TSUBAME 共同利用　平成 23 年度　産業利用　成果報告書

利用課題名　超大規模三次元高周波電磁界シミュレーションへの GPU クラスター適用検証
英文：Verification of GPU cluster acceleration for super large scale electromagnetic simulation with CST STUDIO SUITE.

利用課題責任者
安永　高志

所属
株式会社エーイーティー　技術部
http://www.aetjapan.com

邦文抄録（300 字程度）

三次元電磁界シミュレーションは、普及に伴いその応用範囲の拡大を見せている。高速デジタル回路基板の発する EMC ノイズ対策、実際の自動車と同様の内外装とドライバーを含む生体電磁界解析など日常的に行われている。しかしながら、信号伝送高速化や構造複雑等の主因によって、シミュレーション実行に要するコンピュータリソースは増大の一途を辿っている。今日の民生電気機器や自動車などの解析では、数億メッシュ規模に至ることは珍しくなく、市販のワークステーションやサーバでの計算実行は事実上不可能なケースも多い。その打開策として有望視される大規模 MPI-GPU クラスターコンピューティングの適用可能性を探るため、国内外屈指の性能を誇る GPU グリッドクラスターである TSUBAME2.0 において、超大規模三次元高周波電磁界シミュレーションの実現可能性を検証した。本課題では商用電磁界解析プログラム CST STUDIO SUITE の三次元高周波電磁界シミュレーションモジュールである MW STUDIO を解析エンジンとして用いた。

英文抄録（100 words 程度）

Demands for very large scale 3D electromagnetic simulation is growing due to increasing simulation frequency and complexity of the structure, e.g. for high speed digital circuit module within electrical appliance or antenna propagation simulation of automotive. In order to cover such huge computation resource, GPU-MPI cluster computing is promising method. Multi purposed 3D electromagnetic simulation software CST MW STUDIO is applied for TSUBAME2.0 validation test, which get successful result at last.

Keywords: **大規模電磁界シミュレーション、MPI クラスター、CST STUDIO SUITE, MW STUDIO**

背景と目的

FDTD 法や FIT 法に代表される時間領域差分法に基づく三次元高周波電磁界シミュレーションでは、電磁波がメッシュを越えて伝搬する際発生するグリッド分散と呼ばれる誤差の蓄積を低減するために、解析最小波長 λ（at 最大周波数）に対し、λ /10~λ /20 程度のメッシュ割数を必要とする。その結果、解析最大周波数の三乗に比例してメッシュ数が増加し、それと同時に計算に要するメモリ量も増加する。

また、モデル寸法の大型化は、電磁波伝搬距離の拡大を意味し、電磁波伝搬シミュレート時間（タイムステップ数）もまた増加することも念頭に置く必要がある。

CST STUDIO SUITE の時間領域ソルバ（CST MW STUDIO）は、領域分割法(Domain decomposition method)に基づく MPI(Message Passing Interface)クラスター計算機において、最大 20 億メッシュの超大規模シミュレーションの実行が可能である。

その計算原理は、大容量メモリを要する大規模モデルを、一般的な計算機が解析可能な小区画の一群へと領域分割し、区画境界面の電磁界情報を各計算機ノード間で逐一交換することによって、大規模並列計算を可能とする。

時間領域ソルバによる行列演算は、計算を担うプロセッサとメモリ間のデータ転送帯域が律速条件となることから、CPU よりも数倍大きいメモリ帯域性能を有する GPGPU ユニットを計算サーバノード内で複数台併用することによって、シミュレーション速度を飛躍的に高めることが可能である。

その上で更に、MPI-GPU クラスターの並列計算スループットを確保するには、各計算ノードが低遅延かつ広帯域通信が可能な Infiniband 相当のネットワークで相互接続されていることが不可欠である。Infiniband に比べ通信遅延時間が一桁大きい GbE ネットワークでは、GPGPU によって各ノード内の計算速度が著しく向上したとしても、ノード間通信遅延がボトルネックとなり、ノード数を増やしても総合的なスループットの向上は達せられない。

以上の要件から、1,408 台の CPU ノードに 4,224 台の Tesla M2050 カードを装備し、それらが Infiniband QDRx2 チャンネル（80Gb/s）の超高速ネットワークで相互接続された TSUBAME2.0 は、三次元時間領域電磁界シミュレーション実行環境として最適なプラットフォームと目される。

TSUBAME2.0 上で CST MW STUDIO の実行が可能となれば、これまで数週間を要した大規模電磁界シミュレーションを短時間のうちに終えて結果を得られるようになり、既存の設計フローに明白なブレークスルーをもたらす。また、従来は不可能であった大規模問題への取組が可能ともなれば、新規性の高い研究・開発テーマの創出も期待される。

CST STUDIO SUITE の概要

CST STUDIO SUITE は、1992 年ドイツで設立された Computer Simulation Technology AG 社（CST）により開発された、三次元電磁界シミュレータである。その中核となる高周波電磁界解析モジュールである CST MW STUDIO は、RF, EDA, EMC/EMI 等の幅広いテーマの研究・開発において今日活用が図られている。

CST STUDIO SUITE インストールと動作確認

SUSE Linux Enterprise Server 11 が OS 稼働する TSUBAME において、CST STUDIO SUITE の Linux 版プログラムインストーラを実行し、エラーフリーでセットアップは完了した。

その後、インタラクティブノードにログインし、コマンドライン起動によって MW STUDIO 時間領域ソルバがシングルノード実行可能であることを確認した。

また、t2sub ジョブスケジューラを介した MPI ジョブについても、サブミットスクリプトで指定したキュー種、ノード数、メモリ量、GPU 数等指定リソースに基づき MPI-GPU クラスター計算が実行されることを確認した。

この時、Native Infiniband インターコネクトを介して MPI 並列計算が行われることも確認され、Intel MPI 共々全ての計算機構が正常に動作していることが最終的に確認された。

MPI-GPU クラスターベンチマークテストの概要

CST MW STUDIO と TSUBAME2.0 による大規模電磁界シミュレーションへの適用可能性を探る為、ベンチマークシミュレーションを実施した。

解析モデルとして、ドライバー人体を考慮した自動車モデルを設けた。電磁波の放射源としては、GPS アンテナをフロントダッシュボード中心付近に設置してある。

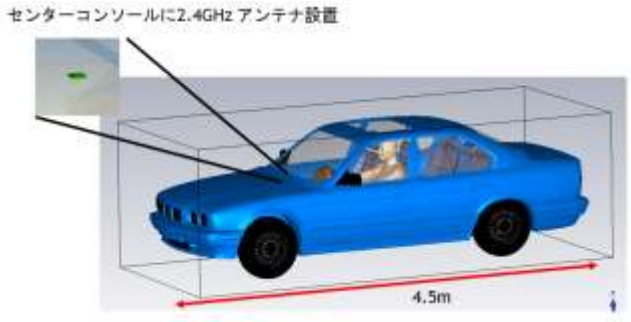

Fig. 1 自動車シミュレーションモデル外観

モデル	自動車(人体モデル含む)
電磁波放射源	GPS アンテナ (フロントダッシュボード中心付近)
解析周波数	2.4GHz
メッシュ数	560,339,364 メッシュ

表1 自動車シミュレーションモデル概要

マイクロ波周波数帯における人体組織の比誘電率は概ね40以上と大きく、自由空間(比誘電率＝1)との比較でいえば、6 分の1以下(～1/√40)の長さまで電磁波の波長は人体内で短縮する。つまり、先述の背景で述べた通り、人体の存在する空間においては、人体無しの場合に比較して 6^3＝216 倍のメッシュ数を割り当てる必要があり、所要メモリ量並びに計算時間が飛躍的に増加する。

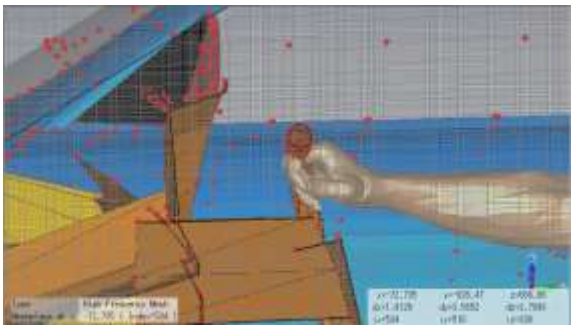

Fig. 2 ドライバー腕部付近のメッシュ分割状況

その為、ドライバー人体は存在しないという仮定の下に電磁界シミュレーションが往々にして行われているのが今日の現状である。

この段階でモデルは約 5.6 億メッシュに至っており、更に 10 億、20 億、50 億、…といった今後取り組むべき大規模モデルへの展望を占う試金石として手頃なモデル規模と判断した。

ベンチマーク実行を担う MPI 計算ノードは、最大 64 ノードの並列動作を受け入れられるキューとして、必然的に S キューが選ばれた。以下 5 パターンの MPI-GPU ノード構成に基いてシミュレーションを実行し、MPI 並列計算スループットと大規模シミュレーションへの適用可能性について検証を行った。

1. 8 MPI ノード並列(CPU のみ)
2. 16 MPI ノード並列(CPU のみ)
3. 32 MPI ノード並列(CPU のみ)
4. 64 MPI ノード並列(CPU のみ)
5. 8 MPI ノード並列(2GPU/ノード併用)

MPI-GPU クラスターベンチマーク結果

MPI-GPU クラスターにおける電磁界シミュレーションは、以下の工程に沿って進捗する。

1. 空間領域分割(所定のノード数相当の区画に分割したモデル情報を、各計算ノードへ送致)
2. マトリクス計算(分割されたモデルを各ノードで行列へ変換)
3. ソルバセットアップ(モデル行列をメインメモリまたは GPU メモリ上に配置)
4. ソルバ実行、電磁波時間発展計算
(Leap-Frog スキームによる電磁波伝搬計算)
5. 計算結果をフロントエンドノードに集約
6. 空間分割データの再結合処理(近傍電磁界分布、遠方界指向性データの一体化)

この順番沿って、各主要工程の所要時間や律速条件等について個々に解説したい。

まず最初の空間領域分割の工程は、短時間の処理でありながら、シミュレーション全体の成否を握っている。メッシュモデルに即してモデルを分割する際、後の MPI 並列計算時にノード間で負荷のアンバランスが生じる事が無きよう、各ノードに割り当てる区画領域の計算量が等しくなるよう最大限配慮される。

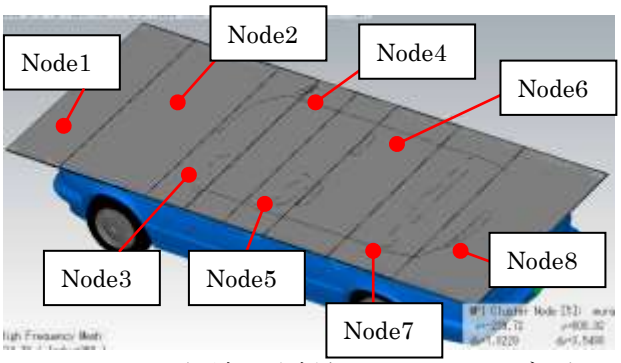

Fig. 3 8 つの領域に分割されたメッシュモデル例

領域分割されて各 MPI ノードに送致されたメッシュモデルは、Maxwell Grid Equation に則して行列に展開される。その処理に要する時間(Matrix calculation time)を下に示す。

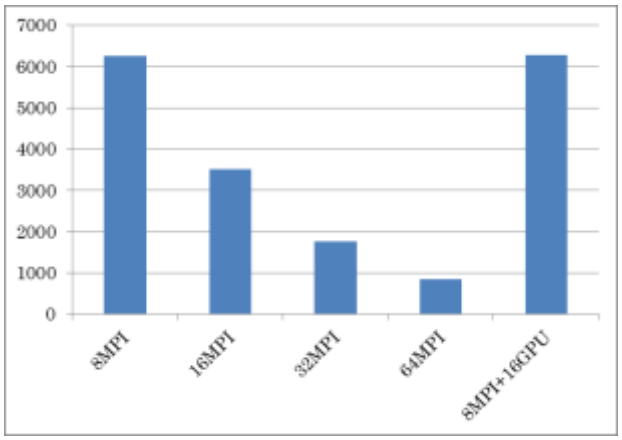

Fig. 4 Matrix calculation time in secs.

本処理に要する時間は、各 MPI ノード担当モデルのメッシュ数に比例するため、MPI ノード数を増やして割り当てモデルがダウンサイズするのに比例して所要時間も減少する傾向が見られる。

次に、本題のソルバ実行に相当する電磁波時間発展計算の性能結果を示しますが、まずは所要時間ではなく、毎秒何百万セルの電磁界を更新出来たか、そのスループット性能を以下に示す。

$$\text{Reflesh rate} = \frac{\text{メッシュセル数*タイムステップ数}}{\text{Soler loop time in secs.}}$$

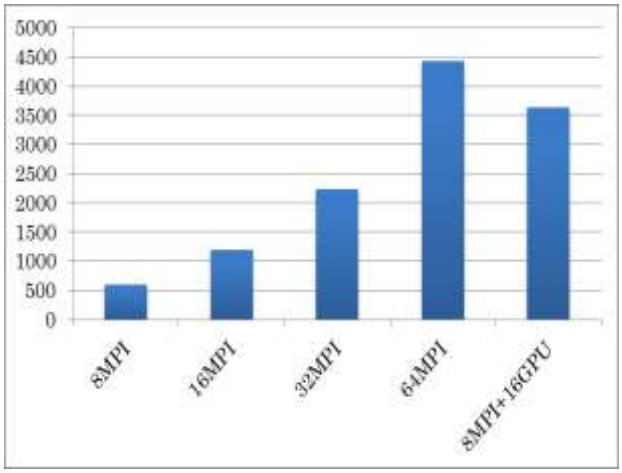

Fig. 5 Reflesh rate in Million cells per secs.

8,16,32,64MPI 並列の結果については、MPI 並列数に比例してスループットが向上する明快な傾向が観測される。ここで特筆すべきは、8MPI+16GPU 時の結果であり、Tesla M2050 カードをノード当たり 2 台併用するだけで、8MPI 比で約 6 倍の性能ゲインを得ている点である。単純なケーススタディで例えると、

1 台の CPU サーバ計算で、96 時間
8 MPI クラスター計算で、12 時間
8MPI+16GPU クラスター計算で、2 時間

といった劇的な時間短縮が図れること示唆しており、TSUBAME の潤沢な GPU リソース活用による巨大なベネフィットの一端が垣間見れた。参考のため、ソルバ実行時間(Solver loop time)も以下に示す。

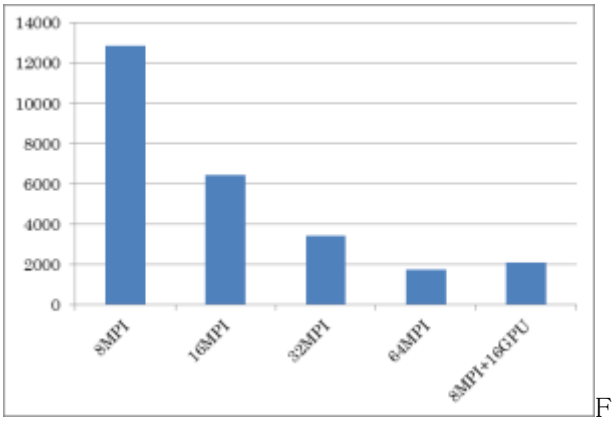

Fig. 6 Solver loop time in secs.

ソルバ実行を終えた後は、空間分割して個々に計算した電磁界分布データの再結合が行われる（MPI merge）。これが完了した後初めて Fig.7 に示すような空間的に連続な電磁界分布がプロット可能となる。

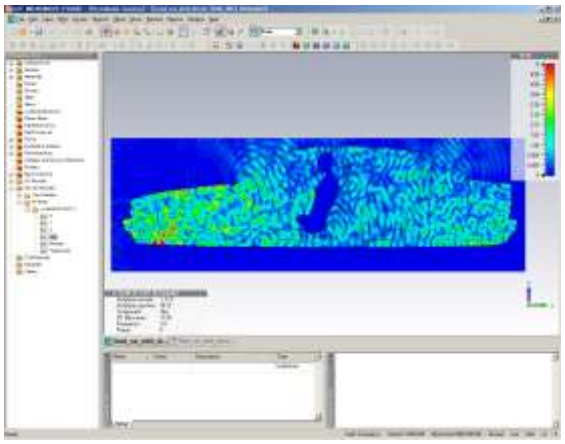

Fig. 7 GPS 通信周波数 2.4GHz における電界分布
（データサイズ 72.4GByte）

Fig.8 に示された通り、MPI merge の所要時間にも MPI 並列数との相関が観察される。これは、各ドメインが互いに接する面積が小さい（＝ドメイン分割数が多い）程に、再結合が速やかに行える為と思われる。

また、MPI merge の対象となる電磁界分布のバリエーション（モニター数）に応じて所要時間が増減することにも留意が必要である。

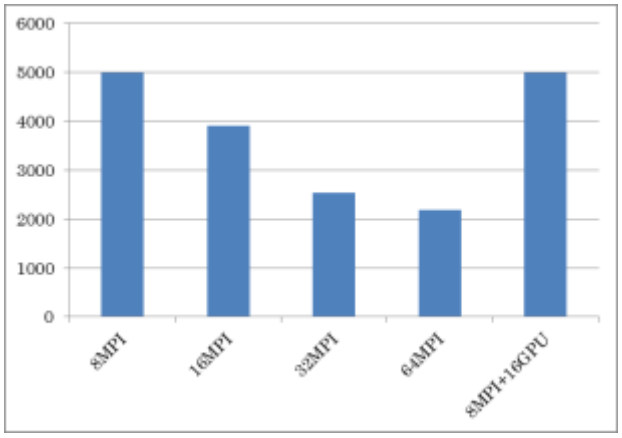

Fig.8 MPI merge time in secs.

最後に一連の MPI-GPU クラスター計算に要したトータルの時間（MPI total time)を Fig.9 に示す。内訳を見やすくするために、個々の工程の所要時間を累積グラフ形式で表す。

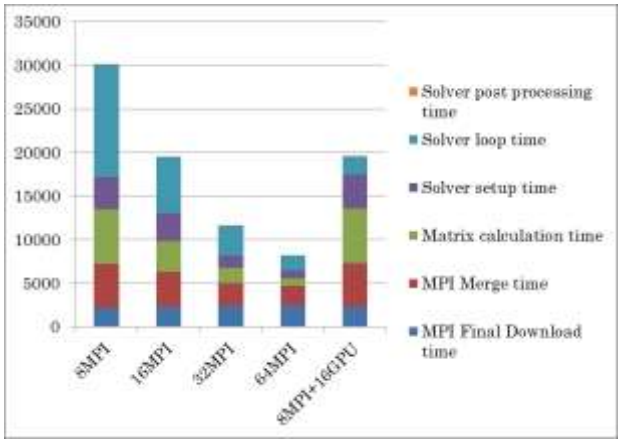

Fig. 9 MPI total time in secs.

今回の検証モデルは、総所要時間においてソルバ実行時間(Solver loop time)の占める割合が比較的小さいケースであり、結果としてそれ以外の工程の寄与が目立つ結果となった。

一般に取り組まれている現実的なモデルでは、ソルバ実行時間が大半を占めるケースが大勢であるため、この様なケースはむしろ少ない。

しかしながら、今回解説した各工程の背景を留意した上で、モデル規模に最適なクラスターノード選択を行うことは重要である。

まとめ、今後の課題

CST MW STUDIO が TSUBAME2.0 において問題なく動作することが確認された。

MPI クラスター計算に際し GPU を併用することで、ソルバ実行工程において、目覚ましい性能ゲインが得られることが確認され、今後取り組むべき数十億メッシュ規模のシミュレーション実行の足がかりを得た。

マトリクス計算、電磁界データ再結合処理を一層効率化する為、ルーチンの最適化あるいはデータ処理簡略化の手法を、プログラムに実装する必要性が確認された。（今後の CST STUDIO SUITE リリースにて改善を図る予定）

目下解析可能なモデル規模上限とされる 20 億セルを克服するための開発が進められており、今後のバージョンリリースについても引き続き検証を行い、更に大規模モデルのシミュレーション可能性を模索していく。（バージョン 2013 リリースでの対応を開発元 CST が公約済）