Whitepaper

Intelligent Storage Solution and Reference Architectures

Trademarks and Permissions

Notice

The purchased products, services and features are stipulated by the contract made between Huawei and the customer. All or part of the products, services and features described in this document may not be within the purchase scope or the usage scope. Unless otherwise specified in the contract, all statements, information, and recommendations in this document are provided "AS IS" without warranties, guarantees or representations of any kind, either express or implied.

The information in this document is subject to change without notice. Every effort has been made in the preparation of this document to ensure accuracy of the contents, but all statements, information, and recommendations in this document do not constitute a warranty of any kind, express or implied.

FUTUREWEI TECHNOLOGIES, INC.

Boston Research Center

Address: 111 Speen Street, Suite 114
Framingham, MA 01701
United States of America

Website: http://www.futurewei.com/

Executive Summary

As the world entered the second decade of the 21st century, AI applications are changing the world rapidly. In many enterprises, AI infrastructure is critical to the future and storage is a key piece of this foundation.

An end-to-end AI infrastructure solution not only provides the raw compute and storage capacity, but also the software capability that can drive values out of the system. Several new types of AI pipelines, such as model pipelines and insight pipelines, can provide value to the business. These pipelines, gradually trained by real data and generated from the combined knowledge of vendors and users, can have a direct impact on business.

Performance is also critical. Insights need to be generated in the shortest amount of time. High bandwidth and low latency are major requirements for AI training infrastructure. The goal is to NOT let other infrastructure as the bottleneck and fully saturate GPU processing capability. By doing this, the ultimate TCO is achieved for customer. GPUDirect Storage (GDS) by Nvidia® is an example to provide the high bandwidth needed by AI applications. Storage vendors are fighting to a higher GDS benchmark score, so they are gain a foothold on the market. The benchmark numbers are refreshed almost monthly.

At the same time, AI applications are demanding more from the vendors supporting them. While vendors are working on providing more powerful machines, engineers and researchers are developing innovative products and algorithms that use more resources. The infrastructure needs to be scalable, reliable, and fast to support more computation and IO requests. As the use cases and AI algorithms being used can change frequently, an AI infrastructure must be open to the ecosystem and adaptive to different algorithms.

At last, as GPUs and fast CPUs are expensive, an AI infrastructure needs to be cost-effective. It must find a balance between the time to finish a job and the cost spend with it. AI infrastructure also needs to be efficient in a way how GPUs (more expensive) can be utilized to their full potentials.

In this white paper, we survey different use cases and classify their workloads. Then, we gave an overview of the benchmark in the industry. Then based on these use cases, a common reference architecture based on open-source and public tools is proposed.

The purpose of this document is to discuss a set of interesting topics related to AI infrastructure, so the reader can obtain an overall picture of AI infrastructure use cases on the market. The document is having the following audiences in mind:

- *IT executives who are interested in AI use cases and AI solutions.*
- *Solution Architects who are interested in AI use cases and requirements.*
- *Sales and marketing professionals who are interested in current industry trends.*

CONTENTS

1 AI ECOSYSTEM AND MARKET TREND

1.1 GENERAL TRENDS

The AI ecosystem development is driven by both industrial companies and academic researchers. As deep learning algorithms found practical usage in our daily life, they are widely adopted in computer vision, content recommendation, autonomous driving, and life science research. AI has become one of the fastest-growing areas in the computer science field.

While big data technologies are having a wide adoption in the field of data analytics, machine learning (deep learning in particular), joined the toolbox to provide better utilization of the large amount of data collected. The abundance of training data and more advanced algorithms together proved that AI could do much more than what people expected 10 years ago. Many companies that used to invest heavily in HPC platforms are now adding AI deep learning platforms into their IT infrastructure.

On the vendor and service provider side, all tech companies are investing heavily in the AI market. Many companies are investing in the AI technology stack from top to bottom. From the lower layer components such as AI chips, chip drivers, math libraries, deep learning frameworks, job scheduling platforms, all the way to AI applications for different industries.

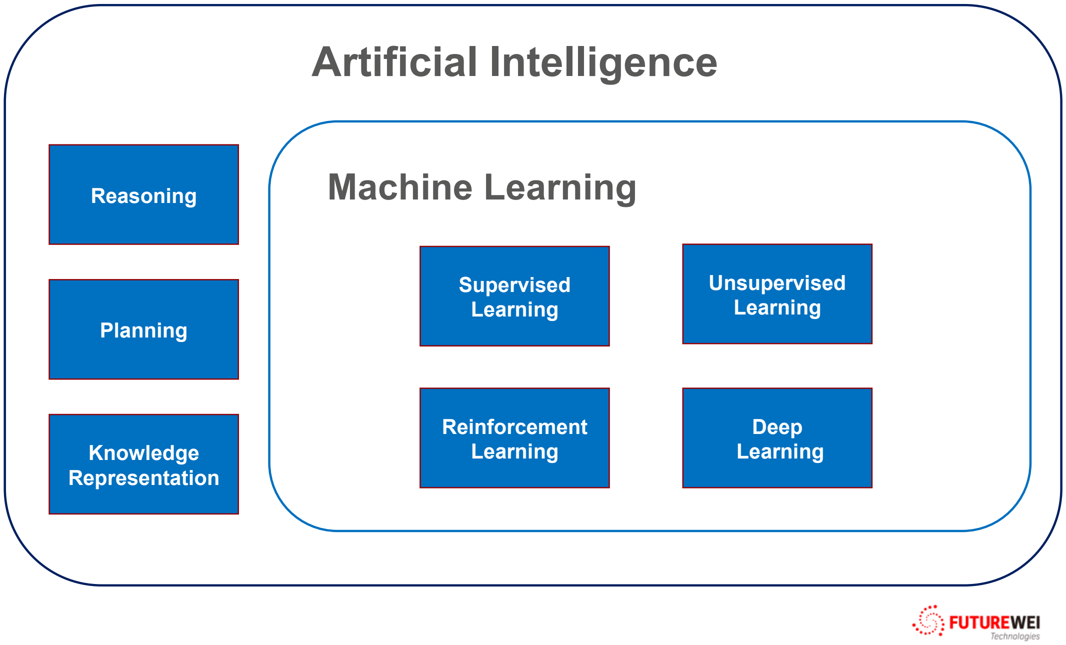
The AI ecosystem evolves around several key players. Nvidia®, which has a dominant foothold on the latest GPU technology, is branching into software libraries and AI applications. Google®, on the other hand, invested heavily in the open-sourced deep learning framework Tensorflow, while developing a proprietary Tensor Processing Unit (TPU) chip. Facebook® is promoting PyTorch, which has obtained a similar market share as Tensorflow. Huawei developed Ascend, a Neural Processing Unit (NPU), and Mindspore, a deep learning framework.

Deep learning algorithms are evolving rapidly such that many capabilities of advanced AI programs 3 years ago are now considered basics. Thanks to the openness and sharing of the AI community, new concepts are adopted quickly into products. Many barriers to industrial adoption are no longer technical but legal and cost.

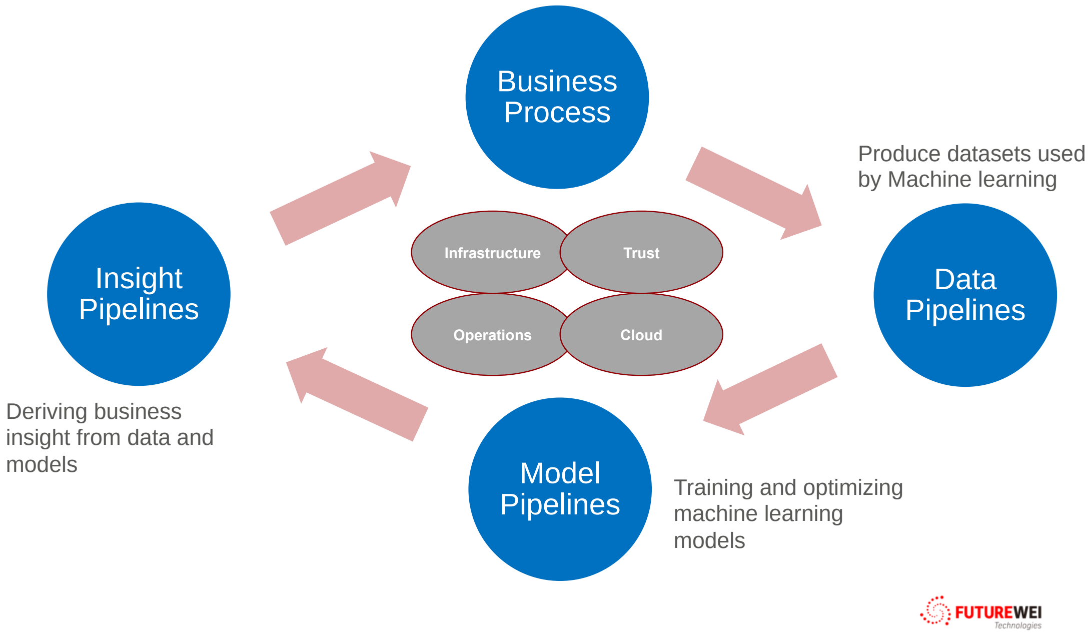
1.2 AI/ML Workflows, Framework

AI/ML has gained extreme popularity over the last 10 years. However, Artificial intelligence research can be traced back to as early as the 1940s. Due to a large amount of dataset availability and extreme progress of hardware computation capability (especially GPUs), deep learning has gained tremendous traction. However, organizations are still leveraging traditional and other AI/ML techniques. It is still necessary to capture all the AI/ML technologies.

As shown in the following diagram, Machine learning is the key part of AI. AI, however, does cover other technologies such as reasoning, planning, and knowledge representation. DL has garnered more attention for the last 10 years due to the reasons mentioned before.



The differences between AI and ML are in semantics. ML is a technique and AI, on the other hand, is capability. AI means the capability of computers to demonstrate intelligent behavior. The realization of the AI capability involves many techniques and components, and ML is just one of them. Specifically, an AI system consists of three distinct building blocks: data pipelines, model pipelines, and insight pipelines. Data pipelines are processes that integrate siloed data sources and produce the datasets to be used by machine learning. Model pipelines are processing that train and tune machine learning models based on the datasets. Insight pipelines are processes that derive business insights from data analytics and model inferencing. Also, production AI must cope with infrastructure deployment and optimization. It must ensure AI decisions are trustworthy. It must operationalize the end-to-end system. And it must support hybrid multi-cloud environments.



All different components of an AI system shall be tacked to enable broad adoption of AI, not just the ML phase (Model pipelines).

1.3 Infrastructure Challenges

None of these innovations is possible without the help of an AI infrastructure. Like other IT infrastructures, an AI infrastructure consists of three major parts: compute, network, and storage. The compute part can be further divided into CPU-based and accelerator-based computations. The software stack can be divided into orchestration software, AI application software, and core AI system software.

The core AI software includes training infrastructure, inferencing infrastructure, framework (e.g., Tensorflow), distributed scheduling (e.g., Horovad [1]), and GPU-acceleration libraries (e.g., CUDA for Nvidia® GPUs).

The application software includes applications that are used to solve real problems. For example, Nvidia® Clara is an AI-powered healthcare application framework. There are many AI application software powering other industries, such as autonomous driving, life sciences, financial industry, etc.

The orchestration software is a component to facilitate the input/output of training, deploy the generated models, etc. This component is platform-dependent, and many orchestration software is built on top of infrastructure software provided by the platform. For example, if models are deployed as containers, then the AI orchestration software is using the container deployment infrastructure.

For cloud deployments, due to the size of the deployment, an AI infrastructure can be dedicated. However, for medium-sized corporations, a mid-size AI infrastructure may share hardware and software with big data infrastructure as both need access to storage and compute resources.

An interesting case for AI infrastructure is the hybrid cloud. Public clouds are known for their elasticity, which means a large set of compute or storage resources can be provisioned right away. The computing resources also include many software components that can be used to form a bigger application environment. Public clouds fit particularly well to use cases in which the workload fluctuates. Users can use dynamic provision and avoid paying for any services that are not used. On the financial statement, users can save on the initial cost of building a data center and turn CAPEX into OPEX.

However, public clouds often face the cost issue when the scale is large and the workload is stable. In this case, buying or renting from on-prem IT vendors is often more cost-effective than using public cloud services. Another case of using on-prem services is when the user would like to use different hardware equipment from the ones provided by cloud vendors. In the AI infrastructure case, some users may prefer a particular Neural Processing Unit (NPU) or FPGA for acceleration, and this option is not available on the cloud.

Hybrid cloud has the potential to combine the pros of both public clouds and on-premise infrastructure. But the key issue to be solved is the data movement problem. It is well known that data gravity exists and customers are making different decisions based on their IT and financial needs.

Moving a large amount of data is very costly. Moving a large amount of data in a short time is even more costly. The network bandwidth is tight, and therefore the classic question was asked about whether it is faster to move the disks or transfer the data via a network. As a result, the center of the data gravity should be considered before the data started accumulation.

There are three types of hybrid clouds: (1) On-prem is the center and cloud is used as development to take advantage of the nimble of fast provisioning and development. But the production remains at the on-prem side for data safety and cost reduction. (2) The public cloud is the center and the on-prem side is used as a stage to make sure data is processed before it is transferred to the cloud. (3) On-prem is one of the multi-cloud centers. Workloads can be switched among these centers as needed.

In this document, we do not assume any of the three types. The computation (e.g., programs and models) can be moved from one data center to another at any time.

2 AI WORKLOAD CLASSIFICATION

One of the major requirements from AI/ML infrastructure providers is how to systematically categorize computing and data transfer demands for AI/MI workloads. Different AI/ML applications put pressure on different parts of the infrastructure.

Based on related statistics, around 62% of total execution time among AI/ML workloads on average is spent in weight and gradient communications. 60% of workloads can be potentially sped up by using AllReduce architecture exploiting high-speed links between GPU interconnects. And 1.7x speedup can be achieved when Ethernet bandwidth is upgraded from 25G to 100Gbps.

AI/ML has rather different phases: ML analytics and big data processing, training, and inference. Each phase has rather different workload characterizations. The table below summarizes high-level storage requirements in a different phase of AI workflow:

	DATA CHARACTERISTICS	STORAGE REQUIREMENT	NETWORK
INGEST	write-heavy, sequential I/O, mixed file sizes	high throughput	10-100GE
PREPARE	read/write heavy, random/sequential I/O, mixed file sizes	random I/O performance	10-100GE

TRAIN	read/write heavy random I/O, small files	highly parallel, high bandwidth (GBs/s), low latency (<1ms)	100GE/IB RDMA preferred
INFERENCE	read-heavy, multi-tenant	extreme low latency	100GE RoCE / IB

2.1 ML Analytics and Big Data Processing

This phase is commonly referred to as the data cleaning and analytics phase, equivalent to “ingest” and “prepare” in the summary table above. It is more related to big data and analytics than ML. This phase can be broken down into several stages: Data ingestion, Data cleaning, and data egress.

The data ingestion phase involves a massive amount of data streaming into a data lake. This is a typical large quantity of files with sequential write workloads.

Data Cleaning plays an important role in the field of Data Management as well as Analytics and Machine Learning. Data cleaning typically demands large or small sequential read/write workloads with a large throughput.

2.2 Training

The most important factor to support training performance is to NOT let other infrastructure components be bottlenecks. Let GPU be! That is easier said than done. Training workloads are normally more I/O intensive. Latency, throughput as well as parallel access are all requirements for storage and network. Random small file read with large quantities is a typical workload for training (image processing, computer vision, etc.). Individual files seldom big enough to deserve any attention, but there could be millions of files at scale. Total aggregate throughput can also be up to more than multiple GB/S. Due to this requirement, a lot of IT infrastructure for AI/ML likes to deploy a parallel file system for AI storage needs. A lot of new technologies like GDS and RDMA are leveraged to speed up data transfer between storage and GPUs.

The outputs of the AI/ML training phase are more easily manageable for IT infrastructure. They are often small enough that there is no issue with modern enterprise IT systems.

2.3 Inference

During the inference phase, AI/ML algorithms apply pre-trained models on incoming data to make a decision/prediction. The amount of data involved in this phase is typically small, however, the decision normally needs to be made as soon as possible, and requests could come frequently. This operation requirement brings a different challenge to the storage system compared to the data processing and training phase. The capacity of the storage systems is not key consideration anymore and read I/O latency has become the number one important factor. For small-scale systems, a local NVME SSD will do the job, assuming the training model is actively synced up with the training side of the pipeline. NVME over RDMA fabric is most suitable for large-scale or distributed systems.

3 Benchmark

There are many popular benchmark tools and platforms in the industry, these generic results provide valuable information as a reference for AI and storage applications. We will go through a few most popular and widely accepted benchmarks in the storage industry in this chapter.

3.1 Storage Benchmark

3.1.1 Primary Storage Benchmark

Primary storage is very sensitive to both latency and throughput in block access. One of the most widely accepted primary storage benchmark scoreboards is SPC-1 [27]. Huawei® OceanStor Dorado is the number 1 performer on this list, with a single cluster offering 21 million IOPS throughput and as low as 0.05ms latency on block access.

Here is the latest published result:

Rank	Performance (SPC-1 IOPS)	Test Sponsor	Price-Performance (per SPC-1 KIOPS™)	Submission Identifier	Tested Storage Product
#1	21,002,561	HUAWEI	CN¥ 2913.78	A32018	OceanStor Dorado 18000 V6
#2	11,000,576	宏杉科技 macrosan	$ 385.64	A32020	MS7000G2-Mach
#3	10,001,522	FUJITSU	$ 644.16	A32009	ETERNUS DX8900 S4
#4	7,520,358	inspur 浪潮	$ 386.50	A32014	Inspur AS5600G2
#5	7,000,565	HUAWEI	$ 376.96	A31017	Huawei OceanStor Dorado18000 V3

3.1.2 File System Benchmark

NAS is a very popular format of storage in the AI/big data area. One of the most popular file storage benchmarks is IO 500 by Virtual Institute [28]. This list covers a range of popular file systems from industry and academy, such as file system type Lustre and DAOS; vendor list includes Intel, WekaIO, and Argonne National Lab. Huawei currently does not have a presence on this list. The figure below shows top entries on the current IO500 list:

#	information								io500		
	list id	institution	system	storage vendor	filesystem type	client nodes	client total procs	data	score	bw	md
										GiB/s	kIOP/s
1	sc20	Pengcheng Laboratory	Pengcheng Cloudbrain-II on Atlas 900	Pengcheng Laboratory	MadFS	255	18360	zip	7043.99	1475.75	33622.19
2	isc20	Intel	Wolf	Intel	DAOS	52	1664	zip	1792.98	371.67	8649.57
3	sc19	WekaIO	WekaIO on AWS	WekaIO	WekaIO Matrix	345	8625	zip	938.95	174.74	5045.33
4	isc20	TACC	Frontera	Intel	DAOS	60	1440	zip	763.80	78.31	7449.56
5	isc20	Argonne National Laboratory	Presque	Argonne National Laboratory	DAOS	16	544	zip	537.31	108.19	2668.57
6	sc19	National Supercomputing Center in Changsha	Tianhe-2E	National University of Defense Technology	Lustre	480	5280	zip	453.68	209.43	982.78
7	sc20	Intel	Endeavour	Intel	DAOS	10	640	zip	353.72	43.59	2870.64
8	isc20	KISTI	NURION	DDN	IME	2048	2048	zip	282.45	515.59	154.74
9	isc20	Oracle Cloud Infrastructure	BeeGFS on Oracle Cloud	Oracle Cloud Infrastructure	BeeGFS	270	3240	zip	267.25	293.05	243.73
10	sc20	JCAHPC	Oakforest-PACS	DDN	IME	2048	4096	zip	253.57	697.20	92.22
11	sc19	NVIDIA	DGX-2H SuperPOD	DDN	Lustre	10	400	zip	249.50	86.97	715.76
12	sc20	EPCC	NextGENIO	BSC & JGU	GekkoFS	10	3800	zip	239.37	45.79	1251.32
13	sc19	University of Cambridge	Data Accelerator	Dell EMC	Lustre	128	2048	zip	229.45	131.25	401.13

3.1.3 Database Benchmark

Major database vendors frequently offer their benchmarking utility to evaluate storage system performance. Some are only available to customers and partners, such as SAP® HANA® standard application benchmark tool. There are also some open tools such as Microsoft® Diskspd and SQLIO. Due to the vast difference between different database software, there is not any meaningful industry-wide scoreboard for storage performance in database applications. Each application tends to maintain its list or leave it up to the customer to decide.

3.1.4 Generic I/O Benchmark Tools

There are plenty of generic storage I/O performance benchmark tools available, the most popular ones for testing enterprise storage arrays include iometer and fio. The fio community has many scripts available for simulating all kinds of common types of workload such as database, big data analytics, etc.

Fio has comprehensive documentation available online [29].

The project source code is hosted on GitHub® [30], with plenty of examples showing how you can use a script [31] to simulate all kinds of workload to match your real application.

3.1.5 GPUDirect Storage Benchmark

In 2020, Nvidia® published a new way for GPU to directly access storage through RDMA fabric without involving CPU and system memory in the middle. This greatly improved high-end AI-storage cluster overall performance. The GPUDirect software suite includes a benchmark tool “gdsio” by Nvidia® to test the I/O performance of the cluster.

Here is a sample command for benchmarking local NVME disk direct access through GDS:

```
/usr/local/cuda/gds/tools# ./gdsio -f /mnt/test/testfile1 -d 0 -w 4 -s 10G -i 1M -I 0 -x 0
IoType: READ XferType: GPUD Threads: 4 DataSetSize: 10212352/10485760(KiB) IOSize: 1024(KiB) Throughput:
2.681944 GiB/sec, Avg_Latency: 1456.483233 usecs ops: 9973 total_time 3.631417 secs
```

Simillarly, GDS can directly access remote NVME over fabric using NVMEoF protocol:

```
/usr/local/cuda/gds/tools# ./gdsio -f /nofmnt/testnof -d 0 -w 4 -s 10G -i 1M -I 0 -x 0
IoType: READ XferType: GPUD Threads: 4 DataSetSize: 1016832/1048576(KiB) IOSize: 1024(KiB) Throughput:
2.236000 GiB/sec, Avg_Latency: 1743.542703 usecs ops: 993 total_time 0.433688 secs
```

See the GDS website [32] for more details on how to use the utility to troubleshoot and benchmark GPUDirect storage.

3.2 AI/ML Compute Benchmark

Due to the complexity and huge variety of AI/ML applications, there is a lack of industry-standard benchmark tools or widely accepted scoreboards. One commonly accepted standard is the raw computing power in terms of how many TFLOPS the system can handle, this does not necessarily have a linear correlation with different applications, but still provide a good reference point.

Since the AI hardware market is mostly dominated by Nvidia®, a very good reference for checking hardware raw computing power is the wiki page for Nvidia® GPU [33]. It contains a detailed specs comparison table of GPUs within each generation/family. For example here is the table for Tesla data center GPU [34], here is a sub-section of the Tesla table with the most important information for AI compute efficiency.

Model	**Micro-architecture**	**Chips**	**Shaders**	**Memory**				
			Cuda cores (total)	**Bus width (bit)**	**Size (GB)**	**Bandwidth (GB/s)**	**Single precision (MAD or FMA)**	**Double precision (FMA)**
P100 GPU accelerator				3072	12	549	8071–9340	4036–4670
V100 GPU accelerator (mezzanine)	Volta	1× GV100-895-A1	5120	4096	16 or 32	900	14899	7450
V100 GPU accelerator (PCIe card)		1× GV100					14028	7014
T4 GPU accelerator (PCIe card)	Turing	1× TU104-895-A1	2560	256	16	320	8100	Unknown
A100 GPU accelerator (PCIe card)	Ampere	1× GA100-883AA-A1	6912	5120	40	1555	19500	9700

Some generic tool suite such as ai-benchmark [35] offers a wide range of coverage on popular algorithms. It can be used as a good reference too. Here is the top part of a detailed list of GPU AI benchmark score ranking provided by ai-benchmark.

Model	TF Version	Cores	Frequency, GHz	Acceleration	Platform	RAM, GB	Year	Inference Score	Training Score	AI-Score
Tesla V100 SXM2 32Gb	2.1.0	5120 (CUDA)	1.29 / 1.53	CUDA 10.1	Debian 10	32	2018	17761	18030	35791
Tesla V100 SXM2 16Gb	2.1.0	5120 (CUDA)	1.31 / 1.53	CUDA 10.1	Red Hat 7.5	16	2017	17251	17836	35086
Tesla V100 PCIE 32Gb	2.1.0	5120 (CUDA)	1.23 / 1.38	CUDA 10.1	Debian 10	32	2018	16530	17865	34394
Tesla V100 PCIE 16Gb	2.1.0	5120 (CUDA)	1.25 / 1.38	CUDA 10.1	Red Hat 7.5	16	2017	16511	17837	34347
NVIDIA Quadro GV100	1.14.0	5120 (CUDA)	1.13 / 1.63	CUDA 10	Debian 10	32	2018	16748	17132	33880
NVIDIA TITAN V	2.1.0	5120 (CUDA)	1.20 / 1.46	CUDA 10.1	Ubuntu 18.04	12	2017	16192	17215	33406
NVIDIA TITAN RTX	2.1.0	4608 (CUDA)	1.35 / 1.77	CUDA 10.1	Ubuntu 18.04	24	2018	16084	17255	33339
GeForce RTX 2080 Ti	2.1.0	4352 (CUDA)	1.35 / 1.55	CUDA 10	Debian 10	11	2018	16042	16828	32870
NVIDIA Quadro RTX 8000	2.1.0	4608 (CUDA)	1.40 / 1.77	CUDA 10.1	Debian 10	48	2018	13014	14637	27651
NVIDIA Quadro GP100	2.0.0	3584 (CUDA)	1.30 / 1.44	CUDA 10	Red Hat 7.4	16	2016	12264	13436	25700
NVIDIA TITAN Xp	2.1.0	3840 (CUDA)	1.41 / 1.58	CUDA 10.2	Debian 10	12	2017	11948	12922	24870
GeForce GTX 1080 Ti	2.1.0	3584 (CUDA)	1.58 / 1.60	CUDA 10.2	Debian 10	11	2017	11914	12473	24386
GeForce RTX 2080 SUPER	2.1.0	3072 (CUDA)	1.65 / 1.82	CUDA 10.1	Windows 10	8	2019	11513	12734	24247
GeForce RTX 2070 SUPER	2.1.0	2560 (CUDA)	1.61 / 1.77	CUDA 10.2	Ubuntu 18.04	8	2019	11472	12710	24182

Tesla V100 is the GPU that is used in the original version of Nvidia® DGX, the latest generation DGX-2 uses a more advanced Tesla A100 card, which is even more powerful.

4 Reference Architecture

4.1 Summary

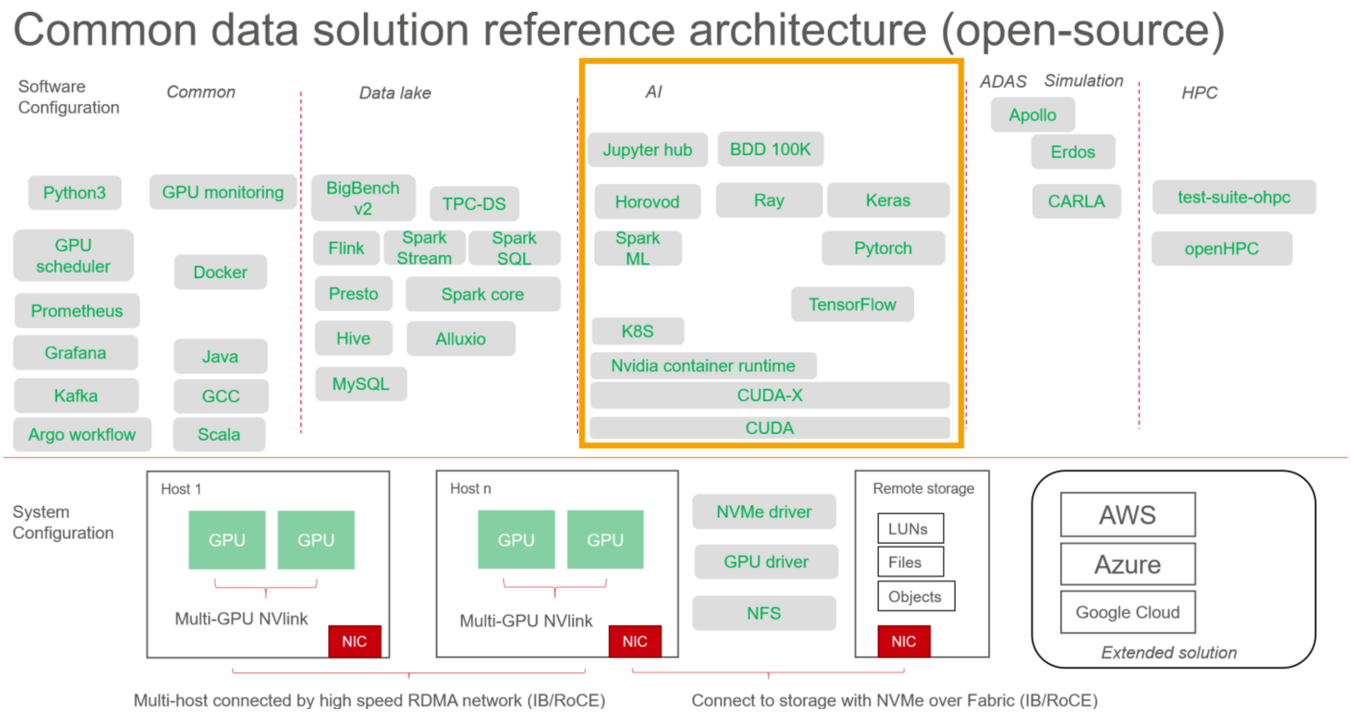
AI/ML especially deep learning has made significant progress during the last 10 years, particularly in the fields of NLP, computer imaging, vision, and autonomous driving. It is powered and accelerated by advanced research of new algorithms and large datasets with higher than ever computational powers.

To fully utilize the full potential of ever-increasing computational power for ML and reduce overall costs of ML, a new infrastructure architecture shall be designed to accommodate those new workloads, reduce overall complexities of infrastructure and strike a balance among computing, network, and storage. This section is trying to classify different choices for overall architecture and give recommendations to build computation, memory, storage, and network as well as software ecosystem.

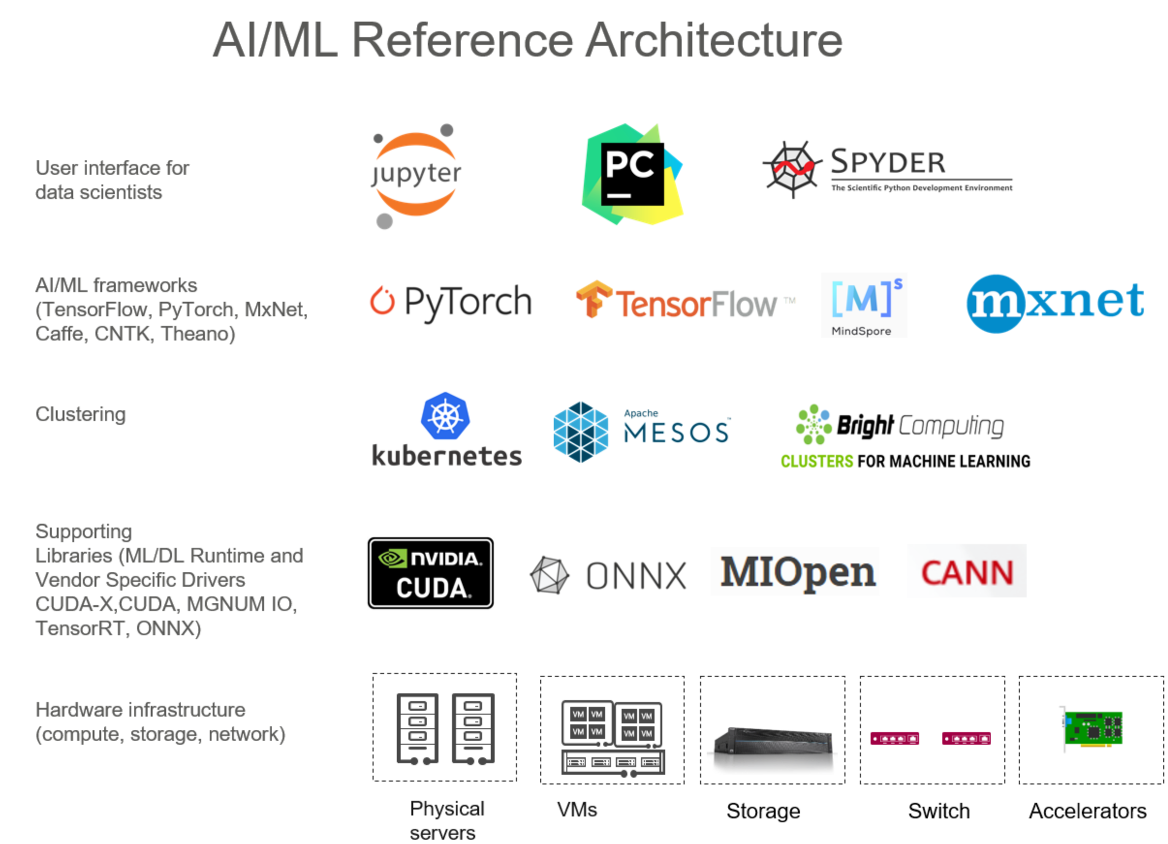
Based on application requirements for performance, storage, and network, we break down AI/ML configuration into 3 typical tiers:

4.2 AI/ML Reference Architecture

Regardless of different tiers, there is a common need for an AI/ML architecture as shown below. Here is a common data solution reference architecture for HPC and AI/ML. As you will know, it is rather complex and overwhelming for data science to choose the right hardware, software, and frontend to start their works. The goal is to not let data scientists worry about those configurations among storage, network, and GPU. Rather let them focus on data science only.



By consolidating different use cases, requirements, and other factors, we propose the following reference architecture. We will describe each sector in detail:



4.2.1 AI/ML Data Science Front End

An AI/ML data science front end allows data scientists to keep track of errors and maintain clean code. It is a one-stop-shop of IDE environment for AI and ML. Jupyter® notebook, Jupyter® lab, Spyder®, Guleviz®, Orange®, RSstudio®, VSC Code are some of the typical front-ends with the Jupyter® series being the most popular ones. The front end should allow end-users to choose the framework to use and which cluster to run the training program. The front end should mask the gory details of the underneath layers and expose the necessary choices to the end-users.

4.2.2 AI/ML frameworks

In this layer, an end-user can choose from many frameworks. The most popular two today are Tensorflow and PyTorch. Many others have their user communities. Many vendors push their own AI/ML library to influence the community. This layer is generally open-source-based.

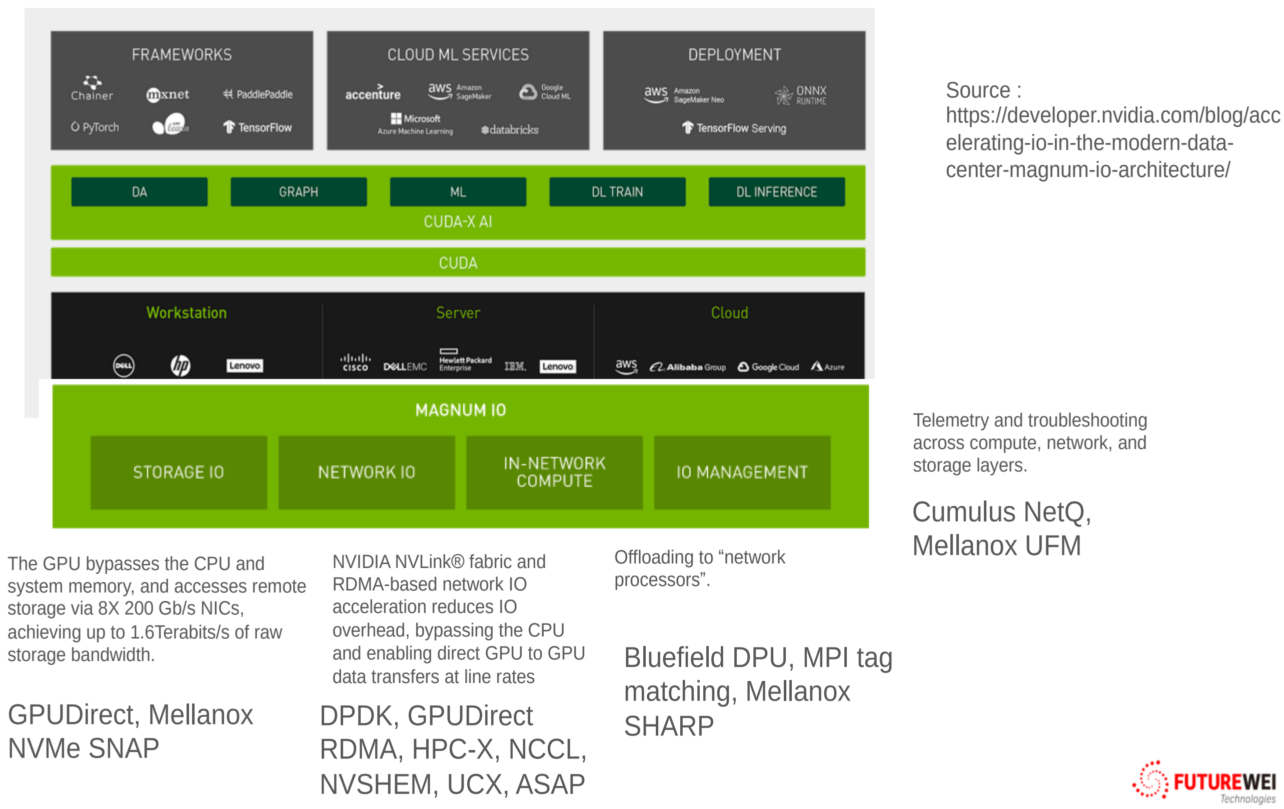
4.2.3 Cluster Resource Management

A cluster resource management layer is to manage different compute cluster resources to provide upper lay resource isolation and provisioning. It also facilitates customer application deployments on top of resource management and orchestration software (Yarn, K8S, Mesos, etc.). The most popular cluster resource management software is Jupyter enterprise gateway and its commercial enhanced variant – Bright Cluster Management for AI/ML.

4.2.4 Supporting libraries

This section is hardware-related because each vendor has its libraries based on the special hardware. For example, CUDA for Nvidia® GPUs and MIOpen for AMD Instinct GPUs. There are many AI/ML solutions that customers can choose from. Being at top of the game, Nvidia® provides the most complete ecosystem in this area with supporting most major frameworks and hardware optimizations.

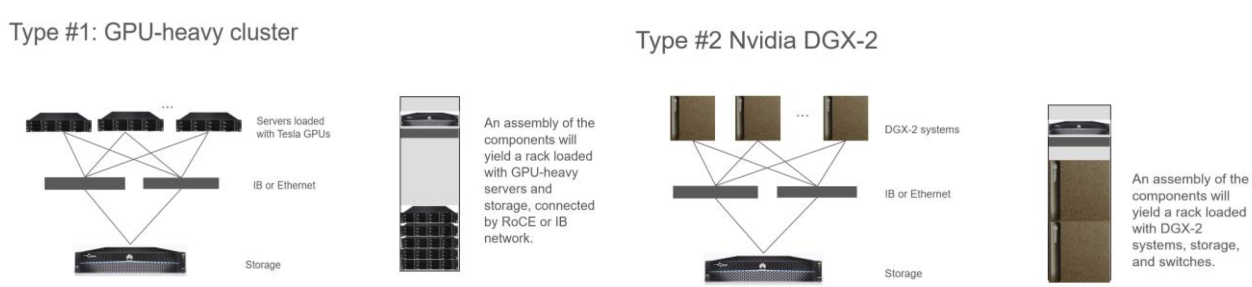
The following diagram shows an optimized stack for Nvidia® AI/ML ecosystem:



4.2.5 Infrastructure

This is a fundamental building block of ML/AI, all the performances and computation power come from this layer. Generally, these layers include 3 major parts: Compute (including GPUs), storage, and network. Compute usually is built with high GPU density high-performance servers with larger memory and faster local NVMe drives. A good example is the Nvidia® DGX system. Fast speed and big throughput with RDMA capability network shall also be provided to satisfy the bandwidth requirement of GPU and applications. For large-scale training, a scale-out NAS storage system shall be provided.

Below are shown 2 typical configurations for this type of infrastructure:



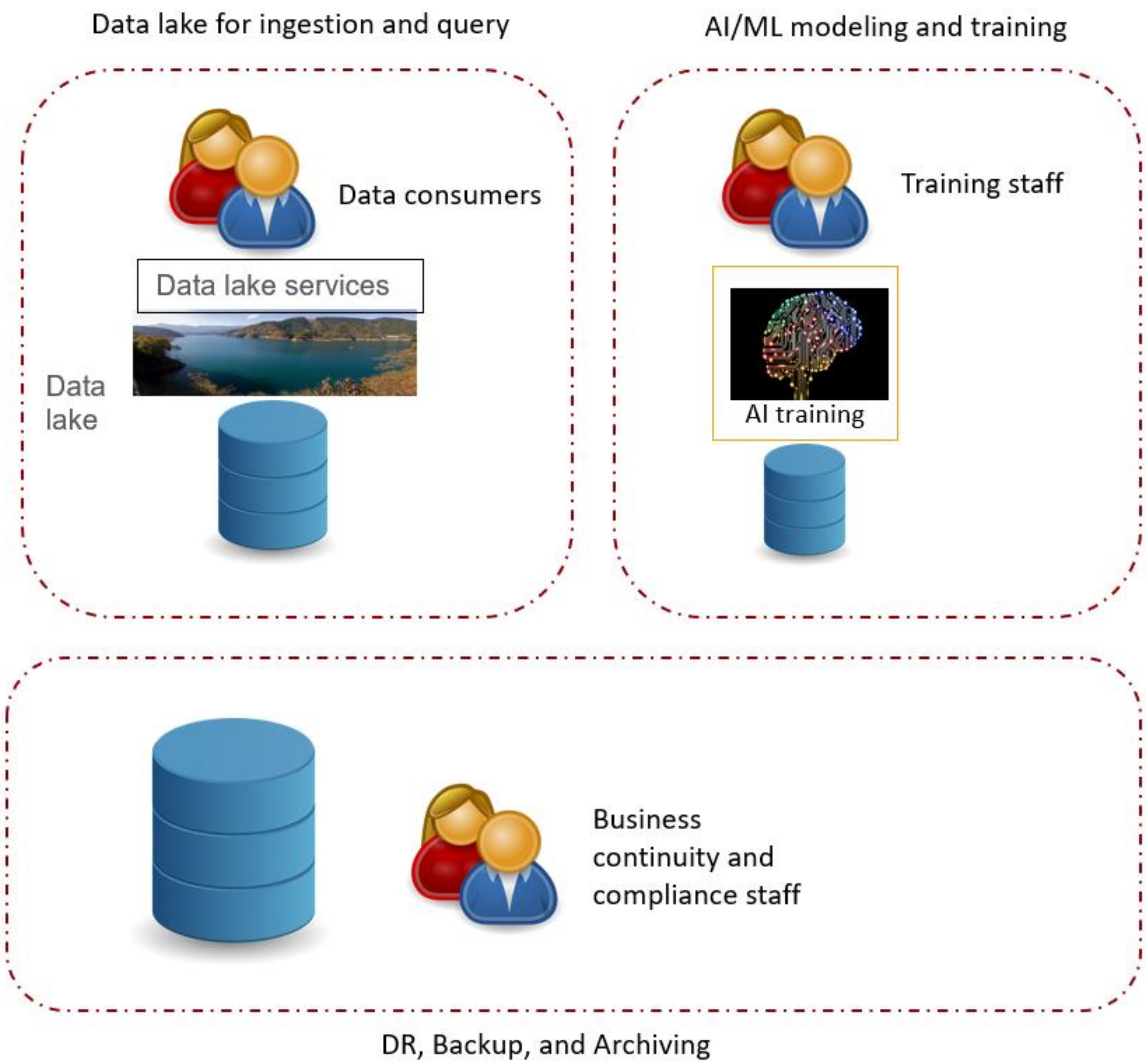
- *Compute*
  In this reference architecture, the compute can be composed of any mainstream CPUs, such as Intel Xeon, AMD, or various ARM-based CPUs. The GPU/TPU can also use different vendors. As Nvidia® is

having a strong position in the GPU world, an Nvidia® GPU accelerated system is a popular choice. But other chips are possible if they support major AI/ML frameworks, such as TensorFlow and Pytorch.

The compute-storage disaggregated architecture allows scale-out of the computing resources and storage resources independently. Therefore, it is well suited for large-scale deployment. For smaller environments, a distributed hyper-converged (dHCI) environment can be deployed. The dHCI solution has both the advantage of the traditional HCI (fast deployment, easy management) and the advantage of the disaggregated architecture (flexibility).

- *Network*
  The network section can use multiple high-speed networks. Higher bandwidth and lower latency are critical to the performance of this disaggregated architecture. However, due to cost reasons, users may not want to use the fastest network. RDMA over fabric and RDMA over Ethernet (RoCE) are the two most popular network standards used in this reference architecture. Users may choose a network that fits the budget and bandwidth needs.

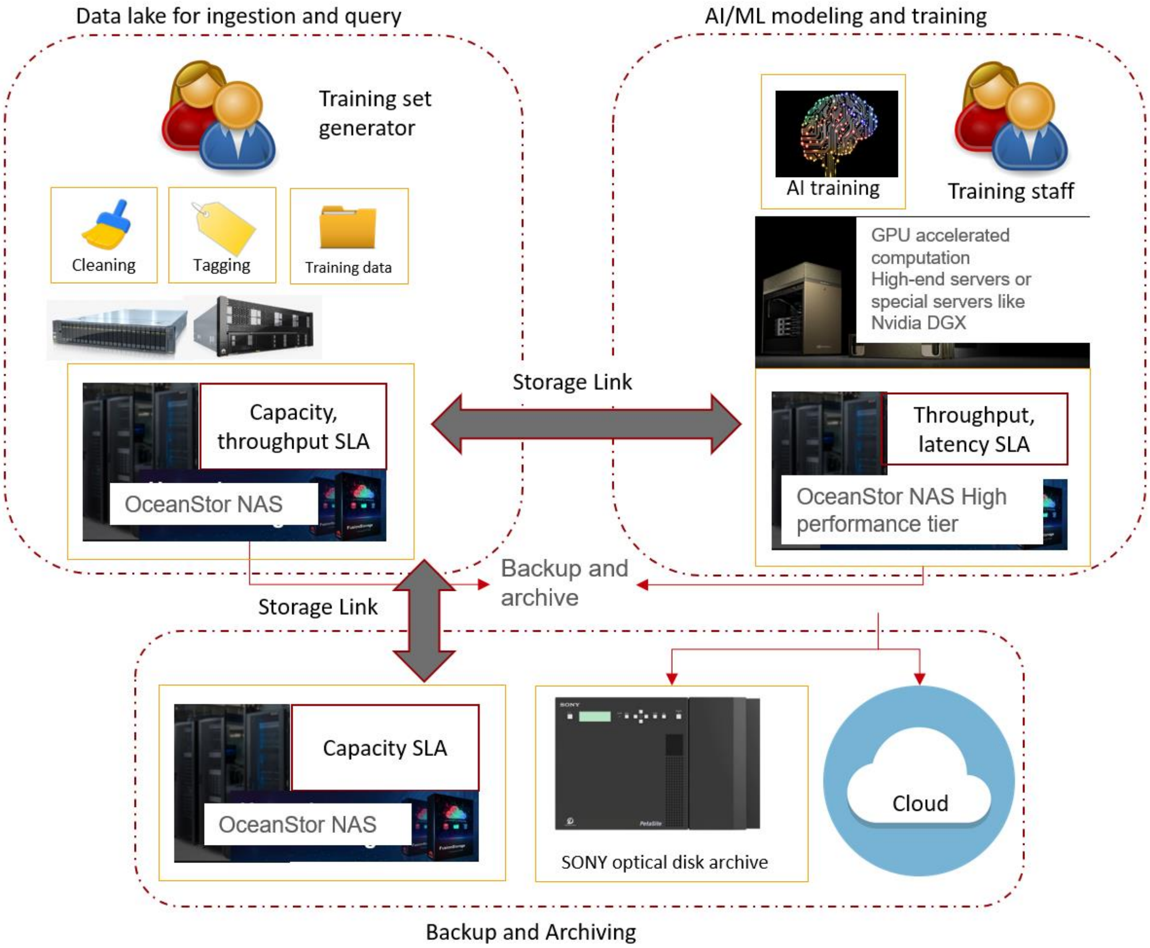
- *Storage*
  There are three areas where storage requirements arise. They are the data lake, AI/ML training, and DR/backup/archiving. These three areas have different focuses, even though they have overlapping goals. The data lake users may have stricter criteria of the unit price of each GB, due to the amount of data needed. But they also put performance and throughput into consideration. The AI/ML training users are concerned with the speed of training and the performance of underlying storage. The DR/backup/archiving users may have an even bigger capacity challenge because multi-year data will be kept for compliance with regulations. But latency is generally not an issue.



4.3 High-end AI Cluster

For mission-critical AI applications that require ASAP response and not very sensitive to cost, the solution should include the following components:

- Compute: A powerful computing node such as Nvidia® DGX-2 or equivalent servers.
- Network: RDMA fabric (100Gb RoCE or InfiniBand) is recommended for connection between the compute node and storage system, so the AI engine can access data through NVMeOF with minimal latency and high throughput.
- Storage: The storage in the solution can be mid to high-end all-flash array depends on the estimated workload and capacity requirements of the targeted application.



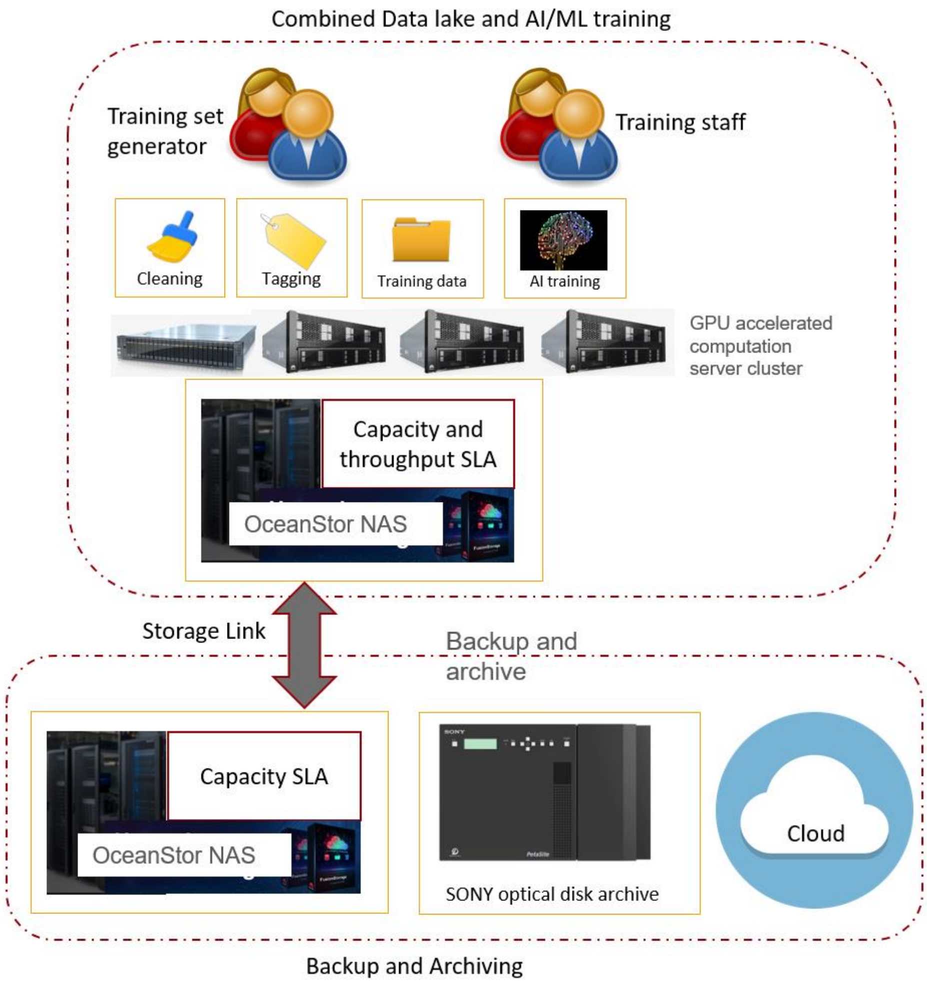
This solution is suitable for tier 0 applications such as hedge funds, where each second delay in making a correct decision may mean millions of dollars. Other use cases include the insurance industry where clients could be waiting online for a quote, and credit card company fraud detection.

4.4 MID-RANGE AI CLUSTER

Many AI application clients are seeking a balance between performance and cost, most of them can be classified into this solution group. The proposed solution should maximize performance per dollar spent while satisfying the application's minimal requirements. Here is an example of components for a mid-range AI compute / storage cluster:

- Compute: Multi-GPU/TPU server cluster.
- Network: 25/40Gbps RoCE RDMA fabric is recommended for connection between the compute node and storage system. This setup will allow NVMeOF access for best performance and cost less than the 100Gbps solution.

- Storage: Mid-range all-flash array with enough capacity and front-end I/O modules to satisfy minimum application requirements.

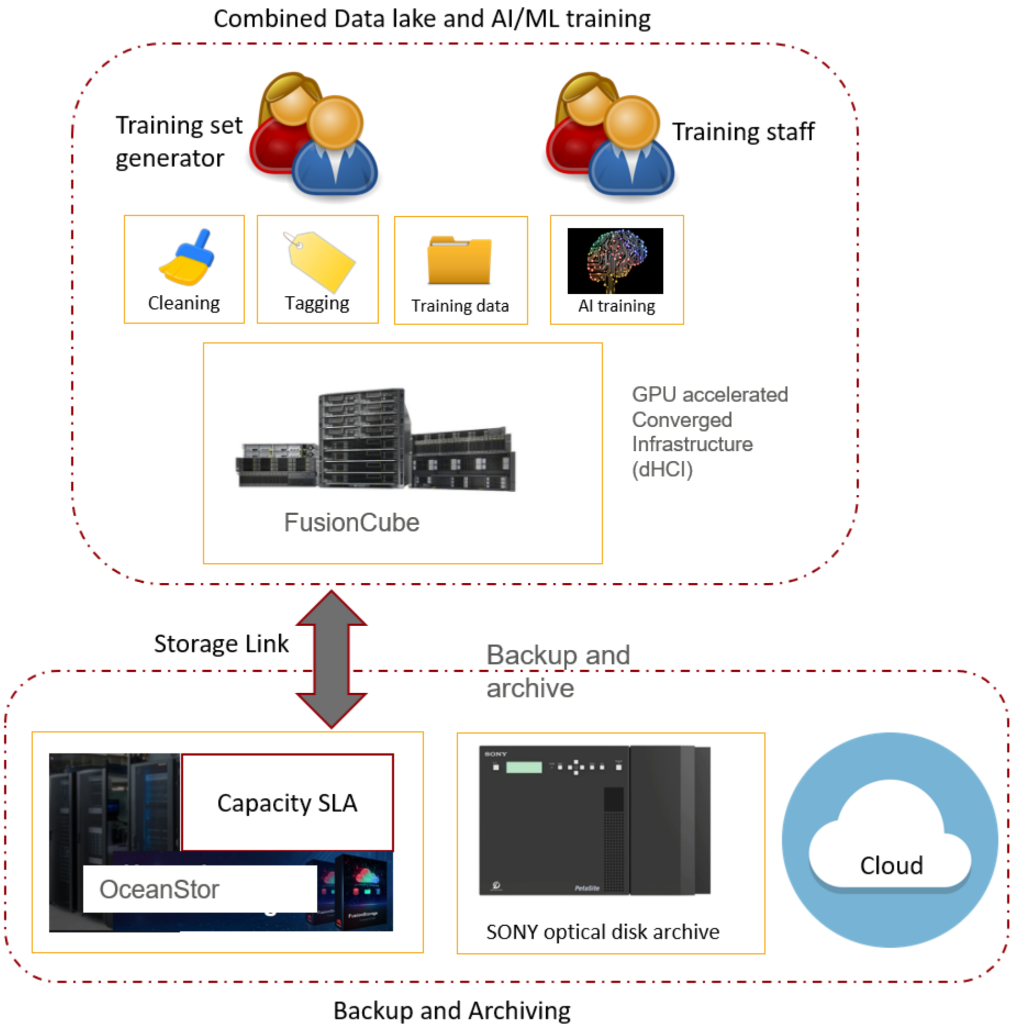


A wide range of applications in enterprise, industry, and academy can use this solution, such as hospitals for AI-assisted diagnosis, advertising companies for offline analysis, automotive industry for ADAS training, etc. See chapters 3 and 4 for more ideas.

4.5 Entry-level AI Cluster / Remote Office

Some AI applications do not need very large scale to work, or by nature deployed in a distributed way with multiple remote sites that each needs to be able to perform AI tasks on their own. For such kinds of small-scale AI applications, flexibility, convenience, and low cost are frequently the most important factors when clients choose their hardware solution. HCI or dHCI solutions should be very good candidates for such scenarios. For example, in the case of HCI, a small HCI cluster with the following characteristics can be a very good candidate:

- 2-3 nodes redundant HCI cluster taking half-rack or shipped as a pre-configured box
- Each node contains 1-4 GPU/TPU
- Key enterprise-grade storage features built-in the HCI stack, such as snapshot, clone
- Inter-node communication uses 10/25Gbps ethernet with optional RoCE configuration
- Integrated management interface for storage and compute



Example of such scenario includes a remote office for financial and insurance companies, life science research labs, medical institute field offices, etc. See chapters 3 and 4 for more market opportunity ideas.

4.6 GDS

4.6.1 Introduction

Nvidia® Corp added GPUDirect Storage (GDS) support to establish storage to GPU data pipes when reading data into Nvidia® GPUs. More details of GDS can be found at the Nvidia® website https://developer.Nvidia.com/blog/gpudirect-storage/

This feature is widely welcomed by the industry as a possible solution to the IO bottleneck of GPU solutions. Using GDS, the CPU utilization ratio will be reduced, and storage bandwidth to GPU will be increased. In this white paper, we used NVMe drives to test local and remote NVMe access.

We tested several GDS modes and compare the results to show GDS could help increase IO performance in some scenarios.

4.6.2 Settings

In the test, we used two Linux hosts, one serves as the computer with a GPU card, and another serves as remote storage. The detailed installation guide can be found at Nvidia® website: https://docs.Nvidia.com/gpudirect-storage/troubleshooting-guide/index.html

The next table shows details of the configuration.

Config	*Value*
Server Model	Dell R740
RAM	
GPU card	One Tesla T4
Network card	Mellanox CX-4
Operating System	Ubuntu 20.14
Mellanox OFED	5.1
Nvidia® Driver	
Nvidia® GDS version	v0.9 beta
Nvidia® CUDA version	11.0
Nvidia® cuDNN version	8.0
Nvidia® NCCL version	2.7.8

4.6.2.1 Test steps

In our tests, we used the open-source Ubuntu operating system, so we followed steps to install DEB packages.

1. Install CUDA (https://docs.Nvidia.com/cuda/cuda-installation-guide-linux/index.html)
2. Install cuDNN (https://docs.Nvidia.com/deeplearning/cudnn/install-guide/index.html)
3. Install NCCL (https://docs.Nvidia.com/deeplearning/nccl/install-guide/index.html)
4. Install Mellanox OFED driver

As requested by GDS (see details in Nvidia® GPUDirect Storage Installation and Troubleshooting Guide https://docs.Nvidia.com/gpudirect-storage/troubleshooting-guide/index.html), the OFED driver installation must contain special options, with-nvmf, with-nfsrdma, enable-gds, and add-kernel-support.

5. Configure IPoIB interface and validate RDMA & network bandwidth

```
root@dsw10:~# ip link | grep ibs
6: ibs5f0: <NO-CARRIER,BROADCAST,MULTICAST,UP> mtu 4092 qdisc mq state DOWN mode DEFAULT
group default qlen 256
7: ibs5f1: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 2044 qdisc mq state UP mode DEFAULT group
default qlen 256

root@dsw10:~# ethtool ibs5f1
Settings for ibs5f1:
        Supported ports: [ ]
        Supported link modes:   Not reported
```

```
        Supported pause frame use: No
        Supports auto-negotiation: No
        Supported FEC modes: Not reported
        Advertised link modes:  Not reported
        Advertised pause frame use: No
        Advertised auto-negotiation: No
        Advertised FEC modes: Not reported
        Speed: 40000Mb/s
        Duplex: Full
        Port: Other
        PHYAD: 0
        Transceiver: internal
        Auto-negotiation: off
        Link detected: yes
```

Add IP address to the 40g IB interfaces on both server:

```
root@dsw10:~# ip addr show dev ibs5f1

7: ibs5f1: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 2044 qdisc mq state UP group default qlen
256

    inet 10.40.1.1/8 brd 10.255.255.255 scope global ibs5f1

root@dsw11:~# ip addr show dev ibs5f1

7: ibs5f1: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 2044 qdisc mq state UP group default qlen
256

    inet 10.40.1.2/8 brd 10.255.255.255 scope global ibs5f1
```

```
Start OpenSM to enable IP service:

    root@dsw11:~# systemctl start opensm

Ensure ping works:

root@dsw11:~# ping 10.40.1.1
PING 10.40.1.1 (10.40.1.1) 56(84) bytes of data.
64 bytes from 10.40.1.1: icmp_seq=1 ttl=64 time=0.947 ms
```

Run RDMA bandwidth test:

```
Run server on 10.40.1.1:
ib_send_bw -d mlx5_1 -i 1 -F --report_gbits
Run client on 10.40.1.2:
ib_send_bw -d mlx5_1 -i 1 -F --report_gbits 10.40.1.1
```

Result:

```
---------------------------------------------------------------------------------------
                    Send BW Test
 Dual-port       : OFF          Device         : mlx5_1
 Number of qps   : 1            Transport type : IB
 Connection type : RC           Using SRQ      : OFF
 PCIe relax order: ON
 ibv_wr* API     : ON
 TX depth        : 128
 CQ Moderation   : 1
 Mtu             : 4096[B]
 Link type       : IB
 Max inline data : 0[B]
 rdma_cm QPs     : OFF
 Data ex. method : Ethernet
```

```
---------------------------------------------------------------------------------------
 local address: LID 0x01 QPN 0x0912 PSN 0xaed5a4
 remote address: LID 0x03 QPN 0x0912 PSN 0xf024ff
---------------------------------------------------------------------------------------
 #bytes     #iterations    BW peak[Gb/sec]    BW average[Gb/sec]   MsgRate[Mpps]
 65536      1000             38.26              38.26              0.072979
---------------------------------------------------------------------------------------
```

6. Install GDS packages

Follow the steps of GDS release notes (https://docs.Nvidia®.com/gpudirect-storage/release-notes/index.html#install-gds)

7. Ready to test GDS performance

Change GDS settings in /etc/cufile.json for different test options.

4.6.3 Local NVMe drive setting

The basic setting of GDS is to use local NVMe drives on the host.

```
/usr/local/cuda-11.0/gds/tools# ./gdscheck -p
 GDS release version (beta): 0.9.0.18
 Nvidia®_fs version:  2.3 libcufile version: 2.3
 cuFile CONFIGURATION:
 NVMe           : Supported
```

1. Mount local NVMe drive in ordered mode, mount -t ext4 -o data=ordered /dev/nvme0n1p1 /mnt
2. gdsio utility benchmark

```
/usr/local/cuda/gds/tools# ./gdsio -f /mnt/test/testfile1 -d 0 -w 4 -s 10G -i 1M -I 0 -x 0
IoType: READ XferType: GPUD Threads: 4 DataSetSize: 10212352/10485760(KiB) IOSize: 1024(KiB)
Throughput: 2.681944 GiB/sec, Avg_Latency: 1456.483233 usecs ops: 9973 total_time 3.631417 secs
```

The result shows that a local NVMe drive can provide <3GB/s bandwidth.

4.6.4 Lustre setting

GDS has a mode that uses Lustre cluster as remote storage. Lustre can be installed as described in https://wiki.lustre.org/Installing_the_Lustre_Software and

https://doc.lustre.org/lustre_manual.pdf. Test based on steps in https://wiki.lustre.org/Testing_HOWTO

4.6.5 NVMeOF setting

NVMe over Fabric (NVMeOF) is an industry standard (https://nvmexpress.org/wp-content/uploads/NVMe_Over_Fabrics.pdf) that allows NVMe protocol to be used on remote links. The NVMeOF support is already in Linux kernel.

In this setting, a host with NVMe drives is configured as NVMeOF target to provide storage. Then an EXT4 file system can be built onto it to provide file service. The result shows that a remote NVMe target provides a bandwidth close to local NVMe devices.

1. Edit /etc/cufile.json. Change IP address in properties.rdma_dev_addr_list.
2. modprobe nvmet
3. modprobe nvmet-rdma
4. Setup NVMeOF target (https://community.mellanox.com/s/article/howto-configure-nvme-over-fabrics#jive_content_id_NVME_Target_Configuration)
5. Setup NVMeOF Client to validate IO path (https://community.mellanox.com/s/article/howto-configure-nvme-over-fabrics#jive_content_id_NVMe_Client_Initiator_Configuration)
   Additional step before client can connect: nvme gen-hostnqn > /etc/nvme/hostnqn

   Discover log success:
   ```
   root@dsw10:~# nvme -discover -t rdma -a 10.40.1.2 -s 4420

   Discovery Log Number of Records 1, Generation counter 2
   =====Discovery Log Entry 0======
   trtype:  rdma
   adrfam:  ipv4
   subtype: nvme subsystem
   treq:    not specified, sq flow control disable supported
   portid:  1
   trsvcid: 4420
   subnqn:  target1
   traddr:  10.40.1.2
   rdma_prtype: not specified
   rdma_qptype: connected
   rdma_cms:    rdma-cm
   rdma_pkey: 0x0000
   ```
6. Baseline benchmark
7. Test GDS
   ```
   /usr/local/cuda/gds/tools# ./gdsio -f /nofmnt/testnof -d 0 -w 4 -s 10G -i 1M -I 0 -x 0
   IoType: READ XferType: GPUD Threads: 4 DataSetSize: 1016832/1048576(KiB) IOSize: 1024(KiB)
   Throughput: 2.236000 GiB/sec, Avg_Latency: 1743.542703 usecs ops: 993 total_time 0.433688 secs
   ```

8. Proposal for Training Infrastructure Based on NVMeOF

Based on the experiments above, we propose the following layout for a highly efficient AI training infrastructure using a NVMeOF storage working with multiple multi-GPU compute nodes to scale up to a powerful training platform.

On a NVMeOF storage array, a read only LUN will be and provisioned exposed to compute as input for their training. All nodes can access this shared set of data directly over fabric, saving data copy and improve efficiency. In addition, a set of writable LUNs can be provisioned for each compute node so they can use them to store their interim data.

This approach offers raw block access to all compute nodes over RDMA fabric, directly consumable to all GPUs through GDS APIs, with the best possible performance and least number of intermediate layers. Meanwhile, this design can easily leverage all advanced features offered by an enterprise storage array. We consider this architecture to be the most ideal design for customers with the highest performance and feature demands.

4.6.6 NFSoverRDMA Setting

NFSoverRDMA is a configuration that allows NFS protocol to run on top of RDMA. Compared to NFS over TCP, NFS over RDMA can provide higher bandwidth. In this example, we used the NFS server provided by the Linux kernel.

Please refer to the Nvidia® GPUDirect (https://docs.Nvidia.com/gpudirect-storage/troubleshooting-guide/index.html#nfs-suppport-gds)

1. Setup NFS server (https://docs.Nvidia.com/gpudirect-storage/troubleshooting-guide/index.html#install-nfs-server-rdma-mofed_5-1)
2. Setup NFS client (https://docs.Nvidia.com/gpudirect-storage/troubleshooting-guide/index.html#install-gds-supp-nfs-client)

```
mount -v -o proto=rdma,port=20049,vers=3 IPADDR:/ExportDir /mnt/nfs_rdma_gds
```

3. Test GDS

   Currently “gdscheck -p” shows that NFS is unsupported.

   We can work around it by setting cufile.json. properties.allow_compat_mode to true.

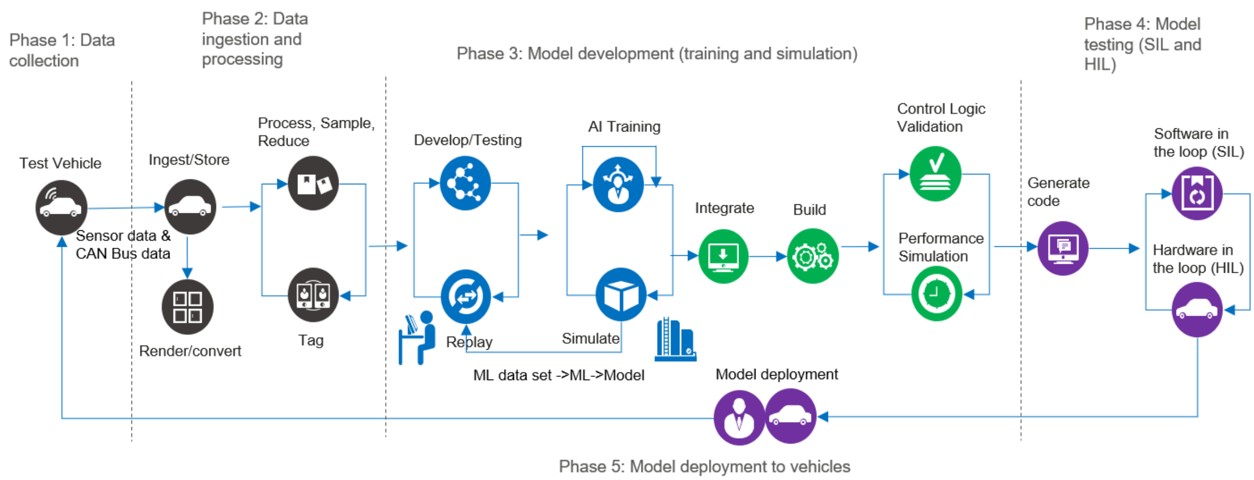
5 Selected Use Cases

In this chapter, we use several use cases to showcase different stages and the infrastructure challenges in each use case.

5.1 ADAS

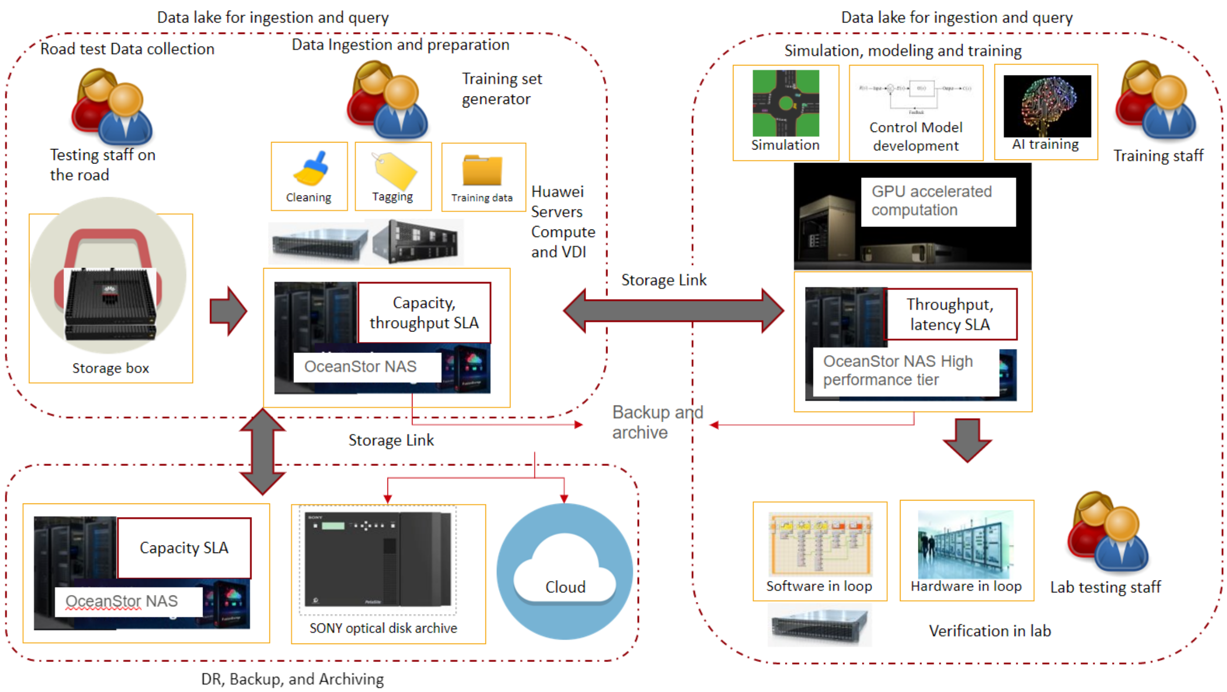
The ADAS development process can be divided into the following phases:

1. Data collection
2. Data ingestion and processing
3. Model development with training and simulation
4. Model testing (SIL and HIL)
5. Model deployment to vehicles



First, data is collected through a fleet of vehicles traveling on target terrain. Data may include time sequence data from sensors, cameras, lidar, and positioning systems. The estimated size of data may reach 60TB per car each day. Second, the collected data is processed to have a clean format and tagged information for machine learning usage. Today the tagging process can be automated but there is still some work left for humans. The third phase, model development with training and simulation, is the core piece of the ADAS infrastructure. The training set is fed into neural networks to detect objects, project trajectories, and plan actions of the vehicle. Computers can use real road data and simulated road data to train the model until the model reaches a satisfactory goal. The companies that can train the model fast and accurately will provide better products and shorter go-to-market turnaround times. In the fourth phase, the trained model will be deployed to labs to have software-in-the-loop (SIL) and hardware-in-the-loop (HIL) testing. SIL testing is performed in a software-based environment, whereas HIL testing emulates the real car model for control systems. In the fifth phase and the final phase, the polished model is deployed on the autonomous vehicle and it can generate more data in future road tests.

Each phase of the process has a different requirement for storage. It is well known that each enterprise has different goals so that the sizing of the capacity could be very different among different car manufacturers. The more data acquired, the more demanding the system will become.



Phase 1 needs on-vehicle equipment to store data from all sensors. There are other requirements for shock resistance and temperature.

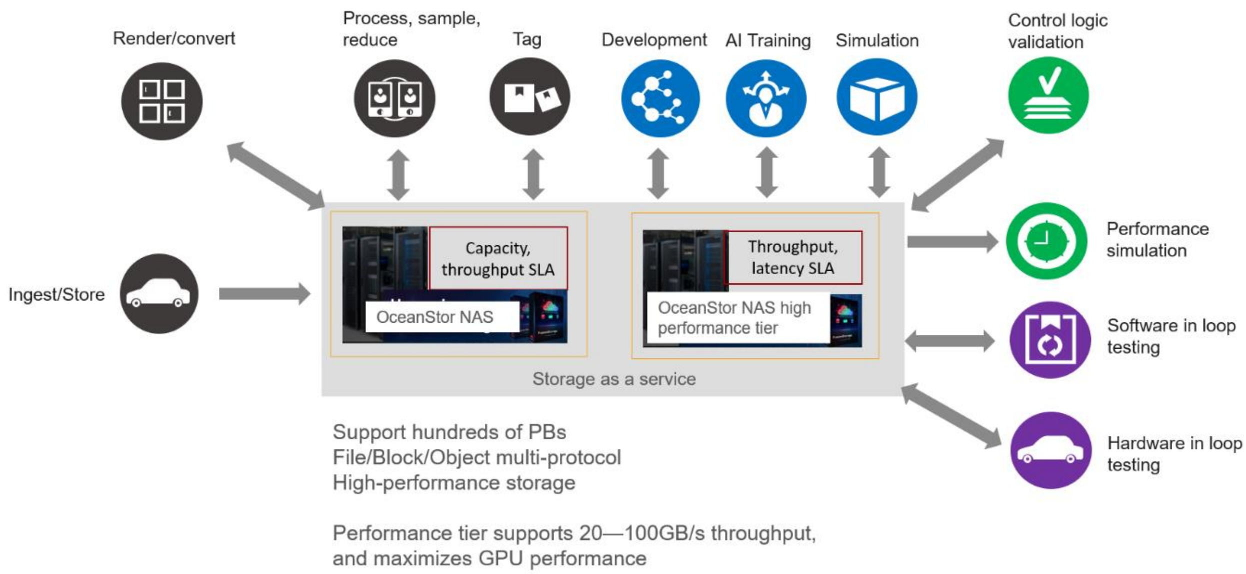
Phase 2 needs a big data processing system, which can be hundreds of PetaBytes. The actual capacity depends on the enterprise's business goals. The reference architecture defines such a system.

Phase 3 needs a training system that can handle millions of pictures, which can go into PetaBytes. Depending on the requirement of training time, the bandwidth of reading can be high. Nvidia® GDS is designed to provide high bandwidth to GPUs.

Phase 4 requires the SIL system to verify a large part of the scenario data (e.g., 30%) and write results to the storage system every day. All of these require a large bandwidth system. A high-performance storage cluster is needed to handle this workload.

Phase 5 does not have specific requirements for storage.

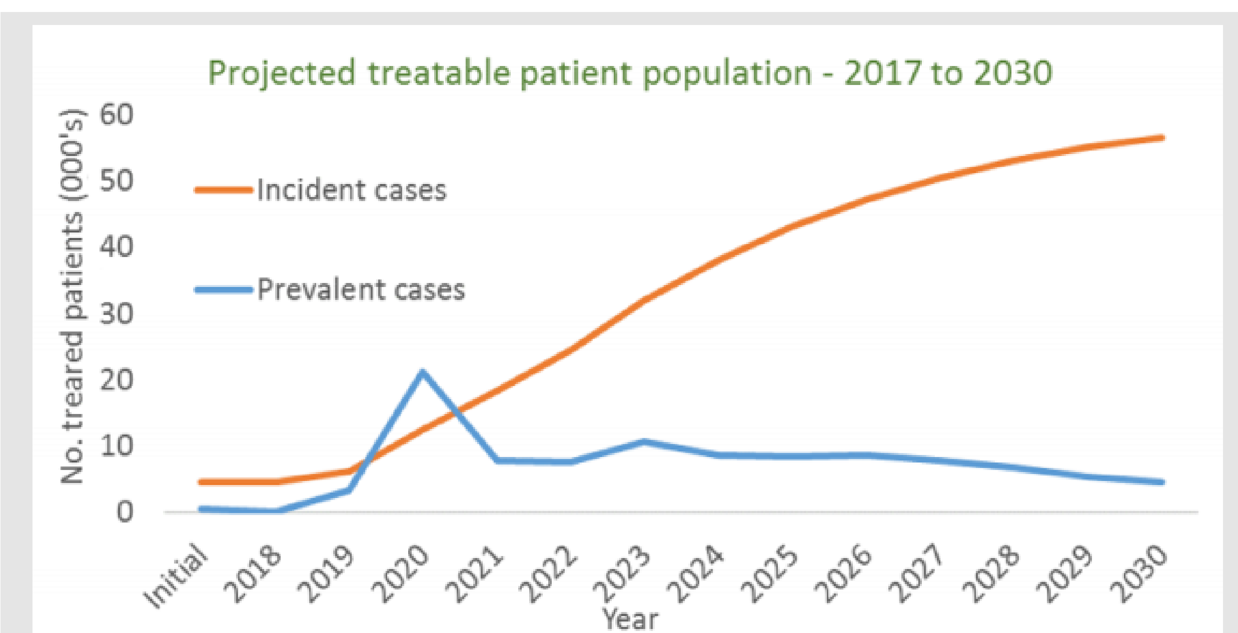
The storage service can be divided into two tiers. One tier is capacity sensitive, and the other tier is performance sensitive, particularly for the AI development environment. Details of each environment are based on customer requirements. One requirement is that the system should be scalable as the environment changes.



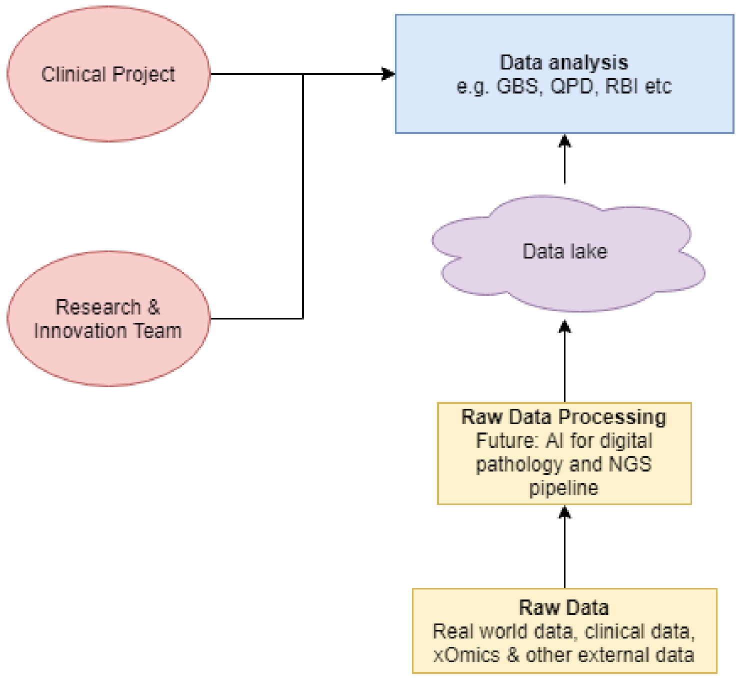
The previous diagram shows an environment using the OceanStor storage platform. The storage as a service layer provides a foundation for persistent storage. The reference architecture can be used to implement the phases.

5.2 HEALTHCARE AND LIFE SCIENCE

AI found extensive use in modern medical and life science, throughout a wide range of areas such as drug discovery, disease study, diagnoses, etc. The next 10 years will likely see significant changes in the US healthcare system with major improvements in the treatment paradigms for numerous diseases that previously had high morbidity and mortality. These will require the overall system to adapt, particularly in terms of how treatments are reimbursed and financed, as we move from chronic palliative therapies to acute curative ones. Below is predicted genetic treatment market growth [36]:



The picture below shows a typical life science research workflow.



There are a few main pain points in the life science AI applications, the first is capacity.

Genetic research had multiple breakthroughs in recent years and has taken off in medical applications such as cancer diagnosis and treatment. The genetic applications market is forecasted to grow explosively in the coming years.

Quoting data from strand-NGS v2.9 (A popular genetic analysis software), here is the computing and storage requirements for analyzing a single genome:

The storage requirement for a single human genome:

	Coverage	No. of Reads	Read Length	BAM File Size	Strand NGS Size
Whole Genome	37.7x	975,000,000	115	82 GB	104 GB
Whole Genome	38.4x	3,200,000,000	36	138 GB	193 GB
Exome	40x	110,000,000	75	5.7 GB	7.1 GB

The 1st phase is the data collection phase, the massive input data need to be somehow stored and made conveniently available for later processing, this is very similar to the ADAS application's data import phase. Most of the data during this phase will be written once, read many times, and the operations are mostly sequential I/O. Low-cost SDS with great scalability is very suitable for this phase. AI applications will consume data from this pool of data.

The 2nd phase is raw data processing, in this phase raw data will go through preliminary processing to get ready for analytic applications and tend to be stored in a more centralized place such as cloud or on-prem primary storage. The data will later be accessed more randomly and frequently, while the amount of data will be around a magnitude lower than raw data but still very large. This will require both high capacity and higher performance. A tiered storage solution with a high-end enterprise primary storage combined with a low-cost capacity tier will be suitable for this purpose. Flexible access protocol is a key value for this phase, as this data lake will be shared by all kinds of different applications.

The 3rd phase is where analytic applications do their job, huge AI computing power is required, as well as high-performance local storage for intermediate results. HCI cluster with SSD-backed local storage is suitable for most scenarios. During this phase, if the pre-processed data is accessible through some high-performance channel such as 100g NVMeOF, applications can directly load data from the data lake, otherwise, data need to migrate from the phase 2 storage to a high performance local or nearby high-speed storage tier.

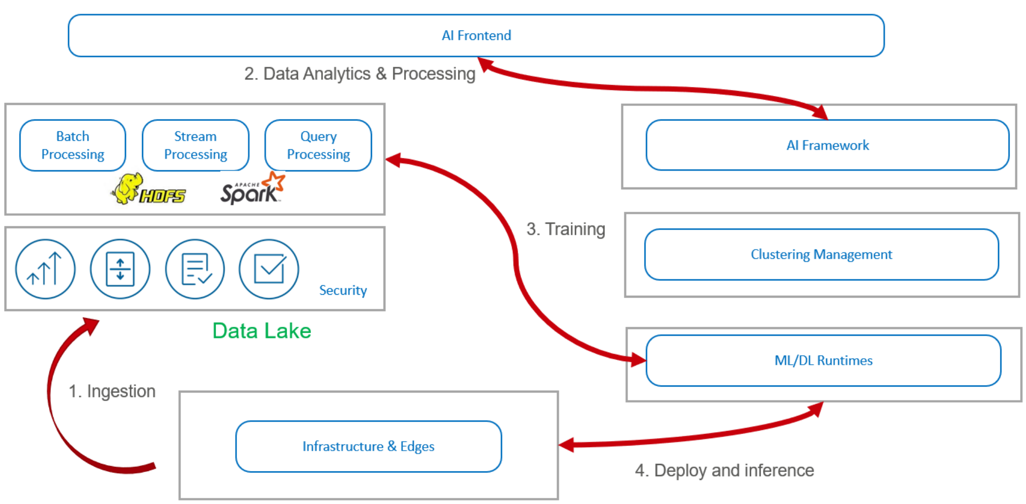
5.3 The Energy Sector

The offshore oil and gas industry has changed rapidly in recent years, with new technologies being adopted by the energy sector to meet the challenges of a digital economic landscape. Artificial intelligence is an exciting new technology field and can be applied to the oil & gas industry to save substantially by simulating real project scenarios.

There are several phases for oil & gas company to explore the capabilities of AI:

1. Data collection and ingestion phase
2. Data analytics and processing
3. Training
4. Central and distributed model deployment and inference

To harvest the capabilities of AI/ML, a large amount of data is needed for training purposes. It is also desirable to build a big data lake for analytics and AI/ML. Phase 1 generally is for this purpose. Once data have been ingested by the data lake, a data analytics and processing data platform shall be provided to preprocess data and clean data for machine learning training purposes. Phase 3 is the actual training phase. At phase 4, either a centralized data model is deployed at a data center or a distributed model is deployed at the edge for inference purposes.



6 Conclusion

In this document, we first introduce the AI ecosystem and market trend. A complete AI solution should provide both the infrastructure and the AI pipelines built on top of the infrastructure. The infrastructure should be optimized based on the AI application needs.

Through these applications, we summarize the types of workloads. Although there is not a definitive workload benchmark for AI workloads, we listed several relevant benchmarks that can help the readers to benchmark an AI infrastructure. The benchmark selection will still be a subject being debated in the industry. Most benchmarks are recognized by the industry for many years. GDS benchmark, as a representative of newer benchmarks, is quite recent.

A general AI infrastructure is proposed. The architecture is open and based on open-source software. Based on each user's needs, the architecture can be evolved from low-end, mid-range, to high-end configurations. We also listed several possible GDS configurations.

Finally, we used several popular applications to show that the reference architecture is valid.

7 References

[1] The Linux Foundation Projects, "Horovod.ai," [Online]. Available: https://horovod.ai/.

[2] M. Turck, "Resilience and Vibrancy: The 2020 Data & AI Landscape," 30 Sept 2020. [Online]. Available: https://mattturck.com/data2020/.

[3] Grand View Research, "Customer Experience Management Market Size $23.6 Billion By 2027," Feb 2020. [Online]. Available: https://www.grandviewresearch.com/press-release/global-customer-experience-management-cem-market.

[4] [Online]. Available: https://www.law.com/legaltechnews/2020/08/12/brave-new-world-how-ai-tools-are-used-in-the-legal-sector/?slreturn=20210027160243.

[5] [Online]. Available: https://www.ibm.com/downloads/cas/NAJXEKE6.

[6] [Online]. Available: https://hbr.org/2020/11/ai-can-make-bank-loans-more-fair.

[7] [Online]. Available: https://www.idc.com/getdoc.jsp?containerId=EUR246532420.

[8] [Online]. Available: https://www.uipath.com/rpa/intelligent-process-automation.

[9] K. Lyons, "YouTube brings in $5 billion in ad revenue as Alphabet and Google® bounce back," 29 10 2020. [Online]. Available: https://www.theverge.com/2020/10/29/21531711/Google®-alphabet-ad-revenue-youtube-waymo-cloud-search.

[10] Grand View Research, "Smart Education And Learning Market Worth $680.1 Billion By 2027," Apr 2020. [Online]. Available: https://www.grandviewresearch.com/press-release/global-smart-education-learning-market#:~:text=Smart%20Education%20And%20Learning%20Market%20Worth%20%24680.1%20Billion

%20By%202027,-
April%202020%20%7C%20Report&text=The%20global%20smart%20education%20and,17.

[11] Grand View Research, "Smart Education And Learning Market Size, Share & Trends Analysis Report By Age, By Component (Hardware, Software, Service), By Learning Mode, By End User, By Region, And Segment Forecasts, 2020 - 2027," Apr 2020. [Online]. Available: https://www.grandviewresearch.com/industry-analysis/smart-education-learning-market.

[12] [Online]. Available: https://www.iflexion.com/blog/artificial-intelligence-real-estate.

[13] [Online]. Available: https://ash.harvard.edu/files/ash/files/artificial_intelligence_for_citizen_services.pdf.

[14] CIODIVE, [Online]. Available: https://www.ciodive.com/news/capital-one-breach-raises-questions-about-security-and-cloud-first-strategi/560129/.

[15] [Online]. Available: https://emerj.com/ai-sector-overviews/artificial-intelligence-in-insurance-trends/.

[16] [Online]. Available: https://www.bankrate.com/insurance/artificial-intelligence-meets-the-insurance-industry/.

[17] [Online]. Available: https://www.forbes.com/sites/robtoews/2020/08/26/ai-will-revolutionize-healthcare-the-transformation-has-already-begun/?sh=2f0c85c9722f.

[18] [Online]. Available: https://www.genengnews.com/magazine/314/ai-in-the-life-sciences-six-applications/.

[19] Grand View Research, "Intelligent Transportation System Market Size, Share & Trends By Type (ATIS, ATMS, ATPS, APTS, EMS), By Application (Traffic Management, Public Transport), By Region, And Segment Forecasts, 2020 - 2027," 2020. [Online]. Available: https://www.grandviewresearch.com/industry-analysis/intelligent-transportation-systems-industry#:~:text=Report%20Overview,5.8%25%20from%202020%20to%202027..

[20] Plunkett Research, "Transportation Industry Statistics and Market Size Overview, Business and Industry Statistics," 2019. [Online]. Available: https://www.plunkettresearch.com/statistics/Industry-Statistics-Transportation-Industry-Statistics-and-Market-Size-Overview/.

[21] Y. D. a. S. S. M. Piero Molino, "Introducing Ludwig, a Code-Free Deep Learning Toolbox," 11 Feb 2019. [Online]. Available: https://eng.uber.com/introducing-ludwig/.

[22] V. Panyam, "Introducing Neuropod, Uber ATG’s Open Source Deep Learning Inference Engine," 8 June 2020. [Online]. Available: https://eng.uber.com/introducing-neuropod/.

[23] A. S. a. M. D. Balso, "Meet Horovod: Uber’s Open Source Distributed Deep Learning Framework for TensorFlow," Uber, 17 10 2017. [Online]. Available: https://eng.uber.com/horovod/.

[24] A. Gale, "Introducing Flyte: A Cloud Native Machine Learning and Data Processing Platform," 7 Jan 2020. [Online]. Available: https://eng.lyft.com/introducing-flyte-cloud-native-machine-learning-and-data-processing-platform-fb2bb3046a59.

[25] C. Donnelly, "Uber backs hybrid cloud as route to business and geographical expansion," 7 11 2018. [Online]. Available: https://www.computerweekly.com/news/252452059/Uber-backs-hybrid-cloud-as-route-to-business-and-geographical-expansion#:~:text=To%20achieve%20this%20scale%2C%20Uber,storage%2C%20database%20and%20GPU%20workloads..

[26] [Online]. Available: https://www.forbes.com/sites/cognitiveworld/2019/07/05/how-ai-is-transforming-agriculture/?sh=616bd49c4ad1.

[27] "SPC-1 Benchmark," [Online]. Available: http://spcresults.org/benchmark-results-spc1.

[28] "IO500 Benchmark," [Online]. Available: https://www.vi4io.org/io500/start?fields=information__system,information__institution,information__storage_vendor,information__filesystem_type,information__client_nodes,information__client_total_procs,io500__score,io500__bw,io500__md,information__data,inf.

[29] "FIO documentation," [Online]. Available: https://fio.readthedocs.io/en/latest/.

[30] "FIO project source," [Online]. Available: https://github.com/axboe/fio.

[31] "FIO script examples," [Online]. Available: https://github.com/axboe/fio/tree/master/examples.

[32] "GPU Direct Storage Guide," [Online]. Available: https://docs.Nvidia.com/gpudirect-storage/troubleshooting-guide/index.html.

[33] "Nvidia® GPU Wiki," [Online]. Available: https://en.wikipedia.org/wiki/List_of_Nvidia_graphics_processing_units.

[34] "Tesla GPU Wiki," [Online]. Available: https://en.wikipedia.org/wiki/List_of_Nvidia_graphics_processing_units#Tesla.

[35] "AI Benchmark," [Online]. Available: https://pypi.org/project/ai-benchmark/.

[36] [Online]. Available: https://www.forbes.com/sites/robtoews/2020/08/26/ai-will-revolutionize-healthcare-the-transformation-has-already-begun/?sh=2f0c85c9722f.

[37] D. H. Lei. [Online]. Available: https://drive.Google.com/file/d/12JY_mIdBMFzaqP-lppHkUdjgWxMVRBXu/view.

[38] [Online]. Available: https://arxiv.org/pdf/1910.05930.pdf.