# HPCG on Tianhe2

Yutong Lu[1], Chao Yang[2], Yunfei Du[1]

1 National University of Defense Technology, Changsha, Hunan, China

2 Institute of Software, CAS, Beijing, China

# Outline

- ❑HPCG result overview on Tianhe-2
- ❑Key Optimization works
  - ➢Hybrid HPCG：CPU+MIC

# HPCG result

- ❒ **Tianhe-2 (Nudt-V11)**
  - ➢ **HPCG version 2.4**
  - ➢ **Hybrid code**
  - ➢ **Whole scale (with 3Mics each node)**
  - ➢ **Problem Size 136*176*176**
- ❒ **Result**
  - ➢ **623280GFlops**
  - ➢ **1.14% of peak performance**
  - ➢ **Efficiency 81.15%**

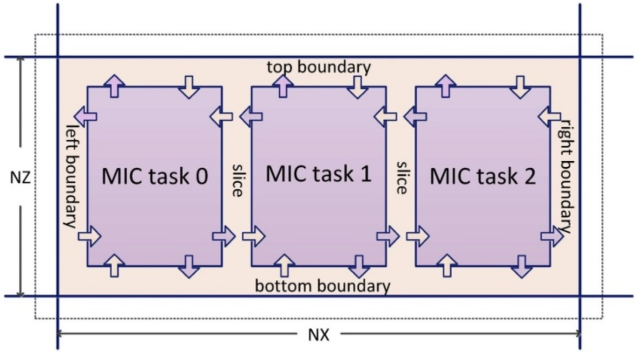
# Optimization

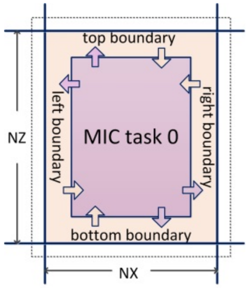
- ❑Intra-node
  - ➢Improve the performance of hybrid single node
- ❑Inter-node
  - ➢Improve the scalability
- ❑Choose the suitable problem size to balance the both aspects

# Optimization: Intra-node Partition

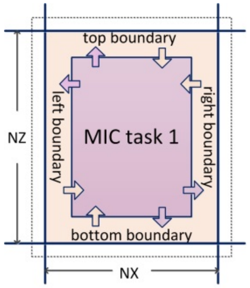
- ❐ **An inner-outer subdomain partition strategy**
  - ➢ **A regular inner parts for each MIC device, an irregular outer part for CPU**
  - ➢ **isolating MIC computation from MPI communication, avoiding data movement between different MIC devices, thus providing a chance for computation-communication overlapping**

- ❐ **Two alternatives**
  - ➢ **1 MPI process per node (old nudt-v06))**
    - ◆ 3 inner tasks + 1 outer task (per process)
    - ◆ Larger optimization overhead, because async memory alloc on MIC works poorly!
  - ➢ **3 MPI processes per node (new nudt-v11)**
    - ◆ 1 inner task + 1 outer task (per process)
    - ◆ 6 out of 8 CPU cores for each outer task

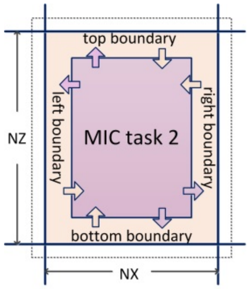




CPU 0~7



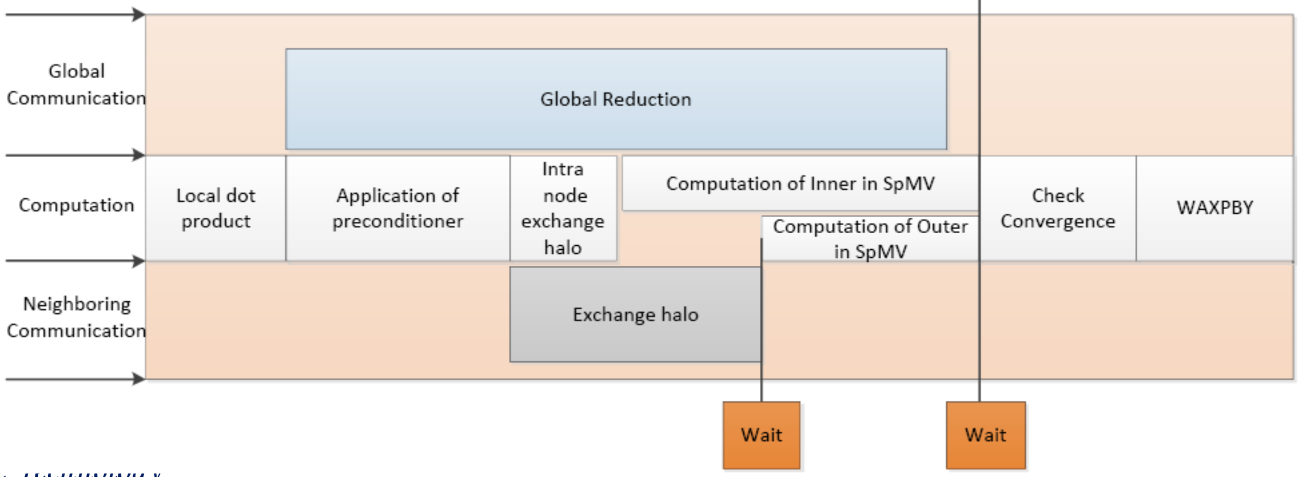
CPU 8~15



CPU 16~23

# Optimization: Optimizing MPI Communication

- ❒ **Pipelined CG (Ghysels 2013 ParCo) for global comm hiding**
  - ➢ Exactly mathematically equivalent to the standard CG
  - ➢ Only need one global communication for two dots and one norm per iteration
  - ➢ The global communication can be overlapped with preconditioner and SpMV
  - ➢ The number of WAXPBY's per iteration is increased from 3 to 8, the increased cost can be reduced by proper kernel fusions.
  - ➢ Overlapping neighboring communication with computation in SpMV
  - ➢ Based on the inner-outer subdomain partition
  - ➢ Halo exchange is overlapped with the computation of the inner part



Global Communication

Global Reduction

Computation

Local dot product

Application of preconditioner

Intra node exchange halo

Computation of Inner in SpMV

Computation of Outer in SpMV

Check Convergence

WAXPBY

Neighboring Communication

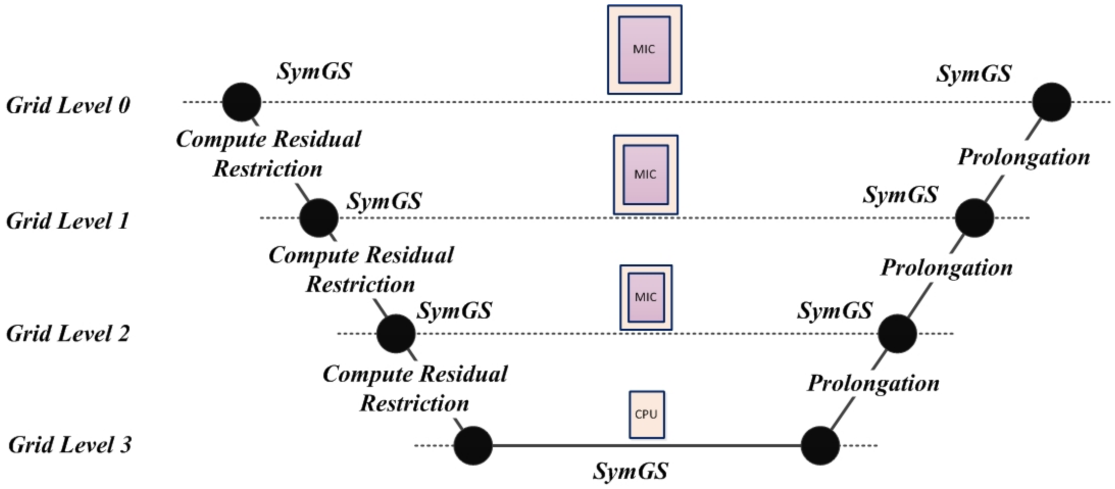
Exchange halo

Wait

Wait

# Optimization: Load Balance between CPU and MIC

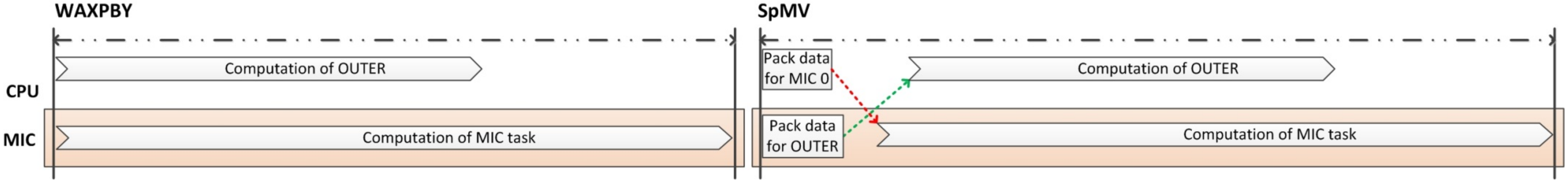
- ❒ 4 level (ratio 2) V-cycle geometric multigrid preconditioner
  - ➢ Load-imbalance exists between CPU and MIC if using inner-outer partitions on all levels (the outer thickness on finest level is at least 8)
  - ➢ Adjusting the outer thickness on finest level to be 4
    - ◆ Hybrid inner-outer partitions on grid levels 0, 1, 2
    - ◆ CPU-only partition on grid level 3
    - ◆ Some extra PCI-express transfer is needed to pull the inner blocks from MIC to CPU



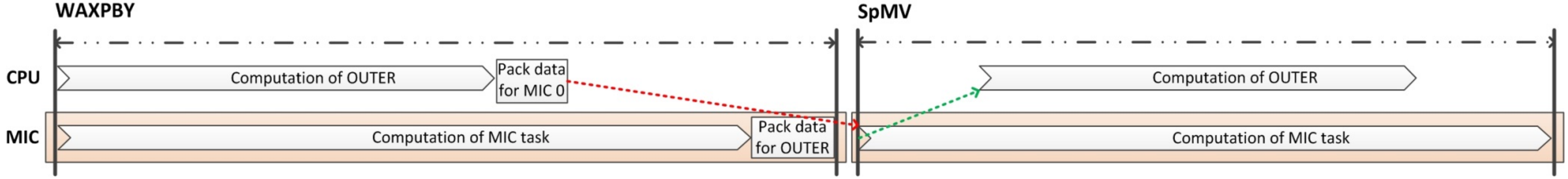
# Optimization: Asynchronous Data Transfer Scheme

- Data movement is needed in SpMV and SymGS.

**Pack and exchange the halo information at the beginning of the current kernel**



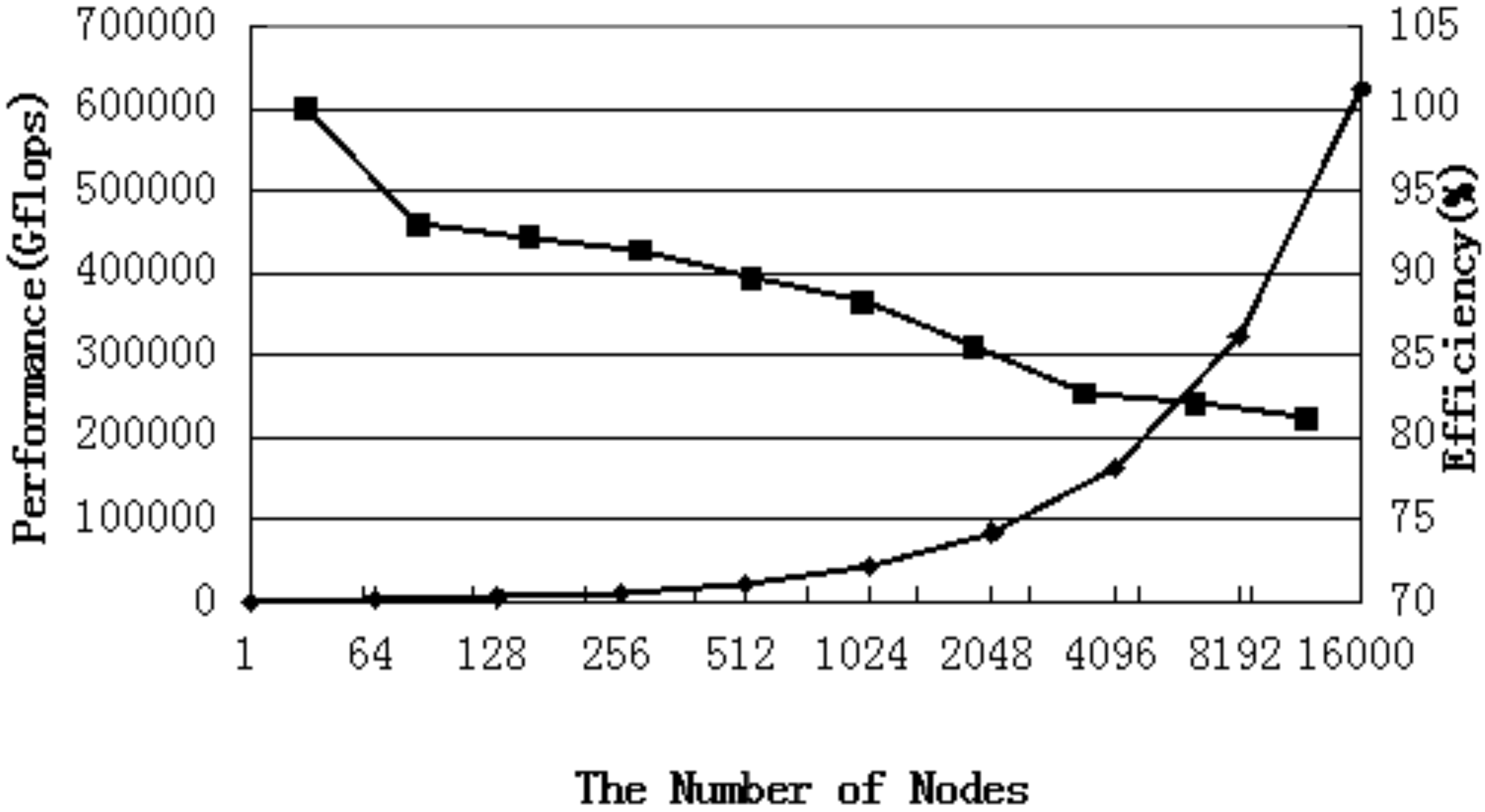
**Exploit the CPU's waiting time to pack and transfer data from CPU to MIC device in the preceding kernel, thus eliminate the MIC's waiting time.**



# Optimization: Others on MIC and CPU

- ❒ Sparse matrix storage format
  - ➢ SELLPACK on MIC, ELLPACK on CPU
- ❒ SIMDization
  - ➢ Using *gather* and *streaming store* instructions on MIC
- ❒ Different red-black reordering methods
  - ➢ Block multi-color ordering on grid level 0
  - ➢ Fusing the forward and backward sweep on other levels
- ❒ Different parallel methods
  - ➢ Multi-level parallelism on MIC side
- ❒ Optimization of communication among OpenMP threads
  - ➢ Employing light-weight kernels such as *WaitNeighbors* and *IntraBarrier* instead of a global barrier

# Scalability



# Future Directions

## ❑ Further Optimization

- Continue to improve the hybrid method
- Communication optimization for network topology aims to improve efficiency

ytlu@nudt.edu.cn
yangchao@iscas.ac.cn