(12) 特許協力条約に基づいて公開された国際出願

(19) 世界知的所有権機関
国際事務局

(43) 国際公開日
2010 年 2 月 11 日(11.02.2010)

PCT

(10) 国際公開番号
WO 2010/016104 A1

(51) 国際特許分類:
G06F 9/50 (2006.01) *G06F 9/46* (2006.01)

(21) 国際出願番号: PCT/JP2008/063977

(22) 国際出願日: 2008 年 8 月 4 日(04.08.2008)

(25) 国際出願の言語: 日本語

(26) 国際公開の言語: 日本語

(71) 出願人 (米国を除く全ての指定国について): 富士通株式会社 **(FUJITSU LIMITED)** [JP/JP]; 〒2118588 神奈川県川崎市中原区上小田中４丁目１番１号 Kanagawa (JP).

(72) 発明者; および

(75) 発明者/出願人 (米国についてのみ): 井村 栄克 **(IMURA, Hidekatsu)** [JP/JP]; 〒2118588 神奈川県川崎市中原区上小田中４丁目１番１号 富士通株式会社内 Kanagawa (JP).

(74) 代理人: 真田 有, 外 **(SANADA, Tamotsu** et al.); 〒1800004 東京都武蔵野市吉祥寺本町１丁目１０番３１号吉祥寺マークビル５階 Tokyo (JP).

(81) 指定国 (表示のない限り、全ての種類の国内保護が可能): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RS, RU, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) 指定国 (表示のない限り、全ての種類の広域保護が可能): ARIPO (BW, GH, GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), ユーラシア (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), ヨーロッパ (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MT, NL, NO, PL, PT, RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

添付公開書類:

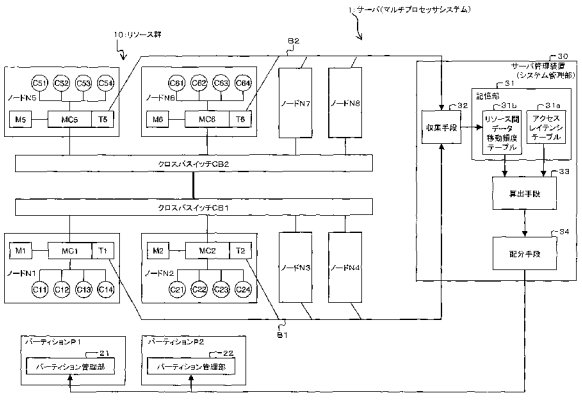
— 国際調査報告（条約第 21 条(3)）

WO 2010/016104 A1

(54) **Title**: MULTIPROCESSOR SYSTEM, MANAGEMENT DEVICE FOR MULTIPROCESSOR SYSTEM, AND COMPUTER-READABLE RECORDING MEDIUM IN WHICH MANAGEMENT PROGRAM FOR MULTIPROCESSOR SYSTEM IS RECORDED

(54) 発明の名称: マルチプロセッサシステム，マルチプロセッサシステム用管理装置およびマルチプロセッサシステム用管理プログラムを記録したコンピュータ読取可能な記録媒体

[図1]



1 SERVER (MULTIPROCESSOR SYSTEM)
10 RESOURCES
N5, N6, N7, N8 NODE
CB2, CB1 CROSSBAR SWITCH
N1, N2, N3, N4 NODE
P1, P2 PARTITION
21, 22 PARTITION MANAGEMENT UNIT
30 SERVER MANAGEMENT DEVICE (SYSTEM MANAGEMENT UNIT)
32 COLLECTION MEANS
31 STORAGE SECTION
31b TABLE OF DATA MOVEMENT FREQUENCY BETWEEN RESOURCES
31a TABLE OF ACCESS LATENCY
33 CALCULATION MEANS
34 ALLOCATION MEANS

(57) **Abstract**: Partitioning is optimized by realizing resource allocation in consideration of system characteristics and the processing capacity of the whole system is improved. A system management unit (30) calculates optimum allocation of resources to each partition according to distance information about the distance between the resources and the frequency of data movement between the resources. The resources are allocated to partitions (P1, P2) through partition management units (21, 22) so as to realize optimum allocation.

(57) 要約: 本件は、システムの特性を意識したリソース配分を実現することによってパーティション分割の最適化をはかり、システム全体の処理性能を向上させる。このため、本件のシステム管理部（３０）において、複数のリソース相互間の距離に係る距離情報と複数のリソース相互間のデータ移動頻度とに基づき、各パーティションに対する複数のリソースの最適配分が算出され、その最適配分の状態になるように複数のパーティション管理部（２１，２２）を介して複数のパーティション（Ｐ１，Ｐ２）に複数のリソースが配分される。

明 細 書

マルチプロセッサシステム，マルチプロセッサシステム用管理装置およびマルチプロセッサシステム用管理プログラムを記録したコンピュータ読取可能な記録媒体

技術分野

[0001] 本発明は、CPU（Central Processing Unit；演算処理部），メモリなどの複数のリソースを複数のパーティションに割り当て分割し、各パーティションに属するリソースを使用してパーティション毎にデータ処理を実行する、マルチプロセッサシステム等の計算機システムに用いて好適の技術に関する。

背景技術

[0002] 一般に、多数のCPU，メモリ，I／O（入出力部）から構成される大規模マルチプロセッサシステムでは、NUMA（Non-Uniform Memory Access）と呼ばれるアーキテクチャが採用されることが多い。このNUMAアーキテクチャでは、メモリのレイテンシが均一でないこと、つまり、「近いメモリ」と「遠いメモリ」とが存在することが特徴である。ここで、レイテンシとは、CPU等がメモリにアクセスした際にメモリからの応答時間に対応するもので、レイテンシの小さいメモリは「近いメモリ」、レイテンシの大きいメモリは「遠いメモリ」と定義することができる。

[0003] また、大規模マルチプロセッサシステムは、上述のように、多数のCPUやメモリ，I／Oをリソースとしてそなえてら構成されている。このような大規模マルチプロセッサシステムでは、多数のリソースを複数のパーティションに分割し各パーティションにおいて独立のOS（Operating System）を動作させるパーティショニング技術が使用されている。

[0004] なお、例えば、下記特許文献1，2には、論理パーティション（ソフトパーティション）技術が開示されている。この論理パーティション技術では、ホストOS（制御ホスト）上で複数のOSが論理パーティション毎に起動される。各論理パーティションには論理プロセッサ等が割り当てられており、ホストOSによって、論理プロセッサ等と物理プロセッサ等とが対応付けられながら、各OSによる処理が論理パーティション毎に実行さ

れる。論理パーティション技術は、仮想的なパーティションを用いるものであるのに対し、本件は、リソースを分割して用いるハードパーティション技術、つまり、パーティション毎に物理的に異なるリソースを用いる技術を前提としている。

特許文献1:特開2006－127462号公報

特許文献2:特開2007－193776号公報

発明の開示

発明が解決しようとする課題

[0005] ところで、NUMAアーキテクチャを採用したマルチプロセッサシステムでパーティション分割を行なう場合、処理性能の低下を招かないためにも、パーティションの構成要素(リソース)が、極力、複数のノードにまたがらないシステム構成とすることが望ましい。従って、通常、ノード単位でパーティション分割を行なう。しかし、分割後、各パーティション内においてCPUやメモリの追加／削減／故障に伴う変更などを行なっているうちに、期せずして、パーティションの構成要素が複数のノードにまたがってしまうことがある(例えば図5参照)。

[0006] パーティション構成が不適切な場合、例えば、上述のごとくパーティションの構成要素が複数のノードにまたがるような場合、以下のような不具合が生じる。つまり、プロセッサ(CPU)が「遠いメモリ」にアクセスすることになり、メモリレイテンシが増加する。また、メモリアクセスを行なうのにより多くの通信路を経由することになり、マルチプロセッサシステム全体でのトラフィックが不必要に増加してしまう。その結果、システム全体の処理性能が低下してしまう。

[0007] 本発明の目的の一つは、システムの特性を意識したリソース配分を実現することによってパーティション分割の最適化をはかり、システム全体の処理性能を向上させることである。

[0008] なお、前記目的に限らず、後述する発明を実施するための最良の形態に示す各構成により導かれる作用効果であって、従来の技術によっては得られない作用効果を奏することも本発明の他の目的の一つとして位置付けることができる。

課題を解決するための手段

[0009] ここに開示されるマルチプロセッサシステムは、複数のリソース，複数のパーティショ

ン管理部およびシステム管理部を有している。該複数のリソースは、複数のパーティションのいずれか一つに対し単独で割当可能なものである。該複数のパーティション管理部は、該複数のパーティションのそれぞれに属するリソースを管理するものである。該システム管理部は、該複数のリソースおよび該複数のパーティション管理部を管理するものである。そして、該システム管理部は、第1テーブル記憶手段，収集手段，第2テーブル記憶手段，算出手段および配分手段を有している。ここで、該第1テーブル記憶手段は、該複数のリソース相互間の距離に係る距離情報を定義する第1テーブルを記憶するものである。該収集手段は、該複数のリソース相互間のデータ移動情報を収集するものである。該第2テーブル記憶手段は、該収集手段によって収集された前記データ移動情報に基づく該複数のリソース相互間のデータ移動頻度を保持する第2テーブルを記憶するものである。該算出手段は、該第1テーブルの距離情報と該第2テーブルのデータ移動頻度とに基づき、各パーティションに対する、該複数のリソースの最適配分を算出するものである。該配分手段は、該複数のパーティションに対する該複数のリソースの配分状態が該算出手段によって算出された前記最適配分の状態になるように、該複数のパーティション管理部を介して該複数のパーティションに該複数のリソースを配分するものである。

[0010] また、ここに開示されるマルチプロセッサシステム用管理装置は、上述した複数のリソースおよび複数のパーティション管理部を有するマルチプロセッサシステムにおいて、該複数のリソースおよび該複数のパーティション管理部を管理するものである。そして、この管理装置は、上述した第1テーブル記憶手段，収集手段，第2テーブル記憶手段，算出手段および配分手段を有している。

[0011] さらに、ここに開示されるマルチプロセッサシステム用管理プログラムは、上述した複数のリソースおよび複数のパーティション管理部を有するマルチプロセッサシステムにおいて、該複数のリソースおよび該複数のパーティション管理部を管理する管理装置（システム管理部）として、コンピュータを機能させるものである。このプログラムは、上述した第1テーブル記憶手段，収集手段，第2テーブル記憶手段，算出手段および配分手段として、該コンピュータを機能させる。なお、ここに開示される、コンピュータ読取可能な記録媒体は、上述したマルチプロセッサシステム用管理プログラムを

記録したものである。

発明の効果

[0012] 開示の技術によれば、マルチプロセッサシステム内のリソース間の距離情報とデータ移動頻度とに基づいて、各パーティションに対するリソースの最適配分が統計的に算出され、その最適配分に応じたリソース配分が行なわれる。これにより、システムの特性を意識したリソース配分が実現され、パーティション分割が最適化され、システム全体の処理性能が大幅に向上する。

図面の簡単な説明

[0013] [図1]本発明の一実施形態としてのマルチプロセッサシステムの構成を示すブロック図である。

[図2]本実施形態のアクセスレイテンシテーブル(第1テーブル)の一例を示す図である。

[図3]本実施形態のリソース間データ移動頻度テーブル(第2テーブル)の一例を示す図である。

[図4]図1に示すマルチプロセッサシステム用管理装置の動作について説明するためのフローチャートである。

[図5]図1に示すマルチプロセッサシステムにおけるパーティション分割の具体的な最適化動作例を説明すべく同システムの最適化前の状態を示す図である。

[図6]図1に示すマルチプロセッサシステムにおけるパーティション分割の具体的な最適化動作例を説明すべく同システムの最適化後の状態を示す図である。

符号の説明

[0014] 1 サーバ(マルチプロセッサシステム)

10 リソース群

21, 22 パーティション管理部

30 サーバ管理装置(マルチプロセッサシステム用管理装置, システム管理部)

31 記憶部(第1テーブル記憶手段, 第2テーブル記憶手段)

31a アクセスレイテンシテーブル(第1テーブル, ノード間距離テーブル)

31b リソース間データ移動頻度テーブル(第2テーブル)

32　収集手段

33　算出手段

34　配分手段

N1～N8　ノード

CB1，CB2　クロスバスイッチ

Ci1～Ci14(i＝1～8)　CPU(演算処理部;リソース)

Mi(i＝1～8)　メモリ(リソース)

MCi(i＝1～8)　メモリコントローラ(収集手段)

Ti(i＝1～8)　データ移動情報収集用テーブル(収集手段)

B1，B2　リソース間データ移動情報収集用バス(収集手段)

P1，P2　パーティション

発明を実施するための最良の形態

[0015] 以下、図面を参照して本発明の実施の形態を説明する。

図1は本発明の一実施形態としてのマルチプロセッサシステムの構成を示すブロック図である。この図1に示す本実施形態のマルチプロセッサシステムの例であるサーバ1は、CPU，メモリなどの複数のリソース(リソース群10参照)を複数のパーティションに割り当て分割し、各パーティションに属するリソースを使用してパーティション毎にデータ処理を実行するものである。なお、本実施形態では、2つのパーティションP1，P2が設定されている場合について説明するがパーティションの数は2に限定されるものではない。以下、マルチプロセッサシステム1を単に「システム1」と称する場合がある。

[0016] 本実施形態のマルチプロセッサシステム1は、リソース群10，パーティション管理部21，22およびサーバ管理装置30を有している。ここで、パーティション管理部21，22およびサーバ管理装置30は、それぞれ、例えば、ボード単位で構成されている。

[0017] リソース群10は、本実施形態ではP1，P2の複数のパーティションのいずれか一つに対し単独で割当可能な、CPUやメモリなどの複数のリソースを含んでいる。より具体的に、本実施形態において、リソース群10は、8個のノードN1～N8と、これら8個のノードN1～N8の相互間を通信可能に接続するクロスバスイッチCB1，CB2とを含

んでいる。

[0018] そして、4個のノードN1～N4は、クロスバスイッチCB1に接続され、クロスバスイッチCB1を介して相互に通信可能になっている。同様に、4個のノードN5～N8は、クロスバスイッチCB2に接続され、クロスバスイッチCB2を介して相互に通信可能になっている。また、クロスバスイッチCB1とCB2とは接続されており、4個のノードN1～N4と4個のノードN5～N8との間では、クロスバスイッチCB1およびCB2を介して相互通信が可能になっている。

[0019] 各ノードN1～N8は、システム1における複数のリソースをその物理的な配置によって分割したリソースの集合であり、各リソースはただ1つのノードに属している。例えば、各ノードNi(i＝1～8)はリソースとして、4個のCPUCi1～Ci4と、1個のメモリMiと、1個のメモリコントローラMCiと、1個のデータ移動情報収集用テーブルTiとを含んでいる。

[0020] ここで、1個のメモリMiは、例えば、複数のDIMM(Double Inline Memory Module)の組合せとして構成されている。また、メモリコントローラMCiは、CPU Ci1～Ci4, メモリMiおよびクロスバスイッチCB1(またはCB2)の相互間のデータ移動を制御する機能を有している。さらに、メモリコントローラMCiは、メモリMiに対するリードリクエストがあった場合、どのCPUからのリードリクエストであるかに関するデータ移動情報をテーブルTiに記録する機能も有している。各テーブルTiに記録されたデータ移動情報は、後述するごとく、リソース間データ移動情報収集用バスB1またはB2を介し、サーバ管理装置30の収集手段32によって収集される。

[0021] なお、図1, 図5および図6では、ノードN1, N2, N5, N6における、CPU C11～C14, C21～C24, C51～C54, C61～C64;メモリM1, M2, M5, M6;メモリコントローラMC1, MC2, MC5, MC6;テーブルT1, T2, T5, T6が図示されている。一方、ノードN3, N4, N7, N8における、CPU C31～C34, C41～C44, C71～C74, C81～C84;メモリM3, M4, M7, M8;メモリコントローラMC3, MC4, MC7, MC8;テーブルT3, T4, T7, T8の図示は省略されている。

[0022] また、ハードウェアの構造によっては、CPUとメモリとの特定の組については、CPUをメモリから分離することが不可となる場合もあるが、ここでは、CPUとメモリとの全て

の組について分離可能であると説明する。ただし、本発明は、CPUとメモリとが分離不可か分離可能かに限定されるものではない。

[0023] さらに、図1，図5および図6に示すシステム1では、ノード数が8であり、クロスバスイッチ数が2であり、また、各ノードNiにおけるCPU数が4でメモリ数が1である場合について説明したが、本発明は、これらの数に限定されるものではない。

[0024] パーティション管理部21，22は、それぞれ、パーティションP1，P2に対応してそなえられ、パーティションP1，P2に属するCPUやメモリといったリソースを管理するものである。また、各パーティション管理部21，22は、各パーティションP1，P2についての条件テーブルに基づいて、各パーティションP1，P2に属するリソースを認識する。その認識結果に従って、各パーティション管理部21，22は、複数のリソースをパーティションP1，P2のそれぞれに割り当て分割配分し、各パーティションP1，P2に属するリソースを管理する。なお、各パーティション管理部21，22における条件テーブルは、サーバ管理装置30から指示・設定される。

[0025] サーバ管理装置（マルチプロセッサシステム用管理装置，システム管理部）30は、リソース群10として示される複数のリソースおよび複数のパーティション管理部21，22を管理するもので、記憶部31，収集手段32，算出手段33および配分手段34を有している。

[0026] 記憶部31は、例えばRAM（Random Access Memory）によって構成され、第1テーブルとしてのアクセスレイテンシテーブル31aを記憶する第1テーブル記憶手段、および、第2テーブルとしてのリソース間データ移動頻度テーブル31bを記憶する第2テーブル記憶手段として機能するものである。

[0027] ここで、アクセスレイテンシテーブル（ノード間距離テーブル）31aは、図1に示すシステム1のリソース群10に属する複数のリソース相互間の距離に係る距離情報を定義するものである。このアクセスレイテンシテーブル31aでは、その距離情報として、例えば図2に示すように、各リソースの属するノード間の距離が、より具体的には、ノード間の実際のアクセスレイテンシ（アクセス待ち時間；単位：nsec）が定義されている。

[0028] このアクセスレイテンシテーブル31aで定義される距離情報、つまりアクセスレイテ

ンシは、システム1、あるいはシステム1に含まれるリソース群10の性能として性能試験によって予め取得されるもので、予め与えられて記憶部31のアクセスレイテンシテーブル31aに登録される。

[0029] ここで、図2は、本実施形態のアクセスレイテンシテーブル31aの一例を示す図である。

例えば、図2に示すアクセスレイテンシテーブル31aにおいて、ノードN1とノードN2との距離は100nsec、ノードN3とノードN8との距離は200nsecと定義されている。なお、第1テーブルとしてのアクセスレイテンシテーブル31aにおける値は、本実施形態のごとくアクセスレイテンシや、そのアクセスレイテンシに比例した値に限定されるものではない。アクセスレイテンシテーブル31aにおける値は、リソース相互間の距離に係るものであれば、実際のアクセスレイテンシのほかに、例えば、データの通過する通信路のスループットなどによって重み付けした値を使用することもできる。

[0030] なお、以下の説明では、アクセスレイテンシテーブル31aによって定義されるノードNiとノートNjとの距離、つまりアクセスレイテンシをdistance(i, j)として、下記の通り記載する。

distance(i, j)= 50 (i=j)

100 ((i≦4 かつ j≦4 かつ i≠j) または

(5≦i かつ 5≦j かつ i≠j))

200 (それ以外)

[0031] リソース間データ移動頻度テーブル31bは、収集手段32によって収集されたデータ移動情報に基づく複数のリソース相互間のデータ移動頻度を保持するものである。

[0032] ここで、収集手段32は、各ノードNiにおけるテーブルTiにメモリコントローラMCiによって記録されたデータ移動情報を、バスB1, B2を介して各テーブルTiから受信して収集するものである。そのデータ移動情報は、上述したように、どのCPUからメモリMiに対するリードリクエストを受けたかを示す情報である。

[0033] そして、収集手段32は、各テーブルTiから受信したデータ移動情報を、リソース間データ移動頻度テーブル31bに登録することにより統合する機能も有している。これにより、リソース間データ移動頻度テーブル31bには、どのCPUがどのメモリに対して

、何回、リードリクエストを発行したかに関する情報、たとえば通信回数／データ移動回数／リードアクセス回数が、データ移動頻度として登録される。

[0034] なお、ここでは、リードリクエストについて計数を行なっているが、ライトリクエスト、つまりCPUからメモリへの書き込み要求についてのデータ移動情報をメモリコントローラMCiやテーブルTiによって収集してもよい。この場合、リソース間データ移動頻度テーブル31bには、どのCPUがどのメモリに対して、何回、ライトリクエストを発行したかに関する情報、つまり通信回数／データ移動回数／ライトアクセス回数が、データ移動頻度として登録される。また、リードリクエストに係るデータ移動回数のみをデータ移動頻度として計数してもよいし、ライトリクエストに係るデータ移動回数のみをデータ移動頻度として計数してもよいし、リードリクエストおよびライトリクエストの両方に係るデータ移動回数の合計値をデータ移動頻度として計数してもよい。

[0035] このように、本実施形態では、メモリコントローラMCi, テーブルTi, バスB1, B2, リソース間データ移動頻度テーブル31bおよび収集手段32によって、複数のリソース相互間のデータ移動情報を収集する収集手段が構成される。この収集手段を用いて、CPUやメモリといったリソース間で通信が行なわれる際に、どこからどこへのデータ移動が行なわれたかが識別されデータ移動頻度としてリソース間データ移動頻度テーブル31bに記録される。

[0036] ここで、図3は本実施形態のリソース間データ移動頻度テーブル31bの一例を示す図である。この図3に示すリソース間データ移動頻度テーブル31bには、各CPUの各メモリに対するアクセス回数の具体例が記録されている。リソース間データ移動頻度テーブル31bからは、例えば、CPU C11は、メモリM1に対し1000回のアクセスを、メモリM2に対し500回のアクセスを行なっていることが分かる。また、例えば、CPU 23は、メモリM2に対し1000回のアクセスを行なっていることが分かる。

[0037] なお、図3に示すリソース間データ移動頻度テーブル31bでは、各CPUの属するノード番号iが＃NODE欄に記入されているとともに、各CPUの属するパーティション番号が＃PART欄に記載されている。ここでは、図5を参照しながら後述する例と同様に、8個のCPU C11, C12, C13, C14, C21, C22, C51, C52および2個のメモリM1, M5がパーティションP1に属し、4個のCPU C23, C24, C61, C62および

1個のメモリM2がパーティションP2に属している。

[0038] 以下の説明では、CPU Cik(i＝1～8, k＝1～4)とメモリMn(n＝1～8)との間のデータ移動回数(アクセス回数)であって、リソース間データ移動頻度テーブル31bに登録されたものを、F(Cik, Mn)と記載する。例えば、CPU C13とメモリM5との間の登録データ移動回数F(C13, M5)＝1500である。

[0039] また、CPU Cikとメモリ Mnとの間の距離、つまりノード間距離あるいはアクセスレイテンシをD(Cik, Mn)と記載すると、D(Cik, Mn)＝distance(CPUの属するノードのノードID番号i, メモリの属するノードのノードID番号n)＝distance(i, n)となる。例えばCPU C61とメモリM2との距離は、D(C61, M2)＝distance(6, 2)であり、図2に示すアクセスレイテンシテーブル31aを参照すると、200である。

[0040] 算出手段33は、アクセスレイテンシテーブル31aの距離情報(アクセスレイテンシ／メモリレイテンシ)とリソース間データ移動頻度テーブル31bのデータ移動頻度とに基づき、各パーティションP1, P2に対する、複数のリソースの最適配分を算出するものである。

[0041] このとき、算出手段33は、まず、アクセスレイテンシテーブル31aの距離情報とリソース間データ移動頻度テーブル31bのデータ移動頻度とに基づき、各パーティションP1, P2に割り当てられる複数のリソースの全ての組み合わせのそれぞれについて平均距離、つまり平均メモリレイテンシを算出する。

[0042] つまり、算出手段33は、前記組み合わせ毎に、リソース間データ移動頻度テーブル31bにデータ移動頻度として記録された各メモリMnに対する各CPU Cikのアクセス回数F(Cik, Mn)と、アクセスレイテンシテーブル31aに距離情報として定義された対応メモリレイテンシ、つまりノード間アクセスレイテンシD(Cik, Mn)＝distance(i,n)との積の総和を算出する。そして、算出手段33は、当該積の総和をアクセス回数の総和で除算した値を、当該組み合わせについての平均距離として算出する。この後、算出手段33は、複数のリソースの全ての組み合わせのうち、その組み合わせについて算出された平均距離が最小になるリソースの組み合わせを最適配分として選択する。

[0043] ここで、上述のように、パーティションP1には8個のCPUおよび2個のメモリが割り当

てられるとともに、パーティションP2には4個のCPUおよび1個のメモリが割り当てられるものとする。このような場合に、例えば、パーティションP2についてみると、ノードN1～N8における32個のCPUおよび8個のメモリから、4個のCPUおよび1個のメモリを選択してパーティションP2に割り当てるとすると、多数の組み合わせが考えられる。算出手段33は、その各組み合わせについての平均距離つまり平均メモリレイテンシを、以下のようにアクセスレイテンシテーブル31a，リソース間データ移動頻度テーブル31bのデータに基づいて算出する。

[0044] ここでは、簡単のために、図5に示すごとく4個のCPU C23，C24，C61，C62および1個のメモリM2を割り当てられたパーティションP2についての平均メモリレイテンシを、図2および図3にそれぞれ示すアクセスレイテンシテーブル31aおよびリソース間データ移動頻度31bのデータに基づいて算出する場合について具体的に説明する。

[0045] まず、パーティションP2でのメモリアクセス回数の総数は、図3に示すノード間距離テーブルテーブル31bに記録された数値に基づき、

F(C23, M2)+F(C24, M2)+F(C61, M2)+F(C62, M2)

＝1000＋4000＋3000＋2000

＝10000

となる。

[0046] 従って、図5に示すリソース組み合わせのパーティションP2での平均メモリレイテンシは、図2に示すアクセスレイテンシテーブル31aに記録されたメモリレイテンシおよび図3に示すリソース間データ移動頻度テーブル31bに記録されたアクセス回数に基づいて算出される。

[図5に示すパーティションP2の平均メモリレイテンシ]

＝Σ'D(C,M)*F(C,M)/10000

＝{D(C23,M2)*F(C23,M2)+D(C24,M2)*F(C24,M2)

+D(C61,M2)*F(C61,M2)+D(C62,M2)*F(C62,M2)}／10000

＝(50*1000+50*4000+200*3000+200*2000)／10000

＝1250000／10000

＝125 nsec

なお、Σ'は、パーティションP2に属するするCPUおよびメモリの全ての組み合わせについて算出されるD(C,M)*F(C,M)の総和を意味している。

[0047] これに対し、パーティションP2に割り当てられるリソースのうち、図5に示すCPU C23, C24およびメモリM2が、図6に示すように、それぞれCPU C63, C64およびメモリM6に置き換えられた場合の平均メモリレイテンシは、以下のように算出される。このとき、メモリM6に対するCPU C63, C64のアクセス回数は、それぞれ、メモリM2に対するCPU C23, C24と同じ値とする。つまり、

F(C63,M6)＝F(C23,M2)＝1000

F(C64,M6)＝F(C24,M2)＝4000

であり、図6に示すパーティションP2での平均メモリレイテンシは、以下のように算出される。

[0048] [図6に示すパーティションP2の平均メモリレイテンシ]

＝Σ'D(C,M)*F(C,M)/10000

＝{D(C63,M6)*F(C63,M6)+D(C64,M6)*F(C64,M6)

+D(C61,M6)*F(C61,M6)+D(C62,M6)*F(C62,M6)}／10000

＝(50*1000+50*4000+50*3000+50*2000)／10000

＝1250000／10000

＝50 nsec

[0049] 図6に示すパーティションP2の平均メモリレイテンシは、図5に示すパーティションP2の平均メモリレイテンシの40％(＝50/125)に減少し、システム1の大幅な性能改善が見込まれる。

[0050] 上述のようにして、算出手段33は、全てのリソースの組み合わせに対し平均距離を算出し、その平均距離を最小にするリソース組み合わせを、最適なパーティション構成(最適配分)として求める。

[0051] つまり、一般的に記載すると、算出手段33は、τ:{CPUの集合}→{CPUの集合}, ρ:{メモリの集合}→{メモリの集合}に対して、

平均距離AvgD(τ, ρ)＝Σ'D(τ(C), ρ(M))*F(C,M)／10000

を計算し、これを最小にするτ, ρを求める。その結果得られたτ(パーティションP2のCPUの集合)およびρ(パーティションP2のメモリの集合)が、平均レイテンシを最小にする、最適なパーティションP2のリソース構成(リソース配分)になる。なお、Σ'は、上述と同様、パーティションP2に属するするCPUおよびメモリの全ての組み合わせについて算出されるD(τ(C), ρ(M))*F(C,M)の総和を意味している。

[0052] なお、パーティションP1およびパーティションP2にそれぞれ属するリソースは、他のパーティションに属すことはできない。従って、実際には、算出手段33は、パーティションP1およびパーティションP2のそれぞれに属する、12個のCPUおよび3個のメモリの組み合わせを逐次選択し、各組み合わせについて、上述と同様にして平均メモリレイテンシを算出し、その平均距離に基づいて最適配分、つまり平均距離が最小となるリソース組み合わせを選択することになる。

[0053] 配分手段34は、各パーティションP1, P2に対するリソース配分状態が算出手段33によって算出された最適配分の状態になるように、各パーティション管理部21, 22を介して各パーティションP1, P2に対しCPU CikおよびメモリMnを配分するものである。このとき、配分手段34は、各パーティション管理部21, 22に対し最適配分に関する情報を通知し、各パーティション管理部21, 22における、各パーティションP1, P2についての条件テーブルの内容を書き換え変更する。ここで、配分手段34から各パーティション管理部21, 21に通知される最適配分に関する情報は、各パーティションP1, P2に含まれるべきCPU CikおよびメモリMnを指定する情報である。

[0054] この配分手段34による配分変更処理は、深夜など、システム1の使用頻度の低い時間帯に、変更対象リソースの属するノードを含むボードの電源を落とした上で実行される。その配分変更処理に際しては、各パーティション管理部21, 22における条件テーブルの書換が行なわれるとともに、変更対象のCPU内データやメモリの記憶データを変更後のCPUやメモリに移動する処理が実行される。これにより、各パーティションP1, P2内のリソースの構成が最適なパーティション構成に変更される。ただし、本発明は、このような配分変更処理に限定されるものでなく、ボードの活性交換等によって配分変更処理を行なってもよい。

[0055] この配分手段34によるリソース配分変更は、現状のパーティション構成での平均距

離よりも小さい平均距離のパーティション構成が存在する場合に実行される。特に、その際、配分変更後のパーティション構成によって、現状、つまり配分変更前のパーティション構成よりも所定基準以上の性能改善が得られる場合にリソース配分変更が実行される。より具体的には、上述のごとく算出される性能改善率[配分変更後の平均距離]／[配分変更前の平均距離]が所定値以下となる場合に、上記リソース配分変更を実行することが好ましい。

[0056] なお、上述した算出手段33および配分手段34による処理は、例えば、新規パーティション追加，所定時間経過，ユーザ（サーバ管理者）のリクエストなどをトリガとして、深夜などのシステム1の使用頻度の低い時間帯に実行される。

[0057] また、算出手段33は、最適配分となるリソース組み合わせが複数存在する場合には、後述する配分手段34によるリソース配分を行なう際にリソース配分変更量が最も少なくなる、リソース組み合わせを、最適配分として選択することが望ましい。これにより、リソース配分変更に伴う、各パーティション管理部21，22における条件テーブルの書換変更や、CPU／メモリにおけるデータ移動などの処理を最小限に抑え、効率的に配分変更を行なうことができる。

[0058] 次に、図4に示すフローチャート（ステップS1～S8）に従い、上述のごとく構成された本実施形態のマルチプロセッサシステム1（サーバ管理装置30）の動作について、図5および図6を参照しながら説明する。なお、図5および図6は、いずれも、図1に示すシステム1におけるパーティション分割の具体的な最適化動作例を説明するためのもので、図5はシステム1の最適化前の状態を示す図、図6はシステム1の最適化後の状態を示す図である。

[0059] ここで、各々のパーティションが同一量のリソースを使用していたとしても、各パーティションP1，P2におけるリソースの組み合わせによっては、システム1の性能は大きく異なる。そこで、本実施形態では、各パーティションP1，P2に対しリソースを再配分し、システム1の処理性能を最適化する。

[0060] 図5に示す例では、8個のCPU C11，C12，C13，C14，C21，C22，C51，C52および2個のメモリM1，M5がパーティションP1に属し、4個のCPU C23，C24，C61，C62および1個のメモリM2がパーティションP2に属している。つまり、パーティショ

ンP1に属するCPUは3つのノードN1, N2, N5に分散配置され、パーティションP1に属するメモリM1, M5は2つのノードN1, N5に分散配置されている。また、パーティションP2に属するCPUは2つのノードN2, N6に分散配置されている。このように同一パーティションにおけるCPUとメモリとが異なるノードに分散配置されていると、ノード間通信を行なう必要があり、メモリレイテンシが悪化することになる。例えば、ノードN6に属するCPU C61は、他ノードN2におけるメモリM2にアクセスする必要があり、メモリレイテンシが悪化してしまう。

[0061] これに対し、図6に示す例は、図5に示すごとく配分されたリソースに対し、サーバ管理装置30が、例えば図2や図3に示すアクセスレイテンシテーブル31aや, リソース間データ移動頻度31bに用い図4に示す手順で最適化処理を行なった結果得られた、最適化後の状態である。この図6に示す例では、8個のCPU C11, C12, C13, C14, C21, C22, C23, C24および2個のメモリM1, M2がパーティションP1に属し、4個のCPU C61, C62, C63, C64および1個のメモリM6がパーティションP2に属している。

[0062] このように再配分を行なうことにより、パーティションP2に属するCPUおよびメモリは一つのノードN6内に配置される。従って、CPUがメモリアクセスを行なう際には、必ず自ノードN6のメモリM6にアクセスすることになり、メモリレイテンシは最小になる。

[0063] また、パーティションP1に属するCPUおよびメモリは、同一のクロスバスイッチCB1に収容された2つのノードN1, N2内に配置される。従って、この場合も、CPUがメモリアクセスを行なう際には、自ノードのメモリもしくは同一クロスバスイッチCB1に収容された他のノードのメモリにアクセスすることになり、メモリレイテンシは最小になる。

[0064] さて、本実施形態のサーバ管理装置30が動作を開始すると、図4に示すように、まず、アクセスレイテンシテーブル31aを初期化してから(ステップS1)、システム1の運用を開始する(ステップS2)。なお、アクセスレイテンシテーブル31aの初期化では、本システム1のリソース群10に対応するアクセスレイテンシテーブル31aが記憶部31に登録格納される。また、動作開始時の初期化に際しては、各ノードNiにおけるテーブルTiの初期化(クリア)も行なわれる。

[0065] この後、収集手段32によって、リソース間のデータ移動情報の収集が開始される(

ステップS3)。この収集処理では、各ノードNiにおけるテーブルTiに記録されたデータ移動情報が、バスB1，B2を介して各テーブルTiから収集され、リソース間データ移動頻度テーブル31bに登録される。これにより、リソース間データ移動頻度テーブル31bには、リソース群10において、どのCPUがどのメモリに対して、何回、リクエストを発行したかに関する情報、つまり通信回数／データ移動回数／リードアクセス回数などが、データ移動頻度として登録される。情報分析のトリガが発生するまでは、上述のようなリソース間のデータ移動情報の収集が継続される(ステップS4のNOルート)。

[0066] そして、例えば、新規パーティション追加，所定時間経過，ユーザ(サーバ管理者)のリクエストなどの何らかのトリガが発生すると(ステップS4のYESルート)、算出手段33によって、アクセスレイテンシテーブル31aの距離情報、つまりアクセスレイテンシ／メモリレイテンシとリソース間データ移動頻度テーブル31bのデータ移動頻度とに基づき、各パーティションP1，P2に対する、リソースの最適配分が算出される(ステップS5)。つまり、算出手段33によって、上述のように、全てのリソースの組み合わせに対し平均距離が算出され、その平均距離を最小にするリソース組み合わせが、最適なパーティション構成(最適配分)として求められる。

[0067] この後、サーバ管理装置30では、算出手段33によって得られた最適なパーティション構成(最適配分)について、上述のごとき性能改善率[配分変更後の平均距離]／[配分変更前の平均距離]が算出される。そして、その性能改善率が所定値以下であるか否かが判断される(ステップS6)。

[0068] 性能改善率が所定値を超えている場合、現状のパーティション構成よりもよいパーティション構成が存在しないと判断され(ステップS6のNOルート)、現状のパーティション構成が維持される。つまり、サーバ管理装置30は、リソース間のデータ移動情報の収集を継続し、ステップS4の処理へ移行する。

[0069] 一方、性能改善率が所定値以下である場合、現状のパーティション構成よりもよいパーティション構成が存在すると判断され(ステップS6のYESルート)、配分手段34による配分変更処理が実行される(ステップS7)。

[0070] その際、例えば図5に示すパーティション構成から図6に示すパーティション構成へ

配分変更する場合には、変更対象となるノードN1，N2，N5，N6の動作が停止される。そして、配分手段34によって、各パーティション管理部21，22における、各パーティションP1，P2についての条件テーブルの内容が書き換えられるとともに、変更対象のCPU内データやメモリの記憶データが変更後のCPUやメモリに移動される。このとき、メモリM2の記憶データがメモリM6に移動されるとともに、CPU C23，C24の内部データがCPU C63，C64に移動される。その後、メモリM5の記憶データがメモリM2に移動されるとともに、CPU C51，C52の内部データがCPU C23，C24に移動される。このようなデータ移動処理を行なってから、ノードN1，N2，N5，N6の電源が投入され、各パーティションP1，P2内のリソースの構成が最適なパーティション構成（最適配分）に変更される。

[0071] パーティション構成の変更を終了すると、リソース間データ移動頻度テーブル31bやテーブルT1，T2，T5，T6において、変更対象となったリソースに係るデータ移動頻度やデータ移動情報といった情報がクリアされ（ステップS8）、サーバ管理装置30はステップS3の処理へ移行する。

[0072] このように、本発明の一実施形態としてのマルチプロセッサシステム1やサーバ管理装置30によれば、マルチプロセッサシステム1内のリソース間の距離情報とデータ移動頻度とに基づいて、各パーティションに対するリソースの最適配分が統計的に算出され、その最適配分に応じたリソース配分が行なわれる。これにより、システム1の特性を意識したリソース配分が実現され、パーティション分割、つまりパーティションへのリソースの割当が最適化され、システム全体の処理性能が大幅に向上する。つまり、システム1のNUMA特性を考慮したリソースの再配置を行なうことにより、同一リソースを使用した場合の処理性能を最大化することができる。

[0073] なお、本発明は上述した実施形態に限定されるものではなく、本発明の趣旨を逸脱しない範囲で種々変形して実施することができる。

また、上述した記憶部（第1テーブル記憶手段，第2テーブル記憶手段）31，収集手段32，算出手段33および配分手段34としての機能（各手段の全部もしくは一部の機能）は、コンピュータ（CPU，情報処理装置，各種端末を含む）が所定のアプリケーションプログラム（マルチプロセッサシステム用管理プログラム）を実行することによ

って実現される。

[0074] そのプログラムは、例えばフレキシブルディスク，CD（CD－ROM，CD－R，CD－RWなど），DVD（DVD－ROM，DVD－RAM，DVD－R，DVD－RW，DVD＋R，DVD＋RW，ブルーレイディスクなど）等のコンピュータ読取可能な記録媒体に記録された形態で提供される。この場合、コンピュータはその記録媒体からマルチプロセッサシステム用管理プログラムを読み取って内部記憶装置または外部記憶装置に転送し格納して用いる。また、そのプログラムを、例えば磁気ディスク，光ディスク，光磁気ディスク等の記憶装置（記録媒体）に記録しておき、その記憶装置から通信回線を介してコンピュータに提供するようにしてもよい。

[0075] ここで、コンピュータとは、ハードウエアとOS（オペレーティングシステム）とを含む概念であり、OSの制御の下で動作するハードウエアを意味している。また、OSが不要でアプリケーションプログラム単独でハードウェアを動作させるような場合には、そのハードウェア自体がコンピュータに相当する。ハードウエアは、少なくとも、CPU等のマイクロプロセッサと、記録媒体に記録されたプログラムを読み取るための手段とをそなえている。上記分散型ストレージシステム用制御プログラムとしてのアプリケーションプログラムは、上述のようなコンピュータに、手段31～34としての機能を実現させるプログラムコードを含んでいる。また、その機能の一部は、アプリケーションプログラムではなくOSによって実現されてもよい。

[0076] さらに、本実施形態における記録媒体としては、上述したフレキシブルディスク，CD，DVD，磁気ディスク，光ディスク，光磁気ディスクのほか、ICカード，ROMカートリッジ，磁気テープ，パンチカード，コンピュータの内部記憶装置（RAMやROMなどのメモリ），外部記憶装置等や、バーコードなどの符号が印刷された印刷物等の、コンピュータ読取可能な種々の媒体を利用することもできる。

請求の範囲

[1] 複数のパーティションのいずれか一つに対し単独で割当可能な、複数のリソースと、

該複数のパーティションのそれぞれに属するリソースを管理する複数のパーティション管理部と、

該複数のリソースおよび該複数のパーティション管理部を管理するシステム管理部とを有し、

該システム管理部は、

該複数のリソース相互間の距離に係る距離情報を定義する第1テーブルを記憶する第1テーブル記憶手段と、

該複数のリソース相互間のデータ移動情報を収集する収集手段と、

該収集手段によって収集された前記データ移動情報に基づく該複数のリソース相互間のデータ移動頻度を保持する第2テーブルを記憶する第2テーブル記憶手段と、

該第1テーブルの距離情報と該第2テーブルのデータ移動頻度とに基づき、各パーティションに対する、該複数のリソースの最適配分を算出する算出手段と、

該複数のパーティションに対する該複数のリソースの配分状態が該算出手段によって算出された前記最適配分の状態になるように、該複数のパーティション管理部を介して該複数のパーティションに該複数のリソースを配分する配分手段とを有していることを特徴とする、マルチプロセッサシステム。

[2] 該第1テーブルにおける前記距離情報として、各リソースの属するノード間のアクセスレイテンシが定義されていることを特徴とする、請求項1に記載のマルチプロセッサシステム。

[3] 該第2テーブルにおける前記データ移動頻度として、前記複数のリソース相互間のデータ移動回数が記録更新されることを特徴とする、請求項1または請求項2に記載のマルチプロセッサシステム。

[4] 該複数のリソースとして複数の演算処理部と複数のメモリとが含まれ、前記データ移動回数として各演算処理部と各メモリとの間の通信回数が記録更新されることを特徴

とする、請求項3に記載のマルチプロセッサシステム。

[5] 該算出手段は、該第1テーブルの距離情報と該第2テーブルのデータ移動頻度とに基づき、各パーティションに割り当てられる該複数のリソースの全ての組み合わせのそれぞれについて平均距離を算出し、当該平均距離が最小になるリソースの組み合わせを前記最適配分として選択することを特徴とする、請求項1～請求項4のいずれか一項に記載のマルチプロセッサシステム。

[6] 該算出手段は、前記最適配分として複数の組み合わせが存在する場合には、該配分手段によるリソース配分を行なう際に配分変更量の最も少ない組み合わせを前記最適配分として選択することを特徴とする、請求項5に記載のマルチプロセッサシステム。

[7] 該複数のリソースとして複数の演算処理部と複数のメモリとが含まれ、

該算出手段は、前記組み合わせ毎に、該第2テーブルに前記データ移動頻度として記録された各メモリに対する各演算処理部のアクセス回数と、該第1テーブルに前記距離情報として定義された対応メモリレイテンシとの積の総和を算出し、当該積の総和を前記アクセス回数の総和で除算した値を、当該組み合わせについての前記平均距離として算出することを特徴とする、請求項5または請求項6に記載のマルチプロセッサシステム。

[8] 複数のパーティションのいずれか一つに対し単独で割当可能な、複数のリソースと、該複数のパーティションのそれぞれに属するリソースを管理する複数のパーティション管理部とを有するマルチプロセッサシステムにおいて、該複数のリソースおよび該複数のパーティション管理部を管理するマルチプロセッサシステム用管理装置であって、

該複数のリソース相互間の距離に係る距離情報を定義する第1テーブルを記憶する第1テーブル記憶手段と、

該複数のリソース相互間のデータ移動情報を収集する収集手段と、

該収集手段によって収集された前記データ移動情報に基づく該複数のリソース相互間のデータ移動頻度を保持する第2テーブルを記憶する第2テーブル記憶手段と、

該第1テーブルの距離情報と該第2テーブルのデータ移動頻度とに基づき、各パーティションに対する、該複数のリソースの最適配分を算出する算出手段と、

該複数のパーティションに対する該複数のリソースの配分状態が該算出手段によって算出された前記最適配分の状態になるように、該複数のパーティション管理部を介して該複数のパーティションに該複数のリソースを配分する配分手段とを有していることを特徴とする、マルチプロセッサシステム用管理装置。

[9] 該第1テーブルにおける前記距離情報として、各リソースの属するノード間のアクセスレイテンシが定義されていることを特徴とする、請求項8に記載のマルチプロセッサシステム用管理装置。

[10] 該第2テーブルにおける前記データ移動頻度として、前記複数のリソース相互間のデータ移動回数が記録更新されることを特徴とする、請求項8または請求項9に記載のマルチプロセッサシステム用管理装置。

[11] 該複数のリソースとして複数の演算処理部と複数のメモリとが含まれ、前記データ移動回数として各演算処理部と各メモリとの間の通信回数が記録更新されることを特徴とする、請求項10に記載のマルチプロセッサシステム用管理装置。

[12] 該算出手段は、該第1テーブルの距離情報と該第2テーブルのデータ移動頻度とに基づき、各パーティションに割り当てられる該複数のリソースの全ての組み合わせのそれぞれについて平均距離を算出し、当該平均距離が最小になるリソースの組み合わせを前記最適配分として選択することを特徴とする、請求項8～請求項11のいずれか一項に記載のマルチプロセッサシステム用管理装置。

[13] 該算出手段は、前記最適配分として複数の組み合わせが存在する場合には、該配分手段によるリソース配分を行なう際に配分変更量の最も少ない組み合わせを前記最適配分として選択することを特徴とする、請求項12に記載のマルチプロセッサシステム用管理装置。

[14] 該複数のリソースとして複数の演算処理部と複数のメモリとが含まれ、

該算出手段は、前記組み合わせ毎に、該第2テーブルに前記データ移動頻度として記録された各メモリに対する各演算処理部のアクセス回数と、該第1テーブルに前記距離情報として定義された対応メモリレイテンシとの積の総和を算出し、当該積の

総和を前記アクセス回数の総和で除算した値を、当該組み合わせについての前記平均距離として算出することを特徴とする、請求項12または請求項13に記載のマルチプロセッサシステム用管理装置。

[15] 複数のパーティションのいずれか一つに対し単独で割当可能な、複数のリソースと、該複数のパーティションのそれぞれに属するリソースを管理する複数のパーティション管理部とを有するマルチプロセッサシステムにおいて、該複数のリソースおよび該複数のパーティション管理部を管理するマルチプロセッサシステム用管理装置として、コンピュータを機能させるプログラムを記録したコンピュータ読取可能な記録媒体であって、

該プログラムは、

該複数のリソース相互間の距離に係る距離情報を定義する第1テーブルを記憶する第1テーブル記憶手段、

該複数のリソース相互間のデータ移動情報を収集する収集手段、

該収集手段によって収集された前記データ移動情報に基づく該複数のリソース相互間のデータ移動頻度を保持する第2テーブルを記憶する第2テーブル記憶手段、

該第1テーブルの距離情報と該第2テーブルのデータ移動頻度とに基づき、各パーティションに対する、該複数のリソースの最適配分を算出する算出手段、および、

該複数のパーティションに対する該複数のリソースの配分状態が該算出手段によって算出された前記最適配分の状態になるように、該複数のパーティション管理部を介して該複数のパーティションに該複数のリソースを配分する配分手段、として、該コンピュータを機能させることを特徴とする、マルチプロセッサシステム用管理プログラムを記録したコンピュータ読取可能な記録媒体。

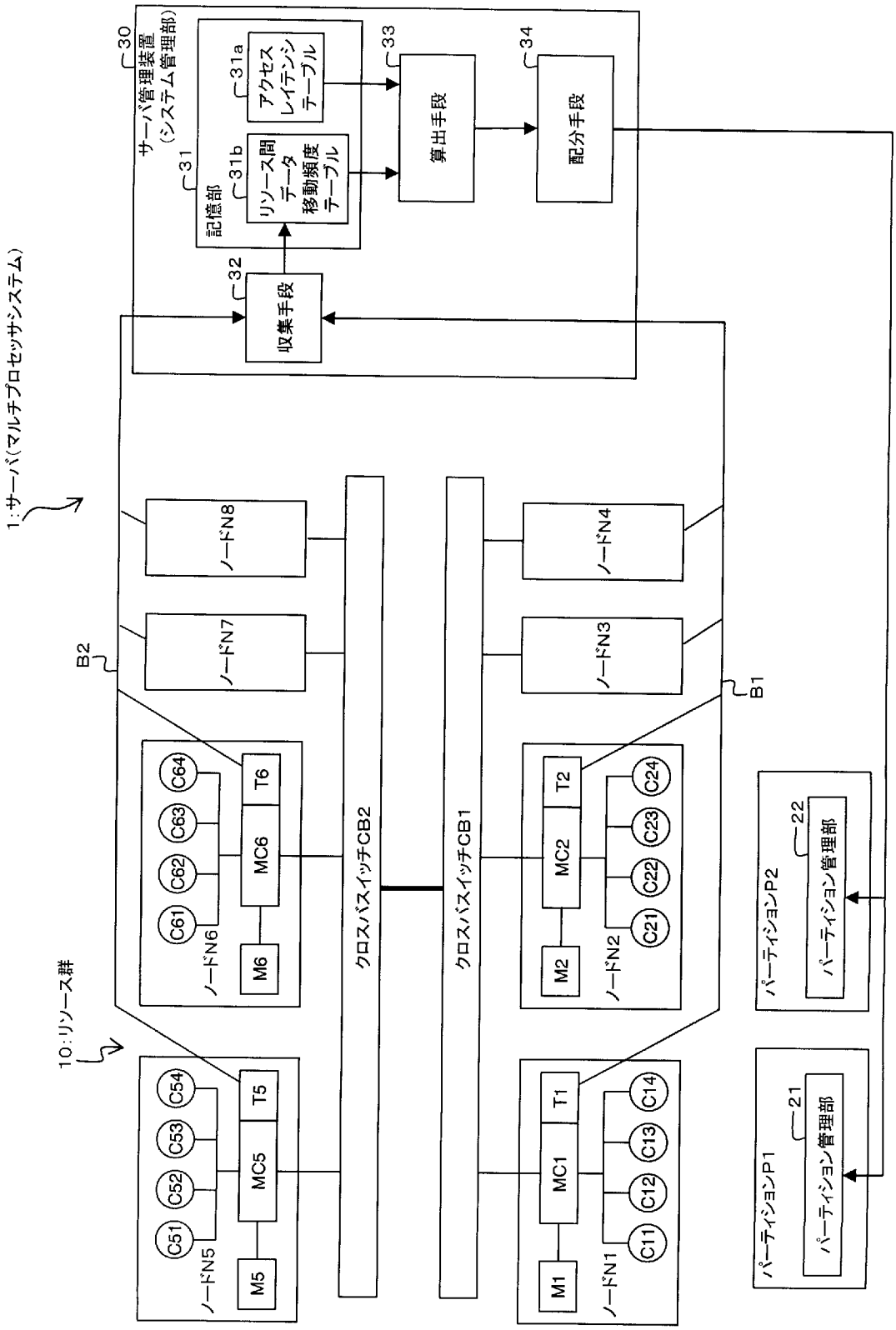
[16] 該プログラムは、該算出手段として該コンピュータを機能させる際に、該第1テーブルの距離情報と該第2テーブルのデータ移動頻度とに基づき、各パーティションに割り当てられる該複数のリソースの全ての組み合わせのそれぞれについて平均距離を算出し、当該平均距離が最小になるリソースの組み合わせを前記最適配分として選択するように、該コンピュータを機能させることを特徴とする、請求項15に記載のマルチプロセッサシステム用管理プログラムを記録したコンピュータ読取可能な記録媒体

。

[17] 該複数のリソースとして複数の演算処理部と複数のメモリとが含まれ、

該プログラムは、該算出手段として該コンピュータを機能させる際に、前記組み合わせ毎に、該第2テーブルに前記データ移動頻度として記録された各メモリに対する各演算処理部のアクセス回数と、該第1テーブルに前記距離情報として定義された対応メモリレイテンシとの積の総和を算出し、当該積の総和を前記アクセス回数の総和で除算した値を、当該組み合わせについての前記平均距離として算出するように、該コンピュータを機能させることを特徴とする、請求項16に記載のマルチプロセッサシステム用管理プログラムを記録したコンピュータ読取可能な記録媒体。

[図1]


1:サーバ(マルチプロセッサシステム)
10:リソース群
30
サーバ管理装置
(システム管理部)
31
記憶部
31a
アクセス
レイテンシ
テーブル
31b
リソース間
データ
移動頻度
テーブル
32
収集手段
33
算出手段
34
配分手段
B1
B2
ノードN1
ノードN2
ノードN3
ノードN4
ノードN5
ノードN6
ノードN7
ノードN8
M1
M2
M5
M6
MC1
MC2
MC5
MC6
T1
T2
T5
T6
C11
C12
C13
C14
C21
C22
C23
C24
C51
C52
C53
C54
C61
C62
C63
C64
クロスバスイッチCB1
クロスバスイッチCB2
パーティションP1
パーティション管理部
21
パーティションP2
パーティション管理部
22

[図2]

31a:アクセスレイテンシテーブル(第1テーブル)

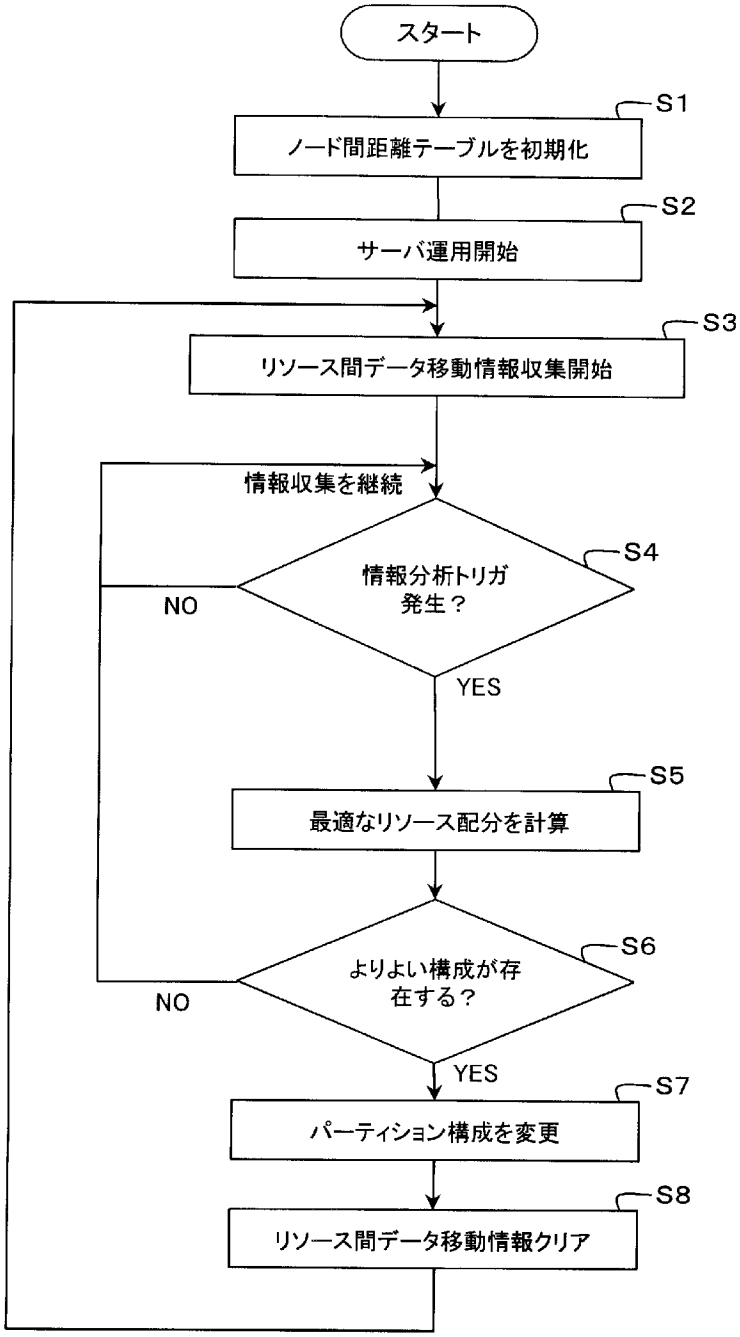
	N1	N2	N3	N4	N5	N6	N7	N8
N1	50	100	100	100	200	200	200	200
N2	100	50	100	100	200	200	200	200
N3	100	100	50	100	200	200	200	200
N4	100	100	100	50	200	200	200	200
N5	200	200	200	200	50	100	100	100
N6	200	200	200	200	100	50	100	100
N7	200	200	200	200	100	100	50	100
N8	200	200	200	200	100	100	100	50

[図3]

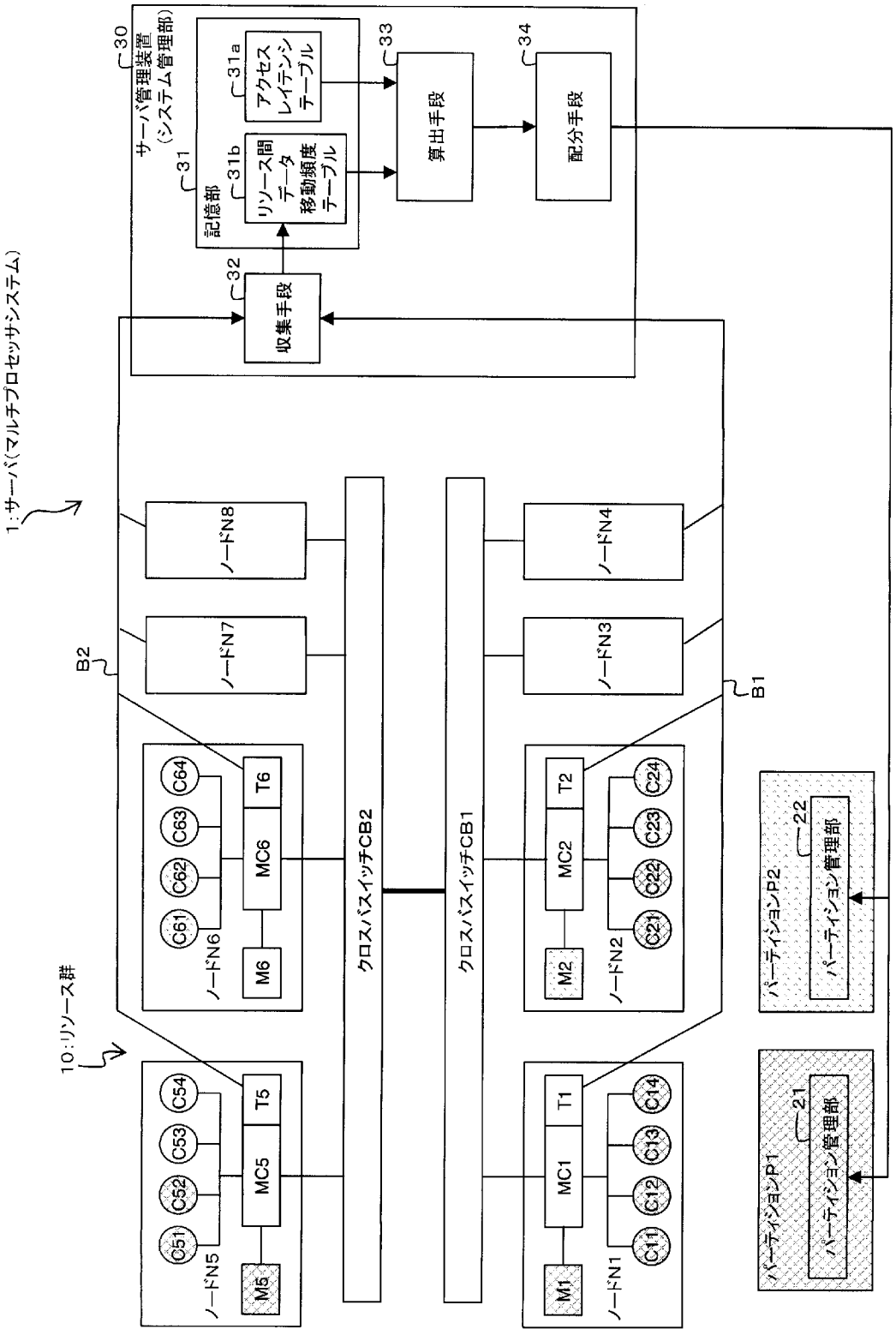
31b:リソース間データ移動頻度テーブル(第2テーブル)

#CPU	#NODE	#PART	M1	M2	M3	M4	M5	M6	M7	M8
C11	1	1	1000	0	0	0	500	0	0	0
C12	1	1	2000	0	0	0	1000	0	0	0
C13	1	1	2000	0	0	0	1500	0	0	0
C14	1	1	1000	0	0	0	500	0	0	0
C21	2	1	1000	0	0	0	500	0	0	0
C22	2	1	3000	0	0	0	1000	0	0	0
C23	2	2	0	1000	0	0	0	0	0	0
C24	2	2	0	4000	0	0	0	0	0	0
・										
・										
C51	5	1	1000	0	0	0	2000	0	0	0
C52	5	1	2000	0	0	0	1500	0	0	0
C53	5	-	0	0	0	0	0	0	0	0
C54	5	-	0	0	0	0	0	0	0	0
C61	6	2	0	3000	0	0	0	0	0	0
C62	6	2	0	2000	0	0	0	0	0	0
C63	6	-	0	0	0	0	0	0	0	0
C64	6	-	0	0	0	0	0	0	0	0
・										
・										

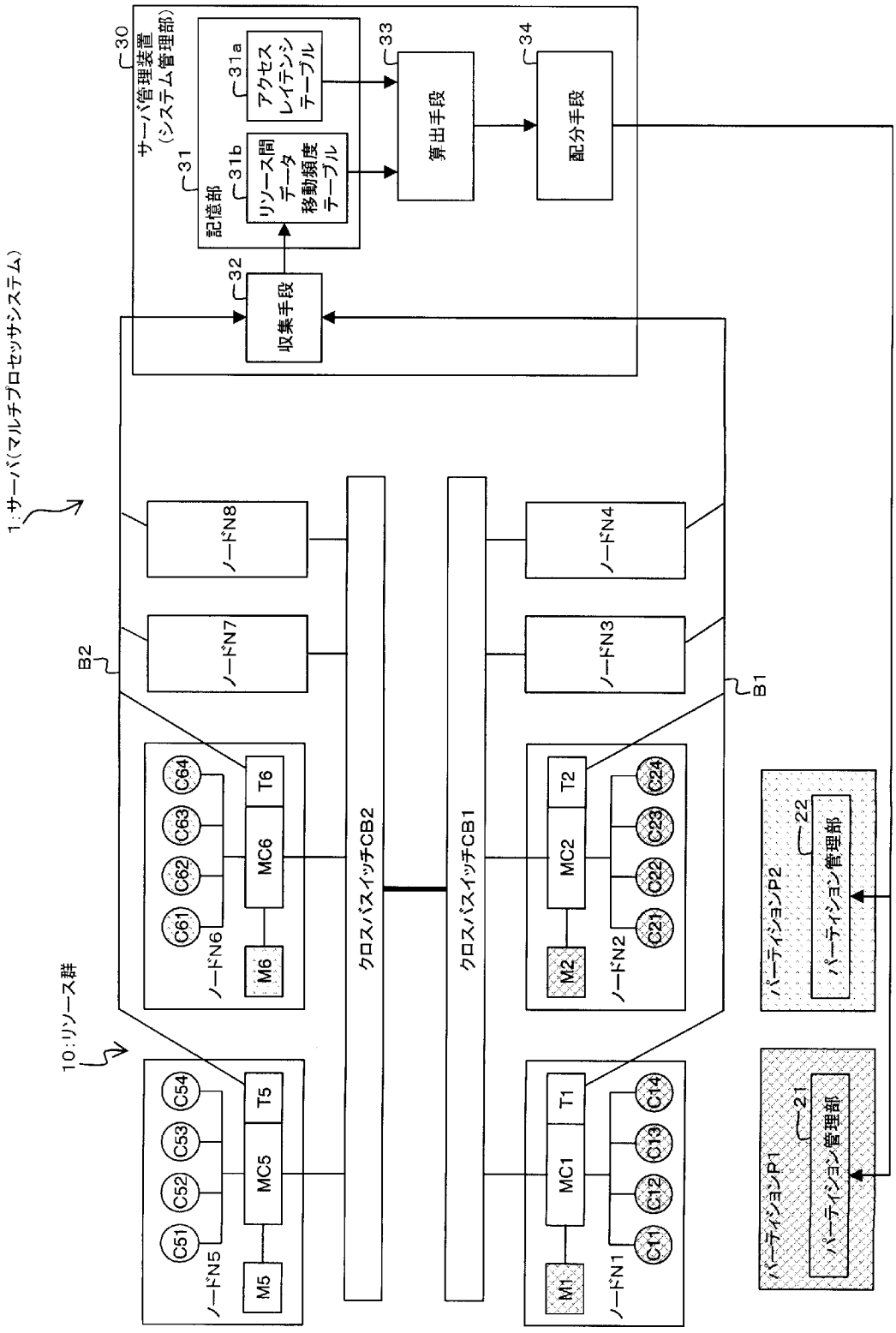
[図4]


スタート
S1
ノード間距離テーブルを初期化
S2
サーバ運用開始
S3
リソース間データ移動情報収集開始
情報収集を継続
S4
情報分析トリガ
発生？
NO
YES
S5
最適なリソース配分を計算
S6
よりよい構成が存
在する？
NO
YES
S7
パーティション構成を変更
S8
リソース間データ移動情報クリア

[図5]


1:サーバ(マルチプロセッサシステム)
10:リソース群
30
サーバ管理装置
(システム管理部)
31
記憶部
31a
アクセス
レイテンシ
テーブル
31b
リソース間
データ
移動頻度
テーブル
32
収集手段
33
算出手段
34
配分手段
B1
B2
ノードN1
ノードN2
ノードN3
ノードN4
ノードN5
ノードN6
ノードN7
ノードN8
M1
M2
M5
M6
MC1
MC2
MC5
MC6
T1
T2
T5
T6
C11
C12
C13
C14
C21
C22
C23
C24
C51
C52
C53
C54
C61
C62
C63
C64
クロスバスイッチCB1
クロスバスイッチCB2
パーティションP1
パーティションP2
21
22
パーティション管理部
パーティション管理部

[図6]


1:サーバ(マルチプロセッサシステム)
10:リソース群
サーバ管理装置
(システム管理部)
30
記憶部
31
31a
アクセス
レイテンシ
テーブル
31b
リソース間
データ
移動頻度
テーブル
32
収集手段
33
算出手段
34
配分手段
B1
B2
ノードN1
ノードN2
ノードN3
ノードN4
ノードN5
ノードN6
ノードN7
ノードN8
M1
M2
M5
M6
MC1
MC2
MC5
MC6
T1
T2
T5
T6
C11
C12
C13
C14
C21
C22
C23
C24
C51
C52
C53
C54
C61
C62
C63
C64
クロスバスイッチCB1
クロスバスイッチCB2
パーティションP1
パーティションP2
21
22
パーティション管理部
パーティション管理部

INTERNATIONAL SEARCH REPORT

International application No.
PCT/JP2008/063977

A. CLASSIFICATION OF SUBJECT MATTER
G06F9/50(2006.01)i, *G06F9/46*(2006.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
G06F9/50, G06F9/46

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched
Jitsuyo Shinan Koho 1922-1996 Jitsuyo Shinan Toroku Koho 1996-2008
Kokai Jitsuyo Shinan Koho 1971-2008 Toroku Jitsuyo Shinan Koho 1994-2008

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	JP 2007-257097 A (NEC Corp.), 04 October, 2007 (04.10.07), Full text; all drawings & US 2007/0226449 A1	1-17
A	JP 2006-003972 A (NEC Corp.), 05 January, 2006 (05.01.06), Par. No. [0013]; Figs. 1, 2 (Family: none)	1-17
A	JP 2004-199561 A (Hitachi, Ltd.), 15 July, 2004 (15.07.04), Par. Nos. [0020] to [0030]; Fig. 1 & US 2004/0143664 A1	1-17

☒ Further documents are listed in the continuation of Box C. ☐ See patent family annex.

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
22 September, 2008 (22.09.08)	30 September, 2008 (30.09.08)
Name and mailing address of the ISA/ Japanese Patent Office Facsimile No.	Authorized officer Telephone No.

INTERNATIONAL SEARCH REPORT

International application No.
PCT/JP2008/063977

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	JP 2002-202959 A (Hitachi, Ltd.), 19 July, 2002 (19.07.02), Par. Nos. [0018] to [0032]; Fig. 1 & US 2002/0087611 A1	1-17

国際調査報告

国際出願番号　ＰＣＴ／ＪＰ２００８／０６３９７７

Ａ．　発明の属する分野の分類（国際特許分類（ＩＰＣ））

Int.Cl.　G06F9/50(2006.01)i, G06F9/46(2006.01)i

Ｂ．　調査を行った分野

調査を行った最小限資料（国際特許分類（ＩＰＣ））

Int.Cl.　G06F9/50, G06F9/46

最小限資料以外の資料で調査を行った分野に含まれるもの

日本国実用新案公報	１９２２－１９９６年
日本国公開実用新案公報	１９７１－２００８年
日本国実用新案登録公報	１９９６－２００８年
日本国登録実用新案公報	１９９４－２００８年

国際調査で使用した電子データベース（データベースの名称、調査に使用した用語）

Ｃ．　関連すると認められる文献

引用文献のカテゴリー＊	引用文献名　及び一部の箇所が関連するときは、その関連する箇所の表示	関連する請求の範囲の番号
X	JP 2007-257097 A (日本電気株式会社) 2007.10.04, 全文, 全図 & US 2007/0226449 A1	１－１７
A	JP 2006-003972 A (日本電気株式会社) 2006.01.05, 段落（００１３）, 第１図, 第２図（ファミリーなし）	１－１７
A	JP 2004-199561 A (株式会社日立製作所) 2004.07.15, 段落（００２０）－（００３０）, 第１図	１－１７

☑　Ｃ欄の続きにも文献が列挙されている。

☐　パテントファミリーに関する別紙を参照。

＊　引用文献のカテゴリー

「Ａ」特に関連のある文献ではなく、一般的技術水準を示すもの

「Ｅ」国際出願日前の出願または特許であるが、国際出願日以後に公表されたもの

「Ｌ」優先権主張に疑義を提起する文献又は他の文献の発行日若しくは他の特別な理由を確立するために引用する文献（理由を付す）

「Ｏ」口頭による開示、使用、展示等に言及する文献

「Ｐ」国際出願日前で、かつ優先権の主張の基礎となる出願の日の後に公表された文献

「Ｔ」国際出願日又は優先日後に公表された文献であって出願と矛盾するものではなく、発明の原理又は理論の理解のために引用するもの

「Ｘ」特に関連のある文献であって、当該文献のみで発明の新規性又は進歩性がないと考えられるもの

「Ｙ」特に関連のある文献であって、当該文献と他の１以上の文献との、当業者にとって自明である組合せによって進歩性がないと考えられるもの

「＆」同一パテントファミリー文献

国際調査を完了した日 ２２．０９．２００８	国際調査報告の発送日 ３０．０９．２００８
国際調査機関の名称及びあて先 日本国特許庁（ＩＳＡ／ＪＰ） 郵便番号１００－８９１５ 東京都千代田区霞が関三丁目４番３号	特許庁審査官（権限のある職員）　５Ｂ　３５６０ 鈴木　修治 電話番号　０３－３５８１－１１０１　内線　３５４５

Ｃ（続き）．　関連すると認められる文献

引用文献のカテゴリー＊	引用文献名　及び一部の箇所が関連するときは、その関連する箇所の表示	関連する請求の範囲の番号
	& US 2004/0143664 A1	
Ａ	JP 2002-202959 A (株式会社日立製作所) 2002.07.19, 段落（００１８）－（００３２), 第１図 & US 2002/0087611 A1	１－１７