HP ProLiant Server + InfiniBand + IO アクセラレータ

+ CLUSTERPRO 検証報告

日本ヒューレット・パッカード株式会社
テクノロジーサービス事業統括
テクノロジーコンサルティング統括本部

日本電気株式会社
第一 IT ソフトウェア事業部
CLUSTERPRO グループ

2011 年 9 月 1 日

1　はじめに

今日のコンピュータシステムにおいて、CPU、メモリ、I/O（ディスク、ネットワーク）のパフォーマンスギャップは拡大の一途を辿っています。

CPU の領域においては ノースブリッジの機能を取り込んだり、コアにより近い位置にキャッシュメモリを配置したりするなど、CPU とメモリとのパフォーマンスギャップを埋めるべく開発が行われています。

一方、ディスクの領域においては、SSD[1]が普及し、ネットワークの領域では、10 ギガビットイーサネット（10GbE）の普及や、InfiniBand の広帯域性・低遅延性が見直されるなど、メモリと I/O とのパフォーマンスギャップ拡大を緩和する動きもあります。

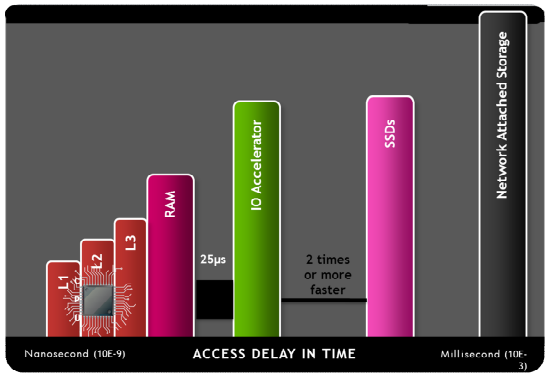
ソフトウェアはこれらの傾向の恩恵を素直に享受できます。つまり「S/W は何もしなくても H/W が速くなれば、システムが速くなる」というわけです。

ノード間でデータミラー処理を行うソフトウェアもこの例に漏れません。データミラー処理とは、ノード1で発生するディスクへの書込み I/O をノード2でも同様に実行することでノード間のデータを同期させるいわばネットワーク RAID を実現するものです。これを応用して、アプリケーションがノード 1 でも 2 でも最新のデータを保持するようにし、ノード 1 で障害が発生した場合にノード 2 でアプリケーションを継続実行させる仕組みがデータミラー型 HA クラスタです。

この仕組みの平常時におけるシステムパフォーマンスは一般的に各ノードでの書込み I/O の処理性能によって上限が決定されます。つまり、各ノードの CPU 性能でもなく、ノード間のネットワーク性能でもなく、ディスク I/O 性能がシステムパフォーマンスのボトルネックになる、ということです。

つまりディスク I/O のパフォーマンスを向上させることがシステム全体のパフォーマンス改善につながります。そのための製品として日本 HP には超高速半導体ストレージである IO アクセラレータがあります。図 1 は通常の SSD に比べ IO アクセラレータが劇的に優れていることを表しています。

図 1．IO アクセラレータ Maximizes Value Potential



日本 HP はこの IO アクセラレータと InfiniBand を使用してデータベース性能を劇的に向上させるソリューションである、「高速 DB ソリューション[2]」を提供しています。このソリューションと NEC の CLUSTERPRO を組み合わせることで、高性能を保ちつつ信頼性をさらに向上させることが可能になります。構成概要は図 2 になります。

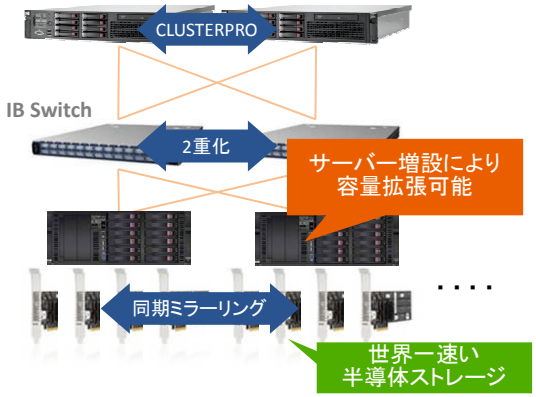
[1] ソリッドステートドライブ（Solid State Drive）

[2] http://h50146.www5.hp.com/services/ci/oracle/pdfs/das_proliant.pdf

図 2．高速 DB ソリューション（内蔵 IO アクセラレータ構成）＋ CLUSTERPRO

また、大きなデータ容量が必要な場合は、InfiniBand を利用した iSCSI 構成にすることで、冗長性を持った超高速 DB 構成が実現できます。構成概要は図 3 になります。

図 3．高速 DB ソリューション ＋ CLUSTERPRO



日本 HP には InfiniBand と IO アクセラレータを使用した高速 DB ソリューションに関する数々のノウハウや実績があり、NEC には CLUSTERPRO に関する数々のノウハウや実績があります。

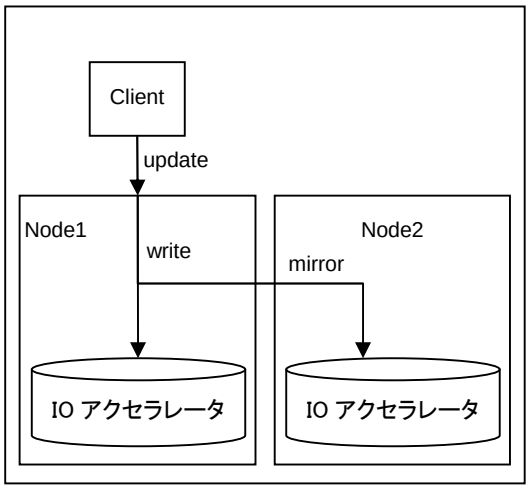
日本 HP と NEC は高速 DB ソリューションにクラスタソフトウェアとして CLUSTERPRO の同期ミラーリング機能を使用することで、PostgreSQL や Microsoft SQL Server をどの程度高速化できるのかを検証しました。

2　検証構成

サーバに HP ProLiant DL370 G6（以下 DL370 G6）、ディスクに IO アクセラレータ（PCIe に接続するタイプの SSD で、通常の SATA、SAS 接続の SSD より高性能な製品）、ネットワークに InfiniBand QDR(40Gbps)[4]、HA クラスタソフトに NEC CLUSTERPRO（データミラー型）[5]を使いました。

接続構成は図 4 の通りです。（構成要素の詳細は付録1を参照）

図 4．内蔵 IO アクセラレータ構成



クライアントからノード 1（DL370 G6）上の DB インスタンスへ更新トランザクション（write）が投入されると、SSD への write I/O は CLUSTERPRO によってノード 2（DL370 G6）の SSD へも投入されます。両ノードでの write が完了するとクライアントは更新トランザクションの完了を認識します。

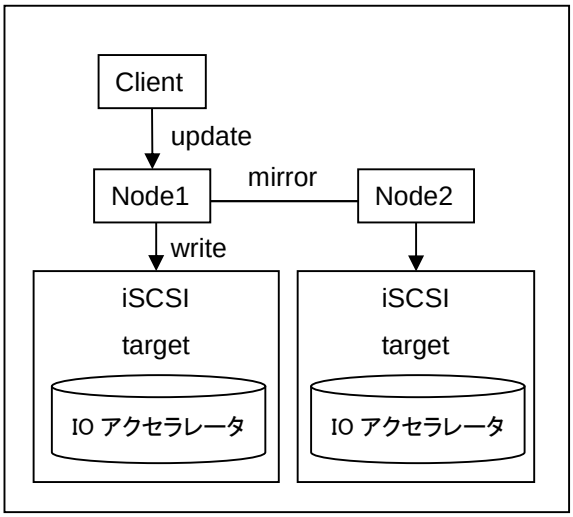
また図 5 に示すような iSCSI を使用した構成も検証しました。

図 5．iSCSI 構成

---

4 http://h18000.www1.hp.com/products/quickspecs/13078_na/13078_na.html

5 http://www.nec.co.jp/clusterpro
今回は同期ミラーリングを用いましたが、CLUSTERPRO は非同期ミラーリングもサポートしています。



両ノード（HP ProLiant DL380 G6（以下 DL380 G6））は iSCSI イニシエータとなり、DB ファイルを iSCSI ターゲット（DL370 G6）に配置します。図 4 との違いは DB ファイルを PCIe 接続の SSD 上に配置するか、iSCSI ターゲット上に配置するか、また、DL370 G6 と DL380 G6 の搭載メモリ量（DL370 G6 は 24GB、DL380 G6 は 6GB）です。両構成ともノード間通信は IPoIB（IP over InfiniBand）を使用しました。

3 動作確認

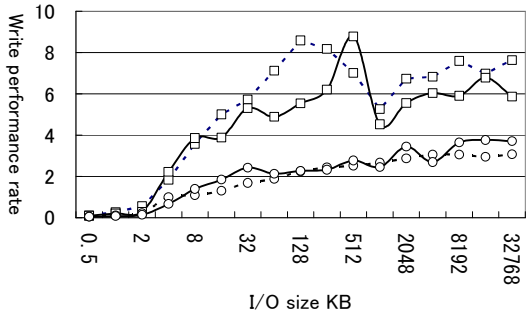
まず組合せそのものに問題が無いかを確認するために、対象 H/W へ CLUSTERPRO をインストール後、障害を発生させ、それを正しく検出してフェイルオーバを行うか検証しました。結果としては、全確認項目をクリアし DL370 G6、DL380 G6、IO アクセラレータ、InfiniBand、と CLUSTERPRO の組合せそのものには全く問題が無い事を確認できました。検証項目と結果の詳細は付録 2 を参照下さい。

4 ディスク I/O 性能測定

単一の業務アプリケーションから見たとき、どの程度のディスク I/O 性能が得られるのか、という観点から、以下の条件で現用系ノードのミラー化されたディスクの I/O 性能を測定しました。

- シングルインスタンス
- I/O サイズ 512Byte ～ 32MByte の シーケンシャル I/O, ランダム I/O
- I/O パターン は 0% read, 100% write[6]

グラフ 1



□ ： シーケンシャル I/O

○ ： ランダム I/O

実線 ： ミラーリング有 （CLUSTERPRO 有）

点線 ： ミラーリング無 （CLUSTERPRO 無）[7]

ミラーリング有りのシーケンシャル I/O で得られたピーク性能は 1GB/sec を軽く超え、実に 100Mbit イーサネットの論理性能比で 100 倍以上の性能が得られました。「単一の業務アプリケーションが得られるディスク I/O 性能」という意味では、非常に高い性能です。

検証構成と一般的なサーバ[8]とで、以下については、同じ傾向が見られました。

- ネットワーク I/O 性能 ≫ ディスク I/O 性能

[6] ミラーリング機能は read の性能劣化は殆ど無く、 write の性能劣化が避けられない、という性質があるため、write 性能の劣化度合いを調べるための条件です。

[7] Y 軸はミラーリング無しの 4KB ランダム I/O の性能を 1.0 としてスケーリングしました。

[8] 2way Nehalem + SATA Disk + 1Gbps Network

であり、ボトルネックはディスク、及び、ディスク I/O に関連するキャッシュ。

- ランダム I/O では「ミラーを行わないときのディスク性能」≒「ミラーを行うときのディスク性能」

シーケンシャル I/O では I/O サイズによってミラーを行うときのディスク性能が行わないときの性能に対して勝ったり劣ったりしており、グラフが全体的にデコボコしています。これは、NAND フラッシュメモリデバイスへ read 0%、 write 100%の I/O パターンを発行した場合に見られる「書込速度の揺れ」が観測されたものと思われます。

この結果から、CLUSTERPRO と Infiniband を組み合わせることにより、IO アクセラレータの性能を高いレベルで引き出していることが伺えます。

IO アクセラレータでは、フォーマット方式を変更し予備領域を増やすことでシーケンシャルライトの性能をより安定化させるチューニングが可能です。しかし、今回の検証ではすでに 1GB/s を超えるシーケンシャル性能が確認できており、IO アクセラレータが標準フォーマットのままでも高いディスク I/O 性能が維持できていることがわかります。

5 トランザクション性能

データベースに PostgreSQL9.0.4、Microsoft SQL Server 2008 R2、ベンチマークソフトには JdbcRunner[10]に含まれる Tiny TPC-C を使用して、

(1) Red Hat Enterprise Linux（以下 RHEL） + IO アクセラレータ構成

(2) RHEL + iSCSI　IO アクセラレータ構成

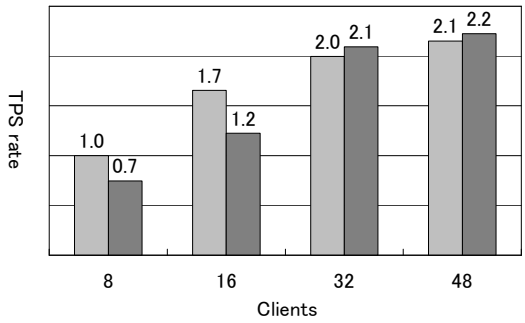
(3) Windows + iSCSI　IO アクセラレータ構成

の各構成に対して性能測定を行いました。

TPC-C とは、TPC によって策定された OLTP ベンチマーク仕様の一つです。卸売業における注文・支払いなどの業務をモデルにしており、概ね write : read = 9 : 1 のトランザクションを実行します。

（1） RHEL + 内蔵 IO アクセラレータ構成

グラフ 2



■(左側) : ミラーリング無（CLUSTERPRO 無）

■(右側) : ミラーリング有（CLUSTERPRO 有）

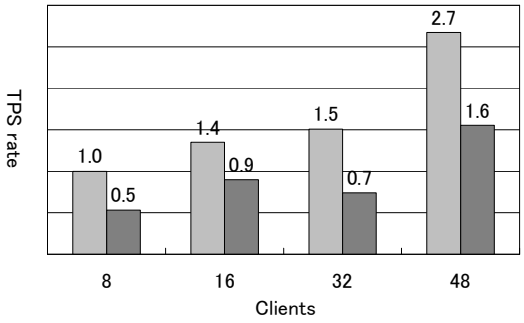
ミラーリング無しの構成に比べてミラーリング有りの構成は、概ね 25%程度の性能低下が起こる、と解釈してよいでしょう。32 クライアントと 48 クライアントのケースは、概ね同じ性能となっていますが、このとき ミラーリング機能の使用有無に関わらず CPU 使用率、ディスク I/O 使用率が同程度になっていたため、ディスクやネットワークではなく CPU がボトルネックになっていると思われます。

今回測定した 3 つのパターンのうち、トランザクション性能の絶対値が最も高いのがこのパターンです。

10 http://hp.vector.co.jp/authors/VA052413/jdbcrunner/

（2） RHEL + iSCSI IO アクセラレータ構成

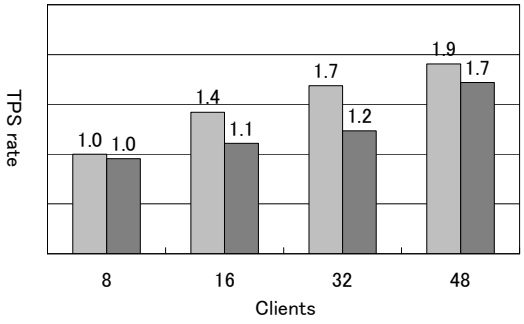
グラフ 3



ミラーリング無しの構成に比べてミラーリング有りの構成は、概ね、35～50%程度の性能低下が起こる、と解釈してよいでしょう。32 クライアントと48 クライアントのミラーリング機能を使用するケースでは、VACUUM の影響なのか、ベンチマーク中に iowait[11]が極端に高くなる(CPU 時間の80～90%)時間帯があり、その影響でパフォーマンスが出ていないと思われます。

（3） Windows + iSCSI IO アクセラレータ構成

グラフ 4

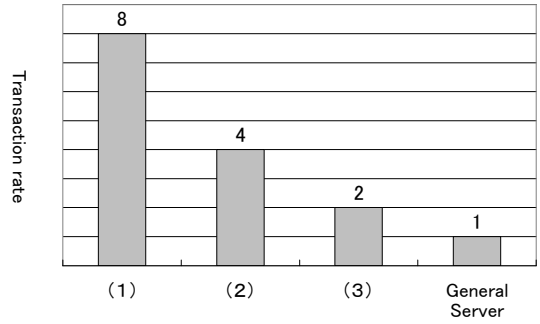


ミラーリング無しの構成に比べてミラーリング有りの構成は、10～25%程度の性能低下が起こると解釈してよいでしょう。トランザクション性能の絶対値が最も低かったのがこのパターンでしたが、それでも一般的なサーバと比べて二倍強の性能であることは特筆に値します。

OS が Windows に変わっただけなので、IO アクセラレータや InfiniBand（IP over InfiniBand）のドライバを含む Windows の I/O 性能にチューニングの余地があるようです。

ミラーリング機能を使用する場合、各構成のトランザクション性能は概ねグラフ 5 のような傾向になりました。

グラフ 5



上記では「思った程パフォーマンスが出なかった理由」を述べた部分がありますが、「一般的なサーバとの比較」という観点に立てば何れも高性能であると言えます。

6 データミラー復旧速度

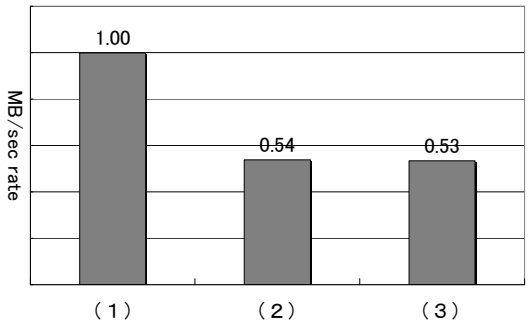
CLUSTERPRO によるデータミラー型クラスタでは、一方のノードが停止している間、稼動しているノードでは更新データが記録されます。

その後、停止していたノードが稼動状態になったら、更新データを転送することでデータを同期状態にする復旧処理が行われます。

今回は更新データ量と復旧処理に掛かった時間から復旧速度を算出し、各構成を比較しました。

[11] CPU 総時間当たりの I/O 待ち時間の割合

グラフ 6



IO アクセラレータ構成の方が iSCSI 構成の倍近い性能になっています。構成上、目立つ差異は、iSCSI の存在と、メモリ量です。RHEL IO アクセラレータ構成は搭載メモリ 24GB、iSCSI 構成は RHEL Windows 何れも 6GB で、バッファに使用されるメモリ量は RHEL IO アクセラレータ構成が最も多く、これが性能に影響していると考えられます。

尚、RHEL IO アクセラレータ構成はもっと性能が出てもよいのではないか とも思えます。これは、復旧処理が単一プロセスで実行され、単一の CPU コアのみ高使用率となり、CPU がボトルネックになりやすいことが影響していると考えられます。

7 まとめ

CLUSTERPRO のミラーリング機能により若干のパフォーマンス劣化は起こるものの、従来型の HDD 構成に比べ十分高いパフォーマンスが得られることが確認できました。

- CLUSTERPRO のミラーリング有りの構成で、IO アクセラレータを使用するパターンは、使用しないパターン（SATA HDD 構成）に比べ約 1.5～3.3 倍のパフォーマンスが得られました。
- IO アクセラレータを使用する構成で、CLUSTERPRO のミラーリング有りのパターンは、無しのパターン（単体サーバ構成）に比べ約 25%～50% のパフォーマンス低下が見られました。

ただし、トランザクションベンチマークは write : read = 9:1 の割合で書き込みが発生しているため、読み込みの割合が増えるほどパフォーマンスの劣化は低くなると考えられます。

また、IO アクセラレータ構成と iSCSI 構成の結果から、メモリサイズが大きい方がディスクに対する負荷が小さくなるため、キャッシュを増やすことで CLUSTERPRO のミラーによる性能劣化の影響を減らすことができると考えられます。

8 今後の展望

今回使用したような高速な I/O が可能なシステムは、BI（ビジネスインテリジェンス）をはじめとする「巨大なデータを素早く扱いたい」というデマンドに対するソリューションとして今後ポピュラーになっていく事が予想されます。

お問合わせ先

日本ヒューレット・パッカード株式会社

jpn_tc_solution_desk@hp.com

日本電気株式会社

info@clusterpro.jp.nec.com

付録1

図 1 で使用するノード 1,2

図 2 で iSCSI target として使用するサーバ

サーバ	HP ProLiant DL370 G6
OS	RHEL 5.5 x86_64
System ROM	P63　08/16/2010
CPU	Intel(R) Xeon(R) CPU　E5520 2.26GHz
CPU 数	2P（8core） HT=on
メモリ	24GB
オンボード NIC	HP NC375i PCIe Dp Gigabit Svr Adptr
InfiniBand HCA	HP IB 4X QDR CX-2 PCI-E G2 DUAL PORT HCA(MT26428)
InfiniBand driver	VoltaireOFED-1.5.2_2
内蔵ディスク	HP 146GB 6G SAS 10K 2.5in DP ENT HDD
ディスクコントローラ	Smart アレイ P410i/256MB コントローラ
IO-A HW	HP 320GB SLC PCIe IO アクセラレータ Duo ×2
IO-A firmware/driver	Firmware v43674/ driver 2.2.0
IO-A フォーマット	Advertised Capacity

図 2 で使用するノード 1,2

サーバ	HP ProLiant DL380 G6
OS	RHEL 5.5 x86_64 Windows Server 2008 R2 x86_64
System ROM	P62 03/30/2010
CPU	Intel(R) Xeon(R) E5540 2.53GHz
CPU 数	1P（4core） HT=on
メモリ	6GB
オンボード NIC	HP NC382i PCIe Dp Gigabit Svr Adptr
InfiniBand HCA	HP IB 4X QDR CX-2 PCI-E G2 DUAL PORT HCA(MT26428)
InfiniBand driver	VoltaireOFED-1.5.2_2(RHEL) MLNX WinOF VPI MSI v2.1.3 for x64 Platforms(Windows)
内蔵ディスク	HP 146GB 6G SAS 10K 2.5in DP ENT HDD
ディスクコントローラ	Smart アレイ P410i/256MB コントローラ
IO-A HW	-
IO-A firmware/driver	-
IO-A フォーマット	-

使用したソフトウェア・バージョン

CLUSTERPRO X3.0 for Linux
CLUSTERPRO X3.0 for Windows
Red Hat Enterprise Linux 5.5
Windows Server 2008 R2
PostgreSQL 9.0.4
Microsoft SQL Server 2008 R2

付録 2

状態遷移試験

項番	操作	確認事項	確認		
			1[12]	2[13]	3[14]
1	クラスタ生成	自動ミラー初期構築が行われること。 自動的にグループが起動すること。	OK	OK	OK
2	グループの停止	グループが正常に停止すること。	OK	OK	OK
3	グループの起動	グループが正常に起動すること。	OK	OK	OK
4	プライマリサーバのシャットダウン	セカンダリサーバに正常にフェイルオーバされること。	OK	OK	OK
5	プライマリサーバの起動	プライマリサーバが正常に起動すること。 自動ミラー復帰が行われること。	OK	OK	OK
6	セカンダリサーバのシャットダウン	プライマリサーバに正常にフェイルオーバされること。	OK	OK	OK
7	セカンダリサーバの起動	セカンダリサーバが正常に起動すること。 自動ミラー復帰が行われること。	OK	OK	OK
8	グループの移動	プライマリサーバからセカンダリサーバへ正常にグループ移動すること。	OK	OK	OK
9	DB のストール	DB モニターが異常を検出し、正常にフェイルオーバされること。	OK	OK	OK
10	クラスタのシャットダウン	両サーバが正常にシャットダウンすること。	OK	OK	OK
11	両サーバの起動	自動的にグループが起動すること。	OK	OK	OK

[12] Linux + IO アクセラレータ構成
[13] Linux + iSCSI 構成
[14] Windows + iSCSI 構成

Linux + IO アクセラレータ構成 障害試験

No	検証項目	内容	実行方法	結果	備考
1	パス障害 （レプリケーション用のセグメント）	DB へのアクセス中（ベンチマークツール実行中）に、InfiniBand HCA の port1 障害が起こることによる挙動	レプリケーション用セグメント Port1 の断線	⇒正常に稼動	
2	パス障害 （レプリケーション用のセグメント）	DB へのアクセス中（ベンチマークツール実行中）に、レプリケーション用のセグメント障害が起こることによる挙動	レプリケーション用セグメント Port2 の断線	17 秒 IO ストール ミラーブレイク FO(*)発生しない	ミラーディスクコネクト(*)の HB(*)タイムアウトが 16 秒 インターバルは 10 秒 ⇒16～26 秒＋α(1～3 秒)くらいで IO を AP に戻す。
3	ディスク障害 （IO アクセラレータ障害）	DB へのアクセス中（ベンチマークツール実行中）に、dd コマンドにより、論理的にプライマリ側の IO-A ディスク障害を起こしたときの挙動	データパーティションに /dev/zero で上書き	PostgreSQL 監視により FO 実行するも、DB が不正状態	再活性させるには、DB 再構築が必要
4	DB サーバ障害 （カーネルパニック）	DB へのアクセス中（ベンチマークツール実行中）に、プライマリ DB サーバ側をカーネルパニックにより、サーバ障害を起こしたときの挙動	echo c > /proc/sysrq-trigger の実行	HB タイムアウトにより FO ⇒正常に稼動	パニックした側で mdagent 起動時に、/dev/fioa1 が開けなかったが、mdagent 再起動で正常に戻る。原因は、SSD のリカバリ処理によりデバイスの準備ができる前に、mdagent サービスが起動したため。
5	レプリケーション再同期時間	セカンダリのファイルシステムをフォーマット後、約 20GB のデータの再同期が完了する時間の測定	Web Manager からデータフルコピー	18 分 22 秒	ボリューム全体サイズ：160GB 実データサイズ：77GB
6	DB ストール検証	DB へのアクセス中（ベンチマークツール実行中）に、DB ストール障害を起こしたときの挙動	SIGSTOP でプロセスストール実行	DBAgent の異常検出により FO ⇒正常に稼動	

Linux + iSCSI 構成 障害試験

No	検証項目	内容	実行方法	結果	備考
1	パス障害 (レプリケーション用のセグメント)	DB へのアクセス中(ベンチマークツール実行中)に、InfiniBand HCA のレプリケーション用セグメントの port 障害が起こることによる挙動	Port の断線	IO ストール 28 秒 ⇒正常に稼動	ミラーディスクコネクトのハートビートタイムアウトは 16 秒 インターバルは 10 秒 ⇒16～26 秒＋α(1～3 秒)くらいで IO を AP に戻す。
2	パス障害 (iSCSI 用のセグメント)	DB へのアクセス中(ベンチマークツール実行中)に、InfiniBand HCA の iSCSI 用セグメントの port 障害が起こることによる挙動	Port の断線	DB サーバにパニックが発生し、HB タイムアウトによる FO(*) ⇒正常に稼動	
3	DB サーバ障害 (カーネルパニック)	DB へのアクセス中(ベンチマークツール実行中)に、プライマリ DB サーバ側をカーネルパニックにより、サーバ障害を起こしたときの挙動	echo c > /proc/sysrq-trigger の実行	HB タイムアウトによる FO ⇒正常に稼動	
4	ストレージサーバ障害 (カーネルパニック)	DB へのアクセス中(ベンチマークツール実行中)に、ストレージサーバ側をカーネルパニックにより、サーバ障害を起こしたときの挙動	ストレージサーバで echo c > /proc/sysrq-trigger の実行	DB サーバにパニックが発生し、HB タイムアウトによる FO ⇒正常に稼動	SSD のリカバリ処理によりデバイスの準備ができる前に、iSCSI ターゲットサービスが起動し、DB サーバ側からディスクが見えない状況に。ストレージサーバの再起動が必要
5	レプリケーション再同期時間	セカンダリのファイルシステムをフォーマット後、約 20GB のデータの再同期が完了する時間の測定	Web Manager からデータフルコピー	10 分 37 秒	ボリューム全体サイズ:320GB 実データサイズ:34GB

Windows + iSCSI 構成 障害試験

No	検証項目	内容	実行方法	結果	備考
1	パス障害 （レプリケーション用のセグメント）	DB へのアクセス中（ベンチマークツール実行中）に、InfiniBand HCA のレプリケーション用セグメントの port 障害が起こることによる挙動	Port の断線	IO ストール 18 秒 ⇒正常に稼動	コネクトタイムアウト 20 秒が起因
2	パス障害 （iSCSI 用のセグメント）	DB へのアクセス中（ベンチマークツール実行中）に、InfiniBand HCA の iSCSI 用セグメントの port 障害が起こることによる挙動	Port の断線	ミラー領域に対するディスク監視が異常を検出し、OS シャットダウンによる FO(*) ⇒正常に稼動	ディスク監視およびミラーディスク監視の監視インターバル＋タイムアウトがミラーコネクトタイムアウトより小さくなるようにチューニングする。 ・ミラーディスクの設定値 ミラーコネクトタイムアウト：120 秒 ・各監視リソースの設定値 インターバル：5 秒　タイムアウト：30 秒 最終動作：OS シャットダウン
3	DB サーバ障害 （電源 OFF）	DB へのアクセス中（ベンチマークツール実行中）に、プライマリ DB サーバ側を電源 OFF により、サーバ障害を起こしたときの挙動	電源 OFF	ハートビートタイムアウトによる FO ⇒正常に稼動	
4	ストレージサーバ障害 （カーネルパニック）	DB へのアクセス中（ベンチマークツール実行中）に、ストレージサーバ側をカーネルパニックにより、サーバ障害を起こしたときの挙動	echo c > /proc/sysrq-trigger の実行	#2 と同じ	#2 と同じ
5	レプリケーション再同期時間	セカンダリのファイルシステムをフォーマット後、約 20GB のデータの再同期が完了する時間の測定	Web Manager からデータフルコピー	10 分 42 秒	ボリューム全体サイズ：320GB 実データサイズ：31GB