Grant Thornton

#GTBharat
SHAPING VIBRANT INDIA

Rise of AI infrastructure

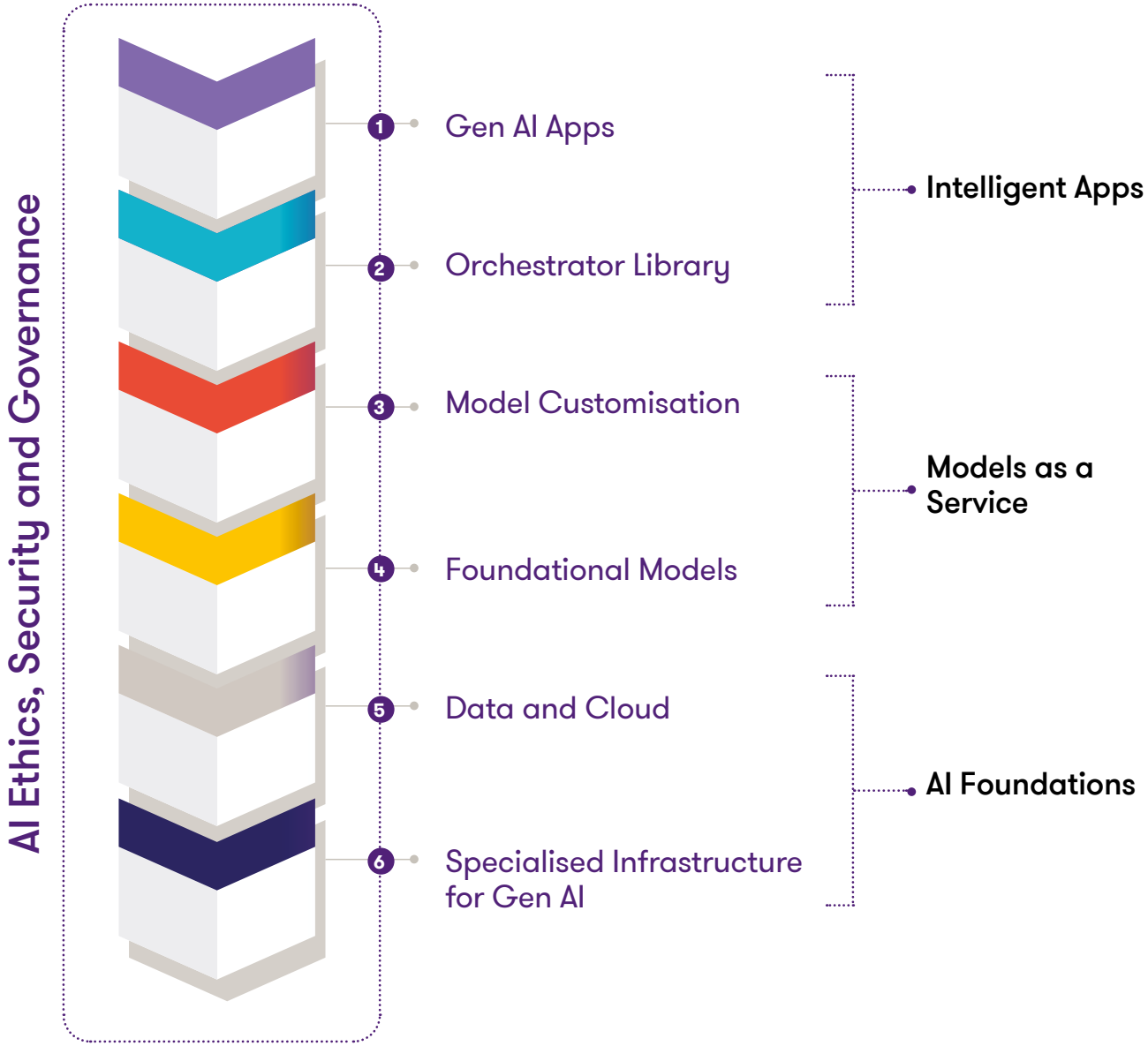
The rapid rise in popularity of ChatGPT and other AI chatbots such as Stable Diffusion, Gemini, Microsoft Copilot, and Midjourney has caught everyone's attention. Generative AI and Foundational Models power these chatbots and AI agents. With such progress, we currently stand at the inflexion point for AI.

Generative AI and traditional Analytical AI can reinvent all industries and business functions. Analytical AI has been applied to classification, regression, clustering, forecasting and predictive use cases. Over the next decade, Gen AI could increase global GDP by **3% to 7%**, potentially adding **USD 2.5 to 8.2 trillion** to the world economy. While Gen AI has use cases and applications across industry and business functions, we have already seen significant adoption in customer operations, IT, software engineering, new products, R&D, HR, sales, marketing and creative functions.

AI Architecture Stack and AI Infrastructure

Recently, the AI Architecture Stack has emerged as increases in Gen AI adoption in enterprises of all sizes. A well-designed and well-thought-out AI Architecture Stack helps scale Gen AI and AI/ML workloads from POC to production and would also help AI adoption across hundreds of use cases across business functions, improving AI security, safety, ethics and governance.

The AI Infrastructure comprises **AI Foundations, Models as a Service** and **AI Ethics, Security and Governance** layers.



The key elements of AI Infrastructure include

AI foundations layer

- **Specialised infrastructure:** Building and training AI and Gen AI Models require specialised hardware and infrastructure including:
  - Specialised chips and GPUs: Graphics processing unit (GPUs), tensor processing unit (TPUs), FPGA, Custom Silica.
  - Fast and efficient networking: InfiniBand, hollow core fibre (HCF), RoCE / RDMA over Converged Ethernet, Remote Direct Memory Access (RDMA) and other fast and efficient networking hardware and protocols.
  - Modern storage solutions: AI optimised file system, cloud-based faster storage like Lustre file system
- **Cloud:** The specialised infrastructure for AI can be easily accessed through a cloud service provider. Besides infrastructure, cloud service providers provide services, including Machine Learning as a Service, Foundational Model/ LLM as a Service, Data services, AI security, governance, and RAI.
- **Data:** The success of your AI programmes depends on your data strategy. You should build a strong data foundation, consider data as a product, and utilise modern data architecture patterns. Think about data lineage, data quality and data governance.

Model as a Service Layer

- **Foundational models:** Foundational models power Gen AI and can be categorised by modality, number of parameters or source. Single-modality models can generate only a single output type, e.g., either text or image while multi-modality models can generate multiple output content such as code, text, image, video. By the number of parameters, they can be classified as large language models (LLM) and small language models (SLM). Models also could be open source or closed source.
- **Model customisation:** Models need to be customised so that they can understand your organisation, client and sector. This could be done in several ways, including prompt engineering, retrieval augmented generation (RAG), fine tuning and pre training.

AI ethics, security and governance layer

- **AI ethics, security and governance:** Building fair and inclusive AI systems that do not have bias, are safe and secure, follow rules, compliance and laws, do not cause IP theft, and provide explainability and traceability is the biggest concern and challenge. Everyone involved in AI should ensure that AI is used for the benefit of humanity and protects human values. AI ethics, security, responsible AI, and explainable AI are emerging and growing fast. AI Infrastructure provides features and services like content safety protection, model monitoring and guardrails, Explainable AI tools, model debugging and tracing, and other safeguards.

Benefits of AI infrastructure

AI infrastructure plays a key role in the recent increase in Gen AI and Analytical AI usage. Its advantages are:

Enables Gen AI and AI/ML use cases: The AI Infrastructure makes the Gen AI use cases such as model training, fine-tuning, customisation, inferencing and AI/ML model building possible.

Superior performance: At the core of the AI Infrastructure is GPUs, TPUs and custom chips made specially for AI workloads. GPUs comprise large numbers of cores (thousands), allowing the parallel processing of calculations on a significant scale. For example, the Nvidia H100 Tensor Core GPU (Grace Hopper series) boasts 16,896 CUDA Cores. The previous generation Nvidia A100 GPUs supported 6,912 CUDA Cores. This provides the raw performance (few peta FLOPS) to the upper level of services and cloud VMs you utilise (running on GPU). This performance is required for training and fine-tuning your LLMs.

Ease of use: Gen AI services (AWS Bedrock and SageMaker JumpStart, Azure AI Studio, GCP Vertex AI Studio, IBM watsonx.ai) helps you get started building Gen AI applications, chatbots and AI agents quickly without knowing the details. These services offer an LLM catalogue, allowing you to customise them using prompt engineering, RAG and fine-tuning. They also help in integration with Data & APIs, workflows, agents, Responsible AI and AI Safety & security. Similarly, Machine Learning as a Service or MLaaS (AWS SageMaker, Azure Machine Learning, GCP Vertex A) allows you to start building machine learning models quickly.

Model lifecycle management: Together, the AI Infrastructure services help in end-to-end model lifecycle management, starting from exploring the models, understanding details using a model card, testing the models using an easy-to-use UI, model evaluation and comparison to customising and fine-tuning the model. These services also help operationalise monitoring and observability, performance and cost analysis.

Scalability: Cloud-based AI Infrastructure provides dynamic and auto-scaling of computing, storage, and other AI/ML services. As your AI needs increase or peak usage of your AI-powered Intelligent Apps increase, you can scale out your AI services.

Cost optimisation: We all know that model training and inferencing costs can rise, but trying to build your own AI Infrastructure on cloud or your own data centre could increase your cost manifold. All AI Infrastructure services are available at a pay-as-you-go model, which helps start AI explorations and experiments at a very low (sometimes completely free) cost. Cloud cost management and optimisation (FinOps) techniques and governance should be used to monitor and optimise AI Infrastructure cost regularly.

AI Infrastructure vs Traditional IT Infrastructure

Area	Traditional IT Infrastructure	AI Infrastructure
Compute	Traditional IT infrastructure uses central processing units (CPUs) for computing. Intel, AMD or cloud service providers usually provide CPUs.	AI Infrastructure primarily uses GPUs, and TPUs for computing. These chips can perform many floating-point operations needed in Gen AI and AI workloads. AI infrastructure also uses FPGA, ASIC, Custom Silicon. In recent times Nvidia has emerged as #1 provider of GPUs. Some of the recent generation Nvidia GPUs (Grace Hopper Series, Blackwell Series) provide PetaFLOPs or 1015 floating point operations per second.
Networking	Traditionally cloud and on-prem IT infrastructure utilises Ethernet and Fiber channel technology for connectivity.	AI infrastructure uses fast & efficient networking, including InfiniBand, hollow core fibre (HCF), Ultra Ethernet, RDMA over Converged Ethernet (RoCE), etc., providing faster networking connectivity than traditional Ethernet technology.
Storage	It consists of cloud storage services including block, object, file, and archival storage.	AI workloads require Specialised Storage Solutions like AI Optimised File System and faster storage. These include Lustre file system (ideal for HPC and AI workloads), Amazon's FSx (scalable and high performant file system) etc.
Datacentre	It uses cloud, on-prem or Colo data centre technologies. While cloud service providers update their data centre technology periodically some of the latest innovations are oriented towards AI infrastructure.	Cloud services and data centre providers have recently optimised their data centres for performance, cost and sustainability. AI Infrastructure utilises some of these innovations, including sustainable, energy-efficient power, advanced liquid cooling mechanisms, next-generation hardware racks, advanced switching and networking systems, etc.

Services	• Historically Hyperscalers (AWS, Azure, GCP) and other cloud service providers (Oracle, IBM, Alibaba) have offered three type of cloud services. • **Infrastructure as a Service (IaaS)** – Infrastructure-related services including Compute (Virtual Machines, Bare metal servers), Networking, Storage (Block storage, Disk storage, File Storage, Archival storage), Load Balancer, Firewall, DDOS protector, Identity and Access Mgmt. (IAM) etc • **PaaS (Platform as a Service)** – Includes Databases, Analytics, Containers, Web app hosting, Serverless, IoT, Integration and APIs, Mobile backends and other services. • **Software as a Service (SaaS)** – Contains Digital workplace services, collaboration and communication services, Contact Centre as a Service etc.	**Hyperscalers and cloud service providers offer the below AI Infrastructure services -** • **GPU-based VMs** – All major cloud service providers offer services for creating VMs with Nvidia and AMD GPUs like Nvidia H100, Nvidia A100 and AMD Radeon • **Gen AI Services and foundational models catalogue** – Cloud service providers have expanded their portfolio to include Gen AI services. This includes AWS Bedrock, SageMaker JumpStart, Azure AI Studio, GCP Vertex AI Studio, IBM watsonx.ai etc. These services provide LLM catalog, model playground, model customisation and fine tuning and other services. • **MLaaS / Machine learning as a Service** includes machine learning services such as hosted notebooks, training and building machine learning models, cognitive APIs like vision, speech and document inference. • AI Frameworks include AI frameworks and libraries like PyTorch, TensorFlow, Keras, scikit-learn, numpy, etc. • **Vector and Search Databases as a Service** – Typically, Vector databases support RAG. Popular Vector databases include Pinecone, Milvus, Chroma, Weaviate etc. CSPs and database vendors have extended their search and other NoSQL databases to store, index and search vector embeddings. This includes AWS OpenSearch, Azure AI Search, MongoDB etc. • **Responsible AI Services** – It helps to build AI systems that are fair, unbiased; systems that follows regulations, compliance and laws. The services in this category include Azure Responsible AI Dashboard, Vertex AI Responsible AI services, SageMaker Clarify, watsonx.governance • **Explainable AI Services** – This covers Explainable AI, AI Transparency, AI logging and debugging. Popular tools in this area includes Fairlearn, Error Analysis, DiCE, Watson OpenScale, Azure Model Debugging etc. • **AI Security and Safety Services** – Gen AI and traditional AI/ML systems face exploitation attacks against them including prompt injection attack, jailbreaking etc. For example, AI researchers have used a jailbreaking technique called Do Anything Now (DAN) to bypass ChatGPT's guardrails to reveal its inner workings. Therefore, strong guardrails and optimum security are essential before production of an AI system. CSPs provide AI Security and Safety services. • **LLMOps** – It includes services to operate and optimise various components of AI and Gen AI projects including observability and monitoring, prompt analytics, performance tuning, cost management and optimisation (FinOps), data exploration and analysis etc.

Sidenote: Foundation model categories

Foundational Models are the workhorse behind Generative AI technology, and Gen AI applications like ChatGPT, Microsoft Copilot, Gemini, Stability Diffusion and Midjourney.

These models could be categorised in different ways – by modality, by model size, whether they are open source vs closed source and by use cases.

By Modality

Single Modality – Can generate only a single output type, e.g., either text or image

Multi Modality – Can generate multiple output types such as code, text, image

Text

Generates text response based on user's input. Response could be improved by changing the input or prompt – this technique is called Prompt Engineering. Example – ChatGpt, Gemini, Copilot in Windows

Image

Can generate image based on textual input by the user. Examples of Gen AI image generation includes Midjourney, Stable Diffusion and Open AI DALL-E

Code

Can generate code in programming languages like Python, SQL or Java. Tools like GitHub copilot, Tabnine or AWS Code Whisper generates code from comment and context.

Audio

Can be used to generate audio/ voice from user's text prompt, speech synthesis, generating music etc. Examples include OpenAI Whisper, Audio PaLM, Google AudioLM etc.

Video

With the increase in compute power and advancements in Gen AI technology, now it can help in generating video from text input. Primary examples include Nvidia GET3D and Google DreamFusion.

Other

Custom Industry, sector and function specific models. For example, Nvidia BioBERT is a model that helps in biomedical text mining and natural language processing tasks.

By Language model size

Large Language Model (LLM) or language models with billions of parameters; for example, GPT-4, a Large Language Model (LLM), purportedly contains 1.76 trillion parameters.

Small Language Model (SLM) or language models with fewer parameters than LLM. This helps the models perform better, answer questions in specific domain better, yet the quality is comparable to LLMs. For example, Microsoft's Phi-3 Mini has 3.8 billion parameters and Mistral 7B has only 7 billion parameters.

Open source vs Closed source models

Open source: Open source LLM has publicly accessible code, documentation and architecture and is often developed by a community. It allows free use, modification and distribution. Examples of open source LLM include - Llama 2/3 (Meta), BLLOM (Hugging Face), BERT (Google), Falcon 180B, XGen-7B (Salesforce), Flan T5, Mistral 7B

Closed source or proprietary: Cloud source models do not publicly expose their code, architecture or data. Such models include OpenAI's GPT 3/4, Google's Gemini, Claude Sonnet/Opus/Haiku, AI21 Jurassic and Stable Diffusion XL.

Note: The traditional definition of open source software (OSS) cannot be easily applied to AI technologies, especially Gen AI models (including LLM). OSS typically consists of code LLM, and foundational models consist of huge amounts of data and weights or tuning parameters apart from code. So, there is still discussion in the industry about understanding and differentiating between open vs. open source models. A truly open source model should disclose everything from code, its weights to the corpus of data used for training the model, the architecture and even the steps taken to build it.

By Use case

General purpose models: These models could be used for a wide range of tasks and are trained on large amounts of general web text. Examples include GPT 3/4, BERT etc.

Task-specific models: These models are tailored for a specific NLP task, like.g., summarising, answering questions, translating, etc. Examples include BART from Meta that is used for text generation tasks like summarisation and question generation and ALBERT from Google, that is used for question answering and sentence classification only.

Domain-specific models: These models perform better for domain-specific tasks like science, medicine or law. They are trained on text from a specific domain like science, medicine or law. Examples include BioBERT, which is trained on biomedical literature and used for biomedical NLP tasks, and SciBERT, which is trained on scientific text and is effective for scientific information extraction and question answering.

Identifying the optimum foundational model for your project, domain and use cause could be daunting. It is essential to have a framework to identify the optimum foundational model or LLM, based on features, quality, performance and cost.

Choosing the optimum foundation model is pivotal

Provider	Models
OpenAI	GPT2, GPT3, GPT4, GPT-4o, Codex, DALL-E, and Whisper
Google (Deepmind)	BERT, LaMDA, PaLM 2, Gemini, Codey, Imagen, Chirp
Meta AI	Llama 2/3, Code Llama
Cohere	Command, Command R, Command Light
Anthropic	Claude, Claude 3 Sonet, Claude Instant
Microsoft	Phi-2, Phi-3 (SLM)
Mistral AI	Mistral 7B, Mixtral 8X7B, Mistral Large
Hugging face	BLOOM
AI21 Labs	Jurassic 2, Jamba
Amazon	Titan Text/ Image/ Embeddings
IBM	Granite chat/ instruct/ code
Snowflake	snowflake-arctic-instruct
Databricks	DBRX
Deci AI	DeciLM, DeciCoder
Stability AI	Stable Diffusion XL
Nvidia	StyleGAN3, EG3D, Megatron

Features and Quality

- Model capability and features
- Model accuracy, coherence, grounding/ hallucination, fluency, relevance, similarity
- Supported by cloud providers, hosting/ deployment model
- Model size (LLM vs SLM)
- Enterprise readiness
- Knowledge cut off
- Sustainability considerations
- Customizability
- Sector/ Industry specificity
- Open source or closed source
- Regional support and Sovereignty
- Ethics, Transparency and Responsible AI considerations

Performance

Latency / Performance

Cost

Cost

Optimum Model or Models chosen for the use case

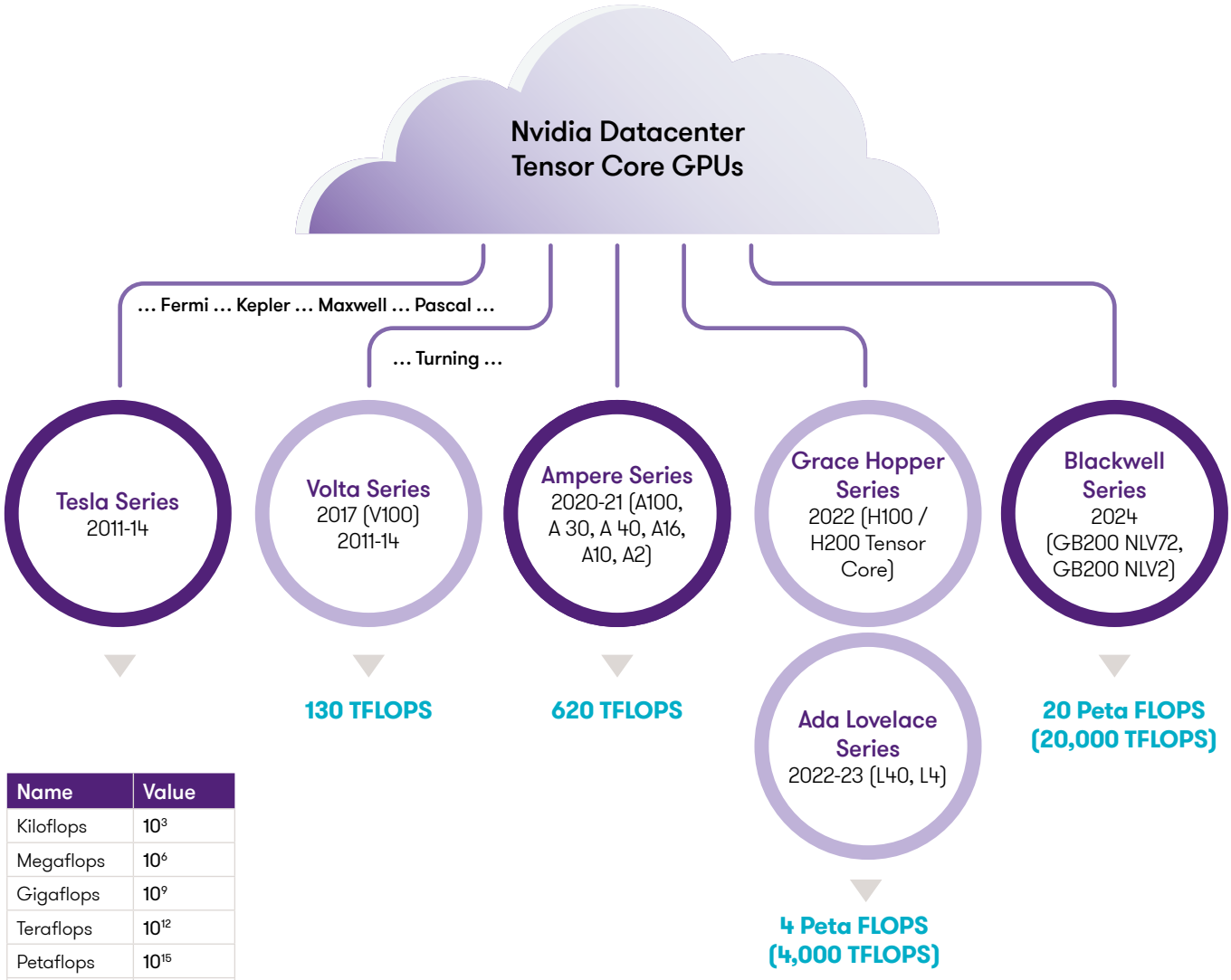
Sidenote: Emergence of GPUs, TPUs and Custom Silicon

GPUs are the workhorse for AI workloads, including model training and inferencing. GPUs as their name indicates stands for graphics processing unit, which was predominantly used for graphics applications like games, CAD etc. Because of their ultra parallel processing capabilities (GPUs can have thousands of core, while CPUs are limited to few cores) later, their use case got extended to VDI and HPC (high processing compute) workloads. GPUs played a huge role in the rise of AI and Gen AI. GPUs help AI workloads perform large number of floating point calculations every second. GPUs are often called as gold of AI/ Gen AI as they made training and inferencing of foundation models and LLMs with billions of parameters possible.

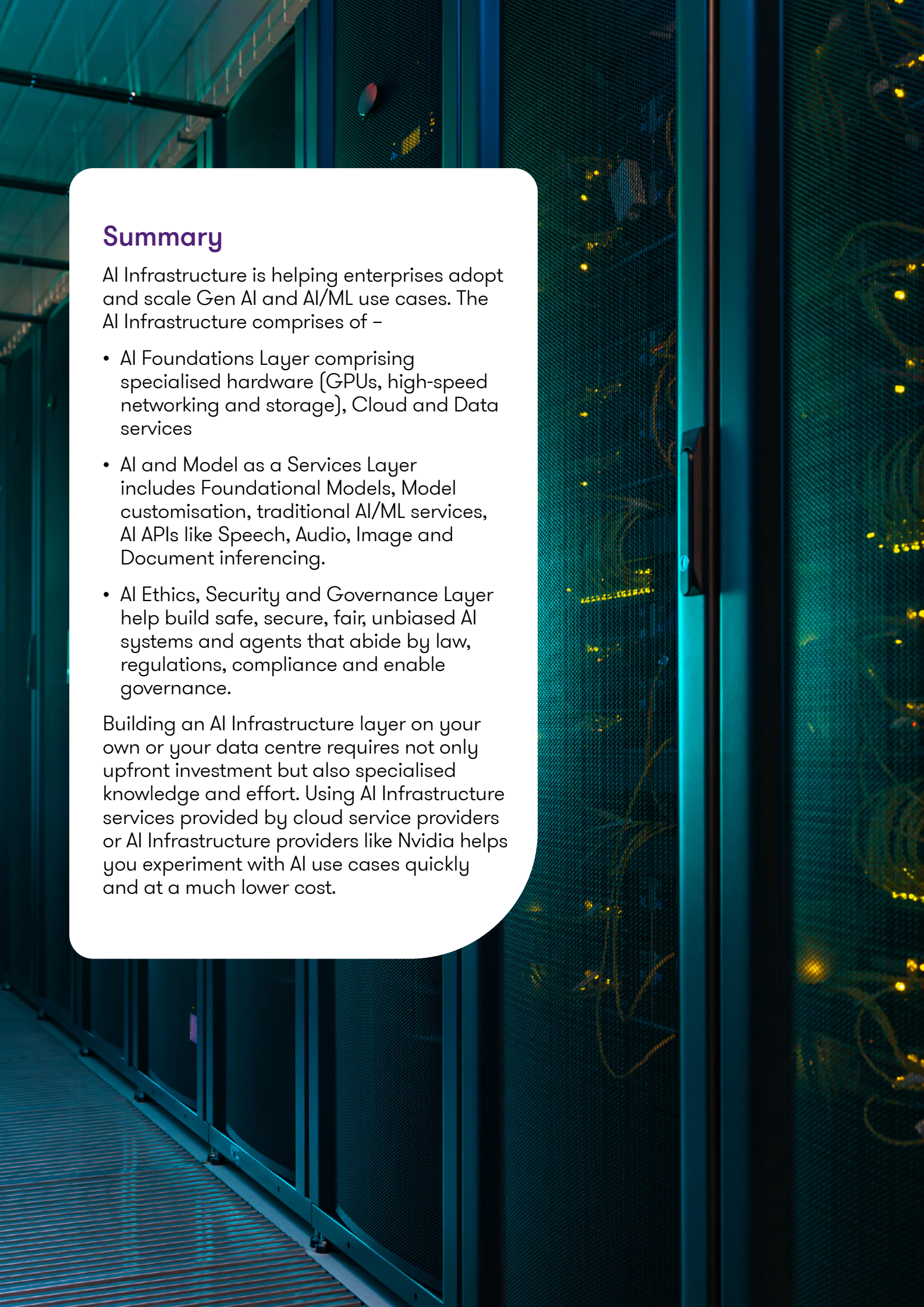
Emergence of Cloud computing and subsequently AI & Gen AI also saw the rise of ASIC chips like TPU or Tensor Processing Unit from Google, and even custom silicon chips and other hardware. These inventions caused the rise in computation capability exponentially, similar to Moore's Law, which predicted that the number of transistors on an integrated circuit would double every two years with a minimal cost increase.

The below diagram shows the increase of floating point calculation capacity of Nvidia GPUs in recent years -

Accelerated Computing - Massive Improvements in Chip, Hardware and Data Center Technologies Optimised for Gen AI



Name	Value
Kiloflops	10^3
Megaflops	10^6
Gigaflops	10^9
Teraflops	10^{12}
Petaflops	10^{15}
Exaflops	10^{18}
Zettaflops	10^{21}



Summary

AI Infrastructure is helping enterprises adopt and scale Gen AI and AI/ML use cases. The AI Infrastructure comprises of –

- AI Foundations Layer comprising specialised hardware (GPUs, high-speed networking and storage), Cloud and Data services
- AI and Model as a Services Layer includes Foundational Models, Model customisation, traditional AI/ML services, AI APIs like Speech, Audio, Image and Document inferencing.
- AI Ethics, Security and Governance Layer help build safe, secure, fair, unbiased AI systems and agents that abide by law, regulations, compliance and enable governance.

Building an AI Infrastructure layer on your own or your data centre requires not only upfront investment but also specialised knowledge and effort. Using AI Infrastructure services provided by cloud service providers or AI Infrastructure providers like Nvidia helps you experiment with AI use cases quickly and at a much lower cost.

Great Place To Work® Certified DEC 2023 – DEC 2024 INDIA ™

Best Workplaces™ in Professional Services
Great Place To Work® INDIA 2024

We are
Shaping Vibrant Bharat

A member of Grant Thornton International Ltd., Grant Thornton Bharat is at the forefront of helping reshape the values in the profession. We are helping shape various industry ecosystems through our work across Assurance, Tax, Risk, Transactions, Technology and Consulting, and are going beyond to shape more **#VibrantBharat.**

Our offices in India

● Ahmedabad ● Bengaluru ● Chandigarh ● Chennai
● Dehradun ● Goa ● Gurugram ● Hyderabad ● Indore
● Kochi ● Kolkata ● Mumbai ● New Delhi ● Noida ● Pune

Scan QR code to see our office addresses

www.grantthornton.in

Connect with us on

@Grant-Thornton-Bharat-LLP
@GrantThorntonBharat
@GrantThornton_Bharat
@GrantThorntonIN
@GrantThorntonBharatLLP
GTBharat@in.gt.com

Authors

Aniruddha Chakrabarti
Partner, Cloud and AI Services
Grant Thornton Bharat
E: aniruddha.c@in.gt.in

Tanya Khatri
Marketing Consultant
Grant Thornton Bharat
E: tanya.khatri@in.gt.com

To know more visit our website

CH 8805 | RD | SJ