壁仞科技
BIREN TECHNOLOGY

壁仞™ BR100 GPGPU:

Accelerating Datacenter Scale AI Computing

Mike Hong, Lingjie Xu
& Team

Hot Chips 34, Aug 2022

Notice and Disclaimer

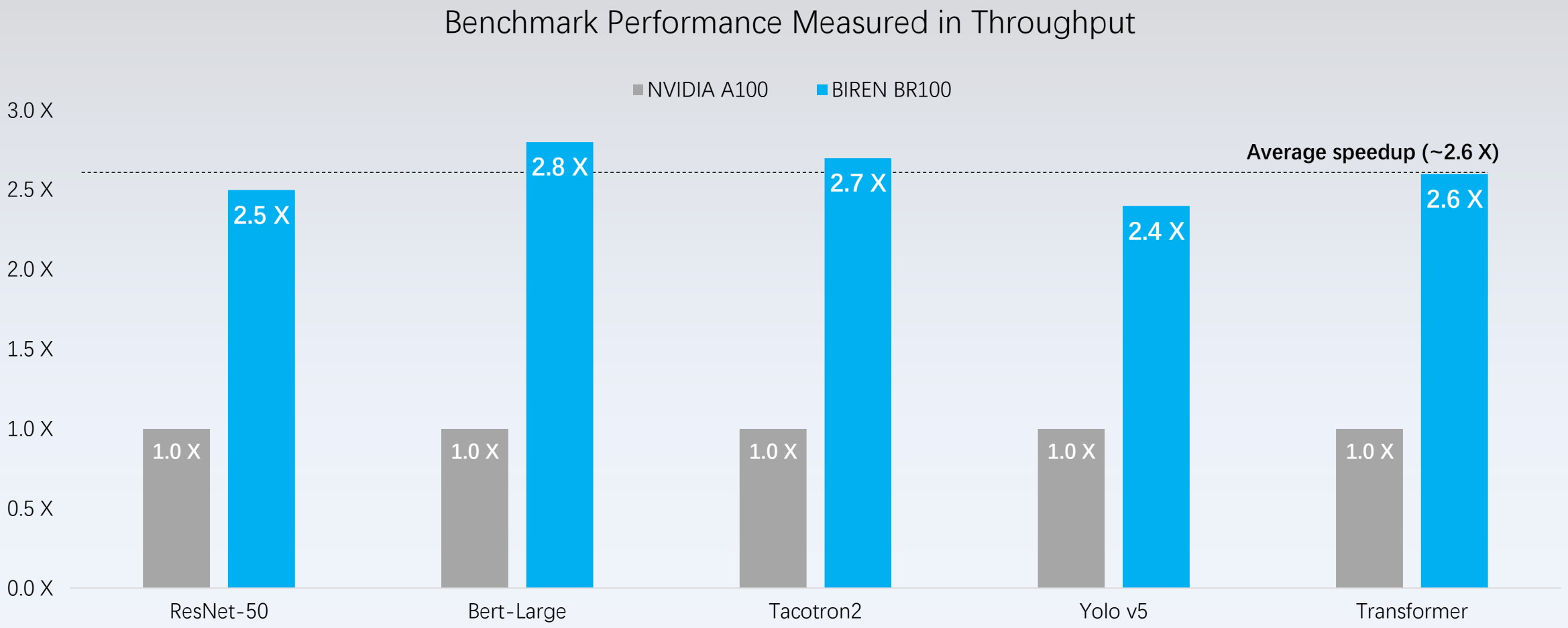
- **Copyright** © 2020-2022 Biren Technology. All rights reserved.
- **Confidentiality.** This document contains confidential information of Biren Technology, which shall not be disclosed to any third party by any means unless explicitly permitted by Biren Technology.
- **Trademark.** All trade names, trademarks, graphical marks and domain names in this document are the properties of Biren Technology. Without prior written consent, they may not be copied, reproduced, modified, published, uploaded, posted, transmitted, or distributed in any way.
- **Forward Looking Statements.** Information in this document, other than statements or description of historical fact, may contain forward-looking statements. The forward-looking statements are made based on certain assumptions, projections and calculations made by us with regards to the industry and management's expertise. These forward-looking statements are subject to significant risks and uncertainties and our actual results may differ materially. Your business decisions shall not be made solely based on the information.
- **Disclaimer.** This material is provided "as is" without any express or implied warranty of any kind, including warranties of merchantability, title, non-infringement of intellectual property, or fitness for any particular purpose.

壁仞科技
BIREN TECHNOLOGY

BIREN BR100

Area	1074mm^2 @7nm
Transistor Count	77Billion
Host Interface	PCIe Gen 5 x16 w/ CXL
Peak Performance	2048 TOPS @ INT8 1024 TFLOPS @ BF16 512 TFLOPS @ TF32+ 256 TFLOPS @ FP32 Also supports FP16, INT32, INT16 and other formats
Memory	64GB HBM2E
Interconnections	8 BLink™ 2.3TB/s external I/O bandwidth
Form Factor	OAM with 550W Max TDP

Accelerating Deep Learning Workloads



Purpose-built for Datacenter-Scale Computing

Compute density

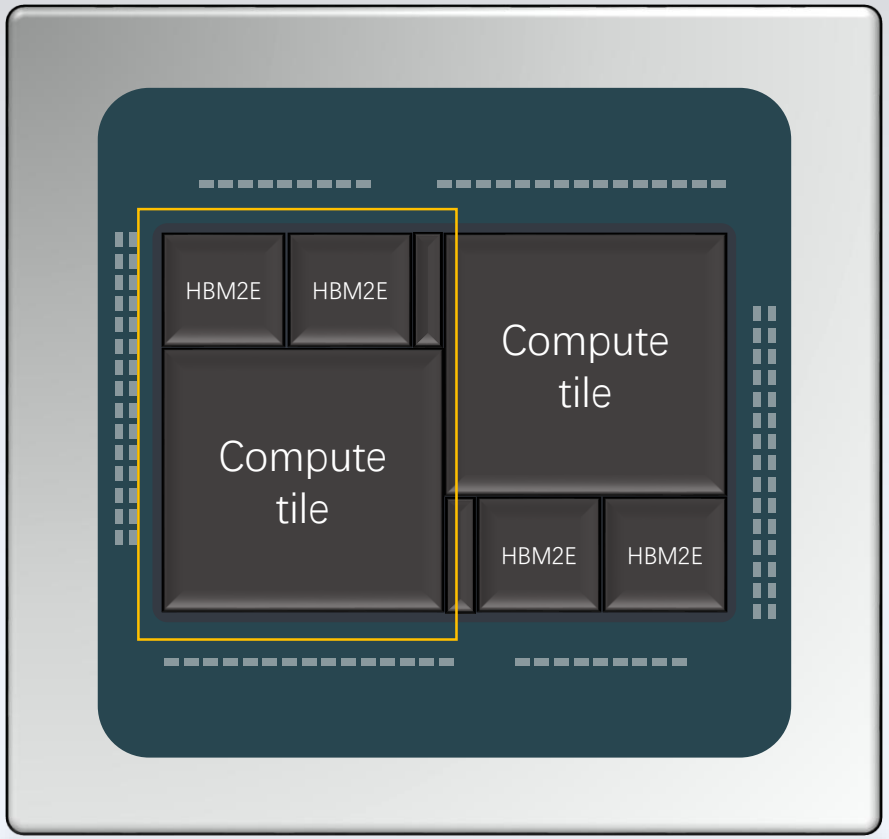
Connectivity

Cost of ownership

Co-design
hardware and software

Compatibility with
datacenter infrastructure

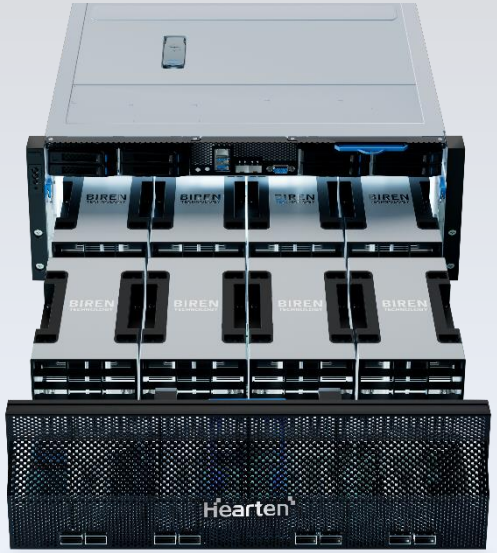
One Tapeout, Multiple Products



- ✓ To break the reticle size limit and integrate more transistors on chip
- ✓ One tapeout to empower multiple SKUs
- ✓ Smaller die for better yield, hence lower cost
- ✓ 896GB/s high speed die-to-die interconnect
- ✓ **30%** more performance, **20%** better yield compared with a monolithic design

BR100 Product Line

BR100



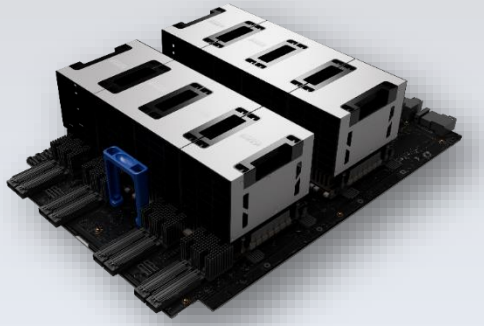
Hearten Server

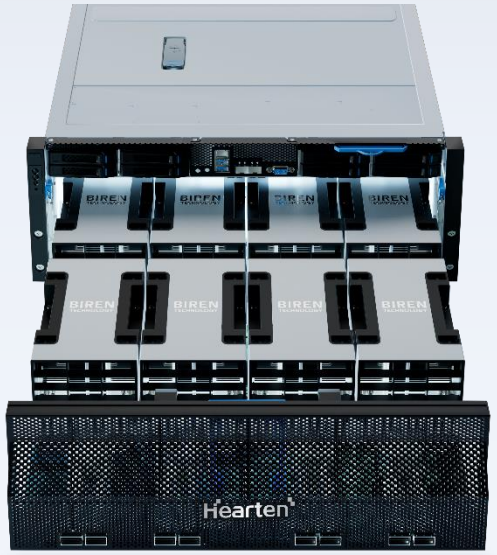
BR104

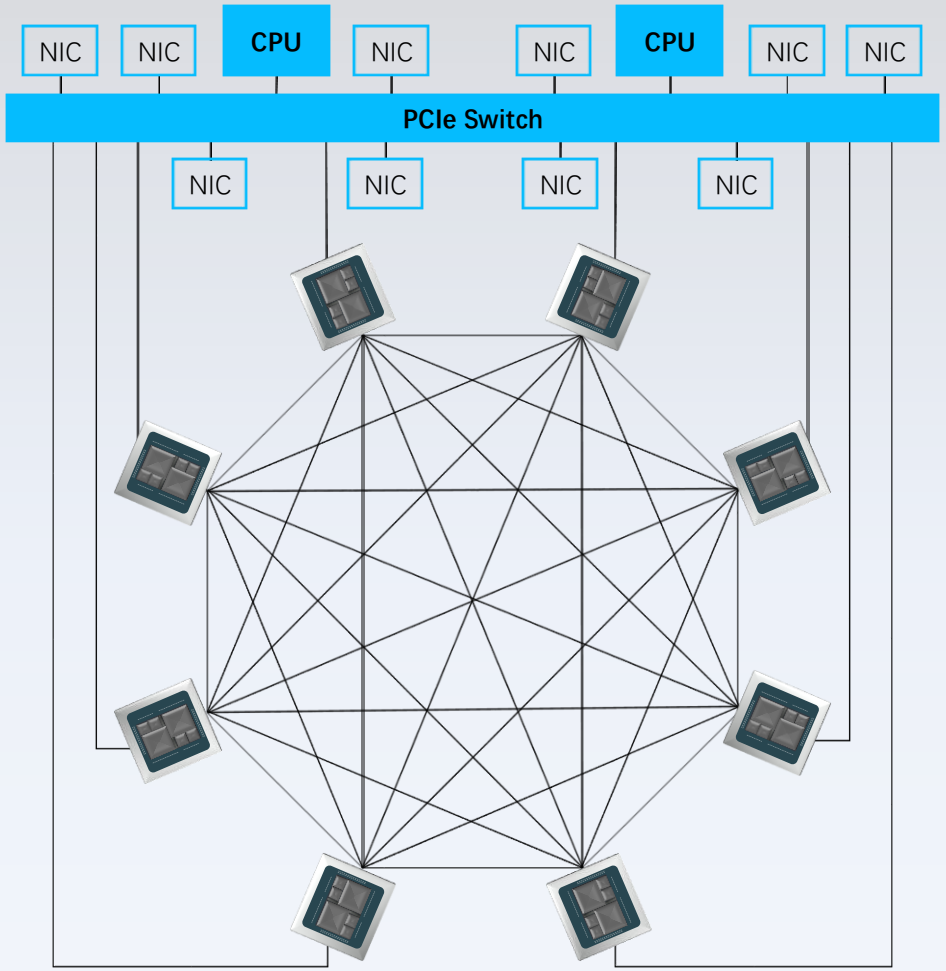
壁砺™ 100 OAM

壁砺™ 104 PCIe

OAM Server Interconnect Topology

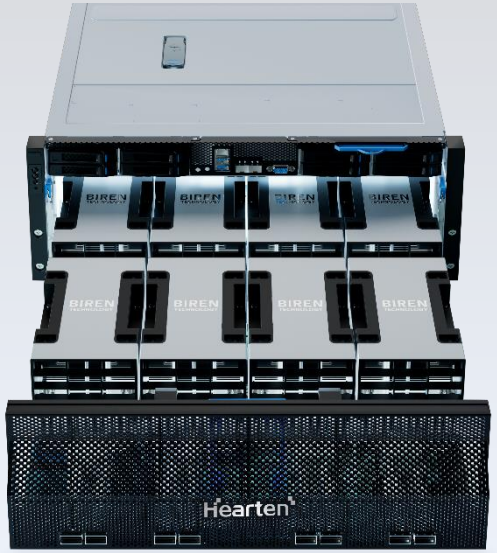






BR100 Product Line

BR100



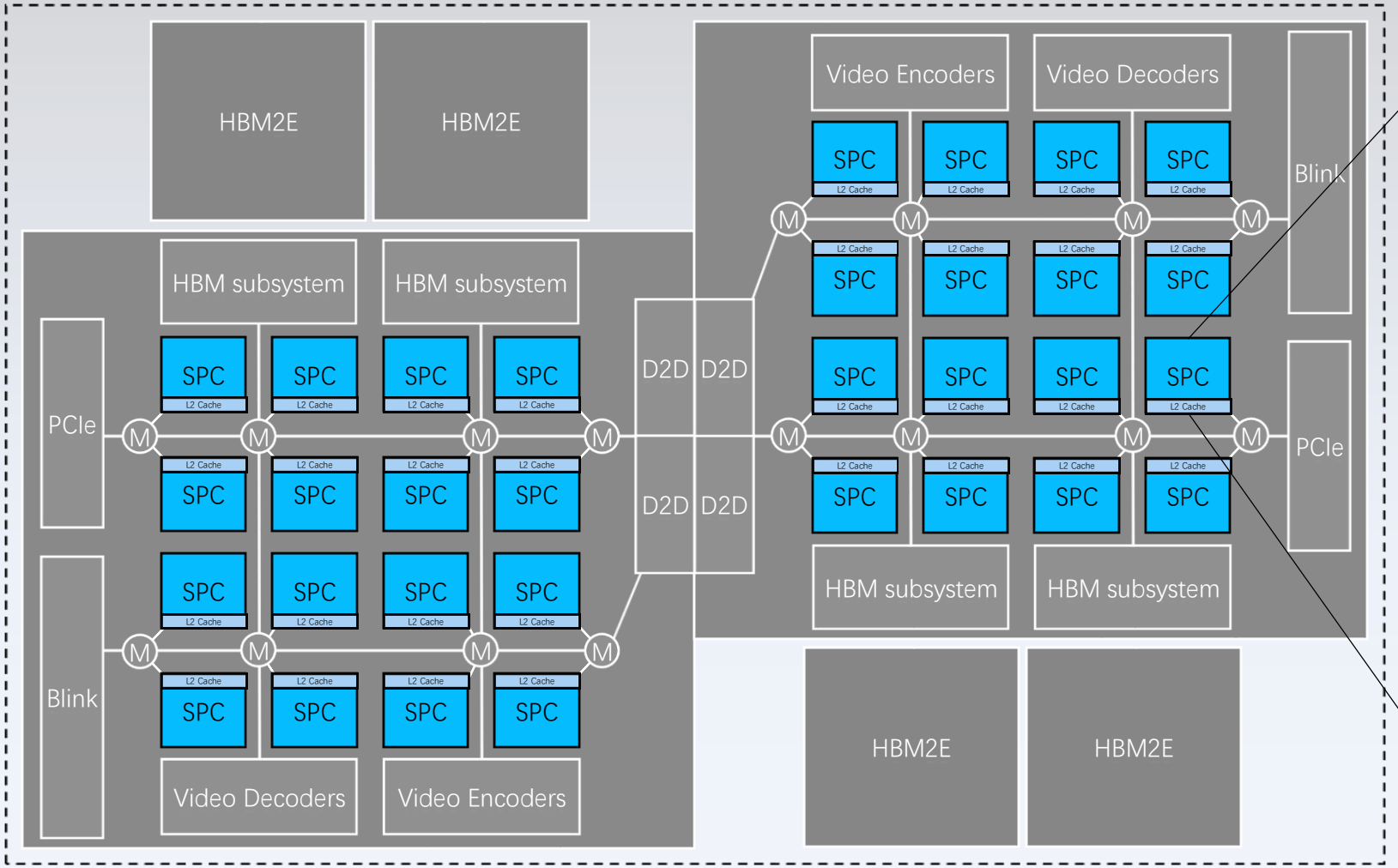
Hearten Server

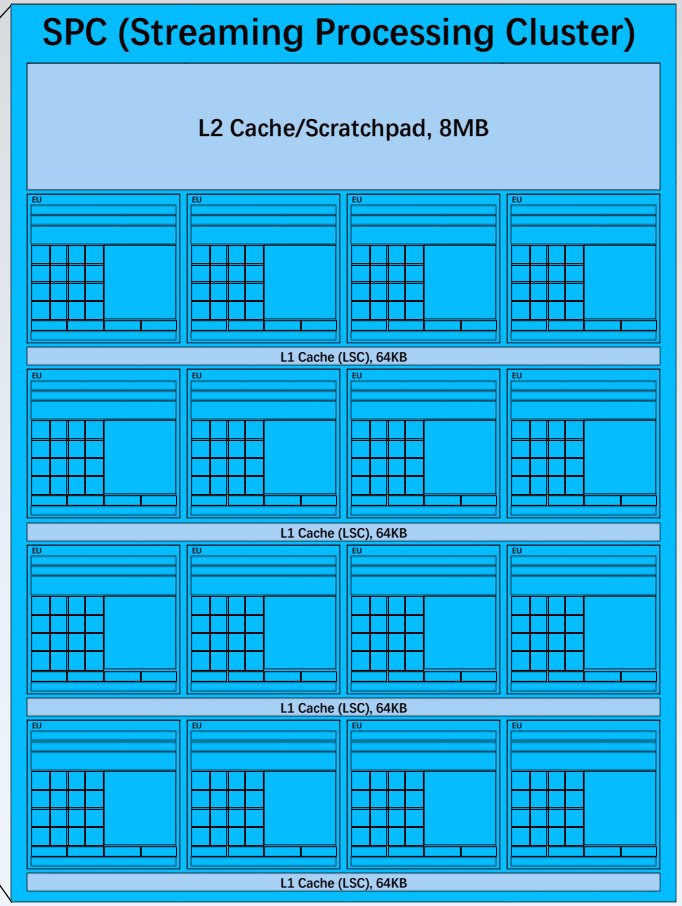
BR104

壁砺™ 100 OAM

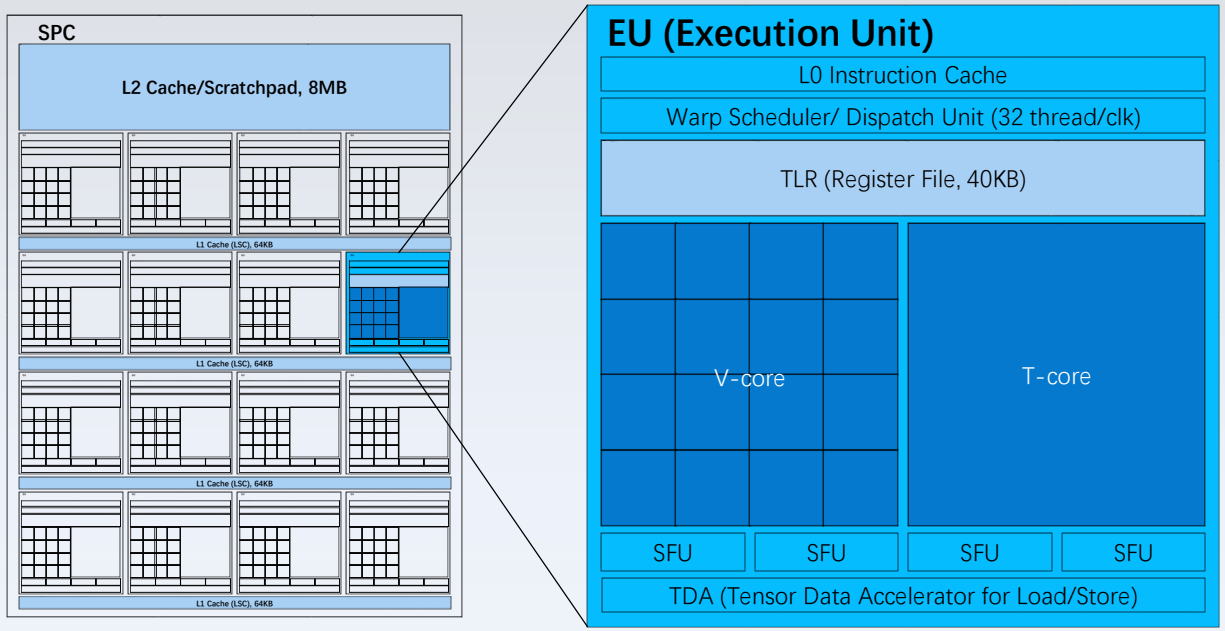
壁砺™ 104 PCIe

BR100 Architecture Diagram





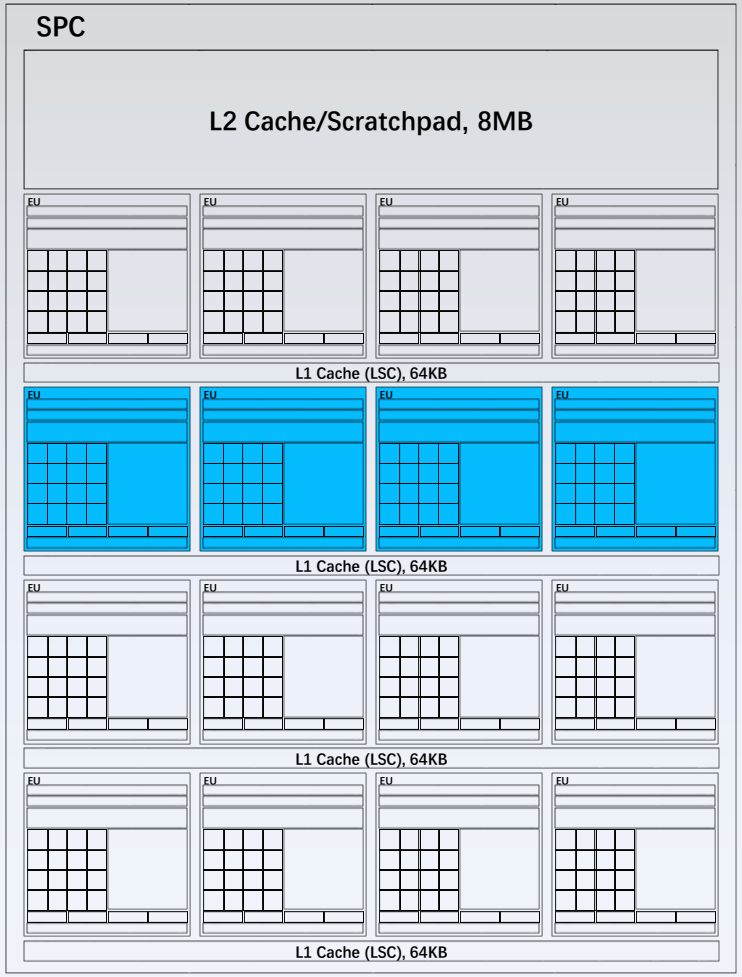
BR100 SPC Architecture



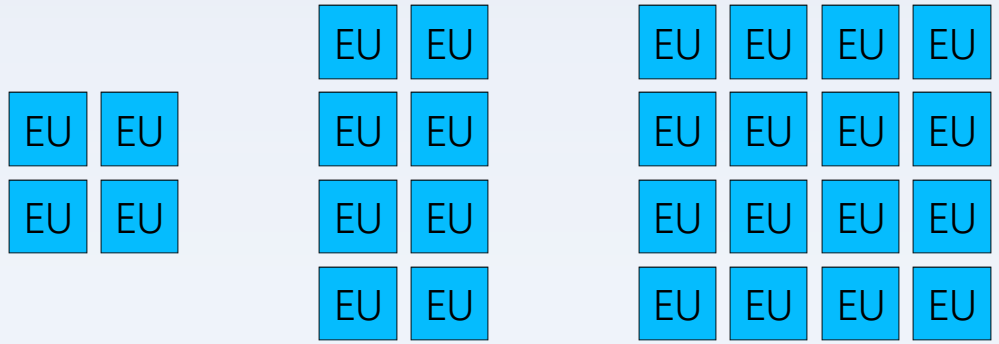
Building blocks of a BR100 SPC:

- ✓ **16 x EU (execution unit)**, each EU has:
  - 16 x streaming processing core (V-core), 1 x tensor engine (T-core)
  - 40KB TLR (Thread Local Register)
  - 4 x SFU
  - TDA (Tensor Data Accelerator)
- ✓ **4 x 64KB L1 Cache/LSC (Load & Store Cache)**
- ✓ Up to **8MB Distributed L2 Cache**
  - Holds shared data for all SPCs
  - Can be configured into **scratchpad**
  - Built-in reduction engine

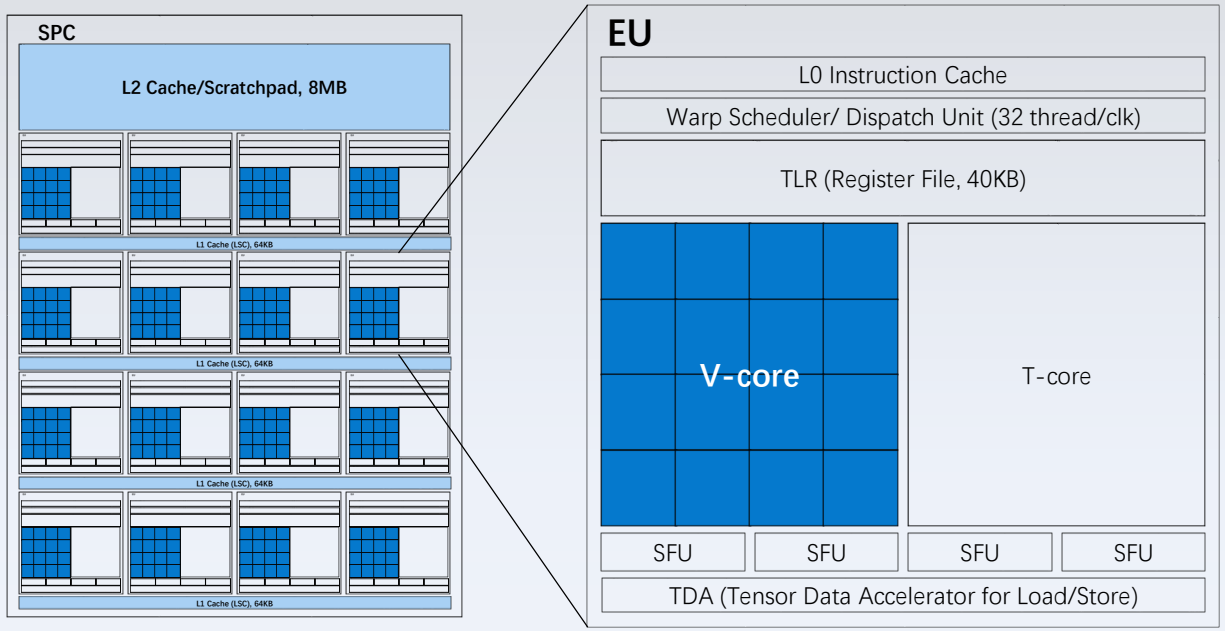
Scalable CU Architecture



- ✓ Multiple EUs form a **CU (compute unit)**
- ✓ Thread groups in a CU are synchronized
- ✓ **Each CU can contain 4/8/16 EUs**



V-core: A General-Purpose SIMT Processor



Full set ISA for general purpose computing

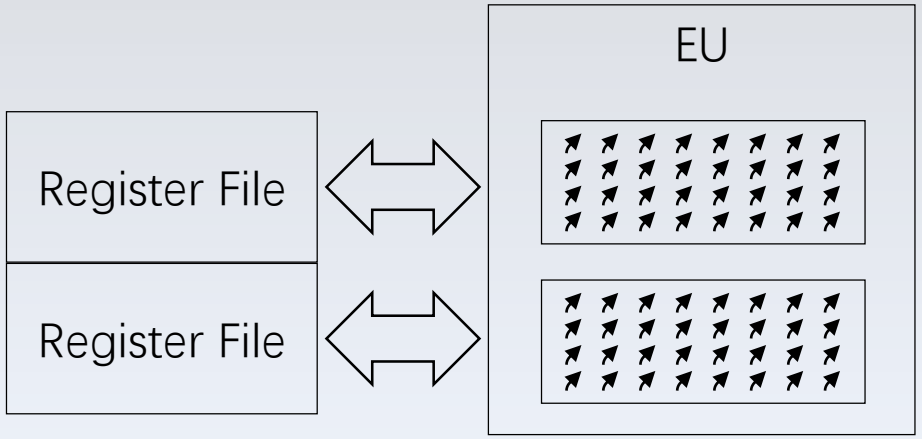
- ✓ 16x cores, supporting FP32, FP16, INT32, INT16
- ✓ SFU
- ✓ Load/Store
- ✓ Data preprocessing
- ✓ Manages T-core with multiple sync channels
- ✓ Handles DL OPs like Batch Norm, ReLu, etc

Enhanced SIMT Model

- ✓ 128K threads run on 32 SPCs
- ✓ Cooperative Warps
- ✓ Super-scaler (static and dynamic)

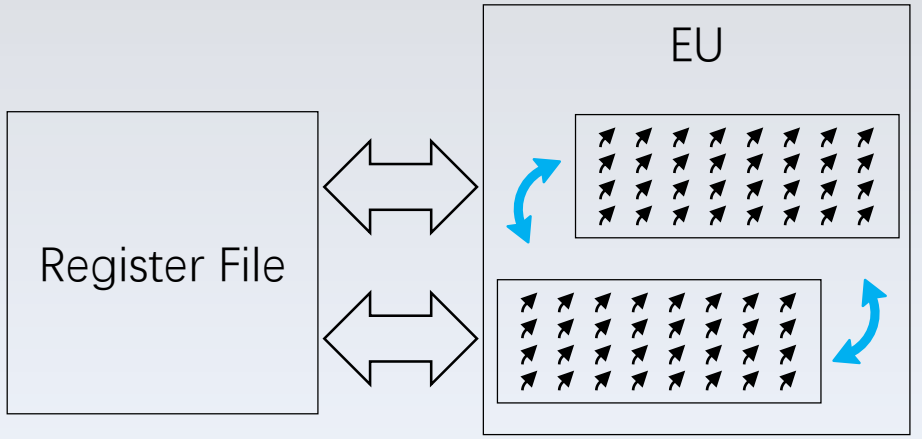
Flexible V-Core Warp Control

Standard Control Mode



Warps run the same code
and partition the register files

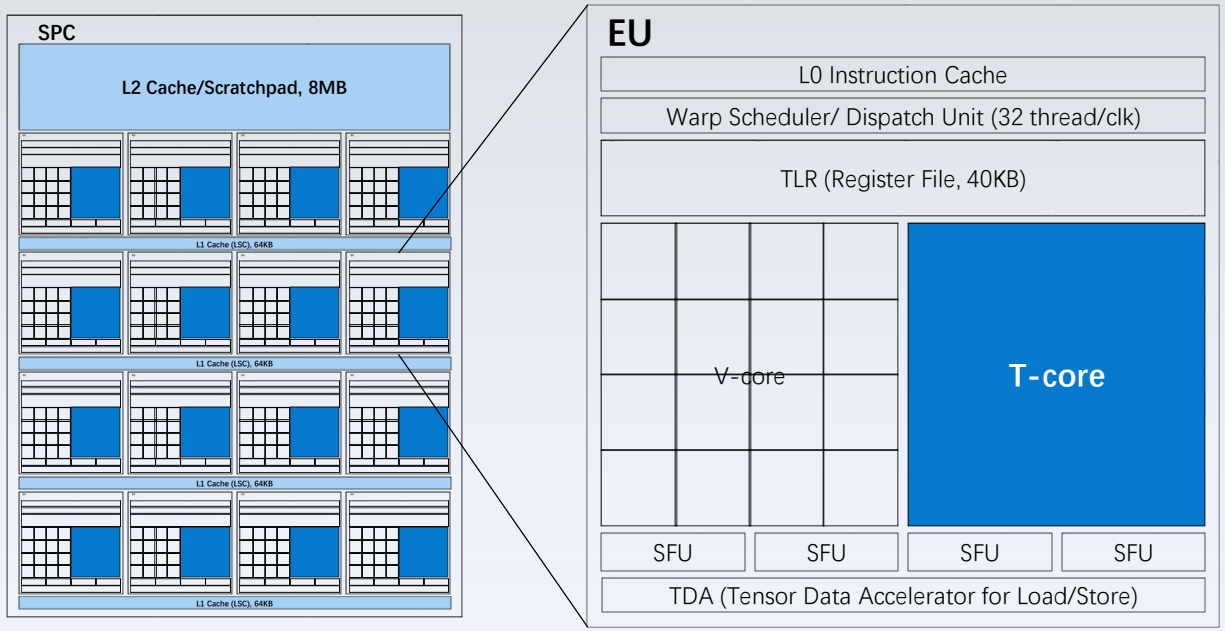
C-Warp/Kernel Coroutine Control Mode



Warps run different codes, collaborate with each other (producer-consumer) and exchange data in register files

- ✓ Enhanced parallelism
- ✓ Highly efficient for OP fusion

T-core: High Level Overview



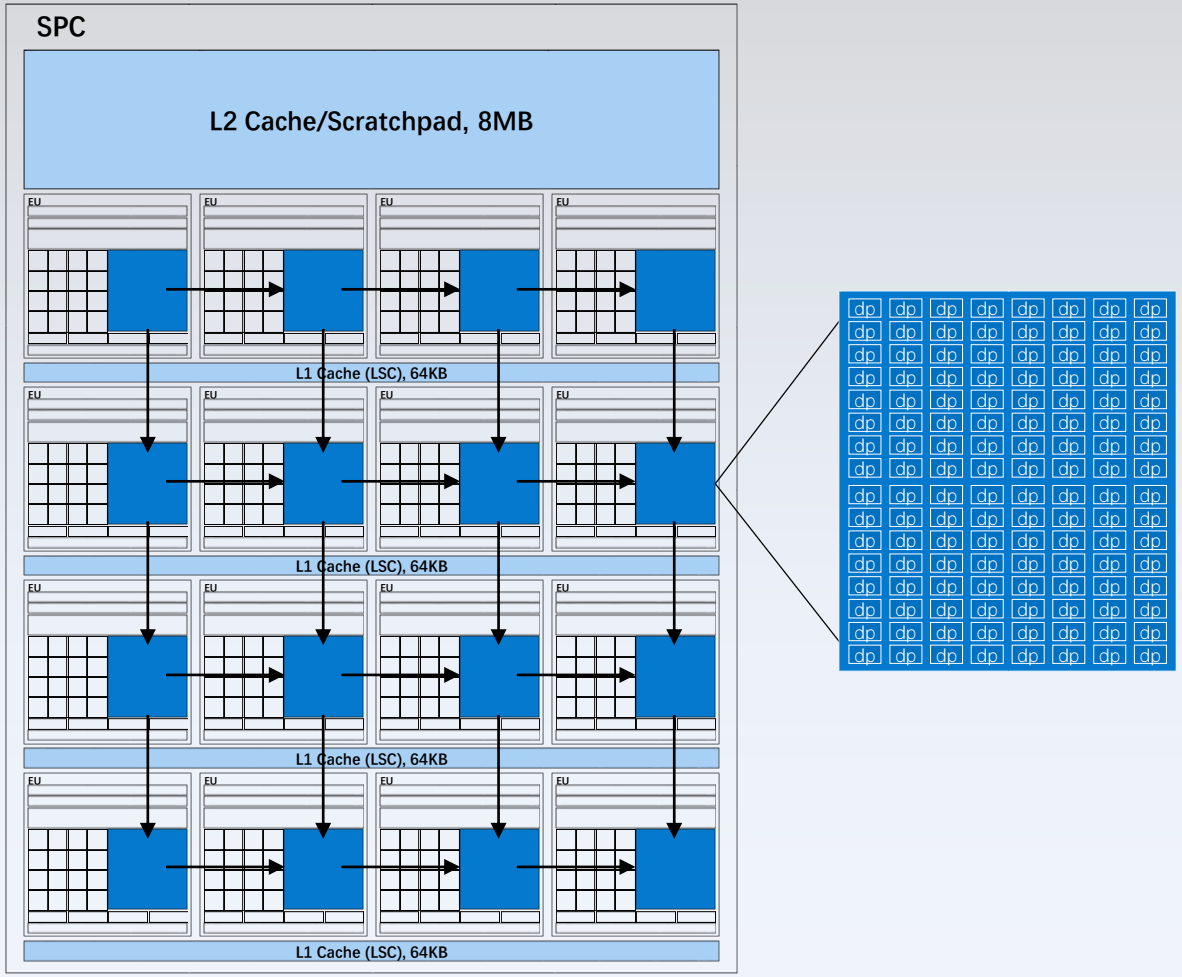
To accelerate common operations in AI:

- ✓ MMA (Matrix Multiplication Addition)
- ✓ Convolution
- ✓ …

Design Goals:

- ✓ To speed up MMA & convolution which account for majority workload in deep learning
- ✓ Best reuse among different dimensions (batch, sample, channel…)
- ✓ Model level and data level parallelism
- ✓ Closely coupled with V-cores and L2 scratchpad to maximize utilization

SPC-Scale 2.5D GEMM Architecture



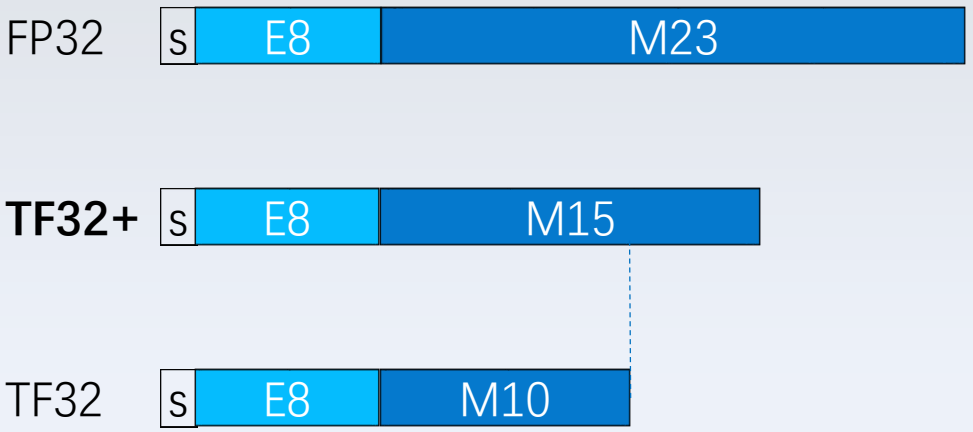
Consists of:

- ✓ 16 T-cores in a **2D** systolic array
- ✓ 2 groups of 8 x 8 dot product (dp) operations per T-core (8 x 8 x dp8 **3D MMA** for BF16)
- ✓ Equivalent to **64 x 64** Matrix Multiplication
  - Supports FP32, TF32+, BF16, INT16, INT8, INT4 tensor formats

Benefits:

- ✓ Higher data reuse rate with a big 2.5D GEMM
- ✓ Less memory bandwidth and cache occupation
- ✓ Better throughput and energy efficiency
- ✓ Lower latency v.s. big 2D GEMM
- ✓ More scalable and easier routing v.s. 3D GEMM

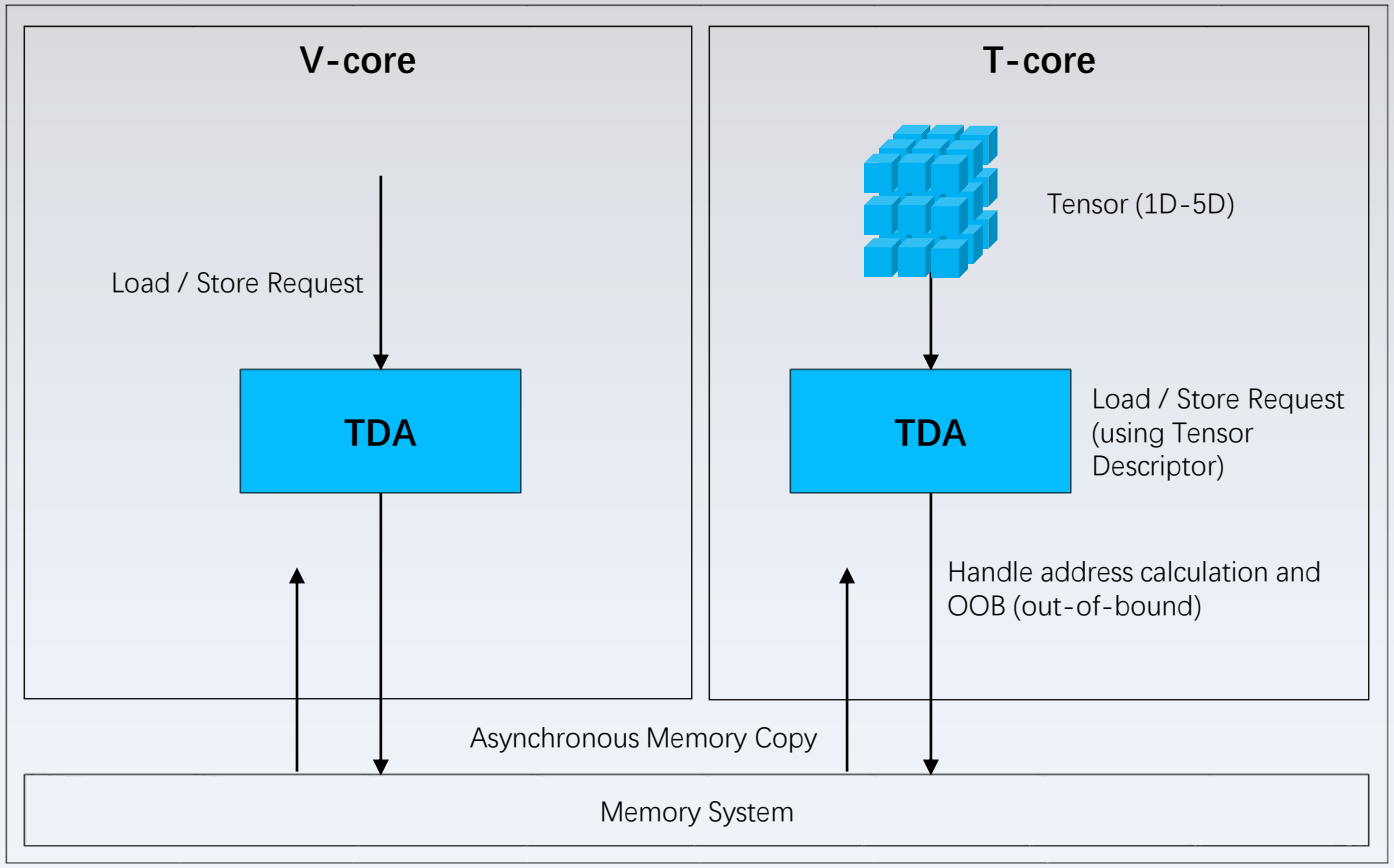
TF32+ Tensor Data Type



Why TF32+?

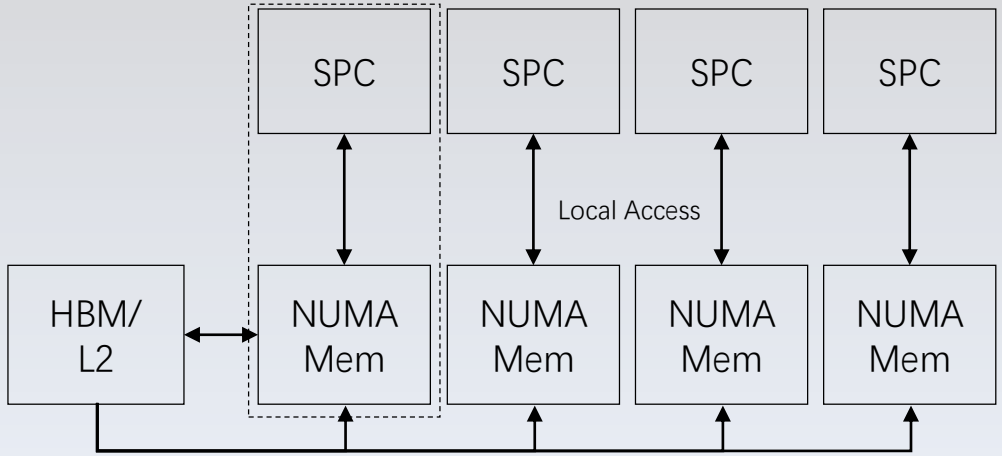
- ✓ E8M15, with 24 bits in total
- ✓ 32x more precise compared to TF32 in AI training
- ✓ To reuse BF16 multiplier (with 1+7 mantissa) and simplify T-core design
- ✓ Automatically kicked in when using tensor acceleration libraries and declared as FP32

Tensor Data Accelerator (TDA)



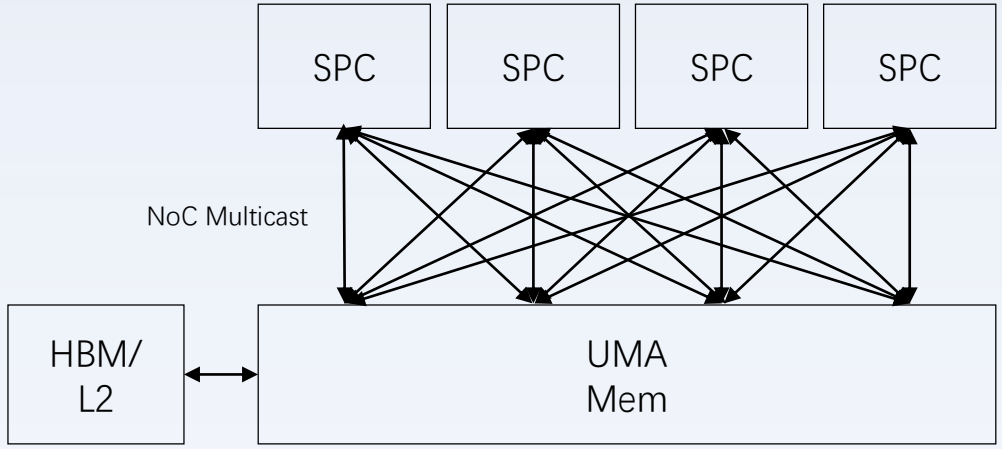
- ✓ TDAs in T-core and V-core are dedicated to accelerate address calculation and OOB using tensor descriptor
- ✓ TDA improves tensor data fetch efficiency by offloading addressing overhead and supporting different tensor layouts

Memory Scheme: NUMA and UMA



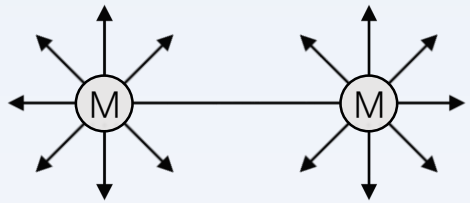
NUMA memory scheme

- ✓ Stores SPC's private data, e.g. activation
- ✓ Local access with high bandwidth
- ✓ Data broadcasted to relevant SPC in model parallelism



UMA memory scheme

- ✓ Stores shared data used by all SPCs, e.g. weights
- ✓ Accelerated by NoC multicasting

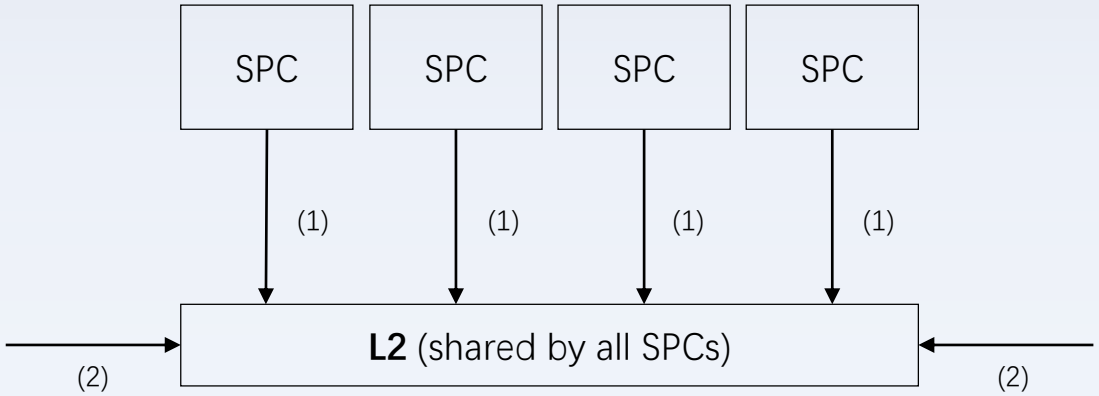


Near Memory Computing

Near Memory Engine: L2 Reduction

- ✓ Reduction operations in L2 by SPC instructions (1)
- ✓ Reduction by DMA (2)

- ✓ To offload SPCs and accelerate computation

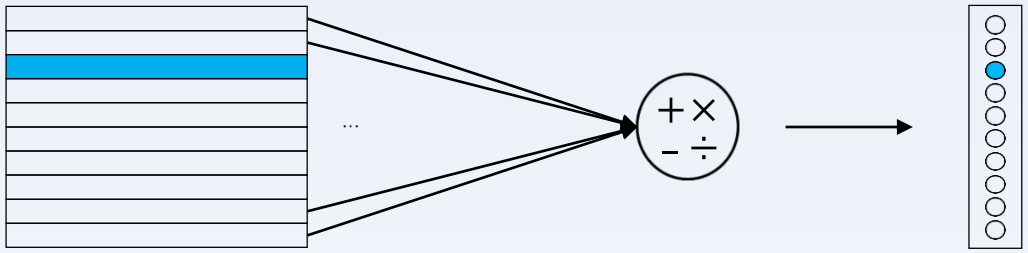


(L2 offloading SPC with reduction engine)

Embedding Accelerator

- ✓ Embedding table processing accounts for major computation in recommendation systems

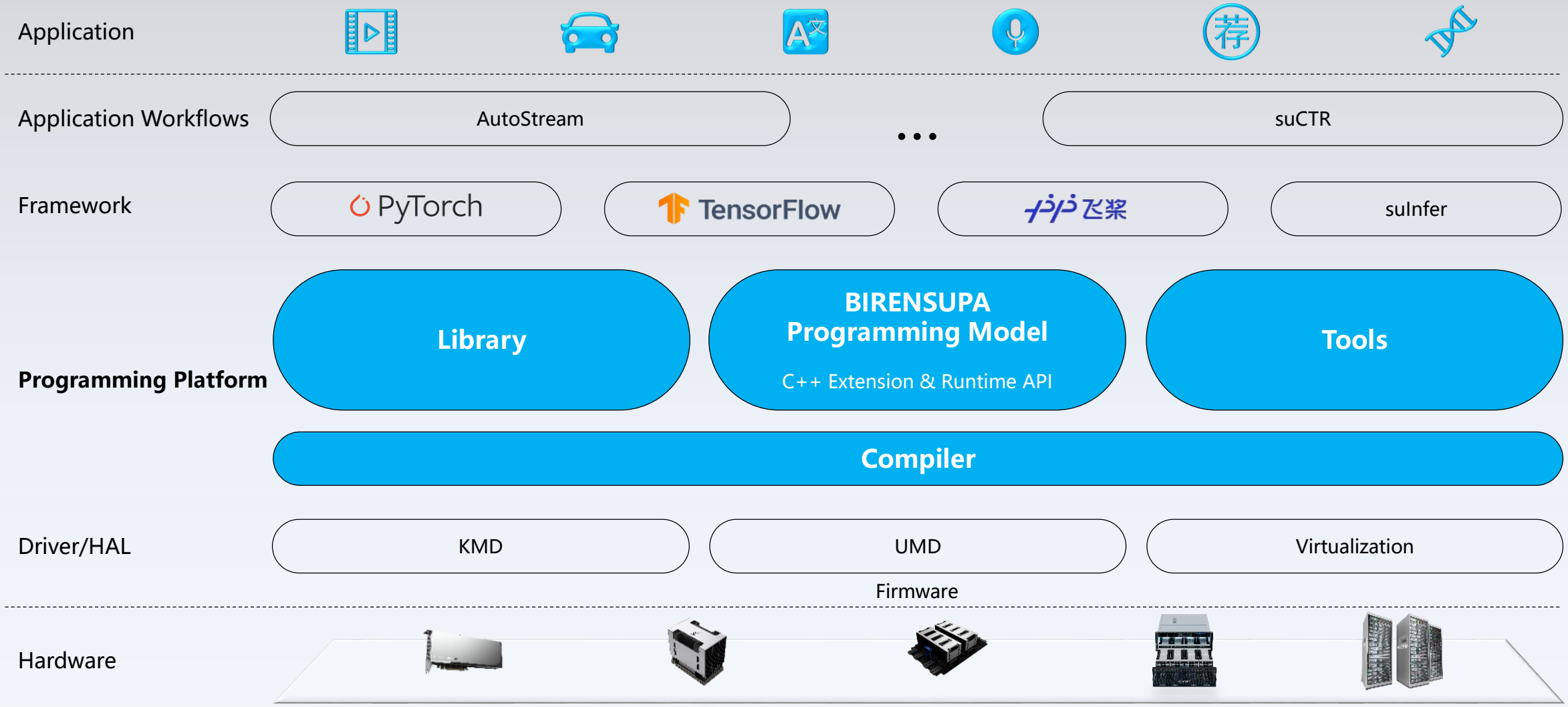
- ✓ Reduces the embedding table in L2 before SPC



Embedding Table

Reduction in L2

BIRENSUPA™ Software Platform



Summary

✓ **We introduced BR100, a GPGPU designed for accelerating datacenter-scale AI computing**

- PFLOPS level compute density, high connection bandwidth (on-die/off-die)
- 7nm chiplet design with CoWoS packaging
- 550W OAM form factor with 8 cards all-to-all interconnection topology
- Over 300MB on-chip SRAM for data cache and reuse
- A general-purpose processor with 2.5D GEMM acceleration
- A new TF32+ tensor data type

✓ **We optimized data streaming by introducing following features in BR100:**

- Special C-Warp control mode
- TDA
- NUMA/UMA memory scheme with NoC multicast
- Near memory computing

✓ **BIRENSUPA software platform allows developers to program on BR100 with ease**