IBM's Telum II & Spyre AI Accelerators

STEVE MCDOWELL, CHIEF ANALYST
8/29/24

CONTEXT

In 2021, IBM made a significant leap forward by introducing the IBM Telum processor, which featured the company's first advanced on-processor AI accelerator for inferencing. This was a major factor behind a bump in the success of the IBM z16 mainframe program, addressing the growing need for AI integration in enterprise workloads.

At the Hot Chips 2024 conference in Palo Alto, California, IBM unveiled the next generationj of enterprise AI solutions: the IBM Telum II processor and the IBM Spyre Accelerator. These new technologies should meet the demands of the AI era, providing enhanced performance, scalability, and AI capabilities. Both are expected to be available in 2025.

IBM TELUM II

The IBM Telum II processor is a significant upgrade over its predecessor. Developed using Samsung's 5nm process technology, it includes eight high-performance cores, each running at 5.5GHz. This is a substantial improvement in processing power aimed at supporting the most demanding enterprise workloads.

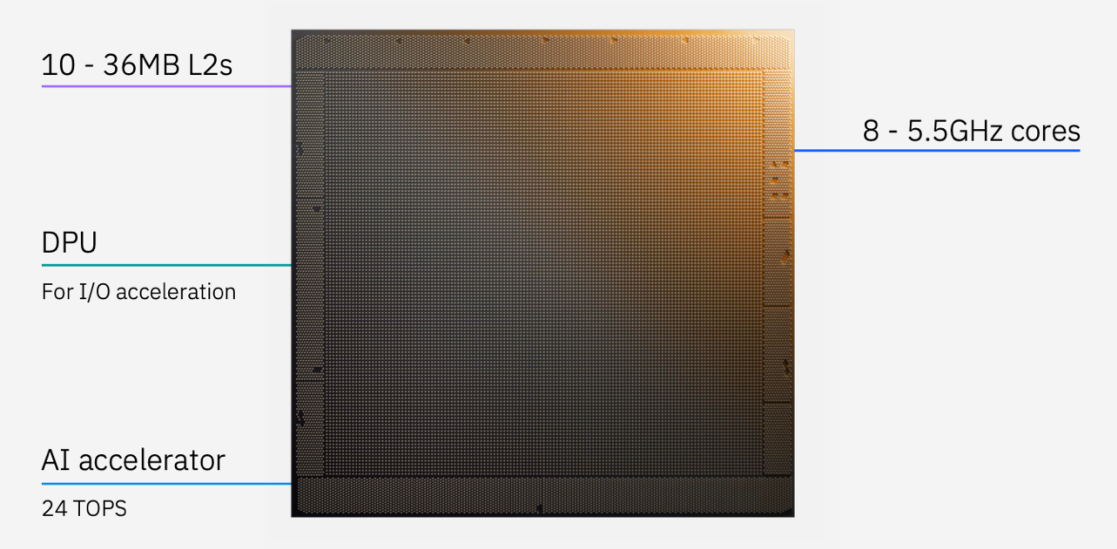

FIGURE 1: IBM TELLUM II (SOURCE: IBM)

The IBM Telum II processor delivers several key advancements and technical features:

- **Development and Technology:**
  - Developed using Samsung's 5nm process technology.
  - Features eight high-performance cores, each running at 5.5GHz.
- **On-Chip Cache Enhancements:**
  - 40% increase in on-chip cache capacity.
  - Virtual L3 cache expanded to 360MB.
  - Virtual L4 cache increased to 2.88GB.
- **AI Acceleration Capabilities:**
  - AI accelerator delivers 4x the compute power of its predecessor, achieving 24 trillion operations per second (TOPS).
  - Architectural improvements enable model runtimes to coexist with demanding enterprise workloads.
  - Enhanced support for INT8 data type, improving compute efficiency for specific AI applications.

- New compute primitives for better support of large language models (LLMs), expanding the range of AI models that can be analyzed.

- **System-Level Improvements:**
  - Enhanced load balancing across AI accelerators, with each accelerator accepting work from any core in the processor drawer.
  - When fully configured across all eight AI accelerators in the drawer, it provides up to 192 TOPS of AI processing power.

IBM SPYRE

The IBM Spyre Accelerator complements the Telum II processor, offering additional AI processing power for enterprise workloads.

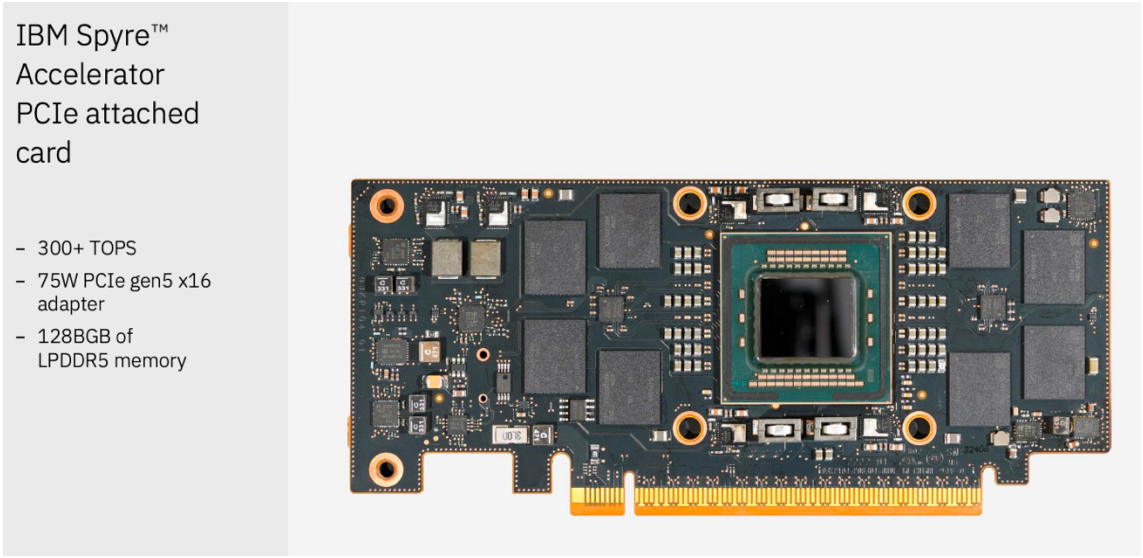

FIGURE 2: IBM SPYRE (SOURCE: IBM)

Key aspects of the Spyre Accelerator include:

- **Technical Specifications:**
  - Contains 32 AI accelerator cores, with a similar architecture to the AI accelerator in the Telum II chip.
  - Multiple Spyre Accelerators can be connected to IBM Z's I/O subsystem via PCIe, allowing for scalable AI processing capabilities.

- **Ensemble AI Methodology:**

- Designed to support a broader set of AI use cases, particularly those involving ensemble AI methods.
- Ensemble AI leverages multiple models to enhance prediction accuracy and performance.
- Applications include fraud detection and compliance monitoring, where traditional neural networks and LLMs can be combined for improved outcomes.

ANALYSIS

IBM's introduction of the Telum II processor and Spyre Accelerator is a strategic move reinforcing the company's leadership in enterprise computing. By focusing on both the hardware advancements and the practical application of AI, IBM is not just keeping pace with the AI revolution—it's helping to drive it forward.

The IBM Telum II processor is a clear evolution of its predecessor, boasting a series of enhancements that address the growing demands of modern enterprise workloads. From a technical standpoint, the move to Samsung's 5nm technology and the expansion of on-chip cache capacities are improvements that will directly impact performance. However, the real story here is the enhanced AI acceleration capabilities.

The Telum II processor, with its 4x increase in AI processing power to 24 TOPS, meets this need. Notably, IBM has focused on raw performance metrics and optimized the architectural design to ensure these AI capabilities can be seamlessly integrated into enterprise workloads.

The IBM Spyre Accelerator adds another layer of sophistication to IBM's AI strategy. By providing a scalable, modular AI processing solution that can be integrated into existing IBM Z systems, IBM addresses a key pain point for many enterprises: how to scale AI capabilities without overhauling existing infrastructure.

NAND Research is a technology-focused industry analyst firm providing research, customer content, market and competitive intelligence, and custom deliverables to technology vendors, investors, and end-customer IT organizations.

Contact NAND Research via email at info@nand-research.com or visit our website at nand-research.com.