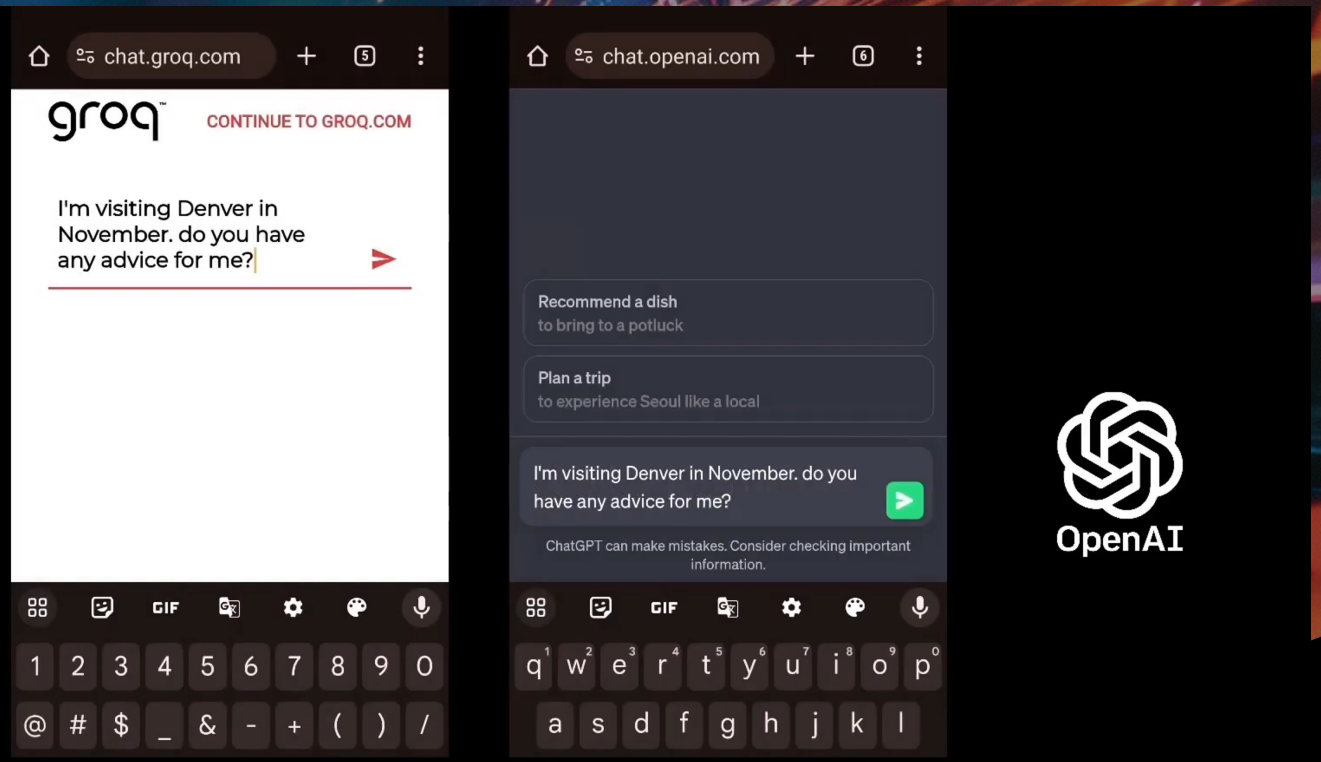
The New Era of AI Computing Power

The Convergence Breakthrough of RISC-V, Domain-Specific Architectures, and Near-Memory Computing

Peng Gao

- Senior VP of SOPHGO
- Bachelor's and Master's degree from BUPT (Beijing University of Posts and Telecommunications)
- Worked in the IC design field for 20 years
- More than 10 invention patents
- Specializing in artificial intelligence and medium to large scale SoC chip design
- Earliest explorers and practitioners in the field of artificial intelligence and RISC-V chips in China

What is Computing Power?



groq™

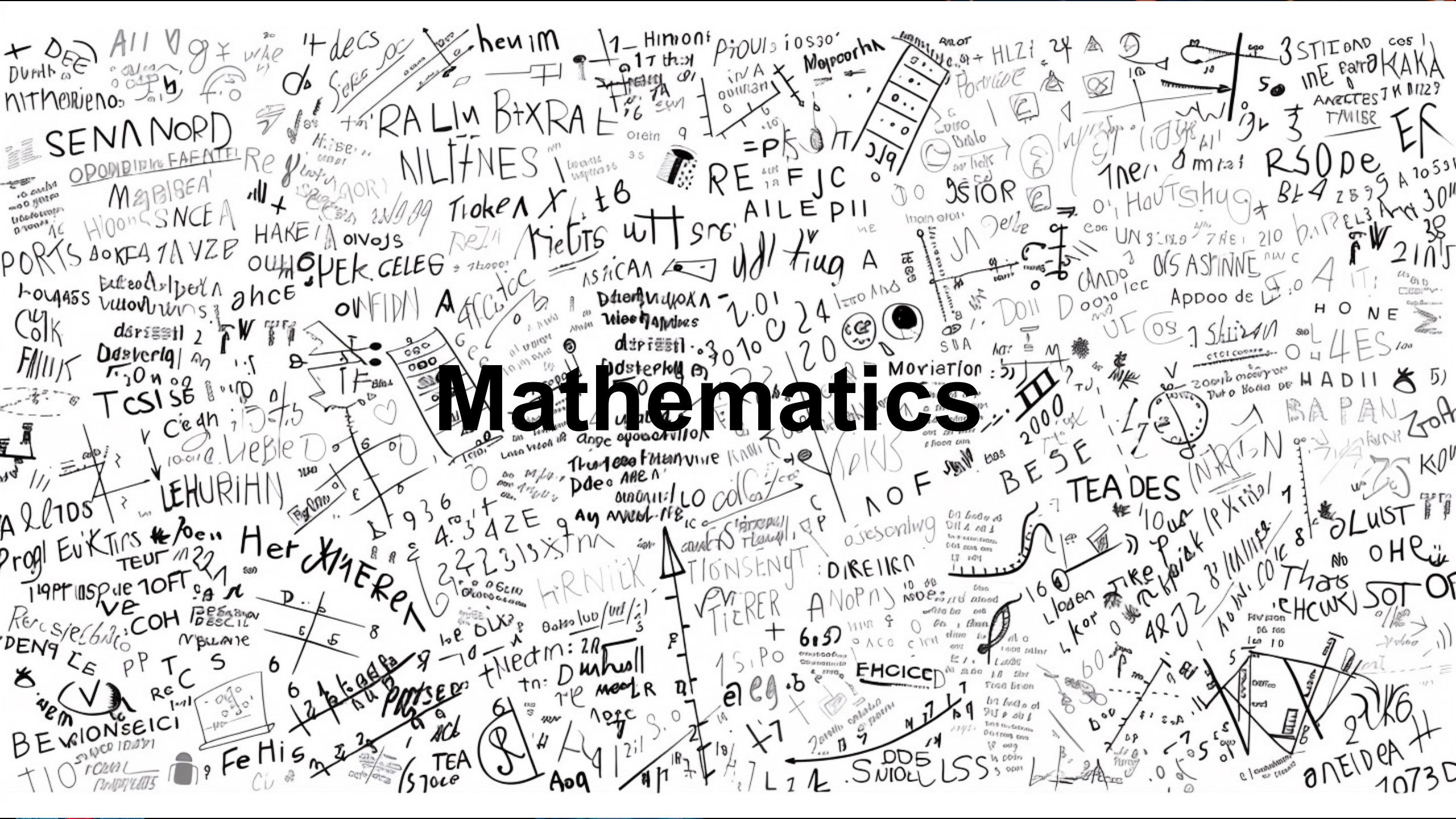
OpenAI

Take a look at this video.

Why?

	Groq	H100
Process Node	14nm	4nm
Single Chip computing power	188@fp16	989@fp16
Intra-node bandwidth GB/s	480	NVLinks 900
Single Chip memory GB	0.22 (SRAM)	80 (HBM)
Memory bandwidth TB/s	80	3.35
Number of chips required for inference (70GB)	319	1

What is the difference between the computation of LLMs and CV in AI computing?


Mathematics

What is LLM? Take LLaMA2-7B as an Example

Main Compute in Prefill

Input	Weight	Output	Times
[len, 4096]	[4096, 4096]	[len, 4096]	96
[len, 128]	[128, len]	[len, len]	1024
[len, len]	[len, 128]	[len, 128]	1024
[len, 4096]	[4096, 4096]	[len, 4096]	32
[len, 4096]	[4096, 11008]	[len, 11008]	64
[len, 11008]	[11008, 4096]	[len, 4096]	32
[len, 4096]	[4096， 32000]	[len, 32000]	1

Main Compute in decode

Input	Weight	Output	Times
[1, 4096]	[4096, 4096]	[1, 4096]	96
[1, 128]	[128, 4096]	[1, 4096]	1024
[1, 4096]	[4096, 128]	[1, 128]	1024
[1, 4096]	[4096, 4096]	[1, 4096]	32
[1, 4096]	[4096, 11008]	[1, 11008]	64
[1, 11008]	[11008, 4096]	[1, 4096]	32
[1, 4096]	[4096， 32000]	[1, 32000]	1

- From a mathematical perspective, AI is all about matrix multiplication.

Human: What is the apple?
LLAMA: Apple is a fruit.

Firstly, “What is the apple?” is embedded to a [5, 4096] matrix, then the matrix is put into the LLAMA and the first output “Apple” is generated. This stage is called prefill.

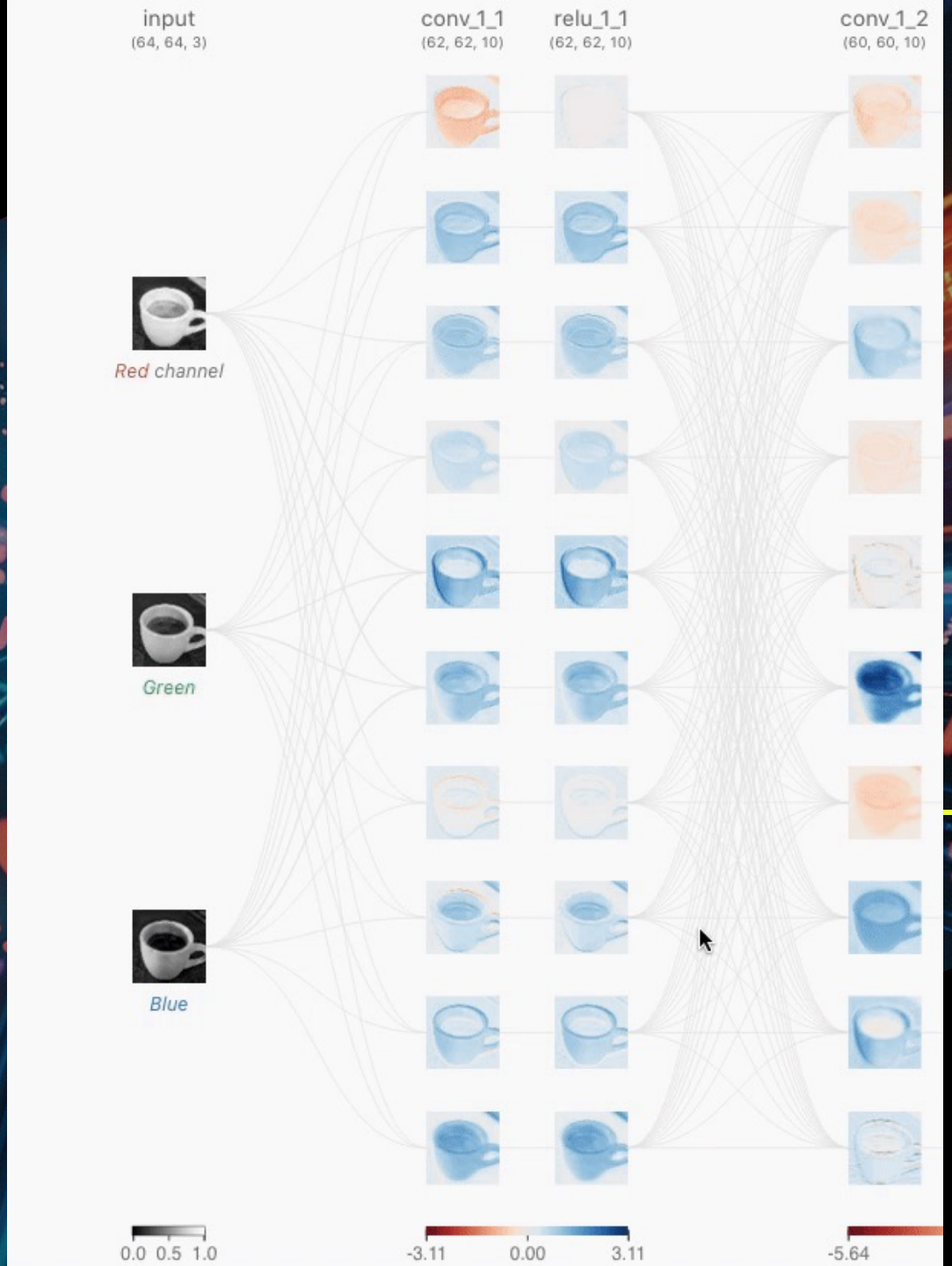
Secondly, “What is the apple? Apple” is embedded to a [6, 4096] matrix, then the matrix is put into the LLAMA and the second output “is” is generated.

There are some repeated compute between these two steps. After saving the repeated compute, only “Apple” is put into the LLAMA. This stage is called decode.

What is CNN?

- Convolutional Layers
- Pooling Layers
- Flatten Layer
- Activation Functions
  - ReLU
  - Softmax

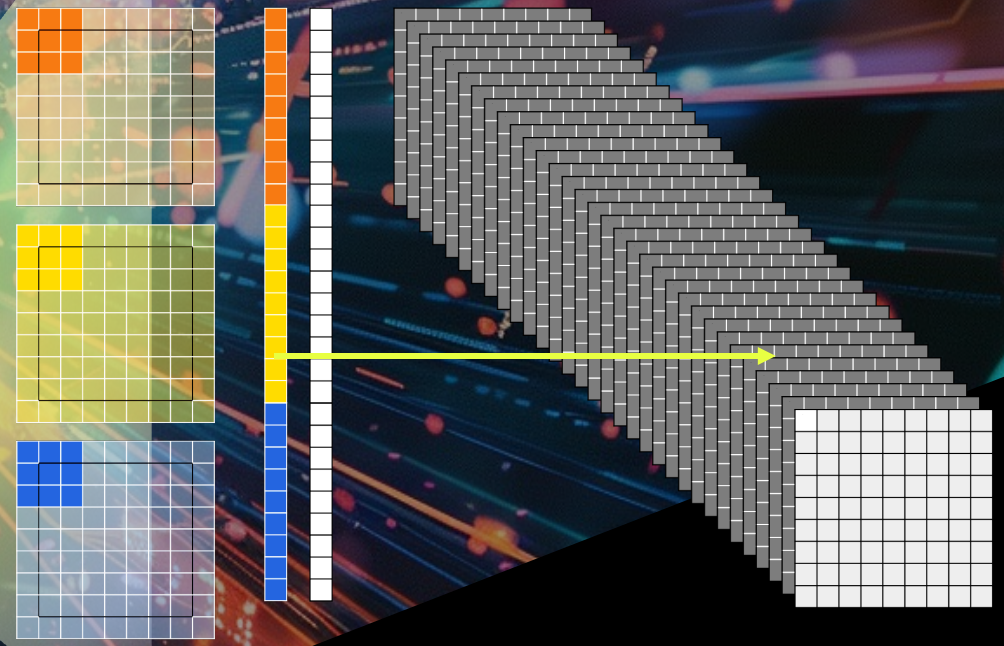
Refer to
https://poloclub.github.io/cnn-explainer/
for more about CNN



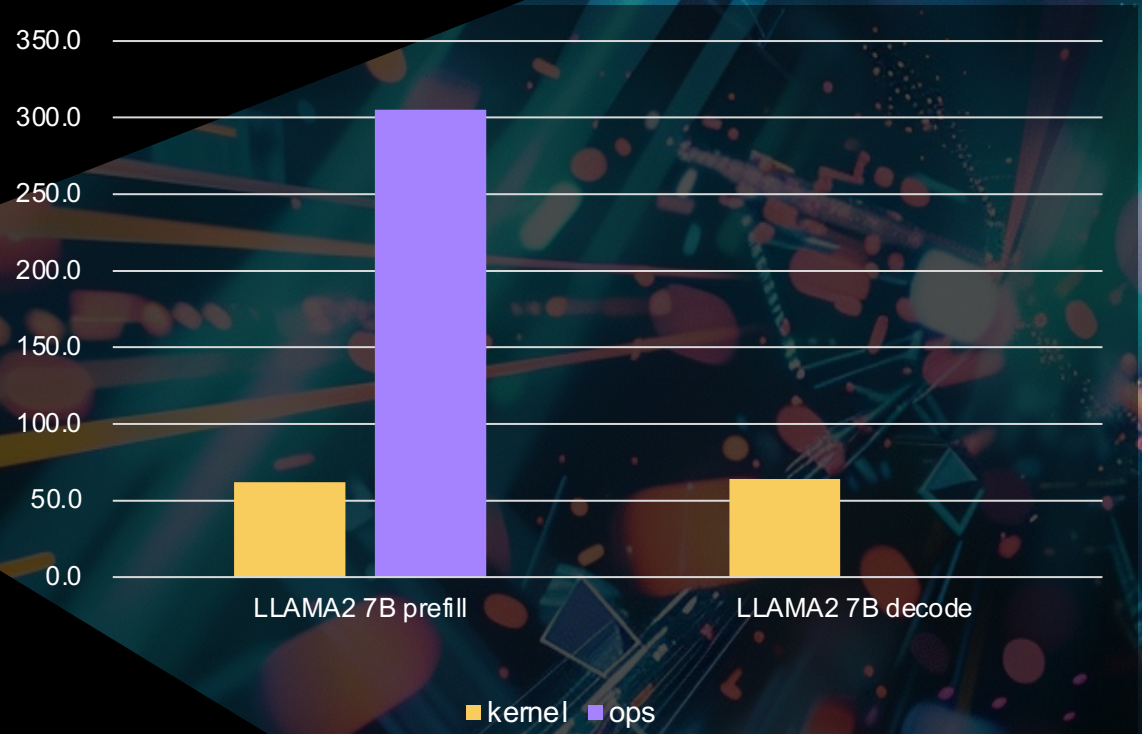
中辰計劃

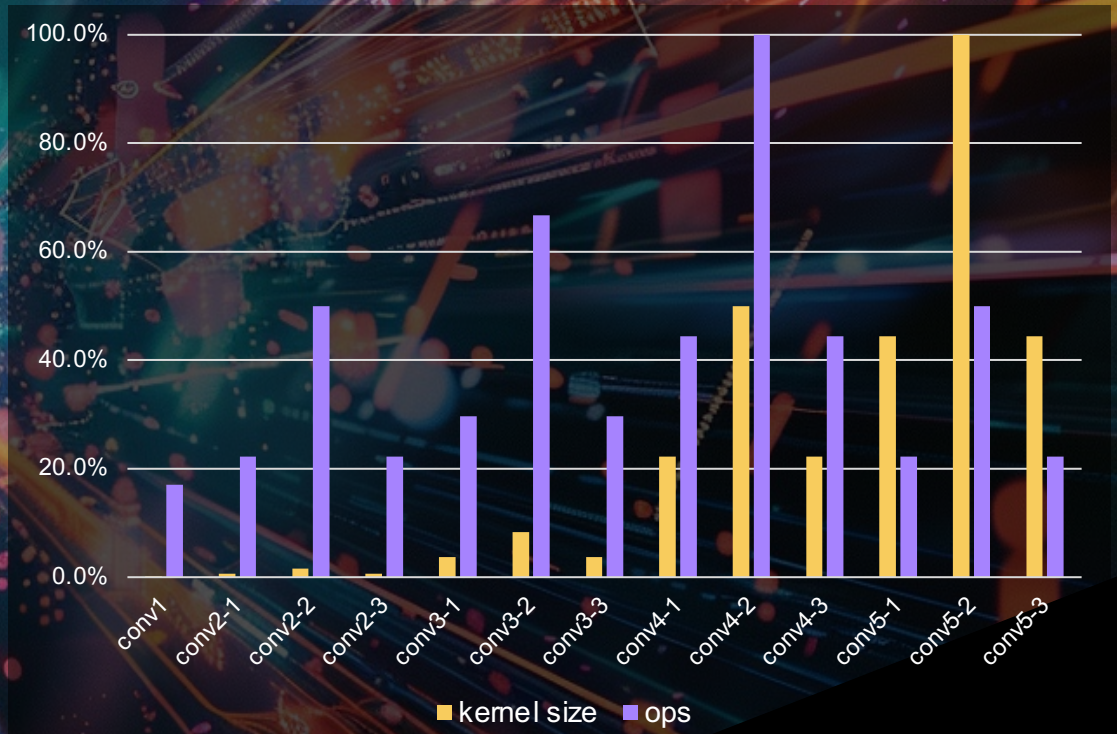
CNN: Resnet50

- A lot of data permuting needed before performing matrix multiplication.
- Sharing weights significantly reduce the amount of data transmission compared to the amount of calculation.



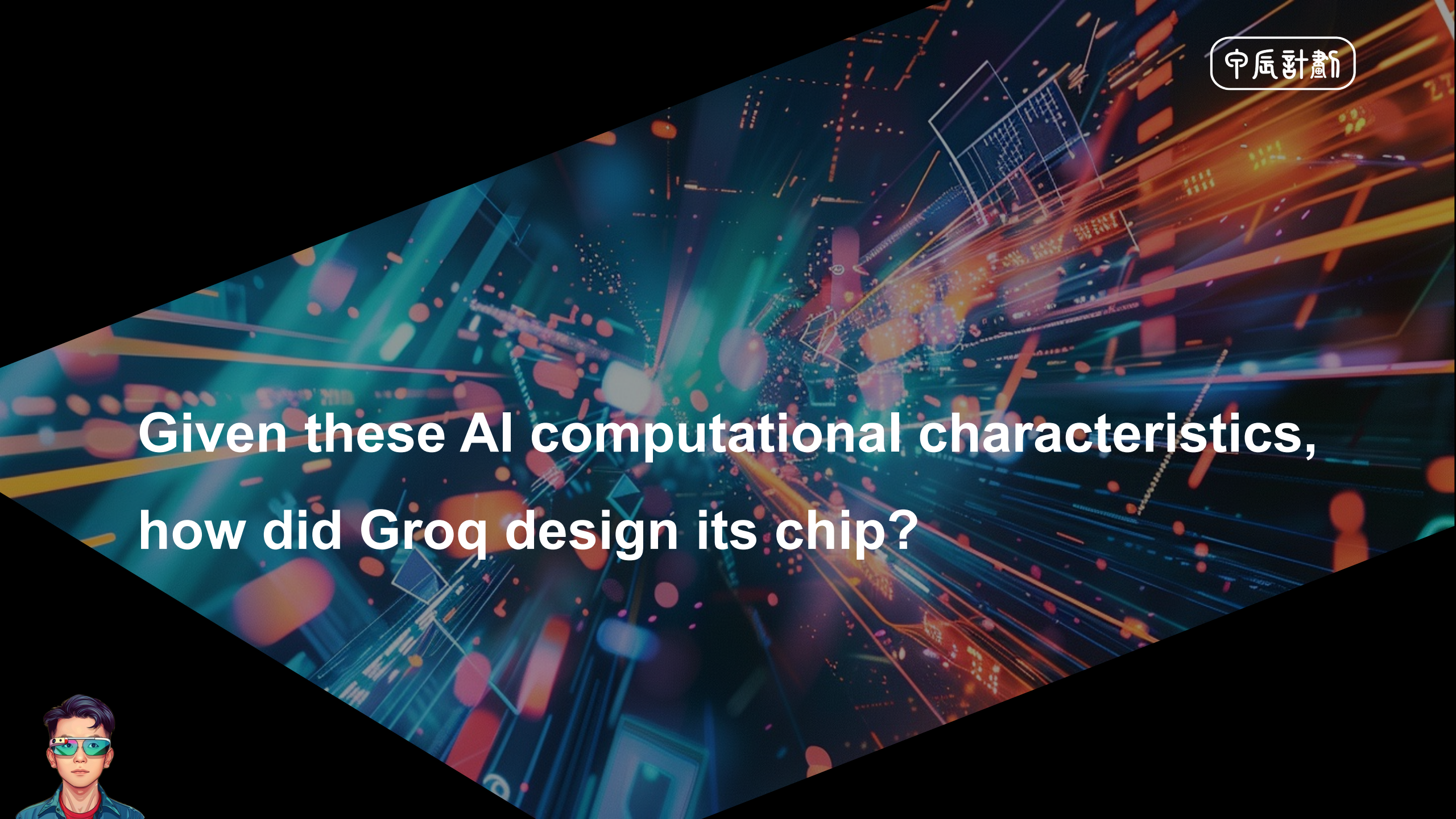
LLaMA2-7B vs Resnet50
Profile of Compute & Bandwidth Requirement

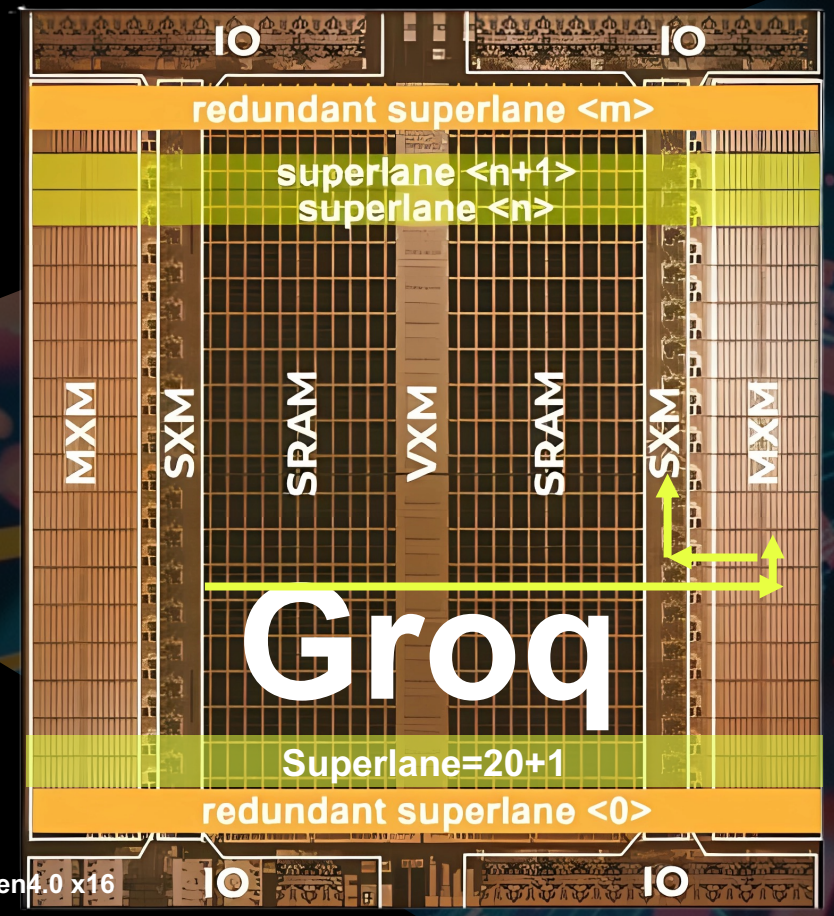




From the perspective of comparing the amount of data transmission and computing

- The prefill and decode stages of llama are completely opposite
- The different stages of the resnet50 execution process vary greatly


Given these AI computational characteristics,
how did Groq design its chip?



14nm process node, size 725mm² (25mm x 29mm)
INT8 720Tops, FP16 188 TFlops
220MB SRAM, no DRAM and HBM controllers
480 GBps Chip-to-Chip Links
PCI-E Gen4.0 x16

DSA:
Focus on AI
Minimalist design philosophy
Software defines hardware

The challenge in the software. The chip was released on 2020, and 4 years were spent on the software.

SIMD Unit

320-element vector

SXM
Data Reshapes

Matrix-Vector
Matrix-Matrix

MEM on-chip SRAM

Vector-Vector

MEM on-chip SRAM

SXM
Data Reshapes

Matrix-Vector
Matrix-Matrix

superlane

Synchronized instruction dispatch across all SIMD units for lockstep execution

Instruction
Flow

Data Flow

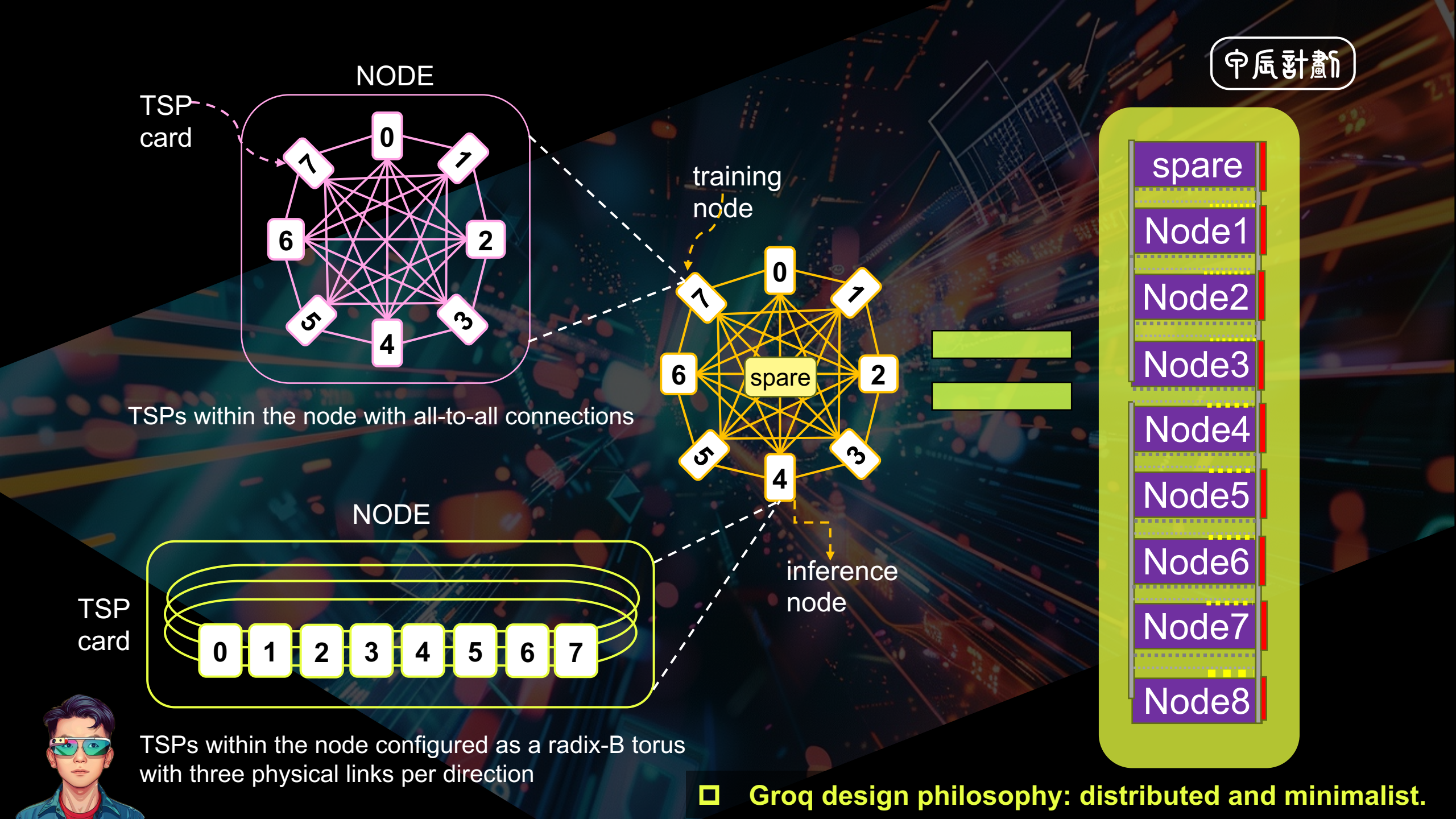
Data Flow

Numeric Mode	Max Size	Supported Density
int8	[N,320]x[320,320]	Two per MXM
float16	[N,320]x[160,320]	One per MXM

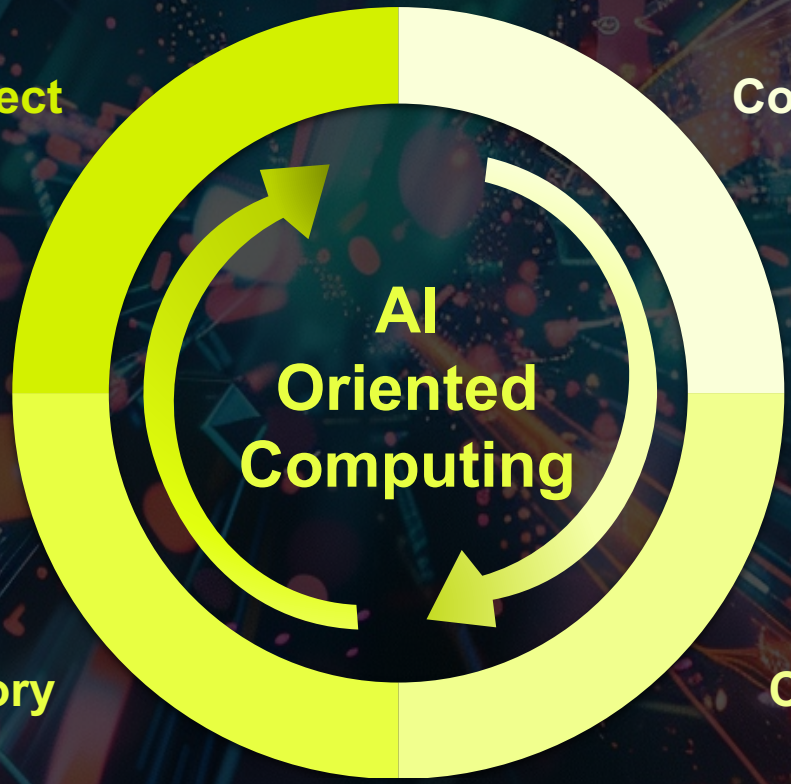
Instruction Dispatch

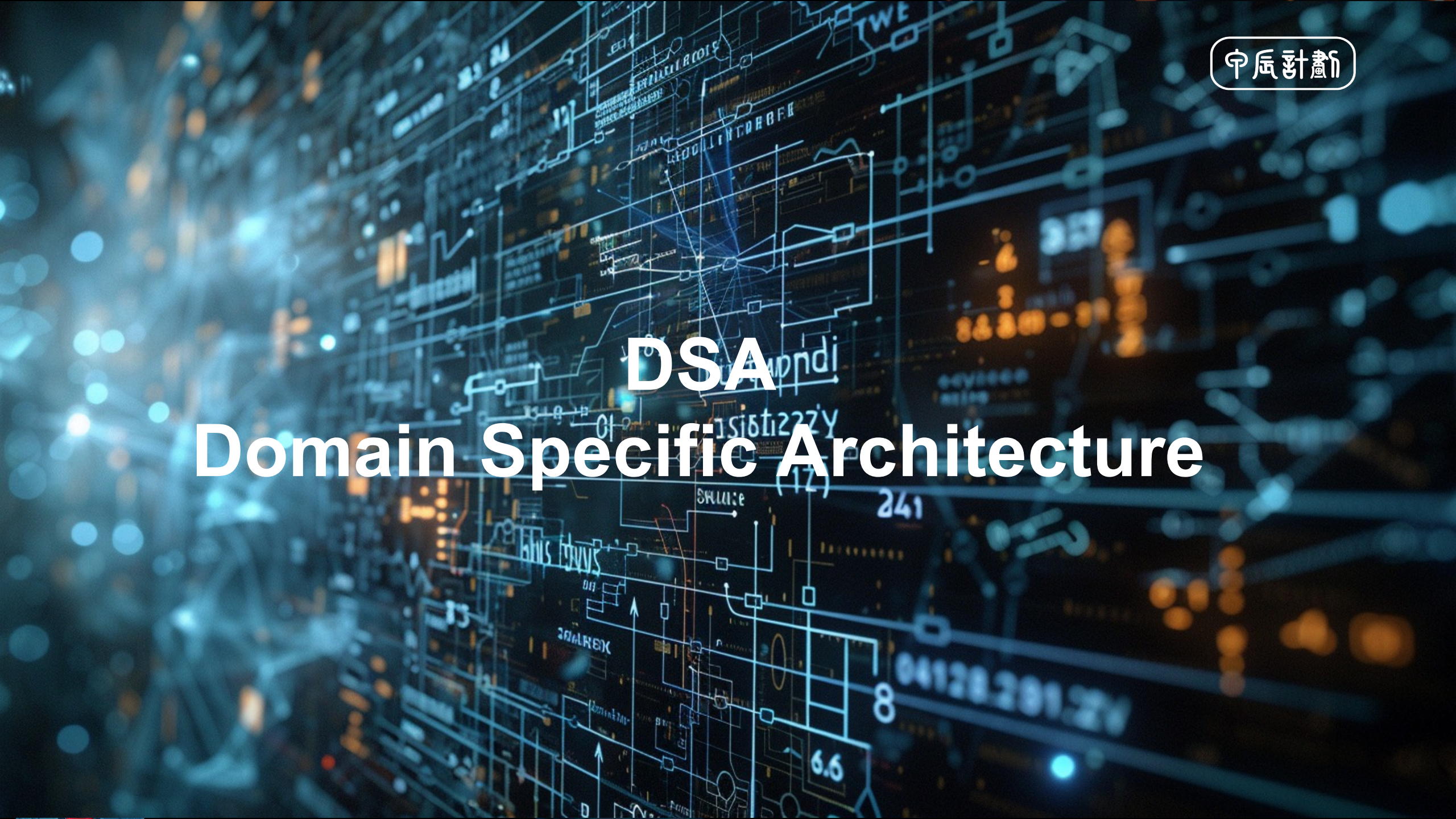
Instruction Dispatch

144 Instructions Dispatch Paths


NODE
TSP card
0
1
2
3
4
5
6
7
TSPs within the node with all-to-all connections
training node
spare
inference node
NODE
TSP card
0 1 2 3 4 5 6 7
TSPs within the node configured as a radix-B torus with three physical links per direction
spare
Node1
Node2
Node3
Node4
Node5
Node6
Node7
Node8
Groq design philosophy: distributed and minimalist.

Important Factors for AI Oriented Computing



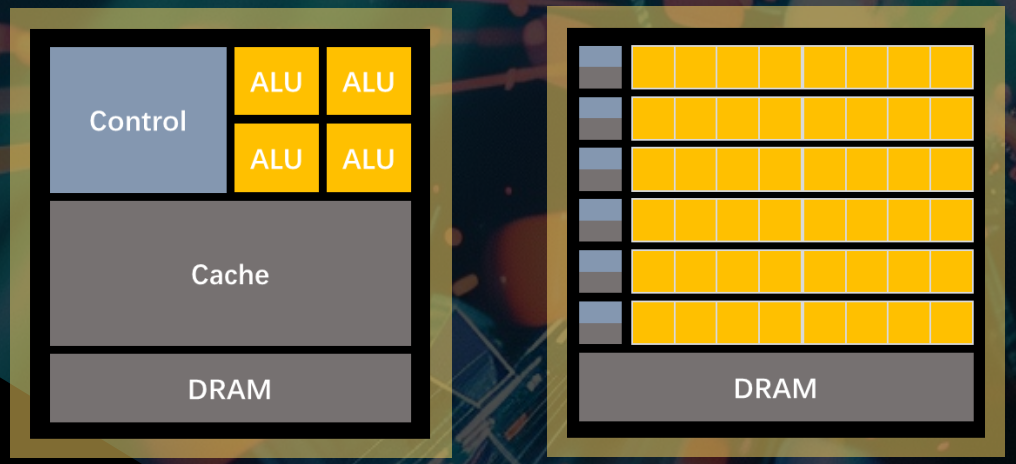


DSA
Domain Specific Architecture

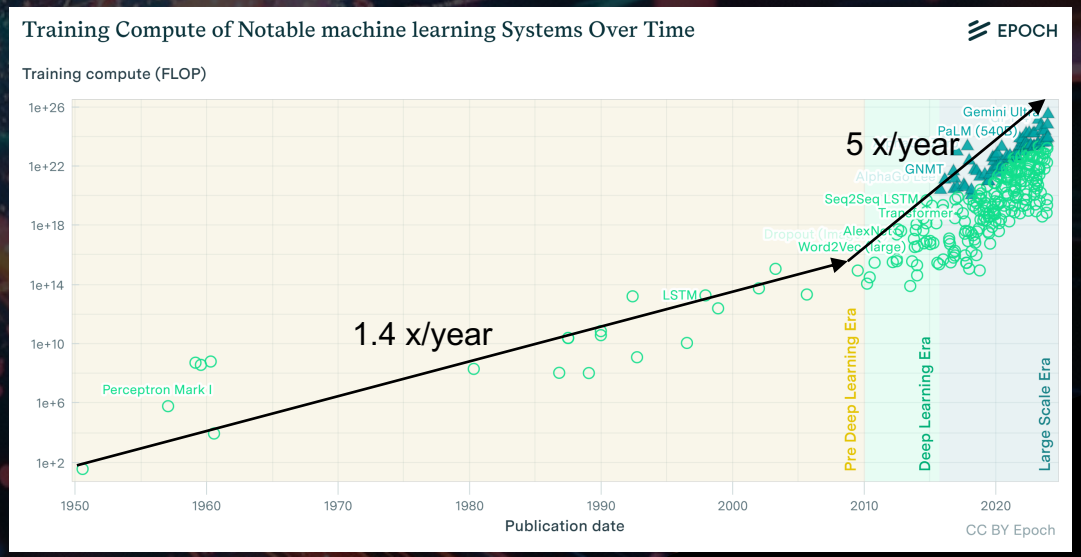
DSA Brings Cost-Efficient And Energy-Efficient AI Computing

- In general-purpose processors, the compute unit occupies much smaller area than control logic and on-chip memory.

CPU vs GPU



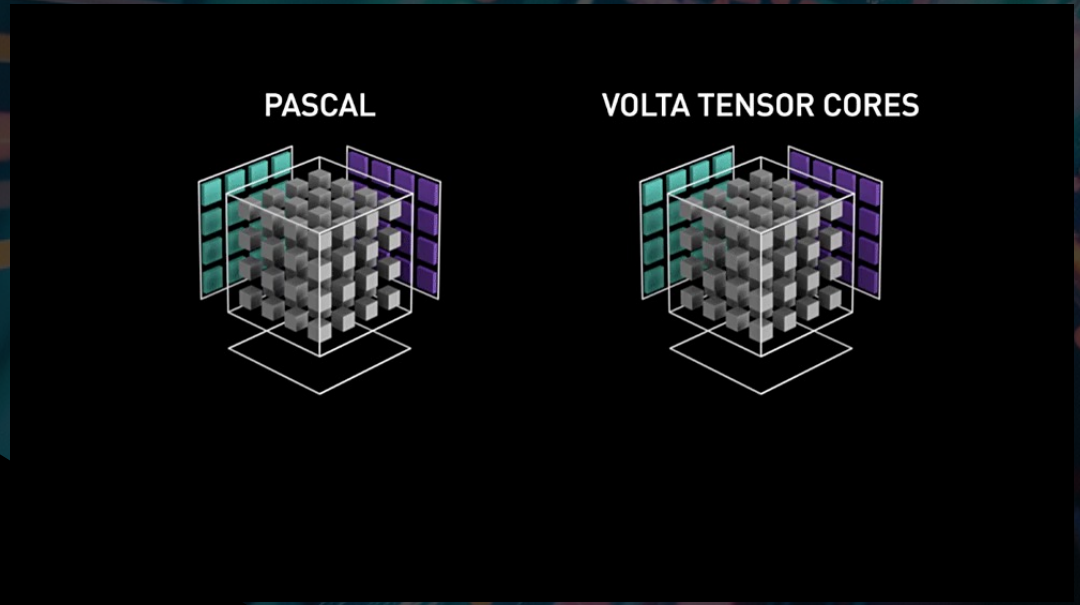
- General-purpose processors have long been unable to meet the needs of exponential growth in the amount of computing.



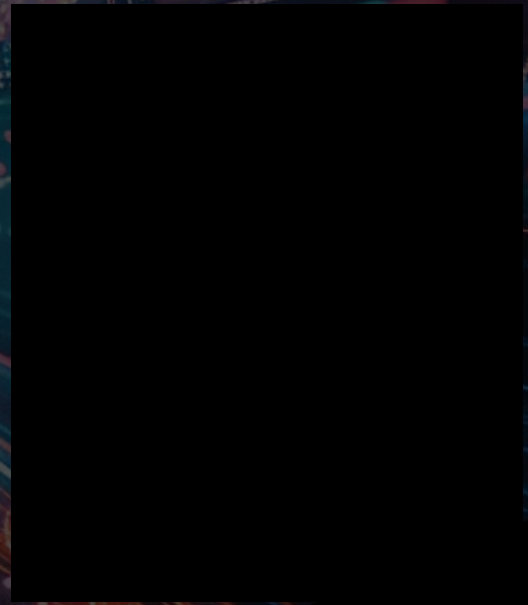
- To achieve cost-efficient and energy-efficient AI Domain Specific Accelerator
  - More compute unit
  - Less control unit
  - More deterministic data flow

Two Typical Compute Component In AI Processor

- Tensor Core

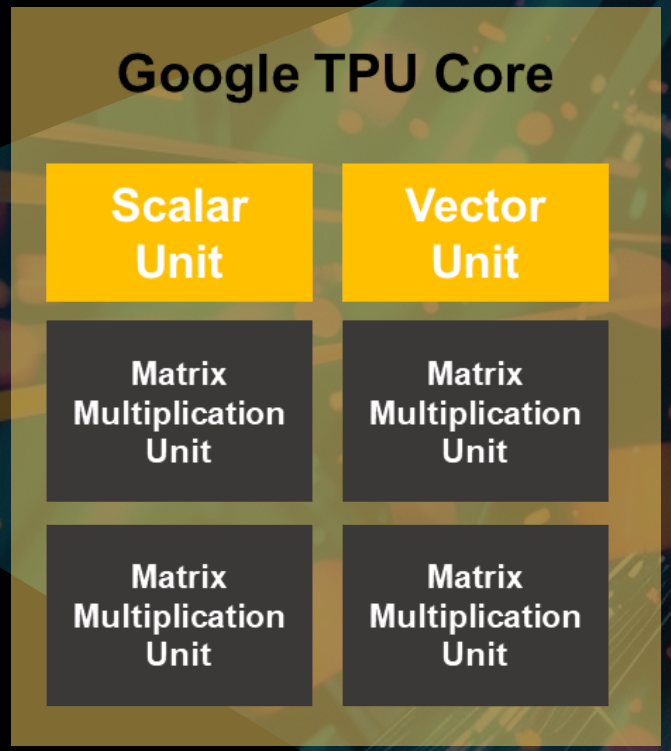


- Systolic Array



- ~1024 MACs is performance/cost boundary of these two architecture.
  - Tensor Core can’t be big, in Hopper one tensor core do 1024 int8 MACs per cycle.
  - Google TPU typical uses 128x128 systolic array, which offers 16,384 MACs per cycle.

Now There are Many MACs, but how to Control?



- AI algorithms require not only matrix multiplication, but also vector compute, scalar compute, data permuting, special functions, and branches.
- How to meet these requirements at low cost and quickly?
  - Implement these features in DSA? High design cost, not much return and less flexibility lead to poor ROI and backwards compatibility.
  - Using a ready Processor? For mainstream X86 and ARM processors, you don't have permission to modify it.
- Software friendliness is very important !

- **RISC-V :** a free, universal, open and consensus ISA.

Control： RISC-V

RISC-V Is Free, Universal, Open And Consensus

Free
for everyone to use

Universal
for various applications

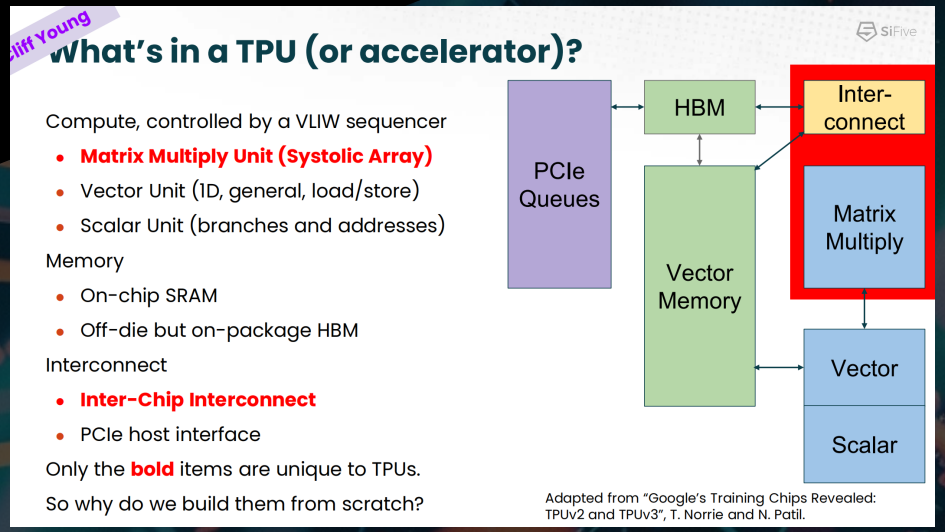
Open
to do customized modifications

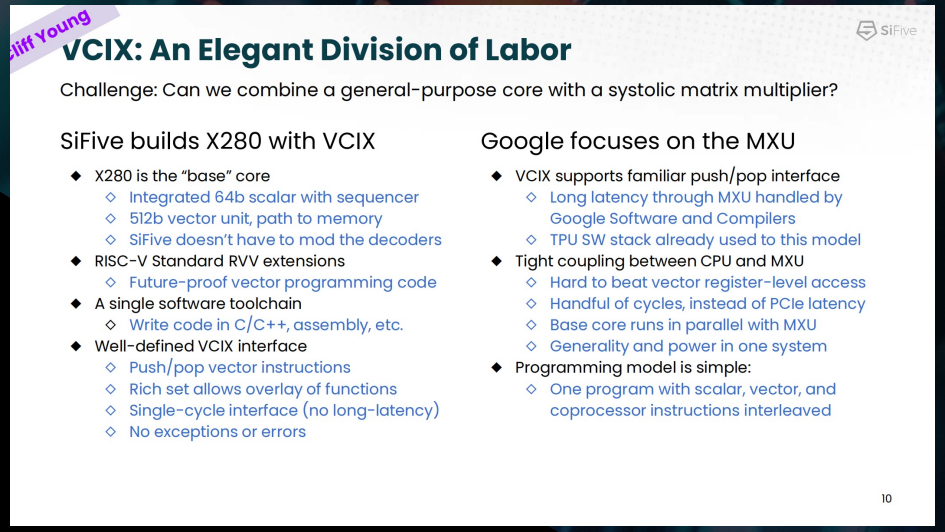
Consensus
by international foundation

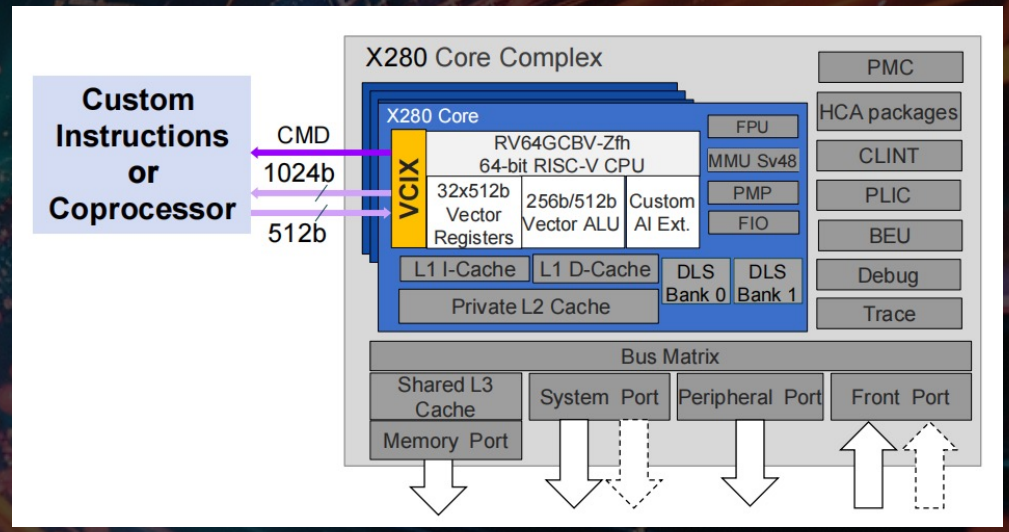
MCU	XuanTie	Nuclei	
CPU	XiangShan	SiFive	Ventana
GPGPU	ChengYing	Vortex	
TPU	Google	Dojo	Tenstorrent

- **RISC-V** is what we need!

Google Adopts RISC-V for Next-generation TPU





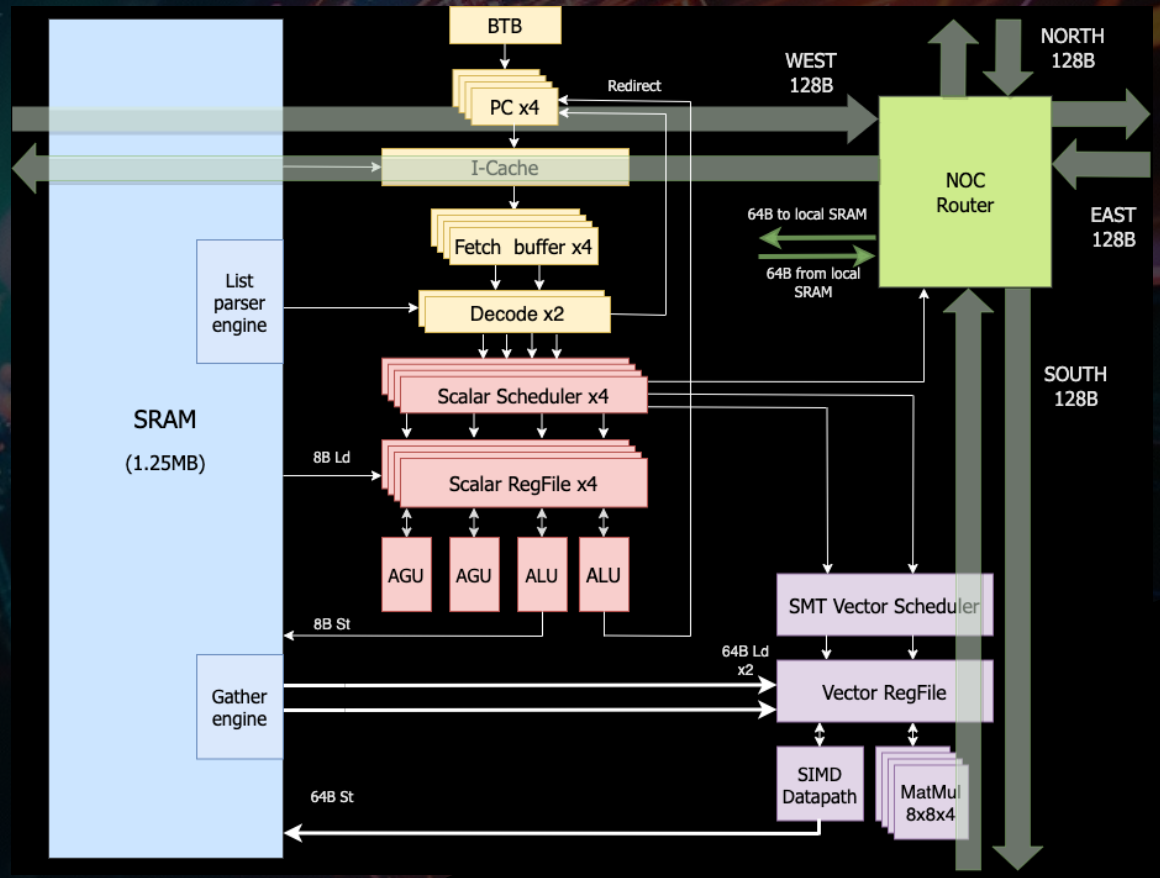


SiFive Intelligence RISC-V Core

- Wider vector ALU
- Faster load/store from PL2$
- VCIX interface enable easily custom DSA design integration
- Supported by AI/ML SDK with open source ML Compiler and Runtime

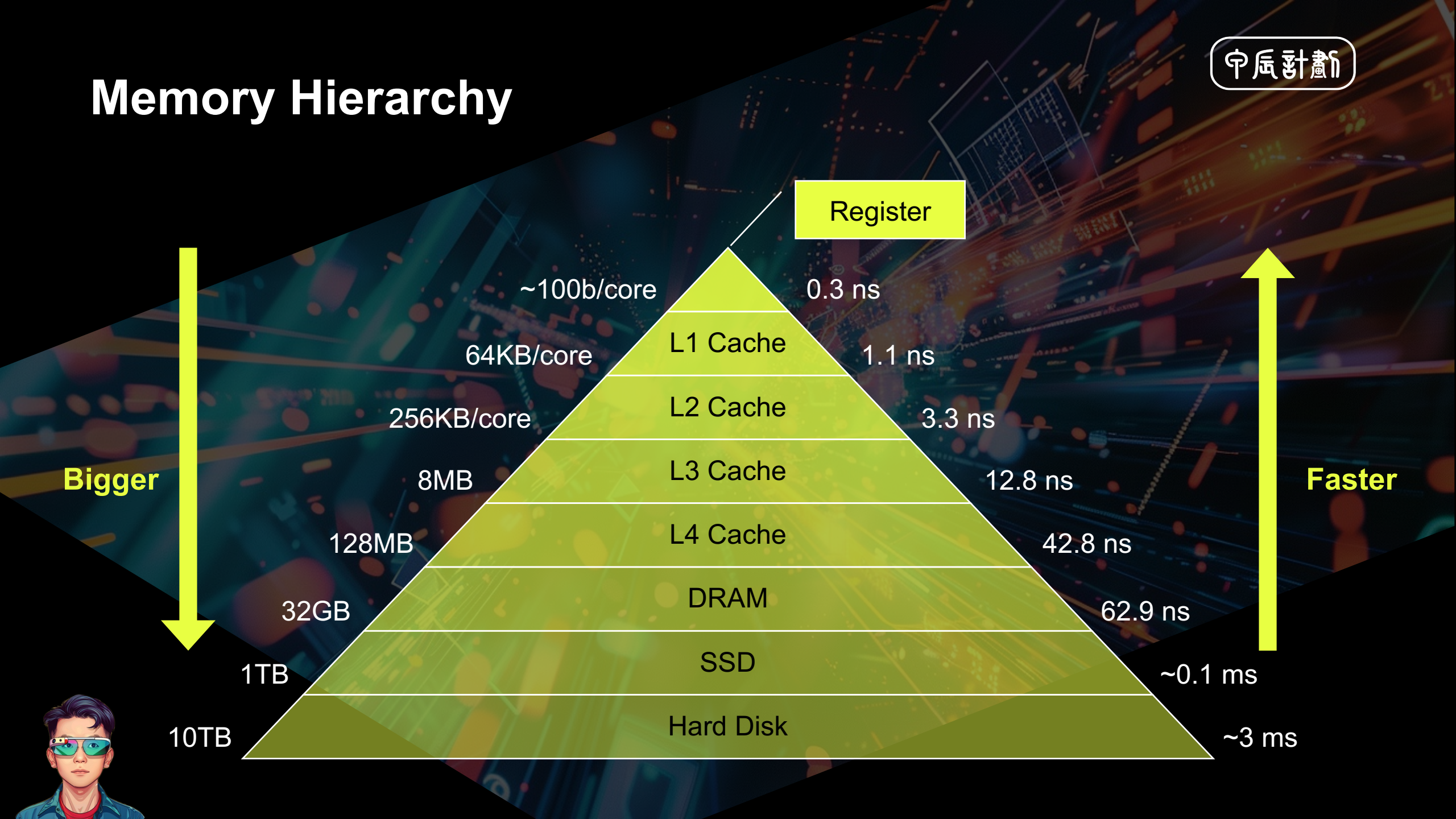
Tesla Dojo Node Designed on RISC-V

- **Superscalar**
  - high-throughput
- **Multi-threaded**
  - Scalar Register file replicated per thread
  - Typical application uses 1 or 2 compute threads and 1-2 communication threads
- **RISC-V + Custom AI ISA**
  - 64B wide SIMD unit
  - 8x8x4 matrix multiplication units



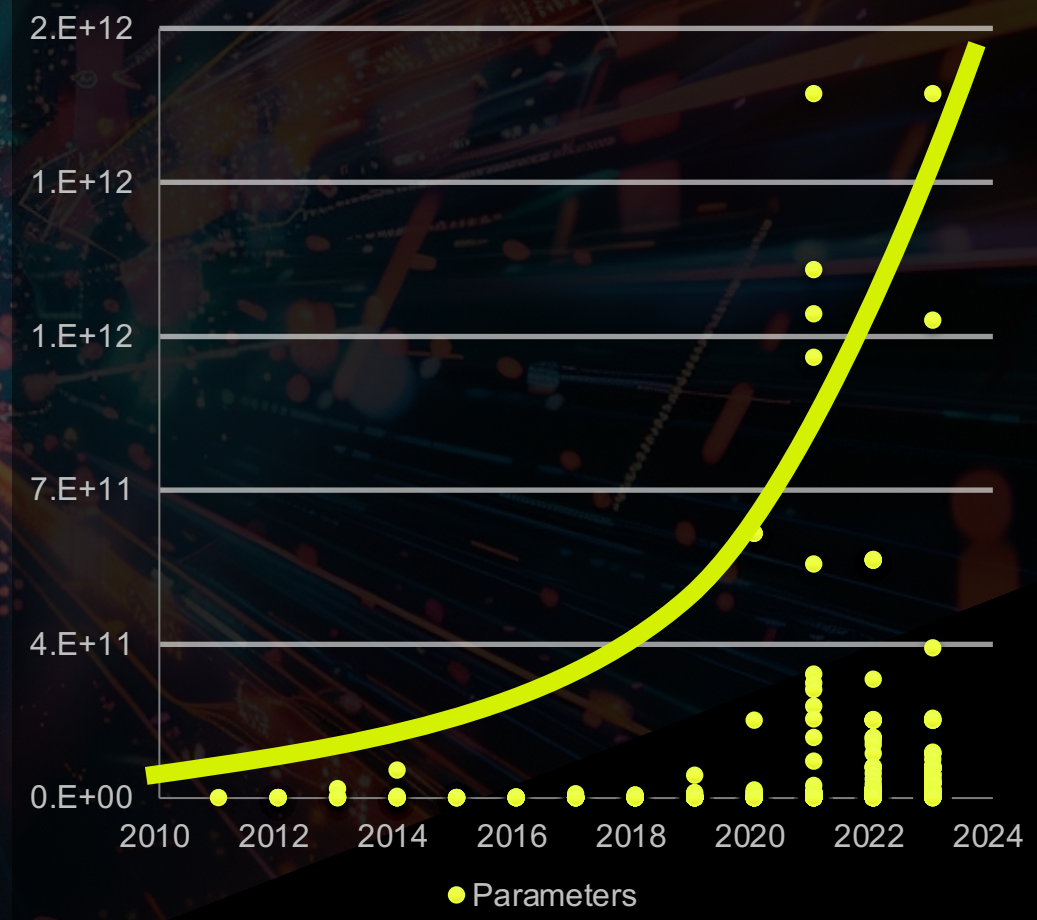

Memory Wall

Memory Hierarchy



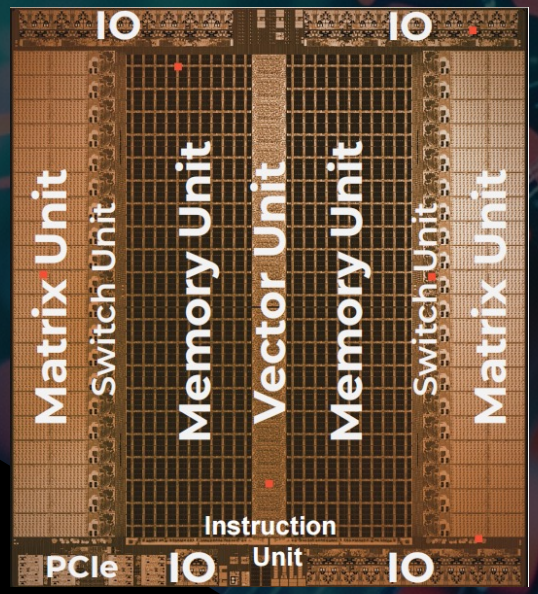
Near-Memory Computing Break the Memory Wall

- Entering the era of large models, with the rapid growth of model size, the memory wall problem becomes more and more serious.
- The academic community has proposed various solutions for In-Memory Computing and Near-Memory Computing.
- Near-Memory Computing is relatively mature and is widely used in the design of various SOTA AI computing chips.
  - SRAM
  - HBM
  - Any new technology?

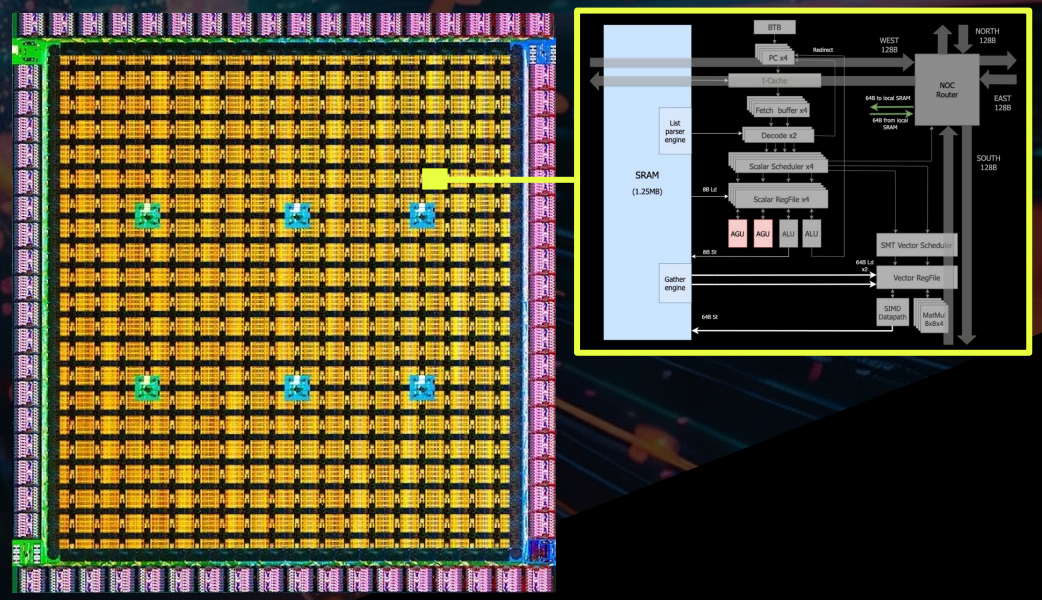


SRAM-Based Near-Memory Computing

- Groq 725 mm2 @14nm
  - **230 MB** memory capacity
  - **80 TB/s** memory bandwidth



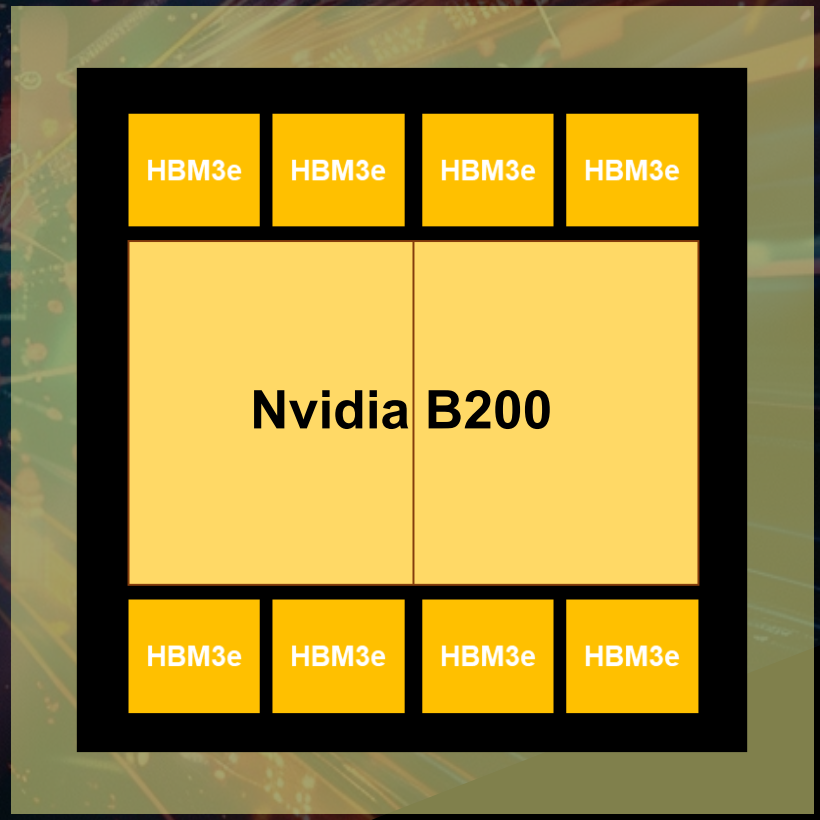
- Dojo 645 mm2 @7nm
  - **440 MB** memory capacity
  - **141.6 TB/s** read and **95.6 TB/s** write



- SRAM offers much bigger bandwidth, but it is too large and expensive

HBM-Based Near-Memory Computing

- **Nvidia Blackwell B200**
  - **8 TB/s** memory bandwidth
  - **192 GB** memory capacity



- Good balance between bandwidth and capacity

Nvidia GPU HBM Usage

	A100 40GB PCIe	A100 80GB PCIe	A100 40GB SXM	A100 80GB SXM	H100 PCIe	H100 SXM	H100 NVL	H100S SXM(1)
Capacity (GB)	40	80	40	80	80	80	192 (2)	120/144
HBM Version	2	2E	2	2E	2E	3	3	3
Stacks	5	5	5	5	5	5	12	5/6
Layers	4/8+1	8+1	4/8+1	8+1	8+1	8+1	8+1	12+1
Speed (GTs)	2.43	3.02	2.43	3.19	3.19	5.23	5.08	5.60
Bandwidth (GB/s)	1,555	1,935	1,555	2,039	2,039	3,350	7,800	3584/4301

Other Accelerator HBM Usage

	Google TPUv4i	Google TPUv4	Google TPUv5i(1)	Google TPUv5(1)	AMD MI250X	AMD MI300A	AMD MI300X	AWS Trainium/ Inferentia 2
Capacity (GB)	8	32	16	64	128	128	192	32
HBM Version	2	2	2E	3	2E	3	3	2E
Stacks	2	4	2	4/6	8	8	8	2
Layers	4+1	8+1	4+1	8+1	8+1	8+1	12+1	4+1
Speed (GTs)	2.29	2.34	3.20	5.2	3.20	5.20	5.60	3.20
Bandwidth (GB/s)	585	1,200	819	2662/3993	3,277	5,325	5,734	819

HBM Performance Evolution

- Now used across multiple applications
  - AI/ML, HPC, Networking

Bandwidth

(2018) HBM2E

Bandwidth **469GB/s**
I/O Speed **3.6Gbps**

(2018) HBM2E

Bandwidth **469GB/s**
I/O Speed **3.6Gbps**

(2022) HBM3

Bandwidth **819GB/s**
I/O Speed **6.4Gbps**

HBM3

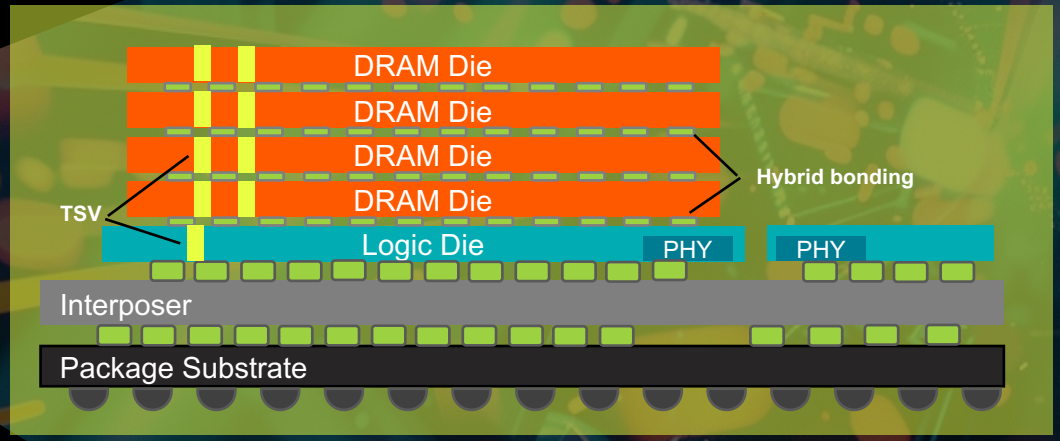
Bandwidth **1075GB/s**
I/O Speed **8.4Gbps**

Rambus HBM3 Memory Subsystem

I/O Speed

Sources: SK hynix and Rambus

WoW - A New Solution For Near-Memory Computing



4 layers DRAM die (~700mm2), totally support:

- Bandwidth: > 20 TB/s
- Capacity: > 50 GB
- Power: ~50W
- Cost: < 15$ / GB

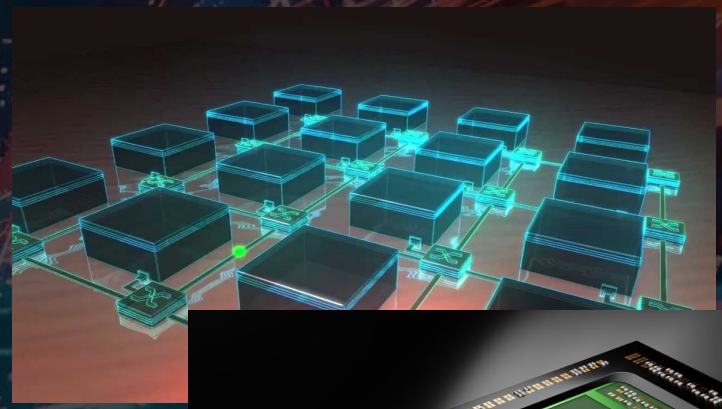
- Potential alternative to HBM

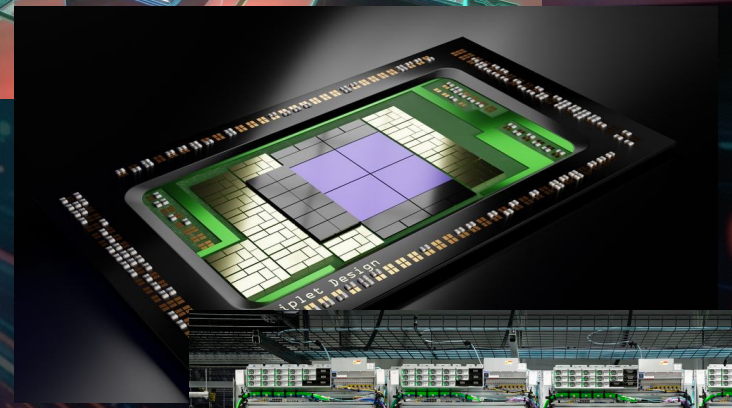
HBM vs WoW

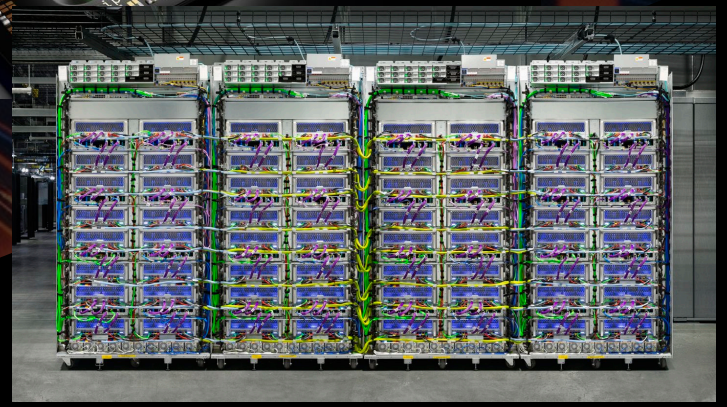
	HBM	WoW
Density	192 GB	50GB
Bandwidth	8 TB/s	20 TB/s
Scalability	Micro bumps significantly limit the application of additional I/O, such as the 2048 I/O of HBM3	Hybrid bonding supports virtually unlimited I/O interconnects.
Energy Efficiency	~5pJ/bit (Micro Bump)	~0.5pJ/bit (hybrid bonding): 5.28W/TB
Heat	The presence of underfill significantly affects the intra-layer heat dissipation of HBM	Without underfill, and with high-density inter-layer metals plus thinner wafers, there is reduced thermal resistance, resulting in extremely high heat dissipation efficiency.
Reliability	Inter-layer MicroBump issues, unable to pass automotive certification.	Supports automotive
Cost	$13 / GB	$ 15 / GB
Supply Chain	Heavily dependent on TSMC's CoWoS capacity and Cadence's 3D integration tools	Domestically feasible


中辰計劃
Connect
NOC, D2D, C2C

Expand The Boundaries Of Collaborative Computing Systems Using NOC, D2D, C2C







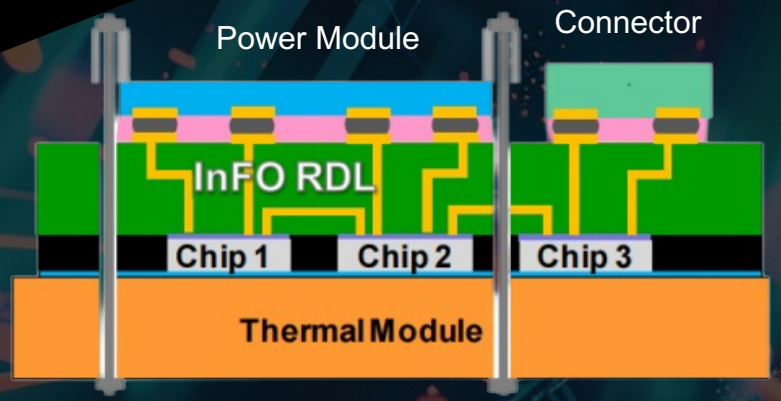
- The explosion in model size and compute amount far exceeding the speed of Moore's Law makes it necessary to implement an algorithm on multi-core or even multi-processor.
- Although hardware cache coherence can greatly simplify software in the field of general-purpose processors, it is not feasible in the field of AI accelerators due to its weak performance.

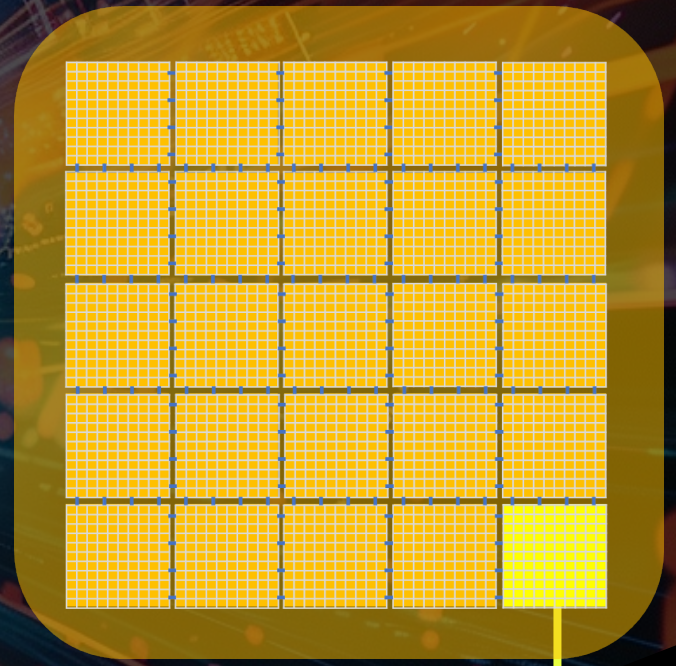
- High-performance, low-cost interconnect technology is needed.

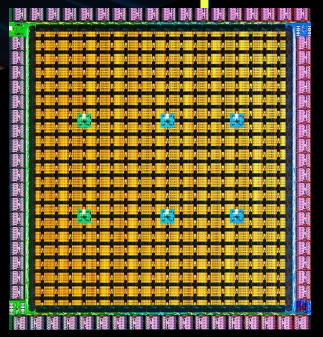
NOC Interconnection in Tesla Dojo

- **2 Tbps** bidirectional bandwidth between adjacent node and single cycle per hop
- **4.5 TB/s** between adjacent row and **5 TB/s** between adjacent column
- Target hardware simplicity!
  - No cache coherence
  - Flat addressing scheme
  - Programable routing table
  - Not guarantee end-to-end traffic ordering

Amazing D2D Interconnection in Tesla Dojo







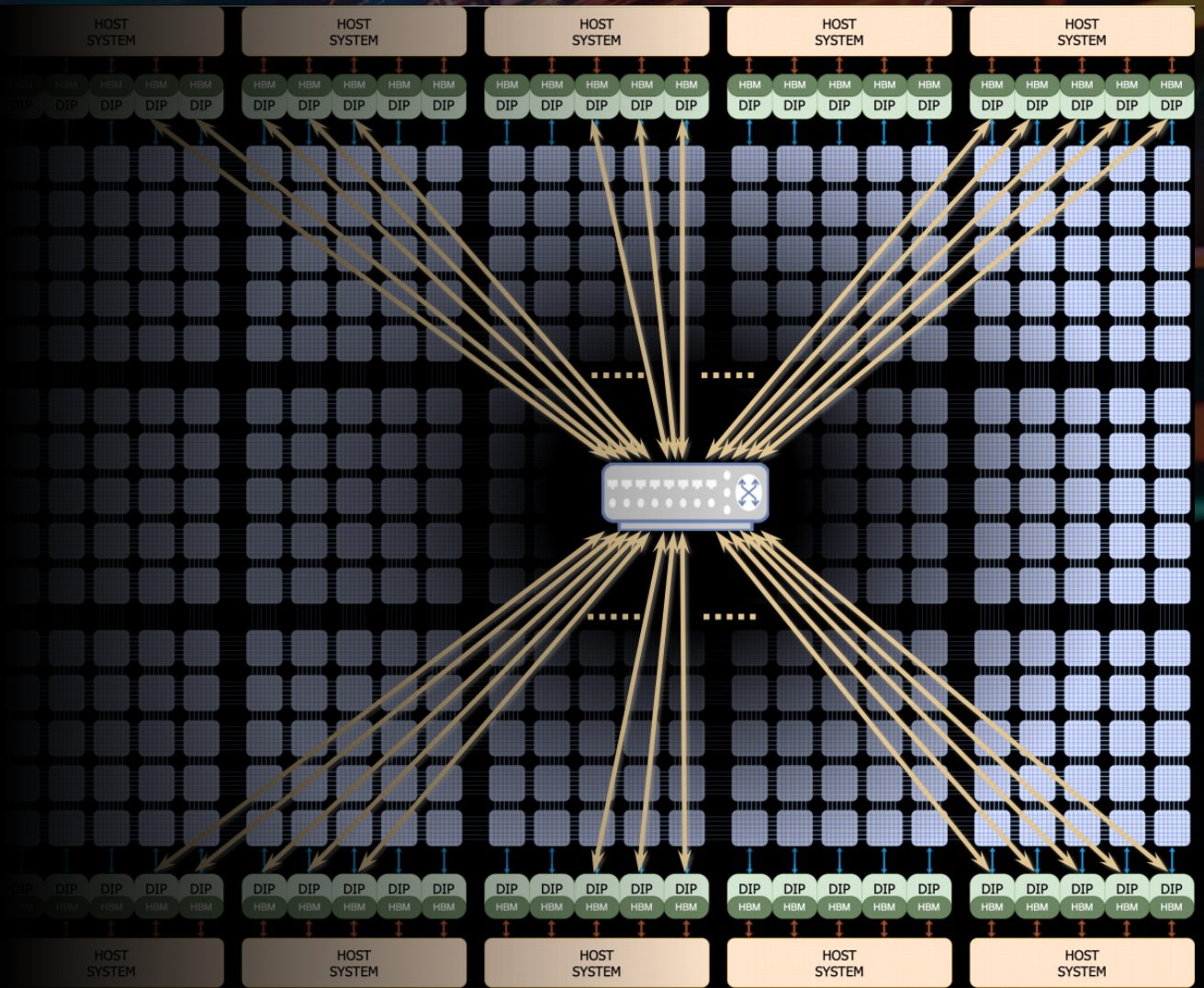
- **InFO_SoW (System-on-Wafer)**
- **2 TB/s** between adjacent D1 chip
- 10 TB/s between adjacent rows and column inside one training tile

- Amazing design! One wafer is one chip!

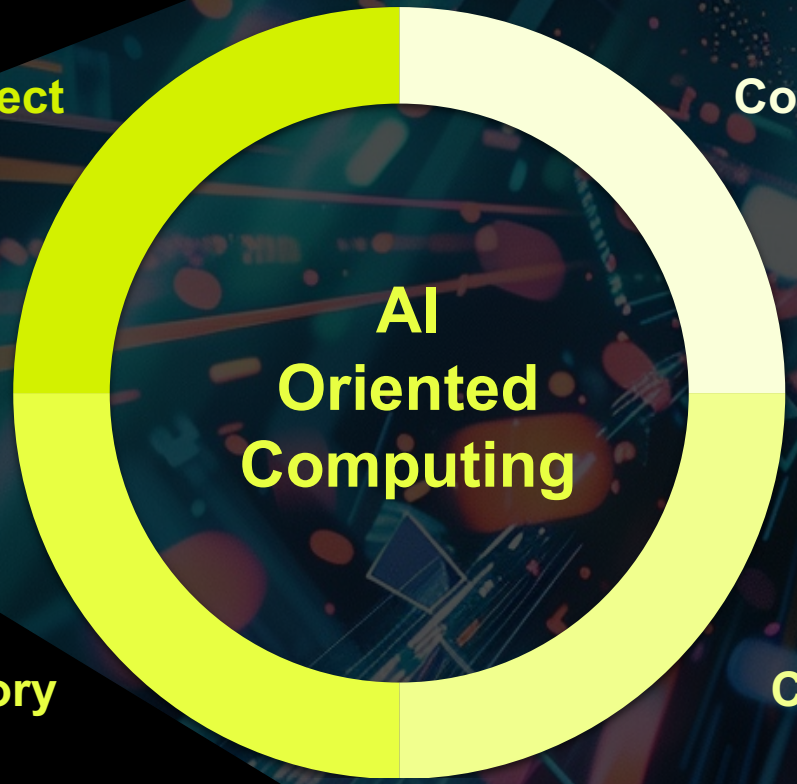
C2C Interconnection in Tesla Dojo

- **9 TB/s** between adjacent tile
- Edge tile connected to 5 DIPs, which offer totally:
  - **5 * 0.9 TB/s** connection bandwidth
  - **5 * 32 GB** HBM memory
  - **5 * 400 Gbps** Ethernet
  - **5 * 32 GB/s** PCIe
- Global communication – Z-plane links between DIPs through an ethernet switch

- **One more ExaFLOPS club member**



Important Factors for AI Oriented Computing



- DSA
- RISC-V
- Near-Memory Computing
- NOC, D2D, C2C

[Oasis]

PC Processor Base on RISC-V

LLM PERFORMANCE

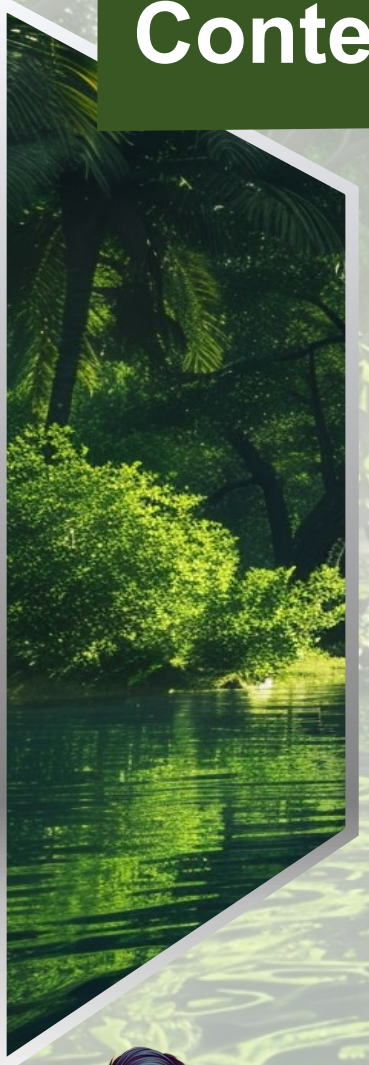
16

TOKENS PER SECOND

Running LLaMA-2 13B INT4

Our Choice in Designing OASIS Gen1 (SG2380), which Focus on One-Chip-Running LLM

Contents

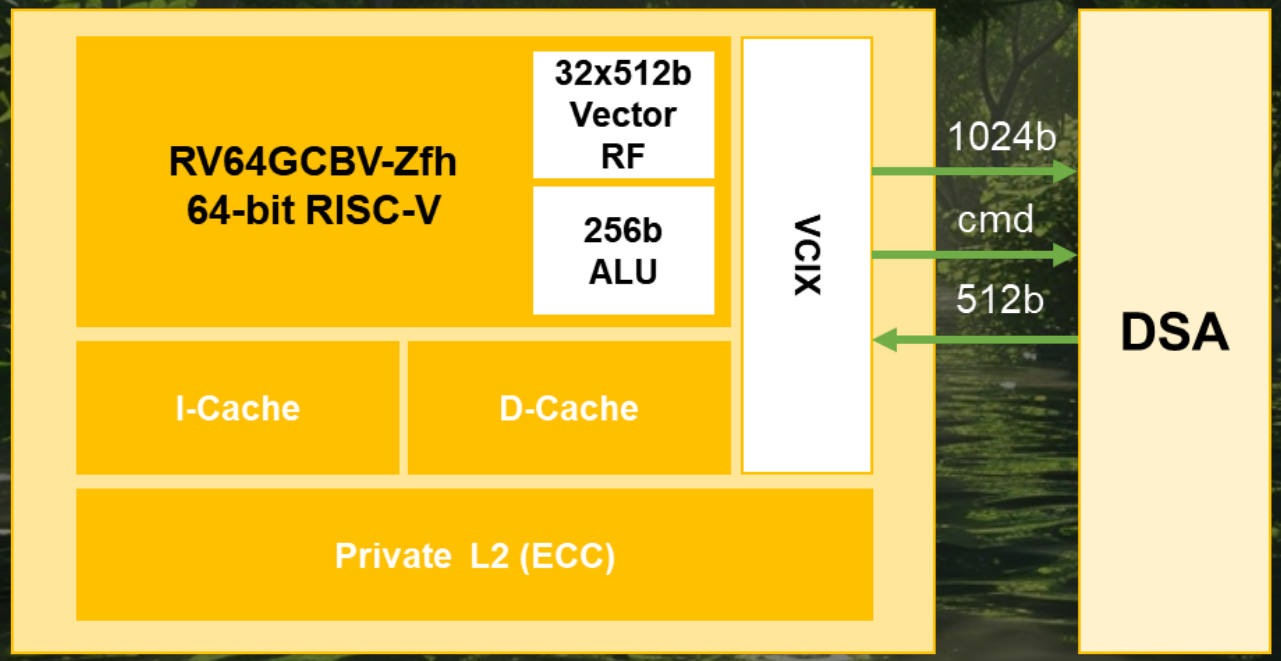


SOPHON TPU

High efficiency AI DSA

- Cube for matrix multiplication
- Near-memory computing
- 1 TIU，many Lanes
- Double layer ring bus to support control and data broadcast and transfer
- Parametric design specifications

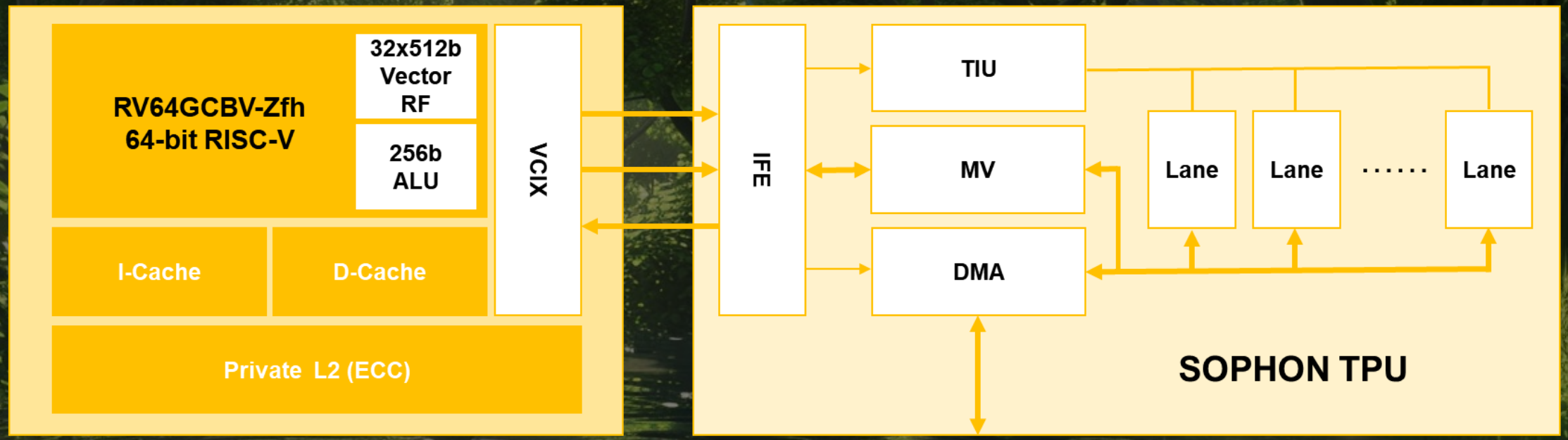
SiFive Intelligence X280



- Lightweight applications processor for Linux and baremetal
- 64-bit RISC-V ISA 8-stage dual-issue in-order pipeline

- 512-bit vector register length processor
- VCIX interface for tight-coupled coprocessor or DSA

X280 + TPU



IFE

- Instruction queue
- Instruction decode
- Dependency check

MV

- Data movement between vector register file and TPU

DMA

- Data movement between TPU and DRAM

Registers

Tensor register (TR): ta0 ~ ta23

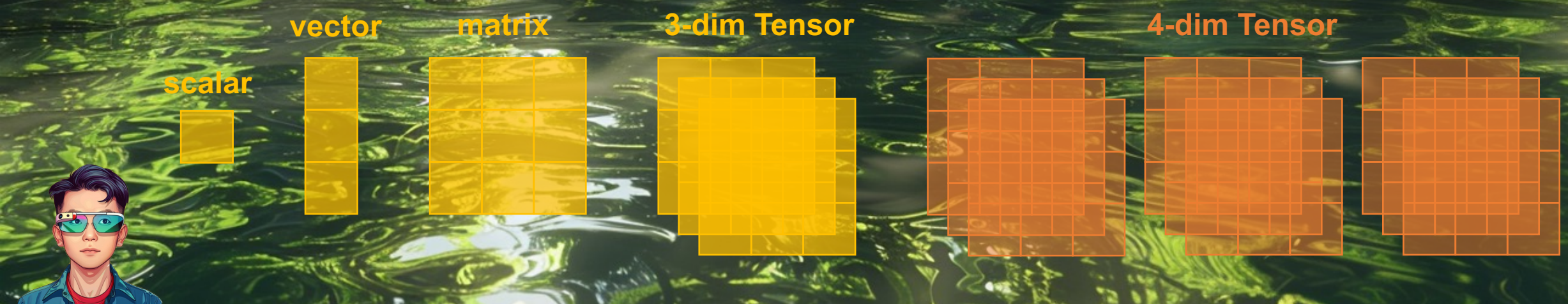
- TR represents the Tensor stored on local memory of TPU.
- Each TR consists of a Bank or half a bank on SRAM.
- 16x 128KB registers and 8x 256KB registers

Global Tensor Register (GR): ga0 ~ ga7

- GR means stored as a Tensor on DRAM.

Constant register (CR): ca0 ~ ca7

- CR is used for storage Constant.



Tensor Instruction

- conv
- conva
- fconv
- fconva
- dwconv
- fdwconv

- pool.avg
- pool.favg
- pool.max
- pool.fmax
- pool.min
- pool.fmin

- add
- sub
- mul
- mac
- abs
- max
- min
- selgt
- seleq
- sellt
- lsh
- ash
- rsh

- fadd
- fsub
- fmul
- fmac
- fabs
- fmax
- fmin
- fselgt
- fseleq
- fsellt
- fsubabs
- fsqradd
- fsqrsub

- cvt.i2f
- cvt.i2i
- cvt.f2i
- cvt.f2f
- and
- xor
- or
- not

- gatherpc
- scatterpc
- gather2d
- scatter2d
- gather
- scatter
- hgather
- hscatter
- masksel

- sfu.norm
- sfu.rsqrt
- sfu.taylor

- mm
- mma
- fmm
- fmma
- mm2
- mm2a
- fmm2
- fmm2a

- roipool.avg
- roipool.favg
- roipool.max
- roipool.fmax
- roipool.min
- roipool.fmin

- mv.collect
- mv.dist
- mv.v.m
- mv.m.v

- msgsend
- msgwait
- dma.msgsend
- dma.msgwait
- nop
- sync.i

- cfg.quant
- cfg.pad
- cfg.insert
- cfg.iter
- cfg.vlc
- cfg.dmaidx
- cfgtr

- cfgcr
- cfggr
- cfgtr.shape
- cfgtr.stride
- cfgtr.shape
- cfgtr.hwstride
- cfgtr.ncstride

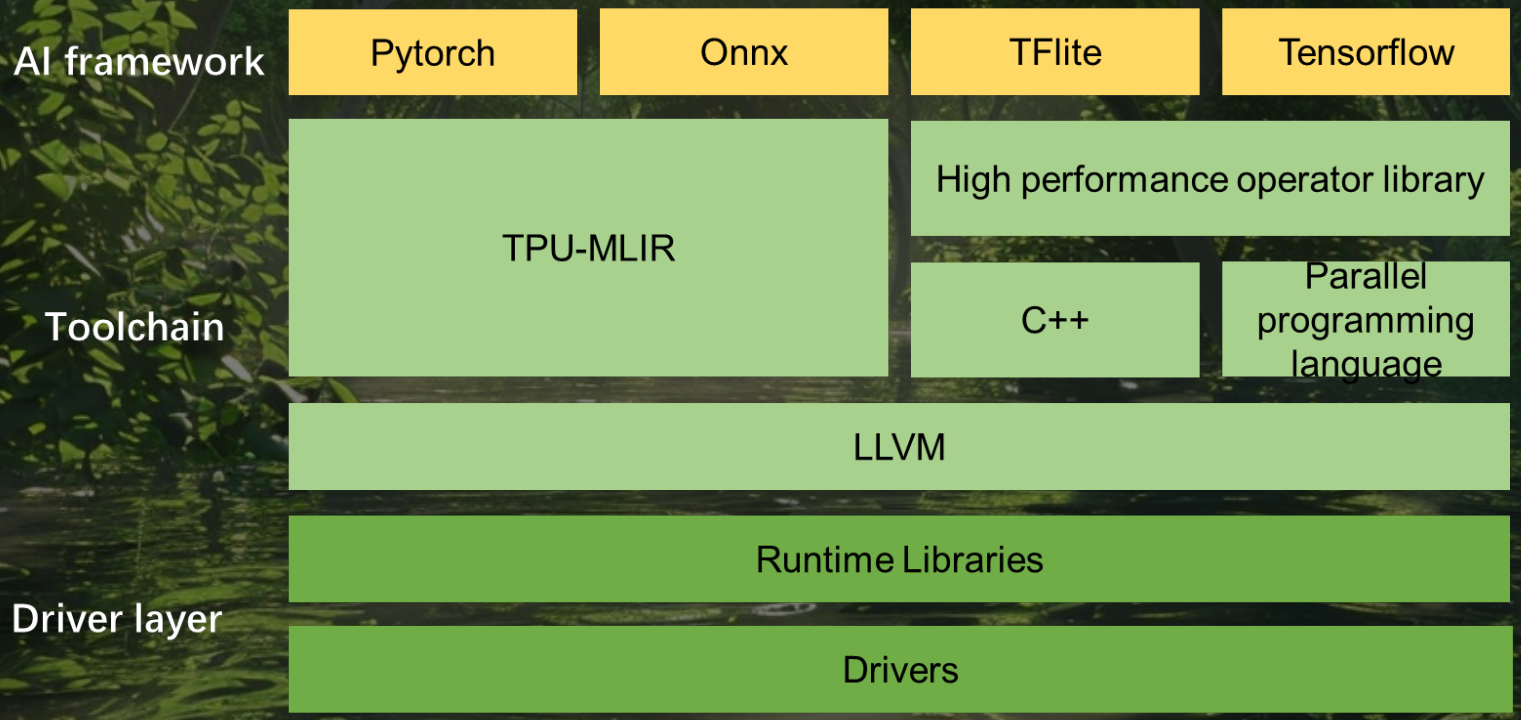
- dma.ld
- dma.ldt
- dma.st
- dma.stt
- dma.cp
- dma.mld
- dma.mldt

- dma.mst
- dma.mstt
- dma.fill
- dma.nct
- dma.cwt

- dma.reverse
- dma.masksel
- dma.fmasksel
- dma.nzidx
- dma.fnzidx
- dma.hgather
- dma.cmpr

- dma.racmpr
- dma.radecmpr
- dma.fcmpr
- dma.fdecmpr
- dma.fracmpr
- dma.fradecmpr

Software Stack



High-performance graph compiler based on MLIR

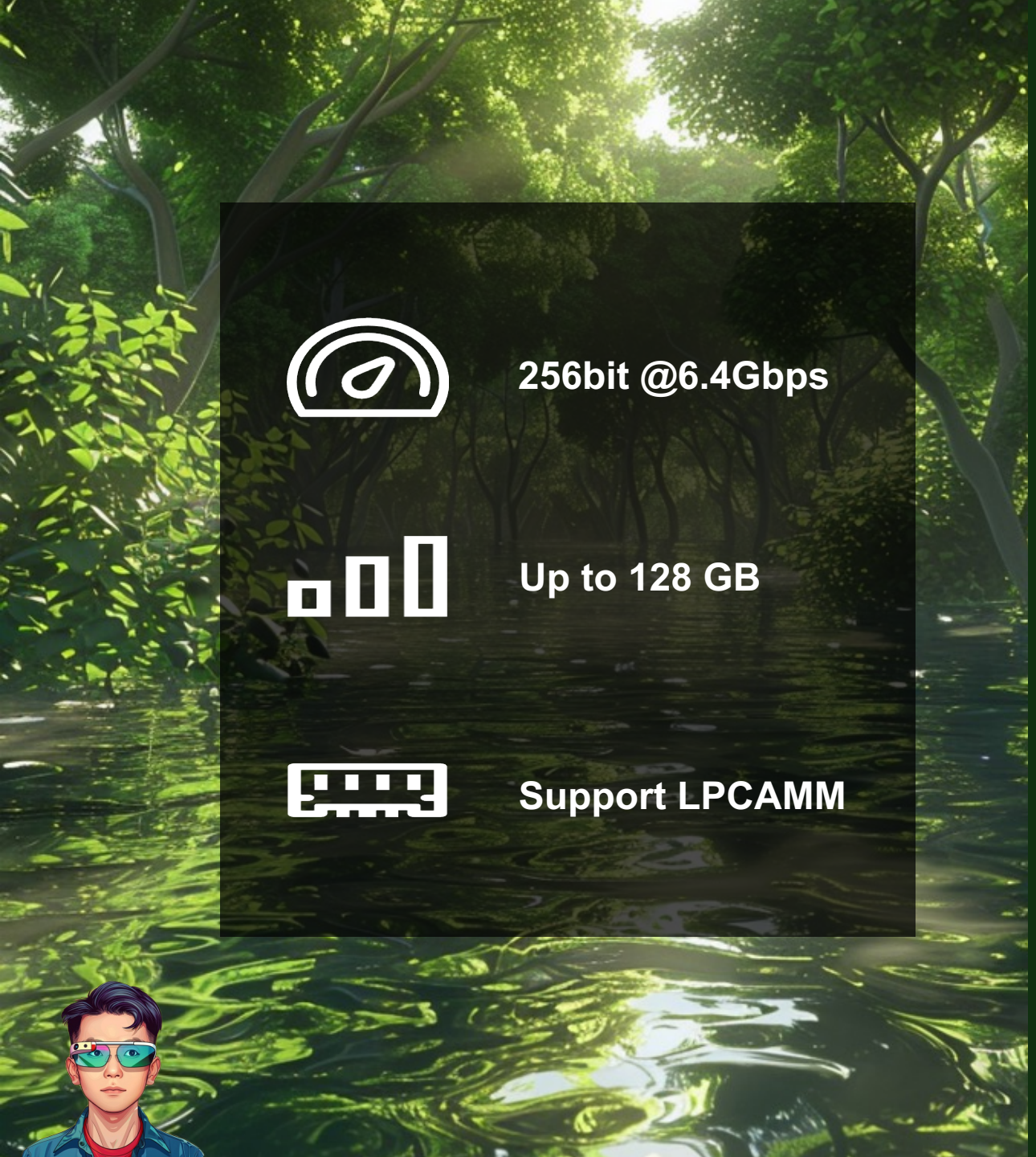
- Support mainstream AI framework
- Operator fusion
- Operator Tiling and Scheduling

High-performance operator library

- Supports 200+ common AI operators

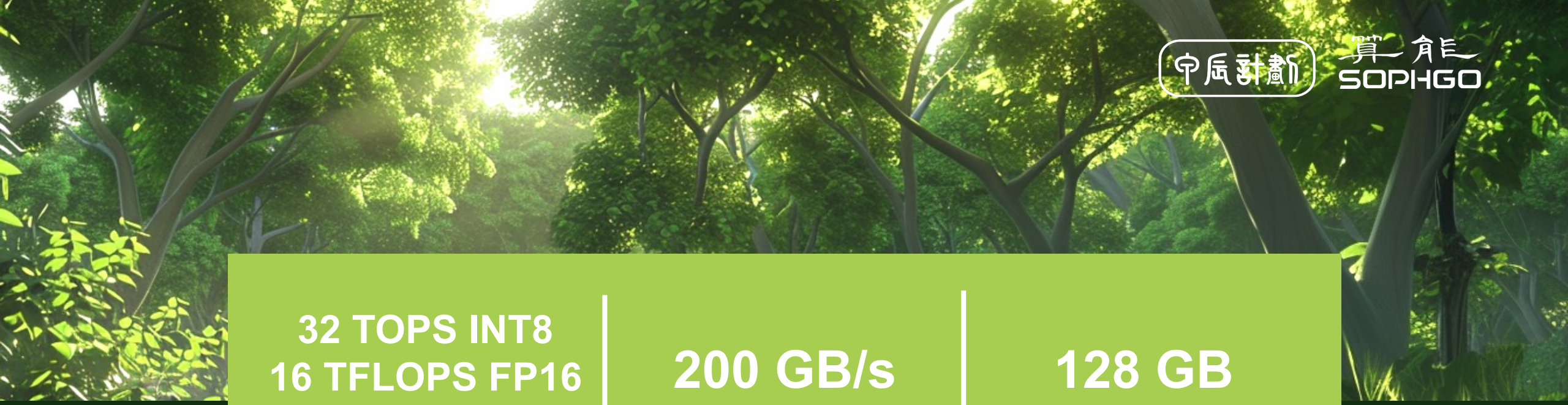
Parallel programming language

- Front-end based on c++ syntax
- Supports intrinsics of vector/tpu extension instructions
- Support loop optimization and software pipelining



Memory

- Bandwidth and capacity is important for running single batch LLM.
- LPDDR is the best choice: relatively big bandwidth, big capacity, low power and low cost.
- As for our selected process point, LPDDR5 is the best option.



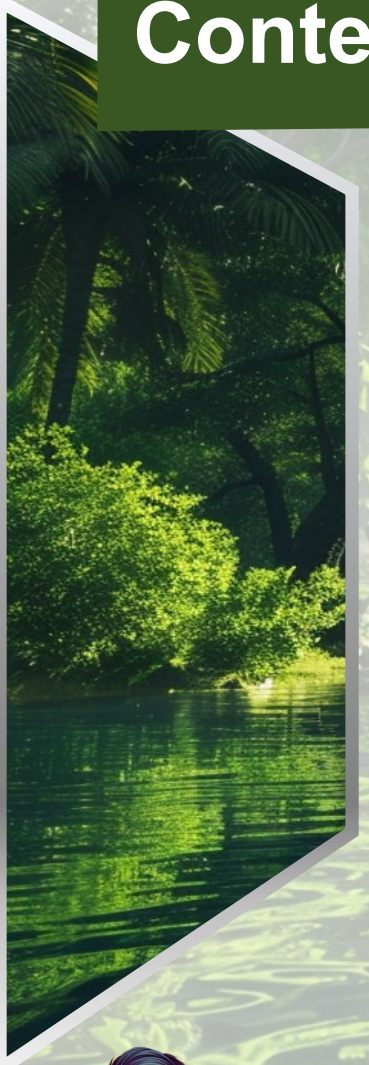
32 TOPS INT8 16 TFLOPS FP16	200 GB/s	128 GB
4-Core RISC-V + TPU	Memory bandwidth	DRAM

Ultimate ML Performance in Terminal Device

Full support for language and visual large models

Llama Qwen ChatGLM StableDiffusion Whisper CodeGeeX

Contents



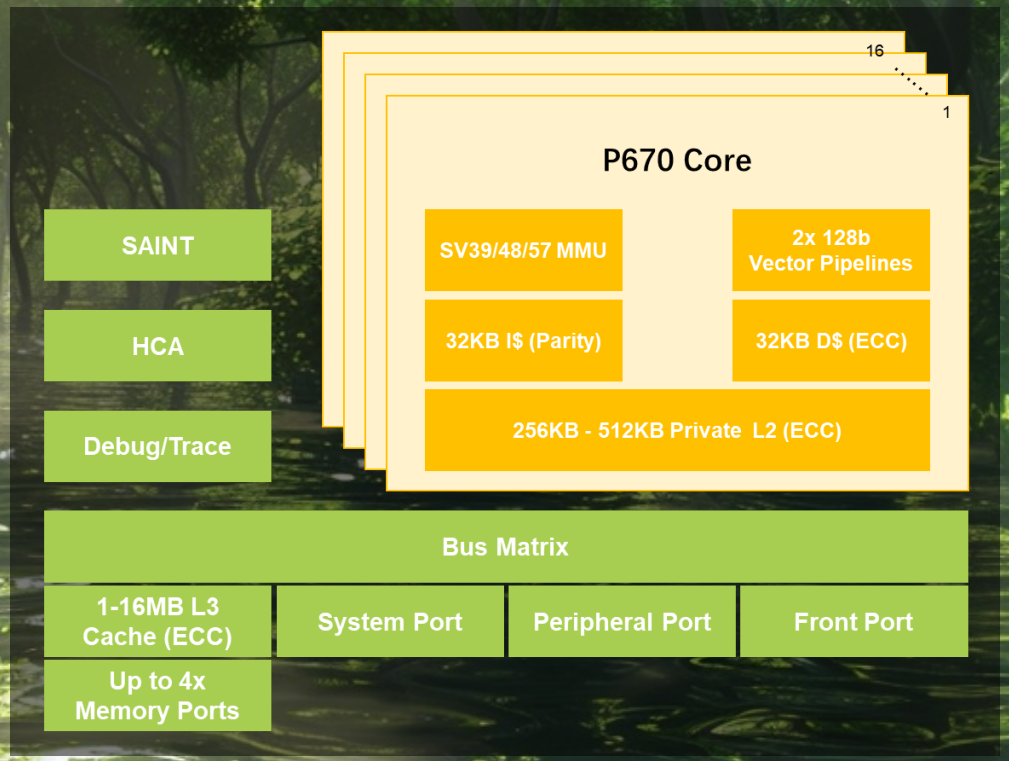
2 SG2380 Other Main Features

High Performance RISC-V Processor

12+ SPECInt2k6 / GHz
20% Higher Performance Than Cortex-A76

64-bit RISC-V ISA with Hypervisor Extension and MSI Interrupt Controller

Support for the RISC-V Vector Extension with 2x 128b Vector ALUs



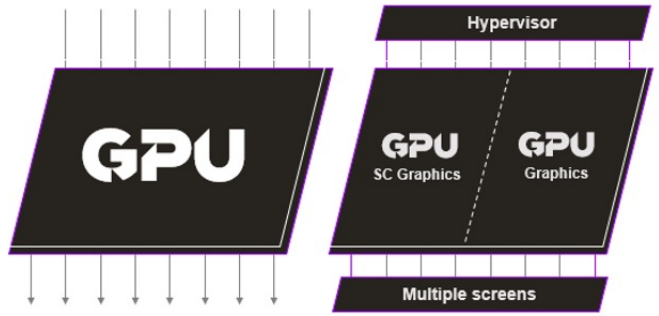
Coherent Multicore with up to 16 Cores in Core Complex and coherent interfaces

Advanced Security Capabilities and Crypto Accelerators

Sophisticated Trace and Debug capabilities

Graphics

IMG AXT
16 - 512



Multi-screen. Multi-OS. One GPU.
HyperLane flexibility and security

Scaling from main-stream mobile to desktop

Advanced compression technology

Hardware virtualization with HyperLane technology

Key use cases:

- Premium mobile gaming, rapid photo editing and super slick user interfaces.
- Gaming, photo editing, video editing and AI workloads on a laptop or desktop.

- 512 FP32 per cycle
- 16 pixel per cycle
- Vulkan 1.3
- OpenGL ES 3.3/2.0/1.1 + Extensions
- OpenCL 3.0

Multimedia & Interconnect

4K@60fps Decode
4K@30fps Encode

1x USB 3.2 Gen2 x1 Type-C,
with DP Alt Mode

1x 2 lane eDP 1.4b
1x HDMI 2.0
1x 4 lane DSI 2.0

1x USB 3.2 Gen2 x1 Type-C,
with DP Alt Mode

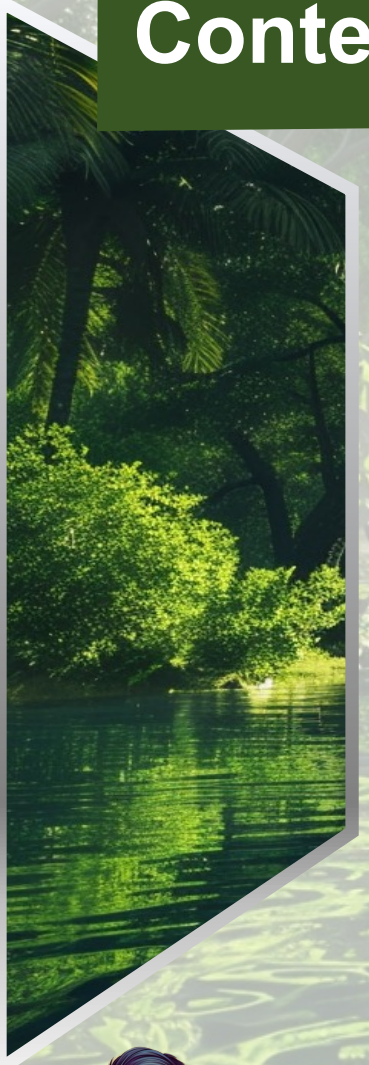
2x USB 3.2 Gen2 x1 Type-C
1x USB 3.2 Gen2 x1 Type-A

x16 / x8 + 2 x4 / 4 x2 + 2 x4
PCIe4 RC / EP

4 x1
PCIe3 RC / EP / SATA

2 x1 PCIe3 RC / EP / 1-5G Eth
2 x1 PCIe3 RC / EP / 10-25G Eth

Contents



3 RISC-V and AI Enthusiasts

RISC-V Prosperity 2036

ASE PLCT

- Based on RISC-V, realize the open standard system and open source system software stack covering the whole information industry from data center to desktop office, from mobile wear to intelligent Internet of Things
- Complete the adaptation and optimization for RISC-V in all basic key industry fields
- Form a network of tens of thousands of top talents

Partners


中辰計劃
算能
SOPHGO
Thank you
for watching