1A-2

グローバルジョブスケジューリングアルゴリズム評価用シミュレータ

渡邊　啓正[†]　本多　弘樹[†]　Massimiliano Rak[††] Umberto Villano[†††]

電気通信大学　大学院情報システム学研究科[†]

Second University of Naples[††]、University of Sannio[†††]

1.　はじめに

グローバルジョブスケジューリングは、グリッドで動作する並列分散プログラム群（ジョブ）に対し、所定の最適化方針にしたがって、実行に割り当てるグリッドの計算資源を計画することである。

グリッドの計算資源は次の点で、不安定な計算基盤である。すなわち、（１）構成が不均質である点、（２）構成や性能が変動する点、（３）複数の組織によって運用される点、である。

一般に、ジョブの実行において、ジョブ投入者の指定したサービス品質（完了期限など）を満たすことが必要である。グローバルジョブスケジューリングでは、前述の資源の不安定性に対応しながら、ジョブ実行におけるサービス品質を維持することが必要である。以降では、ジョブ実行においてサービス品質が満たされないことをジョブの失敗と呼ぶ。

以上の要件を満たすべく、われわれは、ジョブの失敗の少ないグローバルジョブスケジューリングアルゴリズムを開発している。

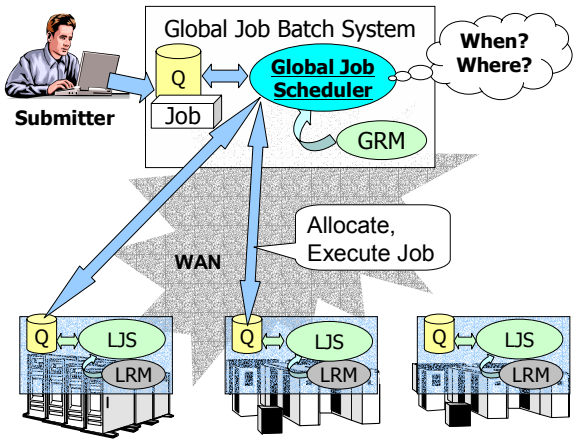
本稿では、提案するグローバルジョブスケジューリングアルゴリズムの概要を述べる。また、アルゴリズム評価用シミュレータの要件、および実装中の評価用シミュレータについて述べる。

2.　提案するグローバルジョブスケジューリングアルゴリズム

われわれはジョブの失敗を減らすべく、グローバルジョブスケジューリングアルゴリズムに次の技術を導入する。

デッドライン保証

ジョブに設定されたデッドライン（完了期限）を遵守するべく、デッドラインの早いジョブから実行する（Earliest Deadline First）。



Q: ジョブキュー
GRM: グローバル資源情報サーバ
LJS: ローカルジョブスケジューラ
LRM: ローカル資源情報サーバ

図１．グローバルジョブスケジューリング

また、ジョブ投入時において、ジョブの実行時間を予測し、デッドラインを遵守できないと推測した場合ジョブを投入できないようにする。

資源の先行予約

ジョブに対して資源を先行予約し、予約絶対優先でジョブを実行する。

耐故障

ジョブに設定された冗長度に基づいて、ジョブを複数の資源で冗長実行する。資源の故障によるジョブの失敗を回避できる可能性が大きくなる。

資源の運用ポリシーを反映した割り当て

休日やメンテナンス等のために、あらかじめ利用不可能と判っている資源を利用候補から除外する。

3.　評価用シミュレータの要件

提案アルゴリズムの有効性を検証するために、シミュレータによって再現される必要のある事項は以下のとおりである。

- 計算機（マルチプロセッサ対応であること）
- クラスタ内ネットワーク、LAN、WAN

Simulator for Evaluating Global Job Scheduling Algorithm

† Hiromasa Watanabe, Hiroki Honda

† Graduate School of Information Systems, University of Electro-Communications

- 計算機の先行予約
- 計算機の故障
- 計算機の運用ポリシー（利用可能性のスケジュール）
- ローカルジョブバッチシステム
- グローバルジョブバッチシステム
- サイト別情報提供サービス
- 並列分散プログラムの動作（複数同時実行が可能であること）

さらに、運用ポリシーが日～週単位で設定されるため、月単位のシミュレーションを実用レベルで短時間に完了できる必要がある。

4. 提案シミュレータ

われわれは既存の並列プログラム実行シミュレータ HeSSE[1]を拡張することで、評価用シミュレータを実現する。

4.1 HeSSE

HeSSE は以下を再現できる。

- SMP クラスタ
- クラスタ内ネットワーク、LAN（Ethernet）
- 任意の並列プログラムの動作
  - 並列プログラムは、計算ブロックの計算量と、通信パターンを記述したメタプログラムで表現される。

また、HeSSE には以下の特徴がある。

- 高精度である
  - 実際の並列プログラムの実行時間と、シミュレーション内のメタプログラムの実行時間の誤差が常に 5%未満である[2]。
- シミュレーションが実時間で短く完了する
  - 実際の実行に 15 分～1 時間要する並列プログラムに対し、メタプログラムのシミュレーション実行では入力パラメータによらず実時間で 30 秒しか要さなかった（30～120 倍の時間短縮）[2]。

4.2 HeSSE の機能拡張

評価用シミュレータの実現のために HeSSE に必要な機能拡張は次のとおりである。

A) WAN（ゲートウェイ）
B) 資源の先行予約
C) 資源の故障
D) 資源の運用ポリシー
E) メタプログラムのスケジュール起動
F) メタプログラムの複数同時実行
G) ローカルジョブバッチシステム
H) グローバルジョブバッチシステム
I) サイト別資源情報提供サービス

本稿執筆時点で、E)・F)・G)の拡張の実装を完了し、正常動作を確認できている。

5. 関連研究

Bricks[3], Anastatia[4]等、グリッドのシミュレータが開発されている。しかしながら、これらは機能やソフトウェア公開状況の点で評価用シミュレータには適さない。Bricks では並列分散プログラムのタスク間通信を再現できない。Anastasia はスケジューラコンポーネントの改変が困難であるため、グローバルジョブスケジューリングアルゴリズムの評価に適さない。

6. おわりに

本稿では、ジョブ失敗率を低減するグローバルジョブスケジューリングアルゴリズムを提案した。また、グローバルジョブスケジューリングアルゴリズム評価用シミュレータの要件、およびわれわれが提案するシミュレータの概要を述べた。

今後は提案シミュレータの実装を進め、機能と精度を評価する。そして提案シミュレータを用いて提案アルゴリズムの有効性を評価する。

参考文献

[1] N. Mazzocca, M. Rak, U. Villano, “The Transition from a PVM Program Simulator to a Heterogeneous System Simulator: The HeSSE Project”, Proceedings in 7th European PVM/MPI Users’ Group Meeting, pp.266-273, 2000.

[2] T. Fahringer, N. Mazzocca, M. Rak, S. Pllana, U. Villano, G. Madsen, “Performance Modeling of Scientific Applications: Scalability Analysis of LAPW0”, Proceedings in 11th Euromicro Conference on Parallel, Distributed and Network-Based Processing (Euro PDP 2003), pp.5-12, 2003.

[3] 竹房 あつ子, 合田 憲人, 松岡 聡, 中田 秀基, 長嶋 雲兵, “グローバルコンピューティングのスケジューリングのための性能評価システム”, 情報処理学会論文誌, vol. 41, No. 5, pp. 1628-1638, 2000.

[4] 門島 正和, 鈴木 雄大, 柴田 俊介, 中島 浩：メガスケールシミュレータ Anastasia における高精度タスクモデルシミュレーション,情報処理学会研究報告 2004-HPC-99, pp.73-78, 2004.