最新の VOLTA アーキテクチャと NVIDIA GPU CLOUD (NGC)

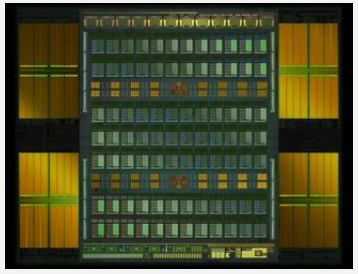
エヌビディア合同会社
佐々木邦暢 (@_ksasaki)

NVIDIA

VOLTA アーキテクチャ

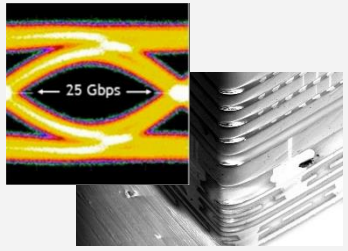
Tesla V100の概要

Volta Architecture



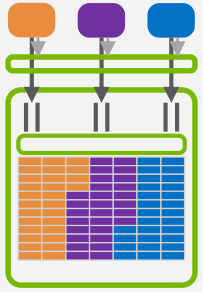
Most Productive GPU

Improved NVLink & HBM2



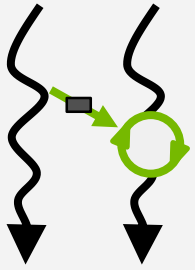
Efficient Bandwidth

Volta MPS



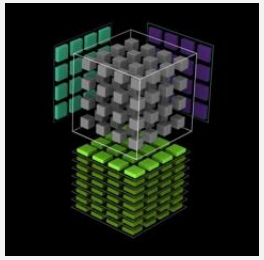
Inference Utilization

Improved SIMT Model



New Algorithms

Tensor Core



125 Programmable TFLOPS Deep Learning

ディープラーニングと HPC 、両方に最適な GPU

Tesla V100

トランジスタ数:21B
815 mm^2

80 SM
5120 CUDAコア
640 Tensorコア

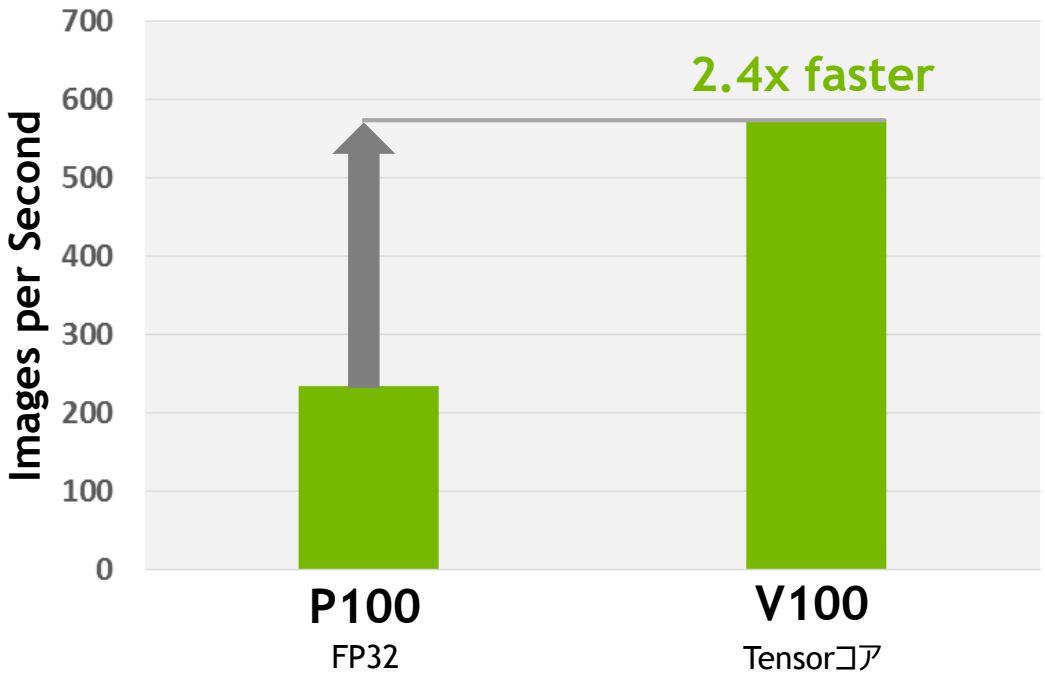
HBM2
16 GB, 900 GB/s

NVLink 300 GB/s

*full GV100 chip contains 84 SMs

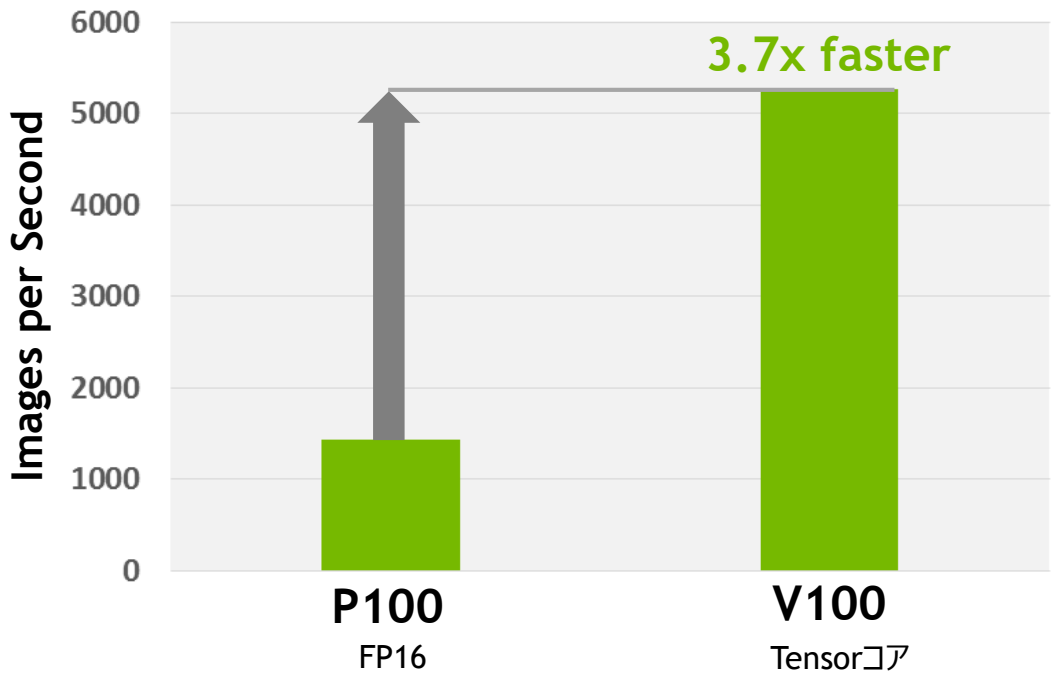
VOLTA
ディープラーニング性能を大幅に向上

トレーニング



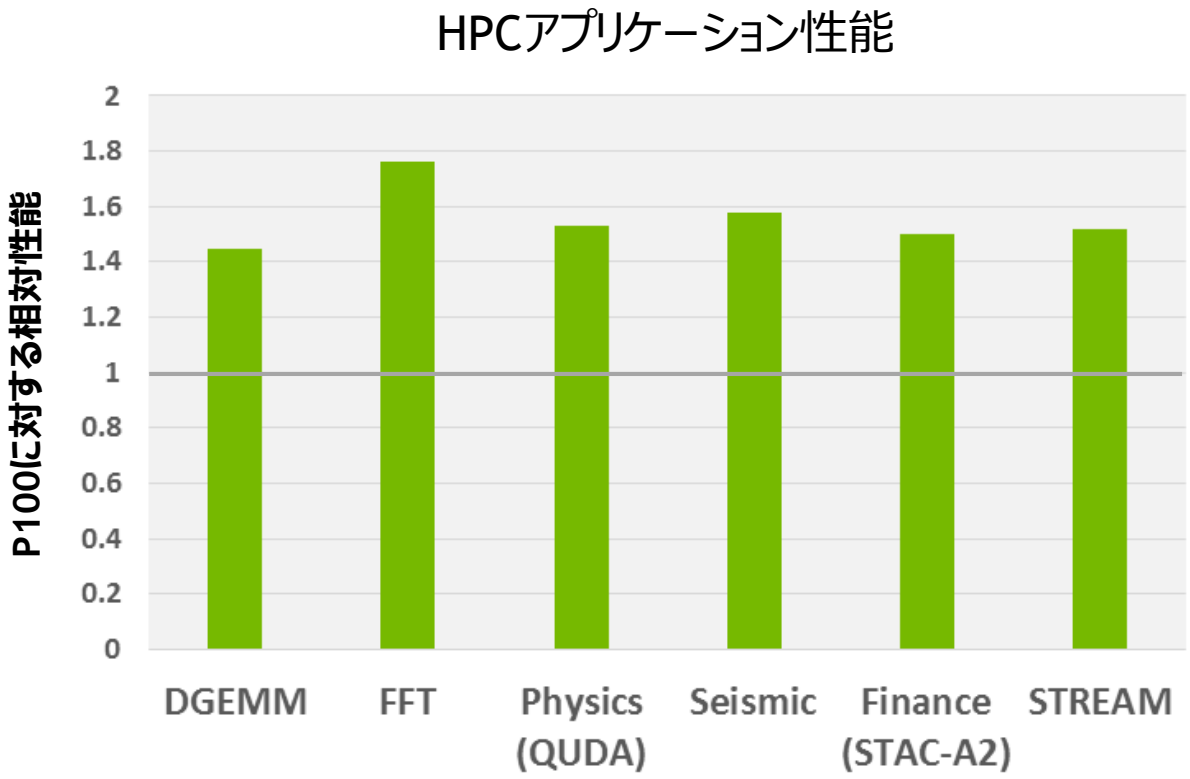
インファレンス

TensorRT - 7ms Latency

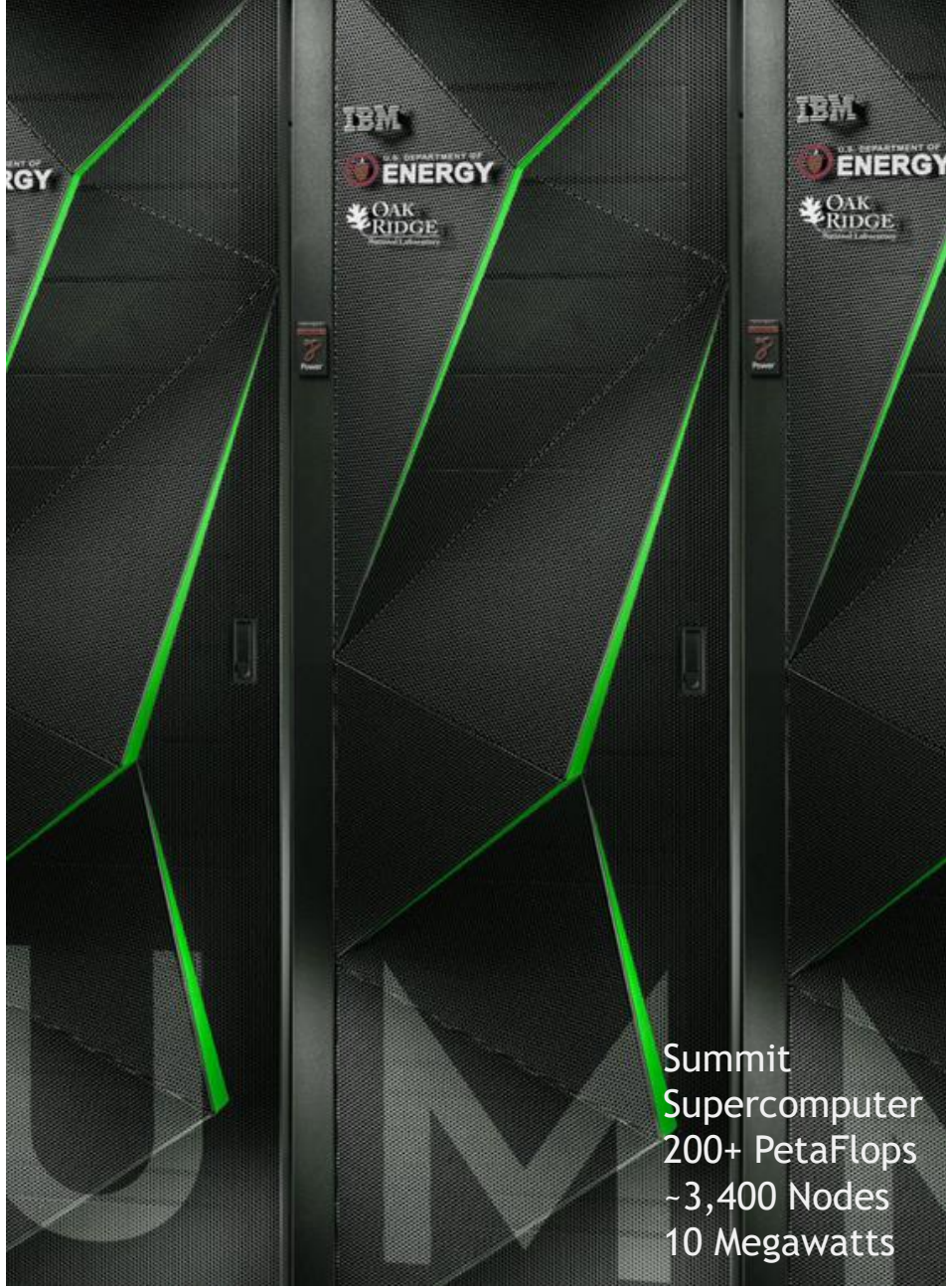


(*) DLモデルはResNet50

VOLTA
HPC性能を大きく向上



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



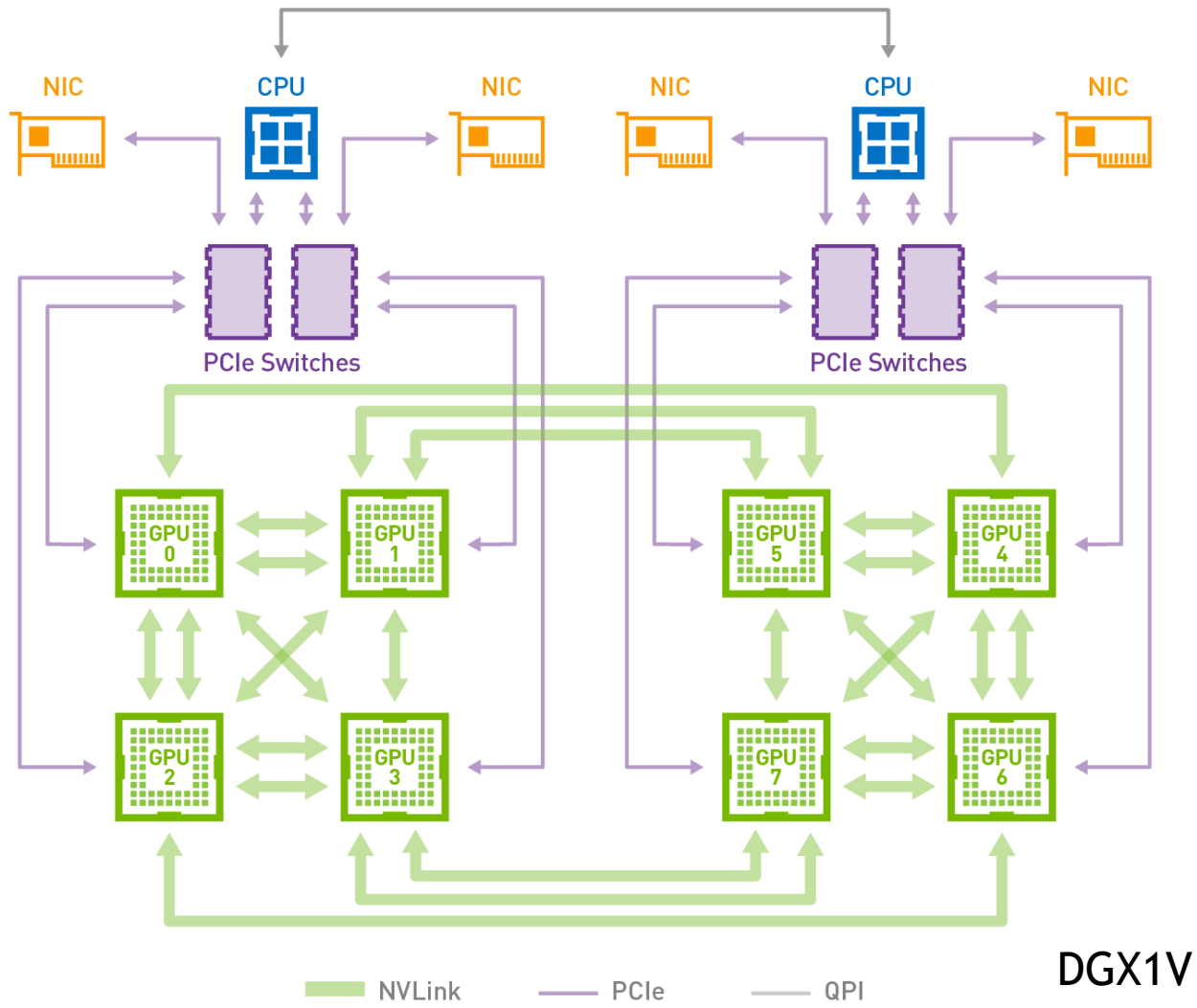
GPU ピーク性能比較: P100 vs v100

	P100	V100	性能UP
トレーニング性能	10 TOPS	125 TOPS	12x
インファレンス性能	21 TFLOPS	125 TOPS	6x
FP64/FP32	5/10 TFLOPS	7.8/15.6 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 キャッシュ	4 MB	6 MB	1.5x
L1 キャッシュ	1.3 MB	10 MB	7.7x

Volta NVLink

	P100	V100
リンク数	4	6
バンド幅 / リンク	40 GB/s	50 GB/s
トータルバンド幅	160 GB/s	300 GB/s

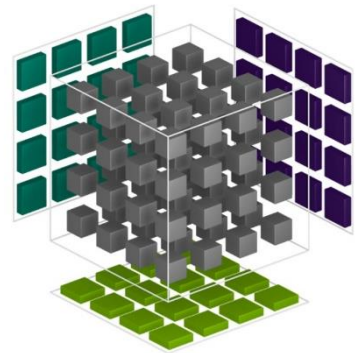
(*) バンド幅は双方向



DGX1V

Tensor コア

混合精度行列計算ユニット



4x4の行列の積和演算を1サイクルで計算する性能 (128演算/サイクル)

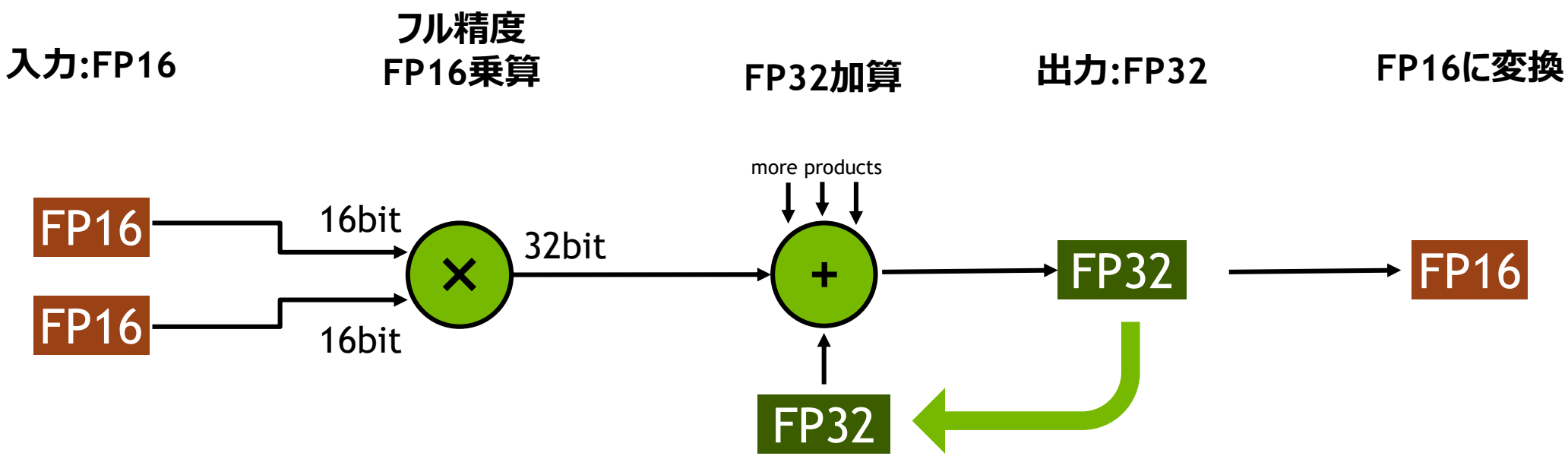
行列のFMA (Fused Multiply-Add)

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

FP16 or FP32 FP16 FP16 FP16 or FP32

$$D = AB + C$$

VOLTA TENSOR演算



FP16加算もサポート (インファレンス用)

TENSORコアの使い方

NVIDIA cuBLAS, cuDNN, TensorRT

Volta向けに最適化された
フレームワーク・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レベル行列演算テンプレート

DLフレームワークの対応状況

NVIDIA CAFFE 0.16

https://github.com/NVIDIA/caffe/tree/caffe-0.16

FP16、Tensorコアに完全対応

ForwardとBackward: それぞれ、データ型、計算型を指定可能 (FP32 or FP16)

ウェイト更新: FP32更新対応

ロス・スケーリング対応

NVIDIA CAFFE 0.16

https://github.com/NVIDIA/caffe/tree/caffe-0.16

```
name: "AlexNet_fp16"

default_forward_type:  FLOAT16
default_backward_type: FLOAT16

# default_forward_math:  FLOAT   # GP100 only
# default_backward_math: FLOAT   # GP100 only

global_grad_scale: 1000.

layer {
  forward_math:  FLOAT16
  backward_math: FLOAT
  …
}
```

```
solver_data_type: FLOAT16
```

TensorFlow

Tensorコア: TensorFlow 1.4で対応

データ型をFP16にすると、Tensorコアを使用

```
tf.cast(tf.get_variable(..., dtype=tf.float32), tf.float16)
```

ウェイトFP32更新: 可能

ロススケーリング: 可能

```
scale = 128
grads = [grad / scale for grad in tf.gradients(loss * scale, params)]
```

PyTorch

Tensorコア: 対応

FP16ストレージにすると、Tensorコアを使用

```
Input = input.cuda().half()
model = model.cuda().half()
```

ウェイトFP32更新: 可能

ロススケーリング: 可能

PYTORCH

Chainer

Tensorコア: Chainer V4で対応予定

データ型をFP16にすると、Tensorコア使用

```
x = F.cast(x, np.float16)
```

FP32パラメータ更新: 対応

```
optimizer = chinaer.optimizers.SGD()
optimizer.use_fp32_update()
```

ロススケーリング: 対応(予定)

```
loss = lossfunc(y, t)
loss.backward(loss_scale=1024)
```

Tesla V100 搭載 NVIDIA DGX-1

AI 研究の必需品

960 Tensor TFLOPS | Tesla V100 8基 | NVLink ハイブリッドキューブ

TITAN X で 8 日かかる計算が 8 時間に

CPU サーバー 400台分の性能がワンボックスに

NVIDIA DGX Station

パーソナル DGX

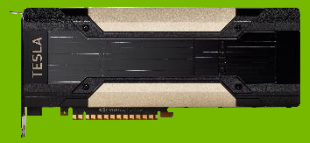
480 Tensor TFLOPS | Tesla V100 4基

NVLink 全結合 | 3つの DisplayPort

1500W | 水冷

AWS P3 インスタンス

Tesla V100 を搭載する最新 GPU インスタンス

インスタンス名		GPU メモリ (GB)	NVLink	vCPU 数	メモリ (GB)
p3.2xlarge	1	16	2,496	8	61
p3.8xlarge	4	64	19,968	32	244
p3.16xlarge	8	128	39,936	64	488

NVIDIA GPU CLOUD (NGC)

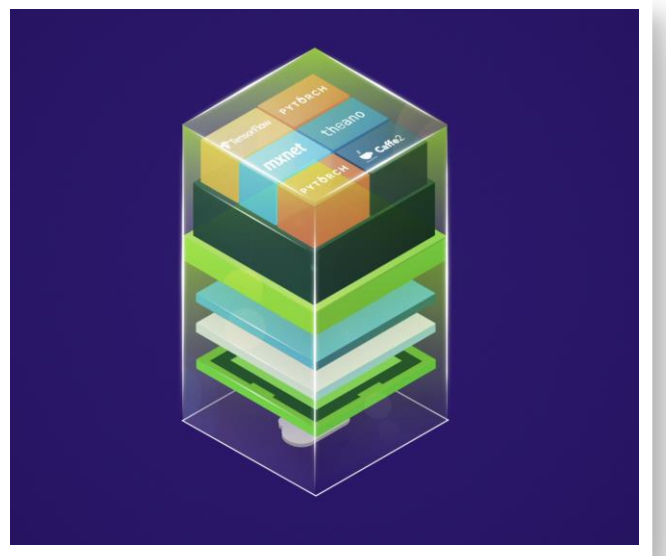
NVIDIA GPU CLOUD

誰でも、どこでも、ディープラーニング

数週間ではなく、数分でイノベーションを生み出す
ディープラーニング ソフトウェア環境を構築するための複雑な作業を、大幅に軽減します。

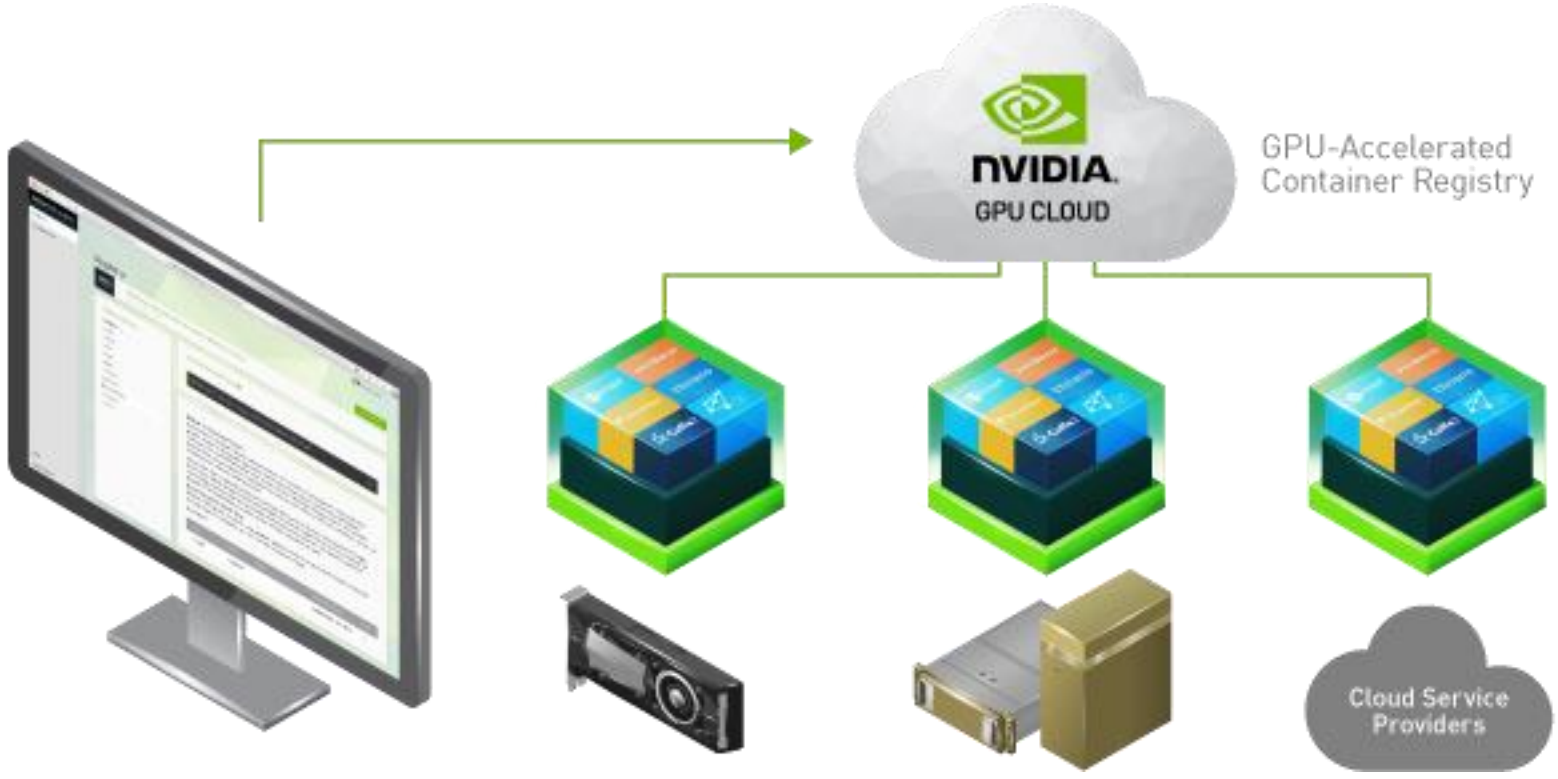
クロス プラットフォーム ディープラーニング
NGC のコンテナは、オンプレミスの DGX Systems や TITAN 搭載 PC だけでなく、各種クラウドの GPU インスタンスでも同じように動きます。

常に最新
コンテナが最高のパフォーマンスを発揮できるように、エヌビディアが毎月アップデートを提供します。



NVIDIA GPU Cloud は GPU に最適化されたディープラーニング フレームワークと各種ライブラリを、すぐに実行できるコンテナとして無料で提供します。

クロス プラットフォーム ディープラーニング



NVIDIA Volta あるいは
Pascal 世代の TITAN GPU

NVIDIA DGX-1 と
DGX Station

NVIDIA Volta 搭載の
Amazon EC2 P3 インスタンス

NGC の GPU 最適化ディープラーニング コンテナ

ディープラーニングソフトウェアの包括的なカタログ

- NVCaffe
- Caffe2
- Chainer
- Microsoft Cognitive Toolkit (CNTK)
- DIGITS
- MXNet
- PyTorch
- TensorFlow
- Theano
- Torch
- CUDA (base level container for developers)
- **NEW!** – NVIDIA TensorRT inference accelerator with ONNX support

NVIDIA GPU CLOUD

3 ステップで始める AWS と NGC でのディープラーニング

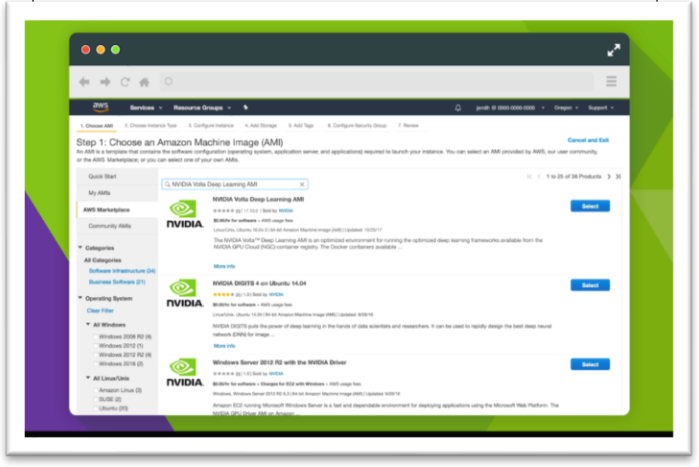
サインアップ

まずは NGC のアカウントを作成してください。

www.nvidia.com/ngcsignup

イメージをデプロイ

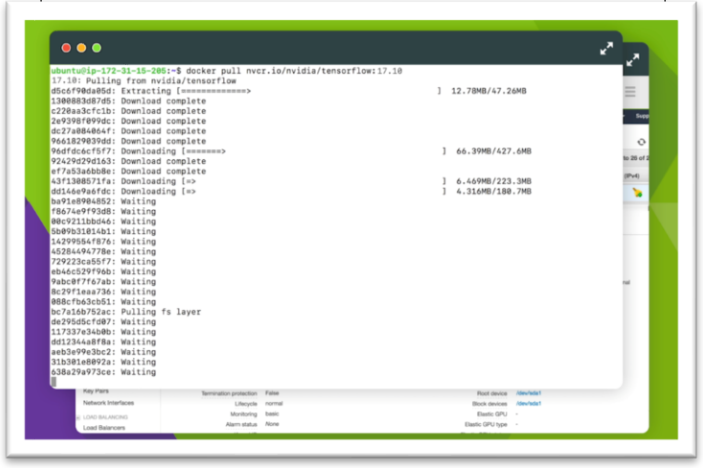
NVIDIA Volta Deep Learning AMI for NGC で P3 インスタンスを作成！



コンテナをPULL

お好みのフレームワーク コンテナをPULLしてください。

これで準備完了です！



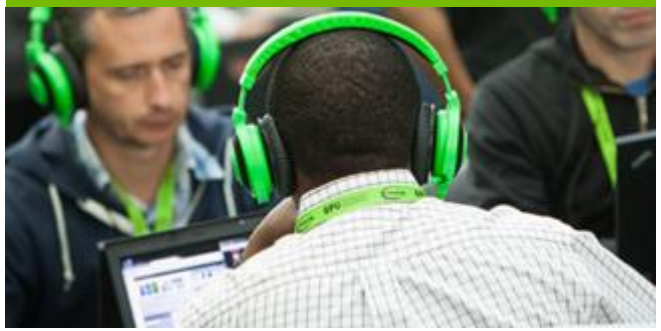
DEEP LEARNING INSTITUTE (DLI)

DEEP LEARNING INSTITUTE

ディープラーニングのハンズオントレーニング

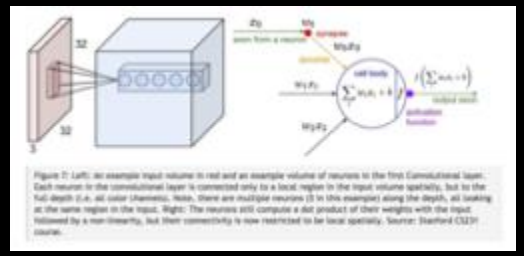
オンサイトワークショップ

自習ラボ



パートナーコース

技術ブログ



QWIKLABS: クラウドベースのハンズオンラボ

手元の PC に GPU は不要。ディープラーニングを簡単に体験。

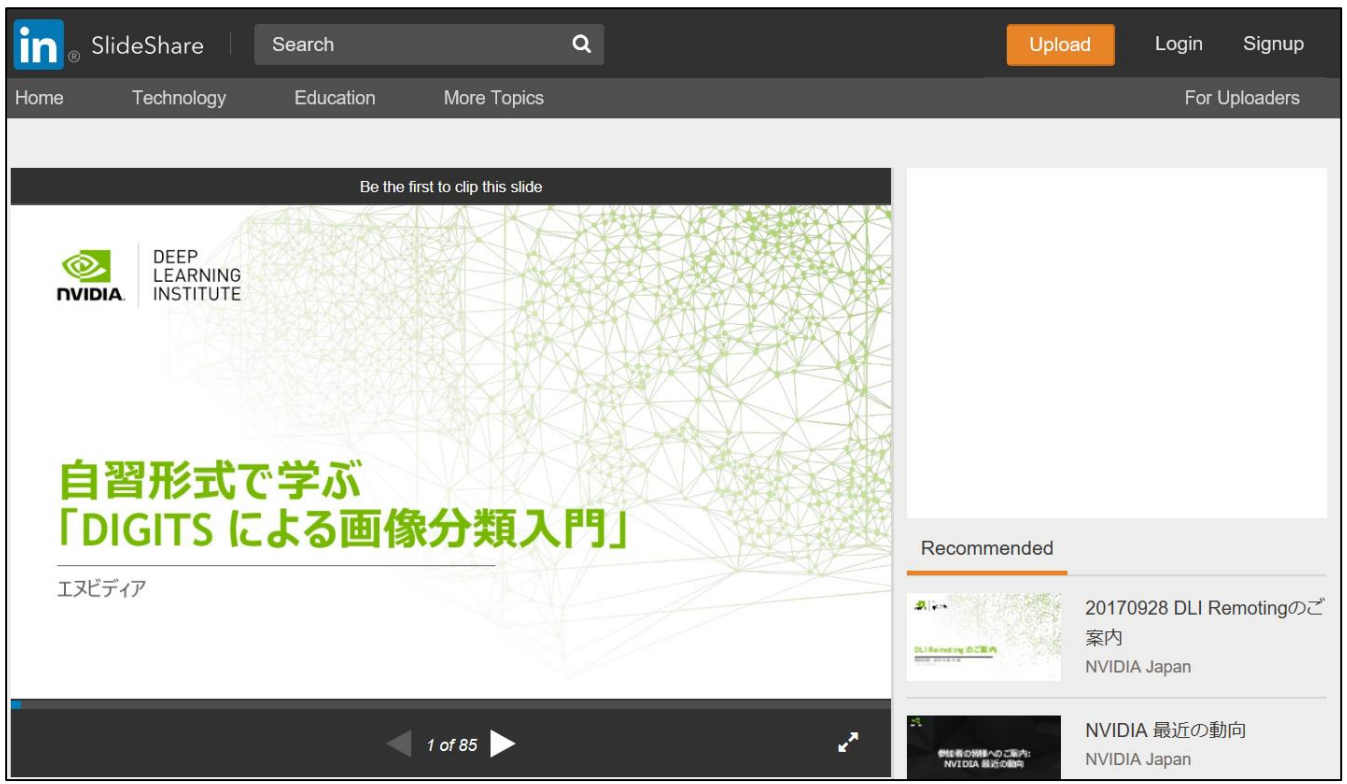
https://nvidia.qwiklab.com に多数のハンズオンラボが揃っています。

「ディープラーニング入門」「Image Classification with DIGITS」の二つは無料（回数制限あり）

ラボタイトル	評価	レベル	価格	完了までの時間
ディープラーニング入門	★★★★☆	Beginner	無料	45分
Deep Learning Workflows with TensorFlow, MXNet and NVIDIA-Docker	★★★★☆	Beginner	30クレジット	1時 30分
Deep Learning for Genomics using DragoNN with Keras and Theano	★★★½☆	Intermediate	30クレジット	1時 30分
Image Classification with DIGITS	★★★★½	Beginner	無料	1時 30分
Image Classification with TensorFlow: Radiomics - 1p19q Chromosome Status Classification	★★★★☆	Intermediate	30クレジット	1時 30分

自習ガイドを公開しました

NVIDIA Japan の SlideShare ページにあります



https://www.slideshare.net/NVIDIAJapan/digits-80145577

「DLI もくもく会」やってます

開催前 2018/01/29（月） 19:00～ 17/20人

DLI もくもく会 #2

yukofuji 他

東京都港区赤坂2-11-7 ATT新館13F

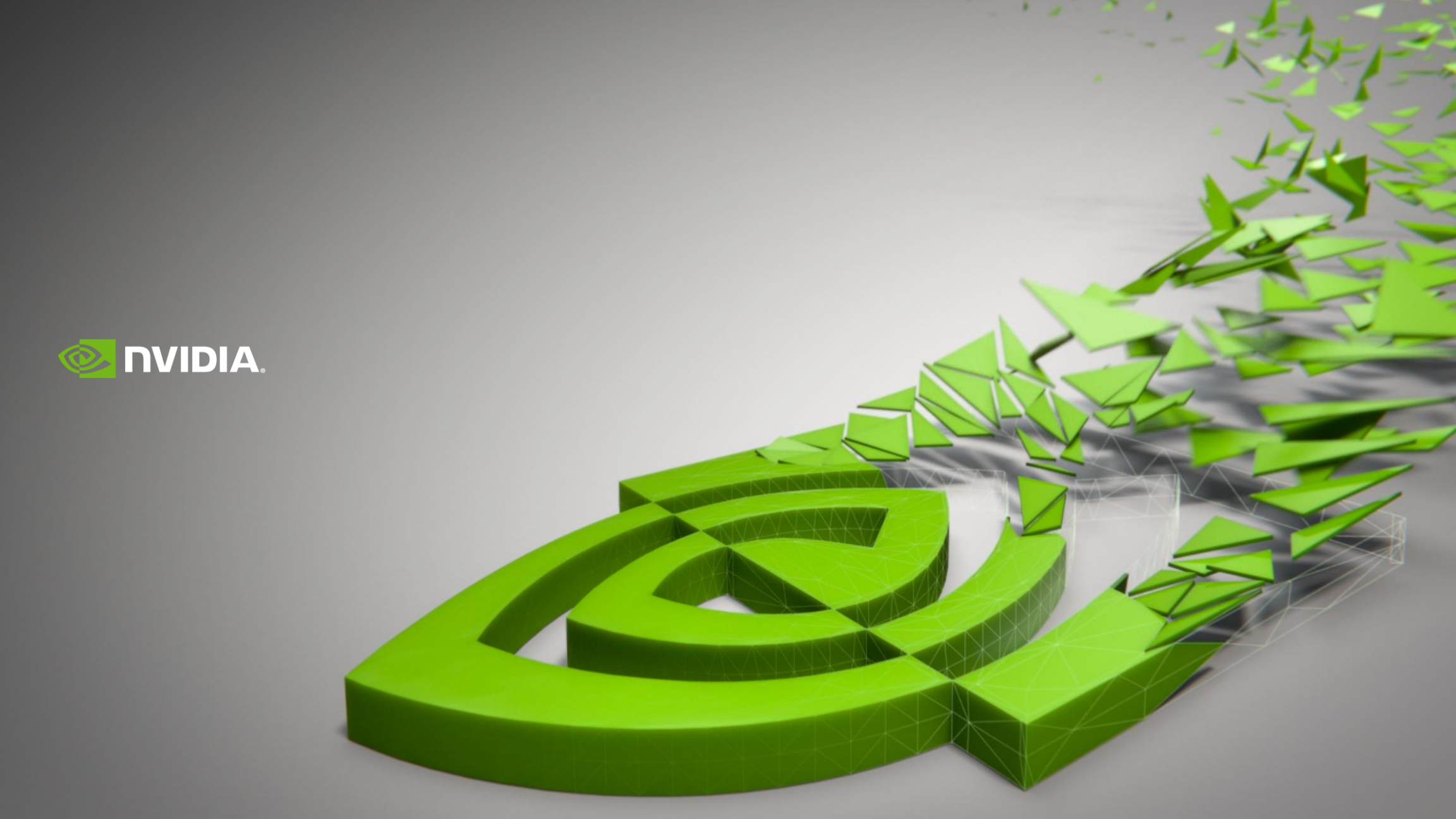
開催前 2018/02/15（木） 19:00～ 18/20人

DLI もくもく会 #3

yukofuji 他

東京都港区赤坂2-11-7 ATT新館13F

https://nvidia.connpass.com/


NVIDIA