SS研科学技術計算分科会

科学技術計算分科会 2003年度 第2回会合

３次元MHDコードによるPRIMEPOWER HPC2500の性能評価：名古屋大学における事例報告

名古屋大学太陽地球環境研究所

荻野竜樹

名古屋大学情報連携基盤センター

津田知子

アウトライン

１．背景

２．地球磁気圏の３次元グローバルMHDモデル

３．VPPにおける並列化

４．PRIMEPOWER HPC2500における並列化

５．HPC2500とVPPにおける３次元MHDコード計算速度

６．計算速度の比較の考察と問題点

７．地球磁気圏の３次元グローバルMHDシミュレーションの将来

８．まとめ

1．背景

・太陽活動とジオスペース（地球外圏）環境変化の因果関係を解明する研究

・外部境界をできるだけ遠くに設定してかつ空間分解能をできるだけ上げる

・計算方法の改良と最大級のスーパーコンピュータの効率的な利用

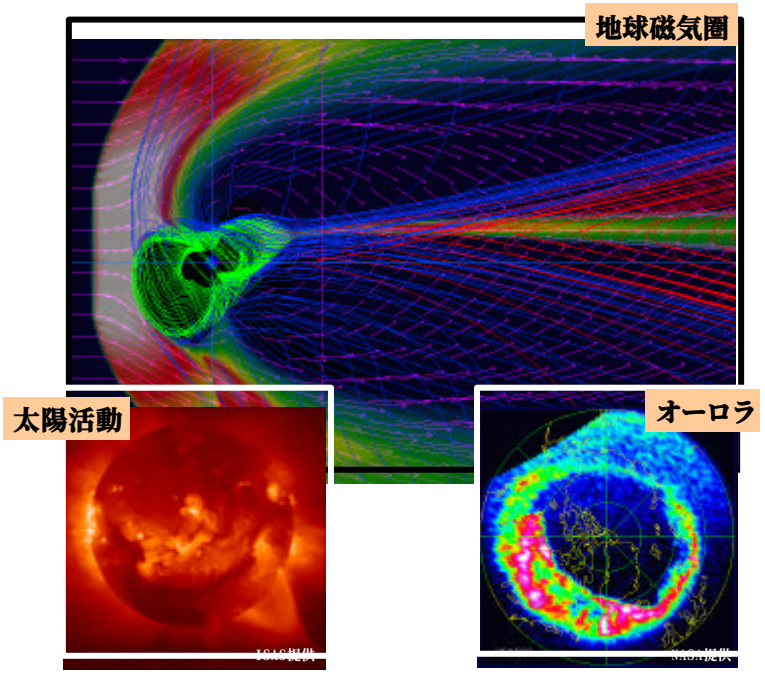
・並列計算共通プログラム言語／手法

High Performance Fortran (HPF)

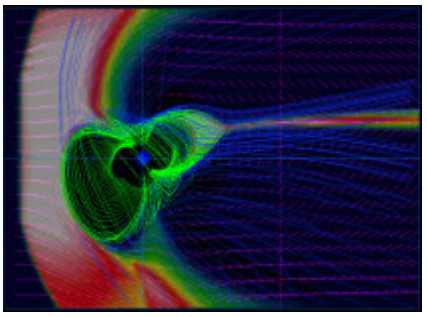
Message Passing Interface (MPI)

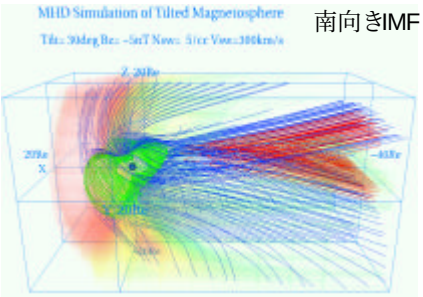
・ベクトル並列機からスカラ並列機への移行

富士通株式会社：VPP5000からPRIMEPOWER HPC2500へ

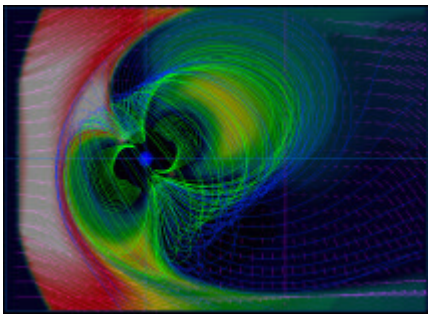


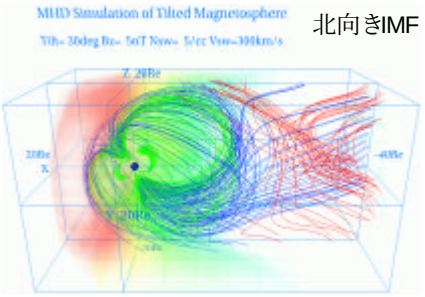
MHD Simulation of the Earth's Magnetosphere with Dipole Tilt for IMF Bz Component



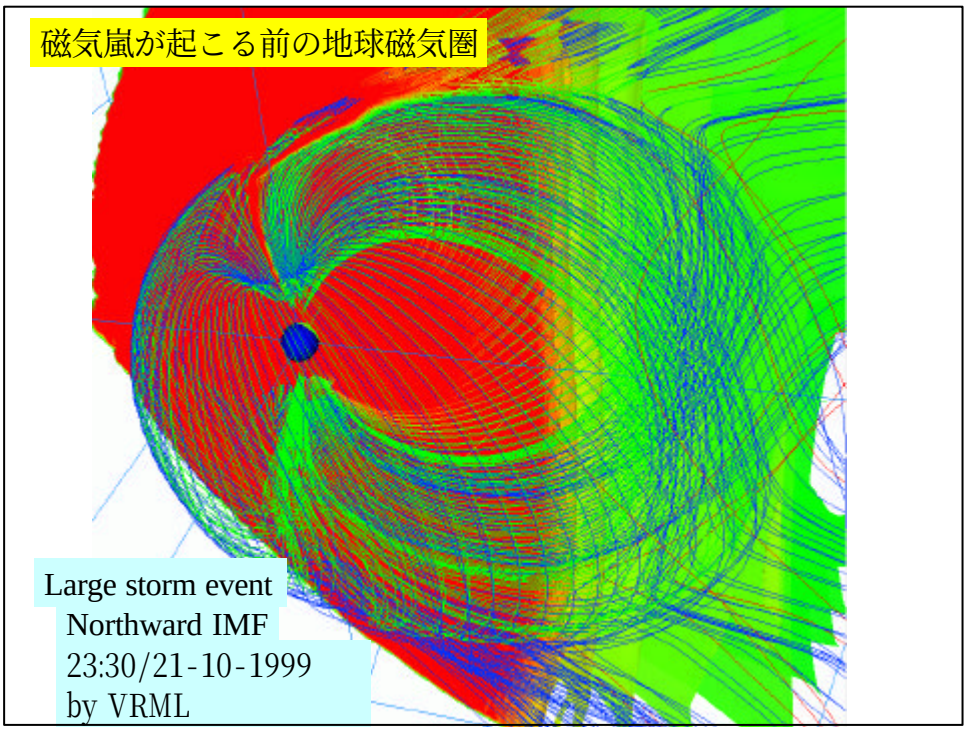


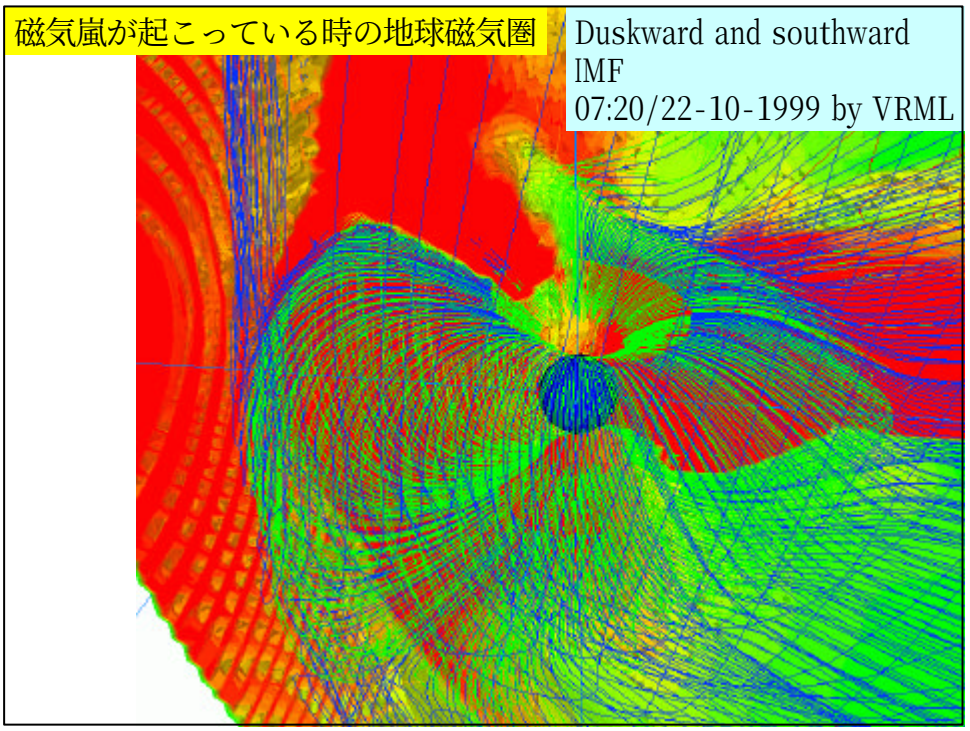
南向きIMF





北向きIMF





2．地球磁気圏の３次元グローバルMHDモデル

・太陽風磁気圏相互作用の３次元グローバルMHDモデル

・MHD方程式とマックスウェル方程式をmodified leap-frog 法を用いて、初期値境界値問題として時間発展を解く

・並列計算では、z方向に１次元の領域分割

・z方向のdo loopは最大限に大きくとり、その出現回数を最少化

・z方向のデータ転送は必要最小限の２カ所にコンパクトにまとめて、一括大量転送を実施

基礎方程式

・MHDモデルの基礎となる規格されたMHD方程式とMaxwell方程式

$$\frac{\partial r}{\partial t} = -\nabla \cdot (\mathbf{v} r) + D \nabla^2 r \qquad (1)$$

$$\frac{\partial \mathbf{v}}{\partial t} = -(\mathbf{v} \cdot \nabla)\mathbf{v} - \frac{1}{r}\nabla p + \frac{1}{r}\mathbf{J} \times \mathbf{B} + \mathbf{g} + \frac{1}{r}\Phi \qquad (2)$$

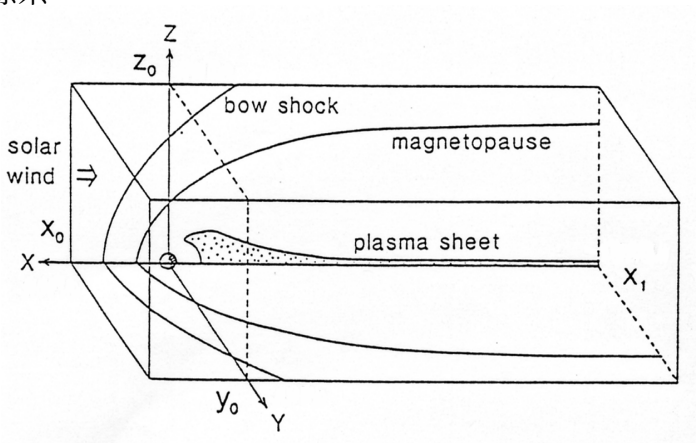
$$\frac{\partial p}{\partial t} = -(\mathbf{v} \cdot \nabla)p - g\ p \nabla \cdot \mathbf{v} + D_p \nabla^2 p \qquad (3)$$

$$\frac{\partial \mathrm{B}}{\partial t} = \nabla \times (\mathbf{v} \times \mathbf{B}) + h\, \nabla^2 \mathbf{B} \qquad (4)$$

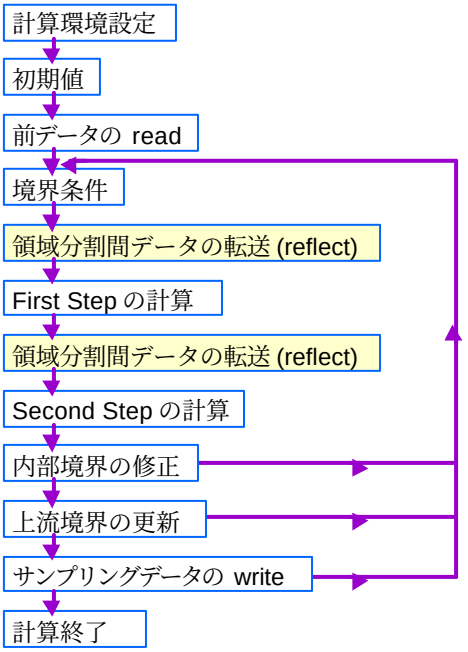
$$\mathbf{J} = \nabla \times (\mathbf{B} - \mathbf{B}_d) \qquad (5)$$

座標系と境界条件

3次元MHDシミュレーションに用いる太陽地球磁気圏座標系



3. MHDコードにおけるHPF/JA



3．VPPにおける並列化

- VP200やCRAY-1などでフルベクトル化されていた３次元MHDコード
- VPP Fortranを用いて、VPP500用にフルベクトル化フル並列化された３次元MHDコードを作成 (1997)
- HPF/JA を用いて、VPP5000用にフルベクトル化フル並列化された３次元MHDコードを作成 (2000)
- VPP FortranとHPF/JAからMPIへ書き換えて、フルベクトル化フル並列化された３次元MHDコードを作成 (2002)
- ３次元MHDコードの基本構造は、MPIとHPF/JAとVPP Fortranで同じ

４．PRIMEPOWER HPC2500における並列化

・VPP Fortranで使用したMPIの３次元MHDコードと基本的に同じ

・配列の順序は、f(nx2,ny2,nz2,nb)のように、x方向、y方向、z方向、8個のMHD方程式の成分の順序で、z方向に１次元の領域分割

・高速化のためにf(nx2,ny2,nz2,nb)からF(nb,nx2,ny2,nz2)への配列の順序の並び替えを宣言

・領域分割の両端のデータをcpu間で通信には、wild cardとmpi_sendrecvの利用及び通信データのコンパクト化を実行

・コンパイルオプションでは、キャッシュのレベルの最適化、largepageの利用、64bit対応用のオプションV9のオフを採用

５．HPC2500とVPPにおける３次元MHDコード計算速度

・VPP5000/64では、VPP FortranとHPF/JAとMPIのいずれでも同等の計算速度がでていて、スケーラビリティもかなり良い

・PRIMEPOWER HPC2500では、32cpuまではかなりよいスケーラビィリティが得られた

・HPC2500 (ngrd)の8cpuで8.36 Gflopsの計算速度を得たように、VPP5000の　1cpuに対してHPC2500 の7cpuで同等の速度を得ることができた。

・33cpu以上で計算速度が急に遅くなったが、HPC2500（ngrd）において33cpu以上では高速ネットワークが働かなくなるからである

・領域分割間の通信停止の場合の計算速度を測定すると48cpuまで確実によいスケーラビィリティが得られている

<< 名古屋大学情報連携基盤センターのPRIMEPOWER HPC2500（grid）の性能 >>

【ＣＰＵ】

32CPU構成×2筐体＝計64CPU

（このうち、Batchには24CPU×2を使用、あとは、TSSとGrid用）

・SPARC64 V(1.3GHz)

・内部キャッシュ(1次キャッシュ:命令128KB+データ128KB、2次キャッシュ:2MB)

・理論計算速度：5.2 GFlops/CPU, 332.8 GFlops/TOTAL

【メモリ】

64GB×2筐体＝計128GB

【高速インターコネクト】

ノード間通信性能：4GB/s×2＝8GB/s（理論性能値）

(A) 加減乗除の四則演算の実測（Mflops）

1. F(I)=FLOAT(I)
2. P(I)=F(I)
3. U(I)=F(I)+P(I)
4. U(I)=F(I)-P(I)
5. U(I)=F(I)*P(I)
6. U(I)=F(I)/P(I)

Computer	1	2	3	4	5	6
PRIMEPOWER HPC2500 (Mflops)	190	465	366	353	366	94
VPP5000 (scalar)	42	291	210	212	217	61
VPP5000 (vector)	2327	8533	3200	3200	3200	2844
HPC P4Linux	384	889	628	630	610	51

(B) 四則演算の速度の比較

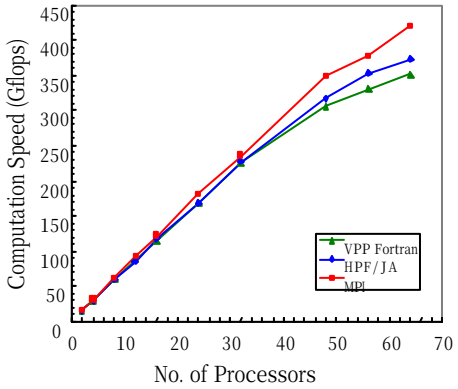
computer	compiler option	processing capability (MFLOPS)	
Fujitsu VPP-500 (1PE)	frtpx, -sc	19.8	
Fujitsu VPP-500 (1PE)	frtpx	730.3	
Fujitsu VPP-5000 (1PE)	frt, -sc	189.8	(1999.12.27)
Fujitsu VPP-5000 (1PE)	frt	3,073.7	(1999.12.27)
SunFireV800 (sv080)	f90 -0	138.1	(2002.08.28)
GP7000Fモデル900(gpcs)	frt -03	95.8	(2003.04.08)
HPC2500 (ngrd)	frt -03	273.3	(2003.04.08)
DEC Alpha DS-20(833MHz)	f90 -fast	246.9	(2002.11.08)
SONY VAIO Pentium4	lf95	547.6	(2002.09.03)
HPC P4Linux	ifc	478.7	(2003.02.06)

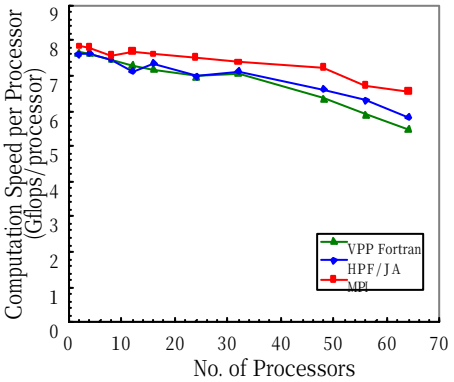
(C) MHDコードを用いての計算速度の比較

computer	compiler	(a)3D-MHD 50x50x26 sec	(MFLOPS)	(b)3D-MHD 62x32x32 sec	(MFLOPS)	(c)2D-MHD 402x102 sec	(MFLOPS)
Fujitsu VPP-5000(1PE)	frt, -sc	0.717	(169)	0.694	(168)	0.1238	(225)
Fujitsu VPP-5000(1PE)	frt	0.0233	(5201)	0.0227	(5110)	0.0044	(6316)
SunFireV800 (sv080)	f90 -0	0.305	(397)	0.336	(348)	0.0781	(355)
GP7000Fモデル900(gpcs)	frt -03	0.2515	(481)	0.2905	(403)	0.0562	(493)
HPC2500 (ngrd)	frt -03	0.0938	(1291)	0.0938	(1246)	0.0234	(1184)
DEC Alpha (833MHz)	f90 -fast	0.1016	(1186)	0.0859	(1355)	0.0234	(1155)
SONY VAIO Pentium4	lf95	0.203	(596)	0.312	(375)	0.0430	(645)
HPC P4Linux	ifc	0.0906	(1335)	0.0943	(1241)	0.0240	(1172)

Computer Processing Capability in the MPI, HPF/JA and VPP Fortran by Using VPP

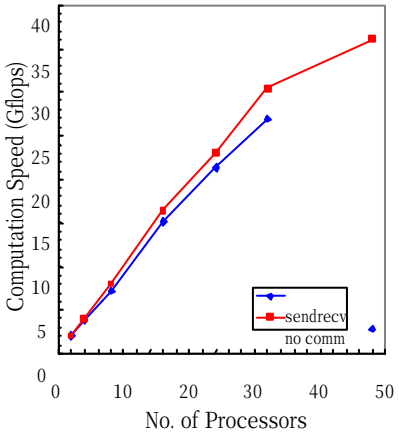
(nx,ny,nz)=(800,200,670)

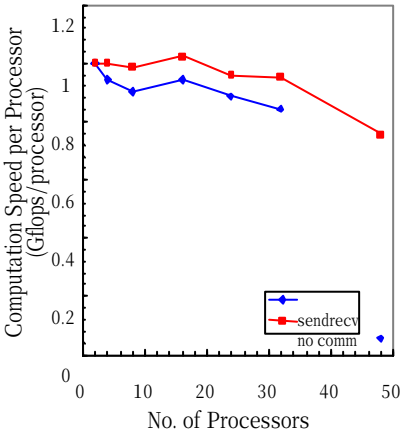




Computer Processing Capability in the MPI by Using HPC2500

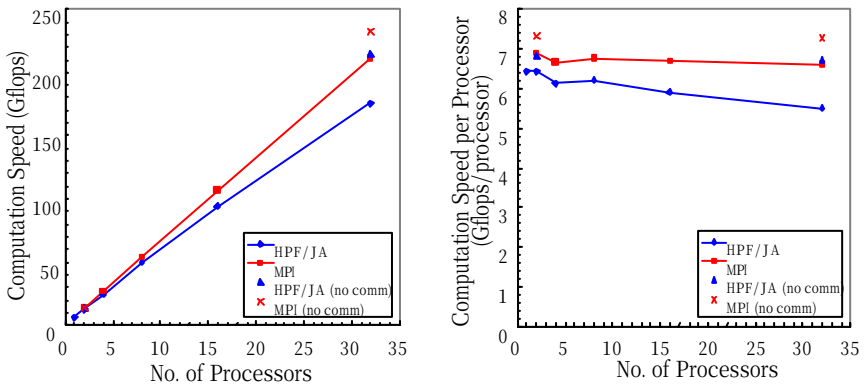
(nx,ny,nz)=(250,158,158)





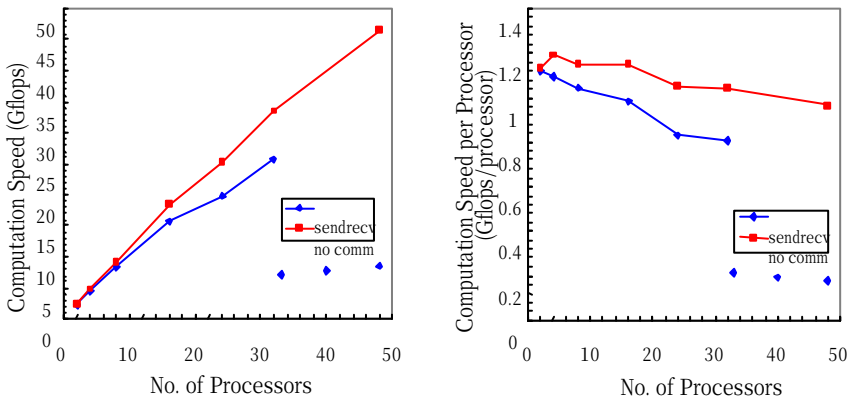
Computer Processing Capability in the MPI and HPF/JA by Using VPP

(nx,ny,nz)=(500,318,318)



Computer Processing Capability in the MPI by Using HPC2500

(nx,ny,nz)=(500,318,318)



６．計算速度の比較の考察と問題点

・PRIMEPOWER HPC2500で、33cpu以上では並列計算の顕著な劣化がみられたが、33cpu以上ではデータ転送に高速光インターコネクト（4GB/sec）を用いずに、GbEを使用しているため(-ldt 0)。

・24cpu使用時のコンパイルオプションの違いによるハードバリアとソフトバリア利用の計算速度は、それぞれ21.82 Gflopsと20.27 Gflopsであり、その差は約7.1%。

・ベクトル並列機VPP5000からスカラ並列機PRIMEPOWER HPC2500への移行で、高効率の並列計算が保証されるのか？

・ベクトル並列機では最内のdoloopでベクトル化して、最外のdo loopで並列化するのが基本。計算効率を上げるためには最内のdo loopは長い方が望ましい。

・スカラ並列機はキャッシュのヒットが高速計算に不可欠。１次キャッシュと２次キャッシュのヒット、largepageの機能で高速計算を保証。従って、計算に関与する変数の番地はあまり離れていないことが望ましい。

７．地球磁気圏の３次元グローバルMHDシミュレーションの将来

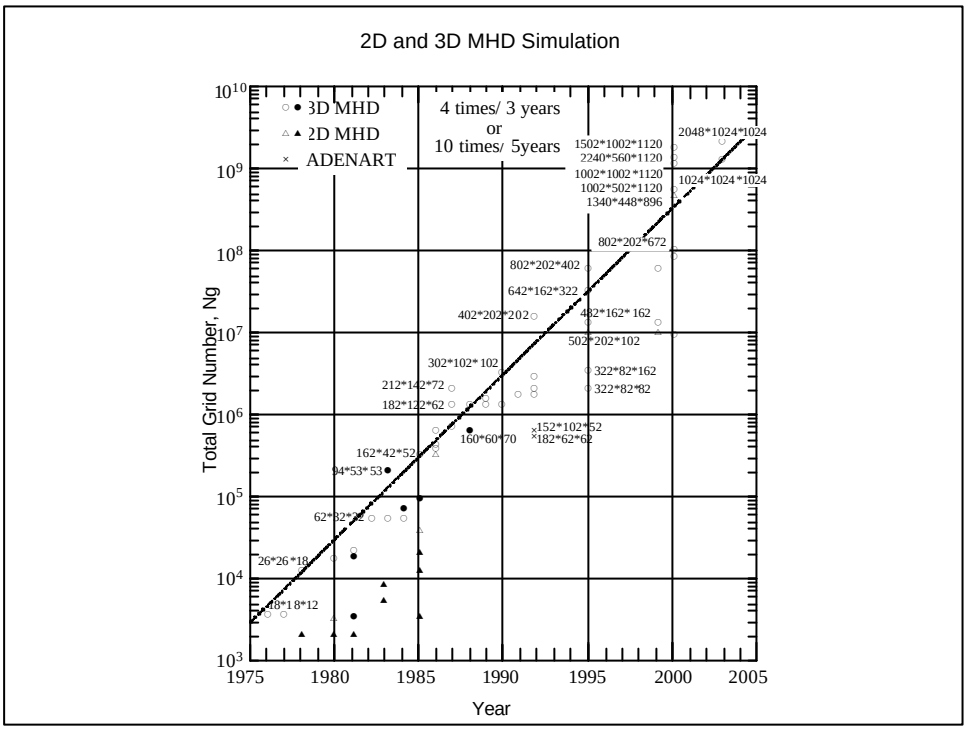
・利用できる総格子点数からみたMHDシミュレーションの年毎の発展拡大の傾向をみると、 この25年間で平均して３年で４倍あるいは５年で１０倍の割合で増大している。

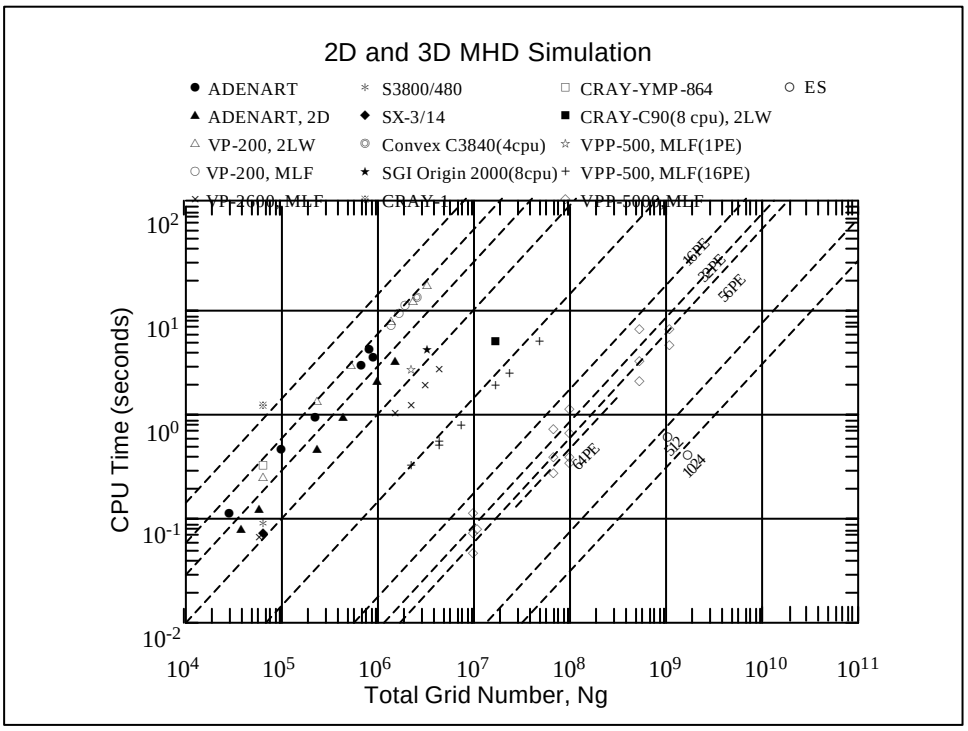
・最近の高性能のワークステーションやパーソナルコンピュータの計算速度の伸びは著しいが、最大級のスーパーコンピュータは常にそれらの約千倍の高速計算処理能力を有している。

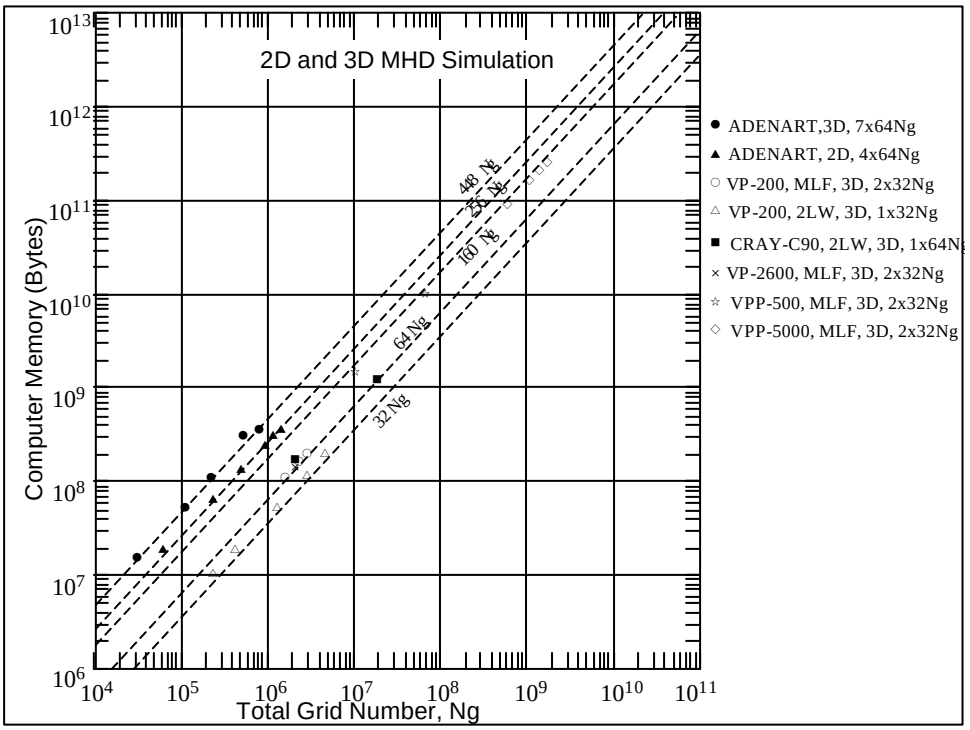
・多数の3次元格子点を用いた大容量シミュレーションでは、メモリ不足やRISC計算機でのcache不足などによって、更に決定的な差が生じる。

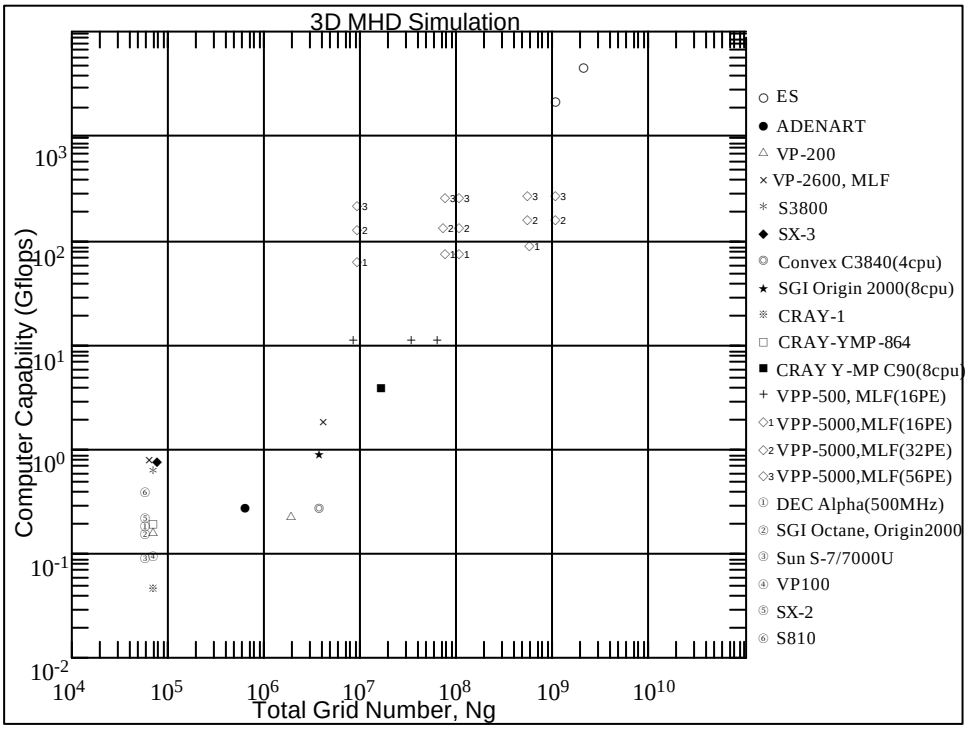
・フルベクトル化／フル並列化されたプログラムを用いて、VPP5000等のスーパーコンピュータで100時間かかる3次元MHDシミュレーションは、スーパーコンピュータ以外の通常のワークステーションやパーソナルコンピュータで実行することは全く不可能である。

・地球磁気圏の3次元グローバルMHDシミュレーションには、最大級のスーパーコンピュータの利用は常に必須の条件と言える。


2D and 3D MHD Simulation
3D MHD
2D MHD
ADENART
4 times/ 3 years
or
10 times/ 5years
Total Grid Number, Ng
Year
2048*1024*1024
1502*1002*1120
2240*560*1120
1002*1002*1120
1002*502*1120
1340*448*896
1024*1024*1024
802*202*672
802*202*402
642*162*322
402*202*202
482*162* 162
502*202*102
302*102* 102
322*82*162
322*82*82
212*142*72
182*122*62
160*60*70
152*102*52
182*62*62
162*42*52
94*53*53
62*32*32
26*26*18
18*18*12


2D and 3D MHD Simulation
ADENART
ADENART, 2D
VP-200, 2LW
VP-200, MLF
VP-2600, MLF
S3800/480
SX-3/14
Convex C3840(4cpu)
SGI Origin 2000(8cpu)
CRAY-1
CRAY-YMP-864
CRAY-C90(8 cpu), 2LW
VPP-500, MLF(1PE)
VPP-500, MLF(16PE)
VPP-5000, MLF
ES
16PE
32PE
56PE
64PE
512
1024
CPU Time (seconds)
Total Grid Number, Ng


2D and 3D MHD Simulation
Computer Memory (Bytes)
Total Grid Number, Ng
448 Ng
256 Ng
160 Ng
64 Ng
32 Ng
ADENART,3D, 7x64Ng
ADENART, 2D, 4x64Ng
VP-200, MLF, 3D, 2x32Ng
VP-200, 2LW, 3D, 1x32Ng
CRAY-C90, 2LW, 3D, 1x64Ng
VP-2600, MLF, 3D, 2x32Ng
VPP-500, MLF, 3D, 2x32Ng
VPP-5000, MLF, 3D, 2x32Ng


3D MHD Simulation
Computer Capability (Gflops)
Total Grid Number, Ng
ES
ADENART
VP-200
VP-2600, MLF
S3800
SX-3
Convex C3840(4cpu)
SGI Origin 2000(8cpu)
CRAY-1
CRAY-YMP-864
CRAY Y-MP C90(8cpu)
VPP-500, MLF(16PE)
VPP-5000,MLF(16PE)
VPP-5000,MLF(32PE)
VPP-5000,MLF(56PE)
DEC Alpha(500MHz)
SGI Octane, Origin2000
Sun S-7/7000U
VP100
SX-2
S810

８．まとめ

・ベクトル並列機VPP5000から、クラスタ型スカラ並列機PRIMEPOWER HPC2500への移行

・地球磁気圏の３次元MHDシミュレーションコードを用いたPRIMEPOWER HPC2500の性能評価

32cpuまでの利用ではスケーラビィティはかなりよい

グリッド数（nx,ny,nz）=(500,318,318)を用いて、
2cpuで21%(=0.116/5.2)の実行効率

8cpu で8.4 Gflops、32cpuで25.8 Gflops の計算速度を実現

・今回はスレッド並列は用いずに、MPIによるプロセス並列ジョブの測定のみ実行

・ノード内自動並列、ノード間利用者による並列化で高効率計算を実現するのは、非常に困難な問題である。

・PRIMEPOWER HPC2500は、設計思想としてスレッド並列（自動並列、あるいはOpenMPを使用したノード内並列）とプロセス並列（MPI等を使用したノード間並列）との併用による並列ジョブの実行を想定していて、最大性能を引き出すにあたってはその構成が望ましい。

・PRIMEPOWER HPC2500で更にcpu数を増やしてTflops の実行速度が実現すること、ベクトル並列機に負けない計算速度を実現することを強く期待する。

８．まとめ - ２

スーパーコンピューティングの４つの機能

計算速度
cpu 610 Gflops

主メモリ
1TB

磁気ディスク
17TB

＋

グラフィックス
動画、３次元可視化

ネットワーク
SuperSINET (10 Gbps)

近隣のアジア諸国の研究者・学生に日本の最大級のスーパーコンピュータを利用してもらう機会を提供できないか。

韓国、台湾、中国、インドなど