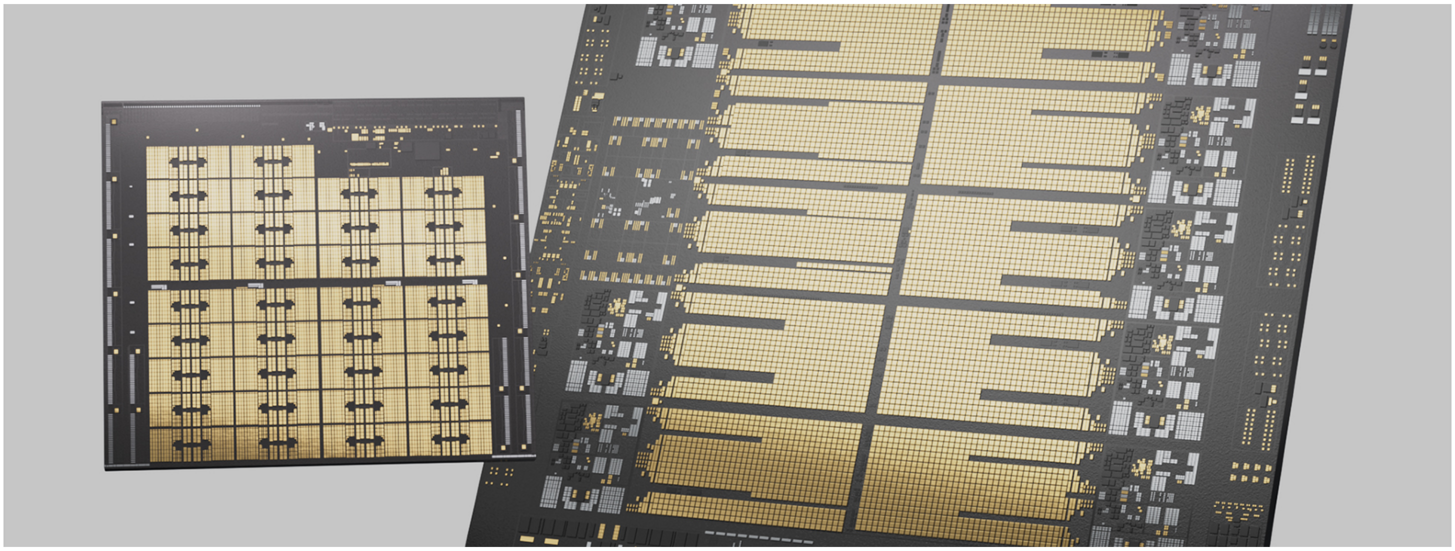
IBM Newsroom

New IBM Processor Innovations To Accelerate AI on Next-Generation IBM Z Mainframe Systems

New IBM Telum II Processor and IBM Spyre Accelerator unlock capabilities for enterprise-scale AI, including large language models and generative AI

Advanced IO technology enables and simplifies a scalable IO sub-system designed to reduce energy consumption and data center footprint



PALO ALTO, Calif., Aug. 26, 2024 /PRNewswire/ -- IBM (NYSE:IBM) revealed architecture details for the upcoming IBM Telum® II Processor and IBM Spyre™ Accelerator at Hot Chips 2024. The new technologies are designed to significantly scale processing capacity across next generation IBM Z mainframe systems helping accelerate the use of traditional AI models and Large Language AI models in tandem through a new ensemble method of AI.

Your browser does not support the video tag.

With many generative AI projects leveraging Large Language Models (LLMs) moving from proof-of-concept to production, the demands for power-efficient, secured and scalable solutions have emerged as key priorities. Morgan Stanley research published in August projects generative AI's power demands will skyrocket 75% annually over the next several years, putting it on track to consume as much energy in 2026 as Spain did in 2022.[1] Many IBM clients indicate architectural decisions to support appropriately sized foundation models and hybrid-by-design approaches for AI workloads are increasingly important.

The key innovations unveiled today include:

- **IBM Telum II Processor:** Designed to power next-generation IBM Z systems, the new IBM chip features increased frequency, memory capacity, a 40 percent growth in cache and integrated AI accelerator core as well as a coherently attached Data Processing Unit (DPU) versus the first generation Telum chip. The new processor is expected to support enterprise compute solutions for LLMs, servicing the industry's complex transaction needs.
- **IO acceleration unit:** A completely new Data Processing Unit (DPU) on the Telum II processor chip is engineered to accelerate complex IO protocols for networking and storage on the mainframe. The DPU simplifies system operations and can improve key component performance.

- **IBM Spyre Accelerator:** Provides additional AI compute capability to complement the Telum II processor. Working together, the Telum II and Spyre chips form a scalable architecture to support ensemble methods of AI modeling – the practice of combining multiple machine learning or deep learning AI models with encoder LLMs. By leveraging the strengths of each model architecture, ensemble AI may provide more accurate and robust results compared to individual models. The IBM Spyre Accelerator chip, previewed at the Hot Chips 2024 conference, will be delivered as an add on option. Each accelerator chip is attached via a 75-watt PCIe adapter and is based on technology developed in collaboration with the IBM Research. As with other PCIe cards, the Spyre Accelerator is scalable to fit client needs.

"Our robust, multi-generation roadmap positions us to remain ahead of the curve on technology trends, including escalating demands of AI," said Tina Tarquinio, VP, Product Management, IBM Z and LinuxONE. "The Telum II Processor and Spyre Accelerator are designed to deliver high-performance, secured, and more power efficient enterprise computing solutions. After years in development, these innovations will be introduced in our next generation IBM Z platform so clients can leverage LLMs and generative AI at scale."

The Telum II processor and the IBM Spyre Accelerator will be manufactured by IBM's long-standing fabrication partner, Samsung Foundry, and built on its high performance, power efficient 5nm process node. Working in concert, they will support a range of advanced AI-driven use cases designed to unlock business value and create new competitive advantages. With ensemble methods of AI, clients can achieve faster, more accurate results on their predictions. The combined processing power announced today will provide an on ramp for the application of generative AI use cases. Some examples could include:

- **Insurance Claims Fraud Detection:** Enhanced fraud detection in home insurance claims through ensemble AI, which combine LLMs with traditional neural networks geared for improved performance and accuracy.
- **Advanced Anti-Money Laundering:** Advanced detection for suspicious financial activities, supporting compliance with regulatory requirements and mitigating the risk of financial crimes.
- **AI Assistants:** Driving the acceleration of application lifecycle, transfer of knowledge and expertise, code explanation as well as transformation, and more.

Specifications and Performance Metrics:

Telum II processor: Featuring eight high-performance cores running at 5.5GHz, with 36MB L2 cache per core and a 40% increase in on-chip cache capacity for a total of 360MB. The virtual level-4 cache of 2.88GB per processor drawer provides a 40% increase over the previous generation. The integrated AI accelerator allows for low-latency, high-throughput in-transaction AI inferencing, for example enhancing fraud detection during financial transactions, and provides a fourfold increase in compute capacity per chip over the previous generation.

The new IO Acceleration Unit DPU is integrated into the Telum II chip. It is designed to improve data handling with a 50% increased IO density. This advancement enhances the overall efficiency and scalability of IBM Z, making it well suited to handle the large-scale AI workloads and data-intensive applications of today's businesses.

Spyre Accelerator: A purpose-built enterprise-grade accelerator offering scalable capabilities for complex AI models and generative AI use cases is being showcased. It features up to 1TB of memory, built to work in tandem across the eight cards of a regular IO drawer, to support AI model workloads across the mainframe while designed to consume no more than 75W per card. Each chip will have 32 compute cores supporting int8 and fp16 datatypes for both low-latency and high-throughput AI applications.

Availability

The Telum II processor will be the central processor powering IBM's next-generation IBM Z and IBM LinuxONE platforms. It is expected to be available to IBM Z and LinuxONE clients in 2025. The IBM Spyre Accelerator, currently in tech preview, is also expected to be available in 2025.

Statements regarding IBM's future direction and intent are subject to change or withdrawal without notice, and represent goals and objectives only.

About IBM

IBM is a leading provider of global hybrid cloud and AI, and consulting expertise. We help clients in more than 175 countries capitalize on insights from their data, streamline business processes, reduce costs and gain the competitive edge in their industries. Thousands of government and corporate entities in critical infrastructure areas such as financial services, telecommunications and healthcare rely on IBM's hybrid cloud platform and Red Hat OpenShift to affect their digital transformations quickly, efficiently and securely. IBM's breakthrough innovations in AI, quantum computing, industry-specific cloud solutions and consulting deliver open and flexible options to our clients. All of this is backed by IBM's long-standing commitment to trust, transparency, responsibility, inclusivity and service.

Additional Sources

- Read more about the IBM Telum II Processor.
- Read more about the IBM Spyre Accelerator.
- Read more about the IO Accelerator

Media Contact:

Chase Skinner
IBM Communications
chase.skinner@ibm.com

Aishwerya Paul
IBM Communications
aish.paul@ibm.com

[1] Source: Morgan Stanley Research, August 2024.

SOURCE IBM

Additional assets available online: Photos (3)
Video (1)

https://stage.mediaroom.com/ibmnewsroom/ai-on-z