TechTarget® Custom Media
White Paper

Top Considerations When Building Out Your AI/ML Infrastructure

intequus®

Many businesses are now experimenting with artificial intelligence applications such as machine learning and deep learning, yet few have fully integrated AI into business and IT operations.

This is changing, though, as AI enters the mainstream: IDC expects 90% of new enterprise applications will use AI by 2025.[1] The productivity-enhancing benefits of intelligent chatbots, predictive maintenance and intelligent business process optimization are too compelling for organizations to ignore any longer.

Up until now, many organizations have encouraged developers to get their feet wet with AI by designating cloud sandboxes for experimentation. The approach is acceptable for investigation, as cloud platforms are easily accessible and support various frameworks and tool sets. However, operationalizing AI requires greater discipline and business alignment. This creates a compelling argument for centering AI development within the data center.

Performance-optimized AI development is best done with purpose-built hardware and software similar to that used in high-performance computing environments.

Getting Serious About AI

Projects that will be deployed into production must be scrutinized for relevance to the company's strategic imperatives and business goals. Approval processes and criteria for moving projects into production pipelines need to be put in place. A cost-benefit analysis should be conducted and policies reviewed or implemented in such areas as compliance, privacy and security. Time frames and budgets must be established. Metrics should be defined in areas such as developer productivity and accountability.

Most cloud development platforms don't provide the level of detailed administrative oversight that is required to operationalize AI. These platforms are initially a cost-effective way to secure quick access to general-purpose infrastructure, but costs can mount up over time. Performance-optimized AI development is best done with purpose-built hardware and software similar to that used in high-performance computing environments.

The companies that build servers understand this and are responding with integrated systems, including right-sized processors, memory and storage designed specifically for AI workloads. These systems are optimized for large-scale in-memory processing and come with high-performance storage, high-speed backplanes and multiple graphics processing units (GPUs). The demand for hardware optimized for AI tasks is expected to grow from $20 billion in 2018 to $234 billion by 2025.[2] Crunchbase lists 17 startups that are building AI-specific hardware.

[1] IDC FutureScape Outlines the Impact 'Digital Supremacy' Will Have on Enterprise Transformation and the IT Industry," IDC, Oct. 29, 2019

[2] "Revenues From the AI-Driven Hardware Market Worldwide From 2018 to 2025," Statista, April 2019

Systems built to process multi-GPU workloads can reduce AI model training times by up to 30% compared with those running only on CPUs. Purpose-built AI systems are also increasingly being designed to work with specific programming languages and frameworks, many of which are already containerized for rapid execution. The more demanding the application, the more value can be realized from custom-tailored hardware.

What's Special About AI Hardware

AI training models can be extensive and may require multiple days to process. The use of purpose-built hardware optimizes memory, bandwidth, compute and storage performance to avoid leaving processors idle.

AI systems are typically more complex than off-the-shelf servers, often employing a combination of CPUs, GPUs, field-programmable gate arrays (FPGAs) and application-specific integrated circuits (ASICs). Parallel coprocessors provide higher data access throughput than CPU clusters. Referred to as heterogeneous computing, these systems separate the processing clusters used for analytics and AI from data storage systems. Accelerated data access enables cost savings on both business intelligence and data science workloads.

Despite such infrastructure complexity, the functionality of AI training models is usually more straightforward than that of conventional data processing applications. AI models work best when the algorithm itself is simple and the volumes of data are significant. As a result, AI servers are optimized to orchestrate many parallel tasks.

GPUs are vital to AI processing because they are built for sizable parallel processing tasks and can handle many process threads simultaneously. The amount of memory on AI-targeted GPUs is also important to allow for larger data sets to be loaded into GPU memory. NVIDIA's NVlink high-speed GPU interconnect allows multiple GPUs to

be linked together, and remote direct memory access (RDMA) enables workloads to be moved directly from one GPU's memory to another to increase performance and decrease the time to a solution. GPUs have many processing cores, while CPUs have relatively few. CPUs are thus better suited to complex tasks, while GPUs are better at repetitive ones.

Factors to Consider When Building AI/Machine Learning Infrastructure

Here are the major hardware components of a system designed for AI workloads and how they impact results.

CPUs

As noted above, most AI workloads require significant amounts of parallel processing. Conventional pipelined CPUs are overkill for parallel processing, so those tasks are better offloaded to coprocessors. These algorithms are best done with multiple GPUs. They offload processing from CPUs and excel at parallel data processing using thousands of small cores. The role of the CPU is to piece together results, a task for which it is well suited.

AI also can use low-precision computing, such as floating-point or integer calculations, which can be significantly faster than CPUs. Again, CPUs are overkill for these tasks because they waste cycles by providing more precision than needed.

> AI also can use low-precision computing, such as floating-point or integer calculations, which can be significantly faster than CPUs.

Microprocessor makers are increasingly building AI-specific functions into off-the-shelf processors. For example, Intel's Xeon line now comes with a group of acceleration features that boosts performance on popular deep learning frameworks for inference processing. Nevertheless, large-scale training models will always work best with parallel GPUs.

FPGAs may be used for parallel processing of specific instructions. Specialty AI processors may also have other features that are optimized for machine learning and deep learning workloads, including parallel processing of large volumes of calculations in a specified sequence and coordinated computations among accelerators.

Memory

AI models use a lot of memory, but not all are the type that is common in standard CPU architectures. Model training requires direct and fast memory to feed relatively simple applications built for repetitive execution. GPU memory and GPUDirect RDMA bypass the CPU and main memory for the best performance.

Conventional computing architectures are too complex and thus may fail to provide the needed performance.[3] For example, the multilevel cache architectures used in popular CPUs aren't necessary for AI training models because there is no need to reuse data.

Extensive models may also require more memory than a conventional server—even one that includes GPUs—can provide. In that case, processing is typically spread across multiple GPUs in a rack, with each having dedicated memory. In some scenarios, entire AI algorithms may be loaded onto a chip to provide for sufficiently fast retrieval.

Storage

Latency is critical in AI training because complex models may take weeks to run. The size of the data sets can be enormous: Microsoft reportedly used five years of continuous speech data

to teach computers to talk. Shortening the path from storage to CPU/GPU translates directly into performance gains.

Commodity object storage provides good scalability but is too slow to meet the I/O needs of multi-GPU systems. To compensate, data lakes based on fast solid-state disk NVMe storage may need to be used as intermediate data pools to serve GPUs as quickly as possible.

NVMe storage is best suited to meet performance needs. Still, sufficient capacity may not be available in cloud platforms, or there may be limits on how many I/O operations per second are allowed on a device each month. Costs can grow significantly as the volume of I/O increases. On-premises systems can be purpose built for much faster data ingestion and parameter tuning than is possible in the cloud, without usage limits.

Bandwidth

AI training is typically conducted on copies of production data sets, which can reach a petabyte in size. Uploading data sets of that size to the cloud can take days. Most cloud providers also charge egress fees if customers want to retrieve their data from the cloud, with a 1-petabyte transfer costing approximately $50,000.

Inside the compute engine, synchronizing nodes requires a high-speed network. Bandwidth can become a significant bottleneck at that point. Processing training workloads in the cloud can be 30% slower than on a local server due to network factors alone.

[3] "Why AI Workloads Require New Computing Architectures–Part 1," Applied Materials, June 20, 2018

Workload characteristics

The more specialized the AI workload is, the more specific the hardware requirement. For example, artificial neural networks, which mimic biological neural networks, use trial and error techniques that produce and continuously run multiple network derivatives. This can consume a lot of cloud processor capacity and increase costs as workloads are rerun. Furthermore, developers often experiment with numerous learning frameworks for a single problem, which requires running tests across each one, thereby driving up usage costs.

Software

Development toolkits are increasingly being written for AI-specific use cases. For example, the NVIDIA DRIVE Sim is a simulation platform built for large-scale, physically accurate multi-sensor simulation scenarios such as autonomous vehicles. These software development kits and frameworks are often bundled into purpose-built platforms, and the platforms themselves can be constructed to optimize performance on specific SDKs and frameworks. While cloud platforms run many of these SDKs, the hardware is rarely optimized for them. Specialized hardware configurations are available for deployment only in the data center.

Administration

Operationalizing AI requires monitoring ongoing projects with KPIs, controls and accountability. An on-premises solution supports this activity by giving IT complete visibility into the work AI developers are doing. A local platform also provides better security than a multitenant cloud instance by limiting the possibility of accidental data exposure, which is the No. 1 cloud security risk today.

The best approach to optimizing cost and control is to use a combination of cloud and on-premises resources. Use the cloud as a sandbox for experimentation and testing, and move operational AI processing to a platform that is easier to fine-tune, control and monitor. Promising projects can be migrated to the data center for training and integration with operating systems. Operational AI shouldn't be an either-or proposition. Align the platform to the task and take AI development to the next level.

The best approach to optimizing cost and control is to use a combination of cloud and on-premises resources.

Bottom-Line Considerations

While apples-to-apples comparisons are difficult to make because of the specialized nature of AI workloads, many experienced developers say building and maintaining AI models in the cloud can cost two to three times as much as building and running them on on-premises infrastructure.[4]

One Forrester Consulting survey found that 42% of companies primarily use cloud providers for training AI models today, but just 12% plan to do so in three years.[5] Most will operationalize AI by moving to purpose-built hardware and a combination of on-premises and cloud platforms. DeterminedAI compared the three-year cost of running AI workloads on premises and in the cloud and concluded that specialized local hardware was more than 75% less expensive for high-utilization scenarios.[6]

AI workloads have become more sophisticated and specialized. Achieving optimal performance requires specifying the ideal combination of hardware components. It is a common misconception that running AI models is just a matter of adding GPUs. As we have discussed in this paper, solutions need to be modified for actual workloads, which may require significant upgrades to memory, storage and bandwidth.

[4] "AI On Premises: Benefits and a Predictive-Modeling Use Case," Canonical, March 24, 2021

[5] "Whose Hardware Will Run Analytics, AI and ML Workloads?" The New Stack, Jan. 31, 2019

[6] "Choosing Your Deep Learning Infrastructure: The Cloud vs. On-Prem Debate," DeterminedAI, July 30, 2020

Data center environments and power sources may also need to be changed. Software drivers, libraries, development frameworks and SDKs must be chosen to match the desired outcomes. Performance tuning is ongoing, and because models are based on historical data, retraining may be required to prevent quality from decaying.

Few IT professionals with general computing backgrounds have the requisite skills to build and install an AI infrastructure. Working in conjunction with partners that have deep experience configuring, installing and managing AI computing environments gets the most out of your investment and provides even more significant savings than running in the cloud.

Using a single integrator for system configuration support and services, organizations can utilize such quality-enhancing and time-saving features as single-source procurement, supply chain management, rack-level integration and seamless expansion.

Equus Compute Solutions is the largest custom computer manufacturer in the channel, with 3.5 million custom-configured servers, software appliances, desktops and notebooks shipped worldwide. Equus has worked with thousands of partners on custom solutions. Its PartnerLink website provides personalized product configurators, packaging and imaging, testing, expedited product delivery and tracking.

Equus offers a wide range of GPU-based computing platforms in rack-mounted and tower configurations, optimized for massively parallel computing environments. These can be tuned to the workload with customer-specified GPU options, a wide variety of high-performance, hybrid and archive storage configurations, and open systems management.

The time to value for AI projects depends on finding the use cases with the most significant business impact, but there is also a less tangible cost in the risk of not moving quickly enough. “By the time a late adopter has done all the necessary preparation, earlier adopters will have taken considerable market share,” according to Vikram Mahidhar and Thomas Davenport in a Harvard Business Review article.[7] “They’ll be able to operate at substantially lower costs with better performance. In short, the winners may take all, and late adopters may never catch up.”

> The time to value for AI projects depends on finding the use cases with the most significant business impact, but there is also a less tangible cost in the risk of not moving quickly enough.

[7] “Why Companies That Wait to Adopt AI May Never Catch Up,” Harvard Business Review, Dec. 6, 2018

This content was commissioned by Intequus, and produced by TechTarget Inc.