(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第6089884号
(P6089884)
(45)発行日　平成29年3月8日(2017.3.8)　(24)登録日　平成29年2月17日(2017.2.17)

(51)Int.Cl.　　F I
G06F 11/20 (2006.01)　　G06F 11/20 692
G06F 11/22 (2006.01)　　G06F 11/22 673J

請求項の数 9　(全 52 頁)

(21)出願番号　特願2013-71904(P2013-71904)
(22)出願日　平成25年3月29日(2013.3.29)
(65)公開番号　特開2014-197266(P2014-197266A)
(43)公開日　平成26年10月16日(2014.10.16)
審査請求日　平成27年12月4日(2015.12.4)

(73)特許権者　000005223
富士通株式会社
神奈川県川崎市中原区上小田中４丁目１番１号
(74)代理人　100092978
弁理士　真田　有
(74)代理人　100112678
弁理士　山本　雅久
(72)発明者　田村　雅寿
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

審査官　三坂　敏夫

最終頁に続く

(54)【発明の名称】情報処理システム，情報処理装置，情報処理装置の制御プログラム，及び情報処理システムの制御方法

(57)【特許請求の範囲】
【請求項１】
相互に接続される複数の情報処理装置を有し、前記複数の情報処理装置間で通信を行なう情報処理システムにおいて、
前記複数の情報処理装置の各々が、
前記複数の情報処理装置のうちの自情報処理装置以外の他の情報処理装置の各々から、前記他の情報処理装置により判定された前記複数の情報処理装置の各々の状態に関する状態情報を受信する受信処理部と、
前記受信処理部が前記他の情報処理装置の各々から受信した前記状態情報に基づいて、前記複数の情報処理装置の各々の状態を判定する判定部と、
前記判定部が判定した前記複数の情報処理装置の各々の状態に関する送信用状態情報を、前記他の情報処理装置の各々へ送信する送信処理部と、を有し、
前記判定部は、
前記受信処理部が受信した前記状態情報が示す前記複数の情報処理装置の各々の状態と、前記他の情報処理装置の各々からの前記状態情報の受信状況とに基づいて、前記複数の情報処理装置の各々の状態を判定し、
前記送信処理部は、
第１所定時間ごとに、前記送信用状態情報を、前記他の情報処理装置の各々へ送信し、
前記判定部は、
前記第１所定時間以上の時間である第２所定時間内に前記状態情報を受信しなかった他

の情報処理装置の状態を、停止の可能性を示す第１状態と判定し、
前記受信処理部が受信した前記状態情報に基づいて、第１所定数以上の前記複数の情報処理装置で前記第１状態であると判定された情報処理装置の状態、又は、前記他の情報処理装置の少なくとも１つから停止を示す第２状態であると判定された情報処理装置の状態を、前記第２状態と判定することを特徴とする、情報処理システム。
【請求項２】
前記判定部は、
前記受信処理部が受信した前記状態情報が示す前記複数の情報処理装置の各々の状態と、前記判定部が含まれる自情報処理装置の状態に関する状態情報に関する自己状態情報とに基づいて、前記複数の情報処理装置の各々の状態を判定することを特徴とする、請求項１記載の情報処理システム。
【請求項３】
前記判定部は、
前記受信処理部が受信した前記状態情報に基づいて、前記第１所定数以上の数である第２所定数以上の前記複数の情報処理装置で前記第２状態であると判定された情報処理装置を、リカバリ処理中を示す第３状態と判定し、
前記複数の情報処理装置のうちの１以上の情報処理装置はさらに、
前記判定部が前記第３状態と判定した情報処理装置に対して、リカバリ処理を実行するリカバリ処理部を有することを特徴とする、請求項１又は請求項２記載の情報処理システム。
【請求項４】
前記複数の情報処理装置が複数のグループに分割され、
前記複数のグループの各々における代表情報処理装置はさらに、
前記複数のグループのうちの自グループ以外の他のグループの各々における他の代表情報処理装置から、前記他の代表情報処理装置により判定された前記複数のグループの代表情報処理装置の各々の状態に関する代表状態情報を受信するグループ間受信処理部と、
前記グループ間受信処理部が前記他の代表情報処理装置の各々から受信した前記代表状態情報に基づいて、前記複数の代表情報処理装置の各々の状態を判定するグループ間判定部と、
前記グループ間判定部が判定した前記複数の代表情報処理装置の各々の状態に関する送信用代表状態情報を、前記他の代表情報処理装置の各々へ送信するグループ間送信処理部と、を有し、
前記複数の情報処理装置の各々において、
前記送信処理部は、
前記送信用状態情報を、前記自グループにおける他の情報処理装置の各々へ送信し、
前記判定部は、
前記受信処理部が前記自グループにおける他の情報処理装置の各々から受信した前記状態情報に基づいて、前記自グループにおける情報処理装置の各々の状態を判定することを特徴とする、請求項１～３のいずれか１項記載の情報処理システム。
【請求項５】
前記複数のグループの各々における代表情報処理装置はさらに、
前記自グループにおける情報処理装置の数が第４所定値を超えた場合、前記自グループから、複数の情報処理装置を分割して新たなグループを作成する管理部を有することを特徴とする、請求項４記載の情報処理システム。
【請求項６】
前記管理部は、
前記自グループにおける情報処理装置の数が第５所定値未満の場合、前記自グループと前記他のグループのうちのいずれかのグループとを統合することを特徴とする、請求項５記載の情報処理システム。
【請求項７】

相互に接続される複数の情報処理装置の各々において、
前記複数の情報処理装置のうちの自情報処理装置以外の他の情報処理装置の各々から、前記他の情報処理装置により判定された前記複数の情報処理装置の各々の状態に関する状態情報を受信する受信処理部と、
前記受信処理部が前記他の情報処理装置の各々から受信した前記状態情報に基づいて、前記複数の情報処理装置の各々の状態を判定する判定部と、
前記判定部が判定した前記複数の情報処理装置の各々の状態に関する送信用状態情報を、前記他の情報処理装置の各々へ送信する送信処理部と、を有し、
前記判定部は、
前記受信処理部が受信した前記状態情報が示す前記複数の情報処理装置の各々の状態と、前記他の情報処理装置の各々からの前記状態情報の受信状況とに基づいて、前記複数の情報処理装置の各々の状態を判定し、
前記送信処理部は、
第１所定時間ごとに、前記送信用状態情報を、前記他の情報処理装置の各々へ送信し、
前記判定部は、
前記第１所定時間以上の時間である第２所定時間内に前記状態情報を受信しなかった他の情報処理装置の状態を、停止の可能性を示す第１状態と判定し、
前記受信処理部が受信した前記状態情報に基づいて、第１所定数以上の前記複数の情報処理装置で前記第１状態であると判定された情報処理装置の状態、又は、前記他の情報処理装置の少なくとも１つから停止を示す第２状態であると判定された情報処理装置の状態を、前記第２状態と判定することを特徴とする、情報処理装置。
【請求項８】
相互に接続される複数の情報処理装置の各々を制御する情報処理装置に、
前記複数の情報処理装置のうちの自情報処理装置以外の他の情報処理装置の各々から、前記他の情報処理装置により判定された前記複数の情報処理装置の各々の状態に関する状態情報を受信し、
前記他の情報処理装置の各々から受信した前記状態情報に基づいて、前記複数の情報処理装置の各々の状態を判定し、
判定した前記複数の情報処理装置の各々の状態に関する送信用状態情報を、前記他の情報処理装置の各々へ送信する、処理を実行させ、
前記判定において、
受信した前記状態情報が示す前記複数の情報処理装置の各々の状態と、前記他の情報処理装置の各々からの前記状態情報の受信状況とに基づいて、前記複数の情報処理装置の各々の状態を判定し、
前記送信において、
第１所定時間ごとに、前記送信用状態情報を、前記他の情報処理装置の各々へ送信し、
前記判定において、
前記第１所定時間以上の時間である第２所定時間内に前記状態情報を受信しなかった他の情報処理装置の状態を、停止の可能性を示す第１状態と判定し、
受信した前記状態情報に基づいて、第１所定数以上の前記複数の情報処理装置で前記第１状態であると判定された情報処理装置の状態、又は、前記他の情報処理装置の少なくとも１つから停止を示す第２状態であると判定された情報処理装置の状態を、前記第２状態と判定することを特徴とする、情報処理装置の制御プログラム。
【請求項９】
相互に接続される複数の情報処理装置を有し、前記複数の情報処理装置間で通信を行なう情報処理システムの制御方法において、
前記複数の情報処理装置の各々が、
前記複数の情報処理装置のうちの自情報処理装置以外の他の情報処理装置の各々から、前記他の情報処理装置により判定された前記複数の情報処理装置の各々の状態に関する状態情報を受信し、

前記他の情報処理装置の各々から受信した前記状態情報に基づいて、前記複数の情報処理装置の各々の状態を判定し、
判定した前記複数の情報処理装置の各々の状態に関する送信用状態情報を、前記他の情報処理装置の各々へ送信し、
前記判定において、
受信した前記状態情報が示す前記複数の情報処理装置の各々の状態と、前記他の情報処理装置の各々からの前記状態情報の受信状況とに基づいて、前記複数の情報処理装置の各々の状態を判定し、
前記送信において、
第１所定時間ごとに、前記送信用状態情報を、前記他の情報処理装置の各々へ送信し、
前記判定において、
前記第１所定時間以上の時間である第２所定時間内に前記状態情報を受信しなかった他の情報処理装置の状態を、停止の可能性を示す第１状態と判定し、
受信した前記状態情報に基づいて、第１所定数以上の前記複数の情報処理装置で前記第１状態であると判定された情報処理装置の状態、又は、前記他の情報処理装置の少なくとも１つから停止を示す第２状態であると判定された情報処理装置の状態を、前記第２状態と判定することを特徴とする、情報処理システムの制御方法。

【発明の詳細な説明】

【技術分野】

【０００１】

本件は、情報処理システム，情報処理装置，情報処理装置の制御プログラム，及び情報処理システムの制御方法に関する。

【背景技術】

【０００２】

従来、複数のノード（ストレージ装置，情報処理装置）をそなえ、データを複数のノードに分散させて保持する分散ストレージシステム（ストレージシステム，情報処理システム）が知られている。

分散ストレージシステムにおいて、例えば、複数のノードのいずれかのノードに故障が発生した場合、分散ストレージシステムを使用するクライアントは、故障したノードへアクセスをすることができなくなる。

【０００３】

また、故障したノードが他のノードと冗長化されていた場合には、クライアントは、故障したノードの代わりに冗長化されたノードへアクセスをすることができる。しかし、冗長化されたノードをそなえる分散ストレージシステムは、ノードに故障が発生する前のデータの多重化状態を回復するリカバリ処理及び故障ノードの交換が行なわれるまで、冗長度が低下した信頼性の低い状態になる。

【０００４】

従って、分散ストレージシステムは、複数のノードの状態を監視し、ノードの故障を速やかに検出することが好ましい。

しかし、分散ストレージシステムでは、ノード又はノード間のリンクの故障により、複数のノードが分断され、分断された一方のノードと他方のノードとが、ノードの故障について異なる判断をすることがある。この状態をスプリットブレイン（Split Brain）状態という。スプリットブレイン状態の一例としては、一方のノードと他方のノードとの間でリンクの故障が発生することにより互いにアクセスができなくなる状態が発生するが、双方のノードは、互いにアクセスができなくなった相手のノードが故障したと判断する場合が挙げられる。

【０００５】

例えば、一方及び他方のノードが、同一データの冗長データを互いに保持する場合に、スプリットブレイン状態に陥ると、双方のノードは、それぞれの保持する冗長データを個別に更新したり、他のノードへリカバリをし、冗長データの一貫性を崩す可能性がある。

分散ストレージシステムにおいて、スプリットブレイン状態に陥ることを防止する手法としては、以下に例示する手法が知られている。
（１）複数のノードの各々が、複数のノードのうちの所定のノード（コントロールノード）へ自ノードの構成情報及び生存報告を通知する。コントロールノードは、複数のノードの各々から得た情報を集約して複数のノードを監視し、監視結果から故障したノードを検出すると、リカバリを行ない、管理者等へノードの故障を通知する。
（２）複数のノードの各々が、互いに生存報告のやり取りを行ない（情報交換フェーズ）、どのノードが監視及び故障ノードの検出を行なうかを、他のノードとの間で合意を取ることで選定する。合意を得たノード（決定ノード）は、複数のノードの各々の状態を監視し、監視結果から故障したノードを検出すると、リカバリを行ない、管理者等へノードの故障を通知する。
（３）複数のノードの各々が、所定のノードへ生存報告を行なう。故障ノードは所定のノードにより即座に検出はされず、管理者等が、所定のノードを参照し手動で故障ノードの検出及びリカバリ等の対応を行なう。

【０００６】

上記（１）の手法では、コントロールノードが故障ノードの検出を行ない、上記（２）の手法では、合意を得た決定ノードが故障ノードの検出を行なう。また、上記（３）の手法では、管理者等が故障ノードの検出を行なう。従って、上記（１）～（３）の手法によれば、複数のノードで判断が行なわれるのではなく、特定のノード又は管理者が判断を行なうため、スプリットブレイン状態に陥ることを防止できる。

【０００７】

なお、関連する技術として、分散ストレージシステムにおいて、データの消失を防ぐため、コンピュータが、複数のストレージノードから収集した属性に基づき、ストレージノードを２以上のグループに分ける技術が知られている（例えば、特許文献１参照）。この技術では、コンピュータは、作成した各グループ内において、データを分散した分散データと、当該データと同一内容の冗長データを分散した冗長分散データとが存在しないように、分散データ及び冗長分散データを各グループに割り当てる。

【０００８】

また、関連する他の技術として、管理サーバが、データを保持する全てのストレージで同一のデータ・プールを構成し、異なるデータをできるだけプール内の複数の異なるストレージに分散して保持させる技術が知られている（例えば、特許文献２参照）。

さらに、関連する他の技術として、ネットワーク監視装置が、複数ノードをグループ単位に分割し、分割したグループの１つのノードから論理回線状態を取得して、論理回線の監視を行なう技術が知られている（例えば、特許文献３参照）。

【０００９】

また、関連する他の技術として、ネットワーク管理システムが、ノードの自装置の情報、ホップ数等の情報に基づいて形成されたグループ毎に、グループ内のノードを監視するグループ管理装置をそなえる技術が知られている（例えば、特許文献４参照）。

【先行技術文献】

【特許文献】

【００１０】

【特許文献１】国際公開第ＷＯ２００８／１１４４４１号パンフレット
【特許文献２】特表２０１１－５０５６１７号公報
【特許文献３】特開２０１０－２５８６１４号公報
【特許文献４】特開２０１１－０５５２３１号公報

【発明の概要】

【発明が解決しようとする課題】

【００１１】

上記（１）の手法では、複数のノードの各々の情報が１点（コントロールノード）に集約されるため、コントロールノードがＳＰＯＦ（Single Point Of Failure；単一障害点

）となる。従って、コントロールノードが故障した場合、クライアントは、コントロールノードが復旧するまで分散ストレージシステムの利用が制限されるという課題がある。

上記（２）の手法では、複数のノード間で合意を形成するために複雑な手順が行なわれるため、上記（１）の手法と比較して、合意を形成するまでの時間が余計にかかる場合がある。また上記（３）の手法では、管理者等による人為的な判断が行なわれるため、ノードの故障が発生してからノードの故障が検出され、リカバリ処理が行なわれるまでに、上記（１）及び（２）の手法と比較して時間がかかる場合がある。つまり、上記（２）及び（３）の手法では、障害が発生したノードに対するリカバリ処理等の開始が遅くなり、クライアントが分散ストレージシステムの利用を制限される時間が長くなるという課題がある。

【００１２】

なお、上述した関連する技術は、いずれも、上記（１）の手法のように管理装置が複数のノードを管理するものであり、上述した課題については考慮されていない。

このように、複数のストレージ装置をそなえるストレージシステムにおいて、複数のストレージ装置の各々の状態を判断する上述した技術では、ストレージシステムの可用性が低下するという課題がある。

【００１３】

ここまで、情報処理システムがストレージシステム（分散ストレージシステム）であるものとして説明したが、これに限定されるものではない。上述した課題は、情報処理システムがそなえる複数の情報処理装置の各々が、分散データではなく他の情報処理装置とは異なるデータを保持する場合であっても、同様に生じ得る。

１つの側面では、本発明は、複数の情報処理装置をそなえる情報処理システムにおいて、可用性の低下を抑止することを目的とする。

【００１４】

なお、前記目的に限らず、後述する発明を実施するため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１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】

【００１５】

本件の情報処理システムは、相互に接続される複数の情報処理装置を有し、前記複数の情報処理装置間で通信を行なう情報処理システムにおいて、前記複数の情報処理装置の各々が、前記複数の情報処理装置のうちの自情報処理装置以外の他の情報処理装置の各々から、前記他の情報処理装置により判定された前記複数の情報処理装置の各々の状態に関する状態情報を受信する受信処理部と、前記受信処理部が前記他の情報処理装置の各々から受信した前記状態情報に基づいて、前記複数の情報処理装置の各々の状態を判定する判定部と、前記判定部が判定した前記複数の情報処理装置の各々の状態に関する送信用状態情報を、前記他の情報処理装置の各々へ送信する送信処理部と、をそなえてよい。また、前記判定部は、前記受信処理部が受信した前記状態情報が示す前記複数の情報処理装置の各々の状態と、前記他の情報処理装置の各々からの前記状態情報の受信状況とに基づいて、前記複数の情報処理装置の各々の状態を判定し、前記送信処理部は、第１所定時間ごとに、前記送信用状態情報を、前記他の情報処理装置の各々へ送信してよい。さらに、前記判定部は、前記第１所定時間以上の時間である第２所定時間内に前記状態情報を受信しなかった他の情報処理装置の状態を、停止の可能性を示す第１状態と判定し、前記受信処理部が受信した前記状態情報に基づいて、第１所定数以上の前記複数の情報処理装置で前記第１状態であると判定された情報処理装置の状態、又は、前記他の情報処理装置の少なくとも１つから停止を示す第２状態であると判定された情報処理装置の状態を、前記第２状態と判定してよい。

【発明の効果】

【００１６】

第１実施形態及び第２実施形態によれば、複数の情報処理装置をそなえる情報処理シス

テムにおいて、可用性の低下を抑止することができる。
【図面の簡単な説明】
【００１７】
【図１】第１実施形態の一例としてのストレージシステムの構成例を示す図である。
【図２】第１実施形態の一例としてのノードのハードウェア構成例を示す図である。
【図３】第１実施形態の一例としてのノードの機能構成例を示す図である。
【図４】第１実施形態の一例としてのノードが送受信するノード状態情報を例示する図である。
【図５】第１実施形態の一例としてのノードが管理するノード状態管理情報を例示する図である。
【図６】第１実施形態の一例としての新規ノードが送信する情報を例示する図である。
【図７】第１実施形態の一例としての新規ノードが受信する情報を例示する図である。
【図８】第１実施形態の一例としてのノードが他ノードの状態を判定するときの状態遷移の一例を示す図である。
【図９】第１実施形態の一例としての複数のノードによるノード状態情報の送受信処理の一例を説明する図である。
【図１０】第１実施形態の一例としてのノードが自ノードの状態を判定するときの状態遷移の一例を示す図である。
【図１１】第１実施形態の一例としての新規ノードによる起動後の動作例を説明するフローチャートである。
【図１２】第１実施形態の一例としてのノードによる他ノードの状態を判定する動作例を説明するフローチャートである。
【図１３】第１実施形態の一例としてのノードによる自ノードの状態を判定する動作例を説明するフローチャートである。
【図１４】第２実施形態の一例としてのノードの機能構成例を示す図である。
【図１５】第２実施形態の一例としてのノードが管理するパーティ管理情報を例示する図である。
【図１６】第２実施形態の一例としての複数のノードによる代表ノード状態情報及びノード状態情報の送受信処理の一例を説明する図である。
【図１７】第２実施形態の一例としてのノードが送受信するノード状態情報を例示する図である。
【図１８】第２実施形態の一例としてのノードが送受信する代表ノード状態情報を例示する図である。
【図１９】第２実施形態の一例としてのノードが管理するノード状態管理情報を例示する図である。
【図２０】第２実施形態の一例としてのストレージシステムにノードが追加される例を示す図である。
【図２１】図２０に示すストレージシステムにおけるパーティの分割処理の一例を説明する図である。
【図２２】図２１に示すストレージシステムにおけるノードの削除処理及びパーティの統合処理の一例を説明する図である。
【図２３】第２実施形態の一例としてのストレージシステムにおけるパーティの分割処理の具体例を説明する図である。
【図２４】第２実施形態の一例としての代表ノードによる他の代表ノードの状態を判定する動作例を説明するフローチャートである。
【図２５】第２実施形態の一例としてのノードによるパーティ内の他ノードが停止した場合の動作例を説明するフローチャートである。
【図２６】第２実施形態の一例としてのノードによるパーティの分割処理及び統合処理の動作例を説明するフローチャートである。
【発明を実施するための形態】

【００１８】
以下、図面を参照して実施の形態を説明する。
〔１〕第１実施形態
〔１－１〕ストレージシステムの構成
以下、図１及び図２を参照して、第１実施形態の一例としてのストレージシステム１の構成について説明する。
【００１９】
図１は、第１実施形態の一例としてのストレージシステム１の構成例を示す図であり、図２は、図１に示すノード１０－１～１０－５のハードウェア構成例を示す図である。
図１に示すように、第１実施形態に係るストレージシステム（情報処理システム）１は、複数（例えば５つ）のノード１０－１～１０－５及び複数（例えば３つ）のスイッチ２０－１～２０－３をそなえる。
【００２０】
なお、以下、ノード１０－１～１０－５を区別しない場合には、単にノード１０といい、スイッチ２０－１～２０－３を区別しない場合には、単にスイッチ２０という。
ストレージシステム１は、複数のノード１０及びスイッチ２０により、ＳＡＮ（Storage Area Network）を形成し、相互に接続される複数のノード１０間で通信を行なう。また、ストレージシステム１は、図示しないクライアントに接続され、クライアントに対してノード１０が有する記憶領域（リソース）を提供する。
【００２１】
ストレージシステム１としては、分散ストレージシステム又はクラスタファイルシステム等の、データを複数のノード１０に分散させて保持する種々のストレージシステムが例として挙げられる。例えば、ストレージシステム１は、Ｗｅｂサーバのデータベースやクラウドストレージ等に用いられることがある。
なお、複数のノード１０の各々が、分散データではなく他のノード１０とは異なるデータを保持してもよい。
【００２２】
ノード（ストレージ装置，ノード装置，情報処理装置）１０は、クライアント（端末装置、図示省略）からの各種要求に応じて、ノード１０がそなえる記憶部１０ｃ（図２参照）に対する各種処理を行なう。なお、ノード１０としてはＰＣ（Personal Computer）サーバ等の情報処理装置が挙げられる。
ノード１０は、図２に示すように、ＣＰＵ（Central Processing Unit）１０ａ、メモリ１０ｂ、記憶部１０ｃ、ネットワークインタフェース１０ｄ、入出力部１０ｅ、記録媒体１０ｆ、及び読取部１０ｇをそなえる。なお、ノード１０－１～１０－５は、互いに同様のハードウェアをそなえることができるため、以下、任意のノード１０がそなえるハードウェアについて説明する。
【００２３】
ＣＰＵ１０ａは、メモリ１０ｂ、記憶部１０ｃ、ネットワークインタフェース１０ｄ、入出力部１０ｅ、記録媒体１０ｆ、及び読取部１０ｇと接続され、種々の制御や演算を行なう演算処理装置（プロセッサ）である。ＣＰＵ１０ａは、メモリ１０ｂ、記憶部１０ｃ、記録媒体１０ｆ、読取部１０ｇに接続又は挿入された記録媒体１０ｈ、又は図示しないＲＯＭ（Read Only Memory）等に格納されたプログラムを実行することにより、ノード１０における種々の機能を実現する。なお、ＣＰＵ１０ａに限らず、プロセッサとしては、ＭＰＵ（Micro Processing Unit）等の電子回路が用いられてもよい。
【００２４】
メモリ１０ｂは、種々のデータやプログラムを格納する記憶装置である。ＣＰＵ１０ａは、プログラムを実行する際に、メモリ１０ｂにデータやプログラムを格納し展開する。なお、メモリ１０ｂとしては、例えばＲＡＭ（Random Access Memory）等の揮発性メモリが挙げられる。
記憶部１０ｃは、例えばＨＤＤ（Hard Disk Drive）等の磁気ディスク装置、ＳＳＤ（S

olid State Drive）等の半導体ドライブ装置、又はフラッシュメモリ等の不揮発性メモリ等の、種々のデータやプログラム等を格納する１以上のハードウェアである。記憶部１０ｃが有する記憶領域は、クライアントにより用いられる。

【００２５】

ネットワークインタフェース１０ｄは、スイッチ２０を介したノード１０又はクライアントとの間の接続及び通信の制御を行なうコントローラである。ネットワークインタフェース１０ｄとしては、例えば、ＬＡＮ（Local Area Network）、ファイバチャネル（Fibre Channel；ＦＣ）、又はインフィニバンド（InfiniBand）（登録商標）等に準拠したインタフェースカードが挙げられる。なお、ネットワークインタフェース１０ｄは、ＬＡＮに準拠する場合、ｉＳＣＳＩ（Internet Small Computer System Interface）に対応することが好ましい。

【００２６】

入出力部１０ｅは、例えばマウスやキーボード等の入力装置及びディスプレイやプリンタ等の出力装置の少なくとも一方を含んでよい。例えば入出力部１０ｅは、ストレージシステム１の管理者等により、後述するノード情報の設定又は参照、ログの参照、その他種々の作業に用いられる。

記録媒体１０ｆは、フラッシュメモリやＲＯＭ等の記憶装置であり、種々のデータやプログラムを記録する。読取部１０ｇは、光ディスクやＵＳＢ（Universal Serial Bus）メモリ等のコンピュータ読取可能な記録媒体１０ｈに記録されたデータやプログラムを読み出す装置である。

【００２７】

記録媒体１０ｆ及び１０ｈの少なくとも一方には、第１実施形態に係るノード１０（及び後述する第２実施形態に係るノード１０Ａ）の機能を実現する制御プログラムが格納されてもよい。すなわち、ＣＰＵ１０ａは、記録媒体１０ｆ、又は読取部１０ｇを介して記録媒体１０ｈから出力された制御プログラムを、メモリ１０ｂ等の記憶装置に展開して実行することにより、ノード１０の機能を実現する。

【００２８】

なお、上述した各ハードウェアは、互いにバスを介して通信可能に接続される。例えば、ＣＰＵ１０ａ、メモリ１０ｂ、及びネットワークインタフェース１０ｄは、システムバスに接続される。また、例えば、記憶部１０ｃ、入出力部１０ｅ、記録媒体１０ｆ、及び読取部１０ｇは、Ｉ／Ｏ（Input/Output）インタフェース等を介してシステムバスに接続される。なお、記憶部１０ｃは、ＳＣＳＩ、ＳＡＳ（Serial Attached SCSI）、ファイバチャネル、ＳＡＴＡ（Serial Advanced Technology Attachment）等に準拠したバス（ケーブル）で、ＤＩ（Disk Interface）等のＩ／Ｏインタフェースに接続される。

【００２９】

なお、ノード１０の上述したハードウェア構成は例示である。従って、ノード１０内でのハードウェアの増減や分割等は適宜行なわれてもよい。

スイッチ（接続装置）２０は、複数のノード１０間又は他のスイッチ２０間に接続され、スイッチ２０に接続されたノード１０間でやり取りされるコマンド又はデータ等の情報を中継する。スイッチ２０としては、例えばＬ２（Layer 2）スイッチ、ＦＣスイッチ等のハードウェアスイッチが挙げられる。

【００３０】

図１に例示するストレージシステム１では、スイッチ２０－１は、スイッチ２０－２及び２０－３に接続される。また、スイッチ２０－２はスイッチ２０－１及びノード１０－１及び１０－２に、スイッチ２０－３はスイッチ２０－１及びノード１０－３～１０－５に、それぞれ接続される。なお、スイッチ２０は、図１に示すものに限られず、ノード１０の数等に応じて、多段に接続されてもよいし、１つのスイッチ２０が用いられてもよい。

【００３１】

なお、クライアントがインターネット又はイントラネット等のネットワークを介してス

トレージシステム１に接続される場合、スイッチ２０とクライアントとの間にルータが介設されてもよい。ルータとしては、例えば、ソフトウェアルータの他、Ｌ３スイッチ等のハードウェアルータ等が挙げられる。

〔１－２〕ノードの説明

第１実施形態の一例としてのストレージシステム１は、上述のように、相互に接続される複数のノード１０を有し、複数のノード１０間で通信を行なう。

【００３２】

具体的には、第１実施形態の一例としてのノード１０の各々は、以下の（ａ）～（ｃ）の処理を行なう。

（ａ）複数のノード１０のうちの自ノード１０以外の他ノード１０の各々から、他ノード１０により判定された複数のノード１０の各々の状態に関するノード状態情報Ｔ１（図４参照）を受信する。

【００３３】

（ｂ）他ノード１０の各々から受信したノード状態情報Ｔ１に基づいて、複数のノード１０の各々の状態を判定する。

（ｃ）判定した複数のノード１０の各々の状態に関するノード状態情報Ｔ１を、他ノード１０の各々へ送信する。

なお、ノード１０の状態とは、ノード１０が正常に動作しているか否かを示す種別であり、詳細は後述する。

【００３４】

ノード１０は、上記（ａ）～（ｃ）の処理を繰り返す。つまり、ノード１０の各々は、自ノード１０が判定（生成）したノード状態情報Ｔ１を、自ノードが正常に稼働していることを示すハートビートとして定期的に他ノード１０の各々へ送信する。そして、ノード１０の各々は、他ノード１０からハートビートとして送信されてきたノード状態情報Ｔ１を受信し、自ノード１０が保持する管理情報を更新する。これにより、ノード１０は、ストレージシステム１内の複数のノード１０間で各ノード１０の状態を共有し、他ノード１０からのノード状態情報Ｔ１に基づいて自律的に複数のノード１０の各々の状態を判定することができる。

【００３５】

なお、ストレージシステム１において、複数のノード１０の接続形態は、図１に例示したものに限定されないが、複数のノード１０間が離間するほど、ノード状態情報Ｔ１の送受信においてレイテンシ又はパケットロス等が生じ得る。従って、ストレージシステム１において、ノード状態情報Ｔ１を送受信する複数のノード１０は、ネットワークの品質が互いに均一となることが望ましい。

【００３６】

〔１－３〕ノードの構成

次に、図３～図１０を参照して、第１実施形態の一例としてのノード１０の構成について説明する。

図３は、第１実施形態の一例としてのノード１０の機能構成例を示す図である。図４は、ノード１０が送受信するノード状態情報Ｔ１を例示する図であり、図５は、ノード１０（特にノード１０－１）が管理するノード状態管理情報Ｔ２を例示する図である。

【００３７】

ノード１０は、上述した処理を行なうため、図３に例示するように、ノード状態保持部１１、受信処理部１２、ノード状態決定部１３、送信処理部１４、リカバリ処理部１５、及び停止処理部１６をそなえる。なお、ノード１０－１～１０－５は、互いに同様の機能をそなえることができるため、以下、任意のノード１０がそなえる機能について説明する。

【００３８】

〔１－３－１〕ノード状態保持部

ノード状態保持部１１は、図５に示すノード状態管理情報Ｔ２を保持する記憶領域であ

り、例えば上述したメモリ１０ｂにより実現される。

〔１－３－２〕受信処理部

受信処理部１２は、上記（ａ）の処理を行なう。具体的には、受信処理部１２は、複数のノード１０のうちの自ノード１０以外の他ノード１０の各々から、図４に例示するノード状態情報Ｔ１を受信し、ノード状態保持部１１が保持するノード状態管理情報Ｔ２（図５参照）を更新する。

【００３９】

ノード状態情報（状態情報）Ｔ１は、送信元のノード１０で判定された各ノード１０の状態を含む情報である。例えば、自ノード１０が送信するノード状態情報Ｔ１には、自ノード１０が判定した各ノード１０の状態が含まれ、自ノード１０が受信するノード状態情報Ｔ１には、受信するノード状態情報の送信元である他ノード１０で判定された各ノード１０の状態が含まれる。なお、ノード１０は、図４に示すようにノード状態情報Ｔ１をテーブルとして生成し、送受信することができる。

【００４０】

図４に示すように、ノード状態情報Ｔ１は、ノード１０の識別情報の一例であるノードＩＤ、ノード１０ごとの状態、ノード１０のアドレスの一例であるＩＰ（Internet Protocol）アドレス、及びノード１０のポート番号を含む。図４に示すノード状態情報Ｔ１は、ノード１０－１～１０－５に対応するノードＩＤ“１”～“５”の状態を含む。

一例として、ノードＩＤ“１”には、状態“Alive”、ＩＰアドレス“192.168.0.1”、ポート番号“12345”が対応付けられる。

【００４１】

なお、ノード１０の識別情報として、ノードＩＤを例に挙げたが、これに限定されるものではない。識別情報は、各ノード１０を特定できるユニークな情報であればよい。例えば、識別情報として、ノード１０のＩＰアドレス、シリアル番号、又はネットワークインタフェース１０ｄのＭＡＣ（Media Access Control）アドレス等が用いられてもよい。

また、ノード１０のアドレスとして、ＩＰアドレスを例に挙げたが、これに限定されるものではない。アドレスは、ＩＰ以外のプロトコルにおいてノード１０を特定可能な種々のアドレスが用いられてもよい。

【００４２】

ノード状態管理情報Ｔ２は、自ノード１０及び他ノード１０で判定された複数のノード１０の各々の状態を管理する情報である。例えば、ノード状態管理情報Ｔ２は、自ノード１０が各ノード１０の状態をどう判断しているか、他ノード１０が各ノード１０をどう判断しているか、及び最後に各ノード１０からハートビートとしてのノード状態情報Ｔ１を受信したのはいつかといった情報を含む。なお、ノード１０は、図５に示すようにノード状態管理情報Ｔ２をテーブルとして生成し、管理することができる。

【００４３】

以下、図５の説明においては、自ノード１０がノード１０－１であるものとする。

図５に示すように、ノード状態管理情報Ｔ２は、図４に示すノード状態情報Ｔ１と同様に、ノード１０の識別情報の一例としてのノードＩＤ、ノード１０ごとの状態、ノード１０のアドレスの一例としてのＩＰアドレス、及びノード１０のポート番号を含む。また、ノード状態管理情報Ｔ２はさらに、他のノード１０から受信したノード状態情報Ｔ１に含まれるノード１０ごとの状態（図５中、“ｂｙ　２”～“ｂｙ　５”と表記）、及び他のノード１０ごとの最終更新情報を含む。図５に示すノード状態管理情報Ｔ２は、ノード１０－１～１０－５に対応するノードＩＤ“１”～“５”の状態を含む。

【００４４】

一例として、ノードＩＤ"１"には、自ノード１０－１が判定した状態"Alive"、他ノード１０－２～１０－５がそれぞれ判定した状態"Alive"、最終更新情報"1 sec ago"（１秒前）、ＩＰアドレス"192.168.0.1"、ポート番号"12345"が対応付けられる。つまり、ノード状態管理情報Ｔ２には、受信処理部１２が受信したノード状態情報Ｔ１が示す複数のノード１０の各々の状態が含まれる。また、ノード状態管理情報Ｔ２には、ノード状態決定

部１３が含まれる自ノード１０の状態に関するノード状態情報Ｔ１に関する自己状態情報が含まれる。

【００４５】

受信処理部１２は、他ノード１０の各々から上述したノード状態情報Ｔ１を受信すると、受信したノード状態情報Ｔ１に含まれるノード１０ごとの状態を、ノード状態管理情報Ｔ２における対応する他ノード１０の列に設定する。つまり、図５に例示する“ｂｙ　２”～“ｂｙ　５”（自ノード１０がノード１０－１の場合）の状態は、対応する他ノード１０からの情報に基づき設定される。なお、ノードＩＤ“４”の状態の説明は後述する。

【００４６】

また、ノード１０－１の受信処理部１２は、ノード１０－２からノード状態情報Ｔ１を受信すると、ノード状態管理情報Ｔ２において、ノード状態情報Ｔ１に含まれるノード１０ごとの状態を、“ｂｙ　２”の列に設定する。また、受信処理部１２は、ノード１０－２に対応するノードＩＤ“２”の最終更新情報を更新する。

なお、最終更新情報は、最後にハートビートを受信したのがいつであるかを示す情報であり、図５に示す例では、最終更新情報として、現在時刻と最後に受信を行なった時刻（最終受信時刻）との差を示しているが、これに限定されるものではない。例えば、ノード１０は、最終更新情報に最終受信時刻そのものを設定することで、最終更新情報を更新してもよい。また、ノード１０は、ノード１０ごとに、時間の経過に応じて値が変化（例えば増加）するタイマを実行し、ノード状態管理情報Ｔ２の最終更新情報では、対応するタイマ値を参照してもよい。最終更新情報にタイマ値が用いられる場合、ノード１０は、最終更新情報の更新の際に、タイマのカウント値をリセットすることで、最終更新情報を更新することができる。

【００４７】

受信処理部１２は、他ノード１０からノード状態情報Ｔ１を受信した都度、受信したノード状態情報Ｔ１に基づきノード状態管理情報Ｔ２を更新してもよい。また、受信処理部１２は、受信したノード状態情報Ｔ１を送信元のノード１０の識別情報と対応付けてメモリ１０ｂ等に保持しておき、後述する第１所定時間ごとに、メモリ１０ｂが保持するノード状態情報Ｔ１に基づきノード状態管理情報Ｔ２を更新してもよい。

【００４８】

また、受信処理部１２は、上述したノード状態情報Ｔ１の受信に加え、ノード１０のＩＰアドレス及びポート番号を受信することができる。

図６は、第１実施形態の一例としての新規ノード１０が送信する情報を例示する図であり、図７は、新規ノード１０が受信する情報を例示する図である。

ノード１０（後述する送信処理部１４）は、起動後、つまりストレージシステム１に追加されると、自ノード１０のＩＰアドレス及びポート番号を含む情報を全てのノード１０へ通知する。具体的には、ストレージシステム１に追加されたノード（新規ノード）１０は、図６に例示する送信情報Ｔ３を、ストレージシステム１内の全てのノード１０へブロードキャスト等により通知する。

【００４９】

図６に示すように、新規ノード１０が送信する送信情報Ｔ３は、新規ノード１０の識別情報の一例としてのノードＩＤ、新規ノード１０の状態、新規ノード１０のアドレスの一例としてのＩＰアドレス、及び新規ノード１０のポート番号を含む。例えば、図６に示す送信情報Ｔ３は、新規ノード１０に対応するノードＩＤ“６”の状態を含む。

一例として、ノードＩＤ“６”には、新規ノード１０が判定した状態“Alive”、ＩＰアドレス“192.168.0.6”、ポート番号“12345”が対応付けられる。

【００５０】

他ノード１０の受信処理部１２は、追加された新規ノード１０から送信情報Ｔ３を通知されると、送信情報Ｔ３に含まれるＩＰアドレス及びポート番号、並びに送信元のノードＩＤの情報をノード状態管理情報Ｔ２に追加する。以後、ノード１０（送信処理部１４）は、追加した新規ノード１０のＩＰアドレス及びポート番号に対してもハートビートを送

信する。
【００５１】
また、新規ノード１０（受信処理部１２）は、新規ノード１０が通知した送信情報Ｔ３を受け取った他ノード１０の各々から、順次ハートビート（ノード状態情報Ｔ１′）を受信する。なお、新規ノード１０が受け取るノード状態情報Ｔ１′は、図４に示すノード状態情報Ｔ１と同様のデータ構造であるが、新規ノード１０の情報が追加されているため、便宜上、ノード状態情報Ｔ１′と表記する。
【００５２】
図７に示すように、新規ノード１０が受信するノード状態情報Ｔ１′は、図４に示すノード状態情報Ｔ１に加えて、新規ノード１０に対応するノードＩＤ“６”の状態を含む。一例として、ノードＩＤ“６”には、他ノード１０が判定した新規ノード１０の状態“Alive”、ＩＰアドレス“192.168.0.6”、ポート番号“12345”が対応付けられる。
新規ノード１０（受信処理部１２）は、受信したノード状態情報Ｔ１′に含まれる他ノード１０のＩＰアドレス及びポート番号、並びに送信元のノードＩＤの情報からノード状態管理情報Ｔ２を作成又は更新する。これにより、新規ノード１０は、ノード状態情報Ｔ１′をハートビートとして定期的に送信する送信処理部１４のサービスを開始することができる。
【００５３】
〔１－３－３〕ノード状態決定部
ノード状態決定部（判定部）１３は、上記（ｂ）の処理を行なう。具体的には、ノード状態決定部１３は、ノード状態管理情報Ｔ２を参照してノード１０ごとの状態を判定し、ノード状態管理情報Ｔ２に設定する。より具体的に、ノード状態決定部１３は、受信処理部１２が受信したノード状態情報Ｔ１が示す複数のノード１０の各々の状態と、他ノード１０の各々からのノード状態情報Ｔ１の受信状況とに基づいて、複数のノード１０の各々の状態を判定する。
【００５４】
ここで、ノード１０の状態及び状態遷移について説明する。
図８は、第１実施形態の一例としてのノード１０が他ノード１０の状態を判定するときの状態遷移の一例を示す図であり、図９は、複数のノード１０によるノード状態情報Ｔ１の送受信処理の一例を説明する図である。図１０は、ノード１０が自ノード１０の状態を判定するときの状態遷移の一例を示す図である。
【００５５】
なお、図９に示す例においては、説明の簡略化のため、ノード１０間の接続状態のみを示し、スイッチ２０の図示を省略している。
〔１－３－３－１〕ノード状態決定部が他ノードについて判定する各状態の説明
はじめに、ノード１０（ノード状態決定部１３）が他ノード１０について判定する各状態について説明する。図８に示すように、ノード１０が他ノード１０について判定する状態には、Ａｌｉｖｅ、Ｓｕｓｐｅｃｔ、Ｄｏｗｎ、及びＺｏｍｂｉｅが含まれる。
【００５６】
Ａｌｉｖｅは、ノード１０が正常に動作している状態（稼動中）を示す。ノード状態決定部１３は、ノード状態管理情報Ｔ２を参照して、最終更新情報が第２所定時間内であり、且つ第１所定数以上のノード１０からＳｕｓｐｅｃｔと判定されていない他ノード１０の状態を、Ａｌｉｖｅと判定する。
なお、他ノード１０がストレージシステム１に追加された場合、ノード状態決定部１３は、追加された他ノード１０に関する最初の判定において、追加された他ノード１０の状態を初期状態であるＡｌｉｖｅと判定する（図８の矢印（Ｉ）参照）。
【００５７】
ここで、第２所定時間としては、ノード１０がノード状態情報Ｔ１を送信する時間周期である第１所定時間以上の時間とすることができる。例えば、各ノード１０がノード状態情報Ｔ１を１秒（第１所定時間）ごとに送信する場合、第２所定時間は、ノード１０の負

荷による送信処理の遅延又は通信経路の輻輳等を考慮して、数倍～数十倍程度の時間（例えば２０秒）とすることができる。

【００５８】

また、第１所定数としては、例えば過半数とすることができる。

以下、第１所定時間は１秒であり、第２所定時間は２０秒であり、第１所定数はノード１０の数の過半数であるものとして説明する。

Ｓｕｓｐｅｃｔ（第１状態）は、ノード１０が故障（停止）している疑いのある状態（停止の可能性）を示す。ノード状態決定部１３は、ノード状態管理情報Ｔ２を参照して、最終更新情報が第２所定時間よりも前である他ノード１０の状態、つまり第２所定時間内にノード状態情報Ｔ１を受信しなかった他ノード１０の状態を、Ｓｕｓｐｅｃｔと判定する。すなわち、ノード状態決定部１３は、ハートビートの不達時間が閾値（第２所定時間）を超えた他ノード１０の状態を、Ｓｕｓｐｅｃｔと判定する。

【００５９】

例えば、ノード状態決定部１３は、Ａｌｉｖｅの状態と判定した他ノード１０から、２０秒よりも長くノード状態情報Ｔ１を受信できない場合、当該他ノード１０の状態をＡｌｉｖｅからＳｕｓｐｅｃｔに遷移させる（図８の矢印（II）参照）。

また、ノード状態決定部１３は、Ｓｕｓｐｅｃｔの状態と判定した他ノード１０の状態が自ノード１０又は他ノード１０によりＤｏｗｎと判定される前に、当該他ノード１０からノード状態情報Ｔ１を受信する場合がある。この場合、ノード状態決定部１３は、当該他ノード１０の状態をＳｕｓｐｅｃｔからＡｌｉｖｅに遷移させる（図８の矢印（III）参照）。

【００６０】

Ｄｏｗｎ（第２状態）は、ノード１０において故障等の障害が発生している状態（停止中）を示す。ノード状態決定部１３は、第１所定数以上のノード１０でＳｕｓｐｅｃｔと判定された他ノードの状態、又は他ノード１０の少なくとも１つからＤｏｗｎと判断された他ノード１０の状態を、Ｄｏｗｎと判定する。

例えば、ノード状態決定部１３がＡｌｉｖｅ又はＳｕｓｐｅｃｔの状態と判定した他ノード１０の状態について、過半数以上のノード１０でＳｕｓｐｅｃｔと判定される場合、又は他ノード１０のうちのいずれかがＤｏｗｎと判定される場合がある。この場合、ノード状態決定部１３は、Ａｌｉｖｅ又はＳｕｓｐｅｃｔの状態と判定した当該他ノード１０の状態を、Ｄｏｗｎと判定する（図８の矢印（IV）又は（Ｖ）参照）。

【００６１】

一例として、図９に示すように、ノード１０－１がノード１０－２、１０－３、及び１０－５からノード状態情報Ｔ１を１秒ごとに受け取る一方、ノード１０－４からノード状態情報Ｔ１を３０秒間受け取っていない場合を考える。このとき、ノード状態管理情報Ｔ２は、図５に例示する状態になる。

つまり、自ノード１０－１は、ノード１０－４から２０秒よりも長くノード状態情報Ｔ１を受け取っていないため、ノード１０－４の状態をＳｕｓｐｅｃｔと判定する。また、他ノード１０－３及び１０－５も、ノード１０－４から２０秒よりも長くノード状態情報Ｔ１を受け取っておらず、他ノード１０－３及び１０－５によるノード１０－４の状態の判定結果もＳｕｓｐｅｃｔとなる。この場合、ノード状態決定部１３は、ノード１０－４の状態が過半数のノード１０によりＳｕｓｐｅｃｔと判定されたため、ノード１０－４の状態をＤｏｗｎに遷移させる。

【００６２】

このように、他ノード１０に障害等が発生した場合、ノード状態管理情報Ｔ２では、障害等が発生した他ノード１０の状態が、当該他ノード１０のノードＩＤの行方向（図５中、横軸方向）に順にＳｕｓｐｅｃｔに遷移する（図５中、ノードＩＤ“４”参照）。そして、ノード状態決定部１３は、Ｓｕｓｐｅｃｔになったノード１０の数が過半数に達した場合に、当該他ノード１０の状態をＤｏｗｎと判定するのである。

【００６３】

なお、図９に示す例では、ノード１０－１は、ノード１０－２～１０－５へノード状態情報Ｔ１を送信するが、ノード１０－４は故障（停止）している（又は疑いのある）状態であるため、ノード状態情報Ｔ１はノード１０－４により受信されない。

Ｚｏｍｂｉｅ（第３状態）は、ノード１０が後述するリカバリ処理部１５によりリカバリ処理が行なわれている状態（リカバリ処理中）を示す。Ｚｏｍｂｉｅは、ノード１０に故障等の障害が発生した後、障害が発生したノード１０のノード情報が削除されるまでの暫定状態である。クライアント及びリカバリ処理に係わるノード１０以外のノード１０は、Ｚｏｍｂｉｅの状態のノード１０へのアクセスが制限される。

【００６４】

具体的には、ストレージシステム１は、障害が発生したノード１０について、障害が発生したノード１０が保持するデータに関連するデータを持つノード１０により、リカバリ処理を実行させる。リカバリ処理は、上述のように、障害が発生したノード１０内のデータの冗長データを保持するノード１０から、冗長データを他ノード１０へコピーし、データの多重化状態を回復する処理である。

【００６５】

例えば、リカバリ処理部１５によるリカバリ処理中に、障害が発生したノード１０が復旧する場合、又は同一のノード名でストレージシステム１に追加される場合があり得る。この場合、ストレージシステム１上で障害が発生したノード１０に古いデータが存在する状態で、古いデータと独立してリカバリ処理が実行される状態が発生し、データの一貫性が崩れる可能性がある。

【００６６】

クライアントは、ストレージシステム１内でどのノード１０にデータが格納されているかを管理するテーブルを保持するが、このテーブルでは、ノード１０に障害が発生したことを即座に検知できない場合がある。仮に、クライアントが、障害が発生したノード１０からデータ（古いデータ）を取得してしまうと、取得したデータと、リカバリ処理が行なわれ、他ノード１０にコピーされた冗長データとの間で不整合が生じることになる。

【００６７】

以上の理由から、ノード状態決定部１３は、障害が発生したノード１０の状態を、リカバリ処理が完了する（古いデータが削除される）まで、Ｚｏｍｂｉｅと判定する。これにより、ノード状態決定部１３は、Ｚｏｍｂｉｅのノード１０に対して、クライアント及びリカバリ処理に係わるノード１０以外のノード１０からアクセスできないようにし、データの一貫性が崩れることを防止する。従って、Ｚｏｍｂｉｅの状態である期間は、リカバリ処理が完了するまで、障害が発生したノード１０から古いデータが読み出されることを抑止するガード期間であるといえる。

【００６８】

ノード状態決定部１３は、第２所定数以上のノード１０でＤｏｗｎと判定された他ノードの状態を、Ｚｏｍｂｉｅと判定する。

ここで、第２所定数としては、第１所定数以上の数、好ましくは、全てのノード１０の数とすることができる。以下、第２所定数は全てのノード１０の数であるものとして説明する。

【００６９】

例えば、ノード状態決定部１３は、自ノード１０を含め全てのノード１０（Ｄｏｗｎ又はＺｏｍｂｉｅの状態のノード１０を除く）でＤｏｗｎの状態と判定されたノード１０の状態を、ＤｏｗｎからＺｏｍｂｉｅに遷移させる（図８の矢印（VI）参照）。

ノード状態決定部１３は、全てのノード１０からＤｏｗｎと判定されたノード１０をＺｏｍｂｉｅとすることで、全てのノード１０の共通認識によって、障害が発生したノード１０をリカバリすべきノード１０であると確実に決定することができる。

【００７０】

なお、リカバリ処理が完了すると、障害が発生したノード１０以外のノード１０のノード状態決定部１３は、自ノード１０が保持するノード状態管理情報Ｔ２から、障害が発生

したノード１０に関する情報を削除する（図８の矢印（VII）参照）。
以上のように、ノード状態決定部１３は、受信処理部１２が受信したノード状態情報Ｔ１が示す複数のノード１０の各々の状態、つまり図５のノード状態管理情報Ｔ２における“他ノードからの情報”に基づいて、複数のノード１０の各々の状態を判定する。
【００７１】
また、ノード状態決定部１３は、以下に説明するように、ノード状態決定部１３が含まれる自ノード１０の状態に関して判断を行なった自己状態情報（図５の"自ノードでの判断"参照）にさらに基づいて、複数のノード１０の各々の状態を判定してもよい。
〔１－３－３－２〕ノード状態決定部が自ノードについて判定する各状態の説明
次に、ノード１０（ノード状態決定部１３）が自ノード１０について判定する各状態について説明する。図１０に示すように、ノード１０が自ノード１０について判定する状態には、Ａｌｉｖｅ、Ｉｓｏｌａｔｅ、及びＤｏｗｎが含まれる。
【００７２】
Ａｌｉｖｅ（初期状態）は、ノード状態決定部１３が他ノード１０について判定するＡｌｉｖｅと同様の状態であり、自ノード１０が正常に動作している状態（稼動中）を示す。
自ノード１０が起動したとき、ノード状態決定部１３は、自ノード１０に関する最初の判定において、自ノード１０の状態をＡｌｉｖｅと判定する（図１０の矢印（ｉ）参照）。
【００７３】
Ｉｓｏｌａｔｅ（第４状態）は、自ノード１０がストレージシステム１から切り離された状態を示す。Ｉｓｏｌａｔｅの状態になる場合としては、例えば自ノード１０からスイッチ２０までの経路で障害が発生した場合や、自ノード１０のネットワークインタフェース１０ｄが故障した場合等が挙げられる。
ノード状態決定部１３は、ノード状態管理情報Ｔ２を参照して、第２所定時間内に第３所定数以上の他ノード１０からノード状態情報Ｔ１を受信しなかった場合、自ノード１０の状態を、ＡｌｉｖｅからＩｓｏｌａｔｅに遷移させる。すなわち、ノード状態決定部１３は、ハートビートの不達時間が閾値（第２所定時間）を超えた他ノード１０の数が第３所定数以上である場合、自ノード１０の状態をＩｓｏｌａｔｅと判定するのである。
【００７４】
ここで、第３所定数としては、第１所定数と同様、例えばノード１０の数の過半数とすることができる。
以下、第３所定数はノード１０の数の過半数であるものとして説明する。
例えば、ノード状態決定部１３は、ハートビートの不達時間が閾値を超えたノード１０の数が過半数に達した場合に、自ノード１０の状態をＩｓｏｌａｔｅに遷移させる（図１０の矢印（ii）参照）。
【００７５】
なお、自ノード１０が経路障害等によりストレージシステム１から切り離された場合、受信処理部１２は、他ノード１０からハートビートを受信しない。その結果、ノード状態管理情報Ｔ２では、自ノード１０で判定したノード１０ごとの状態が列方向（図５中、縦軸方向）に順にＳｕｓｐｅｃｔに遷移する。そして、ノード状態決定部１３は、Ｓｕｓｐｅｃｔになったノード１０の数が過半数に達した場合に、自ノード１０の状態をＩｓｏｌａｔｅと判定するのである。
【００７６】
また、自ノード１０の状態がＩｓｏｌａｔｅに遷移した場合、後述する停止処理部１６による停止処理により、自ノード１０は停止する（図１０の矢印（iii）参照）。
ところで、ノード１０は、自ノード１０の状態がＩｓｏｌａｔｅに遷移した場合、ストレージシステム１から切り離されているため、自ノード１０の状態がＩｓｏｌａｔｅであることを他のノードへノード状態情報Ｔ１により伝えることができない。また、ノード１０は、他ノード１０の状態がＩｓｏｌａｔｅになった場合にも、当該他ノード１０はスト

レージシステム１から切り離されているため、他ノード１０の状態がＩｓｏｌａｔｅになったことをノード状態情報Ｔ１により検知することができない。

【００７７】

自ノード１０の状態がＩｓｏｌａｔｅに遷移した場合、他ノード１０間でやり取りされるノード状態情報Ｔ１内では、自ノード１０の状態としてＳｕｓｐｅｃｔ、Ｄｏｗｎ、Ｚｏｍｂｉｅの順で遷移する。換言すれば、ノード１０は、他ノード１０の状態をＳｕｓｐｅｃｔ、Ｄｏｗｎ、又はＺｏｍｂｉｅと判定する場合、当該他ノード１０自身で判定した状態はＩｓｏｌａｔｅである可能性がある。

【００７８】

Ｄｏｗｎ（第２状態）は、ノード状態決定部１３が他ノード１０について判定するＤｏｗｎと同様の状態であるが、Ｄｏｗｎに遷移するまでの判定内容が、他ノード１０について判定する場合と異なる。ノード１０（例えばノード状態決定部１３）は、自ノード１０内で所定の障害が発生したことを検出した場合、自ノード１０の状態をＡｌｉｖｅからＤｏｗｎに遷移させる。

【００７９】

所定の障害としては、例えば自ノード１０による復旧が不可能又は困難な障害であり、ハードウェア障害等が挙げられる。なお、ノード１０による自ノード１０での障害の発生の検出は、既知の種々の手法により行なうことが可能であり、その詳細な説明は省略する。

ノード状態決定部１３は、自ノード１０に例えば復旧不可能な障害が発生した場合、自ノード１０の状態を、Ｄｏｗｎと判定する（図１０の矢印（ｉｖ）参照）。

【００８０】

また、自ノード１０の状態がＤｏｗｎに遷移した場合、後述する停止処理部１６による停止処理により、自ノード１０は停止する（図１０の矢印（ｖ）参照）。

なお、ノード１０は、自ノード１０の状態をＩｓｏｌａｔｅ又はＤｏｗｎと判定した場合、他ノード１０で判定される自ノード１０の状態は、Ｓｕｓｐｅｃｔ、Ｄｏｗｎ、Ｚｏｍｂｉｅの順で遷移する。

【００８１】

他ノード１０により、自ノード１０の状態がＺｏｍｂｉｅと判定されると、上述のように、自ノード１０に対するリカバリ処理が実行され、自ノード１０以外のノード１０が保持するノード状態管理情報Ｔ２から、自ノード１０に関する情報が削除される。

ノード状態決定部１３は、上述のように、自ノード１０及び他ノード１０の状態を判定し、ノード状態管理情報Ｔ２を更新する。

【００８２】

具体的には、ノード状態決定部１３は、自ノード１０及び他ノード１０の各々について判定した状態を、図５に例示するノード状態管理情報Ｔ２における“状態”の列に設定する。

ノード状態決定部１３は、以上のようにして、ノード状態管理情報Ｔ２に基づき複数のノード１０の各々の状態を判定することができる。つまり、ノード状態決定部１３は、受信処理部１２が受信したノード状態情報Ｔ１が示す複数のノード１０の各々の状態と、ノード状態決定部１３が含まれる自ノード１０の状態に関するノード状態情報Ｔ１に関する自己状態情報とに基づいて、上記判定を行なう。

【００８３】

なお、ノード状態決定部１３による上述した判定は、第１所定時間置きに全ノード１０について一括で行なわれてもよいし、ノード１０ごとに異なるタイミングで、第１所定時間置きに行なわれてもよい。

また、ノード１０は、ノード状態決定部１３により自ノード１０の状態がＤｏｗｎ又はＩｓｏｌａｔｅと判定された場合、ノード状態保持部１１が保持するノード状態管理情報Ｔ２を、記録媒体１０ｆ等の不揮発性メモリに保存してもよい。これにより、リカバリ処理後、作業者等は、ノード１０の停止要因が復旧不可能又は困難な障害（Ｄｏｗｎ）によ

るものか、ストレージシステム１から切り離されたこと（Ｉｓｏｌａｔｅ）によるものかを判断でき、障害復旧を迅速に行なうことができる。

【００８４】

〔１－３－４〕送信処理部

送信処理部１４は、上記（ｃ）の処理を行なう。具体的には、送信処理部１４は、第１所定時間ごとに、ノード状態決定部１３が判定した複数のノード１０の各々の状態に関するノード状態情報Ｔ１を、他ノード１０の各々へ送信する。

より具体的に、送信処理部１４は、ノード状態管理情報Ｔ２を参照して、ＩＰアドレス及びポート番号を取得し、他ノード１０へ送信するノード状態情報Ｔ１の宛先ノードを判定する。また、送信処理部１４は、ノード状態管理情報Ｔ２を参照して、自ノード１０が判定した各ノード１０についてのノードＩＤ、状態、ＩＰアドレス、及びポート番号の情報からノード状態情報Ｔ１を生成する。そして、送信処理部１４は、生成したノード状態情報Ｔ１を、ハートビートとして他ノード１０の各々へ送信する。

【００８５】

以下、受信処理部１２が受信するノード状態情報Ｔ１及び送信処理部１４が送信するノード状態情報Ｔ１は、同様のデータ構造であるが、便宜上、送信処理部１４が送信するノード状態情報Ｔ１を送信用ノード状態情報（送信用状態情報）Ｔ１という場合がある。

また、送信処理部１４は、ノード状態情報Ｔ１の送信に加え、上述のように、自ノード１０の起動後、送信情報Ｔ３（図６参照）をストレージシステム１内の全てのノード１０へブロードキャスト等により通知する。

【００８６】

〔１－３－５〕リカバリ処理部

リカバリ処理部１５は、他ノード１０に対してリカバリ処理を実行する。具体的には、リカバリ処理部１５は、ノード状態決定部１３がＺｏｍｂｉｅと判定したノード１０に対して、リカバリ処理を実行する。

なお、リカバリ処理は、全てのノード１０により行なわれなくても、ノード状態管理情報Ｔ２においてＺｏｍｂｉｅと判定されたノード１０に関係するノード１０により行なわれればよい。

【００８７】

例えば、Ｚｏｍｂｉｅと判定されたノード１０内のデータの冗長データ又は関連するデータを保持するノード１０のリカバリ処理部１５が、上記冗長データ又は関連するデータを他ノード１０の記憶部１０ｃへコピーすればよい。又は、上記冗長データ又は関連するデータのコピー先のノード１０のリカバリ処理部１５が、上記冗長データ又は関連するデータを保持するノード１０からデータを取得し、自ノード１０の記憶部１０ｃ等へ保存してもよい。

【００８８】

リカバリ処理部１５は、リカバリ処理においてコピーが完了すると、Ｚｏｍｂｉｅと判定されたノード１０内のデータを削除して、リカバリ処理を終わる。なお、Ｚｏｍｂｉｅと判定されたノード１０が停止した場合等には、リカバリ処理部１５は、Ｚｏｍｂｉｅと判定されたノード１０内のデータを削除できない可能性がある。この場合、リカバリ処理部１５は、Ｚｏｍｂｉｅと判定されたノード１０内のデータの削除を行なわずに、リカバリ処理を終わってもよい。また、リカバリ処理部１５は、リカバリ処理が完了すると、リカバリ処理の完了をノード状態決定部１３へ通知する。

【００８９】

ノード状態決定部１３は、リカバリ処理の完了が通知されると、Ｚｏｍｂｉｅと判定したノード１０に関するノードＩＤ、自ノード１０及び各ノード１０で判定された状態、最終更新情報、ＩＰアドレス、及びポート番号をノード状態管理情報Ｔ２から削除する。これにより、各ノード１０は、Ｚｏｍｂｉｅと判定されたノード１０をストレージシステム１から完全に切り離すことができる。

【００９０】

なお、ノード状態管理情報Ｔ２から情報を削除されたノード１０は、例えば作業者等により、修理又は交換によりストレージシステム１へ再組込み可能な状態になると、起動され、自ノード１０の状態をＡｌｉｖｅと判定する（図１０の矢印（ｉ）参照）。このとき、上述のように、新規ノード１０はＩＰアドレス及びポート番号を他ノード１０へ通知し、各ノード１０のノード状態管理情報Ｔ２に新規ノード１０の情報が追加されて、使用可能な状態になる。

【００９１】

ところで、ストレージシステム１の運用において、作業者等は、障害が発生したノード（障害ノード）１０が保持する情報の初期化、又は故障個所等の交換（ノード１０全体又は部品交換等の場合）を行ない、障害ノード１０の復旧を行なう。そして、作業者等は、復旧した障害ノード１０をストレージシステム１へ再組込みすることで、他ノード１０に新規ノード１０として認識させることができる。従って、障害により低下した、障害ノード１０に関するデータの多重度及びノード１０の冗長度を回復させることができる。

【００９２】

また、復旧前の障害ノード１０が使用していたＩＰアドレスは、他ノード１０のノード状態管理情報Ｔ２から削除されているため、復旧後の障害ノード１０は、再組込み後も同じＩＰアドレスを使いまわすことができる。従って、ストレージシステム１の管理者は、ストレージシステム１におけるＩＰアドレスの管理を容易に行なうことができ、利便性が高い。

【００９３】

〔１－３－６〕停止処理部

停止処理部１６は、自ノード１０に所定の障害が発生し、ノード状態決定部１３が自ノード１０の状態をＤｏｗｎと判定した場合、又は、ノード状態決定部１３が自ノード１０の状態をＩｓｏｌａｔｅと判定した場合、自ノード１０を停止させる処理を行なう。

なお、停止処理部１６による停止処理は、リカバリ処理部１５によるリカバリ処理が完了した後、具体的にはリカバリ処理において故障ノード１０内のデータが削除された後に行なわれることが好ましい。

【００９４】

また、他ノード１０のリカバリ処理部１５が、障害が発生したノード１０に対するリカバリ処理の完了後又はリカバリ処理の過程で、障害が発生したノード１０の停止処理を行なってもよい。この場合、停止処理部１６を省略することができる。

以上のように、第１実施形態の一例としてのストレージシステム１によれば、複数のノード１０により、メッシュ状に、ノード１０間でハートビートが行なわれる。ハートビートには、各ノード１０で判定された複数のノード１０の各々の状態が含まれ、複数のノード１０間でノード１０の各々の状態が共有される。

【００９５】

従って、ストレージシステム１は、個々のノード１０が自律的に判定した他ノード１０の状態の判定結果に基づき、各ノード１０の状態について、信頼性の高い判定結果を得ることができる。つまり、特定のノード又は監視装置等がノードの状態を集中的に監視する場合、特定のノード等により他ノードの状態について誤った判定がされる場合がある。これに対し、ストレージシステム１によれば、各ノード１０は、複数のノード１０から見た各ノード１０の状態を考慮して、自ノード１０及び他ノード１０の状態を判定することができるため、特定のノード等により誤った判定がされることを防止できる。

【００９６】

また、各ノード１０は、判定結果を共有し、信頼性の高い判定結果を得ることができるため、スプリットブレイン状態に陥ることを抑止できる。なお、ノード１０は、仮にスプリットブレイン状態に陥ったとしても、自ノード１０がＩｓｏｌａｔｅになると自律的に停止するため、冗長データの不整合が発生することを抑止できる。

さらに、各ノード１０は、ノード１０の状態をハートビート等の簡素な手法により共有するため、従来の手法と比較して、高速に、且つ容易に、自ノード１０及び他ノード１０

の状態を判定することができる。

【００９７】

従って、各ノード１０は、例えば、障害が発生したノード１０を高速に検出することが可能となり、クライアントからストレージシステム１へのアクセスの停止時間の短縮や、信頼性が低下する時間の短縮を図ることが可能となる。

〔１－４〕動作例

次に、図１１～図１３を参照して、上述の如く構成された第１実施形態の一例としてのノード１０による動作例を説明する。図１１は、第１実施形態の一例としての新規ノード１０による起動後の動作例を説明するフローチャートである。図１２は、ノード１０による他ノード１０の状態を判定する動作例を説明するフローチャートであり、図１３は、ノード１０による自ノード１０の状態を判定する動作例を説明するフローチャートである。

【００９８】

〔１－４－１〕新規ノードによる起動後の動作例

はじめに、図１１を参照して、新規ノード１０による起動後の動作例を説明する。

図１１に示すように、ノード（新規ノード）１０が起動し（ステップＳ１）、ストレージシステム１内のネットワークに接続されると、新規ノード１０のノード状態決定部１３により、自ノード１０の状態がＡｌｉｖｅと判定される（ステップＳ２）。

【００９９】

次いで、送信処理部１４により、自ノード１０のＩＰアドレス及びポート番号等のノード情報が収集され、送信情報Ｔ３（図６参照）が生成される。そして、送信処理部１４により、生成した送信情報Ｔ３がブロードキャスト等によりストレージシステム１内の全てのノード１０へ送信される（ステップＳ３）。

送信情報Ｔ３を受信した他ノード１０の各々は、新規ノード１０のノード情報をノード状態管理情報Ｔ２に追加し、宛先に新規ノード１０を含めてハートビート（ノード状態情報Ｔ１）を送信する。

【０１００】

新規ノード１０では、受信処理部１２により、ハートビートが待ち受けられる（ステップＳ４，ステップＳ４のＮｏルート）。他ノード１０からハートビートが受信されると（ステップＳ４のＹｅｓルート）、受信処理部１２により、受信したノード状態情報Ｔ１′（図７参照）から他ノード１０の各々のノード情報が抽出され、ノード状態管理情報Ｔ２が作成される（ステップＳ５）。

【０１０１】

そして、新規ノード１０では、ノード状態管理情報Ｔ２に基づいて、送信処理部１４による第１所定時間ごとにハートビートを送信するサービスが開始され（ステップＳ６）、新規ノード１０による起動後に行なわれる処理が終了する。

〔１－４－２〕ノードによる他ノードの状態を判定する動作例

次に、図１２を参照して、ノード１０による他ノード１０の状態を判定する動作例を説明する。

【０１０２】

なお、図１２に示すステップＳ１１～Ｓ２３の処理は、ノード１０の各々において、ノード状態決定部１３により一のノード１０の状態が判定される際に行なわれる処理である。従って、ステップＳ１１～Ｓ２３の処理は、各ノード１０のノード状態決定部１３により、他ノード１０の各々について、定期的（第１所定時間ごと）に実行される。

図１２に示すように、ノード状態決定部１３により、ノード状態管理情報Ｔ２内の“状態”が参照され、判定対象のノード１０について直前に判定した状態がどの状態であるかが判定される（ステップＳ１１，Ｓ１６，Ｓ１９）。

【０１０３】

判定対象のノード１０について直前に判定した状態がＡｌｉｖｅである場合（ステップＳ１１のＹｅｓルート）、ノード状態決定部１３により、判定対象のノード１０からのハートビートの不達時間が閾値を超えたか否かが判定される（ステップＳ１２）。このとき

、ノード状態決定部１３は、ノード状態管理情報Ｔ２の“最終更新情報”の時間が第２所定時間よりも長いか否かを判定する。

【０１０４】

ハートビートの不達時間が閾値を超えた場合（ステップＳ１２のＹｅｓルート）、ノード状態決定部１３により、判定対象のノード１０の状態がＳｕｓｐｅｃｔと判定され（ステップＳ１３）、処理が終了する。このとき、ノード状態決定部１３は、判定対象のノード１０について、ノード状態管理情報Ｔ２内の“状態”にＳｕｓｐｅｃｔを設定する。そして、ノード状態決定部１３は、次の判定対象のノード１０がある場合、次の判定対象のノード１０に係る状態の判定処理に移行する。

【０１０５】

一方、ステップＳ１２において、ノード状態決定部１３により、ハートビートの不達時間が閾値を超えていないと判定された場合（ステップＳ１２のＮｏルート）、処理がステップＳ１４に移行する。ステップＳ１４では、ノード状態決定部１３により、判定対象のノード１０の状態が、過半数（第１所定値）のノード１０からＳｕｓｐｅｃｔと判定されたか否か、又は複数のノード１０のいずれかのノード１０によりＤｏｗｎと判定された否かが判定される。

【０１０６】

判定対象のノード１０の状態が、過半数のノード１０からＳｕｓｐｅｃｔと判定されておらず、複数のノード１０のいずれかのノード１０によりＤｏｗｎとも判定されていない場合（ステップＳ１４のＮｏルート）、判定対象のノード１０に対する処理が終了する。一方、判定対象のノード１０の状態が、過半数のノード１０からＳｕｓｐｅｃｔと判定された又は複数のノード１０のいずれかのノード１０によりＤｏｗｎと判定された場合（ステップＳ１４のＹｅｓルート）、処理がステップＳ１５に移行する。

【０１０７】

ステップＳ１５では、ノード状態決定部１３により、判定対象のノード１０の状態がＤｏｗｎと判定され、処理が終了する。このとき、ノード状態決定部１３は、判定対象のノード１０について、ノード状態管理情報Ｔ２内の“状態”にＤｏｗｎを設定する。

また、判定対象のノード１０について直前に判定した状態がＳｕｓｐｅｃｔである場合（ステップＳ１１のＮｏルートからステップＳ１６のＹｅｓルート）、処理がステップＳ１７に移行する。ステップＳ１７では、ノード状態決定部１３により、判定対象のノード１０から新たなハートビートが受信されたか否か、つまりハートビートの不達時間が閾値未満となったか否かが判定される。このとき、ノード状態決定部１３は、ノード状態管理情報Ｔ２の“最終更新情報”の時間が第２所定時間未満であるか否かを判定する。

【０１０８】

新たなハートビートが受信されていない場合（ステップＳ１７のＮｏルート）、処理がステップＳ１４に移行する。一方、新たなハートビートが受信された場合（ステップＳ１７のＹｅｓルート）、ノード状態決定部１３により、判定対象のノード１０の状態がＡｌｉｖｅと判定され（ステップＳ１８）、処理が終了する。このとき、ノード状態決定部１３は、判定対象のノード１０について、ノード状態管理情報Ｔ２内の“状態”にＡｌｉｖｅを設定する。

【０１０９】

判定対象のノード１０について直前に判定した状態がＤｏｗｎである場合（ステップＳ１１のＮｏルート，ステップＳ１６のＮｏルートからステップＳ１９のＹｅｓルート）、処理がステップＳ２０に移行する。ステップＳ２０では、ノード状態決定部１３により、判定対象のノード１０の状態が第２所定値の数のノード１０（例えば全ノード１０）によりＤｏｗｎと判定されたか否かが判定される。

【０１１０】

全ノード１０によりＤｏｗｎと判定されていない場合（ステップＳ２０のＮｏルート）、判定対象のノード１０に対する処理が終了する。一方、全ノード１０によりＤｏｗｎと判定された場合（ステップＳ２０のＹｅｓルート）、ノード状態決定部１３により、判定

対象のノード１０の状態がＺｏｍｂｉｅと判定される。また、自ノード１０が保持するデータが判定対象のノード１０が保持するデータに関連する場合、リカバリ処理部１５により、判定対象のノード１０に対するリカバリ処理が実行され（ステップＳ２１）、処理が終了する。このとき、ノード状態決定部１３は、判定対象のノード１０について、ノード状態管理情報Ｔ２内の“状態”にＺｏｍｂｉｅを設定する。

【０１１１】

判定対象のノード１０について直前に判定した状態がＺｏｍｂｉｅである場合（ステップＳ１１のＮｏルート，ステップＳ１６のＮｏルートからステップＳ１９のＮｏルート）、処理がステップＳ２２に移行する。ステップＳ２２では、ノード状態決定部１３により、判定対象のノード１０についてリカバリ処理が完了したか否かが判定される。リカバリ処理が完了していない場合、判定対象のノード１０に対する処理が終了する。一方、リカバリ処理が完了した場合（ステップＳ２２のＹｅｓルート）、ノード状態決定部１３により、ノード状態管理情報Ｔ２から、判定対象のノード１０に関する情報が削除され（ステップＳ２３）、処理が終了する。

【０１１２】

以上のように、ノード１０により、一のノード１０の状態の判定処理が行なわれる。

〔１－４－３〕ノードによる自ノードの状態を判定する動作例

次に、図１３を参照して、ノード１０による自ノード１０の状態を判定する動作例を説明する。

なお、図１３に示すステップＳ３１～Ｓ３４の処理は、ノード１０の各々において、ノード状態決定部１３により自ノード１０の状態が判定される際に行なわれる処理である。従って、ステップＳ３１～Ｓ３４の処理は、各ノード１０のノード状態決定部１３により、定期的（第１所定時間ごと）に実行される。

【０１１３】

図１３に示すように、ノード状態決定部１３により、自ノード１０内で所定の障害の発生、例えば修復不可能な障害の発生が検出されたか否かが判定される（ステップＳ３１）。

所定の障害の発生が検出されると（ステップＳ３１のＹｅｓルート）、ノード状態決定部１３により、自ノード１０の状態がＤｏｗｎと判定され（ステップＳ３２）、処理が終了する。このとき、ノード状態決定部１３は、自ノード１０について、ノード状態管理情報Ｔ２内の“状態”にＤｏｗｎを設定する。

【０１１４】

一方、所定の障害の発生が検出されない場合（ステップＳ３１のＮｏルート）、ノード状態決定部１３により、ハートビートの不達時間が閾値を超えたノード数が過半数に達したか否かが判定される（ステップＳ３３）。このとき、ノード状態決定部１３は、ノード状態管理情報Ｔ２の“最終更新情報”の時間が第２所定時間よりも長い他ノード１０が第３所定値以上の数であるか否かを判定する。

【０１１５】

ハートビートの不達時間が閾値を超えたノード数が過半数である場合（ステップＳ３３のＹｅｓルート）、ノード状態決定部１３により、自ノード１０の状態がＩｓｏｌａｔｅと判定され（ステップＳ３４）、処理が終了する。このとき、ノード状態決定部１３は、自ノード１０について、ノード状態管理情報Ｔ２内の“状態”にＩｓｏｌａｔｅを設定する。

【０１１６】

一方、ステップＳ３３において、ノード状態決定部１３により、ハートビートの不達時間が閾値を超えたノード数が過半数未満であると判定された場合（ステップＳ３３のＮｏルート）、自ノード１０の状態に係る判定処理が終了する。そして、ノード状態決定部１３は、次の判定対象のノード１０がある場合、次の判定対象のノード１０に係る状態の判定処理に移行する。

【０１１７】

なお、ステップＳ３２又はＳ３４において、ノード状態決定部１３により、自ノード１０の状態がＤｏｗｎ又はＩｓｏｌａｔｅと判定されると、自ノード１０は、他ノード１０のリカバリ処理部１５からリカバリ処理を受ける。そして、自ノード１０は、停止処理部１６により、又は、他ノード１０のリカバリ処理部１５により、停止処理が行なわれる。

以上のように、ノード１０により、自ノード１０の状態の判定処理が行なわれる。

【０１１８】

〔１－５〕第１実施形態のまとめ

このように、第１実施形態の一例としてのストレージシステム１によれば、複数のノード１０の各々において、受信処理部１２は、他ノード１０の各々から、ノード状態情報Ｔ１を受信する。また、ノード状態決定部１３は、受信処理部１２が他ノード１０の各々から受信したノード状態情報Ｔ１に基づいて、複数のノード１０の各々の状態を判定する。さらに、送信処理部１４は、ノード状態決定部１３が判定した結果に基づき送信用ノード状態情報Ｔ１を、他ノード１０の各々へ送信する。

【０１１９】

従って、各々のノード１０は、特定のノード又は監視装置等によりノード１０の状態を集中的に監視するのではなく、他ノード１０が判定した複数のノード１０の状態に基づいて、自ノード１０及び他ノード１０を監視することができる。従って、特定のノード又は監視装置等の故障により、ストレージシステム１の利用が制限されるといった点を解消できる。また、各々のノード１０が自ノード１０及び他ノード１０を自律的に監視するため、監視を行なうノードを決定せずに済み、さらに管理者等を介入させずに済むため、ノード１０の故障後にストレージシステム１の利用が制限される時間を短縮できる。

【０１２０】

このように、第１実施形態の一例としてのストレージシステム１によれば、複数のノード１０をそなえるストレージシステム１において、複数のノード１０の状態の監視に伴う可用性の低下を抑止することができる。

また、ノード状態決定部１３は、受信処理部１２が受信したノード状態情報Ｔ１が示す複数のノード１０の各々の状態と、他ノード１０の各々からのノード状態情報Ｔ１の受信状況とに基づいて、複数のノード１０の各々の状態を判定する。また、送信処理部１４は、第１所定時間ごとに、送信用ノード状態情報Ｔ１を、他ノード１０の各々へ送信する。

【０１２１】

これにより、各々のノード１０は、第１所定時間ごとの他ノード１０の各々からのノード状態情報Ｔ１の受信状況に応じて、複数のノード１０の各々の状態を判定することができ、ノード状態情報Ｔ１を送信できないノード１０の異常を容易に検出することができる。

さらに、ノード状態決定部１３は、第２所定時間内にノード状態情報Ｔ１を受信しなかった他ノード１０の状態を、Ｓｕｓｐｅｃｔと判定する。また、ノード状態決定部１３は、第１所定数以上の複数のノード１０でＳｕｓｐｅｃｔであると判定されたノード１０の状態、又は、他ノード１０の少なくとも１つからＤｏｗｎであると判定されたノード１０の状態を、Ｄｏｗｎと判定する。

【０１２２】

これにより、各々のノード１０は、自ノード１０でノード状態情報Ｔ１が不達になったノード１０を直ちに障害等が発生したノード１０であると判断せず、他ノード１０の判断結果を考慮して、障害等が発生したノード１０を判定することができる。これにより、ノード１０は、各ノード１０の状態について、信頼性の高い判定結果を得ることができる。

また、ノード状態決定部１３は、第２所定数以上の複数のノード１０でＤｏｗｎであると判定されたノード１０を、Ｚｏｍｂｉｅと判定する。また、リカバリ処理部１５は、ノード状態決定部１３がＺｏｍｂｉｅと判定したノード１０に対して、リカバリ処理を実行する。

【０１２３】

これにより、リカバリ処理部１５は、第２所定数以上、例えば全てのノード１０がＤｏ

ｗｎであると判定したノード１０について、リカバリ処理を行なうため、誤った判断でリカバリ処理が行なわれることを抑止できる。また、障害等が発生したノード１０の状態がリカバリ処理中を示すＺｏｍｂｉｅ状態になることで、クライアント又はリカバリ処理を行なわないノード１０が古いデータを保持するＺｏｍｂｉｅ状態のノード１０へアクセスすることを抑止できる。

【０１２４】

さらに、ノード状態決定部１３は、自ノード１０に所定の障害が発生した場合、自ノード１０の状態をＤｏｗｎと判定する。また、ノード状態決定部１３は、第２所定時間内に第３所定数以上の他ノード１０からノード状態情報Ｔ１を受信しなかった場合、自ノード１０の状態を、Ｉｓｏｌａｔｅと判定する。さらに、停止処理部１６は、ノード状態決定部１３が自ノード１０の状態をＤｏｗｎ又はＩｓｏｌａｔｅと判定した場合、自ノード１０を停止させる。

【０１２５】

これにより、クライアント又はリカバリ処理を行なわないノード１０が、自ノード１０が保持する古いデータへアクセスすることを抑止できる。また、Ｉｓｏｌａｔｅになったノード１０が自律的に停止するため、スプリットブレイン状態に陥ったとしても、冗長データの不整合の発生を抑止できる。

〔２〕第２実施形態

〔２－１〕ノードの説明

次に、第２実施形態の一例としてのノード１０Ａについて説明する。

【０１２６】

第１実施形態及び第２実施形態に係るストレージシステム１は、多数（例えば、数十から数千台）のノードをそなえることがある。

上述のように、第１実施形態に係るストレージシステム１は、全ノード１０対全ノード１０の完全なメッシュ状態でハートビートの通信を行なう。

一方、第２実施形態に係るストレージシステム１は、ノード１０Ａをある程度（例えば数～数十台程度）のまとまり（以下、パーティという）に分割し、パーティ内のノード１０Ａ間では完全メッシュのハートビートの通信を行なう。一方、パーティ間では、各パーティの代表のノード１０Ａ（代表ノード１０Ａ）同士による完全メッシュのハートビートの通信を行なう。

【０１２７】

このように、第２実施形態の一例としてのストレージシステム１は、複数のノード１０Ａにより、階層的なノード１０Ａでの情報交換を行なう。これにより、ストレージシステム１は、全ノード１０Ａによる完全メッシュのハートビートの通信を行なうよりも、ストレージシステム１における通信負荷及び処理負荷を低減させることができる。特に、ストレージシステム１が、例えば数千台もの多数のノード１０Ａをそなえる場合に有効である。

【０１２８】

〔２－２〕ノードの構成

次に、図１４～図２３を参照して、第２実施形態の一例としてのノード１０Ａの構成について説明する。

図１４は、第２実施形態の一例としてのノード１０Ａの機能構成例を示す図である。

第２実施形態に係るノード１０Ａは、第１実施形態に係るノード１０と比べて、パーティ情報保持部１０１、パーティ間受信処理部１０２、パーティ間ノード状態決定部１０３、パーティ間送信処理部１０４、及びパーティ管理部１０５をさらにそなえる。

【０１２９】

また、第２実施形態に係るノード１０Ａは、第１実施形態に係るノード１０がそなえるノード状態保持部１１及び受信処理部１２とは一部の機能が異なるノード状態保持部１１Ａ及び受信処理部１２Ａをそなえる。

さらに、第２実施形態に係るノード１０Ａは、第１実施形態に係るノード１０がそなえ

るノード状態決定部１３及び送信処理部１４とは一部の機能が異なるノード状態決定部１３Ａ及び送信処理部１４Ａをそなえる。

【０１３０】

なお、ノード１０Ａは、上述した以外の点については、以下の説明において特に言及しない限り、ノード１０と同様の構成をそなえる。従って、以下、ノード１０Ａの説明において、ノード１０がそなえる構成と同一の符号の構成についての重複した説明は省略する。

〔２－２－１〕パーティ情報保持部及びノード状態保持部

パーティ情報保持部１０１は、図１５に示すパーティ管理情報Ｔ４を保持する記憶領域であり、例えば上述したメモリ１０ｂにより実現される。

【０１３１】

図１５は、第２実施形態の一例としてのノード１０Ａが管理するパーティ管理情報を例示する図である。

上述のように、第２実施形態の一例としてのストレージシステム１は、複数のノード１０Ａを数～数十台程度の複数のパーティに分割する。

パーティ管理情報Ｔ４は、複数のパーティとパーティに属するノード１０Ａとを対応付けて管理する情報である。なお、ノード１０Ａは、図１５に示すようにパーティ管理情報Ｔ４をテーブルとして生成し、送受信することができる。

【０１３２】

図１５に示すように、パーティ管理情報Ｔ４は、パーティの識別情報の一例であるパーティＩＤ、パーティに属するノード１０Ａの識別情報の一例であるノードＩＤ、及びパーティのバージョン番号を含む。図１５に示すパーティ管理情報Ｔ４は、パーティＩＤ“Ａ”～“Ｅ”についての情報を含む。

一例として、パーティＩＤ“Ａ”には、ノードＩＤ“１～１０”、バージョン番号“１”が対応付けられる。

【０１３３】

なお、パーティの識別情報として、パーティＩＤを例に挙げたが、これに限定されるものではない。識別情報は、各パーティを特定できるユニークな情報であればよい。例えば、識別情報として、アルファベットの他、数値、ノードＩＤの範囲の最小値又は最大値、ＩＰアドレスのマスク等が用いられてもよい。

また、ノード１０の識別情報として、ノードＩＤを例に挙げたが、これに限定されるものではなく、第１実施形態において既述のように、ノード１０Ａを特定できるユニークな情報であればよい。

【０１３４】

なお、図１５に例示するパーティ管理情報Ｔ４において、ノードＩＤにはパーティに属するノード１０Ａのノードの範囲（最小値～最大値）が設定されているが、これに限定されるものではない。例えば、ノードＩＤには、パーティに属するノード１０Ａのノードが複数の範囲、又は一つずつ設定されてもよい。

バージョン番号は、ノード１０Ａにおいて、自ノード１０Ａが持つパーティ管理情報Ｔ４が最新の情報であるか否かを判断するために用いられる。例えば、後述するパーティ管理部１０５により、パーティが分割又は統合される場合がある。この場合、分割又は統合が行なわれたパーティに属するノードＩＤも変化するため、各ノード１０Ａは、バージョン番号を参照して、最新のパーティ管理情報Ｔ４を識別するのである。

【０１３５】

ノード状態保持部１１Ａは、図１９に示すノード状態管理情報Ｔ７を保持する記憶領域であり、例えば上述したメモリ１０ｂにより実現される。

〔２－２－２〕パーティ間受信処理部及び受信処理部

次に、図１６～図１９を参照して、パーティ間受信処理部１０２及び受信処理部１２について説明する。

【０１３６】

図１６は、第２実施形態の一例としての複数のノード１０Ａによる代表ノード状態情報Ｔ５及びノード状態情報Ｔ６の送受信処理の一例を説明する図である。図１７は、ノード１０Ａが送受信する代表ノード状態情報Ｔ５を例示する図であり、図１８は、ノード１０Ａが送受信するノード状態情報Ｔ６を例示する図である。図１９は、ノード１０Ａが管理するノード状態管理情報Ｔ７を例示する図である。

【０１３７】

なお、図１６に示す例においては、説明の簡略化のため、ノード１０Ａ間の接続状態のみを示し、スイッチ２０の図示を省略している。

図１６に例示するように、代表ノード（代表ストレージ装置，代表情報処理装置）１０Ａは、複数のパーティのうちの自パーティ以外の他のパーティの各々における他の代表ノード１０Ａとの間で、代表ノード状態情報Ｔ５を送受信する。また、代表ノード１０Ａは、自パーティのパーティメンバであるメンバノード１０Ａへ代表ノード状態情報Ｔ５を送信し、メンバノード１０Ａは、自パーティの代表ノード１０Ａへノード状態情報Ｔ６を送信する。

【０１３８】

なお、図１６に示す例において、丸で囲われた数字は、ノードＩＤを示す。以下、例えばノードＩＤ“１”の代表ノード１０Ａを特定する場合には、代表ノード１０Ａ－１又はノード１０Ａ－１と表記する。また、例えばノードＩＤ“２”のメンバノード１０Ａを特定する場合には、メンバノード１０Ａ－２又はノード１０Ａ－２と表記する。

代表ノード１０Ａ及びメンバノード１０Ａは、特に言及しない限り互いに同様の機能をそなえることができるため、以下の説明において、任意のノード１０Ａがそなえる機能について説明する。

【０１３９】

パーティ間受信処理部（グループ間受信処理部）１０２は、自ノード１０Ａがパーティの代表ノード１０Ａである場合に、他のパーティの各々の代表ノード１０Ａから、図１７に例示する代表ノード状態情報Ｔ５を受信する。そして、代表ノード１０Ａのパーティ間受信処理部１０２は、受信した代表ノード状態情報Ｔ５に基づいて、ノード状態保持部１１Ａが保持するノード状態管理情報Ｔ７（図１９参照）を更新する。

【０１４０】

受信処理部１２Ａは、自ノード１０Ａが属するパーティ内の自ノード１０Ａ以外の他ノード１０Ａ（自パーティ内の代表ノード１０Ａを含む）の各々から、代表ノード状態情報Ｔ５又は図１８に例示するノード状態情報Ｔ６を受信する。そして、受信処理部１２Ａはノード状態保持部１１Ａが保持するノード状態管理情報Ｔ７（図１９参照）を更新する。

代表ノード状態情報（代表状態情報）Ｔ５は、送信元の代表ノード１０Ａにより判定された複数のパーティの代表ノード１０Ａの各々の状態に関する情報である。例えば、代表ノード１０Ａが送信する代表ノード状態情報Ｔ５には、代表ノード１０Ａが判定した自パーティ内のメンバノード１０Ａの状態と、他のパーティの代表ノード１０Ａから取得した他のパーティに属する全てのノード１０Ａの状態が含まれる。なお、代表ノード１０Ａは、図１７に示すように代表ノード状態情報Ｔ５をテーブルとして生成し、送受信することができる。

【０１４１】

例えば、図１７に示す例では、図１６に示す代表ノード１０Ａ－１は、他の代表ノード１０Ａ－１１及び１０Ａ－２１へ送信する代表ノード状態情報Ｔ５に、自パーティ内で判定した自パーティ内の各ノード１０Ａ－１～１０Ａ－３の状態を含める。また、代表ノード１０Ａ－１は、代表ノード状態情報Ｔ５に、他の代表ノード１０Ａ－１１及び１０Ａ－２１から受信した他のパーティ内のノード１０Ａ－１１～１０Ａ－１３及び１０Ａ－２１～１０Ａ－２３の状態を含める。

【０１４２】

また、代表ノード１０Ａは、自パーティ内のメンバノード１０Ａ－２及び１０Ａ－３に対しても他の代表ノード１０Ａへ送信するものと同様の代表ノード状態情報Ｔ５を送信し

、メンバノード１０Ａ－２及び１０Ａ－３からはノード状態情報Ｔ６を受信する。
つまり、パーティ内の代表ノード１０Ａ及びメンバノード１０Ａは、互いにパーティ内のノード１０Ａの状態の判定結果をハートビートで通知し合い、代表ノード１０Ａは、自パーティ内での判定結果を全パーティの代表ノード１０Ａへ伝達する。

【０１４３】

なお、代表ノード状態情報Ｔ５のデータ構造は、図４に示すノード状態情報Ｔ１と基本的に同様であるため、詳細な説明は省略する。
ノード状態情報（状態情報）Ｔ６は、送信元のノード１０Ａで判定された自パーティにおける他ノード（メンバノード）１０Ａの各々の状態を含む情報である。例えば、図１８に示す例では、図１６に示すメンバノード１０Ａ－２は、自パーティに属するノード１０Ａ－１及び１０Ａ－３へ送信するノード状態情報Ｔ６に、自パーティ内で判定した各ノード１０Ａ－１～１０Ａ－３の状態を含める。なお、ノード１０Ａは、図１８に示すようにノード状態情報Ｔ６をテーブルとして生成し、送受信することができる。

【０１４４】

なお、ノード状態情報Ｔ６のデータ構造は、図４に示すノード状態情報Ｔ１と基本的に同様であるため、詳細な説明は省略する。
以下、代表ノード状態情報Ｔ５及びノード状態情報Ｔ６を、単にノード状態情報Ｔ５及びＴ６と表記する場合がある。
ノード状態管理情報Ｔ７は、自ノード１０Ａ及び全パーティの全ノード１０Ａで判定された複数のノード１０Ａの各々の状態を管理する情報である。なお、ノード１０Ａは、図１９に示すようにノード状態管理情報Ｔ７をテーブルとして生成し、管理することができる。

【０１４５】

以下、図１９の説明においては、自ノード１０Ａが代表ノード１０Ａ－１であるものとする。
図１９に示すように、ノード状態管理情報Ｔ７は、図５に示すノード状態管理情報Ｔ２と同様に、ノード１０ＡのノードＩＤ、ノード１０Ａごとの状態、ノード１０ＡのアドレスのＩＰアドレス、及びノード１０Ａのポート番号を含む。また、ノード状態管理情報Ｔ７はさらに、他のノード１０Ａから受信したノード状態情報Ｔ５又はＴ６に含まれるノード１０Ａごとの状態、及び他のノード１０Ａごとの最終更新情報を含む。例えば、他のノード１０Ａから受信したノード状態情報Ｔ５又はＴ６に含まれるノード１０Ａごとの状態には、“ｂｙ　２”、“ｂｙ　３”、“ｂｙ　１１”～“ｂｙ　１３”、及び“ｂｙ　２１”～“ｂｙ　２３”が含まれる。

【０１４６】

図１９に示すノード状態管理情報Ｔ７は、ノード１０Ａ－１～１０Ａ－３、１０Ａ－１１～１０Ａ－１３、及び１０Ａ－２１～１０Ａ－２３に対応するノードＩＤ“１”～“３”、“１１”～“１３”、及び“２１”～“２３”の状態を含む。
一例として、ノードＩＤ“１”には、自ノード１０Ａが判定した状態“Alive”、他ノード１０Ａ－２、１０Ａ－３、１０Ａ－１１、及び１０Ａ－２１がそれぞれ判定した状態“Alive”、最終更新情報“1 sec ago”が対応付けられる。また、ノードＩＤ“１”にはさらに、ＩＰアドレス“192.168.0.1”、ポート番号“12345”が対応付けられる。

【０１４７】

パーティ間受信処理部１０２は、他の代表ノード１０Ａの各々から上述した代表ノード状態情報Ｔ５を受信すると、ノード状態管理情報Ｔ７を更新する。また、受信処理部１２Ａは、自パーティ内の他ノード１０Ａの各々から上述したノード状態情報Ｔ５又はＴ６を受信すると、ノード状態管理情報Ｔ７を更新する。具体的には、パーティ間受信処理部１０２及び受信処理部１２Ａは、受信したノード状態情報Ｔ５又はＴ６に含まれるノード１０Ａごとの状態を、ノード状態管理情報Ｔ７における対応する他ノード１０Ａの列に設定する。つまり、図１９に例示する他ノード１０Ａが判定した状態は、対応する他ノード１０Ａからの情報に基づき設定される。

【０１４８】
なお、パーティ間受信処理部１０２及び受信処理部１２Ａによる、ノード状態管理情報Ｔ７の更新は、第１実施形態に係る受信処理部１２による処理と同様であるため、重複した説明は省略する。
パーティ間受信処理部１０２及び受信処理部１２Ａは、受信処理部１２と同様に、ノード状態情報Ｔ５又はＴ６を受信した都度、又は第１所定時間ごとに、ノード状態管理情報Ｔ７を更新する。
【０１４９】
なお、受信処理部１２Ａは、上述したノード状態情報Ｔ５又はＴ６の受信に加え、図６及び図７を用いて上述したように、新規に追加されたノード１０ＡのＩＰアドレス及びポート番号を受信することができる。
また、パーティ間受信処理部１０２は、上述した代表ノード状態情報Ｔ５の受信に加え、図１５に示すパーティ管理情報Ｔ４を受信することができる。
【０１５０】
パーティ間受信処理部１０２は、代表ノード１０Ａからパーティ管理情報Ｔ４を受信すると、ノード状態保持部１１Ａが保持するパーティ管理情報Ｔ４と比較する。そして、パーティ間受信処理部１０２は、受信したパーティ管理情報Ｔ４に、新たに追加されたパーティＩＤ、又はバージョン番号が更新されたパーティＩＤがある場合、当該パーティＩＤの情報を用いて自ノード１０Ａが保持するパーティ管理情報Ｔ４を更新する。
【０１５１】
ところで、各パーティの代表ノード１０Ａは、所定のルールに基づいて決定される。例えば、代表ノード１０Ａは、各ノード１０Ａが保持するパーティ管理情報Ｔ４及びノード状態管理情報Ｔ７等に基づいて求められる。
一例として、代表ノード１０Ａは、パーティに属するノード１０Ａの中で、最も小さいノードＩＤを持つノード１０Ａとすることができる。このように、各ノード１０Ａが保持する情報から判断可能な所定のルールを予め定めておくことで、各ノード１０Ａは、代表ノード１０Ａを容易に選出することができる。
【０１５２】
これにより、代表ノード１０Ａに障害等が発生した場合であっても、パーティ内のノード１０Ａは、所定のルールに基づき次の代表ノード１０Ａを選出することができる。また、代表ノード１０Ａは、他のパーティの代表ノード１０Ａが停止した場合であっても、他のパーティの新たな代表ノード１０Ａを推定できるため、新たな代表ノード１０Ａとの間で、パーティ間のハートビートの通信を継続することができる。
【０１５３】
〔２－２－３〕パーティ間ノード状態決定部及びノード状態決定部
パーティ間ノード状態決定部（グループ間判定部）１０３は、パーティ間受信処理部１０２が他の代表ノード１０Ａの各々から受信した代表ノード状態情報Ｔ５に基づいて、複数の代表ノード１０Ａの各々の状態を判定する。
なお、パーティ間ノード状態決定部１０３による、代表ノード１０Ａ間でのノード１０Ａの各々の状態の判定手法は、第１実施形態に係るノード状態決定部１３によるノード１０間でのノード１０の各々の状態の判定と同様である。
【０１５４】
例えば、パーティ間ノード状態決定部１０３は、受信した代表ノード状態情報Ｔ５が示す複数の代表ノード１０Ａの各々の状態と、他の代表ノード１０Ａの各々からの代表ノード状態情報Ｔ５の受信状況とに基づいて、代表ノード１０Ａの各々の状態を判定する。
なお、代表ノード１０Ａは、他のパーティの代表ノード１０Ａの状態が全ての代表ノード１０ＡからＤｏｗｎであると判定された場合、パーティ管理情報Ｔ４及びノード状態管理情報Ｔ７から、当該他のパーティにおいて次に代表ノード１０Ａとなるべきノード１０Ａを判断する。この判断は、上述のように、代表ノード１０Ａを選出する所定のルールに基づいて行なわれる。

【０１５５】
そして、代表ノード１０Ａは、次の代表ノード１０Ａと判断した他のパーティのノード１０Ａへハートビートを送信する。代表ノード１０Ａは、ハートビートが疎通すると（他のパーティのノード１０Ａからハートビートを受信すると）、当該他のパーティのノード１０Ａを新たな代表ノード１０Ａと判断する。一方、他のパーティのノード１０Ａからのハートビートの不達時間が閾値を超えると、さらに次の代表ノード１０Ａとなるべきノード１０Ａを判断する。
【０１５６】
パーティ間ノード状態決定部１０３は、他のパーティ内の全ノード１０Ａに対してハートビートが疎通しなかった場合、当該他のパーティに属する全ノード１０Ａが停止したと判断する。この場合、パーティ間ノード状態決定部１０３は、当該他のパーティに属する全ノード１０Ａの状態をＺｏｍｂｉｅと判定し、リカバリ処理部１５にリカバリ処理を実行させる。
【０１５７】
ノード状態決定部（判定部）１３Ａは、受信処理部１２Ａが自パーティにおける他ノード１０Ａの各々から受信したノード状態情報Ｔ５又はＴ６に基づいて、自パーティにおけるノード１０の各々の状態を判定する。
なお、ノード状態決定部１３Ａによる、自パーティ内のノード１０Ａ間でのノード１０Ａの各々の状態の判定手法は、第１実施形態に係るノード状態決定部１３によるノード１０間でのノード１０の各々の状態の判定と同様である。
【０１５８】
例えば、ノード状態決定部１３Ａは、受信したノード状態情報Ｔ５又はＴ６が示す複数のノード１０Ａの各々の状態と、他ノード１０Ａの各々からのノード状態情報Ｔ５又はＴ６の受信状況とに基づいて、自パーティ内のノード１０Ａの各々の状態を判定する。
なお、ノード状態決定部１３Ａは、自パーティの代表ノード１０Ａの状態をＤｏｗｎと判定した場合、自パーティ内で生存している（Ａｌｉｖｅ状態の）ノード１０Ａ間で、上述した代表ノード１０Ａを選出する所定のルールを適用する。
【０１５９】
そして、各ノード１０Ａは、自ノード１０Ａが代表ノード１０Ａに昇格するか否かを判断し、昇格すると判断した場合、代表ノード１０Ａとして、他のパーティの代表ノード１０Ａとの間でハートビートの通信を開始する。
ここで、パーティ間ノード状態決定部１０３及びノード状態決定部１３Ａによる、ノード状態管理情報Ｔ７の参照箇所及び更新箇所について説明する。なお、この説明では、パーティ間ノード状態決定部１０３及びノード状態決定部１３Ａは、ノード１０Ａ－１にそなえられるものとする。
【０１６０】
図１９に示すように、ノード状態管理情報Ｔ７における“状態”の列の、二重線で囲われた領域は、自パーティ以外の他のパーティにおいて判定された状態である。従って、ノード１０Ａ－１～１０Ａ－３がそなえるパーティ間ノード状態決定部１０３及びノード状態決定部１３Ａは、二重線で囲われた領域（破線で四角く囲われた領域を除く）については基本的に判定及び更新を行なわない。
【０１６１】
また、図１９に示すように、ノード状態管理情報Ｔ７における“状態”の列の、破線で四角く囲われた領域は、複数のパーティの各代表ノード１０Ａで判定された状態である。従って、ノード１０Ａ－１がそなえるパーティ間ノード状態決定部１０３は、破線で四角く囲われた領域を、判定により更新する。
例えば、パーティ間ノード状態決定部１０３は、他の代表ノード１０Ａの最終更新情報を参照し、ハートビート（代表ノード状態情報Ｔ５）の到達の有無に応じてＡｌｉｖｅ又はＳｕｓｐｅｃｔの判定を行なう。また、パーティ間ノード状態決定部１０３は、他の代表ノード１０Ａについて、ノード状態管理情報Ｔ７における破線で丸く囲われた領域を参

照し、Ｓｕｓｐｅｃｔ、Ｄｏｗｎ、又はＺｏｍｂｉｅの判定を多数決等により行なう。
【０１６２】
さらに、図１９に示すように、ノード状態管理情報Ｔ７における“状態”の列の、実線で四角く囲われた領域は、自パーティ内の各ノード１０Ａで判定された状態である。従って、ノード１０Ａ－１がそなえるノード状態決定部１３Ａは、実線で四角く囲われた領域を、判定により更新する。
ノード状態決定部１３Ａは、他ノード１０Ａの最終更新情報を参照し、ハートビート（ノード状態情報Ｔ５又はＴ６）の到達の有無に応じてＡｌｉｖｅ又はＳｕｓｐｅｃｔの判定を行なう。また、ノード状態決定部１３Ａは、他ノード１０Ａについて、ノード状態管理情報Ｔ７における実線で角丸の四角で囲われた領域を参照し、Ｓｕｓｐｅｃｔ、Ｄｏｗｎ、又はＺｏｍｂｉｅの判定を多数決等により行なう。
【０１６３】
なお、パーティ間ノード状態決定部１０３及びノード状態決定部１３Ａによる判定の基準は、第１実施形態において既述のものと同様であり、詳細な説明は省略する。
また、パーティ間ノード状態決定部１０３及びノード状態決定部１３Ａは、上述のように、ノード１０Ａの状態を判定すると、ノード状態管理情報Ｔ７を更新する。
具体的には、パーティ間ノード状態決定部１０３及びノード状態決定部１３Ａは、自ノード１０Ａ及び他ノード１０Ａの各々について判定した状態を、図１９に例示するノード状態管理情報Ｔ７における“状態”の列に設定する。
【０１６４】
なお、パーティ間ノード状態決定部１０３及びノード状態決定部１３Ａによる上述した判定は、第１所定時間置きに判定対象の全ノード１０Ａについて一括で行なわれてもよいし、ノード１０Ａごとに異なるタイミングで、第１所定時間置きに行なわれてもよい。
〔２－２－４〕パーティ間送信処理部及び送信処理部
パーティ間送信処理部（グループ間送信処理部）１０４は、第１所定時間ごとに、パーティ間ノード状態決定部１０３が判定した複数の代表ノード１０Ａの各々の状態に関する代表ノード状態情報Ｔ５を、他の代表ノード１０Ａの各々へ送信する。
【０１６５】
具体的には、パーティ間送信処理部１０４は、パーティ管理情報Ｔ４及びノード状態管理情報Ｔ７を参照して、上述のように所定のルールに基づき、他のパーティの代表ノード１０Ａを特定する。そして、パーティ間送信処理部１０４は、ノード状態管理情報Ｔ７から、他の代表ノード１０ＡのＩＰアドレス及びポート番号を取得し、代表ノード状態情報Ｔ５の宛先ノードを判定する。
【０１６６】
また、パーティ間送信処理部１０４は、ノード状態管理情報Ｔ７を参照して、全ノード１０ＡについてのノードＩＤ、状態、ＩＰアドレス、及びポート番号の情報から代表ノード状態情報Ｔ５を生成する。そして、パーティ間送信処理部１０４は、生成した代表ノード状態情報Ｔ５を、ハートビートとして他の代表ノード１０Ａの各々へ送信する。
また、パーティ間送信処理部１０４は、代表ノード状態情報Ｔ５の送信に加え、後述するパーティ管理部１０５によりパーティ管理情報Ｔ４が更新された場合には、パーティ管理情報Ｔ４（図１４参照）をストレージシステム１内の全てのノード１０Ａへ通知する。なお、この通知は、ブロードキャスト等により行なわれてもよい。
【０１６７】
また、パーティ間送信処理部１０４は、パーティ管理情報Ｔ４が更新されたタイミングに限らず、パーティ管理情報Ｔ４を代表ノード状態情報Ｔ５とともにハートビートとして、他の代表ノード１０Ａへ送信してもよい。
送信処理部１４Ａは、送信用ノード状態情報Ｔ６を、自パーティにおける他ノード１０Ａの各々へ送信する。
【０１６８】
具体的には、送信処理部１４Ａは、パーティ管理情報Ｔ４及びノード状態管理情報Ｔ７

を参照して、自パーティ内の他ノード１０Ａを特定する。そして、送信処理部１４Ａは、ノード状態管理情報Ｔ７から、自パーティ内の他ノード１０ＡのＩＰアドレス及びポート番号を取得し、ノード状態情報Ｔ６の宛先ノードを判定する。

また、送信処理部１４Ａは、ノード状態管理情報Ｔ７を参照して、自ノード１０Ａが判定した各ノード１０ＡについてのノードＩＤ、状態、ＩＰアドレス、及びポート番号の情報からノード状態情報Ｔ６を生成する。そして、送信処理部１４Ａは、生成したノード状態情報Ｔ６を、ハートビートとして自パーティ内の他ノード１０Ａの各々へ送信する。

【０１６９】

また、送信処理部１４Ａは、ノード状態情報Ｔ６の送信に加え、上述のように、自ノード１０Ａの起動後、送信情報Ｔ３（図６参照）をストレージシステム１内の全てのノード１０Ａへブロードキャスト等により通知する。

なお、パーティ間受信処理部１０２が受信する代表ノード状態情報Ｔ５及びパーティ間送信処理部１０４が送信する代表ノード状態情報Ｔ５は、同様のデータ構造である。また、受信処理部１２Ａが受信するノード状態情報Ｔ６及び送信処理部１４Ａが送信するノード状態情報Ｔ６は、同様のデータ構造である。以下、便宜上、パーティ間送信処理部１０４が送信する代表ノード状態情報Ｔ５を送信用代表ノード状態情報（送信用代表状態情報）Ｔ５といい、送信処理部１４Ａが送信するノード状態情報Ｔ６を送信用ノード状態情報（送信用状態情報）Ｔ６という場合がある。

【０１７０】

〔２－２－５〕パーティ管理部

次に、図２０～図２３を参照して、パーティ管理部１０５について説明する。

図２０は、第２実施形態の一例としてのストレージシステム１にノード１０Ａが追加される例を示す図であり、図２１は、図２０に示すストレージシステム１におけるパーティの分割処理の一例を説明する図である。図２２は、図２１に示すストレージシステム１におけるノード１０Ａの削除処理及びパーティの統合処理の一例を説明する図である。図２３は、第２実施形態の一例としてのストレージシステム１におけるパーティの分割処理の具体例を説明する図である。

【０１７１】

なお、図２０～図２２に示す例においては、説明の簡略化のため、ノード１０Ａ間の接続状態のみを示し、スイッチ２０の図示を省略している。

パーティ管理部（管理部）１０５は、自ノード１０Ａが属するパーティに関する管理を行なう。

具体的には、パーティ管理部１０５は、自パーティにおけるノード１０Ａの追加又は削除により、自パーティに属するノード１０Ａの数が所定の上限又は所定の下限を超えた場合、自パーティの分割又は統合を行なう。

【０１７２】

例えば、ストレージシステム１の運用が開始されたとき等の初期状態において、パーティが１つ又は複数ある場合、ストレージシステム１の運用に応じてパーティにノード１０Ａが追加される場合がある。ノード１０Ａの追加により、パーティを構成するノード１０Ａの数が多くなると、パーティ内でのハートビートの通信によりノード１０Ａの処理負荷及びネットワークの負荷が高まり、ストレージシステム１の性能が低下する可能性がある。

【０１７３】

そこで、パーティ管理部１０５は、自パーティにおけるノード１０Ａの数が予め決められた上限（第４所定値）を上回った場合、自パーティから複数のノード１０Ａを分割し、新たなパーティを作成する。

また、逆に、パーティ管理部１０５は、パーティを構成するノード数が下限（第５所定値）を下回った場合、パーティを統合する。パーティを統合する理由は、少数のノード１０Ａを含むパーティが多数あると、代表ノード１０Ａ間のハートビートの通信による代表ノード１０Ａの処理負荷及びネットワーク負荷が高まるためである。また、パーティ管理

情報Ｔ４が肥大化し、パーティの管理に係る処理負荷が増大することも理由の一つである。

【０１７４】

なお、予め定められた上限及び下限としては、ストレージシステム１の規模やポリシー等によって異なるが、例えば上限を数十～数百台程度とし、下限を数～数十台程度とすることができる。以下、説明の簡略化のため、上限を５台、下限を２台として説明する。

パーティ管理部１０５によるパーティの分割又は統合に伴うパーティ管理情報Ｔ４の変更は、各パーティに所属する代表ノード１０Ａが、自パーティのエントリについて行なうことができる。代表ノード１０Ａがそなえるパーティ管理部１０５は、パーティ管理情報Ｔ４を変更すると、パーティ間送信処理部１０４を介して、ハートビートに載せて全ノード１０Ａへ伝達する。

【０１７５】

なお、パーティ管理情報Ｔ４は、ブロードキャスト等により全ノード１０Ａへ伝達されてもよいし、代表ノード状態情報Ｔ５とともにハートビートとして各代表ノード１０Ａへ伝達されてもよい。パーティ管理情報Ｔ４が各代表ノード１０Ａへ伝達される場合、パーティ管理情報Ｔ４を受け取った代表ノード１０Ａは、自パーティのメンバノード１０Ａへ転送することが好ましい。

【０１７６】

以下、パーティ管理部１０５によるパーティの分割処理及び統合処理について説明する。

図２０の紙面左上に示すように、ストレージシステム１が、パーティＡ及びＢをそなえる場合を例に挙げて説明する。なお、パーティＡは、ノードＩＤ“１”、“３”、“５”、“７”、及び“９”の５つのノード１０Ａを有し、パーティＢは、ノードＩＤ“１１”、“１３”、“１５”、“１７”、及び“１９”の５つのノード１０Ａをそなえるものとする。また、図２０の紙面右側に示すように、パーティ管理情報Ｔ４には、パーティＩＤ“Ａ”にノードＩＤ“１～１０”が、パーティＩＤ“Ｂ”にノードＩＤ“１１～２０”が、それぞれ対応付けられているものとする。

【０１７７】

なお、パーティＡ及びＢの代表ノード１０Ａは、それぞれノードＩＤ“１”及び“１１”のノード１０Ａ（以下、代表ノード１０Ａ－１及び１０Ａ－１１という）である。

以上の例において、パーティＡにノードＩＤ“８”のノード１０Ａが追加される場合を想定する（図２０の紙面左下及び図２１の紙面左上参照）。この場合、パーティＡには、６つのノードが含まれる。なお、ノードＩＤ“８”は、パーティＡに対応付けられたノードＩＤの範囲内であるため、パーティ管理情報Ｔ４に変更はない。

【０１７８】

代表ノード１０Ａ－１がそなえるパーティ管理部１０５は、自パーティＡに属するノード１０Ａの数が上限である５つを超えるため、パーティＡを分割する。

図２１の紙面左下に示すように、代表ノード１０Ａ－１のパーティ管理部１０５は、パーティ管理情報Ｔ４及びノード状態管理情報Ｔ７を参照して、パーティＡをノード数が２分の１になるように分割する。例えば、パーティ管理部１０５は、パーティＡに属するノードＩＤのうち、ノードＩＤが小さい順に３つのノード１０Ａをパーティ Ａに残し、それ以外の３つのノードをパーティＣとして分割する。つまり、パーティ管理部１０５は、パーティＡを、ノードＩＤ“１”、“３”、及び“５”をそなえる新たなパーティＡと、ノードＩＤ“７”～“９”をそなえるパーティＣとに分割する。

【０１７９】

なお、パーティ管理部１０５は、パーティの分割において、ノード数を２分の１にする際に余りが出る場合、余りのノード１０Ａを分割に係る２つのパーティのいずれかに割り振る。

代表ノード１０Ａ－１のパーティ管理部１０５は、パーティＡを分割すると、パーティ管理情報Ｔ４のパーティＩＤ“Ａ”のエントリにおいて、ノードＩＤを“１～５”に設定

し、バージョン番号を“２”に変更する。また、代表ノード１０Ａ－１のパーティ管理部１０５は、パーティ管理情報Ｔ４にパーティＩＤ“Ｃ”のエントリを追加し、ノードＩＤ“６～１０”、バージョン番号“１”を対応付ける。

【０１８０】

そして、代表ノード１０Ａ－１のパーティ管理部１０５は、変更したパーティ管理情報Ｔ４′を、パーティ間送信処理部１０４を介して全ノード１０Ａへ通知する。

なお、ノード１０Ａ－１は、新たなパーティＡにおいて、引き続き代表ノード１０Ａとしてパーティ管理情報Ｔ４のパーティＩＤ“Ａ”のエントリの管理を行なう。一方、パーティＣでは、ノードＩＤ“７”～“９”のノード１０Ａで、代表ノード１０Ａを選出する所定のルールが適用され、例えばノードＩＤ“７”のノード１０Ａ（以下、代表ノード１０Ａ－７という）が、代表ノード１０Ａになる。代表ノード１０Ａ－７は、代表ノード１０Ａ－１、１０Ａ－１１とともに、代表ノード１０Ａ間でのハートビートの通信を行なうとともに、パーティＣのエントリを管理する。

【０１８１】

以上のように、代表ノード１０Ａがそなえるパーティ管理部１０５により、パーティの分割処理が行なわれる。

次いで、図２２の紙面左上に示すように、パーティＡ内のノードＩＤ“３”及び“５”のノード１０Ａが障害等の発生により停止した場合を想定する。

代表ノード１０Ａ－１がそなえるパーティ管理部１０５は、自パーティＡに属するノード１０Ａの数が、ノード１０Ａの停止に伴い下限である２つを超える（下回る）ため、パーティＡを他のパーティと統合する。

【０１８２】

図２２の紙面左下に示すように、代表ノード１０Ａ－１のパーティ管理部１０５は、パーティ管理情報Ｔ４′及びノード状態管理情報Ｔ７を参照して、パーティＡと統合する他のパーティを決定する。パーティＡと統合する他のパーティとしては、例えばノード数が最も少ないパーティが挙げられる。この場合、代表ノード１０Ａ－１のパーティ管理部１０５は、自パーティＡ以外でノード数が最も少ないパーティＣを統合対象のパーティに決定する。

【０１８３】

代表ノード１０Ａ－１のパーティ管理部１０５は、統合対象のパーティを決定すると、パーティ管理情報Ｔ４′のパーティＩＤ“Ａ”のエントリにおいて、ノードＩＤをパーティＣとマージさせて、“１～１０”に設定し、バージョン番号を“３”に変更する。また、代表ノード１０Ａ－７のパーティ管理部１０５は、パーティ管理情報Ｔ４′からパーティＩＤ“Ｃ”のエントリを削除する。

【０１８４】

そして、代表ノード１０Ａ－１のパーティ管理部１０５は、変更したパーティ管理情報Ｔ４″を、パーティ間送信処理部１０４を介して全ノード１０Ａへ通知する。

なお、ノード１０Ａ－１は、新たなパーティＡにおいて、引き続き代表ノード１０Ａとしてパーティ管理情報Ｔ４のパーティＩＤ“Ａ”のエントリの管理を行なう。一方、パーティＣの代表ノード１０Ａ－７は、新たなパーティＡにおいて代表ノード１０Ａを選出する所定のルールの敗者であるので、メンバノード１０Ａ－７に降格する。

【０１８５】

また、図２２に示す例では、パーティＡとパーティＣとが統合されたため、ノードＩＤも“１～５”と“６～１０”とがマージされて“１～１０”になった。しかし、パーティ管理情報Ｔ４の状態によっては、統合する２つのパーティのノードＩＤの範囲が離れ、間に存在するノードＩＤが他のパーティを構成する場合も考えられる。このような場合、統合後のパーティに属するノードＩＤは、１つの範囲ではなく、複数の範囲で又は１つずつ設定されてもよい。

【０１８６】

以上のように、代表ノード１０Ａがそなえるパーティ管理部１０５により、パーティの

統合処理が行なわれる。
なお、代表ノード１０Ａのパーティ管理部１０５は、所定時間ごとに、自パーティのノード１０Ａの数が上限に達したか否か、及び下限を下回ったか否かを判定することができる。
【０１８７】
また、代表ノード１０Ａのパーティ管理部１０５は、ストレージシステム１に追加された新規ノード１０Ａから送信される送信情報Ｔ３を受信したことを契機に、自パーティのノード１０Ａの数が上限に達したか否かを判定してもよい。
さらに、代表ノード１０Ａのパーティ管理部１０５は、自パーティ内のノード１０Ａに障害等が発生し、当該ノード１０Ａのリカバリ処理が完了したことを契機に、自パーティのノード１０Ａの数が下限を下回ったか否かを判定してもよい。
【０１８８】
上述したパーティ管理部１０５の説明では、パーティ管理部１０５は、パーティの分割及び統合に係るノード１０Ａの選定を、ノードＩＤの値に基づいて行なうものとして説明した。しかし、ストレージシステム１において、ノード１０Ａ間のハートビートの通信は、ノード１０Ａ間の距離に応じたレイテンシやパケットロスの影響を受ける。
そこで、パーティ管理部１０５は、以下で説明するように、パーティの分割及び統合に係るノード１０Ａの選定を、例えばノード１０Ａが接続されるスイッチ２０に基づいて行なうことが好ましい。なお、以下の説明は、管理者等による、ストレージシステム１の運用開始前のパーティの初期設定の際や、運用中においてパーティの構成が複雑化したことによるパーティの再設定の際にも、同様に考慮されることが好ましい。
【０１８９】
一例として、パーティの初期設定の際に、１つのスイッチ２０に接続されるノード１０Ａ群が同じパーティに設定されることが考えられる。図２３の紙面上側に示す例では、スイッチ２０に、ノード１０Ａ－１～１０Ａ－４が接続され、これらノード１０Ａ－１～１０Ａ－４が１つのパーティを構成する。なお、スイッチ２０のポート数は４つであるものとする。
【０１９０】
ストレージシステム１へのノード１０Ａ－５及び１０Ａ－６の追加に伴い、ノード１０Ａの数が１つのスイッチ２０に収まらなくなると、作業者等により、スイッチ２０の増設及びトポロジの調整が行なわれる。例えば、図２３の紙面下側に示すように、スイッチ２０－１にノード１０Ａ－１～１０Ａ－３が接続され、スイッチ２０－２にノード１０Ａ－４～１０Ａ－６が接続される。また、スイッチ２０－１及び２０－２間が接続される。
【０１９１】
代表ノード１０Ａのパーティ管理部１０５は、ノード１０Ａ－５及び１０Ａ－６が追加されると（ノード１０Ａが図２３の紙面下側に示す接続状態になると）、ノード１０Ａ及びスイッチ２０の接続関係に関する情報を取得する。例えば、パーティ管理部１０５は、スイッチ２０が保持する各ポートの接続先の情報等を取得することで、ノード１０Ａ及びスイッチ２０の接続関係に関する情報を取得（推定）することができる。なお、スイッチ２０からの接続先の情報等の取得は、既知の種々の手法により行なうことが可能であり、その詳細な説明は省略する。また、パーティ管理部１０５は、作業者等から、入出力部１０ｅを介してノード１０Ａ及びスイッチ２０の接続関係に関する情報を入力されてもよい。
【０１９２】
そして、パーティ管理部１０５は、取得したノード１０Ａ及びスイッチ２０の接続関係から、例えば、パーティを、スイッチ２０－１に接続されるノード１０Ａ群と、スイッチ２０－２に接続されるノード１０Ａ群とに分割する。
このように、パーティ管理部１０５は、自パーティにおけるノード１０Ａ及びスイッチ２０の物理的な接続関係に関する情報に基づいて、パーティから分割するノード１０Ａを決定することができる。

【０１９３】
なお、パーティ管理部１０５は、ノード１０Ａ及びスイッチ２０の接続関係に関する情報として、代表ノード１０Ａからパーティ内の他ノード１０Ａの各々までのホップ数を検出してもよい。これは、ホップ数が近いノード１０Ａ同士は、同じスイッチ２０に接続されている可能性が高いと推測できるからである。
ここまで、図２３を参照してパーティ管理部１０５によるパーティの分割処理について説明したが、パーティ管理部１０５によるパーティの統合処理についても同様である。
【０１９４】
すなわち、パーティ管理部１０５は、自パーティと統合する他のパーティとして、ノード数が少ないパーティを選択するのではなく、ノード１０Ａ及びスイッチ２０の接続関係に基づき選択してもよい。
〔２－３〕動作例
次に、図２４～図２６を参照して、上述の如く構成された第２実施形態の一例としてのノード１０Ａによる動作例を説明する。図２４は、第２実施形態の一例としての代表ノード１０Ａによる他の代表ノード１０Ａの状態を判定する動作例を説明するフローチャートである。図２５は、ノード１０Ａによるパーティ内の他ノード１０Ａが停止した場合の動作例を説明するフローチャートである。図２６は、ノード１０Ａによるパーティの分割処理及び統合処理の動作例を説明するフローチャートである。
【０１９５】
〔２－３－１〕代表ノードによる他の代表ノードの状態を判定する動作例
はじめに、図２４を参照して、代表ノード１０Ａによる他の代表ノード１０Ａの状態を判定する動作例を説明する。
なお、図２４に示すステップＳ４１～Ｓ５５の処理は、代表ノード１０Ａの各々において、パーティ間ノード状態決定部１０３により他の一の代表ノード１０Ａの状態が判定される際に行なわれる処理である。従って、ステップＳ４１～Ｓ５５の処理は、各代表ノード１０Ａのパーティ間ノード状態決定部１０３により、他の代表ノード１０Ａの各々について、定期的（第１所定時間ごと）に実行される。
【０１９６】
また、図２４に示すステップＳ４１～Ｓ４９、Ｓ５２、及びＳ５３の処理は、図１２に示すステップＳ１１～Ｓ１９、Ｓ２２、及びＳ２３の処理と比較して、判定対象のノード１０（１０Ａ）が代表ノード１０Ａである点が異なる。以下、ステップＳ４１～Ｓ４９、Ｓ５２、及びＳ５３の処理の説明において、図１２に示すステップＳ１１～Ｓ１９、Ｓ２２、及びＳ２３の処理と同様な部分の詳細は省略する。
【０１９７】
図２４に示すように、パーティ間ノード状態決定部１０３により、ノード状態管理情報Ｔ７内の“状態”が参照され、判定対象の代表ノード１０Ａについて直前に判定した状態がどの状態であるかが判定される（ステップＳ４１，Ｓ４６，Ｓ４９）。
判定対象の代表ノード１０Ａについて直前に判定した状態がＡｌｉｖｅである場合（ステップＳ４１のＹｅｓルート）、処理がステップＳ４２に移行する。ステップＳ４２では、パーティ間ノード状態決定部１０３により、判定対象の代表ノード１０Ａからのハートビートの不達時間が閾値を超えたか否かが判定される。
【０１９８】
ハートビートの不達時間が閾値を超えた場合（ステップＳ４２のＹｅｓルート）、パーティ間ノード状態決定部１０３により、判定対象の代表ノード１０Ａの状態がＳｕｓｐｅｃｔと判定され（ステップＳ４３）、処理が終了する。このとき、パーティ間ノード状態決定部１０３は、判定対象の代表ノード１０Ａについて、ノード状態管理情報Ｔ７内の“状態”にＳｕｓｐｅｃｔを設定する。そして、パーティ間ノード状態決定部１０３は、次の判定対象の代表ノード１０Ａがある場合、次の判定対象の代表ノード１０Ａに係る状態の判定処理に移行する。
【０１９９】

一方、ステップＳ４２において、パーティ間ノード状態決定部１０３により、ハートビートの不達時間が閾値を超えていないと判定された場合（ステップＳ４２のＮｏルート）、処理がステップＳ４４に移行する。ステップＳ４４では、パーティ間ノード状態決定部１０３により、判定対象の代表ノード１０Ａの状態が、過半数（第１所定値）の代表ノード１０ＡからＳｕｓｐｅｃｔと判定されたか否かが判定される。又は、パーティ間ノード状態決定部１０３により、判定対象の代表ノード１０Ａの状態が、複数の代表ノード１０Ａのいずれかの代表ノード１０ＡによりＤｏｗｎと判定された否かが判定される。

【０２００】

判定対象の代表ノード１０Ａの状態が、過半数の代表ノード１０ＡからＳｕｓｐｅｃｔと判定されておらず、いずれかの代表ノード１０ＡによりＤｏｗｎとも判定されていない場合（ステップＳ４４のＮｏルート）、代表ノード１０Ａに対する処理が終了する。一方、判定対象の代表ノード１０Ａの状態が、過半数の代表ノード１０ＡからＳｕｓｐｅｃｔと判定された又はいずれかの代表ノード１０ＡによりＤｏｗｎと判定された場合（ステップＳ４４のＹｅｓルート）、処理がステップＳ４５に移行する。

【０２０１】

ステップＳ４５では、パーティ間ノード状態決定部１０３により、判定対象の代表ノード１０Ａの状態がＤｏｗｎと判定され、処理が終了する。このとき、パーティ間ノード状態決定部１０３は、判定対象の代表ノード１０Ａについて、ノード状態管理情報Ｔ７内の“状態”にＤｏｗｎを設定する。

また、判定対象の代表ノード１０Ａについて直前に判定した状態がＳｕｓｐｅｃｔである場合（ステップＳ４１のＮｏルートからステップＳ４６のＹｅｓルート）、処理がステップＳ４７に移行する。ステップＳ４７では、パーティ間ノード状態決定部１０３により、判定対象の代表ノード１０Ａから新たなハートビートが受信されたか否かが判定される。

【０２０２】

新たなハートビートが受信されていない場合（ステップＳ４７のＮｏルート）、処理がステップＳ４４に移行する。一方、新たなハートビートが受信された場合（ステップＳ４７のＹｅｓルート）、パーティ間ノード状態決定部１０３により、判定対象の代表ノード１０Ａの状態がＡｌｉｖｅと判定され（ステップＳ４８）、処理が終了する。このとき、パーティ間ノード状態決定部１０３は、判定対象の代表ノード１０Ａについて、ノード状態管理情報Ｔ７内の“状態”にＡｌｉｖｅを設定する。

【０２０３】

判定対象の代表ノード１０Ａについて直前に判定した状態がＤｏｗｎである場合（ステップＳ４１のＮｏルート，ステップＳ４６のＮｏルートからステップＳ４９のＹｅｓルート）、処理がステップＳ５０に移行する。ステップＳ５０では、パーティ間ノード状態決定部１０３により、判定対象の代表ノード１０Ａの状態が第２所定値の数の代表ノード１０Ａ（例えば全代表ノード１０Ａ）によりＤｏｗｎと判定されたか否かが判定される。

【０２０４】

全代表ノード１０ＡによりＤｏｗｎと判定されていない場合（ステップＳ５０のＮｏルート）、判定対象の代表ノード１０Ａに対する処理が終了する。一方、全代表ノード１０ＡによりＤｏｗｎと判定された場合（ステップＳ５０のＹｅｓルート）、処理がステップＳ５４に移行する。ステップＳ５４では、ノード状態決定部１３により、該当パーティに他ノード１０Ａが生存しているか否か、つまり該当パーティにＡｌｉｖｅと判定された他ノード１０Ａが存在するか否かが判定される。

【０２０５】

ステップＳ５４において、該当パーティに生存している他ノード１０Ａが存在しないと判定された場合（ステップＳ５４のＮｏルート）、処理がステップＳ５１に移行する。ステップＳ５１では、パーティ間ノード状態決定部１０３により、該当パーティに所属する全ノード１０Ａの状態がＺｏｍｂｉｅと判定される。また、自ノード１０Ａが保持するデータが該当パーティに所属するいずれかのノード１０Ａが保持するデータに関連する場合

、リカバリ処理部１５により、該当ノード１０Ａに対するリカバリ処理が実行され、処理が終了する。このとき、ノード状態決定部１３は、該当パーティに所属する全ノード１０Ａについて、ノード状態管理情報Ｔ７内の“状態”にＺｏｍｂｉｅを設定する。

【０２０６】

一方、ステップＳ５４において、該当パーティに生存している他ノード１０Ａが存在すると判定された場合（ステップＳ５４のＹｅｓルート）、処理がステップＳ５５に移行する。ステップＳ５５では、パーティ間ノード状態決定部１０３により、生存している他ノード１０Ａのうちの次点のノード１０Ａが新たな判定対象の代表ノード１０Ａと判定され、処理が終了する。なお、この判定は、代表ノード１０Ａを選定する所定のルール（例えばノードＩＤが最も小さいノード１０Ａ）に基づいて行なわれる。代表ノード１０Ａは、ステップＳ５５において選定した新たな判定対象の代表ノード１０Ａに対して、次回以降のハートビートの通信を行なう。

【０２０７】

判定対象の代表ノード１０Ａについて直前に判定した状態がＺｏｍｂｉｅである場合（ステップＳ４１のＮｏルート，ステップＳ４６のＮｏルートからステップＳ４９のＮｏルート）、処理がステップＳ５２に移行する。ステップＳ５２では、パーティ間ノード状態決定部１０３により、判定対象の代表ノード１０Ａについてリカバリ処理が完了したか否かが判定される。リカバリ処理が完了していない場合、判定対象の代表ノード１０Ａに対する処理が終了する。一方、リカバリ処理が完了した場合（ステップＳ５２のＹｅｓルート）、パーティ間ノード状態決定部１０３により、ノード状態管理情報Ｔ７から、判定対象の代表ノード１０Ａに関する情報が削除され（ステップＳ５３）、処理が終了する。

【０２０８】

以上のように、代表ノード１０Ａにより、他の一の代表ノード１０Ａの状態の判定処理が行なわれる。

〔２－３－２〕ノードによるパーティ内の他ノードが停止した場合の動作例

次に、図２５を参照して、ノード１０Ａによるパーティ内の他ノード１０Ａが停止した場合の動作例を説明する。

【０２０９】

なお、図２５に示すステップＳ６１～Ｓ６３の処理は、メンバノード１０Ａの各々において、ノード状態決定部１３Ａにより自パーティの代表ノード１０Ａの状態が判定される際に行なわれる処理である。従って、ステップＳ６１～Ｓ６３の処理は、メンバノード１０Ａのノード状態決定部１３Ａにより、自パーティの代表ノード１０Ａについて、定期的（第１所定時間ごと）に実行される。

【０２１０】

図２５に示すように、メンバノード１０Ａによる自パーティ内の他ノード１０Ａの状態の判定により、自パーティの代表ノード１０Ａが停止したか否かが判定される（ステップＳ６１）。

自パーティの代表ノード１０Ａが停止していない場合（ステップＳ６１のＮｏルート）、処理が終了する。ノード状態決定部１３Ａは、次の判定対象のメンバノード１０Ａがある場合、次の判定対象のメンバノード１０Ａに係る状態の判定処理に移行する。

【０２１１】

一方、自パーティの代表ノード１０Ａが停止した場合（ステップＳ６１のＹｅｓルート）、自ノード１０Ａが代表ノード１０Ａになるか否かが判定される（ステップＳ６２）。なお、この判定は、代表ノード１０Ａを選定する所定のルールに基づいて行なわれる。

ノード１０Ａにより、ステップＳ６２において自ノード１０Ａが代表ノード１０Ａになると判定された場合（ステップＳ６２のＹｅｓルート）、他のパーティの代表ノード１０Ａの各々との間のハートビートの通信が開始され（ステップＳ６３）、処理が終了する。

【０２１２】

一方、ノード１０Ａにより、ステップＳ６２において自ノード１０Ａが代表ノード１０Ａにならないと判定された場合（ステップＳ６２のＮｏルート）、処理が終了する。

以上のように、ノード１０Ａによるパーティ内の他ノード１０Ａが停止した場合の処理が終了する。

〔２－３－３〕ノードによるパーティの分割処理及び統合処理の動作例

次に、図２６を参照して、ノード１０Ａによるパーティの分割処理及び統合処理の動作例を説明する。

【０２１３】

図２６に示すように、代表ノード１０Ａのパーティ管理部１０５により、例えば所定時間ごとに、自パーティのノード１０Ａの数が上限を上回ったか否かが判定される（ステップＳ７１）。

ノード１０Ａの数が上限を上回った場合（ステップＳ７１のＹｅｓルート）、代表ノード１０Ａのパーティ管理部１０５により、パーティが２つに分割され、パーティ管理情報Ｔ４が更新される（ステップＳ７２）。なお、パーティ管理部１０５は、上述のように、自パーティをノード数が２分の１になるように分割し、余りが出る場合、余りのノード１０Ａを分割に係る２つのパーティのいずれかに割り振る。また、パーティ管理部１０５は、自パーティのノード１０Ａを分割後のいずれのパーティに割り当てるかを、ノードＩＤ、ノード１０Ａ及びスイッチ２０の接続関係に基づき決定する。

【０２１４】

ステップＳ７１において、ノード１０Ａの数が上限以下である場合（ステップＳ７１のＮｏルート）、パーティ管理部１０５により、自パーティのノード１０Ａの数が下限未満であるか否かが判定される（ステップＳ７３）。

ノード１０Ａの数が下限未満である場合（ステップＳ７３のＹｅｓルート）、代表ノード１０Ａのパーティ管理部１０５により、自パーティと他のパーティとの統合が行なわれる。具体的には、代表ノード１０Ａのパーティ管理部１０５は、自パーティと統合する他のパーティを、ノードＩＤ、ノード１０Ａ及びスイッチ２０の接続関係、又は他のパーティの代表ノード１０Ａまでのホップ数等に基づき決定する。

【０２１５】

そして、代表ノード１０Ａのパーティ管理部１０５により、統合後に自ノード１０Ａが代表ノード１０Ａになるか否かが判定される（ステップＳ７４）。具体的には、代表ノード１０Ａのパーティ管理部１０５は、パーティ管理情報Ｔ４及びノード状態管理情報Ｔ７を参照して、自ノード１０ＡのノードＩＤと決定した他のパーティの代表ノード１０ＡのノードＩＤとを比較する。そして、パーティ管理部１０５は、自ノード１０ＡのノードＩＤが他のパーティの代表ノード１０ＡのノードＩＤよりも小さいか否かを判定する。

【０２１６】

統合後に自ノード１０Ａが代表ノード１０Ａにならないと判定された場合（ステップＳ７４のＮｏルート）、代表ノード１０Ａのパーティ管理部１０５により、パーティ管理情報Ｔ４の自パーティのエントリが削除され（ステップＳ７５）、処理が終了する。

一方、統合後に自ノード１０Ａが代表ノード１０Ａになると判定された場合（ステップＳ７４のＹｅｓルート）、代表ノード１０Ａのパーティ管理部１０５により、自パーティと他のパーティとが統合される。具体的には、代表ノード１０Ａのパーティ管理部１０５により、パーティ管理情報Ｔ４の自パーティのエントリのノードＩＤに、統合する他のパーティのノードＩＤがマージされて、パーティ管理情報Ｔ４が更新される（ステップＳ７６）。そして、パーティ管理部１０５による処理が終了する。

【０２１７】

なお、上述のように、ステップＳ７１の処理は、ストレージシステム１に追加された新規ノード１０Ａから送信される送信情報Ｔ３を受信したことを契機に開始されてもよい。

また、ステップＳ７３の処理は、自パーティ内のノード１０Ａに障害等が発生し、当該ノード１０Ａのリカバリ処理が完了したことを契機に開始されてもよい。

さらに、ステップＳ７１及びＳ７２の処理と、ステップＳ７３～Ｓ７６の処理とは、互いに独立して実行されてもよいし、処理順序が変更されてもよい。

【０２１８】

〔２－４〕第２実施形態のまとめ

上述のように、第２実施形態の一例としてのノード１０Ａによれば、第１実施形態に係るノード１０と同様の効果を奏することができる。

また、第２実施形態の一例としてのノード１０Ａによれば、複数のノード１０Ａが複数のパーティに分割される。そして、各パーティの代表ノード１０Ａの各々において、パーティ間受信処理部１０２は、他のパーティの各々における他の代表ノード１０Ａから、代表ノード状態情報Ｔ５を受信する。また、パーティ間ノード状態決定部１０３は、代表ノード状態情報Ｔ５に基づいて、複数の代表ノード１０Ａの各々の状態を判定する。さらに、パーティ間送信処理部１０４は、パーティ間ノード状態決定部１０３が判定した複数の代表ノード１０Ａの各々の状態に関する送信用代表ノード状態情報Ｔ５を、他の代表ノード１０Ａの各々へ送信する。

【０２１９】

さらに、複数のノード１０Ａの各々において、送信処理部１４Ａは、送信用ノード状態情報Ｔ６を、自パーティにおける他ノード１０Ａの各々へ送信する。また、ノード状態決定部１３Ａは、受信処理部１２Ａが自パーティにおける他ノード１０Ａの各々から受信したノード状態情報Ｔ６に基づいて、自ノード１０Ａにおけるノード１０Ａの各々の状態を判定する。

【０２２０】

これにより、メンバノード１０Ａにより、自パーティ内のノード１０Ａの状態が判定され、代表ノード１０Ａにより、パーティ間（代表ノード１０Ａ間）の状態が判定される。

従って、ストレージシステム１においてノード１０Ａの数が増大した場合でも、ノード間でやり取りされる情報の直接の送信先を絞ることができるため、ハートビートの送受信のコストの増大を抑えることができる。

【０２２１】

つまり、ストレージシステム１は、全ノード１０Ａによる完全メッシュのハートビートの通信を行なうよりも、ストレージシステム１における通信負荷及び処理負荷を低減させることができる。

また、各代表ノード１０Ａにおいて、パーティ管理部１０５は、自パーティにおけるノード１０Ａの数が第４所定値を超えた場合、自パーティから、複数のノード１０Ａを分割して新たなパーティを作成する。

【０２２２】

これにより、パーティ内でのハートビートの通信によるノード１０Ａの処理負荷及びネットワークの負荷に起因した、ストレージシステム１の性能低下を抑止することができる。

さらに、パーティ管理部１０５は、自パーティにおけるノード１０Ａ及びスイッチ２０の接続関係に関する情報に基づいて、自パーティから分割する複数のノード１０Ａを決定する。

【０２２３】

これにより、ノード１０Ａ間の距離に応じたレイテンシやパケットロスの影響を抑止することができる。

また、パーティ管理部１０５は、自パーティにおけるノード１０Ａの数が第５所定値未満の場合、自パーティと他のパーティのうちのいずれかのパーティとを統合する。

これにより、多数の代表ノード１０Ａ間のハートビートの通信による代表ノード１０Ａの処理負荷及びネットワーク負荷に起因した、ストレージシステム１の性能低下を抑止することができる。

【０２２４】

〔３〕その他

以上、本発明の好ましい実施形態について詳述したが、本発明は、係る特定の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内において、種々の変形、変更して実施することができる。

例えば、第１及び第２実施形態に係るストレージシステム１がそなえるノード１０及び１０Ａ、並びにスイッチ２０の構成及び台数は、上述したものに限定されず、任意の構成及び台数とすることができる。

【０２２５】

また、第１及び第２実施形態においては、ストレージシステム１がそなえるノード１０及び１０Ａにおける処理について説明したが、これに限定されるものではない。ノード１０及び１０Ａは、ストレージ装置のほか、情報に対する処理を行なうサーバ等の情報処理装置であってもよく、ストレージシステム１は、複数の情報処理装置をそなえる情報処理システムであってもよい。

【０２２６】

また、第１及び第２実施形態においては、ノード１０及び１０Ａは、例えばストレージシステム１によるサービスの提供に用いられるＩＰラインを介してハートビートを行なうものとして説明したが、これに限定されるものではない。例えば、ノード１０及び１０Ａは、ＬＡＮケーブル等の専用の制御線を介して相互に接続され、専用線を用いてハートビートを行なってもよい。これにより、ＩＰラインのネットワークの負荷を軽減させることができる。なお、ノード１０及び１０Ａは、ＩＰラインを用いる場合、ノード１０及び１０Ａ間のパスの障害検出を行なうことができるため、専用線を用いるよりも監視範囲を拡張することができる。

【０２２７】

さらに、第２実施形態においては、ノード１０Ａは、１段のパーティを構成するものとして説明したが、これに限定されるものではなく、多段のパーティを構成してもよい。つまり、代表ノード１０Ａが数百～数千台等の多数存在する場合、代表ノード１０Ａを複数の上位パーティに分割し、上位パーティ間でハートビートの通信を行なうとともに、各上位パーティ内で、代表ノード１０Ａ間のハートビートの通信を行なってもよい。

【０２２８】

また、第２実施形態においては、全てのノード１０Ａが代表ノード１０Ａになる可能性があったが、これに限定されるものではない。例えば、ノード１０Ａ間で、特定の処理を行なうノード１０Ａ等の処理負荷を増やしたくないノード１０Ａについて、代表ノード１０Ａの候補から除外するＮＧリストを共有してもよい。この場合、各ノード１０Ａは、ＮＧリストに含まれるノード１０Ａについては代表ノード１０Ａに選出しないようにする。

【０２２９】

さらに、第１及び第２実施形態に係るノード１０及び１０Ａがそなえる各機能は、適宜省略してもよく、分割又は統合してもよい。例えば、第２実施形態に係るパーティ間受信処理部１０２及び受信処理部１２Ａを統合し、１つの受信処理部としてもよく、パーティ間送信処理部１０４及び送信処理部１４Ａを統合し、１つの送信処理部としてもよい。また、パーティ間ノード状態決定部１０３及びノード状態決定部１３Ａを統合し、１つのノード状態決定部（判定部）としてもよい。

【０２３０】

また、第２実施形態に係るノード１０Ａは、代表ノード１０Ａとして動作する際に、パーティ間受信処理部１０２、パーティ間ノード状態決定部１０３、パーティ間送信処理部１０４、パーティ管理部１０５の機能を実行する。従って、ノード１０Ａは、代表ノード１０Ａとして動作しない場合（例えば上記ＮＧリストに自ノード１０が登録される場合等）には、これらの機能を無効化又は省略してもよい。

【０２３１】

さらに、第１及び第２実施形態の一例における各処理フローのステップの実行順序を、適宜変更してもよい。

また、第１実施形態に係るノード１０及び第２実施形態に係るノード１０Ａの各種機能の全部もしくは一部は、コンピュータ（ＣＰＵ，情報処理装置，各種端末を含む）が所定のプログラムを実行することによって実現されてもよい。

【０２３２】

そのプログラムは、例えばフレキシブルディスク、ＣＤ、ＤＶＤ、ブルーレイディスク等のコンピュータ読取可能な記録媒体（例えば図２に示す記録媒体１０ｈ）に記録された形態で提供される。なお、ＣＤとしては、ＣＤ－ＲＯＭ、ＣＤ－Ｒ、ＣＤ－ＲＷ等が挙げられる。また、ＤＶＤとしては、ＤＶＤ－ＲＯＭ、ＤＶＤ－ＲＡＭ、ＤＶＤ－Ｒ、ＤＶＤ－ＲＷ、ＤＶＤ＋Ｒ、ＤＶＤ＋ＲＷ等が挙げられる。この場合、コンピュータはその記録媒体から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。

【０２３３】

ここで、コンピュータとは、ハードウェアとＯＳ（Operating System）とを含む概念であり、ＯＳの制御の下で動作するハードウェアを意味している。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。ハードウェアは、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コンピュータプログラムを読み取る手段とをそなえている。上記プログラムは、上述のようなコンピュータに、第１実施形態に係るノード１０又は第２実施形態に係るノード１０Ａの各種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。

【０２３４】

〔４〕付記

以上の第１及び第２実施形態に関し、更に以下の付記を開示する。

（付記１）

相互に接続される複数の情報処理装置を有し、前記複数の情報処理装置間で通信を行なう情報処理システムにおいて、

前記複数の情報処理装置の各々が、

前記複数の情報処理装置のうちの自情報処理装置以外の他の情報処理装置の各々から、前記他の情報処理装置により判定された前記複数の情報処理装置の各々の状態に関する状態情報を受信する受信処理部と、

前記受信処理部が前記他の情報処理装置の各々から受信した前記状態情報に基づいて、前記複数の情報処理装置の各々の状態を判定する判定部と、

前記判定部が判定した前記複数の情報処理装置の各々の状態に関する送信用状態情報を、前記他の情報処理装置の各々へ送信する送信処理部と、を有することを特徴とする、情報処理システム。

【０２３５】

（付記２）

前記判定部は、

前記受信処理部が受信した前記状態情報が示す前記複数の情報処理装置の各々の状態と、前記他の情報処理装置の各々からの前記状態情報の受信状況とに基づいて、前記複数の情報処理装置の各々の状態を判定し、

前記送信処理部は、

第１所定時間ごとに、前記送信用状態情報を、前記他の情報処理装置の各々へ送信することを特徴とする、付記１記載の情報処理システム。

【０２３６】

（付記３）

前記判定部は、

前記受信処理部が受信した前記状態情報が示す前記複数の情報処理装置の各々の状態と、前記判定部が含まれる自情報処理装置の状態に関する状態情報に関する自己状態情報とに基づいて、前記複数の情報処理装置の各々の状態を判定することを特徴とする、付記１記載の情報処理システム。

【０２３７】

（付記４）

前記判定部は、
前記第１所定時間以上の時間である第２所定時間内に前記状態情報を受信しなかった他の情報処理装置の状態を、停止の可能性を示す第１状態と判定し、
前記受信処理部が受信した前記状態情報に基づいて、第１所定数以上の前記複数の情報処理装置で前記第１状態であると判定された情報処理装置の状態、又は、前記他の情報処理装置の少なくとも１つから停止を示す第２状態であると判定された情報処理装置の状態を、前記第２状態と判定することを特徴とする、付記２記載の情報処理システム。
【０２３８】
（付記５）
前記判定部は、
前記受信処理部が受信した前記状態情報に基づいて、前記第１所定数以上の数である第２所定数以上の前記複数の情報処理装置で前記第２状態であると判定された情報処理装置を、リカバリ処理中を示す第３状態と判定し、
前記複数の情報処理装置のうちの１以上の情報処理装置はさらに、
前記判定部が前記第３状態と判定した情報処理装置に対して、リカバリ処理を実行するリカバリ処理部を有することを特徴とする、付記４記載の情報処理システム。
【０２３９】
（付記６）
前記判定部は、
前記自情報処理装置に所定の障害が発生した場合、前記自情報処理装置の状態を前記第２状態と判定し、
前記第２所定時間内に第３所定数以上の前記他の情報処理装置から前記状態情報を受信しなかった場合、前記自情報処理装置の状態を、前記他の情報処理装置から切り離されたことを示す第４状態と判定し、
前記複数の情報処理装置の各々はさらに、
前記自情報処理装置に所定の障害が発生し、前記判定部が前記自情報処理装置の状態を前記第２状態と判定した場合、又は、前記判定部が前記自情報処理装置の状態を前期第４状態と判定した場合、前記自情報処理装置を停止させる処理を行なう停止処理部を有することを特徴とする、付記４又は付記５記載の情報処理システム。
【０２４０】
（付記７）
前記複数の情報処理装置が複数のグループに分割され、
前記複数のグループの各々における代表情報処理装置はさらに、
前記複数のグループのうちの自グループ以外の他のグループの各々における他の代表情報処理装置から、前記他の代表情報処理装置により判定された前記複数のグループの代表情報処理装置の各々の状態に関する代表状態情報を受信するグループ間受信処理部と、
前記グループ間受信処理部が前記他の代表情報処理装置の各々から受信した前記代表状態情報に基づいて、前記複数の代表情報処理装置の各々の状態を判定するグループ間判定部と、
前記グループ間判定部が判定した前記複数の代表情報処理装置の各々の状態に関する送信用代表状態情報を、前記他の代表情報処理装置の各々へ送信するグループ間送信処理部と、を有し、
前記複数の情報処理装置の各々において、
前記送信処理部は、
前記送信用状態情報を、前記自グループにおける他の情報処理装置の各々へ送信し、
前記判定部は、
前記受信処理部が前記自グループにおける他の情報処理装置の各々から受信した前記状態情報に基づいて、前記自グループにおける情報処理装置の各々の状態を判定することを特徴とする、付記１～６のいずれか１項記載の情報処理システム。
【０２４１】

（付記８）
前記複数のグループの各々における代表情報処理装置はさらに、
前記自グループにおける情報処理装置の数が第４所定値を超えた場合、前記自グループから、複数の情報処理装置を分割して新たなグループを作成する管理部を有することを特徴とする、付記７記載の情報処理システム。
【０２４２】
（付記９）
前記情報処理システムはさらに、
前記複数の情報処理装置間に介設され、前記複数の情報処理装置間で送受信される情報を中継する接続装置を有し、
前記管理部は、
前記自グループにおける情報処理装置及び前記接続装置の接続関係に関する情報に基づいて、前記自グループから分割する複数の情報処理装置を決定することを特徴とする、付記８記載の情報処理システム。
【０２４３】
（付記１０）
前記管理部は、
前記自グループにおける情報処理装置の数が第５所定値未満の場合、前記自グループと前記他のグループのうちのいずれかのグループとを統合することを特徴とする、付記８又は付記９記載の情報処理システム。
【０２４４】
（付記１１）
相互に接続される複数の情報処理装置の各々において、
前記複数の情報処理装置のうちの自情報処理装置以外の他の情報処理装置の各々から、前記他の情報処理装置により判定された前記複数の情報処理装置の各々の状態に関する状態情報を受信する受信処理部と、
前記受信処理部が前記他の情報処理装置の各々から受信した前記状態情報に基づいて、前記複数の情報処理装置の各々の状態を判定する判定部と、
前記判定部が判定した前記複数の情報処理装置の各々の状態に関する送信用状態情報を、前記他の情報処理装置の各々へ送信する送信処理部と、を有することを特徴とする、情報処理装置。
【０２４５】
（付記１２）
相互に接続される複数の情報処理装置の各々を制御する情報処理装置の制御プログラムにおいて、
前記情報処理装置に、
前記複数の情報処理装置のうちの自情報処理装置以外の他の情報処理装置の各々から、前記他の情報処理装置により判定された前記複数の情報処理装置の各々の状態に関する状態情報を受信させ、
前記他の情報処理装置の各々から受信した前記状態情報に基づいて、前記複数の情報処理装置の各々の状態を判定させ、
判定した前記複数の情報処理装置の各々の状態に関する送信用状態情報を、前記他の情報処理装置の各々へ送信させることを特徴とする、情報処理装置の制御プログラム。
【０２４６】
（付記１３）
前記情報処理装置に判定させる処理は、
受信した前記状態情報が示す前記複数の情報処理装置の各々の状態と、前記他の情報処理装置の各々からの前記状態情報の受信状況とに基づいて行なわれ、
前記情報処理装置に送信させる処理は、
第１所定時間ごとに行なわれることを特徴とする、付記１２記載の情報処理装置の制御

プログラム。
【０２４７】
（付記１４）
前記情報処理装置に判定させる処理は、
受信した前記状態情報が示す前記複数の情報処理装置の各々の状態と、自情報処理装置の状態に関する状態情報に関する自己状態情報とに基づいて、前記複数の情報処理装置の各々の状態を判定させることを特徴とする、付記１２記載の情報処理装置の制御プログラム。
【０２４８】
（付記１５）
前記情報処理装置に、
前記第１所定時間以上の時間である第２所定時間内に前記状態情報を受信しなかった他の情報処理装置の状態を、停止の可能性を示す第１状態と判定させ、
受信した前記状態情報に基づいて、第１所定数以上の前記複数の情報処理装置で前記第１状態であると判定された情報処理装置の状態、又は、前記他の情報処理装置の少なくとも１つから停止を示す第２状態であると判定された情報処理装置の状態を、前記第２状態と判定させることを特徴とする、付記１３記載の情報処理装置の制御プログラム。
【０２４９】
（付記１６）
前記情報処理装置に、
受信した前記状態情報に基づいて、前記第１所定数以上の数である第２所定数以上の前記複数の情報処理装置で前記第２状態であると判定された情報処理装置を、リカバリ処理中を示す第３状態と判定させ、
前記第３状態と判定した情報処理装置に対して、リカバリ処理を実行させることを特徴とする、付記１５記載の情報処理装置の制御プログラム。
【０２５０】
（付記１７）
前記複数の情報処理装置が複数のグループに分割された前記情報処理システムにおける前記情報処理装置に、
前記複数のグループのうちの自グループ以外の他のグループの各々における他の代表情報処理装置から、前記他の代表情報処理装置により判定された前記複数のグループの代表情報処理装置の各々の状態に関する代表状態情報を受信させ、
前記他の代表情報処理装置の各々から受信した前記代表状態情報に基づいて、前記複数の代表情報処理装置の各々の状態を判定させ、
判定した前記複数の代表情報処理装置の各々の状態に関する送信用代表状態情報を、前記他の代表情報処理装置の各々へ送信させ、
前記情報処理装置に前記送信用状態情報を送信させる処理は、
前記送信用状態情報を、前記自グループにおける他の情報処理装置の各々へ送信させ、
前記情報処理装置に前記複数の情報処理装置の各々の状態を判定させる処理は、
前記自グループにおける他の情報処理装置の各々から受信した前記状態情報に基づいて、前記自グループにおける情報処理装置の各々の状態を判定させることを特徴とする、付記１２～１６のいずれか１項記載の情報処理装置の制御プログラム。
【０２５１】
（付記１８）
前記情報処理装置に、
前記自グループにおける情報処理装置の数が第４所定値を超えた場合、前記自グループから、複数の情報処理装置を分割して新たなグループを作成させることを特徴とする、付記１７記載の情報処理装置の制御プログラム。
【０２５２】
（付記１９）

前記自グループにおける情報処理装置と、前記複数の情報処理装置間に介設され前記複数の情報処理装置間で送受信される情報を中継する接続装置との接続関係に関する情報に基づいて、前記自グループから分割する複数の情報処理装置を決定させることを特徴とする、付記１８記載の情報処理装置の制御プログラム。

【０２５３】

（付記２０）

前記情報処理装置に、

前記自グループにおける情報処理装置の数が第５所定値未満の場合、前記自グループと前記他のグループのうちのいずれかのグループとを統合させることを特徴とする、付記１８又は付記１９記載の情報処理装置の制御プログラム。

【０２５４】

（付記２１）

相互に接続される複数の情報処理装置を有し、前記複数の情報処理装置間で通信を行なう情報処理システムの制御方法において、

前記複数の情報処理装置の各々が、

前記複数の情報処理装置のうちの自情報処理装置以外の他の情報処理装置の各々から、前記他の情報処理装置により判定された前記複数の情報処理装置の各々の状態に関する状態情報を受信し、

前記他の情報処理装置の各々から受信した前記状態情報に基づいて、前記複数の情報処理装置の各々の状態を判定し、

判定した前記複数の情報処理装置の各々の状態に関する送信用状態情報を、前記他の情報処理装置の各々へ送信することを特徴とする、情報処理システムの制御方法。

【０２５５】

（付記２２）

相互に接続される複数の情報処理装置の各々において、

プロセッサを有し、

前記プロセッサが、

前記複数の情報処理装置のうちの自情報処理装置以外の他の情報処理装置の各々から、前記他の情報処理装置により判定された前記複数の情報処理装置の各々の状態に関する状態情報を受信し、

前記他の情報処理装置の各々から受信した前記状態情報に基づいて、前記複数の情報処理装置の各々の状態を判定し、

判定した前記複数の情報処理装置の各々の状態に関する送信用状態情報を、前記他の情報処理装置の各々へ送信することを特徴とする、情報処理装置。

【符号の説明】

【０２５６】

１　ストレージシステム（情報処理システム）

１０，１０－１～１０－５，１０Ａ，１０Ａ－１～１０Ａ－６，１０Ａ－１１～１０Ａ－１３，１０Ａ－２１～１０Ａ－２３　ノード（ストレージ装置，情報処理装置）

１０ａ　ＣＰＵ（プロセッサ）

１０ｂ　メモリ

１０ｃ　記憶部

１０ｄ　ネットワークインタフェース

１０ｅ　入出力部

１０ｆ，１０ｈ　記録媒体

１０ｇ　読取部

１１，１１Ａ　ノード状態保持部

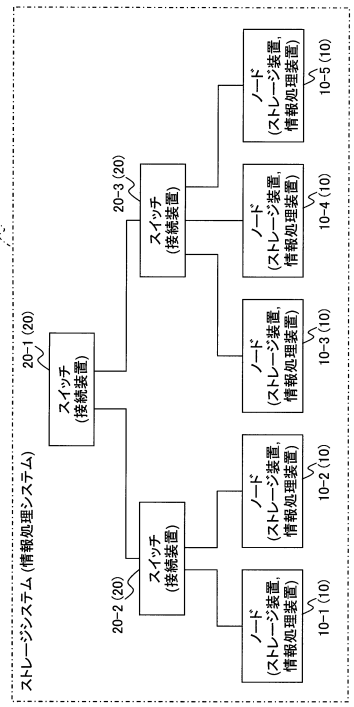
１２，１２Ａ　受信処理部

１３，１３Ａ　ノード状態決定部（判定部）

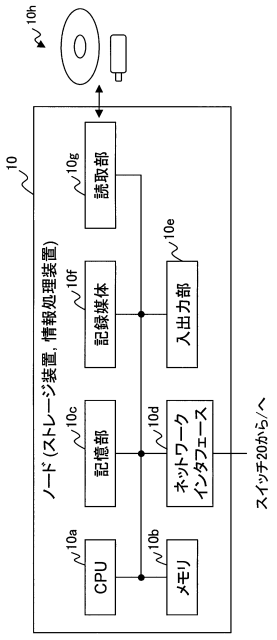
１４，１４Ａ　送信処理部

１５　　リカバリ処理部
１６　　停止処理部
１０１　　パーティ情報保持部
１０２　　パーティ間受信処理部（グループ間受信処理部）
１０３　　パーティ間ノード状態決定部（グループ間判定部）
１０４　　パーティ間送信処理部（グループ間送信処理部）
１０５　　パーティ管理部（管理部）
２０，２０－１～２０－３　　スイッチ（接続装置）

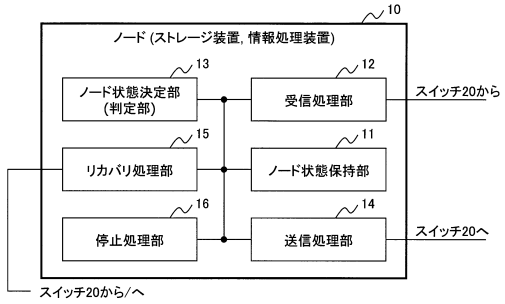
【図１】



【図２】



【図３】



【図４】

ノード状態情報 (状態情報) T1

ノードID	状態	IPアドレス	ポート番号
1	Alive	192.168.0.1	12345
2	Alive	192.168.0.2	12345
3	Alive	192.168.0.3	12345
4	Suspect	192.168.0.4	12345
5	Alive	192.168.0.5	12345

【図５】

ノード状態管理情報 T2

ノードID	状態	by 2	by 3	by 4	by 5	最終更新情報	IPアドレス	ポート番号
1	Alive	Alive	Alive	Alive	Alive	1sec ago	192.168.0.1	12345
2	Alive	Alive	Alive	Alive	Alive	1sec ago	192.168.0.2	12345
3	Alive	Alive	Alive	Alive	Alive	1sec ago	192.168.0.3	12345
4	Suspect	Alive	Suspect	Alive	Suspect	30 sec ago	192.168.0.4	12345
5	Alive	Alive	Alive	Alive	Alive	1sec ago	192.168.0.5	12345

自ノードでの判断
他ノードからの情報

【図６】

送信情報 T3

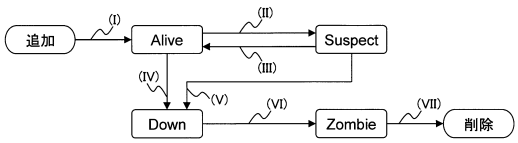
ノードID	状態	IPアドレス	ポート番号
6	Alive	192.168.0.6	12345

【図７】

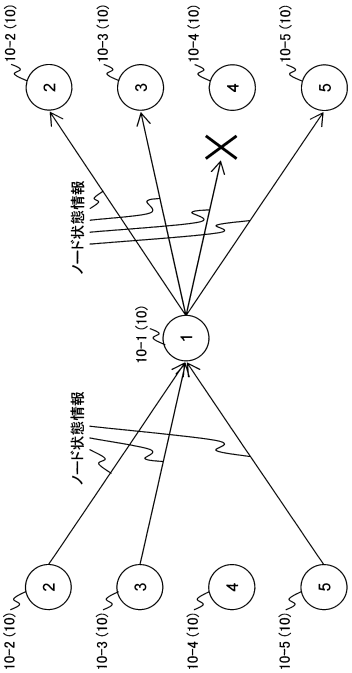
ノード状態情報 (状態情報) T1′

ノードID	状態	IPアドレス	ポート番号
1	Alive	192.168.0.1	12345
2	Alive	192.168.0.2	12345
3	Alive	192.168.0.3	12345
4	Suspect	192.168.0.4	12345
5	Alive	192.168.0.5	12345
6	Alive	192.168.0.6	12345

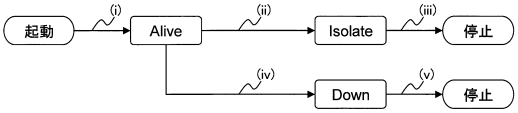
【図８】



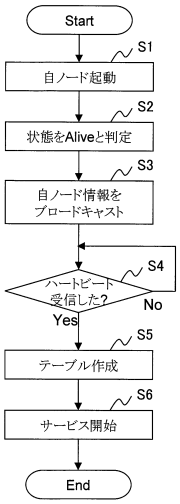
【図９】



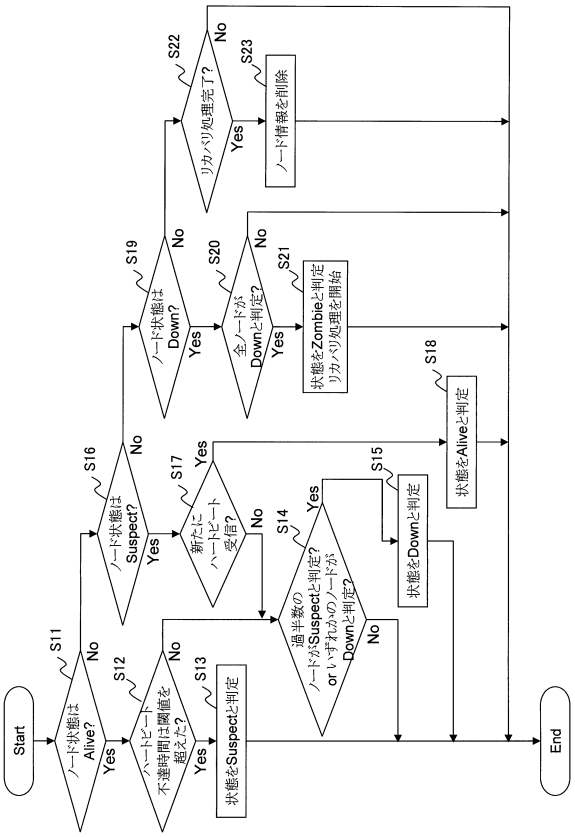
【図１０】



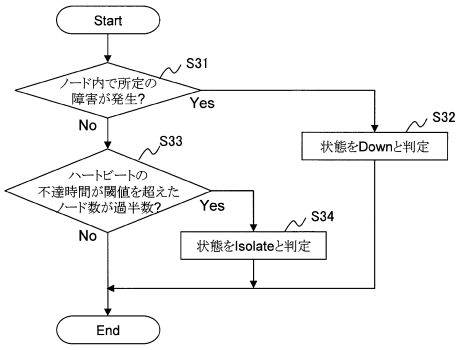
【図１１】



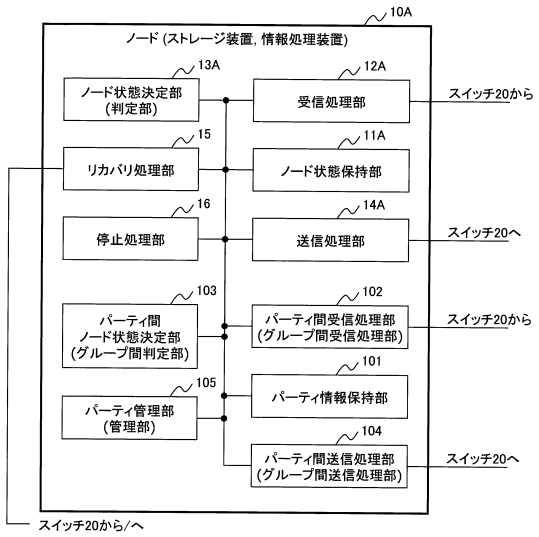
【図１２】



【図１３】



【図１４】

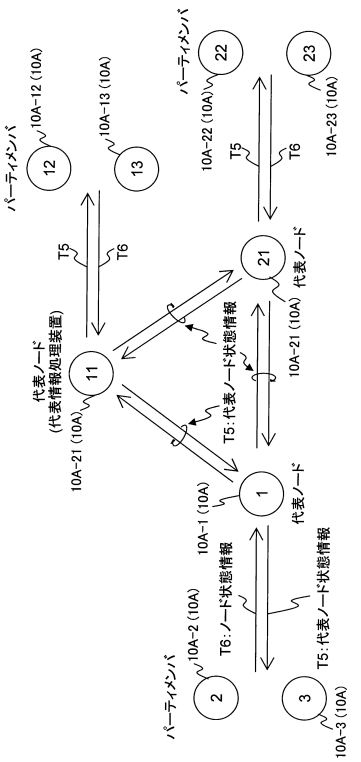


【図１５】

パーティ管理情報 T4

パーティID	ノードID	バージョン番号
A	1～10	1
B	11～20	1
C	21～30	3
D	31～40	4
E	41～50	2

【図１６】



【図１７】

代表ノード状態情報 (代表状態情報)
(代表ノードからメンバノード, 代表ノードから代表ノード)

T5

ノードID	状態	IPアドレス	ポート番号
1	Alive	192.168.0.1	12345
2	Alive	192.168.0.2	12345
3	Alive	192.168.0.3	12345
11	Alive	192.168.0.11	12345
12	Alive	192.168.0.12	12345
13	Alive	192.168.0.13	12345
21	Alive	192.168.0.21	12345
22	Alive	192.168.0.22	12345
23	Alive	192.168.0.23	12345

【図１８】

ノード状態情報 (状態情報)
(メンバノードから代表ノード, メンバノードからメンバノード)

T6

ノードID	状態	IPアドレス	ポート番号
1	Alive	192.168.0.1	12345
2	Alive	192.168.0.2	12345
3	Alive	192.168.0.3	12345

【図１９】

ノード状態管理情報

T7

ノードID	状態	by 2	by 3	by 11	by 12	by 13	by 21	by 22	by 23	最終更新情報	IPアドレス	ポート番号
1	Alive	Alive	Alive	Alive			Alive			1sec ago	192.168.0.1	12345
2	Alive	Alive	Alive							1sec ago	192.168.0.2	12345
3	Alive	Alive	Alive							1sec ago	192.168.0.3	12345
11	Alive			Alive	Alive	Alive	Alive	Alive		1sec ago	192.168.0.11	12345
12	Alive			Alive	Alive	Alive					192.168.0.12	12345
13	Alive			Alive	Alive	Alive					192.168.0.13	12345
21	Alive			Alive			Alive	Alive	Alive	1sec ago	192.168.0.21	12345
22	Alive						Alive	Alive	Alive		192.168.0.22	12345
23	Alive						Alive	Alive	Alive		192.168.0.23	12345

自ノードでの判断

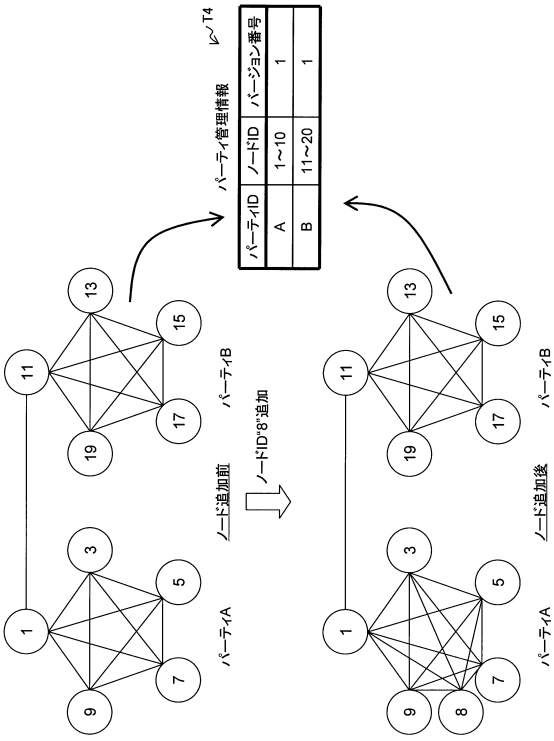
代表ノード間でハートビート不達のときに
代表ノードで判断

他ノードでの判断が利用される

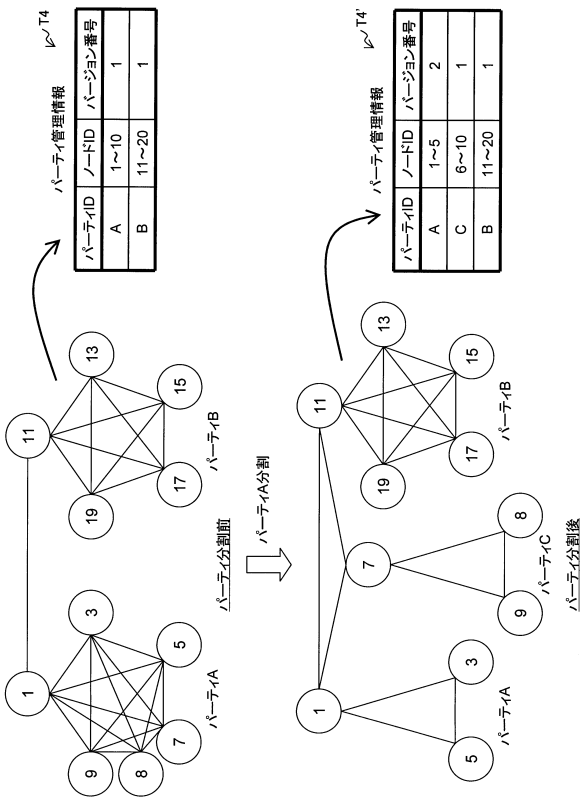
自ノードでの判断で参照される

代表ノード間でハートビート不達のときに
代表ノードでの判断で参照される

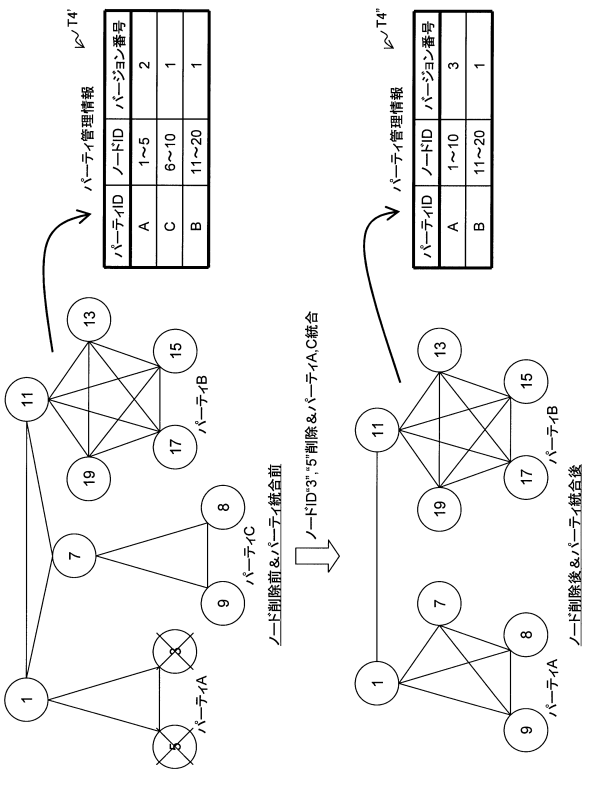
【図２０】



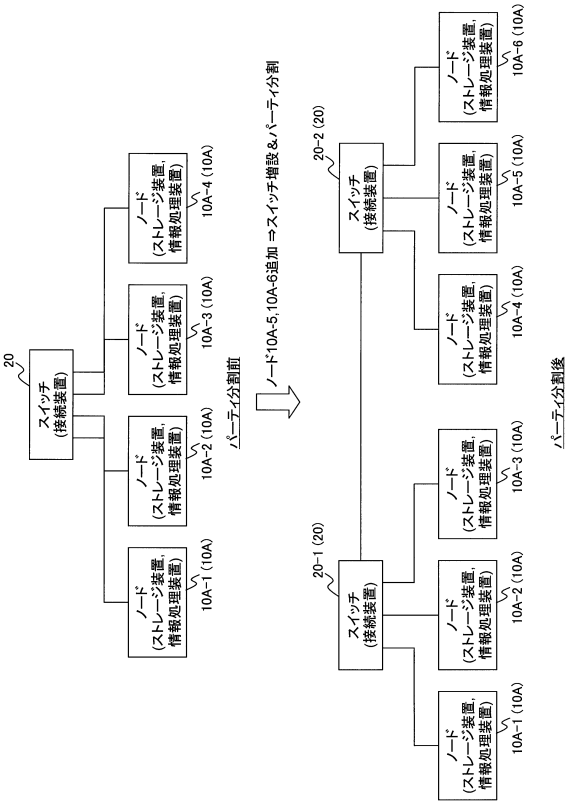
【図２１】



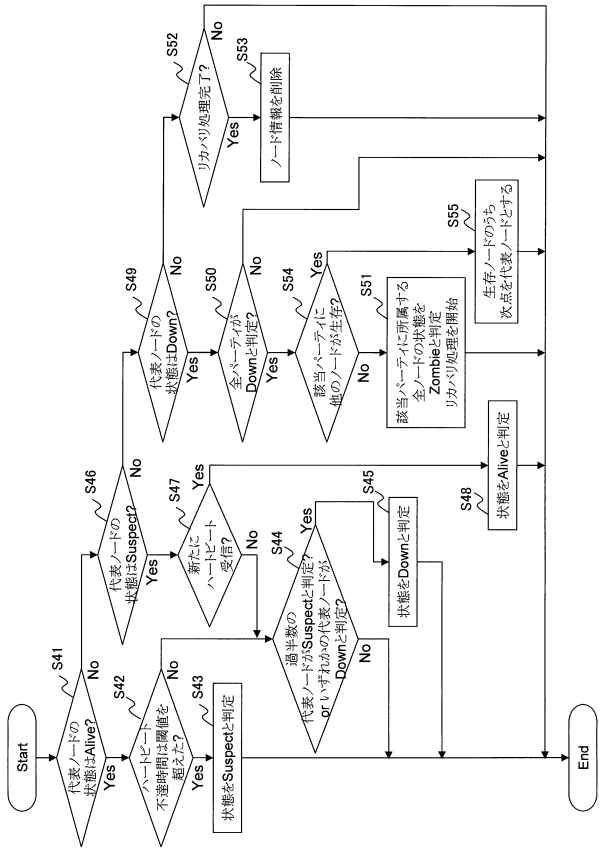
【図２２】



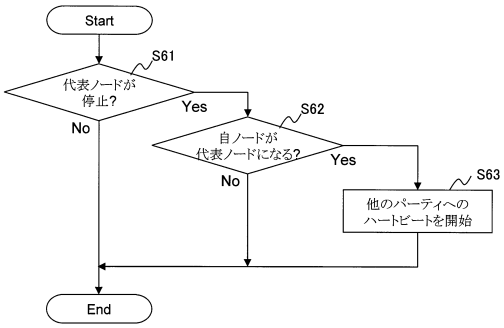
【図２３】



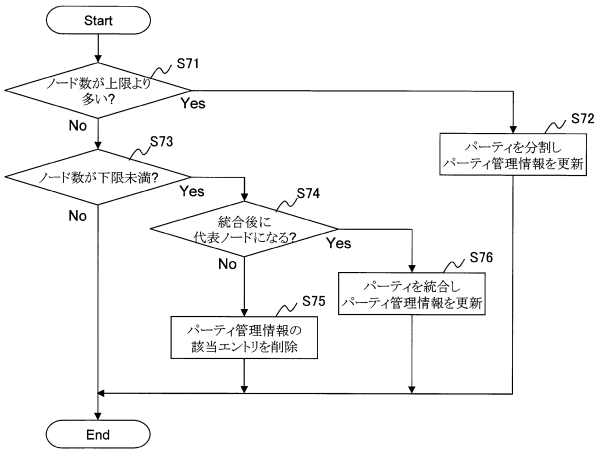
【図２４】



【図２５】


Start
S61
代表ノードが
停止?
Yes
No
S62
自ノードが
代表ノードになる?
Yes
No
S63
他のパーティへの
ハートビートを開始
End

【図２６】


Start
S71
ノード数が上限より
多い?
Yes
No
S72
パーティを分割し
パーティ管理情報を更新
S73
ノード数が下限未満?
Yes
No
S74
統合後に
代表ノードになる?
Yes
No
S76
パーティを統合し
パーティ管理情報を更新
S75
パーティ管理情報の
該当エントリを削除
End

フロントページの続き

(56)参考文献 特開２００２－１３２５３５（ＪＰ，Ａ）
特開昭５７－２０１９４５（ＪＰ，Ａ）
特開２００９－２１７５０４（ＪＰ，Ａ）
特開２００２－３１２１９９（ＪＰ，Ａ）

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ０６Ｆ　１１／２０
Ｇ０６Ｆ　１１／２２