



DDN GfarmScaler

Gfarmを加速するDDN GfarmScalerのご紹介

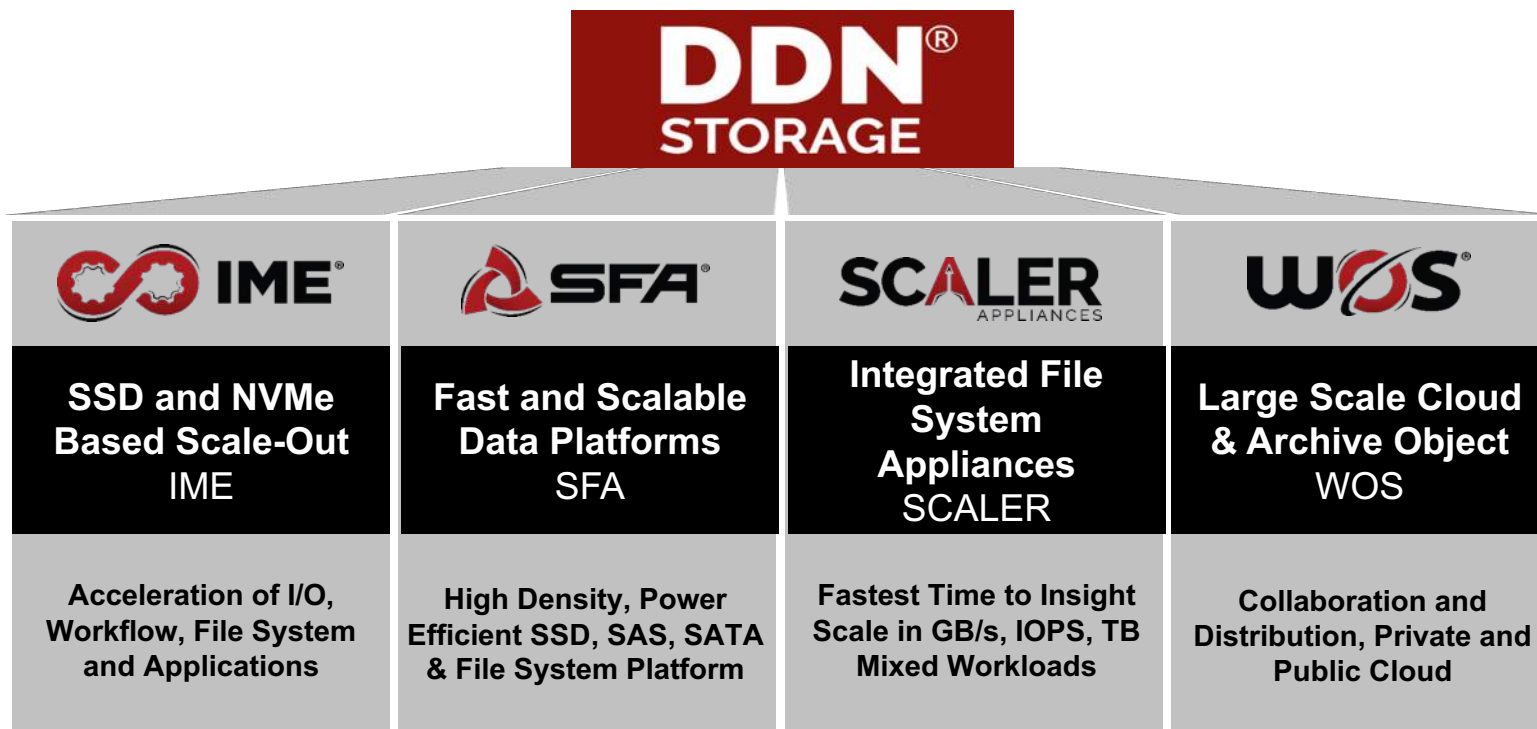
DataDirect Networks Japan, Inc.

橋爪信明

2017/12/22

DDN紹介

- HPCを中心にEnd-To-End Storage Solutionを開発・製造・販売しています



DDN 製品ポートフォリオ

Burst Buffer

Infinite Memory Engine (IME)

クライアントと並列ファイルシステムの間Layerに位置づけられるBurst Bufferであり一時的なBurst IOに対応し、多数プロセスからShared Fileへの小さなランダムIOアクセスなど並列ファイルシステムが不得意としているI/Oを高速化。アプライアンスおよびソフトウェアによる提供。



IME Software



IME14K



IME240

File System

DDN EXAScaler

- オープンソースのLustreファイルシステムをベースに構成
- 膨大な数のクライアントに対し高スループットIO環境を提供
- 多数のHPCサイトで多くの実績

DDN GRIDScaler

- 高いIO性能に加えてマルチプロトコルアクセス環境、Snapshot、Replication、DR、サイト間連携など多彩な機能を提供
- 低速ストレージ、テープ装置等を用いた階層化機能の実現

Storage Platform

SFA14KX/ES14KX/GS14KX



54GB/sec / 1.7M IOPS
Max 912 Drive

SFA7700X/ES7K/GS7K



12GB/sec / 450K IOPS
Max 396 Drive

※ ES14KX/GS14KX/ES7K/GS7KはRAIDコントローラ上でEXAScalerまたはGRIDScalerによるファイルサービスを提供するアプライアンス製品となります。

Drive Enclosure / Drive

SS8460/8412/8462



6/12Gbps 84Drive Enclosure

SS7700



6Gbps 60Drive Enclosure

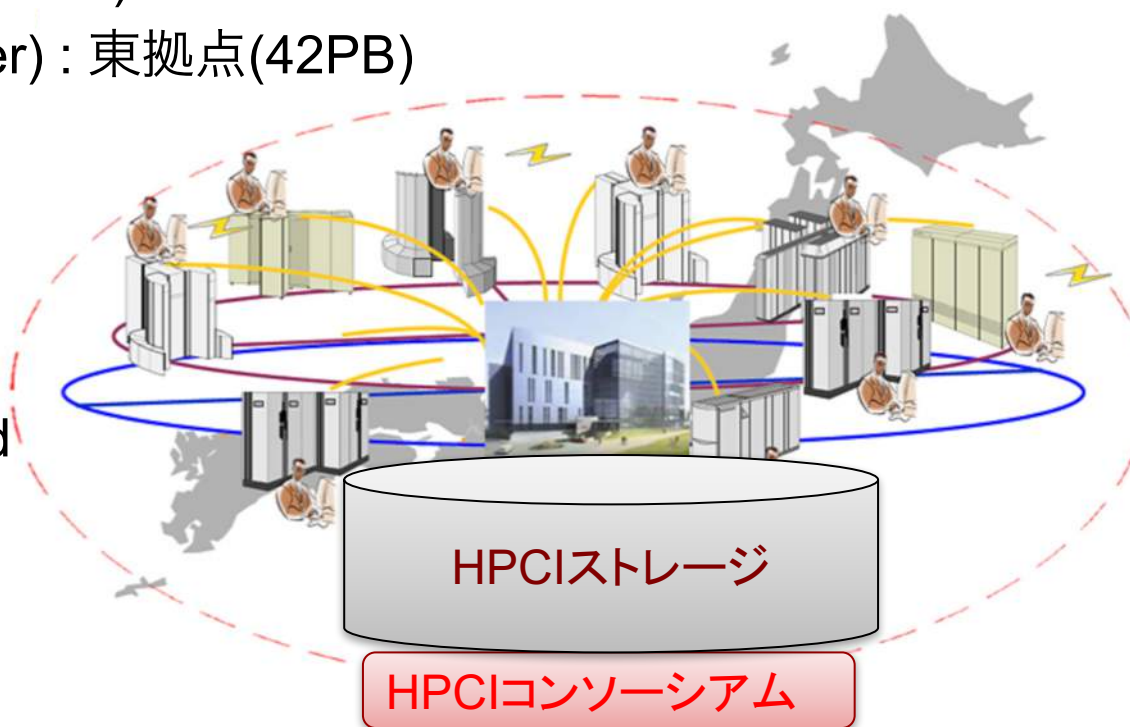
Drive



7.2K CapSAS
10,15K Perf SAS
SSD, NVMe SSD

DDNとGfarmの関わり

- HPCI (High Performance Computing Infrastructure)
 - ▶ HPCI共用ストレージとして2012年度から採用
 - ▶ SFA10K : 西拠点(10PB), 東拠点(12PB)
 - ▶ SFA12K : 東工大拠点(0.4PB)
 - ▶ SFA14KXE(GfarmScaler) : 東拠点(42PB)
- Gfarm
 - ▶ XFSを提供
 - ▶ SFA10K, 12K : Block
 - ▶ SFA14KXE : Embedded



GfarmScalerとは

- HPCI東拠点・西拠点更新調達で提案
 - ▶ サーバを減らすことによるコスト削減
 - ▶ 性能は維持
- SFA14KXE、SFA7700XE上でGfarmを提供
 - ▶ データストレージ用に外部サーバ必要なし
 - ▶ SFA14KXE、SFA7700XEのコントローラ上でメタデータ以外のサービスを提供
- Gfarmバックエンド：高速で大容量のXFS環境
 - ▶ 高スループットXFS
 - ▶ 大容量XFS

DDN SFA14KXE 概要

項目	製品仕様
フォームファクター	4U (Dual ControllerおよびBBU、2.5inch Drive slot x72を内蔵)
コントローラ	Active/Active Dual Controller
対応RAIDレベル	RAID1(1+1)、RAID5 (8+1 or 4+1)、RAID6(8+2 or 4+2)
最大理論性能	シーケンシャル Read/Write 40GB/s
最大ドライブ数	840x 3.5inch ドライブ ※10x SS8462構成時 48 x PCIe NVMe SSD or 72 x 2.5inch SAS SSD
対応ドライブ	7.2Krpm NL-SAS (12Gps SAS 4Kn) 10Krpm SAS (12Gps SAS 4Kn) SSD (12Gps SAS 4Kn) 混在可能
キャッシュ	32GB / Controller (Battery-Backup DDR4 SDRAM)
ホストインターフェイス	4x EDR Infiniband / Controller 2x Intel Omni-Path FC / Controller
バックエンドポート	20x SAS 3.0
冗長部品	RAIDコントローラ、電源、冷却ファン
ホットスワップ交換対象部品	RAIDコントローラ、電源、冷却ファン、ドライブ
ソフトウェア機能	LUNマッピングおよびマスキング、Read QoS、ポートゾーニング、Direct Protect
管理機能	Web GUI管理、CUI管理、メールベースの障害発報、SNMP

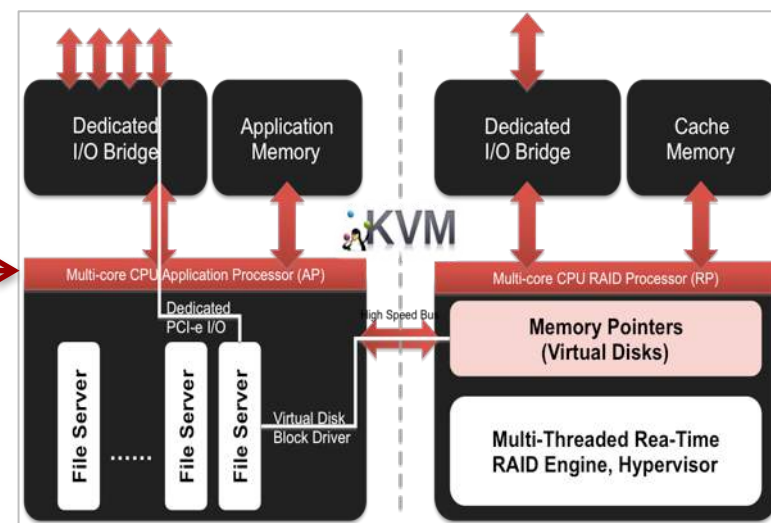
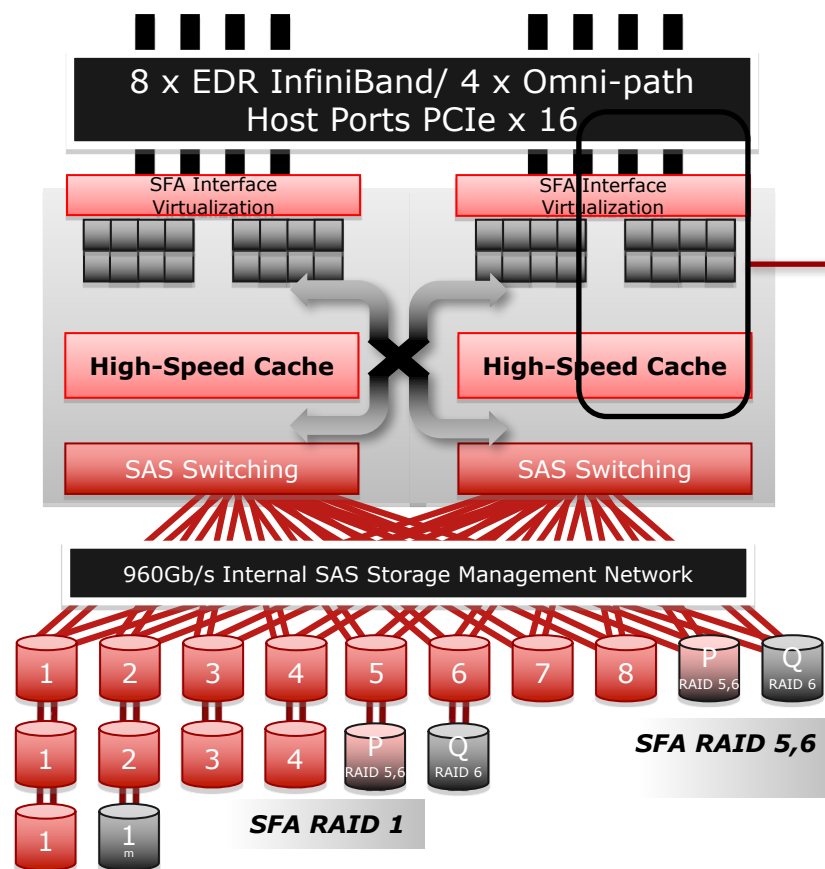


SFA14KXE 詳細

	SFA14KXE
用途	ファイルサーバアプライアンス
アップリンクポート	ファイルサーバからのアップリンクポートとしてインターコネクトネットワークに接続しファイルサービスを提供
サポートネットワークポート	EDR IB Intel OmniPath 40GbE 100GbE
ポート数/コントローラ (ポート数/システム)	EDR IB : 4 (8) Intel OmniPath : 2 (4)
CPU/Memory (14KXEはAP/RP共通)	e5-2695v4 (18-cores) x4 (2コントローラ合計) 16 x 32GB @2400 MHz (512GB)
ファイルサーバ用VM数	1 or 2 / コントローラ, (1CPU内8core利用) 2 or 4 / システム
RAIDキャッシュ	32GB/コントローラ
アプリケーションメモリ(VMで使用)	180GB/コントローラ (90GB/プロセッサ)
最大サポートディスク数	1752(うち拡張エンクロージャは1680)
最大サポートディスク エンクロージャ	10 (Direct接続) 20 (Daisy Chain接続)

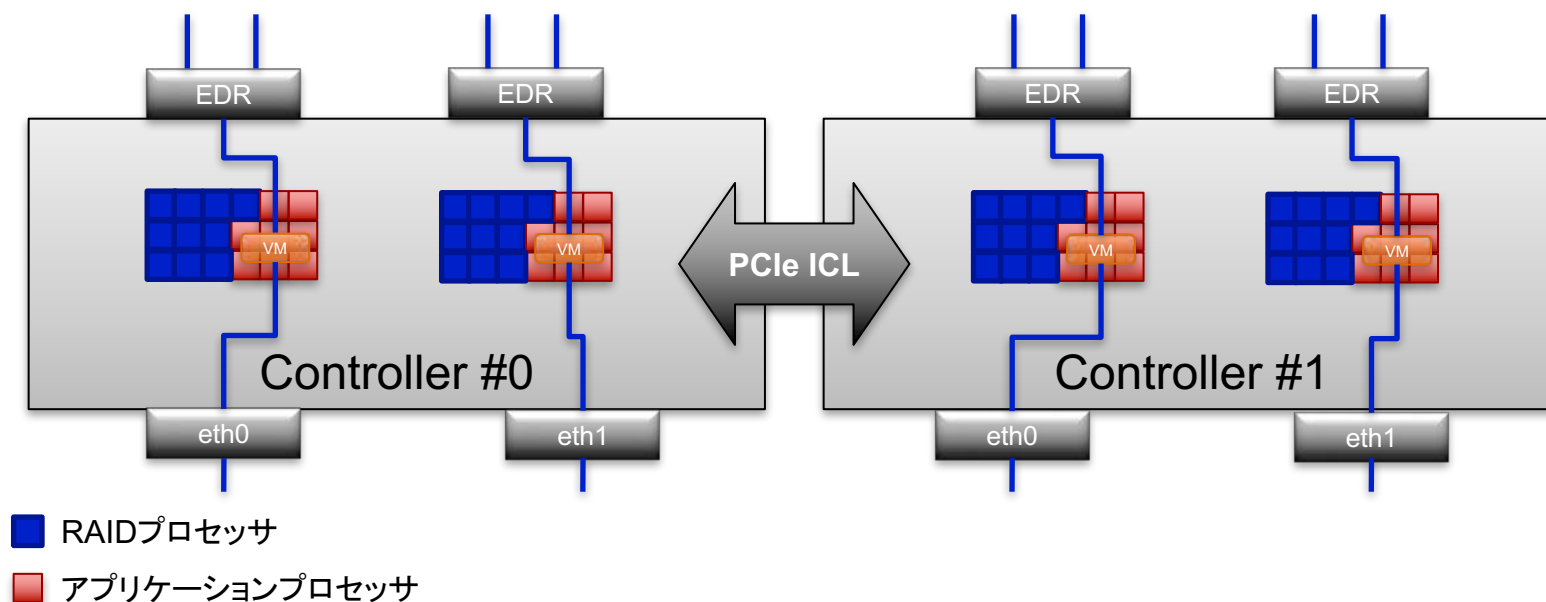
SFA14KXE ブロック図

DDNのファイルサーバプラットフォームであるSFA14KXEプラットフォーム上に、KVMを搭載し、クライアント群にファイルシステムサービスを提供します。



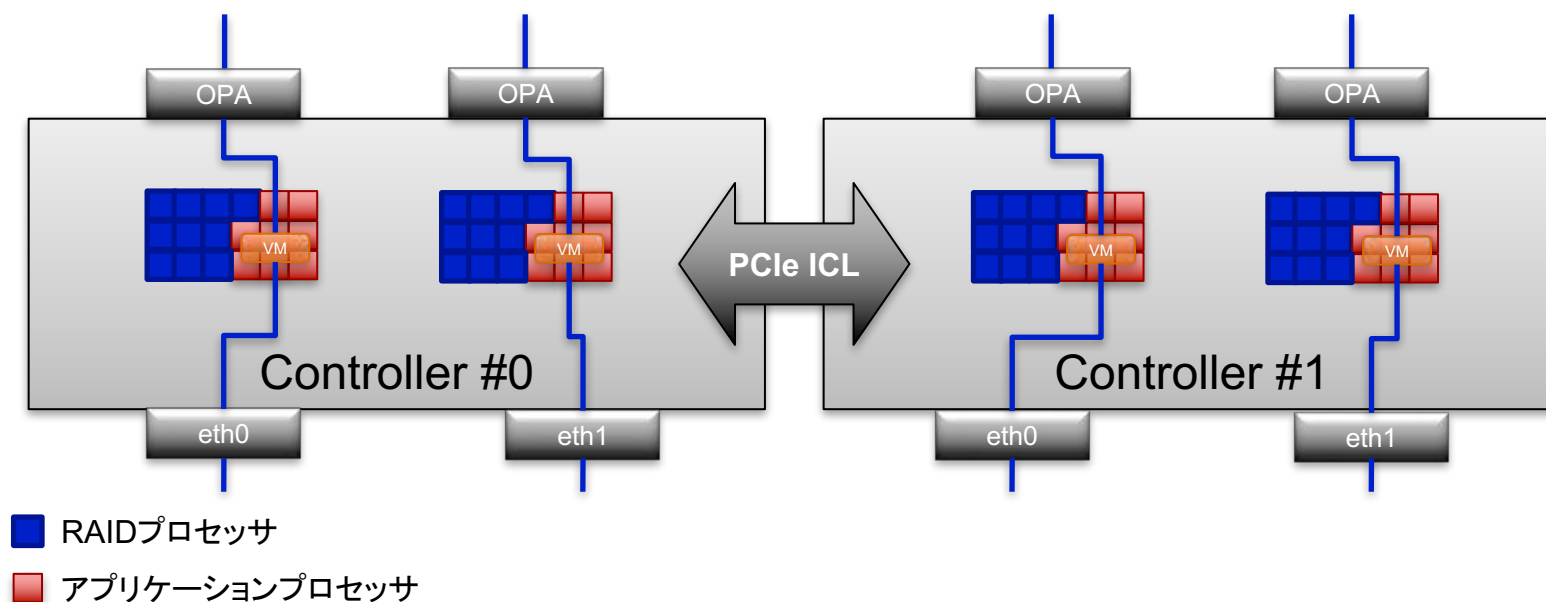
SFA14KXE VM構成例 - EDR

- EDR利用例
- KVM構成
 - 各CPU Socket毎に1x VMを割当
 - 18CPU Core中8CPU Coreを各VMに割当
 - Dual Port EDR-IB HCA x1 を各VMに割当、SR-IOVで使用
 - 各VMへの割当可能メモリは最大90GB
- Mgmt Ethernet構成
 - 2ポート/コントローラ
 - eth0はSFAOS,BMC,VM0が共有
 - eth1はVM1が利用(eth1を未使用とし、VM1をeth0に割り当ててることも可能)



SFA14KXE VM構成例 - OPA

- OPA利用例
- KVM構成
 - 各CPU Socket毎に1x VMを割当
 - 18CPU Core中8CPU Coreを各VMに割当
 - Single Port OPA HFI x1 を各VMに割当、Path Throughで使用
 - 各VMへの割当可能メモリは最大90GB
- Mgmt Ethernet構成
 - 2ポート/コントローラ
 - eth0はSFAOS,BMC,VM0が共有
 - eth1はVM1が利用(eth1を未使用とし、VM1をeth0に割り当ててることも可能)



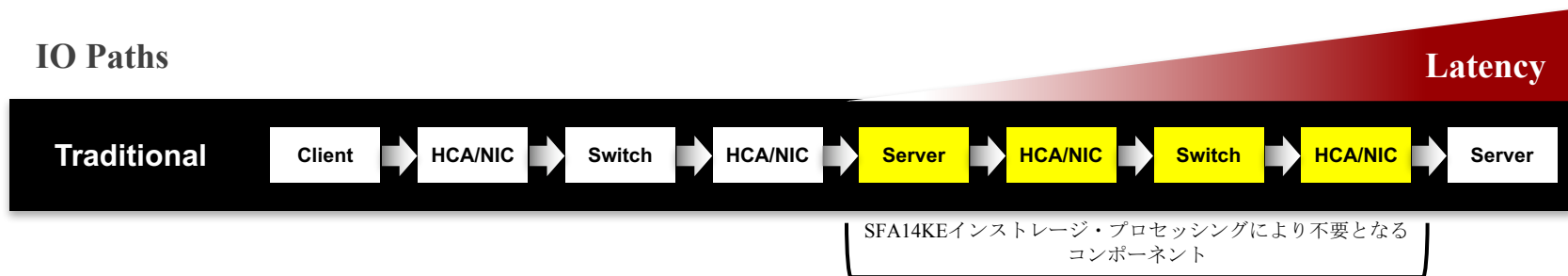
SFA14KXE

インストレージ・プロセッシング

SFA14KXEはストレージコントローラ内でファイルサービスを実装することにより、外部サーバ、ネットワークアダプタ、スイッチなどを必要としません。SFA14KXEの主なメリットは以下の通りです。

- ・ 外部サーバ、ネットワークアダプタ、スイッチ購入コスト削減
- ・ 構成の簡素化による管理コストの削減
- ・ 低遅延でのストレージアクセスを実現（下記図参照）

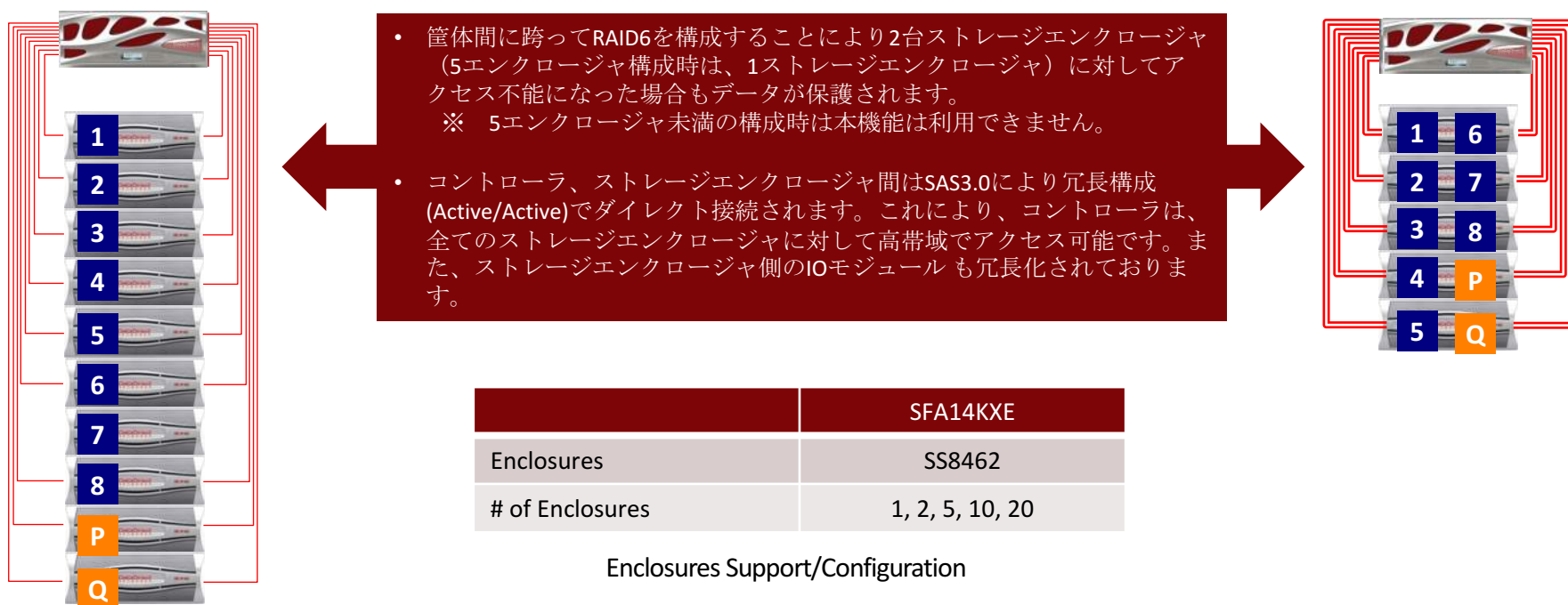
IO Paths



データ/パケットの経路が簡素化することによりIOPSを向上

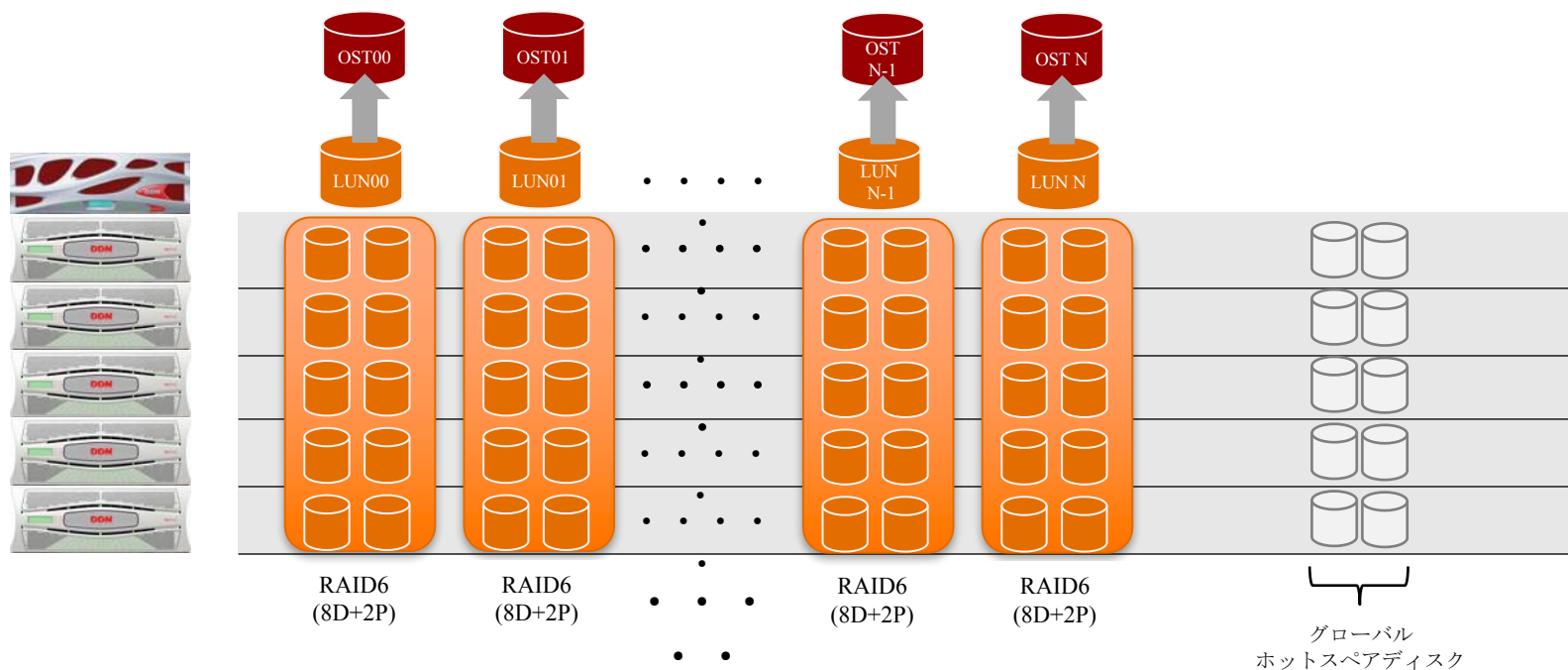
ダイレクト接続モデル

SFA 14KXEではコントローラ～ストレージエンクロージャ間がSAS 3.0インターフェイスにより直接接続されます。これにより、完全な冗長性と高内部帯域幅を提供します。(20エンクロージャ構成時はDaisy Chain接続)



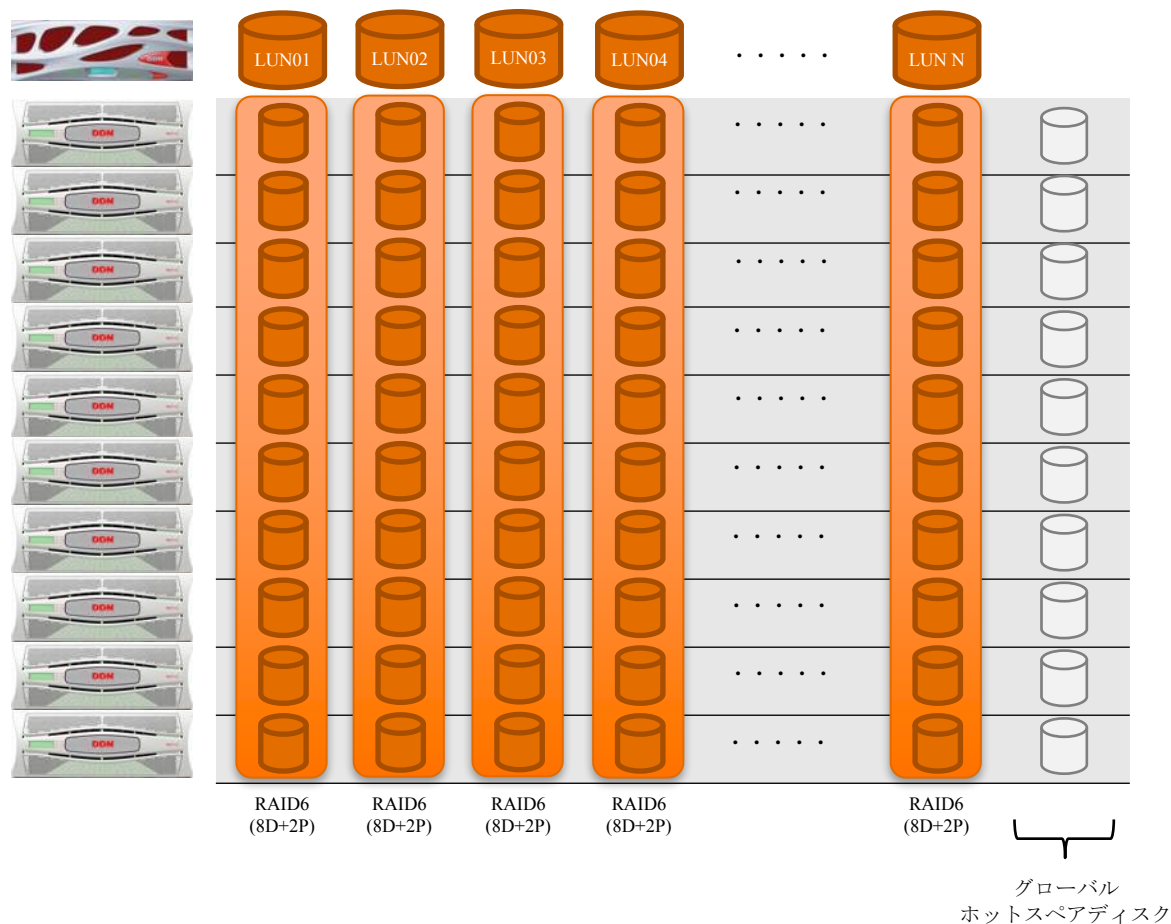
ストレージシステムのHDD構成 – 5enclosure

提案構成では、各LUNはRAID6(8D+2P)で構成されます。5台のエンクロージャに跨ってRAID6を構成することにより1台のエンクロージャがアクセス不能となってもデータが保護されます。



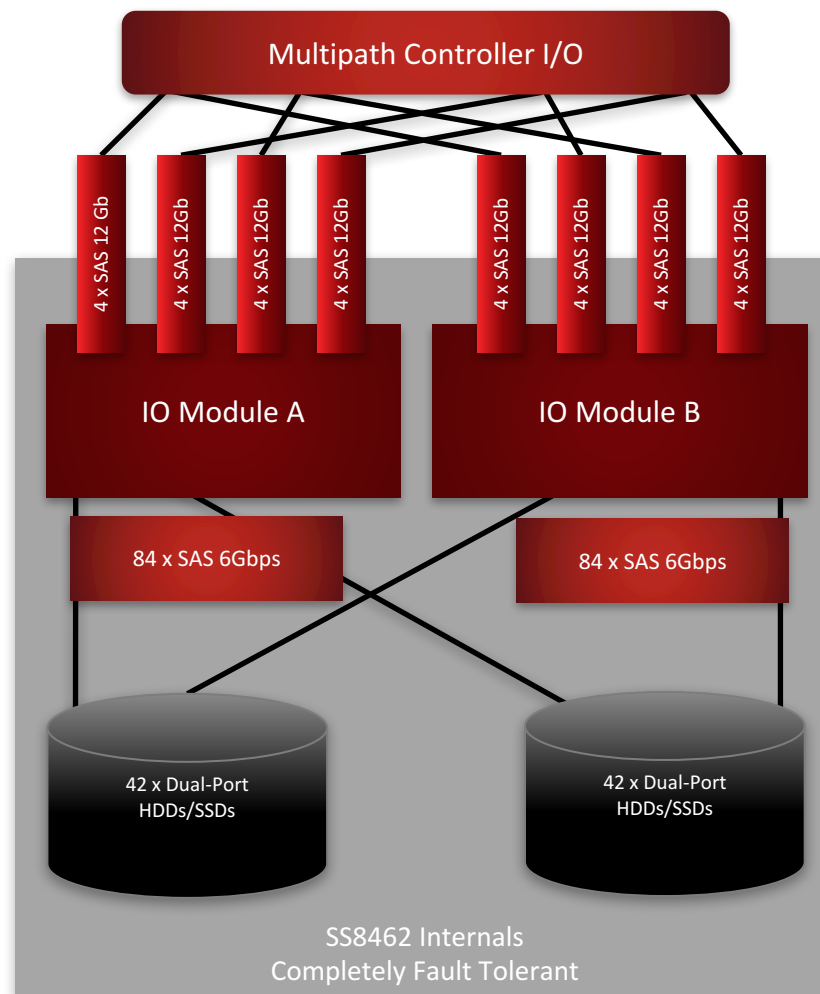
ストレージシステムのHDD構成 – 10enclosure

提案構成ではありませんが、10エンクロージャの場合、各LUNはRAID6(8D+2P)で構成されます。10台のエンクロージャに跨ってRAID6を構成することにより最大2台のエンクロージャがアクセス不能となってもデータが保護されます。



DDN SS8462ストレージエンクロージャ

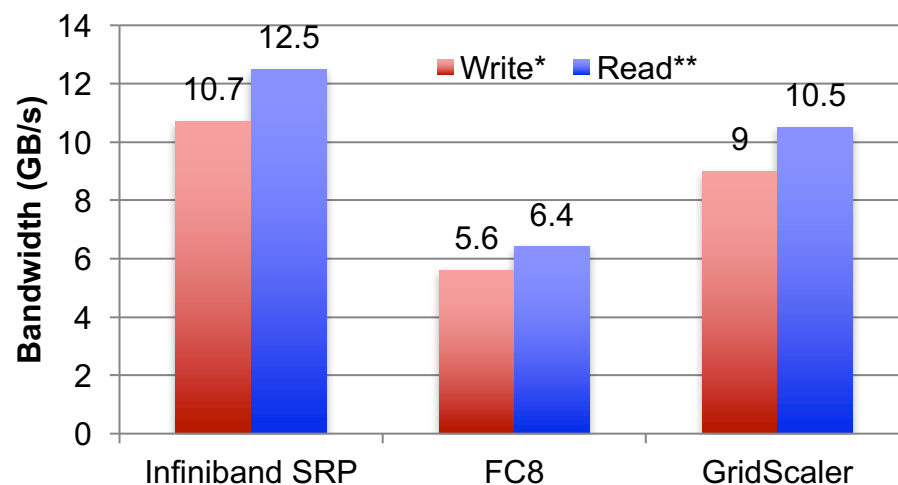
ご提案のSS8462は、84本のドライブを収容可能な高密度ストレージエンクロージャです。SSD、SAS、SATAの各種ドライブを混在可能であり、I/Oパスも含めて全てのコンポーネントが冗長化されています。



SFA7700X概要

SFA7000 Specifications

Appliances	Active/Active Block Servers
Cache	64GB - Battery Backed Cache 22GB Mirred
RAID Levels	RAID 1/5/6
SFA Flash Cache	12.8 TB
Host Connectivity	4 x 56Gb (FDR) または 8x 8Gb/s または 4x16Gb/s FC
Drive Support	2.5"/3.5" SAS, SATA, SSD
Base Unit Drive Slots	60
Max JBOD Expansion	最大 4
Supported JBODs	SS700 60-Drive 6Gb SS8460 84-Drive 6Gb



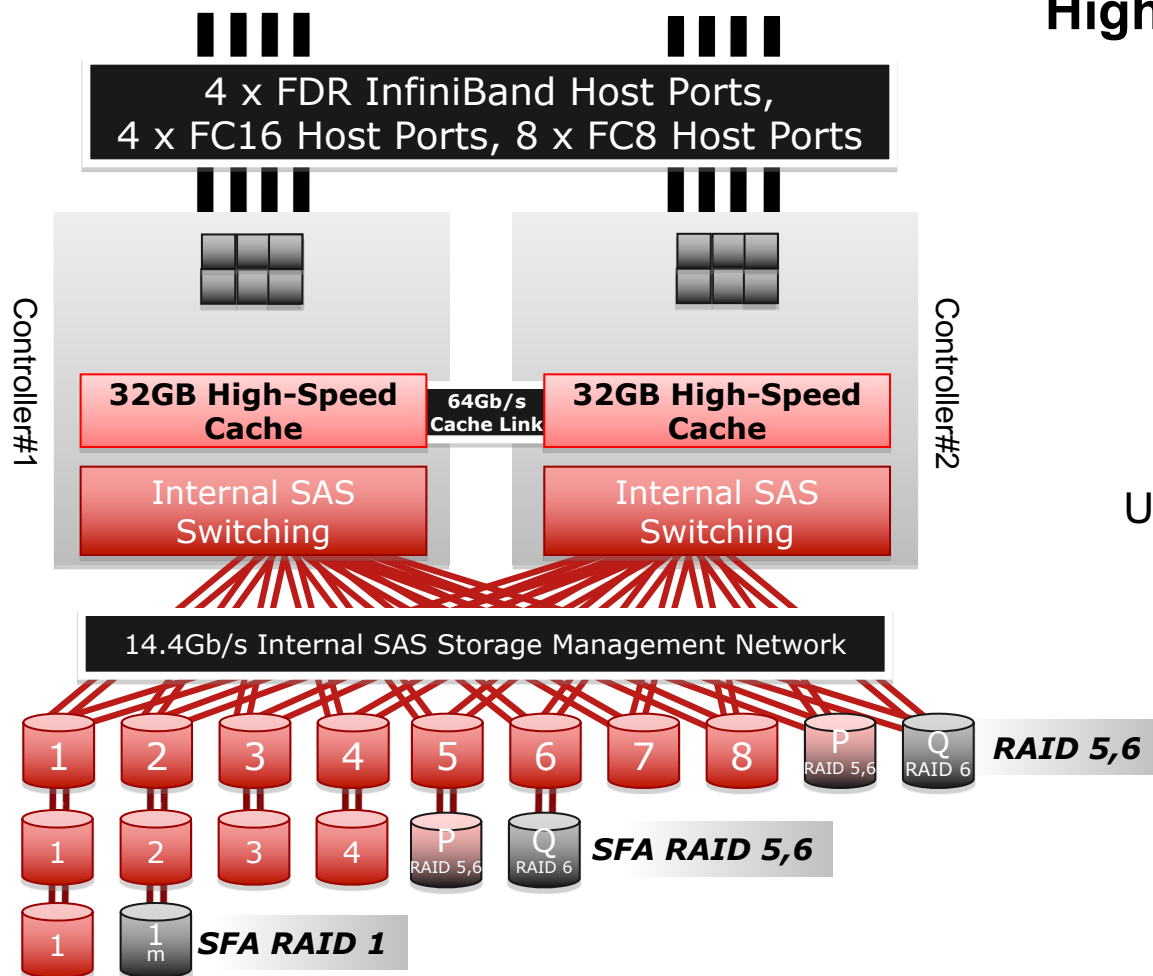
* Large Block, Full Stripe

** Read I/O Includes Parity Verification

SFA7700X詳細

項目	製品仕様
コントローラ	Active/Active Dual Controller
RAID	RAID6(8+2,4+2), RAID5(8+1,4+1), RAID1(1+1)
最大理論スループット	シーケンシャル Read/Write 10GB/s
IOPS	45万 IOPS ※デュアルコントローラにおけるキャッシュに対するランダム 4KB IOPS値
最大ドライブ数	396台(SS8460x4)
キャッシュ	16GB / Controller バッテリバックアップ保護 ECC不揮発性メモリ キャッシュ
ホストインターフェイス	2x FDR Infiniband / Controllerまたは 4 x 8Gbps FC / Controllerまたは 2x 16Gbps FC /Controller
ホットスワップ交換対象部品	RAIDコントローラ、ドライブ、電源、冷却ファン
システム管理	CLI、GUI(Web) SNMP、NTP、Syslog、電子メールによる障害通知

SFA7700X ブロック図



Highly Parallelized SFA Storage Processing Engine

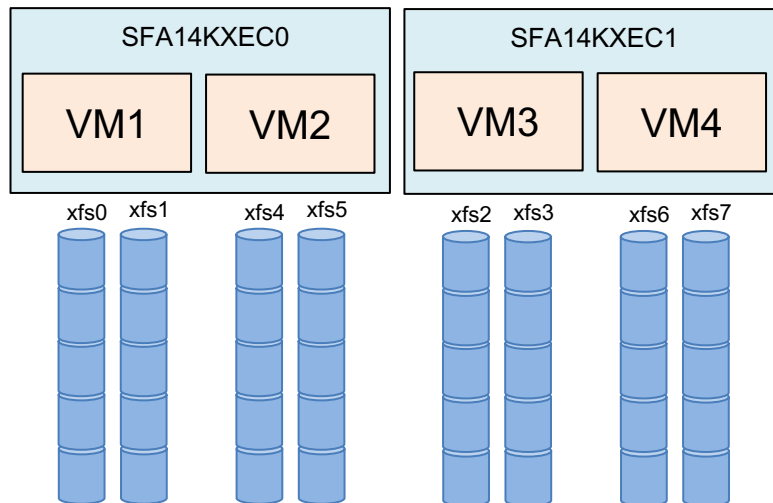
- Active/Active Storage Design
- 10-12GB/s Read & Write Speed
- Up to 396 Disk Drives Expansion
- 600K Burst IOPS
- 350K Sustained Random SSD IOPS
- 32GB Mirrored Cache (Protected)
- Up to 12.8 TB Flash Cache using SFX
- RAID 1/5/6
- Intelligent Block Striping
- DirectProtect™
- GUI, SNMP, CLI, API
- 4 x FDR IB Host-Ports
- 4 x FC16 Host-Ports
- 4RU Height

XFS on GfarmScaler : ベンチマーク環境

- SFA
 - ▶ SFA14KXE + 5x SS8462
 - ▶ 400x 7.2Krpm NL-SAS
- XFS
 - ▶ SFA14KXE VM上にXFSを構成
 - ▶ VM毎に1/2/4 XFS構成を推奨
 - ▶ 1MBにアラインされたSequential IOであれば、コントローラ間のキャッシュミラーを行わず、最大性能を得られる
- 性能
 - ▶ 14KXE: 1VMあたり、Write 約7GB/s、Read 約8.5GB/s までスケール
 - ▶ 7700XE: 未実施だが、Aggregated で6-7GB/s Write, 7-8GB/s Readを想定

Benchmark Configuration

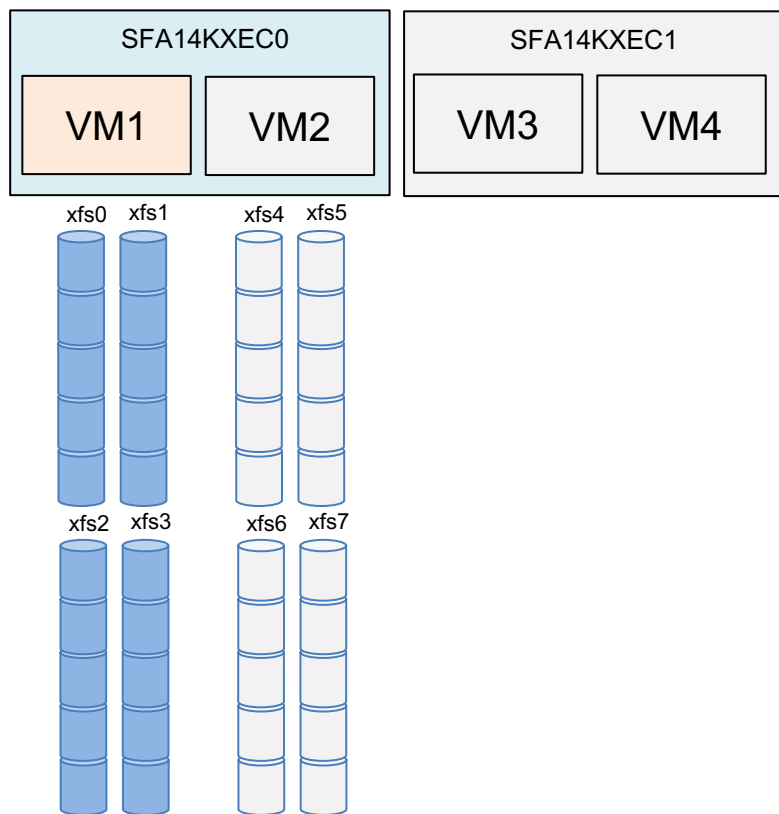
Common Setup



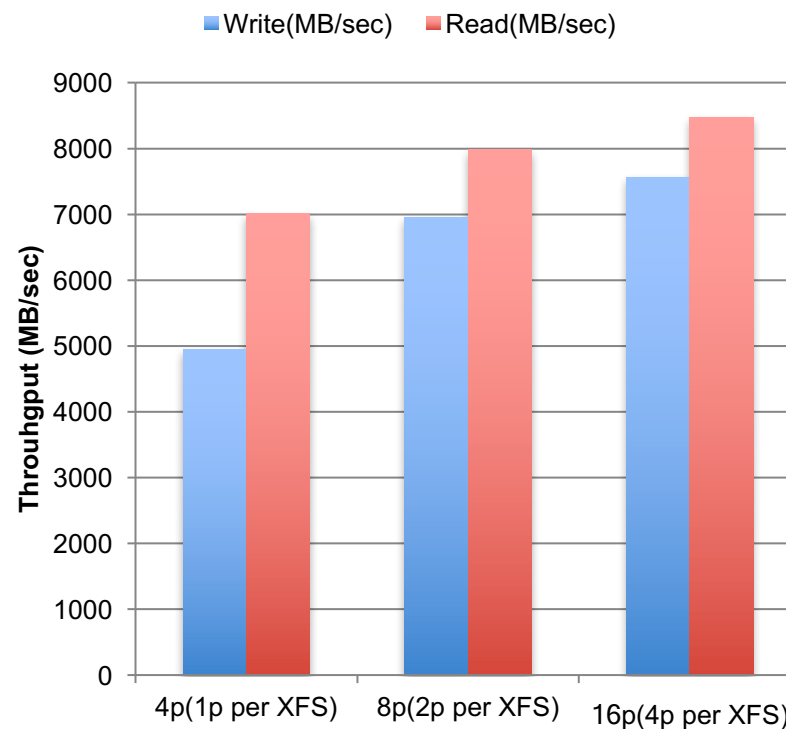
- 1 x SFA14XE + 5 x SS8462
 - ▶ 400 x 6TB NL-SAS
 - ▶ 40 x LUN(8D+2P)
 - ▶ 4 x VM
 - ▶ 5 x LUN per a XFS with LVM
- Tested with IOzone
 - ▶ 1MB sequential Write and Read
 - ▶ Create total 2TB files for test
 - ▶ `iozone -i0 -i1 -+n -r 1m -s XXg -t X -+m hostlist`

Benchmark Results: Backend Filesystem(1)

1/4 H/W resource (1 x VM, 200 x Disk, 4 x XFS)

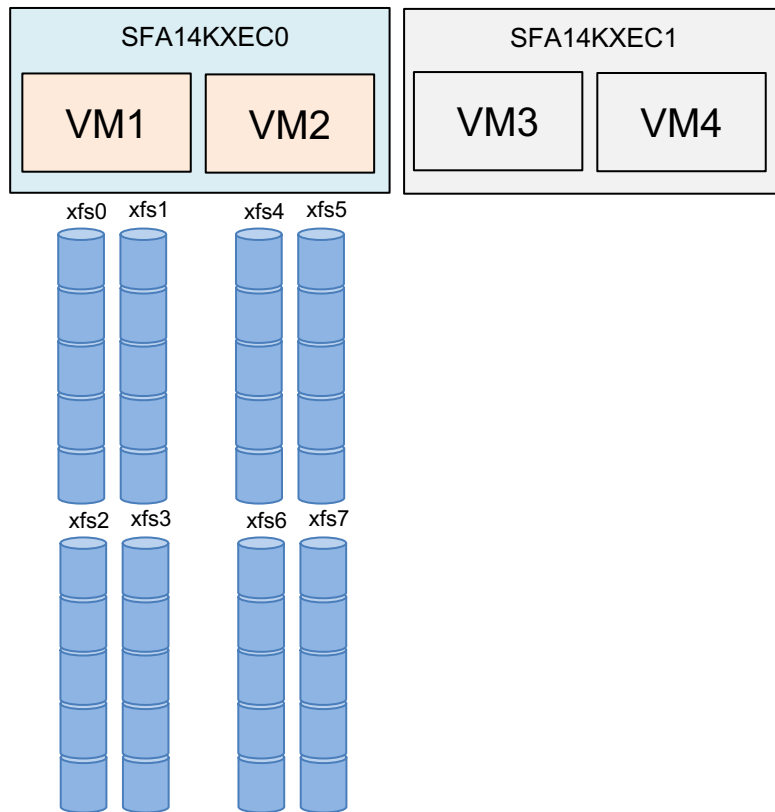


IOZone(1 x VM, 200 x Disk, 4 x XFS)

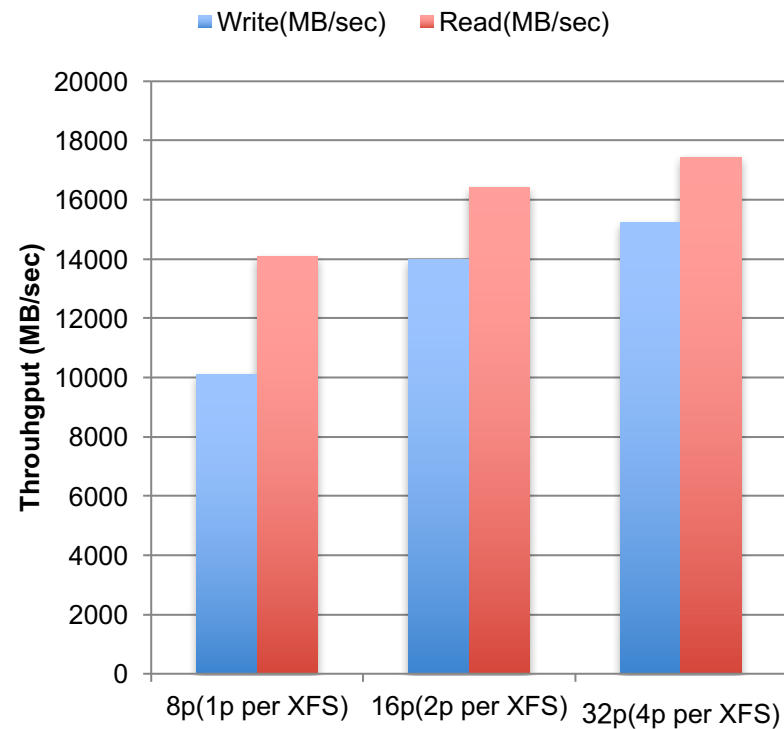


Benchmark Results: Backend Filesystem(2)

2/4 H/W resource (2 x VM, 400 x Disk, 8 x XFS)

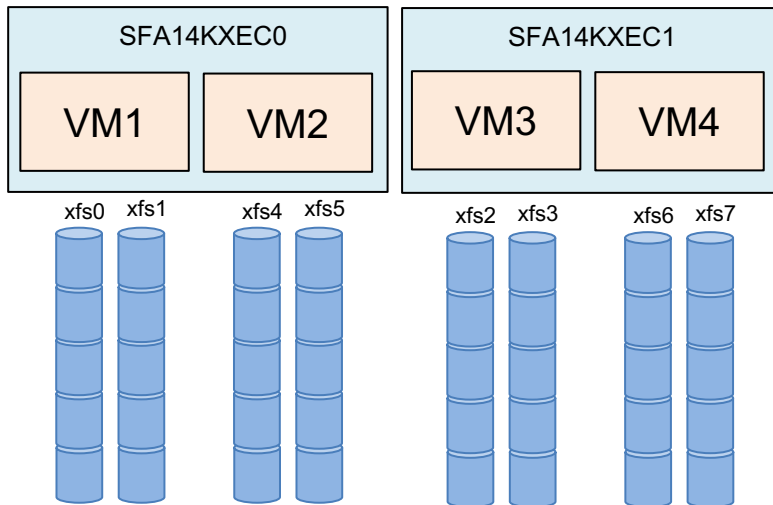


IOZone(2 x VM, 400 x Disk, 8 x XFS)

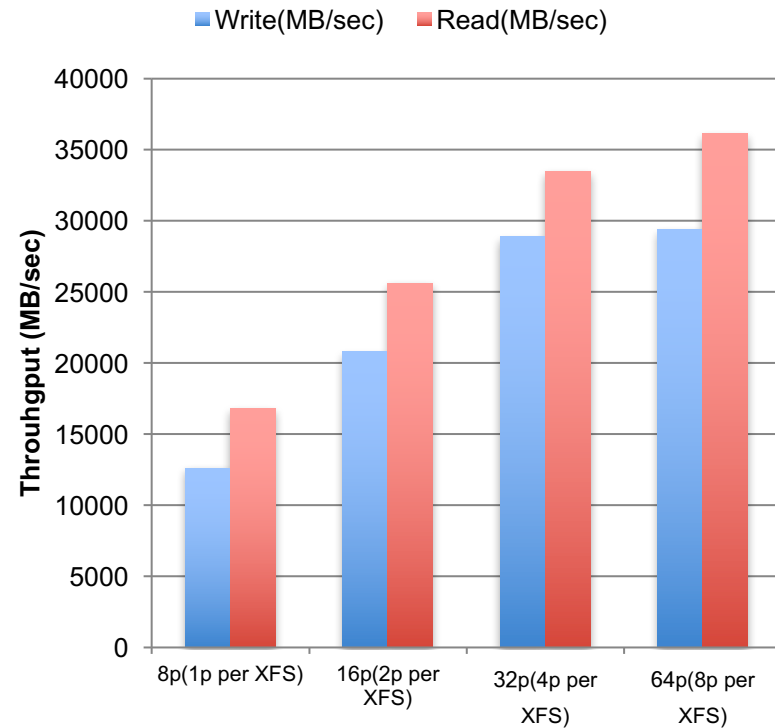


Benchmark Results: Backend Filesystem(3)

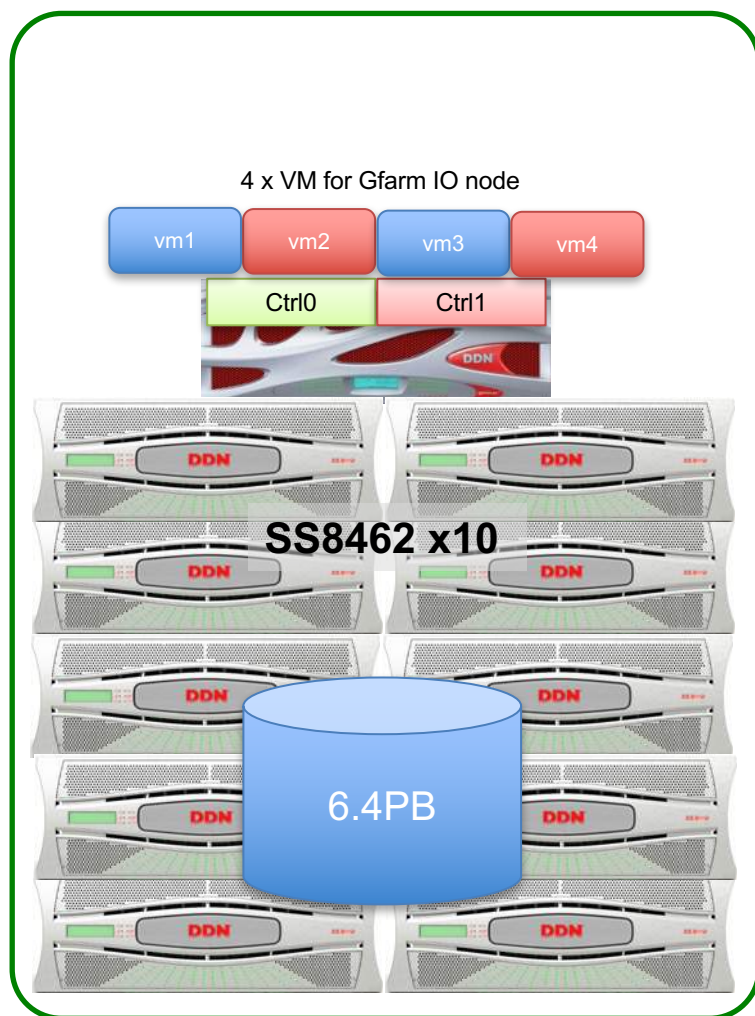
4/4 H/W resource (4 x VM, 400 x Disk, 8 x XFS)



IOZone(4 x VM, 400 x Disk, 8 x XFS)



東拠点でのXFS性能(実測値)



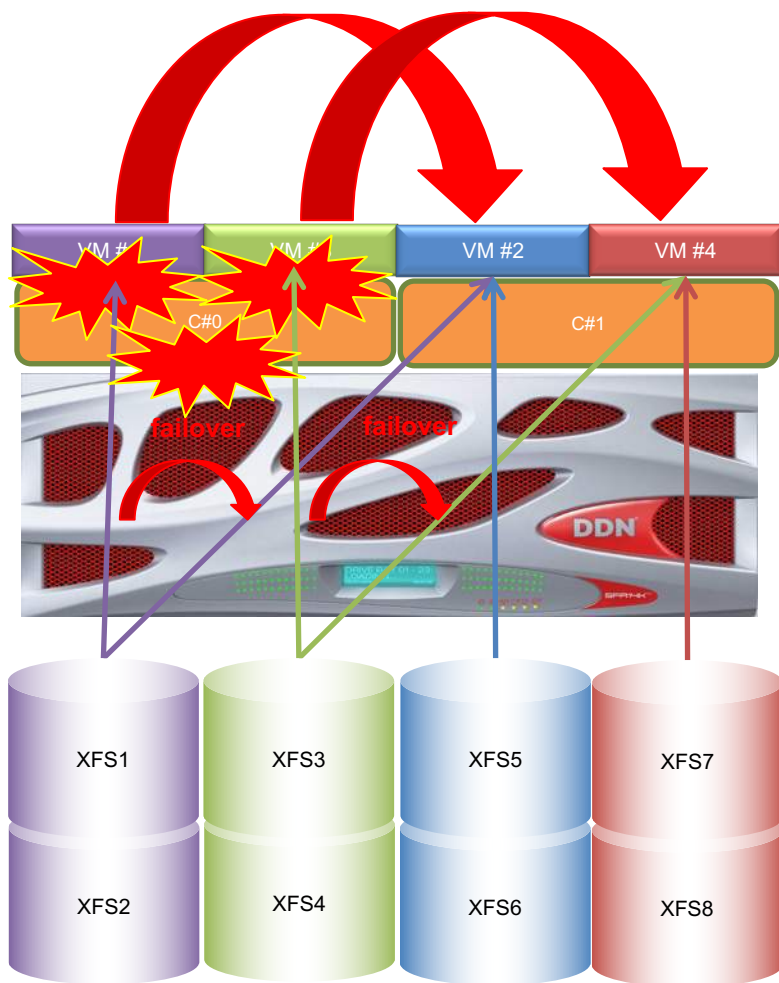
- 7x GfarmScaler(SFA14KXE)
- Single SFA building block
 - ▶ 10x SS8462
 - ▶ 770x HGST 10TB Disk
 - ▶ 8x XFS filesystem (2x XFS per VM)
 - ▶ Aggregate XFS performance from VM
 - Write 31.0GB/sec
 - Read 31.9GB/sec
 - +20% Over achieved against performance target!
- Scalability and manageability
 - ▶ Reduce IO stack and Gfarm talks to SFA RAID engine directly
 - ▶ Online SFA firmware upgrade for robust Gfarm services

Performance Summary

- Performance scales linearly
 - ▶ Single SFA14KXE controller (400 x NL-SAS)
 - 1 x VM: Max 7.5GB/sec (Write), 8.4GB/sec (Read)
 - 2 x VM: Max 15.2GB/sec (Write), 17.4GB/sec (Read)
 - ▶ Dual SFA14KXE controllers (400 x NL-SAS)
 - 4 x VM: Max 29.3GB/sec (Write), 33.4GB/sec (Read)
 - ▶ Dual SFA14KXE controllers (770 x NL-SAS)
 - 4 x VM: Max 31.0GB/sec (Write), 31.9GB/sec (Read)

Failover Configuration for XFS

- XFS Failover by pacemaker/corosync
- Developed new application resource group



Thank You!

Keep in touch with us



Team-JPSales@ddn.com



Tokyu Bancho Bldg. 8F
6-2 Yonbancho Chiyoda-ku,
Tokyo 102-0081



@ddn_limitless



+81-3-3261-9101
+81-3-3261-9140



company/datadirect-networks