

(19) 世界知的所有権機関  
国際事務局

(43) 国際公開日  
2012年8月9日(09.08.2012)



(10) 国際公開番号

WO 2012/105230 A1

- (51) 国際特許分類:  
*G06F 9/50* (2006.01)      *G06F 1/32* (2006.01)
- (21) 国際出願番号: PCT/JP2012/000605
- (22) 国際出願日: 2012年1月31日(31.01.2012)
- (25) 国際出願の言語: 日本語
- (26) 国際公開の言語: 日本語
- (30) 優先権データ:  
特願 2011-020949 2011年2月2日(02.02.2011) JP
- (71) 出願人(米国を除く全ての指定国について): 日本電気株式会社(NEC Corporation) [JP/JP]; 〒1088001 東京都港区芝五丁目7番1号 Tokyo (JP).
- (72) 発明者; および
- (75) 発明者/出願人(米国についてのみ): 大野 善之 (OHNO, Yoshiyuki) [JP/JP]; 〒1088001 東京都港区芝五丁目7番1号 日本電気株式会社内 Tokyo (JP).  
小林 大(KOBAYASHI, Dai) [JP/JP]; 〒1088001 東京都港区芝五丁目7番1号 日本電気株式会社内

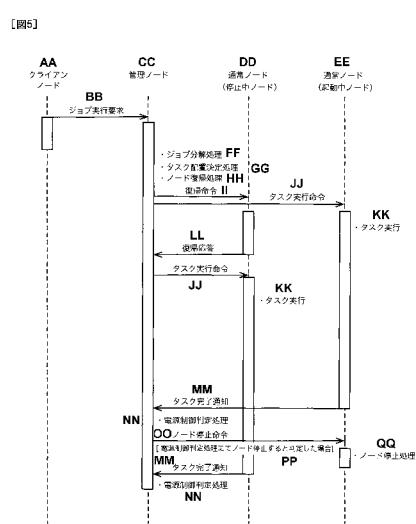
Tokyo (JP). 菅 真樹 (KAN, Masaki) [JP/JP]; 〒1088001 東京都港区芝五丁目7番1号 日本電気株式会社内 Tokyo (JP).

- (74) 代理人: 岩壁 冬樹, 外(IWAKABE, Fuyuki et al.); 〒1040031 東京都中央区京橋二丁目8番7号 読売中公ビル6階 サンライズ国際特許事務所 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, CL, 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, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, 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, LR, LS, MW,

[続葉有]

(54) Title: DISTRIBUTED SYSTEM, DEVICE, METHOD, AND PROGRAM

(54) 発明の名称: 分散システム、装置、方法及びプログラム



- AA CLIENT NODE  
BB JOB EXECUTION REQUEST  
CC ADMINISTRATION NODE  
DD REGULAR NODE (INTERRUPTED NODE)  
EE REGULAR NODE (ACTIVATED MODE)  
FF JOB SEGMENTING PROCESS  
GG TASK POSITION DETERMINATION PROCESS  
HH NODE RESTORATION PROCESS  
II RESTORATION INSTRUCTION  
JJ TASK EXECUTION INSTRUCTION  
KK TASK EXECUTION  
LL RESTORATION RESPONSE  
MM TASK COMPLETION NOTIFICATION  
NN POWER SUPPLY CONTROL ASSESSMENT PROCESS  
OO NODE INTERRUPT INSTRUCTION  
PP (IF NODE INTERRUPT IS ASSESSED IN POWER SUPPLY CONTROL ASSESSMENT PROCESS)  
QQ NODE INTERRUPT PROCESS

**(57) Abstract:** Provided is a distributed system, comprising: regular nodes further having a plurality of power conserving states with differing times to the restoring of a regular operating state; and an administration node which allocates jobs to the regular nodes and causes the execution of same thereon. The administration node further comprises: a node selection means for selecting the regular node to which to allocate a job and cause the execution thereof from among the regular nodes which are in a power conserving state; and a node control means for controlling such that the selected regular node is restored to the regular operating state. The node selection means selects the regular nodes in the power conserving states in ascending order of time to restoration to the regular operating state.

**(57) 要約:** 通常動作状態への復帰時間が異なる複数の省電力状態を有する通常ノードと、ジョブを通常ノードに割り当てる実行させる管理ノードとを備え、管理ノードは、省電力状態にある通常ノードからジョブを割り当てる実行させる通常ノードを選択するノード選択手段と、選択した通常ノードを通常動作状態に復帰するように制御するノード制御手段とを含み、ノード選択手段は、複数の省電力状態のうちの通常動作状態に復帰する時間が短い省電力状態にある通常ノードから順に選択する。



MZ, NA, RW, 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 (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE,  
ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV,  
MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK,

SM, 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)）

## 明 細 書

### 発明の名称：分散システム、装置、方法及びプログラム

#### 技術分野

[0001] 本発明は、複数のノードを有する分散システムに関し、特に、複数のノードを1つの分散システムとして動作させる際に、分散システム全体の省電力化をはかる分散システムに関する。

#### 背景技術

[0002] 分散システムは、プロセッサおよび記憶媒体を備えたノード（計算機・記憶装置）を数十台から数千台規模備え、ネットワークで接続することで1つのシステムとして利用し、1ノードでは得ることのできない計算能力や記憶容量を得ることができるシステムである。

[0003] 分散システムの一例として、ノードをネットワーク結合し、そのHDD（ハードディスクドライブ）やメモリを用いてデータ格納・利用する分散ストレージシステムが存在する。分散ストレージシステムでは、データをどの計算機に配置するか、処理をどの計算機で行うかをソフトウェアや特別なハードウェアにより実現し、システムの状態に対し動作を動的に変更することで、システム内のリソース使用量を調整し、システム利用者（クライアント計算機）に対する性能を向上している。

[0004] 分散システムを構成するノード1台の消費電力は150W/H程度であるが、大規模な分散システムでは、数百・数千ものノードを備えているために、システム全体の消費電力が膨大になるという問題がある。

[0005] しかし、システムに対する負荷量は時々刻々と変化するものであり、常に全てのノードを100%活用するような負荷がかかっているわけではない。ある時にはシステム全体の10%を利用し、またある時にはシステム全体の90%を利用するというように、システム内で必要となるノード数は時間によって変動する。

[0006] このような負荷量の変動に応じ、ジョブを実行していないアイドル状態の

ノードが偶発的に発生したり、また、負荷量に応じて処理ノード数を変化させることでアイドル状態のノードを意図的に作り出したりすることができる。例えば特許文献1には、ジョブ実行中でないアイドル状態のノードを、サスペンド状態へ移行させることで消費電力を削減するクラスタシステムが記載されている。

- [0007] スーパーコンピュータのような複数の計算機を接続して大規模な計算を行う計算機システムでは、複数のノードにジョブを割り当てて実行させ、あるジョブが完了すると次のジョブを複数のノードに計算ジョブを割りつけるというジョブ管理をする。
- [0008] 特許文献2では、上述のスーパーコンピュータのような計算機システムにおいて、システムに与えられたジョブを低消費電力で実行するジョブ管理方法が開示されている。特許文献2に記載されたジョブ管理方法では、システムに対するジョブ状況を保持しておき、未来のジョブ実行時期やジョブ実行に必要なノード数を決定し、個々のジョブ実行ごとに必要なノードをあらかじめジョブ実行開始前に準備し、一方でジョブ実行に必要のないノードを停止し、システム全体の消費電力の削減をはかる。
- [0009] 上記先行技術で利用されるノード停止には、通常、ACPI : Advanced Configuration and Power Interface (非特許文献1) で規定される停止状態がとられることが多い。

## 先行技術文献

### 特許文献

- [0010] 特許文献1：特開2003-162515号公報  
特許文献2：特開2008-225639号公報

### 非特許文献

- [0011] 非特許文献1：「ADVANCED CONFIGURATION AND POWER INTERFACE SPECIFICATION 4.0」 ヒューレットパッカード、インテル、マイクロソフト、フェニックステクノロジー、東芝 2009年6月16日

## 発明の概要

### 発明が解決しようとする課題

- [0012] しかし、上記特許文献1に記載された技術では、一度サスPEND状態などの停止状態に移行させたノードを再度利用するには、ノードに通電し、OSを起動して利用が可能になるまでの待ち時間が必要となる。
- [0013] また、上記特許文献2に記載されたジョブ管理方法では、ジョブの実行予定に基づいて停止ノードをあらかじめ復帰させてくことで、ノード復帰にかかる時間を隠蔽している。スーパーコンピュータのようにジョブを実行キューにためて処理を行うようなシステムにおいては隠蔽効果がある。特に、1ジョブを数百ノードで協調動作して数十分以上のオーダーで処理する場合（並列処理）においては効果的である。しかし、分散ストレージシステムのようにシステムに対する処理要求が予想できない場合や、小さなジョブが多数あったり、大きなジョブを小さなタスクに分割して多くのノードに分配したりして処理する場合（分散処理）、復帰にかかる時間の隠蔽ができないという問題がある。
- [0014] このように上記の技術では、消費電力を抑制するためにノードを停止させるとノード復帰にかかる待ち時間のために、処理性能（処理時間）が低下するという問題がある。
- [0015] そこで本発明は、複数のノードを備えた分散システムにおいて、ノードを停止状態に移行させることで分散システム全体の消費電力を抑制しながらも、負荷が大きくなった時の処理性能の低下の抑制を図ることができる分散システム、情報処理装置、分散方法及び分散プログラムを提供することを目的とする。

### 課題を解決するための手段

- [0016] 本発明による分散システムは通常動作状態への復帰時間が異なる複数の省電力状態を有する通常ノードと、ジョブを通常ノードに割り当てて実行させる管理ノードとを備え、管理ノードは、省電力状態にある通常ノードからジョブを割り当てて実行させる通常ノードを選択するノード選択手段と、ノー

ド選択手段が選択した通常ノードを通常動作状態に復帰させるように制御するノード制御手段とを含み、ノード選択手段は、複数の省電力状態のうちの通常動作状態に復帰するまでの時間が短い省電力状態にある通常ノードから順に選択することを特徴とする。

[0017] 本発明による分散システムは、電源制御手段とタスク実行手段とを備える通常ノードと、ジョブの実行命令を受け付けるジョブ受信手段と、ジョブ受信手段が受け付けたジョブを1つまたは複数のタスクに分解し、1台または複数台の通常ノードにタスクを実行させるジョブ管理手段と、通常ノードの電源状態を管理制御するノード電源制御手段とを備える管理ノードとを備えた分散システムであって、電源制御手段は、省電力状態中の消費電力と省電力状態から通常起動状態に復帰する時間とがそれぞれ異なる複数段階の省電力状態に通常ノードを移行させる機能と、省電力状態にある通常ノードを通常起動状態に復帰させる機能とを有し、ジョブ管理手段は、管理ノードが受け付けたジョブの量に応じて、ジョブを分解したタスクを実行させる通常ノード数を決定し、ノード電源制御手段は、通常ノードのうち通常起動状態にありタスク実行が可能な通常ノード数が、ジョブ管理手段が決定したタスクを実行させる通常ノード数に満たない場合に、複数段階の省電力状態のうちの1つ以上の状態にある通常ノードから、タスクを実行させる通常ノード数を満たすだけ通常ノードを選択する起動ノード選択機能を有し、起動ノード選択機能において、複数段階ある省電力状態のうちの通常起動状態に復帰するまでの時間が短い省電力状態にある通常ノードから順に、通常起動状態に復帰させる通常ノードとして選択し、選択した通常ノードに対して、通常起動状態への移行を指示する命令である起動命令を発行し、ジョブ管理手段は、通常ノードのうち通常起動状態にありタスク実行が可能な通常ノードと、ノード電源制御手段が発行した起動命令に従って通常起動状態に復帰した通常ノードとに対してタスクの実行を指示するタスク実行命令を発行し、タスク実行手段は、ジョブ管理手段が発行したタスク実行命令に従ってタスクを実行することを特徴とする。

[0018] 本発明による情報処理装置は、分散システムにおいてジョブを通常ノードに割り当てて実行させる情報処理装置であって、通常動作状態への復帰時間が異なる複数の省電力状態を有する通常ノードのうちの省電力状態にある通常ノードからジョブを割り当てて実行させる通常ノードを選択するノード選択手段と、ノード選択手段が選択した通常ノードを通常動作状態に復帰させるように制御するノード制御手段とを含み、ノード選択手段は、複数の省電力状態のうちの通常動作状態に復帰するまでの時間が短い省電力状態にある通常ノードから順に選択することを特徴とする。

[0019] 本発明による分散方法は、ジョブを通常ノードに割り当てて実行させる分散方法であって、通常動作状態への復帰時間が異なる複数の省電力状態を有する通常ノードのうちの省電力状態にある通常ノードからジョブを割り当てて実行させる通常ノードを選択し、選択した通常ノードを通常動作状態に復帰させるように制御し、省電力状態にある通常ノードからジョブを割り当てて実行させる通常ノードを選択する際に、複数の省電力状態のうちの通常動作状態に復帰するまでの時間が短い省電力状態にある通常ノードから順に選択することを特徴とする。

[0020] 本発明による分散プログラムは、分散システムにおいてジョブを通常ノードに割り当てて実行させるための分散プログラムであって、コンピュータに、通常動作状態への復帰時間が異なる複数の省電力状態を有する通常ノードのうちの省電力状態にある通常ノードからジョブを割り当てて実行させる通常ノードを選択するノード選択処理と、選択した通常ノードを通常動作状態に復帰させるように制御するノード制御処理とを実行させ、ノード選択処理で、複数の省電力状態のうちの通常動作状態に復帰するまでの時間が短い省電力状態にある通常ノードから順に選択させることを特徴とする。

## 発明の効果

[0021] 本発明によれば、複数のノードを備えた分散システムにおいて、ノードを停止状態に移行させることで分散システム全体の消費電力を抑制しながらも、負荷が大きくなった時の処理性能の低下の抑制を図ることができる。

## 図面の簡単な説明

[0022] [図1]本実施形態における分散システムの構成例を示すブロック図である。

[図2]分散システムにおける管理ノードの構成例を示すブロック図である。

[図3]分散システムにおける通常ノードの電源状態の一例を示す状態遷移図である。

[図4]分散システムにおける通常ノードの各電源状態の消費電力と、各状態遷移にかかる時間との一例を示す表である。

[図5]分散システムのジョブ実行の一例を表すシーケンス図である。

[図6]分散システムにおける管理ノードのノード電源制御部が処理する復帰ノード選択手順の一例を表すフローチャートである。

[図7]分散システムにおいて電源制御ノード決定部が復帰命令を発行した際の一例を示すシーケンス図である。

[図8]分散システムにおいてタスク配置決定部222がタスク実行命令を発行した際の一例を示すシーケンス図である。

[図9]分散システムにおけるノード電源制御部が処理する電源制御判定処理の処理手順の一例を表すフローチャートである。

[図10]分散システムにおいてノード電源制御部が停止命令を発行した際の一例を示すシーケンス図である。

[図11]分散システムにおけるノード電源制御部が処理するノード数調整処理の処理手順の一例を表すフローチャートである。

[図12]分散システムの最小の構成例を示すブロック図である。

## 発明を実施するための形態

[0023] 以下、本発明による分散システムの実施形態について、図面を参照して説明する。図1は、本実施形態における分散システムの構成例を示すブロック図である。図1に示すように、分散システムは、アクセス経路決定手段を有するネットワークに接続された1つ以上のクライアントノード100と、1つ以上の管理ノード200と、1つ以上の通常ノード300とを含む。図1に示す例では、1つのクライアントノード100と1つの管理ノード200

とが示されているが、複数のクライアントノード100と管理ノード200とを含んでいてもよい。

- [0024] 通常ノード300は、1ノードずつ個別のノード番号（ノード001～ノードXXX）を有している。具体的には、通常ノード300は、ノード番号を記憶部に記憶している。
- [0025] クライアントノード100は、ジョブ実行を要求するノードである。クライアントノード100が発行したジョブ実行要求がネットワークを介して管理ノード200に伝えられる。具体的には、クライアントノード100は、ネットワークを介して管理ノード200にジョブ実行要求を送信する。
- [0026] 管理ノード200は、ジョブ受信部210、ジョブ制御部220およびノード電源制御部230を備えている。
- [0027] ジョブ受信部210は、クライアントノード100が発行したジョブ実行要求を受け付ける機能を備えている。具体的には、ジョブ受信部210は、クライアントノード100がネットワークを介して送信したジョブ実行要求を受信する。以下、クライアントノード100がジョブ実行要求を発行するとの表現を用いるが、具体的には、クライアントノード100がネットワークを介して管理ノード200にジョブ実行要求を送信することである。
- [0028] ジョブ制御部220は、ジョブ受信部210が受け付けたジョブを各通常ノード300が実行できる単位のタスクとして分割し、通常ノード300にタスク実行を要求する機能を備えている。具体的には、ジョブ制御部220は、ジョブ受信部210が受信したジョブ実行要求によって示されるジョブを通常ノード300が実行できる単位のタスクに分割し、ネットワークを介して通常ノード300に分割したタスクの実行要求を送信する。
- [0029] ノード電源制御部230は、通常ノード300の電源状態を管理したり、ノード停止／復帰といった電源制御を行う通常ノード300を決定し、電源制御要求を発行したりする機能を備えている。具体的には、ノード電源制御部230は、電源制御を行う通常ノード300に対して、ネットワークを介して電源制御要求を送信する。以下、電源制御要求を発行するとの表現を用

いるが、具体的には、ネットワークを介して通常ノード300に電源制御要求を送信することである。

- [0030] 管理ノード200が発行する電源制御要求には、通常ノード300を省電力状態へ停止させる停止命令（以下、ノード停止命令ともいう）と、通常ノード300を省電力状態から復帰させる復帰命令（以下、ノード復帰命令ともいう）との2種類ある。すなわち、通常ノード300は、停止命令を受信すると起動状態（例えば、後述する実行状態またはアイドル状態。また、通常動作状態や通常起動状態ともいう）から省電力状態に移行し、復帰命令を受信すると省電力状態から起動状態に移行するように制御する。
- [0031] 図2は、管理ノード200の構成例を示すブロック図である。図2に示すようにジョブ制御部220は、ジョブ分解部221と、タスク配置決定部222と、命令通知部223とを備えている。
- [0032] ジョブ分解部221は、ジョブを各通常ノード300が実行できる単位のタスクとして分割する機能を備えている。具体的には、ジョブ分解部221は、ジョブ受信部210が受信したジョブ実行要求によって示されるジョブを、通常ノード300が実行できる単位のタスクに分割する。
- [0033] タスク配置決定部222は、ジョブ分解部221が分解したタスクをどの通常ノード300に実行させるかを決定する機能を備えている。
- [0034] 命令通知部223は、タスク配置決定部222の決定に従って、通常ノード300に対して、タスク実行命令や電源制御要求を通知する機能を備えている。以下、タスク実行命令を発行するとの表現を用いるが、具体的には、管理ノード200が、ネットワークを介して通常ノード300にタスク実行命令を送信することである。
- [0035] また、ノード電源制御部230は、電源制御ノード決定部231を備えている。電源制御ノード決定部231は、ジョブ制御部220から電源制御要求（または後述するノード起動要求）を受けた際、または所定期間ごとに、電源制御対象とする通常ノード300および移行する電源状態を決定する機能を備えている。

- [0036] また、管理ノード200は、記憶部240を備えている。記憶部240は、ジョブ制御部220およびノード電源制御部230にまたがって、各通常ノード300に対するタスクの分配状況や通常ノード300のタスク実行状況を管理するためのタスク配置情報と、各通常ノード300の電源状態を管理するためのノード状態情報を記憶している。
- [0037] また、図1に示すように、通常ノード300は、通信部310、タスク実行部320、復帰命令受信部330および電源制御部340を備えている。
- [0038] 通信部310は、管理ノード200が発行したタスク実行命令や、電源制御要求のうちの停止命令を受信する機能を備えている。
- [0039] タスク実行部320は、通信部310が受信したタスク実行命令に基づいてタスクを実行する機能を備えている。
- [0040] 復帰命令受信部330は、管理ノード200が発行した電源制御要求のうちの復帰命令を受信する機能を備えている。
- [0041] 電源制御部340は、通信部310が受信した停止命令や復帰命令受信部330が受信した復帰命令に従って電源制御を行う機能を備えている。
- [0042] なお、本実施形態では、1つの通常ノード300において、あらかじめ定められた数のタスクを同時処理可能である。以下の例では説明を簡略化するため1つのタスクを実行できるようにする。
- [0043] 図3に、通常ノードで電源制御を行う場合の状態遷移図を示す。通常ノード300は、タスクを実行していないアイドル状態である場合、省電力効果の異なる3種類の省電力状態（停止状態レベル1、停止状態レベル2、停止状態レベル3）に移行することができる。
- [0044] 省電力状態の例として、ACPI(Advanced Configuration and Power Interface)で規定される省電力状態がある。本実施形態では、停止状態レベル1としてACPIで規定されるS1(プロセッサ給電停止)、停止状態レベル2としてACPIで規定されるS3(メモリのみ給電)、停止状態レベル3としてACPIで規定されるS4(メモリ内容のディスク退避、全電源供給停止)に移行するものとする。ただし、これらの省電力状態でも、復帰命令受信部330には電気供給され

ており、復帰命令受信部330は、常に復帰命令を受信できる。また、通常ノード300は、アイドル状態でなくタスク実行状態であったとしても、実行状態を示すデータをメモリやディスク装置などの記憶装置に保持しておき、アイドル状態を経由して省電力状態に移行することもできる。

- [0045] ACPIで規定される電源状態S1、S3、S4では、省電力効果はS1、S3、S4の順に小さい。すなわち、S1が最も小さく、次にS3が小さく、S4が最も大きい。一方でアイドル状態から省電力状態に移行したり、省電力状態からアイドル状態に復帰したりする状態遷移時間はS1、S3、S4の順に短い。すなわち、S1が最も短く、次にS3が短く、S4が最も長い。つまり、状態遷移時間が長い省電力状態ほど省電力効果が高い。参考として、図4に各状態の消費電力と各状態遷移にかかる時間との一例を示す。
- [0046] 本実施形態では、省電力状態をACPIで規定される3状態としたが、状態数は複数であればよく、また、省電力にさせるACPIで規定される方式である必要もなく、本発明で制限するものではない。
- [0047] 次に、本実施形態における分散システムにおける管理ノード200が保持する、タスク配置情報、および、ノード状態情報について説明する。上述のように、タスク配置情報およびノード状態情報は、管理ノード200の記憶部240に記憶されている。
- [0048] タスク配置情報には、各通常ノード300がタスク実行中であるか否かを示す情報が含まれる。ノード状態情報には、各通常ノード300が省電力状態として停止中であるか、または起動中であるかを示す情報が含まれる。例えば、ある通常ノード300が省電力状態として停止中であれば、ノード状態情報には、その通常ノード300が停止状態であることを示す情報が含まれる。また、ノード状態情報には、状態ごとに該当する通常ノード300の個数を示す情報が含まれる。
- [0049] 次に、本実施形態における分散システムのジョブ実行の処理フローについて説明する。
- [0050] 図5は、ジョブ実行の流れを示すシーケンス図である。まず、クライアン

トノード100は、ジョブ実行要求を管理ノード200に送信する。すると、クライアントノード100からジョブ実行要求を受信した管理ノード200は、ジョブ分解部221において、受け付けたジョブを1つ以上のタスクに分解する[図5に示されるジョブ分解処理]。ここで受け付けたジョブとは、受信したジョブ実行要求によって示されるジョブである。

- [0051] ジョブ分解処理が完了すると、管理ノード200は、タスク配置決定部222において、分解したタスクをどの通常ノード300で実行するかを決定する[図5に示されるタスク配置決定処理]。
- [0052] このタスク配置決定処理では、管理ノード200は、通常ノード300の一部が低消費電力状態で停止中であることにより、分解したタスクが起動中の通常ノード300では実行しきれるか否かを判断する。そして、起動中の通常ノード300では実行しきれないと判断した場合、管理ノード200は、低消費電力状態で停止中の通常ノード300の一部または全てを復帰させる[図5に示されるノード復帰処理]。
- [0053] ノード復帰処理では、タスク配置決定部222は、ノード電源制御部230に対して、ノード起動要求とともに復帰すべきノード数を通知する。すると、ノード電源制御部230は、復帰対象ノードを決定し、決定した復帰対象の通常ノード300宛にノード復帰命令を発行する。その後、復帰命令を受信した通常ノード300は、ノード復帰処理を行い、復帰後にノード復帰応答を管理ノード200に返信する。ここでのノード復帰処理とは、例えば、通常ノード300が停止状態からアイドル状態に移行するように制御することである。
- [0054] また、ノード復帰処理と同時に、タスク配置決定部222は、その時点で起動中の通常ノード300に対して、命令通知部223を通じてタスク実行命令を発行する。一方、ノード復帰処理によって復帰させた通常ノード300に対しては、タスク配置決定部222は、ノード復帰応答を受信後に、タスク実行命令を発行する。
- [0055] タスク実行命令を受信した通常ノード300は、タスク実行命令に従って

、タスクを実行し、タスク実行完了後にタスク完了通知を管理ノード200に送信する。

[0056] 管理ノード200は、タスク完了通知を受信すると、タスクが完了した通常ノード300を低消費電力状態で停止させるか否かを判定する〔電源制御判定処理〕。そして、低消費電力状態で停止させると判定した場合には、管理ノード200は、該当する通常ノード300に対してノード停止命令を発行する。すると、ノード停止命令を受信した通常ノード300は、低消費電力状態で停止する。一方で、電源制御判定処理で通常ノード300を停止しないと判定した場合には、管理ノード200はそのまま何もせず、該当する通常ノード300もアイドル状態で待機する。

[0057] 以下に、管理ノード200における、ジョブ分解処理、タスク配置決定処理、ノード復帰処理および電源制御判定処理について説明する。また、管理ノード200がノード復帰命令・タスク実行命令・ノード停止命令の各命令を発行した場合の、管理ノード200と通常ノード300との間の処理シーケンスについて説明する。

[0058] ジョブ分解処理では、ジョブ分解部221は、ジョブを1つ以上、(全通常ノード300の数-タスク実行中の通常ノード300の数)個以下のタスクに分解する。このときジョブ分解部221は、1つあたり通常ノード300が1台で処理するタスクとして分解する。例えば、ジョブとしてN個のファイルに書かれた数列を1つのファイルにマージしてソートするという処理が与えられたとき、N-1個のソートタスクと1個の(ソートandファイルマージ)タスクとに分解するという方法が考えられる。

[0059] タスク数を小さくして、1タスクの処理量を多くすれば、タスク実行する通常ノード300が少なくなり、他の通常ノード300を低消費電力状態に停止することができるので、省電力効果が高い。しかし、ジョブ実行にかかる時間は増大する。逆に、1タスクの処理量を少なくし、タスク数を多くすると、多くの通常ノード300がタスク実行を処理するので、ジョブ実行にかかる時間は短くなるものの、省電力効果は低くなる。なお、本発明において

ては、ジョブ分解部221でのタスク分解の方法については限定しない。

- [0060] タスク配置決定処理では、タスク配置決定部222は、ジョブ分解処理で分解したタスクを実行する通常ノード300を決定する。具体的には、タスク配置決定部222は、記憶部240が記憶するノード状態情報を参照し、タスク数がアイドル状態の通常ノード300の数以下であるか否か判断する。そして、タスク数がアイドル状態の通常ノード300の数以下であると判断した場合には、タスク配置決定部222は、アイドル状態の通常ノード300のうち、タスク数個の通常ノード300をタスク実行するノードとして選択する。
- [0061] 一方、タスク数がアイドル状態の通常ノード300の数より多いと判断した場合には、タスク配置決定部222は、アイドル状態の通常ノード300全てをタスク実行するノードとして選択し、残りの（タスク数－アイドル状態の通常ノード数）個の通常ノード300をノード復帰処理により起動し、タスク実行する通常ノード300として選択する。
- [0062] タスク実行ノードとして選択された通常ノード300は、アイドル状態であればタスク配置決定処理直後に、停止状態であればノード復帰処理にてアイドル状態に移行した後に、タスク実行命令に従ってタスク実行を開始する。なお、本発明においては、アイドル状態の通常ノード300のうち、どの通常ノード300をタスク実行するノードとして選択するかについての決定方法は限定しない。また、どのタスクをどの通常ノード300に配置するかについての決定方法も限定しない。
- [0063] ノード復帰処理では、タスク配置決定部222は、ノード電源制御部230に対して、ノード起動要求とともに起動するノード数を通知する。図6に、ノード電源制御部230が、ノード起動要求を受信し、起動するノード数分の停止状態の通常ノード300を選択し、復帰命令を通常ノード300に対して発行する手順を示す。図6に示す復帰ノード選択手順では、停止状態にある通常ノード300のうち、停止状態レベル1から順に、つまり、省電力効果が低く復帰が早いノードから順に復帰させるノードとして選択し、復

帰命令を発行している。

- [0064] 図6に示すように、ノード電源制御部230は、タスク配置決定部222が出力したノード起動要求を受信する（ステップS10）。ここでは、ノード電源制御部230は、ノード起動要求とともに、起動する通常ノードの数を示す値nを受信する。
- [0065] 次いで、ノード電源制御部230は、停止状態レベルが低い順に、停止状態レベルk（ここでは停止状態レベル1）の通常ノード300の数が、x個（受信した値n）より少いか否かを判断する（ステップS11、12）。
- [0066] 停止状態レベルkの通常ノード300の数がx個より少ないと判断した場合には、ノード電源制御部230は、全ての停止状態レベルkの通常ノード300に対して復帰命令を発行する（ステップS13）。そして、ノード電源制御部230は、xの値をx-通常ノード300数（停止状態レベルk）とし、停止状態レベルが低い順に停止状態レベル3に至るまで、ステップS12、13の処理を繰り返す（ステップS14、15、16）。その後、ノード電源制御部230は、n-x個のノード復帰応答の受信待ちをし（ステップS19）、ノード起動処理を完了する。
- [0067] 一方、停止状態レベルkの通常ノード300の数がx個以上であると判断した場合には、ノード電源制御部230は、停止状態レベルkの通常ノード300のうちのx個に対して復帰命令を発行する（ステップS17）。その後、ノード電源制御部230は、n個のノード復帰応答の受信待ちをし（ステップS18、19）、ノード起動処理を完了する。
- [0068] 図7は、電源制御ノード決定部231が復帰命令を発行して、通常ノード300が復帰するまでの流れの一例を示すシーケンス図である。
- [0069] 電源制御ノード決定部231は、復帰させる通常ノード300のノード番号を指定して復帰命令を発行する。発行した復帰命令は、命令通知部223を経由して、復帰させる対象の通常ノード300の復帰命令受信部330に通知される。
- [0070] 具体的には、電源制御ノード決定部231は、復帰させる通常ノードを特

定し、特定した通常ノードのノード番号を含む復帰命令を命令通知部223に出力する。すると、命令通知部223は、ネットワークを介して、ノード番号によって特定される通常ノード300の復帰命令受信部330に復帰命令を送信する。

- [0071] 通常ノード300の復帰命令受信部330は、復帰命令を受信すると、ノードの通電を開始し、電源制御部340に対して復帰要求を出力する。すると、電源制御部340は、ノード復帰処理を行う。ここでのノード復帰処理とは、例えば、通常ノード300が停止状態からアイドル状態に移行するように制御することである。
- [0072] 電源制御部340によるノード復帰処理が完了すると、通常ノード300は、通信部310を用いて、管理ノード200に対して、ノード復帰処理が完了したことを示すノード復帰応答を自ノード番号とともに通知する。
- [0073] 管理ノード200の電源制御ノード決定部231は、ノード復帰応答を受信すると、記憶部240が記憶するノード状態情報を書き換える。具体的には、電源制御ノード決定部231は、ノード復帰応答を受信した通常ノード300が起動中であることを示すようにノード状態情報を書き換える。
- [0074] また同時に、電源制御ノード決定部231は、タスク配置決定部222に対して、停止状態にあった通常ノード300が復帰したことを探知する。すると、タスク配置決定部222は、該当する通常ノード300に対してタスク実行命令を発行する。
- [0075] 図8は、タスク配置決定部222がタスク実行命令を発行して、通常ノード300がタスク実行を完了するまでの流れの一例を示すシーケンス図である。
- [0076] タスク配置決定部222は、タスクを実行させる通常ノード300に対してタスク実行命令を通知する前に、タスク配置情報を書き換える。具体的には、タスク配置決定部222は、対象の通常ノード300がタスク実行中であることを示すようにタスク配置情報を書き換える。
- [0077] タスク配置情報の書き換え後、タスク配置決定部222は、命令通知部2

23を介して、タスクを実行させる通常ノード300に対してタスク実行命令を通知する。

- [0078] 通常ノード300は、通信部310でタスク実行命令を受信すると、受信したタスク実行命令に従って、タスク実行部320でタスクを実行する。
- [0079] タスク実行部320によるタスク実行が完了すると、通常ノード300は、通信部310を用いて、タスク完了通知をタスクの実行結果とともに管理ノード200に通知する。
- [0080] 管理ノード200のタスク配置決定部222は、タスク完了通知を受信すると、タスク配置情報を、タスクが完了した旨に書き換える。すなわち、タスク配置決定部222は、対象の通常ノード300がタスク実行中でないことを示すようにタスク配置情報を書き換える。そして、タスク配置情報の書き換えが完了すると、管理ノード200は、電源制御判定処理を実行する。
- [0081] 電源制御判定処理では、管理ノード200は、アイドル状態、停止状態レベル1、停止状態レベル2の優先度順に、あらかじめ設定しておいた各状態の必要ノード数を満たすように停止状態を決定する。アイドル状態、停止状態レベル1、停止状態レベル2の各状態においても必要ノード数を満たしている場合には、停止状態3に停止する。
- [0082] 電源制御ノード決定部231は、あらかじめ電源制御判定処理におけるアイドル状態および停止状態レベルごとの必要ノード数の設定値を保持している。この設定値は、固定値でも非固定値でもよい。例えば、固定値で、アイドル状態ノード：0ノード、停止状態レベル1：5ノード、停止状態レベル2：15ノードというように設定する。また、例えば、非固定値で、アイドル状態ノード：(全ノード数-実行状態ノード数)×0%、停止状態レベル1：(全ノード数-実行状態ノード数)×10%、停止状態レベル2：(全ノード数-実行状態ノード数)×30%というように、その時点における各状態におけるノード数から必要ノード数を計算させるという方法もある。
- [0083] 次に、電源制御判定処理の処理手順について説明する。図9は、電源制御判定処理の処理手順の一例を示すフローチャートである。

- [0084] まず、電源制御ノード決定部231は、アイドル状態のノード数とあらかじめ設定しておいた設定値とを比較し、アイドル状態のノード数が設定値よりも小さいか否かを判定する（ステップS21）。そして、設定値よりも小さいと判定した場合には、電源制御ノード決定部231は、何も処理を行わずに、電源制御判定処理を完了する。
- [0085] 一方、アイドル状態のノード数が設定値以上であると判定した場合、電源制御ノード決定部231は、停止状態レベル1から停止状態レベル2へと順に、停止状態レベルkにおいて停止中のノード数とあらかじめ停止状態レベルごとに設定しておいた設定数とを比較する。
- [0086] そして、停止状態レベルkにおいて停止中のノード数が停止状態レベルkの設定数に満たない場合には、電源制御ノード決定部231は、停止状態レベルkで停止することを決定する。停止状態レベル1から停止状態レベル2で停止すると決定しなかった場合には、電源制御ノード決定部231は停止状態レベル3で停止すると決定する。停止させる状態が決定すると、ノード電源制御部230は、停止命令を対象の通常ノード300に対して発行する。
- [0087] 図9に示す例では、電源制御ノード決定部231は、アイドル状態のノード数が停止状態レベルk（ここでは停止状態レベル1）の設定値よりも小さいか否かを判定する（ステップS22、23）。
- [0088] ステップS23においてアイドル状態の通常ノード数が設定値よりも小さいと判定した場合には、電源制御ノード決定部231は、停止状態レベルk（ここでは停止状態レベル1）に移行させることを決定し、対象の通常ノード300に対して停止命令を発行する（ステップS24）。その後、電源制御ノード決定部231は、電源制御判定処理を完了する。
- [0089] 一方、ステップS23においてアイドル状態のノード数が設定値以上であると判定した場合には、電源制御ノード決定部231は、停止状態レベル3に至るまでステップS23の処理を繰り返す（ステップS25、26）。その後、電源制御ノード決定部231は、停止状態レベル3に移行させること

を決定し、対象の通常ノード300に対して停止命令を発行する（ステップS27）、電源制御判定処理を完了する。

[0090] 図10は、ノード電源制御部230の電源制御ノード決定部231が停止命令を発行して、通常ノード300が省電力状態で停止するまでの流れの一例を示すシーケンス図である。

[0091] 電源制御ノード決定部231は、まず停止対象の通常ノード300の状態が停止状態を示すようにノード状態情報を変更する。その後、電源制御ノード決定部231は、命令通知部223を介して、停止対象の通常ノード300に停止命令を通知する。通信部310で停止命令を受信した通常ノード300は、電源制御部340に停止命令を出力し、停止命令に従って指定された停止レベルで停止するように制御する。

[0092] 以上の実行フローで、本実施形態の分散システムはジョブを実行する。

[0093] なお、電源制御ノード決定部231は、アイドル状態および停止状態レベルごとに設定した必要ノード数を満たすように、ジョブ実行とは非同期のタイミングで、停止状態にある通常ノード300を、より復帰時間の短い停止状態、およびアイドル状態に状態変更をする[ノード数調整処理]。

[0094] 図11は、ノード数調整処理で、アイドル状態または停止状態レベルLにn個の通常ノード300を追加する場合の手順の一例を示すフローチャートである。なお、図11では簡単化のため、アイドル状態を停止状態レベル0と記すものとする。

[0095] 図11に示すように、電源制御ノード決定部231は、停止状態レベルの高い順に、停止状態レベルk（ここでは停止状態レベル3）の通常ノード300の数がx個（ここではn個）より小さいか否か判定する（ステップS31、32）。

[0096] 停止状態レベルkの通常ノード300の数がx個より少ないと判断した場合には、電源制御ノード決定部231は、全ての停止状態レベルkの通常ノード300を停止状態レベルLに変更する（ステップS33）。そして、電源制御ノード決定部231は、xの値をx-通常ノード300数（停止状態

レベル k ) とし、停止状態レベルが高い順に停止状態レベル L に至るまで、ステップ S 3 2 、 3 3 の処理を繰り返す (ステップ S 3 4 、 3 5 、 3 6) 。その後、電源制御ノード決定部 2 3 1 は、ノード調整処理を完了する。

- [0097] 一方、停止状態レベル k の通常ノード 3 0 0 の数が x 個以上であると判断した場合には、電源制御ノード決定部 2 3 1 は、停止状態レベル k の通常ノード 3 0 0 のうちの x 個を停止状態レベル L に変更する (ステップ S 3 7) 。その後、ノード電源制御部 2 3 0 は、 x の値を 0 とし (ステップ S 3 8) 、ノード調整処理を完了する。
- [0098] 図 1 1 に示す手順により、より高い停止状態レベルにある通常ノード 3 0 0 のうちの n 個の通常ノード 3 0 0 を停止状態レベル L に移行させることができる。なお、停止状態にある通常ノード 3 0 0 を、より低い停止状態レベルに移行させる場合には、例えば、復帰命令によりアイドル状態に復帰させ (例えば図 7 に示される) 、復帰応答受信後に、停止命令により停止状態に停止させる (例えば図 1 0 に示される) 。
- [0099] ノード数調整処理が実施されるタイミングは任意であるが、例えば最後に電源制御要求 (停止命令や復帰命令) が発行されてから所定期間後にノード数調整処理を実施するというタイミングをとることが望ましい。
- [0100] 以上が本実施形態における分散システムの説明である。このような形態をとることにより、本発明による分散システムは、ノードを停止状態に移行させることで分散システム全体の消費電力を抑制しながらも、負荷が大きくなつた時には、復帰の早い停止状態レベルの低いノードから順に復帰させてタスク実行させることで、処理性能の低下を抑制することができる。
- [0101] なお、本発明による分散システムは、本実施形態に限定されるわけではなく、下記のような変更も可能である。
- [0102] 上記実施形態の分散システムでは、通常ノード 3 0 0 では、同時に 1 つのタスクしか実行できない例について説明したが、複数のタスクを同時に実行できてもよい。その場合、タスク配置情報では、割り当て中のタスクを全て管理する必要がある。また、タスク配置決定部 2 2 2 がタスクの配置を決定

する際には、起動中のノードで実行中タスク数の少ないノードにタスクを多く配置するタスク配置決定方法をとることもできる。

- [0103] 以上の記載は実施形態に基づいて行ったが、本発明は上記実施形態に限定されるものではない。本発明の構成や詳細には、本発明のスコープ内で当業者が理解し得る様々な変更を加えることができる。
- [0104] 次に、本発明による分散システムの最小構成について説明する。図12は、分散システムの最小の構成例を示すブロック図である。図12に示すように、分散システムは、最小の構成要素として、通常動作状態への復帰時間が異なる複数の省電力状態を有する通常ノード10と、ジョブを通常ノード10に割り当てて実行させる管理ノード20とを備えている。また、管理ノード20は、ノード選択手段21と、ノード制御手段22とを含む。
- [0105] 図12に示す最小構成の分散システムでは、ノード選択手段21は、省電力状態にある通常ノード10からジョブを割り当てて実行させる通常ノード10を選択する。ここで、ノード選択手段21は、複数の省電力状態のうちの通常動作状態に復帰するまでの時間が短い省電力状態にある通常ノード10から順に選択する。そして、ノード制御手段22は、ノード選択手段21が選択した通常ノード10を通常動作状態に復帰させるように制御する。
- [0106] 従って、最小構成の分散システムによれば、ノードを停止状態に移行させることで分散システム全体の消費電力を抑制しながらも、負荷が大きくなった時には、復帰の早い停止状態レベルの低いノードから順に復帰させてタスク実行させることで、処理性能の低下を抑制することができる。
- [0107] なお、本実施形態では、以下の(1)～(7)に示すような分散システムの特徴的構成が示されている。
- [0108] (1) 分散システムは、通常動作状態（例えば、実行状態やアイドル状態）への復帰時間が異なる複数の省電力状態を有する通常ノード（例えば、通常ノード300によって実現される）と、ジョブを通常ノードに割り当てて実行させる管理ノード（例えば、管理ノード200によって実現される）とを備え、管理ノードは、省電力状態にある通常ノードからジョブを割り当

て実行させる通常ノードを選択するノード選択手段（例えば、タスク配置決定部222および電源制御ノード決定部231によって実現される）と、ノード選択手段が選択した通常ノードを通常動作状態に復帰させるように制御するノード制御手段（例えば、電源制御ノード決定部231によって実現される）とを含み、ノード選択手段は、複数の省電力状態のうちの通常動作状態に復帰するまでの時間が短い省電力状態にある通常ノードから順に選択することを特徴とする。

[0109] (2) 分散システムにおいて、通常ノードは、演算処理を実行するプロセッサと、情報を記憶するメモリおよび不揮発性の記憶装置とを少なくとも搭載し、移行可能な省電力状態として、プロセッサの電源のみ停止する第1の省電力状態（例えば、停止状態レベル1）と、演算のコンテキストをメモリに保存し、メモリ以外の給電を停止する第2の省電力状態（例えば、停止状態レベル2）と、演算のコンテキストを不揮発性の記憶装置に保存し、全ての給電を停止する第3の省電力状態（例えば、停止状態レベル3）との3種類の省電力状態とを少なくとも含み、ノード選択手段は、省電力状態にある通常ノードからタスクを実行させる通常ノードを選択する際に、第1の省電力状態にある通常ノード、第2の省電力状態にある通常ノード、第3の省電力状態にある通常ノードの順に優先して選択するように構成されていてよい。

[0110] (3) 分散システムは、電源制御手段（例えば、電源制御部340によって実現される）とタスク実行手段（例えば、タスク実行部320によって実現される）とを備える通常ノード（例えば、通常ノード300によって実現される）と、ジョブの実行命令を受け付けるジョブ受信手段（例えば、ジョブ受信部210によって実現される）と、ジョブ受信手段が受け付けたジョブを1つまたは複数のタスクに分解し、1台または複数台の通常ノードにタスクを実行させるジョブ管理手段（例えば、ジョブ制御部220によって実現される）と、通常ノードの電源状態を管理制御するノード電源制御手段（例えば、ノード電源制御部230によって実現される）とを備える管理ノードである。

ド（例えば、管理ノード200によって実現される）とを備えた分散システムであって、電源制御手段は、省電力状態中の消費電力と省電力状態から通常起動状態に復帰する時間とがそれぞれ異なる複数段階の省電力状態に通常ノードを移行させる機能と、省電力状態にある通常ノードを通常起動状態に復帰させる機能とを有し、ジョブ管理手段は、管理ノードが受け付けたジョブの量に応じて、ジョブを分解したタスクを実行させる通常ノード数を決定し（例えば、ジョブ分解部221とタスク配置決定部222が処理を実行することによって実現される）、ノード電源制御手段は、通常ノードのうち通常起動状態にありタスク実行が可能な通常ノード数が、ジョブ管理手段が決定したタスクを実行させる通常ノード数に満たない場合に、複数段階の省電力状態のうちの1つ以上の状態にある通常ノードから、タスクを実行させる通常ノード数を満たすだけ通常ノードを選択する起動ノード選択機能を有し、起動ノード選択機能において、複数段階ある省電力状態のうちの通常起動状態に復帰するまでの時間が短い省電力状態にある通常ノードから順に、通常起動状態に復帰させる通常ノードとして選択し、選択した通常ノードに対して、通常起動状態への移行を指示する命令である起動命令を発行し（例えば、電源制御ノード決定部231が処理を実行することによって実現される）、ジョブ管理手段は、通常ノードのうち通常起動状態にありタスク実行が可能な通常ノードと、ノード電源制御手段が発行した起動命令に従って通常起動状態に復帰した通常ノードとに対してタスクの実行を指示するタスク実行命令を発行し、タスク実行手段は、ジョブ管理手段が発行したタスク実行命令に従ってタスクを実行することを特徴とする。

[0111] (4) 分散システムにおいて、通常ノードは、演算処理を実行するプロセッサと、情報を記憶するメモリおよび不揮発性の記憶装置とを少なくとも搭載し、電源制御手段は、制御する省電力状態として、プロセッサの電源のみ停止する第1の省電力状態（例えば、停止状態レベル1）と、演算のコンテキストをメモリに保存し、メモリ以外の給電を停止する第2の省電力状態（例えば、停止状態レベル2）と、演算のコンテキストを不揮発性の記憶装置

に保存し、全ての給電を停止する第3の省電力状態（例えば、停止状態レベル3）との3種類の省電力状態とを少なくとも含むように構成されていてもよい。

- [0112] (5) 分散システムにおいて、ノード電源制御手段は、通常起動状態に復帰させる通常ノードを、第1の省電力状態にある通常ノードから選択し、次に、第2の省電力状態にある通常ノードから選択し、次に、第3の省電力状態にある通常ノードから選択するように構成されていてもよい。
- [0113] (6) 分散システムにおいて、ノード電源制御手段は、タスク実行が完了した通常ノードを、複数段階の省電力状態のうちのいずれかの状態に移行させることを決定し、通常ノードに対し、決定した省電力状態への移行を指示するノード停止命令を発行するように構成されていてもよい。
- [0114] (7) 分散システムにおいて、ノード電源制御手段は、タスク実行が完了した通常ノードに対してノード停止命令を発行するにあたり、各省電力状態ごとに予め設定された所定数に満たない省電力状態のうち、通常起動状態に復帰するまでの時間が短い省電力状態に移行させることを決定するように構成されていてもよい。
- [0115] 上記の実施形態の一部又は全部は、以下の付記のようにも記載され得るが、以下には限られない。
- [0116] (付記1) ノード電源制御手段は、省電力状態にある通常ノードのうち、通常起動状態に復帰するまでの時間が短い省電力状態にある通常ノード数が所定数以下のときに、通常起動状態に復帰するまでの時間が長い省電力状態にある通常ノードのうちの1つまたは複数を、前記通常起動状態に復帰するまでの時間が短い省電力状態に移行させることを決定し、決定した移行対象の通常ノードに対して復帰命令を発行し、該通常ノードが通常起動状態に復帰後に、該通常ノードに対して、ノード停止命令を発行する請求項3から請求項7のうちのいずれか1項に記載の分散システム。
- [0117] 以上、実施形態及び実施例を参照して本願発明を説明したが、本願発明は上記実施形態および実施例に限定されるものではない。本願発明の構成や詳

細には、本願発明のスコープ内で当業者が理解し得る様々な変更をすることができる。

[0118] この出願は、2011年2月2日に出願された日本特許出願2011-020949を基礎とする優先権を主張し、その開示の全てをここに取り込む。

### 産業上の利用可能性

[0119] 本発明に係る分散システムは、分散コンピュータや分散データベース、分散ストレージ、並列データ処理システム、並列ファイルシステム、並列データベース、データグリッド、クラスタコンピュータに適用することができる。

### 符号の説明

[0120]

- 10 通常ノード
- 20 管理ノード
- 21 ノード選択手段
- 22 ノード制御手段
  - 100 クライアントノード
  - 200 管理ノード
  - 210 ジョブ受信部
  - 220 ジョブ制御部
  - 221 ジョブ分解部
  - 222 タスク配置決定部
  - 223 命令通知部
  - 230 ノード電源制御部
  - 231 電源制御ノード決定部
  - 240 記憶部
- 300 通常ノード
- 310 通信部
- 320 タスク実行部

330 復帰命令受信部

340 電源制御部

## 請求の範囲

- [請求項1] 通常動作状態への復帰時間が異なる複数の省電力状態を有する通常ノードと、  
ジョブを前記通常ノードに割り当てて実行させる管理ノードとを備え、  
前記管理ノードは、  
前記省電力状態にある通常ノードから前記ジョブを割り当てて実行させる通常ノードを選択するノード選択手段と、  
前記ノード選択手段が選択した通常ノードを前記通常動作状態に復帰させるように制御するノード制御手段とを含み、  
前記ノード選択手段は、前記複数の省電力状態のうちの通常動作状態に復帰するまでの時間が短い省電力状態にある通常ノードから順に選択する  
ことを特徴とする分散システム。
- [請求項2] 通常ノードは、演算処理を実行するプロセッサと、情報を記憶するメモリおよび不揮発性の記憶装置とを少なくとも搭載し、移行可能な省電力状態として、前記プロセッサの電源のみ停止する第1の省電力状態と、演算のコンテキストを前記メモリに保存し、前記メモリ以外の給電を停止する第2の省電力状態と、演算のコンテキストを前記不揮発性の記憶装置に保存し、全ての給電を停止する第3の省電力状態との3種類の省電力状態とを少なくとも含み、  
ノード選択手段は、前記省電力状態にある通常ノードからジョブを割り当てて実行させる通常ノードを選択する際に、前記第1の省電力状態にある通常ノード、前記第2の省電力状態にある通常ノード、前記第3の省電力状態にある通常ノードの順に優先して選択する  
請求項1記載の分散システム。
- [請求項3] 電源制御手段とタスク実行手段とを備える通常ノードと、  
ジョブの実行命令を受け付けるジョブ受信手段と、前記ジョブ受信

手段が受け付けたジョブを1つまたは複数のタスクに分解し、1台または複数台の前記通常ノードに該タスクを実行させるジョブ管理手段と、前記通常ノードの電源状態を管理制御するノード電源制御手段とを備える管理ノードとを備えた分散システムであって、

前記電源制御手段は、省電力状態中の消費電力と前記省電力状態から通常起動状態に復帰する時間とがそれぞれ異なる複数段階の省電力状態に該通常ノードを移行させる機能と、省電力状態にある該通常ノードを通常起動状態に復帰させる機能とを有し、

前記ジョブ管理手段は、該管理ノードが受け付けたジョブの量に応じて、該ジョブを分解したタスクを実行させる通常ノード数を決定し、

前記ノード電源制御手段は、前記通常ノードのうち通常起動状態にありタスク実行が可能な通常ノード数が、前記ジョブ管理手段が決定したタスクを実行させる通常ノード数に満たない場合に、前記複数段階の省電力状態のうちの1つ以上の状態にある通常ノードから、前記タスクを実行させる通常ノード数を満たすだけ通常ノードを選択する起動ノード選択機能を有し、前記起動ノード選択機能において、前記複数段階ある省電力状態のうちの通常起動状態に復帰するまでの時間が短い省電力状態にある通常ノードから順に、通常起動状態に復帰させる通常ノードとして選択し、選択した通常ノードに対して、通常起動状態への移行を指示する命令である起動命令を発行し、

前記ジョブ管理手段は、前記通常ノードのうち通常起動状態にありタスク実行が可能な通常ノードと、前記ノード電源制御手段が発行した前記起動命令に従って通常起動状態に復帰した通常ノードとに対してタスクの実行を指示するタスク実行命令を発行し、

前記タスク実行手段は、前記ジョブ管理手段が発行した前記タスク実行命令に従ってタスクを実行することを特徴とする分散システム。

- [請求項4] 通常ノードは、演算処理を実行するプロセッサと、情報を記憶するメモリおよび不揮発性の記憶装置とを少なくとも搭載し、  
電源制御手段は、制御する省電力状態として、前記プロセッサの電源のみ停止する第1の省電力状態と、演算のコンテキストを前記メモリに保存し、前記メモリ以外の給電を停止する第2の省電力状態と、演算のコンテキストを前記不揮発性の記憶装置に保存し、全ての給電を停止する第3の省電力状態との3種類の省電力状態とを少なくとも含む  
請求項3記載の分散システム。
- [請求項5] ノード電源制御手段は、通常起動状態に復帰させる通常ノードを、第1の省電力状態にある通常ノードから選択し、次に、第2の省電力状態にある通常ノードから選択し、次に、第3の省電力状態にある通常ノードから選択する  
請求項4記載の分散システム。
- [請求項6] ノード電源制御手段は、タスク実行が完了した通常ノードを、複数段階の省電力状態のうちのいずれかの状態に移行させることを決定し、前記通常ノードに対し、決定した省電力状態への移行を指示するノード停止命令を発行する  
請求項3から請求項5のうちのいずれか1項に記載の分散システム。  
。
- [請求項7] ノード電源制御手段は、タスク実行が完了した通常ノードに対してノード停止命令を発行するにあたり、各省電力状態ごとに予め設定された所定数に満たない省電力状態のうち、通常起動状態に復帰するまでの時間が短い省電力状態に移行させることを決定する  
請求項6記載の分散システム。
- [請求項8] 分散システムにおいてジョブを通常ノードに割り当てて実行させる情報処理装置であって、  
通常動作状態への復帰時間が異なる複数の省電力状態を有する前記

通常ノードのうちの該省電力状態にある通常ノードから前記ジョブを割り当てて実行させる通常ノードを選択するノード選択手段と、

前記ノード選択手段が選択した通常ノードを前記通常動作状態に復帰させるように制御するノード制御手段とを含み、

前記ノード選択手段は、前記複数の省電力状態のうちの通常動作状態に復帰するまでの時間が短い省電力状態にある通常ノードから順に選択する

ことを特徴とする情報処理装置。

[請求項9]

ジョブを通常ノードに割り当てて実行させる分散方法であって、

通常動作状態への復帰時間が異なる複数の省電力状態を有する前記通常ノードのうちの該省電力状態にある通常ノードから前記ジョブを割り当てて実行させる通常ノードを選択し、

選択した通常ノードを前記通常動作状態に復帰させるように制御し、

前記省電力状態にある通常ノードから前記ジョブを割り当てて実行させる通常ノードを選択する際に、前記複数の省電力状態のうちの通常動作状態に復帰するまでの時間が短い省電力状態にある通常ノードから順に選択する

ことを特徴とする分散方法。

[請求項10]

分散システムにおいてジョブを通常ノードに割り当てて実行させるための分散プログラムであって、

コンピュータに、

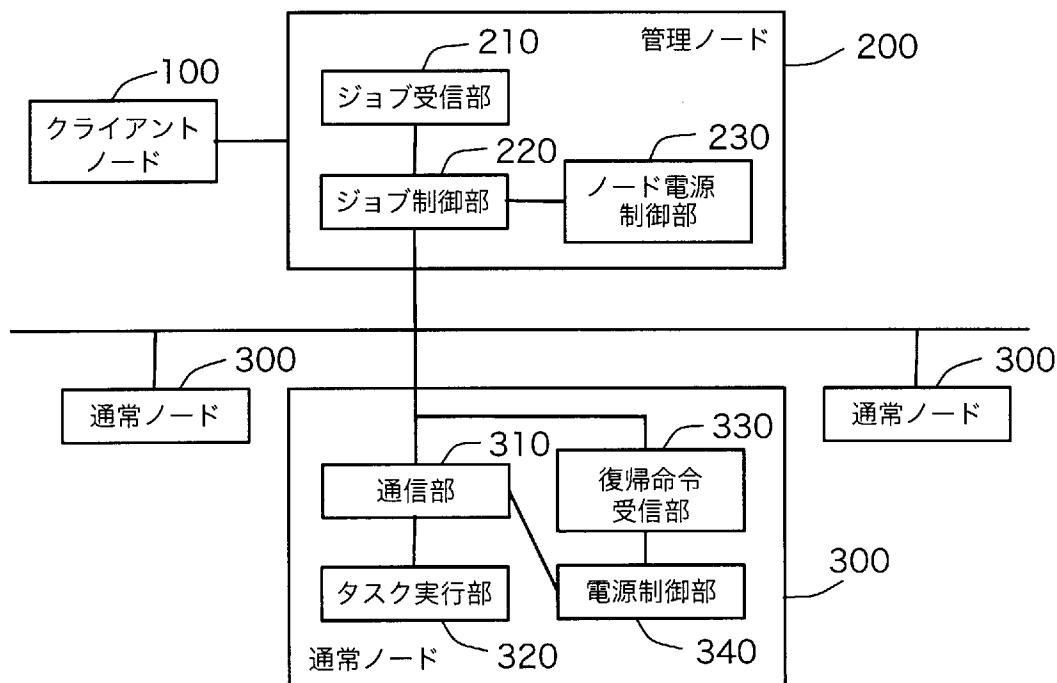
通常動作状態への復帰時間が異なる複数の省電力状態を有する前記通常ノードのうちの該省電力状態にある通常ノードから前記ジョブを割り当てて実行させる通常ノードを選択するノード選択処理と、

選択した通常ノードを前記通常動作状態に復帰させるように制御するノード制御処理とを実行させ、

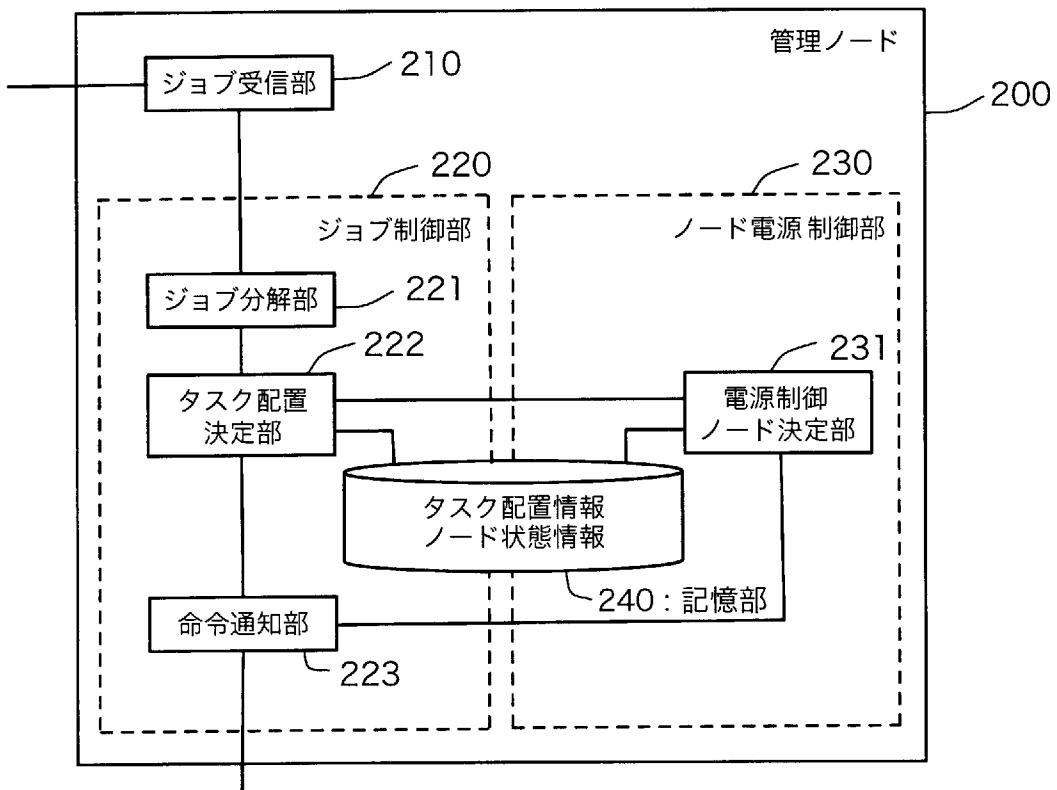
前記ノード選択処理で、前記複数の省電力状態のうちの通常動作状

態に復帰するまでの時間が短い省電力状態にある通常ノードから順に  
選択させる  
ための分散プログラム。

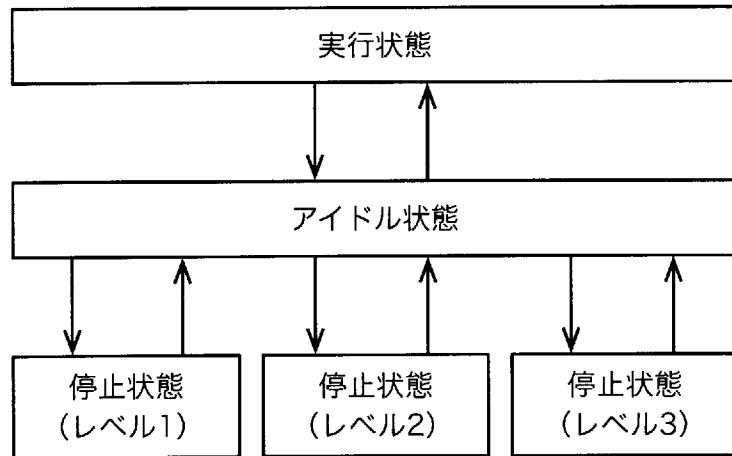
[図1]



[図2]



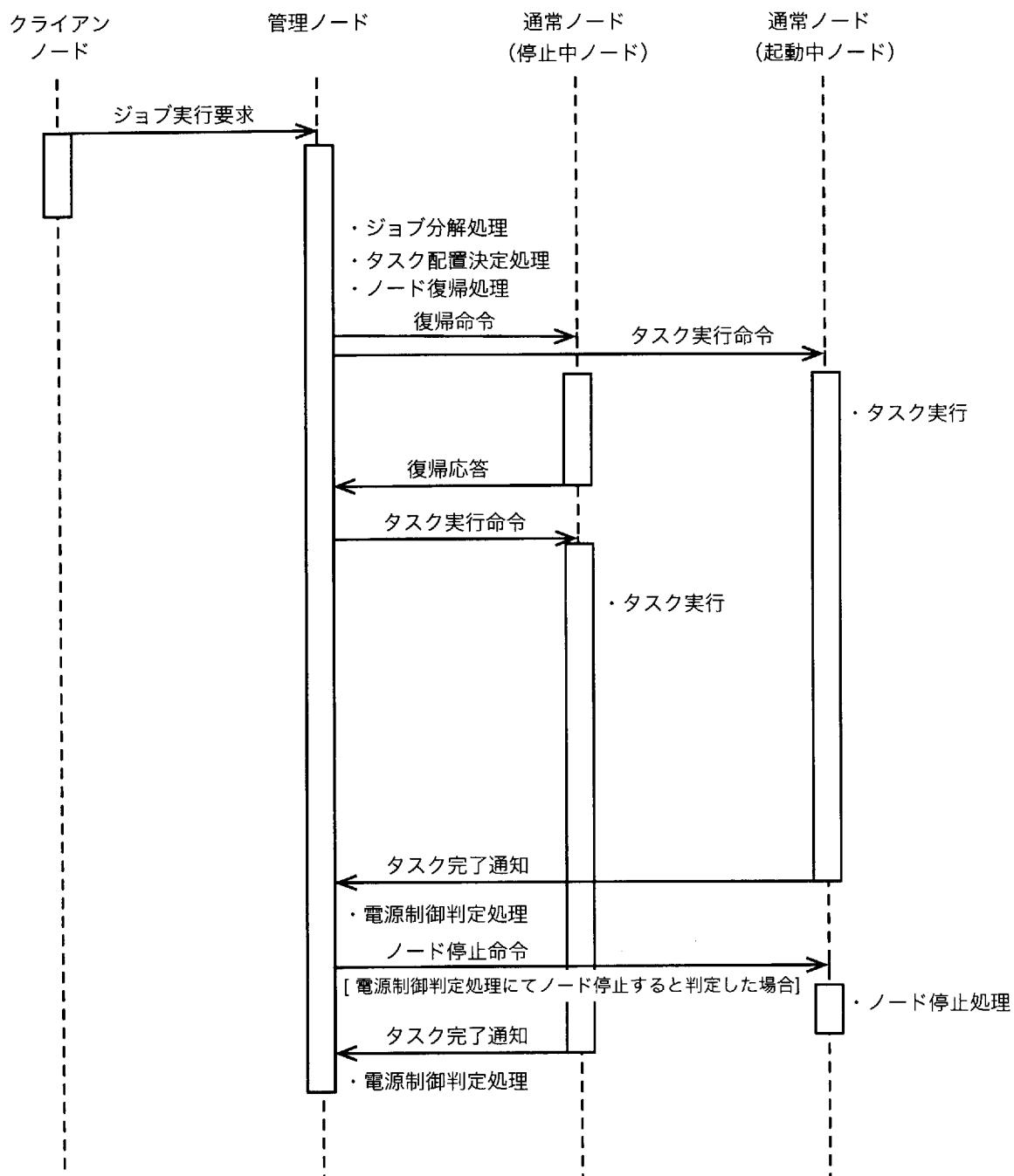
[図3]



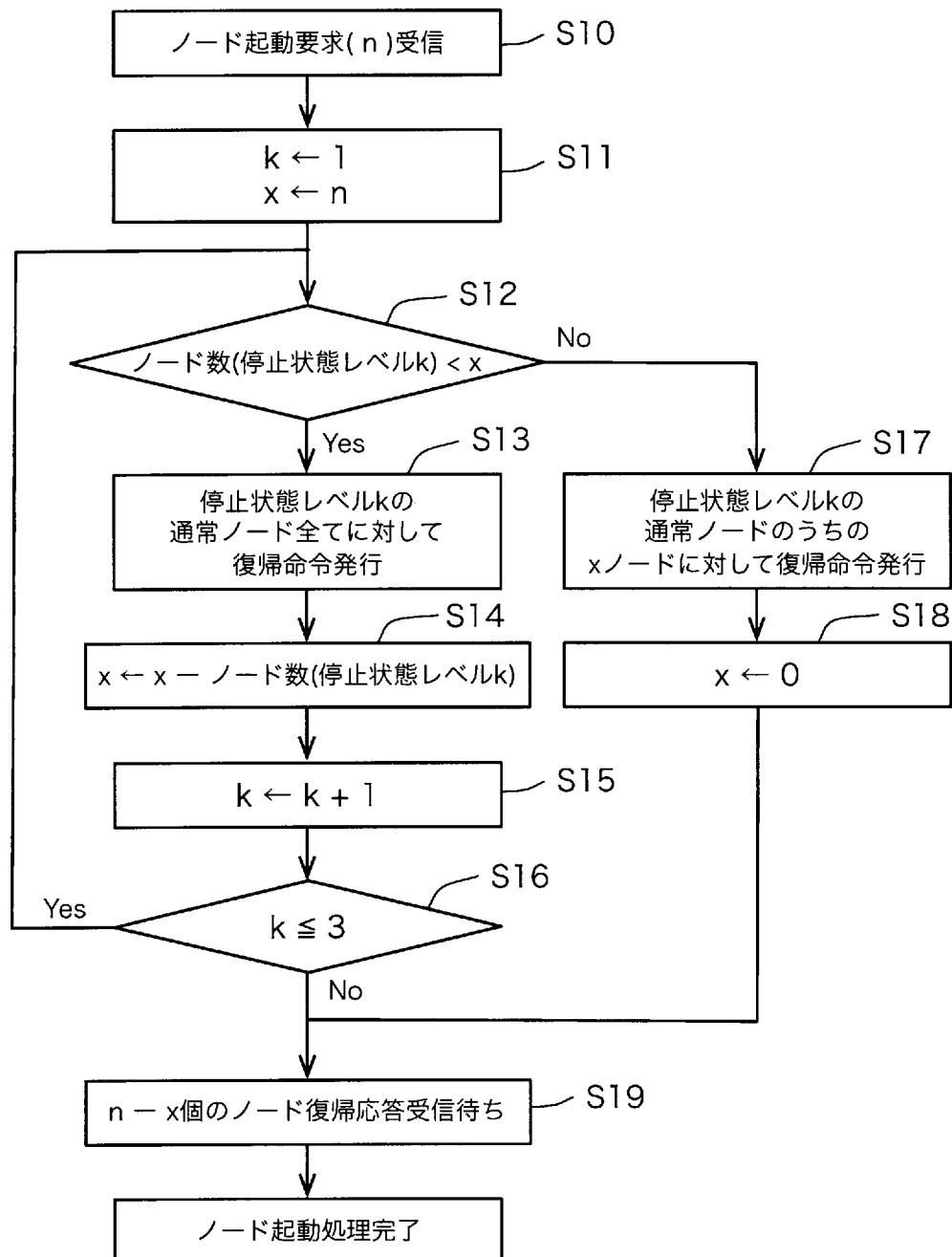
[図4]

状態	消費電力[W]	アイドル状態への遷移時間[sec]	アイドル状態からの遷移時間[sec]
実行状態	180	(タスク完了後) 0	0
アイドル状態	130	—	—
停止状態 (レベル1)	80	1	3
停止状態 (レベル2)	15	3	10
停止状態 (レベル3)	0	20	180

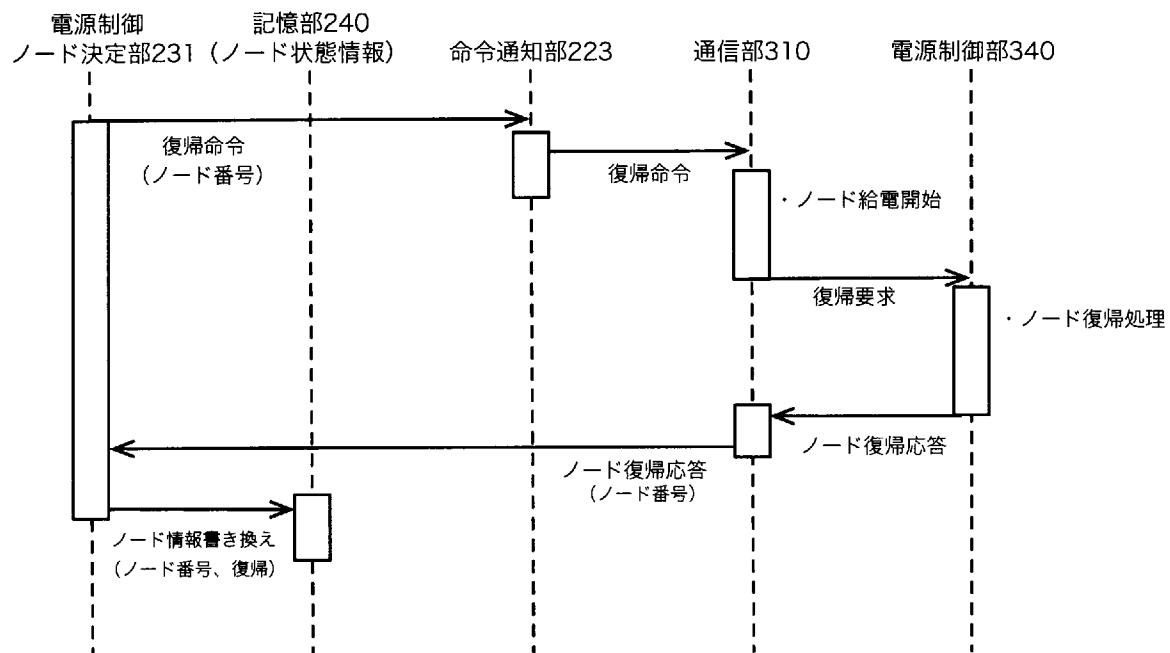
[図5]



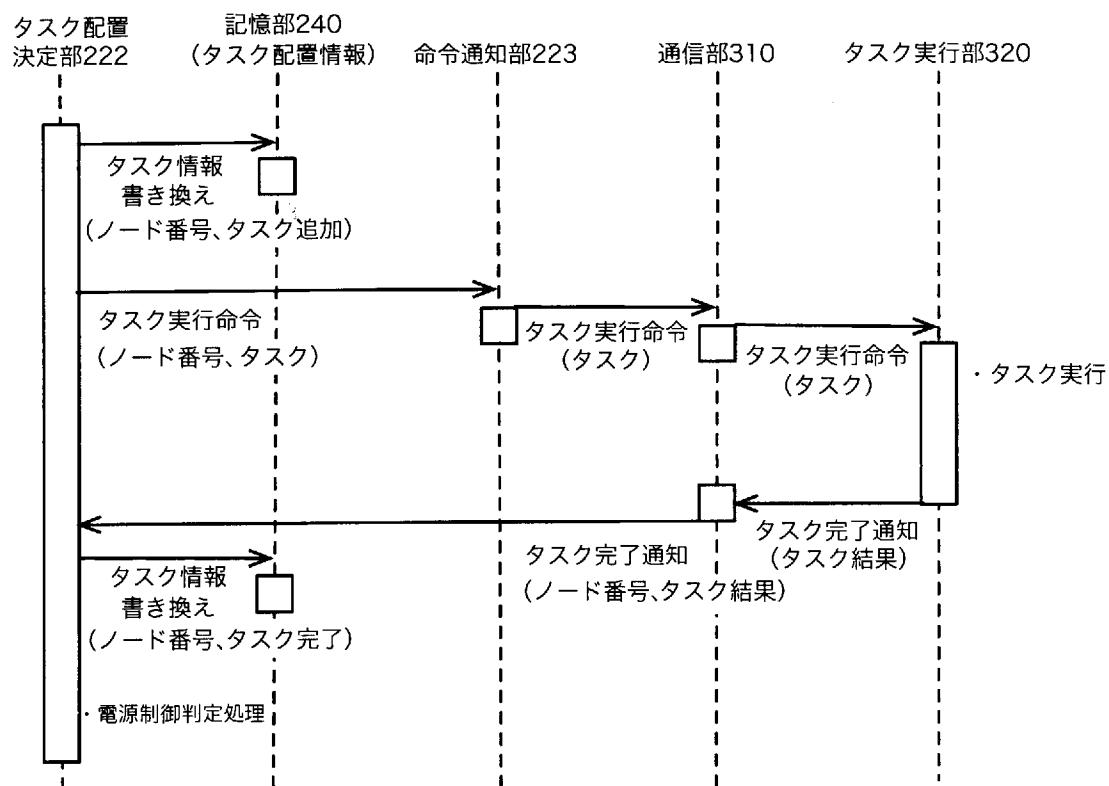
[図6]



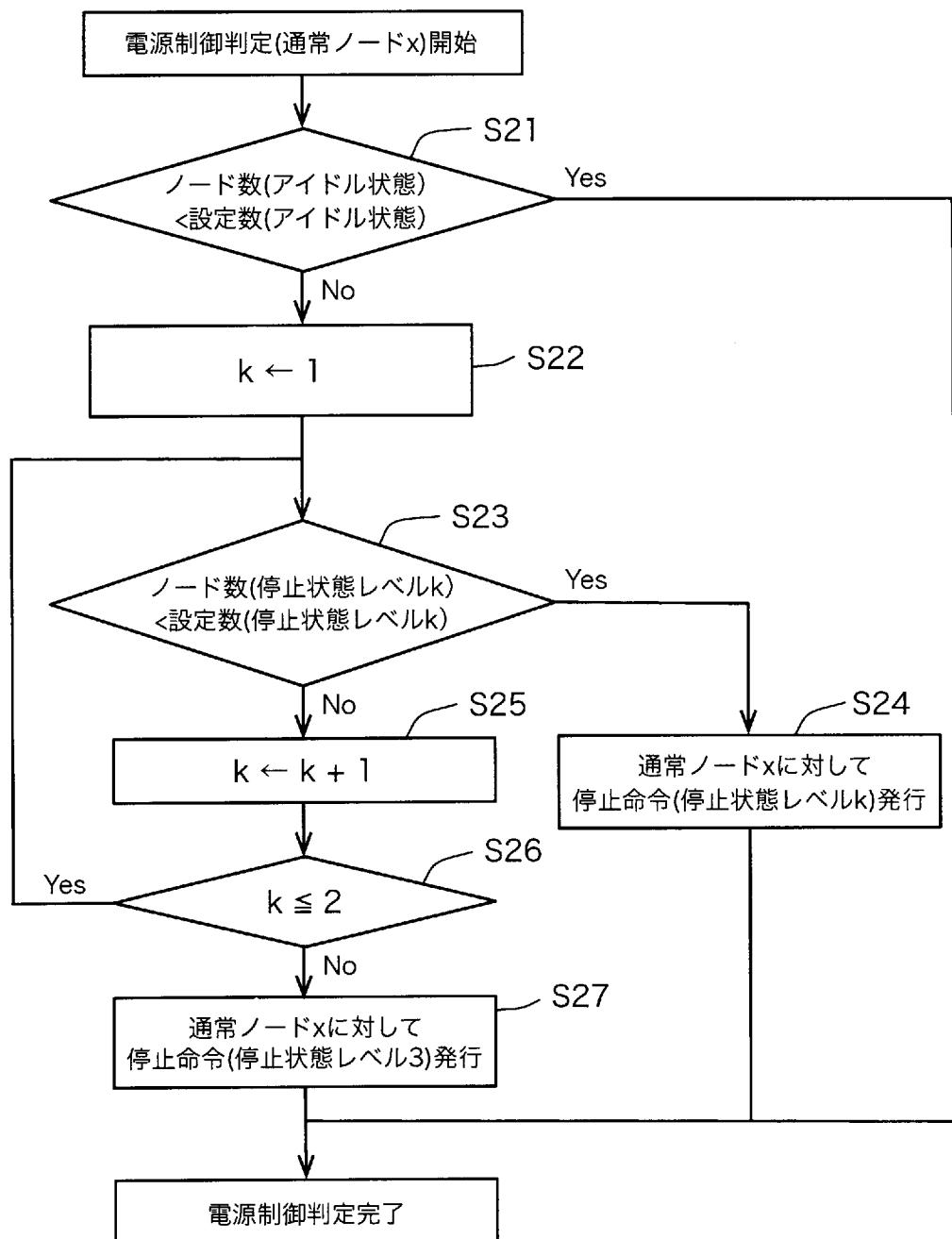
[図7]



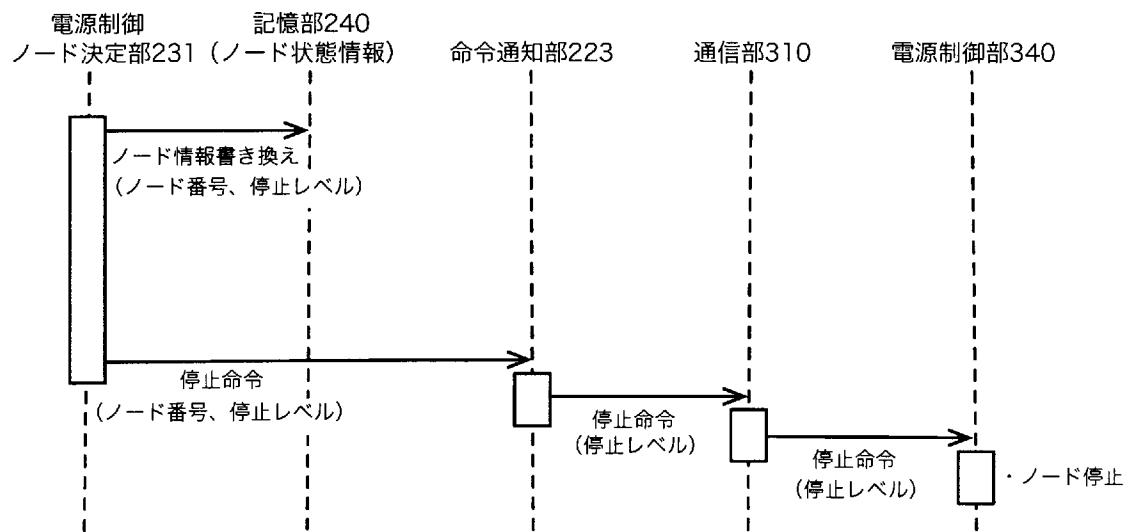
[図8]



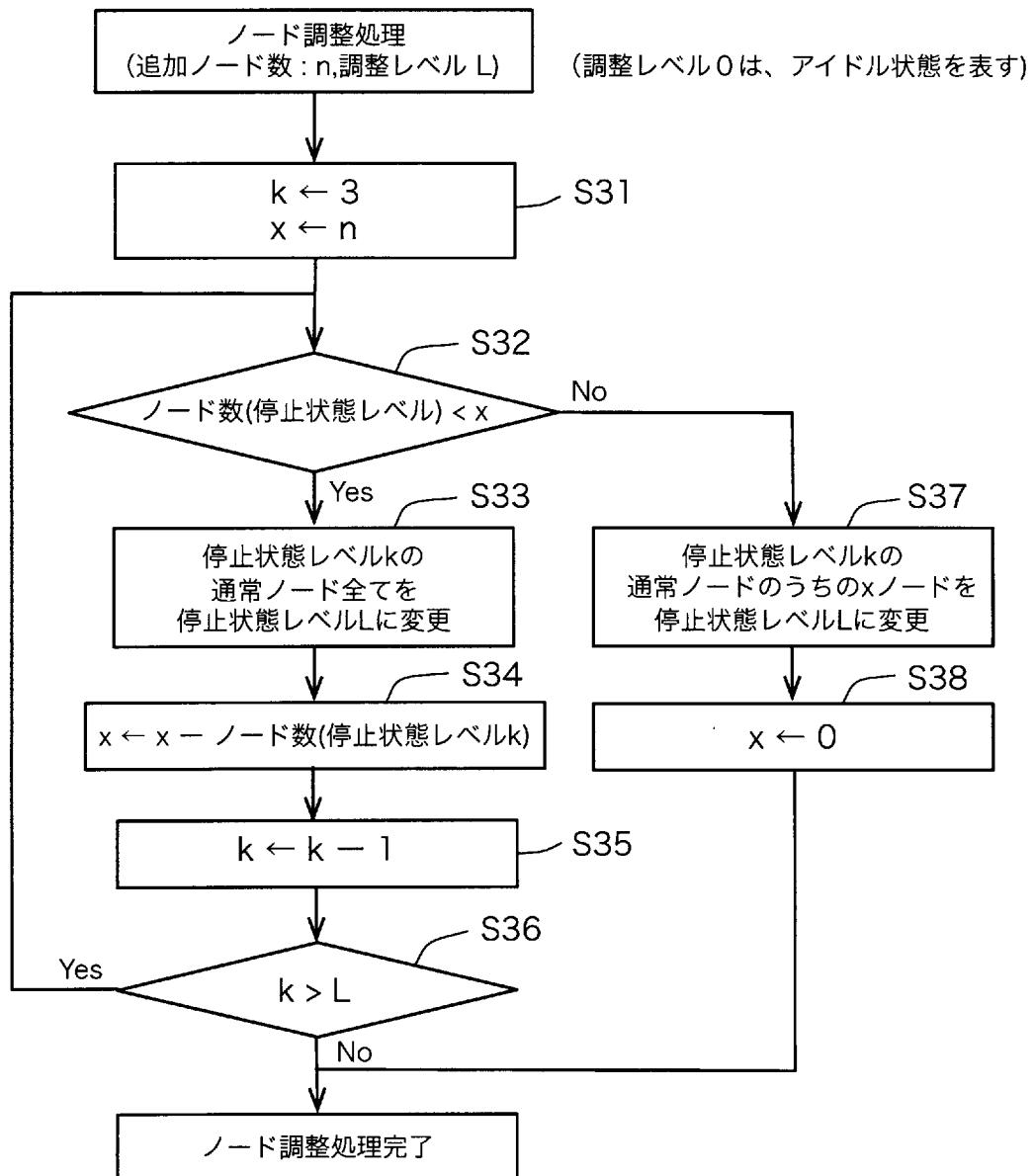
[図9]



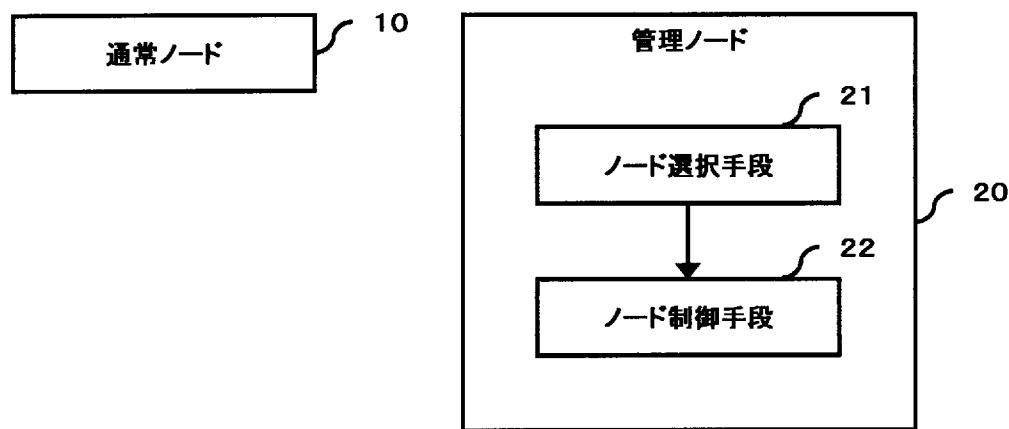
[図10]



[図11]



[図12]



## INTERNATIONAL SEARCH REPORT

International application No.

PCT/JP2012/000605

**A. CLASSIFICATION OF SUBJECT MATTER**  
*G06F9/50(2006.01) i, G06F1/32(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, G06F1/32*

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 1996-2012  
 Kokai Jitsuyo Shinan Koho 1971-2012 Toroku Jitsuyo Shinan Koho 1994-2012

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 2010-165193 A (Fujitsu Ltd.), 29 July 2010 (29.07.2010), entire text; all drawings & US 2010/0185766 A1	1-10
A	JP 2006-343955 A (Canon Inc.), 21 December 2006 (21.12.2006), entire text; all drawings & US 2006/0279766 A1	1-10
P, X	JP 2011-257834 A (Ricoh Co., Ltd.), 22 December 2011 (22.12.2011), fig. 6; paragraphs [0052] to [0055] (Family: none)	1-10

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  
 19 March, 2012 (19.03.12)

Date of mailing of the international search report  
 03 April, 2012 (03.04.12)

Name and mailing address of the ISA/  
 Japanese Patent Office

Authorized officer

Facsimile No.

Telephone No.

## 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, G06F1/32(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, G06F1/32

## 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	1971-2012年
日本国実用新案登録公報	1996-2012年
日本国登録実用新案公報	1994-2012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 請求項の番号
X	JP 2010-165193 A (富士通株式会社) 2010.07.29, 全文、全図 & US 2010/0185766 A1	1-10
A	JP 2006-343955 A (キヤノン株式会社) 2006.12.21, 全文、全図 & US 2006/0279766 A1	1-10
P X	JP 2011-257834 A (株式会社リコー) 2011.12.22, 図6、[0052] - [0055] (ファミリーなし)	1-10

□ 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  19. 03. 2012	国際調査報告の発送日  03. 04. 2012
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	特許序審査官(権限のある職員)  吉田 美彦 電話番号 03-3581-1101 内線 3545 5B 9384