Accelerating time to value is priority #1

What do you do with zettabytes of data?

- **Hybrid IT**: Built-in data analysis & contextually aware
- **Your Apps & Data**: Containerized, automated and orchestrated
- **Intelligent Edge**: Beacons, sensors and geo-positioning

- **Flexible consumption on premise and by cloud**: Driven by agile DevOps
- **Ecosystem of innovation partners**: Artificial intelligence, robust machine, deep learning, algorithms
- **Security & resilience built-in**: Mobile users, apps and devices
- **Containerized, automated and orchestrated**: Ubiquitous connectivity
- **Big, rich, trained datasets**: Reliable performance & experience
- **Built-in data analysis & contextually aware**: Adaptive trust security
- **Adaptive trust security**: Ubiquitous connectivity

---

**HPE PointNext**
What is AI

Artificial Intelligence (AI)
A process where a computer solves a task in a way that mimics human behavior. Today, narrow AI—when a machine is trained to do one particular task—is becoming more widely used, from virtual assistants to self-driving cars to automatic tagging your friends in your photos on Facebook.

Machine Learning (ML)
Algorithms that allow computers to learn from examples without being explicitly programmed.

Deep Learning (DL)
A subset of ML which uses deep artificial neural networks as models and does not require feature engineering.
Types of artificial neural networks

Topology to fit data characteristics

**Convolutional (CNN):**
Images

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

**Recurrent (RNN, LSTM):**
Speech, text, sensor
Phases of machine learning

Ingest → Transform → Train → Validate with inference → Inference in production

Training: Ingest → Transform → Train

Inference: Validate with inference → Inference in production

Tools:
- NVIDIA TensorRT (CNTK)
- TensorFlow
- Caffe
- KALDI
- PaddlePaddle
- Theano
- mxnet
- PyTorch
- The Microsoft Cognitive Toolkit (CNTK)
Deep learning ecosystem

Software
- Caffe
- Keras
- theano
- torch
- KALDI
- Caffe2
- CNTK
- mxnet
- TensorFlow
- Chainer

Hardware
- NVIDIA
- Intel
- ARM
- Google
- Xilinx
- Movidius
### Application
**TB (Tuberculosis) Diagnostic** 肺結核診斷

#### Background & Challenge

- Requires professional trainings
  - Inconspicuous difference on X-ray
  - Irregularity of data
- Advancement in Deep Learning
  - Pattern recognition
  - Computer vision
- Better clinical diagnosis and personalized precision medicine

#### Benefit

- Decrease diagnostic time **60x**
- Diagnostic accuracy reach **96%**
- Decrease error diagnostic **25%**

---

### Hospital

<table>
<thead>
<tr>
<th>Patients</th>
<th>GB Data</th>
<th>Ensemble</th>
<th>Throughput*</th>
</tr>
</thead>
<tbody>
<tr>
<td>662</td>
<td>3.5</td>
<td>5</td>
<td>60x</td>
</tr>
</tbody>
</table>

![AUC Score](chart.png) **96%**

<table>
<thead>
<tr>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>AUC</td>
</tr>
<tr>
<td>Specificity</td>
</tr>
<tr>
<td>Sensitivity</td>
</tr>
</tbody>
</table>

---

![Images](images.png) No Tuberculosis vs Tuberculosis
Image Recognition Deep Learning Algorithm introduction
Face Recognition system architecture

1. Front-end image capture with cameras and facial recognition equipment.
2. Transmit captured images to the server.
4. Facial recognition data is stored.
5. VMS (Video Management System) retrieves and processes facial data to present visitor entry routes and time spent.

1. Present visitors in which area or location entered.
2. Present visitors in which area or location left, as well as time spent.
BigData & AI Solution Blocks

Advisory, professional and operational services

Custom Use Case / ISV Solutions

<table>
<thead>
<tr>
<th>AI Solutions Layer</th>
<th>AI Layer - Platform Focus</th>
<th>Data Platform Layer</th>
<th>Data Deployment Layer</th>
</tr>
</thead>
<tbody>
<tr>
<td>MDI Defect Detect</td>
<td>CME Voice Detection</td>
<td>Hadoop Ecosystem</td>
<td>AI and BigData Deployments</td>
</tr>
<tr>
<td>CME Fraud Detection</td>
<td>PUB Smart Transp. Security</td>
<td>Fast Data &amp; Edge Data</td>
<td>HPE Green Lake Consumption Model</td>
</tr>
<tr>
<td>CME Card Transact</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Machine Learning as a service

- Caffe
- torch
- Kaffeine

Automated AI Tools

- DataRobot
- dataiku

HPE Apollo 6500
The enterprise bridge to accelerated computing

HPE Apollo 2000
The bridge to enterprise scale-out architecture

HPE Apollo 4200
optimized for Big Data Analytics and other data storage intensive workloads

HPE DL 380
unprecedented levels of performance for databases and analytic workloads

HPE Green Lake Consumption Model

HPE Apollo 6500 & HPE Apollo 4200 optimized for Big Data Analytics and other data storage intensive workloads

HPE Apollo 2000 & HPE DL 380 unparalleled levels of performance for databases and analytic workloads

AI Layer - Platform Focus

- MDI Defect Detect
- CME Voice Detection
- PUB Smart Transp. Security
- CME Fraud Detection
- CME Card Transact

Data Platform Layer

- Big Data
- Hadoop Ecosystem
- cloudera
- MAPR
- Spark

Data Deployment Layer

- Fast Data & Edge Data

AI and BigData Deployments

- Scan Streaming
- Percipient

Hewlett Packard Enterprise
Offering you the most comprehensive AI portfolio in the industry
# HPE has a comprehensive portfolio tailored for AI

HPE has a comprehensive portfolio tailored for AI, covering various industries such as Government, academia and industries, Financial services, Government and academia, Life Sciences, Health, and Manufacturing. This portfolio is designed to support AI projects from training models in data centers to edge analytics and inference engines.

## Tools & Resources

### Deep Learning Cookbook
- **Centers of Excellence**
- **HPE/NVIDIA Deep Learning Institute**

### AI-ready software configurations
- **Containers**: NVIDIA GPU Cloud
  - Comprehensive catalog of GPU-accelerated containers for DL software
- **MLOps**: rapid software installation for AI
- **Bare-Metal**: Open Source Distributions
  - combined HPC + AI data analytics workloads

### Choice of fabrics
- **Intel® Omni-Path Architecture**
- **Mellanox InfiniBand**
- **HPE FlexFabric Network**

## Compute ideal for training models in data center
- **HPE SGI 8600**
  - Petaflop scale for deep learning, enterprise platform for accelerated computing
- **HPE Apollo 6500 Gen10**
- **HPE ProLiant**
  - Maximize GPU capacity and performance with lower TCO
- **HPE Apollo 2000 Gen10**
  - The bridge to enterprise scale-out architecture

## Compute for both training models and inference at edge
- **HPE Edgeline EL4000**
  - Unprecedented deep edge compute & high capacity storage; open standards
- **HPE Adaptive Rack Cooling System**
  - Higher power density with less datacenter heat

## Edge analytics and inference engine
- **HPE Data Management Framework**
- **Scalable Storage for Lustre**
- **WekaIO Matrix**
- **HPE Apollo, ProLiant and JBODs**

## AI storage solutions

### AI Transformation Workshop and Digital Prescriptive Maintenance
- **HPE GreenLake Flex Capacity**
- **HPE Datacenter Care Services**

### Confidential
Enterprise platform for accelerated computing
HPE Apollo 6500 Gen10 system

- **Accelerated performance** for GPU intensive workloads
- **Flexibility** for deep learning and HPC environments
- Resilient, secure and simple for lower TCO
Accelerated performance for GPU intensive workloads
Reliable Deep Learning platform with industry leading GPUs and interconnects

HPE's highest GPUs per server
- Up to 125 Tflops* single precision performance
- Eight GPUs per server

Powerful host
- Up to 25% higher performance with 4 InfiniBand network adapters
- NVMe drives
- High speed DDR4 SmartMemory

Leading accelerator technology
- NVLink 2.0 enables dedicated GPU-to-GPU communication

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

* NVIDIA Tesla V100 provides 15.7 TFlops single precision performance with NVLink 2.0 per GPU x 8 GPUs = 125 Tflops
http://www.nvidia.com/content/PDF/Volta-Datasheet.pdf
Flexibility for Deep Learning and HPC environments
Range of configuration and topology options to suit your workloads

**Accelerator choice**
- NVLINK 2.0 for maximum bandwidth
- PCIe for traditional GPU support

**Flexible support and options**
- Many high-speed fabric options
- Ubuntu and Enterprise OS choice
- HPE Pointnext support flexibility

**Accelerator topologies**
- Efficient hybrid cube-mesh for NVLink
- 4:1 or 8:1 GPU:CPU flexibility in PCIe

**Storage options**
- Up to 16 devices
- SAS/SATA SSDs, Up to 4 NVMe drives
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 for resiliency
- 5 fan modules

**Standard rack mount**
- Designed for 1m deep rack
- No special flooring, plumbing needs

Designed for you
Higher performance with NVLink for deep learning / AI

Wide choice of GPUs supported

- PCIe Gen3 x16
- NVLINK (Pascal and Volta) up to 25GBps
- NVLINK for Volta @ 25GB/s

Hewlett Packard Enterprise

For HPE and Channel Partner internal use only
HPE Infrastructure Management Software
HPE iLO 5

**Configuration and Monitoring**
Insure secure transit and locks server hardware configuration

**Remote Management**
Detect and address possible security vulnerabilities in current server setup

**Workload Performance Advisor**
Server tuning recommendations to improve server performance
• **Deploy infrastructure faster** by leveraging template based provisioning

• **Simplifies lifecycle operations** with advanced self monitoring and maintenance tools and frictionless updates for important lifecycle ops like firmware and driver updates

• **Helps increase productivity** by making integrating third party tools a programmatic task through our unified API.
Deploy infrastructure faster

Profile templates built by subject matter experts for your workloads and applications with custom settings for:

- Compute
- Storage
- Networking

90% faster deployment of new configurations

Templates ensure faster deployments with fewer errors whether it’s 1 system or 100
Simplify lifecycle operations

Easy monitoring and maintenance

HPE OneView Global Dashboard

Apply firmware and driver updates within the template

Frictionless updates

97% decrease in time for hypervisor patch update

For HPE and Channel Partner internal use only
Increase productivity

Reduce time to stand up infrastructure from hours to minutes

New-HPOVProfile -name myESXhost -template ESX6.5host

OneView Unified API

Compute  Storage  Network
Single console of management from HPE OneView
Aligns with OneView Global Dashboard

Common look and feel across HPE Management tools
Data storage stages in an artificial intelligence workflow

Optimized storage for various processing needs

**Application Data**
Focus on application integration, resilience, and manageability

**Data Lake**
Storage solutions to provide the right feature set and resiliency for your core business

**Optimization**
Data extraction, cleaning, transformation, reduction to optimize for ML, DL

**Training**
Highly IO-intensive (read), massively parallel, GPU-accelerated learning (training)

Focus on massive throughput, scalability, and parallelism

**Fast Data**
Focus on cost efficiency, reliability and durability for long-term storage

---

Hewlett Packard Enterprise
Hybrid data management with Converged Data Platform Overview

MAPR CONVERGED DATA PLATFORM

HDFS-API | NFS, POSIX, FUSE | S3 | HBase-API | OJAI/JSON | Kafka-API | REST-API

EXISTING ENTERPRISE APPLICATIONS

BATCH & INTERACTIVE ANALYTICS

INTELLIGENT APPLICATIONS

CLOUD-SCALE DATA STORE

ANALYTICS & MACHINE LEARNING

OPERATIONAL DATABASE

GLOBAL EVENT STREAMS

ON-PREMISE, IN THE CLOUD, EDGE
Distributed File and Object store

- Massive Scalability
- Reliability and High Availability

- Deploy Anywhere
- Flexibility to Optimize for Speed and Cost

- Global Namespace
- Automated Data Placement
- Multi-Tenancy and Security

- High Speed Ingest and Data Processing
- Support Different Types of Data
- Naturally Analytics Ready
MapR Data Tiers

KEY BENEFITS

• One solution for different categories of data
• Easy management of data across its life cycle, while maintaining SLAs
• Automated placement and movement of data across tiers
• TCO benefits by leveraging cheap object storage for inactive data
• Monitoring and optimizing space utilization

Hewlett Packard Enterprise
One Global Namespace Across All Tiers

Key Features

- File Rules & Policy Engine
- External Cloud Tiers
- Inline Transparent File Access
- Snapshots & Mirroring Interoperability
- Recalling From Cloud
MapR Database

Advantages of MapR Database
MapR Event Store for Apache Kafka
**Benchmarking Neural Network on Converged Data Platform**

- Eight GPUs: NVIDIA Tesla V100
- Two CPUs: Intel 8167M 2.0 GHz
- Memory: 768 GB
- Network: 1x 25 Gbps. (dedicated)

Five storage nodes
- Two CPUs: Intel E5-2699 2.3 GHz
- Memory: 512 GB
- Network: 1x 10 Gbs (dedicated)
- Eight Local Disks: 3.2 TB NVMe SSD
Running AI Workloads via Containers
NVIDIA DGX-1 Software Stack = NVIDIA GPU Cloud (NGC)
NGC Repository of Software for AI, HPC and more

Setup free login at:
https://ngc.nvidia.com
NGC-Ready Program & Support

HPE Apollo 6500 Gen10 V100 SXM2 and HPE ProLiant DL380 Gen10 are NGC-Ready and HPE continues program validation across NVIDIA-based platforms.

NGC Support Services for HPE Apollo 6500 Gen 10 Systems with Tesla V100 SXM2 and HPE servers with T4 GPUs*:

- Optional enterprise support contract for NGC software with direct access to NVIDIA NGC support specialist

Get started with AI faster with NGC containers

Accelerate AI workload performance with NVIDIA GPU-optimized AI frameworks, libraries and GPU drivers on NGC-Ready HPE systems

Faster time to insight with peace of mind with NGC Support Service from HPE

NVIDIA GPU Cloud (NGC):

- Deep Learning Frameworks
  - Caffe
  - CNTK
  - Theano
  - TensorFlow
- Deep Learning Inference
  - TensorRT
- Deep Learning Libraries
  - CUDA, CUDNN, NCCL
- NVIDIA Container
  - NVDOCKER
- GPU Driver
  - NVIDIA Driver
- System OS
  - UBUNTU, RHEL

*General availability: December 2019
HPE Machine Learning Ops
Bringing DevOps agility to Machine Learning
AI and ML present a huge opportunity for enterprises ….

Business value created by AI

$3.2 Trillion in 2022\(^1\)

Enterprise AI adoption

2.7x growth in last 4 years\(^3\)

Worldwide spending on cognitive and AI systems

$77.6B in 2022

---

1 Gartner Press Release, April 2018
2 IDC – Worldwide Semiannual Artificial Intelligence Systems Spending Guide, 2018
3 Gartner “2019 CIO Survey: CIOs Have Awaken to the Importance of AI”
… but deploying enterprise AI and ML at scale is challenging

“Over 60% of models developed with the intention of operationalizing them were never actually operationalized. There are many reasons for this, but a crucial one is a lack of tools to enable and facilitate operationalization, which is not just about deployment.”

- Gartner

80-85%

Enterprises are running into the ‘last mile’ problem with ML model deployment and management¹

¹ Sumit Pal, Sr Director Analyst, Gartner, “Don't Stumble at the Last Mile: Leveraging MLOps and DataOps to Operationalize ML and AI”
Operationalizing ML models is hard

- A zoo of tools and frameworks for data processing, model development, model training, and visualization
- Not enough effort in standards and processes for operationalizing models
- Multiple barriers to adoption at scale

Only operational ML pipelines deliver business value
Challenges with ML in the enterprise

Security and governance

Data Prep
- Siloed (local) development
- Access to scalable compute
- Data duplication

Build
- Inconsistent processes
- Reproducibility
- Auditability
- Setup of ML/DL environments
- Access to scalable compute
- Support for multiple frameworks
- Inefficient handoff
- Scale

Monitor
- Lack of standards
- Prone to errors
- Interpretability
App store
HPE ML Ops Architecture

Data Engineers

Data Scientists

ML Architects

HPE ML Ops

Data Prep

Build

Train

Deploy

Monitor

Collaborate

BlueData EPIC Platform

Compute

Storage

CPUs

GPUs

NFS

HDFS

aws

Microsoft Azure

Google Cloud Platform

* 3rd party commercial software requires license independent of HPE ML Ops
Optimize investment Using HPE Flexible Capacity Service
How HPE GreenLake Flex Capacity benefits IT

- 59% for compute and
- 48% for storage

Save on costs due to overprovisioning

Enterprises overprovision on average by

- 59% for compute and
- 48% for storage

Capacity ahead of demand
Maintain a safe buffer of capacity
- 50% experienced downtime due to capacity planning
- 57% received complaints of slow performance

1 451 Research November 2016
From Edge to Core - HPE’s Elastic Platform for Analytics

### Intelligent Edge
- Aruba and Edgeline

### Core
#### Workload-optimized compute

<table>
<thead>
<tr>
<th>Synergy</th>
<th>Apollo 2000, 6500 or DL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Streaming, Fast Data Analytics</td>
<td>Interactive</td>
</tr>
<tr>
<td>Apollo 4000</td>
<td>ML / AI</td>
</tr>
<tr>
<td>Batch</td>
<td></td>
</tr>
</tbody>
</table>

#### High speed ToR switching
- 25/100 Gbit Networking

#### Apollo 4000
- Data Lake
- Tiered storage for Big Data Analytics
- Hot, Warm, Cold
- Process, Train
- Data storage for AI workflows

### HPE Enterprise Solutions and Performance validated configurations
- Performance | Security | Best Practices
HPE knows
AI can amplify human capabilities and turn exponentially growing data into insight, action and value. Into competitive edge and new customer experiences.

HPE is
the trusted, global partner enterprises need to navigate this fast-changing field and realize the promise of AI and data analytics everywhere, from Edge to cloud to data center.

Only HPE
is delivering on a roadmap to create more powerful, more flexible, more secure and more efficient computing and data architectures, integrated with the expertise to make AI happen, fast.

HPE is powering AI from edge to cloud to core.
Now and into the future.
Thank you