NVIDIA A100 Tensor Core GPU

80GB HBM2e

NVIDIA A100 Tensor Core GPU による高速化をもって、AI、データ分析、HPC と言った世界で最も困難な計算に挑むことができます。第3世代 Tensor Core と NVIDIA マルチインスタンス GPU (MIG) テクノロジを利用することであらゆるサイズのワークロードを加速できます。

NVIDIA AMPERE ARCHITECTURE

・大小さまざまなワークロードを高速化
・A100 はすべての GPU ユーティリティを最大化

THIRD-GENERATION TENSOR CORES

・A100 は 312 TFLOPS の ディープラーニングパフォーマンスを発揮
・NVIDIA Volta GPU の 20倍

NEXT-GENERATION NVLINK

・前世代に比べ 2倍のスループット
・SXM4 は NVSwitch と組み合わせ最大 8基、PCIe は NVLink Bridge 最大 2基
・A100 GPU を最大 600 GB/s で相互接続

MULTI-INSTANCE GPU (MIG)

・A100 GPUは最大7つのGPUインスタンスにパーティション化し、ハードウェアレベルで完全に分離
・GPU 使用率を最適化し、すべてのユーザとアプリケーションへのアクセスを拡大

HBM2

・前世代よりも 1.7倍高いメモリ帯域幅
・80 GB HBM2 - DRAM の使用効率を95%に高める

HBM2e

・HBM2 をベースに容量とメモリバス帯域幅を引き上げ

STRUCTURAL SPARSITY

・スパースモデルに対して最大約 2倍のパフォーマンスを提供

SPECIFICATION COMPARISON

Product Name	A100-SXM4	A100-PCIe
Peak FP64	9.7 TF	9.7 TF
Peak FP64 Tensor Core	19.5 TF	19.5 TF
Peak FP32	19.5 TF	19.5 TF
Peak TF32 Tensor Core	156 TF | 312 TF*	156 TF | 312 TF*
Peak BFLOAT16 Tensor Core	312 TF | 624 TF*	312 TF | 624 TF*
Peak FP16 Tensor Core	312 TF | 624 TF*	312 TF | 624 TF*
Peak INT8 Tensor Core	624 TOPS | 1,248 TOPS*	624 TOPS | 1,248 TOPS*
GPU Memory	80 GB HBM2e	80 GB HBM2e
GPU Memory Bandwidth	2,039 GB/s	1,935 GB/s
Interconnect	NVIDIA NVLink 600 GB/s** PCIe Gen4 64 GB/s	NVIDIA NVLink 600 GB/s** PCIe Gen4 64 GB/s
Multi-instance GPUs	Up to 7MIGs @10GB	Up to 7MIGs @10GB
Form Factor	SXM4	PCIe
Max TDP Power	400W	300W

* With sparsity ** SXM GPUs via HGX A100 server boards, PCIe GPUs via NVLink Bridge for up to 2-GPUs

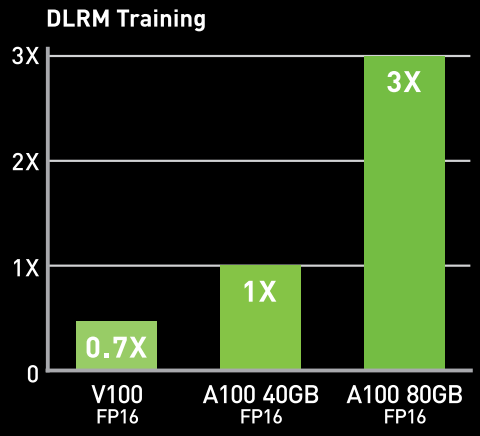
HPC TECH

NVIDIA ELITE PARTNER

A100 40GB vs A100 80GB Benchmark

ディープラーニング トレーニング

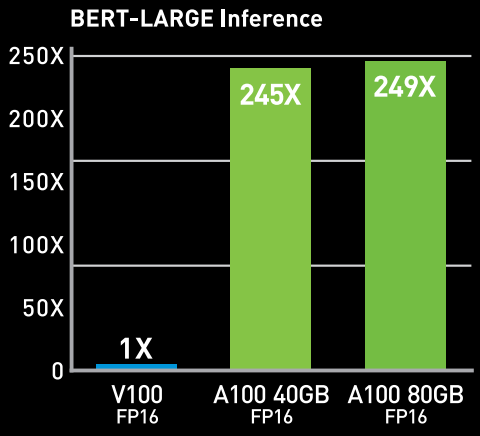
最大級のモデルで最大3倍高速



BERT-Large Inference | CPU only: Dual Xeon Gold 6240 @ 2.60 GHz, precision = FP32, batch size = 128 | V100: NVIDIA TensorRT™ (TRT) 7.2, precision = INT8, batch size = 256 | A100 40GB and 80GB, batch size = 256, precision = INT8 with sparsity.

ディープラーニング推論

CPUと比較して最大 249倍高速

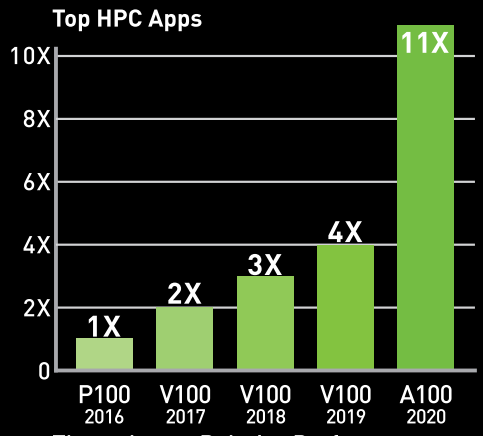


Sequences Per Second - Relative Performance

MLPerf 0.7 RNN-T measured with (1/7) MIG slices. Framework: TensorRT 7.2, dataset = LibriSpeech, precision = FP16.

ハイパフォーマンスコンピューティング

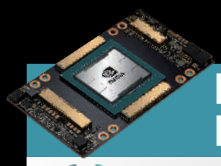
4年間で11倍のHPCパフォーマンス



Throughput - Relative Performance

Geometric mean of application speedups vs. P100: Benchmark application: Amber [PME-Cellulose_NVE], Chroma [szscl21_24_128], GROMACS [ADH Dodec], MILC [Apex Medium], NAMD [stmv_nve_cuda], PyTorch (BERT-Large Fine Tuner], Quantum Espresso [AUSURF 112-jR]; Random Forest FP32 [make_blobs (160000 x 64 : 10)], TensorFlow [ResNet-50], VASP 6 [Si Huge] | GPU node with dual-socket CPUs with 4x NVIDIA P100, V100, or A100 GPUs.

NVIDIA A100 RACKMOUNT SERVER - FLAGSHIP MODEL

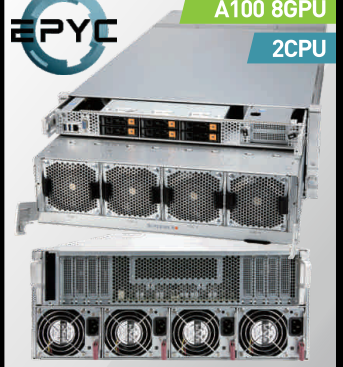


NVIDIA HGX A100 Platform High-End GPU Server

A100 4GPU / 2CPU

Up to 4TB Memory | 4SLOT BAY | 2U Rackmount | 4NVMe U.2 | PCI-Express 4.0

HPCT RS2E32-4GN



A100 8GPU / 2CPU

Up to 4TB Memory | 6SLOT BAY | 4U Rackmount | 6NVMe U.2 | PCI-Express 4.0

HPCT RS4E32-8GN

Single / Dual Root Complex High-End GPU Server

A100 1～4GPU / 1CPU

Up to 1TB Memory | 4SLOT BAY | 2U Rackmount | 2NVMe U.2 | PCI-Express 4.0

HPCT RG2E31-4GP

A100 1～8GPU / 2CPU

Up to 4TB Memory | 24SLOT BAY | 4U Rackmount | 4NVMe U.2 | PCI-Express 4.0

HPCT RS4E32-8GP

	HPCT RS2E32-4GN	HPCT RS4E32-8GN		HPCT RG2E31-4GP	HPCT RS4E32-8GP
	SXM4 モデル例 1	SXM4 モデル例 2		PCIe モデル例 1	PCIe モデル例 2
GPU	NVIDIA A100 SXM4 x4 40GB HBM2	NVIDIA A100 SXM4 x8 40GB HBM2	GPU	NVIDIA A100 PCIe x1 40GB HBM2	NVIDIA A100 PCIe x1 40GB HBM2
CPU	AMD EPYC 7763 x2 (2.45GHz, 64Core) Total128Core	AMD EPYC 7763 x2 (2.45GHz, 64Core) Total128Core	CPU	AMD EPYC 7713P x1 (2.0GHz, 64Core)	AMD EPYC 7763 x2 (2.45GHz, 64Core) Total128Core
RAM	1024GB DDR4-3200	2048GB DDR4-3200	RAM	512GB DDR4-3200	512GB DDR4-3200
SSD	1.92TB NVMe PCIe4.0 x1	1.92TB NVMe PCIe4.0 x1	SSD	960GB SATA 6Gbp/s x1	1.92TB NVMe PCIe4.0 x1
HDD	-	-	HDD	2TB SATA 6Gbp/s x1	-
OS	Linux	Linux	OS	Linux	Linux

上記モデルは一例です。お客様の用途に合うようカスタマイズいたします。

NVIDIA エリートパートナー　BrightComputing 正規代理店　A2ZEON 日本総代理店　ANSYS Discovery Live 代理店

株式会社 HPCテック
https://www.hpctech.co.jp
〒103-0006 東京都中央区日本橋富沢町 7-13 洋和ビル４F TEL:03-5643-2681 FAX:03-5643-2682 MAIL:info@hpctech.co.jp