

(19) 世界知的所有権機関
国際事務局



(43) 国際公開日
2004 年 11 月 18 日 (18.11.2004)

PCT

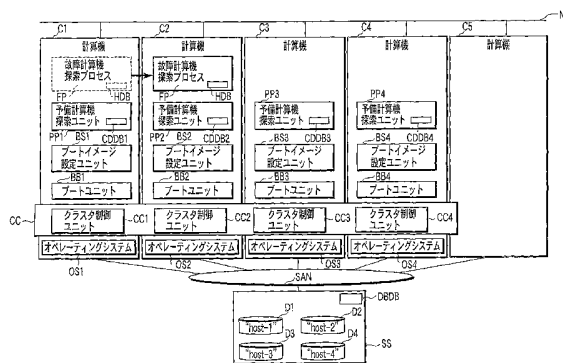
(10) 国際公開番号
WO 2004/099990 A1

- (51) 国際特許分類: G06F 11/20
- (21) 国際出願番号: PCT/JP2004/006500
- (22) 国際出願日: 2004 年 5 月 7 日 (07.05.2004)
- (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2003-132117 2003 年 5 月 9 日 (09.05.2003) JP
- (71) 出願人 (米国を除く全ての指定国について): 株式会社 東芝 (KABUSHIKI KAISHA TOSHIBA) [JP/JP]; 〒1058001 東京都港区芝浦一丁目 1 番 1 号 Tokyo (JP). 東芝ソリューション株式会社 (TOSHIBA SOLUTIONS CORPORATION) [JP/JP]; 〒1056691 東京都港区芝浦一丁目 1 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): 溝口 研一 (MI-ZOGUCHI, Kenichi) [JP/JP].
- (74) 代理人: 鈴江 武彦, 外 (SUZUYE, Takehiko et al.); 〒1000013 東京都千代田区霞が関 3 丁目 7 番 2 号 鈴榮特許綜合法律事務所内 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, AT, AU, AZ, BA, BB, BG, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NA, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RU, SC, SD, SE, SG, SK, SL, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, YU, ZA, ZM, ZW.

/ 続葉有 /

(54) Title: COMPUTER SYSTEM AND DAMAGED COMPUTER REPLACING CONTROL METHOD TO BE APPLIED FOR THE SYSTEM

(54) 発明の名称: 計算機システム及び同システムに適用される故障計算機代替制御方法



C1...COMPUTER
FP...DAMAGED COMPUTER SEARCH PROCESS
PP1...SPARE COMPUTER SEARCH UNIT
BS1...BOOT IMAGE SETTING UNIT
BB1...BOOT UNIT
CC1...CLUSTER CONTROL UNIT
OS1...OPERATING SYSTEM
C2...COMPUTER
PP2...SPARE COMPUTER SEARCH UNIT
BS2...BOOT IMAGE SETTING UNIT
BB2...BOOT UNIT
CC2...CLUSTER CONTROL UNIT
OS2...OPERATING SYSTEM
C3...COMPUTER
PP3...SPARE COMPUTER SEARCH UNIT
BS3...BOOT IMAGE SETTING UNIT
BB3...BOOT UNIT
CC3...CLUSTER CONTROL UNIT
OS3...OPERATING SYSTEM
C4...COMPUTER
PP4...SPARE COMPUTER SEARCH UNIT
BS4...BOOT IMAGE SETTING UNIT
BB4...BOOT UNIT
CC4...CLUSTER CONTROL UNIT
OS4...OPERATING SYSTEM
C5...COMPUTER

(57) Abstract: A damaged computer search process (FP) operating on a computer (C1) is started on another computer such as a computer (C2) if trouble occurs in the computer (C1). When the search process (FP) finds a damaged computer (C1), it searches a spare computer (C5) by using a spare computer search unit (PP2). A boot image setting unit (BS2) selects a storage unit (D1) for storing a boot image which has been executed in the damaged computer (C1). A boot unit (BB2) makes the selected spare computer (C5) perform boot by using the boot image stored in the selected storage unit (D1).

(57) 要約: 計算機 (C1) 上で動作する故障計算機探索プロセス (FP) は、当該計算機 (C1) で障害が発生すると、他の計算機、例えば計算機 (C2) で起動される。探索プロセス (FP) は、故障計算機 (C1) を見つけると、予備計算機探索ユニット (PP2) を用いて予備計算機 (C5) を探す。ブートイメージ設定ユニット (BS2) は、故障計算機 (C1) で実行されていたブートイメージを記憶する記憶ユニット (D1) を選択する。ブートユニット (BB2) は、選択された記憶ユニット (D1) に記憶されているブートイメージを用いて、選択された予備計算機 (C5) をブートさせる。



(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, HU, IE, IT, LU, MC, NL, PL, PT, RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

2文字コード及び他の略語については、定期発行される各PCTガゼットの巻頭に掲載されている「コードと略語のガイダンスノート」を参照。

添付公開書類:

— 国際調査報告書

1

明 細 書

計算機システム及び同システムに適用される故障計算機代替
制御方法

技術分野

本発明は、プロビジョニングノードと呼ばれる予備計算機を含む複数の計算機で構成される計算機システムに係わる。本発明は特に、稼動中の計算機が故障した場合に、その故障した計算機で実行されていたブートイメージを予備計算機で実行可能とするのに好適な計算機システム及び同システムに適用される故障計算機代替制御方法に関する。

背景技術

近年、1つの筐体内に数十乃至数百もの計算機ノードを含む高密度型の計算機システムが市販されている。このような計算機システムには、プロビジョニングノードと呼ばれる予備計算機が含まれることが多い。一般に、予備計算機は通常は使用されず、通常使用している計算機が故障した場合の代替計算機（代替ノード）として使用される。そのためには、故障した計算機で実行されていたブートイメージを、予備計算機のブートイメージとして設定するためのオペレータ操作が必要となる。この設定の後、予備計算機を起動することにより、当該予備計算機を故障した計算機の代わりとして使用することができる。

一方、金子哲夫、森良哉、“クラスタソフトウェア”、東芝レビュー、Vol.54 No.12(1999)、p.18-21（以下、先行技術文献と称する）には、クラスタシステムと呼ばれる計算機シ

2

システムが記載されている。クラスタシステムにおいては、稼動状態にある計算機で障害が発生した場合に、その障害が発生した計算機で実行されていたサービス（業務）を、システム内の他の計算機に引き継がせる（フェイルオーバーする）ことができる。このサービスを引き継ぐ計算機には、例えば待機状態（ホットスタンバイ状態）にある計算機が用いられる。

上記したように、予備計算機（プロビジョニングノード）を含む従来の計算機システムでは、システム内の計算機が故障した場合、予備計算機を故障した計算機の代わりとして使用することができる。しかし、従来の技術では、予備計算機を故障した計算機の代わりとして使用可能とするには、オペレータの介在が必要である。

一方、上記先行技術文献に記載されたクラスタシステムでは、故障した（障害が発生した）計算機で実行されていたサービスを、システム内の他の計算機に自動的に引き継がせる（フェイルオーバーする）ことが可能である。

しかし、このクラスタシステムでは、サービスの引き継ぎ先の計算機は起動（ブート）されている必要がある。このため、クラスタシステムで適用されるサービスを引き継ぐ技術を、予備計算機を故障した計算機の代わりとして使用可能とする仕組みに応用することは困難である。

発明の開示

本発明は上記事情を考慮してなされたものでその目的は、計算機システムに含まれる計算機が故障した場合に、オペレ

一タの介在なしで、予備計算機を故障した計算機の代わりとして使用できるようにすることにある。

本発明の1つの観点によれば、予備計算機を含む複数の計算機を備えた計算機システムが提供される。この計算機システムは、前記複数の計算機をブートするための各ブートイメージをそれぞれ個別に記憶する複数のストレージ装置と、前記複数の計算機のステータスを記憶する第1の記憶ユニットと、前記複数のストレージ装置の各々と当該ストレージ装置に記憶されているブートイメージでブートされている計算機との対応関係を表す情報を記憶する第2の記憶ユニットと、前記計算機システム内で故障した計算機を探索するように構成された故障計算機探索ユニットと、前記故障計算機探索ユニットが故障した計算機を見つけたとき、前記故障した計算機の代替として使用するための予備計算機を、前記第1の記憶ユニットに記憶されている前記複数の計算機のステータスに従って選択するように構成された予備計算機探索ユニットと、前記故障計算機探索ユニットが故障した計算機を見つけたとき、前記予備計算機探索ユニットが選択した予備計算機をブートするためのブートイメージが記憶されたストレージ装置を、前記第2の記憶ユニットに記憶されている情報に従って選択するように構成されたブートイメージ選択ユニットと、前記予備計算機探索ユニットが選択した予備計算機を、前記ブートイメージ選択ユニットが選択したストレージ装置に記憶されているブートイメージを用いてブートするように構成されたブートユニットとを具備する。

図面の簡単な説明

図 1 は本発明の第 1 の実施形態に係る計算機システムの構成を示すブロック図。

図 2 は図 1 中のデータベース C D D B i ($i = 1, 2, 3, 4$) のデータ構造例を示す図。

図 3 は計算機の状態の遷移を示す状態遷移図。

図 4 は図 1 中のデータベース D B D B のデータ構造例を示す図。

図 5 は図 1 中のデータベース H D B のデータ構造例を示す図。

図 6 は同第 1 の実施形態の主として故障計算機探索プロセス F P の処理手順を示すフローチャート。

図 7 は同第 1 の実施形態において故障計算機探索プロセス F P が計算機 C 1 から計算機 C 2 にフェイルオーバーされた状態を示す図。

図 8 は図 6 中のステップ S 5 の実行後のデータベース D B D B の内容を示す図。

図 9 は図 6 中のステップ S 5 の実行後のデータベース C D D B i の内容を示す図。

図 1 0 は故障した計算機 C 1 が実行していたブートイメージを計算機 C 5 によって実行することが可能となった状態を示す図。

図 1 1 は本発明の第 2 の実施形態に係る計算機システムの構成を示すブロック図。

図 1 2 は同第 2 の実施形態の主として故障計算機探索プロ

セス F P の処理手順を示すフローチャート。

図 1 3 は本発明の第 3 の実施形態に係る計算機システムの構成を示すブロック図。

図 1 4 は同第 3 の実施形態の主として故障計算機探索プロセス F P 1 の処理手順を示すフローチャート。

図 1 5 は本発明の第 4 の実施形態に係る計算機システムの構成を示すブロック図。

図 1 6 は同第 4 の実施形態の主として故障計算機探索プロセス F P 1 の処理手順を示すフローチャート。

発明を実施するための最良の形態

以下、本発明の実施の形態につき図面を参照して説明する。図 1 は本発明の第 1 の実施形態に係る計算機システムの構成を示すブロック図である。図 1 の計算機システムは、5 台の計算機 C 1 乃至 C 5 から構成される。計算機 C 1 乃至 C 5 はネットワーク N で相互に接続されている。また計算機 C 1 乃至 C 5 はストレージ・エリア・ネットワーク S A N に接続されている。このストレージ・エリア・ネットワーク S A N には、ストレージ装置 S S も接続されている。ストレージ装置 S S には、ディスク（ディスクドライブ）D 1 乃至 D 4 が設けられている。つまり、計算機 C 1 乃至 C 5 とストレージ装置 S S 内のディスク D 1 乃至 D 4 とはストレージ・エリア・ネットワーク S A N で接続されている。

ディスク D 1 , D 2 , D 3 及び D 4 には、それぞれホスト名として “host-1” , “host-2” , “host-3” 及び “host-4” が割り当てられる計算機のためのブートイメージが予め保存

されている。ディスク D_i ($i = 1, 2, 3, 4$) に保存されているブートイメージを“host- i ”のブートイメージと呼ぶ。“host- i ”のブートイメージは、例えばオペレーティングシステム OS_i 、及び当該オペレーティングシステム OS_i のもとで動作するアプリケーションプログラムを含む。つまりディスク D_i は、“host- i ”のブートディスクとして用いられる。

図 1 には、計算機 C_1 乃至 C_5 のうちの計算機 C_1 乃至 C_4 が、各々ディスク D_1 乃至 D_4 が保存しているブートイメージを用いて起動して稼動している状態が示されている。また、図 1 には、稼動中の計算機 C_1 乃至 C_4 上で、それぞれオペレーティングシステム OS_1 乃至 OS_4 が動作している状態も示されている。これら稼動中の計算機 C_1 乃至 C_4 は、自身のホスト名を、それぞれディスク D_1 乃至 D_4 のブートイメージ中に記録された“host-1”乃至“host-4”と認識しているものとする。一方、計算機 C_1 乃至 C_5 のうちの残りの計算機 C_5 は、図 1 の状態では、プロビジョニングノードと呼ばれる予備計算機として配置されている。つまり計算機 C_5 には、オペレーティングシステムを含むブートイメージはロードされておらず、当該計算機 C_5 はブートされていない。なお、予備計算機は 1 台である必要はなく、複数の予備計算機が配置される構成であっても構わない。

ストレージ装置 SS は、データベース $DBDB$ を有する。データベース $DBDB$ は、ストレージ装置 SS 内のディスク D_1 乃至 D_4 と当該ディスク D_1 乃至 D_4 をブートディスク

とする計算機（ブート計算機）との対応関係を管理するための管理テーブル（管理情報）を保持する。

稼動中の計算機 C 1 乃至 C 4 では、それぞれ、予備計算機探索ユニット P P 1 乃至 P P 4、ブートイメージ設定ユニット B S 1 乃至 B S 4、及びブートユニット B B 1 乃至 B B 4 が動作する。予備計算機探索ユニット P P i（ $i = 1, 2, 3, 4$ ）は、予備計算機を探索する。予備計算機探索ユニット P P i は、データベース C D D B i を有する。データベース C D D B i は、図 1 の計算機システム内の各計算機 C 1 乃至 C 5 の状態とディスク D 1 乃至 D 4 との関係を管理するための管理テーブル（管理情報）を保持する。

ブートイメージ設定ユニット B S i は、ストレージ装置 S S 内のディスク D i に保存されているブートイメージが、ホスト名として“host-i”が割り当てられる計算機で実行されるように、当該ブートイメージと計算機との対応関係の設定（ここでは間接的設定）を行う。このユニット B S i によるブートイメージと計算機との対応関係の設定（ブートイメージ設定）は、故障計算機探索プロセス F P からの指示に従って行われる。ブートユニット B B i は、ホスト名として“host-i”が割り当てられる計算機を、ブートイメージ設定ユニット B S i によって設定されたブートイメージによりブートさせる。

稼動中の計算機 C 1 乃至 C 4 のうちのいずれか 1 つ、例えば計算機 C 1 では、故障計算機探索プロセス F P が動作する。故障計算機探索プロセス F P は、データベース H D B を利用

して故障計算機を探索する故障計算機探索ユニットとして機能する。データベースHDBは、“host-i”が割り当てられている計算機が稼動（生存）しているかを管理するための管理テーブル（管理情報）を保持する。

計算機C_i上で動作する、予備計算機探索ユニットPP_i、ブートイメージ設定ユニットBS_i及びブートユニットBB_iは、例えばプログラムユニットである。本実施形態では、各プログラムユニットは故障計算機代替制御プログラムに含まれている。ユニットPP_i、BS_i及びBB_iは、故障計算機代替制御プログラムを計算機C_i（内の図示せぬCPU）が読み取って実行することにより実現される。計算機C₁上で動作する故障計算機探索プロセスFPも故障計算機代替制御プログラムに含まれるプログラムユニットである。探索プロセスFPも、故障計算機代替制御プログラムを計算機C₁が読み取って実行することにより実現される。したがって、計算機C₁以外の計算機で探索プロセスFPが動作することもあり得る。

稼動中の計算機C₁乃至C₄ではまた、クラスタ制御ユニットCC₁乃至CC₄が動作する。クラスタ制御ユニットCC₁乃至CC₄は、ネットワークNを介して相互に通信を行うことにより、障害が発生した計算機（障害計算機）を検出する。ここでは、クラスタ制御ユニットCC₁乃至CC₄は、ハートビート信号と呼ばれる信号を相互に且つ定期的に送信する。クラスタ制御ユニットCC₁乃至CC₄は、ハートビート信号が予め定められた一定時間（タイムアウト時間）を

超えて途絶えたことをもって、対応するクラスタ制御ユニットが存在する計算機での障害発生を判断する。クラスタ制御ユニット C C 1 乃至 C C 4 は、1 つの仮想的なクラスタ制御システム C C を構成する。クラスタ制御システム C C は、障害が検出された計算機（つまり障害計算機）で実行されていたサービスを他の計算機に引き継がせるための制御を行う。本実施形態では、クラスタ制御システム C C により制御されるサービスの 1 つとして、故障計算機探索プロセス F P が予め定義されている。つまり、探索プロセス F P は、当該プロセス F P が動作している計算機（図 1 では計算機 C 1）が障害等により停止した場合に、他の計算機で起動されるようにクラスタ制御システム C C により制御される。

クラスタ制御ユニット C C i は、クラスタ制御用のソフトウェアプログラム（クラスタソフトウェア）を計算機 C i が読み取って実行することにより実現される。本実施形態では、クラスタソフトウェアと故障計算機代替制御プログラムとは、それぞれ独立のプログラムであるものとする。しかし、例えば故障計算機代替制御プログラムに相当するコード情報をクラスタソフトウェアに予め組み込むことも可能である。

図 2 は、図 1 中のデータベース C D D B i （ $i = 1, 2, 3, 4$ ）のデータ構造例を示す。図 2 に示すように、データベース C D D B i の各レコードは、計算機、当該計算機のステータス、及びディスクの各情報を設定するための項目を含む。計算機の情報には、当該計算機を識別するための計算機識別子が用いられる。ステータスは、対応する計算機の状態

を示し、“稼動”（S），“プロビジョニング”（P），“ダウン”（D）または“リザーブ”（R）を表す。ステータスが“稼動”（S）の場合、対応するレコード中のディスクの項目には、当該レコード中の計算機識別子で示される計算機の起動に用いられたブートイメージを保存しているディスク（ストレージ装置）を識別するためのディスク識別子が設定される。

図2のデータベースC D D B iは、図1の計算機システムの状態を示している。まず、計算機C 1乃至C 4は“稼動”（S）の状態、つまり、それぞれディスクD 1乃至D 4をブートディスクとして稼動している状態にある。一方、計算機C 5は“プロビジョニング”（P）の状態、つまり予備計算機の状態にある。

図3は、計算機の状態の遷移を示す状態遷移図である。ここでは、上記4つの計算機状態の他に、システムに物理的に組み込まれていない状態（－）が示されている。図3に示すように、本実施形態では、“プロビジョニング”（P）の状態にある計算機（予備計算機）は、“リザーブ”（R）の状態を経て“稼動”（S）の状態に切り替え可能である。

図4は、図1中のデータベースD B D Bのデータ構造例を示す。図4に示すように、データベースD B D Bの各レコードは、ブートイメージを保存しているディスク（ブートディスク）の情報（例えばディスク識別子）、及び当該ディスクをブートディスクとする計算機の情報（例えば計算機識別子）を設定するための項目を含む。

図5は、図1中のデータベースHDBのデータ構造例を示す。図5に示すように、データベースHDBの各レコードは、ホストを示す情報（例えばホスト名）、及びカウンタの情報（カウンタ値）を設定するための項目を含む。ここでは、クラスタ制御ユニットCCIからハートビート信号がタイムアウト時間内に送信される毎に、当該制御ユニットCCIが動作しているホストに対応するカウンタが1インクリメントされる。

次に、図1の計算機システムにおける動作について、図6を参照して説明する。図6は主として故障計算機探索プロセスFPの処理手順を示すフローチャートである。ここでは、故障計算機を検出し、その故障計算機に適用されていたブートイメージを用いて予備計算機を起動する動作を例に挙げる。今、図1の計算機システムでは、計算機C1乃至C4が、それぞれ“host-1”乃至“host-4”のブートイメージに従って稼動している状態、つまり“稼動”（S）の状態にあるものとする。“host-1”乃至“host-4”のブートイメージは、それぞれストレージ装置SS内のディスクD1乃至D4に保存されている。また、計算機C5は予備計算機として計算機システム内に配置されており、“プロビジョニング”（P）の状態にあるものとする。この場合、稼動状態にある計算機C1乃至C4内の予備計算機探索ユニットPP1乃至PP4が有するデータベースCddb1乃至Cddb4の内容は、いずれも図2に示したようになっている。また、ストレージ装置SS内のデータベースDddbの内容は、図4に示したよ

うになっている。

図1の計算機システムにおいて、クラスタ制御システムC C内の各クラスタ制御ユニットC C i ($i = 1, 2, 3, 4$)、即ち各計算機C i上で動作しているクラスタ制御ユニットC C iは、相互にハートビート信号を送受信する。クラスタ制御ユニットC C iは他のクラスタ制御ユニットC C j ($j = 1, 2, 3, 4$ 、但し $j \neq i$) からタイムアウト時間内にハートビート信号を受信できた場合、データベースH D B内のレコードのうち、クラスタ制御ユニットC C jが動作しているホストに対応するレコード中のカウンタを1インクリメントする。クラスタ制御システムC Cは、クラスタ制御ユニットC C jからのハートビート信号がタイムアウト時間を超えて途絶えた場合、当該クラスタ制御ユニットC C jが動作している計算機C jの障害発生を判断する。

ここでは、クラスタ制御システムC Cにより障害発生と判断された計算機C jが、故障計算機探索プロセスF Pが動作している計算機C 1、即ちホスト名“host-1”が割り当てられている計算機C 1であるものとする。本実施例において、探索プロセスF Pは、当該プロセスF Pが動作している計算機が障害発生計算機と判断された場合に、他の計算機に引き継ぐべき（フェイルオーバーすべき）サービスとして定義されている。また、本実施形態のように探索プロセスF Pが計算機C 1で動作している場合、計算機C 2乃至C 4は待機系計算機として位置付けられる。これにより、計算機C 1が障害発生計算機と判断された場合、探索プロセスF Pは計算機C

13

2乃至C4のいずれかに引き継がれる。ここでは、探索プロセスFPは、図7に示すように、障害発生計算機C1から計算機C2に引き継がれるものとする。以降、探索プロセスFPは、計算機C2上で動作する。

さて、計算機Cjの障害が発生して、クラスタ制御ユニットCCjからのハートビート信号が途絶えた状態が長時間続くものとする。この場合、計算機Cjに対応する上記HDB内のカウンタは長時間変化しない。故障計算機探索プロセスFPは、このような状態の検出と、計算機Cjから予備計算機への切り替えとを、次の手順で実行する。

まず故障計算機探索プロセスFPは、起動されると、予め定められた一定時間スリープする（ステップS1）。次に探索プロセスFPは、データベースHDBを参照して、上記一定時間の間にカウンタが変化していないホストを探す（ステップS2）。もし、カウンタが変化していないホストがないならば、つまり各ホストに対応するカウンタがいずれも変化しているならば、探索プロセスFPは、故障したホストはないものと判断する。この場合、探索プロセスFPはステップS1に戻ってスリープする。以降、探索プロセスFPは、カウンタが変化していないホストがない限り、ステップS1及びS2の処理を繰り返す。

これに対し、カウンタが変化していないホストがあるならば、探索プロセスFPは、当該ホストのクラスタ制御ユニットからのハートビート信号が途絶えている要因を判別する。この要因は、ホストが故障したためと、当該ホストがリブー

トされるための何れかである。探索プロセス F P は、この要因の判別のために、リブートに必要な一定時間スリープする（ステップ S 3）。その後、探索プロセス F P は、改めてデータベース H D B を参照することにより、先に、変化していないと判定されたホストのカウンタが依然として変化していないかを判定する（ステップ S 4）。もし、上記カウンタが依然として変化していない場合には、探索プロセス F P は、対応するホストは生存しておらず、したがって故障したものと判定する。ここでは、ホスト名 “host-1” が割り当てられている計算機 C 1 の故障が判定されるものとする。

このとき探索プロセス F P は、クラスタ制御システム C C による制御で、計算機 C 1 ではなくて計算機 C 2 で動作している（図 7 参照）。探索プロセス F P は、計算機 C 1 の故障を判定した場合、予備計算機探索ユニット P P 2 により予備計算機（プロビジョニングノード）を探させる。この予備計算機探索ユニット P P 2 による予備計算機の探索は、次のように行われる。

まず予備計算機探索ユニット P P 2 は、データベース C D D B 2 を参照する。そして探索ユニット P P 2 は、データベース C D D B 2 から、ステータスが “プロビジョニング”

（P）の状態にある計算機の計算機識別子を取得する。データベース C D D B 2（C D D B 1 乃至 C D D B 4）が図 2 のようになっている本実施形態では、“プロビジョニング”

（P）の状態にある計算機の計算機識別子は C 5 である。この場合、計算機 C 5 が予備計算機として検出（選択）される。

探索ユニット P P 2 は、計算機 C 5 を予備計算機として検出（選択）すると、データベース C D D B 2 を次のように操作する。即ち探索ユニット P P 2 は、計算機 C 1 が故障計算機で、計算機 C 5 が予備計算機である場合、故障計算機 C 1 及び予備計算機 C 5 のステータスを、それぞれ、“ダウン”（D）及び“リザーブ”（R）に変更する。この探索ユニット P P 2 のデータベース操作内容は、稼動状態にある他の計算機 C 3 及び C 4 の探索ユニット P P 3 及び P P 4 が有するデータベース C D D B 3 及び C D D B 4 に反映される。これにより、データベース C D D B 3 及び C D D B 4 の内容が、データベース C D D B 2 の内容に一致するように変更される。この変更後のデータベース C D D B 2 乃至 C D D B 4（C D D B i）の内容を図 9 に示す。

さて、予備計算機 C 5 が検出（選択）された場合、故障計算機探索プロセス F P は、故障した計算機 C 1 が使用していたブートイメージを当該予備計算機 C 5 に使用させるための制御を行う。ここでは、探索プロセス F P は、計算機 C 1 が使用していたブートイメージを予備計算機 C 5 で実行させるために、ストレージ装置 S S 内のデータベース D B D B をブートイメージ設定ユニット B S 2 により操作させる（ステップ S 5）。

ブートイメージ設定ユニット B S 2 は、データベース D B D B 中のレコードのうち、故障計算機 C 1 がブート計算機として設定されているレコード、つまり計算機 C 1 の情報（計算機識別子）を含むレコードを選択する。この選択されたレ

コードには、図4から明らかなように、計算機C1の情報（計算機識別子）と対をなしてディスクD1の情報（ディスク識別子）が設定されている。したがって、選択されたレコードは、故障計算機C1をブートするのに用いられていたブートイメージがディスクD1に記憶されていることを表す。

次に、ブートイメージ設定ユニットBS2は、ディスクD1に記憶されているブートイメージを、予備計算機C5で実行させるために、選択されたレコードを次のように操作（更新）する。即ち設定ユニットBS2は、選択されたレコードを対象に、ディスクD1の情報（ディスク識別子）と対をなすブート計算機の情報（計算機識別子）を、故障計算機C1から予備計算機C5の情報に変更する。この設定ユニットBS2によるレコードの更新操作、つまりデータベースDBDBの更新操作は、故障計算機C1をブートするのに用いられていたブートイメージを、予備計算機C5をブートするためのブートイメージとして選択することと等価である。つまり、設定ユニットBS2は、ブートイメージ選択ユニットとして機能する。この設定ユニットBS2によるブートイメージ選択操作（データベースDBDBの操作）により、故障計算機C1で実行されていた“host-1”のブートイメージ（ディスクD1に記憶されているブートイメージ）が、予備計算機C5に間接的に設定されたことになる。このときのデータベースDBDBの内容を図8に示す。

故障計算機探索プロセスFPは、ブートイメージ設定ユニットBS2を用いてステップS5を実行すると、ブートユニ

ット B B 2 により、予備計算機 C 5 を上記選択されたブートイメージでブートさせる（ステップ S 6）。以下、この予備計算機 C 5 をブートする動作について詳述する。まず、予備計算機 C 5 は、図示しないインタフェース回路を持っている。このインタフェース回路は、ネットワーク N と接続するための制御回路として動作する。このインタフェース回路は、ネットワーク N から自身宛に送信される特別なパケットの受信をするために待機している。そのためインタフェース回路には、常時スタンバイ電流が供給されている。インタフェース回路は、ネットワーク N を介して自身宛の特別なパケットを受信した場合に、当該インタフェース回路を持つ計算機（ここでは予備計算機 5）を起動（ブート）するための機能を持つ。このようなインタフェース回路を持った予備計算機 C 5 は、常時起動可能な状態（スタンバイ）状態に設定されている。

ブートユニット B B 2 は、予備計算機 C 5 を起動させるために、ネットワーク N を介して予備計算機 C 5 に特別なパケットを送信する。予備計算機 C 5 のインタフェース回路は、この特別なパケットを受信したことに基づいて、予備計算機 C 5 に起動処理（ブートローダーの起動などの処理）を開始させる。この起動処理を開始した予備計算機 C 5 は、自身をブートするためのブートイメージが記録されたディスクを探すために、データベース D B D B を参照する。予備計算機 C 5 は、データベース D B D B を参照して予備計算機 C 5 の識別子が設定されているレコードを探す。予備計算機 C 5 は、

探した予備計算機C5の識別子が設定されているレコードに記録されたディスクD1の識別子に従い、ストレージ装置S5内のディスクD1に保存されている“host-1”のブートイメージを用いてブートする。なお、上記のようにネットワークを介して特別なパケットを特定の計算機に送信して、その特定の計算機を起動させる技術は、Wake on LAN（商標）という名称で広く一般に知られている。

上述した予備計算機C5をブートする動作により、故障した計算機C1が実行していた“host-1”のブートイメージを、計算機C5で実行することが可能になる。つまり、計算機C5が“host-1”として起動される。するとデータベースCDD2が操作されて、計算機C5のステータスが“リザーブ”（R）から“稼動”（S）の状態に変更される。この結果、計算機C1が故障するまで当該計算機C1上で動作していた、オペレーティングシステムOS1、予備計算機探索ユニットPP1、ブートイメージ設定ユニットBS1、ブートユニットBB1及びクラスタ制御ユニットCC1が、図10に示すように計算機C5で動作するようになる。

このように本実施形態では、故障計算機探索プロセスFPが障害発生計算機C1上で動作していても、当該プロセスFPが他の計算機（ここでは計算機C2）に引き継がれる。このため、探索プロセスFPにより計算機C1の故障が確実に判定できる。また、探索プロセスFPの制御により、引き継ぎ先の計算機上で動作する予備計算機探索ユニット、ブートイメージ設定ユニット及びブートユニット（ここでは、予備

計算機探索ユニット P P 2、ブートイメージ設定ユニット B S 2 及びブートユニット B B 2) を利用して、それぞれ、予備計算機の探索、故障計算機 C 1 が使用していたブートイメージを予備計算機が使用可能とするための設定及び予備計算機のブートの自動化を実現できる。

上記実施形態では、ホストの故障検出に、ハートビート信号をタイムアウト時間内に受信する都度インクリメントされるカウンタを用いている。しかし、タイムアウト時間を経過してもハートビート信号が依然として受信されない時間、つまりタイムアウト時間（第 1 のタイムアウト時間）からの経過時間を監視することにより、対応するホストの故障を判定することも可能である。ここでは、第 1 のタイムアウト時間からの経過時間が予め定められた第 2 のタイムアウト時間を超えた場合に、対応するホストの故障を判定すればよい。

[第 2 の実施形態]

図 1 1 は本発明の第 2 の実施形態に係る計算機システムの構成を示すブロック図である。図 1 1 において、図 1 の計算機システムと等価な構成要素には同一符号を付してある。まず、図 1 1 の計算機システムの構成について、図 1 の計算機と相異なる点を中心に説明する。図 1 1 の計算機システムにおいて、ネットワーク N には、リモート配布サーバ R D S が接続されている。リモート配布サーバ R D S は、ディスク（ディスクドライブ）R 1, R 2, R 3 及び R 4 を備えている。このディスク R 1, R 2, R 3 及び R 4 には、それぞれ “host-1”, “host-2”, “host-3” 及び “host-4” のブート

イメージが予め保存されている。このディスク R 1 , R 2 , R 3 及び R 4 に保存されているブートイメージの所定位置には、ホスト名 “host-1” , “host-2” , “host-3” 及び “host-4” が記録されている。ここで、ディスク R 1 , R 2 , R 3 及び R 4 に保存されているブートイメージで、計算機 C 1 , C 2 , C 3 及び C 4 をそれぞれブートするものとする。この場合、計算機 C 1 , C 2 , C 3 及び C 4 のホスト名は、それぞれ “host-1” , “host-2” , “host-3” 及び “host-4” となる。

計算機 C 1 , C 2 , C 3 , C 4 及び C 5 は、それぞれディスク（ローカルディスクドライブ）D 1 , D 2 , D 3 , D 4 及び D 5 を備えている。ディスク D 1 , D 2 , D 3 及び D 4 には、リモート配布サーバ R D S のディスク R 1 , R 2 , R 3 及び R 4 の内容がそれぞれコピーされている。この場合、計算機 C 1 , C 2 , C 3 及び C 4 を起動すると、“host-1” , “host-2” , “host-3” 及び “host-4” で特定されるホストが起動する。

次に、図 1 1 の計算機システムにおける動作について、図 1 2 を参照して説明する。図 1 2 は主として故障計算機探索プロセス F P の処理手順を示すフローチャートである。ここでは、上記第 1 の実施形態と同様に、故障計算機を検出し、その故障計算機に適用されていたブートイメージを予備計算機に設定して当該予備計算機を起動する動作を例に挙げる。今、計算機 C 1 乃至 C 4 が、それぞれ、ディスク D 1 乃至 D 4 にコピーされた “host-1” 乃至 “host-4” のブートイメー

21

ジに従って稼動している状態、つまり“稼動”(S)の状態にあるものとする。また、計算機C5は予備計算機として計算機システム内に配置されており、“プロビジョニング”(P)の状態にあるものとする。更に、計算機C1上では、故障計算機探索プロセスFPが動作しているものとする。探索プロセスFPは、計算機C1乃至C4のいずれにも障害が発生しない通常状態では、図6中のステップS1及びS2に相当する処理(ステップS11及びS12)を繰り返す。

このような状態で、上記第1の実施形態と同様に計算機C1が故障したものとする。この場合、探索プロセスFPは、図6中のステップS1、S2、S3及びS4に相当する処理(ステップS11、S12、S13及びS14)により、計算機C1の故障を判断する。探索プロセスFPは、次に述べる段階で、待機系計算機C2乃至C4の何れか、例えば計算機C2に引き継がれる。この段階とは、計算機C1上で動作するクラスタ制御ユニットCC1からのハートビート信号がタイムアウト時間を越えて途絶えた結果、クラスタ制御システムCCにより当該計算機C1の障害発生が検出された段階である。

探索プロセスFPは、計算機C1の故障を判断すると、予備計算機探索ユニットPP2により予備計算機を探させる。ここでは、上記第1の実施形態と同様に、計算機C5が予備計算機として見つけられた(選択された)ものとする。すると、探索プロセスFPは、故障した計算機C1が使用していた“host-1”のブートイメージを予備計算機C5に使用させ

22

るための制御を行う。ここでは、探索プロセス F P は、“host-1”のブートイメージを予備計算機 C 5 で実行させるために、当該ブートイメージを、ブートイメージ設定ユニット B S 2 により当該予備計算機 C 5 のローカルディスク D 5 にコピーさせる（ステップ S 1 5）。“host-1”のブートイメージは、リモート配布サーバ R D S 内のディスク R 1 に保存されている。そこで、ブートイメージ設定ユニット B S 2 は、リモート配布サーバ R D S からディスク R 1 を選択する。そしてブートイメージ設定ユニット B S 2 は、ディスク R 1 に保存されている“host-1”のブートイメージを、予備計算機 C 5 のローカルディスク D 5 にコピーする（ステップ S 1 5）。このように本実施形態においては、故障計算機 C 1 で実行されていた“host-1”のブートイメージが、故障計算機探索プロセス F P によって検出（選択）された予備計算機 C 5 に直接設定される。この点で、故障計算機 C 1 で実行されていた“host-1”のブートイメージが予備計算機 C 5 に間接的に設定される上記第 1 の実施形態とは異なる。

次に探索プロセス F P は、予備計算機 C 5 のディスク D 5 にコピーされている“host-1”のブートイメージに従って、ブートユニット B B 2 により、当該計算機 C 5 をブートさせる（ステップ S 1 6）。このブートユニット B B 2 による、計算機 C 5 を対象とするブート操作は、上記第 1 の実施形態におけるブート操作と同様に行われる。これにより、計算機 C 5 は“host-1”として起動される。この結果、計算機 C 1 が故障するまで当該計算機 C 1 上で動作していた、オペレー

23

ティングシステムOS 1、予備計算機探索ユニットPP 1、ブートイメージ設定ユニットBS 1、ブートユニットBB 1及びクラスタ制御ユニットCC 1は、当該計算機C 5で動作するようになる。

〔第3の実施形態〕

図13は本発明の第3の実施形態に係る計算機システムの構成を示すブロック図である。図13において、図1の計算機システムと等価な構成要素には同一符号を付してある。まず、図13の計算機システムの構成について、図1の計算機と相異なる点を中心に説明する。図13の計算機システムにおいて、計算機C 1、C 2、C 3及びC 4は、ストレージ装置SS内のディスクD 1、D 2、D 3及びD 4にそれぞれ保存されている“host-1”、“host-2”、“host-3”及び“host-4”のブートイメージに従って稼動している状態、つまり“稼動”（S）の状態にあるものとする。この計算機C 1、C 2、C 3及びC 4上では、故障計算機探索プロセスFP 1、FP 2、FP 3及びFP 4が、それぞれ並行して動作する。探索プロセスFP 1、FP 2、FP 3及びFP 4は、図1中の探索プロセスFPに相当する。

故障計算機探索プロセスFP 1、FP 2、FP 3及びFP 4の特徴は、それぞれホスト名が“host-1”、“host-2”、“host-3”及び“host-4”のブートイメージで起動されている計算機（ここでは計算機C 1、C 2、C 3及びC 4）で動作する点にある。この点で、探索プロセスFP 1、FP 2、FP 3及びFP 4は、上記探索プロセスFP、即ちシステム

24

内のいずれか 1 つの計算機上で唯一動作する探索プロセス F P と異なる。また探索プロセス F P 1 , F P 2 , F P 3 及び F P 4 の特徴は、自身が動作すべき計算機を認識可能な点にもある。つまり探索プロセス F P 1 , F P 2 , F P 3 及び F P 4 は、自身が動作している計算機に割り当てられているホスト名（自身のホスト名）を、それぞれストレージ装置 S S 内のディスク D 1 , D 2 , D 3 及び D 4 に保存されているブートイメージ中に記録された “host-1” , “host-2” , “host-3” 及び “host-4” であると認識している。探索プロセス F P 1 , F P 2 , F P 3 及び F P 4 は、ホスト名の認識機能を利用して、故障計算機を探索する。このため探索プロセス F P 1 , F P 2 , F P 3 及び F P 4 は、探索プロセス F P とは異なり、故障計算機の探索にデータベース H D B を必要としない。

本実施形態では、クラスタ制御システム C C により制御されるサービスとして、故障計算機探索プロセス F P 1 , F P 2 , F P 3 及び F P 4 が予め定義されている。つまり、探索プロセス F P 1 , F P 2 , F P 3 及び F P 4 は、当該プロセス F P 1 , F P 2 , F P 3 及び F P 4 が動作している計算機（図 13 では計算機 C 1 , C 2 , C 3 及び C 4）が障害等により停止した場合に、他の計算機で起動されるようにクラスタ制御システム C C により制御される。

次に、図 13 の計算機システムにおける動作について、図 14 を参照して説明する。図 14 は主として故障計算機探索プロセス F P 1 (F P i) の処理手順を示すフローチャート

である。ここでは、故障計算機探索プロセスF P 1が計算機C 1の故障を検出し、その故障計算機C 1に適用されていたブートイメージを予備計算機に設定して当該予備計算機を起動する動作を例に挙げる。

まず、故障計算機探索プロセスF P 1は、起動されると、自身が動作すべき計算機、つまりホスト名が“host-1”の計算機C 1で動作しているか否かを判定する（ステップS 21）。通常は、探索プロセスF P 1は、図13に示したように計算機C 1で動作している。この場合、探索プロセスF P 1は、次に起動されるまでスリープする（ステップS 28）。

今、探索プロセスF P 1が動作している計算機C 1で障害が発生したものとする。すると、探索プロセスF P 1が、クラスタ制御システムC Cの制御により、障害発生計算機C 1から計算機システム内の他の計算機に移される（フェイルオーバーされる）。つまり、探索プロセスF P 1が起動される箇所が、障害発生計算機C 1から他の計算機に変更される。これにより、探索プロセスF P 1は、自身が動作すべきホスト名“host-1”の計算機（C 1）とは異なる計算機で動作することになる。ここでは、探索プロセスF P 1は、ホスト名が“host-2”の計算機C 2で起動されたものとする。

探索プロセスF P 1は計算機C 2で起動された場合、上記ステップS 21において、自身が動作すべきホスト名が“host-1”の計算機C 1ではない計算機（ホスト名が“host-2”の計算機C 2）で動作しているものと判断する。すると探索プロセスF P 1は、元々自身が動いていたホスト名が

“host-1”の計算機C 1に障害が発生したことを認識する。
この場合、探索プロセスF P 1は、例えば自身が現在動作している計算機C 2上の予備計算機探索ユニットP P 2を用いて、予備計算機を探す（ステップS 2 2，S 2 3）。この探索ユニットP P 2による予備計算機の探索は、上記第1の実施形態と同様に、データベースC D D B 2を参照して、ステータスが“プロビジョニング”（P）の状態にある計算機の計算機識別子を取得することで実現される。もし、“プロビジョニング”（P）の状態にある計算機が存在しない場合には、故障計算機探索プロセスF P 1は、一定時間スリープした後（ステップS 2 4）、予備計算機探索ユニットP P 2を用いて再び予備計算機を探す（ステップS 2 2，S 2 3）。

今、データベースC D D B 2（C D D B i）が図2のようになっているものとする、計算機C 5が予備計算機として検出される。故障計算機探索プロセスF P 1は予備計算機C 5が検出（選択）された場合、図6中のステップS 5及びS 6に相当する処理（ステップS 2 5及びS 2 6）を実行する。即ち探索プロセスF P 1は、故障した計算機C 1が使用していた“host-1”のブートイメージを予備計算機C 5に使用させるための制御を行う。ここでは、探索プロセスF P 1は、“host-1”のブートイメージを予備計算機C 5で実行させるために、ストレージ装置S S内のデータベースD B D Bを、例えば自身が現在動作している計算機C 2上のブートイメージ設定ユニットB S 2により操作させる（ステップS 2 5）。このデータベースD B D Bの操作により、上記第1の実施形

態と同様に、故障計算機 C 1 をブートするのに用いられていたブートイメージが、予備計算機 C 5 をブートするためのブートイメージとして選択されることになる。次に探索プロセス F P 1 は、自身が現在動作している計算機 C 2 上のブートユニット B B 2 により、予備計算機 C 5 を上記選択された“host-1”のブートイメージでブートさせる（ステップ S 2 6）。これにより、故障した計算機 C 1 が実行していた“host-1”のブートイメージを、計算機 C 5 で実行することが可能となる。即ち、計算機 C 5 が“host-1”として起動され、計算機 C 1 が故障するまで当該計算機 C 1 上で動作していた、オペレーティングシステム O S 1、予備計算機探索ユニット P P 1、ブートイメージ設定ユニット B S 1、ブートユニット B B 1 及びクラスタ制御ユニット C C 1 が、当該計算機 C 5 で動作するようになる。

このように、故障した計算機 C 1 が実行していた“host-1”のブートイメージで予備計算機 C 5 が起動されたものとする。つまり、計算機 C 5 が“host-1”として起動されたものとする。すると、本来“host-1”の計算機で動作すべき故障計算機探索プロセス F P 1 が、ホスト名が“host-2”の計算機 C 2 から、“host-1”として起動された、ホスト名が“host-1”の計算機 C 5 に移さ（戻さ）れる（ステップ S 2 7）。この探索プロセス F P 1 を、当該プロセス F P 1 が本来動作すべきホスト名が“host-1”の計算機に移す（戻す）処理（つまりフェイルバック処理）は、クラスタ制御システム C C により行われる。その後、探索プロセス F P 1 は、次

に起動されるまでスリープする（ステップ S 28）。

上記第 3 の実施形態では、予備計算機探索ユニット P P 1（P P i）、ブートイメージ設定ユニット B S 1（B S i）及びブートユニット B B 1（B B i）は故障計算機探索プロセス F P 1（F P i）から独立している。しかし、予備計算機探索ユニット P P 1（P P i）、ブートイメージ設定ユニット B S 1（B S i）及びブートユニット B B 1（B B i）が故障計算機探索プロセス F P 1（F P i）に付随する（含まれる）構成とすることも可能である。この構成では、クラスタ制御システム C C の制御により故障計算機探索プロセス F P 1 が計算機 C 1 から計算機 C 2 に移される際に、予備計算機探索ユニット P P 1、ブートイメージ設定ユニット B S 1 及びブートユニット B B 1 も計算機 C 2 に移される。この場合、故障計算機探索プロセス F P 1 は、図 14 のフローチャートに示す処理を、予備計算機探索ユニット P P 1、ブートイメージ設定ユニット B S 1 及びブートユニット B B 1 を用いて実行する。

また、上記第 3 の実施形態では、計算機 C 1 の一時的な障害の場合にも、故障計算機探索プロセス F P 1 が計算機 C 2 に移されて、計算機 C 1 で実行されていたブートイメージで予備計算機 C 5 が起動される可能性がある。そこでクラスタ制御システム C C は、故障計算機探索プロセス F P 1（F P i）に関しては、他のサービスとは異なり、計算機障害の検出に用いられるタイムアウト時間より長いタイムアウト時間を用いて、障害発生計算機から他の計算機への移動制御（フ

ェイルオーバ)を行うとよい。

〔第4の実施形態〕

図15は本発明の第4の実施形態に係る計算機システムの構成を示すブロック図である。図15において、図11または図13の計算機システムと等価な構成要素には同一符号を付してある。図15の計算機システムが第3の実施形態で適用された図13の計算機システムと相異なる点は、第2の実施形態で適用された図11の計算機システムが、第1の実施形態で適用された図1の計算機システムと相異なる点に一致する。この図15の計算機システム内の故障計算機探索プロセスFP1 (FPi) の処理手順を図16のフローチャートに示す。図16から明らかなように、故障計算機探索プロセスFP1 (FPi) は、第3の実施形態で適用された図14のフローチャートのステップS21乃至S28に相当する処理(ステップS31乃至S38)を実行する。但し、ステップS37は、図14中のステップS27と同様に、クラスタ制御システムCCにより実行される。図16のフローチャートが図14のフローチャートと相異なるのは、図14中のステップS25に相当する処理(ステップS35)である。つまり、故障計算機C1が使用していたブートイメージを予備計算機C5に使用させるための処理が異なる。ステップS35では、故障計算機C1が使用していたブートイメージ、即ちリモート配布サーバRDS内のディスクR1に保存されているブートイメージを、予備計算機C5のローカルディスクD5にコピーする処理が行われる。この処理は、図12中の

ステップ S 1 5 と同様である。

上記第 1 乃至第 4 の実施形態では、あるブートイメージを実行していた計算機が故障した場合に、その計算機の故障検出をトリガとして、当該計算機が実行していたブートイメージで予備計算機を起動する技術が適用される。この技術を応用して、例えば計算機の負荷が基準値より大きくなったことをトリガとして、当該計算機が実行していたブートイメージで予備計算機を起動する構成とすることも可能である。この場合、故障計算機探索プロセスの引き継ぎは必ずしも必要ない。また、予め定められた時刻の到来をトリガとして、同一計算機を、時間帯によって異なるブートイメージで起動する構成とすることも可能である。つまり、同一計算機を、例えば昼間は第 1 のオペレーティングシステムを含む第 1 のブートイメージで起動し、夜間は第 2 のオペレーティングシステムを含む第 2 のブートイメージで起動することも可能である。この構成は、特に、ブートイメージのコピーを必要としない図 1 または図 1 3 に示した計算機システムに適する。

なお、本発明は、上記実施形態そのままに限定されるものではなく、実施段階ではその要旨を逸脱しない範囲で構成要素を変形して具体化できる。また、上記実施形態に開示されている複数の構成要素の適宜な組み合わせにより種々の発明を形成できる。例えば、実施形態に示される全構成要素から幾つかの構成要素を削除してもよい。更に、異なる実施形態に亘る構成要素を適宜組み合わせてもよい。

産業上の利用可能性

31

本発明によれば、計算機が故障した場合に、オペレータの介在なしで、予備計算機を故障した計算機の代わりとして使用できるよう、当該予備計算機にブートイメージを設定して当該予備計算機を起動することができる。

32

請 求 の 範 囲

1. 予備計算機を含む複数の計算機を備えた計算機システム、前記計算機システムは、

前記複数の計算機をブートするための各ブートイメージをそれぞれ個別に記憶する複数のストレージ装置と、

前記複数の計算機のステータスを記憶する第1の記憶ユニットと、

前記複数のストレージ装置の各々と当該ストレージ装置に記憶されているブートイメージでブートする計算機との対応関係を表す情報を記憶する第2の記憶ユニットと、

前記計算機システム内で故障した計算機を探索するように構成された故障計算機探索ユニットと、

前記故障計算機探索ユニットが故障した計算機を見つけたとき、前記故障した計算機の代替として使用するための予備計算機を、前記第1の記憶ユニットに記憶されている前記複数の計算機のステータスに従って選択するように構成された予備計算機探索ユニットと、

前記故障計算機探索ユニットが故障した計算機を見つけたとき、前記予備計算機探索ユニットが選択した予備計算機をブートするためのブートイメージが記憶されたストレージ装置を、前記第2の記憶ユニットに記憶されている情報に従って選択するように構成されたブートイメージ選択ユニットと、

前記予備計算機探索ユニットが選択した予備計算機を、前記ブートイメージ選択ユニットが選択したストレージ装置に記憶されているブートイメージを用いてブートするように構

33

成されたブートユニットと

を具備する。

2. 請求項1に従う計算機システムにおいて、前記故障計算機探索ユニットは、前記複数の計算機の各々について、当該計算機が正常に稼動しているかを定期的に判定するように構成された第1の判定ユニットと、前記第1の判定ユニットにより正常に稼動していないと判定された計算機が、その判定時点から一定時間を経過しても正常に稼動していない場合に、当該計算機が故障していると判定するように構成された第2の判定ユニットとを含む。

3. 請求項1に従う計算機システムにおいて、前記故障計算機探索ユニットは、前記複数の計算機のうちの稼動状態にあるいずれか1つの計算機で動作して、前記計算機システム内で故障した計算機を探索するように構成された特定プロセスであって、当該プロセスが動作する計算機で障害が発生した場合には、前記予備計算機以外の他の計算機で動作するように構成された特定プロセスを含む。

4. 請求項3に従う計算機システムにおいて、前記計算機システム内で障害が発生した計算機を検出し、当該障害が発生した計算機で実行されていたサービスを前記予備計算機以外の他の計算機に引き継がせるように構成されたクラスタ制御システムであって、前記障害が発生した計算機が前記特定プロセスが動作する計算機の場合に、前記特定プロセスを前記他の計算機に引き継がせるように構成されたクラスタ制御システムを更に具備する。

34

5. 請求項4に従う計算機システムにおいて、前記複数のストレージ装置に記憶されている各ブートイメージでブートされた計算機の各々が稼動しているかを管理するための管理情報を記憶する第3の記憶ユニットを更に具備し、

前記クラスタ制御システムは、前記複数のストレージ装置に記憶されている各ブートイメージでブートされた計算機を定期的に監視することにより、前記第3の記憶ユニットに記憶されている管理情報を更新し、

前記故障計算機探索ユニットは、前記第3の記憶ユニットに記憶されている管理情報に基づいて、前記計算機システム内で故障した計算機を探索する。

6. 請求項5に従う計算機システムにおいて、前記クラスタ制御システムは、前記複数の計算機のうちの稼動状態にある計算機でそれぞれ動作して相互に定期的に通信を行うことにより、障害が発生した計算機を検出するように構成された複数のクラスタ制御ユニットであって、前記通信の結果に応じて前記第3の記憶ユニットに記憶されている管理情報を更新するように構成された複数のクラスタ制御ユニットを含む。

7. 請求項1に従う計算機システムにおいて、前記故障計算機探索ユニットは、前記複数の計算機のうちの稼動状態にある計算機でそれぞれ動作して、前記計算機システム内で故障した計算機を探索するように構成された複数の特定プロセスであって、当該プロセスが動作する計算機で障害が発生した場合には、前記予備計算機以外の他の計算機で動作する

ように構成されている複数の特定プロセスを含む。

8. 請求項7に従う計算機システムにおいて、前記複数の特定プロセスの各々は、自身が前記複数の計算機のうちのいずれの計算機で動作すべきか認識可能であり、自身が動作すべき計算機で動作していない場合、自身が動作すべき計算機は故障していると判定する。

9. 請求項7に従う計算機システムにおいて、前記計算機システム内で障害が発生した計算機を検出し、当該障害が発生した計算機で実行されていたサービスを前記予備計算機以外の他の計算機に引き継がせるように構成されたクラスタ制御システムであって、前記障害が発生した計算機が前記特定プロセスが動作する計算機の場合に、前記特定プロセスを前記他の計算機に引き継がせるように構成されたクラスタ制御システムを更に具備する。

10. 請求項9に従う計算機システムにおいて、前記クラスタ制御システムは、前記複数の計算機のうちの稼動状態にある計算機でそれぞれ動作して相互に定期的に通信を行うことにより、障害が発生した計算機を検出するように構成された複数のクラスタ制御ユニットを含む。

11. 予備計算機を含む複数の計算機で構成される計算機システムに適用される、予備計算機を故障した計算機の代わりとして使用するための方法、前記方法は、

前記計算機システム内で故障した計算機を探索することと、

前記探索することによって、故障した計算機がみつけれられたとき、前記故障した計算機の代替として使用するための予

36

備計算機を、第1の記憶ユニットに記憶されている前記複数の計算機のステータスに従って選択することと、

前記複数の計算機をブートするための各ブートイメージをそれぞれ個別に記憶する複数のストレージ装置の中から、第2の記憶ユニットに記憶されている、前記複数のストレージ装置の各々と当該ストレージ装置に記憶されているブートイメージでブートする計算機との対応関係を表す情報に従って、前記故障した計算機の代替として選択された予備計算機をブートするためのブートイメージが記憶されたストレージ装置を選択することと、

選択された予備計算機を、選択されたストレージ装置に記憶されているブートイメージを用いてブートすることと

を具備する。

12. 請求項11に従う方法において、前記探索することは、前記複数の計算機のうち稼動状態にあるいずれか1つの計算機で動作する特定プロセスにより実行される。

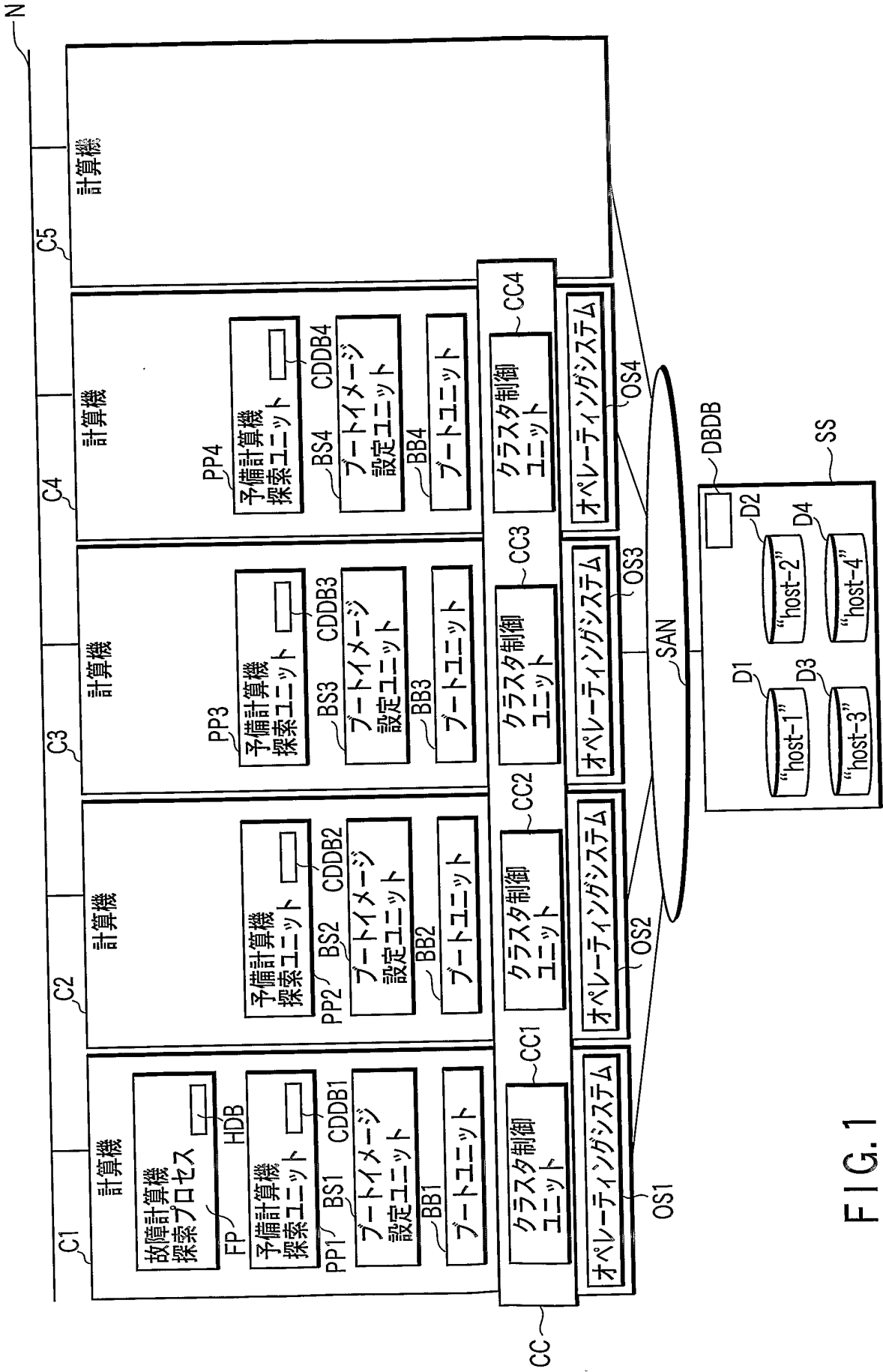
13. 請求項12に従う方法において、前記特定プロセスが動作する計算機で障害が発生した場合に、前記特定プロセスを前記予備計算機以外の他の計算機に移動することを更に具備する。

14. 請求項11に従う方法において、前記探索することは、前記複数の計算機のうち稼動状態にある全ての計算機でそれぞれ動作する複数の特定プロセスにより実行される。

15. 請求項14に従う方法において、前記複数の計算機のうちのいずれかの計算機で障害が発生した場合に、当該障

害が発生した計算機で動作している特定プロセスを前記予備計算機以外の他の計算機に移動することを更に具備する。

16. 請求項15に従う方法において、前記探索することは、当該探索することが実行されるべき計算機で実行されていない場合に、当該探索することが実行されるべき計算機は故障していると判定することを含む。



CDDBi
(i=1,2,3,4)

計算機	ステータス	ディスク
C1	S	D1
C2	S	D2
C3	S	D3
C4	S	D4
C5	P	—

FIG. 2

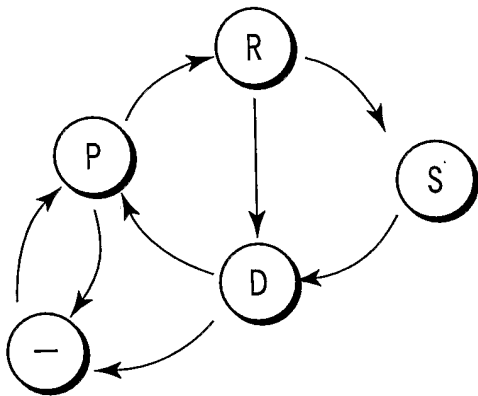


FIG. 3

DBDB

ディスク	ブート計算機
D1	C1
D2	C2
D3	C3
D4	C4

FIG. 4

HDB

ホスト	カウンタ
host-1	721
host-2	1044
host-3	1045
host-4	1046

FIG. 5

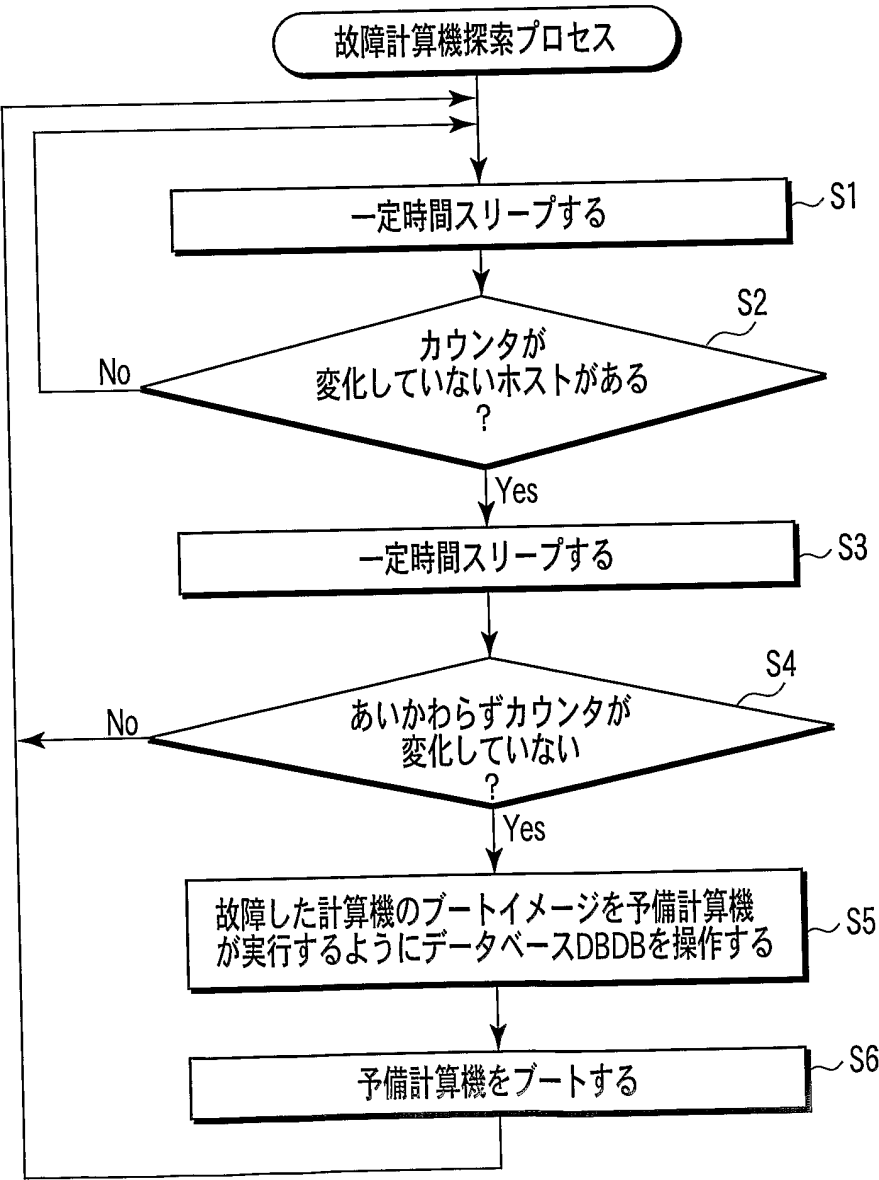


FIG. 6

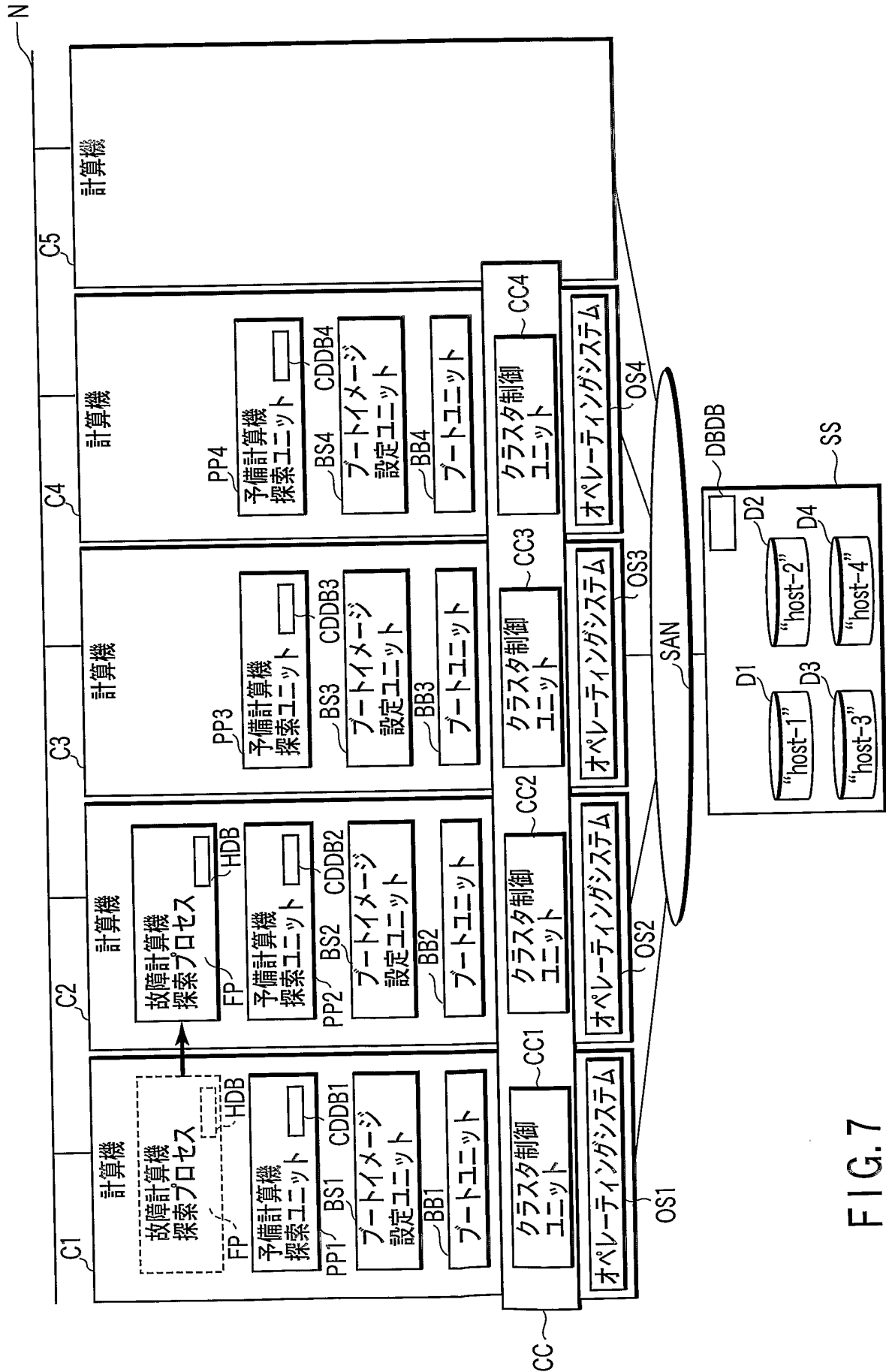


FIG. 7

5/12

DBDB

ディスク	ブート計算機
D1	C5
D2	C2
D3	C3
D4	C4

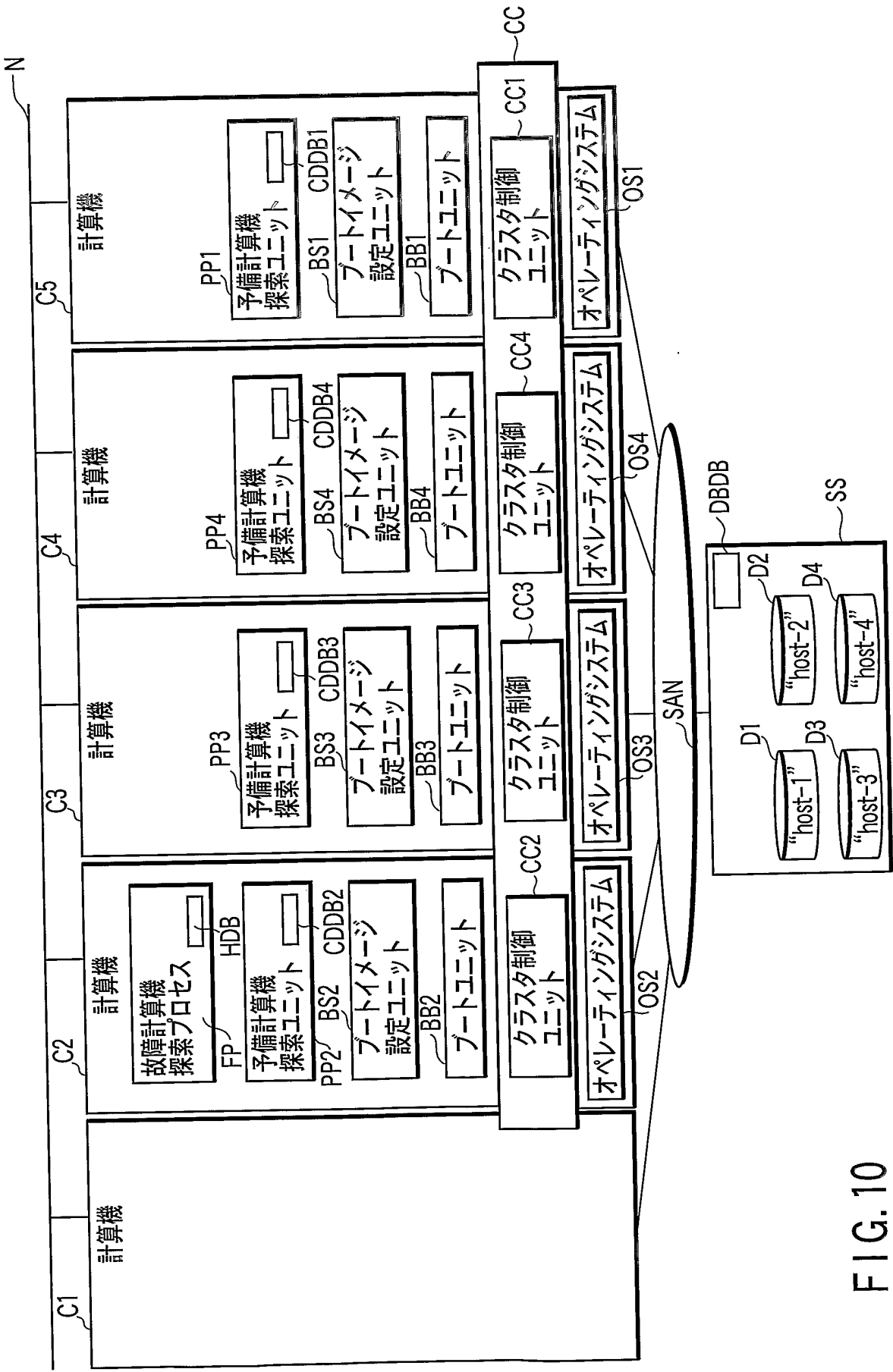
FIG. 8

CDDBi

(i=1,2,3,4)

計算機	ステータス	ディスク
C1	D	—
C2	S	D2
C3	S	D3
C4	S	D4
C5	R	D1

FIG. 9



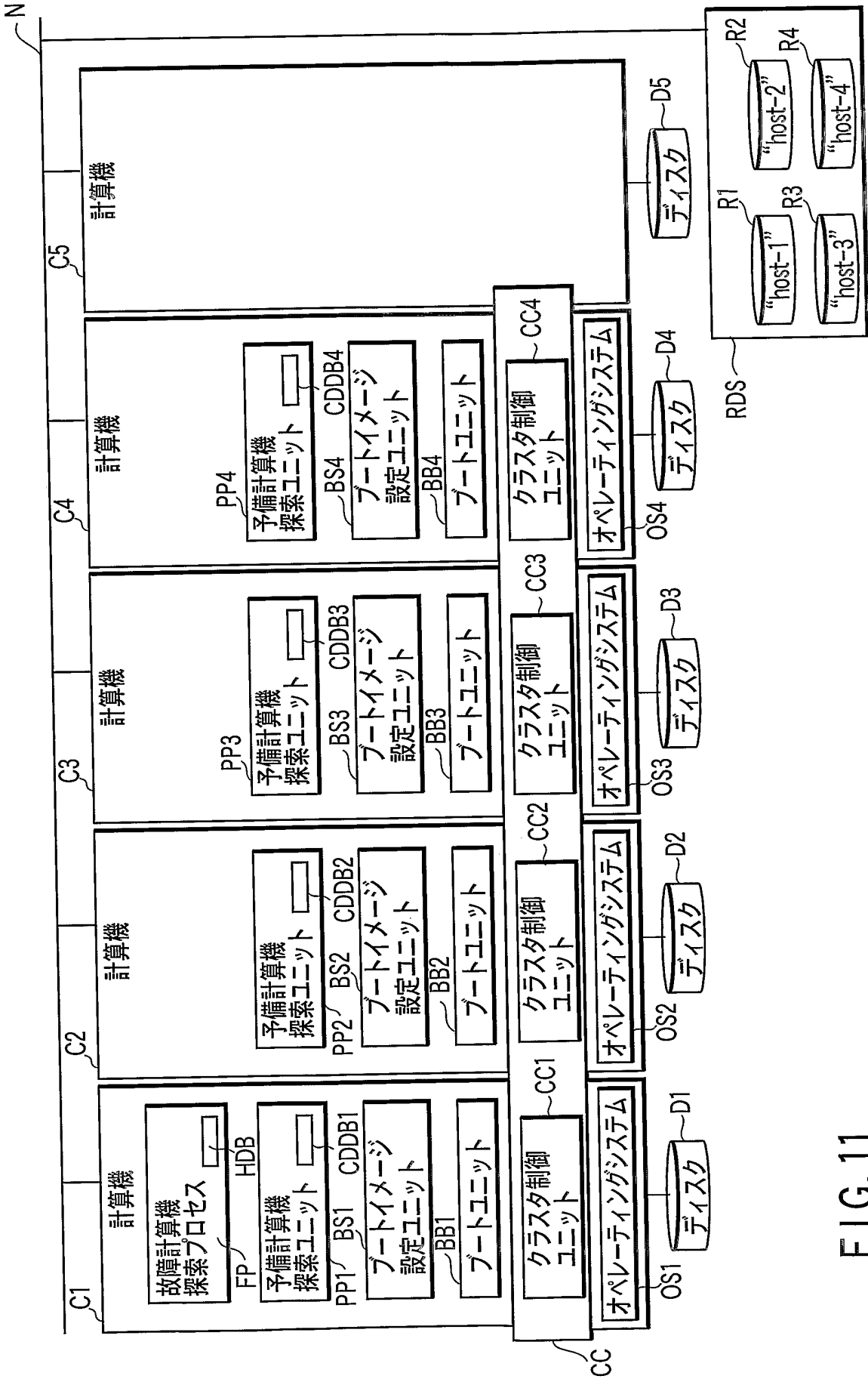


FIG. 11

8/12

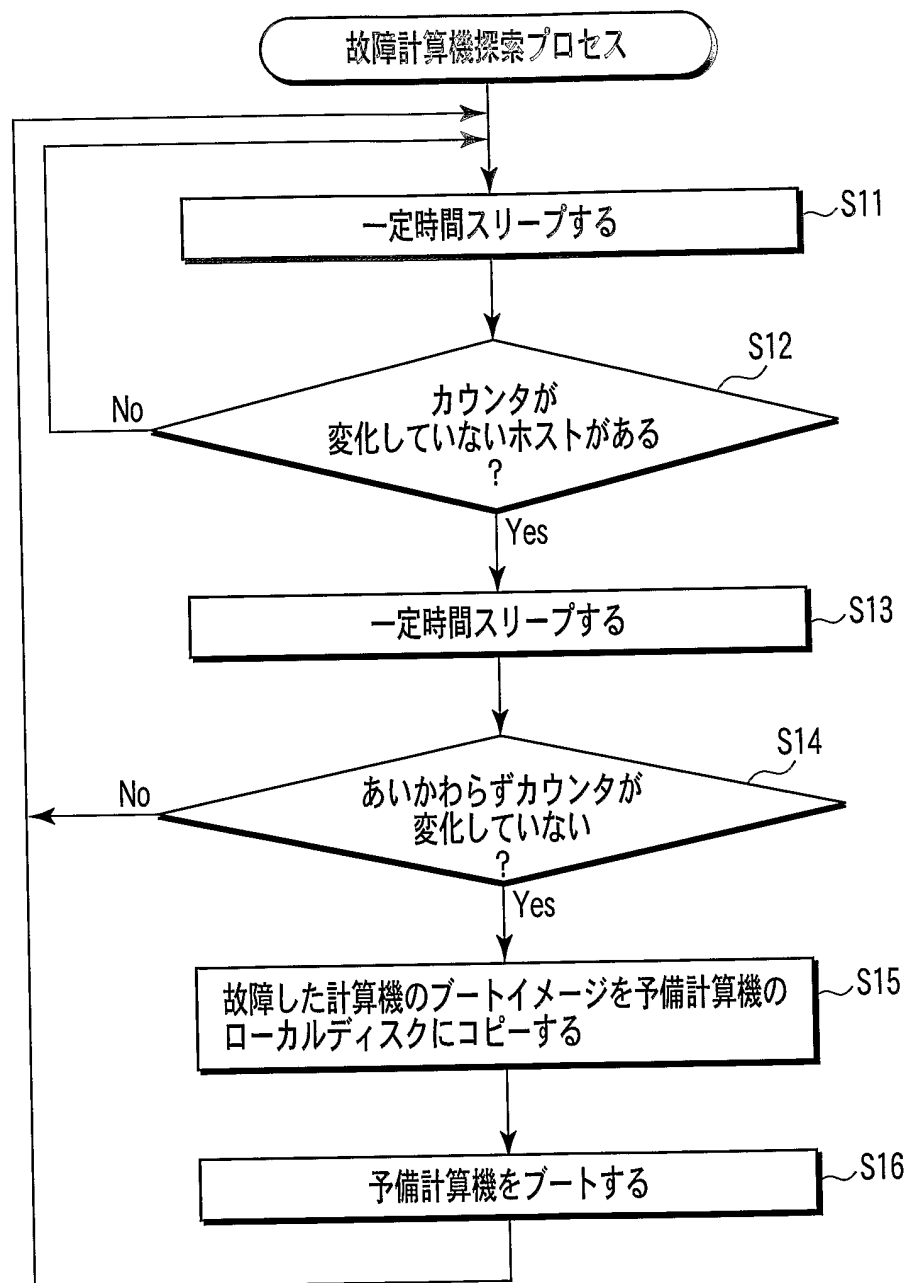


FIG. 12

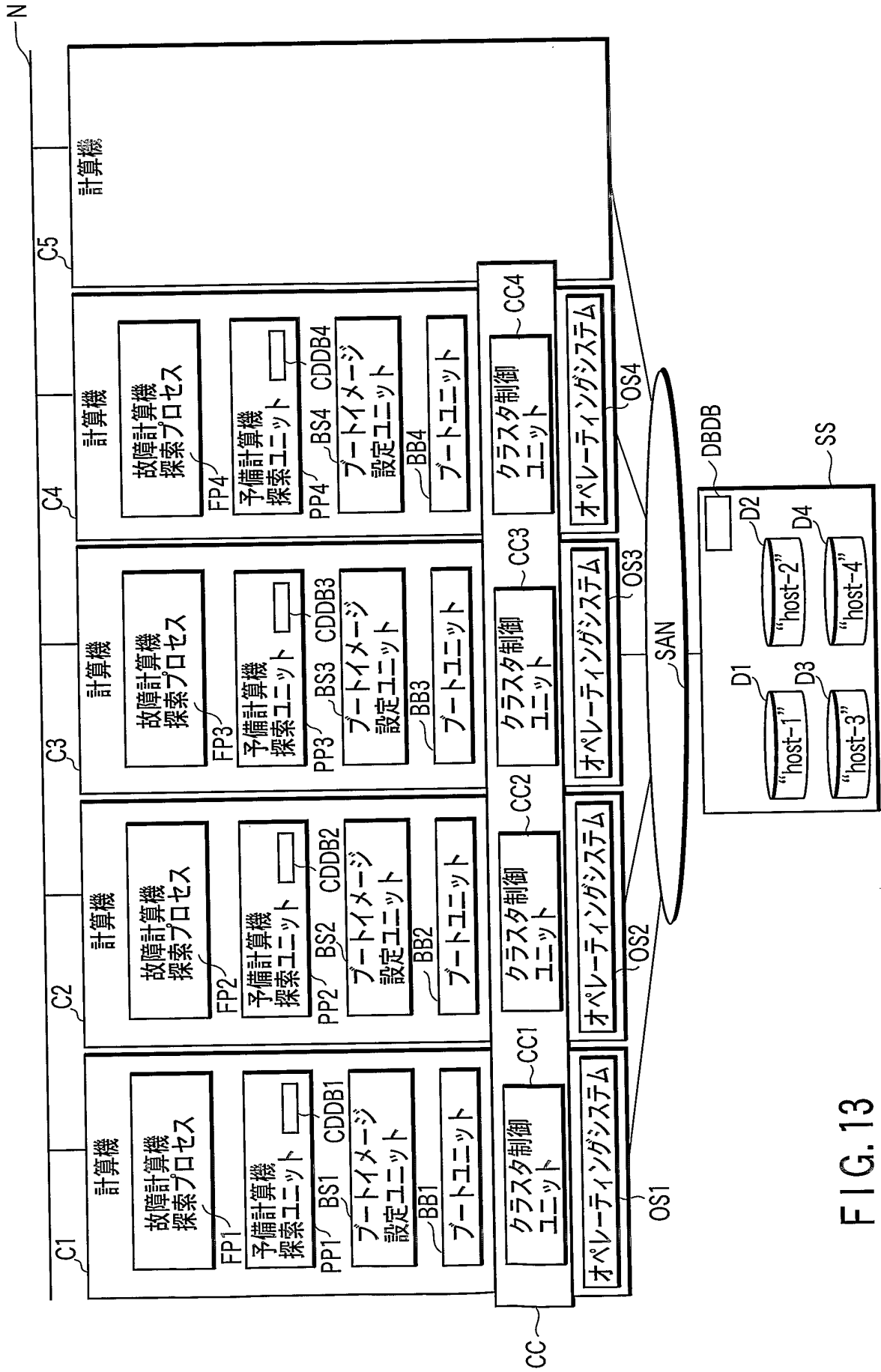


FIG. 13

10/12

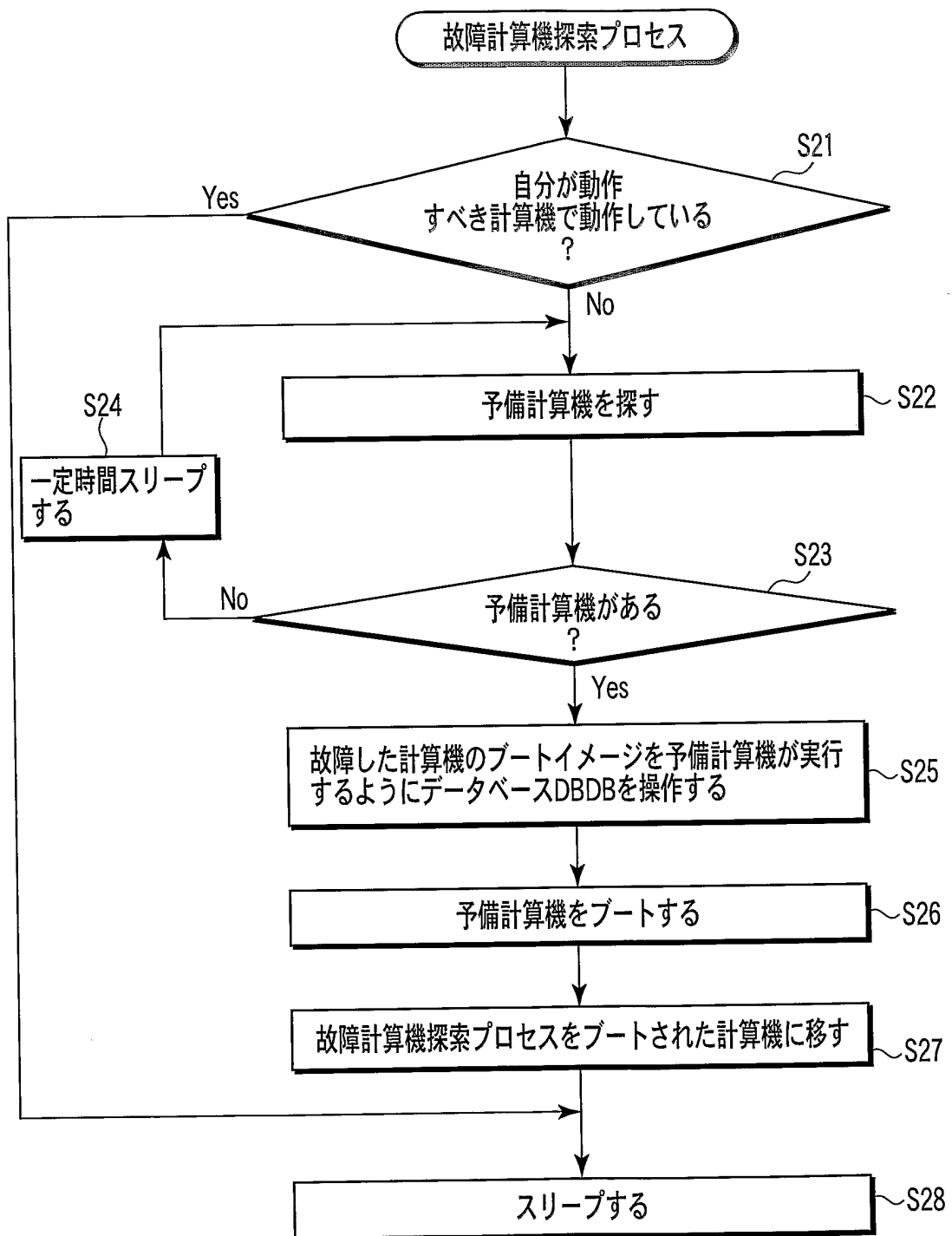


FIG. 14

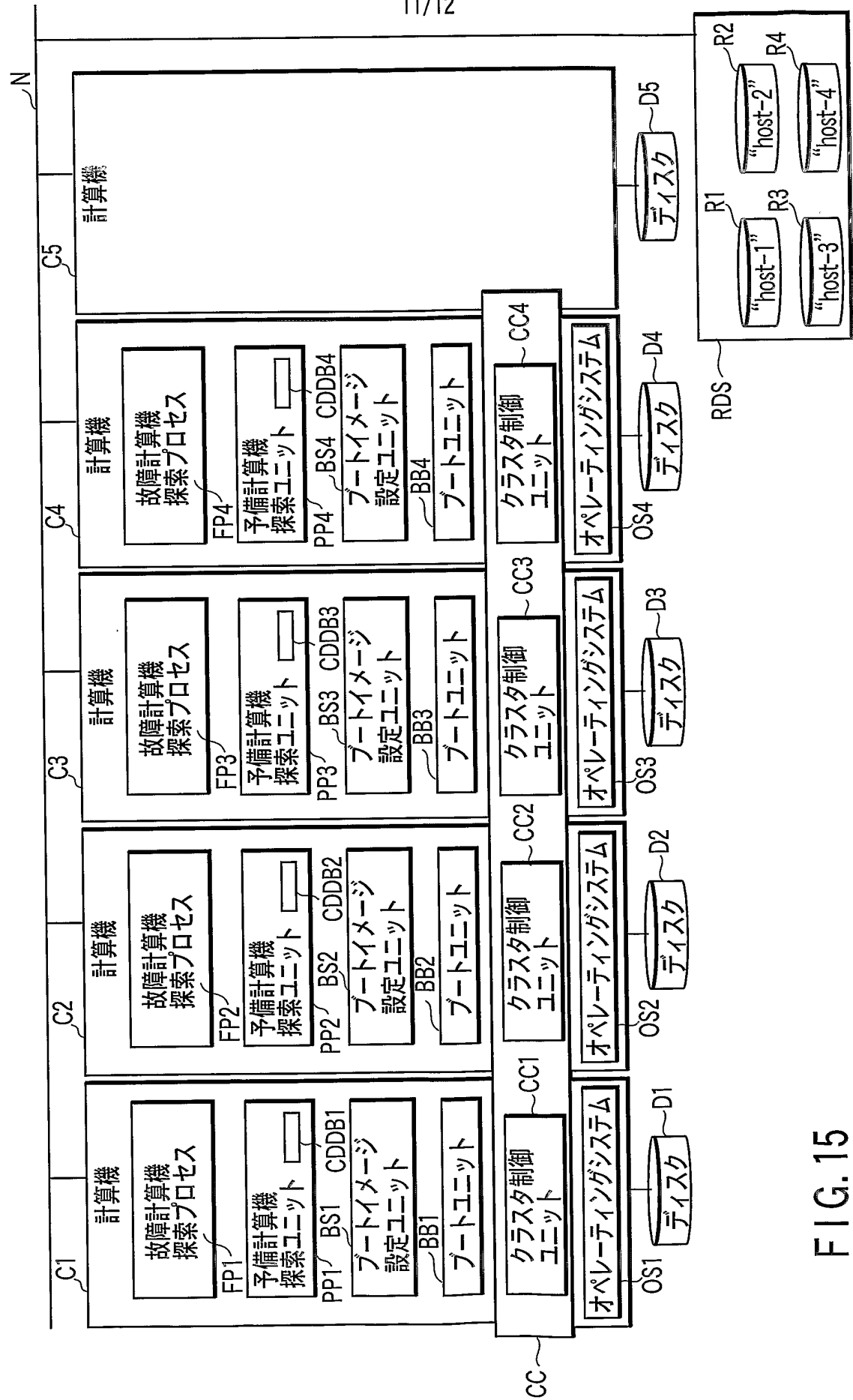


FIG. 15

12/12

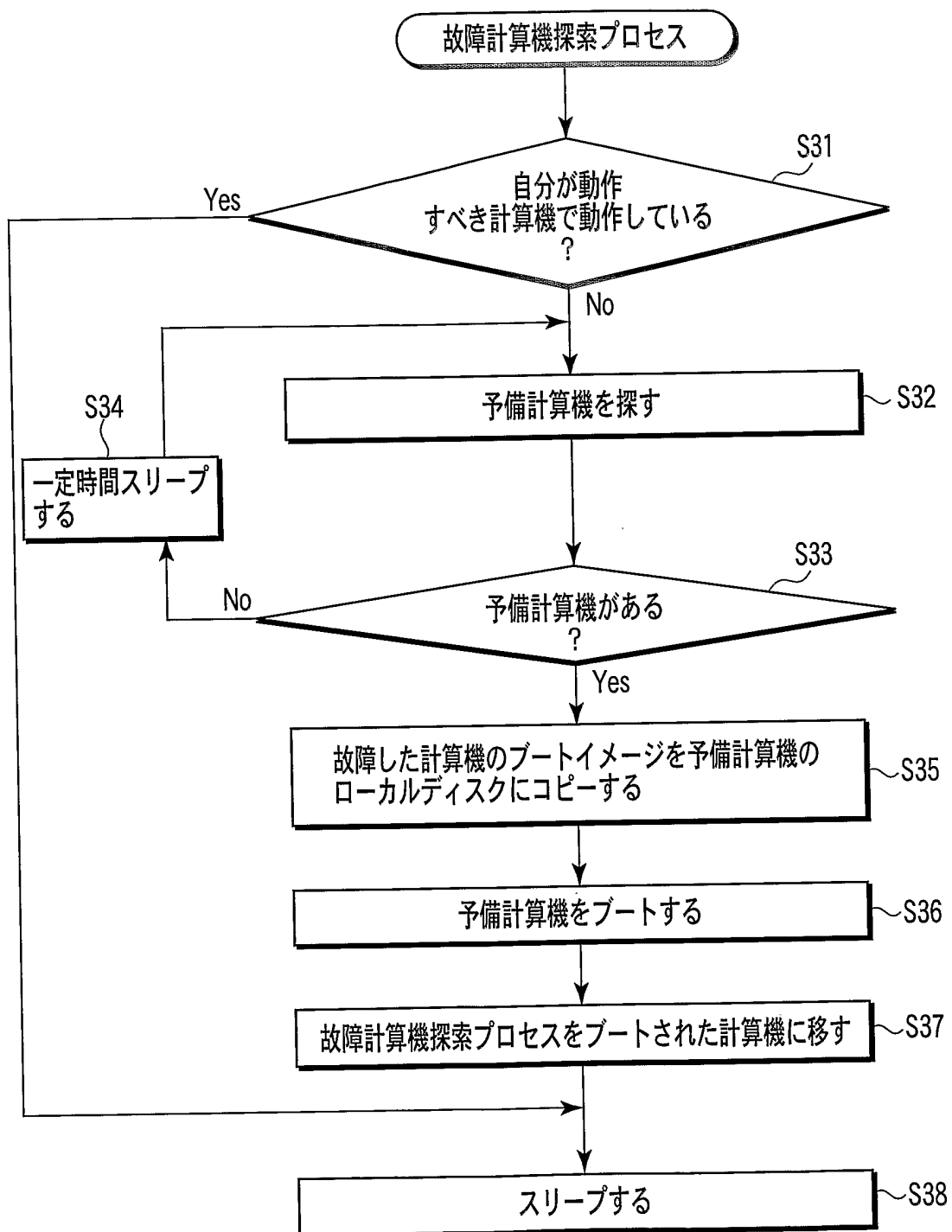


FIG. 16

INTERNATIONAL SEARCH REPORT

International application No.

PCT/JP2004/006500

A. CLASSIFICATION OF SUBJECT MATTER

Int.Cl⁷ G06F11/20

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)

Int.Cl⁷ G06F11/20, G06F15/177

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	1996-2004
Kokai Jitsuyo Shinan Koho	1971-2004	Toroku Jitsuyo Shinan Koho	1994-2004

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 Y A	JP 2001-256071 A (Fuji Electric Co., Ltd.), 21 September, 2001 (21.09.01), Par. No. [0012]; Fig. 1 (Family: none)	1, 11 2, 12, 14 3-10, 13, 15, 16
Y	JP 6-231101 A (Director-General of National Aerospace Laboratory of Science and Technology Agency), 19 August, 1994 (19.08.94), Par. Nos. [0047] to [0051] (Family: none)	2
Y	JP 10-105423 A (NEC Corp.), 24 April, 1998 (24.04.98), Full text; all drawings (Family: none)	12, 14

☐ 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
10 August, 2004 (10.08.04)

Date of mailing of the international search report
24 August, 2004 (24.08.04)

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. ⁷ G 06 F 11/20

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. ⁷ G 06 F 11/20、G 06 F 15/177

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 1971-2004年

日本国実用新案登録公報 1996-2004年

日本国登録実用新案公報 1994-2004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	J P 2001-256071 A (富士電機株式会社)	1, 11
Y	2001. 09. 21, 段落【0012】, 第1図 (ファミリーなし)	2, 12, 14
A		3-10, 13, 15, 16

☒ 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

10. 08. 2004

国際調査報告の発送日

24. 8. 2004

国際調査機関の名称及びあて先

日本国特許庁 (ISA/J P)

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)

久保 正典

5 B

9642

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
Y	JP 6-231101 A (科学技術庁航空宇宙技術研究所長) 1994. 08. 19, 段落【0047】～【0051】 (ファミリーなし)	2
Y	JP 10-105423 A (日本電気株式会社) 1998. 04. 24, 全文, 全図 (ファミリーなし)	12, 14