GPU TECHNOLOGY CONFERENCE

# エヌビディアが加速する AI 革命

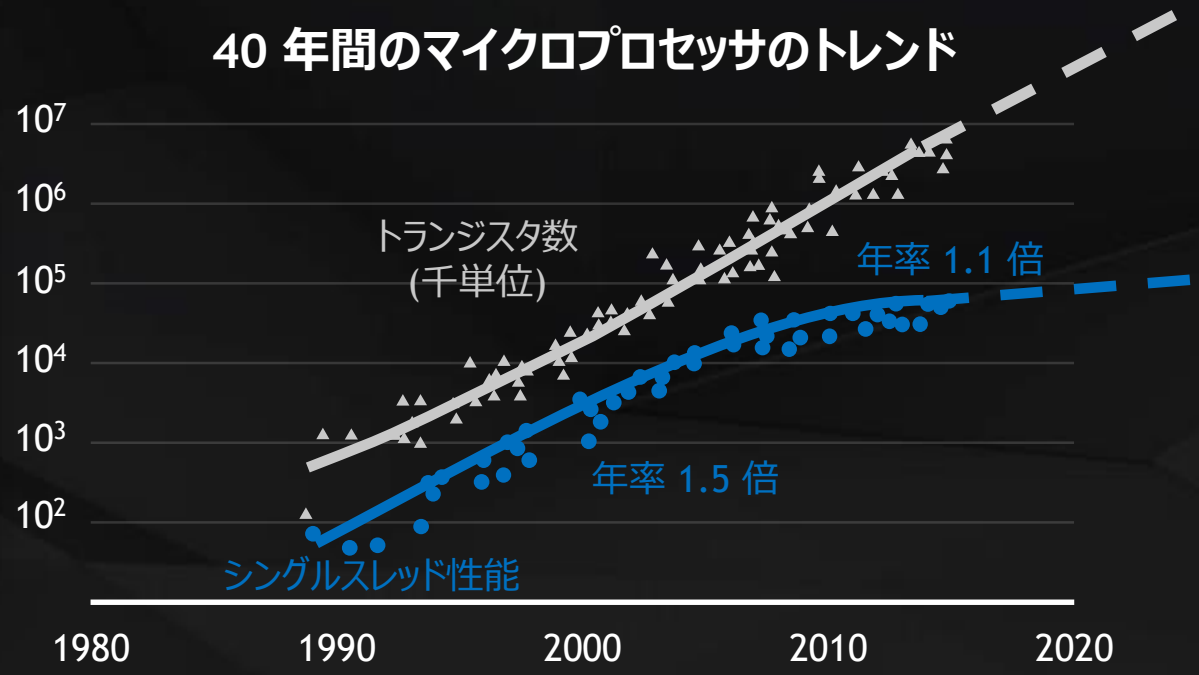
エヌビディア合同会社

エンタープライズマーケティング本部長　林 憲一

# ムーアの法則後の世界

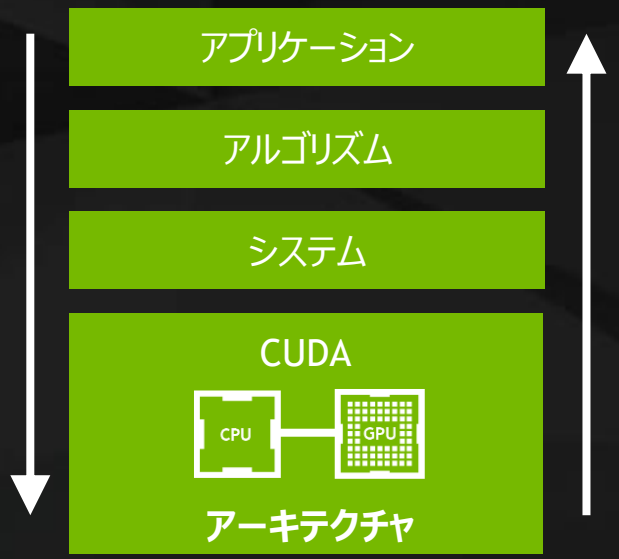
The End of Road for General Purpose Processors and the Future of Computing

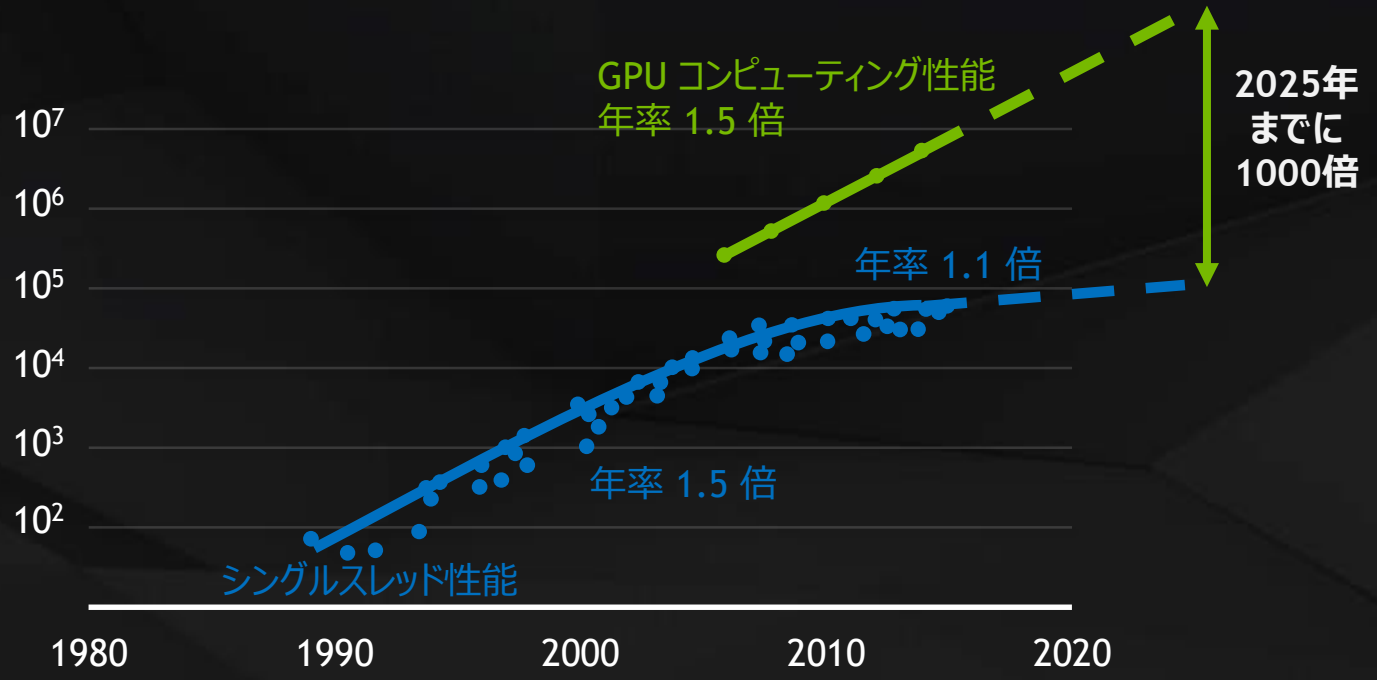
John Hennessy
Stanford University
March 2017




Original data up to the year 2010 collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, and C. Batten New plot and data collected for 2010-2015 by K. Rupp

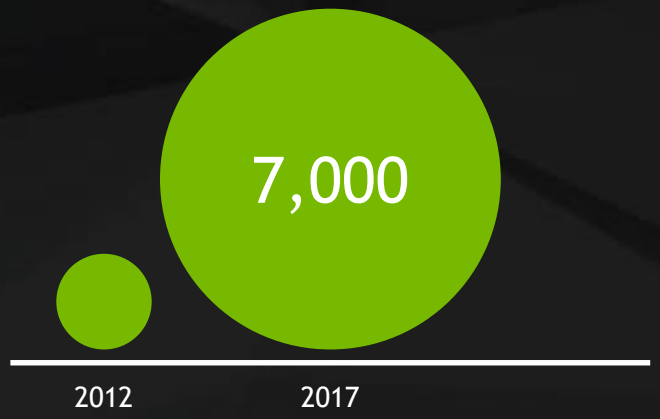
# GPU コンピューティングの登場



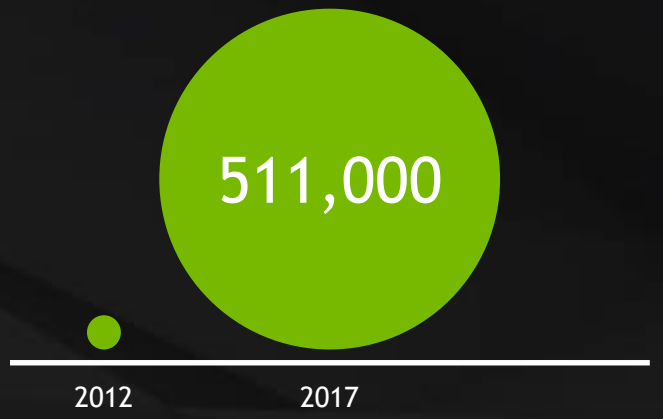




Original data up to the year 2010 collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, and C. Batten New plot and data collected for 2010-2015 by K. Rupp

# GPU コンピューティングの登場



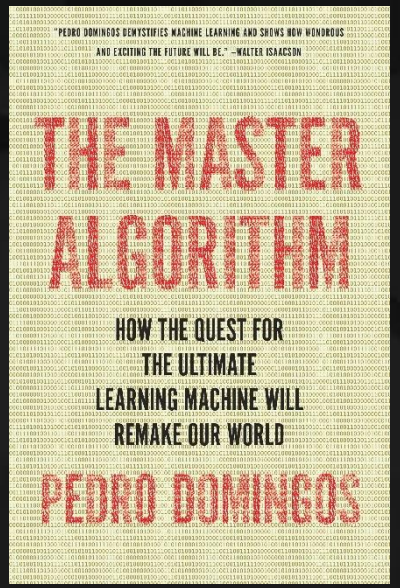

GTC 参加者
5 年で 3 倍



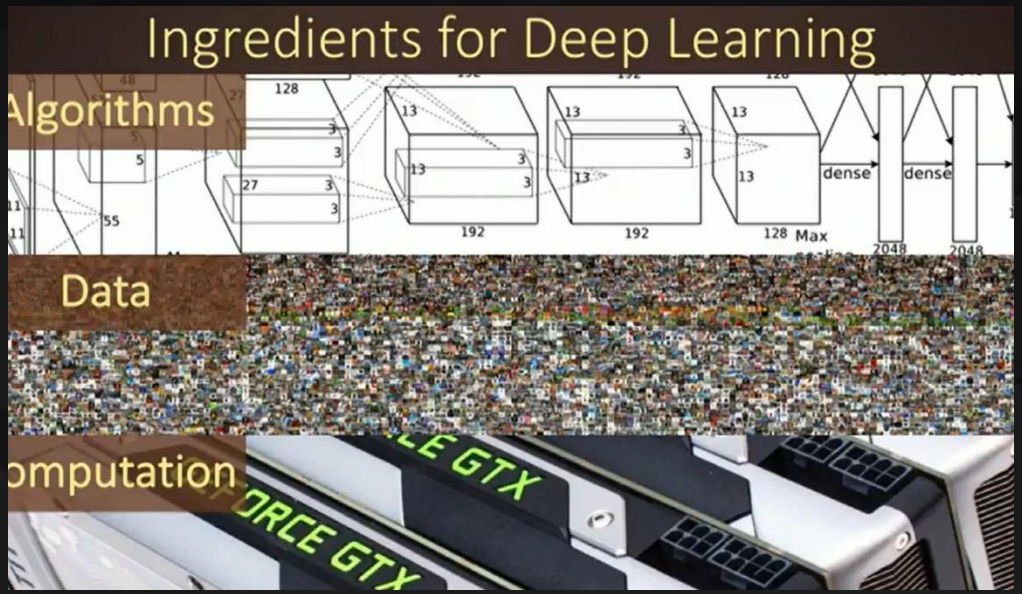

GPU コンピューティング開発者
5 年で 11 倍


1年間 の CUDA ダウンロード数
2016 年

# マシンラーニングの時代




“The Master Algorithm”
— Pedro Domingos




“A Quest for Intelligence”
— Fei-Fei Li

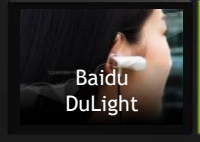
# 現代の AI のビッグバン

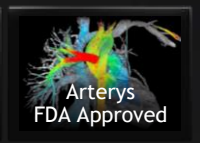
IDSIA
CNN on GPU

U Toronto
AlexNet
on GPU

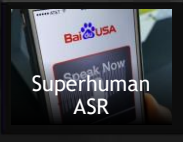

Stanford &
NVIDIA
Large-scale
DNN on GPU






LSTM

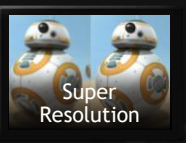



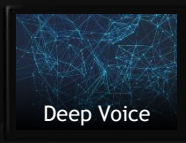

Reinforcement
Learning



Auto
Encoders

GAN

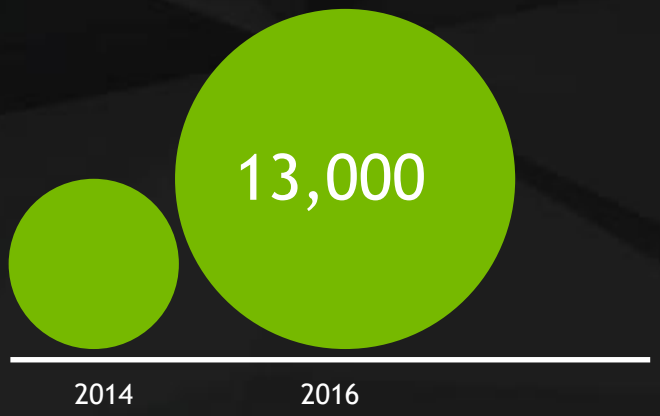


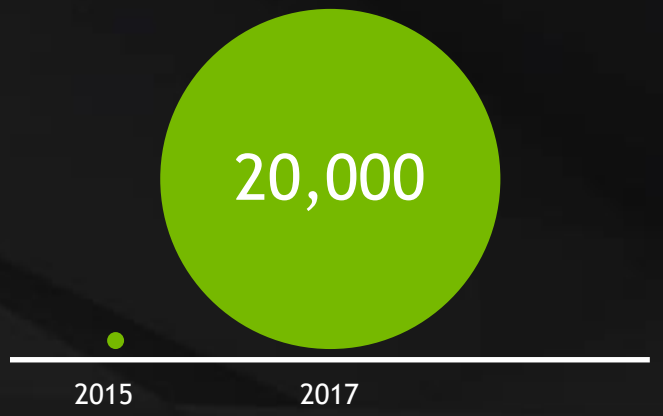


Transfer
Learning

# 現代 AI のビッグバン



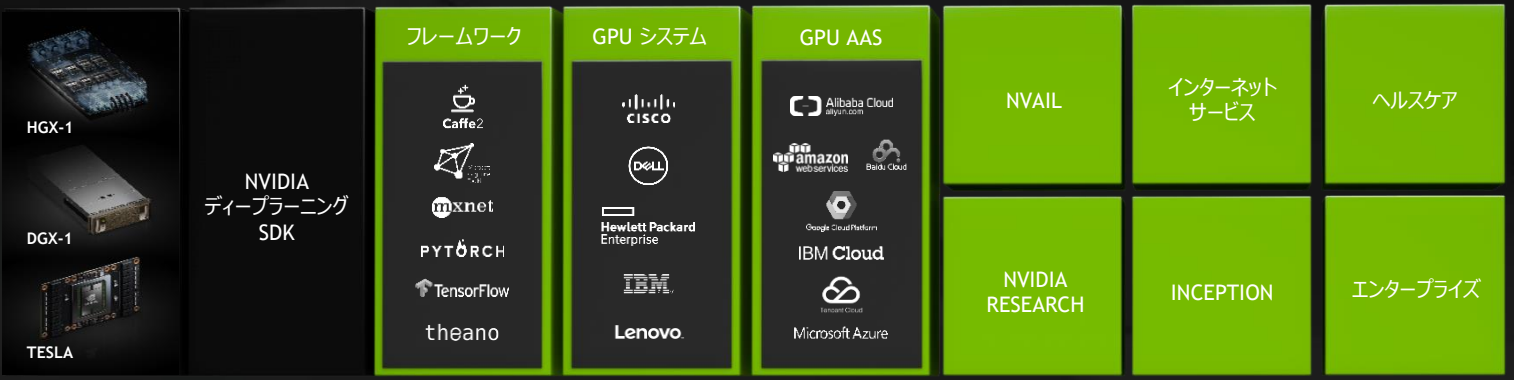

NIPS、ICML、CVPR、ICLR 参加者数
2 年で 2 倍




Udacity でのAI プログラム
2 年で 100 倍


AI スタートアップ投資
4 年で 9 倍

# エヌビディアが加速する AI 革命



# NVIDIA INCEPTION プログラム
# 1300 のディープラーニングスタートアップを支援

## ヘルスケア

athelas
Atomwise
ARTERYS
babylon
BAYLABS
CLOUD MEDX
deep genomics
Genetesis
imagia
lumiata
PHENOMIC AI
zebra

## リテール eテール

focal
GOFIND
GROKSTYLE
huew
MARIANA
mashgin
netra
ProductAI
shoppr

## 金融

Alpaca
CAPE
Deep Learni.ng
Orbital Insight
Quantenstein

## セキュリティ

aiRXsys
ENTROPIX
herta
KUNA
umoo cv

## IOT 及び 製造

3DSignals
ABEJA
CLOUDBRAIN
IFM
KONUX
Preferred Networks
PREDII
Reliability Solutions.
VERDIGRIS
VIMOC Technologies

## 自律動作機械

AIMOTIVE
ARGO AI
BLUE RIVER TECHNOLOGY
dispatch
drive.ai
marble
nuTonomy
OPTIMUS RIDE
pilot.ai
图森 tuSimple
ZENUITY
ZOOX

## サイバー

CYLANCE
deepinstinct
UNIFYID
GreatHorn

## 建築

CUBICASA
HOVER
PRENAV
SMARTVID.IO

## プラットフォーム 及び API

clarifai
deep gram
Gridspace
KITT.AI
7INVENSUN
VisionLabs

## データマネージメント

CrowdAI
DATALOGUE
understand.ai

## 開発プラットフォーム

bonsai
SIGOPT
FLOYD
$H_2O.ai$
skymind

## ビジネスインテリジェンス 及び 可視化

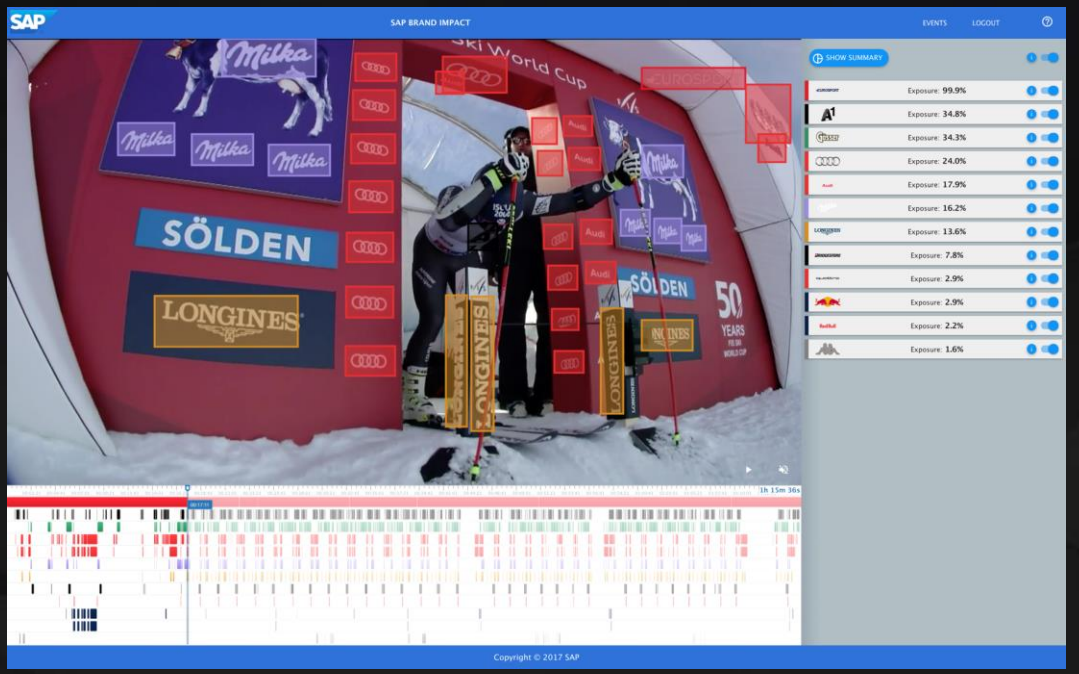
Blazing
graphistry
kinetica
MAPD
SQREAM TECHNOLOGIES

# エンタープライズのための SAP AI

SAP から提供される最初の商業的 AI オファリング

Brand Impact、Service Ticketing、Invoice-to-Record アプリケーション

DGX-1 と AWS で NVIDIA GPU を利用



# 年々複雑さを増すモデル

# 発表 Tesla V100

## AI と HPC のための大きな飛躍
## Tensor コアを搭載した Volta アーキテクチャ

210 億トランジスタ | TSMC 12nm FFN | 815mm$^2$

5120 CUDA コア

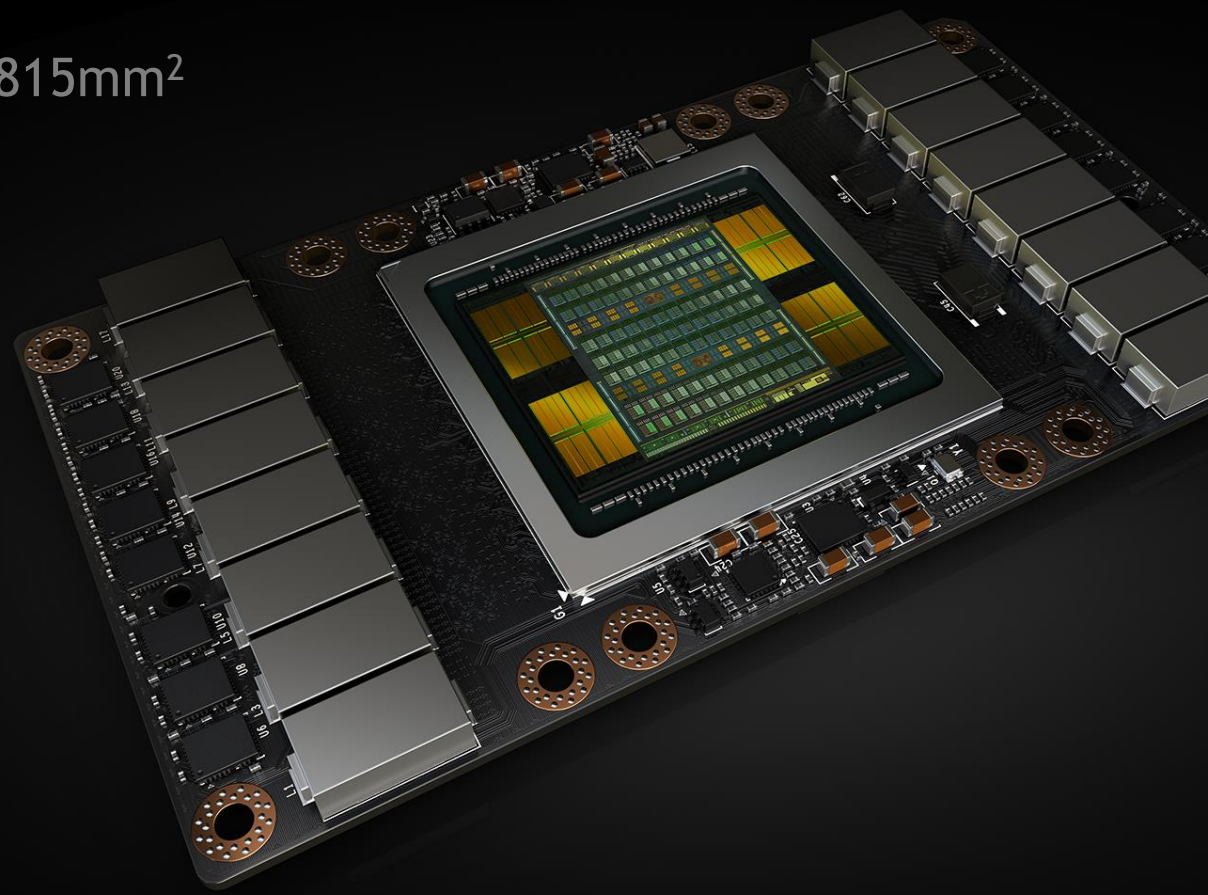
7.5 FP64 TFLOPS | 15 FP32 TFLOPS

120 Tensor TFLOPS

総レジスタファイル 20MB | 16MB キャッシュ

900 GB/s の 16GB HBM2

300 GB/s NVLink



# 新開発 Tensor コア

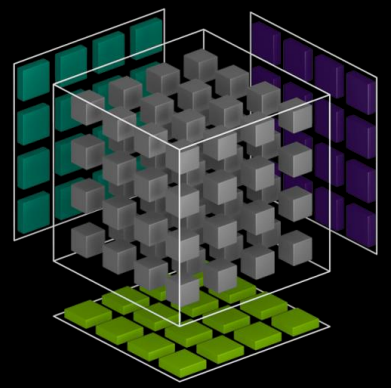
CUDA Tensor 演算命令 及び データフォーマット

4x4 行列処理配列

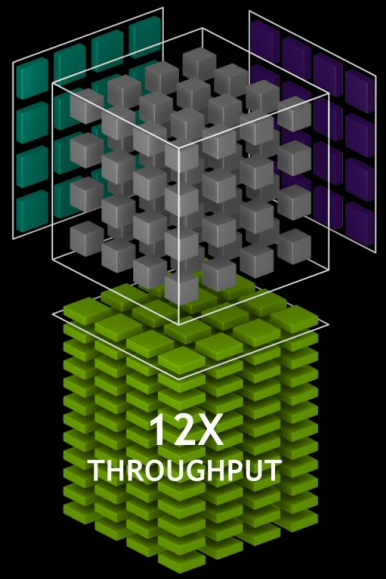
D[FP32] = A[FP16] * B[FP16] + C[FP32]

ディープラーニングに最適化

PASCAL



VOLTA TENSOR CORES



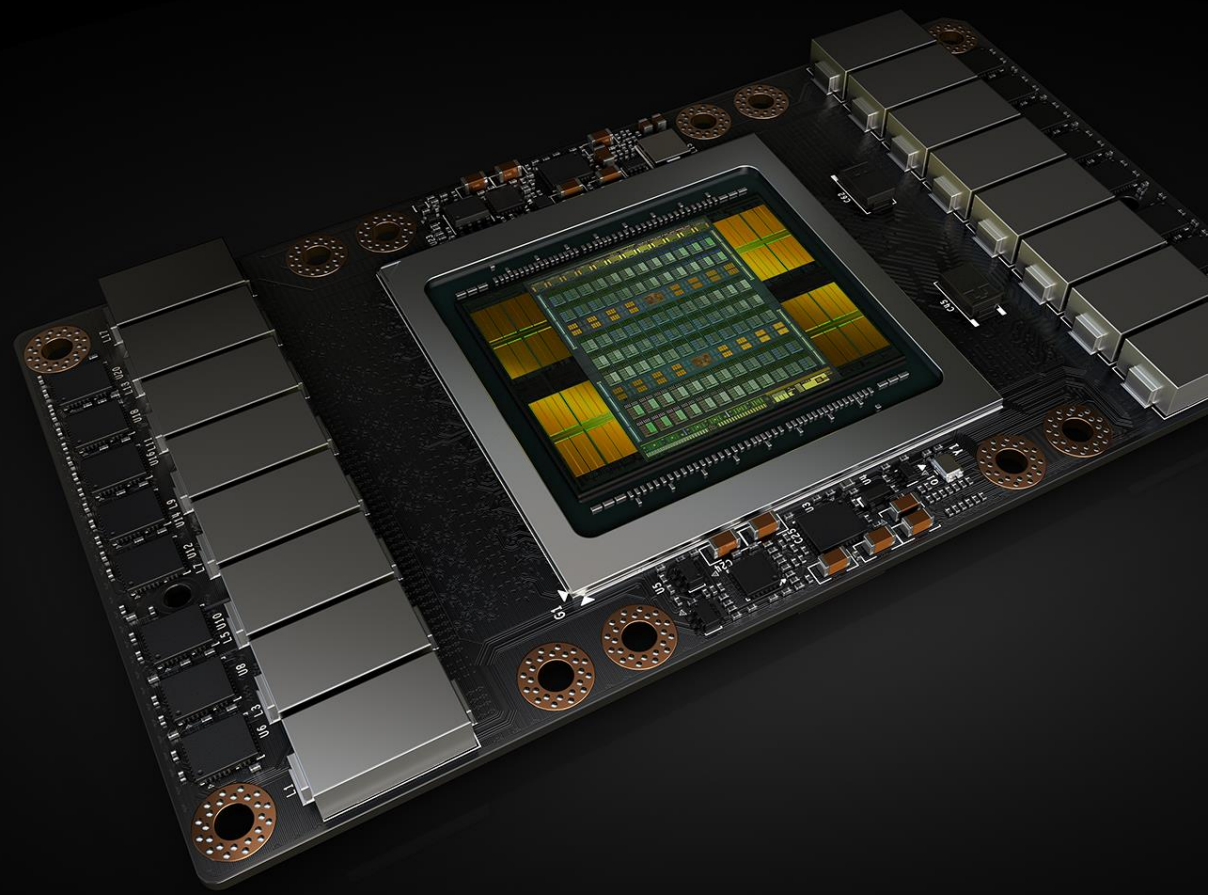

アクティベーション入力　重み入力　出力結果

# 発表 Tesla V100

## AI と HPC のための大きな飛躍
## Tensor コアを搭載した Volta アーキテクチャ

Pascal 世代と比較して

- HPC のための汎用演算性能 1.5 倍
- ディープラーニングトレーニングのための
  - Tensor 演算性能 12 倍
- ディープラーニング推論のための
  - Tensor 演算性能 6 倍



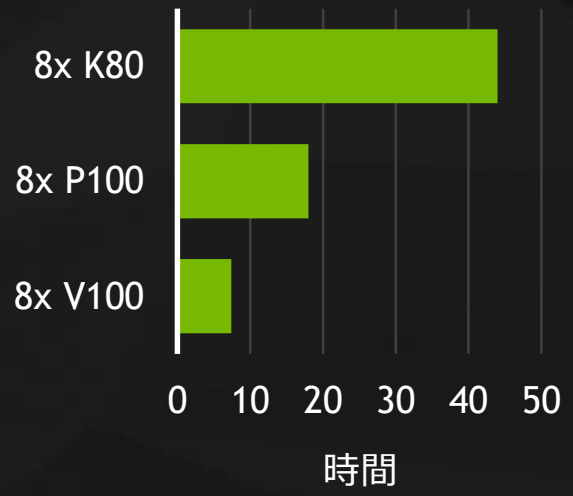
# 発表
# Volta 対応フレームワーク

Caffe2

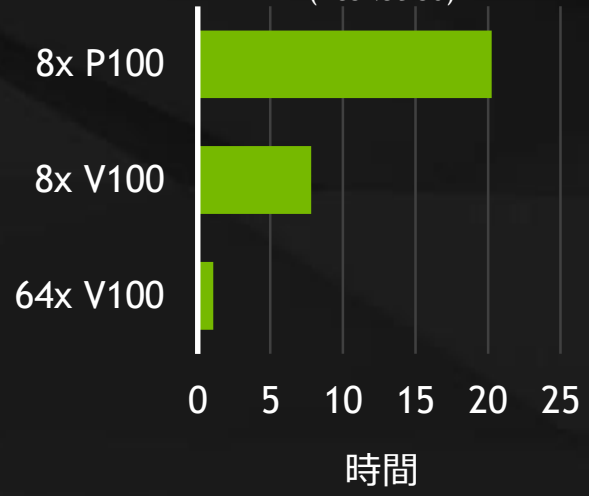
Microsoft Cognitive Toolkit

mxnet

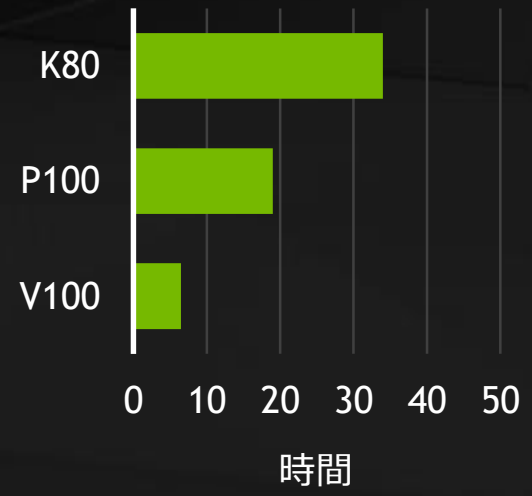
**CNN トレーニング**
(ResNet-50)




**NCCL 2.0 を利用したマルチノード トレーニング**
(ResNet-50)




**LSTM トレーニング**
(ニューラル機械翻訳)



# 発表
# Tesla V100 搭載 NVIDIA DGX-1
## AI 研究に必須の道具

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

# 発表
# Tesla V100 搭載 NVIDIA DGX-1

## AI 研究に必須の道具

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

ご注文は: nvidia.com/DGX-1

DEP
HITACHI Inspire the Next
HPC SYSTEMS
HPC TECH
Panasonic

# 発表
# NVIDIA DGX ステーション
## パーソナル DGX

480 Tensor TFLOPS | Tesla V100 4基

NVLink 全結合 | 3つの DisplayPort

1500W | 水冷

# 発表
# NVIDIA DGX ステーション
## パーソナル DGX

480 Tensor TFLOPS | Tesla V100 4基

NVLink 全結合 | 3つの DisplayPort

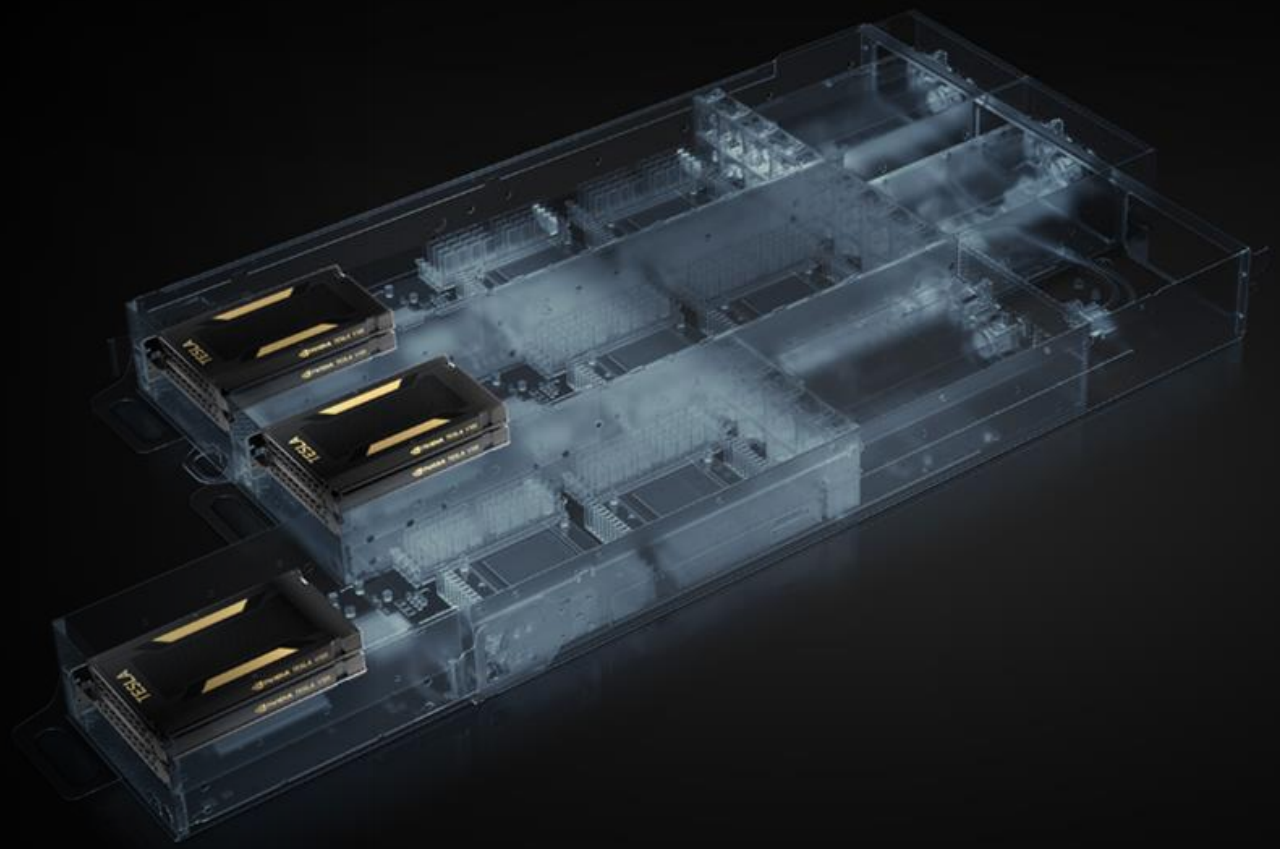
1500W | 水冷

ご注文は: nvidia.com/DGX-Station

# 発表
# ハイパースケール推論のための Tesla V100

Skylake に対して 15～25 倍の推論性能

150W | FHHL PCIE



# GPU で加速されたデータセンターの例

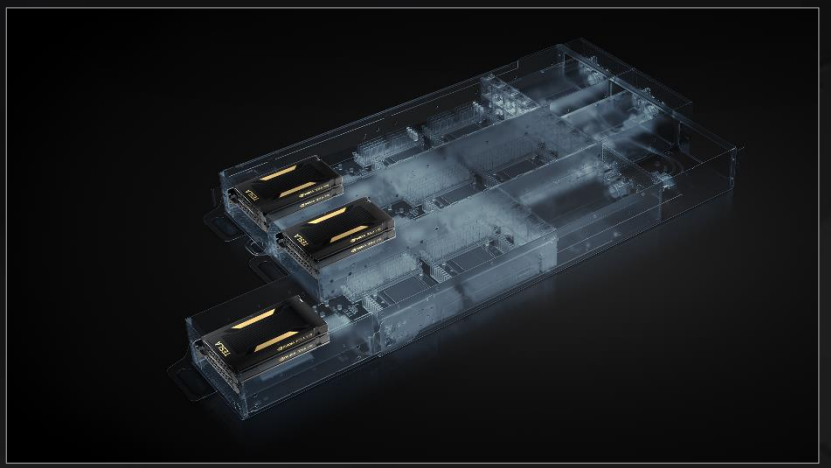
データセンター全体で300,000 推論/秒

CPU 当り 300 推論/秒 ➡ 1000 CPU

1000 CPU ➡ 500 ノード

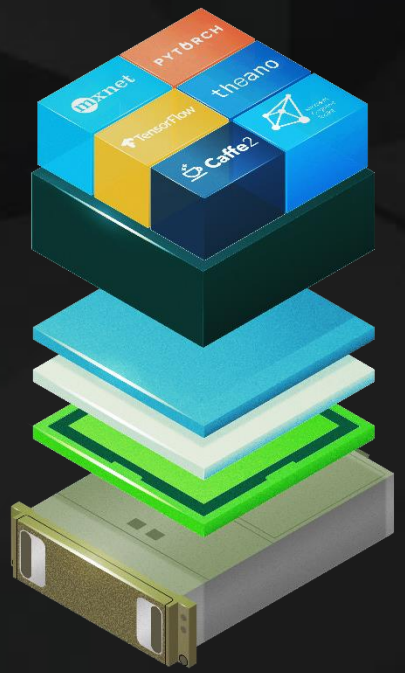
ノード 3000ドル ➡ 150万ドル

ノード 500W ➡ 250KW



500 ノードの CPU サーバー ⟶ Tesla V100 によって 1/15に ⟶ 33 ノードの GPU で加速されたサーバー

# NVIDIA ディープラーニングスタック




ディープラーニングフレームワーク

ディープラーニングライブラリ
NVIDIA cuDNN, NCCL,
cuBLAS, TensorRT

CUDA ドライバ

オペレーティングシステム

GPU

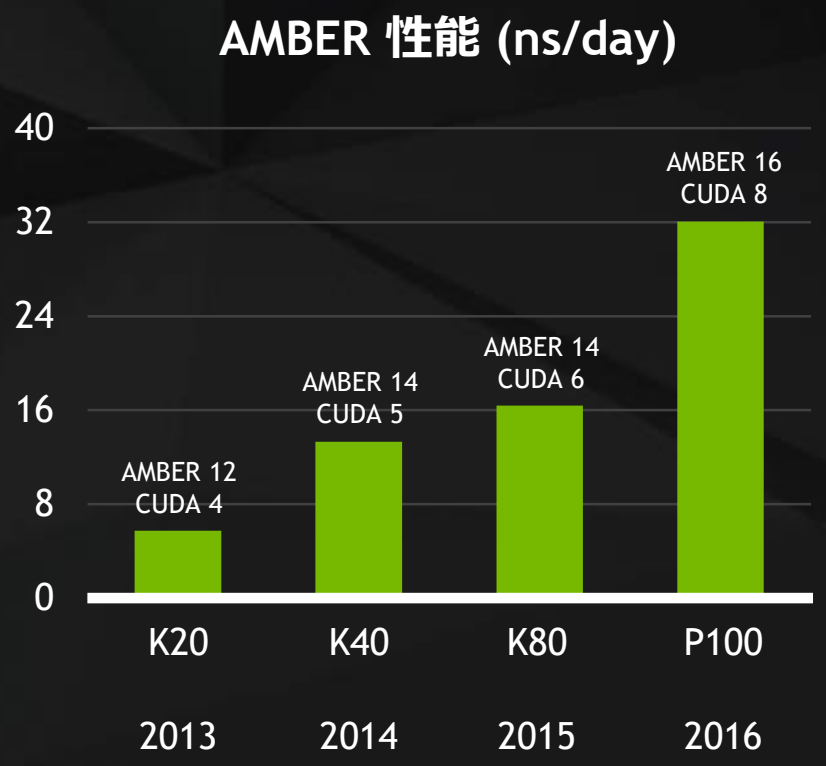
システム

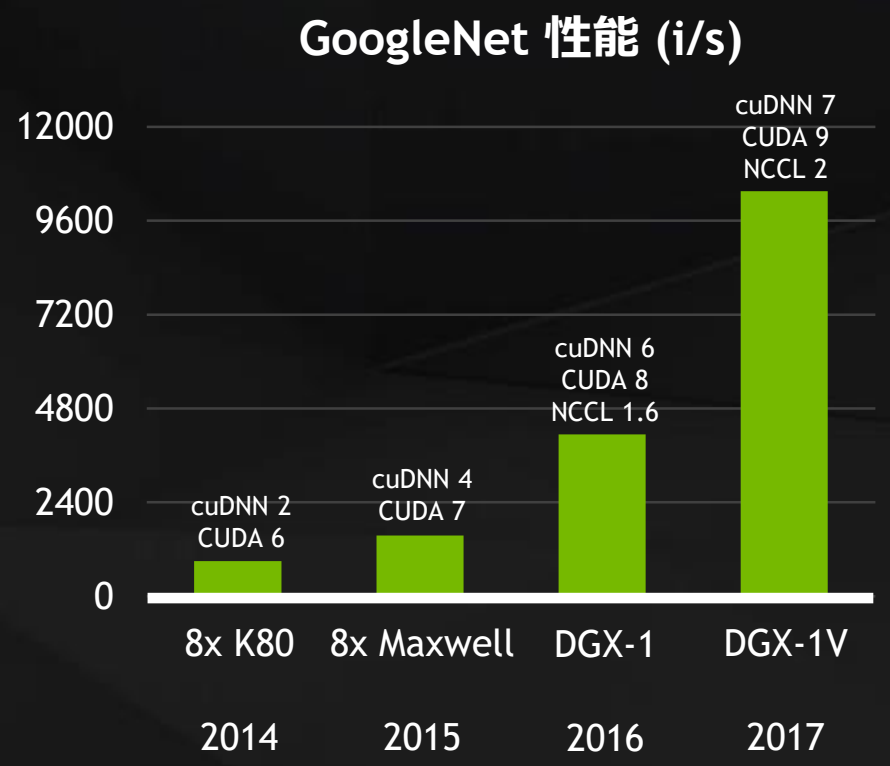
# 発表
# NVIDIA GPU クラウド

## ディープラーニングに最適化された GPU で加速されたクラウドプラットフォーム


NVDocker のコンテナとして提供 | フルスタックで最適化
常に最新 | エヌビディアによって完全にテストおよびメンテナンス | 7月にベータ提供

# GPU コンピューティング性能



# エヌビディアが加速する AI 革命

エッジでの AI

# AI が革新するトランスポーテーション

年間 4500億キロ

米国では 2億5000万台の車のために
8億の駐車場

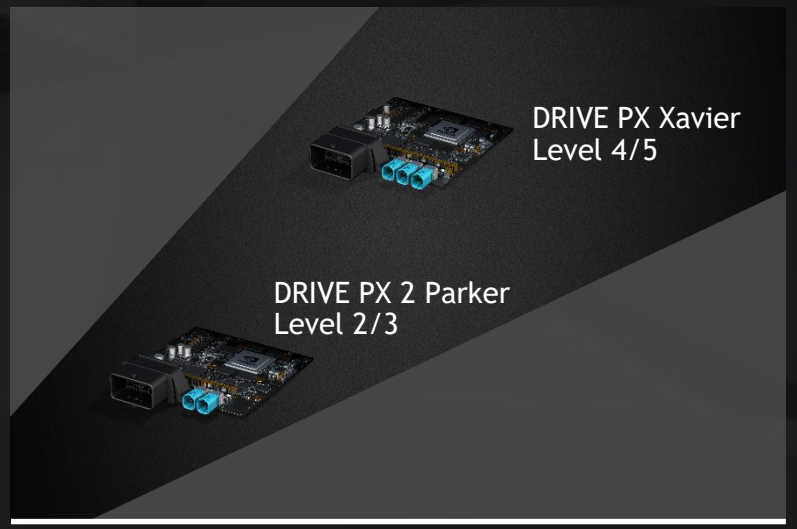
ドミノピザ: 一日100万個のピザを配達

# NVIDIA DRIVE — AI カープラットフォーム

100 TOPS

10 TOPS

1 TOPS




自己位置推定

パスプランニング

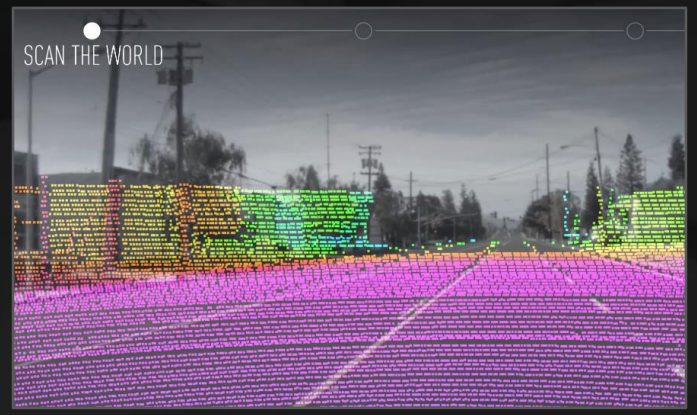
認識 AI

コンピュータビジョンライブラリ

CUDA、cuDNN、TensorRT

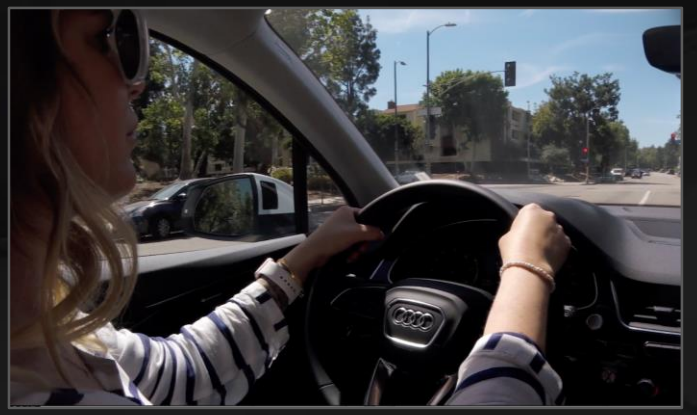
OS

# NVIDIA DRIVE




マッピングから運転へ

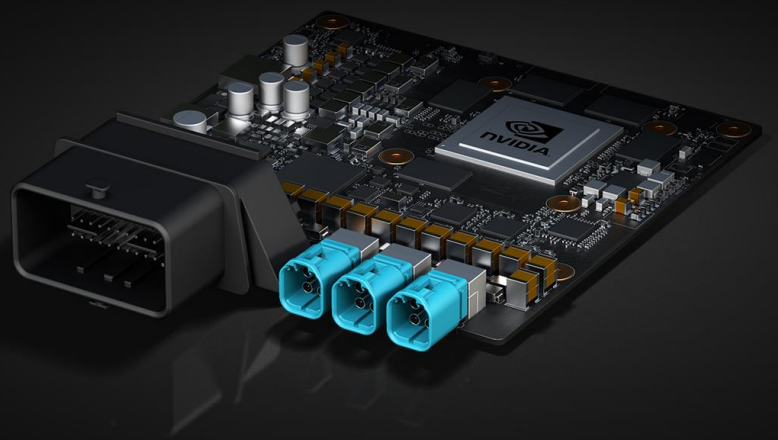

コ パイロット



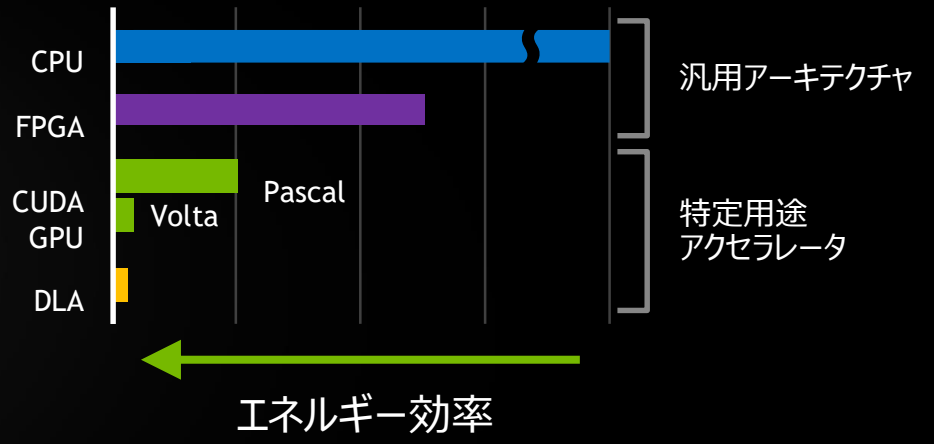
ガーディアン エンジェル

# 発表
# トヨタ、自動運転車向けに NVIDIA DRIVE PX を選択

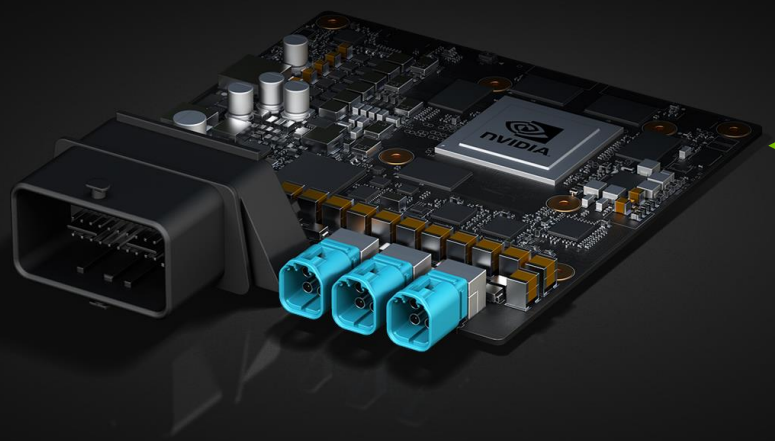
# 自動運転のための AI プロセッサ



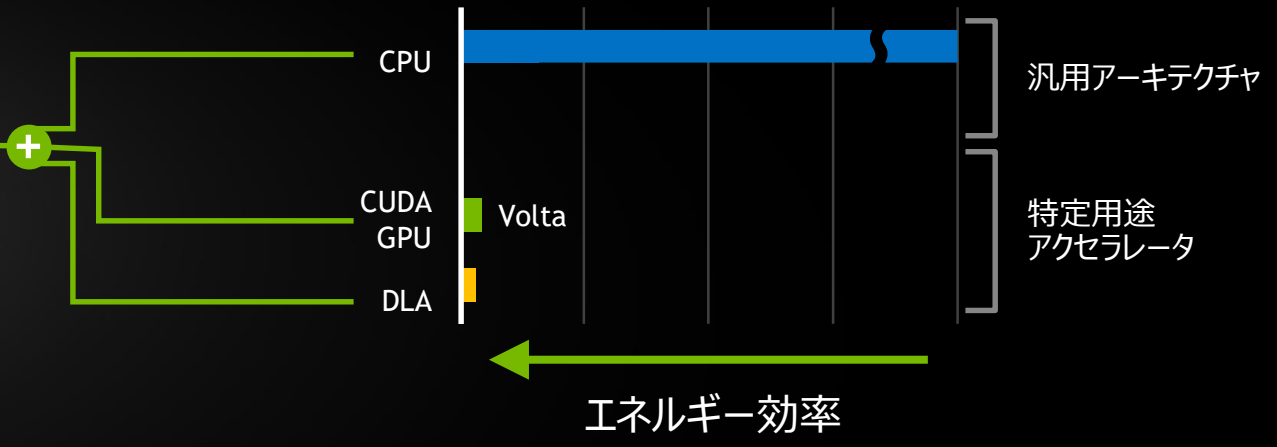

XAVIER
30 TOPS DL
30W
カスタム ARM64 CPU
512 コア Volta GPU
10 TOPS DL アクセラレータ



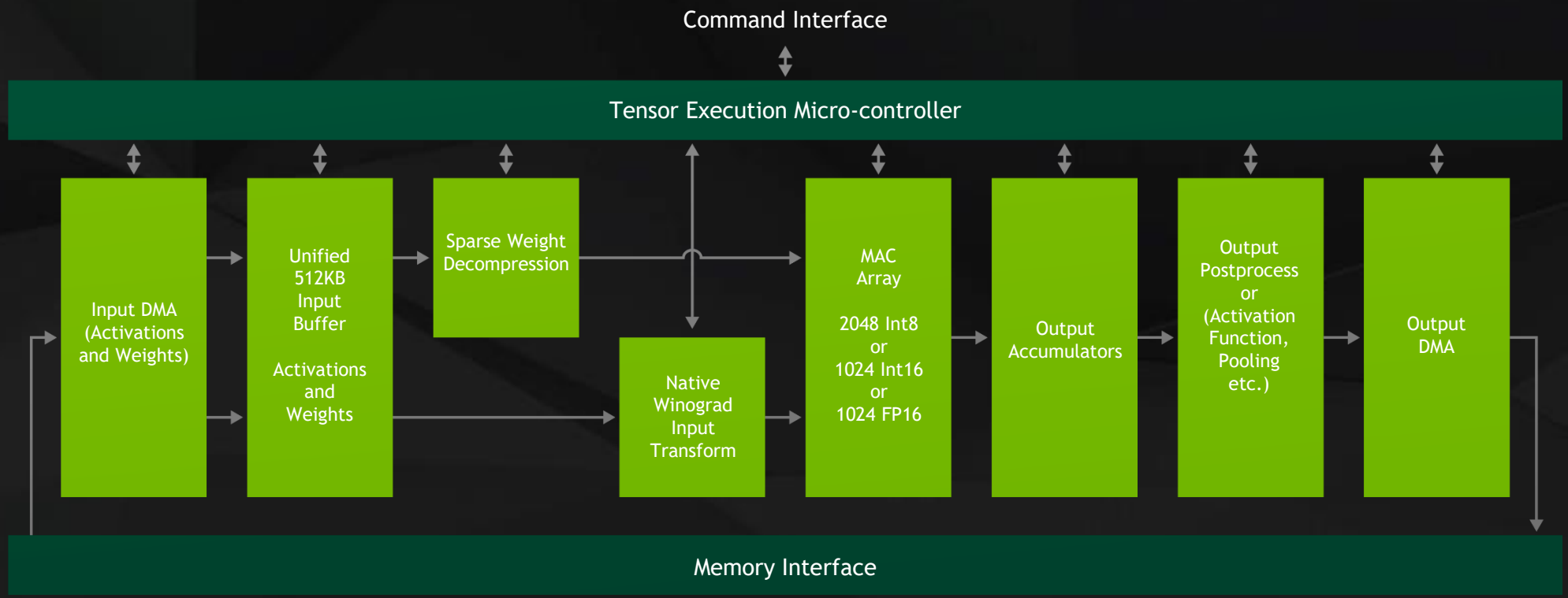
# 自動運転のための AI プロセッサ



XAVIER
30 TOPS DL
30W
カスタム ARM64 CPU
512 コア Volta GPU
10 TOPS DL アクセラレータ

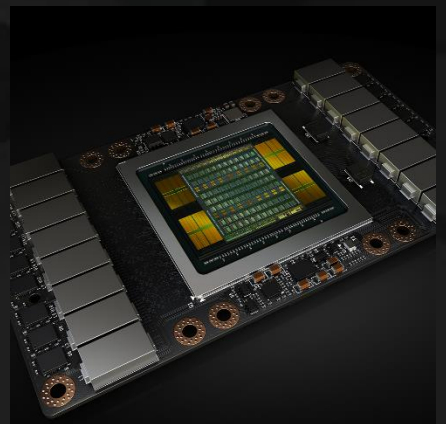


# 発表
# Xavier DLA オープンソース化

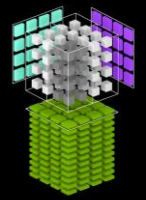



アーリーアクセス予定: 7 月 | 一般リリース予定: 9 月

# エヌビディアが加速する AI 革命



Tesla V100



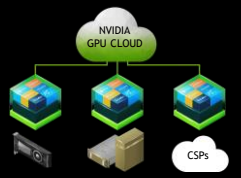
Tensor コア

TensorRT

Alibaba amazon Baidu 百度 facebook

Google Microsoft Tencent

全てのクラウドに NVIDIA GPU




NVIDIA GPU クラウド

DGX-1 及び DGX ステーション

Xavier DLA
オープンソース化

TOYOTA


GPU TECHNOLOGY
CONFERENCE