スパコンTSUBAME利用促進シンポジウム

GPUで始めるディープラーニング入門

エヌビディア合同会社

ディープラーニング　ソリューションアーキテクト兼CUDAエンジニア　村上真奈

NVIDIA

AGENDA

エヌビディアについて

ディープラーニングとは?

GPUディープラーニング / ディープラーニングSDK

エヌビディアについて

創業1993年

共同創立者兼CEO ジェンスン・フアン
（Jen-Hsun Huang）

1999年 NASDAQに上場（NVDA）

1999年にGPUを発明
その後の累計出荷台数は1億個以上

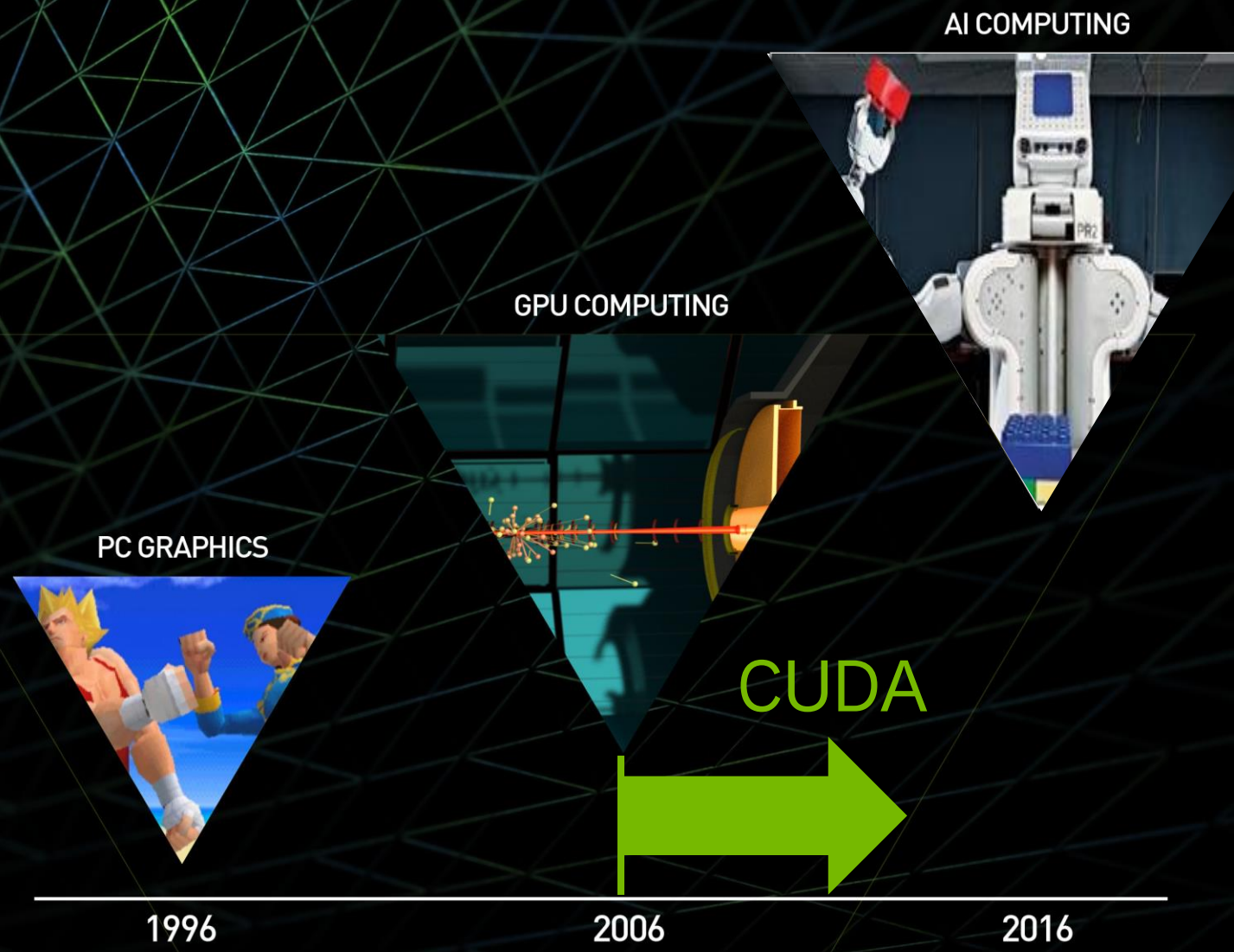
2016年度の売上高は50億ドル

社員は世界全体で9,227人

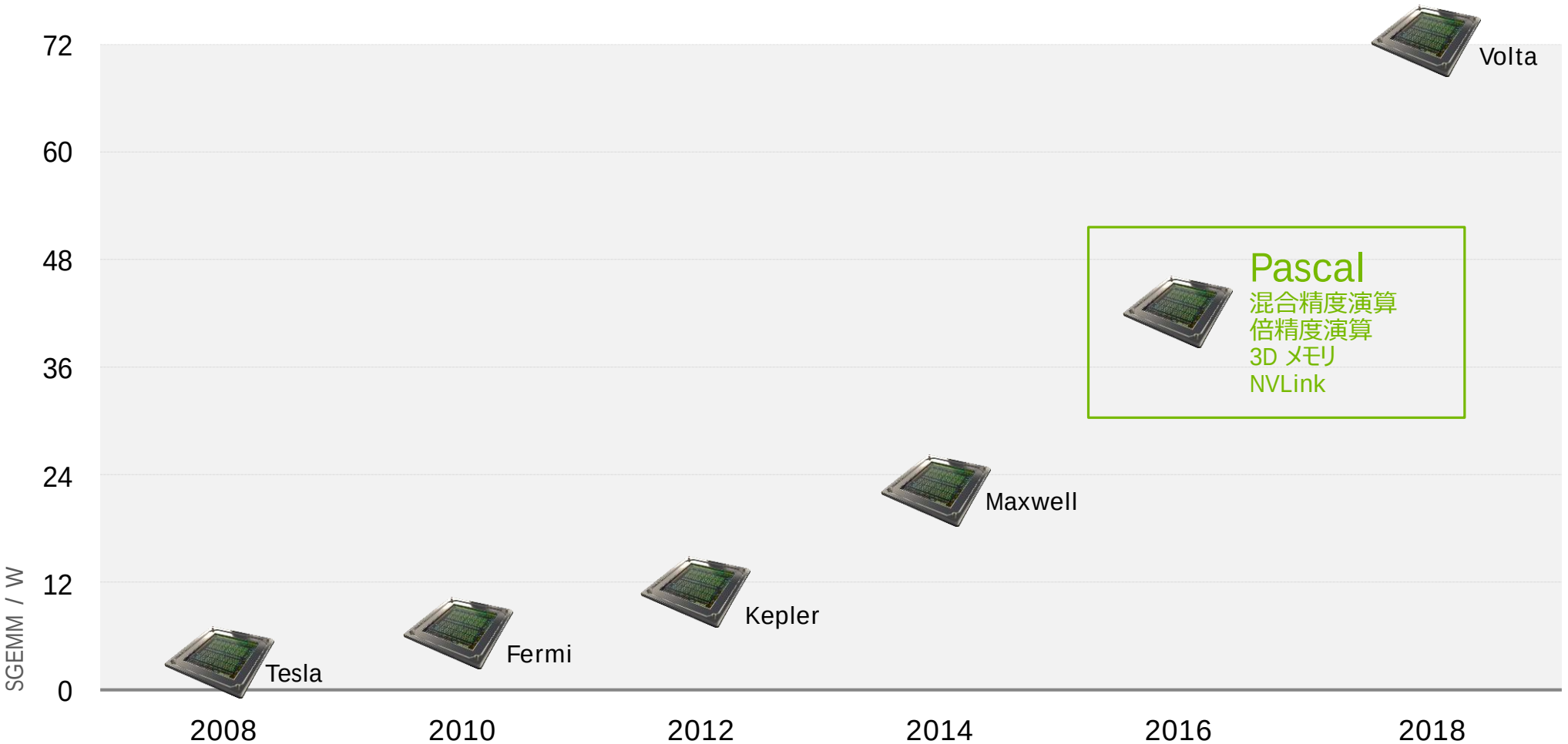
約7,300件の特許を保有

本社は米国カリフォルニア州サンタクララ

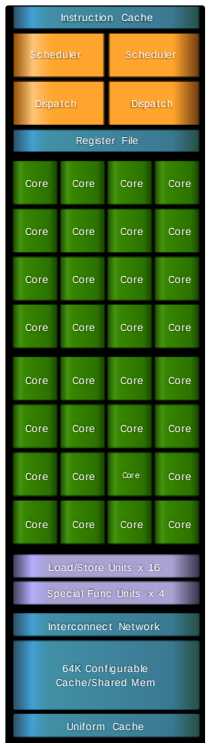
NVIDIA GPU の歴史



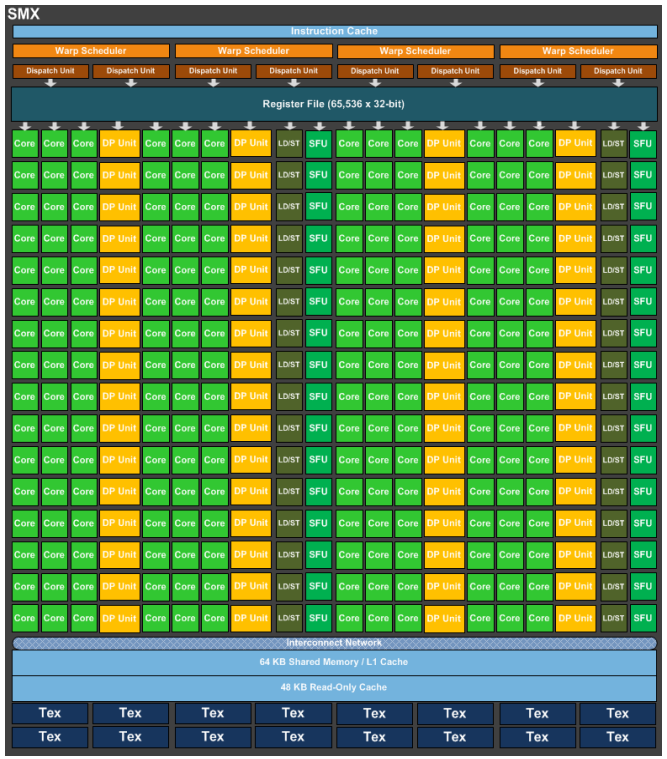
GPU ロードマップ


72
60
48
36
24
12
0
SGEMM / W
2008
2010
2012
2014
2016
2018
Tesla
Fermi
Kepler
Maxwell
Pascal
混合精度演算
倍精度演算
3D メモリ
NVLink
Volta

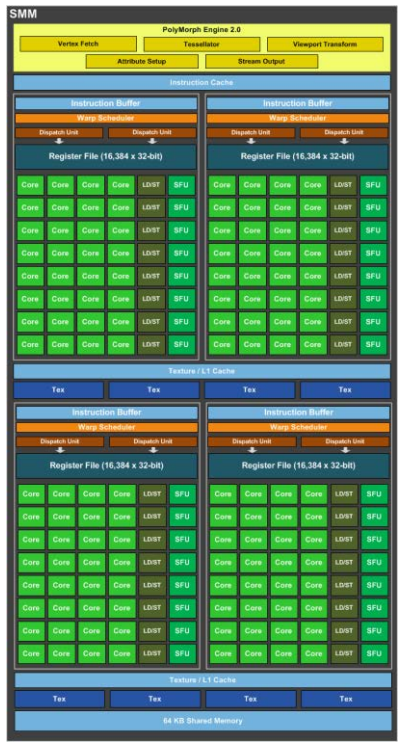
Compute Capability (CC)



Fermi
CC 2.0
32 cores / SM



Kepler
CC 3.5
192 cores / SMX



Maxwell
CC 5.0
128 cores / SMM

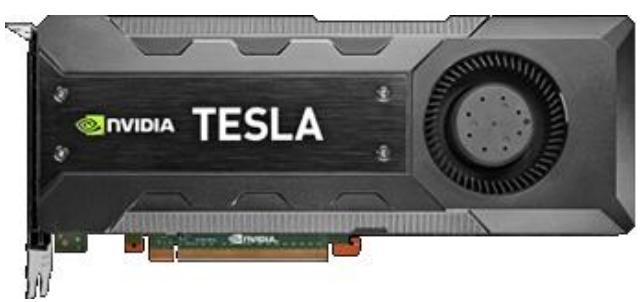
Pascal
CC 6.0
64 cores / SMM

https://developer.nvidia.com/cuda-gpus

Tesla K20x (CC 3.5)

2688 CUDA Cores
- FP64: 1.31 TF
- FP32: 3.95 TF
- FP16: …
- INT8: …

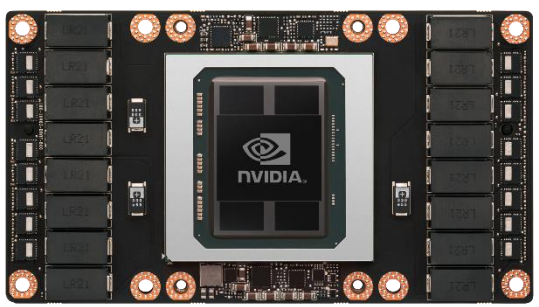
GDDR5
- 384 bit width
- 6GB
- 250 GB/s

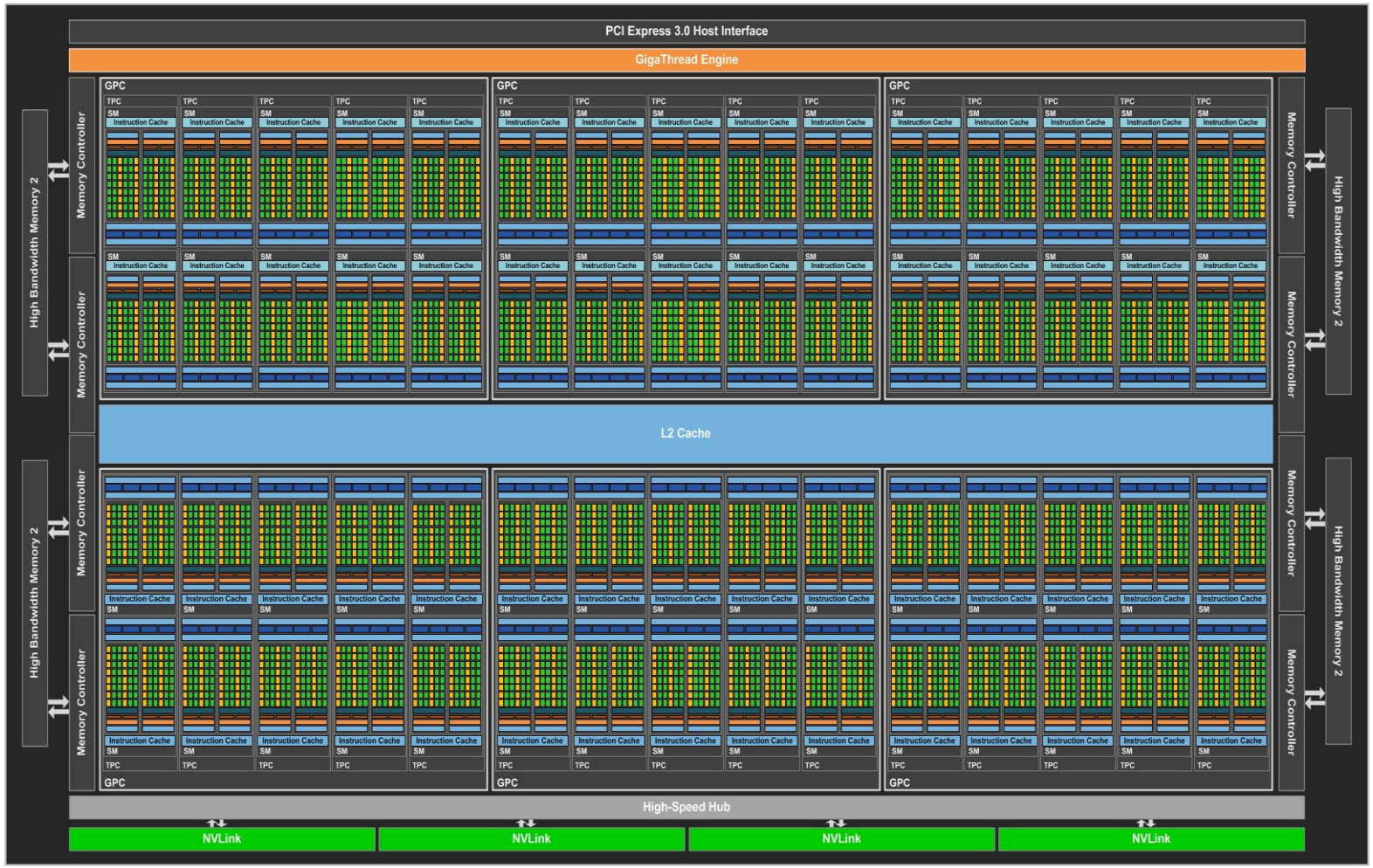


Tesla P100 SXM2 (CC 6.0)

3584 CUDA Cores
- FP64: 5.3 TF
- FP32: 10.6 TF
- FP16: 21.2 TF
- INT8: …

HBM2
- 4096 bit width
- 16 GB
- 732 GB/s

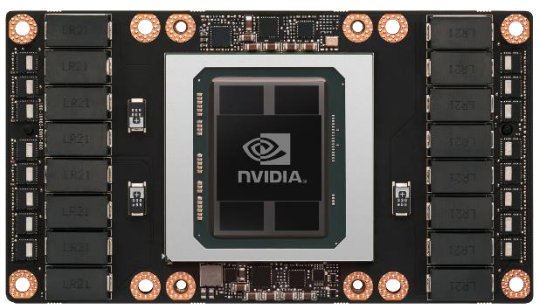


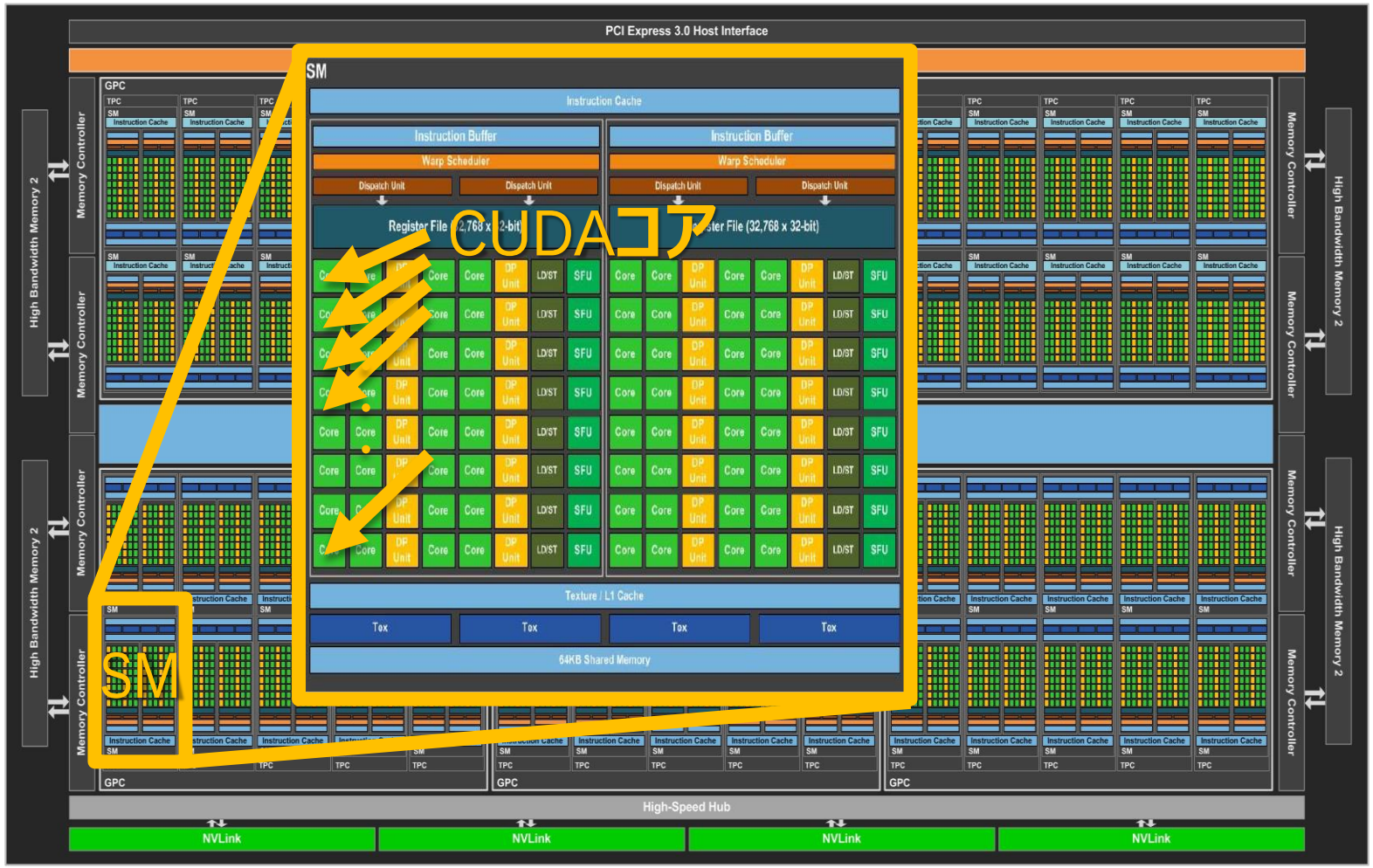


Tesla P100 SXM2 (CC 6.0)

3584 CUDA Cores
- FP64: 5.3 TF
- FP32: 10.6 TF
- FP16: 21.2 TF
- INT8: …

HBM2
- 4096 bit width
- 16 GB
- 732 GB/s

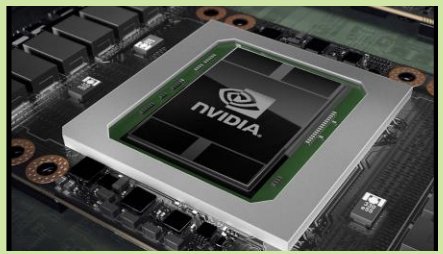




TESLA P100

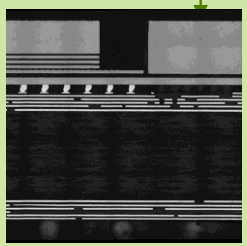
最速の計算ノードを実現する新しい GPU アーキテクチャ

Pascalアーキテクチャ



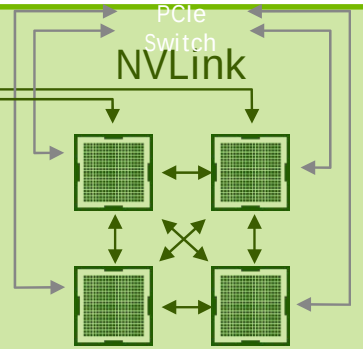
最高の演算能力

HBM2 積層メモリ



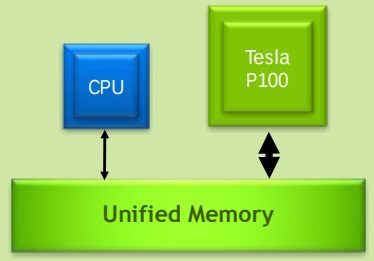
演算とメモリを一つの パッケージに統合

NVLink

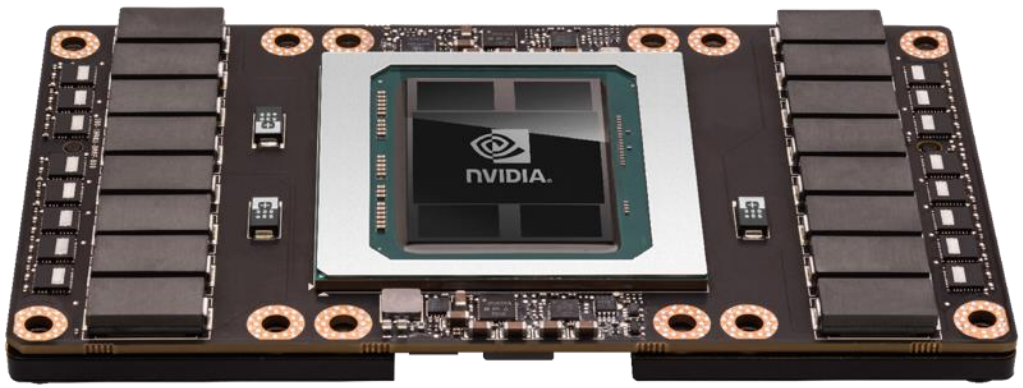


スケーラビリティを最大化する GPU インターコネクト

ページマイグレーションエンジン



広大な仮想アドレス空間で シンプルな並列プログラミング



CUDA

GPU超並列コンピューティングプラットフォーム

CUDA8.0が最新

マルチOSで動作
(Windows,Linux(x86,power,arm),MacOS)

様々なライブラリや開発環境を提供

ライブラリ・ディレクティブ

OpenACC / NVIDIA CUDA Libraries /
3rd Party Libraries

プログラム言語

C++ / Fortran / Python / R / Matlab etc..

開発環境

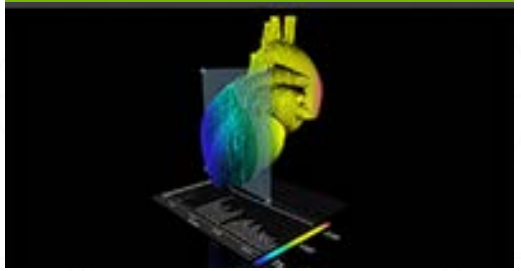
Debugger / Analyzer / IDE etc..

GPUアプリケーションの例

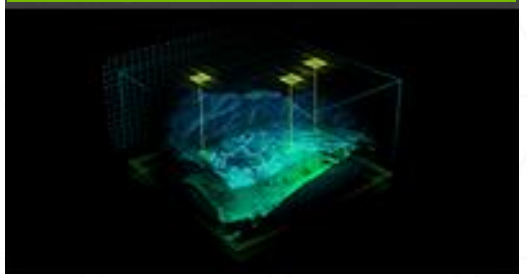
http://www.nvidia.com/object/gpu-applications.html

画像処理　コンピュータビジョン

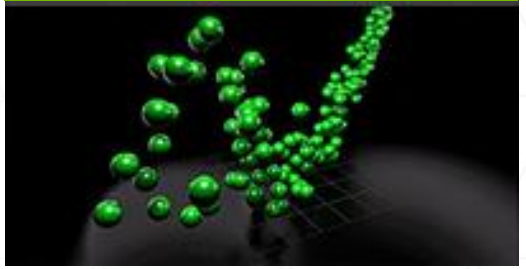
医療画像



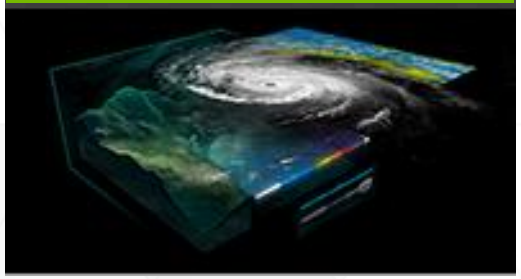
防衛



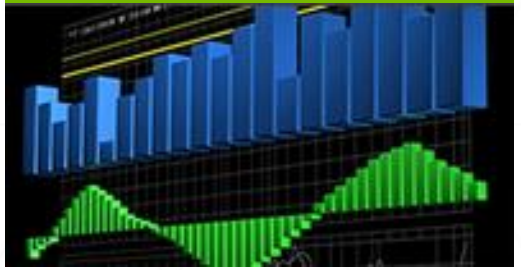
計算化学



気象

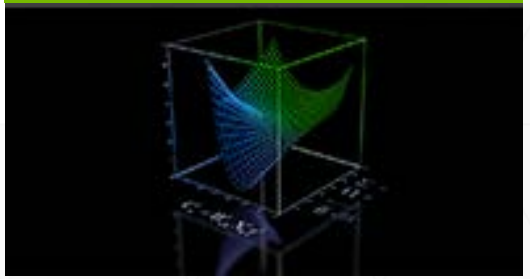


金融工学



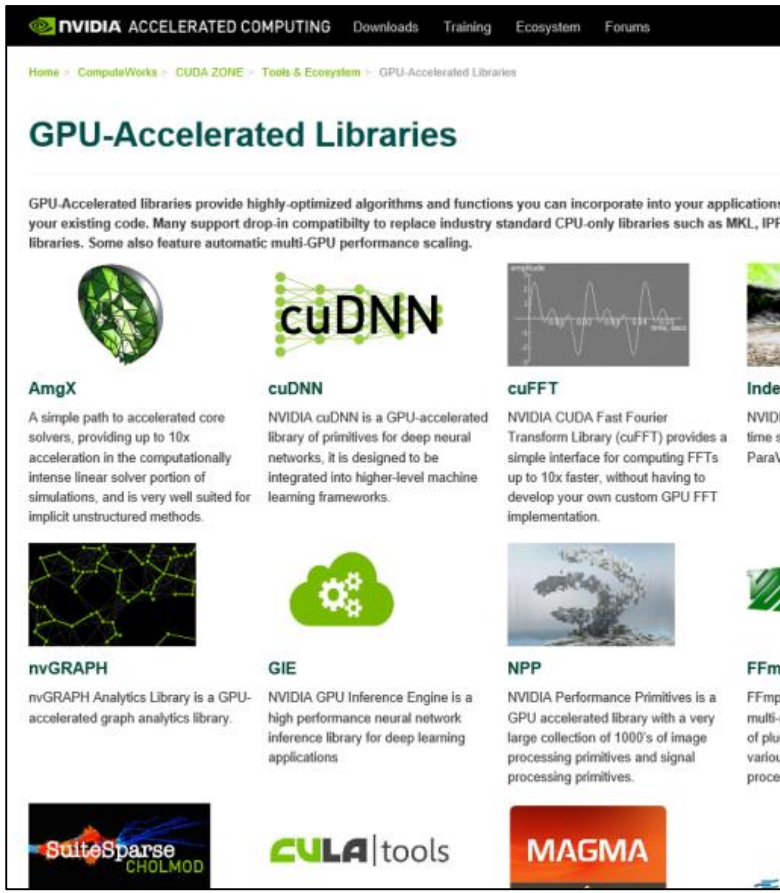
バイオ

数値解析



NVIDIA CUDA ライブラリ

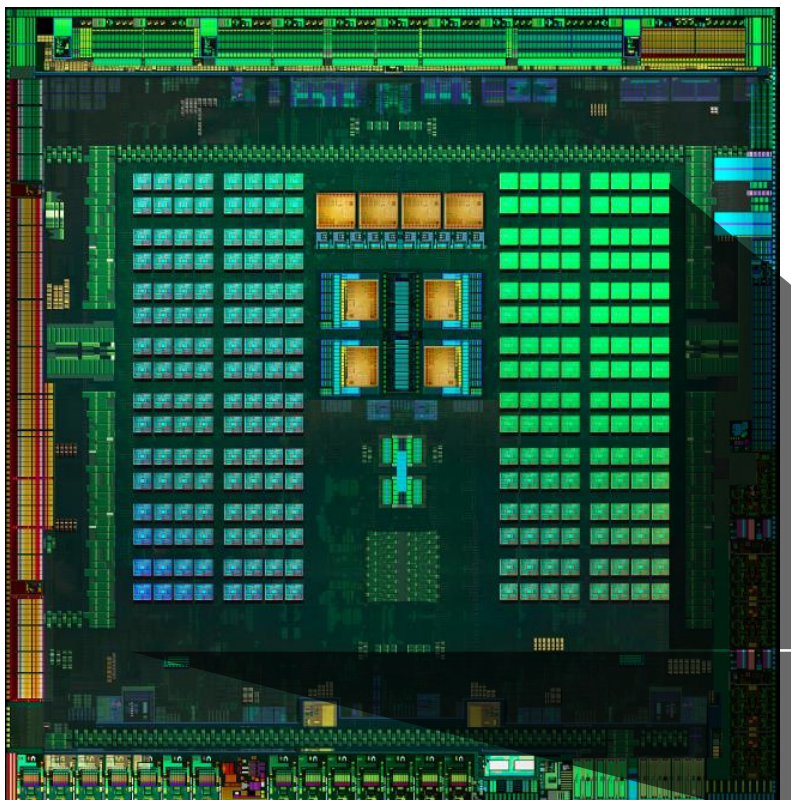
- cuFFT　フーリエ変換
- cuBLAS　行列演算(密行列)
- cuSPARSE　行列演算(疎行列)
- cuSOLVER　行列ソルバ
- cuDNN　ディープラーニング
- cuRAND　乱数生成
- Thrust　C++テンプレート(STLベース)
- NPP　画像処理・信号処理プリミティブ

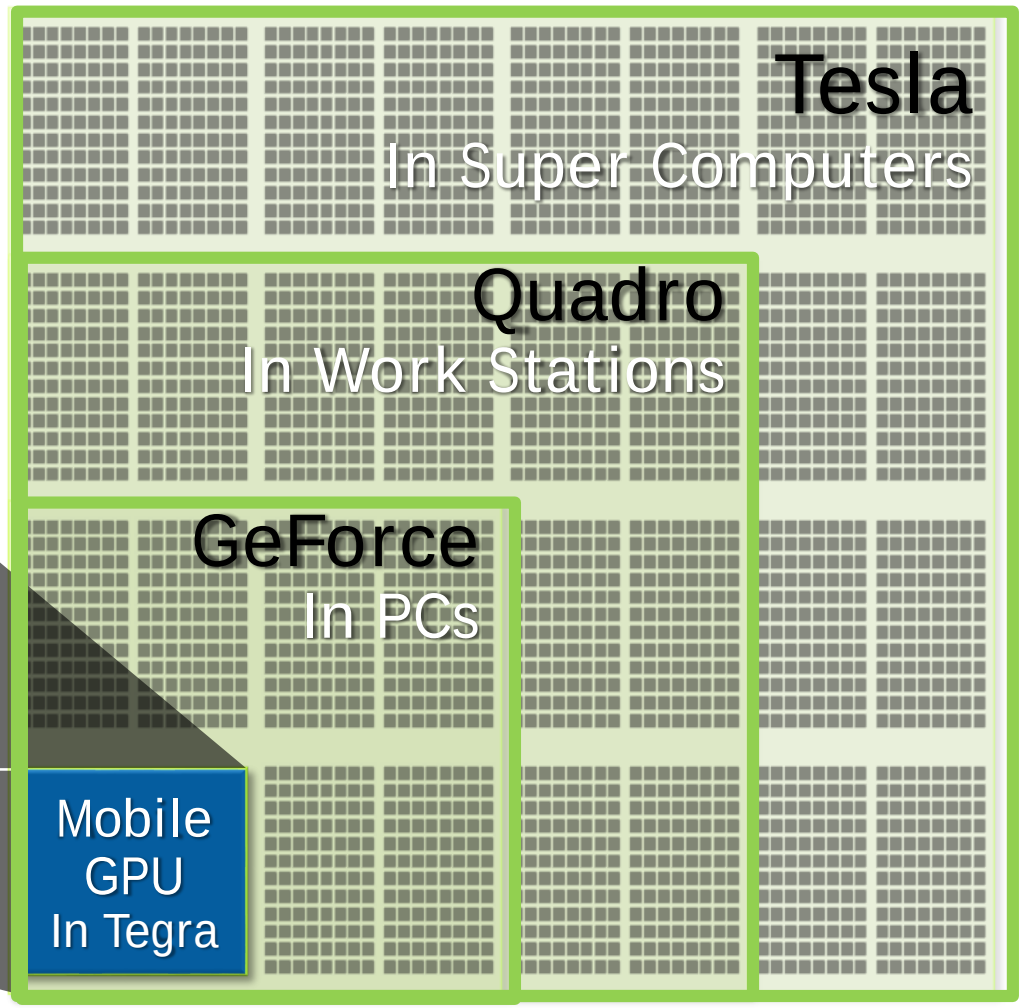


NVIDIA ONE-ARCHITECTURE

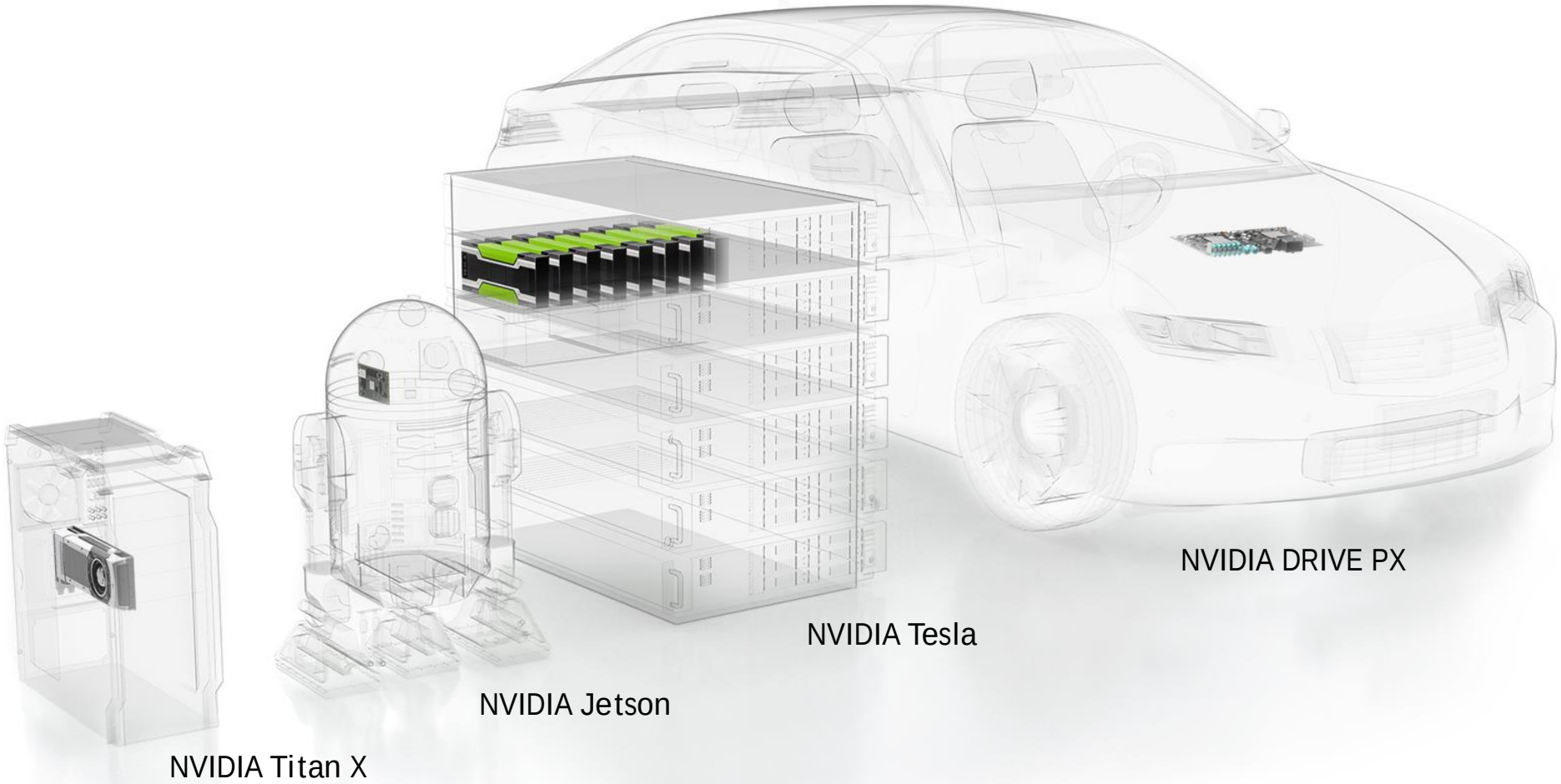
from Super Computer to Automotive SoC

Automotive Tegra



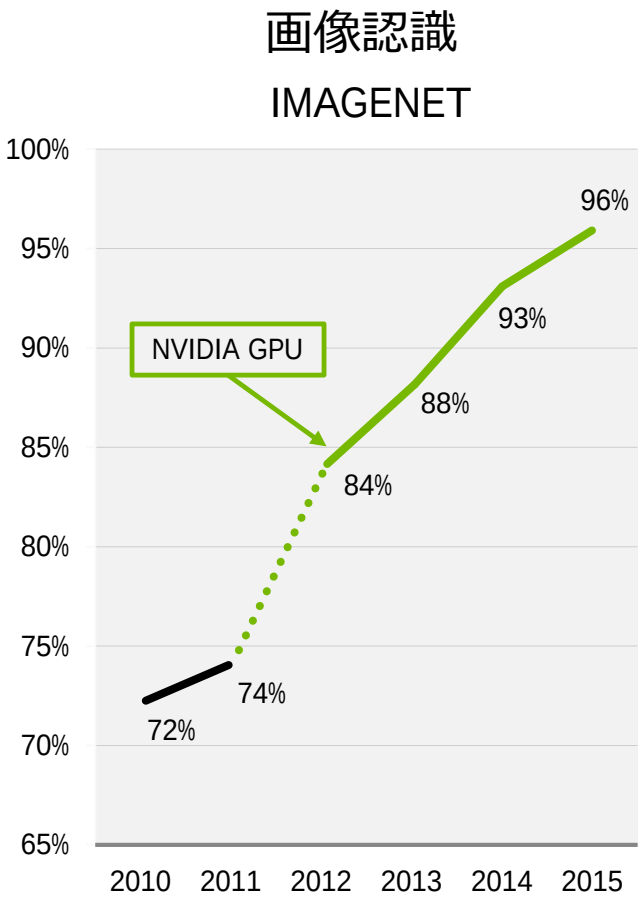


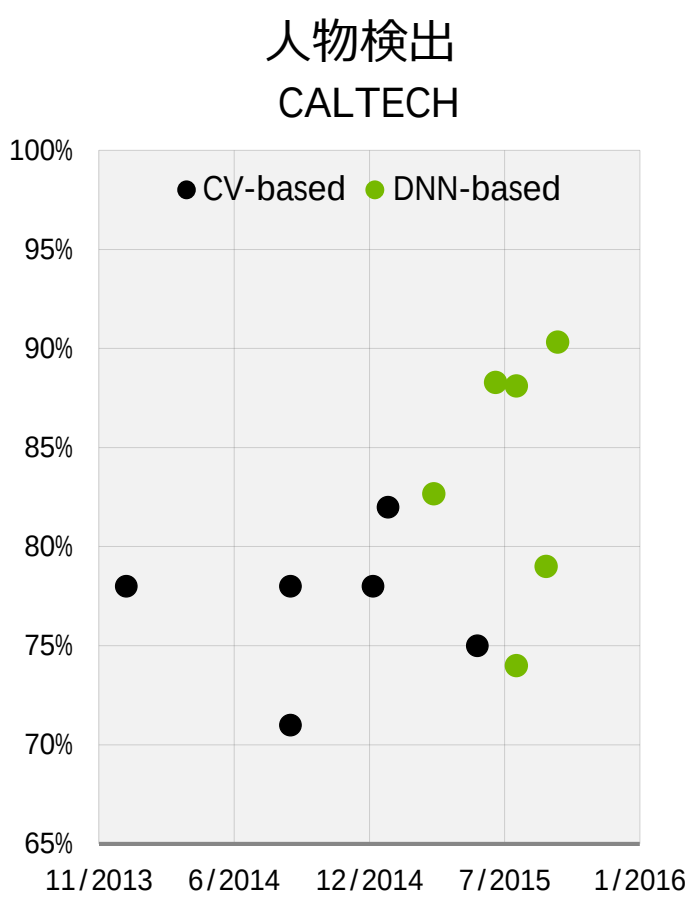
deep learning EVERYWHERE

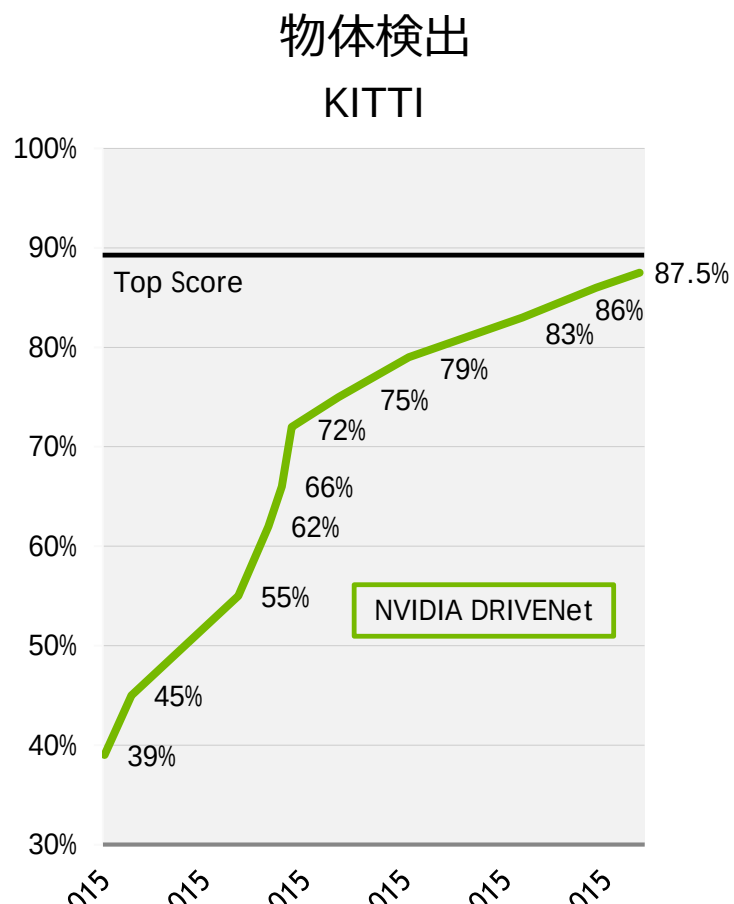


ディープラーニングとは?

驚くべき認識精度の向上



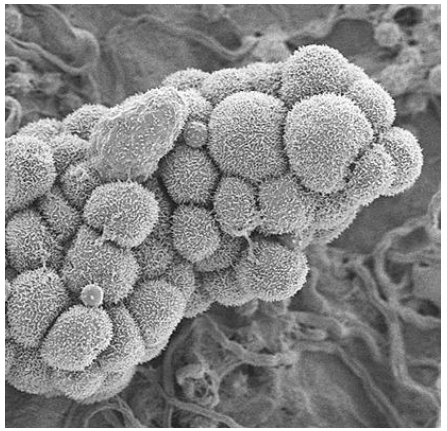




様々な分野でディープラーニングを応用

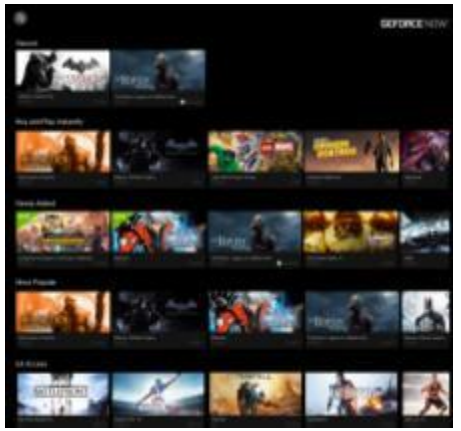
インターネットとクラウド

画像分類
音声認識
言語翻訳
言語処理
感情分析
推薦



医学と生物学

癌細胞の検出
糖尿病のランク付け
創薬

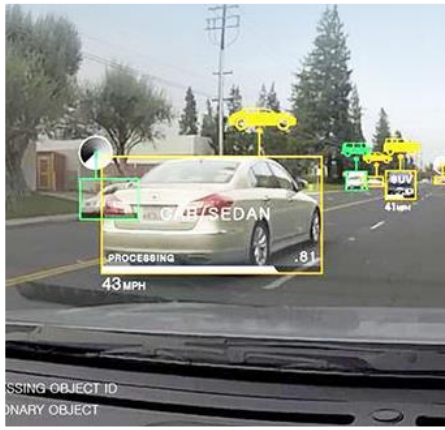


メディアとエンターテイメント

字幕
ビデオ検索
リアルタイム翻訳

セキュリティと防衛

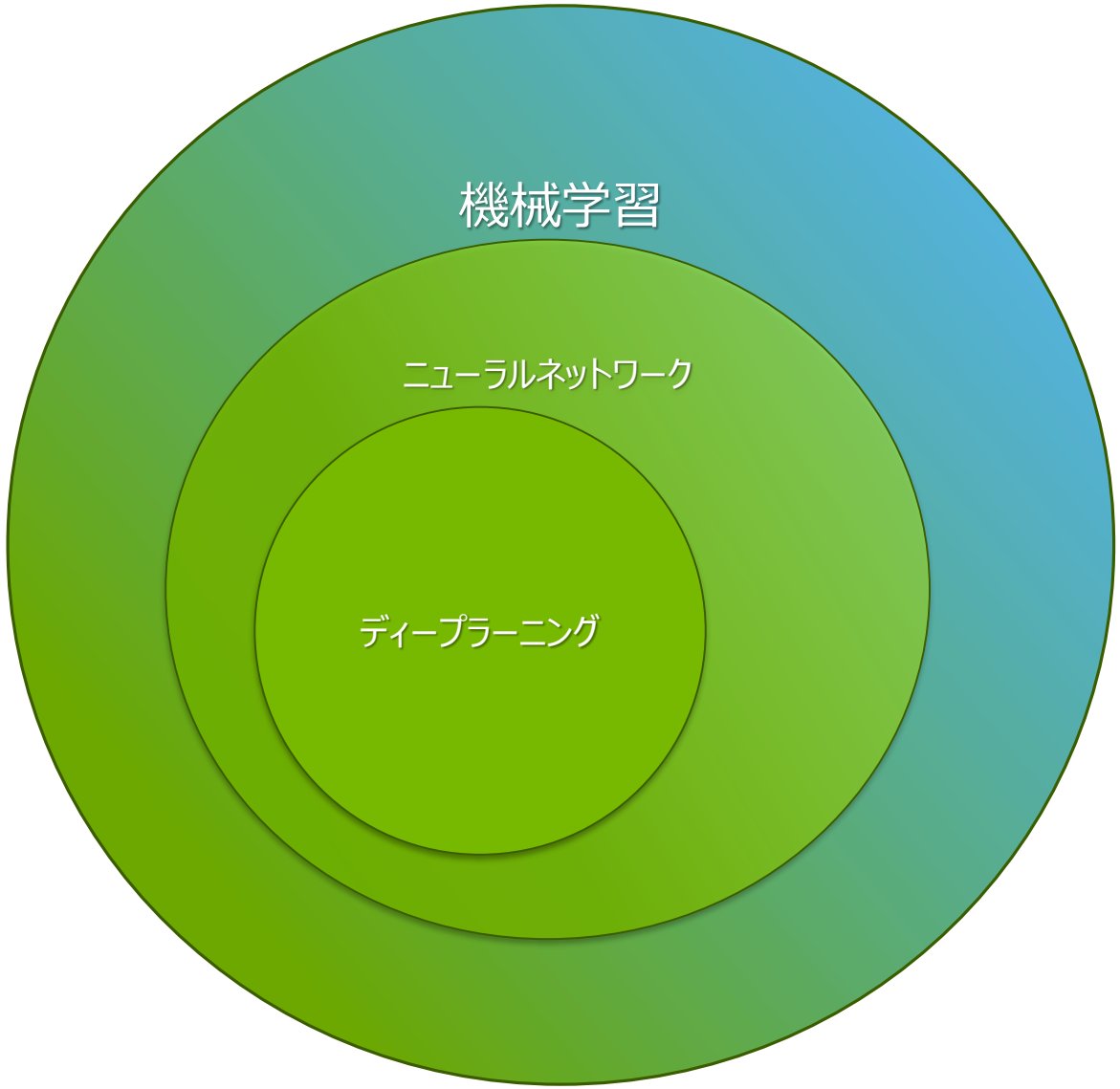
顔検出
ビデオ監視
衛星画像



機械の自動化

歩行者検出
白線のトラッキング
信号機の認識

機械学習とディープラーニングの関係

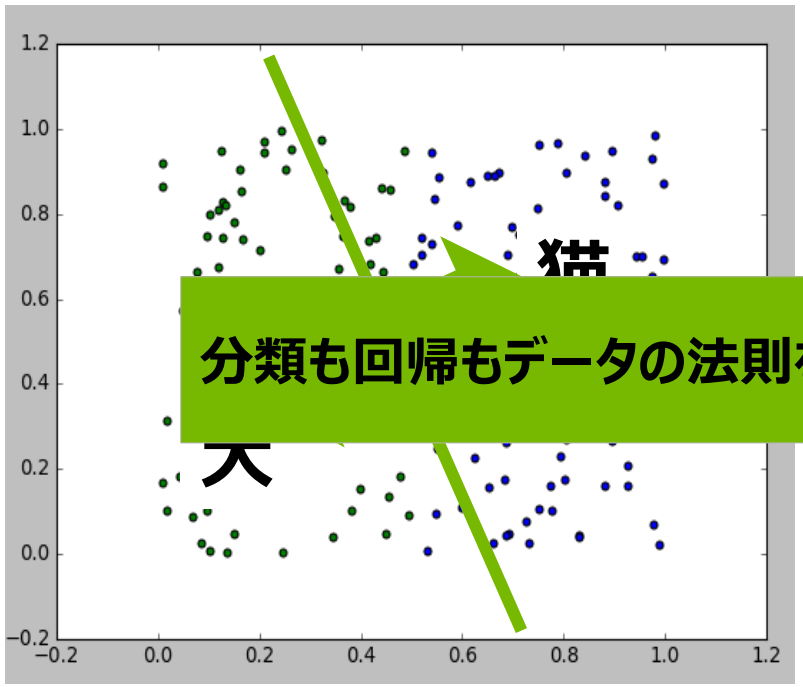


分類(Classification)**と回帰**(Regression)

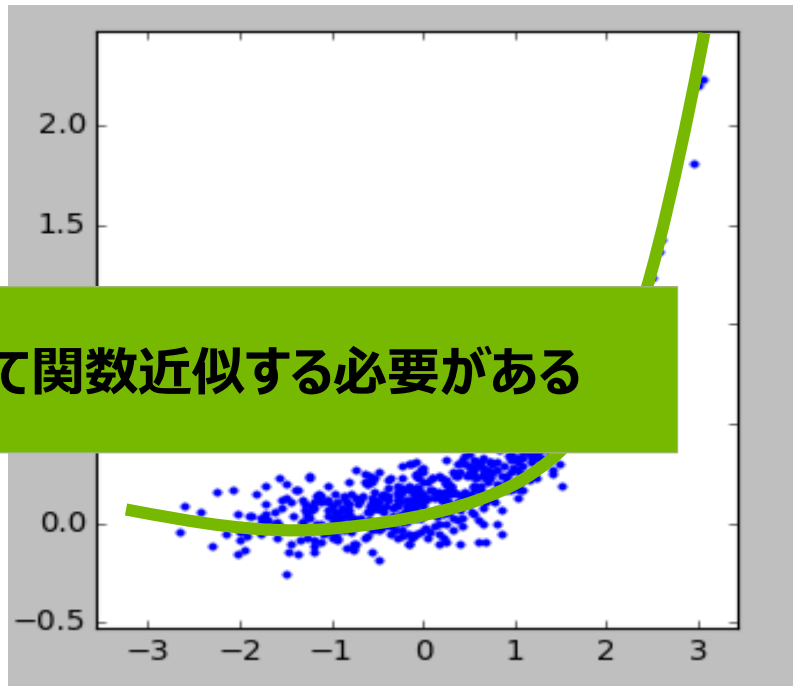
- 分類問題の例
  - 画像に写っているのが猫か犬の画像か推論する
- 回帰問題の例
  - N社の1ヶ月後の株価を予想する

分類(Classification)と回帰(Regression)

分類問題

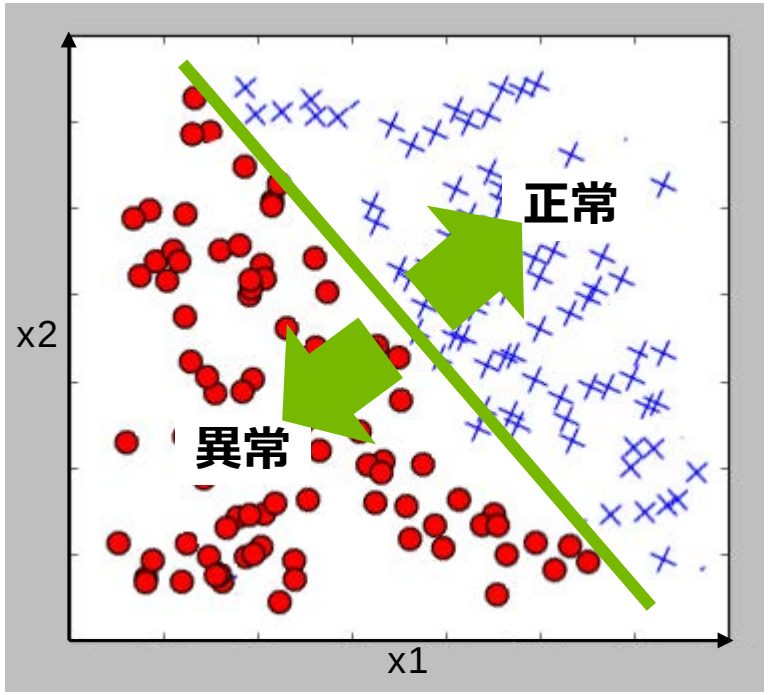


回帰問題



分類も回帰もデータの法則を見つけて関数近似する必要がある

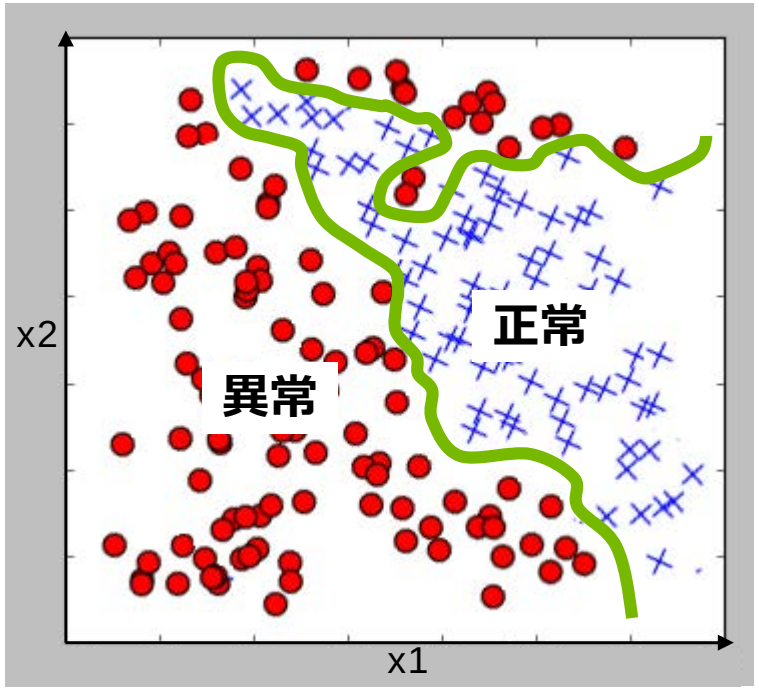
分類問題の例



例) 正常と異常に分類する事を考える
赤:異常
青:正常

$$Y = \alpha X + \beta$$

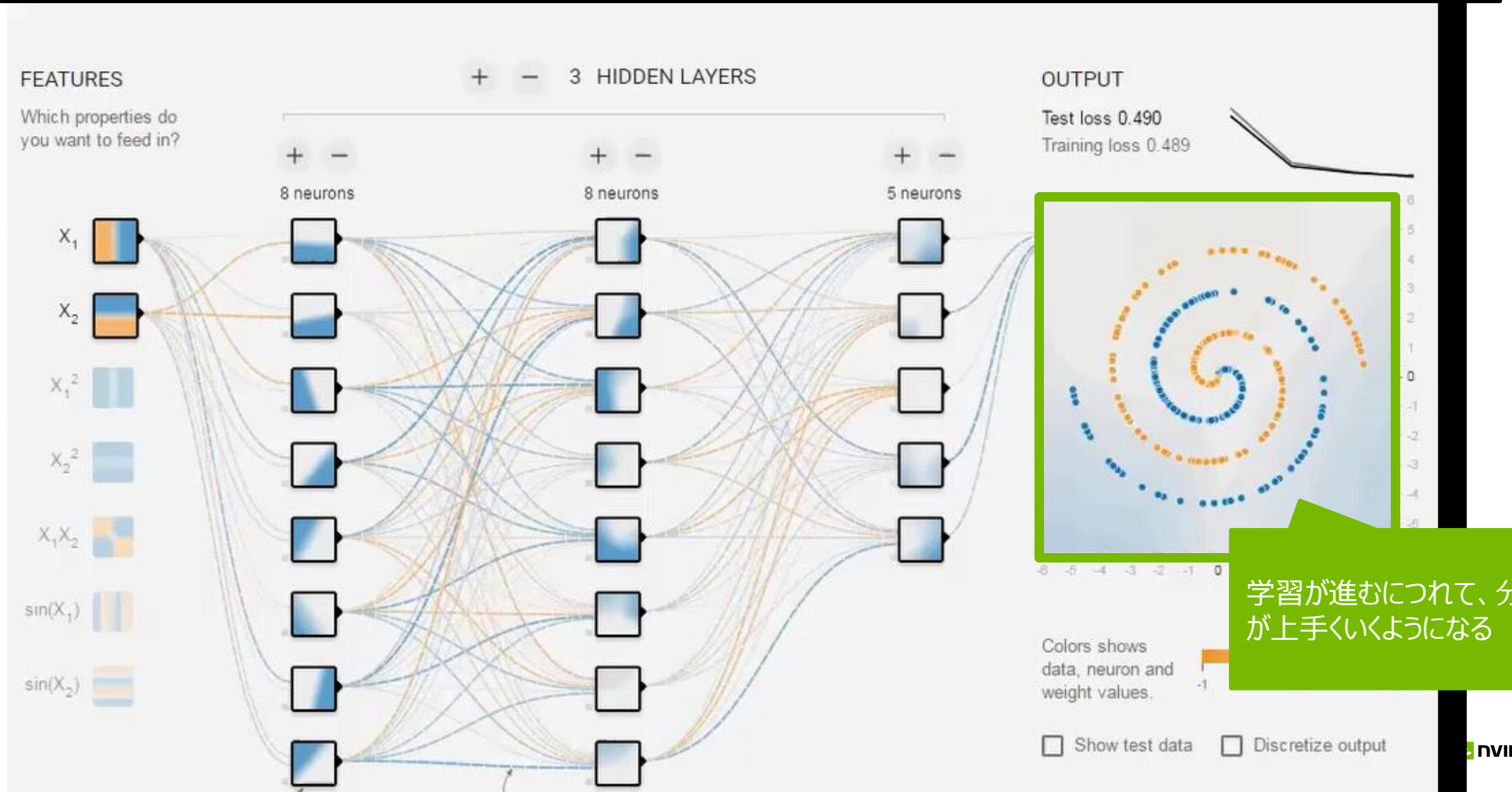
分類問題の例



例) 正常と異常に分類する事を考える
赤:異常
青:正常

複雑な関数近似が必要になる

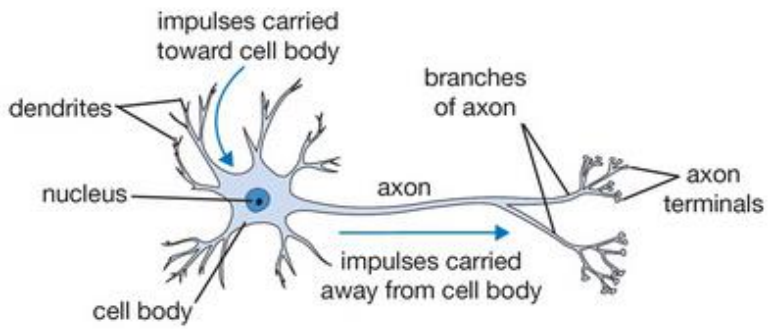
ディープラーニング手法による分類の例



人工ニューロン

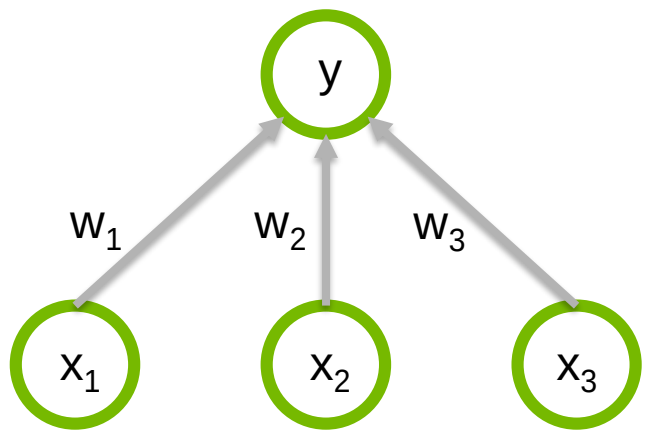
神経回路網をモデル化

神経回路網



スタンフォード大学cs231講義ノートより

人工ニューロン

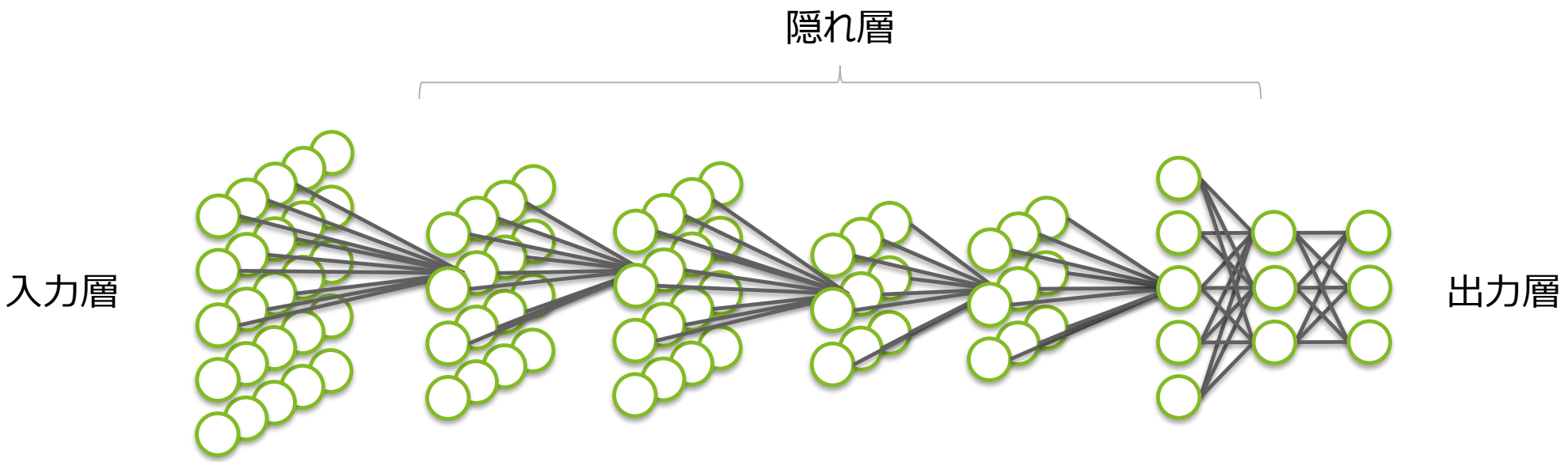


$$y=F(w_1x_1+w_2x_2+w_3x_3)$$

$$F(x)=\max(0,x)$$

人工ニューラルネットワーク

単純で訓練可能な数学ユニットの集合体
ニューラルネットワーク全体で複雑な機能を学習



十分なトレーニングデータを与えられた人工ニューラルネットワークは、入力データから判断を行う
複雑な近似を行う事が出来る。

ディープラーニングの恩恵

ディープラーニングとニューラルネットワーク

- ロバスト性
  - 特徴量の設計を行う必要がない。 - 特徴は自動的に獲得される学習用データのバラつきの影響を押さえ込みながら、自動的に学習していく
- 一般性
  - 同じニューラルネットワークのアプローチを多くの異なるアプリケーションやデータに適用する事が出来る
- スケーラブル
  - より多くのデータで大規模並列化を行う事でパフォーマンスが向上する

ディープラーニングのアプローチ

学習(トレーニング):

エラー

モデル

犬
猫
アライグマ

推論(インファレンス):

モデル

犬

畳込みニューラルネットワーク(CNN)

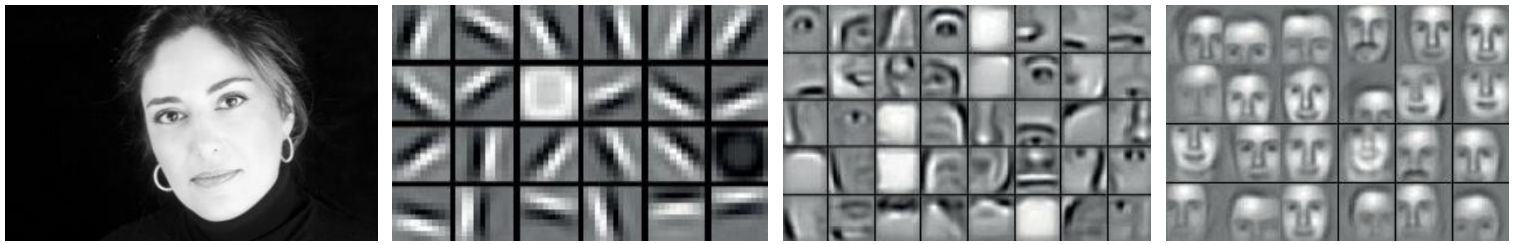
多量なトレーニングデータと多数の行列演算

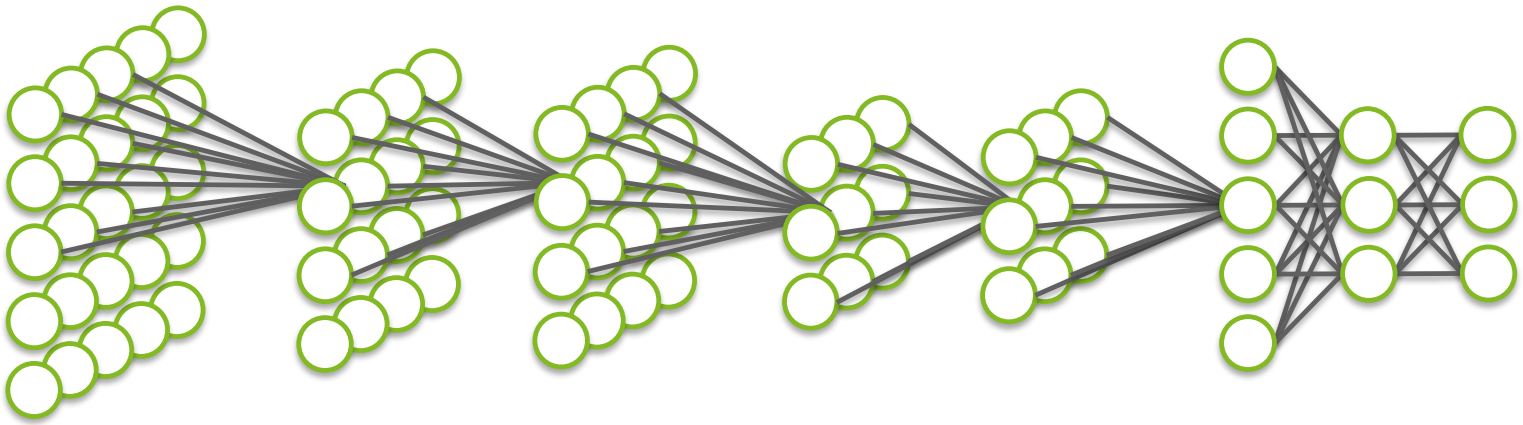
目的
顔認識

トレーニングデータ
1,000万～1億イメージ

ネットワークアーキテクチャ
10 層
10 億パラメータ

ラーニングアルゴリズム
30 エクサフロップスの計算量
GPU を利用して30日

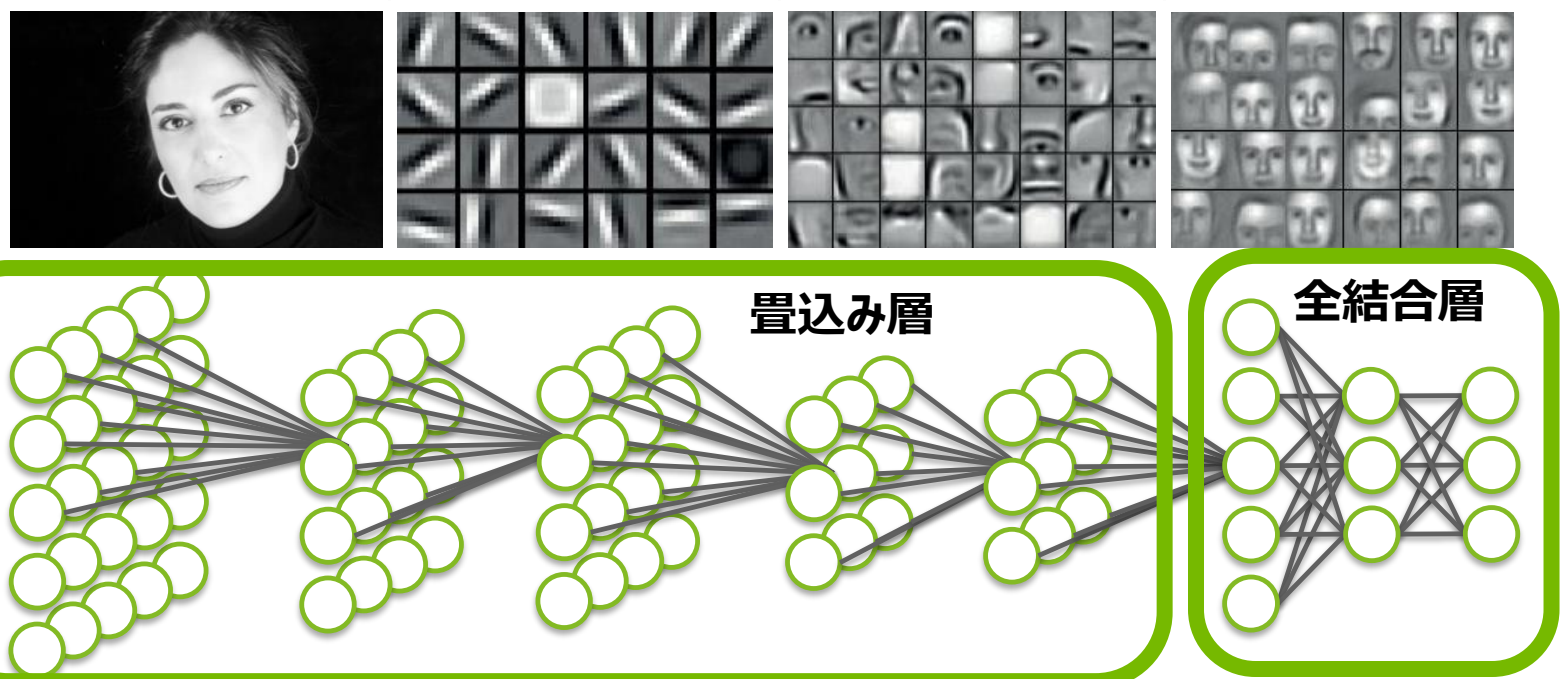




畳込みニューラルネットワーク(CNN)

- 画像認識・画像分類で使われる、高い認識精度を誇るアルゴリズム。畳込み層で画像の特徴を学習

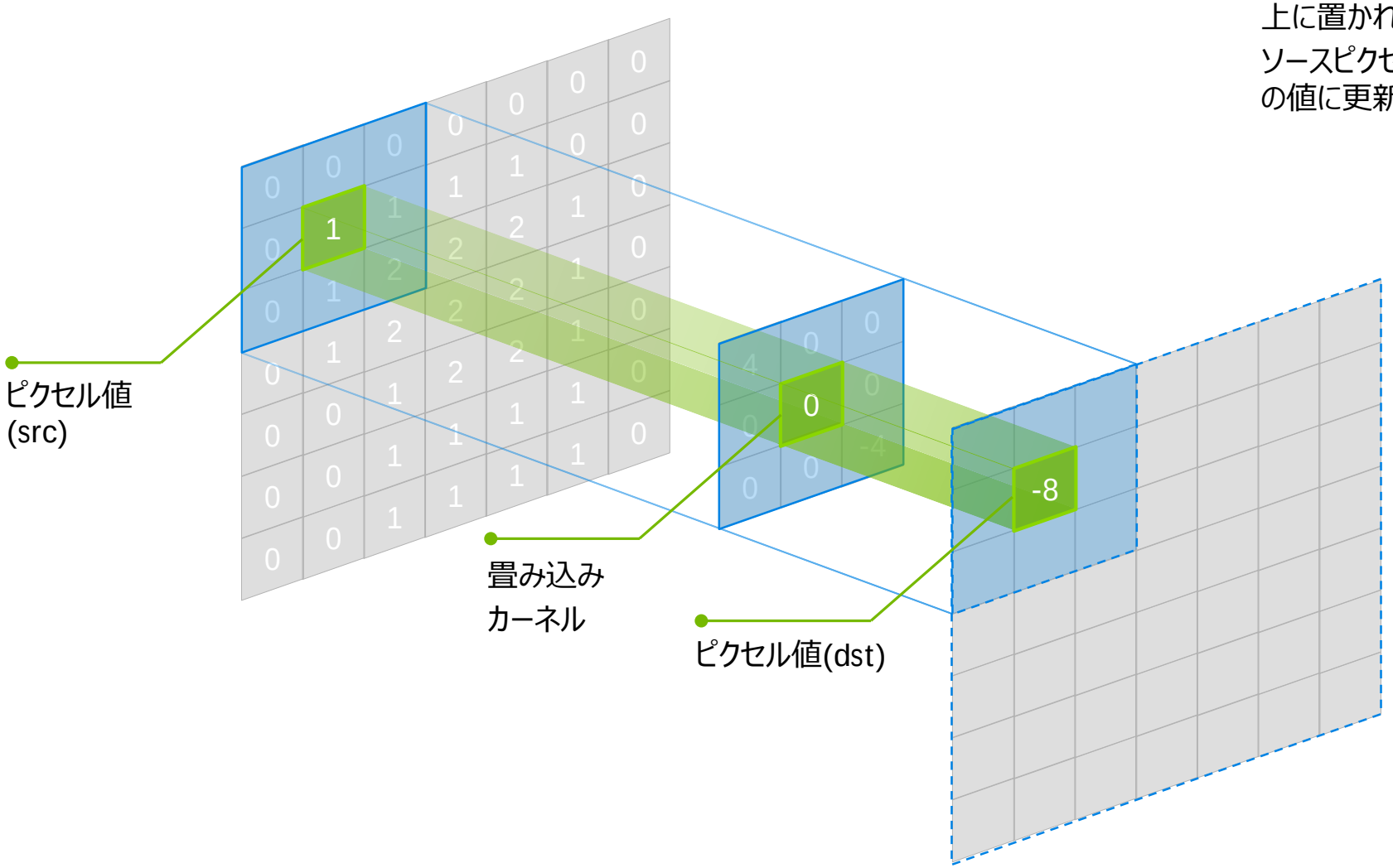
目的	トレーニングデータ	ネットワークアーキテクチャ	ラーニングアルゴリズム
顔認識	1,000万～1億イメージ	10 層 10 億パラメータ	30 エクサフロップスの計算量 GPU を利用して30日



畳込み層

カーネルの中心の値はソースピクセル上に置かれている。

ソースピクセルはフィルタを自身の積の値に更新される。



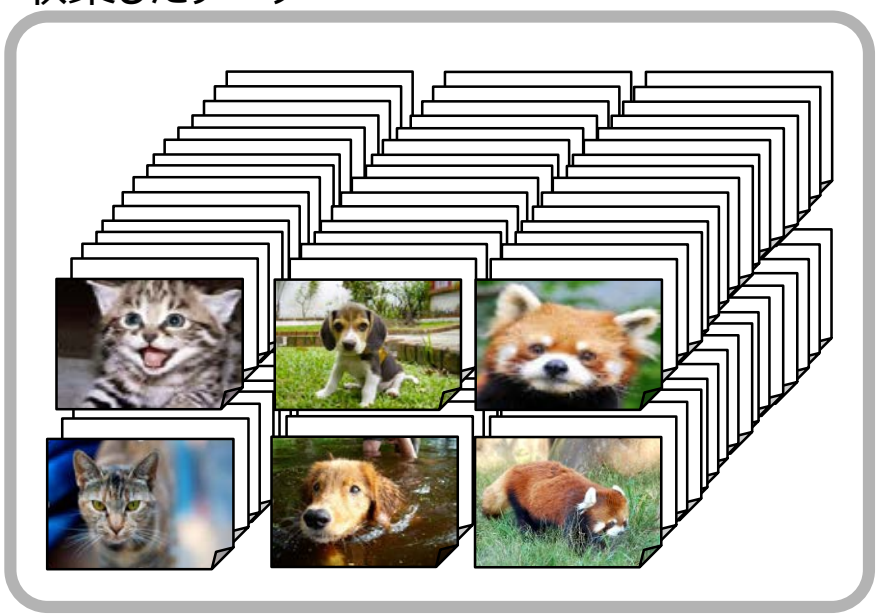
ディープラーニングの学習の流れ

訓練データと検証データ

ディープラーニングの学習

- データを**訓練データ**(training data)と**検証データ**(validation data)に分割する

収集したデータ



訓練データ(training data)

重みの更新
(学習)に使用

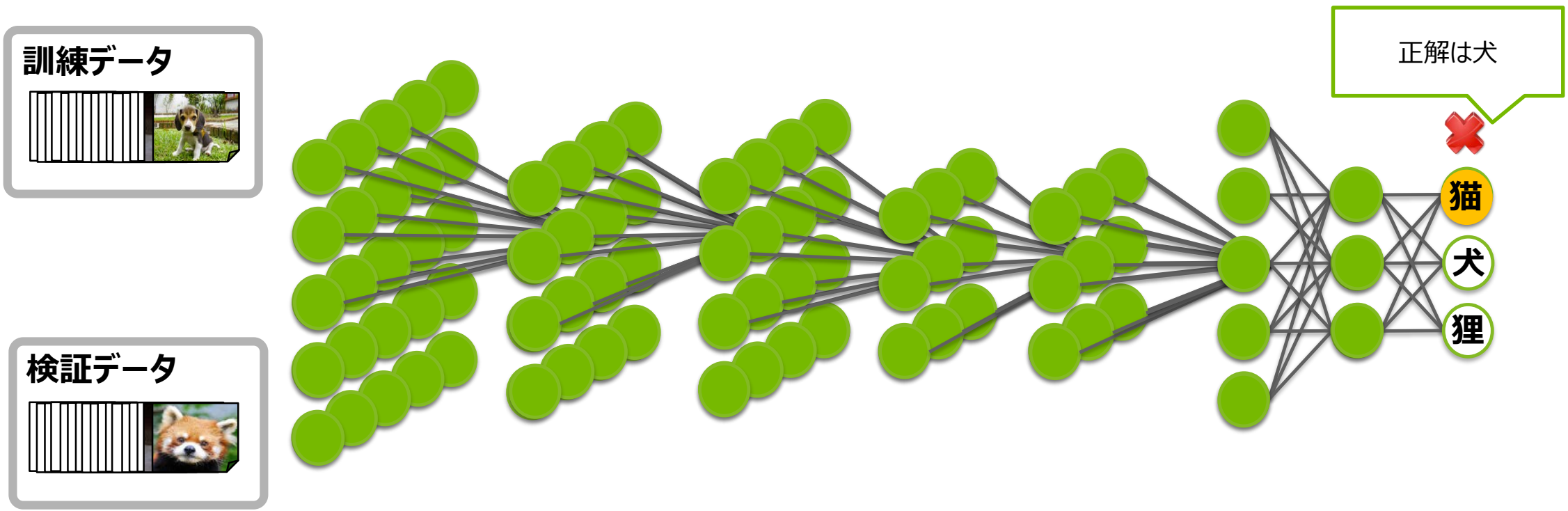
検証データ(validation data)

汎化性能の
確認に使用

訓練データによる重みの更新

ディープラーニングの学習

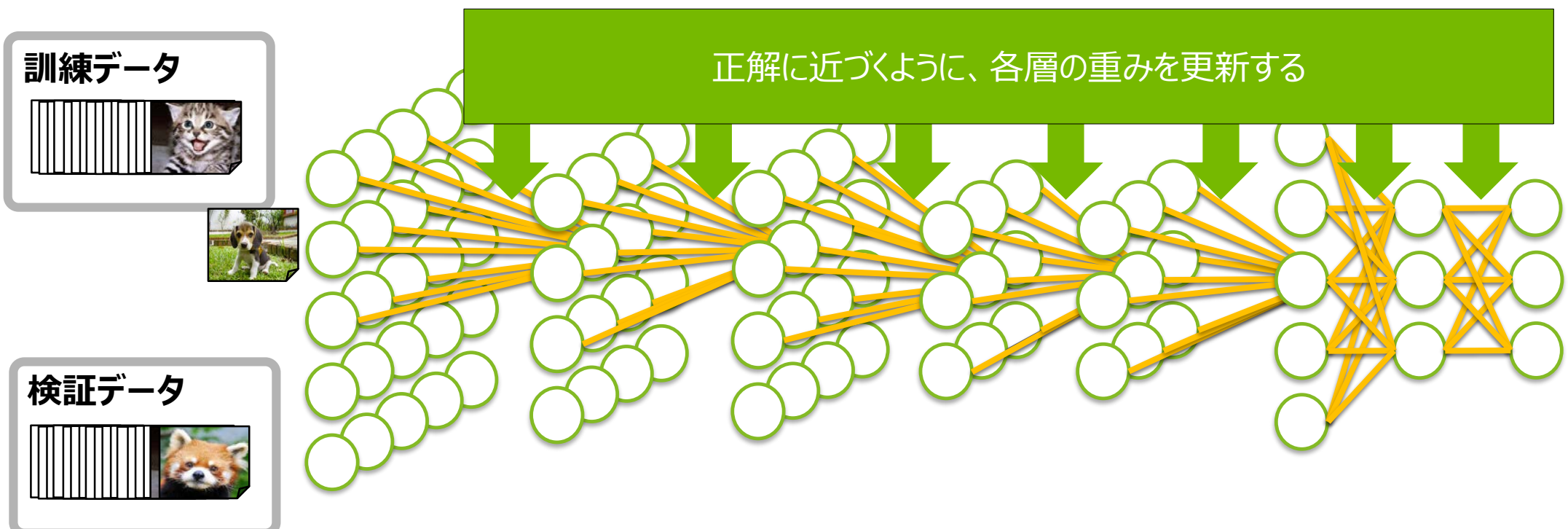
訓練データをニューラルネットワークに与え、正解ラベルと出力結果の誤差が無くなるように重みWの更新を繰り返す



訓練データによる重みの更新

ディープラーニングの学習

訓練データをニューラルネットワークに与え、正解ラベルと出力結果の誤差が無くなるように重みWの更新を繰り返す



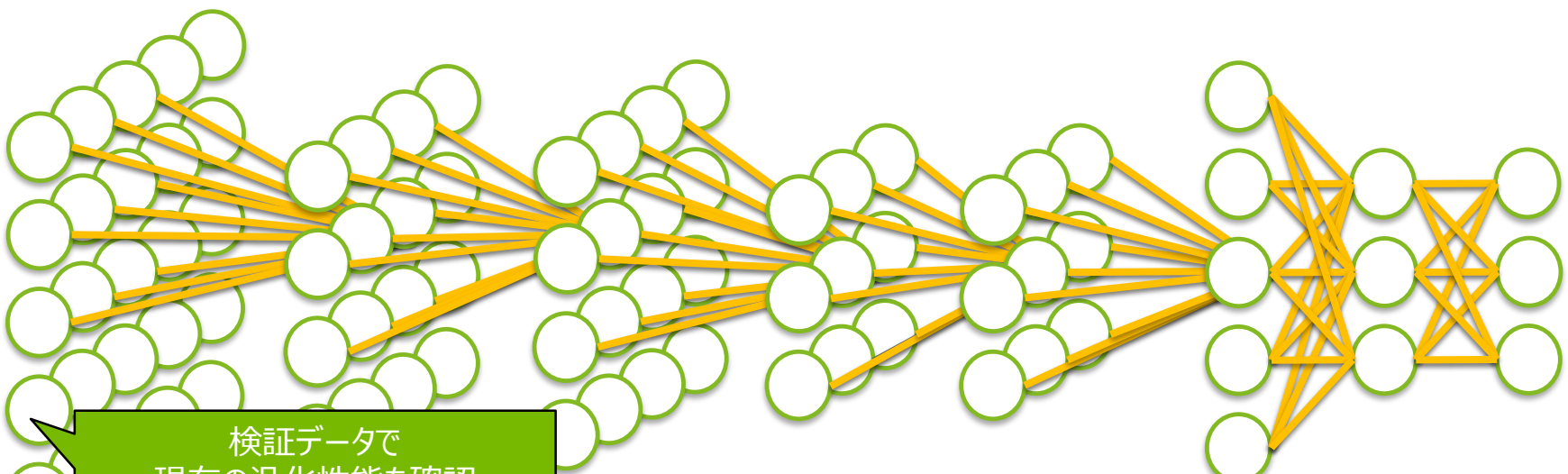
学習ループ

訓練データと検証データの役割

すべての訓練データを用いて重み更新を行う + すべての検証データで汎化性能を確認
⇒ 1 **エポック**(epoch)**と呼ぶ**

訓練データ

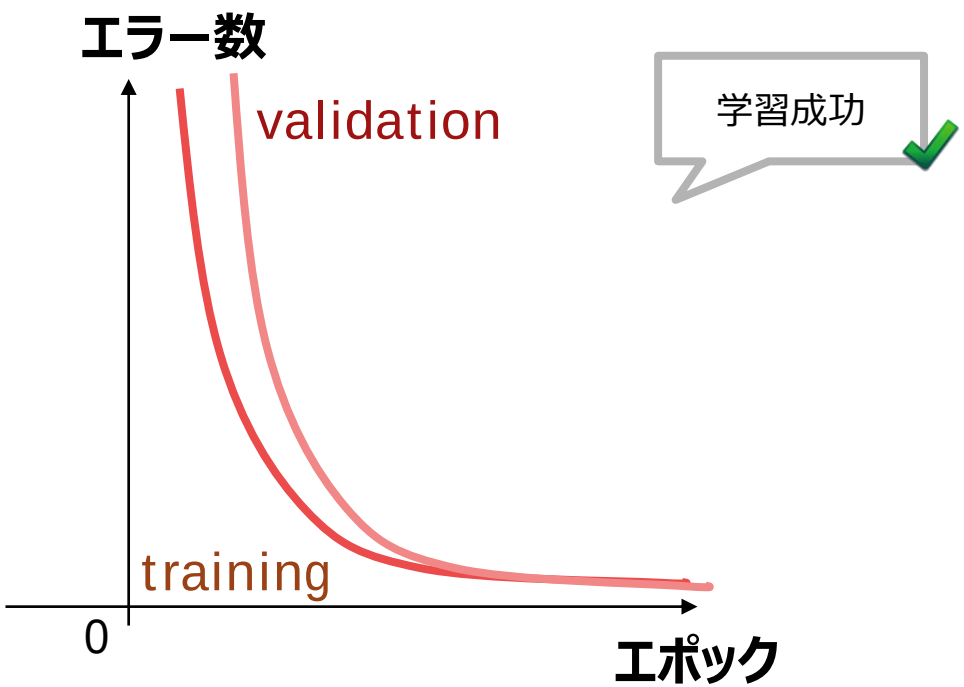
検証データ

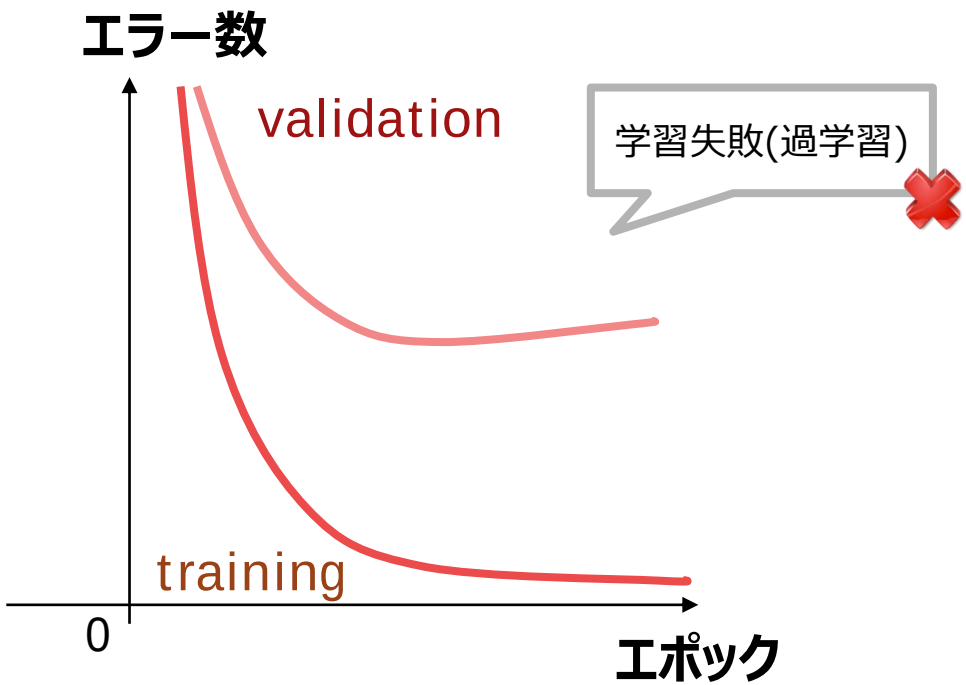


学習時の性能の確認

訓練データと検証データの役割

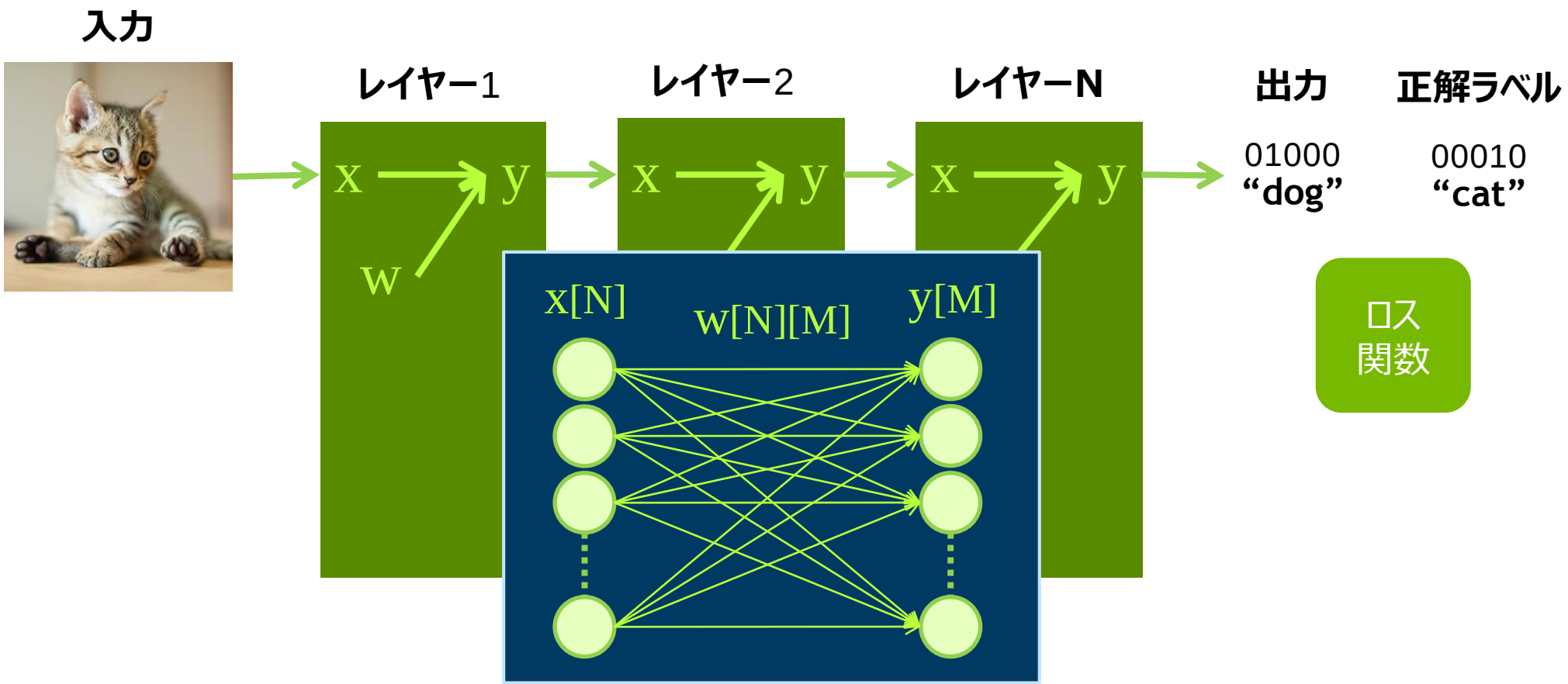
各エポックで訓練データをニューラルネットワークに与えた際の間違い率と検証データを与えた際の間違い率を確認しながら学習を進める必要がある





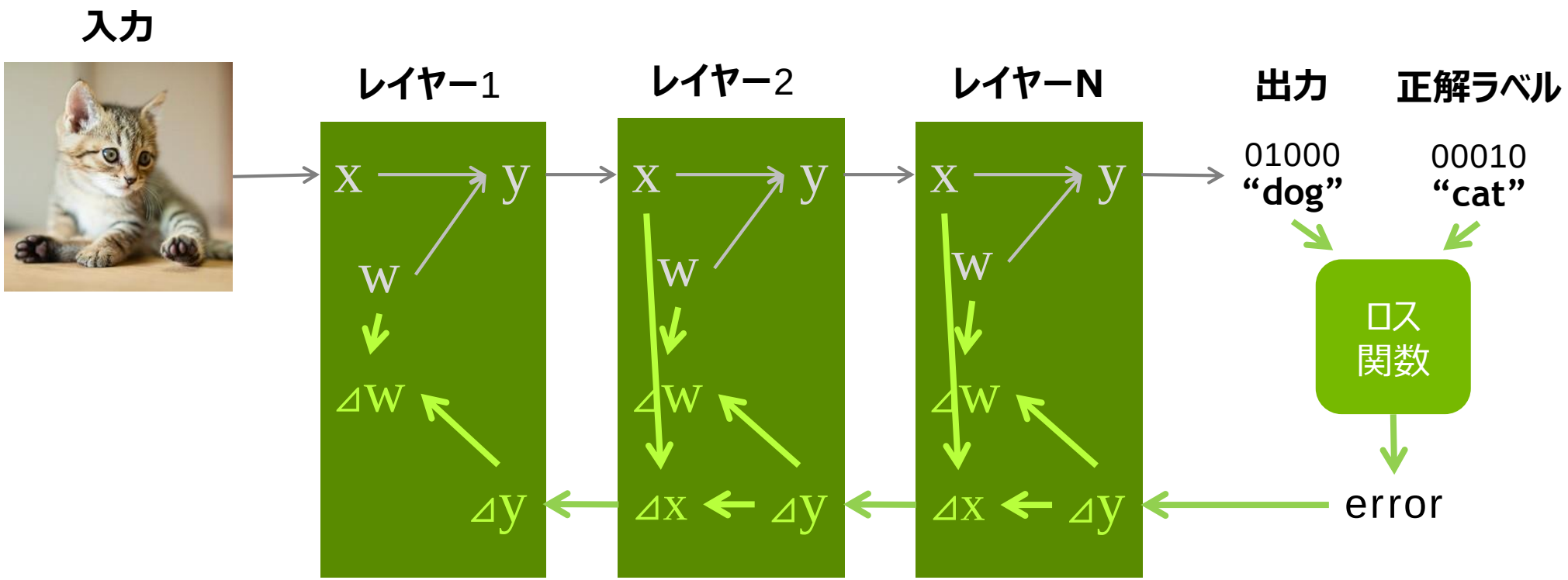
順伝播(フォワードプロパゲーション)

誤差逆伝播法



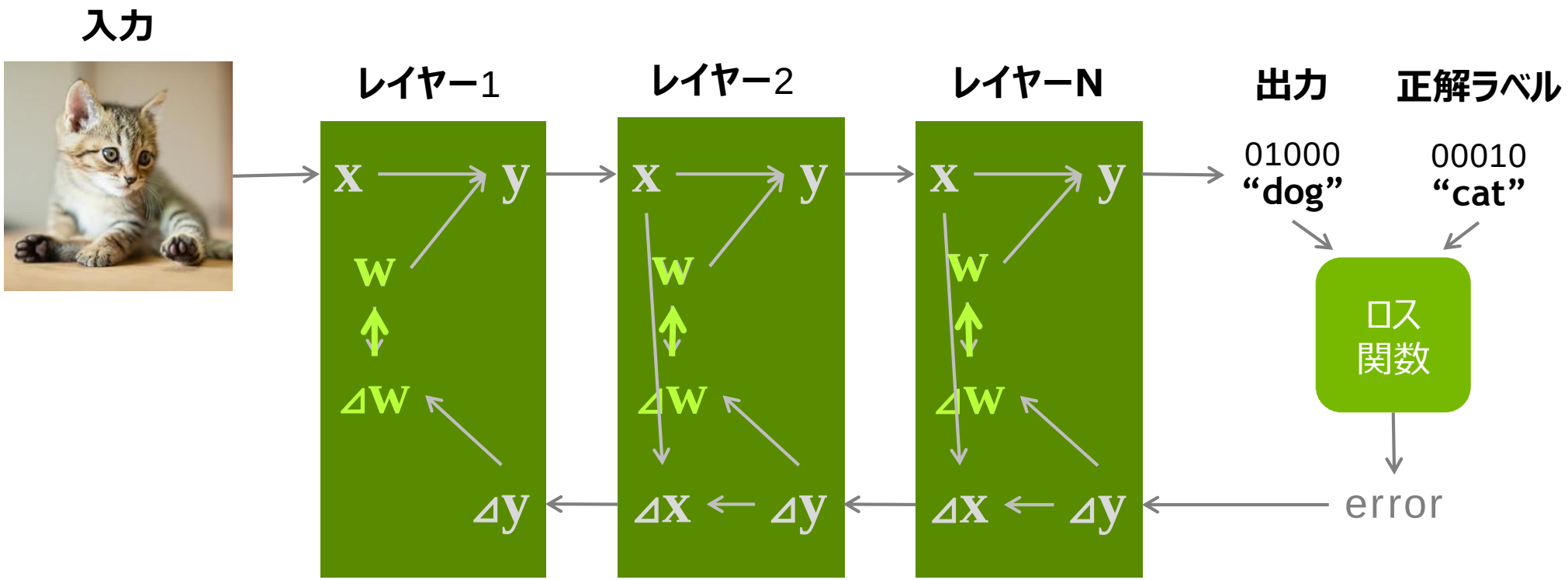
逆伝播(バックプロパゲーション)

誤差逆伝播法



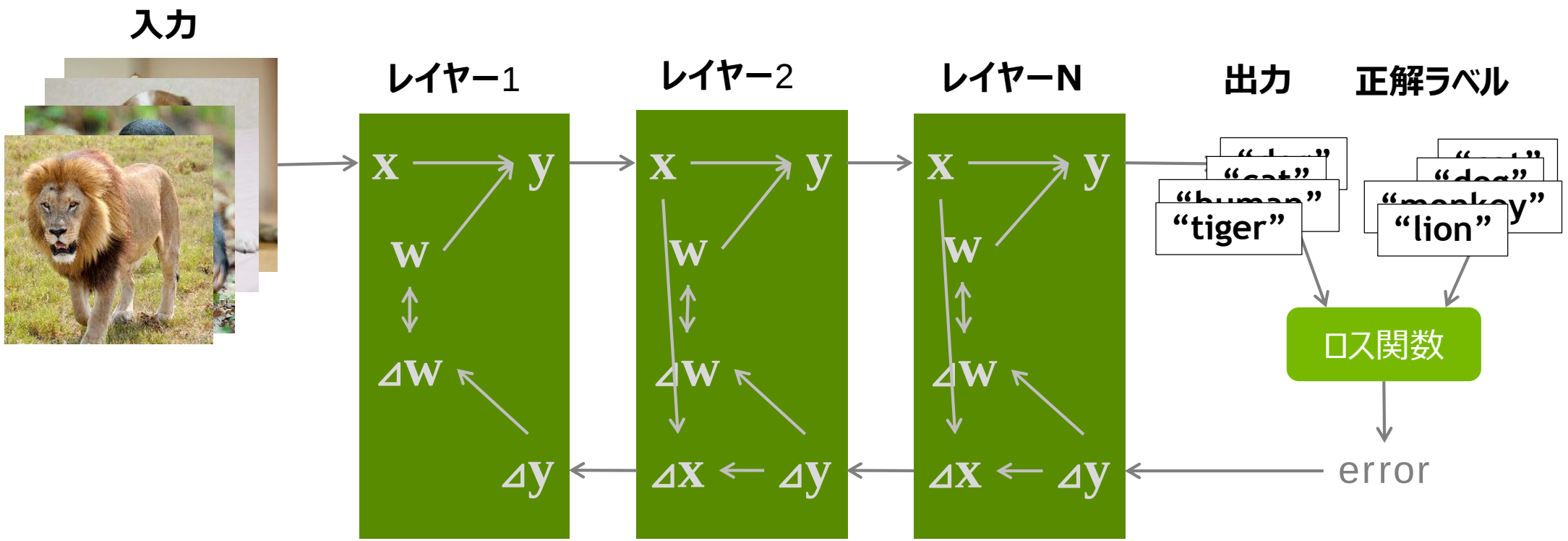
重みの更新

誤差逆伝播法



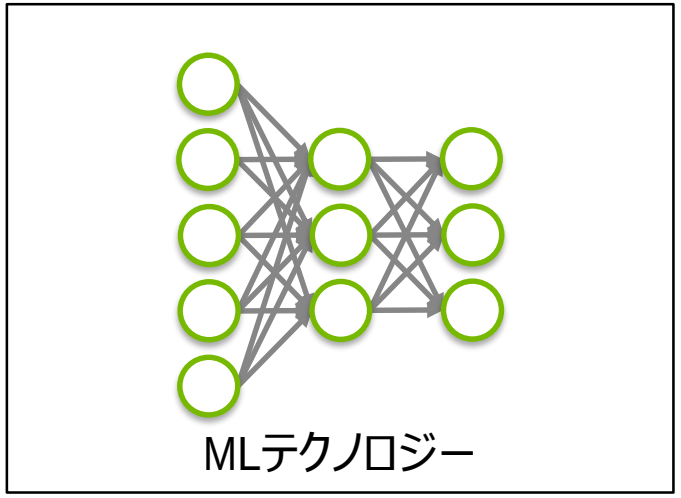
ミニバッチ

誤差逆伝播法



GPUディープラーニング

ディープラーニングを加速する3つの要因

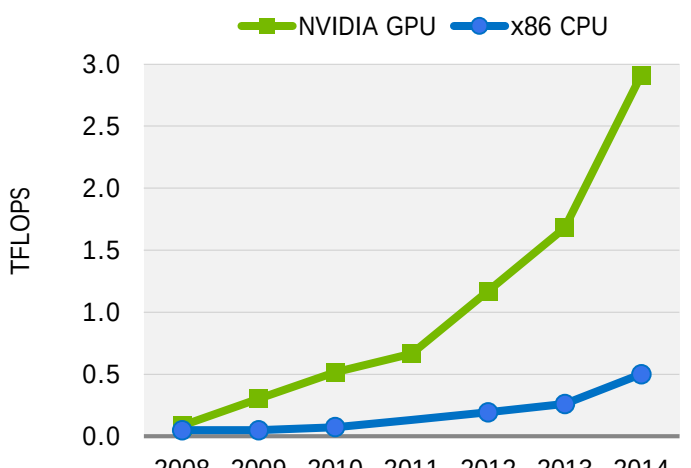


TORCH	CAFFE
NYU facebook	Berkeley UNIVERSITY OF CALIFORNIA
THEANO	MATCONVNET
Université de Montréal	UNIVERSITY OF OXFORD
MOCHA.JL	PURINE
Massachusetts Institute of Technology	NUS National University of Singapore
MINERVA	MXNET*
NYU	UNIVERSITY of WASHINGTON Carnegie Mellon University

facebook 350 million images uploaded per day

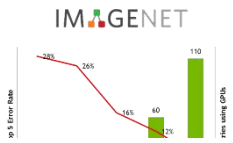
Walmart 2.5 trillion transactions per hour

YouTube 100 hour video uploaded per minute



ディープラーニング SDK

ディープラーニングを加速するディープラーニングライブラリ



画像分類

物体検出

コンピュータビジョン

音声認識

言語翻訳

ボイス&オーディオ

推薦エンジン

感情分析

自然言語処理

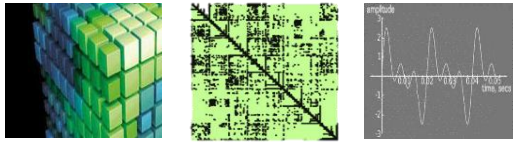
Caffe
Chainer
DL4J
Deeplearning4j
Mocha.jl
julia
K
KERAS
Microsoft
CNTK
MatConvNet
mxnet
MINERVA
Purine
OpenDeep
Pylearn2
TensorFlow
torch
theano

ディープラーニングフレームワーク

cuDNN

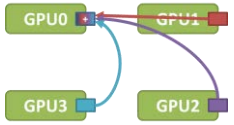
ディープラーニング

cuBLAS cuSPARSE cuFFT



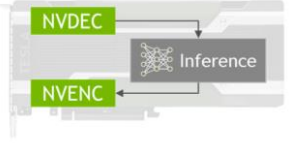
CUDA 数学ライブラリ

NCCL



マルチGPU間通信

DeepStream SDK



ビデオ解析

TensorRT

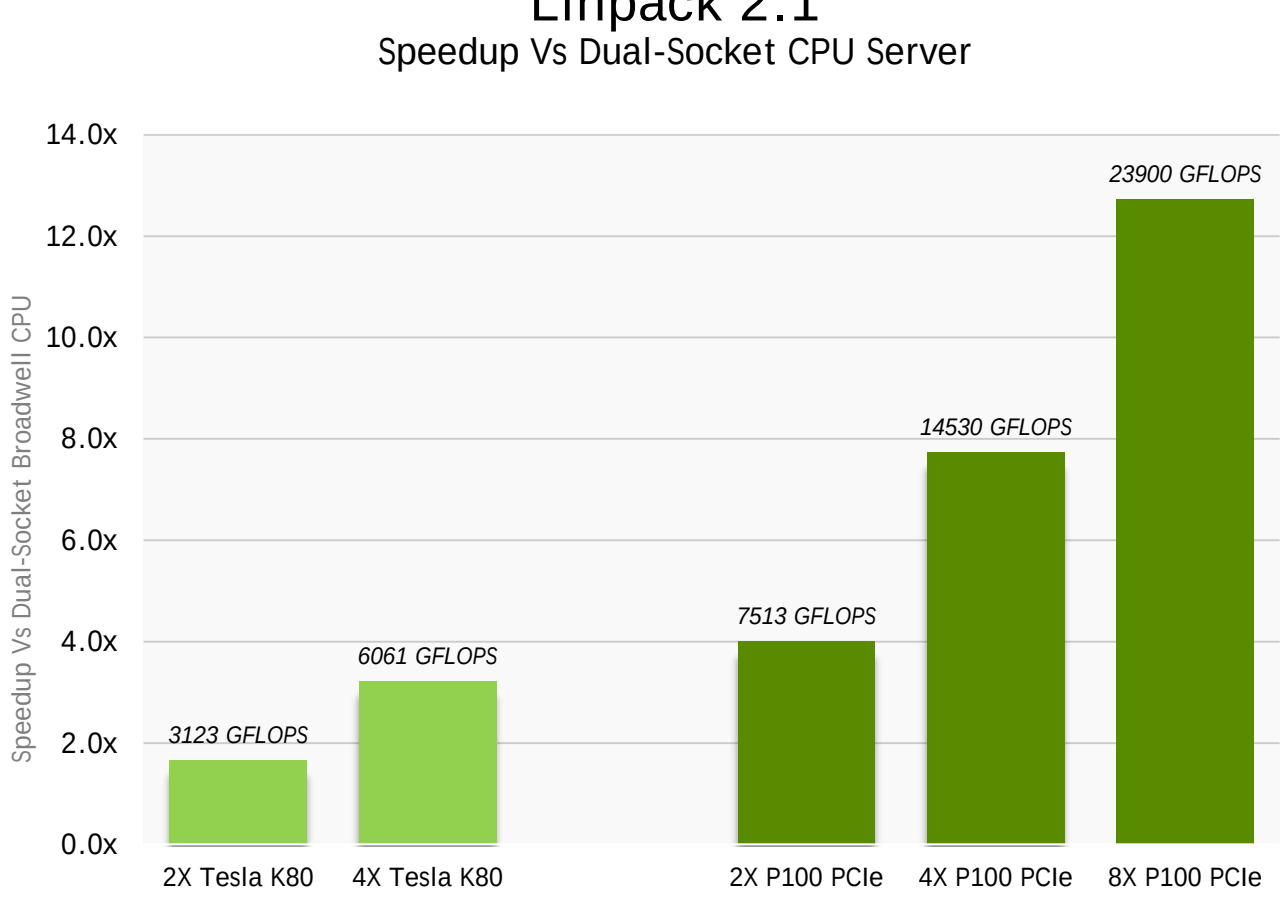
インファレンス

LINPACK

Benchmark
Measures floating point computing power

Accelerated Features	Metric	Scalability
All	GFLOPS	Multi-GPU, Multi Node

https://www.top500.org/project/**linpack**/



CPU Server: Dual Xeon E5-2699 v4@2.2GHz (44-cores)
GPU Servers: Dual Xeon E5-2699 v4@2.2GHz (44-cores) with Tesla K80s or P100s PCIe
CUDA Version: CUDA 8.0.44
Dataset: HPL.dat

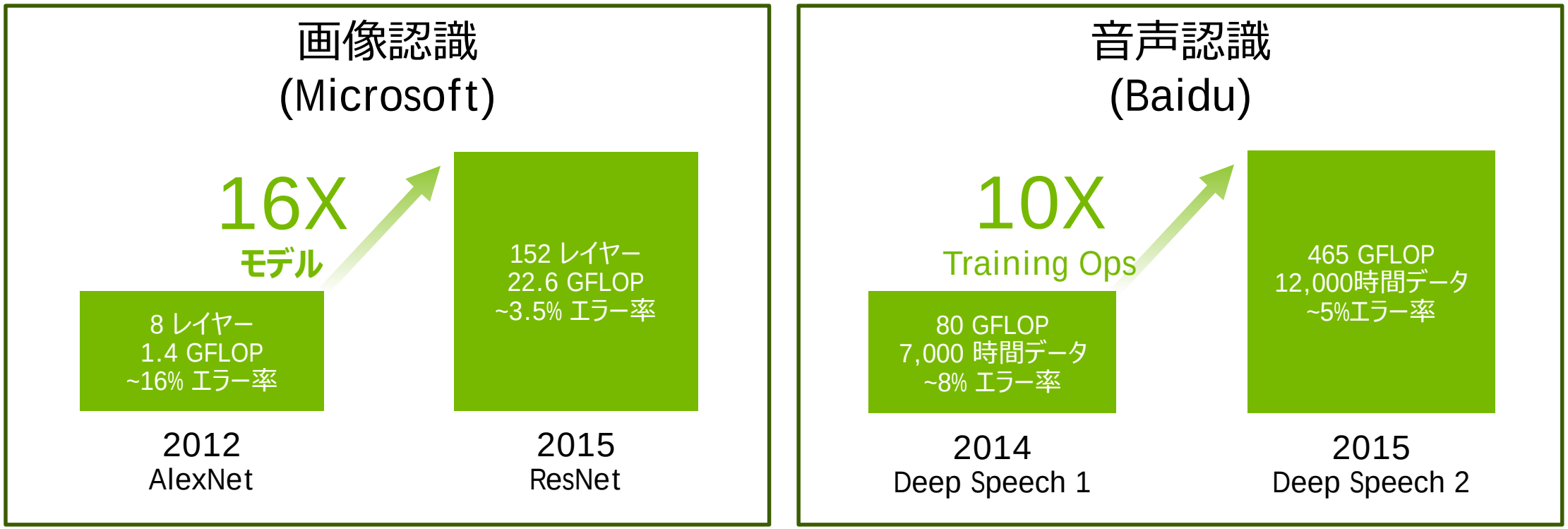
P100 SXM2による高速な学習

FP32 Training, 8 GPU Per Node

Refer to "Measurement Config." slide for HW & SW details
AlexnetOWT/GoogLenet use total batch size=1024, VGG-D uses total batch size=512, Incep-v3/ResNet-50 use total batch size=256

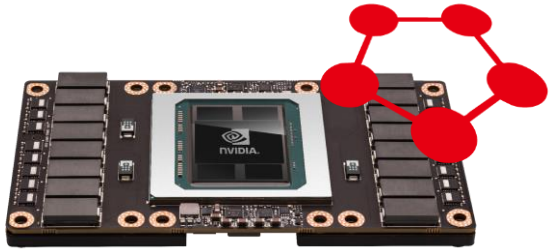
認識精度向上のため
モデルはよりディープに、データはより大きく

強力な計算パワーが必要に

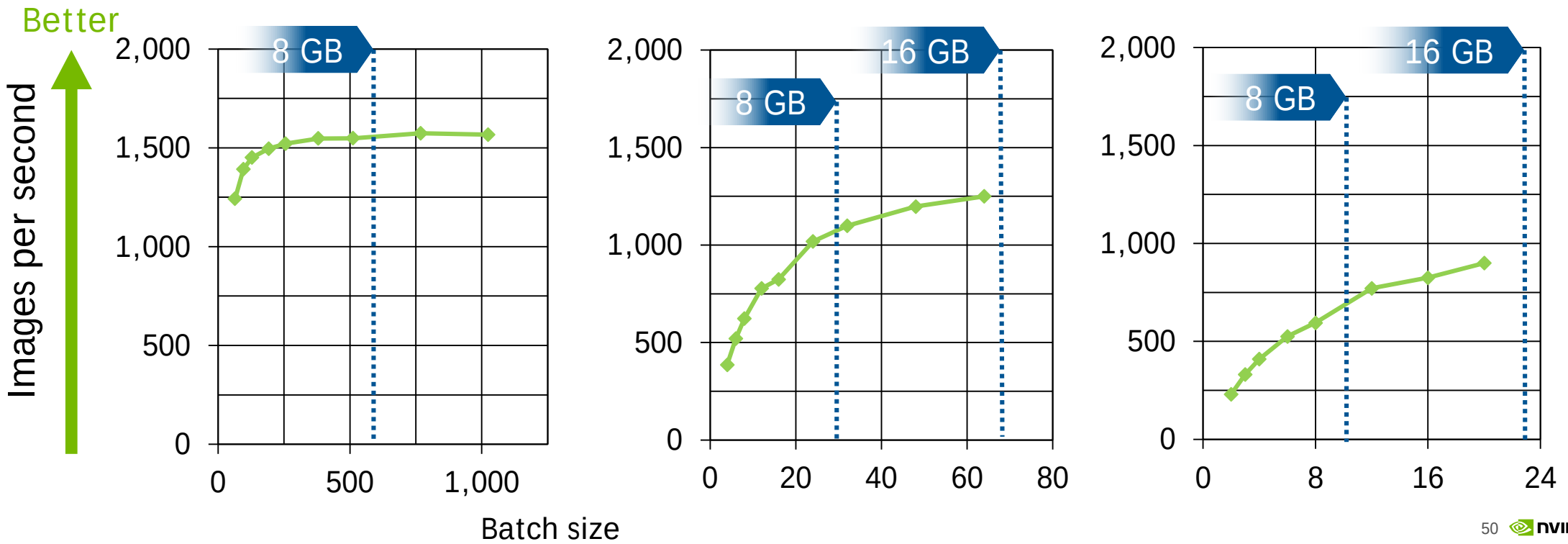


バッチサイズと計算性能

TESLA P100, Chainer 1.17.0



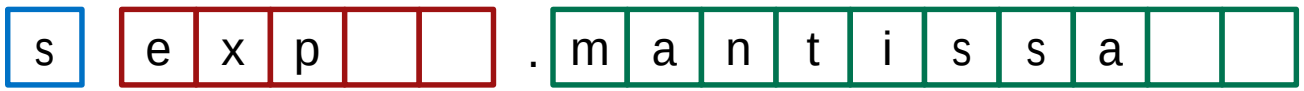
AlexNet (7 layers) 2012 | VGG-D (16 layers) 2014 | ResNet (152 layers) 2015



FP16

[TESLA P100] FP16: **半精度浮動小数点**

IEEE 754準拠

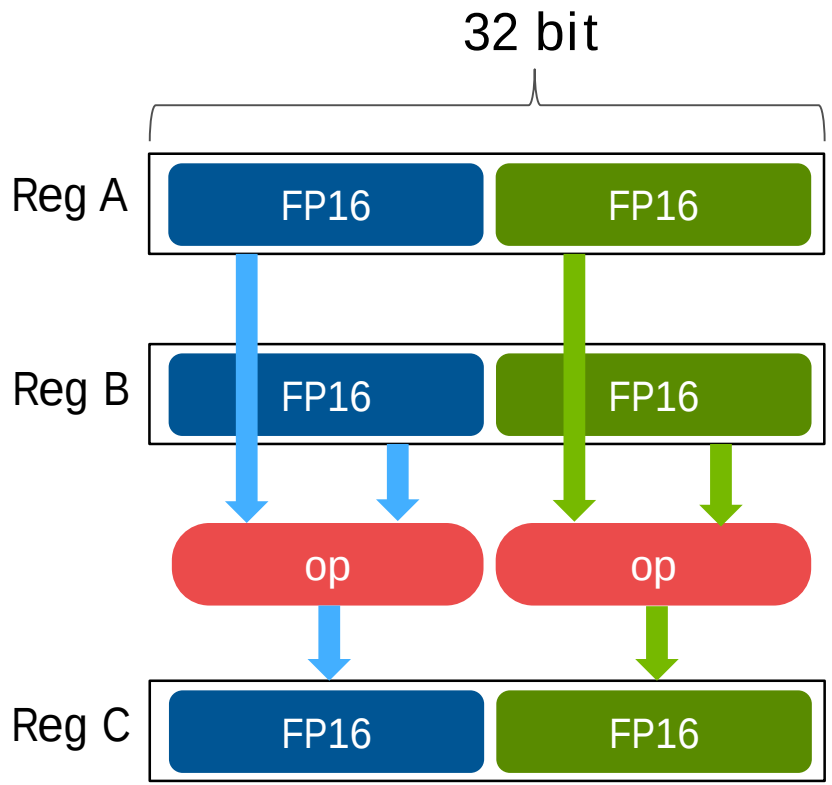
- 16-bit 

  - 符号部:1-bit, 指数部:5-bits, 仮数部:10-bit
- ダイナミックレンジ: 2^{40}
  - Normalized values: 2^{-14} ～ 65504 (≈2^{16})
  - Subnormal at full speed

ユースケース
ディープラーニング
画像処理
信号処理

[TESLA P100] FP16: **半精度浮動小数点**

- 2-way SIMD
  - 32ビットレジスタ内に2x FP16
  - **メモリフットプリント**: FP32**の半分**
- Instructions
  - Add, Sub, Mul
  - Fma (Fused Multiply and Add)
- 1サイクルあたり4つの算術
  - FP32より2倍高速



cuda_fp16.h cuBLAS cuDNN

マルチGPU学習

データ並列(同期型)

Copy Model, Assigne different data

Inputs

GPU 1

Layer 1

W

Layer 2

W

Layer N

W

Labels

“cat”

LossFunc

GPU 2

W

W

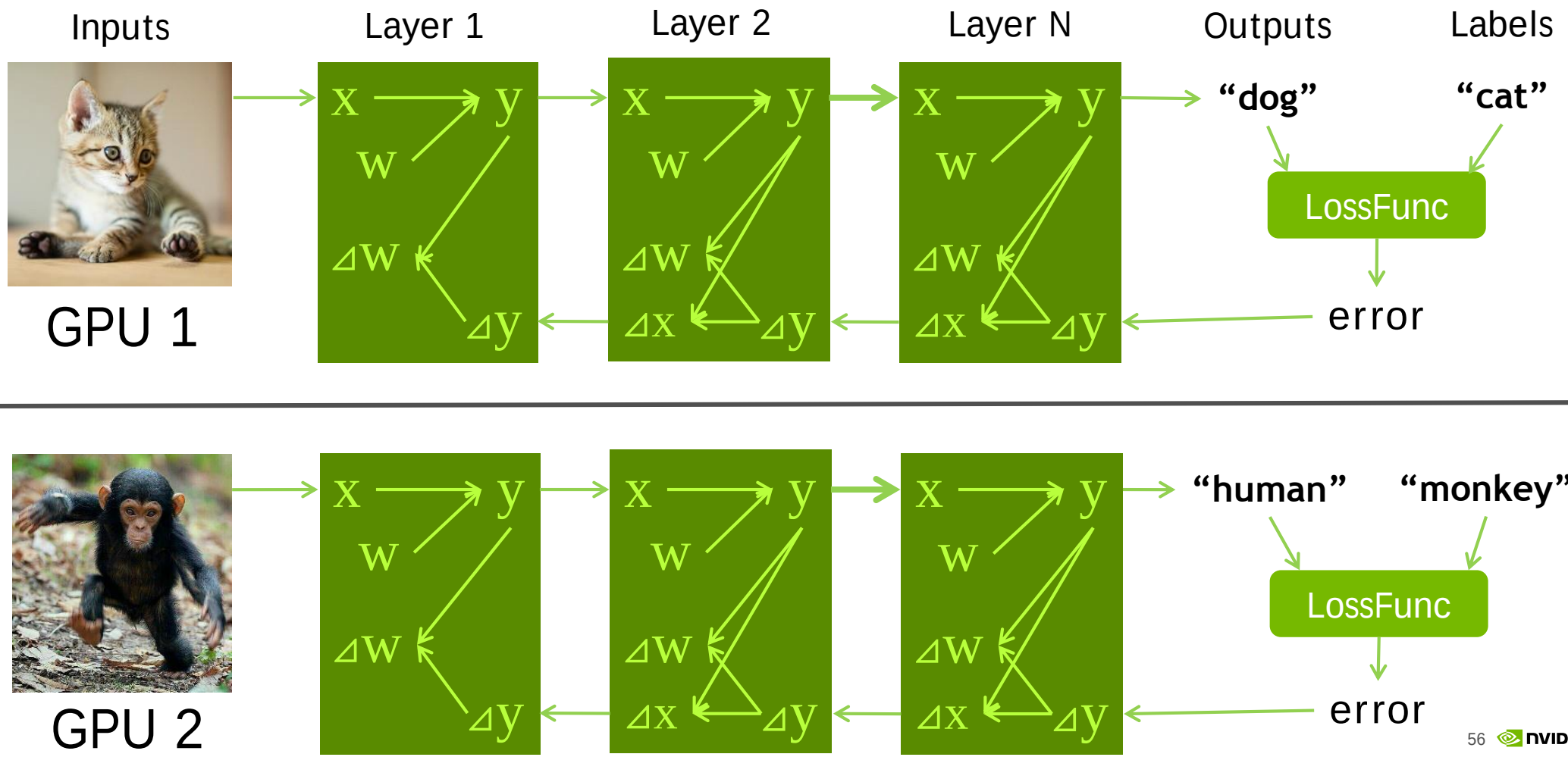
W

“monkey”

LossFunc

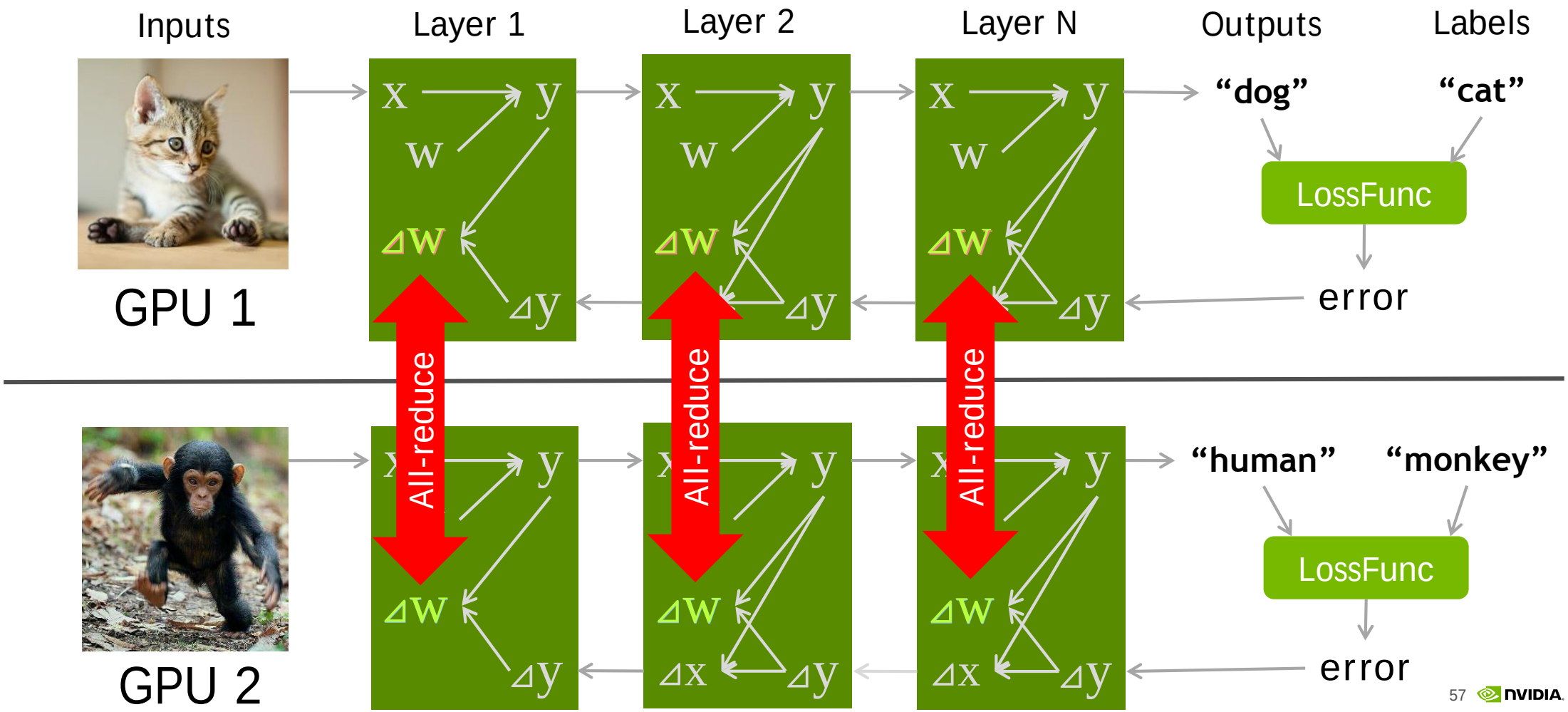
データ並列(同期型)

Forward & Backward Independently



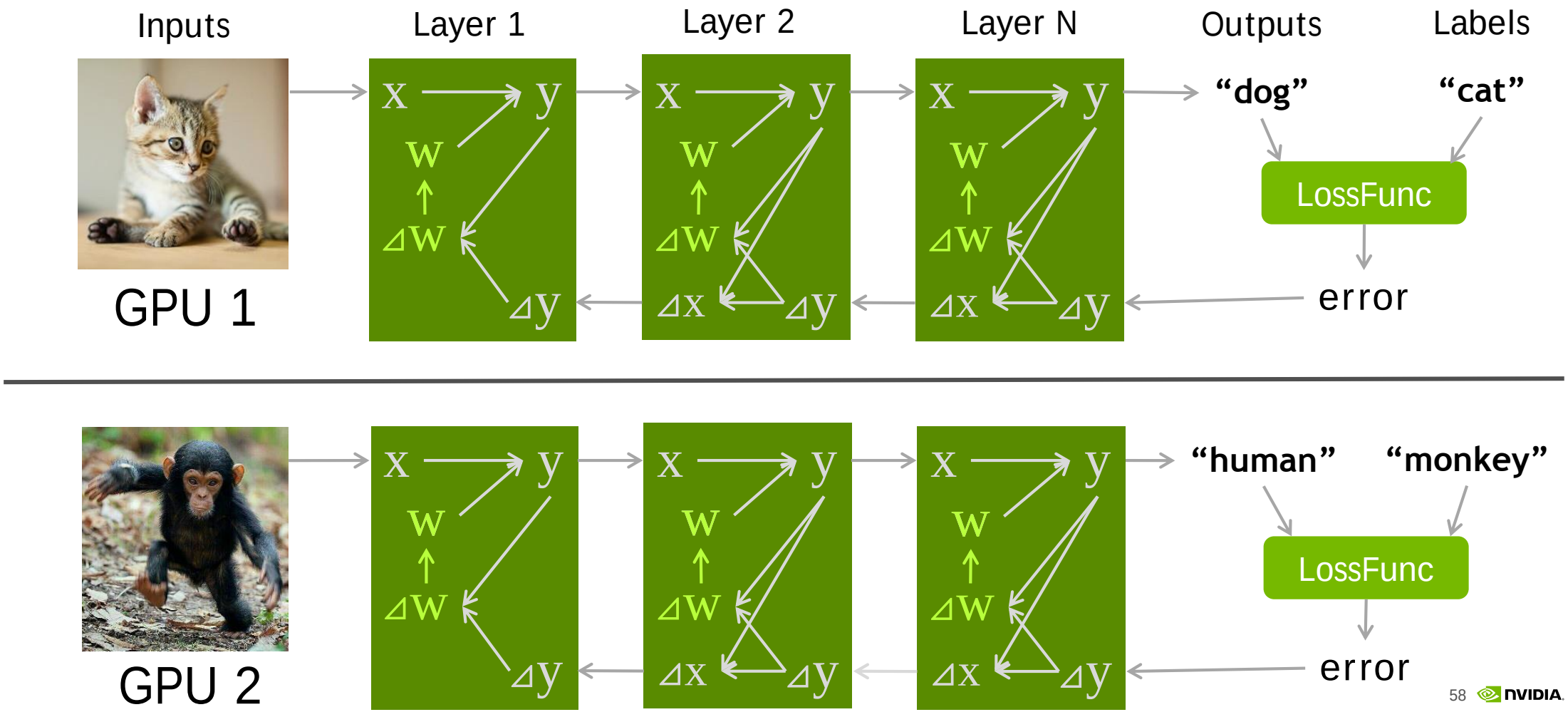
データ並列(同期型)

Combine ⊿w over multi-GPU



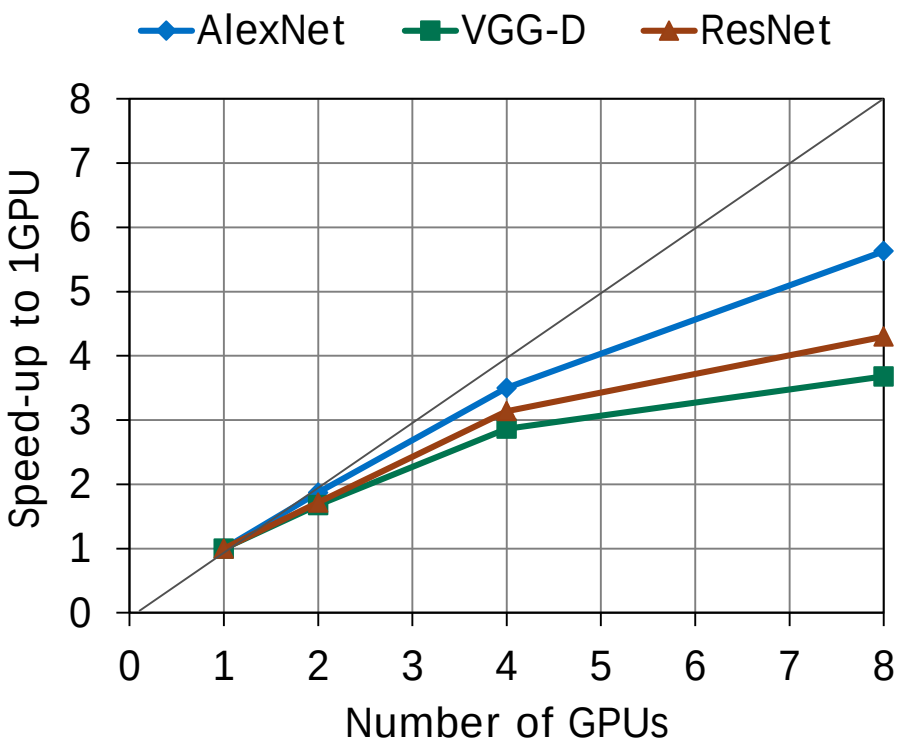
データ並列(同期型)

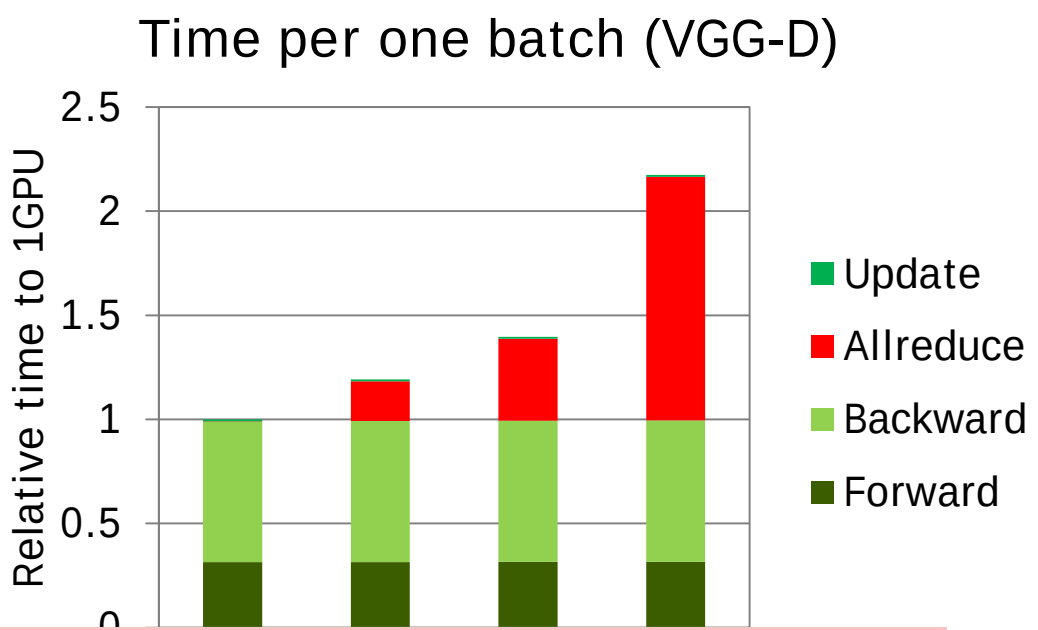
Update Weights Independently



マルチGPU学習のパフォーマンス

NVIDIA DGX-1, Chainer 1.17.0 with multi-process patch



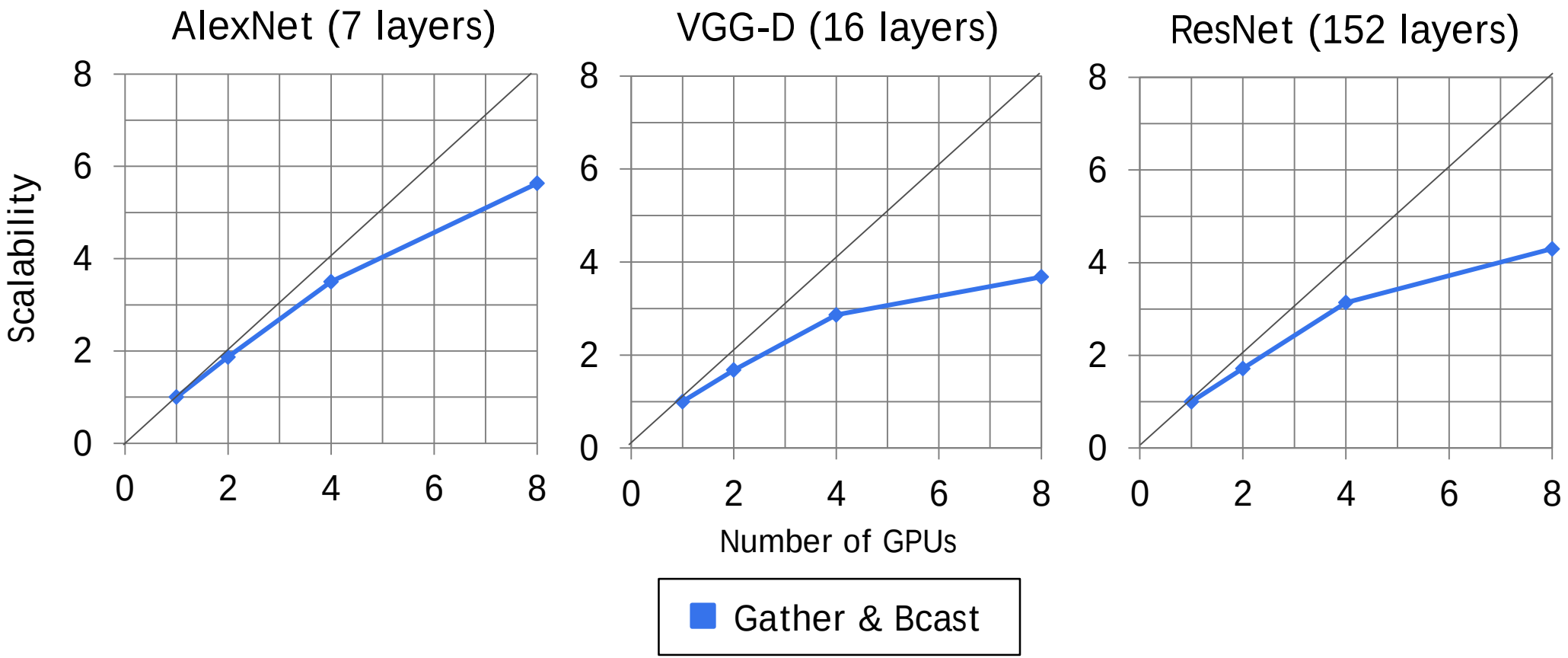


[Batch size per GPU] AlexNet:768, VGG-D:32, ResNet:12

DGX-1's NVLink is not well utilized.
Chainer's all-reduce implementation is naïve "gather and broadcat".

マルチGPU学習のパフォーマンス(NCCL使用なし)

NVIDIA DGX-1, Chainer 1.17.0 with multi-process patch

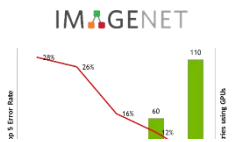


[Batch size per GPU] AlexNet:768, VGG-D:32, ResNet:12

NCCL

ディープラーニング SDK

ディープラーニングを加速するディープラーニングライブラリ



画像分類

物体検出

コンピュータビジョン

音声認識

言語翻訳

ボイス&オーディオ

推薦エンジン

感情分析

自然言語処理

Caffe
Chainer
DL4J
Deeplearning4j
Mocha.jl
julia
K
KERAS
Microsoft
CNTK
MatConvNet
mxnet
MINERVA
Purine
OpenDeep
Pylearn2
TensorFlow
torch
theano

ディープラーニングフレームワーク

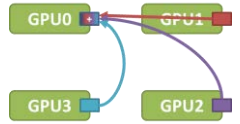
cuDNN

ディープラーニング

cuBLAS
cuSPARSE
cuFFT

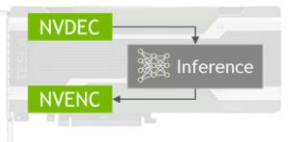
CUDA 数学ライブラリ

NCCL



マルチGPU間通信

DeepStream SDK



ビデオ解析

TensorRT

インファレンス

NCCL(NVIDIA Collective Collection Library)

ディープラーニング SDK

マルチGPU集合通信ライブラリ

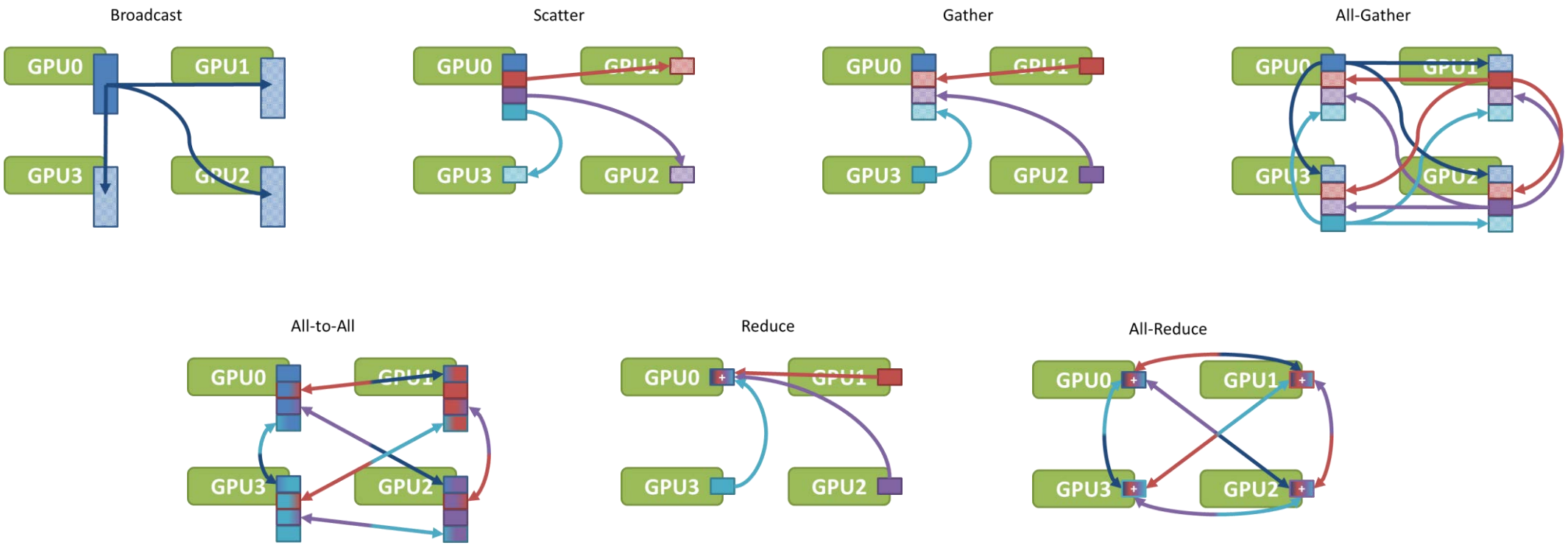
- 最新リリースはv1.2.3
- https://github.com/NVIDIA/nccl

all-gather, reduce, broadcast など標準的な集合通信の処理をバンド幅が出るように最適化

シングルプロセスおよびマルチプロセスで使用する事が可能

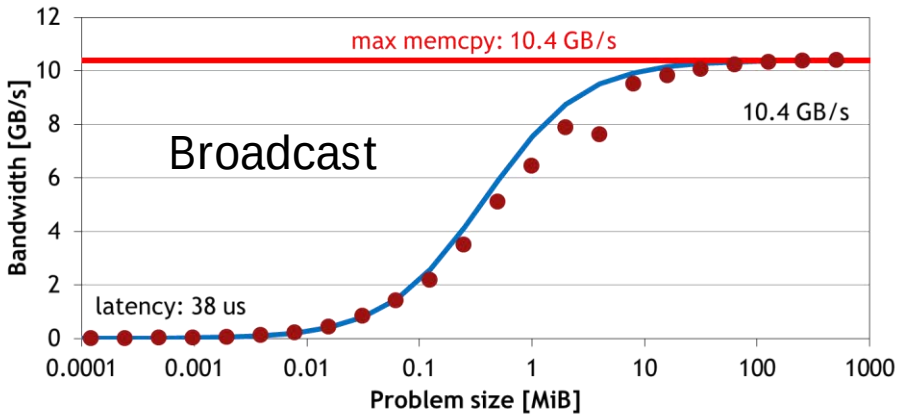
NCCL(NVIDIA Collective Collection Library)

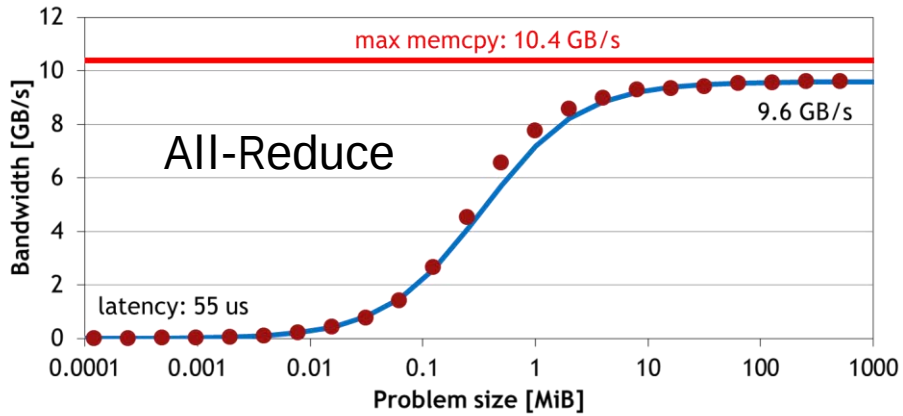
NCCLの集合通信処理

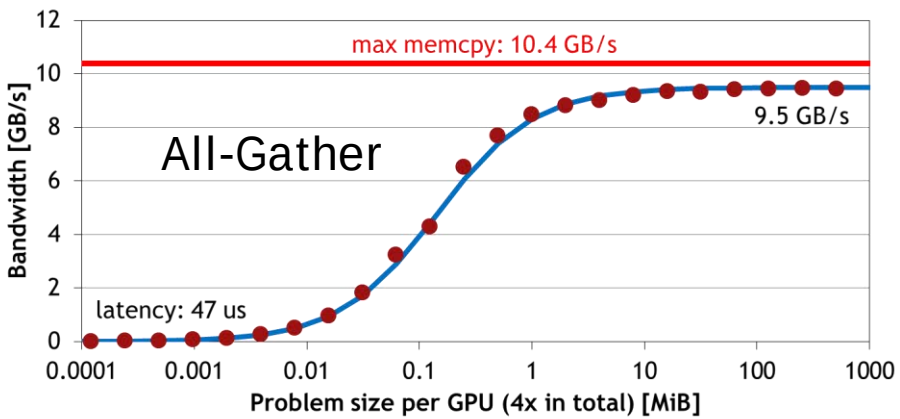


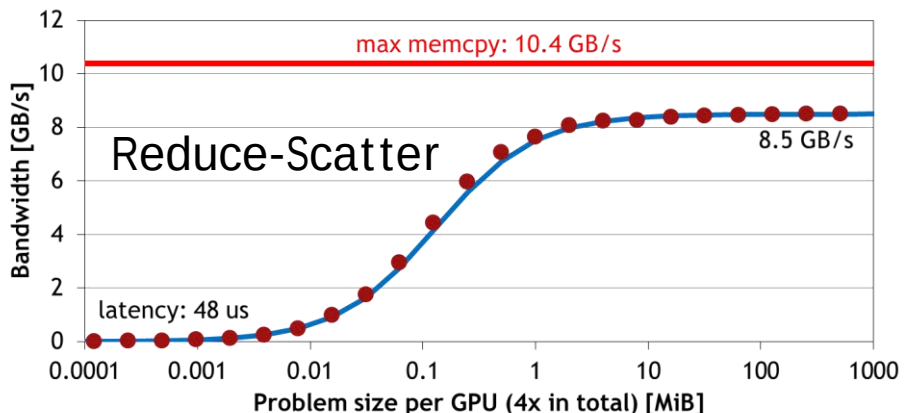
NCCL パフォーマンス

Bandwidth at different problem sizes (4 Maxwell GPUs)



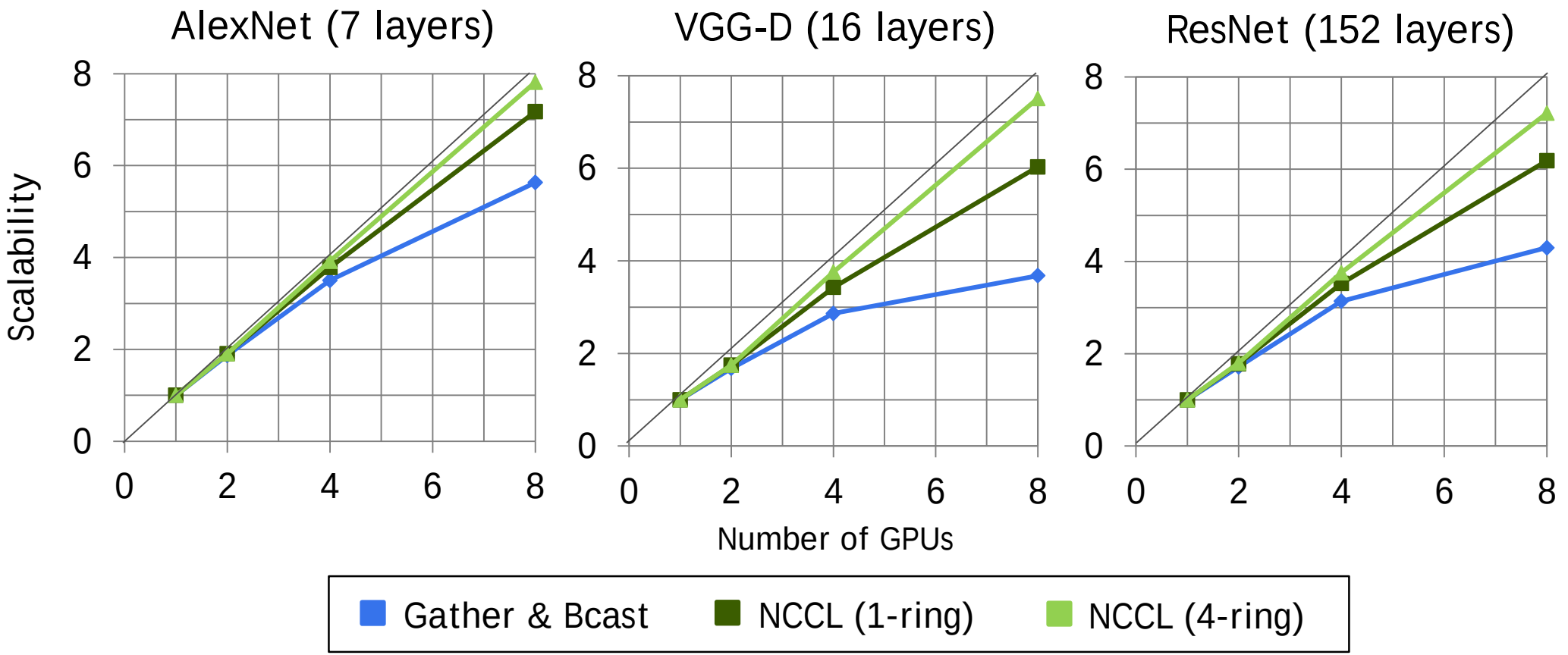






マルチGPU学習のパフォーマンス(NCCL使用)

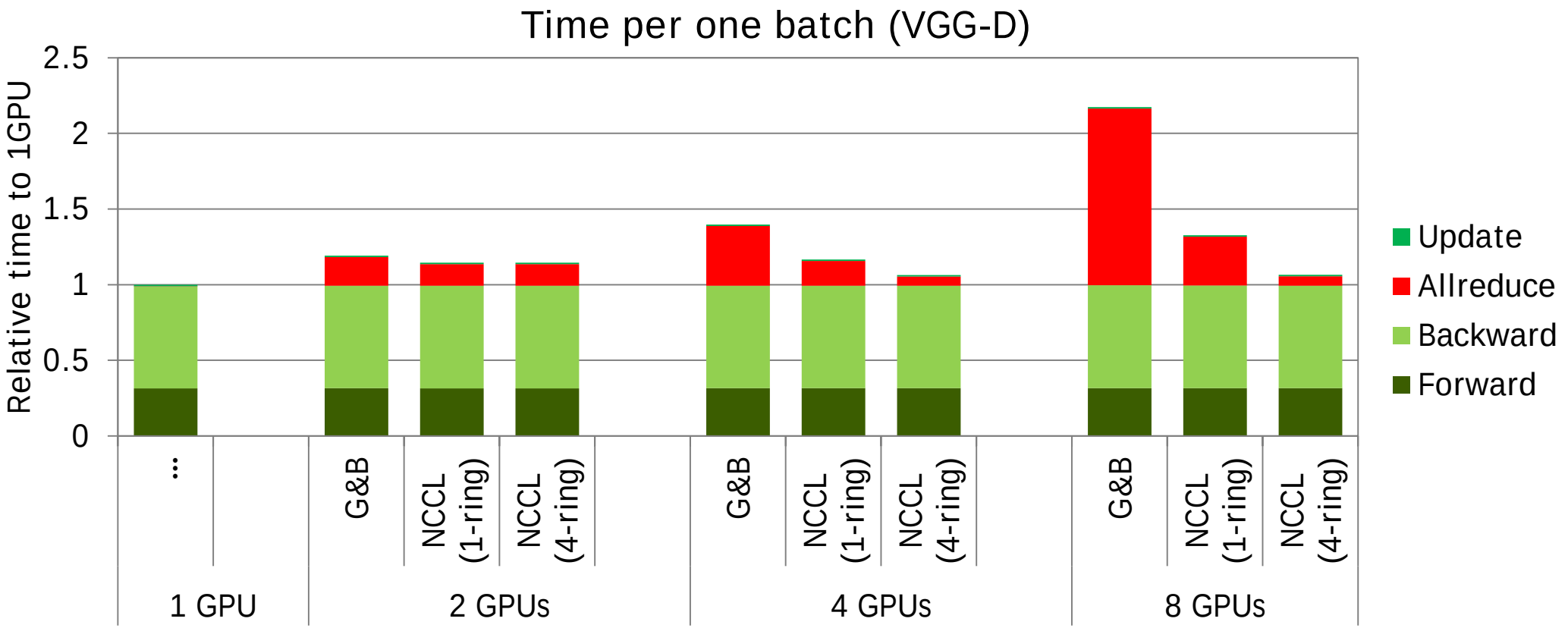
NVIDIA DGX-1, Chainer 1.17.0 with NCCL patch



[Batch size per GPU] AlexNet:768, VGG-D:32, ResNet:12

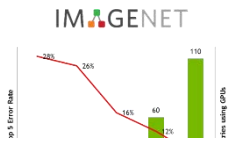
マルチGPU学習のパフォーマンス(NCCL使用)

NVIDIA DGX-1, Chainer 1.17.0 with NCCL patch



ディープラーニング SDK

ディープラーニングを加速するディープラーニングライブラリ



画像分類

物体検出

コンピュータビジョン

音声認識

言語翻訳

ボイス&オーディオ

推薦エンジン

感情分析

自然言語処理

Caffe
Chainer
DL4J
Deeplearning4j
Mocha.jl
julia
K
KERAS
Microsoft
CNTK
MatConvNet
mxnet
MINERVA
Purine
OpenDeep
Pylearn2
TensorFlow
torch
theano

ディープラーニングフレームワーク

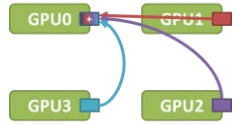
cuDNN

ディープラーニング

cuBLAS
cuSPARSE
cuFFT

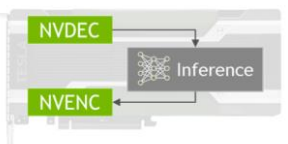
CUDA 数学ライブラリ

NCCL



マルチGPU間通信

DeepStream SDK



ビデオ解析

TensorRT

インファレンス

NVIDIA　ディープラーニング学習コース

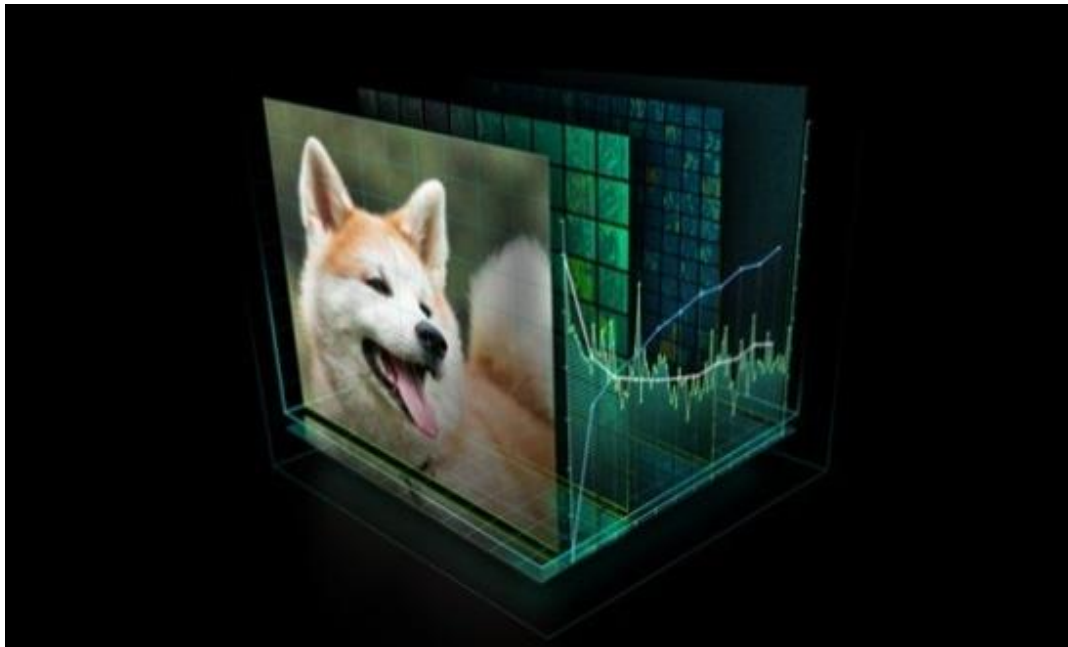
NVIDIA ディープラーニング・インスティチュート

ディープラーニングの為の自習型のクラス。ハンズオンラボ、講義資料、講義の録画を公開。

ハンズオンラボは日本語で受講可能

https://developer.nvidia.com/deep-learning-courses

1. ディープラーニング入門
2. DIGITSによる画像分類入門
3. DIGITSによる物体検出入門
4. etc…



GPU TECHNOLOGY CONFERENCE

#GTC17

GTC 2017 参加登録受付中

2017/5/8 - 11 **サンノゼで開催**

基調講演

テクニカルセッション

ハンズオンラボ

ポスター展示

専門家との交流

スペシャルイベント

http://www.gputechconf.com/

THANK YOU!