GPU TECHNOLOGY CONFERENCE

GTC 2016の基調講演から

エヌビディア合同会社 プラットフォームビジネス本部

部長　林 憲一

PRESENTED BY NVIDIA.

GTC 2016

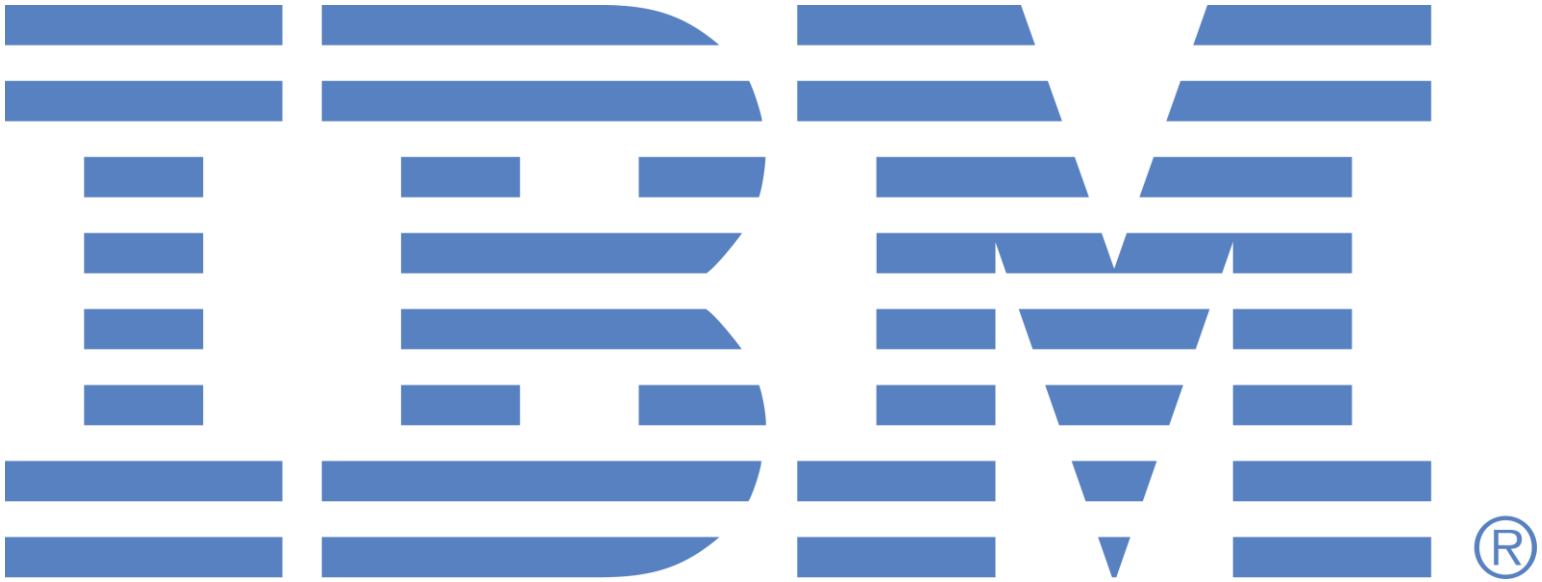
- 2016年 4月 4～7日　米サンノゼコンベンションセンター
- 世界 54 ヵ国から参加者 5519人 ＋ エヌビディア社員 805人
- 608 セッション　150ポスター
- 208 出展社

GTC 2016 基調講演

ジェンスン・ファン
共同創設者、社長兼CEO
4月5日

ロブ・ハイ
IBM Watson CTO
4月6日

ギル・プラット
トヨタリサーチインスティテュート CEO
4月7日



日本アイ・ビー・エム株式会社
ハイエンド・システム事業部 IBM Distinguished Engineer
清水　茂則様

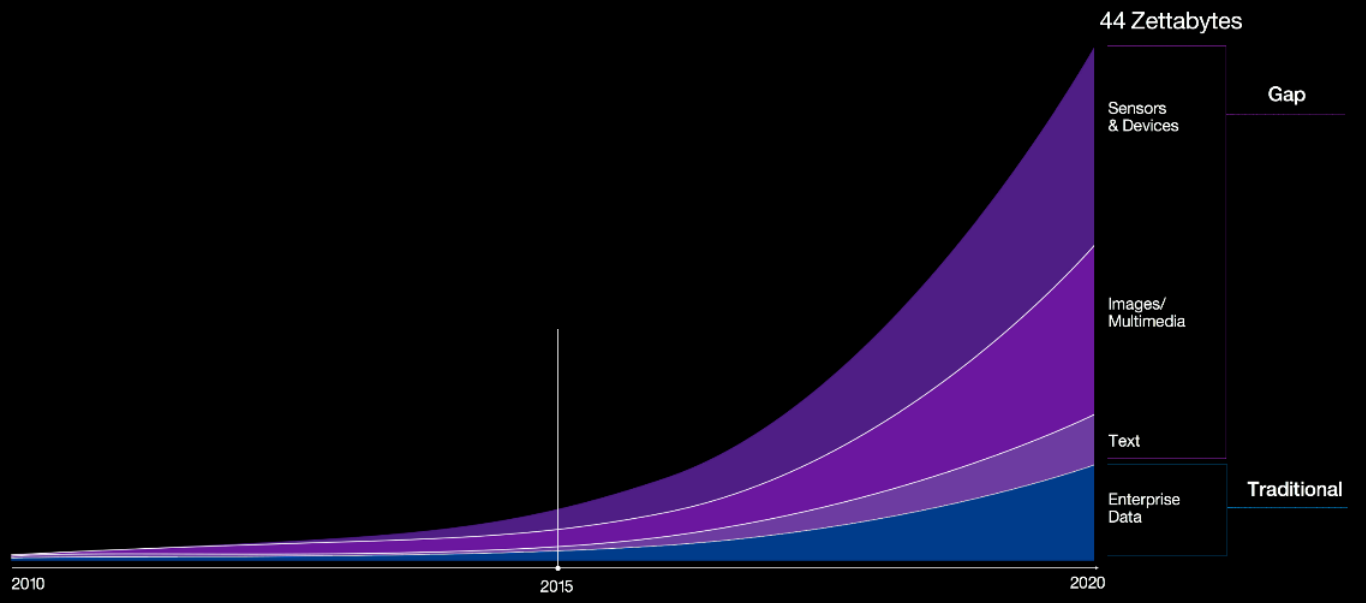
IBM Watson
Advances in Artificial Intelligence

Rob High, Jr.
IBM Fellow, Vice President
Chief Technology Officer
IBM Watson

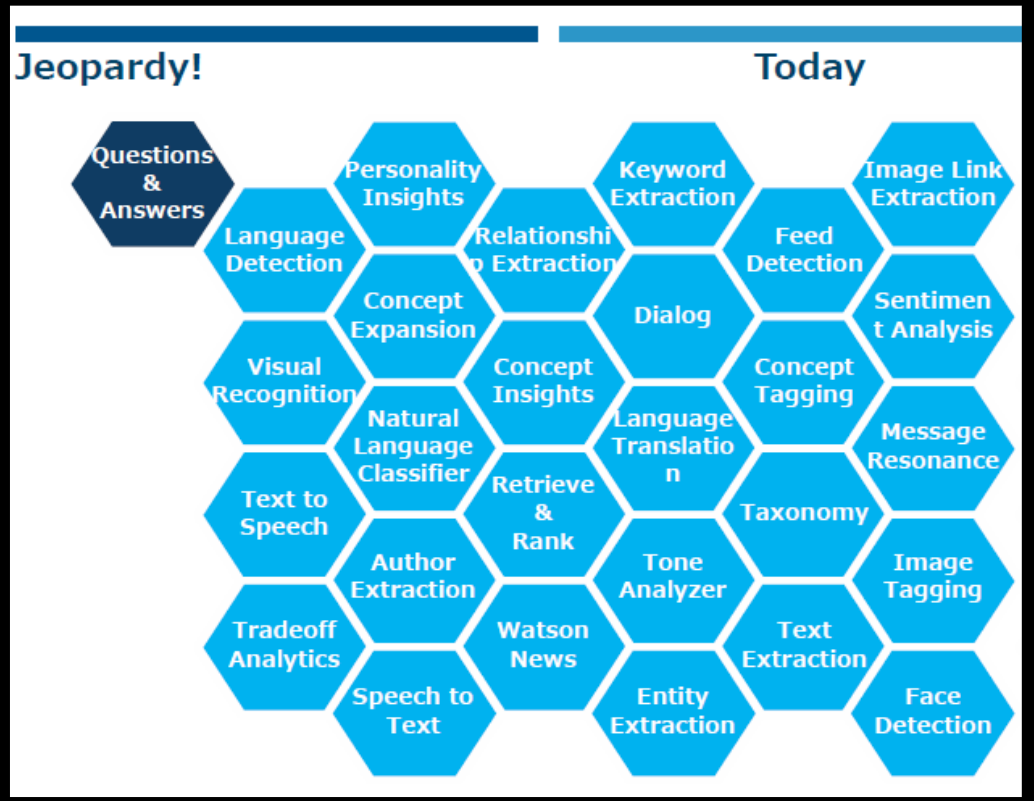
Shigenori Shimizu
IBM Distinguished Engineer
Data Centric Computing
IBM Systems, Hardware

Watson was Introduced to Jeopardy! Audiences in Feb 2011

What is driving the need for Cognitive Computing?



Watson Cognitive Services built on Bluemix

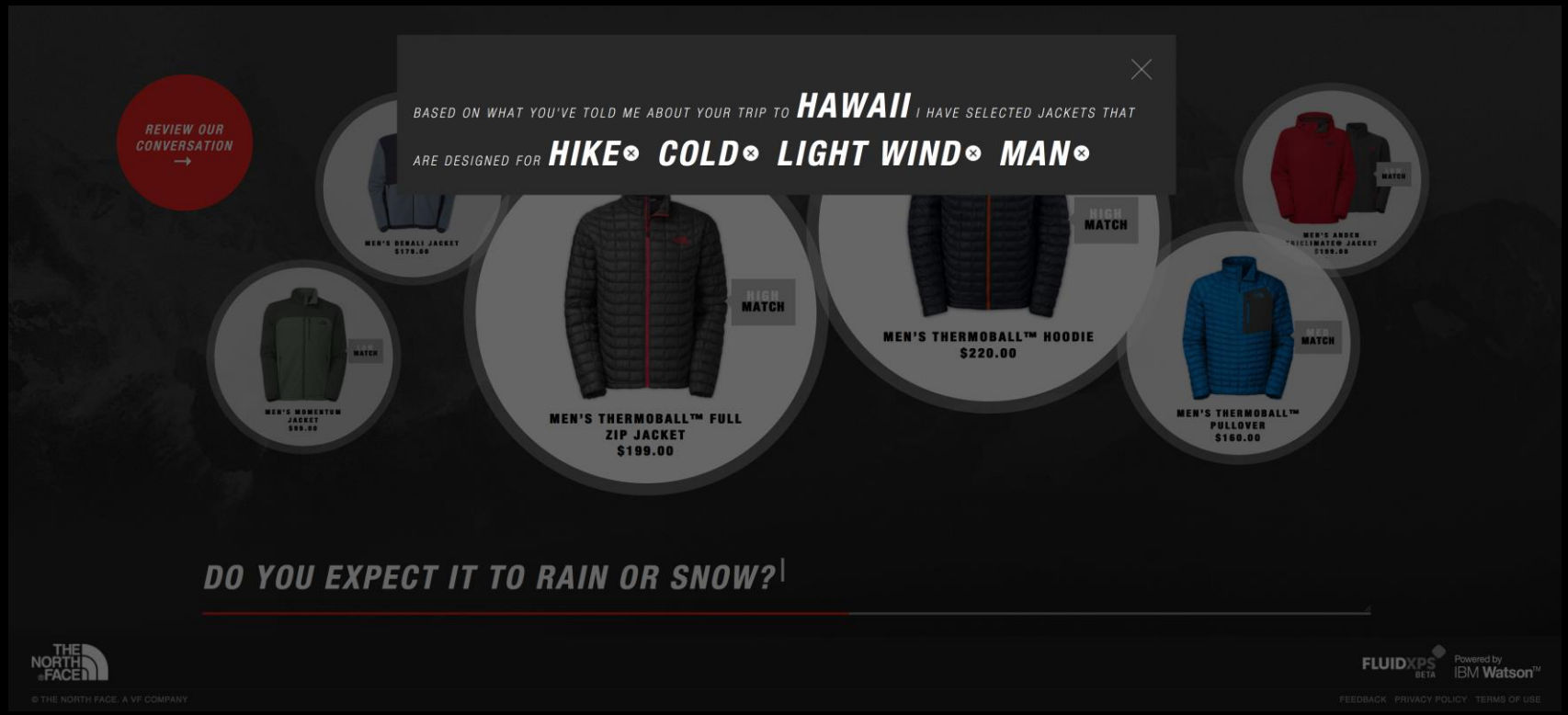


- Build your application using callable Watson Service APIs at **ibm.com/bluemix**
  - AlchemyLanguage
  - AlchemyVision
  - AlchemyNews
  - Concept Expansion
  - Concept Insights
  - Language Identification
  - Language Translation
  - Natural Language Classifier
  - Personality Insights
  - Relationship Extraction
  - Speech to Text
  - Text to Speech, ····

  Can be combined with the 100s of other available services on Bluemix

Fluid
working with The North Face

Changing the on-line shopping experience

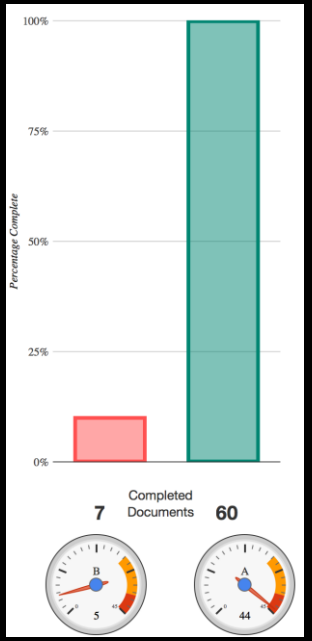


Watson Robotics
Empowering human-machine interaction

- Experiments on integrating Watson with Aldebaran NAO robots (http://www.aldebaran.com/en)
- Anthropomorphic animation
- Vocal/auditory interactions
- Responses augmented with anatomical gesturing to punctuate key points

To achieve Cognitive Computing we need bigger, faster, cheaper compute power

- Using GPUs we have improved training time 8.5x



In 10 years, cognitive systems will be to computing what transaction processing is today

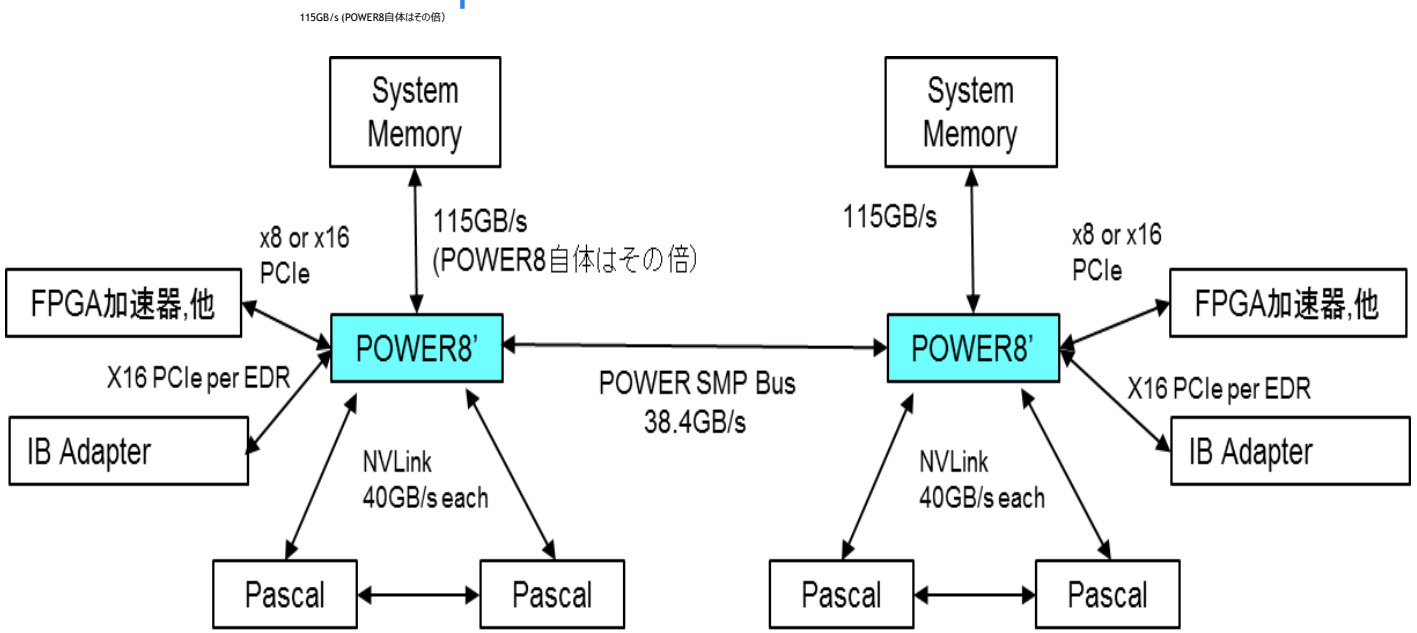
- Amplify human creativity
- Learn their behavior through formal and informal training processes
- Interact with humans on our terms – in the language of humans
- Demonstrate their expertise through trust and depth of character
- Evolve strategies of success – adapting to ever changing knowledge and understanding
- Establish transformative relationships between humans and machines

NVIDIA Pascal 搭載のIBM次期サーバー製品（ご参考）

Exhibited at OpenPOWER Summit 2016

Deep Learningに最適な設計

- 4 GPUs per Node
- NVLink for CPU-GPU and GPU-GPU
- FPGA, IBにも余裕のPCI、さらにCAPI
- 2U Cluster Optimized



Thanks for your attention!
IBM

ibmwatson.com facebook.com/ibmwatson @ibmwatson

TOYOTA
RESEARCH INSTITUTE

POWER, PARALLEL AUTONOMY, AND PEOPLE

Gill Pratt | CEO at Toyota Research Institute | GTC 2016

1.2 Million People

Part 2:
Parallel Autonomy

Must we achieve Level 4 to deal with the handoff problem?

AKIO TOYODA'S PRIORITIES

- Safety
- Environment
- Mobility for All
- Fun to Drive

SERIES (CHAUFFER) VS. PARALLEL (GUARDIAN ANGEL)

Aspect	Chauffer (Series Autonomy)	Guardian Angel (Parallel Autonomy)
Duty Cycle	100%	< 1% : only if accident imminent
Liability	Manufacturer	Mostly Driver
Required Competence	All of Driving	Do No Harm
Development	All or Nothing lives lost until done	Incremental lives saved sooner
Driver's Skills	Ignored	Utilized as much as possible
Fun + Love of Car	Decreased car becomes train	Increased allows high performance experience by novice drivers
Handoff Problem ?	Yes	No

Note: Technology Supporting Guardian Angel + Chauffer are similar

SIMULATION

- Repeatable Studies of Human – Machine Interface
- Regression Testing for Software Development
- Amplification of Physical Testing

TOYOTA RESEARCH INSTITUTE

NVIDIA

TRI ANN ARBOR (TRI-ANN)

Palo Alto
~150 People
Guardian Angel

Ann Arbor
~50 People
Chauffer

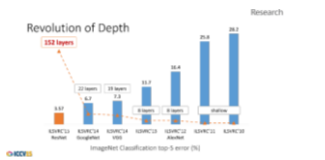
Cambridge
~50 People
Simulation

新しいコンピューティングモデル

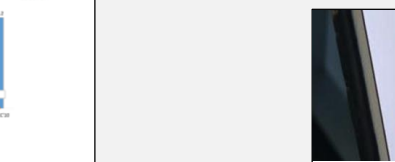
人工知能にとって驚くべき一年

IMAGENET

マイクロソフトとグーグルが
画像認識で人間を超える



マイクロソフト
スーパーディープネットワーク



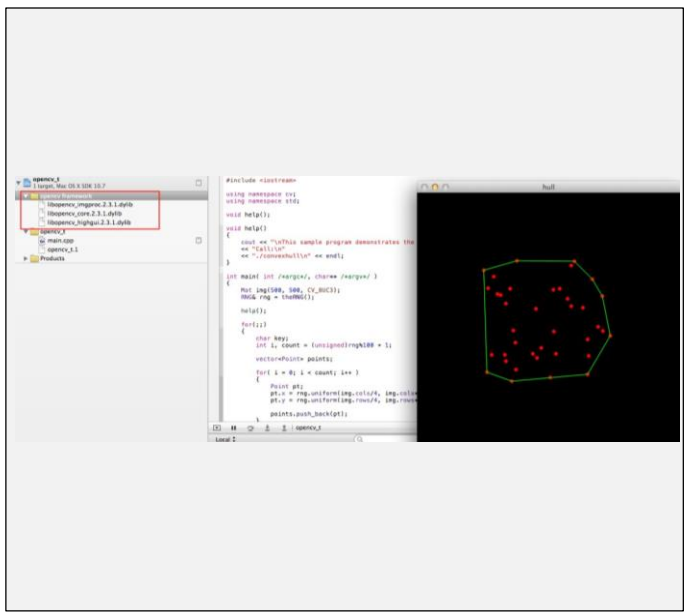
Deep Speech 2
二つの言語を
一つのネットワークで

バークレーのブレット
全てのロボットを
一つのネットワークで

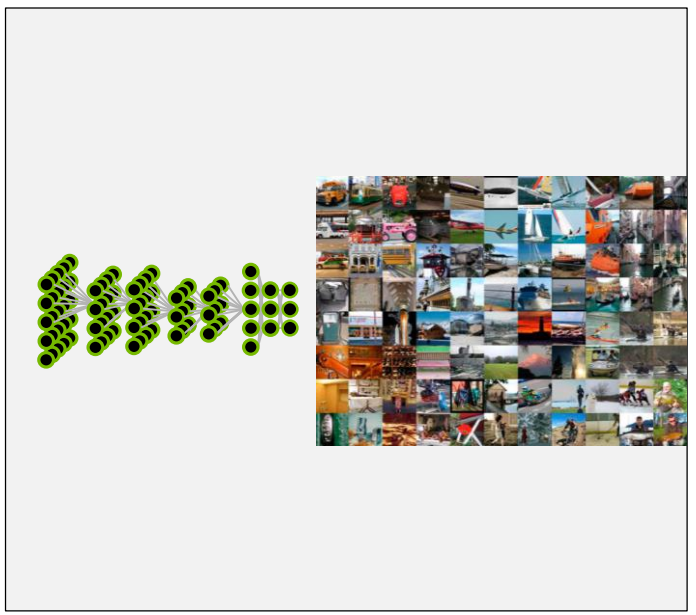
新コンピューティングモデル
がポップカルチャーにも

AlphaGo
世界チャンピオンを倒す

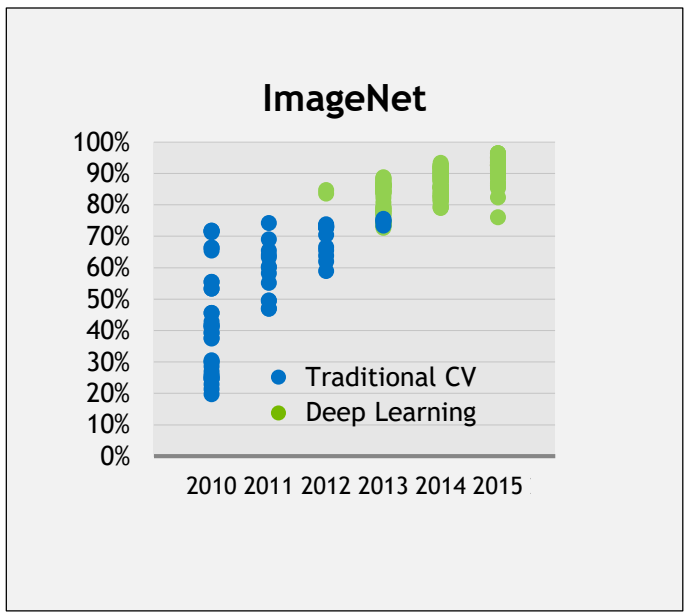
新しいコンピューティングモデル



従来からのコンピュータービジョン
専門家 + 時間



ディープラーニングによる物体認識
DNN + データ + HPC

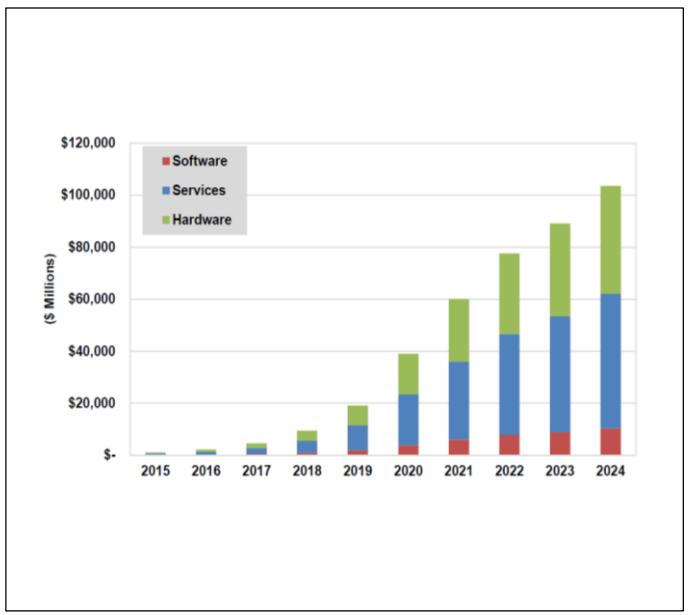


ディープラーニングが
人間を超える成果を達成

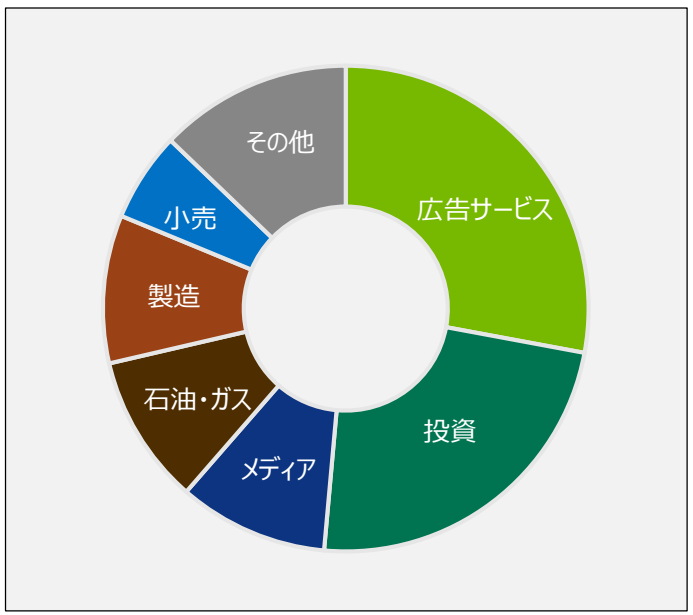
拡がり続けるモダンAIの地平

拡がり続けるモダンAIの地平

今後10年間で50兆円の市場創出



セグメント毎の
ディープラーニングの売上



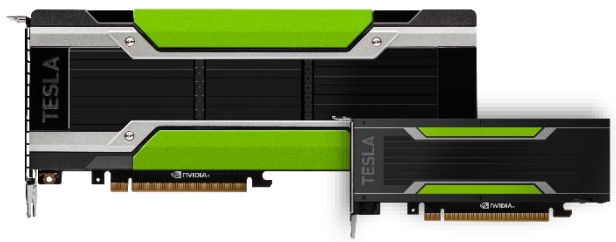
産業別ディープラーニング
ソフトウェアの売上

IBM コグニティブビジネスは
200兆円市場

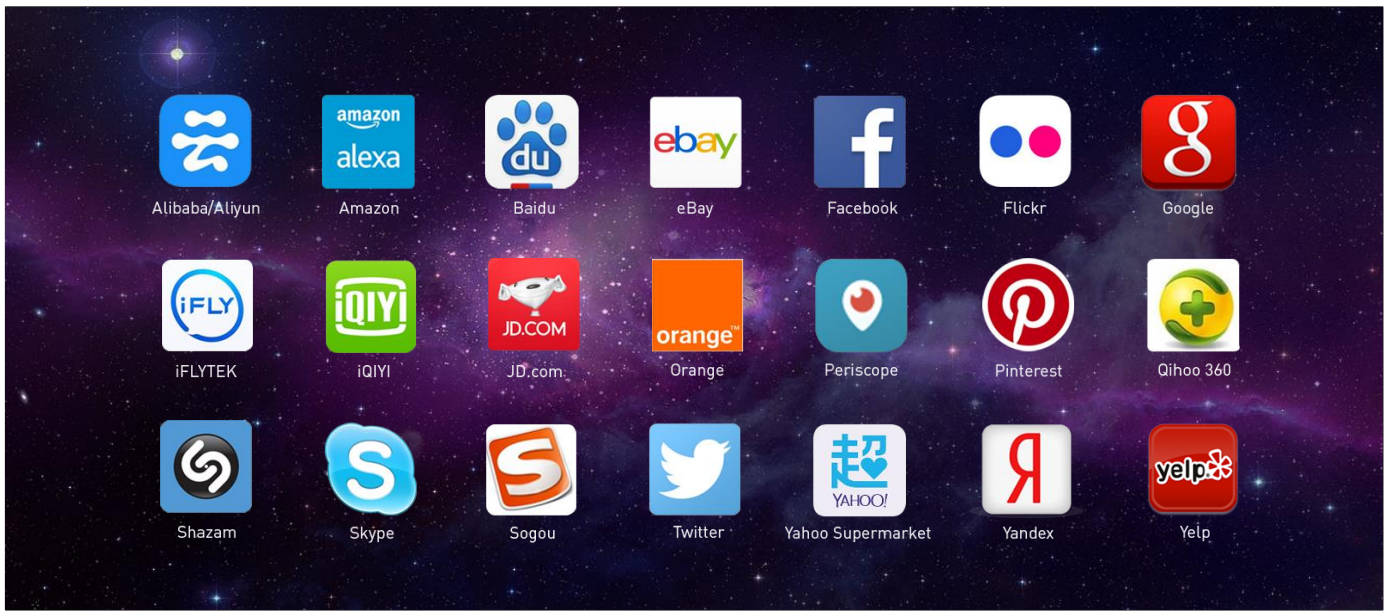
SOURCE: "Deep Learning for Enterprise Applications," 4Q 2015, Tractica

ハイパースケールのための NVIDIA GPU

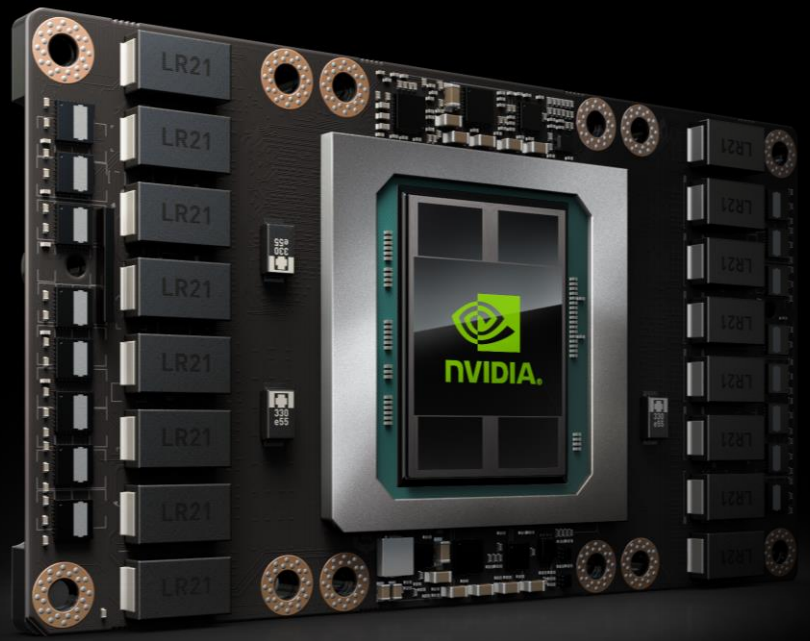
TESLA M40 & TESLA M4



10倍スピードアップ
20 イメージ/秒/ワット



AIを利用したクラウドサービス

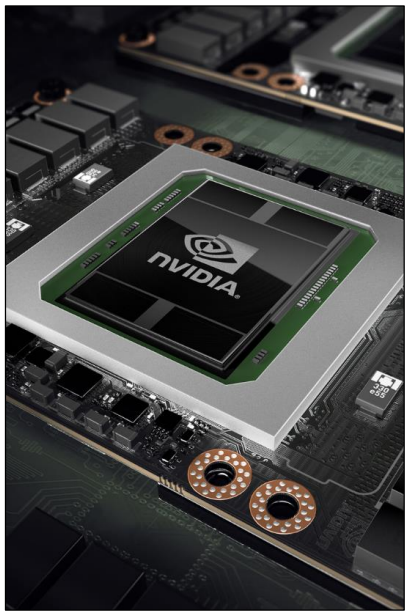


TESLA P100

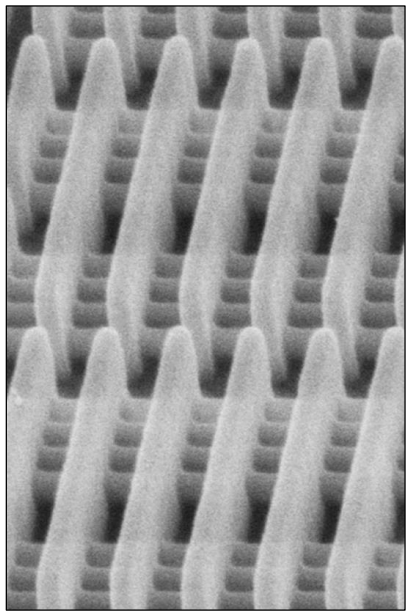
ハイパースケールデータセンターのための
世界で最も先進的な GPU

倍精度 5.3TF | 単精度 10.6TF | 半精度 21.2TF

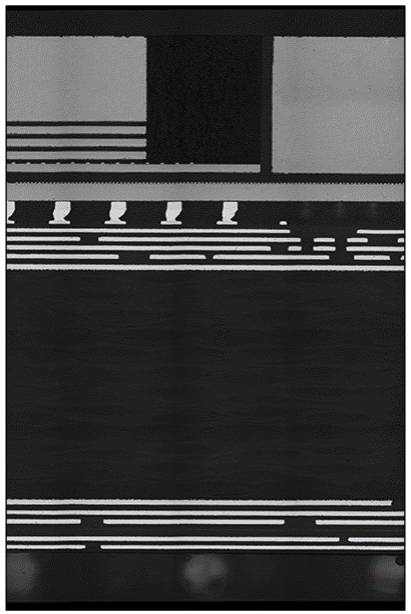
TESLA P100 の先進テクノロジー



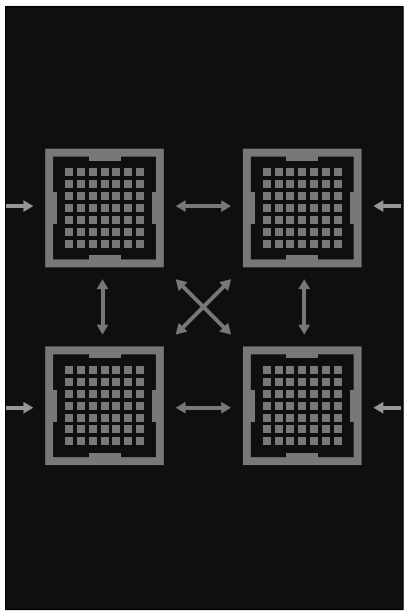
Pascal アーキテクチャ



16nm FinFET

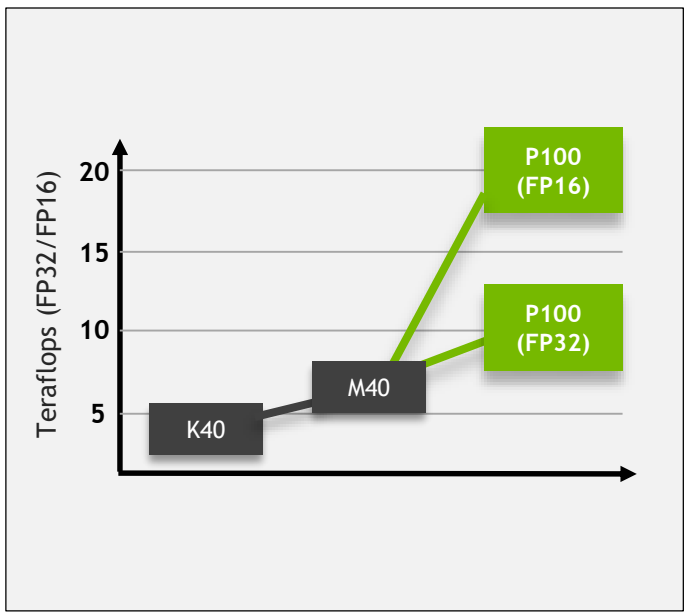


HBM2 積層メモリ

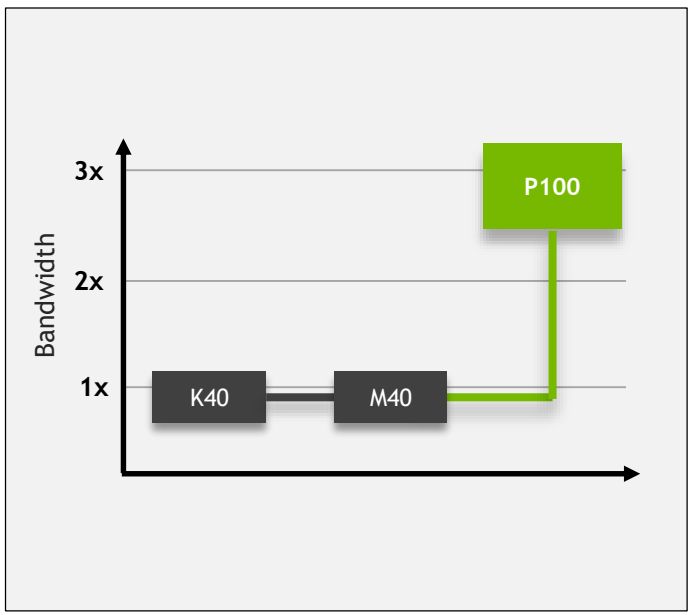


NVLink システム
インターコネクト

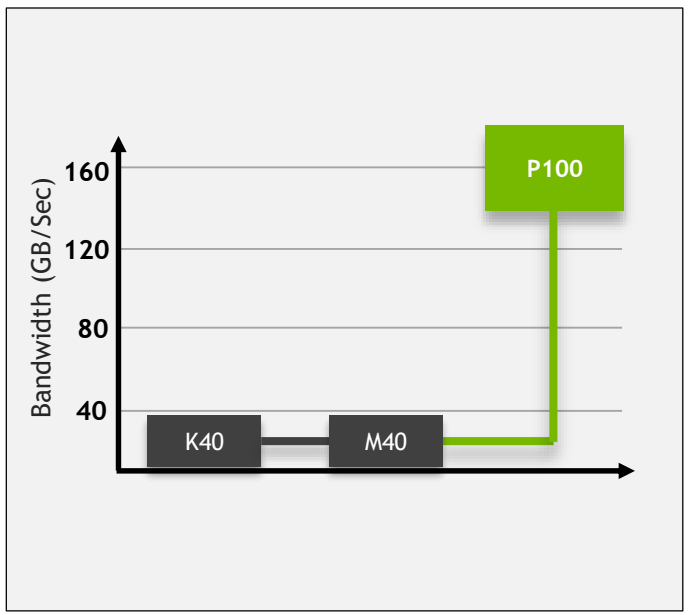
あらゆる面で大きな飛躍



3倍の演算性能



3倍のメモリバンド幅



5倍のGPU間通信速度

TESLA P100 搭載サーバー

2017年第一四半期

NVIDIA DGX-1

世界初のディープラーニング用スーパーコンピュータ

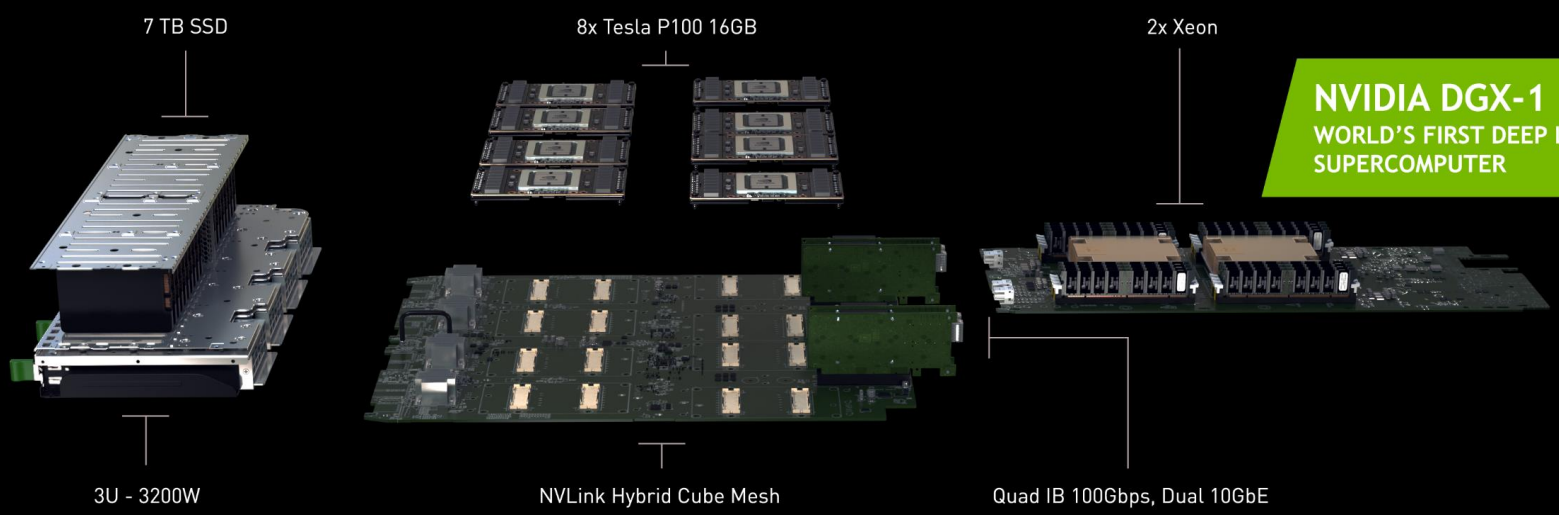
ディープラーニングに最適化

8基の Tesla P100

NVLink システムインターコネクト

半精度 170 テラフロップス

主要AIフレームワークを加速



NVIDIA DGX-1
WORLD'S FIRST DEEP LEARNING SUPERCOMPUTER
170 TFLOPS

“250 台のサーバーがワンボックスに”

	DUAL XEON	DGX-1
FLOPS (CPU + GPU)	3 TF	170 TF
ノード当りの総帯域幅	76 GB/s	768 GB/s
ALEXNET トレーニング時間	150 時間	2 時間
2時間でトレーニングを終えるのに必要なノード数	250 ノード以上*	1 ノード

**Caffe Training on Multi-node Distributed-memory Systems Based on Intel® Xeon® Processor E5 Family (extrapolated)*
Gennady Fedorov (Intel)'s picture Submitted by Gennady Fedorov (Intel), Vadim P. (Intel) on October 29, 2015
https://software.intel.com/en-us/articles/caffe-training-on-multi-node-distributed-memory-systems-based-on-intel-xeon-processor-e5

日本での販売

NVIDIA DGX-1: 世界初のディープラーニング用スーパーコンピュータ

GDEP

HPC SYSTEMS

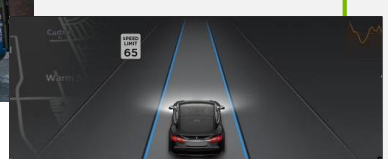
株式会社 日立製作所

セルフドライビングカーへの飛躍の年

Uber の参入

トヨタ自動車がAI研究に
1000億円投資

Audi、BMW、ダイムラー
HERE 買収



Tesla Model S
オートパイロット

Baidu の参入

ボルボが2017年に
自動運転 Drive Me

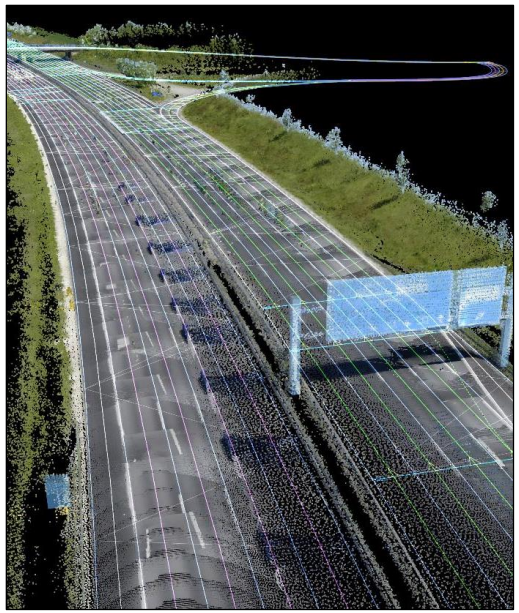
米運輸省、コンピュータを
ドライバとみなす

トヨタ、日産、ホンダなど6社
自動運転で共同研究

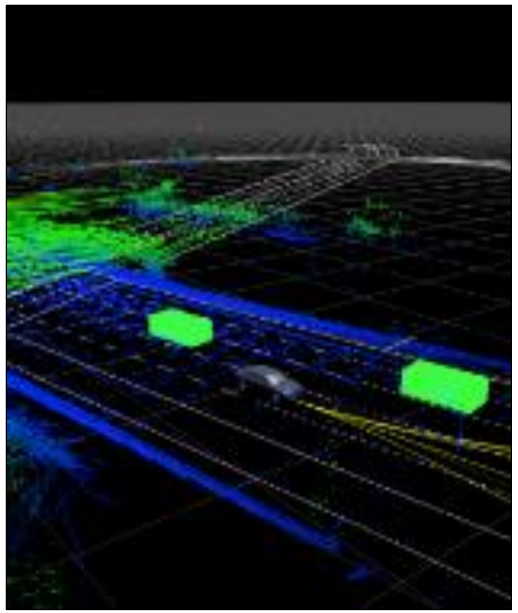
Tesla Model 3
30万台プレオーダー

GM が Cruise 買収

セルフドライビングループ

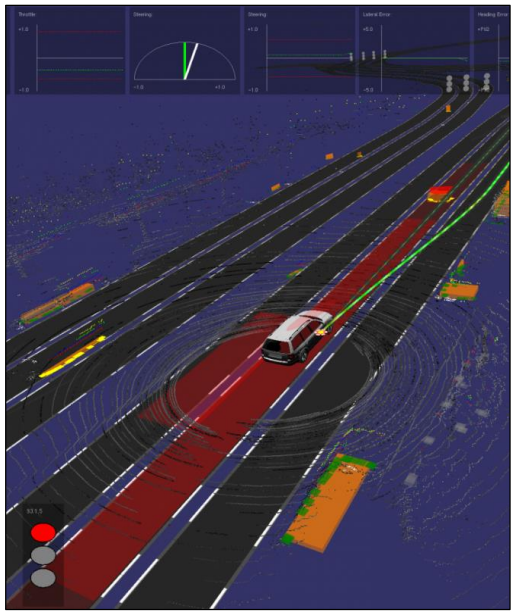


MAP



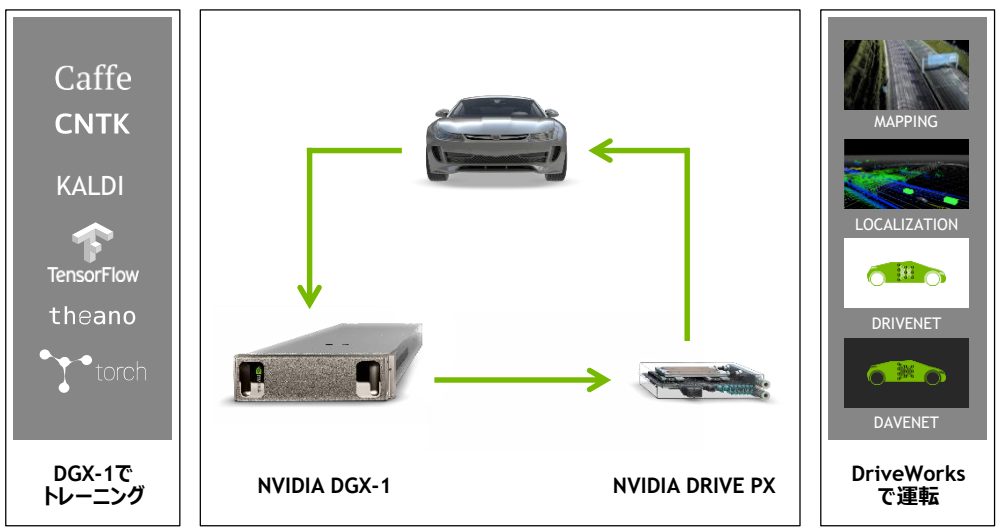
LOCALIZE

SEE



DRIVE

NVIDIA DRIVE PX
AI カー コンピュータ

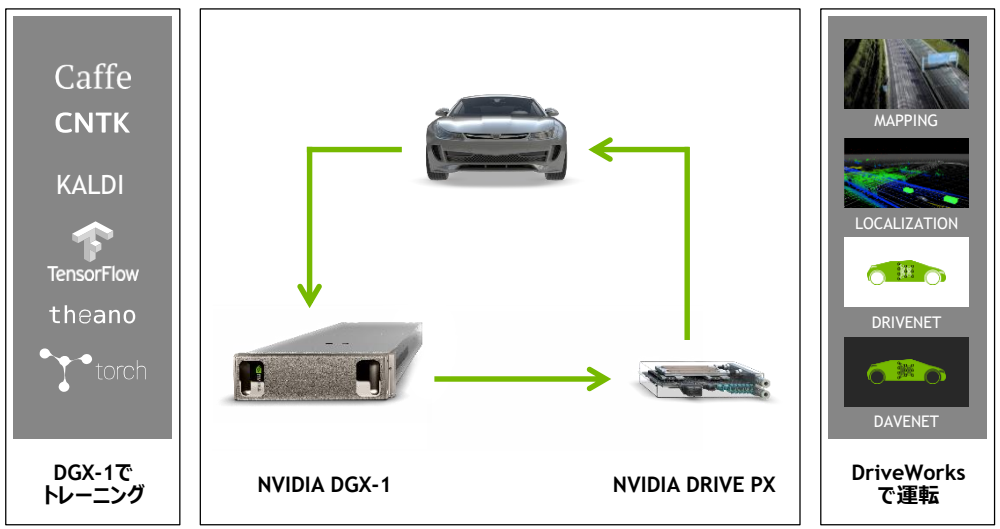


世界初のディープラーニング カー コンピュータ プラットフォーム

End to End スケーラブルアーキテクチャ

オープンプラットフォーム

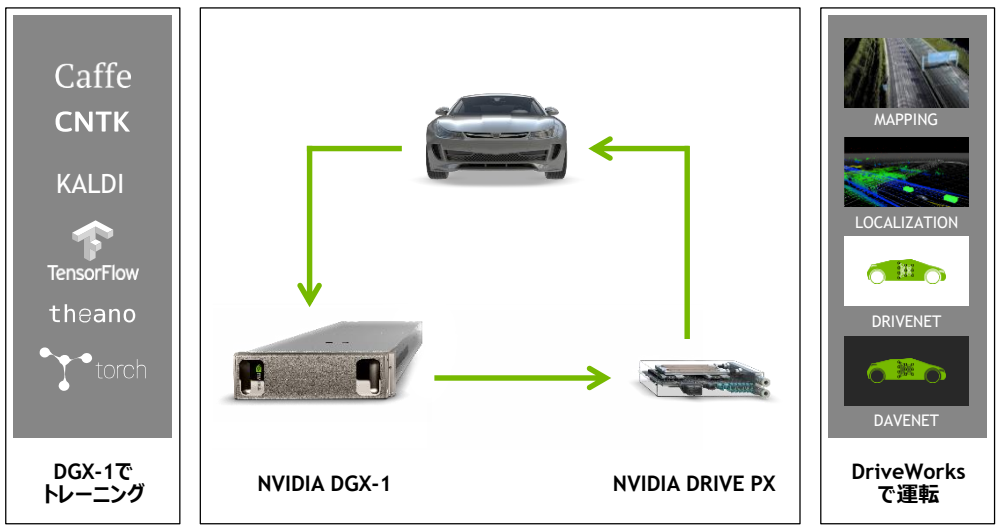
NVIDIA DRIVE PX
パーセプション

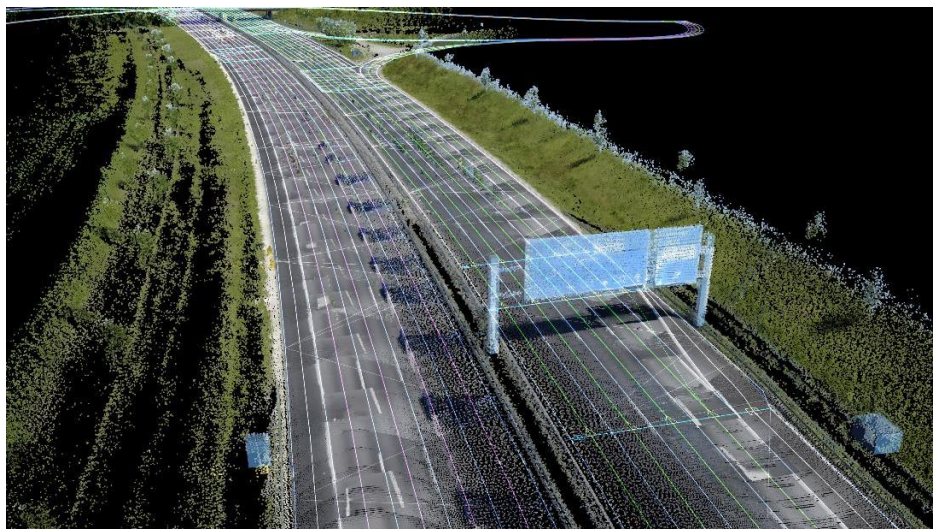


NVIDIA DRIVENET
KITTI 自動車認識で最高スコア

	Method	Hard	Moderate	Easy	Environment
1	NVDriveNet-H	83.76 %	89.81 %	90.92 %	GPU @ 2.5 Ghz (Python + C/C++)
2	sensekitti	79.99 %	89.72 %	91.42 %	GPU @ 2.5 Ghz (Python + C/C++)
3	SDP+RPN	78.38 %	88.85 %	90.14 %	GPU @ 2.5 Ghz (Python + C/C++)
4	Mono3D	78.96 %	88.66 %	92.33 %	GPU @ 2.5 Ghz (Matlab + C/C++)
5	3DOP	79.10 %	88.64 %	**93.04 %**	GPU @ 2.5 Ghz (Matlab + C/C++)

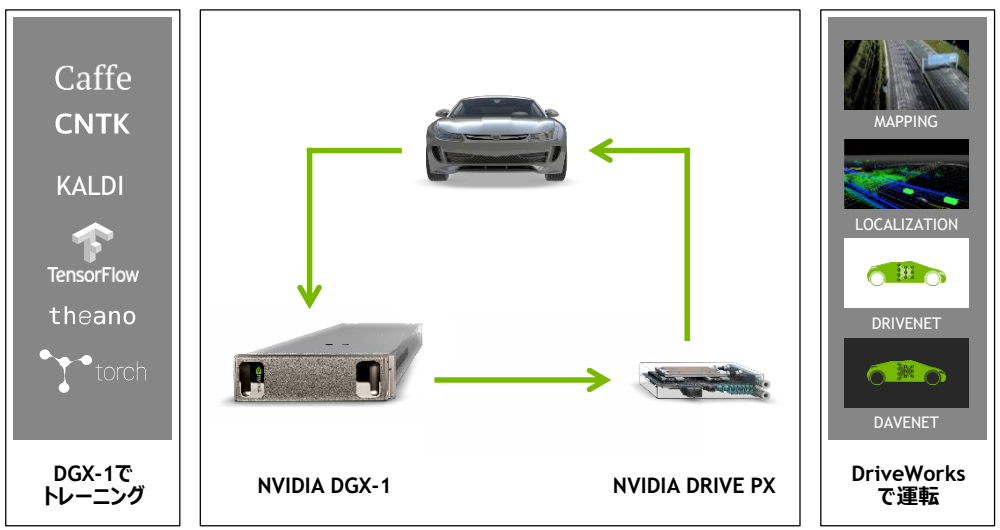
新しい END-TO-END HD マッピング





マッピングプラットフォーム

AI 運転の新たな試み


Caffe
CNTK
KALDI
TensorFlow
theano
torch
DGX-1で
トレーニング
NVIDIA DGX-1
NVIDIA DRIVE PX
MAPPING
LOCALIZATION
DRIVENET
DAVENET
DriveWorks
で運転

世界初の自動運転カーレース

10 チーム 20 台 | NVIDIA DRIVE PX 2が頭脳に | 2016/17 Formula E シーズン

ROBORACE