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

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

IBM Z

# You probably used IBM Z today!

# The IBM Telum Processor Design

**Performance and Scale**

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

**Embedded Accelerators**

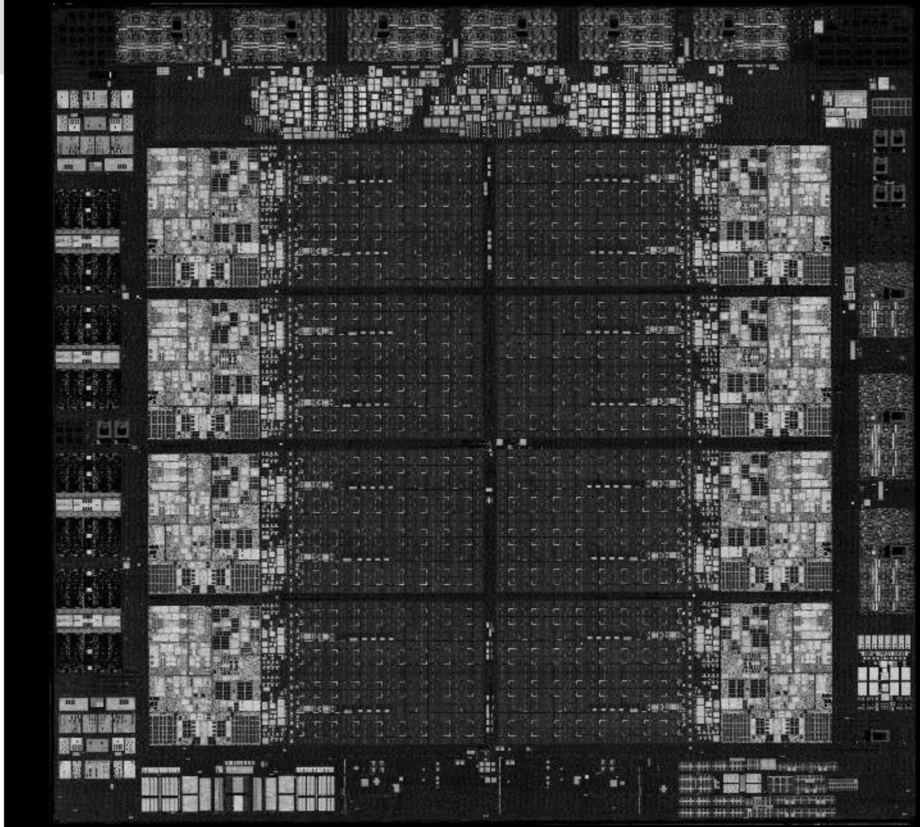
- Sort, Compression, Crypto
- AI

**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)



# The IBM Telum Processor Design

**Performance and Scale**

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

**Embedded Accelerators**

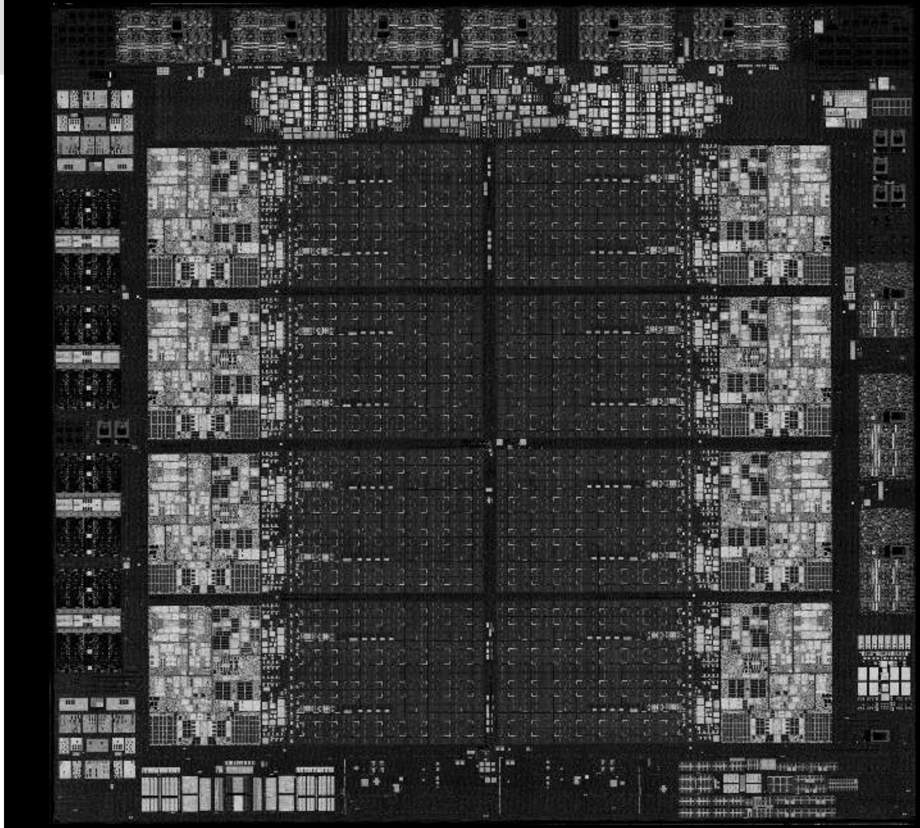
- Sort, Compression, Crypto
- AI

**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

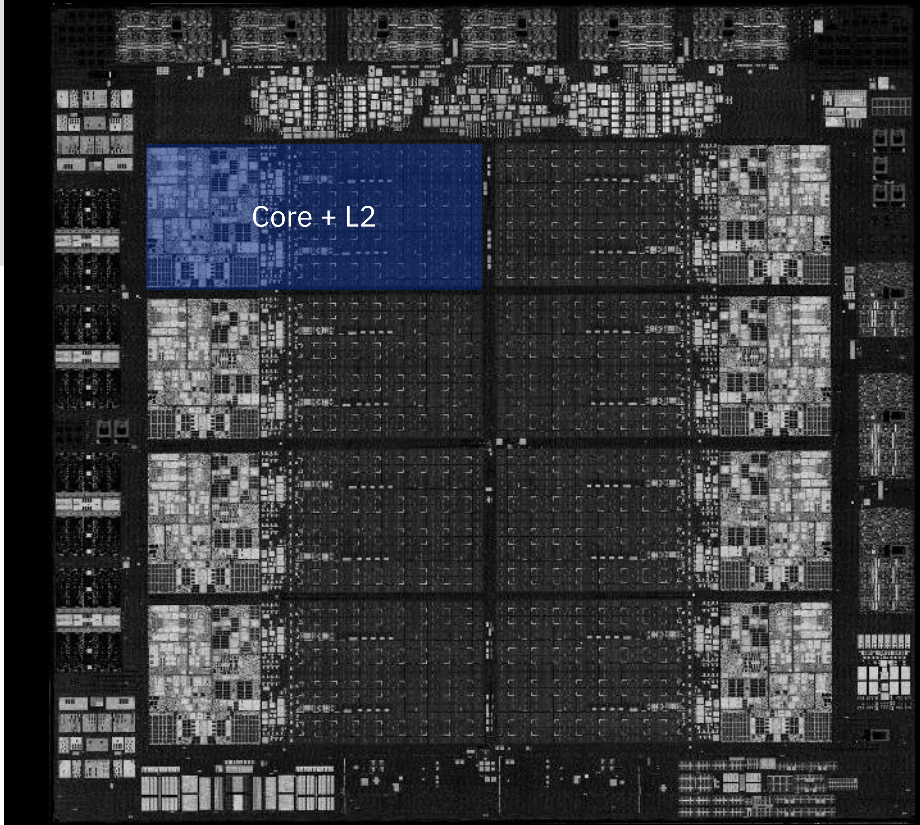
5+ GHz out-of-order pipeline with SMT2

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

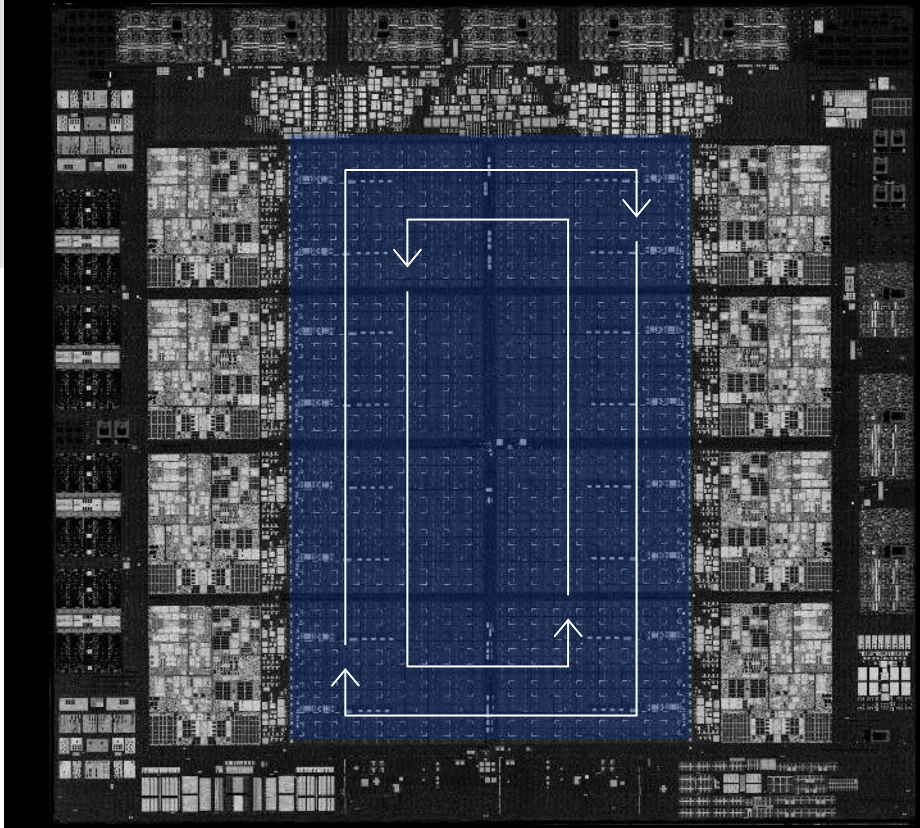
- >320 GB/s ring bandwidth

On-chip Horizontal Cache Persistence

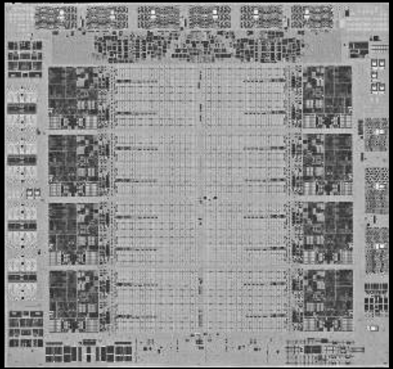
- Virtual on-chip 256MB L3 through L2 cooperation
- 256MB distributed cache with avg ~12ns latency

Across-chip Horizontal Cache Persistence

- Virtual 2GB L4 cache across up to 8 chips

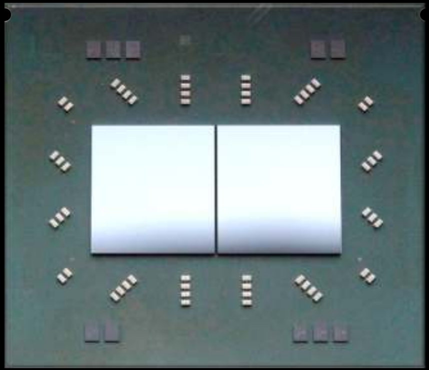


# Building large scale systems: connecting up to 32 chips



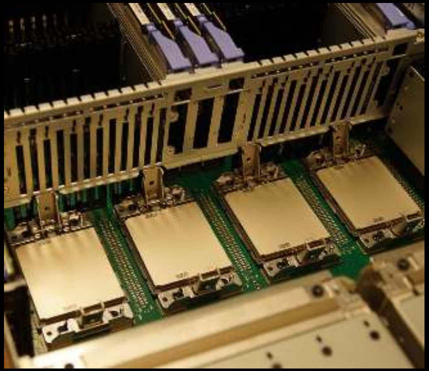
## Single Chip

1 chip
256MB cache



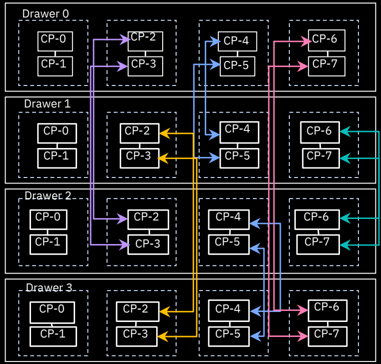
## Dual Chip Module

2 chips
512MB cache



## 4-Socket Drawer

8 chips
2GB cache



Drawer 0
CP-0 CP-1 CP-2 CP-3 CP-4 CP-5 CP-6 CP-7

Drawer 1
CP-0 CP-1 CP-2 CP-3 CP-4 CP-5 CP-6 CP-7

Drawer 2
CP-0 CP-1 CP-2 CP-3 CP-4 CP-5 CP-6 CP-7

Drawer 3
CP-0 CP-1 CP-2 CP-3 CP-4 CP-5 CP-6 CP-7

## 4-drawer system

32 chips
8GB cache

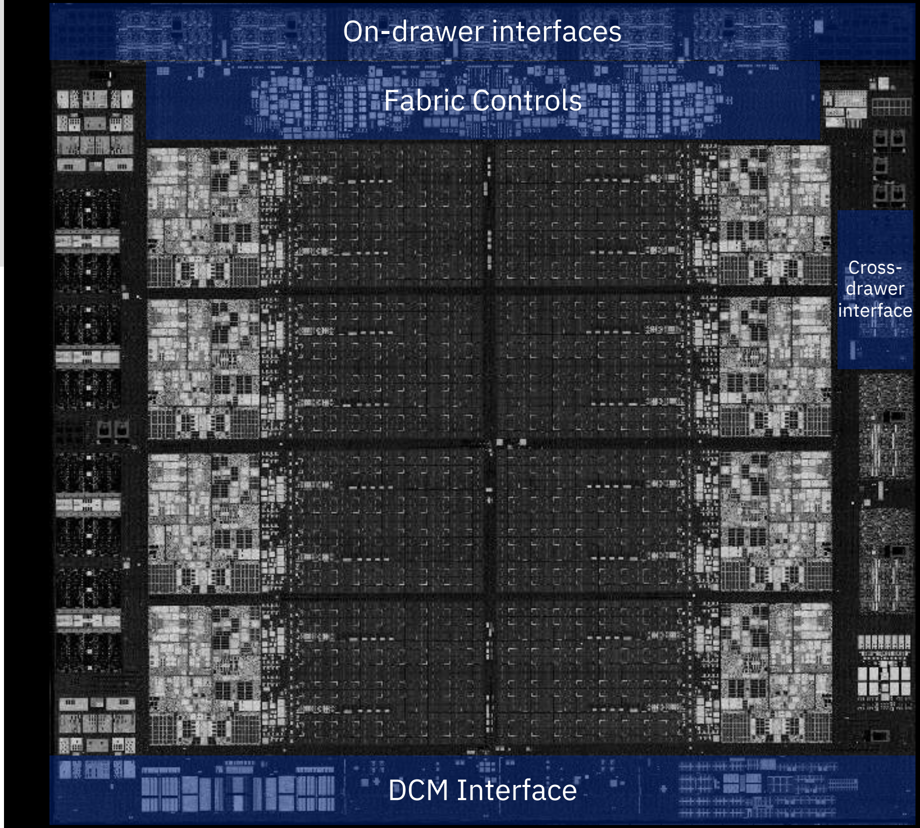
# Building large scale systems: Fabric & interface optimizations

**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



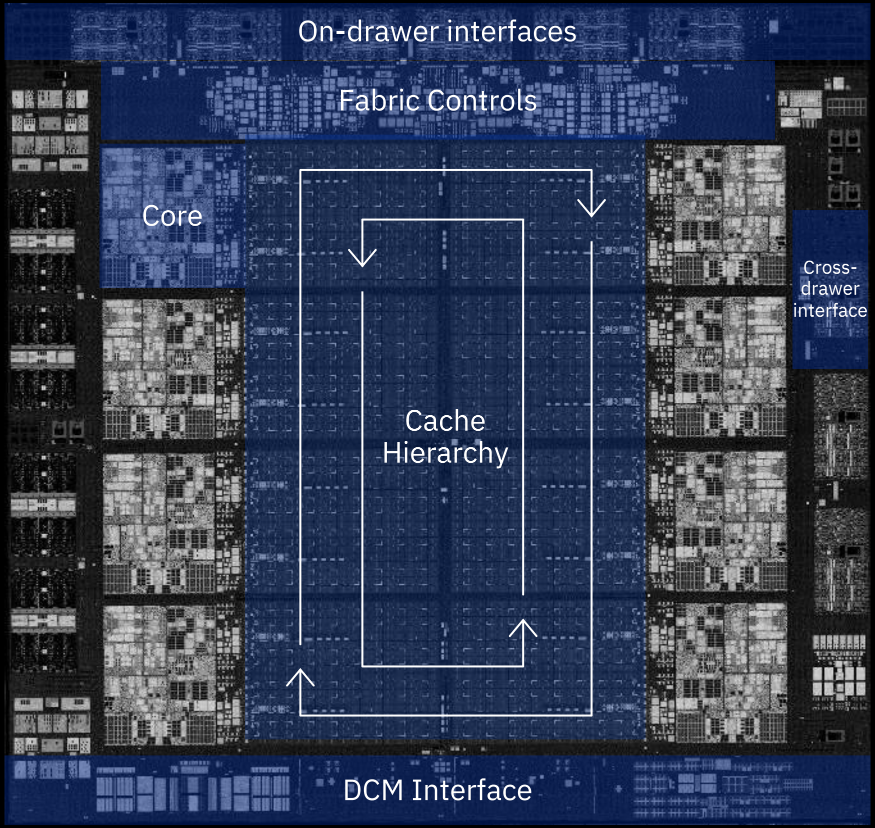
# Enterprise workload performance

### 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**

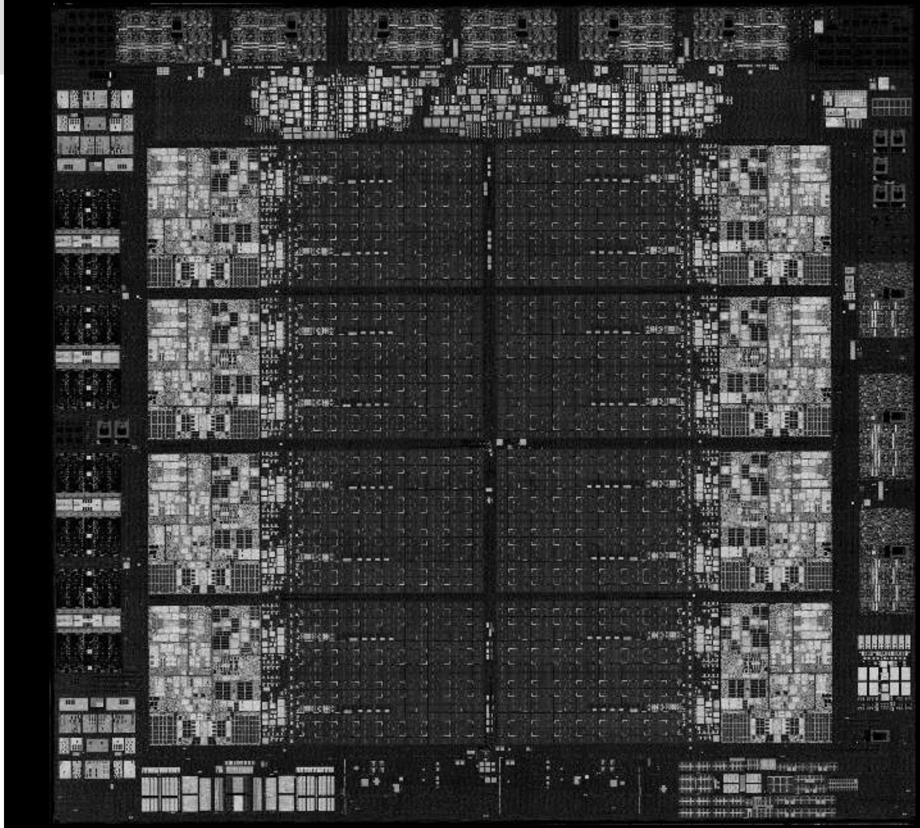
- Sort, Compression, Crypto
- AI

**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

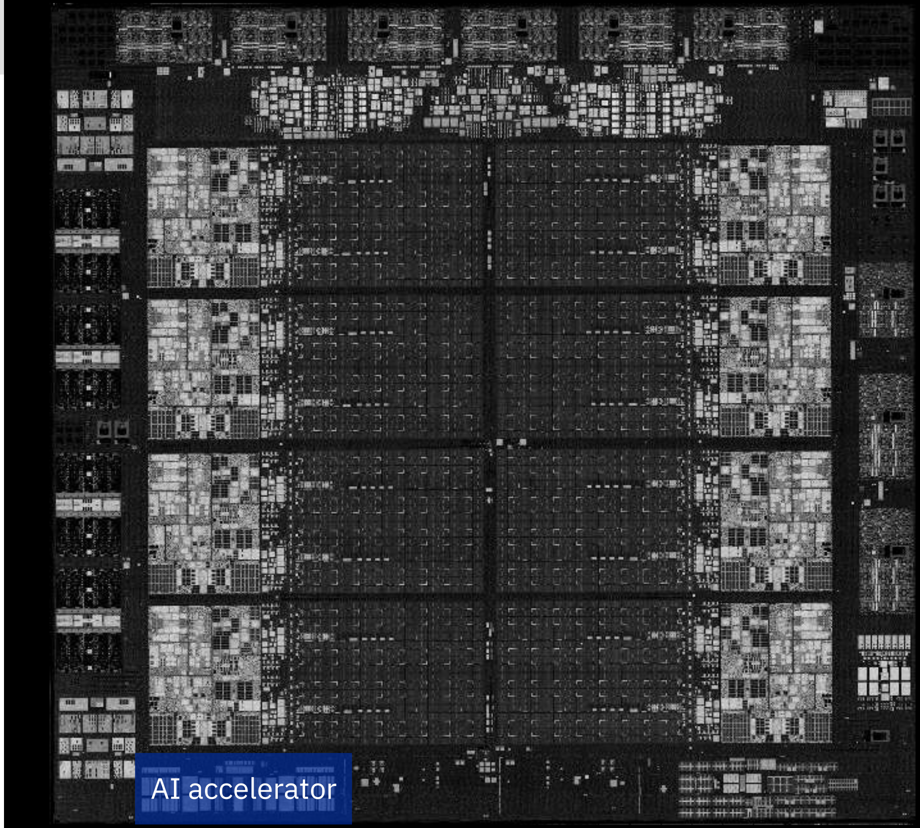
Minimize security exposure for sensitive data

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

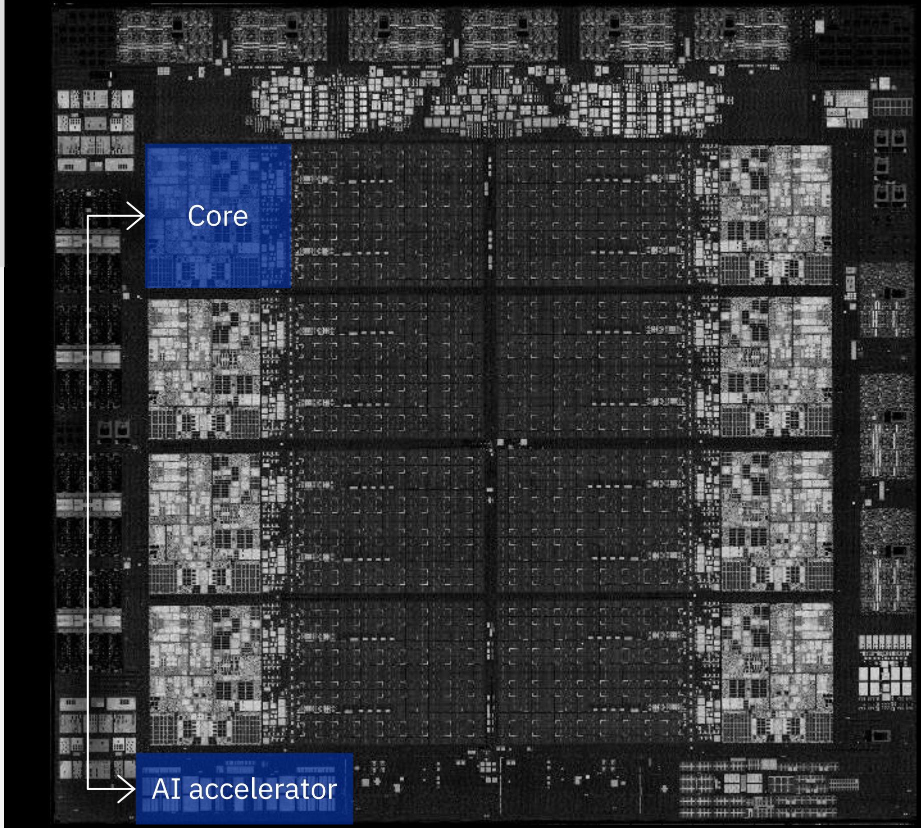
- Memory-to-memory CISC instruction
- Operates directly on tensor data in user space
- Matrix Multiplication, Convolution, Pooling, Activation Functions

## 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

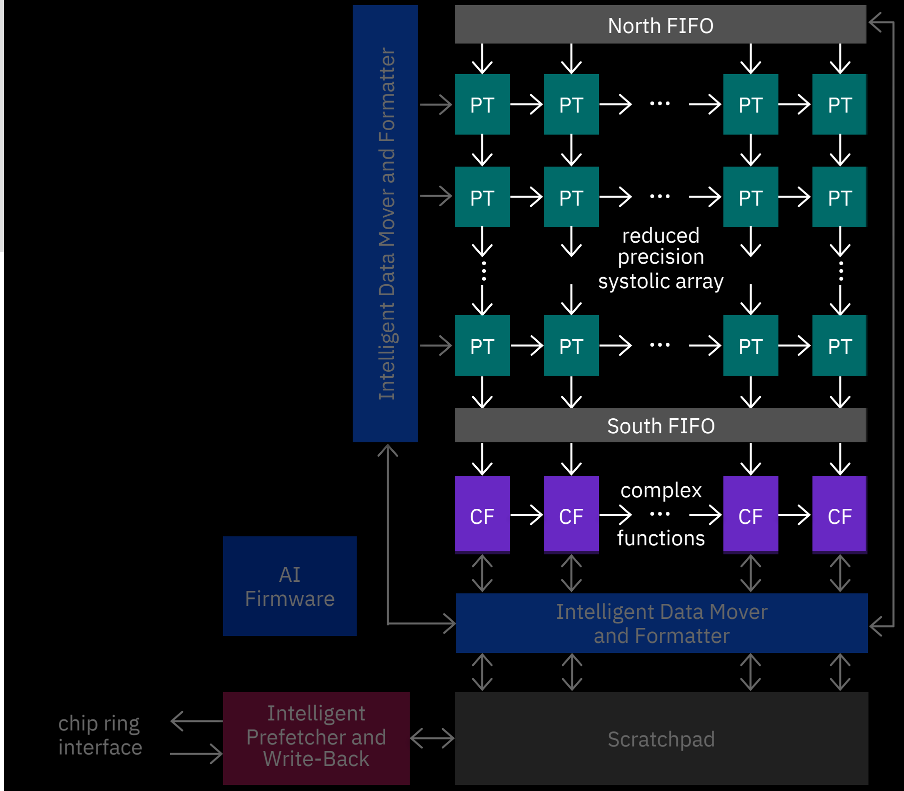
- Over 200 TFLOPS on 32-chip system

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



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

# Integrated AI Accelerator – data movers

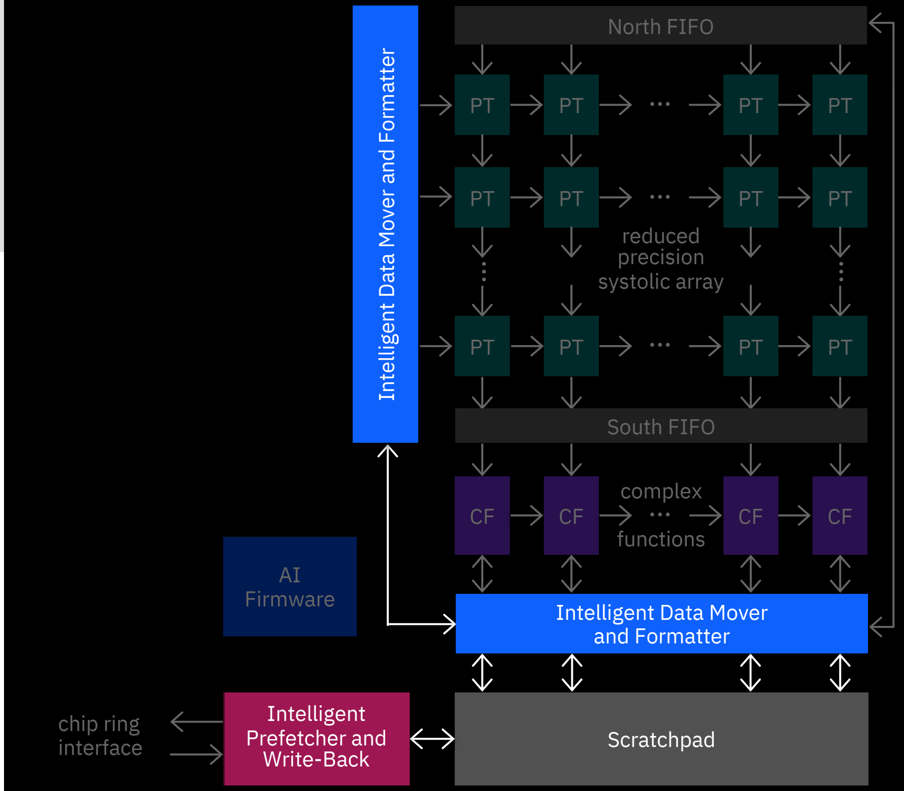
**On Chip AI Accelerator**

## 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

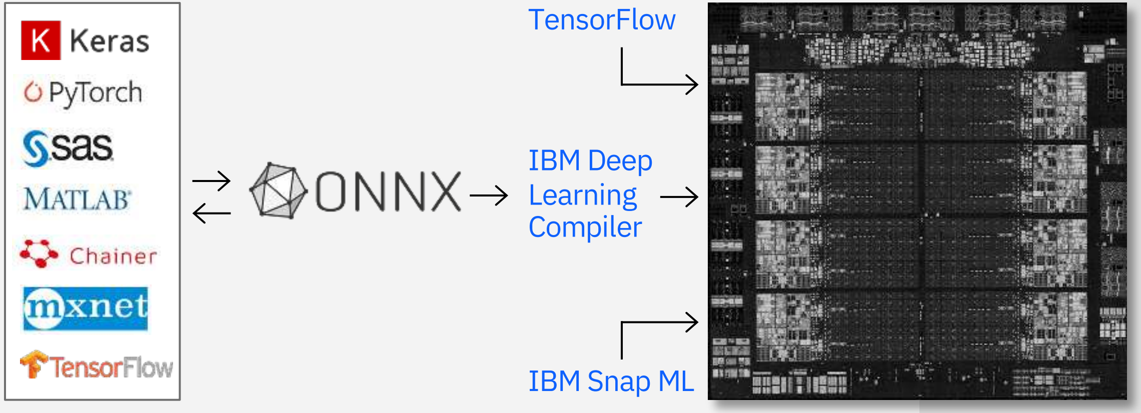
- 600+ GB/s bandwidth
- Format and prepare data on the fly for compute and write-back



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

# 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
Financial
Insurance

Retail
Hospitality
Transportation

Healthcare
Government
...

### 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 by Red Hat

### Database

IBM Db2   Db2 AI for z/OS   mongoDB.

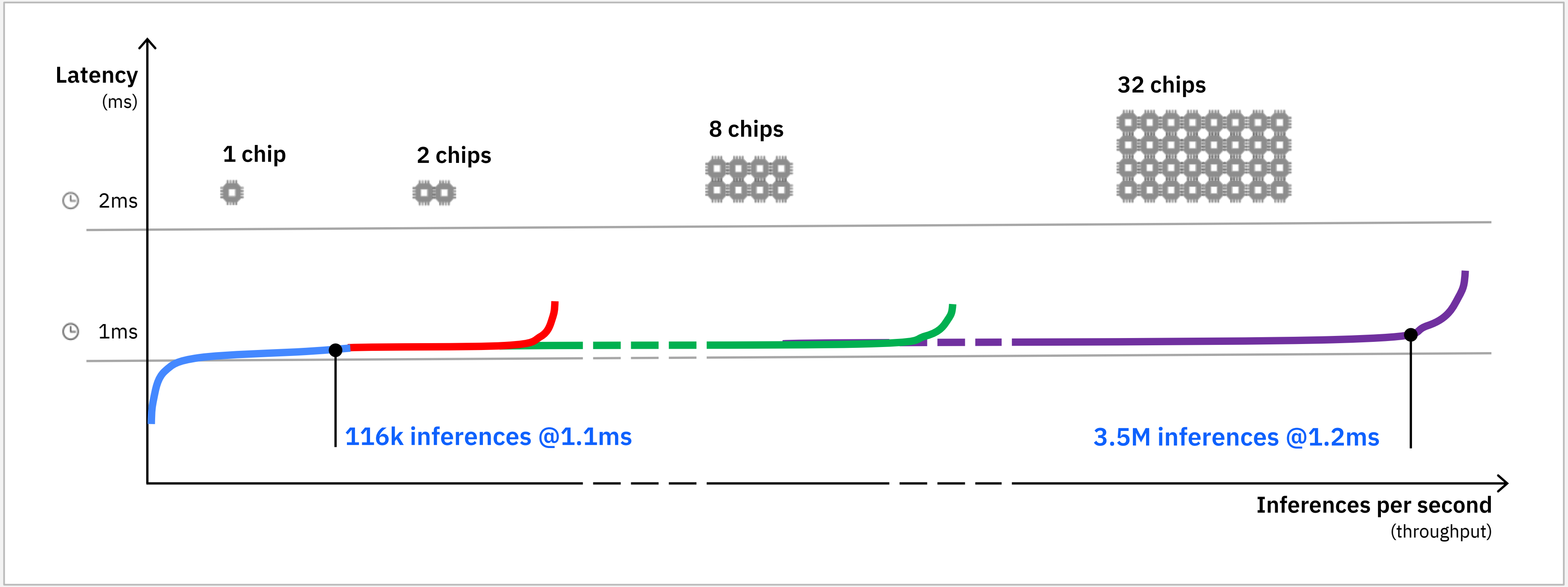
IMS   VSAM   PostgreSQL   MariaDB Foundation

### Operating Systems, Containers

z/OS   OPENSHIFT   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

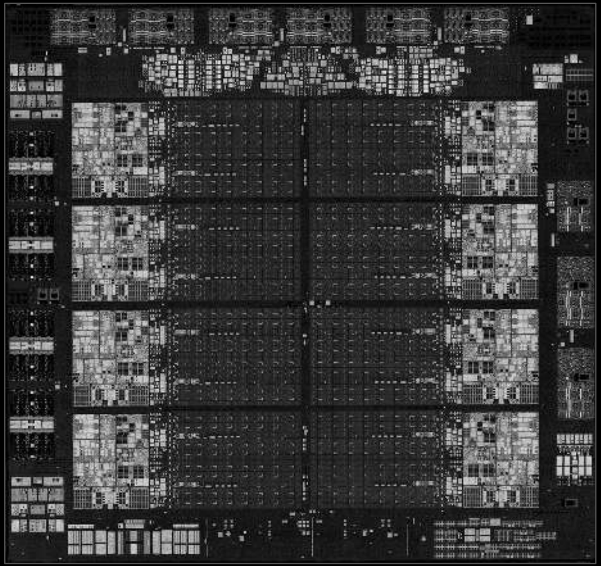
Next generation Z processor is optimized to run enterprise workloads with embedded real time AI insights.

**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

# Please note

IBM’s statements regarding its plans, directions, and intent are subject to change or withdrawal without notice and at IBM’s sole discretion.

Information regarding potential future products is intended to outline our general product direction and it should not be relied on in making a purchasing decision.

The information mentioned regarding potential future products is not a commitment, promise, or legal obligation to deliver any material, code or functionality. Information about potential future products may not be incorporated into any contract.

The development, release, and timing of any future features or functionality described for our products remains at our sole discretion.

Performance is based on measurements and projections using standard IBM benchmarks in a controlled environment. The actual throughput or performance that any user will experience will vary depending upon many factors, including considerations such as the amount of multiprogramming in the user’s job stream, the I/O configuration, the storage configuration, and the workload processed. Therefore, no assurance can be given that an individual user will achieve results similar to those stated here.

# Notices and disclaimers

**U.S. Government Users Restricted Rights — use, duplication or disclosure restricted by GSA ADP Schedule Contract with IBM.**

This document is current as of the initial date of publication and may be changed by IBM at any time. Not all offerings are available in every country in which IBM operates.

Information in this presentation (including information relating to products that have not yet been announced by IBM) has been reviewed for accuracy as of the date of initial publication and could include unintentional technical or typographical errors. IBM shall have no responsibility to update this information. **This document is distributed "as is" without any warranty, either express or implied. In no event, shall IBM be liable for any damage arising from the use of this information, including but not limited to, loss of data, business interruption, loss of profit or loss of opportunity.** IBM products and services are warranted per the terms and conditions of the agreements under which they are provided.

The performance data and client examples cited are presented for illustrative purposes only. Actual performance results may vary depending on specific configurations and operating conditions.

**Any statements regarding IBM's future direction, intent or product plans are subject to change or withdrawal without notice.**

Performance data contained herein was generally obtained in a controlled, isolated environments. Customer examples are presented as illustrations of how those customers have used IBM products and the results they may have achieved. Actual performance, cost, savings or other results in other operating environments may vary.

References in this document to IBM products, programs, or services does not imply that IBM intends to make such products, programs or services available in all countries in which IBM operates or does business.

It is the customer's responsibility to ensure its own compliance with legal requirements and to obtain advice of competent legal counsel as to the identification and interpretation of any relevant laws and regulatory requirements that may affect the customer's business and any actions the customer may need to take to comply with such laws. IBM does not provide legal advice or represent or warrant that its services or products will ensure that the customer follows any law.

Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products about this publication and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products. IBM does not warrant the quality of any third-party products, or the ability of any such third-party products to interoperate with IBM's products. IBM expressly disclaims all warranties, expressed or implied, including but not limited to, the implied warranties of merchantability and fitness for a purpose.

The provision of the information contained herein is not intended to, and does not, grant any right or license under any IBM patents, copyrights, trademarks or other intellectual property right.

IBM, the IBM logo, and ibm.com are trademarks of International Business Machines Corporation, registered in many jurisdictions worldwide. Other product and service names might be trademarks of IBM or other companies. A current list of IBM trademarks is available on the Web at "Copyright and trademark information" at: www.ibm.com/legal/copytrade.shtml.

# Trademarks

The following are trademarks of the International Business Machines Corporation in the United States and/or other countries.

IBM, IBM 8-bar Logo, ibm.com, and IBM Z

Adobe, the Adobe logo, PostScript, and the PostScript logo are either registered trademarks or trademarks of Adobe Systems Incorporated in the United States, and/or other countries.

Cell Broadband Engine is a trademark of Sony Computer Entertainment, Inc. in the United States, other countries, or both and is used under license therefrom.

IT Infrastructure Library is a Registered Trade Mark of AXELOS Limited.

ITIL is a Registered Trade Mark of AXELOS Limited.

Linear Tape-Open, LTO, the LTO Logo, Ultrium, and the Ultrium logo are trademarks of HP, IBM Corp. and Quantum in the U.S. and other countries.

Intel, Intel logo, Intel Inside, Intel Inside logo, Intel Centrino, Intel Centrino logo, Celeron, Intel Xeon, Intel SpeedStep, Itanium, and Pentium are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.

The registered trademark Linux® is used pursuant to a sublicense from the Linux Foundation, the exclusive licensee of Linus Torvalds, owner of the mark on a worldwide basis.

Java and all Java-based trademarks and logos are trademarks or registered trademarks of Oracle and/or its affiliates.

Microsoft, Windows, Windows NT, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.

OpenStack is a trademark of OpenStack LLC. The OpenStack trademark policy is available on the OpenStack website.

Red Hat®, JBoss®, OpenShift®, Fedora®, Hibernate®, Ansible®, CloudForms®, RHCA®, RHCE®, RHCSA®, Ceph®, and Gluster® are trademarks or registered trademarks of Red Hat, Inc. or its subsidiaries in the United States and other countries.

RStudio®, the RStudio logo and Shiny® are registered trademarks of RStudio, Inc.

UNIX is a registered trademark of The Open Group in the United States and other countries.

VMware, the VMware logo, VMware Cloud Foundation, VMware Cloud Foundation Service, VMware vCenter Server, and VMware vSphere are registered trademarks or trademarks of VMware, Inc. or its subsidiaries in the United States and/or other jurisdictions.

Zowe™, the Zowe™ logo and the Open Mainframe Project™ are trademarks of The Linux Foundation.

Other product and service names might be trademarks of IBM or other companies.

**Notes:**

Performance is in Internal Throughput Rate (ITR) ratio based on measurements and projections using standard IBM benchmarks in a controlled environment. The actual throughput that any user will experience will vary depending upon considerations such as the amount of multiprogramming in the user's job stream, the I/O configuration, the storage configuration, and the workload processed. Therefore, no assurance can be given that an individual user will achieve throughput improvements equivalent to the performance ratios stated here.

IBM hardware products are manufactured from new parts, or new and serviceable used parts. Regardless, our warranty terms apply.

All customer examples cited or described in this presentation are presented as illustrations of the manner in which some customers have used IBM products and the results they may have achieved. Actual environmental costs and performance characteristics will vary depending on individual customer configurations and conditions.

This publication was produced in the United States. IBM may not offer the products, services or features discussed in this document in other countries, and the information may be subject to change without notice. Consult your local IBM business contact for information on the product or services available in your area.

All statements regarding IBM's future direction and intent are subject to change or withdrawal without notice, and represent goals and objectives only.

Information about non-IBM products is obtained from the manufacturers of those products or their published announcements. IBM has not tested those products and cannot confirm the performance, compatibility, or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

Prices subject to change without notice. Contact your IBM representative or Business Partner for the most current pricing in your geography.

This information provides only general descriptions of the types and portions of workloads that are eligible for execution on Specialty Engines (e.g, zIIPs, zAAPs, and IFLs) ("SEs"). IBM authorizes customers to use IBM SE only to execute the processing of Eligible Workloads of specific Programs expressly authorized by IBM as specified in the “Authorized Use Table for IBM Machines” provided at www.ibm.com/systems/support/machine_warranties/machine_code/aut.html (“AUT”). No other workload processing is authorized for execution on an SE. IBM offers SE at a lower price than General Processors/Central Processors because customers are authorized to use SEs only to process certain types and/or amounts of workloads as specified by IBM in the AUT.

# Thank you

**Dr. Christian Jacobi**
Distinguished Engineer
Chief Architect IBM Z Processor Design
IBM Systems, Poughkeepsie, NY
cjacobi@us.ibm.com