(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第3815967号
(P3815967)
(45)発行日　平成18年8月30日(2006.8.30)　(24)登録日　平成18年6月16日(2006.6.16)

(51)Int.Cl.　　F I
G06F 12/00 (2006.01)　　G06F 12/00 535Z
G06F 12/08 (2006.01)　　G06F 12/00 514M
　　G06F 12/00 545A
　　G06F 12/08 531Z
　　G06F 12/08 557

請求項の数 190　(全 42 頁)

(21)出願番号　特願2000-531781 (P2000-531781)
(86)(22)出願日　平成11年2月12日 (1999.2.12)
(65)公表番号　特表2002-503846 (P2002-503846A)
(43)公表日　平成14年2月5日 (2002.2.5)
(86)国際出願番号　PCT/US1999/002965
(87)国際公開番号　WO1999/041664
(87)国際公開日　平成11年8月19日 (1999.8.19)
審査請求日　平成16年7月28日 (2004.7.28)
(31)優先権主張番号　60/074,587
(32)優先日　平成10年2月13日 (1998.2.13)
(33)優先権主張国　米国 (US)
(31)優先権主張番号　09/199,120
(32)優先日　平成10年11月24日 (1998.11.24)
(33)優先権主張国　米国 (US)

(73)特許権者　502303739
オラクル・インターナショナル・コーポレイション
アメリカ合衆国、９４０６５　カリフォルニア州、レッドウッド・ショアーズ、オラクル・パークウェイ、５００
(74)代理人　100064746
弁理士　深見　久郎
(74)代理人　100085132
弁理士　森田　俊雄
(74)代理人　100083703
弁理士　仲村　義平
(74)代理人　100096781
弁理士　堀井　豊

最終頁に続く

(54)【発明の名称】あるノードのキャッシュから別のノードのキャッシュへデータを転送するための方法および装置

(57)【特許請求の範囲】

【請求項１】

第１のキャッシュから第２のキャッシュへ資源を転送するための方法であって、

前記第１のキャッシュから永続的記憶装置に前記資源を最初に永続的に記憶することなしに、前記第１のキャッシュ内に資源の第１のコピーを保持する一方で資源の第２のコピーを第１のキャッシュから第２のキャッシュへと転送するステップと、

資源の前記第１のコピーまたはそのサクセサが永続的に記憶されるまで前記第１のキャッシュ内に資源の少なくとも１つのコピーを保持するステップとを含む、方法。

【請求項２】

前記第１のキャッシュは第１のデータベースサーバによって維持されるキャッシュであり、前記第２のキャッシュは第２のデータベースサーバによって維持されるキャッシュである、請求項１に記載の方法。

【請求項３】

前記第２のコピーを前記第２のキャッシュに転送するよりも前に前記資源の前記第１のコピーが前記第１のキャッシュ内で変更されることを可能にするステップと、

前記第２のコピーを前記第２のキャッシュに転送した後に前記資源の前記第１のコピーが変更されることを防ぐステップとをさらに含む、請求項１に記載の方法。

【請求項４】

前記第２のコピーを前記第２のキャッシュに転送した後に、前記第１のコピーの開放の許可の要求を送信するステップと、

前記要求に応答して、前記第１のコピーまたはそのサクセサが永続的に記憶されるようにするステップと、
前記サクセサが永続的に記憶されたことに応答して、前記第１のコピーが開放可能であることを示すメッセージを送信するステップとをさらに含む、請求項１に記載の方法。
【請求項５】
前記第１のコピーの開放の許可の要求を送信するステップは、送信側プロセスによって実行され、
前記第１のコピーまたはそのサクセサが永続的に記憶されるようにするステップは、送信側プロセス以外のプロセスが前記資源の前記第１のコピーのサクセサを記憶するようにするステップを含む、請求項４に記載の方法。
【請求項６】
前記第１のキャッシュ内に資源の少なくとも１つのコピーを保持するステップは、
前記第１のコピーを永続的に記憶しようとする前に、前記資源の永続的に記憶されたコピーが前記第１のコピーより最近のものであるかどうか決定するステップと、
もし前記永続的に記憶されたコピーが前記第１のコピーより最近のものであれば、前記第１のコピーを永続的に記憶することなしに前記第１のコピーを開放するステップと、
もし前記永続的に記憶されたコピーが前記第１のコピーより最近のものでなければ、前記第１のコピーを永続的に記憶するステップとを含む、請求項１に記載の方法。
【請求項７】
変更許可を、第１のキャッシュに関連付けられる送信側プロセスから第２のキャッシュに関連付けられる受信側プロセスに、前記資源の前記第２のコピーとともに転送するステップをさらに含む、請求項３に記載の方法。
【請求項８】
前記資源にアクセスする許可はマスタによって管理され、
前記変更許可を前記受信側プロセスに転送するステップは、前記変更許可の前記受信側プロセスへの転送についての確認を前記マスタから受信するより前に実行される、請求項７に記載の方法。
【請求項９】
変更許可を、前記第１のキャッシュに関連付けられる送信側プロセスから前記第２のキャッシュに関連付けられる受信側プロセスに、前記資源の第２のコピーとともに転送するステップをさらに含み、
前記資源にアクセスする許可はマスタによって管理され、
前記変更許可を前記受信側プロセスに転送するステップは、前記変更許可の前記受信側プロセスへの転送の確認を前記マスタが受信するより前に実行される、請求項１に記載の方法。
【請求項１０】
前記第２のキャッシュに関連付けられる受信側プロセスが前記資源の要求を前記資源のマスタに送信するステップと、
前記受信側プロセスからの前記要求に応答して、前記資源の前記マスタが前記第１のキャッシュに関連付けられる送信側プロセスにメッセージを送信するステップと、
前記送信側プロセスが前記マスタからの前記メッセージに応答して、前記第２のコピーを前記受信側プロセスに転送するステップとをさらに含む、請求項１に記載の方法。
【請求項１１】
前記第２のコピーを前記第２のキャッシュに転送するステップの後に、
前記第１のキャッシュに関連付けられる送信側プロセスが、ロックマネージャにロックを要求するステップを含み、前記ロックは前記資源をディスクに書込む許可は与えるが前記資源を変更する許可は与えず、さらに、
前記ロックマネージャが、前記第１のコピーと少なくとも同じほど最近である前記資源のバージョンを有するプロセスを選択するステップと、
前記ロックマネージャが、前記ロックを前記選択されたプロセスに与えるステップと、

前記選択されたプロセスが、前記資源の前記バージョンをディスクに書込むステップとをさらに含む、請求項１に記載の方法。

【請求項１２】

前記資源のバージョンがディスクに書込まれるのに応答して、前記ロックマネージャが前記バージョンよりも古い前記資源のすべてのバージョンが開放されるようにするステップをさらに含む、請求項１１に記載の方法。

【請求項１３】

前記資源のダーティコピーを保持するキャッシュの障害の後に、

前記障害の発生したキャッシュが資源の最新バージョンを保持していたかどうかを決定するステップと、

もし前記障害の発生したキャッシュが資源の最新バージョンを保持していたならば、

資源の最新パストイメージをディスクに書込むステップと、

資源の以前のすべてのパストイメージを開放するステップと、

前記障害の発生したキャッシュの復旧ログを適用して資源の最新バージョンを再構築するステップとをさらに含む、請求項１に記載の方法。

【請求項１４】

もし前記障害の発生したキャッシュが資源の最新バージョンを保持していなかったならば、

資源の最新バージョンをディスクに書込むステップと、

資源のすべてのパストイメージを開放するステップとをさらに含む、請求項１３に記載の方法。

【請求項１５】

前記資源のダーティバージョンを保持する複数のキャッシュの障害の後に、

前記障害の発生したキャッシュのいずれかが資源の最新バージョンを保持していたかどうかを決定するステップと、

もし前記障害の発生したキャッシュのいずれかが資源の最新バージョンを保持していたならば、

前記障害の発生したキャッシュの復旧ログをマージし適用して資源の最新バージョンを再構築するステップをさらに含む、請求項１に記載の方法。

【請求項１６】

複数のキャッシュが１つ以上の共有されるディスクからの資源のダーティバージョンを保持するシステムにおいてデータを管理する方法であって、

前記複数のキャッシュのうち１つのキャッシュが、前記複数のキャッシュのうち別のキャッシュにある資源のダーティバージョンを要求するとき、前記資源のダーティバージョンが存在するキャッシュから、前記ダーティバージョンを、前記ダーティバージョンを要求するキャッシュに、前記ダーティバージョンを最初に永続的に記憶することなしに転送するステップと、

前記複数のキャッシュの各キャッシュに対して別々の復旧ログを維持するステップと、

前記複数のキャッシュのうち１つのキャッシュに障害が発生したとき、前記障害の発生したキャッシュを、前記障害の発生したキャッシュに関連付けられた復旧ログに基づいて、前記複数のキャッシュのうちの他のキャッシュの前記別々の復旧ログを検査することなしに、復旧させるステップとを含む、方法。

【請求項１７】

前記複数のキャッシュの各キャッシュは、複数のデータベースサーバの別々のデータベースサーバによって維持されるキャッシュである、請求項１６に記載の方法。

【請求項１８】

前記ダーティバージョンが存在するキャッシュは第１のキャッシュであり、

前記第１のキャッシュ内の資源のダーティバージョンは資源の第１のコピーであり、

前記ダーティバージョンを要求するキャッシュは第２のキャッシュであり、

前記ダーティバージョンを転送するステップは、前記資源の第２のコピーを前記第２の

キャッシュに転送することにより実行され、
前記方法はさらに、
前記第２のコピーを前記第２のキャッシュに転送するよりも前に前記資源の前記第１のコピーが前記第１のキャッシュ内で変更されることを可能にするステップと、
前記第２のコピーを前記第２のキャッシュに転送した後に前記資源の前記第１のコピーが変更されることを防ぐステップとをさらに含む、請求項１６に記載の方法。
【請求項１９】
複数のノードによって共有される資源を管理するための方法であって、
前記複数のノードのうち第１のノードの第１のキャッシュ内で前記資源を変更して前記資源の変更されたバージョンを作成するステップと、
前記第１のノードに障害が発生したときにどこで作業を開始するかを示す、前記第１のノードについてのチェックポイントを維持するステップと、
前記第１のキャッシュから永続的記憶装置に前記変更されたバージョンを最初に永続的に記憶することなしに、前記第１のキャッシュ内に前記変更されたバージョンの第１のコピーを保持する一方で、前記変更されたバージョンの第２のコピーを前記第１のキャッシュから前記複数のノードのうち第２のノードの第２のキャッシュに転送するステップと、
前記複数のノードのうち別のノードが、前記変更されたバージョンと少なくとも同じほど最近のものである前記資源のバージョンを永続的に記憶したことが示されることに応答して、前記チェックポイントを進めるステップとを含む、方法。
【請求項２０】
前記変更されたバージョンの前記第１のコピーまたはそのサクセサが永続的に記憶されるまで、前記第１のキャッシュ内に前記変更されたバージョンの少なくとも１つのコピーを保持するステップをさらに含む、請求項１９に記載の方法。
【請求項２１】
前記複数のノードのうち別のノードが、前記変更されたバージョンと少なくとも同じほど最近のものである前記資源のバージョンを永続的に記憶したことが示されることに応答して、前記第１のノードで前記資源を開放するステップを含む、請求項１９に記載の方法。
【請求項２２】
第１のキャッシュから第２のキャッシュへ資源を転送するための方法であって、
前記第１のキャッシュから永続的記憶装置に前記資源を最初に永続的に記憶することなしに、前記第１のキャッシュ内に資源の第１のコピーを保持する一方で資源の第２のコピーを前記第１のキャッシュから前記第２のキャッシュへと転送するステップと、
前記資源の第１のコピーまたはそのサクセサが永続的に記憶されるまで前記第１のコピーが前記第１のキャッシュ内で置換されることを防ぐステップとを含む、方法。
【請求項２３】
命令の１つ以上のシーケンスを搬送して第１のキャッシュから第２のキャッシュへ資源を転送するためのコンピュータ読出可能媒体であって、前記命令の１つ以上のシーケンスが１つ以上のプロセッサによって実行されると、前記１つ以上のプロセッサは、
前記第１のキャッシュから永続的記憶装置に前記資源を最初に永続的に記憶することなしに、前記第１のキャッシュ内に資源の第１のコピーを保持する一方で資源の第２のコピーを第１のキャッシュから第２のキャッシュへと転送するステップと、
資源の前記第１のコピーまたはそのサクセサが永続的に記憶されるまで前記第１のキャッシュ内に資源の少なくとも１つのコピーを保持するステップとを実行する、コンピュータ読出可能媒体。
【請求項２４】
前記第２のコピーを前記第２のキャッシュに転送するよりも前に前記資源の前記第１のコピーが前記第１のキャッシュ内で変更されることを可能にするステップと、
前記第２のコピーを前記第２のキャッシュに転送した後に前記資源の前記第１のコピーが変更されることを防ぐステップとを実行するための命令のシーケンスをさらに含む、請

求項２３に記載のコンピュータ読出可能媒体。
【請求項２５】
前記第２のコピーを前記第２のキャッシュに転送した後に前記第１のコピーの開放の許可の要求を送信するステップと、
前記要求に応答して、前記第１のコピーまたはそのサクセサが永続的に記憶されるようにするステップと、
前記サクセサが永続的に記憶されたことに応答して、前記第１のコピーが置換可能であることを示すメッセージを送信するステップとを実行するための命令のシーケンスをさらに含む、請求項２３に記載のコンピュータ読出可能媒体。
【請求項２６】
前記第１のコピーの開放の許可の要求を送信するステップは送信側プロセスによって実行され、
前記第１のコピーまたはそのサクセサが永続的に記憶されるようにするステップは、前記送信側プロセス以外のプロセスが前記資源の前記第１のコピーのサクセサを記憶するようにするステップを含む、請求項２５に記載のコンピュータ読出可能媒体。
【請求項２７】
前記第１のキャッシュ内に資源の少なくとも１つのコピーを保持するステップは、
前記第１のコピーを永続的に記憶しようとする前に、前記資源の永続的に記憶されたコピーは前記第１のコピーより最近のものであるかどうか決定するステップと、
もし前記永続的に記憶されたコピーが前記第１のコピーより最近のものであれば、前記第１のコピーを永続的に記憶することなしに、前記第１のコピーを開放するステップと、
もし前記永続的に記憶されたコピーが前記第１のコピーよりも最近のものでなければ、前記第１のコピーを永続的に記憶するステップとを含む、請求項２３に記載のコンピュータ読出可能媒体。
【請求項２８】
変更許可を、前記第１のキャッシュに関連付けられる送信側プロセスから前記第２のキャッシュに関連付けられる受信側プロセスに、前記資源の前記第２のコピーとともに転送するステップを実行するための命令のシーケンスをさらに含む、請求項２４に記載のコンピュータ読出可能媒体。
【請求項２９】
前記資源にアクセスする許可はマスタによって管理され、
前記変更許可を受信側プロセスに転送するステップは、前記変更許可の前記受信側プロセスへの転送についての確認を前記マスタから受信するより前に実行される、請求項２８に記載のコンピュータ読出可能媒体。
【請求項３０】
変更許可を、前記第１のキャッシュに関連付けられる送信側プロセスから前記第２のキャッシュに関連付けられる受信側プロセスに、前記資源の第２のコピーとともに転送するステップを実行するための命令のシーケンスをさらに含み、
前記資源にアクセスする許可はマスタによって管理され、
前記変更許可を前記受信側プロセスに転送するステップは、前記変更許可の前記受信側プロセスへの転送の確認を前記マスタが受信するより前に実行される、請求項２３に記載のコンピュータ読出可能媒体。
【請求項３１】
前記第２のキャッシュに関連付けられる受信側プロセスが前記資源の要求を前記資源のマスタに送信するステップと、
前記受信側プロセスからの前記要求に応答して、前記資源の前記マスタが前記第１のキャッシュに関連付けられる送信側プロセスにメッセージを送信するステップと、
前記送信側プロセスが前記マスタからの前記メッセージに応答して、前記第２のコピーを前記受信側プロセスに転送するステップとを実行するための命令のシーケンスをさらに含む、請求項２３に記載のコンピュータ読出可能媒体。

【請求項３２】
前記第２のコピーを前記第２のキャッシュに転送するステップの後に、
前記第１のキャッシュに関連付けられる送信側プロセスが、ロックマネージャにロックを要求するステップを実行するための命令のシーケンスをさらに含み、前記ロックは前記資源をディスクに書込む許可は与えるが前記資源を変更する許可は与えず、さらに、
前記ロックマネージャが、前記第１のコピーと少なくとも同じほど最近である前記資源のバージョンを有するプロセスを選択するステップと、
前記ロックマネージャが、前記ロックを前記選択されたプロセスに与えるステップと、
前記選択されたプロセスが、前記資源の前記バージョンをディスクに書込むステップとを実行するための命令のシーケンスを含む、請求項２３に記載のコンピュータ読出可能媒体。
【請求項３３】
前記資源の前記バージョンがディスクに書込まれるのに応答して、前記ロックマネージャが前記バージョンよりも古い前記資源のすべてのバージョンが開放されるようにするステップを実行するための命令のシーケンスをさらに含む、請求項３２に記載のコンピュータ読出可能媒体。
【請求項３４】
前記資源のダーティコピーを保持するキャッシュの障害の後に、
前記障害の発生したキャッシュが資源の最新バージョンを保持していたかどうかを決定するステップと、
もし前記障害の発生したキャッシュが資源の最新バージョンを保持していたならば、
資源の最新パストイメージをディスクに書込むステップと、
資源の以前のすべてのパストイメージを開放するステップと、
前記障害の発生したキャッシュの復旧ログを適用して資源の最新バージョンを再構築するステップとを実行するための命令のシーケンスをさらに含む、請求項２３に記載のコンピュータ読出可能媒体。
【請求項３５】
もし前記障害の発生したキャッシュが資源の最新バージョンを保持していなかったならば、
資源の最新バージョンをディスクに書込むステップと、
資源のすべてのパストイメージを開放するステップとを実行するための命令のシーケンスをさらに含む、請求項３４に記載のコンピュータ読出可能媒体。
【請求項３６】
前記資源のダーティバージョンを保持する複数のキャッシュの障害の後に、
前記障害の発生したキャッシュのいずれかが資源の最新バージョンを保持していたかどうかを決定するステップと、
もし前記障害の発生したキャッシュのいずれかが資源の最新バージョンを保持していたならば、
前記障害の発生したキャッシュの復旧ログをマージし適用して資源の最新バージョンを再構築するステップとを実行するための命令のシーケンスをさらに含む、請求項２３に記載のコンピュータ読出可能媒体。
【請求項３７】
前記第１のキャッシュ内に資源の第１のコピーを保持するための命令は、第１のデータベースサーバによって維持される第１のキャッシュ内に資源の第１のコピーを保持するための命令を含み、
前記第１のキャッシュから永続的記憶装置に前記資源を最初に永続的に記憶することなしに資源の第２のコピーを第１のキャッシュから第２のキャッシュへと転送するための命令は、前記第１のデータベースサーバによって維持される前記第１のキャッシュから永続的記憶装置に前記資源を最初に永続的に記憶することなしに資源の第２のコピーを前記第１のデータベースサーバによって維持される第１のキャッシュから第２のデータベースサ

ーバによって維持される第２のキャッシュへと転送するための命令を含み、
資源の前記第１のコピーまたはそのサクセサが永続的に記憶されるまで前記第１のキャッシュ内に資源の少なくとも１つのコピーを保持するための命令は、資源の前記第１のコピーまたはそのサクセサが永続的に記憶されるまで前記第１のデータベースサーバによって維持される前記第１のキャッシュ内に資源の少なくとも１つのコピーを保持するための命令を含む、請求項２３に記載のコンピュータ読出可能媒体。
【請求項３８】
命令を搬送し、複数のキャッシュが１つ以上の共有されるディスクからの資源のダーティバージョンを保持するシステムにおいてデータを管理するためのコンピュータ読出可能媒体であって、前記命令は、
前記複数のキャッシュのうち１つのキャッシュが、前記複数のキャッシュのうち別のキャッシュにある資源のダーティバージョンを要求するとき、前記資源のダーティバージョンが存在するキャッシュから、前記ダーティバージョンを、前記ダーティバージョンを要求するキャッシュに、前記ダーティバージョンを最初に永続的に記憶することなしに転送するステップと、
前記複数のキャッシュの各キャッシュに対して別々の復旧ログを維持するステップと、
前記複数のキャッシュのうち１つのキャッシュに障害が発生したとき、前記障害の発生したキャッシュを、前記障害の発生したキャッシュに関連付けられた復旧ログに基づいて、前記複数のキャッシュのうちの他のキャッシュの前記別々の復旧ログを検査することなしに、復旧させるステップとを実行するための命令を含む、コンピュータ読出可能媒体。
【請求項３９】
前記複数のキャッシュの各キャッシュは、複数のデータベースサーバの別々のデータベースサーバによって維持されるキャッシュである、請求項３８に記載のコンピュータ読出可能媒体。
【請求項４０】
前記ダーティバージョンが存在するキャッシュは第１のキャッシュであり、
前記第１のキャッシュ内の資源のダーティバージョンは資源の第１のコピーであり、
前記ダーティバージョンを要求するキャッシュは第２のキャッシュであり、
前記ダーティバージョンを転送するステップは、前記資源の第２のコピーを前記第２のキャッシュに転送することにより実行され、
前記コンピュータ読出可能媒体はさらに、
前記第２のコピーを前記第２のキャッシュに転送するよりも前に前記資源の前記第１のコピーが前記第１のキャッシュ内で変更されることを可能にするステップと、
前記第２のコピーを前記第２のキャッシュに転送した後に前記資源の前記第１のコピーが変更されることを防ぐステップとを実行するための命令をさらに含む、請求項３８に記載のコンピュータ読出可能媒体。
【請求項４１】
命令の１つ以上のシーケンスを搬送して複数のノードが共有する資源を管理するためのコンピュータ読出可能媒体であって、前記命令の１つ以上のシーケンスが１つ以上のプロセッサによって実行されると、前記１つ以上のプロセッサは、
前記複数のノードのうち第１のノードの第１のキャッシュ内で前記資源を変更して前記資源の変更されたバージョンを作成するステップと、
前記第１のノードに障害が発生したときにどこで作業を開始するかを示す、前記第１のノードについてのチェックポイントを維持するステップと、
前記第１のキャッシュから永続的記憶装置に前記変更されたバージョンを最初に永続的に記憶することなしに、前記第１のキャッシュ内に前記変更されたバージョンの第１のコピーを保持する一方で、前記変更されたバージョンの第２のコピーを前記第１のキャッシュから前記複数のノードのうち第２のノードの第２のキャッシュに転送するステップと、
前記複数のノードのうち別のノードが、前記変更されたバージョンと少なくとも同じほど最近のものである前記資源のバージョンを永続的に記憶したことが示されることに応答

して、前記チェックポイントを進めるステップと実行する、コンピュータ読出可能媒体。
【請求項４２】
前記１つ以上のプロセッサによって実行されると、前記１つ以上のプロセッサが、
前記変更されたバージョンの前記第１のコピーまたはそのサクセサが永続的に記憶されるまで、前記第１のキャッシュ内に前記変更されたバージョンの少なくとも１つのコピーを保持するステップを実行するようにする命令をさらに含む、請求項４１に記載のコンピュータ読出可能媒体。
【請求項４３】
前記１つ以上のプロセッサによって実行されると、前記１つ以上のプロセッサが、
前記複数のノードのうち別のノードが、前記変更されたバージョンと少なくとも同じほど最近のものである前記資源のバージョンを永続的に記憶したことが示されることに応答して、前記第１のノードで前記資源を開放するステップを実行するようにする命令をさらに含む、請求項４１に記載のコンピュータ読出可能媒体。
【請求項４４】
命令の１つ以上のシーケンスを搬送して第１のキャッシュから第２のキャッシュへ資源を転送するためのコンピュータ読出可能媒体であって、前記命令の１つ以上のシーケンスが１つ以上のプロセッサによって実行されると、前記１つ以上のプロセッサは、
前記第１のキャッシュから永続的記憶装置に前記資源を最初に永続的に記憶することなしに、前記第１のキャッシュ内に資源の第１のコピーを保持する一方で資源の第２のコピーを前記第１のキャッシュから前記第２のキャッシュへと転送するステップと、
前記資源の第１のコピーまたはそのサクセサが永続的に記憶されるまで前記第１のコピーが前記第１のキャッシュ内で置換されることを防ぐステップとを実行する、コンピュータ読出可能媒体。
【請求項４５】
資源を転送するためのシステムであって、
第１のキャッシュを有するノードを含み、前記第１のキャッシュは、１つ以上の他のノードに含まれる１つ以上の他のキャッシュのうち第２のキャッシュに通信可能に結合され、
前記ノードは、前記第１のキャッシュから永続的記憶装置に前記資源を最初に永続的に記憶することなしに、前記第１のキャッシュ内に資源の第１のコピーを保持する一方で資源の第２のコピーを第１のキャッシュから第２のキャッシュへと転送するように構成され、
前記ノードは、資源の前記第１のコピーまたはそのサクセサが永続的に記憶されるまで前記第１のキャッシュ内に資源の少なくとも１つのコピーを保持するように構成される、システム。
【請求項４６】
前記ノードは第１のデータベースサーバであり、前記１つ以上の他のノードのうち少なくとも１つのノードは前記第２のキャッシュを含む第２のデータベースサーバである、請求項４５に記載のシステム。
【請求項４７】
前記ノードは、前記第２のコピーを前記第２のキャッシュに転送するよりも前に前記資源の前記第１のコピーが前記第１のキャッシュ内で変更されることを可能にするように構成され、
前記ノードは、前記第２のコピーを前記第２のキャッシュに転送した後に前記資源の前記第１のコピーが変更されることを防ぐように構成される、請求項４５に記載のシステム。
【請求項４８】
前記ノードは、前記第２のコピーを前記第２のキャッシュに転送した後に、前記第１のコピーの開放の許可の要求をマスタノードに送信するように構成され、
前記ノードは、前記要求に応答して、前記マスタノードが前記第１のコピーまたはその

サクセサが永続的に記憶されるようにした後に、前記第１のコピーが開放可能であることを示すメッセージを前記マスタノードから受信するように構成される、請求項４５に記載のシステム。

【請求項４９】

前記ノードは、前記第１のコピーの開放の許可の要求を送信するように構成される送信側プロセスを含み、

前記送信側プロセス以外のプロセスは前記資源の前記第１のコピーのサクセサを記憶する、請求項４８に記載のシステム。

【請求項５０】

前記ノードは、前記第１のキャッシュ内に資源の少なくとも１つのコピーを、

前記第１のコピーを永続的に記憶しようとする前に、前記資源の永続的に記憶されたコピーが前記第１のコピーより最近のものであるかどうか決定し、

もし前記永続的に記憶されたコピーが前記第１のコピーより最近のものであれば、前記第１のコピーを永続的に記憶することなしに前記第１のコピーを開放し、

もし前記永続的に記憶されたコピーが前記第１のコピーより最近のものでなければ、前記第１のコピーを永続的に記憶することにより、保持するように構成される、請求項４５に記載のシステム。

【請求項５１】

前記ノードは、変更許可を、第１のキャッシュに関連付けられる送信側プロセスから第２のキャッシュに関連付けられる受信側プロセスに、前記資源の前記第２のコピーとともに転送するように構成される、請求項４７に記載のシステム。

【請求項５２】

前記資源にアクセスする許可はマスタノードによって管理され、

前記ノードは、前記変更許可の前記受信側プロセスへの転送についての確認を前記第１のノードが前記マスタノードから受信するより前に、前記変更許可を受信側プロセスに転送するように構成される、請求項５１に記載のシステム。

【請求項５３】

前記ノードは、変更許可を、前記第１のキャッシュに関連付けられる送信側プロセスから前記第２のキャッシュに関連付けられる受信側プロセスに、前記資源の前記第２のコピーとともに転送するように構成され、

前記資源へのアクセス許可はマスタによって管理され、

前記変更許可の前記受信側プロセスへの転送は、前記変更許可の前記受信側プロセスへの転送の確認を前記マスタが受信するより前に実行される、請求項４５に記載のシステム。

【請求項５４】

前記ノードは前記第１のキャッシュに関連付けられる送信側プロセスを含み、前記送信側プロセスは、前記第２のキャッシュに関連付けられる受信側プロセスから前記資源の要求を受信したマスタノードからメッセージを受信するように構成され、

前記送信側プロセスは、前記マスタノードからの前記メッセージに応答して、前記第２のコピーを前記受信側プロセスに転送する、請求項４５に記載のシステム。

【請求項５５】

前記ノードは、前記第２のコピーが前記第２のキャッシュに転送された後にロックマネージャにロックを要求するように構成され、前記ロックは前記資源をディスクに書込む許可は与えるが前記資源を変更する許可は与えず、

前記ロックマネージャは、前記第１のコピーと少なくとも同じほど最近である前記資源のバージョンを有するプロセスを選択し前記ロックを前記選択されたプロセスに与えて、前記選択されたプロセスが前記資源の前記バージョンをディスクに書込むようにする、請求項４５に記載のシステム。

【請求項５６】

前記資源の前記バージョンがディスクに書込まれるのに応答して、前記ロックマネージ

ャが、前記バージョンよりも古い前記資源のすべてのバージョンが開放されるようにする、請求項５５に記載のシステム。

【請求項５７】

前記資源のダーティコピーを保持するキャッシュの障害の後に、障害の発生したキャッシュが資源の最新バージョンを保持していたかどうかを決定するように構成されるマスタノードを含み、

もし前記障害の発生したキャッシュが資源の最新バージョンを保持していたならば、マスタノードは、資源の最新パストイメージをディスクに書込み、資源の以前のすべてのパストイメージを開放し、前記障害の発生したキャッシュの復旧ログを適用して資源の最新バージョンを再構築するように構成される、請求項４５に記載のシステム。

【請求項５８】

前記マスタノードは、もし前記障害の発生したキャッシュが資源の最新バージョンを保持していなかったならば、資源の最新バージョンをディスクに書込み資源のすべてのパストイメージを開放するように構成される、請求項５７に記載のシステム。

【請求項５９】

前記資源のダーティバージョンを保持する複数のキャッシュの障害の後に、前記障害の発生したキャッシュのいずれかが資源の最新バージョンを保持していたかどうかを決定し、もし前記障害の発生したキャッシュのいずれかが資源の最新バージョンを保持していたならば、前記障害の発生したキャッシュの復旧ログをマージし適用して資源の最新バージョンを再構築するように構成されるマスタノードをさらに含む、請求項４５に記載のシステム。

【請求項６０】

１つ以上の共有されるディスクからの資源のダーティバージョンを保持する複数のキャッシュの間でデータを管理するためのシステムであって、

前記複数のキャッシュのうち１つのキャッシュを含むノードを含み、前記複数のキャッシュの各キャッシュに対して別々の復旧ログが維持され、前記複数のキャッシュのうち別のキャッシュが前記キャッシュ内に存在する資源のダーティバージョンを要求し、前記ノードは、前記ダーティバージョンを永続的に記憶することなしに前記ダーティバージョンを他のキャッシュに転送するように構成され、前記システムはさらに、

前記複数のキャッシュのうち障害の発生したキャッシュを、前記障害の発生したキャッシュに関連付けられた復旧ログに基づいて、前記複数のキャッシュのうち他のキャッシュの前記別々の復旧ログを検査することなしに、復旧するように構成されたマスタノードを含む、システム。

【請求項６１】

前記複数のキャッシュの各キャッシュは、複数のデータベースサーバの別々のデータベースサーバによって維持されるキャッシュである、請求項６０に記載のシステム。

【請求項６２】

前記ダーティバージョンが存在するキャッシュは第１のキャッシュであり、

前記第１のキャッシュ内の資源のダーティバージョンは資源の第１のコピーであり、

前記ダーティバージョンを要求するキャッシュは第２のキャッシュであり、

ノードは、資源の第２のコピーを第２のキャッシュに転送することにより資源のダーティバージョンを転送するように構成され、

第１のノードは、前記第２のコピーを前記第２のキャッシュに転送するよりも前に前記資源の前記第１のコピーが前記第１のキャッシュ内で変更されることを可能にし、

前記第１のノードは、前記第２のコピーを前記第２のキャッシュに転送した後に前記資源の前記第１のコピーが変更されることを防ぐように構成される、請求項６０に記載のシステム。

【請求項６３】

複数のノードによって共有される資源を管理するためのシステムであって、

第１のキャッシュを有するノードを含み、前記第１のキャッシュは、１つ以上の他のノ

ードに含まれる１つ以上の他のキャッシュのうちの第２のキャッシュに通信可能に結合され、
前記ノードは、
前記ノードの第１のキャッシュ内で前記資源を変更して前記資源の変更されたバージョンを作成し、
前記ノードに障害が発生したときにどこで作業を開始するかを示す、前記ノードについてのチェックポイントを維持し、
前記第１のキャッシュから永続的記憶装置に前記変更されたバージョンを最初に永続的に記憶することなしに、前記第１のキャッシュ内に前記変更されたバージョンの第１のコピーを保持する一方で、前記変更されたバージョンの第２のコピーを前記第１のキャッシュから第２のキャッシュに転送し、
前記複数のノードのうち別のノードが、前記変更されたバージョンと少なくとも同じほど最近のものである前記資源のバージョンを永続的に記憶したことが示されることに応答して、前記チェックポイントを進めるように構成される、システム。
【請求項６４】
前記ノードはさらに、
前記変更されたバージョンの前記第１のコピーまたはそのサクセサが永続的に記憶されるまで、前記第１のキャッシュ内に前記変更されたバージョンの少なくとも１つのコピーを保持するように構成される、請求項６３に記載のシステム。
【請求項６５】
前記ノードはさらに、
前記複数のノードのうち別のノードが、前記変更されたバージョンと少なくとも同じほど最近のものである前記資源のバージョンを永続的に記憶したことが示されることに応答して、前記第１のノードで前記資源を開放するように構成される、請求項６３に記載のシステム。
【請求項６６】
資源を転送するためのシステムであって、
第１のキャッシュを有するノードを含み、前記第１のキャッシュは、１つ以上の他のノードに含まれる１つ以上の他のキャッシュのうち第２のキャッシュに通信可能に結合され、
前記ノードは、前記第１のキャッシュから永続的記憶装置に前記資源を最初に永続的に記憶することなしに、前記第１のキャッシュ内に資源の第１のコピーを保持する一方で資源の第２のコピーを前記第１のキャッシュから前記第２のキャッシュへと転送するように構成され、
前記ノードは、資源の前記第１のコピーまたはそのサクセサが永続的に記憶されるまで前記第１のコピーが前記第１のキャッシュ内で置換されることを防ぐように構成される、システム。
【請求項６７】
資源のダーティバージョンを保持する１つ以上のキャッシュの障害の後に前記資源の復旧を管理するための方法であって、
障害の発生しなかった複数のキャッシュの各キャッシュにおいて資源のどのバージョンが保持されているのかを決定するステップと、
前記複数のキャッシュから、前記障害の発生しなかった複数のキャッシュ内の資源の他のいずれかのバージョンと少なくとも同じほど最近である資源の特定のバージョンを識別するステップと、
前記資源の特定のバージョンがディスクに書込まれるようにするステップとを含む、方法。
【請求項６８】
前記資源の特定のバージョンは資源の現在のコピーである、請求項６７に記載の方法。
【請求項６９】

前記資源の特定のバージョンは資源のパストイメージコピーである、請求項６７に記載の方法。
【請求項７０】
前記資源のパストイメージコピーは、資源の他のいずれかのパストイメージコピーと少なくとも同じほど最近のものである、請求項６９に記載の方法。
【請求項７１】
前記資源のパストイメージコピーは、前記障害の発生しなかった複数のキャッシュに現在保持されている資源の他のいずれかのパストイメージコピーと少なくとも同じほど最近のものである、請求項７０に記載の方法。
【請求項７２】
資源の最新バージョンがディスクに書込まれた後、前記複数のキャッシュが資源のパストイメージを開放するようにするステップをさらに含む、請求項６７に記載の方法。
【請求項７３】
前記決定するステップ、識別するステップおよびディスクに書込まれるようにするステップは、ロックマネージャによって実行される、請求項６７に記載の方法。
【請求項７４】
前記ロックマネージャは、前記１つ以上のキャッシュの障害の前に前記資源のマスタに指定された、生き残ったマスタである、請求項７３に記載の方法。
【請求項７５】
前記ロックマネージャは、前記１つ以上のキャッシュの障害の後に前記資源のマスタに指定された、新しいマスタである、請求項７３に記載の方法。
【請求項７６】
資源のダーティバージョンを保持する複数のキャッシュの障害の後に前記資源を復旧させるための方法であって、
前記障害の発生した複数のキャッシュのいずれかが資源の最新バージョンを保持していたかどうかを決定するステップと、
もし前記障害の発生した複数のキャッシュのいずれかが資源の最新バージョンを保持していたならば、
障害の発生したキャッシュのサブセットを決定するステップを含み、前記障害の発生したキャッシュのサブセットは、永続的に記憶された資源のバージョンに続いて前記資源を更新した、障害の発生したキャッシュのみを含み、さらに、
前記障害の発生したキャッシュのサブセットの復旧ログをマージし適用して資源の最新バージョンを再構築するステップを含む、方法。
【請求項７７】
命令の１つ以上のシーケンスを搬送して資源のダーティバージョンを保持する１つ以上のキャッシュの障害の後に前記資源の復旧を管理するためのコンピュータ読出可能媒体であって、前記命令の１つ以上のシーケンスが１つ以上のプロセッサによって実行されると、前記１つ以上のプロセッサは、
障害の発生しなかった複数のキャッシュの各キャッシュにおいて資源のどのバージョンが保持されているのかを決定するステップと、
前記複数のキャッシュから、前記障害の発生しなかった複数のキャッシュ内の資源の他のいずれかのバージョンと少なくとも同じほど最近である資源の特定のバージョンを識別するステップと、
前記資源の特定のバージョンがディスクに書込まれるようにするステップとを実行する、コンピュータ読出可能媒体。
【請求項７８】
前記資源の特定のバージョンは資源の現在のコピーである、請求項７７に記載のコンピュータ読出可能媒体。
【請求項７９】
前記資源の特定のバージョンは資源のパストイメージコピーである、請求項７７に記載

のコンピュータ読出可能媒体。
【請求項８０】
前記資源のパストイメージコピーは、資源の他のいずれかのパストイメージコピーと少なくとも同じほど最近のものである、請求項７９に記載のコンピュータ読出可能媒体。
【請求項８１】
前記資源のパストイメージコピーは、前記障害が発生しなかった複数のキャッシュに現在保持されている資源の他のいずれかのパストイメージコピーと少なくとも同じほど最近のものである、請求項８０に記載のコンピュータ読出可能媒体。
【請求項８２】
１つ以上のプロセッサによって実行されると、前記１つ以上のプロセッサが、
資源の最新バージョンがディスクに書込まれた後、前記複数のキャッシュが資源のパストイメージを開放するようにするステップを実行するようにする命令をさらに含む、請求項７７に記載のコンピュータ読出可能媒体。
【請求項８３】
前記命令の１つ以上のシーケンスが実行されると、前記１つ以上のプロセッサは、ロックマネージャに、前記決定するステップ、識別するステップおよび書込まれるようにするステップを実行させる、請求項７７に記載のコンピュータ読出可能媒体。
【請求項８４】
前記ロックマネージャは、前記１つ以上のキャッシュの障害の前に前記資源のマスタに指定された、生き残ったマスタである、請求項８３に記載のコンピュータ読出可能媒体。
【請求項８５】
前記ロックマネージャは、前記１つ以上のキャッシュの障害の後に前記資源のマスタに指定された、新しいマスタである、請求項８３に記載のコンピュータ読出可能媒体。
【請求項８６】
命令の１つ以上のシーケンスを搬送して資源のダーティバージョンを保持する複数のキャッシュの障害の後に前記資源を復旧させるためコンピュータ読出可能媒体であって、前記命令の１つ以上のシーケンスが１つ以上のプロセッサによって実行されると、前記１つ以上のプロセッサは、
前記障害の発生したキャッシュのいずれかが資源の最新バージョンを保持していたかどうかを決定するステップを実行し、
もし前記障害の発生したキャッシュのいずれかが資源の最新バージョンを保持していたならば、
障害の発生したキャッシュのサブセットを決定するステップを実行し、前記障害の発生したキャッシュのサブセットは、永続的に記憶された資源のバージョンに続いて前記資源を更新した、障害の発生したキャッシュのみを含み、さらに、
前記障害の発生したキャッシュのサブセットの復旧ログをマージし適用して資源の最新バージョンを再構築するステップを実行する、コンピュータ読出可能媒体。
【請求項８７】
資源のダーティバージョンを保持する１つ以上のキャッシュの障害の後に前記資源の復旧を管理するための装置であって、前記装置は、
障害の発生しなかった複数のキャッシュの各キャッシュにおいて資源のどのバージョンが保持されているのかを決定し、
前記複数のキャッシュから、前記障害の発生しなかった複数のキャッシュ内の資源の他のいずれかのバージョンと少なくとも同じほど最近である資源の特定のバージョンを識別し、
前記資源の特定のバージョンがディスクに書込まれるようにするように構成される、装置。
【請求項８８】
前記資源の特定のバージョンは資源の現在のコピーである、請求項８７に記載の装置。
【請求項８９】

前記資源の特定のバージョンは資源のパストイメージコピーである、請求項８７に記載の装置。

【請求項９０】

前記資源のパストイメージコピーは、資源の他のいずれかのパストイメージコピーと少なくとも同じほど最近のものである、請求項８９に記載の装置。

【請求項９１】

前記資源のパストイメージコピーは、前記障害の発生しなかった複数のキャッシュに現在保持されている資源の他のいずれかのパストイメージコピーと少なくとも同じほど最近のものである、請求項９０に記載の装置。

【請求項９２】

前記装置はさらに、資源の最新バージョンがディスクに書込まれた後、前記複数のキャッシュが資源のパストイメージを開放するように構成される、請求項８７に記載の装置。

【請求項９３】

前記装置はロックマネージャである、請求項８７に記載の装置。

【請求項９４】

前記ロックマネージャは、前記１つ以上のキャッシュの障害の前に前記資源のマスタに指定された、生き残ったマスタである、請求項９３に記載の装置。

【請求項９５】

前記ロックマネージャは、前記１つ以上のキャッシュの障害の後に前記資源のマスタに指定された、新しいマスタである、請求項９３に記載の装置。

【請求項９６】

資源のダーティバージョンを保持する複数のキャッシュの障害の後に前記資源を復旧させるための装置であって、前記装置は、

前記障害の発生したキャッシュのいずれかが資源の最新バージョンを保持していたかどうかを決定し、

もし前記障害の発生したキャッシュのいずれかが資源の最新バージョンを保持していたならば、

障害の発生したキャッシュのサブセットを決定し、前記障害の発生したキャッシュのサブセットは、永続的に記憶された資源のバージョンに続いて前記資源を更新した、障害の発生したキャッシュのみを含み、さらに、

前記障害の発生したキャッシュのサブセットの復旧ログをマージし適用して資源の最新バージョンを再構築するように構成される、装置。

【請求項９７】

資源のダーティコピーを保持する第１のキャッシュの障害の後に前記資源を復旧させるための方法であって、

前記第１のキャッシュから永続的記憶装置に前記資源を最初に永続的に記憶することなしに、前記第１のキャッシュ内に資源の第１のコピーを保持する一方で資源の第２のコピーを前記第１のキャッシュから第２のキャッシュへと転送するステップと、

前記障害の発生したキャッシュが資源の最新バージョンを保持していたかどうかを決定するステップと、

もし前記障害の発生したキャッシュが資源の最新バージョンを保持していたならば、前記障害の発生したキャッシュの復旧ログを資源の以前のバージョンに適用して資源の最新バージョンを再構築するステップとを含む、方法。

【請求項９８】

前記資源の以前のバージョンはディスク上に永続的に記憶される、請求項９７に記載の方法。

【請求項９９】

前記資源の以前のバージョンは第３のキャッシュ内の資源のパストイメージである、請求項９７に記載の方法。

【請求項１００】

前記第３のキャッシュ内の資源の前記パストイメージは、障害の発生しなかった複数のキャッシュのうちいずれかのキャッシュ内に現在保持されている資源の他のいずれかのパストイメージと少なくとも同じほど最新のものである、請求項９９に記載の方法。

【請求項１０１】

前記障害の発生したキャッシュの復旧ログは、永続的記憶装置に前記資源の以前のバージョンを最初に永続的に記憶することなしに、前記第３のキャッシュ内の資源の以前のバージョンに適用される、請求項９９に記載の方法。

【請求項１０２】

資源のダーティバージョンを保持する複数のキャッシュの障害の後に前記資源を復旧させるための方法であって、

第１のキャッシュから永続的記憶装置に前記資源を最初に永続的に記憶することなしに、前記第１のキャッシュ内に資源の第１のコピーを保持する一方で資源の第２のコピーを前記第１のキャッシュから第２のキャッシュへと転送するステップと、

前記障害の発生したキャッシュのいずれかが資源の最新バージョンを保持していたかどうかを決定するステップと、

もし前記障害の発生したキャッシュのいずれかが資源の最新バージョンを保持していたならば、

前記障害の発生したキャッシュの復旧ログをマージするステップと、

前記マージされた復旧ログを資源の以前のバージョンに適用して資源の最新バージョンを再構築するステップとを含む、方法。

【請求項１０３】

前記資源の以前のバージョンはディスク上に永続的に記憶される、請求項１０２に記載の方法。

【請求項１０４】

前記資源の以前のバージョンは、前記障害の発生した複数のキャッシュに含まれない第３のキャッシュ内の資源のパストイメージである、請求項１０２に記載の方法。

【請求項１０５】

前記第３のキャッシュ内の資源の前記パストイメージは、前記障害の発生した複数のキャッシュに含まれないいずれかのキャッシュ内に現在保持されている資源の他のいずれかのパストイメージと少なくとも同じほど最新のものである、請求項１０４に記載の方法。

【請求項１０６】

前記マージされた復旧ログは、永続的記憶装置に前記資源を最初に永続的に記憶することなしに、前記第３のキャッシュ内の資源の以前のバージョンに適用される、請求項１０４に記載の方法。

【請求項１０７】

命令の１つ以上のシーケンスを搬送して資源のダーティコピーを保持する第１のキャッシュの障害の後に前記資源を復旧させるためのコンピュータ読出可能媒体であって、前記命令の１つ以上のシーケンスが１つ以上のプロセッサによって実行されると、前記１つ以上のプロセッサは、

前記第１のキャッシュから永続的記憶装置に前記資源を最初に永続的に記憶することなしに、前記第１のキャッシュ内に資源の第１のコピーを保持する一方で資源の第２のコピーを前記第１のキャッシュから第２のキャッシュへと転送するステップと、

前記障害の発生したキャッシュが資源の最新バージョンを保持していたかどうかを決定するステップと、

もし前記障害の発生したキャッシュが資源の最新バージョンを保持していたならば、前記障害の発生したキャッシュの復旧ログを資源の以前のバージョンに適用して資源の最新バージョンを再構築するステップとを実行する、コンピュータ読出可能媒体。

【請求項１０８】

前記資源の以前のバージョンはディスク上に永続的に記憶される、請求項１０７に記載のコンピュータ読出可能媒体。

【請求項１０９】
前記資源の以前のバージョンは第３のキャッシュ内の資源のパストイメージである、請求項１０７に記載のコンピュータ読出可能媒体。
【請求項１１０】
前記第３のキャッシュ内の資源の前記パストイメージは、障害の発生しなかった複数のキャッシュのうちいずれかのキャッシュ内に現在保持されている資源の他のいずれかのパストイメージと少なくとも同じほど最新のものである、請求項１０９に記載のコンピュータ読出可能媒体。
【請求項１１１】
前記障害の発生したキャッシュの復旧ログは、永続的記憶装置に前記資源の以前のバージョンを最初に永続的に記憶することなしに、前記第３のキャッシュ内の資源の以前のバージョンに適用される、請求項１０９に記載のコンピュータ読出可能媒体。
【請求項１１２】
命令の１つ以上のシーケンスを搬送し、資源のダーティバージョンを保持する複数のキャッシュの障害の後に前記資源を復旧させるためのコンピュータ読出可能媒体であって、前記命令の１つ以上のシーケンスが１つ以上のプロセッサによって実行されると、前記１つ以上のプロセッサは、
第１のキャッシュから永続的記憶装置に前記資源を最初に永続的に記憶することなしに、前記第１のキャッシュ内に資源の第１のコピーを保持する一方で資源の第２のコピーを前記第１のキャッシュから第２のキャッシュへと転送するステップと、
前記障害の発生したキャッシュのいずれかが資源の最新バージョンを保持していたかどうかを決定するステップと、
もし前記障害の発生したキャッシュのいずれかが資源の最新バージョンを保持していたならば、
前記障害の発生したキャッシュの復旧ログをマージするステップと、
前記マージされた復旧ログを資源の以前のバージョンに適用して資源の最新バージョンを再構築するステップとを実行する、コンピュータ読出可能媒体。
【請求項１１３】
前記資源の以前のバージョンはディスク上に永続的に記憶される、請求項１１２に記載のコンピュータ読出可能媒体。
【請求項１１４】
前記資源の以前のバージョンは、前記障害の発生した複数のキャッシュに含まれない第３のキャッシュ内の資源のパストイメージである、請求項１１２に記載のコンピュータ読出可能媒体。
【請求項１１５】
前記第３のキャッシュ内の資源の前記パストイメージは、前記障害の発生した複数のキャッシュに含まれないいずれかのキャッシュ内に現在保持されている資源の他のいずれかのパストイメージと少なくとも同じほど最新のものである、請求項１１４に記載のコンピュータ読出可能媒体。
【請求項１１６】
前記マージされた復旧ログは、永続的記憶装置に前記資源を最初に永続的に記憶することなしに、前記第３のキャッシュ内の資源の以前のバージョンに適用される、請求項１１４に記載のコンピュータ読出可能媒体。
【請求項１１７】
資源のダーティコピーを保持する第１のキャッシュの障害の後に前記資源を復旧させるための装置であって、前記装置は、
前記第１のキャッシュから永続的記憶装置に前記資源を最初に永続的に記憶することなしに、前記第１のキャッシュ内に資源の第１のコピーを保持する一方で資源の第２のコピーを前記第１のキャッシュから第２のキャッシュへと転送し、
前記障害の発生したキャッシュが資源の最新バージョンを保持していたかどうかを決定

し、
もし前記障害の発生したキャッシュが資源の最新バージョンを保持していたならば、前記障害の発生したキャッシュの復旧ログを資源の以前のバージョンに適用して資源の最新バージョンを再構築するように構成される、装置。
【請求項１１８】
前記資源の以前のバージョンはディスク上に永続的に記憶される、請求項１１７に記載の装置。
【請求項１１９】
前記資源の以前のバージョンは第３のキャッシュ内の資源のパストイメージである、請求項１１７に記載の装置。
【請求項１２０】
前記第２のキャッシュ内の資源の前記パストイメージは、障害の発生しなかった複数のキャッシュのうちいずれかのキャッシュ内に現在保持されている資源の他のいずれかのパストイメージと少なくとも同じほど最新のものである、請求項１１９記載の装置。
【請求項１２１】
前記障害の発生したキャッシュの復旧ログは、永続的記憶装置に前記資源の以前のバージョンを最初に永続的に記憶することなしに、前記第３のキャッシュ内の資源の以前のバージョンに適用される、請求項１１９に記載の装置。
【請求項１２２】
資源のダーティバージョンを保持する複数のキャッシュの障害の後に前記資源を復旧させるための装置であって、前記装置は、
第１のキャッシュから永続的記憶装置に前記資源を最初に永続的に記憶することなしに、前記第１のキャッシュ内に資源の第１のコピーを保持する一方で資源の第２のコピーを前記第１のキャッシュから第２のキャッシュへと転送し、
前記障害の発生したキャッシュのいずれかが資源の最新バージョンを保持していたかどうかを決定し、
もし前記障害の発生したキャッシュのいずれかが資源の最新バージョンを保持していたならば、
前記障害の発生したキャッシュの復旧ログをマージし、
前記マージされた復旧ログを資源の以前のバージョンに適用して資源の最新バージョンを再構築するように構成される、装置。
【請求項１２３】
前記資源の以前のバージョンはディスク上に永続的に記憶される、請求項１２２に記載の装置。
【請求項１２４】
前記資源の以前のバージョンは、前記障害の発生した複数のキャッシュに含まれない第３のキャッシュ内の資源のパストイメージである、請求項１２２に記載の装置。
【請求項１２５】
前記第３のキャッシュ内の資源の前記パストイメージは、前記障害の発生した複数のキャッシュに含まれないいずれかのキャッシュ内に現在保持されている資源の他のいずれかのパストイメージと少なくとも同じほど最新のものである、請求項１２４に記載の装置。
【請求項１２６】
前記マージされた復旧ログは、永続的記憶装置に前記資源を最初に永続的に記憶することなしに、前記第３のキャッシュ内の資源の以前のバージョンに適用される、請求項１２４に記載の装置。
【請求項１２７】
複数のノードが使用する資源を管理するための方法であって、
前記複数のノードのうち第１のノードから資源の要求を受信するステップを含み、前記第１のノードは第１のキャッシュを含み、さらに、
前記複数のノードのうち第２のノードを識別するステップを含み、前記第２のノードは

資源の第１のコピーを有する第２のキャッシュを含み、さらに、
前記第２のノードが、前記第２のキャッシュから永続的記憶装置に前記資源を最初に永続的に記憶することなしに、資源の第２のコピーを前記第２のキャッシュから前記第１のキャッシュに転送するようにするステップと、
資源の前記第１のコピーまたはそのサクセサが永続的に記憶されるまで前記第２のキャッシュ内に資源の少なくとも１つのコピーが保持されるようにするステップとを含む、方法。
【請求項１２８】
前記第１のノードは第１のデータベースサーバであり、前記第２のノードは第２のデータベースサーバである、請求項１２７に記載の方法。
【請求項１２９】
前記複数のノードのうち第３のノードからさらに許可の要求を受信するステップをさらに含み、前記第３のノードは第３のキャッシュ内において資源の第３のコピーを含み、前記許可は前記第３のノードが前記資源をディスクに書込むことを許可するが前記第３のノードが前記資源を変更することは許可せず、さらに、
前記複数のノードのうち第４のノードを識別するステップを含み、前記第４のノードは第４のキャッシュ内において資源の第４のコピーを含み、前記第４のコピーは前記第３のコピーと少なくとも同じほど最近のものであり、さらに、
前記第４のノードが前記第４のコピーをディスクに書込むようにするステップを含む、請求項１２７に記載の方法。
【請求項１３０】
前記第４のノードが前記第４のコピーをディスクに書込むことに応答して、前記第４のコピーよりも古い前記資源のすべてのコピーが開放されるようにするステップをさらに含む、請求項１２９に記載の方法。
【請求項１３１】
前記第４のノードは前記第３のノードである、請求項１２９に記載の方法。
【請求項１３２】
前記第４のノードは前記第３のノードでない、請求項１２９に記載の方法。
【請求項１３３】
前記資源の前記第４のコピーは資源のパストイメージである、請求項１２９に記載の方法。
【請求項１３４】
前記資源のパストイメージは、前記複数のノードが現在保持している資源の他のいずれかのパストイメージと少なくとも同じほど最近のものである、請求項１３３に記載の方法。
【請求項１３５】
前記資源の前記第４のコピーは資源のカレントバージョンである、請求項１２９に記載の方法。
【請求項１３６】
前記第４のノードが前記第４のコピーをディスクに書込むようにするステップは、
前記第４のノードに、前記第４のノードが前記第４のコピーをディスクに書込むことを許可するさらなる許可を与えるステップを含み、前記さらなる許可は前記第４のノードが前記第４のコピーを変更することを許可しない、請求項１２９に記載の方法。
【請求項１３７】
前記第４のノードに与えられる前記さらなる許可は、前記第４のノードが前記第４のコピーをディスクに書込むことを許可する書込ロックであり、前記書込ロックは、前記第４のノードが前記第４のコピーを変更することは許可しない、請求項１３６に記載の方法。
【請求項１３８】
前記第４のノードに与えられる前記さらなる許可は、前記第４のノードが前記第４のコピーをディスクに書込むことを許可する書込トークンであり、前記書込トークンは、前記

第４のノードが前記第４のコピーを変更することは許可しない、請求項１３６に記載の方法。

【請求項１３９】

前記第４のノードが前記第４のコピーをディスクに書込むようにするステップは、

状態を前記第４のコピーに関連付けるステップを含み、前記状態は前記第４のノードが前記第４のコピーをディスクに書込むことを許可する、請求項１２９に記載の方法。

【請求項１４０】

前記要求を受信する前に、前記第２のノードに、前記第２のノードが前記第１のコピーを変更することを許可する第１の許可を与えるステップをさらに含み、前記第１の許可は前記第２のノードが前記第１のコピーをディスクに書込むことは許可せず、さらに、

前記第２のノードが前記第２のコピーを前記第１のキャッシュに転送するようにする前に、前記第２のノードに、前記第２のノードが前記第２のコピーを保持することを要求する第２の許可を与えるステップを含み、前記第２の許可は前記第２のノードが前記第２のコピーを変更することを許可せずかつ前記第２のノードが前記第２のコピーをディスクに書込むことを許可しない、請求項１２７に記載の方法。

【請求項１４１】

前記第２のノードに与えられる前記第１の許可は、前記第２のノードが前記第１のコピーを変更することを許可する変更ロックであり、

前記変更ロックは前記第２のノードが前記第１のコピーをディスクに書込むことは許可せず、

前記第２の許可は前記第２のノードが前記第２のコピーを保持することを要求する保持ロックであり、

前記保持ロックは前記第２のノードが前記第２のコピーを変更することを許可せずかつ前記第２のノードが前記第２のコピーをディスクに書込むことを許可しない、請求項１４０に記載の方法。

【請求項１４２】

前記第２のノードに与えられる前記第１の許可は、前記第２のノードが前記第１のコピーを変更することを許可する変更トークンであり、

前記変更トークンは前記第２のノードが前記第１のコピーをディスクに書込むことは許可せず、

前記第２の許可は前記第２のノードが前記第２のコピーを保持することを要求する保持トークンであり、

前記保持トークンは前記第２のノードが前記第２のコピーを変更することを許可せずかつ前記第２のノードが前記第２のコピーをディスクに書込むことを許可しない、請求項１４０に記載の方法。

【請求項１４３】

前記要求を受信する前に、前記第１のコピーを、前記第２のノードが前記第１のコピーを変更することを許可しかつ前記第２のノードが前記第１のコピーをディスクに書込むことを許可する第１の状態と関連付けるステップと、

前記第２のノードが前記第２のコピーを前記第１のキャッシュに転送するようにする前に、前記第１のコピーを、前記第２のノードが前記第１のコピーをディスクに書込むことを許可する第２の状態と関連付けるステップとをさらに含み、前記第２の状態は前記第２のノードが前記第１のコピーを変更または上書きすることを許可しない、請求項１２７に記載の方法。

【請求項１４４】

前記第１の状態は、前記第２のノードが前記第１のコピーを変更することを許可しかつ前記第２のノードが前記第１のコピーをディスクに書込むことを許可する現在の状態であり、

前記第２の状態は、前記第２のノードが前記第１のコピーをディスクに書込むことを許可するパストイメージ書込可能状態であり、前記パストイメージ書込可能状態は前記第２

のノードが前記第１のコピーを変更または上書きすることを許可しない、請求項１４３に記載の方法。

【請求項１４５】

前記第２のノードが変更許可を前記第２のノードから前記第１のノードに転送するようにするステップをさらに含む、請求項１２７に記載の方法。

【請求項１４６】

前記第２のノードから、前記第２のノードが前記変更許可を前記第１のノードに転送したというメッセージを受信するステップをさらに含む、請求項１４５に記載の方法。

【請求項１４７】

前記第２のノードが前記変更許可を前記第１のノードに転送したという前記メッセージは、前記第２のノードが前記変更許可を前記第１のノードに転送した後に受信される、請求項１４６に記載の方法。

【請求項１４８】

前記第１のノードが前記変更許可を保持していることを示すデータを記憶するステップをさらに含む、請求項１４６に記載の方法。

【請求項１４９】

複数のノードの共有資源を管理するための方法であって、

前記共有資源に変更許可を与えるのとは別に、前記共有資源に書込許可を与えるステップを含み、

前記書込許可を保持することは、前記書込許可を保持するノードが、ディスクに、前記書込許可を保持するノードのキャッシュに保持される共有資源のコピーを書込むことを許可し、前記書込許可を保持することは、前記書込許可を保持するノードが、前記書込許可を保持するノードのキャッシュに保持される共有資源のコピーを変更することは許可せず、

前記変更許可を保持することは、前記変更許可を保持するノードが、前記変更許可を保持するノードのキャッシュに保持される共有資源のコピーを変更することを許可し、前記変更許可を保持することは、前記変更許可を保持するノードが、前記変更許可を保持するノードのキャッシュに保持される共有資源のコピーを書込むことは許可しない、方法。

【請求項１５０】

前記共有資源についての前記書込許可は前記複数のノードのうち１つのノードのみに時を選ばず与えられ、前記共有資源についての前記変更許可は前記複数のノードのうち１つのノードのみに時を選ばず与えられる、請求項１４９に記載の方法。

【請求項１５１】

前記書込許可および前記変更許可は前記複数のノードのうち異なるノードによって保持される、請求項１４９に記載の方法。

【請求項１５２】

前記書込許可および前記変更許可は前記複数のノードのうち同一のノードによって保持される、請求項１４９に記載の方法。

【請求項１５３】

前記共有資源に前記書込許可を与えるのとは別にかつ前記共有資源に前記変更許可を与えるのとは別に、前記共有資源に保持許可を与えるステップをさらに含み、

前記保持許可を保持することは、前記保持許可を保持するノードが、前記保持許可を保持するノードのキャッシュに保持される資源のコピーを保持することを要求する、請求項１４９に記載の方法。

【請求項１５４】

前記複数のノードのうち１組のノードの各ノードに対し、前記共有資源に前記書込許可を与えるのとは別にかつ前記共有資源に前記変更許可を与えるのとは別に、前記共有資源に保持許可を与えるステップをさらに含み、

前記保持許可を保持することは、前記保持許可を保持する１組のノードの各ノードが、前記保持許可を保持する１組のノードの各ノードのキャッシュに保持される共有資源のコ

ピーを保持することを要求する、請求項１４９に記載の方法。
【請求項１５５】
前記保持許可を保持することは、前記保持許可を保持するノードが、前記保持許可を保持するノードのキャッシュに保持される共有資源のコピーまたはそのサクセサが永続的に記憶されるまで、前記保持許可を保持するノードのキャッシュに保持される共有資源のコピーを保持することを要求する、請求項１５３に記載の方法。
【請求項１５６】
命令の１つ以上のシーケンスを搬送して複数のノードが使用する資源を管理するためのコンピュータ読出可能媒体であって、前記命令の１つ以上のシーケンスが１つ以上のプロセッサによって実行されると、前記１つ以上のプロセッサは、
前記複数のノードのうち第１のノードから資源の要求を受信するステップを実行し、前記第１のノードは第１のキャッシュを含み、さらに、
前記複数のノードのうち第２のノードを識別するステップを実行し、前記第２のノードは資源の第１のコピーを有する第２のキャッシュを含み、さらに、
前記第２のノードが、前記第２のキャッシュから永続的記憶装置に前記資源を最初に永続的に記憶することなしに、資源の第２のコピーを前記第２のキャッシュから前記第１のキャッシュに転送するようにするステップと、
資源の前記第１のコピーまたはそのサクセサが永続的に記憶されるまで前記第２のキャッシュ内に資源の少なくとも１つのコピーが保持されるようにするステップとを実行する、コンピュータ読出可能媒体。
【請求項１５７】
前記第１のノードは第１のデータベースサーバであり、前記第２のノードは第２のデータベースサーバである、請求項１５６に記載のコンピュータ読出可能媒体。
【請求項１５８】
前記１つ以上のプロセッサによって実行されると、前記１つ以上のプロセッサが、
前記複数のノードのうち第３のノードからさらに許可の要求を受信するステップを実行するようにする命令をさらに含み、前記第３のノードは第３のキャッシュ内において資源の第３のコピーを含み、前記許可は前記第３のノードが前記資源をディスクに書込むことを許可するが前記第３のノードが前記資源を変更することは許可せず、さらに、
前記複数のノードのうち第４のノードを識別するステップを実行するようにする命令を含み、前記第４のノードは第４のキャッシュ内において資源の第４のコピーを含み、前記第４のコピーは前記第３のコピーと少なくとも同じほど最近のものであり、さらに、
前記第４のノードが前記第４のコピーをディスクに書込むようにするステップを実行するようにする命令を含む、請求項１５６に記載のコンピュータ読出可能媒体。
【請求項１５９】
前記１つ以上のプロセッサによって実行されると、前記１つ以上のプロセッサが、
前記第４のノードが前記第４のコピーをディスクに書込むことに応答して、前記第４のコピーよりも古い前記資源のすべてのコピーが開放されるようにするステップを実行するようにする命令をさらに含む、請求項１５８に記載のコンピュータ読出可能媒体。
【請求項１６０】
前記第４のノードは前記第３のノードである、請求項１５８に記載のコンピュータ読出可能媒体。
【請求項１６１】
前記第４のノードは前記第３のノードでない、請求項１５８に記載のコンピュータ読出可能媒体。
【請求項１６２】
前記資源の前記第４のコピーは資源のパストイメージである、請求項１５８に記載のコンピュータ読出可能媒体。
【請求項１６３】
前記資源のパストイメージは、前記複数のノードが現在保持している資源の他のいずれ

かのパストイメージと少なくとも同じほど最近のものである、請求項１６２に記載のコンピュータ読出可能媒体。

【請求項１６４】

前記資源の前記第４のコピーは資源のカレントバージョンである、請求項１５８に記載のコンピュータ読出可能媒体。

【請求項１６５】

前記第４のノードが前記第４のコピーをディスクに書込むようにするための命令は、前記１つ以上のプロセッサによって実行されると、前記１つ以上のプロセッサが、

前記第４のノードに、前記第４のノードが前記第４のコピーをディスクに書込むことを許可するさらなる許可を与えるステップを実行するようにする命令を含み、前記さらなる許可は前記第４のノードが前記第４のコピーを変更することを許可しない、請求項１５８に記載のコンピュータ読出可能媒体。

【請求項１６６】

前記第４のノードに与えられる前記さらなる許可は、前記第４のノードが前記第４のコピーをディスクに書込むことを許可する書込ロックであり、前記書込ロックは、前記第４のノードが前記第４のコピーを変更することは許可しない、請求項１６５に記載のコンピュータ読出可能媒体。

【請求項１６７】

前記第４のノードに与えられる前記さらなる許可は、前記第４のノードが前記第４のコピーをディスクに書込むことを許可する書込トークンであり、前記書込トークンは、前記第４のノードが前記第４のコピーを変更することは許可しない、請求項１６５に記載のコンピュータ読出可能媒体。

【請求項１６８】

前記第４のノードが前記第４のコピーをディスクに書込むようにするための命令は、前記１つ以上のプロセッサによって実行されると、前記１つ以上のプロセッサが、

状態を前記第４のコピーに関連付けるステップを実行するようにする命令を含み、前記状態は前記第４のノードが前記第４のコピーをディスクに書込むことを許可する、請求項１５８に記載のコンピュータ読出可能媒体。

【請求項１６９】

前記１つ以上のプロセッサによって実行されると、前記１つ以上のプロセッサが、

前記要求を受信する前に、前記第２のノードに、前記第２のノードが前記第１のコピーを変更することを許可する第１の許可を与えるステップを実行するようにする命令をさらに含み、前記第１の許可は前記第２のノードが前記第１のコピーをディスクに書込むことは許可せず、さらに、

前記第２のノードが前記第２のコピーを前記第１のキャッシュに転送するようにする前に、前記第２のノードに、前記第２のノードが前記第２のコピーを保持することを要求する第２の許可を与えるステップを実行するようにする命令を含み、前記第２の許可は前記第２のノードが前記第２のコピーを変更することを許可せずかつ前記第２のノードが前記第２のコピーをディスクに書込むことを許可しない、請求項１５６に記載のコンピュータ読出可能媒体。

【請求項１７０】

前記第２のノードに与えられる前記第１の許可は、前記第２のノードが前記第１のコピーを変更することを許可する変更ロックであり、

前記変更ロックは前記第２のノードが前記第１のコピーをディスクに書込むことは許可せず、

前記第２の許可は前記第２のノードが前記第２のコピーを保持することを要求する保持ロックであり、

前記保持ロックは前記第２のノードが前記第２のコピーを変更することを許可せずかつ前記第２のノードが前記第２のコピーをディスクに書込むことを許可しない、請求項１６９に記載のコンピュータ読出可能媒体。

【請求項１７１】
前記第２のノードに与えられる前記第１の許可は、前記第２のノードが前記第１のコピーを変更することを許可する変更トークンであり、
前記変更トークンは前記第２のノードが前記第１のコピーをディスクに書込むことは許可せず、
前記第２の許可は前記第２のノードが前記第２のコピーを保持することを要求する保持トークンであり、
前記保持トークンは前記第２のノードが前記第２のコピーを変更することを許可せずかつ前記第２のノードが前記第２のコピーをディスクに書込むことを許可しない、請求項１６９に記載のコンピュータ読出可能媒体。
【請求項１７２】
前記１つ以上のプロセッサによって実行されると、前記１つ以上のプロセッサが、
前記要求を受信する前に、前記第１のコピーを、前記第２のノードが前記第１のコピーを変更することを許可しかつ前記第２のノードが前記第１のコピーをディスクに書込むことを許可する第１の状態と関連付けるステップと、
前記第２のノードが前記第２のコピーを前記第１のキャッシュに転送するようにする前に、前記第１のコピーを、前記第２のノードが前記第１のコピーをディスクに書込むことを許可する第２の状態と関連付けるステップとを実行するようにする命令をさらに含み、
前記第２の状態は前記第２のノードが前記第１のコピーを変更または上書きすることを許可しない、請求項１５６に記載のコンピュータ読出可能媒体。
【請求項１７３】
前記第１の状態は、前記第２のノードが前記第１のコピーを変更することを許可しかつ前記第２のノードが前記第１のコピーをディスクに書込むことを許可する現在の状態であり、
前記第２の状態は、前記第２のノードが前記第１のコピーをディスクに書込むことを許可するパストイメージ書込可能状態であり、前記パストイメージ書込可能状態は前記第２のノードが前記第１のコピーを変更または上書きすることを許可しない、請求項１７２に記載のコンピュータ読出可能媒体。
【請求項１７４】
前記１つ以上のプロセッサによって実行されると、前記１つ以上のプロセッサが、
前記第２のノードが変更許可を前記第２のノードから前記第１のノードに転送するようにするステップを実行するようにする命令をさらに含む、請求項１５６に記載のコンピュータ読出可能媒体。
【請求項１７５】
前記１つ以上のプロセッサによって実行されると、前記１つ以上のプロセッサが、
前記第２のノードから、前記第２のノードが前記変更許可を前記第１のノードに転送したというメッセージを受信するステップを実行するようにする命令をさらに含む、請求項１７４に記載のコンピュータ読出可能媒体。
【請求項１７６】
前記第２のノードが前記変更許可を前記第１のノードに転送したという前記メッセージは、前記第２のノードが前記変更許可を前記第１のノードに転送した後に受信される、請求項１７５に記載のコンピュータ読出可能媒体。
【請求項１７７】
前記１つ以上のプロセッサによって実行されると、前記１つ以上のプロセッサが、
前記第１のノードが前記変更許可を保持していることを示すデータを記憶するステップを実行するようにする命令をさらに含む、請求項１７５に記載のコンピュータ読出可能媒体。
【請求項１７８】
命令の１つ以上のシーケンスを搬送して複数のノードの共有資源を管理するためのコンピュータ読出可能媒体であって、前記命令の１つ以上のシーケンスが１つ以上のプロセッ

サによって実行されると、前記１つ以上のプロセッサは、
前記共有資源に変更許可を与えるのとは別に、前記共有資源に書込許可を与えるステップを実行し、
前記書込許可を保持することは、前記書込許可を保持するノードが、ディスクに、前記書込許可を保持するノードのキャッシュに保持される共有資源のコピーを書込むことを許可し、前記書込許可を保持することは、前記書込許可を保持するノードが、前記書込許可を保持するノードのキャッシュに保持される共有資源のコピーを変更することは許可せず、
前記変更許可を保持することは、前記変更許可を保持するノードが、前記変更許可を保持するノードのキャッシュに保持される共有資源のコピーを変更することを許可し、前記変更許可を保持することは、前記変更許可を保持するノードが、前記変更許可を保持するノードのキャッシュに保持される共有資源のコピーを書込むことは許可しない、コンピュータ読出可能媒体。
【請求項１７９】
前記共有資源についての前記書込許可は前記複数のノードのうち１つのノードのみに時を選ばず与えられ、前記共有資源についての前記変更許可は前記複数のノードのうち１つのノードのみに時を選ばず与えられる、請求項１７８に記載のコンピュータ読出可能媒体。
【請求項１８０】
前記書込許可および前記変更許可は前記複数のノードのうち異なるノードによって保持される、請求項１７８に記載のコンピュータ読出可能媒体。
【請求項１８１】
前記書込許可および前記変更許可は前記複数のノードのうち同一のノードによって保持される、請求項１７８に記載のコンピュータ読出可能媒体。
【請求項１８２】
前記１つ以上のプロセッサによって実行されると、前記１つ以上のプロセッサが、
前記共有資源に前記書込許可を与えるのとは別にかつ前記共有資源に前記変更許可を与えるのとは別に、前記共有資源に保持許可を与えるステップを実行するようにする命令をさらに含み、
前記保持許可を保持することは、前記保持許可を保持するノードが、前記保持許可を保持するノードのキャッシュに保持される資源のコピーを保持することを要求する、請求項１７８に記載のコンピュータ読出可能媒体。
【請求項１８３】
前記１つ以上のプロセッサによって実行されると、前記１つ以上のプロセッサが、
前記複数のノードのうち１組のノードの各ノードに対し、前記共有資源に前記書込許可を与えるのとは別にかつ前記共有資源に前記変更許可を与えるのとは別に、前記共有資源に保持許可を与えるステップを実行するようにする命令をさらに含み、
前記保持許可を保持することは、前記保持許可を保持する１組のノードの各ノードが、前記保持許可を保持する１組のノードの各ノードのキャッシュに保持される共有資源のコピーを保持することを要求する、請求項１７８に記載のコンピュータ読出可能媒体。
【請求項１８４】
前記保持許可を保持することは、前記保持許可を保持するノードが、前記保持許可を保持するノードのキャッシュに保持される共有資源のコピーまたはそのサクセサが永続的に記憶されるまで、前記保持許可を保持するノードのキャッシュに保持される共有資源のコピーを保持することを要求する、請求項１８２に記載のコンピュータ読出可能媒体。
【請求項１８５】
複数のノードの共有資源を管理するためのシステムであって、
前記共有資源に変更許可を与えるのとは別に、前記共有資源に書込許可を与えるためのメカニズムを含み、
前記書込許可は、前記書込許可を保持するノードが、ディスクに、前記書込許可を保持

するノードのキャッシュに保持される資源のコピーを書込むことを許可し、前記書込許可は、前記書込許可を保持するノードが、前記書込許可を保持するノードのキャッシュに保持される資源のコピーを変更することは許可せず、

変更許可は、前記変更許可を保持するノードが，前記変更許可を保持するノードのキャッシュに保持される資源のコピーを変更することを許可し、前記変更許可は、前記変更許可を保持するノードが、前記変更許可を保持するノードのキャッシュに保持される資源のコピーを書込むことは許可しない、システム。

【請求項１８６】

前記メカニズムはさらに、前記複数のノードに書込許可を１つのみ１度だけ与え、前記複数のノードに変更許可を１つのみ１度だけ与えるように構成される、請求項１８５に記載のシステム。

【請求項１８７】

前記メカニズムはさらに保持許可を与えるように構成され、前記保持許可は、前記保持許可を保持するノードが、前記保持許可を保持するノードの第３のキャッシュに保持される資源の第３のコピーを保持することを要求する、請求項１８５に記載のシステム。

【請求項１８８】

前記保持許可は、前記保持許可を保持するノードが、前記資源の第３のコピーまたはそのサクセサが永続的に記憶されるまで前記資源の第３のコピーを保持することを要求する、請求項１８７に記載のシステム。

【請求項１８９】

前記メカニズムはさらに複数の保持許可を与えるように構成され、前記複数の保持許可の各保持許可は、前記複数のノードのうち１つのノードに与えられ、各保持許可は、前記複数の保持許可のうち１つを保持するノードが、前記複数の保持許可のうち１つを保持するノードのキャッシュに保持される資源の別のコピーを保持することを要求する、請求項１８５に記載のシステム。

【請求項１９０】

前記複数の保持許可の各保持許可は、前記複数の保持許可のうち１つを保持するノードが、資源の他のコピーを、前記資源の他のコピーまたはそのサクセサが永続的に記憶されるまで保持することを要求する、請求項１８９に記載のシステム。

【発明の詳細な説明】

【０００１】

【発明の分野】

この発明は、あるノードがデータストアからデータを要求するとき要求されたデータの最も最近のバージョンが別のノードのキャッシュ内にあるときことに関連付けられるペナルティを低減するための技術に関する。

【０００２】

【発明の背景】

スケーラビリティを向上させるため、データベースシステムの中には、（各々が別個に稼動する）２つ以上のデータベースサーバがディスクメディア上に記憶されるなど、共有の記憶装置に同時にアクセスすることを可能にするものがある。各データベースサーバは、ディスクブロックなどの、共有資源をキャッシュするためのキャッシュを有する。そのようなシステムをここではパラレルサーバシステムと呼ぶ。

【０００３】

パラレルサーバシステムに関連付けられる問題の１つに、「ピング」と呼ばれるものの可能性がある。あるサーバのキャッシュ内にある資源のバージョンが異なったサーバのキャッシュに与えられなければならないとき、ピングは起きる。したがって、ピングが起きるのは、データベースサーバＡがそのキャッシュ内の資源ｘを変更した後、データベースサーバＢが資源ｘの変更を要求するときである。データベースサーバＡおよびＢは、典型的には、異なったノード上で稼動するが、場合によっては同じノード上で稼動することもあり得る。

【０００４】
ピングを処理するアプローチの１つを、ここでは「ディスク介入」アプローチと呼ぶ。ディスク介入アプローチは、中間的記憶装置としてディスクを使用して２つのキャッシュ間で資源の最新バージョンを転送する。したがって、上記の例では、ディスク介入アプローチは、データベースサーバ１が資源Ｘのそのキャッシュバージョンをディスクに書込み、データベースサーバ２がこのバージョンをディスクからそのキャッシュへと検索することを必要とする。ディスク介入アプローチは、資源のサーバ間転送ごとに２ディスクＩ／Ｏを必要とするので、パラレルサーバシステムのスケーラビリティが制限される。具体的には、ピングを処理するために必要とされるディスクＩ／Ｏは、比較的不経済で時間がかかり、システムに加えられるデータベースサーバの数が多ければ多いほど、ピングの数も多くなる。

【０００５】
しかしながら、ディスク介入アプローチは、単一のデータベースサーバの障害からの比較的効率よい復旧を提供する、というのもそのような復旧が必要とするのは障害の発生したデータベースサーバの復旧（再実行）ログを適用するだけであるからである。障害の発生したデータベースサーバの再実行ログを適用すると、障害の発生したデータベースサーバ上のトランザクションが障害の発生したサーバのキャッシュ内の資源に加えた、かかわった変更はすべて確実に復旧される。復旧の間の再実行ログの使用は、１９９７年１月２１日出願の「復旧可能オブジェクト内のキャッシングデータ」（“CACHING DATA IN RECOVERABLE OBJECTS”）と題する米国特許出願連続番号第０８／７８４，６１１号に詳細に記載される。

【０００６】
ディスク介入アプローチを採用するパラレルサーバシステムは、典型的には、資源アクセスおよび変更に関するグローバルな調停のすべてが分散ロックマネージャー（ＤＬＭ）によって行なわれる場合のプロトコルを使用する。例示的ＤＬＭの動作は、１９９６年６月２４日出願の「ロックキャッシングのための方法および装置」（“METHOD AND APPARATUS FOR LOCK CACHING”）と題する米国特許出願連続番号第０８／６６９，６８９号に詳細に記載され、その内容はここに引用により援用される。

【０００７】
典型的な分散ロックマネージャーシステムでは、任意の所与の資源に属する情報は、資源に対応するロックオブジェクト内に記憶される。各ロックオブジェクトは、単一のノードのメモリ内に記憶される。ロックオブジェクトが記憶されているノード上にあるロックマネージャーは、そのロックオブジェクトおよびそれがカバーする資源のマスタと呼ばれる。

【０００８】
ピングを処理するためにディスク介入アプローチを採用するシステムでは、ピングは、さまざまなロックが関係する通信においてＤＬＭを必要とする。具体的には、データベースサーバ（「要求サーバ」）が資源のアクセスを必要とするとき、データベースサーバは、それが適切なモード、すなわち、読出の場合には共有され、書込の場合には排他的であるモードにロックされた所望の資源を有するかどうかをチェックする。もし要求データベースサーバが正しいモードにロックされた所望の資源を有していなければ、または、資源に全くロックがされていなければ、要求サーバは、資源のマスタに要求を送信して特定のモードのロックを獲得する。

【０００９】
要求データベースサーバによってなされた要求は、資源の現在の状態と競合することがある（たとえば、別のデータベースサーバが資源に対する排他的なロックを現在保持している可能性がある）。もし競合がなければ、資源のマスタは、ロックを許可し許可を登録する。競合の場合には、資源のマスタは、競合解決プロトコルを開始する。資源のマスタは、競合するロックを保持するデータベースサーバ（「ホルダ」）に、下位の互換性のあるモードにそのロックをダウングレードするよう命令する。

【００１０】
不幸にも、もしホルダ（たとえば、データベースサーバＡ）が所望の資源の更新された（「ダーティ」）バージョンをそのキャッシュ内に現在持っていなければ、それはそのロックを即座にダウングレードできない。そのロックをダウングレードするために、データベースサーバＡは、「ハードピング」プロトコルと呼ばれるものを経る。ハードピングプロトコルに従って、データベースサーバＡは、ディスクに書込まれるべき更新に関連付けられる再実行ログを強制し、資源をディスクに書込み、そのロックをダウングレードし、データベースサーバＡが完了したことをマスタに通知する。通知を受取ると、マスタは、ロック許可を登録し、要求されたロックが許可されたことを要求サーバに通知する。この時点で、要求サーバＢは、ディスクからそのキャッシュ内に資源を読出す。

【００１１】
上述したとおり、ディスク介入アプローチによっては、あるデータベースサーバによって更新された資源（「ダーティ資源」）を別のデータベースサーバに直接発送することはできない。そのような直接発送は、復旧に関連する問題のために、実行可能性がないと考えられる。たとえば、資源がデータベースサーバＡで変更されてから、データベースサーバＢに直接発送されたと仮定する。データベースサーバＢでも、資源は変更され、データベースサーバＡに発送し返される。データベースサーバＡで、資源は３度目に変更される。各サーバが、別のサーバに資源を送る前にすべての再実行ログをディスクに記憶することによって、受信側が先の変更可能となると仮定する。

【００１２】
３度目の更新の後に、データベースサーバＡがだめになったと仮定する。データベースサーバＡのログは、穴のあいた資源への変更のレコードを含む。具体的には、サーバＡのログは、データベースサーバＢによってなされたこれらの変更を含んでいない。正確には、サーバＢによってなされた変更は、データベースサーバＢのログ内に記憶されている。この時点で、資源を復旧するために、２つのログは適用される前にマージされなければならない。このログマージ動作は、もし実現されれば、障害の発生しなかったデータベースサーバを含む、データベースサーバの総数に比例して時間および資源を必要とするであろう。

【００１３】
上述したディスク介入アプローチは、障害の後の復旧ログのマージに関連付けられる問題を回避するが、簡単で効率のよい復旧を支持する定常状態のパラレルサーバシステムの性能にペナルティを科す。直接発送アプローチは、ディスク介入アプローチに関連付けられるオーバーヘッドを回避するが、障害の発生した場合に複雑で非スケーラブルな復旧動作を伴う。

【００１４】
以上に基づいて、復旧動作の複雑性または持続時間を激しく増大させることなしに、ピングに関連付けられるオーバーヘッドを低減するためのシステムおよび方法を提供することが明らかに望まれる。

【００１５】
【発明の概要】
最初に資源をディスクに書込むことなしに、あるデータベースサーバのキャッシュから別のデータベースサーバのキャッシュへと資源を転送するための方法および装置が提供される。データベースサーバ（要求者）が資源を変更したい場合、要求者は資源の現在のバージョンを要求する。現在のバージョンを有するデータベースサーバ（ホルダ）は現在のバージョンを要求者に直接発送する。バージョンを発送すると、ホルダは資源を変更する許可を失うが、メモリ内に資源のコピーを引続き保持する。資源の保持されたバージョンまたはその後のバージョンがディスクに書込まれると、ホルダは資源の保持されたバージョンを廃棄することができる。他の態様では、ホルダは保持されたバージョンを廃棄しない。サーバ障害の場合には、障害の発生したサーバの再実行ログ内の変更のあったすべての資源の先のコピーを、必要に応じて、障害の発生したサーバの再実行ログを適用するため

の開始点として使用する。この技術を用いて、単一のサーバ障害（障害の最もよくある形態）は、資源へアクセスしていなかったさまざまなデータベースサーバの復旧ログをマージする必要なしに、復旧される。

【００１６】

この発明は、同じ参照番号が同様の要素を指している添付の図面に例として示されるが、これに限定されるものでない。

【００１７】

【好ましい発明の詳細な説明】

ピングに関連付けられるオーバーヘッドを低減するための方法および装置を記載する。以下の記載では、説明のため、この発明を完全に理解するために、多くの具体的な詳細を述べる。しかしながら、この発明がこれらの具体的な詳細なしに実施可能であることは当業者には明らかであろう。他のデータベースサーバでは、この発明を不要にわかりにくくすることを避けるために、周知の構造および装置がブロック図の形で示される。

【００１８】

機能概要

この発明のある局面に従うと、最初にディスクに記憶することなしに、データベースサーバ間で直接資源の更新されたバージョンを発送することによってピングを処理し、これによってディスク介入アプローチに関連付けられるＩ／Ｏオーバーヘッドを回避する。さらに、単一の場合の障害復旧に関連付けられる問題は、資源が別のキャッシュに転送されたとしても、変更された資源またはその何らかのサクセサがディスクに書込まれるまで資源の変更されたバージョンがキャッシュ内で置換されることを防ぐことによって、回避される。

【００１９】

説明のため、キャッシュ内で置換不可能である資源のコピーを、ここでは「留められた」資源と呼ぶ。留められた資源を置換可能にする動作を、資源を「解放する」と呼ぶ。

【００２０】

ＭおよびＷロックアプローチ

この発明のある局面に従うと、資源に対する変更許可とディスクへの書込許可とは分離される。したがって、キャッシュからディスクへ資源の更新されたバージョンを書込む許可を有するデータベースサーバが、必ずしも資源を更新する許可を有するとは限らない。逆に、資源のキャッシュされたバージョンを変更する許可を有するデータベースサーバが、そのキャッシュされたバージョンをディスクに書込む許可を有するとは限らない。

【００２１】

ある実施例に従うと、許可のこの分離は、特殊なロックを使用することによって実施される。具体的には、資源を変更する許可は、「Ｍ」ロックによって与えられるであろうし、ディスクに資源を書込む許可は、「Ｗ」ロックによって与えられるであろう。しかしながら、ここに記載するようなＭロックおよびＷロックの使用は、資源の転送されたバージョンが、その資源またはそのサクセサがディスクに書込まれるまでキャッシュ内で置換されることを防ぐためのメカニズムの１つにすぎないことが注目される。

【００２２】

図２を参照すると、この発明のある実施例に従って、ＭロックおよびＷロックを使用するデータシステムにおいてピングに応答して実行されるステップを示す。ステップ２００で、資源を変更したいデータベースサーバは、資源のマスタ（すなわち、資源のロックを管理するデータベースサーバ）にＭロックを要求する。ステップ２０２で、マスタは、資源のＭロックを現在保持するデータベースサーバ（「ホルダ」）に、２つのサーバを接続する通信チャネル（「相互接続」）を介する直接転送によって、Ｍロックを資源のそのキャッシュされたバージョンとともに転送するよう命令する。

【００２３】

ステップ２０４で、ホルダは、資源の現在のバージョンおよびＭロックを要求者に送る。ステップ２０６で、ホルダは、Ｍロックの転送をマスタに通知する。ステップ２０８で、

マスタは、資源のロック情報を更新して要求者がＭロックを現在保持していることを示す。

【００２４】

ＰＩ資源

Ｍロックのホルダは、必ずしもＷロックを有しているとは限らず、このためそのキャッシュ内に含まれる資源のバージョンをディスクに書出す許可を有していない可能性がある。したがって、転送データベースサーバ（すなわち、Ｍロックを最後に保持していたデータベースサーバ）は、未来のある時点でそのバージョンをディスクに書出すよう要求される可能性があるので、資源のそのバージョンをダイナミックメモリに留め続ける。転送データベースサーバ内に留まる資源のバージョンは、もし受信データベースサーバが資源のそのコピーを変更すれば、古くなる。転送データベースサーバは、受信データベースサーバ（またはそのサクセサ）が資源をいつ変更するかわかっているとは限らないので、転送データベースサーバは資源のコピーを送信した時点から、その保持されたバージョンを「古い可能性のあるデータ」として扱う。資源のそのような古い可能性のあるバージョンを、ここではパストイメージ資源（ＰＩ資源）と呼ぶ。

【００２５】

ＰＩ資源の解放

資源のキャッシュされたバージョンが解放された後、これは新しいデータで上書される可能性がある。典型的には、資源のダーティバージョンは、資源をディスクに書込むことによって解放され得る。しかしながら、キャッシュ内にＰＩ資源を有するデータベースサーバが必ずしも、ＰＩ資源をディスクに記憶する権利を持っているとは限らない。これらの状況下でＰＩ資源を解放するためのある技術が、図３に示される。

【００２６】

図３を参照すると、データベースサーバがそのキャッシュ内のＰＩ資源を解放したいとき、これはＷロックの要求を分散ロックマネージャー（ＤＬＭ）に送信する。ステップ３０２で、ＤＬＭは次に、要求データベースサーバ、または、資源のより新しいバージョン（サクセサ）をそのキャッシュ内に有する何らかのデータベースサーバに、資源をディスクに書出すよう命令する。こうして、資源をディスクに書込むよう命令されたデータベースサーバは、ダブルロックを許可される。ダブルロックを許可されたデータベースサーバか資源をディスクに書込んだ後、データベースサーバはＷロックを解放する。

【００２７】

次に、ＤＬＭは、すべてのデータベースサーバにメッセージを送信して書出された資源のバージョンを示し（ステップ３０４）、この結果、資源のこれ以前のＰＩバージョンはすべて解放可能となる（ステップ３０６）。たとえば、ディスクに書込まれたバージョンが時間Ｔ１０で変更されたと仮定する。それより前の時間Ｔ５で最後に変更された資源のバージョンを有するデータベースサーバは、ここで、これが記憶されているバッファを他のデータのために使用することができるであろう。しかしながら、それより後の時間Ｔ１１で変更されたバージョンを有するデータベースサーバは、資源のそのバージョンをそのメモリ内に保持し続けなければならないであろう。

【００２８】

ＭおよびＷロックアプローチの下でのピング管理

この発明のある実施例に従って、図１を参照して記載するように、ＭおよびＷロックアプローチを実現化してピングを処理してもよい。図１を参照すると、４つのデータベースサーバＡ、Ｂ、ＣおよびＤのブロック図が示され、これらのサーバはすべて特定の資源を含むデータベースへのアクセスを有している。例示される時点では、データベースサーバＡ、ＢおよびＣはすべて、資源のバージョンを有する。データベースサーバＡのキャッシュ内に保持されるバージョンは、資源の最も最近に変更されたバージョンである（時間Ｔ１０で変更された）。データベースサーバＢおよびＣに保持されるバージョンは、資源のＰＩバージョンである。データベースサーバＤは、資源のマスタである。

【００２９】

この時点で、別のデータベースサーバ（「要求者」）が資源を変更したいと仮定する。要求者は、マスタに変更ロックを要求する。マスタは、要求者からの競合する要求のために、データベースサーバＡにコマンドを送信してロック（「ＢＡＳＴ」）をダウンコンバートする。ダウンコンバートコマンドに応答して、資源の現在のイメージ（クリーンまたはダーティのいずれでも）が、データベースサーバＡから要求者に、資源を変更する許可とともに発送される。こうして発送された許可は、資源をディスクに書込む許可を含んでいない。

【００３０】

データベースサーバＡがＭロックを要求者に送ると、データベースサーバＡはそのＭロックを「保持」ロック（「Ｈロック」）にダウングレードする。Ｈロックは、データベースサーバＡが留められたＰＩコピーを保持していることを示す。Ｈロックの所有権は、オーナーにＰＩコピーをそのバッファキャッシュ内に維持することを強制するが、ＰＩコピーをディスクに書込むいかなる権利もそのデータベースサーバに与えない。同じ資源に対して複数の同時的Ｈホルダがあり得るが、一度に資源の書込ができるデータベースサーバは１以下であり、したがって資源のＷロックを保持することのできるデータベースサーバはたった１つである。

【００３１】

資源を発送するより前に、データベースサーバＡはログが確実に強制されるようにする（すなわち、データベースサーバＡによって資源になされた変更について生成された復旧ログが永続的に記憶されるようにする）。変更許可を送ることによって、データベースサーバＡは、資源を変更する自らの権利を失う。資源のコピー（発送の時点ではそうであったような）は、発送データベースサーバＡになおも維持されている。資源の発送の後に、データベースサーバＡ内に保持される資源のコピーは、ＰＩ資源である。

【００３２】

優遇書込

データベースサーバがダーティ資源を別のデータサーバに直接発送した後、資源の保持されたコピーは留められたＰＩ資源となり、解放されるまでそのバッファを別の資源に使用することはできない。ＰＩ資源を含むバッファをここでは、ＰＩバッファと呼ぶ。これらのバッファは、データベースサーバのキャッシュ内に有効な空間を占有しており、やがては他のデータのために再利用されなければならない。

【００３３】

バッファキャッシュ内のＰＩバッファ（古くなったまたはチェックポイントされた）を置換するために、ここでは「優遇書込」と呼ぶ新しいディスク書込プロトコルを採用する。優遇書込プロトコルに従って、データベースサーバが資源をディスクに書込む必要があるとき、データベースサーバは要求をＤＬＭに送信する。ＤＬＭは、ディスクに書込まれるべき資源のバージョンを選択し、選択されたバージョンを有するデータベースサーバを見つけ、書込要求を開始したデータベースサーバに代わって、そのデータベースサーバにディスクへの資源の書込をさせる。資源をディスクに実際に書込むデータベースサーバは、資源の最新の軌跡に依存して、書込を要求したデータベースサーバであっても、または、らかの他のデータベースサーバであってもよい。

【００３４】

資源の選択されたバージョンをディスクに書込むことによって、ディスクに書込まれた選択されたバージョンと同じ古さまたはそれよりも古い、クラスタのすべてのバッファキャッシュ内の資源のＰＩバージョンはすべて解放される。ディスクに書込まれるべきバージョンを選択するために使用される規準を、以下により詳細に記載する。しかしながら、選択されたバージョンは、マスタに知られている最新のＰＩバージョンか、または、資源のカレントバージョン（「ＣＵＲＲ」）のいずれかであり得る。カレントバージョン以外のバージョンを選択する利点の１つは、この別のバージョンの選択によって現在のコピーが妨害されることなく変更可能となることである。

【００３５】

ＰＩ資源を保持しているデータベースサーバは、資源のＷロックを獲得しているならば、そのＰＩコピーを書出すことができる。資源の書込は、さまざまなデータベースサーバ間でのＣＵＲＲ資源イメージの移動から切離される。

【００３６】

効率的要因

資源を別のデータベースサーバに発送するたびにＰＩコピーを書込む必要はない。したがって、資源を永続的に記憶する目的は、ディスクコピーを十分最近のものにしておくことと、バッファキャッシュ内の置換不可能な資源の数を妥当なものにしておくこととである。さまざまな要因が、上述した優遇書込プロトコルを採用するシステムの効率性を決定する。具体的には、

（１） ディスクにダーティ資源を書込むことによって起きるＩ／Ｏ動作を最低限にすることと、

（２） 資源のディスクバージョンを十分に現在のものにしておくことによって障害後の復旧動作を迅速化することと、

（３） 留められたＰＩ資源でバッファキャッシュがオーバーフローすることを防ぐこととが望まれる。

【００３７】

第１の規準を最大化すると第２および第３の規準に否定的影響が及び、その逆もまたある。したがって、トレードオフが必要である。この発明のある実施例に従うと、総ＩＯ経費に対する制御と併せてチェックポイントのさまざまな技術（臨時的継続的チェックポイントと混合されたＬＲＵ）を組合せるセルフチューニングアルゴリズムを使用してもよい。

【００３８】

最新書込アプローチ

上述した優遇書込プロトコルの代替を、ここでは最新書込アプローチと呼ぶ。最新書込アプローチに従うと、すべてのデータベースサーバが、そのＰＩ資源をディスクに書込む許可を有する。しかしながら、そうする前に、データベースサーバは資源のディスクベースのコピーに対するロックを獲得する。ロックを獲得した後、データベースサーバは、ディスクバージョンを、これが書込みたいＰＩバージョンと比較する。もしディスクバージョンの方が古ければ、ＰＩバージョンがディスクに書込まれる。もしディスクバージョンの方が新しければ、ＰＩバージョンは廃棄されてもよく、それが占有していたバッファは再利用可能である。

【００３９】

優遇書込プロトコルと違って、最新書込アプローチは、データベースサーバが、自己のＰＩバージョンをディスクに書込むことによって、またはディスクバージョンの方がより新しいことを決定することによって、自己のＰＩバージョンを解放可能にする。しかしながら、最新書込アプローチは、ディスクベースのコピーのロックに対する競合を増大させ、優遇書込アプローチでは起きなかったであろうディスクＩ／Ｏを招く可能性がある。

【００４０】

許可ストリング

典型的なＤＬＭは、限られた数のロックモードを使用することによって資源へのアクセスを管理し、ここではモードは互換性があるか競合しているかのいずれかである。ある実施例に従うと、資源へのアクセスを管理するメカニズムは、ロックモードを異なった種の許可および義務の集合と代用するよう拡張される。許可および義務は、たとえば、資源を書込み、資源を変更し、キャッシュ内の資源を維持するなどの許可を含んでもよい。具体的な許可および義務を以下により詳細に記載する。

【００４１】

ある実施例に従うと、許可および義務は、許可ストリングに符号化される。多くの許可は資源自体にではなく資源のバージョンに相関するので、許可ストリングは資源バージョン数によって増大するであろう。もし２つの異なった許可ストリングが、資源の同じバージョン（たとえば、変更のための現在バージョンまたは書込のためのディスクアクセス）に

対する同じ排他的許可を要求するならば、これらは競合する。そうでなければこれらは互換性がある。

【００４２】

許可転送を使用する同時実行性

上述したとおり、資源があるデータベースサーバで変更され、別のデータベースサーバによってさらなる変更を要求されると、マスタは、資源のカレントコピー（ＣＵＲＲコピー）を保持するデータベースサーバに、そのＭロック（変更する権利）を資源のＣＵＲＲコピーとともに他のデータベースサーバに送るように命令する。重要なことには、Ｍロックの要求はマスタに送信されるが、許可は何らかの他のデータベースサーバ（先のＭロックホルダ）によってなされる。この三者間メッセージングモデルは、ロック要求が最初にアドレスされたロックマネージャーを含むデータベースサーバからロック要求に対する応答が期待される、従来の双方向通信とはかなり異なる。

【００４３】

この発明のある実施例に従うと、資源のＣＵＲＲコピーのホルダ（たとえば、データベースサーバＡ）がＭロックを別のデータサーバに送ると、データベースサーバＡは、Ｍロックが転送されたことをマスタに通知する。しかしながら、データベースサーバＡは、マスタが通知を受取ったという確認を待つことなく、そのような確認を受取る前にＣＵＲＲコピーおよびＭロックを送信する。待たないことによって、マスタとデータベースサーバＡとの間の往復通信は転送に遅延をもたらすことなく、これによってプロトコルレイテンシがかなり節約される。

【００４４】

許可は許可の現在ホルダから許可の要求者に直接転送されるので、マスタが常に、ロック許可の正確な全体像を知っているとは限らない。むしろ、マスタは、任意の所与の時間でのロックの正確な位置についてではなく、Ｍロックの軌跡についてのみ、「これを最近保持した」データベースサーバについてのみ知っている。ある実施例に従うと、この「レージーな」通知方式は、Ｍロックに適用可能であるが、Ｗロック、Ｘロック、またはＳロック（またはその対応物）には適用可能でない。ロック方式のさまざまな実施例を以下により詳細に記載する。

【００４５】

障害復旧

この発明のコンテクストにおいては、サーバに関連付けられるキャッシュがアクセス不可能となった場合、データベースサーバに障害が発生したという。ここに記載する技術を用いるダーティ資源の直接のサーバ間発送を採用するデータシステムは、単一サーバの障害に応答して復旧ログをマージする必要性を回避する。ある実施例に従って、単一のサーバの障害は、図４に示すとおり処理される。図４を参照して、単一のデータベースサーバに障害が発生すると、復旧プロセスは、障害の発生したデータベースサーバのキャッシュ内に保持される各資源について、以下のステップを実行する。

【００４６】

（ステップ４００）　資源の最新バージョンを保持するデータベースサーバをを決定し、

（ステップ４０２）　ステップ４００で決定されたデータベースサーバが障害の発生したデータベースサーバでなければ、（ステップ４０４）決定されたデータベースサーバは資源のそのキャッシュされたバージョンをディスクに書込み、（ステップ４０６）資源のＰＩバージョンはすべて解放される。このバージョンは、資源に加えられた、かかわった変更（障害の発生したデータベースサーバによってなされたものを含む）を有するため、いかなるデータベースサーバの復旧ログも適用される必要がない。

【００４７】

もしステップ４０２で決定されたデータベースサーバが障害の発生したデータベースサーバであれば、（ステップ４０８）資源の最新ＰＩバージョンを保持するデータベースサーバは、資源のそのキャッシュされたバージョンをディスクに書出し、（ステップ４１０）先のＰＩバージョンはすべて解放される。ディスクに書出されたバージョンは、障害の発

生したデータベースサーバ以外のすべてのデータサーバによって資源に加えられたかかわった変更を有する。障害の発生したデータベースサーバの復旧ログを適用して（ステップ４１２）障害の発生したデータベースサーバによって加えられたかかわった変更を復旧する。

【００４８】

代替的に、資源の最新ＰＩバージョンを、ディスク上ではなくキャッシュ内の現在のバージョンを復旧するための開始点として使用してもよい。具体的には、障害の発生したデータベースサーバの復旧ログから適切なレコードを、キャッシュ内にある最新ＰＩバージョンに直接適用して、最新ＰＩバージョンを保持するデータベースサーバのキャッシュ内のカレントバージョンを再構築してもよい。

【００４９】

複数のデータベースサーバの障害

複数のサーバ障害の場合に、最新ＰＩコピーもいかなるＣＵＲＲコピーも生き残らなかったとき、資源になされた変更が障害の発生したデータベースサーバの複数のログにわたって広がっていることが起こり得る。この状況下では、障害の発生したデータベースサーバのログはマージされなければならない。しかしながら、すべてのデータベースサーバのログではなく、障害の発生したデータベースサーバのログのみがマージされなければならない。したがって、復旧のために必要とされる作業量は、構成全体のサイズにではなく障害の程度に比例する。

【００５０】

どの障害の発生したデータベースサーバが資源を更新したかを決定することが可能なシステムにおいては、資源を更新した障害の発生したデータベースサーバのログのみがマージされ適用される必要がある。同様に、どの障害の発生したデータベースサーバが、資源の永続的に記憶されたバージョンの後に資源を更新したかを決定することのできるシステムにおいては、資源の永続的に記憶されたバージョンの後に資源を更新した、障害の発生したデータベースサーバのログのみがマージされ適用される必要がある。

【００５１】

例示的動作

説明のために、例示的な一連の資源転送を図１を参照して記載する。一連の転送の間、資源は複数のデータベースサーバでアクセスされる。具体的には、資源がクラスタノードに沿って発送され変更されると、データベースサーバの１つでのチェックポイントによってこの資源の物理的Ｉ／Ｏが起こる。

【００５２】

再び図１を参照すると、４つのデータベースサーバ、Ａ、Ｂ、ＣおよびＤがある。データベースサーバＤが資源のマスタである。まずデータベースサーバＣが資源を変更する。データベースサーバＣは資源バージョン８を有する。この時点で、データベースサーバＣは、この資源に対するＭロック（排他的変更権）も有する。

【００５３】

この時点で、データベースサーバＢが、データベースサーバＣが現在保持している資源を変更したいと仮定する。データベースサーバＢは、資源のＭロックの要求（１）を送信する。データベースサーバＤは、資源に関連付けられるモディファィアキュー上に要求を置き、

（ａ）　変更許可（Ｍロック）をデータベースサーバＢに送り、

（ｂ）　資源の現在イメージをデータベースサーバＢに送信し、

（ｃ）　データベースサーバＣのＭロックをＨロックにダウングレードするよう、データベースサーバＣに命令する（メッセージ２：ＢＡＳＴ）。

【００５４】

このダウングレード動作の後に、Ｃは、そのバッファキャッシュ内に資源のそのバージョン（ＰＩコピー）を維持させられる。

【００５５】

データベースサーバＣは、要求された動作を実行し、新しい変更に対してログをさらに強制してもよい。加えて、データベースサーバＣは、これが動作を実行したこと（ＡＳＴ）をマスタにレージーに通知する（３ＡｃｋＭ）。この通知は、データベースサーバＣがバージョン８を維持していることもマスタに知らせる。データベースサーバＣは、マスタからの確認を待たない。従って、データベースサーバＢは、マスタがそれを知る前に、Ｍロックを得ることが可能である。

【００５６】

一方で、データベースサーバＡもまた資源を変更することを決定したとする。データベースサーバＡは、メッセージ（４）をデータベースサーバＤに送信する。このメッセージは、データベースサーバＣからデータベースサーバＤへの非同期の通知の前に、到着し得る。

【００５７】

データベースサーバＤ（マスタ）は、（Ｂがこれを得て変更した後に）資源をデータベースサーバＡに送るよう、データベースサーバＢ、すなわちこの資源の最新と知られているモディファィアにメッセージ（５）を送信する。なお、データベースサーバＤは、資源がそこにあるのかまだなのかを知らない。しかし、データベースサーバＤは、資源がやがてＢに到着することは知っている。

【００５８】

データベースサーバＢが資源を得て意図された変更をした後（現在Ｂは資源のバージョン９を有している）、これは自己のロックをＨにダウングレードし、データベースサーバＡに、資源のカレントバージョン（「ＣＵＲＲ資源」）をＭロックとともに送信する（６）。データベースサーバＢはまた、レージーな通知（６ＡｃｋＭ）をマスタに送信する。

【００５９】

この資源はデータベースサーバＡで変更されつつあるが、データベースサーバＣでのチェックポイントメカニズムが、資源をディスクに書込むことを決定したとする。上記の非同期の事象に関しては、３ＡｃｋＭおよび６ＡｃｋＭの両方が既にマスタに到着していると仮定する。チェックポイント動作に応答して実行された動作を図５を参照して示す。

【００６０】

図５を参照すると、データベースサーバＣは、書込権限を含まない、バージョン８に対するＨロックを保持しているので、データベースサーバＣは、メッセージ１をマスタ（Ｄ）に送信してそのバージョンについてのＷ（書込）ロックを要求する。この時点ではもう、マスタは、（確認が到着したと仮定して）資源がデータベースサーバＡに発送されたことを知っている。データベースサーバＤは、資源書込の命令とともに、（非請求の）Ｗロックをデータベースサーバに送信する（２ＢａｓｔＷ）。

【００６１】

一般的な場合においては、この命令は、送信通知が到着している最新のデータベースサーバへ（または、最新であると知られている送信者から資源を受取ると考えられるデータベースサーバへ）送られる。データベースサーバＡは、資源のそのバージョンを書込む（３）。データベースサーバによって書込まれた資源は、資源のバージョン１０である。このときまでに、もしさらなる要求者が資源を要求していれば、資源のカレントコピーはどこか他にあるであろう。ディスクは、書込が完了したとき確認する（４ＡｃｋＷ）。

【００６２】

書込が完了すると、データベースサーバＡは、データベースサーバＤに、バージョン１０が現在ディスク上にあるという情報を与える（５ＡｃｋＷ）。データベースサーバＡは、（これは最初には要求していなかった）そのＷロックを自発的にダウングレードする。

【００６３】

マスタ（Ｄ）はデータベースサーバＣに行って、要求されたＷロックを許可する代わりに、書込が完成したことをＣに通知する（６）。マスタは、現在のディスクバージョン数を全てのＰＩコピーのホルダに知らせ、これによってＣでのこれ以前のＰＩコピーはすべて解放可能となる。このシナリオでは、データベースサーバＣは、１０より古いＰＩコピー

を有していないので、これはデータベースサーバＣのロックをＮＵＬＬにダウンコンバートする。

【００６４】

マスタはまた、確認メッセージをデータベースサーバＢに送信してデータベースサーバＢに１０より以前のそのＰＩコピーを解放するよう命令する（７ＡｃｋＷ（１０））。

【００６５】

分散ロックマネージャ

従来のＤＬＭ論理と対照的に、ここに記載する直接発送技術を実現するシステムでのマスタは、データベースサーバでのロック状態について不完全な情報を有することがある。ある実施例に従うと、資源のマスタは、以下の情報およびデータ構造を維持する。

【００６６】

（１）　（変更または共有アクセスのいずれかのための）ＣＵＲＲコピー要求者のキュー（キューの長さの上限は、クラスタ内のデータベースサーバの数である）。このキューをここでは、カレント要求キュー（ＣＱ）と呼ぶ。

【００６７】

（２）　資源が別のＣＵＲＲ要求者に送信されると、送信側はレージーに（これが確認を待たないという意味では非同期に）マスタに事象について通知する。マスタは、最新のいくつかの送信者を追跡し続ける。これがＣＱ上のポインタである。

【００６８】

（３）　ディスク上の最新資源バージョンのバージョン数。

（４）　Ｗロック許可およびW要求キュー。

【００６９】

ある実施例に従うと、W許可は同期する。すなわち、これはマスタによってのみ許可され、マスタは、この資源についてのクラスタ内の書込要求者が１以下であることを確実にする。マスタが次の許可を出すことができるのは、先の書込が完了しWロックが解放されたと通知された後のみである。もし２以上のモディファィアがあれば、Wロックは書込の持続時間の間与えられ、書込の後に自発的に解放される。もしモディファィアが１つだけであれば、モディファィアはW許可を維持可能である。

【００７０】

（５）　そのそれぞれの資源バージョン数を備えるＨロックホルダのリスト。これは、バッファキャッシュ内のＰＩコピーについての情報を（おそらく不完全であるが）与える。

【００７１】

ディスクウォームアップ

ここに記載する直接発送は、資源のバッファキャッシュイメージとディスクイメージとのライフサイクルを大きく引き離すので、復旧の際にこのギャップを埋める必要がある。ある実施例に従うと、ＤＬＭ復旧とバッファキャッシュ復旧との間に、復旧の新しいステップが加えられる。この新しい復旧ステップをここでは「ディスクウォームアップ」と呼ぶ。

【００７２】

通常のキャッシュ動作の間、資源のマスタは、（キャッシュ復旧に先行する）ＤＬＭ復旧の際に、資源の位置とＰＩコピーおよびＣＵＲＲコピーの利用可能性とについておおよそしか知らないが、資源のマスタは、生き残ったデータベースサーバのバッファキャッシュ内の最新ＰＩおよびＣＵＲＲコピーの利用可能性について完全な情報を収集する。資源のマスタが、（もし障害より前に資源が障害の発生したデータベースサーバ上にマスタされていれば）新しいマスタであっても生き残ったマスタであっても、これは当てはまる。

【００７３】

情報を収集した後、マスタは、どのデータベースサーバが資源の最新コピーを所有しているかを知る。「ディスクウォームアップ」段では、マスタは、資源のこの最新コピー（もし利用可能であればＣＵＲＲ、および、もしＣＵＲＲコピーが障害の発生したデータベースサーバとともに消失していれば最新ＰＩコピー）のオーナーにWロックを発行する。次

に、マスタは、このデータベースサーバに、資源をディスクに書込むよう命令する。書込が完了すると、すべての他のデータベースサーバは、そのＨロックをＮＵＬＬロックに変換する（なぜなら書込まれたコピーが最新の利用可能なものであるからである）。これらのロックがコンバートされた後、キャッシュ復旧は通常通り続行可能である。

【００７４】

ディスクウォームアップ段の間、いくつかの最適化が可能である。たとえば、もし最新イメージが復旧を実行するデータベースサーバのバッファキャッシュ内にあれば、資源は必ずしもディスクに書込まれる必要はない。

【００７５】

ロックベース方式の代替

データベースサーバ間での資源のダーティコピーを直接発送するためのさまざまな技術を、特殊なタイプのロック（Ｍロック、ＷロックおよびＨロック）使用するロッキング方式をコンテクストとして記載した。具体的には、これらの特殊ロックを使用して、（１）資源のカレントバージョンを有するサーバのみが資源を変更することと、（２）資源の同じバージョンまたはより新しいバージョンがディスクに書込まれるまで、すべてのサーバが資源のそのＰＩバージョンを維持することと、（３）資源のディスクベースのバージョンが資源のより古いバージョンによって重ね書きされないこととを確実にする。

【００７６】

しかしながら、ロックベースのアクセス制御方式は、この発明が実施可能であるコンテクストの１つにすぎない。たとえば、任意のさまざまなアクセス制御方式を用いてこれらの同じ３つの規則を実施してもよい。したがって、この発明は、特定のタイプのアクセス制御方式に限定されるものではない。

【００７７】

たとえば、ロックをベースとして資源へのアクセスを管理する代わりに、アクセスは、トークンによって管理されてもよく、この場合各トークンが特定のタイプの許可を表わす。特定の資源のためのトークンが、上述した３つの規則が確実に実施されるように、パラレルサーバ間で転送されてもよい。

【００７８】

同様に、規則は、状態ベースの方式を用いて実施され得る。状態ベースの方式では、資源のバージョンは、事象に応答して状態を変化させ、バージョンの状態がそのバージョンに対して実行可能である動作のタイプを決定する。たとえば、データベースサーバは、その「現在の」状態での資源のカレントバージョンを受取る。現在の状態は、資源の変更および資源のディスクへの書込を可能とする。データベースサーバが資源のカレントバージョンを別のノードに転送すると、保持されているバージョンは「ＰＩ書込可能」状態に変る。ＰＩ書込可能状態では、バージョンは、（１）変更不可能であり、（２）量ね書き不可能であるが、（３）ディスクへの書込は可能である。資源の任意のバージョンがディスクに書込まれると、ディスクに書込まれたバージョンと同じまたはそれよりも古い、ＰＩ書込可能状態にあるバージョンのすべてが、「ＰＩ解放」状態に置かれる。ＰＩ解放状態では、バージョンは量ね書き可能であるが、ディスクへの書込または変更は不可能である。

【００７９】

ハードウェア概要

図６は、この発明の実施例が実現可能であるコンピュータシステム６００を示すブロック図である。コンピュータシステム６００は、情報を受け渡しするためのバス６０２または他の通信メカニズムと、バス６０２に結合され情報を処理するためのプロセッサ６０４とを含む。コンピュータシステム６００はまた、ランダムアクセスメモリ（ＲＡＭ）または他のダイナミック記憶装置などの主メモリ６０６を含み、これはバス６０２に結合されプロセッサ６０４によって実行されるべき命令および情報を記憶する。主メモリ６０６はまた、プロセッサ６０４によって実行されるべき命令の実行の間、一時的変数または他の中間情報を記憶するために使用され得る。コンピュータシステム６００は、リードオンリメモリ（ＲＯＭ）６０８または他の静的記憶装置をさらに含み、これはバス６０２に結合さ

れ静的情報およびプロセッサ６０４のための命令を記憶する。磁気ディスクまたは光学ディスクなどの記憶装置６１０が設けられこれはバス６０２に結合され情報および命令を記憶する。

【００８０】

コンピュータシステム６００は、バス６０２を介して陰極線管（ＣＲＴ）などのディスプレイ６１２に結合されもよく、これはコンピュータユーザに情報を表示する。英数字および他のキーを含む入力デバイス６１４は、バス６０２に結合されプロセッサ６０４に情報およびコマンド選択を与える。ユーザ入力デバイスの別のタイプは、マウス、トラックボールまたはカーソル方向キーなどのカーソルコントロール６１６であって、これは方向情報およびコマンド選択をプロセッサ６０４に与え、かつ、ディスプレイ６１２上のカーソルの動きを制御する。この入力デバイスは典型的には、第１の軸（たとえばｘ）および第２の軸（たとえばｙ）の、２軸での２自由度を有し、これによってデバイスは画面での位置を特定することが可能となる。

【００８１】

この発明は、ピングに関連付けられるオーバーヘッドを低減するコンピュータシステム６００の使用に関する。この発明のある実施例に従うと、ピングに関連付けられるオーバーヘッドは、プロセッサ６０４が主メモリ６０６に含まれる１つ以上の命令の１つ以上のシーケンスを実行することに応答して、コンピュータシステム６００によって低減される。そのような命令は、記憶装置６１０などの、別のコンピュータ読出可能媒体から主メモリ６０６に読出されてもよい。主メモリ６０６内に含まれる命令のシーケンスを実行することによって、プロセッサ６０４はここに記載するプロセスステップを実行する。代替の実施例では、ハードワイア回路をソフトウェア命令の代わりにまたはこれと組合せて使用してこの発明を実現してもよい。したがって、この発明の実施例は、ハードウェア回路およびソフトウェアの特定の組合せに限定されない。

【００８２】

ここに用いる「コンピュータ読出可能媒体」という言葉は、プロセッサ６０４に命令を与えて実行させることに関与する任意の媒体を指す。そのような媒体は、不揮発性媒体、揮発性媒体および伝送媒体を含むがこれに限られるものではない、多くの形態を取ってもよい。不揮発性媒体は、たとえば、記憶装置６１０などの、光学ディスクまたは磁気ディスクを含む。揮発性媒体は、主メモリ６０６などの、ダイナミックメモリを含む。伝送媒体は、バス６０２を含むワイアを含む、同軸ケーブル、銅線および光ファイバを含む。伝送媒体はまた、電波および赤外データ通信の間生成されるものなど、音波または光波の形態を取ってもよい。

【００８３】

コンピュータ読出可能媒体の通常の形態は、たとえば、フロッピー、フレキシブルディスク、ハードディスク、磁気テープ、またはその他の磁気媒体、ＣＤ－ＲＯＭ、その他の光学媒体、パンチカード、紙テープ、孔のパターンを備えるその他の物理的媒体、ＲＡＭ、ＰＲＯＭおよびＥＰＲＯＭ、ＦＬＡＳＨ－ＥＰＲＯＭ、その他のメモリチップまたはカートリッジ、以下に記載する搬送波、またはコンピュータが読出可能なその他の媒体の形態を含む。

【００８４】

コンピュータ読出可能媒体のさまざまな形態は、１つ以上の命令の１つ以上のシーケンスをプロセッサ６０４に搬送して実行することにかかわり得る。たとえば、命令は最初に、遠隔コンピュータの磁気ディスク上に担持されてもよい。遠隔コンピュータは、命令をそのダイナミックメモリにロードし、モデムを使用して電話線を介して命令を送信することができる。コンピュータシステム６００にローカルなモデムは、電話線上のデータを受信し、赤外送信器を使用してデータを赤外信号に変換することができる。赤外検出器は赤外信号で搬送されるデータを受信可能であり、適切な回路がデータをバス６０２上に与えることができる。バス６０２は、データを主メモリ６０６に搬送し、プロセッサ６０４はそこから命令を検索し実行する。主メモリ６０６によって受取られた命令は、プロセッサ６

０４によって実行される前またはその後に、記憶装置６１０上にオプションとして記憶されてもよい。

【００８５】

コンピュータシステム６００は、１つ以上の記憶装置（たとえばディスクドライブ６５５）がコンピュータシステム６００と１つ以上の他のＣＰＵ（たとえばＣＰＵ６５１）の両方にアクセス可能である、共有ディスクシステムに属する。例示のシステムでは、ディスクドライブ６５５への共有アクセスは、システムエリアネットワーク６５３によって与えられる。しかしながら、さまざまなメカニズムを代替的に使用して共有アクセスを与えてもよい。

【００８６】

コンピュータシステム６００はまた、バス６０２に結合される通信インターフェイス６１８を含む。通信インターフェイス６１８は、双方向のデータ通信を与え、これはネットワークリンク６２０に結合し、ネットワークリンクはローカルネットワーク６２２に接続される。たとえば、通信インターフェイス６１８は、統合サービスデジタル網（ＩＳＤＮ）カードまたはモデムであってもよく対応するタイプの電話線にデータ通信接続を与える。別の例として、通信インターフェイス６１８は、互換性のあるＬＡＮにデータ通信接続を与えるローカルエリアネットワーク（ＬＡＮ）カードであってもよい。ワイアレスリンクが実現されてもよい。いかなるそのような実現化例でも、通信インターフェイス６１８は、さまざまなタイプの情報を表わすデジタルデータストリームを搬送する電気信号、電磁波信号または光学信号を送信し受信する。

【００８７】

ネットワークリンク６２０は、典型的には、１つ以上のネットワークを介して他のデータデバイスにデータ通信を与える。たとえば、ネットワークリンク６２０は、ローカルネットワーク６２２を介してホストコンピュータ６２４またはインターネットサービスプロバイダ（ＩＳＰ）６２６によって動作するデータ装置に接続してもよい。ＩＳＰ６２６は、現在通常「インターネット」６２８と呼ばれるワールドワイドパケットデータ通信ネットワークを介して、データ通信サービスを提供する。ローカルネットワーク６２２とインターネット６２８とはどちらも、デジタルデータストリームを搬送する電気信号、電磁波信号または光学信号を使用する。さまざまなネットワークを通る信号と、ネットワークリンク６２０上および通信インターネット６１８を通る信号とは、デジタルデータをコンピュータシステム６００へかつそこから搬送するものであるが、情報を転送する搬送波の例示的形態である。

【００８８】

コンピュータシステム６００は、ネットワーク、ネットワークリンク６２０および通信インターフェイス６１８を介して、プログラムコードを含め、メッセージを送信しデータを受信することが可能である。インターネットの例では、サーバ６３０は、インターネット６２８、ＩＳＰ６２６、ローカルネットワーク６２２および通信インターフェイス６１８を介して、アプリケーションプログラムのために要求されたコードを伝送可能である。

【００８９】

受信されたコードは、受信されたときにプロセッサ６０４によって実行されてもよいし、かつ／または記憶装置６１０または他の不揮発性装置に記憶されて後に実行されてもよい。このようにして、コンピュータシステム６００は、搬送波の形でアプリケーションコードを獲得可能である。

【００９０】

複数のデータベースサーバが共通の永続性記憶装置へのアクセスを有するときに生じるピングを参照してピングを処理するための技術が記載されるが、この技術はこのコンテクストに限定されるのではない。具体的には、これらの技術は、あるキャッシュに関連付けられるプロセスが現在のバージョンが他のキャッシュ内に位置する資源を要求する可能性のあるいかなる環境に適用されてもよい。そのような環境は、たとえば、異なったノード上のテキストサーバが同じテキスト材料へのアクセスを有するような環境、異なったノード

上のメディアサーバが同じビデオデータへのアクセスを有するような環境、などを含む。

【００９１】

ここに記載する技術を用いてピングを処理すれば、資源のデータベースサーバ間の転送は効率よくなるので、動作可能時間性能は、データベースサーバの数およびデータベースサーバあたりのユーザの増加にあわせて増大する。加えて、この技術によって、データベースサーバの数の増加にあわせて増大する単一のデータベースサーバの障害（障害の最もよくあるタイプ）からの効率的な復旧が得られる。

【００９２】

重要なことには、ここに記載する技術は、ディスク介入によってではなく、ＩＰＣトランスポートを介して資源を送信することによってピングを処理する。したがって、ピングをもたらす、資源についてのディスクＩ／Ｏは、かなり解消される。同期Ｉ／Ｏを伴うのは、これがログ強制のために必要とされる場合においてのみである。加えて、ディスクＩ／Ｏはチェックポイントおよびバッファキャッシュ置換のために生じるが、そのようなＩ／Ｏは、クラスタにわたるバッファ発送を減速させることはない。

【００９３】

ここに記載する直接発送技術はまた、ピングによって生じるコンテクスト切換の数を低減させるようにもなる。具体的には、プロトコルの関与者（要求者およびホルダ）とマスタとの間の往復メッセージのシーケンスは、要求者、マスタ、ホルダ、要求者からなる通信トライアングルによって代用される。

【図面の簡単な説明】

【図１】　資源の最新バージョンのキャッシュからキャッシュへの転送を例示するブロック図である。

【図２】　この発明の実施例に従ってディスク介入なしにあるキャッシュから別のキャッシュへ資源を伝送するステップを例示するフローチャートである。

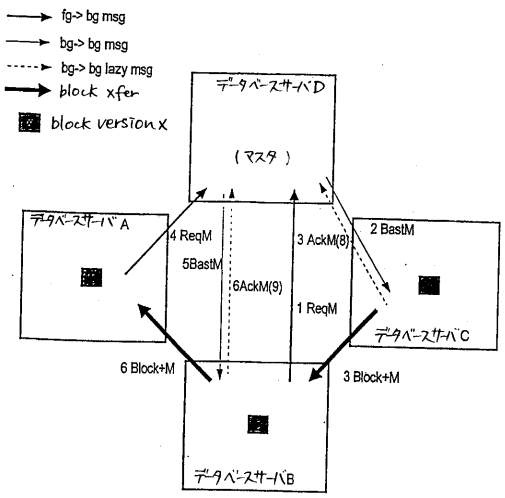
【図３】　この発明の実施例に従って、資源のパストイメージを解放するステップを例示するフローチャートである。

【図４】　この発明の実施例に従って単一のデータベースサーバの障害の後に復旧するステップを例示するフローチャートである。

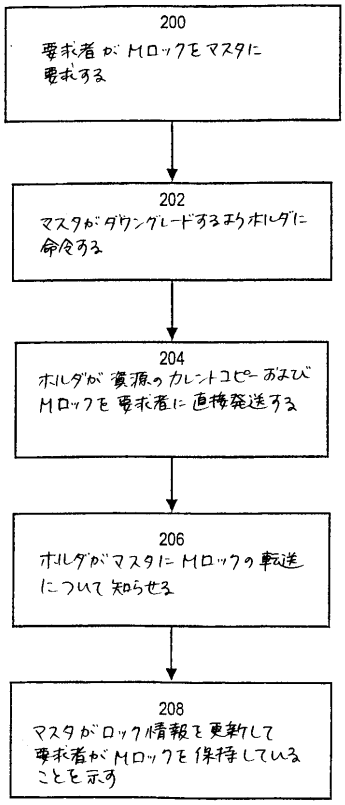
【図５】　この発明の実施例に従ってチェックポイントサイクルを例示するブロック図である。

【図６】　この発明の実施例が実現可能であるコンピュータシステムのブロック図である。

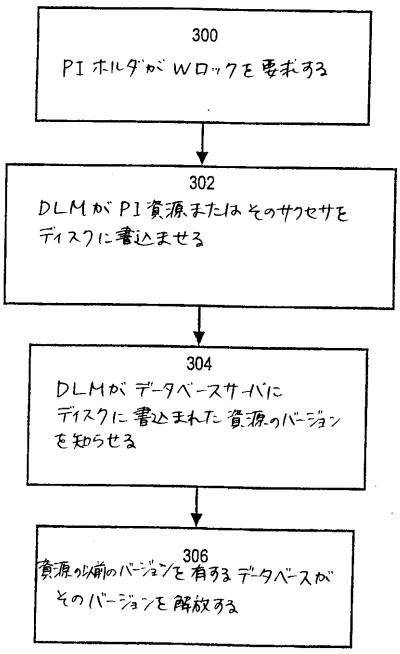
【図１】


fg-> bg msg
bg-> bg msg
bg-> bg lazy msg
block xfer
block version x
データベースサーバD
（マスタ）
データベースサーバA
データベースサーバC
データベースサーバB
4 ReqM
5BastM
6AckM(9)
3 AckM(8)
2 BastM
1 ReqM
6 Block+M
3 Block+M

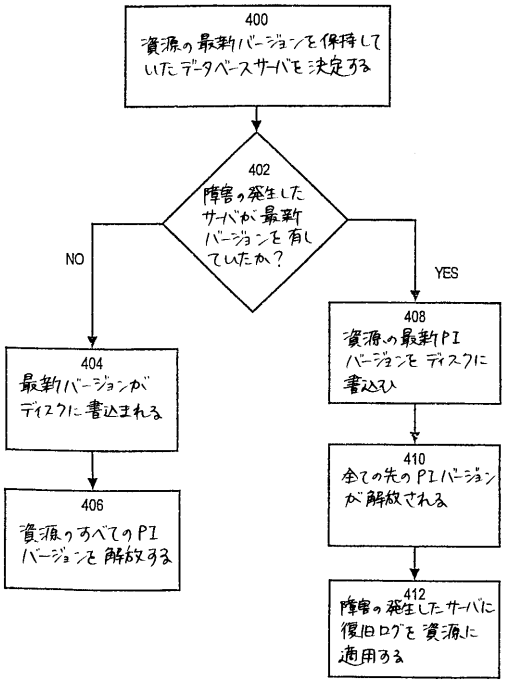
【図２】


200
要求者がMロックをマスタに要求する
202
マスタがダウングレードするようホルダに命令する
204
ホルダが資源のカレントコピーおよびMロックを要求者に直接発送する
206
ホルダがマスタにMロックの転送について知らせる
208
マスタがロック情報を更新して要求者がMロックを保持していることを示す

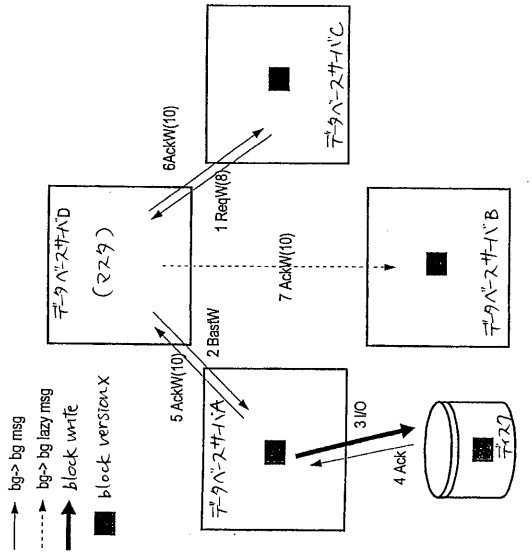
【図３】


300
PIホルダがWロックを要求する
302
DLMがPI資源またはそのサクセサをディスクに書込ませる
304
DLMがデータベースサーバにディスクに書込まれた資源のバージョンを知らせる
306
資源の以前のバージョンを有するデータベースがそのバージョンを解放する

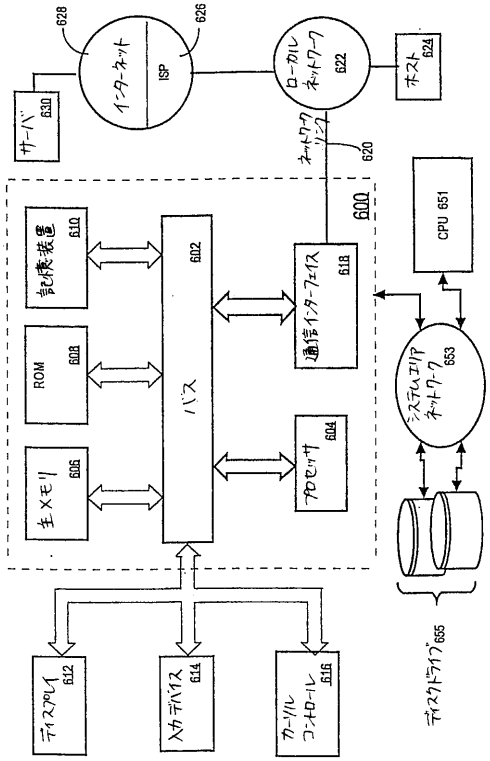
【図４】


400
資源の最新バージョンを保持していたデータベースサーバを決定する
402
障害の発生したサーバが最新バージョンを有していたか？
NO
YES
404
最新バージョンがディスクに書込まれる
406
資源のすべてのPIバージョンを解放する
408
資源の最新PIバージョンをディスクに書込む
410
全ての先のPIバージョンが解放される
412
障害の発生したサーバに復旧ログを資源に適用する

【図５】


データベースサーバC
データベースサーバD
（マスタ）
データベースサーバB
データベースサーバA
ディスク
6AckW(10)
1 ReqW(8)
7 AckW(10)
2 BastW
5 AckW(10)
3 I/O
4 Ack
bg-> bg msg
bg-> bg lazy msg
block write
block version.x

【図６】


インターネット
ISP
628
626
ローカルネットワーク
622
ホスト
624
サーバ
630
ネットワークリンク
620
600
記憶装置
610
ROM
608
主メモリ
606
バス
602
通信インターフェイス
618
プロセッサ
604
CPU 651
システムエリアネットワーク
653
ディスクドライブ
655
ディスプレイ
612
入力デバイス
614
カーソルコントロール
616

フロントページの続き

(74)代理人 100098316
弁理士 野田 久登
(74)代理人 100109162
弁理士 酒井 將行
(72)発明者 バンフォード，ロジャー・ジェイ
アメリカ合衆国、９４１０９ カリフォルニア州、サンフランシスコ、ハイド・ストリート、２４３０
(72)発明者 クロッツ，ボリス
アメリカ合衆国、９４００２ カリフォルニア州、ベルモント、ウィンディング・ウェイ、１５６６

審査官 相崎 裕恒

(56)参考文献 特開平０４－８４３５０（ＪＰ，Ａ）
特開平０４－３６５１５２（ＪＰ，Ａ）
特開平１１－８５６００（ＪＰ，Ａ）

(58)調査した分野(Int.Cl.，ＤＢ名)
G06F 12/00,13/00
G06F 12/08-12
G06F 9/46/54
G06F 15/16-177