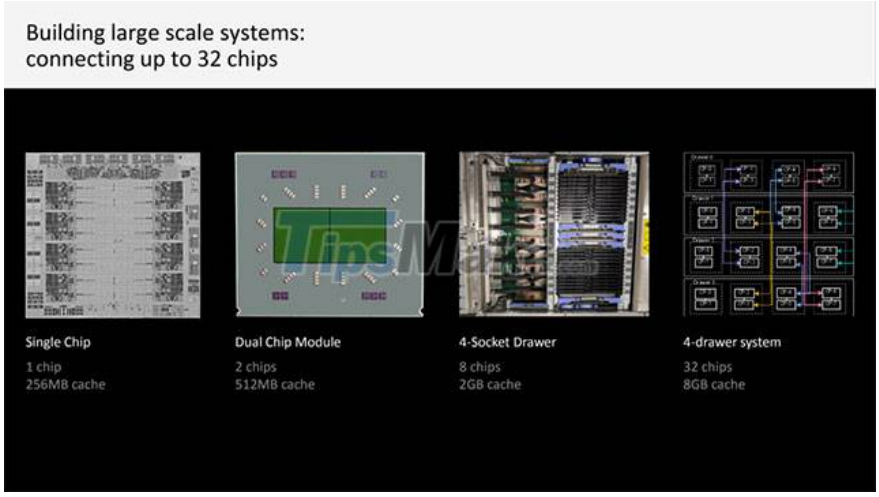
IBM announces next generation Z processor: 7nm Telum chip, 22.5 billion transistors, 8 cores running on 5GHz clock

IBM has just released relatively detailed information about its next-generation Telum chipset, which is part of the new Z series of processors. The Telum chip carries a completely new core architecture design, aimed at radically enhancing AI processing capabilities.

In fact, Telum is IBM's first microprocessor to integrate Artificial Intelligence (AI) on the processor chip.

Specifically, according to IBM, the next-generation Z cores are optimized along with an all-new cache and multi-chip hierarchy that enables up to 40% performance improvement per socket. Telum includes a total of 8 cores with dedicated L2 cache. This CPU model has a built-in SMT2 feature that provides 16 on-chip threads, while a maximum configuration of 32 cores and 64 threads is available with a 4-chamber system.

The clock speed is said to be over the 5GHz threshold, while each individual L2 cache is 32MB in size and the latency uses a 19-cycle load (~3.8 ns including CTR access). The dual-chip modular design contains 22 billion transistors and more than 30,000 meters of wiring across 17 metal layers.



Moving on to shared L3 and L4 caching across 8 cores, the IBM Z Telum chip comes with 256MB virtual 'on-chip' L3 cache and 2GB virtual L4 cache on up to 8 cores. L2 cache uses a 320 GB/s dual ring topology, while L3 cache is distributed through association with L2, and has an average latency of 12ns.

This chip's performance in AI Acceleration is rated at more than 6 TFLOPs per chip and over 200 TFLOPs in an IBM Z 4-chamber 4-chamber system. The internal Matrix array has 128 cells with 8-way FP-16 SIMD, high density multiplier and FPU accumulation.

While the Activation array consists of 32 cells with 8 dimensions FP16/FP-32 SIMD. The dual-chip configuration yields 116,000 inferences (1.1ms) while the 32-chip configuration yields 3,600,000 inferences (1.2ms).

The IBM Z Telum on-chip AI accelerator provides:

1. Very low and consistent inference latency
2. Calculating usability on a specific scale
3. Supports various AI models, from traditional ML to RNN and CNN
4. Security - provides enterprise-class virtualization and memory protection
5. Flexible scalability with future hardware and firmware updates.

The IBM Z Telum chip will be fabricated on Samsung's 7nm process and will have a die size of 530mm2. This chip will be geared towards enterprise & embedded workloads. A server system using Telum chip will be launched in the first half of 2022.