(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第6326062号
(P6326062)
(45)発行日　平成30年5月16日(2018.5.16)　(24)登録日　平成30年4月20日(2018.4.20)

(51)Int.Cl.　F I
G06F 9/50 (2006.01)　G06F 9/46 465A
G06F 9/46 465C

請求項の数 18　(全 13 頁)

(21)出願番号　特願2015-544196(P2015-544196)
(86)(22)出願日　平成25年11月26日(2013.11.26)
(65)公表番号　特表2016-506557(P2016-506557A)
(43)公表日　平成28年3月3日(2016.3.3)
(86)国際出願番号　PCT/US2013/072094
(87)国際公開番号　WO2014/082094
(87)国際公開日　平成26年5月30日(2014.5.30)
審査請求日　平成27年10月1日(2015.10.1)
(31)優先権主張番号　61/729,930
(32)優先日　平成24年11月26日(2012.11.26)
(33)優先権主張国　米国(US)

(73)特許権者　514318459
サイクル　コンピューティング，エルエルシー
アメリカ合衆国　０６８３０　コネチカット州，グリーンウィッチ，スイート　３エフ，レイルロード　アベニュー　１５１
(74)代理人　100140109
弁理士　小野　新次郎
(74)代理人　100118902
弁理士　山本　修
(74)代理人　100106208
弁理士　宮前　徹
(74)代理人　100120112
弁理士　中西　基晴

最終頁に続く

(54)【発明の名称】異なる環境どうし間でのジョブ実行依頼のトランスペアレントなルーティング

(57)【特許請求の範囲】
【請求項１】
作業負荷を管理する方法であって、
（ａ）プロセッサによって、ルーティングするための作業負荷を受け取ること、
（ｂ）前記プロセッサによって、ローカルコンピュータクラスタ及び外部コンピュータクラスタについての連続的に取得したリアルタイムのパフォーマンス及び使用データに基づいて前記作業負荷をどこへルーティングするかを決定し、それに応じて前記作業負荷をルーティングし、その際、作業負荷に関連する出力データを、(i)作業負荷実行中にシームレスに、又は、(ii)自動的に修正された作業負荷工程として、作業負荷がルーティングされた先に転送すること、
前記ローカルコンピュータクラスタに部分的に余裕がある場合に、前記外部コンピュータクラスタへ実行依頼された作業負荷の部分集合を、前記ローカルコンピュータクラスタへ戻すこと
を備えた方法。
【請求項２】
請求項１に係る方法であって、受け取った前記作業負荷についての完了最終期限及びバッチ/非バッチ・タイプを確認すること、を更に備えた方法。
【請求項３】
請求項２に係る方法であって、前記ルーティング工程が、前記完了最終期限までに前記作業負荷を完了するのに充分な容量を有するローカルコンピュータクラスタについての前

記連続的に取得したリアルタイムのパフォーマンス及び使用データに応じて、前記ローカルコンピュータクラスタへ前記ルーティングを行うことを含む、方法。

【請求項４】

請求項２に係る方法であって、前記ルーティング工程が、実行依頼されたバッチタイプの作業負荷の第１部分を前記ローカルコンピュータクラスタへ、かつ前記バッチタイプの作業負荷の第２部分を少なくとも１つの外部コンピュータクラスタへルーティングすることを含む、方法。

【請求項５】

請求項２に係る方法であって、前記ルーティング工程が、完了最終期限が前記ローカルコンピュータクラスタの容量からみた完了時期よりも長い非バッチタイプの作業負荷については、前記外部コンピュータクラスタへルーティングすることを含む、方法。

【請求項６】

請求項４に係る方法であって、実行依頼された前記バッチタイプの作業負荷の前記第１及び前記第２部分をどこへルーティングするかを、前記プロセッサによって調整すること、を更に備えた、方法。

【請求項７】

（ａ）少なくとも１つのプロセッサ、及びコンピュータープログラムを格納した少なくとも１つの記憶部を備えた装置であって、前記コンピュータープログラム付きの少なくとも１つの記憶部は、前記少なくとも１つのプロセッサとともに、前記装置を動作させ、少なくとも、

（ｂ）ルーティングのための作業負荷を受け取り、

（ｃ）ローカルコンピュータクラスタ及び外部コンピュータクラスタについての連続的に取得したリアルタイムのパフォーマンス及び使用データに基づいて、前記作業負荷をどこにルーティングするかを決定し、それに応じて作業負荷をルーティングし、その際、作業負荷に関連する出力データを、(i)作業負荷実行中にシームレスに、又は、(ii)自動的に修正された作業負荷工程として、作業負荷がルーティングされた先に転送し、

前記ローカルコンピュータクラスタに部分的に余裕がある場合に、前記外部コンピュータクラスタへ実行依頼された作業負荷の部分集合を、前記ローカルコンピュータクラスタへ戻す

ように構成されている装置。

【請求項８】

請求項７に係る装置であって、前記装置が、受け取った前記作業負荷についての完了最終期限及びバッチ／非バッチタイプを、前記少なくとも１つのプロセッサによって確認するように、前記コンピュータープログラム付きの前記少なくとも１つの記憶部が構成されている装置。

【請求項９】

請求項８に係る装置であって、前記ルーティング動作が、前記完了最終期限までに前記作業負荷を完了するのに充分な容量を有するローカルコンピュータクラスタについての前記連続的に取得したリアルタイムのパフォーマンス及び使用データに応じて、前記ローカルコンピュータクラスタへ前記ルーティングを行うことを含む装置。

【請求項１０】

請求項８に係る装置であって、前記ルーティング動作が、実行依頼されたバッチタイプの作業負荷の第１部分を前記ローカルコンピュータクラスタへ、かつ前記バッチタイプの作業負荷の第２部分を少なくとも１つの外部コンピュータクラスタへルーティングすることを含む装置。

【請求項１１】

請求項８に係る装置であって、前記ルーティング動作が、完了最終期限が前記ローカルコンピュータクラスタの容量からみた完了時期よりも長い非バッチタイプの作業負荷については、前記外部コンピュータクラスタへルーティングすることを含む装置。

【請求項１２】

前記装置が、実行依頼された前記バッチタイプの作業負荷の前記第１及び前記第２部分をどこへルーティングするかを、前記少なくとも１つのプロセッサによって調整するように、前記コンピュータープログラム付きの少なくとも１つの記憶部が構成されている、請求項１０に係る装置。

【請求項１３】

少なくとも１つのプロセッサによって実行可能なコンピュータープログラムを格納した非一時的コンピュータ読取可能記憶媒体であって、前記コンピュータープログラムが前記少なくとも１つのプロセッサによって、

（ａ）ルーティングのための作業負荷を受け取る動作、

（ｂ）ローカルコンピュータクラスタ及び外部コンピュータクラスタについての連続的に取得したリアルタイムのパフォーマンス及び使用データに基づいて、前記作業負荷をどこへルーティングするか決定し、それに応じて前記作業負荷をルーティングし、その際、作業負荷に関連する出力データを、(i)作業負荷実行中にシームレスに、又は、(ii)自動的に修正された作業負荷工程として、作業負荷がルーティングされた先に転送する動作、

前記ローカルコンピュータクラスタに部分的に余裕がある場合に、前記外部コンピュータクラスタへ実行依頼された作業負荷の部分集合を、前記ローカルコンピュータクラスタへ戻す動作

を実行する非一時的コンピュータ読取可能記憶媒体。

【請求項１４】

請求項１３に係る非一時的コンピュータ読取可能記憶媒体であって、前記コンピュータープログラムによって、前記プロセッサが、受け取った前記作業負荷についての完了最終期限及びバッチ／非バッチ・タイプを確認する動作を、更に行う、非一時的コンピュータ読取可能記憶媒体。

【請求項１５】

請求項１４に係る非一時的コンピュータ読取可能記憶媒体であって、前記ルーティング動作が、前記完了最終期限までに前記作業負荷を完了するのに充分な容量を有する前記ローカルコンピュータクラスタについての前記連続的に取得したリアルタイムのパフォーマンス及び使用データに応じて、前記ローカルコンピュータクラスタへ前記ルーティングを行うことを含む、非一時的コンピュータ読取可能記憶媒体。

【請求項１６】

請求項１４に係る非一時的コンピュータ読取可能記憶媒体であって、前記ルーティング動作が、実行依頼されたバッチタイプの作業負荷の第１部分を前記ローカルコンピュータクラスタへ、かつ前記バッチタイプの作業負荷の第２部分を少なくとも１つの外部コンピュータクラスタへルーティングすることを含む、非一時的コンピュータ読取可能記憶媒体。

【請求項１７】

請求項１４に係る非一時的コンピュータ読取可能記憶媒体であって、前記ルーティング動作が、完了最終期限が前記ローカルコンピュータクラスタの容量からみた完了時期よりも長い非バッチタイプの作業負荷については、前記外部コンピュータクラスタへルーティングすることをさらに含む、非一時的コンピュータ読取可能記憶媒体。

【請求項１８】

請求項１６に係る非一時的コンピュータ読取可能記憶媒体であって、前記コンピュータープログラムが、実行依頼された前記バッチタイプの作業負荷の前記第１及び前記第２部分をどこへルーティングするかを、前記プロセッサに調整させるように構成されている、非一時的コンピュータ読取可能記憶媒体。

【発明の詳細な説明】

【背景技術】

【０００１】

本発明は、高パフォーマンス演算又はビッグデータ処理システムに関し、かつ負荷をビジー状態の分散コンピューティング環境から、余裕のある環境及び／又は動的に拡張可能

な環境へ自動送信する方法及びシステムに関する。
【０００２】
ジョブスケジュールシステムは、異なるコンピュータ作業負荷を巨大コンピュータ環境に分散可能である。大企業内のコンピュータ環境は、以下の特徴を有する傾向がある：
・静的サイズ
・一般的に物理的マシンによって構築
・概略均一な構成
・同じクラスタ内でしっかり接続
・他の部署のクラスタとはゆるく接続
・地理的に別の場所のクラスタとの接続は貧弱
・クラスタどうし間では共有ストレージ部へのアクセスは一般に困難
・あるクラスタは混み別のクラスタは余裕が有るとのホットスポットが出来るのが通例
【０００３】
部署のクラスタの規模及び部署の作業負荷の要求量が変化した場合、他の部署や場所にジョブを流すようにすると魅力的である。しかし、作業負荷はネットワーク及びストレージにより強く制約されやすいから、これら高パフォーマンスコンピュータ、ネットワーク、及びストレージ源の領域間に作業負荷が機能的に跨るようにするのは難しい。
【発明の概要】
【０００４】
本発明の実施形態は、高パフォーマンスコンピュータ又は「ビッグデータ」又は「マップレデュース」ソフトウェアの開発者が、企業内部及び／又は複数の情報インフラサービス（イアースIaaS）クラウド環境どうし間に跨るコンピュータ資源を切れ目無く使用するシステム及び方法を提供するものである。これは、パフォーマンス及び信頼性が不確実なゾーンの中にあって、周知のパフォーマンス特性を有するコンピュータ出力領域をなす、閉じた企業ネットワークの内部又は外部における、コンピュータ群からなる個々のクラスタを処理することによってなされる。このシステムは、データを転送し、作業負荷をあたかもそれが実在して手元で動作しているように遠隔のクラスタに移す。該システムは、バッチ・タイプの高パフォーマンスジョブのために、又は本発明の目的からして等価であるマップレデュースジョブにおける「マップ」処理のために、データを動かし、分離可能なジョブを分割して別のコンピューティング環境へ送り、完了した結果を戻す。以下に述べるすべての、実行依頼されたバッチタイプの作業負荷の一部分は、実行依頼されたマップレデュースタイプの作業負荷におけるマップ部分と等価である。このワークフローの決定動作及び実行は、開発者及びアプリケーションがまったく意識する必要のないプロセスとして行われる。開発者又はアプリケーションが、データ及びジョブを移動させるという煩雑さに見舞われることはない。開発者は、そのアプリケーションを一つの領域で作動させるだけでよく、それを本発明によって自動的に他の領域へ移すという複雑な作業が行われる。
【０００５】
現行のアプローチでは、スケジューリング環境が同じで場所的には分離されたコンピュータ資源を用いて処理しており、かつローカル環境とリモート環境を等しく扱っている。現行アプローチには、本発明がその問題に対する優れた解決手段となるような２つのファクターがある。ファクター１： 低レイテンシ及び高周波数帯域との仮定の下で実行する、不確かなＷＡＮリンク間の処理は、たいがい失敗する。ファクター２： 全世界的に共有されたストレージ装置はパフォーマンス特性が一般的に遅いために、作業負荷内のジョブの進展速度の不足によって失敗に終わる。本発明によれば、これらの落し穴の両方とも避けることによって、確実にジョブがシステム環境間でより早く流れることができ、これらのジョブを実行することで、開発者が１つの地域にある内部クラスタ上で動いていると思うようなスピードと信頼性のある処理を続行することができる。本発明を使用した場合に生じる唯一の追加負担は、データをローカルからリモートのクラスタへ移動させてジョブ実行を補助してもらうことであり、このリモート領域で演算が完了後、データ結果が元

の領域へ返送される。

【０００６】

本発明の実施形態では、詳細なパフォーマンス及び使用データをクラスタから連続的に集め、このデータを用いて、以下のパラメータに基づき、ジョブルーティングに関する決定をする：

セキュリティ、パフォーマンス、及び規制遵守上の問題から、内部環境でジョブをやらせたいというユーザー又は自動ワークロードの要望;

アマゾン・ウェブサービス（AWS ）のような、外部のダイナミックなコンピューティング環境における耐ランニングコスト;

リモートクラスタの演算に要求されるデータ群における、すでに同期している部分又は区画の存在；

すべてのコンピューター分野にわたるすべてのコンピュータ資源の利用現況;

往復伝送される必要があるデータの量に関する情報で結び付き得る、データ転送のためのクラスタ間で利用可能な周波数帯域

【０００７】

最終的なコンピュータジョブルーティングを決定するのに使われる仲介アルゴリズムは、種々の動特性に応じて設定変更可能である。本発明の実施形態では、実行依頼されているジョブ、及びジョブを実行出来たであろうポテンシャルクラスタに関して有するすべての知識を適用することによって、アプリケーション、開発者、またはエンドユーザの介在無しで、作業負荷のための補助スケジューリングを実行し、適切な領域にジョブを自動的にルーティングする。補助スケジューリングの決定は、実行依頼時又はその後定期的に行われる。そして、決定後直ちにジョブを、内部又は外部の、静的又は動的な特定領域のマシン上での仕事を受け持つスケジューラに送出する。

【０００８】

本発明の実施形態では、スケジュールに入れられた複数クラスタ及び複数ジョブが、互いに完全に独立して動くのを許容し、常に必要な十分大きな安定性を与える。機能性環境を保つために低レイテンシー通信が求められることがない。本発明の実施形態では、これらクラスタが、当該アーキテクチャの範囲外でも完全に機能するのを許容し、完全にローカルの作業負荷と、本発明の補助スケジューリング・アルゴリズムを介して他のクラスタから入って来るジョブを混合する。これによって、セキュリティに関して、旧来の相互運用性及び柔軟性が確保される：　スケジュールに入れられたジョブが本発明によって実行依頼された場所へ転送されるのが望ましくない場合、エンドユーザは、簡単にジョブを通常通りローカル領域で実行されるよう指示できる。 本発明によれば、その補助スケジューリングアルゴリズムによって作業負荷をより多く実行依頼することで、コンピュータ資源の広く分散された共用部の使用率をも高め、全体的な利用度が多くなる。

【図面の簡単な説明】

【０００９】

図１は、エンドユーザ、又はサイクル・サーバーのようなジョブ実行依頼ポータルによるジョブ実行依頼プロセスを示すブロック図である。

【００１０】

図２は、SubmitOnceのルーティング・エンジン及び決定を操作可能な変数を示すブロック図である。

【００１１】

図３は、リモートへの、データ転送及びスケジューラ・インタラクションを含むジョブ実行依頼のワークフローを示すブロック図である。

【００１２】

図４は、リモートのクラスタへの実行依頼時、部分的に余裕がある内部クラスタへ作業を戻す処理を示すブロック図である。

【００１３】

図５は、SubmitOnceの作業負荷のルーティングを示すフローチャートである。

【００１４】
図６は、SubmitOnceアプリケーションの作業負荷のルーティングの概念構造を示すフローチャートである。
【発明の詳細な説明】
【００１５】
以下、本発明に係る処理及びシステムの実施形態を説明する。なお、いかなる場合も本開示が以下の実施形態に限定されるものでないことを注記する。下記実施形態は、本発明をここで実行するための非限定的な例示に過ぎない。
【００１６】
実施形態では、スケジューラ・ベースのジョブの実行依頼の動作を正確にまねたクラウド内において作業負荷を実行依頼するためのシステムを備えている。システムは、ジョブ・スケジューラの操作に関する知識を使って、実行依頼されている作業負荷についてできるだけ多くのメタデータを引き出す。
【００１７】
他の実施形態は、ジョブルーティングを集中的に決めるための変動可能な数の環境パラメータを決めるスケジューラ監視ソリューションと結合されたジョブルーティング機構を備えている。実施形態では、ジョブルーティングの決定がなされたとき、フレームワークが自動リモートアクセスを使うことで、データの継続伝送、リモートでのコマンド実行、及びジョブ監視を実行する。
【００１８】
実施形態では、異なるスケジューラ、又はマップレデュース、又は「ビッグデータ」のワークフレームを用いて、一組のジョブが、複数の異なる環境上で動くことができ、完了時には、気付かれることなく１つの統合されたエリアとなるアーキテクチャを備えている。実施形態では、クラウド・コンピューティング環境内で作業負荷を実行依頼するためのシステムを更に含む。このシステムは、ジョブ・スケジューラの操作に関する知識を使って、スケジューラ・ベースのジョブの実行依頼の動作を正確にまねる。このシステムは、実行依頼されている作業負荷に対応する有用なメタデータの少なくとも一部を引き出す。
【００１９】
この発明の実施形態は、複数のクラスタ環境内で、幾つかの又はすべての実行依頼があったことを周期的に確認する手段を使って、ジョブルーティングをリアルタイムに決定するための変動可能な個数の環境パラメータを決めるスケジューラ監視ソリューションと結合されたジョブルーティング機構を備えている。さらに他の実施形態は、ジョブルーティングの決定がなされたとき、データの継続伝送、リモートでのコマンド実行、及びジョブ監視を実行するための自動リモートアクセスを使うためのフレームワークを含む。更なる実施形態では、一組のジョブが、複数の異なる環境上で動くことができ、完了時には、気付かれることなく１つの統合されたエリアとなるアーキテクチャを備えている。ある実施形態では、分散コンピューティング環境どうし間で作業負荷を指定する方法を含む。前記方法は、複数のコンピュータクラスタの各々からパフォーマンス及び使用データを連続的に得ることを含み、前記複数のコンピュータクラスタの第１サブセットは第１領域に在り、前記複数のコンピュータクラスタの第２サブセットは第２領域に在り、各領域は、公知のパフォーマンス特性、パフォーマンスゾーン、及び信頼性ゾーンを有している。前記方法は、分散コンピューティング環境へルーティングするためのジョブを受け取ることを更に含む。前記方法は、得られたパフォーマンス及び使用データに応じて、ジョブを所定のコンピュータクラスタへルーティングすることを更に含み、前記領域は、前記所定のコンピュータクラスタを包含する。
【００２０】
更なる実施形態は、分散コンピューティング環境どうし間で作業負荷を指定する方法を含む。前記方法は、電子的に実行依頼された作業負荷についての完了最終期限及びバッチ/非バッチ・タイプを確認することを含む。前記方法は、次のうち少なくとも１つによって実行依頼された作業負荷を処理することを更に含む;（ｉ）完了最終期限までに処理を

完了するのに充分な容量を有するローカルコンピュータクラスタに応じて、実行依頼された作業負荷をローカルコンピュータクラスタへルーティングすること；（ｉｉ）実行依頼されたバッチ・タイプの作業負荷の第１部分、又は等価的にはマップレデュースタイプの実行依頼された作業負荷のマップ部の部分を有効容量のローカルコンピュータクラスタへルーティングすること、及び、実行依頼されたバッチ・タイプの作業負荷の第２部分、又は等価的にはマップレデュースタイプの実行依頼された作業負荷のマップ部の第２部分を少なくとも１つのリモートコンピュータクラスタへルーティングすること；　及び（ｉｉｉ）ローカルコンピュータクラスタの容量からみた完了時期よりも長い完了最終期限を有する非バッチタイプの実行依頼された作業負荷をリモートコンピュータクラスタへルーティングすること。明確さのために、「マップレデュース」作業負荷は、バッチタイプの作業負荷であり、マップ部分及び個々のレデュースジョブを構成要素とする。同様に、いわゆる「はっきりしない」又は「はっきりとした」並列な作業負荷は、要するに、どんな作業負荷でも、多くの独立した演算を含んでいれば、たとえ個々が厳密に並列でなくても、バッチタイプの作業負荷である。

【００２１】

他の実施形態は、分散コンピューティング環境どうし間で作業負荷を指示する方法を含む。前記方法は、アプリケーション作業負荷ルータで作業負荷の実行依頼を受け取ることを含む。前記方法は、以下のうち少なくとも１つの工程によって作業負荷の実行依頼をルーティングすることを更に含む。（ｉ）実行依頼された作業負荷の第１部分、又は等価的にはマップレデュースタイプの実行依頼された作業負荷のマップ部の部分を、ローカルコンピュータクラスタへルーティングし、実行依頼された作業負荷の前記第１部分、又は等価的にはマップレデュースタイプの実行依頼された作業負荷のマップ部の前記第２部分は、ローカルコンピュータクラスタの有効満了パラメータ内にあり、実行依頼された作業負荷の第２部分をリモート（非ローカル）のコンピュータクラスタへルーティングすること、及び（ｉｉ）ローカルコンピュータクラスタが無い場合、リモートコンピュータクラスタへ作業負荷の実行依頼をルーティングすること。

【００２２】

本実施方法は、さらにワークフロー自動修正工程を含むことができる。この工程は、１又はそれ以上のリモート（非ローカル）コンピュータへルーティングされた作業負荷の実行依頼に関連するデータのための、出力データ転送とその後の入力データ転送を含む。

【００２３】

前記ジョブ実行依頼コマンド-ライン／ＡＰＩ／ウェブページ／ウェブサービスは、ブロック１０６の環境情報、ブロック１０８のルーティング／実行依頼の試行から導き出された変数、及びブロック１０４のユーザー入力メタデータに蓄積する。ブロック１１２に在るサーバー環境が、利用できないか、応答が長くかかり過ぎる場合には、実行可能なジョブ実行依頼は、ブロック１１０の通り、常にローカルで前記実行依頼を実行する。このように、ジョブ実行依頼は、常に所定の時間間隔内で発生する。

【００２４】

タスクが複数のクラスタ上で動作する場合には、出力は、そのシステムで定義されたクラスタ名によって示される接頭辞を用いたワークフローによって区別される。ジョブ実行依頼の出力は、生成場所のスケジューラ・コマンドによって作られた出力と同一でなければならない。このように、ユーザー、作業負荷、又はＡＰＩは、当該システムを活用することで、意識することなく当該システムと互いに連携することができる。前記出力は、ブロック１１４に在る一般的な処理中にサーバー環境によって返信される。

【００２５】

図２は、本発明の他の実施形態を示す。図２においては、SubmitOnceのルーティング・エンジン、及び決定手順を操作可能な変数を表すブロック図が示されている。

【００２６】

図２において、本システムの処理及び構成要素は、ブロック２０２で受け取られるジョブルーティング及び実行依頼動作を構成して管理してモニターするのに使用されるサーバ

ー・アーキテクチャ内のＧＵＩダッシュボードを含む。当然に、前記システムを特定の方針及び予想に適合させために管理者によって使用することのできる初期環境設定の実行依頼をも含む。ブロック２０４では、利用可能な共有記憶場所、広告アプリケーション、現在の容量、予約超過閾値、及び動的実行ノード能力等（ただしこれらに限定されない）のスケジューリング環境のために定義可能なメタデータを取り込むことができる。前記メタデータは、環境設定中、及び／又はブロック２１２に示すようにモニタリング環境によってリアルタイムに引き出されている時、入力可能である。ブロック２０６、２０８，及び２１０に示すように、ジョブをローカルに送信することも、それが最も好都合であれば、選択可能である。ブロック２１２のように、ローカルルーティングがすぐに明らかにならない場合に、完全な仲介ルーチンに入る。ブロック２１４において、どんなクラスタ単位が選択されたとしても、最終的には実際の実行依頼プロセスのためにルーティングが決定される。

【００２７】

図３は、本発明における他の実施形態を示す。図３は、データ転送及びスケジューラ・インタラクションを含むリモートジョブの実行依頼のワークフローを表したブロック図を示す。

【００２８】

図３の実施形態におけるプロセス及び構成要素は、中央サーバがブロック３０２に在る１又はそれ以上のクラスタ単位と通信するハブ・アンド・スポークデザインを含む。実行依頼時のカギとなる決定事項は、ルーティングがローカルであるかリモートであるかである。なぜなら、それによって、データをブロック３０４とブロック３０８のどちらへ移動させる必要があるかが決まるからである。クラスタ単位は、コンピュータ資源における静的に割当られたか又は動的に割当られたリストで、マシンの内部クラスタとマシンの外部クラスタの両方を表し得る点に留意する必要がある。図３は、ブロック（３１０，３１６）及び（３１２，３１４）において用いられているようにスケジュールベース又はオンデマンドベースで、内部に潜在するデータ及び外部に潜在するデータの両方を転送可能なチケット・ベースのデータ転送メカニズムを必要とする。このプロセスでも、ブロック（３０６，３１８）のステップのためのコマンド実行のために、中央サーバとリモートノード間の安全な信頼できる通信を必要とする。また、委託されたジョブ実行依頼の前、指示中、又は指示後の何らかの潜在的不具合に対し適切にエラー処理する必要はある。エラーに遭遇したときは、前記システムは、ブロック３０６のように、フェイルセーフとして、ローカルでの実行を指示すべきある。

【００２９】

図４は、本発明のさらに他の実施形態を示す。図４のブロック図は、リモートクラスタへ実行依頼するときは、部分的に余裕がある内部クラスタへ作業を戻すプロセスを表している。

【００３０】

この実施形態のプロセス及び構成要素は、ジョブ実行依頼が特定のクラスタ単位に委託されるとき始まり、更なるロード・バランシングの機会がある。大半の作業負荷はリモートのクラスタ単位に指定されるが、作業負荷の部分集合を、すぐに利用できるローカル資源で動作させるために切り出してもよい。そうすると、ブロック４０２のように全体のランタイムが減少する。この動作部分は、以下に該当する場合に採用される：（１）実行依頼がきつく連結された並列ジョブではない　（２）実行依頼が、一列のジョブである　（３）タスク列の分割作業は、システム内で行うことが出来る。　前記システムは、利用可能な実行スロット数を数え、実行中のジョブを数え、ブロック４０４で利用可能なスロットを計算する。ブロック４０６では、ローカルのクラスタが先ず一杯になるように、ジョブ列が分割され、ジョブの残りは、ブロック４０８において、選択されたリモートのクラスタへ実行依頼される。これら二つ以上の実行依頼によって、ブロック４０６及びブロック４０８の処理が進められる、前述したように、ワークフローが進行する。

【００３１】

図５は、本発明の他の実施形態を示す。図５は、SubmitOnceのワークフロールーティングのフローチャートを表す。ブロック５０２において、ユーザー／アプリケーションがジョブを実行依頼することで前記処理が開始する。ブロック５０４において、ジョブが「完了」最終期限を有するか判定する処理へ続く。「完了」最終期限が無ければ、ブロック５０６において、クラスタの空きは十分かどうかを判定する処理を行なう。イエスであれば、ブロック５１０において、ローカルクラスタへルーティングする処理をする。ノーであれば、ブロック５１２において、バッチジョブかを判定する。イエスであれば、ブロック５１４において、ローカルの入り口を塞ぎ、外部へルーティングする。ノーであれば、ブロック５１６において、単に外部へルーティングする。一方、「完了」最終期限が有る場合、ブロック５０８において、ローカルへ行く時間は十分かどうかを判定する。イエスであれば、ブロック５１０において、ローカルクラスタへルーティングされる。ノーであれば、ブロック５１２において、バッチジョブかどうかを判定し、その後、前述した処理へ続く。

【００３２】

図６は、本発明の一実施形態を示す。図６は、SubmitOnceアプリケーションの作業負荷をルーティングするアーキテクチャを図示したものである。前記ルーティングアーキテクチャは、ブロック６０２において、ユーザー又はアプリケーションによるジョブの実行依頼から始まる。ブロック６０４において、アプリケーション・ワークロード・ルータがジョブを受け取る。ローカル環境のクラスタが無ければ、ブロック６０８において、ジョブは、内部／外部のクラウドへ送られる。ローカルへのルーティングが可能であれば、ジョブがブロック６０６へ送られる。一方、必要ならば、ブロック６０８において、ジョブは、ローカルのクラスタから必要に応じてクラウドへ拡張できる。

【００３３】

本発明の開示を実行する実施方法は、作業負荷を管理する方法を含む。前記方法は、プロセッサが、ルーティングのための作業負荷を受け取ること；　並びに前記プロセッサが、ローカルコンピュータクラスタ及び外部コンピュータクラスタから連続的に取得したリアルタイムのパフォーマンス及び使用データに基づいて作業負荷をどこへルーティングするかを決定し、それに応じて前記作業負荷をルーティングすることを備えている。前記方法は、受け取った作業負荷に関する完了最終期限、及びバッチ／非バッチ・タイプを確認することを更に含む。

【００３４】

前記実施方法は、前記ルーティング工程が、完了最終期限までに作業負荷を完了するのに十分な容量を有するローカルコンピュータクラスタについての前記連続的に取得したリアルタイムのパフォーマンス及び使用データに応じて、前記ローカルコンピュータクラスタへルーティングすることを、更に含む。前記方法は、前記ルーティング工程が、実行依頼されたバッチタイプの作業負荷の第１部分、又は等価的にはマップレデュースタイプの実行依頼された作業負荷のマップ部をローカルコンピュータクラスタへ、　かつバッチタイプの実行依頼作業負荷の第２部分、又は等価的にはマップレデュースタイプの実行依頼された作業負荷のマップの第２部分を、少なくとも１つの外部コンピュータクラスタへルーティングすること、をも含む。

【００３５】

前記実施方法は、前記ルーティング工程が、完了最終期限がローカルコンピュータクラスタの容量からみた完了時期よりも長い非バッチタイプの作業負荷については、外部のコンピュータ作業負荷へ送信することをも含む。前記方法は、実行依頼されたバッチタイプの作業負荷の前記第１及び第２部分をどこへルーティングするかを、前記プロセッサによって調整すること、を更に含むことができる。

【００３６】

前記実施方法は、結果的に、非一時的コンピュータ読取可能記憶媒体、例えば非一時的コンピュータ読取可能メモリに格納されたコンピュータープログラムを実行することであってもよく、電子装置の構成要素によって前記電子装置を動作させるある特別な方法を実

行することであってもよい。前記実施方法は、コンピュータープログラム付きの１つの記憶部、及び１つのプロセッサを含む装置によって実行することであってもよい。前記コンピュータープログラム付きの前記記憶部は、前記装置が前記プロセッサによって前記実施方法を実行するように構成されている。

【００３７】

まとめると、本発明の種々の実施形態は、プロセッサと記憶部を含むコンピュータのような種々の電子装置によって実行できる。

【００３８】

大規模コンピュータ環境のユーザは、多くの別のコンピュータ環境を利用する能力を必要とする。その根っ子にあるスケジューラ、サーバ、ネットワークの構成については完全に理解する必要はない。

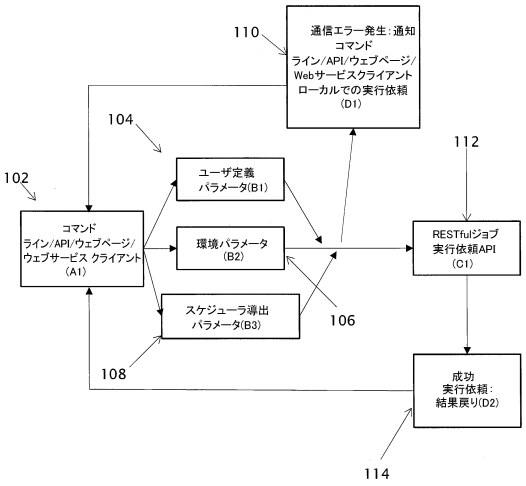
【００３９】

本発明開示手段は、ローカル及び外部のクラスタへジョブを自動送信している間、エンドユーザが典型的なジョブを実行依頼できるインターフェースを提供することを含む。これによって、複雑な構成及びプロセスによる負担をかけることなく、エンドユーザが出来る事が増える。アプリケーション・ワークロード・ルータ内の自動化によって、エンドユーザに過剰な複雑さを隠したまま、通常は１つの独立したクラスタに束縛されるはずの可能範囲を増やせる。

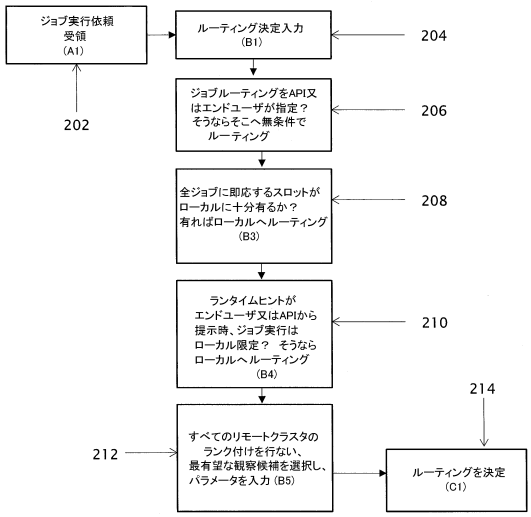
【００４０】

ひとたびこの自動ジョブルーティング環境が完全に構成されたときは、エンドユーザは、適切なルーティングのための自らの作業負荷をすべて記述する方法を必要とする。この記述の大部分は、スケジューリングレイヤを使ってなされる。エンドユーザにとって最も重要なパラメータは、実行信頼性を除くと、すべての作業負荷のための経過時間である。本発明開示内容によって、エンドユーザから更に２つの重要な情報の提供を受けられるジョブルーティング環境を実現できる。１つは、個々のタスクの平均実行時間である。もう１つは、作業負荷の全体的な所望実行時間である。この情報は、タスクの数、動的ＶＭノードのスピンアップ時間、データ転送時間、及びその作業負荷が全くのバッチなのか否か等の、既知のパラメータに沿って考慮される。この結果、ジョブを、複数のクラスタにわたるように分割でき、要求が満たされる間、内部クラスタを最大使用できる。

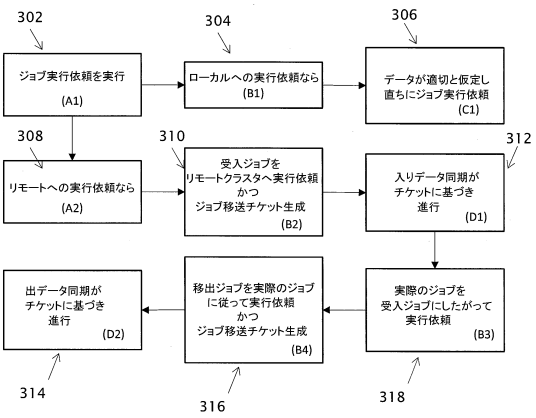
【図１】


110
通信エラー発生：通知
コマンド
ライン/API/ウェブページ/
Webサービスクライアント
ローカルでの実行依頼
(D1)
104
112
102
ユーザ定義
パラメータ(B1)
コマンド
ライン/API/ウェブページ/
ウェブサービス クライアント
(A1)
環境パラメータ
(B2)
RESTfulジョブ
実行依頼API
(C1)
スケジューラ導出
パラメータ(B3)
106
108
成功
実行依頼：
結果戻り(D2)
114

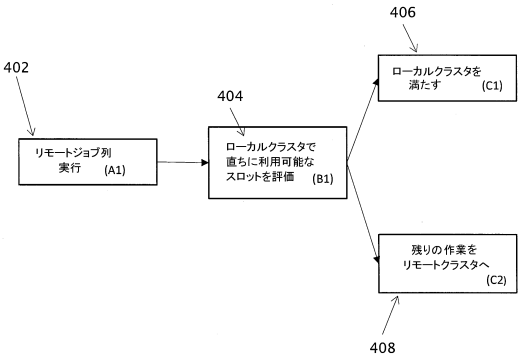
【図２】


ジョブ実行依頼
受領
(A1)
ルーティング決定入力
(B1)
204
202
ジョブルーティングをAPI又
はエンドユーザが指定？
そうならそこへ無条件で
ルーティング
206
全ジョブに即応するスロットが
ローカルに十分有るか？
有ればローカルへルーティング
(B3)
208
ランタイムヒントが
エンドユーザ又はAPIから
提示時、ジョブ実行は
ローカル限定？ そうなら
ローカルへ ルーティング
(B4)
210
214
すべてのリモートクラスタの
ランク付けを行ない、
最有望な観察候補を選択し、
パラメータを入力 (B5)
212
ルーティングを決定
(C1)

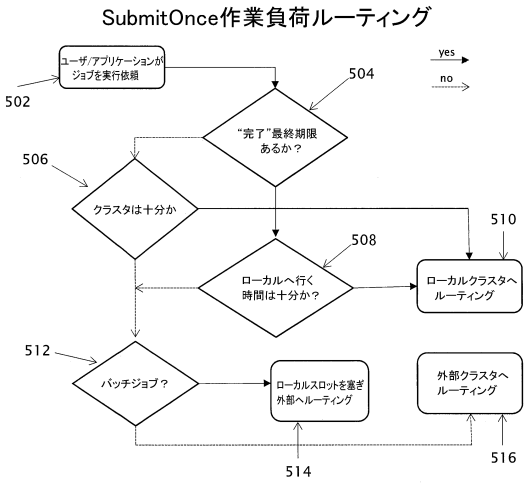
【図３】


302
304
306
ジョブ実行依頼を実行
(A1)
ローカルへの実行依頼なら
(B1)
データが適切と仮定し
直ちにジョブ実行依頼
(C1)
308
310
312
リモートへの実行依頼なら
(A2)
受入ジョブを
リモートクラスタへ実行依頼
かつ
ジョブ移送チケット生成
(B2)
入りデータ同期が
チケットに基づき
進行
(D1)
出データ同期が
チケットに基づき
進行
(D2)
移出ジョブを実際のジョブ
に従って実行依頼
かつ
ジョブ移送チケット生成
(B4)
実際のジョブを
受入ジョブにしたがって
実行依頼
(B3)
314
316
318

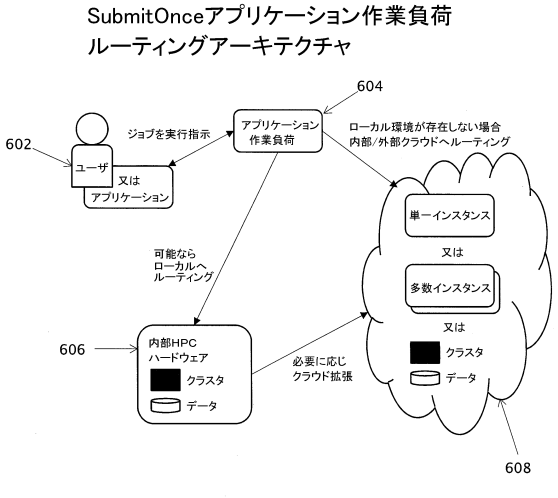
【図４】


406
402
ローカルクラスタを
満たす (C1)
404
リモートジョブ列
実行 (A1)
ローカルクラスタで
直ちに利用可能な
スロットを評価 (B1)
残りの作業を
リモートクラスタへ
(C2)
408

【図５】


SubmitOnce作業負荷ルーティング
ユーザ/アプリケーションが
ジョブを実行依頼
502
504
yes
no
"完了"最終期限
あるか？
506
クラスタは十分か
510
508
ローカルへ行く
時間は十分か？
ローカルクラスタへ
ルーティング
512
バッチジョブ？
ローカルスロットを塞ぎ
外部へルーティング
外部クラスタへ
ルーティング
514
516

【図６】


SubmitOnceアプリケーション作業負荷
ルーティングアーキテクチャ
604
602
ユーザ
又は
アプリケーション
ジョブを実行指示
アプリケーション
作業負荷
ローカル環境が存在しない場合
内部/外部クラウドへルーティング
単一インスタンス
又は
多数インスタンス
又は
クラスタ
データ
可能なら
ローカルへ
ルーティング
606
内部HPC
ハードウェア
クラスタ
データ
必要に応じ
クラウド拡張
608

フロントページの続き

(74)代理人 100162846
弁理士 大牧 綾子
(72)発明者 ストウ, ジェイソン, エー.
アメリカ合衆国 ０６８３０ コネティカット州, グリニッジ, １０７番, グレン ストリート ５
(72)発明者 カクゾレック, アンドリュー
アメリカ合衆国 ４６０３２ インディアナ州, カーメル, グリーブ ストリート ２３４７

審査官 原 忠

(56)参考文献 米国特許出願公開第２０１２／００５４７７１（ＵＳ, Ａ１）
特表２０１２－５２３０３８（ＪＰ, Ａ）
特開２００２－２５９３５３（ＪＰ, Ａ）
特開２００２－３４２０９８（ＪＰ, Ａ）
米国特許出願公開第２００８／００５２７１２（ＵＳ, Ａ１）
米国特許出願公開第２００４／０１９９９１８（ＵＳ, Ａ１）

(58)調査した分野(Int.Cl., ＤＢ名)
Ｇ０６Ｆ ９／４６－９／５４