SunwayLB: Enabling Extreme-Scale Lattice Boltzmann Method Based Computing Fluid Dynamics Simulations on Advanced Heterogeneous Supercomputers (*TPDS'23*)

Zhao Liu , *Associate Member, IEEE*, Xuesen Chu , Xiaojing Lv , Hongsong Meng , Hanyue Liu , Guanghui Zhu , Haohuan Fu , *Senior Member, IEEE*, and Guangwen Yang

OUTLINE

目 录

- 背 景
- SunwayLB优化方法
- 性能结果

背景

- 计算流体力学（Computing Fluid Dynamics, CFD）
  - 控制方程
  - CFD软件求解流程
- CFD 数值方法与模拟模型
  - 格子玻尔兹曼方法（LBM）
- 神威•太湖之光超级计算机（Sunway TaihuLight）
  - SW26010
  - SW26010-Pro

背景

控制方程

- 计算流体力学是通过数值模拟的方式对控制方程求解，一般是纳维-斯托克斯（Navier-Stokes）方程，预测流体流动时的状态和变化
- 给出积分形式的N-S方程的动量守恒：

$$\frac{\partial}{\partial t}\int_V Q dV + \oint_{\partial V}\left(F_c - F_v\right)dS = 0.$$

- 其中源项Q，对流通量F_c，粘性通量F_v

$$Q = \begin{bmatrix} \rho \\ \rho u \\ \rho v \\ \rho w \\ \rho e \end{bmatrix}.$$

$$F_c = \begin{bmatrix} \rho U \\ \rho u U + n_x p \\ \rho v U + n_y p \\ \rho w U + n_z p \\ \rho U\left(e + \frac{u^2+v^2+w^2}{2} + \frac{p}{\rho}\right) \end{bmatrix}.$$

$$F_v = \begin{bmatrix} 0 \\ n_x\tau_{xx} + n_y\tau_{yx} + n_z\tau_{zx} \\ n_x\tau_{yx} + n_y\tau_{yy} + n_z\tau_{yz} \\ n_x\tau_{zx} + n_y\tau_{zy} + n_z\tau_{zz} \\ n_x\theta_x + n_y\theta_y + n_z\theta_z \end{bmatrix}.$$

背景

计算流体力学软件的工作流程

- **网格前处理**
  - 网格格式转化、分区、重构等
- **求解控制方程**
  - 空间离散模型(FVM, LBM)
  - 数值计算方法(DNS, LES, RANS)
- **后处理**
  - 可视化处理和数据分析

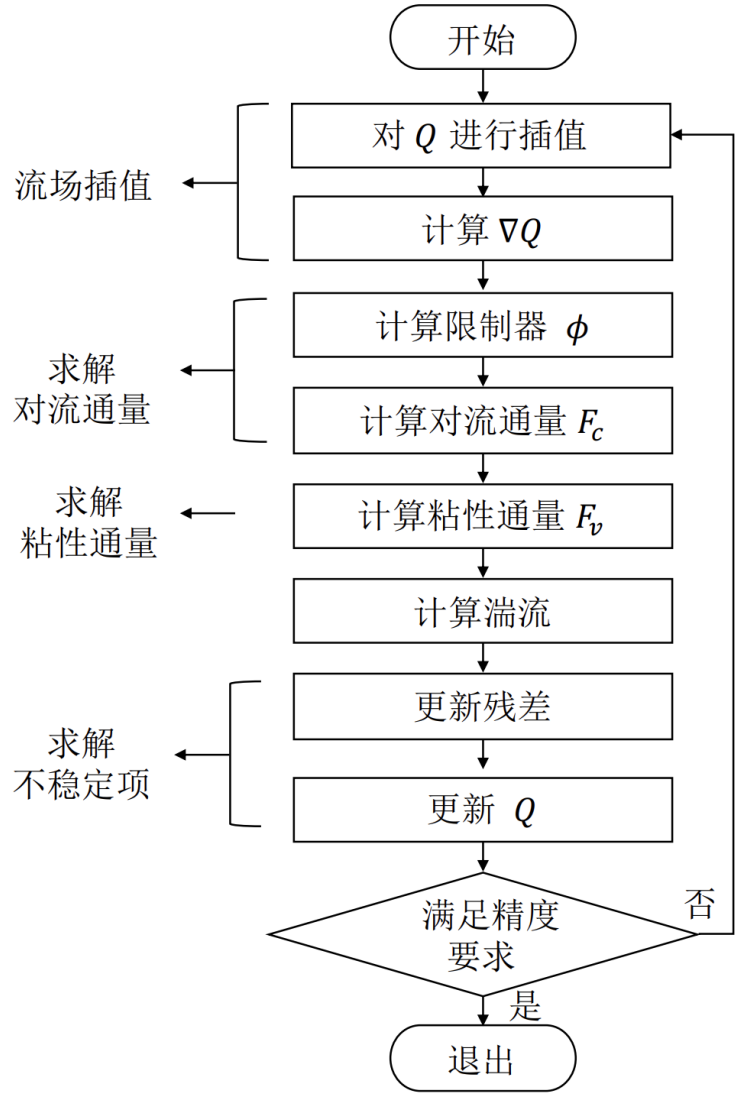


图 2-4　可压缩 Navier-Stokes 方程的 FVM 算法流程图

CFD 数值方法与模拟模型

模拟方法

- DNS（Direct Numerical Simulation，直接数值模拟）——精度高，速度慢
- LES（Large Eddy Simulation，大涡模拟）——介于DNS与RANS之间
- RANS（Reynolds-Averaged Navier-Stokes，雷诺平均NS方程）精度较低,速度块

空间离散方法

- FVM（Finite Volume Method，有限体积法）——体积单元
- FDM（Finite Difference Method，有限差分法）——网格点
- FEM（Finite Element Method，有限元法）——有限元元素
- LBM（Lattice Boltzmann Method，格子-玻尔兹曼方法）——微观粒子

LBM—格子-玻尔兹曼方法概述

- 格子-玻尔兹曼方法基于玻尔兹曼方程，通过粒子的碰撞*collide*与传播*stream*来描述粒子的演化过程；流体被分割为一个个规则的格子*lattice*，流体的运动被视为格子点上的流体粒子群的运动，以速度分布函数 f 为基本变量，方程可写为：

$$\frac{\partial f}{\partial t} + \boldsymbol{u} \cdot \nabla f = \Omega$$

离散后的玻尔兹曼方程为：

$$f_i(\boldsymbol{x} + \boldsymbol{\xi}\delta t, t + \delta t) - f_i(\boldsymbol{x}, t) = \Omega_i$$

格子-玻尔兹曼方法 碰撞模型

LBGK (Lattice Bhatnagar-Gross-Krook)

- 用于简化碰撞项 Ω，在晶格上模拟离散的速度分布函数，通过碰撞算子模拟流体粒子的碰撞过程

速度分布函数在一个时间步长中发生了变化 Ω

$$f_i(\boldsymbol{x}+\boldsymbol{\xi}\delta t, t+\delta t) - f_i(\boldsymbol{x}, t) = \Omega_i$$

碰撞模型：碰撞导致粒子的速度分布函数与平衡速度分布函数产生了差，这个差的变化速度由τ决定

$$f_i\left(\vec{x}+\vec{c}_i\Delta t, t+\Delta t\right) - f_i\left(\vec{x}, t\right) = -\frac{1}{\tau}\left[f_i\left(\vec{x}, t\right) - f_i^{(eq)}\left(\vec{x}, t\right)\right]$$

CFD 数值方法与模拟模型

离散速度模型 DnQm

- 常用的有D2Q9
- 三维D3Q15, D3Q19, D3Q27
- 维度D与速度方向数Q

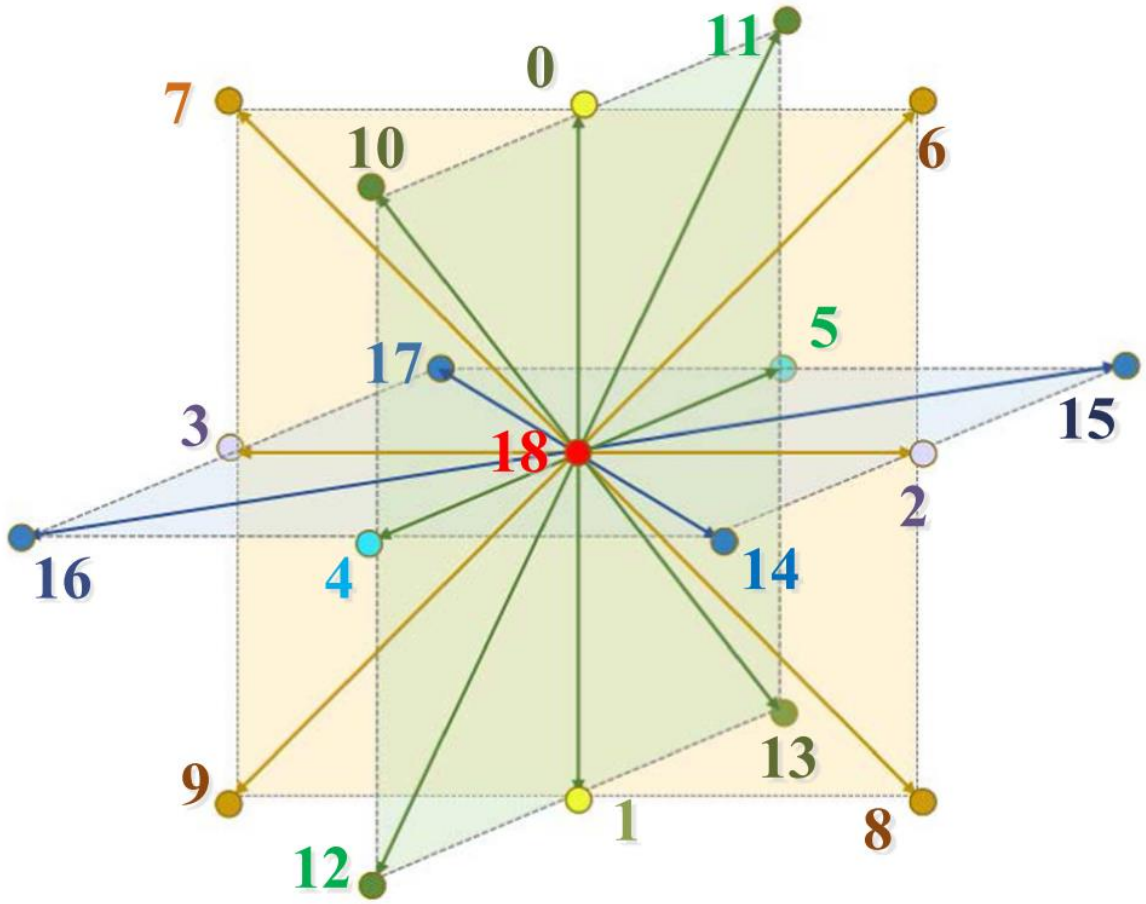


Fig. 3. D3Q19 discretization scheme.

CFD 数值方法与模拟模型

LBM迭代计算流程

- **迭代过程：**按照时间步长进行迭代，每个时间步长都包括**传播**和**碰撞**两个步骤
- **传播操作：**在传播操作中，遍历域中的每个晶格单元，从其邻近的晶格单元接收数据，以更新当前晶格单元的分布函数。传播操作将邻近单元的数据传播到当前单元，以准备进行下一步的碰撞操作
- **碰撞操作：**在碰撞操作中，根据 LBGK 模型，计算每个晶格单元的分布函数的新值。这个过程考虑了流体粒子之间的碰撞影响，根据模型的碰撞算子来更新分布函数
- **迭代：**传播和碰撞操作交替进行，直到达到设定的迭代次数或满足停止条件为止。这样，程序可以模拟流体流动的演化过程，逐步更新分布函数，从而得到流场的解

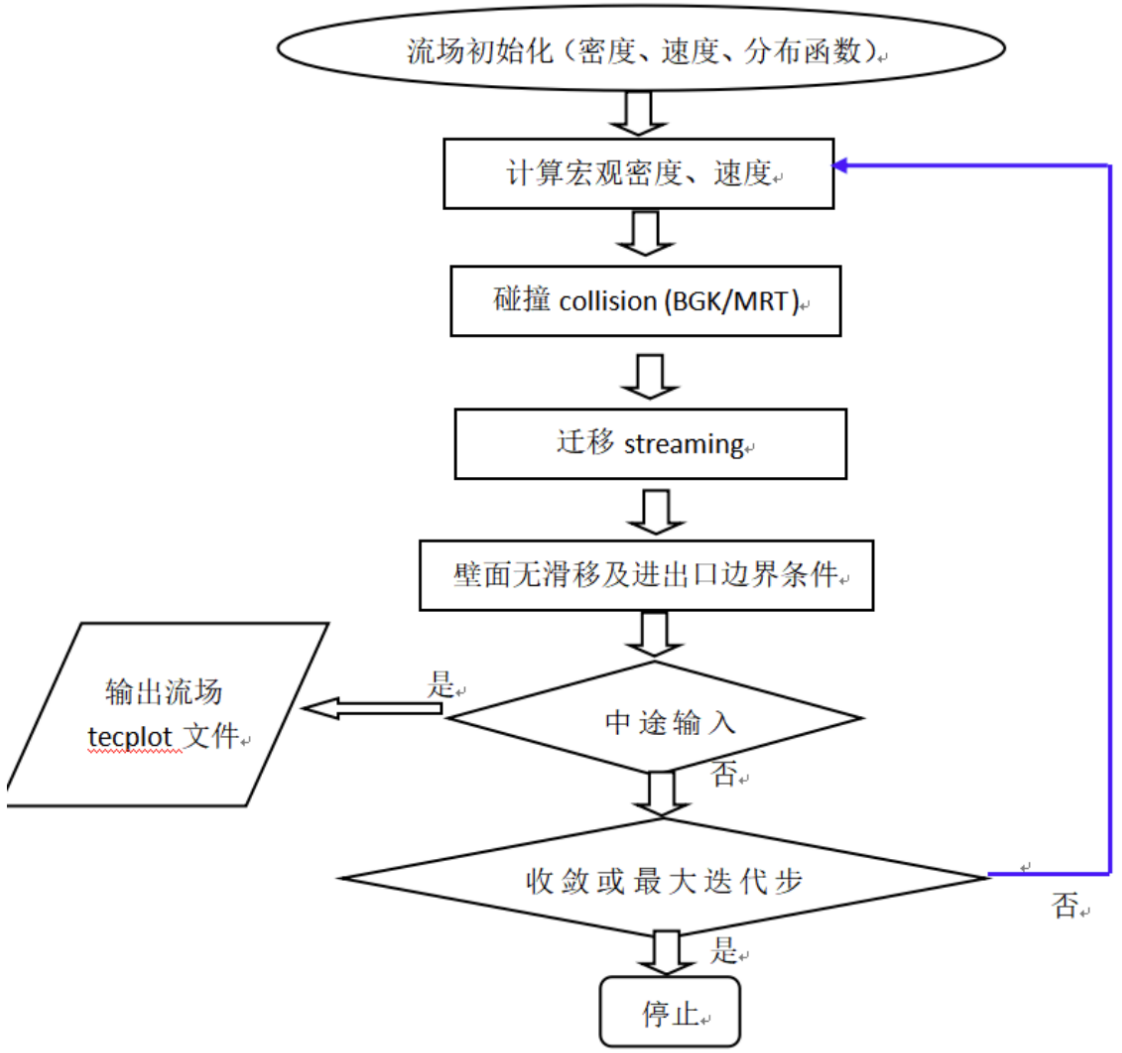


图 10-2 LBM模块架构

Sunway

Sunway一代和二代的硬件

- 40960个SW26010
  - 260-core, 125PFlops峰值性能，93PFlops持续LINPACK性能
- 107520个SW26010-Pro
  - 390-core，93PFlops峰值性能，性能超过1.5EFlops，

1个Core-Groups(CG)

- 1个管理处理元(Management processing element, MPE)
- 1个智能内存处理元(Intelligent memory processing element, IMPE)
- 64个计算处理元(Computing processing elements, CPEs)

Sunway 处理器

SW26010

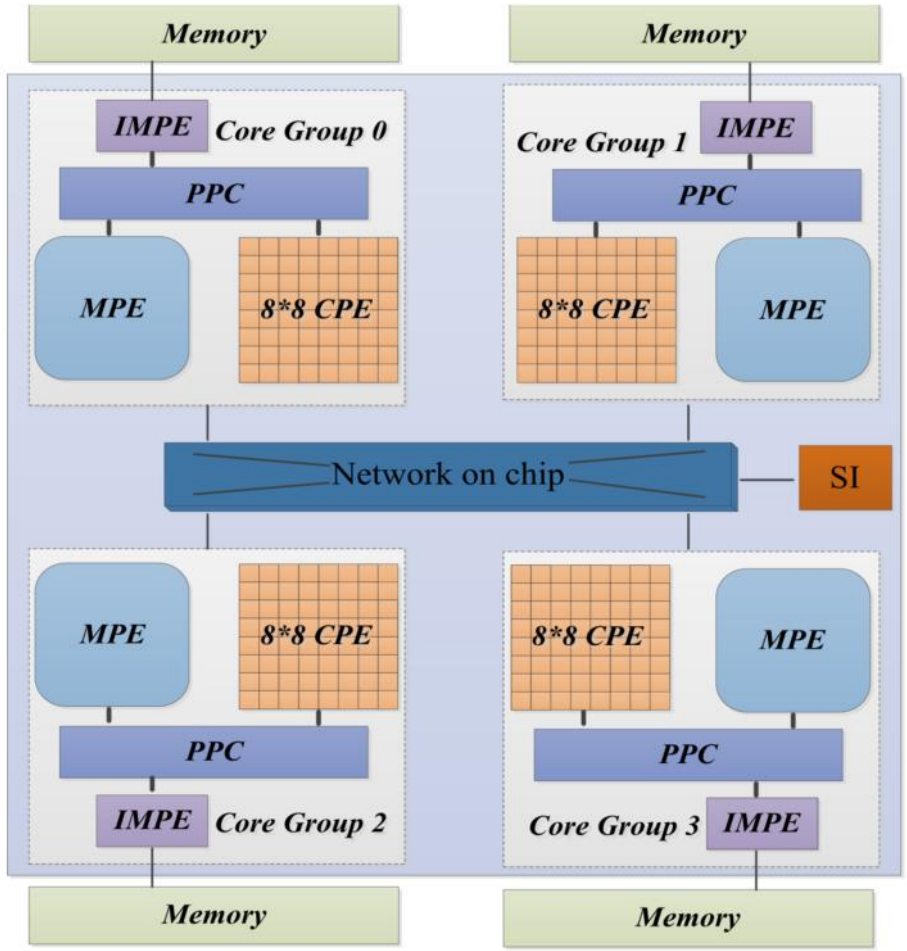
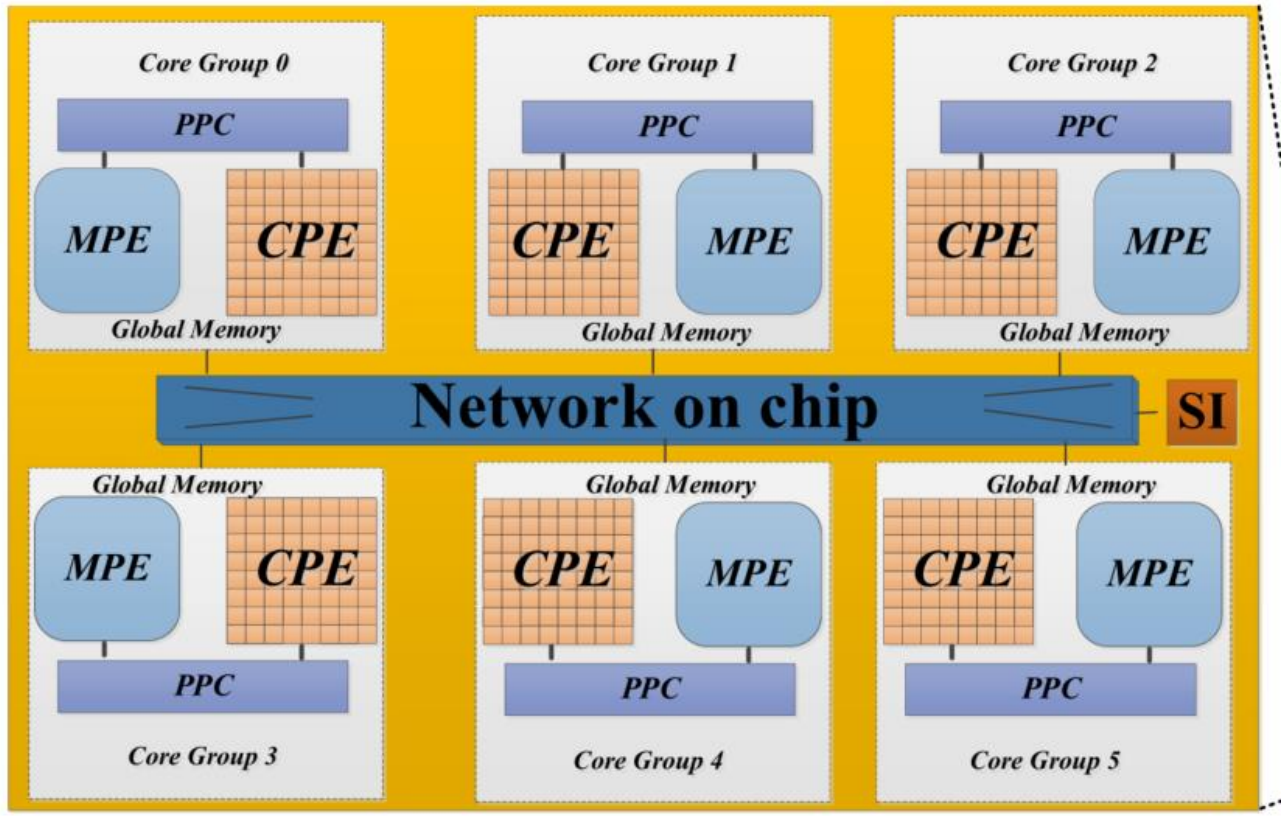


Fig. 1. Architecture of SW26010.

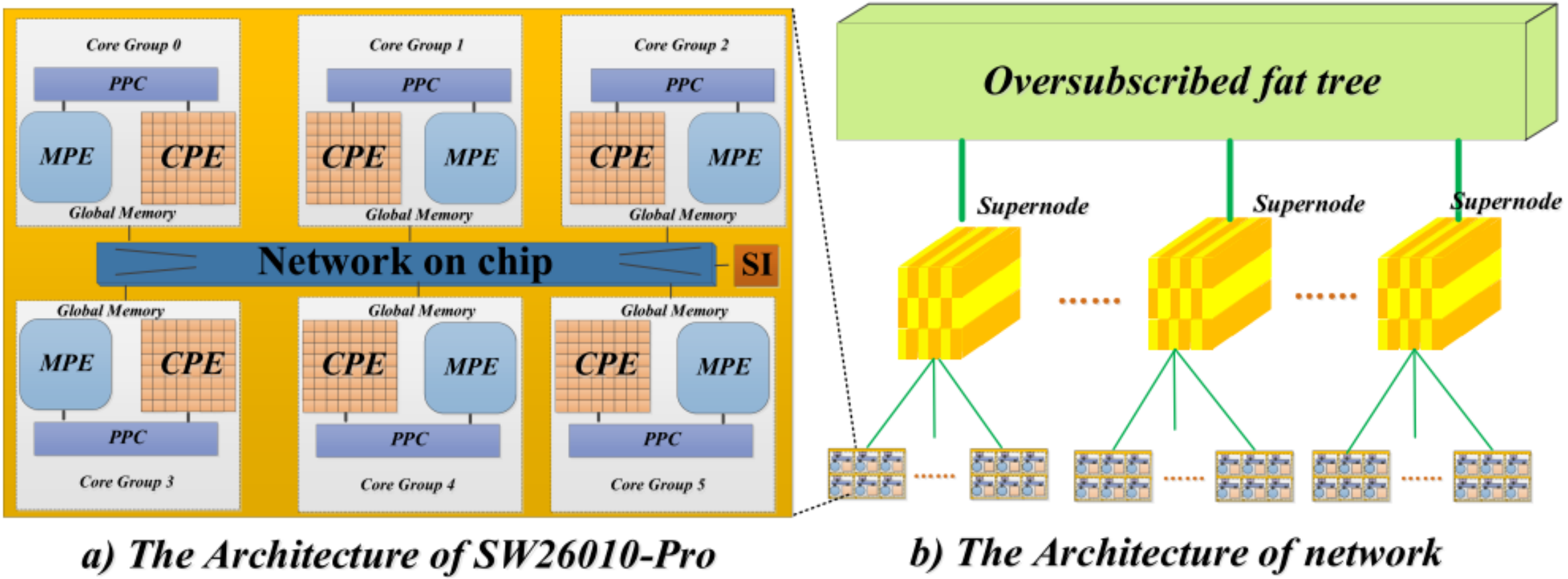
SW2610-Pro



a) The Architecture of SW26010-Pro

Sunway 处理器

SW2610-Pro



a) The Architecture of SW26010-Pro

b) The Architecture of network

SunwayLB

- Motivation
- 论文主要工作
  - 多级并行化策略
  - New Sunway优化
  - SunwayLB-GPU
- 实验结果
  - SunwayLB 弱扩展和强扩展测试
  - 工程实验

SunwayLB

Motivation

- 基于FVM-DNS/LES的湍流任务计算耗时很长；
- LBM的并行计算能力优于FVM；
- LBM 可以在具有专用加速器的异构超级计算系统上并行化；
- 利用先进的异构架构（如Sunway超算）构建一个基于LBM的大规模工业级CFD应用 SunwayLB，从软硬件层面提出优化策略，减少大规模模拟中的计算成本开销

SunwayLB

论文的主要工作

- 开发了综合软件框架 SunwayLB，提供完整的大规模仿真解决方案；
- 在Sunway上实现多级并行优化、内核策略、手动循环展开以及指令重排序技术以提高处理器众核的计算潜能；
- 新代Sunway硬件，新的通信方式，带来内存带宽利用率的提升；
- 将优化技术移植到GPU集群上进行性能评估

SunwayLB 总体框架

- 预处理模块
- 求解器
  - 基于D3Q19
  - 多级并行化方案
  - 动态边界数据交换
- 后处理模块
  - 数据格式和可视化工具
- I/O层支持

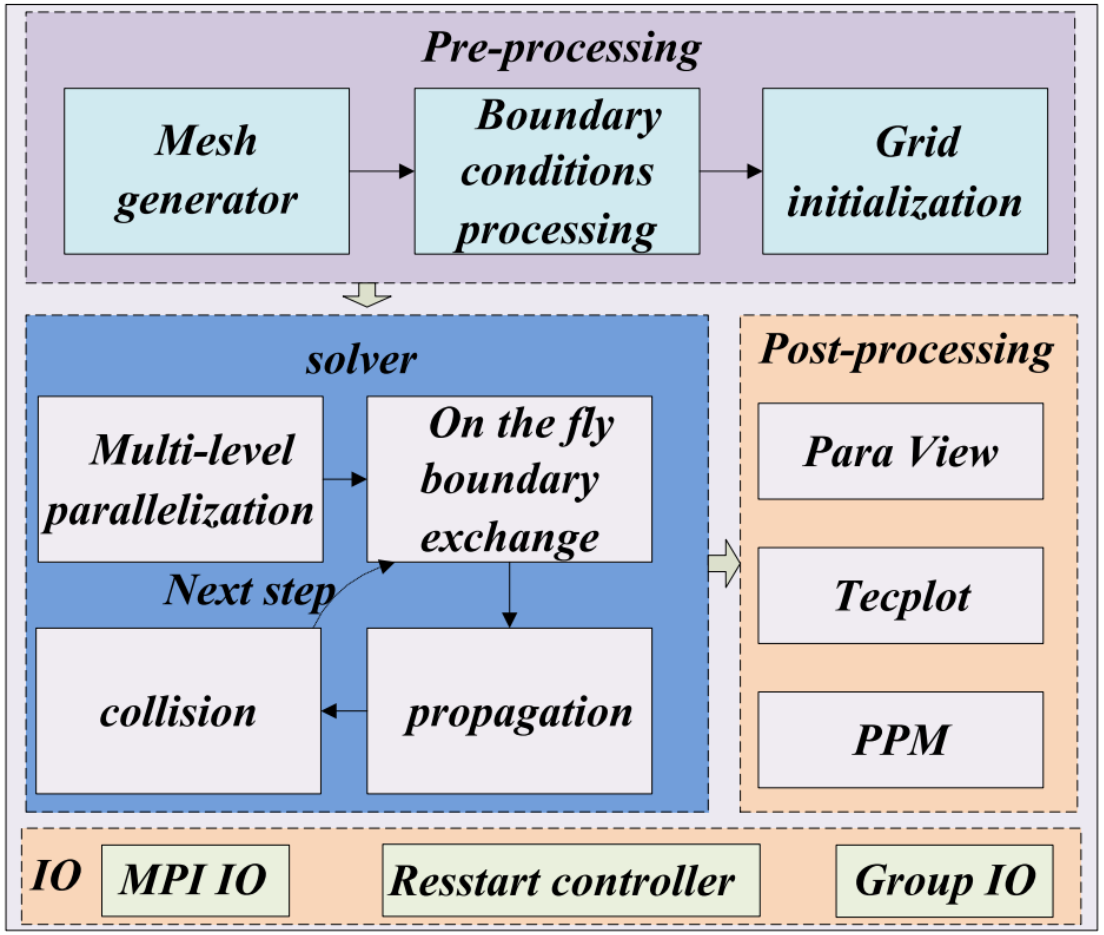


Fig. 4. Holistic framework of SunwayLB.

SunwayLB——多级并行化方法 I

多级并行化方法——数据结构

- D3Q19模型，更新一个单元时，将需要来自其邻居的19个粒子群，如果我们采用结构数组AoS，这些粒子群在内存中不连续，导致大量的随机内存访问和频繁的DMA启动，DMA开销严重影响性能
- 所以采用数组结构(SoA)模型将粒子群数据连续存储在内存中，提高性能

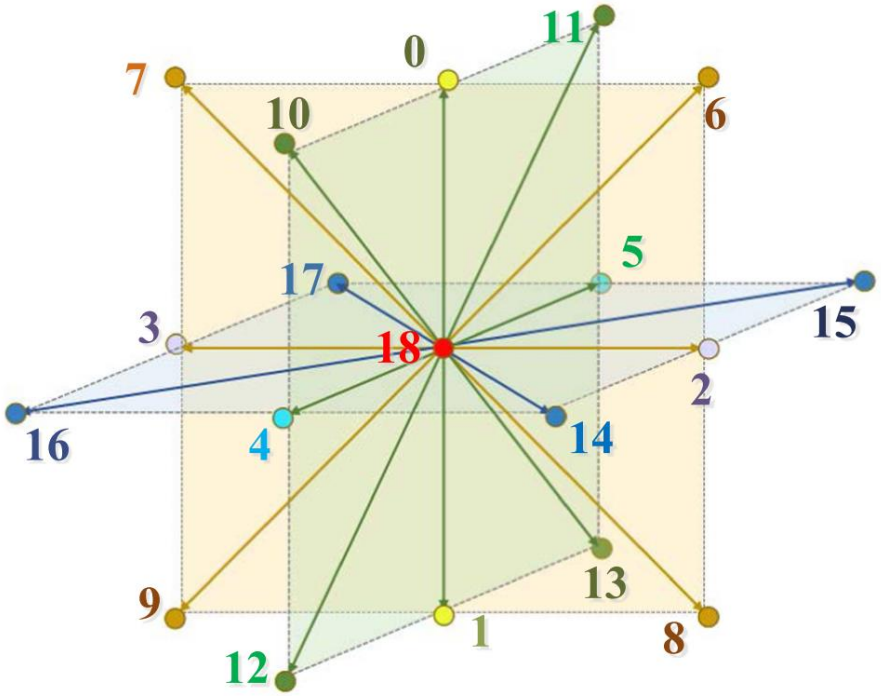


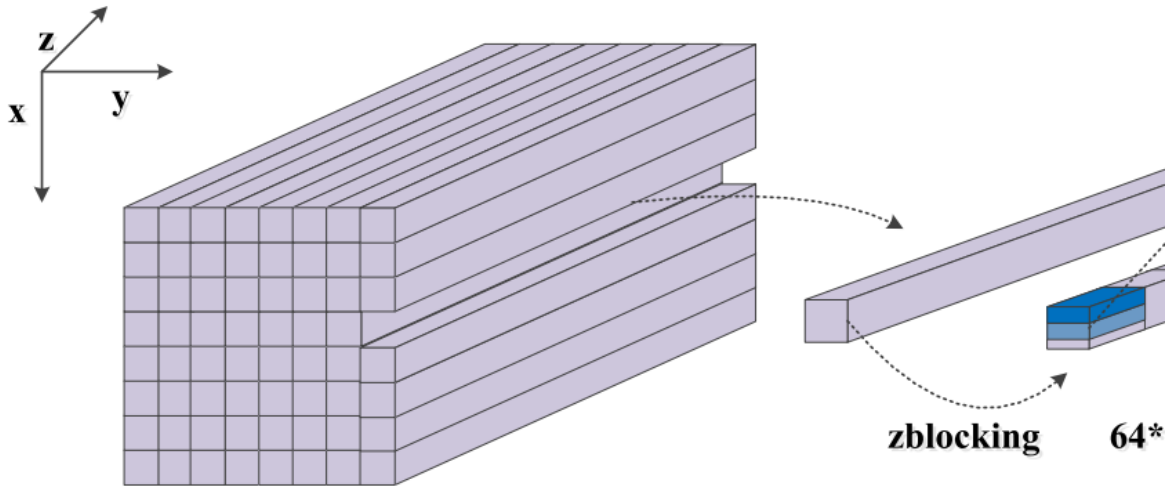
Fig. 3. D3Q19 discretization scheme.

SunwayLB——多级并行化方法 I

1) *Domain Decomposition at the MPI Level and On-the-Fly Halo Exchange Scheme*

➢ *MPI Domain Decomposition*

- x-y方向的2D域分解
- z轴保留完整的子域
- 1个MPI进程最多需要和8个方向的邻居通信

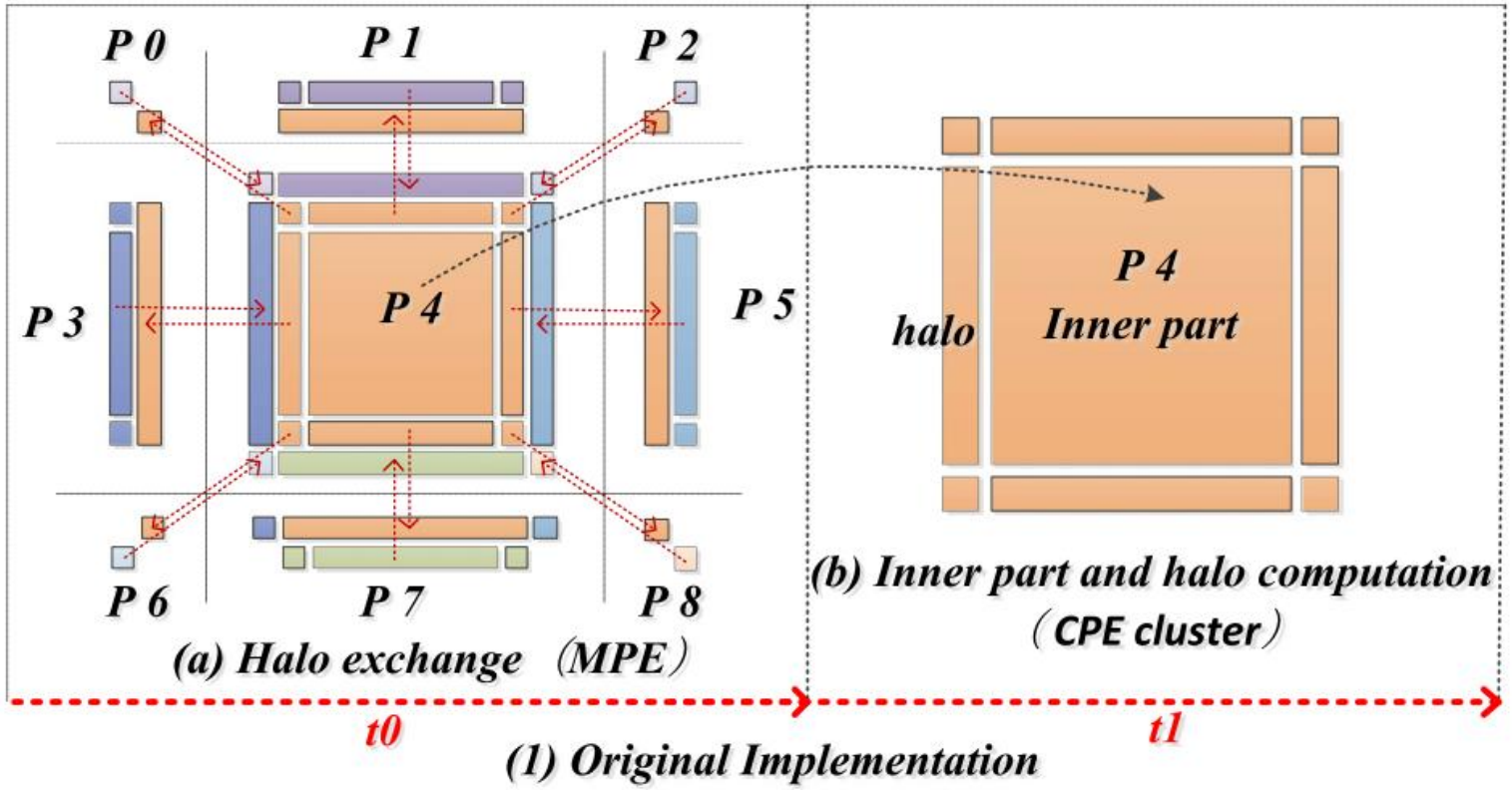


(1) Domain decomposition at the MPI level

SunwayLB——多级并行化方法 I

1) *Domain Decomposition at the MPI Level and On-the-Fly Halo Exchange Scheme*

➢ ***On-the-Fly Halo Exchange Scheme***

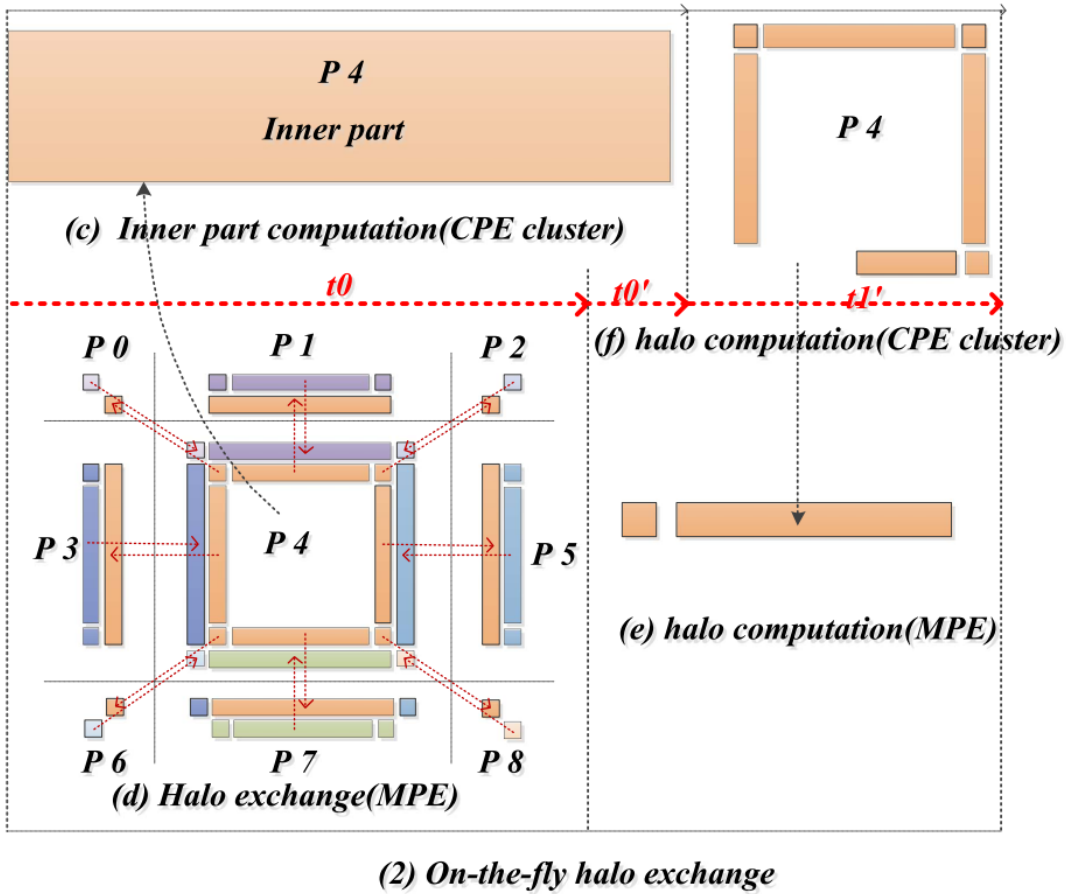


SunwayLB——多级并行化方法 I

1) *Domain Decomposition at the MPI Level and On-the-Fly Halo Exchange Scheme*

➢ *On-the-Fly Halo Exchange Scheme*

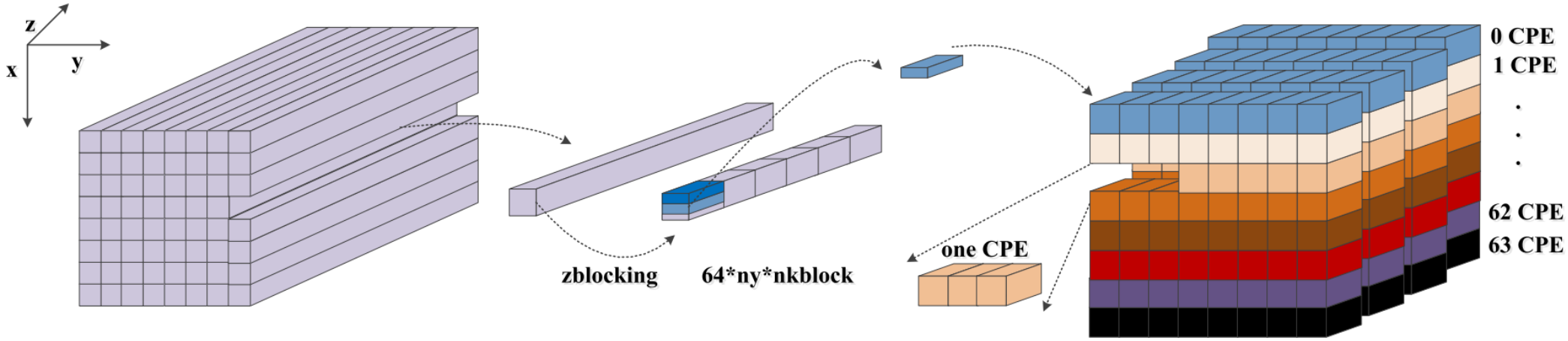
- **MPE和CPEs并行**
- **MPE只负责halo边界数据交换和小部分计算任务**
- ***CPEs* 复制执行内部计算任务**



SunwayLB——多级并行化方法 II

2)Data Blocking and Sharing at the Level of CPEs

- *CPE* 三级内存：REG-LDM-MEM（寄存器-本地数据存储-主内存）
- 由于所以数据需要DMA复制到每一个CPE的64KB大小的LDM中，64个CPE的LDM总大小为4MB，远小于一个MPI进程的数据大小，需要针对CPE集群定制数据阻塞以最大化DMA带宽并减少调用DMA接口的次数，即将尽可能多的连续块大小复制到LDM上

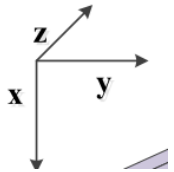


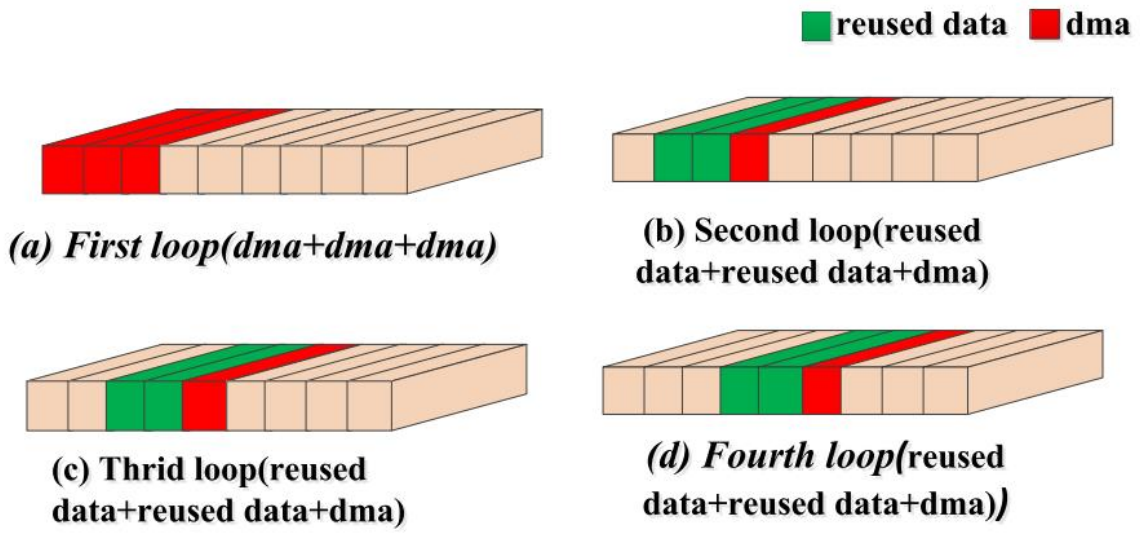
(1) Domain decomposition at the MPI level　　(2) Blocking data to core group and CPE

SunwayLB——多级并行化方法 II

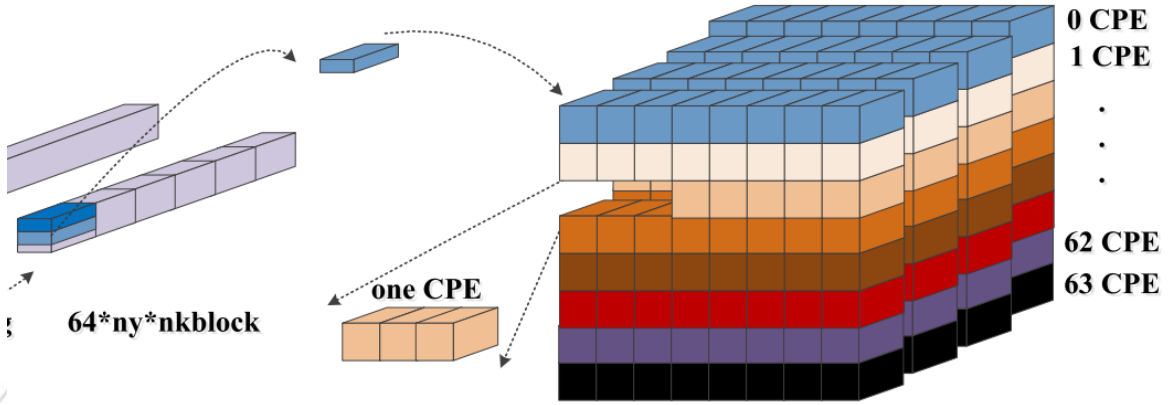
2)Data Blocking and Sharing at the Level of CPEs

- 策略： CPE内部沿 *x* 轴数据循环复用；沿y轴相邻CPE之间共享数据
  - *每个CPE簇在 x-y-z 轴上数据分块为64*3*70，每个CPE在 z-x 平面上有3*70个块*
  - 在x轴上，每个CPE线程需要相邻的数据来执行本地传播，则相邻数据可以重复使用
  - 在y轴上，相邻CPE之间也有共享数据的机会，采用寄存器通信机制完成CPE之间的共享数据





(3) Data reuse inside one CPE



(2) Blocking data to core group and CPE

SunwayLB——多级并行化方法 III

3) *Fusing Kernels With Different Performance Constraints*

- 碰撞 *Propagation and* 传播*Collision steps* 是LBM代码中耗时最长的 Kernels
- 在每个时间步长需要从相邻单元加载并存储19个粒子群进行传播（D3Q19）
- 困难：更新一次流体单元需要从主存中获取380字节的数据，在最大测试中每个CG包含3500万个单元，LDM-MAM传输数据高达12GB，传播过程不涉及浮点计算
- 传播步骤结束后执行碰撞Kernels，进行大部分浮点数学计算，但每个单元仅用自身相关的密度数据，即每个格子的计算是相互独立的，这使得碰撞Kernels可以完全并行化

SunwayLB——多级并行化方法 III

3)*Fusing Kernels With Different Performance Constraints*

- 利用IMPE异步内存访问能力和CPE低延迟DMA操作实现内核映射到核心组（Core group）的不同处理单元上同时执行Kernel
- 并利用A-B model内存布局，能够将传播和碰撞融合到同一个循环中进行，并将耗时的操作映射到不同的处理单元中，实现计算-DMA重叠
- 在一个时间步长中能够减少4次DMA，提高约30%性能

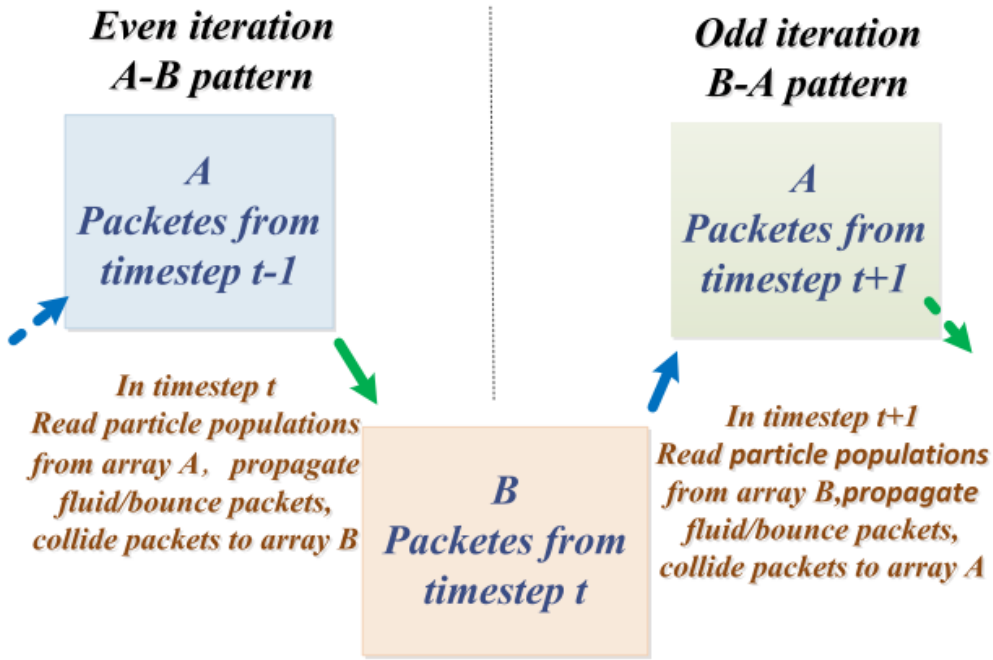
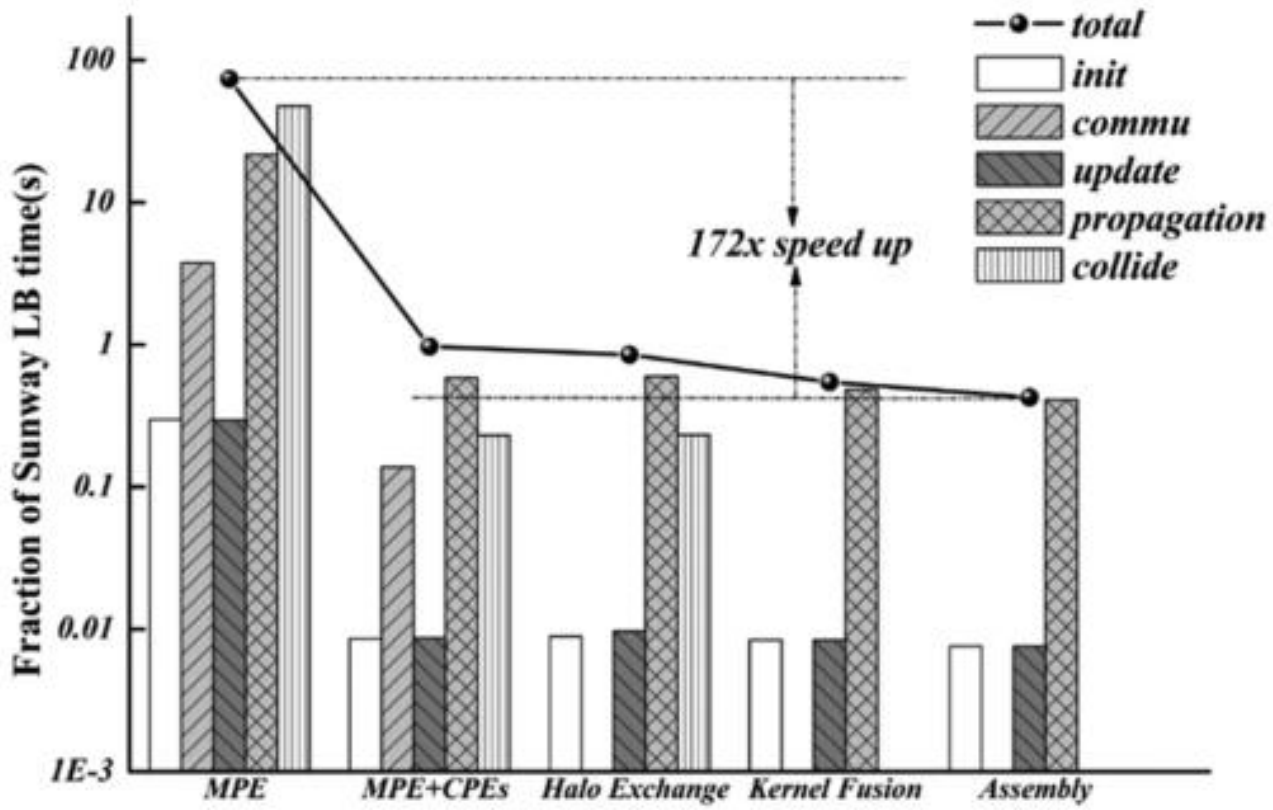


Fig. 7. Execution patter of A-B memory layout.

SunwayLB——多级并行化方法 IV

4) Assembly Code Level Optimizations

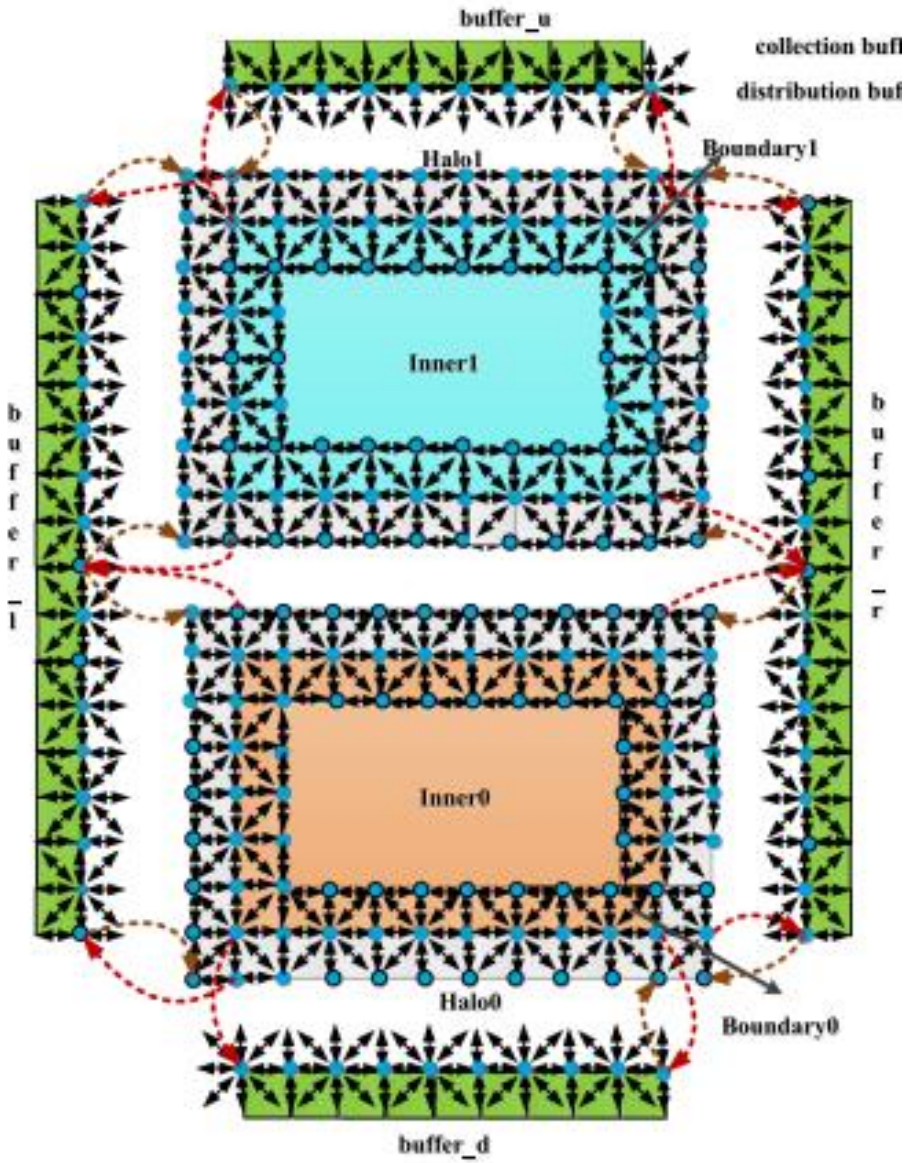
- 手动循环展开和指令调度技术重写内核
- 高效利用CPE管道和256bit向量化指令
- 4种多级并行化方法总共带来172x性能提升!



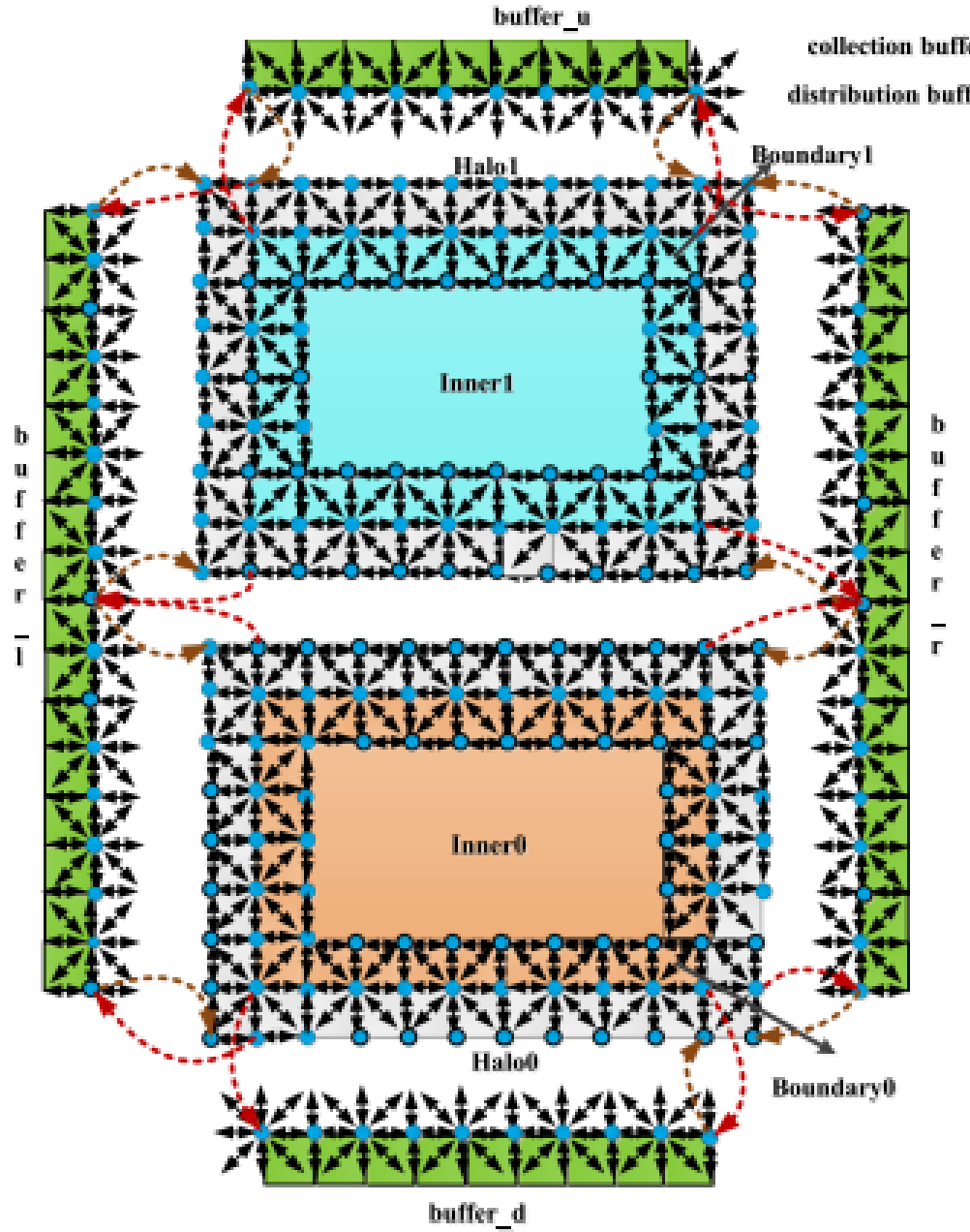
SunwayLB

新一代Sunway的并行和优化策略

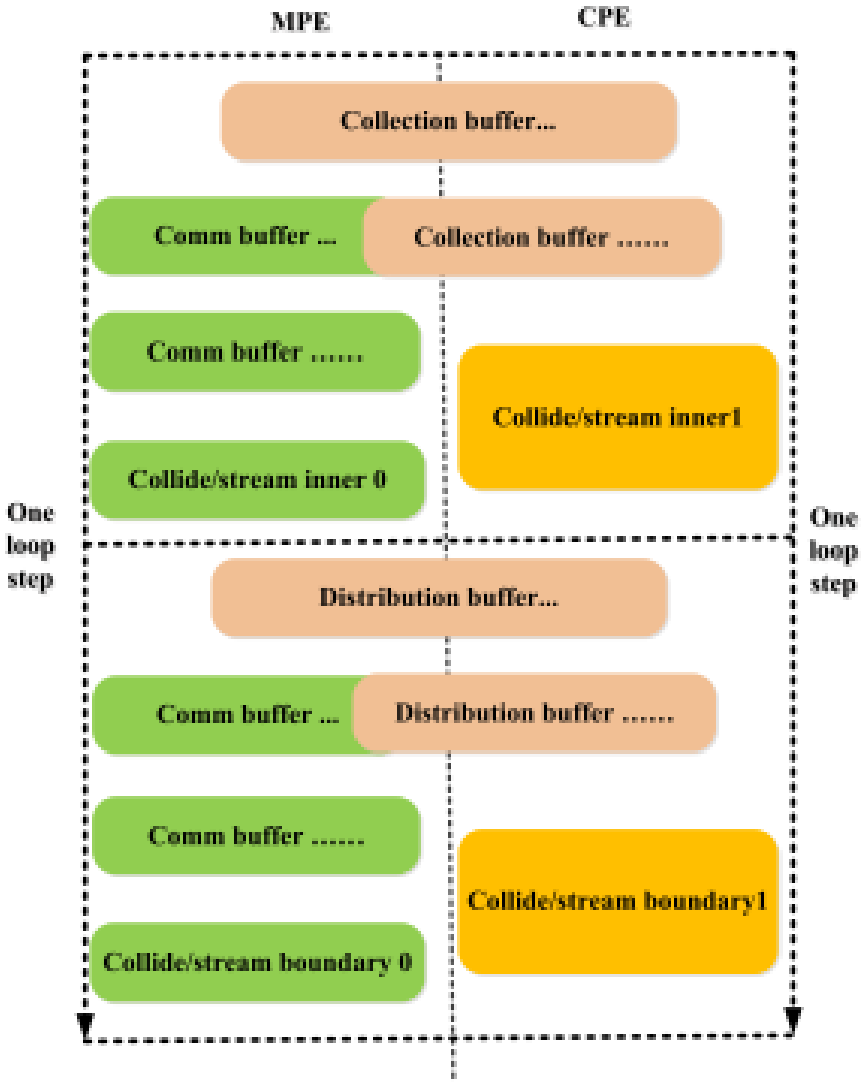
- SW26010-Pro的内存容量和LDM容量显著提高
- 新的通信方案：Halo scheme
- 由于LBM的粒子性质，分配给每个CG的子域需要与相邻子域交换数据
- 需要存储和更新相邻子域的数据，在子域周围放置了一层halo cells
  - Halo cell 分为内部cells和边界cells
  - 内部cells更新子域内的cells
  - 边界cells位于子域的最外层需要halo cells的数据进行传播和碰撞操作



- 新一
  - SW
  - 新


buffer_u
collection buffer
distribution buffer
Halo1
Boundary1
Inner1
buffer_l
buffer_r
Inner0
Halo0
Boundary0
buffer_d

(1)


MPE
CPE
Collection buffer...
Comm buffer ...
Collection buffer .....
Comm buffer .....
Collide/stream inner1
Collide/stream inner 0
One loop step
Distribution buffer...
Comm buffer ...
Distribution buffer .....
Comm buffer .....
Collide/stream boundary1
Collide/stream boundary 0
One loop step

(2)

New SunwayLB

新一代Sunway的并行和优化策略——通信部分

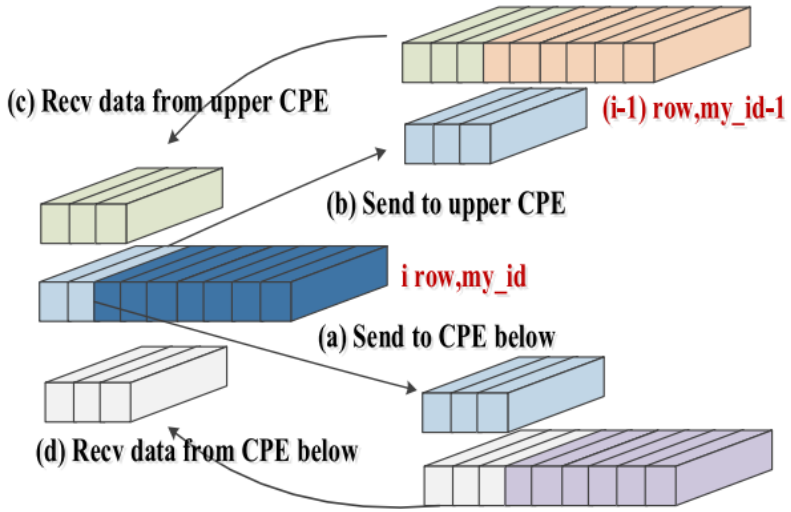
- 一个MPI进程需要与最多8个相邻节点建立通信
- 方法1：通过MPE启动数据交换操作，更新完成后将数据卸载到CPE执行传播和碰撞操作，但这种方法没有发挥异构架构的并行性
- 方法2：协同MPE与CPE
  - 平衡MPE与CPE的工作量
  - 采用非阻塞通信技术，MPE只负责halo cell的交换和小部分计算任务的通信
  - 同时CPE集群将数据DMA到LDM上，本地数据并行协助计算

New SunwayLB

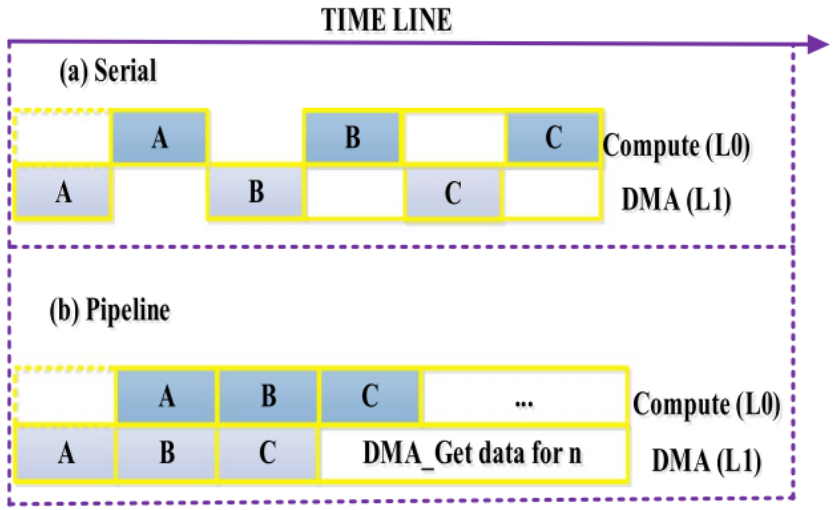
新一代Sunway的并行和优化策略——通信部分

方法2：协同MPE与CPE，将通信开销与内部单元计算重叠

- 平衡MPE与CPE的工作量
- 采用非阻塞通信技术，MPE只负责halo cell的交换和小部分计算任务的通信
- 同时CPE集群将数据DMA到LDM上，本地数据并行协助计算



(1)

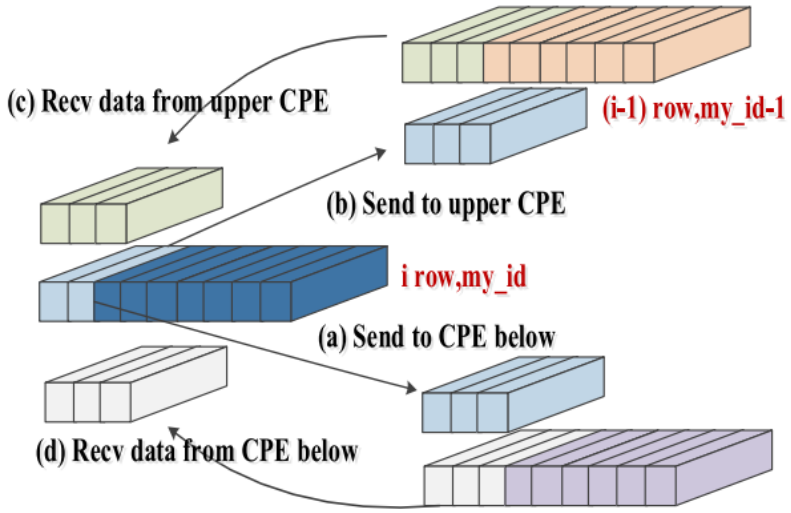


(2)

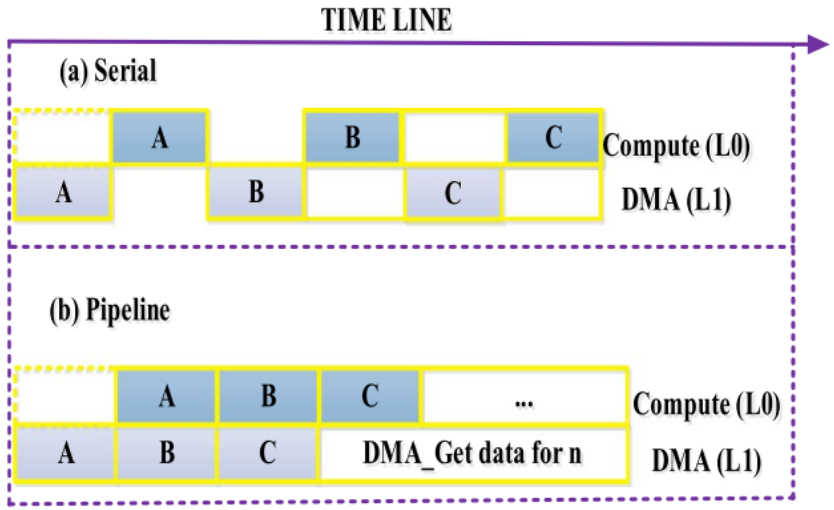
New SunwayLB

Data Sharing and Pipeline Optimization

- SW2610-Pro的CPE集群集成了网状网络和RMA的低延迟通信机制
- CPE网状网络强大带宽，利用RMA进行相邻CPE之间的边界数据共享
- 并且利用CPE的双管道执行异步DMA



(1)



(2)

SunwayLB—GPU集群优化技术

SunwayLB移植到GPU集群上测试

- GPU集群平台
  - 2个Intel Xeon 6248R CPU 和8个NVIDIA GEFORCE RTX 3090 GPU
- CUDA API重构SunwayLB中耗时代码函数
- 利用固定内存cudaMallocHost() 避免复制数据额外步骤和开销，提高带宽利用率

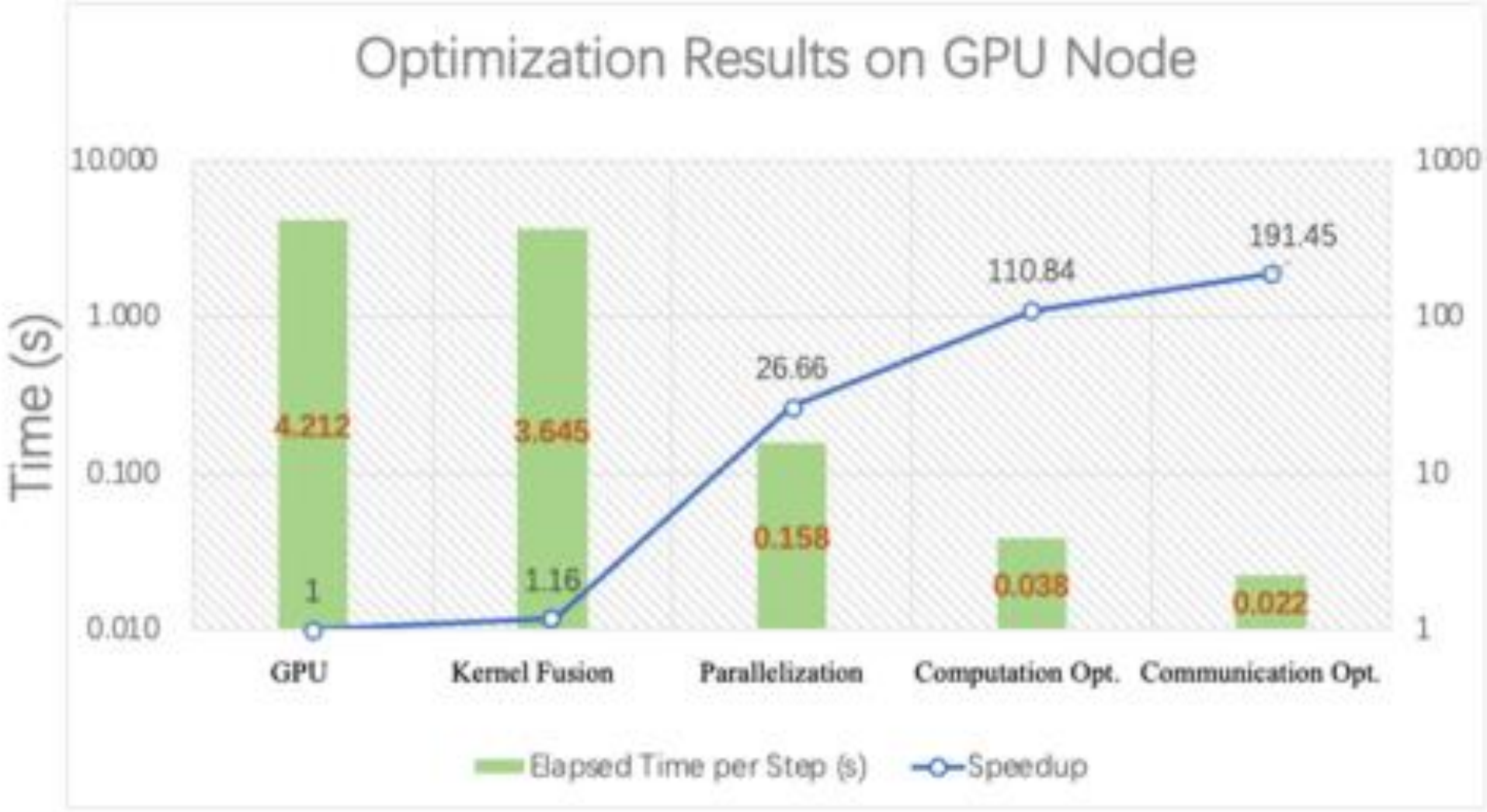
- GPU版本比CPU版本提供了约200倍的加速（1CPU+1GPU对比1CPU）

SunwayLB—GPU集群优化技术

其他优化技术

- 数据分块和Q维度循环移至到Kernel上
- 开销步预计算减少冗余计算步骤
- 用NCCL函数替换MPI函数进行节点内数据通信

SunwayLB—GPU集群优化技术



SunwayLB

性能结果

- LBM软件性能衡量方法：GLUPS, MLUPS
  - *P*单位是LUPS，M是晶格数，t是单个时间步长

$$P = M/t_s$$

- 最大模拟实验
  - 网格尺寸40000*40000*3500，5.6万亿个晶格
  - 模拟速度11245 GLUPS和4.7 Pflops
  - 模拟的规模大小是此前进行的最大规模DNS的2倍

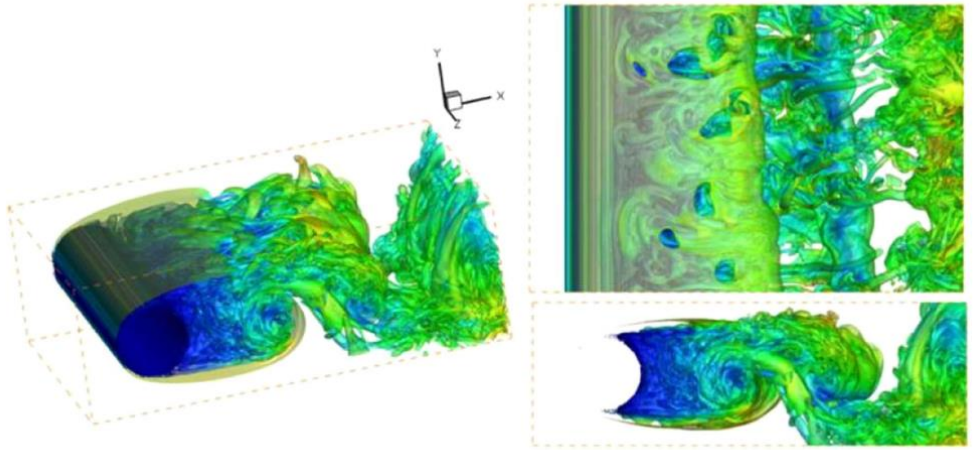
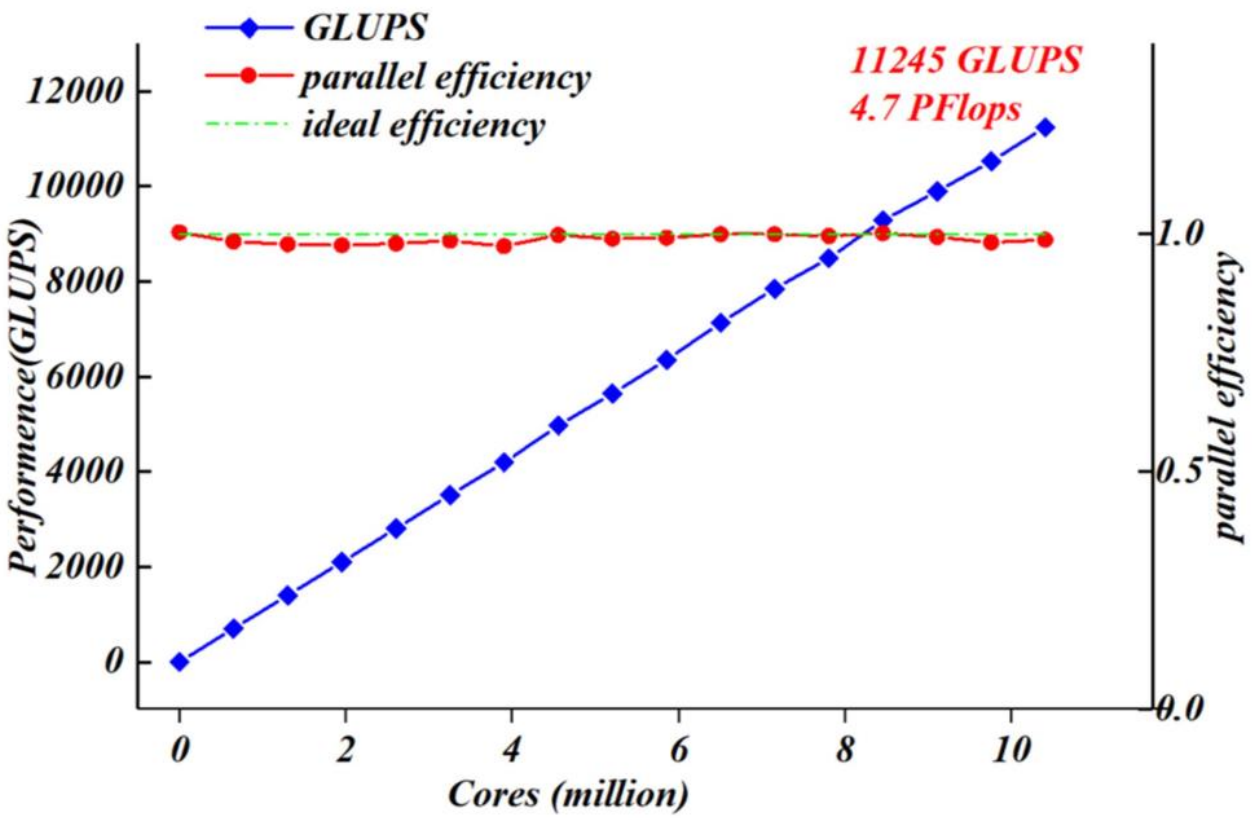


Fig. 12. Direct numerical simulation of instantaneous isosurface of Q-Criterion for flow past circular cylinder scenario at Re = 3900r following computation to realize pipelining.

SunwayLB

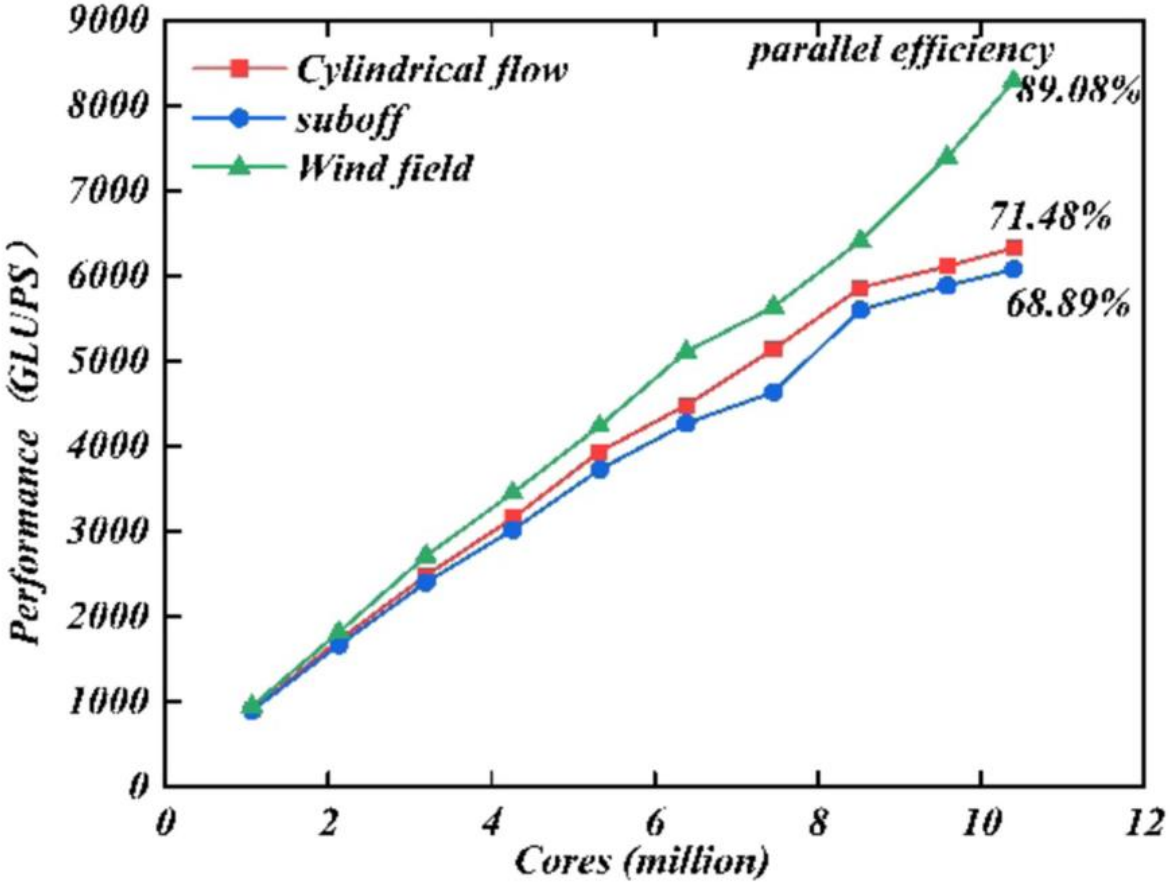
弱扩展性性能和并行效率

- 达到理论最大性能的77%



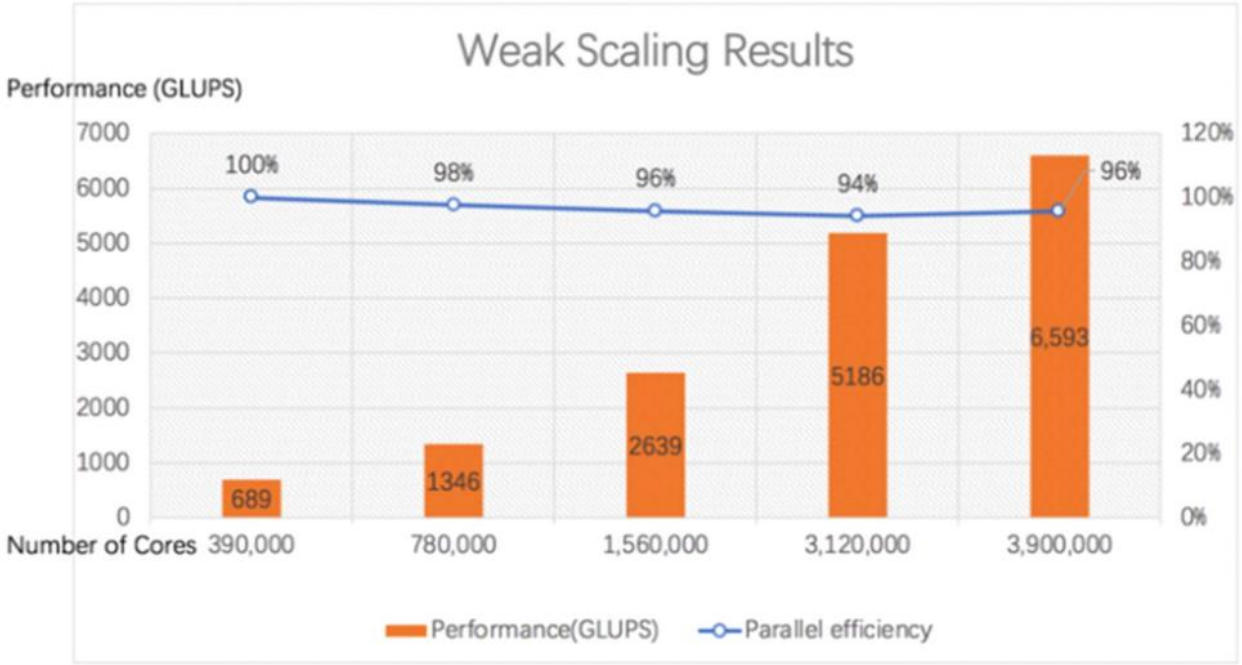
三种强扩展性模拟

- 104e5个核心时实现71.48%并行效率



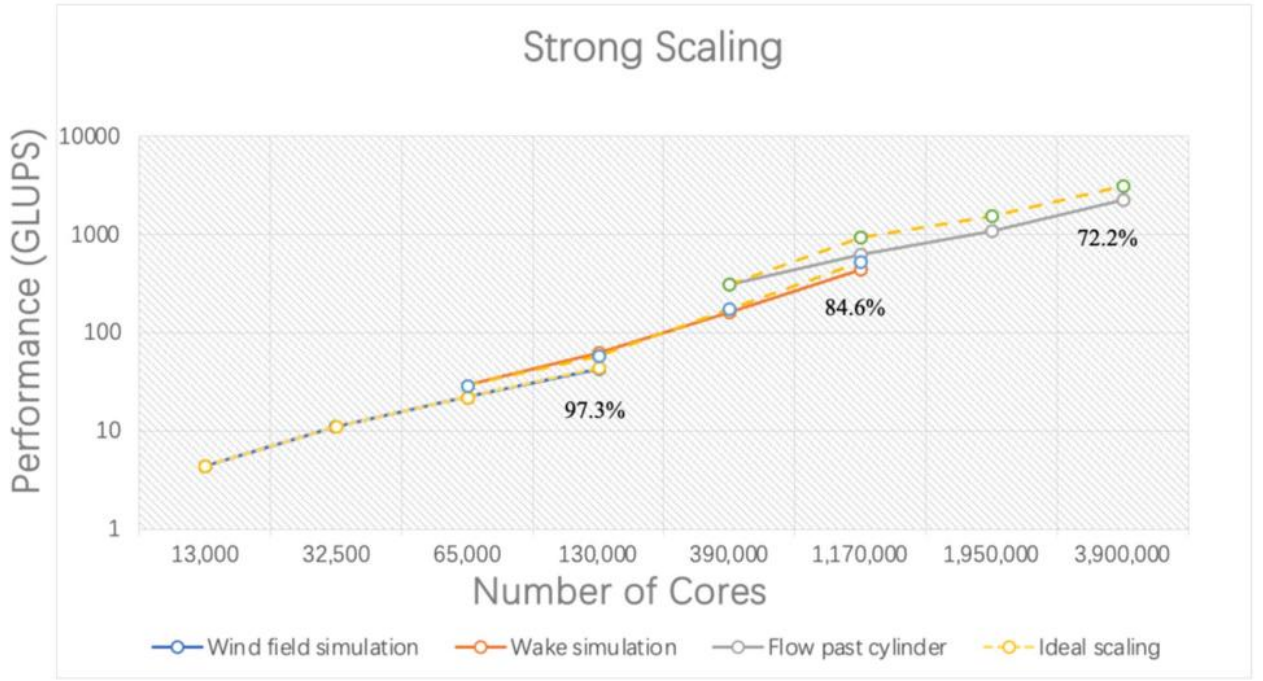
New-SunwayLB

弱扩展性性能和并行效率



强扩展性

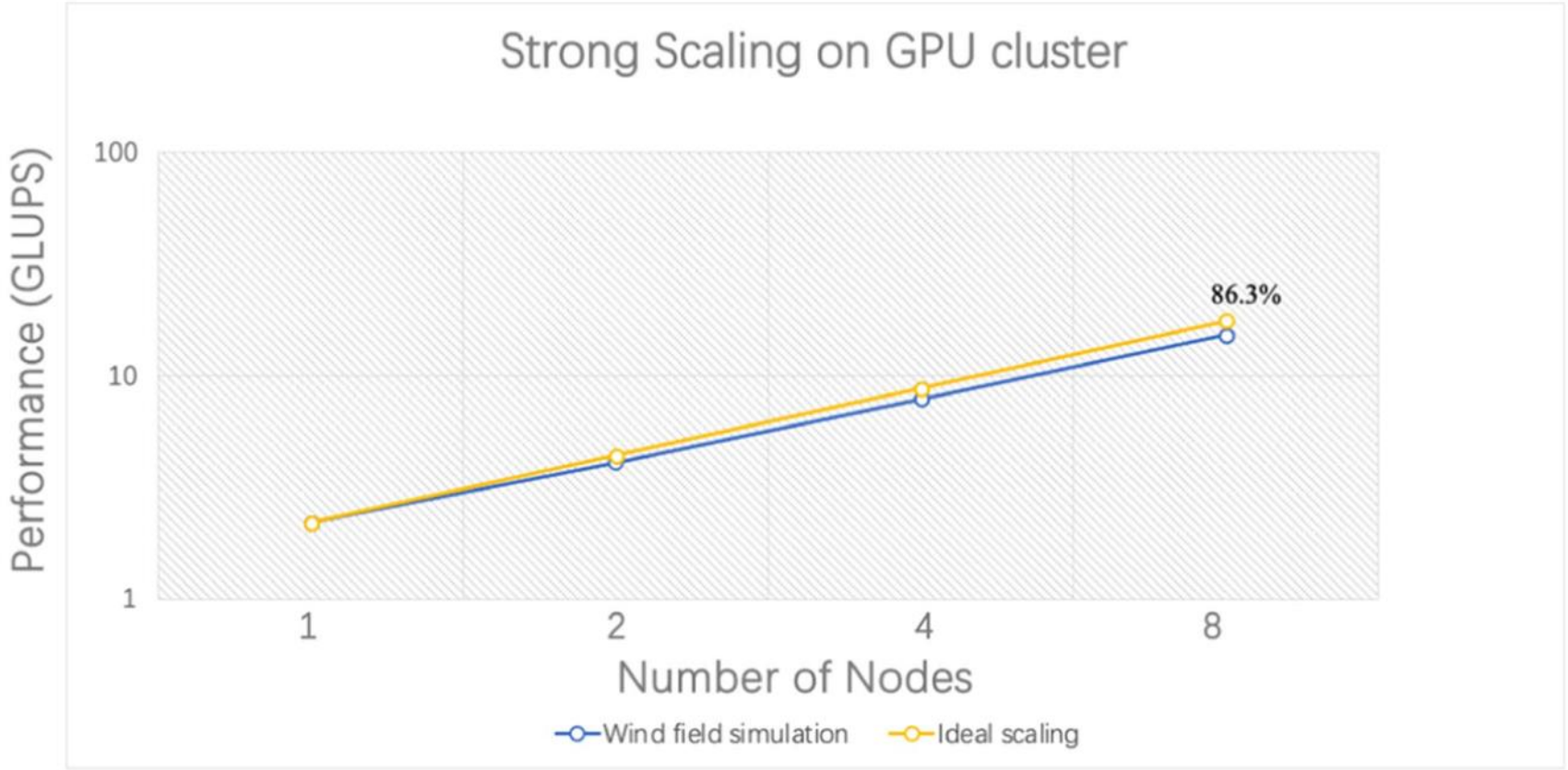
- 最大模拟实现72.2%并行效率



SunwayLB-GPU

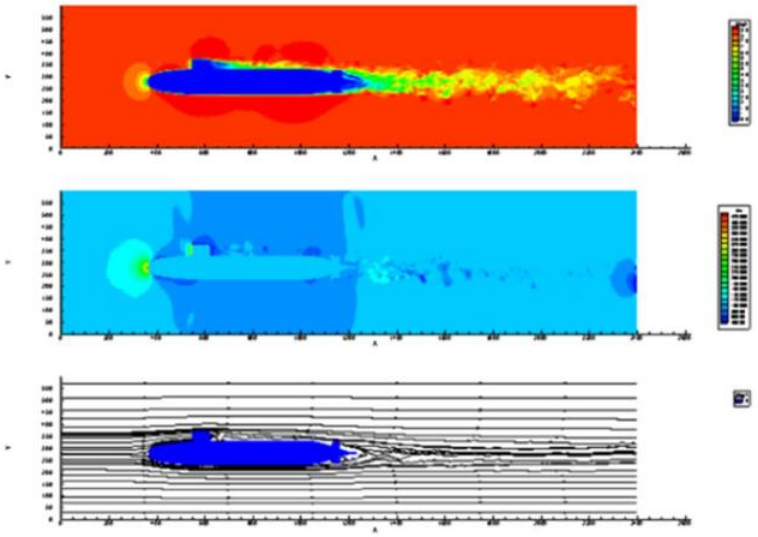
GPU集群上扩展性结果

- 1节点（8个GPU）扩展到8个节点（64个GPU），实现86.3%强扩展性

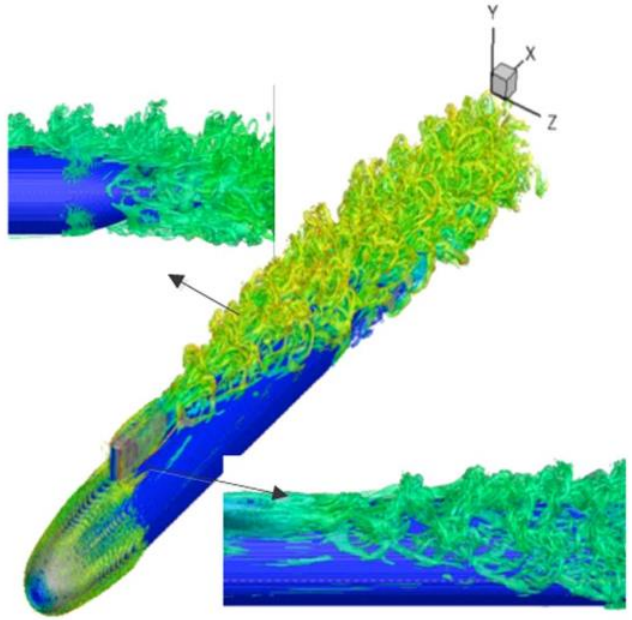


SunwayLB-工程实验

- **Suboff潜艇模型仿真，捕捉湍流用于分析潜艇的阻力机动性和噪声**
  - **两代系统分别实现了68.89%和84.6%并行效率**



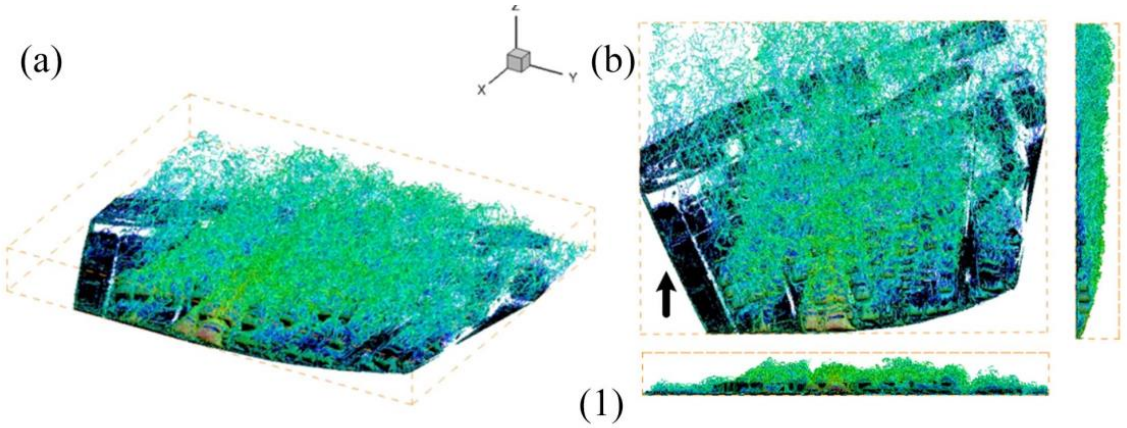
(1)

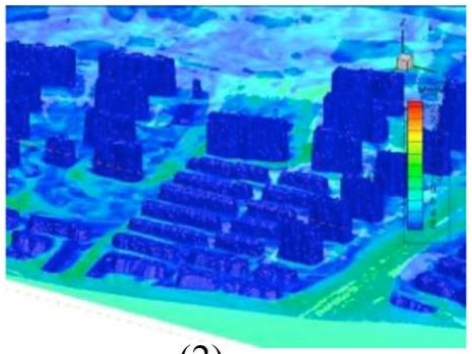


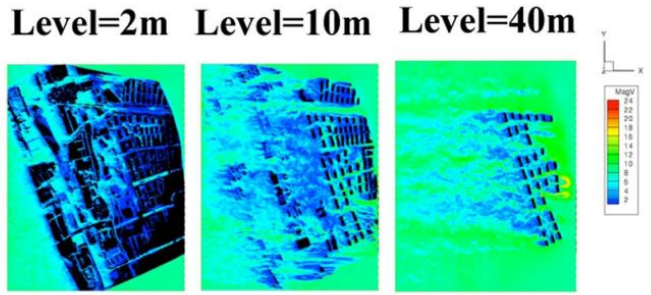
(2)

SunwayLB-工程实验

- 复杂城市地区风流模拟，了解改善湍流城市的风能利用
  - 实现了89%的并行效率，超过8000 GLUPS的性能




(2)



(3)

总 结

- 设计和构建了高度可扩展的LBM框架 SunwayLB
- 软硬件结合提出多种并行化策略
- 实现了基于SunwayLB的GPU版本并提出不同的优化技术
- 性能测试结果表明SunwayLB的优秀并行效率和内存带宽利用率
- 针对大规模工业应用的LBM的解决方案正成为现实

国家超级计算广州中心
NATIONAL SUPERCOMPUTER CENTER IN GUANGZHOU

Thanks