VPP システムにおける並列ジョブスケジューリング*

4Z－4

宇野俊司† 青島直人‡ 坂井賢一§
富士通株式会社¶

1 はじめに

近年，分散並列計算機の個々の Processor Element (PE) の持つ資源 (CPU パワーやメモリ等) はますます増加しており，並列ジョブを含んだ多数のジョブを同時に実行する技術がより重要となっている。本稿では，同時に複数の並列ジョブを効率よく処理するためのジョブスケジューリングの要件，及び実現のための技術と効果について，VPP システムの方式を例に考察する。

2 VPP システムの概要

VPP システムは図 1 に示すような分散メモリ型並列システムである。OS は各 PE 独立に存在し，並列ジョブは各 PE にまたがって実行される。

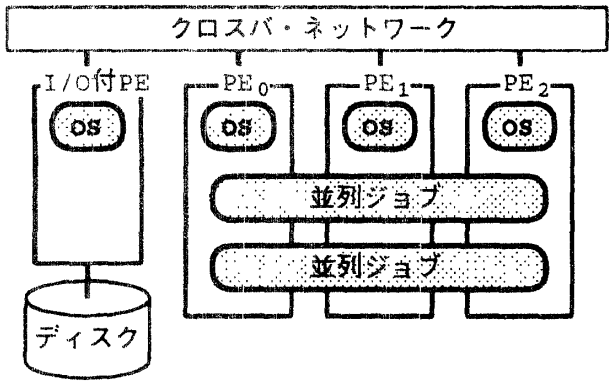

図 1: VPP システム

3 並列ジョブスケジューリングの要件

複数の並列ジョブが PE を共有して動作する場合に，以下のスケジューリングの要件がある。

[要件 1] PE 占有の場合と同等なジョブ実行効率:
PE を共有して並列ジョブを実行する場合，同期の処理が効率に与える影響が最も大きく，同期を効率良く行なうことが重要である。データパラレルのような並行して行う処理の量がほぼ等しいジョブは，PE を占有した場合，同期待ちが短い。VPP にはハードウェアによる高速な同期 (バリア) 機構があるため，短い同期は CPU ループによって効率良く処理できる。このため，PE を共有する場合でも，占有した場合と同等な同期待ち時間でジョブ実行できる必要がある。

[要件 2] 同期待ち時間の有効利用:
並列ジョブが並行して行う処理の量に大きな格差がある時，PE を占有した実行であっても長い同期待ち時間を要する。CPU の使用効率の向上のために，この同期待ちの間の CPU を他ジョブで使用可能にする必要がある。

4 VPP システムの技術

前述の要件を満たすため，VPP システムでは以下のような 2 つの技術を要件 1, 2 の各々に対して導入した。

[技術 1] 並列ジョブの同期スケジューリング (Syncronized Parallel[SP] スケジューリング)

[技術 2] 同期待ち解除機能

SP スケジューリングとは，各 PE 上に分散して実行される並列ジョブに対して，同じ時刻に CPU を与えることにより，同期のポイントに到達するタイミングをずらさないようにするスケジューリング技術である (図 2 にイメージを示す)。

並列ジョブに CPU を与えるタイミングを PE 間の通信によって決定すると，オーバーヘッドが無視できない。VPP システムでは，各 PE 間で高精度に同期したハードウェア・タイマを実装しており，絶対時刻を基準にスケジューリングすることでオーバーヘッドなく実現した。

同期待ち解除機能とは，同期完了までスリープし，同期完了の検出とスリープの解除の処理を行なう機

*Parallel Job Scheduling on VPP System
†Shunji Uno
‡Naoto Aoshima
§Ken-ichi Sakai
¶FUJITSU LIMITED

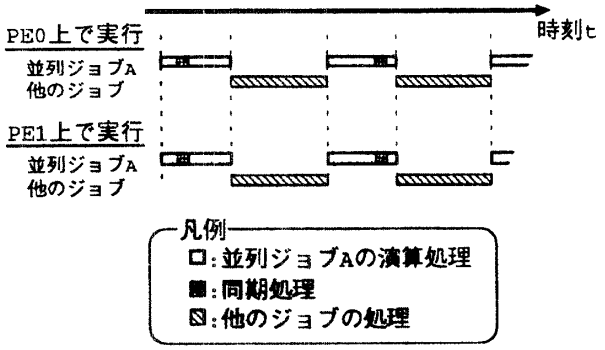

図 2: SP スケジューリングのイメージ

能である。実行効率の観点から，各 PE において高速，かつ，同時にスリープの解除を行なう必要があり，VPP では，同期完了時に各 PE で同時に割り込みを発生させるハードウェア機構を追加して機能を実現した。図 3 に本機能によって同期待ちの間に他のジョブが実行可能になるイメージを示す。

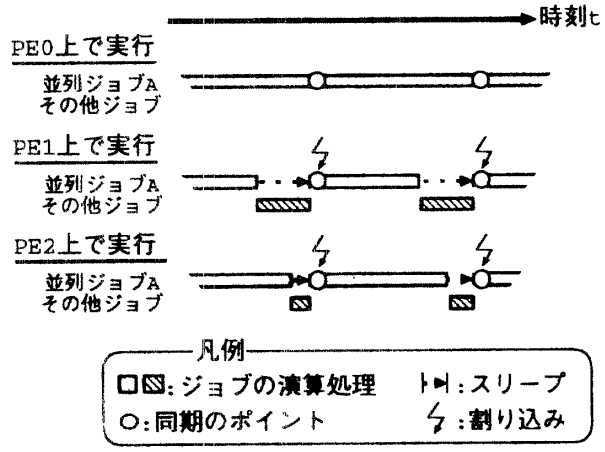

図 3: 同期待ち解除機能のイメージ

5　効果

図 4 に SP スケジューリングの効果を，図 5 に同期待ち解除機能の効果を示す。

図 4 は CPU ループによる同期を 40 万回行なう 3 並列のジョブを PE 共有で 2 多重実行するモデルで，ラウンドロビンでスケジュールした場合と SP スケジューリングを使用した場合のエラップス時間と消費 CPU 時間の結果である。比較のため，PE 占有で 1 多重実行した場合の測定結果を示した。これを見ると，SP スケジューリングを実施した場合は，PE 占有で実行した場合と同等の実行効率が達成できることが分かる。

図 5 は 1PE だけ I/O を発行する 3 並列のジョブを 2 多重で実行するモデルで，同期待ち解除機能によって同期した場合と CPU ループによって同期した場合の，エラップス時間と消費 CPU 時間の結果である。同期待ち解除機能を使用した場合，CPU ループの場合よりも短い時間でジョブが実行できていることが分かる。また，消費 CPU 時間が少なくなっており，PE を共有している場合の実行効率の向上に有効であることが分かる。

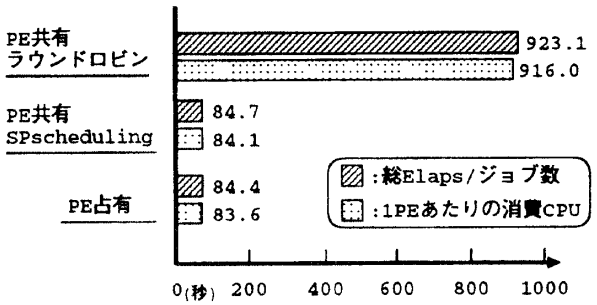

図 4: SP スケジューリングの効果

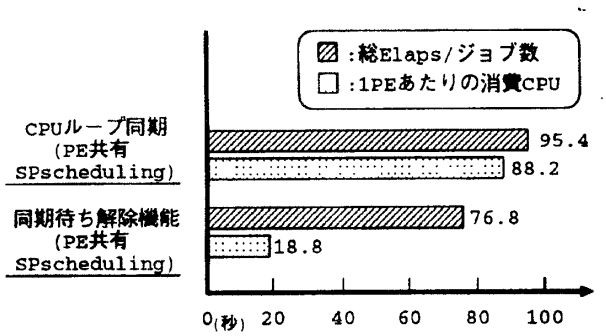

図 5: 同期待ち解除機能の効果

6　まとめ

本稿では分散並列計算機の資源をより効率良く使用するための並列ジョブスケジューリングの要件，及びそれらを実現した VPP システムの技術と効果について述べた。

なお，現状の VPP システムにおいては，1PE あたり 2 多重までの並列ジョブについてこれらの技術が適用されているが，今後はより大きな多重度が必要となる。そのため，多重度の増加，及び並列ジョブの組合せの複雑化に対して，更に効率的なスケジュールの方式を検討していく必要があると考えている。