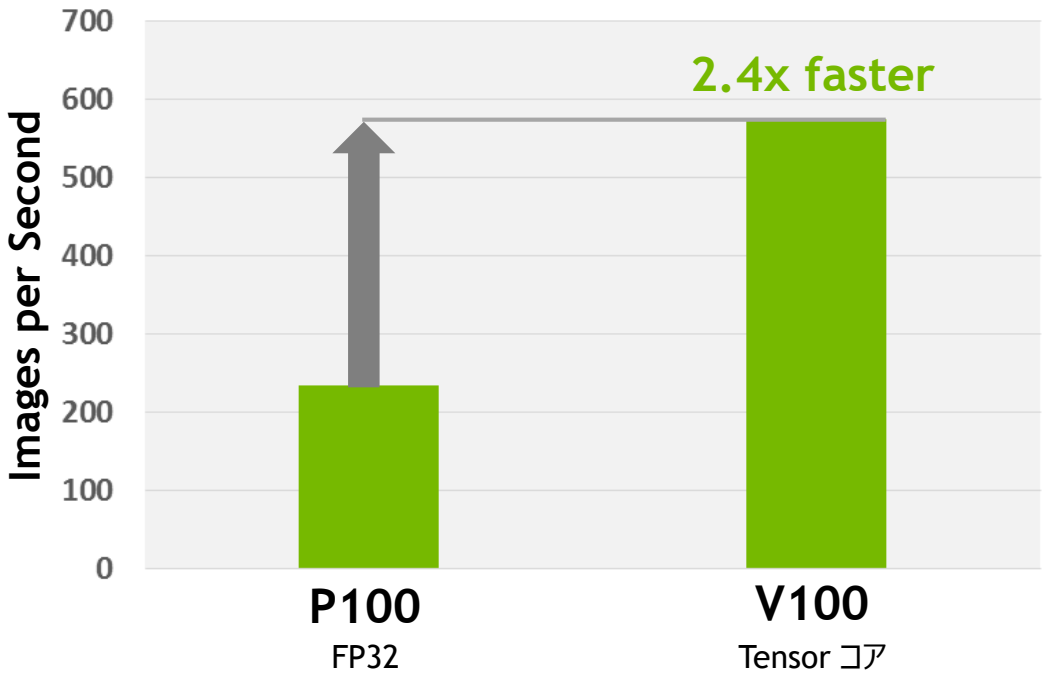
NVIDIA TESLA V100
CUDA 9 のご紹介

森野慎也, シニアソリューションアーキテクト (GPU-Computing)
NVIDIA

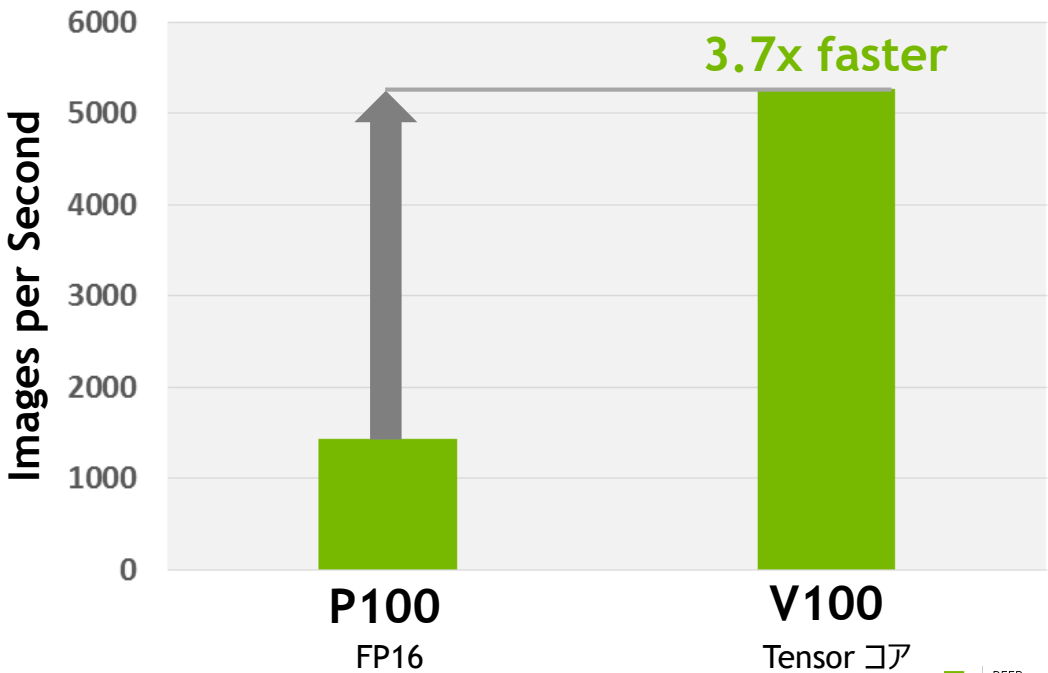
VOLTA: ディープラーニングにおける大きな飛躍

ResNet-50 トレーニング



ResNet-50 推論

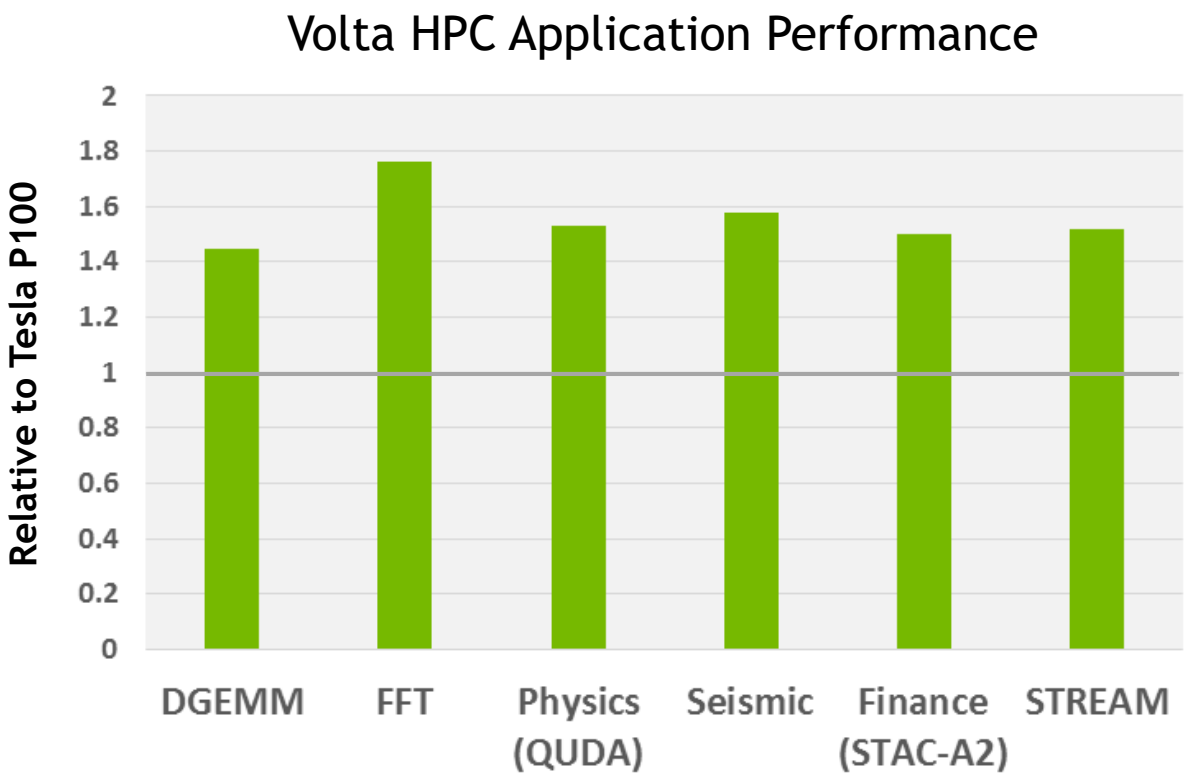
TensorRT - 7ms レイテンシ



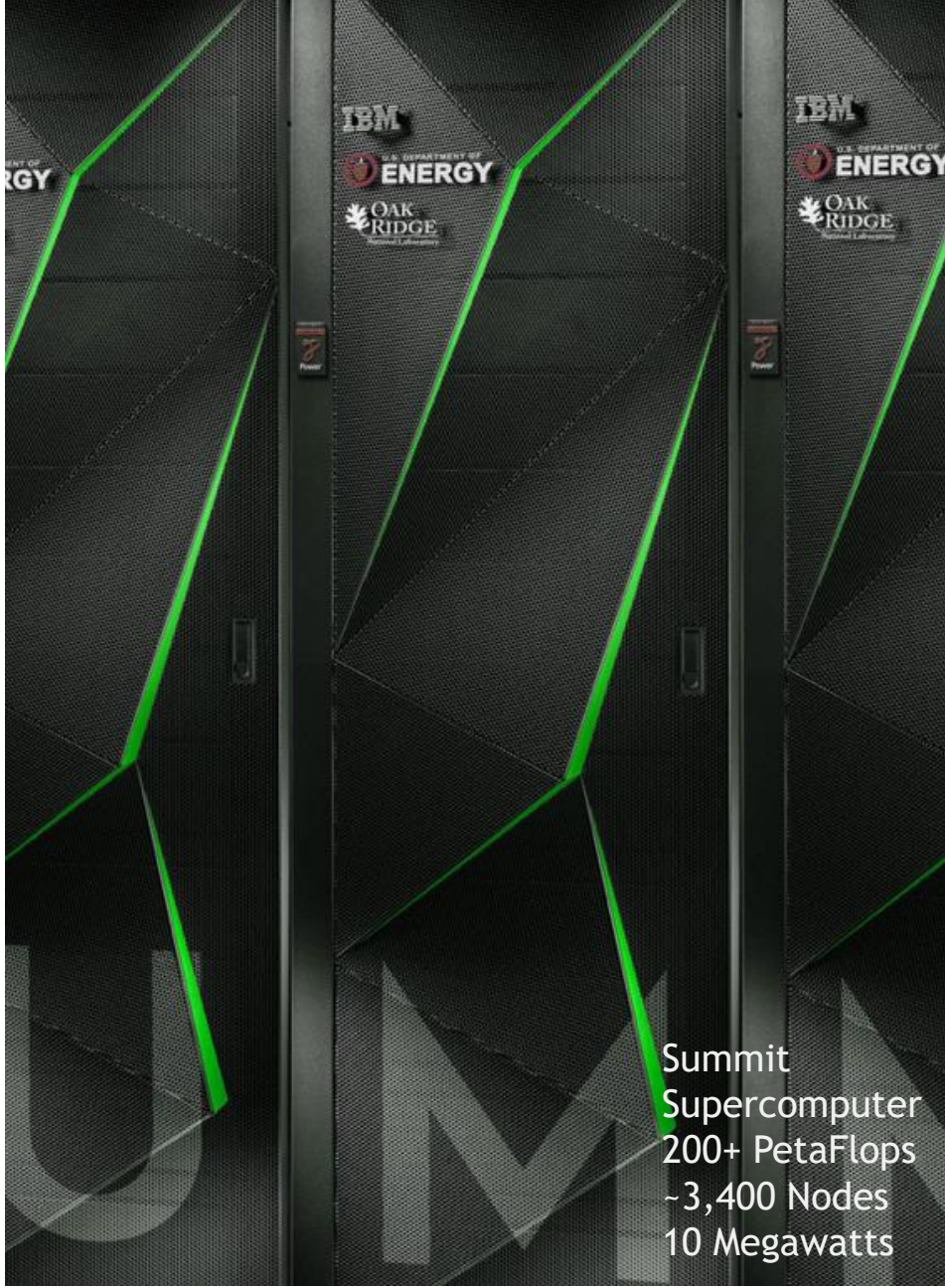
V100 measured on pre-production hardware.

EXASCALEへの歩み

USで最も強力なスーパーコンピュータを Voltaの演算性能で実現

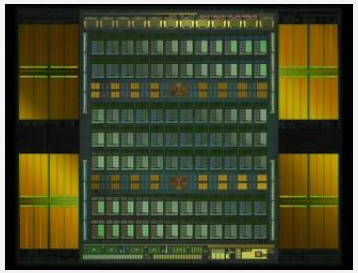


System Config Info: 2X Xeon E5-2690 v4, 2.6GHz, w/ 1X Tesla P100 or V100. V100 measured on pre-production hardware.



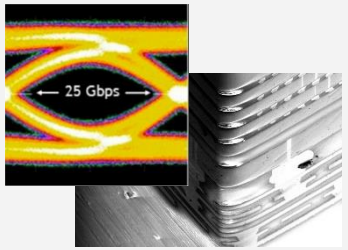
TESLA V100のご紹介

Volta アーキテクチャ



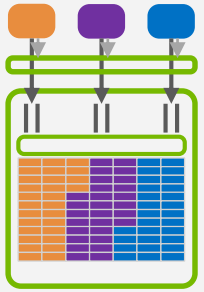
最も生産性の高いGPU

改善された NVLink と HBM2



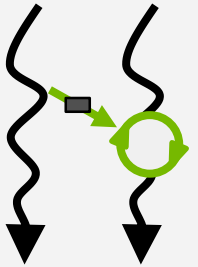
広帯域バンド幅

Volta MPS



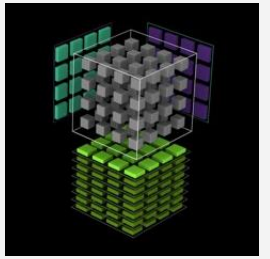
推論での活用

改善されたSIMTモデル



新しいアルゴリズム

Tensor コア



プログラマブルなディープラーニング演算エンジン

ディープラーニングとHPCにおける、最も高速で、最も生産性の高いGPU

TESLA V100

21B transistors
815 mm²

80 SM
5120 CUDA Cores
640 Tensor Cores

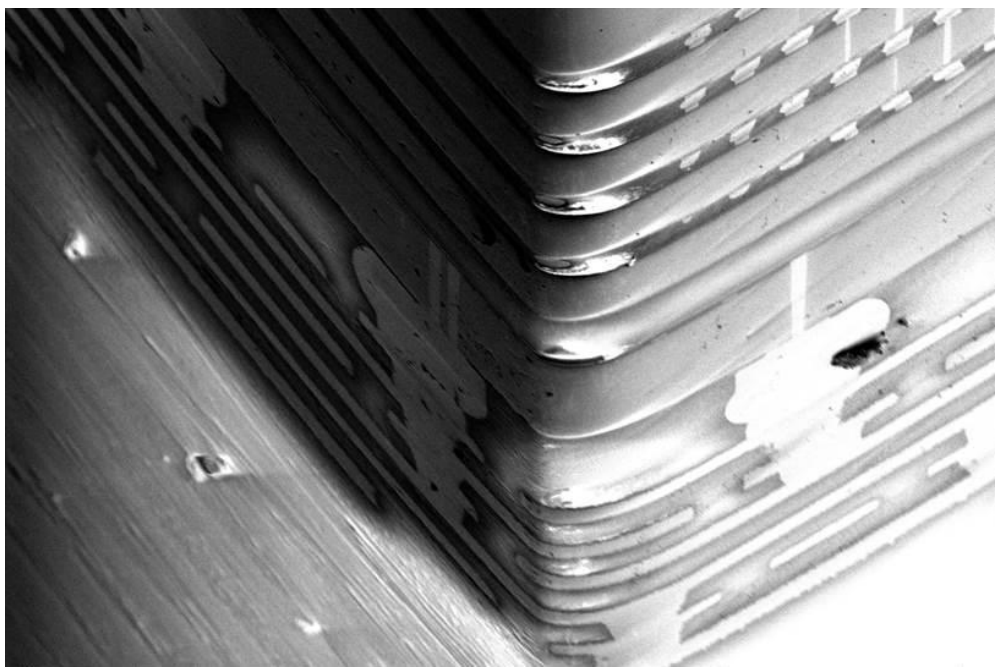
16 GB HBM2
900 GB/s HBM2
300 GB/s NVLink

*full GV100 chip contains 84 SMs

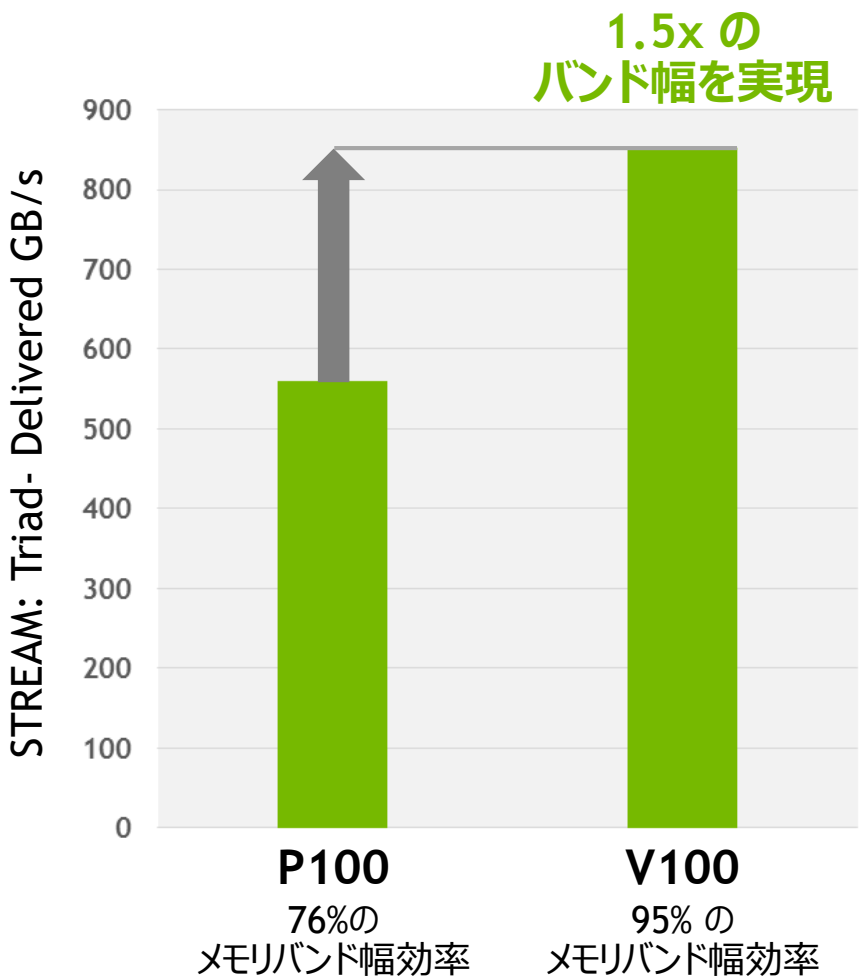
PASCAL / VOLTA GPUの性能比較

	P100	V100	Ratio
トレーニングの高速化	10 TOPS	120 TOPS	12x
推論の高速化	21 TFLOPS	120 TOPS	6x
FP64/FP32	5/10 TFLOPS	7.5/15 TFLOPS	1.5x
HBM2 バンド幅	720 GB/s	900 GB/s	1.2x
NVLink バンド幅	160 GB/s	300 GB/s	1.9x
L2 Cache	4 MB	6 MB	1.5x
L1 Caches	1.3 MB	10 MB	7.7x

新しい HBM2 メモリアーキテクチャ



HBM2 stack



V100 measured on pre-production hardware.

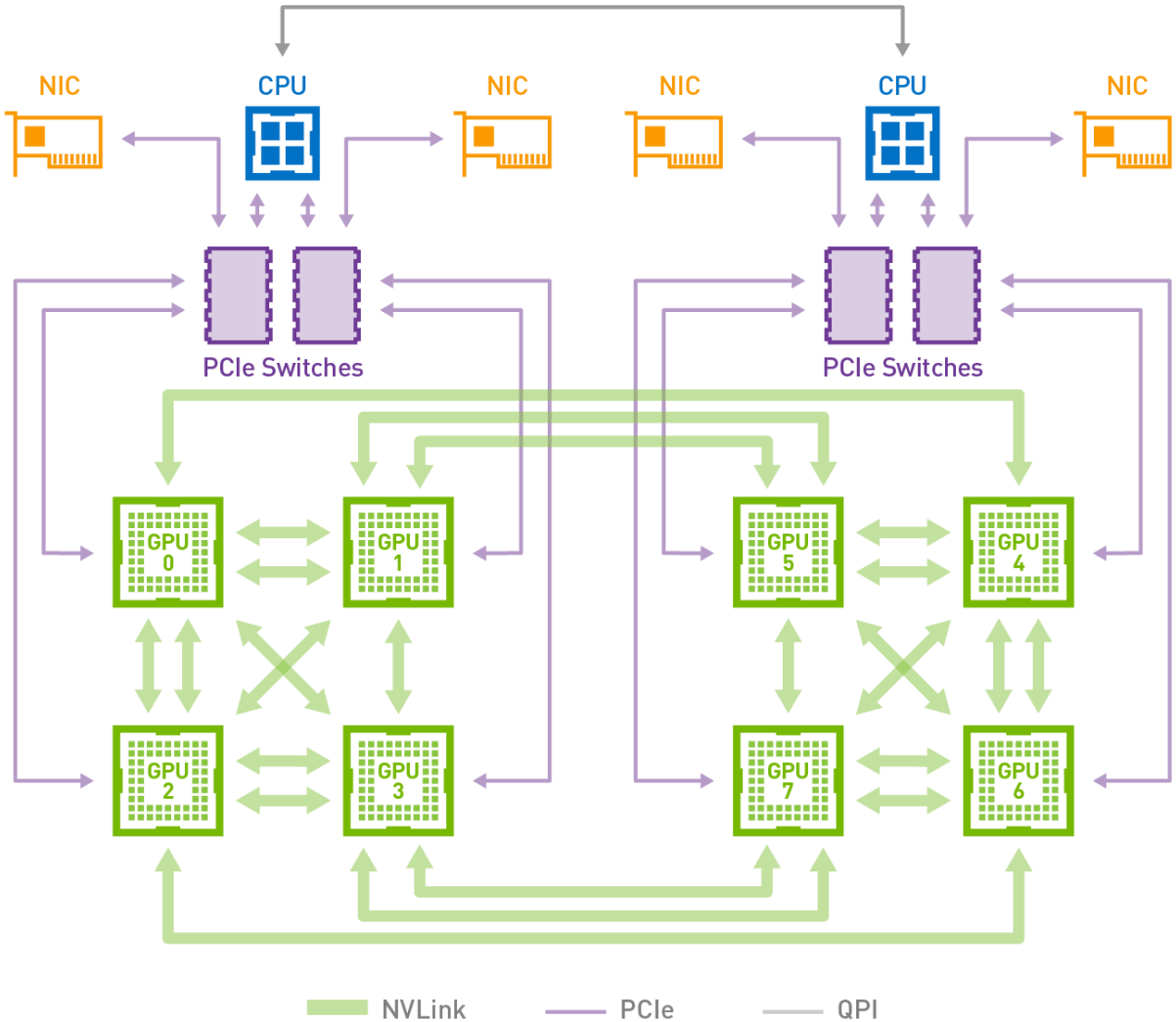
VOLTA NVLINK

300GB/sec

リンク数: 4本 → 6本

伝送速度(片方向) :

20 → 25 GB/sec



刷新されたSMマイクロアーキテクチャ

VOLTA GV100 SM

	GV100
FP32 units	64
FP64 units	32
INT32 units	64
Tensor Cores	8
Register File	256 KB
Unified L1/Shared memory	128 KB
Active Threads	2048

VOLTA GV100 SM

生産性のために刷新された設計

大容量、高速な L1 キャッシュ

テンソル演算の加速化

完全に新しいISA

スケジューラを倍増

簡素化された命令発行ロジック

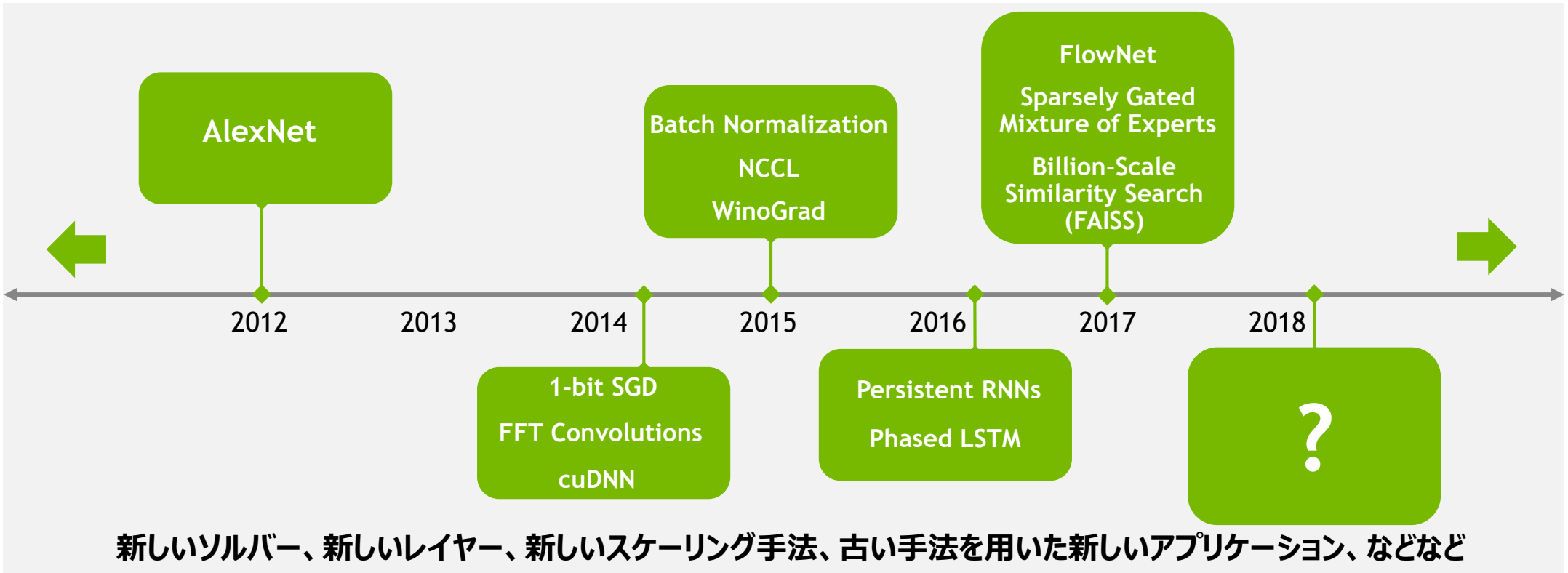
改善された SIMT モデル

=

もっとも容易にプログラミングできるSM

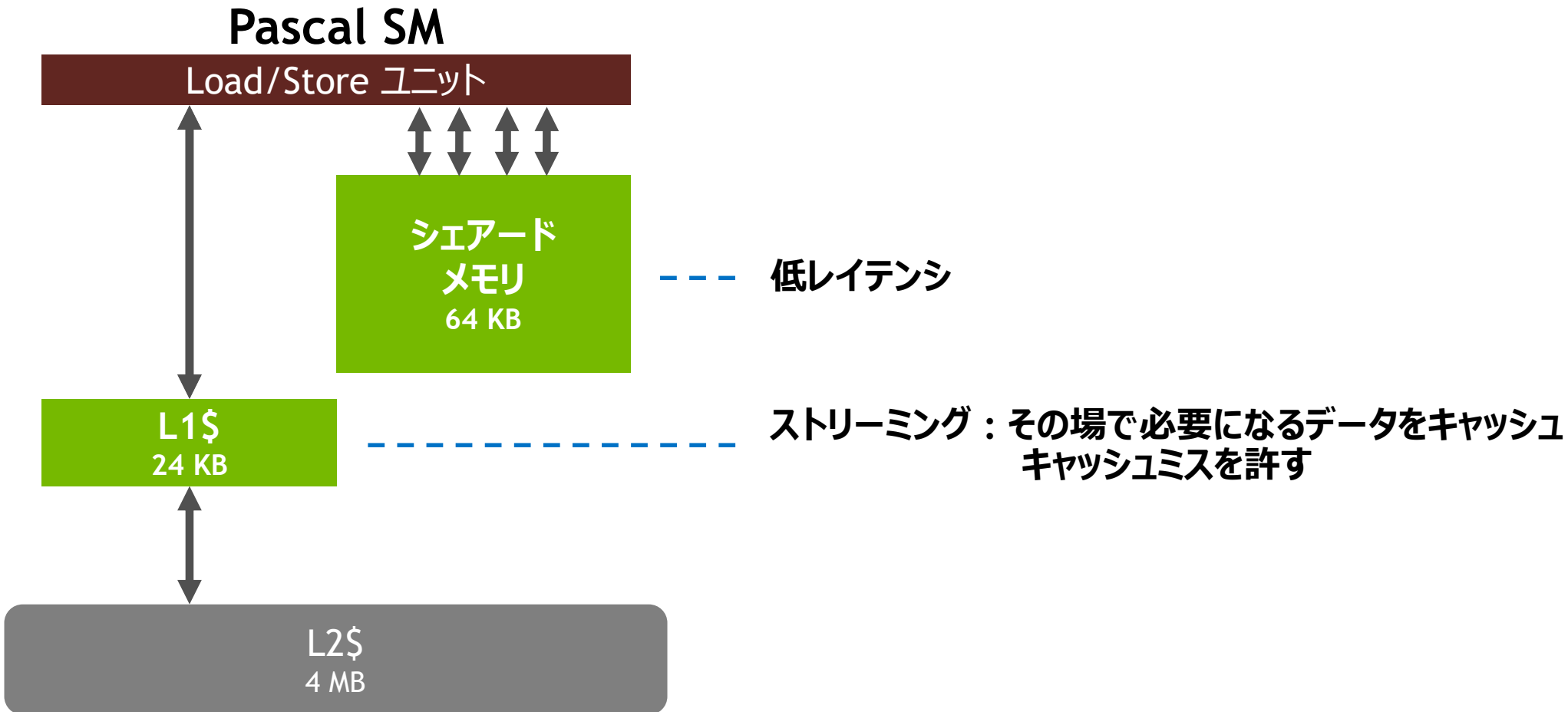
容易なプログラミングでディープラーニングを推進する

*CUDA*で実装されたディープラーニングの手法ディープラーニングの手法

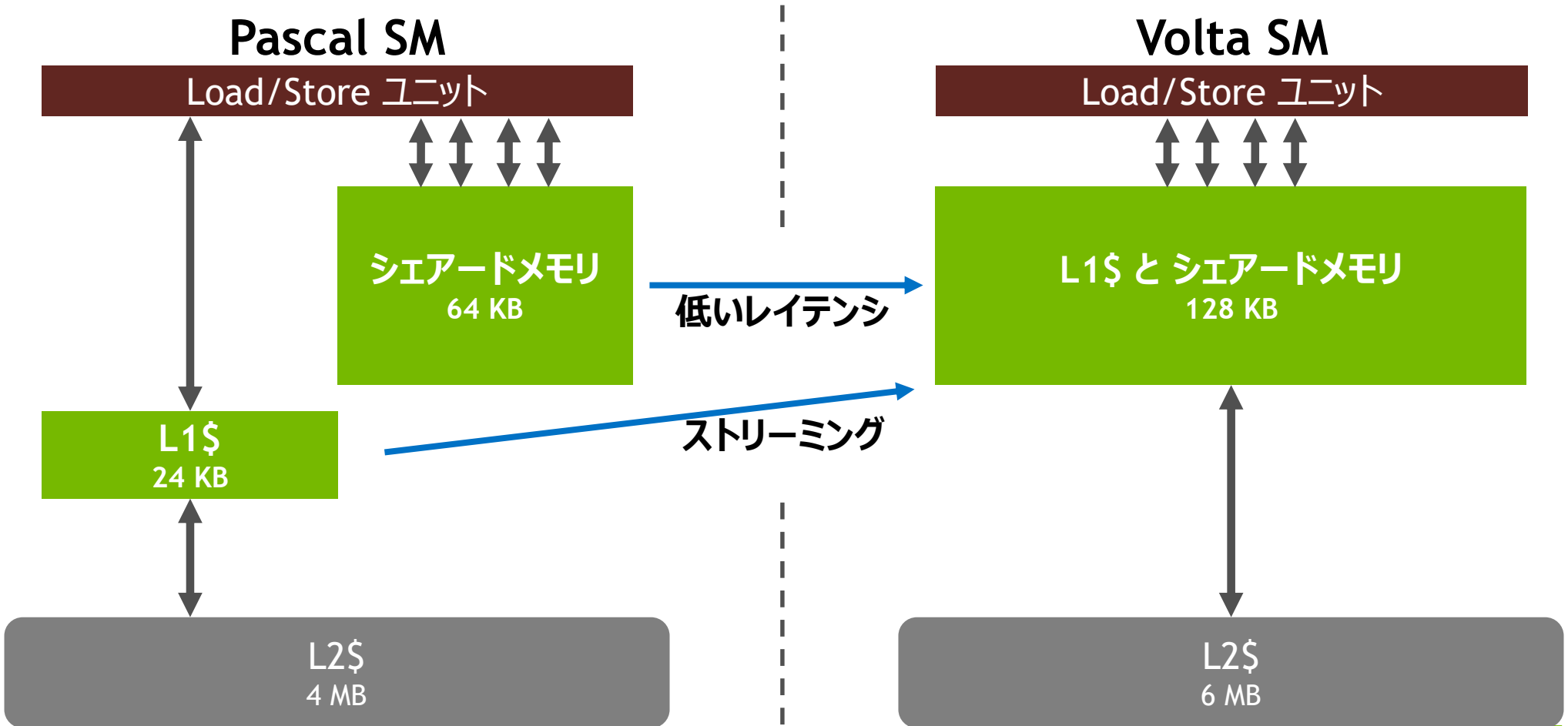


新しいソルバー、新しいレイヤー、新しいスケーリング手法、古い手法を用いた新しいアプリケーション、などなど

再録：PASCALの L1 と シェアードメモリ



UNIFYING KEY TECHNOLOGIES



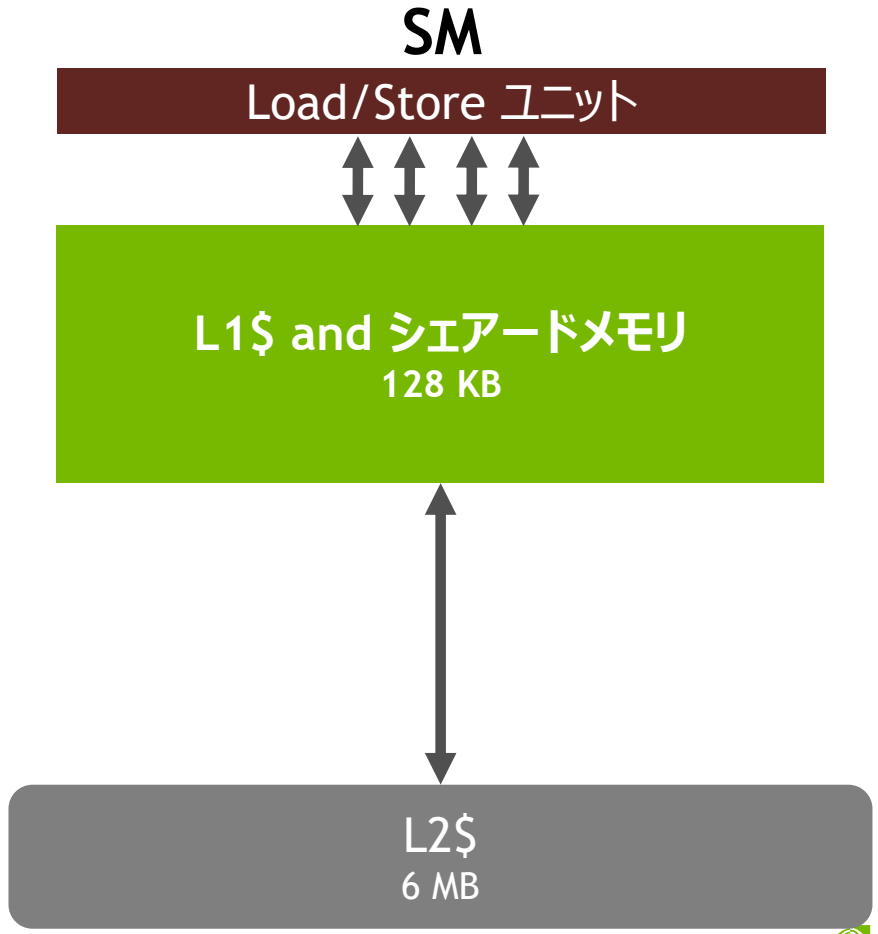
VOLTA L1 AND シェアードメモリ

Volta ストリーミングL1$:

処理中のキャッシュミスを許す
キャッシュヒット時のレイテンシが低い
4倍のバンド幅
5倍の容量

Volta シェアードメモリ:

L1キャッシュと統合された記憶域
最大96 KBまで構成可能



小さくなったシェアードメモリとの性能差

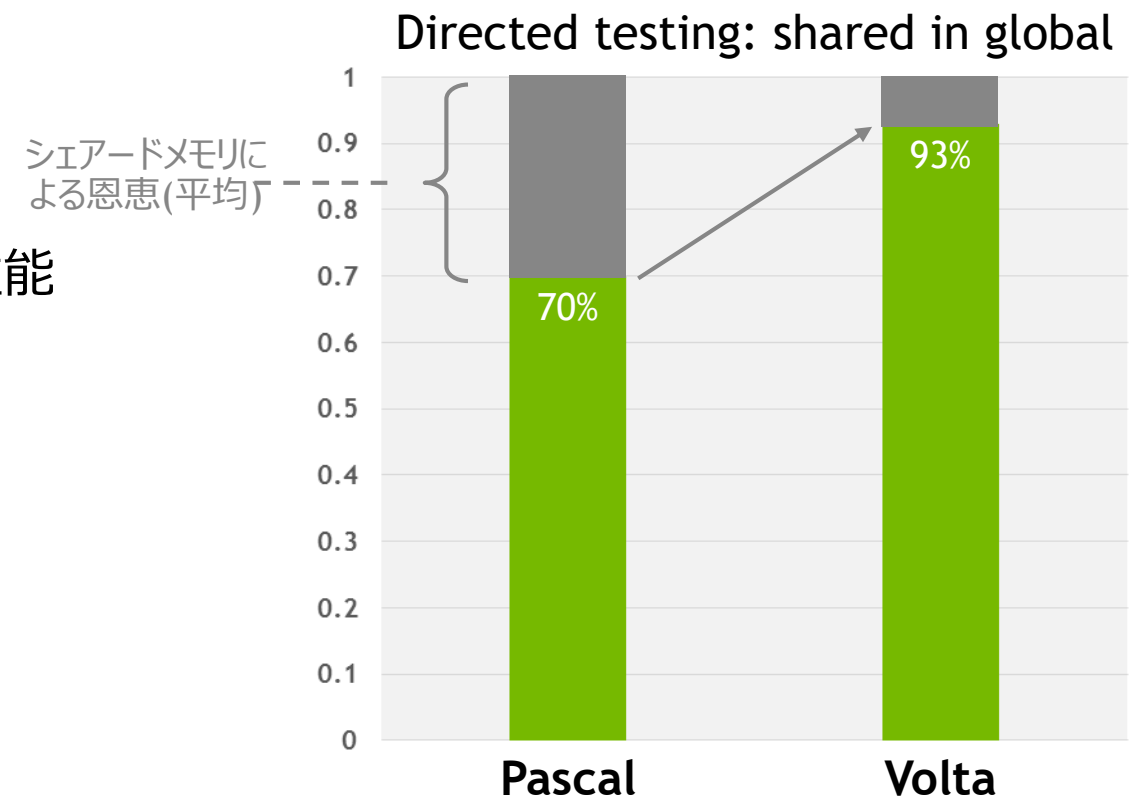
with the GV100 L1 cache

Cache: vs shared

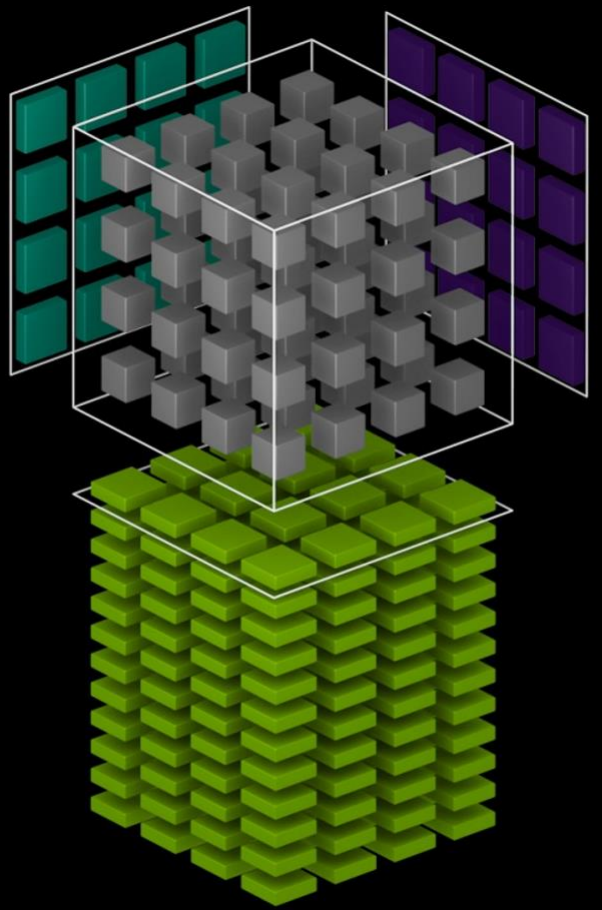
- 簡単に使用できる
- 90%以上の場合で十分な性能

Shared: vs cache

- より高速なアトミクス
- より多くのバンク
- 性能を予測しやすい

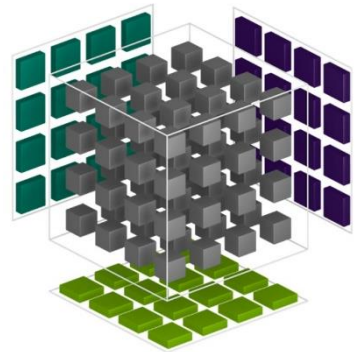


VOLTA TENSOR コア



TENSOR コア

混合精度行列演算
4x4 行列

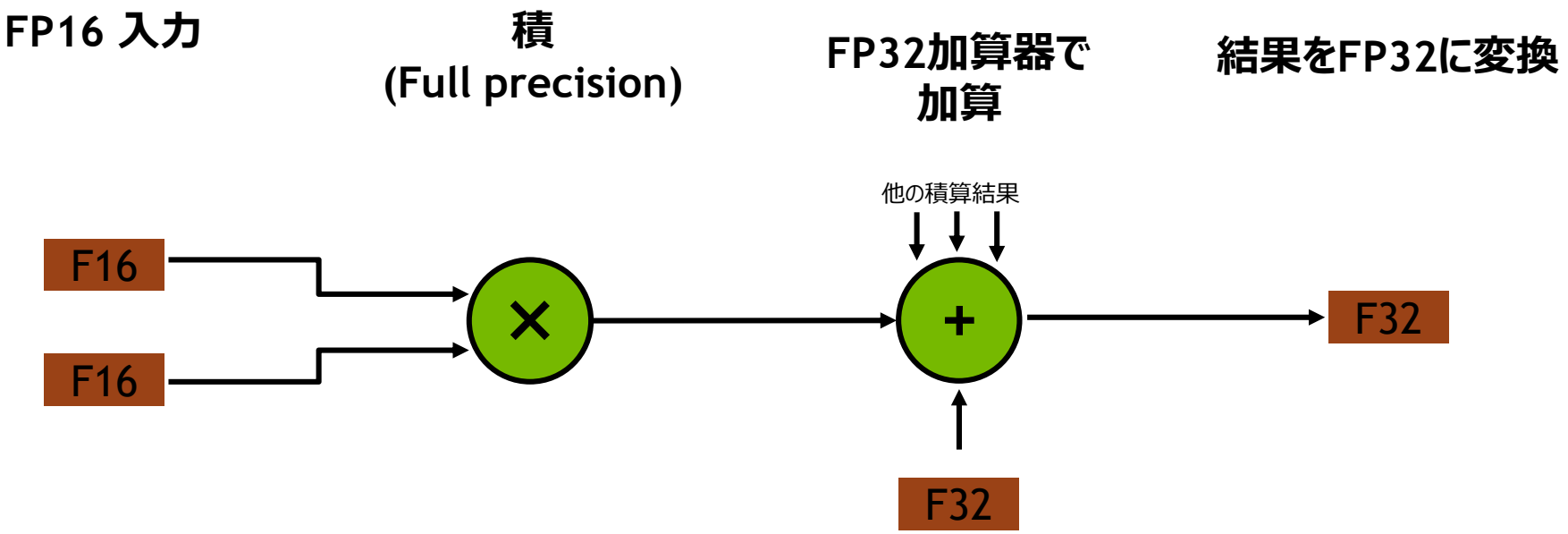


$$D = \begin{pmatrix} A_{0,0} & A_{0,1} & A_{0,2} & A_{0,3} \\ A_{1,0} & A_{1,1} & A_{1,2} & A_{1,3} \\ A_{2,0} & A_{2,1} & A_{2,2} & A_{2,3} \\ A_{3,0} & A_{3,1} & A_{3,2} & A_{3,3} \end{pmatrix} \begin{pmatrix} B_{0,0} & B_{0,1} & B_{0,2} & B_{0,3} \\ B_{1,0} & B_{1,1} & B_{1,2} & B_{1,3} \\ B_{2,0} & B_{2,1} & B_{2,2} & B_{2,3} \\ B_{3,0} & B_{3,1} & B_{3,2} & B_{3,3} \end{pmatrix} + \begin{pmatrix} C_{0,0} & C_{0,1} & C_{0,2} & C_{0,3} \\ C_{1,0} & C_{1,1} & C_{1,2} & C_{1,3} \\ C_{2,0} & C_{2,1} & C_{2,2} & C_{2,3} \\ C_{3,0} & C_{3,1} & C_{3,2} & C_{3,3} \end{pmatrix}$$

FP16 or FP32 　 FP16 　 FP16 　 FP16 or FP32

$$D = AB + C$$

VOLTA TENSOR OPERATION



推論のために*FP16*加算モードも サポート

USING TENSOR CORES

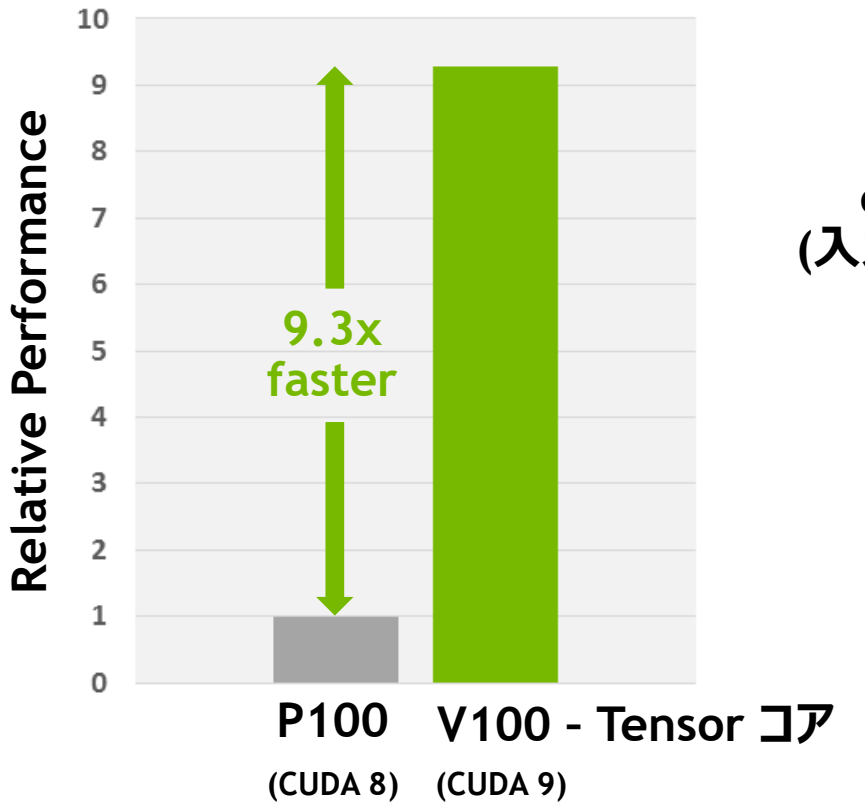
NVIDIA cuDNN, cuBLAS, TensorRT

Voltaに最適化された
フレームワークとライブラリ

```
__device__ void tensor_op_16_16_16(
    float *d, half *a, half *b, float *c)
{
  wmma::fragment<matrix_a, …> Amat;
  wmma::fragment<matrix_b, …> Bmat;
  wmma::fragment<matrix_c, …> Cmat;

  wmma::load_matrix_sync(Amat, a, 16);
  wmma::load_matrix_sync(Bmat, b, 16);
  wmma::fill_fragment(Cmat, 0.0f);

  wmma::mma_sync(Cmat, Amat, Bmat, Cmat);

  wmma::store_matrix_sync(d, Cmat, 16,
                          wmma::row_major);
}
```

CUDA C++
Warpレベル行列演算

A GIANT LEAP FOR DEEP LEARNING



cuBLAS 混合精度演算
(入力 FP16, 演算結果 FP32)
行列積 (M=N=K=2048)

V100 measured on pre-production hardware.

マルチプロセス実行時のスケジューリング

GPU上のマルチプロセススケジューリング

背景

CPU プロセス

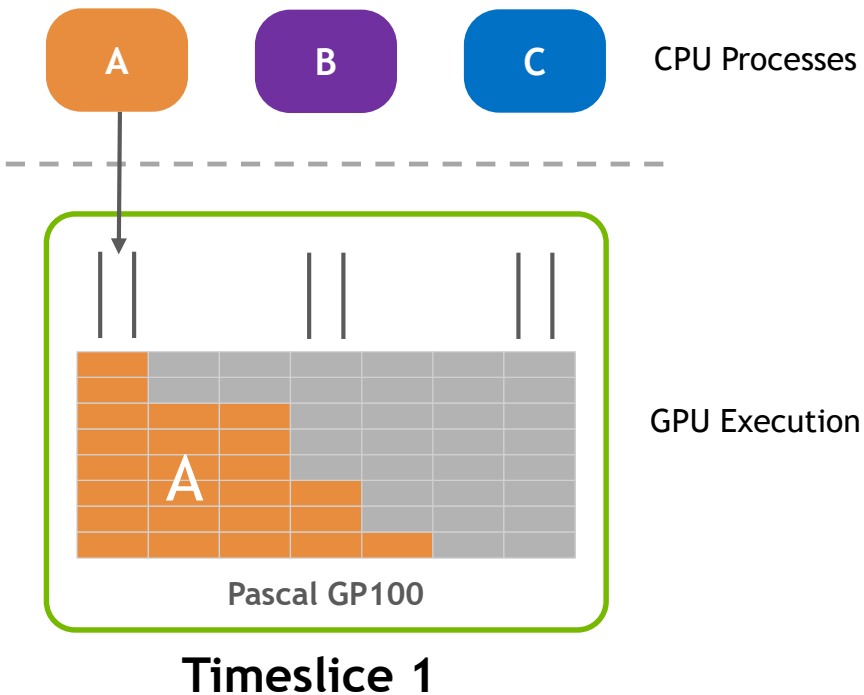
スケジューリングはタイムスライス

単一プロセス実行の場合
スループットが最適化される

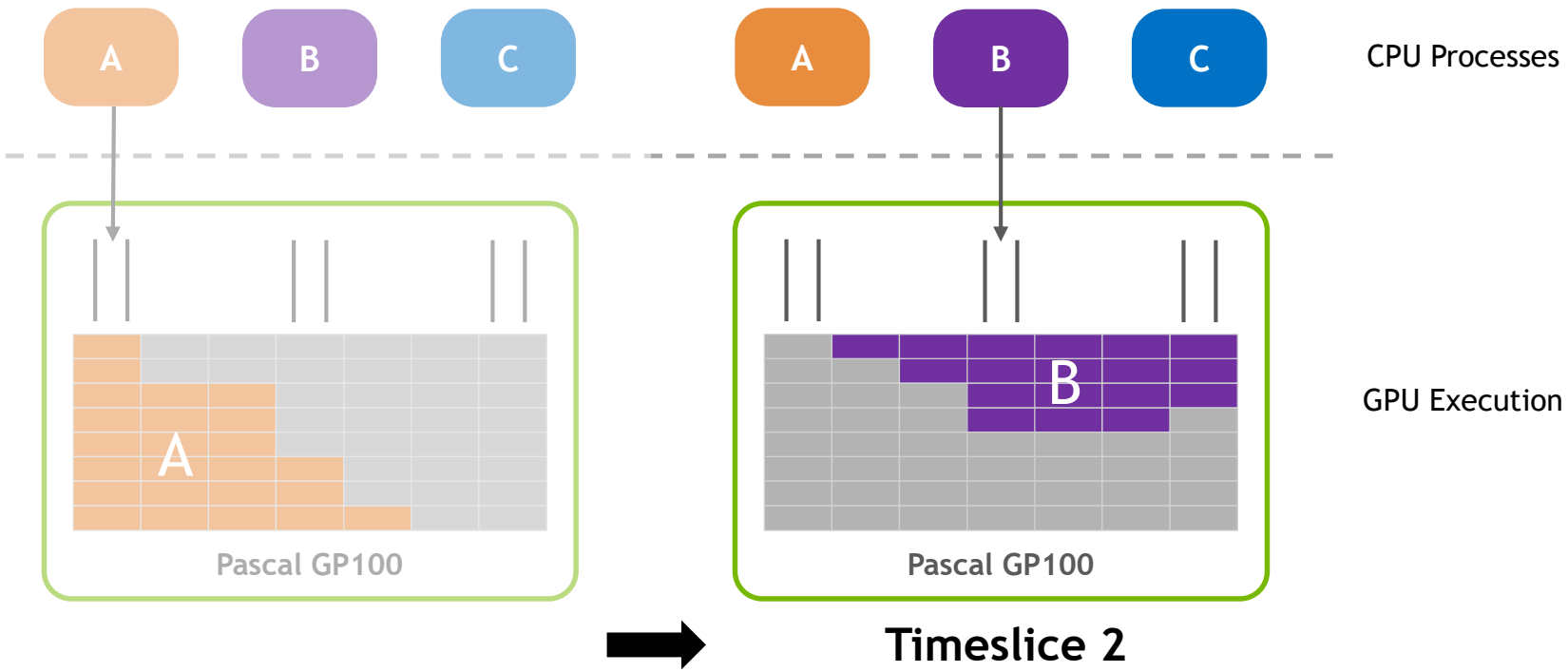
マルチプロセスサービス(MPS)

マルチプロセスのスループットが最適化される

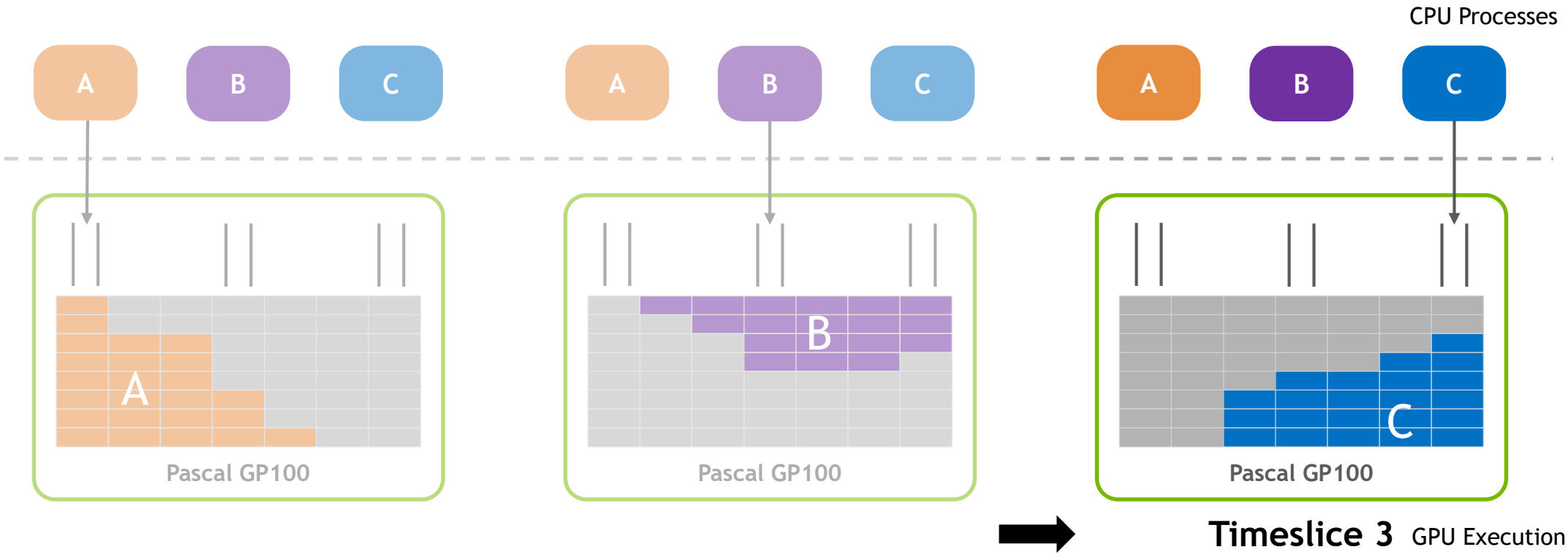
マルチプロセスの実行はタイムスライス



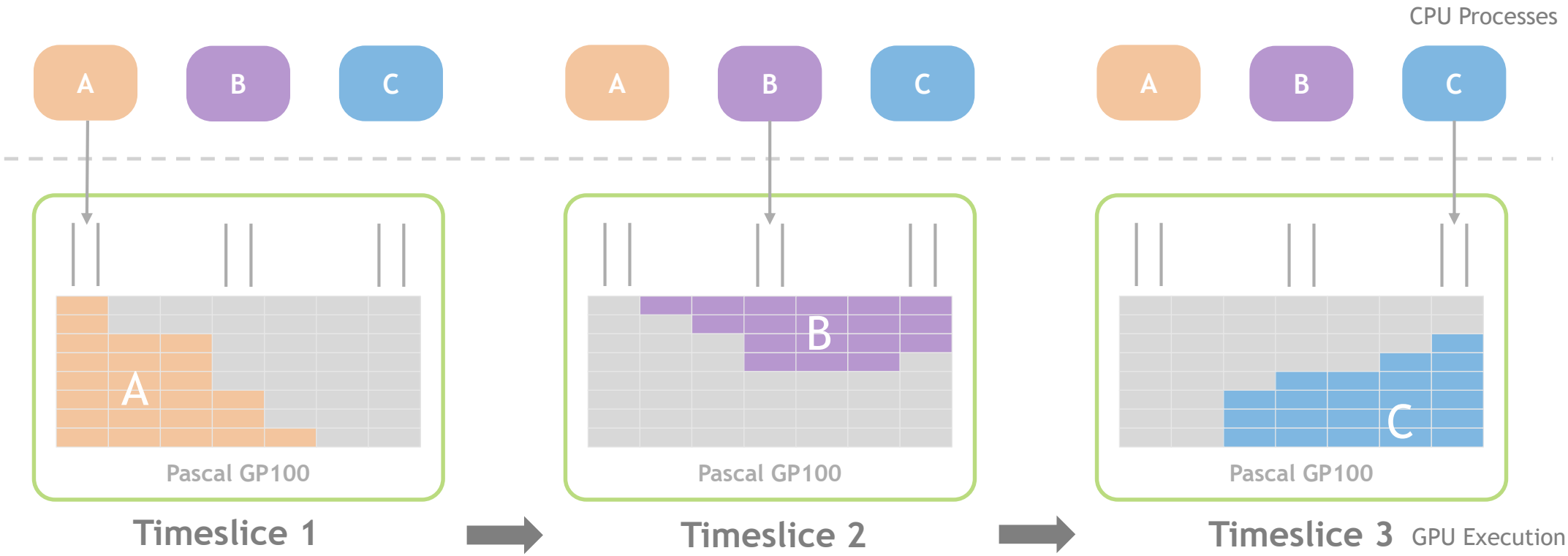
マルチプロセスの実行はタイムスライス



マルチプロセスの実行はタイムスライス



マルチプロセスの実行はタイムスライス

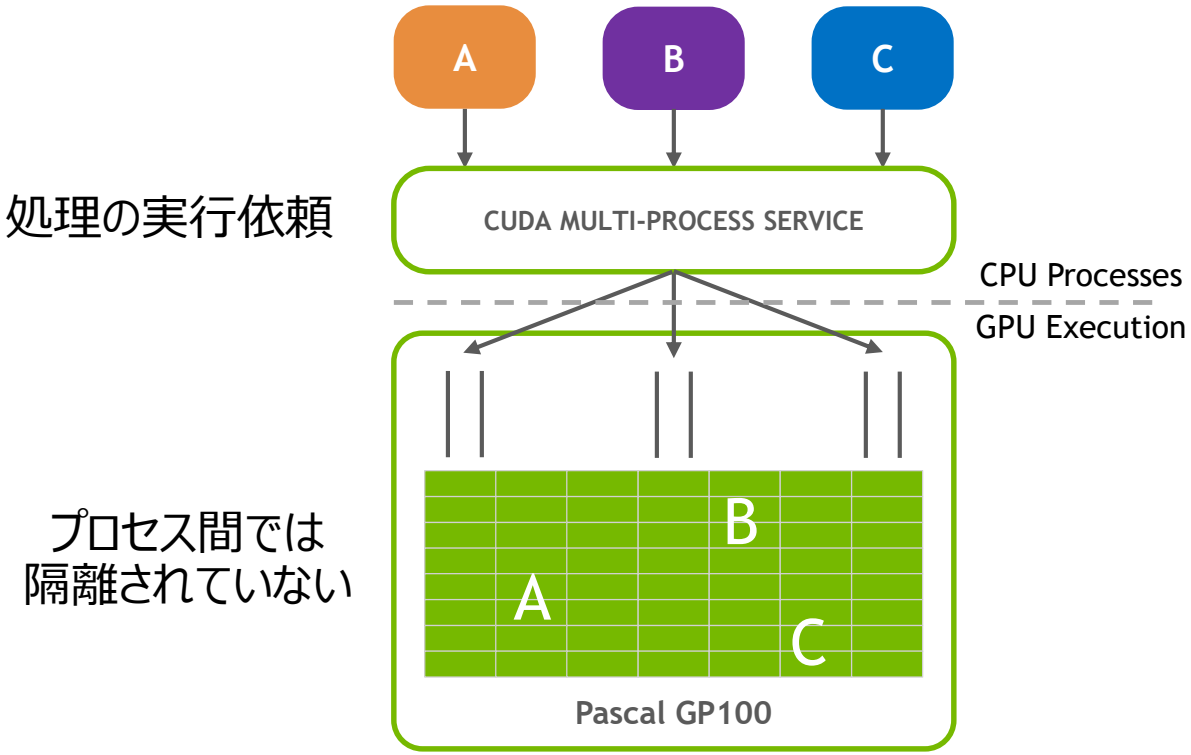


プロセスがアイソレートされている、それぞれのプロセスで最高性能を発揮することができる

PASCALマルチプロセスサービス

CUDAマルチプロセスサービス:

小さなジョブ間で、演算リソースを共有し、
GPUの利用率を改善

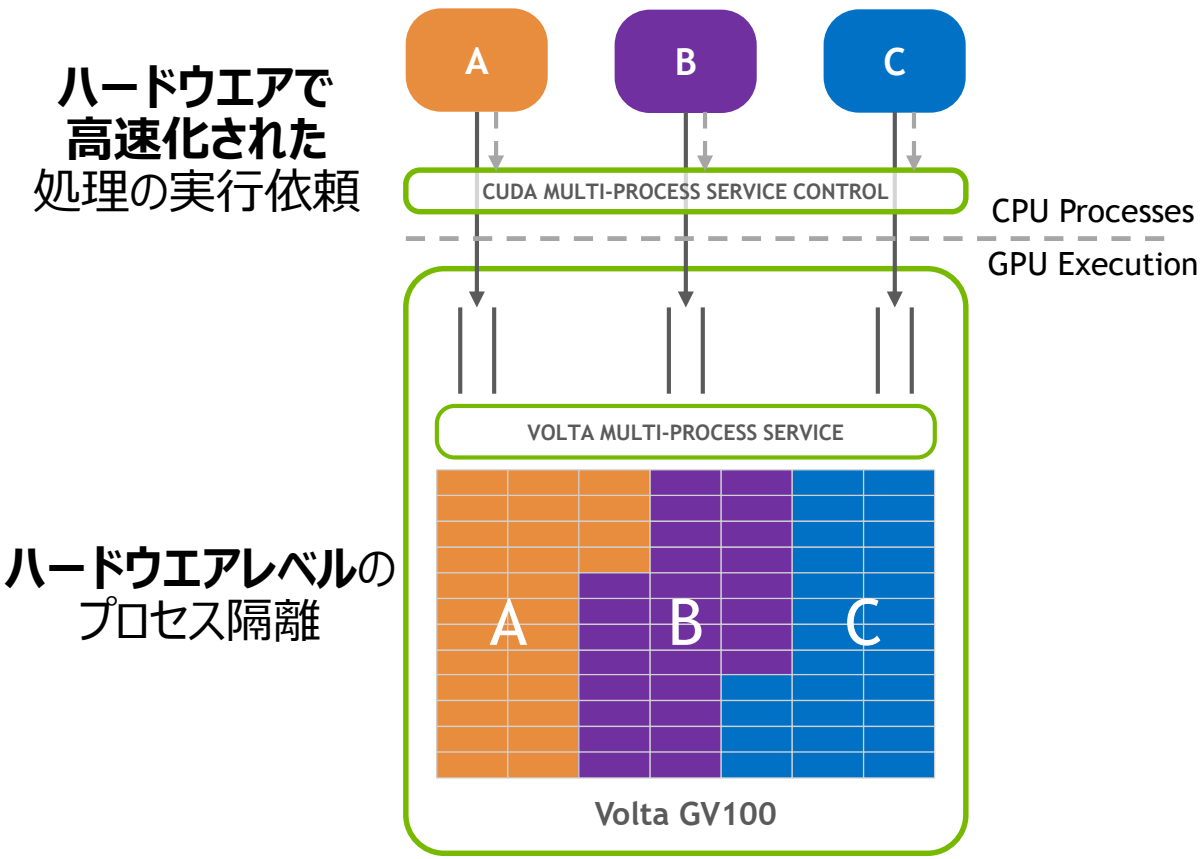


Opt-in: プロセス間の隔離は限定的。プロセスを束ねることでピークスループットを実現

VOLTA マルチプロセスサービス

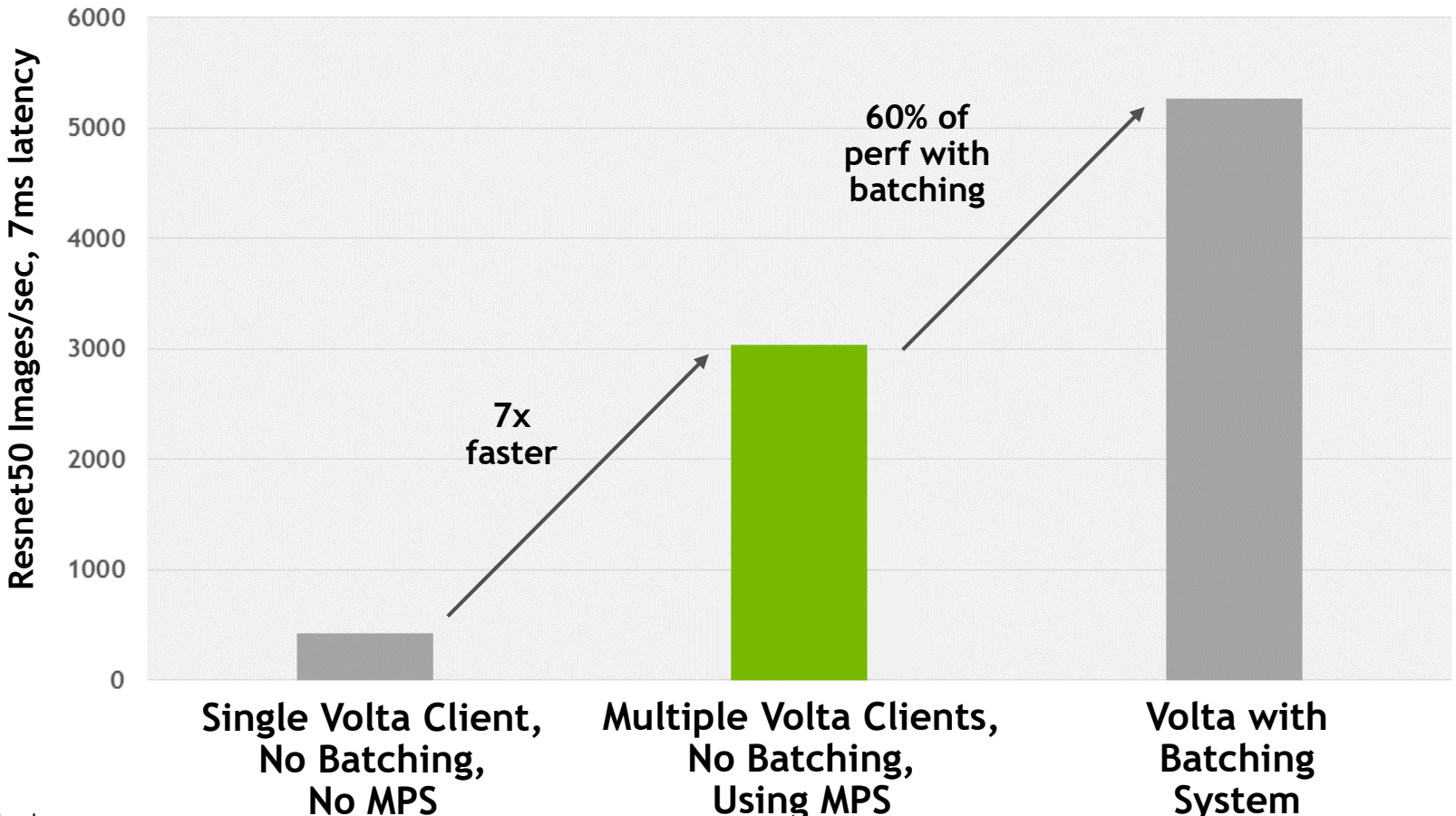
Volta における改善:

- 処理実行のレイテンシを削減
- 処理実行のスループットを改善
- プロセス単位でのスケジューリング
  - より安定したパフォーマンス
- Pascalに比べ3倍のクライアント



VOLTA MPS の推論時性能

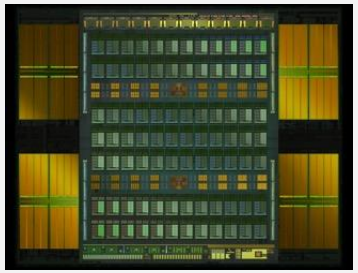
バッチ処理を行わなくとも高速な推論



V100 measured on pre-production hardware.

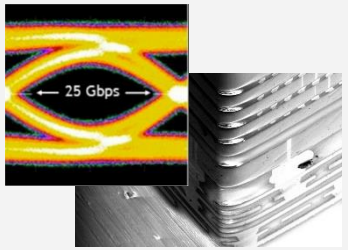
TESLA V100のご紹介

Volta アーキテクチャ



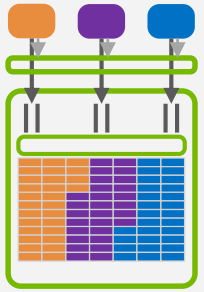
最も生産性の高いGPU

改善された NVLink と HBM2



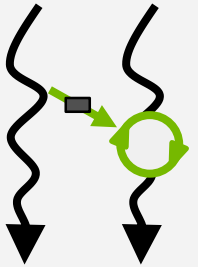
広帯域バンド幅

Volta MPS



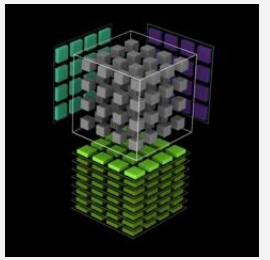
推論での活用

改善されたSIMTモデル



新しいアルゴリズム

Tensor コア



プログラマブルなディープラーニング演算エンジン

他の*V100*の新機能: *2x L2* アトミクス, *int8*, 新しいメモリモデル, コピーエンジンページマイグレーション, などなど ...

ディープラーニングとHPCにおける、最も高速で、最も生産性の高いGPU

www.nvidia.com/dli