HPE is simplifying AI and deep learning analytics across all organizations

2018 May GTC
Taipei
HPE HIT AI/HPC
沈仲杰 robert.sheen@hpe.com
+886-955766078
HISTORY OF AI
AI has been around for more than 50 years; for as long as computing has been around--
Artificial intelligence (人工智慧)
- Any techniques that makes computers simulate human behavior or intelligence 使用電腦模擬人類行為和智慧的科學

Machine learning (機器學習)
- A subset of AI where machines can learn without explicit programming 機器學習是人工智慧的一種方法, 摒除傳統編寫條列式軟體, 尋找用‘計算統計’的方式從大量的資料中找出可預測結果的方法. 資料科學家是整體的重心

Deep learning (深度學習)
- A subset of machine learning dealing with algorithms inspired by the structure and function of the brain called artificial neural networks 深度學習專注於採用類神經網路的架構從大量資料中找出抽象特徵質方法的研究與開發
“every click, voice, swipe, share, like – is being used to make decisions about our future”

2.5 Quintillion bytes of data per day

Every 3 years the volume of data doubles
21% of Americans are constantly online

90% of the data has been collected in last 2 years
SO, WHY IS IT THAT EVERYONE IS TALKING ABOUT IT, NOW?

01 DATA (大數據)
Enormous amount of data of all types
Critical for Machines to Learn

02 ALGORITHMS
Availability of Open Source Algorithms, Neural Networks, Parallel Processing, Deep Learning

03 COMPUTE POWER
GPU based computing with super compute power
Relatively low cost
BIG DATA

opportunity to find insights in new and emerging types of data with speed, and trust at scale

- **Volume (量)**
  - Data at Scale
  - Terabyte to petabytes of data to be managed

- **Velocity (緩急)**
  - Data in Motion
  - Enable decisions within fractions of a second in real time

- **Variety(異同)**
  - Data in many Forms
  - New insights are found when analyzing heterogeneous data types

- **Veracity(真假)**
  - Data in Doubt
  - Manage reliability & predictability of inherently imprecise data.
### Artificial Intelligence – Horizontal use cases

<table>
<thead>
<tr>
<th>Images (像)</th>
<th>Text(字)</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Image analysis (e.g. digital pathology such as tissue classification (cancerous vs. non-cancerous), facial recognition, vehicle number plate recognition)</td>
<td>- Sentiment analysis to automatically understand the perception of customers towards a product or service based on their comments</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Video(動影)</th>
<th>Sensor(感測器)</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Video analytics (e.g. security, surveillance, traffic flow, detect car accidents, monitor pedestrian congestion in public spaces, manufacturing)</td>
<td>- Predictive maintenance – Analyze data that includes a timestamp, a set of sensor readings, and device identifiers. Identify anomalies.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Data(資料)</th>
<th>Security(安全)</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Analytics such as optimization, simulation and most importantly predictive analytics, fraud analysis, preemptive therapy, CRM campaigns</td>
<td>- Advanced threat protection, threat ID and behavior anomalies, correlation and predictive analytics on sensor, image, video data</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Speech(話)</th>
<th>Other(其他)</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Speech recognition (e.g. voice dialing, call routing, search, simple data entry, chat bots for diagnosis and advanced assistance)</td>
<td>- Identification of items, events or observations which do not conform to an expected pattern or other items in a dataset (e.g. fraud detection)</td>
</tr>
</tbody>
</table>
AI CookBook
Select optimal hardware/software configurations for AI Problems
With HPE Deep Learning Cookbook

“Book of recipes” for deep learning workloads

Expert advice to get you started

Availability of complete toolset

- **Comprehensive tool set** based on extensive benchmarking
- **Includes** 11 workloads with 8 DL frameworks and 8 HPE hardware systems
- **Estimates workload performance** and recommends an optimal HW/SW stack for that workload

- **Informed decision making** Optimal hardware and software configurations
- **Eliminates the “guesswork”** Validated methodology and data
- **Improves efficiency** Detects bottlenecks in deep learning workloads

- Deep Learning Benchmarking Suite available on GitHub Dec 2017
- Cookbook Audio Overview: listen to the audio replay
- Learn More – HPE Deep Learning Cookbook Sales Play
- Link to more documentation
- https://psnow.ext.hpe.com/doc?id=a0028521enw

<<Back to Interaction Model
Is deep learning a ‘universal algorithm’?
Is deep learning a ‘universal algorithm’?
One size does NOT fit all

Model (topology of artificial neural network):
- How many layers
- How many neurons per layer
- Connections between neurons (types of layers)
A quick introduction to (Deep) Neural Networks - Training

To solve useful problems we have to connect multiple neurons together. The output from a neuron in one layer becomes the input to neurons in the next layer.

**Computationally intensive**
- Matrix-matrix multiplies
- Vector operations
- FFTs / Convolutions

Training deep networks requires high performance computing hardware and techniques, especially GPU.
Types of artificial neural networks

Topology to fit data characteristics

**Convolutional:**
- Images

**Fully connected:**
- Speech, text, sensor

**Recurrent:**
- Speech, text, sensor
HPE Deep Learning Cookbook Overview

Comprehensive set of tools to guide the choice of the best hardware and software environment for different deep learning workloads

— Eliminate the “guesswork” with Deep Learning Performance Guide: tap into a massive pool of performance data to reason on optimal hardware configuration for your workload

— Validate the configuration of your hardware and software environment with Deep Learning Benchmarking Suite

— Get started fast with one of our Reference Designs as a default technology recipe

https://developer.hpe.com/platform/deep-learning-cookbook/home
HPE Deep Learning Cookbook

A set of tools to guide the choice of the best hardware/software environment for a given deep learning workload.
HPE Deep Learning CookBook

HPE Deep Learning Performance Guide

HPE Deep Learning Benchmark Suite

HPE Deep Learning github

HPE Deep Learning Benchmark Suite

HPE Deep Learning Webinar
HPE Deep Learning Benchmarking Suite

GitHub: https://github.com/HewlettPackard/dlcookbook-dlbs
Documentation: https://hewlettpackard.github.io/dlcookbook-dlbs
How to use HPE Deep Learning Cookbook

1. To help customers define an optimal hardware/software stack for their workloads
   - Identify properties of customer workloads
     - Data type (images, video, audio, text, time series/sensor data)
     - Type of task (detection, classification, content generation, anomaly detection)
     - Volume of available training data
   - Use Reference Configurations as a default option, if a reference configuration is available for a similar workload
     - Image Classification Reference Configuration is available, was identified with help of Deep Learning Cookbook
     - Natural Language Processing, Speech Recognition, Video Analysis and others to come later
   - Use Performance Guide tool (to be publicly available starting March 30, 2018) to input properties of customer workloads and get recommended hardware/software stack
   - Use Performance Guide tool (to be publicly available starting March 30, 2018) to explore how changes in hardware/software configuration influence performance

2. To validate and tune hardware/software configuration
   - Suggest customers to use Benchmarking Suite (available on GitHub) to collect performance measurements on their systems, compare with performance data collected by HPE and available in the Performance Guide, and tune their system configuration
APAC case study
Recommend Startup Environment – Reference Architecture
HPE AI solution from Core to Edge
Data Flow in Edge to Cloud

Stage 1: Sensors / Actuation
- Data is sensed, things controlled

Stage 2: Access / Control
- Aggregated, digitized, pre-processed

Stage 3: Edge IT
- Early analytics and compute

Stage 4: Cloud / Data Center
- Deep analytics, compute, Archive

Data Flow

Control Flow
7 reasons to **shift IT, and compute at the edge**, and **NOT** send the data to the cloud:

- Latency
- Bandwidth
- Cost
- Threats
- Duplication
- Reliability
- Compliance
HPE offers an end-to-end Edge to Cloud Portfolio (HW and SW) Compute (Edgeline, Apollo), Connectivity (Aruba)
HPE has built Converged Appliances for Video Mgmt & AI Analytics

**VIDEO MANAGEMENT TIER**
- Live Video Streams
- Camera x N
- HPE ProLiant m510 / HPE ProLiant m710x
- Video Recording, Indexing, Retrieval SW

**VIDEO ANALYTICS TIER**
- Video Analytics Server
- HPE ProLiant m510
- NVIDIA Tesla P4 x3
- Video Decoding
- Inference Accel.

**Processing Server**
- Database
- 

**Thin Client**
- Video Analytics Client

**Storage System (JBOD)**
- HPE 3PAR Storage
- Cartridge
- Cables

**HPE Edgeline EL4000**
- Video Management Client

**HPE has built Converged Appliances for Video Mgmt & AI Analytics**

**HPE 3PAR Storage**

**Video Streams**
- (Live or Recorded)

**HBA Card**
- x3

**HPE Confidential 26**
HPE Apollo6500 Gen10 and AI Platform
Accelerated performance for GPU intensive workloads
Reliable deep learning platform with industry leading GPUs and interconnects

HPE’s highest GPUs per node
- Up to 125 Tflops single precision performance
- Eight GPUs per node

Powerful host
- High speed / low latency network
- NVMe drives
- High speed DDR4 SmartMemory

Dependable system
- System designed around 350W accelerators
- Reliable due to signal integrity consistency

Leading accelerator technology
- NVLink 2.0 enables dedicated GPU-to-GPU communication
Resilient, secure and simple for **lower TCO**
Designed for ease of use by enterprises

**Easy to service, upgrade**
- Easy access modular design
- Rear cabled fabrics

**Efficient management, security**
- Save time and cost with HPE iLO5
- World’s most secure industry standard server using HPE iLO5

**Resilient power, cooling**
- 2+2 power redundancy
- Hot-plug fans

**Standard rack mount**
- Designed for 1m deep rack
- No special flooring, plumbing needs
HPE has a comprehensive, purpose-built portfolio for deep learning

### HPE SGI 8600
- Petaflop scale for deep learning and HPC

### HPE Apollo 6500 Gen10
- Enterprise platform for accelerated computing

### HPE Apollo sx40
- Maximize GPU capacity and performance with lower TCO

### HPE Apollo 2000 Gen10
- The bridge to enterprise scale-out architecture

### HPE Edgeline EL4000
- Unprecedented deep edge compute and high capacity storage; open standards

<table>
<thead>
<tr>
<th>Compute ideal for training models in data center</th>
<th>Compute for both training models and inference at edge</th>
<th>Edge analytics and inference engine</th>
</tr>
</thead>
<tbody>
<tr>
<td>HPE SGI 8600</td>
<td>HPE Apollo 6500 Gen10</td>
<td>HPE Edgeline EL4000</td>
</tr>
<tr>
<td>Petaflop scale for deep learning and HPC</td>
<td>Enterprise platform for accelerated computing</td>
<td>Unprecedented deep edge compute and</td>
</tr>
<tr>
<td></td>
<td>Maximize GPU capacity and performance with lower TCO</td>
<td>high capacity storage; open standards</td>
</tr>
</tbody>
</table>

### AI software framework

**Easy Setup and Flexible OS**

- Using Bright Computing’s distribution of deep learning software development components and workload management tool integration

### AI storage

**HPE Apollo 4510**

- Large-scale, storage virtualization & tiered data management platform

### Choice of fabrics

- Intel® Omni-Path Architecture
- Mellanox InfiniBand
- HPE FlexFabric Network
Thanks!