IBM announces new z16 systems with new Telum processor chips

By Ray Lucchesi, Silverton Consulting

One could say that IBM z or the mainframe was the first IT platform solution that made it big in the modern age. As such, it commanded significant market share and revenue in its time, not unlike where Apple, Microsoft, AWS et al, are today.

Platforms never really die, as long as R&D dollars are devoted to their upkeep, they can live forever. IBM continues to invest heavily in the z ecosystem, coming out with new z systems on a regular cadence.

The z16 and its Telum processor chip are just the latest iteration from IBM R&D to extend the life of this platform. The Telum processor chip was announced previously, at HotChips33 conference, last year.

IBM z16 Telum processor chip

The Telum chip is an 8 core, 5Ghz CPU, running the z/Architecture (complex) instruction set, built on 7nm process technology, with over 22B transistors/chip. A z16 processor complex can have up to 32 Telum chips, for a total of 256 CPU cores, of which up to 200 can be for customer use, the rest are devoted to system processing and spares.

Each Telum CPU core has 32MB of L2 cache which can be aggregated and shared across the 8 chip CPU cores, via a 320GB/sec ring (bus), to supply 256MB of virtual L3 cache. Similarly, the z16 offers up to 2GB of virtual L4 cache across the up to 32 Telum processor chips. This provides 1.5x more cache than the previous generation, z15 system

Furthermore, IBM has included on the **Telum Chip** an **on-chip AI accelerator capable of 1000s of parallel AI operations**. The Telum AI Accelerator is designed for low-latency high volume transactional workloads supporting up to 6 TFLOPs of FP16 precision compute per chip or ~200 TFLOPs per system. This supports near real time (sub msec. latency), inferencing for activities like, AI driven, credit card transaction validation. IBM mentioned that the z16 Telum AI Accelerator can scale up to 300B near real time inferences per day.

The z16 also has a special cryptographic processor that supports new quantum resistant crypto. NIST, the National Institute of Standards and Technology, initiated a worldwide process in 2016 to identify algorithms that are quantum resistant for public key cryptography which is reaching the end of the analysis phase. Today's approved public key algorithms are secure against

attacks from today's conventional and supercomputers but this will change when *Cryptographically Relevant Quantum Computers* using Shor's algorithm which will very easily break either ECC or RSA public key algorithms. This is why NIST, and others are developing new quantum resistant algorithms.

IBM and current cryptographic research have shown that **lattice-based mathematics** are quantum resistant. Algorithms which IBM helped develop are finalists in the NIST standardization process for next gen, quantum resistant cryptography. The z16 Crypto Express 8S (CEX8S) cryptographic coprocessor has special functionality to perform new lattice-based cryptography.

Significance

IBM for most of the last century led the way with AI. Watson and their Deep Blue chess playing systems were all leading-edge technology when they came out. Since then, Machine Learning and Deep Neural Networks (ML/DNNs) have emerged which have advanced state of the AI art way beyond those older technologies.

DNNs are all about vector (matrix) and floating-point arithmetic using single instruction, multiple data (SIMD) processing functionality. GPUs were invented to do this for graphics IBM has had SIMD support since the IBM z13. The Telum Integrated AI Accelerator, new with z16, provides acceleration for complex deep learning operations such as LSTM, GRU, RELU, etc. in addition to accelerating matrix multiplication operations.

The Telum AI Accelerator is focused on latency sensitive high volume transactional workloads. Although AI training is possible with the AI accelerator it is tailored for high volume inferencing. Customers today wishing to deploy AI in their high transactional activity have had very few options until z16 and its Telum AI accelerator came along, to support low latency, high volume inferencing.

But the AI marketplace is not standing still. Even lower precision FP arithmetic (FP8) are being used for DNN. training and inferencing. And while 6 TFLOPS per Telum chip or ~200 TFLOPS per z16 is impressive, current GPUs are capable of 20 to 30 TFLOPS/chip and can scale much higher when aggregated within systems

Nonetheless, IBM's z16 represents yet another significant step, keeping the z systems alive and well for the foreseeable future.
