sw26010众核处理器

——国防科大2020年高性能评测与优化课程小组讨论

组员：高光杰、孙萌婧

指导：龚春叶、甘新标、杨博

目录

- 简介
- Sw26010介绍
- 性能分析
- 总结

时间	型号	简介
2006年	SW-1	130nm制程工艺的单核心CPU，主频900MHz，集成5700万晶体管（第二代SW-1）
2008年	SW-2	130nm制程工艺的双核CPU，主频1.4GHz（第二代SW-2）
2010年	SW1600	65nm制程工艺的16核CPU，主频1.1G，双精浮点140G。申威1600被用于神威蓝光超算（第三代SW-3）
2012年	SW1610	40nm制程的16核CPU，集成10亿晶体管，主频1.6G，最大功耗50W，双精浮点运算200G,用于服务器，支持中标麒麟操作系统
2012年	SW410	40nm制程的4核CPU，集成2.7亿晶体管，主频1.6G,用于PC，支持中标麒麟操作系统
2014年	SW-5	集成了4个管理核心和256个运算核心的高性能众核CPU，双精浮点运算超过1T，核内 linpack效率93%，并有很高的性能功耗比
时间不详	SW26010	频率1.45GHz，260个核心，包括4个MPE管理单元、4个CPE计算单元及4个MC内存控制器单元。作为国际首款万亿次异构众核处理器，集成260个运算核心，峰值速度突破每秒3万亿次，达到国际领先水平

SW26010简介

- 面向构建十亿亿次超级计算系统
- 自主知识产权的申威指令集（SW-64）
- 片上融合异构众核架构
- 集成4个运算控制核心和256个运算核心
- 核心根据需求扩展了256位向量指令集

为什么研究SW26010?

- 超级计算机的的需要

cpu关键部件技术必须自主可控，避免技术封锁，2015年4月美国就禁止对nudt和几个超算中心出口芯片

- 上一代sw1600芯片已经不能满足新的超级计算机的需要，急需一款新的cpu

Sw26010结构

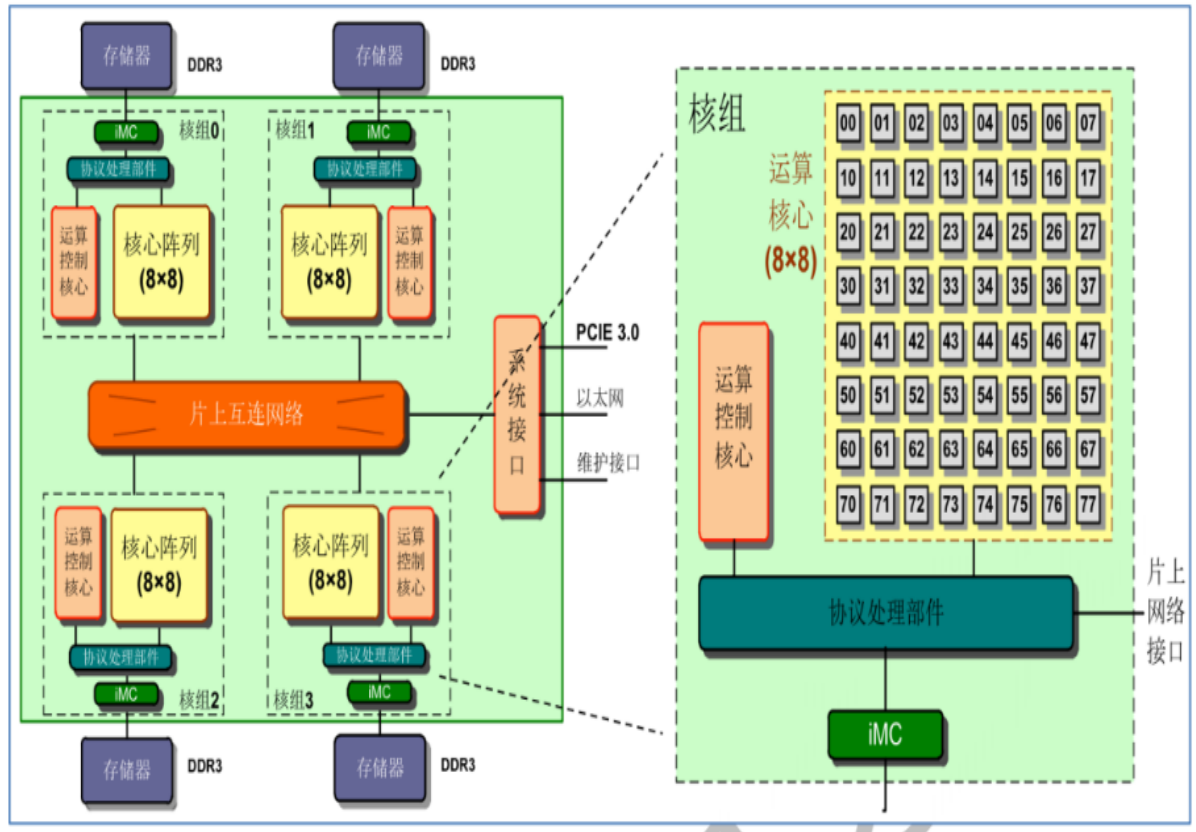


图 1-2“申威 26010”异构众核处理器架构图

- 4个核组（CG）
- 每个核组包括：一个mpe（控制核心-主核）、一个运算核心阵列（从核阵列），从核由64个（8x8）的运算核（从核）
- MPE和CPE都是64位RISC、单线程内核，工作频率为1.45GHz，支持256位矢量（包含4个单/双精度浮点数）指令（包括融合乘加FMA）和32个矢量寄存器（从32个64位通用寄存器扩展而来）

Sw26010主核从核

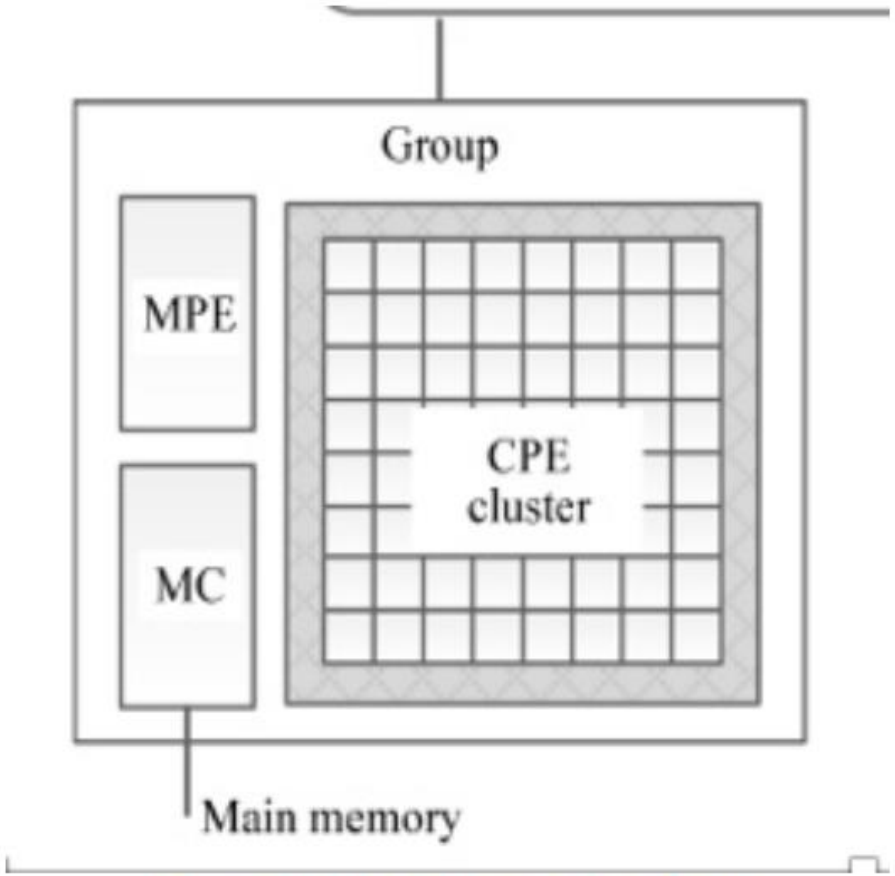
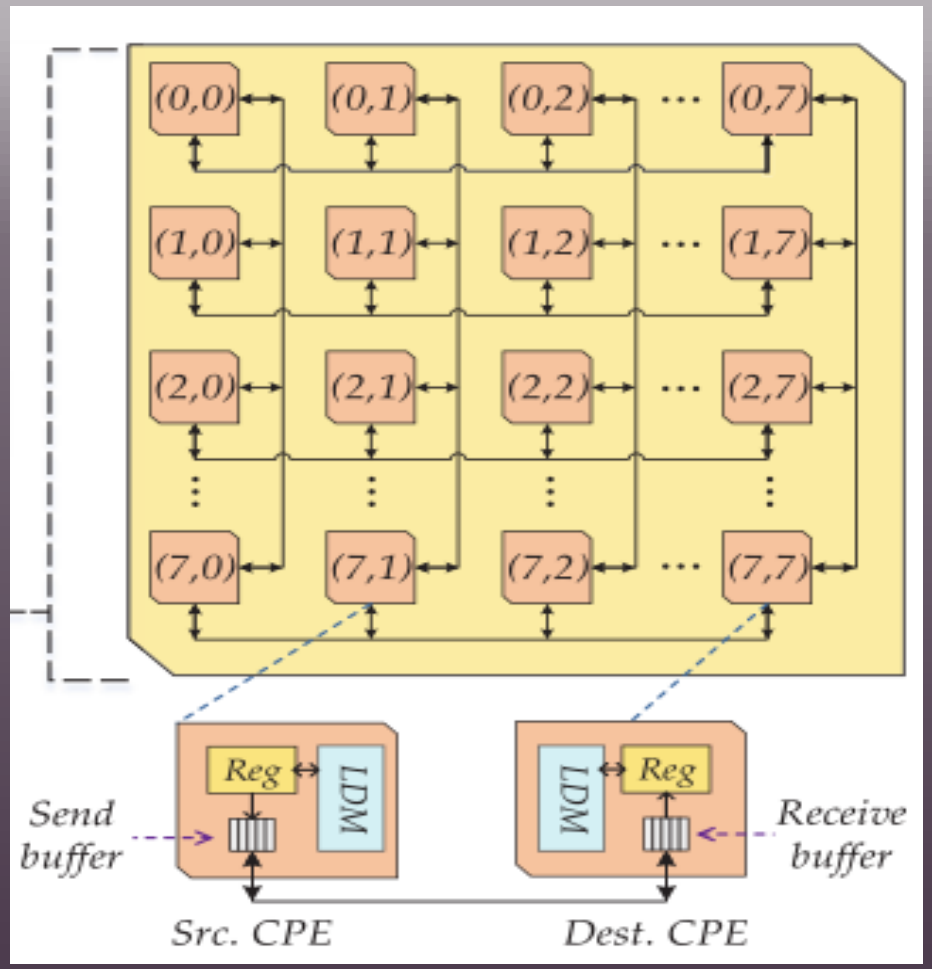


Figure 1: Core Group for Node

•主核：主核作为一个完整的64位 risc核，允许在用户态以及系统态下使用。主核和其他商用的 cpu 一样完全提供函数中断、内存管理、乱序执行等操作以及256-bit的向量化操作。因此，主核是管理和通信的核心模块。

•MPE具有两个浮点流水线，每个浮点流水线每个周期执行8个32位/ 64位浮点运算，每个内核的峰值性能为23.2 GFlops（即，1个处理器中的4个MPE，共0.0928 TFlops 。关于在内存层次结构中，每个MPE都拥有一个32 KB的L1 数据Cache，32 KB L1 指令Cache和256 KB L2 cache

Sw26010主核从核

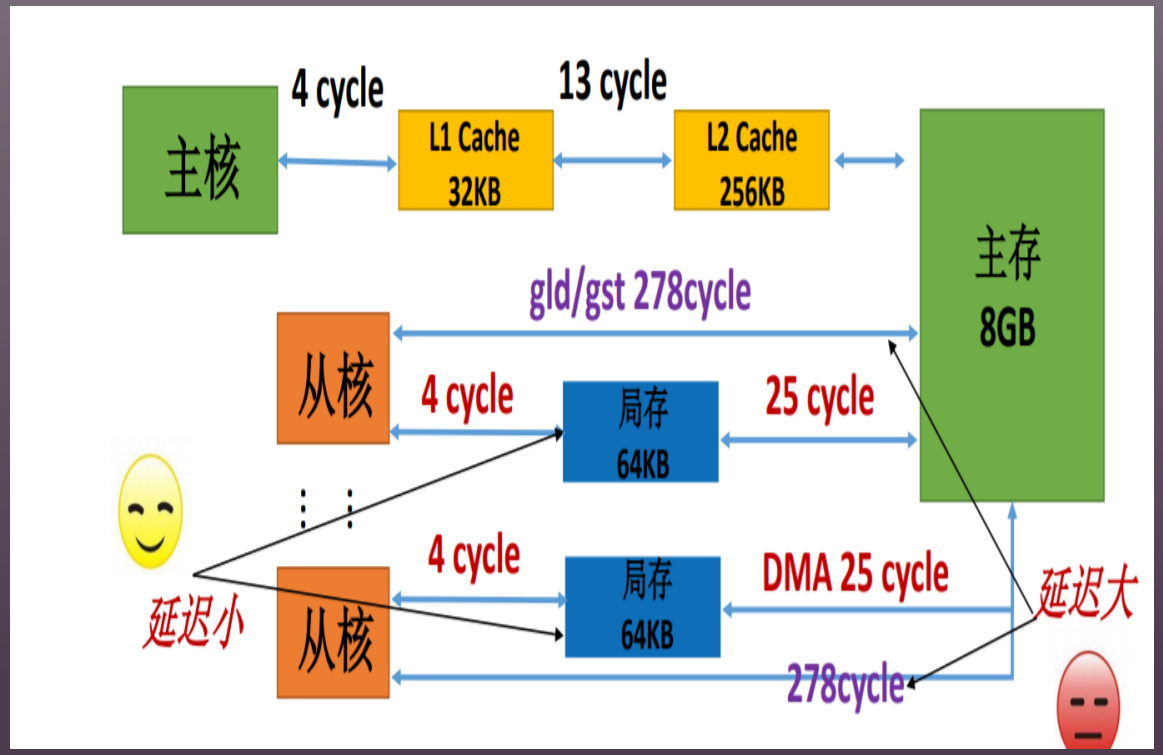


•从核：从核只能运行在用户态下而不能运行在系统态下，也不支持函数中断。从核的设计原理是在最小化核心的微架构的复杂度的同时最大化处理器的计算能力。从核通过8*8的网孔网络连接，来实现核间通信的低延迟以及核间同步。从核亦支持256-bit的向量化操作。因此，从核是计算的核心模块

•每个从核包含两条流水线：负责浮点运算的流水线P0，以及负责存储访问的流水线 P1。从核拥有16kb的L1数据cache，没有L1指令cache以及L2 cache，但每一个从核提供了64kb的scratch-pad memory（SPM），SPM有两种使用方式，一种是作为软件cache自动实现数据的缓存，一种是完全由用户组织数据的传输与换入换出，第二种方式多数情况下更高效。

•RLC，这是一种轻量级的向量寄存器的快速数据通信机制，由网络连接的64个CPE中提供。

Sw26010存储结构



•每个从核有一个64KB的局部存储（LDM），LDM以草稿本 (Scratch Pad Memory, SPM) 的方式组织，由软件控制管理，属于一种可编程存储器（cache是由硬件控制，对程序员不可见）

•从核可通过 gld/gst 离散访主存，也可以通过 DMA （Direct Memory Access，直接内存访问）批量式访主存

Sw26010异构并行方式

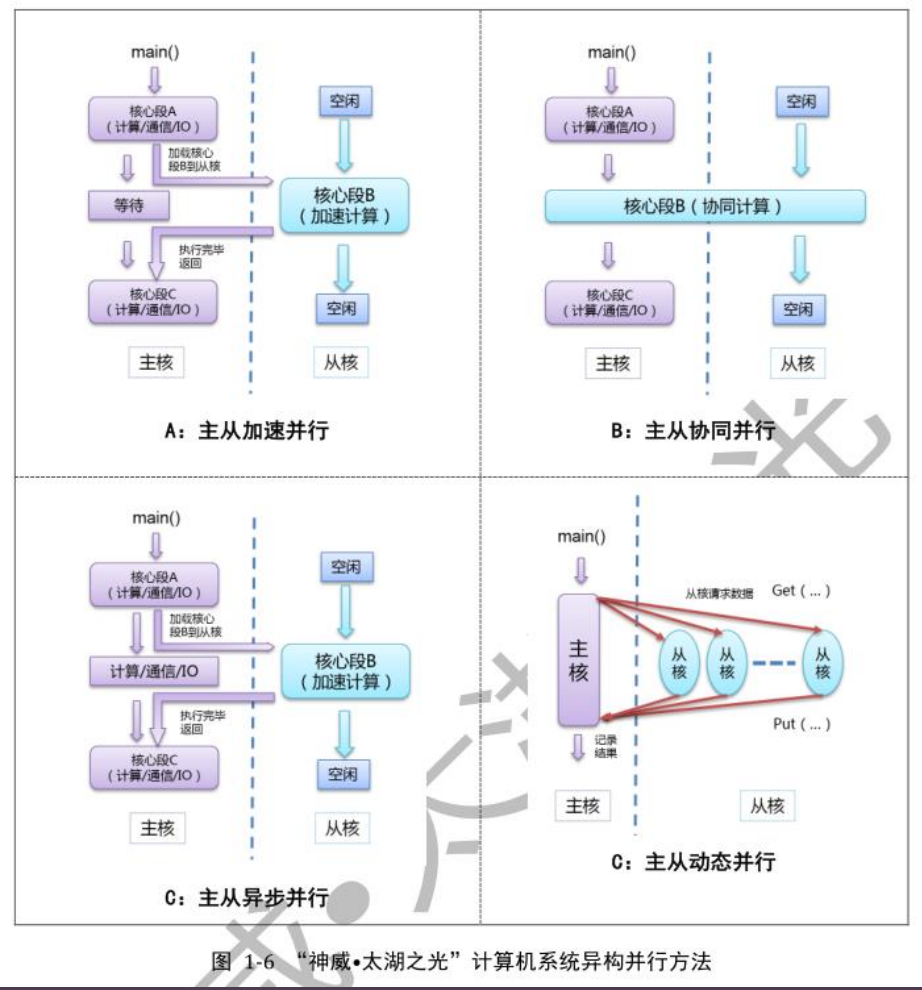


图 1-6 “神威•太湖之光”计算机系统异构并行方法

- 主从加速并行：主核负责通信、IO，从核计算加速，从核在计算核心段过程中，主核处于等待状态，直到从核完成该核心段的计算任务
- 主从协同并行：主核和从核作为对等的个体进行并行计算，根据各自计算能力进行负载分配，共同完成核心段的计算
- 主从异步并行：在从核进行加速计算的同时，主核不等待，而是进行其他如计算、通信或I/O 等操作，从而提高主从协作的计算效率
- 主从动态并行：主核负责任务分配，从核负责任务计算并回写计算结果

Sw26010主要性能指标

- 主核和从核工作频率 1.45GHz;
- 计算能力:
- a) 主核的峰值速度为双精度浮点 24GFLOPS、单精度定点16.5GIPS;
- b) 从核的峰值速度为双精度浮点 12GFLOPS、单精度定点13.5GIPS;
- c) 单芯片峰值速度为双精度浮点 3.168TFLOPS、 单精度定点 3.522TIPS
- 集成 4 路 128 位 DDR3 存储控制器，访存带宽为 136.51GB/s，支持最大主存容量配置 128GB;
- 芯片采用 PCIe 3.0 标准接口，8 通道，双向峰值带宽 16GB/s;

Table 2: Comparison with Top 3 Machines

	ORNL Titan	NUDT Tianhe-2	Sunway TaihuLight
Theoretical Peak	27 Pflop/s = (2.6 CPU + 24.5 GPU) Pflop/s	54.9 Pflop/s = (6.75 CPU + 48.14 Coprocessor) Pflop/s	125.4 Pflop/s = CPEs +MPEs Cores per Node = 256 CPEs + 4 MPEs Supernode = 256 Nodes System = 160 Supernodes Cores = 260 * 256 * 160 = 10.6M
HPL Benchmark Flop/s	17.6 Pflop/s	30.65 Pflop/s	93 Pflop/s
HPL % Peak	65.19%	55.83%	74.16%
HPCG Benchmark	0.322 Pflop/s	0.580 Pflop/s	.371 Pflop/s
HPCG % Peak	1.2%	1.1%	0.30%
Compute Nodes	18,688	16,000	40,960
Node	AMD Optron Interlagos (16 cores, 2.2 GHz) plus Nvidia Tesla K20x (14 cores, .732 GHz)	2 – Intel Ivy Bridge (12 cores, 2.2 GHz) plus 3 - Intel Xeon Phi (57 cores, 1.1 GHz)	256 CPEs + 4 MPEs
Sockets	18,688 Interlagos + 18,688 Nvidia boards	32,000 Ivy Bridge + 48,000 Xeon Phi boards	40,960 nodes with 256 CPEs and 4 MPEs per node
Node peak performance	1.4508 Tflop/s = (.1408 CPU + 1.31 GPU) Tflop/s	3.431 Tflop/s = (2*.2112 CPU + 3*1.003 Coprocessor) Tflop/s	3.06 Tflop/s CPE: 8 flops/core/cycle (1.45 GHz*8*256 = 2.969 Tflop/s) MPE (2 pipelines) 2*4*8 flops/core/cycle (1.45 GHz*1= 0.0928Tflop/s)
Node Memory	32 GB CPU + 6 GB GPU	64 GB CPU + 3*8 GB Coprocessor	32 GB per node
System Memory	.710 PB = (.598 PB CPU and .112 PB GPU)	1.4 PB = (1.024 PB CPU and .384 PB Coprocessor)	1.31 PB (32 GB*40,960 nodes)
Configuration	4 nodes per blade, 24 blades	2 nodes per blade, 16 blades per	Node peak performance is 3.06 Tflop/s, or 11.7 Gflop/s per core.

基于汇编微基准测试集分析sw26010性能

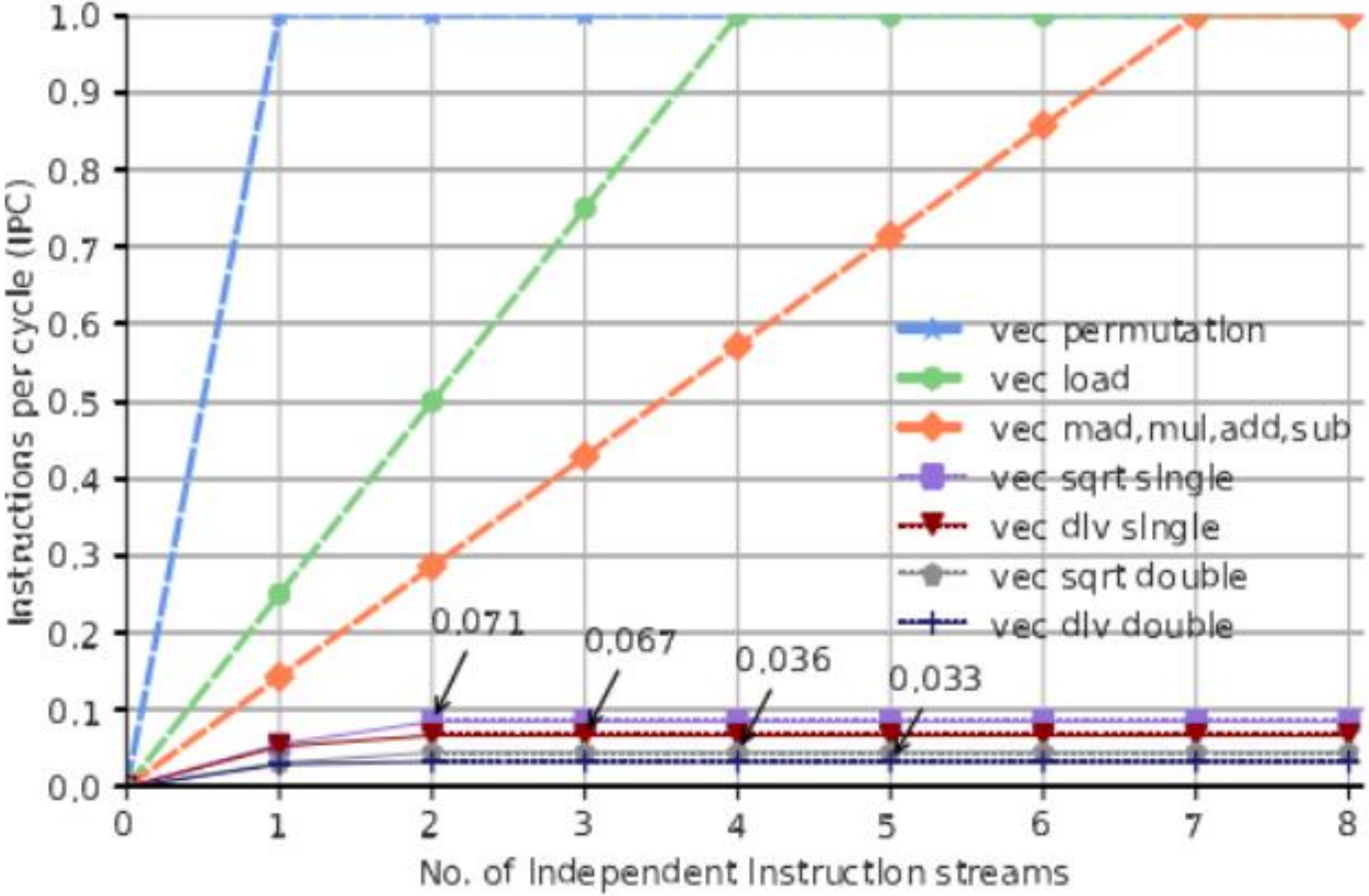
- 1、流水线性能
  - 1.1 指令延迟
  - 1.2 指令吞吐
- 2 、主存访问性能
  - 2.1 DMA
  - 2.2 全局离散访存
  - 2.3 片上spm局部存储性能
- 3、RLC（寄存器通信）性能

流水线性能分析

- 利用了写后读（Read After Write，RAW）相关性来构建指令延迟测试序列：

表 3-1 指令延迟测试结果

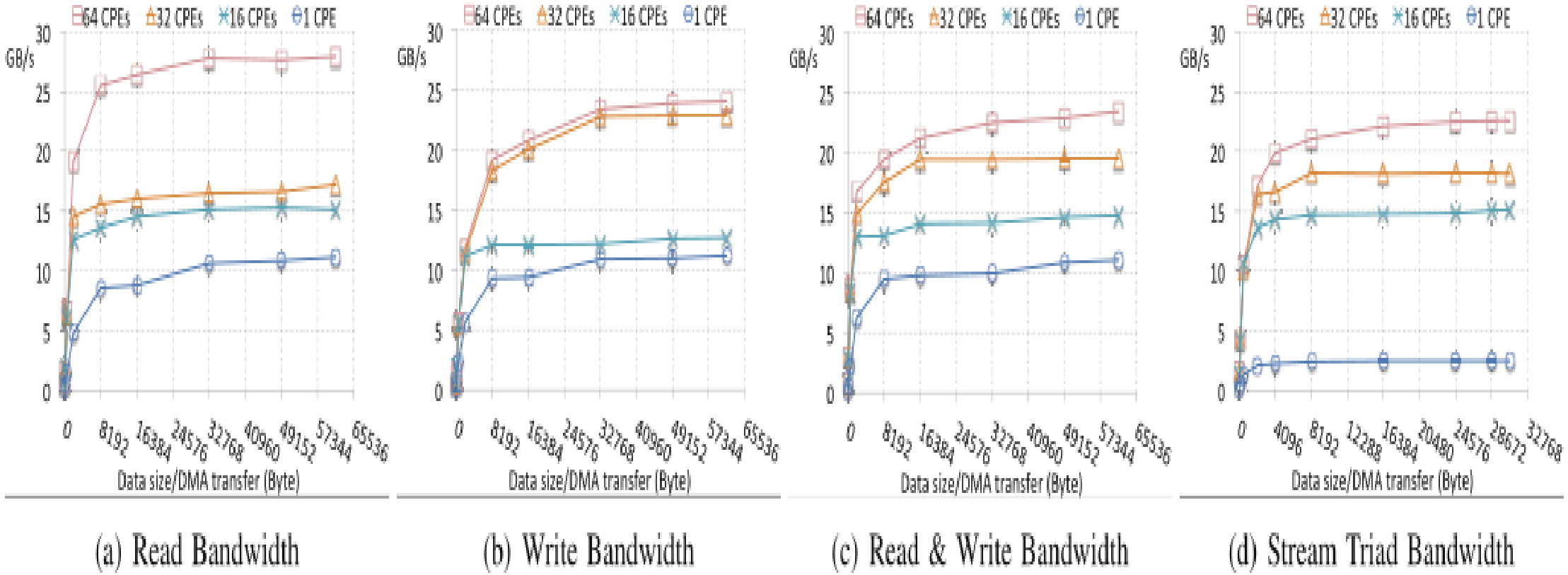
指令类型	指令	延迟 (时钟周期数)
算术运算	vmuld, vmuls, vaddd, vadds, vsubd, vsubs, vmad, vmas	7
	vsqrtd/vsqrts	32/18
	vdivd/vdivs	34/19
SPM 访存	vldd, vlds	4
混洗	vinsw, vinsf, vextw, vextf, vshff, vshfw	1
同步	sync, synr	14


Instructions per cycle (IPC)
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0
1
2
3
4
5
6
7
8
No. of Independent Instruction streams
vec permutation
vec load
vec mad,mul,add,sub
vec sqrt single
vec dlv single
vec sqrt double
vec dlv double
0.071
0.067
0.036
0.033

访存性能分析

- SW26010 CPE的内存层次结构包括三个级别:
- 每核组8GB的主存，通过DMA批量地或者全局load/store离散地访问
- 每个CPE上的64KB Scratch-Pad Memory（SPM）
- 32个64-bit通用寄存器（可扩展为256bit向量寄存器）

DMA访存性能


64 CPEs
32 CPEs
16 CPEs
1 CPE
GB/s
Data size/DMA transfer (Byte)
(a) Read Bandwidth
(b) Write Bandwidth
(c) Read & Write Bandwidth
(d) Stream Triad Bandwidth

DMA广播模式

- 广播模式适用于主存向从核广播数据，因此是单向传输。广播模式下任意一个从核发起广播请求后，数据会从主存传输到从核阵列上，然后再通过从核阵列网络将数据传输到所有从核上
- 实测发现广播模式带宽为 6.97GB/s，考虑到广播模式避免了 63
- 次重复的数据传输，因此有效带宽为 446.1 GB/s(6.97 × 64)。

DMA行模式

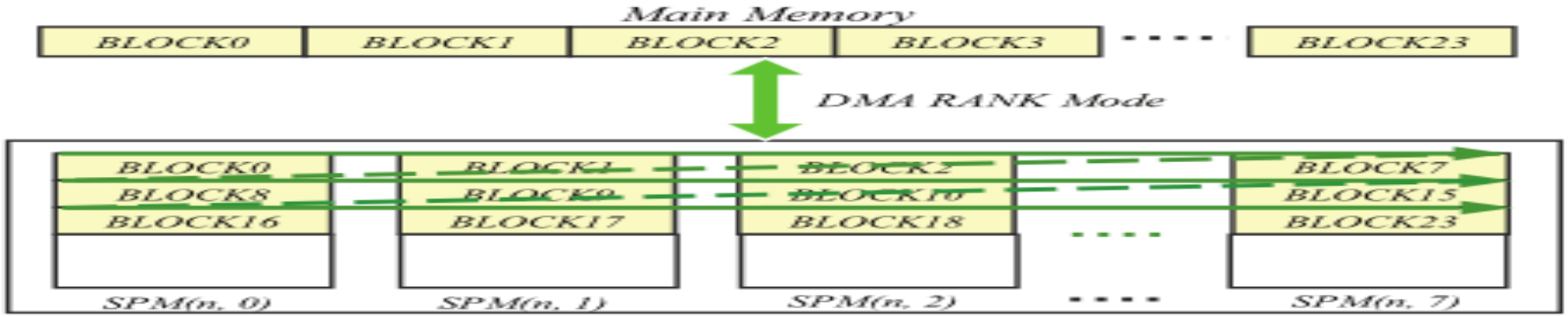


Figure 4: DMA RANK Mode

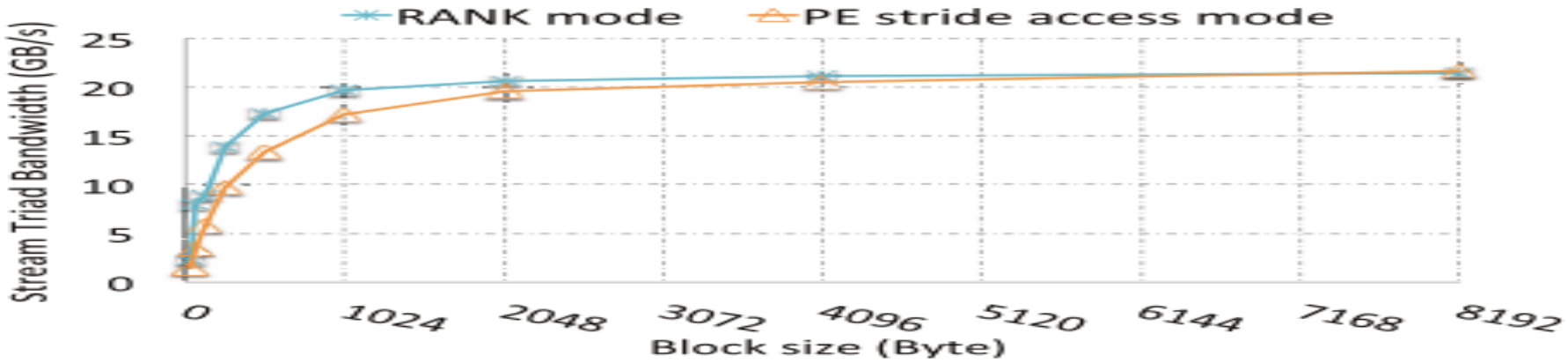


Figure 5: DMA Rank Mode vs. PE unit-stride access Mode

全局离散访存

- 全局离散访存提供了较为便捷的编程模式，用户无需考虑 DMA 传输的显示调用，可以以访问片上存储的方式直接读写主存
- Copy、Scale、Add、Traid 四种访存模式下带宽仅为 3.88 GB/s、1.61 GB/s、1.45 GB/s和 1.48 GB/s，与 DMA 模式相比非常低。
- 全局离散访问的延迟和同时发起数据读取请求的从核个数直接相关：当从核个数 n 小于 2 时，每个从核的访存延迟为 134-138ns（194-200 个时钟周期）；当从核个数大于 2 时，每个从核的访存延迟为 58 × n ns（84 × n 个时钟周期）。

寄存器通信

- 片上寄存器通信提供了从核阵列网络上 CPE 之间点对点/广播 256 位数据的通信方式直接的寄存器通讯只允许在同一行或同一列内的从核之间，遵循包含先进先出（First-In First-Out, FIFO）发送/接收缓冲区的匿名生产者-消费者协议，也即发送指令是异步的
- 点对点模式聚合带宽: 32/2.33 字节每时钟周期 × 1.45 GHz × 32 从核对每核组 × 4 核组 =2549 GB/s (平均来说, 接受方每隔 1 + t1 +t23= 2.33 个时钟周期可以接收到 32 字节的数据)。
- 广播模式聚合带宽: 32/2.33 字节每时钟周期 × 1.45 GHz × 7 条通信路径每次广播 × 8 行/列每核组 × 4 核组 = 4461 GB/s。

访存性能

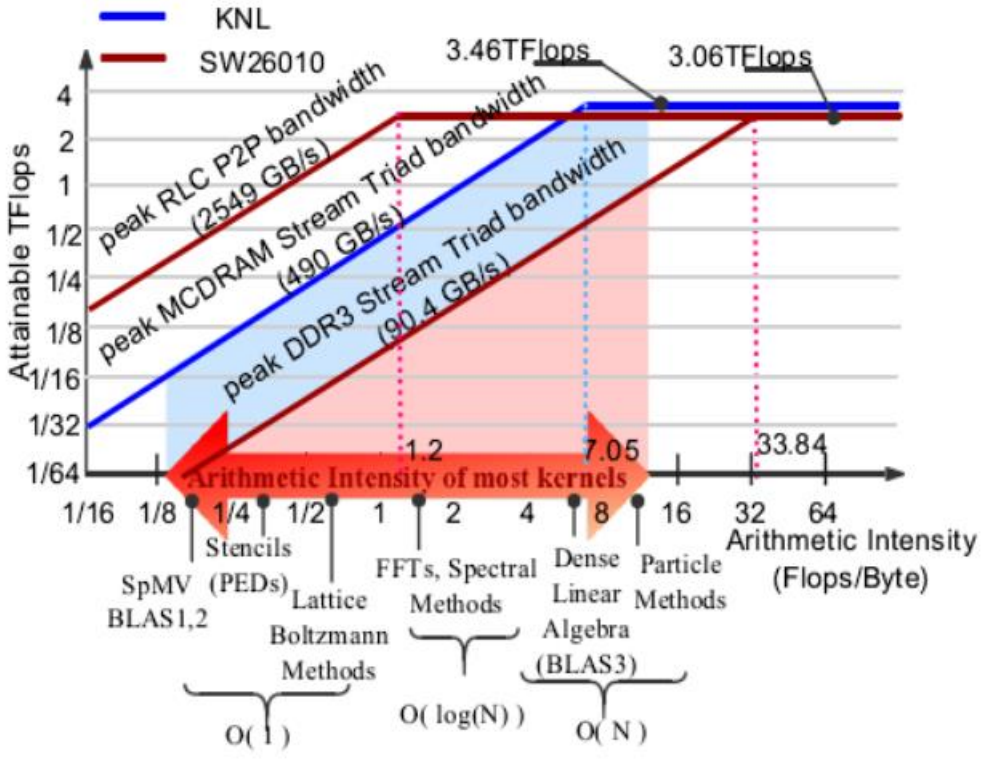


图 3–10 SW26010、KNL 架构的 Roofline model 以及常见科学计算核心的计算强度

- 主存带宽实测值仅为 90.4 GB/s(22.6GB/s 每核组 ×4 核组), SW26010 处理器的浮点运算与访存带宽比高达 33.84，这一值远高于传统处理器（一般为 10），导致绝大多数应用将受限于访存。因此，为在 SW26010 上获得较高的浮点运算能力，需要对应用的计算强度进行优化，同时降低对内存带宽的需求

总结

- 1、自研指令集sw64，可以保证技术自主可控，但这种独有的指令集也带来了不便，所有的软件组件都需要为独有的指令集重新编译，并为Mesh架构和SPM架构进行优化
- 2、4pe+64x4ce的结构，有效提高了芯片的计算能力
- 3、SPM将复杂性抛给了应用程序，避免了大量核心环境下的同步问题
- 4、多数应用受限于访存，每个申威SW26010提供4个128位DDR3-2133内存控制器，不支持硬件多线程，在持续数据供给方面比较很欠缺

谢谢