How to Utilize multi-GPU system better for AI model training

AIモデルのトレーニング、マルチGPUシステムの活用方法

Jerry Huang,シニアディレクター

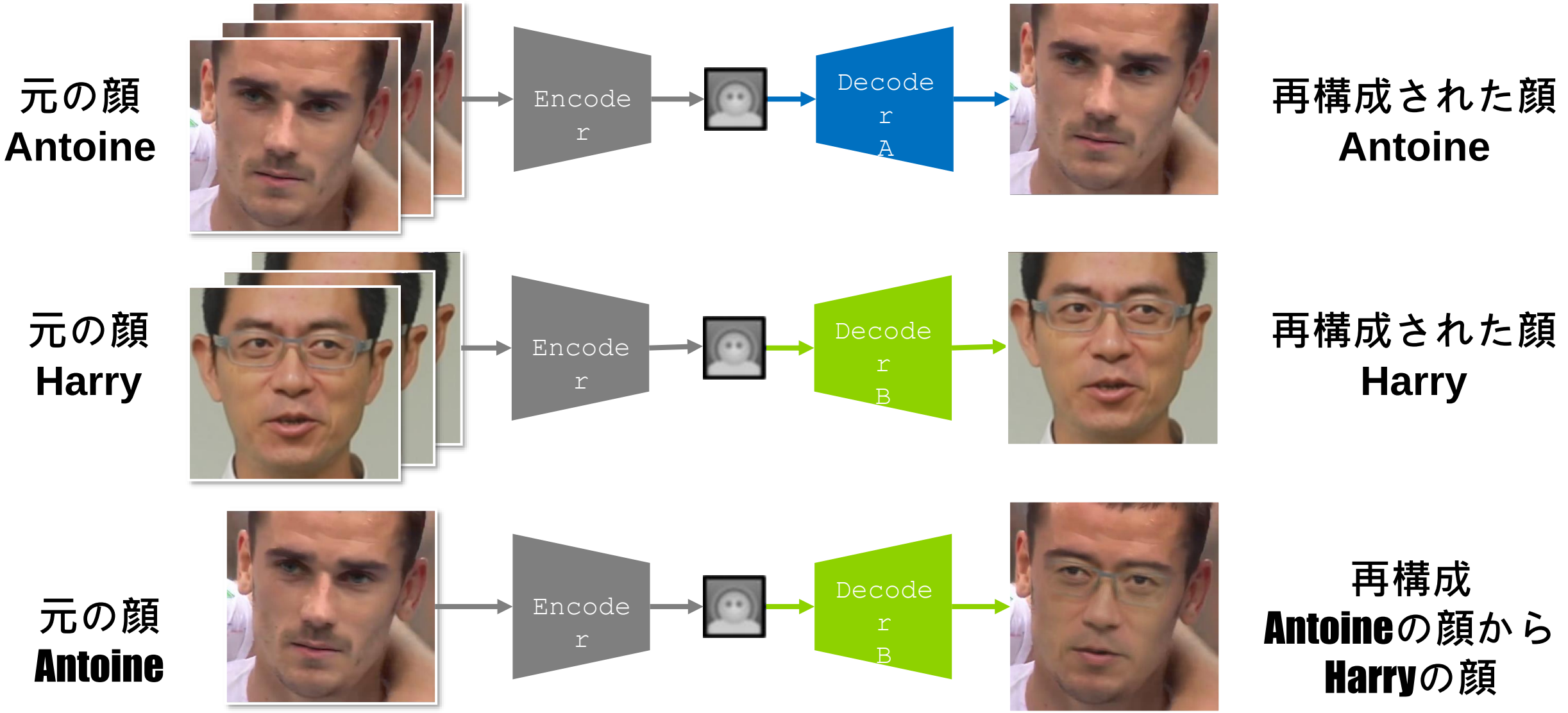
2018/10/17

wiwynn®

Faceswap – マルチGPU構成の例

Faceswap – マルチGPU構成の例

Faceswap – マルチGPU構成の例



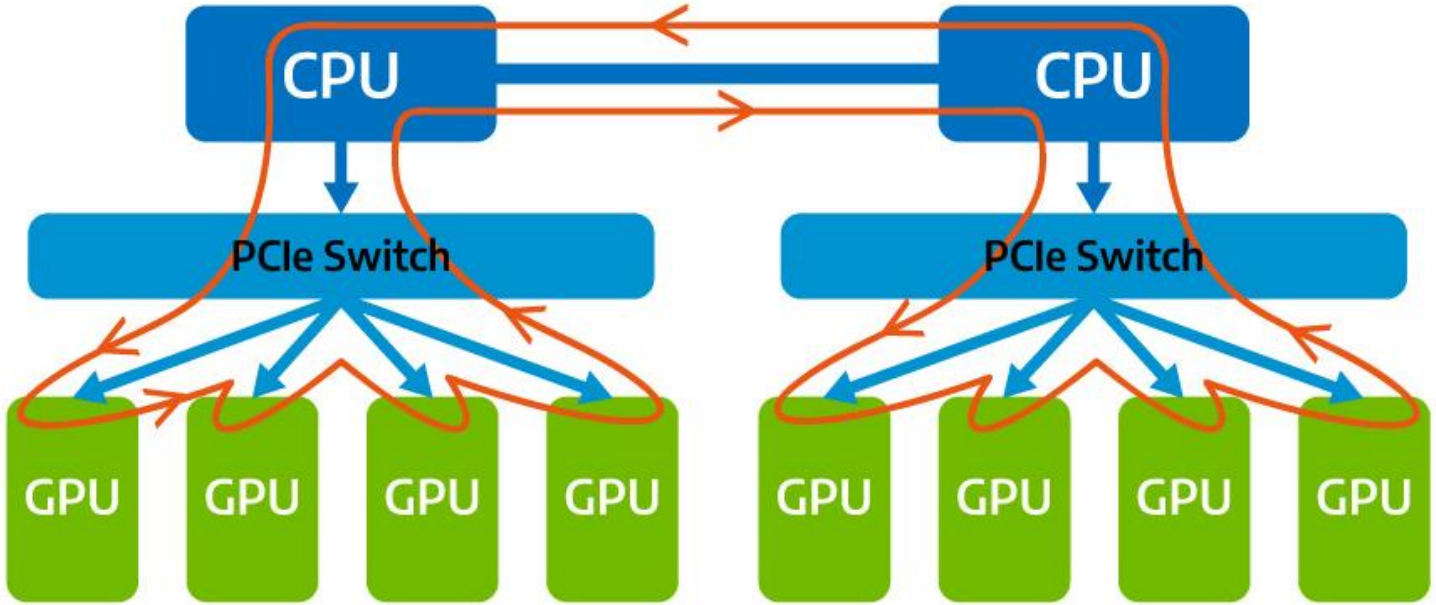
AIトレーニングを行う方法

- 1人につき10万枚以上の写真
- 1人につき12～15時間
- トレーニング時間を短縮する方法

複数のGPUアクセラレータを展開する

機械学習用のマルチGPU

- データの並列性
  - 同じモデルを各GPUカードに入れますが、データの異なる部分でフィードします。



Wiwynn GPU ソリューション

必要ですか？

・21インチ製品

・19インチ製品

・更なるGPU

SV7400シリーズ
4U8G GPUサーバー

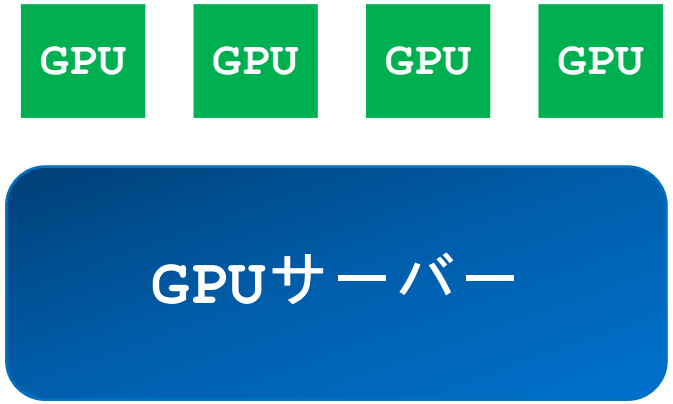
SV500シリーズ
4U8G GPUサーバー

XC200シリーズ
4U16X GPUアクセラレータ

アプリケーションのワークロード

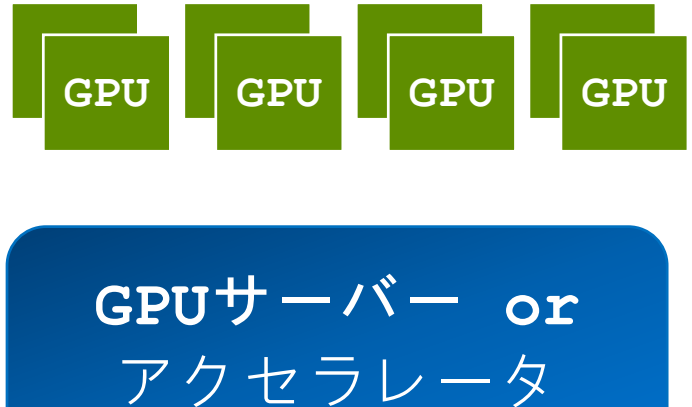
- ワークロードは計算能力 **(GPU数)** に依存する

ワークロード **1:4**



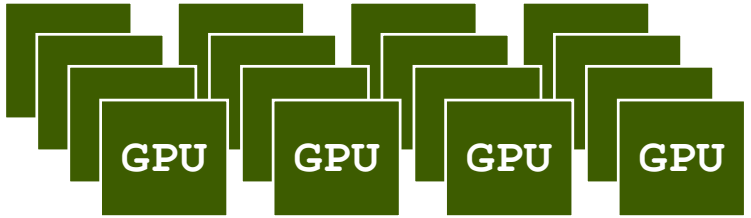
サーバー: GPU = 1 : 4

ワークロード **1:8**



サーバー: GPU = 1 : 8

ワークロード **1:16**



サーバー: GPU = 1 : 16

システム構成

Mini SAS HDケーブル(PCIe x16)

XC200 (4U16X GPU

PCIe スイッチ	Broadcom PEX9797
GPU	NVIDIA V100 *16

SV300G3 (1U Multi-purpose

CPU	Intel Xeon Gold 6130 *2
メモリ	DDR4 2400MHz 32GB *16
SSD	Intel® SSD DC S3500シリーズ480GB *1

XC200シリーズ

19インチ 4U16X GPU アクセラレータ

Disaggregated PCIeアクセラレータ

16 PCIe 3.0 x 16アドインカード

最大4台のサーバー接続用の柔軟な構成

工具不要でメンテナンスが簡単な4つのドロア

ワークロード1:4、1:8、1:16をサポート

4U; 176 (H) x 448 (W) x 900 (D) mm

ソフトウェアスタック

Dockerイメージ

アプリケーション: Faceswap

フレームワーク: Keras, Tensorflow, Horovod

Library
- numpy
- Math Kernel Library
- scipy
- Scikit-learn
- cudnn
- mpi4py
- OpenCV
- NCCL

Parallel Computing Toolkits
- CUDA
- OpenMPI

Tools
- jupyter notebook
- Tensor Fusion

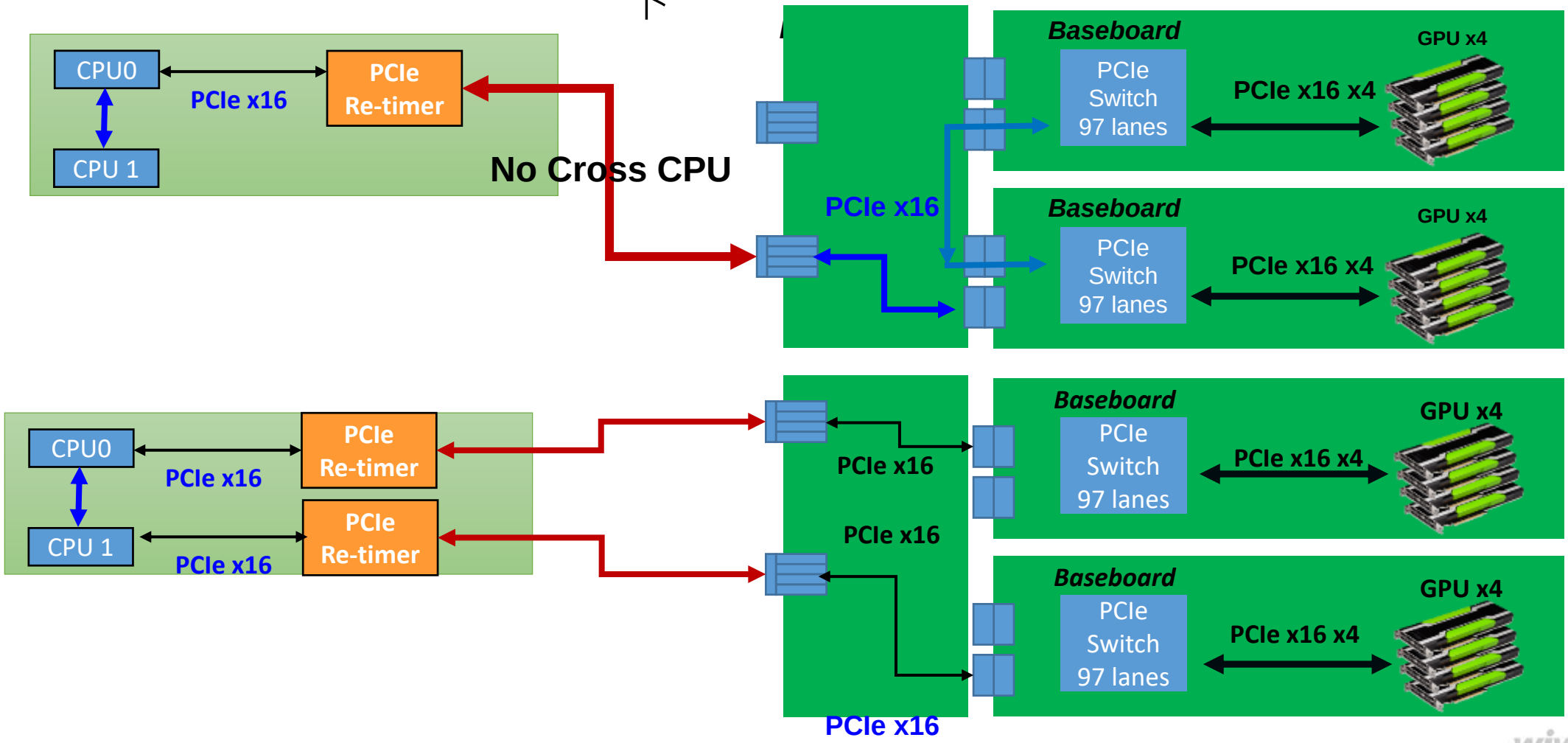
Nv-docker

OS (Linux /64 ビット/カーネル4.4)

ドライバ: GPU GPU GPU GPU GPU GPU GPU GPU

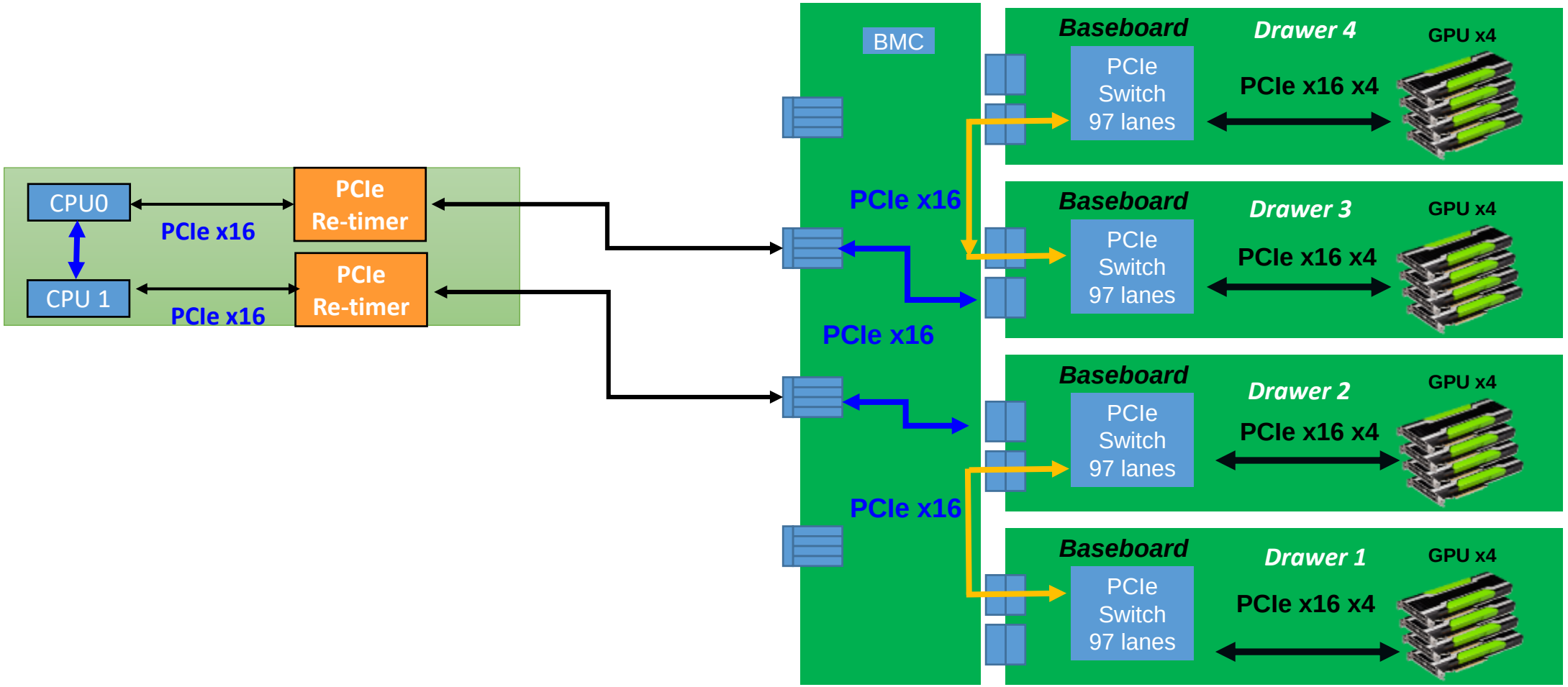
アプリケーションのワークロード–1:8

各ノードに専用の8つのGPUアクセラレータカード



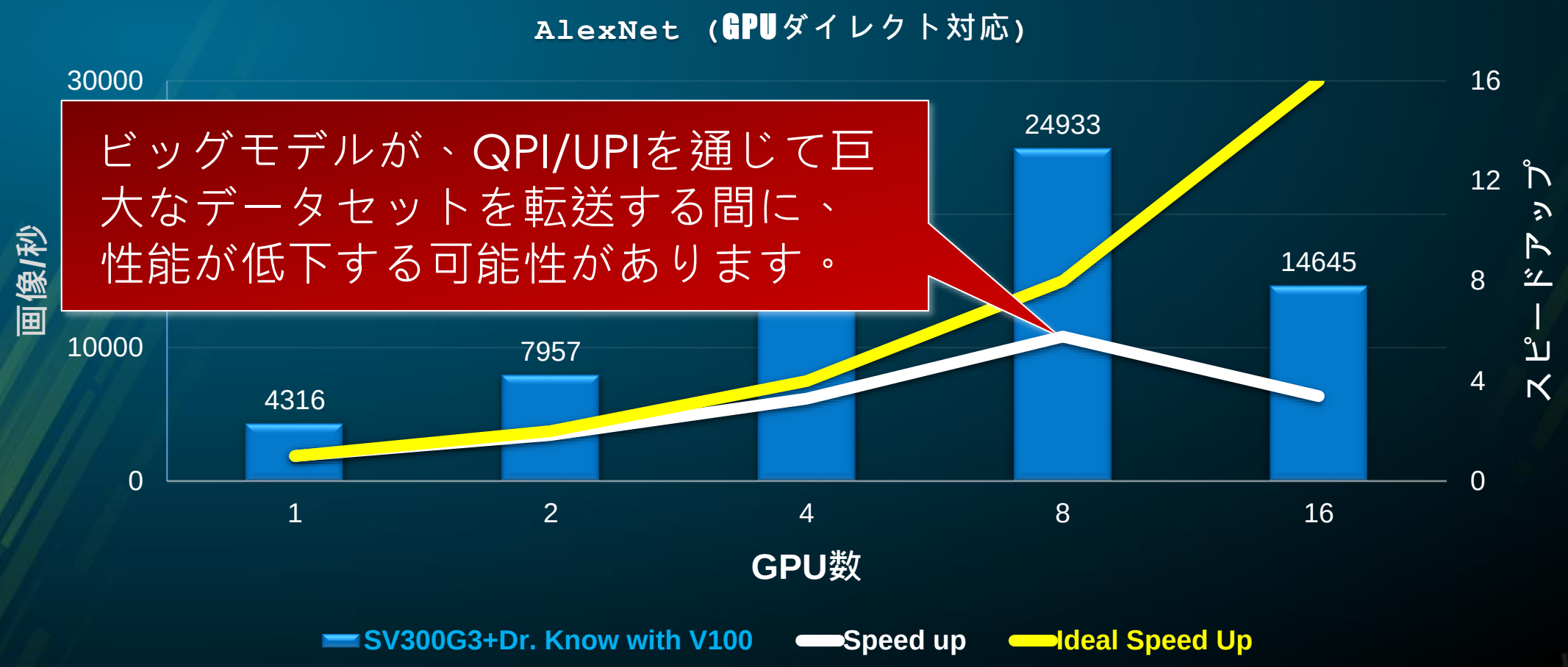
アプリケーションのワークロード–1:16

各ノードに専用の16つのGPUアクセラレータカード



ベンチマーク

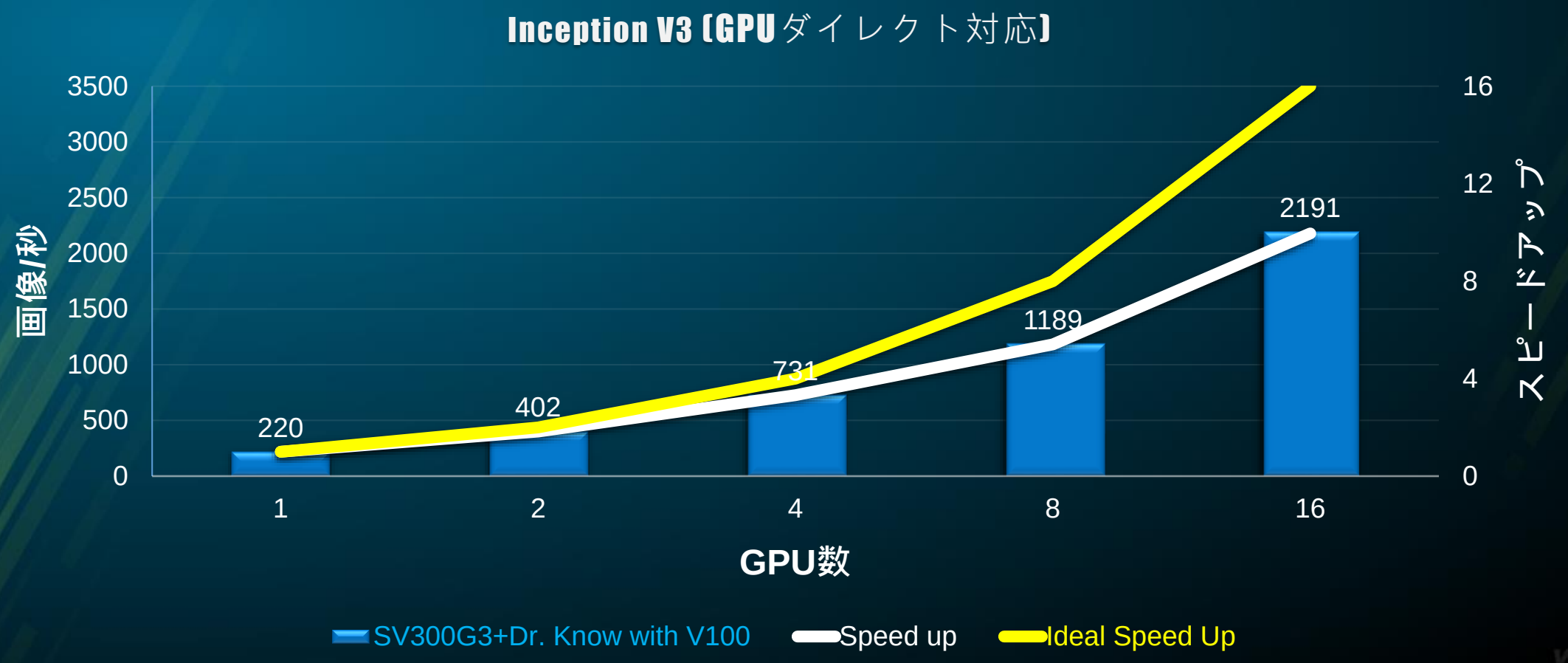
2倍のGPUを使用している間、コンピューティング能力は、約1.7～1.8倍になる可能性があります。



ビッグモデルが、QPI/UPIを通じて巨大なデータセットを転送する間に、性能が低下する可能性があります。

ベンチマーク

CPUをまたぐ、良好な性能を実現したい場合は、より小さいモデルやパラメータサーバーモードを使用することをお勧めします。



White Paper

wiwynn®

Accelerating Your AI/Deep Learning Model Training with Multiple GPUs

White Paper

データセンターのワークロードに最適化されたITソリューション
と最高のTCOを提供します。