White Paper

Intel® TDX: Empowering Baidu AI Cloud for Confidential and Efficient LLM Applications

BAIDU AI CLOUD

Contents

Data Security Challenges in LLM Applications

Large language models (LLMs) are at the epicenter of transformation in the artificial intelligence era, profoundly reshaping the way we live, learn, and work. The release of ChatGPT sounded the trumpet for the large-scale industrialization of LLMs, and the in-depth integration of large language models in various industry applications was a key topic at the 2024 WAIC (World Artificial Intelligence Conference), where it was acknowledged that the adoption of large language models faces two major challenges:

- Constrained by high costs and explosive demand, AI (Artificial Intelligence) computing power is difficult to obtain, especially for the large number of small and medium enterprises that are potential drivers of application innovations.
- The industry needs high-quality data and smooth data flow to support the integration of LLMs into diversified sectors. However, the lack of infrastructure to ensure the secure and seamless flow of data assets hinders the process.

As regards the first challenge, the popularity of public clouds has made it easier to utilize AI computing power everywhere. The Baidu Baige heterogeneous computing platform offers a range of computing power options catering to diverse needs to meet AI computing demand, including availability in various regions, with different scales and capabilities. At the same time, Baige enhances the performance of AI computing capability and alleviates industry computing power shortages through technological innovations such as chip optimization and "multiplexing and scheduling" technology, effectively reducing computing power costs for enterprises.

However, the second challenge – namely, the acquisition of data assets and their secure and efficient circulation – has become a bottleneck issue for the industrial application of LLM, one that exists across various processes of large language models, with typical examples being during the training stage and inference stages:

- **Training stage:** It is very challenging to securely and efficiently gather industry data owned by different enterprises and institutions. For example, in the financial industry, various banks, insurance companies and other such institutions have a great deal of industry knowledge and customer data, but "due to strict data access procedures in the financial industry, it is highly challenging to build specialized large language models using data from different organizations in the absence of effective data security and technologies," a customer's senior expert told Baidu AI Cloud.
- **Inference stage:** Similarly, the lack of a mutual trust-based data security supporting infrastructure across cloud service providers, model vendors, and end users hinders the deployment of LLM inference services. For model vendors, high-value models are important assets of their enterprise, and they are worried about the leakage of their models. On the other hand, enterprise customers are worried about their proprietary trade secrets and data being stolen. In addition, individual end users also worry about their personal data being misused by service providers, resulting in data leaks and consequently privacy breaches.

Building Confidential AI Infrastructure with Intel® TDX

Privacy-preserving computing technologies provide a secure and trustworthy technical guarantee for the deployment of LLM services. These technologies, such as multi-party computation (MPC), homomorphic encryption (HE), and trusted execution environments (TEEs), ensure that service participants can complete the required task while protecting the security of their respective data. Among these, confidential computing characterized by trusted execution environments is better suited to today's industrial deployment demands in terms of performance, application compatibility, and operational management.

Cloud service providers provide an IaaS base that supports AI computing and confidential computing, which further improves the ability to commercialize LLMs at scale and enables their widespread use in various industries. Traditional data security technologies focus more on data in rest and transit, whereas confidential computing provides protection for data in use by building a trusted execution environment, making the data available and invisible.

The 5th Gen Intel® Xeon® Processor provides a built-in Intel security engine while maintaining excellent performance, and supports both Intel® Software Guard Extensions (Intel® SGX) and Intel® Trust Domain Extensions (Intel® TDX) confidential computing technologies, enabling customers to choose confidential computing technologies that better meet their business needs and regulatory requirements, and enhancing the protection of data confidentiality and code integrity.

Intel® SGX provides application-level and function-level data isolation, whereas Intel® TDX provides isolation boundaries and confidentiality protection at the virtual machine level. Furthermore, Intel® TDX is a hardware-based TEE facilitating the deployment of trust domains (TD), which are isolated and encrypted virtual machines (VM) designed to protect sensitive data and applications from unauthorized access (Figure 1).

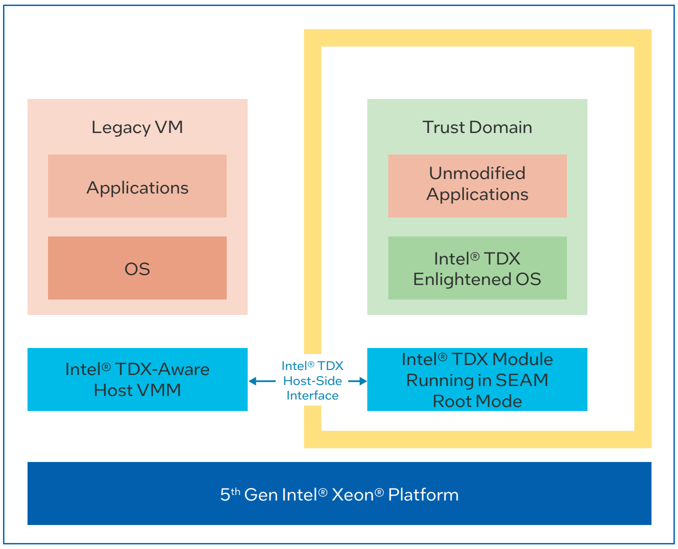
Legacy VM
Applications
OS
Trust Domain
Unmodified Applications
Intel® TDX Enlightened OS
Intel® TDX-Aware Host VMM
Intel® TDX Host-Side Interface
Intel® TDX Module Running in SEAM Root Mode
5th Gen Intel® Xeon® Platform

Figure 1.

Moreover, users can directly run their applications in TD to obtain a runtime security guarantee without any code modification or integration work. Intel® TDX helps ensure data integrity, confidentiality, and authenticity, which empowers engineers and tech professionals to create and maintain secure systems, enhancing trust in virtualized environments.

In line with industry changes, Baidu AI Cloud has developed confidential computing IaaS products for AI and LLM applications based on the 5th Gen Intel® Xeon® Processor and Intel® TDX technology.

Intel® AMX to optimize confidential LLM inference

Intel® TDX technology establishes a virtualized Trusted Execution Environment (TEE) that seamlessly supports various LLM applications while ensuring the security of sensitive data and proprietary models required by applications. Taking advantage of Intel® Advanced Matrix Extensions (Intel® AMX), a built-in accelerator, Intel® Xeon® Processors can accelerate LLM inference to maximize the value of data security and ensure high performance:

Intel® AMX BF16/INT8 quantization boost performance

Quantization with Intel® AMX brings an infrastructure bonus. It can run LLM inference in a smaller hardware footprint, conserving the memory and computing resources needed to operate.

Intel® TDX and Intel® AMX provide impressive performance with security

General LLM workloads like Llama 2-7B/13B produce substantial throughput and latencies.

What is Intel® AMX?

Intel® AMX is a built-in accelerator that enables Intel® Xeon® processors to optimize deep learning training and inferencing workloads. Intel® AMX architecture consists of two components: Tiles and Tile Matrix Multiplication (TMUL) (see Figure 2). With this new tiled architecture, Intel® AMX generation-on-generation performance gains are significant.

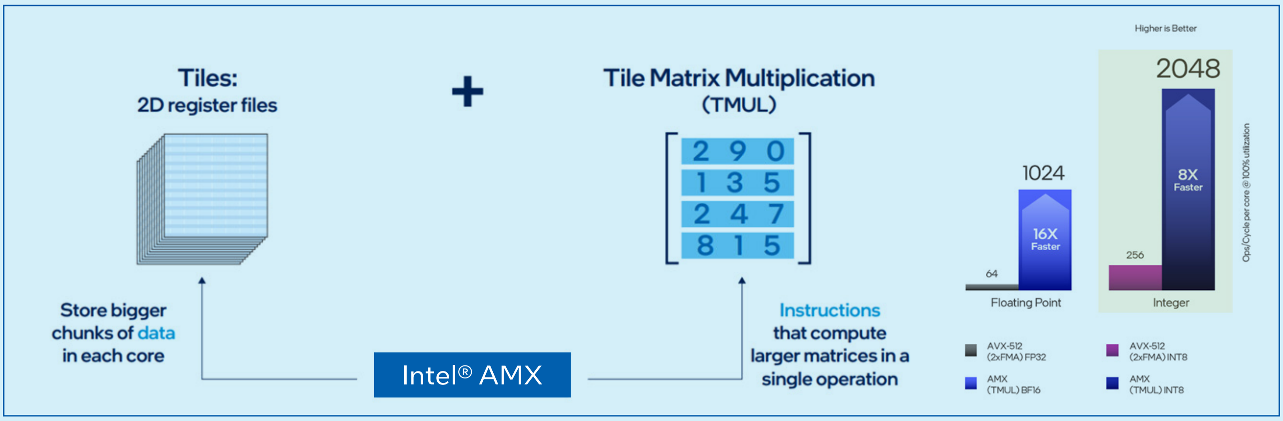

Figure 2. Intel® AMX Architecture

Intel® TDX for Protecting Heterogeneous AI Workloads and Infrastructure

Heterogeneous acceleration is a key requirement in the deployment of large language models. The Intel's virtualization-based Intel® TDX technology offers robust support for confidential computing needs in a heterogeneous environment by extending trust. It enables the creation of a heterogeneous confidential computing environment by binding a confidential computing accelerator with the TEE running on the main processor.

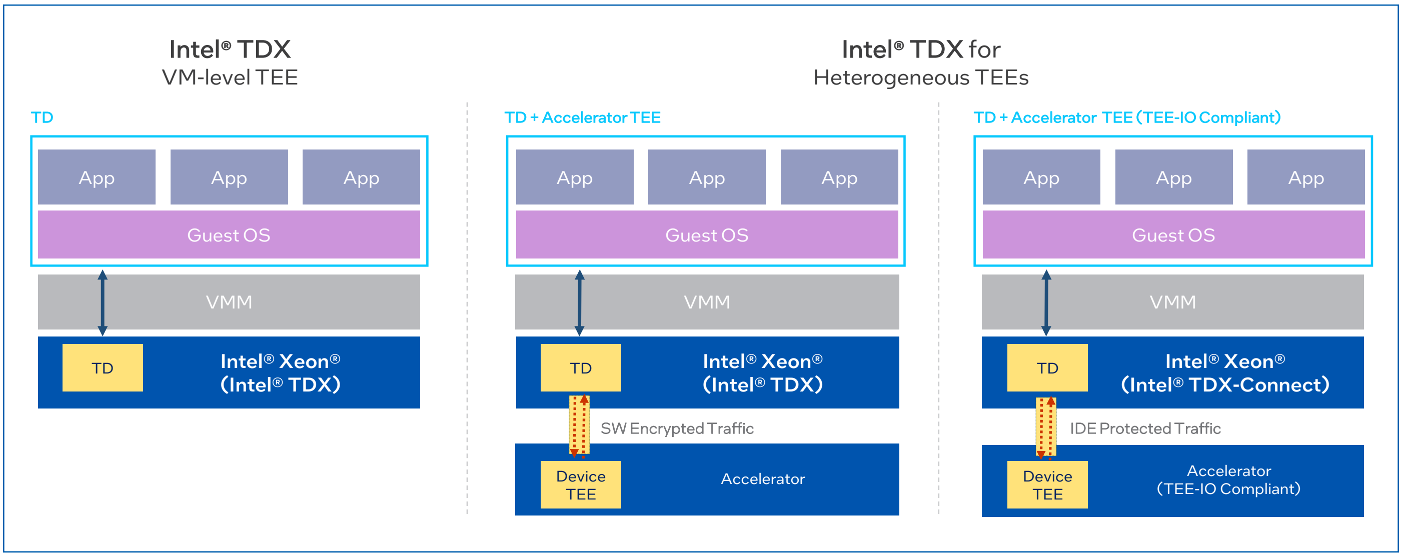

Figure 3. Intel® TDX Roadmap

Leveraging 5th Gen Intel® Xeon® processors with Intel® TDX technology and the exceptional support for heterogeneous accelerators, Baidu AI Cloud expands the AI computing infrastructure with confidential computing capabilities. This not only addresses data security requirements in artificial intelligence applications, but also meets the performance requirements for various LLMs. In this confidential mode:

- The Intel® TDX-based host TEE cooperates with the accelerator TEE to facilitate the secure transfer of data to and from the encrypted memory of the CPU and the accelerator.
- Upon loading the accelerator driver within the TD, it creates a secure communication channel with the accelerator, utilizing this channel for all subsequent data exchanges between the CPU and the accelerator.
- This process utilizes an encrypted buffer, known as a bounce buffer, which is designated within TD shared memory and is accessible by the accelerator.
- Furthermore, customers have the option to request attestation to confirm that both the VMs and accelerator are operating within a correctly configuration state before initiating security-sensitive applications and uploading confidential information, such as proprietary model parameters or data.

The heterogeneous TEE enabled by Intel® TDX excludes host operation system software, system firmware, and IO links from trusted boundaries of sensitive applications and data, ensuring data security in LLM applications. Baidu AI Cloud for confidential AI built on Intel® TDX can provide confidentiality computing infrastructure for data assets in LLM applications – including model vendors' model data, enterprise customers' domain knowledge database, and privacy information – during the user inference service.

Baidu AI Cloud offers customers more options and flexibility to run LLM workloads securely and efficiently on the cloud, keeping private customer data and the LLM model protected end-to-end. Customers – whether they are model providers, enterprises, or individual users – can use remote attestation to verify the trustworthiness of remote execution environments before granting access to proprietary data and sensitive information.

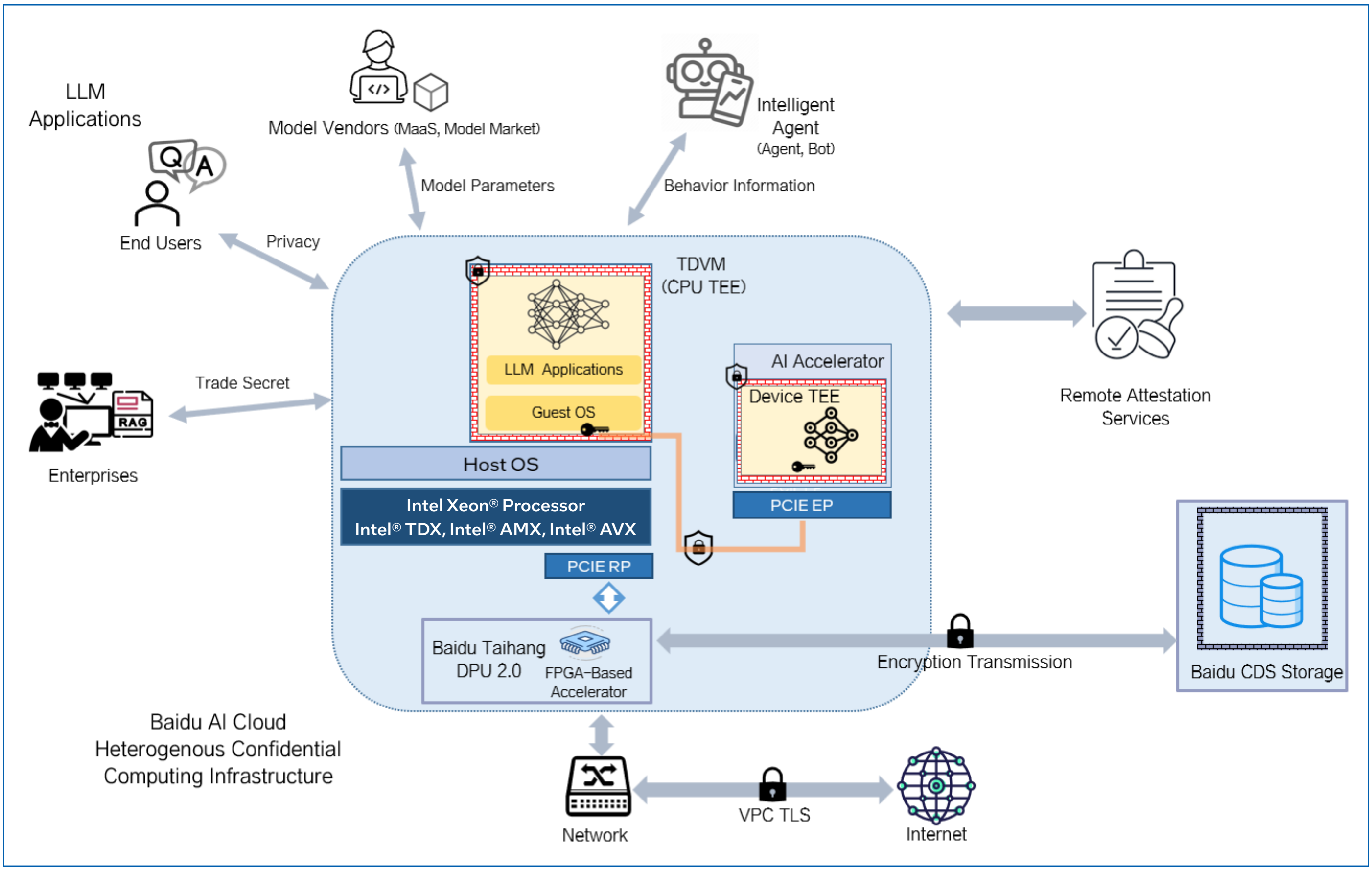

Figure 4. Confidential Inference Based on Baidu AI Cloud Heterogeneous Instance

Intel® AVX instructions can effectively improve data encryption performance during the data exchange process between host TEE and accelerator TEE. As illustrated in Figure 5, end-to-end performance testing of large models running on Baidu AI Cloud Confidential AI Instance shows that introducing confidentiality protection has a minimal impact, well within acceptable performance limits for production, thereby ensuring both confidentiality and high performance.

The 5th Gen Intel® Xeon® processor and Intel® TDX technology offer a crucial opportunity to maximize AI capabilities, especially for industries that need to deal with security-sensitive data, such as healthcare and finance.

Furthermore, future Intel® TDX Connect technology will introduce hardware-level protection capabilities for heterogeneous acceleration, such as PCIe IDE for link confidentiality and IOMMU enhancement on trusted MMIO and DMA access, to provide enhanced security and efficiency for confidentiality in AI applications.

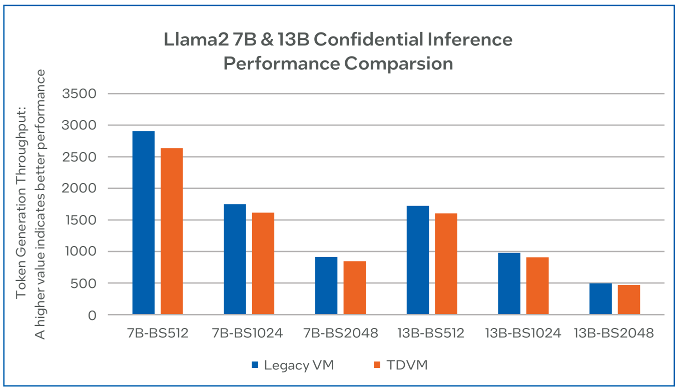

Figure 5. Confidential Inference Performance Comparison[1]

Intel® TDX Connect improves workload performance by removing the additional work needed for data encryption and decryption as well as back-and-forth memory movement.

Summary

Today, the application of artificial intelligence (AI) is becoming ubiquitous along with highly developed internet and cloud computing capabilities, and generative AI (GenAI) and large language models (LLMs) are reshaping the productivity of enterprises and society. Despite the various benefits and advantages of artificial intelligence, growing concerns about data security and privacy leakage have become obstacles for AI adoption in various industries.

Baidu AI Cloud has been closely collaborating with Intel to build secure, reliable and highly efficient cloud and AI infrastructure. Its confidential computing AI IaaS capabilities built on the 5th Gen Intel® Xeon® Processor and Intel® TDX technology provide flexible support for data security across various AI usages, with particular emphasis on safeguarding data and privacy with optimized performance in the rapidly evolving landscape of large language models.

Baidu and Intel will continue this in-depth cooperation to promote the application of confidential computing and future IO acceleration technology in GenAI and LLMs, building secure and easy-to-use AI infrastructure for industry applications.

In view of the fact that the widespread use of large language models has increased standards and demand for more secure and more efficient cloud infrastructure, Intel is dedicated to helping cloud service providers meet these higher requirements while ensuring data protection and performance.

To learn more, visit intel.com/tdx

[1] Test Configuration: Intel® Xeon® Platinum 8563C, 2.6GHZ, VM: 22VCPU, 352 GB. Performance varies by use, configuration and other factors. Learn more at www.Intel.com/PerformanceIndex

Baidu AI Cloud

Established in 2015, Baidu AI Cloud provides enterprises and developers with leading-edge artificial intelligence, big data, cloud computing services and user-friendly development tools. Over the years, Baidu AI Cloud has been deep in the realm of industrial intelligence, serving over 5 million enterprise customers and developers, holding the largest domestic market share of AI public cloud services for four consecutive years. In 2020, Baidu AI Cloud took the lead in initiating the 'Integration of Cloud and AI' strategy. As the world witnessed a surge of interest in generative AI in 2023, Baidu AI Cloud responded by launching the 'Baidu AI Cloud AI Native Stack', offering access to its efficient 'Baidu-Baige AI Heterogeneous Computing Platform', its powerful ERNIE Foundation Model, its user-friendly, one-stop Foundation Model Development Platform 'Qianfan ModelBuilder' and its AI-Native Application Development Platform 'Qianfan AppBuilder', along with abundant AI native application prototypes and industry solutions. With these offerings, Baidu AI Cloud is dedicated to meeting clients' diverse business needs in the AI-Native era, empowering thousands of industries and in doing so, accelerating the advancement of industrial intelligence.

Notices & Disclaimers

Performance varies by use, configuration and other factors. Learn more on the Performance Index site.

Performance results are based on testing as of dates shown in configurations and may not reflect all publicly available updates. See backup for configuration details. No product or component can be absolutely secure.

Your costs and results may vary.

Intel technologies may require enabled hardware, software or service activation.