

(19) 日本国特許庁 (JP)

(12) 特 許 公 報 (B2)

(11) 特許番号

特許第5041805号
(P5041805)

(45) 発行日 平成24年10月3日 (2012. 10. 3)

(24) 登録日 平成24年7月20日 (2012. 7. 20)

(51) Int. Cl.	F I
G 0 6 Q 50/10 (2012. 01)	G 0 6 F 17/60 1 2 4
G 0 6 F 3/06 (2006. 01)	G 0 6 F 17/60 1 3 2
	G 0 6 F 3/06 3 0 1 E
	G 0 6 F 3/06 3 0 2 Z

請求項の数 27 (全 17 頁)

(21) 出願番号	特願2006-509277 (P2006-509277)	(73) 特許権者	511076424
(86) (22) 出願日	平成16年3月25日 (2004. 3. 25)		ヒューレット・パカード デベロップメント カンパニー エル. ビー.
(65) 公表番号	特表2006-521640 (P2006-521640A)		Hewlett-Packard Development Company, L. P.
(43) 公表日	平成18年9月21日 (2006. 9. 21)		アメリカ合衆国 テキサス州 77070
(86) 国際出願番号	PCT/US2004/009119		ヒューストン コンパック センタ ド
(87) 国際公開番号	W02004/088547		ライブ ウェスト 11445
(87) 国際公開日	平成16年10月14日 (2004. 10. 14)	(74) 代理人	110000039
審査請求日	平成18年11月7日 (2006. 11. 7)		特許業務法人アイ・ピー・エス
審判番号	不服2010-9125 (P2010-9125/J1)	(72) 発明者	マーチャント・アリフ
審判請求日	平成22年4月28日 (2010. 4. 28)		アメリカ合衆国カリフォルニア州 ロスア
(31) 優先権主張番号	10/400, 556		ルトス トラバーソアベニュー 439
(32) 優先日	平成15年3月27日 (2003. 3. 27)		
(33) 優先権主張国	米国 (US)		

最終頁に続く

(54) 【発明の名称】 データストレージシステムのサービス品質コントローラ及びサービス品質方法

(57) 【特許請求の範囲】

【請求項 1】

データストレージのサービス品質制御方法であって、

QoSコントローラ(104)が、複数のワークロードのそれぞれについての複数の要求を、外部から受け入れて、受け入れた要求それぞれに優先順位を付けることと、

QoSコントローラ(104)が、前記複数の要求から、前記優先順位が高い順に選択して、前記データストレージのキュー(118)へ、前記選択した順に要求を転送することと、

QoSコントローラ(104)が、前記キューが保持する要求の数を、保持できる要求の目標数を示す目標キュー深度に維持することと、

ストレージデバイス(106)が、前記要求に応じた処理が完了した場合に、この要求を前記キューから削除することと、

QoSコントローラ(104)が、前記要求それぞれに応じた処理に要する待ち時間の目標値を示し、予め定められた期間に受け入れる要求の数を示す要求レートに応じて定められる各目標待ち時間と、前記要求それぞれに応じた処理に要する各待ち時間値とに基づいて、前記目標キュー深度を調整することと

を含み、

前記各待ち時間値は、前記QoSコントローラが前記要求を外部から受け入れた受入時間と、前記受け入れた要求に応じた前記データストレージにおける処理が完了した完了時間との差に基づいて、前記QoSコントローラにより計算され、

10

20

前記目標待ち時間それぞれは、この目標待ち時間が定められる時点の要求レートに対応する予め定められた読み出し要求の最大待ち時間と、この目標待ち時間が定められる時点の要求レートに対応する予め定められた書き込み要求の最大待ち時間と、全要求に対する読み込み要求の割合と、全要求に対する書き込み要求の割合とに基づいて、前記QoSコントローラにより、予め定められた間隔で定期的に定められる

方法。

【請求項2】

前記優先順位を付けることは、
要求に応じた処理が完了する時間の目標値を示す目標デッドラインを計算すること
を含む請求項1に記載の方法。

10

【請求項3】

前記QoSコントローラが、前記目標デッドラインを経過した要求を、前記キューに転送すること

をさらに含む請求項2に記載の方法。

【請求項4】

前記QoSコントローラが、前記要求の受入時間をモニタすること
をさらに含む請求項2に記載の方法。

【請求項5】

前記要求の前記目標デッドラインを前記計算することは、
前記目標待ち時間と、前記要求の前記受入時間とに基づいて計算すること
を含む
請求項4に記載の方法。

20

【請求項6】

前記ワークロードの要求レートをモニタすること
をさらに含む請求項5に記載の方法。

【請求項7】

前記キューは、
ストレージデバイスの中に含まれるデバイスキュー(118)である
請求項1に記載の方法。

【請求項8】

前記QoSコントローラが、前記デバイスキュー(118)への前記要求の転送に先立って、前記要求それぞれを、対応する入力キュー(110)に割り当てる
請求項7に記載の方法。

30

【請求項9】

前記計算された待ち時間値は、目標待ち時間を有するワークロード用であり、
前記目標キュー深度を前記調整することは、
前記ワークロードの前記目標待ち時間が前記計算された待ち時間値よりも短いときに、
前記目標キュー深度を削減すること
を含む
請求項1に記載の方法。

40

【請求項10】

前記QoSコントローラが、前記目標待ち時間と、前記計算された待ち時間との比較値に基づいて、前記目標キュー深度を、比例的に削減する
請求項9に記載の方法。

【請求項11】

前記計算された待ち時間値は、目標待ち時間を有するワークロード用であり、
ワークロードそれぞれは、
目標待ち時間及び計算された待ち時間値
を有し、
前記目標キュー深度値を前記調整することは、

50

目標待ち時間それぞれが計算された待ち時間値それぞれよりも長いときに、前記目標キュー深度を増加させること

を含む請求項 1 に記載の方法。

【請求項 1 2】

前記要求をキューに転送することの結果として、前記デバイスキューは、前記転送された要求の数を示す到達キュー深度を有し、

前記到達キュー深度が、前記目標キュー深度により制限されるときにのみ、前記 QoS コントローラが、前記目標キュー深度の値を増加させること

をさらに含む請求項 7 に記載の方法。

【請求項 1 3】

前記目標キュー深度の値を増加させることは、所定の乗数により前記目標キュー深度を増加させる

請求項 1 2 に記載の方法。

【請求項 1 4】

ストレージシステムのサービス品質制御装置であって、

複数のワークロードから要求を受け取る複数の入力キューであって、ワークロードそれぞれは、対応する前記入力キューに割り当てられ、入力キューそれぞれの要求は優先順位が付けられる複数の入力キューと、

ワークロードそれぞれの要求の性能値をモニタするモニタと、

前記入力キューから優先順位の高い順に前記要求を選択して、ストレージデバイスキューへ、前記選択した順に要求を転送するスケジューラであって、前記ストレージデバイスキューが保持する要求の数を、保持できる要求の目標数を示す目標キュー深度に維持するスケジューラと、

前記要求それぞれに応じた処理に要する待ち時間の目標値を示し、予め定められた期間に受け入れる前記要求の数を示す要求レートに応じて定められる各目標待ち時間と、ワークロードそれぞれの前記要求の前記性能値とに基づいて、前記ストレージデバイスキューの前記目標キュー深度を調整するコントローラと

を有し、

前記目標待ち時間それぞれは、この目標待ち時間が定められる時点の要求レートに対応する予め定められた読み出し要求の最大待ち時間と、この目標待ち時間が定められる時点の要求レートに対応する予め定められた書き込み要求の最大待ち時間と、全要求に対する読み込み要求の割合と、全要求に対する書き込み要求の割合とに基づいて、前記 QoS コントローラにより、予め定められた間隔で定期的に定められる

装置。

【請求項 1 5】

前記キューそれぞれの中の前記要求は、要求それぞれに割り当てられた、要求に応じた処理が完了する時間の目標値を示す目標デッドラインに従って優先順位が付けられる

請求項 1 4 に記載の装置。

【請求項 1 6】

要求それぞれの目標デッドラインは、前記要求の受入時間と、この要求の前記目標待ち時間とから決められる

請求項 1 5 に記載の装置。

【請求項 1 7】

前記モニタは、前記要求レートをモニタする

請求項 1 4 に記載の装置。

【請求項 1 8】

ワークロードそれぞれの要求への前記性能値は、

前記各要求に応じた処理に要する待ち時間値

を含む

請求項 1 4 に記載の装置。

10

20

30

40

50

【請求項 19】

前記コントローラは、前記要求それぞれに対する前記目標待ち時間が、いずれの要求に対する待ち時間値よりも短いときに、前記目標キュー深度を削減する

請求項 18 に記載の装置。

【請求項 20】

前記目標待ち時間と、前記要求に対する待ち時間値との比較値に基づいて、前記計算された待ち時間との比較値に基づいて、前記目標キュー深度が、比例的に削減される

請求項 19 に記載の装置。

【請求項 21】

前記コントローラは、前記要求それぞれに対する目標待ち時間それぞれが、対応する待ち時間値よりも長いときに、前記目標キュー深度を増加させる

請求項 18 に記載の装置。

【請求項 22】

前記ストレージデバイスキューは、
到達キュー深度
を有し、

前記コントローラは、前記目標キュー深度により前記到達キュー深度が制限されるときにのみ、前記目標キュー深度値を増加させる

請求項 18 に記載の装置。

【請求項 23】

前記目標キュー深度は、所定の乗数により増加される

請求項 22 に記載の装置。

【請求項 24】

前記コントローラは、前記データストレージデバイスの前記目標キュー深度を、下式に従って調整し、

【数 1】

$$E = \min_k \frac{\text{latencyTarget}(W_k)}{L(W_k)}$$

$$Q_{\text{new}} = \begin{cases} E * Q_{\text{old}} & \text{if } E < 1, \\ (1+\varepsilon) Q_{\text{old}} & \text{else if } Q_{\text{max}} \geq Q_{\text{old}}, \\ Q_{\text{old}} & \text{otherwise.} \end{cases}$$

ここで、latencyTarget(W_k)は、ワークロードW_kの要求に対する目標待ち時間であり、Q_{new}は、上記式に従って調整された新たな目標キュー深度であり、Q_{old}は、上記式に従って調整される前の元の目標キュー深度であり、Q_{max}は、先の所定の期間において到達したストレージデバイスキューの最大の深度であり、L(W_k)は、ワークロードW_kの要求に対する前記待ち時間値であり、 ε は、所定の乗数である

請求項 18 に記載の装置。

【請求項 25】

前記ワークロードそれぞれは、

1つ以上の対応するデータ記憶

を有し、

記憶領域に前記データ記憶を割り当てるキャパシティプランナー

をさらに有する請求項 22 に記載の装置。

【請求項 26】

機械により読み取り可能なプログラムストレージデバイスであって、

ストレージシステムの複数のワークロードそれぞれについての複数の要求を、外部から受け入れて、受け入れた要求それぞれに優先順位を付けるステップと、

前記複数の要求から、前記優先順位が高い順に選択して、前記ストレージシステムのキュー（118）へ、前記選択した順に前記要求をキューイングするステップであって、キューイングする要求の数を、保持できる要求の目標数を示す目標キュー深度に維持するステップと、

前記要求それぞれに応じた処理に要する待ち時間の目標値を示す各目標待ち時間と、前記要求それぞれに応じた処理に要する各待ち時間値に応じて、前記目標キュー深度を調整するステップと

10

を含む方法のステップを、前記機械により実行させるプログラムを記憶し、

前記待ち時間値は、前記要求を外部から受け入れた受入時間と、前記受け入れた要求に応じた処理が完了した完了時間との間の差に基づいて計算され、

前記目標待ち時間それぞれは、この目標待ち時間が定められる時点の要求レートに対応する予め定められた読み出し要求の最大待ち時間と、この目標待ち時間が定められる時点の要求レートに対応する予め定められた書き込み要求の最大待ち時間と、全要求に対する読み込み要求の割合と、全要求に対する書き込み要求の割合とに基づいて、前記 QoS コントローラにより、予め定められた間隔で定期的に定められる

プログラムストレージデバイス。

20

【請求項 27】

前記優先順位付けは、前記要求の受入時間と、前記目標待ち時間から決められ、要求それぞれに割り当てられる、要求に応じた処理が完了する時間の目標値を示す目標デッドラインに従って実行される

請求項 26 に記載のプログラムストレージデバイス。

【発明の詳細な説明】

【技術分野】

【0001】

本発明は、データストレージシステムに関する。

より詳細には、本発明は、サービス品質（QoS）についてのデータストレージシステムの制御に関する。

30

【0002】

この出願は、2003年3月27日に出願された「Data Storage System Emulation」という発明の名称の米国出願第10/400,232号に関連したものである。

この米国出願の内容は、参照により本明細書に援用される。

【背景技術】

【0003】

〔発明の背景〕

技術の進歩といった要因により、コンピュータハードウェアのコストの削減、及び、デジタルデータの量を増加させるワールドワイドウェブの成長が、世界規模で起こっている。

40

例えば、企業、行政、及び家庭におけるコンピュータシステムは、テキスト形式及び他の文書形式のデータ、データベース、マルチメディアファイル、電子メール通信、ウェブページ、トランザクションレコード等を生成するのに使用される。

その結果、データストレージ要求は、非常に大きなものとなり、次第に増大している。

【0004】

データストレージシステムは、この増加する要求により推進されて、より大きく、且つより複雑になっている。

例えば、現代のデータセンタは、数十個の大きなアレイならびに数千個の論理ボリューム及びファイルシステムを含む場合がある。

50

このようなデータセンタは、大きな組織、さらには複数の組織のストレージ要求に役立つことができる。

【 0 0 0 5 】

組織が、そのデータ記憶及び管理の要求をストレージサービスプロバイダ (S S P) に外部委託することがますます増加している。

S S P は、その顧客に利用可能とするストレージを自身のディスクアレイ上に割り当てる。

組織は、機器を購入し、適切な人材を雇うことによって内部で自身のストレージ要求を満たすことができると同時に、その別々の部門にストレージサービスを提供するための S S P モデルに内部で従うこともできる。

10

【 0 0 0 6 】

各顧客が受け取るサービスレベルを、例えば容量、可用性、及び性能レベルの点で指定できることが望ましい。

しかしながら、これは、各顧客、キャッシュ空間、ディスク、バス、ネットワーク帯域幅等のストレージ資源、及び、ストレージコントローラのプロセスサイクルを得るために競争する S S P に独自の負荷を提供することから、S S P との関連で困難になる可能性がある。

さらに、ストレージシステムは、通常、要求の発生元にかかわらず、すべての要求に対してベストエフォート型のサービスを提供するように設計される。

【 0 0 0 7 】

20

各仕事負荷が十分なサービスを受け取ることを実にすることを試みる 1 つの手法は、過剰な容量及び帯域幅を有するストレージシステムを過剰供給することである。

しかしながら、この技法は、追加された資源がシステムのコストを増加させるので、一般にコスト効率が良くない。

別の手法は、各仕事負荷に別々の物理資源 (例えば、1 つ又は複数の専用ディスクアレイ) を割り当てることである。

この技法は、追加される容量又は帯域幅を一般に任意の増加量で追加できない代わりに、仕事負荷の要件が変化すると、広範囲の再構成が必要とされる場合がある点で、柔軟性に欠ける傾向がある。

【 特 許 文 献 1 】 米 国 特 許 第 6 1 7 0 0 4 2 号

30

【 特 許 文 献 2 】 米 国 特 許 第 6 4 3 4 6 3 1 号

【 特 許 文 献 3 】 米 国 特 許 第 5 6 8 0 5 3 9 号

【 特 許 文 献 4 】 米 国 特 許 第 6 1 5 7 9 6 3 号

【 発 明 の 開 示 】

【 発 明 が 解 決 し よ う と す る 課 題 】

【 0 0 0 8 】

したがって、特定の仕事負荷が、複数の仕事負荷をハンドリングするストレージシステムから受け取るサービスレベルを指定できる改善された機能が必要とされている。

本発明が対象とするのはこの目的である。

【 課 題 を 解 決 す る た め の 手 段 】

40

【 0 0 0 9 】

【 発 明 の 概 要 】

本発明は、データストレージシステムのサービス品質コントローラ及び関連した方法である。

1 態様では、本発明は、データストレージのサービス品質方法及びサービス品質装置を提供する。

複数のストレージシステムの仕事負荷のそれぞれについての要求に優先順位が付けられる。

これらの要求は、ストレージデバイスキューを目標キュー深度に維持するために、その優先順位に従ってストレージデバイスキューへ選択的に転送される。

50

この目標キュー深度は、要求の待ち時間値に応じて調整される。この待ち時間値は、各仕事負荷について要求の到着時刻と完了時刻との差に基づいて計算される。

【 0 0 1 0 】

要求に優先順位を付けることは、要求の監視された到着時刻と仕事負荷の目標待ち時間とに基づいて要求の目標デッドラインを計算することにより行うことができる。

仕事負荷の目標待ち時間は、その仕事負荷の要求レートに基づくことができる。

待ち時間を削減するために、仕事負荷の目標待ち時間がその計算された待ち時間値よりも小さい場合に、目標キュー深度を削減することができる。

スループットを増加させるために、各仕事負荷の目標待ち時間が、計算された各待ち時間値よりも大きい場合に、目標キュー深度を増加させることができる。

10

要求をデバイスキューに上記転送した結果として、デバイスキューは到達キュー深度を有する。

したがって、目標キュー深度値を増加させることは、到達キュー深度が目標キュー深度によって制限される場合に実行することができる。

【 0 0 1 1 】

別の態様では、各仕事負荷が複数の要求を含んでいた場合に、関数が、複数の仕事負荷のそれぞれについて許容可能な待ち時間の統計値を指定する。

要求は、スケジューリングされて、データストレージデバイスへ転送され、監視された待ち時間統計値は、各仕事負荷の許容可能な待ち時間統計値の範囲内に概ね維持される。

各要求の目標デッドラインに従って要求に優先順位を付けることによって、それら要求をスケジューリングすることができる。

20

データストレージデバイスにおいて目標キュー深度を維持するために、要求を転送することができる。

仕事負荷の許容可能な待ち時間統計値が、監視された待ち時間統計値よりも小さい場合に、目標キュー深度を削減することができる。

さらに、許容可能な各待ち時間統計値が、監視された各待ち時間統計値よりも大きい場合に、目標キュー深度を増加させることができる。

【 0 0 1 2 】

本発明のこれらの態様及び他の態様は、本明細書でより詳細に説明される。

【発明を実施するための最良の形態】

30

【 0 0 1 3 】

[好ましい実施の形態の詳細な説明]

本発明は、データストレージシステムのサービス品質コントローラ及び関連した方法を提供する。

本発明は、ホストからの入出力（Ｉ／Ｏ）要求を途中で捕捉して、それらＩ／Ｏ要求をストレージデバイスへ転送すると同時に、それらＩ／Ｏ要求が転送されるレートを選択的に調節することによって、１つ又は複数のホストと１つ又は複数のストレージデバイスとの間の仮想化レベルを追加する。

各要求は、ストレージデバイスへの配信についてスケジューリングされると同時に、その完了が監視される。

40

完了した要求から得られた統計的な情報に基づいて、その後、各仕事負荷の所望の性能を維持しようとして、スケジューリングを変更することができる。

本発明は、一定の性能分離（performance isolation）を提供して、或る仕事負荷が経験する性能が他の仕事負荷の変動による影響をあまり受けず、性能目標が満たされることをより確実にする。

ストレージ資源間で複数の仕事負荷をバランスさせることによって、資源が効率的に利用される。

【 0 0 1 4 】

図１は、本発明の１態様によるデータストレージ管理システム１００を示している。

このシステム１００は、容量計画器（capacity planner）１０２、サービス品質コント

50

ローラ 104、及び 1 つ又は複数のハードディスクアレイ等のストレージデバイス（複数可）106を含む。

【0015】

容量計画器 102 は、1 つ又は複数の仕事負荷の仕様を入力として受け取る。

例えば、システムアドミニストレータがこの入力を提供することができる。

仕事負荷の仕様は、例えば、各仕事負荷によって必要とされる複数のストア（すなわち、仮想化されたストレージ空間）、及び、各仕事負荷によって必要とされる容量（すなわち、ストレージ空間の量）を含むことができる。

これに加えて、要求生成速度に基づく許容可能な I/O 要求待ち時間等の仕事負荷の性能目標もシステム 100 に入力することができる。

10

容量計画器 102 は、デバイス 106 にストアを割り当てることによって、各仕事負荷のストレージをストレージデバイス 106 に割り当て、また、デバイス 106 が仕事負荷の集約した要求を満たすのに十分な容量及び帯域幅を確実に有するようにする。

この割り当ては、仕事負荷の変化する要件及びデバイス 106 の構成の変更を満たすように定期的に変更することができる。

容量計画器 102 は、容量計画ステップを実行するようにプログラミングされた汎用コンピュータシステムとして実施することができる。

容量計画器によって示されたデバイス 106 の物理構成は、例えば、システムアドミニストレータが実行することができる。

【0016】

20

1 態様では、容量計画器 102 は、2001 年 10 月 23 日に出願された「Automated System Adaptation Technique Particularly for Data Storage Systems」という発明の名称の米国特許出願第 10/046,463 号に従って動作する。

この米国特許出願の内容は、参照により本明細書に援用される。

この計画器 102 では、仕事負荷の下でのストレージデバイス 106 のオペレーションを解析すること、その解析に基づいて新たな設計を生成すること、及び、その新たな設計に既存のシステムを移行させることを含む一連のステップが、繰り返しループで実行される。

これらのタスクをプログラムとして繰り返すことによって、容量計画は、過剰供給されることなく、仕事負荷をサポートする容量計画に収束する。

30

【0017】

別の態様では、容量計画器 102 は、2001 年 8 月 7 日に出願された「Simultaneous Array Configuration and Store Assignment for a Data Storage System」という発明の名称の米国特許出願第 09/924,735 号に従って動作する。

この米国特許出願の内容は、参照により本明細書に援用される。

この計画器 102 では、複数のノードを有するデータ構造体が、コンピュータ可読メモリに記憶される。

それらのノードの少なくともいくつかは、各属性を有する物理的なデータストレージデバイス 106 に対応する。

仕事負荷によって使用される複数のデータストアは、それぞれ、入力として提供される容量要件や帯域幅要件等のデータストレージ要件を有する。

40

これらのデータストアは、再帰的に階層に割り当てられ、各ノードにおいて、ストアの要件が属性のいずれも上回らないことがチェックされる。

ストアが割り当てられている間、ストアによりよく適合するように階層を変更することができる。

どの属性にも違反しない可能な割り当てを、システムの目標に従って互いに比較することができる。

【0018】

図 1 のシステム 100 に使用される上述した容量計画器 102 は例示である。

したがって、別の計画器を選択できることは明らかである。

50

さらに、この容量計画は、従来の手動による技法に従って実行することもできる。

【 0 0 1 9 】

ホストシステム 1 0 8 は、例えば、1 つ又は複数のアプリケーションが実行されて、それらのアプリケーションがストレージデバイス 1 0 6 に対する I / O 要求（すなわち、読み出しオペレーション及び書き込みオペレーション）を生成するコンピュータシステムを含むことができる。

各仕事負荷は、1 つ又は複数のホスト 1 0 8、及び / 又は、1 つ又は複数のホスト 1 0 8 で実行されているアプリケーションによって生成された一続きの I / O 要求を含む。

例えば、特定の仕事負荷は、ホスト 1 0 8 のうちの単一のホストによって、又は、ホスト 1 0 8 のうちのいくつかで実行されている特定のアプリケーションによって生成された要求のすべてから成ることがある。

別の例として、ホスト又はアプリケーションによって生成された要求は、読み出し要求及び書き込み要求に分割することができる。

この場合、読み出し要求が或る仕事負荷を表し、書き込み要求が別の仕事負荷を表す。

【 0 0 2 0 】

サービス品質コントローラ 1 0 4 は、1 つ又は複数のホストシステム 1 0 8 によって生成された I / O 要求を受け取る。

これに応答して、コントローラ 1 0 4 は、ストレージデバイス 1 0 6 と通信して、それらの要求を完了する。

これに加えて、コントローラ 1 0 4 は、容量計画器 1 0 2 からストレージ割り当て情報を受け取ることもできる。

コントローラ 1 0 4 は、このストレージ割り当て情報を使用して、各要求が属する特定の仕事負荷を特定することができる。

コントローラ 1 0 4 が仕事負荷の性能目標を満たすのに使用可能でない場合、コントローラ 1 0 4 は、過負荷アラームを通信すること等によって、容量計画器 1 0 2 に信号で伝えることができる。

これが行われると、容量計画器 1 0 2 は、デバイス 1 0 6 にストアを再割り当てすること、デバイス 1 0 6 を、より多くの容量及び / 又は帯域幅を有する他のデバイスと交換すること、又はデバイス 1 0 6 の個数又はデバイス 1 0 6 のコンポーネントの個数を増加させてデバイス 1 0 6 の全体の容量及び / 又は帯域幅を増加させること等によって、適切な動作を取ることができる。

【 0 0 2 1 】

各仕事負荷の性能目標は、1 対の曲線によって指定することができる。

この 1 対の曲線の各曲線は、時間ウィンドウ w の期間で平均化されて提供された要求レートの関数としての読み出し待ち時間及び書き込み待ち時間を指定する。

図 2 A 及び図 2 B は、仕事負荷の例示の性能要件を示している。

より具体的には、図 2 A は、横軸に、毎秒の I / O の個数による読み出し要求の発行レートを有し、縦軸に、ミリ秒による上限の最大待ち時間（期間 w で平均されたもの）を有するグラフを示している。

待ち時間は、所与の要求を処理するのに必要とされる時間である。

この待ち時間は、コントローラ 1 0 4 がホスト 1 0 8 から要求を受信した時と、ストレージデバイス 1 0 6 がその要求を完了した時との間に経過した時間として計測することができる。

図 2 A に示すように、読み出し要求レートが毎秒 1 0 個の I / O 未満である場合、1 0 m s の最大待ち時間が認められる。

読み出し要求レートが毎秒 1 0 個の I / O と 2 0 個の I / O との間である場合、1 5 m s の最大待ち時間が認められ、読み出し要求レートが毎秒 2 0 個の I / O と 3 0 個の I / O との間である場合、2 0 m s の最大待ち時間が認められる。

これに加えて、読み出し要求レートが毎秒 3 0 個の I / O と 4 0 個の I / O との間である場合、2 5 m s の最大待ち時間が認められる。

必要ではないが、I/O要求レートが毎秒40個のI/Oを超える場合、待ち時間に対して制限は与えられない。

【0022】

図2Bは、書き込み要求の待ち時間の制限のグラフを示している。

図2Bに示すように、書き込み要求レートが毎秒10個のI/O未満である場合、20msの最大待ち時間が認められる。

書き込み要求レートが毎秒10個のI/Oと20個のI/Oとの間である場合、30msの最大遅延時間が認められ、書き込み要求レートが毎秒20個のI/Oと30個のI/Oとの間である場合、40msの最大待ち時間が認められる。

これに加えて、書き込み要求レートが毎秒30個のI/Oと40個のI/Oの間である場合、50msの最大待ち時間が認められる。

10

図2Aと同様に、I/O要求レートが毎秒40個のI/Oを超える場合、待ち時間に対して制限は与えられない。

読み出し要求についての図2Aに示す性能要件と書き込み要求についての図2Bに示す性能要件とは異なるが、同じ要件が双方に適用される場合があることは明らかであろう。

さらに、待ち時間の制限をより少なく指定することもできるし、より多く指定することもできる。

例えば、図3は、毎秒40個のI/O未満のすべての要求レートに対して、20msの単一の待ち時間制限が適用されるグラフを示している。

【0023】

20

一般に、これらの2つの曲線は、3要素のベクトル(vector of triples) $((r_1, tr_1, tw_1), (r_2, tr_2, tw_2), \dots, (r_n, tr_n, tw_n))$ として表すことができる。

ここで、 $0 < r_1 < \dots < r_n$ であり、 r は要求レートであり、 tr は読み出し要求の最大待ち時間であり、 tw は書き込み要求の最大待ち時間である。

r_n を超える要求レートについての待ち時間には制限が与えられないので、これは、 $tr_{n+1} = tw_{n+1} = \infty$ として表すことができる。

したがって、図2Aの曲線及び図2Bの曲線は、 $((10 \text{ IO/s}, 10 \text{ ms}, 20 \text{ ms}), (20 \text{ IO/s}, 15 \text{ ms}, 30 \text{ ms}), (30 \text{ IO/s}, 20 \text{ ms}, 40 \text{ ms}), (40 \text{ IO/s}, 25 \text{ ms}, 50 \text{ ms}))$ として表すことができる。

30

【0024】

時間は、 w のウィンドウ(エポック(epoch))に分割することができる。

各ウィンドウ w は、約1秒程度の長さとしてすることができるが、別のウィンドウの長さを選択することもできる。

したがって、最大待ち時間を超えるかどうかを判断するために、計測された待ち時間は、 w の期間にわたって平均化される。

したがって、全I/O要求のうち読み出し要求の割合が f_r である仕事負荷が性能要件を満たすには、どの時間ウィンドウ w_i の平均待ち時間も、前のウィンドウ w にわたる平均要求レートが r_i よりも小さい場合に、 $f_r * tr_i + (1 - f_r) tw_i$ を超えるべきではない。

40

この公式は、仕事負荷が読み出しのみである場合には待ち時間の限度が tr_i であることを意味し、仕事負荷が書き込みのみである場合には待ち時間の限度が tw_i であることを意味し、仕事負荷が読み出し/書き込みの混合したものである場合には待ち時間の限度が2つの限度を線形補間したものであることを意味する。

しかしながら、計測された待ち時間をこのように組み合わせる必要はない。

特に、仕事負荷が、観測された読み出し待ち時間又は観測された書き込み待ち時間によって支配される場合には、例えば、要求の調節は、その一方又は他方のみに基づくことができる。

また、待ち時間の目標が読み出し要求及び書き込み要求について同じである場合、それぞれについて観測された待ち時間を区別する必要もない。

50

【 0 0 2 5 】

一般に、仕事負荷の性能目標は、その仕事負荷の計測可能な特性の任意の組に基づいた、必要とされる待ち時間を特定する任意の計算可能な関数とすることができる。

例えば、待ち時間は、I / Oサイズの関数又はI / Oサイズの組み合わせとして、且つ、要求が読み出しであるのか、それとも書き込みであるのかで指定することができる。

この関数は、数学関数とすることもできるし、この場合のように、テーブルによって指定することもできる。

読み出し及び書き込みの性能目標は、ここでは、単一の目的関数に組み合わされるが、それら性能目標は別々にしておくこともできる。

例えば、仕事負荷からの読み出し要求及び書き込み要求は、別々の入力キューに入れることができ、別々の待ち時間の限度を各キューに関連付けることができる。

同様に、他の基準に基づいて、要求を別々の入力キューに分離することもできる。

例えば、性能目標が、小さなI / O（例えば、64 K B未満）及び大きなI / O（例えば、64 K B以上）について別々の待ち時間関数を指定する場合、大きなI / O要求及び小さなI / O要求を別々の入力キューに分離することができ、対応する性能要件をそれらI / O要求に別々に適用することができ、それによって、それらI / O要求を別々の仕事負荷として扱うことができる。

【 0 0 2 6 】

図4は、図1のサービス品質ストレージコントローラ104の1実施の形態をより詳細に示している。

コントローラ104は、入力キュー110、I / Oスケジューラ112、I / Oモニタ114、及びI / Oコントローラ116を含む。

キュー110は、バッファとして構成されたメモリデバイスによって実施することができる一方、スケジューラ112、モニタ114、及びコントローラ116は、汎用若しくは専用のハードウェア及び／又はソフトウェアによって実施することができる。

【 0 0 2 7 】

また、図4は、従来通りストレージシステムの一部であり、したがって、デバイス106に含まれるデバイスキュー118も示している。

QoSコントローラ104がない場合、要求は、ホスト108からデバイスキュー118へ直接非知的に配信される。

デバイスキュー118からの要求の削除及びデバイスキュー118内での要求のソートは、デバイス106の制御下で行うことができ、したがって、デバイスキュー118は、QoSコントローラ104から独立して動作する。

例えば、一般のストレージデバイス（デバイス106で利用できる）は、デバイスキュー118の要求の中でそれら要求をソートするための、デバイスの製造者に独自の方式を実施することができる。

したがって、本発明は、商用のストレージデバイスと互換性を有する。

一方、QoSコントローラ104は、デバイスキュー118への要求の流れを制御する。

【 0 0 2 8 】

（例えば、ホスト108から）QoSコントローラ104に到着した要求は、入力キュー110に入れられる。

入力キュー110では、各仕事負荷からの要求は、好ましくは、キュー100のうちの特定の1つに割り当てられる。

I / Oモニタ114及びI / Oコントローラ116からの繰り返される入力に基づいて、スケジューラ112は、デバイスキュー118の目標のキュー深度値、及び、スケジューラ112が維持しようと試みる仕事負荷ごとの待ち時間目標値を保持する。

【 0 0 2 9 】

スケジューラ112は、各仕事負荷についての要求に、その同じ仕事負荷における他の要求に対して優先順位を付ける。

10

20

30

40

50

これは、アーリストデッドラインファースト (EDF (earliest deadline first)) スケジューリングを使用して行うことができる。

このEDFスケジューリングでは、仕事負荷 W_k からの要求のデッドラインは、その要求の到着時刻 $arrivalTime(W_k)$ に、その仕事負荷の待ち時間目標値 $latencyTarget(W_k)$ を加えたものとして計算される。

したがって、仕事負荷 W_k からの要求のデッドラインは、 $arrivalTime(W_k) + latencyTarget(W_k)$ として計算することができる。

この仕事負荷のデッドラインは、その仕事負荷における最も古い保留中の要求と同じであると考えられる。

【0030】

10

スケジューラ112は、デバイスキュー118を定期的にポーリングして、その時点のキュー深度（すなわち、キュー118において保留中の要求の個数）を求める。

代替的に、例えば、デバイス106がデバイスキュー118の深度の報告をサポートしていない場合に、スケジューラ112は、モニタ114からのI/O要求の到着及び完了の情報に基づいて現在のデバイスキュー深度を計算することもできる。

好ましくは、キュー深度は、定期的に（例えば、1msごとに）ポーリングされ、また、I/O要求の完了時にもポーリングされる。

仕事負荷からの要求は、指定された状況下でデバイスキュー118へ転送される。

すなわち、デバイスキュー118の現在の深度がI/Oコントローラ116によって指定された目標値よりも小さい場合に、要求をデバイスキュー118へ転送することができる。

20

この場合、スケジューラ112は、最も早いデッドラインを有する仕事負荷を選択して、その対応するキュー110の最初の要求（すなわち、最も早いデッドラインを有する要求）をデバイスキュー118へ転送する。

スケジューラ112は、デバイスキュー118の現在の深度が目標深度に達するまで、次の最も早いデッドラインを有する仕事負荷を選択するステップ、及び、その仕事負荷の最初の要求をデバイスキュー118へ転送するステップを繰り返す。

このように、デバイスキュー118をその目標深度に維持するために、要求は、入力キュー110において優先順位付けされて、要求のうち最も高い優先順位の要求がデバイスキュー118へ転送される。

30

また、要求のデッドラインがすでに経過している時も、その要求をデバイスキュー118へ転送することができる。

期限が経過したすべての要求は、たとえ、この要求をデバイスキュー118へ転送することによって、デバイスキュー118の深度がその目標深度を超えさせることになっても、デバイスキュー118へ転送されることが好ましい。

これによって、新たに到着した、小さな待ち時間要件を有する仕事負荷の要求に直に対応することができる。

【0031】

I/Oモニタ114は、I/O要求の到着及び完了を監視する。

より具体的には、モニタ114は、各仕事負荷についての新たな読み出し要求及び書き込み要求の、QoSコントローラ104への到着レートを監視することができ、また、この情報をI/Oコントローラ116へ報告することもできる。

40

性能目標が要求レートに基づいて最大待ち時間を指定できることを思い出されたい。

したがって、各仕事負荷の目標待ち時間は、その時点の要求レートに基づいて定期的に求めることができる。

より具体的には、性能目標が $((r_1, tr_1, tw_1), (r_2, tr_2, tw_2), \dots, (r_n, tr_n, tw_n))$ として与えられる場合、読み出し要求及び書き込み要求を含む仕事負荷 W_k の待ち時間目標値は、次のように計算することができる。

$latencyTarget(W_k) = fr * tr_i + (1 - fr) tw_i, \text{ if } r_{i-1} > readRate(W_k) + writeRate(W_k) < r_i$

50

ここで、 $readRate(W_k)$ は仕事負荷 W_k の読み出し要求レートであり、 $writeRate(W_k)$ は仕事負荷 W_k の書き込み要求レートであり、 $r_0 = 0$ 、 $r_{n+1} = \infty$ 、 $tr_{n+1} = tw_{n+1} = \infty$ である。

目標値は、好ましくは、 $P = 0.05$ 秒ごとに定期的に求められて報告される。

ただし、別の期間 P を選択できることは明らかである。

【0032】

I/O モニタ 114 は、要求到着レートを監視することに加えて、好ましくは、要求が QoS コントローラ 104 内に受信されると、各 I/O 要求の到着時刻も監視する。

これに加えて、モニタ 114 は、デバイス 106 によって報告された各要求の完了時刻も監視する。

10

モニタ 114 は、これから、各要求の待ち時間及び各仕事負荷の (w の期間にわたる) 平均待ち時間を計算することができる。

この平均待ち時間は、次に、コントローラ 116 に報告される。

待ち時間の平均値は、好ましくは、 $P = 0.05$ 秒ごとに定期的に計算されて報告される。

ただし、別の期間を選択できることは明らかである。

【0033】

コントローラ 116 は、平均待ち時間情報に応じて、デバイスキュー 118 の目標深度を調整することができる。

この制御は、デバイスキュー 118 の深度を削減すると、デバイス 106 における待ち時間が削減されることにつながるが、スループットも削減されることにつながるという仮定に基づいている。

20

逆に言えば、デバイスキュー 118 の深度を増加させると、待ち時間が増加することにつながり、スループットも増加することにつながる。

この仮定は、ほとんどのディスク及びディスクアレイに当てはまると予想される。

【0034】

コントローラ 116 は、デバイスキュー 118 をフル (すなわち、その目標深度に) 維持しようと試みる。

これは、デバイスキュー 118 に多くの要求を有すると、デバイス 106 の利用が改善され、したがって、スループットが最大になるからである。

30

しかしながら、デバイスキュー 118 にあまりにも多くの要求を有すると、待ち時間が許容できないものとなるおそれがある。

例えば、仕事負荷が小さな待ち時間を要求する場合、デバイスキュー 118 に多くの要求を有することは、仕事負荷からの次の要求が、キュー 118 の他の多くの未処理の要求と競合することを意味し、したがって、その要求は、キュー 118 が有する未処理の要求が少ない場合よりも遅く (大きな待ち時間で) 完了することを意味する。

このように、デバイスキュー 118 をその目標深度に維持することは、待ち時間が過大になることなく、ストレージデバイス 106 の利用を最大にすることにつながる。

目標キュー深度は、このバランスを維持するように調整される。

【0035】

40

特定のインプリメンテーションでは、コントローラ 116 は、各仕事負荷の現在の目標待ち時間を、前の期間中に計測されたその平均待ち時間と比較して、デバイスキュー 118 の目標深度を調整する。

いずれかの仕事負荷が、その前の計測された待ち時間よりも小さな新しい目標待ち時間を有すると、そのデバイスキュー 118 の目標深度は削減される。

目標深度の削減量は、好ましくは、各仕事負荷についての目標待ち時間と計測された待ち時間との差の中で最大の差に比例する。

例えば、2つの仕事負荷 W_1 及び W_2 を有するシステムにおいて、仕事負荷 W_1 の新たな目標待ち時間が、その計測された値よりも 10% 小さく、且つ、仕事負荷 W_2 の新たな目標待ち時間が、その計測された目標値よりも 5% 小さい場合、デバイスキュー 118 の

50

目標深度は、好ましくは、デバイスキュー 118 の前の目標深度から 10 % だけ削減される。

しかしながら、デバイスキュー 118 の目標値の削減は、別の方法で求めることもできることは明らかである。

例えば、削減量は、仕事負荷の 2 つ又は 3 つ以上についての目標値と計測された値との差から統計的に（例えば、平均化することによって）求めることができる。

逆に言えば、新たな目標待ち時間のすべてが、仕事負荷のすべてについて計測された待ち時間よりも長い場合、デバイスキュー 118 の目標深度は、より大きなスループットを可能にさせるために増加させることができる。

例えば、目標深度は、所定の乗数によって増加させることができる。

10

例えば、乗数が 0.1 である場合、目標深度は、その前の値の 110 % に増加される。

ただし、別の乗数を選択することもできる。

デバイスキュー 118 の目標深度を増加させると、実際の到達キュー深度がその目標深度によって制限された場合にのみ、スループットが増加すると予想される。

したがって、目標深度を増加させるかどうかは、到達キュー深度が目標深度によって制限されたかどうかに基づいて決定することができる。

これらの条件のいずれも当てはまらない場合には、デバイスキュー 118 の目標深度は変更しないでおくことができる。

【0036】

このように、コントローラ 116 は、非線形フィードバックを実施して、デバイスの目標深度を調整することができる。

20

形式的に、上述したフィードバック方式は、以下によって表すことができる。

【0037】

【数 1】

$$E = \min_k \frac{\text{latencyTarget}(W_k)}{L(W_k)}$$

30

$$Q_{new} = \begin{cases} E * Q_{old} & \text{if } E < 1, \\ (1+\epsilon) Q_{old} & \text{else if } Q_{max} \geq Q_{old}, \\ Q_{old} & \text{otherwise.} \end{cases}$$

【0038】

ここで、 Q_{new} は、デバイスキュー 118 の新たな目標深度であり、 Q_{old} は、デバイスキュー 118 の前の目標深度であり、 Q_{max} は、前の期間に到達したデバイスキュー 118 の最大深度であり、 $L(W_k)$ は、仕事負荷 W_k の計測された平均待ち時間であり、 ϵ は、所定の小さな正の値（例えば、0.1）である。

40

最初に、目標キュー深度は、初期値（例えば、200 エントリー）に設定される。

【0039】

上記のように、データストレージのサービス品質システム及びサービス品質方法を説明してきた。

本発明は、一定の性能分離を提供して、或る仕事負荷が経験する性能が他の仕事負荷の変動による影響をあまり受けず、性能目標が満たされることをより確実にする。

ストレージ資源間で複数の仕事負荷をバランスさせることによって、資源が効率的に利用される。

50

【 0 0 4 0 】

上記は、本発明の特定の実施の形態に関するものであるが、本発明の原理及び精神から逸脱することなく、これらの実施の形態に変更を行えることが当業者には理解されよう。

本発明の範囲は、添付の特許請求の範囲によって画定される。

【図面の簡単な説明】

【 0 0 4 1 】

【図 1】本発明の 1 態様によるデータストレージ管理システムを示す図である。

【図 2 A】仕事負荷の例示の性能要件を示す図である。

【図 2 B】仕事負荷の例示の性能要件を示す図である。

【図 3】単一の待ち時間制限がしきい値未満のすべての要求レートに適用される代替的な例示の性能要件を示す図である。 10

【図 4】本発明の 1 態様によるサービス品質ストレージコントローラを示す図である。

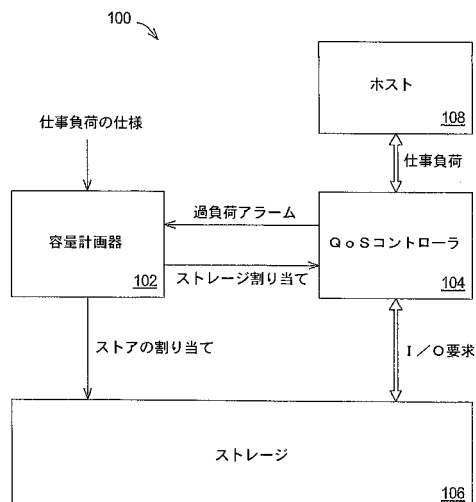
【符号の説明】

【 0 0 4 2 】

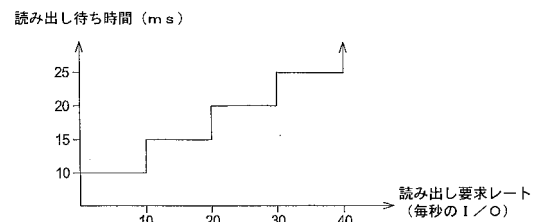
1 0 2 . . . 容量計画器 ,
1 0 4 . . . Q o S コントローラ ,
1 0 6 . . . ストレージ ,
1 0 8 . . . ホスト ,
1 1 0 . . . 入力キュー ,
1 1 2 . . . I / O スケジューラ ,
1 1 4 . . . I / O モニタ ,
1 1 6 . . . I / O コントローラ ,

20

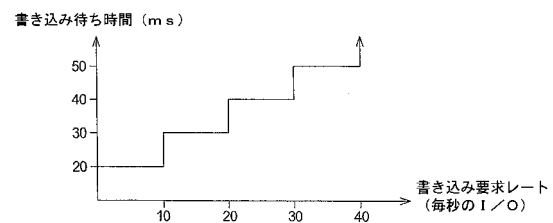
【図 1】



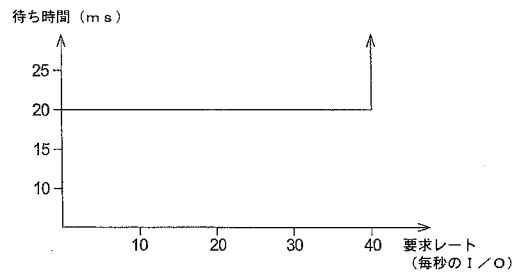
【図 2 A】



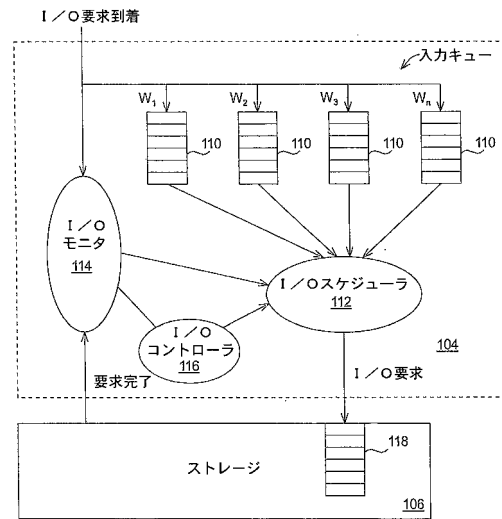
【図 2 B】



【図 3】



【図 4】



フロントページの続き

- (72)発明者 ラム・クリストファー
アメリカ合衆国カリフォルニア州 サンフランシスコ チャーチストリート 1181
- (72)発明者 ギラーム・アルバーツ
アメリカ合衆国カリフォルニア州 サンノゼ ブロッサムヒルロード 1615

合議体

審判長 清水 稔
審判官 衣川 裕史
審判官 大野 克人

- (56)参考文献 特開平8-55071(JP,A)
特開2002-132698(JP,A)
金井秀明 他, Webを用いたネットワーク構成視覚化システムの構築, 情報処理学会研究報告, 社団法人情報処理学会, 1998年11月20日, 第98巻, 第104号, p.1~6

- (58)調査した分野(Int.Cl., DB名)
G06F 17/60, G06F 3/06