity	61 PB (内1PBはバックアップ領域やソフトウェア環境に利用)
Maximum inode	22 billion

30PB

30PB

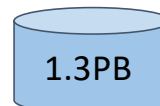
- 2パーティションで運用
- BW 400 GB/s, Meta 630 kIOPS /1parts.
- MDT 8ユニット&OST 80ユニット/1parts.
- ユーザDirはDNE Phase1がデフォルト



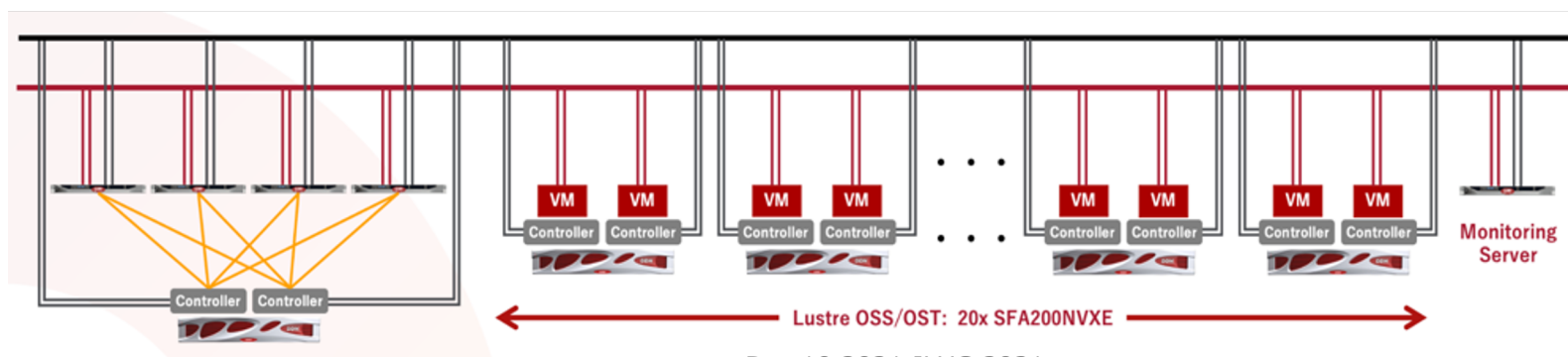
Dec.10 2021 JLUG 2021

ES4ストレージの詳細 ワーク領域(work)

File System	DDN EXAScaler (Lustre 2.12.6)
Lustre MDS	4x DDN 1U Sever
Lustre MDT	1x DDN SFA200NV, 21x 1.92TB NVMe SSD
Lustre OSS/OST	20x DDN SFA200NVXE, 480x 3.84TB NVMe SSD
Monitoring Server	1x DDN 1U Server
Usable Capacity	1.3 PB
Maximum inode	11 billion



- 1パーティションで運用
- BW 400 GB/s, Meta 315 kIOPS
- MDT 4ユニット&OST 80ユニット
- ユーザDirはDNE Phase1がデフォルト



Dec.10 2021 JLUG 2021

ES3ストレージとES4ストレージの比較（ハードウェア）



	ES3	ES4
容量	4.5 PB (data; 7 partitions) 12.8 PB (work with file staging)	60 PB (data合計; 2 partitions) 1.3 PB (work; 1 partitions)
#IO server (OSS)	28 (data; 7 partitions) 80 (work; 40 partitions)	40 (data合計; 2 partitions) 20 (work)
#metadata server(MDS)	↑	16 (data合計) 4 (work)
ファイルステージングシステム	○	×
ファイルシステム	NEC ScaTeFS	DDN EXAscaler (Lustre)
IO Network BW / node	1.0 GB/s	25 or 50 GB/s
Total IO BW	112 GB/s (data合計) 320 GB/s (work合計)	800 GB/s (data合計) 400 GB/s (work)

ファイルシステムモニタリング (DDN LustreMon)



DDN LustreMon

- リアルタイムのI/Oモニタリングが可能
- IO size**, File System Usage, Load Average
- Throughput, Metadata operation**
- 計算ノード毎・**ユーザ毎**・**グループ毎**・**バッチジョブ毎**・**キュー毎**

モニタリング・障害対応以外にも、
ユーザサポートの原因切り分けにも活用中



実際のユーザサポート時の原因切り分け例

「MPIプログラムの実行時間が数か月前より2倍以上遅くなった」

- ①HW障害の有無やSW(コンパイラやライブラリ)の確認
- ②**ユーザ毎やグループ毎、バッチジョブ毎のIO量を確認**
→該当MDSで大量のメタデータアクセス有
- ③ソースコードにおけるIOルーチン部分の確認
→MPIIOも利用している特徴があったが、該当箇所に問題無し
→原因は、全MPIプロセスでstdoutへ詳細なログをLustreへ書き出していること+ジョブの同時実行と判明



ES4ストレージの利用状況（Lustreモニタリングによる利用方法の分類）



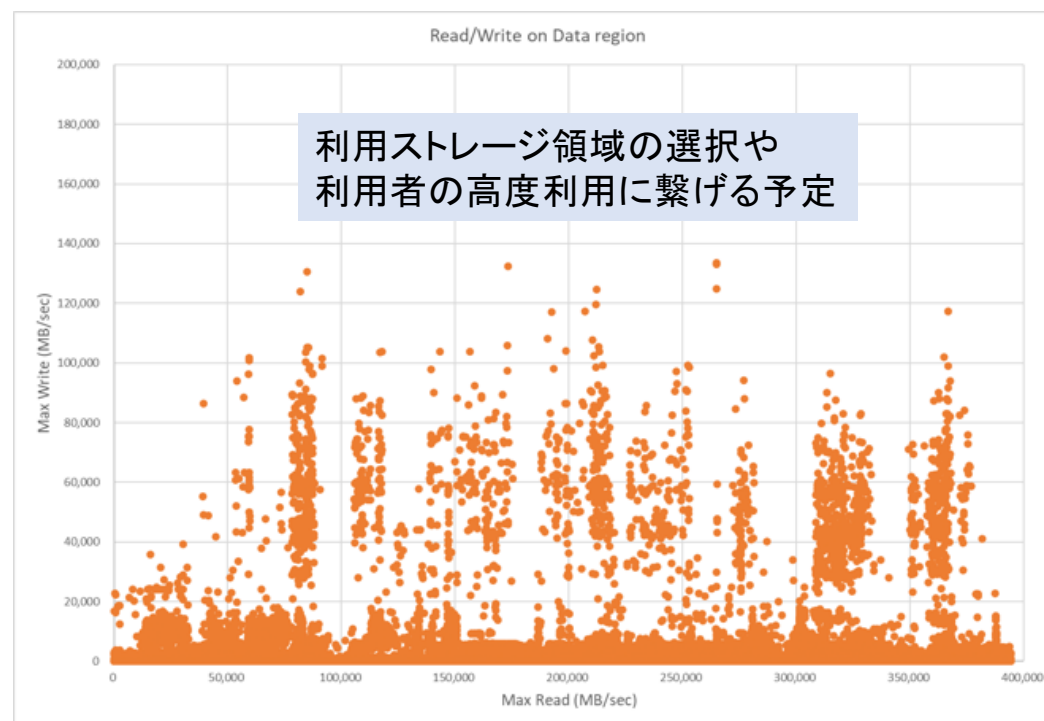
計算ノードにおけるユーザジョブ単位で以下の情報を
Lustreモニタリングで取得
※利用者も自身のジョブに関して、ログインサーバでチェック可能

ユーザジョブIOの見える化

- ・Total IOサイズ (MB)
- ・Read/Writeサイズ (MB)
- ・Max Read/Write (MB/sec)
- ・メタデータ(ops)情報

Start Time	End Time	Total Size (MB)	Write Size (MB)	Read size (MB)	Max Write (MB/sec)	Max Read (MB/sec)	Meta Total (Ops)
2021/12/09 13:57:33	2021/12/09 14:46:49	59462.60	47638.39	11824.21	307.33	68.67	2635511.46
2021/12/09 13:58:15	2021/12/09 14:47:08	59717.86	54241.45	5476.41	609.02	48.00	2540497.78
2021/12/09 13:58:39	2021/12/09 14:47:14	344549.74	77822.37	64746.37	990.91	326.38	401.99
2021/12/09 14:40:41	2021/12/09 14:47:16	57560.91	39904.26	17656.65	380.33	166.71	63467.70
2021/12/09 13:58:39	2021/12/09 14:47:51	42312.57	32017.47	10295.10	261.05	98.57	2065983.11
2021/12/09 13:59:20	2021/12/09 14:48:01	55003.68	45292.74	9710.94	475.66	100.09	2584480.69
2021/12/09 14:47:53	2021/12/09 14:48:32	2080.22	376.87	1623.35	17.76	19.61	69730.85
2021/12/09 14:01:42	2021/12/09 14:48:35	37989.21	31705.04	6284.17	397.14	56.97	3086452.21
2021/12/09 14:01:42	2021/12/09 14:48:38	59424.56	48964.19	10460.37	426.69	90.80	2970634.90
2021/12/09 14:01:42	2021/12/09 14:48:48	40399.48	36974.83	4324.65	272.18	74.25	283656.70
2021/12/09 14:01:42	2021/12/09 14:48:52	47779.56	44228.11	3551.45	287.09	70.40	2336751.03
2021/12/09 14:01:42	2021/12/09 14:49:01	55508.68	50616.83	4891.85	580.02	61.44	2079734.82
2021/12/09 14:01:42	2021/12/09 14:49:04	29040.91	25702.44	3338.47	353.26	63.01	1630184.86
2021/12/09 14:01:02	2021/12/09 14:49:06	51877.37	43012.22	9165.15	492.38	75.04	2533259.29
2021/12/09 14:01:41	2021/12/09 14:49:13	48027.31	39109.11	8918.20	289.74	120.03	2785768.97
2021/12/09 14:01:20	2021/12/09 14:49:14	46603.15	41716.93	4886.22	402.80	69.27	2055430.30
2021/12/09 14:01:41	2021/12/09 14:49:15	34027.83	29095.77	4932.06	350.11	70.50	2783921.04
2021/12/09 14:01:42	2021/12/09 14:49:17	54455.60	43835.50	10620.10	404.05	142.23	2341801.51
2021/12/09 14:01:34	2021/12/09 14:49:28	30780.65	26689.66	4090.99	465.06	52.60	1843550.60
2021/12/09 14:01:34	2021/12/09 14:49:28	44829.90	37781.48	7048.42	355.64	75.82	3679528.08
2021/12/09 14:42:58	2021/12/09 14:49:31	40282.94	38971.30	9211.64	872.00	111.77	44029.40
2021/12/09 14:00:44	2021/12/09 14:49:37	51740.88	43726.31	8014.57	320.60	91.82	2293561.64
2021/12/09 14:01:42	2021/12/09 14:49:37	41484.15	33831.21	7652.94	520.73	62.47	2032564.42
2021/12/09 14:01:15	2021/12/09 14:49:45	60982.21	52574.64	8407.57	420.90	51.84	2623703.08
2021/12/09 14:01:41	2021/12/09 14:49:50	50761.88	45123.69	5638.19	310.98	52.11	2811699.89
2021/12/09 14:01:41	2021/12/09 14:49:52	59238.00	54262.57	4975.43	487.77	88.01	2382342.63
2021/12/09 14:01:09	2021/12/09 14:49:55	44834.55	35664.96	9169.59	413.99	54.40	1826210.65

Lustremon コマンドのアウトプット例



Dec.10 2021 JLUG 2021

※モニタリングデータ3ヵ月分(2021 9月-11月)

ストレージ利用状況（Lustreモニタリングより）

※モニタリングデータ3ヵ月分(2021 9月-11月)



Home領域

利用率 25 %
(2021年12月時点)



BW < 10GB/s & Meta < 25Kops

Data領域

利用率 23 %



BW 50-100GB/s & Meta100-200Kops

Work領域

利用率 35 %
※運用部門利用やテスト利用者も多い



BW < 10GB/s & Meta < 10Kops

複数ストレージの使い分けが課題！
Work領域への誘導など

ES4ストレージの性能比較



複数ノードでの評価

データ領域構築時(暫定4パーティション環境)を用いた性能試験で1パーティション基礎性能を確認

IOバンド幅(IOR; FilePerProcess) :

VE搭載ノード: 12ノード192MPI(16 per node)

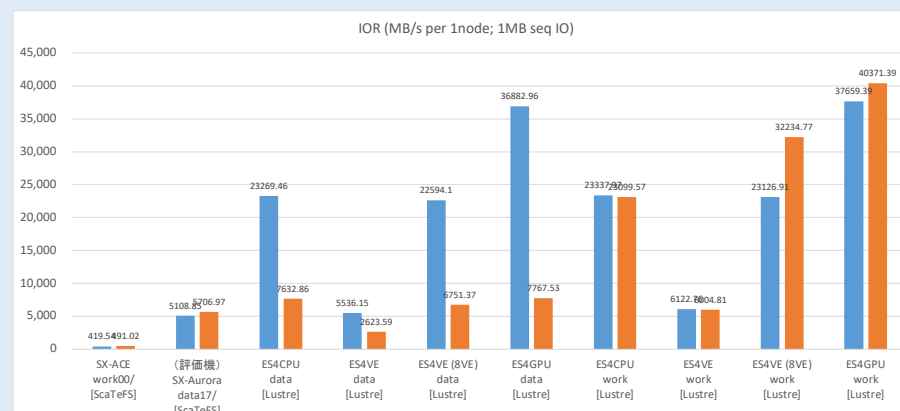
Write: 200 GB/s
Read: 200 GB/s

メタデータ性能(MDTEST) :

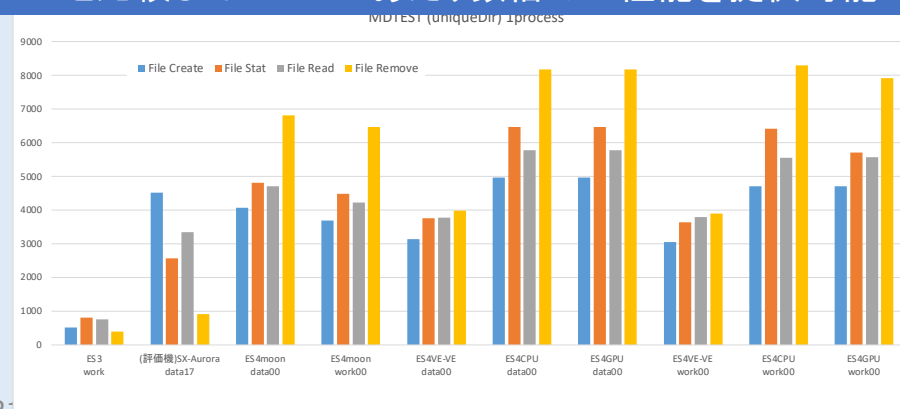
VE搭載ノード: 32ノード1024MPI(32 per node)

File creation: 250,000 ops
File stat: 1,100,000 ops
File read: 430,000 ops
File removal: 180,000 ops

1ノード内での評価



ES3と比較してプロセスあたり数倍のIO性能を提供可能に



Dec.10 2021

16

ES4ストレージ評価（安定性評価）



複数の実ユーザプログラム（シミュレーションコード）に基づいた小～中サイズ問題を、ジョブスケジューラを用いて、それぞれ平均1000本程度連続的にランダム同時実行

条件: VE搭載ノードまたはCPUノードを利用

データ領域

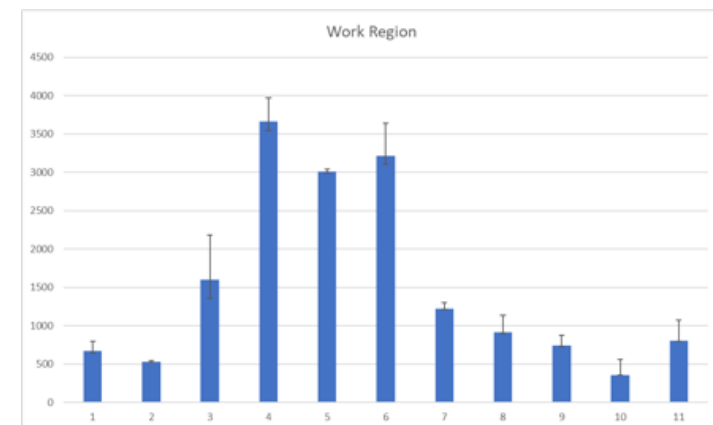
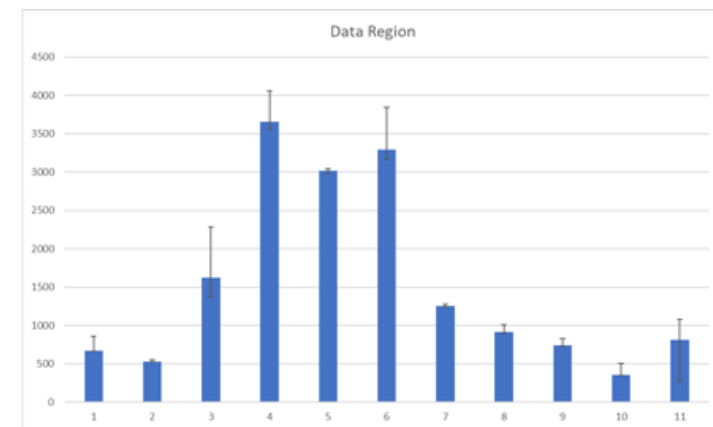
- 平均で17%の計算時間のブレあり

ワーク領域

- 平均で17%の計算時間のブレあり
- データ領域よりもわずかに速い傾向があるが、ほぼ同等

計算ノードのアーキテクチャの違いやIBネットワーク自体の影響も考えられるが、計算時間への影響の観点からは、データ領域とワーク領域で大きな差は無かった

ワーク領域の容量制限があるため、データ解析、もしくは、シミュレーションとデータ解析の連携などに利用する？



ES4ストレージ性能確認（ユーザアプリでの大規模実行）



スペクトル法を用いた地震波形シミュレーションコード(Fortran)での
メッシュ生成部およびソルバー部の大規模実行(エレメント数4200万)を評価

条件: VE搭載ノードのVE部(646ノード;5166VE)・Data領域・DNE PHASE1
実行時オプション(エンディアン変換無・入出力バッファ10MB設定)

メッシュ生成部

総プロセス数: 41334MPI

実行時間: 29分12秒

Input: 数MB程度

Output: 合計41 TB（合計約50万ファイル）

Write時間: 2分程度(推定値)(=350-400 GB/s)

ソルバー部

総プロセス数: 41334MPI

必要メモリ量: 111TB

実行時間: 5時間(波形ファイル25分間分)

Input: 合計41TB（合計約50万ファイル）

Output: 数GB程度

Read時間: 98.86秒(=414 GB/s)



実アプリでの大規模MPI実行においても、設計値通りのIO性能を発揮

ES4でのIO500ベンチマーク測定結果



IO500の10ノードチャレンジをData領域およびWork領域で測定

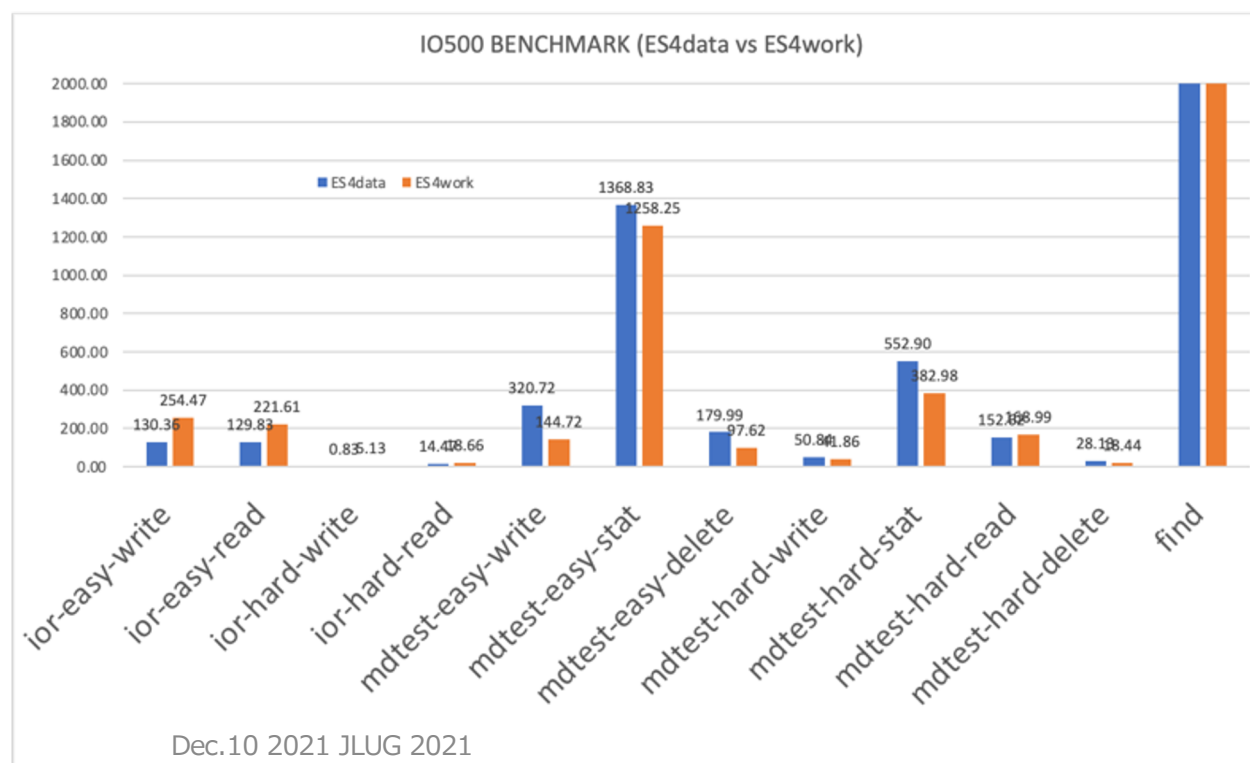
条件: VE搭載ノードのVH部で測定・DNE PHASE2・メンテナンス時のサイレント環境で実行

IOバンド幅(GiB/s) :

ワーク領域がデータ領域の
約2倍の性能

メタデータ性能(kIOPS) :

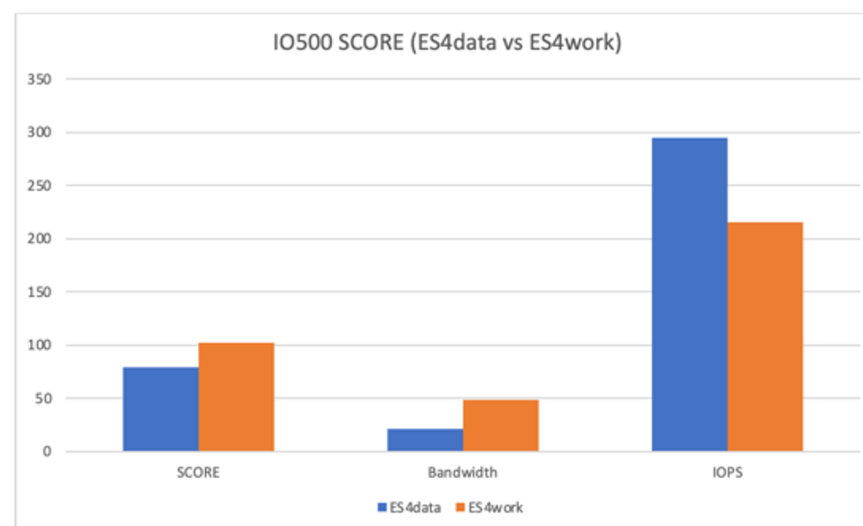
MDSの台数が多いデータ領域の方が、ワーク領域より性能が高い（項目によっては最大2倍）



ES4でのIO500ベンチマーク測定結果(スコア)



	データ領域	ワーク領域
総合スコア	79.14	101.87
BWスコア [GiB/s]	21.26	48.19
MDスコア [KIOPS]	294.61	215.38



総合スコアは、ワーク領域の方が大きい。

ワーク領域の結果をISC21リストへ登録し、10ノードチャレンジで18位（SC21リストでは23位）
更なる性能改善と大規模環境での測定評価を目指し、結果を分析中。

• ES4計算ノードおよびストレージ紹介

- 計算ノード: マルチアーキテクチャ構成 (x86/VE/GPU)
- ストレージ: 研究開発を行う上での十分なディスク容量
 - データ領域: 60 PB, 1パーティション 30PB, 400 GB/s, 630 kIOPS
 - ワーク領域: 1.3 PB, 400 GB/s, 315 kIOPS

• ストレージのIO性能評価

- IO500 10ノードチャレンジ(SC21リスト)でWork領域23位、Data領域26位相当
- 実ユーザアプリにおいても、計算能力向上分をカバーするストレージ能力を確認

• IOモニタリング情報の活用

- 障害監視に加えて、ユーザサポートの原因切り分けにも利用

• 利用方法の検討

- 分散共有ファイルシステム(HDD or NVMe SSD)と計算ノードローカルストレージとの効果的な使い分けと利用者の高度利用促進のための情報提供