TAKSHASHILA
INSTITUTION

A Primer on AI Chips

The Brains Behind the Bots

Ashwin Prasad and Satya S. Sahu

Takshashila Discussion Document No. 2024-21
Version 1.0, November, 2024

The paper provides an understanding of the various chips used to run AI models. It explains the characteristics and limitations of these chips which makes them suitable for certain AI applications, and not others. It also provides an overview of the market landscape and cautions against overreliance on a single vendor. It emphasises on exploring alternative solutions, and fostering open-source software ecosystems necessary for a diverse and resilient AI hardware landscape.

Recommended Citation:
Ashwin Prasad, Satya S Sahu, "A Primer on AI Chips," Takshashila Discussion Document No. 2024-21, The Takshashila Institution.

Executive Summary

The rise of Machine Learning, Deep Learning, and Natural Language Processing has driven unprecedented demand for specialised AI chips. These systems require substantial computational resources and can be deployed either in cloud data centres for maximum processing power or at the network edge for reduced latency and enhanced privacy.

The AI chip ecosystem comprises three critical components: accelerators (including CPUs, GPUs, FPGAs, and ASICs), memory and storage systems, and networking infrastructure. Each component plays a vital role in handling AI workloads, with different architectures offering varying trade-offs between performance and efficiency. The market for these technologies is heavily concentrated among a few key players: NVIDIA, Intel, AMD, Google, and TSMC.

A particular concern is NVIDIA's dominance in the GPU market and its proprietary software ecosystem, which creates significant dependencies for organisations and nations seeking to build sovereign AI infrastructure. As AI becomes increasingly critical to techno-national strategies worldwide, policymakers must understand these technological dependencies and support the development of alternative hardware and software solutions to ensure a more diverse and resilient AI chip ecosystem.

This document has been formatted to be read conveniently on screens with landscape aspect ratios. Please print only if absolutely necessary.

Authors

Ashwin Prasad is a Research Analyst with the High-Tech Geopolitics Programme at the Takshashila Institution. He can be reached at ashwin@takshashila.org.in

Satya S Sahu is a Research Analyst with the High-Tech Geopolitics Programme at the Takshashila Institution. He can be reached here.

Disclosures

The authors have utilised Claude and ChatGPT in the course of developing the index, abbreviations and glossary of this paper.

Acknowledgements

The authors would like to thank Pranay Kotasthane for his valuable comments and feedback.

Table of Contents

1. Abbreviations

AI	Artificial Intelligence
ASIC	Application-Specific Integrated Circuits
CGI	Computer Generated Imagery
CPU	Central Processing Unit
CUDA	Compute Unified Device Architecture
CXL	Compute Express Link
DRAM	Dynamic Random-Access Memory
FPGA	Field-Programmable Gate Arrays
GDDR	Graphics Double Data Rate
GPU	Graphics Processing Units
HBM	High Bandwidth Memory
HDD	Hard Disk Drive
HPC	High-Performance Computing
ISA	Instruction Set Architecture
LAN	Local Area Network
LLM	Large Language Model
ML	Machine Learning

NLP	Natural Language Processing
NVMe	Non-Volatile Memory Express
NPU	Neural Processing Unit
PCIe	Peripheral Component Interconnect Express
PIM	Processing-in-Memory
ROCm	Radeon Open Compute
SMRs	Small Modular Reactors
SoC	Systems-on-Chip
SSD	Solid State Drive
TPU	Tensor Processing Unit
UALink	Ultra Accelerator Link
UEC	Ultra Ethernet Consortium
UCIe	Universal Chiplet Interconnect Express
UPI	Ultra Path Interconnect

2. Background

The emergence of AI marks a significant milestone in the information age. As a General-Purpose Technology, AI holds the potential to have a transmuting effect on different sectors in different ways—autonomous driving in the automotive industry[1], fraud detection and risk assessment in finance[2], personalised marketing in retail[3], AI-driven diagnosis and personalised medicine in healthcare[4], AI-driven weapons and decision support systems[5]—the list is endless.

There is a pervasive interest in leveraging AI technologies for their economic, social, and strategic benefits. The size of the AI hardware market was valued at over $50 billion in 2023, and it is estimated to grow almost tenfold by 2030.[6] As AI permeates across various sectors, all of this comes with massive computational needs that the hardware has to enable and sustain.

A big chunk of this computational need is being met using GPUs. NVIDIA is the world's largest GPU company. With its AI-centric GPUs and extensive software ecosystem, NVIDIA has emerged as the world leader in AI computing.[7] It has positioned GPUs as the default choice for companies, government organisations, universities or any other entity that wants to deploy AI solutions. Case in point - about half of the India’s AI mission outlay of over ₹10,000 crores[8] has been earmarked for procuring GPUs to build AI computational infrastructure.[9]

Why is such a large portion of the budget earmarked to build AI computing capacity? Why did the Indian government choose GPUs? How do GPUs compare to other accelerators like the CPUs, FPGAs and ASICs for AI workloads? Does the growing complexity of AI algorithms challenge the traditional reliance on GPUs? Are there scenarios where FPGAs and ASICs outperform GPUs in AI applications? What implications does the choice of hardware architecture have on cost-effectiveness, energy consumption, flexibility and scalability?

As AI technologies evolve, policymakers should have a clear and thorough understanding of the available AI hardware options and their suitability for different use cases. Informed decision-making is necessary to build effective, efficient, and future-proof AI computing infrastructure under national missions like INDIAai.

This discussion document serves as a primer to understand the key elements of AI Chips, and is divided into three broad sections. The first section explains the workloads involved in AI tasks in order to understand the computational requirements that the hardware has to fulfil. The second section provides a comprehensive overview of the key elements of AI computing hardware.

These elements include AI accelerators (also called processing units), memory, storage, interconnects and networking systems. The section also distinguishes AI-specific hardware from other general-purpose computing hardware. The third section discusses the linkages between AI accelerators and software development ecosystems.

3. Understanding AI and its hardware requirements

Artificial Intelligence as a diverse bundle of technologies has existed for decades now. As such, the underlying hardware that run these technologies is also similarly disparate, and continuously evolving. For instance, the computer systems that ran the first image recognition algorithms operated differently from those running today's state-of-the-art facial recognition models.[10]

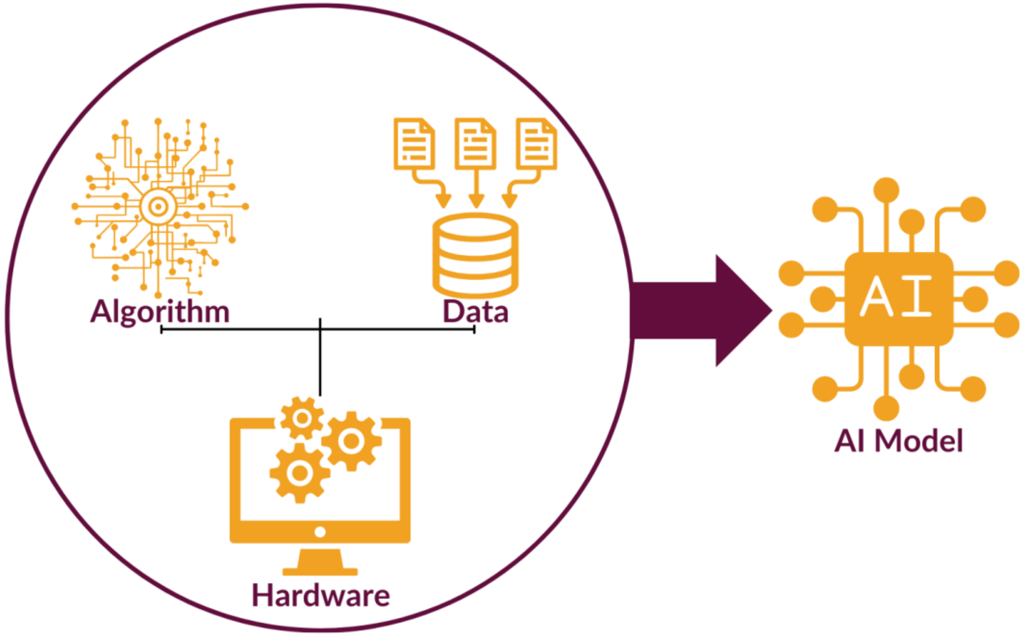
AI hardware, therefore, encompasses a wide range of computing systems but it has recently gained prominence in the public eye due to a dramatic progress in fields such as machine learning, and an exponential growth in digitised data. At the same time, the ability for algorithms to crunch massive amounts of data is directly attributable to the drastic increase in computing power seen over the past few decades.[11] While other subdomains of AI are still used, whenever AI is mentioned today, chances are that it refers to Machine Learning (ML). Machine Learning and associated subdomains such as Natural Language Processing (NLP), and Deep Learning, form the most significant chunk of the global AI market. The scope of this paper is restricted to computing hardware relevant to ML and associated fields.

Machine learning is a technology that allows computers to learn on their own instead of being programmed for every scenario.

Instead of following rigid rules, these systems analyse patterns in large amounts of data to make decisions or predictions. For example, rather than programming specific rules about what makes up a cat photo, a machine learning system learns to recognise cats by studying thousands of cat pictures. This approach has revolutionised computing, enabling applications like speech recognition, recommendation systems, and autonomous vehicles. The quality of results typically improves with more data and computing power, which is why modern AI requires such powerful processors.

Three main technological inputs[12] come together to make these models work:

1. The algorithms that form the brains of the AI models,
2. The data that these algorithms learn from,
3. And finally, the hardware that enables the algorithms to learn and run.



Source: Authors' Visualisation

Understanding the interaction between algorithms and the data in machine learning models provides a useful background to realise the computational requirements that the hardware has to fulfil.

These interactions can be broadly divided into two stages: training and inference. Algorithms undergo *training* where they learn from existing data. Once sufficiently trained, they can be used for *inference*, that is, to make predictions and draw conclusions about new data.[13]

Supervised data is labelled with the corresponding output, allowing the AI system to learn from the relationship between the two. Unsupervised data does not have a corresponding output, allowing the AI to identify patterns and clusters in the data.

3.1. Training

The foundation of AI is created during the training phase. It creates a mathematical model that can process new data to make valid predictions and draw accurate conclusions.[14] Training enables AI models to learn from supervised and unsupervised data and improve over time—essentially self-program.

There are various training types, all of which generally train over multiple stages or *iterations*. In each iteration, the model is taught from some data points in the dataset. The nature of the data varies depending on the model. For a model trained to interpret visual information, the datasets consist of images and videos. For an AI model that is trained to understand human language, called a Language Model (Large or Small LM, or LLM/SLM), the data consists of the language in textual form.

The process of training an AI model and the corresponding computational demands is explained using the example of an LLM below.[15]

Collecting the data points in the dataset: The dataset should have numerous data points sufficient to capture the nuances of human language. The data sources can come from user-generated content on the internet, books, web pages, etc. The datasets for recent LLMs may consist of terabytes of text.[16]

Pre-processing the datasets: The datasets have to be converted into a format that AI models can process. The data is cleaned of irrelevant content and converted from text to numbers—the format that the AI understands and interprets.[17]

Training and Testing: The model starts with an initial understanding of the language that may be random gibberish. A partial sentence of a certain length is fed into the model. The model uses this input to predict the next part of the sentence. Based on the accuracy of the prediction, the AI algorithm readjusts its initial understanding.

Just a single one of these iterations can require up to billions of mathematical calculations.[18] One complete pass of an entire dataset through the model is called an *epoch*. Given that a single iteration of processing one input requires billions of calculations, a single epoch may require many billion *billion* calculations. Training typically involves multiple epochs, as many as several thousand in some cases.

These calculations are mostly independent matrix multiplications. Each matrix multiplication does not always require the result of another matrix multiplication which means they can be run independently and in parallel. According to OpenAI's estimations, the training of the GPT-3 model took over 300 billion *trillion* floating point calculations.[19] Considering that running these operations on a single NVIDIA Tesla V100 GPU would take 355 years[20] and that GPT-3 was trained on 10,000 V100 GPUs,[21] the total training time is estimated to be around 34 days.

Floating point operations (or 'flops') are calculations involving decimal numbers in computers. They are particularly crucial for AI applications, which require billions or trillions of these calculations per second. When you hear about an AI chip performing at 'teraflops', it means it can handle trillions of these floating point calculations every second.

Therefore, training involves processing massive amounts of data, necessitating significant computational resources. While sustaining computing power at this scale, AI hardware's cost-effectiveness, performance, and power consumption are important considerations.

3.2. Inference

AI models are deployed in real-world environments after training. This stage is called inference. The models process new, real-world data to make valid predictions and draw accurate conclusions. While less demanding computationally, AI inference use cases have *different* computational requirements. These may include *latency, performance, memory, storage,* *energy efficiency, privacy*, and *scalability* requirements. The type of inference use case has a bearing on the type of hardware infrastructure required to run the AI models.

3.2.1 AI on the cloud

LLM chatbots like OpenAI's ChatGPT, for instance, are computationally very intensive, requiring significant processing power and memory. Such extensive utilisation of compute hardware also has extensive energy demands and heat generation. Therefore, these AI systems are deployed in data centres that have the necessary high-performance hardware along with dedicated cooling and power infrastructure to service high volume, sustained AI workloads.[22]

3.2.1.1 AI run by Data Centres

Data centres have emerged as an integral part of AI-on-the-cloud infrastructure. Data centres are large-scale facilities that host hardware at scale and thus efficiently provide computational resources.[23] They are increasingly used to train and run large AI models, and barring a few considerations such as software platform support, most customers running AI workloads do not have to worry about the minutiae of the computing hardware of these data centres.[24]

Optimised not just for AI workloads but also other HPC tasks, data centres house dense clusters of specialised hardware.[25] The image below shows a cluster of GPUs in a data centre.

A GPU Cluster at a Data Centre © CSIRO [26]

Along with specialised processing chips, they have to also feature enormous amounts of high-performance memory and storage systems. All these components need to be connected via highly performant networking and interconnect solutions to enable rapid data transfers without delays.[27] These networked clusters of compute infrastructure are so well-coordinated that they are usually considered a *single* unit of high-performance compute themselves. Consequently, as data centres composed of these compute clusters become the sole choice available to customers running AI workloads, they have become the *de facto* unit of AI compute infrastructure.[28]

NVIDIA's DGX200 cluster is a good example of this. They use multiple interconnected nodes, each powered by NVIDIA's latest GPU technology, to deliver high-performance compute for AI training and inference tasks.

As a unit of compute, scalability is also a key requirement of these centres.[29] They have to be designed to be modular with easy expansion. They also need to be amenable to quick upgrades to newer generations of specialised hardware. The goal is to have data centres that can evolve further as AI workloads change, without needing to be redesigned or rebuilt.

All of the operations in the data centre require large amounts of uninterrupted energy. The largest data centres can consume up to tens of hundreds of megawatts,[30] and therefore, also require effective thermal management with advanced cooling solutions. Additionally, they need sufficient levels of redundancy built in to ensure reliability and minimise downtime. These high standards of operation impose significant hardware investments and innovation.

Microsoft is powering its data centres with Small Nuclear Reactors (SMRs) to meet the consistently high-power requirements.

An AI-focused Data Centre © Free Malaysia [31]

However, the AI in the cloud, powered by data centres that are often large distances away from end-users, cannot cover all AI use cases. For instance, highly powerful computing resources of a data centre will not always be able to meet the ultra-low latency operations required for ML workloads for a fitness tracker or a smartwatch, since user data needs to contend with the transmission and processing times associated with cloud computing. These use cases and devices require the processing and source of data input to be in close proximity.

3.2.2 AI on the Edge

There are many use cases such as autonomous driving or traffic management, where it would be impractical to centralise the computing power due to the AI algorithm needing to operate with very low latency or to preserve data privacy. In such a scenario, computing resources are instead placed closer to the source of inputs—the users at the network's edge. These systems are referred to as Edge AI.[32]

Edge AI can vastly increase the scope of AI applications in the real world. This has easy and innovative use cases across multiple sectors. These possibilities have emerged due to advances in computing infrastructures, which have become small, fast, efficient, and specialised enough.

For instance, Edge AI systems can assess sensor data in industrial equipment to detect anomalies and potential issues early and minimise downtime.[33] In these systems, the Edge AI hardware must continuously monitor and assess data in harsh conditions and balance edge and cloud communication.
Remote monitoring through continuous health data analysis using Edge AI with secure processing and low power consumption can be used to improve patient outcomes.[34] Edge AI can enable user health data to be stored and

processed locally as opposed to being sent to the cloud. This ensures privacy and gives users greater control over their personal data.

In the case of autonomous vehicles, Edge AI, with low latency, high energy efficiency and bespoke processing capabilities for interpreting multiple sensor inputs, is necessary to process data from cameras, GPS and other sensors for real-time decision-making.[35] Similar use cases exist in traffic management, agriculture, customer analytics, etc.

Even consumer laptops have AI models deployed on-device that run on customised mobile processors.[36] Each of these steps has a bearing on the hardware consideration for the computations. Low latency becomes essential to ensure a real-time, seamless experience. High processing power and memory efficiency are necessary to handle high-traffic applications.

4. Understanding the hardware that makes AI computation possible

The interaction of data and algorithms that results in functional AI models is powered by the underlying hardware architectures, enabling the computational prowess required for AI tasks ranging from facial recognition in smartphone photography to climate modelling on supercomputers.[37]

This document focuses on three key categories of AI hardware components: processing units or accelerators, memory and storage systems, and networking and interconnects infrastructure.

4.1. Processors/Accelerators: The 'engines'

The processors, or accelerators are the "engines" of the AI systems. They are responsible for crunching the complex mathematical and algorithmic operations over the course of the training as well as the inference stages. There are broadly four types of processing units.

4.1.1. Central Processing Units (CPUs)

CPUs are general-purpose accelerators that can handle a wide range of tasks, including AI workloads. These are at the heart of almost all consumer computing electronics like PCs, smartphones, and laptops.[38]

While they are flexible and easy to program, they may not provide the optimal performance for AI applications compared to more specialised accelerators. In an effort to remedy this, newer CPU Systems-on-Chips (SoCs) dedicate a section of the silicon die to a specialised architecture meant only for running AI workloads locally. Often referred to as Neural Processing Units (NPUs), they are intended to address the traditional weakness of CPUs at training or inference tasks.[39]

American companies, Intel and AMD, are the major players in the CPU market. While Intel maintains a significant market share and has long-standing established relationships with OEMs, AMD's processors have wrested market share away in recent years with gains in performance and power/thermal efficiency.[40] Both AMD and Intel maintain a quasi-duopoly on the x86 Instruction Set Architecture (ISA) that forms the foundation of their chip designs. NVIDIA, known for their GPUs, are a recent entrant into the CPU space, with their ARM-based "Grace" processor.[41]

An Instruction Set Architecture (ISA) is the low-level language that a computer processor understands - essentially a set of commands that software can use to control the chip. ISAs are specific to particular processor families, such as x86 for Intel and AMD's CPUs, and ARM for many mobile and embedded processors. Processors using the same ISA can run the same software, regardless of who manufactured it.

GPUs also have their own proprietary ISAs defined by their manufacturers (e.g., NVIDIA's CUDA). However, these GPU ISAs are not as widely used or licensed as CPU ISAs like x86 and ARM. Companies developing AI accelerators must either license an existing ISA (like ARM), create their own, or use an open standard like RISC-V. This choice has significant business implications - using x86 means dealing with Intel and AMD's duopoly, while creating a new ISA requires building an entirely new software ecosystem. Many AI chip startups opt for ARM or RISC-V architectures, as these provide these established ecosystems.

In the mobile devices market, the ARM-based accelerators dominate. UK-based ARM creates design blueprints for processing units and licences them out to other companies like Apple, Samsung and Qualcomm. The latter modify or incorporate the licensed ARM designs, make necessary customisations and create their own CPUs within Systems-on-Chips (SoCs). These customisations include adding unique features like NPUs.[42]

NPUs first began being integrated with CPUs in mobile chips. Apple's Neural Engine in A11 Bionic and Huawei's Kirin 970 are some examples. More recently, Intel and AMD have begun implementing NPUs on their desktop CPUs also.

Taiwan's TSMC is a major manufacturer of most of these cutting-edge accelerators. Samsung in Korea also manufactures advanced accelerators but has lower volumes. Meanwhile, Intel, which used to manufacture its own chips in the US and Israel, has encountered difficulties in remaining abreast of the manufacturing capabilities of TSMC. The US is therefore also trying to catch up and build capabilities to fabricate the most advanced processors.

While Intel has traditionally been a company that fabricates its own chips in-house, they have recently outsourced fabrication of some advanced chips to TSMC to take advantage of the latter's fabrication process advantages. Chip fabrication is an exceedingly expensive endeavour with fabrication facilities for leading-edge chips costing anywhere between $10 billion to $28 billion.

An Intel CPU © Intel [43]

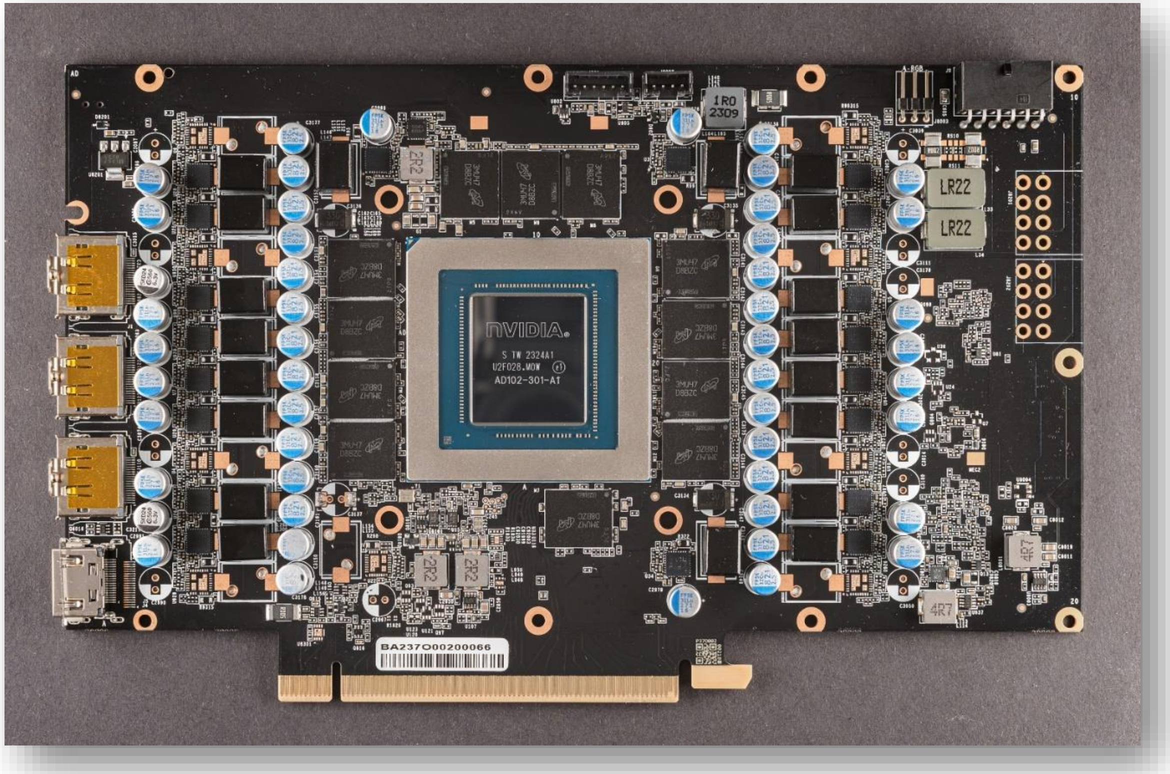
4.1.2. Graphics Processing Units (GPUs)

GPUs have become the dominant processing units for AI, particularly for training highly complex AI models.[44] Originally coined as a term for accelerators meant for rendering graphics in video games, and other Computer-Generated Imagery (CGI),[45] they are designed to perform parallel computations on large datasets, making them well-suited for the matrix operations of AI algorithms.[46]

GPUs have been instrumental in the rapid advancement of AI capabilities in recent years, partly due to their accepted prevalence in scientific computing tasks that have leveraged parallel processing capabilities, as well as increased ease of use in programming them.[47] Their use is prevalent in data centres specialised to run AI workloads.

NVIDIA has a monopoly in the GPU market and has captured nearly all of the market share, with AMD in a distant second place.[48] It has also pivoted majorly to AI and data centre-based GPUs away from its traditional gaming GPU roots. Intel also makes GPUs of its own but remains a minor player.[49]
GPU design is heavily concentrated in the US where NVIDIA, AMD and Intel are headquartered. NVIDIA and AMD are fabless companies that design and sell their own chips but do not manufacture them. The manufacturing is outsourced to East Asia where it is concentrated, particularly at TSMC in Taiwan. This combination of factors has prompted

investments in the US, Europe, China, and India to explore production capabilities and enhance supply chain resilience.[50]



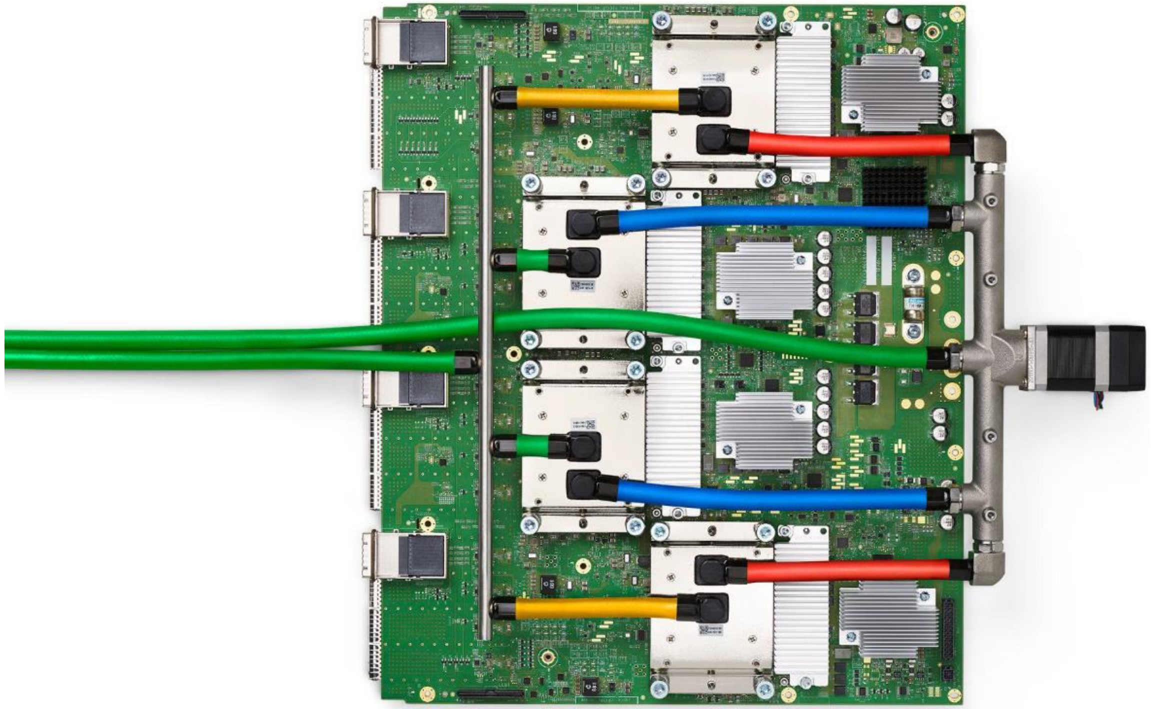
A GPU © Nvidia [51]

4.1.3. Application-Specific Integrated Circuits (ASICs)

Being the most specialised of the processing units, ASICs are chips that are designed from the outset for a specific set of tasks, such as AI inference, sometimes within more-specific temperature, power, and space thresholds as compared to CPUs, and GPUs etc.[52]

As such, they provide the highest performance and energy efficiency for their target workloads,[53] but as a trade-off, they lack the flexibility of other processing units. Examples of AI ASICs include Google's Tensor Processing Units (TPUs),[54] and Cerebras' Wafer-Scale Engine.[55] TPUs, for instance, are designed at the silicon level for AI workloads that take advantage of Google's TensorFlow framework. Since the specific range of operations enabled by this framework is known beforehand, the chip design can be optimised only for them. Because of this, TPUs can deliver high performance and energy efficiency for both AI inference and training tasks in Google's data centres.[56] Vast number of TPUs are used in compute clusters like data centres to train and run AI models. Because of their custom-nature, ASICs can prove to be expensive.[57]

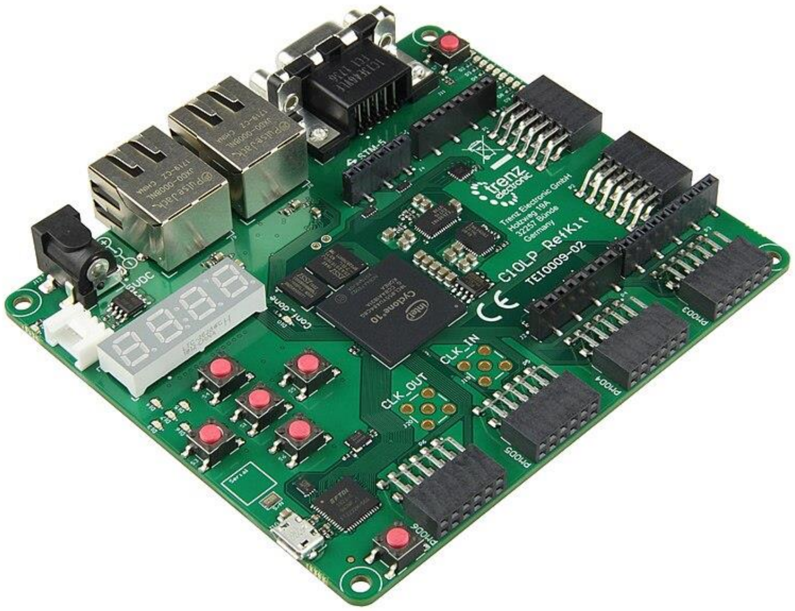
The market for ASICs Is less concentrated than that of GPUs and CPUs.[58] There are many big tech companies as well as startups building ASICs for AI workloads, with demand also similarly disaggregated in terms of volumes shipped. Given the large size of the market and scope for specialisation, companies try to find their own niche.[59] Google and Intel are the notable big players with other emerging ones like Graphcore, Cerebras Systems, Groq, Tenstorrent, Mythic and Blaize. However, like all advanced semiconductor manufacturing, there is significant market concentration in East Asia for manufacturing ASICs.[60]



A TPU © Google [61]

4.1.4. Field-Programmable Gate Arrays (FPGAs)

FPGAs are reconfigurable integrated circuits that can be programmed to perform specific tasks, including AI workloads. Their versatility lies in being able to switch between different types of workloads post manufacturing, treading a middle-ground between the flexibility of CPUs and the performance of ASICs. This makes them attractive for certain AI applications, especially in scenarios where algorithms evolve rapidly, but without the costs of leveraging ASICs.[62]



An FPGA Board © Altera [63]

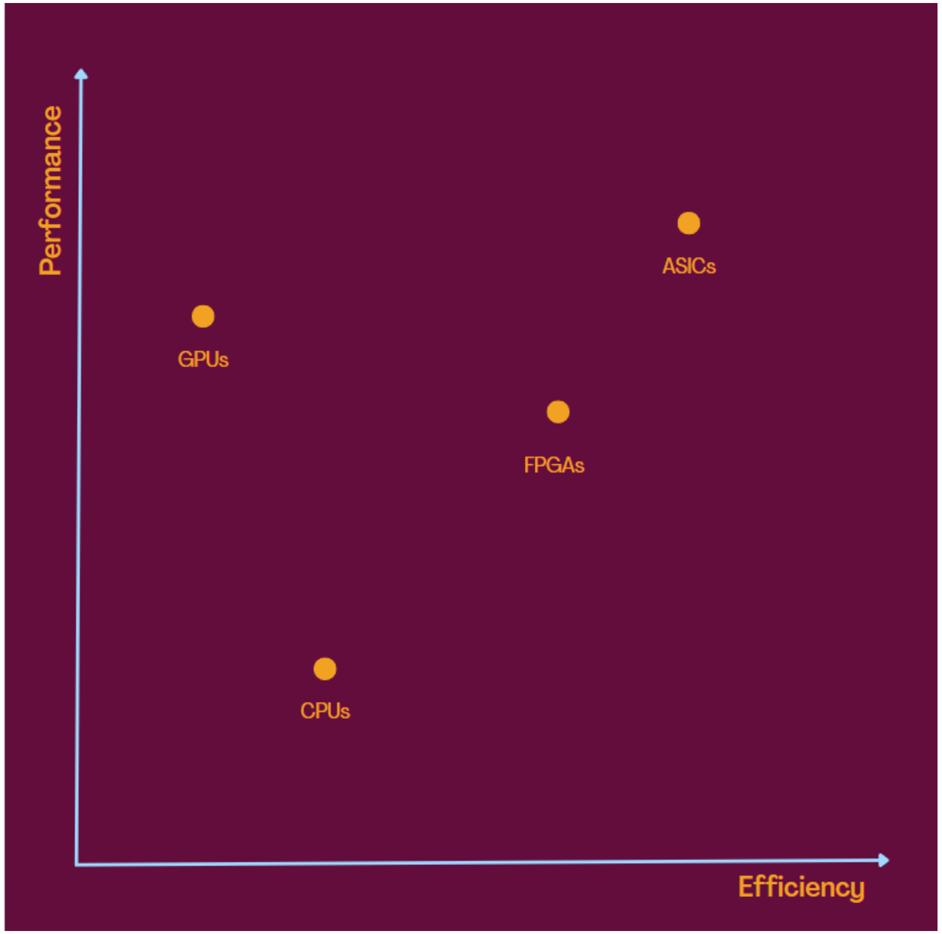
It also makes FPGAs an essential input in AI hardware and software R&D, allowing for experimentation and prototyping by researchers and students.[64] FPGAs are predominantly seen in areas like defence electronics, networking, space research and exploration etc where adaptability of functions is an important factor.[65]

Like the CPU market, the FPGA market is dominated by Intel and AMD again.[66] While Intel fabricates some FPGAs in-house, TSMC is again a major manufacturer of FPGAs.

We can evaluate different types of AI-specific accelerators across two key dimensions: performance, and efficiency.

"Performance" encompasses considerations that enable an accelerator to quickly process high volumes of data and complex AI workloads. This captures multiple metrics: raw processing power to crunch precise mathematical operations; high memory bandwidth and capacity to feed data to the processing unit; low latency to provide fast response times to end-users; scalability for tackling massive datasets and models as per use cases; and finally, the ease of programmability.

"Efficiency" encapsulates the cost and sustainability aspects of operating AI systems built using different accelerators. This covers metrics such as the energy consumption of powering and cooling the units, the upfront purchase costs and long-term operating expenses, and also the overall environmental footprint.



Source: Authors' Visualisation: A Rule of Thumb comparison of some popular accelerators

Criteria	CPUs	GPUs	FPGAs	ASICs
Processing Peak Power	Moderate	High	Very High	Highest
Power Consumption	High	Very High	Very Low	Low
Flexibility	Highest	Medium	Very High	Lowest
Training	Poor at training	The only production-ready training hardware	Not efficient	Potentially, best for training, but not available yet
Inference	Poor for inference at scale, useful for smaller workloads on the edge	Average for inference	Best for inference	Efficient at inference for highly tailored workloads

Source: Author's Recreation[67]

4.2. Memory and Storage Systems: The 'fuel lines' and 'fuel tanks'

If processing units are visualised as the "engines" of AI systems, memory and storage systems are the veritable fuel lines and the fuel tanks that ensure that these systems run properly. AI models need to be fed with extremely large volumes of data as they are being trained.[68] The models need to be able to store this data. Further, during inference, they need to be able to access input and return output data rapidly, consistently, and reliably. This storage, access and transfer of data is made possible by different types of memory and storage systems.

4.2.1. Memory

4.2.1.1. Random-Access Memory (Dynamic RAM)

Dynamic RAM (DRAM) is the most common type of main memory used in AI systems. It offers relatively high capacity and bandwidth but compared to other types of memory, it can be a bottleneck for data-intensive AI workloads.[69] It is usually leveraged by processing units like CPUs, and is usually physically situated away from the latter. A combination of CPUs and DRAM is the most common configuration found in consumer computing devices like smartphones and PCs.[70]

4.2.1.2. High Bandwidth Memory (HBM)

HBM is a specialised type of memory that provides much higher bandwidth than traditional DRAM. It is increasingly used in high-performance AI hardware. HBM stacks memory chips vertically and places them closer to the accelerator.

HBM modules take advantage of advanced packaging technologies to stack modules vertically, and are placed much closer to the logic processing unit, on the silicon die itself. Therefore, HBM is better placed for latency-sensitive tasks since the physical distance that electrical signals need to cross between processing and storage is lower.[71] This significantly reduces the time and energy required to move data, enabling more complex AI models to run efficiently. HBM has been crucial in advancing areas like real-time video analysis and scientific simulations.[72]

4.2.1.3. Graphics Double Data Rate (GDDR)

GDDR is a type of DRAM that offers much higher bandwidth and lower latency than standard DRAM. GDDR can be used in GPUs for AI workloads when cost is a primary concern or when the workloads and datasets are relatively small.

In contrast to HBM, GDDR memory modules are situated on the board instead of on the GPU's silicon die. Therefore, it is generally less efficient than HBM in terms of power consumption and suffers comparatively on latency and bandwidth metrics. On the flip side, GDDR offers lower cost, wider availability, and lower memory requirements.[73]

Given the high barriers to entry, the global market for memory systems is oligopolistic. Samsung, SK Hynix and Micron Technology have almost all of the market share. China's YMTC has suffered in its capability to develop HBM production facilities due to US sanctions.[74] Samsung and SK Hynix are South Korean companies while Micron is American. Due to the highly commoditised nature of memory chips, all the major players in the DRAM market are Integrated Device Manufacturers, and manufacture their own memory chips.[75]

4.2.2. Storage

While memory systems focus on rapid data access for active computations, data storage is crucial for maintaining vast amounts of data that AI systems need for training and inference. AI datasets can easily reach into the petabytes,[76] requiring massive storage capacity and fast access speeds. The viability of a data storage architecture is dependent on a variety of factors such as scalability, availability, security, performance, and resiliency.[77] Storage solutions can therefore be configured to favour one or some of these factors based on system requirements. Key components include:

4.2.2.1. Solid State Drives (SSDs)

These offer faster read and write speeds compared to traditional hard disk drives, making them valuable for AI workloads that require frequent data access. NVMe (Non-Volatile Memory Express) SSDs, in particular, provide high-speed storage and retrieval, low latency, and high-throughput. NVMe SSDs are a popular choice in data centres to host the datasets required to run AI models. [78]

In addition to Samsung, SK Hynix and Micron, Western Digital, Seagate and Kioxia (Japan) are some notable players in SSD manufacturing. The production is mainly concentrated in South Korea, Japan, China and the US.[79]

4.2.2.2. Hard Disks Drives (HDDs)

While slower than SSDs, HDDs offer larger capacities at lower costs, making them suitable for storing vast datasets used in AI training. They are often used in tiered storage systems, where frequently accessed data is stored on faster SSDs while less frequently used data resides on HDDs.[80]

The HDD industry has consolidated into three major players–Seagate, Western Digital and Toshiba that have almost all of the market share.[81]

4.2.2.3. Tape Storage

For archival purposes and extremely large datasets that do not require quick retrieval, tape storage provides a cost-effective solution. While access times are slow, tape storage can be useful for storing historical data or backups of AI models and datasets.[82]

Characterised by limited suppliers of key components, most of the tape storage production happens in Japan and the US by companies like IBM, Quantum and Fujifilm.[83]

Memory and storage solutions may not directly form part of the calculus when it comes to taking investment decisions for AI infrastructure. Usually, the choice of the processing unit and use-cases will also determine the choice of memory and storage solutions due to tight integration.[84] However, as mentioned, memory chips are also highly commoditised and subject to intense geopolitical and geoeconomic pressures, due to the nature of global value chains as well as a steadily escalating US-China trade competition.[85]

Emerging memory technologies and architectures, such as processing-in-memory (PIM), are also being explored to address the unique challenges of AI workloads, such as the need for high capacity, low latency, and energy efficiency. The choice of architecture, and storage configurations like In-Memory Data Grids, and Distributed Storage Systems, can significantly impact the speed of data ingestion for AI training, the responsiveness of AI inference systems, and the overall cost and energy efficiency of AI infrastructure.

Criteria	DRAM	HBM	GDDR	SSDs (NVMe)
Purpose	General-purpose main memory primarily used by CPUs	High-performance AI and enterprise workload-specific memory	Graphics rendering	Long-term data storage and retrieval
Bandwidth (higher is better)	Moderate	Very High	High	Low
Latency (lower is better)	Moderate	Low	Low	High
Strengths	High capacities and supply, low cost	Highest bandwidth, low latency	Good balance of performance and cost	Large capacity, fast for storage compared to legacy options like HDDs and Tape
Weaknesses	Can be a bottleneck for data-intensive AI workloads	Expensive, complex manufacturing; concentrated supply chains	Less efficient than HBM	Not as fast as RAM for active computations
Supply Chain Considerations	Oligopolistic market dominated by Samsung, SK Hynix, Micron	Same as DRAM; China's YMTC facing challenges due to US sanctions	Similar to DRAM, produced by same major players	Many suppliers including Samsung, SK Hynix, Micron, WD, Seagate, Kioxia

Source: Authors' Visualisation

4.3. Interconnects and Networking Capabilities: The 'highways'

Interconnects serve as the metaphorical highways of AI hardware, enabling data transfer between processing units, memory, and storage. Interconnects and networking capabilities are essential for enabling efficient communication between AI hardware components, both within a single system and across multiple data centres.[86] Broadly, there are three kinds of interconnect and networking technologies leveraged across the AI hardware technology stack:

Chip-to-chip interconnects like chiplets are becoming inextricably linked with advanced packaging technologies, and they are part of an emerging technological shift in chip design, fabrication, and packaging. Alongside chiplets, vertical (3D) packaging technologies etc., are also becoming more commonplace in newer AI processors. As processing units become more and more complex, on-chip interconnects become critical for their performance.

4.3.1. On-Chip Interconnects

On-chip interconnects facilitate communication between different parts within a system-on-chip (SoC). An SoC contains various aforementioned components like a CPU, a GPU, memory and storage.

As Moore's law[87] hits the limitations of physics,[88] chips have pivoted towards using separate parts of the chip for separate tasks, integrating them on the same foundational structure of the chip.[89] The resulting *chiplet* would be able to retain better performance, improvements in thermal and power requirements and easier manufacturing.[90] The on-chip interconnects enable fast and efficient communication within the chiplets.[91]

As of now, very few firms such as TSMC (Taiwan), Samsung (South Korea), and Intel (US) etc, have the ability to develop advanced packaging technologies.[92] The fabrication stage of the semiconductor global value chain (GVC) is already dominated by these players.[93] This dominance is further amplified in AI chips.[94]

4.3.2. Chip-to-Chip Interconnects

Chip-to-chip interconnects enable high-speed communication channels between multiple chips within the same system.[95] Technologies like NVIDIA's NVLink or Intel's Ultra Path Interconnect (UPI) enable high-bandwidth, low-latency connections between multiple NVIDIA GPUs or Intel CPUs,[96] respectively. This is what allows for the creation of powerful compute clusters of processing units,[97] and therefore, for scaling up computations and performance to tackle complex AI workloads.

However, due to the degree of their integration with the architecture of the processing units themselves, these interconnect technologies are typically proprietary. For instance, AMD's GPU offerings are not compatible with NVLink, as the latter requires specific hardware and software support from NVIDIA.[98] Therefore, the choice of processing unit also determines the nature, and capabilities of downstream technologies essential for the functioning of AI compute infrastructure.

In an effort to create an open industry standard for GPU-to-GPU interconnect, an industry consortium has proposed the development of an Ethernet-based interconnect called the Ultra Accelerator Link (UALink).

Expansion standards such as Peripheral Component Interconnect Express (PCIe) is an extensively used interconnect standard that enables communication between a wide range of hardware components, including processing units, storage, and other peripherals such as networking cards etc. It is maintained by a consortium of 900+ companies. While not as fast as specialised interconnects like NVLink, its ubiquity, high bandwidth and low latency make it suitable for workloads that require fast data transfer. ASICs and FPGAs are usually dependent on PCIe for connectivity.

The capabilities of processing units, motherboards, and storage solutions are often tied to their support of the newest iteration of the PCIe standard. The total number of PCIe lanes supported by an AI system directly affects its scalability as more GPUs etc can be connected to it.[99]

Therefore, the choice of platforms and other peripherals directly correlates with the choice of processing units or memory solutions. Higher end platforms, which can support a higher number of interconnects, will cost more, and may be in higher demand, amidst potential supply-chain constraints. The question of vendor and ecosystem lock-in also becomes pertinent, since the ability to mix-and-match similar processing units sourced from different vendors remains limited.

4.3.3. Node-to-Node Interconnects

These refer to communication channels between different "nodes" of a compute cluster. Examples of a compute cluster could range from simple arrangements of home PCs connected via LAN to more complex ones such as server clusters in data centres.[100] Node-to-Node interconnects enable this communication providing high-speed data transfers and low-latency responses.

The most common technologies used here are Ethernet and InfiniBand. The former is an open 50-year-old connectivity technology that sees both commonplace use for providing broadband internet to homes, as well as for communications between data centres and enterprises. The latter was developed to be a more performant replacement to Ethernet in the 1990s and initially found success only in High-Performance Computing (HPC) environments such as supercomputers.[101] It has since become a widely deployed interconnect technology in HPC data centres, and cloud computing.[102]

While Ethernet has shown its age, the technology is being optimised for high performance computing and AI networking. To that end, the Ultra Ethernet Consortium (UEC) initiative has been proposed with the backing of major players in the AI hardware and software industry alike, such as AMD, Broadcom, Intel, Microsoft, and Meta etc.

Estimates suggest that upwards of 90% of scalable AI systems use this networking architecture.[103] InfiniBand's main strength over existing Ethernet solutions is its relatively high data integrity during the transfer of data between nodes. A lack of data integrity can slow down AI training workloads, and therefore, has a direct impact on costs and efficiency in the AI value chain.[104] On the other hand, Ethernet has a slightly higher bandwidth ceiling,[105] and it has relatively lower implementation costs.

Since InfiniBand is an open industry standard interconnect specification,[106] it means that other firms can still produce InfiniBand solutions to enable high-bandwidth HPC networking for their AI system offerings. However, NVIDIA's partnerships (with leading server vendors and data centre operators),[107] continued innovation of the standard,[108] and its dominant share in the upstream GPU market[109] have ensured that its InfiniBand-based products command the lion's share of the downstream networking solutions market as well.[110] The openness of this standard theoretically makes customers of InfiniBand technology less susceptible to vendor lock-in; in practice however, the combination of the above factors create network effects that make it difficult for competitors to unseat NVIDIA's market dominance.

Investments in advanced interconnect technologies can be as important as investments in processing units themselves for building a presence in the AI global value chain (AI GVC). Standards like CXL, UALink, UCIe, and UEC are expected to play a significant role in the future of AI hardware, providing a standardised, interoperable foundation for high-performance, multi-vendor systems. A long-term policy strategy to incentivise homegrown hyperscalers to add to, and implement open standards like UEC can lower entry barriers for small AI data centre players. The presence of a large number of such networking-solution providers can potentially exert a countervailing pressure on NVIDIA's market concentration in this downstream market.[111]

Criteria	On-Chip Interconnects	Chip-to-Chip Interconnects	Node-to-Node Interconnects
Scale of Integration	Within SoC or across chiplets	Between chips in a system	Between separate compute nodes (ex: data centres)
Key Technologies	Chiplets, Through-Silicon Vias (TSVs), Silicon interposers	NVLink, Intel UPI, PCIe, CXL, UCIe	InfiniBand, Ethernet
Proprietary vs Open	Mostly proprietary	Mix of proprietary and open standards	Predominantly open standards
Market Concentration	High (TSMC, Samsung, Intel)	Moderate to High (NVIDIA, Intel dominate proprietary solutions)	Moderate (NVIDIA dominates InfiniBand-based solutions market)
Impact on AI Performance	Critical for chip efficiency	Enables multi-GPU/CPU systems and affects scalability for workloads	Crucial for distributed AI training where workloads are spread across multiple nodes
Future Trends	3D packaging, advanced chiplet integration	Adoption of CXL and UCIe standards	Ultra Ethernet Consortium (UEC) development
Supply Chain Considerations	Limited to advanced foundries; potential bottleneck in AI chip production	Dependent on GPU/CPU manufacturers, PCIe offers some flexibility for vendor choices	Broader supplier base for Ethernet; InfiniBand solutions largely from NVIDIA

Source: Authors' Visualisation

Box 1: Distinguishing AI Hardware from General-Purpose Computing Hardware

General-purpose hardware, which can include CPUs and GPUs meant to be end-consumers, can be considered as a superset that includes hardware components used for a wide range of computing tasks, including AI workloads. AI hardware, on the other hand, is a subset of this superset, specifically designed and optimised for AI applications.

For instance, both the NVIDIA RTX 4090 gaming GPU, and the AI-specific NVIDIA H100 GPU use 16 lanes of the common PCIe Gen 5 interconnects to interface with the system's other components. However, the RTX 4090 is not designed at the architecture level for relatively high levels of precision in matrix arithmetic operations needed for AI workloads. On the other hand, the H100 is designed to do just that at very high levels of precision. GPUs like the H100 also support arithmetic instruction formats that can be leveraged for processing AI workloads faster; this is something that consumer-oriented GPUs like the RTX 4090 are not designed to support, and are consequently, significantly slower at the same tasks.

Similarly, the RTX 4090 is equipped with only 24 GB of slower GDDR6X memory as compared to the H100's 80 GB of much faster HBM3e memory. Importantly, the RTX 4090 can only interact with other GPUs on the same system via PCIe, while the H100 has NVLink, which as mentioned earlier, is a much faster chip-to-chip interconnect.

Therefore, despite sharing similar characteristics and components, general-purpose computing hardware may be inadequate for the purposes of running AI workloads. Consumer-grade GPUs may be able to support AI workloads at a smaller scale, but they are not designed to fully replace dedicated AI hardware in large-scale deployments or HPC environments.

Despite the limitations of consumer-grade computing hardware like the RTX 4090, Chinese firms attempted to use these GPUs as a substitute for dedicated AI hardware like the H100 in the wake of the October 2022 unilateral US export controls on advanced chips. Subsequently, the export control thresholds were expanded to cover the computing capabilities of both the H800 and the RTX 4090 GPUs.

5. Understanding the software that supports AI hardware

While the interaction of data and algorithms that results in functional AI models is powered by the underlying hardware, the hardware itself is dependent on certain software components. Much like how consumer PCs and smartphones are dependent on their operating systems, the performance of AI processing units, also known as AI accelerators, can only be realised through the software ecosystems that support their deployment.[112] Software frameworks, libraries, and programming languages harness the processing capabilities of accelerators and simplify the development process for models and applications that run on them. As such, whether or not an AI accelerator meets with widespread success and adoption in the industry is heavily dependent on the maturity and ease of use of their compatible software ecosystems.

5.1. The AI Software Ecosystem

The AI software ecosystem broadly consists of: **AI frameworks**, **programming languages**, and **programming platforms**. This paper focuses on the software ecosystem relevant in the downstream stages of the AI value chain, i.e., models, and applications. While outside the scope of this paper, various software and data analysis tools also exist for processing data before it is used for training and inference.

The term **AI Frameworks** broadly refers to the pre-made tools and libraries that developers can use to create, train, and test AI models.[113] Frameworks relieve developers of the need to be minutely aware of the complexities of managing the hardware's low-level operations (like memory management) and are usually hardware-agnostic – which means they can run on CPUs and GPUs as well as other specialised accelerators. That said, many frameworks have optimisations for specific chip architectures.[114] Prominent examples of AI frameworks include TensorFlow, PyTorch, and MXNet.

Programming languages (such as Python and Julia),[115] used in AI development serve as the interface between developers and AI frameworks. Python, in particular, has become the de facto standard for AI developers since it is simple and easy to learn, and has a mature and extensive ecosystem of libraries useful for scientific computing.[116]

As mentioned earlier, high-level languages and frameworks aim to be hardware-agnostic, but developers often rely on lower-level, accelerator-specific features to achieve optimal performance. This is where a custom **software development platform** can come into play.

TensorFlow is heavily optimised for Google's TPUs allowing it to excel at matrix multiplication algorithms that are fundamental to many AI workloads. AMD/Xilinx's Vitis AI is designed to simplify the deployment of AI inference workloads on Xilinx FPGAs.

An AI research project might prefer PyTorch due to its popularity in the academic community and support for a wide range of NVIDIA's GPUs (that have been a mainstay in scientific computing community); however, a production-focused project might choose TensorFlow for its hassle-free integration with Google's cloud services. The framework chosen for an AI workload, therefore, influences the choice of cloud service platforms as well as the underlying choice of accelerators being offered by the cloud platform.

Programming platforms like NVIDIA's proprietary CUDA (Compute Unified Device Architecture) are a prime example.[117] CUDA encapsulates a suite of software tools, libraries, and APIs specifically designed for NVIDIA GPUs. It provides a familiar programming interface to developers using common languages like C, C++, and Fortran, and allows them to write code that can directly access the parallel computing capabilities of the GPU to greatly speed up computing tasks.

AI developers typically interact with CUDA indirectly through frameworks like TensorFlow or PyTorch. Since most popular AI frameworks are open-source, this creates an interesting situation where developers seeking to use them to benefit from the open-source benefits of community contributions, transparency, and rapid innovation must rely on a proprietary hardware and software platform to achieve peak performance.

5.2. CUDA and its absent competition

CUDA was developed to address the challenges in programming GPUs for general-purpose computing tasks. GPUs could potentially accelerate heavily parallelised workloads (graphics rendering was just such a task), but before CUDA, programming for them required low-level coding skills and a deep understanding of the underlying chip architecture.[118]

NVIDIA tackled this problem in two ways.[119] First, NVIDIA introduced a GPU chip design architecture that was composed of smaller programmable units, termed generally in the industry as "shader units". Second, NVIDIA created the CUDA software development platform that specifically allowed coders to write programs for these units (now referred to as "CUDA cores") on its GPUs.[120]

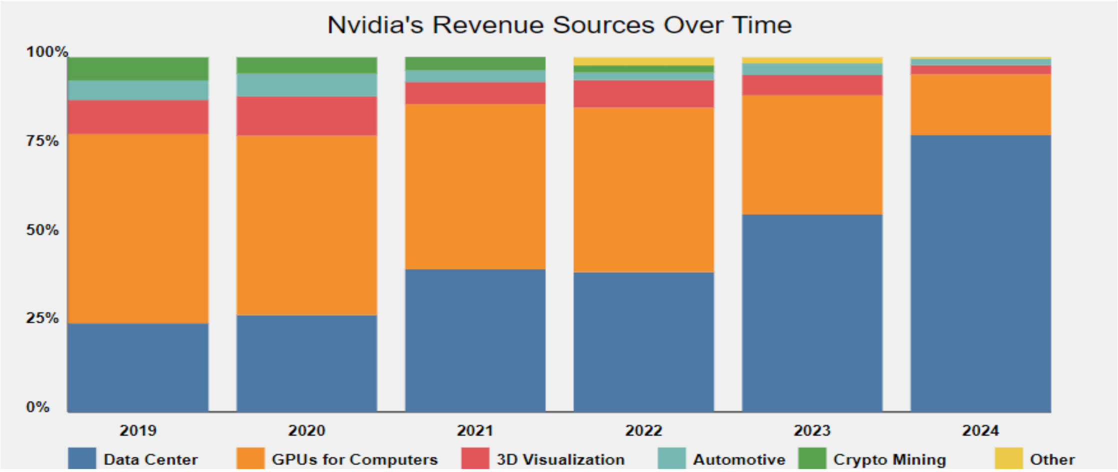
The CUDA platform was designed to attract developers by advertising the massive parallel computing power of GPUs on the back of very little in the way of learning barriers, by highlighting its similarities with other common programming languages. In a nutshell, CUDA as a software platform is inextricably integrated with the silicon-level hardware architecture.

While CUDA and NVIDIA GPUs saw its initial steady adoption by the global scientific computing community, they are now essential tools in any field engaged in AI applications. As the CUDA ecosystem has matured, it has created a virtuous feedback loop of innovation and adoption, with an increasing number of developers contributing to an increasing number of optimised libraries and tools. This continues to further reinforce NVIDIA's market position across both consumer (graphics rendering and gaming-oriented) and enterprise/AI segments.

This closed CUDA-GPU integration means that potential competitors are prevented from leveraging the CUDA platform, as NVIDIA's hardware architecture IP remains proprietary.[121] CUDA itself is free to use, and NVIDIA invested in optimising different sub-platforms of CUDA meant for specific use-cases in industry and research, such as Robotics, Machine Learning, Data Centres etc. The commonality afforded by the platforms ensured that applications across a wide range of domains would also be compatible with all NVIDIA GPUs. NVIDIA invested heavily in training courses and outreach in this regard (and continues to do so), ensuring that both academia and industry adopted its GPUs for their needs.[122]

The CUDA ecosystem has therefore created two-sided network effects stemming from both developers (supply) and industry (demand) utilising the same GPUs and software platform.[123]

CUDA's exclusivity has been a key factor in NVIDIA's dominance in the AI hardware market. The closed integration of the software development ecosystem with the hardware has enabled NVIDIA to charge supra-competitive prices for its GPUs across both gaming, and enterprise sectors.[124]



Source: Authors' Visualisation (Data from Visualcapitalist)[125]

As of now, AMD's GPUs such as the Instinct MI300X cost substantially less than NVIDIA's flagship offerings, such as the H100. However, market trends suggest that the uptake of AMD GPUs has been primarily due to the global demand for AI compute, which NVIDIA's production runs cannot fulfil.

5.3 CUDA alternatives

Several alternative software ecosystems to CUDA exist; however, these have struggled to match CUDA's maturity and performance stemming from NVIDIA's first-mover advantage and the network effects created by its large user-base. The most prominent competitor to CUDA is AMD's Radeon Open Compute (ROCm).

ROCm is a platform designed for use with AMD's GPUs, providing a suite of software tools and libraries to developers, similar to CUDA. While relatively new and lacking in overall support,[126] ROCm has two key benefits: **first**, it includes an abstraction layer, HIP (Heterogeneous-Compute Interface for Portability),[127] that allows developers to convert CUDA applications easily to run on AMD GPUs in a short timeframe. **Second**, its open-source nature potentially allows for long-term developer buy-in, and crowdsourced additions to its range of libraries. These two factors offer a major value proposition for developers and organisations concerned about vendor lock-in.

CUDA has undoubtedly accelerated the adoption and innovation of AI. However, from a policy perspective, it is a case study that highlights the unsavoury implications of proprietary software ecosystems in the AI hardware market. Besides market concentration risks, vendor lock-in, and other competition barriers, nation-states seeking to build sovereign AI infrastructure using GPUs will have to contend with the strategic dependency associated with being reliant on a single provider like NVIDIA.

Box 2: Translation Layers

Translation layers are software that allow code written for a particular hardware architecture to run on a different architecture. They essentially "translate" this code between disparate systems, and therefore, enable application compatibility across GPUs from different vendors.

As mentioned earlier, AMD's HIP can be considered a translation layer; however, it requires developers to manually port CUDA applications to run on AMD GPUs. However, a true translation layer allows for CUDA applications to interface with a non-NVIDIA GPU as if it were one, on-the-fly.

The most prominent example of a translation layer is ZLUDA, which allowed first, Intel, and subsequently, AMD GPU users, to run CUDA applications natively without the need for developer intervention, or source-code generation as an intermediate step. Despite not providing 100% compatibility or performance, ZLUDA received developer interest, and AMD funded the open-source project until recently.

AMD's withdrawal of support has been linked to NVIDIA's reiteration of CUDA licensing terms, which prohibits its use for the development of ZLUDA-like translation layers. While no overt legal action has been undertaken by NVIDIA, it is clear that the development of CUDA translation layers threatens its market position and lowers the value proposition of its GPUs on price-to-performance metrics. However, ZLUDA development continues with plans to support multiple GPU architectures.

6. Conclusion

This primer hopes to serve as a foundational resource for understanding the key facets and components of AI compute hardware. Policymakers must gain a comprehensive understanding of this hardware that powers transformative AI technologies.

The long-term implications of their hardware choices are magnified when we consider that computing infrastructure under national missions like INDIAai are expected to not only be effective, versatile, and efficient, but also future-proof. This document demonstrates how factors like performance, efficiency, cost, and the availability of a robust and developer-friendly software ecosystem play crucial roles in determining the suitability of different hardware options for various AI applications.

GPUs remain the popular choice for AI computing. Overreliance on a single GPU vendor or proprietary technologies can lead to strategic dependencies for nation-states, high switching costs and vendor lock-in, as well as a reduced scope for competition and innovation. It is useful to consider alternatives like ASICs and FPGAs while taking note of their technical characteristics, trade-offs, and market dynamics.

Given the importance of this hardware, long-term national strategies for building compute infrastructure should encompass exploring and supporting the development of alternative hardware and software solutions to mitigate the aforementioned risks. Future research documents will identify specific policy levers for AI compute governance and pathways through which nation-states can develop and maintain strategic footholds in the compute hardware global value chain.

7. Glossary

A

- **AI (Artificial Intelligence): A broad term encompassing technologies that enable computers to mimic human intelligence, such as learning, problem-solving, and decision-making. The sources primarily focus on AI powered by Machine Learning.**
- **AI Accelerator:** See *Processing Unit.*
- **Algorithm:** A set of instructions or rules that a computer follows to solve a problem or complete a task. **In the context of AI, algorithms form the "brains" of AI models, learning from data to make predictions.**
- **Application-Specific Integrated Circuit (ASIC): A type of processing unit custom-designed for a specific task, such as AI inference. ASICs offer the highest performance and energy efficiency for their target workloads but lack flexibility.** Examples: Google's Tensor Processing Units (TPUs), Cerebras' Wafer-Scale Engine.
- **ARM: A UK-based company that designs processing unit blueprints and licenses them to other companies like Apple, Samsung, and Qualcomm. ARM processors dominate the mobile device market.**

C

- **Central Processing Unit (CPU): A general-purpose processor that can handle a wide range of tasks, including AI workloads. CPUs are found in most consumer electronics like PCs, smartphones, and laptops.** While flexible, CPUs may not be as performant as specialised processors for AI. Key manufacturers: Intel, AMD.
- **Chiplet: A modular approach to chip design where separate parts of a chip are dedicated to specific tasks and integrated onto the same foundational structure.** This allows for better performance, improved thermal and power efficiency, and easier manufacturing.
- **Cloud AI: AI systems deployed in data centres, providing computational resources remotely.** Suitable for computationally-intensive tasks that require significant processing power and memory.
- **Compute Cluster: A group of interconnected computers (nodes) that work together to perform complex computations.** Examples range from home PCs connected via LAN to server clusters in data centres.
- **Compute Unified Device Architecture (CUDA): NVIDIA's proprietary software platform specifically designed for programming NVIDIA GPUs.** CUDA provides a familiar programming interface and allows developers to access the parallel computing capabilities of GPUs. **It has played a significant role in NVIDIA's dominance in the AI hardware market.**

- **CXL (Compute Express Link): An open industry standard for high-speed CPU-to-device and CPU-to-memory interconnects, expected to play a significant role in the future of AI hardware.**

D

- **Data Centre: A large-scale facility that houses and efficiently provides computational resources, often used to train and run large AI models.** Data Centres contain clusters of specialised hardware, including processing units, memory, storage, and networking solutions.
- **Deep Learning**: A subfield of Machine Learning that uses artificial neural networks with multiple layers to analyse and learn from data.
- **Dynamic Random-Access Memory (DRAM):** The most common type of main memory used in AI systems. **It offers relatively high capacity and bandwidth but can be a bottleneck for data-intensive AI workloads.**

E

- **Edge AI: AI systems where computational resources are located closer to the source of data, such as on edge devices.** This enables low-latency operation and data privacy. Examples: autonomous vehicles, fitness trackers, smartwatches.

- **Epoch: One complete pass of an entire dataset through an AI model during training.**
- **Ethernet: An open standard networking technology used for communication between computers and other devices.** It's widely used in both home and enterprise networks, including data centres. **While not as performant as InfiniBand for AI training, it offers higher bandwidth and lower implementation costs.**

F

- **Field-Programmable Gate Array (FPGA): A reconfigurable integrated circuit that can be programmed to perform specific tasks, including AI workloads. FPGAs offer a balance between CPU flexibility and ASIC performance.** They are commonly used in defence electronics, networking, and space research. Key manufacturers: Intel, AMD.
- **Floating Point Calculation: A mathematical operation involving decimal numbers, performed by computers.** AI training often requires billions of trillions of floating-point calculations.

G

- **General-Purpose Technology:** A technology with the potential to have a transformative impact across various sectors and industries. **AI is considered a General-Purpose Technology.**
- **Global Value Chain (GVC):** The interconnected network of activities involved in the design, production, distribution, and use of a product or service across different geographical locations. **The AI hardware market has a complex GVC, with concentration in certain regions and companies.**
- **Graphics Double Data Rate (GDDR): A type of DRAM that offers higher bandwidth and lower latency than standard DRAM.** Often used in GPUs for AI workloads when cost is a concern or datasets are relatively small. **It's generally less efficient than HBM.**
- **Graphics Processing Unit (GPU): A type of processing unit originally designed for graphics rendering but now widely used for AI, particularly for training complex models. GPUs excel at parallel processing, making them well-suited for AI's matrix operations. NVIDIA dominates the GPU market.**
- **GPT-3: A large language model (LLM) developed by OpenAI, demonstrating the massive computational requirements of AI training.** Training GPT-3 involved quadrillions of calculations and took an estimated 34 days using 10,000 NVIDIA V100 GPUs.

H

- **Hard Disk Drive (HDD): A storage device that uses magnetic disks to store data.** HDDs offer large storage capacities at lower costs compared to SSDs but are slower.
- **Heterogeneous-Compute Interface for Portability (HIP): An abstraction layer in AMD's ROCm platform that allows developers to easily convert CUDA applications to run on AMD GPUs.**
- **High Bandwidth Memory (HBM): A specialised type of memory that provides higher bandwidth and lower latency than traditional DRAM.** HBM stacks memory chips vertically and places them closer to the processor, improving data transfer speed and efficiency.
- **High-Performance Computing (HPC): The use of supercomputers and parallel processing techniques to solve complex computational problems.** AI training often requires HPC infrastructure.

I

- **Inference: The stage where a trained AI model processes new data to make predictions or draw conclusions.** Less computationally demanding than training but has different requirements, such as latency, performance, and efficiency.

- **InfiniBand: A high-speed networking technology commonly used in HPC and data centre environments.** It offers high bandwidth and low latency, making it suitable for data-intensive AI workloads. **InfiniBand is known for its high data integrity, which is crucial for AI training.**
- **Instruction Set Architecture (ISA): The fundamental set of instructions that a processor can understand and execute.** Intel and AMD processors are based on the x86 ISA.
- **Interconnect: A communication channel that enables data transfer between different hardware components, such as processing units, memory, and storage.** Types: on-chip, chip-to-chip, node-to-node.

L

- **Latency:** The time delay between a request for data and the data's arrival. **Low latency is critical for real-time AI applications.**
- **Large Language Model (LLM): A type of AI model trained on massive text datasets to understand and generate human-like language.** Examples: OpenAI's ChatGPT.

M

- **Machine Learning (ML): A type of AI that enables computers to learn from data without explicit programming.** ML algorithms identify patterns and make predictions based on data. **The sources primarily focus on AI powered by ML.**

- **Memory: A temporary storage space where a computer stores data that it is actively using.** Different types of memory are used in AI systems, including DRAM, HBM, and GDDR.
- **Micron Technology: A leading manufacturer of memory and storage solutions, including DRAM, HBM, and SSDs.**

N

- **Natural Language Processing (NLP): A subfield of AI focused on enabling computers to understand, interpret, and generate human language.**
- **Networking: The interconnection of computers and other devices to enable communication and data exchange.** Networking technologies like Ethernet and InfiniBand are essential for AI hardware, especially in data centre environments.
- **Neural Processing Unit (NPU): A specialised processing unit integrated into some CPUs, specifically designed for AI workloads.** NPUs aim to improve the performance of CPUs for AI tasks.
- **Node: A single computer or server within a compute cluster.** Nodes are interconnected to enable distributed computing for AI workloads.
- **Non-Volatile Memory Express (NVMe): A communication protocol for SSDs that offers high-speed storage and retrieval, low latency, and**

high-throughput. NVMe SSDs are commonly used in data centres for AI workloads.

- **NVLink: NVIDIA's proprietary chip-to-chip interconnect technology that enables high-bandwidth, low-latency connections between multiple NVIDIA GPUs.**

O

- **On-Chip Interconnect: A type of interconnect that facilitates communication between different components within a System-on-Chip (SoC).** Chiplets and advanced packaging technologies are examples of on-chip interconnects.
- **OpenAI: An AI research and deployment company known for developing large language models, including GPT-3.**

P

- **Packaging Technology: Techniques used to enclose and connect semiconductor chips to other components on a printed circuit board.** Advanced packaging technologies, such as chiplets and 3D packaging, are critical for improving the performance and efficiency of AI hardware.

- **Parallel Processing: The ability to execute multiple computations simultaneously, significantly speeding up complex tasks. GPUs excel at parallel processing, making them suitable for AI workloads.**
- **Peripheral Component Interconnect Express (PCIe): A widely used expansion standard that enables communication between various hardware components, including processing units, storage, and networking cards.**
- **Performance: A measure of how quickly and efficiently a processing unit can handle AI workloads.** Factors considered include processing power, memory bandwidth, latency, scalability, and programmability.
- **Processing-in-Memory (PIM): An emerging memory technology that integrates processing capabilities into memory itself, reducing data movement and improving efficiency.**
- **Processing Unit: The "engine" of an AI system responsible for executing the mathematical and algorithmic operations involved in training and inference.** Types: CPUs, GPUs, ASICs, FPGAs. Also known as an *AI accelerator*.
- **Programmability:** The ease with which a processing unit can be programmed to perform specific tasks. **A developer-friendly programming environment is crucial for AI hardware adoption.**
- **Programming Language: A formal language used to write instructions for computers to execute.** Python is a popular programming language for AI development due to its simplicity and extensive libraries.

- **Programming Platform: A set of software tools, libraries, and APIs that provide a framework for developing and deploying AI applications.** Examples: NVIDIA's CUDA, AMD's ROCm.
- **Proprietary Technology:** Technology that is owned and controlled by a specific company, limiting access and competition. **NVIDIA's CUDA is an example of proprietary technology that has contributed to its market dominance but also raised concerns about vendor lock-in.**
- **PyTorch: An open-source AI framework known for its flexibility and research-oriented features.**

R

- **Radeon Open Compute (ROCm): AMD's open-source software platform for programming AMD GPUs, designed to compete with NVIDIA's CUDA.** ROCm offers an abstraction layer (HIP) for easier porting of CUDA applications and benefits from community contributions.
- **Random-Access Memory (RAM): A type of computer memory that allows data to be accessed randomly, regardless of its physical location on the storage medium.**

S

- **Scalability: The ability of an AI system to handle increasing workloads or larger datasets by adding more resources.** Scalability is crucial for data centres and cloud AI platforms.
- **Samsung: A South Korean multinational conglomerate that is a leading manufacturer of memory chips, SSDs, and advanced packaging technologies.**
- **SK Hynix: A South Korean memory semiconductor manufacturer, specialising in DRAM, NAND flash, and other memory products.**
- **Small Modular Reactor (SMR): A type of nuclear reactor that is smaller and more scalable than traditional reactors.** Microsoft is exploring the use of SMRs to power its data centres.
- **Software Ecosystem: The collection of software components, including frameworks, programming languages, and platforms, that support the development and deployment of AI applications. A robust and developer-friendly software ecosystem is crucial for the success of AI hardware.**
- **Solid State Drive (SSD): A type of storage device that uses flash memory to store data.** SSDs offer significantly faster read and write speeds compared to HDDs. **NVMe SSDs are commonly used in data centres for AI workloads.**
- **System-on-Chip (SoC): An integrated circuit that combines multiple components of a computer system, such as a CPU, GPU, memory, and storage, onto a single chip.** SoCs are commonly used in mobile devices and embedded systems.

T

- **Taiwan Semiconductor Manufacturing Company (TSMC): A Taiwanese multinational semiconductor contract manufacturing and design company.** TSMC is a major manufacturer of CPUs, GPUs, and other advanced processors.
- **Tape Storage: A storage technology that uses magnetic tape to store data.** Tape storage is often used for archival purposes and for storing extremely large datasets that do not require quick retrieval.
- **Tensor Processing Unit (TPU): Google's custom-designed ASIC specifically optimised for AI workloads, particularly those using Google's TensorFlow framework.**
- **TensorFlow: An open-source AI framework developed by Google, known for its scalability and performance.** TensorFlow is heavily optimised for Google's TPUs.
- **Training: The process of creating an AI model by "teaching" an algorithm using data.** During training, the AI model learns to identify patterns and make predictions based on the provided data.

U

- **Ultra Accelerator Link (UALink): A proposed industry standard for an Ethernet-based interconnect designed for high-speed GPU-to-**

GPU communication. UALink aims to create an open alternative to proprietary solutions like NVIDIA's NVLink.

- **Ultra Ethernet Consortium (UEC): An industry initiative backed by major players in the AI hardware and software industry to optimise Ethernet for high-performance computing and AI networking.**
- **Ultra Path Interconnect (UPI): Intel's proprietary chip-to-chip interconnect technology that enables high-speed communication between multiple Intel CPUs.**
- **Universal Chiplet Interconnect Express (UCIe): An open industry standard for interconnecting chiplets from different vendors, promoting interoperability and innovation in AI hardware.**

V

- **Vendor Lock-In: A situation where a customer becomes reliant on a single vendor for products or services, making it difficult and costly to switch to a competitor.** Proprietary technologies can lead to vendor lock-in.

W

- **Wafer-Scale Engine: A massive AI processor developed by Cerebras Systems, known for its large size and processing power by virtue of having been fabricated on an entire silicon wafer.**

This glossary provides a starting point for understanding the key terms and concepts related to AI hardware. Further exploration of the sources and other resources is encouraged for a deeper understanding.

8. References

1. IEEE. "Document 9016391." Accessed September 15, 2024. https://ieeexplore.ieee.org/abstract/document/9016391
2. TUDR. "AI-Driven Approaches." Accessed September 22, 2024. https://tudr.org/3078/1/AI-Driven%20Approaches.pdf
3. ScienceDirect. "Article on AI Hardware." Accessed September 3, 2024. https://www.sciencedirect.com/science/article/pii/S2666603022000136
4. Springer. "AI and Machine Learning Applications." Accessed September 28, 2024. https://link.springer.com/chapter/10.1007/978-3-030-49186-4_27
5. MIT Science Policy Review. "Science Policy Review." Accessed October 12, 2024. https://sciencepolicyreview.org/wp-content/uploads/securepdfs/2022/08/MITSPR-v3-191618003019.pdf
6. Verified Market Research. "Global Artificial Intelligence (AI) Hardware Market." Accessed October 2, 2024. https://www.verifiedmarketresearch.com/product/global-artificial-intelligence-ai-hardware-market/
7. Nasdaq. "NVIDIA's Market Dominance Intact Amid AI Chip Launch Concerns." Accessed September 29, 2024. https://www.nasdaq.com/articles/NVIDIAs-market-dominance-intact-amid-ai-chip-launch-concerns
8. Press Information Bureau. "Press Release on AI Initiatives." Accessed September 18, 2024. https://pib.gov.in/PressReleaseIframePage.aspx?PRID=2012355
9. Hindustan Times. "Centre Earmarks ₹5,000 Crore in AI Plan for Computing Power." Accessed October 10, 2024. https://www.hindustantimes.com/india-news/centre-earmarks-5k-cr-in-ai-plan-for-computing-power-101720118109856.html
10. CERN. "CERN Document." Accessed October 25, 2024. https://cds.cern.ch/record/400313/files/p21.pdf; Turing Post. "Computer Vision History." Accessed September 27, 2024. https://www.turingpost.com/p/cvhistory2; ScienceDirect. "Performance of CPUs and GPUs on Deep Learning." Accessed October 1, 2024. https://www.sciencedirect.com/science/article/pii/S0167926019301762

11. ResearchGate. "Hardware-Enabled Artificial Intelligence." Accessed October 26, 2024. https://www.researchgate.net/publication/328994883_Hardware-Enabled_Artificial_Intelligence
12. China Talk. "AI Compute 101: The Geopolitics of AI." Accessed September 25, 2024. https://www.chinatalk.media/p/ai-compute-101-the-geopolitics-of
13. NVIDIA Blog. "Deep Learning Training vs. Inference." Accessed October 14, 2024. https://blogs.NVIDIA.com/blog/difference-deep-learning-training-inference-ai/
14. ScienceDirect. "Article on AI Hardware." Accessed October 18, 2024. https://www.sciencedirect.com/science/article/pii/S2666675821001041
15. Borealis AI. "A High-Level Overview of Large Language Models." Accessed October 23, 2024. https://www.borealisai.com/research-blogs/a-high-level-overview-of-large-language-models/
16. arXiv. "Attention is All You Need." Accessed October 9, 2024. https://arxiv.org/pdf/2005.14165
17. Turing. "Data Collection and Preprocessing in Python." Accessed September 20, 2024. https://www.turing.com/kb/how-data-collection-and-data-preprocessing-in-python-help-in-machine-learning
18. Understanding AI. "Large Language Models Explained." Accessed October 17, 2024. https://www.understandingai.org/p/large-language-models-explained-with
19. Understanding AI. "Large Language Models Explained." Accessed October 21, 2024. https://www.understandingai.org/p/large-language-models-explained-with
20. Lambda Labs. "Demystifying GPT-3." Accessed October 5, 2024. https://lambdalabs.com/blog/demystifying-gpt-3
21. Microsoft News. "OpenAI and Azure Supercomputer." Accessed October 11, 2024. https://news.microsoft.com/source/features/innovation/openai-azure-supercomputer/
22. Construction Physics. "How to Build an AI Data Center." Accessed October 29, 2024.
https://www.construction-physics.com/p/how-to-build-an-ai-data-center
23. arXiv. "AI Data Center White Paper." Accessed October 15, 2024. https://arxiv.org/pdf/2311.02651
24. Construction Physics. "How to Build an AI Data Centre", Accessed October 15, 2024. https://www.construction-physics.com/p/how-to-build-an-ai-data-center

25. Ibid.
26. The CSIRO GPU cluster at the data centre © CSIRO (https://commons.wikimedia.org/wiki/File:CSIRO_ScienceImage_11313_The_CSIRO_GPU_cluster_at_the_data_centre.jpg). CC BY 3.0 AU.
27. arXiv. "AI Data Center White Paper." Accessed October 26, 2024. https://arxiv.org/pdf/2311.02651
28. Next Platform. "Data Center Networking." Accessed September 27, 2024. https://www.nextplatform.com/2020/08/04/datacenter-is-the-new-unit-of-compute-open-networking-is-how-to-automate-it/
29. Seagate. "Data Center Scalability and Cost Savings." Accessed October 28, 2024. https://www.seagate.com/in/en/blog/data-center-scalability-efficiency-meets-cost-savings/
30. arXiv. "AI Data Center White Paper." Accessed October 29, 2024. https://arxiv.org/pdf/2311.02651
31. NVIDIA Blog. "What is Edge AI?" Accessed October 30, 2024. https://blogs.NVIDIA.com/blog/what-is-edge-ai/
32. Springer. "AI and Machine Learning Applications." Accessed October 1, 2024. https://link.springer.com/article/10.1007/s10462-024-10748-9
33. Today Australia-based AirTrunk opens AI-ready data centre in Johor © Free Malaysia Today. Available at https://www.freemalaysiatoday.com/category/business/2024/07/30/australia-based-airtrunk-opens-ai-ready-data-centre-in-johor/. Licensed under CC BY 4.0.
34. ScienceDirect. "AI and Machine Learning Applications." Accessed September 2, 2024. https://www.sciencedirect.com/science/article/abs/pii/S0045790621004699
35. IEEE. "IEEE Document 10213996." Accessed September 3, 2024. https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=10213996
36. Apple. "Apple Intelligence." Accessed September 4, 2024. https://www.apple.com/in/apple-intelligence/
37. Takshashila. "A Pathway to AI Governance." Accessed September 5, 2024. https://takshashila.org.in/research/a-pathway-to-ai-governance

38. Semantic Scholar. "Performance of CPUs and GPUs on Deep Learning." Accessed September 6, 2024. https://www.semanticscholar.org/paper/Performance-of-CPUs-and-GPUs-on-Deep-Learning-For-N.-Rao/13b89fca8836e4e1855ba864a65b238646c3ec1e
39. Forbes. "At the Heart of the AI PC Battle Lies the NPU." Accessed September 7, 2024. https://www.forbes.com/sites/moorinsights/2024/04/29/at-the-heart-of-the-ai-pc-battle-lies-the-npu/; Dookeran, Jason. "Intel, AMD, and Qualcomm: What Do Their Next-Gen NPUs Have to Offer?" *How-To Geek*, September 12, 2024. https://www.howtogeek.com/intel-amd-and-qualcomm-what-do-their-next-gen-npus-have-to-offer/.
40. PC Viewed. "AMD vs Intel CPU Market Share." Accessed September 8, 2024. https://pcviewed.com/amd-vs-intel-cpu-market-share/
41. Next Platform. "NVIDIA's Grace ARM CPU Holds Its Own Against x86 for HPC." Accessed September 9, 2024. https://www.nextplatform.com/2024/02/06/NVIDIAs-grace-arm-cpu-holds-its-own-against-x86-for-hpc/
42. Forbes. "At the Heart of the AI PC Battle Lies the NPU." Accessed September 10, 2024. https://www.forbes.com/sites/moorinsights/2024/04/29/at-the-heart-of-the-ai-pc-battle-lies-the-npu/
43. Intel, Intel CPU Core i7 6700K Skylake top (2015). https://commons.wikimedia.org/wiki/File:Intel_CPU_Core_i7_6700K_Skylake_top.jpg. Licensed under CC BY-SA4.0.
44. Rao, N. "Performance of CPUs and GPUs on Deep Learning." Semantic Scholar. Accessed October 29, 2024. https://www.semanticscholar.org/paper/Performance-of-CPUs-and-GPUs-on-Deep-Learning-For-N.-Rao/13b89fca8836e4e1855ba864a65b238646c3ec1e; Nvidia Chopper GPUs Expand Reach as Demand for AI Grows." NVIDIA News. Accessed October 29, 2024. https://nvidianews.nvidia.com/news/nvidia-hopper-gpus-expand-reach-as-demand-for-ai-grows
45. Peddie, Jon. 2022. "Introduction." Springer EBooks. https://doi.org/10.1007/978-3-031-10968-3_1. Accessed October 29, 2024.

46. Understanding AI, "Large Language Models Explained with Matrix Operations", accessed September 21, 2024, https://www.understandingai.org/p/large-language-models-explained-with; NVIDIA Blogs, Why GPUs are Great for AI, accessed October 12, 2024, https://blogs.NVIDIA.com/blog/why-gpus-are-great-for-ai/.
47. Peddie, Jon. 2022. “Introduction.” Springer EBooks. https://doi.org/10.1007/978-3-031-10968-3_1. Accessed
October 29, 2024.
48. SGA Analytics, "NVIDIA AI Dominance: How Long Will This Bull Run?", accessed September 27, 2024, https://us.sganalytics.com/blog/nvidia-AI-dominance-how-long-will-this-bull-run/.
49. Intel, "Discrete GPUs, accessed October 7", 2024, https://www.intel.com/content/www/us/en/products/details/discrete-gpus.html.
50. U.S. Chamber of Commerce, "U.S. Supply Chain Vulnerabilities and Resilience" , accessed September 15, 2024, https://www.uscc.gov/sites/default/files/2022-11/Chapter_2_Section_4--U.S._Supply_Chain_Vulnerabilities_and_Resilience.pdf; European Commission, "Chips Act", accessed October 2, 2024, https://digital-strategy.ec.europa.eu/en/factpages/chips-act.; CNN, "China Semiconductor Investment Fund", accessed September 30, 2024, https://edition.cnn.com/2024/05/27/tech/china-semiconductor-investment-fund-intl-hnk/index.html ; IEEE Spectrum, "Indian Semiconductor Manufacturing", accessed October 19, 2024, https://spectrum.ieee.org/indian-semiconductor-manufacturing.
51. Nvidia, Nvidia GeForce RTX 4090 GPU (2023), available at: https://en.wikipedia.org/wiki/File@5nm@AdaLovelace@AD102@GeForce_RTX_4090@S_TW_2324A1_U2F028.MOW_AD102-301-A1_DSCx3@VIS.jpg. Licensed under CC BY 2.0
52. PCMag, "ASIC," accessed September 21, 2024, https://www.pcmag.com/encyclopedia/term/asic.
53. Cerebras Systems, "Cerebras Wafer-Scale Engine," accessed October 12, 2024, https://cerebras.ai/product-chip/.
54. SemiEngineering, "Tensor Processing Unit (TPU)," accessed September 27, 2024,

https://semiengineering.com/knowledge_centers/integrated-circuit/ic-types/processors/tensor-processing-unit-tpu/.
55. Cerebras Systems, "Wafer-Scale Processors: The Time Has Come," accessed October 7, 2024, https://cerebras.ai/blog/wafer-scale-processors-the-time-has-come/.
56. TechCrunch, "Google's Custom Machine Learning Chips," accessed September 15, 2024, https://techcrunch.com/2017/04/05/google-says-its-custom-machine-learning-chips-are-often-15-30x-faster-than-gpus-and-cpus/.
57. Electronic Design, "The Economics of ASICs," accessed October 2, 2024, https://www.electronicdesign.com/technologies/embedded/article/21808278/ensilica-the-economics-of-asics-at-what-point-does-a-custom-soc-become-viable.
58. Coherent Market Insights, "ASIC Chip Market," accessed September 30, 2024, https://www.coherentmarketinsights.com/industry-reports/asic-chip-market.
59. Ibid.
60. The Information, "To Reduce AI Costs, Google Wants to Ditch Broadcom as Its TPU Server Chip Supplier,"accessed October 19, 2024, https://www.theinformation.com/articles/to-reduce-ai-costs-google-wants-to-ditch-broadcom-as-its-tpu-server-chip-supplier
61. Google, TPU v4 (2021), available at: https://commons.wikimedia.org/wiki/File:TPU_v4.png. Licensed under CC BY 2.0.
62. Xilinx, "FPGA for AI," accessed October 12, 2024, https://www.xecor.com/blog/fpga-for-ai
63. Altera, TEI0009 Altera Cyclone 10 LP FPGA RefKit (2017), available at: https://commons.m.wikimedia.org/wiki/File:TEI0009_Altera_Cyclone_10_LP_FPG_RefKit.jpgLicensed under CC BY-SA 4.0.
64. FPGA Insights, "FPGA-Based Prototyping: Accelerating Innovation and Development," accessed September 21, 2024, https://fpgainsights.com/fpga/fpga-based-prototyping-accelerating-innovation-and-development/
65. Xilinx, "FPGA for AI," accessed November 12, 2024, https://www.xecor.com/blog/fpga-for-ai

66. HPCwire, "FPGA Development Brings AMD and Intel Competition to the Forefront," accessed October 12, 2024, https://www.hpcwire.com/2023/06/28/fpga-development-brings-amd-and-intel-competition-to-the-forefront/
67. Embedded.com, "Leveraging FPGAs for Deep Learning," accessed September 27, 2024,
https://www.embedded.com/leveraging-fpgas-for-deep-learning/
68. Understanding AI, "Large Language Models Explained with Matrix Operations," accessed October 7, 2024,
https://www.understandingai.org/p/large-language-models-explained-with
69. Renesas, "Closing the Performance Gap Between DRAM and AI Processors," accessed September 15, 2024,
https://www.renesas.com/us/en/blogs/closing-performance-gap-between-dram-and-ai-processors
70. Red Hat, "CPU Components and Functionality," accessed October 2, 2024,
https://www.redhat.com/sysadmin/cpu-components-functionality
71. Cadence, "Riding the AI Wave Using HBM (High-Bandwidth Memory)," accessed September 30, 2024,
https://community.cadence.com/cadence_blogs_8/b/fv/posts/riding-the-ai-wave-using-hbm-high-bandwidth-memory
72. Mason, "Memory Subsystems 2018," accessed October 19, 2024,
https://mason.gmu.edu/-spudukot/Files/Conferences/Mem_Subsys18.pdf
73. TechTarget, "AI Model Training Drives Up Demand for High-Bandwidth Memory," accessed November 12, 2024,
https://www.techtarget.com/searchstorage/feature/AI-model-training-drives-up-demand-for-high-bandwidth-memory
74. Korea Times, "South Korea to Become a Global Semiconductor Powerhouse," accessed September 21, 2024,
https://www.koreatimes.co.kr/www/world/2024/10/501_356500.html
75. Core, "A Survey of Deep Learning Architectures," accessed October 2, 2024,

https://core.ac.uk/download/pdf/4423431.pdf
76. Forbes, "Why AI Teams Need a Unified Data Format for Machine Learning Datasets," accessed October 12,
2024, https://www.forbes.com/councils/forbestechcouncil/2021/11/11/why-ai-teams-need-a-unified-data-format-for-machine-learning-datasets/
77. MinIO, "The Architects' Guide to Storage for AI," accessed September 27, 2024,
https://blog.min.io/the-architects-guide-to-storage-for-ai/
78. Microchip, "How AI is Transforming NVMe-SSDs," accessed October 7, 2024,
https://www.microchip.com/en-us/about/media-center/blog/2022/how-ai-is-transforming-nvme-ssds
79. Mordor Intelligence, "NAND Flash Memory Market - Companies," accessed September 15, 2024,
https://www.mordorintelligence.com/industry-reports/nand-flash-memory-market/companies
80. Western Digital, "A Balancing Act: HDDs and SSDs in Modern Data Centers," accessed October 2, 2024,
https://blog.westerndigital.com/a-balancing-act-hdds-and-ssds-in-modern-data-centers/
81. Evertiq, accessed September 30, 2024, https://evertiq.com/news/19546
82. TechRadar, "The Advantages of Tape for Backup and Archive," accessed October 19, 2024,
https://www.techradar.com/news/the-advantages-of-tape-for-backup-and-archive
83. Persistence Market Research,
"Tape Storage Market," accessed September 12, 2024,
https://www.persistencemarketresearch.com/market-research/tape-storage-market.asp
84. C.F.A.U.K., "From Zero to Hero - A Data Scientists' Guide to Hardware," accessed October 20, 2024,
https://www.cfauk.org/pi-listing/from-zero-to-hero-a-data-scientists-guide-to-hardware;

85. CSIS, "Chip Shortages Light Geopolitics and Climate Change," accessed September 11, 2024, https://www.csis.org/blogs/strategic-technologies-blog/chip-shortages-light-geopolitics-and-climate-change;
Fortune, "Memory Chip Sector Suffering Rout Despite Vows to Escape Boom-Bust Cycle," accessed September 11, 2024, https://fortune.com/2023/01/29/memory-chip-sector-suffering-rout-despite-vows-to-escape-boom-bust-cycle/
86. Marvell, "Scaling AI Means Scaling Interconnects," accessed September 21, 2024, https://www.marvell.com/blogs/scaling-ai-means-scaling-interconnects.html
87. Intel, "Moore's Law," accessed October 12, 2024, https://www.intel.com/content/www/us/en/newsroom/resources/moores-law.html
88. Imperial College London, "The End of Moore's Law," accessed September 27, 2024, https://www.imperial.ac.uk/news/193270/the-moores-law/ ;
89. AMD, "Zen Core," accessed October 7, 2024, https://www.amd.com/en/technologies/zen-core.html
90. Ibid.
91. EE Times, "The Role of Interconnection in the Evolution of Advanced Packaging Technology," accessed September 15, 2024, https://www.eetimes.com/the-role-of-interconnection-in-the-evolution-of-advanced-packaging-technology/;
Synopsys,"What is Die-to-Die Interface," accessed October 2, 2024, https://www.synopsys.com/glossary/what-is-die-to-die-interface.html
92. Reuters, "TSMC Leads Advanced Chip Packaging Wars, LexisNexis Patent Data Says," accessed September 30, 2024, https://www.reuters.com/technology/tsmc-leads-advanced-chip-packaging-wars-lexisnexis-patent-data-says-2023-08-01/
93. Tom's Hardware, "TSMC and Samsung Foundry Becoming Dominant Makers of Advanced Chips," accessed October 19, 2024, https://www.tomshardware.com/news/tsmc-and-samsung-foundry-becoming-dominant-makers-of-advanced-chips
94. CNBC, "South Korea's Dominance in Memory Chips: An Advantage in the AI Race," accessed November 12, 2024, https://www.cnbc.com/2023/07/06/south-

koreas-dominance-in-memory-chips-an-advantage-in-ai-race.html; Business Korea, "South Korea's Memory Chip Industry Faces Challenges in the AI Era," accessed November 12, 2024,
https://www.businesskorea.co.kr/news/articleView.html?idxno=217353
95. University of Wisconsin-Madison, "NVLink GPU," accessed Oct 15, 2024,
https://pages.cs.wisc.edu/-yxy/cs764-120/papers/nvlink-gpu.pdf
96. Ibid.
97. NVIDIA, "NVLink," accessed September 21, 2024,
https://www.NVIDIA.com/en-in/data-center/nvlink-c2c/;
University of Wisconsin-Madison, "NVLink GPU," accessed October 12, 2024,
https://pages.cs.wisc.edu/-yxy/cs764-120/papers/nvlink-gpu.pdf
98. HPC Tech, "GTC22: Whitepaper Hopper v1.02," accessed September 27, 2024,
https://www.hpctech.co.jp/assets/images/info/catalog/pdf/gtc22-whitepaper-hopper_v1.02.pdf
99. HPC Tech, "GTC22: Whitepaper Hopper v1.02," accessed October 12, 2024,
https://www.hpctech.co.jp/assets/images/info/catalog/pdf/gtc22-whitepaper-hopper_v1.02.pdf
100. IBM, "Cluster Computing," accessed October 7, 2024,
https://www.ibm.com/think/topics/cluster-computing
101. Naddod, "What is InfiniBand," accessed September 15, 2024,
https://www.naddod.com/blog/what-is-infiniband
102. Ibid.
103. The Register, "AI Networks: InfiniBand vs. Ethernet," accessed October 2, 2024,
https://www.theregister.com/2024/01/24/ai_networks_infiniband_vs_ethernet/
104. Ibid.
105. Next Platform, "Greasing the Skids to Move AI from InfiniBand to Ethernet," accessed October 19, 2024,
https://www.nextplatform.com/2024/05/09/greasing-the-skids-to-move-ai-from-infiniband-to-ethernet/
106. NVIDIA, "InfiniBand Technology Overview," accessed September 12, 2024,
https://network.NVIDIA.com/related-docs/whitepapers/WP_InfiniBand_Technology_Overview.pdf

107. Fabricated Knowledge, "NVIDIA Earnings Overview - Networking," accessed September 12, 2024, https://www.fabricatedknowledge.com/p/NVIDIA-earnings-overview-networking; Data Gravity, "NVIDIA's $10B Revenue Networking Business," accessed September 12, 2024, https://www.datagravity.dev/p/NVIDIAs-10b-revenue-networking-business
108. Constellation Research, "NVIDIA's Uncanny Knack for Staying Ahead," accessed September 21, 2024, https://www.constellationr.com/blog-news/insights/NVIDIAs-uncanny-knack-staying-ahead ; Tech Blog, "Will AI Clusters Be Interconnected via InfiniBand or Ethernet? NVIDIA Doesn't Care, But Broadcom Sure Does,"accessed October 12, 2024, https://techblog.comsoc.org/2024/08/28/will-ai-clusters be-interconnected-via-infiniband-or-ethernet-NVIDIA-doesnt-care-but-broadcom-sure-does/
109. NVIDIA News, "NVIDIA Completes Acquisition of Mellanox, Creating Major Force Driving Next-Gen Data Centers," accessed September 27, 2024, https://NVIDIAnews.NVIDIA.com/news/NVIDIA-completes-acquisition-of-mellanox-creating-major-force-driving-next-gen-data-centers
110. The Register, "AI Networks: InfiniBand vs. Ethernet," accessed October 7, 2024, https://www.theregister.com/2024/01/24/ai_networks_infiniband_vs_ethernet/
111. Next Platform, "Greasing the Skids to Move AI from InfiniBand to Ethernet," accessed September 15, 2024, https://www.nextplatform.com/2024/05/09/greasing-the-skids-to-move-ai-from infiniband-to-ethernet/
112. Medium, "A Future of AI Through the Semiconductor Looking Glass," accessed October 2, 2024, https://medium.com/@adi.fu7/a-future-of-ai-through-the-semiconductor-looking-glass-ca24451517ae; Rakuten, "Interplay of Hardware and AI: Growth of AI Capabilities with Advancements in Hardware," accessed September 30, 2024, https://corp.rakuten.co.in/rakathon-2024-blog-interplay-of-hardware-and-ai-growth-of-ai-capabilities-with-advancements-in-hardware/; State of AI Report 2019,"Chapter 3: Why Has AI Come of Age," accessed October 19, 2024, https://www.stateofai2019.com/chapter-3-why-has-ai-come-of-age/
113. Medium, "A Beginner's Guide to 10 AI Frameworks and Libraries: AI for Absolute Beginners," accessed November 12, 2024, https://medium.com/ai-for-absolute-

beginners/a-beginners-guide-to-10-ai-frameworks-and-libraries-ai-for-absolute-beginners-8a21c0196ab2

114. ACM, "Domain-Specific Hardware Accelerators," accessed November 12, 2024, https://cacm.acm.org/research/domain-specific-hardware-accelerators/

115. JuliaLang, " JuliaLang," accessed October 16, 2024, https://julialang.org/; Python,"Python," accessed October 17, 2024, https://www.python.org/

116. arXiv, "NVIDIA 40S," accessed October 17, 2024, https://arxiv.org/html/2405.16956v1

117. NVIDIA, "CUDA Zone," accessed October 17, 2024, https://developer.NVIDIA.com/cuda-zone

118. Fabien Sanglard, "Demystifying GPU-Compute Architectures," accessed October 19, 2024, https://fabiensanglard.net/cuda/index.html

119. The Chip Letter, "Demystifying GPU-Compute Architectures," accessed October 19, 2024, https://thechipletter.substack.com/p/demystifying-gpu-compute-architectures

120. Supermicro, "CUDA," accessed October 19, 2024, https://www.supermicro.com/en/glossary/cuda; Medium, "The CUDA Advantage: How NVIDIA Came to Dominate AI and the Role of GPU Memory in Large-Scale Models," accessed October 19, 2024, https://medium.com/@aidanpak/the-cuda-advantage-how-NVIDIA-came-to-dominate-ai-and-the-role-of-gpu-memory-in-large-scale-model-e0cdb98a14a0

121. Harvard University, "NVIDIA's Winning Platform Strategy with CUDA," accessed October 19, 2024, https://d3.harvard.edu/platform-digit/submission/NVIDIAs-winning-platform-strategy-with-cuda/

122. Ibid.

123. Ibid.

124. Fortune, "What Does NVIDIA Do: Chips AI Jensen Huang," accessed October 20, 2024, https://fortune.com/2024/02/22/what-does-NVIDIA-do-chips-ai-jensen-huang/

125. Visual Capitalist, "NVIDIA Revenue by Product Line," accessed October 20, 2024, https://www.visualcapitalist.com/nvidia-revenue-by-product-line/

126. Shihab Shahriar, "CUDA vs. ROCm: A Case Study Through Random Number Libraries" accessed October 20; 2024, https://shihab-shahriar.github.io//blog/2023/Cuda-vs-Rocm-A-Case-Study-Through-Random-Number-Libraries/
127. ROCm, "Software Tools Optimization," accessed October 20, 2024, https://rocm.blogs.amd.com/software-tools-optimization/hipify/README.html

TAKSHASHILA
INSTITUTION

The Takshashila Institution is an independent centre for research and education in public policy. It is a non-partisan, non-profit organisation that advocates the values of freedom, openness, tolerance, pluralism, and responsible citizenship. It seeks to transform India through better public policies, bridging the governance gap by developing better public servants, civil society leaders, professionals, and informed citizens.

Takshashila creates change by connecting good people, to good ideas and good networks. It produces independent policy research in a number of areas of governance, it grooms civic leaders through its online education programmes and engages in public discourse through its publications and digital media.