(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第6576324号
(P6576324)
(45)発行日　令和1年9月18日(2019.9.18)　(24)登録日　令和1年8月30日(2019.8.30)

(51)Int.Cl.　F I
G06F 3/06 (2006.01)　G06F 3/06 301Z
G06F 13/10 (2006.01)　G06F 13/10 340A

請求項の数 17　(全 26 頁)

(21)出願番号　特願2016-502215 (P2016-502215)
(86)(22)出願日　平成26年3月13日 (2014.3.13)
(65)公表番号　特表2016-517588 (P2016-517588A)
(43)公表日　平成28年6月16日 (2016.6.16)
(86)国際出願番号　PCT/US2014/026689
(87)国際公開番号　WO2014/151928
(87)国際公開日　平成26年9月25日 (2014.9.25)
審査請求日　平成29年3月8日 (2017.3.8)
(31)優先権主張番号　61/784,282
(32)優先日　平成25年3月14日 (2013.3.14)
(33)優先権主張国・地域又は機関
米国 (US)

(73)特許権者　508032284
カリフォルニア　インスティチュート　オブ　テクノロジー
アメリカ合衆国，カリフォルニア州　９１１２５，パサデナ，イースト・カリフォルニア・ブールヴァード　１２００，エムシー　６－３２
(74)代理人　100147485
弁理士　杉村　憲司
(74)代理人　100164471
弁理士　岡野　大和
(74)代理人　100188307
弁理士　太田　昌宏

最終頁に続く

(54)【発明の名称】非均質システム用の分散ストレージ配分

(57)【特許請求の範囲】
【請求項１】
非均質ストレージシステムにおいてストレージ量を配分するコンピュータベースの方法であって、該コンピュータベースの方法は：
既知であり非均質なアクセス確率を有するハードウェアストレージノードのセットを提供するステップと、
コンピュータによって、目的関数を提供するステップと、
コンピュータによって、制約を提供するステップと、
前記制約及び前記目的関数に基づいて、コンピュータによって、前記ハードウェアストレージノードのセットについてのｋ個の互いに素であるサブセットを選択するステップであって、ｋは１以上の整数であり、前記ｋ個の互いに素であるサブセットの各々は、複数のハードウェアストレージノードを含む、ステップと、
前記選択するステップに基づいて、コンピュータによって、前記ストレージ量を前記ｋ個の互いに素であるサブセットに対して配分するステップと
を含む、方法であって、
前記配分するステップは、前記ｋ個の互いに素であるサブセットの前記複数のハードウェアストレージノードの前記非均質なアクセス確率及び前記制約に基づいており、かつ、
前記配分するステップは、前記ｋ個の互いに素であるサブセットのうちのｋ番目のサブセットの各ハードウェアストレージノードに、等量に配分されたストレージ量S_kを基準として、前記ｋ個の互いに素であるサブセットのうちの１番目、２番目、・・・、（ｋ－１

）番目のサブセットの各ハードウェアストレージノードに、それぞれ、ｋ*Ｓ$_k$、（ｋ－１）*Ｓ$_k$、…、２*Ｓ$_k$のストレージ量が等量に配分されることとなるように、前記ストレージ量を前記ｋ個の互いに素であるサブセットの前記複数のハードウェアストレージノードに跨がって分散させることによって達成される、方法。

【請求項２】

前記目的関数は、ａ）データ回収の信頼度、ｂ）データのアクセス可能性、及びｃ）ストレージのコスト、の１以上に対応する、請求項１に記載のコンピュータベースの方法。

【請求項３】

前記制約は、ａ）前記ハードウェアストレージノードのセットに関しての利用可能な総ストレージ以下であるストレージ量、ｂ）ストレージのコスト、ｃ）データ回収の信頼度、及びｄ）データのアクセス可能性、の１以上に対応する、請求項１又は２に記載のコンピュータベースの方法。

【請求項４】

コンピュータによって、ファイルを提供するステップと、

コンピュータによって、前記ファイルを複数のファイルに分割するステップと、

前記提供するステップと前記分割するステップに基づいて、コンピュータによって、前記複数のファイルをエンコードするステップと、

前記エンコードするステップに基づいて、コンピュータによって、前記ストレージ量を表す総サイズとなる複数のエンコードされたファイルを取得するステップと、

前記取得するステップに基づいて、コンピュータによって、前記複数のエンコードされたファイルを、前記選択されたｋ個の互いに素であるサブセット上に格納するステップとをさらに含む請求項１乃至３のいずれか一項に記載のコンピュータベースの方法であって、

各エンコードされたファイルのサイズは、前記配分するステップによって決定された配分済みストレージ量に合致される、方法。

【請求項５】

前記エンコードするステップは、ａ）最大距離分離（ＭＤＳ）符号、及びｂ）ランダム線形符号、のいずれか１つを用いて行われる、請求項４に記載のコンピュータベースの方法。

【請求項６】

前記ｋ個の互いに素であるサブセットのあるサブセットのハードウェアストレージノードは、ａ）ハードディスク、ｂ）コンピュータシステム、ｃ）データサーバ、ｄ）データセンタ、ｅ）インターネットサービスプロバイダ、及びｆ）仮想プライベートサーバ（ＶＰＳ）の１つを備える、請求項１乃至５のいずれか一項に記載のコンピュータベースの方法。

【請求項７】

前記選択するステップ及び前記配分するステップはコンピュータベースのアルゴリズムによって行われ、該アルゴリズムのステップは、ａ）ｋ＝１である１レベル対称配分アルゴリズム、ｂ）ｋ＝２である２レベル対称配分アルゴリズム、ｃ）ｋ＝３である３レベル対称配分アルゴリズム、及び一般化されたｄ）ｋレベル対称配分アルゴリズム、の１以上によって提供され、

ハードウェアストレージノードのセットについての前記ｋ個の互いに素であるサブセットのそれぞれは、レベルと呼ばれ、前記１、２、３、kレベル対称配分アルゴリズムは、それぞれ、１、２、３、ｋレベルを選択することを含む、請求項１乃至６のいずれか一項に記載のコンピュータベースの方法。

【請求項８】

前記ｋレベルの対称配分アルゴリズムは前記２レベル及び３レベル対称配分アルゴリズムについての拡張であり、

配分されるストレージ量は、

【数１】

$$S_k = \frac{T}{km_1+(k-1)m_2+\cdots+2m_{k-1}+m_k}$$

によって表され、Ｔは前記ストレージ量であり、m_iは前記ｋレベルのうちのレベルｉのハードウェアストレージノードの数である、請求項７に記載のコンピュータベースの方法。

【請求項９】

コンピュータベースのソースを備える分散ストレージ配分のためのコンピュータベースのシステムであって、該ソースは１以上の通信リンク上で既知であり非均質なアクセス確率を有する複数のハードウェアベースのストレージノードと通信するように構成されており、該ソースはストレージ配分アルゴリズムを実行して前記複数のハードウェアベースのストレージノード上に配分されたストレージを得るように構成されており、前記アルゴリズムは、

ｉ）提供されたバジェットと提供された目的関数とに基づいて、前記複数のハードウェアベースのストレージノードについてのｋ個の互いに素であるサブセットを選択するタスクであって、ｋは１以上の整数であり、前記提供されたバジェットは前記複数のハードウェアベースのストレージノードに関しての利用可能な総ストレージサイズについての部分を指定する、タスクと、

ｉｉ）前記提供されたバジェットを前記ｋ個の互いに素であるサブセットの前記複数のハードウェアベースのストレージノードに跨がって分散させることによって前記提供されたバジェットを前記ｋ個の互いに素であるサブセットに配分するタスクであって、前記ｋ個の互いに素であるサブセットのうちのｋ番目のサブセットの各ハードウェアストレージノードに、等量に配分されたストレージ量S_kを基準として、前記ｋ個の互いに素であるサブセットのうちの１番目、２番目、・・・、（ｋ－１）番目のサブセットの各ハードウェアストレージノードに、それぞれ、$k*S_k$、$(k-1)*S_k$、…、$2*S_k$のストレージ量が等量に配分されることとなる、タスクと

を行うアルゴリズムであって、

前記提供されたバジェットを配分するタスクは前記非均質なアクセス確率及び前記提供されたバジェットに基づく、システム。

【請求項１０】

前記目的関数は、ａ）データ回収の信頼度、ｂ）データのアクセス可能性、ｃ）ストレージのコスト、及びｄ）利用可能ストレージのサイズ制限の１以上に対応する、請求項９に記載のコンピュータベースのシステム。

【請求項１１】

前記ストレージ配分アルゴリズムは、ａ）ｋ＝１である１レベル対称配分アルゴリズム、ｂ）ｋ＝２である２レベル対称配分アルゴリズム、ｃ）ｋ＝３である３レベル対称配分アルゴリズム、及び一般化されたｄ）ｋレベル対称配分アルゴリズム、の１以上によって提供されるステップをさらに含み、

ハードウェアストレージノードのセットについての前記ｋ個の互いに素であるサブセットのそれぞれは、レベルと呼ばれ、前記１、２、３、ｋレベル対称配分アルゴリズムは、それぞれ、１、２、３、ｋレベルを選択することを含む、請求項９又は１０に記載のコンピュータベースのシステム。

【請求項１２】

前記ｋレベルの対称配分アルゴリズムは前記２レベル及び３レベル対称配分アルゴリズムについての拡張であり、

配分されるバジェット量は、

【数２】

$$S_k = \frac{T}{km_1+(k-1)m_2+\cdots+2m_{k-1}+m_k}$$

によって表され、Ｔは前記提供されたバジェットであり、m_iは前記ｋレベルのうちのレベルｉのハードウェアベースのストレージノードの数である、請求項１１に記載のコンピュータベースのシステム。

【請求項１３】

１以上の通信リンク上で既知であり非均質な信頼度を有する複数のハードウェアベースのストレージノードを含むハードウェアベースのストレージノードのセットと通信するように構成された請求項９に記載のコンピュータベースのシステムと、

既知であり非均質な信頼度を有する複数のハードウェアベースのストレージノードと

を備える分散ストレージ配分システムであって、

前記コンピュータベースのシステムは、前記提供されたバジェットに等しいサイズのエンコードされたファイルを、前記ストレージ配分アルゴリズムによって提供された配分済みストレージに基づいて、前記複数のハードウェアベースのストレージノードに跨がって分散させるように構成された、システム。

【請求項１４】

前記エンコードされたファイルは、前記コンピュータベースのソース上で実行されるコンピュータベースのエンコーディングアルゴリズムであって、ａ）最大距離分離（ＭＤＳ）符号、及びｂ）ランダム線形符号、のいずれか１つに基づいているアルゴリズムによってエンコードされる、請求項１３に記載の分散ストレージ配分システム。

【請求項１５】

前記複数のハードウェアベースのストレージノードのあるハードウェアベースのストレージノードは、ａ）ハードディスク、ｂ）コンピュータシステム、ｃ）データサーバ、ｄ）データセンタ、ｅ）インターネットサービスプロバイダ、及びｆ）仮想プライベートサーバ（ＶＰＳ）の１つを備える、請求項１３又は１４に記載の分散ストレージ配分システム。

【請求項１６】

複数の独立しているメッセージを通信リンク上でリアルタイムストリーミングするためのコンピュータベースの方法であって、該コンピュータベースの方法は：

ｉ）前記複数の独立しているメッセージのメッセージサイズｓを、コンピュータによって提供するステップと、

ｉｉ）時間ステップの数に基づいてメッセージ作成間隔ｃを、コンピュータによって提供するステップであって、前記メッセージ作成間隔は２つの連続するメッセージの作成時刻間の時間間隔を規定する、ステップと、

ｉｉｉ）バジェットを指定する制約を、コンピュータによって提供するステップであって、前記バジェットは各時間ステップにて送信されるエンコードされたパケットについての最大サイズに対応する、ステップと、

ｉｖ）時間ステップの個数で表された固定されたデコーディング遅延ｄを、コンピュータによって提供するステップであって、前記固定されたデコーディング遅延は前記複数の独立しているメッセージのうちのあるメッセージの作成時刻との関係での遅延を規定し、前記メッセージは該遅延内にてコンピュータベースのデコーダによって１以上の送信されたパケットに基づいてデコードされなければならない、ステップと、

ｖ）前記通信リンク上での送信されたパケットについての非均質遅延依存性損失確率を、規定する非均質アクセス確率モデルを、コンピュータによって提供するステップと、

ｖｉ）前記複数の独立しているメッセージのうちのあるメッセージを、コンピュータによってエンコードするステップと、

ｖｉｉ）前記ステップｉ）～ｖｉ）に基づいて、前記エンコードされたメッセージに対

応する複数のパケットをコンピュータによって生成して、かつ、前記複数のパケットを前記通信リンク上で逐次的に送信するステップ
とを含む、方法であって、
前記複数の独立しているメッセージのうちの時間ステップｉにおいて作成されたメッセージに、非均質なアクセス確率モデルに従って選択された時間ステップｉ，ｉ＋１，…，ｉ＋ｄにおいて送信されたパケット内の領域の部分S_i、S_{i+1}、…、S_{i+d}が、配分され、
時間ステップｉ，ｉ＋１，…，ｉ＋ｄにおいて送信された前記パケット内の領域の配分された部分S_{i+1}、…、S_{i+d}は、それぞれ、$2*S_i$、…、$(d+1)*S_i$に等しく、
前記メッセージは、時間ステップｉ，ｉ＋１，…，ｉ＋ｄにおけるパケット内の領域の配分された部分に跨がって削除修正コードを用いてコーディングされ、
前記メッセージは、コンピュータベースのデコーダによって、前記メッセージの作成時刻から測って前記固定されたデコーディング遅延内にてデコードされる、方法。
【請求項１７】
前記パケット内の領域の前記配分された部分は、コンピュータベースの配分アルゴリズムによって得られるのであって、該アルゴリズムのステップは、ａ）前記パケット内の領域の配分された部分が１の配分された部分である場合における１レベル対称配分アルゴリズム、ｂ）前記パケット内の領域の配分された部分が２の配分された部分である場合における２レベル対称配分アルゴリズム、ｃ）前記パケット内の領域の配分された部分が３の配分された部分である場合における３レベル対称配分アルゴリズム、及びｄ）前記パケット内の領域の配分された部分がｋの配分された部分である場合におけるｋレベル対称配分アルゴリズム、の１以上によって提供される、請求項１６に記載のコンピュータベースの方法。

【発明の詳細な説明】
【技術分野】
【０００１】
〔関連出願の相互参照〕
本願は、「非均質システム用の分散ストレージ配分」（Distributed Storage Allocation for Heterogeneous Systems）と題する２０１３年３月１４日に出願された米国仮特許出願第６１／７８４，２８２号の優先権を主張するのであり、ここにてその全体が参照によって取り込まれる。
【０００２】
〔政府援助についての陳述〕
この発明は、空軍から授与された助成第ＦＡ９５５０－１０－１０１６６号に基づく政府援助の下でなされた。政府は、この発明に関して一定の権利を有する。
【背景技術】
【０００３】
本願の開示は、ネットワーク通信及び例えばコンテンツデリバリ又は無線通信において用いられる分散ネットワークアーキテクチャの分野に関する方法及びアルゴリズムに関する。特に本願の開示は、新規な分散ストレージ配分手法を提供し、これは、例えば非均質なアクセス確率を有するストレージノードを伴う分散ストレージを備えるネットワークを介した通信に用いることができる。目標は、分散ストレージ上で、リカバリが成功する確率を最大化する態様で、所与のデータオブジェクト（例えば、ファイル等）を格納することである。データオブジェクトは、複数のストレージノードに跨がって分割及びコーディングされることができる。適切な符号（例えば、最大距離分離（ＭＤＳ）符号、ランダム線形符号等）を用いると仮定すれば、アクセスされたデータの総量が少なくとも元のデータオブジェクトのサイズである場合、元のデータオブジェクトをリカバーすることができる。

【先行技術文献】
【非特許文献】
【０００４】
【非特許文献１】D. Leong、A. G. Dimakis及びT. Ho；「Distributed storage allocations」（IEEE Trans. Inf. Theory, vol. 58, no. 7, pp. 4733-4752、２０１２年７月）
【非特許文献２】W. K. Lin、D. M. Chiu及びY. B. Lee；「Erasure code replication revisited」（Proc. Int. Conf. Peer-to-Peer Comput. (P2P)、２００４年９月）
【非特許文献３】A. Tsirigos及びZ. J. Haas；「Analysis of multipath routing-Part I: The effect on the packet delivery ratio」（IEEE Trans. Wireless Commun., vol. 3, no. 1, pp. 138-146、２００４年１月）
【非特許文献４】「Analysis of multipath routing, Part 2: Mitigation of the effects of frequently changing network topologies」（IEEE Trans. Wireless Commun., vol. 3, no. 2, pp. 500-511、２００４年３月）
【非特許文献５】J.-S. Wu及びR.-J. Chen；「An algorithm for computing the reliability of weighted-k-out-of-n systems」（IEEE Trans. Rel., vol. 43, no. 2, pp. 327-328、１９９４年６月）
【非特許文献６】V. Ntranos、G. Caire及びA. G. Dimakis；「Allocations for heterogeneous distributed storage」（Proc. IEEE Int. Symp. Inf. Theory (ISIT)、２０１２年７月）
【非特許文献７】D. Leong及びT. Ho；「Erasure coding for real-time streaming」（Proc. IEEE Int. Symp. Inf. Theory (ISIT), Cambridge, Massachusetts, USA、２０１２年７月）
【非特許文献８】D. Leong、A. Qureshi、及びT. Ho；「On coding for real-time streaming under packet erasures」（Proc. IEEE Int. Symp. Inf. Theory (ISIT), Istanbul, Turkey、２０１３年７月）
【非特許文献９】Y. Chen、J. Chen及びY. Yang；「Multi-hop delay performance in wireless mesh networks」（MOBILE NETW APPL, vol. 13, pp. 160-168、２００８年４月）
【発明の概要】
【発明が解決しようとする課題】
【０００５】
本願開示の様々な実施形態によれば、分散ストレージシステム内の様々なストレージノードは、それらのアクセス確率について降順で配列されうる（例えば、最も高い故障確率のノードが最後とされる）。１以上のレベルに関しての対称配分に基づく新規な手法を、本願開示にて提供する。基本的な１レベル対称配分では、最初のｍ個のノード上で均一に全ストレージバジェットを分散し、ｍの値は、例えば故障確率に基づいて決定される。２レベル対称配分では、バジェットは２分されて、一方の部分は最初のm_1個のノードにわたって均一に分散されて、他方の部分は後続のm_2個のノードにわたって均一に分散され、第１のサブセット内の各ノードに配分される量は第２のサブセットのそれの２倍とされる。ｋ＝２とした２レイヤー配分から導出されるｋレベル対称配分についても説明する。本願開示の様々な実施形態によれば、アクセス確率を与えられた場合において良好な１レベル、２レベル、及びｋレベルの対称配分を発見するための効率的なアルゴリズムが提供される。
【０００６】
本願開示の様々なグラフにて示すように、本願開示による１レベル、２レベル及びｋレベルの対称配分は、その複雑度が低けれども、大偏差不等式及び凸最適化に基づく既存の方法に勝る性能を、同様のパラメータ設定下にて、数値実験においてもたらすことができる。さらに、本願開示による新規な２レベル及びｋレベルの対称配分は、新規な１レベル対称配分よりも高いリカバリ確率を達成することができる。また、本願開示によれば、少数のノード個数（例えば、ｎ≦４）については、網羅的に決定することができる正確な最適配分を、１レベル及び２レベルの対称配分の中から発見することができる。

【０００７】
分散ストレージに加えて、上記手法は、例えばリアルタイムストリーミング用の符号設計の問題等についても適用されることができ、該問題においては、メッセージが逐次的にソース（例えば、コンピュータベースのワークステーション等）に到着することができ、パケット削除チャネルを介したシンク（sink）（例えば、コンピュータベースのワークステーション等）への送信を行うためにソースにてメッセージがエンコードされ、該シンクは指定された遅延の間にメッセージを逐次的に復号する必要がある。様々なネットワークシナリオにおいては、（例えば、syncノードによって受信された場面等における）パケット遅延が変動し、これによって遅延と共にパケット受信の確率が上昇し得る。リアルタイムストリーミング問題における送信パケットをストレージ配分問題のノードとして同視すれば、対応する非均質遅延依存性の損失確率が与えられれば、対応する非均質ストレージ配分問題に関しての解法は対応するストリーミング問題に関してのイントラセッションコードに転じることになる。

【０００８】
本願開示の第１の側面によれば、非均質ストレージシステムにおいてストレージを配分するコンピュータベースの方法であって、該コンピュータベースの方法は：既知であり非均質な信頼度を有するハードウェアストレージノードのセットを提供するステップと、コンピュータによって、目的関数を提供するステップと、コンピュータによって、制約を提供するステップと、制約及び目的関数に基づいて、コンピュータによって、ハードウェアストレージノードのセットについての１以上の互いに素であるサブセットを選択するステップと、選択するステップに基づいて、コンピュータによって、ストレージ量を１以上の互いに素であるサブセットに対して配分するステップとを含む、方法であって、配分するステップは非均質な信頼度及び制約に基づいており、かつ、配分するステップは、１以上の互いに素であるサブセットの各サブセットの各ハードウェアストレージノードに等量のストレージ量であって異なるサブセットのハードウェアストレージノードに配分される量とは異なる量のストレージ量が配分されることとなるようにストレージ量を１以上の互いに素であるサブセットに跨がって分散させることによって達成される、方法、が提供される。

【０００９】
本願開示の第２の側面によれば、分散ストレージ配分のためのコンピュータベースのシステムが提供されるのであって、該コンピュータベースのシステムは：コンピュータベースのソースを備える分散ストレージ配分のためのコンピュータベースのシステムであって、該ソースは１以上の通信リンク上で既知であり非均質な信頼度を有する複数のハードウェアベースのストレージノードと通信するように構成されており、該ソースはストレージ配分アルゴリズムを実行して複数のハードウェアベースのストレージノード上に配分されたストレージを得るように構成されており、アルゴリズムは、ｉ）提供されたバジェットと提供された目的関数とに基づいて、複数のハードウェアベースのストレージノードについての１以上の互いに素であるサブセットを選択するタスクであって、バジェットは複数のハードウェアベースのストレージノードに関しての利用可能な総ストレージサイズについての部分を指定する、タスクと、ｉｉ）バジェットを１以上の互いに素であるサブセットに跨がって分散させることによってバジェットを１以上の互いに素であるサブセットに配分するタスクであって、１以上の互いに素であるサブセットの各サブセットの各ハードウェアベースのストレージノードに等量のバジェットであって異なるサブセットのハードウェアベースのストレージノードに配分される量とは異なる量のバジェットが配分されることとなる、タスクとを行うアルゴリズムであって、バジェットについての配分タスクは非均質な信頼度及びバジェットに基づいている、システムが提供される。

【００１０】
本願開示の第３の側面によれば、複数の独立しているメッセージを通信リンク上でリアルタイムストリーミングするためのコンピュータベースの方法であって、該コンピュータベースの方法は：ｉ）複数の独立しているメッセージのメッセージサイズｓを、コンピュ

ータによって提供するステップと、ｉｉ）時間ステップの数に基づいてメッセージ作成間隔ｃを、コンピュータによって提供するステップであって、メッセージ作成間隔は２つの連続するメッセージの作成時刻間の時間間隔を規定する、ステップと、ｉｉｉ）バジェットを指定する制約を、コンピュータによって提供するステップであって、バジェットは各時間ステップにて送信されるエンコードされたパケットについての最大サイズに対応する、ステップと、ｉｖ）時間ステップの個数で表された固定されたデコーディング遅延ｄを、コンピュータによって提供するステップであって、固定されたデコーディング遅延は複数の独立しているメッセージのうちのあるメッセージの作成時刻との関係での遅延を規定するのであり、メッセージは該遅延内にてコンピュータベースのデコーダによって１以上の送信されたパケットに基づいてデコードされなければならない、ステップと、ｖ）通信リンク上での送信されたパケットについての非均質遅延依存性損失確率を、規定する非均質信頼度モデルを、コンピュータによって提供するステップと、ｖｉ）複数の独立しているメッセージのうちのあるメッセージを、コンピュータによってエンコードするステップと、ｖｉｉ）ステップｉ）～ｖｉ）の対応に基づいて、エンコードされたメッセージに対応する複数のパケットをコンピュータによって生成して、かつ、複数のパケットを通信リンク上で逐次的に送信するステップとを含む、方法であって、複数の独立しているメッセージのうちの時間ステップｉにおいて作成されたメッセージに、非均質信頼度モデルに従って選択された時間ステップｉ，ｉ＋１，…，ｉ＋ｄにおいて送信されたパケット内の領域の部分が、配分され；メッセージは、時間ステップｉ，ｉ＋１，…，ｉ＋ｄにおけるパケット内の領域の配分された部分に跨がって削除修正コードを用いてコーディングされ；メッセージは、コンピュータベースのデコーダによって、メッセージの作成時刻から測って固定されたデコーディング遅延内にてデコードされる、方法が提供される。

【図面の簡単な説明】

【００１１】

【図１】例示的な分散ストレージシステム内での情報フローを示す図であり、ソースｓはｎ個のストレージノードにまたがってコーディング及び格納されるべき正規化された単位サイズの単一のデータオブジェクトを有しており、データコレクタｔはノードについてのランダムなサブセットｒ内に格納されたデータのみをアクセスすることによって元のデータオブジェクトをリカバーしようと試みる、図である。

【図２】$p_1 \geqq p_2 \geqq p_3$を満たす３つの非均質ノード（ｎ＝３）を備えるシステムについての最適配分表Ｉを示す図である。

【図３】$p_1 \geqq p_2 \geqq p_3 \geqq p_4$を満たす４つの非均質ノード（ｎ＝４）を備えるシステムについての最適配分表ＩＩを示す図である。

【図４】ｎ＝４及び（p_1, p_2, p_3, p_4）＝（0.8479，0.6907，0.6904，0.6725）の場合について、本願開示のアルゴリズム１及び３の各々によって得られた１レベル及び２レベル対称配分の性能を、精密最適解と比較して示す図である。

【図５】本願の開示によって提案されるアルゴリズム及び従来技術のアルゴリズムの性能を、ｎ＝３０の場合のリカバリの失敗確率の観点から表す様々なグラフを示す図である。

【図６】総バジェットＴ＝２及びｎ＝３０の場合の様々な方法におけるノード当たりの分布量を表す様々なグラフを示す図である。

【図７】提案されるアルゴリズムの性能を、ｎ＝５０の場合のリカバリの失敗確率の観点から表す様々なグラフを示す図である。

【図８】総バジェットＴ＝１：７及びｎ＝５０の場合の様々な方法におけるノード当たりの分布量を表す様々なグラフを示す図である。

【図９】様々なアルゴリズムに関して時間コストをノード個数との関係で表す様々なグラフを示す図である。

【図１０】メッセージレートｓ＝１とした場合の、復号失敗確率を１から４までのパケットレートＰとの関係で表す様々なグラフを示す図である。

【図１１】メッセージレートｓの範囲が0.1から１であり、且つパケットレートＰ＝１である場合における、複合失敗確率を表す様々なグラフを示す図である。

【図１２】本願開示の実施形態に係る１レベル対称配分アルゴリズムを示す図である。
【図１３】本願開示の実施形態に係る１レベル対称配分アルゴリズムであって図１２に示されるアルゴリズムの簡略版であるアルゴリズムを、示す図である。
【図１４】本願開示の実施形態に係る２レベル対称配分アルゴリズムを示す図である。
【図１５】本願開示の実施形態に係る３レベル対称配分アルゴリズムを示す図である。
【図１６】メッセージが逐次的にソースに到着し、該メッセージはパケット削除チャネルを介した送信のためにエンコードされ、該メッセージはシンクによって作成時刻からの指定された遅延の間に復号されるべきものとされる、例示的なリアルタイムストリーミングシステムを、示す図である。
【図１７Ａ】パケットが逐次的に送信されるストリーミングパケット構成における、連続する２つの時間ステップにてストリーミングパケットに配分されるストレージを表す図である。
【図１７Ｂ】パケットが逐次的に送信されるストリーミングパケット構成における、連続する２つの時間ステップにてストリーミングパケットに配分されるストレージを表す図である。
【図１８】本願開示の様々な分散ストレージシステムの様々なストレージノードに用いられる例示的ターゲットハードウェアを示す図である。
【発明を実施するための形態】
【００１２】
〔序論〕
本願開示による様々な観点は、所与の総ストレージバジェットについてリカバリが成功する確率を最大化させるように、データオブジェクト（例えば、ファイル、ビデオストリーム、セルラー音声メッセージ等）を、非均質なアクセス確率を有するストレージノード（例えば、ハードウェア依存を有するもの等）のセットに跨がって格納するという問題について検討する。データオブジェクトは、複数のストレージノードに跨がって分割及びコーディングされることができる。例えば、最大距離分離（ＭＤＳ）符号等を用いることによって、アクセスされたデータの総量が少なくとも元のデータオブジェクトのサイズである場合、元のデータオブジェクトをリカバーすることができる。目標は、後に続くリカバリの確率を最大化させるように、ストレージノードのセットに跨がってのストレージバジェットの配分に関しての最適配分を、決定することである。
【００１３】
ストレージ配分の問題は、例えば、非均質なノードを有するピアツーピアクラウドストレージ、コンテンツデリバリネットワーク、遅延トレラントネットワーク、及び無線センサネットワーク等の実際的なストレージに関する問題によって提起される。例えば、各ノードは、アップタイム／ダウンタイム統計が知られている仮想プライベートサーバ（ＶＰＳ）であることができ、該統計をそれぞれに対応するアクセス確率についての指標として用いることができる。別の例では、サーバ用のデータを格納するために用いられるハードディスクラックは、異なる信頼度のストレージハードウェア（例えば、ハードディスク等）を備えることができ、該信頼度は所与のストレージハードウェアの年齢及び／又はそれに対応する製造者による信頼度データ（例えば、平均故障間隔即ちＭＴＢＦ等）によって測られることができ、これは非均質なストレージを伴う等価な分散ストレージをモデリングするための基礎として用いることもでき、本願開示によって提供される様々な解をこれに適用することができる。
【００１４】
可能な配分については大きな探索空間があり、また、所与の配分に関してのリカバリ確率を計算することは#P-hardな問題であるため、ノードが均質（例えば、同一）なアクセス確率を有するという特殊な場合であっても、ストレージ配分は複雑な組み合わせ最適化問題となる（非特許文献１）。この問題のバリエーションは、Ｐ２Ｐネットワーキング（非特許文献２）、無線通信（非特許文献３、非特許文献４）、及び信頼性エンジニアリング（非特許文献５）を含む幾つかの異なる分野において研究されている。ノードが異なる

アクセス確率を有するという非均質な場合については、Ntranosら（非特許文献６）が大偏差不等式及び凸最適化に基づくアルゴリズムを提唱している。

【００１５】

本願開示の様々な実施形態によって提供される（基本的な）対称配分は、適切に選択されたノードについてのサブセットに跨がってバジェットを分散させ、このような配分は例えば計算コストを低減可能であるという点で効率的に構築されることができ、非特許文献６の既存の方法に勝る性能をもたらすことができる。本願開示のさらなる実施形態によれば、基本的な対称配分及びそれに対応する構築アルゴリズムは、複数のｋレベル対称配分へと拡張されることができ、該拡張においては総バジェットがｋ個の部分に分割されて、各部分はノードについての異なるサブセットにわたって均一に分散されて、高位レベル内の各ノードに配分される量が直近レベルの配分量の倍数とされる。換言すれば、直近レベル内のノード当たりの量が単位量とされる。

【００１６】

本願開示による新規な２レベル及び３レベル対称配分についての所与の個数のノードについての性能分析は、本願開示の後続のセクションにて提供される。少数のノード個数ｎ≦４については、網羅的に（例えば、網羅的コンピューティング等によって）決定することができる正確な最適配分は、（例えば、表Ｉ及びＩＩにて与えられる）１レベル及び２レベルの対称配分の中から発見することができる。より大きなノード個数については、本願開示の様々な実施形態によって提供されるこれらの対称配分は、既存の非対称配分の性能に勝る性能をもたらすことができる。さらに、これらの性能は、関連付けられるレベルの個数を増加させること即ち対称配分の柔軟性を増加させることによって、さらに向上させうる。本願開示の後続のセクションにて提示するように、対称配分に関しての様々な手法は、ストリーミングコードの設計に応用されうる。

【００１７】

〔問題についての説明〕

元のデータオブジェクトが単位サイズに正規化されていると仮定する。図１に示すように、データオブジェクトは、適切なコーディングによりｎ個の非均質なノード（１，２，...ｎ）に跨がって格納され、ｎ個のリンク（１０５）を介してソースノードＳ（１１０）がデータオブジェクトをｎ個のノードに供給する。各ノードｉはコーディングされたデータの量x_iを格納し、これは所与の総ストレージバジェットＴに基づく。ストレージバジェットに対しての制約は、例えば次のことに関連付けられることができる：即ち、制限された送信帯域及び／又は制限された格納領域、さらには、各ノードにおいてデータオブジェクトの全体をミラーリングすることが過度な負担をもたらすという意味におけるストレージコストについての制約。データ回収の時点で、データコレクタ（１１５）は、確率p_iで各ノードに独立的にアクセスする。ここで、$0<p_i<1$である。ｒは、データコレクタによってアクセスされたノードについてのランダムなサブセットを表し、ｒの確率分布は、例えばアクセスモデル及び／又は故障モデル（例えば、ノードやリンクが確率論的に故障する可能性があるということ）等によって指定されることができる。この最適化問題を一般化したものは、リカバリを最大化する目的関数と利用可能なストレージに対して限度を付す制約とで表すことができ、次の数式（１）において示される。

【００１８】

【数１】

$$\left.\begin{aligned}&\max_{x_1,\ldots,x_n}\sum_{\mathrm{r}\in\mathcal{P}(\{1,\ldots,n\})}\left(\prod_{i\in\mathrm{r}}p_i\right)\left(\prod_{i\notin\mathrm{r}}1-p_i\right)\cdot 1\left[\sum_{i\in\mathrm{r}}x_i\geq 1\right]\\ &\text{ここで:}\\ &\sum_{i=1}^{n}x_i\leq T\\ &x_i\geq 0\qquad\forall\, i\in\{1,\ldots,n\}\end{aligned}\right\}\quad(1)$$

【００１９】

ここで、１[.]は指標関数（特性関数とも称する）を表す。各ノードは確率ｐ$_{i}$で利用可能となり、結果としてもたらされる利用可能ノードｒ（例えば、アクセス可能なノード等）についてのセットはランダムなものであり、また、サイズもランダムである。数式（１）の目的関数は、リカバリ（あるいはアクセス）確率とみなすことができ、サブセットｒに対応する確率の総和として表すことができ、これは、この場合では総バジェットＴを規定する固定関数であるバジェット関数によって制限的条件（例えば、制約等）が付加された場合におけるデータオブジェクトのリカバリを成功させるサブセットｒに対応する。換言すれば、数式（１）で表された最適化問題は、利用可能なストレージバジェットについて特定の制限が課された状況下でシステムの信頼性を最大化しようとする。本願開示による様々な教示事項及び後続のセクションで提示される様々な教示事項は、このような最適化問題についての解法を提供する。これらの挟持事項は、異なる信頼度の異なるノード内に如何ほどのストレージをプット（配分）すべきかの問題について最適化を図り、これらのノードを備えるシステムの総体的信頼性が最大化される。このような解法は、最適解を得るためのステップバイステップアルゴリズムを表す流れ図においても提供される。当業者は、このような最適化問題についての定式化及び対応する方法に関して熟知している。最適化問題（１）において検討される制約は固定されているが、当業者は、提示される様々な実施形態による教示事項を依然用いたままで、どのようにしてこのような制約を変更して課される要求に、より良く適合させて最適化された配分を得るか、を知っている。例えば、バジェット（ストレージ）に関連付けられているコストを最小化することに興味があるとする。この場合、より多くのストレージを使用すれば、システムは、よりコストのかさむものとなる。このような場合、制約がコスト（目的）関数となり、また、信頼性目標が信頼性制約となり、これらについて最適化問題が解かれることになる。換言すれば、所望の目的に応じて、先述の目的関数及び制約を入れ替えることができる。代替的には、目的関数をコスト及び信頼性についての重み関数とすることができる。

【００２０】

簡略化のため、Ｔ＜ｎであると仮定する。なぜならば、Ｔ≧ｎについては最適な配分が自明なもの：{1,…,1}となるからであり、ここでは各ノードは元のデータオブジェクトについてコピーを１つ格納する。また、ｘ$_{i}$≦１であると仮定することができる。なぜならば、ノードがデータについて１よりも多い個数のコピーを格納することは無益であるからである。

【００２１】

ｎが大きい場合においては数式（１）によって定義された最適化問題は計算的に困難となるが、小さいｎについては、サブセット（例えば、ノードのそれ等）の可能な組合せ全てに関して網羅的なサーチを行うことによって最適配分を決定することが可能となる。異なるシナリオについて網羅的サーチによって得られた最適配分に関するリストは、それぞれ図２及び図３の表Ｉ及び表ＩＩに示されている。表Ｉはｎ＝３であり、ｐ$_{1}$≧ｐ$_{2}$≧ｐ$_{3}$についての最適配分を示し、表ＩＩはｎ＝４であり、ｐ$_{1}$≧ｐ$_{2}$≧ｐ$_{3}$≧ｐ$_{4}$についての最適配分を示す。

【００２２】
〔１レベル対称配分〕
本願開示の実施形態によれば、ノードのサブセットにわたって総バジェットＴを均一に分散することによって、基本的な１レベル対称配分を得る。より高い確率のノードを用いることによってシステムの信頼性が向上する。したがって、ノードは、$p_1 \geqq p_2 \geqq \cdots \geqq p_n$となるように、アクセス確率について降順で配列され、数式（１）によって示された最適化問題は、サイズｍのサブセットを選択することに単純化され、これは、次の配分に対応する。
【００２３】
【数２】

$$\left\{x_1 = \frac{T}{m}, \ldots, x_m = \frac{T}{m},\ x_{m+1} = 0, \ldots, x_n = 0\right\}$$

【００２４】
リカバリを成功させるには、ｍ個の空でないノードについて少なくとも

$\left\lceil \frac{m}{T} \right\rceil$

のノードにアクセスすることが必要となる。Ｍ個の空でないノードにアクセスする確率は、次の通りである。
【００２５】
【数３】

$$P_M = \sum_{\mathbf{r} \in \mathcal{P}(\{\binom{m}{M}\})} \left(\prod_{k \in \mathbf{r}} p_k\right)\left(\prod_{l \notin \mathbf{r}} (1 - p_l)\right) \qquad (2)$$

【００２６】
ここで、

$\mathcal{P}\left(\left\{\binom{m}{M}\right\}\right)$

はサイズがＭである{1,…, ｍ}のサブセットについてのセットを表す。
【００２７】
したがって、リカバリが成功する確率は次の数式（３）によって表される。
【００２８】
【数４】

$$\begin{aligned} P_{\text{successful}} &= \sum_{M=\lceil \frac{m}{T} \rceil}^{m} P_M \qquad (3) \\ &= \sum_{M=\lceil \frac{m}{T} \rceil}^{m} \sum_{\mathbf{r} \in \mathcal{P}(\{\binom{m}{M}\})} \left(\prod_{k \in \mathbf{r}} p_k\right)\left(\prod_{l \notin \mathbf{r}} (1 - p_l)\right) \end{aligned}$$

【００２９】
本願開示の実施形態によれば、図４に示されるように、アルゴリズム１は、数式（２）及び（３）を用いることによって最適な１レベル対称配分を発見する。本願における「レベル」との用語は、ゼロでないバジェットで配分されたノードの各々が総バジェットの量と同量を含むことになる配分を意味することができる。
【００３０】
ｍの取り得る値全てを検討することに代えて、［１，セクションＩＩ－Ｃ］で提示したものに類似しているアプローチを考えることができ、

$\left\lceil \frac{m}{T} \right\rceil$

についての個別の値各々に対応する最大のｍにアルゴリズムの範囲を限定することができ、それによって候補値

$$m \in \{\lfloor T \rfloor, \lfloor 2T \rfloor, \ldots, \lfloor K_{max} T \rfloor\}$$

が得られる。ここでＫｍａｘは、

$$\lfloor KT \rfloor \leq n$$

となる最大の整数Ｋである。しかし、依然として大きい値のｎについては、アルゴリズム１は計算的に困難であり続ける。なぜならば、ｎの値に応じて関連付けられる計算的複雑性が急速に増大するからである。

【００３１】

本願開示の実施形態によれば、アルゴリズム１の複雑性をさらに減少させるために、各確率ｐｉを確率ｐｉの平均値で代替することができる。これは、

$$p_i' = p_{avg} = \frac{\sum_{i=1}^{m} p_i}{m}$$

と表され、結果として、図５に示すアルゴリズム２が得られ、これはアルゴリズム１の簡略版であるとみなすことができる。元のアルゴリズム１と比べて、この簡略化（ｐａｖｇを用いること）によってアルゴリズム（例えば、ハードウェアプロセッサ上で実行される対応するプログラムコード等）の実行時間を大幅に減少させることができる。本願開示の後続のセクションで提示される数値実験によると、図５のアルゴリズム２によって説明された簡略化版が、その単純さにもかかわらずすこぶる良好な性能を提供することが分かる。図５の対応するステップ（１３１）～（１３８）で説明されたアルゴリズム２（及び本願開示で説明された他の新規なアルゴリズム）が、本願開示の現セクションにおいて提供された数式についての解法及び関連する簡略化手法についての例示的実装例にすぎないことを当業者は容易に理解するであろう。

【００３２】

〔２レベル及び３レベル対称配分〕

表Ｉ及びＩＩ内の最適配分に関して得た知見に基づいて、そして、一部の最適解は所与のバジェットＴについて２レベルの量配分を有するということに着眼して、発明者は２レベル対称配分についてのより一般的なクラスを検討した。これにおいては、各レベルは利用可能なノードの総体についての互いに素となるサブセットを備え、第１レベルの各ノードに配分される量は第２のレベルの各ノードに配分される量の２倍であり、次の通りである。

【００３３】

【数５】

$$\Big(\underbrace{x_1, \ldots, x_{m_1}}_{\text{level 1}}, \underbrace{x_{m_1+1}, \ldots, x_{m_1+m_2}}_{\text{level 2}}, x_{m_1+m_2+1}, \ldots, x_n\Big), \tag{4}$$

ここで

$$\begin{aligned} &x_1 = \cdots = x_{m_1} = S_1, \\ &x_{m_1+1} = \cdots = x_{m_1+m_2} = S_2, \\ &x_{m_1+m_2+1} = \cdots = x_n = 0, \\ &S_1 = \frac{2T}{2m_1 + m_2}, \\ &S_2 = \frac{T}{2m_1 + m_2}. \end{aligned} \tag{5}$$

【００３４】

各ノード内に格納されるデータの量は、Ｓ２の倍数として表されることができるため、

$S_2 = \frac{T}{2m_1+m_2}$

であり、S_2は定量又は基礎的ストレージブロックとして扱われることができる。先のセクションの手法を踏襲すると、$2m_1+m_2$個の基礎的ストレージブロックのうち少なくとも

$$\left\lceil \frac{2m_1+m_2}{T} \right\rceil$$

がアクセスされるとリカバリが成功するということに発明者が気付いた。ここで、

$$m_{acc} \in \{0, \dots, 2m_1 + m_2\}$$

は、アクセスされた基礎的ストレージブロックの総個数を表し、また、

$$m_{acc,L} \in \{0, \dots, m_L\}$$

は、レベルＬ内でアクセスされたノードの個数を表すものとする。したがって、$m_{acc}=2_{macc,1}+m_{acc,2}$である。よって、リカバリの成功を可能とする全てのペア（$m_{acc,1}$，$m_{acc,2}$）についてのセットである

$$\mathcal{M}_{m_1,m_2}$$

は、次の数式（６）で表されることができる。

【００３５】

【数６】

$$\left.\begin{aligned} \mathcal{M}_{m_1,m_2} = \{(m_{acc,1}, m_{acc,2}) : \quad & 2m_{acc,1} + m_{acc,2} \geq \left\lceil \frac{2m_1+m_2}{T} \right\rceil, \\ & m_{acc,1} \in \{0, \dots, m_1\}, \\ & m_{acc,2} \in \{0, \dots, m_2\}\}. \end{aligned}\right\} \quad (6)$$

【００３６】

ここで発明者は、簡略化のため、２レベル対称配分の第１レベル（例えば、レベル１等）に配分されるノードの量である$mOpt_1$が、先述の１レベル対称配分スキームにおける$mOpt_{oneLevel}$よりも小さい、という仮定を導入する。簡略化のための仮定及び１レベル対称配分にて得られた$mOpt_{oneLevel}$の値に基づいて、並びに本願開示のさらなる実施形態によれば、図６にて説明されるアルゴリズム３が得られる。アルゴリズム３は、アルゴリズム２に大筋で類似した態様で２つのレベル（例えば、レベル１、レベル２等）における確率を計算することによって、（$mOpt_1$，$mOpt_2$）で与えられる良好な２レベル対称配分を効率的に探索する。

【００３７】

１レベル対称配分のケースと同様にして、アルゴリズム３は（m_1，m_2）についての特定の選択肢のみを検討することによって探索空間を狭める。具体的には、m_1についての所与の選択肢については、該アルゴリズムは、

$$\left\lceil \frac{2m_1+m_2}{T} \right\rceil$$

の個別の値各々に対応する最大のm_2に注目範囲を限定し、これによって以下の候補値が得られる。

【００３８】

【数７】

$$m_2 \in \{\lfloor K_{min}T - 2m_1 \rfloor, \lfloor (K_{min}+1)T - 2m_1 \rfloor, \dots, \lfloor K_{max}T - 2m_1 \rfloor\},$$

【００３９】

ここでK_{min}は、

$$\lfloor KT - 2m_1 \rfloor \geq 0$$

となる最小の整数Ｋであり、K_{max}は、

$$\lfloor KT - 2m_1 \rfloor \leq n - m_1$$

となる最大の整数Ｋである。

【００４０】

発明者の知見によると、ＭＯｐｔ$_2$の値は多数のループ内では希にしか変化しないものであり、ＭＯｐｔ$_2$についての値の変化が２つあれば、これらは通常ループのサイズの３分の２の距離Ｄの範囲内に存在する。したがって、アルゴリズム３では、アルゴリズムのステップ（１４０２）～（１４１３）で定義されるインナーループ内のＭＯｐｔ$_2$がループのサイズの３分の２にわたって変化しない場合、アルゴリズムはこのＭＯｐｔ$_2$が現在ループ内の最適選択肢であるとみなした上で次のループへと続行し、これによってアルゴリズムの（例えば、ハードウェアプロセッサ上で実行された場合等の）実行時間が大幅に削減され、同時にほぼ同じ性能が得られることになる。また、Ｄはより小さい値に設定することができ、これによってアルゴリズムの実行は加速されるが、準最適解が発生することもあり得る。

【００４１】

本願開示のさらなる実施形態によれば、３つのレベルをサポートする対称配分についてのさらなる一般化手法が、図７のアルゴリズム４にて例示的な形態で説明及び提示されている。ここで、３つのレベルのサブセットのサイズをそれぞれｍ$_1$，ｍ$_2$，ｍ$_3$とし、第１レベルでのノード当たりの量は第３レベルでのノード当たりの量の３倍であり、第２レベルでのノード当たりの量が第３レベルでのノード当たりの量の２倍である。リカバリの成功を可能にする全てのトリプル（ｍ$_{acc,1}$，ｍ$_{acc,2}$，ｍ$_{acc,3}$）についてのセット

$$\mathcal{M}_{m_1,m_2,m_3}$$

は、次式で表すことができる。

【００４２】

【数８】

$$\mathcal{M}_{m_1,m_2,m_3} = \left\{ (m_{acc,1}, m_{acc,2}, m_{acc,3}) : \begin{array}{l} 3m_{acc,1} + 2m_{acc,2} + m_{acc,3} \geq \left\lceil \frac{3m_1+2m_2+m_3}{T} \right\rceil, \\ m_{acc,1} \in \{0, \ldots, m_1\}, \\ m_{acc,2} \in \{0, \ldots, m_2\}, \\ m_{acc,3} \in \{0, \ldots, m_3\} \}. \end{array} \right. \quad (7)$$

【００４３】

発明者は、簡略化のため、３レベル対称配分におけるｍ$_1$が、本願開示にて先述した２レベル対称配分のｍＯｐｔ$_{1,twoLevel}$＋ｍＯｐｔ$_{2;twoLevel}$よりも小さいという仮定を導入する。簡略化のための仮定及び本願開示の実施形態によれば、図７にて説明されるアルゴリズム４が得られる。アルゴリズム４は、数式（１）で表される最適化問題についての良好な３レベル対称配分を効率的に探索する。

【００４４】

本願開示の先のセクションにて提示された対称配分でなされたと同様に、特定の選択肢（ｍ$_1$，ｍ$_2$，ｍ$_3$）のみを検討することによって探索空間を狭めることができる。具体的には、ｍ$_1$及びｍ$_2$についての所与の選択肢について、発明者は

$$\left\lceil \frac{3m_1+2m_2+m_3}{T} \right\rceil$$

の個別の値各々に対応する最大のｍ$_3$に注目範囲を限定し、これによって、次の候補値が得られる。

【００４５】

【数９】

$$m_3 \in \{\lfloor K_{min}T - 3m_1 - 2m_2 \rfloor, \lfloor (K_{min}+1)T - 3m_1 - 2m_2 \rfloor, \ldots, \lfloor K_{max}T - 3m_1 - 2m_2 \rfloor\},$$

【００４６】

ここで、K_{min}は、

$$\lfloor KT - 3m_1 - 2m_2 \rfloor \geq 0$$

を満たす最小の整数Ｋであり、K_{max}は、

$$\lfloor KT - 3m_1 - 2m_2 \rfloor \leq n - m_1 - m_2$$

を満たす最大の整数Ｋである。また、距離閾値Ｄは、アルゴリズムのステップ（１５０３）～（１５１５）で定義された（インナー）ループのサイズの３分の２となるように設定される。

【００４７】

本願開示の様々な実施形態によって提案される対称配分は、容易にｋレベルをサポートするように一般化されることができる。ここでｋ＝４，５，６，...等である。このようなｋレベル対称配分では、空でないノードが各々ｃ、２ｃ，...，ｋｃのデータの量を格納するようにストレージバジェットＴを配分することができる。ここでｃは対応する定量又は基礎的ストレージブロックS_kのサイズであり

$$S_k = \frac{T}{km_1 + (k-1)m_2 + \cdots + 2m_{k-1} + m_k}$$

である。

【００４８】

〔数値実験〕

このセクションでは、上述のアルゴリズムの性能を、リカバリ失敗確率の観点から評価する。

【００４９】

〔Ａ．　ｎ＝４の場合における表ＩＩの最適解との比較〕

ｎ＝４については、アルゴリズム１及びアルゴリズム３によって得られた１レベル及び２レベル対称配分の性能を、網羅的に計算された表ＩＩ（図３）から得られた精密最適解と、それぞれ比較することができる。発明者は、１≦Ｔ≦４となるストレージバジェット範囲と、０．５から１の間で独立的かつ画一的に選出されたノードアクセス確率p_i（１≦ｉ≦ｎ）について検討した。図８は、アクセス確率について、典型的なケースにおける結果を示す。本願開示の実施形態による１レベル及び２レベル対称配分は、一般的には精密最適配分良好に近似するものとなることを発明者は発見した。また、特定の限定されたＴの範囲については、最適解を得るためには２レベル対称配分が必要となることをも発明者は発見した（例えば、本件においてはm_1＝１，m_2＝３の場合）。

【００５０】

〔Ｂ．　ｎが大きい値となる場合における他の既存のアルゴリズムとの比較〕

このセクションでは、大きい値のｎについての性能に関して、アルゴリズム２、３、４を、非特許文献６で提示される既存のアルゴリズム（即ち、最大分散、チェルノフ閉形式、チェルノフ反復）と、比較する。比較においては、バジェット範囲は１．２≦Ｔ≦２とし、また、ノードアクセス確率p_iは０．５から１の間で独立的かつ画一的にランダムに選出されたものとする。故障確率曲線は、数値的モンテカルロシミュレーション実験を用いて得られる（なお、グラフもこれによって作成される）。

【００５１】

異なる問題スケール（ｎ＝３０，５０）についての結果は図９、１０、１１、１２に示される。図９及び１１では、本願開示の様々な実施形態による１レベル、２レベル及び３レベル対称配分によって生成された故障確率曲線、並びに、先行技術文献である非特許文献６で提示されたチェルノフ反復法及びチェルノフ閉形式法によって生成された曲線が、

グラフに示されている。これらの曲線によれば、バジェットＴが小さい場合、対称配分は常にチェルノフ系手法に勝る性能をもたらす。Ｔが大きい場合（例えば、ｎ＝３０についてＴ＝２の場合等）、１レベル及び２レベル対称配分がチェルノフ系手法に劣る場合があっても、チェルノフ系手法に勝る性能をもたらし得る３レベル対称配分は存在する。図１０及び図１２にて、バジェットＴ＝２及びバジェットＴ＝１．７についての、異なる手法についての量の分布を示すプロットがそれぞれ示されている。これらのプロットは、対称系の手法がより柔軟なものとされてより多くのレベルについてのサポートが設けられれば、性能がさらに向上し得ることを示唆している。

【００５２】

〔Ｃ．　様々な方法の時間コスト〕

図１３は、様々なアルゴリズムについて、ノード個数との関係での時間コストを示す。この図によれば、１レベル対称配分は他の配分手法よりも常に早く、また、ｎ＜５５の場合に関しては２レベル対称配分がチェルノフ反復アルゴリズムよりも早い、ということが分かる。３レベル対称配分が多くのシナリオでは最も遅く、より高い性能を求める場合の妥協案として用いられることができる。

【００５３】

〔Ｄ．　アルゴリズムの応用：対称的時間不変イントラセッションコード〕

分散ストレージ用途に加えて、対称配分アルゴリズムは、リアルタイムストリーミング用のコード設計の問題にも応用されることができ（非特許文献７、８）、例示的な概略的構成図が図１６に示されている。この事例では、メッセージが逐次的にソース（１６１０）に到着し、パケット削除チャネル（１６２０）を介したシンク（１６３０）への送信を行うために該メッセージは（例えば、先行技術においていう所謂削除修正コードを用いて）エンコードされ、該シンクは指定された遅延内にてメッセージを逐次的に復号する必要がある。様々なネットワークシナリオでは、パケット遅延は変動し得るものであるため、遅延と共にパケット受信の確率が上昇する。ストリーミング問題におけるメッセージ作成時刻とデコーディング期限との間での送信パケットを、ストレージ配分問題のノードとして同視すれば、対応する非均質遅延依存性の損失確率（例えば、非均質信頼度モデル等）を与えられれば、対応する非均質ストレージ配分問題に関しての解法は対応するストリーミング問題に関してのイントラセッションコードに転じることになり、また、ストレージ配分問題におけるストレージバジェットが総利用可能パケットサイズに転じる。

【００５４】

図１６をさらに参照する例として次のケースを想定する。即ち、画一的なサイズｓ＞０の独立しているメッセージ（１，２，３...）がソース（１６１０）にて規則的な間隔たるｃ＝１時間ステップ毎に作成されており、また、該メッセージはレシーバにてそれぞれの作成時刻から遅延ｄ＝１０時間ステップ内にてデコードされなければならない。非特許文献９の図５（ａ，ｂ）の遅延束縛違反確率を、各パケットの異なるフェイリア確率として用いる。ソースはサイズＰ＞ｓの単一のデータパケットを、時間ステップ毎に、リンク上に送信することを許されている。したがって、対応する仮想バジェットは、＝Ｐ／Ｓとして表すことができる。

【００５５】

固定されたメッセージレートをｓ＝１として、パケットレートＰを１～４の範囲内とした場合について、図１４の故障確率プロットが得られる。また、固定されたパケットレートをｐ＝１として、メッセージレートｓを０．１～１の範囲内とした場合について、図１５の故障確率プロットが得られる。

【００５６】

本願開示の様々な実施形態による例示的なストリーミング構成に関しては、図１７Ａ及び１７Ｂが、ｃ＝１かつｄ＝４のケースについての例示的なパケット分布を示す。図１７Ａ及び１７Ｂのグラフの列は逐次的に送信されたパケットを表し、任意の時刻（例えば、ｔ軸等）に送信される全てのパケットの総和がバジェット（例えば、総パケットサイズ等）によって制約される。図１７Ａ及び１７Ｂの網掛けされたボックスは時間ステップ１に

て作成されて幾つかの送信されたパケット（p_0，p_1，p_2，p_3）に分散されたデータ（x_1，x_2，x_3，x_4）の量を表す。これらのパケットは、これらがデコードされなければならない遅延の範囲に従って送信される。ハッチングが掛けられた図１７Ｂのボックスは、各送信パケット内の時間ステップ２（例えば、時間ステップ２において作成された場合）におけるデータの量を表す。時間ステップｉにて送信されたパケットが時間ステップｊまでに受信される確率はp_{j-i}であり、遅延の値Δ＝ｊ－ｉについて、p_Δは遅延Δについて増加する関数である。各送信事象にて利用可能な最大総パケットサイズ（例えば、バジェット等）とデコーディング確率p_Δを与えられれば、各送信されたパケット（p_0，p_1，p_2，p_3）についての最適配分パケットサイズ（x_1，x_2，x_3，x_4）を、最適化問題（１）を用いて導出することができる。よって、これに関連する解法は、非均質ストレージ配分問題との関係で提示した解法に従うこととなる。本願開示によるこのような教示事項とこれら２つの問題間での相似性とを与えられれば、当業者はこのような教示事項を、非均質な構成要素を備える他のシステムにも適用できる。

【００５７】

本願開示の様々な実施形態及び先のセクションで述べた様々な実施形態において、非均質分散ストレージ配分最適化の問題に関して、複数レベルの対称配分を探索する数式（１）によってもたらされる様々なアルゴリズムが提供されている。レベル数及び対称配分の柔軟性が増大するにつれて、アルゴリズムの実行時間（例えば、コンピュータプロセッサを用いる場合の実行時間）が延びる。しかし、提示した数値実験によれば、本願開示の様々な実施形態に基づいて提案された対称配分は、少ないレベル（例えば、１レベルや２レベル等）を伴うものであって、比較的低い複雑性を伴い、比較的短い実行時間を伴うものであったとしても、多くのシナリオにて既存のより複雑な配分スキームに勝る性能をもたらすことができることが示されている。

【００５８】

〔ハードウェア考慮事項〕

本願開示の図１にて提示され、かつ前述のセクションで説明された分散非均質システムについて再度言及すると、図１内の対応するボックス（例えば、「ストレージノードｉ」と称されるボックス）によって表されるストレージノードが様々なハードウェア／ファームウェア／ソフトウェアの組合せの形をとることができ、これによって提供される様々なアルゴリズムの効率性に影響を与えずに、図１に表されるシステムの非均質的性質を作出することができることを、当業者は容易に理解することができる。各ストレージノードは、実質的に、例えば、単一のハードディスクドライブ、単一のコンピュータベースのワークステーション、ネットワーク化されたサーバを備えるデータセンタ、ストレージアレイ又は単一のインターネットプロバイダ等のデータを格納することができるエンティティとみなされることができる。

【００５９】

１つの例示的な実装例によれば、各ストレージノードは異なるクラウドサービスプロバイダであることができ、関連付けられた仮想プライベートサーバ（ＶＰＳ）及び対応するストレージシステムを用いて、データオブジェクトx_iの部分についてのソース（１１０）とストレージと通信することができる。各クラウドサービスプロバイダ（例えば、ストレージノード等）は既知の統計的アップタイム／ダウンタイム（例えば、ノードが８０％又は７０％又は...％の時間だけ利用可能となっていること等）を有していることができ、これをさらに用いて、データコレクタ（１１５）を介してのデータオブジェクトリカバリに関しての関連付けられたアクセス／信頼度モデル（例えば、確率p_iでクラウドサービスが利用可能となっていること等）を導出することができる。

【００６０】

別の例示的実装例によれば、各ストレージノードは、ソース（１１０）のための主ストレージとして用いられるデータストレージラック内に用いられる物理的ハードディスクであることができる。例えば、ソースはサーバコンピュータであることができる。追加的には、各ハードディスクは、ハードディスク製造者によって提供される例えば平均故障間隔

（ＭＴＢＦ）パラメータ等に基づいている関連付けられた故障曲線を有していることができる。このような各ハードディスクについてのＭＴＢＦ及びハードディスクの年齢を用いて、データコレクタワークステーションによるデータオブジェクトのリカバリに関しての対応するアクセス／信頼度モデルを導出することができる。

【００６１】

一部の事例では、ストレージノードの信頼度／アクセス可能性は、ノードの様々な構成要素の品質についての関数として表され得る。例えば、再生品ドライブや低グレード構成要素を用いて製造されたドライブ等の安価なハードディスクドライブは、最新のハードディスクに比してより低い信頼度を有することができる。別の例を挙げるならば、古めのインフラストラクチャ（例えば、ハードウェア、コンピュータ、ストレージシステム、ケーブル類及び通信リンク等...）を用いているサービスプロバイダは、最新のインフラストラクチャを用いているプロバイダよりも安いが、劣った信頼度／アクセス可能性を有し得る。このような場合において数式（１）によって記述された最適化問題は、分散された非均質ストレージ空間に関連付けられるコストを勘定に入れて、かつ、データオブジェクトのリカバリ可能性を確保しつつそのようなコストの最小化を図ることができる。したがって、本願開示による様々なアルゴリズムにて説明されて提供された方法は、総システムコストを削減するため及び信頼度を維持するために用いられることができる。

【００６２】

図１に示され、様々なストレージノード及びステーション（１１０、１１５）を接続する様々な通信リンクは、ツイステッド導線、同軸ケーブル、イーサネットケーブル、電気的／光学的信号を搬送するために導体を用いる任意のタイプのケーブル等の導体を用いる有線接続及び／又は空中に送出されるＲＦ信号を用いる無線リンクであることができる。このようなリンクの幾つかは、有線及び無線のリンクの組合せを用いて実現することができる。

【００６３】

本願開示の様々な実施形態では、ソース（例えば、図１の１１０、図１６の１６１０）は、本願開示によって提供される様々なアルゴリズム（例えば図４乃至７）の１以上を包含するプログラムコード（例えば、ソフトウェア等）を実行しているコンピュータプロセッサであることができる。このようなプログラムコードは、本願開示の様々なアルゴリズムによる様々なステップを包含することができる。このようなコンピュータプロセッサは、合理的な程度の演算能力及びメモリサイズを有する、メインフレーム、マイクロコンピュータ、デスクトップ機（ＰＣ、Ｍａｃ等）、ラップトップ機、ノートパソコン等の市販品又は具体的なタスクのために設計されたプロプライエタリハードウェアを含む任意のターゲットハードウェア（例えば、後述の図１８）によって実装されることができ、これらはマイクロプロセッサやデジタル信号プロセッサ（ＤＳＰ）や様々なＦＰＧＡ／ＣＰＬＤ等を含むことができる。当業者は、このようなコンピュータプロセッサについての、異なるオペレーティングシステム（ＯＳ）及び／又はハードウェア構成を用いる多種多様な構成をよく知っている。このため、１以上の様々なアルゴリズムを包含するプログラムコードは、ソースの任意の特定のターゲットハードウェアに向けて適合されることができる。

【００６４】

本願開示にて説明された方法（例えば、図４乃至７に示されている単一及び複数レベル対称配分アルゴリズム並びに関連するステップバイステップチャート等）及び対応する分散ストレージ通信システムは、ハードウェア、ソフトウェア又はこれらの組合せによって実装されることができる。モジュール、ノード又は構成要素として説明された要素は、集合的に又は個別的に、ハードウェア、ソフトウェア及び／又はファームウェアの組合せを用いて実装されることができる。本願開示の方法（例えば、フローチャートやアルゴリズム等）のソフトウェア部分は、コンピュータ可読媒体を備えることができ、該媒体は命令（例えば、実行可能プログラム等）を備えることができ、該命令の実行によって説明された方法の少なくとも一部が行われ、本願開示の様々な実施形態に基づく対称配分アルゴリズムの一部の又は全体的な構築等が該命令の実行によって行われる。コンピュータ可読媒

体は、例えばランダムアクセスメモリ（ＲＡＭ）及び／又は読み出し専用メモリ（ＲＯＭ）を含むことができる。命令は、プロセッサ（例えば、デジタル信号プロセッサ（ＤＳＰ）等）、特定用途向け集積回路（ＡＳＩＣ）、フィールドプログラマブル論理アレイ（ＦＰＧＡ）又はこれらの組合せによって実行されることができ、これらは単一の集積回路（ＩＣ）と統合されていることができる。

【００６５】

図１８は、本願開示の様々な教示内容による、１以上のコード／サブグラフ構築方法を実装するための（例えば、ソースノード等に関する）及び／又はそのようにしてエンコードされたデータを復号するための（例えば、リレー／シンクノード等に関する）ターゲットハードウェア（１０）（例えば、コンピュータシステム等）についての例示的実施形態である。このターゲットハードウェアは、プロセッサ（１５）と、メモリバンク（２０）と、ローカルインターフェースバス（３５）と、１以上の入出力装置（４０）を備える。プロセッサは、メモリ（２０）内に格納されている何らかの実行可能プログラムに基づいて、提供される様々なコーディング構築方法の実施形態に関連する１以上の命令を、オペレーティングシステム（２５）が提供されることにより（例えば、全体的に又は部分的に）実行することができる。これらの命令は、ローカルインターフェース（３５）を介して、ローカルインターフェース及びプロセッサ（１５）に特有な何らかのデータインターフェースプロトコルに準拠して、プロセッサ（２０）へと伝達される。ローカルインターフェース（３５）は、プロセッサベースのシステムの複数の要素間におけるアドレス、制御及び／又はデータ接続を提供することに一般的に向けられた、コントローラ、バッファ（キャッシュ）、ドライバ、リピータ及びレシーバ等の複数の要素に関しての象徴であることに留意されたい。幾つかの実施形態においては、幾らかのローカルメモリ（キャッシュ）をプロセッサ（１５）に設けることができ、実行されるべき命令の一部をここに格納して実行速度を幾らか向上させることができる。プロセッサによる命令の実行に際しては何らかの入出力装置（４０）の使用が必要となり得るのであり、その場合においては、エンコード及び／又はデコードされるべきメッセージを含むビットストリームデータを入力すること、キーボードからのコマンドを入力すること、表示装置へデータを出力することまたは、通信チャネルを介して送出されるべきエンコードされたデータパケット（例えば、本願にて提供される方法によるもの）を出力すること、通信チャネルからのデータパケットを入力すること等が必要とされ得る。一部の実施形態では、オペレーティングシステム（２５）は、プログラムの実行のために必要とされる様々なデータ及び命令を集約するための中央要素となることによってこれらのタスクを促進して、これらをマイクロプロセッサに提供する。一部の実施形態では、オペレーティングシステムは存在しない場合があり、全てのタスクはプロセッサ（１５）による直接的制御の下に置かれるが、ターゲットハードウェア装置（１０）の基本的アーキテクチャは図１８に示したもののままとされる。一部の実施形態では、複数のプロセッサを並列構成において用いて実行速度を増すことができる。このような場合においては、実行可能プログラムは、並列実行のために特に適合させたものとされることができる。一部の実施形態では、プロセッサ（１５）が本願開示にて提供される方法の一部を実行することができ、他の部分は、ローカルインターフェース（３５）を介してターゲットハードウェア（１０）によってアクセス可能な入出力ロケーションに配置された専用のハードウェア／ファームウェアを用いて、実装されることができる。ターゲットハードウェア（１０）は、複数の実行可能プログラム（３０）（例えば、特殊な通信プログラム等を含む）を含むことができ、各々は互いの関係で独立的に又は複合的に実行されることができる。これらの実行可能プログラムは、プロセッサによって実行されると、本願開示において提示された方法の少なくとも一部（例えば、構築アルゴリズム、エンコーディング、デコーディング等）を行うための命令を備えることができる。

【００６６】

図１８に示した例示的コンピュータハードウェアは、集積回路（ＩＣ）内にて実装されることができる。本願開示の幾つかの実施形態によれば、本願開示の様々な実施形態（例

えば、アルゴリズム等）を全体的又は部分的に実装する対称配分モジュールは、ＩＣ内にて部分的又は全体的に実装されることができる。このようなＩＣは、システムの一部として用いられて、本願開示の様々な実施形態による非均質システムについての対称配分を提供するために用いられることができる。ＩＣ性能及び／又は他のシステム要件に基づいて、例えば配分及び／又はアルゴリズム実行速度についての特定の所望性能を満たすように、本願開示のアルゴリズムを適合させるようにして、ＩＣのメモリ（例えば、プログラマブルメモリ等）内に格納されているプログラムをアップグレードすることができる。当業者は、提供されている対称配分の方法及び図４乃至７にフローチャートが示されているアルゴリズムを用いることによって、図１及び図１６に示されたシステムについての様々な他のハードウェア／ソフトウェア／ファームウェアに関しての実装例を、部分的に又は全体的に着想することができる。

【００６７】

上述の例は、非均質システム用の分散ストレージ配分並びに関連する単一及び複数レベル対称配分アルゴリズムについての実施形態を如何にして作成及び使用するかについての完全な開示及び説明を、当業者に与えるために提供されているのであって、発明者が自己の開示として認識する範囲を限定するものとしては意図されていない。本願開示を実施するための上述した実施態様に関して施される変更は、情報／コーディング／通信についての理論分野及び処理分野の当業者によって用いられることができ、また、後述の請求の範囲の範疇内となるものと意図されている。本願明細書にて言及した全ての特許及び刊行物は、本願開示が関連する分野の当業者の技術水準を指し示すものであることができる。本願開示にて引用された全ての参照文献は、仮にそれらが個別的にそれらの全体について参照によって取り込まれたときと同じ効果をもって参照により取り込まれる。

【００６８】

本願開示は、特定の方法又はシステムに限定されないものと理解されるべきであり、もちろんこれらは多種多様である。また、本願にて用いる用語は具体的な実施形態を説明するためだけに用いられているのであり、限定的なものとしては意図されていない。本願明細書及び添付の請求の範囲において用いられる不定冠詞及び定冠詞を伴う単数形は、内容が反対を明確に指示している場合を除いて、複数形の指示対象を含む。「複数の」との用語は、内容が反対を明確に指示している場合を除いて、２以上の指示対象を含む。明確に他の定義がされていない限り、本願にて用いられた技術的及び科学的な用語の全ては、本願開示が属する技術分野の当業者が一般的に理解する意味と同じ意味を持つ。

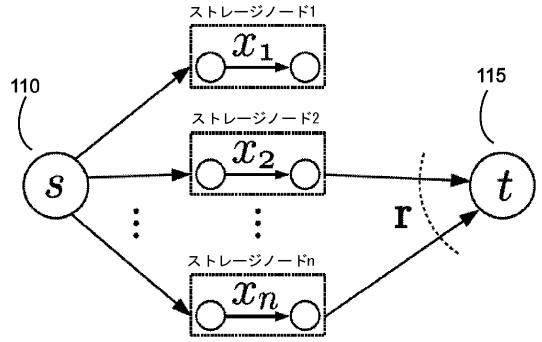
【００６９】

本願開示において、参考文献欄で引用されて［.］で示される９つの文献であって題名・著者・出版情報が提供される文献は、各々の全体について、参照によって取り込まれる。

【００７０】

本願開示について幾つかの実施形態を説明した。もっとも、本願開示の趣旨及び範囲から逸脱せずに様々な変更を加えることができるということは理解されるべきことである。したがって、他の実施形態も後述の請求の範囲の範疇内に含まれることになる。

【図１】



【図２】

表Ｉ

バジェットT	最適配分 (x_1, x_2, x_3)
$1 \le T < \frac{3}{2}$	(1,0,0)
$\frac{3}{2} \le T < 2$	(1,0,0) if $-p_1 + p_1p_2 + p_1p_3 + p_2p_3 - 2p_1p_2p_3 \le 0$
	$(\frac{1}{2}, \frac{1}{2}, \frac{1}{2})$ if $-p_1 + p_1p_2 + p_1p_3 + p_2p_3 - 2p_1p_2p_3 \ge 0$
$2 \le T < 3$	(1,1,0)

【図３】

表ＩＩ

バジェットT	最適配分 (x_1, x_2, x_3, x_4)
$1 \le T < \frac{4}{3}$	(1, 0, 0, 0)
$\frac{4}{3} \le T < \frac{3}{2}$	(1, 0, 0, 0) if $-p_1 + p_1p_2p_3 + p_1p_2p_4 + p_1p_3p_4 + p_2p_3p_4 - 3p_1p_2p_3p_4 \le 0$
	$(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ if $-p_1 + p_1p_2p_3 + p_1p_2p_4 + p_1p_3p_4 + p_2p_3p_4 - 3p_1p_2p_3p_4 \ge 0$
$\frac{3}{2} \le T < \frac{5}{3}$	(1, 0, 0, 0) if $-p_1 + p_1p_2 + p_1p_3 + p_2p_3 - 2p_1p_2p_3 \le 0$
	$(\frac{1}{2}, \frac{1}{2}, \frac{1}{2}, 0)$ if $-p_1 + p_1p_2 + p_1p_3 + p_2p_3 - 2p_1p_2p_3 \ge 0$
$\frac{4}{3} \le T < 2$	(1, 0, 0, 0) if $-p_1 + p_1p_2 + p_1p_3 + p_2p_3 - 2p_1p_2p_3 \le 0$ and $-p_1 + p_1p_2 + p_1p_3 - p_1p_2p_3 + p_1p_4 - p_1p_2p_4 - p_1p_3p_4 + p_2p_3p_4 \le 0$
	$(\frac{1}{2}, \frac{1}{2}, \frac{1}{2}, 0)$ if $-p_1 + p_1p_2 + p_1p_3 + p_2p_3 - 2p_1p_2p_3 \ge 0$ and $-p_2p_3 + p_1p_2p_3 + p_1p_4 - p_1p_2p_4 - p_1p_3p_4 + p_2p_3p_4 \le 0$
	$(\frac{2}{3}, \frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ if $-p_1 + p_1p_2 + p_1p_3 - p_1p_2p_3 + p_1p_4 - p_1p_2p_4 - p_1p_3p_4 + p_2p_3p_4 \ge 0$ and $-p_2p_3 + p_1p_2p_3 + p_1p_4 - p_1p_2p_4 - p_1p_3p_4 + p_2p_3p_4 \ge 0$
$2 \le T < \frac{5}{2}$	(1, 1, 0, 0) if $-p_1 - p_2 + 2p_1p_2 + p_1p_3 + p_2p_3 - 2p_1p_2p_3 + p_1p_4 + p_2p_4 - 2p_1p_2p_4 + p_3p_4 - 2p_1p_3p_4 - 2p_2p_3p_4 + 3p_1p_2p_3p_4 \le 0$
	$(\frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2})$ if $(p_2p_3 + p_2p_4 + p_3p_4 - 2p_2p_3p_4)(1 - p1) + p_1 \le p_1 + p_2 - p_1p_2$
$\frac{5}{2} \le T < 3$	(1, 1, 0, 0) if $p_2p_3 + p_2p_4 + p_3p_4 - 2p_2p_3p_4 + p_1 - p1p2 \ge p_1 + p_2 - p_1p_2$
	$(1, \frac{1}{2}, \frac{1}{2}, \frac{1}{2})$ if $-p_2 + p_2p_3 + p_2p_4 + p_3p_4 - 2p_2p_3p_4 \ge 0$
$3 \le T < 4$	(1, 1, 1, 0)

【図４】

アルゴリズム1（1レベル対称配分）

for each subset size choice $m = 1, \ldots, n$ **do**
  for each $M = \lceil \frac{m}{T} \rceil, \ldots, m$ **do**
    空でないノードm個のうちちょうどM個にアクセスする確率を計算する：
    $P_M = \sum_{\tau \in \mathcal{P}(\{(\frac{m}{M})\})} (\prod_{k \in \tau} p_k)(\prod_{l \notin \tau}(1 - p_l))$.
  end for
  サブセットサイズmについての成功確率を計算する：
  $P_{\mathrm{succ},m} = \sum_{M=\lceil \frac{m}{T} \rceil}^{m} P_M$.
end for
最大の$P_{\mathrm{succ},n}$を伴う最適サブセットサイズ選択肢mを取得する。

【図５】

アルゴリズム2（アルゴリズム1の簡略版）

for each subset size choice $m = \lfloor T \rfloor, \ldots, \lfloor K_{\max} T \rfloor$ **do**
  m個ノードについての平均アクセス確率を計算する：
  $p_{\mathrm{avg}} = \frac{\sum_{i=1}^{m} p_i}{m}$.
  for each $M = \lceil \frac{m}{T} \rceil, \ldots, m$ **do**
    Compute $P'_M = \binom{m}{M} \cdot p_{\mathrm{avg}}^{M} \cdot (1 - p_{\mathrm{avg}})^{(m-M)}$.
  end for
  サブセットサイズmについての成功確率を計算する：
  $P_{\mathrm{succ},m} = \sum_{M=\lceil \frac{m}{T} \rceil}^{m} P'_M$.
end for
最大の$P_{\mathrm{succ,mOpt}}$を伴う最適サブセットサイズ選択肢mOptを取得する。

【図６】

アルゴリズム3（2レベル対称配分）

for each subset size choice $m_1 = 1, \ldots, mOpt_{\mathrm{oneLevel}}$ **do**
  第1レベル内のm_1個のノードについての平均アクセス確率を計算する：$p_{\mathrm{avg},1} = \frac{\sum_{i=1}^{m_1} p_i}{m_1}$.
  for each subset size choice $m_2 = \lfloor K_{\min} T - 2m_1 \rfloor, \lfloor (K_{\min} + 1)T - 2m_1 \rfloor, \ldots, \lfloor K_{\max} T - 2m_1 \rfloor$ **do**
    第2レベル内のm_2個のノードについての平均アクセス確率を計算する：$p_{\mathrm{avg},2} = \frac{\sum_{i=m_1+1}^{m_1+m_2} p_i}{m_2}$.
    リカバリを成功させる全ペア$(m_{\mathrm{acc},1}, m_{\mathrm{acc},2})$についての集合$\mathcal{M}_{m_1,m_2}$を、数式（6）によって計算する。
    現在の選択肢である(m_1, m_2)に対応する成功確率を計算する：

$$P_{\mathrm{succ},m_1,m_2} = \sum_{(m_{\mathrm{acc},1}, m_{\mathrm{acc},2}) \in \mathcal{M}_{m_1,m_2}} P_{\mathrm{acc},1} \cdot P_{\mathrm{acc},2},$$

ここで

$$P_{\mathrm{acc},1} = \binom{m_1}{m_{\mathrm{acc},1}} \cdot p_{\mathrm{avg},1}^{m_{\mathrm{acc},1}} \cdot (1 - p_{\mathrm{avg},1})^{(m_1 - m_{\mathrm{acc},1})},$$

$$P_{\mathrm{acc},2} = \binom{m_2}{m_{\mathrm{acc},2}} \cdot p_{\mathrm{avg},2}^{m_{\mathrm{acc},2}} \cdot (1 - p_{\mathrm{avg},2})^{(m_2 - m_{\mathrm{acc},2})}.$$

    $P_{\mathrm{succ},m_1,m_2}$を現在の最適$P_{\mathrm{succ},mOpt_1,mOpt_2}$と比較する。
    if $P_{\mathrm{succ},m_1,m_2} > P_{\mathrm{succ},mOpt_1,mOpt_2}$ **then**
      Set $P_{\mathrm{succ},mOpt_1,mOpt_2} = P_{\mathrm{succ},m_1,m_2}$.
      Set $mOpt_1 = m_1$, and $mOpt_2 = m_2$.
    end if
    $D = \frac{2(n-m_1)}{3}$段のステップについて$mOpt_2$が同じ値を保持する場合、この$mOpt_2$の値を、現在の選択肢であるm_1に関してのm_2の最適な選択肢とみなす。
  end for
end for
最大の$P_{\mathrm{succ},mOpt_1,mOpt_2}$値を伴う$(mOpt_1, mOpt_2)$についての最適な選択肢を取得する。

【図７】

アルゴリズム4（3レベル対称配分）

1501: **for each subset size choice** $m_1 = 1, \ldots, mOpt_{1,twoLevel} + mOpt_{2,twoLevel}$ **do**

1502: 第1レベル内のm_1個のノードについての平均アクセス確率を計算する: $p_{avg,1} = \frac{\sum_{i=1}^{m_1} p_i}{m_1}$.

1503: **for each subset size choice** $m_2 = 1, \ldots, n - m_1 - 1$ **do**

1504: 第2レベル内のm_2個のノードについての平均アクセス確率を計算する: $p_{avg,2} = \frac{\sum_{i=m_1+1}^{m_1+m_2} p_i}{m_2}$.

1505: **for each subset size choice** $m_3 = \lfloor K_{\min}T - 3m_1 - 2m_2 \rfloor, \lfloor (K_{\min}+1)T - 3m_1 - 2m_2 \rfloor, \ldots, \lfloor K_{\max}T - 3m_1 - 2m_2 \rfloor$ **do**

1506: 第3レベル内のm_3個のノードについての平均アクセス確率を計算する: $p_{avg,3} = \frac{\sum_{i=m_1+m_2+1}^{m_1+m_2+m_3} p_i}{m_3}$.

1507: リカバリを成功させる全トリプル$(m_{acc,1}, m_{acc,2}, m_{acc,3})$ついての集合$\mathcal{M}_{m_1,m_2,m_3}$を、数式(7)によって計算する。

1508: 現在の選択肢である(m_1, m_2, m_3)に対応する成功確率を計算する:

$$P_{succ,m_1,m_2,m_3} = \sum_{(m_{acc,1},m_{acc,2},m_{acc,3}) \in \mathcal{M}_{m_1,m_2,m_3}} P_{acc,1} \cdot P_{acc,2} \cdot P_{acc,3},$$

ここで

$$P_{acc,L} = \binom{m_L}{m_{acc,L}} \cdot p_{avg,L}^{m_{acc,L}} \cdot (1 - p_{avg,L})^{(m_L - m_{acc,L})}$$

for $L \in \{1, 2, 3\}$.

1509: P_{succ,m_1,m_2,m_3}を現在の最適$P_{succ,mOpt_1,mOpt_2,mOpt_3}$と比較する。

1510: **if** $P_{succ,m_1,m_2,m_3} > P_{succ,mOpt_1,mOpt_2,mOpt_3}$ **then**

1511: Set $P_{succ,mOpt_1,mOpt_2,mOpt_3} = P_{succ,m_1,m_2,m_3}$.

1512: Set $mOpt_1 = m_1$, $mOpt_2 = m_2$, and $mOpt_3 = m_3$.

1513: **end if**

1514: $D = \frac{2(n-m_1-m_2)}{3}$段のステップについて$mOpt_3$が同じ値を保持する場合、この$mOpt_3$の値を、現在の選択肢である$(m_1, m_2)$に関しての$m_3$の最適な選択肢とみなして、$m_2$の次の値に移行する。

1515: **end for**

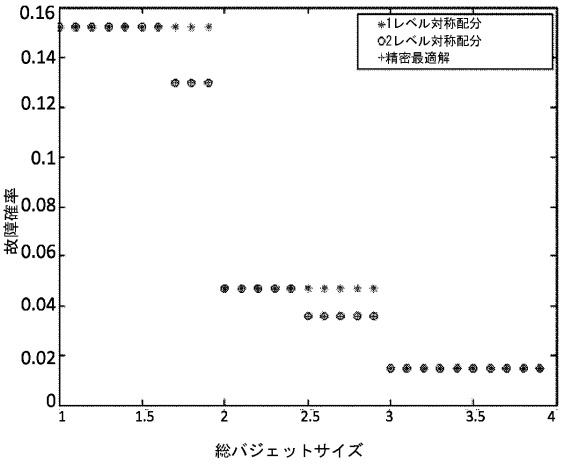
1516: ...

1517: **end for**

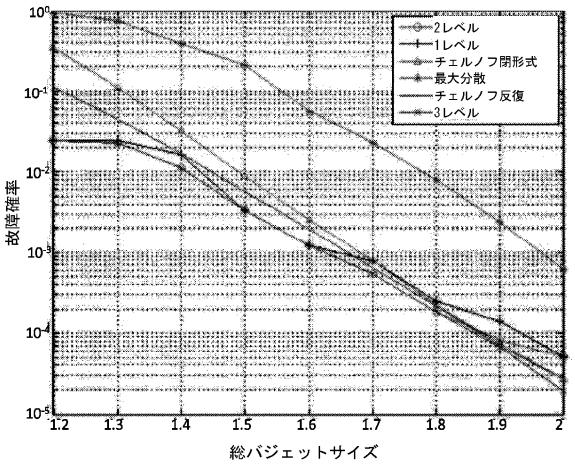
1518: **end for**

1519: 最大の$P_{succ,mOpt_1,mOpt_2,mOpt_3}$値を伴う$(mOpt_1, mOpt_2, mOpt_3)$についての最適な選択肢を取得する。

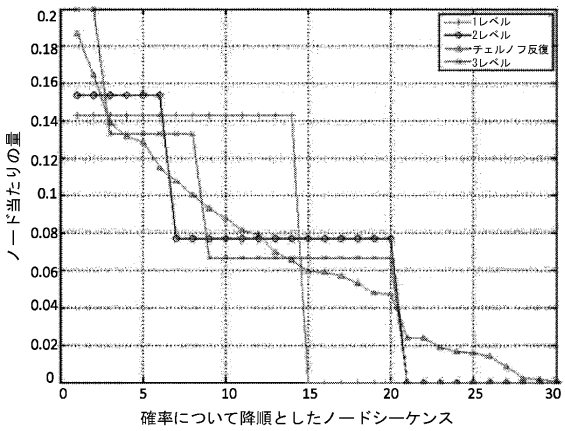
【図８】



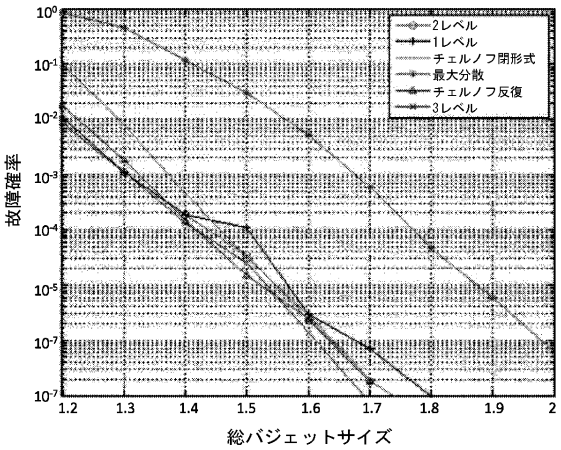
【図９】



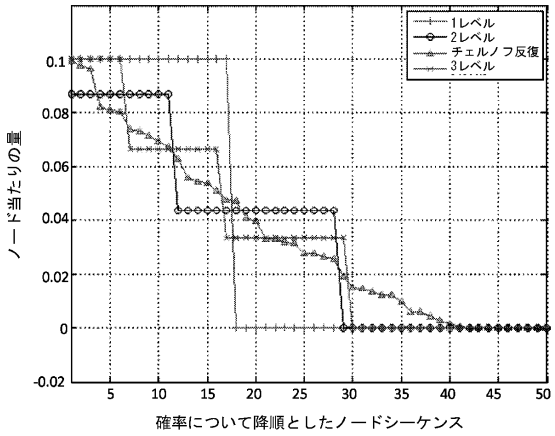
【図１０】



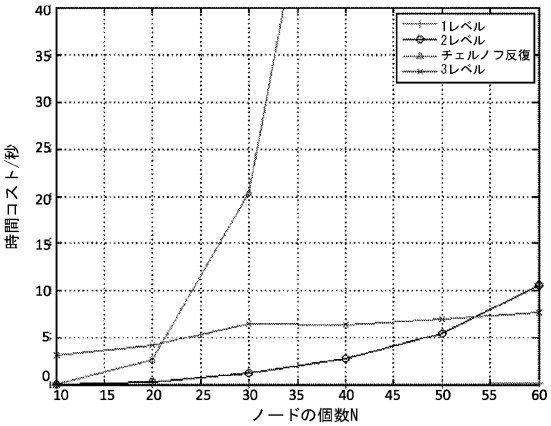
【図１１】



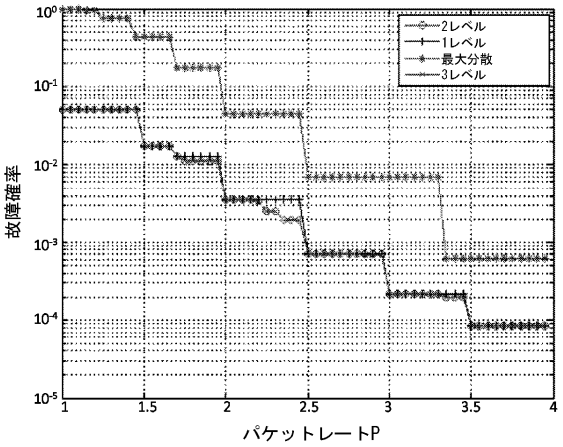
【図１２】



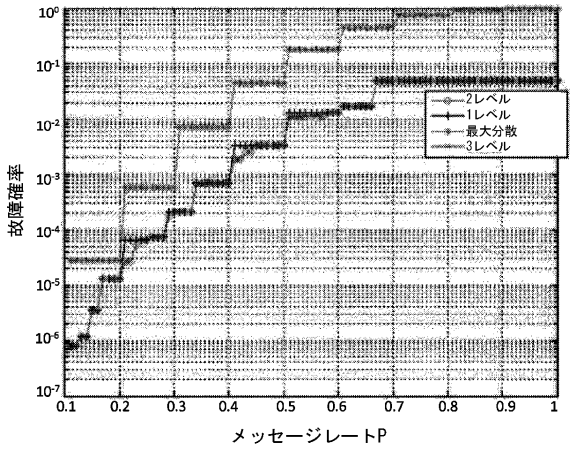
【図１３】



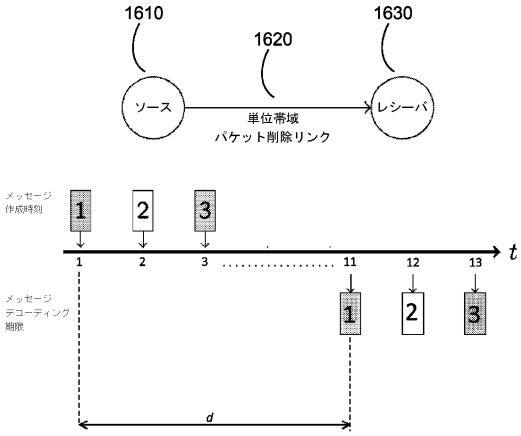
【図１４】



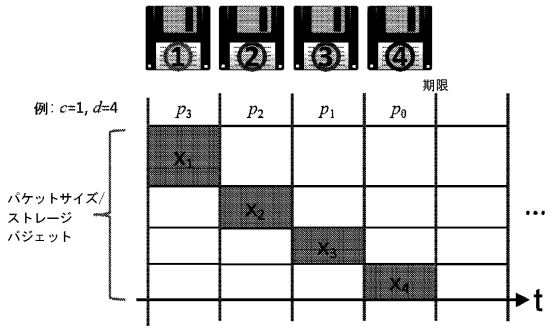
【図１５】



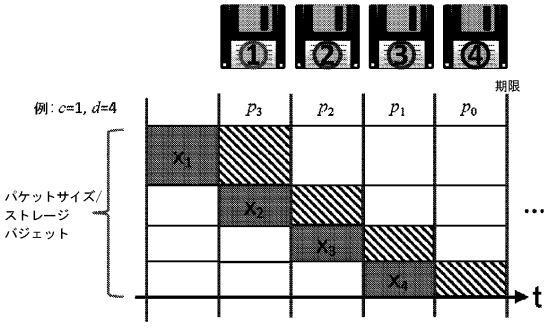
【図１６】



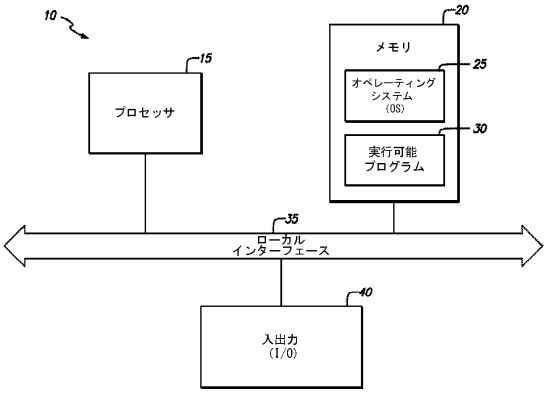
【図１７Ａ】



【図１７Ｂ】



【図１８】



フロントページの続き

(72)発明者　ジャオ　リー
アメリカ合衆国　カリフォルニア州　９４４０３　サン　マテオ　サウス　ノーフォーク　ストリート　２７２３　アパートメント　２０６
(72)発明者　トレイシー　シー　ホー
アメリカ合衆国　カリフォルニア州　９１１０５　パサデナ　ラグーナ　ロード　１１０７
(72)発明者　ディレク　レオン
アメリカ合衆国　カリフォルニア州　９１１０６　パサデナ　エス　チェスター　アヴェニュー　１４５　１／２
(72)発明者　ホンギ　ヤオ
アメリカ合衆国　カリフォルニア州　９４０６５　レッドウッド　シティ　キー　レーン　３５０

審査官　田名網　忠雄

(56)参考文献　米国特許出願公開第２０１３／００５４５３６（ＵＳ，Ａ１）
特開２００３－２９６１５０（ＪＰ，Ａ）

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ０６Ｆ　　３／０６－３／０８
Ｇ０６Ｆ　１３／１０－１３／１４
Ｇ０６Ｆ　１６／００－１６／９５８