中国科学院近代物理研究所
Institute of Modern Physics, Chinese Academy of Sciences

GPU高性能加速器仿真程序设计

报告人：田园
中国科学院近代物理研究所
直线加速器中心

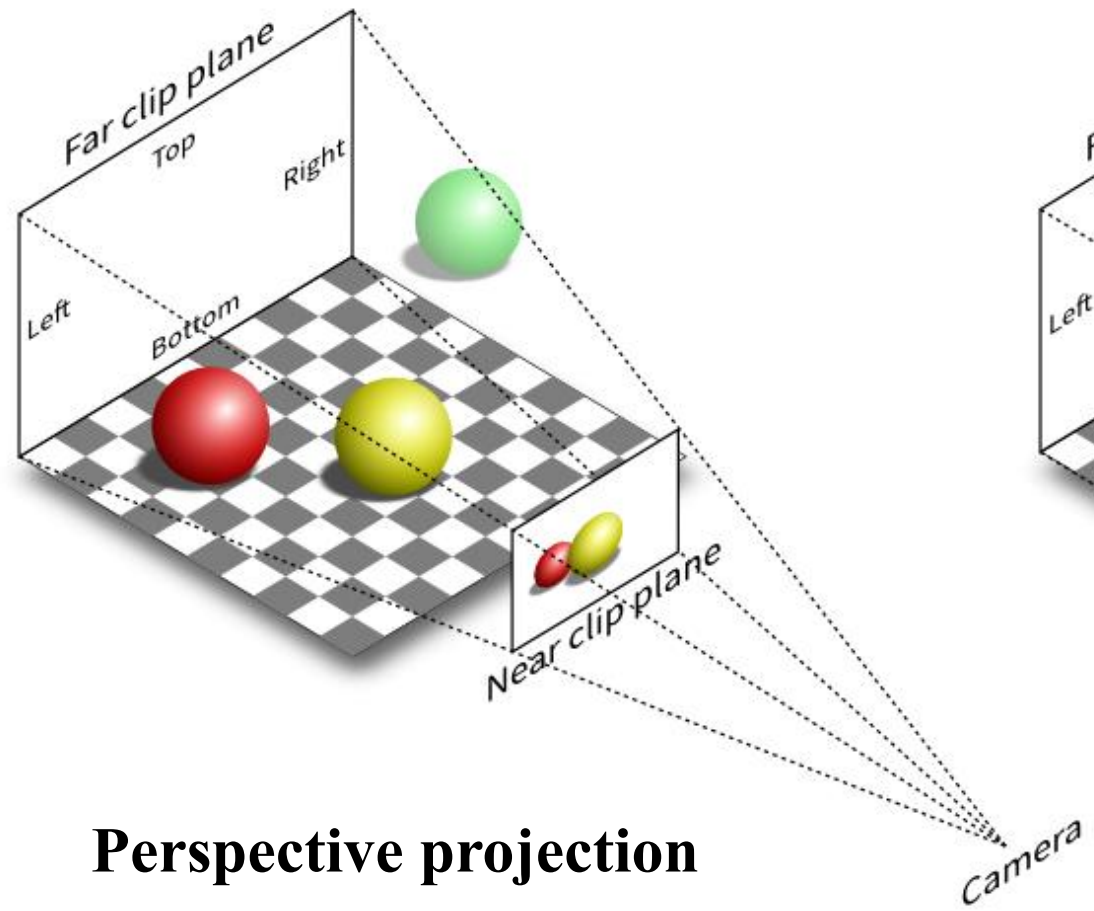
主要内容

- 图形处理器GPU与CUDA
- 基于GPU的AVASX加速器多粒子程序设计与优化
- AVASX的测试与验证

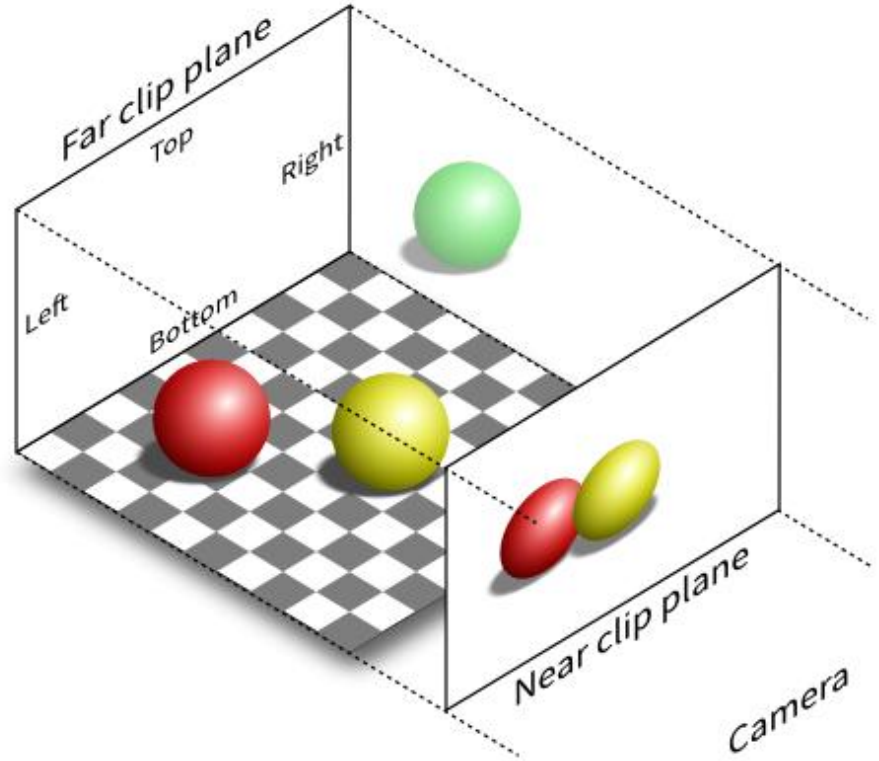
1. 图形处理器GPU与CUDA

• Graphics Processing Units，图形处理器

- GPU诞生之初的作用：将数字化的三维立体图形经过顶点计算、着色计算后，最终获得由二维离散像素组成的图像，并显示在显示设备上。



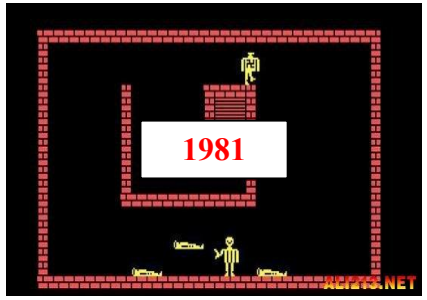
Perspective projection



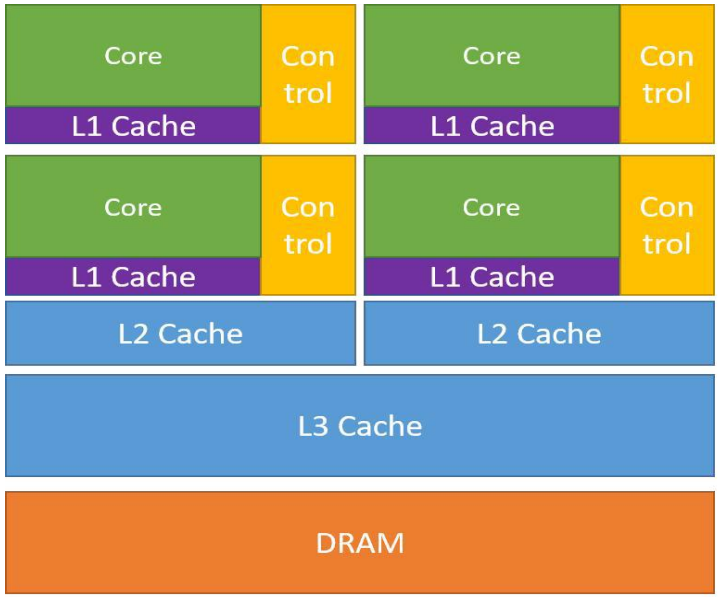
Orthographic projection

• 为什么GPU计算速度比CPU要快?

- 对游戏画质的追求，推动了GPU的快速发展，如果处理不够快，就只能看幻灯片~

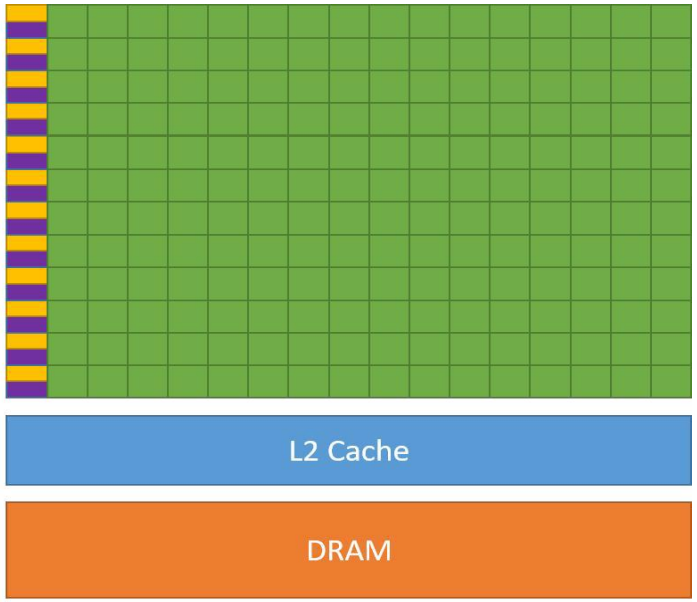


- 低计算密度;
- 复杂逻辑控制;
- 大缓存;
- 低访存延迟;



CPU

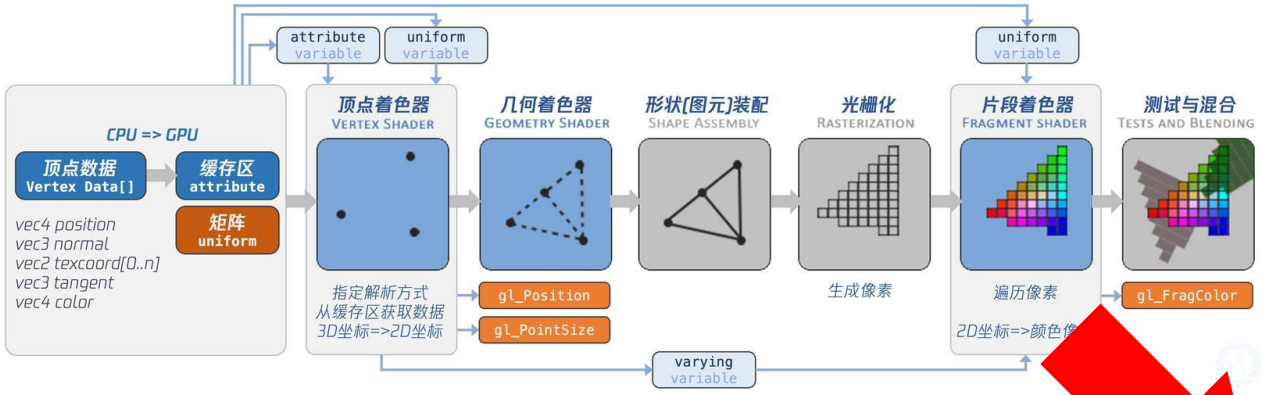
VS.



GPU

- 高计算密度;
- 高数据吞吐量;
- 深流水线;
- 高访存延迟;

• 从管线式GPU到统一渲染GPU

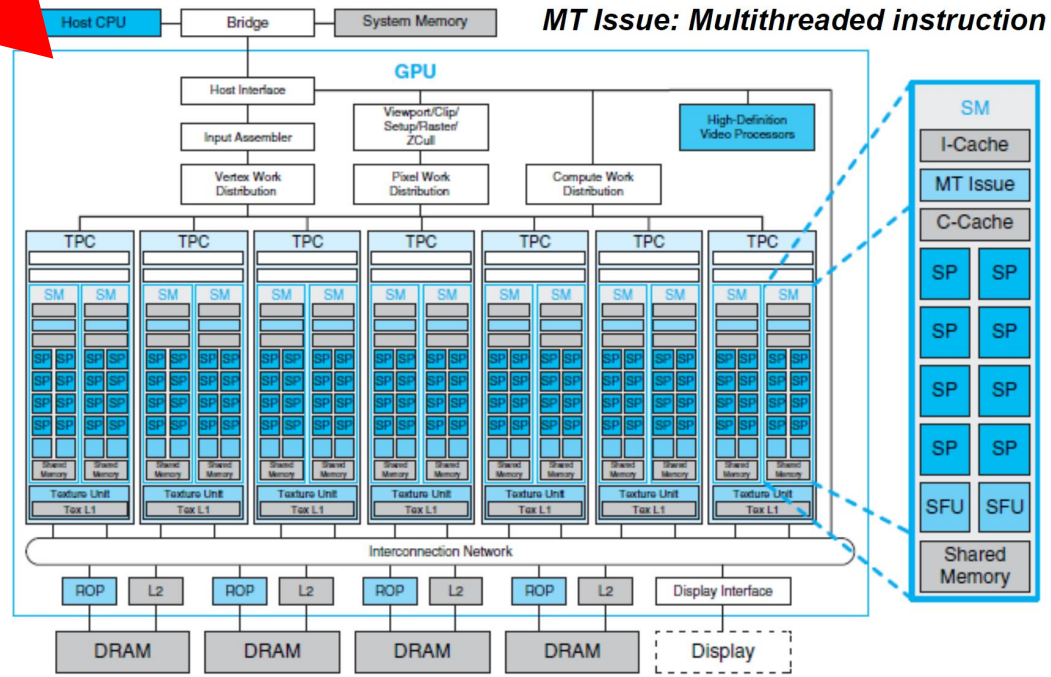


管线式GPU架构：

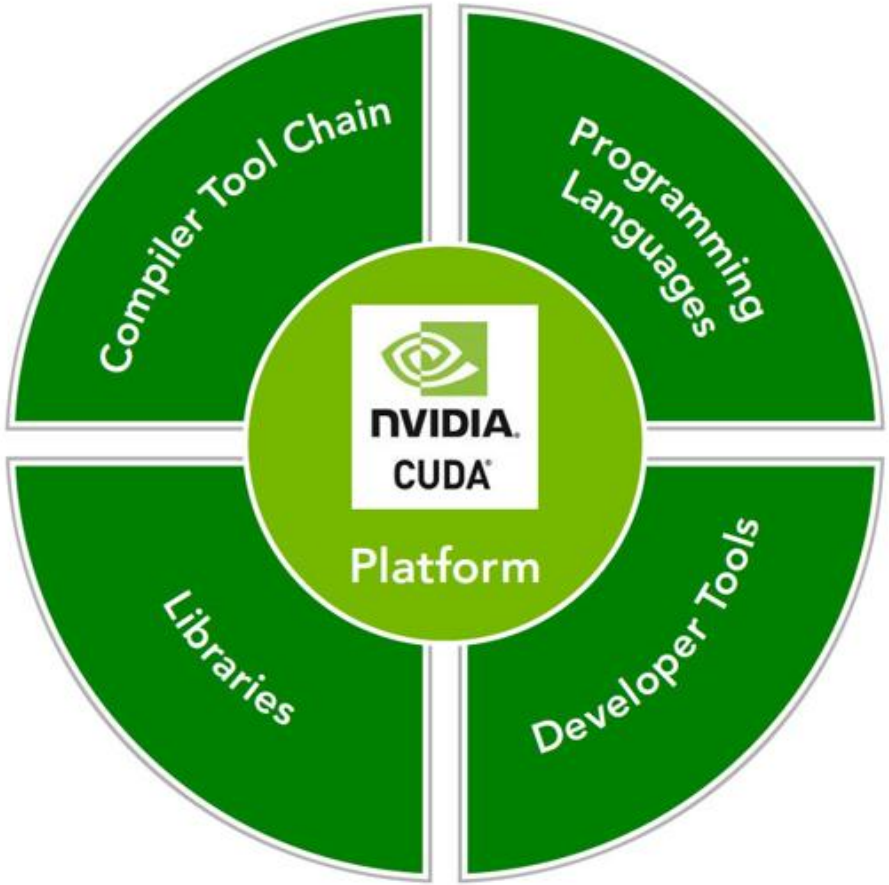
- 1999 ~ 2005；
- 硬件T&L（Transforming & Lighting）；
- 顶点着色器与像素着色器分离；
- 不灵活，难以平衡顶点运算需求和像素运算需求；

统一渲染架构GPU：

- 2006 ~ 至今；
- 硬件T&L（Transforming & Lighting）；
- 不再分离顶点着色器与像素着色器，统一为shader着色器；
- 非常灵活，shader可编程，既能用于顶点运算也能用于像素运算；
- GPU可编程、可科学计算的基础；



• Compute Unified Device Architecture，CUDA



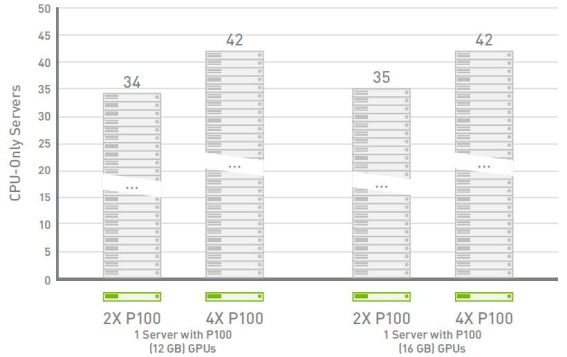
LAMMPS Performance Equivalence

Single GPU Server vs Multiple CPU-Only Servers

CPU Server: Dual Xeon E5-2690 v4 @ 2.6 GHz, GPU Servers: Same as CPU server with NVIDIA® Tesla® P100 for PCIe (12 GB or 16 GB) | NVIDIA CUDA® Version: 8.0.44 | Dataset: EAM | To arrive at CPU node equivalence, we used measured benchmarks with up to 8 CPU nodes and linear scaling beyond 8 nodes.

AMBER Performance Equivalence

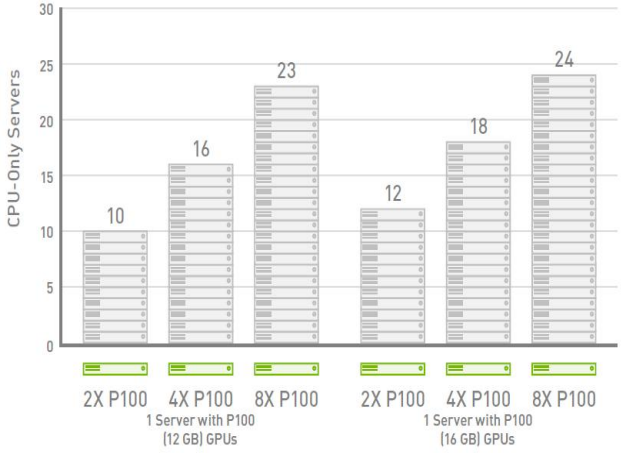
Single GPU Server vs Multiple CPU-Only Servers



CPU Server: Dual Xeon E5-2690 v4 @ 2.6 GHz, GPU Servers: Same as CPU server with NVIDIA® Tesla® P100 for PCIe (12 GB or 16 GB) | NVIDIA CUDA® Version: 8.0.44 | Dataset: GB-Myoglobin | To arrive at CPU node equivalence, we used measured benchmarks with up to 8 CPU nodes and linear scaling beyond 8 nodes.

VASP Performance Equivalence

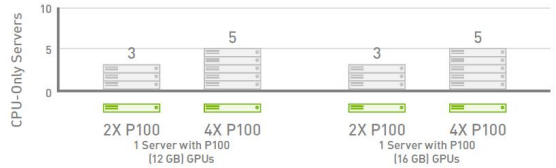
Single GPU Server vs Multiple CPU-Only Servers



PU Server: Dual Xeon E5-2690 v4 @ 2.6 GHz, GPU Servers: Same as CPU server with NVIDIA® Tesla® P100 for PCIe (12 GB or 16 GB) | NVIDIA CUDA® Version: 8.0.44 | Dataset: B_hR105| To arrive at CPU node equivalence, we used measured benchmarks with up to 8 CPU nodes and linear scaling beyond 8 nodes.

SIMULIA Abaqus Performance Equivalency

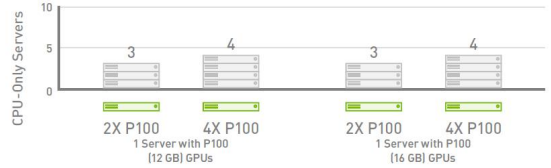
Single GPU Server vs Multiple CPU-Only Servers



CPU Server: Dual Xeon E5-2690 v4 @ 2.6GHz, GPU Servers: Same as CPU server with NVIDIA® Tesla® P100 for PCIe (12 GB or 16 GB) | NVIDIA CUDA® Version: 7.5 | To arrive at CPU node equivalence, we use measured benchmark with up to 8 CPU nodes. Then we use linear scaling to scale beyond 8 nodes.

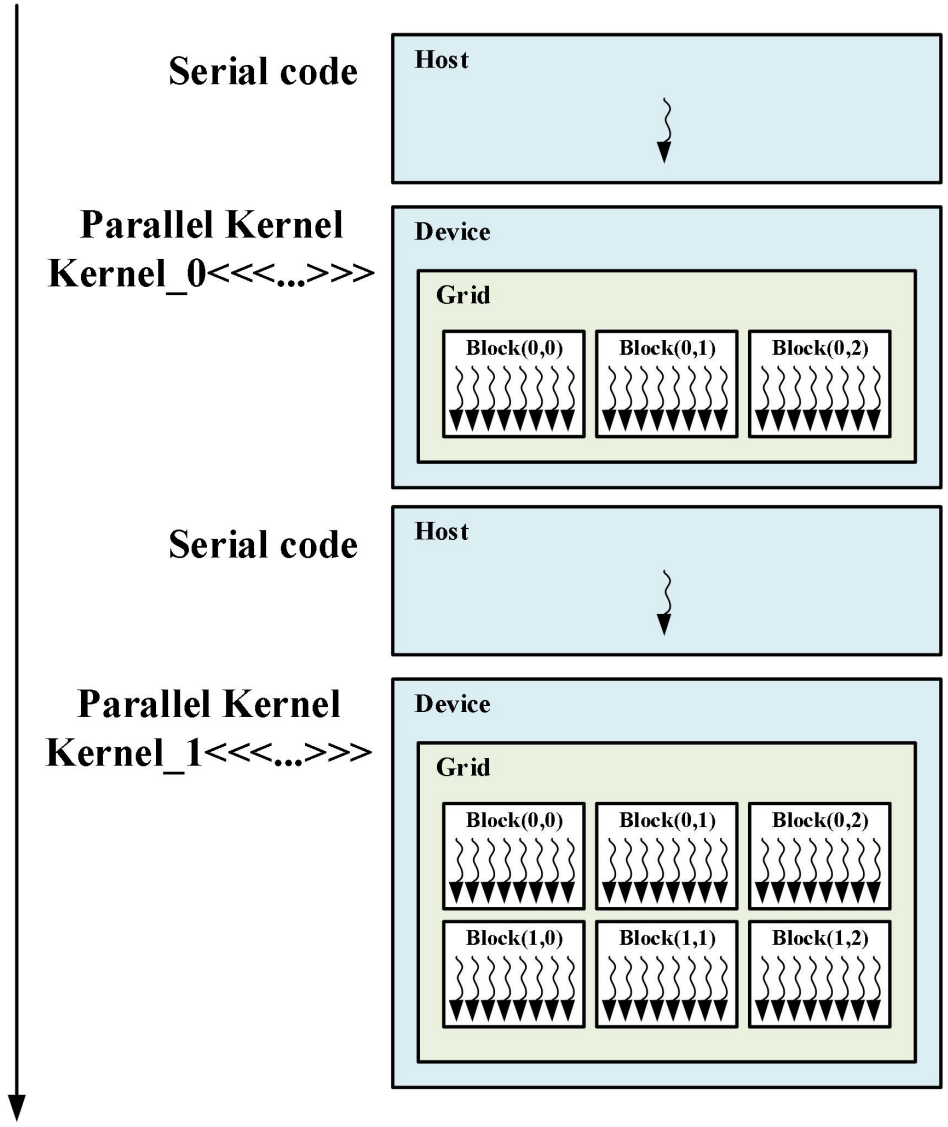
ANSYS Fluent Performance Equivalency

Single GPU Server vs Multiple CPU-Only Servers

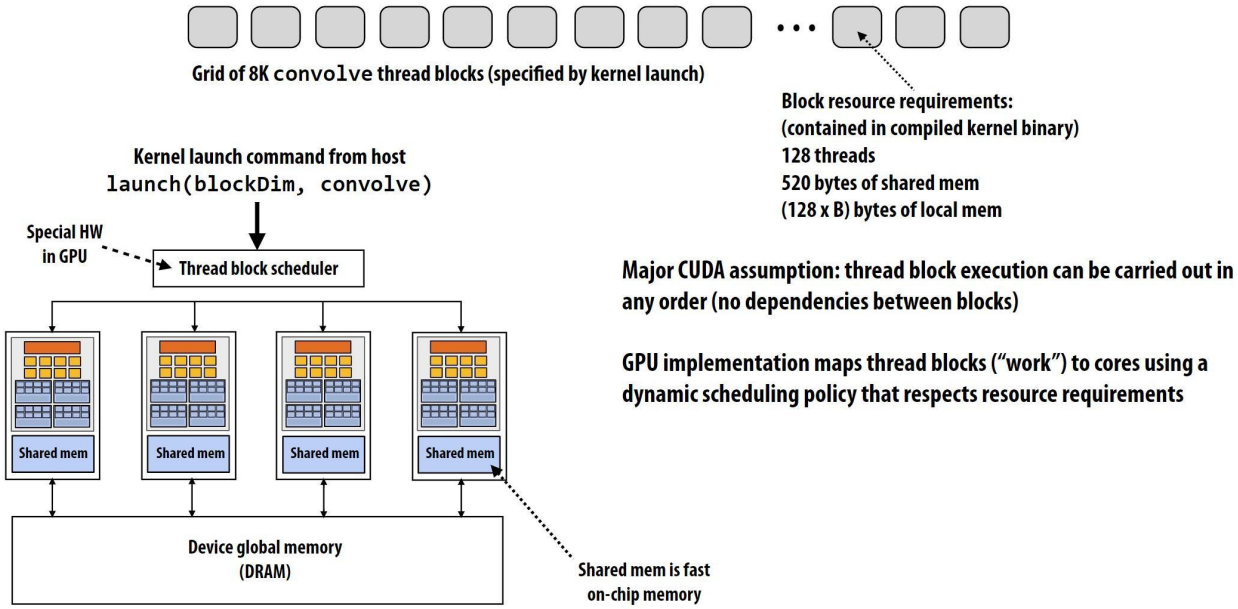


CPU Server: Dual Xeon E5-2690 v4 @ 2.6GHz, GPU Servers: Same as CPU server with NVIDIA® Tesla® P100 for PCIe (12 GB or 16 GB) | NVIDIA CUDA® Version: 6.0 | Dataset: Water Jacket | To arrive at CPU node equivalence, we use measured benchmark with up to 8 CPU nodes. Then we use linear scaling to scale beyond 8 nodes.

• CUDA的线程组织和执行

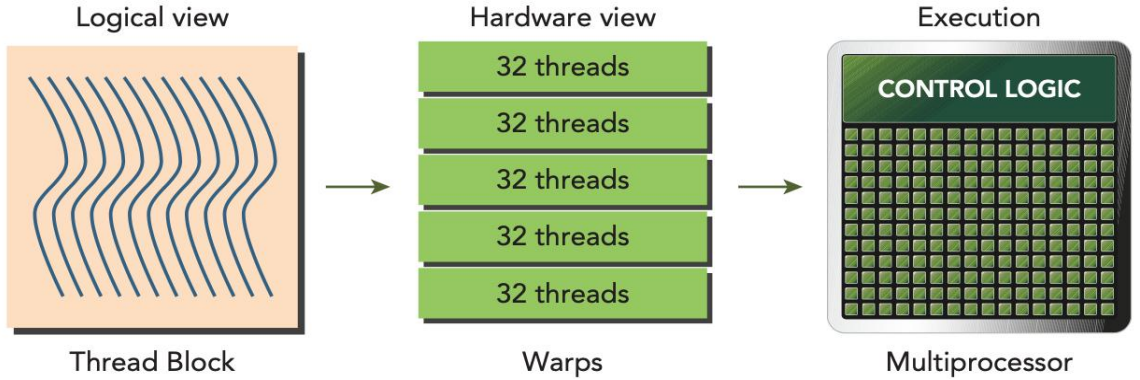


CUDA thread-block assignment



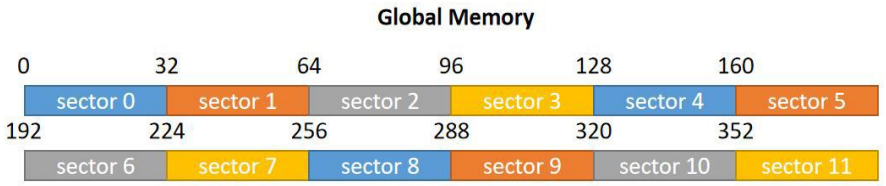
Major CUDA assumption: thread block execution can be carried out in any order (no dependencies between blocks)

GPU implementation maps thread blocks ("work") to cores using a dynamic scheduling policy that respects resource requirements

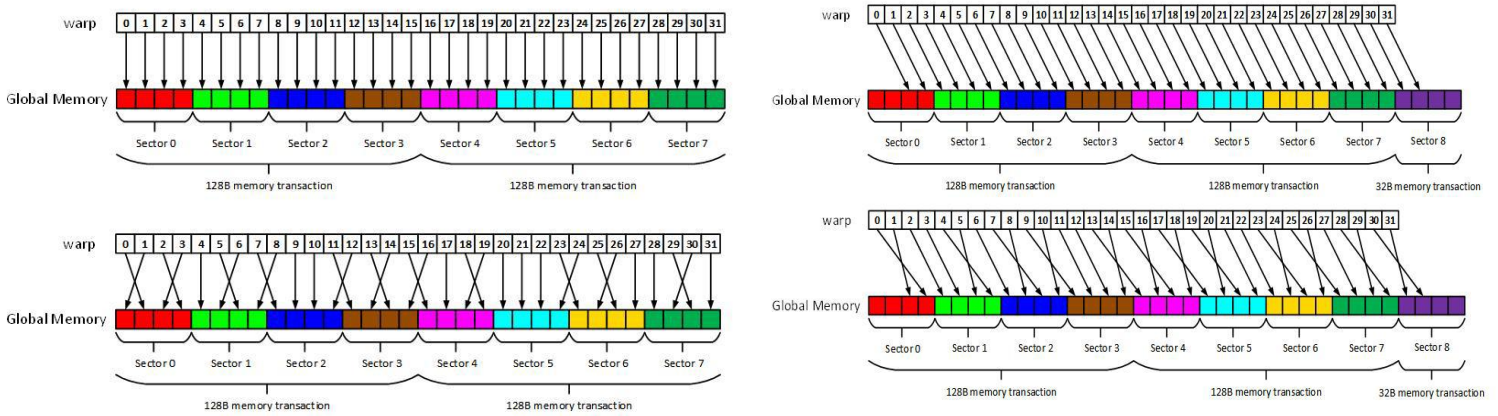


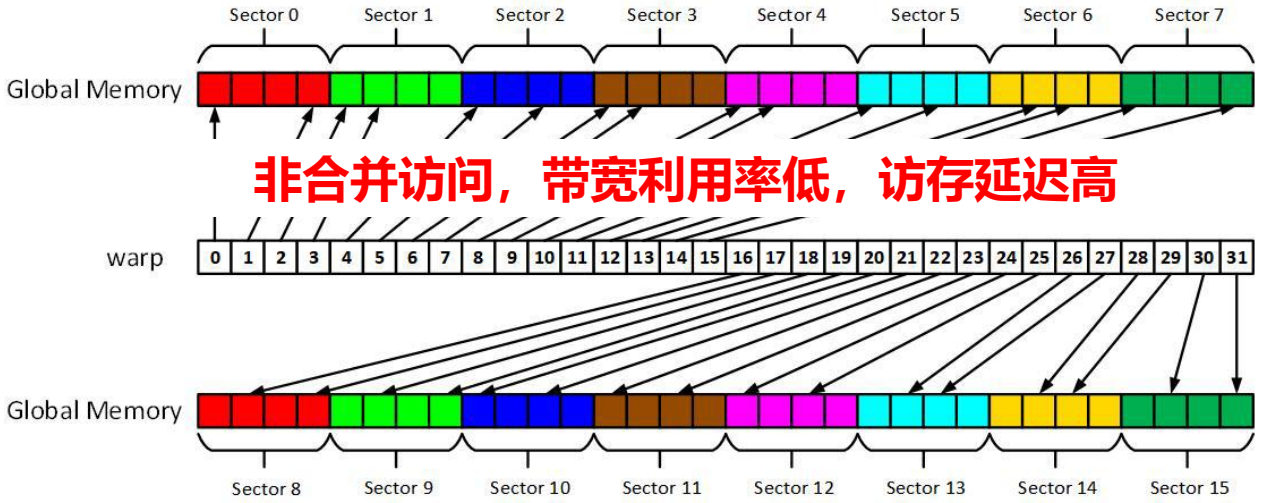
· CUDA的内存系统

内存系统	存储位置	是否缓存	访问延迟	作用域
寄存器	on chip	N/A	极低	thread
本地内存	off chip	no	高	thread
共享内存	on chip	N/A	低	block
常量内存	off chip	yes	高	grid
全局内存	**off chip**	**yes**	**高**	**grid**
纹理内存	off chip	yes	高	grid



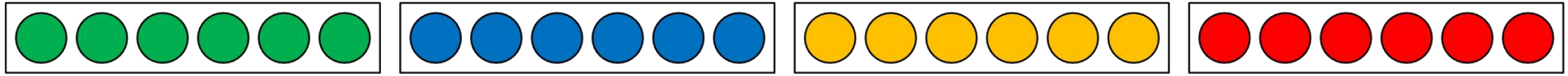
全局内存合并访问，内存带宽利用率高，访存延迟小



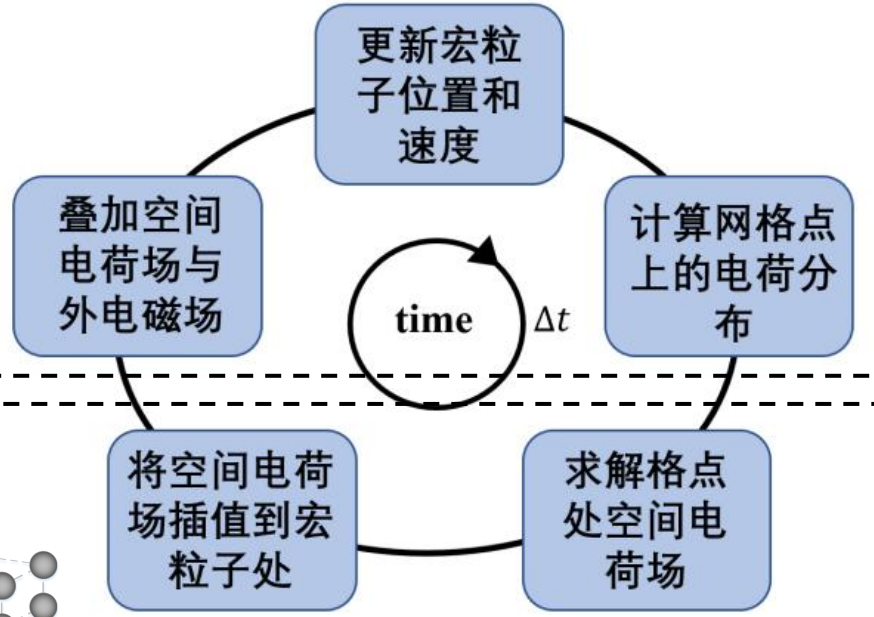


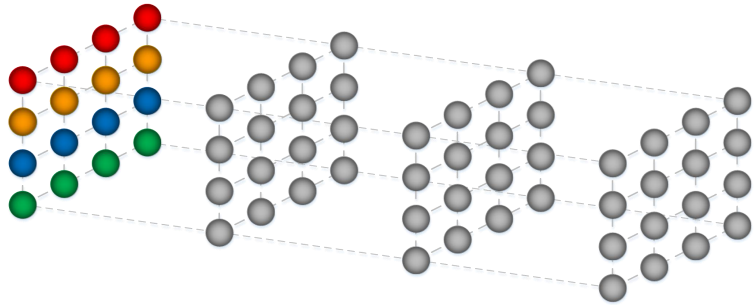
3. 基于GPU的AVASX程序设计

• GPU加速器多粒子仿真算法概述



- 一维Grid（Block在x方向排列）、一维Block（线程在x方向排列）、线程与粒子一一对应。

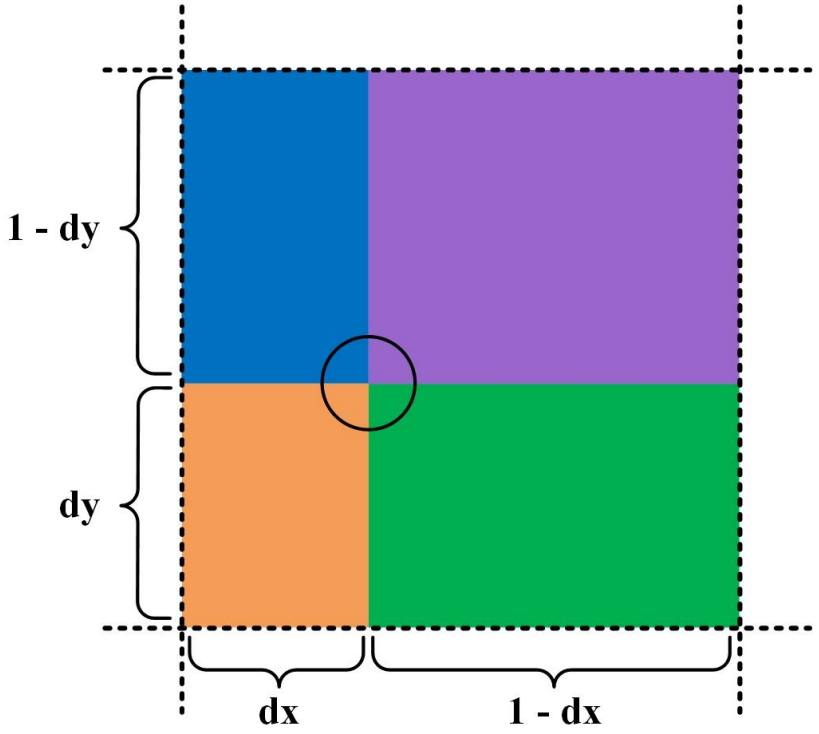


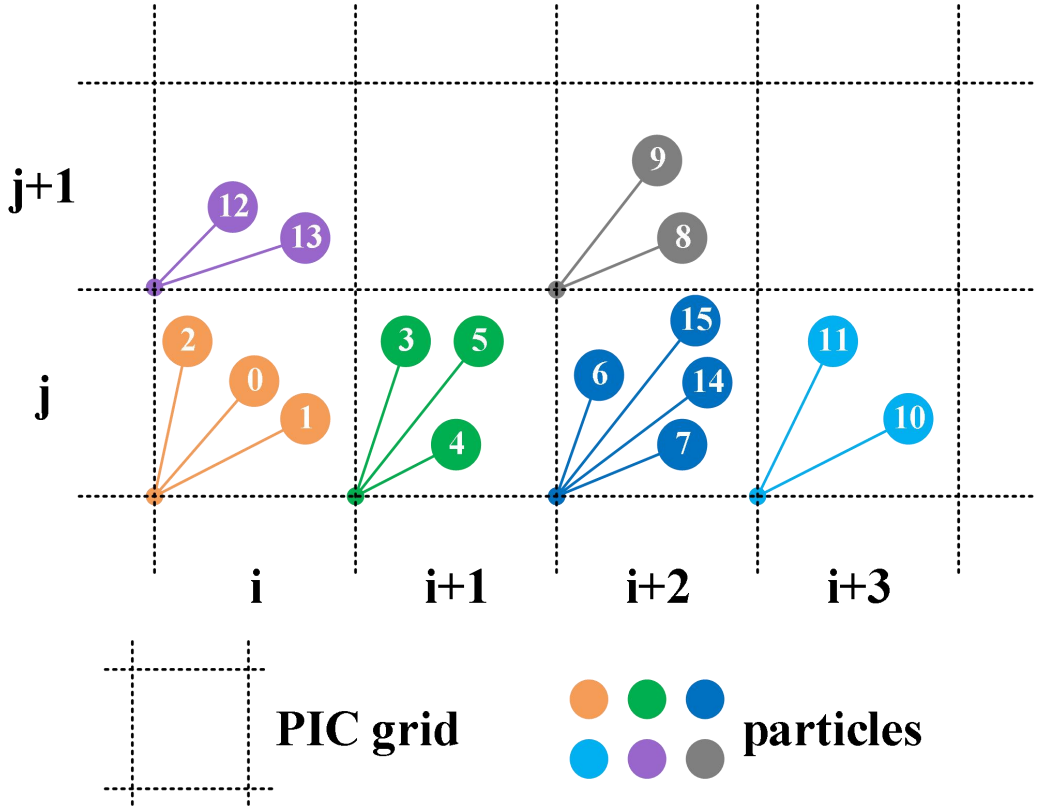


- 二维Grid、一维Block（线程在x方向排列）、线程与PIC网格点一一对应。

• 电荷沉积

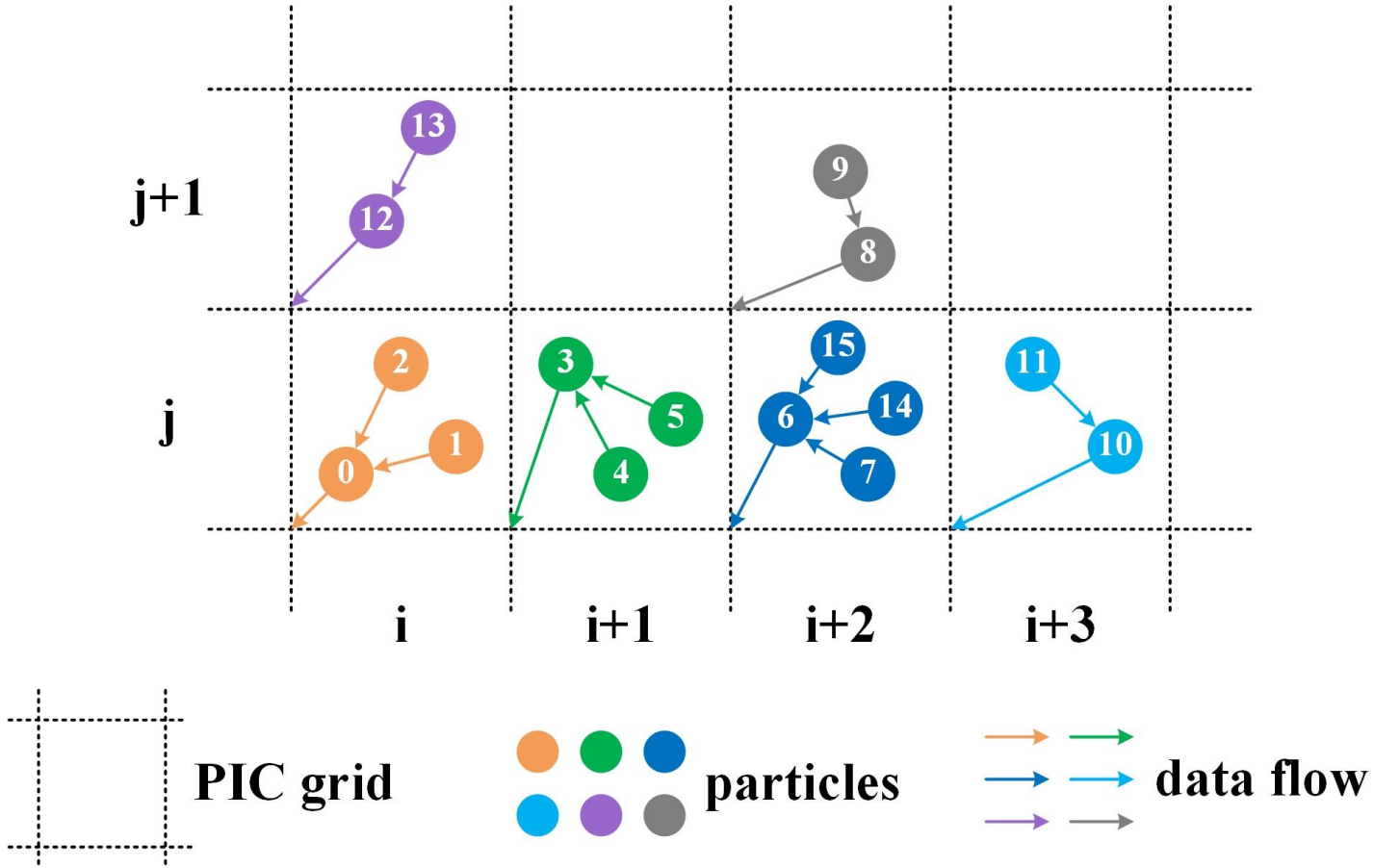
- 存在问题：在GPU并行计算模式下，多个线程将沉积电荷写入相同PIC格点，**存在内存竞争写入问题**，需要使用互斥的atomicAdd原子加法操作，大量线程执行原子操作会降低并发度（串行化执行），从而降低电荷沉积的计算性能。





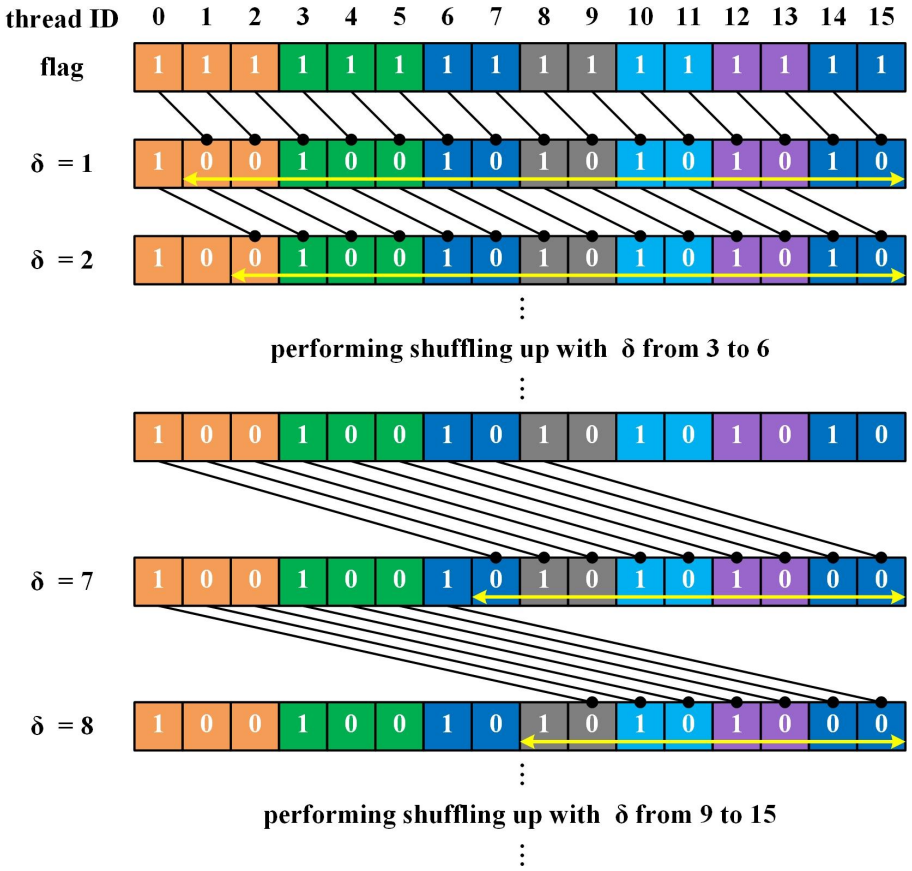
• 线程冲突优化策略1： warp线程聚合

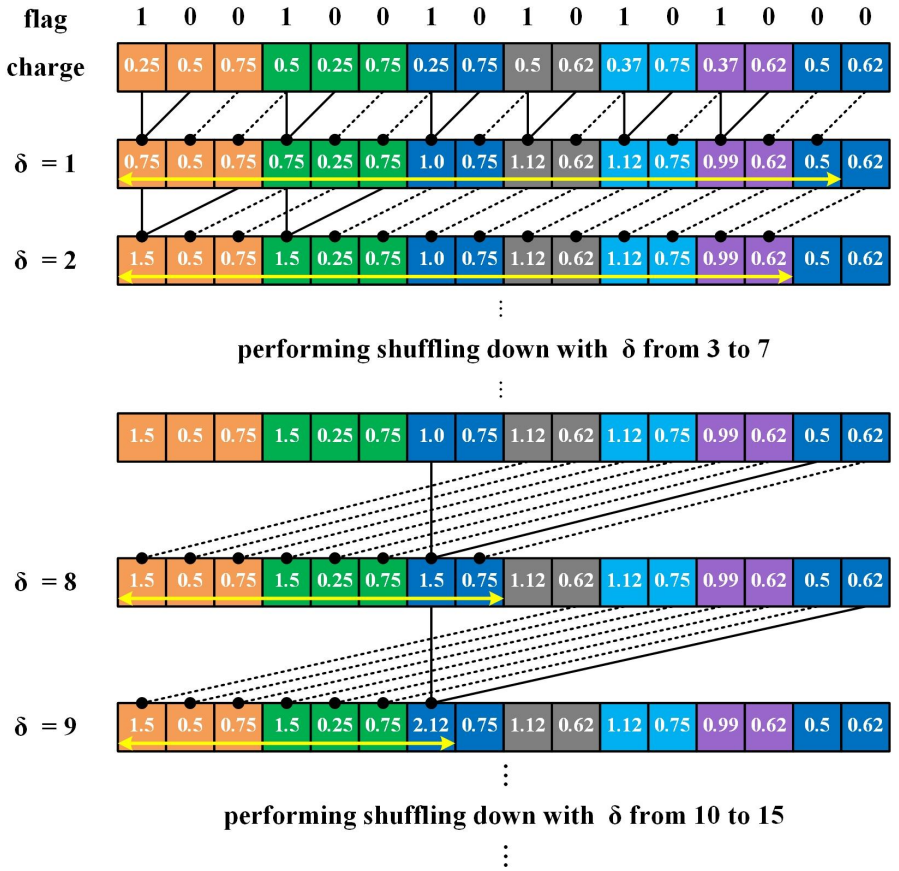
- warp线程聚合：将warp内线程按粒子所在网格分组，每组指派一个收集线程，收集组内其他线程的分配电荷，最终仅由收集线程执行原子操作，从而减少原子指令数。



• warp线程聚合：收集线程选定与电荷收集

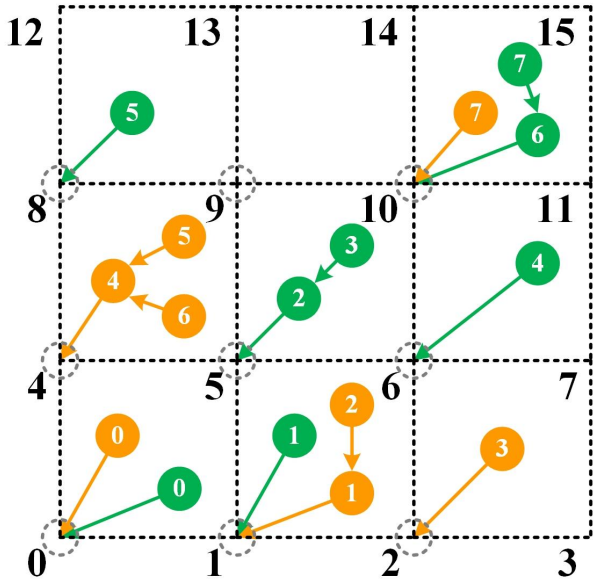
- 线程聚合实现方案：（1）使用__shfl_up_sync线程洗牌函数，在每个分组中筛选出来一个收集线程；（2）使用__shfl_down_sync洗牌函数，由每个收集线程收集同组线程数据。

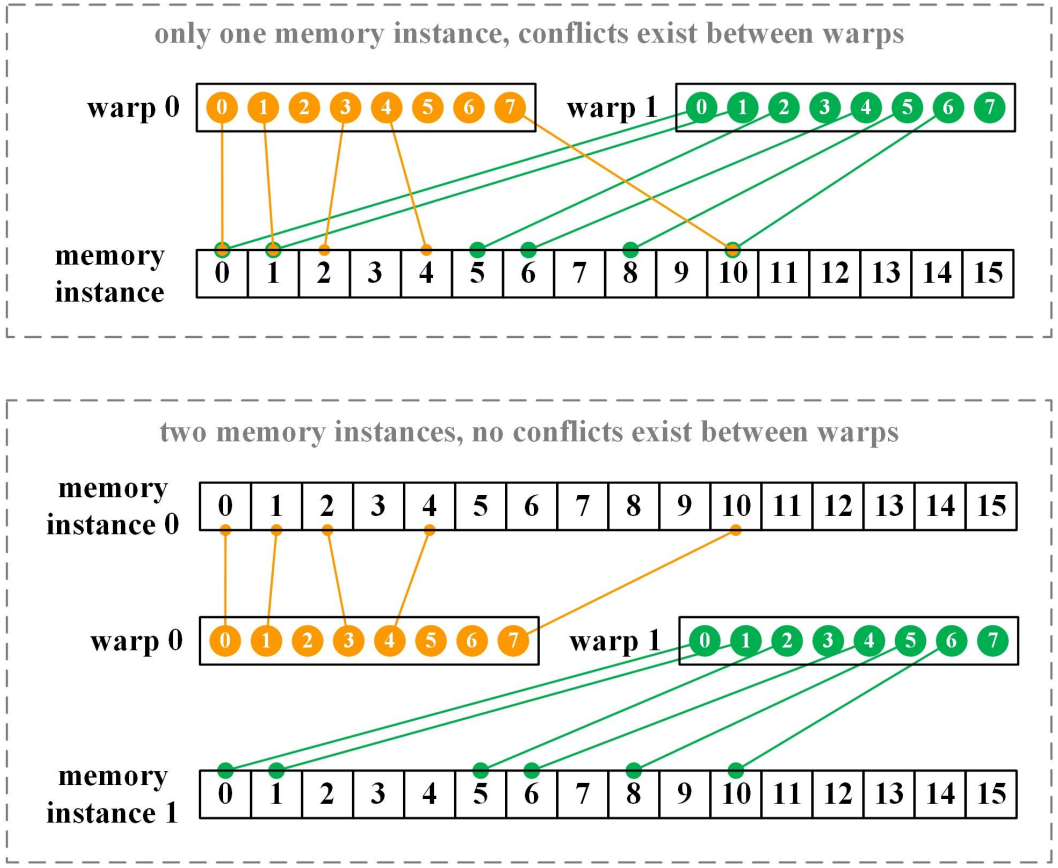




• 线程冲突优化策略2：多PIC网格内存副本

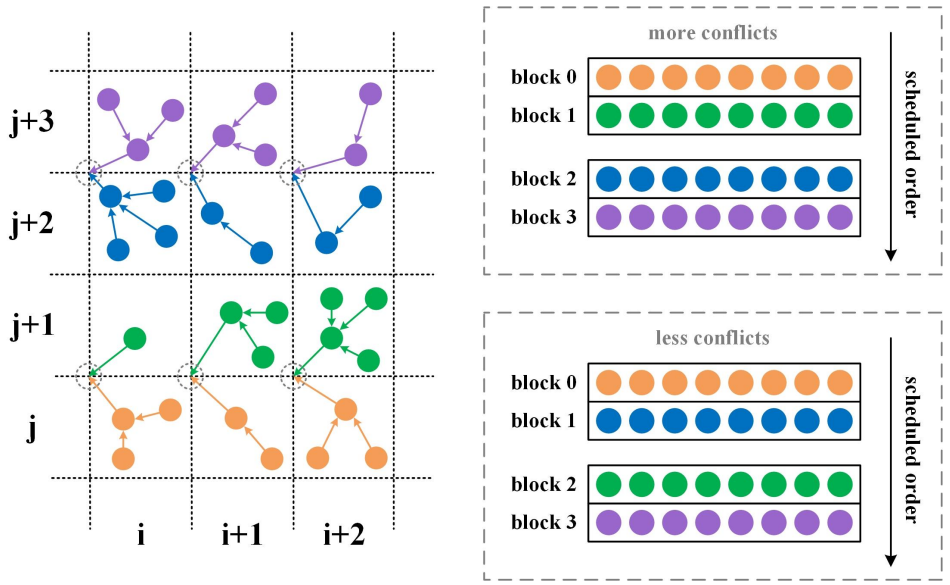
- 多PIC网格内存副本：按照block内部的warp数量，在全局内存上分配多个网格点内存副本，将不同的内存副本指派到不同的warp，从而降低warp之间的内存写入冲突。

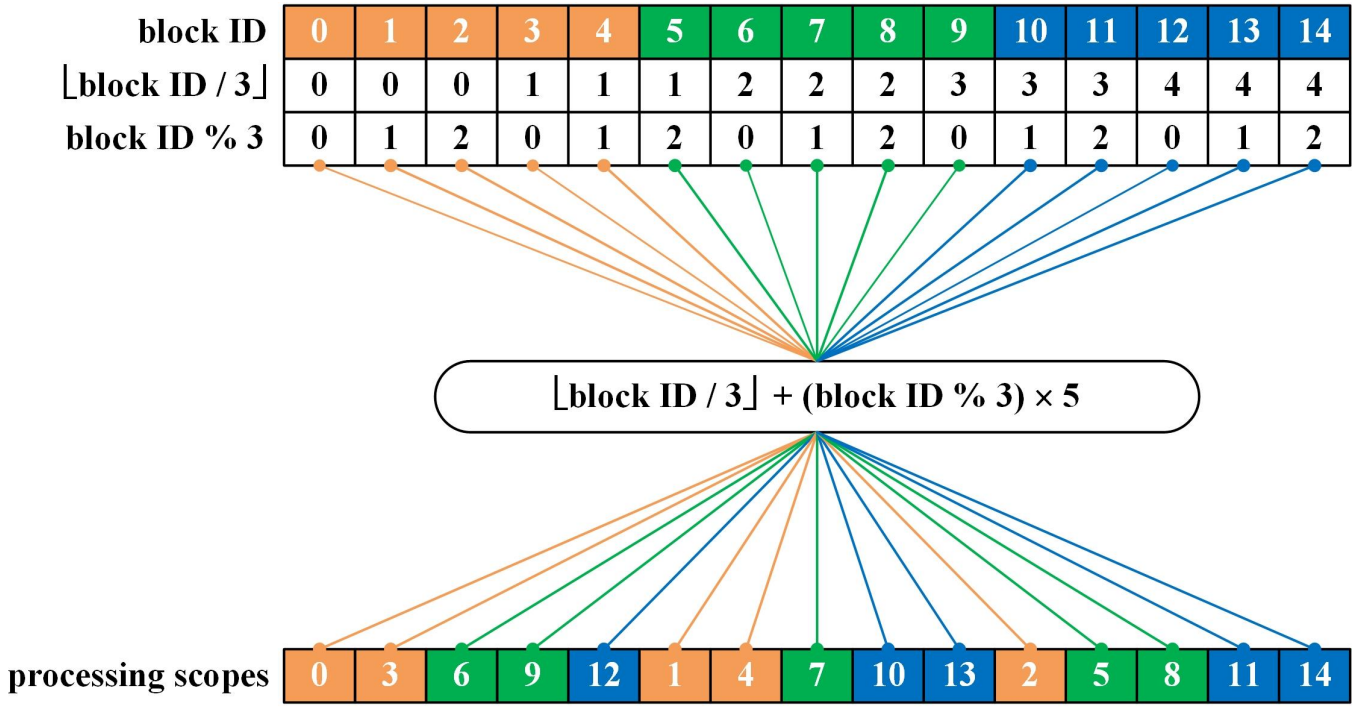




• 线程冲突优化策略3：block数据处理范围发散

- 数据处理范围发散：改变同一批次被调度的block的数据处理范围，使得被调度block之间的粒子不再位于相邻网格，从而降低内存写入冲突。

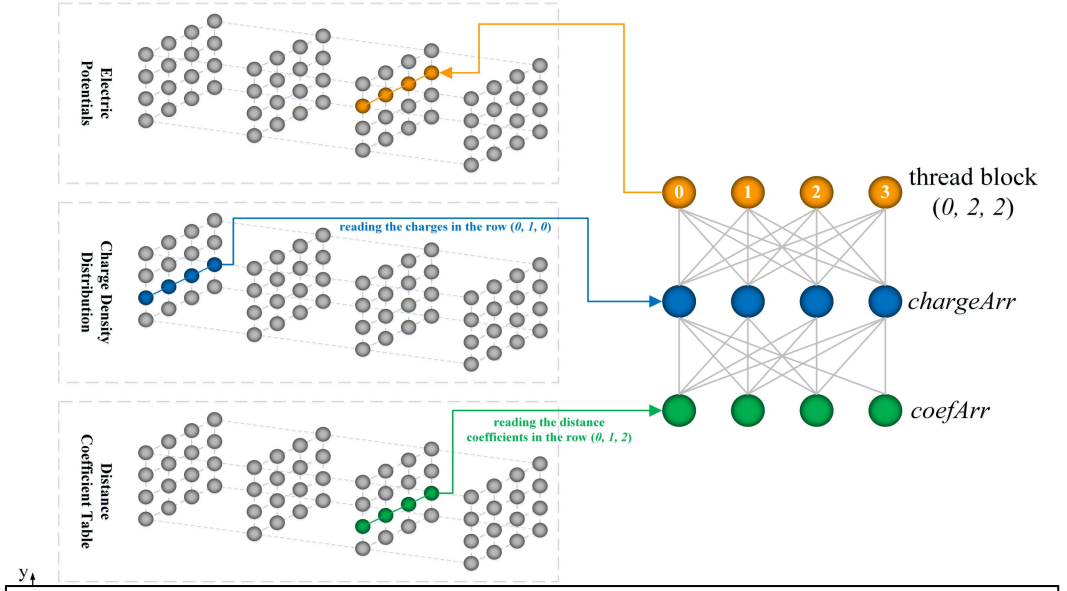




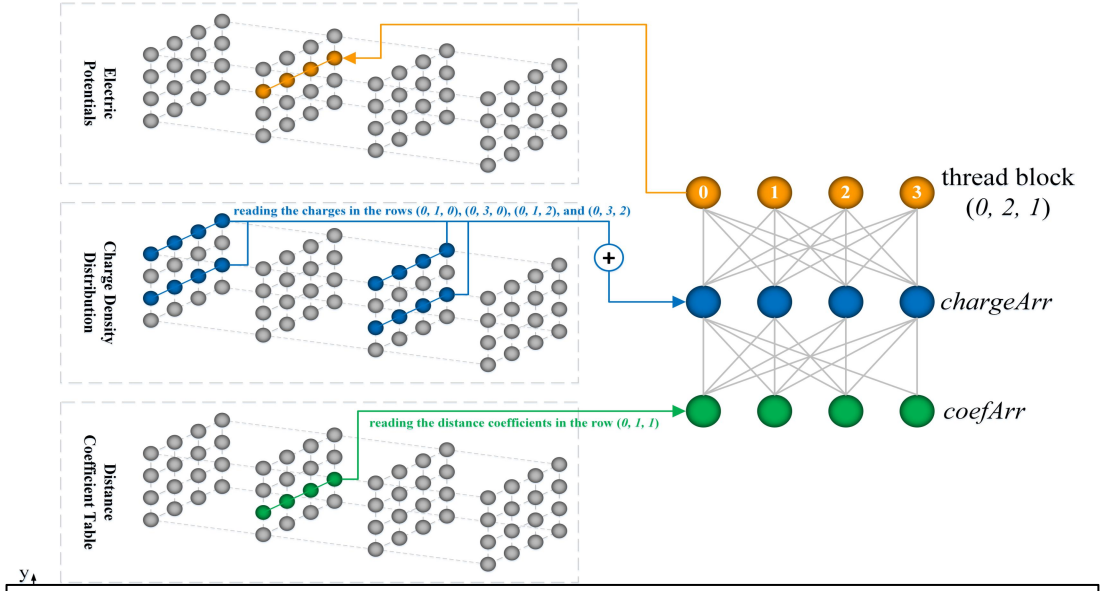
• 空间电荷场求解：PICNIC

$$\varphi_{(u,v,w)} = \sum_{i=0}^{N_x}\sum_{j=0}^{N_y}\sum_{k=0}^{N_z}\frac{\kappa q_{(i,j,k)}}{r_{(|u-i|,|v-j|,|w-k|)}}, \begin{cases} u \neq i \\ v \neq j \\ w \neq k \end{cases}$$

- **需求分析**：非周期边界（开放边界）的空间电荷效应求解，例如大能散仿真、多束团仿真；
- **算法依据**：根据库仑定律，任意PIC格点电势都是由其他格点电荷在该点产生的电势积分得到（如上公式所示）；
- **算法难点**：在三维模拟中，积分求解空间电荷场需要**六重循环**，时间复杂度为**O(N⁶)**，相较于FFT求解器的时间复杂度**O(NlogN)**，PICNIC算法是一种高计算负载且低计算效率的方法，严重制约了该方法在仿真中的应用；



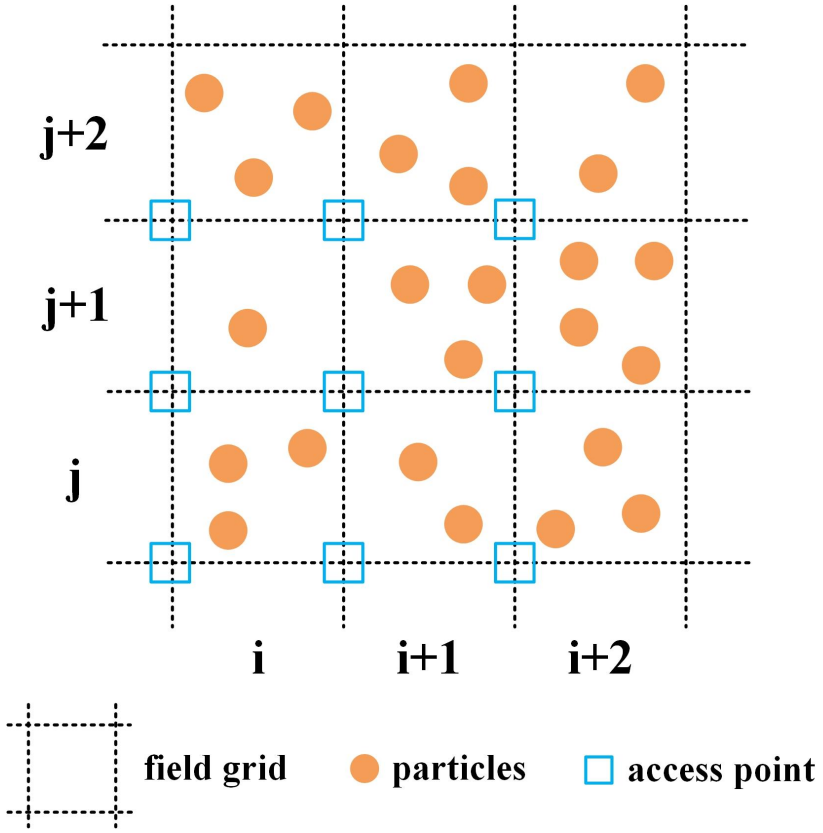
根据GPU计算架构特点，设计了并行PICNIC算法，优化GPU全局内存的合并访问，降低访存的延迟

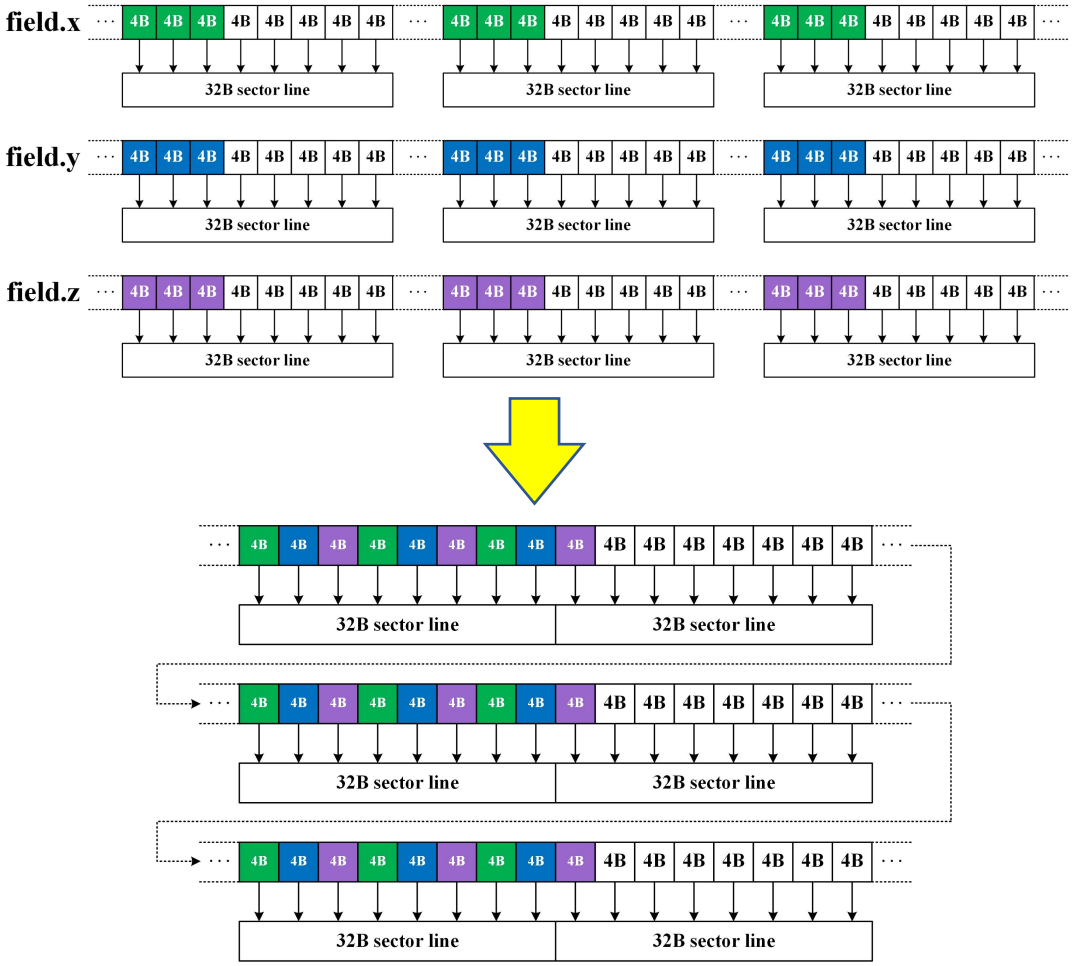


提出了计算对称性优化方法，利用GPU共享内存降低内存的冗余访问，减少循环次数

• 外场内存布局优化

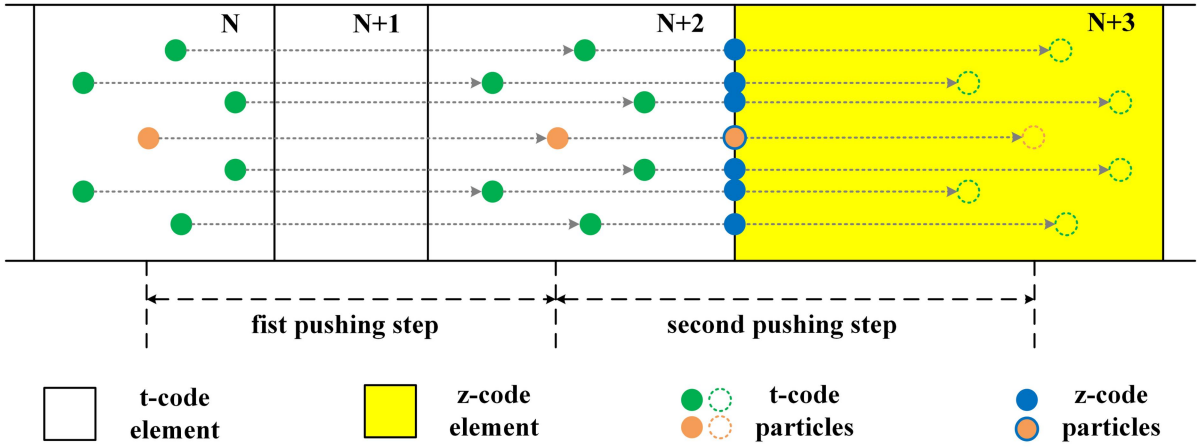
- 外场内存布局优化：使用float6 {Bx, By, Bz, Ex, Ey, Ez}自定义数据结构保存外场数据，提高全局内存访问带宽利用率。

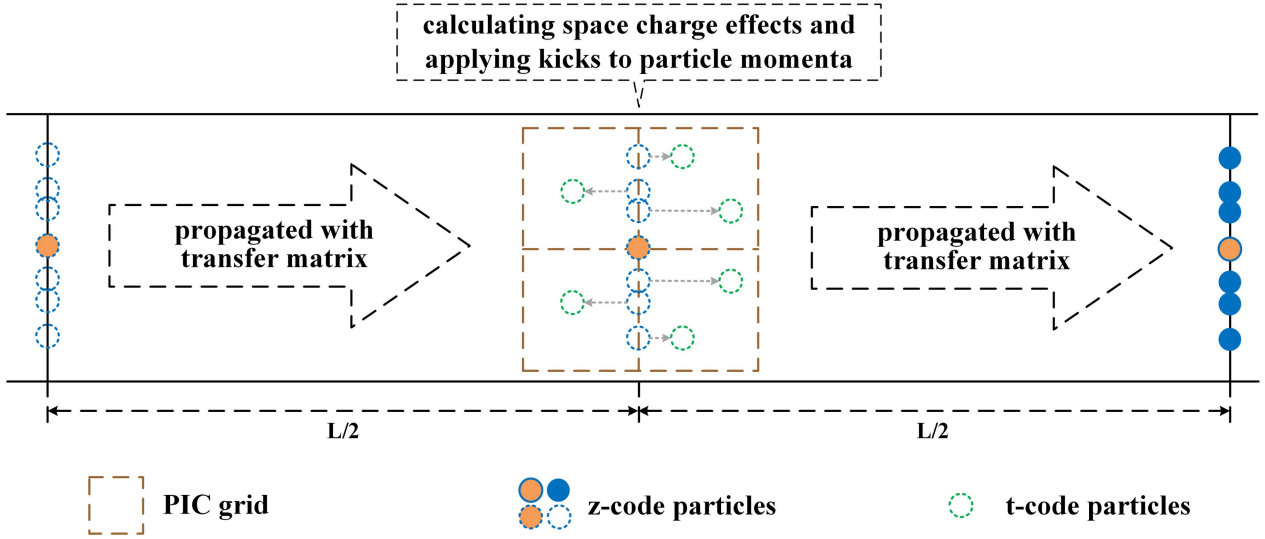




• 粒子推进

- 多种推进模式：支持全t-code推进、全z-code推进和t-code、z-code动态切换推进；
- 元件误差模拟：在t-code推进模式下，支持元件的平移、旋转等误差模拟；
- 叠加场模拟：在t-code推进模式下，支持多个外场叠加模拟；



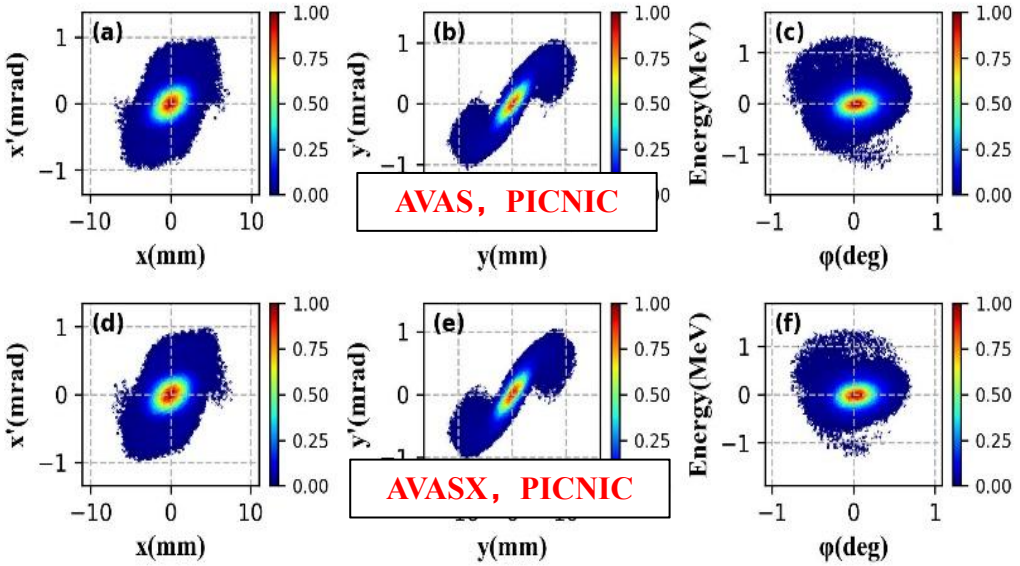


4. AVASX的测试与验证

• GPU加速的PICNIC算法校验

Method	Emittance (π·mm·mrad)			Bunch size (mm)			Energy (MeV)
	$Emit_x$	$Emit_y$	$Emit_z$	$Size_x$	$Size_y$	$Size_z$	
AVAS	0.537	0.671	0.298	1.763	3.057	0.833	630.547
AVASX	0.543	0.662	0.298	1.773	3.026	0.833	630.547

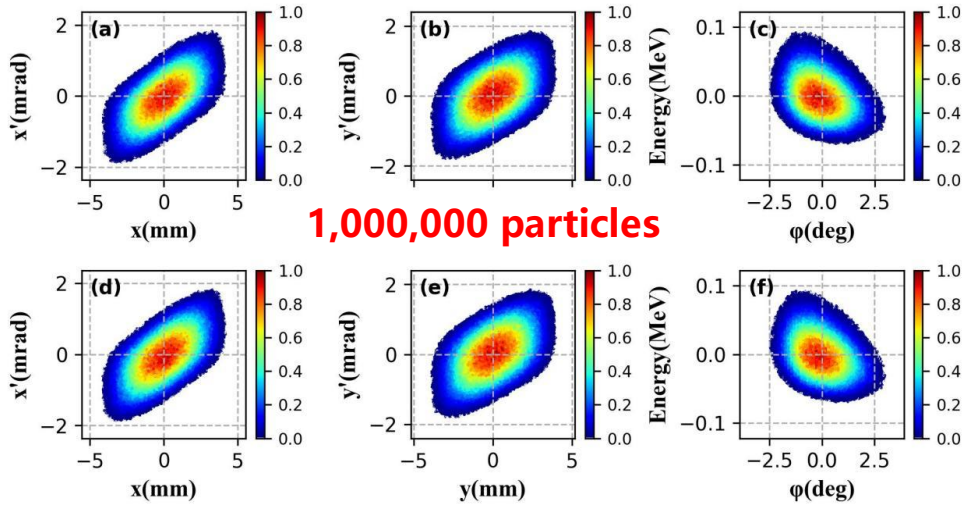
Grid point dimensions	Per step durations of PICNIC on CPUs & GPUs			
	CPU 1T (ms)	CPU 56T (ms)	GPU (ms)	GPU optimized (ms)
24 × 24 × 24	277	13	0.27	0.20
32 × 32 × 32	1546	45	0.65	0.44
48 × 48 × 48	19549	528	4.41	2.28
64 × 64 × 64	109919	3326	23.26	10.46

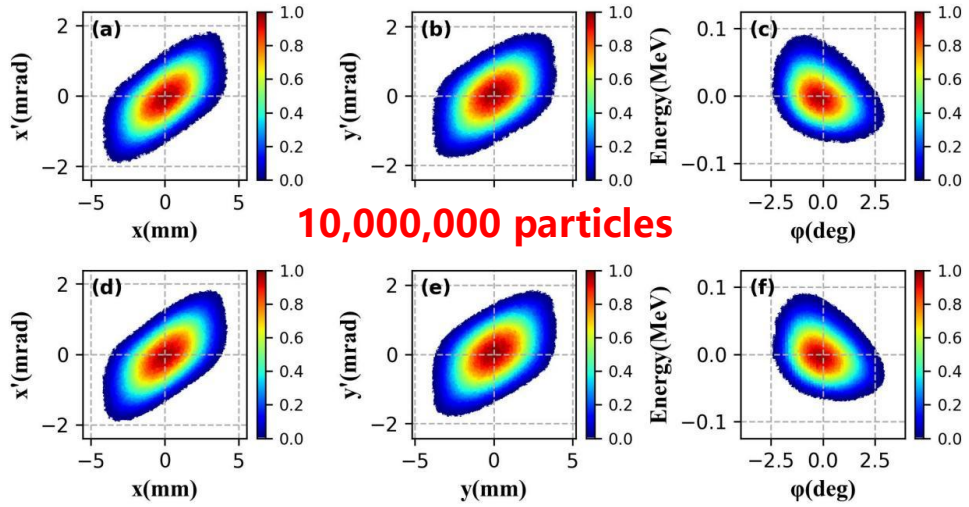


• AVASX的仿真结果校验——CAFe

- CAFe（Chinese ADS Front-end Demo Linac）仿真测试：初始束团能量1.36 MeV，流强0.27 mA，频率162.5 MHz。

Code	Particle number	Emittance (π·mm·mrad)			Bunch size (mm)			Energy (MeV)
		$Emit_x$	$Emit_y$	$Emit_z$	$Size_x$	$Size_y$	$Size_z$	
AVASX	1,000,000	0.148	0.152	0.134	1.585	1.486	0.914	16.950
	10,000,000	0.149	0.152	0.134	1.585	1.486	0.914	16.950
AVAS	1,000,000	0.148	0.152	0.134	1.584	1.485	0.914	16.950
	10,000,000	0.149	0.152	0.134	1.584	1.486	0.914	16.950

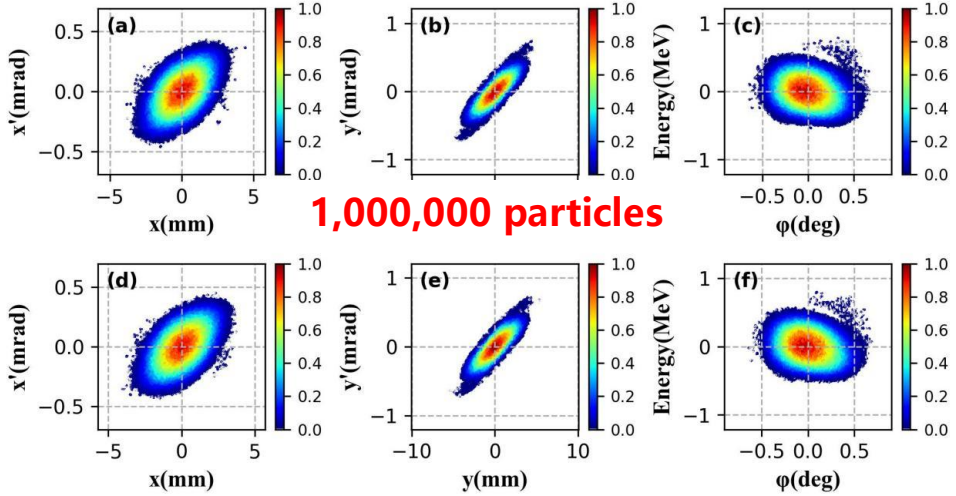


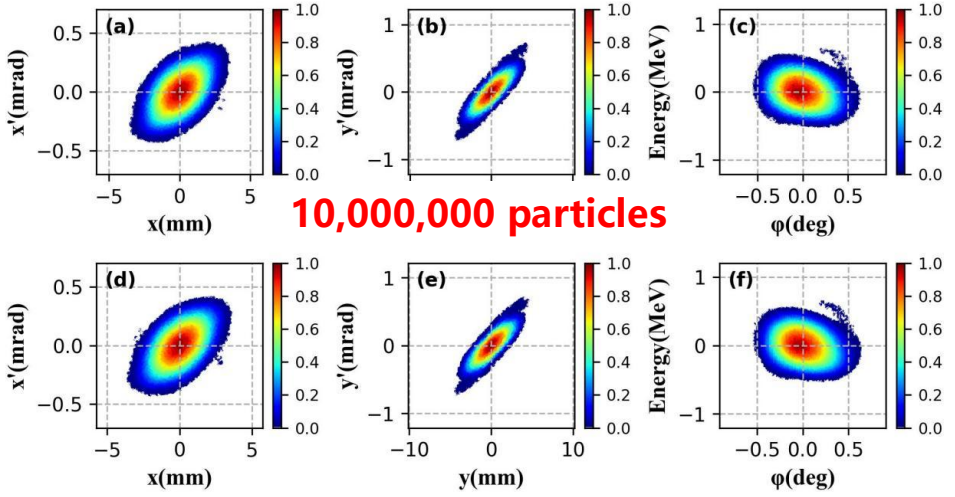


• AVASX的仿真结果校验——CiADS

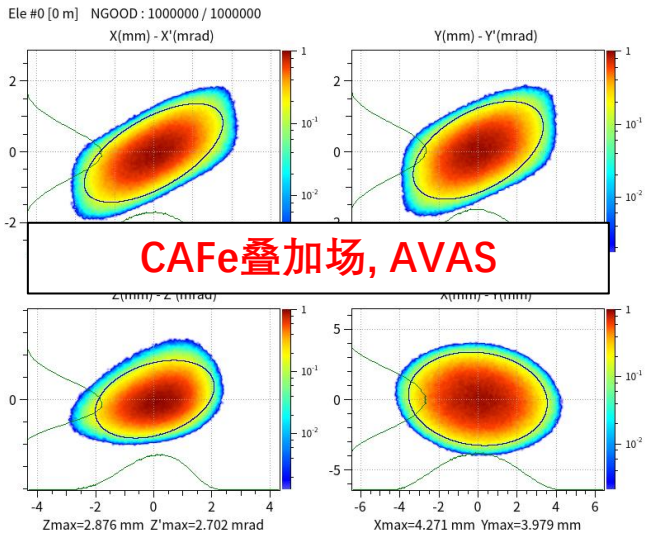
- CiADS（China initiative Accelerator Driven System）仿真测试：初始束团能量2.1 MeV，流强5.0 mA，频率162.5 MHz。

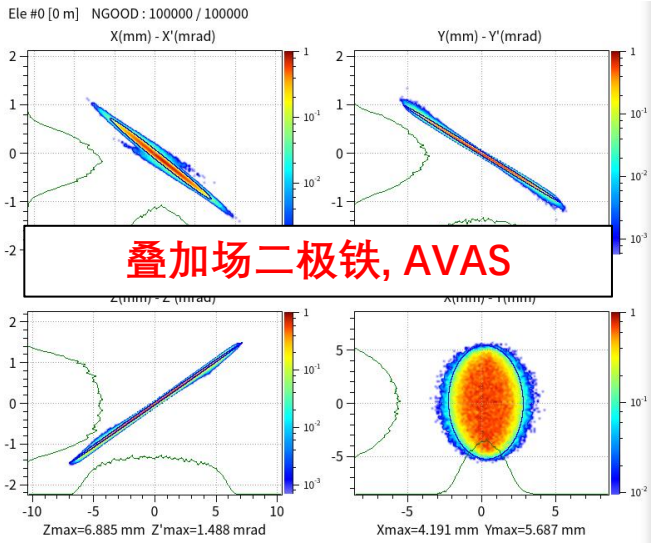
Code	Particle number	Emittance (π·mm·mrad)			Bunch size (mm)			Energy (MeV)
		$Emit_x$	$Emit_y$	$Emit_z$	$Size_x$	$Size_y$	$Size_z$	
AVASX	1,000,000	0.216	0.222	0.225	1.216	1.521	0.863	630.576
	10,000,000	0.216	0.222	0.225	1.216	1.522	0.862	630.570
AVAS	1,000,000	0.216	0.222	0.225	1.214	1.524	0.870	630.575
	10,000,000	0.216	0.222	0.224	1.215	1.525	0.869	630.574

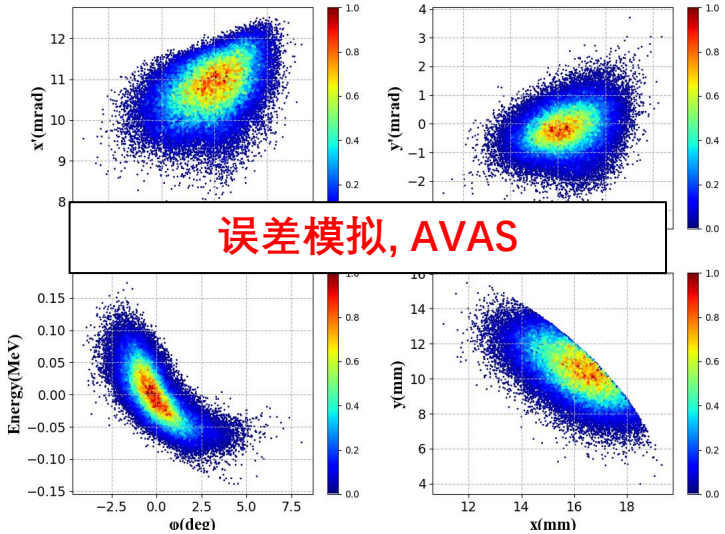


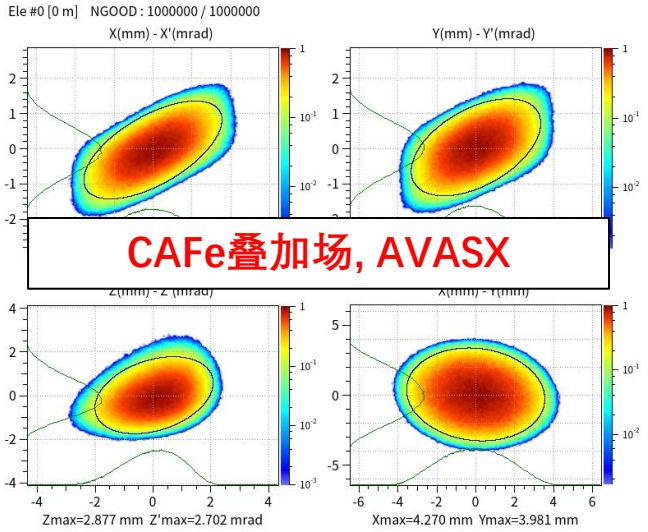


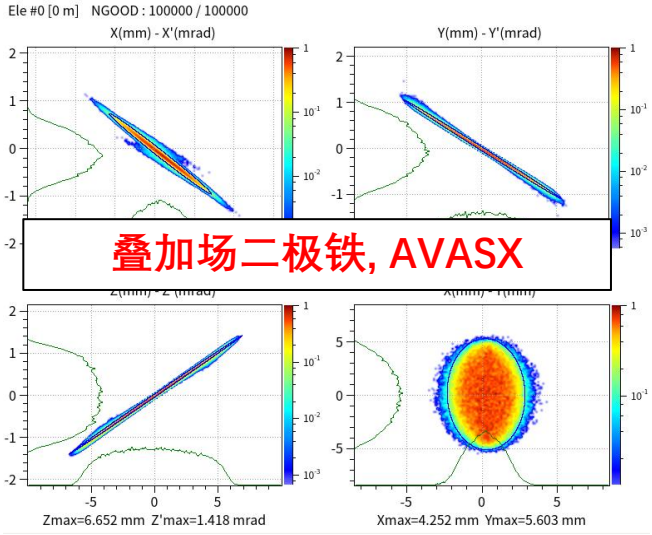
• 误差分析、叠加场元件测试

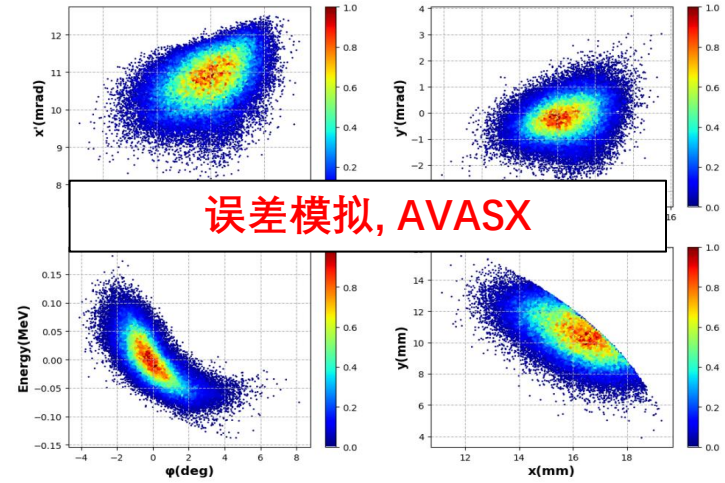












• 外场内存布局性能测试

- **kernel-1：外场以SOA（structure of arrays）布局存储**
- **kernel-2：外场以AOS（array of structures）布局存储**

表1. kernel-1与kernel-2在加载外场时，L1缓存、L2缓存、device内存的访存请求数以及内核执行时间

Employed kernel	Threads to L1 cache			L1 cache to L2 cache		L2 cache to device memory		Duration (us)
	Memory requests	Requested sectors	Hit rate (%)	Memory requests	Requested sectors	Requested sectors	Requested bytes	
kernel-1	2,812,500	6,008,627	80.95	300,060	1,144,828	1,125,396	36,012,672	101.25
kernel-2	1,656,250	4,501,828	74.77	290,704	1,135,932	1,125,176	36,005,632	86.72

表2. kernel-1与kernel-2在对CAFe仿真性能的影响

Employed kernel	Duration per step (us)	Computing performance (Giga particles per second)	Performance improvement (%)
kernel-1	523.720486	1.909415	-
kernel-2	516.982222	1.934302	1.30

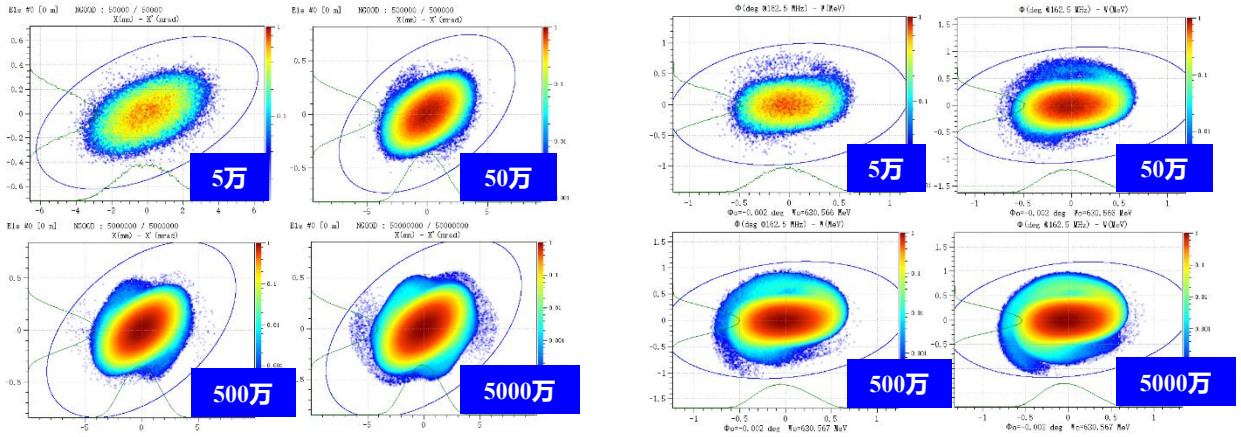
• 电荷沉积内核性能分析结果

Employed kernel	Runtime options	Threads to L1 cache			Excessive sector rate (%)	Compute throughput (%)	Memory throughput (GB/s)	Duration (us)	Performance improvement (%)
		Memory requests	Requested sectors	Hit rate (%)					
kernel-3	不排序、无优化	250,000	7,832,161	0%	9.90	6.92	51.96	463.01	-
kernel-4	不排序、仅聚合	250,016	7,814,976	0%	9.90	14.31	51.96	463.74	- 0.16
kernel-5	不排序、聚合+发散范围	251,904	7,822,528	0%	9.90	14.37	52.07	463.49	- 0.10
kernel-6	不排序、仅内存副本	250,000	7,832,737	0%	9.90	7.55	56.82	426.50	8.56
kernel-7	仅排序、无优化	250,000	7,869,016	0%	not provided	5.87	19.48	1230.00	- 62.36
kernel-8	排序、聚合	250,016	997,324	0%	0.01	29.95	167.78	143.55	222.54
kernel-9	排序、聚合+发散范围	251,904	1,004,876	0%	0.01	56.05	307.71	78.40	490.57
kernel-10	排序、聚合+内存副本	250,016	997,324	0%	0.01	45.79	250.32	96.67	378.96
kernel-11	排序、所有优化策略	251,904	1,004,876	0%	0.01	64.63	356.47	68.06	580.30

• GPU仿真程序与CPU仿真程序性能对比

测试环境和设备：

- CPU设备：双路28核Intel Gold 6330处理器，最多使用56核计算
- GPU设备：NVIDIA Tesla A100-PCIe 40GB
- PIC网格数：128 × 128 × 128， 256 × 256 × 256
- 测试束线：CAFe、CiADS



Code	Simulation durations (s)							
	1,000,000 particles		10,000,000 particles		100,000,000 particles (8 GPUs)		500,000,000 particles (8 GPUs)	
	CAFe	CiADS	CAFe	CiADS	CAFe	CiADS	CAFe	CiADS
AVASX	16.58	45.98	70.87	222.24	52.42	243.09	324.55	1499.73
AVAS	7832.72	25331.75	12359.02	39188.10	-			
Speedup	472.42	550.93	174.39	176.33	-	-	-	-

谢谢