C-002

拠点間高可用ストレージを実現するデータ二重化方式
Design of Data Duplication for Multi-Site High-Availability Storage Function

長尾 尚† **斎藤 秀雄**† **川口 智大**†
Takashi Nagao Hideo Saito Tomohiro Kawaguchi

1. はじめに

基幹システムでは，ストレージの障害時でも業務を継続できるよう，業務データを遠隔地にコピーする．この実現手段として，2 台のストレージ間で常にデータを二重化して整合性を保証する拠点間高可用ストレージ[1]機能が挙げられる(図 1)．一般に，サーバからストレージへのアクセスの応答時間が長くなると，アプリケーションの処理性能が低下する．このため，二重化のための通信時間の増加を抑えることが課題であった．本稿では，応答時間の増加を抑えるデータ二重化方式を述べる．

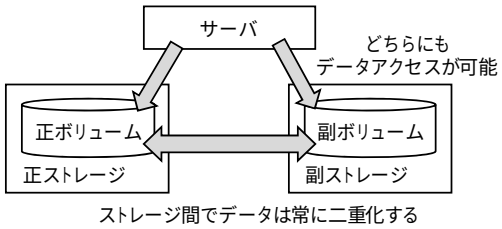

図 1　拠点化高可用ストレージの概要

2. 拠点間高可用ストレージにおけるコマンド処理

本章では，ストレージのアクセス応答性能に影響するストレージのコマンド処理を説明し，コマンド処理に要する時間を定式化する．

2.1 コマンド処理の流れ

図 2 に，正ストレージと副ストレージ間のデータの二重化方式を示す．データ更新時は，ストレージ間のデータの整合性を保証するため，必ず正ストレージ側からデータを更新する．例えば，正ストレージがコマンドを受領した場合，正ストレージは正ストレージのデータ更新後に副ストレージへデータを送信し，副ストレージにデータを更新させる．また，副ストレージがコマンドを受領した場合，副ストレージはまず正ストレージへデータ送信し，正ストレージにデータ更新させた後に副ストレージはデータを更新する．このように正副のストレージが通信しあうことで，整合性を保証する．

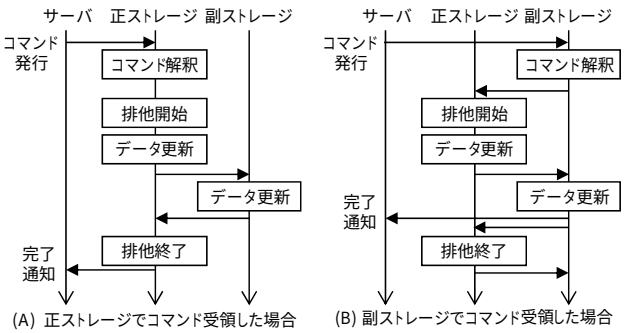

図 2　更新系コマンドの処理

図 3 に，データ参照時のコマンド処理方式を示す．データ参照時は，コマンドを受領したストレージが自身のデータを返却するため，ストレージ間の通信は発生しない．

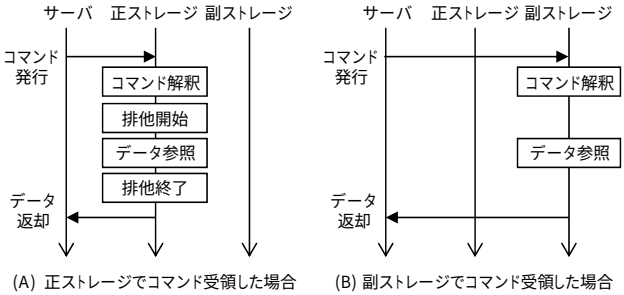

図 3　参照系コマンドの処理

2.2 コマンド応答時間の定式化

ストレージ間の通信が発生するデータ更新時に着目し，コマンド応答時間を定式化する．どちらのストレージでコマンドを受領しても処理の内容は同じため，正ストレージでコマンドを受領した場合を想定する．データ更新時の処理時間を図 4 に示す．サーバからストレージへの更新データの送信時間は，データ転送時間(T_{rcv})と通信遅延(T_{d1})の和である．ここで，送信データのサイズを x_1，各装置間の通信帯域を R_s，ストレージ内のメモリ帯域を R_m とすると，データ転送時間(T_{rcv})は式(1)となる．次に，ストレージは，受信したデータをキャッシュに書き込むことでデータを更新する．このため，書き込みデータのサイズを x_2 とすると，データ更新時間(T_{wrt})は式(2)となる．相手ストレージへのデータ送信は，データ転送時間(T_{snd})と通信遅延(T_{d2})の和となる．相手ストレージへの送信データのサイズ x_3 とすると，データ転送時間(T_{snd})は式(3)となる．さらに，完了通知における通信遅延 T_{d1} と T_{d2} を加え，コマンド応答時間(T)は式(4)となる．

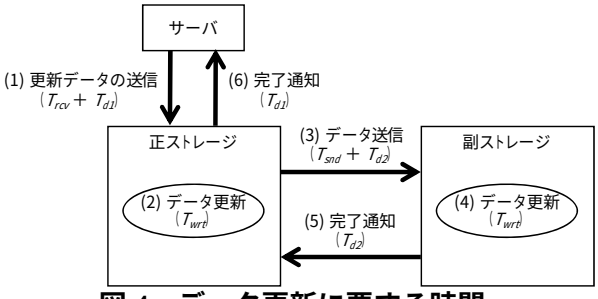

図 4　データ更新に要する時間

$$T_{rcv} = x_1 / R_s + x_1 / R_m \quad \cdots(1)$$

$$T_{wrt} = x_2 / R_m \quad \cdots(2)$$

$$T_{snd} = x_3 / R_m + x_3 / R_s \quad \cdots(3)$$

$$T = T_{rcv} + 2T_{wrt} + T_{snd} + 2T_{d1} + 2T_{d2} \quad \cdots(4)$$

本稿では，表 1 に示す値を定数に使用する．各通信遅延は，サーバと正ストレージは同じ拠点に，副ストレージは別拠点に設置されている想定としている．

表 1　応答時間算出のパラメタ

項目	例	値
R_m	DDR3-2133[2] (Dual Channel)	34.1 GB/s
R_s	8G FC (Fibre Channel)	0.8 GB/s
T_{d1}		0 ms
T_{d2}		1 ms

3. SCSI コマンドの処理方式

本章では，主要な SCSI コマンド[3][4]のうち，ストレージ間の通信が発生するものについて，通信処理による応答時間への影響を極小化する方法について説明する．

基幹システムで利用される主要なアプリケーションのひとつである VMware から発行される I/O コマンドをサンプルとして評価した．

[5]によると，VMware®では SCSI コマンドの平均応答時間を 10ms 以内にするようガイドラインが設定されているため，10ms 以下の応答時間を目標とした．

3.1 WRITE コマンド

WRITE コマンドは指定するデータに更新する基本的なコマンドである．VMware VMFS-5 の標準ブロックサイズは 1MB[6]であり，VMware はブロックサイズ単位にデータを更新する．このとき，x1 = x2 = x3 = 1MB であるため，T = 約 3.3ms となり，目標を達成できる．

3.2 PERSISTENT RESERVE OUT コマンド

PERSISTENT RESERVE OUT コマンドは共有ボリュームの排他を開始／終了するコマンドである．サーバは Key と呼ぶ 8B の制御情報を指定する．このとき，x1 = x2 = x3 = 8B であるため，T = 約 2.0ms となり，目標を達成できる．

3.3 COMPARE AND WRITE コマンド

COMPARE AND WRITE コマンドは，ストレージがサーバから受信した更新前データと現在のデータが一致する場合にのみ，更新後データに更新するコマンドである．VMware は共有ボリュームへの排他確保のために本コマンドを使用する．本コマンドの処理の流れを図 5 に示す．

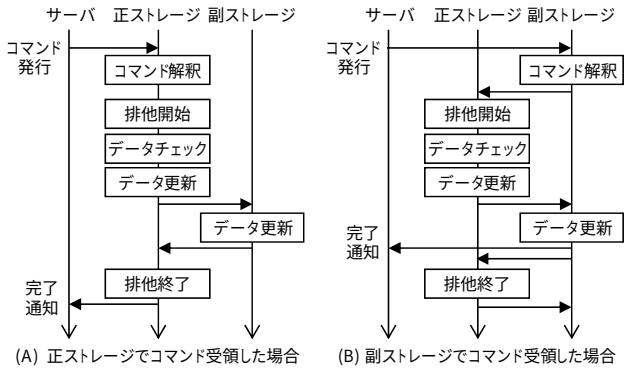

図 5　COMPARE AND WRITE コマンドの処理

正ストレージで排他開始後にデータの一致チェックを実施する．一致していれば処理を続行し，副ストレージではチェックしない．VMware は共有ボリュームの排他に本コマンドを利用する．更新データのサイズは 512B のため，$x_1 = x_2 = x_3$ =512B であるため，T = 約 2.0ms となり，目標を達成できる．

3.4 WRITE SAME コマンド

WRITE SAME コマンドは指定されたパタンでデータを広範囲に繰り返し書き込むコマンドである．サーバは 512B でパタンを指定する．書き込み範囲は，最大 256MB である．ここで，16MB の更新データを相手ストレージに送信すると，x1＝512B，x2＝x3 = 256MB であるため，T = 約 329.1ms になり目標時間を超過する(図 6)．このため，512B のパタンを相手ストレージに送信して，相手ストレージで繰り返し書き込みを行う．これにより，x1＝x3＝512B，x2=256MB とすることができ，T = 約 9.3ms となり，目標を達成できる．

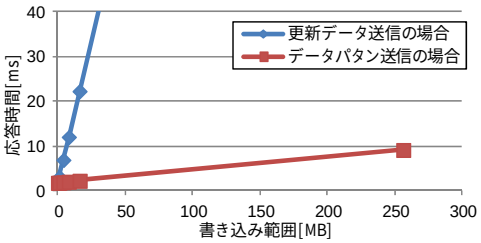

図 6　WRITE SAME コマンドの応答時間

3.5 EXTENDED COPY コマンド

EXTENDED COPY コマンドは任意の 2 領域間のデータコピーを実施するコマンドである．VMware では，非同期のデータコピーを期待するため，本コマンドの応答時間は問題とならない．

4. おわりに

今回，拠点間高可用ストレージにおけるコマンド応答時間を定式化し，VMware 環境を想定して主要なデータ更新コマンドの応答時間が目標時間内に収まるかを検証した．この結果，目標応答時間内にコマンド処理を完了できる見通しを得た．

参考文献

[1] HDS, “Hitachi Global-Active Device: Continuous Operation and Availability for Key Applications”, HDS Datasheet (2015).
[2] JEDEC, “DDR3 SDRAM Standard”, JESD79-3F(2012).
[3] T10 Technical Committee, “Information technology - SCSI Block Commands – 4 (SBC-4) ”, INCITS 506 (2015).
[4] VMware, “VMware vSphere Storage APIs – Array Integration (VAAI)”, VMware Technical Marketing Documentation (2012)
[5] VMware, “Using esxtop to identify storage performance issues for ESX / ESXi (multiple versions)”, VMware Knowledge Base 1008205 (2014).
[6] VMware, “Block size limitations of a VMFS datastore”, VMware Knowledge Base 1003565 (2013)

†(株) 日立製作所 情報通信イノベーションセンタ, Center for Technology Innovation, Hitachi Ltd.