# 可重构芯片加速超大AI模型训练与推理

于义

清微智能
TSING MICRO
北京清微智能科技有限公司

# 主要内容

❶ **清微智能简介**

❷ CGRA基本原理

❸ 清微大算力可重构芯片介绍

# 公司简介——十六年技术积累

2006

2014

2015

国家技术发明
二等奖


2017

国际低功耗设计
竞赛冠军


2018

清微智能
公司成立

2020

中国电子学会
技术发明一等奖

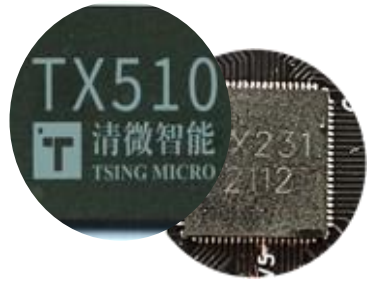

2021

中国唯一通过谷歌认
证的智能芯片

Google

2022

芯片销售超千万颗，客户含海康，
国网，商汤，腾讯，华为等



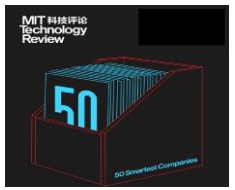

北京未来芯片技术高精尖创新中心
BEIJING INNOVATION CENTER FOR FUTURE CHIPS

获北京未来芯片高精
尖创新中心支持


《麻省理工科技评论》
“皇冠级的成就”

Silicon 100

两次入围Silicon 100




中国IC独角兽
CHINA IC UNICORN

IMPACT
FORTUNE

《麻省理工科技评论》“50家聪明公司”
《财富》“最具影响力创业企业”
中国IC 独角兽企业

# 技术积累

## 总体情况

围绕可重构智能计算芯片在以 IEEE JSSC、TCAS-I 为代表的 SCI 期刊上共发表论文 126 篇、在顶级会议(ISSCC/VLSI/ISCA/HPCA/DAC)等共发表论文 114 篇，出版《人工智能芯片设计》专著一部，申请/授权发明专利 170/100 项、国际发明专利 21/4 项。候选人相关成果被全球 6 大洲 40 个国家的 300 多个研究机构 (包括 MIT、Stanford、UC Berkeley、Google、Intel 等) 跟踪引用，相关论文谷歌学术他引达 **3000** 余次

## 发表论文 240 篇

集成电路TOP 1期刊 **IEEE JSSC** 6 篇 引用次数达**190** 次

集成电路奥林匹克会议 **ISSCC** 3 篇 引用次数达 48次

集成电路奥林匹克会议 **VLSI** 6 篇 引用次数达90次

集成电路领域核心期刊 **IEEE TVLSI** 最受欢迎论文（2017.8 - Now) 引用次数达 **159** 次

人工智能领域期刊 **Neurocomputing**（2018）引用次数达 **173** 次

计算机体系结构TOP 1会议 **ISCA** 4 篇 引用次数达 72次

计算机体系结构顶级会议 **HPCA 21** 1 篇

## 专著《可重构计算》

我国本领域**第一部**

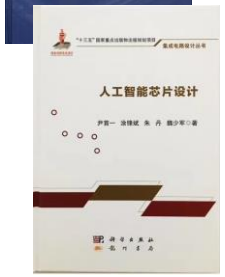
2014年7月/科学出版社


## 专著《人工智能芯片设计》

我国 **AI** 芯片领域

**第一部**系统性专著

2020年3月/科学出版社



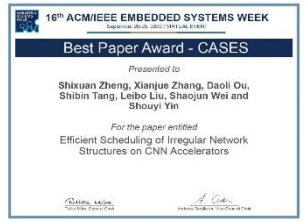

## 申请发明专利 170 项，已获授权 100 项

硬件架构：申请 63 项，授权 34
软件方法：申请 65 项，授权 39
系统应用：申请 52 项，授权 27


## 获奖及报道

2014年教育部技术发明一等奖
2015年国家技术发明二等奖
2015年中国发明专利金奖
2017年ACM/IEEE ISLPED Design Contest Award
2018 年 MIT Technology Review 专题报道
2019 年DAC 低功耗目标检测挑战赛全球亚军
2019年中国电子信息领域优秀科技论文
2019年 SRE 语音算法大赛全球前十，亚洲第二
2020 年中国电子学会科学 技术发明一等奖
2020 年16th ACM/IEEE 嵌入式系统 可重构编译 最佳论文奖
2021年中国电子信息领域优秀科技论文
2021年入选全球最值得关注新创半导体公司排行榜silicon 100







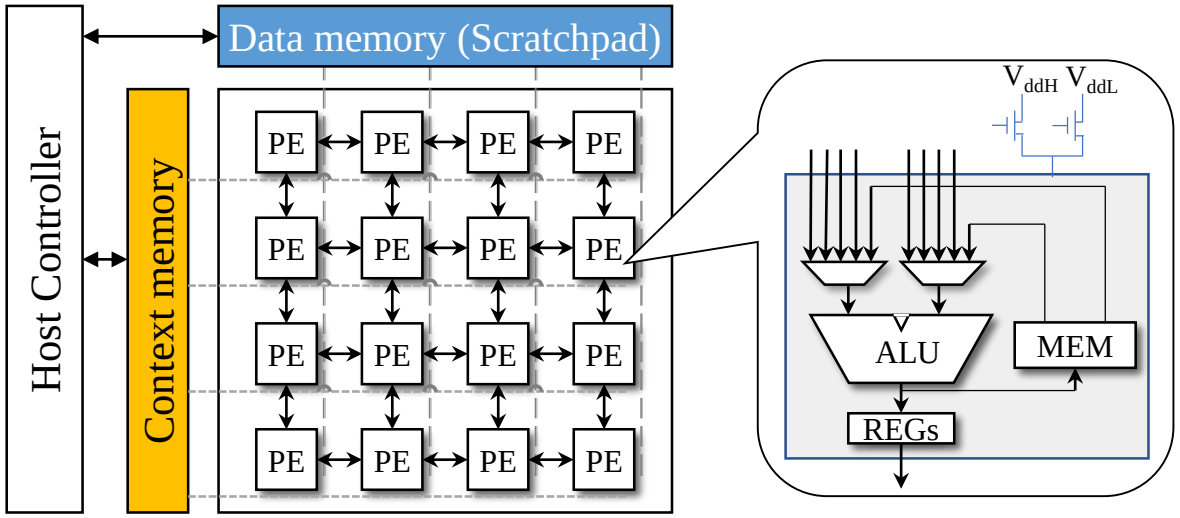

# 主要内容

❶ 清微智能简介

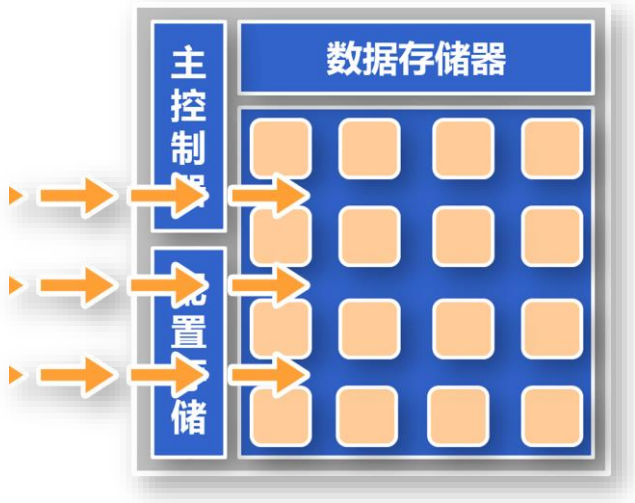
❷ **CGRA基本原理**

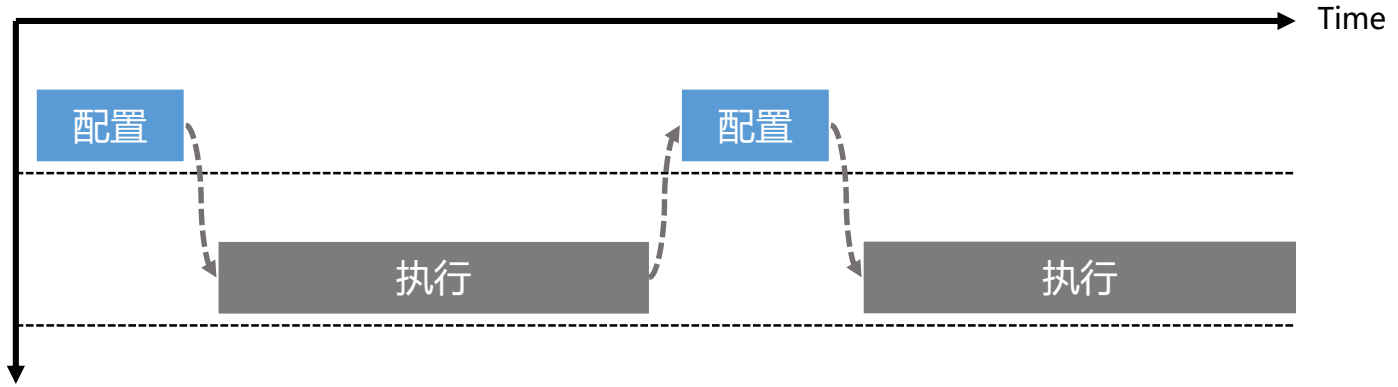
❸ 清微大算力可重构芯片介绍

# CGRA基本结构和配置机制




| | | | |
|---|---|---|---|
| A + B | A >> B | A > B | (A+B)>>C |
| A - B | A + (B>>C) | A == B | (A+B)<<C |
| A & B | A | A < B | (A-B)>>C |
| A \| B | A + (B<<C) | A >= B | (A+B)<<C |
| A ^ B | A - (B<<C) | A <= B | A×B_H |
| A ~^ B | \|A-B\| | A != B | Clip(A, -B, B) |
| ~A | (A>>C)-B | A - (B>>C) | Clip(A, 0, B) |
| A << B | A×B_L | (A<<C)-B | C?A:B |



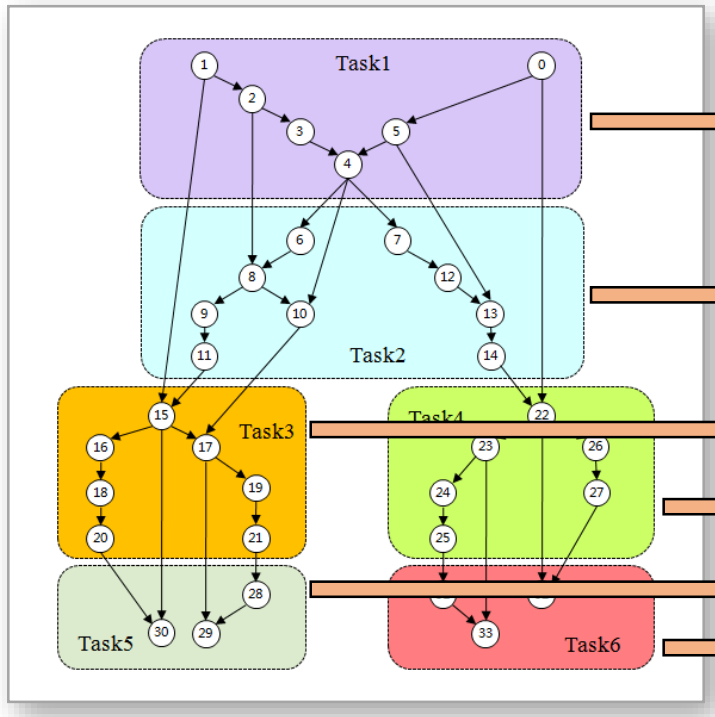



# 程序基础映射示例-空域执行，数据驱动

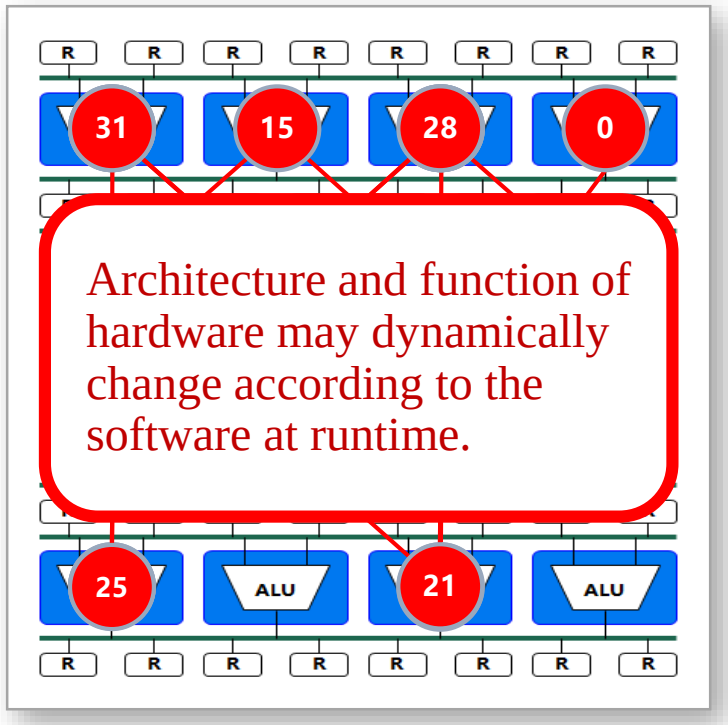
软件程序

```
IF (condition = 1) THEN
   r1 = b * b;
   r2 = a * c;
   r2 = r2 * 4;
   r1 = r1 - r2;
   IF (r1 < 0) THEN
      state = '1';
   ELSE
      state = '0';
      r3 = r1 * 4;
      r3 = r3 + 1;
      r3 = r3 * 0.2;
      FOR i = 1 to 3 LOOP
         r4 = r1 / r3;
         r3 = r3 + r4;
         r3 = r3 / 2;
      END LOOP;
      r1 = 0 - b;
      r2 = a + a;
      r4 = r1 + r3;
      x1 = r4 / r2;
      r5 = r1 - r3;
      x2 = r5 / r2;
   END IF;
END IF;
```

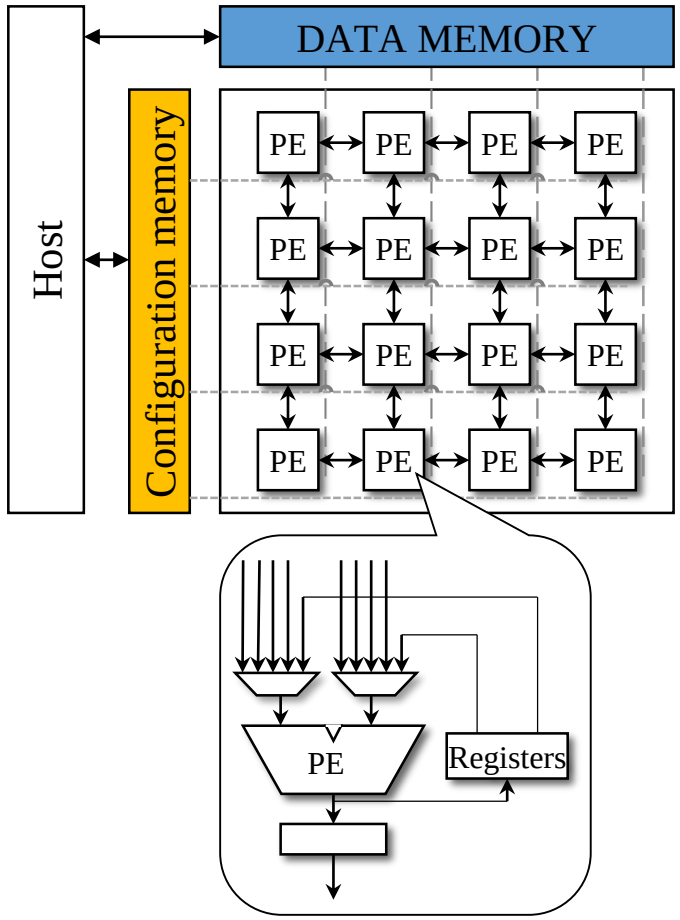
数据流图




数据通道



# 可重构计算架构ASIC执行方式-示例

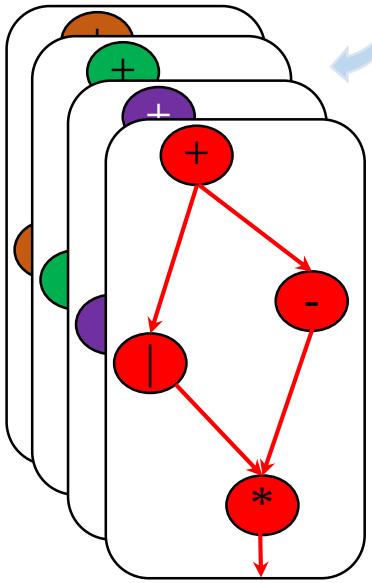



CGRA 基础架构

程序

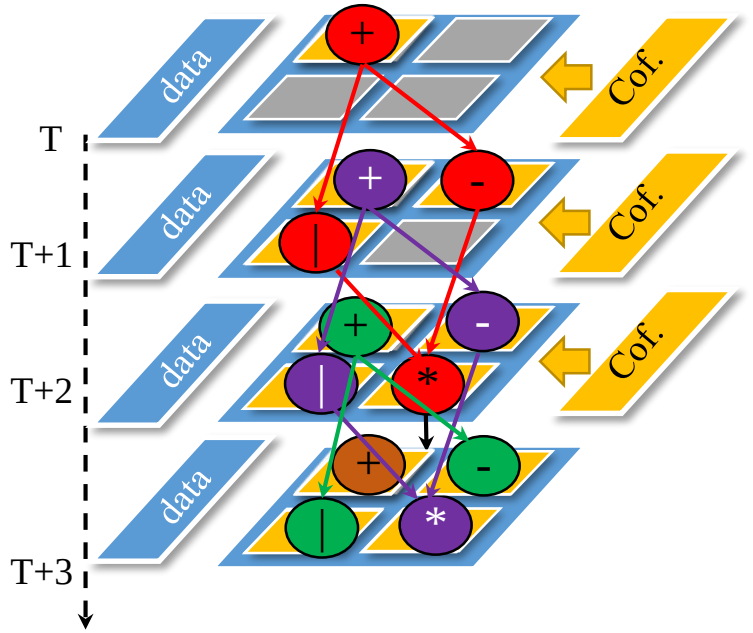
```
for(i=1; i<N; i++){
  a[i]=b[i]+c[i];
  d[i]=a[i] | k[i];
  e[i]=a[i]-f[i];
  g[i]=d[i]*e[i]; }
```

数据流图



循环体

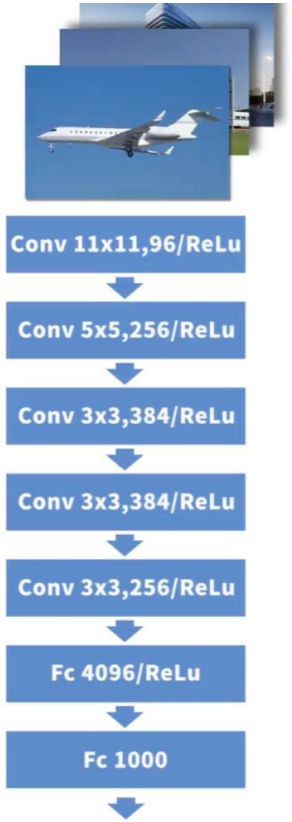
执行

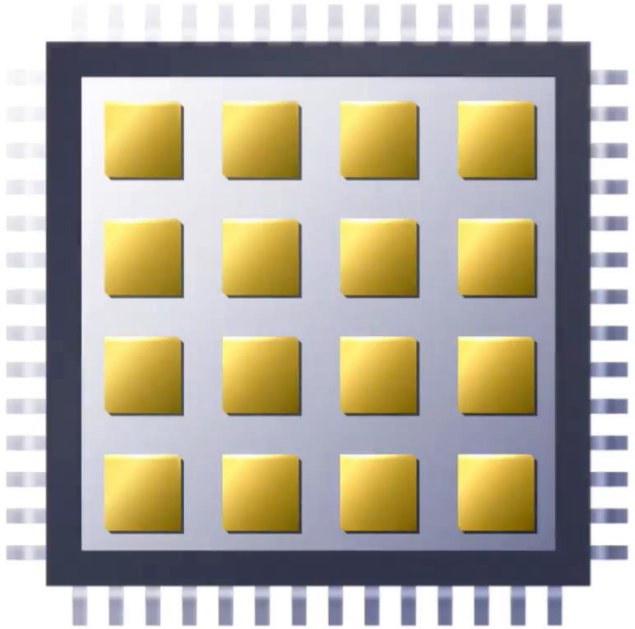



配置为专用电路在数据流驱动下反复执行

# 核心技术-四元编程重构计算模式

- 在人工智能任务计算过程中，架构在“数据位宽-算子操作-数据通道-计算模式”四个维度按需动态重构，实现最优计算模式。
- 效果：计算总能效显著提高。

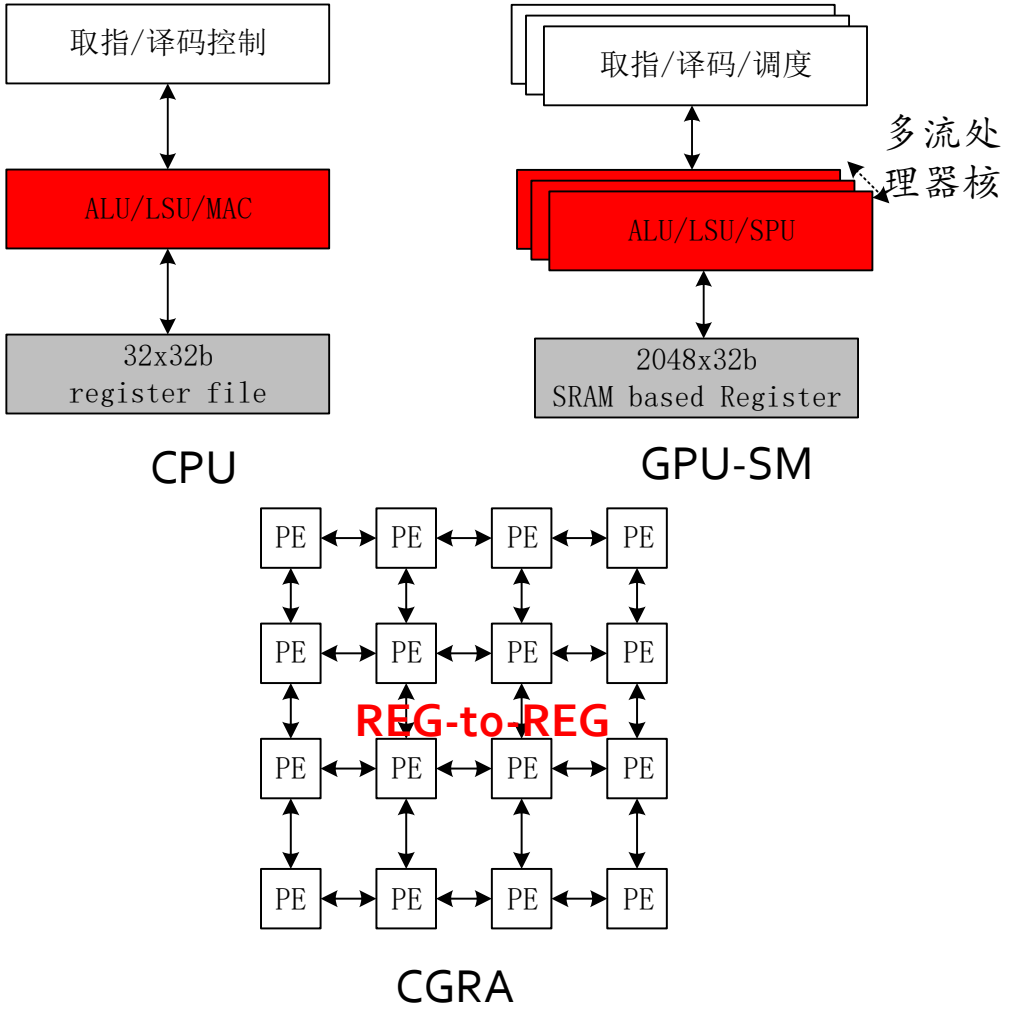


2021电子学会技术发明**一等奖**

# 相比GPU/CPU架构的核心原理优势

| 基础架构 | 模块 | Power(Uw)(28nm工艺数据) | 考虑计算占比的总功耗 | 计算能效(以CPU为基准 | Comments |
|---|---|---|---|---|---|
| CPU | 32x32b register file | 1024 | 1024/20%=5120 | 1 | CPU中ALU通过通用寄存器堆进行数据交互，如RISCV为32个32bit的通用寄存器 |
| GPU | 2048x32b SRAM | 2412 | 2412/80%=3012 | 1.69 | GPU中为了实现SIMT，ALU需要在上千个线程之间选择数据；A100单个SM最大支持的线程数为2048； |
| CGRA | 32bit PE register | 64 | 64/1=64 | 80 | CGRA的数据流大量的数据传输发生在PE之间，PE内的寄存器仅为32b |




**CGRA从核心原理上相比CPU和GPU在同样算力下，有显著能效提升。**

# 相比FPGA架构的核心原理优势

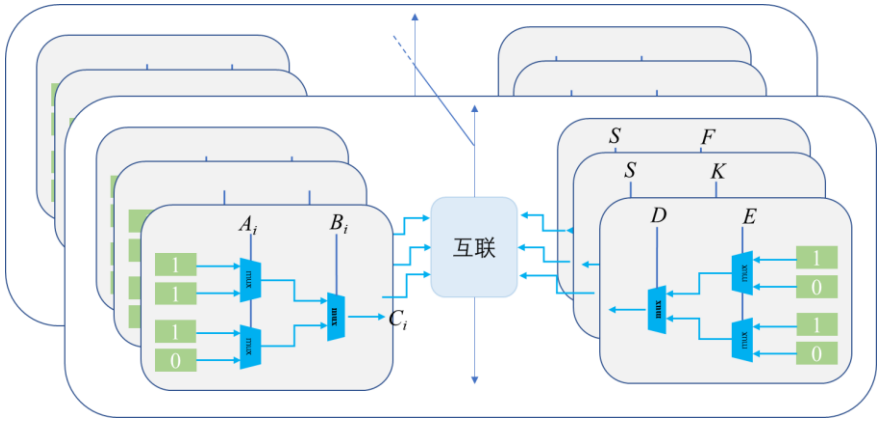
用户程序

```
For(i=1; i<N; i++){
    S[i]=A[i]+B[i];
    D[i]=S[i] | K[i];
    E[i]=S[i]-F[i];
    G[i]=D[i]*E[i]; }
```

## FPGA的实现方式

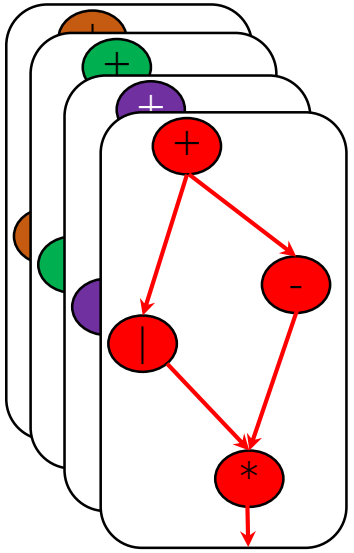
$S_i=A_i \oplus B_i \oplus C_{i-1}$

$C_i = A_i B_i + C_{i-1}(A_i + B_i)$



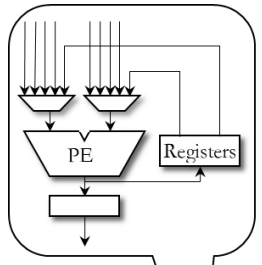

Bit配置

数据流图

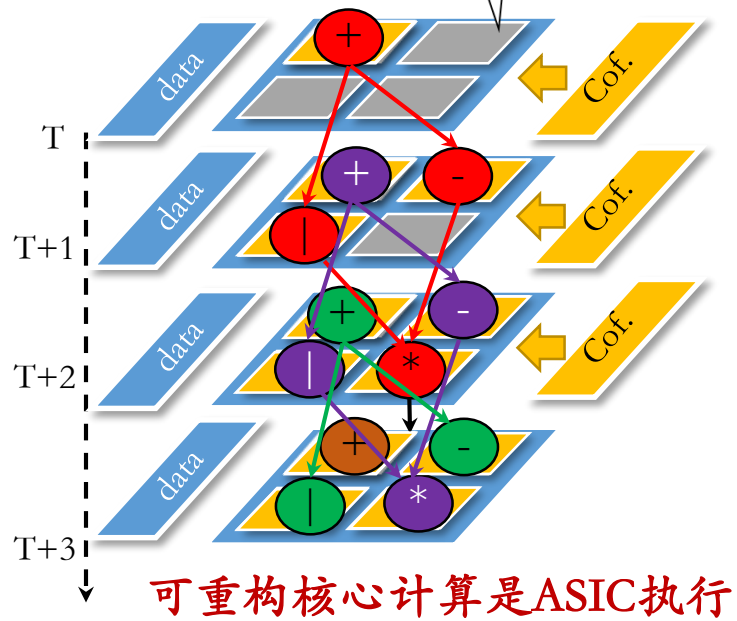


执行

## 可重构计算方式




CGRA 基础 PE架构




1.$C_i=\bar{C}_{i\_1}(Ai+Bi)+A_iB_i$, Bit级结果预存,频繁访存

2. Bit级互联资源开销大，逻辑时延长，主频不高

3. Bit级配置，配置信息多，无法动态配置，
存储开销大

**CGRA相比FPGA在同样算力下，计算效率有数量级提升**

**可重构核心计算是ASIC执行方式**

# 可重构芯片优势

效率最低，性能已经瓶颈。

CPU

第一代计算芯片

intel AMD ARM

资源效率低，能效低，编程时间长，成本高。

FPGA

第二代计算芯片

XILINX ALTERA LATTICE SEMICONDUCTOR

能效较低，性能接近瓶颈。

GPGPU

第三代计算芯片

AMD NVIDIA Imagination

CGRA

**第四代计算芯片**

清微智能 TSING MICRO

SambaNova SYSTEMS

**1、性能。相比FPGA性能再提升10+倍。**

**2、灵活性。接近CPU的通用灵活性。**

**3、资源效率。接近ASIC的资源效率。**

**4、设计规模。能够实现大规模系统。**

**5、生态。自主开放的逐渐成熟的生态。**

# 主要内容

❶ 清微智能简介

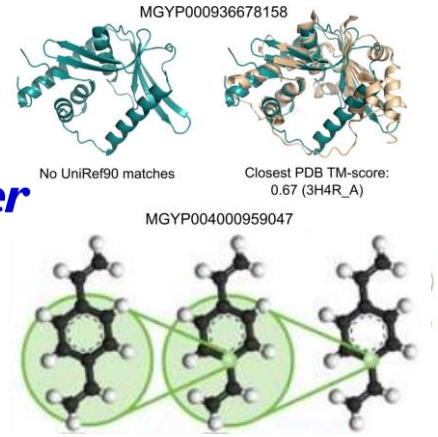
❷ CGRA基本原理

❸ **清微大算力可重构芯片介绍**

# CGRA大算力芯片-迎接算力大爆炸时代

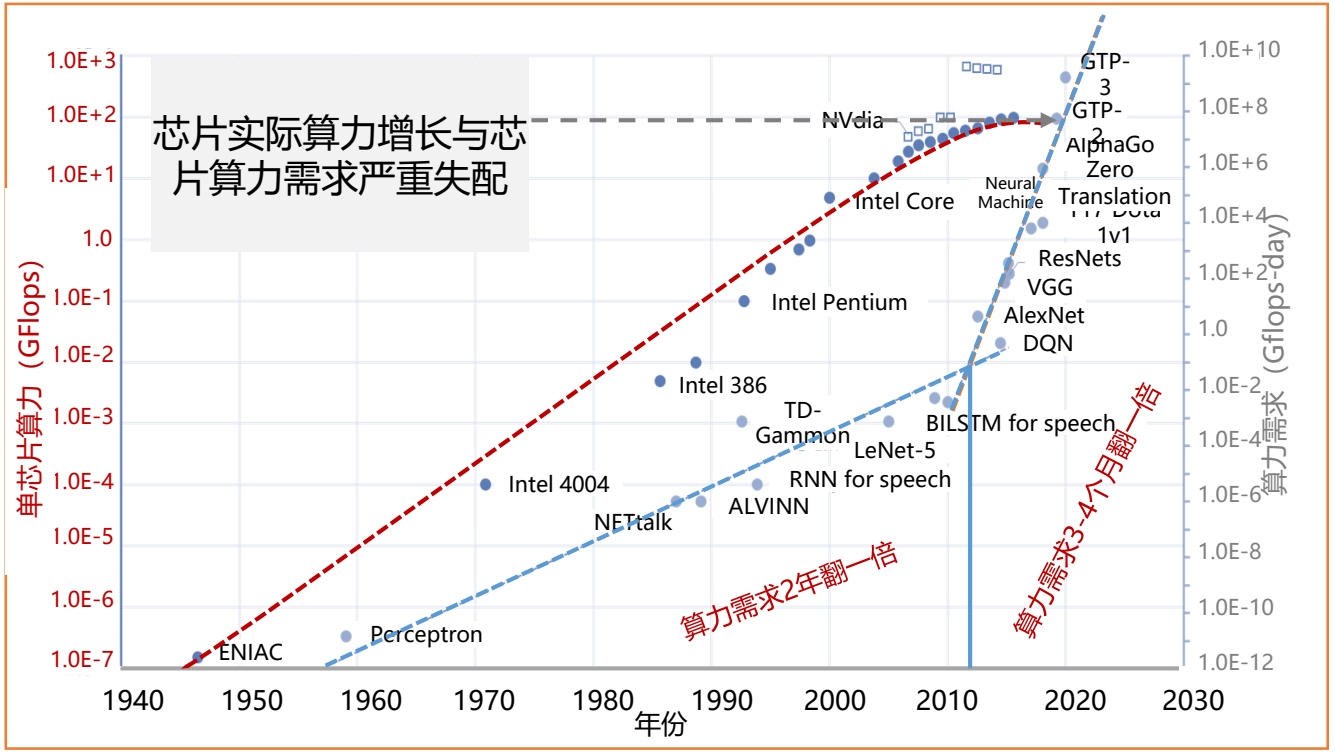
AI超大模型快速演进

蛋白质分析 DeepMind
药物研发 Pfizer
自动驾驶 TESLA
人机对话 Google PaLM
… …




**算力占据绝对主导地位**

**超大模型需要超大算力支撑**



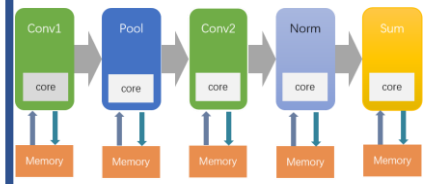

自 2010 年以来，ML 模型的训练计算量惊人地增长了 100 亿。

例如PaLM-540B模型的训练需求算力是**2.56e24 FLOPs**，在**840万**个TPUv4上训练需要**64天**，花费近**2000万美元**!

**亟需开发新型大算力芯片减少超大模型训练时间和成本。**

# 实现超大算力面临的挑战

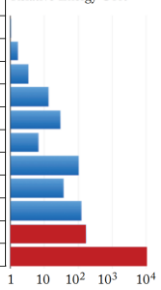
## 访存墙问题




CPU、GPU等共享存储式计算架构，数据在计算单元和存储器中来回搬运，造成访存瓶颈，限制算力提升。

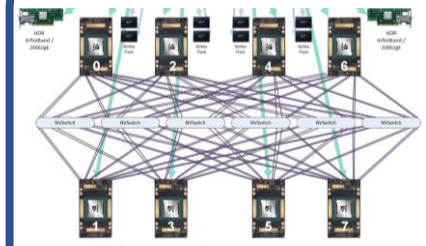
## 能量效率低

| Operation: | Energy (pJ) |
|---|---|
| 8b Add | 0.03 |
| 16b Add | 0.05 |
| 32b Add | 0.1 |
| 16b FP Add | 0.4 |
| 32b FP Add | 0.9 |
| 8b Multiply | 0.2 |
| 32b Multiply | 3.1 |
| 16b FP Multiply | 1.1 |
| 32b FP Multiply | 3.7 |
| 32b SRAM Read (8KB) | 5 |
| 32b DRAM Read | 640 |




频繁的访存和依赖复杂的网络互联，造成大量能量没有充分的利用在真实计算上，造成能量效率很低。

## 算力拓展性差



算力的大规模拓展依赖于大量的高性能网络设备，多层网络设备造成拓展成本高和性能瓶颈。

## 通信墙问题

基于交换机等网络设备进行通信，会使系统性能受限于通信带宽和延时，形成通信瓶颈，限制算力的提升。

## 编程墙问题

现有的编程模型和开发环境无法适应超大算力计算系统，多种层次的并行度难以被充分挖掘。



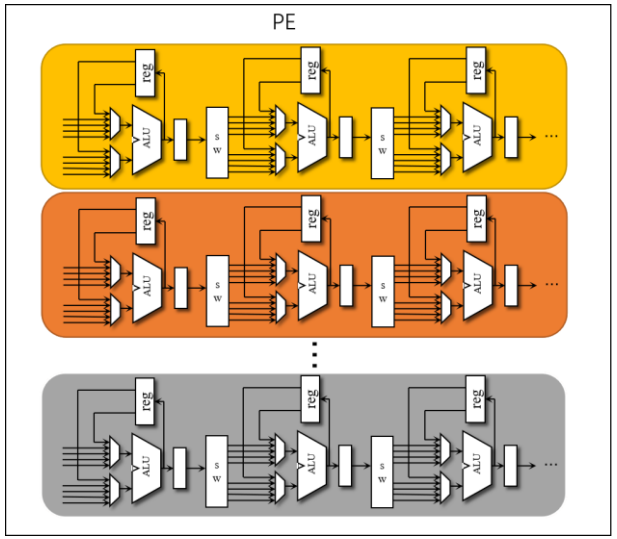
TSING MICRO

# 主流人工智能算法特点

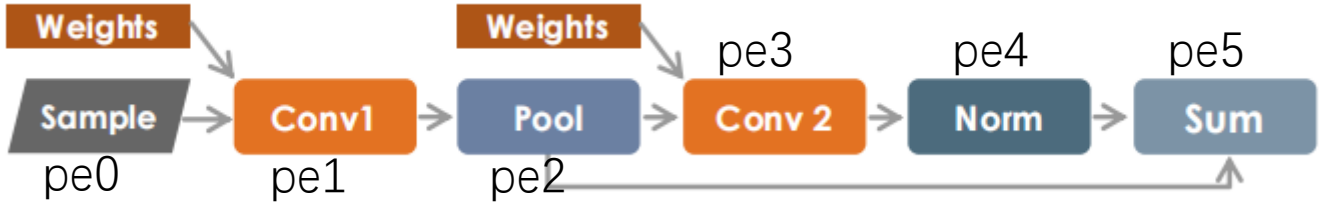
- 数据密集 √
- 控制密集 ×

- 蕴含的并行度高
  - 指令级并行
  - 数据级并行
  - 线程级并行

- 数据流计算
  - 数据块在层间传递
  - 规则的访存模式




CGRA PE内部为“类SIMD流水线“，可以充分释放程序蕴含的指令级并行和数据级并行能力。

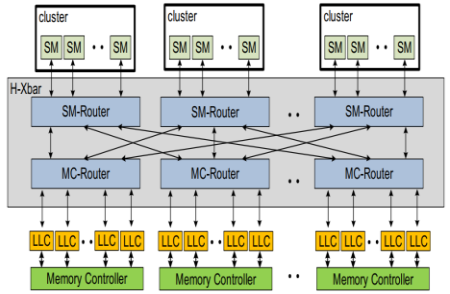



CGRA PE之间天然的数据流计算范式，很好匹配人工智能算法特征。

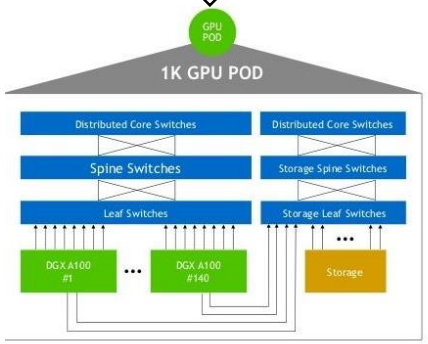


TSING MICRO

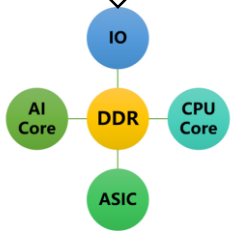
# 传统GPU共享存储架构方式-算力拓展性差




共享存储架构




通过主机网卡以及交换机实现通信




存储为中心计算模式

## 传统GPU架构 VS TX8可重构架构

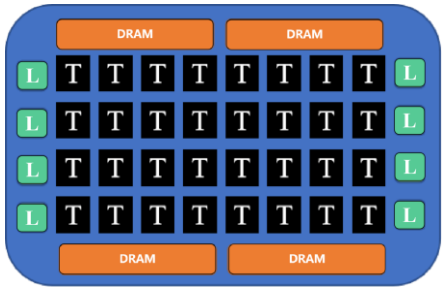
- 共享存储式架构
- 访存瓶颈
- 能效低下
- 交换机拓展算力
- 大算力拓展困难
- 成本高昂
- 带宽瓶颈
- 延迟大
- 部署困难
- 编程不友好

- 消息传递式架构
- 点对点通信
- 高能效
- 芯片直接互联
- 算力无限拓展
- 低成本
- 低延时
- 快速部署
- 编程扁平化

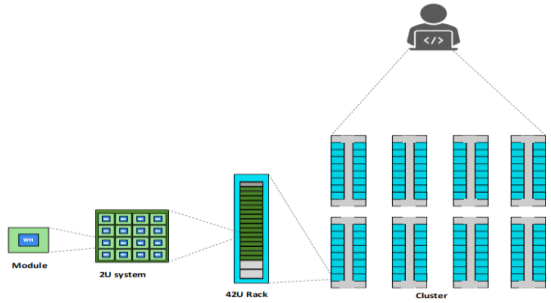
| 模型 | 参数量 | GPU数量 | 单卡峰值性能(TFLOPs) | 训练时间(天) |
|---|---|---|---|---|
| GPT-3 | 1746亿 | 384 | 144 | 90 |
| | | 768 | 88 | 74 |
| | | 1536 | 44 | 74 |
| | 5296亿 | 640 | 138 | 169 |
| | | 1120 | 98 | 137 |
| | | 2240 | 48 | 140 |

GPU单卡峰值算力随集群增大而降低

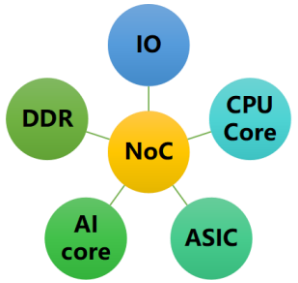
GPU数量翻倍，训练时长没有缩短




mesh网络+tsmlink

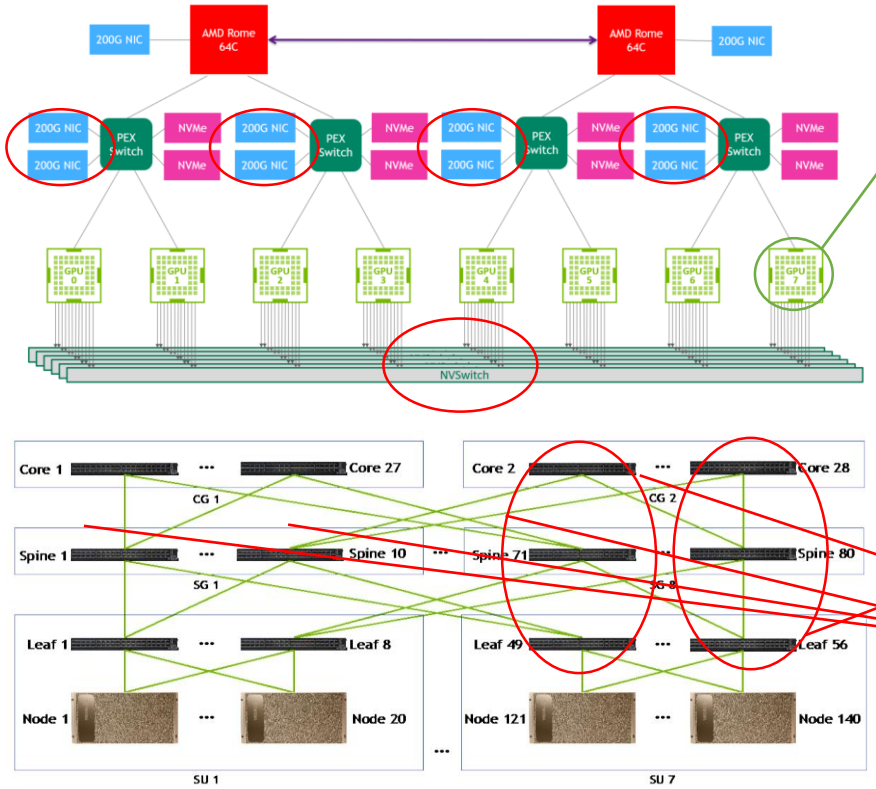



芯片之间直接互联通信




互联通信为中心计算模式

# 传统GPU架构无法很好scale out，交换机成本高



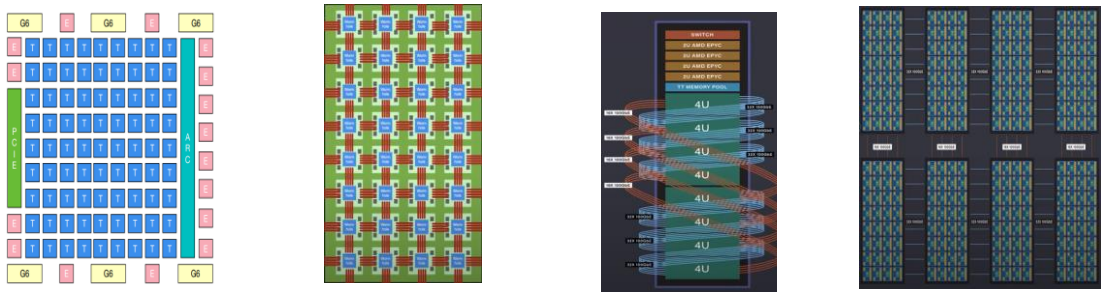

$12,000
Per card

$199,000
Per DGX

$7 million
Per Superpod

## 清微智能——无网卡和交换机解决方案

chip → Server → rack → superComputer



CGRA天然数据流计算 → 天然适合大规模拓展

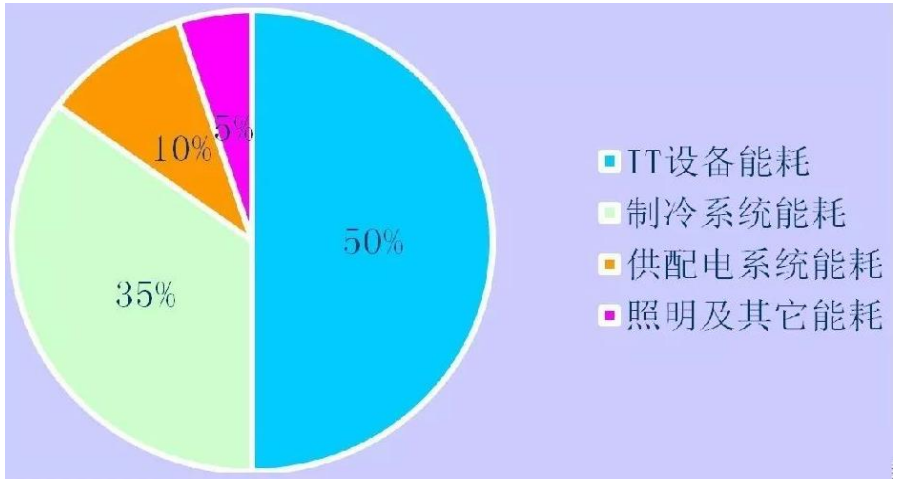
芯片与芯片，server与server直接互联，无需网卡、交换机等网络设备。

**基于传统GPU架构搭建supercomputer，交换机成本近40%，大量资源浪费在交换机上。**

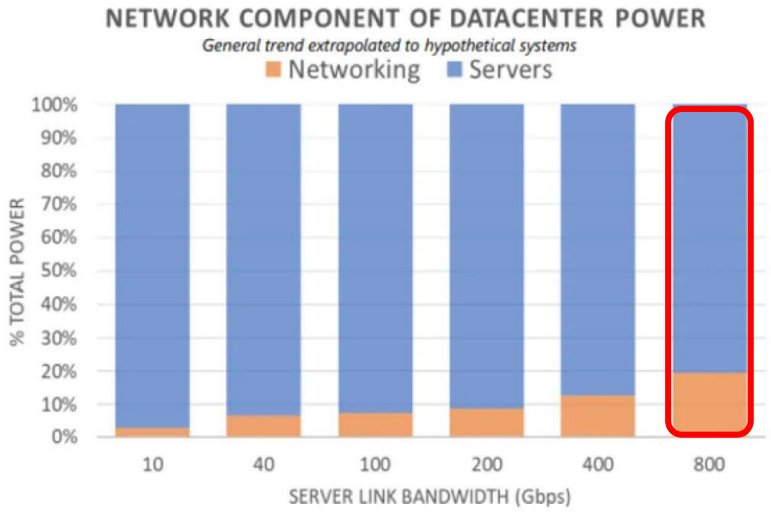


TSING MICRO

# 计算中心能耗-交换机等网络组件能耗可观








网络功耗占比高达20%

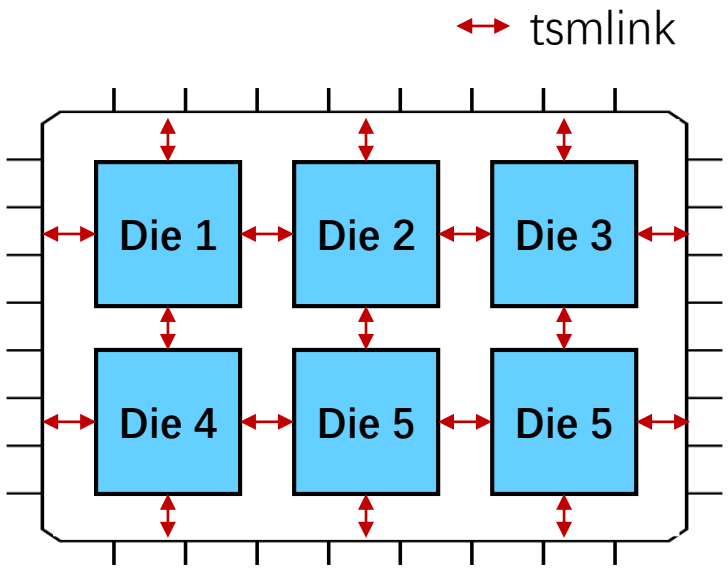
| 具体案例 | TOP500最新排名 | 算力（Eops） | 计算能耗（MW） | 网络系统能耗（MW） | 计算能耗效率 Eops/MW |
|---|---|---|---|---|---|
| 天河-2A | 第7 | **0.1(fp32)** | **18.5** | NA | **0.0059（fp32）** |

**大量的网卡和交换机等网络设备也消耗了可观的能量，降低了计算系统的能效。**

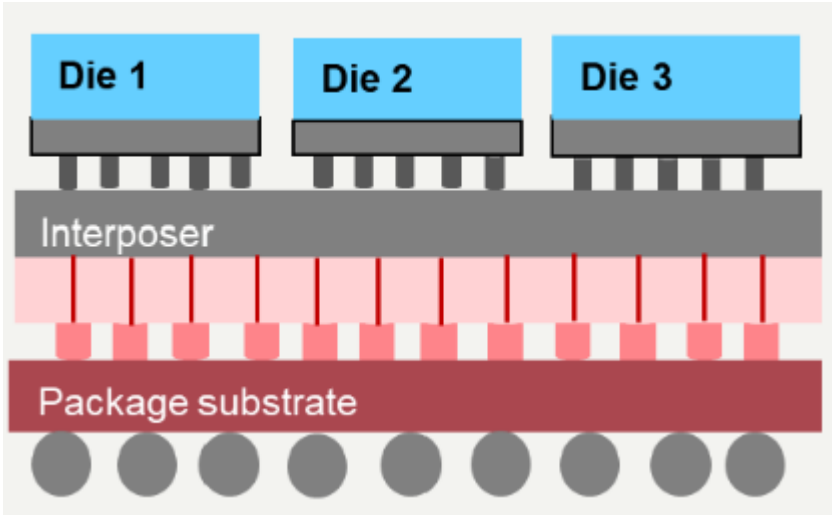
**TX8大算力可重构芯片能够实现芯片之间直接扩展，无需借助交换机等网络设备，实现能效大幅提升。**

# 大算力芯片架构方案：多CGRA die 集成-高算力

**多个CGRA Die通过tsmlink互联，利用2.5D chiplet集成chip，实现大算力拓展。**




顶视图



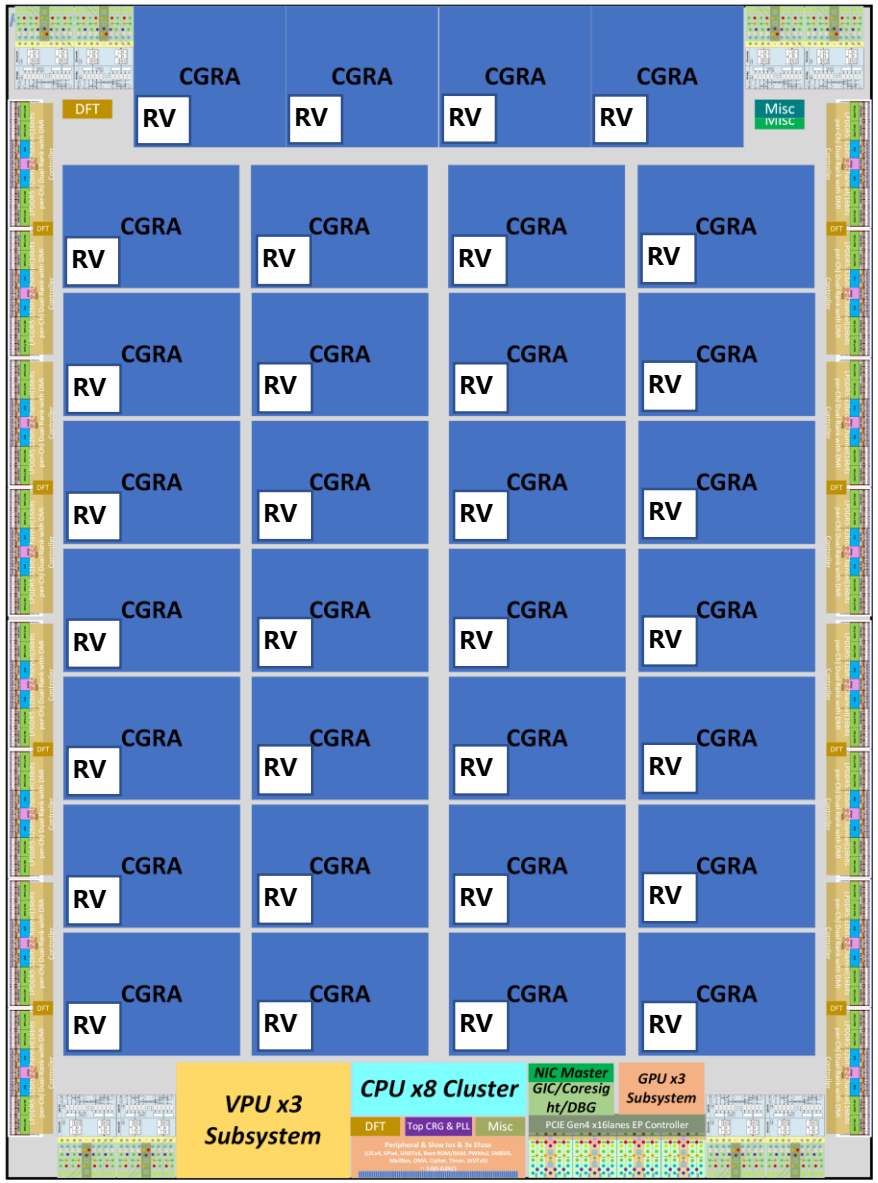

切面图

**tsmlink是清微智能独创的互联接口，能够同时支持die2die互联和chip2chip互联。**

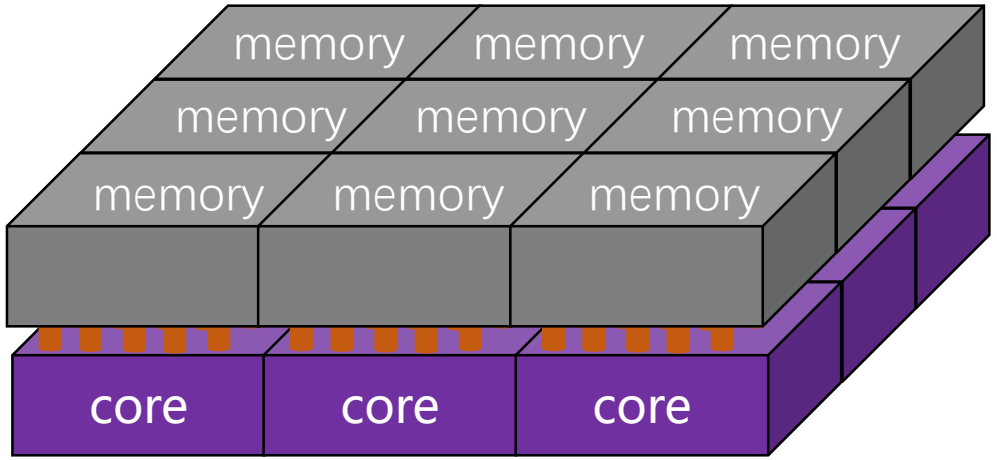


TSING MICRO

# 大算力芯片架构方案：CGRA + 3D chiplet集成



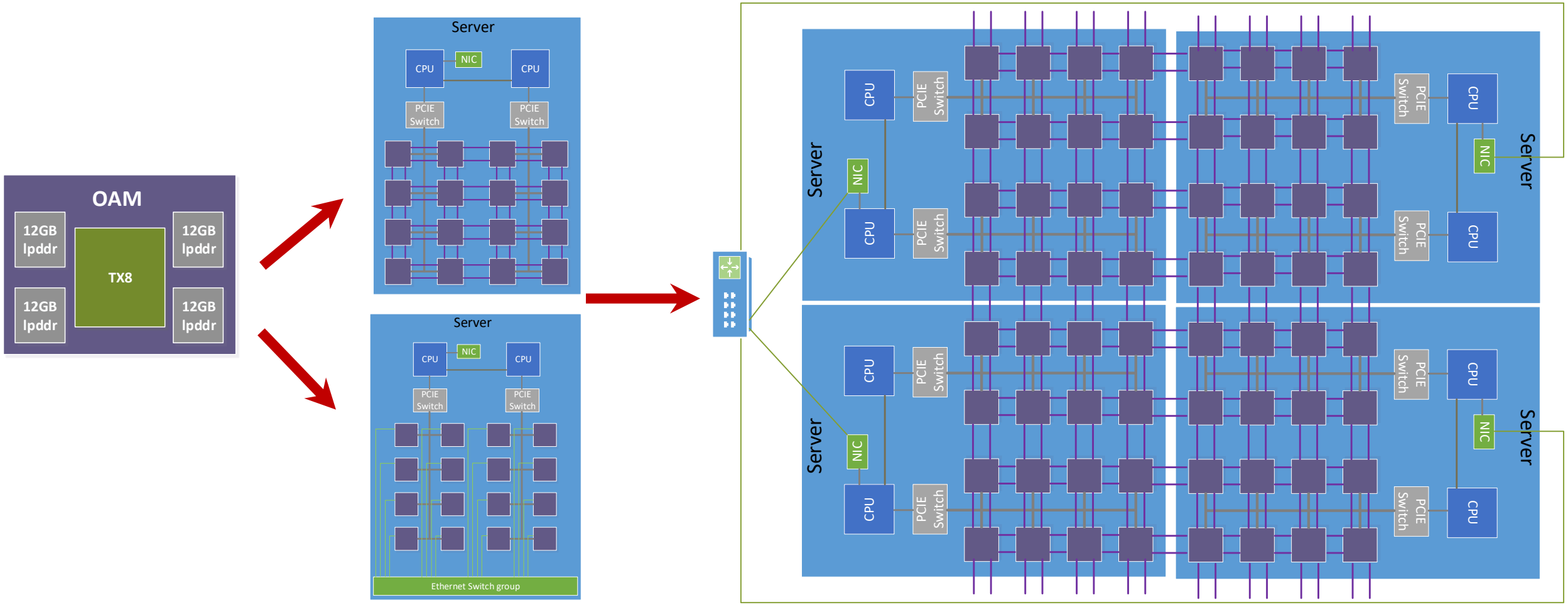

**高带宽3D集成方案**




CGRA core与存储器垂直方向互联，数据传输距离大幅减少，有效增加访存带宽。

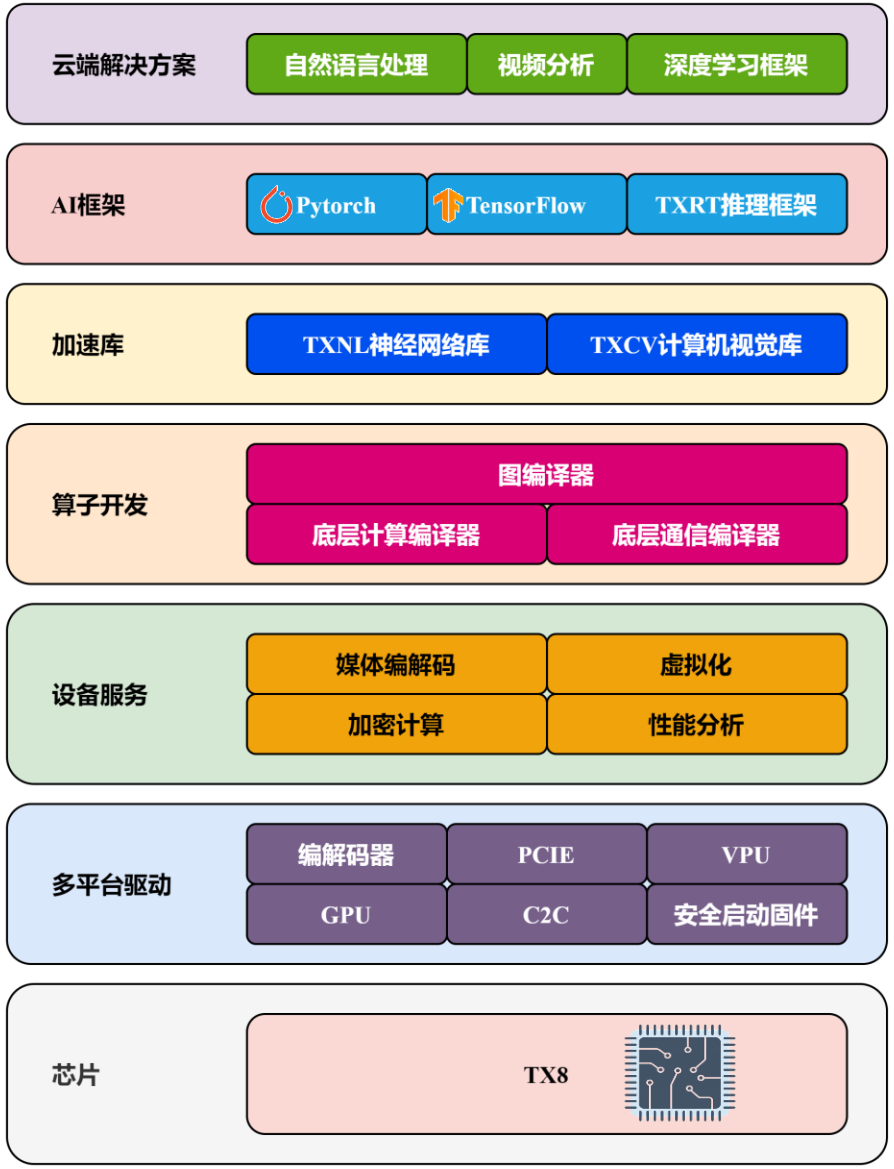
**将CGRA logic Die 和 memory Die 3D集成**

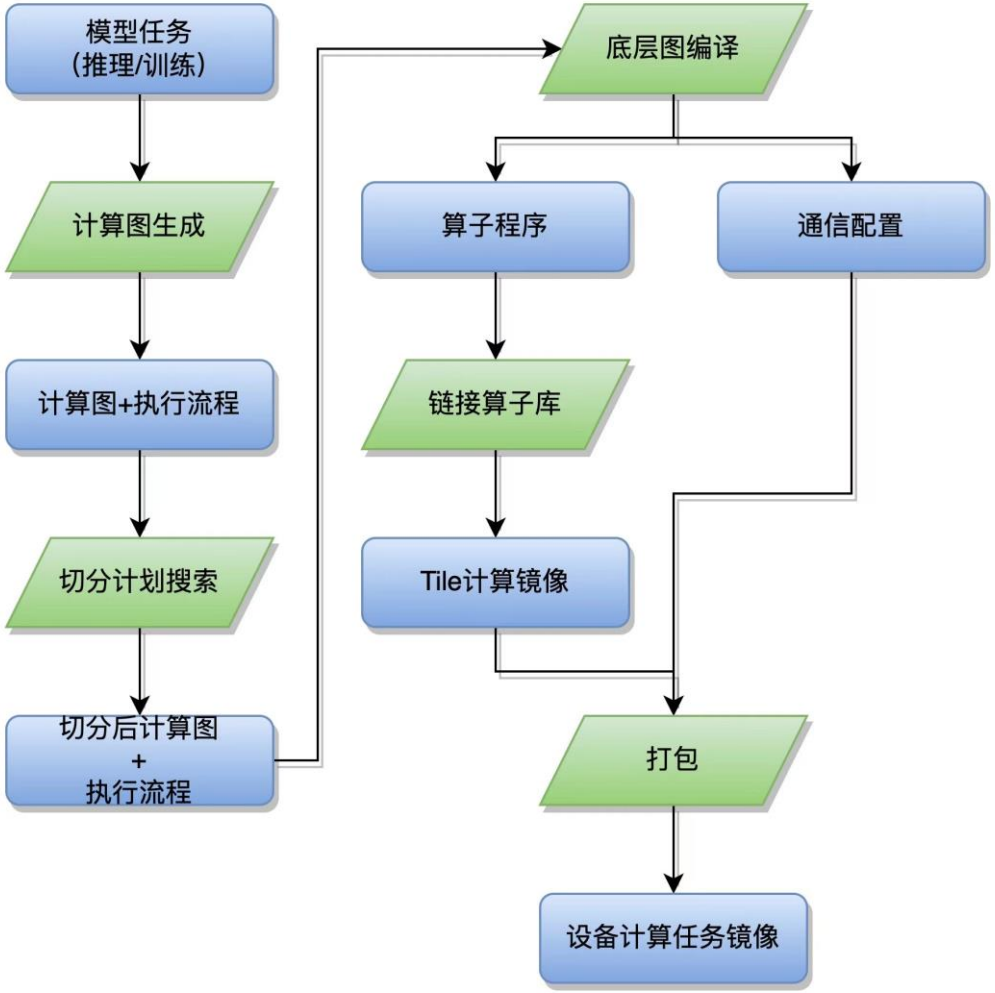
# 大算力解决方案——大算力训推一体系统




**极佳的算力拓展性，计算芯片TSMlink直接mesh连接，无需交换机，能够拓展到E级算力**（1 EFLOPs=1000000 TFLOPs）

# 支撑大模型训练的完善软件栈

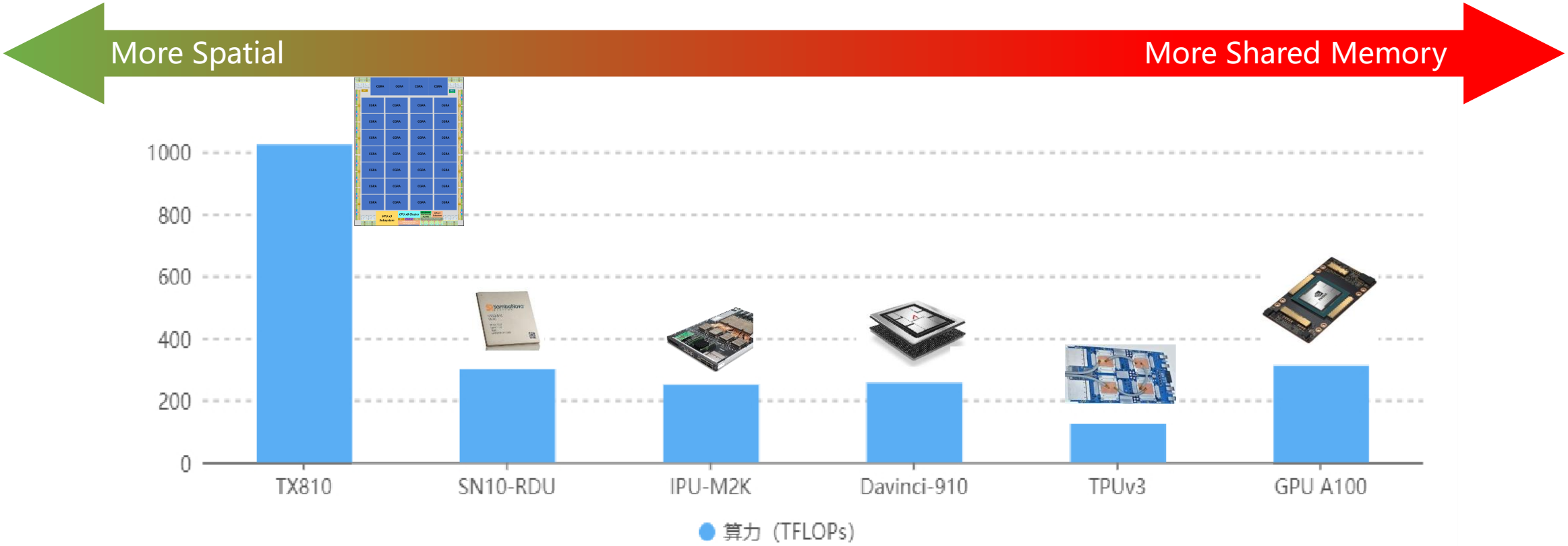





TSING MICRO

# 可重构计算天然空域执行方式，性能成本优势明显




| 可扩展性 | 灵活<br>不需要交换机 | | 灵活<br>不需要交换机 | | 灵活<br>不需要交换机 | | 效率低<br>需要交换机 | | 效率低<br>需要交换机 | | 效率低<br>需要交换机 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 计算能效 | 最高 | | 高 | | 高 | | 中 | | 中 | | 低 | |

**More spatial的Tsingmicro和Sambanova架构处理器（可重构处理器）相比More share memory的Nvidia GPU在处理万亿参数模型时，可以获得功耗大幅减少。**



TSING MICRO

清微智能
TSING MICRO

# THANKS

感 谢 观 看

北京清微智能科技有限公司