Recent Trends in Edge Computing

Feb. 2020

~ Agenda ~

1. What is Edge-Computing?
2. AI Chips
3. Edge AI products

~ Summary ~

1. What is Edge-Computing?

- Put heavy calculations on the cloud side to the edge side!

2. AI Chips

- Not only GAFAM, but also the rise of Chinese companies (BATIS)!

3. Edge AI products

- Jetson Nano, which can be bought for $99, is the better choice!

*GAFAM (Google, Amazon, Facebook, Apple, Microsoft)
*BATIS (Baidu, Alibaba, Tencent, iFlytek, SenseTime)

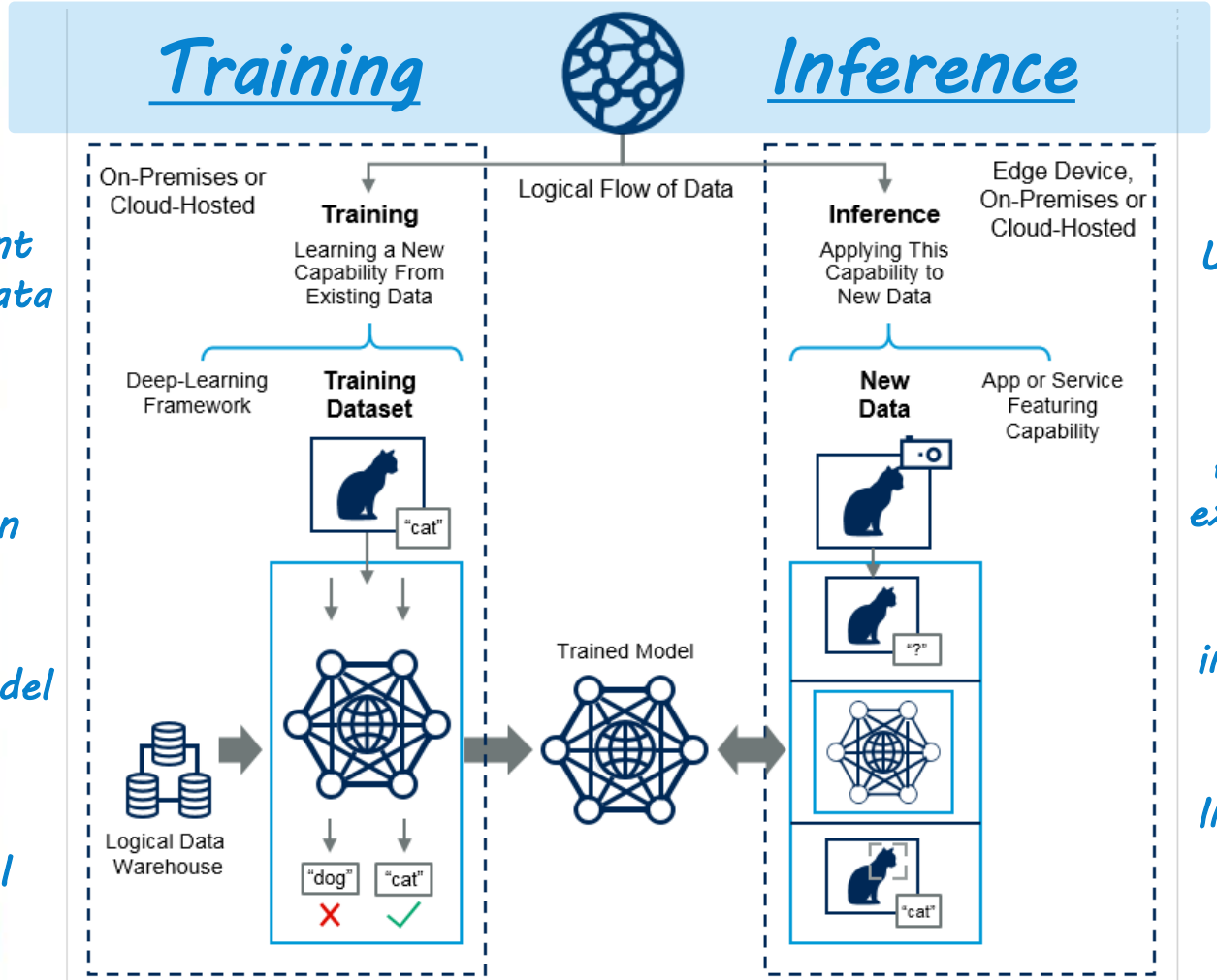
【Premise】 Machine learning process consists of Training and Inference

Large amount of training data

Feature extraction

Generate inference model

Save model



Unknown data

Feature extraction

Apply inference model

Inference result

【Gartner(Feb·2019):Training versus Inference】

~ Agenda ~

1. What is Edge-Computing?
2. AI Chips
3. Edge AI products

Three hurdles facing the cloud

① Data volume

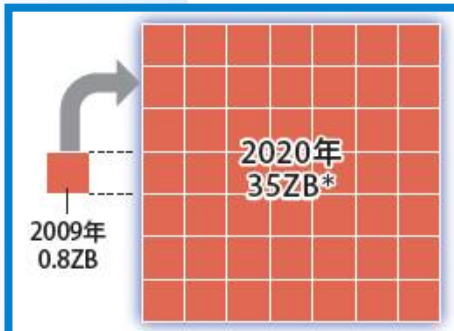
Data volume is expected to increase 44 times in about 10 years!

② Energy

Global electricity consumption continues to increase!

③ Real-time

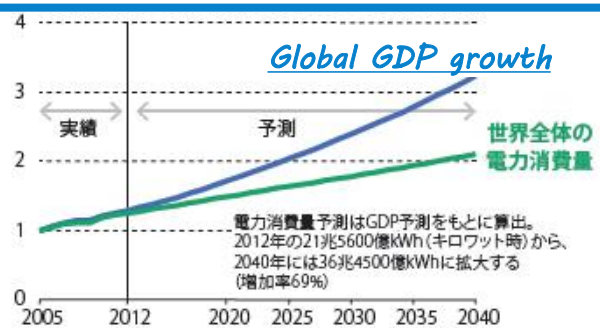
Communication speed is a bottleneck and accidents cannot be avoided!



Insufficient file capacity in simple IoT systems

Needs to select data at the edge terminal!

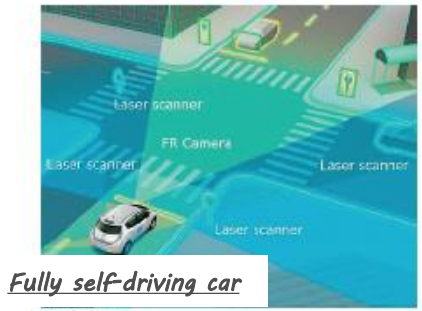
＊ZB:ゼタバイト＝10の21乗バイト
出典：IDC『The Digital Universe Decade – Are You Ready?』



Insufficient energy in simple IoT systems

Needs to store necessary data at the edge terminal!

出典：米Energy. Information Administration
『International Energy Outlook2016』



Fully self-driving car

Real-time processing is not possible with IoT systems

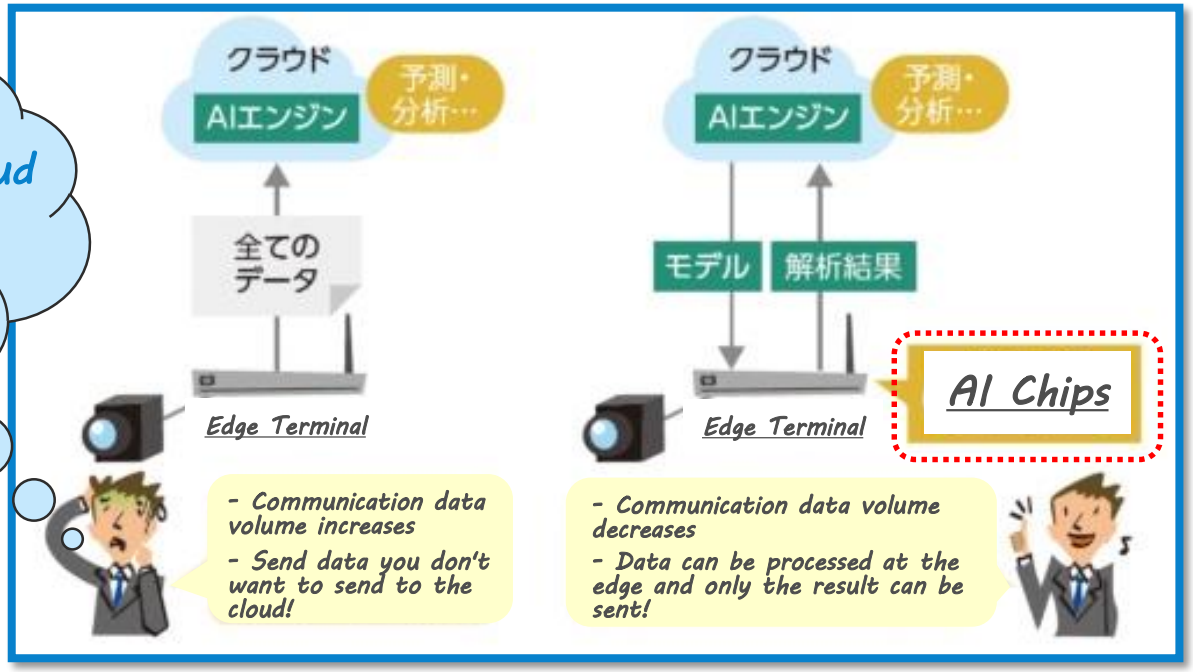
Need intelligence to judge the situation at the edge!

【Three hurdles facing the cloud】

Move heavy calculations to the edge to balance the load on the cloud

Three hurdles for cloud
① Data volume
② Energy
③ Real-time

Cloud Computing → Edge Computing

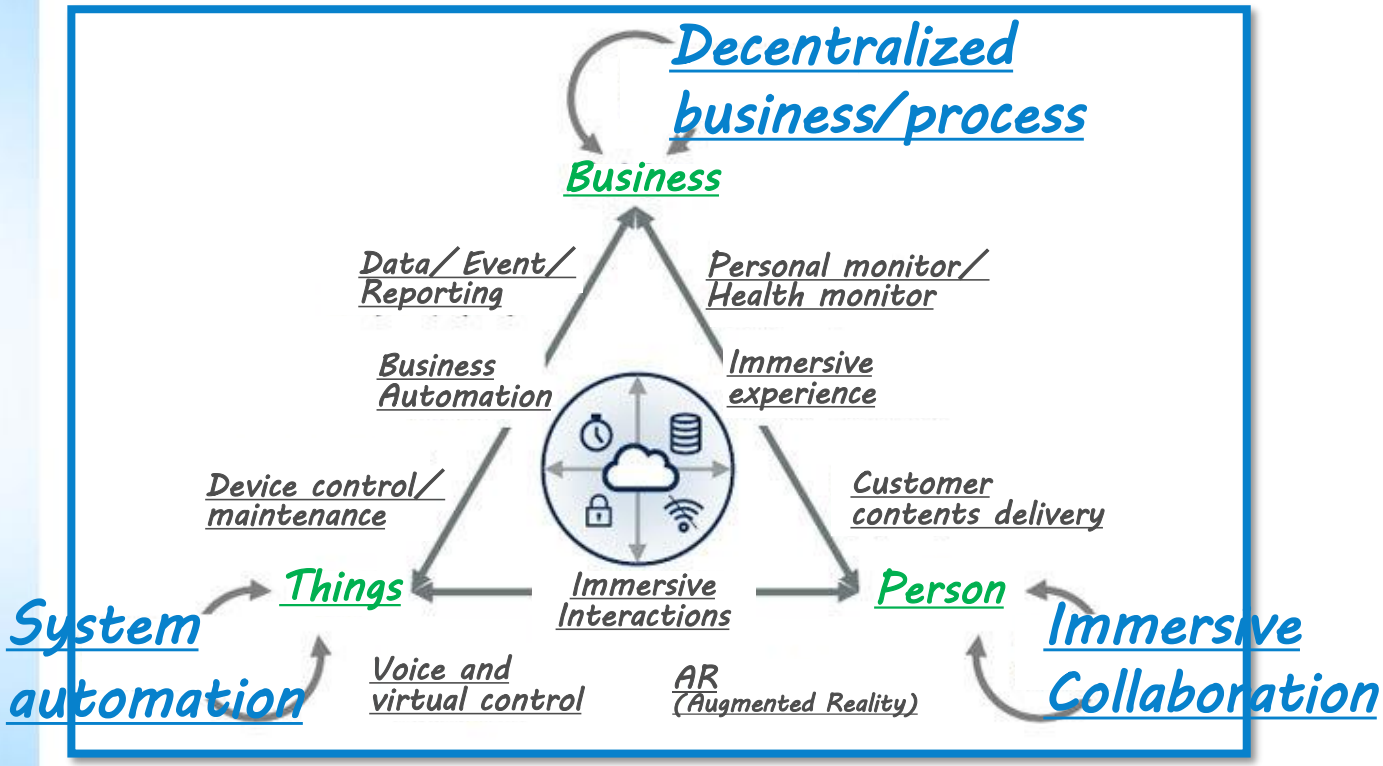


Note) Edge computing and IoT are different concepts!

Edge computing is expected to be active when "real-time judgment" is required near the edge

- **Case1): Automatic car operation**
  - Detects a person from the onboard camera and automatically stops immediately
- **Case2): Failure diagnosis/detection**
  - Detect abnormalities on the production line of the factory and thin out on the spot
- **Case3): Speech conversion such as automatic translation**
  - Translate and convert speech in real time

Edge computing use cases are unknown and 12 business-based categories proposed by Gartner can be helpful



"Twelve Categories in Edge Computing Use Cases"
[Source]: Gartner

~ Agenda ~

1. What is Edge-Computing?
2. AI Chips
3. Edge AI products

The performance of the AI chip is directly linked to the competitiveness of business

- **What's AI Chip?**
  - Semiconductors specialized in AI computation
  - Sometimes called "AI Accelerator Chip"
  - The idea of turning AI functions into hardware

- **Toward No.1 growth Semiconductors!**
  - The top semiconductor is being replaced by for AI from for automotive
  - AI chip market will be expected to double even more in the five years from 2020 to 2025!

AI chips are roughly classified into two types, learning and inference

- **[A] For learning use:**
  - For server devices used on the cloud side
  - For processing that requires extremely high computing power, such as DL learning
  - Providing inference by trained AI to many users

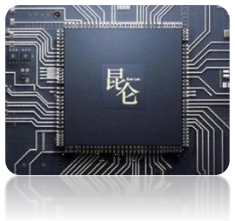
- **[B] For inference use:**
  - For embedded devices on the edge side
  - For inference processing in real time
  - Need to run with less power consumption

Leading companies in the world are developing own AI chips (1/2)

- **NVIDIA:** Tesla, DGX, Jetson series
- **Google:** "TPU (Tensor Processing Unit)"
- **Intel:** "NNP-L1000 (codename: Spring Crest)"
- **Microsoft:** Project Brainwave FPGA based
- **Apple:** "Neual Engine" A12 Bionic
- **Facebook:** AI chip specialized in natural language processing
- **Amazon:** "AWS Inferentia" (acquired Annapurna Labs technology)
- **Tesla:** New chip for autonomous driving (shown as "World's Best")
- **IBM:** From Watson to AI chip development!

Leading companies in the world are developing own AI chips (2/2)

- **China jumps into AI powers (BATIS):**
  - Baidu: "Kunlun" [Autonomous driving vehicle]
  - Alibaba: "Ali NPU" [Smart City]
  - Tencent: [Medical]
  - iFlytek: World Top Level in Speech Recognition [Speech Recognition]
  - SenseTime: [Face Recognition]



- **Other than BATIS:**
  - Huawei (Huawei): "Kirin 980" "Ascend 910" "Ascend 310"
  - DeePhi Tech: Acquired "DeePhi DPU" by Xilinx (Dylinks) (2018/07)
  - Bitmain / Cambricon: AI chip for bitcoin mining
  - Horizon Robotics

R&D of AI chips are also actively conducted in Japan

- **Famous companies:**
  - PFN (Preferred Networks): "MN-Core" (Dec. 2018)
  - Fujitsu: "DLU (Deep Learning Unit)"
  - Denso's semiconductor subsidiary NSITEXE:
    - Automotive dedicated AI chip DFP (Data Flow Processor)

- **Start-up companies:**
  - AISing (Venue from Iwate University): "AiiR Help" (Jan. 2019)
  - Idein
  - ABEJA

Movement to support AI chip development business in Japan

- **[NEDO]: Innovation promotion business to accelerate the development of AI chips**
  - Project period: 2018~22 (5years)
  - FY2019 budget: ¥1.68 billion

【NEDO】

- The development of AI chips and the like requires advanced skills and expensive design tools
- Especially for small and medium-sized companies and venture companies, they have high hurdles for new entrants, even though they have innovative ideas
- Therefore, we will conduct a business to support design and development to make the ideas of small and medium-sized venture companies practical

Consortium "SCAiLE" promoting the advancement and practical application of edge AI

【SCAiLE】

- **SCAiLE:**
  - SCalable AI for Learning at the Edge
  - Established in April 2019

- **Joining four companies (Japan and US):**
  - 1) Crossbar (US):
    "Resistive RAM" technology that enables extremely fast searches
  - 2) Gyrfalcon Technology (US):
    "AI accelerator" technology that speeds up AI processing
  - 3) mtes Neural Networks (mtesNN):
    Technologies and platforms to connect edge devices and AI
  - 4) RoboSensing (a group company of mtesNN):
    Neural network related technology

Which company will break out of the AI chip battle and the melee?

- **1. De facto standard:**
  - Joint development with industry-leading user companies
  - Breakthrough in industrial applications
  - User perspective (Apple: Focus on privacy considerations)
- **2. Focusing strategy:**
  - For learning or for inference
  - Processor type (ASIC, FPGA, GPU, CPU)
  - Portfolio of chip type
- **3. Securing technical skills:**
  - Low latency, cost effective
  - Acquisition of high-tech companies

~ Agenda ~

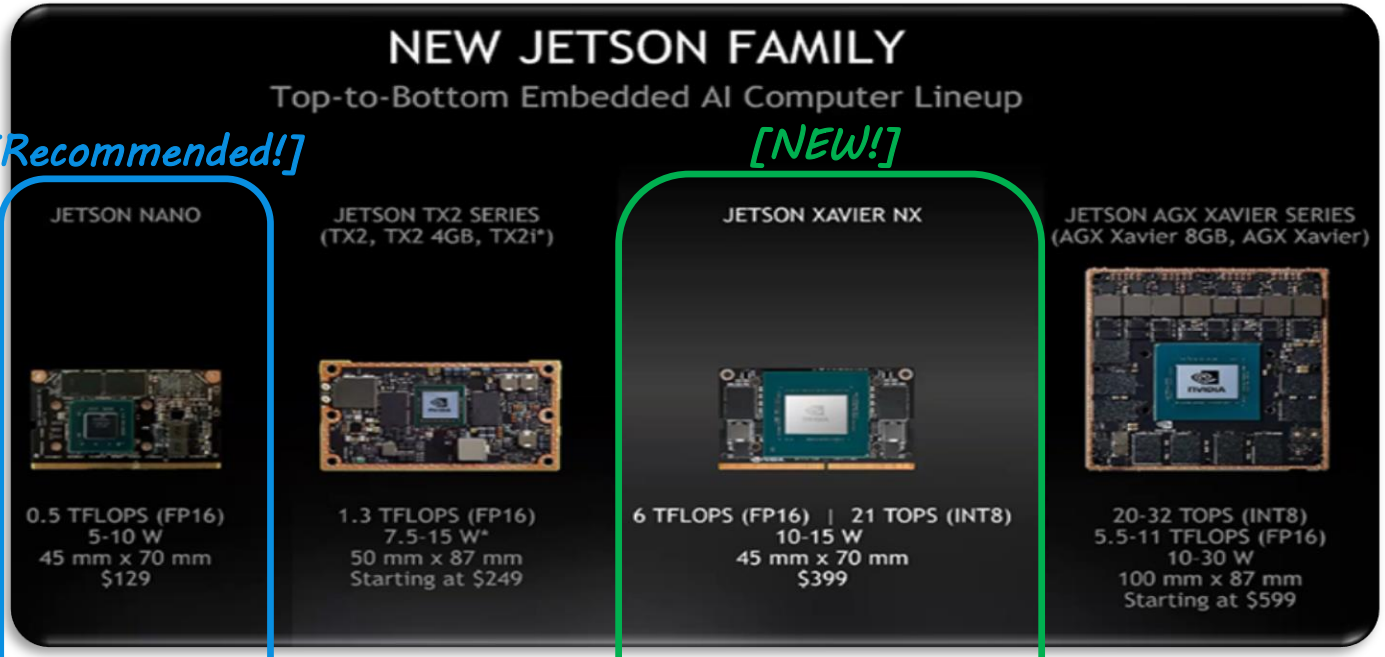
1. What is Edge-Computing?
2. AI Chips
3. Edge AI products

GPU machines can be used for $99!

NVIDIA: Jetson Series

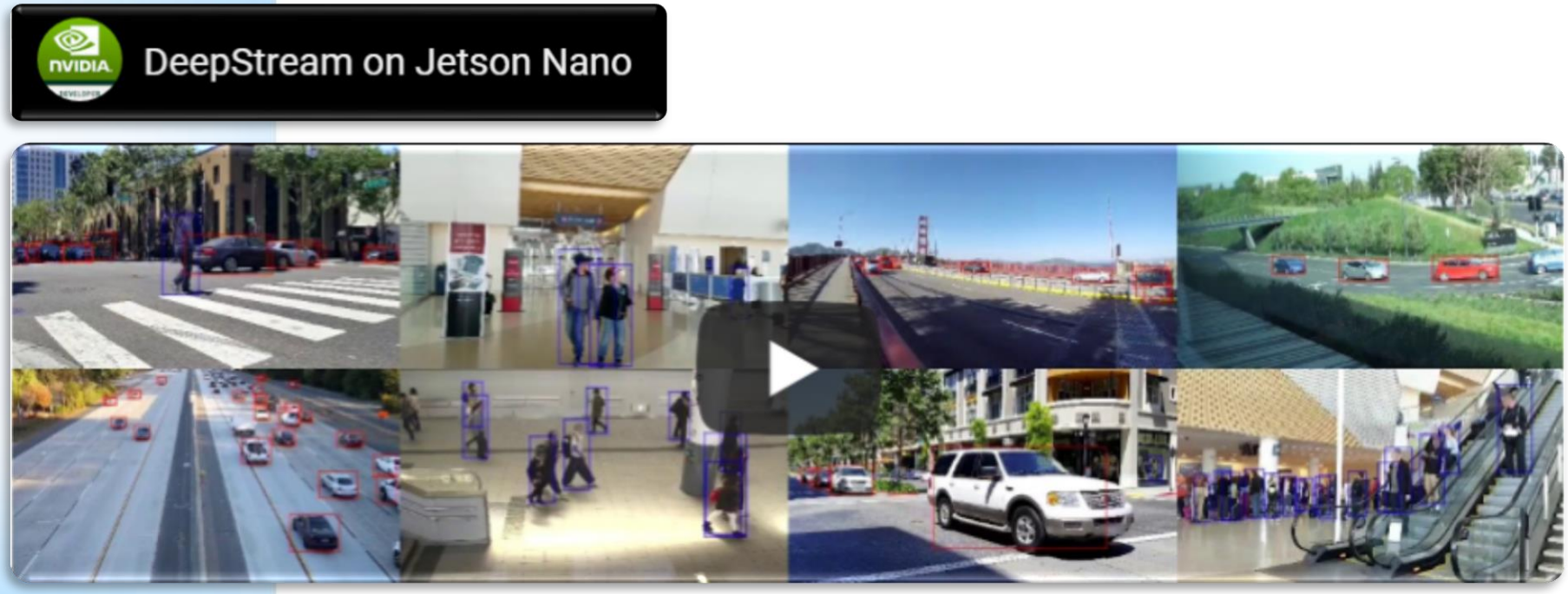
Developer Kit

JetBot



[Recommended!]		[NEW!]	
Nano	TX2	XAVIERNX	XAVIER AGX
$99	$599	$399	$1,299
2019/4	2017y	2020y	2018y

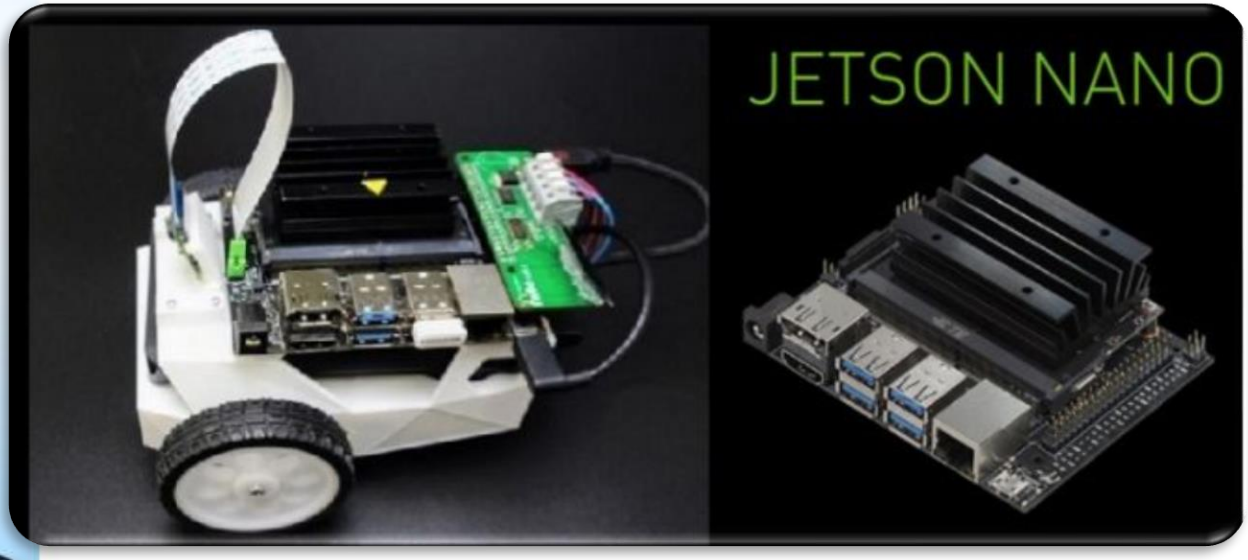
Jetson Nano enables parallel real-time processing of up to 8 streams



【DeepStream on Jetson Nano】

Jetson Nano is also used for AI learning materials!

Utsunomiya University develops with NVIDIA and FaBo, the world's first practical introduction to class!



【Utsunomiya University develops with NVIDIA and FaBo】

Five AI Trends to Watch in 2019
(As of 2019/01/01)

1. The rise of AI chips

2. Fusion of IoT and AI at the edge

3. ONNX is the key to interoperability

4. ML automation(AutoML) evolves

5. Automation by "AIOps" advances

【Forbes Japan】

End of Document