

【公報種別】特許法第17条の2の規定による補正の掲載

【部門区分】第6部門第3区分

【発行日】平成20年8月21日(2008.8.21)

【公表番号】特表2008-516346(P2008-516346A)

【公表日】平成20年5月15日(2008.5.15)

【年通号数】公開・登録公報2008-019

【出願番号】特願2007-535843(P2007-535843)

【国際特許分類】

G 0 6 F 9/50 (2006.01)

【F I】

G 0 6 F 9/46 4 6 5 Z

【手続補正書】

【提出日】平成20年7月3日(2008.7.3)

【手続補正1】

【補正対象書類名】特許請求の範囲

【補正対象項目名】全文

【補正方法】変更

【補正の内容】

【特許請求の範囲】

【請求項1】

並列処理システム上の2つまたはそれ以上の並列計算ノードにわたってソフトウェア・アプリケーションの割り振りを最適化するためのコンピュータ実施方法であって、

通信ファブリックによって互いに結合された並列処理システム上の2つまたはそれ以上の並列計算ノードにわたって実行するための複数のソフトウェア計算を有する、ソフトウェア・アプリケーションの通信プロファイルを受け取るステップであって、前記複数のソフトウェア計算のうちの第1のソフトウェア計算が、前記複数のソフトウェア計算のうちの第2のソフトウェア・アプリケーションからアプリケーション・データを送り、前記通信プロファイルが、前記ソフトウェア計算のそれぞれの間で送られるソフトウェア・アプリケーション・データの量を含む、受け取るステップと、

ノード間で転送されるソフトウェア・アプリケーション・データの合計量について、ノード間通信コスト関数を数学的に最小にするステップであって、前記通信コストは、前記ソフトウェア・アプリケーションが前記ソフトウェア計算間でのソフトウェア・アプリケーション・データ通信に対して前記通信ファブリックを使用する方法に依存する、最小にするステップと、

を有する、コンピュータ実施方法。

【請求項2】

ノード間で転送されるソフトウェア・アプリケーション・データの前記合計量が、前記ソフトウェア・アプリケーションの前記通信コストを削減するように、グローバルおよび離散時間間隔のうちの少なくとも1つで最小化される、請求項1に記載のコンピュータ実施方法。

【請求項3】

前記通信コストが、使用可能通信トポロジの使用、ネットワーク・トポロジのタイプ、バッファの可用度、ネットワーク・トポロジでのコンテンツ、ソフトウェア・アプリケーション・オーバーヘッド、および通信待ち時間のうちの、少なくとも1つに依存する、請求項1に記載のコンピュータ実施方法。

【請求項4】

前記コスト関数を数学的に最小にすることに応答して、少なくとも前記ソフトウェア計算で、前記複数のノードのうちの1つから前記複数のノードのうちの他のノードへ再配置

するステップをさらに有する、請求項 1 に記載のコンピュータ実施方法。

【請求項 5】

前記ソフトウェア・アプリケーションの前記通信プロファイルを受け取るステップが、以前にマッピングされたノードと共にアプリケーション・データを送る前記ソフトウェア計算のそれぞれが、最近隣にあるノードとして定義された近隣ノードにマッピングされるように、前記ソフトウェア計算のそれぞれを前記ノードのうちの 1 つに順次マッピングするヒューリスティック・マップを受け取るステップを含む、請求項 1 に記載のコンピュータ実施方法。

【請求項 6】

前記ソフトウェア・アプリケーションの前記通信プロファイルを受け取るステップが、それらの間で送られるソフトウェア・アプリケーション・データの量に関して最大限に独立したソフトウェア計算を識別する、グラフ区画マップを受け取るステップを含む、請求項 1 に記載のコンピュータ実施方法。

【請求項 7】

前記グラフ区画マップが、グラフ内のクリークおよびバイクリークを決定するために少なくとも 1 つのアルゴリズムを使用するステップを含む、請求項 6 に記載のコンピュータ実施方法。

【請求項 8】

前記グラフ区画マップが M E T I S ツールを使用するステップを含む、請求項 6 に記載のコンピュータ実施方法。

【請求項 9】

前記ソフトウェア・アプリケーションの前記通信プロファイルを受け取るステップが、並列処理システム上の 2 つまたはそれ以上の並列計算ノード j にわたって実行するための複数のソフトウェア計算 i を有し、前記複数のソフトウェア計算のうちの前記第 1 のソフトウェア計算が、前記複数のソフトウェア計算のうちの前記第 2 のソフトウェア・アプリケーションからアプリケーション・データを送り、前記通信プロファイルが、 $C(i, j)$ によって与えられた前記ソフトウェア計算のそれぞれの間で送られるソフトウェア・アプリケーション・データの量を含み、

前記ノード間通信コスト関数が、前記第 1 のソフトウェア・アプリケーションを前記 2 つまたはそれ以上のノードのうちの第 2 のノードへ再配置することによって、前記 2 つまたはそれ以上のノードのうちの第 1 のノードに割り振られる前記第 1 のソフトウェア・アプリケーションについて、 $F = \sum_{i, j} C(i, j) H(i, j)$ によって与えられ、前記ノード間通信コスト関数が、前記第 1 のノードで送られるソフトウェア・アプリケーション・データの前記量と、 $H(i, j)$ によって与えられる、前記第 1 のノードがソフトウェア・アプリケーション・データを送信した際に使用した前記第 1 のノードと残りのすべてのノードとの間の通信パス・コスト／データ単位との、積によって与えられる、請求項 1 に記載のコンピュータ実施方法。

【請求項 10】

前記ソフトウェア・アプリケーションの前記通信プロファイルを受け取るステップが、各ノードがそれらの間にトラス通信ネットワークを備えて 3 D 格子に配置構成される、並列処理システム上で前記プロファイルを受け取るステップを含む、請求項 1 に記載のコンピュータ実施方法。

【請求項 11】

前記ソフトウェア・アプリケーションの前記通信プロファイルを受け取るステップが、Blue Gene/L スーパーコンピュータである並列処理システム上で前記プロファイルを受け取るステップを含む、請求項 1 に記載のコンピュータ実施方法。

【請求項 12】

請求項 1-11 のいずれか 1 項の方法の各ステップをコンピュータに実行させるためのコンピュータ・プログラム。

【請求項 13】

複数の独立ネットワークによって相互接続された少なくとも2つの並列計算ノードを備え、各ノードは、通信プロファイルに従って複数の並列アルゴリズム操作を含むソフトウェア・アプリケーションを実行する場合に、計算および通信アクティビティのうちの少なくとも1つを必要に応じて実行するための1つまたは複数の処理要素を含み、前記複数の独立ネットワークは、少なくとも1つのノードおよびその独立した区画サブセットの間で、ポイント・ツー・ポイント、グローバル・ツリー通信、ならびにグローバル・バリアおよび通知操作を実行可能にするためのネットワークを含み、

前記並列計算ノードを相互接続する前記複数の独立ネットワークの組み合わせは、前記ソフトウェア・アプリケーションの帯域幅および待ち時間の要件に従って、協調的および単独の少なくとも1つで利用され、前記複数のソフトウェア計算のうちの第1のソフトウェア計算は、前記複数のソフトウェア・アルゴリズムのうちの第2のソフトウェア・アプリケーションからアプリケーション・データを送り、通信プロファイルは、前記ソフトウェア・アルゴリズムのそれぞれの間で送られるソフトウェア・アプリケーション・データの量を含み、

前記並列計算ノードを相互接続する前記複数の独立ネットワークの組み合わせは、ノード間で転送されるソフトウェア・アプリケーション・データの合計量について、ノード間通信コスト関数を数学的に最小にすることによって利用され、前記通信コストは、前記ソフトウェア・アルゴリズムが前記ソフトウェア・アルゴリズム間でのソフトウェア・アプリケーション・データ通信に対して前記通信ファブリックを使用する方法に依存する、超並列処理システム。

【請求項14】

ノード間で転送されるソフトウェア・アプリケーション・データの前記合計量が、前記ソフトウェア・アプリケーションの前記通信コストを削減するように、グローバルおよび離散時間間隔のうちの少なくとも1つで最小化される、請求項23に記載の超並列処理システム。

【請求項15】

前記通信コストが、使用可能通信トポロジの使用、ネットワーク・トポロジのタイプ、バッファの可用度、ネットワーク・トポロジでのコンテンション、ソフトウェア・アプリケーション・オーバーヘッド、および通信待ち時間のうちの、少なくとも1つに依存する、請求項13に記載の超並列処理システム。

【請求項16】

前記コスト関数を数学的に最小にすることに応答して、少なくとも前記ソフトウェア計算で、前記複数のノードのうちの1つから前記複数のノードのうちの他のノードへ再配置するための手段をさらに有する、請求項13に記載の超並列処理システム。

【請求項17】

前記ソフトウェア・アプリケーションの前記通信プロファイルが、以前にマッピングされたノードと共にアプリケーション・データを送る前記ソフトウェア計算のそれぞれが、最近隣にあるノードとして定義された近隣ノードにマッピングされるように、前記ソフトウェア計算のそれぞれを前記ノードのうちの1つに順次マッピングするヒューリスティック・マップを受け取るステップを含む、請求項13に記載の超並列処理システム。

【請求項18】

前記ソフトウェア・アプリケーションの前記通信プロファイルが、それらの間で送られるソフトウェア・アプリケーション・データの量に関して最大限に独立したソフトウェア計算を識別する、グラフ区画マップを受け取るステップを含む、請求項13に記載の超並列処理システム。

【請求項19】

前記グラフ区画マップが、グラフ内のクリークおよびバイクリークを決定するために少なくとも1つのアルゴリズムを使用するステップを含む、請求項18に記載の超並列処理システム。

【請求項20】

前記グラフ区画マップが M E T I S ツールを使用するステップを含む、請求項 18 に記載の超並列処理システム。

【請求項 21】

前記ソフトウェア・アプリケーションの前記通信プロファイルが、並列処理システム上の 2 つまたはそれ以上の並列計算ノード j にわたって実行するための複数のソフトウェア計算 i を有し、前記複数のソフトウェア計算のうちの前記第 1 のソフトウェア計算が、前記複数のソフトウェア計算のうちの前記第 2 のソフトウェア・アプリケーションからアプリケーション・データを送り、前記通信プロファイルが、 $C(i, j)$ によって与えられた前記ソフトウェア計算のそれぞれの間で送られるソフトウェア・アプリケーション・データの量を含み、

前記ノード間通信コスト関数が、前記第 1 のソフトウェア・アプリケーションを前記 2 つまたはそれ以上のノードのうちの第 2 のノードへ再配置することによって、前記 2 つまたはそれ以上のノードのうちの第 1 のノードに割り振られる前記第 1 のソフトウェア・アプリケーションについて、 $F = \sum_{i, j} C(i, j) H(i, j)$ によって与えられ、前記ノード間通信コスト関数が、前記第 1 のノードで送られるソフトウェア・アプリケーション・データの前記量と、 $H(i, j)$ によって与えられる、前記第 1 のノードがソフトウェア・アプリケーション・データを送信した際に使用した前記第 1 のノードと残りのすべてのノードとの間の通信パス・コスト／データ単位との、積によって与えられる、請求項 18 に記載の超並列処理システム。

【請求項 22】

前記ソフトウェア・アプリケーションの前記通信プロファイルが、各ノードがそれらの間にトーラス通信ネットワークを備えて 3 D 格子に配置構成される、並列処理システム上で前記プロファイルを受け取るステップを含む、請求項 18 に記載の超並列処理システム。

【請求項 23】

前記ソフトウェア・アプリケーションの前記通信プロファイルが、Blue Gene / L スーパーコンピュータである並列処理システム上で前記プロファイルを受け取るステップを含む、請求項 18 に記載の超並列処理システム。

【請求項 24】

並列アルゴリズム操作を実行する場合、各ノードが必要に応じて計算または通信アクティビティを実行するための 1 つまたは複数の処理要素を含む、独立ネットワークによって相互接続された複数の処理ノードと、

ネットワークのノードまたはノードのサブセットの間で、高速、低待ち時間の、ポイント・ツー・ポイントおよびマルチキャスト・パケット通信を提供するために最適化された方法でノードを相互接続する、通信リンクを含む n 次元トーラス・ネットワークを備える、第 1 の独立ネットワークと、

前記ネットワークのノードまたはノードのサブセットの間で、同時グローバル操作を容易にするノード相互接続を備える、スケーラブルなグローバル・ツリー・ネットワークを含む、第 2 の前記複数の独立ネットワークと、

通信プロファイルに従って、ならびに、前記複数のソフトウェア計算のうちの前記第 1 のソフトウェア計算が、前記複数のソフトウェア計算のうちの前記第 2 のソフトウェア・アプリケーションからアプリケーション・データを送る要件に従って、独立処理ネットワークの 1 つまたは複数の組み合わせを動的に構成するための区画化手段であって、前記通信プロファイルは、前記ソフトウェア計算のそれぞれの間で送られるソフトウェア・アプリケーション・データの量を含む、区画化手段と、

前記第 1 および第 2 のネットワークの分割可能部分によって相互接続された処理ノードの構成可能サブセットを含む各独立ネットワークであって、前記構成された独立処理ネットワークのそれぞれが、ソフトウェア・アプリケーション処理性能を最適化するための同時協調処理を実行可能にするために利用される、各独立ネットワークと、
を備える、スケーラブルな超並列コンピュータであって、

前記並列計算ノードを相互接続する前記複数の独立ネットワークの組み合わせは、ノード間で転送されるソフトウェア・アプリケーション・データの合計量について、ノード間通信コスト関数を数学的に最小にすることによって協調的または単独で利用され、前記通信コストは、前記ソフトウェア・アルゴリズムが前記ソフトウェア計算間でのソフトウェア・アプリケーション・データ通信に対して前記通信ファブリックを使用する方法に依存する、スケーラブルな超並列コンピュータ。