TODAY WE’LL LOOK FROM SEMICONDUCTOR INDUSTRY POINT OF VIEW—THE CRITICAL INGREDIENTS IN THE EVOLUTION OF **AI AT THE EDGE**

- BACKGROUND: AI SYSTEMS IN ASIA
- DEFINITION OF ‘NETWORK EDGE’
- TRENDS, MARKETS
- ACCELERATING INTELLIGENT VISION AT THE EDGE
- CHALLENGES AHEAD
NOVEMBER 2015 | THIRD ARROW

Artificial Intelligence—Can Japan Lead the Way?
By Richard Jokey

There are plenty of examples of sci-fi writers imagining products that eventually become commonplace. Think video calling, touch-screen computers, and earbud headphones, the latter imagined by Ray Bradbury in his 1950s classic novel Fahrenheit 451.

On the face of it, the sci-fi genre seems to be a good divider of things to come, as well as being a popular kind in literature. And we can imagine dedicated research and development departments of companies and governmental divisions reading the latest sci-fi in an effort to peer into the future and outmaneuver rivals.

Given recent concerns about the economy sometimes given the name sometimes given the country itself, the economy wouldn’t be the ideal place to be reading department of its economic脖子.

It’s clear the economy needs a boost. On industrial production a strong south korean economy has been showing a consecutive quarter of negative growth. Moreover, Abenomics is struggling, with the third attempt to reanimate the economy seeing very limited growth.

Indeed, a sci-fi ministry in Japan has been proposed, and the government sees technology as a key area to push forward.

Artificial Intelligence in India—Opportunities, Risks, and Future Potential
LAST UPDATED ON SEPTEMBER 16, 2018 BY RACHAV BHARADWAJ

South Korean Government Announces Nearly $1 Billion in AI Funding
The government plans to support the establishment of a high-profile research center that will serve AI

Boosting Al In South Korea

Yesterday, the South Korean government announced its plan to boost the artificial intelligence set of matches between Lee and

China announces goal of AI leadership by 2030
July 21, 2017 by Joe Mcdonald

China's government has announced a goal of becoming a global leader in artificial intelligence in just over a decade, putting political muscle behind growing investment by Chinese companies in developing self-driving cars and other advances.

Communist leaders see AI as key to making China an "economic powerhouse" and as a "core technology" for major breakthroughs by 2025 and beyond.
SOME BACKGROUND IN COMPUTER VISION AND WHAT WE ARE DOING AT INTEL TO ENABLE NEW AI SYSTEMS AT THE EDGE
COMPUTER VISION THEN AND NOW

- Heuristic Algorithms
- Human Engineered
- Server Based
- AI / Deep Learning
- Real-time, Higher Accuracy
- Edge Based
EDGE-TO-CLOUD: DATA IS KEY DRIVER

By 2019, 45% of data will be stored, analyzed, and acted on at the edge.

**Drivers for Edge**
- Latency
- Bandwidth
- Security
- Connectivity

**Devices & Things**
- Transportation • Retail
- Public Sector • Logistics • Smart Cities
- Video • Healthcare • Manufacturing
- Smart Buildings • Energy

**Network Hub or Regional Data Center**

**Core Network**

**Cloud Data Center**

Intel IoT Group
EDGE TO CLOUD: MOVING DATA CRITICALLY IMPACTS POWER

END POINT DEVICE

MEMORY
IMAGE SENSOR(S)
VISION PROCESSING
WIRELESS TRANSCEIVER

5~20 W
MEMORY
IMAGE SENSOR(S)
APPLICATION PROCESSOR
WIRELESS TRANSCEIVER
4G/LTE MODEM

EDGE COMPUTE NODE / EDGE SERVER

< 2 W

1000s W

CLOUD

EDGETO CLOUD: MOVING DATA CRITICALLY IMPACTS POWER

END POINT
DEVICE

MEMORY
IMAGE SENSOR(S)
VISION PROCESSING
WIRELESS TRANSCEIVER

5~20 W
MEMORY
IMAGE SENSOR(S)
APPLICATION PROCESSOR
WIRELESS TRANSCEIVER
4G/LTE MODEM

EDGE COMPUTE NODE / EDGE SERVER

< 2 W

1000s W

CLOUD

Intel IoT Group
THERE IS A GROWING OPPORTUNITY FOR **AI AT THE EDGE**

- **25%-35% CAGR** in **AI** for silicon industry
- **Growth larger at the Edge**, compared to Data center

Note: Endpoint and Edge excludes smartphones and client PCs
Source: Intel, IDC, Gartner
THE VIRTUOUS CYCLE DRIVING INNOVATION IN AI AT THE EDGE

EDGE PROCESSING REQUIREMENTS
• Huge Compute Requirements
• Low Latency
• Power Efficiency
• Local Data / Privacy

CHIP TECHNOLOGY IMPROVEMENTS
• Efficient Data Flow
• Better Algorithm Processing
• Efficient Memory Use
• Silicon Process Technology
EDGE AI CAPABILITIES ENABLE AUTONOMOUS FEATURES IN MANY APPLICATIONS

CITIES • STATE • FEDERAL
Public Safety & Surveillance
Traffic, Parking and LPR
Emergency Response

FINANCE • BANKING
People Counting Customer
(i.e. Gender, Wait Time)
ATM Facial Recognition

INDUSTRIAL
Machine Vision Asset Inspection
(i.e. Pipeline)
Augmented Reality

CASINO GAMING
Public Safety & Surveillance
Facial Recognition

TRANSPORTATION
Autonomous Vehicles
Public Safety (i.e. Bus/Rail)
Traffic & People Counting

HOME • RETAIL • SURVEILLANCE
Security & Surveillance
Responsive Retail Advertising
Digital Home Assistant

ROBOTICS
Manufacturing Automation
Industrial (i.e. Pipeline Welding)

DRONES
Emergency Response
Asset Inspection (i.e. Windmill)
INDUSTRIAL EXAMPLE: SMART FACTORY

Smart Factory:
Will Generate an Estimate of 1PB data/day by 2020
In smart factory configuration, fast inspection data is actionable. Vision+DL in industrial PC feeds into edge server.

Local Dashboard/HMI Software

Gateway (Machine Data)
- 3rd Party Agents
- Intel® Data/Stream Manager
- Local OS (e.g. AndroidOS)

Scalable Edge Server
- Intel® OpenVINO™
- 3rd Party Applications
- Intel® Data/Stream Manager
- Clear Linux (built for IoT use)

Cloud-Based Dashboard

Industrial PC (Vision/DL)
- Intel® Algorithms
- Intel® OpenVINO™
- 3rd Party Agents
- Intel® Data/Stream Manager
- Local OS (e.g. AndroidOS)
RETAIL EXAMPLE: SMART STORES

Retail Store:

*Huge Data Processing Requirements*

**Things & Edge Node**
- Item Tracking
- Mobile Payment
- Store Analytics
- Inventory Management
- Customer Tracking

**Edge Network & Core Network**

**Cloud**
NEW INNOVATIVE CHIP ARCHITECTURES ENABLE COMPUTE-EFFICIENT DEEP LEARNING INFERENCE AT THE EDGE WITH MORE PERFORMANCE/WATT/$
AI AT THE EDGE: NOT JUST A TECHNICAL PROBLEM...

HOW TO **SCALE** TO DEVELOPERS WORLDWIDE?*

*HINT: OPPORTUNITIES IN ASIA
TO MAKE ENHANCED DEEP LEARNING AT THE EDGE MORE ACCESSIBLE, WE JUST ANNOUNCED: INTEL® VISION ACCELERATOR DESIGNS

<table>
<thead>
<tr>
<th>EXAMPLE CARD BASED ON VISION ACCELERATOR DESIGNS</th>
<th>INTERFACE</th>
<th>CURRENTLY MANUFACTURED BY*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intel® Vision Accelerator Design with Intel® Movidius™ VPU</td>
<td>M.2, Key E</td>
<td>Intel® Rugged Systems, Advantech, AAEON, uYZI</td>
</tr>
<tr>
<td>Intel® Vision Accelerator Design with Intel® Arria® 10 FPGA</td>
<td>miniPCIe</td>
<td>IEI, Advantech, AAEON, iECOM</td>
</tr>
<tr>
<td></td>
<td>PCIe x4</td>
<td>IEI, Advantech, AAEON, uYZI</td>
</tr>
<tr>
<td></td>
<td>PCIe x4</td>
<td>IEI, Advantech, AAEON, uYZI</td>
</tr>
</tbody>
</table>

- 1 Movidius MA2485 VPU
- 2 Movidius MA2485 VPUs
- 8 Movidius MA2485 VPUs
- Intel® Arria® 10 FPGA 1150GX/660GX

SOFTWARE TOOLS
- OpenVINO™ Toolkit
- Develop NN Model; Deploy across Intel® CPU, GPU, VPU, FPGA; Leverage common algorithms

*Please contact Intel representative for complete list of ODM manufacturers. Other names and brands may be claimed as the property of others.
エッジコンピューティングとディープラーニング

IoTの普及に伴い各種センサーが生み出すデータ量は、通信インフラやストレージインフラの進化を超える勢いで増え続けており、また、IoTサービスの質の面からも処理の即時性が求められる中、センサーデータの収集・分析にクラウドを使わないデータの地産地消を可能にするエッジコンピューティングが注目されています。

当社のディープラーニング推論アクセラレータボードは、既存のエッジサーバーやゲートウェイにアドオンする事で、デ
AI Core – Artificial Intelligence On The Edge

2018-03-06

The first embedded ultra-compact Artificial Intelligence processing card for on the edge computing

UP Bridge the Gap – a brand of AAEON Europe – is proud to launch AI Core: the first embedded ultra-compact Artificial Intelligence processing card for on the edge computing.
OTHER EXAMPLE PARTNER ODM COMPANIES FOR INTEL® VISION ACCELERATOR DESIGNS

- AAEON
- an ASUS assoc. co.
- NEXCOM
- ADLINK TECHNOLOGY INC.
- iEi
- ADVANTECH
MANY OTHER PARTNERS TO BRING SCALE TO AI AT THE EDGE

INTEL® AI: IN PRODUCTION

A LEADING ECOSYSTEM FOR AI AT THE EDGE DEPLOYMENTS

*Partners as of October 10, 2018

EXAMPLE PARTNERSHIP: ALIBABA (CHINA)

Intel and Alibaba Group are:
• Launching of a Joint Edge Computing Platform to accelerate edge computing development
• Establishing the Apsara Stack Industry Alliance targeting on-premises enterprise cloud environments
• Deploying latest Intel technology in Alibaba to prepare for the 11/11 shopping festival
• Bringing volumetric content to the Olympic Games Tokyo 2020 via OBS Cloud
• Accelerating the commercialization intelligent roads

SILICON INDUSTRY HAS CHALLENGES TO OVERCOME TO MAKE AUTONOMY AT THE EDGE POSSIBLE

**Deep Neural Network Performance**

*Inferences Per Second for ResNet50, Batch Size = 1

**AI Performance**
For power held at 5~10 W

---

**SEMI Industry Progress Goals:**

**Reduced Precision Neural Networks**
Improved performance moving from fp16 to binary weights, for example

**Neural Network Compression / Sparsity**
Reducing compute requirement for NN structure, and taking advantage of zeros in matrix computation

**Efficient Memory**
Power efficient memory access with improved bandwidth

**Accelerators**
Balancing fixed function performance with flexibility/programmability

**Compiler Innovation**
Tools making best use of new hardware
THE FUTURE IS IN OUR HANDS