ScaleWorXのHPC/AIソリューション「ScalePOD®」

2023/12/1

株式会社ScaleWorX
吉岡 祐二
yyoshi@scalewx.com

Giga Computing Technology
中村 広志
h.nakamura@gigacomputing.com

about ScaleWorX

ストレージとデータを20年以上扱ってきた視点からのソリューション提供

- DataDirect Networks, Inc. (US) による出資を受け、2021年1月に設立
  - 2008年設立の株式会社データダイレクト・ネットワークス・ジャパン(DDNジャパン)と連携
  - ストレージ製品を含めたトータルなHPC/AIシステムを提供
  - 代表取締役社長兼CEO 山田 昌彦 (2022年9月就任)
- 目指すビジネス
  - 世界の最新テクノロジーを駆使して、コモディティ、オープンソース時代に最適なシステム環境をデザインし、トータルソリューションとして提供できる次世代型のPlatform companyを目指す
- Why ScaleWorX

高度なシステムインテグレーション能力

データ集約型コンピューティングの専門知識

グローバルネットワークとコミニュケーション能力

体制強化とGCTとの戦略的パートナーシップ

- HPC/AI市場向けに本格的なビジネス展開を目指して体制強化
  - 国内で不可欠な設計・構築から運用・保守に至るまで一貫したサポート体制をGCT社と共同で確立
- Giga Computing Technology社 (GCT) とのHPC/AI市場における戦略的提携
  - 多様なHPC/AIワークロードに対応できる多種サーバ製品が提供可能なGCT社と戦略的パートナーシップを締結
  - 国内HPC/AI市場におけるビジネスを共同で推進
- NVIDIA Grace Hopper, Graceサーバ製品を日本市場に先行投入
  - 空冷・水冷関わらず共同で提案を推進
- AMD GPUサーバでも協力

ScalePOD® AI

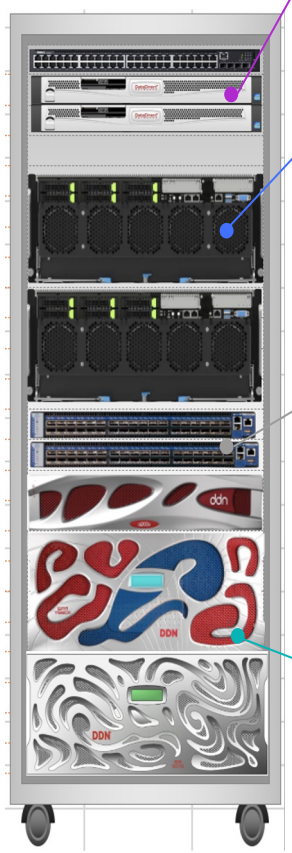
AI用途に最適化された最大1024ノードまで拡張可能なビルディングブロック

- AI業務のライフサイクル・マネジメントに最適なデータセントリック・システム
- 様々な業界におけるITインフラの知見を活かした、シンプルかつ最適化された構成で、システム導入の簡素化と時間短縮に寄与
- 1ラックからスモールスタートして、投資効果を評価してから簡単にスケールアウトできる

最大構成時：Grace Hopper 水冷サーバの場合

- OS：Ubuntu, RHEL
- GPUプラットフォーム：NVIDIA CUDA
- 機械学習ライブラリ：TensorFlow
- 深層学習ライブラリ：PyTorch
- 大規模言語モデル：Llama 2
- グラフィカルな対話操作：Open OnDemand

ScalePOD appliance - Overview



System Management	TOR GbE switch Management server
Computing Node	POD-A：Arm server POD-G：GPU server POD-L：Large Memory server POD-X：Distributed Memory server ※any combination of PODs
Highspeed Network	InfiniBand 200Gbps IB network
Highspeed/Large Storage	550TB highspeed NVMestorage 3M IOPS, 90 GB/sec Read
	2.5 -5 PB general purpose NAS HDD storage
	nexenta NAS, Backup & Archive

NVIDIA NGC Catalog

HPC Collection | End-to-End AI Development | Deep Learning Frameworks

Interactive(k8s), Batch(Slurm)

Cluster Management

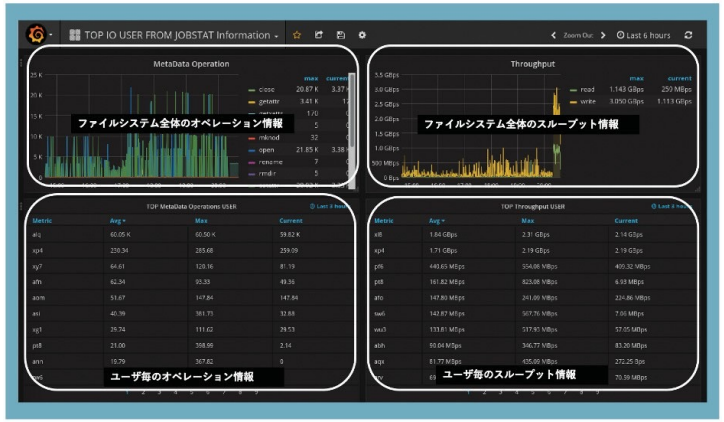
OS(RHEL, ubuntu)

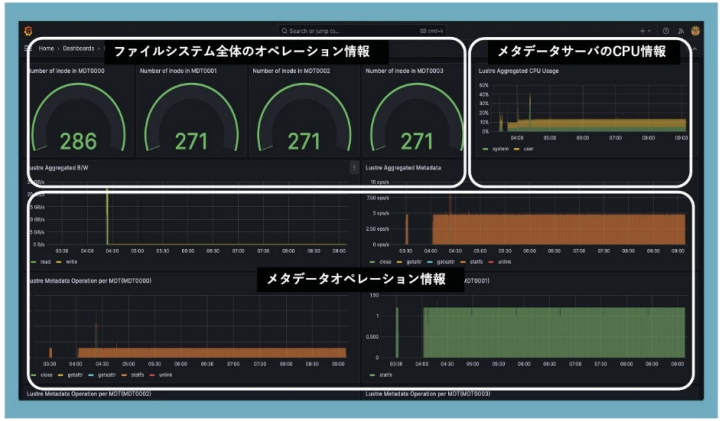
システム最適化ツール：バレルアイ

データ分析I/Oのための詳細なモニタリング製品

トラブルシューティング　サイジング　レポーティング

- 大規模Lustreファイルシステムのモニタリングシステム(ES-Mon) として筑波大学、JAMSTEC、OISTと協力して2013年からDDNジャパンが開発
- Collectd、InfluxDB、VictoriaMetrics、Grafana等のオープンソースを利用
- 2021年からScaleWorXが商品化
- 国内外の大規模な大学、研究機関、製造業を含む100社程度に導入済み
- 個々のアプリケーションやジョブレベルでのI/Oモニタリング
- 詳細なクライアント側のモニタリングも可能





まとめ

データ集約型コンピューティングのための柔軟なインフラソリューション

- クラウドシステムは普及してきているが、ビッグデータを定常的に扱うお客様にはクラウドはまだ高額と認識
- クラウドのような使いやすさや利便性がありつつ、オンプレミスのような経済的なシステムの提供を目指す
- ScalePODが目指すソリューション
  - お客様のアプリケーションとワークフローに合わせたテーラーメイドソリューションを提供
  - ラックレベルソリューションのため、導入期間を数日に短縮
  - 遠隔監視、自動化、プロアクティブサポートによるシステムの効率化・ダウンタイムの最小化または排除

GIGABYTE™

GIGA
COMPUTING

GIGABYTEは、1986年に設立、今年で37周年を迎えます。台北市内から交通至便な新北市に本社構え、世界各国ワールドワイドにて大展開しています。Giga Computingは2023年1月に新たに設立され、100%サーバーに注力したビジネスを手掛けます。

GIGABYTE並びにGiga Computingは、共にNVIDIA社のファーストティアレベルの協業パートナーとしてワールドワイドでビジネス協業を進めています。HPC・AI需要向けのエンタープライズ市場向けGPU製品では、世界で一番積極的な展開を図ってます。

GIGABYTE™

強み特徴【Advantage】

- Intel・AMD・Arm・NVIDIA等、全てのプラットフォームを1stティアで手掛ける事が強み特徴です
- No.1 Server Vendor promoting all of Processors and GPU.

intel　AMD　arm　NVIDIA

GCT	✓ ＊Rackmount ＊Multi Node ＊GPU Server	✓ ＊Rackmount ＊Multi Node ＊GPU Server	✓ ＊Rackmount ＊Multi Node ＊GPU Server	✓ ＊PCI-Ex Solution ＊SMX5 Solution ＊Grace Platform

- We’re 1st Tier Server Vendors for these Processor Makers.
- No1 Server Vendor to strongly promote AMD EPYC Platform since its launch.
- No1 Server Vendor for NV H100 Certificated Server Platoforms.

GIGA COMPUTING

GIGABYTE™

GPU活用したあらゆるワークロードに採用・応用

- 生成AI用途、LLM大規模言語モデル、あらゆる産業のデジタル化基盤、3DCGコンテンツ等のレンダリング等
- 各種GPUをサポートする多種多様なサーバー製品ラインナップで、ジャストサイズなGPUリソースを提供

Generative AI

Develop new services, insights, and original content.

LLM Training and Inference

Accelerate AI training and inference workloads.

Industrial Digitalization

Create and operate metaverse applications.

Rendering and 3D Graphics

Power high-fidelity creative workflows with NVIDIA RTX™ graphics.

GIGA COMPUTING

GIGABYTE™

業界最高密度2U8GPUサーバー

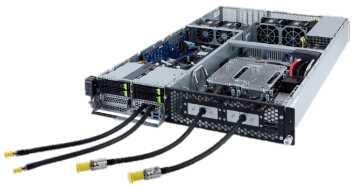
- 1U・2U・3U・4U・5U規格でGPU搭載密度を追求
- 2U8GPUは業界最高密度を誇り、他社の追従を許さないGPU搭載密度

AI & HPC
Universal AI & Graphics
Cloud Gaming

NVIDIA HGX

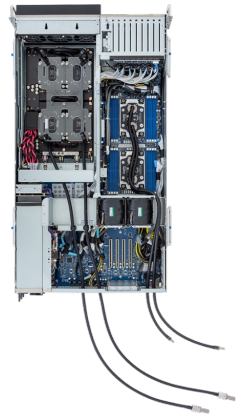
OVX
AI Training
LLM Inferencing
Universal AI & Graphics

NVIDIA HGX



HGX A100
2U4GPU
DLC Type

HGX A100
4U8GPU
DLC Type



HGX H100
3U4GPU
DLC Type

GIGA COMPUTING

GIGABYTE™

アドバンスド・クーリングで最適な冷却方法を提案

- Giga Computingでは、空冷タイプ・水冷タイプ・液浸タイプの異なる冷却ソリューションをスタンダード製品として展開！
- 2U8GPU規格のGPUサーバーから2U4N規格のマルチノードサーバーまで、標準でこれらを展開します！

空冷タイプ

2U8GPU
2U4Node
OCP Server

水冷タイプ

In Rackタイプ
Rack形状タイプ
内部循環タイプ

液浸タイプ

GIGABYTE Brand
冷媒オイル検証
設備工事・設営

GIGA COMPUTING

GIGABYTE™

NVIDIA No.1パートナー

- NVIDIA 認定システムカタログに、H100対応サーバーをいち早く掲載
- 54種類の構成で「Certified」取得し、350種類の構成で「Qualified」取得

Edge | Graphic | Data Analytics | 3D Graphic | HPC | AI

54
Certified

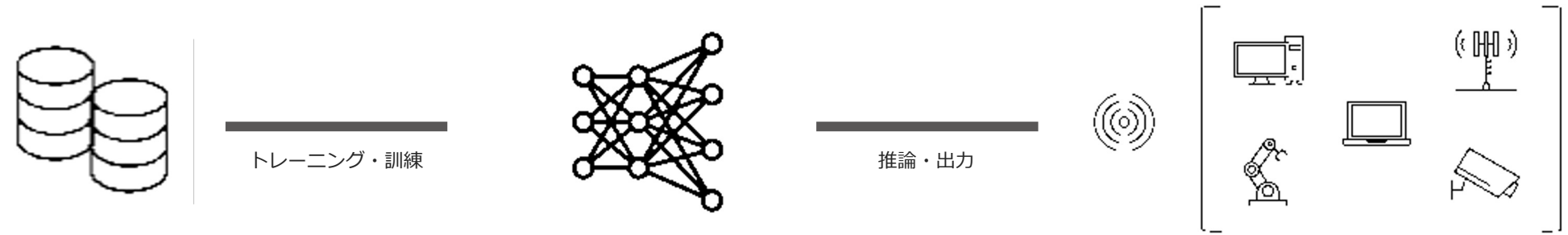
350
Qualified

GIGA COMPUTING

GIGABYTE™

幅広い製品ラインナップでAI需要全般をカバー

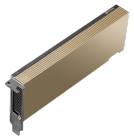
- NVIDIA SXM5タイプH100搭載サーバーやPCI-Exカードタイプ H100対応サーバーを、1U規格～5U規格まで
- 推論・出力向けのNVIDIA L4やL40/L40SといったGPUカードも幅広い製品ラインナップで対応



5U8GPU

SXM5 H100

2U16GPU



L4 * 16

2U8GPU

PCI-Ex H100

2U8GPU

L40/L40S * 8

GIGA COMPUTING

導入事例【海外】

- Living 4.0を推進する財団法人台湾建築センター（TAIWAN ARCHITECTURE & BUILDING CENTER）でのBIMプラットフォーム
- 米国を代表するGPUクラウドプロバイダー・CoreWeave社のGPU基盤

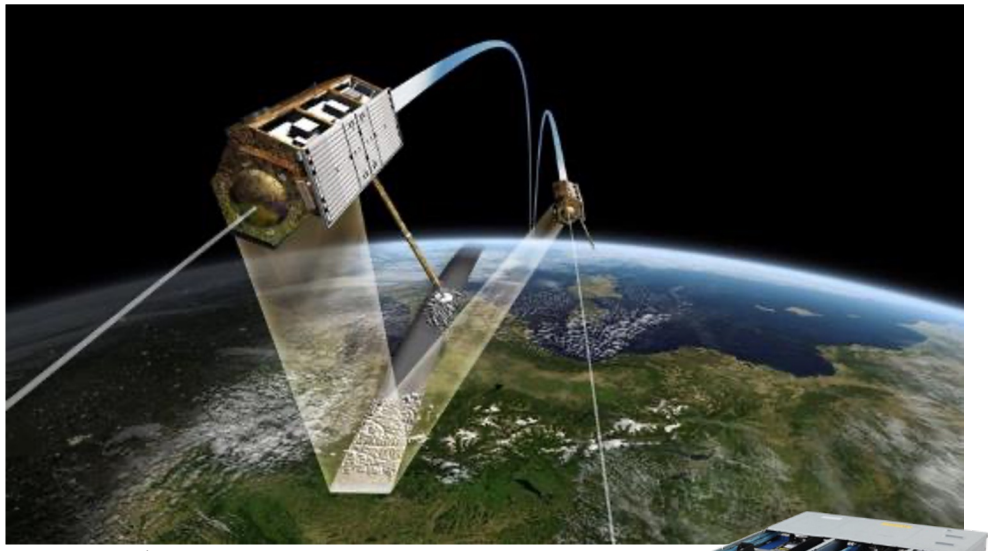
Leading BIM technology revolution in Taiwan
The “Living 4.0 Intelligent Living Space”, slated for a September 2023 launch, aims to advance BIM technology in Taiwan's construction sector. Founded by TABC, the initiative also functions as an educational hub for schools, companies, and research centers. With GIGABYTE's powerful W771-Z00 GPU workstation serving as the virtual host and GIGABYTE BRIX Extreme GB-BEi7HS-126 and GB-BER7HS-5800 managing the connections, the setup enables real-time 3D design collaboration through NVIDIA Omniverse, propelling Taiwan's construction industry into a new generation.

Leading the GPU Cloud Industry with the Latest H100 GPUs
As a cloud computing startup, CoreWeave foresaw the potential of the powerful Nvidia H100 GPU and began deploying H100 GPUs on its cloud computing service much earlier than other companies. With assistance from Nvidia, the company leads the rising trend of GPU computing in cloud services. The heart of the service is GIGABYTE’s G593 servers built with a configuration that uses Intel Xeon CPUs and Nvidia HGX H100 GPUs, creating one of the most powerful and advanced structures designed for large-scale HPC and AI workloads.

GIGABYTE™

導入事例【海外・国内】

- ドイツ連邦共和国の航空宇宙技術を担うドイツ航空宇宙センター（DLR）での2U4N規格・水冷対応サーバー
- 日本を代表する大手ゲーム会社のゲームサービスプラットフォーム基盤



DLR

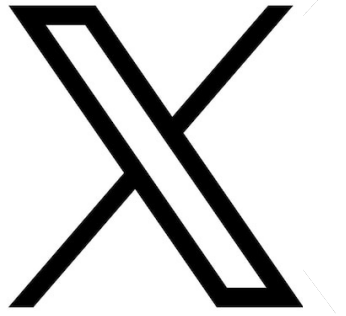
Liquid-Cooled Data Center for Aerospace Research
German National Aerospace Center (DLR) chooses GIGABYTE servers to build their new liquid-cooled data center, to explore new energy sources and develop technologies to protect the environment, and to conduct R&D in aviation, aerospace and transportation.
DLR adopts GIGABYTE H261, a 2U 4 node system combined with a CoolIT's liquid cooling system to achieve extreme compute density and GIGABYTE R281, a 2U system together with AMD EPYC processor and 3x NVIDIA GPUs for virtualization and collaborative computing capabilities in a wide number of research units.

大手ゲーム会社

日本を代表する大手ゲーム会社が手掛けるオンラインゲームのサービスプラットフォーム基盤として採用。人気ゲームコンテンツをマルチタイトル抱える同社の、安定・堅牢が必須なサービスプラットフォーム基盤にAMD EPYC準拠の1U規格シングルソケットタイプサーバーが採用。数十万人が同じ世界でプレイする事もあるこれらのゲームタイトルでは、常にプレイヤーが快適に遊べる環境を提供するためのインフラ構築が必要不可欠で、これらの実現にGiga Computingのサーバーが貢献しています。

GIGA COMPUTING

GIGABYTE™



GIGA COMPUTING


GIGA
COMPUTING


ScaleWorX

Thank you!

www.scalewx.co.jp