

**inspur**

# NF5888M5 (AGX-5)のご提案



## Inspur NF5468M5

最も強力で柔軟性なGPUクラウド・サーバーです  
AIクラウドコンピューティング、オフライントレーニング、オンライン推論、動画加速の応用に適している

### 高度な計算能力

インテル®Xeon®2個の拡張型プロセサ、3UPIインターコネクション、TDP 205W、8ブロック全高全長ダブルワイドGPUカードまたは8枚NVLink GPUカードまたは16枚の半高半長のワンワイドGPUカード

### 最適なトポロジ構成

common、balance、cascadeトポロジをサポート、ソフトウェア切替トポロジをサポート、NVLinkインターコネクト技術をサポート、GPU peer-to-peerをサポート、Tesla P4カードアップ通信帯域収束比4:1をサポートする

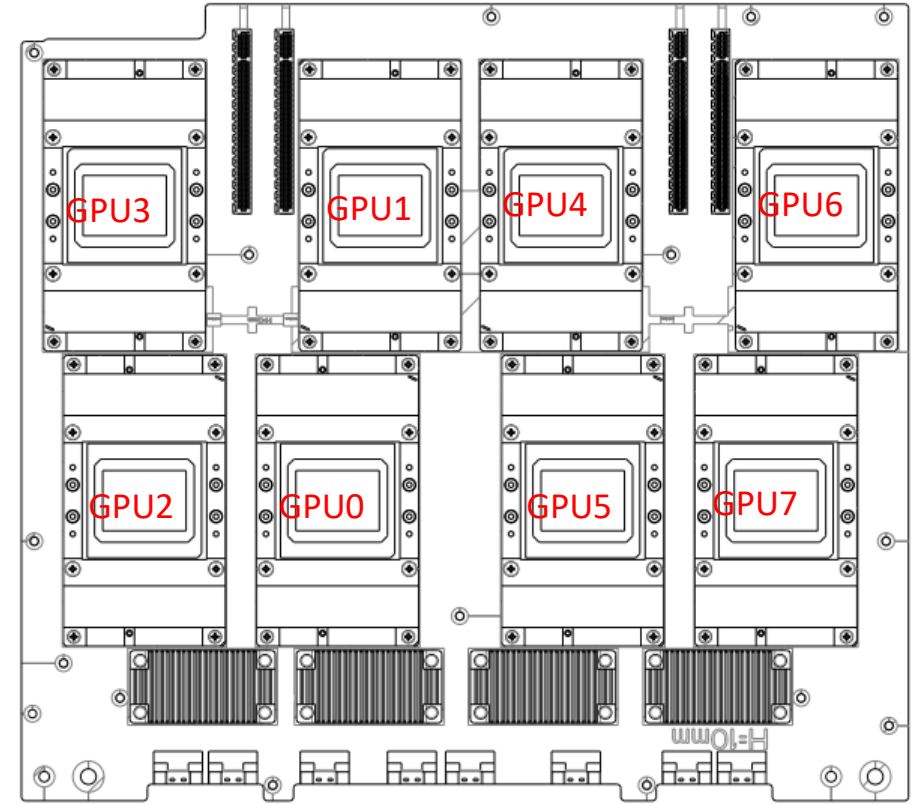
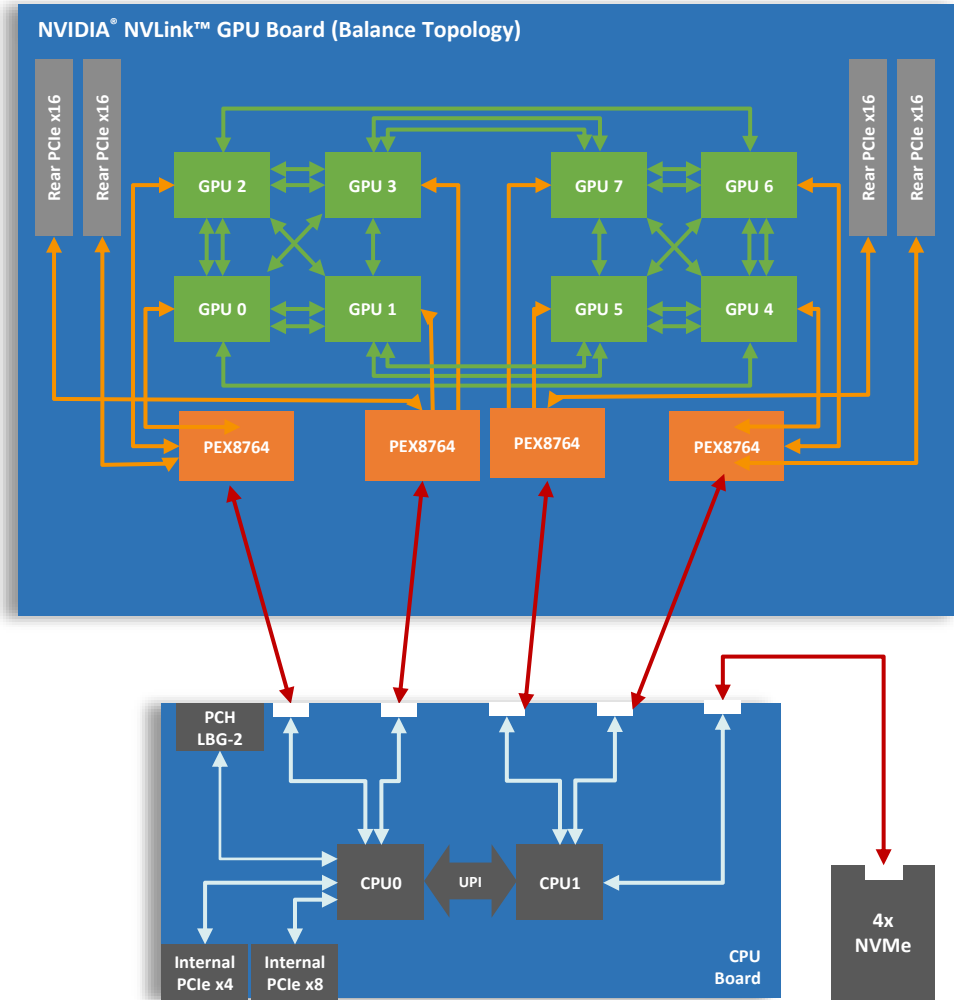
### 高速なIO拡張

PCIEx16スロットを背面に4つ搭載、100Gのリモートダイレクトメモリアクセス(RDMA)が可能な高速ネットワークをサポート、GPUDirect RDMAをサポートする



# NF5468M5-S Topology(SXM2 Switch無し)

NVIDIA® NVLink™ 50GB/s  
Fixed PCIe Gen3  
Flexible PCIe Gen3



NVLink GPU基盤



Inspur NF5888M5(AGX-5)は  
最速AIスーパーコンピューティングプラットフォーム  
現代AIとディープラーニング需要の拡張難題を解決するため  
、データセンターの大規模拡張とエンタープライズ企業のAI  
インフラ拡張に向けて設計した

## 世界最速の計算能力

インテル®Xeon®2個の拡張型プロセッサ、  
3UPIインターコネクト、TDP 205W;  
スタンドアロンで16個のSXM3を搭載可能  
Tesla®V100 32GB若しくは  
次世代のより強力なGPUアクセラレーター  
計算性能は最高2PFLOPS

## 全チップ間的高速インター コネクト

NVIDIA最新のHGX-2™プラットフォームに基づいて、業界で最先端の  
NVSwitch™連携構造、48経路の実現  
2.4TB/sの全チップ間的高速接続

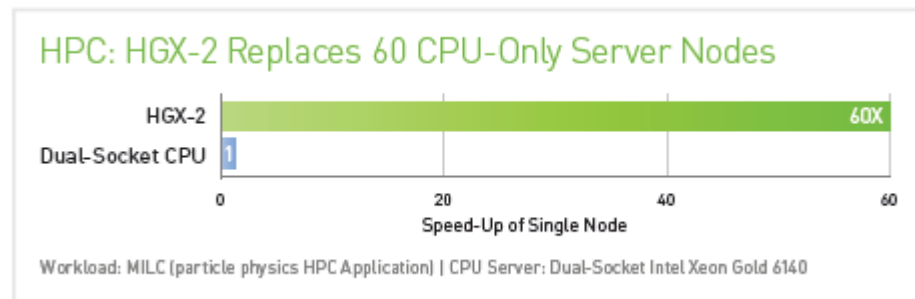
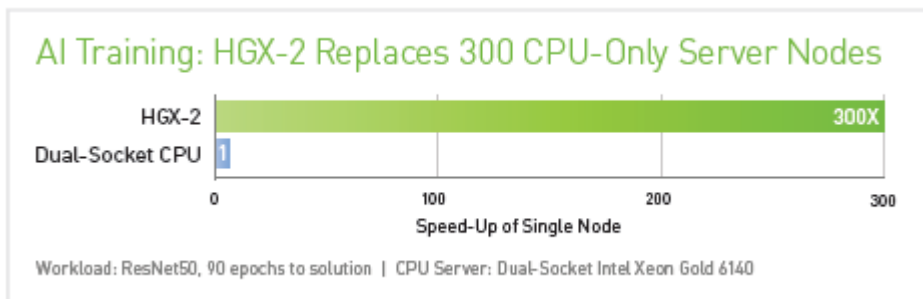
## データ・スループットが全面的に 向上

512GB HBM2グローバルに共有された超高速グラフィックキャッシュ  
3TBの永続メモリは、ハイパーデータへの  
高速アクセスを提供する



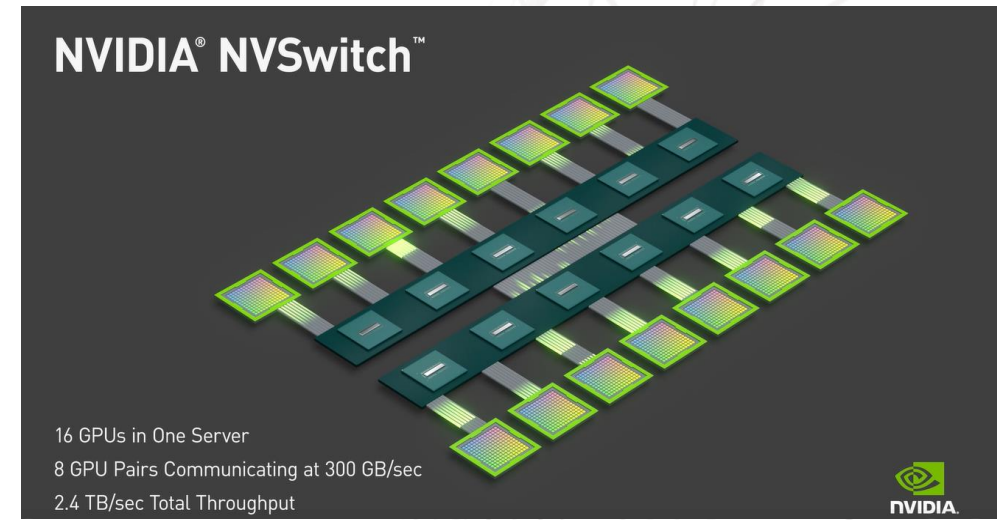
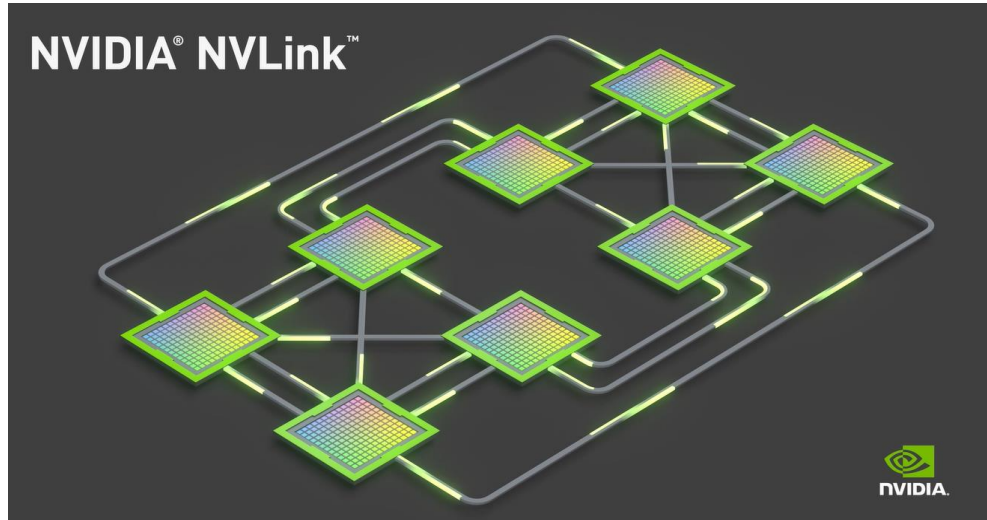
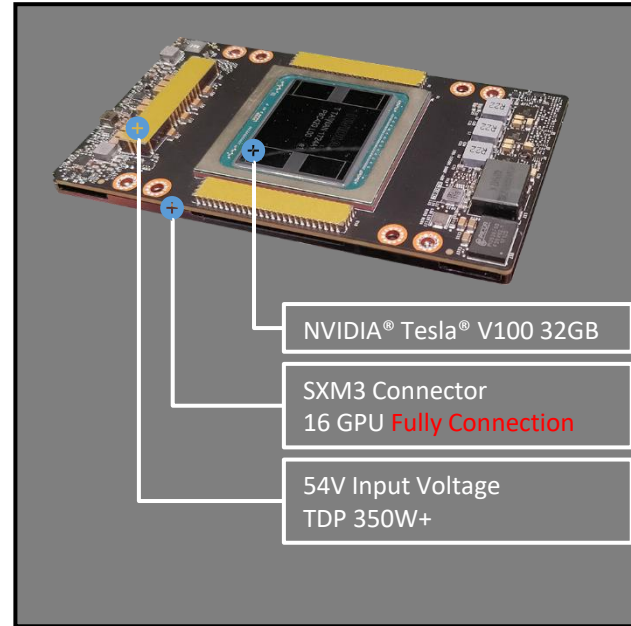
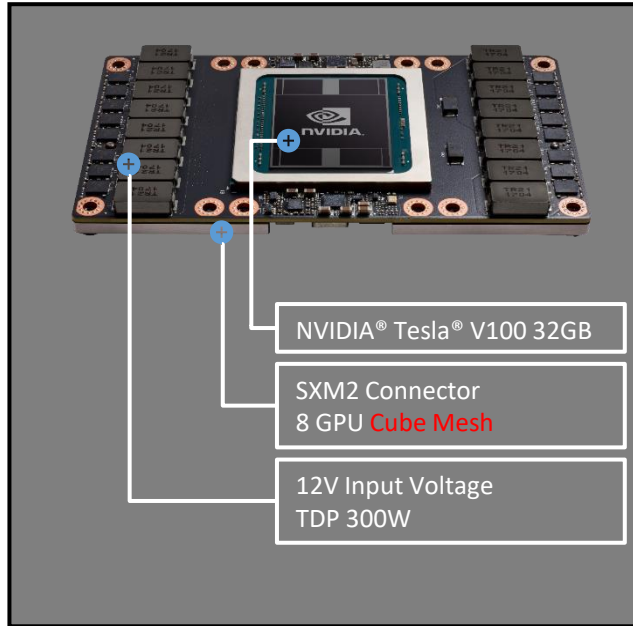
# 製品の位置づけ

番号	名称	説明	応用場面	顧客の悩みと特性価格	顧客ニーズ
1	AIトレーニング	スタンドアロン <b>最速</b> 性能のスーパーサーバ NVlink 2.0に対応している 計算性能は2PFLOPSである	プロAIトレーニング	現代AIとディープラーニング需要の拡張問題を解決し、TCOを向上させる	データセンターの大規模拡張顧客とエンタープライズ企業向けのAIインフラ拡張顧客
2	HPC計算	スタンドアロン <b>最速</b> 性能のスーパーサーバ NVlink 2.0に対応している 計算性能は2PFLOPSである	HPC計算	HPCの複雑なWorkload計算の難題を解決し、TCOを向上させる	HPCは複雑なアプリケーション計算を要求される顧客向け





# SXM2からSXM3





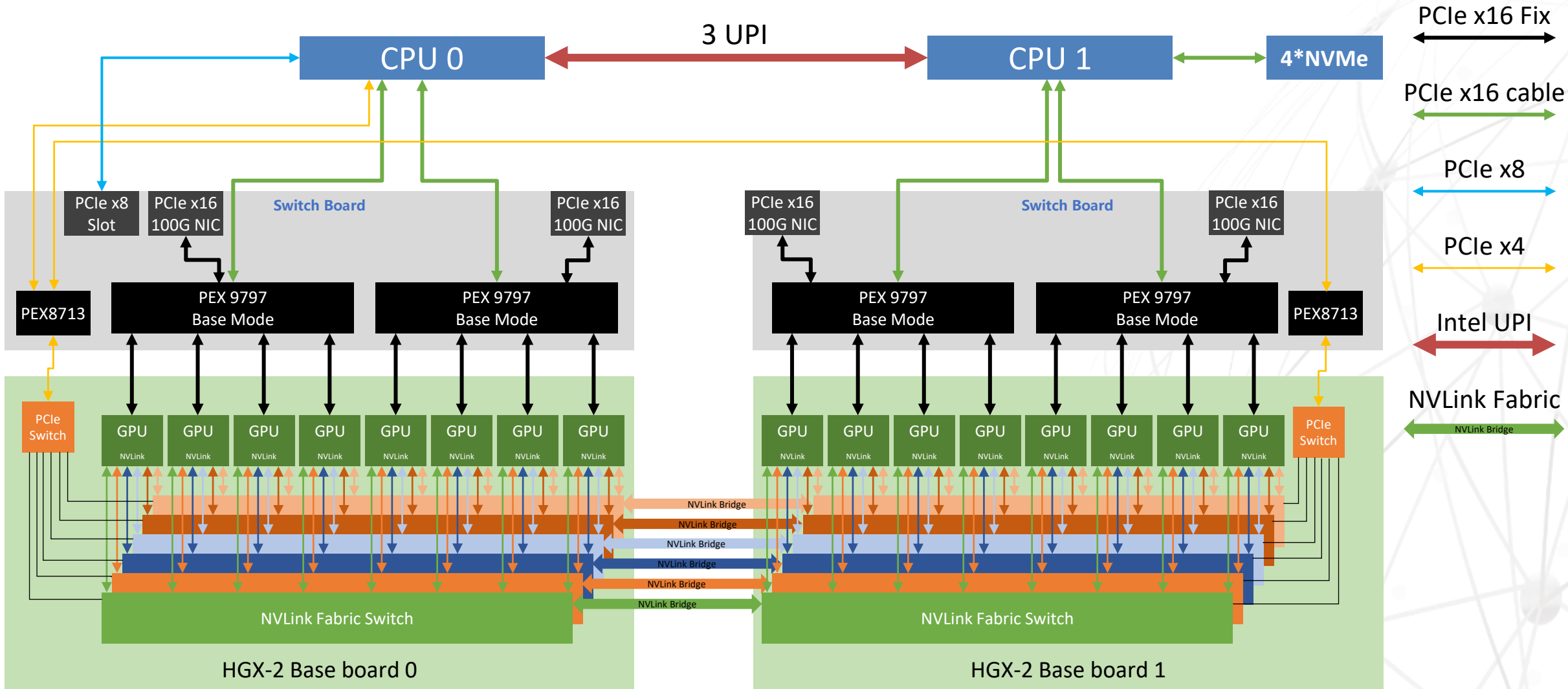
# NF5888M5 SKU 1

- **SKU 1 (First priority 量産支援)**
  - 2 Socket SKL Processor with 3UPI
  - 24DIMMs
  - 2\*HGX-2 Baseboard with 16 Volta GPU
  - NVLink Fabric Bridge enabled
  - 4\*PCIe x16 for 100G NIC





# NF5888M5 SKU 1 System Topology





# 競合規格比較



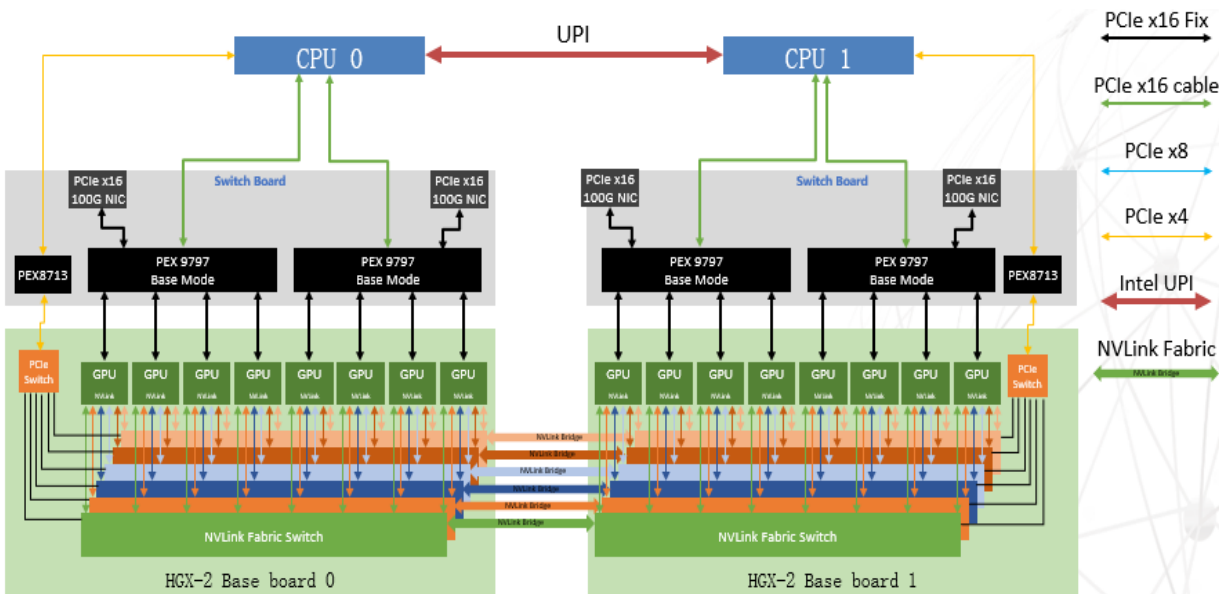
	メーカー	Inspur	N社	S社
基本システム情報	メーカー	Inspur	N社	S社
	型番	NF5888M5 (AGX-5)	DGX-2	9029GP-TNVRT
	フォームファクター	8U	10U	10U
トポロジー	トポロジー	CPUからGPU間1 PCIE Switch hop	CPUからGPU間2 PCIE Switch hops	CPUからGPU間2 PCIE Switch hops
CPU	プロセッサ	インテル®Xeon®の拡張型プロセッサ Up to 2*28 core CPUs	インテル®Xeon®の拡張型プロセッサ Up to 2*816(24cores)	インテル®Xeon®の拡張型プロセッサ Up to 2*28 core CPUs
	メモリスロット	24	24	24
GPU	GPU	SXM3 V100 *16	SXM3 V100 *16	SXM3 V100 *16
拡張 I/O	拡張 PCIe スロット	5*25W low-profile PCIe cards (4* x16, 1* x8)	8 PCI-E x16	16 PCI-E 3.0 x16 for RDMA via IB EDR; 2 PCI-E 3.0 x16 on board
メモリ	HDDスロット	前面HDD: 8x2.5"SATA又は 4x2.5"SATA+4xNVME 内部HDD:2xM.2	10xNVME	16 Hot-swap 2.5" NVMe drives, 6 Hot-swap 2.5" SATA3 drive bays
	NVMe(U.2)	最大4*NVMe	最大10*NVMe	16*NVMe
ネットワーク インターフェース	ネットワーク	4x100Gb/s IBカード ;1x50/25/10G NIC / NVMe Onboard ダブル10Gb/s Ethernet	8x100Gb/s Iカード ダブル10/25Gb/s Ethernet	Dual Port 10GBase-T from Intel X540 Ethernet Controller
電源	電源	3000w (2+2) *2冗長	3000W*6非冗長	3000W*6非冗長
サポート動作環境	動作温度	10° C-35°C	5° C-35° C	10° C-35°C



# 競合品のトポロジー比較分析

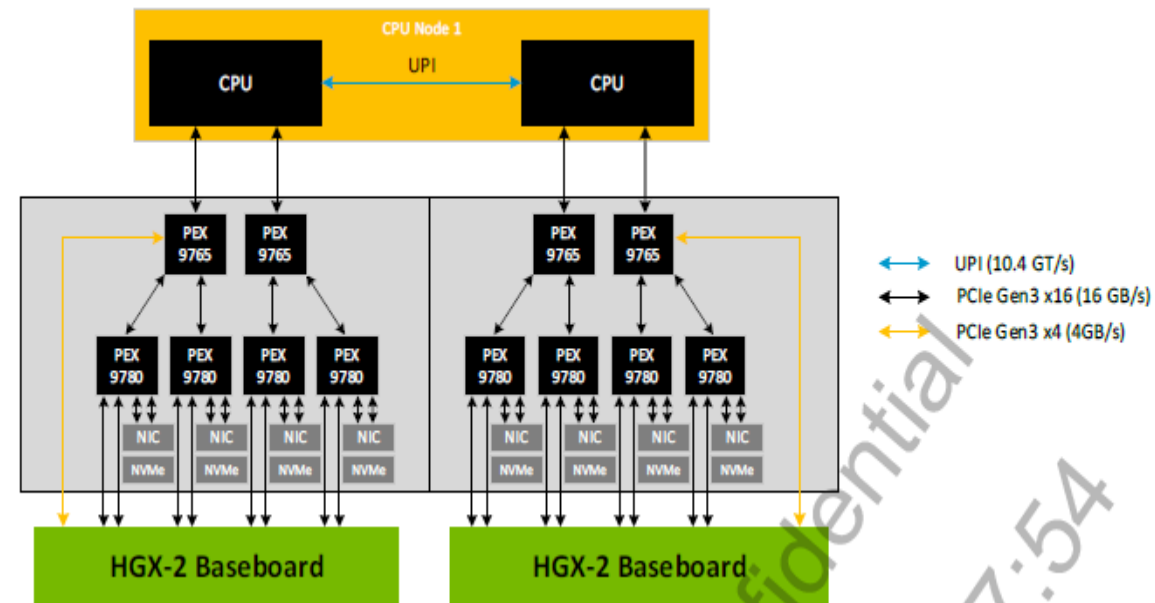
### AGX-5

One PCIE switches hop  
CPUからGPUまでの遅延を低減できる  
PCIE switches数



### DGX-2/9029GP-TNVRT

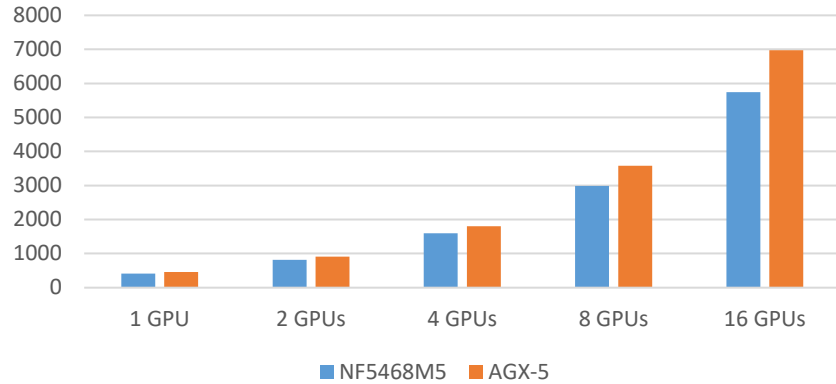
Two Cascaded PCIE switches hops





# アプリケーション性能例1

AGX-5 vs NF5468M5x2



画像分類評価場面では,Imagenetデータセットに基づいて TensorFlow+horovodのVGG-16モデルの評価を行った。AGX-5は16枚のNVLink Switchで相互接続されたV100を持ち,2台のシングル8カード(1台あたり8枚のNVLink V100)で2枚100GのInfinibandで相互接続されたGPUでクラスタを計算するのに対し,AGX-5の評価速度は2台のNF5468M5の**122%**に達する。

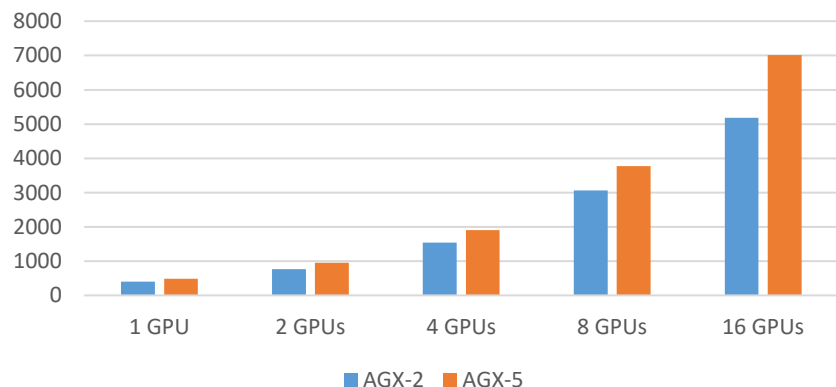
TensorFlow 1.12		
Batch Sizes=256, synthetic(images/sec)	性能単位(images/s)	
vgg16	NF5468M5	AGX-5
1 GPU	407.95	456.81
2 GPUs	810.23	907.39
4 GPUs	1590.63	1805.03
8 GPUs	2982.6	3581.96
16 GPUs	5742.6	6976.06

テスト環境		
Server	Inspur AGX-5	NF5468M5
CPU	Intel(R) Xeon(R) 8168 CPU @ 2.7GHz 24C	Intel(R) Xeon(R) 8168 CPU @ 2.7GHz 24C
memory	12*32GB-2666Mhz	12*32GB-2666Mhz
TensorFlow	1.12	1.12
GPU	NVIDIA Tesla SXM3 V100 32GB	NVIDIA Tesla SXM2 V100 32GB



# アプリケーション性能例2

AGX-5 vs AGX-2 x2



画像分類評価場面では,Imagenetデータセットに基づいてTensorFlow+horovodのVGG-16モデルの評価を行った。AGX-5は16枚のNVLink Switchで相互接続されたV100を持ち,2台のシングル8カード(1台あたり8枚のNVLink V100)で100GのInfinibandで相互接続されたGPUでクラスタを計算するのに対し,AGX-5の評価速度は2台のAGX-2の**135%**に達する。

## TensorFlow 1.12

Batch Sizes=64,  
synthetic(images/sec)

性能単位(images/s)

	AGX-2	AGX-5
vgg16		
1 GPU	406.8	488.22
2 GPUs	773.3	959.21
4 GPUs	1547.7	1904.33
8 GPUs	3067	3769.99
16 GPUs	5187.2	7009.34

## テスト環境

Server	Inspur AGX-5	Inspur AGX-2
CPU	Intel(R) Xeon(R) 8168 CPU @ 2.7GHz 24C	Intel(R) Xeon(R) 8168 CPU @ 2.7GHz 24C
memory	12*32GB-2666Mhz	12*32GB-2666Mhz
TensorFlow	1.12	1.12
GPU	NVIDIA Tesla SXM3 V100 32GB	NVIDIA Tesla SXM2 V100 32GB

*inspur*

**THANK YOU**

