Nexenta
Enterprise class storage for everyone

BizStor
Core Micro Systems
Partner Program

SUPER STORAGE

STORAGE APPLIANCE

ZFS オープンストレージ OS NexentaStor 及びアプライアンス製品のご紹介

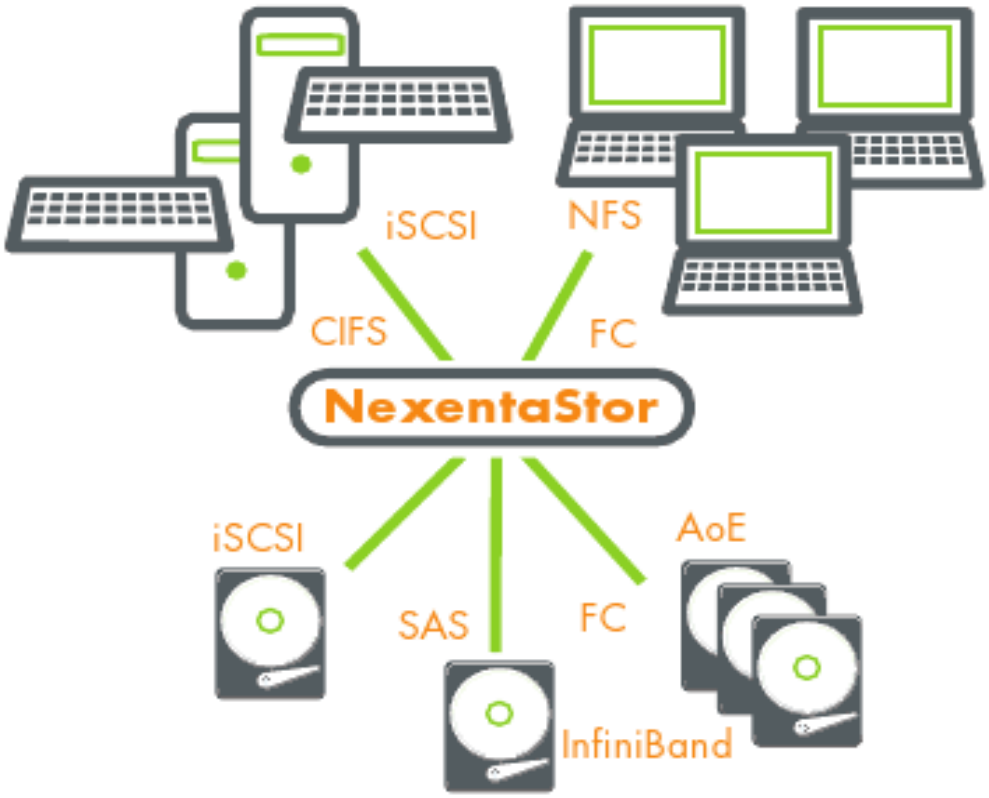
= Contents =

コアマイクロシステムズ株式会社
技術部
Mail : sales@cmsinc.co.jp

平成 21年 11月 13日

NexentaStor

Nexenta Stor とは?



統合ストレージ環境：block and file

オープンストレージOSの先鋭

IA サーバハードウェアで動作

OpenSolaris / ZFS によりエンタープライズストレージシステムを従来の三割程度のコストで実現

- 強力なデータ保護機能
- ファイルサイズ無制限
- スナップショット無制限

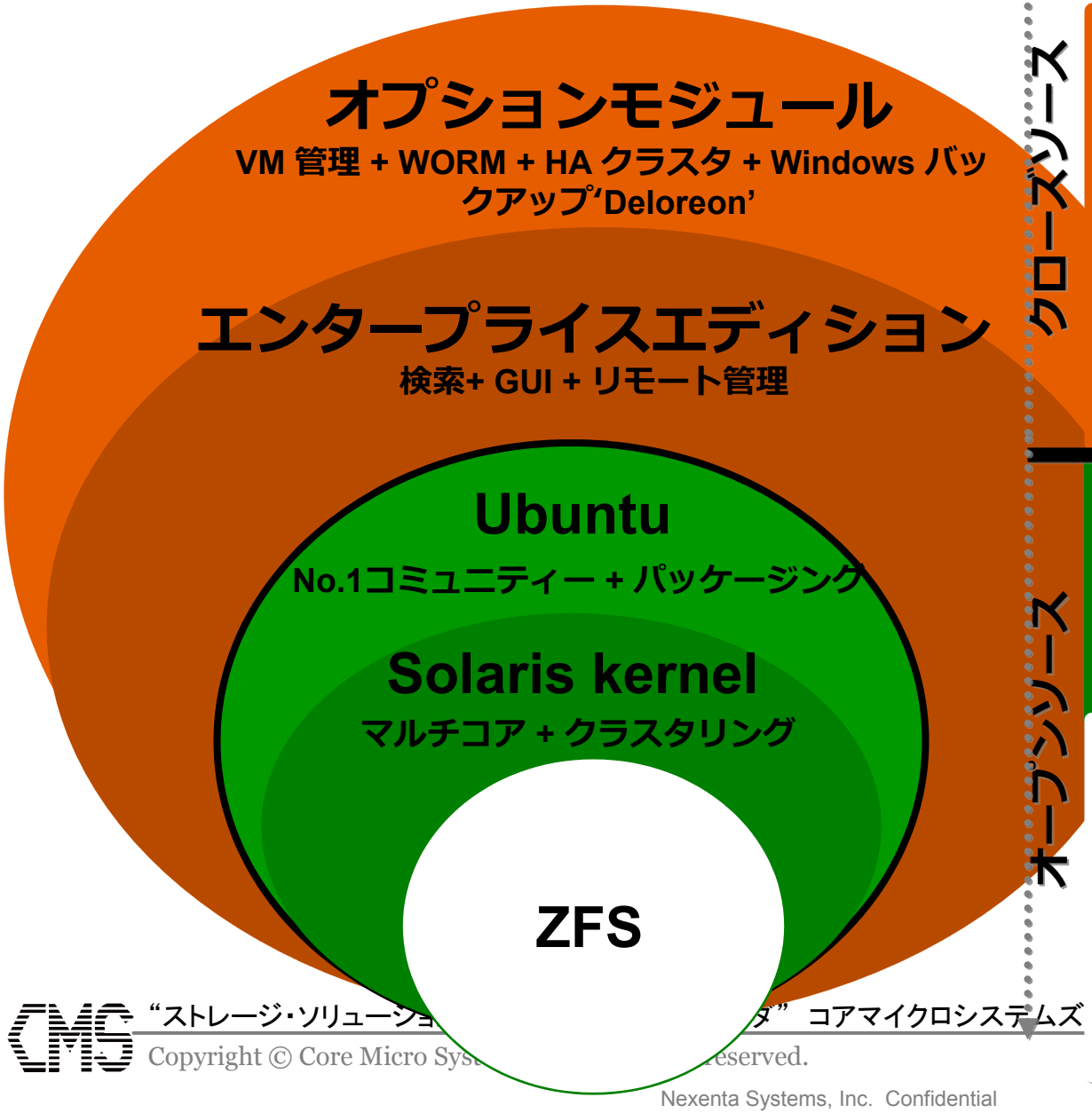
仮想環境対応ストレージ

Windows Server 2008 Hyper-V

vmware technology alliance PARTNER

Xen

オープンストレージOS構造



クローズソース

NexentaStor

- ハードウェアに対する独立型
- NAS/SAN/iSCSI
- ZFSスナップショットによるCDP機能
- ブロック同期によるCDP機能
- イベントベース API

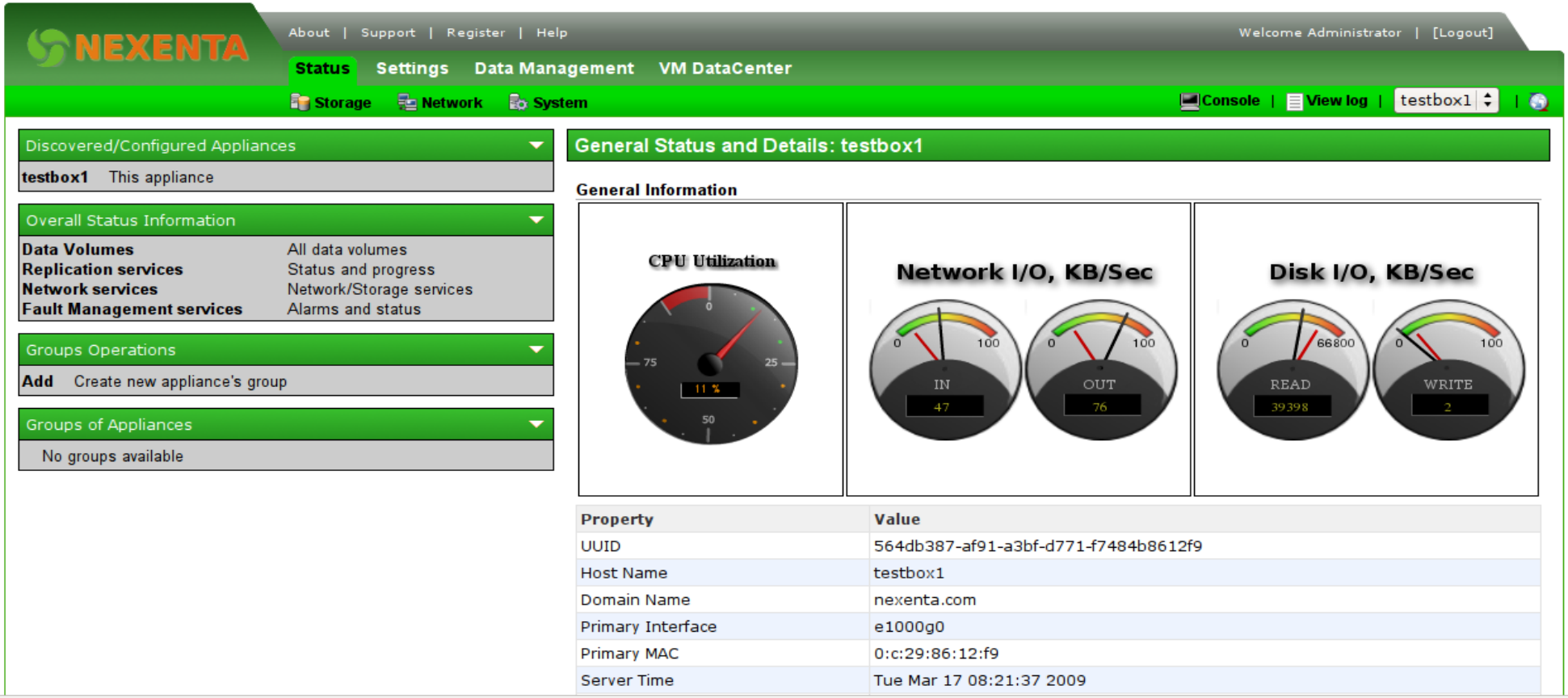
オープンソース

NexentaOS:

Loves multiple cores
Boot レベル ZFS
>ダウンロード数1,000,000

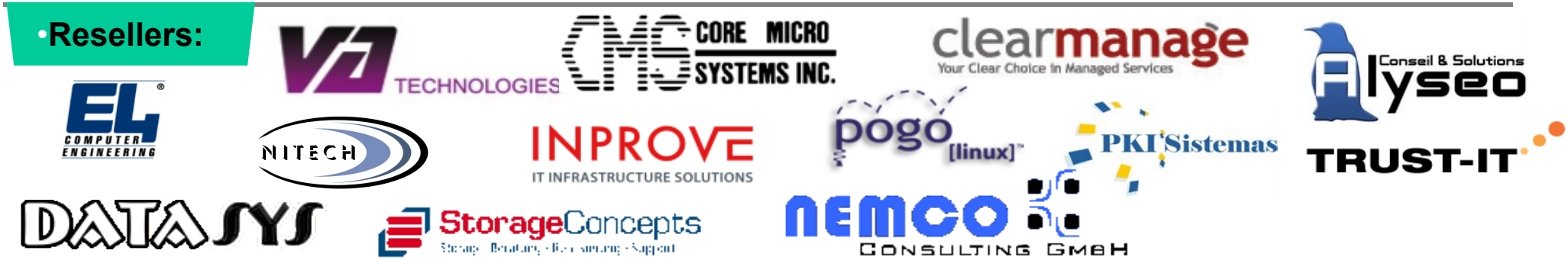
ZFS：ファイルシステム
OpenSolaris：カーネル

管理画面イメージ



顧客とパートナー

•Resellers:

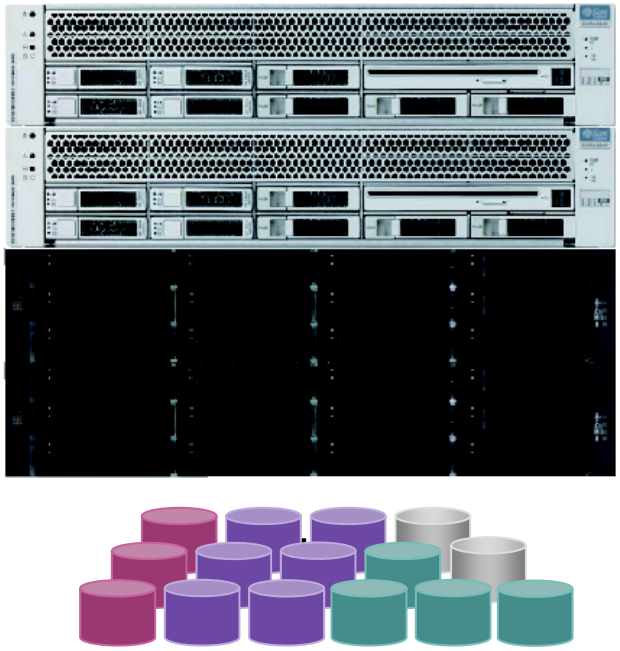


•Partners:

ZFS の特徴

従来のファイルシステムとの違い

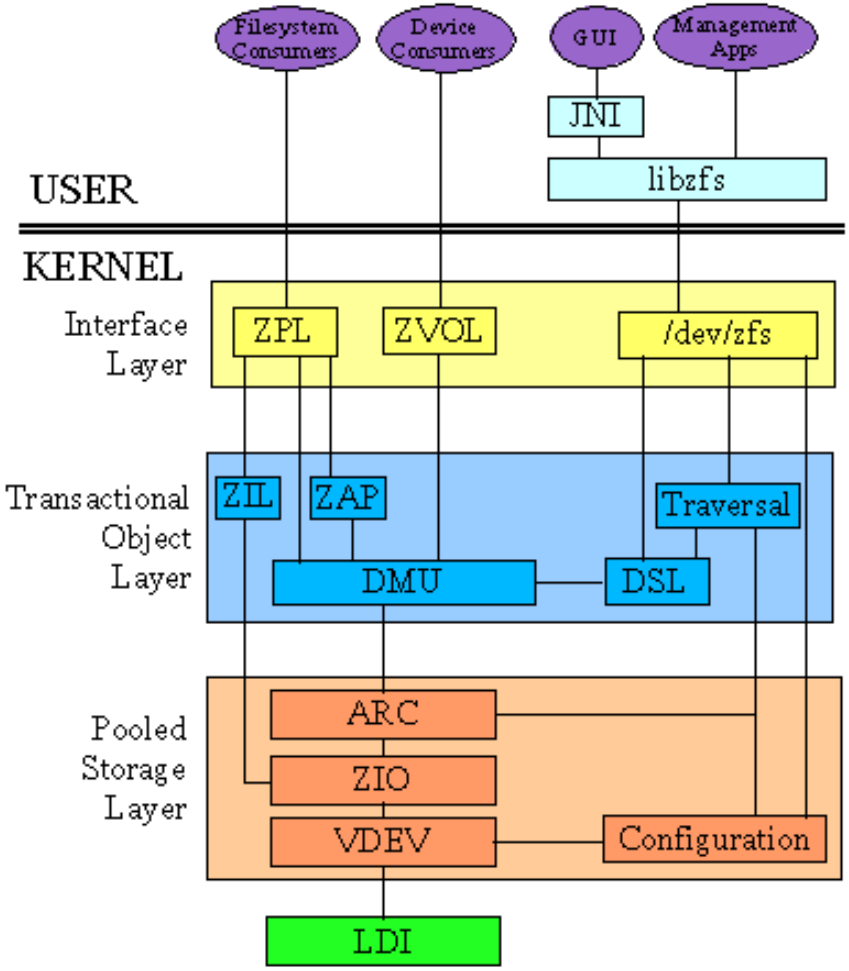
1. ストレージプール
   - 物理デバイス固定割り当てからの解放
   - ハードディスク、ストレージユニットを追加するだけで使用領域が増える
   - ストライプ幅自動調整 RAID
   - 管理の容易さ
2. トランザクションファイルシステム
   - 常にデータの一貫性が保たれる
   - 障害耐性が高い
3. 自己修復データ保護機構
   - チェックサム
4. スケーラビリティ
   - 128 bit ファイルシステム
5. ポータビリティ



Storage Pool

ZFS の構造

1. Interface Layer
   - POSIX インターフェース
   - ボリュームインターフェース
2. Transaction Object Layer
   - DMU オブジェクト
   - トランザクション
   - ZIL
   - スナップショット
3. Pooled Storage Layer
   - I/O パイプライン
   - キャッシュ (ARC)
   - ボリュームマネージャ

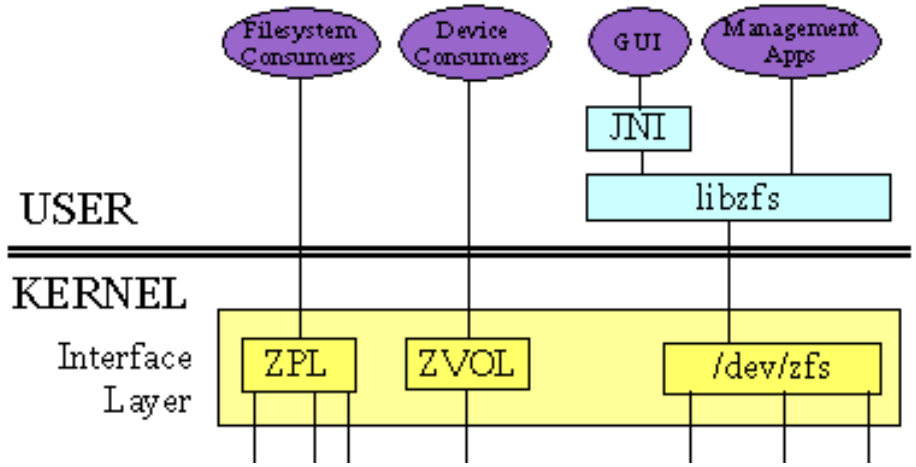


Interface Layer

1. ZPL (ZFS POSIX Layer)
   - VFS インターフェース (mount、umount)
   - vnode インターフェース (open()、read()、write()、mmap() 、 fsync())
   - ACL

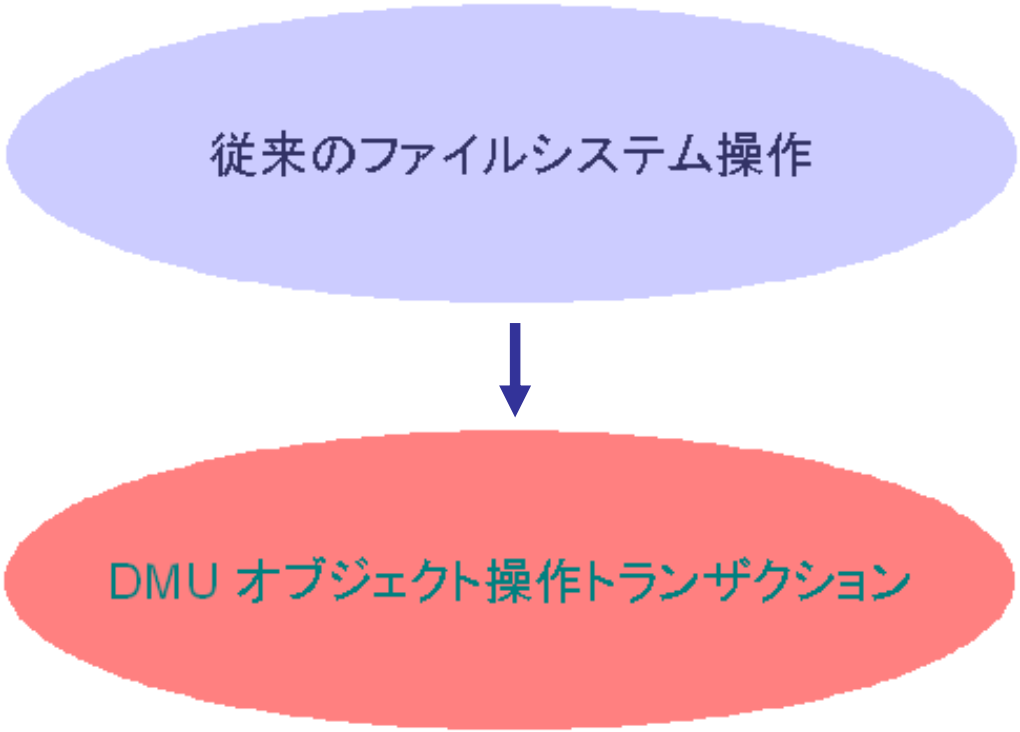
2. ZVOL (ZFS Volume)
   - ボリュームインターフェース (/dev/zvol/[r]dsk/….)
   - ZPL とは別に vnodeops を定義

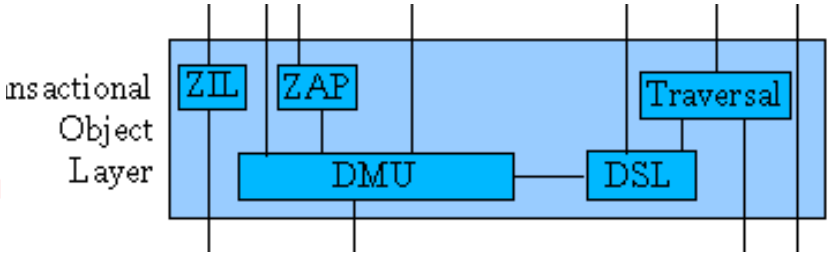
3. /dev/zfs
   - ioctl() による ZFS の制御



Transaction Object Layer

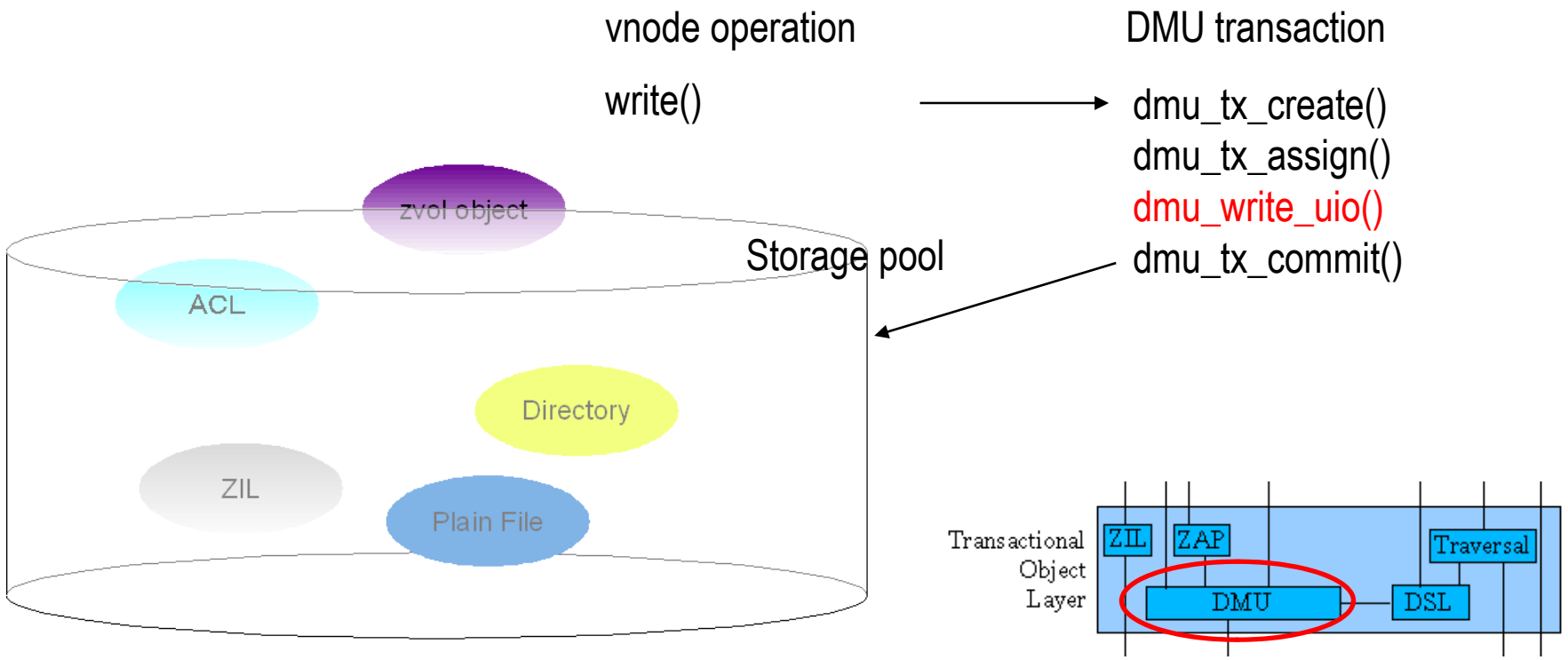
1. オブジェクトベースのファイルシステム
   - メタデータとデータをオブジェクトとして扱う
     - ファイル、ディレクトリ、ACL、ZIL





Transaction Object Layer

1. インターフェースレイヤに DMU オブジェクト操作を提供
   - トランザクション管理
   - DMU オブジェクト操作 (ZIO へマップ)



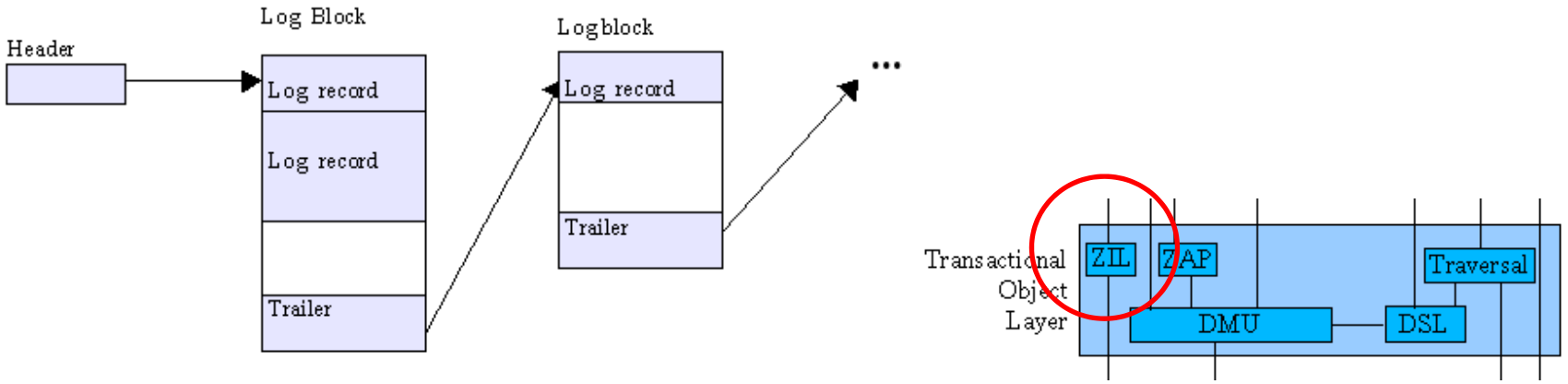
Transaction Object Layer

1. DMU オブジェクト例

```
# zdb -v tank/test
Dataset tank/test [ZPL], ID 34, cr_txg 501, 1.56G, 7 objects

   Object  lvl   iblk   dblk  lsize  asize  type
        0    7    16K    16K    16K  15.0K  DMU dnode
        1    1    16K    512    512     1K  ZFS master node
        2    1    16K    512    512     1K  ZFS delete queue
        3    1    16K    512    512     1K  ZFS directory
        4    1    16K    512    512     1K  ZFS directory
        5    3    16K   128K   800M   800M  ZFS plain file

# zdb -v tank/vol
Dataset tank/vol [ZVOL], ID 40, cr_txg 541, 802M, 3 objects

   Object  lvl   iblk   dblk  lsize  asize  type
        0    7    16K    16K    16K  14.0K  DMU dnode
        1    4    16K     8K   800M   802M  zvol object
        2    1    16K    512    512     1K  zvol prop
```

Transaction Object Layer

1. ZIL (ZFS Intent Log)
   - ログとしてシーケンシャルに書き込む (ジャーナルログ)
   - 専用の高速ログデバイスを指定可能
     - フラッシュディスク
     - ミラー構成
   - データベースなど同期書き込みが多いアプリケーション向け

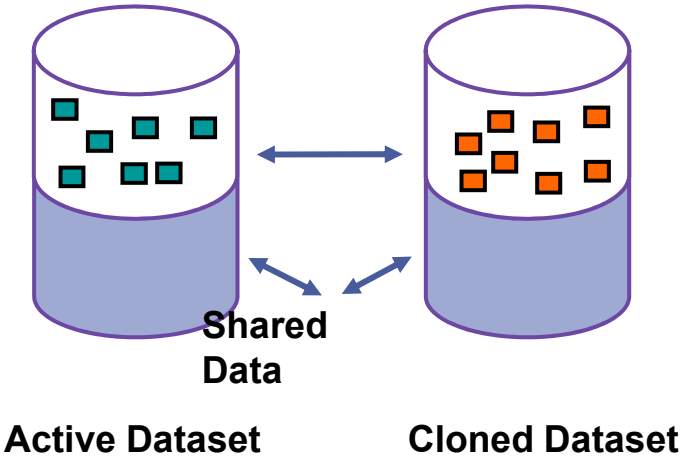


Transaction Object Layer

1. トランザクション
   - txg_time 以内に vdev へ反映
   - 障害発生時のトランザクションは破棄される (チェックサム)
   - システムコールと一対一に対応しない (大きな write() は分割される)
   - メタ構造とデータの対応について常に一貫性が保たれている
     - アプリケーションのデーター貫性が保護されるという意味ではない
     - fsck 不要？ → それでも整合性チェックはした方が良い - scrub

Transaction Object Layer

1. スナップショット
   - 静止点を記録するのみ
   - パフォーマンスへの影響がない
   - 無駄に容量が増えない (COW)

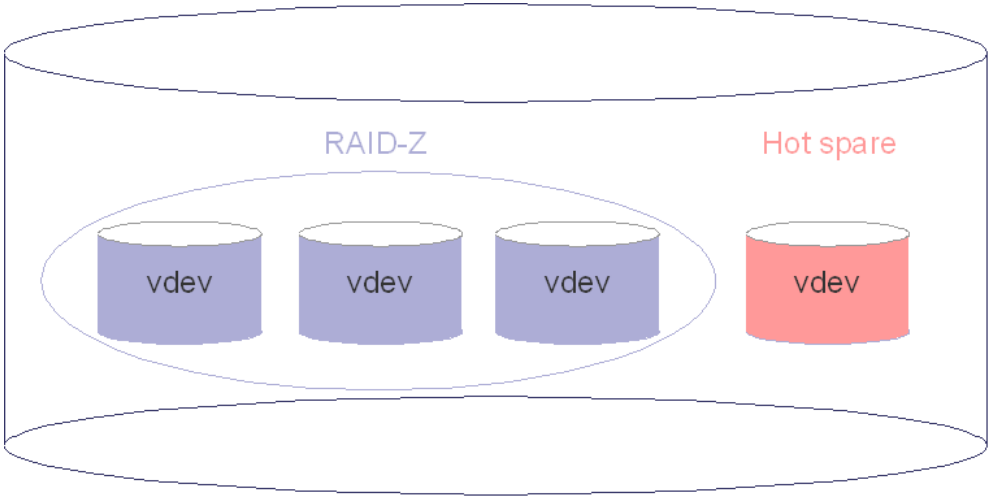


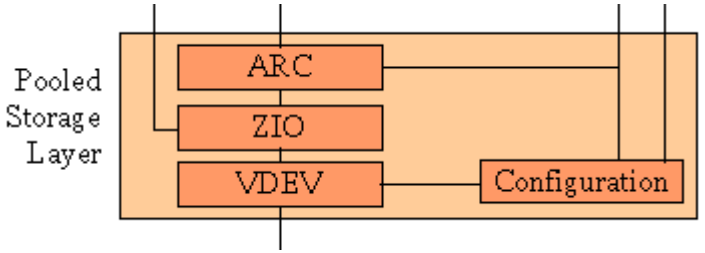
2. スナップショット操作
   - ロールバック
   - 書き込み可能なクローンとしてマウント

```
# zfs snapshot tank@snap1
# zfs list -t snapshot
NAME                    USED  AVAIL  REFER  MOUNTPOINT
tank@snap1                 0      -    19K  -
```

Pooled Storage Layer

1. vdev に依存しないフラットアドレス空間を提供
2. vdev – ハードディスク、ファイル (通常は rdsk)
3. キャッシュ (ARC)
4. I/O パイプライン (ZIO)
5. ボリュームマネージャ (VDEV : ストライプ、ミラー、RAID 5/6)





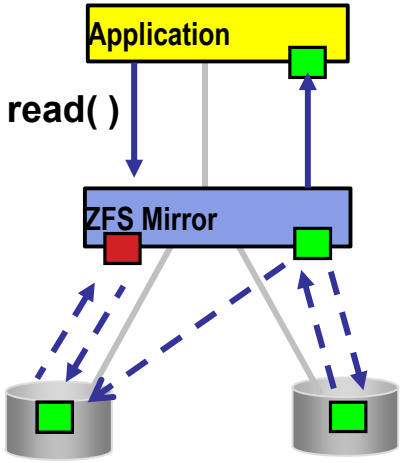
Pooled Storage Layer

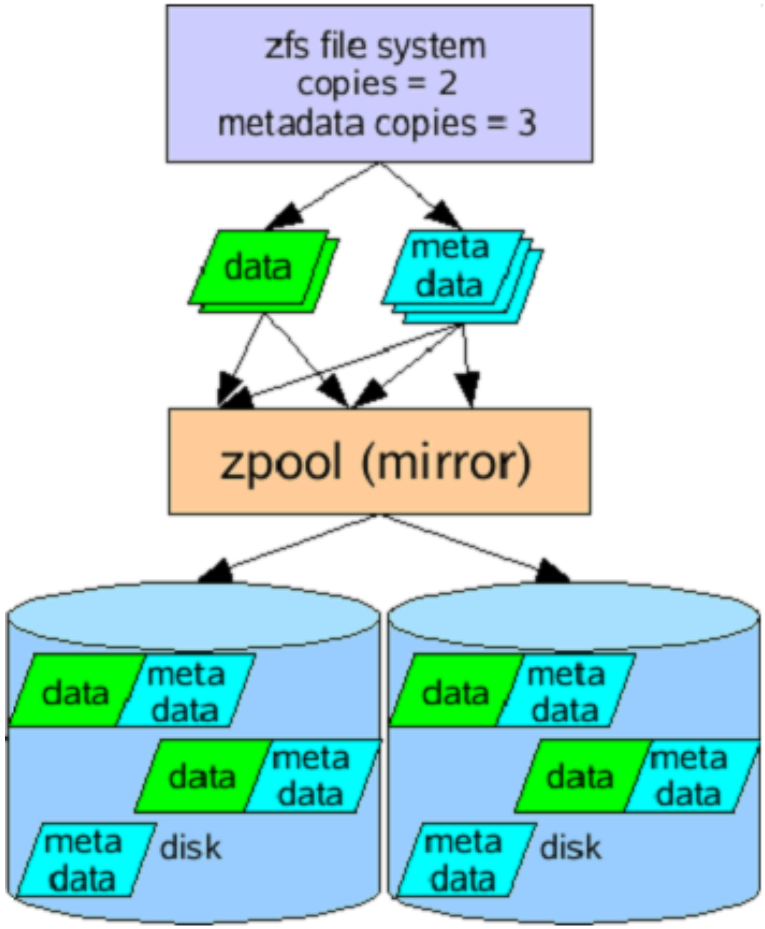
1. プールの作成

```
# zpool create tank mirror c1d1 c2d0 spare c2d1
# zpool status tank
pool: tank
 state: ONLINE
 scrub: none requested
config:
        NAME        STATE     READ WRITE CKSUM
        tank        ONLINE       0     0     0
          mirror    ONLINE       0     0     0
            c1d1    ONLINE       0     0     0
            c2d0    ONLINE       0     0     0
        spares
          c2d1      AVAIL
errors: No known data errors
```

Pooled Storage Layer

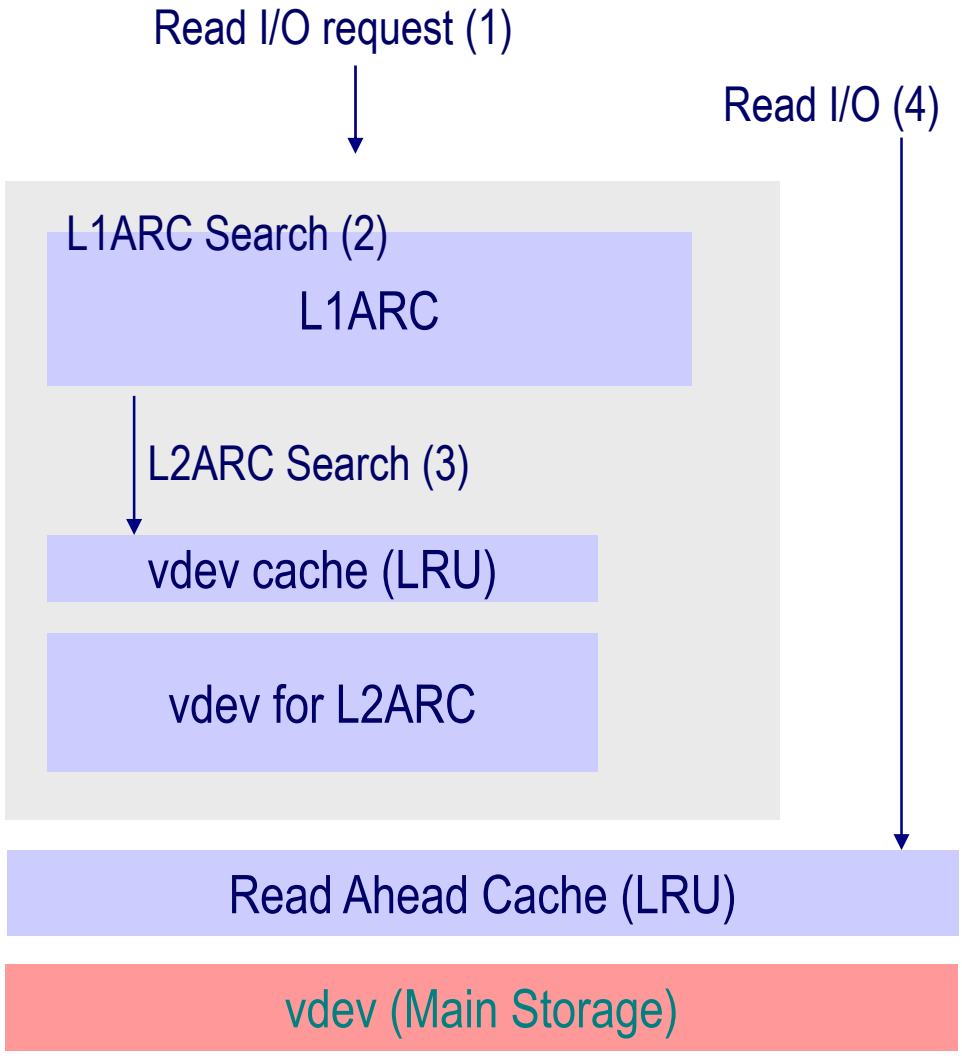
1. 冗長構成によるデータ保護
   - メタデータは n + 1
   - データは指定数だけ冗長化
   - copies = n はディスク障害を救済しない
     - mirror、RAID-Z を使う
2. チェックサム
   - fletcher2、fletcher4、SHA256





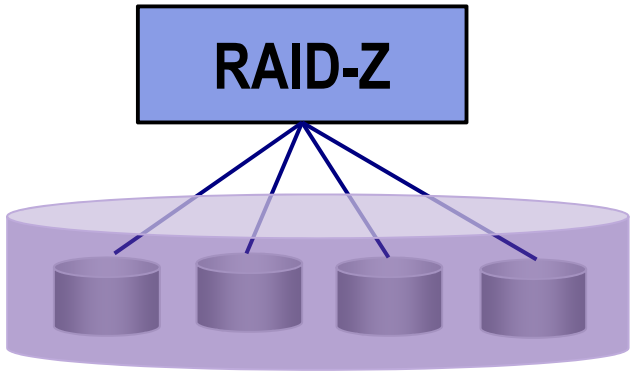
Pooled Storage Layer

1. 読み込み キャッシュ
   - メインメモリを可能な限り使用
     - Read ahead cache (10 MB)
     - Adaptive Replacement Cache
   - L2ARC – 専用外部デバイス



Pooled Storage Layer

1. サポートする RAID 構成
   - RAID 0
   - RAID 1 (mirror)
   - RAID 5 (RAID-Z)
   - RAID 6 (RAID-Z2)
   - RAID 7 (RAID-Z3 – NexentaStor 次期バージョン対応予定)
   - ホットスペア
2. RAID-Z
   - 可変ストライプ幅
3. チェックサムによる自己データ修復
   - 静かなデータ崩壊 (Silent Corruption) 抑制



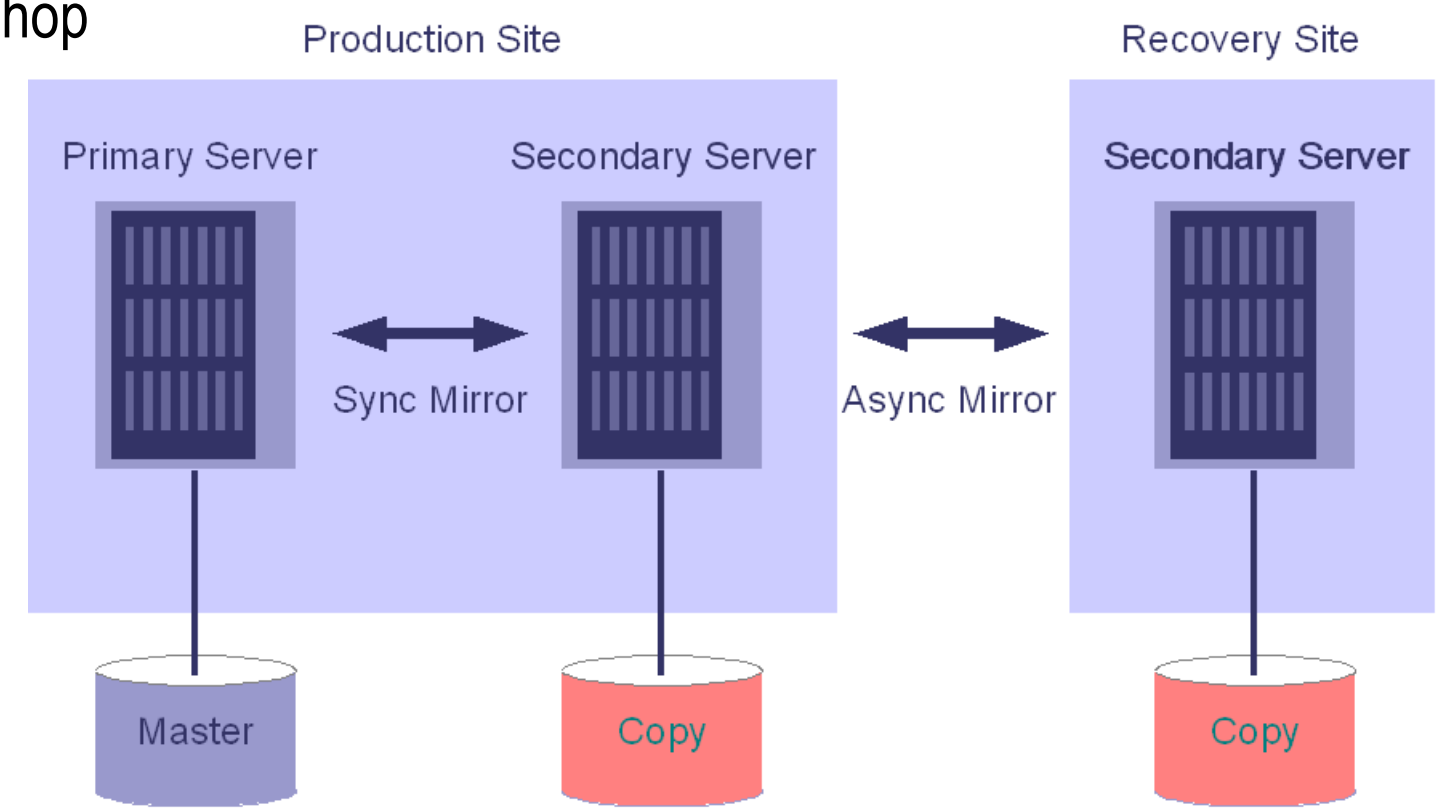
ポータビリティ (ベンダロックインからの解放)

1. オープン仕様
   - お客様のデータはお客様のもの
2. スナップショットの送受信 (send/receive)
   - 標準入出力
   - SSH
   - 遠隔バックアップ
3. 物理デバイスの移動
   - エクスポート・インポート
   - 適応型エンディアン制御
4. オープンソース
   - FreeBSD (10 日でポーティング)
   - Linux (FUSE)

NexentaStor プラグイン

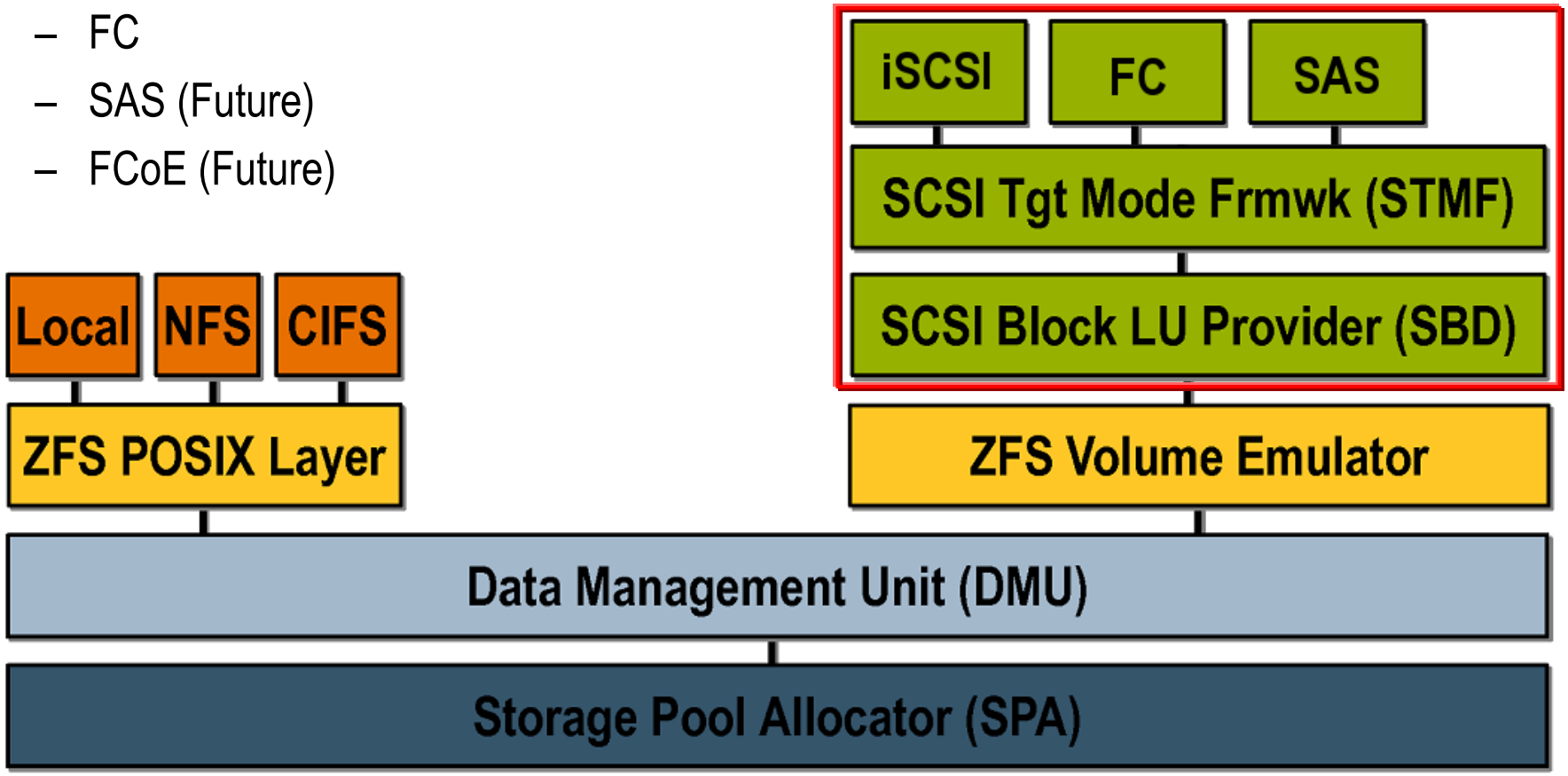
Auto-CDP

1. ブロックレベルレプリケーションソフトウェア
2. Sync / Async
3. Forward / Reverse
4. Multihop



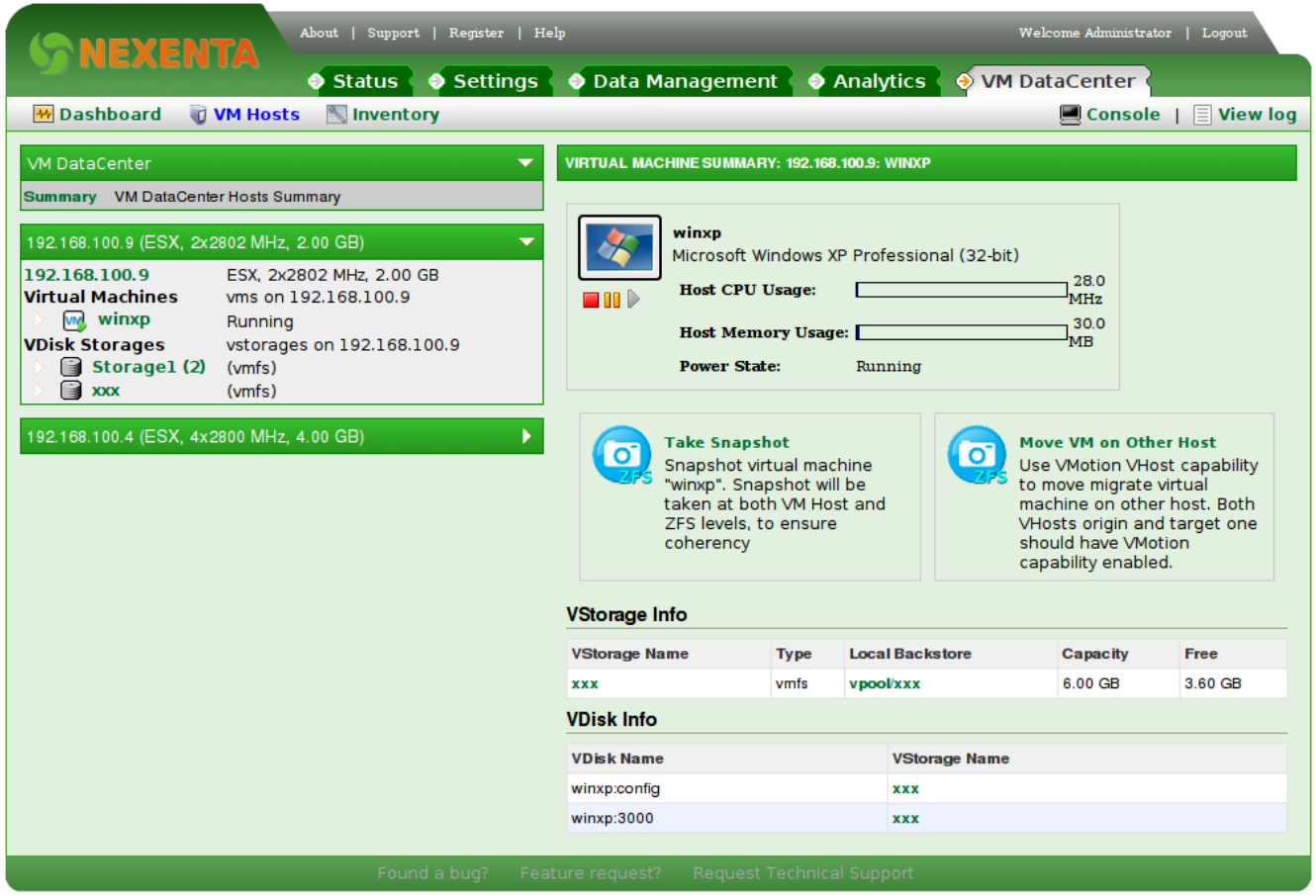
COMSTAR (Target 2.0)

1. SCSI ターゲット機能
   - iSCSI
   - FC
   - SAS (Future)
   - FCoE (Future)

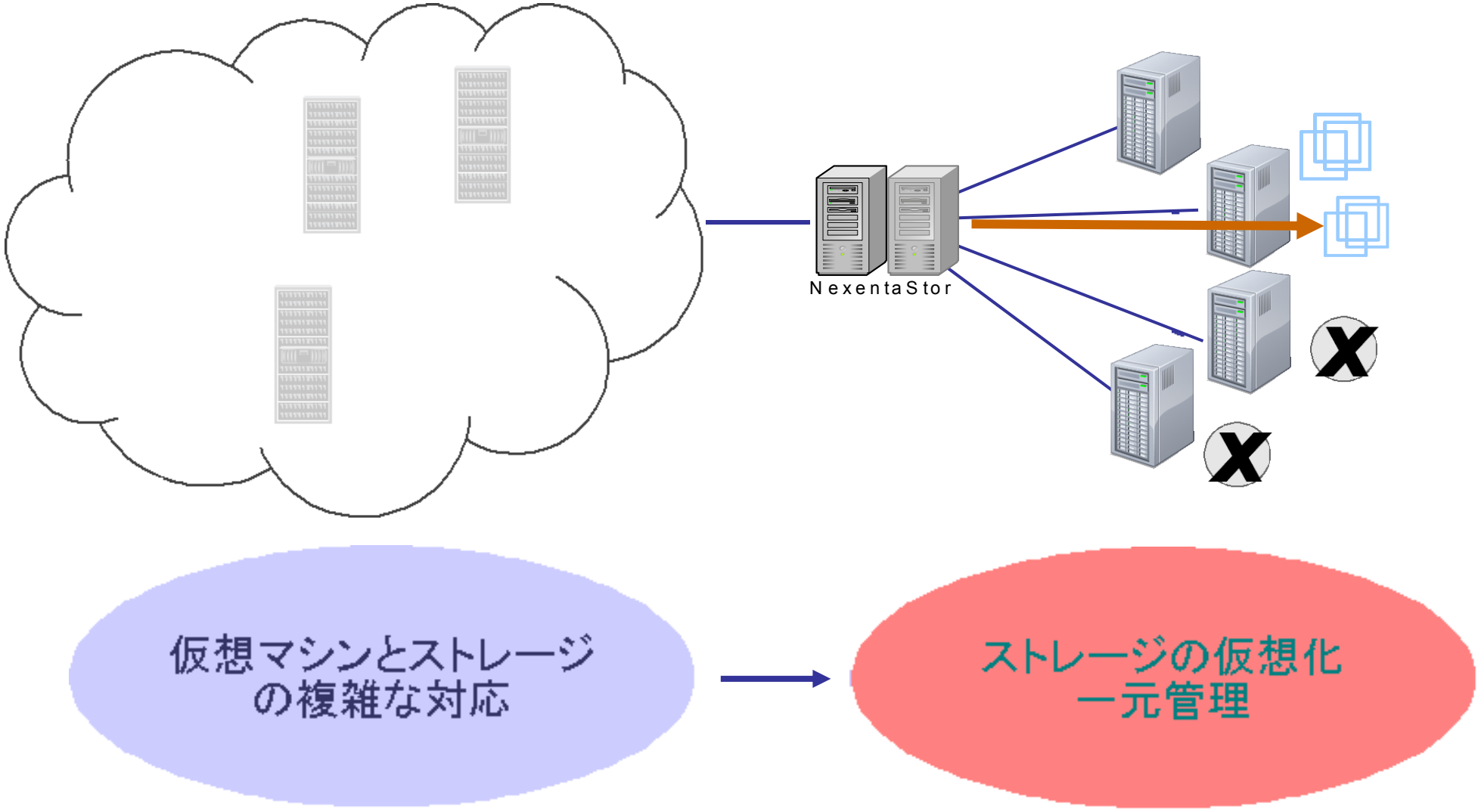


VM Data Center

1. 仮想環境向けストレージサービス統合管理機能
   - VMWare
   - Xen
   - Hyper-V

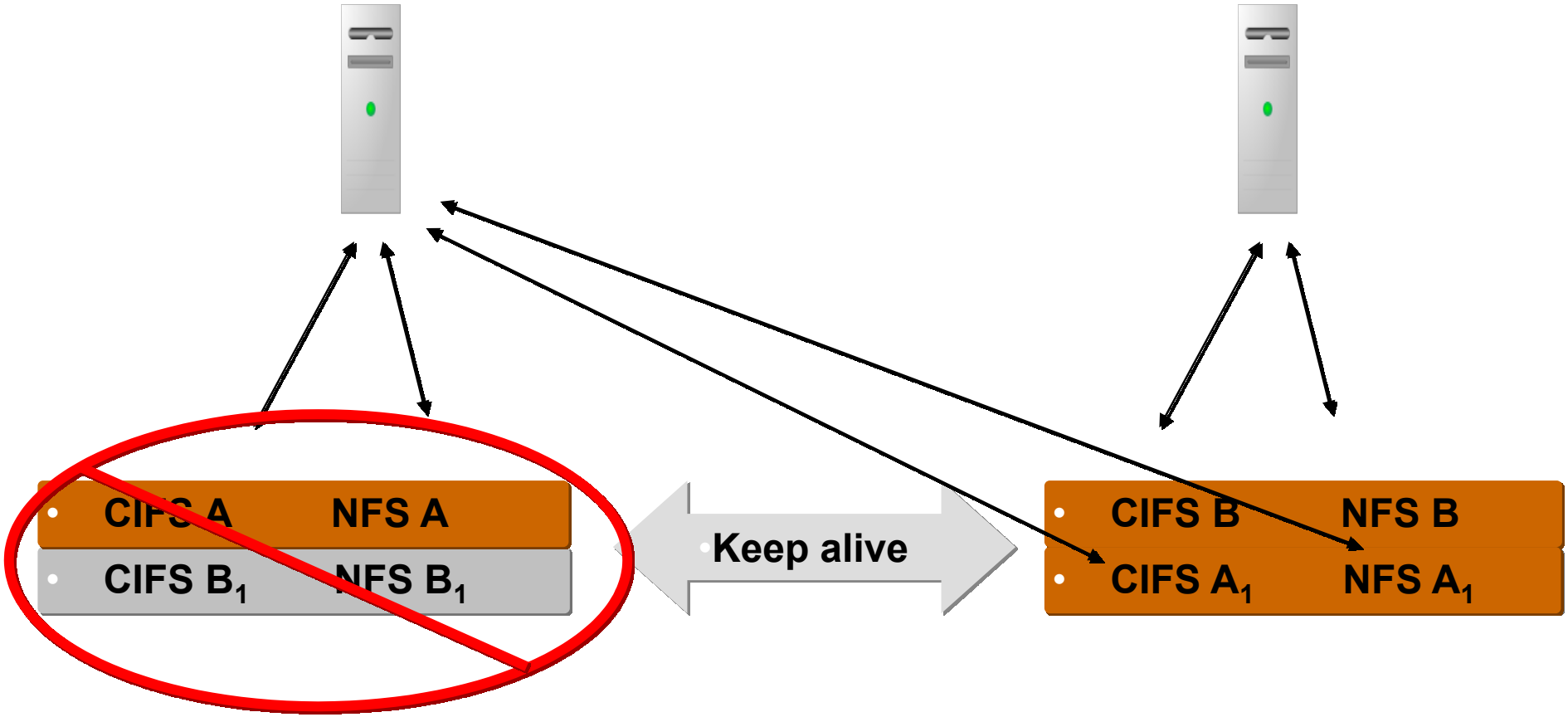


VM Data Center



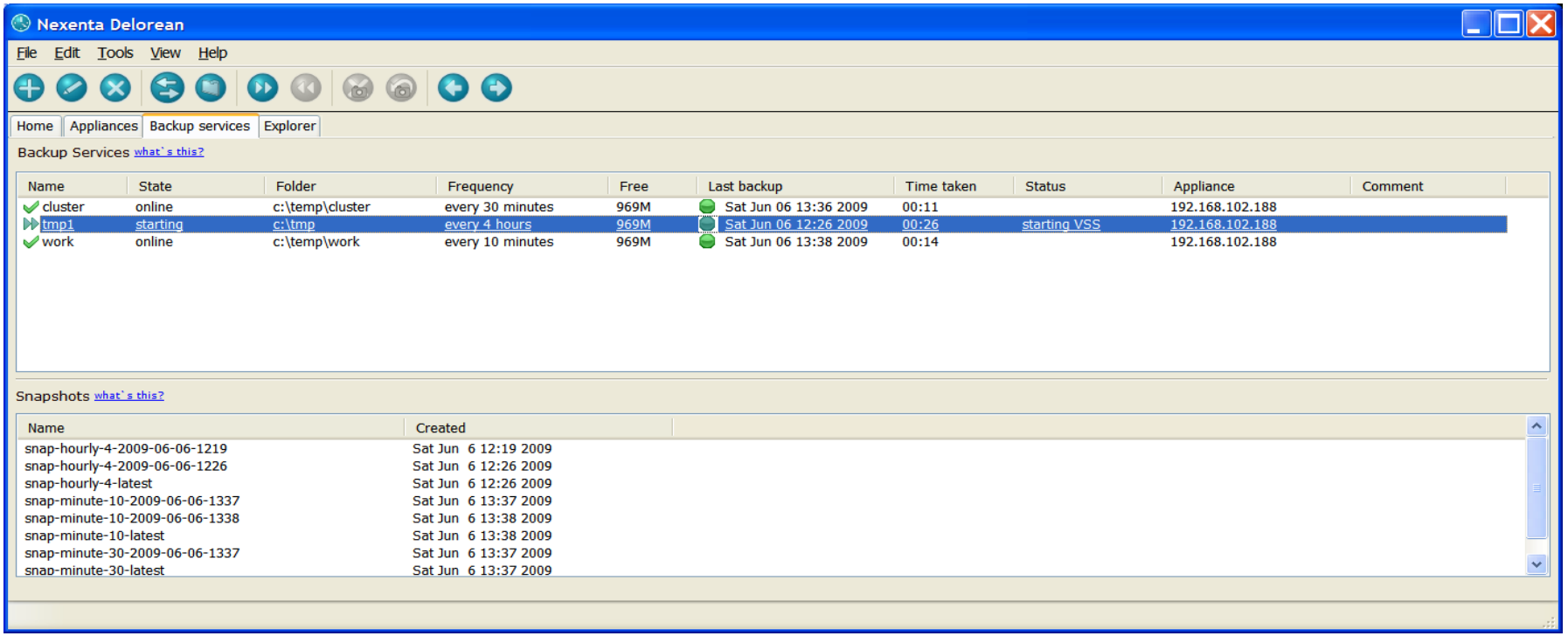
HA Cluster 1.0

1. HA ソフトウェア
   - High-Availability.Com 社 RSF-1 ベース



Delorean

1. Windows ホストの CDP バックアップソフトウェア
2. VSS 対応
3. Windows エクスプローラ風操作 GUI



機能比較

機能比較

Features	NexentaStor	N社同クラス製品	E社同クラス製品
HA Cluster	有り (ファイルサービスのみ)	有り	有り
Protocols	NFS v3/v4, CIFS, iSCSI, HTTP, FTP, NDMP, WebDAV		
Hybrid Storage Pool	有り	無し	無し
重複排除	実装される予定	有り	有り
データ圧縮	有り	無し	無し
スケーラビリティ	128-bit (Zettabyte)	64-bit	64-bit
RAID	有り	有り	有り
シンプロビジョニング	有り	有り	有り
スナップショット	有り	有り	有り
スナップショットリストア	有り	有り (オプション)	有り (オプション)
非同期ミラー	有り (オプション)	有り (オプション)	有り (オプション)
クローン	有り	有り (オプション)	無し
リモートアーカイビング	有り	有り (オプション)	有り (オプション)
電子メール通知	有り	有り	有り
データ検索	有り (MS Office、PDF etc..)	無し	無し

ロードマップ

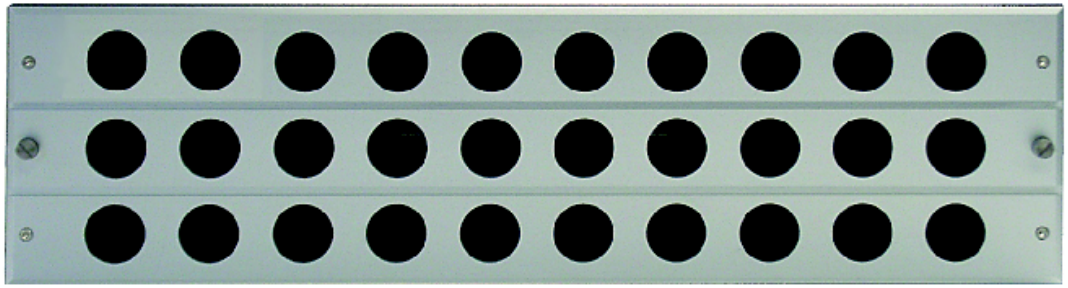
NexentaStor 2.2

1. ウィルスチェック
2. 多言語対応 (NMV)
3. VMDC
   - Xen Live Motion
   - FC (現行 VMDC は iSCSI のみ)
4. 将来の機能
   - 重複排除
   - pNFS
   - SAS、FCoE ターゲット
   - RAID-Z3

Prime STOR ZFS

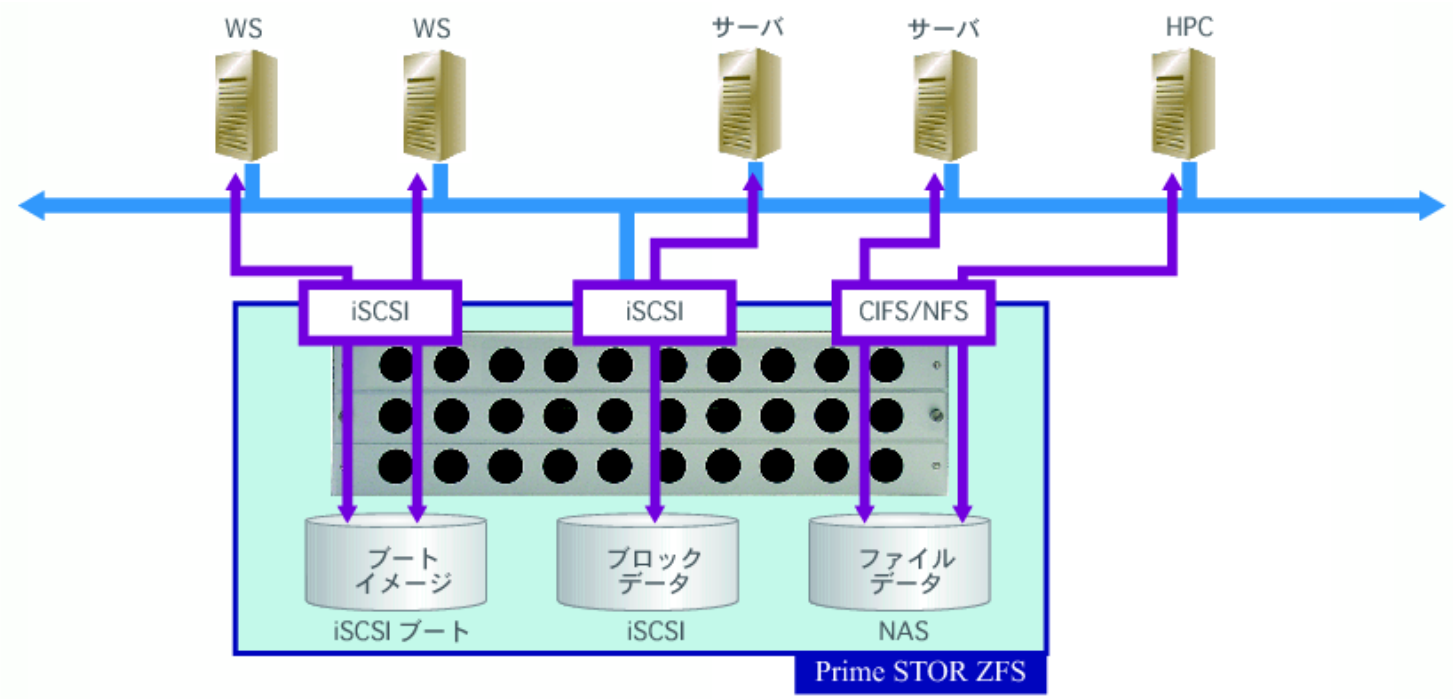
大規模同時アクセスに応える先進のユニファイドストレージ

- トランザクションI/O を最適化する階層キャッシュ構造
- スケーラブルストレージプール
- シンプロビジョニング
- 次世代 10 GbE 対応 iSCSI / NAS ユニファイド I/O



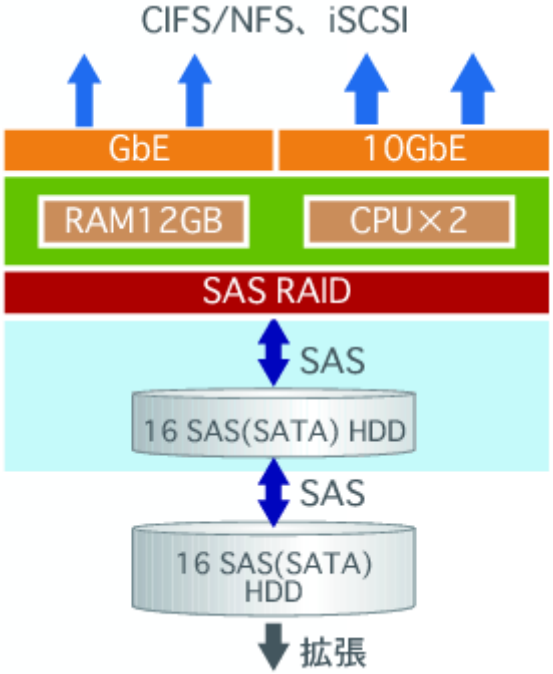
Prime STOR ZFS

1. 用途
   - 仮想サーバ / 仮想マシン用
   - EDA / CAE 大規模ワークグループ
   - メディカルイメージング / CG レンダリング



Prime STOR ZFS

1. DRAM / SSD / HDD を階層化
2. 書き込みキャッシュ (DRAM / SSD オプション)
3. 読み出しキャッシュ (DRAM /SSD オプション)
4. レプリケーションソフトウェアオプション



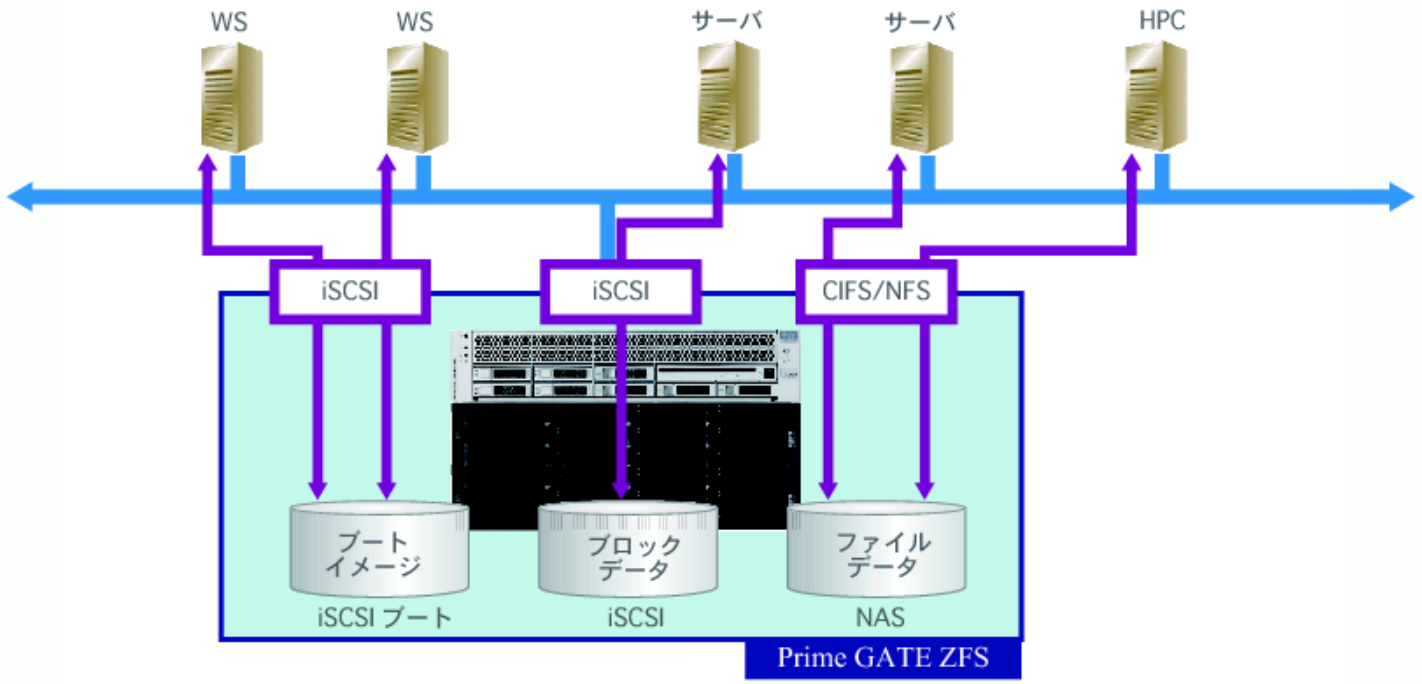
Prime GATE ZFS

大規模同時アクセスに応える先進のストレージ仮想化ゲートウェイ

- トランザクションI/O を最適化する階層ストレージアーキテクチャ
- ダイナミック＆シンプロビジョニング仮想化ストレージプール
- iSCSI 及び FC SAN 及び NAS ユニファイド I/O サービス

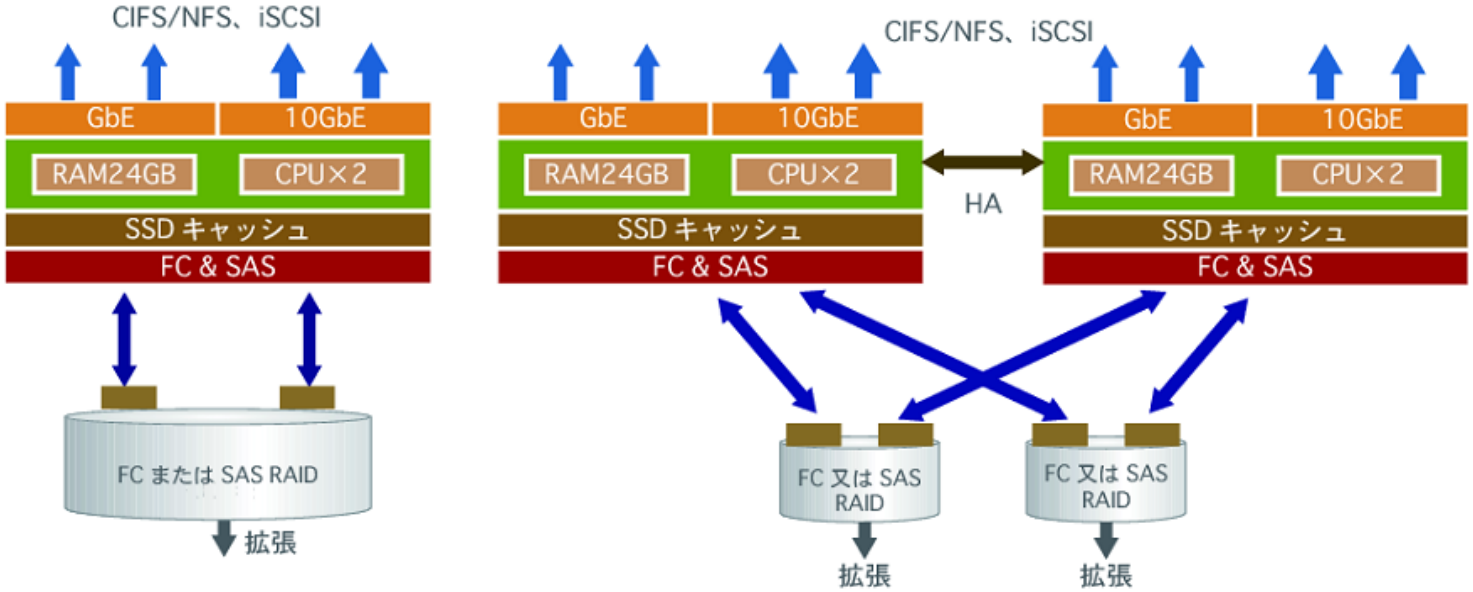
Prime GATE ZFS

1. 用途
   - エンタープライズ
   - データセンター
   - 仮想サーバ用バックストレージ



Prime GATE ZFS

1. DRAM / SSD / HDD を階層化
2. 書き込みキャッシュ (DRAM / SSD)
3. 読み出しキャッシュ (DRAM /SSD)
4. HA オプション
5. レプリケーションソフトウェアオプション



コアマイクロシステムズ株式会社

Core Micro Systems, Inc.

URL：http://www.cmsinc.co.jp/　Mail：sales@cmsinc.co.jp

TEL：03-5917-6451　IP Phone：050-5558-5410　FAX 03-5917-6452

本社 〒173-0026 東京都板橋区中丸町11-2 ワコーレ要町ビル9F