groq®

深度分析 & 产业洞察

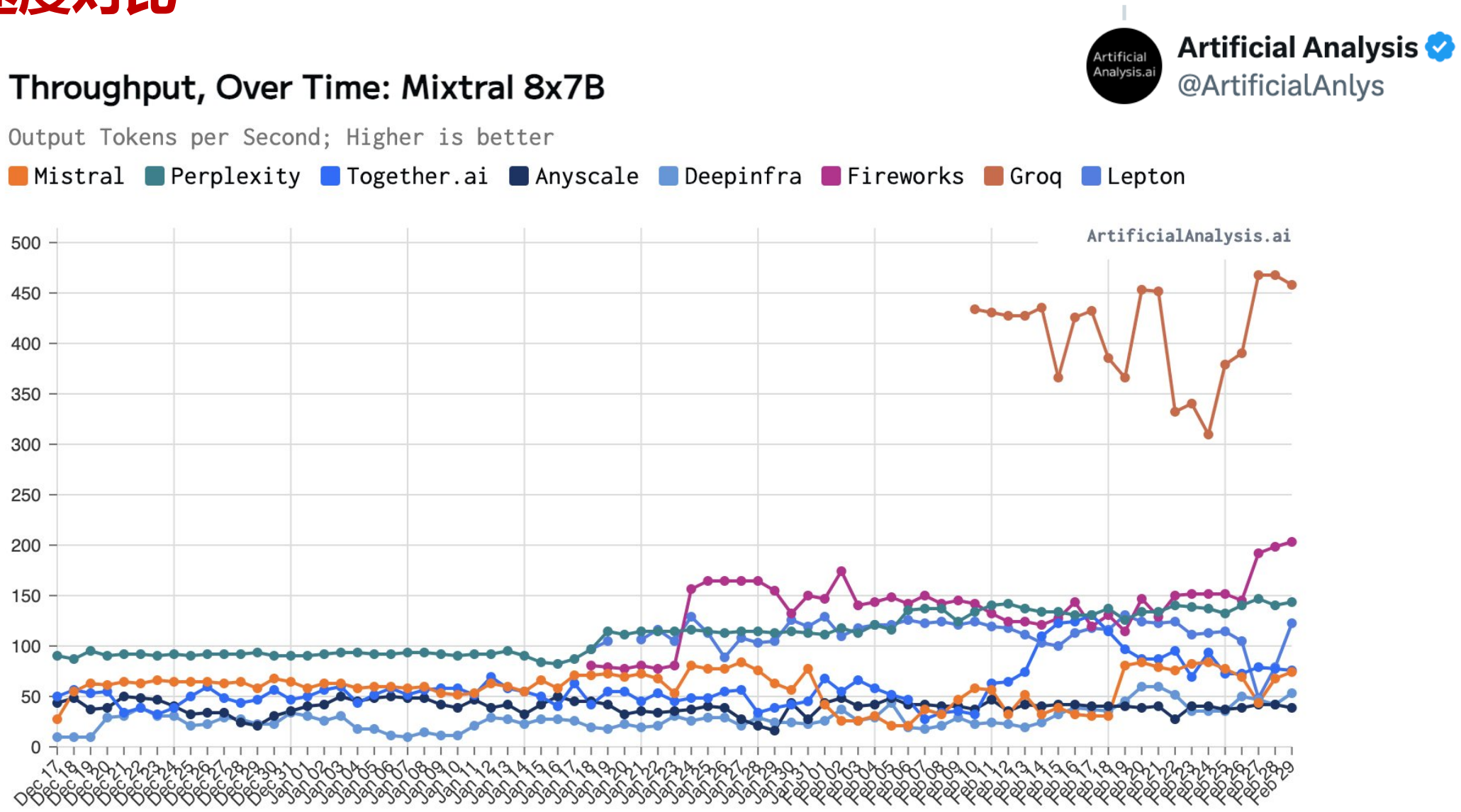
ZOMI

Groq Talk Overview

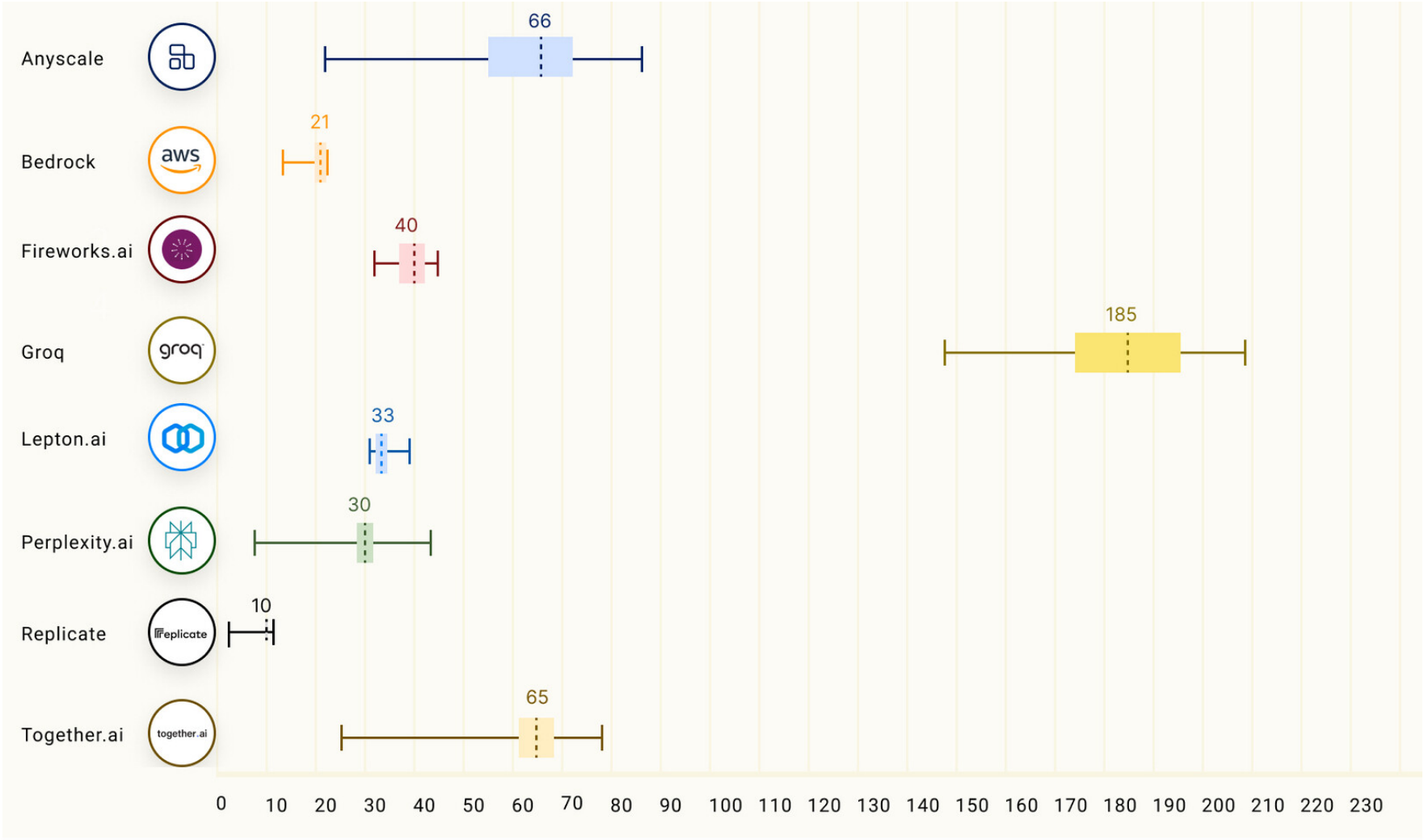
1. **延迟与吞吐**：Groq 成功背后推理的关键指标
2. **背景与概况**：Groq 技术总结 & 效果 & 背景
3. **技术架构**：Groq 软件定义硬件 & TSP 内核解读
4. **一些思考**：推理算力发展 & 市场预测

Groq 效果

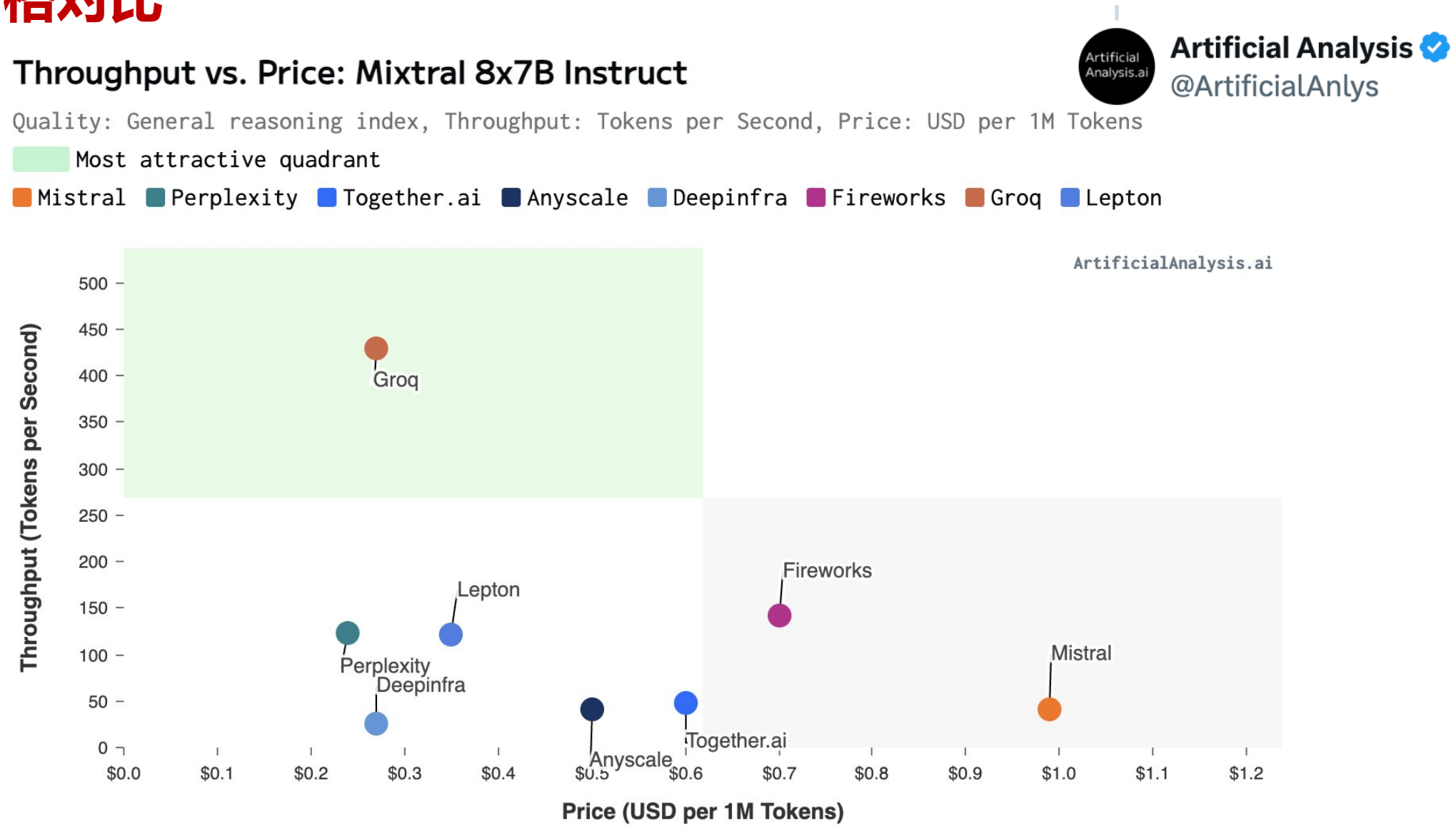
推理速度对比


Artificial Analysis
@ArtificialAnlys
Throughput, Over Time: Mixtral 8x7B
Output Tokens per Second; Higher is better
Mistral
Perplexity
Together.ai
Anyscale
Deepinfra
Fireworks
Groq
Lepton
ArtificialAnalysis.ai
500
450
400
350
300
250
200
150
100
50
0

推理速度对比


Anyscale
66
Bedrock
21
Fireworks.ai
40
Groq
185
Lepton.ai
33
Perplexity.ai
30
Replicate
10
Together.ai
65
0 10 20 30 40 50 60 70 80 90 100 110 120 130 140 150 160 170 180 190 200 210 220 230

推理价格对比


Artificial Analysis
@ArtificialAnlys
Throughput vs. Price: Mixtral 8x7B Instruct
Quality: General reasoning index, Throughput: Tokens per Second, Price: USD per 1M Tokens
Most attractive quadrant
Mistral
Perplexity
Together.ai
Anyscale
Deepinfra
Fireworks
Groq
Lepton
ArtificialAnalysis.ai
Throughput (Tokens per Second)
500
450
400
350
300
250
200
150
100
50
0
Groq
Lepton
Perplexity
Deepinfra
Anyscale
Together.ai
Fireworks
Mistral
$0.0
$0.1
$0.2
$0.3
$0.4
$0.5
$0.6
$0.7
$0.8
$0.9
$1.0
$1.1
$1.2
Price (USD per 1M Tokens)

1. 延迟与吞吐

真正影响大模型推理的指标

延迟 Latency

- 延迟 Latency 输入文本 → 获取结果时间（以每词元每秒 Tokens/s ）：
  - 正常情况下，延迟越低越好；人类阅读速度有限，所以 LLM 的输出有一个合理延迟区间
  - 确保流畅用户体验，LLM 输出速度 >10-15 Tokens/s，即单 Token 延迟 <100ms，H100 ~50 Tokens/s

吞吐 Throughput

- 吞吐指 LLM 单位时间内能处理数据量（每秒查询数 Query/second ）：
  - 吞吐直接影响 LLM 在实际应用中效率、成本、用户体验；
  - 吞吐量越大越好，但是因为在有限的硬件资源下，吞吐量大可以支持更多并发的用户访问；

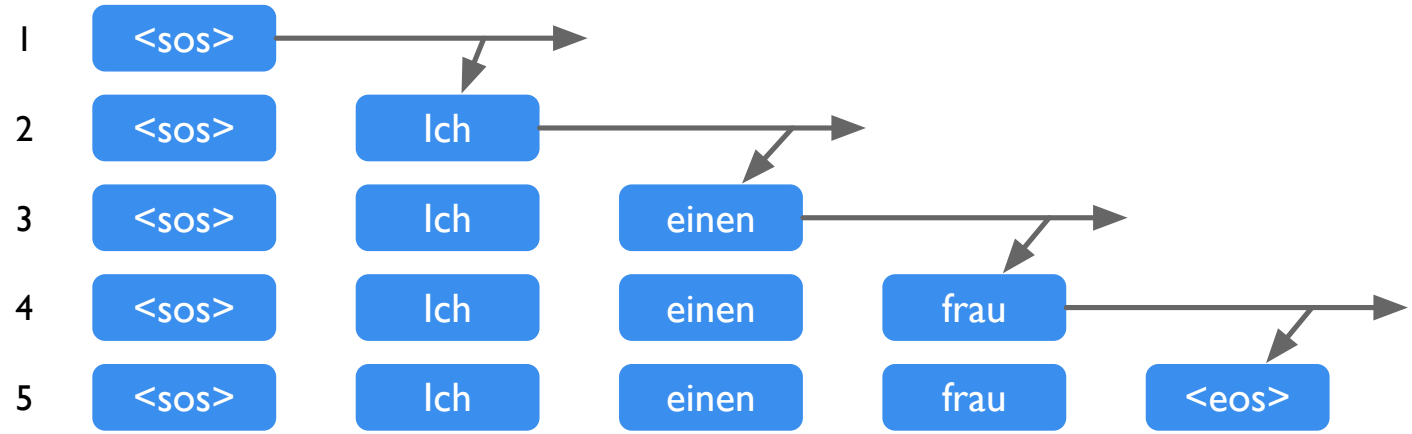
大模型推理阶段

1. Prefill：提示词 Prompt 处理被称为预填充（prefill），一次性将大量 Tokens 预先进行计算输入到模型中去生成 KV-Cache，通过一次 Forward 就可以完成。
2. Decoding：从生成第一个 Token 后开始，采用自回归方式一次生成一个 Token，直到生成一个特殊的 Stop Token 为止，常规的大模型推理阶段。

- Example Prompt：
  - 我是你老爸，你是我儿子，请你说话放尊重点，回答问题前都用尊称。我给出的问题是 {Prompt}，你需要知道的上下文是 {content}

大模型推理阶段

1. Prefill：提示词 Prompt 处理被称为预填充（prefill），一次性将大量 Tokens 预先进行计算输入到模型中去生成 KV-Cache，通过一次 Forward 就可以完成。
2. Decoding：从生成第一个 Token 后开始，采用自回归方式一次生成一个 Token，直到生成一个特殊的 Stop Token 为止，常规的大模型推理阶段。



大模型推理影响因素

- LLM 推理时延由 Prefill 阶段时延和 Decoding 阶段时延决定：
  - Prefill 时延对应第一个 Token 的时延，与输入序列中 Token 数量有关；
  - Decoding 阶段时延决定整体时延，由输出序列的 Token 数决定 → **Groq 性能提升的重点；**

Question

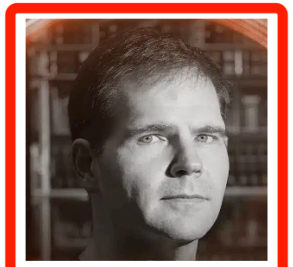
- 为什么 Groq 这么挫的AI 芯片，推理时延能比 NV 超过10X？

	Groq	H100
制造工艺	14nm	4nm
单芯片算力 FP16 TFLOPs	188	989
节点内带宽 GB/s	RealScale：800	NVLinks：900
单节点内存	1.76GB	640GB
80GB 推理所需节点数	45	0.125
80GB 推理所需卡数	347	1

2. Groq 背景与概况

Groq 的创立

- 2016 年，Google TPU 架构师 Jonathan Ross 和 TPU 团队的其他成员创立了 Groq



Jonathan Ross
CEO & Founder

Adam Tachner
CLO & VP of Corporate Development

Estelle Hong
Head, Planning & Execution (PLEX)

Tobi Crabtree
VP of People & Culture

Mark Heaps
VP of Brand & Creative

Aileen Black
President, Government & Public Sector

Sheryl Savage
General Counsel

Jim Miller
VP, Hardware Engineering

John Barrus
VP of Product

Groq 的发展

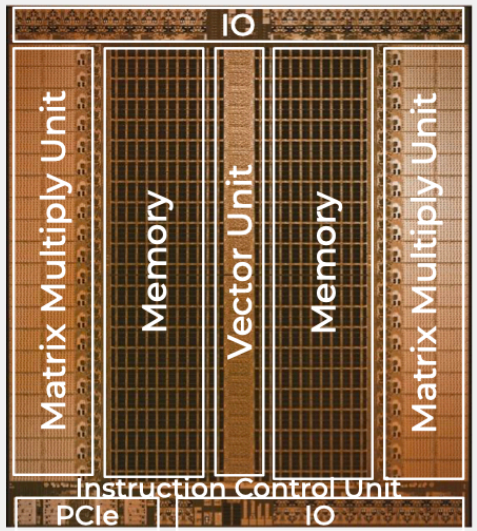
- 2016 年，Google TPU 架构师 Jonathan Ross 和 TPU 团队的其他成员创立了 Groq
- 2020, January 《GROQ ROCKS NEURAL NETWORKS 》
- 2020 《Think Fast: A Tensor Streaming Processor (TSP) for Accelerating Deep Learning Workloads 》
- 2022, June 《A Software-defined Tensor Streaming Multiprocessor for Large-scale Machine Learning》
- 2023, hot-chip 34 《The Groq Software-defined Scale-out Tensor Streaming Multiprocessor》
- 2024, February 因为 Mistral-MOE 7*8B 500 tokens/s 真正火起来

Groq 芯片形态

SRAM Memory
Massive concurrency
80 TB/s of BW
Stride insensitive

Groq TruePoint™ Matrix
4x Engines
320x320 fused dot product
Integer and floating point

Programmable Vector Units
5,120 Vector ALUs* for high performance



GroqChip 1

Networking
480 GB/s bandwidth
Extensible network scalability
Multiple topologies

Data Switch
Shift, transpose, permuter for improved data movement and data reshapes

Instruction Control
Multiple instruction queues for instruction parallelism

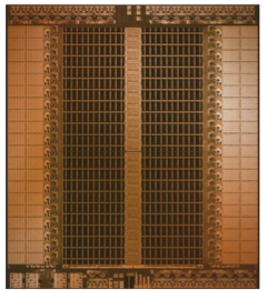
Groq 的定义

神经网络在推理时：

- 计算内容可根据网络架构提前定义
- 软件提前调度计算资源 Schedule
- 减少硬件控制单元开销
- 将更多晶体管留给计算 & Memory内存

Groq 思路：

- 硬件极致简化
- 调度交给编译器
- 硬件根据指令执行



The TSP Simplifies Data Flow Through Stream Programming

Predictable performance at scale

A large, single-level scratchpad SRAM, fixed, deterministic latency

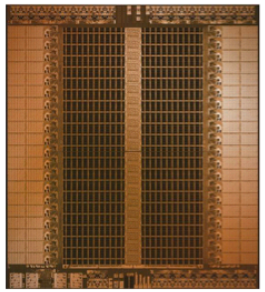
Explicitly allocate tensors in space and time unlocking

massive memory concurrency, and compute flexibility along multiple dimensions:

- Device, hemisphere, memory slice, bank, address offset

Groq 的定义

- Groq 称架构为 Software-defined Tensor Streaming Multiprocessor，突出软件定义。



The TSP Simplifies Data Flow Through Stream Programming

Predictable performance at scale

A large, single-level scratchpad SRAM, fixed, deterministic latency

Explicitly allocate tensors in space and time unlocking

massive memory concurrency, and compute flexibility along multiple dimensions:

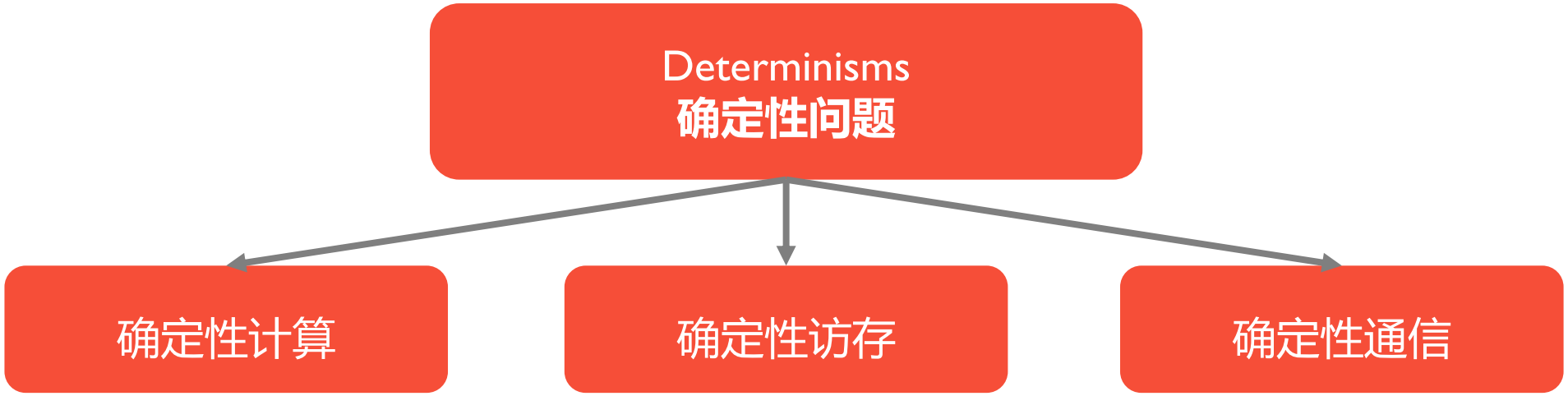
- Device, hemisphere, memory slice, bank, address offset

Groq 软件定义硬件

1. 如何让计算过程完全由软件定义和调度？

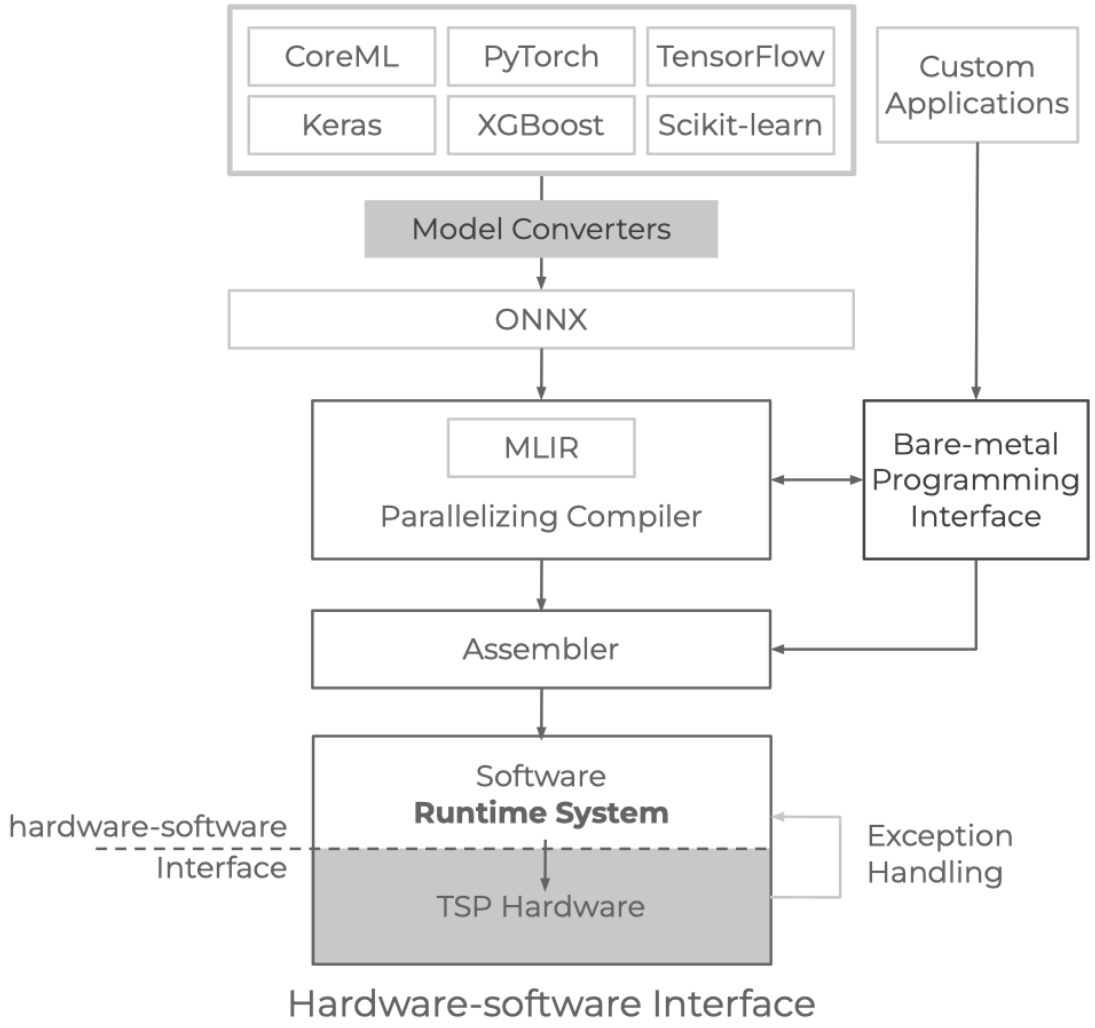
Groq 软件定义硬件

- 编译器提前规划好所有计算逻辑，芯片 chip 上执行一切指令都有确定延迟，即保证**确定性的计算，确定性的访存和确定性的通信。**
- ① **确定性计算**需要明确所有执行指令的执行序和执行时间；② **确定性访存**要求避免使用DRAM作为存储单元（DRAM 访问延迟不固定）；③ **确定性通信**要求通信的逻辑和时序严格同步。



Groq 软件定义硬件

- **软件接口层面：**
  - 动静态接口：“编译时”（静态）与“执行时”（动态）执行的内容，由 runtime 统一管理。
  - 软硬件接口：确定芯片的状态对编译器来说是可见的，这样可以进行正确性预测
- **编译器层面：**
  - 通过计算图进行确定性计算预测
  - 图中“节点”表示 OP，“边”表示张量数据和结果



Groq 芯片形态

ICU: instruction control units

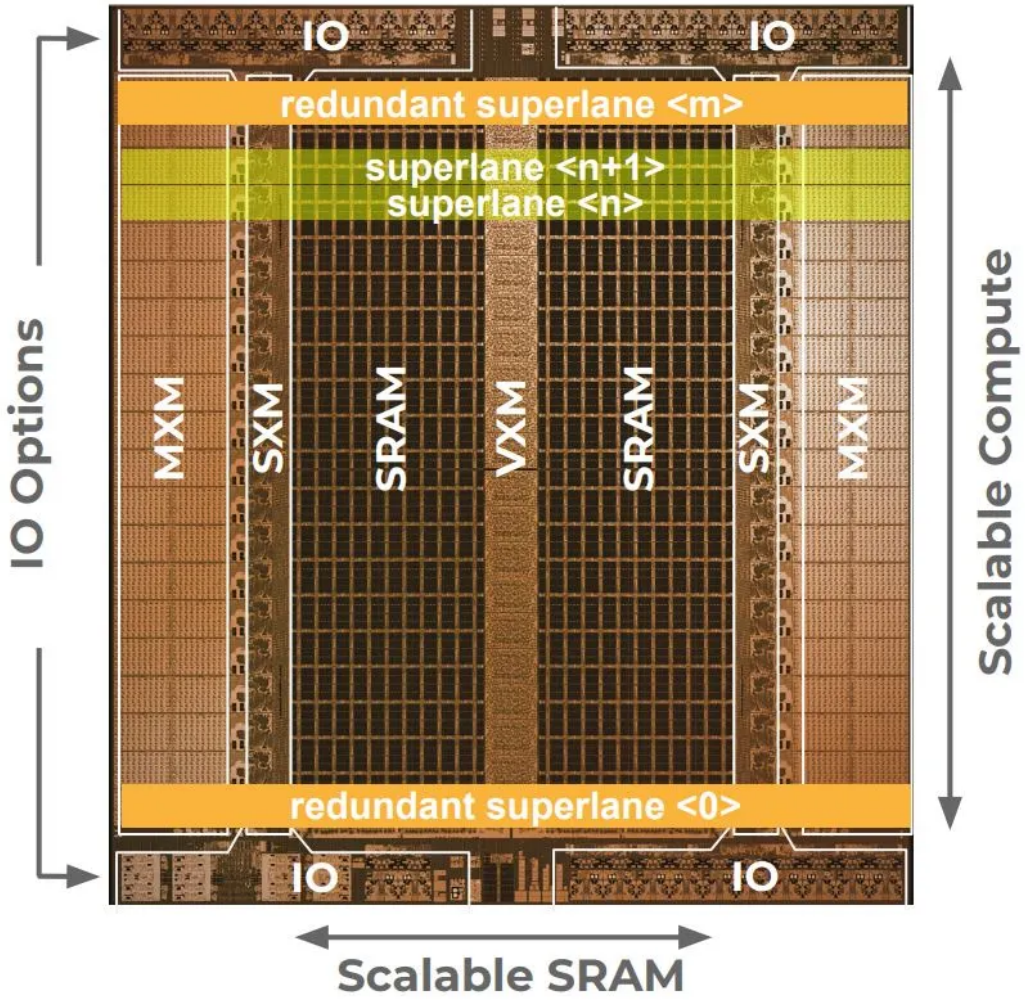
MEM: on-chip memory (SRAM)

VXM: vector processing

MXM: matrix operations

SXM: data reshapes and IO

- 张量的数据水平地流动
- 指令以SIMD的方式垂直执行



Groq 产品形态

GroqChip™

GroqCard™

GroqNode™

GroqRack™

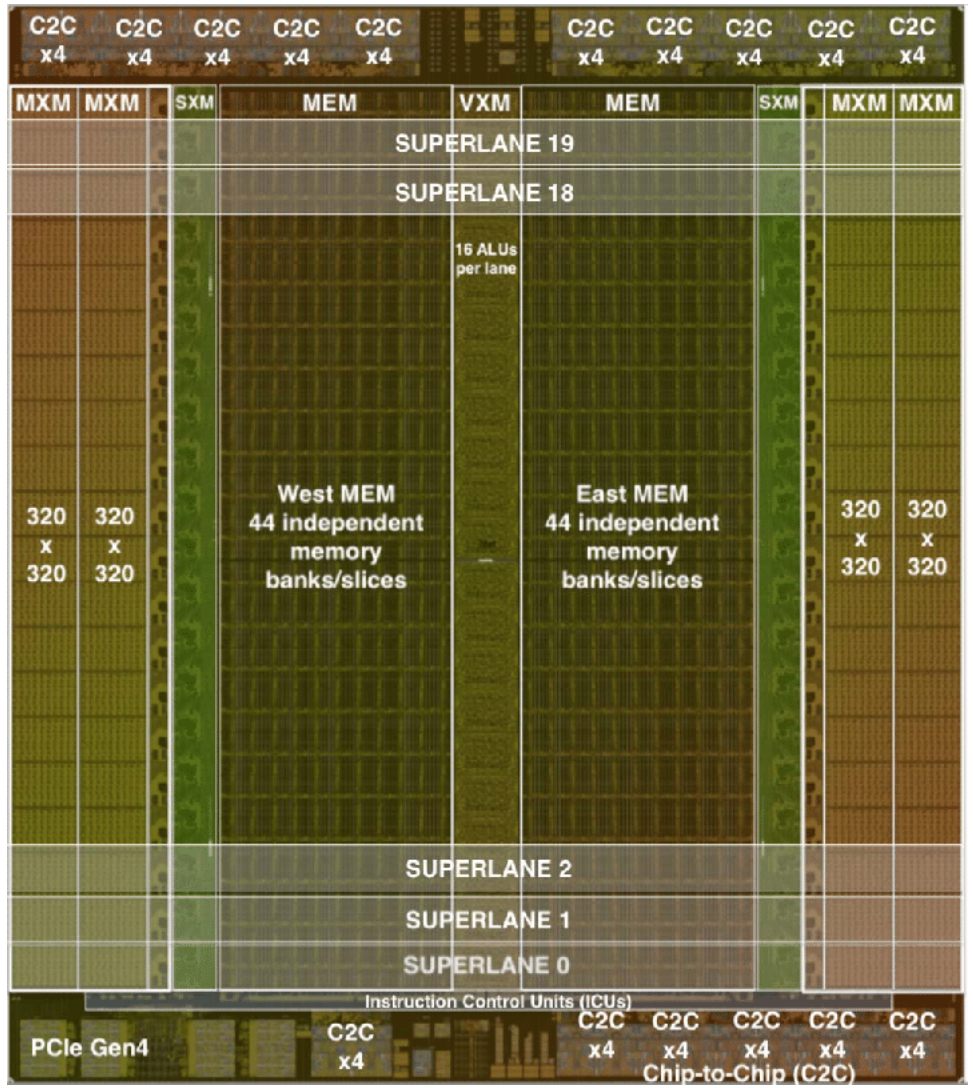
Cloud

3.1 Groq 技术架构

软件定义硬件

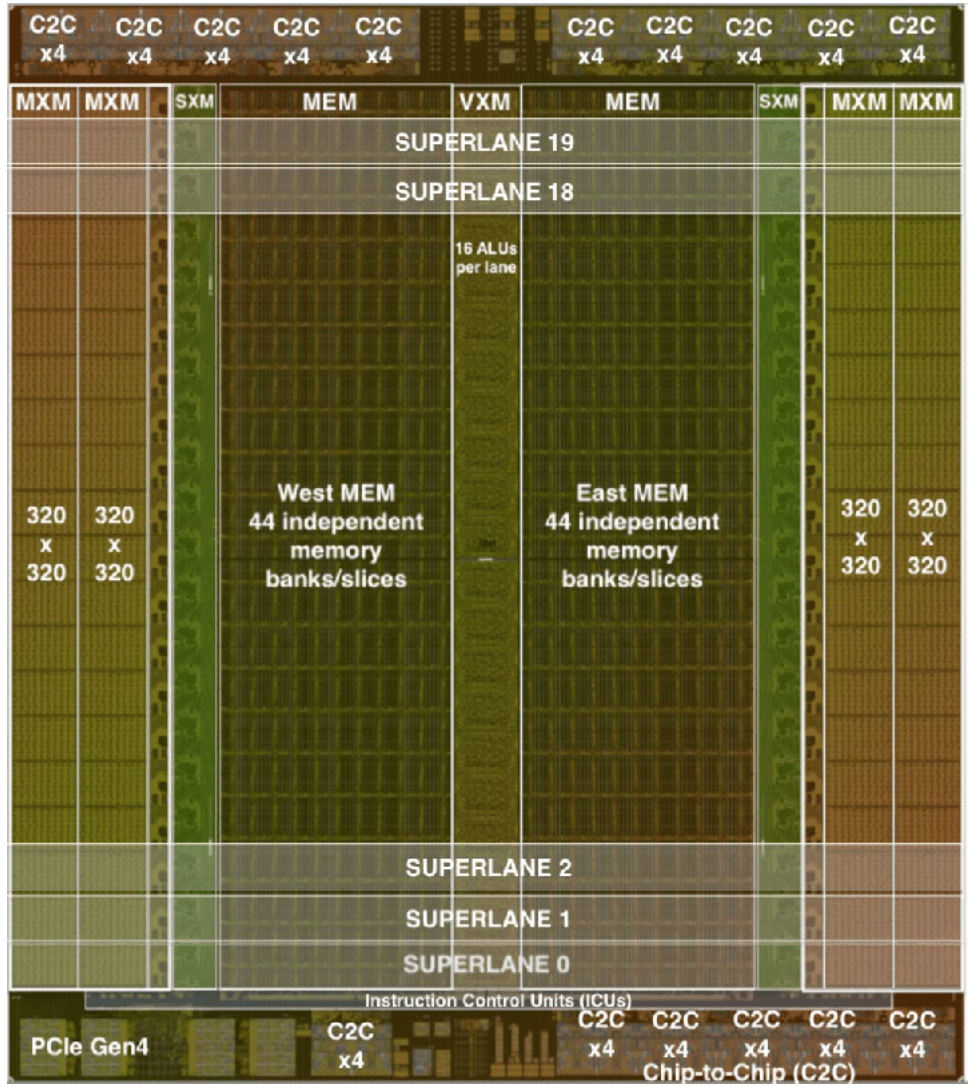
TSP 芯片

- 一个 PCI Express Gen4 x16 接口连接 host CPU
- 片上 220MB SRAM，没有DRAM控制器和接口
- 14nm 工艺，面积725mm2，含 268 亿个晶体管
- 16个 C2C 模块，支持芯片间 320B 向量传输
- 3.84 Tb/s 片外互联带宽，每路 C2C 链路带宽 30 Gbps，片外总带宽 16 ×4 ×30Gb/s ×2 路方向



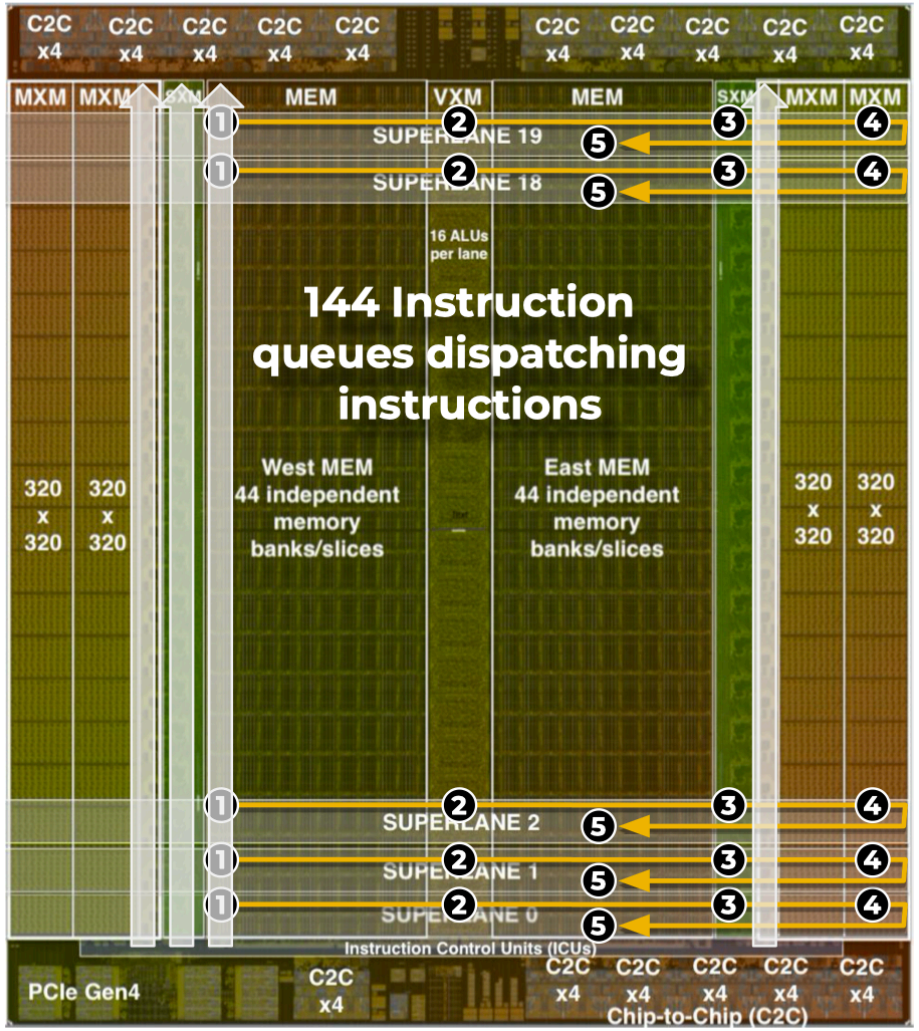
TSP 芯片

- 20 个 Superlane 可用
- 一个冗余 Superlane 提高芯片良率
- 每周期执行 400,000 次整数乘法累加（MAC）
- 可以处理 FP32/FP16 数据，兼顾推理和训练场景



TSP 芯片数据流动

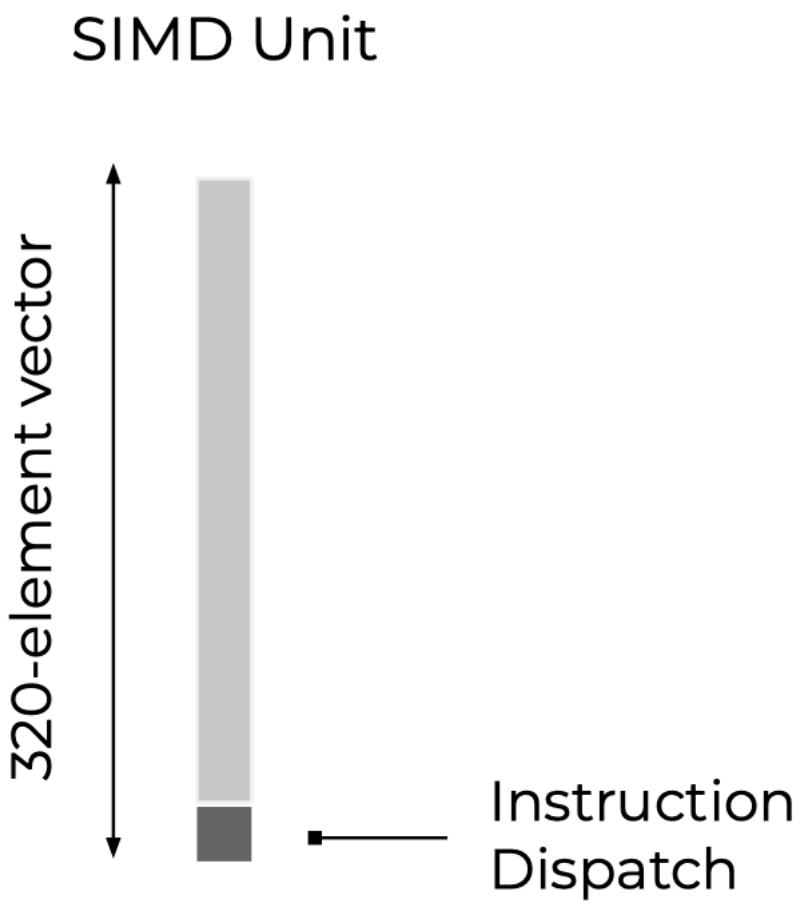
- 存储和计算单元 east、west 左右均匀分布
- 峰值算力是 INT8 750TOPs ， FP16 188 TFLOPs
- VXM（Vector）位于芯片的中间：
  - VXM 每个通道有 16 个 PE，总计 5120（16*320）个
  - 32-bit computations/cycle or 4xINT8 computations/cycle
- MXM（Matrix）为 320 x 320 features:
  - 每个 MXM 单元可以存放 102,400 个 FP16 权重参数
  - 409,600 MACCs (multiply-accumulators)



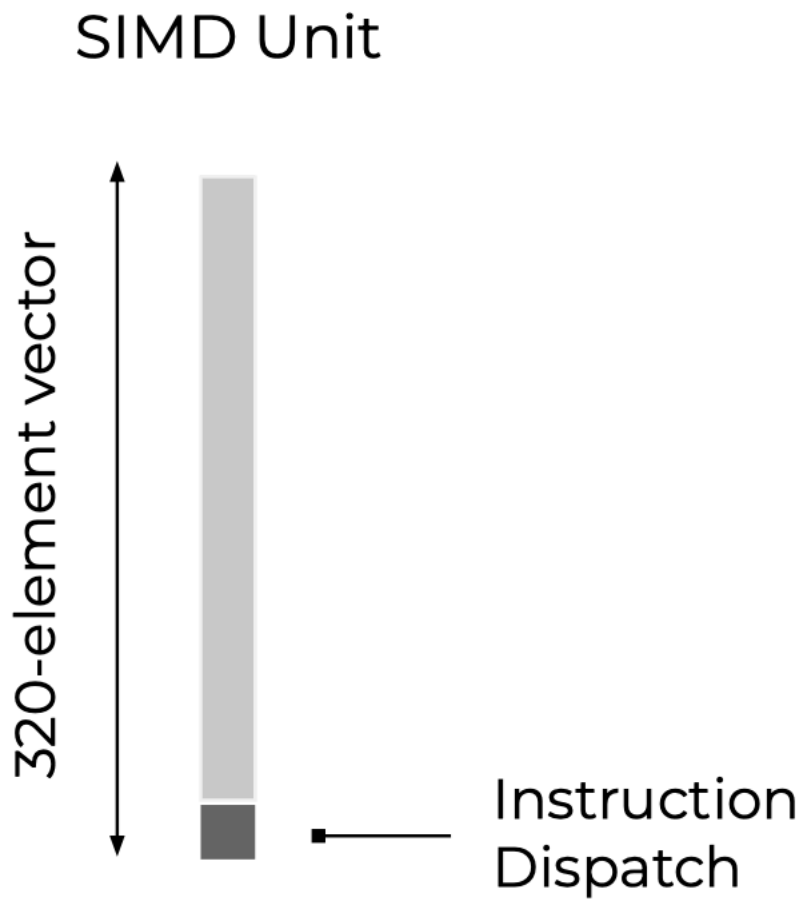
TSP 芯片

- TSP（Tensor Streaming Processor）硬件确定性的，实际吞吐量取决于编译器调度
- 编译器必须协调 144 宽 VLIW 执行单元，每个 VLIW 有 320B 的 SIMD

- Question：TSP 320x320 MAC 阵列来计算各种 Shape 张量对编译器挑战有多大？



TSP 软件定义



TSP 软件定义

Build different types of specialized SIMD units

MXM
Matrix-Vector /
Matrix-Matrix Multiply

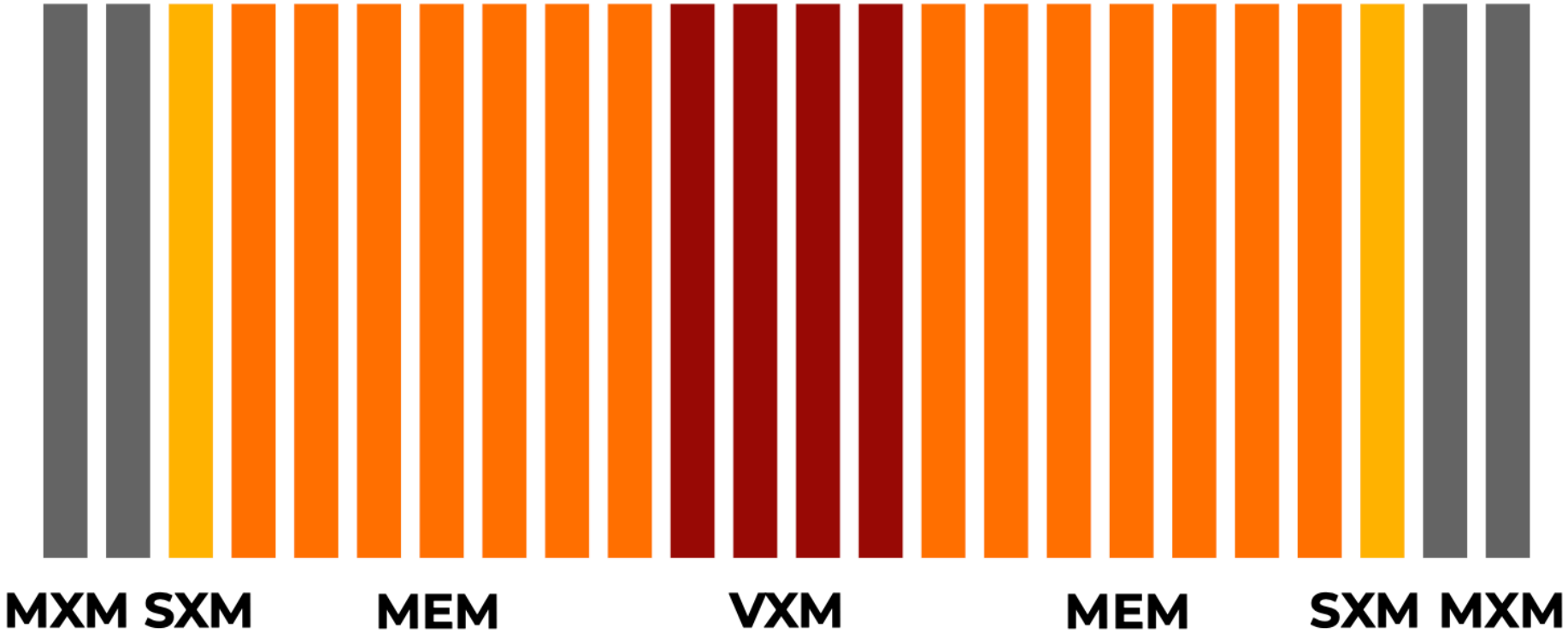
VXM
Vector-Vector
Operations

SXM
Data Reshapes

MEM
On-chip SRAM

TSP 软件定义

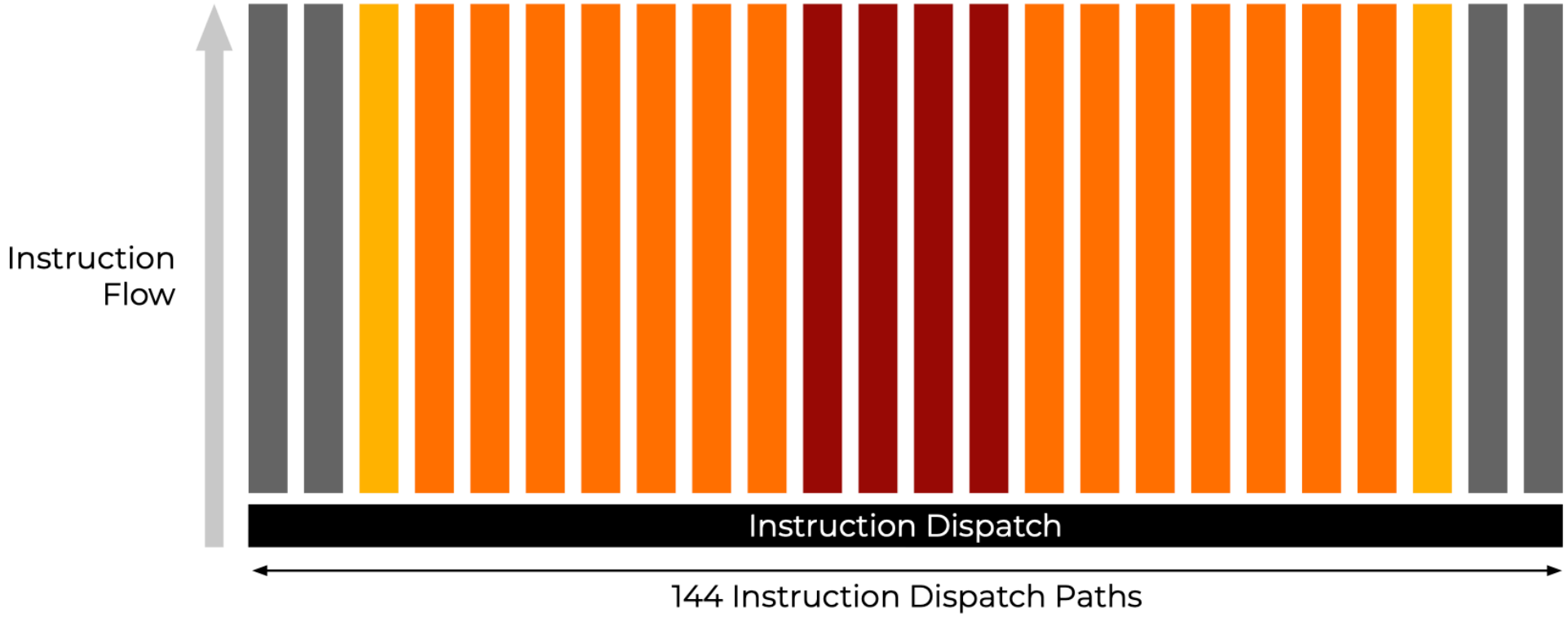
Lay out SIMD units across chip area



TSP 软件定义

Synchronized instruction dispatch across all SIMD units for lockstep execution

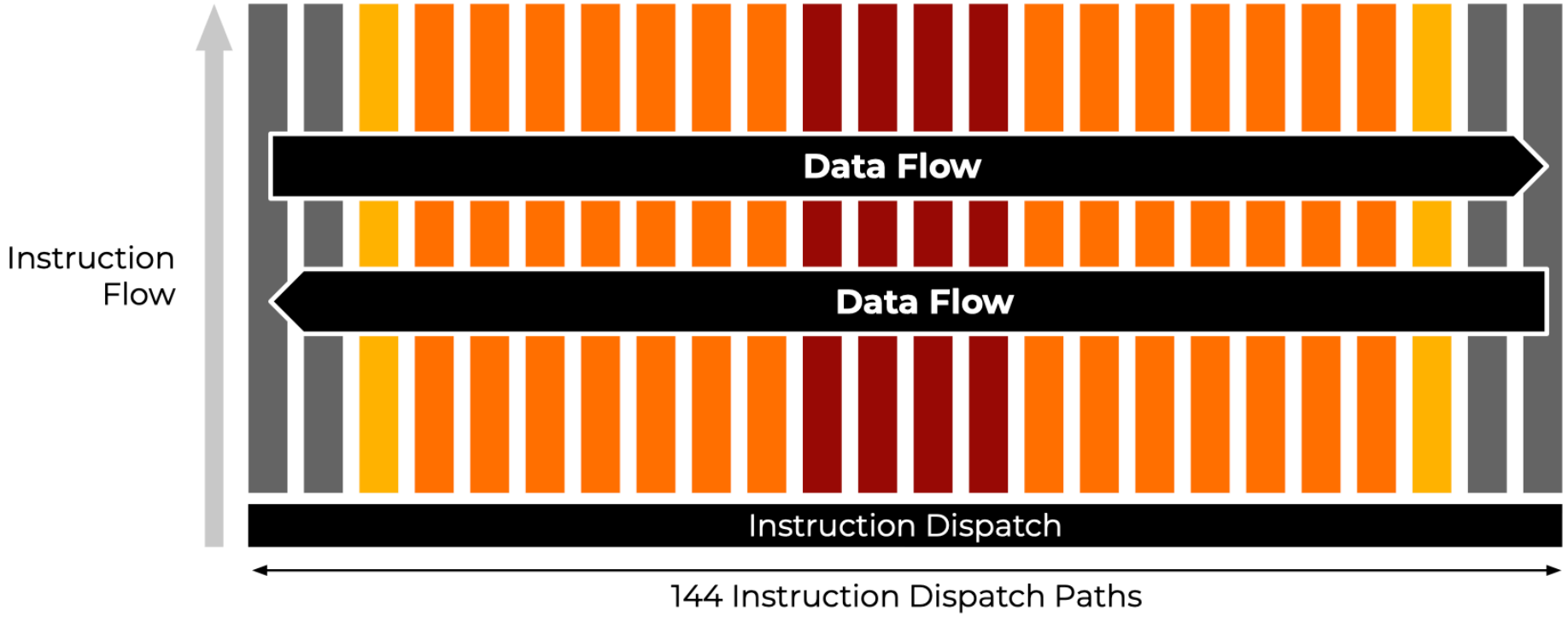
同步指令分发到所有 SIMD 单元，实现计算同步执行



TSP 软件定义

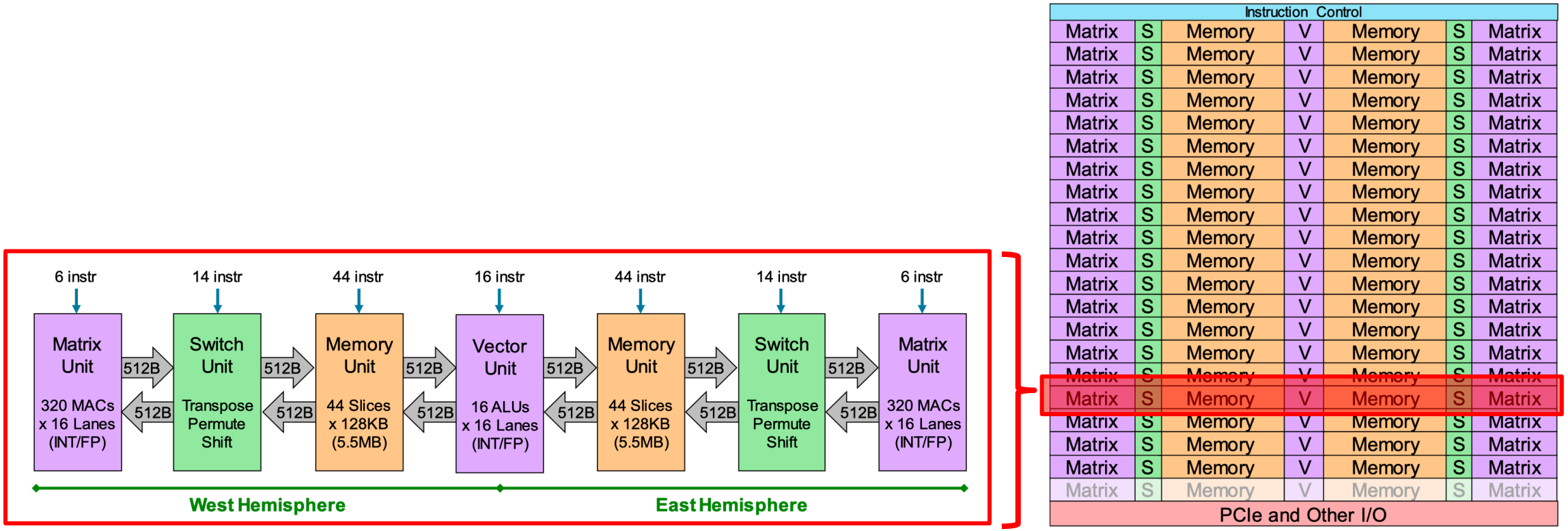
High-bandwidth “Stream Registers” for passing data between units.

提供高带宽的流寄存器（SR）在不同的单元之间流式地传递数据



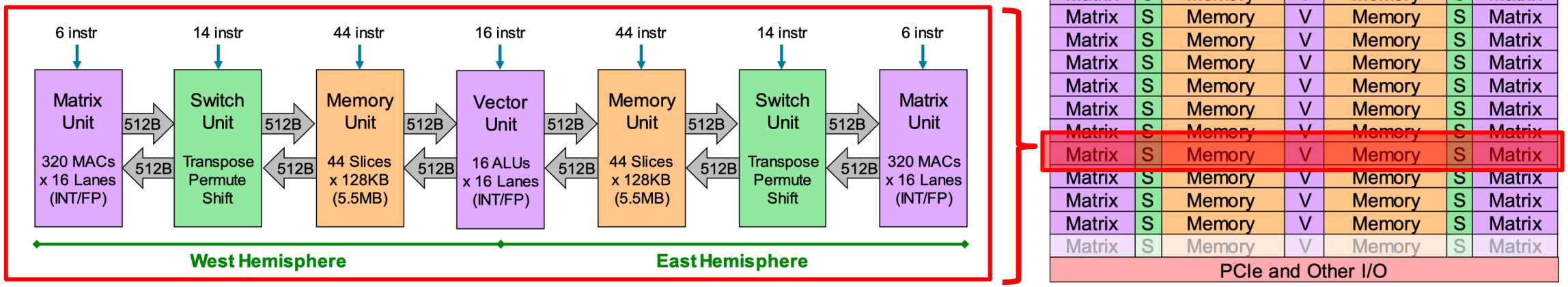
TSP Superlane

- 由 20 条 Superlane 组成，每条 Superlane 即一条数据处理的流水线，包含：
  - 矩阵乘法单元 MXM（x2），内存单元 MEM (x2)，向量处理单元 VXM(x1) 和数据交换单元 SXM (x2)



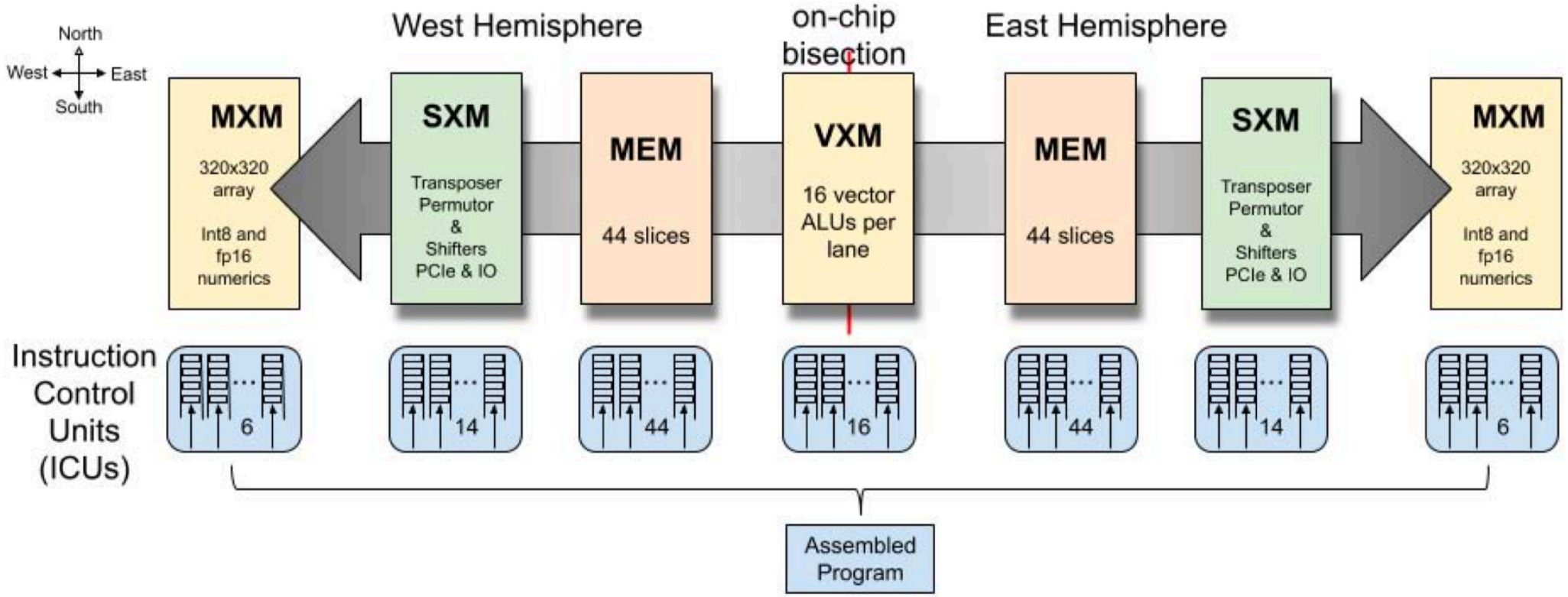
TSP Superlane

- 数据在流水线向东西流动，指令由 ICU 指令控制单元从 Superlane 0 流入：
  - 指令在传入 ICU 后对数据流进行处 or 从 MEM 中读取/写入数据到数据流；
  - Superlane 分为 16 lane，lane/Cycle 接受同样指令，执行同样操作，即16x32 SIMD 处理单元；
  - 20条 Superlane 组成一个 SIMD；



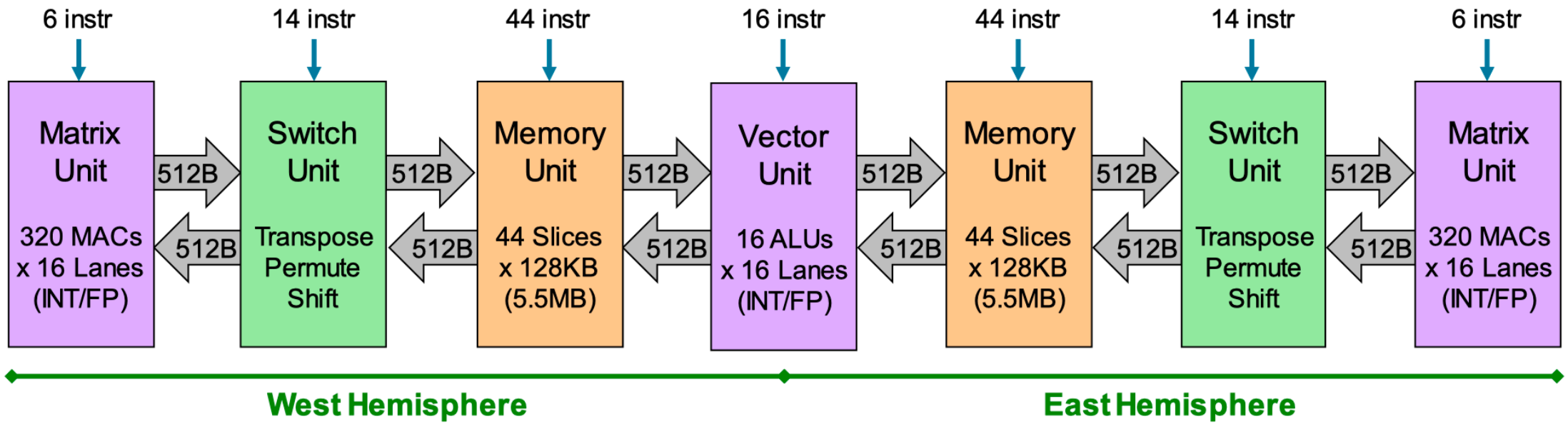
TSP Superlane 位宽

- TSP 执行单指令流 → 把 Superlane 当作微处理器内核
- 实际上是 144 位宽 VLIW 架构，每周期发出一条指令控制 Superlane



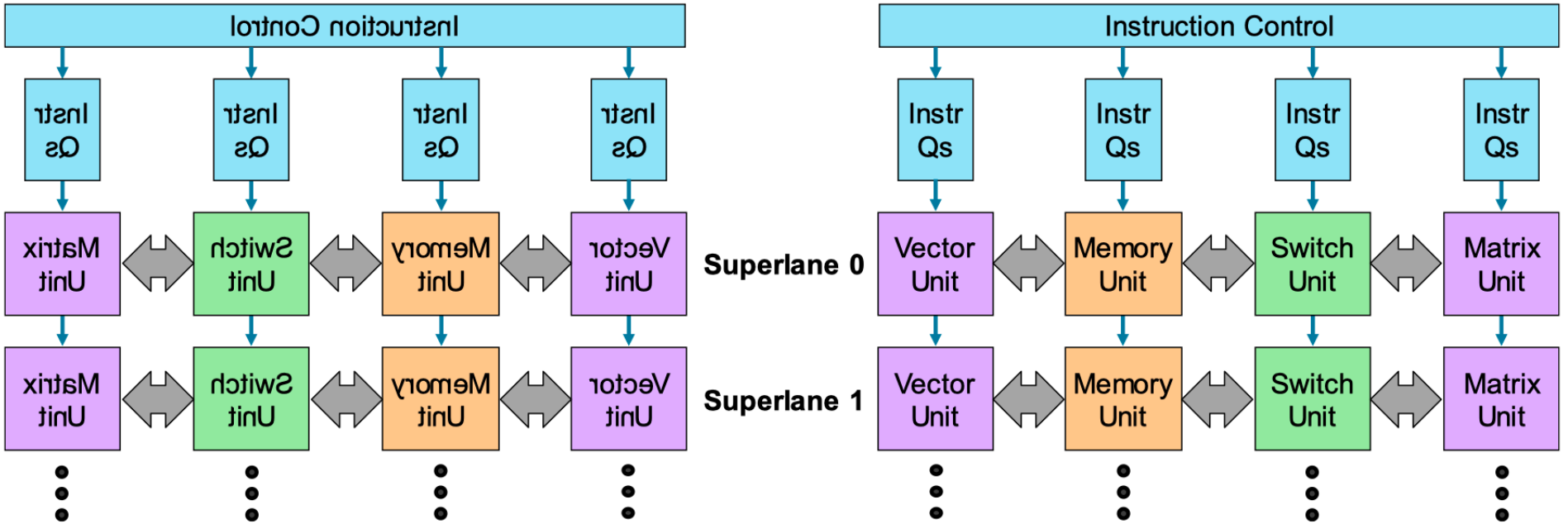
TSP Superlane

- East hemisphere 和 West hemisphere：
  - 每个功能单元包含多个接受指令的子单元，数据从东向西和从西向东流动。
  - 左右对称，每 lane 数据流位宽为 32Byte，总数据流位宽为 512B。



TSP Superlane

- 数据在不同 SIMD 上流动，不同时间交给不同单元执行：
  - 指令流入 SL0 并执行，下一 Cycle 指令流入 SL1，SL0 执行第二组指令。
  - 类脉冲阵列：数据水平移动，指令垂直移动 → 简化设计和布线，消除同步需求，并且易于扩展。
  - MEM 嵌在计算单元中，可以提供高带宽数据流，从而无需外部存储器。



TSP Superlane Pons

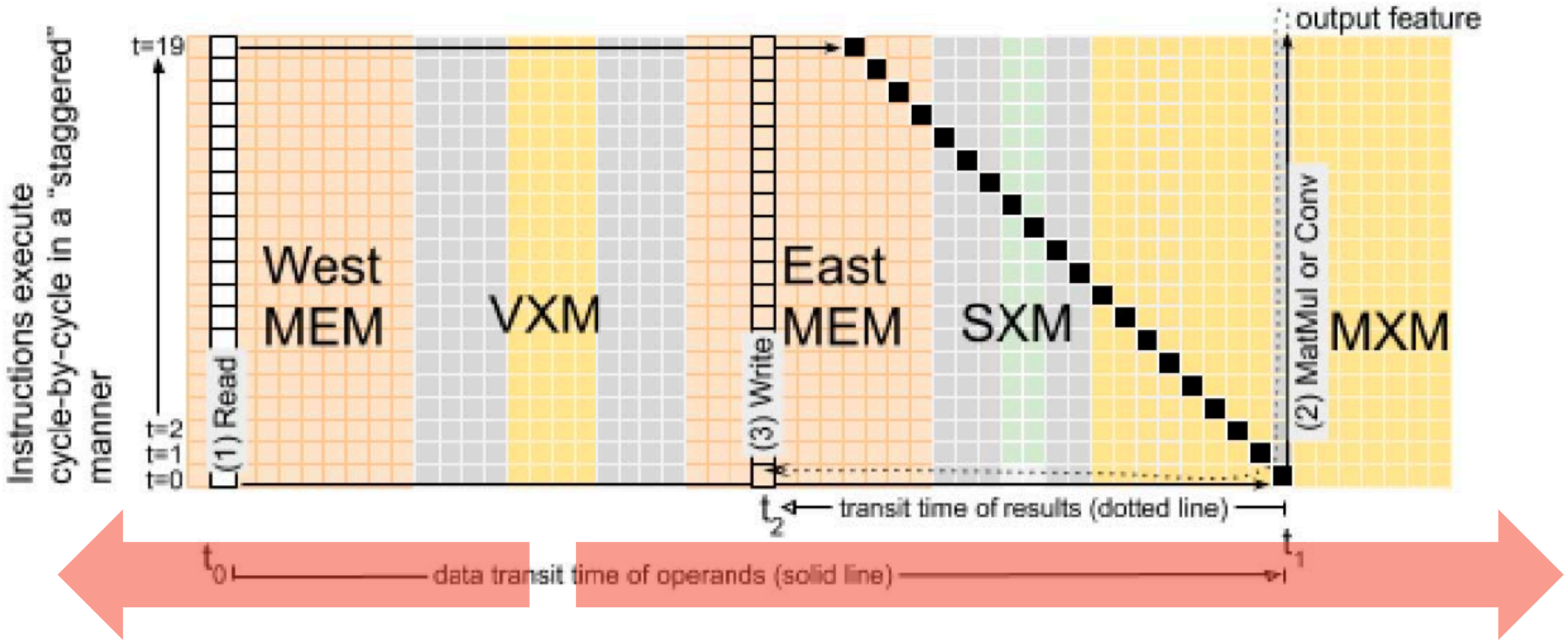
- 数据在不同 SIMD 上流动，不同时间交给不同单元执行：
  - TSP 缺乏缓存、分支预测等逻辑，执行是完全确定性的 → 编译器必须理解指令流和数据流，优化单元利用率。
  - 编译器必须安排数据移动 & 管理内存 & 功能单元 & 非自动取指。

3.2 Groq 技术架构

微内核模块

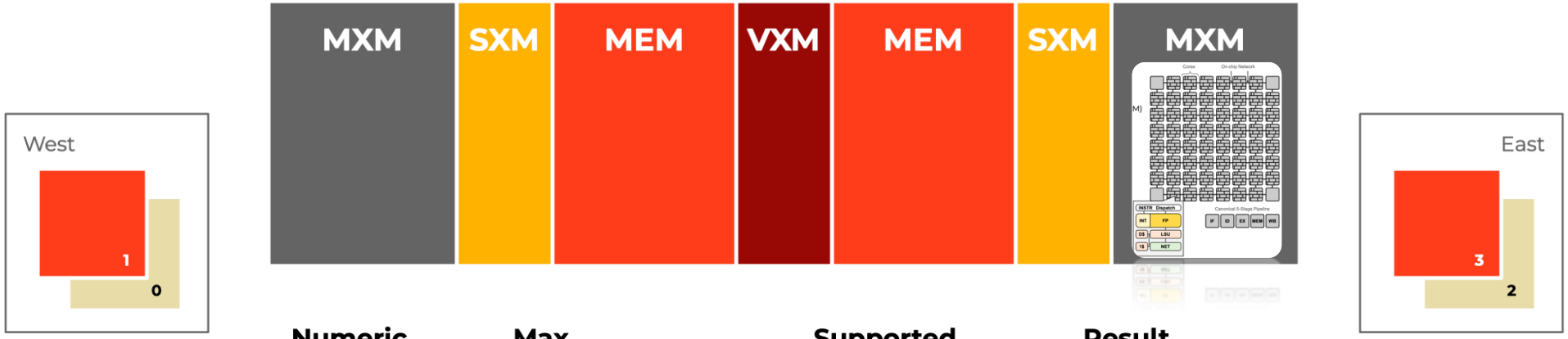
Superlane 细节展开

- 一个 Superlane 由 矩阵乘法单元 MXM（x2），内存单元 MEM (x2)，向量处理单元 VXM(x1) 和数据交换单元 SXM (x2) 组成。



矩阵乘法单元 MXM

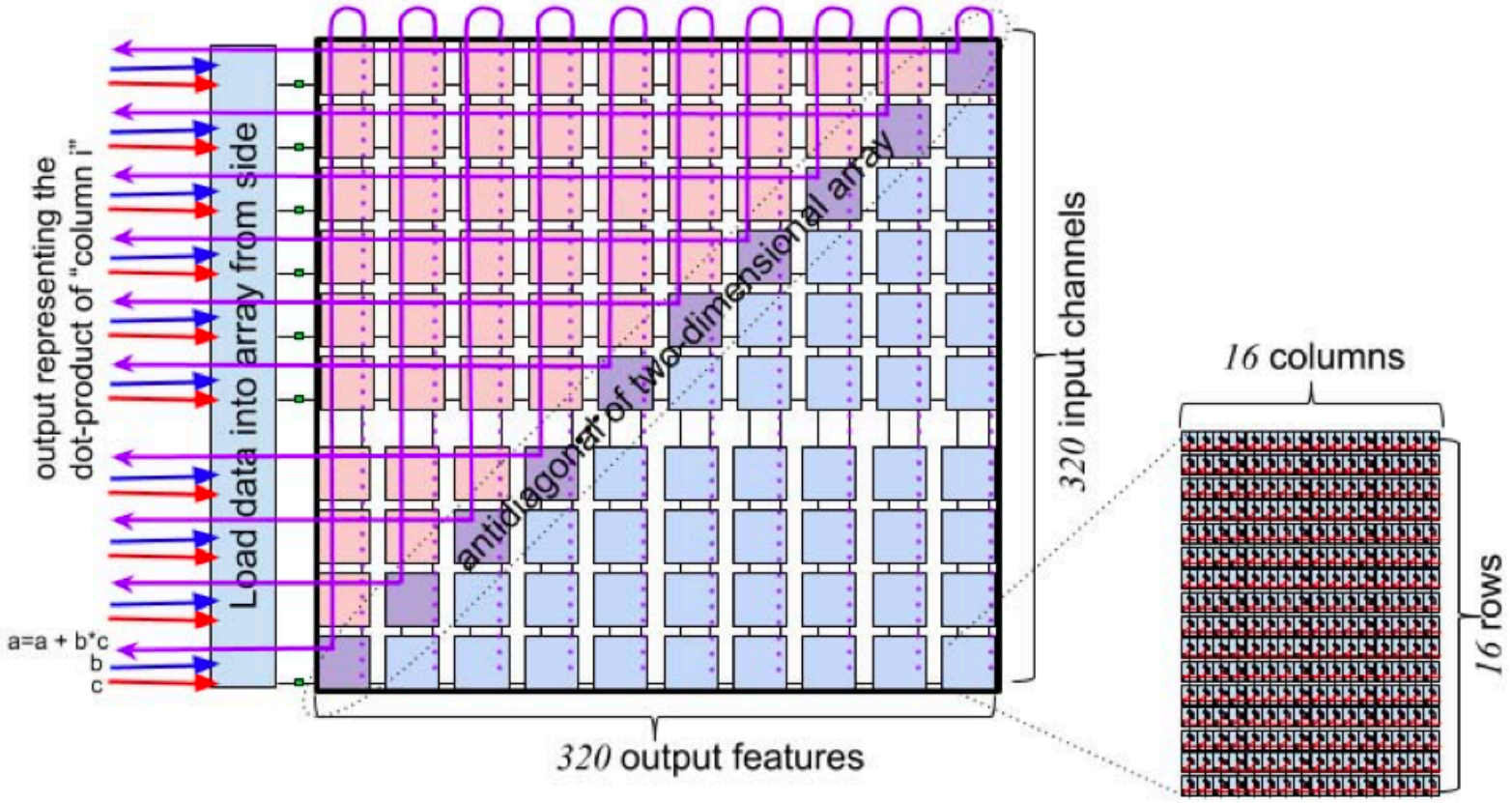
- 矩阵乘法单元包含 320 个 MAC，分成 20 个超级单元 Supercell，16 个 MAC 组成一个 Supercell。
- 每个hemisphere有 320x320 个 MAC 单元，每个 MAC 有两个 8B 权重寄存器和两个 32B 累加器。



Numeric Mode	Max Size	Supported Density	Result Tensor
int8	[N, 320] x [320, 320]	Two per MXM	int32
float16	[N, 320] x [160, 320]	One per MXM	float32

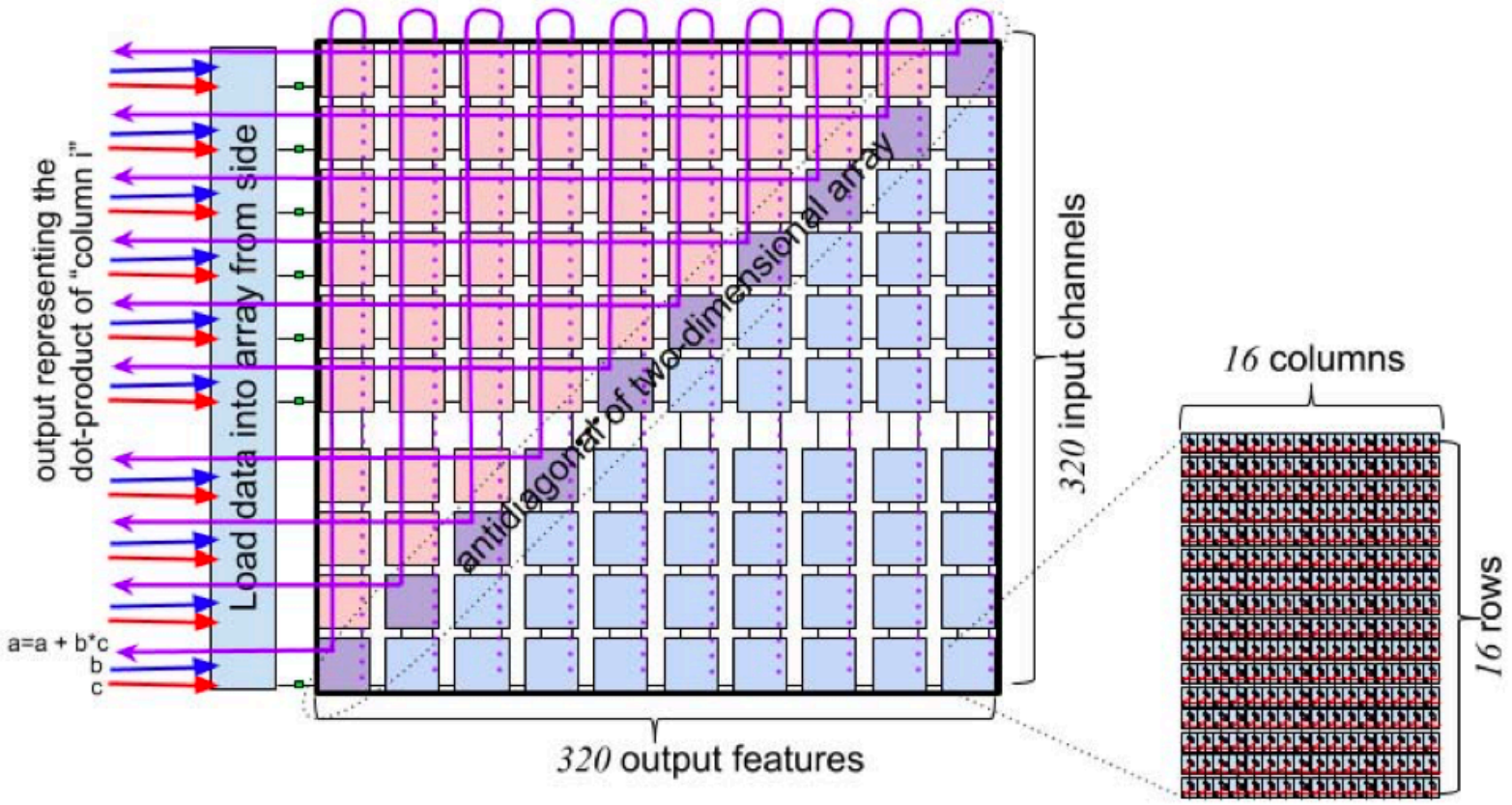
矩阵乘法单元 MXM

- 时钟周期内，权重值和数据流激活值相乘。每 16x16 Supercell 在一周期内计算一个整数部分和，20 周期内计算一个完整的 320 个元素点积。
- MAC 单元可执行单个 FP16/2 Cycle，相对 INT8 吞吐降低 75%。
- 每个 MAC 每 Cycle 产生 409,600 INT8 /102,400 FP16 计算。



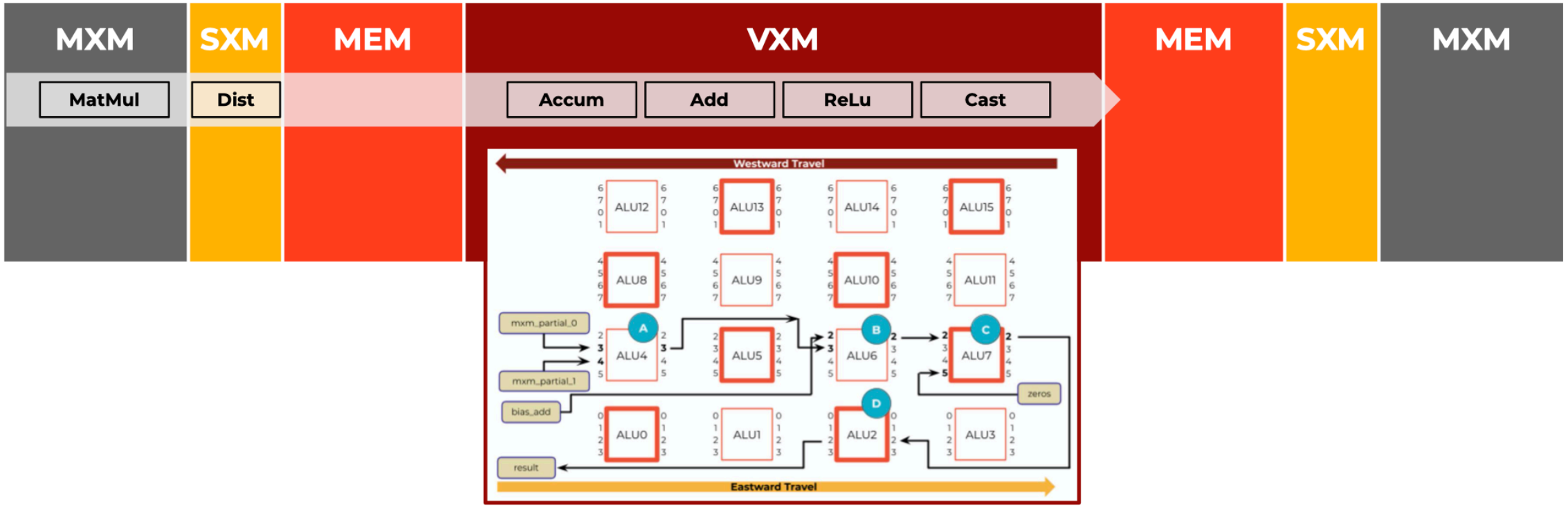
矩阵乘法单元 MXM

- 每个方向上使用 Superlane 的 32 个数据流 → <40 个周期内加载 409,600 个权重寄存器。
- 在阵列侧面加载的激活和权重，INT32 或 FP32 结果从内边缘流回.



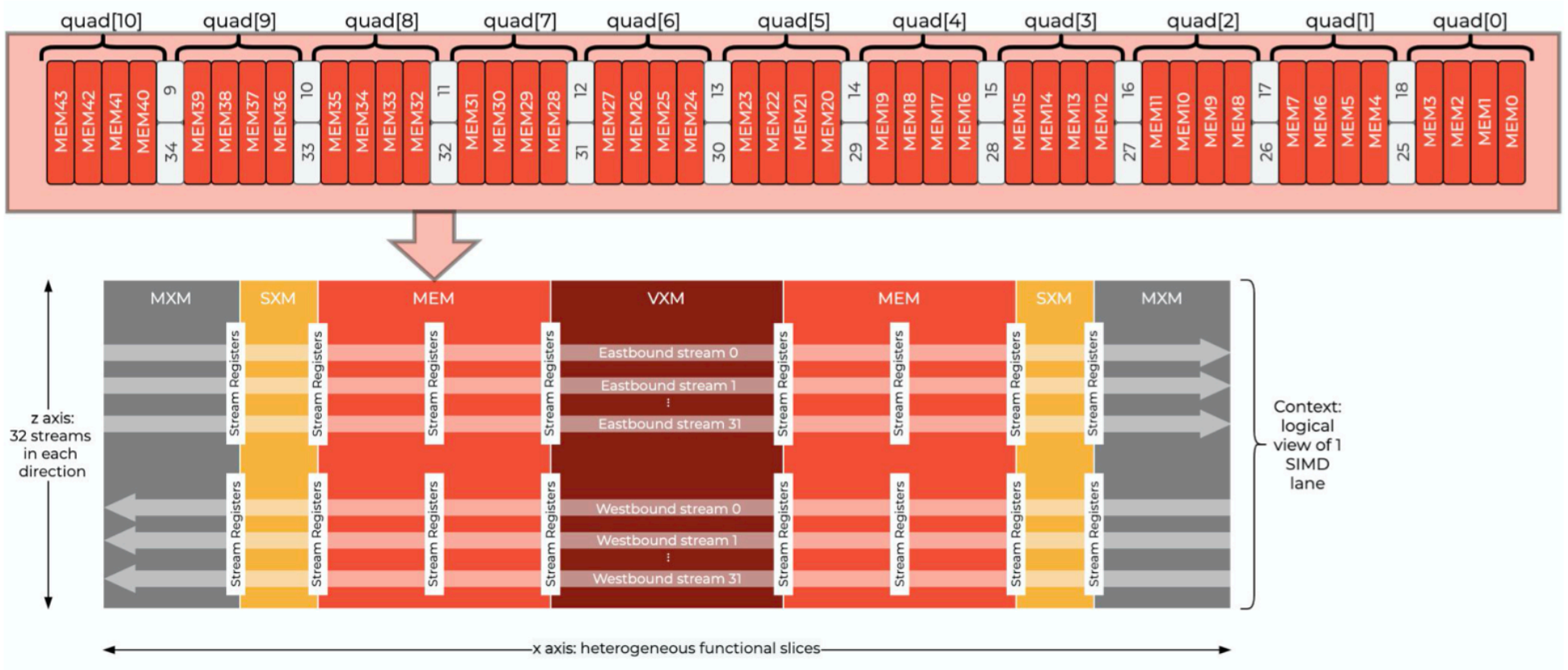
向量执行单元 VXM

- 每条通道 16 个 ALU，每个 ALU 可使用数据流对齐的 4 Byte 作为操作数来执行 32B 计算。
- 计算：常规算术和逻辑运算，整数和浮点格式转换，归一化和激活函数。



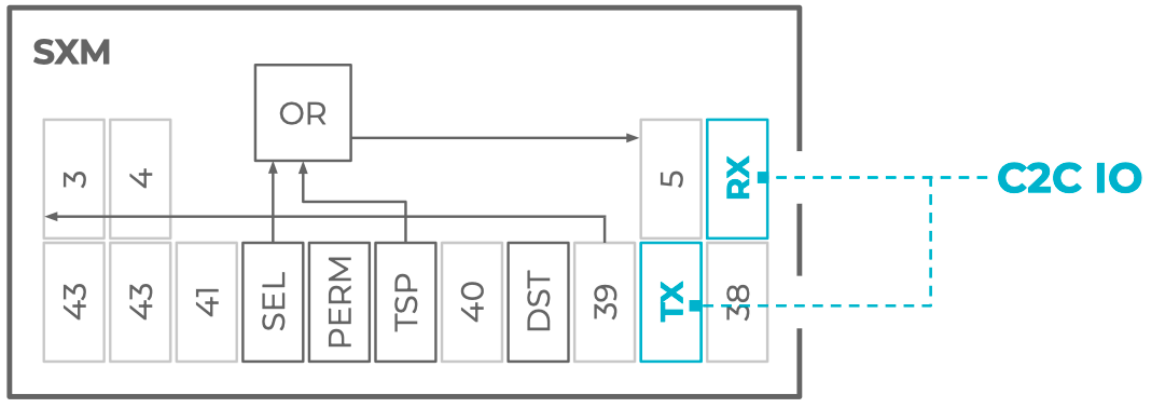
内存单元 MEM (x2)

- **组成**：每个 MEM 5.5MB SRAM(44 个切片，每切片 128KB)，20 个 Superlane 组成 Hemisphere 110MB SRAM。
- 每周期可执行 2x16Byte 读和 2x16Byte 写，允许东西两方向上跨 Superlane 提供和接收数据；



数据交换单元 SXM (x2)

- 对 Tensor 数据进行 Reshape（旋转 or 转置），以更好地适应脉冲的下一计算单元。
- 唯一可以在 Superlane 间通信的单元，每单元有两组通道可将数据向上或向下移动到相邻 Superlane。



数据交换单元 SXM (x2)

- 唯一可以在 Superlane 间通信的单元，每单元有两组通道可将数据向上或向下移动到相邻 Superlane。

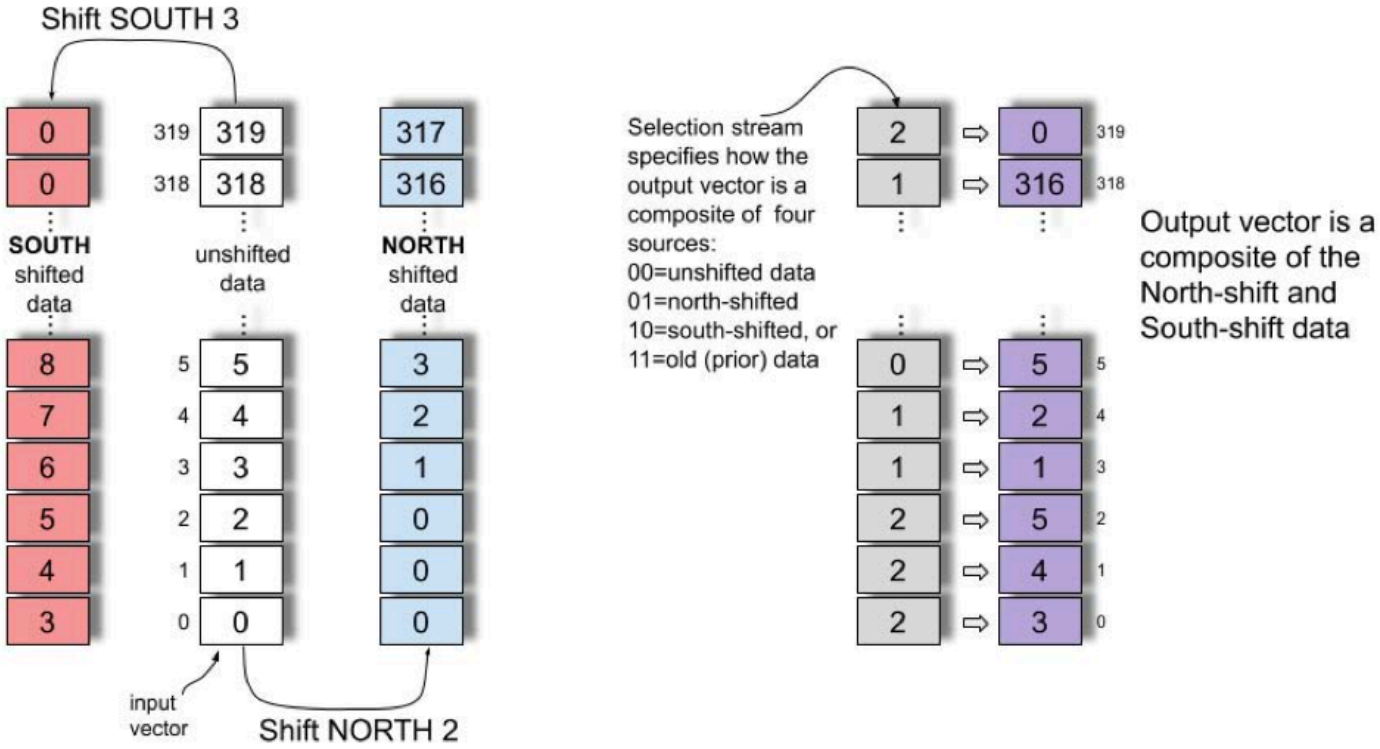
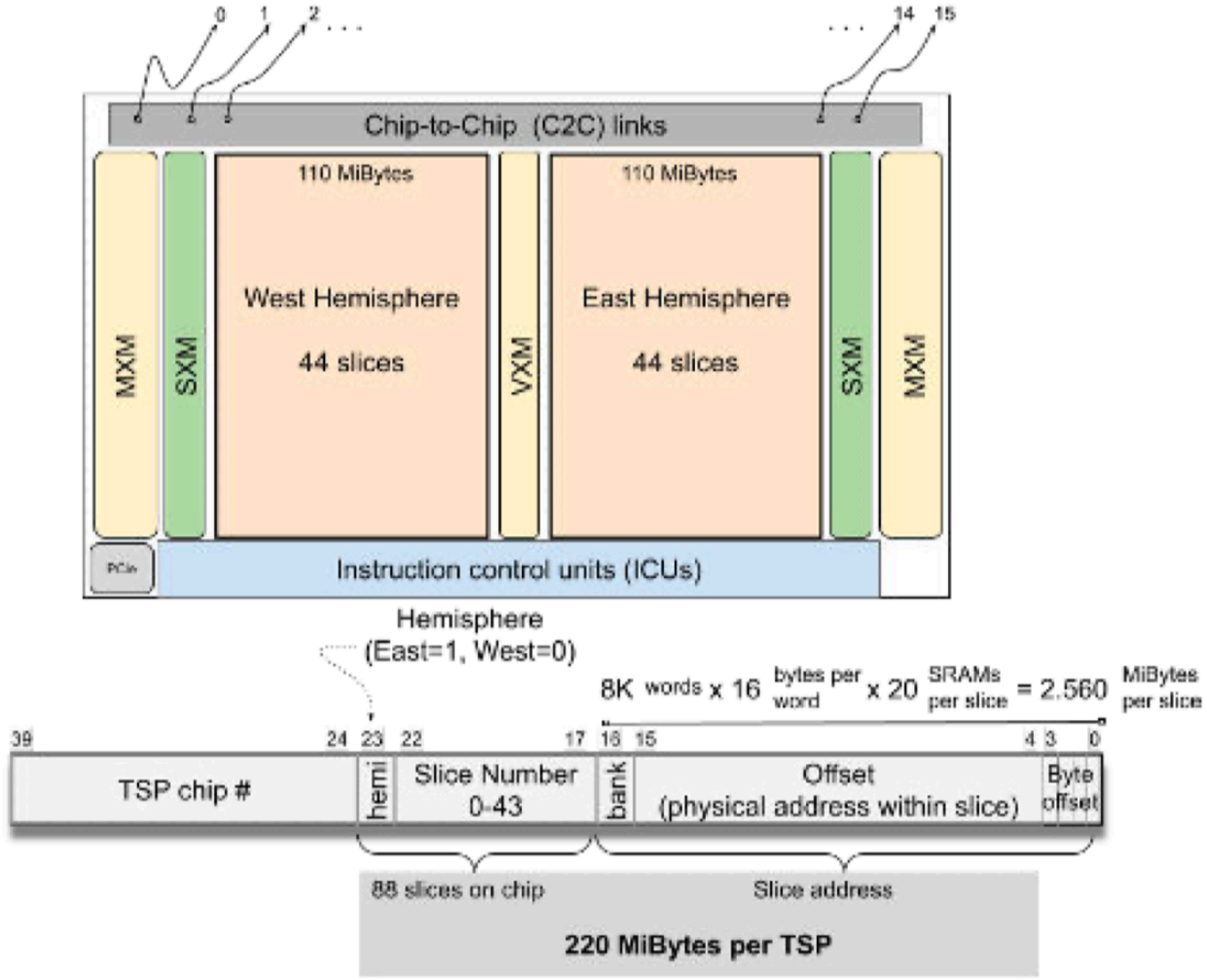


Fig. 8. The SXM provides a combination of North (up) and South (down) shifting operations which can be combined using the `select`.

存储结构

- 内存层次结构按 Rank-5 Tensor 来寻址 → [Device, Hemisphere, Slice, Bank, Address Offset]，形状为 [N， 2， 44， 2， 4096]；

- MEM 还需存储 VLIW 指令，宽度 2,304 Byte（144x16），取指占 ~10% SRAM 总带宽。



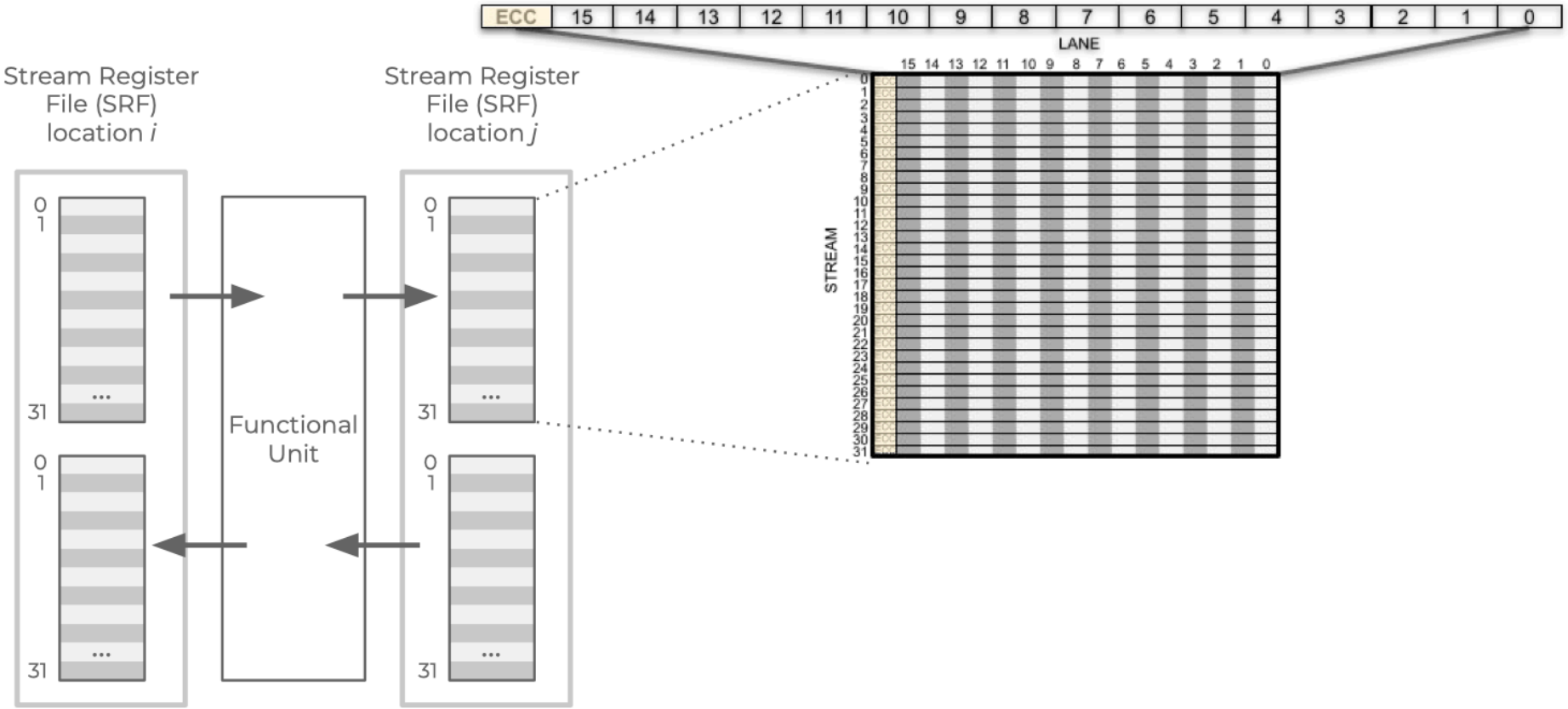
流寄存器 Stream Register

- 流寄存器位于功能单元间，通过编号来标识在 Superlane 内位置。

Fig. 4. Stream registers are numbered to show their locations between the functional slices within a superlane.

流寄存器 Stream Register

- SRF 用来保存操作数和结果，45 个 SRF 横跨 Superlane，用于片上数据传输。ECC进行保护，提高稳定性。



功能单元支持指令

Function	Instruction	Description
ICU	`NOP` N	No-operation, can be repeated N times to delay by N cycles
	`Ifetch`	Fetch instructions from streams or local memory
	`Sync`	Parks at the head of the instruction dispatch queue to await barrier notification
	`Notify`	Releases the pending barrier operations causing instruction flow to resume
	`Config`	Configure low-power mode
	`Repeat` n, d	Repeat the previous instruction n times, with d cycles between iterations
MEM	`Read` a,s	Load vector at address a onto stream s
	`Write` a,s	Store stream s register contents into main memory address a
	`Gather` s, *map*	Indirectly read addresses pointed to by *map* putting onto stream s
	`Scatter` s, *map*	Indirectly store stream s into address in the *map* stream
VXM	unary operation	z = *op* x point-wise operation on 1 operand, x, producing 1 result, z (eg. `mask`, `negate`)
	binary operation	z = x *op* y point-wise operations with 2 operands x and y producing 1 result, z (e.g. `add`, `mul`, `sub`)
	type conversions	Converting fixed point to floating point, and vice versa
	`ReLU`	Rectified linear unit activation function `max`(0,x)
	`TanH`	Hyperbolic tangent - activation function
	`Exp`	exponentiation e^x
	`RSqrt`	Reciprocal square root
MXM	`LW`	Load weights (LW) from streams to weight buffer
	`IW`	Install weights (IW) from streams or LW buffer into the 320×320 array
	`ABC`	Activation buffer control (ABC) to initiate and coordinate arriving activations
	`ACC`	Accumulate (ACC) either INT32 or FP32 result from MXM
SXM	`Shift` *up/down* N	Lane-shift streams up/down by N lanes, and `Select` between North/South shifted vectors
	`Permute` *map*	Bijective permute 320 inputs $\xrightarrow{map}$ outputs
	`Distribute` *map*	Rearrange or replicate data within a superlane (16 lanes)
	`Rotate` *stream*	Rotate $n \times n$ input data to generate n^2 output streams with all possible rotations (n=3 or n=4)
	`Transpose` *sg16*	Transpose 16×16 elements producing 16 output streams with rows and columns interchanged
C2C	`Deskew`	Manage skew across plesiochronous links
	`Send`	Send a 320-byte vector
	`Receive`	Receive a 320-byte vector, emplacing it in main memory

3.3 Groq 技术架构

系统集群组网

集群系统扩展 TSP → LPU

- 从单 GroqChip 扩展到 GroqNode 和 GroqRack 时，RealScale 支持的性能和延迟：

Scale (# of GroqChip processors)	**Peak INT8/FP16 Performance**	**System SRAM** (GBytes)	**Number of Dimensions**	**Diameter** (hops)	**End-to-end Latency** (µs)
1	750 TeraOps 189 TeraFlops	0.2	N/A	N/A	N/A
8 (1 GroqNode)	6 PetaOps 1.5 PetaFlops	1.76	0 (1 node)	1	0.6
16	12 PetaOps 3 PetaFlops	3.5	1 (2 nodes)	2	1.2
64 (1 GroqRack)	48 PetaOps 12 PetaFlops	14	1 (8 nodes)	3	1.8

Figure 2: Performance and latency enabled by RealScale as users scale from a single GroqChip to a GroqNode, and a GroqRack.

确定性网络

- 确定性网络 → 编译器需要考虑每个物理链路上信道带宽和延迟，TSP 中需要显式调度：
  - Tensor 数据在网络拓扑中逐跳流动，使用 TSP Local SRAM 作为缓冲；
  - 网络拓扑和传输路径是确定性，接收 TSP 可以立即将 Tensor 发送到明确的下一跳；
  - TSP 都使用 producer-consumer stream programming model 进行编程，允许将功能单元链接在一起；
  - TSP 功能单元被组织为 320 个元素 SIMD 指令执行，由 20 个 tile 组成，每个 tile 对数据执行 16 路 SIMD 计算。

确定性网络

- 单个 TSP 指令执行是确定性，TSP 间 C2C 可能会引入非确定性：
  - 编译器需要准确估计链路延迟 & 可能得变化
  - TSP 之间缺乏全局同步时钟 → 需要全局时钟
  - 网络间时钟漂移 → 补偿时钟漂移

确定性网络

- TSP 集群中需要软硬协同来确保多个 TSP 间 C2C 同步通信：
  - **通信正确性**：集群共享全局 SRAM，抛弃互斥锁的原子性 → 使用带时间戳发送/接收指令进行通信；
  - **公开状态**：TSP 软硬件接口公开 SRAM 和 SRF 状态，runtime 时模型计算图在集群中显式感知；
  - **通过指令调度**：静态计算图表示为 DAG，使用 ISA 指令显式地对通信流调度；

Name	Description
HAC	hardware aligned counter
SAC	software aligned couner
SYNC	intra-chip pause instruction
NOTIFY	intra-chip global signal to functional units to restart execution
DESKEW	pause instruction until HAC overflows
TRANSMIT	instruction to send notification to child through C2C link
RUNTIME_DESKEW t	delay TSP for $t \pm \delta_t$

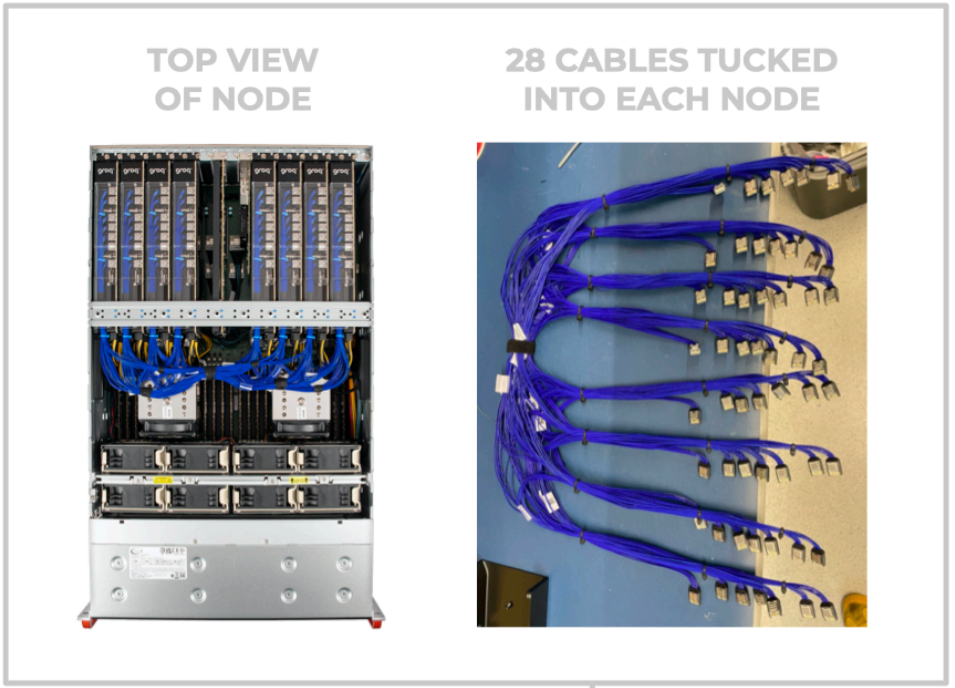
ISA support for a deterministic scale-out system.

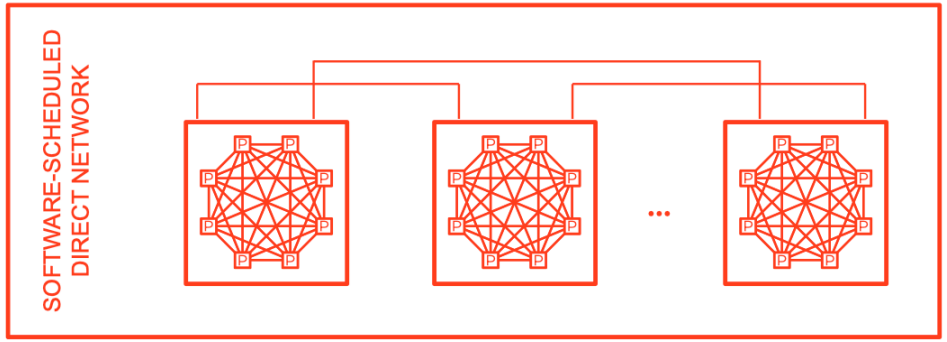
确定性网络

- TSP 集群建立和保持同步机制：
  - TSP 硬件对齐计数器，每 256 个 Cycle 交换状态保持全局时间；
  - 初始对齐程序，利用链接确保每个 TSP 同时开始执行指令；
  - 运行时重同步，补偿长时间计算单 TSP 时钟漂移；

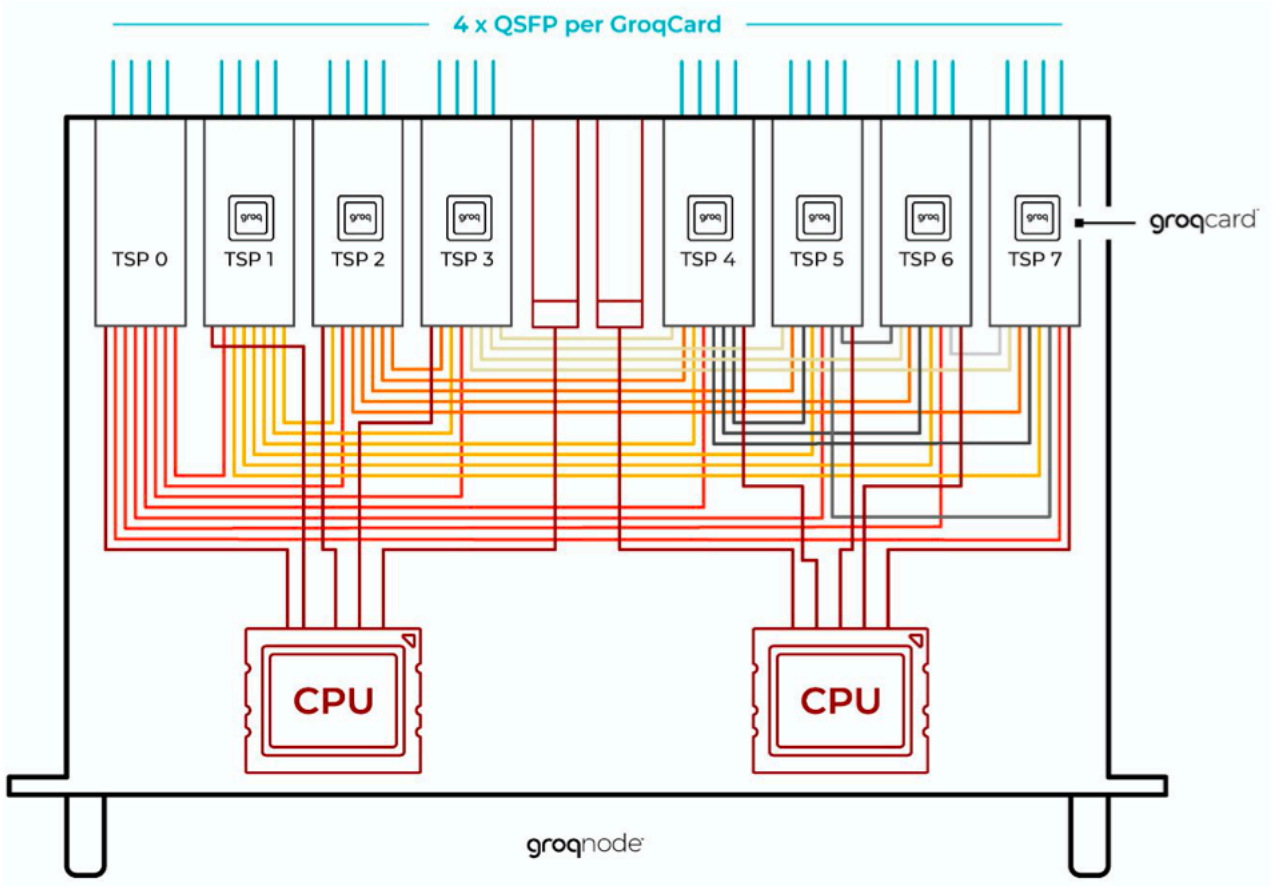
GroqNode

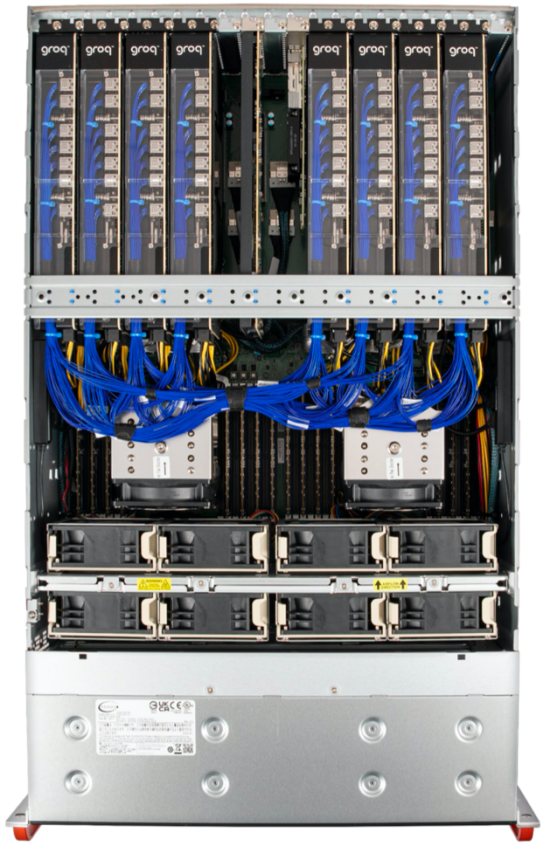
- GroqNode 由8个 TSP 全连接 Full Mesh；
- TSP 互联带宽分为 7 个本地链路 & 4 个全局链路；
- 芯片对芯片（C2C）链路和流量控制；





GroqNode

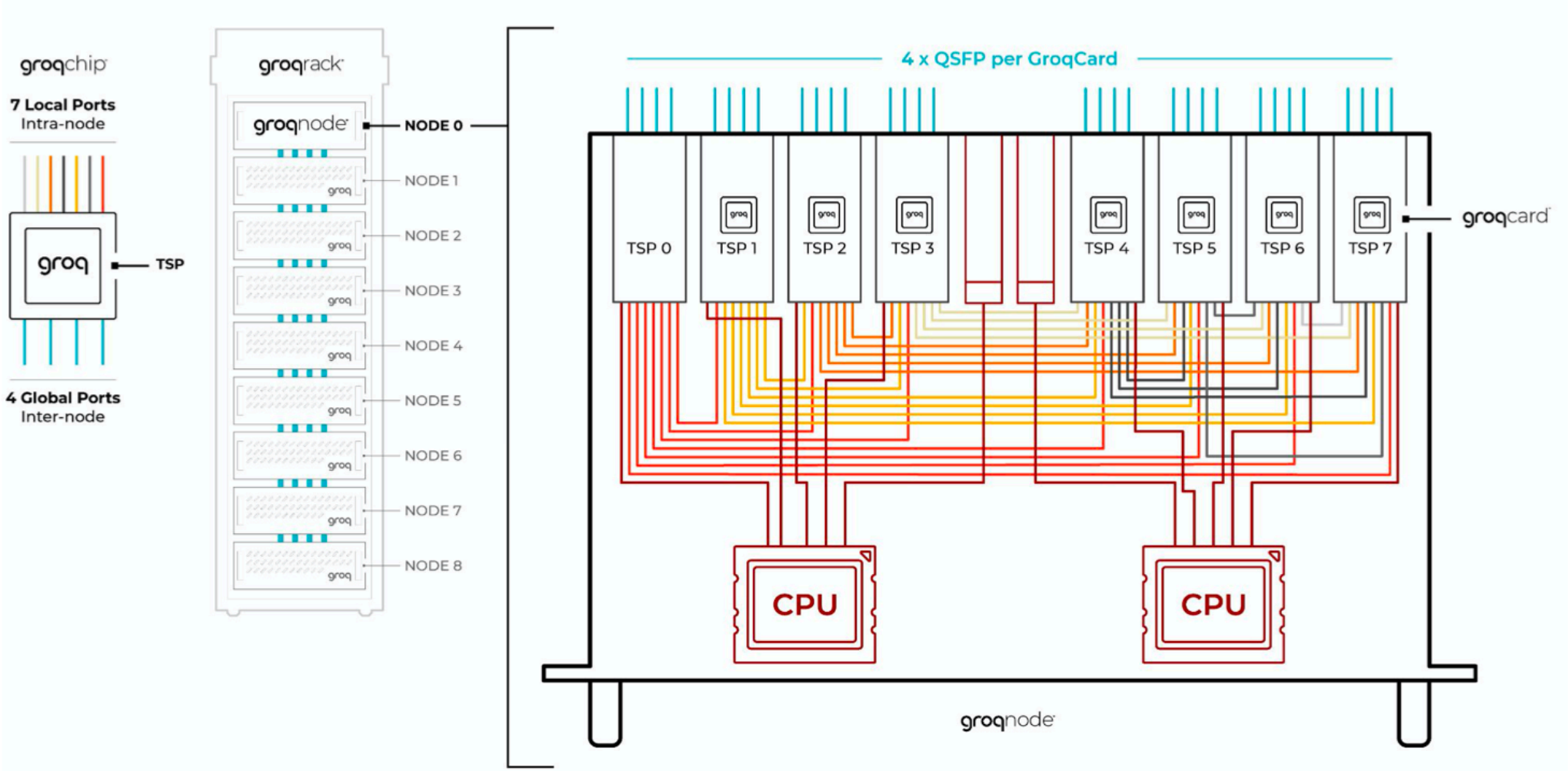

4 x QSFP per GroqCard
TSP 0
TSP 1
TSP 2
TSP 3
TSP 4
TSP 5
TSP 6
TSP 7
groqcard
CPU
CPU
groqnode



GroqNode → GroqRack

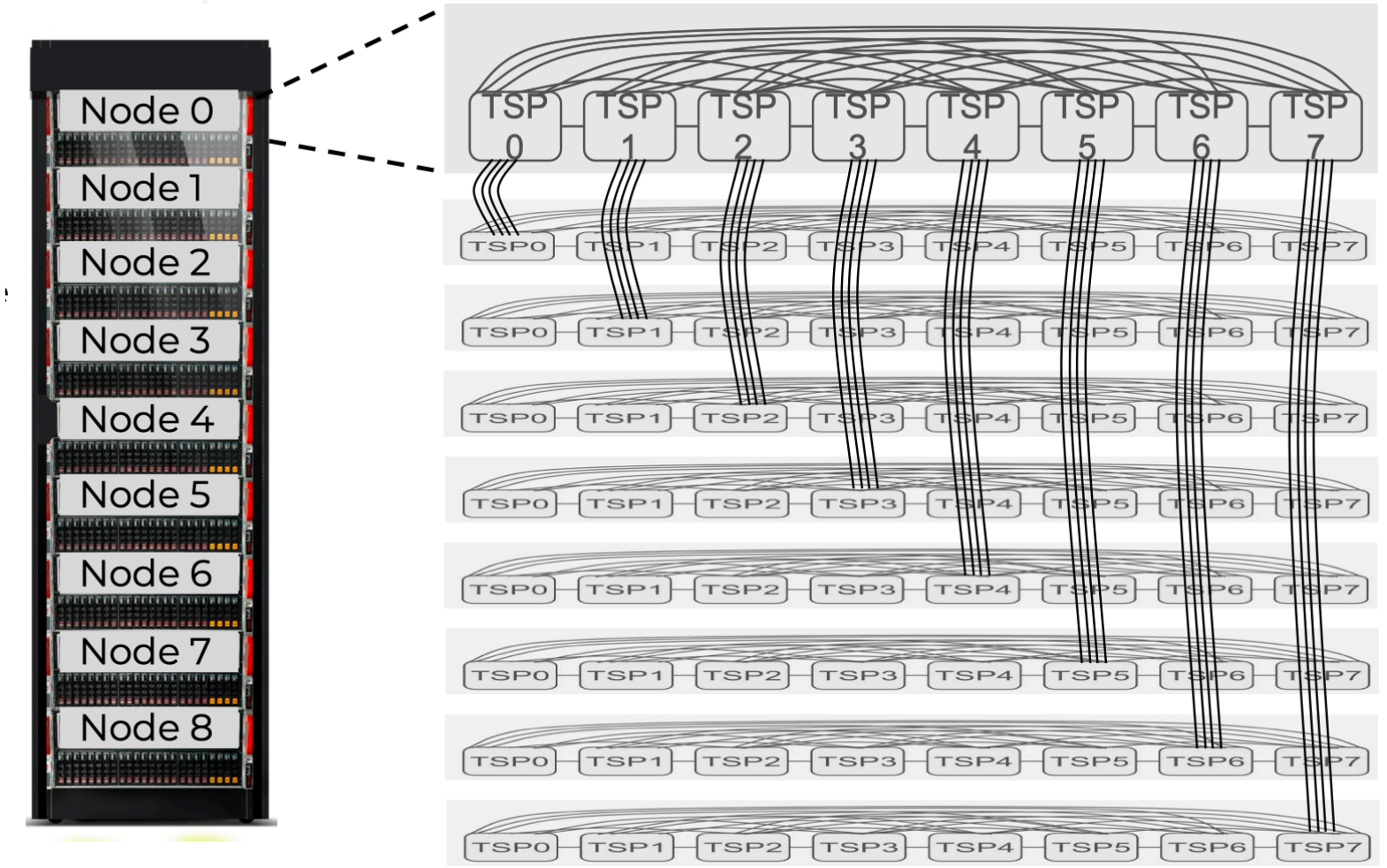
- GroqNode 节点内每个 TSP 所有全局链路可组合创建一个 GroqRack。同时，GroqNode 节点作为 32 个端口虚拟路由器，可作为横向扩展基本单位。
- GroqRack 架构中，每个 GroqNode 节点都连接到集群中其他节点，单个 GroqRack 使用 Dragonfly 拓扑，任何节点都可以当作备用节点。

GroqRack


groqchip
7 Local Ports
Intra-node
groq
TSP
4 Global Ports
Inter-node
groqrack
groqnode
NODE 0
NODE 1
NODE 2
NODE 3
NODE 4
NODE 5
NODE 6
NODE 7
NODE 8
4 x QSFP per GroqCard
TSP 0
TSP 1
TSP 2
TSP 3
TSP 4
TSP 5
TSP 6
TSP 7
groqcard
CPU
CPU
groqnode

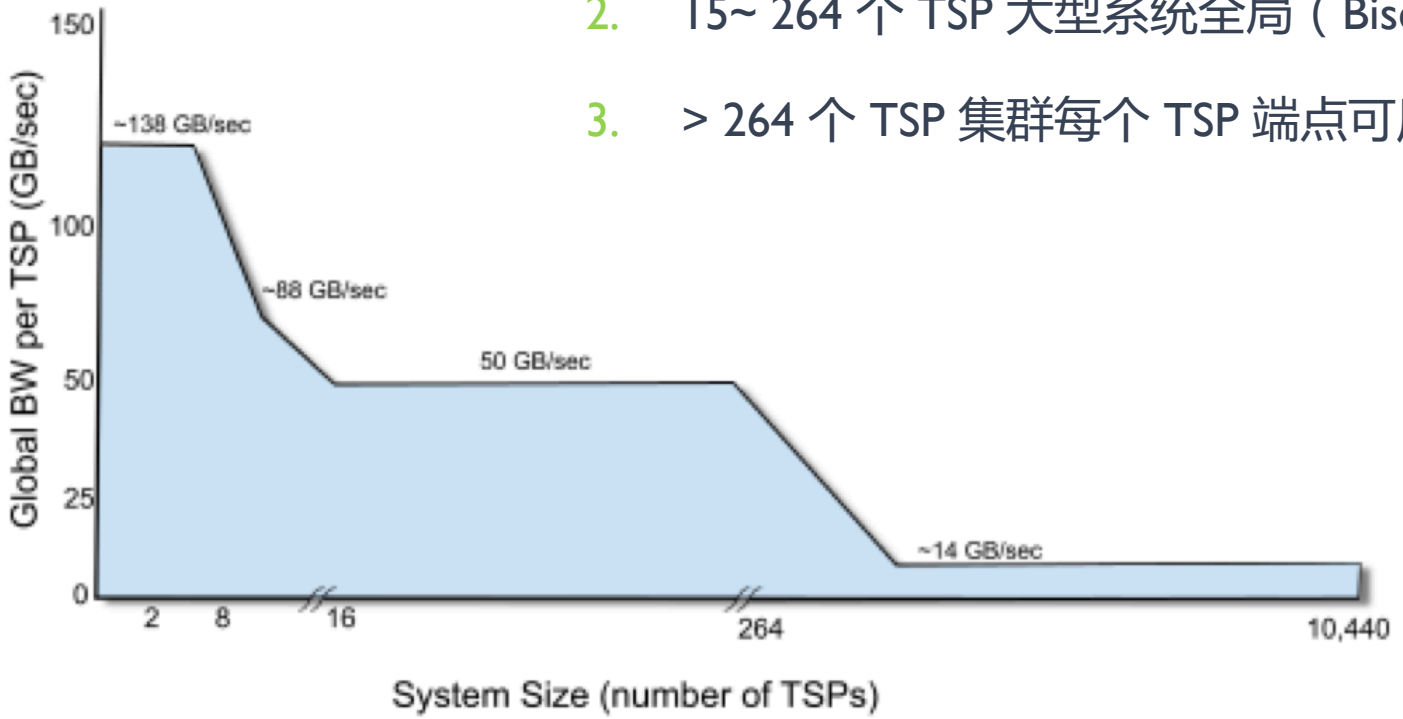
GroqRack

- GroqRack 由 9 个节点组成，每节点 8个 TSP，通过每个 TSP 4 个全局链路相互连接，共 32 × 9 = 288 个全局链路端口。
- 可将机架用作本地分组，以使用144 个端口将每个机架 9 个节点进行双重连接。其余 144 个端口用于连接到系统中其他机架。
- 最大配置集群中最多提供 145 个机架，共 10,440(145×72) 个 TSP，使用最小路由且最多 5-hop（源机架两个，一个全局跃点， 两个目标机架中）。



Scale out 带宽性能

1. <16 个 TSP 小型系统可利用节点内链路提供高带宽通信；
2. 15~ 264 个 TSP 大型系统全局（Bisection）带宽约 50 GB/s；
3. > 264 个 TSP 集群每个 TSP 端点可用全局带宽约为 14 GB/秒；

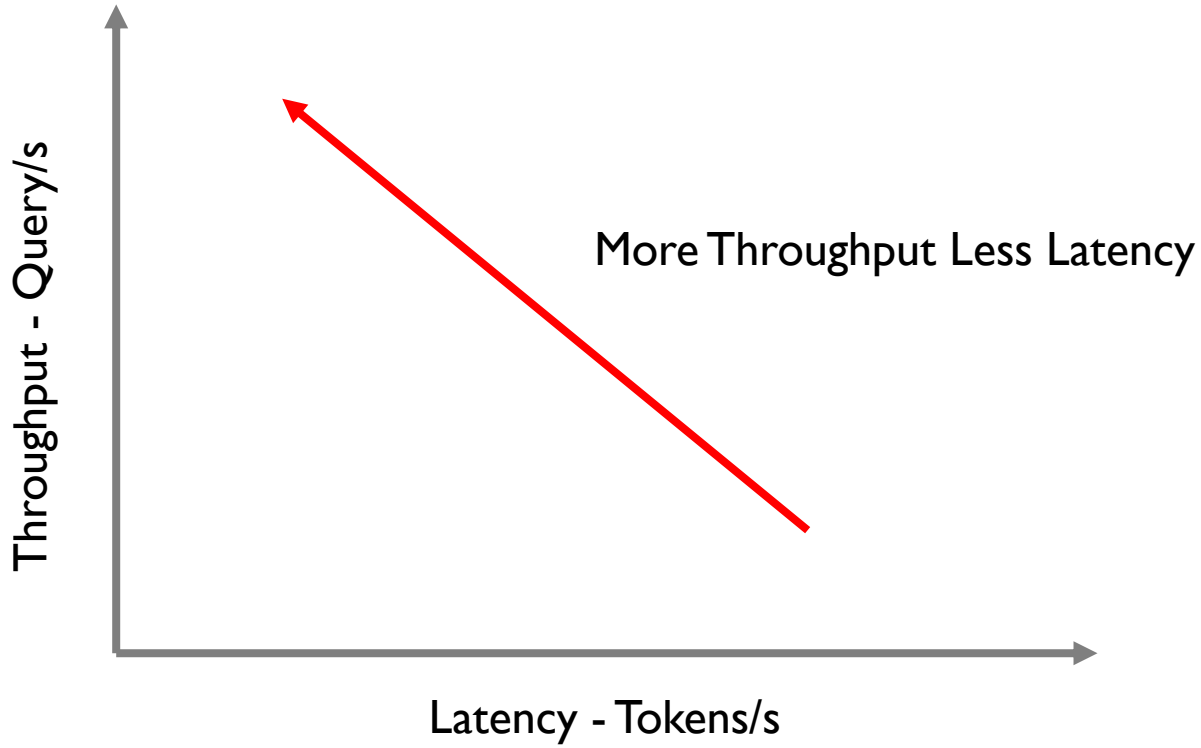


3.4 Groq 技术架构

性能吊打NV总结

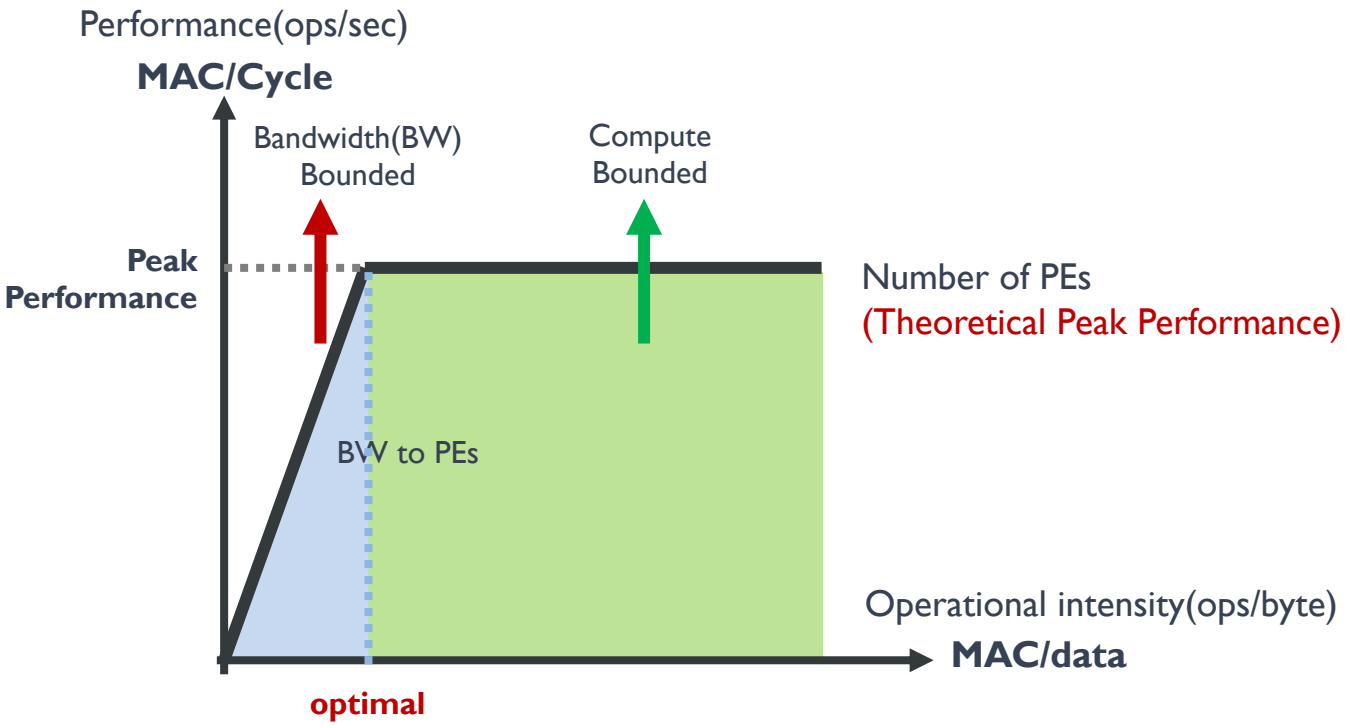
吞吐量 - 时延性能曲线

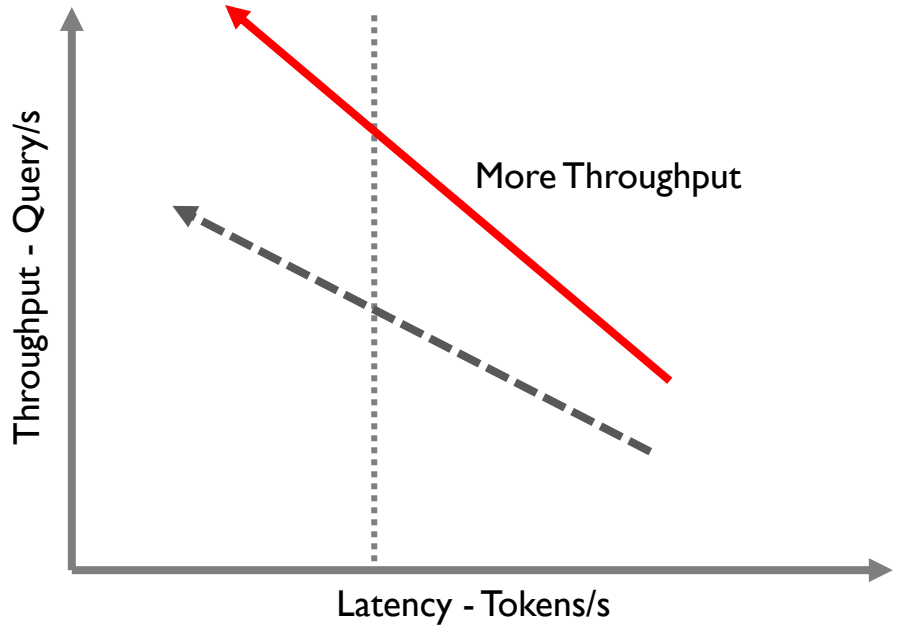
- 真正影响大模型推理的指标：吞吐 vs 时延；
- 推理的目标：更好的吞吐和更低的时延；



吞吐量 - 时延性能曲线

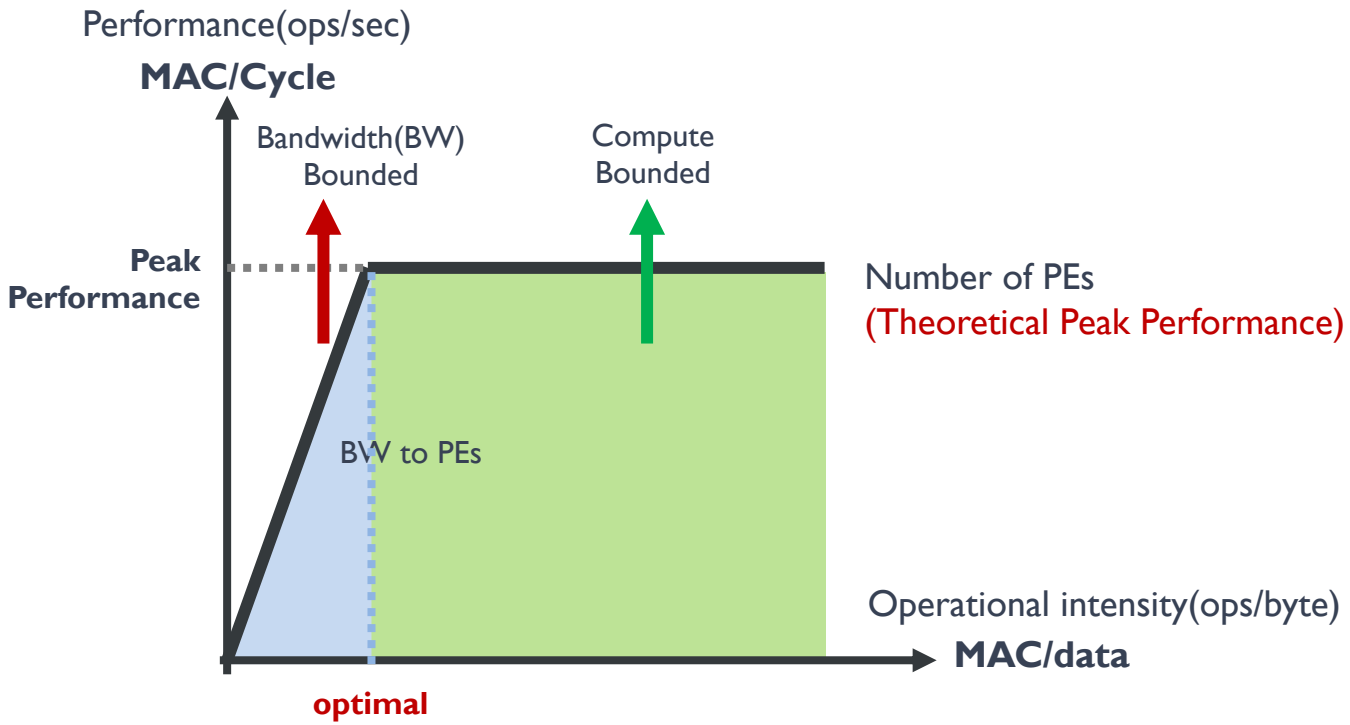
固定硬件 & 固定时延：随着 Batch Size 增加，系统从带宽受限区域 Bandwidth Bound 转变为计算受限区域 Compute Bound。需要增加峰值算力，吞吐-时延曲线会向左上方移动。

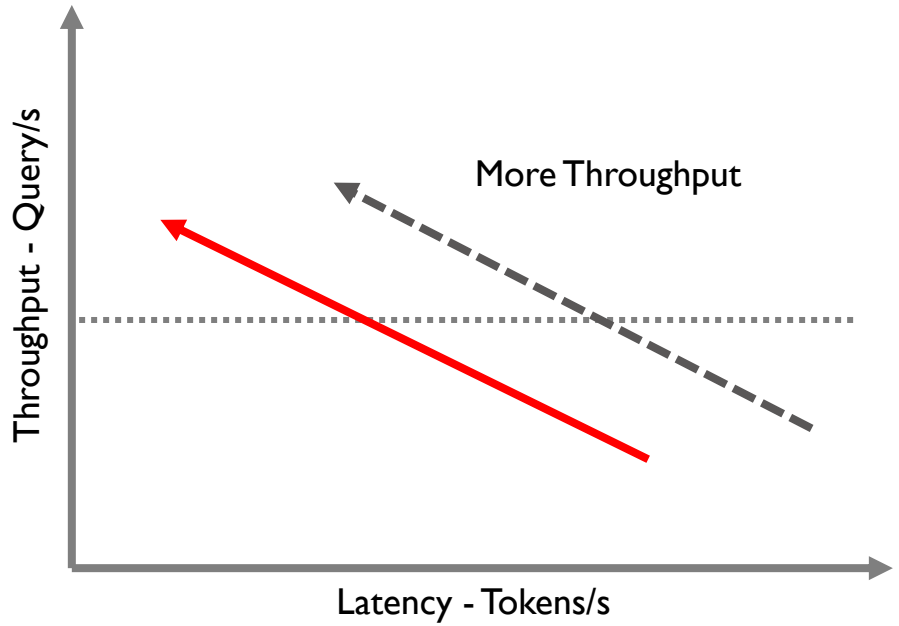




吞吐量 - 时延性能曲线

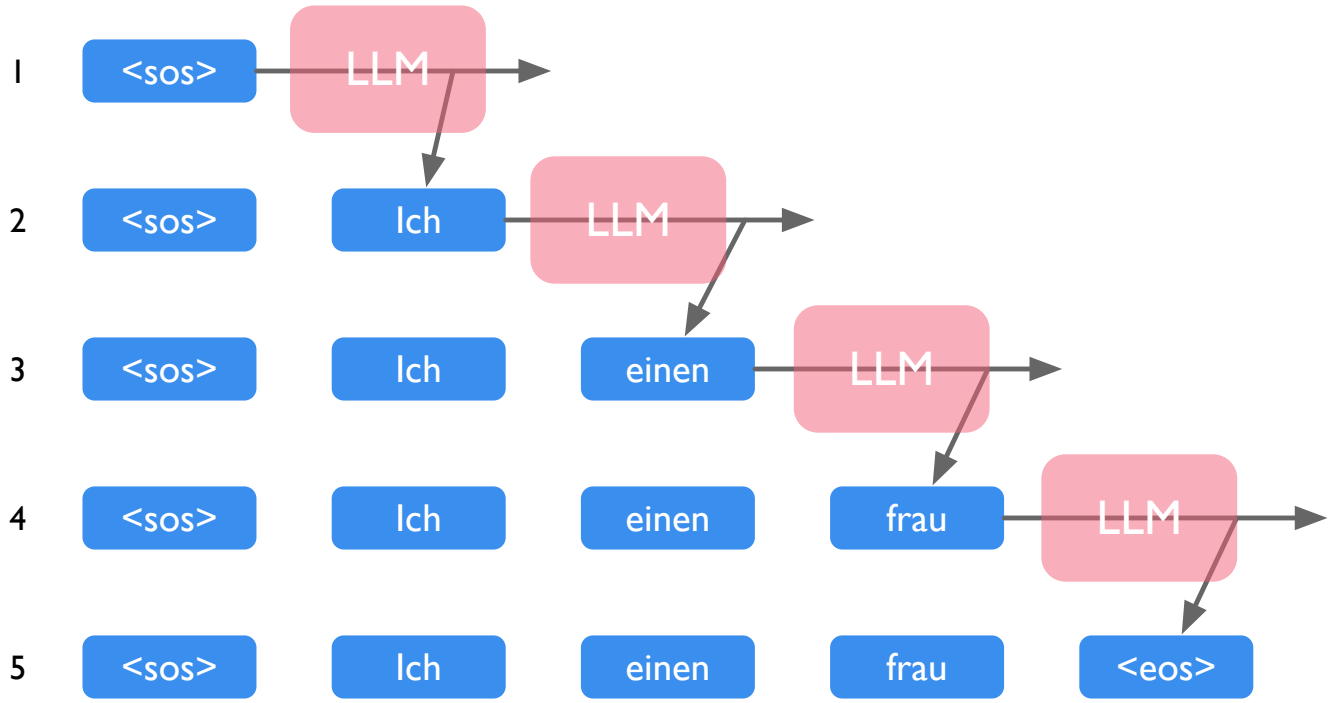
固定硬件 & 固定吞吐：模型 Batch Size 不变（ 1/2/4 较小的值），因为 LLM 推理过程采用自回归方式一次生成一个，循环输入到模型中，因此对带宽要求极高，推理服务处于带宽受限区域 Bandwidth Bound。需要增加片内带宽，吞吐-时延曲线会向左平移。





Groq 成功的关键

- Groq 这么挫的AI 芯片，推理时延能比 NV 超过10X？取消 L3/L2/L1 Cache → 所有数据放 SRAM



	Grq	H100
制造工艺	14nm	4nm
单芯片算力 FP16 TFLOPs	188	989
节点内带宽 GB/s	RealScale：800	NVLinks：900
单节点内存	1.76GB	640GB
80GB 推理所需节点数	45	0.125
80GB 推理所需卡数	347	1
访存带宽	80TB/s	3.35TB/s

技术总结 Pons

- 取消 L3/L2/L1 Cache → 所有数据放 SRAM，实现内存高带宽，进而提高单请求时延 Token/s：
  - 8xH100 难达 500 token/s 输出速度，对输出速度高要求场景，以 Groq 为代表空间计算很有前景
- Groq 芯片做到真正意义 Software-Defined Scale-out Tensor Streaming Multi-Processor：
  - 软件能够1 代打 3 代，2020年芯片2024年集群组网后大放光彩，比大部分DSA强
  - Groq 的架构为 CNN 网络设计而非 Transformer，通过编译实现“地表最快 Token 输出速度”
  - 编译器需要根据 NN 翻译成 [Device, Hemisphere, Slice, Bank, Address Offset] 工程量大
- 对于端侧产品受限于面积&功耗等因素，编译器精细设计数据流，DSA 架构是趋势：
  - 减少 DRAM 会对时延带来收益，端侧产品场景稳定对于 DSA 接受度够高，甚至越 DSA 化越好

技术总结 Cons

- 每片 SRAM 容量为 220MB，KV Cache 占用的内存很大，需要很大的集群规模：
  - 云端场景模型会遵循 scaling low 发展，缺乏外部 DRAM 存储 AI 加速器属于异端，场景不适综合征；
  - 模型趋势越来越大，除了片内需要足够存储空间才能提供足够数据给计算单元，组网的开销会占据成本近半；

思考：推理算力 & 市场预测

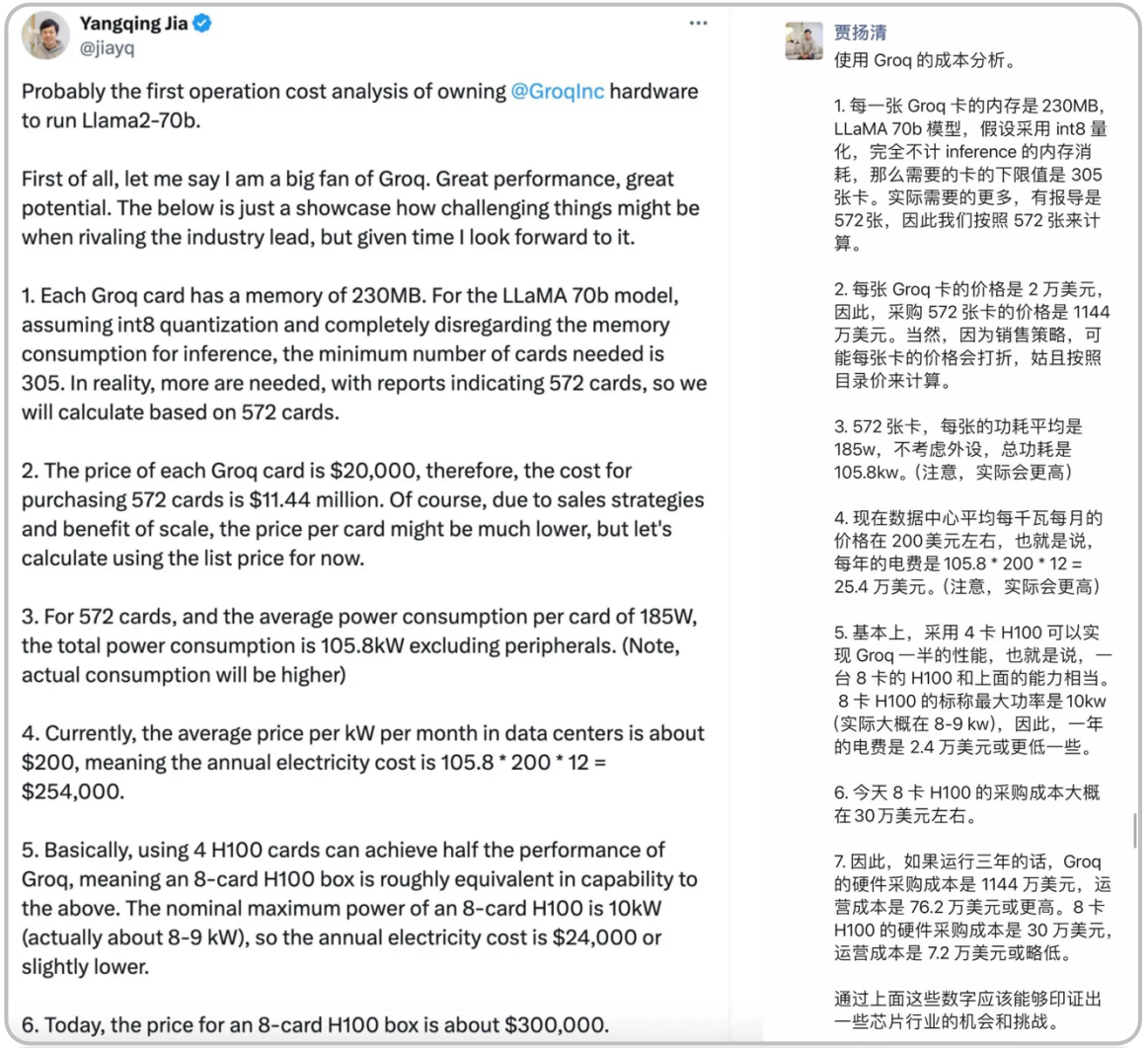
业界的通用看法

- https://www.zhihu.com/question/645010090/answer/3403413472

业界的通用看法

- **贾扬清**：运行三年的话，Groq 的硬件采购成本是 1133 万美元，运营成本是 76.2 万美元或者更高。8 卡 H100 的硬件采购成本是 30 万美元，运营成本是7.2 万美元或略低。

- —— 贵，太贵了，不划算，没有性价比和市场机会。



Yangqing Jia
@jiayq

Probably the first operation cost analysis of owning @GroqInc hardware to run Llama2-70b.

First of all, let me say I am a big fan of Groq. Great performance, great potential. The below is just a showcase how challenging things might be when rivaling the industry lead, but given time I look forward to it.

1. Each Groq card has a memory of 230MB. For the LLaMA 70b model, assuming int8 quantization and completely disregarding the memory consumption for inference, the minimum number of cards needed is 305. In reality, more are needed, with reports indicating 572 cards, so we will calculate based on 572 cards.

2. The price of each Groq card is $20,000, therefore, the cost for purchasing 572 cards is $11.44 million. Of course, due to sales strategies and benefit of scale, the price per card might be much lower, but let's calculate using the list price for now.

3. For 572 cards, and the average power consumption per card of 185W, the total power consumption is 105.8kW excluding peripherals. (Note, actual consumption will be higher)

4. Currently, the average price per kW per month in data centers is about $200, meaning the annual electricity cost is 105.8 * 200 * 12 = $254,000.

5. Basically, using 4 H100 cards can achieve half the performance of Groq, meaning an 8-card H100 box is roughly equivalent in capability to the above. The nominal maximum power of an 8-card H100 is 10kW (actually about 8-9 kW), so the annual electricity cost is $24,000 or slightly lower.

6. Today, the price for an 8-card H100 box is about $300,000.

贾扬清

使用 Groq 的成本分析。

1. 每一张 Groq 卡的内存是 230MB，LLaMA 70b 模型，假设采用 int8 量化，完全不计 inference 的内存消耗，那么需要的卡的下限值是 305 张卡。实际需要的更多，有报导是 572张，因此我们按照 572 张来计算。

2. 每张 Groq 卡的价格是 2 万美元，因此，采购 572 张卡的价格是 1144 万美元。当然，因为销售策略，可能每张卡的价格会打折，姑且按照目录价来计算。

3. 572 张卡，每张的功耗平均是 185w，不考虑外设，总功耗是 105.8kw。(注意，实际会更高)

4. 现在数据中心平均每千瓦每月的价格在 200美元左右，也就是说，每年的电费是 105.8 * 200 * 12 = 25.4 万美元。(注意，实际会更高)

5. 基本上，采用 4 卡 H100 可以实现 Groq 一半的性能，也就是说，一台 8 卡的 H100 和上面的能力相当。8 卡 H100 的标称最大功率是 10kw (实际大概在 8-9 kw)，因此，一年的电费是 2.4 万美元或更低一些。

6. 今天 8 卡 H100 的采购成本大概在30万美元左右。

7. 因此，如果运行三年的话，Groq 的硬件采购成本是 1144 万美元，运营成本是 76.2 万美元或更高。8 卡 H100 的硬件采购成本是 30 万美元，运营成本是 7.2 万美元或略低。

通过上面这些数字应该能够印证出一些芯片行业的机会和挑战。

Semianalysis 分析

- https://www.semianalysis.com/p/groq-inference-tokenomics-speed-but

Groq vs Nvidia BOM Cost (ignores margins, costs for networking, CPUs, racks, power, etc.)			
	Groq LPU System	Nvidia H100 Latency Optimized (SpecDec)	Nvidia H100 Throughput Optimized
Active Silicon & Package Cost	$650	$3,200	$3,200
PCIe/SXM Module Costs	$400	$500	$500
Card/Module Cost	**$1,050**	**$3,700**	**$3,700**
Number of Chips per Inference Unit	576	8	2
Total Chip Modules BOM	**$604,800**	**$29,600**	**$7,400**
Tokens Per Second Per User	**500**	**420**	**30**
Batch Size	3	2	64
Pipeline Parallelism	16	1	1
Concurrent Users	48	2	64
Tokens Per Second Per Inference Unit	**24,000**	**840**	**1,920**
Tokens Per Second Per Chip	42	105	960
$ BOM Per Tokens Per Second	**$25.20**	**$35.24**	**$3.85**

Note: The more sophisticated analysis with all costs accounted for is below in the article, this is a simplified analysis.

Semianalysis 分析

GPU Server Colocation Simple Total Cost Of Ownership (TCO)					
	Unit	Bull Case GPU Server	Realistic Colo Cost	Realistic Colo+Capital Cost	Realistic Depreciation Colo+Capital Cost
Server Capital Costs					
Upfront Server Capex	USD	$350,000	$350,000	$350,000	$350,000
Useful Life in Years	Years	6	6	6	4
Cost of Capital	%	13.0%	13.0%	18.0%	18.0%
Total Server Capital Costs per Month	**USD/mth**	**$7,025.9**	**$7,025.9**	**$7,982.7**	**$10,281.2**
Server Hosting Costs					
Electricity Cost	USD/kWh	$0.087	$0.087	$0.087	$0.087
Utilization Rate	%	80.0%	80.0%	80.0%	80.0%
Power Usage Effectiveness (PUE)	Ratio	1.25	1.25	1.25	1.25
Electricity Cost per kW per mth	USD/kW/mth	$63.5	$63.5	$63.5	$63.5
Colocation Cost (i.e. rack space rental cost, infra, install,	USD/kW/mth	$120.0	$190.0	$190.0	$190.0
Total Hosting Cost per kW per Month	**USD/kW/mth**	**$183.5**	**$253.5**	**$253.5**	**$253.5**
Server Power Consumption	W	10,200	10,200	10,200	10,200
Total Server Hosting Costs per Month	**USD/mth**	**$1,871.8**	**$2,585.8**	**$2,585.8**	**$2,585.8**
Comprehensive Server Cost per Month	**USD/mth**	**$8,897.7**	**$9,611.7**	**$10,568.5**	**$12,867.1**
Number of vCPU/GPUs per Server	vCPU/GPUs	8	8	8	8
Capital Cost per Unit, per Hour	USD/hr	$1.203	$1.203	$1.367	$1.760
Hosting Cost per Unit, per Hour	USD/hr	$0.321	$0.443	$0.443	$0.443
Rental Cost per Unit per Hour	**USD/hr**	**$1.524**	**$1.646**	**$1.810**	**$2.203**
% Cost in Hosting vs Capex	%	26.6%	36.8%	32.4%	25.2%

W = Watts. kW = Kilowatts. kWh = Kilowatt-hours.

semianalysis

成本核算？

1. GPU 成本30~40% 为 HBM， Groq 按 SRAM 成本估算预计单卡 1200$。500 x Groq 总成本 60 万美金，60 万美金只能采购两台 NVIDIA H100。两台 NVIDIA H100 能跑出 500 tokens/s ？

业界的通用看法

- mackler：LLM 很不一样，直接摧毁了上面的美好故事的基本逻辑，不只是 scale 的问题，还有 KV-Cache 热数据。XXX，最大的劣势是 SRAM 容量实在太小了，成本也太高了。

- —— 不妥，太不妥了，不符合 LLM 大语言模型场景。

市场策略的问题

1. AI手机场景要求 2ms/token 推理时延 → H100 大吞吐低时延，只能使用 Groq；
2. 场景对推理时延要求不严格（5 Tokens/s），H100 和 Groq 都满足，需要考虑性价比；

- 推理聚焦 C 端用户已时延 Latency 优先， Groq 优势巨大；不以 时延Latency 重点则 XXX；
- 目前以时延 Latency 优先的应用还没有起来，需要等待大模型真正商业化落地；

用户场景

1. 推理服务提供商提供自有芯片，芯片成本则是推理硬件成本。购买 NV 提供推理服务，其售价则是成本价，当卖铲人还是铲沙者？

思考

- 不能看GPU H100和Groq的直接对比，因为两者不等的。

	Groq	H100
制造工艺	14nm	4nm
芯片尺寸	25*29 mm	814*814 mm
Wafer 倍数	900	1
访存速度	80TB/s	3.35 TB/s
latency 时延	10	1
HMB	None	96G
场景	推理	训练 + 推理

Groq可以打3折以上成本，SRAM 比 STDcell shrink 系数更小。4nm SRAM 密度比 14nm 高，4nm工艺情况下不需要572卡。

1张wafer大约6000美元成本。1颗不到100美元。

Groq主打的是快，不完全是便宜（虽然也比Amazon Bedrock的API定价便宜了60% ）

Groq 技术发展的问题

- **软件生态：**2020 年投产流片的芯片，2023/24 适配 LLM 用 2/3年构建软件体系，AI 系统配套；
- **功耗和散热：**成为核心成本，如比特币矿机工厂相似，3-10年电力成本将会超过芯片成本；

对 Groq 洞察总结

1. 短期内不看好 Groq 此类的 DSA，长期大模型应用落地特别是 AI-Agent 是个很好的技术方向；
2. 主要矛盾并不是 LLM 的技术和 Tokens/s 成本，而是应用短期无法落地；

对产业的思考与洞察

1. **展望 AI Agent**：LLM 未来通过 AI-Agent 落地（具身智能），如 Groq 的 ASIC 将会成为 LLM 主流芯片方案，单应用成本以 10X 级别下降；
2. **下一代芯片**：执行 LLM 最大问题内存不够，未来与 DDR 一起封装而非 HBM，在极低时延下进一步提升芯片吞吐，能够解决目前 LLM 的大部分技术问题；
3. **服务提供商**：自产芯片 + 推理服务的商业模式看来是比较具有经济价值（成本和消耗），所以推测百度 + 昆仑芯、腾讯 + 燧原、阿里 + 含光 对于新的技术架构蠢蠢欲动。

Thank you

把AI系统带入每个开发者、每个家庭、每个组织，构建万物互联的智能世界

Bring AI System to every person, home and organization for a fully connected, intelligent world.

Copyright © 2023 XXX Technologies Co., Ltd.
All Rights Reserved.

The information in this document may contain predictive statements including, without limitation, statements regarding the future financial and operating results, future product portfolio, new technology, etc. There are a number of factors that could cause actual results and developments to differ materially from those expressed or implied in the predictive statements. Therefore, such information is provided for reference purpose only and constitutes neither an offer nor an acceptance. XXX may change the information at any time without notice.

ZOMI

Course chenzomi12.github.io

GitHub github.com/chenzomi12/DeepLearningSystem