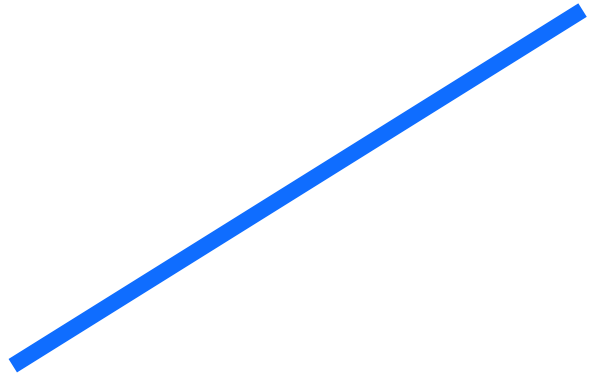
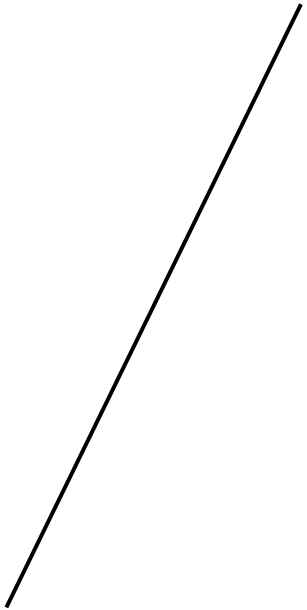
Real-time AI for Enterprise Workloads: the IBM Telum Processor

Dr. Christian Jacobi
IBM Distinguished Engineer
Chief Architect Z Processor Design





IBM

You probably used IBM Z today!

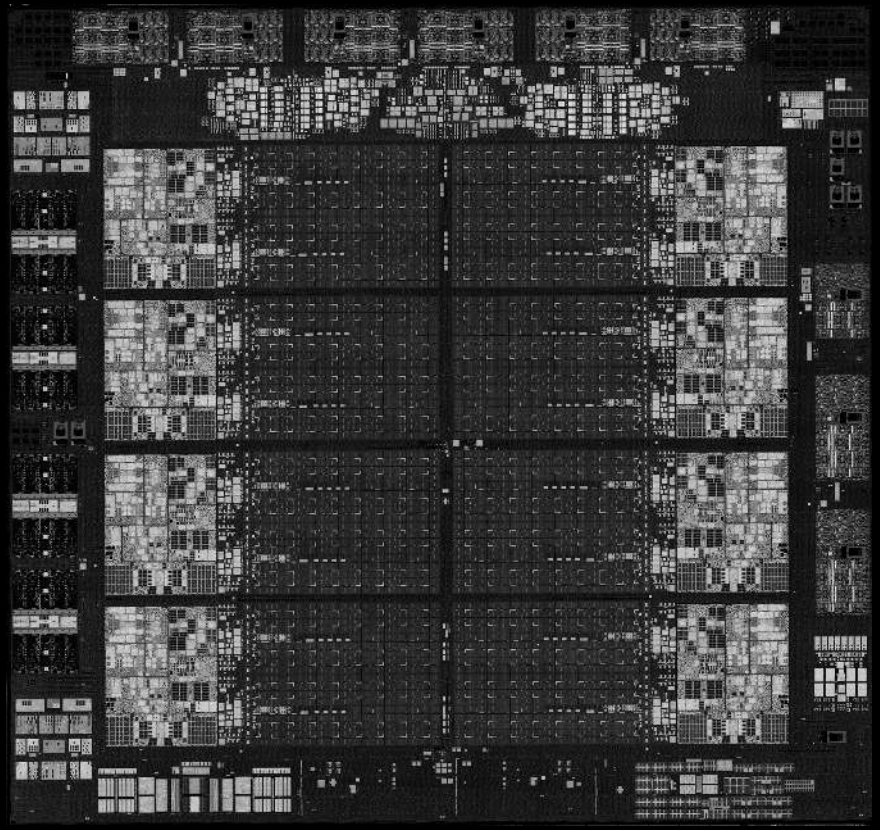
The IBM Telum Processor Design

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

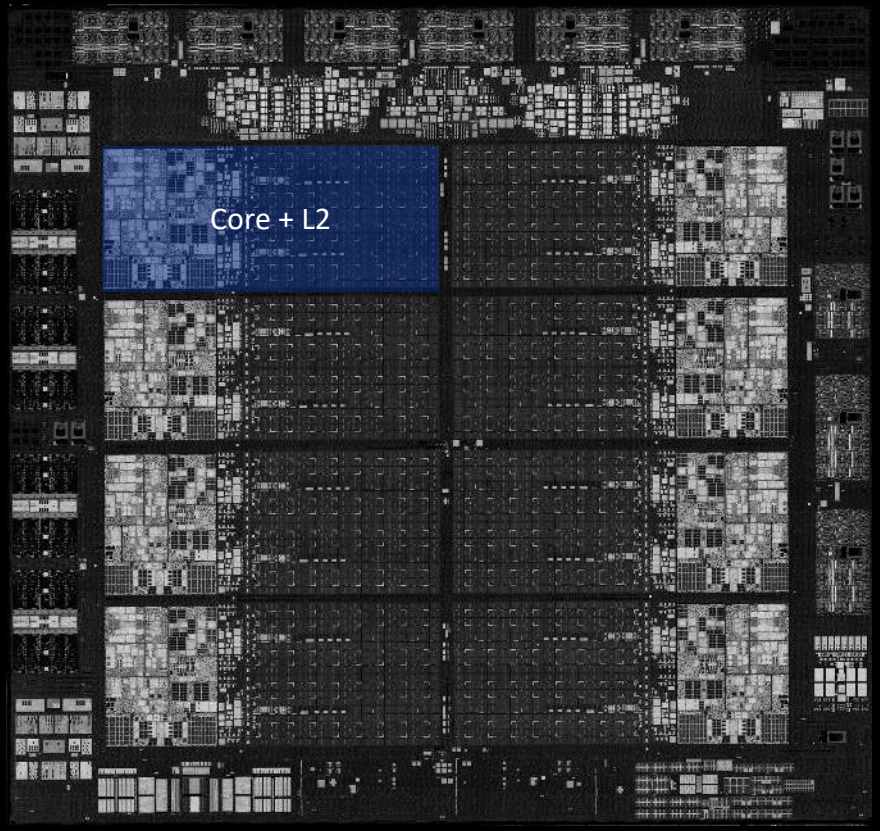
Performance and Scale
- Optimized core
- New cache hierarchy & multi-chip fabric

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



Bigger and faster caches: Horizontal cache persistence

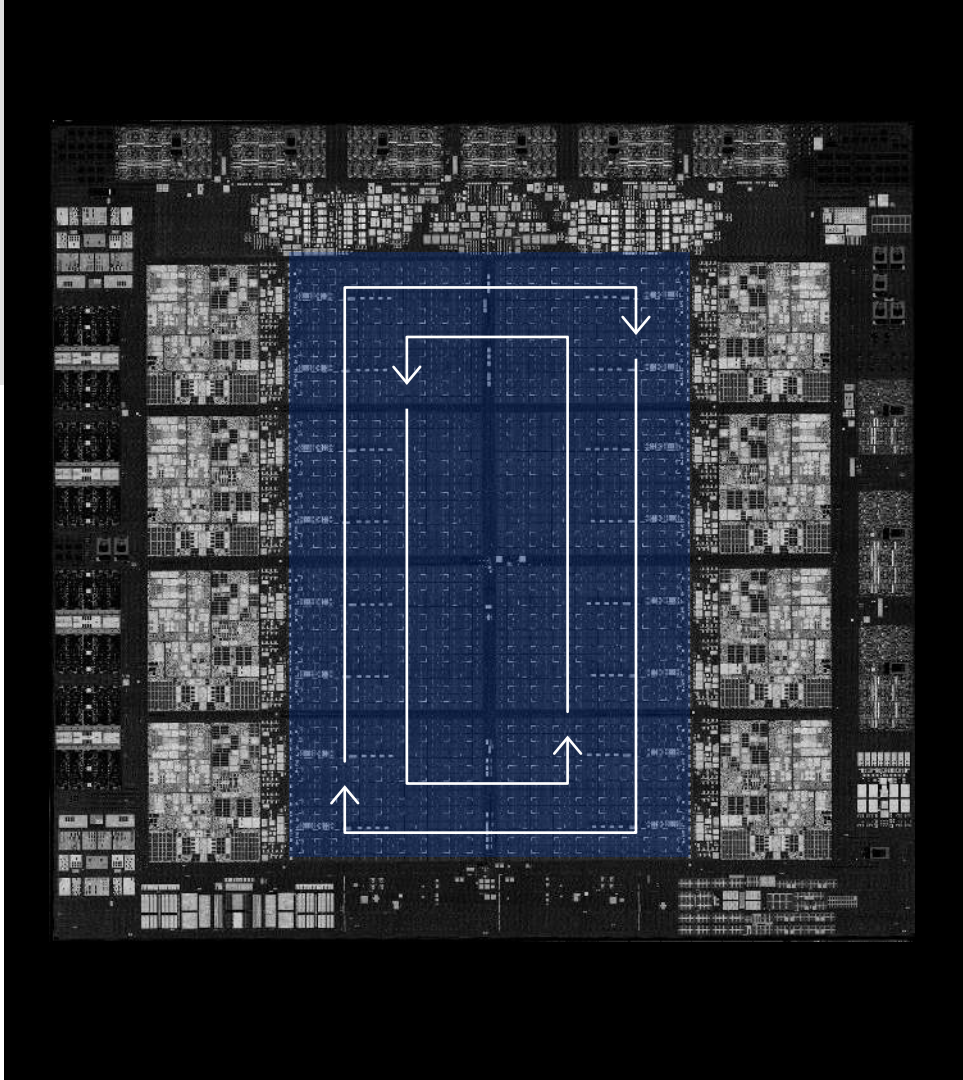
Performance and Scale
- Optimized core
- New cache hierarchy & multi-chip fabric

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

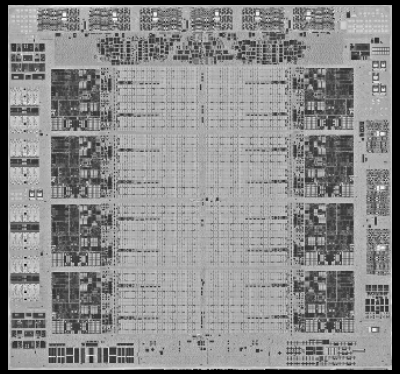
L2 caches interconnected with dual direction rings
- >320 GB/s ring bandwidth

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

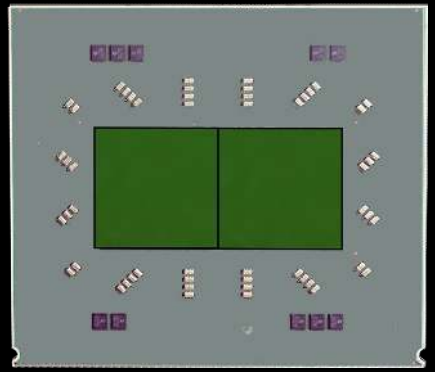


Building large scale systems: connecting up to 32 chips



Single Chip

1 chip
256MB cache

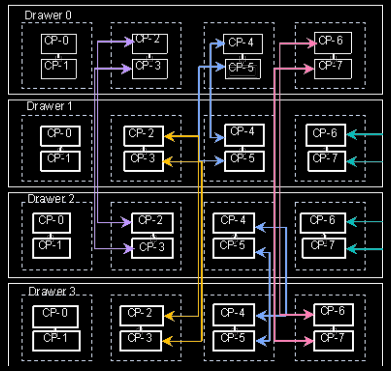


Dual Chip Module

2 chips
512MB cache

4-Socket Drawer

8 chips
2GB cache



4-drawer system

32 chips
8GB cache

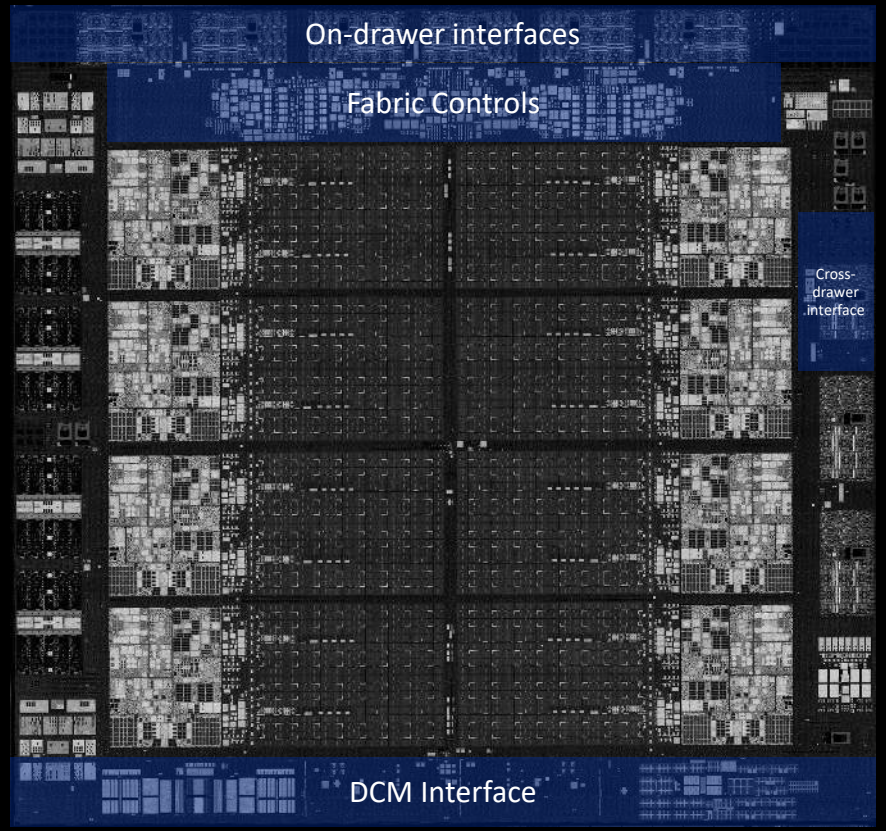
Building large scale systems: Fabric & interface optimizations

Performance and Scale

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

Optimization for latency and bandwidth at every layer

- DCM uses 2 cycle synchronous transfer for minimal latency
- Flat topology within drawer improves latency over z15

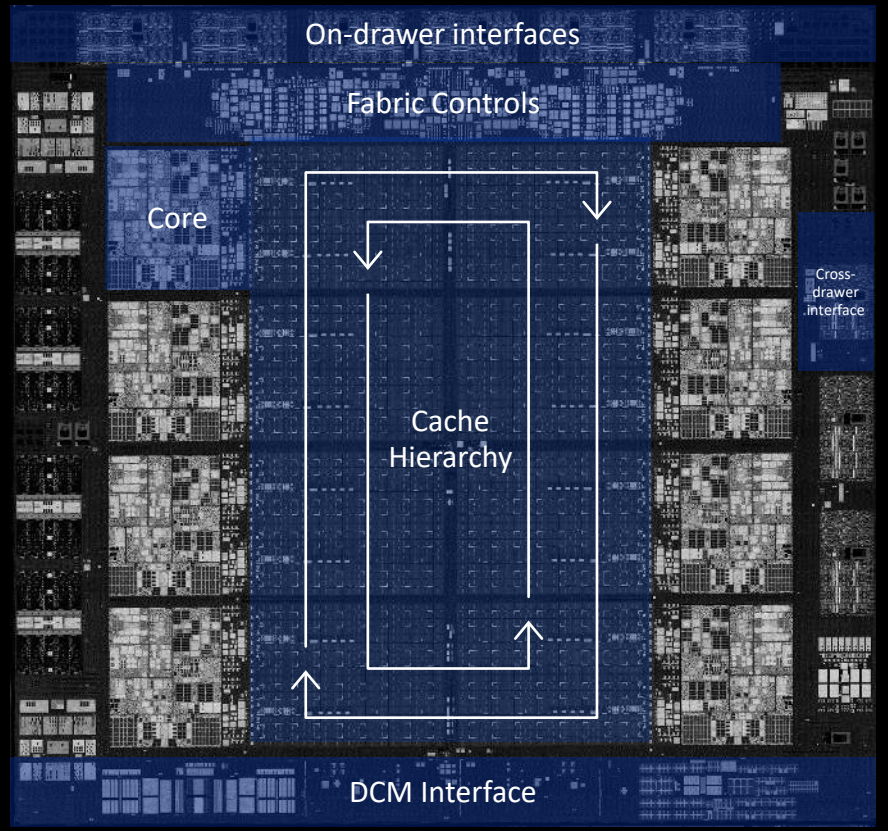


Enterprise workload performance

Performance and Scale

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

Over 40% per socket performance growth



Performance projection based upon pre-silicon engineering analysis of Telum DCM socket vs z15 processor socket

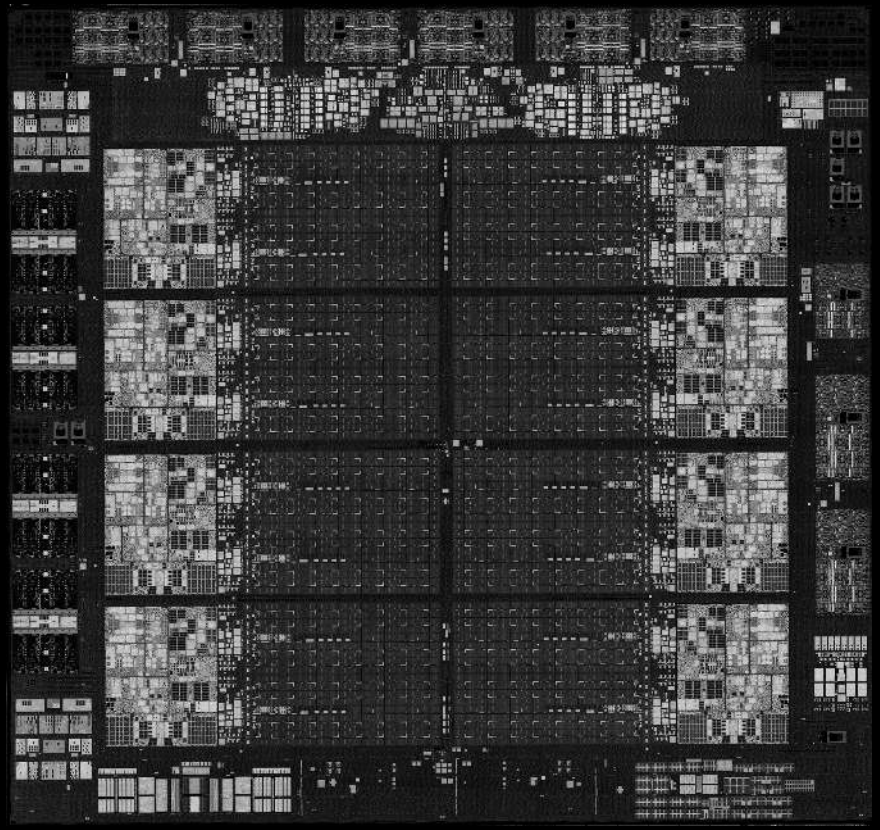
The IBM Telum Processor Design

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)



World-class AI inference platform for enterprise workloads

Business Insights

- Fraud detection
- Customer behavior prediction
- Supply chain optimization

Intelligent Infrastructure

- Workload placement
- Database query plans
- Anomaly detection for security

Maximize AI value with low & consistent latency, enabling real-time application

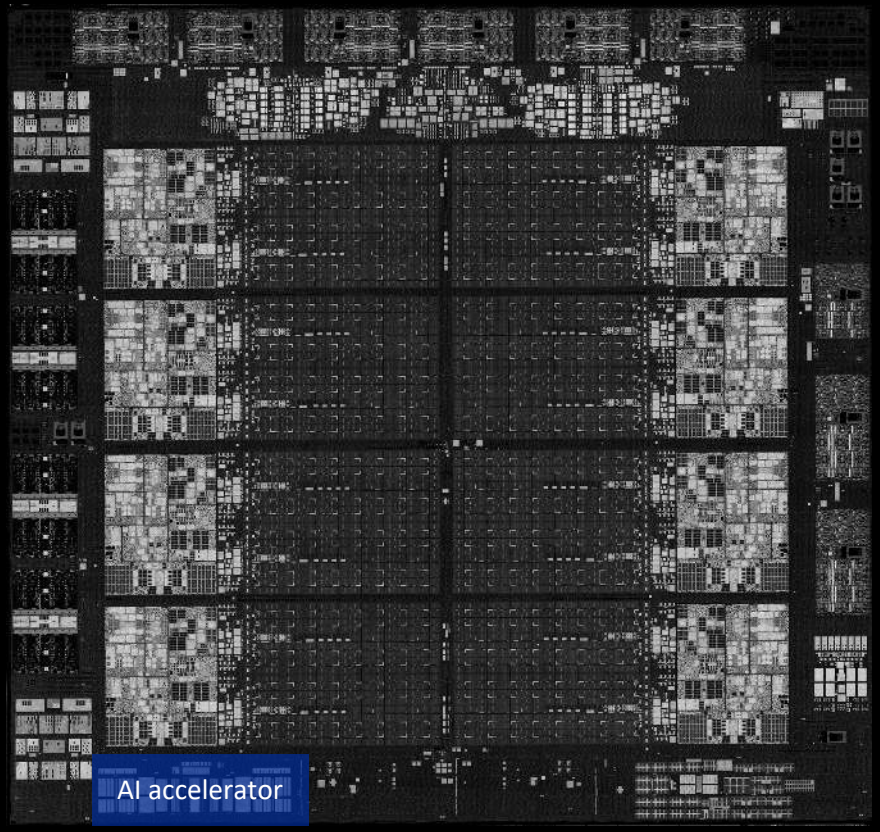
Minimize security exposure for sensitive data

Inference tasks directly embedded into transaction workload on IBM Z

Embedded AI Inference with central low-latency accelerator

Centralized On-chip accelerator shared by all cores

- Very low and consistent inference latency
- Compute capacity for utilization at scale
- Variety of AI models ranging from traditional ML to RNNs and CNNs
- Security – provide enterprise-grade memory virtualization and protection
- Extensibility with future firmware and hardware updates



Integrated AI Accelerator – integration with Z processor cores

On Chip AI Accelerator

New *Neural Network Processing Assist* instruction

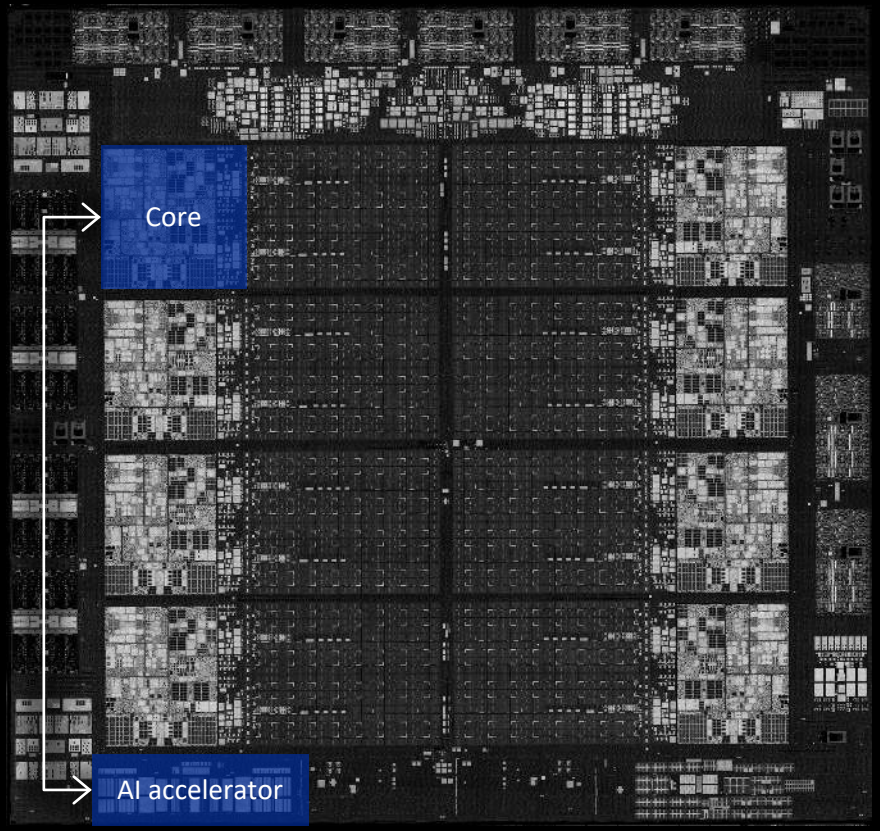
- Memory-to-memory CISC instruction
- Operates directly on tensor data in user space
- Matrix Multiplication, Convolution, Pooling, Activation Functions

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



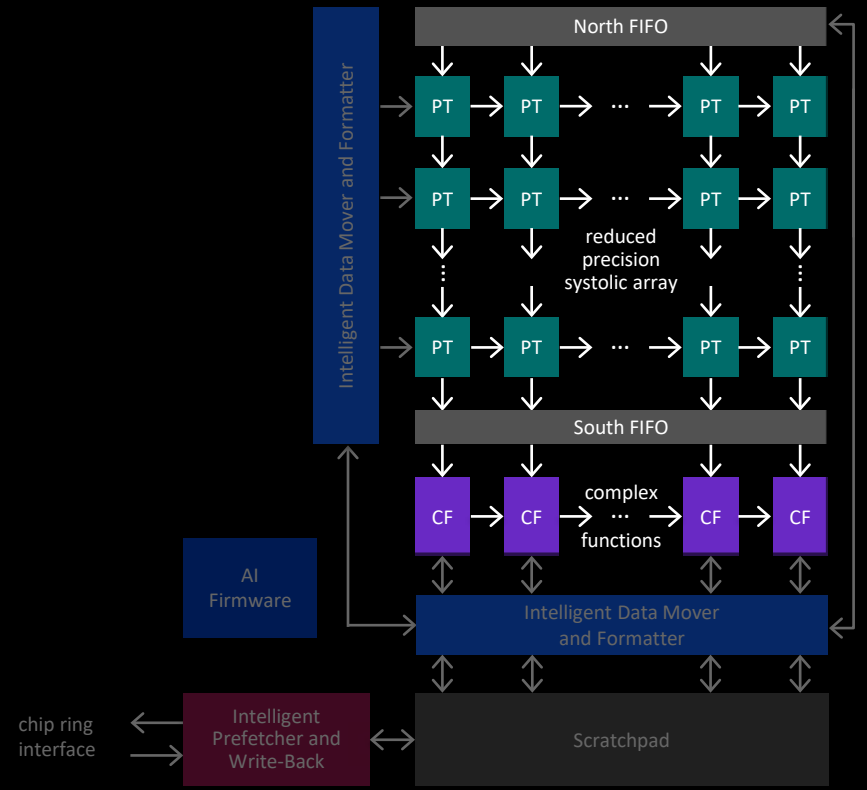
Integrated AI Accelerator – compute arrays

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



Integrated AI Accelerator – data movers

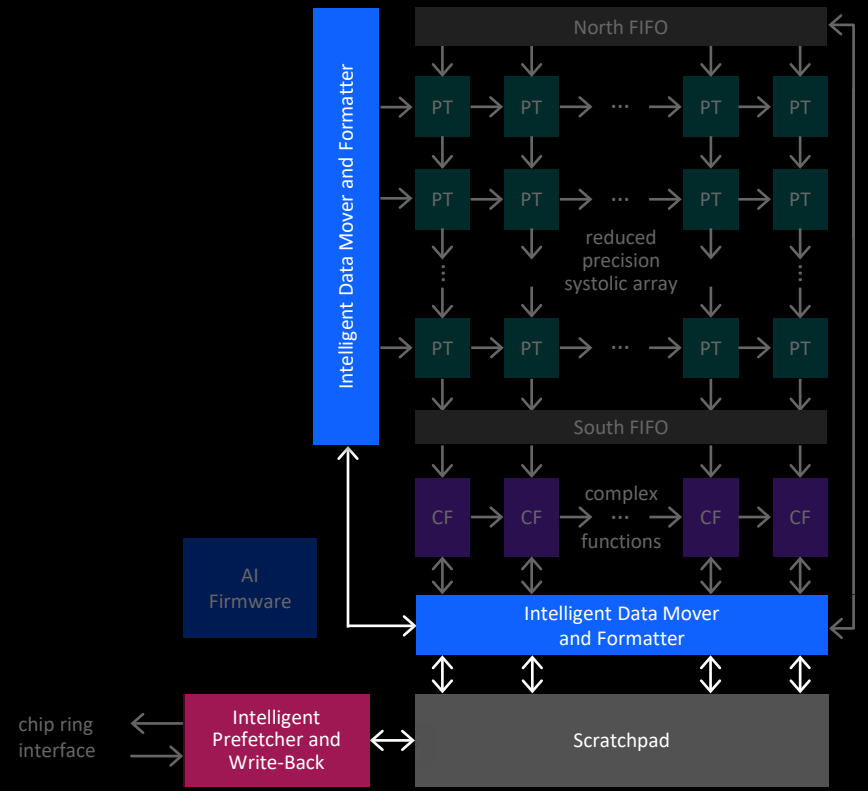
On Chip AI Accelerator

Intelligent Prefetcher and Write-Back

- 120+ GB/s read bandwidth to internal scratchpad
- 80+ GB/s store bandwidth
- Multi-zone scratchpad for concurrent data load, execution and write-back

Intelligent Data Mover and Formatter

- 600+ GB/s bandwidth
- Format and prepare data on the fly for compute and write-back



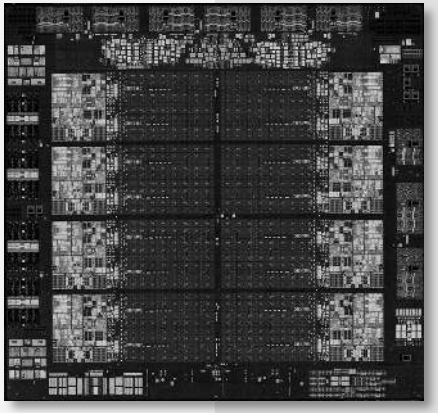
Seamlessly integrate AI into existing enterprise workload stacks

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

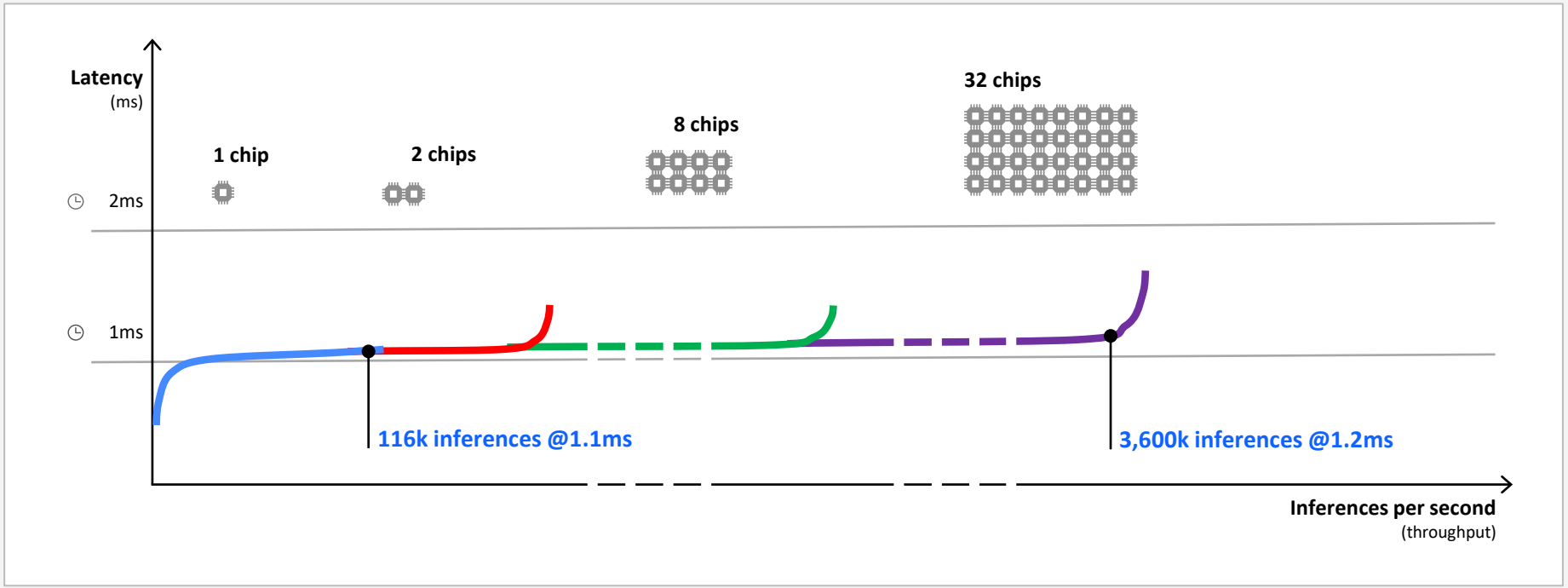
IBM Db2
Db2 AI for z/OS
mongoDB.
IMS
VSAM
PostgreSQL
MariaDB

Operating Systems, Containers

z/OS
TPF

AI Accelerator performance

RNN multi-layer model for Credit Card Fraud – proxy model developed with global bank



Performance projection from cycle accurate simulation model on RNN proxy for Credit Card Fraud detection.

Summary

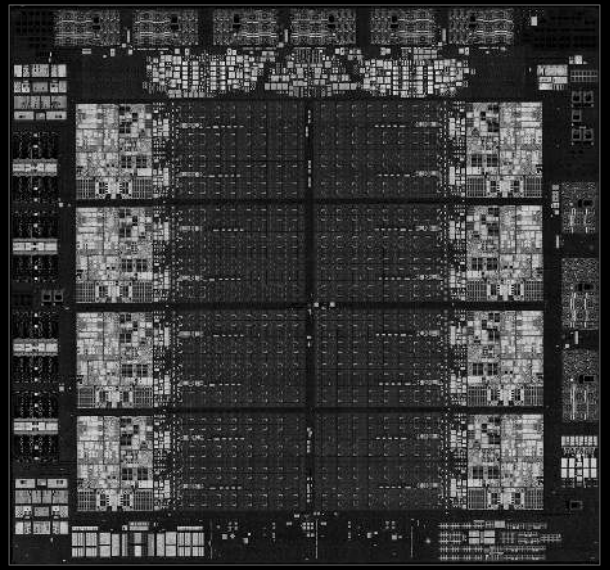
Next generation Z processor is optimized to run enterprise workloads with embedded real time AI insights.

Performance and Scale

Security

Availability

Low-latency accelerator for AI



IBM Telum chip

7nm Samsung technology

530sqmm chip size

22.5 Billion transistors

5+ GHz base clock frequency