

(19) 日本国特許庁(JP)

(12) 特 許 公 報(B2)

(11) 特許番号

特許第3962394号
(P3962394)

(45) 発行日 平成19年8月22日(2007.8.22)

(24) 登録日 平成19年5月25日(2007.5.25)

(51) Int. Cl. F I
G06F 15/173 (2006.01) G O 6 F 15/173 6 4 O M
G06F 11/20 (2006.01) G O 6 F 11/20 3 1 O K

請求項の数 8 (全 17 頁)

<p>(21) 出願番号 特願2004-131894 (P2004-131894) (22) 出願日 平成16年4月27日 (2004.4.27) (65) 公開番号 特開2004-326809 (P2004-326809A) (43) 公開日 平成16年11月18日 (2004.11.18) 審査請求日 平成16年4月27日 (2004.4.27) (31) 優先権主張番号 10/424278 (32) 優先日 平成15年4月28日 (2003.4.28) (33) 優先権主張国 米国 (US)</p>	<p>(73) 特許権者 390009531 インターナショナル・ビジネス・マシー ズ・コーポレーション INTERNATIONAL BUSIN ESS MASCHINES CORPO RATION アメリカ合衆国10504 ニューヨーク 州 アーモンク ニュー オーチャード ロード (74) 代理人 100086243 弁理士 坂口 博 (74) 代理人 100091568 弁理士 市位 嘉宏 (74) 代理人 100108501 弁理士 上野 剛史</p>
--	--

最終頁に続く

(54) 【発明の名称】 ホットプラグ可能な問題のあるコンポーネントの動的検出および問題のあるコンポーネントからのシステムリソースの再割り当て

(57) 【特許請求の範囲】

【請求項1】

ホットプラグ処理をサポートするデータ処理システムであって、
 第1のプロセッサおよび第1のメモリを含む第1の動作コンポーネントセットと、
 前記第1の動作コンポーネントセットを相互接続する接続機構であって、前記接続機構
 は、前記第1の動作コンポーネントセットの処理に介入することなく、ホットプラグコネ
 クタを介して動作コンポーネントの追加および除去が可能であり、前記接続機構は、前記
 動作コンポーネントの追加および除去によって生じる前記データ処理システムの動的な変
 更に対応するための構成設定論理機構を含み、前記構成設定論理機構は、前記接続機構の
 ルーティング及び通信動作を制御するための第1の構成設定及び第2の構成設定を備えて
 いる、前記接続機構と、

前記ホットプラグコネクタを介して前記第1の動作コンポーネントセットに物理的に結
 合された第2の動作コンポーネントセットと、

前記第1の動作コンポーネントセットおよび前記第2の動作コンポーネントセットの双
 方のシステムチェックを自動的に実行するための手段であって、前記システムチェックに
 よって、前記第1の動作コンポーネントセットおよび前記第2の動作コンポーネントセッ
 トのいずれかにおいて、問題のある動作コンポーネントが識別される、前記実行するた
 めの手段と、

前記第2の動作コンポーネントセット内で前記問題のある動作コンポーネントが検出さ
 れた場合、少なくとも前記問題のある動作コンポーネントの除去を動的に開始するための

サービス手段を含むサービス要素と、

を含み、前記データ処理システムが、前記第1の動作コンポーネントセットおよび前記ホットプラグコネクタの1つを介して接続された第2の動作コンポーネントセットの双方を含む場合、前記構成設定論理機構は第2の構成設定を選択し、前記第2の動作コンポーネントセットが前記問題のある動作コンポーネントとして識別された場合、前記構成設定論理機構は前記第1の動作コンポーネントセットをサポートする第1の構成設定を選択する、前記データ処理システム。

【請求項2】

前記第2の動作コンポーネントセットの前記除去を示す出力を発生するための手段を更に含む、請求項1に記載のデータ処理システム。

10

【請求項3】

前記出力は、前記問題のある動作コンポーネントの問題の種類および識別の具体的な指示を含む、請求項2に記載のデータ処理システム。

【請求項4】

前記データ処理システムの実行時に、第3の動作コンポーネントセットを前記データ処理システムにそれぞれ追加および除去して、前記データ処理システムの拡張および縮小を可能にする論理機構を更に含む、前記第3の動作コンポーネントセットは、前記ホットプラグコネクタを介して接続され、前記第1の動作コンポーネントセットが動作している間に、前記第1の動作コンポーネントセットの現在の性能を妨害することなく、前記第3の動作コンポーネントが追加および除去される、請求項1に記載のデータ処理システム。

20

【請求項5】

前記第1の動作コンポーネントセットの現在の動作に介入することなく、前記ホットプラグコネクタを介して前記第2の動作コンポーネントセットの電気的および論理的接続を実行することによって、前記データ処理システムの実行時の拡張を可能として、前記第2の動作コンポーネントセットを含ませるための手段を更に含む、請求項1に記載のデータ処理システム。

【請求項6】

前記システムチェックを自動的に実行するための手段は、前記第2の動作コンポーネントセットの前記システムチェックを自動的に開始および完了させる、請求項1に記載のデータ処理システム。

30

【請求項7】

前記構成設定論理機構は検出論理機構を含み、前記構成設定論理機構は、前記第1の構成設定及び第2の構成設定を選択するためのラッチと、前記ラッチ内の値によって選択される、特定のルーティングおよび動作プロトコルを実施するための複数の構成レジスタとを含み、前記問題のある動作コンポーネントが前記ホットプラグコネクタから除去されていると検出された場合はいつでも、前記検出論理機構によって前記ラッチ内の値が設定される、請求項1に記載のデータ処理システム。

【請求項8】

前記ホットプラグコネクタへの前記第2の動作コンポーネントセットの結合および前記ホットプラグコネクタからの前記第2の動作コンポーネントセットの除去のための一連のホットプラグ接続ポートを提供する接続バックプレーンを更に含む、請求項1に記載のデータ処理システム。

40

【発明の詳細な説明】

【技術分野】

【0001】

本発明は、一般にデータ処理システムに関し、具体的には、データ処理システムのホットプラグ可能なコンポーネントに関する。更に具体的には、本発明は、データ処理システムからホットプラグ可能な問題のあるコンポーネントを非介入的かつ自動的に検出しホット除去することを可能とする方法、システム、およびデータ処理システム構成に関する。

【背景技術】

50

【 0 0 0 2 】

個人用および商用の双方において、より優れた、リソースの豊富なデータ処理システムが要望されていることによって、業界では、顧客利用のために設計されているシステムの改善が続いている。一般に、商用および個人用の双方において、プロセッサの高速化、上位レベルキャッシュの増大、読み取り専用メモリ（ROM）の大容量化、ランダムアクセスメモリ（RAM）スペースの増大等に焦点を当てて改善が行われている。

【 0 0 0 3 】

顧客の要望を満たすためには、顧客が、ハードウェアリソースを含めて、追加のリソースによって既存のシステムを向上または拡張可能であることが必要である。例えば、CD-ROMを搭載したコンピュータを有する顧客は、後に、DVDドライブに「アップグレード」したり、DVDドライブを追加したりしようとする場合がある。あるいは、顧客は、64KバイトメモリのPentium1プロセッサを有するシステムを購入し、後に、チップをPentium3チップにアップグレード/変更して、メモリ容量を256Kバイトに増大させようとする場合がある。

10

【 0 0 0 4 】

現在のデータ処理システムは、わずかな努力でシステムのハードウェア構成にこれらの基本的な変更を加えられるように設計されている。当業者には既知であるように、プロセッサやメモリをアップグレードするには、コンピュータの外箱を外して、マザーボード上で利用可能なプロセッサデッキまたはメモリスロットに新しいチップまたはメモリスティックを「留める」ことが必要である。同様に、DVDプレーヤは、マザーボード上の内部入出力（I/O）ポートの1つに接続することができる。システムによっては、外部DVDドライブを、シリアルポートまたはUSBポートの1つに接続することも可能である。

20

【 0 0 0 5 】

更に、特に商用システムでは、処理リソースを増やすこと、すなわち、現在のプロセッサをもっと高速なものと置換するのではなく、同じ処理システムを更にいくつか購入し、それらを共にリンクさせて全体の処理能力を高くすることを含む改善が行われている。最新の商用システムは、単一のシステムにおいて多数のプロセッサを有するように設計されている。多くの商用システムは、分散型またはネットワーク化システムであり、多数の個別のシステムが互いに相互接続され、処理タスク/作業負荷を共有している。しかしながら、これらの「大規模」商用システムであっても、顧客の要望が変化すれば、頻繁にアップグレードまたは拡張を行わなければならない。

30

【 0 0 0 6 】

とりわけ、システムをアップグレードまたは変更する場合、特に内部に追加したコンポーネントについては、インストールを完了する前にシステムの電源を切る必要があることが多い。しかしながら、外部に接続したI/Oコンポーネントでは、システムを起動し実行している間に、単にコンポーネントをプラグインすれば良い場合がある。コンポーネントを追加する（内部追加または外部追加）ために用いる方法には無関係に、システムは、ファブリックと呼ばれる接続機構に関連した論理を含み、これによって、追加のハードウェアが追加されたこと、または単にシステム構成の変更が行われたことを認識する。次いで、この論理は、ユーザにプロンプトを出力して（または自動的に）、システム構成のアップグレードを開始させ、必要な場合には、必要なドライバをロードして新しいハードウェアのインストールを完了することができる。とりわけ、システム構成のアップグレードは、システムからコンポーネントを除去する場合にも必要である。

40

【 0 0 0 7 】

新しいI/Oハードウェアをデータ処理システムによってほぼ即座に利用可能とするプロセスは、当技術分野では一般に「プラグアンドプレイ」と呼ばれている。この現システムの機能によって、いったんコンポーネントが認識され、適切な動作のために必要なドライバ等がインストールされると、システムは自動的に、システムによるコンポーネントの利用を可能とする。

【 0 0 0 8 】

50

図1は、商用SMPを示す。これは、第1プロセッサ101、第2プロセッサ102、メモリ104、および入出力(I/O)デバイス106を備え、これらは全て相互接続機構108によって接続されている。相互接続機構108は、ワイヤおよび制御論理を含み、これによって、コンポーネント間の通信をルーティングすると共に、ハードウェア構成における変更に対するMP100の応答を制御する。このため、新しいハードウェアコンポーネントは、相互接続機構108を介して既存のコンポーネントにも(直接的または間接的に)接続される。

【0009】

図1に例示するように、MP100は、点線で示される論理パーティション110(すなわちソフトウェアによって実施されるパーティション)を備え、これが、第2プロセッサ102から第1プロセッサ101を論理的に分けている。MP100内で論理パーティション110を利用することによって、第1プロセッサ101および第2プロセッサ102は、互いに独立して動作することができる。また、論理パーティション110は、他のプロセッサの動作問題およびダウンタイムから各プロセッサを実質的に遮断する。

【0010】

SMP100等の商用システムは、上述のように、顧客の要望を満たすように拡張することができる。更に、商用システムに対する変更は、コンポーネントが故障して、システムがフルに動作することができなくなったり、最悪の場合には動作不能になった場合に行われることもある。その場合は、故障したコンポーネントを取り替えなければならない。ある顧客は、システムの製造業者/供給業者に、必要な修理またはアップグレードの管理を依頼する。他の顧客は、サービス技術者(または技術サポート員)を採用する。そのようなサービス技術者の主な仕事は、確実にシステムを機能させること、ならびに、顧客の社員がシステムにアクセスする能力およびシステムが処理時間に影響を受けやすい作業を継続する能力を大きく損ねることなく、システムに必要なアップグレードや修理を完了させることである。

【発明の開示】

【発明が解決しようとする課題】

【0011】

現在のシステムでは、顧客(すなわち技術サポート員)が、図1のシステムから1つのプロセッサ(例えば第1プロセッサ101)を取り外したい場合、顧客は以下の一連のステップを完了させなければならない。

(1) 第1プロセッサ101上での命令の実行を停止させ、全てのI/Oを阻止する。

(2) プロセッサ間にパーティションを置く。

(3) 次いでシステムをシャットダウンする(電源を切る)。顧客の見地からは、システムがいかなる処理も可能でなくなったので(すなわち第2プロセッサ102上の動作も停止する)、故障停止に見える。

(4) 第1プロセッサ101を取り外し、システムの電源を再び入れる。

(5) 次いで、システム(第2プロセッサ102)を休止解除させる。休止解除プロセスは、システムの再起動、OSのリブート、I/O動作の再開、および命令の処理を伴う。

【0012】

同様に、顧客が、第2プロセッサ102のみを有するシステムにプロセッサ(例えば第1プロセッサ101)を追加したい場合、前とは逆の一連のステップを実行しなければならない。

(1) 第2プロセッサ102上での命令の実行を停止させ、全てのI/Oを阻止する。顧客の見地からは、システムがいかなる処理も可能でなくなったので(すなわち第2プロセッサ102上の動作が停止する)、故障停止に見える。

(2) 次いでシステムをシャットダウンする(電源を切る)。

(3) 第1プロセッサ101を追加し、システムの電源を再び入れる。第1プロセッサ101はこの時点で初期化する。初期化は、通常、BIST(組み込み自己診断テスト)

10

20

30

40

50

等を含む一連のテストを行うことを伴う。

(4) 次いで、システムを休止解除する。休止解除プロセスは、システムの再起動、I/O動作の再開、および双方のプロセッサ上での命令の処理の再開を伴う。

【0013】

大規模商用システムでは、上述のプロセスは、極めて長い時間を要する可能性があり、状況によっては完了に数時間から何時間もかかる。このダウンタイム中、顧客はシステムを利用/アクセスすることができない。従って、故障停止は、業界またはシステムの特定の使用によっては、著しい経済的損失となる恐れがある。また、上述のように、プロセスの追加または除去のいずれかを完了させるために、システムの小規模リブートまたは完全リブートが必要である。上述の故障停止は、実際の物理パーティションを有するシステム

10

【0014】

図2は、物理パーティションを有するMPサーバクラスタの一例を示す。MPサーバクラスタ120は、バックプレーンコネクタ128を介して相互接続された3台のサーバ121、122、123を備える。各サーバは、図1のMP100と同様、プロセッサ131、メモリ136、およびI/O138を有する完全な処理システムである。点線で示す物理パーティション126は、サーバ121および122からサーバ123を分ける。サーバ121および122は、最初に相互に結合することができ、後にサーバ123を追加する。あるいは、全てのサーバを最初に相互に結合することができ、後にサーバ123を除去する。サーバ123を追加するか除去するかにかかわらず、システム全体の再構成を

20

【0015】

より大きなシステムからサーバまたはプロセッサを除去することは、多くの場合、そのコンポーネントが動作中に問題を生じたことが契機になっている。これらの問題は、不良のトランジスタ、故障した論理または配線等、様々な理由で生じ得る。通常、システム/リソースを製造すると、システムが正しく動作しているか否かを判定するため、システムは一連の試験を受ける。これは、図2のようなサーバシステムについて特に当てはまる。試験においてほぼ100パーセントの正確さであっても、製造中にいくつかの問題が検出されない場合がある。更に、製造後しばらくしてから内部コンポーネント(トランジスタ

30

【0016】

極めて大きい複雑なシステムでは、既存のシステムおよび新しい追加したシステム上で試験を実行するタスクは、多くの場合、技術者の時間の大きな部分を占めている。問題が生じた場合、この問題は通常、問題が生じてからしばらく(おそらく数日)経った後まで認識されない。問題が特定のリソースで見出されると、多くの場合、このリソースを取り替えなければならない。上述のように、リソースを置換するには、置換/除去されているリソースが残りのシステムから論理的または物理的に分けられている場合であっても、技術者はシステム全体の再構成を行わなければならない。

40

【0017】

問題のあるコンポーネントがシステムの作業負荷を共有していると、結果として、そのコンポーネントを有しないシステムよりも作業生成の効率が低くなり得る。あるいは、問題のあるコンポーネントは、処理エラーを引き起こし、これによってシステム全体が非効率的になる恐れがある。現在、かかるコンポーネントを除去するには、技術者が、最初にシステム全体の試験を行い、問題を起こしているコンポーネントを分離し、次いで上述の

50

除去ステップシーケンスを開始する必要がある。このため、システム保守の大部分では、技術者が継続的にシステムの診断試験を行う必要がある。システムの監視は、多数の工数を費やし、顧客に対して極めてコスト高となる恐れがある。また、問題のあるコンポーネントは、技術者が診断を実行するまで識別されず、システムによって処理されている動作を損なうまで識別されない可能性がある。処理結果を廃棄して、システムを最後の正しい状態にバックアップしなければならないことがある。

【0018】

本発明は、システムが、動作の問題を生じている主なホットプラグ可能なハードウェアコンポーネントを自動的に識別し、問題のあるコンポーネントから他の機能コンポーネントに動作を移すことによって問題のあるコンポーネントに動的に応答することができれば望ましいということ認識している。顧客に見えないように問題のあるコンポーネントを自動的に除去するが、問題のあるコンポーネントの存在および除去を顧客に自動的に警告することができるシステムおよび方法は、望ましい改善であろう。これらおよび他の利点は、ここに記載する本発明によって提供される。

10

【課題を解決するための手段】

【0019】

開示されるのは、システム全体の処理に実質的に介入することなく、ホットプラグ処理システムにおいて問題のあるコンポーネントを動的に検出し、この問題のあるコンポーネントをホット除去方法によって自動的に除去するための方法、システム、およびデータ処理システムである。高度化した相互接続機構およびサービス要素等の他の論理コンポーネントによって非介入ホットプラグ機能を提供するデータ処理システムは、ホットプラグ可能コンポーネントに工場レベルの試験シーケンスを開始および完了させてコンポーネントが適切に機能しているか否かを判定するための追加の論理と共に設計されている。

20

【0020】

コンポーネントが適切に機能していない場合、サービス要素およびオペレーティングシステム(OS)に知らせて、システム全体からコンポーネントの除去を開始する。OSは、そのコンポーネントの作業負荷をシステムの他のコンポーネントに再割り当てし、OSが再割り当てを完了すると、サービス要素がコンポーネントのホット除去を開始する。サービス要素は、コンポーネントの除去を考慮する構成ファイルの選択を行う。次いで、サービス要素は、パーティションを設定し、このため、除去されているコンポーネントはシステムの残り部分とは相互作用しない。そして、コンポーネントは、システムから論理的かつ電気的に分離される。1つの実施形態では、サービス要素は、出力デバイスにメッセージを出力させて、サービス技術者またはシステム管理者にコンポーネントの除去を知らせる。

30

【0021】

本発明の上述および追加の目的、特徴、および利点は、以下の詳細な説明において明らかとなる。

【0022】

本発明の新規の特性と考えられる特徴は、特許請求の範囲において述べる。しかしながら、本発明自体は、その好適な使用形態、更に別の目的および利点と共に、添付図面と関連付けて読む例示的な実施形態の以下の詳細な説明を参照することによって、最も良く理解されよう。

40

【発明を実施するための最良の形態】

【0023】

本発明は、現在のシステムでは避けられないダウンタイムを結果として生じることなく、処理システムの主コンポーネントの機能のホットプラグ追加/除去を可能とする方法およびシステムを提供する。具体的には、本発明は、データ処理システム業界に3つの大きな進歩をもたらす。(1)進行中のシステム動作に介入することのない、対称マルチプロセッサシステム(SMP)におけるホットプラグ可能なプロセッサ/サーバ。(2)進行中のシステム動作に介入することのない、マルチプロセッサシステム(MP)における、

50

メモリ、異種プロセッサ、および入出力（I/O）拡張デバイスを含むホットプラグ可能コンポーネント、および（3）他のシステムコンポーネントの動作を停止させない、システムのホットプラグコンポーネントに影響を与える問題の自動検出および問題のあるコンポーネントの動的除去。

【0024】

簡略化のため、上述の3つの改善は、別個の見出しで識別するセクションとして提示し、一般的なホットプラグ機能は、ホット追加のセクションおよび別個のホット除去のセクションに分ける。これらのセクションの内容は重複する場合がある。しかしながら、実施形態の機能において生じる重複は、最初に発生した場合および後に参照する場合に詳細に記載する。

10

【0025】

1. ハードウェア構成

ここで図面、特に図3を参照すると、本発明の様々な機構の実施を可能とする接続機構および他のコンポーネントによって設計されたマルチプロセッサシステム（MP）が示されている。MP200は、プロセッサ201および202を備える。また、MP200は、メモリ204および入出力（I/O）コンポーネント206も備える。様々なコンポーネントは、ホットプラグコネクタ220を備える相互接続機構208を介して相互接続されている。新しいホットプラグ可能コンポーネントの追加は、相互接続機構208のホットプラグコネクタ220を介して（直接的または間接的に）行われるが、これについては以下で更に詳細に説明する。

20

【0026】

相互接続機構208は、配線および制御論理を含み、これによって、コンポーネント間の通信をルーティングすると共に、ハードウェア構成の変更に対するMP200の応答を制御する。制御論理は、ルーティング論理207および構成設定論理209を備える。具体的には、MP200の左側に示すように、構成設定論理209は、第1および第2の構成設定、すなわちコンフィギュレーションA214およびコンフィギュレーションB216を備える。コンフィギュレーションA214およびコンフィギュレーションB216は、ラッチ217によって制御されるモード設定レジスタ218に結合されている。構成設定論理209内のコンポーネントの実際の動作については、以下で更に詳細に述べる。

【0027】

上述のコンポーネントに加えて、MP200は、サービス要素（S.E.）212も備える。S.E.212は、小さいマイクロコントローラであり、（オペレーティングシステム（OS）とは別個の）特別なソフトウェア符号化論理を備え、これを用いて、システムのコンポーネントを維持し、大規模システムに対するインタフェース動作を完了させる。このため、S.E.212は、MP200を制御するために必要なコードを実行する。S.E.212は、OSに、MP内の追加のプロセッサリソース（すなわちプロセッサ数の増加/削減）を通知し、他のシステムリソース（すなわちメモリ、I/O等）の追加/除去を通知する。

30

【0028】

図4および5は、図3の200に類似した2つのMPを示し、これらは、ホットプラグコネクタ220を介して共に結合されて、より大きな対称MP（SMP）システムを形成する。MP200は、要素0および要素1と示されるが、かかる表示は説明の目的のために必要なものである。要素1は、別個のMPのホットプラグコネクタ220を結合するために設計された配線、コネクタピン、またはケーブル接続を介して要素0に結合することができる。1つの実施形態では、MPを実際に背景プロセッサ拡張ラックにプラグインし、これによって顧客のSMPを拡張して追加のMPを収容することができる。

40

【0029】

一例として、要素0は、顧客の主システム（またはサーバ）であり、この顧客が主システムの処理機能/リソースの増大を望んでいるものとする。要素1は、システム技術者によって主システムに追加される二次システムである。本発明によれば、要素1の追加は、

50

ここに提供するホットプラグ動作によって行われ、要素 1 を接続している間、顧客は要素 0 のダウンタイムを経験することはない。

【 0 0 3 0 】

図 4 および 5 内に示すように、SMP 300 は、点線で示す物理パーティション 210 を備え、これが要素 1 から要素 0 を分けている。物理パーティション 210 によって、各 MP 200 は互いにある程度独立して動作することができる。ある実施では、物理パーティション 210 は、他の MP 200 の動作上の問題およびダウンタイムから各 MP 200 をほぼ遮断する。

【 0 0 3 1 】

II. SMP におけるプロセッサの非介入ホットプラグ可能追加

10

図 6 は、要素 0 に要素 1 を追加する非介入ホットプラグ動作を行うプロセスのフローチャートを示す。以下に説明する「ホット追加」の例では、MP 200 の最初の動作状態は以下の通りである。

要素 0：相互接続機構 208 上でコンフィギュレーション A 214 を用いて OS およびアプリケーションを実行している。また、要素 0 は要素 1 から電気的および論理的に分かれている。

サービス要素 0：単一の MP すなわち要素 0 のコンポーネントを管理している。

接続機構：コンフィギュレーション A 214 を介したルーティング制御等。ラッチ位置はコンフィギュレーション A に設定されている。

要素 1：まだ存在していないか、または存在しているが、まだシステムにプラグインされて

20

【 0 0 3 2 】

図 3、4、および 5 に示すもの以外に他のハードウェアコンポーネントが可能である。設けられているものは例示の目的のためのみに示し、本発明を限定することを意図していない。本実施形態では、MP 200 は、設定されたサイクル数内での切り替えを実行可能とするための論理も備えるので、顧客には動作時間の明らかな損失は見られない。ある数のサイクルを割り当てて、切り替えを行うことができる。接続機構制御論理は、構成切り替えを実行するため、アービタからそのサイクル量を要求する。ほとんどの実施では、実際の必要な時間は、1 秒の約 100 万分の 1 (1 マイクロ秒) であり、これは顧客の観点からは無視できる (または見えない)。

30

【 0 0 3 3 】

図 6 に戻ると、プロセスはブロック 402 において開始し、ここでサービス技術者は、要素 0 (EL0) が実行している間に、要素 0 のホットプラグコネクタ 220 に要素 1 (EL1) を物理的にプラグインする。次いで、ブロック 404 に示すように要素 1 に電力を印加する。1 つの実施では、技術者は、要素 1 を物理的に電源に接続する。しかしながら、本発明では、ホットプラグコネクタ 220 を介して電力を供給することも考えられるので、電源に直接接続しなければならないのは主システムすなわち要素 0 のみである。これは、全ての MP をプラグ接続するバックプレーンコネクタを介して達成可能である。

【 0 0 3 4 】

いったん要素 1 が電力を受容すると、要素 1 内の S.E. は、要素 1 を初期化するためのチェックポイントステップのシーケンスを完了させる。1 つの実施形態では、要素 1 に一組の物理ピンを設け、これらをサービス技術者によって選択してチェックポイントプロセスを開始する。しかしながら、ここで説明する実施形態では、ブロック 406 に示すように、S.E. 0 が、要素 0 に対する別の要素のプラグインの自動検出を完了する。次いで、S.E. 0 は、マスタの役割を負い、S.E. 1 をトリガして、ブロック 408 に示すように、要素 1 のパワーオンリセット (POR) を開始する。POR の結果、クロックがオンし、BISS T を実行し、要素 1 のプロセッサ、メモリおよび接続機構を初期化する。

40

【 0 0 3 5 】

1 つの実施形態では、S.E. 1 は、試験アプリケーションを実行して、要素 1 が適切

50

に動作していることを保証する。このため、ブロック410において、上述の試験に基づき、要素1が「クリーン」か、すなわち主システム(要素0)に統合される準備ができていのか否かを判定する。要素1が統合のためにクリアされていると仮定すると、次いで、ブロック412に示すように、S.E.0およびS.E.1は、双方のMP200が動作/実行している間に、各MP200の接続機構間の相互接続を初期化する。このプロセスは、通信ハイウェイを開放するので、双方の接続機構はタスクを共有することができ、情報のルーティングを効率的に調整することができる。このプロセスには、電氣的に接続されたドライバおよび受信器をイネーブルすること、および、必要な場合には、ブロック414に示すように、この結合システムの最も効率的な動作のためにインタフェースを調整することが含まれる。1つの実施形態では、インタフェースの調整は内部プロセスであり、接続機構の制御論理によって自動的に完了する。システム全体で動作を同期させるため、要素0の制御論理がマスタの役割を負う。すると、要素0の制御論理は、要素0および要素1の双方の全ての動作を制御する。要素1の制御論理は、要素0の動作パラメータ(例えば構成モード設定)を自動的に検出し、それ自身の動作パラメータを同期させて、要素0のものを反映させる。相互接続機構208は、要素0の論理の制御のもとで、論理的および物理的に結合される。

10

【0036】

インタフェースの調整を行っている間、ブロック416に示すように、双方の要素のモード設定レジスタ218にコンフィギュレーションB216をロードする。同じ構成モードをロードすることによって、この結合システムは、接続機構レベルで同じルーティングプロトコルにより動作することができる。どちらか一方の構成モード/プロトコルを選択するプロセスは、ラッチ217によって制御される。動的な例では、S.E.によって、次の要素がプラグインされ、初期化を完了し、システム内に組み込まれる準備ができたことが示されると、新しいトポロジのため、既存の要素および新しい要素の双方で構成レジスタをセットアップする。次いでSEは、ハードウェアに「ゴー」コマンドを発する。例示の実施形態では、ゴーコマンドを実行すると、自動化状態機械は接続機構の動作を一時的に停止し、ラッチ217を変更してコンフィギュレーションBを用い、接続機構の動作を再開する。代替的な実施形態では、SEのゴーコマンドは、全要素上のラッチ217を同期して変更する。いずれの実施形態でも、コンピュータシステムにおけるOSおよびI/Oデバイスは、故障停止を経験しない。なぜなら、構成切り替えは、ほぼプロセッササイクルで生じるからである(この実施形態ではマイクロ秒未満)。ラッチの値は、SMP上でどのように情報をルーティングするかをハードウェアに示し、接続機構上で実施されるルーティング/動作プロトコルを決定する。1つの実施形態では、ラッチはマルチプレクサ(MUX)のための選択入力として機能し、そのデータ入力ポートは構成レジスタの一方に結合されている。ラッチ内の値は、一方の構成レジスタまたは他方の構成レジスタをMUX出力として選択させる。MUXの出力は、モード設定レジスタ218にロードされる。次いで、自動化状態機械コントローラは、システムが実行している間にプロトコルを実施する。

20

30

【0037】

ホットプラグ動作の後のシステムの動作状態は以下の通りである。

40

要素0：接続機構208上でコンフィギュレーションB216を用いてOSおよびアプリケーションを実行している。要素0は、電氣的および論理的に要素1に接続されている。

要素1：接続機構208上でコンフィギュレーションB216を用いてOSおよびアプリケーションを実行している。要素1は、電氣的および論理的に要素0に接続されている。

サービス要素0：要素0および要素1の双方のコンポーネントを管理する。

接続機構：コンフィギュレーションBを介したルーティング制御等。ラッチ位置はコンフィギュレーションBに設定されている。

【0038】

50

ブロック 418 に示すように、この結合システムは、増大した処理能力、分散メモリ等を考慮した新しいルーティングプロトコルで動作を続ける。顧客は、主システムのダウンタイムを経験することなく、更にシステムをリブートする必要もなく、すぐに結合システムの増大した処理リソース/能力の利点を得る。

【0039】

上述のプロセスは、一度に1つ、または同時に複数のいずれかで、多数の追加要素の接続を含むように拡張可能である。一度に1つを完了すると、選択された構成レジスタは、要素を新しく追加（または除去）するたびに切り替えられる。また、別の実施形態では、異なる構成レジスタ範囲を設けて、ある特定の数までのホットプラグされた要素を処理することができる。例えば、システムが1、2、3、または4個の要素を含むことに基づいて、4個の異なるレジスタファイルを選択のために利用可能である。構成レジスタは、メモリ内で特定の位置を示すが、この位置に、特定のハードウェア構成用に設計されたより大きな動作/ルーティングプロトコルが格納され、処理システムの現在の構成に基づいて活性化される。

【0040】

III. メモリ、I/Oチャネル、および異種プロセッサの非介入ホットプラグ

図8に、ホットプラグ機能の1つの追加的な拡張を示す。具体的には、図8は、上述の非介入ホットプラグ機能の機構を拡張して、追加メモリおよびI/Oチャネルならびに異種プロセッサのホットプラグ追加に対応する。MP500は、図2のMP200と同様の主コンポーネントを含み、新しいコンポーネントは500番台の参照番号で識別する。主コンポーネント（すなわち、相互接続機構208を介して共に結合されたプロセッサ201および202、メモリ504A、ならびにI/Oチャネル506A）に加えて、MP500は、接続機構208上にいくつかの追加のコネクタポートを含む。これらのコネクタポート間に、ホットプラグメモリ拡張ポート521、ホットプラグI/O拡張ポート522、およびホットプラグプロセッサ拡張ポート523が含まれる。

【0041】

各拡張ポートは、対応する構成論理509A、509B、および509Cを有し、それぞれのコンポーネントのためのホットプラグ動作を制御する。メモリ504Aに加えて、追加のメモリ504Bを、MP300ならびに要素0および要素1に対して上述したプロセスと同様に、接続機構208のメモリ拡張ポート521に「プラグイン」することができる。アドレス0からNまでの初期メモリ範囲を拡張して、N+1からMまでのアドレスを含ませる。いずれのサイズのメモリの構成モードも、ラッチ517Aによって選択可能である。ラッチ517Aは、追加のメモリ504Bを付加する場合、S.E.212によって設定される。また、I/Oチャネル506B、506CをホットプラグI/O拡張ポート522にホットプラグすることによって、追加のI/Oチャネルを提供可能である。ここでも、追加のI/Oチャネル506B、506Cを追加する場合、I/Oチャネルのサイズの構成モードは、S.E.212によって設定されるラッチ517Cにより選択可能である。

【0042】

最後に、非対称プロセッサ（すなわちMP200内のプロセッサ201および202とは異なる方法で構成/設定されたプロセッサ）を、ホットプラグプロセッサ拡張ポート523にプラグインし、サーバ/要素1について上述したプロセスと同様に初期化することができる。しかしながら、利用可能なメモリおよびI/Oリソースの量の増大のみを検討しなければならない他の構成論理509A、509Bとは異なり、プロセッサ追加のための構成論理509Cでは、より多くのパラメータを考慮する必要がある。なぜなら、プロセッサは非対称であり、正しい構成モードの選択において作業負荷の分割、割り当て等を考慮しなければならないからである。

【0043】

上述の構成によって、システムは、MP500上の処理に著しい障害を生じることなく、プロセッサ、メモリ、I/Oチャネル等を縮小/拡張することができる。具体的には、

10

20

30

40

50

上述の構成によって、メモリおよびI/Oの双方で利用可能なアドレス空間を拡張（および縮小）することが可能となる。アドオンまたは除去の各々は互いに独立して、すなわちプロセッサ対メモリまたはI/Oとして処理され、図示のように別個の論理によって制御される。従って、本発明は、「ホットプラグ」の概念を、従来の言葉の意味ではホットプラグすることができないデバイスに拡大する。

【0044】

図8に示すシステムの初期状態は、メモリ空間量N、I/O空間の数（すなわち、I/Oデバイスを接続するチャンネル）R、速度Zでの処理能力量Yなどを含む。

【0045】

システムの最終状態は、上述の初期状態から、メモリ空間量M（ $M > N$ ）、I/Oチャンネル数T（ $T > R$ ）、ならびに速度ZおよびZ+Wでの処理能力量Y+Xまでの範囲である。

10

【0046】

上述の変数は、例示の目的のためにのみ用い、特定のパラメータ値を示したり本発明を限定したりする意図はない。

【0047】

上述の実施形態では、サービス技術者が、追加のメモリ、プロセッサまたはI/Oを物理的にプラグインすることによって新しいコンポーネント（複数可）をインストールし、次いでS.E.212が自動検出および初期化/構成プロセスを完了させる。追加のメモリがインストールされると、S.E.212は信頼性試験を実行し、全てのコンポーネントで、S.E.212はBISTを実行する。次いで、S.E.212は、インタフェース（点線で表す）を初期化し、代替の構成レジスタ（複数可）をセットアップする。S.E.212は、1マイクロ秒未満で全ハードウェアの切り替えを完了し、次いで、OSに新しいリソースの可用性を通知する。次いでOSは、どのコンポーネントが利用可能で、どの構成が実行しているかに従って、作業負荷の割り当てを完了させる。

20

【0048】

IV. 処理システムにおけるホットプラグされたコンポーネントの非介入除去

図7は、ホットプラグコンポーネントの非介入除去を完了させるプロセスのフローチャートを示す。以下、図4および図5も参照して、要素1および要素0の双方を備える処理システムにおける要素1の除去について説明する。図7に示す除去の例では、SMPの最初の動作状態は、図6のホットプラグ動作の後の上述の動作状態である。

30

【0049】

要素1を除去するためには、サービス技術者が、最初に何らかの方法で待ち状態の除去を知らせる必要がある。1つの実施形態では、各要素の外面上にホット除去ボタン225を設ける。ボタン225は、発光ダイオード（LED）またはその他の信号手段を含み、これによって、動作中の要素が、「オンライン」すなわちプラグインおよび機能しているか、またはオフラインであるとして、サービス技術者によって視覚的に識別可能である。従って、図7において、サービス技術者が要素1を除去したい場合、ブロック452に示すように、技術者は最初にボタン225を押す。別の実施形態では、各要素は何らかの種類のバックプレーンコネクタに留められていると仮定し、要素1を適所に保持しているクランプの除去によって、S.E.212に再構成プロセスの開始を知らせる。更に別の実施形態では、システム管理者が、S.E.212をトリガして、特定のコンポーネントの除去動作を開始させることができる。トリガは、システム上で実行しているソフトウェア構成ユーティリティ内で除去オプションを選択することで行われる。以下のセクション5において、サービス技術者またはシステム管理者による開始を必要としない自動除去方法について説明する。

40

【0050】

いったんボタン225が押されると、顧客からは隠されて（すなわち要素0は実行したままで）、再構成プロセスが背景で開始する。ブロック454に示すように、S.E.212は、OSに、要素1のリソースの喪失処理を通知する。これに回答して、OSは、ブ

50

ロック456に示すように、要素1から要素0にタスク/作業負荷を再割り当てし、要素1を解放する。S.E.212は、OSが要素1から要素0に全処理(およびデータ格納)の再割り当てを完了したという指示を監視し、ブロック458において、再割り当てが完了したか否かについて判定を行う。いったん再割り当てが完了すると、ブロック460に示すように、OSはS.E.212にメッセージを送り、ブロック462に示すように、S.E.212は代替的な構成設定を構成レジスタ218にロードする。代替的な構成設定のロードを行うには、S.E.212が、その構成設定を選択するためラッチ217内で値を設定する。別の実施形態では、ラッチ217は、ボタン225が最初に押されて除去をトリガした場合に設定される。要素1は、要素0を中断することなく、SMP接続機構から論理的に除去され、電氣的に除去される。次いで、ブロック464に示すように、S.E.212はボタン225を明るくする。この照明によって、サービス技術者に再構成プロセスが完了したことを知らせる。次いで、ブロック466に示すように、技術者は電源を切り要素1を物理的に除去する。

10

【0051】

上述の実施形態では、ボタン225内のLEDを利用してサーバの動作状態を知らせる。このため、予め確立されたカラーコードを設定して、要素をオン(ホットプラグ)またはオフ(除去)した場合に顧客または技術者に認識させる。例えば、青い色は、要素が十分に機能し、電氣的および論理的に取り付けられていることを示し、赤い色は、要素が再構成の過程にあり、まだ物理的に除去してはいけなことを示し、緑色(または照明なし)は、要素の再構成が済んでおり(またはもはや論理的にも電氣的にも存在しない)、物理的に除去可能であることを示す。

20

【0052】

V. 問題のあるコンポーネントの非介入自動検出および除去

ホットプラグコンポーネントによる上述の手動除去機能が与えられれば、本発明を拡張することによって、問題のある要素(またはコンポーネント)の非介入自動検出、および、予め確立された(または所望の)動作レベルで機能していない要素または不良の要素の自動分離が行われる。本発明の非介入ホットプラグ機能により、技術者は、処理システム全体を分解することなく問題のある要素を除去することができる。本発明は、この機能を更にもう一段階拡張して、システムにプラグインされたコンポーネントの自動的な問題検出を可能とし、その後、非介入的に(システムが動作している間に)システムから問題のある/不良のコンポーネントを動的に除去する。技術者が開始する再構成とは異なり、この問題要素/コンポーネントの検出およびこれに応じた再構成は自動的に行われ、残りの処理システムで顕著な故障停止を生じることなく背景で行われる。本実施形態によって、問題のある/不良のコンポーネントの効率的な検出が可能となり、問題のあるコンポーネントを処理タスクに用いた場合の、システム全体の完全性に対する潜在的な問題を軽減する。この実施形態は、更に、残りのシステムに故障停止を生じることなく適時に不良のコンポーネントを置換することに役立つ。

30

【0053】

図9は、ホットプラグ環境内における問題のあるコンポーネントの自動検出および動的割り当て解除のプロセスを示す。このプロセスはブロック602において開始し、S.E.が、システムに追加されている新しいコンポーネントを検出し、システムの現在の有効動作状態(プロセッサ、構成レジスタ等の構成状態)をセーブする。あるいは、自動的に、S.E.は、システム動作中の予め確立された時間間隔で、更に、新しいコンポーネントがシステムに追加された場合はいつでも、動作状態をセーブする。ブロック604に示すように、新しい動作状態を入力し、システムハードウェア構成(新しいコンポーネントを含む)を試験する。ブロック606において、新しい動作状態およびシステム構成の試験がOK信号を生成するか否かを判定する。システム構成の試験には、システム全体に対するBISTまたは新しいコンポーネントのみに対するBIST、および新しいコンポーネントの信頼性試験等の他の構成試験が含まれ得る。試験がOK信号を戻した場合、ブロック608に示すように、新しい動作状態を現在の状態としてセーブする。次いで、プロ

40

50

ック610に示すように、新しい動作状態がシステム全体に実装される。プロセスループは、変更があった場合または所定の時間間隔が経過した場合に、いずれかの新しい動作状態の試験に戻る。

【0054】

試験が問題ありの指示を戻した場合、例えばBISTが失敗したか、または実行時エラーチェック回路が活性化した場合、検出および割り当て解除プロセスの割り当て解除段階を開始する。S.E.は、図7に示したステップと同様の一連のステップを経るが、サービス技術者が除去プロセスを開始した図7とは異なり、この実施形態の除去プロセスは自動化されており、あるレベルで試験が失敗したことの指示を受信したことの直接の結果として開始する。S.E.は、ブロック612に示すように除去プロセスを開始する。ブロック614に示すように、出力デバイスにメッセージを送信して、顧客またはサービス技術者に、特定のコンポーネントで問題が見つかり、そのコンポーネントを除去したこと（または除去していること）（すなわちオフラインとすること）を知らせる。1つの実施形態では、出力デバイスは、処理システムに接続されたモニタであり、これによってサービス技術者はシステム全体の動作パラメータを監視する。別の実施形態では、問題は、製造業者または供給業者に（ネットワーク媒体を介して）メッセージとして送られ、次いでその業者が、ブロック616に示すように、不良のコンポーネントを置換または修理するための迅速な処置を取ることができる。

10

【0055】

1つの実施形態では、検出段階は、チップレベルでの試験を含む。このため、製造業者レベルの試験が、システムが動作している間およびシステムを顧客に出荷した後に、システム上で行われる。上述のプロセスによって、システムは、製造品質自己試験機能およびそれらの試験に基づく自動的な非介入動的再構成が可能である。ある1つの特定の実施形態は、パーティションのパーティシャル化を伴う。パーティション切り替え時に、パーティションの状態をセーブする。製造業者品質自己試験は、様々なコンポーネントにおいて専用ハードウェアによって実行される。試験は、上述のように非介入でパーティションを切り替えるのに要するのとほぼ同じ時間（1マイクロ秒）のみを要する。試験によってパーティションが悪いことが示されると、S.E.は、自動的に悪いコンポーネントから作業負荷を再割り当てし、セーブされた以前の良好な状態を復元する。

20

【0056】

本発明について好適な実施形態を参照して具体的に図示し説明したが、当業者には、本発明の精神および範囲から逸脱することなく、形態および詳細において様々な変更を行い得ることは理解されよう。

30

【図面の簡単な説明】

【0057】

【図1】従来技術によるマルチプロセッサシステム(MP)の主なコンポーネントのブロック図である。

【図2】従来技術によるサーバクラスタの複数のサーバを示すブロック図である。

【図3】本発明の1実施形態に従って、様々なホットプラグ機構を提供するように用いられる接続機構制御論理によって設計されたデータ処理システム(サーバ)のブロック図である。

40

【図4】本発明の1実施形態に従ってホットプラグのために構成された図3の2つのサーバを含むMPのブロック図である。

【図5】本発明の1実施形態に従ってホットプラグのために構成された図3の2つのサーバを含むMPのブロック図である。

【図6】本発明の1実施形態に従って図4のMPにサーバを追加するプロセスを示すフローチャートである。

【図7】本発明の1実施形態に従って図4のMPからサーバを除去するプロセスを示すフローチャートである。

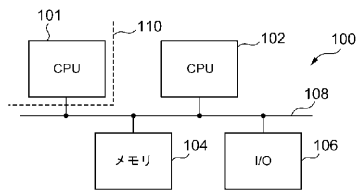
【図8】本発明の1実施形態に従って全ての主コンポーネントのホットプラグ拡張を可能

50

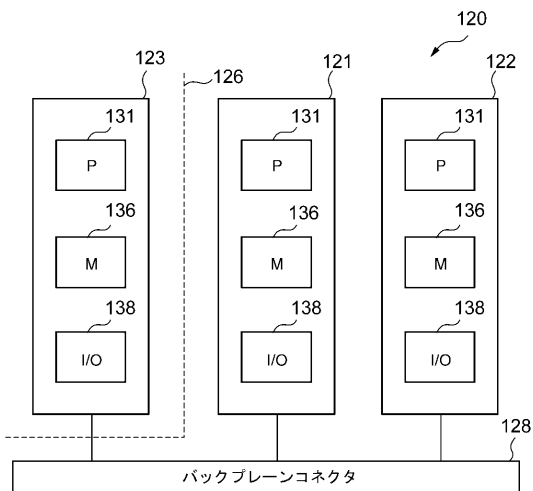
とするデータ処理システムのブロック図である。

【図9】本発明の1実施形態に従って、検出可能な問題を生じているホットプラグされたコンポーネントの自動検出および動的除去を完了するプロセスを示すフローチャートである。

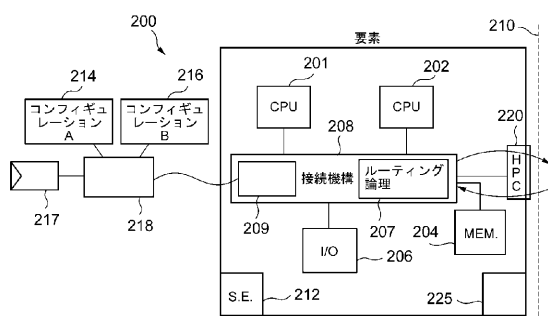
【図1】



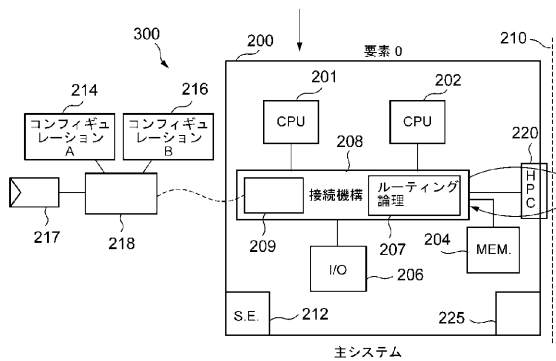
【図2】



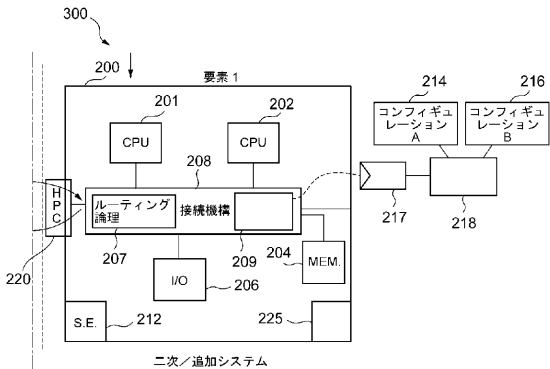
【図3】



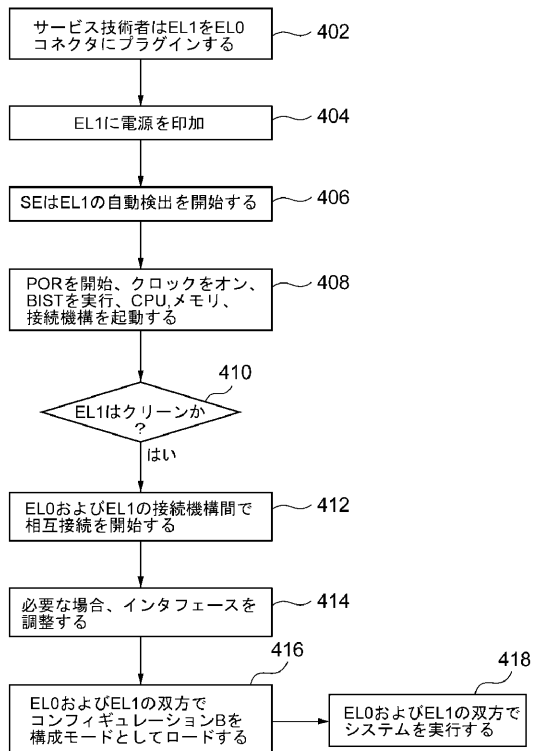
【図4】



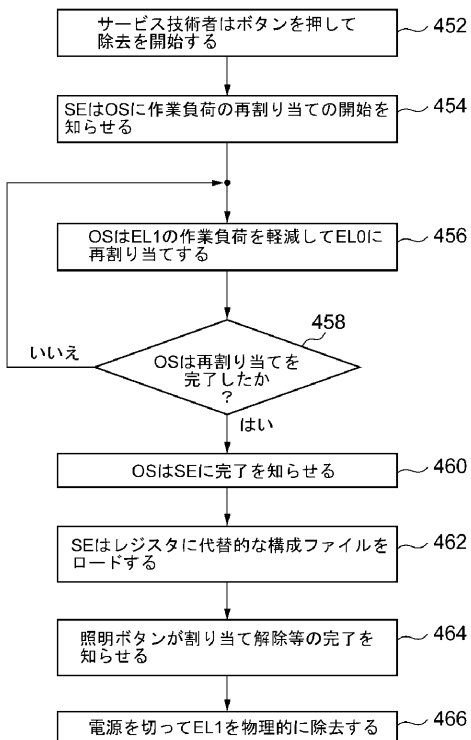
【図5】



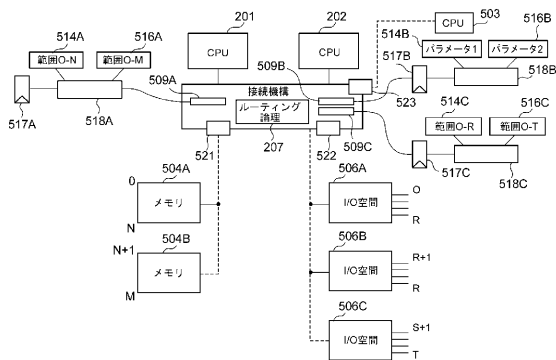
【図6】



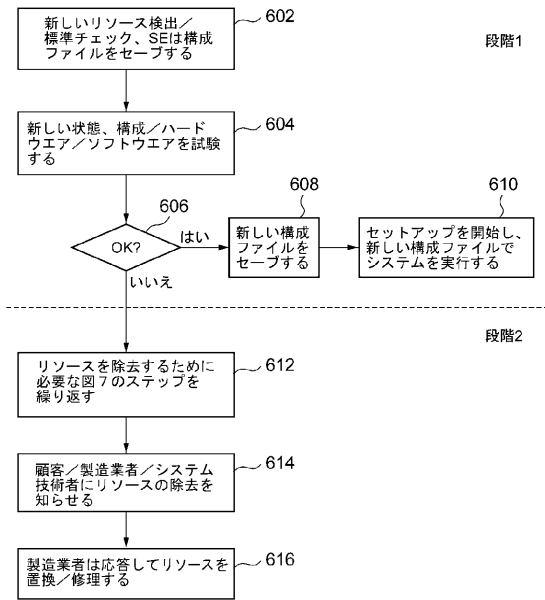
【図7】



【図8】



【 図 9 】



フロントページの続き

- (72)発明者 ラヴィ・クマル・アリミリ
アメリカ合衆国78759 テキサス州オースティン スパイスブラシ・ドライブ 9221
- (72)発明者 マイケル・スティーブン・フロイド
アメリカ合衆国78717 テキサス州オースティン テラ・ベルデ・ドライブ 15108
- (72)発明者 ケヴィン・フランクリン・リック
アメリカ合衆国78681 テキサス州ラウンドロック レグホーン・コーヴ 5100

審査官 北川 純次

- (56)参考文献 特開平10-011319(JP,A)
特開平09-319719(JP,A)
特開2000-311035(JP,A)

- (58)調査した分野(Int.Cl., DB名)
- | | |
|---------|-----------|
| G06F | 15/173 |
| G06F | 11/20 |
| JST7580 | (JDream2) |
| JSTPlus | (JDream2) |