nasscom

Deloitte.

Sovereign GPU Cloud for National Competitiveness in the AI Economy: A Backgrounder

September 2023

Sovereign GPU Cloud for National Competitiveness in the AI Economy: A Backgrounder

1. Introduction

Generative AI, a subset in the field of Artificial Intelligence, has grown by leaps and bounds over the past year, capturing public attention and imagination, and creating fervent discussions over its usability, caveats, and the broader impact on society and the global economy. Generative AI is regarded as one of the most significant breakthroughs of the fourth industrial revolution, with the ability to significantly alter our lives and the way we work and play, one way or the other.

Broadly characterizing the concept, Generative AI refers to advanced artificial intelligence models that, through rigorous training on petabytes of data and employing advanced Deep Learning & Neural Network models, can generate content across modalities such as text, image, audio, video, code, 3D renders etc. Additionally, Generative AI has the unique capability to add contextual awareness and emulate human-like decision-making and reasoning, enabling it to become a vital tool in the way we conduct business.

2. Generative AI and GPUs – Exploring the Infrastructure

Generative AI models need GPUs (Graphics Processing Units) to run and are fundamentally different from traditional computations, which modern CPUs are designed to do. Generative AI models, which are advanced neural networks, need parallel processing capabilities from the processor to perform matrix operations.

CPU vs GPU

The primary distinction between GPUs (Graphics Processing Units) and CPUs (Central Processing Units) lies in their transistor allocation. GPUs allocate a higher proportion of transistors to arithmetic logic units (ALUs) and fewer to caches and flow control mechanisms compared to CPUs.

CPUs are best suited for tasks involving intricate logic interpretation and code parsing. On the other hand, GPUs were designed as specialized workhorses for rendering graphics in computer games. Over time, they evolved to accelerate various geometric computations, such as polygon transformations and coordinate system rotations in 3D graphics.

In terms of physical size, GPUs are smaller than CPUs, yet they tend to incorporate more logical cores (ALUs, control units, and memory caches) than CPUs.

GPUs for Deep Learning

GPUs are well-suited for training artificial intelligence models due to their ability to handle multiple computations concurrently. They excel in processing numerous parallel tasks, thanks to their substantial core count. This is especially advantageous for deep learning, which involves processing large volumes of data, leveraging the high memory bandwidth of GPUs.

In contrast, CPUs handle tasks sequentially and possess fewer cores, making them less efficient for complex and data-intensive computations. When employing GPUs for generative AI tasks, several benefits arise [2]:

- **Speed:** GPUs excel at parallel computations, drastically reducing training and inference times for generative AI models
- **Scalability:** High-speed interconnects like NVLink (NVIDIA's proprietary system interconnect) allow efficient scaling of generative AI models across multiple GPUs or nodes
- **Flexibility:** GPUs support diverse precision levels (e.g., Floating Point (FP)32, FP16, Tensor Float (TF)32, FP8), enabling users to balance accuracy and performance in Generative AI models
- **Innovation:** GPUs continuously evolve, introducing new prospects and solutions for Generative AI tasks

3. Data Centre Infrastructure – Is it Optimal for Generative AI?

The demand for data centers (DCs) is set to grow across the world as data generation scales up exponentially. However, these DCs will not be of

much practical application in their current infrastructure state for Generative AI applications.

Traditionally, DCs relied on CPUs to perform general-purpose computations, mostly related to data storage, indexation, and retrieval. However, for any applications of Generative AI, the requirement is for higher and parallel computing power, which can only be provided by GPUs. ChatGPT, for example, reportedly used 10,000 Nvidia GPUs to train the model [35].

Globally, Google, Microsoft, and Amazon, the three major hyperscale cloud providers, are making significant efforts to accommodate the deployment of enterprise-grade generative AI infrastructure within their existing data centers. The expansion of generative AI at scale necessitates substantial data storage and computational power, placing stress on existing infrastructure. To address these issues, these companies are enhancing their cloud infrastructure and optimizing their platforms [4] [5].

Google: Alphabet is expanding its Google Cloud data centres and reallocating workloads to accommodate AI computing. Google Cloud is collaborating with NVIDIA to integrate the L4 Tensor Core GPU and Vertex AI into its G2 virtual machines, to provide developers with cutting-edge technology, enhancing the speed, scalability, and energy efficiency required for a wide range of AI workloads [27]. Google has also developed a supercomputer using over 4,000 fourth-gen Tensor Processing Units (TPUs), which are Google's proprietary hardware specifically designed for accelerating AI workloads. It states that its new supercomputer demonstrates significant advantages over a system based on Nvidia's A100.[28]

Microsoft: Microsoft is concentrating on optimizing its Azure infrastructure to support the requirements of Generative AI. Along with NVIDIA, they have unveiled blueprints for a new hyperscale GPU accelerator to drive AI cloud computing. To provide hyperscale data centres with a fast and flexible path for AI, they are developing a new HGX-1 hyperscale GPU accelerator [33], which provides extreme performance scalability to meet the requirements of fast-growing AI workloads, and its unique design allows it to be easily adopted into existing data centres

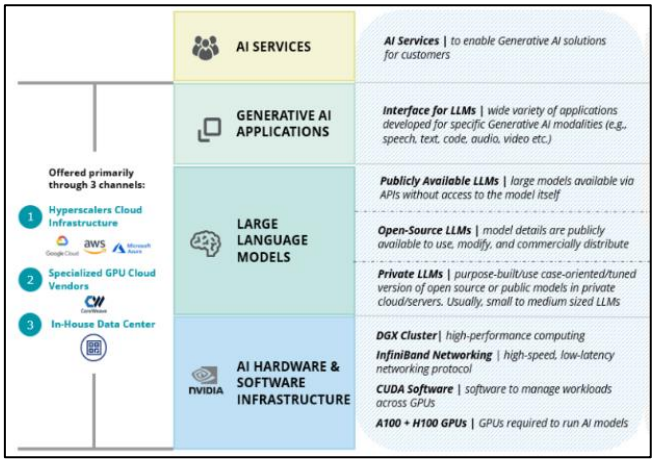
Amazon (AWS): As Nvidia GPUs get expensive and difficult to procure, Amazon is introducing Inferentia and Trainium [52], its microchips, offering more powerful and efficient cloud computing hardware for machine learning inference and model building.

These companies are moving from traditional processors to specialized computing options, incorporating GPUs and proprietary AI accelerators. This shift in infrastructure is driven by the need to balance space utilization, optimize costs, and accommodate increasing demand for generative AI capabilities.

The future of data centres, therefore, is visualized as a mix of CPU and GPU environments. CPU handles the traditional roles of a data centre while the GPU comes in to fulfil the specific AI compute load requests.

4. Generative AI Conceptual Framework

To establish, train, run, and deploy Generative AI, there needs to be continual progress and innovation across the interconnected ecosystems of hardware, software, and data. Typically, a Generative AI framework can be broken down into four distinct layers: **Infrastructure, Platform (Cloud), Model, and Application / Services.** The graphic below illustrates the Conceptual Generative AI Framework using Nvidia as an example for the infrastructure layer:



Source: Deloitte

Infrastructure

At the core of any Generative AI solution are the underlying silicon hardware, software, and

computing resources required to train, run, and deploy sophisticated models. Cloud computing infrastructure is essential to support complex model training and deployment.

The infrastructure layer is comprised of global players such as Nvidia, AMD, and Intel, whose GPUs form the base for the computing infrastructure. Within these players, Nvidia is the dominant player with ~70-80% share in AI GPU space [30] (see "Comparison of Major Chip Companies" in the Appendix).

Within the infrastructure layer, the flow and process can be broken down into the following sub-components:

- **Hardware Acceleration:** Components like GPUs (NVIDIA A100, H100, AMD Radeon VII, Intel Habana Gaudi2) that deliver the required computational power for training and running Generative AI models efficiently
- **Parallel Computing Framework:** Architecture that allows for several processors to simultaneously execute multiple, smaller calculations broken down from an overall larger, complex problem. CUDA (Compute Unified Device Architecture) developed by NVIDIA, DPC++ (Data Parallel C++) developed by Intel for data-parallel programming, and ROCm by AMD are some notable examples
- **Networking and Communication:** Protocols to ensure seamless and rapid communication between GPUs and other components, supporting the collaborative processing required in Generative AI. Nvidia's InfiniBand is a high-speed, low-latency networking protocol designed for data-intensive and high-performance computing environments

Platforms

Data and cloud platforms provide the infrastructure and tools necessary for running AI workloads, including Generative AI models. They offer a range of services that support various aspects of the AI workflow, from data pre-processing and model training to deployment and monitoring. Major cloud platforms offer integration with AI frameworks like TensorFlow and PyTorch, making it easier to develop, train, and deploy generative AI models.

Models

Foundation Models form the heart of Generative AI consisting of highly advanced neural network architectures incorporating deep learning and trained on vast amounts of data to generate outputs across modalities. Some of the most popular model variants include Generative Adversarial Networks (GANs) and Transformers. ChatGPT, OpenAI's language model, runs on GPT (Generative Pretrained Transformer) 3.5/4 which is some of the largest and most powerful in this category. Large Language Models or LLMs are a variant of these Foundation Models which have been specifically trained on massive amounts of text data – including but not limited to books, articles, websites, code etc. There are three discernible archetypes of models currently – publicly available LLMs (e.g., Open AI GPT, Google PaLM, Nvidia NeMo), private LLMs (e.g., Hugging Face, MosaicML) and open source LLMs (e.g., Falcon from UAE).

Applications

At the other end of the Generative AI framework are the user-facing applications that deliver tangible use cases building on top of the models and infrastructure underneath. Applications span across modalities such as text, image, video etc. and have custom interfaces to provide the users a platform to leverage the capabilities of Generative AI (e.g., ChatGPT, Dall E, Microsoft Co-Pilot).

Illustrative GPU Cloud

To better understand and visualize the overall framework and the structural components of Generative AI, an example of a cloud start-up in the market – CoreWeave, which specializes in GPU-accelerated workloads, can be reviewed. CoreWeave offers a modern, Kubernetes-native cloud architecture tailored for large-scale, GPU accelerated workloads. At the core hardware level, distributed training clusters leverage NVIDIA A100 NVLINK GPUs, paired with NVIDIA Infiniband GPUDirect RDMA networking to power deep learning at scale. CPU computing instances use Intel Xeon and AMD EPYC architectures. A fully managed Kubernetes solution ensures performance without infrastructure complexities, supporting rapid instance creation and responsive auto-scaling across thousands of GPUs. Other similar companies operating in this space include Paperspace, RunPod etc. amongst others.

5. Generative AI and Sovereign GPU Cloud

While the Generative AI race heats up globally, most countries are behind the frontrunners – US, China and Israel. There is a need to rapidly capitalize on the emerging technologies and applications of Generative AI. Most countries currently have limited involvement in AI chip hardware design and are missing significant investments in this area, while also lacking audited training and fine-tuning data sets. Furthermore, they need to develop their foundational models akin to GPT-4 or Wu Dao.

Establishing the Need for Sovereign GPU Compute

Every country's priority should be to enhance its AI innovation landscape and establish a globally competitive ecosystem in this domain. To effectively compete with leading nations such as the US, China, and Israel it is crucial for countries to swiftly gain access to the limited supply of GPU chips. This strategic move would empower the countries to bolster their AI capabilities and bridge the existing gap in innovation.

Any significant innovation and acceleration in the field of AI, more so in Generative AI, depends on four aspects – having sufficient budgetary resources, gathering a critical volume of data, establishing the necessary computing power to process it and developing skills to develop and deploy the models and software. Establishing control of and shaping the infrastructure enables the countries to further develop the following paradigms:

1. Build a Globally Competitive AI Ecosystem

Globally, investments and innovations in Generative AI have been concentrated in and around the United States, which is currently the market leader in this space [61]. Other early movers in this space, including China and Israel, have been making significant strides to further increase investment in establishing the necessary compute infrastructure and building actionable and practical use cases in the ecosystem.

2. Gain Access to Computing Infrastructure

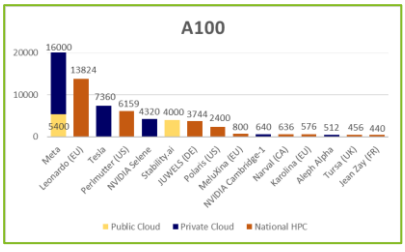
As established previously, Generative AI needs significant GPU computing power as opposed to the traditional CPU infrastructure. Reviewing the GPU market globally in the context of Generative AI, ~70-80% of it is cornered by the US-based Nvidia, with smaller shares of AMD and Intel [30].

There are currently ~40,000 Nvidia A100 chips left in China, according to a recent report by Chinese media company Caixin, citing Yin Qi, CEO of AI unicorn Megvii [32]. Due to the recent export restrictions by the United States on A100/H100 chips, major tech companies in China have placed orders worth US$ 5 Bn for A800 Nvidia chips to continue powering their AI ambitions (the A800 chipset is a weaker version of the A100 with slower data transfer rates than the A100 and is not included in export restrictions) [9] [10] [26].

Nvidia is undergoing a transition towards adopting the H100, their latest advancement in AI processors. Performance evaluations indicate that the H100 demonstrates an order of magnitude higher performance compared to its predecessor, the A100 [37]. An inbuilt Transformer Engine in the H100 is particularly advantageous for the advancement of generative AI models like GPT-3. Additionally, the incorporation of dynamic programming instructions (DPX) contributes to expediting code execution.

Keeping in view the overwhelming global demand for high-performance AI specific GPUs, and the absolute criticality of these chips to train any large language model, countries must build the right partnership with leading hardware providers like Nvidia, AMD, and Intel, to get access to the latest GPU computing.

Global Nvidia A100 and H100 Count, 2022[31]



3. Provide Impetus to the Generative AI Startups

Generative AI startups across the world have lagged compared to peers like the United States and Israel, both in terms of funding and investment, as well as access to capable infrastructure to scale up their offerings.

Generative AI Startup Scenario [36]

	USA	Israel	India
Total Funding in AI, 2013-2022	US$ 249 Bn+	~US$ 11 Bn+	US$ 8 Bn+
# Generative AI Startups, May 2023	400+	70+	60+
Total Funding in Generative AI Startups, to date	US$ 22 Bn+	US$ 1 Bn+	US$ 590 Mn+
Generative AI Unicorns, to date	14	Nil	Nil
Foundation Models, to date	30+	2+	Nil
# Generative AI Startups Funded, Disclosed	65%+	60%+	30%+

Source: nasscom

Global Examples of Sovereign GPU Stack

USA

Investment in GPU compute infrastructure has been driven by private companies including NVIDIA, AMD, Amazon, Microsoft, and Intel. Further, various business models have emerged to provide easier access to GPU computing power

- **VCs are buying GPUs to ensure that their portfolio companies have access to the GPUs even amid a chip shortage**. E.g., Index Ventures signed a deal with Oracle to provide its portfolio companies with access to GPUs [57]
- **Fractional access to GPU computing power.** E.g., Vultr, a cloud computing company, offers fractional access to Nvidia's A100 Tensor Core GPUs where customers can rent a portion of the GPU, rather than purchasing the entire GPU [58]
- **Major companies and startups are working on ways to reduce the use of GPUs and make them more efficient.** E.g., NVIDIA uses cuOpt software, to optimize the performance of ML models; Google Cloud uses using Vertex AI platform for optimizing GPU use [60]

In addition to private investments, the US government has taken several initiatives to develop the high-performance computing infrastructure to support the Generative AI ecosystem. Further, trade restrictions have been imposed on NVIDIA and AMD to limit the export of cutting-edge AI tech to China and Russia, in a bid to restrict their Generative AI ambitions [71].

Israel

US tech giant Nvidia announced plans to construct Israel-1, a cutting-edge Generative AI cloud supercomputer built on a new locally developed high-performance ethernet platform in Israel [14]. Valued at several hundred million dollars, Israel-1 is set to be one of the world's fastest AI supercomputers, expected to commence early production by the close of 2023. This strategic move by Nvidia aligns with its emphasis on AI's transformative potential and its commitment to innovation through local partnerships.

This supercomputer is designed to achieve an extreme performance level of eight exaflops (1 exaflop = 1000 petaflops), denoting the capability to execute a staggering one quintillion calculations per second. Alongside this, Israel-1 is anticipated to reach peak performance levels exceeding 130 petaflops, adept at handling 100 trillion operations per second, specifically geared towards traditional scientific computing tasks.

Further, Israel's Shin Bet security service has created its own Generative AI platform to successfully thwart national threats [19].

Japan

SoftBank is embarking on the development of a domestic Generative AI stack tailored to Japan's language and culture. Unlike existing models based on English and Chinese data, SoftBank's approach aims to capture local nuances. Collaborating with Microsoft, SoftBank will also provide secure data environments for Japanese businesses interested in AI. This plan involves an investment of US$138 million in computing infrastructure, equipped with Nvidia's GPUs [11].

Additionally, Japan is taking significant steps towards enhancing its Generative AI capabilities by co-funding a state-of-the-art supercomputer in Hokkaido, aiming to triple the nation's processing power dedicated to Generative AI [12]. Sakura Internet, a cloud service

provider, is spearheading this initiative to construct the supercomputer in Ishikari City, slated to commence operations next year. The Ministry of Economy, Trade and Industry (METI) is contributing 6.8 billion yen (US$ 48.2 million) of the 13.5-billion-yen (US$ 96 million) budget for the supercomputer which will have over 2,000 Nvidia GPUs. This initiative aims to foster domestic generative AI progress, reducing reliance on foreign solutions like OpenAI's ChatGPT, and fortifying economic security [12] [13].

UAE & Saudi Arabia

UAE and Saudi Arabia are buying thousands of the latest Nvidia GPUs via state-owned enterprises. These Gulf countries are keen on developing an AI industrial base independent of the U.S.-based OpenAI or Google offerings.

"The UAE has made a decision that it wants to own and control its own computational power and talent, have their own platforms, and not be dependent on the Chinese or the Americans,"- a source familiar with Abu Dhabi's policy [22]

UAE has released an open-source LLM called "Falcon 40B" and its scaled-up version "Falcon 180B" available for research and commercial use. Falcon 40B is a model with 40 billion parameters and trained on one trillion tokens developed by the Technology Innovation Institute (TII) [20] [21]. The latest version is a model with 180 billion parameters and was trained on 3.5 trillion tokens using Amazon SageMaker for a total of ~7,000,000 GPU hours.

United Kingdom

UK is planning to spend US$ 126 million to buy AI chips as a part of the country's plans to improve its AI resources. They are already in talks with major AI chip makers such as NVIDIA, AMD and Intel [68].

Reviewing the above examples, it can be inferred that different types of engagement models exist for scaling the sovereign GPU infrastructure:

- Government-led investment: UAE is scaling up through state-owned entities in collaboration with GPU OEMs
- Private sector-led investment: Investment in developing the GPU compute infra in the United States is largely driven by the private sector
- Hybrid or PPP Model: Japan and Israel are developing and scaling their sovereign GPU clouds in partnership with the industry

The strategy is to be adapted and contextualized to each country's needs and aspirations of Generative AI.

6. Key Considerations for India

Given the fast-paced developments and entrenched geopolitics, winning in the AI race requires India to place the right strategic bets in terms of acquiring the necessary computing infrastructure, identifying the most optimal modes of engagement with stakeholders, and further integrating the infrastructure into the India stack. To begin with, it is imperative to understand where India is today on a global landscape with regards to computing infrastructure, cloud services, and other relevant aspects.

MeghRaj – Cloud Computing Initiative of GoI

"MeghRaj", the cloud platform introduced by the government, functions as an integrated e-marketplace for cloud services that adhere to standardized protocols. Its central objective is to expedite the delivery of digital services while optimizing governmental ICT expenditures.

The architecture of Government Interoperability Cloud (MeghRaj) is structured as a federated model, comprising multiple interconnected autonomous clouds in three layers: National Cloud hosted by the National Informatics Centre (NIC), providing Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS) offerings; State Clouds managed by state governments, catering to departmental needs within respective states; and Departmental Clouds, maintained by individual departments to serve their employees and users [72].

While the initiative has a broad set of applications and is beneficial for collaboration, it is primarily tailored for conventional cloud services and might not offer the highly optimized environment optimally suited for Generative AI.

Computing Infrastructure in India

India's existing high-end GPU compute infrastructure, e.g., AIRAWAT and PARAM Siddhi-AI developed by CDAC, currently consists of a total of 656 GPUs (compared to 10,000 GPUs

used by ChatGPT). AIRAWAT PoC of 200 AI Petaflops integrated with PARAM Siddhi – AI of 210 AI Petaflops gives a total peak compute of 410 AI Petaflops Mixed Precision and sustained compute capacity of 8.5 Petaflops (Rmax) Double Precision and a peak compute capacity (Double Precision, Rpeak) of 13 Petaflops, which is insignificant compared to global supercomputers [55]. For example, Frontier in the US has an Rpeak of ~1680 Petaflops, and Fugaku in Japan has an Rpeak of ~537 Petaflops [56].

AI infrastructure is as critical as any other physical infrastructure element and utilities. India needs to set up GPU infrastructure with at least Exaflop AI capacity and 25,000 high performance H100 GPUs [69] or above in collaboration with a reputed cloud provider and system integrator. Hence, a focal shift towards continuous build-out of AI infrastructure and incentives is paramount to ensure India's long-term national competitiveness and achieve leadership in AI. Therefore, the Indian Government and Industry should collaborate to prioritize the establishment of an India GPU cloud, with a comprehensive strategy catering to the Indian AI context.

Acquire Necessary Computing Infrastructure

Considering the lack of computing infrastructure as established in the above section, there is a need to invest in acquiring the required GPU computing capacity within the country, especially as other countries are significantly scaling up their AI ambitions and racing to establish dominance in the field. AI-focused compute infrastructure can be established as:

I. **Greenfield initiative**: Would require collaboration with GPU hardware OEMs and cloud service providers as well as investment in data security and skilling

II. **Leveraging existing high-performance computing infrastructure with modifications to suit the AI workloads:** This would require evaluation and upgradation of existing infra, research collaboration and investment in skillset development

Collaborate with Industry

To acquire the necessary computing infrastructure on priority for the development of a Sovereign Generative AI stack, the government should collaborate with the industry, while aligning with the national data access and data security goals.

As established from the global examples above, three types of engagement models exist: Government-led, private sector-led and hybrid or PPP models. However, there is no single "best" answer for India or any other sovereign state.

Take into account the Total Cost of Ownership (TCO)

One of the key considerations as India sets up its computing infrastructure is to keep the Total Cost of Ownership (TCO) in mind. TCO can be broken down into capital expenditure in the form of acquiring required GPU infrastructure and operational expenses accrued in training and inference of the Generative AI models developed on top of the base infrastructure.

Considering the significant investments required for building out compute infrastructure, government needs to strategically plan long-term incentive programs that accelerate industry investment. Due consideration also needs to be given to the high rate of technological obsolescence and thereby continuous investments are required given the dynamic Gen AI ecosystem which requires constant innovation and adaptation.

Integrate into India Stack

To further bolster the practicality and inclusiveness of the GPU compute cloud infrastructure, India can explore the idea of integrating the new infrastructure as part of the Indian Digital Public Infrastructure (DPI), colloquially known as India Stack [25].

By integrating Generative AI into this ecosystem, India can further enhance its technological capabilities, ensuring inclusive access and fostering innovation across sectors. This addition acts as a natural progression, complementing the foundational DPIs already in place - digital identity (Aadhar), real-time fast payment (UPI) and a platform to safely share personal data without compromising privacy (Account Aggregator built on the Data Empowerment Protection Architecture)

The Generative AI layer brings forth the power of artificial intelligence that can understand, interpret, and create content, thereby amplifying the efficiency and effectiveness of public service delivery and business solutions. This technology democratizes access to advanced AI capabilities, levelling the playing field for individuals and businesses regardless

of their size or resources, thus ensuring inclusivity. Training complex AI models, generation of localized content, empowering start-ups to build next-gen AI apps and access to specialized tools for custom AI models can be potential tangible benefits of integration of GPU cloud with India stack.

To summarize, AI infrastructure is a key component of India's national economic growth strategy. Hence, urgent access to computing resources at the lowest total cost will be a critical factor in attaining leadership in AI.

Based on the above key considerations, India needs to prioritize the establishment of a Sovereign GPU Cloud in close collaboration with the industry. While the initial investments in GPU infra can be led by the government, however, in the long term, India can explore private sector participation to scale out the GPU infrastructure. Given the large scale of CAPEX and OPEX investments involved, these projects need to be supported with strategically planned long-term incentive programs to accelerate industry investment in the core building blocks of the AI stack: data, computing and talent – to spur growth across the broader AI ecosystem, including start-ups and academia.

Next Steps for India

To conclude, the arguments and rationale outlined in this paper illustrate that AI is a key enabler of national competitiveness. To make an impact and step into the ever-competitive Generative AI ecosystem, India has the following strategic points to action upon in a short time frame:

- Understand and establish the need for a sovereign GPU compute infrastructure to enable Generative AI locally
- Review engagement models for developing computing infra (govt-led, private-sector funded, hybrid or PPP model) and select the most viable option considering TCO, data access and data security needs
- Initiate the tasks necessary to acquire the required infrastructure in collaboration with the industry based on the selected engagement model, and significantly accelerate the setup process supported with potential incentives. Initiate small-scale pilot projects showcasing the potential of Generative AI by collaborating with infrastructure providers to create pilot setups, test tailored models, and assess impact, paving the way for wider infrastructure implementation
- Work with developers, both individual and corporate entities, to begin leveraging the Generative AI infrastructure and build tailored models suited to the local ecosystem
- Identify and develop use cases of the indigenous foundation models across the spectrum, from government ministries to public and private industries to the social sector, with the ultimate objective of providing value to the citizens, enhancing national security and providing competitive advantage to local businesses.

7. Appendix

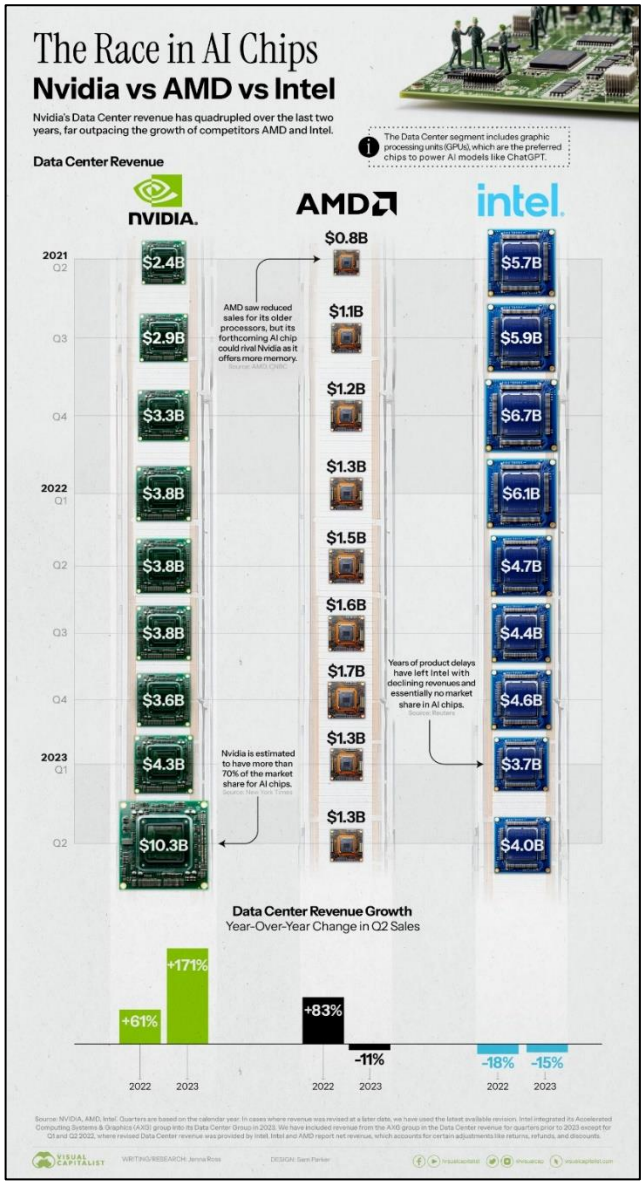
Chipset Selection

Among the high-performance GPU chipsets available in the market, the most adopted for Generative AI applications is Nvidia's Tesla A100, given their higher capability over comparable chipsets from AMD and Intel.

Chipset		Parameter	Detail	Use cases
Nvidia	A100 [42]	Bandwidth	2 TBps	• ~25K in GPT-4 [38] • Alphafold2 AI [39] • Google Bard [40] • Microsoft Bing's AI chat [41]
		HBM	80 GB	
		Interconnects	600 GBps	
		TFLOPS	624	
	V100 [43]	Bandwidth	0.9 TBps	
		HBM	32 GB	
		Interconnects	300 GBps	
		TFLOPS	125	
	Quadro RTX 8000 [44]	Bandwidth	0.67 TBps	
		HBM	48 GB	
		Interconnects	100 GBps	
		TFLOPS	130	
	K80 [45]	Bandwidth	0.48 TBps	
		HBM	24 GB	
		Interconnects	-	
		TFLOPS	5.6	
	P100 [46]	Bandwidth	0.73 TBps	
		HBM	16 GB	
		Interconnects	160 GBps	
		TFLOPS	5.6	
	H100 [48]	Bandwidth	2 TBps	• OpenAI GPT-3 LLM train [47]
		HBM	80 GB	
		Interconnects	600 GBps	
		TFLOPS	3026	
AMD	Radeon VII [49]	Bandwidth	1 TBps	
		HBM	16 GB	
		Interconnects	-	
		TFLOPS	3.46	
Intel	Habana Gaudi2 [50] [51]	Bandwidth	2.45 TBps	• OpenAI GPT-3 LLM train [47]

Source: Company Websites, Published News Articles

Major Companies in AI Chips



Source: Visual Capitalist [30]

Global Examples of GPU-based Clouds

The surge of Generative AI and its applications is increasing demand for high-performance computing from cloud Graphics Processing Units (GPU) providers, which are commonly used for running intensive AI applications. For the leaders of the race to harness the potential of AI, access to GPUs is key. Beyond the traditional cloud service providers (CSPs) like Amazon AWS, Google Cloud Platform, Microsoft Azure, NVIDIA has partnered with or invested in vendors like CoreWeave, Sustainable Metal Cloud, EscherCloud and more to address the demand of GPU-based cloud platforms.

What are GPU-based cloud platforms?

At the core of any Generative AI solution are the underlying silicon hardware, software, and computing resources required to train, run, and deploy sophisticated models. GPU Based Cloud computing infrastructure is essential to support complex model training and deployment. GPUs are well-suited for training AI models due to their ability to handle multiple computations concurrently. They excel in processing numerous parallel tasks, thanks to their substantial core count. This is especially advantageous for deep learning, which involves processing large volumes of data, leveraging the high memory bandwidth of GPUs. Below are three examples of cloud providers who offer NVIDIA powered GPU based cloud platforms across the world.

CoreWeave Cloud is a specialized cloud provider headquartered in New York, that offers high-performance computing infrastructure. They rely on a wide range of high-end NVIDIA GPUs, coupled with native cloud platforms and advanced networking, promising better performance compared to traditional cloud providers. This approach, combined with their infrastructure design, support system, and computing fleet, also aims to optimize cost efficiency for users. They have partnerships with AI startups and cloud providers (which they also compete with) to build clusters to power AI workloads. The company has unique access to the most advanced NVIDIA chips that are in short supply, giving it an edge in competition with traditional cloud providers like Microsoft, Amazon and Google, which are facing supply restraints while working on developing their chips [53] [62].

Sustainable Metal Cloud (SMC) is a large-scale, high-performance, cost-effective, and sustainable AI metal cloud service, run on NVIDIA GPUs, hosted securely in Singapore within a facility operated and staffed by ST Telemedia Global Data Centers (STT GDC) and Firmus. SMC offers a managed compute service with a commercial Service Level Agreement and uses efficient green hosting technology that cuts specific compute CO_2 by up to 50% in Singapore.

SMC is scaling rapidly through its data center platform in Asia, India, and Europe. Its security extends from physical infrastructure to compliance with strict policy and Information and Communication Technology (ICT) frameworks. It also complies with a range of country-specific criteria as well. Firmus, STT & NVIDIA (the three committed partners to deliver SMC) have a platform-agnostic architecture that does not limit its users to an ecosystem lock-in. Models trained on AI Enterprise can be deployed on-premises or in any cloud environment. SMC is available at large scale in Singapore, India, and Australia. Capacity is specifically available in Bengaluru and Pune in India [63].

EscherCloud is an engineering and technology innovation enterprise headquartered in the Netherlands, powered by NVIDIA, that aims to support the vision of a sustainable European virtual economy. They are combining next-generation GPU technology and world-class energy renewal facilities to become the backbone for sustainable technical infrastructure. Hyperion Lab, an initiative space aiming to bring companies, talent, and startups together to accelerate AI and High-Performance Compute within Europe's GDPR compliance framework, is powered by EscherCloud and NVIDIA [64] [65].

In addition to the three examples above, several other independent GPU-based Clouds are starting to mushroom around the world such as Lambda, Applied Digital, Northern Data, Nexgen, OVH, Nebius, Sakura Internet, etc. [66]

8. References

1. https://venturebeat.com/ai/how-nvidia-dominated-ai-and-plans-to-keep-it-that-way-as-generative-ai-explodes/
2. https://www.icdrex.com/the-power-of-generative-ai-and-gpus/
3. https://towardsdatasciece.com/what-is-a-gpu-and-do-you-need-one-in-deep-learning-718b9597aa0d
4. https://www.datacenterdynamics.com/en/analysis/generative-ai-the-future-of-data-centers-part-v-the-chips/
5. https://www.ciodive.com/news/AWS-Microsoft-Google-cloud-infrastructure-AI-ML-compute/649712/
6. https://www.leewayhertz.com/generative-ai-tech-stack/
7. https://techcrunch.com/2023/05/03/where-is-india-in-the-generative-ai-race/
8. https://indianexpress.com/article/opinion/columns/artificial-intelligence-global-generative-ai-market-large-language-models-generative-ai-8901395/
9. https://www.ft.com/content/9dfee156-4870-4ca4-b67d-bb5a285d855c
10. https://nathanbenaich.substack.com/p/your-guide-to-ai-july-2023
11. https://techcrunch.com/2023/08/04/softbank-launches-an-openai-for-japan-sb-intuitions-building-llms-and-generative-ai-in-japanese/
12. https://asia.nikkei.com/Business/Technology/Japan-to-pay-for-half-of-100m-generative-AI-supercomputer
13. https://www.wsj.com/articles/tech-giants-are-chipping-in-fast-on-ai-8899208?mod=article_inline
14. https://www.timesofisrael.com/nvidia-taps-into-israeli-innovation-to-build-generative-ai-cloud-supercomputer/
15. https://a16z.com/2023/04/27/navigating-the-high-cost-of-ai-compute/
16. https://www.wsj.com/articles/the-next-challengers-joining-nvidia-in-the-ai-chip-revolution-e0055485
17. https://www.wsj.com/articles/amds-superchips-face-a-trillion-dollar-hurdle-904fa5ad?mod=article_inline
18. https://www.wsj.com/articles/intel-acquires-habana-labs-for-about-2-billion-11576508408?mod=article_inline
19. https://www.reuters.com/technology/israels-shin-bet-spy-service-uses-generative-ai-thwart-threats-2023-06-27/
20. https://www.reuters.com/technology/abu-dhabi-makes-its-falcon-40b-ai-model-open-source-2023-05-25/
21. https://gulfbusiness.com/tii-makes-its-falcon-40b-ai-model-open-source/
22. https://www.ft.com/content/c93d2a76-16f3-4585-af61-86667c5090ba
23. https://coingeek.com/saudi-arabia-uae-compete-to-buy-nvidia-chip-as-global-ai-race-heats-up/
24. https://www.goldmansachs.com/intelligence/pages/ai-investment-forecast-to-approach-200-billion-globally-by-2025.html
25. https://www.orfonline.org/research/digital-public-infrastructure-lessons-from-india/
26. https://www.theverge.com/2022/11/8/23447886/nvidia-a800-china-chip-ai-research-slowed-down-restrictions
27. http://nvidianews.nvidia.com/news/nvidia-and-google-cloud-deliver-powerful-new-generative-ai-platform-built-on-the-new-l4-gpu-and-vertex-ai
28. https://cacm.acm.org/news/271791-google-says-its-ai-supercomputer-with-tpu-v4-chips-outperforms-nvidias-a100-in-speed/fulltext
29. https://www.eejournal.com/article/intel-oneapi-and-dpc-one-programming-language-to-rule-them-all-cpus-gpus-fpgas-etc/
30. https://www.visualcapitalist.com/nvidia-vs-amd-vs-intel-comparing-ai-chip-sales/
31. https://docs.google.com/presentation/d/1WrkeJ9-CjuotTXoa4ZZlB3UPBXpxe4B3FMs9R9tn34I/edit
32. https://www.scmp.com/tech/article/3224061/chinas-big-tech-firms-scramble-advanced-chips-amid-us-sanctions-and-chatgpt-craze
33. https://www.nvidia.com/en-sg/data-center/hgx-1/
34. https://www.thehindubusinessline.com/data-stories/visually/stellar-growth-in-indias-data-centre-capacity/article66116916.ece
35. https://venturebeat.com/ai/nvidia-will-bring-ai-to-every-industry-says-ceo-jensen-huang-in-gtc-keynote-we-are-at-the-iphone-moment-of-ai/
36. https://nasscom.in/knowledge-center/publications/generative-ai-startup-landscape-india-2023-perspective
37. https://www.forbes.com/sites/moorinsights/2022/09/14/nvidias-new-h100-gpu-smashes-artificial-intelligence-benchmarking-records/?sh=50087e33e728
38. https://mpost.io/gpt-4s-leaked-details-shed-light-on-its-massive-scale-and-impressive-architecture/
39. https://towardsdatascience.com/how-to-deploy-and-interpret-alphafold2-with-minimal-compute-9bf75942c6d7

40. https://www.cnbc.com/2023/04/05/google-reveals-its-newest-ai-supercomputer-claims-it-beats-nvidia-.html#:~:text=AI%20models%20and%20products%20such,clock%20for%20weeks%20or%20months.
41. https://www.theverge.com/23649329/nvidia-dgx-cloud-microsoft-google-oracle-chatgpt-web-browser
42. https://www.nvidia.com/content/dam/en-zz/Solutions/Data-Center/a100/pdf/nvidia-a100-datasheet-us-nvidia-1758950-r4-web.pdf
43. https://images.nvidia.com/content/technologies/volta/pdf/tesla-volta-v100-datasheet-letter-fnl-web.pdf
44. https://www.nvidia.com/content/dam/en-zz/Solutions/design-visualization/quadro-product-literature/quadro-rtx-8000-us-nvidia-946977-r1-web.pdf
45. https://www.nvidia.com/content/dam/en-zz/Solutions/Data-Center/tesla-product-literature/TeslaK80-datasheet.pdf
46. https://images.nvidia.com/content/tesla/pdf/nvidia-tesla-p100-datasheet.pdf
47. https://spectrum.ieee.org/large-language-models-training-benchmark
48. https://resources.nvidia.com/en-us-tensor-core/nvidia-tensor-core-gpu-datasheet
49. https://www.amd.com/en/support/graphics/amd-radeon-2nd-generation-vega/amd-radeon-2nd-generation-vega/amd-radeon-vii
50. https://habana.ai/products/gaudi2/
51. https://www.intel.com/content/www/us/en/developer/articles/technical/habana-gaudi2-processor-for-deep-learning.html
52. https://press.aboutamazon.com/2022/11/aws-announces-three-amazon-ec2-instances-powered-by-new-aws-designed-chips
53. https://www.coreweave.com/products/gpu-compute
54. https://techcrunch.com/2023/04/20/coreweave-a-gpu-focused-cloud-compute-provider-lands-221m-investment/
55. https://www.cdac.in/index.aspx?id=hpc_nsf_siddhi-spec
56. https://techmonitor.ai/hardware/fugaku-quantum-japan-fujitsu
57. https://ai2.news/how-index-ventures-jumped-to-the-front-of-the-ai-gpu-line/
58. https://www.businesswire.com/news/home/20230209005178/en/Independent-Cloud-Computing-Leader-Vultr-Adds-NVIDIA-A16-to-its-A40-A100-and-Fractional-GPU-Offerings
59. https://blogs.nvidia.com/blog/2023/03/21/cuopt-world-record-route/
60. https://dataintegration.info/faster-distributed-gpu-training-with-reduction-server-on-vertex-ai
61. https://dealroom.co/guides/generative-ai
62. https://www.reuters.com/technology/coreweave-raises-23-billion-debt-collateralized-by-nvidia-chips-2023-08-03/
63. Sustainable Metal Cloud https://smc.co/
64. Eshercloud - https://www.ai.nl/companies/eschercloud/
65. EscherCloud https://www.eschercloud.com/
66. https://geekflare.com/best-cloud-gpu-platforms/
67. nvidia.com/en-us/data-center/gpu-cloud-computing/
68. https://analyticsindiamag.com/could-the-gpu-crisis-put-indias-innovation-at-stake/
69. nasscom approach paper on India GPU cloud
70. https://www.cnbc.com/2023/03/13/chatgpt-and-generative-ai-are-booming-but-at-a-very-expensive-price.html
71. https://www.al-monitor.com/originals/2023/08/us-bans-nvidia-amd-ai-chips-export-some-mideast-countries-amid-china-fears#:~:text=The%20United%20States%20has%20broadened,access%20to%20critical%20AI%20resources.
72. https://cloud.gov.in/about.php

Contact us:

nasscom

Debjani Ghosh
Kalyan Mangalapalli

Deloitte

Nitin Mittal (nmittal@deloitte.com)
Saurabh Kumar (sakumar@deloitte.com)
S Anjani Kumar (anjanikumar@deloitte.com)

Contributors:

Deloitte

Ayush Rungta
Pavan Kumar Reddy
Alisha Gupta

nasscom

Deloitte