東京大学
THE UNIVERSITY OF TOKYO

メニィコアクラスタ向け並列多重格子法アルゴリズム

中島　研吾

東京大学情報基盤センター

東京大学大学院情報理工学系研究科数理情報学専攻

2012年8月29日

日本応用数理学会2012年度年会@稚内

TOC

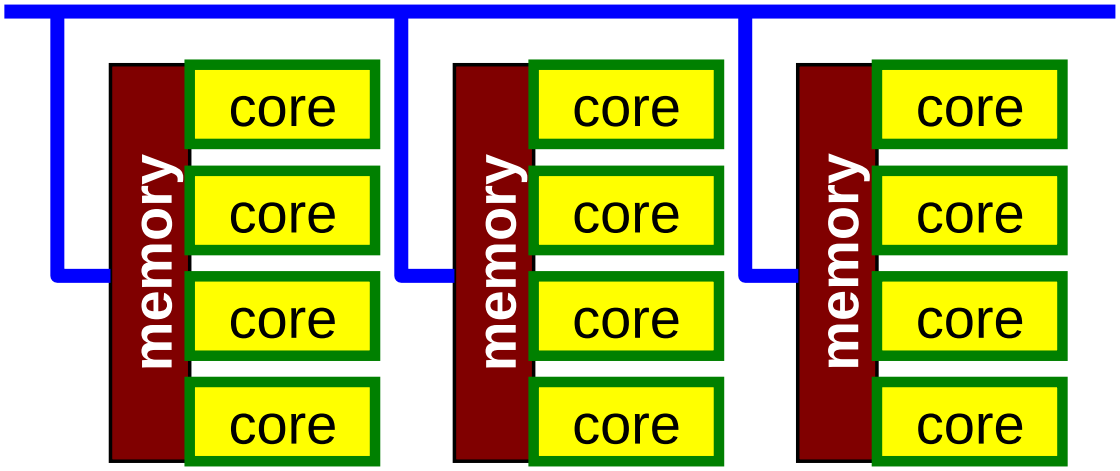
本研究の背景

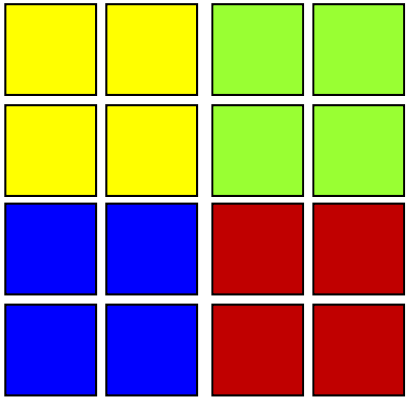
- マルチコアクラスタ向け並列多重格子法，有限体積法
  - T2K東大
  - Fujitsu PRIMEHPC FX10（Oakleaf-FX）

- Flat MPI vs. Hybrid（OpenMP+MPI）
- Hybrid並列プログラミングモデルへの期待
  - MPIプロセス数（幾何学的領域数）を減らすことができる
  - O(10^8-10^9)-way MPIはスケールしないだろう（Exascaleシステム）
  - ヘテロジニアスなアーキテクチャへの拡張が容易
    - CPU+GPU，CPU+ManyCore（e.g. Intel Xeon Phi）
    - MPI+X： OpenMP，OpenACC，CUDA，OpenCL

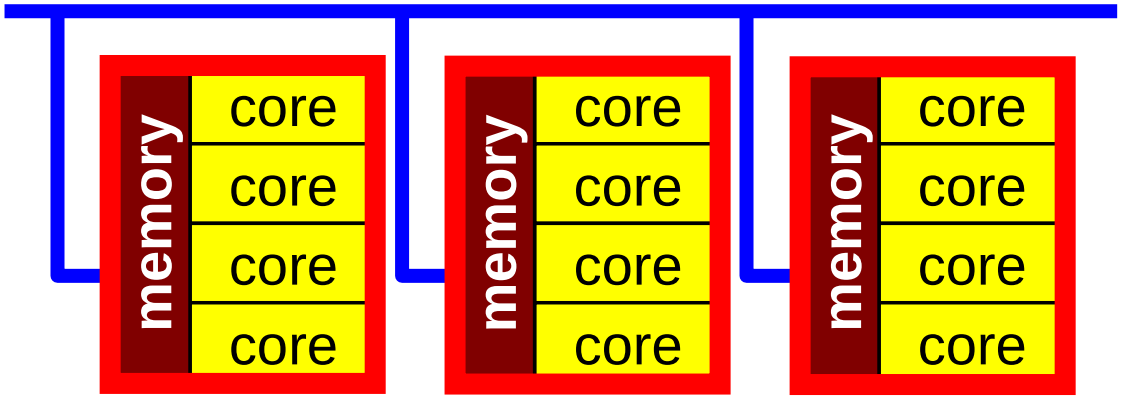
Flat MPI vs. Hybrid

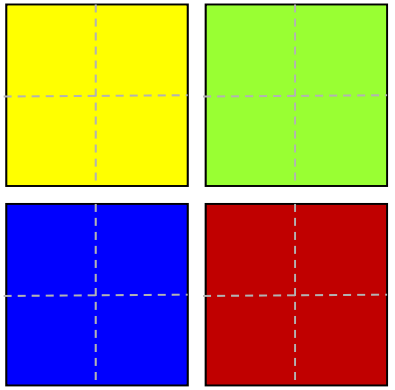
Flat-MPI：Each PE -> Independent





Hybrid：Hierarchal Structure

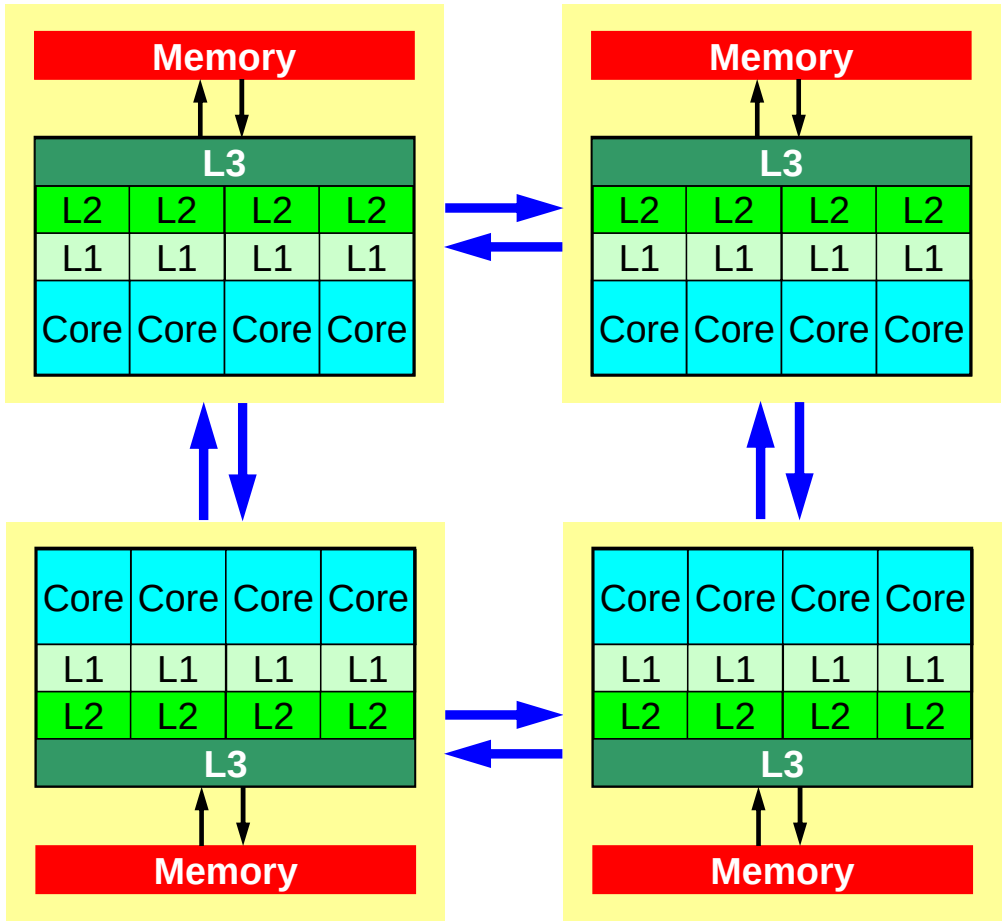


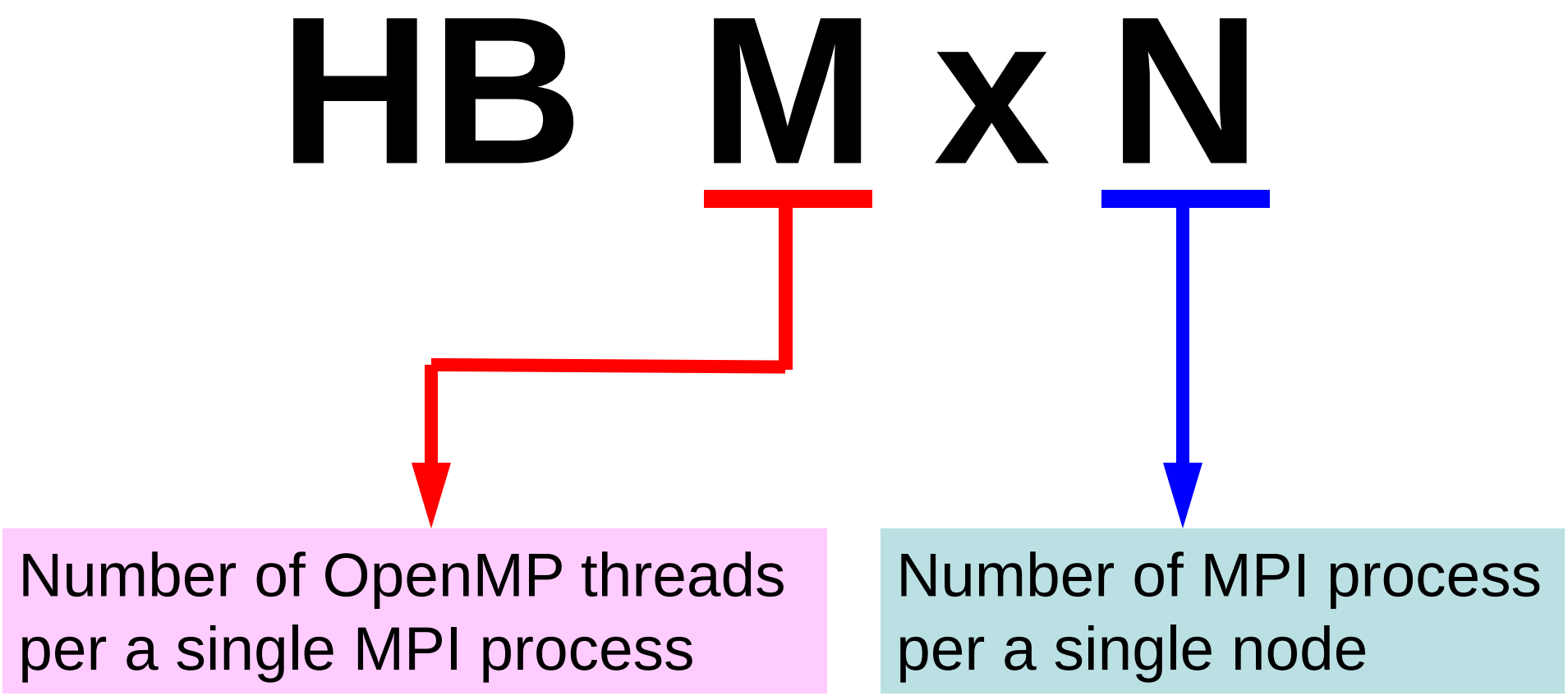


- 背景
- ハードウェア・ソフトウェアの概要
- T2K東大による計算結果
- 更なる最適化：Coarse Grid Aggregation
- まとめ

T2K東大の計算ノード

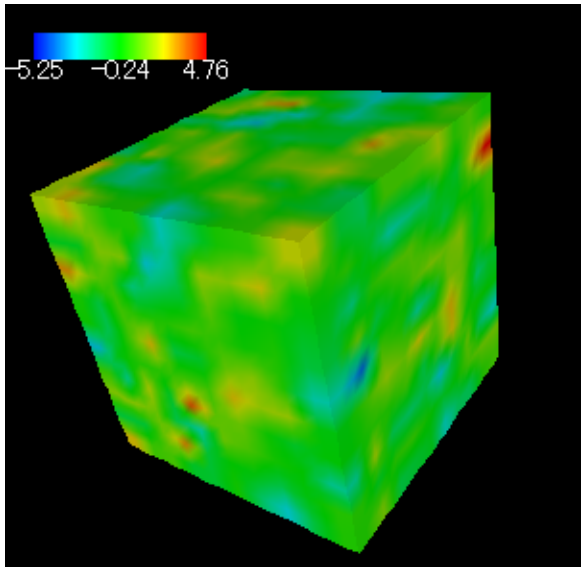
- AMD Quad-core Opteron (Barcelona) 2.3GHz
- 4 “sockets” per node
  - 16 cores/node
- Multi-core, multi-socket system
- cc-NUMA architecture
  - careful configuration needed
    - local data ~ local memory




HB M x N
Number of OpenMP threads per a single MPI process
Number of MPI process per a single node

Target Application

- 3D Groundwater Flow Heterogeneous Porous Media
  - Poisson's equation
  - Randomly distributed water conductivity
  - Distribution of water conductivity is defined through methods in geostatistics〔Deutsch & Journel, 1998〕
- Finite-Volume Method on Cubic Voxel Mesh
- Cyclic Structure of Heterogeneity for every 128^3 grids in each direction

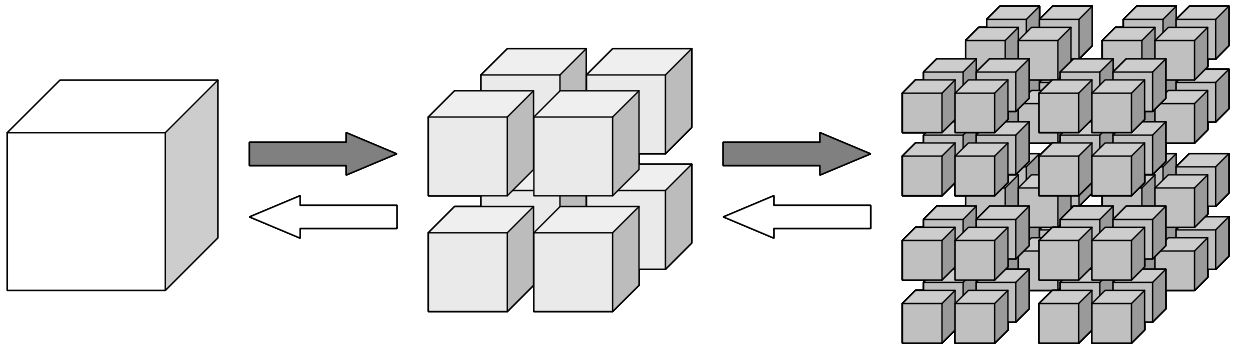


Movie

Linear Solvers

Multigrid

- Preconditioned CG Method
  - Multigrid Preconditioning (MGCG)
  - IC(0) for Smoothing Operator (Smoother)
- Parallel Geometric Multigrid Method
  - 8 fine meshes (children) form 1 coarse mesh (parent) in isotropic manner (octree)
  - V-cycle
  - Domain-Decomposition-based: Block-Jacobi
  - Operations using a single core at the coarsest level (redundant)



Hardware/Software

ASDD

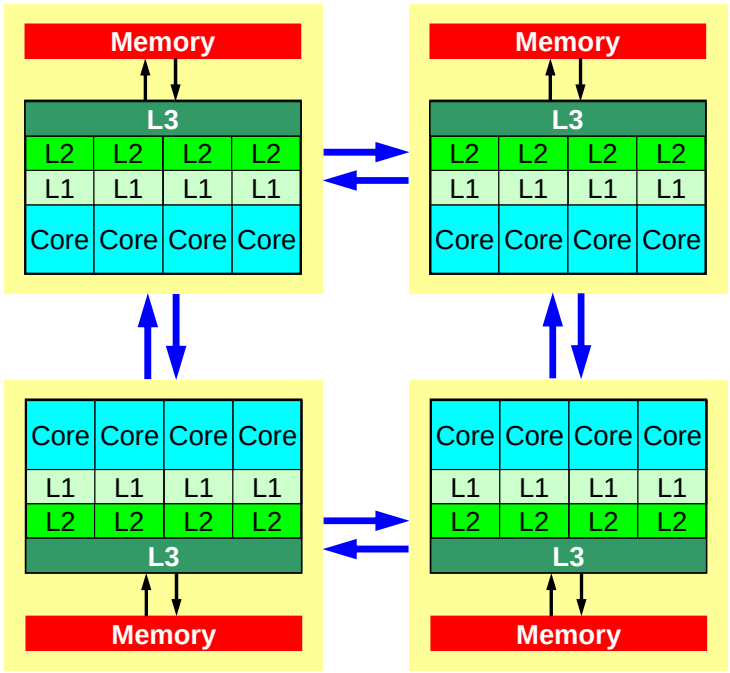
- T2K/Tokyo
  - up to 512 nodes (8,192 cores, 75.4 TFLOPS)
- Program
  - Hitachi FORTRAN90 + MPI: Isend/Irecv/Waitall
  - CRS matrix storage
  - CM-RCM Reordering for OpenMP

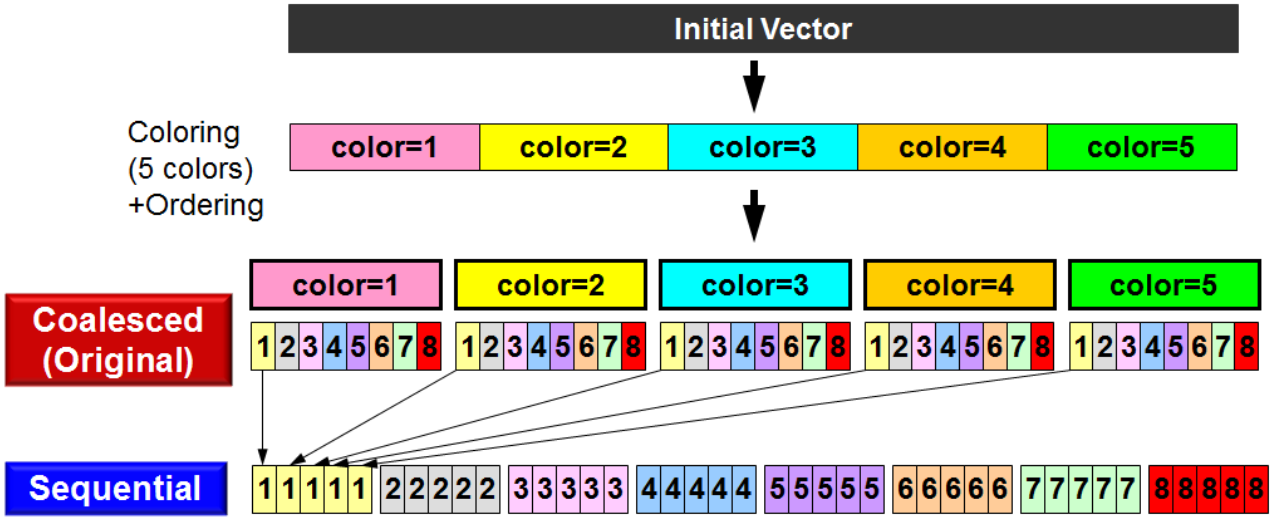
```
for (i=0; i<N; i++) {
  for (k=Index(i-1); k<Index(i); k++{
    Y[i]= Y[i] + A [k]*X[Item[k]];
  }
}
```

- $|Ax-b|/|b|=10^{-12}$ for Convergence
- Heterogeneity
  - Ratio of MAX/MIN water conductivity= 10^{10} (10^{-5}~10^{+5})
- Multigrid Cycles
  - 1 V-cycle/iteration
  - 2 smoothing iterations for restriction/prolongation at every level
  - 1 ASDD iteration cycle for each restriction/prolongation

様々な最適化

- cc-NUMA Architecture
  - NUMA Control
  - First Touch Data Placement
  - Further Reordering with Contiguous/Sequential Memory Access
- Coarse Grid Solver





関連研究

- 近年，OpenMP/MPIハイブリッド並列プログラミングモデルを使用した並列多重格子法の研究は盛ん（LLNL, Sandia）
  - Scalable method
  - Mulricore clusters, Heterogeneous architecture

- Alison Baker (LLNL) et al., “Challenge of Scaling Algebraic Multigrid across Modern Multicore Architectures” (IPDPS 2011)
  - Hypre Library (BoomerAMG), weak scaling
  - IBM BG/P, Cray XT5 using $O(10^5)$ cores
  - MultiCore SUPport library (MCSup)：cc-NUMA向け自動最適化
  - HB 4×4 is the best

Weak Scaling

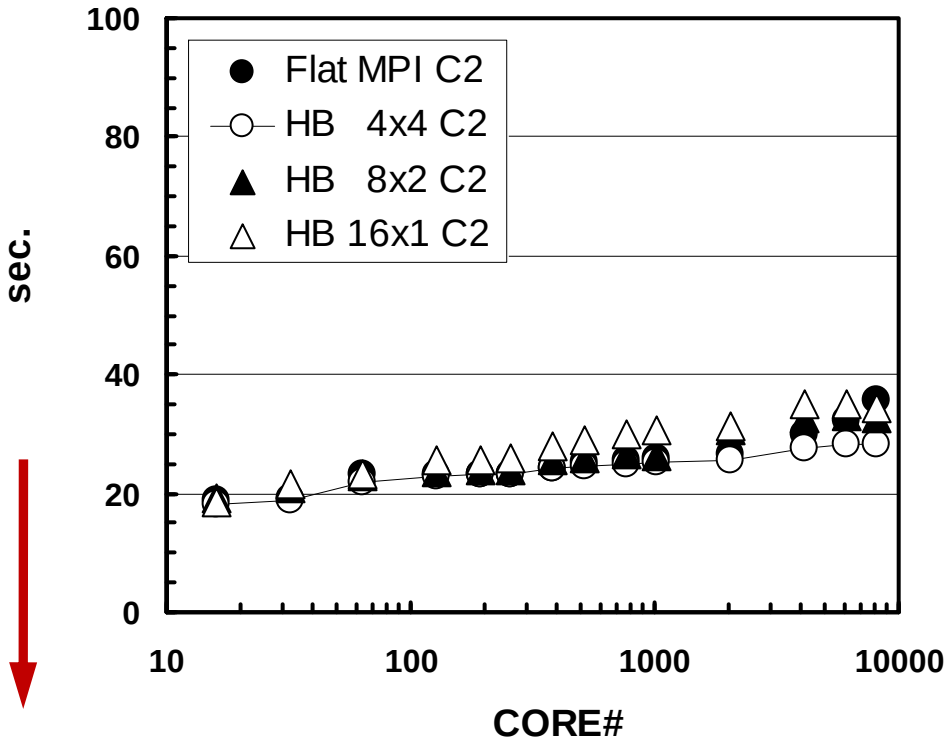
- Up to 8,192 cores (512 nodes)
  - 64^3 cells/core
  - 2,147,483,648 cells
- CM-RCM(2)

Weak Scaling

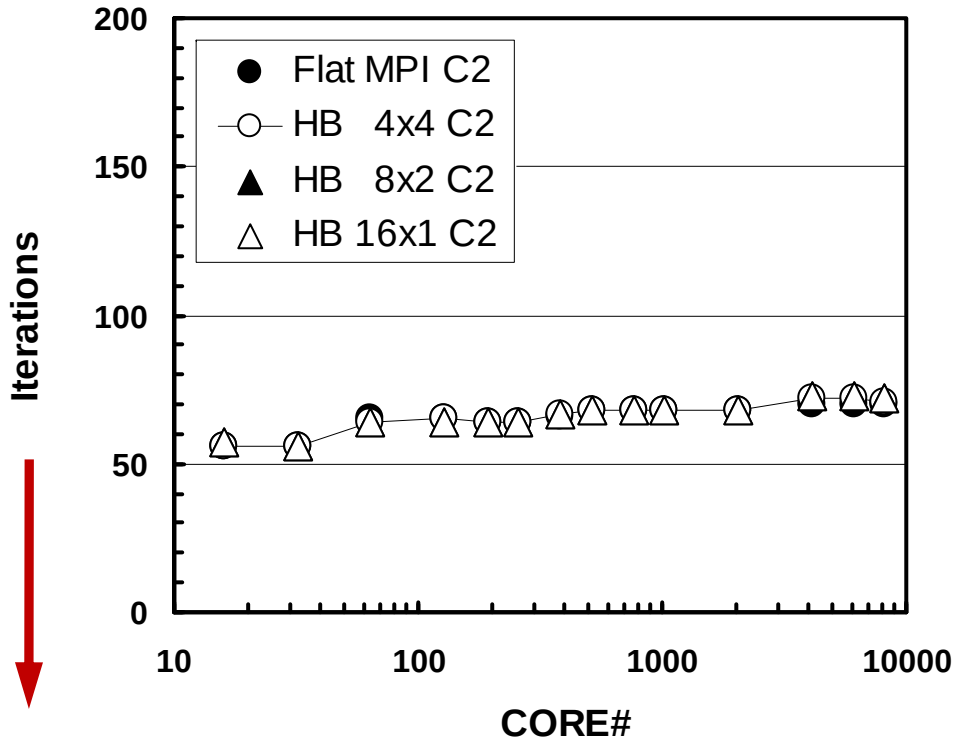
64^3cells/core, up to 8,192 cores (2.15×10^9 cells)

at 8,192 cores: Flat MPI(35.7sec), HB 4x4(28.4), 8x2(32.8), 16x1(34.4)

sec.



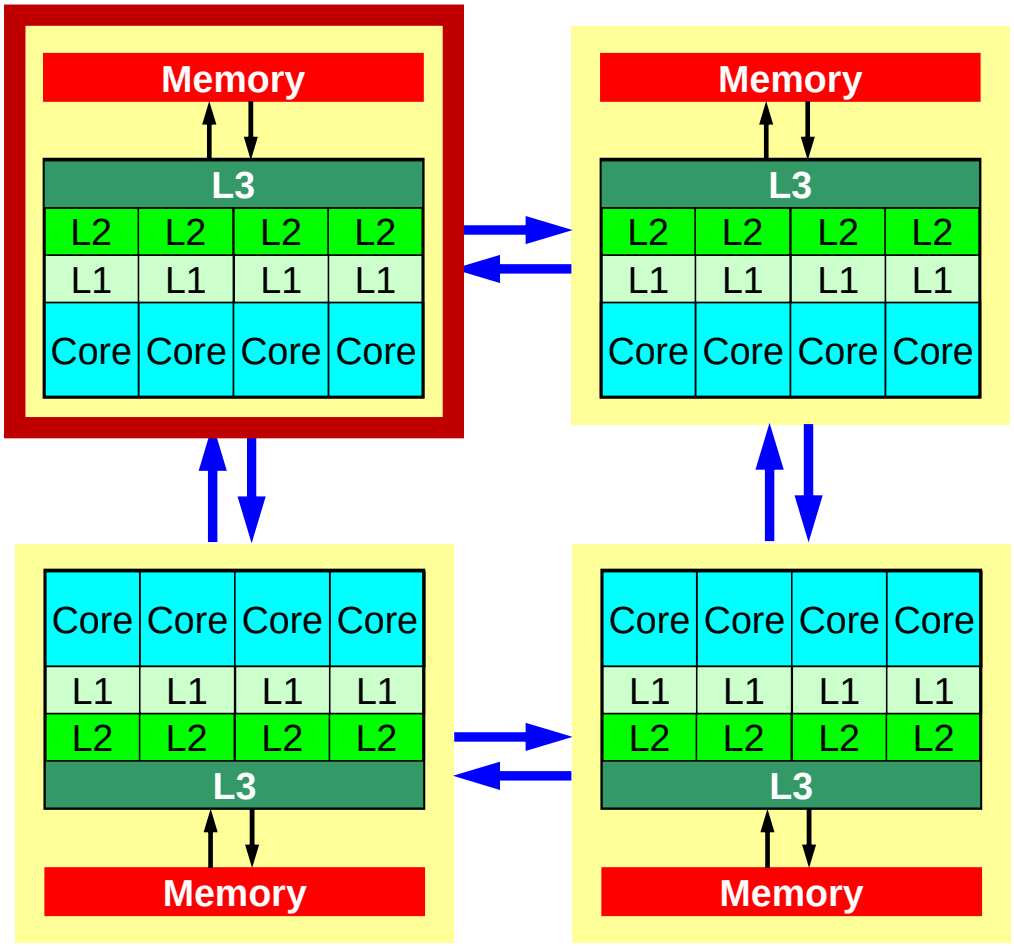
Iterations



Down is good

HB 4x4が最も性能が良い

- ソケット当たり1MPIプロセス
- 各プロセスのデータが同じメモリ上にあることが保証されている



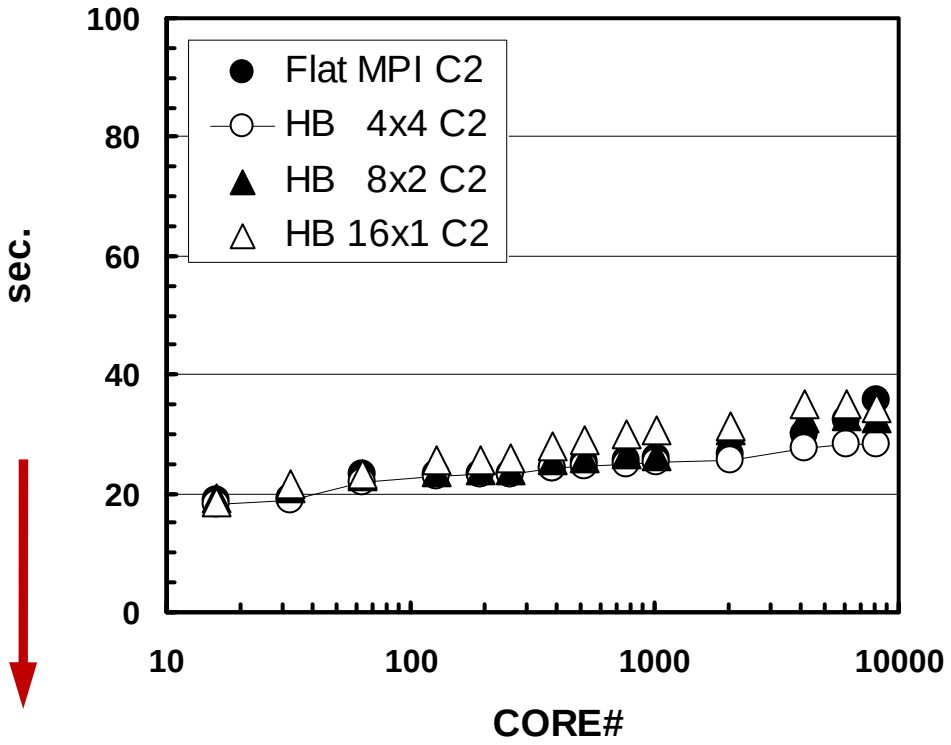
- 背景
- ハードウェア・ソフトウェアの概要
- T2K東大による計算結果
- 更なる最適化：Coarse Grid Aggregation
- まとめ

It's not really scalable

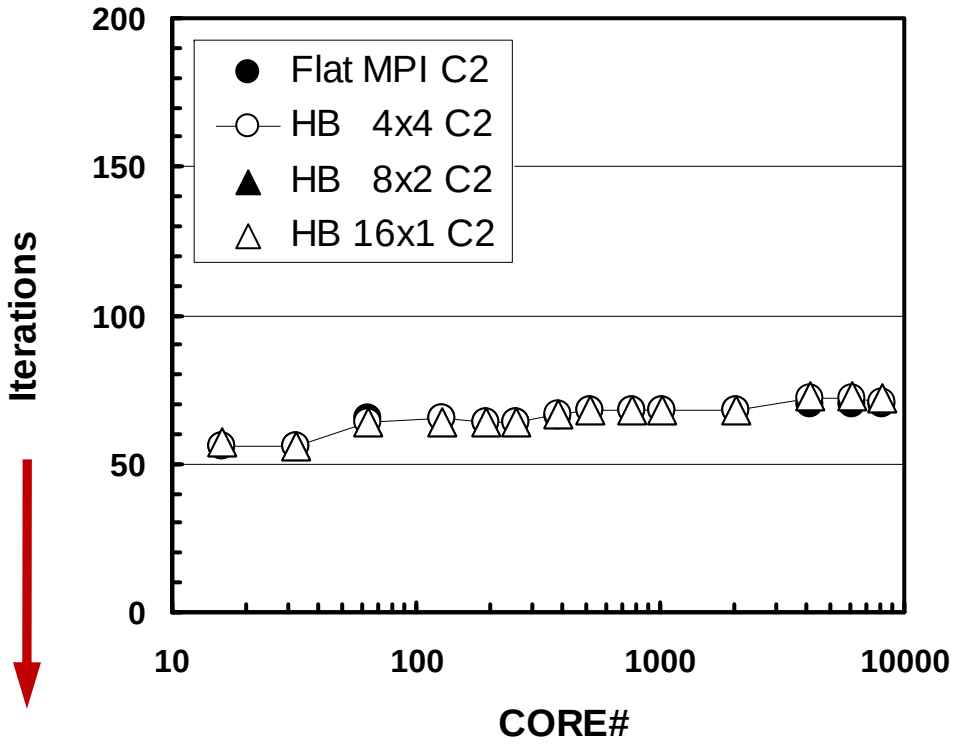
64^3cells/core, up to 8,192 cores (2.15×10^9 cells)

at 8,192 cores: Flat MPI(35.7sec), HB 4x4(28.4), 8x2(32.8), 16x1(34.4)

sec.

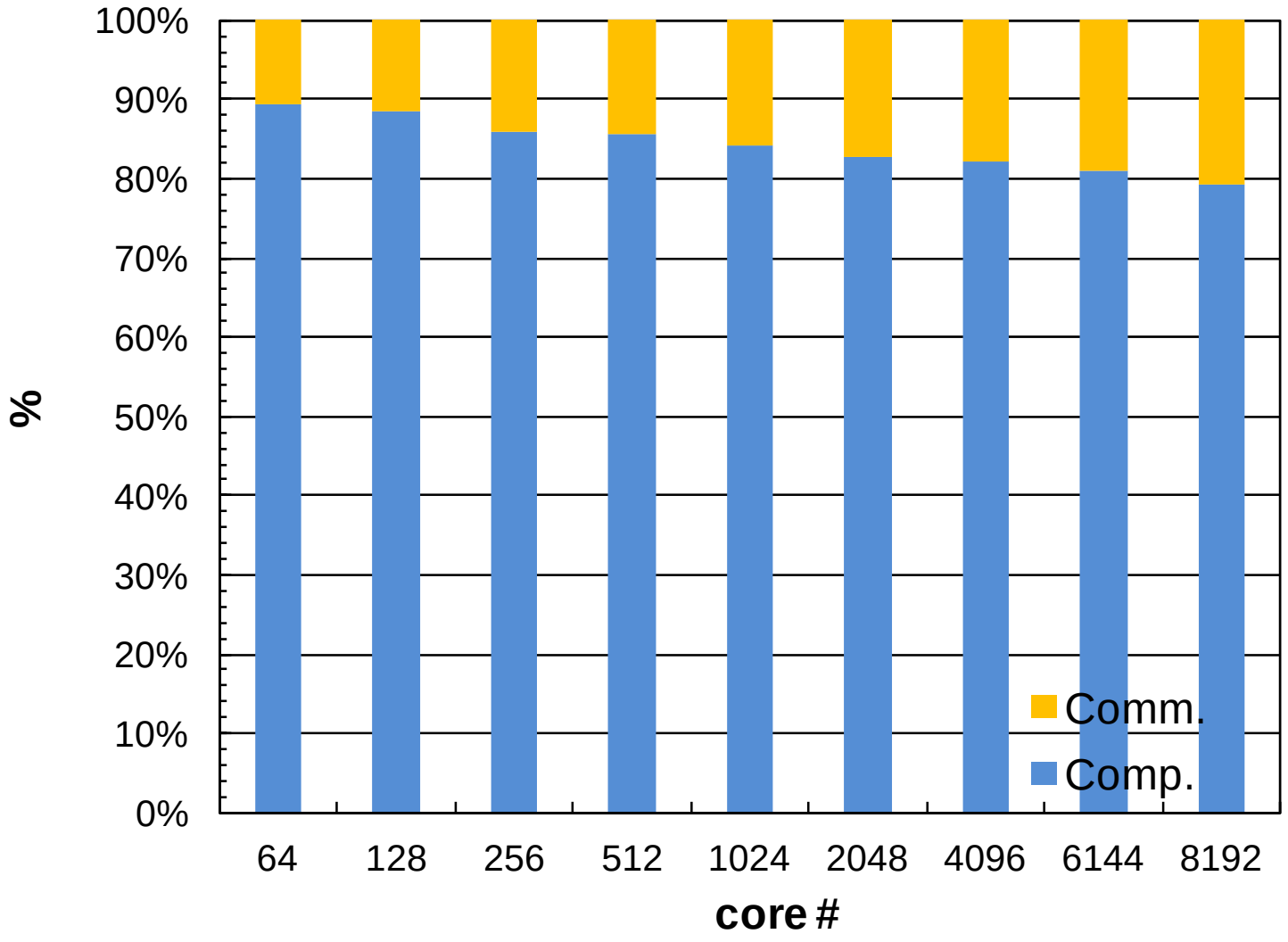


Iterations



Down is good

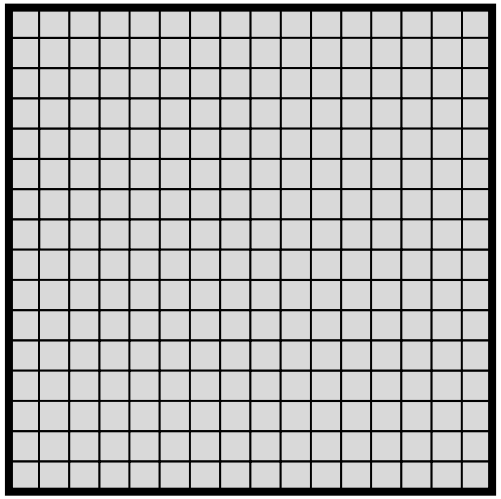
Weak Scaling: HB 4x4 Computation vs. Communication



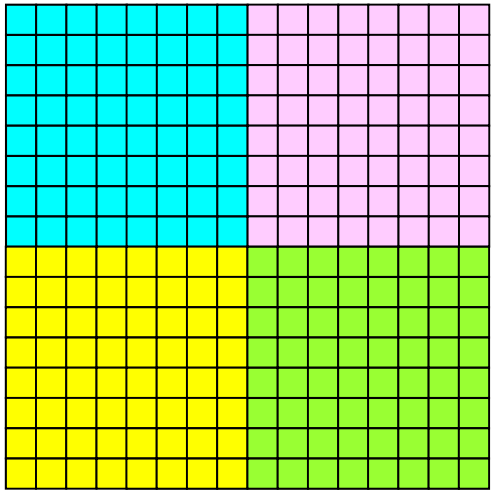
Remedies for Communication Reducing

- Development of new algorithms with reduced communications, where trade-off between efficiency and robustness is very critical.
  - skipping communications: unstable
- More practical approach is *aggregating* MPI processes at coarser levels.
  - Coarse grid solvers of the current work only utilized a single core of 16 cores on a multi-core/multi-socket node.
  - This approach has been adopted.

Initial Fine Grid

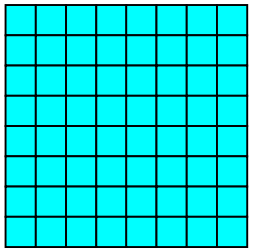


4 Regions (= MPI Processes)

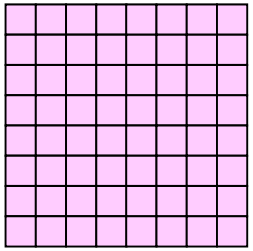


4 Regions (= MPI Processes)

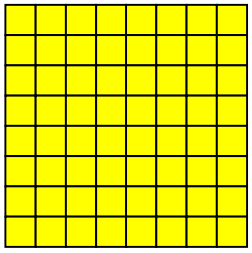
PE#0



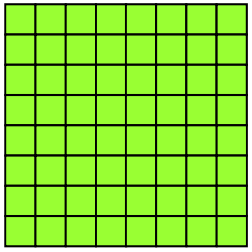
PE#1



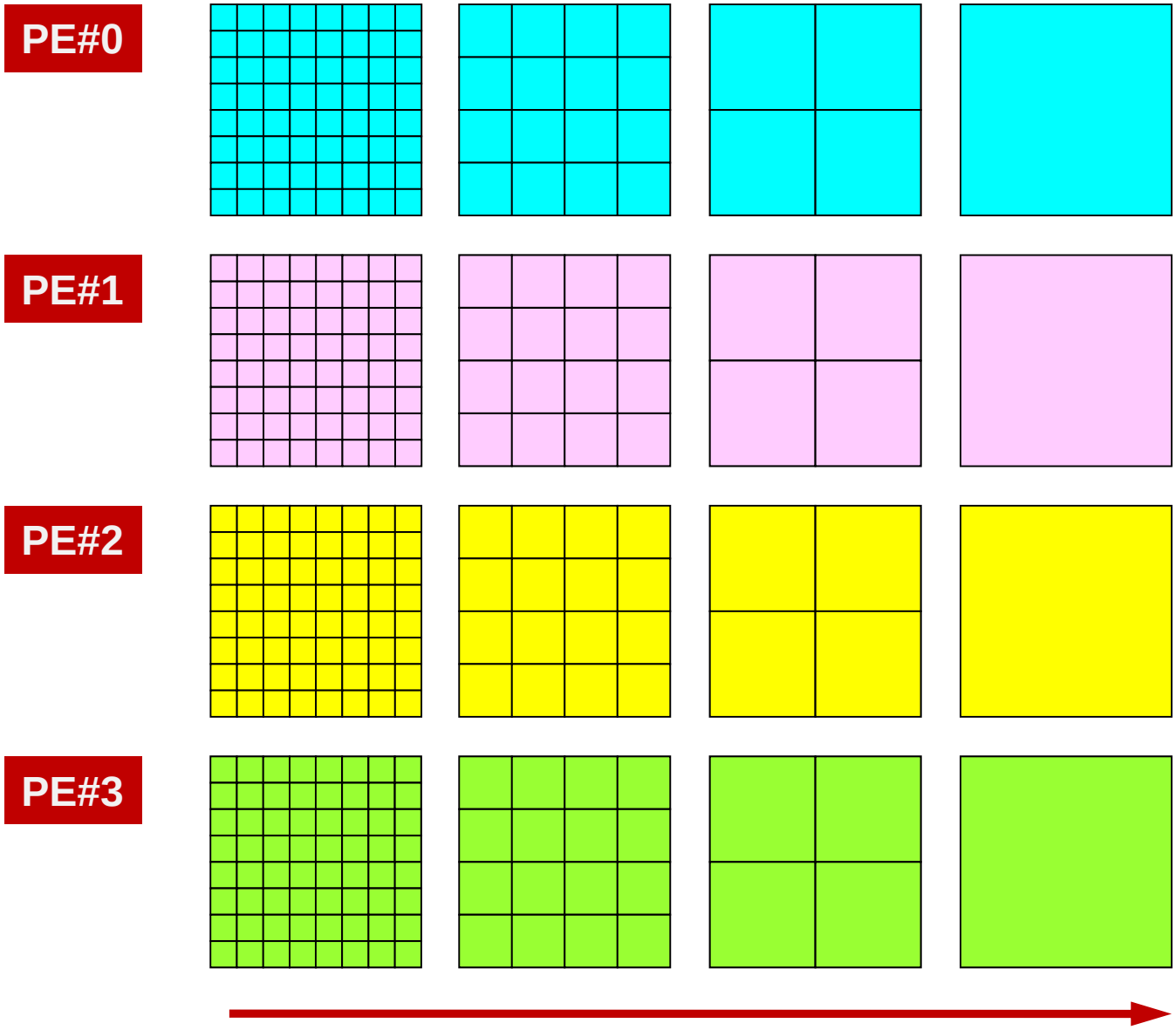
PE#2



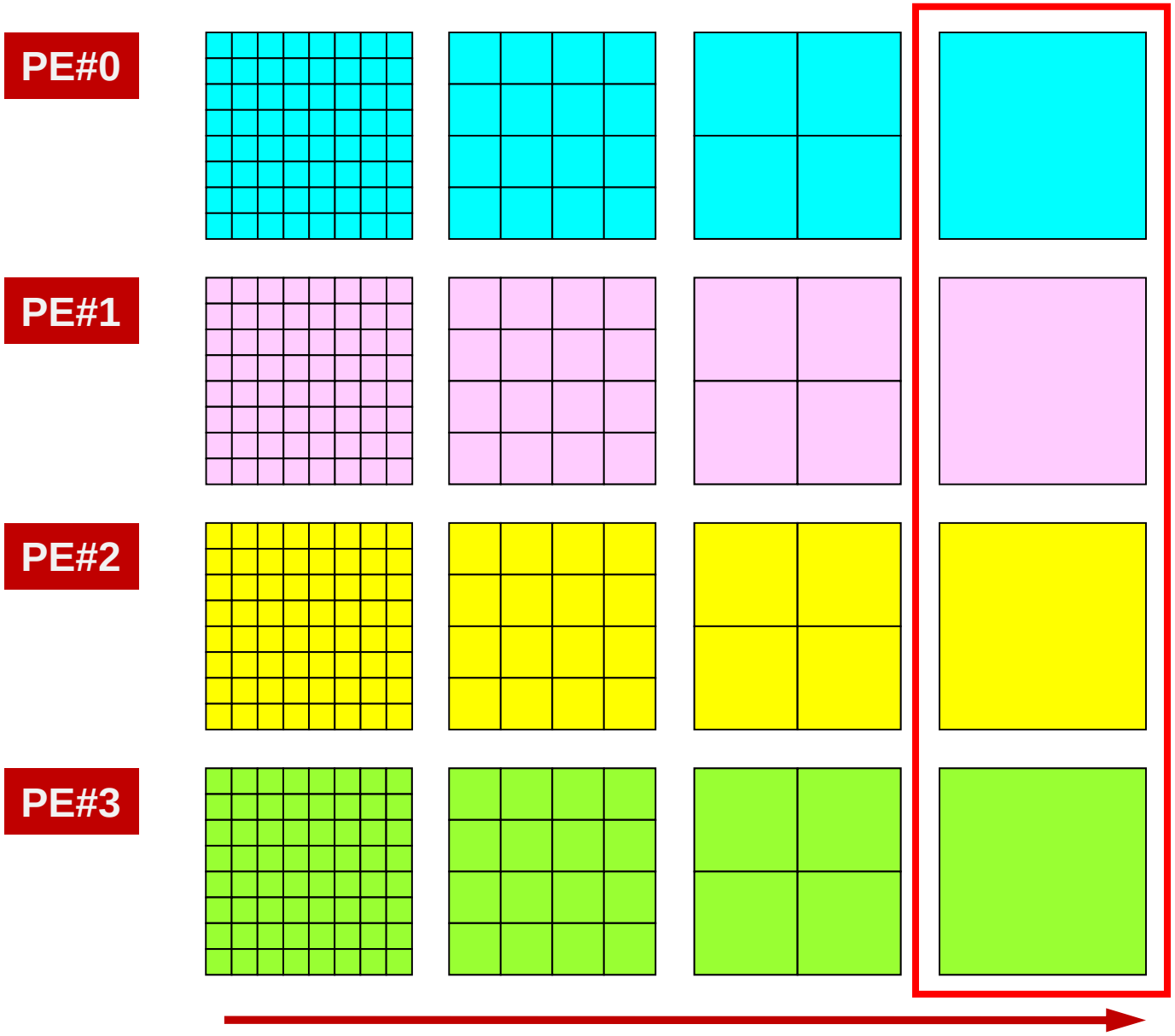
PE#3



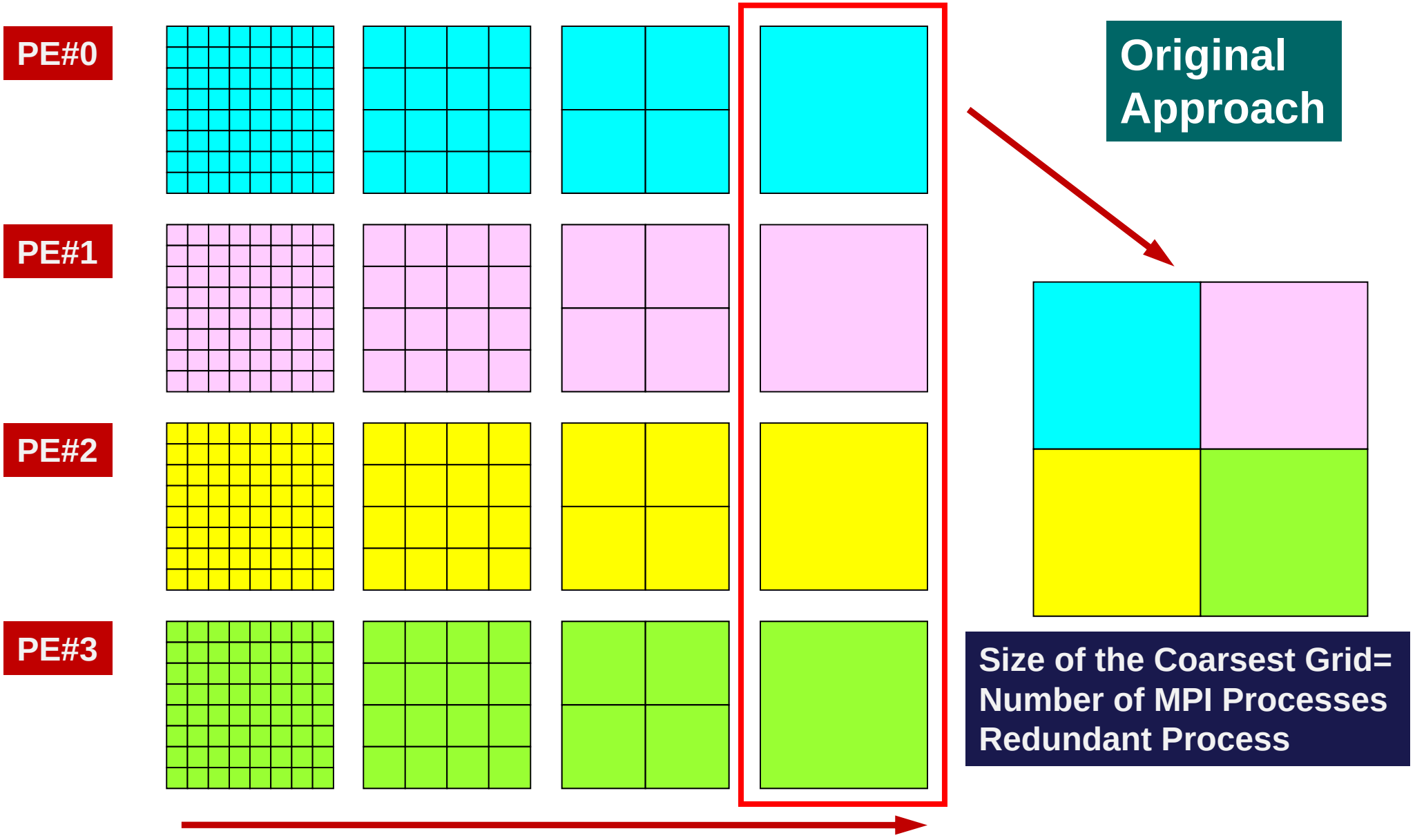
Coarsening, Restriction



Coarsening, Restriction

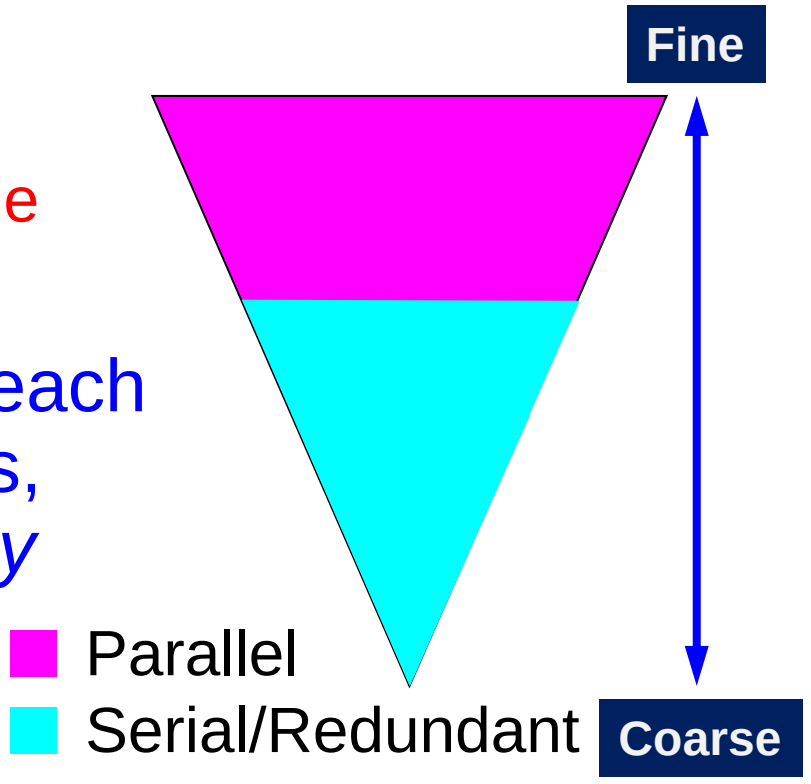


Coarse Grid Solver on a Single Core

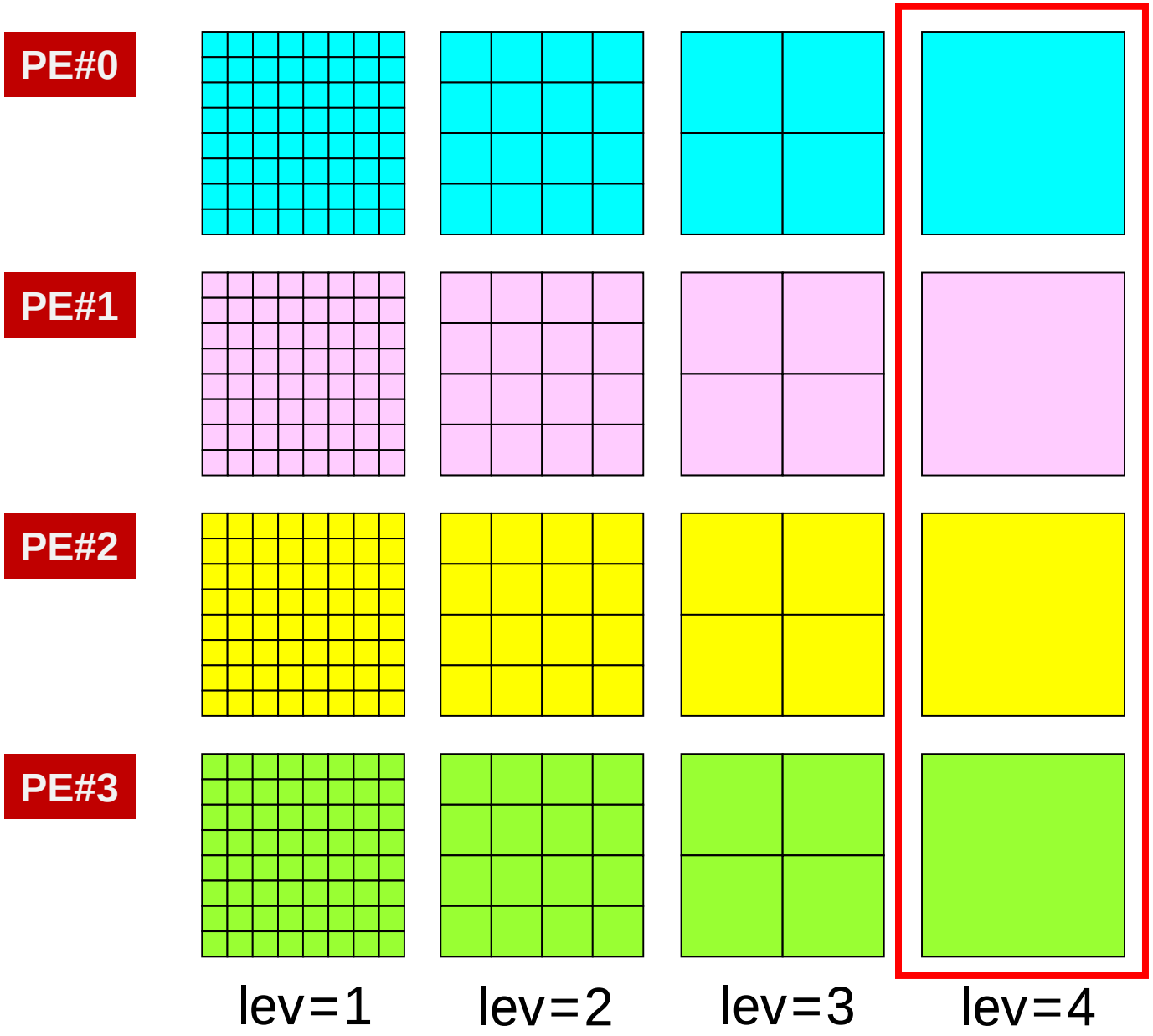


Strategy: Coarse Grid Aggregation

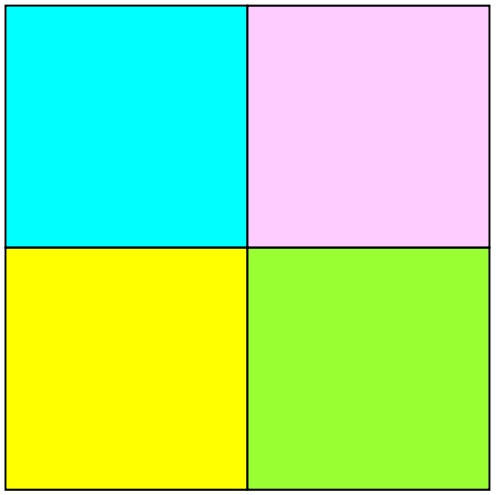
- 粗いレベルでMPIプロセス数を減らす
- Serial／Redundant処理をするレベルを拡張する
  - 粗いレベルでのノード間通信を減らすことができる
- Utilize a node or socket, not a single core
  - HB 4x4: Single Socket
  - HB 8x2: Two Sockets
  - HB 16x1: Four Sockets, Single Node
    - OpenMP is needed
- In post-peta/exa-scale systems, each node will consists O(10^2) of cores, therefore utilization of these *many cores* on each node should be considered.



Coarse Grid Solver on a Single Core

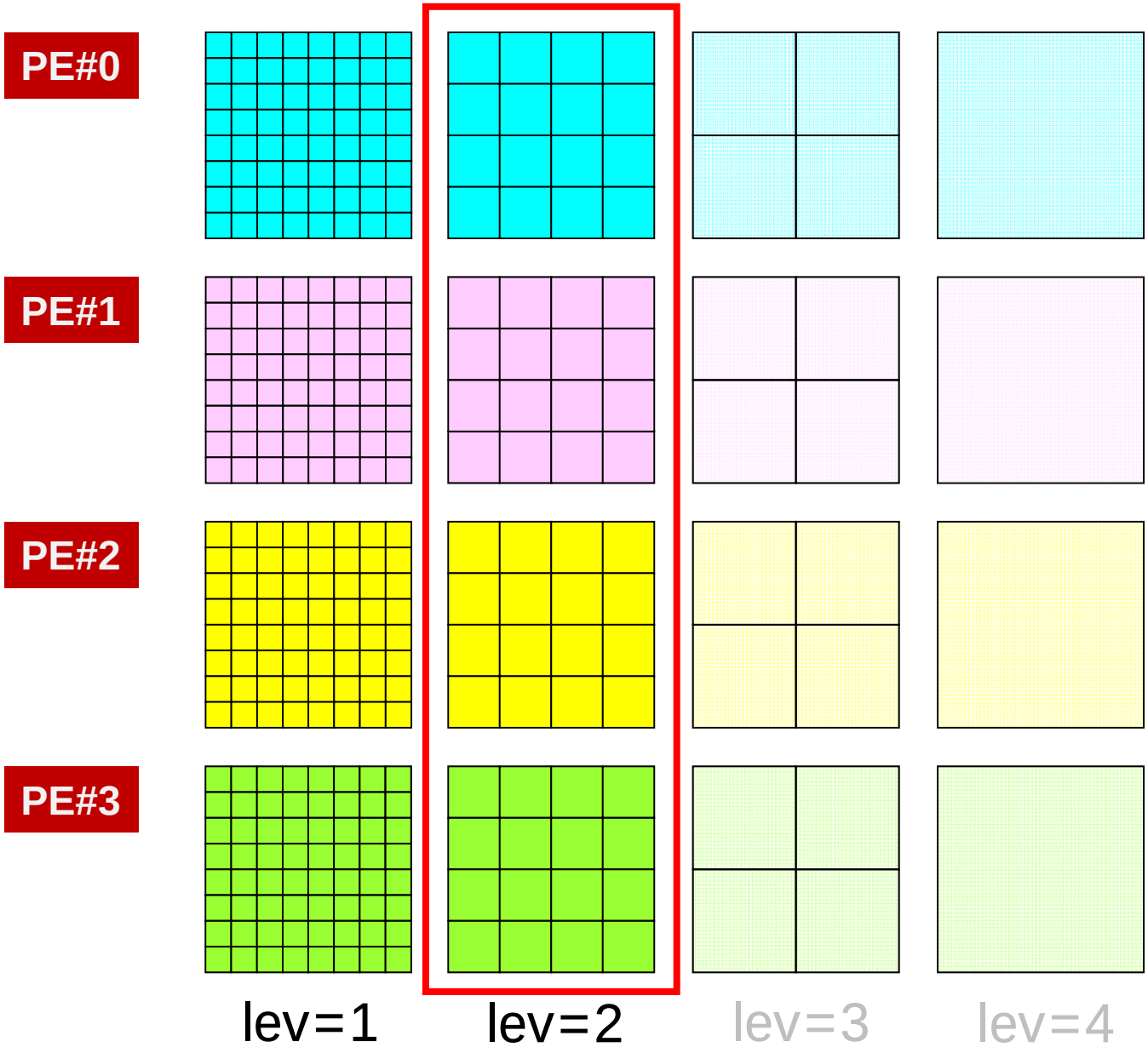


Original Approach

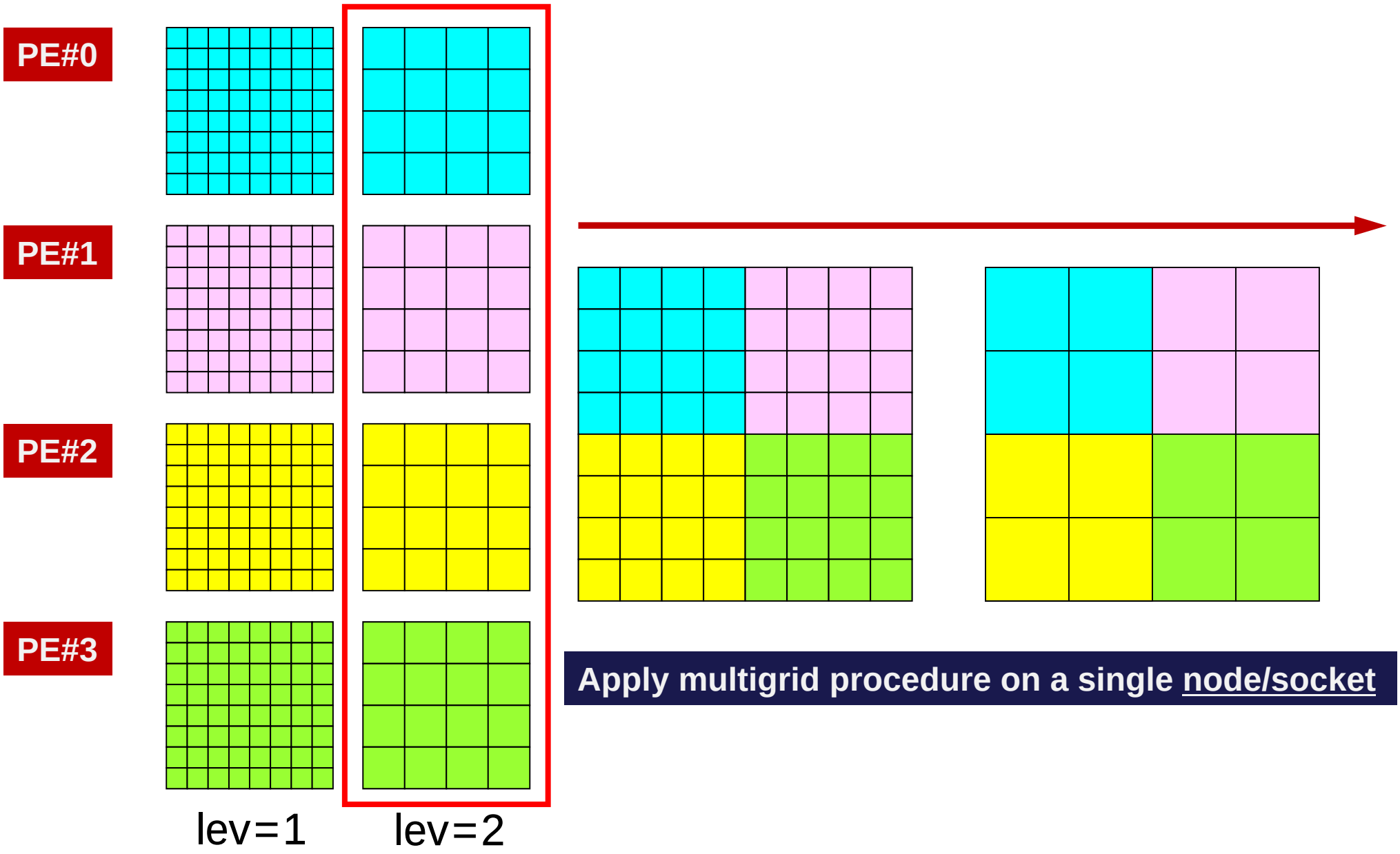


Size of the Coarsest Grid= Number of MPI Processes Redundant Process

Coarse Grid Aggregation: at lev=2



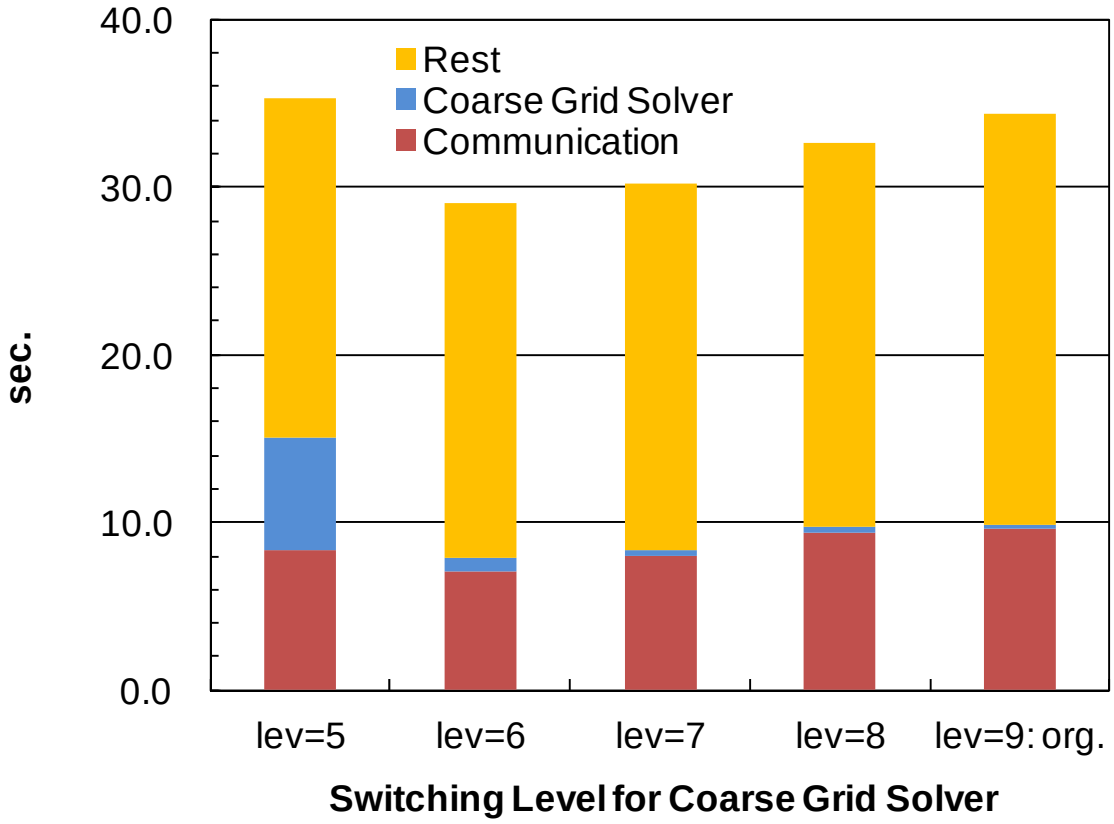
Coarse Grid Aggregation: at lev=2

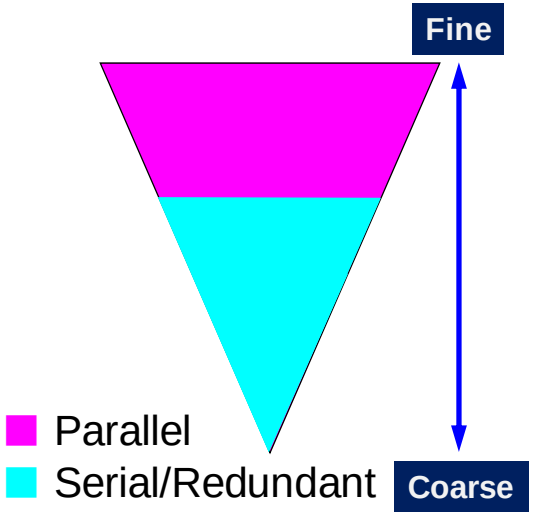


Apply multigrid procedure on a single node/socket

Results: T2K, HB 16x1 512 nodes, 8,192 cores

lev: switching level to “coarse grid solver”

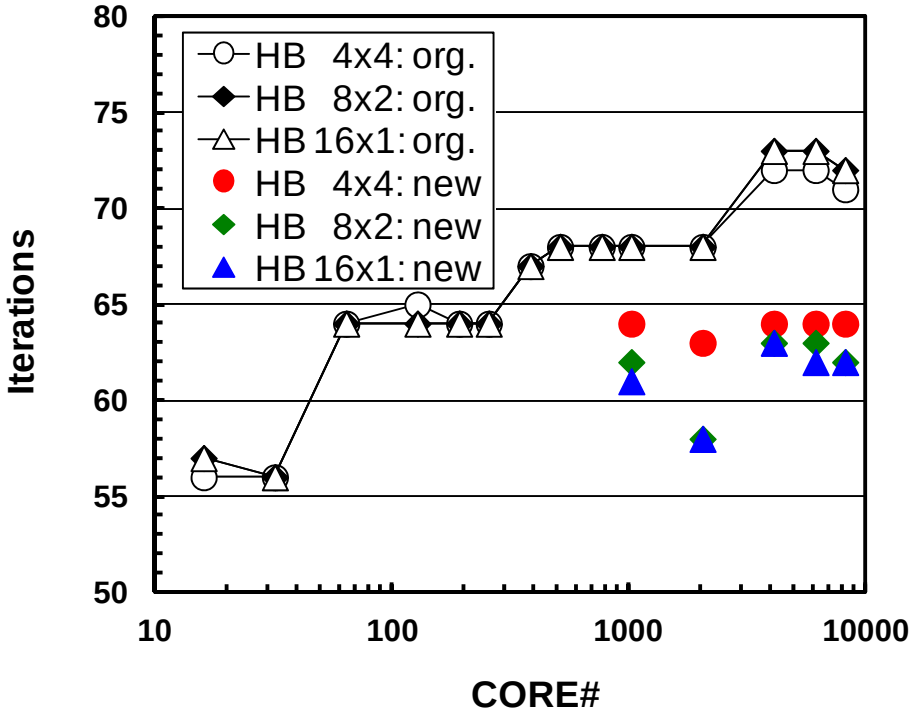


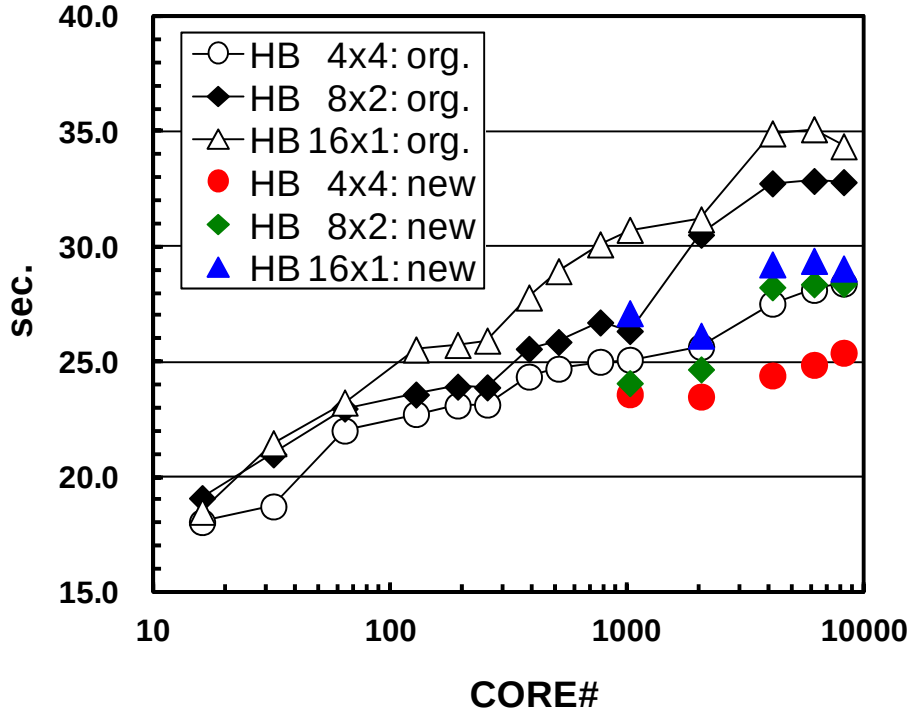


通信の割合を
減らすのが
目的だったが，
あまり減っておらず

Weak Scaling: T2K, 512 nodes, 8,192 cores

Optimum Cases

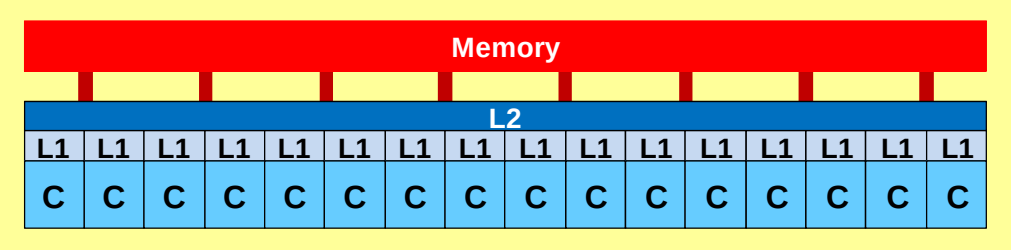




反復回数の減少，安定化に効果があった。

Larger Cases on Fujitsu FX10

- Fujitsu PRIMEHPC FX10 at U.Tokyo (Oakleaf-FX)
  - 16 cores/node, flat/uniform access to memory
- Up to 4,096 nodes (65,536 cores) (Large-Scale HPC Challenge)
  - Max 17,179,869,184 unknowns
  - Flat MPI, HB 4x4, HB 8x2, HB 16x1
    - HB MxN: M-threads x N-MPI-processes on each node
- Weak Scaling
  - 64^3 cells/core
- Strong Scaling
  - $128^3 \times 8 = 16{,}777{,}216$ unknowns, from 8 to 4,096 nodes
- Network Topology is not specified
  - 1D

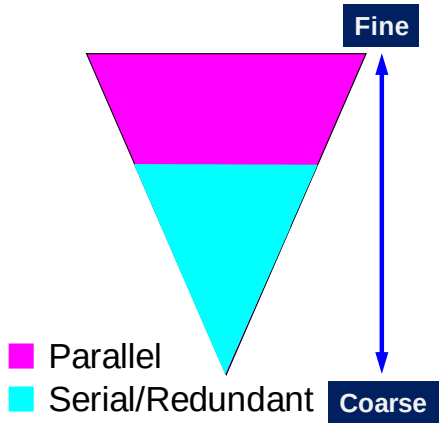


Results at 4,096 nodes

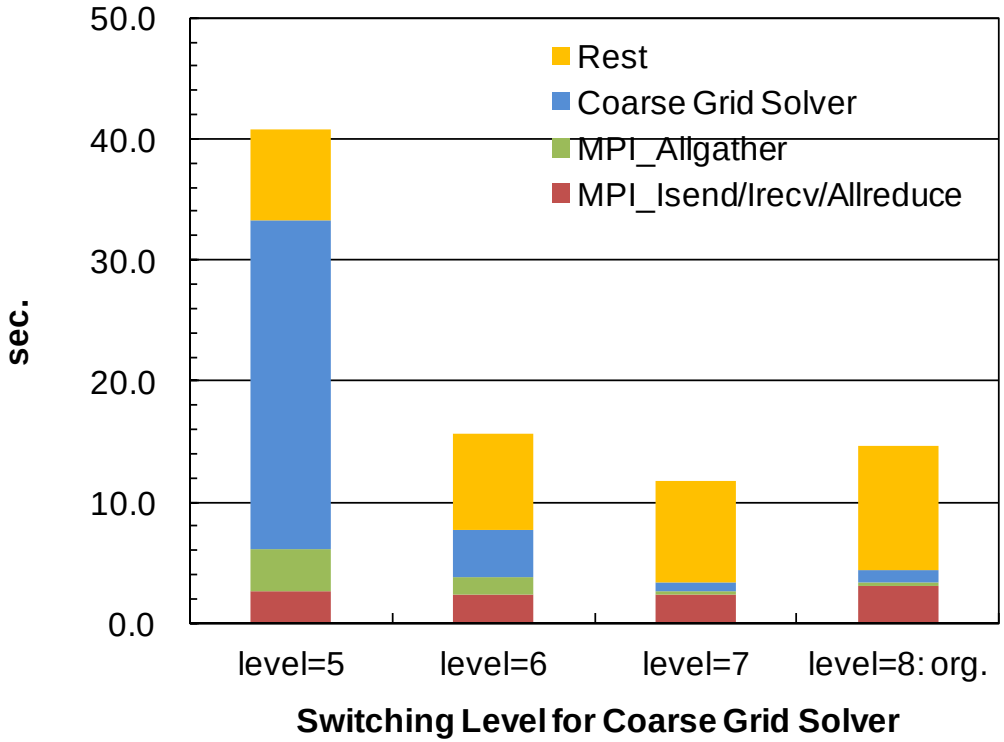
lev: switching level to “coarse grid solver”

Opt. Level= 7

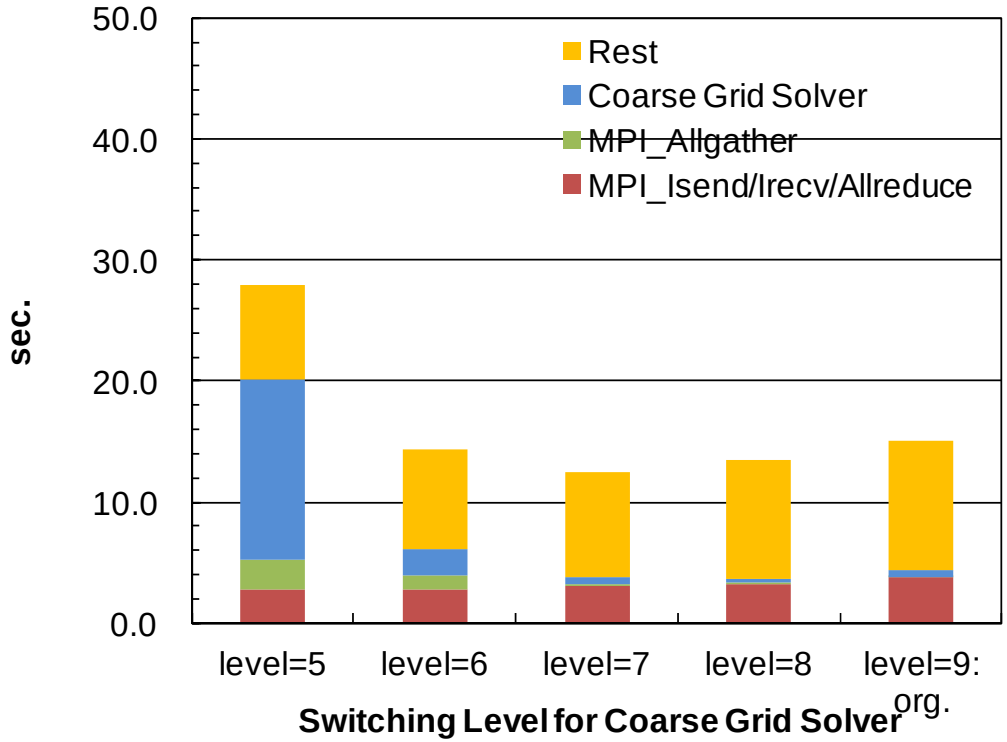
DOWN is GOOD



HB 8x2

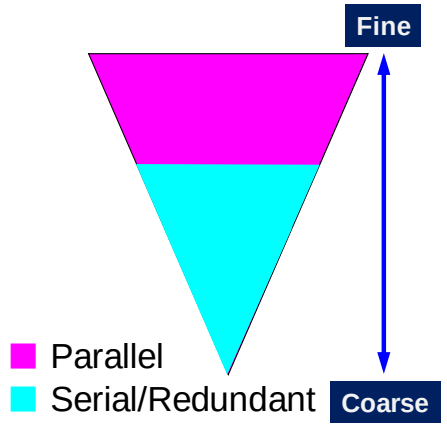


HB 16x1



Improvement of Performance

by “Coarse Grid Aggregation”



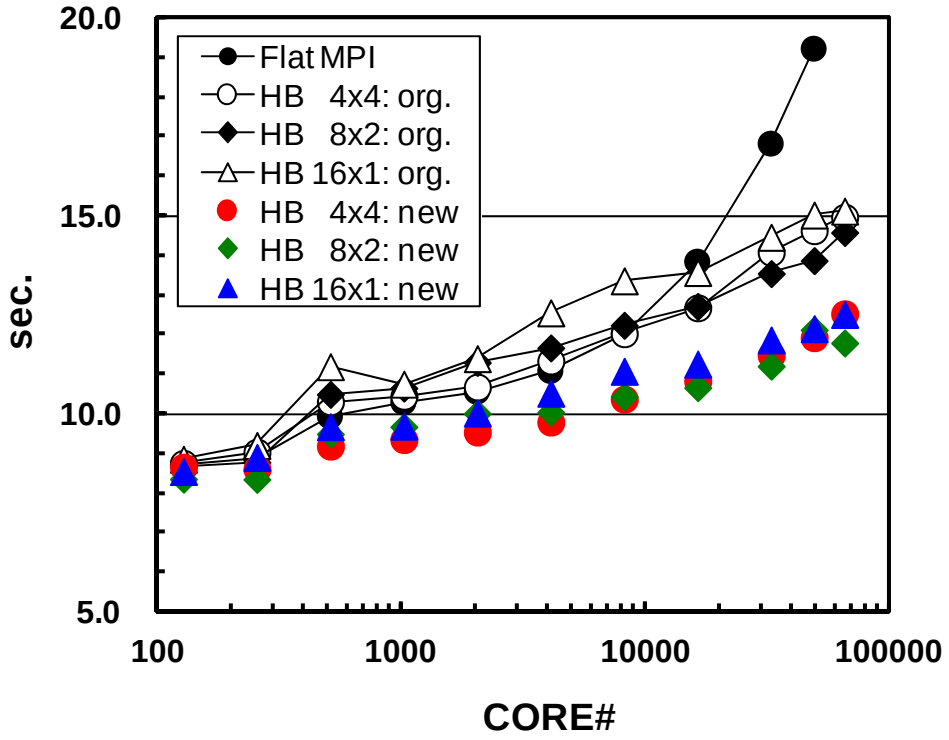
	HB 4x4	HB 8x2	HB 16x1
T2K 512 nodes	11.8 %	15.5 %	18.2 %
FX10 512 nodes	15.9 %	17.2 %	20.6 %
FX10 4,096 nodes	19.4 %	23.7 %	20.9 %

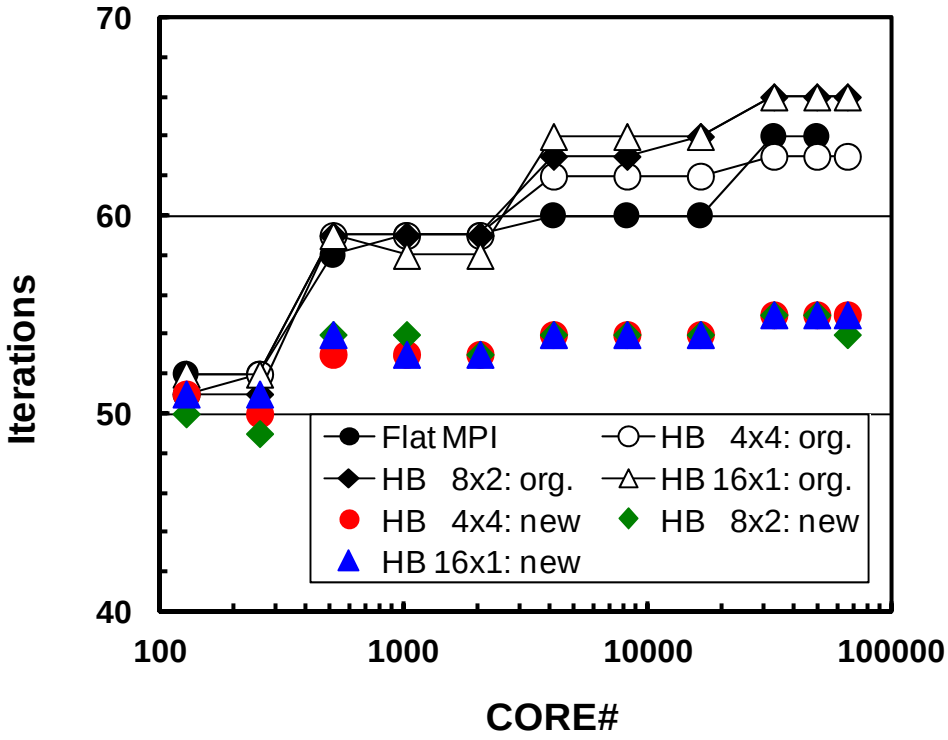
Weak Scaling: up to 4,096 nodes

Flat MPI was terrible for > 3,072 nodes

Switching Level: optimum value at 4,096 nodes

DOWN is GOOD





まとめ

- 「Coarse Grid Aggregation」はO（10^4）コアで収束安定化に効果があった
  - 通信はそれほど減少せず
  - 4,096ノード@FX10ではHB 8×2がベスト
    - MPIプロセス数が少ない方が大規模問題では有利
    - Coarse Grid Solverにおける問題規模が小さくなる
- 更なる最適化
  - FX10
  - 計算と通信のオーバーラップ
  - 最適レベル *lev* の自動決定
  - プロセス数の段階的減少 (e.g. 8192-512-32-1)