FUJITSU | SUSE

Generative AI Whitepaper

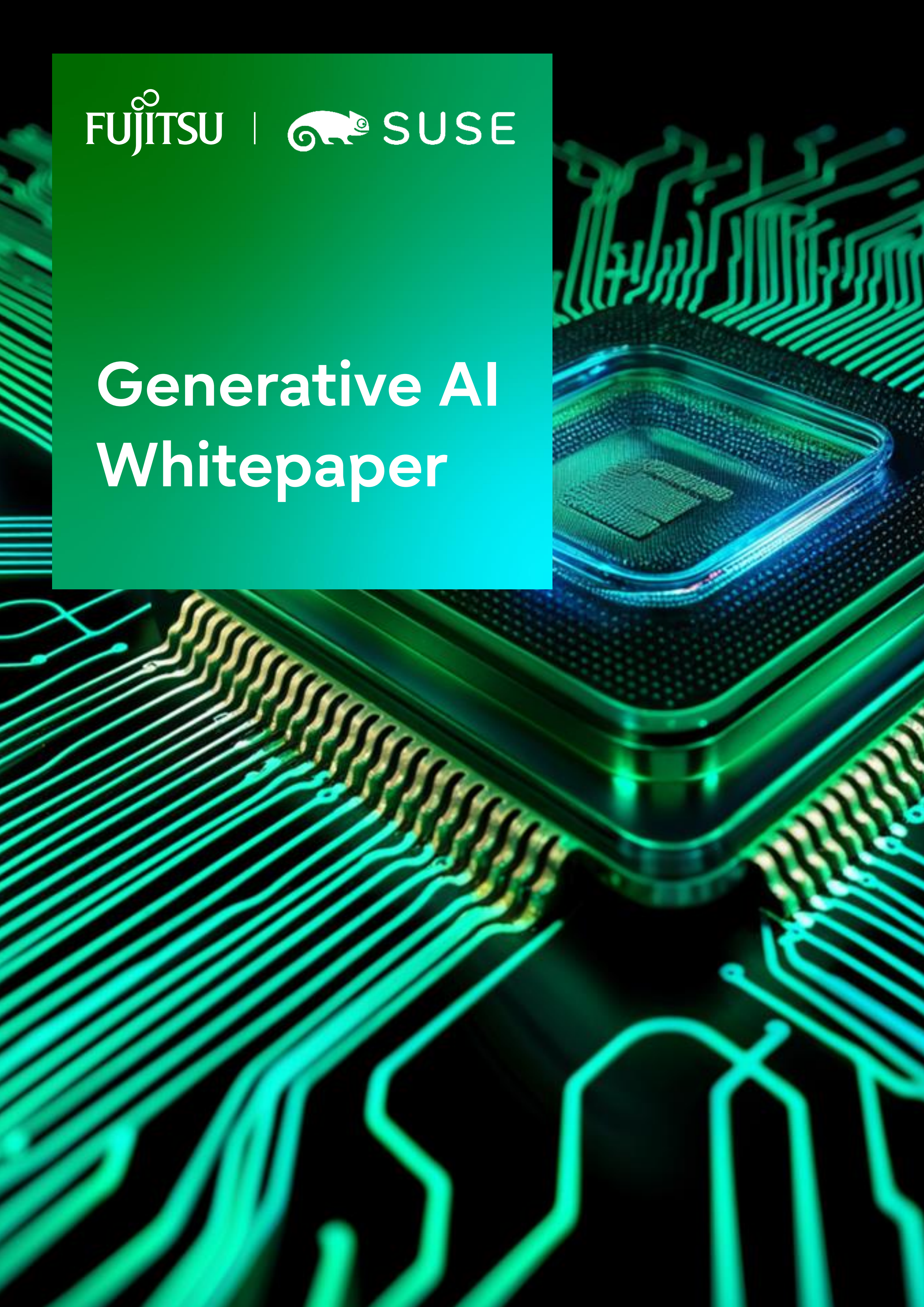


Table of Contents

1 Introduction

Artificial intelligence (AI) is transforming industries across the globe by enabling machines to learn, infer, and make decisions based on data. This document explores how organizations can implement AI solutions with the help of Fujitsu, focusing on training, inferencing, solution architectures, and services that support AI deployments. From technical considerations to ethical guidelines, this guide serves as a comprehensive overview of Fujitsu AI's integration into enterprise systems.

1.1 Fujitsu and AI-Overview

Artificial Intelligence is set to become fundamental to the operations of every industry sector-wide, it is a technology that offers great promise not just for business improvements but for human-centric innovation. AI utilizes relevant data to automate and amplify human capabilities, eliminate mundane and repetitive tasks, generate additional revenue streams, and help with decision making and faster time to market. However, despite all the potential benefits, many organizations have a poor understanding as to what AI can deliver, the effective return on investment and are unsure about what to deploy and where. There is a general inability to identify the right data sets to derive benefits and a lack of in-house skills to build an AI infrastructure or foundation across its life cycle requirements, all of which is leaving organizations unsure as to how to proceed. This is where we at Fujitsu can offer immense value. Working together we help you to develop an AI strategy to overcome the challenges you face in:

- developing an integrated foundation to obtain all relevant data.
- understanding the right skill sets needed
- executing AI in a phased manner without disrupting business
- gaining visibility as to the ROI investment benefits

Our AI Products and AI teams can enable you to lay the right AI technology foundations to meet both your immediate and future business needs.

Understanding the AI value chain to deliver data-driven transformation solutions.

Using our wide-range experience within the field of AI and data-driven business transformation, we apply what we know to advise our customers where they should be deploying AI and which platforms and solutions they should use.

As experts in implementing deep learning solutions, we create reference architectures that can be tuned to specific deployment scenarios. We provide end-to-end AI solutions and guidance based on their use cases.

Why choose Fujitsu as your expert AI partner?

As experts in AI deployment, we understand that the secret to any successful AI implementation is the creation of a foundation based on systems design and business needs. We have been involved in developing and deploying artificial intelligence and associated technologies and take a use-case driven incremental approach to a data-driven transformation that is agile, manageable, and focused on results. We work with you to define a clear transformation strategy centred around a business problem or opportunity in the value chain to develop an AI solution that will deliver value to your organization.

Audience

This technical document is designed for a diverse audience that includes sales teams, solution architects, IT Managers, marketing professionals and other key stakeholders involved in AI solution deployment and strategy. It provides a comprehensive understanding of the technical processes, such as AI training, inferencing, and solution architecture, while also addressing practical concerns like hardware selection, sizing, and validation. For solution architects, it offers insights into building AI infrastructure, while marketing and management teams can gain a better grasp of how AI integrates into business strategies, security, and ethical considerations. The document serves as both a technical guide and a strategic resource, tailored to empower decision-makers to successfully navigate AI solutions within their organizations.

1.2 Contents Overview

Chapter 2: Training, Inferencing, and Finetuning

This chapter provides a deep dive into the core components of AI—training, inferencing, and finetuning. It defines the processes behind training machine learning models and explores how models are finetuned for specific tasks. Additionally, it introduces inferencing, which is the application of trained models to make predictions in real-world scenarios. Use cases along with common challenges and solutions, are also discussed.

Chapter 3: Intel Confidential Computing

Intel's Confidential Computing technologies offer enhanced security measures for sensitive data processing. This chapter explains the importance of secure computing environments and how Intel's Confidential Computing can protect data during AI model training and inferencing, particularly in multi-tenant or cloud environments.

Chapter 4: Solution Components

To build robust AI solutions, high-performance hardware such as PRIMERGY Servers and NVIDIA GPUs are essential. This chapter examines the role of these components in accelerating AI workloads, along with the necessary NVIDIA licensing options that support scalable AI deployments.

Chapter 5: Private GPT – Fujitsu, Systemhaus Ulm and DASU

Fujitsu Private GPT solution is designed to provide an intelligent private chat based local knowledgebase. This chapter explains the on-premise solution that brings GenAI technology within the private scope of your enterprise, creating an environment where your specific data can be accessed in private and securely by your employees. Since all processing is done locally, the security of your enterprise data is assured.

Chapter 6: SUSE® Rancher Prime, SUSE® Virtualization, and Red Hat

Open-source platforms like SUSE® Rancher Prime, SUSE® Virtualization, and Red Hat play a crucial role in managing containerized environments for AI workloads. This chapter explores the specific capabilities these platforms offer in orchestrating AI models, managing resources, and enabling seamless integration across different cloud and on-premises systems.

Chapter 7: Customer Use cases

Accurately sizing AI infrastructure is crucial for ensuring optimal performance. This chapter outlines Fujitsu's sizing guidelines, helping organizations determine the correct infrastructure needed to support different AI workloads, from training models to deploying inferencing pipelines.

Chapter 8: Validation Results and Benchmarks

Generative AI models require rigorous validation to ensure their performance and accuracy. This chapter presents real-world benchmarks and validation results for generative AI models, providing performance insights that help optimize AI systems for various use cases.

Chapter 9: Sizing

Accurately sizing AI infrastructure is crucial for ensuring optimal performance. This chapter outlines Fujitsu's sizing guidelines, helping organizations determine the correct infrastructure needed to support different AI workloads, from training models to deploying inferencing pipelines.

Chapter 10: Fujitsu and AI Ethics

As AI technologies continue to evolve, ethical considerations become more important. This chapter delves into Fujitsu's approach to AI ethics, including guidelines and principles designed to ensure responsible AI development and deployment. Issues such as bias, fairness, transparency, and accountability are discussed in detail.

Chapter 11: Fujitsu Professional Services for AI

This chapter presents the range of services offered by Fujitsu to support AI initiatives, from consulting and solution design to managed services. It highlights Fujitsu's Gen AI Test Drive, which allows clients to build their own proof of concept (POC), along with other AI-validated solutions and infrastructure services. Key services for deployment and ongoing support are also outlined.

Chapter 12 Conclusion

2 Training, Fine-Tuning, and Inferencing

2.1 What is training?

The training of machine learning (ML) models generally involves feeding a model a vast amount of data and iteratively adjusting its internal parameters (weights and biases) to minimize a defined loss function. This process, often employing gradient descent-based optimization algorithms, aims to learn underlying patterns and relationships within the data, enabling the model to make accurate predictions or generate outputs. The complexity of this process scales dramatically with the size and complexity of the model and the dataset. Hardware such as PRIMERGY play a critical role, with powerful GPUs and specialized hardware accelerators becoming increasingly essential for training even moderately sized models.

When it comes to training generative AI (GenAI) the core principles of training remain consistent, but there are unique challenges and opportunities.

Data Scale: GenAI models, particularly large language models (LLMs) and large diffusion models, require datasets orders of magnitude larger than those used for many other ML tasks. We are talking terabytes, petabytes, and even exabytes of data. This necessitates distributed training across multiple machines, often utilizing high-bandwidth, low-latency interconnects like InfiniBand.

Model Size

GenAI models are notoriously large, boasting billions or even trillions of parameters. This massive parameter count directly impacts the computational resources required for training. Specialized hardware like Tensor Processing Units (TPUs) from Google and other custom silicon designed for matrix multiplications are frequently employed to accelerate the training process. The memory capacity of these systems also becomes a critical bottleneck, demanding sophisticated memory management techniques.

Computational Intensity

Training GenAI models is computationally extremely demanding. The sheer number of calculations required for each training iteration necessitates massive parallel processing capabilities. Clusters of high-end GPUs, often interconnected with NVLink or similar technologies for faster communication, are commonly used, but even these can struggle with the largest models. This often leads to training times measured in weeks or even months.

Architectural Complexity

GenAI models often employ sophisticated architectures like Transformers, which are particularly computationally intensive. Optimizations like mixed-precision training (using lower-precision floating-point numbers) and model parallelism (splitting the model across multiple devices) are essential to manage the computational burden.

Re-i nforcement Learning from Human Feedback (RLHF)

This powerful technique is used to align the behavior of large language models and other generative AI models with human preferences. It is particularly crucial because directly training a GenAI model on a massive dataset does not guarantee desirable behavior; the model might generate outputs that are factually incorrect, toxic, or simply unhelpful, even if statistically probable within the training data. RLHF addresses this by incorporating human feedback into the training loop, guiding the model towards generating more aligned and desirable outputs.

2.2 What is fine-tuning?

Once a foundational GenAI model is trained, fine-tuning allows for adaptation to specific tasks or datasets. It enables developers to tailor the model's behavior to specific needs, such as generating text in a particular style, translating languages with higher accuracy, or classifying images with greater precision. This process leverages the pre-trained model's knowledge, significantly reducing the computational cost and data requirements compared to training from scratch.

Fine-tuning typically involves:

Transfer Learning: The pre-trained model's weights are used as a starting point, and only a subset of the parameters is adjusted during fine-tuning. This dramatically reduces training time and data needs.

Targeted Datasets: Smaller, task-specific datasets are used to adapt the model's behavior. For example, a pre-trained LLM might be fine-tuned on a corpus of medical texts to generate medically accurate responses.

Hardware Considerations: While fine-tuning requires significantly less computational power than training, it still benefits from the use of GPUs or TPUs. However, the scale is often smaller, allowing for the use of less powerful, but still high-performance, hardware compared to training.

2.3 What is inferencing?

Inferencing refers to the process of using an existing machine learning model to make predictions or decisions based on new data inputs. In other words, it involves applying a trained model to unseen data and generating outputs that can be used for various purposes such as classification, regression, or recommendation systems. In the context of large language models (LLMs) like Mistral, inferencing involves feeding input text into the model and retrieving relevant responses based on the patterns and relationships learned during training. This process is also known as "model serving" or "deployment."

Inferencing is a critical component of many AI applications, including natural language processing (NLP), computer vision, and predictive analytics. It enables organizations to leverage the power of ML models in real-world scenarios, such as chatbots, recommendation systems, and decision support tools.

Inferencing, in the context of LLMs, may be executed on a laptop, server, server farm, with or without NPU, one or multiple GPUs, depending on the size of the model used and the requirements on response time (latency).

Typical performance metrics are:

- **Time to first token:** This time includes the need to load the LLM model into main memory or GPU, if that has not been done before.
- **Tokens per second:** Once the first token is replied to, the speed of following tokens varies depending on the hardware platform used, the size of the LLM, the complexity of the request and the context size.

In this context, a token is the smallest chunk of data an LLM can operate on. Tokens are derived by a process called tokenization. It turns text and other characters into tokens the LLM can use as input. This process is often also called embedding. Tokens are floating point tensors (multi-dimensional vectors) pointing into a vector space (or “embedding space”). Typical vector spaces have 292, 1024, 4096 or even higher dimensions.

Advantages of using tensors are:

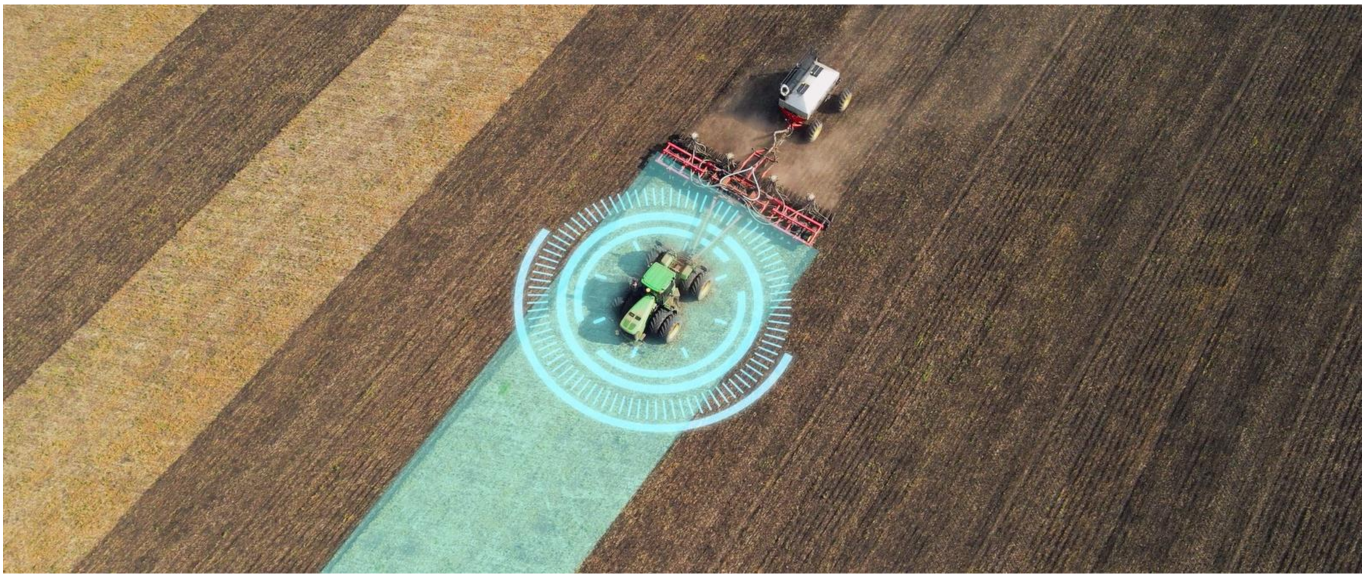
- Tokens with similar meaning are located close together.
- Simple matrix operations, such as translation and rotation along certain axes can be used to translate one language to another.

A single inference (or “prompt execution”) requires main memory space to store the prompt, the embeddings (tokens) and some temporary memory. In practice, LLMs typically use a technique called "memory reuse" or "tensor reuse" to minimize memory usage during processing. This involves reusing the same memory buffer for different layers, by overwriting the previous layer's output with the new one. This approach helps reduce memory requirements and improve efficiency.

2.4 General Machine Learning and AI Use Cases

Machine learning and AI have numerous applications across different industries and domains. Some common use cases include:

- **Natural Language Processing (NLP):** Analyzing text data for tasks such as sentiment analysis, language translation, or question answering using transformer-based models.
- **Recommendation systems**: Predicting user preferences and suggesting relevant items based on historical behavior patterns using collaborative filtering algorithms or deep learning models.
- **Anomaly detection:** Identifying unusual events or behaviors in data streams, such as fraudulent transactions or network intrusions, by comparing them against normal patterns learned from training data.
- **Image recognition:** Identifying objects, faces, or scenes in images using pre-trained VLM (Vision Language Models).
- **Image segmentation:** Creating image masks which can be used for training VLMs.
- Fraud Detection: Inferencing is used in financial institutions to detect fraudulent activities by analyzing transaction patterns and identifying anomalies.
- **Predictive Maintenance:** Inferencing can be employed in industrial settings to predict equipment failures or maintenance needs by analyzing sensor data and historical patterns.
- **Autonomous Vehicles:** Self-driving cars rely on inferencing models for object detection, path planning, and decision-making while navigating through complex environments.
- **Healthcare Diagnostics:** Inferencing can assist in medical diagnosis by analyzing patient data, images, or signals to identify potential health issues or abnormalities.
- **Speech Recognition:** Voice assistants use inferencing models (e.g., 'Whisper' by OpenAI) to transcribe spoken words into text and understand user commands for various tasks.
- **Personalized Advertising:** Online advertising platforms use inferencing models to deliver targeted ads based on users' browsing history and interests.



2.5 Challenges with Training and Fine-tuning

Training and fine-tuning generative AI models, while yielding impressive results, present significant challenges across various domains:

2.5.1 Data Challenges

Data Scarcity

While GenAI models thrive on massive datasets, obtaining sufficient high-quality data for specific tasks can be incredibly difficult. This is particularly true for niche domains or tasks with limited publicly available information. Also, the copyright for data must be considered. Data augmentation techniques can help, but they have limitations and can introduce biases.

Data Bias and Fairness

Training data often reflects existing societal biases, leading to models that perpetuate or even amplify these biases in their outputs. Mitigating bias requires car eful data curation, preprocessing, and the development of fairness-aware training techniques. Identifying and addressing these biases is an ongoing and complex research area.

Data Quality and Consistency

The quality and consistency of training data are paramount. Inconsistent formatting, noisy data, or errors can significantly impact model performance and lead to unreliable outputs. Data cleaning and preprocessing are crucial but time-consuming steps.

Data Privacy and Security

Many datasets used for training GenAI models contain sensitive information. Protecting data privacy and security during training and deployment is crucial, requiring robust anonymization techniques and secure data handling practices.

2.5.2 Computational Challenges

Computational Cost

Training large GenAI models requires immense computational resources and energy. The cost of training can be prohibitive for many organizations, limiting access to this technology. This also contributes significantly to the carbon footprint of AI development.

Hardware Limitations

Even with powerful hardware like TPUs and high-end GPUs, training the largest models can take weeks or months. Memory limitations and communication bottlenecks between devices can further slowdown the training process.

Scalability

Scaling training to larger models and datasets requires sophisticated distributed training techniques. Maintaining efficiency and stability across large clusters of machines presents a significant engineering challenge.

2.5.3 Model Challenges

Overfitting

Large models can easily overfit the training data, performing well on seen examples but poorly on unseen data. Regularization techniques and careful hyperparameter tuning are essential to mitigate overfitting.

Interpretability and Explainability

Understanding why a GenAI model produces a particular output can be extremely difficult. The lack of interpretability hinders debugging, bias detection, and trust in the model's predictions.

Controllability and Safety

Controlling the outputs of GenAI models and ensuring their safety is a major challenge. Models can generate outputs that are toxic, biased, or factually incorrect. Developing techniques to improve controllability and safety is crucial for responsible AI development.

Generalization

Ensuring that a model generalizes well to unseen data and different contexts is crucial. Models might perform well on the training data but fail to generalize to real-world scenarios.

2.5.4 Fine-tuning Specific Challenges

Catastrophic Forgetting

Fine-tuning can sometimes lead to catastrophic forgetting, where the model forgets information learned during the pre-training phase. Techniques like regularization and knowledge distillation can help mitigate this issue.

Data Efficiency

While fine-tuning requires less data than training, obtaining sufficient high-quality data for specific fine-tuning tasks can still be challenging.

Transferability

The effectiveness of transfer learning depends on the similarity between the pre-training and fine-tuning tasks. Fine-tuning might not be effective if the tasks are too dissimilar.

Addressing these challenges requires a multi-faceted approach involving advancements in algorithm design, hardware development, data management techniques, and ethical considerations. Ongoing research and development are crucial for overcoming these hurdles and unlocking the full potential of generative AI.

2.6 Challenges with Inferencing

While inferencing offers many benefits for real-world applications, there are several challenges associated with its implementation and deployment at scale:

- **Model accuracy vs latency trade-off:** There is often a balance between achieving high prediction accuracy while maintaining low inference latency (time taken to generate predictions). This requires careful selection of model architectures and optimization techniques.
- **Hardware requirements:** Inferencing can be computationally intensive, especially for large-scale deployments or complex models like deep neural networks. Specialized hardware such as NPUs or GPUs may be required to achieve acceptable performance levels.
- **Data privacy concerns:** When dealing with sensitive user data (e.g., healthcare records), ensuring compliance with regulations related to data protection and security becomes crucial during the inference process. Organizations must ensure that sensitive data is protected during inferencing, adhering to regulatory requirements such as GDPR and HIPAA. Running, e.g., LLMs locally instead of at a cloud service are highly recommended when working with sensitive company data.
- **Model drift over time:** As new data is collected; existing models may become less accurate due to changes in underlying patterns or distributions. Regular monitoring and updating of deployed models are necessary for maintaining their effectiveness.
- **Scalability issues:** Handling large volumes of incoming requests efficiently can pose challenges when scaling up inferencing systems across multiple servers/nodes without compromising on response times.

3 Intel Confidential Computing

Imagine a world where you can safely unleash the power of AI on your most sensitive data without fear of breaches or misuse. That is the promise of confidential computing with Intel TDX and Nvidia AI GPUs. Briefly, confidential computing with Intel TDX based on Intel Xeon 5th Generation Scalable CPUs provides a secure foundation for the future of AI, enabling organizations to harness its full potential without compromising data privacy and security.

Here is the breakdown

The Problem

AI algorithms thrive on data, often extremely sensitive data. But processing this data in the public cloud or private on-prem datacenters raises significant privacy and security concerns. What if the cloud provider or the private cloud / on-prem datacenter itself is compromised? Confidential computing is a revolutionary approach to data security that encrypts data not just at rest or in transit, but while it is being processed. This means that even if a cloud provider or malicious actor gains access to the server, they will not be able to see or tamper with the data.

Think of it like this

Imagine a locked box where you can perform calculations without opening it. Only you have the key to unlock the box and access the results.

Here is the key takeaway

- Data is encrypted and processed within a secure "enclave" - a protected environment isolated from the rest of the system.
- Only authorized parties with the right keys can access and decrypt the data.
- This ensures data confidentiality, integrity, and privacy.

3.1 Benefits of Confidential Computing

- **Enhanced security:** Protects data from unauthorized access and manipulation.
- **Improved privacy:** Allows sensitive data to be processed in the cloud while maintaining privacy.
- **Increased trust:** Enables collaboration and sharing of sensitive data with greater confidence.
- **Compliance:** Helps organizations meet regulatory requirements for data protection.

Confidential computing is still a modern technology, but it has the potential to transform how we think about data security and privacy in the cloud. It opens new possibilities for innovation and collaboration while ensuring data remains protected.

3.2 The Solution: Intel TDX for Confidential AI

Fort Knox for Your AI

Intel TDX creates secure enclaves, like virtual vaults, within the cloud environment or on your on-prem datacenter / private cloud. These enclaves isolate your AI workloads (data and models) from everything else, even the cloud provider's operating system or the on-prem hypervisors of your virtualization environments.

Hardware-Enforced Protection

Unlike software-only solutions, TDX leverages Intel's hardware to create these impenetrable enclaves, making them extremely resistant to attacks.

Proof of Trustworthiness

TDX includes attestation mechanisms, providing verifiable proof that your AI workload is running within a secure and unmodified enclave.

Benefits

1

Unleash AI on Sensitive Data

Confidently analyze patient records, financial transactions, or classified information with unparalleled security.

2

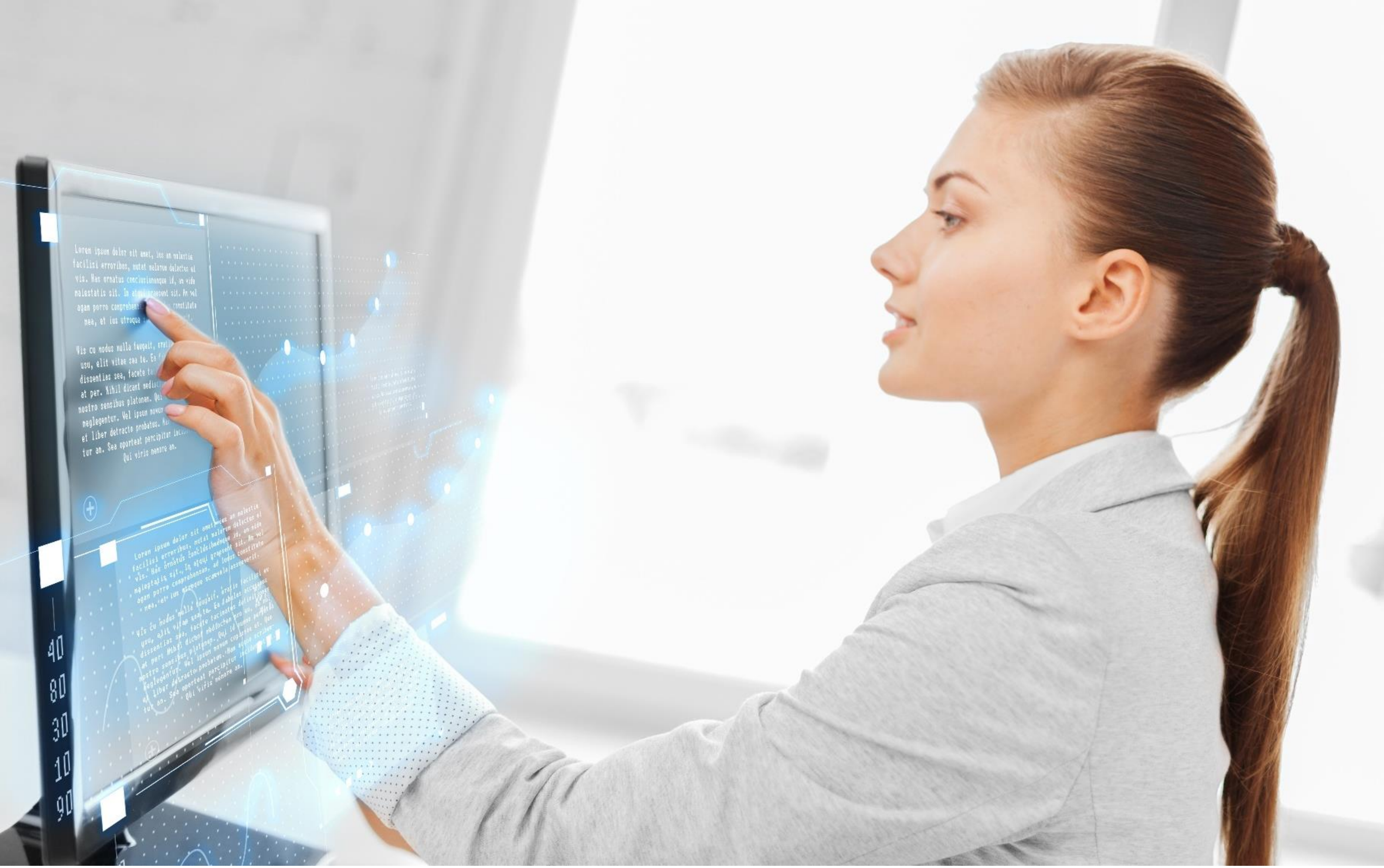
Protect Valuable AI Models

Safeguard your proprietary AI models from theft or tampering, preserving your competitive edge.

3

Build Trust with Users

Demonstrate your commitment to data privacy and security, fostering trust with customers and partners.



Use Cases

Healthcare

Develop AI-powered diagnostic tools that preserve patient privacy.

Finance

Combat fraud and analyze risk with confidential AI models.

Government

Process sensitive data for intelligence and defense applications with enhanced security.

Research

Enabling collaboration and data sharing in sensitive research projects.

Intel is a major player in the confidential computing space, actively driving its development and adoption through various initiatives:

Intel SGX (Software Guard Extensions)

- This is Intel's flagship confidential computing technology, implemented in their processors since 2015.
- SGX creates secure enclaves within the CPU, where code and data are protected from even the operating system and hypervisor.
- This allows sensitive operations to occur within a trusted environment, ensuring data confidentiality and integrity.

Intel Software Guard Extensions (SGX) SDK:

- Intel provides a Software Development Kit (SDK) to simplify the development of applications that leverage SGX.
- This SDK includes libraries, tools, and documentation to help developers create secure and confidential applications.

Intel SGX Enclave Manager

- This is a software component that manages the creation, provisioning, and lifecycle of SGX enclaves.
- It simplifies the process of deploying and managing confidential applications.

Collaboration with Cloud Providers

- Intel works closely with major cloud providers like Microsoft Azure, Amazon Web Services (AWS), and Google Cloud Platform (GCP).
- They collaborate to integrate SGX into their cloud platforms, making confidential computing accessible to a wider audience.

Open Enclave SDK

- Intel is a contributor to the Open Enclave SDK, an open-source project that aims to standardize the development of confidential applications.
- This promotes interoperability and adoption of confidential computing across different platforms.

Research and Development

- Intel continues to invest in research and development to enhance SGX and explore new confidential computing technologies.
- This includes efforts to improve performance, security, and scalability of confidential computing solutions.

In summary, Intel is actively shaping the future of confidential computing by

- **Providing hardware and software solutions:** Intel SGX and related tools enable developers to build secure and confidential applications.
- **Collaborating with industry leaders:** Intel works with cloud providers and other organizations to promote the adoption of confidential computing.
- **Investing in research and development:** Intel continues to innovate and improve confidential computing technologies.

These efforts are crucial for driving the widespread adoption of confidential computing and ensuring its potential to revolutionize data security and privacy in the future.

Intel is a new confidential computing technology that builds upon Intel SGX, offering enhanced security and flexibility. Here is how it is utilized in confidential computing platforms:

1

Enhanced Security

Isolation at the Virtual Machine (VM) Level

TDX isolates entire VMs within a secure enclave, providing a stronger security boundary compared to SGX, which protects only specific code and data segments.

Protection Against Hypervisor Attacks

TDX enclaves are protect ed from the hypervisor, making them resistant to attacks that could compromise the underlying virtualization layer.

Secure Boot and Attestation

TDX enclaves support secure boot mechanisms and attestation features, ensuring that the enclave environment is trustworthy before execution.

2

Flexibility & Scalability

VM-Based Approach

TDX's VM-based approach allows for greater flexibility in deploying and managing confidential applications. Developers can use existing virtualization technologies and tools to manage their confidential workloads.

Scalability

TDX can be deployed on a wide range of Intel platforms, including servers and edge devices, enabling scalable confidential computing solutions.

3

Integration with Existing Platforms

Compatibility with SGX

TDX is designed to be compatible with existing SGX applications, allowing for a smooth transition to the enhanced security features.

Support for Cloud Providers

Intel is working with cloud providers to integrate TDX into their platforms, making it accessible to a wider audience.

How TDX is Utilized in Confidential Computing Platforms

- **Secure Cloud Environments:** TDX can be used to create secure cloud environments where sensitive data and applications are protected from unauthorized access and manipulation.
- **Data-Centric Workloads:** TDX is ideal for data-centric workloads, such as financial services, healthcare, and government applications, where data privacy and security are paramount.
- **Edge Computing:** TDX can be deployed on edge devices to protect sensitive data and applications at the edge of the network.

As Intel continues to develop and refine TDX, it has the potential to further revolutionize confidential computing and make it a more accessible and powerful security solution for a wide range of applications.

Hardware Requirements

- Dual PRIMERGY M7 Servers with 5th Generation Intel Xeon Scalable CPUs
- Drives capable of SED for local Data storage
- RAID Controller with Encryption capabilities (FUJITSU PRAID EP640/EP680) in case of local data Storage → remote storage requires remote Encryption capabilities.
- Nvidia GPU with at least Nvidia H100 or Nvidia H100 NVL GPU. (when a GPU is required for AI inferencing or AI learning)
- Network protocols, which encrypt data during transfer like TLS or SMB encryption or other protocols.

3.3 Summary

Confidential computing, powered by Intel TDX, is revolutionizing the way we think about AI security and privacy. Confidential computing with Intel TDX offers a powerful approach to securing AI workloads. It allows for the development and deployment of AI applications that are both innovative and trustworthy, enabling the responsible use of AI in sensitive domains.

4 Solution Components

Our Fujitsu servers are built to excel in AI-driven operations, offering robust solutions across the entire AI lifecycle. Offering best-in-class performance and energy efficiency on industry-standard CPU technology, PRIMERGY servers provide the simplicity and cost profile needed for backbone operations as well as the power to bring affordable AI operations within reach for sustainability transformation programs. With high-performance components for AI training and inference and storage solutions optimized for I/O-intensive AI workloads, they ensure seamless performance for even the most demanding tasks. These features make our servers ideal for handling diverse AI workloads with unmatched efficiency and reliability, driving innovation in any AI environment.

4.1 Server Systems

4.1.1 PRIMERGY RX2540 M7

The PRIMERGY RX2540 M7 is well-suited for AI workloads due to its powerful dual Intel Xeon Scalable processors, high memory capacity and flexible storage options. It supports GPU acceleration, a key requirement for AI model training and inference. With its advanced I/O capabilities, including PCIe 5.0, and enhanced cooling technology, it ensures efficient data processing and system reliability, making it ideal for handling demanding AI applications and high-performance computing tasks. The PRIMERGY RX2540 M7 generation x86 server based on dual sockets delivers the latest in performance, improved usability, and flexible expandability in an optimized compact 2U chassis. The PRIMERGY RX2540 M7 forms the valuable standard in every modern data center, using the latest technology developments to run every workload from the most basic to business-critical applications depending on the chosen configuration.

Equipped with the latest 4th or 5th generation of Intel® Xeon® Scalable Processors with up to 60 / 64 cores and 4x UPI 2.0 links, there are resulting performance improvements of more than 40% compared to the previous generation processors. Along with enhanced DDR5 memory technology supporting up to 4,800 MT/s or 5,600 MT/s, the server features a flexible, large amount of memory capacity. Configurable in 32 DIMM slots are in total 8TB memory with latest DDR5 modules supported. The support of Compute Express Link (CXL) with 4x 16 devices is included. The modular design of the server offers excellent expandability with up to 12x 3.5" SAS/SATA, up to 24x 2.5" SAS/SATA/NVMe storage drives. In addition, 6 further 2.5" storage devices SAS/ SATA/NVMe are available as an option on the rear of the chassis. Additional expansion options are provided by up to 8x PCIe 5.0 slots and SAS 24G for upcoming devices. Moreover, the server can be equipped with two double-width or up to six single-width NVIDIA GPU cards.

Thus, the server also provides optimized performance for AI and HPC workloads. An onboard OCP v3 LAN connection completes the overall picture. The server system also includes the latest security technologies to help secure sensitive workloads and enable new opportunities to unleash the power of data. PRIMERGY RX2540 M7 always provides Platform Firmware Resilience (PFR) to help protect against platform firmware attacks and is designed to detect and correct them before they can compromise or disable the machine.

Where the right performance, expandability, and efficiency are essential, the PRIMERGY RX2540 M7 is the ideal server for business-critical workloads such as collaboration, business processing, AI, machine learning, graphics rendering, or in-memory databases.

4.1.2 PRIMERGY GX2560 M7

The PRIMERGY GX2560 M7 is a powerhouse designed to elevate your computing experience. Equipped with the 4th or 5th generation of Intel® Xeon® Scalable Processor, combined with 4x NVIDIA HGX H100 GPUs this server delivers unrivalled generative AI and HPC performance with high efficiency. With DDR5 memory support, you can achieve even greater processing power for your demanding workloads. Featuring up to 4 UPI and 32 DDR5 memory slots operating at a blazing bandwidth of up to 5600 MT/s, the GX2560 M7 ensures seamless multitasking and accelerated data processing. Its front-end can accommodate up to 6x SATA/SAS and 2x NVMe storage devices, providing ample space for your expanding data needs.

Unlocking new levels of computational prowess, this server is equipped with 4x NVIDIA HGX H100 GPUs (SXM5 type). Prepared for enhanced visualizations and accelerated computations. With the addition of the 5th generation of PCIe slots, offering up to 6 slots, you can optimize your system for high-speed data transfers and expandability. The PCIe 5.0 IB card further enhances connectivity and data exchange. Rest easy knowing that the PRIMERGY GX2560 M7 comes with comprehensive BMC support, including the security features, API integration and TLS1.2 compatibility. Your data is safeguarded, and you have complete control over your server's operations. Experience the next generation of performance and reliability with the PRIMERGY GX2560 M7. Elevate your business to new heights and stay ahead of the competition.

4.2 Storage Systems

4.2.1 ETERNUS AX Series

The Fujitsu Storage ETERNUS AX series is a highly scalable, all-flash storage solution designed to power AI workloads and digital transformation with unmatched flexibility and performance. Equipped with innovative NVMe technology, it accelerates data-inte nsive AI applications while ensuring seamless data management across edge, core, and cloud environments. Ideal for AI-driven tasks, it guarantees continuous availability and robust data protection, making it perfect for hybrid IT infrastructures. Its ability to scale and adapt to demanding AI workloads ensures businesses can harness the full potential of their data efficiently and securely.

4.2.2 NetApp AFF A800

The AFF A800 is the world's fastest, most cloud-connected all-flash array. You can connect to more clouds, with more intelligent cloud services. Running on ONTAP software, AFF A800 allows you to move your data and applications where they run best, eith er on the premises or in the cloud, leveraging the same ONTAP data management. Embrace multi cloud by using Amazon Web Services (AWS), Azure, Google Cloud, and other leading cloud provide rs for backup, disaster recovery, tiering, cloud analytics, and workload bursting for business agility. ONTAP lets you harness the power of the hybrid cloud.

Simplicity and efficiency

The AFF A800 is simple to deploy. You can have the new system up and running in just 10 minutes. And deduplication, compression, and compaction, the AFF A800 provides capacity savings up to 10x, increasing effective capacity. The AFF A800 offers leading density with an effective capacity of over 2.5PiB in a single 4U system, for a storage footprint that is up to 37 times smaller.

Industry-leading performance by every measure

The AFF A800 delivers sub-200 s latency, 1.3M IOPS at 500 s latency and massive throughput of up to 34GB/s with an HA pair. A NAS cluster delivers up to 11.4M IOPS at 1ms latency, 300 GB/s of throughput and 316PB of effective capacity. A SAN cluster delivers up to 7.8M IOPS at 500 s latency, 204GB/s of throughput, and 158PB of effective capacity.

NVMe Performance and Connectivity in a Compact Design

The NetApp AFF A800 is designed to deliver extreme performance from a compact package. Each 4U chassis accommodates dual controllers for high availability (HA) and includes 48 slots for NVMe SSDs. In addition to 32Gb and 16Gb FC, network options include the storage industry's first 100GbE connectivity, as well as 40GbE and 10GbE. An NVMe-powered SAN scale-out cluster supports up to 12 nodes (6 HA pairs) with 1,440 drives and 160PB of effective capacity. NAS scale-out clusters support up to 24 nodes (12 HA pairs)

4.3 GPUs

Graphics processing technology has significantly evolved, delivering unique benefits in the world of computing. Designed for parallel processing, Graphics Processing Units (GPUs) are used in a wide range of applications, including graphics and video rendering. While best known for their capabilities in gaming, GPUs are increasingly popular for use in creative production and artificial intelligence. Their massively parallel architecture and high memory bandwidth, especially with the adoption of advanced memory technologies like High Bandwidth Memory (HBM) and HBM3, make them well-suited for handling the intensive computations required in training deep neural networks.

Over time, Central Processing Units (CPUs) and the software libraries that run on them have also evolved to become more capable in deep learning tasks. Through extensive software optimizations and the addition of dedicated AI hardware—such as Intel® Deep Learning Boost (Intel® DL Boost) in the latest Intel® Xeon® Scalable processors—CPU-based systems have enjoyed improvements in deep learning performance. CPUs offer versatility and are effective for certain AI workloads, particularly in inference and in applications involving non-image-based deep learning, such as language processing, text analysis, and time-series data.

While GPUs excel in high-performance training tasks due to their high memory bandwidth and parallel processing capabilities, CPUs can support larger overall memory capacities through system RAM, which can be beneficial for specific workloads that require extensive memory. Additionally, CPUs are often more cost-effective for certain applications and offer simplicity in deployment. For instance, CPU-based systems are generally easier to implement in edge environments where power consumption, cooling, and physical space are critical factors. CPUs typically have lower power requirements and heat generation compared to high-performance GPUs, enhancing their suitability in these contexts.

Moreover, Intel's commitment to developing one API open standard helps ensure maximum code reuse across different stacks and architectures. Tools like OpenVINO™ simplify deep learning inference deployment for hundreds of pre-trained models on CPU-based systems, reducing development time and enhancing efficiency.

In conclusion, while GPUs—especially with advancements like HBM3—play a crucial role in accelerating AI workloads, particularly in training complex models that involve high-resolution imagery and require extensive parallel computations, CPUs also have a significant place in the AI landscape. The choice between GPUs and CPUs should be guided by specific application needs, performance requirements, memory capacity and bandwidth considerations, deployment environments, and cost factors. By leveraging the strengths of both CPUs and GPUs, organizations can build balanced and effective AI solutions tailored to their unique requirements.

4.3.1 NVIDIA GPUs

	NVIDIA L40S	NVIDIA H100 NVL
Recommended Use Case	NVIDIA vWS-Performance Optimized (High end). Compute optimized	Compute optimized
Number of GPUs	1 NVIDIA L40S	1 NVIDIA H100
FP32 Cores / GPU	18,176	16,896
Tensor Cores / GPU	568	528
RT Cores	142	
Total Memory Size / GPU	48GB GDDR6	94GB HBM3
MIG Instances / GPU	-	7
Max GPU Power / GPU	350W	400W
Form Factor	PCIe 4.0 Dual-Slot FHFL	PCIe 5.0 2-Slot FHFL
Board Dimensions	26.67cm x 11.18cm	26.67cm x 11.18cm
Cooling Solution	Passive	Passive

NVIDIA AI Enterprise Licensing

Nvidia AI Enterprise is an end-to-end, cloud-native software platform that accelerates data science pipelines and streamlines development and deployment of production-grade co-pilots and other generative AI applications. Easy-to-use microservices provide optimized model performance with enterprise-grade security, support, and stability to ensure a smooth transition from prototype to production for enterprises that run their businesses on AI.

The NVIDIA AI Enterprise Essentials edition includes both AI and infrastructure software stacks for the development and deployment of production AI. In addition to providing security, reliability, manageability and support, NVIDIA AI Enterprise also includes features only available with subscription such as production branches, long-term support branches, Base Command Manager Essentials, NVIDIA vGPU software and more.

5 Fujitsu Private GPT – Chat with Your Own Data

Private GPT by Fujitsu is a cutting-edge AI solution designed to interact seamlessly with your company's data. This AI-driven platform leverages advanced natural language processing capabilities to deliver precise and relevant information, enhancing your organization's knowledge management and decision-making processes. Built on Fujitsu's robust technology infrastructure, Private GPT ensures high performance, scalability, and security, making it an ideal choice for enterprises looking to optimize their data utilization.

Benefits

1

Custom AI model
Relevant and precise results, as it is optimized for your specific tasks and languages.

2

Sustainable
Low energy costs, as designed for sustainability.

3

Performance
Leverage the latest AI technologies on site, without internet dependency.

4

User-friendly interface
Features a modern, intuitive portal for effortless interaction with the AI model.

5

Data sovereignty and security
Your sensitive data remains protected and under your control.

6

Enhanced knowledge management
Utilize advanced AI to streamline information retrieval and improve access to critical data within your organization.

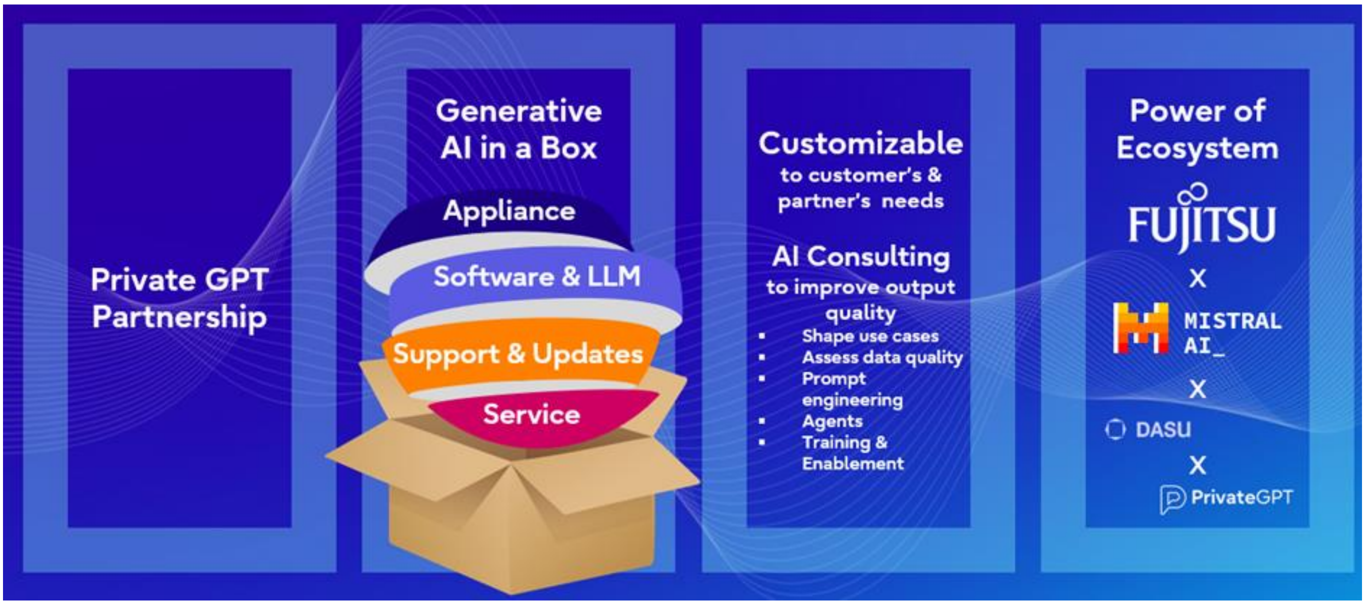
7

Scalability
Designed to grow with your business needs, offering flexible hardware and software options.

8

Continuous improvement
Regular updates and optimizations ensure the system remains at the forefront of AI technology.

Private GPT – What is it?

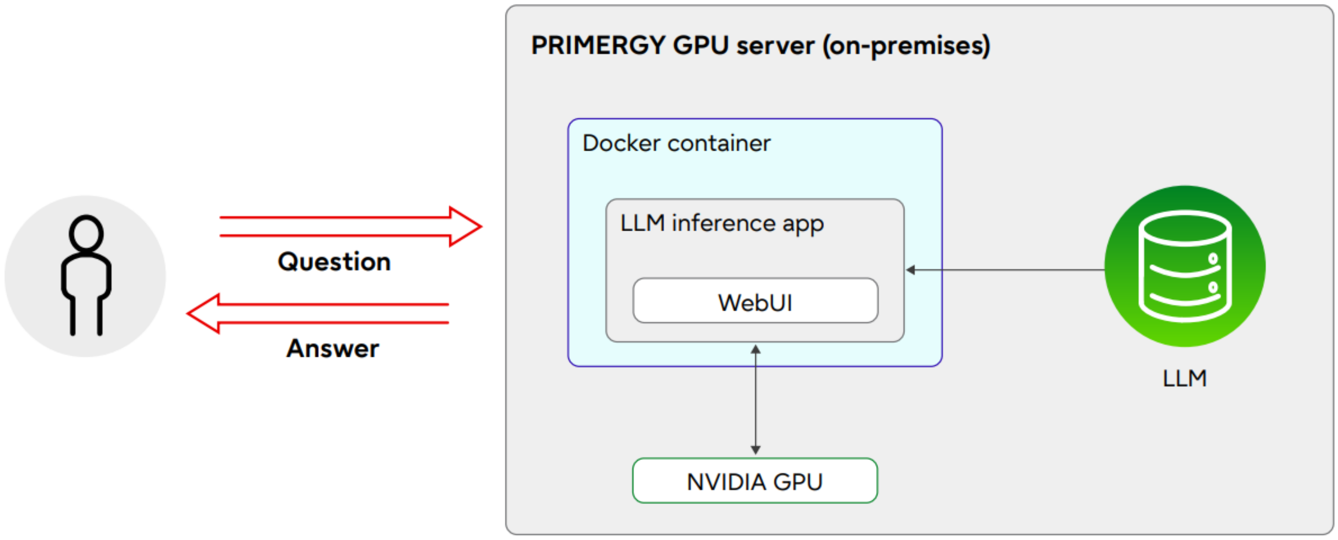


5.1 Prerequisites

On the hardware side a two unit high 19“rack slot to install the PRIMERGY server is needed. An Ethernet connection is required for access to the web portal and SFTP file upload. On the user side a web browser is needed for access the web portal. Also, the network needs to be configured to allow users HTTPS access (port 443) from their working place to the web application and SFTP (2222) for file uploads, SSH (port 4422) for admins respectively.

5.2 Components and subsystems

The solution runs on standard Fujitsu PRIMERGY servers, specifically tuned to the overall requirements. The operating system is the Fujitsu certified SUSE SLES 15 SP6. The Private GPT solution includes a web application in the form of a browser-based chat and administration client, as well as the AI and language subsystem.



The Private GPT solution is based on the Fujitsu PRIMERGY server platforms, offering a highly reliable server environment for business-critical applications. This initial blueprint uses a PRIMERGY RX2540 M7 server with two Intel Xeon Gold 6442Y Processors and one NVIDIA L40S Ada Lovelace GPU. This system’s design is characterized by a balanced performance of GPU and CPU enabling optimal responsiveness for the GenAI engine.

Hardware configuration PRIMERGY RX2540 M7

Item	Description
System	RX2540 M7 8x 2.5 mixed for graphics
CPU	2x Intel® Xeon® Gold 6542Y 24C 2.9 GHz
Memory	16x 16GB (1x16GB) 1Rx8 DDR5-5600 R ECC
OS & Data Storage	
Controller	PRAID EP640
Disks	8x SSD SAS 24G 1.92TB RI 2.5´ Non-/SED H-P
NIC	PLAN EP X710-T2L 2x10GBASE-T
GPU	
Model	NVIDIA L40S
Architecture	Ada Lovelace
Video Memory (VRAM)	48GB GDDR6
Base OS	SUSE SLES 15 SP

Configuration notes

- Operating system and data storage is based on 8x 1.92TB SSDs configured in a highly available RAID5. The minimum configuration offers approximately 12TB of storage capacity for user data.
- Additional capacity up to approx. 46TB can be accommodated in the server.
- One of the 8 disks is reserved as a global hot spare for additional resiliency.
- Network connectivity is supported via 2x 10GbE interfaces (copper)

5.3 Mistral NeMo

Private GPT incorporates the Mistral NeMo Model. Mistral NeMo: the new best small model. A state-of-the-art 12B model with 128k context length, built in collaboration with NVIDIA.

Mistral NeMo beats Meta's Llama 3 8B and Google's Gemma 2 9B in common benchmarks:

Model	Mistral NeMo 12 B	Gemma 2 9B	Llama 3 8B
Context Window	**128k**	8k	8k
HellaSwag (0-shot)	**83.5%**	80.1%	80.6%
Winograd (0-shot)	**76.8%**	74.0%	73.5%
NaturalIQ (5-shot)	**31.2%**	29.8%	28.2%
TriviaQA (5-shot)	**73.8%**	71.3%	61.0%
MMLU (5-shot)	**68.0%**	71.5%	62.3%
OpenBookQA (0-shot)	**60.6%**	50.8%	56.4%
CommonSenseQA (0-shot)	**70.4%**	60.8%	66.7%

HellaSwag - Testing humanlike thinking and common sense

WinoGrande - Testing conclusions and text comprehension

NaturalQ - Testing the ability to answer natural questions

Trivia QA - Testing reading comprehension and answering questions

MMLU - Testing general knowledge and problem-solving skills

OpenBookQA - Testing advanced question-answering

CommonsenseQA - Testing commonsense knowledge

With support for a 128K context length, the model demonstrates improved comprehension and the ability to process large volumes of complex information, resulting in more coherent, accurate, and contextually appropriate outputs. Mistral NeMo is trained on Mistral's proprietary dataset, which contains a significant portion of multilingual and code data, enhancing feature learning, reducing bias, and increasing its capacity to manage diverse and intricate scenarios effectively.

5.4 AI Core

The AI core incorporates a RAG (Retrieval Augmented Generation) engine and a LLM (large language model). Documents provided to the system are preprocessed to extract all text content. The text itself is then divided into smaller sections by a process called chunking. The chunks are then further split up into smaller segments that are then vectorized with the multilingual embedding model BGE-M3 (BGE-M3, kein Datum).

Any question the system receives from a user also undergoes the exact same process, which allows for a vector-based similarity search which is threshold driven and results, via a post processing stage, in the retrieval of the relevant chunks where the best vector matches could be found. These retrieved chunks are then further processed by a reranking mechanism (bge-reranker-v2-m3, kein Datum) that is ranking the chunks regarding their respective content to push chunks by pertinence. Those reranked chunks are then combined with the question along with additional commands and are then sent to the LLM which generates the answer. The commands add influences on the answering behaviour of the LLM to match the desired use case.

5.5 OCR (Optical Character Recognition)

The solution supports OCR (Optical Character Recognition) which is a technology used to identify and extract printed text from digital images of paper documents, such as those generated from scanned files.

5.6 Languages

The Solution can understand documents and give answers in various languages including English, German, French, Spanish, Italian, Bulgarian, Danish, Czech, Spanish, Estonian, Finnish, Hebrew, Hungarian, Indonesian, Polish, Portuguese, Romanian, Russian, Slovenian, Turkish, Ukrainian and many more.

5.7 Uploading information to the system

Custom data is provided by a variety of file formats, including PDF documents, Markdown and Microsoft Word files which can either be uploaded via the web application or by SFTP for bulk uploads. The documents can be assigned to data-groups to achieve data segregation. On the filesystem the documents are separated within folders that are named according to the groups within the web portal for easy handling. The admin user defined within the installation process has SFTP access and is directed to the corresponding folder after login.

5.8 Web application subsystem

The web application is the main user interface for the solution. It is access restricted by username and password. It provides different sections which covers an administrative, maintenance and a chat section. To gain information provided by the system the chat is available for all approved users. All configurations are covered in the administrative section which contains user management, data-groups, and access information regarding Active Directory (AD) and mail server respectively.
(Private GPT – a Fujitsu AI Solution Whitepaper, 2024)

5.9 Conclusion

Private GPT brings state-of-the-art GenAI technology within the private scope of your enterprise, creating an environment where your specific data can be accessed in private and securely by your employees. Since all pruseocessing is done locally, the security of your enterprise data is assured.

6 SUSE Products as Platforms for AI: Harnessing SUSE® Rancher Prime and SUSE® Virtualization

In the rapidly evolving landscape of artificial intelligence (AI), businesses and organizations require robust, scalable, and efficient platforms to support their AI workloads. SUSE, a leader in open-source solutions, offers a range of products that cater to these needs, with SUSE® Rancher Prime and SUSE® Virtualization emerging as pivotal components in this ecosystem. This section of the whitepaper explores how these SUSE products provide a powerful foundation for AI, enhancing deployment, management, and scalability of AI applications. Additionally, it includes a reference architecture to illustrate the optimal integration of these tools within an AI environment.

6.1 SUSE® Rancher Prime: Streamlining Kubernetes Management for AI

Overview

SUSE® Rancher Prime is an open-source Kubernetes management platform designed to simplify the deployment, management, and scaling of containerized applications. As AI workloads often involve complex and resource-intensive processes, SUSE® Rancher Prime's capabilities are particularly advantageous.

Key Features for AI

Multi-Cluster Management

SUSE® Rancher Prime enables centralized management of multiple Kubernetes clusters. For AI deployments, this means organizations can efficiently manage clusters that handle various stages of AI workflows, from data preprocessing to model training and inference.

Scalability

SUSE® Rancher Prime supports horizontal scaling of AI workloads by seamlessly scaling Kubernetes clusters. This elasticity is crucial for handling the dynamic demands of AI processes, such as training deep learning models that require significant computational resources.

Security and Compliance

AI applications often deal with sensitive data. SUSE® Rancher Prime provides robust security features, including role-based access control (RBAC) and integrated security policies, ensuring that data and applications remain secure and compliant with industry standards.

Operational Efficiency

By offering a unified interface for managing Kubernetes clusters, SUSE® Rancher Prime reduces operational overhead and complexity. This efficiency allows AI teams to focus on developing and optimizing models rather than managing infrastructure

Use Case: AI Model Training and Deployment

In a typical AI workflow, training models require substantial computational resources. SUSE® Rancher Prime's ability to manage and scale clusters ensures that resources are allocated dynamically, optimizing performance and cost. Additionally, its integration with popular CI/CD tools facilitates smooth deployment and updates of AI models, enhancing agility and responsiveness.

6.2 SUSE® Virtualization: Simplifying Hyperconverged Infrastructure for AI

Overview

SUSE® Virtualization is a modern hyperconverged infrastructure (HCI) solution built on open-source technologies. It combines compute, storage, and networking into a single platform, leveraging Kubernetes for orchestration. SUSE® Virtualization's design aligns well with the needs of AI applications, offering a versatile and scalable environment.

Key Features for AI

Integrated Storage and Compute

SUSE® Virtualization's HCI approach integrates storage with compute resources, simplifying the management of data-intensive AI workloads. This integration ensures high-performance data access and reduces latency, which is critical for tasks such as large-scale data processing and real-time inference.

Ease of Deployment

SUSE® Virtualization streamlines the deployment of AI infrastructure by providing a unified platform for managing both compute and storage. This ease of deployment accelerates the setup of AI environments, allowing organizations to rapidly prototype and deploy AI solutions.

Resource Optimization

With SUSE® Virtualization, resources are dynamically allocated based on demand. This capability is particularly beneficial for AI workloads that experience variable resource requirements, ensuring that computational and storage resources are used efficiently.

Operational Simplicity

The integrated nature of SUSE® Virtualization reduces the complexity associated with managing separate storage and compute resources. This simplicity is advantageous for AI teams, allowing them to focus on model development and experimentation rather than infrastructure management.

Use Case: AI Data Management and Inference

For AI applications that require large datasets, SUSE® Virtualization's integrated storage solutions enable efficient data management and processing. Its ability to manage both compute and storage resources ensures that data-intensive operations, such as large-scale inference and real-time analytics, are performed with minimal latency and high efficiency.

6.3 Reference Architecture

Overview

The reference architecture presented here illustrates the integration of SUSE® Rancher Prime and SUSE® Virtualization within an AI environment. This architecture is designed to provide scalability, efficiency, and ease of management for AI workloads, leveraging the strengths of both products.

Components

Kubernetes Clusters Managed by SUSE® Rancher Prime

- **AI Development Cluster:** Used for developing and testing AI models. This cluster can be scaled dynamically based on resource demands during model training.
- **Production Inference Cluster:** Dedicated to deploying AI models for real-time inference. This cluster ensures high availability and low latency for production AI applications.
- **Data Processing Cluster:** Handles preprocessing and transformation of large datasets, preparing them for model training and inference.

Security and Monitoring

- **Security Policies**: Managed centrally via SUSE® Rancher Prime, ensuring that all clusters adhere to industry standards and best practices.
- **Monitoring and Logging:** Integrated with both SUSE® Rancher Prime and SUSE® Virtualization, providing real-time insights into system performance, resource utilization, and security events.

Hyperconverged Infrastructure Powered by SUSE® Virtualization

- **Compute Nodes:** Provide the computational power required for AI workloads, including CPUs and GPUs optimized for deep learning.
- **Storage Nodes:** Integrated with compute nodes, these handle the storage of large datasets, model artifacts, and intermediate results. The storage is managed by SUSE® Virtualization's built-in storage management system, ensuring high performance and reliability.
- **Networking Layer:** Facilitates communication between compute and storage nodes, ensuring low latency and high throughput, essential for AI workloads.

CI/CD Pipeline Integration

- **Version Control and CI/CD Tools**: Integrated with SUSE® Rancher Prime, these tools automate the deployment of AI models and infrastructure updates, ensuring continuous delivery and integration of AI solutions.

Workflow

Model Development

Data scientists develop and train models in the AI Development Cluster. The SUSE® Rancher Prime manages the scaling and resource allocation needed for training.

Data Processing

Large datasets are processed in the Data Processing Cluster, utilizing SUSE® Virtualization's storage and compute integration for efficient data handling.

Model Deployment

Once trained, models are deployed to the Production Inference Cluster. The CI/CD pipeline ensures that updates are seamlessly integrated.

Inference and Monitoring

The deployed models handle real-time inference with high availability, while SUSE® Rancher Prime and SUSE® Virtualization monitor performance and manage resources dynamically.

Conclusion

SUSE® Rancher Prime and SUSE® Virtualization provide a powerful and complementary set of tools for AI applications. SUSE® Rancher Prime's Kubernetes management capabilities streamline the deployment and scaling of containerized AI workloads, while SUSE® Virtualization's hyperconverged infrastructure simplifies the management of compute and storage resources. Together, these SUSE products offer a robust platform that enhances the efficiency, scalability, and security of AI operations, positioning organizations to effectively leverage AI technologies in their business strategies. The reference architecture outlined here serves as a guide for deploying these solutions, ensuring that AI initiatives are built on a solid, scalable foundation.

7 AI Customer Use Cases

7.1 Client 1 - AI Supercomputer

Challenge

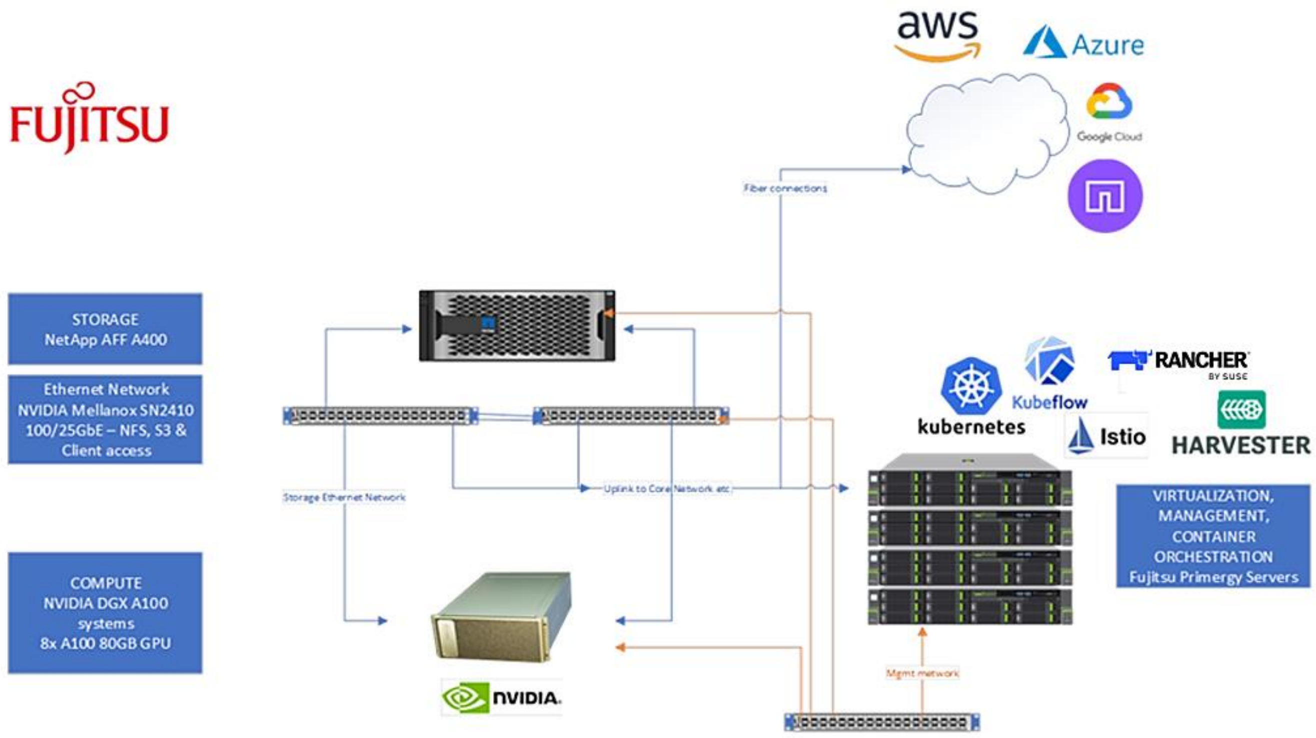
With supercomputing and AI set to play fundamental roles in working and domestic lives, our client wanted to build a powerful and accessible resource to empower both the university and the local economy.

Solution

We worked with our client and a broad range of other suppliers to create a high-power on-premises facility with key AI capabilities and a future-ready infrastructure.

- Powerful supercomputing resource further raises our client's profile
- Local businesses and internal research staff can access AI capabilities
- The new system acts as a template for other supercomputing projects

Architecture



Key features & values

- Based on NVIDIA's & NetApp's reference architecture
    - Verified compatibility, ready-made configurations, designs and best practices.
    - Simple whole environment level support
- Each component is scalable – modular architecture
    - Supports large data models, generative AI use cases e.g., ChatGPT, Private GPT
- Automation & support for container based applications.
- NVIDIA's software – NVIDIA GPU Cloud (NGC), DGX experts and support for tools with enablement consultants
- Futureproof

Fujitsu PRIMERGY servers

- 4x Fujitsu PRIMERGY servers
- Runs CPU based AI workloads, virtualization platform and management and operational applications
- Kubernetes cluster and master node

NVIDIA DGX A100

- 8x NVIDIA A100 80GB GPUs
    - GPUs clustered, total RAM 640GB
    - MIG technology – slices one GPU to 7 parts
    - Accelerates the demanding computational tasks
- AMD CPU with total of 128 cores.
- System memory 2TB.
- High speed 200Gb/s networking for clustering - scaling & storage connections
- NVMe drives for OS (Linux) and 30TB internal NVMe storage
- Running Linux OS
- Performance: 5petaFLOPS (AI), 10 petaFLOPS INT8

NetApp AFF A400

- NetApp All Flash storage with dual controllers and 18x 3.8TB NVMe drives
- Protocols: NFS, S3, SMB, NVMe/FC & TCP, FC & iSCSI
- Inline features to shrink data footprint: deduplication, compression & compaction
- Rich data management features: snapshots, cloning, replication, cold data tiering and support for data science workflows / tools
- Public Cloud Integrations

NVIDIA Networking

- NVIDIA (Mellanox) 100GbE & 25GbE high speed ethernet switches
- Runs Cumulus Linux OS which allows automation, customization and scalability using web-scale principles
- NVIDIA management switch
- Fortinet firewall

Computing power comparison

- Middle range server with 2 CPUs: 1500 GFLOPS = 0,0015 PFLOPS
- Laptop with 12th GEN i5: 300 FLOPS = 0,0003 PFLOPS
- NVIDIA DGX A100: 5 PFLOPS

NVIDIA DGX A100 = 16 666 laptops or 3 333 servers

- Fugaku, Fujitsu's supercomputer: 2150 PFLOPS
- Lumi, CSC's supercomputer: 375 PFLOPS

(Xamk Hippu makes AI supercomputing accessible, 2024)

7.2 Client 2 - Knowledge Centralization

The primary objective of this initiative is to transform access to internal knowledge repositories by leveraging a Large Language Model (LLM), deployed locally. By implementing an LLM-driven interface, we aim to significantly enhance workforce productivity by enabling faster, more intuitive access to internal resources stored in knowledge databases. This will empower employees to efficiently locate and utilize relevant information, minimizing redundant queries and maximizing operational efficiency.

In the current setup, a diverse range of roles—including consultants, project managers, quality assurance (QA) personnel, documentation experts, partners, and pre-sales colleagues—depend on tools like Confluence and Jira as integral components of their daily routines. These platforms host vast amounts of knowledge and documentation crucial for ongoing projects, compliance, troubleshooting, and customer support. Meanwhile, Large Language Models have demonstrated notable effectiveness in enhancing business workflows and knowledge management, as illustrated by the widespread use of popular GenAI tools. The popularity of these AI tools highlights the benefits of natural language processing in knowledge retrieval, particularly for non-technical users. Deploying a specialized LLM tailored to our internal knowledge base would allow employees to harness these capabilities, facilitating a more streamlined, accessible, and user-friendly approach to locating information. This approach would leverage familiar, conversational interfaces to improve day-to-day workflows across departments.

Challenges

Despite the critical role of Confluence and Jira, employees face challenges when trying to quickly locate precise information. Given the extensive and sometimes redundant or overlapping information in these repositories, effective searching often requires time, experience, or advanced skills in crafting complex query strings. Consequently, newer employees frequently seek help from veteran team members, leading to a high dependency on these individuals to locate specific knowledge or historical data, thus creating a bottleneck. To address this, the company has already experimented with solutions like meta-directories and navigational guides. While useful, these resources further illustrate the gap in efficient information retrieval and underscore the need for a more powerful, adaptable search solution. The goal is to reduce dependency on “knowledge holders” and eliminate the need for repetitive, manual guidance, which can be costly and time-consuming.

Solution

To tackle these challenges, we propose the deployment of a local Large Language Model that would be trained on and integrated with the existing Confluence and Jira knowledge databases. This model would operate within the company’s secure IT environment, ensuring all data remains internal and intellectual property (IP) is protected. The LLM would be accessible to employees via a user-friendly, chatbot-style web interface, designed to emulate the conversational ease of consumer-facing AI solutions. This setup would allow employees to enter natural language queries and be directed to the most relevant information or documents quickly and accurately. Users would have options to activate or deactivate specific document sets based on contextual needs, allowing them to narrow their search results and improve output relevance. Ultimately, this solution would reduce the time employees spend searching for information, alleviate the demand on experienced personnel, and streamline knowledge sharing across the organization.

Key Architecture Components

- Private GPT Software
- PY RX2540 M7 8x 2.5 mixed for graphics.
- Intel Xeon Gold
- NVIDIA L40S

7.3 Client 3 - Machine Vison System to Detect Electric Vehicles

Challenge

The European motorway system is set to implement new regulations that will demand enhanced surveillance of electric vehicles (EVs). As part of these changes, e-charging stations will need to be strategically positioned, with no more than 60 km between each, to accommodate EV traffic. A comprehensive analysis of EV mobility patterns is necessary to optimize infrastructure placement based on real-time data of travel behaviors across the European motorway network. Large amounts of data must be processed from the live streams of cameras all over the motorways, which need huge amounts of computing power.

Solution

Leveraging machine learning and predictive analytics, Fujitsu can streamline this process, offering automated insights into EV distribution and usage patterns. Together with Brainpool, Fujitsu developed a machine vision system that can identify electric vehicles in large datasets of CCTV data from cameras placed on European motorways.

1. **Outperforming Humans**: During the creation of the benchmarking. It was near impossible for a human to detect the green through the cameras, detection was made on certain models however the AI had a much higher success rate.
2. **Good Car Detection**: During our human benchmarking 158 cars were detected. The model detected and maintained detection 164.

The project is in progress with an objective to optimize the allocation of charging stations for electric vehicles. Data does not leave the customer data center and personalized data is deleted according to data protection regulations.

Architecture

The end-to-end solution comprises:

- License plate recognition system developed by Brainpool AI
- Servers by Fujitsu: PRIMERGY RX 2540 M7
- Cameras: AXIS Q1700-LE License Plate Camera

8 Sizing

Correct sizing of the AI infrastructure, which implies the servers, network and storage that make up the infrastructure, is needed when using an LLM for inferencing due to the computational demands and high memory requirements it places on the subsystem. Other unique characteristics of these models influence the size of a server needed to extract the correct level of performance. Network and storage requirements must also match the expected data volumes and concurrency that the end-user environment will generate It is essential to size the server appropriately to enable optimal operation of the model and to avoid operational issues such as out-of-memory problems or latency issues.

8.1 Sizing for the LLM

Deploying Large Language Models (LLMs) in data centers necessitates careful consideration of hardware resources, specifically GPUs capable of handling the substantial memory and compute demands of large models. This white paper provides guidance on selecting appropriate data center GPUs for LLM inference tasks, focusing on large models that require enterprise-grade solutions. It integrates critical insights on performance verification, memory requirements, service quality parameters, and the benefits of optimized inference engines like NVIDIA TensorRT-LLM and vLLM. Detailed calculations and formulas are included to aid in informed decision-making.

Proper infrastructure sizing for an LLM ensures optimal performance, scalability, and user experience while effectively managing operational costs.
To understand the performance of a large language model (LLM), you can measure several metrics, each of which is relevant for assessing various aspects of the model's performance. Each metric contributes to understanding LLM performance as follows:

- **First token latency:** This metric, also known as time to first token (TTFT), measures the time it takes for the model to generate the first token of a response after receiving an input prompt. It reflects the initial processing time of the model and can be important for real-time applications where low latency is crucial.
- **Tokens per second:** This metric measures the throughput of the model, indicating how many tokens (words or characters) the model can generate per second on average. It provides insight into the overall speed of the model and its ability to process input data efficiently.
- **Overall response latency:** Unlike first token latency, this metric measures the total time it takes for the model to generate a complete response to a given input prompt. It includes the time for processing the entire input sequence and generating the output sequence. It is crucial for assessing the end-to-end latency experienced by users interacting with the model.
- **Number of concurrent users:** This metric measures the model's ability to handle multiple simultaneous requests or users. It helps determine the scalability and resource requirements of deploying the model in production environments with varying levels of user concurrency.

8.1.1 Understanding the Demands of Large LLMs

Large LLMs, such as those with tens to hundreds of billions of parameters, have transformed natural language processing tasks. However, their deployment poses significant challenges:

- **Massive Memory Requirements**: Storing model weights and managing key-value (KV) caching during inference demand substantial GPU VRAM.
- **Compute Performance Needs:** Efficient inference requires high computational capabilities to handle complex operations, especially in the autoregressive decode phase.
- **Scalability Considerations:** Supporting numerous concurrent requests while maintaining low latency is critical.
- **Inference Engine Optimization:** Utilizing optimized inference engines enhances performance and resource utilization.
- **Power Consumption Management:** Modern LLM deployments require careful power planning based on GPU tier selection:
- **Mid-Range Enterprise GPUs:**
  Base Power Draw: 200-300W per accelerator at peak load
  Cooling Requirements: Additional 15-25% power allocation for cooling systems
  Workload Efficiency: Optimized for inference tasks with moderate batch sizes
  Ideal for: Medium-scale inference workloads and development environments
- **High-End Data Center GPUs:**
  Base Power Draw: 300-700W per accelerator at peak load
  Cooling Requirements: Additional 25-40% power allocation for cooling infrastructure
  Workload Efficiency: Designed for maximum throughput and large batch processing
  Ideal for: Large-scale production deployments and training workloads

8.1.2 Key Factors in GPU Selection for LLM Inference

1. **Memory Capacity (VRAM)**
   - **Model Size:** Models with 70B parameters or more require significant VRAM (typically 48 GB or more)
   - **Precision Levels:** FP16 or BF16 precisions are preferred for their balance of performance and accuracy
   - **KV Cache Management:** Memory usage scales with sequence length (L) and batch size (B)
2. **Compute Performance**
   - **High Throughput:** GPUs must handle extensive computations efficiently.
   - **Precision Support:** Efficient processing of FP16 and BF16 without compromising performance.
   - **Hardware Optimization:** GPUs like NVIDIA H100 are designed to maximize performance with optimized inference engines
3. **Memory Bandwidth**
   - **Data Transfer Rates:** High bandwidth reduces inference latency
4. **Scalability and Interconnect**
   - **Multi-GPU Support:** Necessary for models exceeding single GPU capacity
   - **Futureproofing:** GPUs should accommodate growing model sizes and user demand
5. **Inference Engine Optimization**
   - **NVIDIA TensorRT-LLM:** Optimized for NVIDIA GPUs to accelerate LLM inference
   - **vLLM Inference Engine:** Introduces Paged Attention for improved throughput and memory efficiency
6. **Service Quality Parameters**
   - Throughput and Latency
   - Concurrent Requests
   - Input/Output Tokens

8.1.3 Memory Requirements of Large LLMs

Model Memory Requirements

Model Memory Requirements loaded huggingface safetensors.

Model Name	File Size FP32 (GB)	GPU MemoryFP32 (GB)	GPU Memory FP16 (GB)
Mixtral-8×7B-Instruct-v0.1	~ 91.2	~ 93.4	~ 46,7
Mixtral-8×22B-Instruct-v0.1	~ 274.7	~ 273.4	~ 136,7
Meta-Llama-3-8B-Instruct	~ 32.4	~ 33,6	~ 16,8
Meta-Llama-3-70B-Instruct	137.8	135.7	~ 67,85

Notes

- **FP16 Conversion:** Memory requirements are approximately halved when using FP16 precision (e.g., Mixtral-8x7B requires ~47GB in FP16 vs 93.4GB in FP32)
- **FP16 Precision:** Preferred for balance between performance and accuracy
- **File Size:** Storage size on disk
- **GPU Memory:** VRAM required during inference

Insights

- Models like Meta-Llama-3-70B-Instruct require approximately 135.7 GB of GPU memory
- Multi-GPU configurations are necessary for models exceeding a single GPU capacity

8.1.4 Implications for GPU Selection

1. **Memory Capacity Considerations**

   Large Models

   - Require substantial GPU memory (over 135 GB)
   - GPUs with at least 80 GB VRAM (e.g., NVIDIA H100) are necessary
   - Multi-GPU configurations may be required

2. **Preference for FP16/BF16 Precision**

   - **Data Center Standards:** Preferred for computational efficiency and accuracy
   - **Hardware Support:** Modern GPUs are optimized for these precisions.

3. **Inference Engine Optimization**

   - **NVIDIA TensorRT-LLM:** Optimized to leverage advanced capabilities of NVIDIA GPUs like the H100
   - **vLLM Inference Engine:** Improves throughput and memory efficiency with PagedAttention

8.1.5 Service Quality Considerations

Importance of TTFT and Latency

- **User Experience:** Critical for perceived service quality
- **Performance Targets:** Low latency and high concurrency

Example Service Requirement Calculation

- **Objective:** Support 100 concurrent users with 600 words per minute each

Calculations

1. **Words per Second per User**

$$\frac{\frac{600\ words}{min}}{\frac{60\ sec}{min}} = \frac{10\ words}{sec}$$

2. **Tokens per Second per User (assuming ~2 tokens per word)**

$$\frac{10\ words}{sec} \times \frac{2\ tokens}{words} = \frac{20\ tokens}{sec}$$

3. **Total Tokens per Second**

$$\frac{20\ tokens/_{sec}}{users} \times 100\ users = 2000\ {}^{tokens}/_{sec}$$

Hardware Implications: GPU(s) must handle 2,000 tokens/sec with low latency.

8.1.6 Challenges in Sizing and Performance Verification

- **Model and Library Variations:** Significant performance and memory usage differences
- **Necessity of Testing:** Verify performance for each model and software combination
- **Inference Engine Impact:** Different engines affect performance and resource utilization

8.1.7 GPU Options for Large LLM Deployment

NVIDIA H100 PCI

- **VRAM:** 80 GB HBM3 300W – 350W, (SXM up to 700W)
- **Compute Performance:** Hopper architecture optimized for FP16/BF16
- **Memory Bandwidth:** Ultra-high, reducing latency and power consumption
- **Optimization with TensorRT-LLM:** Maximizes performance on H100 GPUs
- **Ideal Use Cases:** Large-scale deployments requiring maximum performance and accuracy

NVIDIA L40S

- **VRAM:** 48 GB GDDR6 300W - 350W
- **Compute Performance:** Ada Lovelace architecture, optimized for AI inference
- **Memory Bandwidth:** High, suitable for data-intensive operations
- **Ideal Use Cases**
  - Medium to large model deployments
  - Balance between performance and efficiency

NVIDIA RTX 6000 Ada Generation

- **VRAM:** 48 GB GDDR6
- **Compute Performance:** High-performance for professional workloads
- **Ideal Use Cases**
  - Development and testing environments
  - Scenarios not requiring ultra-high-end GPUs

8.1.8 Optimization Strategies for LLM Deployment

1. **FP16/BF16 Precision Optimization**

- **Benefits:** Balances speed and accuracy
- **Hardware Alignment:** GPUs like NVIDIA H100 excel with these precisions

2. **Efficient Inference Engines**

- **TensorRT-LLM**
  - Optimizes inference on NVIDIA GPUs
  - Enhances throughput and reduces latency
- **vLLM**

PagedAttention Mechanism

- **Reduces memory overhead**
- **Manages attention states efficiently**
- **Benefits**
  - High throughput
  - Memory efficiency
  - Scalability

3. **Memory Optimization Techniques**

- **KV Cache Management**

Formula

KV Cache Size (bytes) = 2 × *B* × *L* × *H* × Precision (bytes)

B: Batch size
L: Sequence length
H: Hidden size

Example

B = 100 *L* = 512 *H* = 8192, Precision = 2 bytes (FP16)
KV Cache Size ≈ 1.56 GiB

Implications

- Manageable with optimized B and L
- Inference engines like vLLM optimize memory usage

4. **Efficient Batching Strategies**

- **Dynamic Batching:** Optimizes throughput
- **Concurrency Management:** Maximizes utilization without compromising latency

8.1.9 Performance Verification and Testing

Critical Need: Test under realistic workloads

Key Metrics

- Throughput
- Latency
- Resource Utilization

Testing Scenarios

- Varying batch sizes
- Sequence length variations.
- Different concurrency levels

Inference Engine Evaluation

- Compare TensorRT-LLM and vLLM performance

8.1.10 Guidelines for Selecting GPUs in Data Centers

Define Service Level Objectives

- Set targets for throughput, latency, and concurrency

Assess Model Requirements

- **Memory Needs:** Include model size and KV cache
- **Compute Needs:** Required performance levels
- **Inference Engine Compatibility:** Ensure GPU support

Select Appropriate GPUs

- **Memory Capacity:** GPUs like NVIDIA H100
- **Compute Capability:** Optimized for FP16/BF16 and compatible with inference engines
- **Scalability:** Support multi-GPU configurations

Consider Infrastructure Compatibility

- **Server Support:** Power, cooling, space
- **Interconnects:** Availability of NVLink
- **Software Ecosystem:** Compatibility with TensorRT-LLM, vLLM, and AI frameworks

Performance Verification

- **Test Configurations:** With selected GPUs and inference engines
- **Monitor Metrics:** Adjust based on results

Plan for Future Growth

- **Scalability:** Hardware and engines accommodating increasing demands
- **Flexibility:** Support for various models and updates
- **Budget Alignment:** Balance performance and costs

8.1.11 Conclusion

Deploying large LLMs in data centers requires GPUs with substantial memory and compute capabilities, optimized for FP16/BF16 inference. Successful deployment hinges on:

- **Strategic Planning:** Aligning requirements with hardware and inference engines
- **Inference Engine Selection:** Leveraging optimized engines like TensorRT-LLM and vLLM
- **Performance Verification:** Rigorous testing to meet objectives
- **Scalability:** Preparing for growth in model size and demand

By focusing on data center-grade solutions like NVIDIA H100, NVIDIA L40S, and NVIDIA RTX 6000 Ada Generation GPUs, and utilizing advanced inference engines, organizations can deliver high-quality AI services that meet user expectations and business goals.

8.2 Storage System Sizing Guidelines

NetApp provides a variety of storage platforms that are appropriate for use with AI LLM's and our Private GPT solution. They have successfully tested these systems using NVIDIA's DGX BasePOD, and it was found that two A900 HA pairs (i.e., four A900's) can easily support a cluster of eight DGX H100 systems.

For larger deployments with higher storage performance requirements, additional AFF systems can be added to the NetApp ONTAP cluster using up to 12 HA pairs (24 nodes) in a single cluster.

By using the FlexGroup technology, a 24-node cluster can provide over 40 PB and up to 300 GB/s throughput in a single namespace. With smaller or less demanding deployments, other NetApp storage systems such as the AFF A400, A250 and C800 can be used at a lower cost point.
Customers could even start with a smaller initial footprint and add more or larger storage systems to the cluster as capacity and performance requirements grow since ONTAP 9 supports mixed model clusters.

The table below shows a rough estimate of the number of A100, and H100 GPUs supported for each AFF subsystem.

NetApp storage system sizing guidance

NetApp Model	# of units	Throughput	Raw Capacity (typical / max.)	Connectivity	# of NVIDIA A100 supported	# of NVIDIA H100 supported
AFF A900	1 HA pair	28GB/s	182TB/ 14.7PB	100GbE	1 – 64	1 – 32
	12 HA pairs	336GB/s	2.1PB / 176.4PB		768	384
AFF A800	1 HA pair	25GB/s	368TB / 3.6PB	100 GbE	1 – 64	1 – 32
	12 HA pairs	300 GB/s	4.4PB / 43.2PB		768	384
AFF C800	1 HA pair	21 GB/s	368TB / 3.6PB	100GbE	1 – 48	1 – 24
	12 HA pairs	252GB/s	4.4PB / 43.2PB		576	288
AFF A400	1 HA pair	11GB/s	182TB / 14.7PB	40/100GbE	1 – 32	1 – 16
	12 HA pairs	132GB/s	2.1PB / 176.4PB		384	192
AFF C400	1 HA pair	8GB/s	182TB / 14.7PB	40/100GbE	1 – 16	1 – 8
	12 HA pairs	128GB/s	2.1PB / 176.4PB		192	96
AFF A250	1 HA pair	7.4GB/s	91.2TB / 4.4PB	25GbE 40/100GbE	1 – 16	1 – 8
	4 HA pairs	29.6GB/s	364.8TB / 17.6PB		64	32
AFF C250	1 HA pair	5GB/s	91.2TB / 4.4PB	25GbE 40/100GbE	1 – 8	1 – 4
	4 HA pairs	20GB/s	364.8TB / 17.6PB		32	8

9 LLM Inferencing Benchmarks

To understand the performance of a large language model there are several relevant metrics to be measured for assessing how a model might perform in various scenarios. These metrics were already described in detail in the Sizing chapter, namely the first token latency, tokens per second, response latency and the number of concurrent users. In this chapter the following use case scenarios will be highlighted:

Q&A Bot

A system designed to answer questions posed by users in natural language. It leverages natural language processing and machine learning algorithms to understand and interpret user queries, retrieving relevant information from a database or the internet to provide accurate and concise answers. They are commonly used to help find information on a specific topic.

Chatbot

An application that simulates human conversation through text or voice interactions. It is designed to engage users in dialogue, aiding, information, or entertainment. Chatbots can keep track of questions asked and engage in long running conversations. They are used in more generic contexts and not only for retrieving information about specific topics.

Document Summarization

Document Summarization is an AI technique that automatically generates a concise and coherent summary of a larger text document. It involves extracting or abstracting key information from the original content while preserving its main ideas and context. This technology is useful for quickly understanding large volumes of text without having to read through all of it oneself.

Document Proofing

Document Proofing involves using AI to review and correct written content for errors in grammar, spelling, punctuation, and style. Advanced proofing tools can also provide suggestions for improving clarity, coherence, and tone. By leveraging NLP and machine learning, these tools help writers produce polished and professional documents.

9.1 Test Setup

For the evaluation a general purpose PRIMERGY RX2540 M7 server with the following configuration was used:

Server	PRIMERGY RX2540 M7
CPU	2x Intel® Gold 6430 (32 core / 64 thread)
Memory	128 GB (DDR5 4800 RDIMM)
GPU 1	NVIDIA L40s (48GB GDDR6 VRAM)
GPU 2	NVIDIA L40s (48GB GDDR6 VRAM)
Operating System	Ubuntu 22.04.4 LTS
GPU Driver	535.161.08 (CUDA Version 12.2)

The benchmark tests were performed using the open-source tool LLMPerf, which is a library that allows to measure the throughput, request latency and other metrics of large language models by creating load tests and stressing the LLM's API with it. The following parameters were used and continuously adjusted during the benchmarks:

- Number of concurrent users → simulate multiple users utilizing the LLM
- Input token size → specifies, how much data the LLM must process
- Output token size → the amount of data to be generated by the LLM as a result

9.2 Benchmark Results

In this section the performance values for each of the different benchmark scenarios are shown. For each scenario the chosen parameters are presented before listing the results that were achieved in relation to the amount of concurrent user requests.

9.2.1 Q&A Bot

- Input tokens: 100
- Output tokens: 800

Q&A Bot LLM Performance – Mistral-7B-Instruct

	1x NVIDIA L40S			2x NVIDIA L40S		
Concurrent Users	TTFT (sec)	Latency (sec)	Token/sec	TTFT (sec)	Latency (sec)	Token/sec
10	0.6013	22.5486	35.4539	0.7691	21.0621	38.0898
50	0.5075	33.2351	24.1126	0.8141	31.5363	25.4225
100	0.4871	42.1174	19.0707	0.8048	41.4587	19.3907
200	0.5450	57.1577	14.2434	0.9449	53.6258	15.2670
400	0.8495	84.1070	10.0764	1.7968	75.7539	11.2384
800	13.8388	110.4049	7.8243	14.4487	108.2538	8.2865
1,200	35.1761	137.9730	6.5131	31.9969	133.3078	7.0495

Q&A Bot LLM Performance – Llama-3-8B-Instruct

	1x NVIDIA L40S			2x NVIDIA L40S		
Concurrent Users	TTFT (sec)	Latency (sec)	Token/sec	TTFT (sec)	Latency (sec)	Token/sec
10	0.5548	24.0407	33.3059	0.5695	22.5841	35.4196
50	0.5842	36.7061	21.8310	0.6303	35.1140	22.8305
100	0.6199	45.1766	17.9050	0.6767	44.0070	18.4869
200	0.6558	54.8835	14.9107	0.8157	57.6095	14.3925
400	0.6862	64.5438	12.8138	0.9452	77.5390	10.9799
800	0.7295	79.4346	10.5047	1.2036	107.2801	8.3553
1,200	0.8100	88.3061	9.4577	2.5602	111.2049	8.1115

9.2.2 Chatbot

- Input tokens: 100
- Output tokens: 200

Chatbot LLM Performance – Mistral-7B-Instruct

	1x NVIDIA L40S			2x NVIDIA L40S		
Concurrent Users	TTFT (sec)	Latency (sec)	Token/sec	TTFT (sec)	Latency (sec)	Token/sec
10	0.6843	5.8571	34.3962	0.7358	5.9121	34.0677
50	0.4819	7.0194	28.7931	0.5913	6.3307	32.0122
100	0.4445	7.2147	27.9966	0.6209	6.5211	31.0180
200	0.4251	7.4039	27.2512	0.6435	6.4774	31.2558
400	0.4512	7.4935	26.9688	0.6835	6.6235	30.7120
800	0.4979	7.6071	26.6923	0.7708	6.7838	30.2541
1,200	0.5764	7.8001	26.2533	1.1117	7.1468	29.6631

Chatbot LLM Performance – Llama-3-8B-Instruct

	1x NVIDIA L40S			2x NVIDIA L40S		
Concurrent Users	TTFT (sec)	Latency (sec)	Token/sec	TTFT (sec)	Latency (sec)	Token/sec
10	0.5423	6.2844	31.8343	0.5516	5.4273	36.8940
50	0.5625	6.8164	29.3691	0.5702	6.2309	32.3239
100	0.5435	6.8717	29.1274	0.5938	6.7134	30.0541
200	0.5587	6.9139	28.9409	0.5853	6.5525	30.7962
400	0.5505	6.9664	28.7189	0.6015	6.7140	30.1754
800	0.5718	7.0291	28.5171	0.6076	6.6100	30.6303
1,200	0.5520	7.0094	28.5713	0.6152	6.7322	29.9988

9.2.3 Document Summarization

- Input tokens: 1200
- Output tokens: 200

Document Summary LLM Performance – Mistral-7B-Instruct

	1x NVIDIA L40S			2x NVIDIA L40S		
Concurrent Users	TTFT (sec)	Latency (sec)	Token/sec	TTFT (sec)	Latency (sec)	Token/sec
10	0.6090	6.5367	30.8271	0.7554	6.2315	32.4518
50	0.6100	9.3125	21.8812	0.7848	9.7469	21.0632
100	0.5608	10.0009	20.3491	0.7719	10.5779	19.5148
200	0.6161	10.3798	19.5540	0.7466	10.3605	20.0128
400	0.6404	10.6794	18.9418	0.7827	11.3482	18.1310
800	0.5399	10.6872	18.9459	0.8138	11.2635	18.2764
1,200	0.5343	10.7517	18.8774	0.8222	10.9743	18.8750

Document Summary LLM Performance – Llama-3-8B-Instruct

	1x NVIDIA L40S			2x NVIDIA L40S		
Concurrent Users	TTFT (sec)	Latency (sec)	Token/sec	TTFT (sec)	Latency (sec)	Token/sec
10	0.5874	7.0545	28.3924	0.6245	6.0744	33.1383
50	0.5937	8.1416	24.6482	0.6563	7.9980	25.2680
100	0.5972	8.2913	24.1723	0.6614	8.3865	24.0538
200	0.5935	8.3508	23.9768	0.6628	8.3542	24.0700
400	0.5975	8.3887	23.8616	0.6651	8.5062	23.6040
800	0.5944	8.4726	23.6257	0.6707	8.4265	23.7952
1,200	0.5972	8.4483	23.7172	0.6711	8.1076	24.8332

9.2.4 Document Proofing

- Input tokens: 1200
- Output tokens: 1200

Document Proofing LLM Performance – Mistral-7B-Instruct

	1x NVIDIA L40S			2x NVIDIA L40S		
Concurrent Users	TTFT (sec)	Latency (sec)	Token/sec	TTFT (sec)	Latency (sec)	Token/sec
10	0.6211	37.1784	32.3156	0.7424	35.2583	34.0811
50	0.6636	66.7076	18.0136	0.7792	62.5281	19.2236
100	0.5851	98.4089	12.2266	0.8176	93.0574	12.9370
200	29.0748	144.2303	8.7612	0.8194	163.7618	7.3731
400	100.9385	227.1071	6.2448	39.5071	260.8938	4.7402
800	258.2552	394.4962	4.1715	152.9416	414.6994	3.3844
1,200	409.5277	549.2678	3.2777	294.6056	567.5714	2.6806

Document Proofing LLM Performance – Llama-3-8B-Instruct

	1x NVIDIA L40S			2x NVIDIA L40S		
Concurrent Users	TTFT (sec)	Latency (sec)	Token/sec	TTFT (sec)	Latency (sec)	Token/sec
10	0.5945	39.7526	30.1886	0.6424	37.2655	32.2109
50	0.7126	71.7789	16.7362	0.6961	66.1605	18.1676
100	3.5166	104.1159	11.5816	0.8091	99.8628	12.0689
200	34.6360	147.7788	8.5907	1.0002	168.4184	7.3117
400	102.2531	224.1502	6.2701	36.2983	282.2755	4.5971
800	237.1773	371.2095	4.3254	156.2622	467.1519	3.1407
1,200	382.1296	518.6529	3.3922	301.0651	633.4619	2.4802

10 Fujitsu and AI Ethics

Artificial intelligence businesses are becoming more active in every industry around the world, partly due to the rapid advancement of generative AI technology. AI is also being used to make important decisions. However, using AI without understanding its potential negative impact on stakeholder values and characteristics can induce risks. Many irresponsible AI services have caused widespread harm and abuse in society. This has led to a breakdown of social trust in AI. To rebuild this social trust, and to prevent further harm, many countries are considering regulation of AI deployment and use itself. AI must be safe, secure, and trustworthy to maximize its value for the benefit of society. And it is important for all AI stakeholders to practice AI Ethics. AI Ethics must now be practiced not only by some companies, but by society, including developers, providers, and users. AI Ethics is not intended to regulate technology. It fosters the creation of technological solutions that promote safety.

10.1 Why Do We Need AI Ethics?

1. For a Sustainable Society

Artificial intelligence (AI) businesses are becoming more active in every industry around the world, partly due to the rapid advancement of generative AI technology. AI is also being used to make important decisions. However, using AI without understanding its potential negative impact on stakeholder values and characteristics can induce risks. Many irresponsible AI services have caused widespread harm and abuse in society. This has led to a breakdown of social trust in AI. To rebuild this social trust, and to prevent further harm, many countries are considering regulation of AI deployment and use itself.
AI must be safe, secure, and trustworthy to maximize its value for the benefit of society. And it is important for all AI stakeholders to practice AI Ethics. AI Ethics must now be practiced not only by some companies, but by society, including developers, providers, and users. (Fujitsu AI Ethics, n.d.)

AI Ethics is not intended to regulate technology. It fosters the creation of technological solutions that promote safety. Practicing AI Ethics contributes to gaining trust from users, resulting in taking advantage of business. To maximize the value of useful AI that brings innovation and to continue supporting a sustainable society, Fujitsu promotes “AI Ethics and Governance” for the implementation of AI Ethics.

2. AI Ethics: The Fujitsu Perspective

The relationship between business interests and AI Ethics and how to implement the same might remain yet unclear to many companies considering an AI solution. At Fujitsu, our approach to AI Ethics has been advocating for human-centric technology generation processes. We have been working on AI Ethics from this perspective, and accumulating know-how from an early stage – all to build trust in society about the safe and responsible use of AI.

Here are four key points of the AI Ethics works we are promoting together with our users and other stakeholders to realize safe, secure, and trustworthy AI.

History

Advocating for the "Human-centric" perspective for over 10 years

Since 2009, Fujitsu has been promoting the philosophy of "Human Centric", which aims to realize a sustainable digital society centred on the rights and agencies of people, and which utilizes technology for the benefit of people.

Principle

Ensuring Objectivity Through Cooperation with External Experts

"Fujitsu Group AI Commitment" was developed based on the bioethical principles and referring to the 5 principles designed by AI4People. We also actively participate in discussions on AI Ethics around the world, including the Global Partnership on AI and standardizing bodies such as ISO.

Organization

Ensuring Governance on AI Ethics Involving Management and All Employees

Under the leadership of "AI Ethics and Governance Office", departments in R&D, internal compliance, and user implementation work together to establish and maintain the governance system.

The "Fujitsu Group External Advisory Committee on AI Ethics", which consists of outside experts, objectively evaluates Fujitsu's activities on AI Ethics and the results of the discussion shall be reported to the Board of Directors to link AI Ethics to corporate governance.

Practice

Embedding Principles into Practices

We are working to promote AI Ethics practices in both aspects of governance and technology such as to develop guidelines in each business group and to establish the in-house consultation desk and implement ethical review process to quickly identify and respond to ethical risks, as well as R&D that leads to realize AI Ethics.

3. Fujitsu Group AI Commitment

Our Promises with the Customers and the Society

Fujitsu has long advocated a 'human centric' approach and argued that information technology should fundamentally be used to focus on the needs and value of, and for the benefit of, people. In March 2019, as a reflection of the rapid recent development of AI technologies, Fujitsu formulated and announced the "Fujitsu Group AI Commitment" which is our purpose detailed from an AI Ethics perspective. As the companies which carry out AI-related businesses including research, development, implementation, and operation activities, we aim to emphasize the importance of communication with a wide range of stakeholders in the community, including users and consumers, as we distribute the enormous value of AI to the society. For that purpose, the commitment outlines our promises with the customers and the community.

5 Principals of "Fujitsu Group AI Commitment"

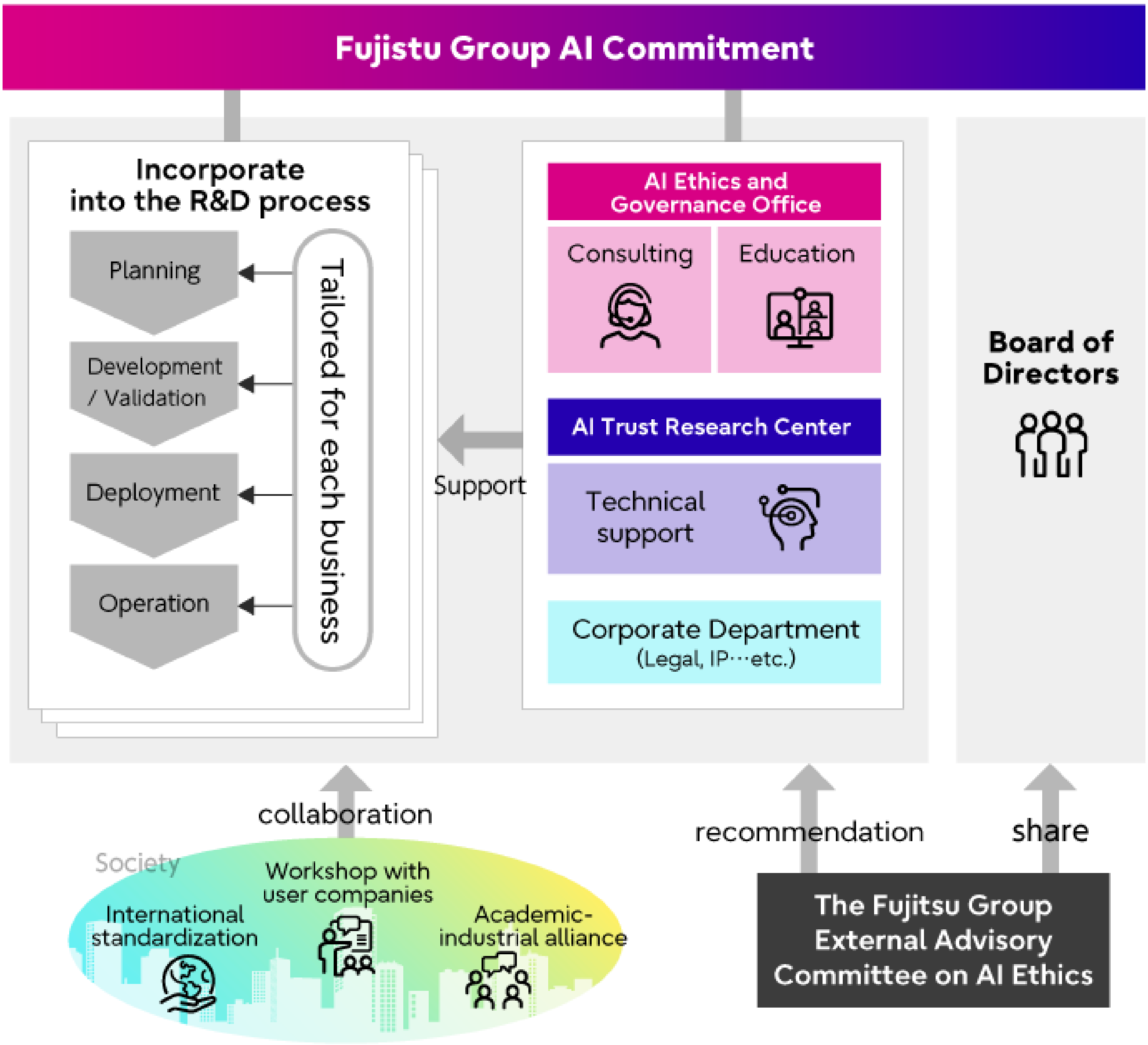
1. **Provide value to customers and society with AI**
2. **Strive for Human Centric AI**
3. **Strive for a sustainable society with AI**
4. **Strive for AI that respects and supports people's decision making**
5. **As corporate responsibility, emphasize transparency and accountability for AI**

10.2 Fujitsu Group External Advisory Committee on AI Ethics

10.2.1 Link AI Ethics to “Corporate Governance”

Fujitsu has been established a committee of external experts as a way of receiving objective, third-party evaluations of the Fujitsu Group's AI Ethics. The committee aims to regularly discuss issues relating to AI Ethics with CDXO and CDPO, thereby enhancing the Fujitsu Group's corporate governance in AI Ethics.

1. **Governance System**


Fujistu Group AI Commitment
Incorporate into the R&D process
Planning
Development / Validation
Deployment
Operation
Tailored for each business
Support
AI Ethics and Governance Office
Consulting
Education
AI Trust Research Center
Technical support
Corporate Department
(Legal, IP···etc.)
Board of Directors
collaboration
recommendation
share
Society
International standardization
Workshop with user companies
Academic-industrial alliance
The Fujitsu Group External Advisory Committee on AI Ethics

Organization

“AI Ethics and Governance Office”, newly established on February 2022, leads, and focuses on implementing measures to actively promote AI Ethics to the society based on Fujitsu Group AI Commitment.

Consulting / Incorporating into the R&D Process

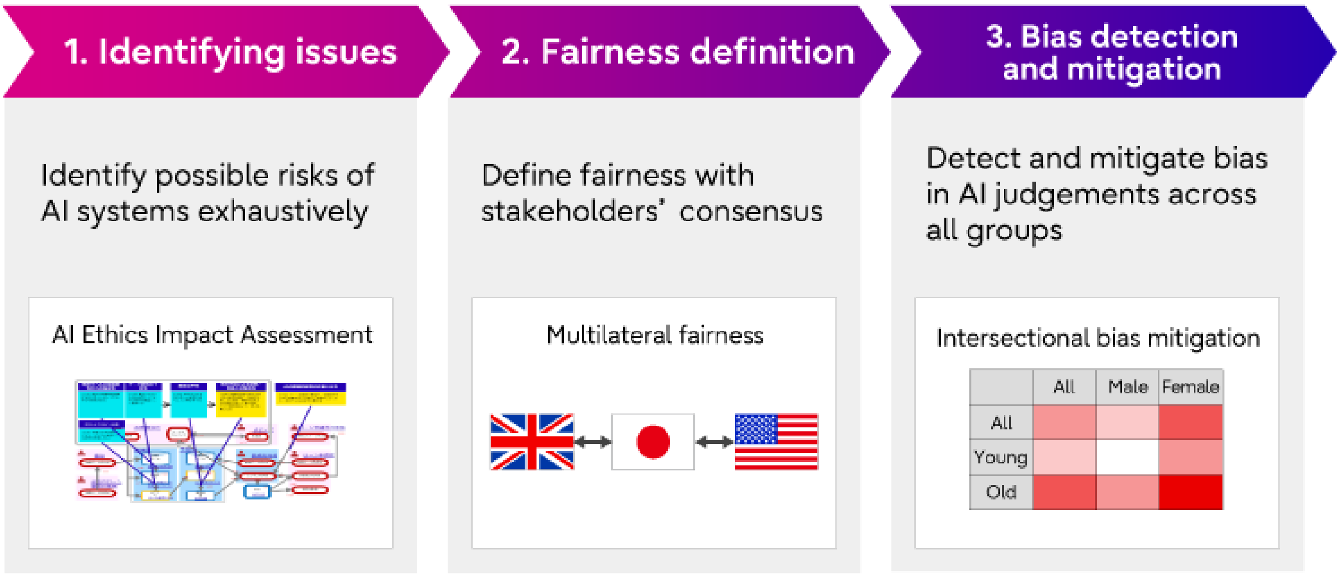
We have established a consultation desk for concerns such as human rights, privacy, and Ethics at each R&D process to avoid AI Ethics issues in advance.
In addition, we are working on various measures to ensure trustworthy AI in each process, such as developing ethical review process and guidelines.

2. Technical Effort to Practice Al Ethics

R&D to help design and audit for trustworthy AI

From a technical perspective, we are trying various initiatives to promote AI ethics from principles to practice. “AI Trust Research Center” works with “AI Ethics Governance Office” to conduct research and development to realize a sustainable world.

Research activities by "AI Trust Research Center”.



1. Identifying issues

Identify possible risks of AI systems exhaustively

AI Ethics Impact Assessment

2. Fairness definition

Define fairness with stakeholders’ consensus

Multilateral fairness

3. Bias detection and mitigation

Detect and mitigate bias in AI judgements across all groups

Intersectional bias mitigation

	All	Male	Female
All			
Young			
Old			

Offering “AI Ethics Impact Assessment”

AI Trust Research Centre developed “AI Ethics Impact Assessment”, a resource toolkit offering developers guidance for evaluating the ethical impact and risks of AI systems based on international AI Ethics guidelines. We offer these resources free of charge to promote the safe and secure deployment of AI systems in society.

3. Academic-industrial collaboration with a social science perspective

One of the activities we are working on together with society is the Academic-industrial collaboration on AI Ethics. By working together with educational and research institutions on AI Ethics, it is expected to have great social significance, by further developing Fujitsu's AI Ethics technologies and applying them to solving social problems, and by returning our company's knowledge to younger generations such as students.
One of the characteristics of our academic-industrial collaboration on AI Ethics is the involvement of various stakeholders across academic fields of science, social science, humanities and so on, as well as gender, age, and national boundaries. For example, in addition to researchers in the field of technology, many experts in the fields of social science, such as law and sociology, are conducting research in our academic-industrial collaboration. AI Ethics is a new field of research, so depending on the culture in which AI is provided and the values of the people, there are different ideas about how AI should be used. Therefore, it is essential to mature the discussion from multiple perspectives with people from diverse backgrounds.

4. Let Us Work Together to Create a Trustworthy AI Society

For further Discussions on AI Ethics Governance

AI Ethics can be realized not only by companies who develop / provide AI, but also by all stakeholders involved in AI must practice AI Ethics in each process. Let us work together and deepen our efforts to realize a sustainable society by secure and trustworthy AI. As a partner working together, please feel free to contact us.

11 Fujitsu Professional Services For AI

11.1 AI Validated Solutions and Infrastructure

We offer a range of AI reference architectures to cater varied business needs:

- Fujitsu Private GPT Solution
- Sentiment Analysis
- Video analysis
- Project specific cases

11.2 Fujitsu Private GPT Solution

As a pioneer in AI technologies, Fujitsu stands among the top holders of AI-related patents globally. We orchestrate a robust ecosystem of AI capabilities, encompassing AI consultancy, innovative technology, managed services, systems integration, and partnerships with local IT providers. This comprehensive network enables us to support your entire generative AI and large language model (LLM) lifecycle—from use case identification and model training to secure, scalable implementation.

Generative AI is transforming business operations by offering unparalleled advantages in automation, innovation, and data-driven decision-making. During this AI-driven revolution, data sovereignty remains a critical priority. On-premises AI solutions allow organizations to harness the full potential of generative AI and LLMs (Large Language Models) while maintaining strict control over data security, privacy, and regulatory compliance.

Fujitsu's Private GPT solution is designed to build customized LLM instances from the ground up, integrating the full software and hardware stack for a secure, on-premises GPT deployment. Our solution is optimized for your specific datasets and business use cases, leveraging the full power of the Fujitsu AI ecosystem.

We provide comprehensive AI lifecycle management, covering everything from initial consultancy, solution design, and model development to testing, deployment, and ongoing maintenance. Our approach ensures a deep understanding of your unique data landscape and business needs, enabling us to deliver fully integrated, end-to-end solutions that empower your organization to thrive with AI.

11.3 The Fujitsu Gen AI Test Drive

In collaboration with industry-leading partners such as Intel, SUSE, NetApp, and Juniper, we have developed the Fujitsu DX Innovation Platform—a comprehensive solution for advancing digital transformation. A key component of this platform is the Fujitsu AI Test Drive, offering a unique opportunity to trial innovative, purpose-built AI infrastructure. With full control over your own datasets, you can experiment with every stage of the AI lifecycle, including data curation, ingestion, and preprocessing (data cleaning).

The AI Test Drive leverages a broad range of reference architectures, designed by expert solution architects, ensuring flexibility to configure the ideal infrastructure components for your specific use cases. This enables the rapid prototyping and deployment of AI workflows tailored to your business needs, offering a scalable environment for testing machine learning models, deep learning algorithms, or generative AI applications with real-world datasets.

The AI Test Drive provides you with a complete infrastructure comprising the following three main layers:

- Containerization
- Virtualization
- 3-tier hybrid infrastructure (Edge-Core-Cloud)

The basis is Kubernetes managed via SUSE® Rancher Prime, with SUSE® Virtualization implemented as the virtualization layer. The entire stack is an open source based 3-tier hybrid infrastructure with multi-tenancy and RBAC (Role-Based Access Control).

11.4 Why Should You Try Out the AI Test Drive?

The Fujitsu AI Test Drive allows data scientists to assess their requirements, validate data integrity, and benchmark performance metrics against the latest AI technologies. Users can test their specific use cases or evaluate pre-configured scenarios, and experiment with different hardware configurations to identify the optimal AI platform for their needs. Additionally, the Test Drive enables users to identify system dependencies and assess the interoperability of various components.

Developed from a range of reference architectures and supported by a network of ecosystem partners, the AI Test Drive ensures comprehensive support from AI specialists who are available to assist with projects daily. As AI complexities evolve, leveraging our ecosystem enables users to bridge skill gaps and refine their AI infrastructure strategy with expert guidance.

Engaging with Fujitsu's AI Test Drive accelerates time-to-value for AI projects by consolidating hardware and software resources in a unified environment. This streamlined approach facilitates rapid access to AI capabilities, allowing for effective validation of models and theories, thereby providing a competitive advantage in the global AI adoption landscape.

12 Conclusion

In conclusion, this whitepaper has explored the multifaceted components and technologies involved in the development, deployment, and management of AI systems. We began by examining the fundamental aspects of AI—training, inferencing, and finetuning—and how these processes are applied in real-world use cases. Intel's Confidential Computing emerged as a key enabler for securing sensitive data during AI operations, particularly in cloud and multi-tenant environments. Robust PRIMERGY Servers and NVIDIA GPUs were highlighted for their critical role in scaling AI workloads, while open-source platforms such as SUSE® Rancher Prime, SUSE® Virtualization, and Red Hat showcased their capabilities in managing containerized environments. Fujitsu's Private GPT solution illustrated the value of on-premise AI for ensuring data security within enterprises, and customer use cases underscored the importance of precise infrastructure sizing. We also explored the validation and benchmarking of AI models to optimize performance, discussed Fujitsu's ethical guidelines for responsible AI development, and highlighted the range of professional services available to help organizations maximize the potential of their AI initiatives. Ultimately, this whitepaper provides a comprehensive roadmap for businesses seeking to harness the power of AI in a secure, scalable, and ethical manner.

Start your AI journey with Fujitsu today.

https://www.fujitsu.com/global/products/data-transformation/data-driven/ai-test-drive/

13 References

BGE-M3. (n.d.). Retrieved from https://huggingface.co/BAAI/bge-m3

bge-reranker-v2-m3. (n.d.). Retrieved from https://huggingface.co/BAAI/bge-reranker-v2-m3

Fairness by Design. (2021, May 14). Retrieved from https://www.fujitsu.com/global/about/research/article/202105-ai-ethics.html

Fujitsu AI Ethics. (n.d.). Retrieved from https://global.fujitsu/en-global/technology/key-technologies/ai/aiethics/governance

Private GPT – a Fujitsu AI Solution Whitepaper. (2024, September). Retrieved from https://sp.ts.fujitsu.com/dmsp/Publications/public/wp-private-gpt-en.pdf

Xamk Hippu makes AI supercomputing accessible. (2024, March). Retrieved from https://sp.ts.fujitsu.com/dmsp/Publications/public/cs-xamk-2024-en.pdf

14 Abbreviations

Abbreviation	Explanation
AD	Active Directory
AI	Artificial Intelligence
API	Application Programming Interface
AWS	Amazon Web Services
CDPO	Certified Data Protection Officer
CDXO	Chief Digital Transformation Officer
CI/CD	Continuous integration / Continuous delivery
CPU	Central Processing Unit
CXL	Compute Express Link
EU	European Union
FAISS	Facebook AI Similarity Search
FAQ	Frequently Asked Questions
FC	Fibre Channel
FHFL	Full-height Full-length
FLOPS	Floating Point Operations Per Second
GCP	Google Cloud Platform
GDPR	General Data Protection Regulation
GenAI	Generative Artificial Intelligence
GPT	Generative Pre-trained Transformer
GPU	Graphics Processing Unit
HBM	High Bandwidth Memory
HCI	Hyperconverged Infrastructure
HIPAA	Health Insurance Portability and Accountability Act
HPC	High-Performance Computing
HTTPS	Hypertext Transfer Protocol Secure
I/O	Input/Output
IoT	Internet of Things
ISO	International Organization for Standardization
IT	Information Technology
KV	Key-Value
LLM	Large Language Model
MIG	Multi-instance GPU
ML	Machine Learning
NFS	Network File System
NLP	natural language processing
NPU	neural processing unit
OCR	Optical Character Recognition
PCIe	Peripheral Component Interconnect Express

PFR	Platform Firmware Resilience
POC	Proof Of Concept
R&D	Research and Development
RAG	Retrieval Augmented Generation
RAID	Redundant Array of Independent Disks
RBAC	Role-based Access Control
RLHF	Reinforcement Learning from Human Feedback
ROI	Return On Investment
S3	Simple Storage Service
SDK	Software Development Kit
SFTP	Secure File Transfer Protocol
SGX	Software Guard Extensions
SLES	SUSE Linux Enterprise Server
SMB	Server Message Block
SMTP	Simple Mail Transfer Protocol
SOME	Sparse Mixture-of-Experts
SSH	Secure Shell
SVM	Support Vector Machine
TCP	Transmission Control Protocol
TDX	Trust Domain Execution
TFTT	Time to First Token
TPU	Tensor Processing Unit
UI	User Interface
VLM	Vision Language Model
VM	Virtual Machine
VRAM	Video Random Access Memory