The IBM Telum Processor: Integrated AI for Enterprise Workloads

IBM'S NEXT-GENERATION PROCESSOR INTEGRATES AI INTO IBM Z

INTRODUCTION

Transaction processing is the lifeblood of the modern enterprise, and many use IBM® Z for these mission-critical applications. Soon, these enterprises will be able to run accelerated Artificial Intelligence (AI) processing on those systems, providing real-time analytic insights. There are significant potential benefits when applying AI in situ where the data resides, and the transactions are taking place, making it possible to inference every transaction to enrich it with trusted, actionable insights.

Today, conducting AI processing during transactions relies on using the Z cores for machine learning or offloading the processing to other servers with GPU accelerators for batch processing. Of course, the latter adds high cost and injects latencies that inhibit real-time analytics and expose an additional security surface to intrusion.

This paper investigates the upcoming Telum processor that IBM previewed at the recent Hot Chips conference, which IBM believes will enable new innovations in the world's most venerable and stable enterprise computing platform.

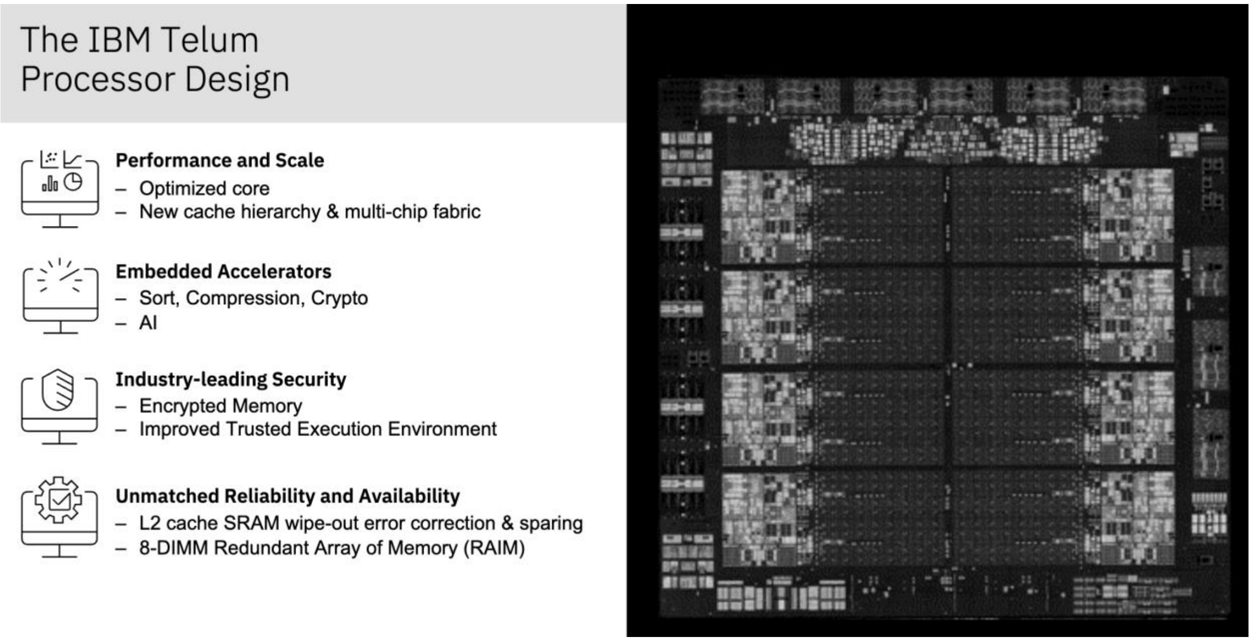
ENTERPRISE APPLICATIONS AND AI

Many enterprise workloads are beginning to leverage machine learning to improve robustness and customer service. Credit card fraud detection and trading settlements are examples of how machine learning delivers excellent results on Z cores. Many enterprises are taking the next step with deep learning, a.k.a. deep neural networks (DNNs), which significantly improve accuracy. However, the computational characteristics of DNNs require significant parallel processing and reduced precision to deliver the required performance. Hence the IBM "Telum" processor.

THE NEW Z PROCESSOR WITH ON-BOARD AI ACCELERATION

The next-generation Z processor, named "Telum," is designed to provide real-time AI insights to enterprise workloads. Along with improvements in performance, scalability, security, and availability, the capability to run low-latency DNN inference processing in real-time puts mainframe applications on a renewed path of innovation that will benefit IBM Mainframe owners for years to come. While IBM previewed the chip at the Hot

Chips conference, IBM has not set a date for availability. We assume the next-generation IBM Z will use the Telum processor sometime in 2022.



The IBM Telum Processor Design

Performance and Scale

- Optimized core
- New cache hierarchy & multi-chip fabric

Embedded Accelerators

- Sort, Compression, Crypto
- AI

Industry-leading Security

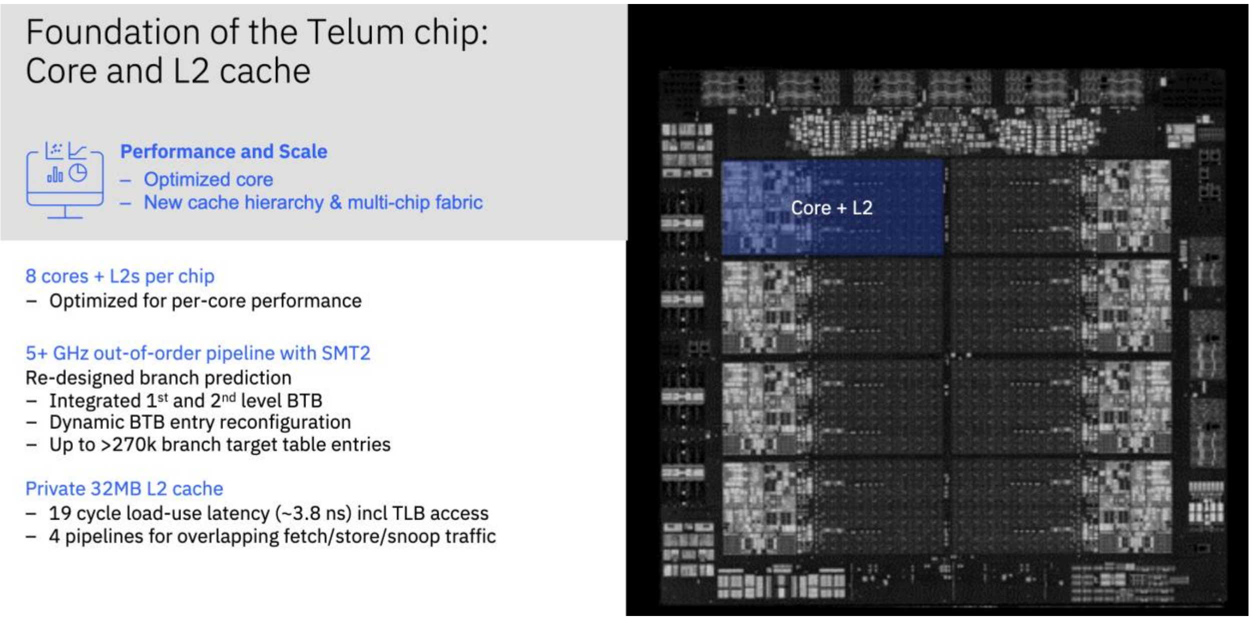
- Encrypted Memory
- Improved Trusted Execution Environment

Unmatched Reliability and Availability

- L2 cache SRAM wipe-out error correction & sparing
- 8-DIMM Redundant Array of Memory (RAIM)

FOUNDATION OF THE TELUM CHIP: CORE AND L2 CACHE WITH CACHE PERSISTENCE

Before diving into the computation accelerator, let’s look at the 7nm Telum’s new core and cache design. The design team decided to add more cache to each core while reducing the number of cores per chip from 12 to 8, and offering a dual-chip module packaging. This design frees up enough die area to increase the Level 2 cache by four-fold, to 32 MB from the z15’s 8MB (4 for instructions, and 4 for data). The larger cache should significantly increase the per-core performance for most applications, especially when coupled with the innovative L3/L4 virtual cache, which spans all cores and chips across the 32-chip fabric.



Foundation of the Telum chip: Core and L2 cache

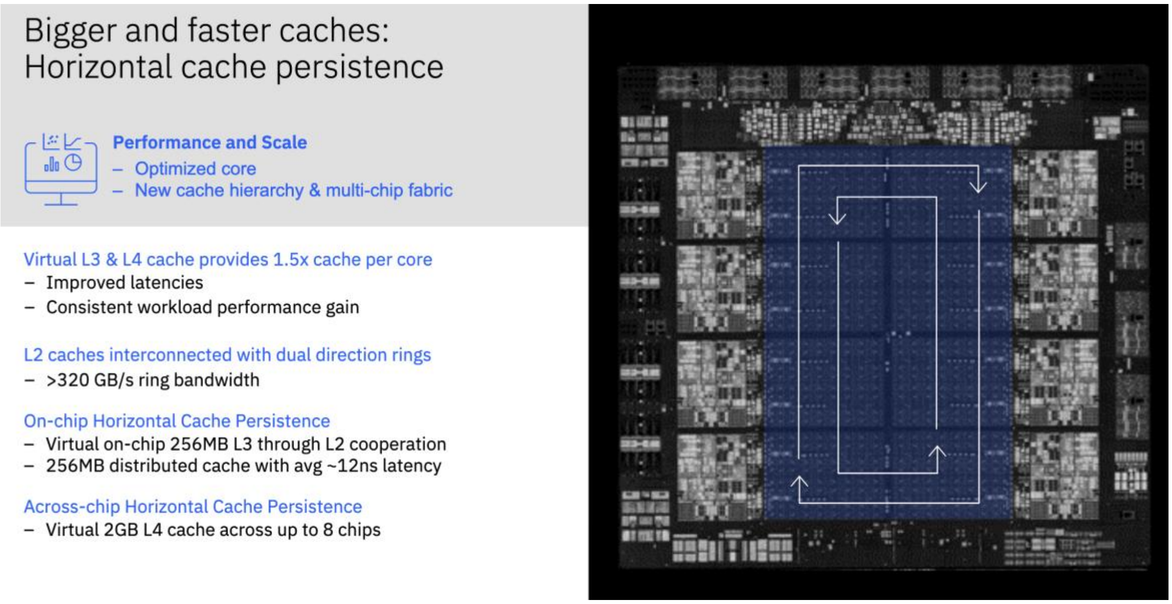
Performance and Scale
- Optimized core
- New cache hierarchy & multi-chip fabric

8 cores + L2s per chip
- Optimized for per-core performance

5+ GHz out-of-order pipeline with SMT2
Re-designed branch prediction
- Integrated 1st and 2nd level BTB
- Dynamic BTB entry reconfiguration
- Up to >270k branch target table entries

Private 32MB L2 cache
- 19 cycle load-use latency (~3.8 ns) incl TLB access
- 4 pipelines for overlapping fetch/store/snoop traffic

The L3 and L4 virtual design across the fabric provides 50% more cache per core, with improved latencies. The idea is that software and microcode still see two separate caches. These caches are shared and distributed across all eight cores via a 320 GB/s ring and across the integrated fabric. Horizontal cache persistence should further reduce cache misses as well. Specifically, when a cache line is ejected, the system looks for available cache capacity on other caches, first on the chip and then even across the 32-chip fabric. We wonder if different architectures such as Power will adopt a similar scheme in the future.



Bigger and faster caches: Horizontal cache persistence

Performance and Scale
- Optimized core
- New cache hierarchy & multi-chip fabric

Virtual L3 & L4 cache provides 1.5x cache per core
- Improved latencies
- Consistent workload performance gain

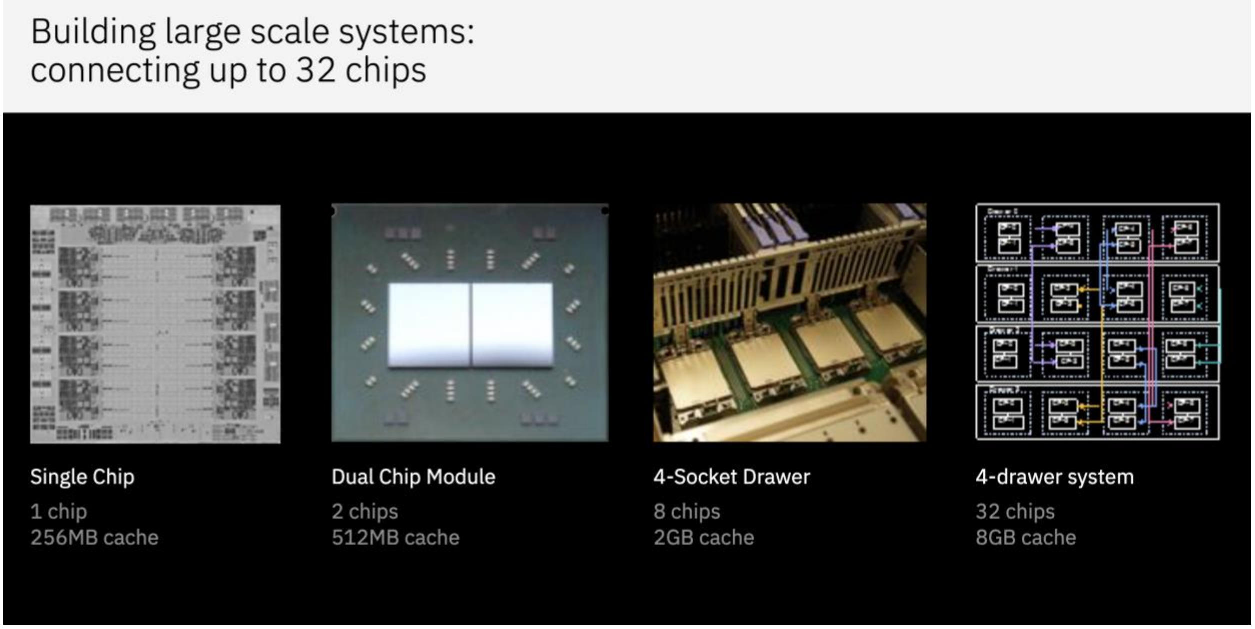
L2 caches interconnected with dual direction rings
- >320 GB/s ring bandwidth

On-chip Horizontal Cache Persistence
- Virtual on-chip 256MB L3 through L2 cooperation
- 256MB distributed cache with avg ~12ns latency

Across-chip Horizontal Cache Persistence
- Virtual 2GB L4 cache across up to 8 chips

BUILDING LARGE SCALE SYSTEMS: NEW FABRIC INTERCONNECTS UP TO 32 CHIPS

IBM engineers envisioned that a faster, low-latency interconnect could also improve bandwidth, and the fabric controller on the Telum chip interconnects up to 32 chips in a 4-drawer system. This custom fabric enables strong availability and very low latency across a flat topology and creates an 8-GB cache pool across the complex.



Building large scale systems:
connecting up to 32 chips

Single Chip	Dual Chip Module	4-Socket Drawer	4-drawer system
1 chip 256MB cache	2 chips 512MB cache	8 chips 2GB cache	32 chips 8GB cache

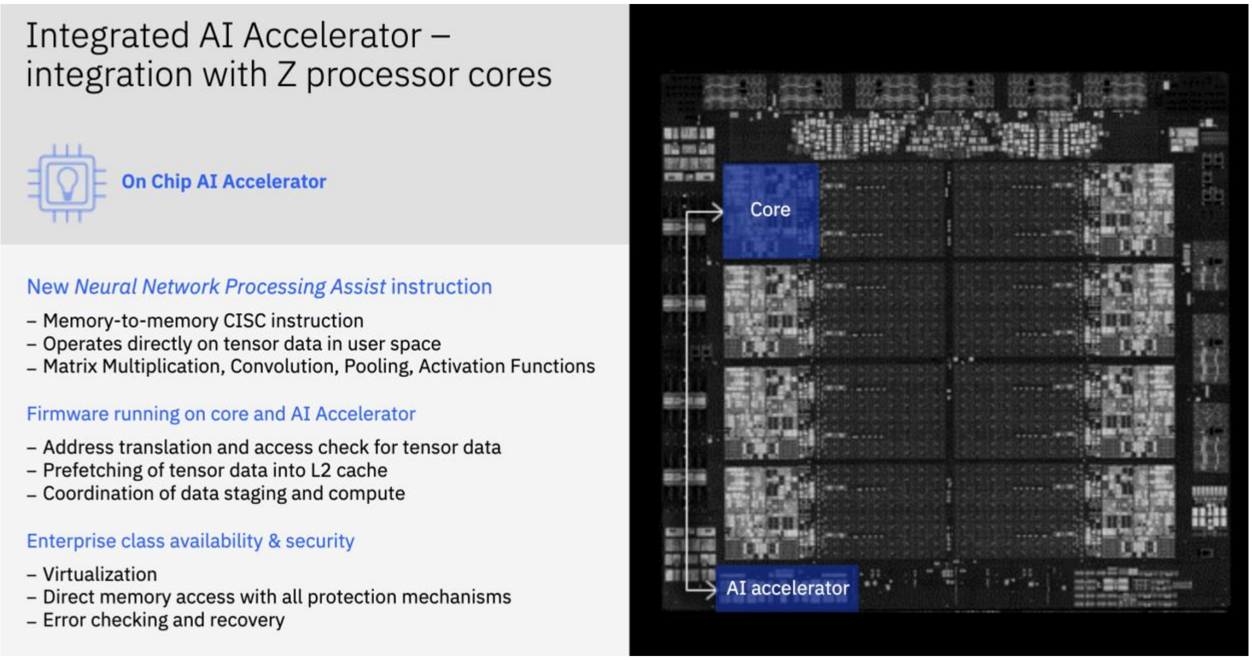
INTEGRATED ACCELERATOR FOR PARALLEL ALGORITHMS IN ENTERPRISE WORKLOADS

As one can see, the new Telum offers significant advantages in both the cores, cache, and interconnect fabric. But the real excitement for Z owners will probably stem from the integrated on-chip parallel processing accelerator for sort, crypto, compression, and especially AI inference processing. We believe this accelerator marks a significant opportunity for application modernization and expansion on Z.

For the first time, Z applications can infuse real-time AI insights into every transaction for customer insights, fraud detection, supply chain optimization, and many other applications. In addition, the AI engine can improve IT operations, including security, privacy, workload placement, database query plans, and anomaly detection. AI is adding new functionality to nearly all modern applications, and the inclusion of an AI engine to the Z processor delivers inference directly on the Z processor. Let’s look further into the design.

Integrated AI Accelerator – Integration with Z Processor Cores

The AI accelerator is fully integrated with the Z cores to minimize latency and maximize throughput and security with minimal and consistent latency. Each Telum die connects the accelerator to all eight cores and cache memory with a Neural Network Processing Assist instruction, enabling matrix multiplication, convolution, pooling, and activation functions. In a departure from other AI acceleration schemes we have explored, the Telum processor uses firmware on the Z core and the AI engine to control execution and data movement from the cache. By using firmware, the platform can support new AI models and computational techniques. Telum is the only accelerator we have reviewed that enables this level of adaptability (outside of FPGAs, of course).



Integrated AI Accelerator – integration with Z processor cores

On Chip AI Accelerator

New *Neural Network Processing Assist* instruction

- Memory-to-memory CISC instruction
- Operates directly on tensor data in user space
- Matrix Multiplication, Convolution, Pooling, Activation Functions

Firmware running on core and AI Accelerator

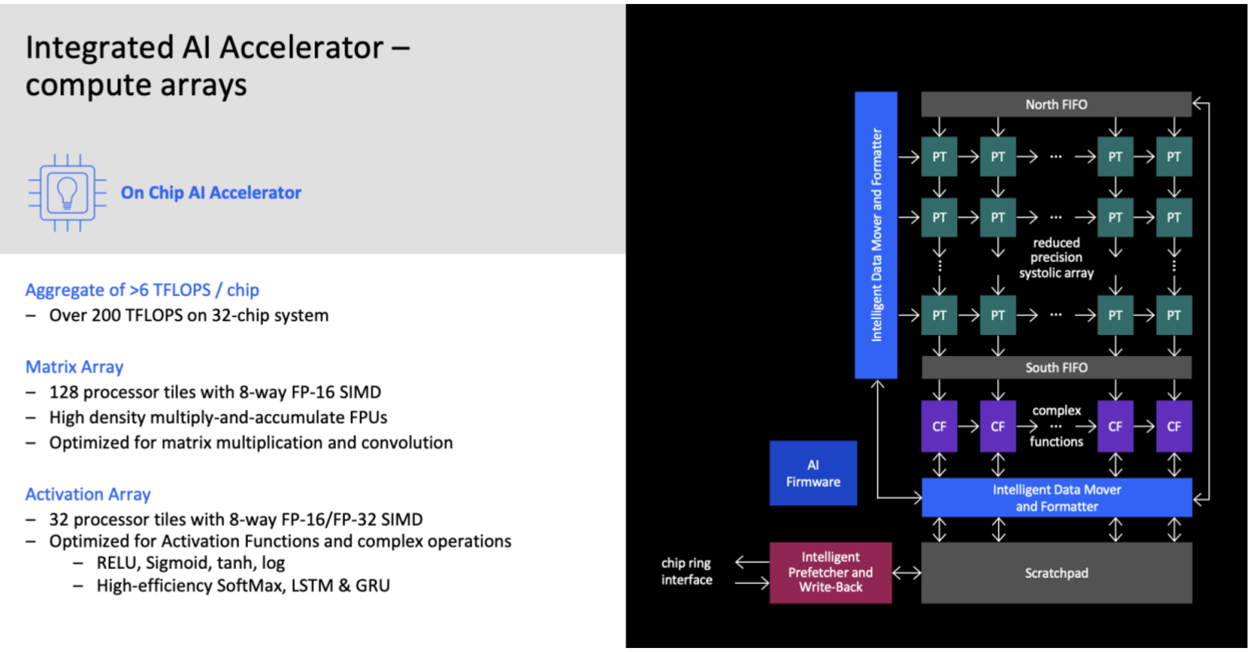
- Address translation and access check for tensor data
- Prefetching of tensor data into L2 cache
- Coordination of data staging and compute

Enterprise class availability & security

- Virtualization
- Direct memory access with all protection mechanisms
- Error checking and recovery

Integrated AI Accelerator – Compute Arrays and Data Movers

The accelerator itself delivers an aggregate of over 6 TFLOPS of 16-bit floating-point throughput per chip to scale up to roughly 200 TFLOPS per system. 1024 processor tiles in a systolic array make up the matrix array, and 256 fp16/32 tiles make up the accelerator for computing activations and include built-in functions for RELU, tanH, and log. The platform also provides enterprise-class availability and security, as one should expect in a Z, with virtualization, error checking/recovery, and memory protection mechanisms. While 6 TFLOPS does not sound impressive, keep in mind that this accelerator is optimized for transaction processing. Most data are in floating-point and are highly structured, unlike in voice or image processing. Consequently, we believe this accelerator will deliver adequate performance and is undoubtedly much faster than offloading to another GPU-equipped server or running the inference on a Z core. The latency of off-platform inference can cause transactions to time out, and inference does not complete.



Integrated AI Accelerator – compute arrays

On Chip AI Accelerator

Aggregate of >6 TFLOPS / chip
- Over 200 TFLOPS on 32-chip system

Matrix Array
- 128 processor tiles with 8-way FP-16 SIMD
- High density multiply-and-accumulate FPUs
- Optimized for matrix multiplication and convolution

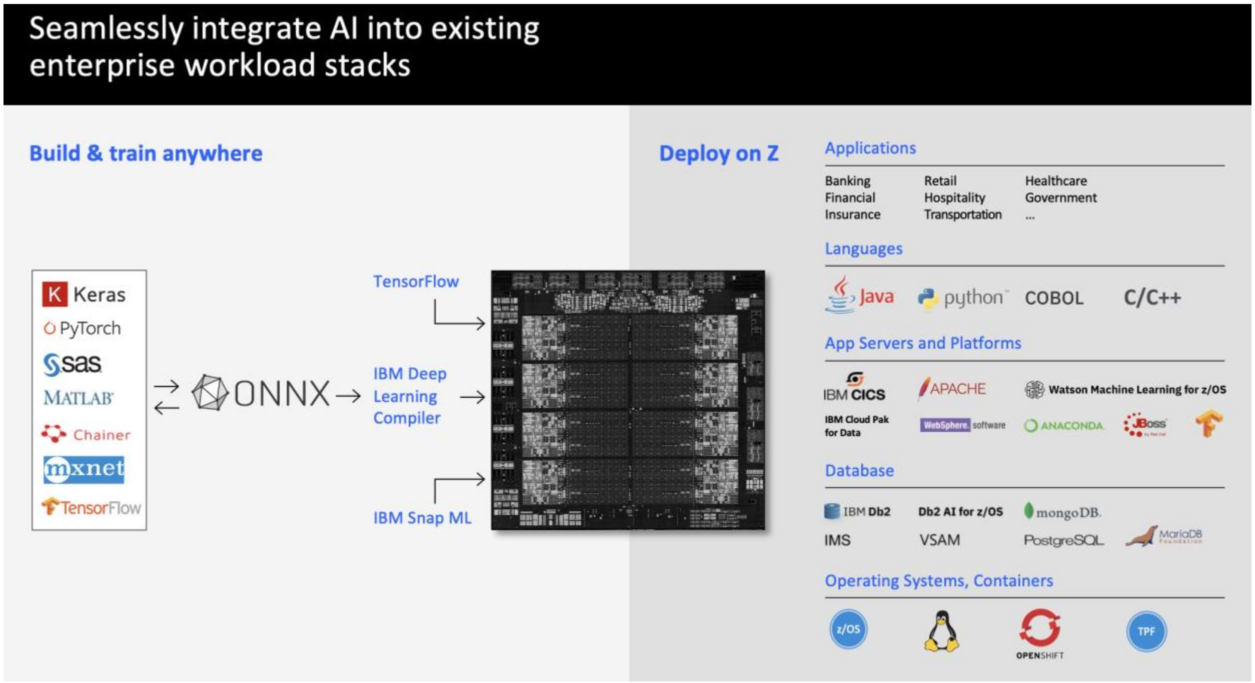
Activation Array
- 32 processor tiles with 8-way FP-16/FP-32 SIMD
- Optimized for Activation Functions and complex operations
  - RELU, Sigmoid, tanh, log
  - High-efficiency SoftMax, LSTM & GRU

Intelligent Data Mover and Formatter
North FIFO
PT
reduced precision systolic array
South FIFO
CF
complex functions
AI Firmware
Intelligent Data Mover and Formatter
chip ring interface
Intelligent Prefetcher and Write-Back
Scratchpad

The accelerator includes a data movement controller for intelligent pre-fetch and writeback into the Telum cache. This logic moves data to/from a scratchpad memory at 60 GB/s, supporting concurrent load, execution, and writeback, and can transform data on the fly.

PROGRAMMING THE TELUM AI ACCELERATOR

The user development process is straightforward, using popular AI frameworks such as TensorFlow and Pytorch and an IBM Deep Learning compiler. No changes to the models are required. Models can be trained anywhere and then converted to the industry standard ONNX exchange format". The IBM Deep Learning Compiler (DLC) then compiles the model from ONNX and creates the binaries for the Telum cores and the AI accelerator. In addition to optimizing AI models for Telum via ONNX and DLC, TensorFlow models will be transparently accelerated without having to be converted to ONNX and complied with the IBM Deep Learning Compiler. Both the TensorFlow plug-ins and the underlying run-time computational libraries will be open source, facilitating community collaboration. The IBM Snap Machine Learning library (a library for speeding up popular ML algorithms) has also been enhanced to provide users with transparent acceleration by leveraging the Telum processor. The system can be running either z/OS, Linux, containers like z/CX, or TPF.



Seamlessly integrate AI into existing enterprise workload stacks

Build & train anywhere

Keras
PyTorch
sas
MATLAB
Chainer
mxnet
TensorFlow

ONNX

TensorFlow
IBM Deep Learning Compiler
IBM Snap ML

Deploy on Z

Applications

Banking	Retail	Healthcare
Financial	Hospitality	Government
Insurance	Transportation	...

Languages

Java, python, COBOL, C/C++

App Servers and Platforms

IBM CICS, APACHE, Watson Machine Learning for z/OS, IBM Cloud Pak for Data, WebSphere software, ANACONDA, JBoss

Database

IBM Db2, Db2 AI for z/OS, mongoDB, IMS, VSAM, PostgreSQL, MariaDB Foundation

Operating Systems, Containers

z/OS, OPENSHIFT, TPF

Enterprise-class availability & security

Enabling Deep Learning processing on the Z gives users access to state-of-the-art AI running on the industry’s most secure and trusted computing platform. Encrypted memory, performance-optimized Trusted execution and enhanced bus & memory error recovery all ensure that the analytics are running in a secure environment. The alternative offloading of data to run AI applications on a less robust platform is not an attractive option and introduces significant and costly risks to the enterprise.

Example Use Cases

Let’s look at a few use cases that exemplify the value and benefits of the integrated AI accelerator on Telum. Many are customers have been using machine learning but believe the next step is deep learning AI. The first example is credit card transaction processing. Large banks often use machine learning on the Z to enable real-time fraud detection for credit card transactions. On-board deep learning processing could enable deeper analytics and increased transaction coverage. Being able to run the analytics on-platform should reduce cost and limit security exposures.

A second example is the clearing and settlement of stock transactions after the market closes, so this would be a batch process. Using AI to predict which trades or transactions have high-risk exposures allows the system to propose a more efficient settlement process. The expedited remediation of questionable transactions can prevent costly consequences, regulatory violations, and negative business impacts.

PERFORMANCE, SCALABILITY, AND LATENCY

The scalability of the solution is critical to achieving the client's business objectives. The IBM development team has tested decision tree models and recurrent neural networks, measuring scalability as an increasing number of cores processing the model to deliver higher throughput. The results show that the shared AI accelerator on the Telum chip maintains linearly low (~1ms) latencies as throughput increases, up to the 32-chip limit of the internal fabric. This latency compares quite favorably to the current state-of-the-art in cloud AI processing, which typically operates at a 5ms threshold to maximize batch throughput at acceptable latency.

CONCLUSIONS

The upcoming Telum processor for Z and LinuxONE will provide a vital step function to enterprise workloads seeking to take advantage of the benefits of using deep learning neural networks. Real-time scoring of credit card transactions is a prime example where the addition of an AI accelerator on the Z processor will transform business practices, perhaps even analyzing 100% of transactions for credit card fraud. In addition to the AI accelerator, the Telum design brings features such as new cache architecture that should improve throughput, perhaps significantly, all while maintaining or improving performance, scalability, security, and availability. We believe Telum may be the most impactful Z processor in recent memory.

Important Information About This Paper

Author: Karl Freund, Founder, and Principal Analyst at Cambrian-AI Research

Inquiries:

Contact us if you would like to discuss this report, and Cambrian-AI Research will respond promptly.

CITATIONS

This paper can be cited by accredited press and analysts but must be cited in-context, displaying the author's name, author's title, and "Cambrian-AI Research." Non-press and non-analysts must receive prior written permission from Cambrian-AI Research for any citations.

LICENSING

DISCLOSURES

This document was developed with IBM funding. Although the document may utilize publicly available material from various vendors, including IBM, it does not necessarily reflect the positions of such vendors on the issues addressed in this document.

DISCLAIMER

The information presented in this document is for informational purposes only and may contain technical inaccuracies, omissions, and typographical errors. Cambrian-AI Research disclaims all warranties as to the accuracy, completeness, or adequacy of such information and shall have no liability for errors, omissions, or inadequacies in such information. This document consists of the opinions of Cambrian-AI Research and should not be construed as statements of fact. The opinions expressed herein are subject to change without notice.

Cambrian-AI Research provides forecasts and forward-looking statements as directional indicators and not as precise predictions of future events. While our forecasts and forward-looking statements represent our current judgment on what the future holds, they are subject to risks and uncertainties that could cause actual results to differ materially. You are cautioned not to place undue reliance on these forecasts and forward-looking statements, which reflect our opinions only as of the date of publication for this document. Please keep in mind that we are not obligating ourselves to revise or publicly release the results of any revision to these forecasts and forward-looking statements in light of new information or future events.