A Survey On NeoClouds, Hyperscalers, and Optimization Techniques for Large Language Models

Aarush Agrawal
Hiranandani Foundation School, Powai, India
aarush.social@gmail.com

***Abstract*—The rapid rise of large language models (LLMs) has created unprecedented demand for specialized AI infrastructure. "Neoclouds," GPU-as-a-Service (GPUaaS) providers, complement traditional hyperscalers by delivering AI-optimized compute, storage, and networking. This study systematically evaluates neoclouds (CoreWeave, Lambda Labs, Together AI) versus hyperscalers (AWS, GCP, Azure, Oracle Cloud), focusing on GPU architectures (NVIDIA H100/H200, AMD MI300, TPUs), interconnects (Ethernet vs. InfiniBand), and AI optimization techniques including quantization, LoRA, and Adam variants. Using LLM training and inference benchmarks, we assess multi-tenant efficiency, memory utilization, and energy consumption. Results from industrial reports show that neoclouds achieve up to 75 per cent faster LLM performance and significantly lower energy costs through hardware-software co-optimization, including tailored memory allocation, mixed-precision computation, and optimized inter-node communication. This study highlights trade-offs in latency, cost, and scalability between neoclouds and hyperscalers, providing actionable insights for enterprises seeking high-performance, cost-effective AI infrastructure.**

***Index Terms*—NeoClouds, Hyperscalers, GPUaaS, AI Accelerators, LLM Optimization, Ethernet, InfiniBand**

I. Introduction

The artificial intelligence (AI) revolution has driven the emergence of neoclouds, cloud providers specializing in GPU-as-a-Service (GPUaaS) within the Infrastructure-as-a-Service (IaaS) model [1]. Unlike hyperscalers such as Amazon Web Services (AWS), Microsoft Azure, Google Cloud Platform (GCP), and Oracle Cloud, which offer broad cloud services, neoclouds focus on GPU-centric infrastructure for AI, machine learning (ML), and analytics workloads. The term "neocloud," though of unclear origin, is now an industry standard for GPUaaS providers delivering reliable, high-performance AI infrastructure. Early neoclouds aimed to capitalize on GPU demand but struggled with unoptimized infrastructure. Successful providers prioritize stable, high-performance systems, addressing the computational needs of large language models (LLMs) with billions to trillions of parameters.

This paper analyzes neoclouds, hyperscalers, their networking architectures (Ethernet vs. InfiniBand), AI accelerators (TPUs, GPUs), and LLM optimization techniques.It compares their infrastructure, scalability, and performance through comprehensive evaluations and highlighting their roles in enterprise AI adoption. Key challenges, such as GPU costs and multi-tenant performance, are addressed with detailed infrastructure descriptions.

II. NeoClouds: Architecture and Services

A. Key Characteristics of Neoclouds

Neoclouds represent a specialized class of cloud providers focused exclusively on AI workloads, characterized by several distinguishing features:

- **GPU-Centric Infrastructure**: The computational foundation relies exclusively on high-performance AI accelerators. As noted by industry analysis, these providers "leverage advanced GPUs from NVIDIA (Hopper H100/H200, Blackwell B200/GB200) and AMD (Instinct MI300), delivering teraflops to petaflops for AI tasks" [2]. For instance, the NVIDIA H100 delivers 312 TFLOPS at FP16 precision, forming the basis for their performance advantage.
- **Full-Stack AI Optimization**: Unlike general-purpose clouds, neoclouds optimize their entire technology stack specifically for AI workloads. Research indicates that "infrastructure, networking, and software stacks are tailored for AI, achieving significantly faster training than hyperscalers" [3].
- **Rapid Innovation Cycle**: Intense market competition drives quick adoption of cutting-edge technologies. Providers consistently integrate the latest hardware and software innovations, such as CoreWeave's Tensorizer and Together AI's FlashAttention-3 [5], [8].
- **Operational Flexibility**: The GPU-as-a-Service (GPUaaS) model operates on pay-as-you-go pricing, eliminating substantial capital expenditures while enabling dynamic resource allocation across multiple training experiments.
- **Developer Control**: Virtual machines remain crucial for AI development. As established in virtualization research, "VMs provide interfaces for hyperparameter tuning and training monitoring, enhancing user control" [4].
- **Pricing Transparency**: Neoclouds typically offer clear, predictable pricing models for GPU access, contrasting with the complex cost structures of traditional hyperscalers.

B. Neocloud Market Leaders

The current neocloud ecosystem features several specialized providers:

- **CoreWeave**: The largest specialized GPUaaS provider, offering comprehensive NVIDIA GPU access including H100, H200, GH200, and GB200 models, complemented by managed Kubernetes services [5].
- **Nebius**: A prominent European provider featuring diverse NVIDIA GPUs (L40s, H100, H200, B200) and high-performance computing clusters leveraging InfiniBand technology [7].
- **Lambda Labs**: Provides both cloud and on-premises NVIDIA GPU solutions (A100, H100, RTX 6000) with integrated software stacks for AI research [6].
- **Vultr AI**: Offers bare-metal servers with NVIDIA A100/H100 GPUs alongside Intel/AMD CPUs, with particular emphasis on sovereign cloud solutions [9].
- **Together AI**: Focuses on open-source LLM development with NVIDIA H100 GPUs, utilizing FlashAttention-3 and RedPajama datasets [8].

C. Detailed Service Comparison

1) CoreWeave: CoreWeave's service portfolio demonstrates comprehensive AI infrastructure coverage:

- **Compute**: GPU instances spanning NVIDIA's entire performance spectrum (H100, H200, GH200, GB200, RTX PRO 6000) with Dell PowerEdge CPU servers and specialized Kubernetes orchestration [5].
- **Storage**: Implements NVIDIA GPUDirect Storage with VAST Data clusters and dedicated AI Object Storage for high-speed data access [5].
- **Networking**: Utilizes BlueField-3 DPUs with VPC and Direct Connect for high-bandwidth connectivity [5].
- **Managed Services**: Offers automated cluster validation, managed Kubernetes, and proprietary tools like Tensorizer for model checkpointing [5].
- **Specialized Services**: Supports VFX rendering (Autodesk Maya, Houdini), large-scale AI training (multi-trillion parameters), and high-throughput inference (800 TPS on Llama 3.1 405B) [5].

2) Nebius: Nebius provides vertically integrated AI infrastructure:

- **Compute**: Features NVIDIA L40s, H100, H200, and B200 GPUs with bare-metal CPUs and Slurm-optimized Kubernetes [7].
- **Storage**: Offers enhanced object storage (100 Gbps, 1M IOPS), fast storage for GPU clusters, and elastic network disks [7].
- **Networking**: Employs high-speed InfiniBand (3.2 Tbit/s) alongside VPC and Direct Connect [7].
- **Managed Services**: Includes managed MLflow, PostgreSQL, Apache Spark, and specialized operators for GPU clusters [7].

D. Why Businesses Need Neoclouds

Neoclouds address critical gaps in traditional cloud infrastructure, which often provides suboptimal GPU configurations and generic networking for specialized AI workloads. By offering dedicated AI accelerators, high-speed interconnects, and optimized software stacks, neoclouds significantly reduce training times compared to hyperscalers [10].

E. Challenges and Mitigation Strategies

The neocloud sector faces several significant challenges:

- **High GPU Costs**: NVIDIA H100 GPUs cost $30,000-40,000$, with rapid obsolescence cycles requiring frequent upgrades [11].
- **Energy Consumption**: AI workloads drive substantial electricity costs, necessitating sustainable solutions and efficient cooling [12].
- **Multi-Tenant Performance**: Achieving predictable job completion times in shared environments remains challenging due to resource contention [13].
- **Vendor Lock-in**: NVIDIA's dominance in GPUs, networking, and software creates significant platform dependency [15].

Common mitigation strategies include vendor partnerships for bulk discounts, energy-efficient data center designs (e.g., Nebius's 1.1 PUE), advanced AI scheduling algorithms that can improve job completion times by 25%, and modular hardware designs that facilitate easier upgrades [13], [16].

III. Networks: Ethernet vs. InfiniBand

A LAN is a data communication network connecting various terminals or computers within a building or limited geographical area. The connection between the devices could be wired or wireless. Although Ethernet has been largely replaced by wireless networks, wired networking still uses Ethernet more frequently. Ethernet, and Wireless LAN using IEEE 802.11 are examples of standard LAN technologies. [17]. Besides Ethernet and Wi-Fi, another historical standard for LANs is the Token Ring. In a Token Ring network, devices are connected in a ring shape, and a special data packet called a "token" is passed around to control which device can transmit data. This method helps prevent data collisions. However, Token Ring technology has been largely replaced by the more flexible and faster Ethernet.

A. Ethernet

Ethernet is the most widely used LAN technology and is constantly evolving according to IEEE standards 802.3. It is the traditional technology for connecting devices in a wired local area network (LAN) or a wide area network. It enables devices to communicate with each other via a protocol, which is a set of rules or a common network language.The reason behind its widespread adoption is that Ethernet is easy to maintain and allows low-cost network implementation. In addition, Ethernet offers flexibility in terms of the topologies that are allowed. Ethernet generally uses a bus topology. Ethernet operates in two layers of the OSI model, the physical layer and the data link layer. For Ethernet, the protocol data unit is a frame since we mainly deal with DLLs. In order to handle collisions, the access control mechanism used in Ethernet is CSMA/CD. [18].

1) Types of Ethernet:

- **Fast Ethernet**: This type of Ethernet network uses cables called twisted pair or CAT5. It can transfer data at a speed of around 100 Mbps (megabits per second). Fast Ethernet uses both fiber optic and twisted pair cables to enable communication. There are three categories of Fast Ethernet: 100BASE-TX, 100BASE-FX, and 100BASE-T4. [19].
- **Gigabit Ethernet**: This is an upgrade from Fast Ethernet and is more common nowadays. It can transfer data at a speed of 1000 Mbps or 1 Gbps (gigabit per second). Gigabit Ethernet also uses fiber optic and twisted pair cables for communication. It often uses advanced cables like CAT5e, which can transfer data at a speed of 10 Gbps. [20].
- **10-Gigabit Ethernet**: This is an advanced and high-speed network that can transmit data at a speed of 10 gigabits per second. It uses special cables like CAT6a or CAT7 twisted-pair cables and fiber optic cables. With the help of fiber optic cables, this network can cover longer distances, up to around 10,000 meters. [21].
- **Switch Ethernet**: This type of network involves using switches or hubs to improve network performance. Each workstation in this network has its own dedicated connection, which improves the speed and efficiency of data transfer. Switch Ethernet supports a wide range of speeds, from 10 Mbps to 10 Gbps, depending on the version of Ethernet being used. [22].

2) Features:

- **Speed**: When compared to a wireless connection, Ethernet provides significantly more speed. Because Ethernet is a one-to-one connection, this is the case. As a result, speeds of up to 10 Gigabits per second (Gbps) or even 100 Gigabits per second (Gbps) are possible. [23].
- **Efficiency**: An Ethernet cable, such as Cat6, consumes less electricity, even less than a wifi connection. As a result, these ethernet cables are thought to be the most energy-efficient. [24].
- **Data Quality**: Because it is resistant to noise, the information transferred is of high quality.[25].

3) Ethernet in AI Workloads: Ethernet's high bandwidth supports AI data centers, but its 20–80 microsecond latency can hinder distributed training. RoCEv2 reduces latency by 15%, improving performance [23].

B. InfiniBand

InfiniBand is a high-speed, low-latency interconnect for HPC and AI, enhancing CPU utilization and data center management [26].Infiniband operates using a point-to-point bidirectional serial link to transfer data. It employs a switched fabric architecture, allowing multiple devices to connect through switches, which manage data traffic efficiently.

1) Features and Benefits:

- **High Bandwidth**: Up to 800 Gbps [27].
- **QoS**: Prioritizes traffic and allocates bandwidth [28].
- **Virtualization**: Virtual lanes isolate tasks [29].
- **Low Latency**: 3–5 microseconds, ideal for LLMs [30].
- **Low Power**: Uses one-third the power of Ethernet adapters [30].

2) InfiniBand in AI Workloads: InfiniBand's native RDMA and NCCL-based load balancing reduce training times by 30% for 175-billion-parameter LLMs [30].

TABLE I: Comparison of Ethernet and InfiniBand Architectures

	Ethernet	InfiniBand
Max Bandwidth	800 Gbps	800 Gbps
MTU	9216 bytes	4096 bytes
Layer 3 Support	Yes	No
Delivery	Near-lossless	Lossless
Load Balancing	Hash Values	Deterministic (NCCL)
RDMA Support	RoCEv2	Native
Enhancements	Dynamic Load Balancing, Weighted ECMP	Adaptive Routing
Pros	Multi-workload fabrics, engineer-friendly	Easy to install, self-optimizing
Cons	Requires QoS tuning	Rare skillset, complex support

IV. Comparison of Cloud Services

The following section compares five cloud service providers: A. CoreWeave B. Nebius C. Lambda.ai D. Vultr Cloud E. Together.ai

A. Compute

CoreWeave (CRWV): Offers GPU Compute with NVIDIA H100, H200, GH200, GB200 NVL72, and RTX PRO 6000 Blackwell GPUs for AI workloads, alongside CPU Compute using Dell PowerEdge servers, and a specialized CoreWeave Kubernetes Service (CKS) for AI orchestration.[5]

Nebius (NBIS): Provides GPU Compute with NVIDIA L40s, H100, H200, and B200 Tensor Core GPUs, CPU Compute with bare-metal performance, and Managed Kubernetes with node termination handlers.[7]

Lambda.ai: Features GPU Compute with NVIDIA A100, H100, and RTX 6000 Ada GPUs, and CPU Compute with Intel Xeon processors, plus Managed Kubernetes for AI workloads.[6]

Vultr: Includes Cloud Compute (VPS) with Intel Xeon and AMD EPYC CPUs, GPU Compute with NVIDIA A100 and H100 GPUs, Bare Metal Servers with high-frequency options, and Managed Kubernetes with GPU support.[9]

Together.ai: Offers GPU Compute with NVIDIA H100 SXM5 GPUs and Serverless GPU Inference, focusing on AI-specific workloads with high-speed InfiniBand networking.[8]

B. Storage

CoreWeave (CRWV): Provides Storage Services with NVIDIA GPUDirect Storage, Dedicated Storage Clusters using VAST Data, and CoreWeave AI Object Storage integrated with Kubernetes.[5]

Nebius (NBIS): Offers Enhanced Object Storage with up to 100 GBps and 1M IOPS, Fast Storage for GPU clusters, and Network Disks for elastic storage.[7]

Lambda.ai: Includes High-Performance Storage for datasets and model checkpoints, and Object Storage integrated with GPU clusters.[6]

Vultr: Features Block Storage with NVMe SSDs, Object Storage with S3 compatibility, File Storage with NFS, and Network-Attached NVMe SSD Storage.[9]

Together.ai: Provides High-Performance Storage for datasets and RedPajama Dataset Storage for open-source AI model training.[8]

C. Networking

CoreWeave (CRWV): Utilizes High-Performance Networking with NVIDIA BlueField-3 DPUs, VPC Networking, and Direct Connect with multiple on-ramp locations.[5]

Nebius (NBIS): Employs InfiniBand Network with up to 3.2 Tbit/s, VPC Networking, and Direct Connect for private connections.[7]

Lambda.ai: Offers High-Speed Networking with up to 400 Gbps, VPC Networking, and Direct Connect for on-premises links.[6]

Vultr: Provides VPC Networking, Direct Connect with BGP support across 32 data centers, CDN, and Load Balancer services.[9]

Together.ai: Features High-Speed InfiniBand Networking with up to 3200 Gbps and Private Networking for secure traffic.[8]

D. Managed Services

CoreWeave (CRWV): Includes Managed Software Services with automated validations, and Managed Environments with proprietary Kubernetes enhancements.[5]

Nebius (NBIS): Offers Managed Kubernetes with GPU-Operator, Slurm-based Clusters with Soperator, Managed MLflow, PostgreSQL, Apache Spark, and Soperator for GPU clusters.[7]

Lambda.ai: Provides Managed Kubernetes and Managed TensorFlow and PyTorch environments for ML frameworks.[6]

Vultr: Features Managed Databases (MySQL, PostgreSQL, Apache Kafka®, Valkey™), Vultr Kubernetes Engine, and Marketplace Applications.[9]

Together.ai: Includes Managed Kubernetes with Slurm integration and Managed Inference Engine with custom FP8 kernels.[8]

E. Virtual and Bare Metal Servers

CoreWeave (CRWV): Offers Virtual Servers with VMware virtualization and Bare Metal Servers with Dell PowerEdge and NVIDIA HGX platforms.[5]

Nebius (NBIS): Provides Virtual Servers with container images and Bare Metal Servers designed in-house.[7]

Lambda.ai: Includes Virtual Servers for AI development and Bare Metal Servers with NVIDIA DGX systems.[6]

Vultr: Features Virtual Servers with NVMe SSDs and Bare Metal Servers with customizable configurations.[9]

Together.ai: Offers Dedicated Instances with NVIDIA GPUs and Virtual Servers for flexible deployment.[8]

F. Specialized Services

CoreWeave (CRWV): Provides VFX and Rendering for creative tools, AI Model Training with Tensorizer, AI Inference with 800 TPS on GB200, and Mission Control for hardware management.[5]

Nebius (NBIS): Offers AI Model Training and Inference with MLPerf® v5.0, Full ML Lifecycle Support, and NVIDIA DGX Cloud Serverless Inference.[7]

Lambda.ai: Includes AI Model Training for large-scale models, AI Inference with low latency, and Custom AI Workstations.[6]

Vultr: Provides DDoS Mitigation, Global AI Model Deployment, Composable Cloud, and Sovereign/Private Cloud options.[9]

Together.ai: Features AI Model Training with FlashAttention-3, AI Fine-Tuning, and AI Inference with QTIP and Turbo Mode.[8]

G. Development Tools

CoreWeave (CRWV): Offers Dataset Optimization Tool, Tensorizer for model management, and Weights & Biases (W&B) Weave for AI observability.[5]

Nebius (NBIS): Includes Terraform, API, CLI, Intuitive Console, MCP Server with Anthropic Claude, JupyterLab Notebook, and SkyPilot.[7]

Lambda.ai: Provides Lambda Labs Console, Lambda Stack with CUDA and ML frameworks, and API/CLI for automation.[6]

Vultr: Offers API, Terraform Provider, Vultr Control Panel, and Container Registry.[9]

Together.ai: Features Together Inference API, Together CLI, and Jupyter Notebook Integration.[8]

V. Hyperscalers and Their Features

A. Definition

Hyperscalers are large-scale data centers that provide a wide range of cloud computing and data solutions for businesses that need vast digital infrastructure, processing, and storage.[31]

B. Distinguishing factors

It is not only their size—they're capable of hosting millions, if not billions, of users with speed and efficiency—but also their capacity to dynamically scale on demand.

C. Difference Between Hyperscalers and co-location Data centers

Hyperscalers use distributed computer systems, teams of independent computers splitting up tasks across multiple machines to process high volumes of data. This is called hyperscale computing. Traditional data centers, on the other hand, prioritize centralized computer systems, which can often mean fewer but more powerful computers. The big difference is how they adjust their capacity to handle increasing workloads: Traditional data centers "scale up" by upgrading existing hardware. Hyperscalers "scale out" by adding more GPUs thus leading in sheer quantity. These methods of scaling make a huge difference when businesses need to grow fast, dramatically increasing their data capacity and workloads. Hyperscalers can usually meet this demand seamlessly, while traditional co-location data centers will often hit a limit due to hardware constraints. [32].

D. Today's Biggest Hyperscalers

The "big three" hyperscalers dominating the cloud market are Amazon Web Services (AWS), Microsoft Azure, and Google Cloud Platform (GCP). These are the companies that provide the most extensive cloud infrastructure solutions worldwide. Among the wide array of services hyperscalers provide, here are a few that stand out:
Infrastructure-as-a-service (IaaS): Virtualized servers, networks, and storage that companies can "rent" from a hyperscaler.
Platform-as-a-service (PaaS): Platforms for developers to build and manage applications without the need for hardware.
Software-as-a-service (SaaS): Some hyperscalers offer computer applications that are accessible via Web browsers.
Content delivery network (CDN): Faster Web content delivery through the use of multiple servers across the globe.
Big data analytics: Tools to process and analyze data to detect patterns, trends, and other data-based insights.

E. Advantages for Enterprises

Companies are increasingly looking to avoid the burden, costs, and risks of having to set up, manage, maintain, and scale up their own hardware-based infrastructure. This is why cloud computing, in general, has gained traction over the last few decades. But when it comes to larger operations dealing with immense data volumes and workloads, hyperscalers' economies of scale dwarf those of traditional data centers, allowing businesses to save on costs and resources while optimizing performance. Hyperscalers may also offer value-added information technology services such as database management, data security, and artificial intelligence (AI) integration, among others. When AI exploded into the mainstream in the early 2020s, hyperscalers became key players in a game that was rapidly changing the nature of work.

F. The Effect of Generative AI

Hyperscalers such as Amazon, Microsoft, and Alphabet—are ramping up their investment in generative AI technologies: Amazon is integrating AI into its retail and cloud services for a number of tasks, from customer experience optimization to machine learning and code generation for its business customers.
Microsoft, through its partnership with OpenAI, has integrated ChatGPT into its Bing search engine and Azure ecosystem, allowing users to access a wide suite of AI solutions including workflow automation, coding, customer service, app development, and data analysis. Alphabet integrated Gemini, its most powerful generative AI model, into business and enterprise-level workspaces. Its multimodal capabilities—processing text, images, audio, and video across multiple platforms—can enable businesses to streamline and automate a wide range of complex tasks.
The widespread adoption of AI offerings from these firms—and other hyperscalers such as IBM (IBM) and Oracle (ORCL)—suggests that companies leveraging this new technology may be gaining a competitive advantage. The assumption among companies—and those who analyze and invest in them—is that the more advanced a company's AI tools, the more competitive its production potential will be. Hence, hyperscalers that are able to provide more advanced AI tools might not only attract more commercial users, but may also provide them with a significant technological edge over their competitors. Hyperscalers are the big players in the global digital infrastructure. The demand for their products and services will likely continue to grow, which, for investors, signals potential long-term investment opportunities.
It is important to consider, however, that many hyperscalers (particularly Amazon, Microsoft, and Alphabet) are already some of the largest companies on the planet. Their market capitalizations have seen decades of growth, driven not only by their cloud services (and other products and services), but also by investor expectations of more innovation in the future. In prior eras, a company or industry would be priced for growth, then slowly fade to value as it matured. Today's hyperscalers, however, can boast price-to-earnings (P/E) ratios that are double and even triple those of a typical value stock. In other words, continued growth and innovation is a presupposition in share prices, which makes investing in hyperscalers a high risk to reward activity.

G. Services

- **IaaS**: Virtualized servers, storage, and networking (e.g., AWS EC2, Azure VMs) [33].
- **PaaS**: Development platforms (e.g., AWS CodePipeline, Azure DevOps) [34].
- **SaaS**: Cloud-hosted applications (e.g., Oracle Cloud Applications) [35].
- **CDN**: Global content delivery (e.g., AWS CloudFront, Azure CDN) [36].
- **Analytics**: Data processing tools (e.g., AWS Redshift, Google BigQuery) [37].

H. Generative AI Integration

- **Amazon**: SageMaker and AI-driven retail optimization [38].
- **Microsoft**: ChatGPT integration in Azure and Bing via OpenAI [39].
- **Alphabet**: Gemini for multimodal AI tasks [40].

I. Comparison of Hyperscalers

This section compares cloud services such as AWS, Azure, Oracle Cloud Infrastructure (OCI), and Oracle Cloud (OC): AWS (Amazon Web Services) is a leading cloud computing platform provided by Amazon, offering a broad range of Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS) solutions. Launched in 2006, AWS provides scalable and cost-effective services such as compute (e.g., EC2), storage (e.g., S3), networking (e.g., VPC), databases (e.g., RDS), AI/ML (e.g., SageMaker), and serverless computing (e.g., Lambda). With a global infrastructure spanning over 30 regions and 100+ availability

zones, AWS supports businesses of all sizes, enabling them to build, deploy, and manage applications with high availability, security, and flexibility.
Azure (Microsoft Azure) is a cloud computing platform developed by Microsoft, offering a wide array of IaaS, PaaS, and SaaS services. Introduced in 2010, Azure provides solutions for computing (e.g., Virtual Machines), storage (e.g., Blob Storage), networking, databases (e.g., Azure SQL), AI/ML (e.g., Azure Machine Learning), and serverless computing (e.g., Azure Functions). With a global presence in over 60 regions, Azure integrates seamlessly with Microsoft's ecosystem (e.g., Windows Server, Active Directory) and emphasizes hybrid cloud capabilities, allowing organizations to extend on-premises infrastructure to the cloud while ensuring security and compliance. OCI (Oracle Cloud Infrastructure) is a comprehensive cloud computing platform that provides a wide range of services, including Infrastructure as a Service (IaaS) and Platform as a Service (PaaS). It offers scalable, secure, and high-performance solutions for building, deploying, and managing applications and workloads. OCI includes services like compute, storage, networking, databases, AI, and more, with a global network of over 50 regions to support enterprise-grade needs. OCP (Oracle Cloud Platform) refers to the broader cloud platform services provided by Oracle, encompassing both OCI and additional offerings. It includes tools and services for developing, integrating, and extending applications, such as Oracle Cloud Applications (SaaS) and various PaaS capabilities. OCP leverages OCI's infrastructure to deliver a fully integrated stack for businesses, supporting diverse workloads and enabling innovation with consistent pricing and performance across regions.

VI. AI Accelerator Architectures

A. Tensor Processing Unit (TPU)

One of the most efficient architectures for Machine Learning hardware is the Tensor Processing Unit made by Google and is available through their cloud developer platform. It utilizes the principles of a systolic array and achieves significant gains in compute time for matrix multiplications and feature map development. It is optimized for deep learning tasks and intensive ML workloads. TPUs achieve 1100 TFLOPS (4-chip) and 460 PetaFLOPS (8960-chip TPU v5p pod) with 160 MB SRAM.[41]

A Systolic array is a simple and energy-efficient architecture for accelerating general matrix multiplication (GEMM) operations in hardware. They provide an alternative way to implement these operations and support parallel data streaming to improve memory access and promote data reuse. This architecture forms the basis of many commercial accelerator offerings like the Google TPU (tensor processing unit), Intel NPU (neural processing unit), IBM AIU, etc.These arrays comprise MAC (multiply-and-accumulate) units that perform the actual operations. Serving the MAC units are the row and column SRAM buffers that feed these units with data.
Each MAC unit will save the incoming data in an internal register and then forward the same data to the outgoing connection in the next cycle. [42] This significantly improves the SRAM bandwidth requirement which was initially a major bottleneck that drained computing resources severely. This was because even though a GPU like an NVIDIA H100 can perform 67 Teraflops at FP64 precision, it only has 80 GB of VRAM which is not sufficient for intense foundation workloads. TPUs optimize on this by their unique MAC approach that significantly resolves this bottleneck achieving 1100 Teraflops in a 4 chip configuration powering Gemini and other foundation models running on the cloud with a TPU v5p projected to deliver 460 Petaflops per pod with an 8960 chip configuration. All of this is achieved with barely 160mb of SRAM. Each core is optimised for 128x128 matrix multiplications. The efficiency of such a device can be understood by observing the systolic array matrix multiplication cycle equation. The cycle time for a 2D weight stationary array is:

$$T = 2S_r + S_c + t - 2 \tag{1}$$

where S_r, S_c are required dimensions, Where Sr and Sc are the dimensions of the systolic array and the IFMAP along with the FILTER respectively, which is compared with the weights to obtain the output tensor with the T on the right representing the total number of computations and the one on the left giving the number of cycles required. However since the pods do not have an infinite amount of TPUs for AI workloads. As a result a constrained version is necessary that takes into account the computational bottleneck offered by a finite amount of hardware.

$$T = (2S_r + S_c + t - 2) \left\lceil \frac{S_r}{R} \right\rceil \left\lceil \frac{S_c}{C} \right\rceil \tag{2}$$

R, C are actual dimensions [43].
Here R and C are the actual dimensions of the systolic array whereas Sr and Sc are the required dimensions. To decrease this time, we can increase the number of MAC units, a process we can call "scaling up". Another approach is to have multiple MAC array units that perform the computation in parallel, which can be called "scaling out". This further reduces the time needed to complete the operation. TPUs take advantage of this and are usually scaled up, not just out in the pods. This significantly boosts the computations - per - watt metric making them ideal for cost efficient hosting of LLMs.

B. Graphical Processing Units (GPUs)

A Graphics Processor Unit (GPU) is mostly known for running applications that weigh heavy on graphics, i.e. 3D modeling software or VDI infrastructure. In the consumer market, a GPU is mostly used to accelerate gaming graphics. Today, GPU's (General Purpose GPU) are the primary choice in High Performance Computing (HPC) landscapes. HPC in itself is the platform serving workloads like Machine Learning (ML), Deep Learning (DL), and Artificial Intelligence (AI). Using a GPU is not only about Machine Learning computations that require image recognition anymore. Calculations on tabular data is also a

common exercise in i.e. healthcare, insurance and financial industry verticals.

Consider the main differences between a Central Processing Unit (CPU) and a GPU. A common CPU is optimized to be as quick as possible to finish a task at a as low as possible latency, while keeping the ability to quickly switch between operations. It's objeCtive is processing tasks in a serialized way. A GPU's main purpose is throughput optimization, allowing to push as many as possible tasks through its internals at once. It does so by being able to process mutliple tasks at the same time.

However, it is not only about the number of cores. And when we speak of cores in a NVIDIA GPU, we refer to CUDA cores that consist of ALU's (Arithmetic Logic Unit). Modern CPU architectures are optimized low latency memory access by using cache layers. The objective of a GPU is High Performance Computing.
A single GPU device consists of multiple Processor Clusters (PC) that contain multiple Streaming Multiprocessors (SM). Each SM accommodates a layer-1 instruction cache layer with its associated cores. Typically, one SM uses a dedicated layer-1 cache and a shared layer-2 cache before pulling data from global GDDR-5 (or GDDR-6 in newer GPU models) memory. Its architecture is tolerant of memory latency. Compared to a CPU, a GPU works with fewer, and relatively small, memory cache layers. Reason being is that a GPU has more transistors dedicated to computation meaning it;s performance does matter significantly on the time taken for data retrieval. If one were to consider the Tesla V100, formerly offered by NVIDIA, there are 80 SM's containing 64 cores each making 5120 cores that are capable of parallel processing. GPUs in general employ a greater amount of transistors resulting in very small node sizes (approaching 2nm) with high densities. The NVIDIA RTX 3090 Ti contains approximately 28.3 billion transistors. The NVIDIA RTX 4090 uses the AD102 chip, which has a 609 mm² die and contains 76.3 billion transistors built using TSMC's custom 4NM node. [44], [45].

VII. Techniques to Reduce ML Workloads

A. Optimizers and the Vanishing Gradient Problem:

In machine learning, optimizers and loss functions are two fundamental components that help improve a model's performance. A loss function evaluates a model's effectiveness by computing the difference between expected and actual outputs. Common loss functions include log loss, hinge loss and mean square loss. An optimizer improves the model by adjusting its parameters (weights and biases) to minimize the loss function value. Examples include RMSProp, ADAM and SGD (Stochastic Gradient Descent). The optimizer's role is to find the best combination of weights and biases that leads to the most accurate predictions.
Let us first consider one of the most popular strategies for loss management and parameter updating. Gradient descent. [46] Key Steps in Gradient Descent Initialize parameters: Randomly initialize the model parameters. Compute the gradient(derivative) of the loss function with respect to the parameters. Adjust the parameters by moving in the opposite direction of the gradient, scaled by the learning rate. Gradient descent is mathematically refined by the Armijo Full Relaxation Condition involving the computation of the Hessian matrix (a square matrix consisting of the scalar partial order second derivatives of a function). This refining ensures the step size is neither too large nor too small but comes with a trade - off. It requires a small number of iterations with a greater number of computations per iteration, increasing the cost of training runs. However, it is conventionally accepted to be better than a normal Gradient Descent.[47].

Now one problem with gradient descent is the learning rate decay which results in wrong parameter values reducing accuracy. This is fixed by advanced algorithms that are critical for maintaining model performance metrics on large datasets. The one that is most commonly used is Adam (Adaptive Moment Estimation). [48]

The postulate of the vanishing gradient problem in stochastic modeling is the phenomenon in which the loss function returns values so small that the process of adjusting parameter values slows, resulting in infinitesimal changes as it approaches the global minimum, significantly reducing training speed and the time needed to approach the convergence. This paradox is eliminated by the Adam optimizer (Adaptive Moment Estimation) with the help of the following operations:

$$m_t = \beta_1 m_{t-1} + (1-\beta_1)\nabla J(\theta_t) \tag{3}$$

$$v_t = \beta_2 v_{t-1} + (1-\beta_2)[\nabla J(\theta_t)]^2 \tag{4}$$

$$\theta_{t+1} = \theta_t - \eta \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon} \tag{5}$$

Adam ensures fast convergence, critical for enterprise LLMs [49] [50]

B. Quantization of Weights and Biases

Quantization is a technique used to reduce the computational and memory costs of running inference by representing weights and activations using low-precision data types, such as 8-bit integers (`int8`), instead of the conventional 32-bit floating point (`float32`). Reducing the bit width leads to models that:

- Require less memory storage,
- Consume less energy (theoretically),
- Enable faster operations like matrix multiplication through integer arithmetic,
- Are deployable on embedded systems, which often support only integer data types.

Types of Quantization:

a) Post-Training Quantization (PTQ):: PTQ is applied after a model has been trained using 32-bit precision. No retraining is required, making it straightforward to implement. Weights—and optionally activations—are converted from `float32` to lower-precision formats like `int8`. The common PTQ types include:

- **Dynamic Quantization:** Only the weights are quantized to `int8`, while activations remain in `float32` during inference. Activations are dynamically quantized based on their runtime range.
- **Static Quantization (Full Integer Quantization):** Both weights and activations are quantized to `int8`. Calibration is required using a representative dataset to determine the dynamic range of activations.
- **Float16 Quantization:** Weights and activations are quantized to `float16`, preserving more of the dynamic range than `int8`, while still accelerating inference.

b) Quantization-Aware Training (QAT):: QAT simulates quantization during the training process, allowing the model to adapt its weights to the target lower-precision format. While forward passes are simulated using quantized values (via "fake quantization"), gradient computations and updates still use `float32`. This often results in less accuracy degradation compared to PTQ.

QAT is especially effective for complex models—such as large CNNs or transformers—where maintaining accuracy is critical.

c) Edge AI Applications:: Ultra low-precision quantization is increasingly essential for deploying large-scale language or vision models on memory-constrained devices, such as microcontrollers. This technique is foundational to Edge AI applications in:

- Drones,
- Internet of Things (IoT),
- Self-driving vehicles.

Quantization Mathematics: Quantization can be described as an affine transformation in Euclidean space. It does not preserve angular or rotational properties. The quantized value x_q is computed as:

$$x_q = \text{round}\left(\frac{x}{s}\right) \tag{6}$$

where:

- x is the original floating-point value,
- s is the scale factor,
- Z (not shown in the formula above) is the zero-point used in asymmetric quantization to map zero in floating point to an integer value.

[51]

d) Dequantization:: The inverse operation is used post-quantization to recover approximate floating-point values during inference. This helps mitigate accuracy loss.

e) Symmetric vs. Asymmetric Quantization::

- **Symmetric Quantization:** Maps a floating-point range, clipped symmetrically around zero (e.g., $[-\alpha, \alpha]$), to an integer range like $[-127, 127]$. No zero-point is used. This method preserves weight symmetry and simplifies computations. [52]
- **Asymmetric Quantization (Affine):** Uses a zero-point offset Z and can represent ranges not centered at zero. It's more flexible for activations but introduces additional arithmetic.

Symmetric quantization is preferred in foundation model training due to its simplicity and computational efficiency. The dequantization for symmetric quantization is typically:

$$x \approx s \cdot x_q \tag{7}$$

This method reduces memory usage by up to 75% [52], while maintaining computational integrity in deep learning pipelines since it simplifies thirty two bit weights to 8 bit, thus reducing the memory space occupied.

C. Low-Rank Adaptation (LoRA)

Low-rank adaptation (LoRA) is a method for rapidly adapting machine learning models to new use cases without retraining them. It enables developers to customize models for specific contexts. LoRA works by appending a lightweight addition, called a low-rank matrix, to the original model. This matrix tweaks the outputs of the model. In this approach the initial weights of the model are frozen and new low rank matrices are created to process the results obtained from the model to match user requirements. They can be compared to "adapters" in an analogical sense as they link a device with a certain port configuration to a device of another port configuration. Similarly here the model being fine tuned is kept as is, with new tensors being introduced to enhance accuracy in a specific use case. It significantly reduces computation related cost and infrastructure requirements. [53]

$$W' = W + AB \tag{8}$$

where $A \in \mathbb{R}^{d \times r}$, $B \in \mathbb{R}^{r \times k}$, and $r \ll \min(d, k)$. LoRA reduces parameters by 90% [53]. The formula given above clearly depicts how fine tuning is far superior to re-training since it relies on elementary matrix multiplication and can be achieved with a standard parallelism enable GPU or even a CPU.

VIII. Conclusion

In conclusion, while AI hardware and algorithms have advanced significantly in recent years, serious challenges remain—particularly the "illusion" of true understanding and the readiness of these models for enterprise use, as highlighted in the MIT AI adoption study[54]. The coming years will be critical in shaping the architectural improvements needed to make large transformer models truly enterprise-ready and capable of meaningfully enhancing workforce productivity.

Acknowledgment

The author would like to thank **Mr. Sharad Sanghi** of **Neysa AI** for his invaluable guidance and dedicated support, which were instrumental in the completion of this paper.

REFERENCES

[1] D. Patel and D. Nishball, "AI NeoCloud Playbook and Anatomy," *SemiAnalysis Blog*, Oct. 3, 2024. [Online]. Available: https://semianalysis.com/2024/10/03/ai-neocloud-playbook-and-anatomy/. [Accessed: Sep. 12, 2025].

[2] NVIDIA Corporation, "H100 Tensor Core GPU – Product Brief," NVIDIA, Sep. 2022. [Online]. Available: https://www.nvidia.com/content/dam/en-zz/Solutions/products/accelerators/h100/pdf/nvidia-h100-datasheet.pdf. [Accessed: Sep. 28, 2025].

[3] J. Sekar, "Optimizing Cloud Infrastructure for AI Workloads: Challenges and Solutions," *Int. J. All Res. Educ. Sci. Methods*, vol. 12, no. 8, pp. 296–307, Aug. 2024.

[4] T. Doi and S. Tanino, "Virtual-machine management program and method for managing virtual machines," U.S. Patent 8 856 234, Sep. 30, 2014.

[5] CoreWeave, "Drive AI Innovation at Scale," Whitepaper, Mar. 2025. [Online]. Available: https://www.coreweave.com/blog/semianalysis-whitepaper. [Accessed: Sep. 5, 2025].

[6] "Lambda Builds AI Factories with Supermicro NVIDIA HGX B200 Server Clusters to Deliver Production-ready Next-Gen AI Infrastructure at Scale," *Lambda Deep Learning Blog*, Aug. 25, 2025. [Online]. Available: https://lambda.ai/blog/lambda-builds-ai-factories-with-supermicro-cologix. [Accessed: Sep. 21, 2025].

[7] Nebius, "GPU Clusters," *Nebius Docs*, 2025. [Online]. Available: https://nebius.com/docs. [Accessed: Sep. 18, 2025].

[8] "Together AI: Frontier AI Factory," *Together AI technical documentation*, 2025. [Online]. Available: https://www.together.ai/ai-factory. [Accessed: Sep. 9, 2025].

[9] R. Smith, "Vultr AI: GPUs," *Vultr Docs*, Sep. 10, 2025. [Online]. Available: https://docs.vultr.com/products/compute/cloud-gpu. [Accessed: Sep. 15, 2025].

[10] STL Partners, "Neoclouds: The new cloud players revolutionising enterprise AI infrastructure," *STL Partners Edge Computing*, Apr. 10, 2025. [Online]. Available: https://stlpartners.com/articles/edge-computing/neoclouds-the-new-cloud-players/. [Accessed: Sep. 2, 2025].

[11] "H200 vs H100 GPU Price Difference?," *Cyfuture Cloud*. [Online]. Available: https://cyfuture.cloud/kb/gpu/h200-vs-h100-gpu-price-difference. [Accessed: Sep. 24, 2025].

[12] "We Did the Math on AI's Energy Footprint. Here's the Story You Haven't Heard.," *MIT Technology Review*, May 20, 2025. [Online]. Available: https://www.technologyreview.com/2025/05/20/1116327/ai-energy-usage-climate-footprint-big-tech/. [Accessed: Sep. 1, 2025].

[13] F. Yu, D. Wang, L. Shangguan, M. Zhang, C. Liu, T. Soyata, and X. Chen, "A Survey of Multi-Tenant Deep Learning Inference on GPU," *arXiv preprint arXiv:2203.09040*, Mar. 2022. [Online]. Available: https://arxiv.org/abs/2203.09040. [Accessed: Sep. 19, 2025].

[14] J. Martin, "Price Wars in the Neocloud Market," *J. Cloud Econ.*, vol. 7, pp. 29–36, 2024.

[15] "What is Vendor Lock-in? — Vendor Lock-in and Cloud Computing," *Cloudflare*. [Online]. Available: https://www.cloudflare.com/learning/cloud/what-is-vendor-lock-in/. [Accessed: Sep. 11, 2025].

[16] F. A. Schulze, H. K. Arndt, and H. Feuersenger, "Obsolescence as a Future Key Challenge for Data Centers," in *Advances and New Trends in Environmental Informatics*, A. Kamilaris, V. Wohlgemuth, K. Karatzas, and I. N. Athanasiadis, Eds. Cham, Switzerland: Springer, 2021, doi: 10.1007/978-3-030-61969-5_5.

[17] S. Banerji and R. S. Chowdhury, "On IEEE 802.11: Wireless Lan Technology," *Int. J. Mobile Netw. Commun. Telematics*, vol. 3, no. 4, pp. 45–62, Aug. 2013, doi: 10.5121/ijmnct.2013.3405.

[18] D. Law, D. Dove, J. D'Ambrosia, M. Hajduczenia, M. Laubach, and S. Carlson, "Evolution of Ethernet Standards in the IEEE 802.3 Working Group," *IEEE Commun. Mag.*, vol. 51, no. 8, pp. 88–96, Aug. 2013, doi: 10.1109/MCOM.2013.6576344.

[19] A. Naz *Improving LAN Performance & Speed by Fast Ethernet Technology: A Synopsis*. New Delhi, India: Concept Books Publication, 2020.

[20] F. B. Tan, N. I. Sarkar, C. Byrne, and N. Al-Qirim, "Gigabit Ethernet Implementation," in *Global Information Technologies: Concepts, Methodologies, Tools and Applications*, M. D. Khosrow-Pour, Ed. Hershey, PA, USA: IGI Global, 2008, pp. 2308–2324, doi: 10.4018/978-1-59904-939-7.ch169.

[21] W. Feng, J. Hurwitz, H. Newman, S. Ravot, R. Cottrell, O. Martin, F. Coccetti, C. Jin, X. Wei, and S. Low, "Optimizing 10-Gigabit Ethernet for Networks of Workstations, Clusters, and Grids: A Case Study," in *Proc. ACM/IEEE SC '03*, Phoenix, AZ, USA, 2003, p. 50, doi: 10.1109/SC.2003.10030.

[22] Y. Song, A. Koubaa, and F. Simonot-Lion, "Switched Ethernet For Real-Time Industrial Communication: Modelling And Message Buffering Delay Evaluation," in *Proc. 28th IEEE Int. Conf. Emerg. Technol. Factory Autom. (WFCS)*, Castelldefels-Barcelona, Spain, 2002, pp. 27–35, doi: 10.1109/WFCS.2002.1159697.

[23] D. Valencic, V. Lebinac, and A. Skendzic, "Developments and current trends in Ethernet technology," in *Proc. 36th Int. Conv. Inf. Commun. Technol. Electron. Microelectron. (MIPRO)*, Opatija, Croatia, 2013, pp. 431–436.

[24] P. Reviriego, K. Christensen, J. Rabanillo, and J. A. Maestro, "An Initial Evaluation of Energy Efficient Ethernet," *IEEE Commun. Lett.*, vol. 15, no. 5, pp. 578–580, May 2011, doi: 10.1109/LCOMM.2011.040111.102259.

[25] D. S. Rodrigues and D. Jos'e, "Performance Analysis of Ethernet Networks Through Quality of Service (QoS) Metrics Using Real and Virtual Machines," *Rev. Bras. Comput. Aplicada*, vol. 17, no. 2, pp. 64–77, 2025, doi: 10.5335/rbca.v17i2.16481.

[26] Mellanox Technologies, "InfiniBand Technology Overview," Whitepaper, Oct. 2008. [Online]. Available: https://network.nvidia.com/pdf/whitepapers/WP_InfiniBand_Technology_Overview.pdf. [Accessed: Sep. 26, 2025].

[27] J. Van Schaik, "The Battle of AI Networking: Ethernet vs InfiniBand," *World Wide Technology Blog*, Nov. 3, 2024. [Online]. Available: https://www.wwt.com/blog/the-battle-of-ai-networking-ethernet-vs-infiniband. [Accessed: Sep. 13, 2025].

[28] NVIDIA, "InfiniBand QoS," *NVIDIA DOCA SDK Documentation*, 2024. [Online]. Available: https://docs.nvidia.com/doca/sdk/InfiniBand+QoS/index.html. [Accessed: Sep. 23, 2025].

[29] NVIDIA Networks, "I/O Virtualization Using Mellanox InfiniBand and Channel I/O Virtualization (CIOV) Technology," Whitepaper, 2010. [Online]. Available: https://network.nvidia.com/pdf/whitepapers/WP_Virtualize_with_IB.pdf. [Accessed: Sep. 10, 2025].

[30] Nebius Team, "InfiniBand in focus: Bandwidth, Speeds and High-Performance Networking," *Nebius Blog*, 2025. [Online]. Available: https://nebius.com/blog/posts/what-is-infiniband. [Accessed: Sep. 27, 2025].

[31] "What is a Hyperscaler?," *Red Hat*. [Online]. Available: https://www.redhat.com/en/topics/cloud-computing/what-is-a-hyperscaler. [Accessed: Sep. 4, 2025].

[32] P. Powell, "Hyperscale vs. Colocation: Go Big or Go Rent?," *IBM Think*, Mar. 27, 2024. [Online]. Available: https://www.ibm.com/think/topics/cloud-computing/hyperscale-vs-colocation. [Accessed: Sep. 16, 2025].

[33] M. Suliman, "A Brief Analysis of Cloud Computing Infrastructure as a Service (IaaS)," *Int. J. Innov. Sci. Res. Technol.*, vol. 6, no. 2, pp. 1409–1412, Feb. 2021.

[34] S. Pastore, "The Platform as a Service (PaaS) Cloud Model: Opportunity or Complexity for a Web Developer?," *Int. J. Comput. Appl.*, vol. 81, no. 18, pp. 29–37, Nov. 2013, doi: 10.5120/14225-2435.

[35] "Software As A Service," *ScienceDirect Topics*. [Online]. Available: https://www.sciencedirect.com/topics/computer-science/software-as-a-service. [Accessed: Sep. 28, 2025].

[36] Z. X. Lim, X. Q. Ho, D. Z. Tan, and W. Goh, "Ensuring Web Integrity Through Content Delivery Networks," in *Proc. IEEE World AI IoT Congr. (AIIoT)*, Seattle, WA, USA, 2022, pp. 494–500, doi: 10.1109/AIIoT54504.2022.9817199.

[37] Y. Yetis, R. G. Sara, B. A. Erol, H. Kaplan, A. Akuzum, and M. Jamshidi, "Application of Big Data Analytics via Cloud Computing," in *Proc. World Autom. Congr. (WAC)*, Rio Grande, PR, USA, 2016, pp. 1–5, doi: 10.1109/WAC.2016.7582986.

[38] R. Manasa and A. J. Devi, "Amazon's Artificial Intelligence in Retail Novelty - Case Study," *Int. J. Case Studies Bus., IT, Educ.*, pp. 787–804, Dec. 2022, doi: 10.47992/IJCSBE.2581.6942.0233.

[39] E. Boyd, "ChatGPT is now available in Azure OpenAI Service," *Microsoft Azure Blog*, Mar. 1, 2023. [Online]. Available: https://azure.microsoft.com/en-us/blog/chatgpt-is-now-available-in-azure-openai-service/. [Accessed: Sep. 20, 2025].

[40] A. Ramachandran, "Unveiling Google's Gemini 2.0: A Comprehensive Study of its Multimodal AI Design, Advanced Architecture, and Real-World Applications," Dec. 2024. [Online]. Available: https://www.researchgate.net/publication/387089907_Unveiling_Google's_Gemini_20_A_Comprehensive_Study_of_its_Multimodal_AI_Design_Advanced_Architecture_and_Real-World_Applications. [Accessed: Sep. 8, 2025].

[41] "TPU V5P," *Google Cloud*. [Online]. Available: https://cloud.google.com/tpu/docs/v5p. [Accessed: Sep. 28, 2025].

[42] H. T. Kung, "Algorithms for VLSI processor arrays," in *Introduction to VLSI Systems*, C. A. Mead and L. Conway, Eds. Reading, MA, USA: Addison-Wesley, 1980, ch. 8.3, pp. 271–280.

[43] T. Charitopoulos *et al.*, "ArrayFlex: A Systolic Array Architecture with Configurable Transparent Pipelining," in *Proc. Design, Autom. Test Eur. Conf. Exhib. (DATE)*, Antwerp, Belgium, 2023, pp. 1234–1239.

[44] NVIDIA, "NVIDIA Tesla V100 GPU Accelerator," Datasheet, Jul. 2017. [Online]. Available: https://images.nvidia.com/content/technologies/volta/pdf/tesla-volta-v100-datasheet-letter-fnl-web.pdf. [Accessed: Sep. 6, 2025].

[45] NVIDIA, "GeForce RTX 4090: Graphics Cards for Gaming," NVIDIA, 2022. [Online]. Available: https://www.nvidia.com/en-in/geforce/graphics-cards/40-series/rtx-4090/. [Accessed: Sep. 7, 2025].

[46] A. Tapkir, "A Comprehensive Overview of Gradient Descent and its Optimization Algorithms," *Int. Adv. Res. J. Sci., Eng. Technol.*, vol. 10, no. 11, Nov. 2023, doi: 10.17148/IARJSET.2023.101106.

[47] L. Armijo, "Minimization of Functions Having Lipschitz Continuous First Partial Derivatives," *Pacific J. Math.*, vol. 16, no. 1, pp. 1–3, 1966.

[48] S. Hochreiter, "The Vanishing Gradient Problem During Learning Recurrent Neural Nets and Problem Solutions," *Int. J. Uncertainty, Fuzziness Knowl.-Based Syst.*, vol. 6, no. 2, pp. 107–116, 1998, doi: 10.1142/S0218488598000094.

[49] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[50] S. Dereich, A. Jentzen, and A. Riekert, "Sharp Higher Order Convergence Rates for the Adam Optimizer," *arXiv preprint arXiv:2504.19426*, Apr. 2025, doi: 10.48550/arXiv.2504.19426.

[51] J. Lang, Z. Guo, and S. Huang, "A Comprehensive Study on Quantization Techniques for Large Language Models," in *Proc. 4th Int. Conf. Artif. Intell., Robot., Commun. (ICAIRC)*, Xiamen, China, 2024, pp. 224–231, doi: 10.1109/ICAIRC64177.2024.10899941.

[52] J. Faraone, N. Fraser, M. Blott, and P. H. W. Leong, "SYQ: Learning Symmetric Quantization For Efficient Deep Neural Networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Salt Lake City, UT, USA, 2018, doi: 10.48550/arXiv.1807.00301.

[53] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-Rank Adaptation of Large Language Models," *arXiv preprint arXiv:2106.09685*, 2021.

[54] A. Challapally *et al.*, "The GenAI Divide: State of AI in Business 2025," Project NANDA, MIT Media Lab, Cambridge, MA, USA, 2025.