



HGST社(旧Virident) Virident Solutions2.0のソフトウェア検証

2014年11月

株式会社アルティマ

▶ 検証目的

- ◆ 富士通社PRIMERGYサーバと、HGST社FlashMAXII(PCIe FlashSSD)とVirident Solutionsソフトウェアを使った機能の確認、及び性能検証を実施しました。

▶ 検証期間

- ◆ 2014年10月27日～11月7日

▶ 検証場所

- ◆ 富士通検証センター（東京・浜松町）

▶ 検証担当

- ◆ 株式会社HGSTジャパン 大黒義弘
- ◆ 株式会社アルティマ 北島佑樹

今回使用する製品について

▶ FlashMAX II



4800GB, 2200GB, 1100GB and 550GB
MLC | HH-HL | PCIe 2.0

HGST社が提供するFlashMAXIIはPCI Express直結の超高速ストレージ製品で、製品は550GB / 1.1TB / 2.2TB / 4.8TBのラインアップを揃えています。

ハイライト

- ・ 安定したアプリケーションパフォーマンスを提供
- ・ HH/HLサイズで大容量(最大4.8TB)
- ・ エンタープライズの信頼性とデータ保護(Hardware RAID)
- ・ インテリジェンスコントローラーによる最適化と安定した性能を実現

▶ Virident Solutions 2.0 ※ソフトウェア製品



HGST Virident Solutions 2.0はPCIeSSDデバイスのストレージファブリックを構築するソフトウェアソリューションです。超高速で柔軟なFlashベースのSAN/NAS環境を提供し、オンデマンドで高速アプリケーションの構築が可能となります。

ハイライト

- ・ Software Definedスケールアウトフラッシュストレージ
HA機能、キャッシュ機能、Sharedボリューム機能
- ・ 1クラスタ最大38TBのフラッシュボリューム管理
- ・ InfiniBand(RDMA対応), Ethernet対応
- ・ CLI/GUIインターフェースによる統合管理

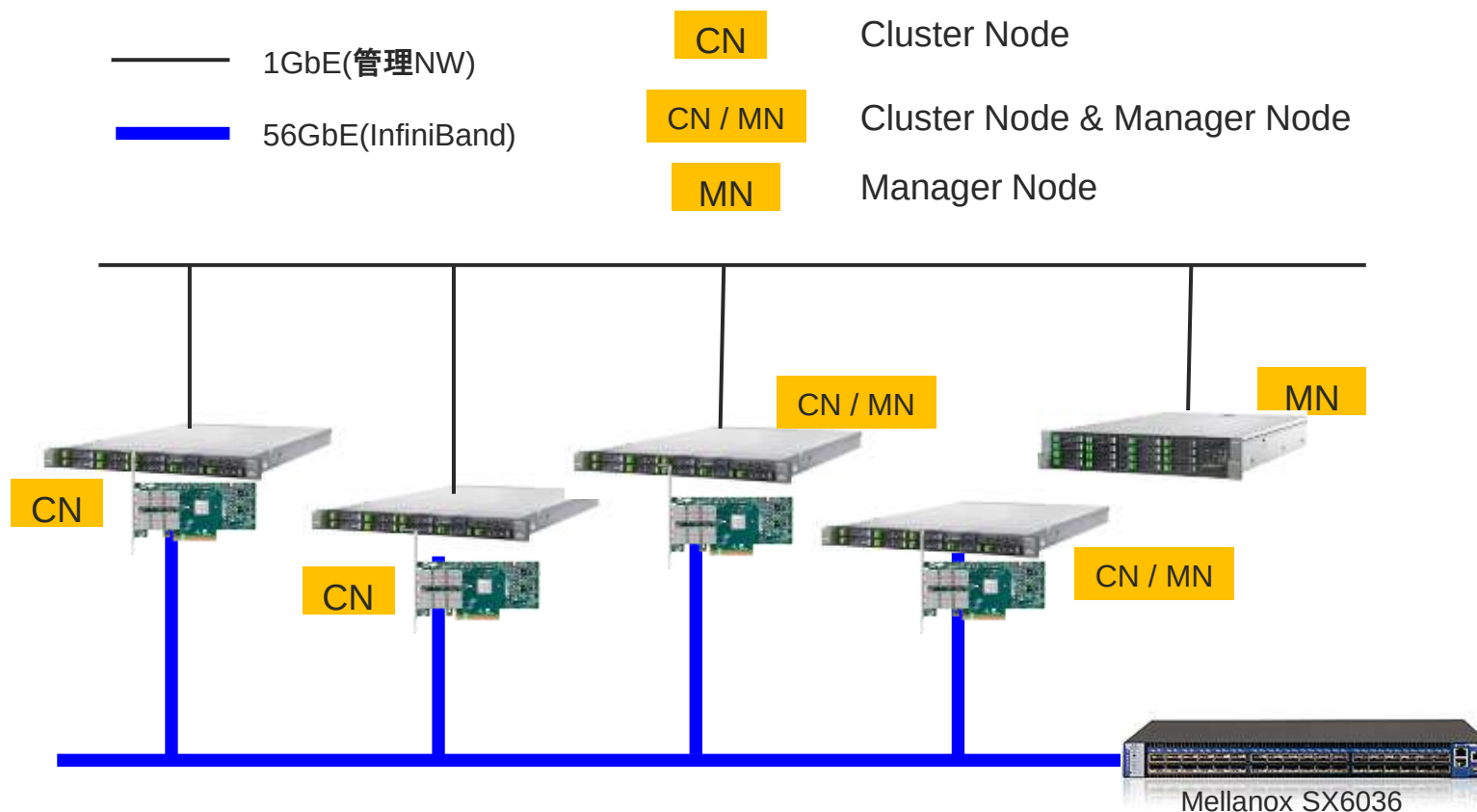
補足:FlashMAX II 性能データ

FlashMAX II Lineup (GB)	550, 1100	1100, 2200	4800
Read throughput (max MB/s, sequential 64K)	1,600	2,700	2,600
Write throughput (max MB/s, sequential 64K)	550	1,000	900
Read IOPS (max IOPS, random 4K)	174,000	345,000	269,000
Write IOPS (max IOPS, random 4K)	27,000	57,000	51,000
Peak write IOPS (max IOPS, random 4k)	109,000	245,000	213,000
Mixed IOPS (70/30 R/W, random 4k)	72,000	138,000	128,000
Peak mixed IOPS (70/30 R/W, random 4k)	161,000	315,000	264,000
Read IOPS (max IOPS, random 8k)	125,000	250,000	214,000
Write IOPS (max IOPS, random 8k)	13,000	28,000	27,000
Latency 512B (μs)	21	22	19

- ▶ Virident Solutions2.0機能検証
 - ◆ ソフトウェアインストール
 - ◆ ボリューム作成/追加/削除
 - ◆ ノード追加/削除

- ▶ Virident Solutions2.0性能検証
 - ◆ ローカル性能
 - ◆ 複数ノード性能

HW	Server	Type	CPU	MEM	PCIe
		PRIMERGY RX200 S8 x 4	Xeon E5-2697v2 2.70GHz 12core x 2	32GB	Gen3
		PRIMERGY RX300 S7 x 1	Xeon E5-2697v2 2.70GHz 12core x 2	32GB	Gen3
	PCIe SSD	Flash MAX II			
	InfiniBand HCA	Mellanox ConnectX3 FDR(56Gbps) Dual-port (PY-HC301)			
	InfiniBand Switch	Mellanox SX6036 36port 56Gbps			
SW	Virident Solutions 2.0	version 2.0			
	Mellanox OFED Driver	MLNX_OFED_LINUX-2.2-1.0.1-rhel6.4-x86_64			
	OS	Red Hat Enterprise Linux 6.4 (for Intel64) Kernel : 2.6.32-358.el6			



- クラスターノード 4ノード構成
- クラスターマネージャーノード 3ノード構成(冗長 : Act / Stb)
- クラスター間通信はInfiniBand(RDMA)を使用
- マネジメントノード間は1Gbps専用NW

Virident Solutions 2.0について

Virident Solutions 2.0概要



▶ 製品名

- ◆ HGST Virident Solutions 2.0

▶ 製品概要

- ◆ フラッシュベースのストレージファブリック (Software Defined Storage) を提供するソフトウェア
- ◆ 性能、高可用性、耐障害性、柔軟性を提供するソフトウェアベースのソリューション

▶ サポートデバイス

- ◆ HGST社製(旧Virident) FlashMAX II
- ◆ FlashMAX III サポート予定

▶ サポートOS

- ◆ RHEL6.x
- ◆ OEL6.x

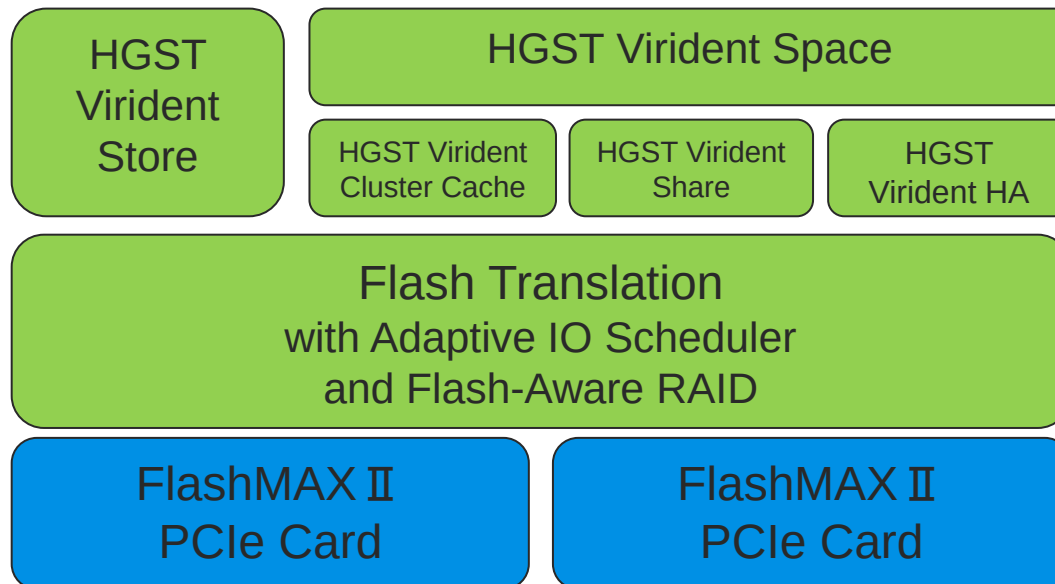
▶ プロトコル

- ◆ InfiniBand(RDMA)サポート
- ◆ Ethernetサポート

Model	PCI Express	Airflow	Processor	System memory per card (vCache)	System memory per card (vHA/vShare/vStore)
FlashMAX II 550 GB	Gen 2 or 3 x8 electrical	150 LFM @ 45 C inlet	2 or 4 Socket, Multicore Intel Xeon® or AMD Opteron	3 GB	1.5 GB
FlashMAX II 1.1 TB Standard	Gen 2 or 3 x8 electrical	150 LFM @ 45 C inlet	2 or 4 Socket, Multicore Intel Xeon® or AMD Opteron	6 GB	2.5 GB
FlashMAX II 1.1 TB Performance	Gen 2 or 3 x8 electrical	200 LFM @ 45 C inlet	2 or 4 Socket, Multicore Intel Xeon® or AMD Opteron	6 GB	2.5 GB
FlashMAX II 2.2 TB	Gen 2 or 3 x8 electrical	200 LFM @ 45 C inlet	2 or 4 Socket, Multicore Intel Xeon® or AMD Opteron	12 GB	6 GB
FlashMAX II 4.8 TB	Gen 2 x8 electrical	200 LFM @ 45 C inlet	2 or 4 Socket, Multicore Intel Xeon® or AMD Opteron	10 GB	10 GB

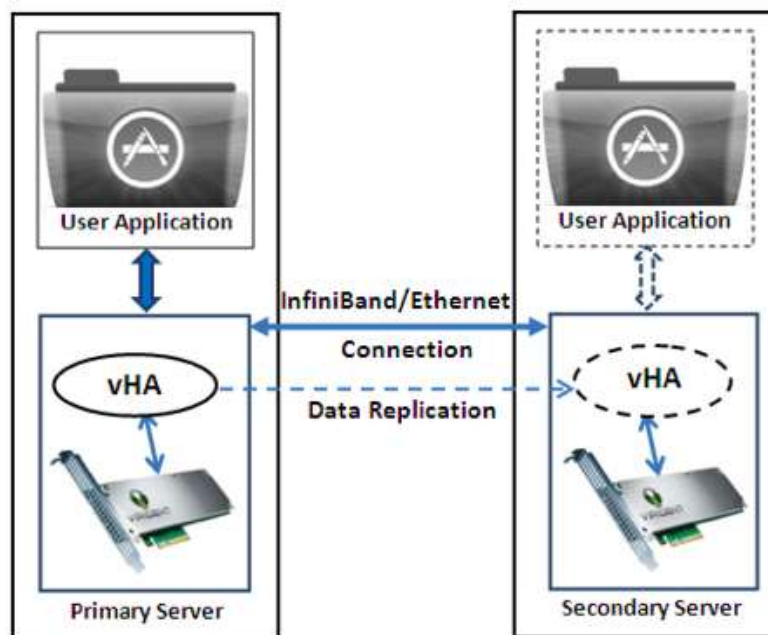
Virident Solutions

ソフトウェアアーキテクチャ



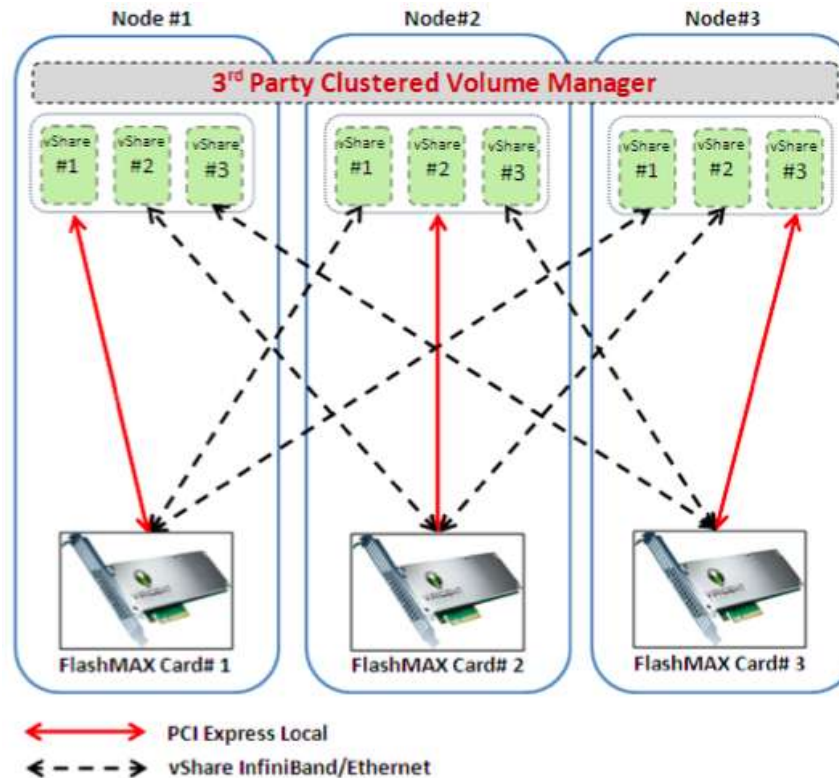
Virident Solutions2.0は上図の要素から構成され、
以下のような機能を提供している

- HA, Cluster Cache, Space *詳細後述
- 複数ノード間のIOスケジューリング
- データ保護のためのFlash-Aware RAID構成
- Fail over / Fail back



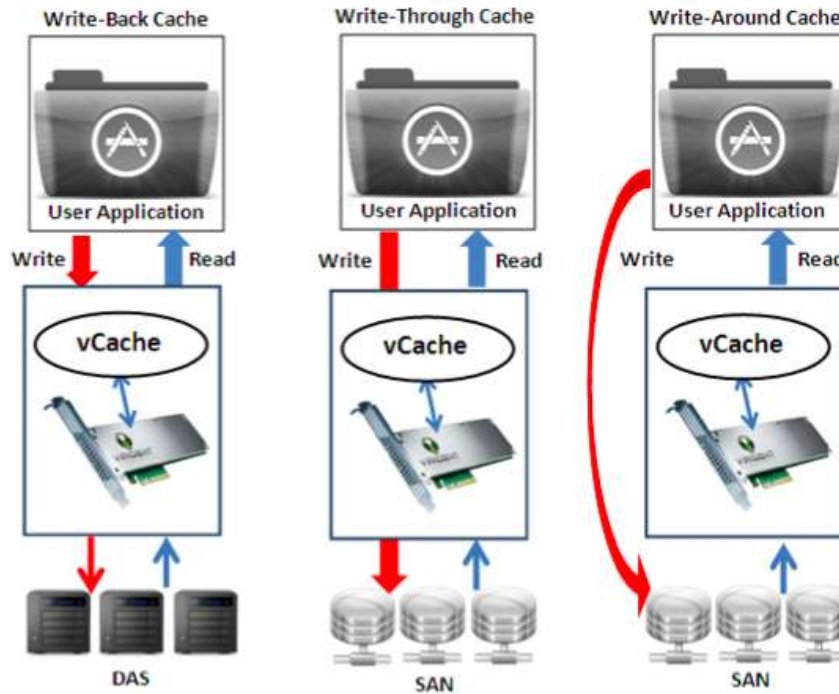
- ノード間のレプリケーション機能
- Active / Standbyで動作し、障害時は自動切り替わり
- ネットワークプロトコルでEthernetとInfiniBand(RDMA)をサポート
 - InfiniBandの場合はローカルに近い性能を得られるのが特徴

HGST Virident Share

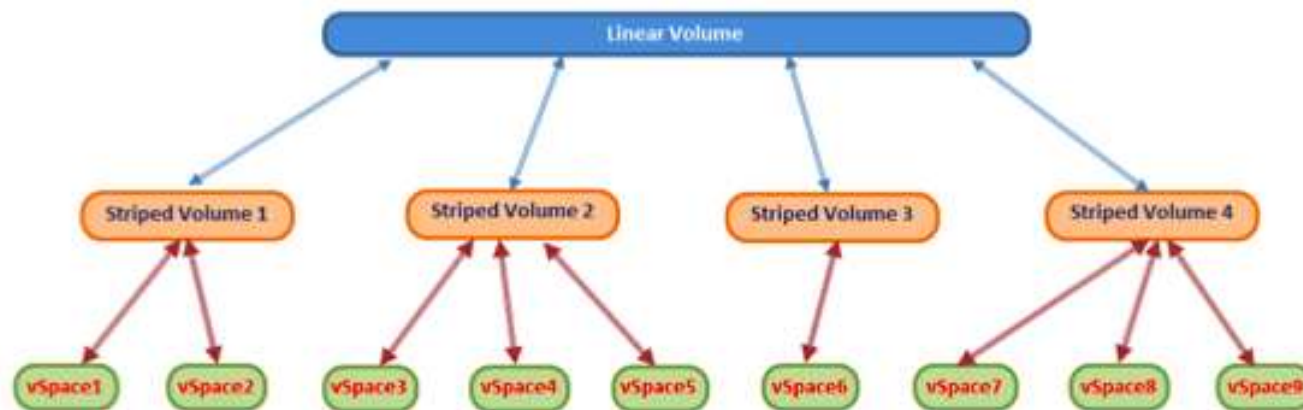


- 複数ノード間でのボリューム共有
- マルチパスのサポート
- InfiniBandサポートでローカルレベルのリモートアクセス性能を実現

HGST Virident Cluster Cache



- キャッシュ機能(Local, Remote対応)
- Write-back, Write-through, Write-aroundの3タイプから選択可能



- HA, Share, Cluster Cache機能の統合管理
- 複数ノード間でのリソース管理
 - ボリューム作成、追加、削除
 - ノード追加、削除
 - IO負荷分散
- 必要なリソース(Block-level volume, cache volume)をダイナミックに提供可能

Virident Solutions 2.0機能検証

▶ デバイスの確認

```
# lspci -d 1a78:  
05:00.0 FLASH memory: Virident Systems Inc. Device 0040 (rev 01)
```

▶ ドライバインストール

- (管理ノードの場合)vgc-clustermgr-2.0.CS-75617.V6B.x86_64.rpm
- vgc-clusternode-2.0.CS-75617.V6B.x86_64.rpm
- vgc-rdma-2.6.32-358.el6.x86_64-2.0.CS-75617.V6B.x86_64.rpm
- vgc-rdma-3.8.13-16.2.1.el6uek.x86_64-2.0.CS-75617.V6B.x86_64.rpm
- vgc-rdma-utils-redhat6-2.0.CS-75617.V6B.x86_64.rpm
- vgc-utils-redhat6-2.0.CS-75617.V6B.x86_64.rpm

```
]# rpm -ihv [ドライバrpm]
```

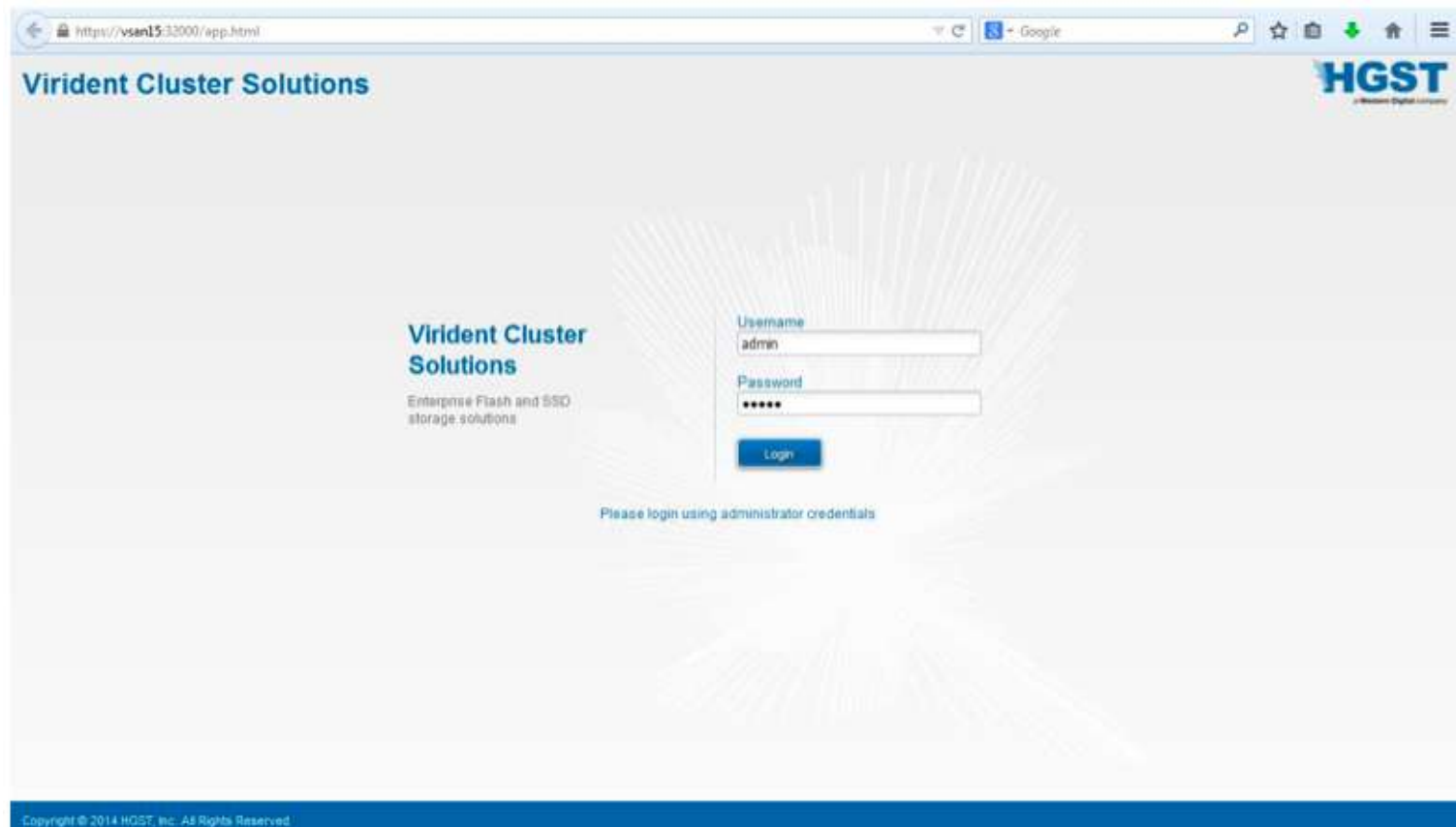
▶ サービス起動

```
# /etc/init.d/vgcd start  
Loading kernel modules...           [ OK ]  
Rescanning SW RAID volumes...       [ OK ]  
Rescanning LVM volumes...           [ OK ]  
Enabling swap devices...            [ OK ]  
Rescanning mount points...          [ OK ]
```

GUIコンソール(クラスター管理画面)

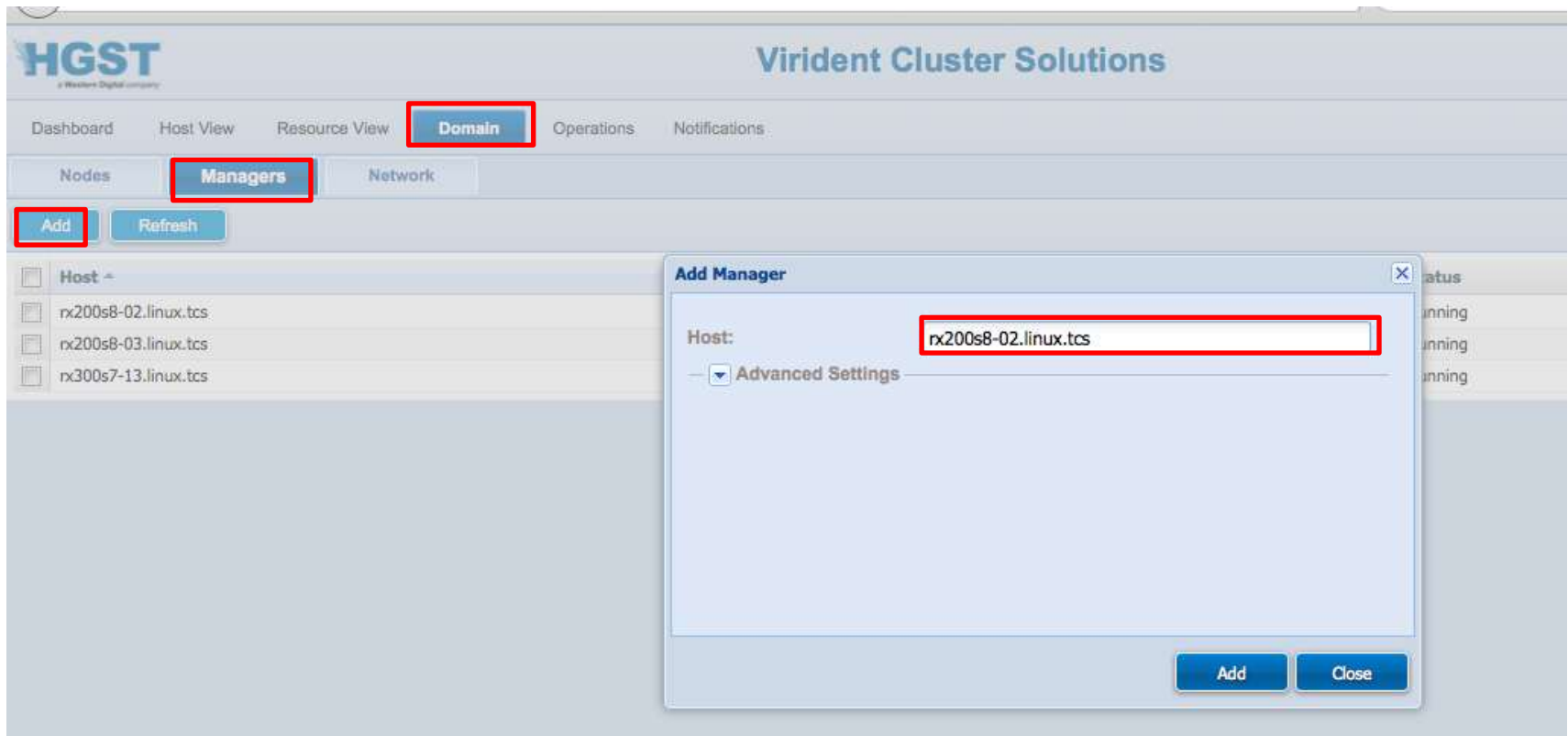
▶ GUIコンソールへのアクセス

https:// 管理ノードIP:32000



ノード追加1

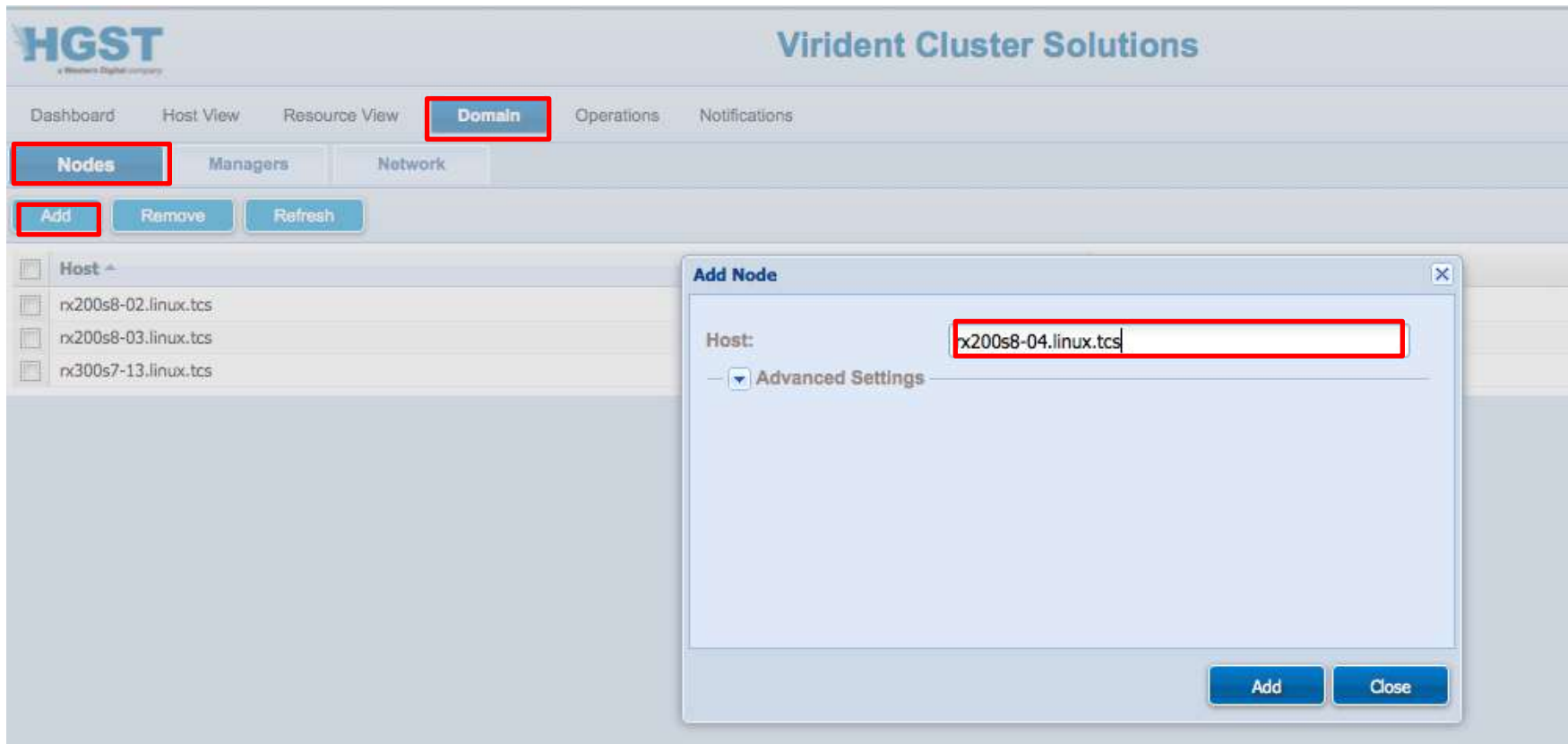
- ▶ マネージャーノードを追加できることを確認
- ▶ 「Domain」 → 「Managers」 → 「Add」
 - ◆ 追加するマネージャーノードのホストドメイン名を指定して追加



The screenshot displays the HGST Virident Cluster Solutions web interface. The top navigation bar includes 'Dashboard', 'Host View', 'Resource View', 'Domain', 'Operations', and 'Notifications'. The 'Domain' tab is selected and highlighted with a red box. Below this, the 'Managers' sub-tab is also selected and highlighted with a red box. On the left side, there are 'Add' and 'Refresh' buttons, with the 'Add' button highlighted by a red box. A list of hosts is visible, including 'rx200s8-02.linux.tcs', 'rx200s8-03.linux.tcs', and 'rx300s7-13.linux.tcs'. An 'Add Manager' dialog box is open in the foreground, featuring a 'Host:' label and a text input field containing 'rx200s8-02.linux.tcs', which is highlighted with a red box. Below the input field is an 'Advanced Settings' section with a dropdown arrow. At the bottom of the dialog are 'Add' and 'Close' buttons.

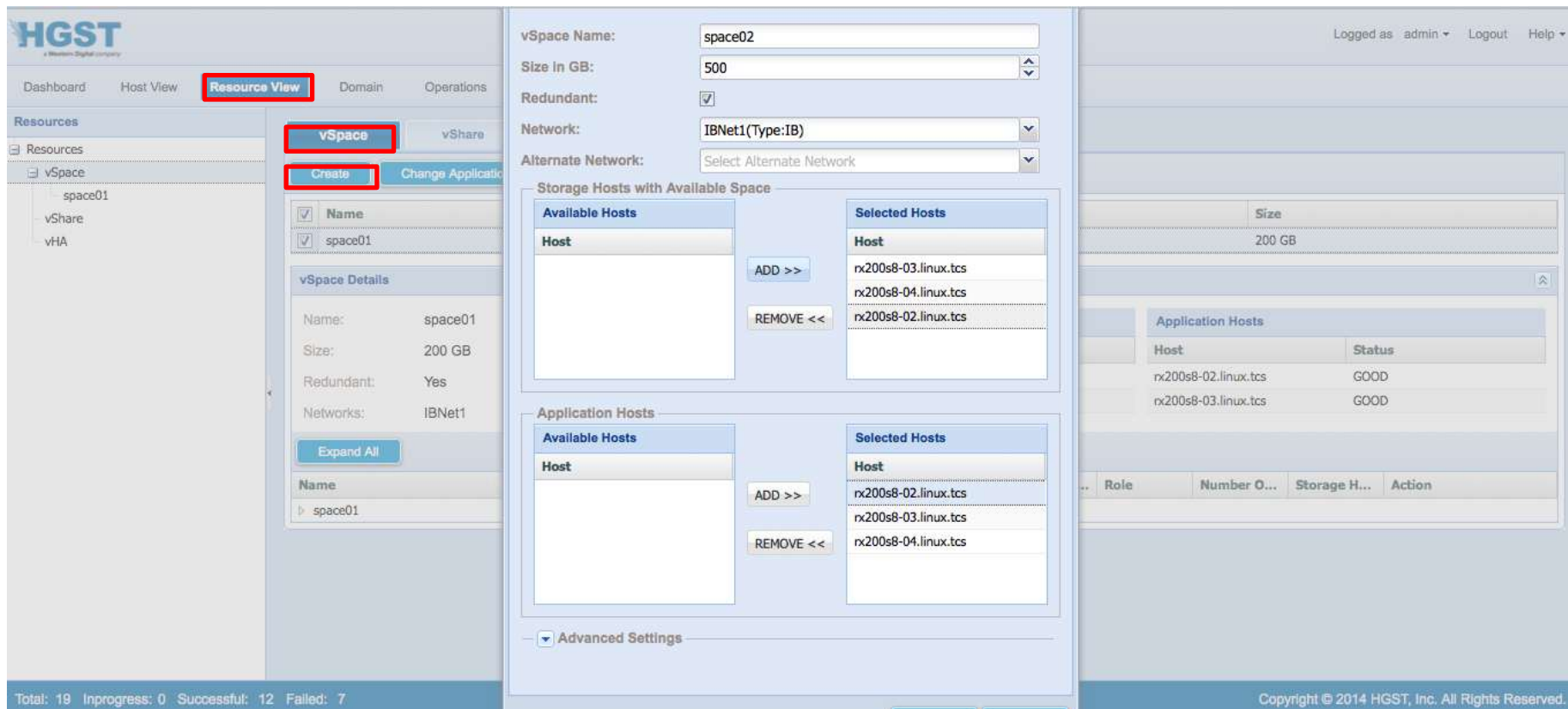
ノード追加2

- ▶ クラスターノードを追加できることを確認
- ▶ 「Domain」 → 「Nodes」 → 「Add」
 - ◆ 追加するクラスターノードのホストドメイン名を指定して追加



The screenshot displays the HGST Virident Cluster Solutions web interface. The top navigation bar includes 'Dashboard', 'Host View', 'Resource View', 'Domain' (highlighted with a red box), 'Operations', and 'Notifications'. Below this, the 'Nodes' tab is selected and highlighted with a red box. Under the 'Nodes' tab, there are three buttons: 'Add' (highlighted with a red box), 'Remove', and 'Refresh'. On the left side, a list of hosts is shown under the 'Host' header, including 'rx200s8-02.linux.tcs', 'rx200s8-03.linux.tcs', and 'rx300s7-13.linux.tcs'. A modal dialog box titled 'Add Node' is open in the foreground. It contains a 'Host:' label and a text input field where 'rx200s8-04.linux.tcs' has been entered (highlighted with a red box). Below the input field is a section for 'Advanced Settings' with a dropdown arrow. At the bottom right of the dialog are 'Add' and 'Close' buttons.

- ▶ ボリューム(vSpace)作成ができることを確認
 - ◆ 「Resource View」 「vSpace」 「Create」
 - ◆ vSpace名、ストレージ容量、ネットワークタイプ、対象クラスターノード、アプリケーションノードを指定



The screenshot displays the HGST vSpace creation interface. The 'Resource View' tab is selected, showing a list of resources on the left. The 'vSpace' section is highlighted, and the 'Create' button is visible. The 'vSpace Details' section shows fields for Name, Size, Redundant, and Networks. The 'Storage Hosts with Available Space' section shows a list of available hosts and selected hosts. The 'Application Hosts' section shows a list of available hosts and selected hosts. The 'Advanced Settings' section is expanded.

vSpace Name: space02
Size in GB: 500
Redundant: ☒
Network: IBNet1(Type:IB)
Alternate Network: Select Alternate Network

Storage Hosts with Available Space

Available Hosts	Selected Hosts
Host	Host
	rx200s8-03.linux.tcs
	rx200s8-04.linux.tcs
	rx200s8-02.linux.tcs

Application Hosts

Available Hosts	Selected Hosts
Host	Host
	rx200s8-02.linux.tcs
	rx200s8-03.linux.tcs
	rx200s8-04.linux.tcs

Application Hosts

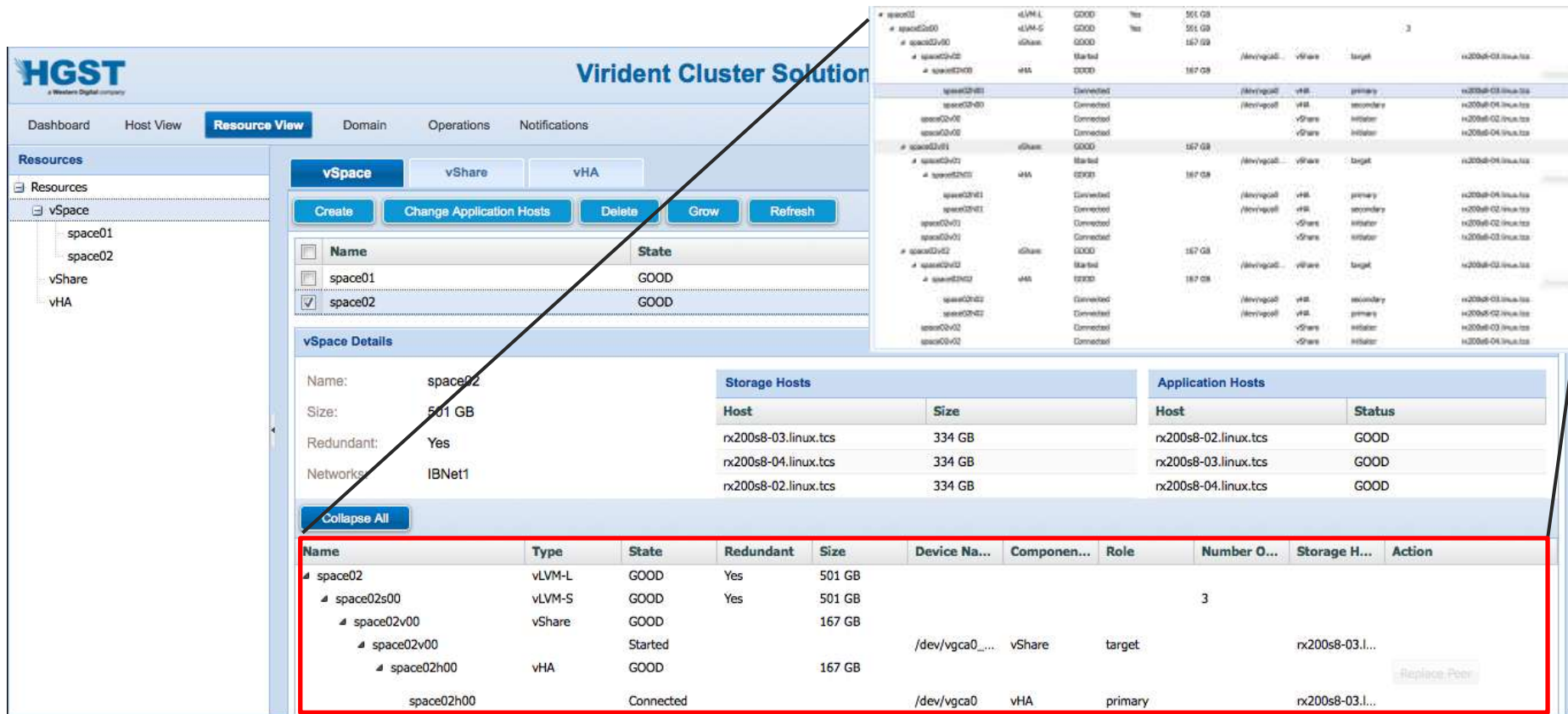
Host	Status
rx200s8-02.linux.tcs	GOOD
rx200s8-03.linux.tcs	GOOD

Total: 19 Inprogress: 0 Successful: 12 Failed: 7

Copyright © 2014 HGST, Inc. All Rights Reserved.

ボリューム作成

- ▶ vHA, vShare, vLVM(Space)によりストレージリソースが分散していることが確認できる



Resources

- Resources
 - vSpace
 - space01
 - space02
 - vShare
 - vHA

vSpace

Create Change Application Hosts Delete Grow Refresh

Name	State
space01	GOOD
space02	GOOD

vSpace Details

Name: space02
Size: 501 GB
Redundant: Yes
Networks: IBNet1

Storage Hosts

Host	Size
rx200s8-03.linux.tcs	334 GB
rx200s8-04.linux.tcs	334 GB
rx200s8-02.linux.tcs	334 GB

Application Hosts

Host	Status
rx200s8-02.linux.tcs	GOOD
rx200s8-03.linux.tcs	GOOD
rx200s8-04.linux.tcs	GOOD

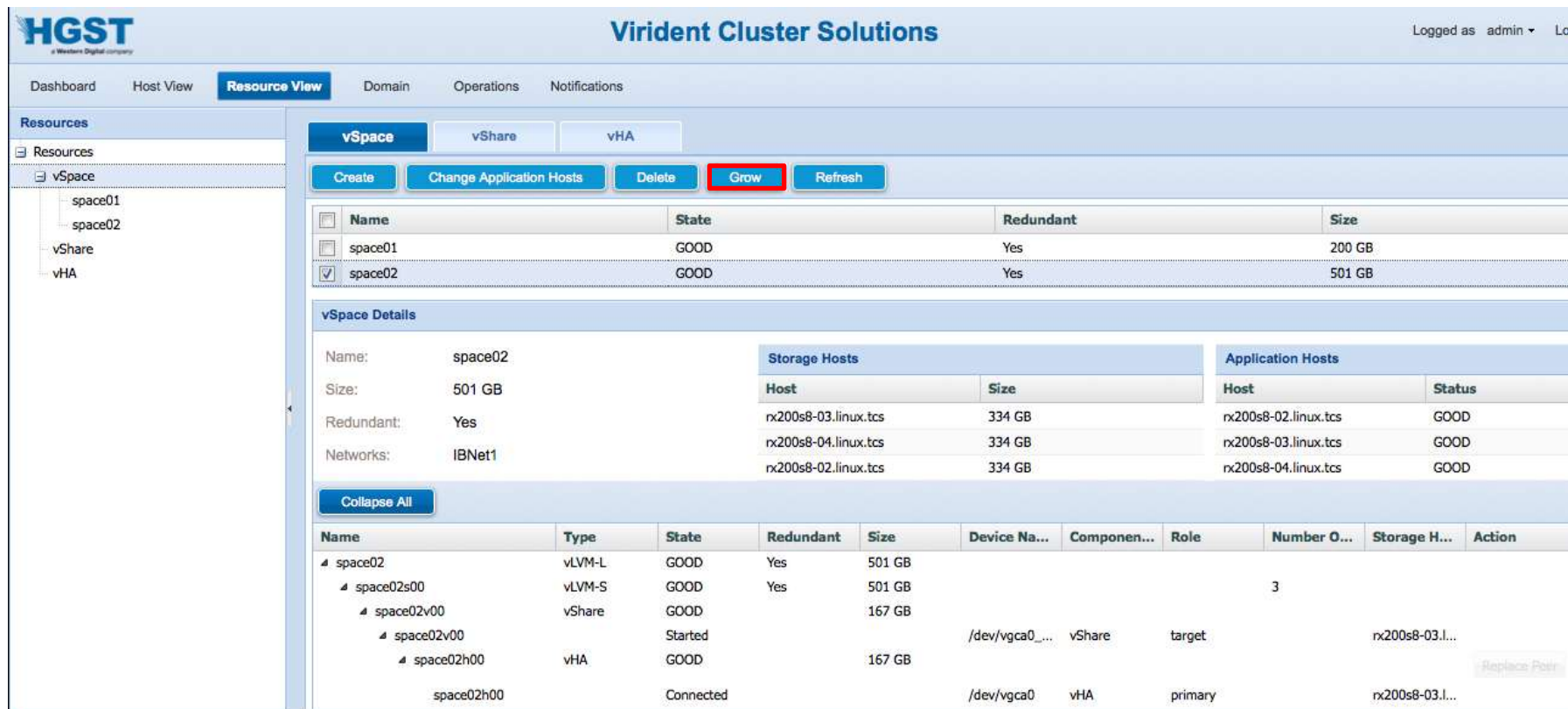
Name	Type	State	Redundant	Size	Device Na...	Componen...	Role	Number O...	Storage H...	Action
space02	vLVM-L	GOOD	Yes	501 GB						
space02s00	vLVM-S	GOOD	Yes	501 GB				3		
space02v00	vShare	GOOD		167 GB						
space02v00		Started			/dev/vgca0...	vShare	target		rx200s8-03.l...	
space02h00	vHA	GOOD		167 GB						
space02h00		Connected			/dev/vgca0	vHA	primary		rx200s8-03.l...	Replace Peer

Total: 26 Inprogress: 0 Successful: 19 Failed: 7

Copyright © 2014 HGST, Inc. All Rights Reserved.

ボリューム追加

- ▶ ボリューム(vSpace)を追加できることを確認
 - ◆ 「Resource View」 「vSpace」 「Grow」
 - ◆ 追加するストレージ容量を指定



The screenshot displays the HGST Virident Cluster Solutions management interface. The 'Resource View' tab is active, showing a list of resources on the left and a detailed view of the selected 'vSpace' (space02) on the right. The 'vSpace' details include its name, size (501 GB), redundant status (Yes), and network (IBNet1). Below this, a table shows the storage hosts and application hosts for the vSpace. The bottom section shows a hierarchical tree of the vSpace components, including space02s00, space02v00, and space02h00.

Resources

- Resources
 - vSpace
 - space01
 - space02
 - vShare
 - vHA

vSpace

Create Change Application Hosts Delete **Grow** Refresh

Name	State	Redundant	Size
space01	GOOD	Yes	200 GB
space02	GOOD	Yes	501 GB

vSpace Details

Name: space02
Size: 501 GB
Redundant: Yes
Networks: IBNet1

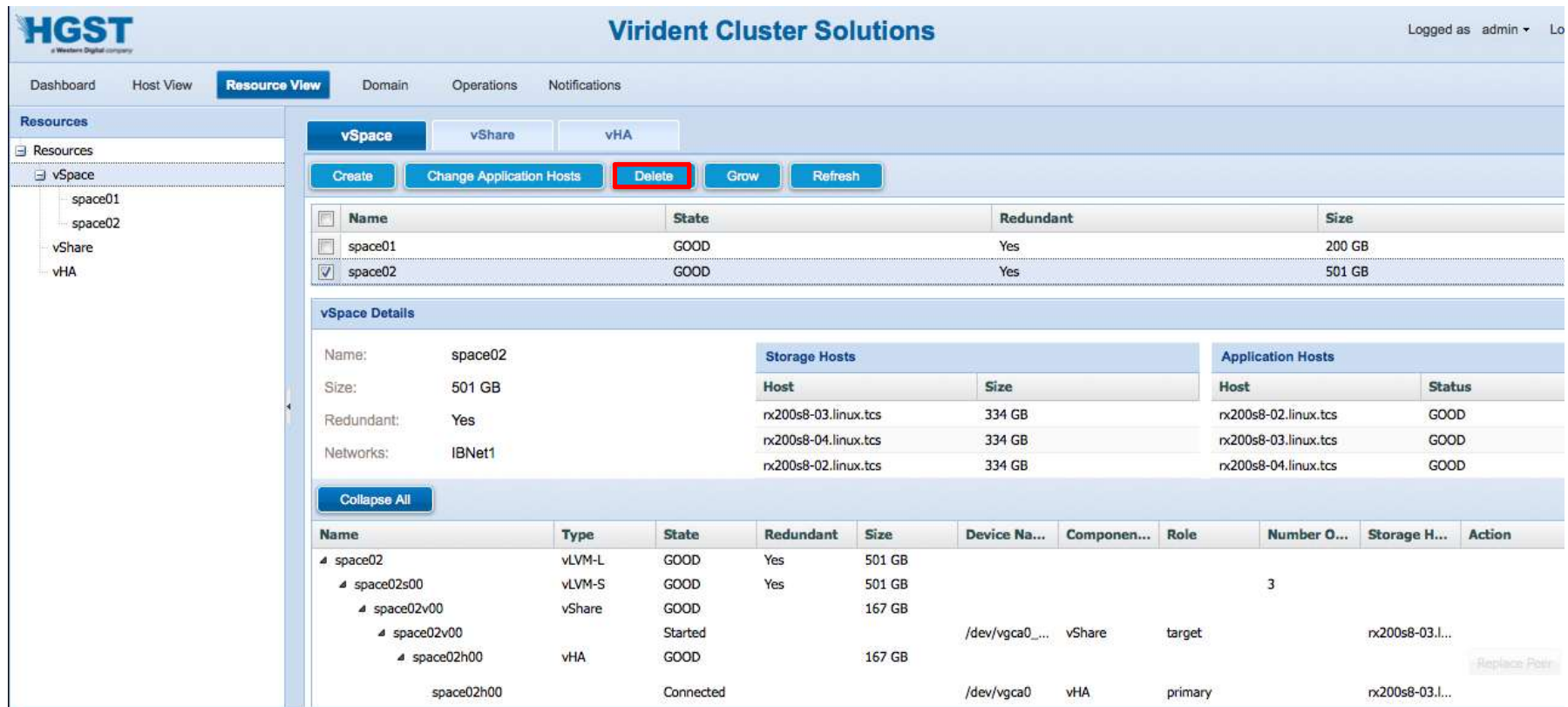
Storage Hosts		Application Hosts	
Host	Size	Host	Status
rx200s8-03.linux.tcs	334 GB	rx200s8-02.linux.tcs	GOOD
rx200s8-04.linux.tcs	334 GB	rx200s8-03.linux.tcs	GOOD
rx200s8-02.linux.tcs	334 GB	rx200s8-04.linux.tcs	GOOD

Collapse All

Name	Type	State	Redundant	Size	Device Na...	Componen...	Role	Number O...	Storage H...	Action
space02	vLVM-L	GOOD	Yes	501 GB						
space02s00	vLVM-S	GOOD	Yes	501 GB				3		
space02v00	vShare	GOOD		167 GB						
space02v00		Started			/dev/vgca0_...	vShare	target		rx200s8-03.l...	
space02h00	vHA	GOOD		167 GB						
space02h00		Connected			/dev/vgca0	vHA	primary		rx200s8-03.l...	Replace Pair

ボリューム削除

- ▶ ボリューム(vSpace)を削除できることを確認
 - ◆ 「Resource View」 「vSpace」 「Delete」



The screenshot displays the HGST Virident Cluster Solutions management interface. The 'Resource View' tab is active, showing a tree of resources on the left and a detailed view of the selected 'space02' vSpace on the right.

Resources Tree:

- Resources
 - vSpace
 - space01
 - space02
 - vShare
 - vHA

vSpace Management Buttons: Create, Change Application Hosts, **Delete** (highlighted), Grow, Refresh.

vSpace Details for space02:

- Name: space02
- Size: 501 GB
- Redundant: Yes
- Networks: IBNet1

Storage Hosts:

Host	Size
rx200s8-03.linux.tcs	334 GB
rx200s8-04.linux.tcs	334 GB
rx200s8-02.linux.tcs	334 GB

Application Hosts:

Host	Status
rx200s8-02.linux.tcs	GOOD
rx200s8-03.linux.tcs	GOOD
rx200s8-04.linux.tcs	GOOD

Volume Details Table:

Name	Type	State	Redundant	Size	Device Na...	Componen...	Role	Number O...	Storage H...	Action
space02	vLVM-L	GOOD	Yes	501 GB						
space02s00	vLVM-S	GOOD	Yes	501 GB				3		
space02v00	vShare	GOOD		167 GB						
space02v00		Started			/dev/vgca0_...	vShare	target		rx200s8-03.l...	
space02h00	vHA	GOOD		167 GB						
space02h00		Connected			/dev/vgca0	vHA	primary		rx200s8-03.l...	Replace Pair

アプリケーションノードから見たボリューム

- ▶ 通常のローカルブロックデバイスとして扱うことができる
- ▶ Example
 - ◆ space01, space02という名前でそれぞれボリューム作成した場合、アプリケーションノードからローカルのブロックデバイスとして見える
 - ◆ アプリケーションサーバ毎にボリュームへのアクセス権を指定することが可能

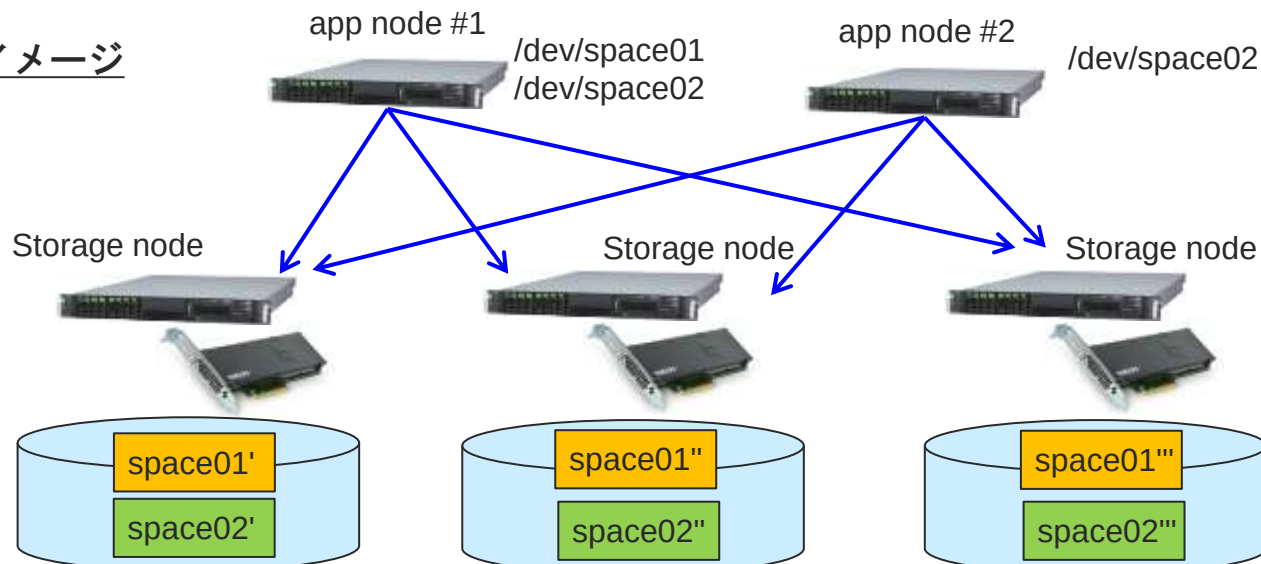
app node #1

```
# ls /dev/space*  
space01  
space02  
.....
```

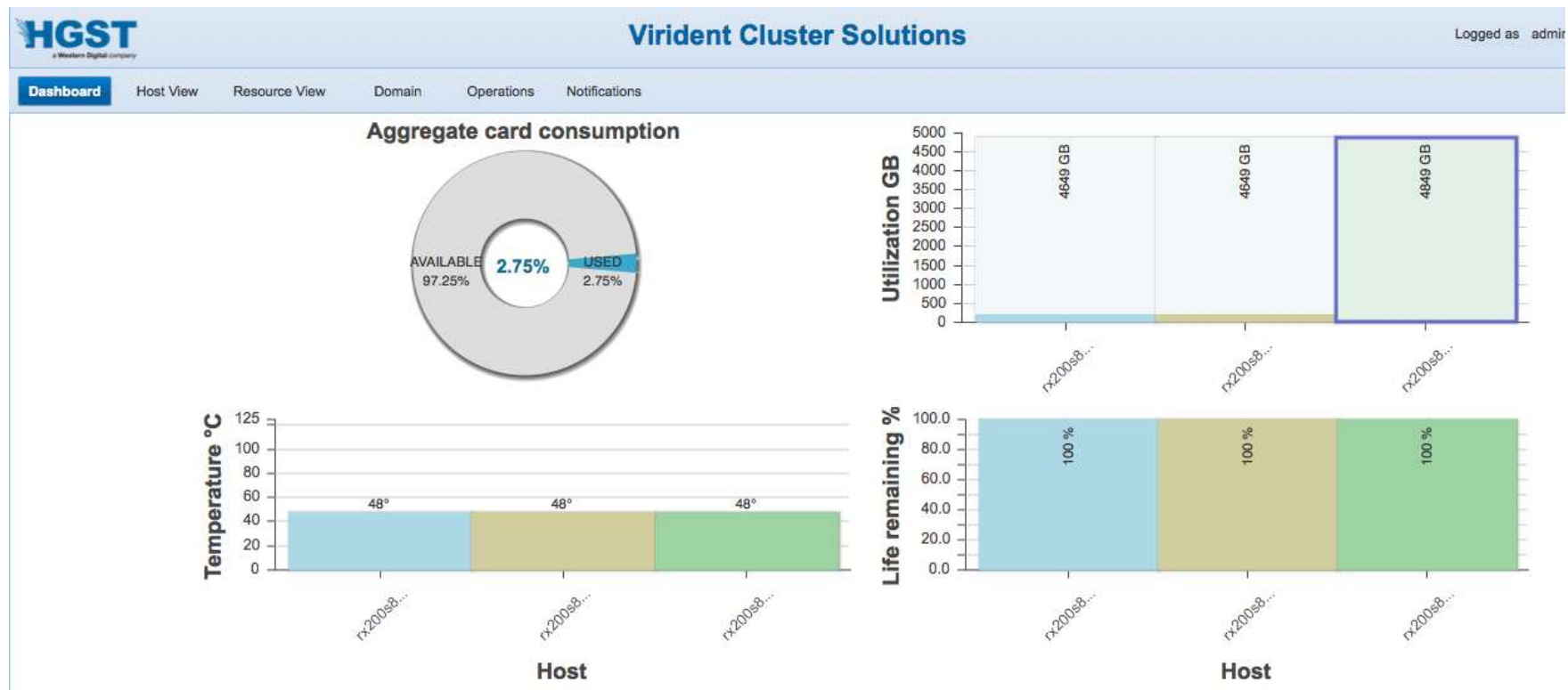
app node #2

```
# ls /dev/space*  
space02  
.....
```

リソースイメージ



- ▶ リソース状況の確認
 - ◆ 各ホストのストレージ容量確認
 - ◆ 全体の稼働率確認



- ▶ ホスト毎のリソース詳細確認
 - ◆ カード情報, vHA, vShare, vCache, vStore


Virident Cluster Solutions
Logged as admin ▾ Logout Help ▾

Dashboard
Host View
Resource View
Domain
Operations
Notifications

Hosts

- Hosts
- rx200s8-02.linux.tcs
- rx200s8-03.linux.tcs
- rx200s8-04.linux.tcs
- rx300s7-13.linux.tcs

Summary
Cards
vShare
vHA
vCache
vStore
Packages

Summary of host rx200s8-04.linux.tcs

Refresh

Host: rx200s8-04.linux.tcs

Host Address: 10.40.15.14

Operating System: Linux x64

OS Version: 2.6.32-358.el6.x86_64

Processor: Intel(R) Xeon(R) CPU E5-2697 v2 @ 2.70GHz

Memory: 31GB

Node Agent Status: Online

Driver up Since: 2014/10/27 14:35:31

vgdrive Driver Loaded: Yes

vHA Driver Loaded: Yes

vCache Driver Loaded: Yes

vSpace Driver Loaded: Yes

IP Data Provider Loaded: Yes

IB Data Provider Loaded: Yes

FMC Version: Virident Cluster Solutions 2.0.75617.V6B

Namespaces

Namespace ^	Device	Size	Role	State	Type
space02h00	/dev/vgca0	167 GB	secondary	Connected	vHA
space02h01	/dev/vgca0	167 GB	primary	Connected	vHA
space02v00		167 GB	initiator	Connected	vShare
space02v01	/dev/vgca0_space02h01	167 GB	target	Started	vShare
space02v02		167 GB	initiator	Connected	vShare

Virident Solutions 2.0性能検証

▶ テストケース

- ◆ case1 : ローカル性能
- ◆ case2 : 1アプリケーションノード + 2ストレージノード
- ◆ case3 : 2ストレージノード
- ◆ case4 : 3ストレージノード
- ◆ case5 : 4ストレージノード

※RDMAとIPoIB(TCP/IPベース)の比較

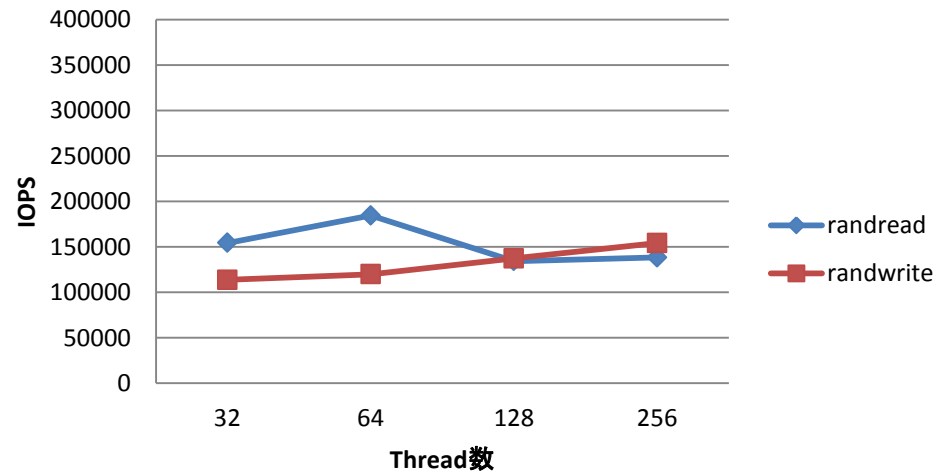
▶ ベンチマークツール

- ◆ fio ver 2.1.7
- ◆ パラメータ
 - block size : 4KB
 - number of jobs : 32,64,128,256
 - Queue depth : 1
 - Random read 100% and Random write 100%

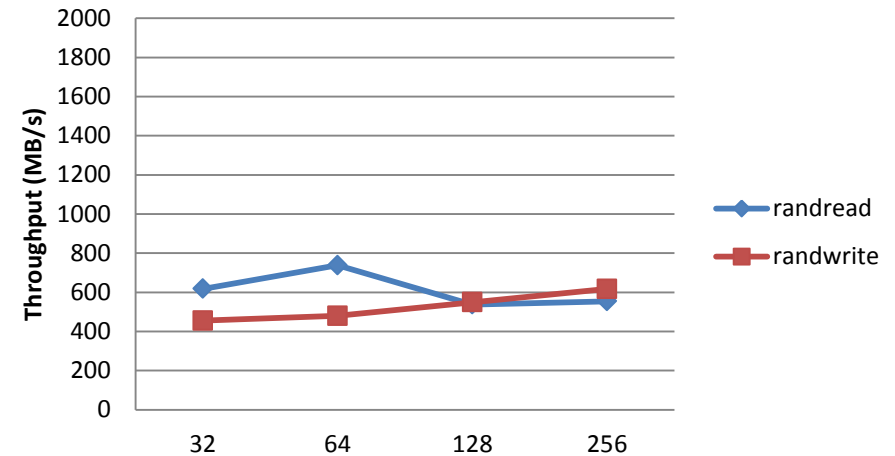
case1 : ローカル性能



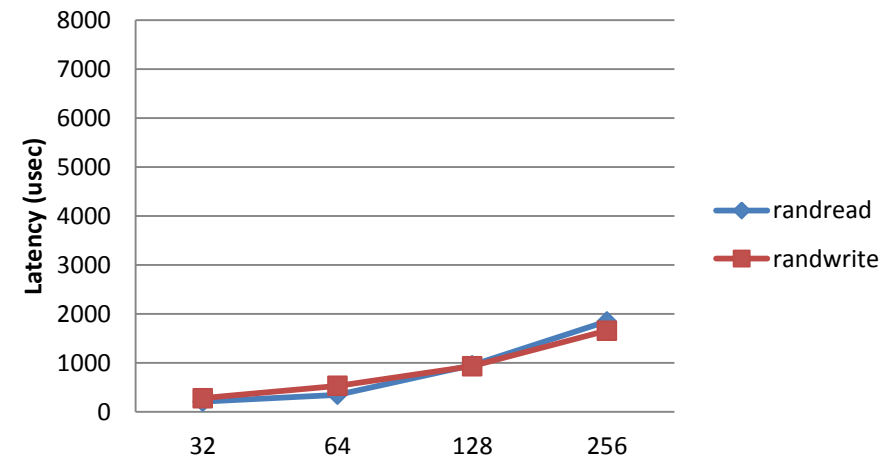
IOPS



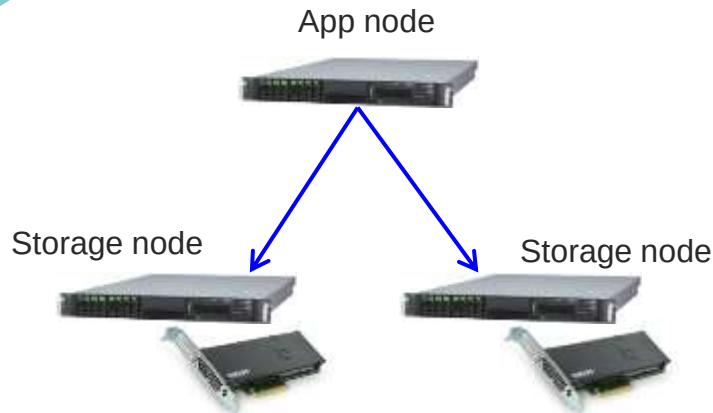
スループット



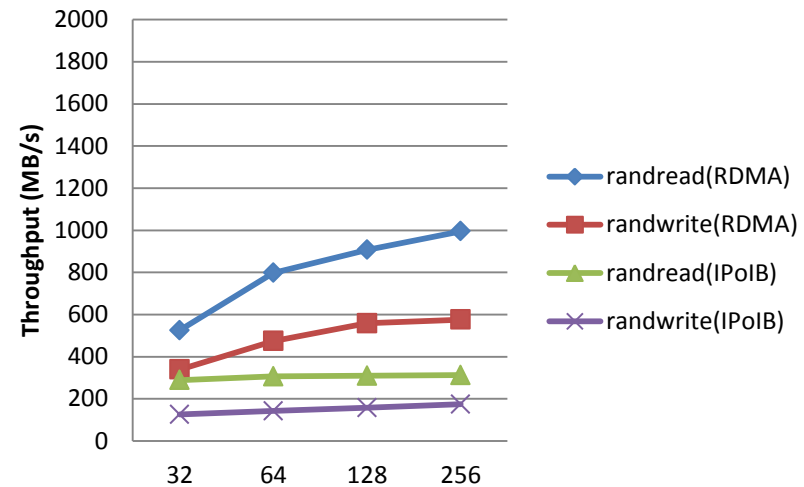
レイテンシー



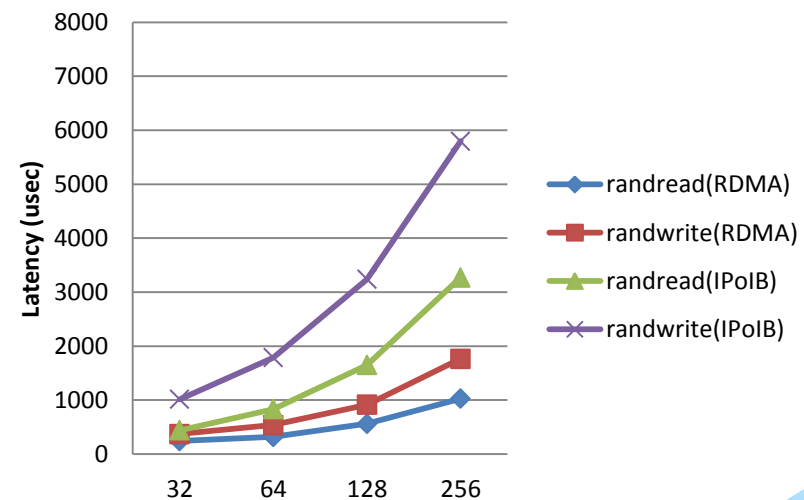
case2 : 1アプリケーションノード + 2ストレージノード



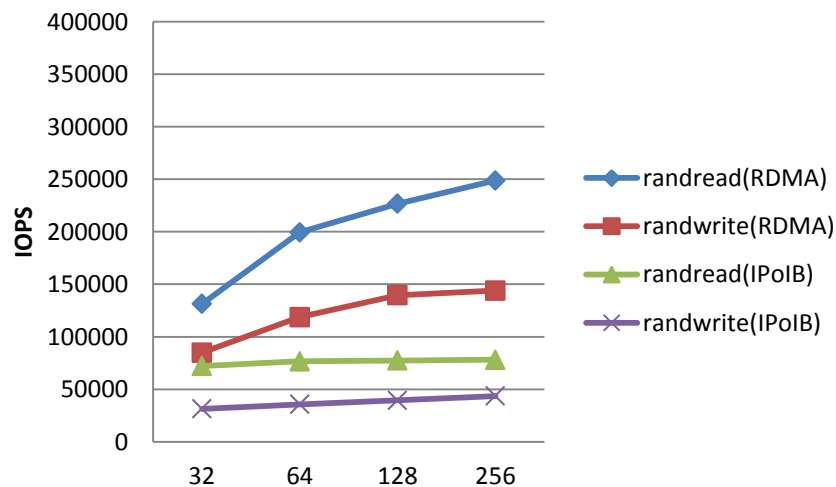
スループット



レイテンシー



IOPS



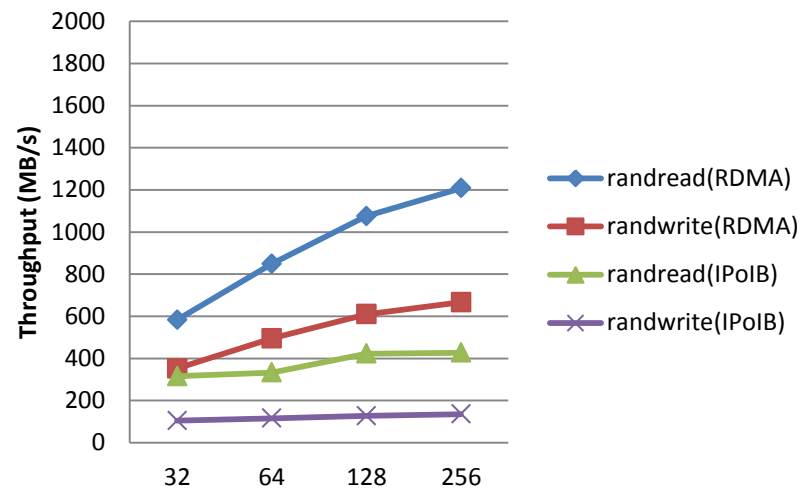
case3 : 2ストレージノード

Storage node & app node

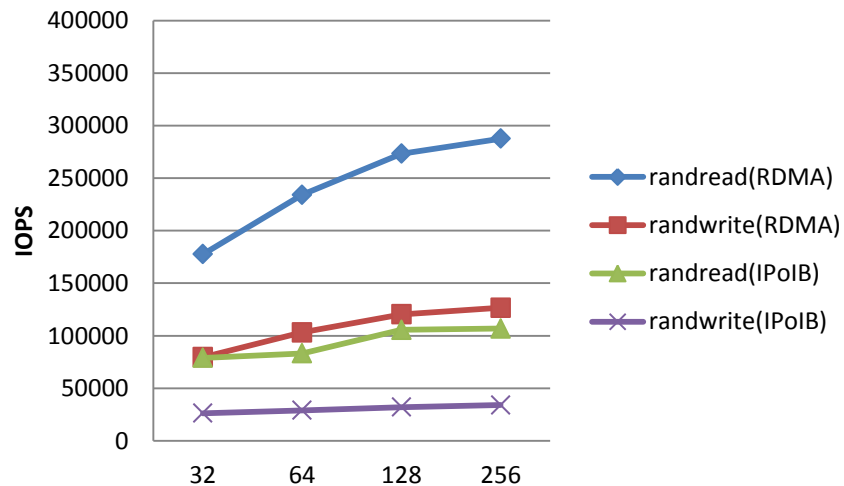
Storage node



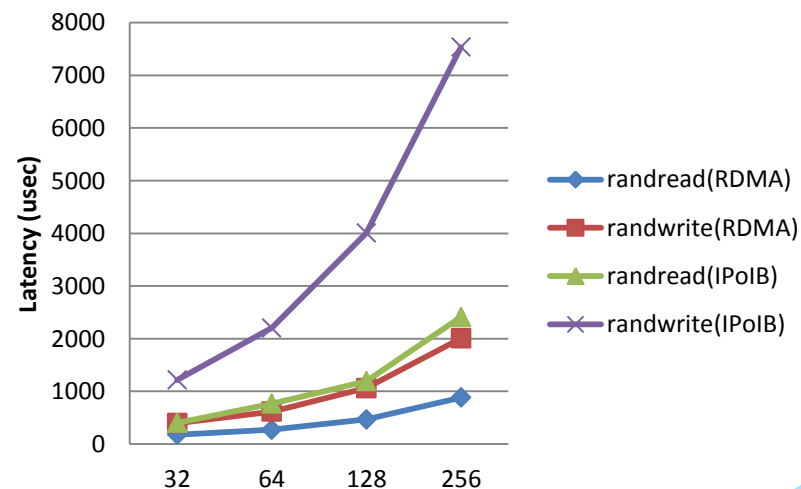
スループット



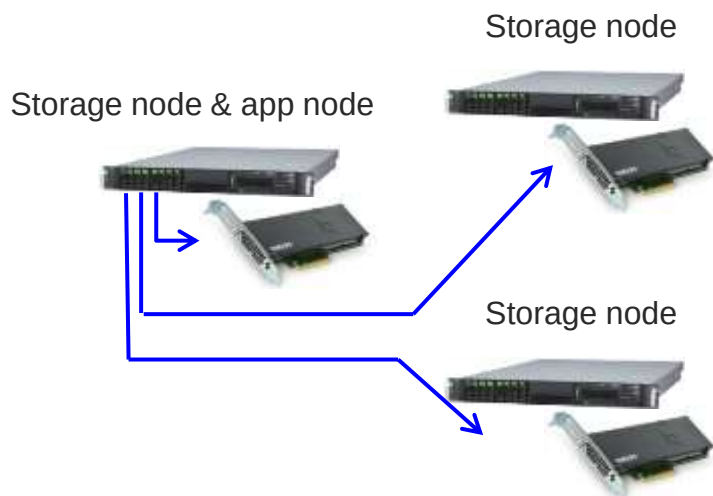
IOPS



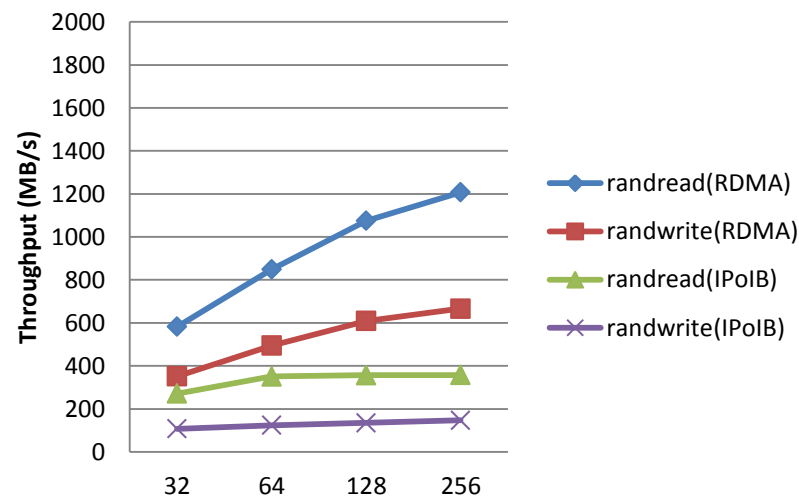
レイテンシー



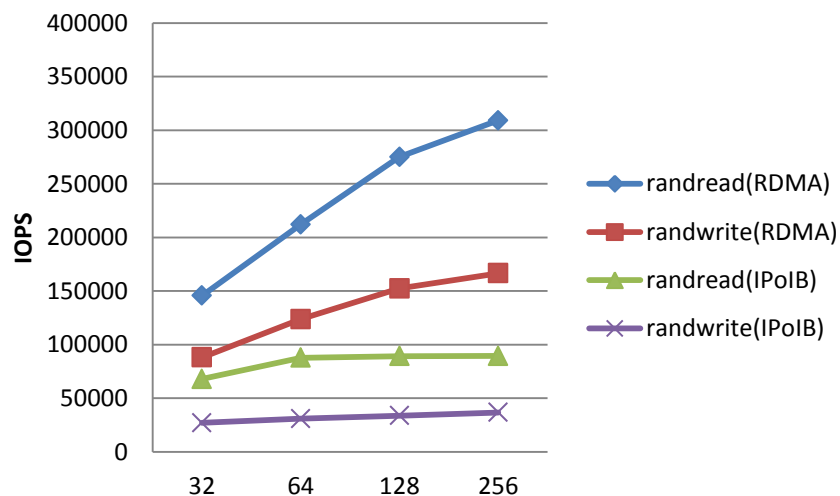
case4 : 3ストレージノード



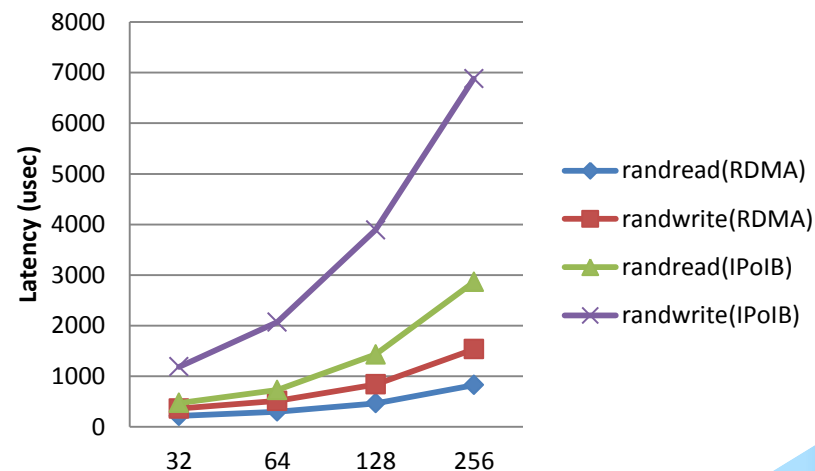
スループット



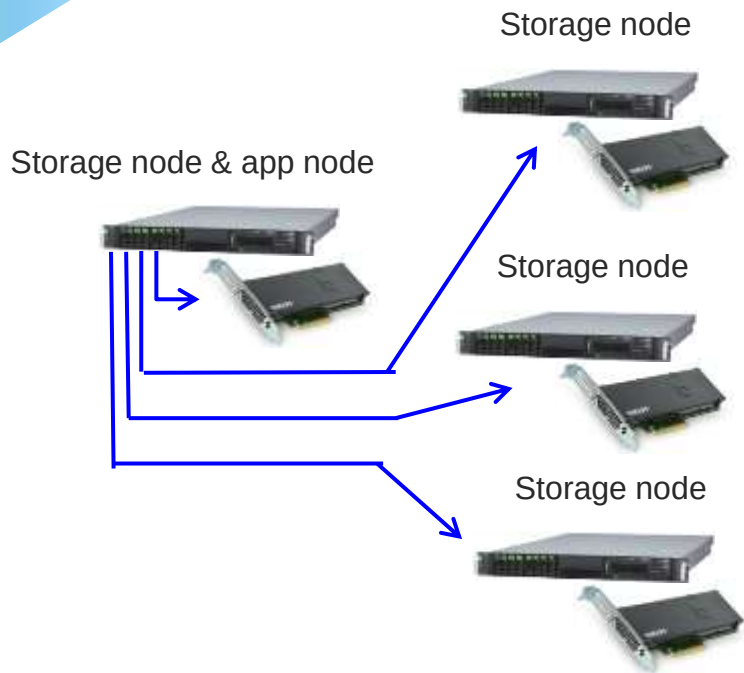
IOPS



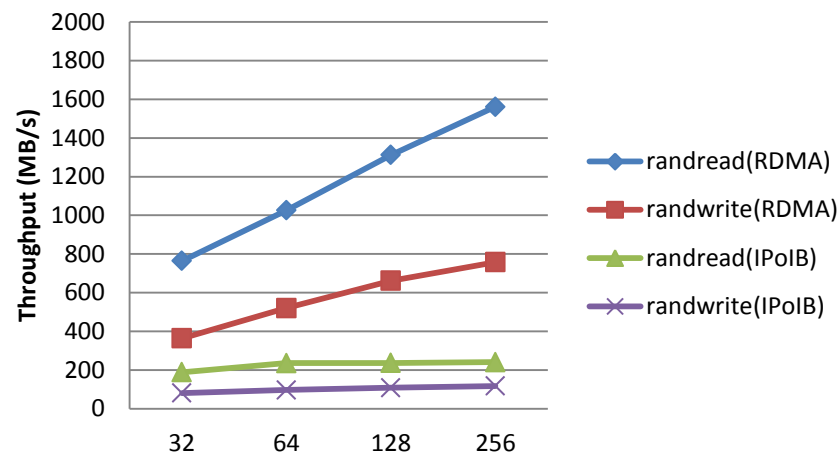
レイテンシー



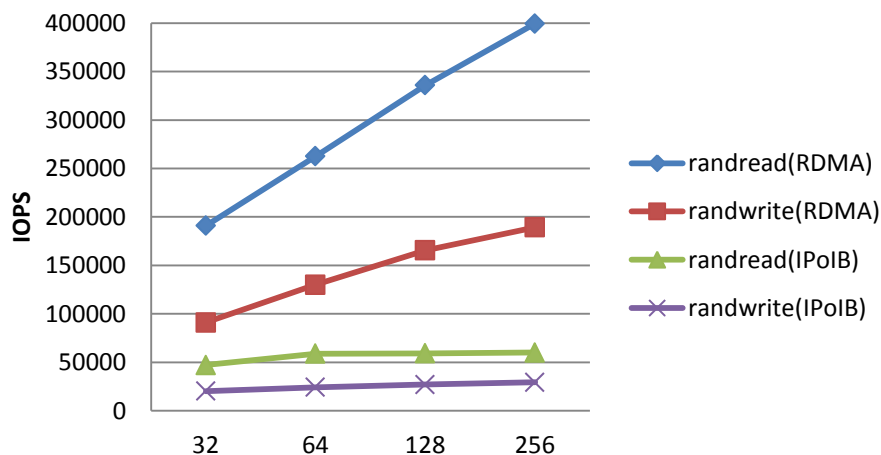
case5 : 4ストレージノード



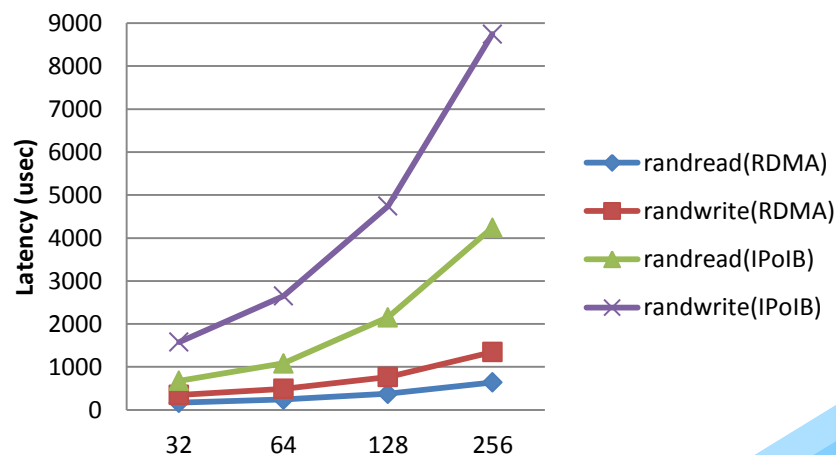
スループット



IOPS



レイテンシー



- ▶ Virident Solutions 2.0の機能全般を確認し動作することを確認できた
 - ◆ ソフトウェアインストール
 - ◆ ボリューム作成、追加、削除
 - ◆ ノード追加

- ▶ ノード数を増やすことで、IOPS/スループット性能がローカル性能以上にスケールすることを確認できた

- ▶ ネットワークプロトコルがRDMAのときIPoIB(TCP/IP)に比べ最大8倍^(*1)の性能差(IOPS/スループット)があることを確認できた

*1 : P34 case5グラフ参照

- ▶ レポート内容、ご質問は下記までご連絡ください。

株式会社アルティマ

TEL : 0454762197

[メール問い合わせ]

virident-support@altima.co.jp

営業 清水宏樹

SE 北島佑樹