GPU: The Catalyst and Accelerator of AI Systems

Chwee Chua
AVP Analytics, Big Data and Cognitive Computing

Sep 2016
The New Industrial Revolution

First Industrial Revolution 1760
- **Turning point in History**
  - Technology innovations
  - Machines transformed all forms of manufacturing
- **Results:**
  - Machine production instead of slow hand production
  - Living standards of the masses of ordinary people increased

Today
- Intelligent Insights/Automation in Highly Complex Processes
- **Results:**
  - Machine decision assistance with increasing precision & accuracy
  - *Solving problems faster than humanly possible*
Digital Transformation: Disrupt or be Disrupted

By 2018:
- One-third of top 20 market share leaders will be significantly disrupted by new competitors.
- Incumbents would also seek to reinvent themselves.

3rd Platform:
- Create New Services
- Evolve Business Models
Cognitive/AI Computing Attributes

- Performs deep natural language processing and analysis
- Conducts learning in real time as data arrives
- Predicts and recommends outcomes
- Scoring with evidence
- Cycle back to the start to make system better over time
IDC FutureScape Prediction:

By 2020, 40% of All Business Analytics Software Will Incorporate Prescriptive Analytics Built On Cognitive Computing Functionality

80% of AP organizations are currently in the ad-hoc and opportunistic phases

Source: IDC FutureScape Worldwide Big Data and Analytics 2016 Predictions – APEJ Implications (IDC #AP40492015)
Cognitive/AI Drivers:

Paradigm Shift:

• Storage & Cloud
• New Frameworks/Libraries
• Speed in Parallel Processing
Data is Fuel to Cognitive/AI solutions

- On average **57.2%** of data from multiple sources are analyzed by companies in China.

- Rich media (e.g. video, audio, images) and social network data are integrated by over **40%** of the companies.

"Data doubling every two years"  
- IDC

"Data doubling every two years"  
- IDC

2 years
Hardware & Software Accelerators

- **Cognitive/AI Solution**
  - Training from data
  - Inference to real-time decision making

**API**

- **Hadoop**
- **CUDA**
- **Cloud**
- **R**

**Data** → **Empower** → **Business Process**

*GPU Accelerates*
- Training from data
- Inference to real-time decision making
Leading in Competitive Advantage

Next 12 – 24 months:
• Companies in China have prioritized Cognitive/AI solutions.

• Cognitive/AI Computing
  • 34% in China (21.8% APeJ)

• Real time decision support
  • 33% in China (40.5% APeJ)
Why is Cognitive/AI Computing Important?

Data-intensive industries seek new business opportunities

- Many industries are acquiring vast data about their customers, operations and markets.

Cognitive/AI Use Cases:

- Customer Service
- Maintenance
- Operations
- Banking
- Sports
- Research
Cognitive/AI in Public Cloud

- In APeJ, **37%** of companies are already running analytical tools on public cloud.

- A further **28%** have plans or are evaluating migration over to public cloud.

- Cloud based services (e.g. IBM Watson, HPE Haven, Microsoft Azure Machine Learning)

- Common Services include: Sentiment Analysis, Object Recognition, Face & Speech Recognition
Cognitive/AI in Automated Devices

By 2020, **80%** of Big Data & Analytics deployment will be Micro Analytics & Data Manipulation

**Intelligent Video Analytics (IVA)**
- Growing importance in public areas/services
- Provides 24x7 monitoring
- Conducts inspections at dangerous hard to reach sites

**Connected Cars**
- China to have **16.2%** share of advanced driver assistance program by 2019.
- Intelligence is needed when vehicle is part of the traffic ecosystem.

**Robotics**
- More than **25%** of robotics exports are from China
- Driven by increasing labor costs, shortage of skill workers
Cognitive/AI in Internet+

Finance
- Within **6 to 12 months** digital services and offerings can easily be replicated.
- Personalized experience is **KEY** to customer retention.

Healthcare
- It took just **10 minutes** for IBM Watson to diagnose a rare leukemia which baffled doctors for months.
- Perform large scale diagnostics with access to medical research data
Cognitive/AI is an Iterative Journey

- Cognitive/AI model adapts with changes in the underlying data.
- Periodic retraining and update of inference model is essential.
- Training REMAINS the most time-consuming component in Cognitive/AI System.
AI Success Story (*thus far*)

Medical Research (*Lung Cancer*)
- Pathologists agree only about **60%** of the time.
- Stanford University’s cognitive/AI program performed **15%** better on prognosis than pathologist

Speech to Text (*English/Mandarin*)
- Deep Speech 2 by Baidu **3x** faster than typing on mobile phone
- Significant **lower error rate** compared to humans (English – 20.4%, Chinese – 63.4%)
GPU’s role in Cognitive/AI System

• Provides 2x to 1,000x improvement from CPU-only solutions

• Software Frameworks (CUDA, cuDNN) eases development

• 1st choice at improving system’s performance

• Silent worker behind background
# Cognitive/AI Accelerator Options

<table>
<thead>
<tr>
<th>Development Focus</th>
<th>GPU</th>
<th>FPGA</th>
<th>ASIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Development Focus</td>
<td>Software</td>
<td>Software/Hardware</td>
<td>Hardware</td>
</tr>
<tr>
<td>Acceleration Domain</td>
<td>Parallel Matrix Computation</td>
<td>Logic Blocks on Hardware</td>
<td>Signal Transmission on Circuit</td>
</tr>
</tbody>
</table>

- **GPU**
  - Acceleration Domain: Parallel Matrix Computation
  - Power Efficiency: High
  - Accessibility: Industrial
  - Hardware Design Specifications: Vendor (Non-customizable)
  - Usage Quantity Factor: Linear scalability

- **FPGA**
  - Acceleration Domain: Logic Blocks on Hardware
  - Power Efficiency: High
  - Accessibility: Open
  - Hardware Design Specifications: Vendor (Customized through programming)
  - Usage Quantity Factor: Low - medium volume usage for optimal results. Large volume production necessary for cost effectiveness.

- **ASIC**
  - Acceleration Domain: Signal Transmission on Circuit
  - Power Efficiency: Very High
  - Accessibility: Industrial
  - Hardware Design Specifications: 100% Customizable
  - Usage Quantity Factor: None (upon production)

- **Ease of Development**
  - GPU: High
  - FPGA: Medium
  - ASIC: Low

- **Update Flexibility**
  - GPU: Yes
  - FPGA: Yes
  - ASIC: No
Trends: Golden Age of AI?

Funding for AI Start-Ups to exceed **US$1B** in 2016.

Recent AI acquisitions:

- **Amazon**
  - Orbeus
- **Apple**
  - Turi
  - Perceptio
  - Emotient
  - Gliimpse
- **Google**
  - DeepMind
  - Moodstocks
- **Microsoft**
  - Genee
  - Wand
  - Swiftkey
Opportunities

GPU benefits are *cross industry* and *cross research* domain

Cognitive/AI Enablement Technology:
- **H/W**: Footprint, Processing Power
- **S/W**: cuDNN, GPU Inference Engine

High impact topics:
- Next Generation Analytics & Computing
- Smart Cities: Public Safety & Security
- Upgrade Cycles
Guidance

• Plan GPU adoption against use cases, data, and skills

• Be conscious of performance vs power consumption

• Identify data sources and algorithms
Future Outlook

• NVIDIA is the undisputed leader of Cognitive/AI innovation

accelerator

• Catalyst for the Golden Age of AI

• GPU usage in Cognitive/AI is most matured compared to other solutions

• Talent shortage remains a major challenge
Just as how the abacus augmented human intelligence, GPUs will augment artificial intelligence and thereby revolutionize how we innovate.

Chwee Chua
ckchua@idc.com