No.ALT-20140326006312

富士通社製サーバ『PRIMERGY RX200S8』と HGST(旧 Virident)社製ソフトウェア 『FlashMAX Connect』の機能検証報告書

2014 年 3 月 26 日
株式会社アルティマ

目次

1. 検証目的

本検証は富士通社製サーバ PRIMERGY RX200S8 と HGST 社製 FlashMAX II (PCIe Storage Class Memory)及び FlashMAX Connect (ソフトウェアスイート)をご安心してご利用して頂く為に、事前に接続性と基本動作を確認することを目的としております。

2. FlashMAX II について

FlashMAX II は PCIe インターフェースを利用したストレージクラスメモリ(SCM)で、最小スペースで非常に高いストレージ容量を搭載し、様々な負荷環境下においても性能劣化を防ぎ安定した高い性能を維持できるエンタープライズ向けの製品となります。FlashMAX II は最小 550GB から最大で 4.8TB の容量をロープロファイルの形状で提供しており、補助電源を必要としないため、サーバ搭載の際の自由度が高く、データセンターのスペース・消費電力の削減に高く貢献することができます。

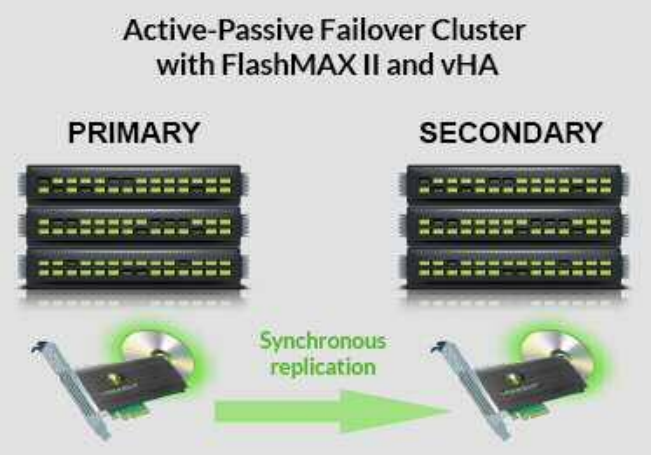
3. FlashMAX Connect について

FlashMAX Connect は FlashMAX デバイス上で動作するソフトウェアスイートで、vHA、vShare、vCache、vStore の 4 つのモジュールから構成されております。標準サーバに搭載された FlashMAX デバイスと組み合わせることで、高い性能とスケーラビリティ、また信頼性を実現したエンタープライズストレージを容易に構築することができ、既存の高価格なプロプライエタリの SAN ストレージネットワークを置き換えることができます。

それぞれのモジュールの機能を下記致します。

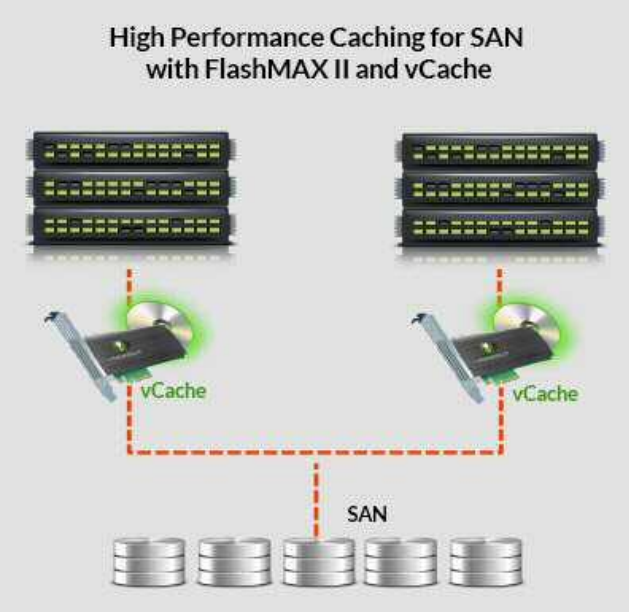
3-1 High Availability

vHA は FlashMAX デバイス上に格納されているデータに対して、サーバ間で高帯域・低レイテンシの同期ミラーリングを実現します。PCIe フラッシュの性能とレイテンシを犠牲にすることなくフェイルオーバークラスタを作ることができます。



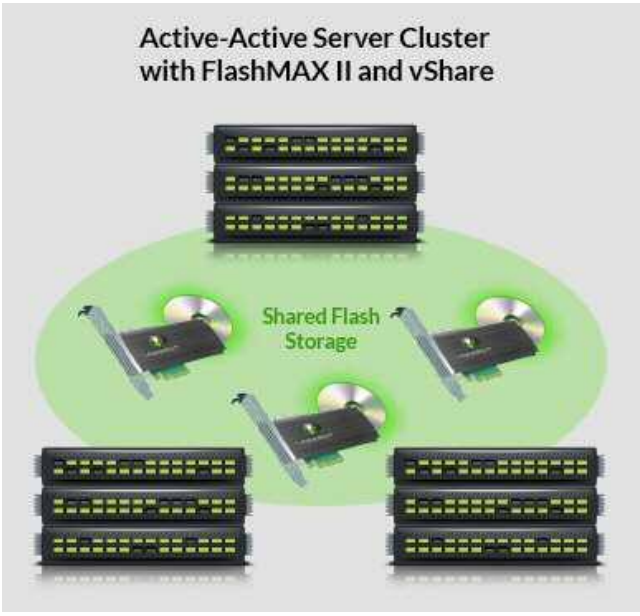
3-2 Caching

vCache は FlashMAX デバイスを高速なフラッシュキャッシュとして使用することができ、ブロックアドレスマッピングの重複を排除し、3rd Party の SSD のキャッシングソリューションと比較し最大で 6 倍の性能とより長いフラッシュの寿命を提供します。キャッシングモードとしては write-back、write-through、write-around から選択することができます。



3-3 Shared Access

vShare は複数のサーバから FlashMAX デバイスをブロックレベルに共有できる機能です。低レイテンシな共有フラッシュストレージネットワークとしてアプリケーションクラスタを構築できます。サービス規模が増大した場合にも性能とフラッシュストレージの容量を簡単に拡張することができます。



3-4 Base Storage Access

vStore は、FlashMAX デバイスのうち、vHA, vCache, vShare を必要としない領域を、搭載されている OS に対してローカルアクセスストレージデバイスを提供する機能になります。今回の検証では通常の FlashMAX デバイスで提供できる機能と同等であるために対象としておりません。

4. 検証構成

4-1 検証場所

場所:株式会社アルティマ
期間:2014 年 2 月 5 日～13 日

4-2使用機材

4-2-1 Hardware

名称	説明
PRIMERGY RX200S8 2 台	CPU : Xeon E5-2640v2(8 コア)　2.00GHz[2CPU] メモリ : 64GB SAS HDD : 300GB(2.5"SAS)×8(RAID5) IB HCA : MCX354A-FCBT (Mellanox 56Gbps ConnectX-3 HCA) ※FW version : 2.30.3110
FlashMAX II 2 枚	容量 : 2.2 TB 形状 : Low Profile

4-2-2 Software

マシン名	オペレーティングシステム	ソフトウェア
PRIMERGY RX200S8	Red Hat Enterprise Linux 6.4	IB Driver ・MLNX_OFED_LINUX-2.0-3.0.0 Oracle ・ Oracle Real Application Clusters 11gR2 (11.2.0.4)
FlashMAX II	Red Hat Enterprise Linux 6.4	kmod-vgc-redhat6.1+-1.2.FC-65244 vgc-rdma-2.6.32-358.el6.x86_64-1.2.FC-65244 vgc-rdma-utils-redhat6-1.2.FC-65244 vgc-utils-redhat6-1.2.FC-65244

4-2-3 サーバ構成

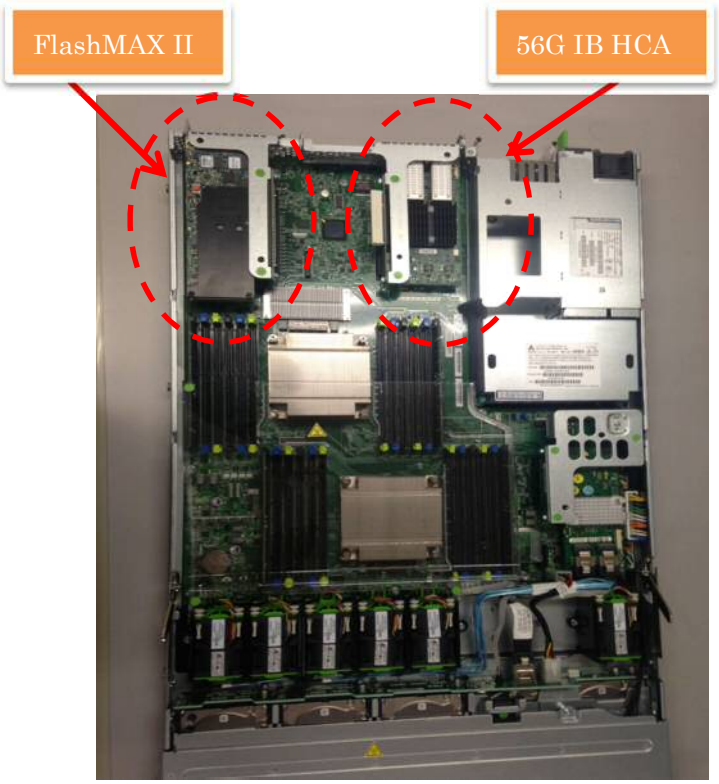


図 1 サーバ構成

4-2-4 ネットワーク構成

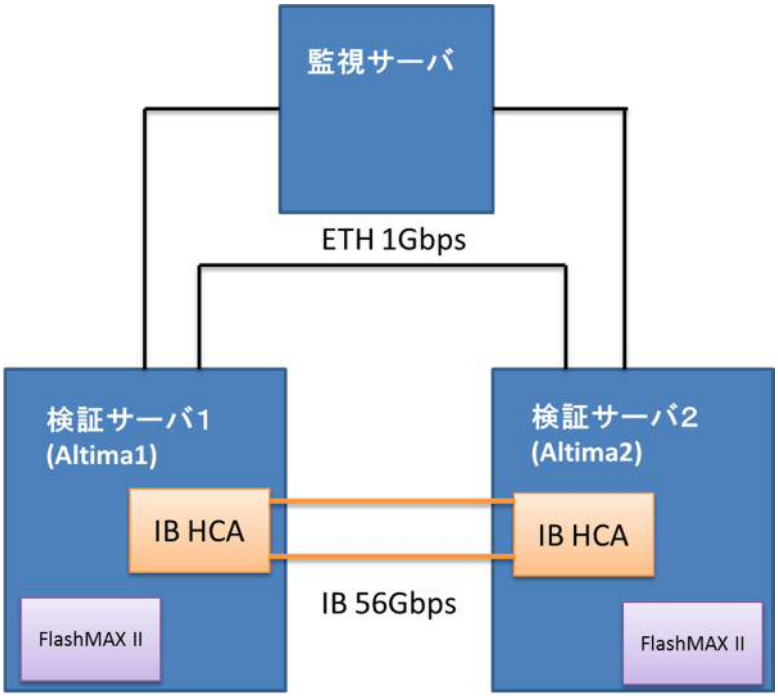


図 2 ネットワーク構成

5 動作確認及び結果

5-1 基本動作確認

1) デバイスの認識を確認する。

FlashMAX II 及び IB HCA が lspci で表示されることを確認した。

```
[root@altima1 ~]# lspci | grep Virident
04:00.0 FLASH memory: Virident Systems Inc. Virident FlashMAX Drive V2 (rev 01)
[root@altima1 ~]# lspci | grep Mellanox
02:00.0 Network controller: Mellanox Technologies MT27500 Family [ConnectX-3]
```

2) ソフトウェアのインストールを確認する。

MLNX OFED と FlashMAX Connect が正常にインストールされることを確認した。

```
[root@altima1 MLNX_OFED_LINUX-2.0-3.0.0-rhel6.4-x86_64]# ./mlnxofedinstall
This program will install the MLNX_OFED_LINUX package on your machine.
Note that all other Mellanox, OEM, OFED, or Distribution IB packages will be removed.
Do you want to continue?[y/N]:y
        .
[root@altima1 Downloads]# rpm -ivh kmod-vgc-redhat6.1+-1.2.FC-65244.V5A.x86_64.rpm
vgc-rdma-2.6.32-358.el6.x86_64-1.2.FC-65244.V5A.x86_64.rpm vgc-rdma-utils-redhat6-1.2.FC-65244.V5A.x86_64.rpm
vgc-utils-redhat6-1.2.FC-65244.V5A.x86_64.rpm
Preparing...                ########################################### [100%]
   1:kmod-vgc-redhat6.1+    ########################################### [ 25%]
   2:vgc-rdma-2.6.32-358.el6########################################### [ 50%]
   3:vgc-utils              ########################################### [ 75%]
   4:vgc-rdma-utils         ########################################### [100%]
[root@altima1 Downloads]#
```

3) FlashMAX II のドライバ起動を確認する。

FlashMAX II のドライバが正常に起動しデバイスが認識されることを確認した。

```
[root@altima1 Downloads]# service vgcd start
Loading kernel modules...                  [  OK  ]
Rescanning SW RAID volumes...              [  OK  ]
Rescanning LVM volumes...                  [  OK  ]
Enabling swap devices...                   [  OK  ]
Rescanning mount points...                 [  OK  ]
[root@altima1 Downloads]#
```

4)　OpenSM を起動しインフィニバンドネットワークのリンクアップを確認する。

ポートの状態が Active、また 56Gbps (FDR) でリンクアップしていることを確認した。

```
[root@altima1 ~]# ibstat
CA 'mlx4_0'
        CA type: MT4099
        Number of ports: 2
        Firmware version: 2.30.3110
        Hardware version: 1
        Node GUID: 0x0002c90300a51370
        System image GUID: 0x0002c90300a51373
        Port 1:
                State: Active
                Physical state: LinkUp
                Rate: 56
                Base lid: 1
                LMC: 0
                SM lid: 1
                Capability mask: 0x0251486a
                Port GUID: 0x0002c90300a51371
                Link layer: InfiniBand
```

5-2 FlashMAX Connect 機能確認

vgc-config ユーティリティにより vHA、vCache、vShare の機能が有効になることを確認した。

```
root@altima1 ~]# vgc-config -p /dev/vgca0 -m maxperformance --enable-vha --enable-vcache --enable-vshare
vgc-config: FlashMAX Connect Software Suite 1.2(65244.V5A)

*** WARNING: this operation will erase ALL data on this FlashMAX drive. Do you want to continue? [yes/no]: yes
*** Formatting drive.    Please wait... ***
[root@altima1 ~]# vgc-config
vgc-config: FlashMAX Connect Software Suite 1.2(65244.V5A)

Current Configuration:
/dev/vgca    1 partition(s)
        /dev/vgca0                  mode=maxperformance          sector-size=512        raid=enabled
                vcache=enabled                  vha=enabled          vshare=enabled
```

5-2-1 High Availability (vHA)

1) 検証サーバ 1 と検証サーバ 2 で 100GB の HA デバイスを作成しデバイスが正常に同期していることを確認した。

```
[root@altima1 ~]# vgc-vha-config --create --role primary --peer altima2 --size 100 vha1 /dev/vgca0
vgc-vha-config: FlashMAX Connect Software Suite 1.2(65244.V5A)

[root@altima2 ~]# vgc-vha-config --create --role secondary --peer altima1 --size 100 --uuid
6e855b57-9581-481e-93b4-ed1cef51b01d vha1 /dev/vgca0
vgc-vha-config: FlashMAX Connect Software Suite 1.2(65244.V5A)
[root@altima2 ~]# vgc-vha-monitor --list
vgc-vha-monitor: FlashMAX Connect Software Suite 1.2(65244.V5A)

  ------------------------------------------------------------------------------------------------------------
  vHA Device                          Local Device      Role              State              UUID
  ------------------------------------------------------------------------------------------------------------
  /dev/vgca0_vha1      /dev/vgca0        secondary         Connected         6e855b57-9581-481e-93b4-ed1cef51b01d
```

2) 障害復旧時の Rsync（再同期）動作を確認する。

IB ケーブルを抜去した状態で Primary(検証サーバ 1)にデータを書き込み、IB ケーブルを再挿入後 Rsync プロセスの開始と終了後に Secondary(検証サーバ 2)でデータの整合性を確認した。Rsync 動作またデータの整合性に問題がないことを確認した。

```
[root@altima1 ~]# vgc-vha-monitor --resync-progress /dev/vgca0_vha1
vgc-vha-monitor: FlashMAX Connect Software Suite 1.2(65244.V5A)

 Resync Progress: [##################################################################] 100 percent done.
```

3) 同期しているデバイス間のデータ整合性を確認する。

1GB のランダムファイルを 100 個生成し、Primary での書き込み、Secondary でデータの読み出しを行い、データの整合性に問題がないことを確認した。また、Primary と Secondary の役割(role)を入れ替えて同様に検証を行いデータの整合性に問題がないことを確認した。

5-2-2 Caching (vCache)

1) Caching の有効性を確認する。

検証サーバ 1 の/dev/sdb デバイスに対して FlashMAX II から 100GB のキャッシュを適応し、正常に設定が反映されていることを確認した。

```
[root@altima1 ~]# vgc-vcache-config --create --mode write-back --size 100 vchace1 /dev/vgca0 /dev/sdb
vgc-vcache-config: FlashMAX Connect Software Suite 1.2(65244.V5A)
 [root@altima1 ~]# vgc-vcache-monitor --list
vgc-vcache-monitor: FlashMAX Connect Software Suite 1.2(65244.V5A)
  --------------------------------------------------------------------------------
  vCache Device                  Mode          Frontend        Backend         State
  --------------------------------------------------------------------------------
  /dev/vgca0_vchace1             Write-back    /dev/vgca0      /dev/sdb        GOOD
```

2) 書き込み性能とデータの Flush 動作を確認する。

fio ツールを使用しキャッシュデバイスに対して書き込みを実施し、期待される性能が出ることまた Flush 機能が動作正常に HDD に Write-back されることを確認した。

```
[root@altima1 ~]# vgc-vcache-monitor --detail /dev/vgca0_vchace1
vgc-vcache-monitor: FlashMAX Connect Software Suite 1.2(65244.V5A)
  vCache Device                  Mode          Frontend        Backend         State
  /dev/vgca0_vchace1             Write-back    /dev/vgca0      /dev/sdc        GOOD
  Cache Details:
    vCache device          : /dev/vgca0_vchace1
    Backend size           : 500 GB
    Cache size             : 100 GB
    Dirty threshold        : 50 GB
    Total cached data      : 9062248448 (8 GiB)
    Total dirty data       : 1193185280 (1 GiB)
.............
```

3)　キャッシュされたデータとローカルディスクのデータの整合性を確認する。

1GB のランダムファイルを 100 個生成しキャッシュデバイスへの書き込み・読み込みをそれぞれ行い、元データと Write-back されたローカルディスクのデータの整合性に問題ないことを確認した。

5-2-3　Shared access (vShare)

vShare は Oracle Real Application Clusters と組み合わせることで最適なソリューションを実現します。今回は実際のサービスで使用することを考慮し Oracle Real Application Clusters をインストールして動作、また性能を検証する。

1)　ネットワーク構成

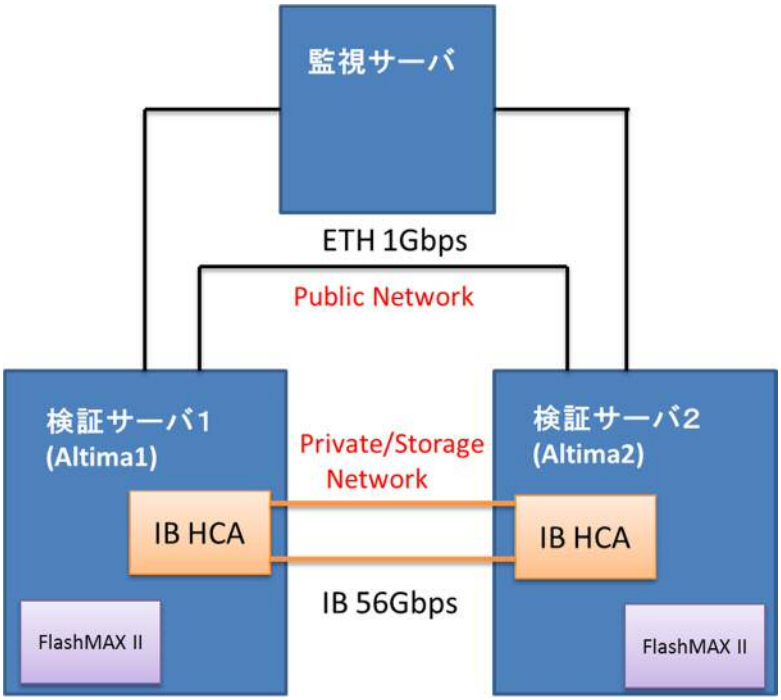


図 3　ネットワーク構成図

2)　フラッシュボリューム(Flash Volumes)構成

Altima1、Altima2 に搭載されている FlashMAX II (2.2 TB)デバイスに対して下記の 3 種類のボリュームを作成した。

サーバ	ボリューム名	容量（GB）	用途
Altima1	vshare-rac1-a	1	クラスタウェア
	vshare-rac2-a	500	データベース
	vshare-rac3-a	500	データベース
Altima2	vshare-rac1-b	1	クラスタウェア
	vshare-rac2-b	500	データベース
	vshare-rac3-b	500	データベース

3)　論理構成

Altima1、Altima2 で作成したフラッシュボリュームに対して、vShare 機能で各々のボリュームを互いに共有(Share)することで Active/Active(2-way mirroring)の高可用性を担保する構成をとる。

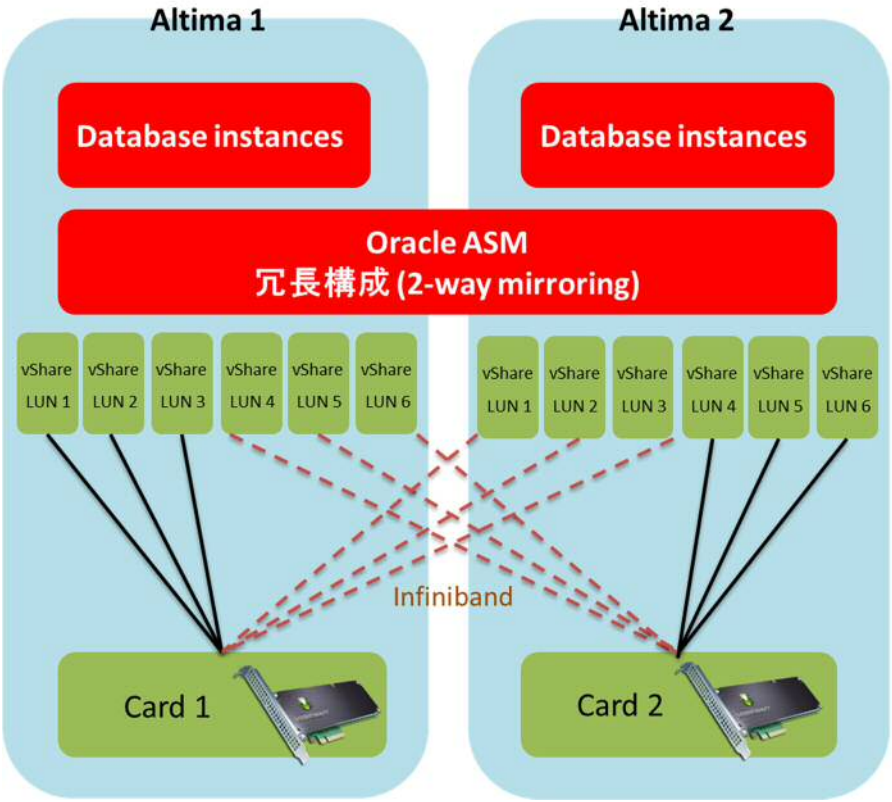


図 4　論理構成図

4)　ボリュームの作成と vShare の接続性を確認する。

今回はオートコンフィグコマンドを使用しコンフィグレーションファイルからボリュームの生成と vShare の接続を行い、正常に動作することを確認した。

```
[root@altima1 ~]# vgc-vshare-auto-config --configure Altima.conf
[root@altima1 ~]# vgc-vshare-monitor
vgc-vshare-monitor: FlashMAX Connect Software Suite 1.2(65244.V5A)
  ------------------------------------------------------------------------------------------------------------
  vShare Device              Local Device        Role        State        UUID
  ------------------------------------------------------------------------------------------------------------
  /dev/vshare-rac2-a         /dev/vgca0          target      Started      36c64553-9d0d-4591-a099-b3c7465d60c5
  /dev/vshare-rac3-a         /dev/vgca0          target      Started      3de14c36-ed89-42d2-b2c5-31073d6a5aa2
  /dev/vshare-rac1-a         /dev/vgca0          target      Started      fb7058b6-33b3-4f9f-b973-c0c9b3c1e2d3
  /dev/vshare-rac1-b         NA                  initiator   Connected    d5d153ea-4afe-46a9-a0e0-5413d2245b79
  /dev/vshare-rac2-b         NA                  initiator   Connected    9cb1407c-0418-44e9-ac96-257f9a6c36b8
  /dev/vshare-rac3-b         NA                  initiator   Connected    2fa4838f-11c7-4d47-ba80-33eb59d61117
```

```
[root@altima1 ~]# fdisk -l | grep /dev/vshare
Disk /dev/vshare-rac1-a: 1000 MB, 1000000000 bytes
Disk /dev/vshare-rac2-a: 500.0 GB, 500000000000 bytes
Disk /dev/vshare-rac3-a: 500.0 GB, 500000000000 bytes
Disk /dev/vshare-rac1-b: 1000 MB, 1000000000 bytes
Disk /dev/vshare-rac2-b: 500.0 GB, 500000000000 bytes
Disk /dev/vshare-rac3-b: 500.0 GB, 500000000000 bytes
```

5) Oracle Real Application Clusters 11gR2 のインストールと動作確認

Altima1、Altima2 に対して Oracle Clusterware 11gR2, Oracle Real Application Clusters 11gR2 をインストールし使用するディスクとして上記で作成されたフラッシュボリュームを適応した。フラッシュボリューム間の可用性については、Oracle ASM によるサーバ間ミラーにより冗長性を確保している[1]。また、ベンチマークツールとして DBMS_RESOURCE_MANAGER.CALIBRATE_IO プロシージャを使用して簡易的な性能測定を実施し、帯域・レイテンシの面で期待される結果になることを確認した。

6. まとめ

本検証で、富士通社製 PRIMERGY RX200S8 と HGST 社製 FlashMAX II 及び FlashMAX Connect の接続性と基本動作に問題がないことを確認することができました。また、Oracle Real Application Clusters 11gR2 をインストールしソフトウェア製品構成までを想定した検証を実施することで FlashMAX Connect の操作面、性能面の優位性を確認することができました。

性能面、機能面に関する詳細な情報に関しては下記のお問い合わせ先までご連絡いただけますと誠に幸いです。

7. 謝辞

本検証は、富士通様及び HGST 様のご協力により検証作業を完了させることができました。検証及び準備に際しましてご協力いただき誠にありがとうございました。

8. お問い合わせ先

株式会社アルティマ 清水 宏樹

Mail : virident-support@altima.co.jp

TEL : 045-476-2197

[1] 投票ディスクについてはオラクル社のホワイトペーパー「Oracle Clusterware 11g Release 2 (11.2) - Using standard NFS to support a third voting file for extended cluster configurations」に従って構成した。