九州大学情報基盤研究開発センター新スーパーコンピュータシステムの紹介

大島 聡史 [1)], 南里 豪志 [1)], 美添 一樹 [1)], 平島 智将 [1)], 原田 浩睦 [1)], 池田 嗣穂 [1)]

1) 九州大学 情報基盤研究開発センター

ohshima@cc.kyushu-u.ac.jp

Introduction of the New Supercomputer System at Research Institute for Information Technology of Kyushu University

Satoshi Ohshima[1)], Takeshi Nanri[1)], Kazuki Yoshizoe[1)], Tomoyuki Hirashima[1)], Hiroyoshi Harada[1)], Tsuguho Ikeda[1)]

1) Research Institute for Information Technology, Kyushu University

概要

九州大学情報基盤研究開発センターでは、2024 年 7 月より新スーパーコンピュータシステムの運用を開始する。本稿では、このシステムの概要を、現有システム ITO からの改善点を踏まえながら紹介する。

1 背景

九州大学情報基盤研究開発センター（以下、本センター）は、大学等の教員、大学院学生、その他の研究者が学術研究等のために利用する全国共同利用施設であり、1969 年に設置されて今日に至っている。本センターでは、科学技術シミュレーションに加えて第 5 期科学技術基本計画に示された AI（人工知能・機械学習）、ビッグデータ、さらにデータサイエンスの研究及びこれらを活用した研究にも対応した研究基盤の提供を目指し、2017 年度にスーパーコンピュータシステム ITO を導入した。ITO は、国内の大学・高等専門学校・大学共同利用機関の構成員の他、センター長が利用を認めた民間企業の構成員や国外の研究者にも利用されている。

このシステムは 2023 年 2 月までの需要を想定したものであり、その後、半導体や部材の世界的な不足に伴って運用を 2024 年 2 月まで延長したものの、計算資源の需給が逼迫していた。また、第 6 期科学技術・イノベーション基本計画では、データ駆動型研究やオープンサイエンスの推進が掲げられており、これを支援する新しい計算基盤の整備が求められている。そこで単なる計算能力増強だけでなく、データサイエンスに代表されるように、急速に多様性を増しつつある様々な需要にも耐えるシステムへの更新が必要であった。

そこで、2024 年 7 月から 5 年間強で予想される計算需要を想定するとともに、従来の大規模科学技術シミュレーションに加え、データ駆動型研究の展開を支援するスーパーコンピュータシステムの調達を実施し、入札の結果、富士通社の FUJITSU Server PRIMERGY シリーズを中核とするシステムとなることが決定した。本稿では、まず現有システム ITO の課題を説明し、その課題を解決する新スーパーコンピュータシステムについて、特徴と概要を紹介する。

2 現有スーパーコンピュータシステム ITO の課題

新スーパーコンピュータシステムの調達にあたり、本センターの現有スーパーコンピュータシステム ITO における主な課題として以下を提起し、これらを解決できるシステムとなるよう仕様を策定した。

計算能力の需給逼迫解消

ユーザの計算需要に計算リソースが追い付いておらず、特に CPU 搭載ノード群で待ち時間の長期化が頻発している。一方、GPU 搭載ノード群については、2017 年の導入当時のアーキテクチャのままであり、最近のデータ駆動型研究や機械学習での需要にこたえるには不十分である。そのため、今後の需要を想定した計算能力の全般的な向上が求められている。

外部ストレージとの連携

ITO のようなオンプレミス型のスーパーコンピュータは冗長性が低いため、保守作業等により

全サービスを停止する期間が発生し、その期間はファイルにアクセスもできないため、可用性に問題が生じている。特に今後重要性を増すオープンサイエンスでは、外部システムとの間で頻繁にデータの相互参照を行うことが想定されており、従来のようなオンプレミスストレージ領域のみによる構成では対応が困難である。そのため、パブリッククラウドのストレージサービスなど、24 時間稼働が想定されている外部のストレージとの同期機能や連携機能を有する、オープンなストレージシステムが必要である。

新しい利用形態に対応したソフトウェアツール

ITO には、仮想マシンを時間予約して対話的に利用する機能が用意されているが、Linux での利用を前提としている点や、仮想マシンの起動に時間を要する点等の課題があった。一方、特に最近スーパーコンピュータ利用が進みつつあるデータ駆動型研究の分野では、ブラウザベースの GUI インタフェースやワークフローツールによるプロジェクト管理などが一般的に利用されている。そのため、これらのソフトウェアツールを用意することにより、PC やパブリッククラウドなどと同様の使い勝手を提供可能となると期待できる。

3 新スーパーコンピュータシステムの概要

3.1 新スーパーコンピュータシステムの構成

新システムのシステム構成図を図 1 に示す。本システムは主に、計算ノード群（ノードグループ A、ノードグループ B、ノードグループ C）、ログインノード群、共有ストレージ、管理サーバ群から構成される。本章では主なハードウェア仕様と導入予定ソフトウェアについて述べる。

新システムと旧システム (ITO) の性能比較を表 1 に、新システムの各ノードグループの詳細を表 3, 表 4, 表 5 にそれぞれ示す。

3.2 ノードグループ A

ノードグループ A は、GPU などの演算加速装置を搭載しない計算ノードグループであり、製品としては FUJITSU Server PRIMERGY CX2550 M7 により構成される。

ノードグループ A は各ノードに Intel 社の Xeon Platinum 8490H（Sapphire Rapids アーキテクチャ）を 2 基ずつと 512 GiB の DDR5 メモリを搭載しており、全ノードグループ中で最も多い 1024 の計算ノードにより構成される。このノードグループは、GPU の利用に適していない汎用のアプリケーションや、多数のノードを用いた大規模な分散計算への利用が想定されている。

同様に GPU を搭載しない ITO のサブシステム A と比較すると、ノード単体の理論演算性能は 2.11 倍に増加しているが、ノード数は 0.512 倍に減少している。ノードグループ（サブシステム）全体としては 1.07 倍に性能が向上している。

各ノードに搭載されているストレージは OS 起動用の HDD のみであり、ユーザが一時領域として使用することを想定した高速ストレージは搭載していない。ノード間の接続には InfiniBand NDR 200Gbps をノードあたり 1 ポート備えている。

3.3 ノードグループ B

ノードグループ B は、ノード内に 4 基の GPU を搭載した計算ノードグループであり、製品としては FUJITSU Server PRIMERGY CX2560 M7 により構成される。

ノードグループ B は各ノードに CPU としてノードグループ A と同じ Intel Xeon Platinum 8490H を 2 基と、1024 GiB の DDR5 メモリ、さらに GPU として NVIDIA H100 (SXM5 版, HBM2e 94GiB 搭載) を 4 基ずつ搭載している。このノードグループは、非常に高い演算性能とメモリ転送性能を有する GPU を搭載しており、データ科学アプリケーションにも数値シミュレーションにも高い性能が期待される。

同様に GPU を搭載した ITO のサブシステム B と比較すると、ノード単体の GPU の理論演算性能は FP64 で 6.32 倍、FP32 や FP16 などでは Tensor Core を含めると 46.67 倍に大きく増加しているが、ノード数は 0.29 倍に大きく減少している。ノードグループ（サブシステム）全体の GPU の性能としては、FP64 で 1.88 倍、FP32 や FP16 などでは Tensor Core を含めると 13.87 倍に大きく上昇している。メモリ転送性能については、GPU 単体では 3.27 倍に上昇しているものの、ノード数が減少したためノードグループ（サブシステム）全体ではほぼ同性能（1.03 倍）に留まっている。（表 1 では GPU の総性能にノードグループ C も含んでいる点に注意されたい。）

ユーザが利用できるノード内ストレージとしては、ノードあたり容量 12.8 TB, アクセス性能 4 GB/s（書き込み性能）の NVMe SSD を搭載している。ノード間の接続には InfiniBand NDR 400Gbps をノードあたり 2 ポート備えており、CPU と GPU を繋ぐ 2 つ

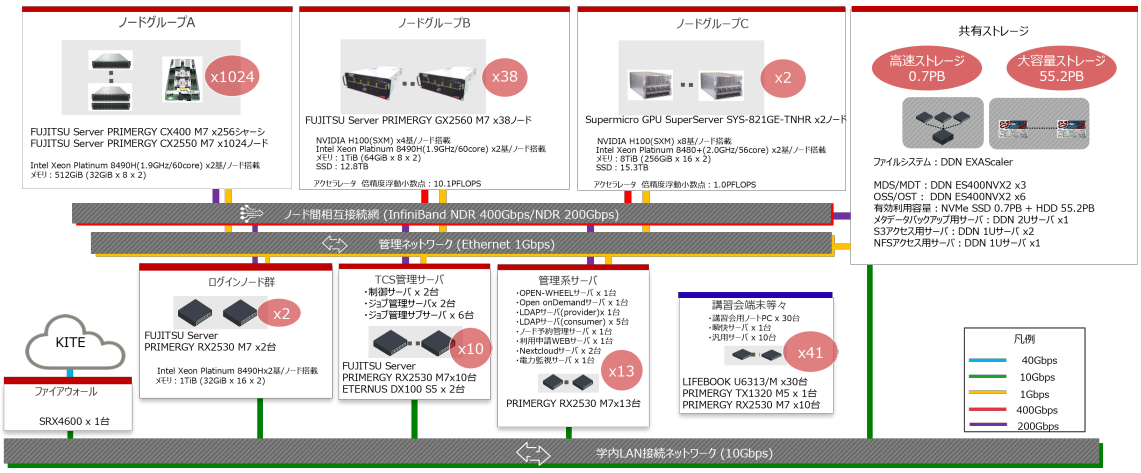

図 1　新スーパーコンピュータシステムの全体構成

表 1　ITO と新システムの全体性能の比較

	ITO	新システム	比率（対 ITO）
総理論演算性能			
- CPU FP64	7.38 PF *1	7.76 PF *3	1.05
- GPU FP64 (TC 除く)	2.71 PF *2	5.63 PF *4	2.08
- GPU FP16, BF16 (TC 含む)	10.84 PF *2	166.22 PF *4	15.33
総メモリ容量			
- ホストメモリ	542 TiB *1	566 TiB *3	1.04
- デバイスメモリ	8,192 GiB *2	15,568 GiB *4	1.90
総メモリ転送性能			
- ホストメモリ	341.18 TB/s *1	653.72 TB/s *3	1.92
- デバイスメモリ	374.78 TB/s *2	437.82 TB/s *4	1.17
総計算ノード数	2,292 *1	1,064 *3	0.46
ストレージ容量	HDD 24.6 PB	大容量 HDD 55.2 PB 高速 SSD 737.28 TB	2.24
ログインノード数	2	2	1.00

(*1: サブシステム A,B とフロントエンドシステムの合算。*2: サブシステム B のみ。*3: ノードグループ A,B,C の合算。*4: ノードグループ B,C の合算。)

の PCI-Express スイッチに 1 ポートずつ接続される。

3.4　ノードグループ C

ノードグループ C は、ノード内に 8 基の GPU を搭載した計算ノードグループであり、製品としては Supermicro SYS-821GE-TNHR により構成される。

ノードグループ C の各ノードには CPU として Intel Xeon Platinum 8480+ を 2 基と、8 TiB の DDR5 メモリ、さらに GPU として NVIDIA H100 (SXM5 版, HBM3 80GiB 搭載) を 8 基ずつ搭載している。このノードグループは、非常に高い演算性能とメモリ転送性能を有する GPU を多数搭載しており、単体ノードの性能やメモリ容量を多く必要とする分野での利用に適している。大規模言語モデル (Large Language Model, LLM) などの計算需要にも応えることが期待される。

ノードグループ C は、ITO における大容量フロントエンドノードが担っていた大容量メモリ環境の提供という役割を担っている。基本フロントエンドノードとともに提供していた予約利用については、新システムではノードグループ A, B, C のスケジューラが提供する予定である。

ユーザが利用できるノード内ストレージとしては、ノードあたり 15.3 TB, アクセス性能 5.6 GB/s（書き込み性能）の NVMe SSD を搭載している。ノード間

の接続には InfiniBand NDR 400Gbps をノードあたり 4 ポート備えており、CPU と GPU を繋ぐ 4 つの PCI-Express スイッチに 1 ポートずつ接続される。

3.5 ログインノード群

ログインノード群は、利用者が ssh ログインやプログラムのコンパイル、ジョブの投入などを行うためのノード群であり、製品としては FUJITSU Server PRIMERGY CX2530 M7 により構成される。

ログインノード群は 2 ノードによって構成されており、各ノードには CPU としてノードグループ A と同じ Intel Xeon Platinum 8490H を 2 基と、1024 GiB の DDR5 メモリを搭載している。

3.6 共有ストレージ

新システムは大容量ストレージと高速ストレージの 2 種類の共有ストレージを有する。製品としてはいずれも DDN EXAScaler により構成される。ファイルシステムはいずれも Lustre である。

大容量ストレージは HDD で構成され、680 本の HDD からなる OSS/OST を 6 セット、合計で 4,080 本の HDD を搭載しており、物理容量としては合計 69.1 PB、RAID6 による実効容量としては合計 55.2 PB の容量を備える。取り扱い可能なファイル数（inode 数）は約 450 億個である。書き込み性能は合計 360GB/s、読み込み性能は合計 420GB/s である。

高速ストレージは NVMe SSD で構成され、21 本の SSD からなる OSS/OST を 6 セット、合計で 126 本の SSD を搭載しており、物理容量としては合計 921.6 TB、RAID6 による実効容量としては合計 737.28 TB の容量を備える。取り扱い可能なファイル数（inode 数）は約 110 億個である。書き込み性能は合計 360GB/s、読み込み性能は合計 480GB/s である。

これらのストレージに対し、NFS や S3 による外部からのアクセス機能も提供する予定である。

3.7 インターコネクト

各システムを相互接続するインターコネクトは InfiniBand NDR によるフルバイセクションバンド幅の Fat Tree トポロジーにより接続される。ネットワークの上流に配置する Spine スイッチとしては NVIDIA の QM9790 を 13 台と QM9700 を 3 台、下流に配置する Leaf スイッチについては QM9790 を 21 台使用する。QM9790, QM9700 ともに InfiniBand NDR 400Gbps を 64 ポート搭載している。このネットワーク（NIC およびスイッチ）は MPI 集団通信の一部を CPU の代わりに実行する機能を有している。

制御および管理用のネットワークについては 1GbE または 10GbE の Ethernet により接続される。詳細は省略する。

表 2　ソフトウェア一覧

分類	ソフトウェア
コンパイラ・開発ツール	Intel oneAPI base&HPC toolkit, NVIDIA HPC SDK, NVIDIA CUDA SDK, GCC, Julia, Mathematica, MATLAB
数値計算ライブラリ	FFTW, PETSc, BLAS, LAPACK, ScaLAPACK
その他のライブラリ	HDF5, NetCDF, NAG
計算化学	GAMESS, GROMACS, LAMMPS, Quantum ESPRESSO, CP2K, Gaussian, GaussView, Molpro, VASP
流体・構造解析	OpenFOAM, Amber, MSC Marc/Nastran, FIELDVIEW, MicroAVS
機械学習・データ解析	MLflow, TensorFlow, PyTorch, Jupyter Notebook, R
ワークフロー・外部連携	OpenWheel, Open OnDemand, Nextcloud, AWS CLI, Azure CLI, gcloud CLI, OCI CLI, AWS ParallelCluster, Azure CycleCloud, GSISSH, Gfarm
その他	Singularity

3.8 ソフトウェア

新システムの提供する主なソフトウェアを表 2 に示す。新ステムでは ITO が提供してきたソフトウェアに加えて、Open OnDemand、Wheel、MLflow といった利便性や生産性の向上に関するソフトウェアや、NextCloud、AWS/Azure/gcloud/OCI のコマンドラインインタフェースといったスーパーコンピュータ外部との連携に関するソフトウェアなど、スーパーコンピュータの新しい使い方をサポートするソフトウェアを充実させている。今後も利用者からの要望に応じてソフトウェアの拡充をする予定である。

3.9 新スーパーコンピュータシステムによる課題解決

2 章で挙げた課題について、新スーパーコンピュータシステムでは、それぞれ以下の形で解決を図っている。

計算能力の需給逼迫解消

まず CPU ノード群について、ノードグループ A のノード数は ITO のサブシステム A の半分程度であるものの、ノード当たりのコア数は 3.3 倍、搭載メモリ量は 2.6 倍、メモリバンド幅は 2.4 倍であり、より大量のジョブを同時に処理できる構成となっている。また、計算性能について、単純なクロック周波数とコア数の積で計算される FP64 の理論演算性能では 3.2 節に記述した通り ITO のサブシステム A に対して 1.07 倍にとどまるものの、実際に科学技術計算に用いられる浮動小数点演算にもとづいたベンチマークプログラム群の性能値である SPECrate 2017 FP では、ITO のサブシステム A での 2000 ノードの合計値 414000 に対し、ノードグループ A は 1024 ノードの合計で 1050624 となり、約 2.5 倍に向上している。これにより、性能面でも、CPU ジョブの大幅なスループット性能向上が達成できている。一方 GPU ノード群については、3.3 節、および 3.4 節に記述した通り、ノードグループ B、C に最新の GPU アーキテクチャである NVIDIA H100 を搭載することで、Tensor Core を含めた理論演算性能が FP64 で 4.1 倍、FP16 で 15.3 倍と、ITO のサブシステム B に搭載された GPU に対して大幅な性能向上を達成している。さらに、特にノードあたり 8 基の NVIDIA H100 と 8TiB のメモリを搭載するノードグループ C は、LLM のように高い演算性能と大量のデバイスメモリを必要とする計算需要にも十分対応可能となっている。

外部ストレージとの連携

共有ストレージシステムには、クラウドコンピューティングにおいて標準的に用いられているストレージインタフェース Amazon S3 API、及び外部のシステムとのファイル連携を容易にする NextCloud によるアクセスを可能としている。さらにログインノード群では、主なクラウドコンピューティングサービス (Amazon Web Service, Microsoft Azure, Google Cloud, Oracle Cloud) を呼び出し可能な API 群が利用可能となっている。これらを用いて外部のストレージやコンピューティングサービスと連携することで、オープンサイエンスに求められる高可用性の実現が可能となる。

新しい利用形態に対応したソフトウェアツール

まず、ジョブスケジューラの対話型ジョブを、指定した時刻に利用可能とする予約機能を提供する。これにより短時間での対話型環境提供が可能となる。さらに、近年計算機センターでの導入が進んでいる Open OnDemand を提供することにより、ウェブブラウザ上の GUI での操作を支援する。さらに、研究プロジェクトにおけるタスクフローの管理を支援するツールとして MLflow と WHEEL を導入する。これらにより、今までスーパーコンピュータや Linux になじみのなかった利用者に対し、PC やパブリッククラウドなどに近い利用環境を提供する。

4 今後の予定

本センターでは、2024 年 7 月の新システム運用開始に向け、現在準備を進めている。現在のところ、最初の 2 カ月程度は試験運用とし、本運用の開始は 9 月を想定している。なお、現有の ITO の運用は 2024 年 2 月末で運用を停止するものの、ログインノードと共有ストレージは 2024 年 9 月末まで利用可能とする。これにより ITO の利用者はシステム更新の作業期間中（2024 年 3 月〜6 月）もファイル参照が可能となる予定である。また、新コンピュータシステムと ITO を 3 カ月程度並行運用することにより、ITO のファイルのうち必要なものを利用者自身で新コンピュータシステムにコピー可能とする。

新システムでの運用方針についても現在検討を進めており、特に利用負担金について、ITO で採用している月額定額制ではなく、予め購入したトークンを利用計算資源に応じて消費する従量課金制に移行することを検討している。これは、定額制では混雑状況によって計算資源単価が変動することが問題となっていたため、従量課金制の導入によって単価を安定させることが目的である。

なお、新システムの名称については、今後公募で決定する予定である。公募の要領など詳細はセンター Web ページ等でアナウンスする。

5 おわりに

本稿では、2024 年 7 月に運用を開始する新スーパーコンピュータシステムの概要を紹介した。現有システム ITO に対して、システム全体のジョブスループット向上や GPU 性能向上などのハードウェア性能向上に加え、対話型利用の利便性向上、ワークフロー管理支援およびクラウドコンピューティング連携のための

新しいソフトウェア導入により、今後予想される多様な研究分野からの計算需要に十分応えることのできるシステムを調達できたと考えている。利用者に質の高い研究環境を提供できるよう、導入業者の富士通株式会社と協力して稼働準備を進めたい。

表 3　ノードグループ A の構成

機種名		FUJITSU Server PRIMERGY CX2550 M7
CPU	型番とノードあたり数量	Intel Xeon Platinum 8490H (Sapphire Rapids) × 2
	コア数と動作周波数	60 コア, 1.90 GHz - 3.50 GHz
	1CPU あたり理論演算性能	FP64: 3,648 GFLOPS
	1CPU あたりメインメモリ	DDR5 4800 MHz, 256 GiB （32 GiB × 8 枚）
	1CPU あたり理論メモリバンド幅	307.2 GB/s （4800 MHz × 8 Byte × 8 チャネル）
ノード数		1024
ノード間接続		InfiniBand NDR (200Gbps) × 1 ポート/ノード
冷却方式		水冷
総理論演算性能		FP64: 7.47 PFLOPS
総メインメモリ容量		512 TiB
総理論メモリバンド幅		629 TB/s
ユーザ用ローカルストレージ		なし

表 4　ノードグループ B の構成

機種名		FUJITSU Server PRIMERGY CX2560 M7
CPU	型番とノードあたり数量	Intel Xeon Platinum 8490H (Sapphire Rapids) × 2
	コア数と動作周波数	60 コア, 1.90 GHz - 3.50 GHz
	1CPU あたり理論演算性能	FP64: 3,648 GFLOPS
	1CPU あたりメインメモリ	DDR5 4800 MHz, 512 GiB （64 GiB × 8 枚）
	1CPU あたり理論メモリバンド幅	307.2 GB/s （4800 MHz × 8 Byte × 8 チャネル）
GPU	型番とノードあたり数量	NVIDIA H100 (Hopper) × 4
	1GPU あたり理論演算性能	FP64: 33.5 TFLOPS FP64 (TC): 66.9 TFLOPS FP16, BF16 (TC): 989.4 TFLOPS
	1GPU あたりメモリ	HBM2e 94 GiB, 2,396 GB/s
	ホストとの接続	PCI-Express Gen.5, 総帯域 128GB/s（片方向あたり 64GB/s）
	GPU 間の接続	NVLink × 18/GPU（1 本あたり片方向 25GB/s、総帯域片方向あたり 450GB/s）
ノード数		38
ノード間接続		InfiniBand NDR (400Gbps) × 2 ポート/ノード
冷却方式		水冷
総理論演算性能		CPU FP64: 277.25 TFLOPS GPU FP64: 5.09 PFLOPS GPU FP16, BF16 (TC): 150.39 PFLOPS
総メモリ容量		ホストメモリ 38 TiB デバイスメモリ 14.2 TiB
総理論メモリバンド幅		ホストメモリ 23.35 TB/s デバイスメモリ 364.19 TB/s
ユーザ用ローカルストレージ		NVMe SSD 12.8 TB / ノード

表 5　ノードグループ C の構成

機種名		Supermicro SYS-821GE-TNHR
CPU	型番とノードあたり数量	Intel Xeon Platinum 8480+ (Sapphire Rapids) × 2
	コア数と動作周波数	56 コア, 2.00 GHz - 3.80 GHz
	1CPU あたり理論演算性能	FP64: 3,584 GFLOPS
	1CPU あたりメインメモリ	DDR5 4800 MHz, 4 TiB （256 GiB × 8 枚）
	1CPU あたり理論メモリバンド幅	307.2 GB/s （4800 MHz × 8 Byte × 8 チャネル）
GPU	型番とノードあたり数量	NVIDIA H100 (Hopper) × 8
	1GPU あたり理論演算性能	FP64: 33.5 TFLOPS
		FP64 (TC): 66.9 TFLOPS
		FP16, BF16 (TC): 989.4 TFLOPS
	1GPU あたりメモリ	HBM3 80 GiB, 3,352 GB/s
	ホストとの接続	PCI-Express Gen.5, 総帯域 128GB/s（片方向あたり 64GB/s）
	GPU 間の接続	NVLink × 18/GPU（1 本あたり片方向 25GB/s、総帯域片方向あたり 450GB/s）
ノード数		2
ノード間接続		InfiniBand NDR (400Gbps) × 4 ポート/ノード
冷却方式		水冷
総理論演算性能		CPU FP64: 14.34 TFLOPS
		GPU FP64: 536.00 TFLOPS
		GPU FP16, BF16 (TC): 2140.80 TFLOPS
総メモリ容量		ホストメモリ 16 TiB
		デバイスメモリ 1.28 TiB
総理論メモリバンド幅		ホストメモリ 1.23 TB/s
		デバイスメモリ 53.63 TB/s
ユーザ用ローカルストレージ		NVMe SSD 15.3 TB / ノード