run:ai

MLOps: Do You Have the Hardware to Make AI Work?

Five Infrastructure Components Essential to MLOps Success

As organizations progress along the maturity model and scale their data science initiatives, they must embrace a strategic approach to their AI infrastructure stack - from AI accelerators, to storage and file systems, networking and the many tools available for modeling and building data pipelines... The stack also straddles the entire AI process, which starts with exploration and training, and then moves into production and inference.

In an on-premises AI infrastructure stack, all of the compute, data and machine learning/deep learning (ML/DL) components and resources are deployed and managed in-house where researchers, ops professionals and IT work together to build, train, and deploy the ML/DL models.

When building an on-premise AI infrastructure stack, there are five main components that must be chosen carefully in order to ensure scalability and high performance:

- **Hardware accelerators** optimized for high-speed, parallel computing. GPUs (graphic processing units) currently dominate the AI landscape.
- **A distributed data storage solution** optimized for fast throughput that can keep up with the GPUs' virtually insatiable demand for data.
- **A networking infrastructure** that delivers very high performance and ultra low latency across a lossless fabric.
- **A set of data science tools** to support the development and deployment of ML/DL-powered applications and services.
- **An orchestration platform** that stitches together all components of the stack and seamlessly synchronizes and manages resources and jobs.

Whether harnessing machine learning for business intelligence or to build AI-powered products and services, MLOps teams must learn how to deploy a purpose-built infrastructure stack that accelerates (rather than constrains) data science initiatives.

Introduction

Data science and digital transformation go hand in hand. Big data mining, data analytics, artificial intelligence (AI), machine learning (ML), and deep learning (DL) are driving a wide range of business-critical use cases—from predictive maintenance and customer behavior analysis to personalized marketing and data-driven business intelligence. According to Statista, "by 2023, digitally transformed organizations are forecast to contribute to more than half of global GDP."

The path to AI maturity starts with exploration and forming a data science team, and it runs through a series of standardizations and optimizations until achieving repeatable production at scale. For a thoughtful analysis of the benchmarks that lead to transformative data science value, see our ebook Ready for Success with AI? A Benchmarking Model for AI Infrastructure Maturity.

As organizations progress along the maturity model and scale their data science initiatives, they must embrace a strategic approach to their AI infrastructure. This guide provides a framework for a well-architected on-premises AI infrastructure stack.

An Overview of the Modern AI Stack

A stack is a set of software and, in some cases, hardware components that interoperate as a logical platform for running a service or application.

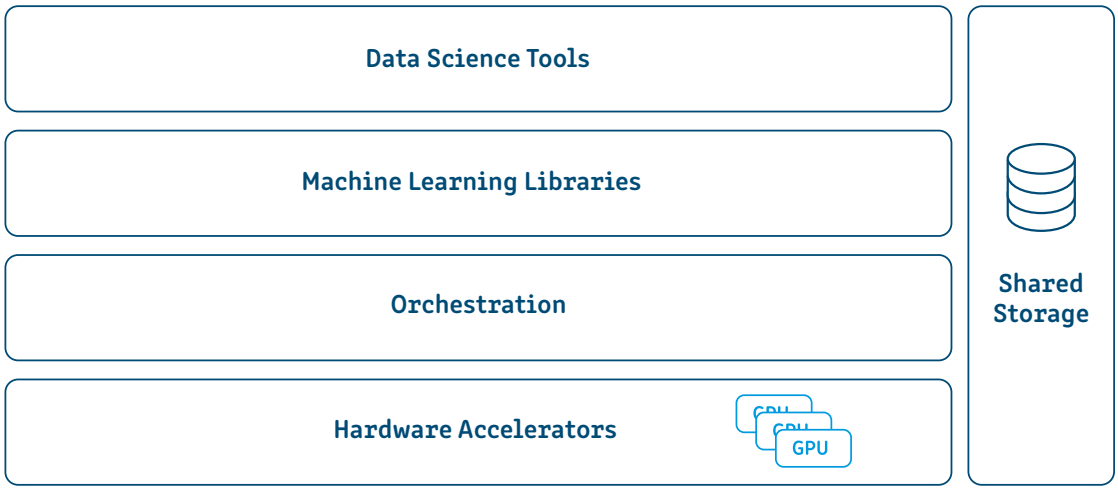

Figure 1: The AI Stack

As shown in Figure 1, the AI stack is comprised of several layers, starting with hardware at the bottom and moving up through:

- Hardware accelerators
- Kubernetes and orchestration layer to manage container access to accelerators and provide specific resource management and virtualization for GPUs
- Machine learning libraries and frameworks
- Data science tools for design and training of models, data management, versioning, model management, packaging, deploying and executing inference models, and more
- Storage built for model training and inferencing that interfaces with the data science tools, with the orchestration layer and with the hardware accelerators.

If we drill down to a model of the AI infrastructure stack specifically, it is comprised of three main layers:

- **Compute:** The computational power provided by physical and virtual servers, containers, and specialized hardware, such as GPUs.
- **Data:** All of the structured and unstructured data that is used to train AI-based models.
- **Machine learning algorithms and platforms:** The supervised and unsupervised algorithms used to train inference models and the frameworks that manage the machine learning process.

In the on-premise AI infrastructure stack, all the components are deployed and managed by the organization, who takes full responsibility for building, training, and deploying the ML/DL models (as opposed to the consumption of fully managed cloud services and platforms).

Considerations When Building Your AI Infrastructure Stack

When building an on-premise AI infrastructure stack, there are five main components that must be chosen carefully in order to ensure scalability and high performance:

- Hardware accelerators for enhanced compute capacity.
- A data storage solution that is optimized for fast throughput.
- A low-latency networking infrastructure.
- A set of data science tools.
- A platform that seamlessly synchronizes and manages resources and jobs.

In this section we provide guidelines for choosing the components that, together, comprise a well-architected and scalable infrastructure that accelerate your data science initiatives.

Hardware Accelerators for Compute-Intensive Operations

Hardware acceleration is the process by which compute-intensive tasks are offloaded onto specialized components that are optimized for high-speed, parallel computing. There are a number of hardware-based approaches to accelerate compute capabilities, including FPGAs (field-programmable gate array) integrated circuits, ASICs (application specific integrated circuits), and multiple-core CPUs. However GPUs (graphic processing units) are the hardware accelerators that currently dominate the AI infrastructure landscape.

As the name suggests, GPUs were originally designed for handling heavy graphical processing tasks, primarily in response to the needs of the video gaming industry. Today GPUs are also used for accelerating parallel calculations carried out on very large quantities of data, and new GPUs have been developed that are optimized specifically for deep learning.

GPUs, with their SIMD (single instruction, multiple data) architecture, are well-suited to deep learning processes, which require the same process to be performed for numerous data items. Since the introduction of the NVIDIA CUDA API framework in 2007, it has become easier for developers to implement GPU processing for computational and AI-related tasks. In addition, deep learning frameworks such as Pytorch and TensorFlow have emerged that abstract the complexities of programming directly with CUDA and make GPU processing even more accessible to modern data science implementations.

With their high bandwidth memory designed specifically for accelerating deep learning computations as well as their inherent scalability, GPUs support distributed training processes and, in general, can significantly speed up ML/DL operations. Table 1 summarizes the various GPU options to consider when building out your AI infrastructure stack.

Vendor & Model	Memory (GB)	Performance (teraflops)	Underlying Technologies
Consumer-Grade GPUs: Cost-effectively supplement existing systems; useful for model building or low-level testing.			
NVIDIA Titan V	12-32	110-125	Tensor Cores, NVIDIA's Volta
NVIDIA Titan RTX	24	130	Tensor and RT Core, NVIDIA's Turing GPU architecture
NVIDIA GeForce RTX 2080 Ti	11	120	Designed more for gaming enthusiasts than professional use
Data Center GPUs: The standard for production data science implementations; designed for large-scale projects and provide enterprise-grade performance			
NVIDIA A100	40	624	Multi-instance GPU (MIG) technology for massive scaling
NVIDIA V100	32	149	Volta technology, designed for HPC, ML, DL
NVIDIA Tesla P100	16	21	Based on Pascal architecture, designed for HPC and ML
NVIDIA Tesla K80	24	8.73	Based on the Kepler architecture, designed for data analytics, scientific computing

Table 1: GPU options for the AI infrastructure stack

In addition, NVIDIA DGX servers are enterprise-grade, full-stack solutions that can be deployed on bare metal servers or using containers. Table 2 summarizes their specifications.

Model	CPUs	GPUs	Underlying Technologies
DGX-1	Intel Xeon x 2	Up to 8 V100 Tensor Cores, each with 32GB memory	Based on the Ubuntu Linux Host OS. Includes the CUDA toolkit, NVIDIA's Deep Learning SDK, the Docker Engine Utility, and the DIGITS deep learning training application.
DGX-2	Intel Xeon Platinum x 2	V100 Tensor Cores x 16, each with 32 GB memory	Based on the NVSwitch networking fabric for 195x faster training than the DGX-1. Provides significant scalability and parallelism
DGX A100	AMD 64-core x 2	A100 x 8, with 320GB memory, 5 petaflops performance	Fully optimized for CUDA-X. Can combine multiple DGX A100 units to create a super cluster. Designed for ML training, inference, and analytics.

Table 2: NVIDIA DGX servers

Other significant players in the hardware acceleration market include AMD, with their Instinct™ MI series accelerators, as well as Graphcore, whose IPUs (Intelligence Processing Units) are designed from the ground up for AI compute. Intel® Deep Link technology makes the GPUs on its Intel® Core™ processors accessible for AI workloads. In addition, Habana Labs, an Intel company, offers a family of programmable deep learning accelerators for the data center.

Open Up Bottlenecks with Fast, Elastic Storage

High storage performance is critical to deep learning at scale and at velocity. One NVIDIA DGX-1 server alone is equivalent to a supercomputer in a box, providing 1 PetaFLOPS of performance.

One of the biggest challenges, therefore, in the AI infrastructure stack is a storage solution that is fast enough to keep up with the GPUs' virtually insatiable demand for data. The GPU servers can theoretically process tens of GBs of data per second, but their onboard storage is a performance bottleneck. The DGX-1, for example, has a bandwidth of 7.8GB/s, but its onboard storage of 4 SATA SSDs effectively reduces that throughput to approximately 2.2GB/s. Similarly, the DGX-1 can theoretically process 2 million random IOPS, but the local storage only provides 400K IOPS.

Another storage-related challenge is the data access pattern of AI workloads. Training models, for example, requires many reads of small files such as individual images or small documents. Each data access has to take into account the overhead of opening and closing a small file. Thus, low read latency and the ability to support many small read operations per second is also essential for keeping the GPUs fed with data.

Without a fast and agile storage solution, a lot of the GPU computing power goes to waste. The NVM Express® (Non Volatile Memory Express, or NVMe®) specification, managed by a non-profit industry consortium, was developed to address such bottlenecks. The NVMe architecture provides a more efficient, lower latency, and more scalable interface for host software to communicate with non-volatile memory across a PCI Express® (PCIe®) bus. However, onboard NVMe can still be a bottleneck. The DGX-2 has 30TB (8 x 3.84TB) of local NVMe storage, but it is not optimized to use that storage efficiently.

Another storage bottleneck to be considered is the need to share very large quantities of data across a GPU cluster. Given the limited available storage capacity of each local server, in large-scale DL projects it is often necessary to copy smaller chunks of the dataset to each server as needed. This can make cluster-wide consistency a real problem. The solution is to implement a file system in a large and expandable namespace, which is shared across all cluster nodes. In order to be able to support high levels of I/O performance, including meeting the high demand for metadata, the shared file system must be highly parallelized.

Our recommendation for storage in the AI infrastructure stack is to take advantage of the significant network connectivity of GPU nodes to implement a distributed and scalable storage architecture. NVIDIA's DGX servers, for example, provide up to 48GB/s of networking bandwidth through up to 8 100Gb ports. And there are a number of data storage management vendors that provide elastic storage solutions that are optimized for AI data pipelines, including:

- NetApp ONTAP AI: A proven architecture powered by NVIDIA DGX servers and NetApp cloud-connected, all-flash storage to simplify, accelerate, and integrate AI and DL data pipelines.
- Weka AI Reference Architecture: A combined solution based on NVIDIA DGX A100 servers and WekaIO's high-performance, scalable WekaFS file storage system. The solution accelerates the training of even the most demanding DL models.
- DDN (DataDirect Networks): DDN's scalable flash storage architecture consolidates data in one place and integrates with NVIDIA DGX A100 to accelerate data science processing speeds by up to 10x.
- PureStorage AIRI: A fully integrated AI-ready scalable infrastructure based on NVIDIA DGX A100 and Pure FlashBlade® file and object storage arrays hardware, as well as the NVIDIA DGX software stack and NVIDIA optimized containers.
- Excelero NVMesh: A hardware-agnostic software-defined block storage solution that enables shared NVMe across any network. It delivers low-latency, scalable centralized storage across any local or distributed file system.
- VAST Data: Its Universal Storage solution delivers exabyte-level scalability allowing enterprises to consolidate all data and applications onto a single scale-out flash tier.

Deploy Networks that Accelerate Training

There's no point in having GPU-based supercomputing power and a superfast storage solution if data transfer within the network is slow. InfiniBand and Ethernet are the two leading high-speed technology frameworks that support intelligent software-defined networking to deliver very high performance and ultra low latency across a lossless fabric.

Mellanox, which was acquired by NVIDIA in April 2020, is a leading provider of both InfiniBand and Ethernet ASICs and switches. Mellanox's high-performance Ethernet switches and network adapters run across an Ethernet Storage Fabric (ESF) that offers unique storage-aware features, such as support for a wide range of CPU-optimizing storage offload protocols (TCP, RDMA, NVMe over Fabrics, etc.), erasure coding, and encryption. Other benefits include support for flexible and scalable software-defined networking, zero packet loss reliability, and peer-to-peer communication among GPUs within the network.

In a case study where Mellanox's Open Ethernet 25/100GbE Spectrum switches were deployed in an enterprise-grade machine learning data center, Mellanox was shown to accelerate training by a factor of 6.5. Mellanox's Ethernet solutions are an integral part of the NetApp, Weka, and DDN AI infrastructure reference architectures described above. Your AI infrastructure stack must also include an enterprise-grade networking solution of this nature.

Machine Learning Libraries and Toolset

The AI infrastructure stack also includes the data science tools that support the development and deployment of ML/DL-powered applications and services. These frameworks seamlessly manage workflows and resources so that data scientists can focus on building, training, and deploying ML models. Some of the leading tools and frameworks in this ecosystem are:

- Tensorflow, originally developed by the Google Brain team, is an end-to-end open source platform of tools, libraries, and community resources for state-of-the-art ML applications. It is written in Python, C++, and CUDA, and works across multiple platforms.
- Apache Spark is a unified analytics engine for large-scale and fast data processing that runs on Hadoop, Apache Mesos, Kubernetes, standalone, or in the cloud.
- Pytorch allows fast, flexible experimentation and efficient production with a user-friendly front-end, scalable distributed training and performance optimization, and a wide range of tools and libraries.
- MLflow is an open source ML lifecycle management platform that integrates with data science tools—like Tensorflow, Apache Spark, and Pytorch—and seamlessly tracks and queries experiments, packages data science code in a platform-agnostic format, deploys ML models across diverse environments, and provides a central model repository.
- Kubeflow is a machine learning toolkit for the deployment of best-of-breed open source ML systems on Kubernetes for portable and scalable ML applications.

Automate and Orchestrate to Optimize Resources and Workflows

Given the scale and complexity of the AI infrastructure stack, automation and orchestration are essential to ensure that resources are optimized (i.e., not under- or over-utilized) and jobs are orchestrated automatically. As shown in Figure 2, Run:AI is an orchestration and virtualization layer that "glues" together the other areas of the AI infrastructure stack. It abstracts workloads from the underlying infrastructure, creating a shared pool of resources that can be dynamically provisioned.

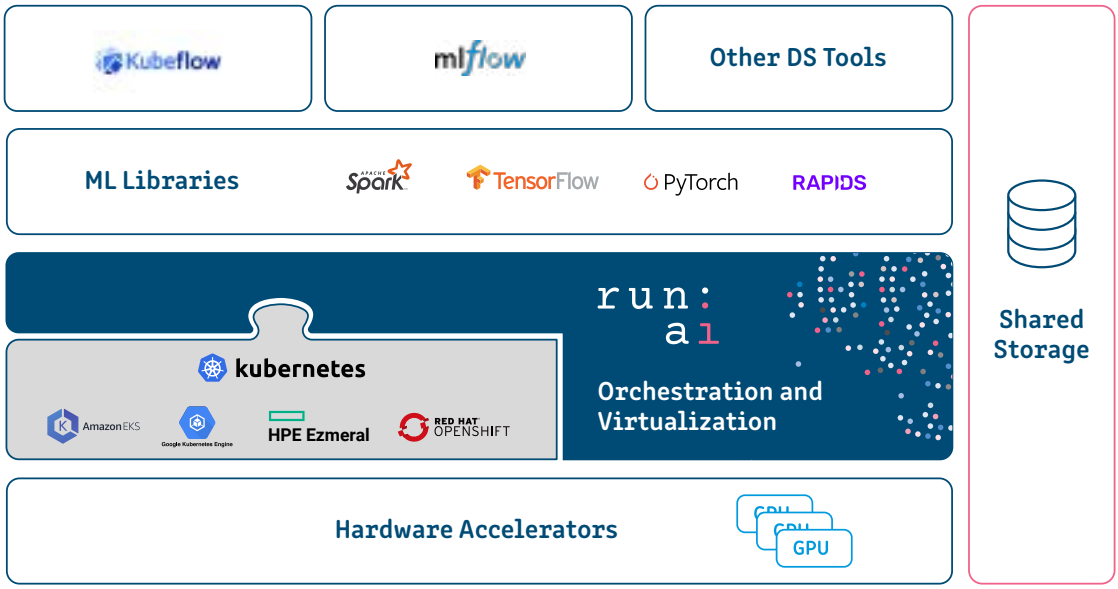

Figure 2: Run:AI orchestrates the AI infrastructure stack

Run:AI simplifies AI infrastructure pipelines, helping data scientists optimize expensive compute resources, accelerate their productivity, and improve the quality of their models. Some of the capabilities you gain when using Run:AI include:

- **Advanced visibility:** With advanced monitoring and cluster management tools, you can see which GPU resources are not being used and dynamically adjust the size of a job to run on available capacity.
- **No more bottlenecks:** You can set up guaranteed quotas of GPU resources, to avoid allocation bottlenecks and optimize billing. By dynamically allocating pooled GPU to workloads, hardware resources are shared more efficiently.
- **Higher level of control:** Run:AI enables you to dynamically change resource allocation, ensuring each job gets the resources it needs at any given time. For example, large ongoing workloads can be scheduled to run during low-demand times, allowing shorter, higher-priority workloads to run simultaneously during peak times.

A real-life example can be found in The London Medical Imaging & AI Centre for Value Based Healthcare. The Centre uses medical images and electronic healthcare data to train sophisticated deep learning computer vision and natural language processing algorithms for more effective screening, faster diagnosis, and more personalized therapies. Their AI infrastructure stack included on-premises DGX-1s and DGX-2s. While overall GPU utilization was below 30%, there were times when not enough GPUs were available for running jobs. Scheduling issues also arose, with bigger experiments that required a large number of GPUs not being able to begin because smaller jobs using only a few GPUs were blocking them from the required resources.

When Run:AI was deployed, along with NetApp storage and Mellanox networking, over the next forty days GPU utilization increased by 2.1x, experiments ran 31x faster, elastic workloads eliminated delays and bottlenecks, and 1.85x more experiments could be run.

Conclusion

To compete successfully in today's economy, companies must innovate faster and continuously deliver value to their customers and partners. These business-critical demands are the core drivers of digital transformation and data science initiatives.

Whether using big data for enhanced business intelligence, personalized marketing, or AI-powered products and services, Deep Learning researchers and the operations teams that support them must learn how to deploy a purpose-built infrastructure stack that accelerates data science initiatives. If any component(s) in the ecosystem create a bottleneck—compute power, storage performance, network latency—these initiatives will not be able to achieve their full transformative potential.

To learn more how Run:AI can optimize your deep learning workflows, sign up for a free trial.