

異機種大規模クラスタ対応型 ホンダグリッドシステム

Honda Grid System for Large-scale Heterogeneous PC-cluster Group

前 川 史 人*
Fumito MAEKAWA

山 本 義 弘* 中 田 秀 基**
Yoshihiro YAMAMOTO Hidemoto NAKADA

要 旨

ABSTRACT

CAE計算環境の充実を図るべく、異機種大規模クラスタ対応型ホンダグリッドシステムを開発した。CAE計算環境が、ハイパフォーマンスCPUを積んだスーパーコンピュータからラック内にPCを何台も換装したPCクラスタへ移行しており、それに伴って急増したリソース(CPU)稼働率を向上し有効活用するために、このようなシステムの構築が望まれていた。

高品質なスケジューラを基盤とし、独自の機能を開発しつつ、環境に合わせたチューニングを施すことで、CPU稼働率を向上でき、これにより、リソース補強の効果と相まって、JOB数やJOB規模が増加したにも関わらず、待ち時間を含めた平均計算時間を約1/10に短縮できた。

また、同様の手法を使い、CAD用のデスクトップPCの有効活用も可能となり、PC上でのCAE計算環境も充実させた。

In order to enhance the CAE computation environment, a Honda grid system for a large-scale heterogeneous PC-cluster group was developed. This system is necessary for increasing the operating rate of and effectively utilizing resources (CPUs), the number of which increases rapidly, along with the shift of the CAE computation environment from that of a supercomputer containing a high performance CPU to a PC cluster consisting of a large number of PCs installed on a rack.

By carrying out tuning that matches the environment based on a high quality scheduler while developing unique functions, the operating rate of the CPU was able to be increased. This, together with the effect of reinforcing the resources, enabled the mean computation time, including the waiting time, to be reduced to approximately one-tenth, regardless of the increase in the job number and job scale.

Also, using the same method, a CAD desktop PC could be effectively utilized, and the CAE computation environment on the PC was also enhanced.

1. まえがき

CAE計算用として、数年前までは少数の強力なCPUを使って高速計算をおこなうスーパーコンピュータ(スパコン)が主流であった。しかし、このシステムは高価であり、規模を拡大するには莫大な投資が必要となる。

これに対して、PCの性能は日進月歩で向上しており、これらを複数台まとめてクラスタ化したPCクラスタを導入することにより、スパコン性能に匹敵する能力を確保しつつ、投資を抑えることが可能となってきた。

しかし、同等性能を維持しようとした場合、当時スパコ

ン8CPUに対し、PCクラスタ64CPUが必要となり、システムはそれまでの数十CPU規模から、数百CPU規模へと激増した。また、CAEへの解析精度の向上や開発期間短縮への期待が高まった結果、モデルのデータ量が増大し、それらを同等時間で計算するためにさらにPCクラスタを導入した結果、1万CPU(コア)に迫るまでに激増した。

このような状況になると、今までおこなっていたマニュアルでのJOB管理運用手法ではシステム運用が破綻するのは明らかであり、早急に運用の自動化と効率化を図らなければならない。

そこで、これらの課題を解決する手法として、グリッド

* 四輪開発センター

**独立行政法人 産業技術総合研究所

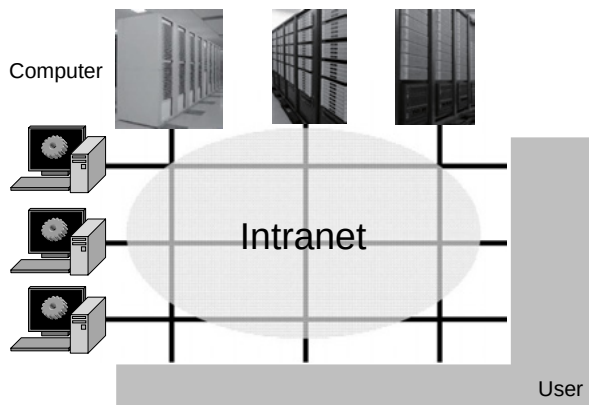


Fig. 1 Honda grid system

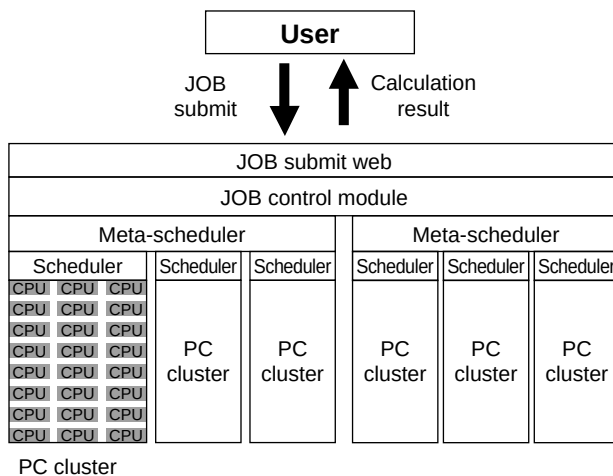


Fig. 2 Outline of Honda grid system

コンピューティング⁽¹⁾技術を応用した。計算要求があれば、社内LANに繋がっている不特定のリソースを使用し、計算して結果を返すというシステムである(Fig. 1)。

本稿では、数千 CPU の高効率稼働を実現しつつ、運用管理工数の低減に成功したホンダグリッドシステム(Fig. 2)に関する研究内容と技術について述べる。

2. システムのねらい

ホンダグリッドシステムの開発のねらいは、導入時期やメーカ、能力が単一ではない PC クラスタ群を、運用工数を掛けずに効率良く稼働させ、JOB 計算時間を低減させることにある。

目標として、社内大規模解析の基準計算において

- (1) CPU 稼働率 85%以上
- (2) 待ち時間を含めた計算時間を7時間以内
- (3) 管理運用工数の低減
- (4) CAD 用 PC の有効活用を定めた。

3. システム構成と実装機能

3.1. JOB スケジューリング機能

大規模 PC クラスタ環境でのグリッドシステム構築において、最も重要な機能はJOBスケジューリングである。投入された JOB を、管理下の CPU の状況や JOB 処理のポリシーに応じて各計算用 CPU に配分する機能であり、その役目を担うのがスケジューラと呼ばれるソフトウェアである。

ホンダグリッドシステムを構築する上でも、スケジューラの選定が成功の鍵を握るので、その評価項目を決め、その中でホンダグリッドの要件として

- (1) Linux と Windows の両方に対応できる
 - (2) 並列計算に強い(または強くできる)
 - (3) 稼働実績があり、安定運用できる
 - (4) CPU の増加に対しても、コスト増加が抑えられる
- を定めて十分に検討した結果、最終的にCondor⁽²⁾を選択した。

3.2. チェックポイントリスタート機能

この機能は、計算の途中の情報(チェックポイント)を一時的に保存し、後にその途中情報から計算の再実行(リスタート)ができるというものである(Fig. 3)。

従来の計算環境では、CPU の数が少なく JOB の数も少なかったので、ユーザごとの実行本数制限だけの単純な先入れ先出しで処理をおこなっていた。

しかし、CPU が増大し、それに伴って JOB 数も激増してくると、単純処理では JOB 処理や CPU 使用状況共に無駄が発生するようになってきた。この無駄を省き、効率の良い JOB 処理 /CPU 使用を達成するため、チェックポイントリスタート機能を導入するに至った。CPUの空きがある場合は、本数制限を解除し溜まっている JOB を処理する。その後、優先度の高いJOBが入ってくると、いったんJOBのチェックポイントを取り、その JOB を一時停止し CPU を空けて、高優先度JOBを処理する。他の計算が終わって CPU が空くと、チェックポイントから再実行する。

ここで課題となるのが、どのJOBを一時停止するかである。一度処理が始まったJOBを途中で止めるのは、ユーザに対してシステムに対する不信感を生む可能性がある。そこで、JOB を処理する順位を決める処理優先度とは逆の視点で、JOBを一時停止する退避優先度というものを考えた。処理開始時間が遅い(処理優先度が低い)JOB ほど退避優先度を高くすることで、後から処理開始された方から必要な時に一時停止する流れになる。

この考え方をベースに、処理のアルゴリズムを具現化し、ユーザに PR した結果、不信感を与えない機能としての導入が可能になった。また、この機能の実現はアプリケーションごとに方法が異なるので、それぞれに適合した処理方法を確立した。

本機能の実装により、全体の JOB 処理能力が 16%アップした。

3.3. メタスケジューラ機能

これまで、複数の計算サーバ(PCクラスタ)をそれぞれ部署やアプリケーション単位で使用していたので、業務の負荷に合わせる形でサーバの負荷も変動する傾向であった。ゆえにサーバ間に負荷のばらつきが発生し、効率の良くな

い運用となっていたが、そのばらつきを解消するために JOB の割り振りを手作業でおこなうのは事実上不可能であった。

そこで、自動的にそのばらつきを解消する(ロードバランスを図る)目的で、複数のサーバを一括管理し JOB の適正配分をおこなうメタスケジューラ機能を開発した(Fig. 4)。

その機能の流れは、

(1) JOB が投入されると、メタスケジューラ(親)が配下に

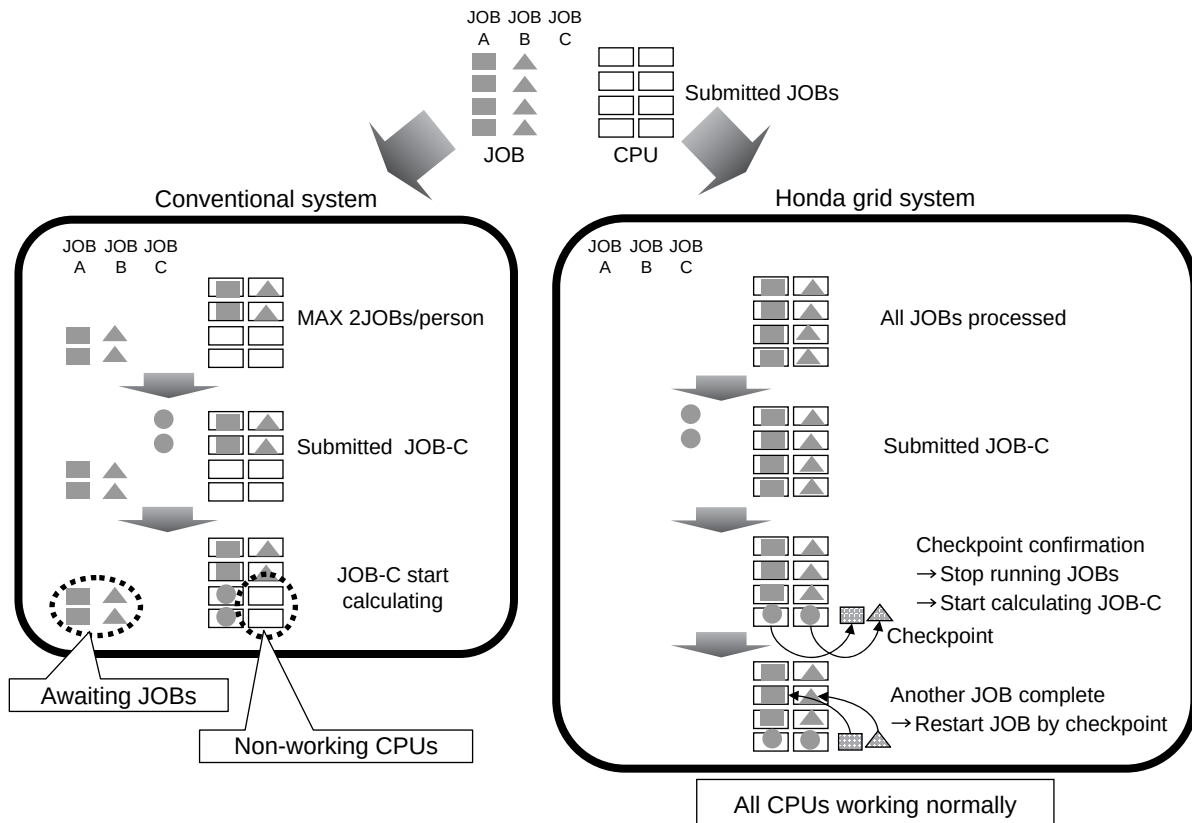


Fig. 3 Checkpoint and restart

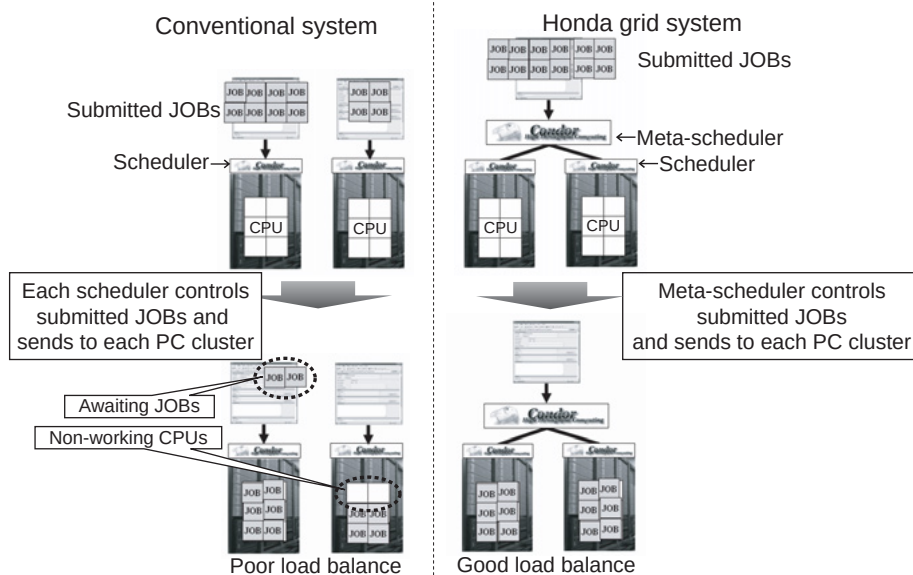


Fig. 4 Meta-scheduling

有る各サーバのスケジューラ(子)と情報交換し、サーバの負荷情報を基にJOBを配下の最適なスケジューラに渡す。

- (2) 渡されたスケジューラは、そのJOBを配下の空きCPUに配信し計算が始まる。
- (3) 計算が終了すると、子スケジューラがJOB終了情報をメタスケジューラに伝え、最終的にそのJOBをプールから削除する。

となる。

この機能によりサーバ間のロードバランスを図れるようになり、結果、CAE 計算環境全体の JOB 処理能力が向上した。

3.4. 簡易 JOB 投入機能

JOBを投入する際、使用アプリケーションごとにJOB投入手法や使用する並列計算用ソフトウェア(Message Passing Interface: MPI)が違うので、投入スクリプトの作成が煩雑になり、かつユーザ本人が作成していたことに絡んでスクリプト記述ミスが発生し、その結果JOBエラーが発生していた。

この工数の掛かるJOB投入をWeb化し、投入スクリプトを自動作成することで記述ミスを無くし、ユーザのJOB投入工数やJOBの計算エラーを削減した。

さらに、投入されたJOBをどのような順番で処理していくかの基準となるポリシーを定め、それに沿った優先順位付けをするプログラムを開発した。これにより、ユーザ間のJOB処理に対する不公平な状況は改善され、全てのユーザに等しいサービスを提供できるようになった。

3.5. システム自動監視機能

大規模環境上で膨大な並列計算を処理するホンダグリッドシステムにおいて、異常の早期発見がシステム全体の安定稼働に対しての最重要課題である。しかし、規模の拡大が

進むにしたがい、これまでのようなマニュアルでの監視対応は不可能である。さらに、何か異常を発見した際の早期復旧に対しても、新しい手法を構築する必要がある。

そこで本システムでは、以下のような対策を講じた。

- (1) 常にグリッドシステムを自動監視するツールを立ち上げ、それによる24時間監視システムを構築し、異常発見の際は担当者にメールを送る。
- (2) グリッドシステム監視業務を分割し、システム全体部門とグリッド関連部門に分けることで、異常発生時の原因追究と対策をスムーズにおこなう。
- (3) ハードウェア関連トラブルの早期解決を図るため、ハードウェアベンダとのホットラインを開設し、ベンダ側に専門担当者を置き、原因追究と対策をスムーズにおこなえるようにした。

これらの対策により、異常発生から回復までの工数と時間を短縮できた。

3.6. ホンダグリッドシステムの効果

本システム構築による効果を Fig. 5 に示す。

4. デスクトップ PC 群へのグリッド応用展開

今日、Windows下でさまざまなCAE計算アプリケーションが使われているが、PC上で計算中は他の作業ができない程パフォーマンスが低下する。しかし、周囲を見渡すと会議などで担当者が席を外し使われていないPCがあるが、それらに代わりに計算をさせることは容易にできない。また、人が使わない夜間に、計算を連続的にこないたいが環境が整っていない。

これらの要望を解決するため、デスクトップPC群に対して、前述のCondorを使ったグリッドの応用展開をおこなった(Fig. 6)。

ここでの課題は、PCの遊休状態の判断である。つまり、

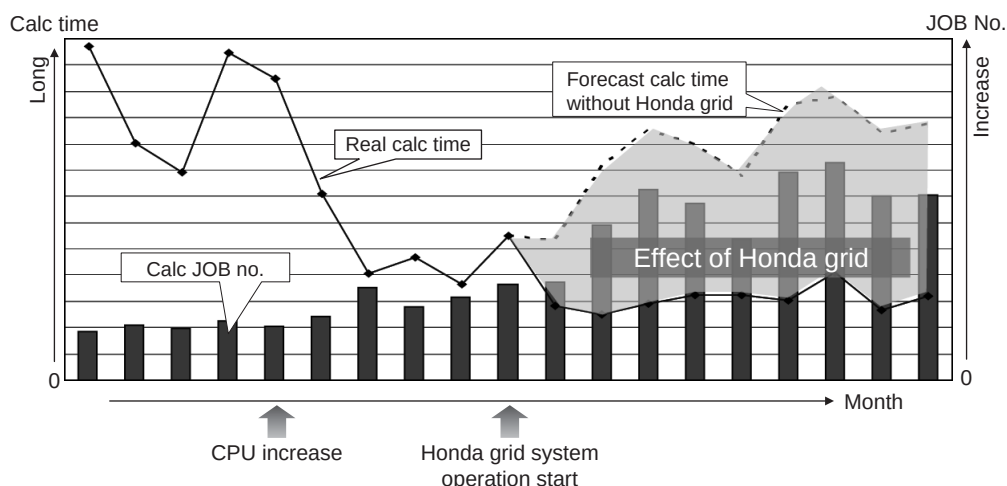


Fig. 5 Result of Honda grid system

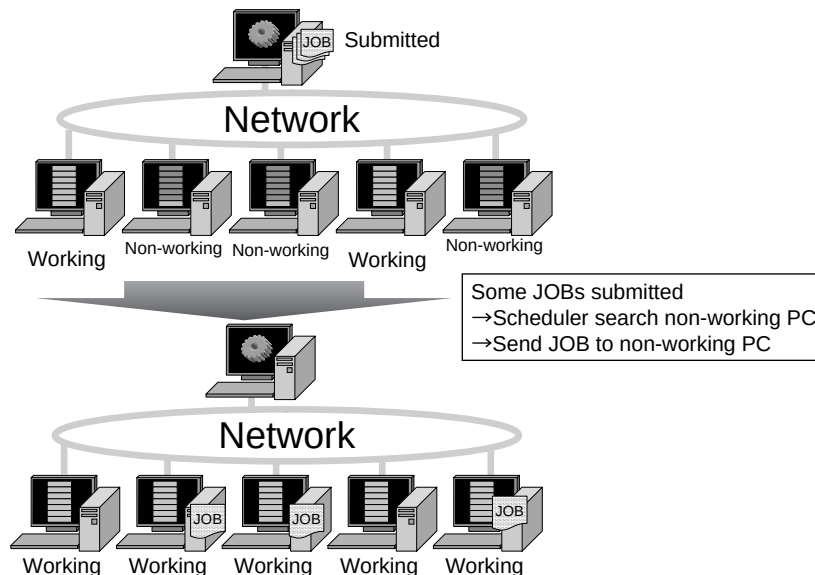


Fig. 6 Honda grid system on CAD PC

ちょっと席を外しているだけなのか、会議等で長時間使わないのかの判断をどうおこなうかである。

そこで本システムでは、

- (1)一定時間内にマウスやキーボードが触られていない。
- (2)一定時間内に CPU が稼働していない。

マシンを遊休 PC とみなし、それらに対して JOB を振り分けるようにした。

そして、JOB 処理中のマシンがユーザに使われた場合はすぐに計算を止め、別の遊休マシンで再計算をおこなうようにした。また、夜間に関しては、PCがユーザに使われることは無いので、投入されたJOBを処理するのに必要な台数を予測し、それらの PC だけを夜間稼働し、計算を実行させるようにした。

この仕様により、PCユーザに負担を掛けずに、効率的かつ信頼性高く JOB を処理する環境が整った。

また、JOB の投入も Web 化することで JOB 投入工数を 1/10 に削減できた。

5. まとめ

大規模 PC クラスタ環境を有効に機能させるいくつかの機能を開発し、ホンダグリッドシステムを構築した。その効果を以下にまとめる。

- (1)CPU 稼働率を 50%から 85%に向上させ、計算能力を 1.7 倍にした。
- (2)CPU増強分と合わせて、計算時間(待ち時間を含む)を約 1/10 に短縮した。
- (3)デスクトップ PC 群をグリッド化し、Windows 環境下での CAE 計算サービスを充実させることが可能になった。

参考文献

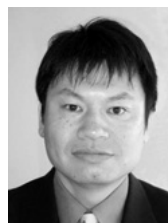
(1) グリッド協議会

<http://www.jpgrid.org/>, 2008/5/20

(2) Condor Project

<http://www.cs.wisc.edu/condor/>, 2008/4/1

■ 著 者 ■



前川 史人



山本 義弘



中田 秀基