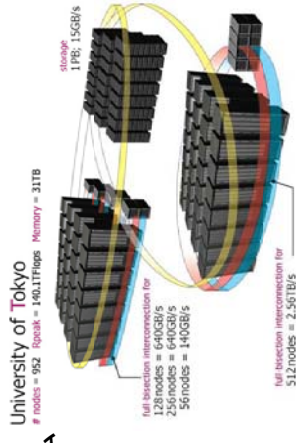


並列Multigrid on T2K(東大)

- T2Kオープンスーパーコンピュータ(東大)
- 並列多重格子法(Multigrid)前処理付きCG法
 - MGCG
- Flat MPI vs. Hybrid (OpenMP+MPI)
 - Hybrid
 - MPIのプロセス数を減らせる→通信オーバーヘッド減少
 - メモリ的には厳しくなる: 特に疎行列ソルバー

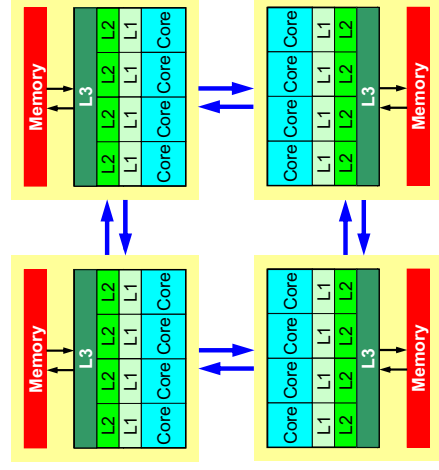
T2K(東大) (1/2)

- T2Kオープンスーパーコンピュータ仕様
 - <http://www.open-supercomputer.org/>
 - 筑波大, 東大, 京大
- T2Kオープンスーパーコンピュータ(東大)
 - Hitachi HA8000クラスシステム
 - 2008年6月～
 - 952ノード (15,232コア), 141 TFLOPS peak
 - Quad-core Opteron (Barcelona)
 - TOP500 53位 (Jun 2010)



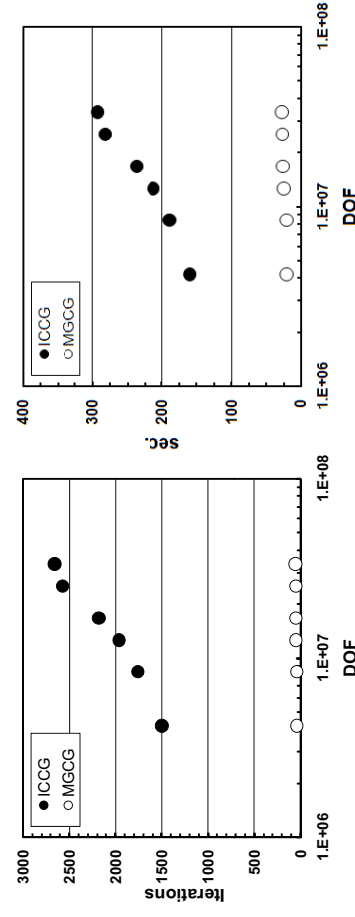
T2K(東大) (2/2)

- AMD Quad-core Opteron (Barcelona) 2.3GHz
 - 4 “sockets” per node
 - 16 cores/node
- マルチコア, マルチソケット
- cc-NUMA (cache coherent Non-Uniform Memory Access)
 - ローカルメモリ上のデータで
きるだけ使用する
 - 陽的なコマンドラインスイッチ
 - NUMA control



Multigrid is scalable

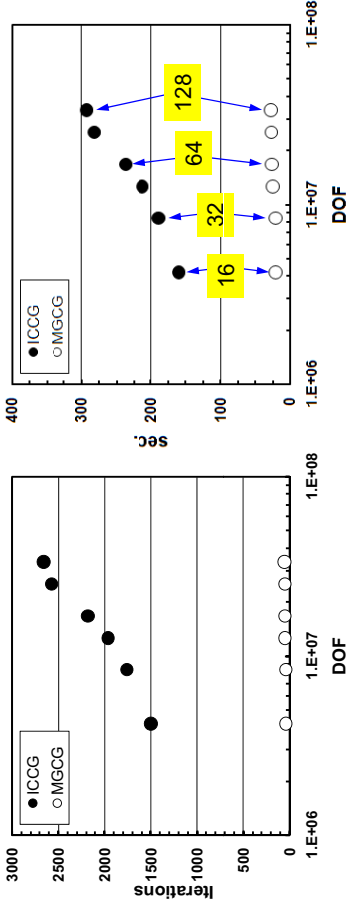
Weak Scaling: Problem Size/Core Fixed 三次元ポアンソン方程式(一様)



Multigrid is scalable

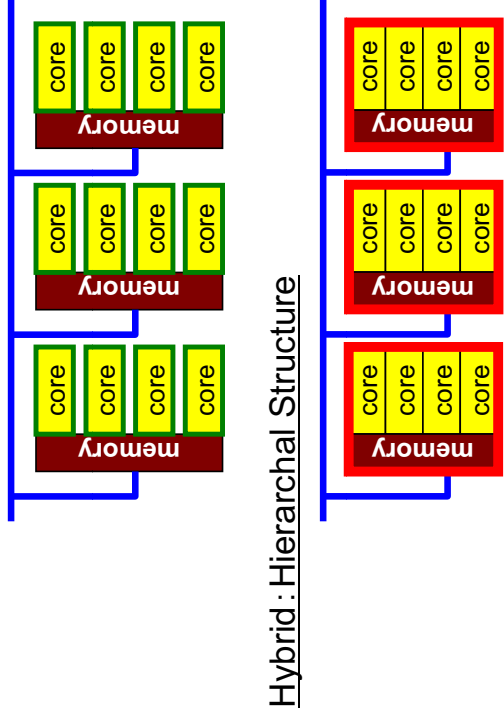
Weak Scaling: Problem Size/Core Fixed

MGCG法の計算時間は「Weak Scaling」では一定=Scalable



Flat MPI vs. Hybrid

Flat-MPI: Each PE -> Independent

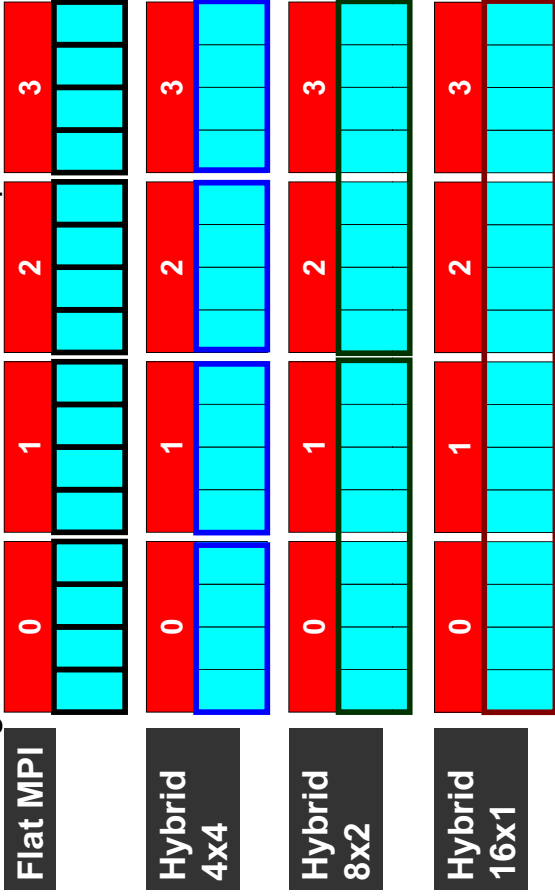


Flat MPI vs. Hybrid

- 性能は様々なパラメータの組み合わせによって決まる
- ハードウェア
 - コア, CPUのアーキテクチャ
 - ピーク性能
 - メモリ性能(バンド幅, レイテンシ)
 - 通信性能(バンド幅, レイテンシ)
 - それらのバランス
- アプリケーション
 - 特性: memory bound, communication bound
 - 問題サイズ

Flat MPI, Hybrid (4x4, 8x2, 16x1)

Higher Performance of HB16x1 is important

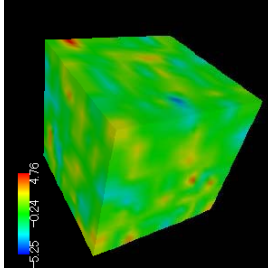


解析対象

- 透水係数が空間的に分布する三次元地下水流れ
 - ポアソン方程式
 - 透水係数は地質統計学的手法によって決定 [Deutsch & Journel, 1998]
- 規則正しい立方体ボクセルメッシュを使用した有限体積法
 - 局所細分化を考慮
- 周期的な不均質性: 128³

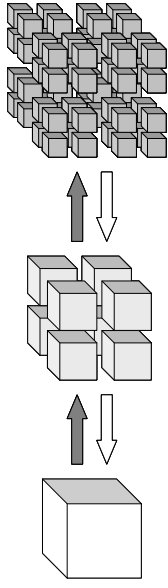
$$\frac{\partial}{\partial x} \left(\lambda \frac{\partial \phi}{\partial x} \right) + \frac{\partial}{\partial y} \left(\lambda \frac{\partial \phi}{\partial y} \right) + \frac{\partial}{\partial z} \left(\lambda \frac{\partial \phi}{\partial z} \right) = q$$

$$\phi = 0 @ x = x_{\max}$$

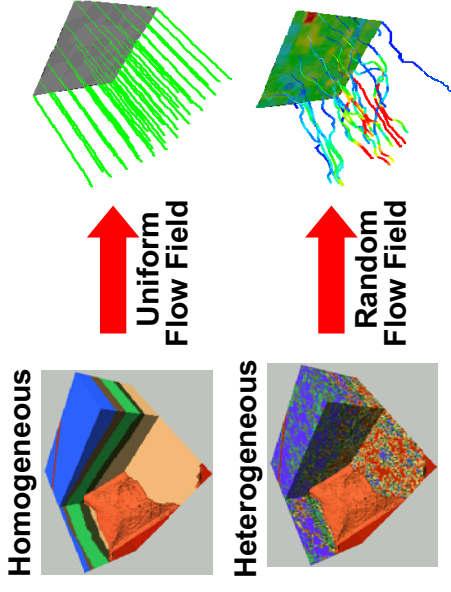


線形ソルバーの概要

- 前処理付きCG法
 - Multigrid 前処理
 - IC(0) for Smoothing Operator (Smoother)
 - Additive Schwarz Domain Decomposition
- 並列 (幾何学的) 多重格子法
 - 当方的な8分木
 - V-cycle
 - 領域分割型: Block-Jacobi局所前処理, 階層型領域間通信
 - 最も粗い格子 (格子数=プロセッサ数) は1コアで実施



Groundwater Flow through Heterogeneous Porous Media



Hardware/Software

- T2K/Tokyo
 - up to 512 nodes (8,192 cores)
- Program
 - Hitachi FORTRAN90 + MPI
 - CRS matrix storage
 - CM-RCM Reordering for OpenMP**
- $|Ax-b|/|b|=10^{-12}$ for Convergence
- 不均質性
 - 最大最小透水係数の比 = 10^{10} ($10^{-5} \sim 10^{+5}$)
- Multigrid Cycles
 - 1 V-cycle/iteration
 - 2 smoothing iterations for restriction/prolongation at every level
 - 1 ASDD iteration cycle for each restriction/prolongation

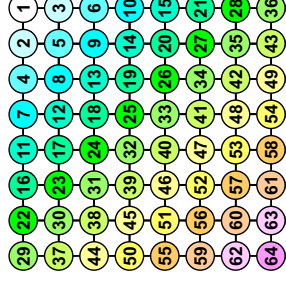
```
for (i=0; i<N; i++) {
  for (k=Index(i-1); k<Index(i); k++) {
    Y[i]= Y[i] + A [k]*X[Index[k]];
  }
}
```

前処理付き反復法のSMP/MulticoreでのOpenMPによる並列化

- DAXPY, SMVP, Dot Products
 - 簡単
- 前処理: ILU系分解, 前進後退代入
 - 大域的な依存性 (Global dependency)
 - 並び替え (Reordering) による並列性の抽出
 - Multicolor Ordering (MC), Reverse-Cuthill-McKee (RCM)
 - 同じ色内の要素は独立⇒並列化可能
 - 「地球シミュレータ」向け最適化 [KN 2002,2003]
 - 並列及びベクトル性能
 - 並列性高く安定なCM-RCMを採用

Ordering Methods

Elements in “same color” are independent: to be parallelized



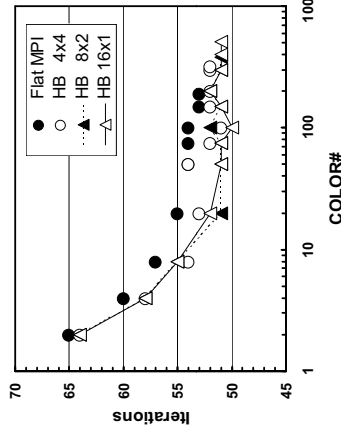
Effect of Number of Colors

色数の効果 (CM-RCM)

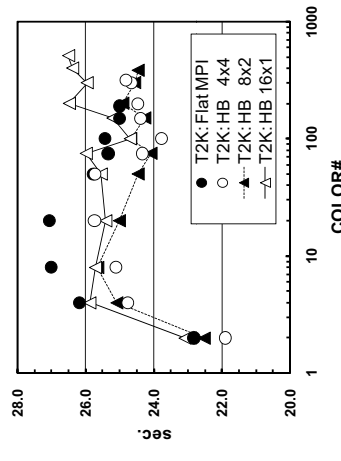
16,777,216 = 64x64³ cells, 64 cores

色数が増えると収束は改善, 計算時間はCM-RCM(2)が最も短い

Iterations



sec.



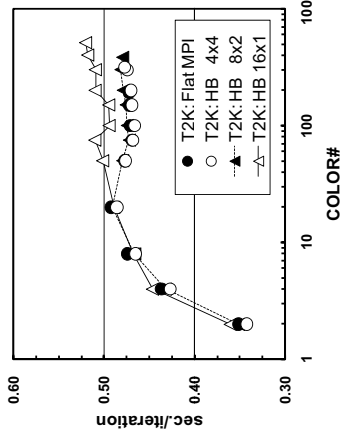
色数の効果 (CM-RCM)

16,777,216 = 64×64^3 cells, 64 cores

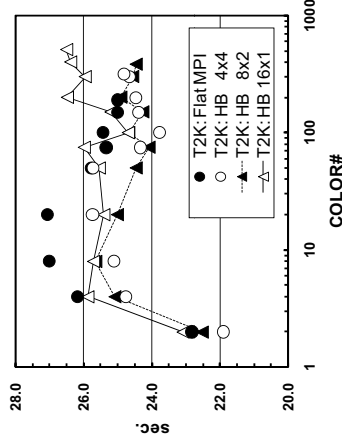
色数が増えると収束は改善, 計算時間は

CM-RCM(2)が最も短い: 反復あたり計算時間短い

sec./iter



sec.



色数の効果 (CM-RCM)

RCM: 前進後退代入時に変数値が変わるため, キャッシュラインからメモリに戻されてしまう可能性がある

29	22	16	11	7	4	2	1
37	30	23	17	12	8	5	3
44	38	31	24	18	13	9	6
50	45	39	32	25	19	14	10
55	51	46	40	33	26	20	15
59	56	52	47	41	34	27	21
62	60	57	53	48	42	35	28
64	63	61	58	54	49	43	36

RCM

45	10	39	5	35	2	33	1
17	46	11	40	6	36	3	34
53	18	47	12	41	7	37	4
24	54	19	48	13	42	8	38
59	25	55	20	49	14	43	9
29	60	26	56	21	50	15	44
63	30	61	27	57	22	51	16
32	64	31	62	28	58	23	52

CM-RCM(2)

61	29	62	30	63	31	64	32
25	57	26	58	27	59	28	60
53	21	54	22	55	23	56	24
17	49	18	50	19	51	20	52
45	13	46	14	47	15	48	16
9	41	10	42	11	43	12	44
37	5	38	6	39	7	40	8
1	33	2	34	3	35	4	36

MC(2)

Weak Scaling

- コアあたりの問題規模を固定
 - Scalableであれば, コア数を増やして全体の問題規模が大きくなっても反復回数が変わらないため, 計算時間は変わらないはず (実際はそううまく行かない)。
 - 全体の問題規模を固定するのが Strong Scaling

- Up to 8,192 cores (512 nodes)

- 64³ cells/core

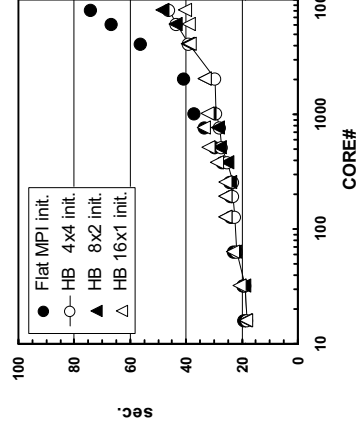
- 2,147,483,648 cells

- CM-RCM(2)

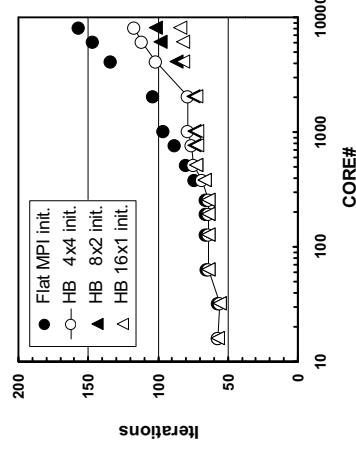
Weak Scaling

64³ cells/core, up to 8,192 cores (2.05×10^9 cells)

sec.

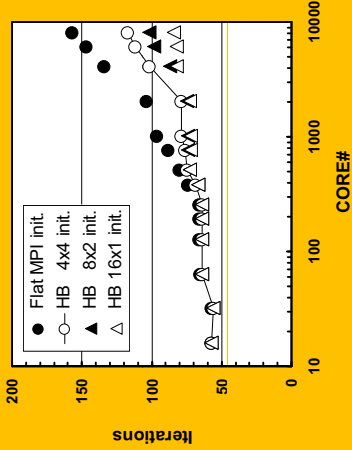


Iterations



Coarse Grid Solverの改良

- 領域数が増えたと反復回数が増加(特にFlat MPI)
- 最も粗い格子 (Coarse Grid Solver)
 - 各領域1メッシュになった状態で1コアに集める
 - IC(0)スムージングを一回施す
- Coarse Grid Solver改良
 - IC(0)スムージングを収束 ($\epsilon=10^{-12}$) まで繰り返す : C1
 - マルチグリッド(V-cycle)を用し, 収束 ($\epsilon=10^{-12}$) まで繰り返す (8,192=32 × 16 × 16) : C2

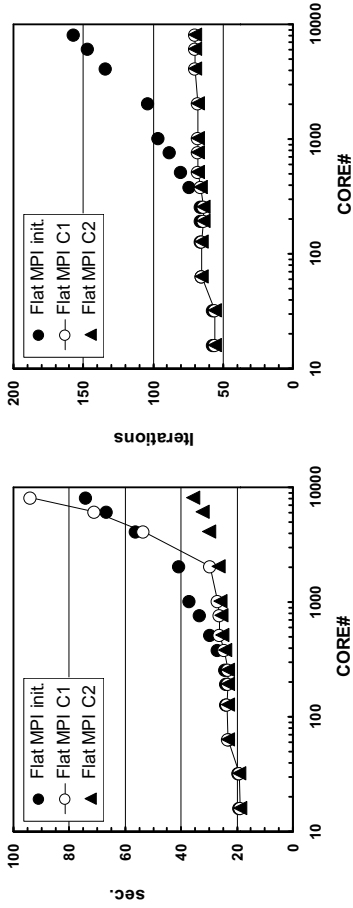


Weak Scaling: Flat MPI

64³cells/core, up to 8,192 cores (2.05 × 10⁹ cells)

sec.

Iterations

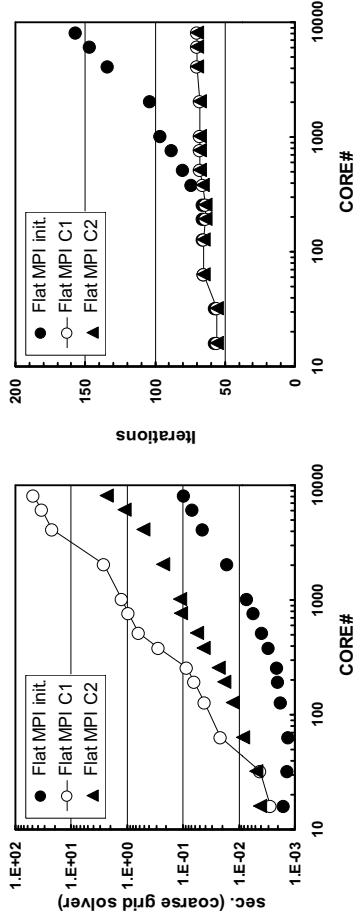


Weak Scaling: Flat MPI

64³cells/core, up to 8,192 cores (2.05 × 10⁹ cells)

Coarse Grid Solver

Iterations

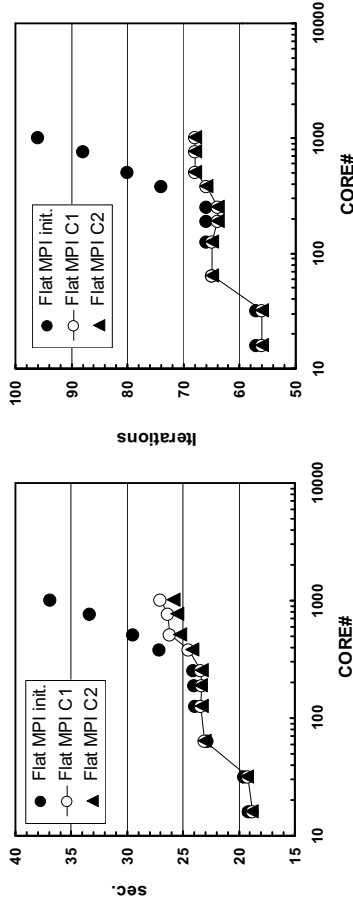


Weak Scaling: Flat MPI

64³cells/core, up to 8,192 cores (2.05 × 10⁹ cells)

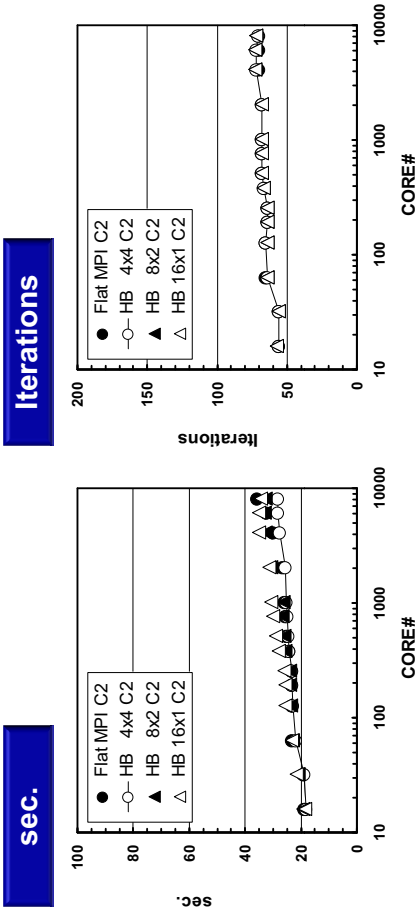
sec.

Iterations

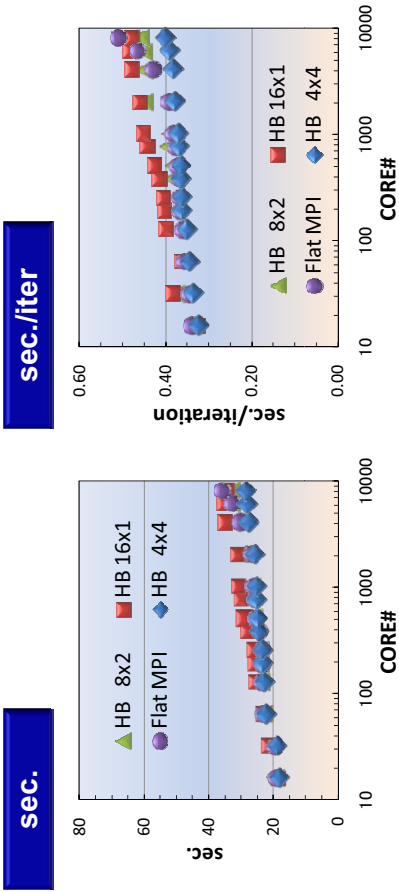


Weak Scaling

64³cells/core, up to 8,192 cores (2.05 × 10⁹ cells)
at 8,192 cores: Flat MPI(35.7sec), HB 4x4(28.4), 8x2(32.8), 16x1(34.4)

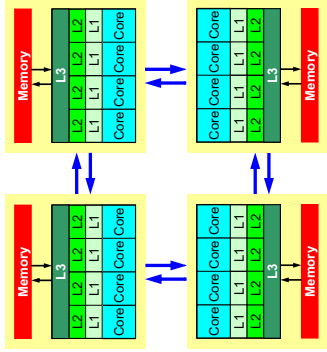


でもよく見るとScalableではない 通信のオーバーヘッド等



まとめ

- (多重格子法 (MG) 前処理 + CG 法)
 - 不均質多孔質媒体中の三次元地下水流れ, 有限体積法
 - IC(0) smoother + ASDD, 幾何学的MG
- OpenMP/MPI Hybrid 並列プログラミングモデル on T2K (東大)



- NUMA Policy
- First Touch Data Placement + Sequential Reordering
- Coarse Grid Solver改良
 - HB 4x4 (a single MPI process per socket) が最も効率がよい: メモリを最も効率よく使っている. 通信オーバーヘッドも少ない
 - 反復回数は並列プログラミングモデルによってほとんど変化しない

今後の課題

- 粗い格子レベルにおけるコア数の漸減
 - 全体のコア数, 領域数が増えると通信オーバーヘッドが増加
- Hybridにおける領域内並べ替え
 - CM-RCM
 - HID
 - 並列化: 結構時間がかかる
- Communication Reducing Algorithms
 - 並列MG: とにかく通信多い