IBM Telum® II processor and IBM Spyre™ Accelerator chip for AI

Chris Berry
IBM Distinguished Engineer

70% of all financial transactions go through an IBM mainframe

IBM mainframes run an estimated 70% of all world transactions by value.
Source: Operationalizing Fraud Prevention on IBM z16, Celent, 2022. ibm.com/downloads/cas/DOXY3Q94

99.999999%

available

itic-corp.com/itic-2023-reliability-survey-ibm-z-results/

Enable in-transaction AI for millions of transactions per second

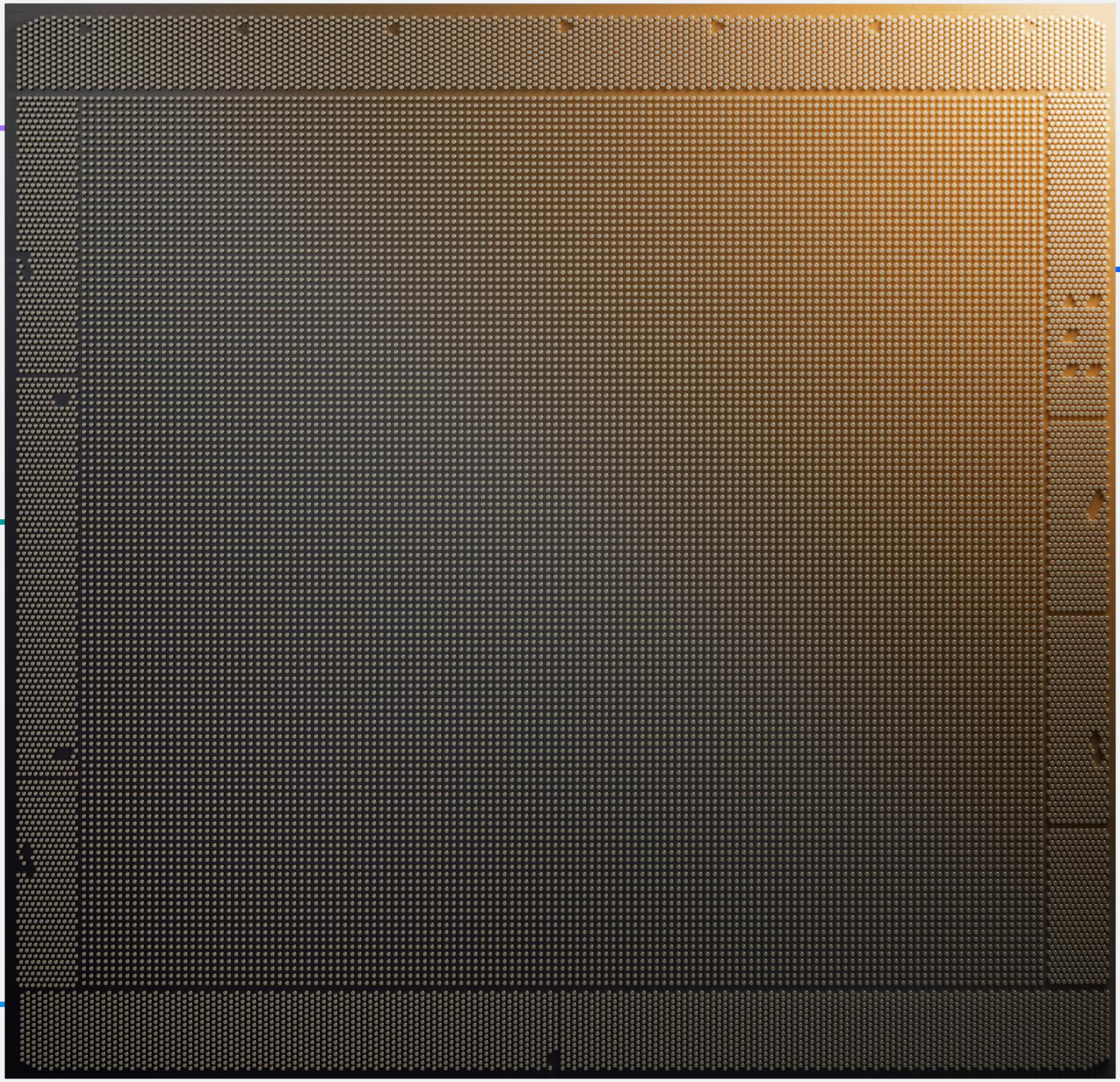
10 - 36MB L2s

DPU

For I/O acceleration

AI accelerator

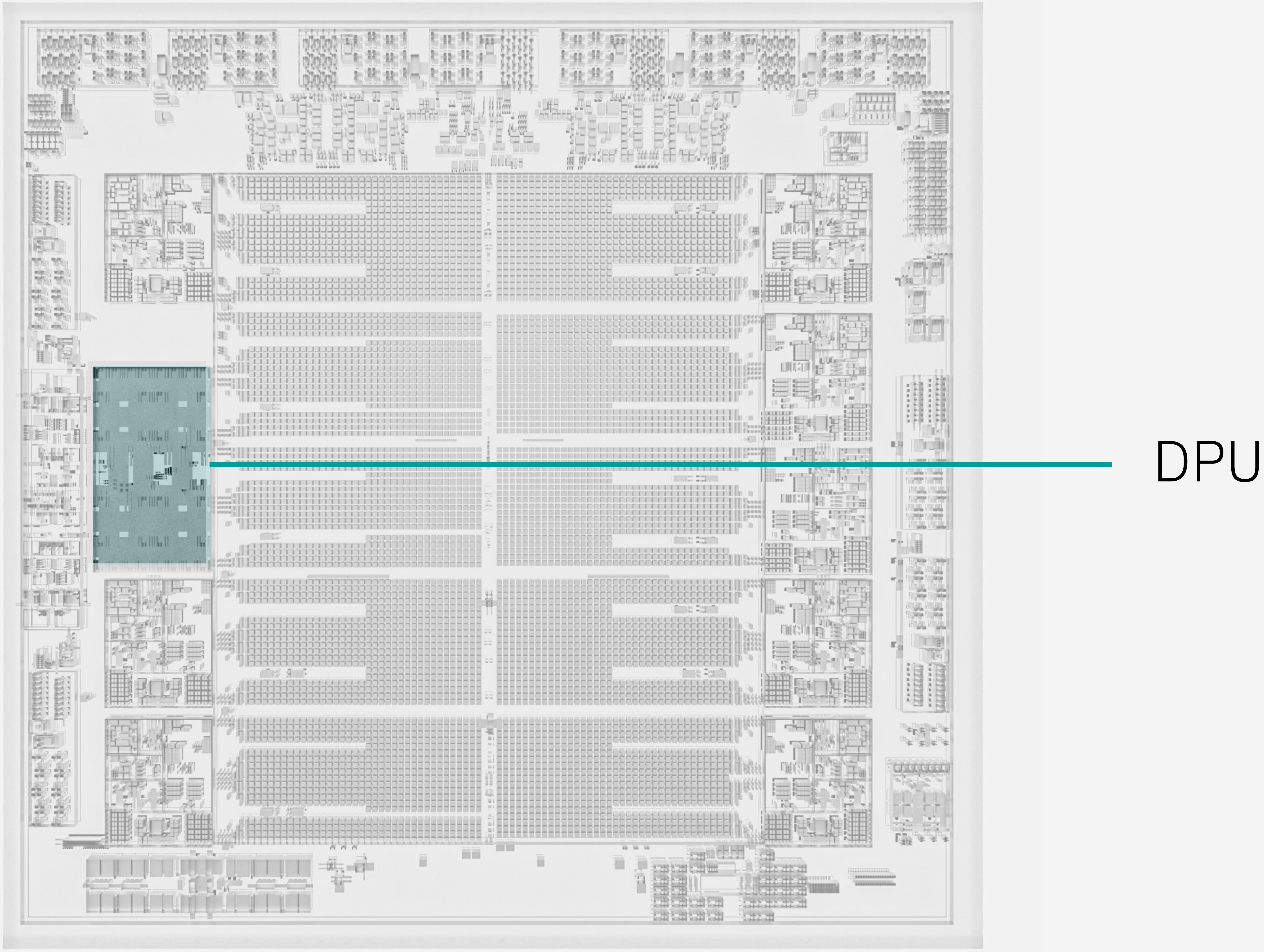
24 TOPS



8 - 5.5GHz cores

On-chip Data Processing Unit implements complex I/O protocols and reduces latency

Reduced power for IO management by 70%



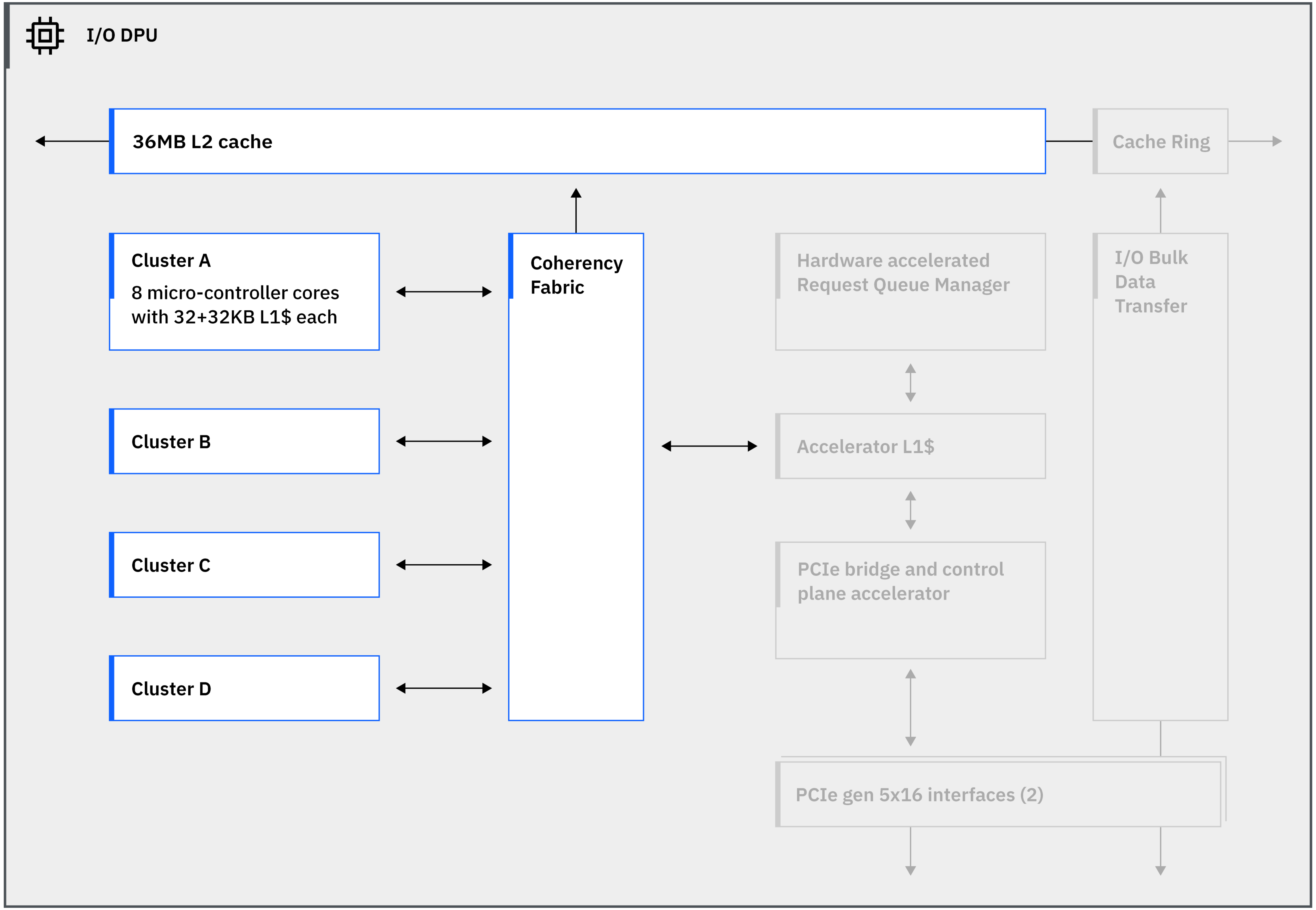
All projections are based on pre-release hardware measurements and configurations and comparisons are made against previous versions of the IBM Telum processor.

4 processing clusters

8 customized micro-controllers per cluster

Coherently connected

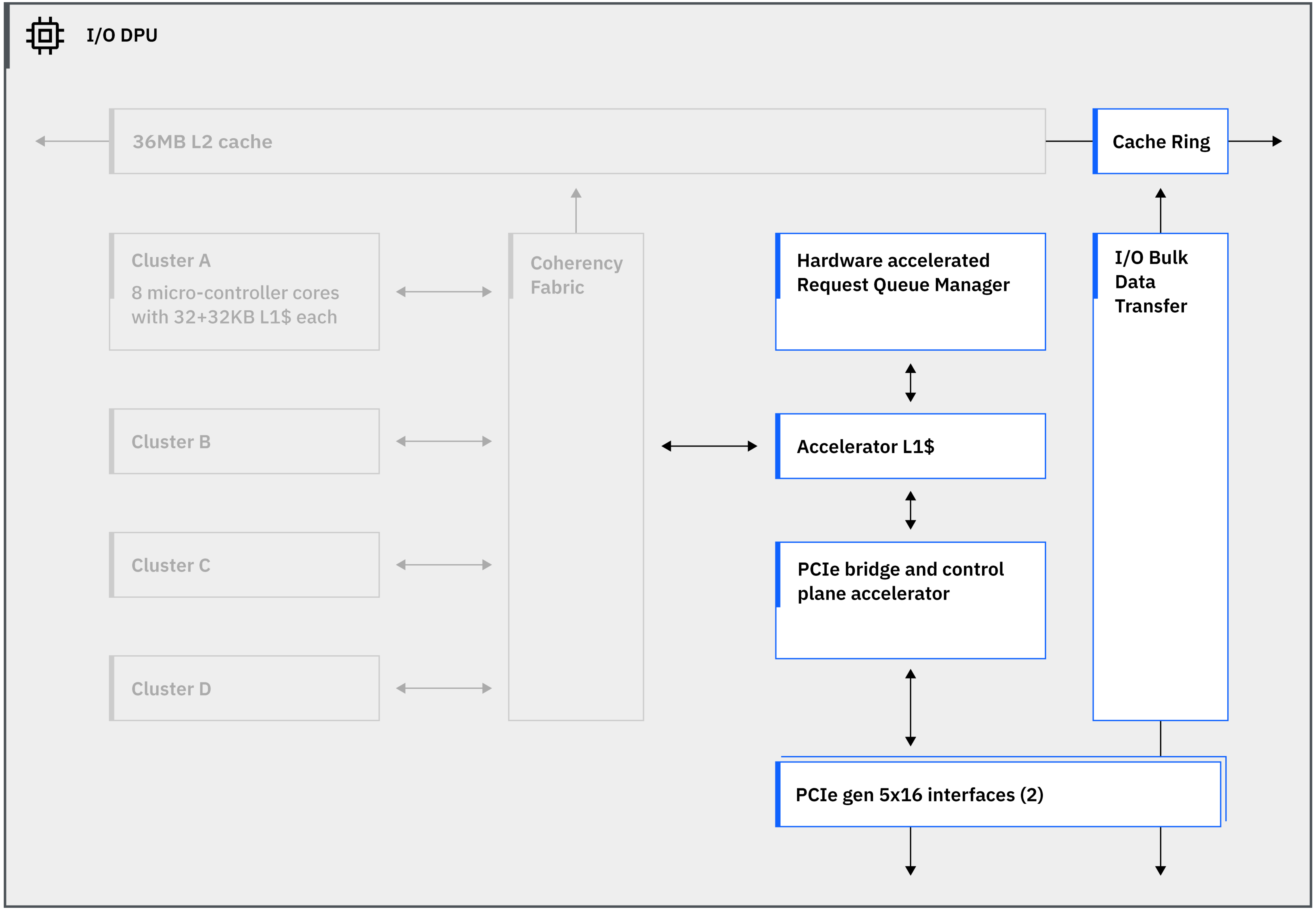
Connected amongst each other and with the overall cache system



Dedicated I/O acceleration logic

Queue manager manages scheduling and work transfer within the DPU

Connects to PCIe off-chip interfaces

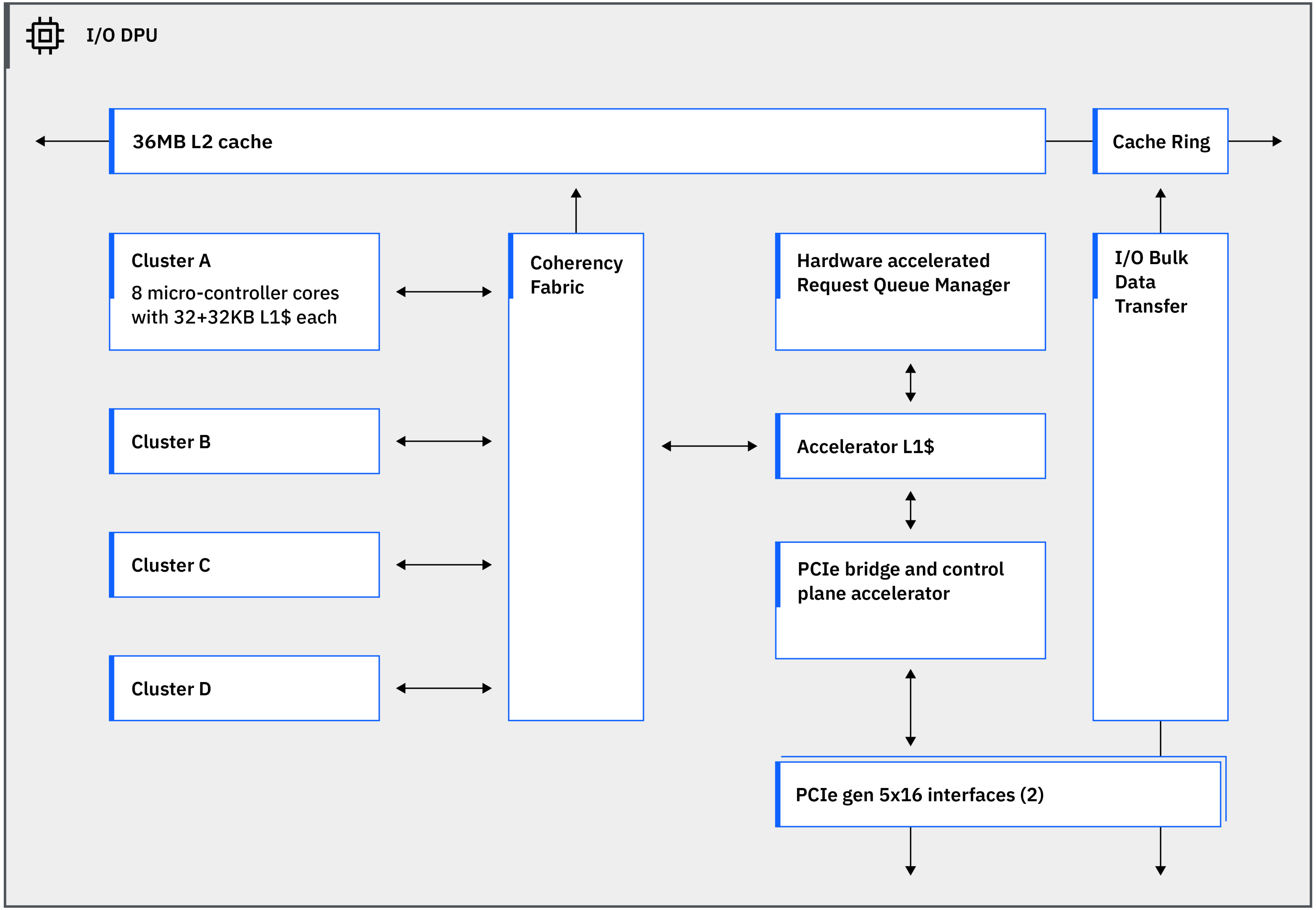


Massive scale-out I/O subsystem

Up to 32
Telum II in a coherent SMP system

12 I/O expansion drawers with 16 PCIe slots

Up to 192 PCIe cards in a system

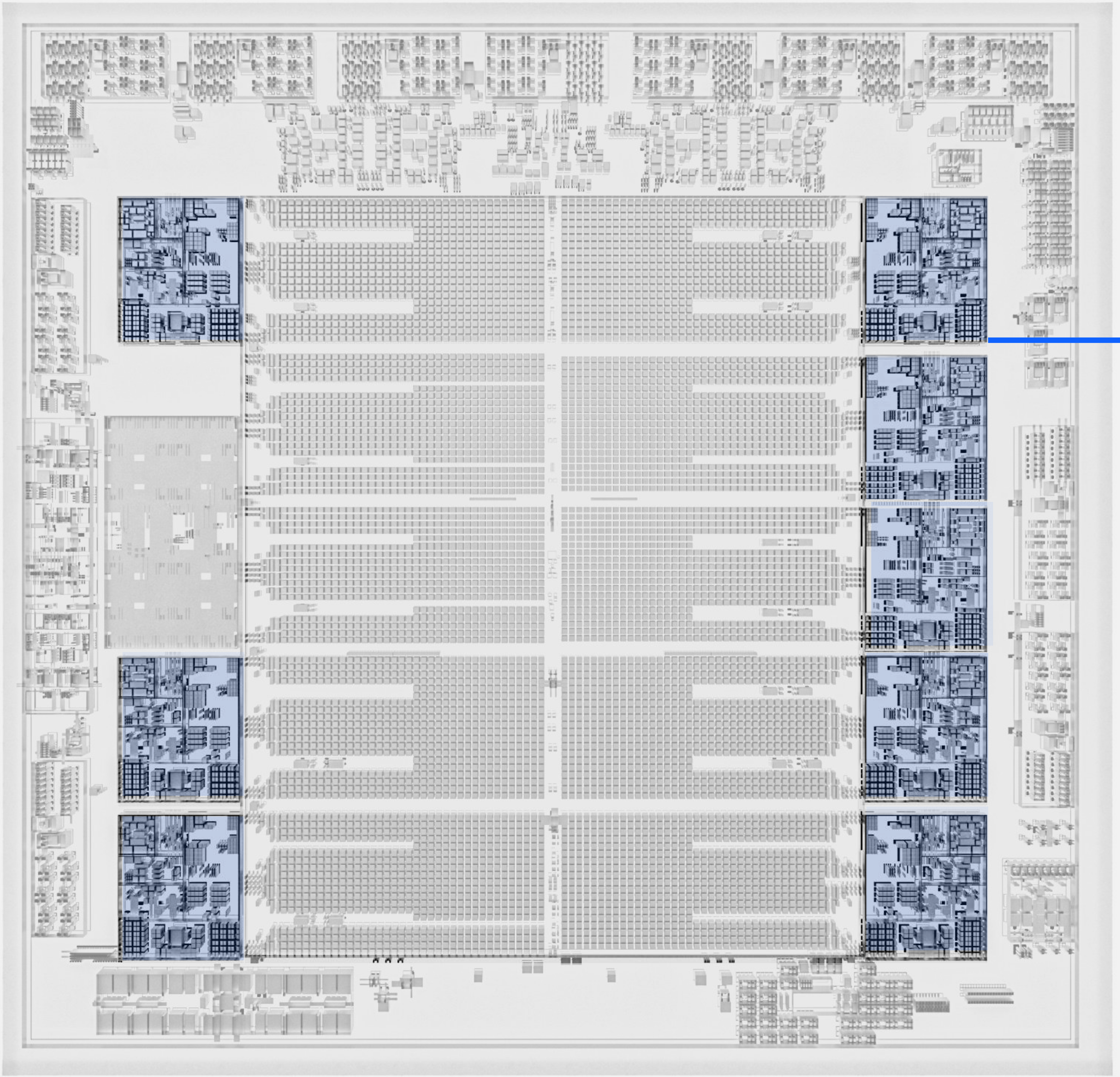


Enhanced Central Compute Complex

Optimized cores

- Branch prediction improvements
- Rename register growth 128 to 160
- Improved store writeback and address translation

- 20% area reduction
- 15% power reduction



8 cores

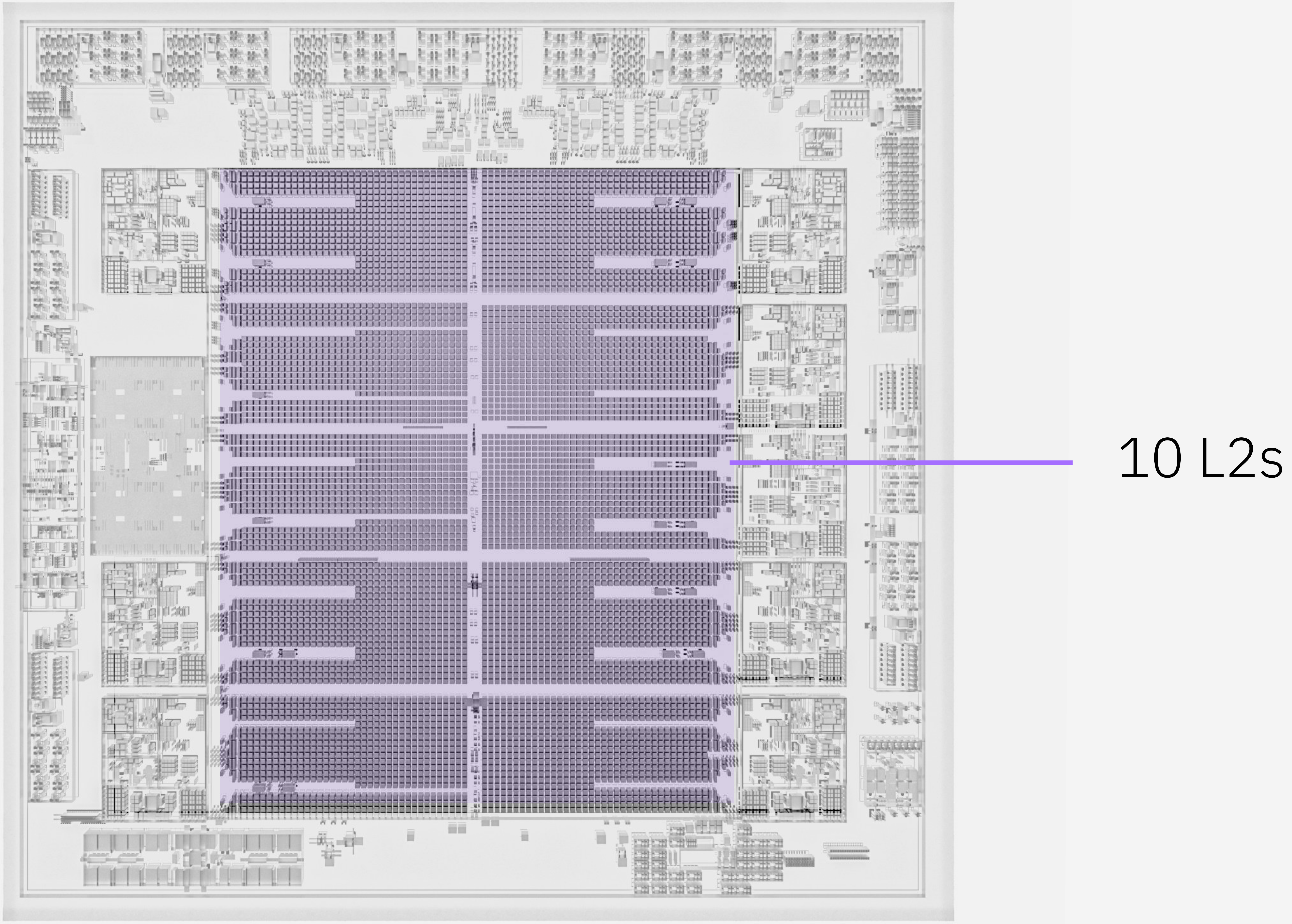
All projections are based on pre-release hardware measurements and configurations and comparisons are made against previous versions of the IBM Telum processor.

40% cache growth

Improved caches

- 36MB L2—3.6ns
- 360MB virtual L3—11.5ns
- 2.8 GB virtual L4—48.5ns

- Improved horizontal persistence
- 30% improvement off-chip bandwidth
- Drawer-to-drawer encryption

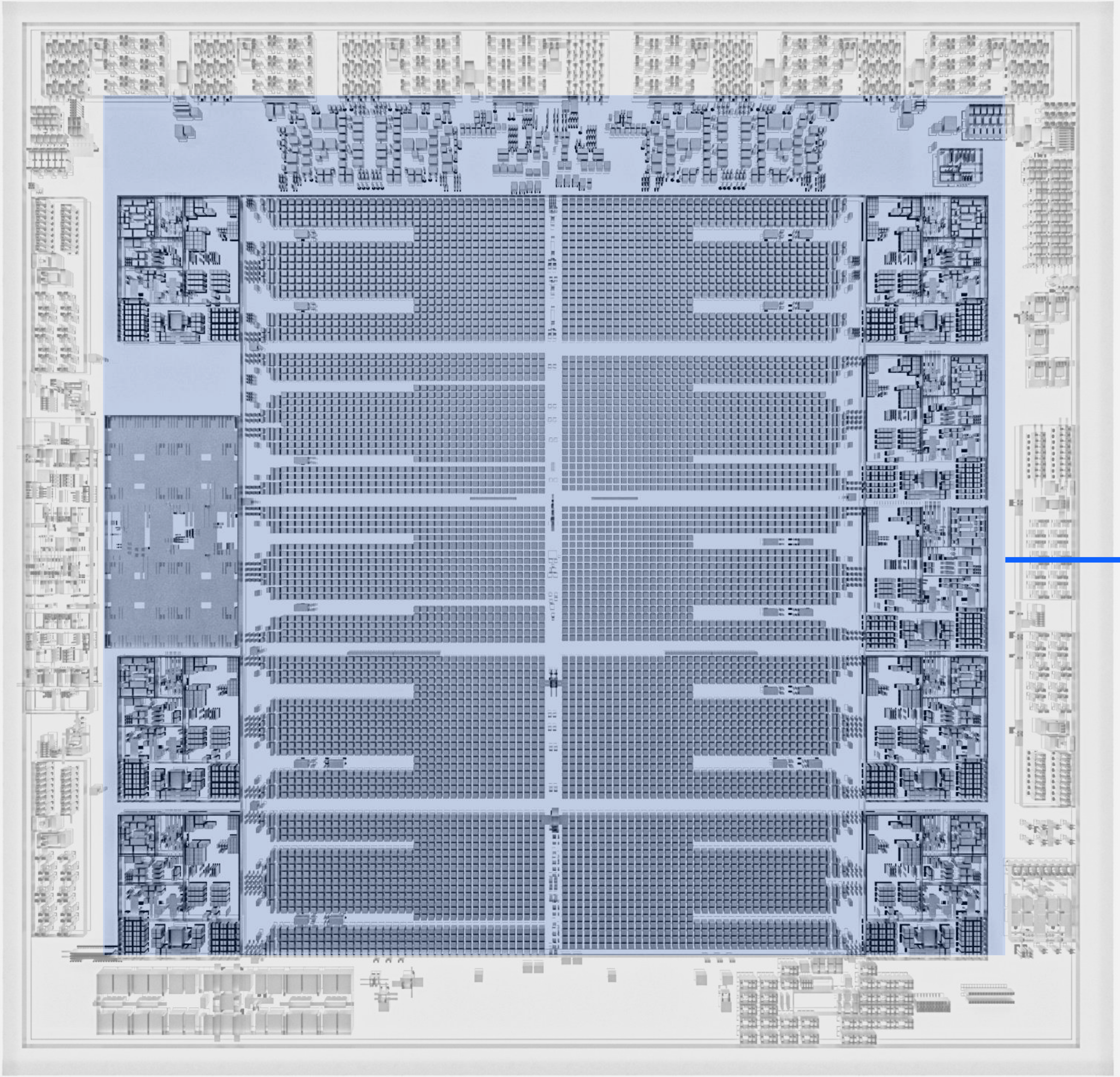


All projections are based on pre-release hardware measurements and configurations and comparisons are made against previous versions of the IBM Telum processor.

5nm
technology

↑5.5 GHz
per core

↑20%
socket
performance



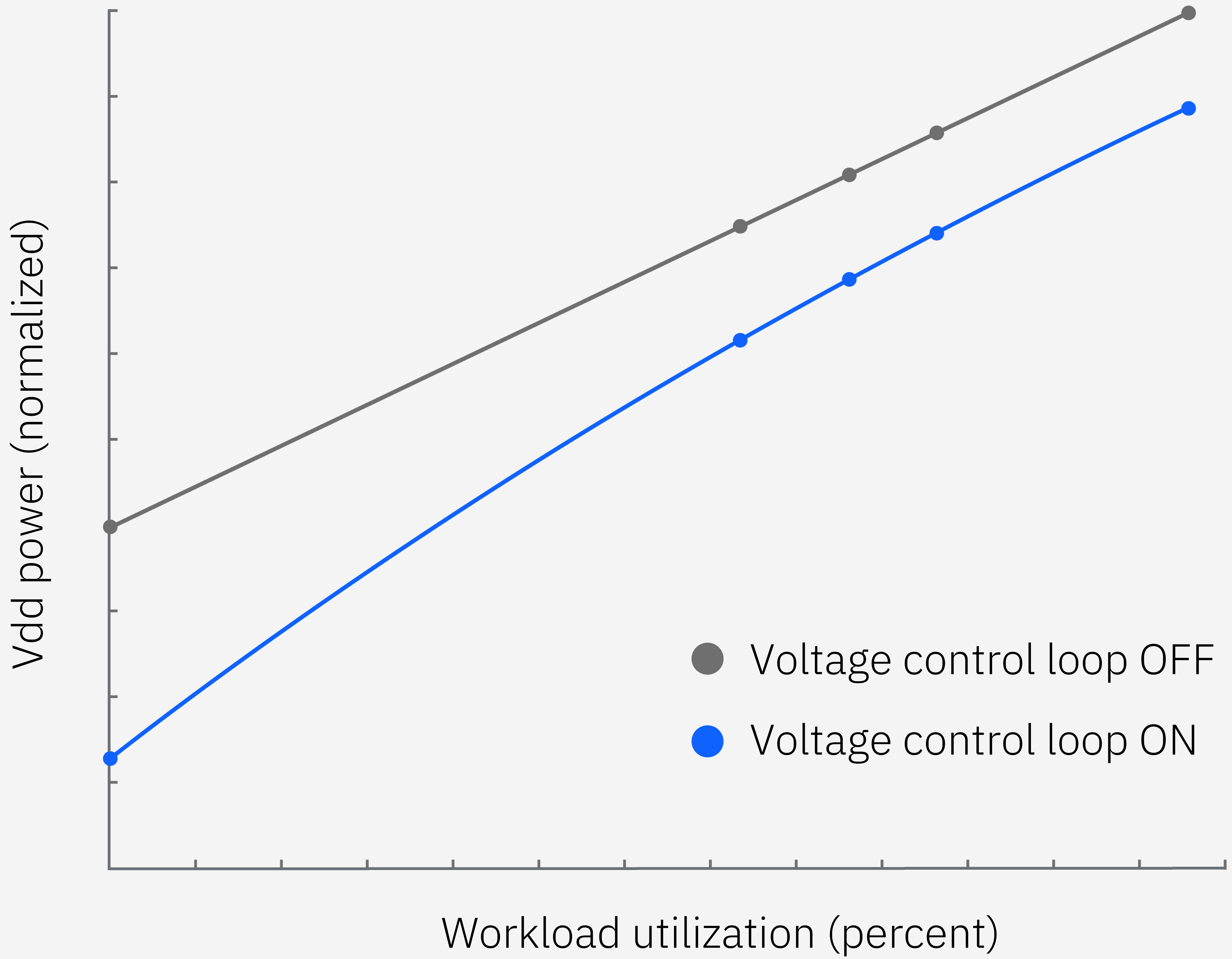
Voltage
control loop

All projections are based on pre-release hardware measurements and configurations and comparisons are made against previous versions of the IBM Telum processor.

↓18%
processor power consumption

Fine grain voltage measurement and workload monitoring circuits

Control loop tightly coupled to power conversion subsystem



AI inferencing for enterprise

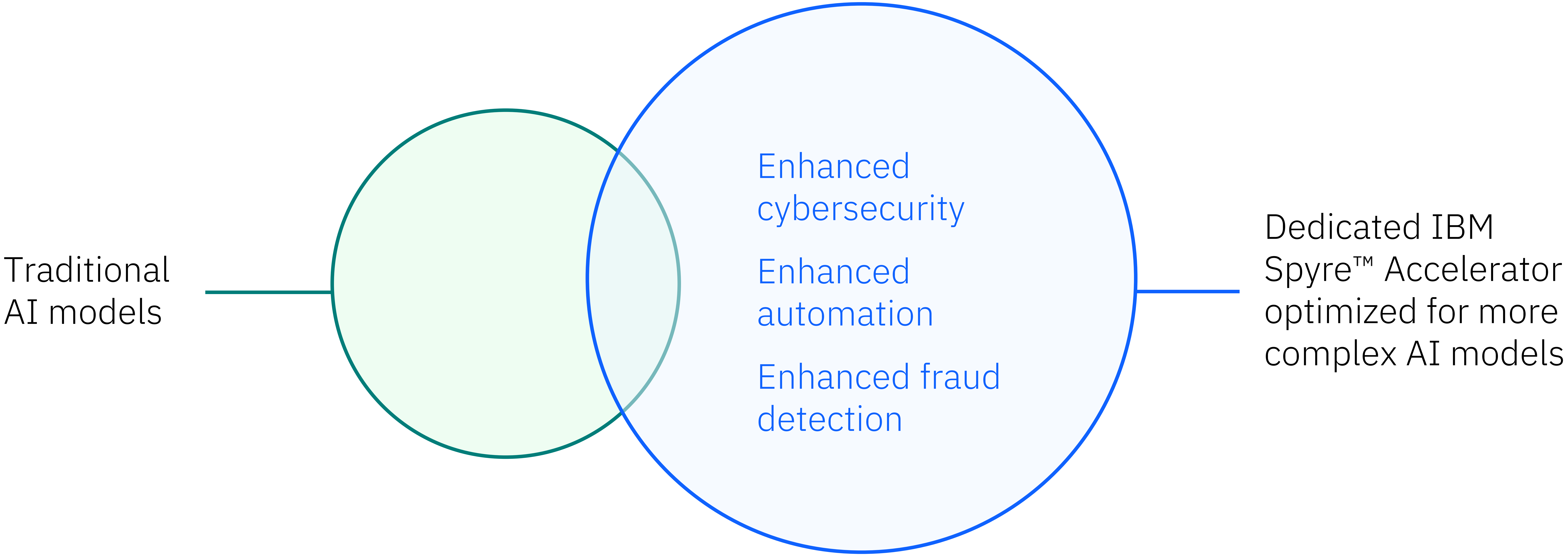
On-processor AI accelerator optimized for traditional AI models

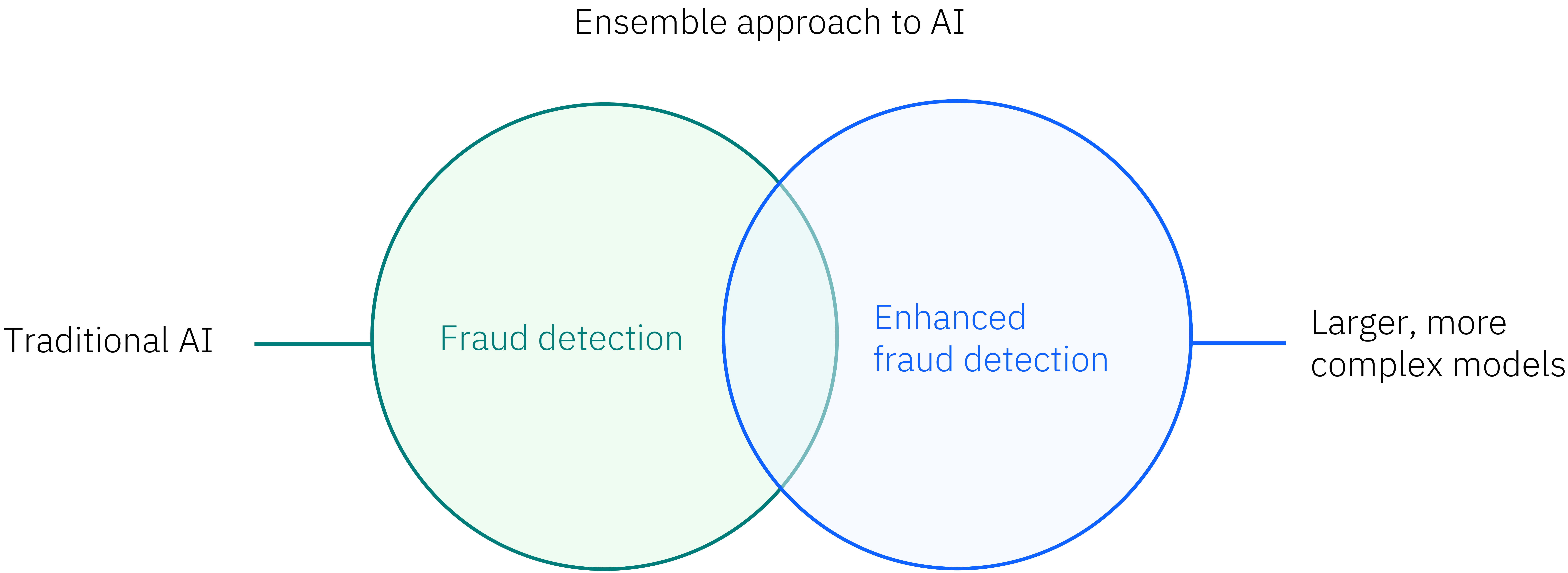
Clearance settlement

Compliance

Insurance claims

Fraud detection


Traditional AI models
Enhanced cybersecurity
Enhanced automation
Enhanced fraud detection
Dedicated IBM Spyre™ Accelerator optimized for more complex AI models


Ensemble approach to AI
Traditional AI
Fraud detection
Enhanced fraud detection
Larger, more complex models

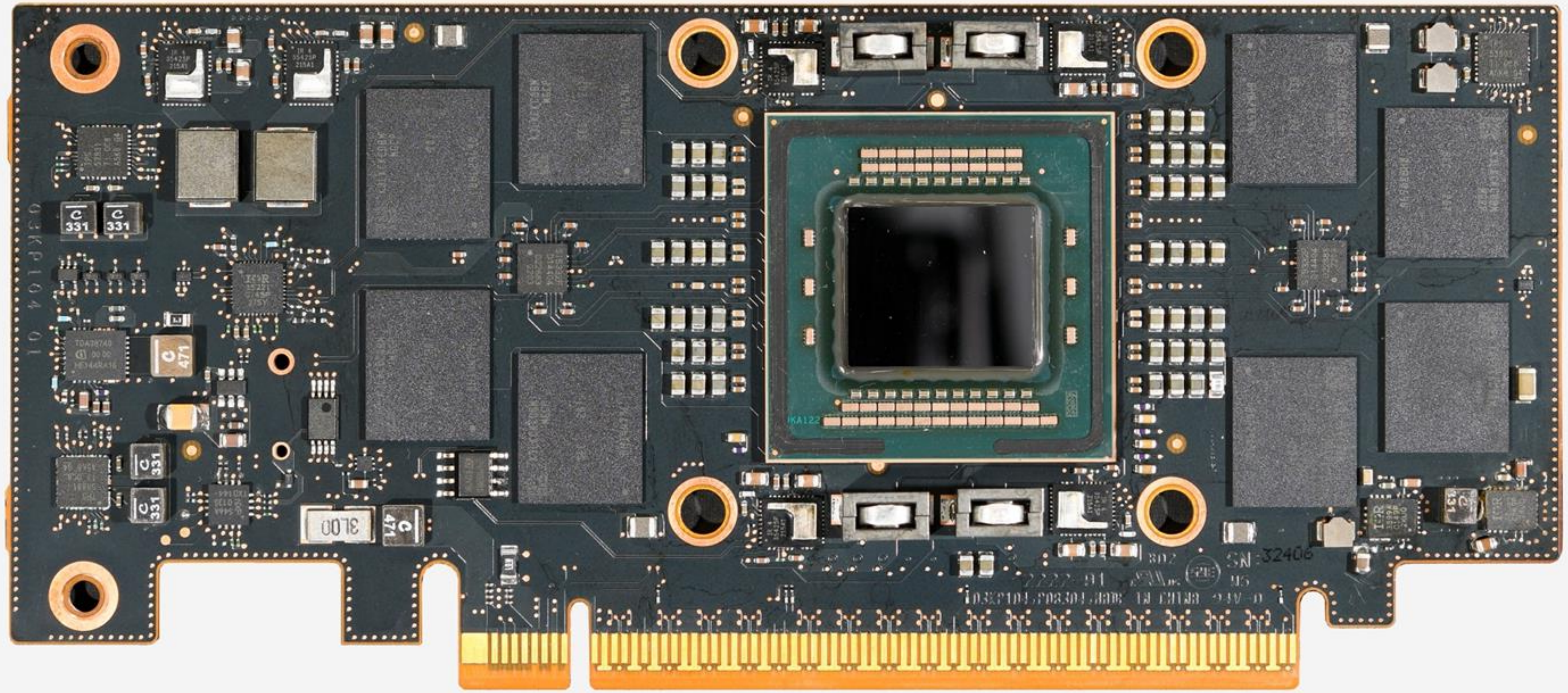
Improved On-processor AI Acceleration

- 24 TOPS per chip
- 192 TOPS per drawer
- 768 TOPS per system

- Integrated as a CISC instruction
- Remote AI accelerator
- Support for LLM compute primitives
- Int8, FP16 datatypes

IBM Spyre™ Accelerator PCIe attached card

- 300+ TOPS
- 75W PCIe gen5 x16 adapter
- 128BGB of LPDDR5 memory

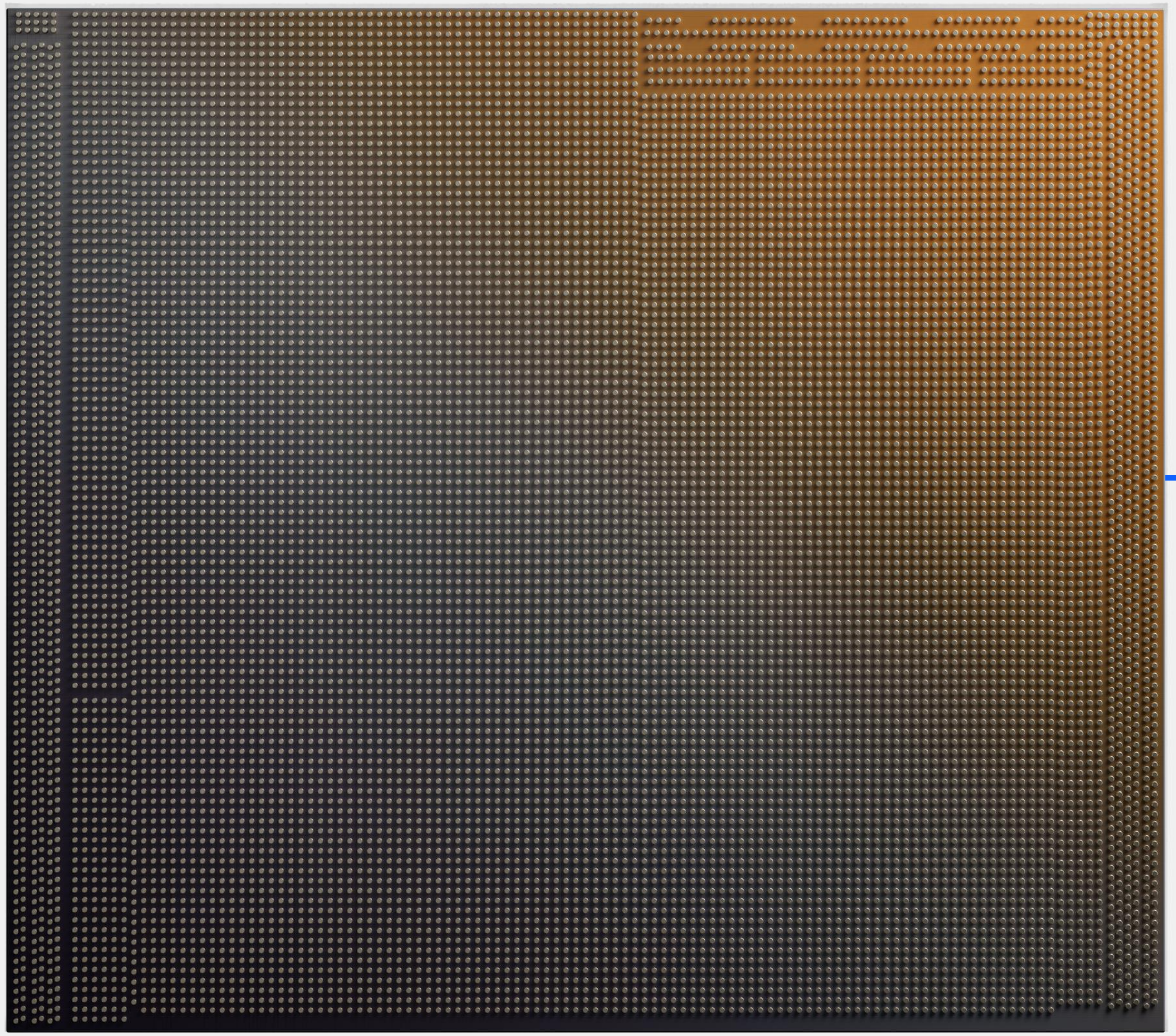


All projections are based on preview hardware.

32 cores

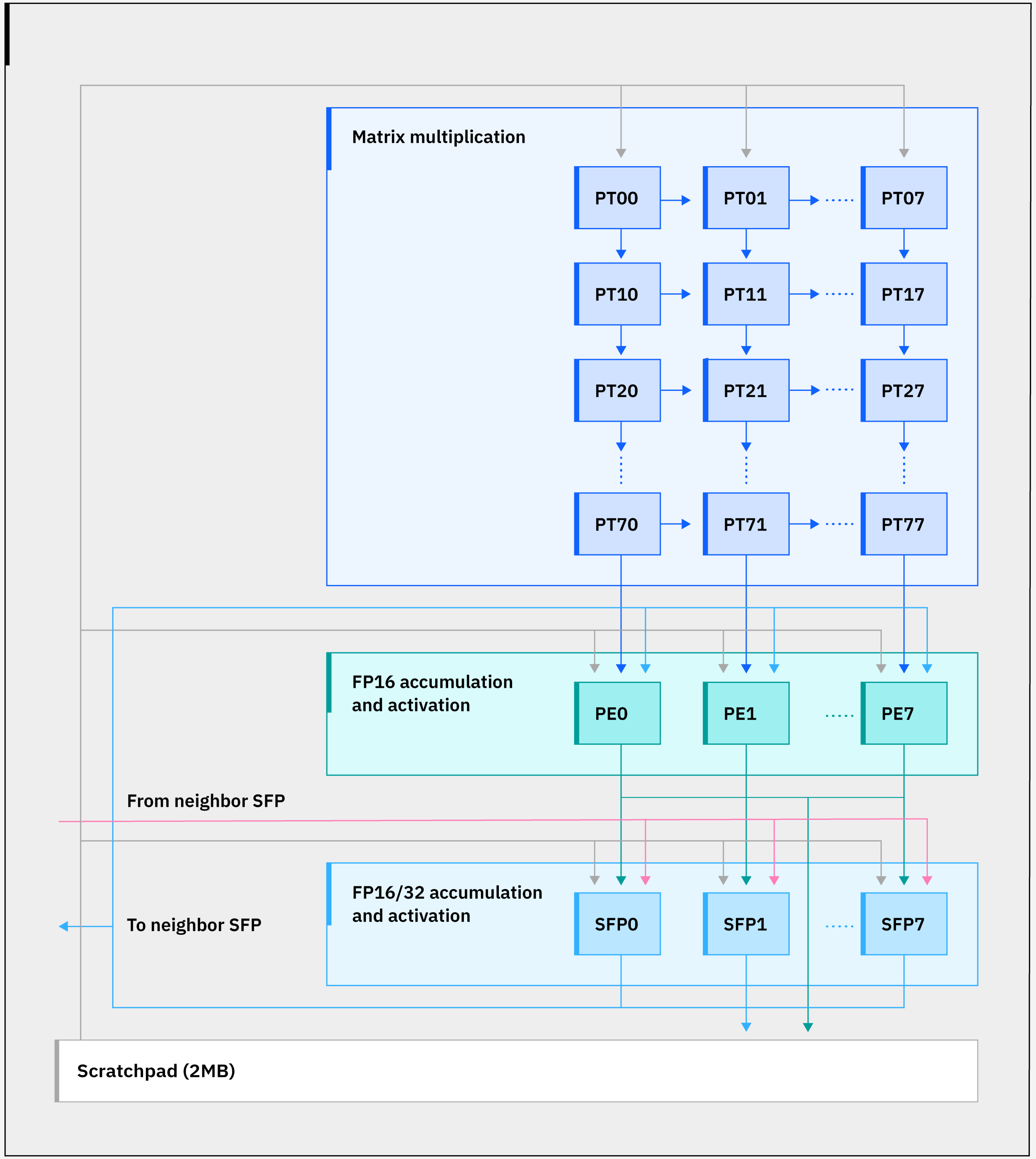
2MB scratchpad on each

55%+ effective TOPS utilization



All projections are based on preview hardware.

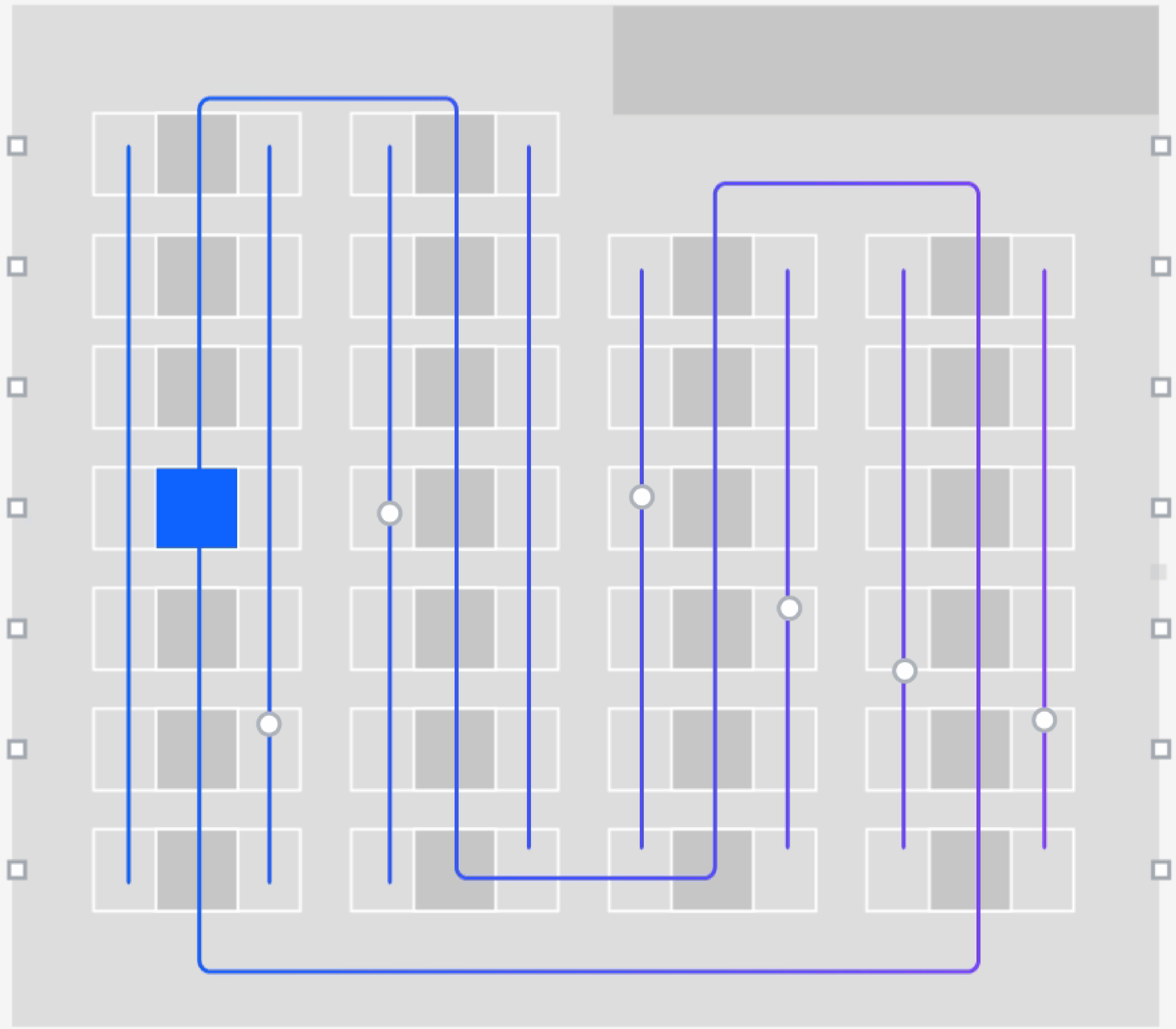
Corelet structure

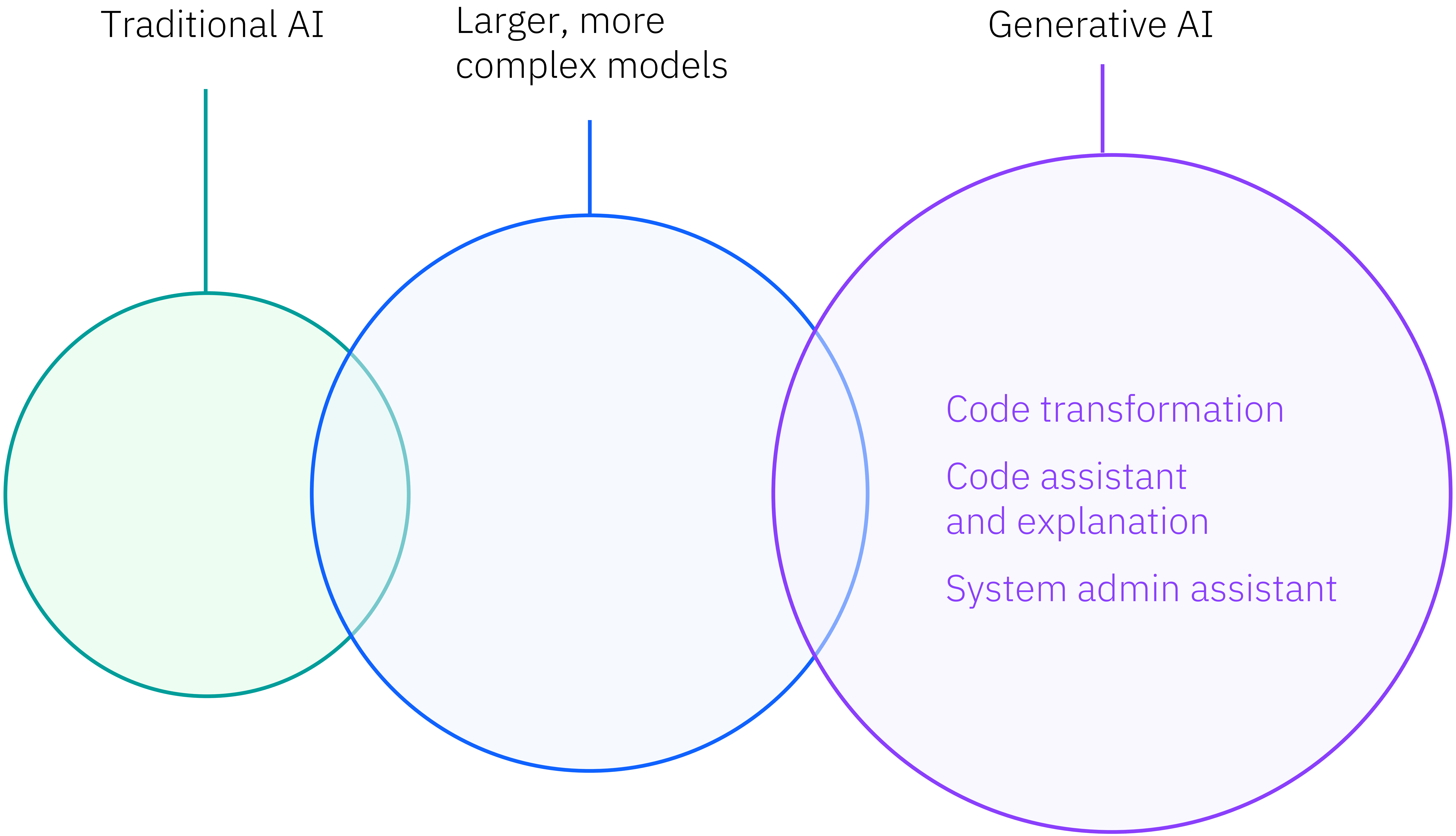


32B bidirectional ring connects the cores

128B bidirectional ring connects scratchpads

200GB/s LPDDR5 connects to scratchpad ring



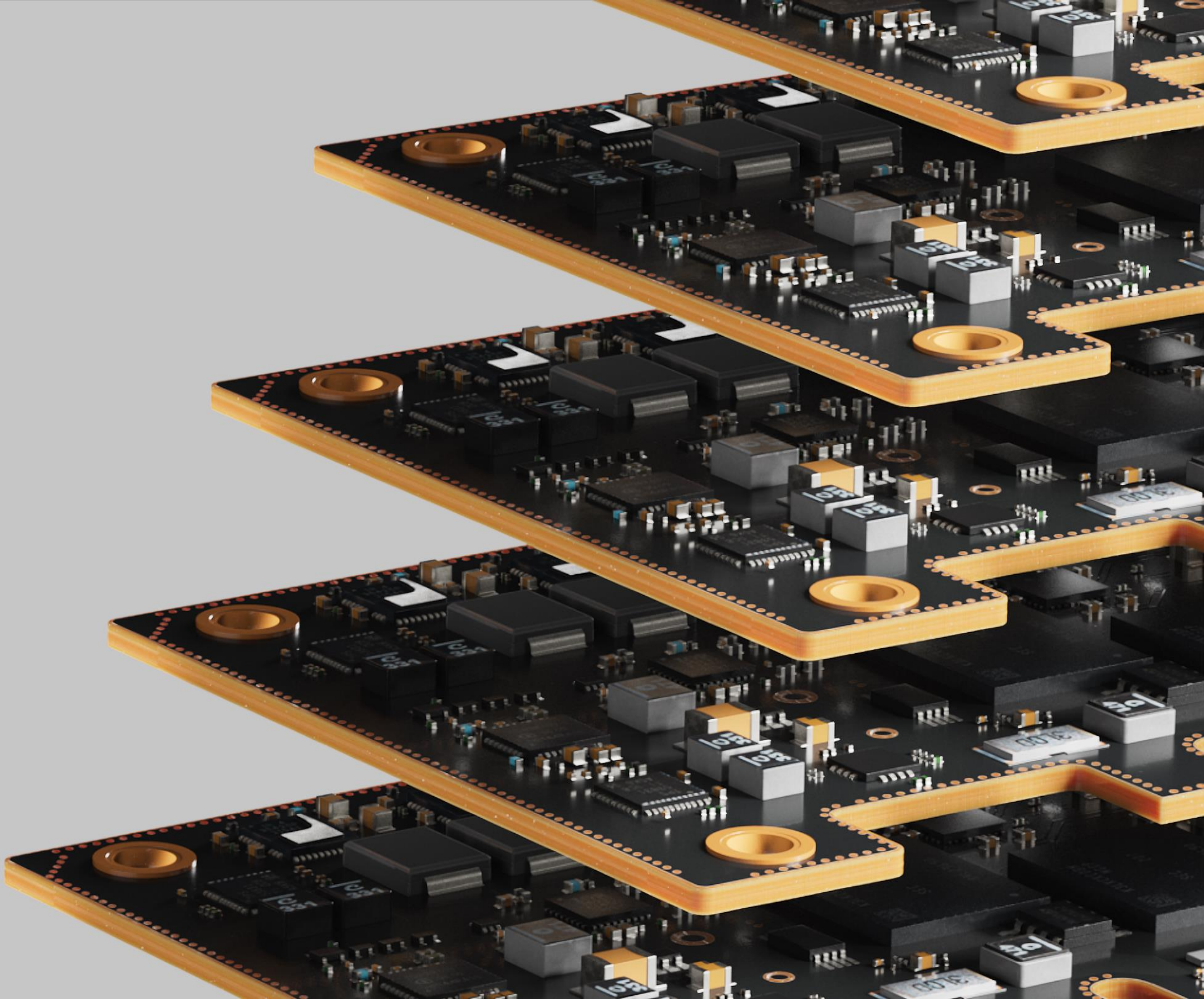

Traditional AI
Larger, more complex models
Generative AI
Code transformation
Code assistant and explanation
System admin assistant

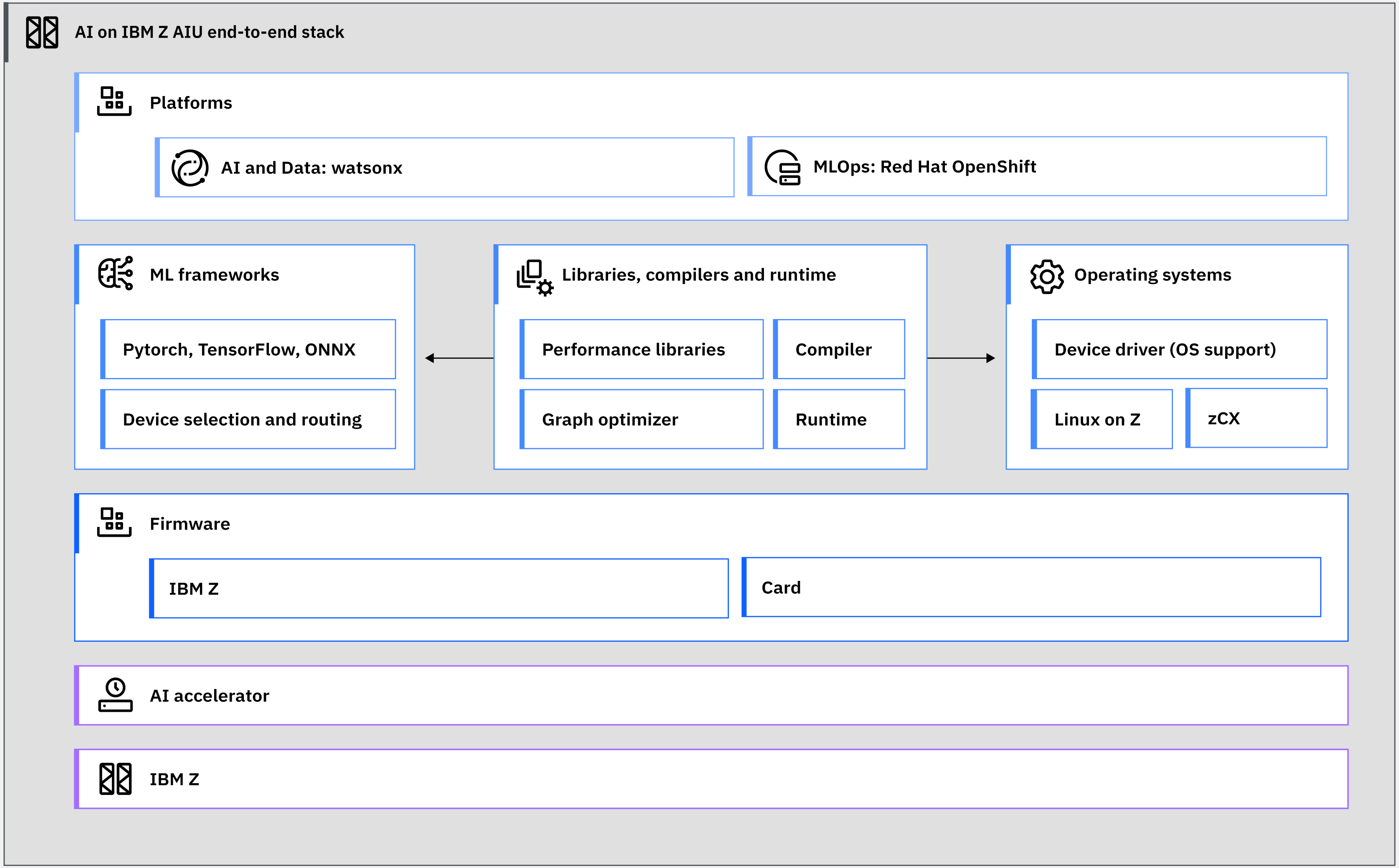
Designed to handle generative AI use cases

8 cards in I/O drawer form a logical cluster

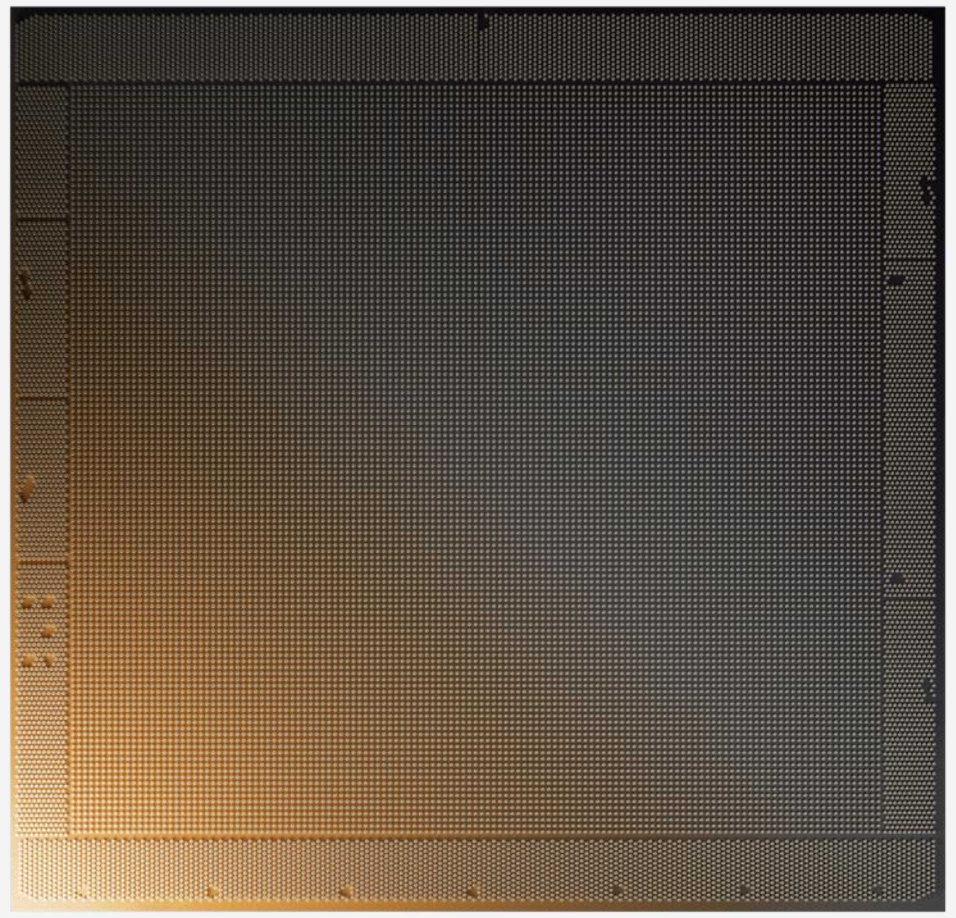
1TB of memory

1.6TB per second aggregate memory bandwidth



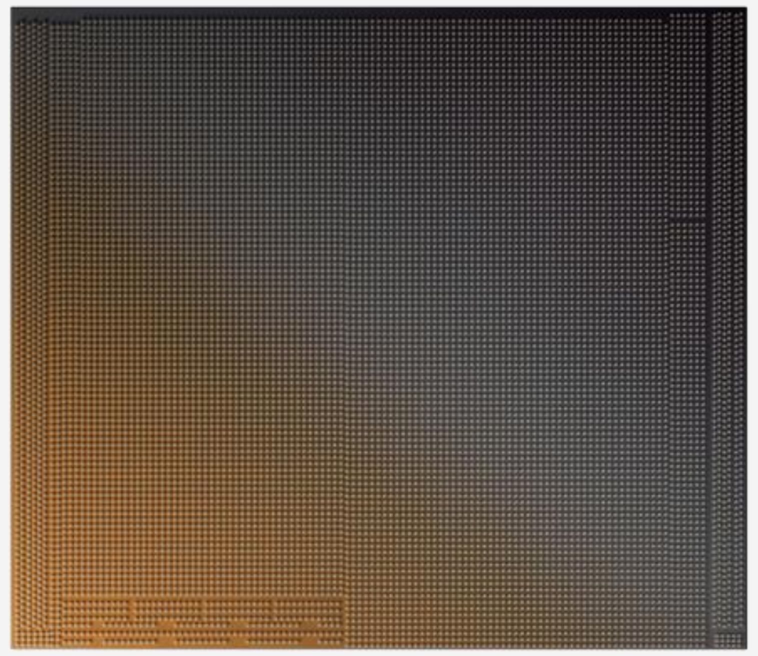

AI on IBM Z AIU end-to-end stack
Platforms
AI and Data: watsonx
MLOps: Red Hat OpenShift
ML frameworks
Pytorch, TensorFlow, ONNX
Device selection and routing
Libraries, compilers and runtime
Performance libraries
Compiler
Graph optimizer
Runtime
Operating systems
Device driver (OS support)
Linux on Z
zCX
Firmware
IBM Z
Card
AI accelerator
IBM Z

Telum® II



Samsung 5HPP technology
43 billion transistors
600 sqmm
24 miles of wire

Spyre™



Samsung 5LPE technology
26 billion transistors
330 sqmm
14 miles of wire

All projections are based on preview hardware.

30 Peta OPS in a system

Next generation AI acceleration
for next generation AI use cases

IBM

Thank you

References

1 - DISCLAIMER: Performance result is extrapolated from IBM internal tests running local inference operations in an IBM z16 LPAR with 48 IFLs and 128 GB memory on Ubuntu 20.04 (SMT mode) using a synthetic credit card fraud detection model (https://github.com/IBM/ai-on-z-fraud-detection) exploiting the IBM Integrated Accelerator for AI. The benchmark was running with 8 parallel threads each pinned to the first core of a different chip. The lscpu command was used to identify the core-chip topology. A batch size of 128 inference operations was used. Results were also reproduced using a z/OS V2R4 LPAR with 24 CPs and 256GB memory on IBM z16. The same credit card fraud detection model was used. The benchmark was executed with a single thread performing inference operations. A batch size of 128 inference operations was used. Results may vary.

IBM Spyre™ Accelerator is currently in tech preview