Panasasストレージクラスタ アーキテクチャ概要

スケーラブルシステムズ株式会社

Panasasストレージクラスタ

ストレージシステムの課題

ストレージシステムの課題

- ストレージアーキテクチャの課題
  - 分散・パラレルへの対応
- ITインフラでの課題
  - クラスタ環境でのストレージシステム
  - ボトルネック
- ワークフローでの課題
  - Windows/Linux環境
  - バッチ・対話処理
- データセンターでの課題
  - スケーラビリティと運用管理

ストレージアーキテクチャの基盤

ストレージアーキテクチャの基盤の変化

従来のストレージ	新しい基盤技術	利点
一体型(Monolithic) 集中・中央管理 (Central)	分散・パラレル (Distributed)	・スケーラビリティ ・クライアント（サーバ）が直接ストレージにアクセス可能 ・メタデータを管理することで、広範囲なデータコピーを避ける
ブロックベース (Block Base)	オブジェクトベース (Object Base)	・デバイス内でオブジェクトのレイアウトを管理

ストレージに関する課題

クライアント（エンドユーザ）

クライアント
- ジョブの実行終了を待つ
- ユーザ数が増えた場合のスケーラビリティの問題
- ユーザ間でのコラボレーションやデータの共有の問題

クラスタ
- 計算クラスタはI/O処理の終了まで計算を中断
- I/O処理は、クラスタの利用率の低下を引き起こす
- ノード数を増やした場合のスケーラビリティの維持の問題

BOTTLENECK

クラスタ

BOTTLENECK

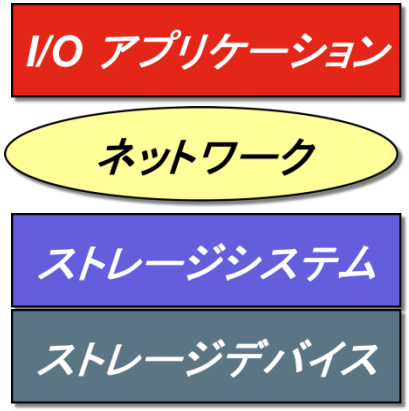
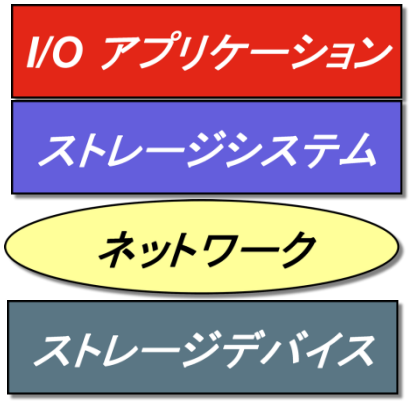
従来のネットワークストレージ

BOTTLENECK

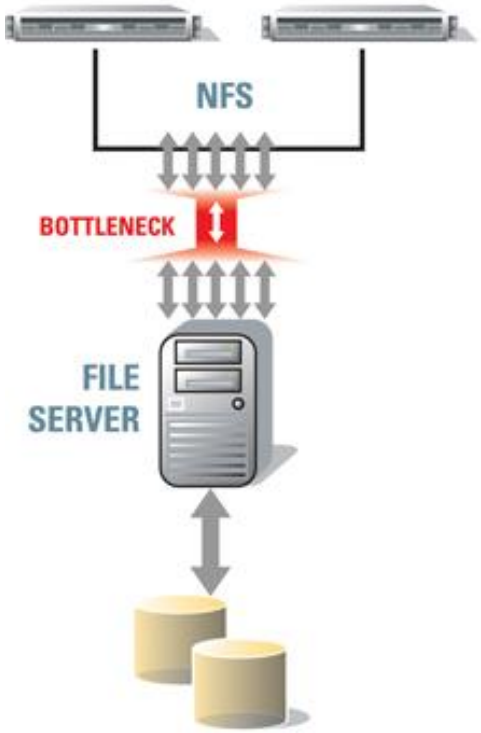
バックアップ/リストア

バックアップ/リストア
- バックアップ処理のためのストレージシステムの負担
- バックアップ実施のタイミング
- 高速でのバックアップの問題

ストレージアーキテクチャ

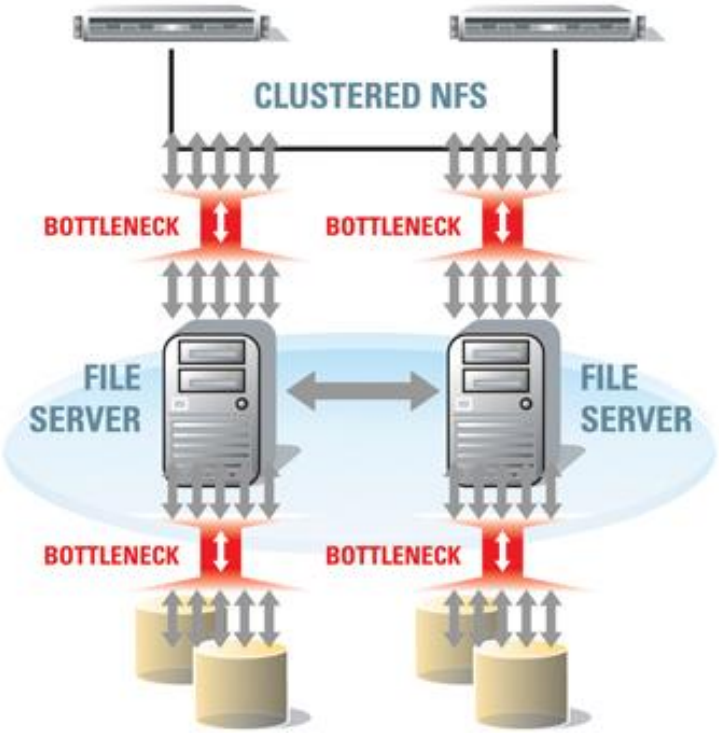
	DAS (Direct Attached Storage)	NAS (Network Attached Storage)	SAN (Storage Area Network)
システム構成図			
利点	高い性能	複数のホスト間でのデータ共有 ファイルシステムをオフロード可能	スケーラブルな性能と容量
課題	ホスト（サーバ）間でのデータ共有の制限	性能とスケーラビリティ ホスト（クライアント）数での制限	コスト

NFS(ネットワークファイルサーバ)



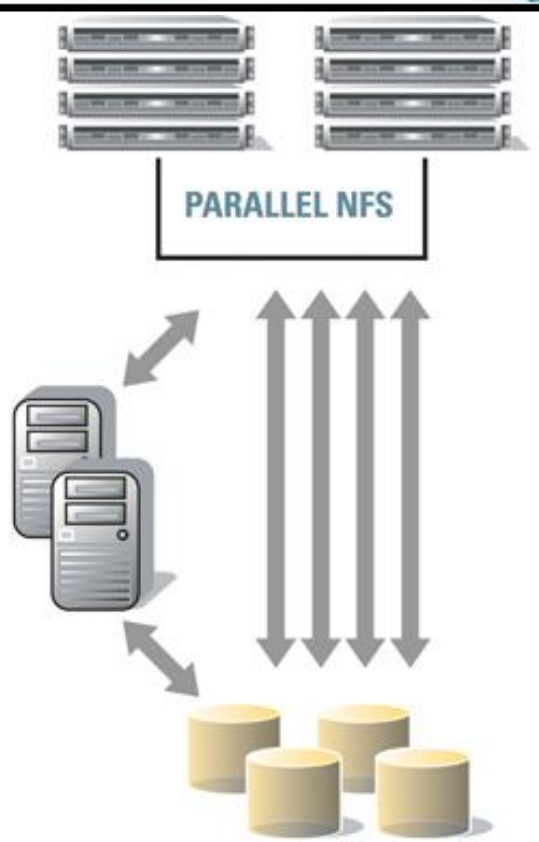
NAS

Network Attached Storage
シリアルI/Oがボトルネック



CLUSTERED STORAGE

複数のNASを統合的に運用管理
個々のNASサーバでのシリアルI/O
がボトルネック



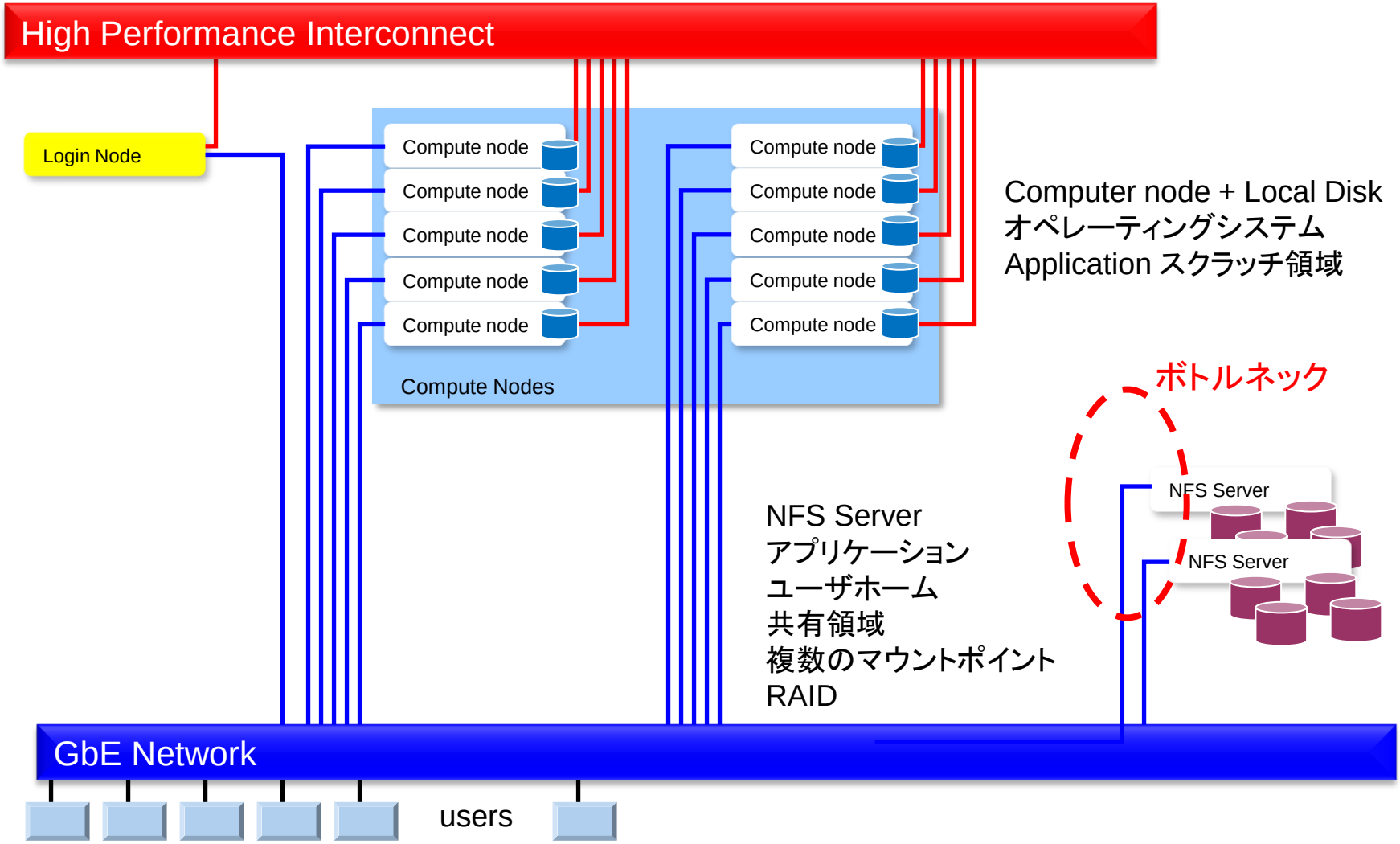
PARALLEL STORAGE

ファイルサーバを経由しないデータ
転送パス
シリアルI/Oのボトルネックの解消と
容易なシステム全体の運用管理

NFS(ネットワークファイルサーバ)

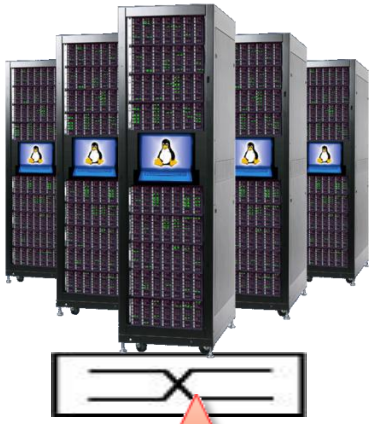
機能	利点	従来型NFS	クラスタNFS	パラレルNFS
直ぐに利用可能なアプライアンス	容易なインストール	✔	✔	✔
統合された管理インターフェイス	システムの運用管理をシンプルに実現	✔	✔	✔
既存のバックアップインフラとの互換性	ワークフローの変更を必要としない	✔	✔	✔
ボリュームのフェイルオーバーとネットワークの冗長性	データの可用性	✔	✔	✔
グローバル共有ファイルシステムでのスケーラビリティ（ペタバイト以上）	低いシステムマネージメントでのオーバーヘッド		✔	✔
数十ギガバイト/秒のトータルバンド幅と十万IOPの性能	要求されるワークフローに対するスケーラブルな性能の提供と多くのクライアントからのアクセス時の小規模ファイルへのアクセス性能		✔	✔
大規模なファイルへのクライアントからの同時パラレルアクセスが可能	大規模ファイルの処理におけるパラレルI/Oの利用による大幅な性能向上		✔	✔
ペタバイト以上の容量までのグローバル共有ファイルシステムのスケーラビリティ	10TBクラスから大規模システムまで低いシステムの管理オーバヘッドを提供			✔
トータルバンド幅のスケーラビリティ(50ギガバイト以上)	大部分のアプリケーションにおいて最高の性能を提供可能			✔
大容量ドライブでのリカバリー出来ない読み込みエラーに対する対応	大容量ドライブにおける信頼性の向上			✔

クラスタシステム

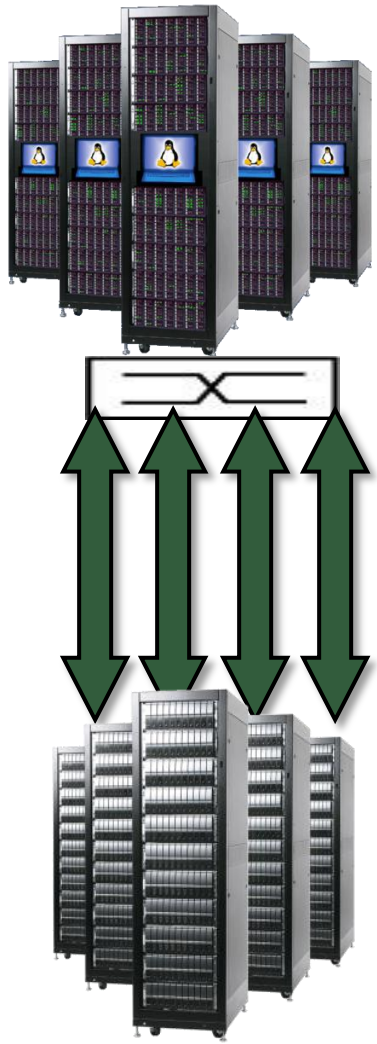

High Performance Interconnect
Login Node
Compute node
Compute node
Compute node
Compute node
Compute node
Compute node
Compute node
Compute node
Compute node
Compute node
Compute Nodes
Computer node + Local Disk
オペレーティングシステム
Application スクラッチ領域
ボトルネック
NFS Server
NFS Server
NFS Server
アプリケーション
ユーザホーム
共有領域
複数のマウントポイント
RAID
GbE Network
users

クラスタ利用時のボトルネック

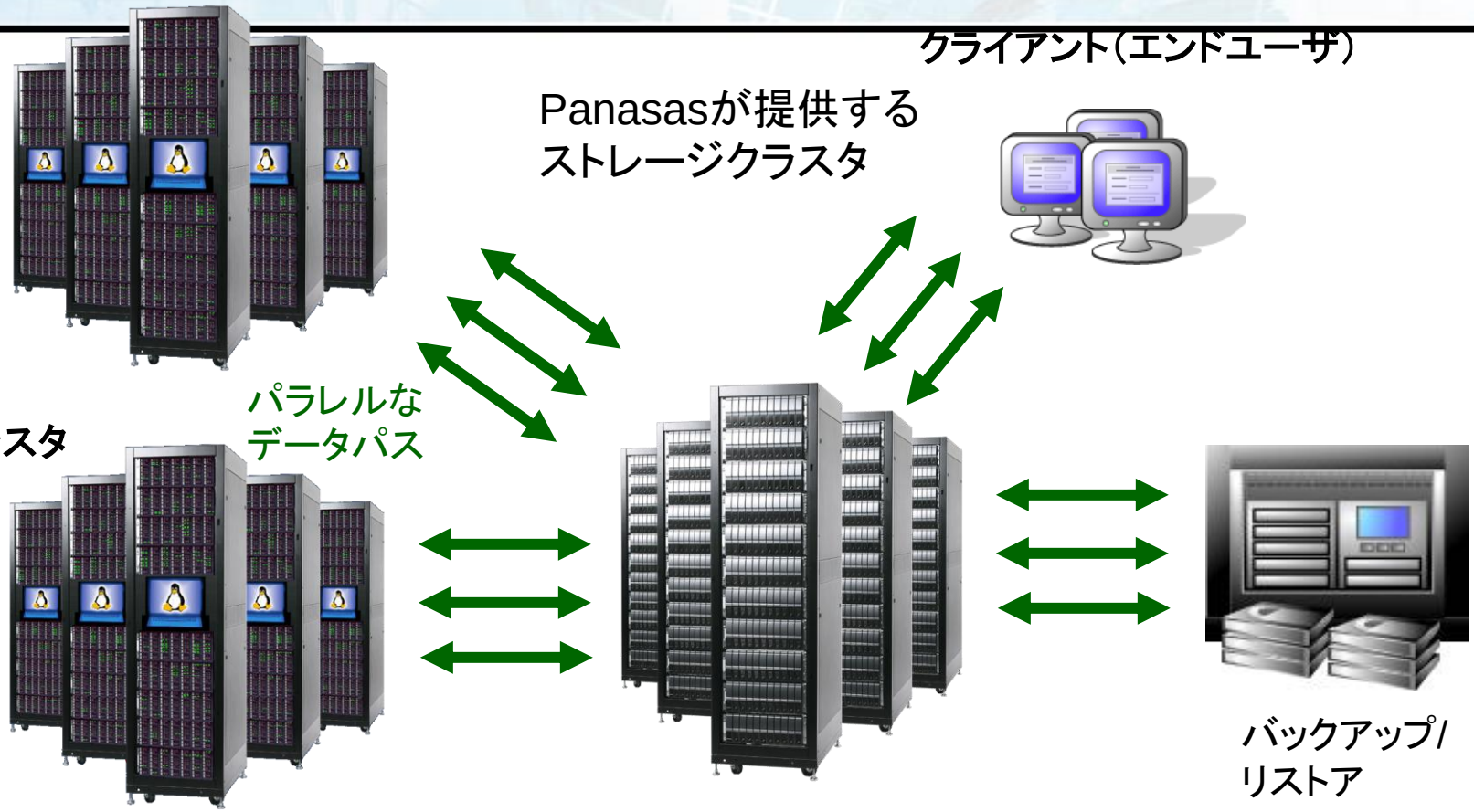
クラスタ⇒パラレルコンピューティング⇒ パラレルIOが必要



一般的なNAS	パラレルストレージ
ストレージに対する単一のデータパス	ストレージに対するパラレルなデータパス
スケーリング	スケーラビリティ
限定されたIOバンド幅	高いIOバンド幅
システム拡張の限界	グローバルネームスペース
柔軟性の欠如	容易な運用管理
高価なシステム構成	低価格



Panasasによるソリューション
クラスタでのボトルネックの解決



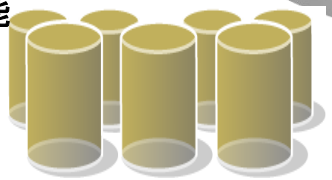
- パラレル分散ファイルシステム
- グローバルネームスペース によるシステムの拡張時の容易な運用管理の実現
- オブジェクトベースストレージによるスケーラブルな性能を実現するアーキテクチャ
- インテリジェントなハードウエアアーキテクチャによる容易なシステム実装と拡張性
- 統合されたストレージソリューション と一般商用製品による低コスト

ワークフローの効率化
ストレージに対する要求

対話処理

ユーザの要求
大容量ドライブは不要
小ー中規模のファイル
ランダムなファイルアクセス
高いIOスループット
高いバンド幅
高い可用性
スナップショット機能

“Run, Evaluate, Re-Run”



NASファイルシステム

“Run & Done”

SANファイルシステム

バッチ処理

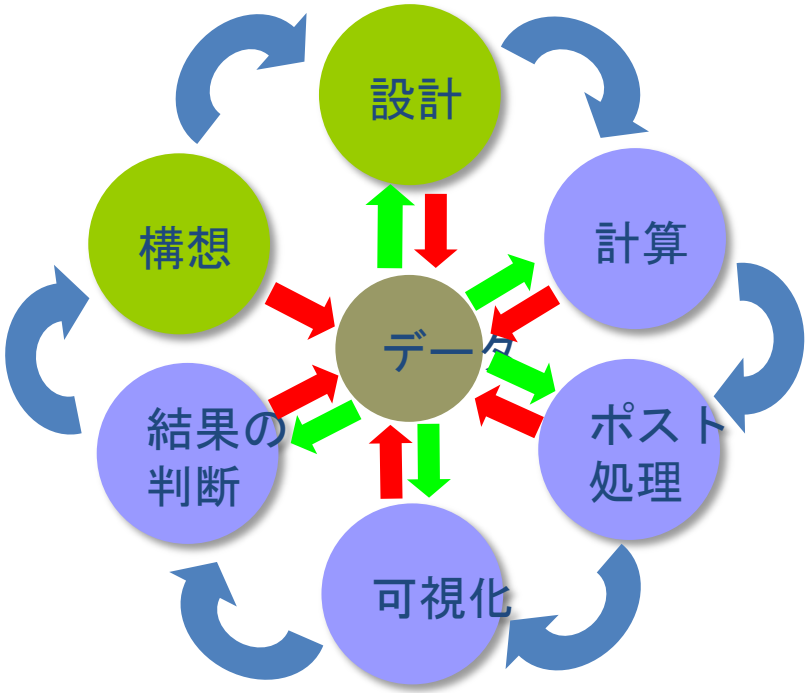
ユーザの要求
大容量のドライブ
大規模なファイル
順次アクセス
高いバンド幅
一貫した可用性
シンプルなSW構成

対話処理	バッチ処理
地下資源探査結果の評価	地質探査解析（大規模データ処理）
EDA デザイン	チップシュミレーション&Tapeout
モデル化、解析結果評価	空力解析、衝突解析
アニメーション処理	レンダリング
トレーディング/ポートフォリオ	リスクマネージメント

理想的なデータの流れの提案

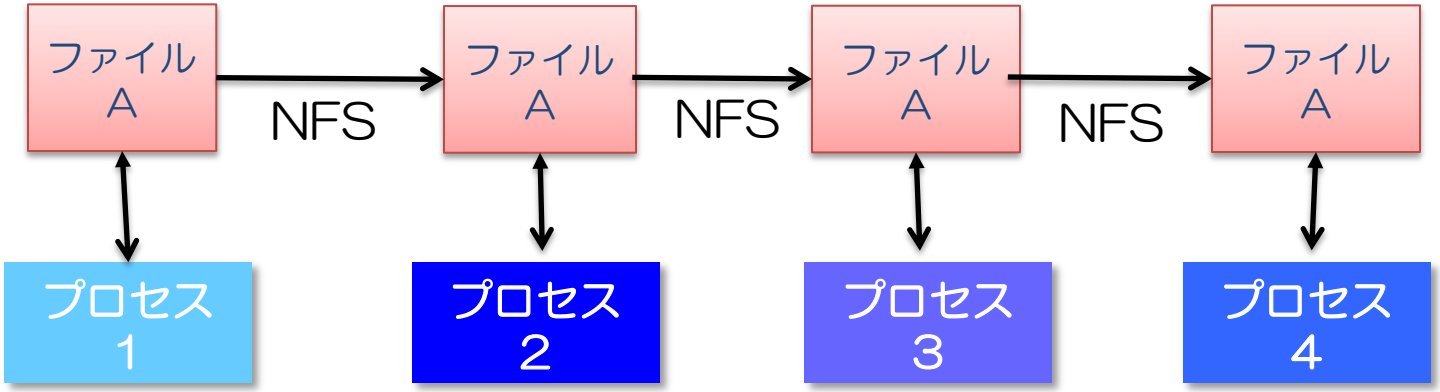
- 一般的な業務の流れを考えてみると..
  - コンセプトの段階から、最終的な製品化の段階まで、データは、業務の流れの中心に位置する
  - 情報は、多くのグループで共有され、データは、ホストコンピュータ間で移動する
  - データセットのサイズは、各ステップ毎により大きくなる

もし、各ステップ毎で、データをコピーし、移動する必要がないのであれば、業務の効率を向上させ、本質的な問題の理解に時間を割ける



アプリケーションワークフロー

ファイル共有をNFSで行った場合

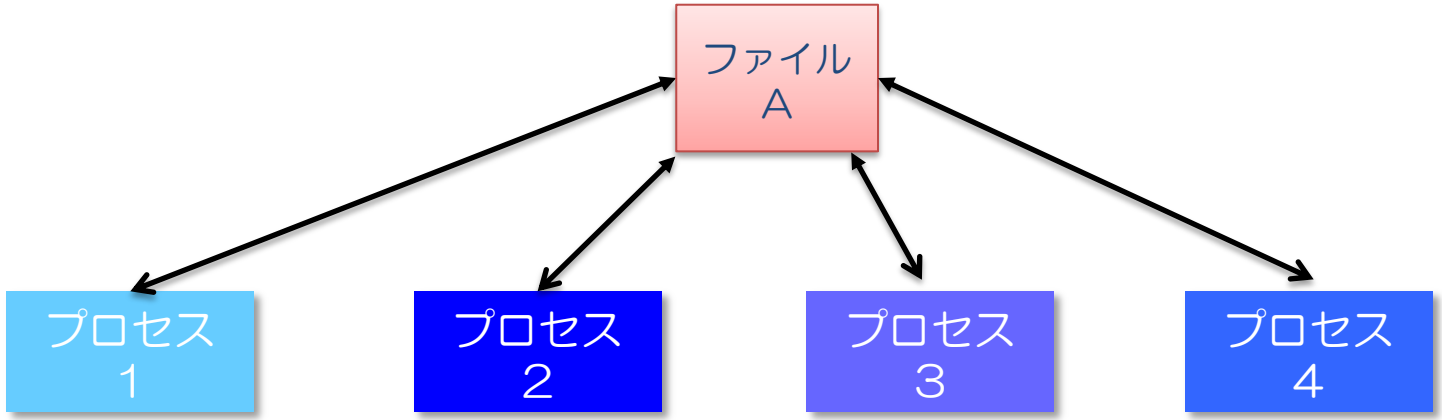


メディア	デジタル化	色補正	効果	合成
製造業.	デザイン	可視化	構造解析	衝突解析

ファイル共有が無い場合には、ネットワーク介してデータの移動が必要
データの保管場所や移動作業などのオーバヘッド
NFSサーバのボトルネックがワークフロー全体に影響する

アプリケーションワークフロー

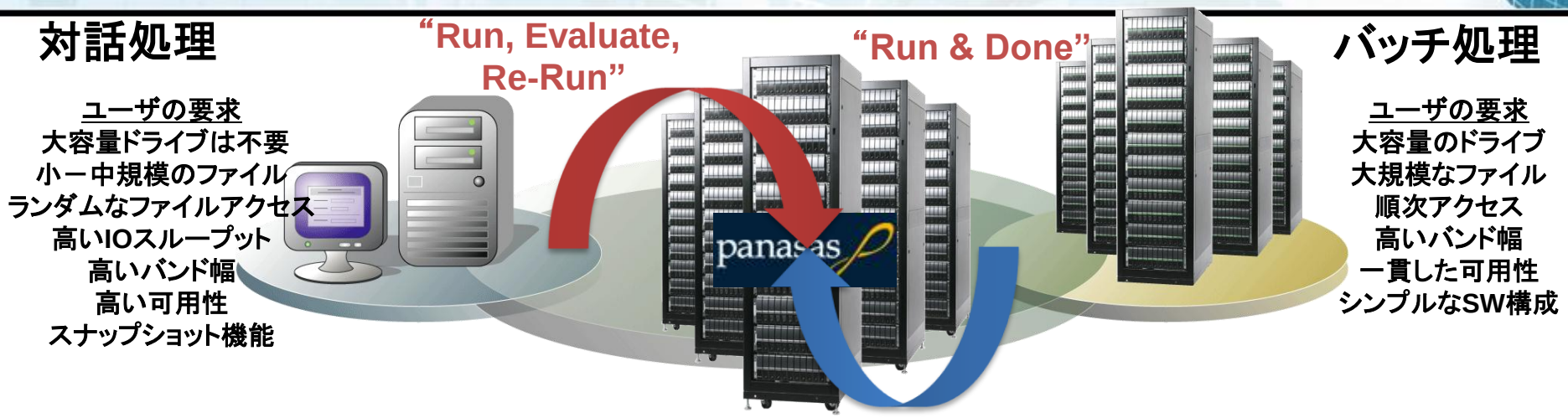
ファイル共有を使用した場合のワークフロー
データ集約ワークフローへの即時共有アクセス



メディア	デジタル化	色補正	効果	合成
製造業.	デザイン	可視化	構造解析	衝突解析

ファイル共有によって、ネットワークを介しての、大規模なファイルを移動が不要となる一時間短縮、ワークフローの効率化スピードアップに貢献する

Panasasによるソリューション
ワークフロー統合ストレージ



共有データへの高速で、容易なアクセスが可能
結果が得られるまでの時間を短縮・データの多重保持が不要

対話処理	バッチ処理
地下資源探査結果の評価	地質探査解析（大規模データ処理）
EDA デザイン	チップシュミレーション&Tapeout
モデル化、解析結果評価	空力解析、衝突解析
アニメーション処理	レンダリング
トレーディング/ポートフォリオ	リスクマネージメント

データセンターの課題

- 継続した性能向上の要求
  - クライアント数、クラスタノード数の増加
  - マルチコア化による計算リソースの強化
- 予測困難なストレージ利用量
  - 導入時に予測することは困難
- 24x7の連続稼働
- 障害からの迅速な復旧
- IT投資の最大活用
- アプリケーションの実行がシンプル

データセンターにおけるストレージ

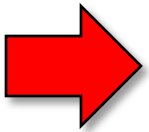
インテル社の推定では、2009年のデータセンターのワークロードの50%はIOに強く依存しています。

ストレージシステムの性能は、データセンターにおいては、現在、非常に重要なITインフラとなっています。これは、図で示すように時間が経過するに従って、そのビジネス上の価値を高めるものになります。

Source: Intel

ストレージシステムの課題

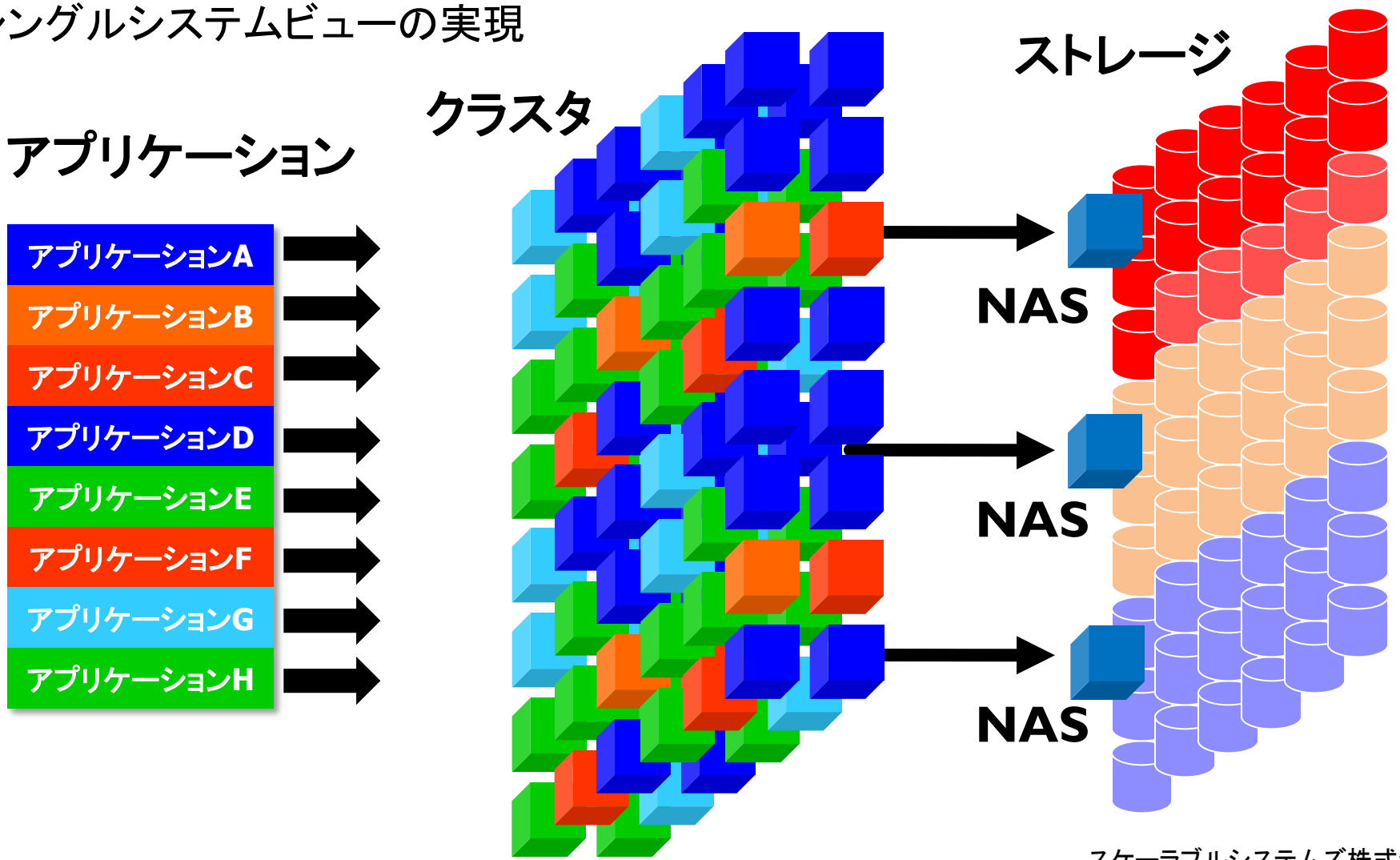
- データアクセスの問題
  - 複数のアーキテクチャ（ハードウエア、ソフトウエア）からのデータへのアクセス
  - 地理的にも分散したストレージの効率的な管理
- 管理運用の問題
  - データ移動を容易に行うことが可能であり、移動したデータに対して、透過的なアクセスが可能
  - ストレージの容量不足の場合などに容易に増設が可能
  - ユーザに負担をかけること無しで、ストレージの運用管理が可能



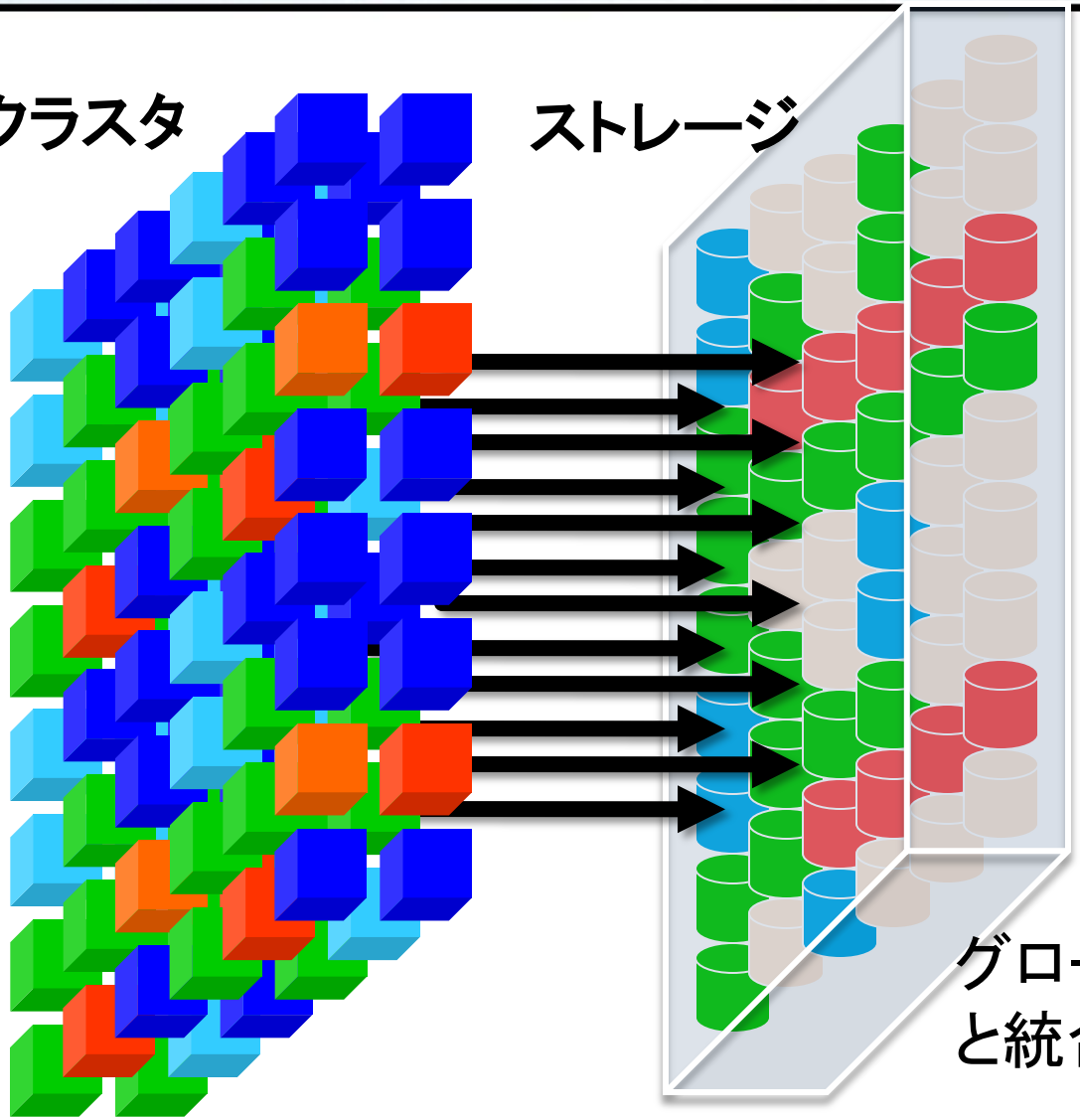
グローバルネームスペースによるソリューション

NASサーバでのシステム構成

すべての点でスケーラブル：パフォーマンス、容量
シングルシステムビューの実現



データセンター
グローバルネームスペースによるソリューション

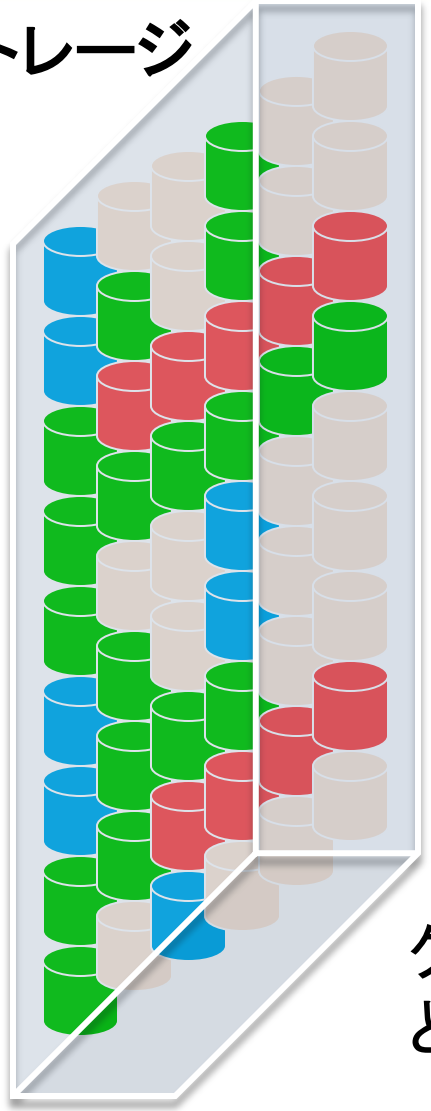


ストレージプール

- 単一の仮想リソース
- 透過的なデータアクセス
- システムの再構築が容易
- 様々なデータの格納が可能
- 可用性

グローバルネームスペースと統合された運用管理環境

Panasasによるソリューション
グローバルネームスペース



- 利用の簡便さ
  - 全クライアントが全データを見ることが可能
  - マウント・ポイント管理が不要
  - クライアント側の変更が不要
- 透過性
  - 容易な拡張
  - Failover
- スケーラビリティ
  - ネームスペースをペタバイトにまで拡張可能
  - 大規模ボリュームの容易な管理

グローバルネームスペース
と統合された運用管理環境

Panasasストレージクラスタ

導入事例

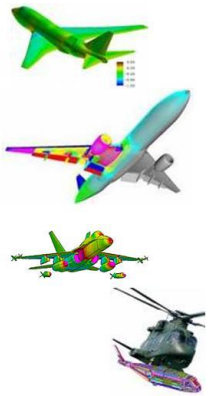
スケーラブルシステムズ株式会社

多くの先進的なユーザが採用

intel

- ✓ **Every processor since 2006** developed, simulated and produced with Panasas storage systems
- ✓ Faster time to market due to **5X** simulation performance improvement with Panasas

BOEING



- ✓ Faster and more accurate design certification on the Boeing 787 – improved air flow simulations at a **cost reduction of over 90%** from previous methods with Panasas

BNP PARIBAS

- ✓ Superior portfolio risk and pricing analysis by running **100X more simulations in the same amount of time** with Panasas

U.S. NCBI/NIH
National Center for Biotechnology Information
NCBI

- ✓ Saw **5X performance improvement** in genetic modeling simulations that store, process, calculate, index and search genetic research for global access with Panasas

LSU TIGERS

- ✓ Panasas provided a **10X performance improvement** for weather modeling and hurricane prediction.

CITADEL

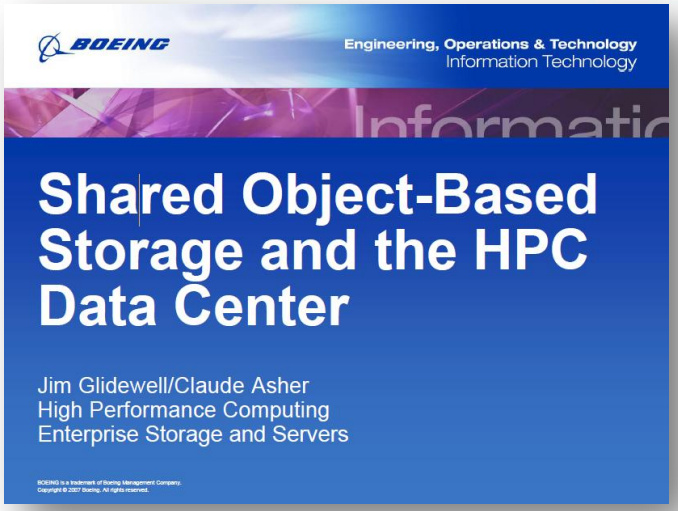
- ✓ "Our ROI with Panasas was about **8 hours**."

Los Alamos
NATIONAL LABORATORY
EST.1943

- ✓ Implemented on RoadRunner: 7000 nodes - 28K core cluster + Infiniband 4X + 3PBs + 216GB/s of IO and achieved 1.3 PF/s.

Pansas採用事例
米ボーイング社

- 利用用途
  - HPCクラスタシステムでの利用
  - 非常に多くのユーザの様々なCAEシュミレーション
    - CFD (計算流体力学)
    - CSM (構造解析)
    - CEM (電磁波解析)
- HPCシステム
  - Linuxクラスタ＋Cray社製ベクトル計算機
  - Panasasストレージクラスタ
- 利用効果
  - 高いスケーラビリティと複数ジョブ、複数ユーザの様々なワークロードに対する効率的な処理



資料の内容については弊社にお尋ねください

Pansas採用事例
米ボーイング社

- Panasasの採用理由
- パラレルファイルシステムが要求要件
  - I/O負荷の大きなジョブと複数のジョブのI/O処理を同時に効率良く処理可能
  - システム全体で高いI/Oバンド幅の要求
- “Production-Ready” ソリューション
  - 導入が容易で直ぐに既存のコンピュータ環境に組み込み利用可能
  - 増設が容易でシステムがスケーラブル
  - システムの負荷分散を動的に実行可能
  - 高い可用性
- TCO削減
  - 導入コスト（コモディティコンポーネント）
  - GbE,10GbE,InfiniBandなどの選択肢
  - 管理運用が容易

Pansas採用事例
インテル

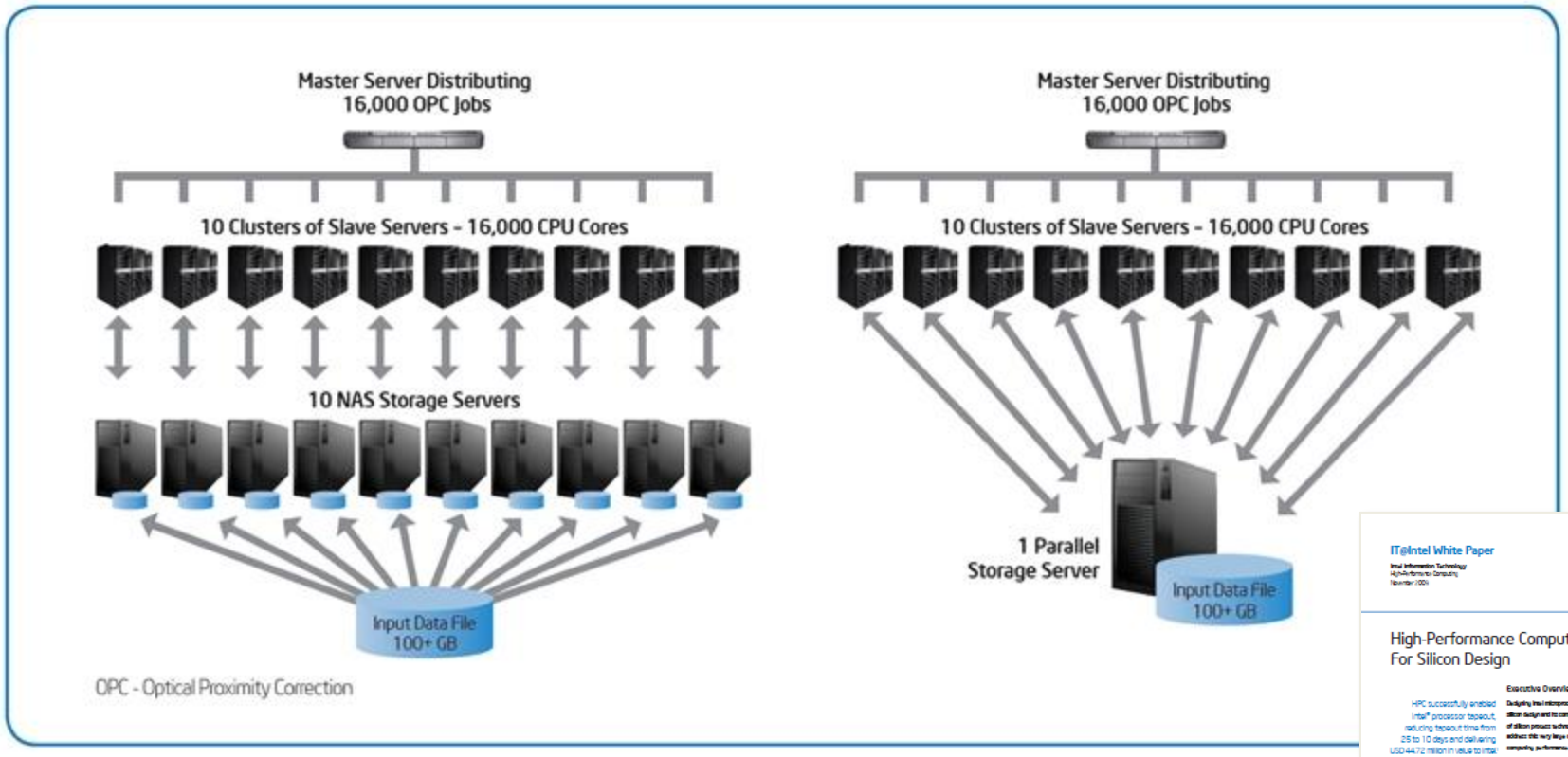


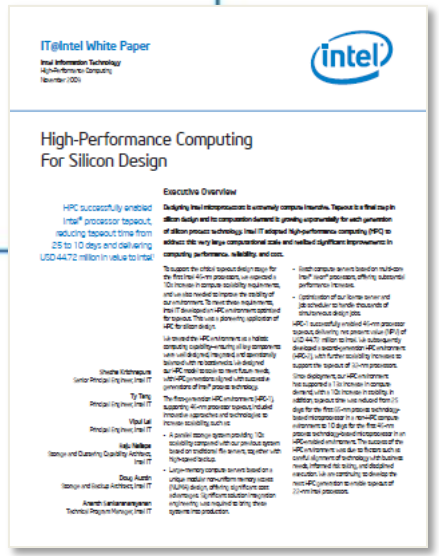
Figure 4. Consolidation with the high-performance computing (HPC) parallel storage environment.

パラレルストレージの利点

スケーラビリティ:コストとスペースの節約にも貢献

性能:従来のシステムに対して300%の性能向上

ボリュームサイズ:利用出来る最大のボリュームサイズが16倍に拡張


IT@Intel White Paper

High-Performance Computing
For Silicon Design

IT@Intel White Paper Intel Information Technology High-Performance Computing November 2009
High-Performance Computing For Silicon Design

スケーラブルシステムズ株式会社

クラスタシステムの構築

Print Story

Close Window

Print Story

Linux.SYS-CON.com Cover Story: Rapid Cluster Deployment

After building a number of clusters from the ground up –including one that made it to the Top500 Supercomputer list – I decided to try a service that many vendors now offer – having a system racked and stacked at the factory then shipped to us. Such a service saves a huge amount of time, not to mention my back, not having to build the cluster and cable all the equipment together. I've been a fan of well-cabled systems and have found the quality control to be acceptable. The key component is the pre-build requirements and verification before the system is built. This will ensure the system shipped is what is expected when it arrives at your front door. There can still be a fair amount of cabling that has to be done once it arrives, if you have a multi-rack configuration, but it's usually limited to plugging in the system's power and public network.

Once this is done, the fun begins…
I've tried a few cluster distribution toolkits, and the one that works for me is the Rocks Cluster Distribution from the San Diego Supercomputing Center. I came across the package in a simple Google search in 2002 and was immediately sold on it. I use the term "sold" loosely since it's under an Open Source BSD-style license available for download and supported by a broad range of technical people who answer most questions on the Rocks user list. I've found support on the list to be better than most commercial distributions, but this may be because there are over 500 registered systems on the Rocks Register.

Here's how simple it is – insert the boot CD, complete a few screens worth of configuration data, and grab a coffee because it's a fairly simple base installation. The Rocks solution is extensible, with a mechanism for users and software vendors to ensure customizations are correctly installed on the system at setup. The mechanism is called a Roll.

The Roll typically consists of packages (RPMS/SRPMS/source) that have to be installed and scripts that are needed to ensure the packages are properly installed and distributed on the cluster. The Rocks team has extensive documentation for the Roll developer in the user manual.

Rocks 4.0.0 is a "cluster on a CD" set. That is it contains all the bits and configuration to build a cluster from "naked" hardware. The core OS bundled with Rocks is CentOS 4, which is a freely downloadable rebuild of Red Hat Enterprise Linux 4. As a side note, in Rocks CentOS 4 is encapsulated as the "OS Roll" and this OS Roll can be substituted with *any* Red Hat Enterprise Linux 4 rebuild (e.g., Scientific Linux) including the official bits from Red Hat. Rolls are used in Rocks to customize your cluster. For example, the HPC Roll contains cluster-specific packages, such as an MPI environment for developing and running parallel programs. Two other examples are the Ganglia Roll, which provides cluster-monitoring tools, and the Area51 Roll, which provides security tools such as Tripwire and chkrootkit.

The Software
The core OS we used for the cluster in this article is CentOS 4.0 and the rolls we used to customize the cluster to our needs were the Compute Roll and the PBS Roll from University of Tromso in Norway.

The Hardware

- **1** – Front-end node – a Dell PowerEdge 2850 with dual 3.6GHz Intel Xeon EM64T processors and 4GB RAM
- **48** – Compute nodes – Dell PowerEdge SC 1425s with dual 3.4GHz Intel Xeon EM64T processors, 2GB RAM and a Topspin PCI-X Infiniband HCA card
- **1** – Topspin 270 Infiniband chassis with modules
- **4** – Dell PowerConnect 5324 Gigabit Ethernet switches
- **1** – Panasas Storage Cluster with one DirectorBlade and 10 StorageBlades
- **2** – Dell 19-inch racks

Start the build process ***time 0:00:00***
Setting up the front-end:
– Insert Compute Roll and boot the system
– Select hpc, kernel, ganglia, base, java, and area51 as the rolls to install
– Select "Yes" for additional roll
– Insert CentOS disk 1
– Select "Yes" for additional roll
– Insert CentOS disk 2
– Select "Yes" for additional roll
– Insert PBS roll
– Select "No" for additional rolls
– Input data on the configuration screen (e.g., fully qualified domain name, root password, IP addresses)
– Select "Disk Druid" to create partitions
– Create/partition ext3 64GB
– Create swap partition 4GB
– Create/export partition 64GB
– Insert CDs as requested to merge them into the distribution

The most important step…grab a mocha and enjoy it while the install runs.

After the-front end installation completes, the site-specific customization of the front-end starts. The base installation of CentOS 4.0 x86_64 has the 2.6.9-5.0.5.ELsmp kernel and we need the 2.6.9-11.ELsmp for many of the packages that will be included with our cluster. Below we'll describe how we do this key upgrade then continue with many package and mount point customizations.

http://linux.sys-con.com/node/193368?page=0,1

Panasas実装時間

“Panasasストレージクラスタが3つのボックスとして到着し、それらのボックスの開封を行ってから、システムが完全に利用できるまでに必要としたのは、僅か 1時間 55 分でした....”

Time for Panasas Integration
The Panasas Storage Cluster arrived in three boxes on one pallet. From the time I clipped the first band on the pallet to having the system fully operational was only 1 hour 55 minutes. Here's how the process went.

ペタスケールスケーラビリティ

CASE STUDY:
Panasas Parallel Storage Powers the World's First Petaflop Supercomputer

HIGHLIGHTS

First Petaflop Supercomputer

- #1 on the Top-500 list in 2009
- Over 3,250 Compute Nodes
- Over 156 I/O Nodes
- Over 12,000 Core Processors
- Hundreds of Thousands of Cell Processors

Panasas Parallel Storage Solutions

- 100 Panasas Storage Shelves
- 2 Petabytes Capacity
- 55 GB/s Throughput
- Throughput Scales Linearly with Capacity
- Non-Stop Availability & Simple to Deploy

ABSTRACT

Scientists want faster, more powerful high-performance supercomputers to simulate complex physical, biological, and socioeconomic systems with greater realism and predictive power. In May 2009, Los Alamos scientists doubled the processing speed of the previously fastest computer. Roadrunner, a new hybrid supercomputer, uses specialized Cell coprocessors to propel performance to petaflop speeds capable of more than a thousand trillion calculations per second.

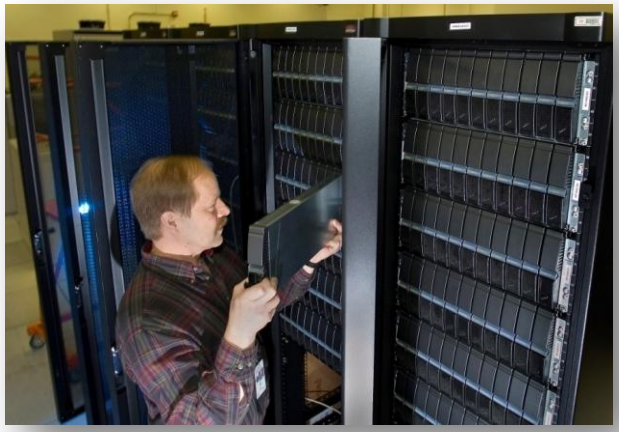
One of the keys to the project's success was the need for a highly reliable storage subsystem that could provide massively parallel I/O throughput with linear scalability that was simple to deploy and maintain. Los Alamos National Laboratory deployed the Panasas ActiveStor Parallel storage to meet the stringent needs of the Roadrunner project. Panasas provides scalable performance with commodity parts providing excellent price/performance, scalable capacity and performance that scale symmetrically with processor, caching, and network bandwidth.

ACTIVESTOR™ PARALLEL STORAGE POWERS THE FIRST PETAFLOP SUPERCOMPUTER AT LOS ALAMOS NATIONAL LABORATORY CASE STUDY | FEBRUARY 2010

アーキテクチャ概要

Panasas ActiveScale ストレージクラスタとは?

- ハードウエアとソフトウエアによるアプライアンスストレージシステム
  - 非同期、パラレル、オブジェクトベース、POSIXコンプライアントファイルシステム
  - グローバルネームスペース
  - クライアントキャッシュコヒレンシ



ストレージに関する課題の解決

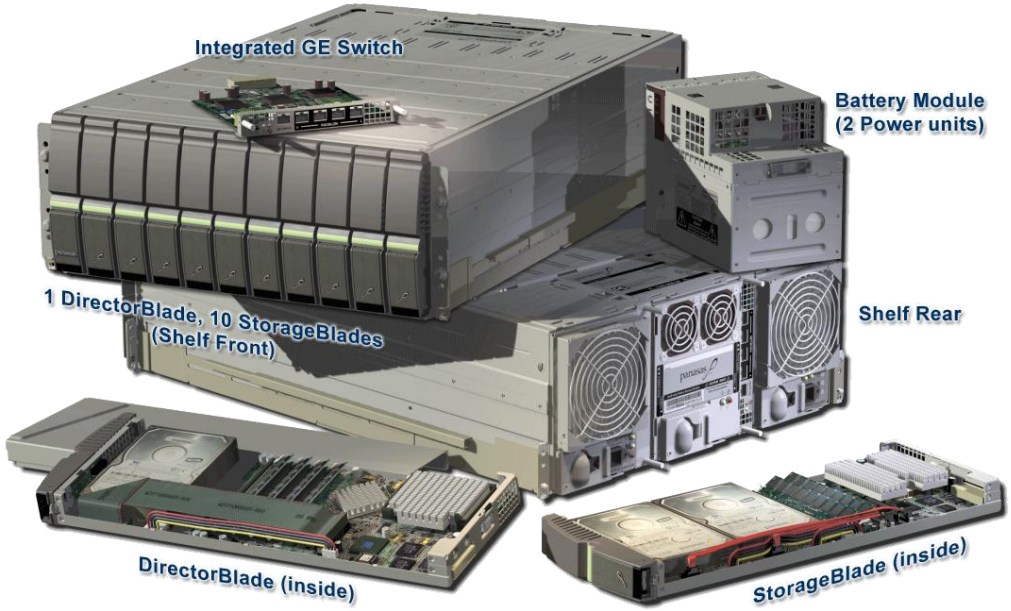
- 従来のストレージシステムのボトルネックの解消
- RAIDに関するボトルネックの解消
  - クライアント数が増えた場合のRAIDのスケーラビリティ
  - 複数ボリューム構成とスケーラブルなRAID再構成（複数のDirectorBladeによるパラレル再構成）
- ネットワーク接続のボトルネック解消
  - 4ポート GbE　リンク（シェルフあたり）
  - 10GbE リンク（シェルフあたり）
- 柔軟性（ファイル毎の設定が可能）
  - RAID1/5 （大きなファイルのストリーミング）
  - RAID10 （N-to-1 書き出し、もしくは、ランダム I/O）
- グローバルネームスペース
- WEBからの一括管理

利用環境への柔軟な対応

- 利用プロトコルの選択
  - DirectFLOW, NFS, CIFS（これらの任意の組み合わせも可能）
    - NFS / CIFS 性能（DiectorBladeを増設）
    - DirectFLOW 性能（StorageBladeを増設）
- インタラクティブとバッチ処理
  - インタラクティブ処理には、キャッシュサイズの大きなStorageBladeの選択が有効
- フォールトトレランス （Fault tolerance）
  - StorageBlade障害に対応するためのスペアブレードの設定
  - ブレードセット(bladeset) 設定による障害対策
  - 冗長ネットワーク構成（オプション）
- ストレージ容量オプション
  - 小容量ブレード
    - より多くのブレードによる性能向上、再構成時に少ないデータ処理、より多くのシェルフ
  - 大容量ブレード
    - より少ないシェルフ構成が可能、ディスク障害の確率低下

データの保護と可用性
ハードウエア設計

- 電源とファンの冗長化
- 各ブレードに対するネットワーク接続の冗長化
- ECCメモリ
- Shelf内にバックアップのネットワークを内蔵



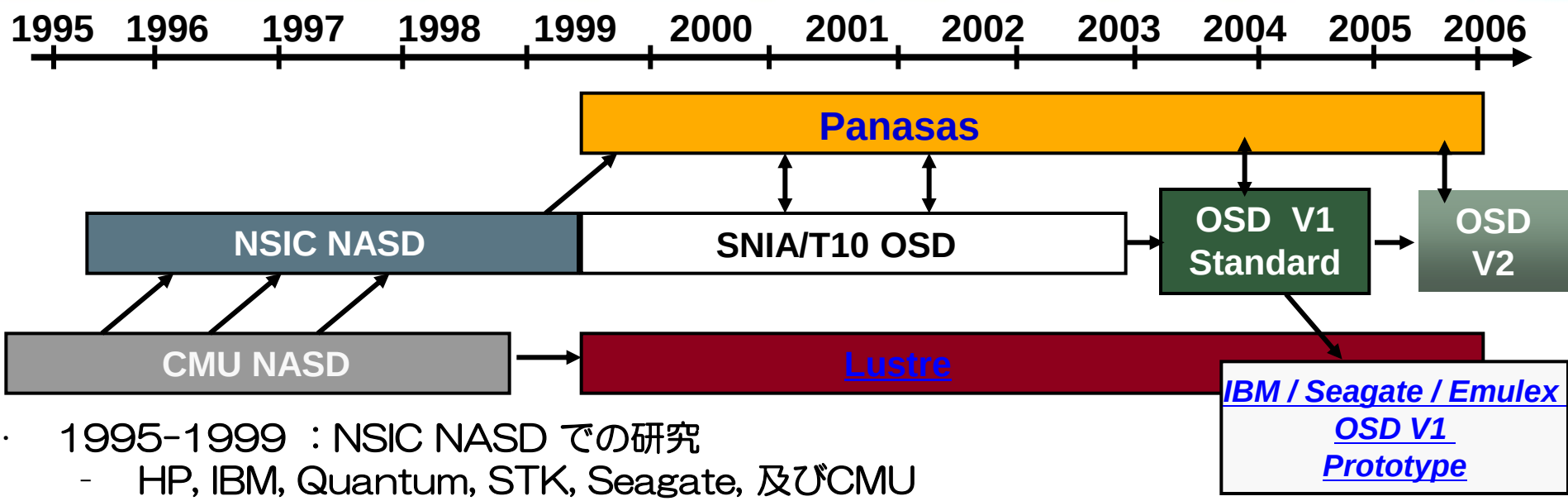
データの保護と可用性
ソフトウエアでの信頼性向上

- 高速な再構成が可能なRAID 10 及び 5 を提供
  - ファイル毎にRAID構成とデータの分散が可能であり、障害時の影響を最小化
  - 障害時の再構成はオンライン中にバックグランドで実行されるため、サービスの停止は不要
- Panasas Tiered Parity
- 信頼性の高いオペレーティングシステム
  - システムのサービスのフェイルオーバ、ファイルサービスのフェイルオーバ、メタデータ処理のフェイルオーバなど
  - 各ブレードで動作しているOSについては、ミラーリングを行う
  - ディスクの状況や熱、ファンの動作などを細かくモニターして、予防診断を行い、障害に対応する
  - スケーラブルな高速バックアップのサポート

Panasas社が提供する構築ブロック

1. オブジェクトストレージデバイス（OSD)
   - データと属性のコンテナ
   - iSCSI/OSDインターフェイスとしてのSNIA T10 を標準インターフェイル
   - Panasas社のStorageBlade は、商用OSDとして初めての製品
2. 分散&パラレルファイルシステム
   - ブロックマネージメントは、オブジェクトストレージインターフェイスのバックで動作
   - クライアントからのIOは直接、パラレルにオブジェクトストレージデバイスに送られる
   - ファイルマネージメントは、メタデータマネージャ全体で処理される
   - 障害発生時の対応
3. スケーラブルなPanasas社のRAIDシステム
   - ファイルを複数のコンテナオブジェクトに分割
   - パラレルRAIDの再構築

オブジェクトストレージの発展（歴史）



- 1995-1999 ：NSIC NASD での研究
  - HP, IBM, Quantum, STK, Seagate, 及びCMU
  - 1999年にSNIAの技術ワーキンググループとなる
    - 45 の会社が参加
- 1999 に SNIA/T10 ワーキンググループに発展
- 1/2005: ANSI が V1 T10 OSD 規格を批准 (ANSI/INCITS 400-2004)
  - SNIA TWG は OSD V2 の内容について活動
  - スナップショット、import/exportやマルチ-オブジェクトの機能や属性の拡張などが議論される
- 重要な点: これらの規格によって、顧客がオブジェクトストレージの採用を検討するためのオプションを築くことになる

オブジェクトストレージ アーキテクチャ

- 標準のSCSIストレージインターフェイスに関する革新的な改善
- データの抽象化のレベル：オブジェクトには、‘関係する’データの格納単位（オブジェクトは、データベースの一つのレコード又はテーブルでも、また、データベース全体とすることも出来る）
  - ストレージをブロックやファイルでなくオブジェクトとして扱う
  - OSD (Object-Based Storage Device)は、オブジェクトの属性、ブロックポインタ、データブロックの割り当てを管理
  - OSDは、各オブジェクト毎にアクセスコントロールを実施
- プラットフォーム固有のデバイス管理をデバイスにオフロード

オペレーション:

Read block
Write block

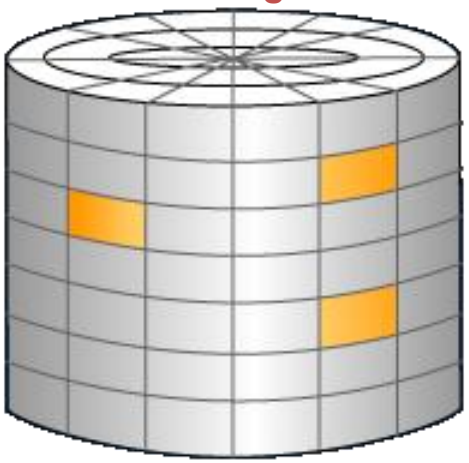
アドレッシング:

Block range

割り当て:

External

Block Storage Device



オペレーション:

Create object
Delete object
Read object
Write object
Get Attributes
Set Attributes

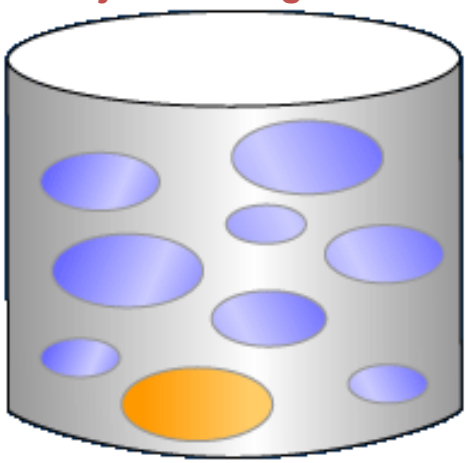
アドレッシング:

[object, byte range]

割り当て:

Internal

Object Storage Device

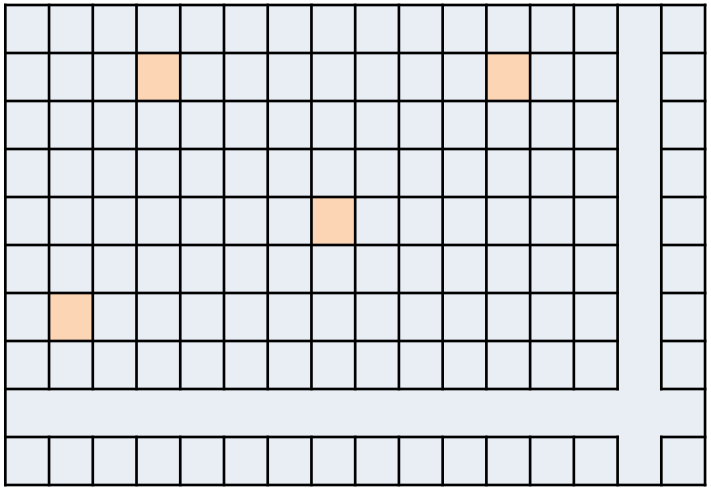


ブロック・ベースと
オブジェクトベースの違い

個々のブロックと通信するプロトコルを利用（SCSI,ATA)

ブロックサイズは固定

データとメタデータの両方が含まれるブロックのコレクション

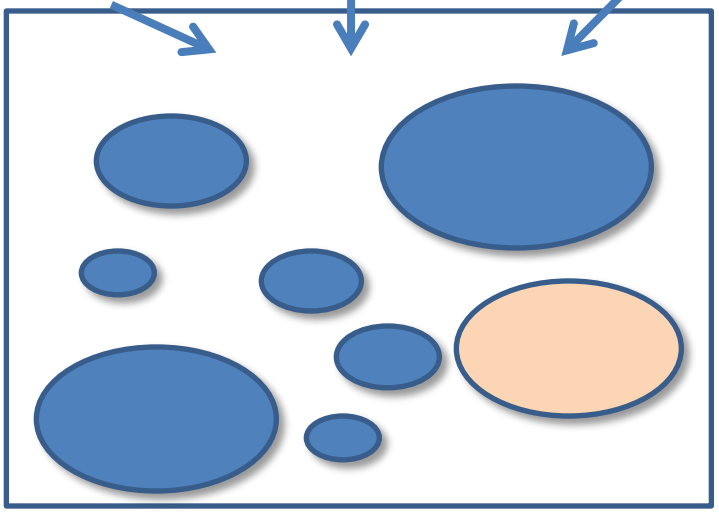


ブロックベースストレージシステム

個々のオブジェクトと通信するプロトコルを利用（OSDなど)

オブジェクトサイズは可変

オブジェクトとそのオブジェクトに関するメタデータ

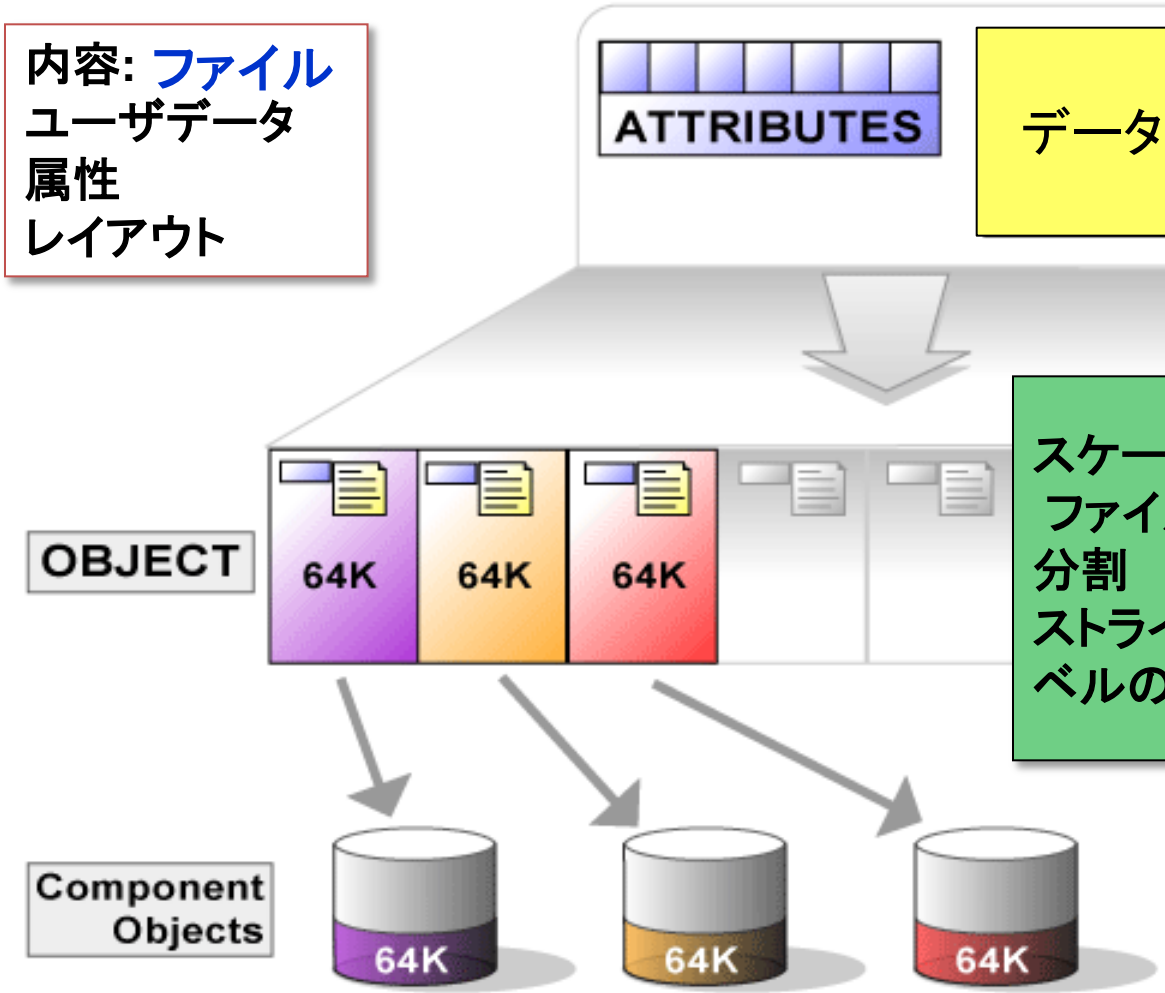


オブジェクトベースストレージシステム

Pansasでのオブジェクトの取り扱い

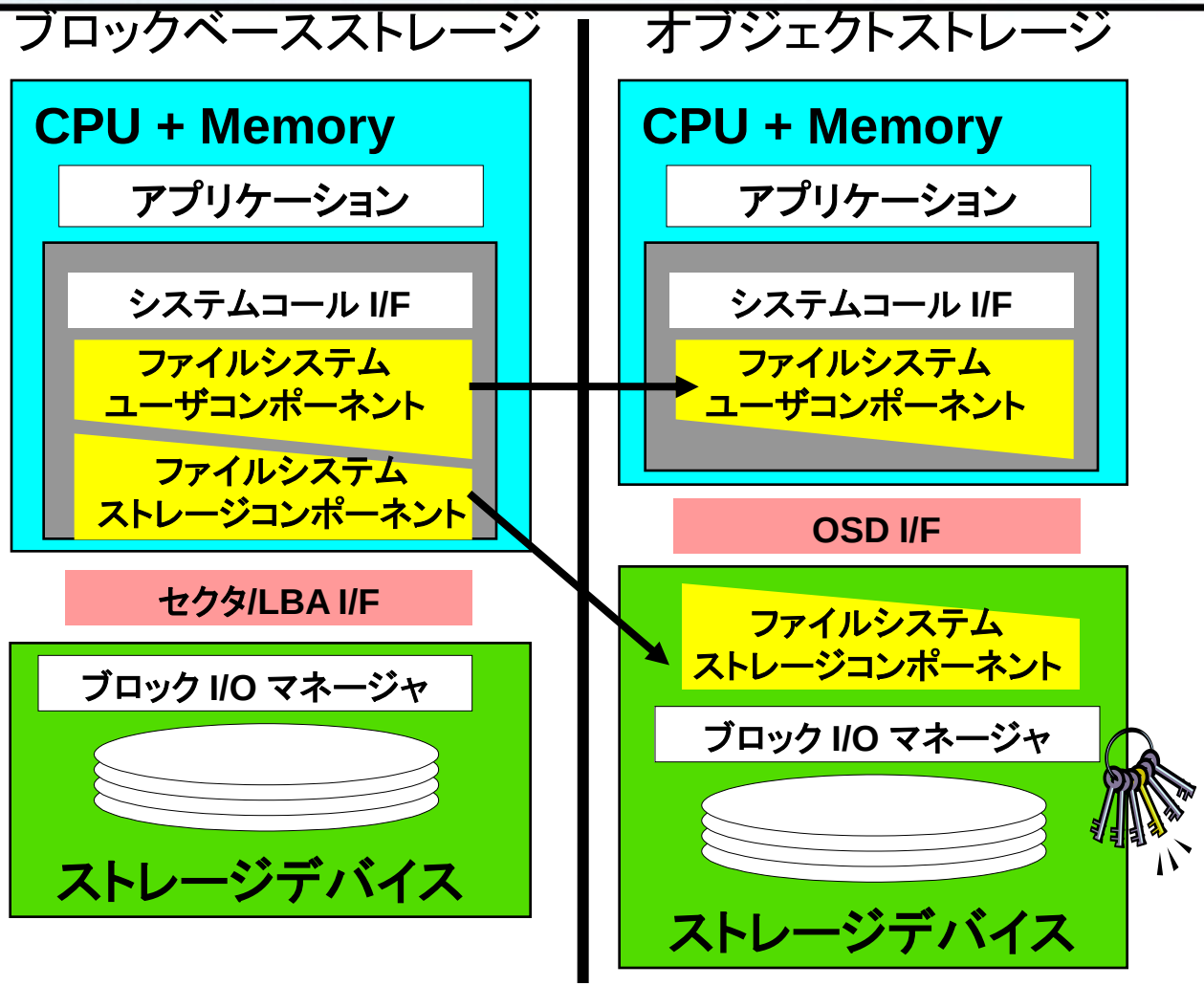
ファイルを小さなマップに分割することで、容量、バンド幅、信頼性の向上を図る

内容: ファイル
ユーザデータ
属性
レイアウト



スケーラブルなオブジェクトマップ
ファイルは、個々のオブジェクトに分割
ストライプ（分割）サイズとRAIDレベルの情報も付加

オブジェクトストレージモデル



- クライアントのファイルシステムの一部の機能をストレージデバイスで実行
- ストレージデバイスは、単にセクターの情報だけでなく、データのレイアウトなどの情報を処理
- ストレージの処理単位をブロックからオブジェクトに変更し、ストレージ自身が、より多くのデータ管理を行う

LBAとは、ハードディスク内のすべてのセクタに対してゼロからの通し番号を振ることで、その通し番号によってセクタを指定する方式のことである。論理ブロックアドレスと呼ばれることもある。

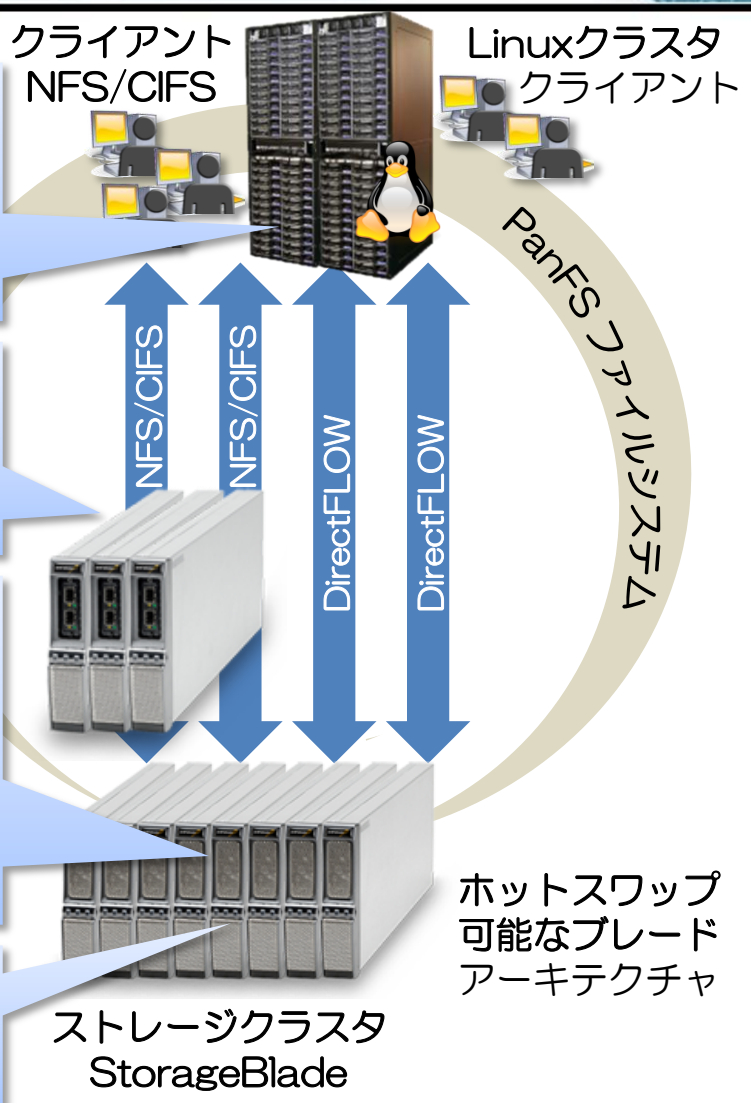
Panasasストレージクラスタ

DirectFLOW クライアントソフトウエア
- クライアントからの同時アクセスを並列に処理可能
- RedHat、SUSEなどの主要なLinuxディストリビューションで利用可能
- pNFSにも対応

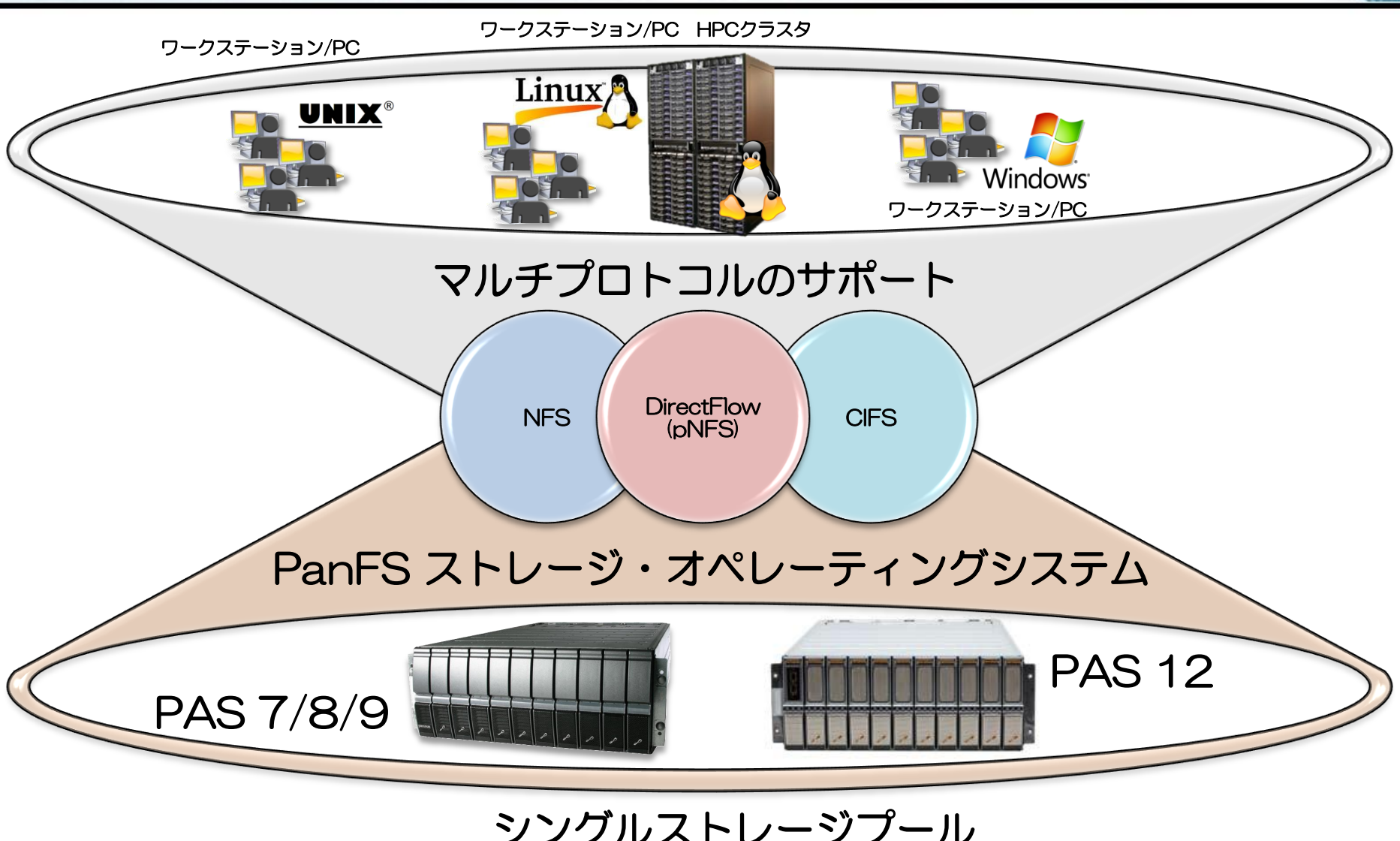
スケーラブルな NFS/CIFS/NDMPサーバ
- 負荷を自動的にストレージクラスタ全体に分散
- クライアント数の増加に合わせてスケーラブルな性能拡張
- 全てのDirectorBladeが全てのファイルにアクセス可能

シングルネームスペース
- 同一データへのいずれのプロトコルでのアクセスも可能
- シングルファイルシステム
- DirectFLOW/NFS/CIFS/NDMP間の完全なコヒレンシの実現
- 非Linuxのデバイスをシステムに統合
- グローバルネームスペースによるシステムの容易な拡張と運用の容易さ

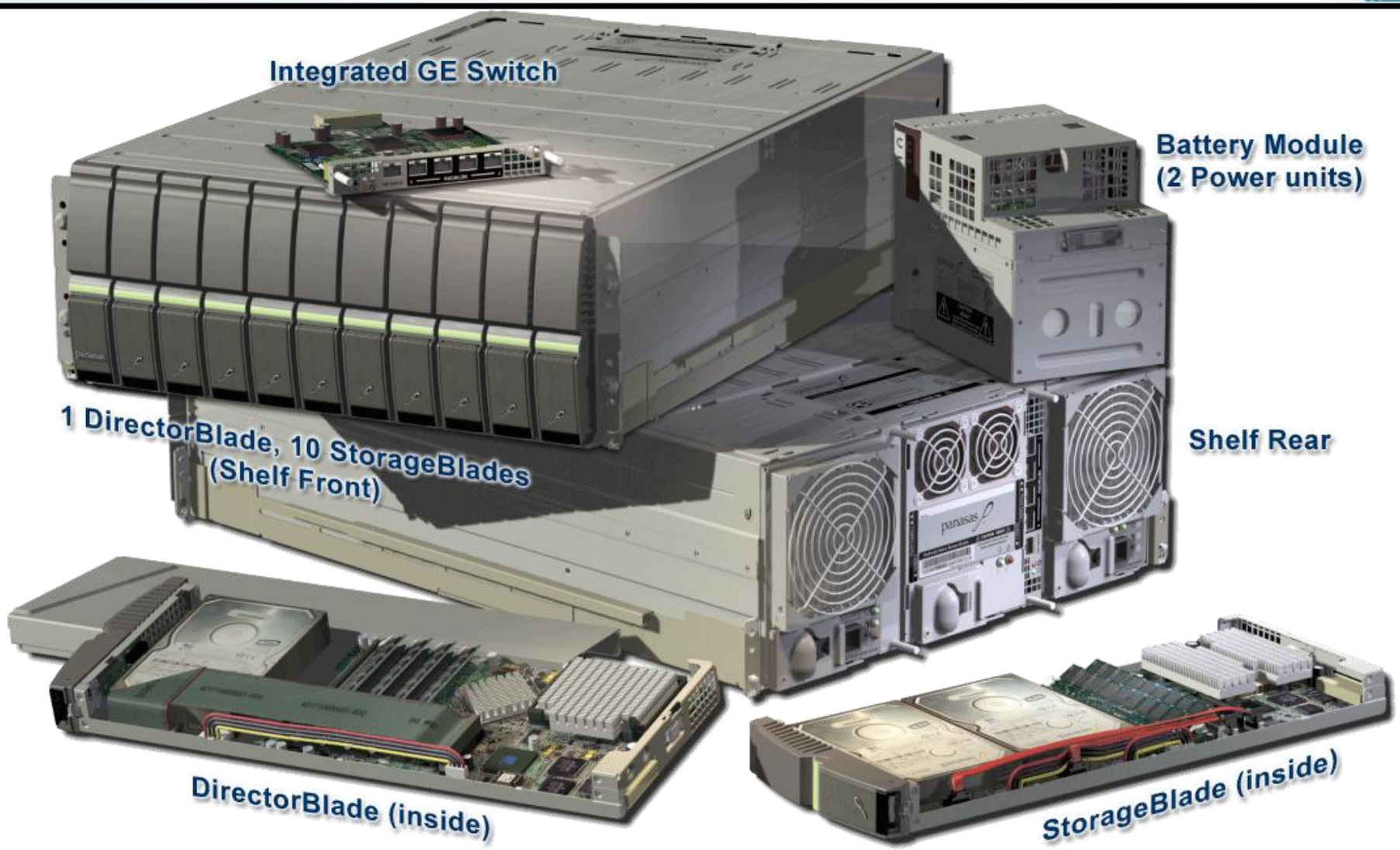
オブジェクトベース
- 優れたスケーラビリティ、信頼性、運用管理
- Panasas Tiered Parityによるデータ保護の強化



Panasasストレージクラスタ



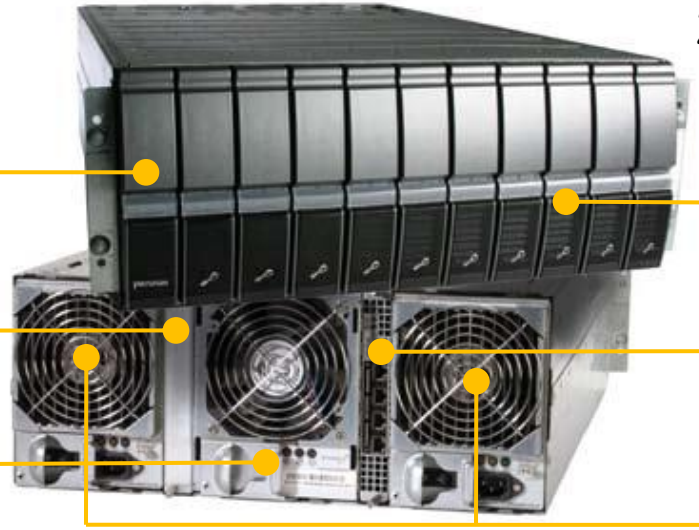
Panasas ストレージクラスタ
業界標準のコンポーネントでのシステム構築



Panasas ActiveStor Performance Module

Panasas ActiveStorストレージクラスタ
Performance Module

ホットスワップ可能
ブレードアーキテクチャ
20TB - 60TB / Module



DirectorBlade
メタデータ処理

StorageBlade
オブジェクトデータ処理

セカンドネットワーク
スイッチ（オプション）

ＧｂＥ、10GbE
ネットワークスイッチ

バッテリモジュール
電源バックアップ

冗長化電源
ホットスワップ可能

ストレージクラスタ構成要素

StorageBlade
- プロセッサ、メモリ、2つのNIC、2つのHDD
- オブジェクトストレージシステム
- ブロックマネージメント

DirectorBlade
- プロセッサ、メモリ、2つのNIC、１つのHDD
- 分散ファイルシステム
- ファイルとオブジェクトマネージメント
- クラスタマネージメント
- NFS/CIFS 再エクスポート

統合されたハードウエアとソフトウエアによるソリューション
- 4Uのシェルフに11のブレード（20-60 TB/シェルフ）
- 現在:1 から 30台のシェルフでシステムを構築
- 将来:1 から 300台のシェルフでシステムを構築

オブジェクトベース　スマートに商用製品を活用
クラスタファイルシステム　したハードウエア構成

Panasas ActiveScale
ストレージクラスタ

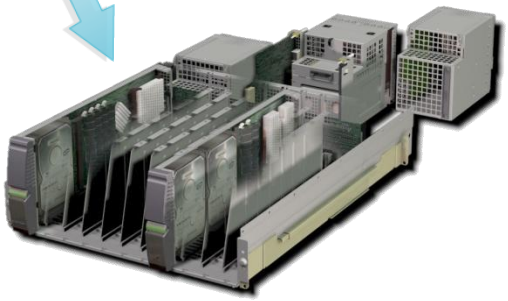
アプライアンデザイン

DirectorBlade

- ストレージクラスタの管理
- ブレード間でのオブジェクトデータの最適な利用

StorageBlade

- SATAドライブ
- 1TB、1.5TB 、2TB



- Shelf あたり10から40TB搭載可能
- ラックあたり100 から400 TB搭載可能

- 16-Port GbE/18-Port 10GbEスイッチ
- 冗長電源 + バッテリ

- ホットスワップ可能
- **No single point** of **failure**（単一機器の障害がシステム全体の障害とならない構成）

Panasas社が提供する構築ブロック

1. オブジェクトストレージデバイス（OSD）
   - データと属性のコンテナ
   - iSCSI/OSDインターフェイスとしてのSNIA T10 を標準インターフェイル
   - Panasas社のStorageBlade は、商用OSDとして初めての製品
2. 分散&パラレルファイルシステム
   - ブロックマネージメントは、オブジェクトストレージインターフェイスのバックで動作
   - クライアントからのIOは直接、パラレルにオブジェクトストレージデバイスに送られる
   - ファイルマネージメントは、メタデータマネージャ全体で処理される
   - 障害発生時の対応
3. スケーラブルなPanasas社のRAIDシステム
   - ファイルを複数のコンテナオブジェクトに分割
   - パラレルRAIDの再構築

Panasasファイルシステムモデル

クラスタファイルシステム

- 多くのサーバでのストレージの共有が可能
- クラスタシステムでのファイルシステムの実装が容易

パラレル

- 同時に複数の読み込みと書き出しが可能
- 一つのファイルを分割して高速に処理することが可能

オブジェクトベース

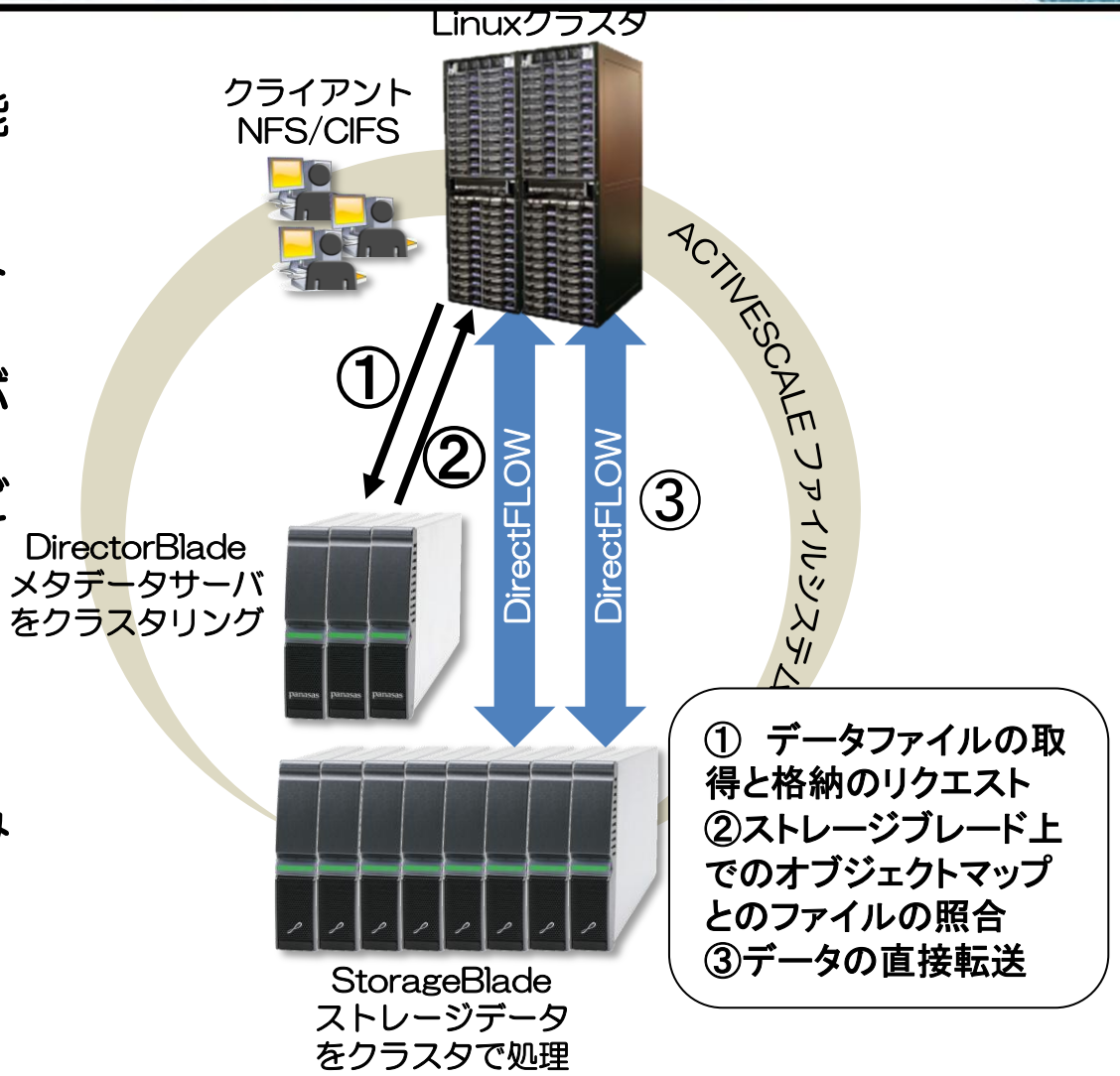
- 非ブロックベース
- IO処理の負荷分散・機能分散が容易

分散ファイルシステム

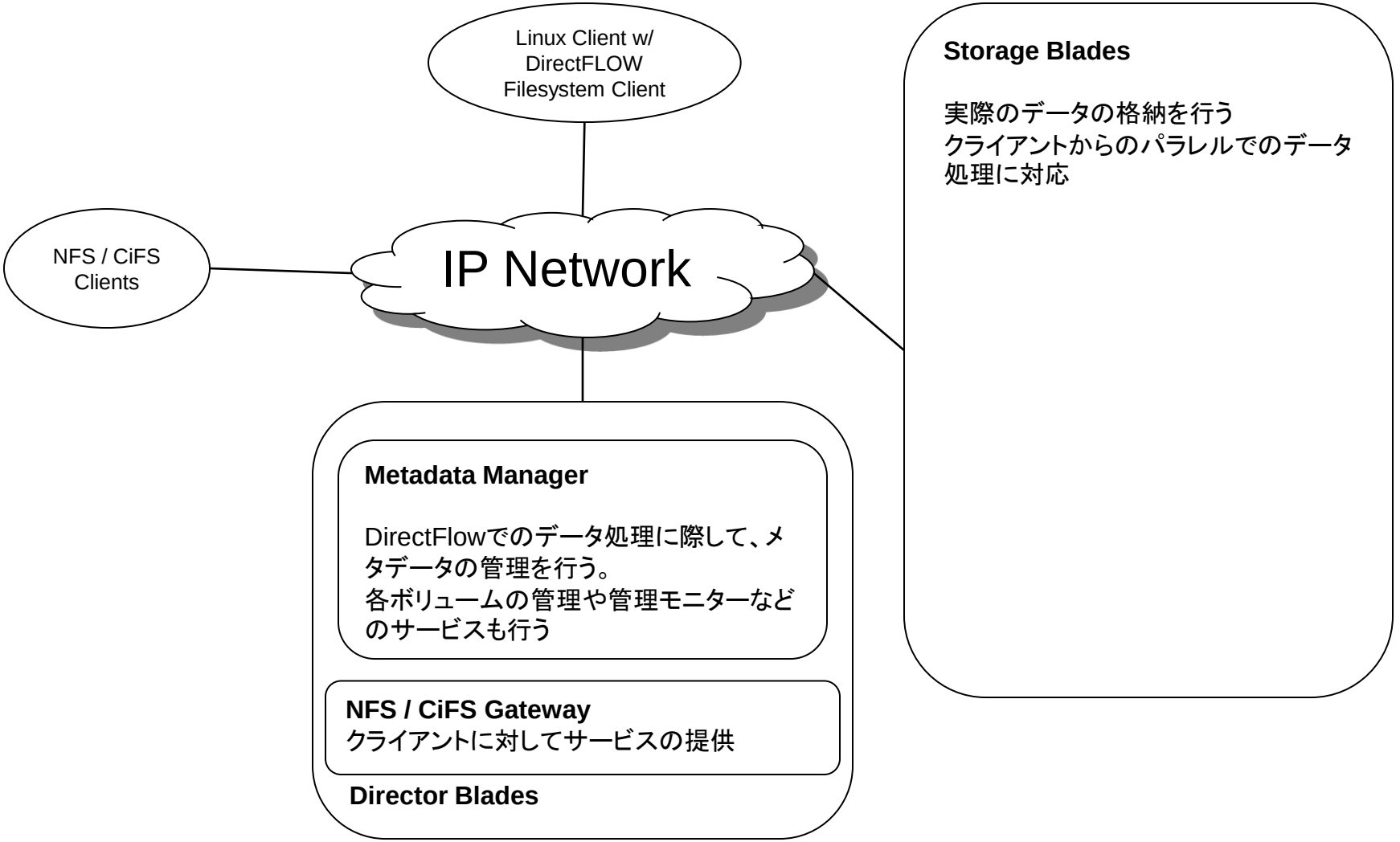
- ストレージはクライアントからアクセス可能なネットワーク上に分散
- 高速ネットワークを利用することで容易に性能向上を図ることが可能

DirectFLOW による最大性能の実現

- DirectFLOW クライアント
  - 標準にインストール可能なファイルシステム
  - 一般のLinuxディストリビューションをサポート
- DirectorBlade クラスタ
  - ネームスペースを仮想ボリュームに分割
  - メタデータのスケーラビリティを実現（ボトルネックの解消）
- StorageBlade クラスタ
  - 大規模ファイルのスプリットが可能
  - 小さなファイルの先読み/後書きが可能

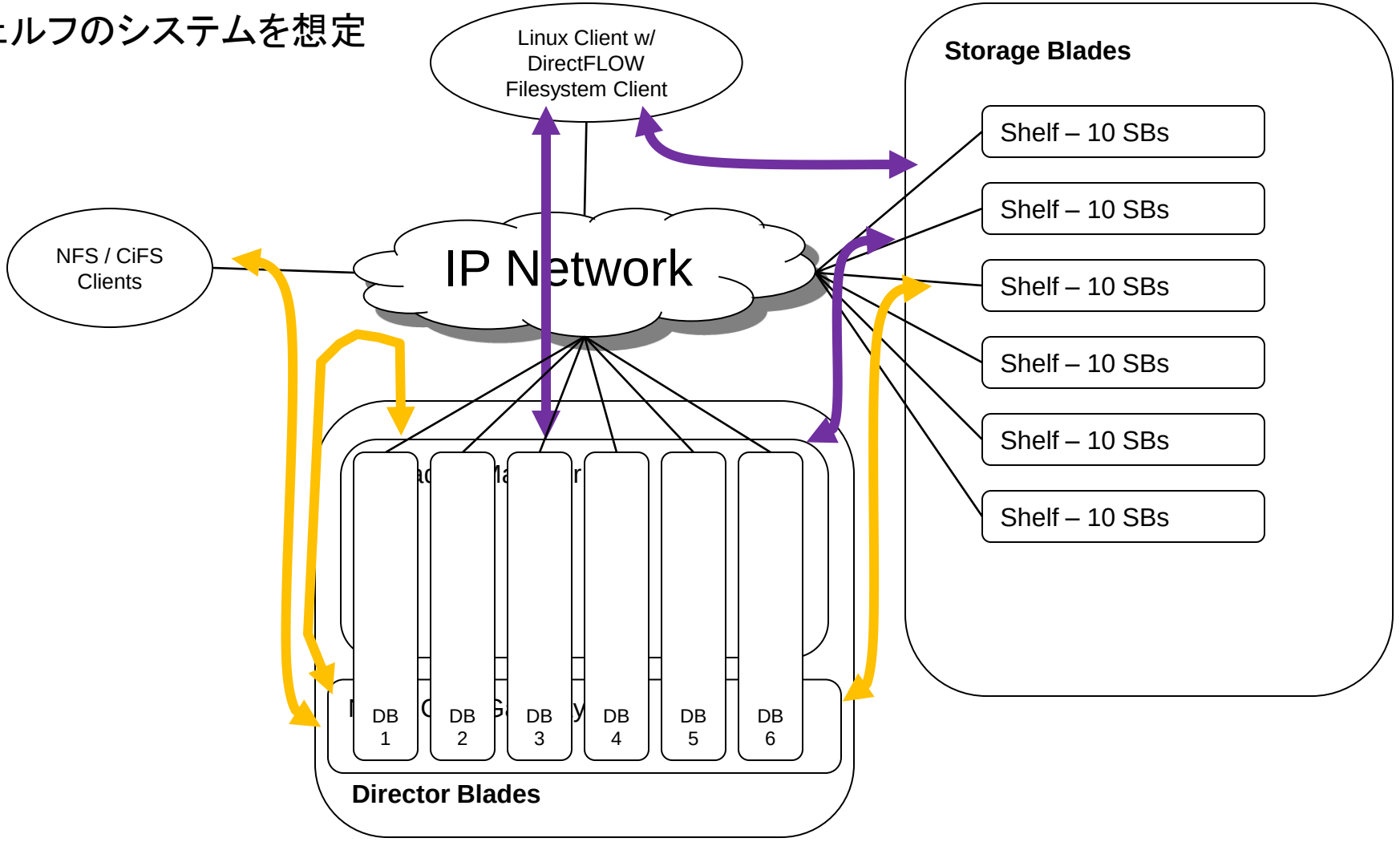


データフローの詳細


Linux Client w/
DirectFLOW
Filesystem Client
NFS / CiFS
Clients
IP Network
Storage Blades
実際のデータの格納を行う
クライアントからのパラレルでのデータ処理に対応
Metadata Manager
DirectFlowでのデータ処理に際して、メタデータの管理を行う。
各ボリュームの管理や管理モニターなどのサービスも行う
NFS / CiFS Gateway
クライアントに対してサービスの提供
Director Blades

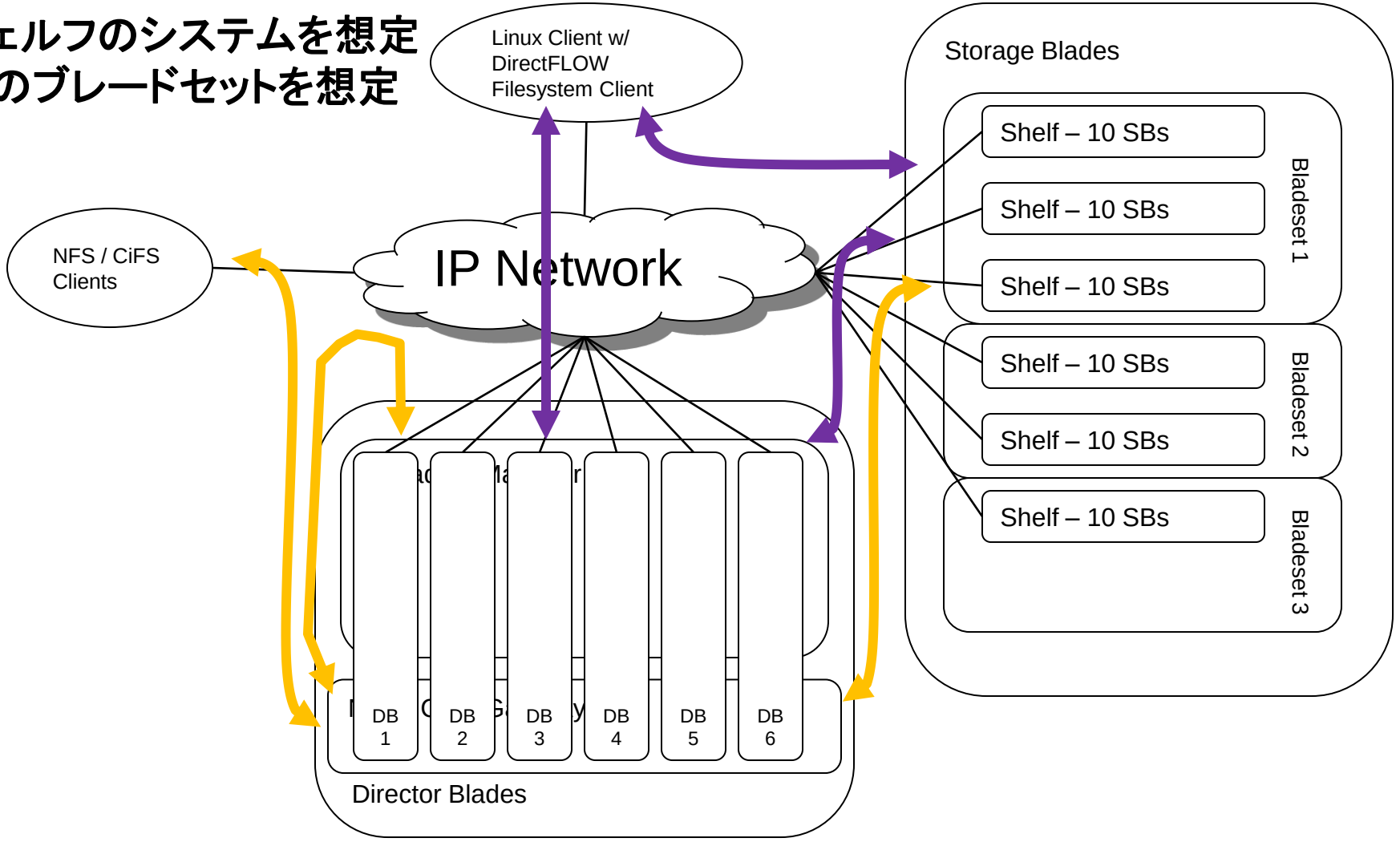
データフロー
DirectFlow & NFS/CIFS

6シェルフのシステムを想定



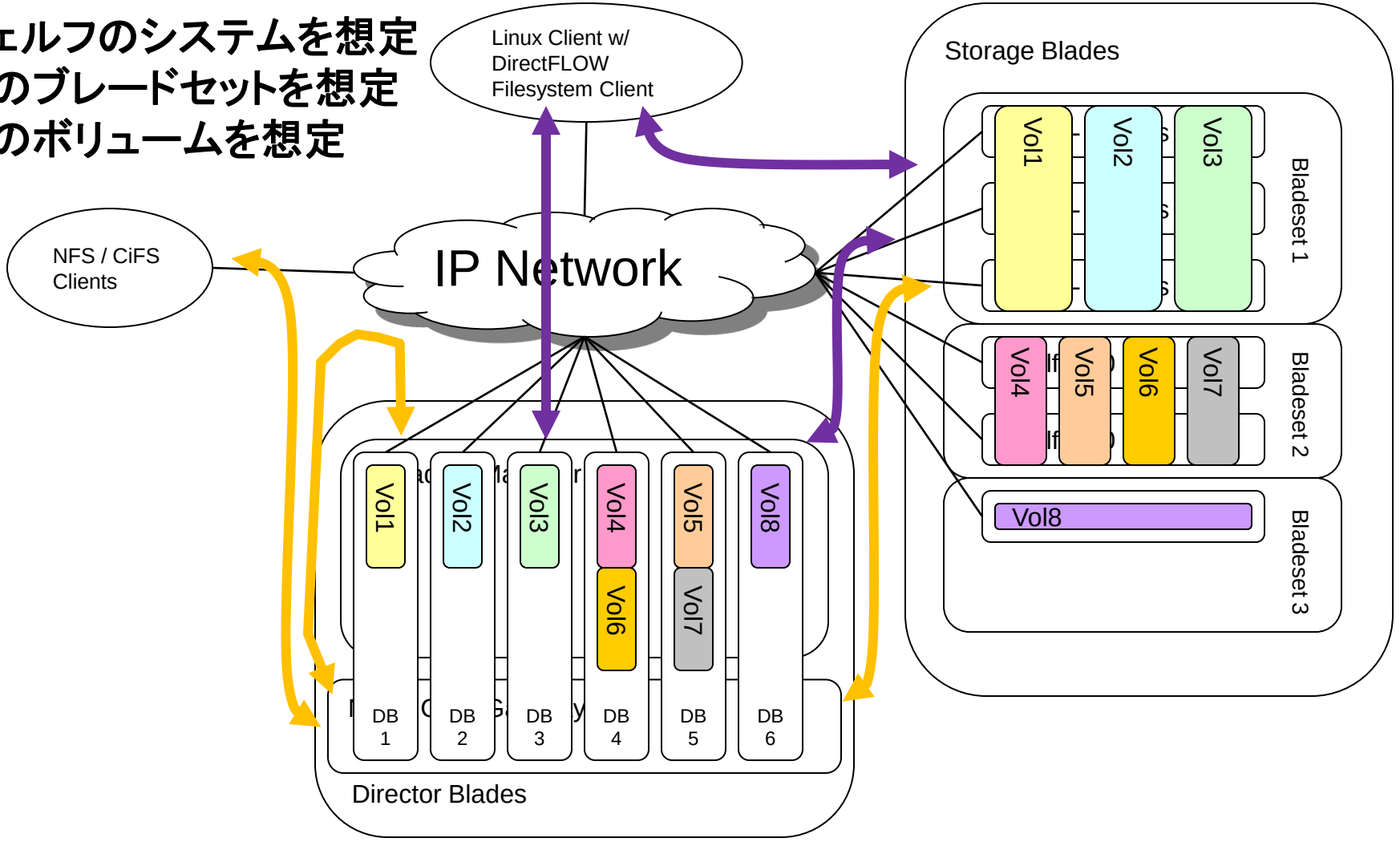
データフローの詳細
DirectFlow & NFS/CIFS

6シェルフのシステムを想定
3つのブレードセットを想定

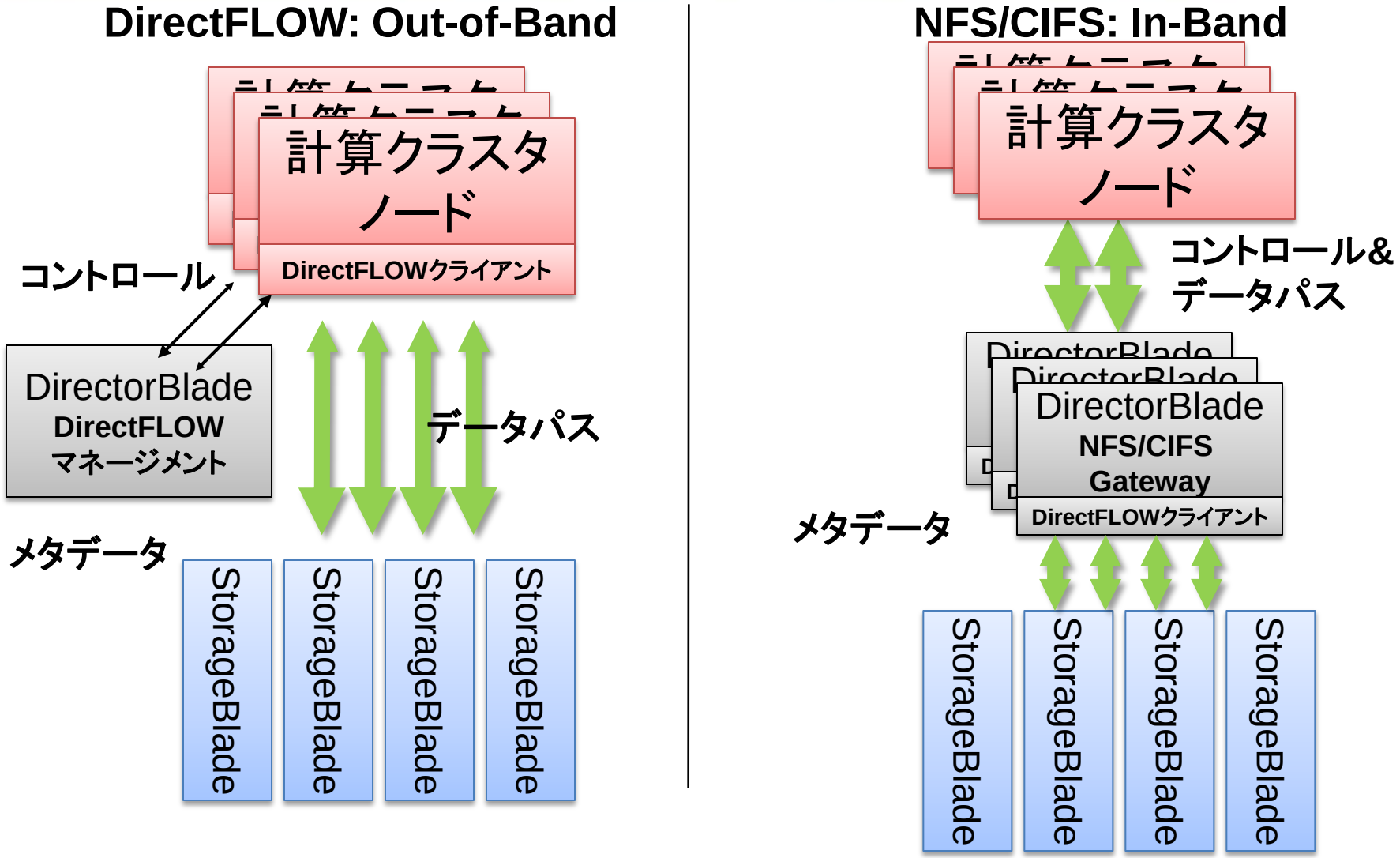


データフローの詳細

6シェルフのシステムを想定
3つのブレードセットを想定
8つのボリュームを想定

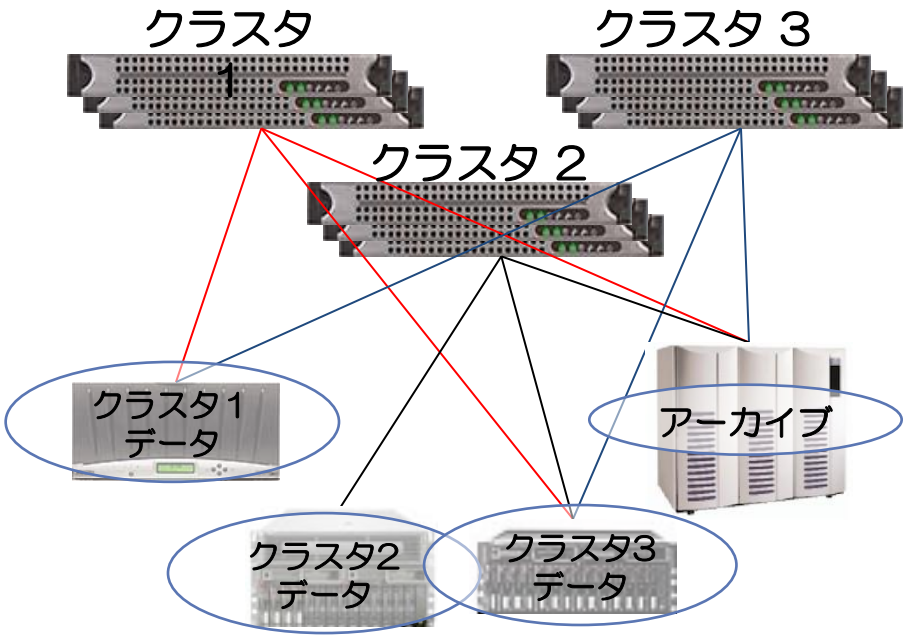


Panasasシステムモデル



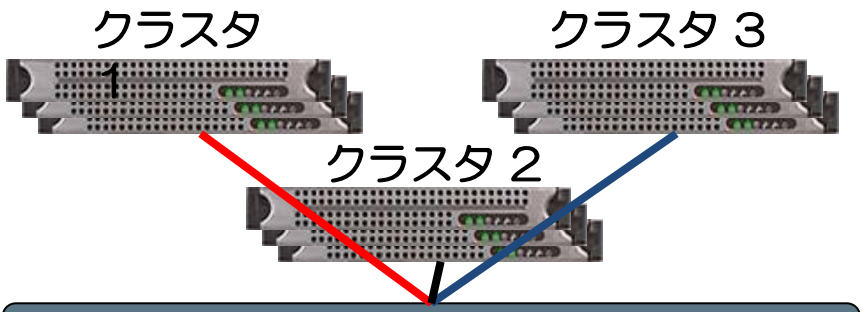
シングルグローバルネームスペース

- 物理的な境界も論理的な境界も存在しない
- クラスタ間でのクロスマウントやデータの移動の排除
- 自動的プロビジョニング：追加したブレードは自動認識され、ストレージプールに追加される

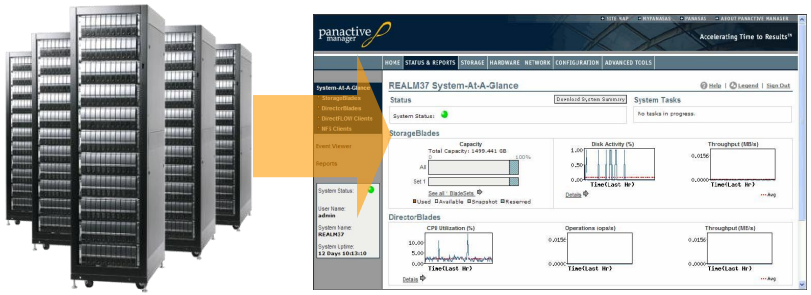


従来のストレージネットワーク

Panasasストレージクラスタ



シングルグローバルネームスペース



全てのデータを共有

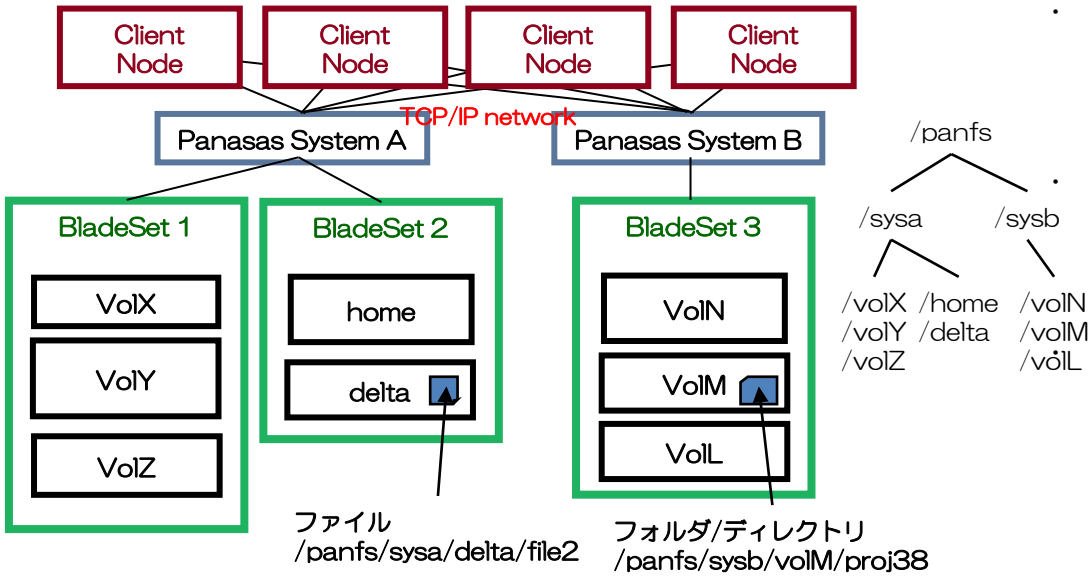
グローバルネームスペース

透過的なデータアクセス

- すべてのクライアントから同じパス名でファイル（例：/panfs/sysa/delta/file2）、フォルダ/ディレクトリ（例：/panfs/sysb/volM/proj38）へのアクセスが可能

柔軟なデータ管理

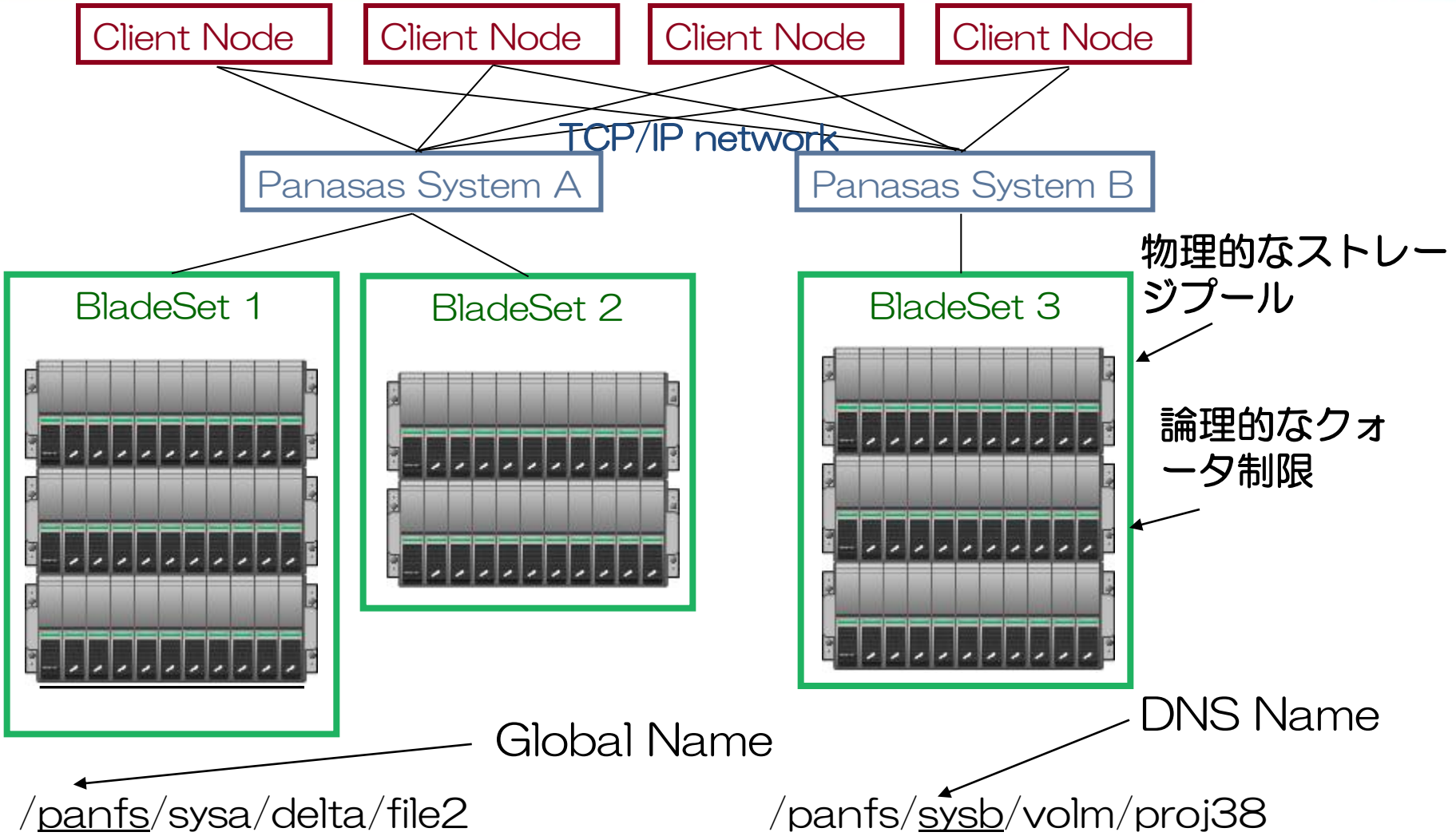
- 管理者はユーザのアクセス方法や利用方法に影響を与えることなく、ストレージの拡張や移動を行うことが可能
- データの管理業務における物理的な作業を大幅に減らすことを可能とし、作業に要する時間を短縮
- 管理者は一つのWEBページで、ロケーションが異なるストレージデバイスのデータ管理を行うことが可能



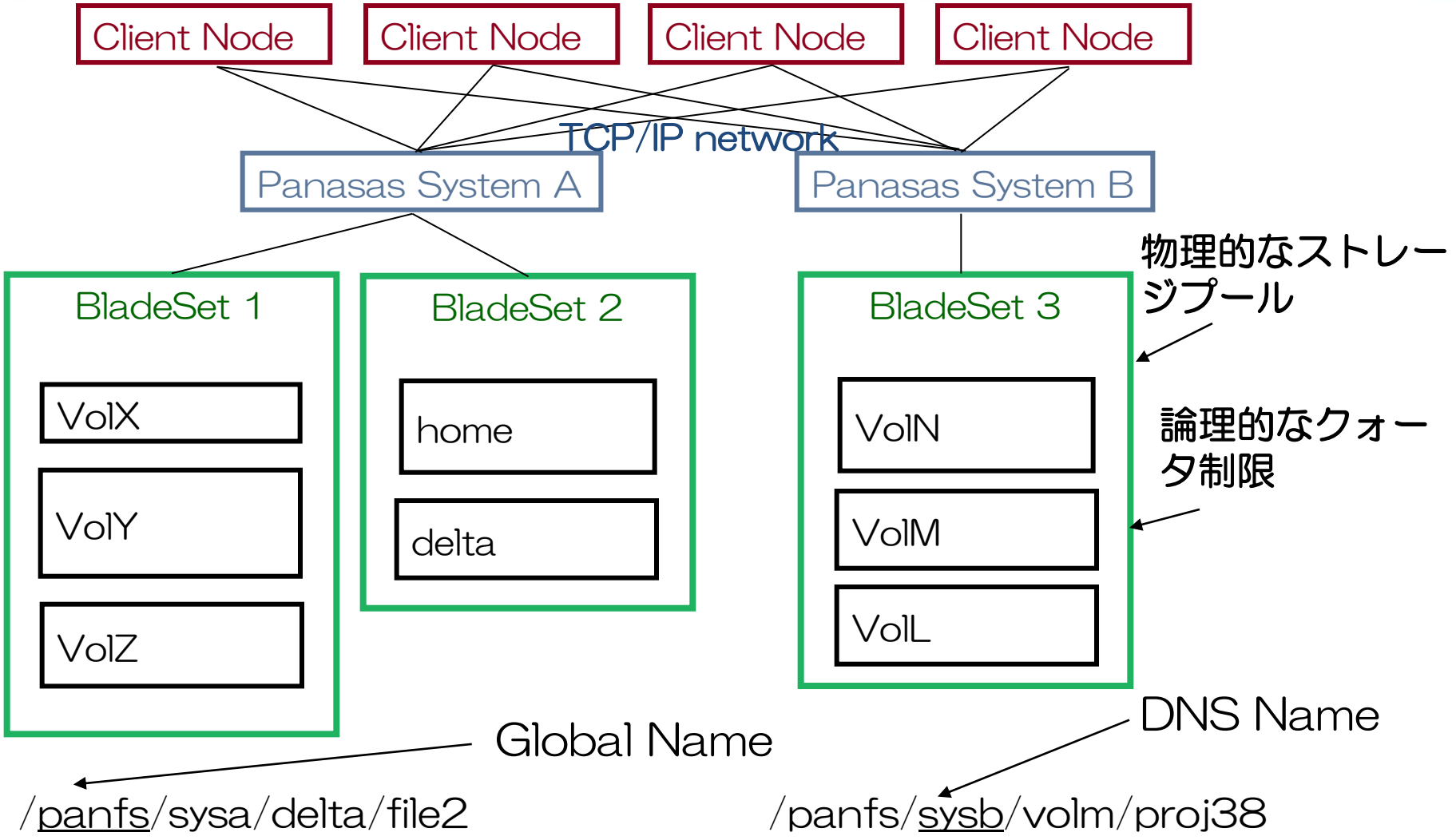
透過的な拡張

- Panasasのグローバルネームスペースは、ストレージ容量について制約のないプラットフォームを実現
- システムの再構成などをオンライン中に実行することも可能であり、ダウンタイムを最小化することを可能
- データ管理や移動はユーザに対して、透過的に行われ、データの保管場所などを気にすることなくデータへのアクセスが可能

グローバルストレージモデル



グローバルストレージモデル



BladeSetとは？

- ユニットとして管理される１つ以上のシェルフの集合体
- 少なくとも一台のDirectorBladeを含む必要がある
- BladeSetに含まれるStorageBladeはすべて同一の容量を持つことが必要（BladeSetが異なれば異なった容量のStorageBladeでも構成可能）
- BladeSetは、障害隔離として機能し、一つのBladeSet内での障害は他のBladeSetに伝播、影響しない
- Bladeの障害は、BladeSet単位で再構成を行うことが対処
- BladeSetはストレージクラスタ内の仮想的なストレージアレイ

BladeSetのサイズの検討

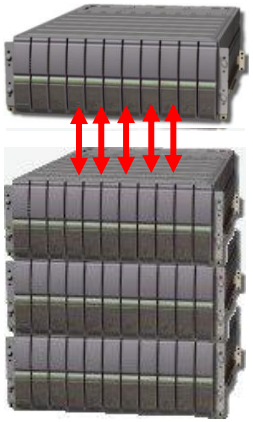
- 大きなBladeSet(多くのシェルフ）の利点
  - 少ないBladeSet数で運用が可能
  - 容量不足になる可能性がより少ない
  - 大規模なファイルについては、より高速に処理が可能
- 小さなBladeSet(少ないシェルフ）の利点
  - ‘ダブル・ディスク障害’の可能性がより低い
  - 一つのBladeSetの障害は他のBladeSetに影響しない（障害時などの対応がより容易）
- 小さなBladeSetをマージすることはできますが、大きなBladeSetを分割することは出来ない

統合されたシングルネームスペース

- シングルポイントでのシステム管理
  - データの孤立化の排除
- 全てのシステムデータに対して、一つのマウントポイント
  - DirectFLOW, CIFS and NFS
  - ローカルとリモートストレージシステム
- ネームスペースは、ボリュームによって柔軟にパーティションに分割可能
  - 個々に RAIDレベルと容量制限(Quota)の設定が可能(ActiveRAID)
  - Quotaの設定によって、顧客は、各ボリュームに割り当てるスペースの制限の設定が可能

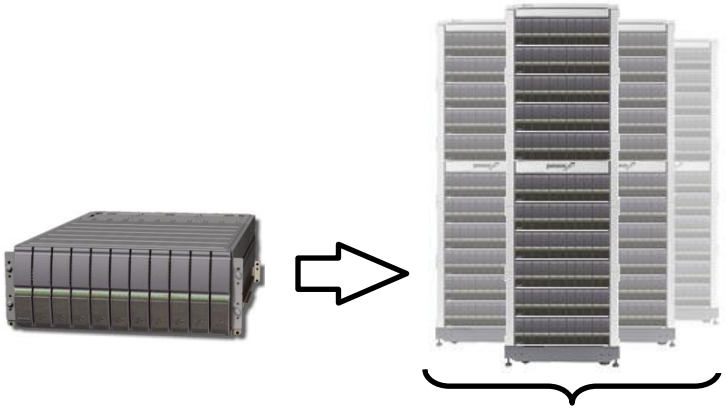
自動プロビジョニングによる容易な拡張

- オンラインプロビジョニング
  - 一つのDirectorBladeの設定を行ない、他の構成は、プライベートポート経由でのDHCPによって、構成を決定する
  - 新規ストレージは、シームレスにシステムに統合可能
  - オブジェクトベースのシステムは、古いデータの新しいストレージへの容易な移行を可能とする
- 制限なしでの拡張性
  - テラバイトからペタバイトまでの拡張性
  - シングルのシームレスなネームスペース



プライベートポート上でのDHCP構成

構成の読み込み
IPアドレスの設定
バージョンの適合



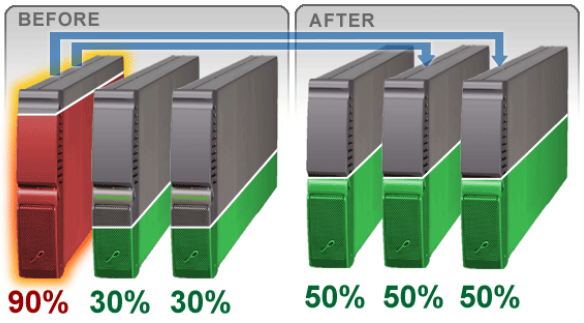
シームレスな
シングルネームスペース!

動的な負荷分散

StorageBlade容量

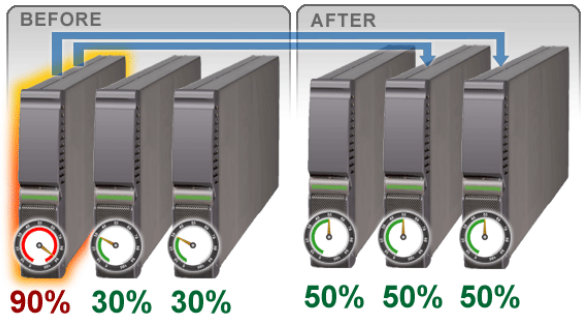
- 新しいデータは、より利用率の低いブレードに格納
- 必要な場合には、動的にデータを移動し、ブレード間での利用率の均一化を図る

StorageBlade性能



- 最大の性能が得られるようにデータオブジェクトの分割を行う
- 動的に利用率の高い”hot” ブレードからオブジェクトを移動する

DirectorBlade性能



- ストレージクラスタは、DirectorBladeの利用率に応じて、クライアントからのデータ処理を各DirectorBladeに配置する
- 必要な場合には、再配置する

Panasasが提供する運用管理機能

インテグレートされたHW/SW

- 既存のITインフラに容易に組み込むことが可能で、その導入を短時間で可能とする

スケーラブルなネームスペース

- スケーラブルなシングルストレージプールを実現することで、数テラバイトからペタバイトまで容易に拡張し、運用管理も容易
- アプリケーション開発をよりシンプルにすることが可能

エンタープライズクラスのストレージソリューション

- 自動的なプロビジョニング、動的なロードバランス、最先端のRAID技術などを提供
- 仮想ボリュームとディスク・クォータ、スナップショット、容量管理レポート、障害やシステムリソースに関する警告など

システム管理と高可用性機能

- 予防的システムマネージメント
  - データとディスクのスキャンを継続的にバックグランドで実施
  - 問題発生の可能性のあるブレードのシステムからの切り離し
- リアルタイムでのクライアントのモニター
  - クライアントからのI/O要求と処理性能をモニターし、ボトルネックを解析
- クォーラム(Quorum)ベースでのクラスタマネージメント
  - 3台のクラスタマネージャによるシステム運用
  - システム状態のレプリケーション
  - クラスタマネージャはブレードとクライアント状態のモニター

システム管理と高可用性機能

- メタデータフェイルオーバー
  - クラスタマネージャによるプライマリーバックアップコントロール
  - ジャーナル処理のための低レイテンシログレプリケーション
  - アプリケーション透過なクライアント認識フェイルオーバー
- クライアントフェイルオーバー
  - DirectFLOW は、フェイルオーバー時にアプリケーションの状況を維持
  - NFS/CIFS サーバは、 DirectorBladeをマイグレート
  - ロックサービス(lockd/statd) はフェイルオーバーシステムと統合
- オンライン中のクライアントアップグレード
  - 利用中でもクライアントソフトウエアのアップグレードが可能

データの保護と可用性
ソフトウエアでの信頼性向上

- 高速な再構成が可能なRAID 10 及び 5 を提供
  - ファイル毎にRAID構成とデータの分散が可能：障害時の影響を最小化
  - 障害時の再構成はオンライン中にバックグランドで実行されるため、サービスの停止は不要
- Panasas Tiered Parity
  - マルチレベルでのパリティ設定によるデータ保護
- 信頼性の高いオペレーティングシステム
  - システムのサービスのフェイルオーバ、ファイルサービスのフェイルオーバ、メタデータ処理のフェイルオーバなど
  - 各ブレードで動作しているOSについては、ミラーリングを行う
  - ディスクの状況や熱、ファンの動作などを細かくモニターして、予防診断を行い、障害に対応する
  - スケーラブルな高速バックアップのサポート

System-at-a-Glanceページ

- Panasasの状態を概要的に参照するページです
- 過去のデータを表示させたい場合は統計レポート機能（Reports）を利用します

Panasas Backup

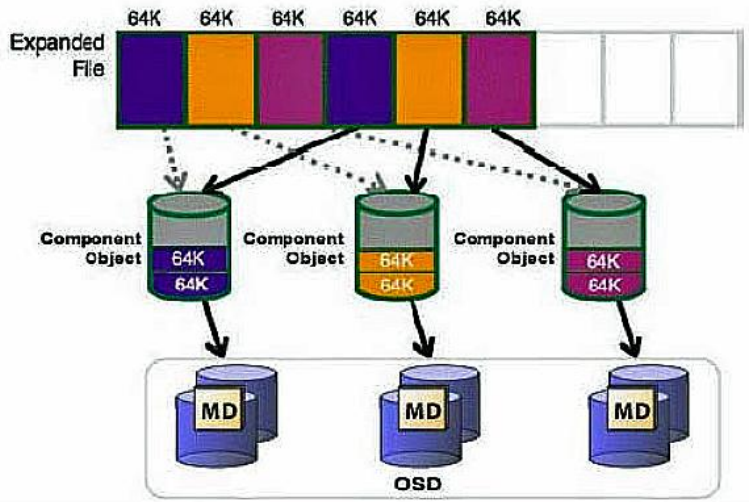
- Panasas ActiveImage
  - スナップショット機能の提供
  - PAS8/9（デフォルト）、PAS7（オプション）
- Panasas間データ転送
  - PanasasのパラレルIO機能を最大限に活用するパラレルコピーツールの提供
  - Linuxのrsync なども複数プロセスで同時実行可能
  - 専用のレプリケーションサーバの提供も可能（Option）

Panasas社が提供する構築ブロック

1. オブジェクトストレージデバイス（OSD）
   - データと属性のコンテナ
   - iSCSI/OSDインターフェイスとしてのSNIA T10 を標準インターフェイル
   - Panasas社のStorageBlade は、商用OSDとして初めての製品
2. 分散&パラレルファイルシステム
   - ブロックマネージメントは、オブジェクトストレージインターフェイスのバックで動作
   - クライアントからのIOは直接、パラレルにオブジェクトストレージデバイスに送られる
   - ファイルマネージメントは、メタデータマネージャ全体で処理される
   - 障害発生時の対応
3. スケーラブルなPanasas社のRAIDシステム
   - ファイルを複数のコンテナオブジェクトに分割
   - パラレルRAIDの再構築

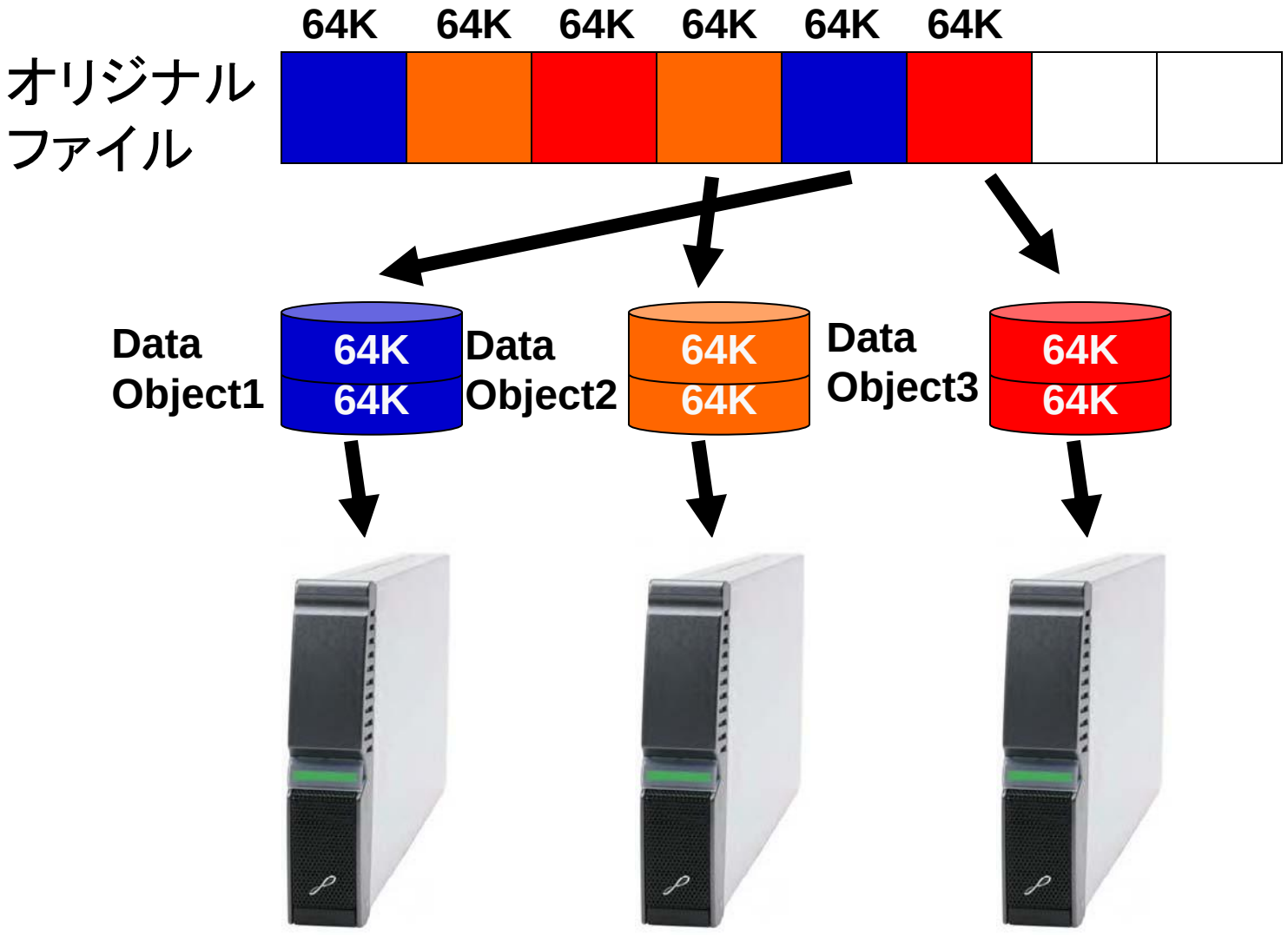
Panasas RAID
PanFS - Panasasファイルシステム

- ストライピング/RAID
  - 個々のファイル毎に複数のOSD上にファイルを分割
  - 各ファイル毎に異なったデータレイアウトとRAIDレベルの設定が可能



- ストライプユニット
  - 一つのOSDにアサイン（64Kがデフォルト）
- RAIDレベル（0/1/5)
- データ分割幅
  - ストライピングされるOSDの数
  - ファイルの最大の転送速度（バンド幅）が得られるように設定
- パリティストライプ幅（RAID 5設定）
  - パリティの値は、クライアントがデータの書き出し時に計算

Panasasファイルシステムモデル



Panasas RAID – Advanced RAID

- Panasas RAID – Advanced RAID
  - Panasasが提供するRAIDシステムは、ディスク単位で管理するものではなく、ファイル単位で設定される
  - 特定のStorageBladeをパリティとはしない
- ファイルの取り扱い
  - ファイルは、ひとつの仮想オブジェクトとして取り扱われる
  - この仮想オブジェクト（ファイル）は、複数のコンポーネントオブジェクト上に格納される
  - 一つのコンポーネントオブジェクトが、StorageBladeに格納される

Panasas RAID
RAIDスペアと再構成の取り扱い

従来のRAID

- ホットスタンバイされたスペアを利用してのファイルシステムの再構成が必要
- 残ったディスクからデータを読み込み、（ホット/コールド）スタンバイのスペアにデータを書く込む必要がある
- したがって、システム内の全ドライブを利用しての再構築となるため、システムに大きな負荷をかけることになる
- 再構成に要する時間は、交換したディスクへのデータの書き込みの要する時間によって決まる

Panasas RAID

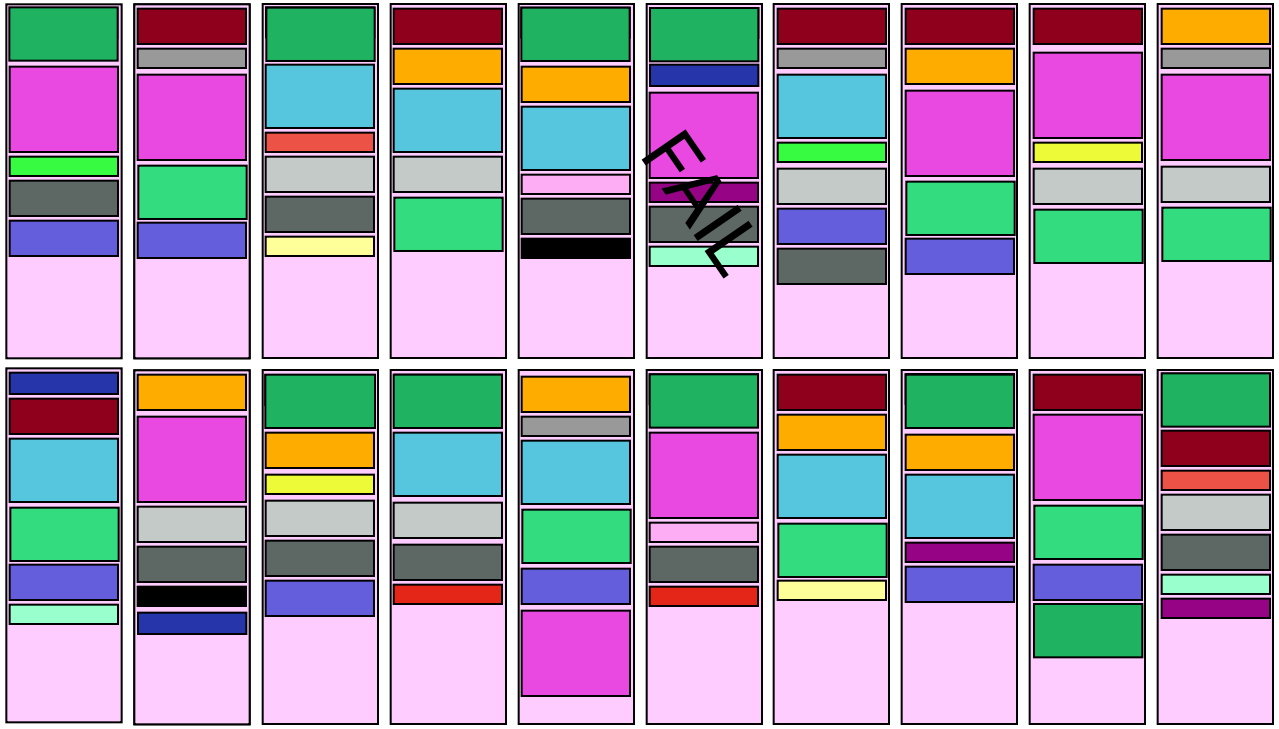
- 一つのスペアに対して再構成を行うのではなく、Pansasのストレージクラスタは、BladeSetで定義される全てのStorageBladeにスペア領域を分散する
- スペア領域を分散させることで、処理性能の向上を図る（全StorageBladeが利用可能）
- 再構成は全StorageBladeでその処理を行うことが可能であり、特定の部分がボトルネックとなる可能性が低い

Panasas RAID
ファイルシステム再構成

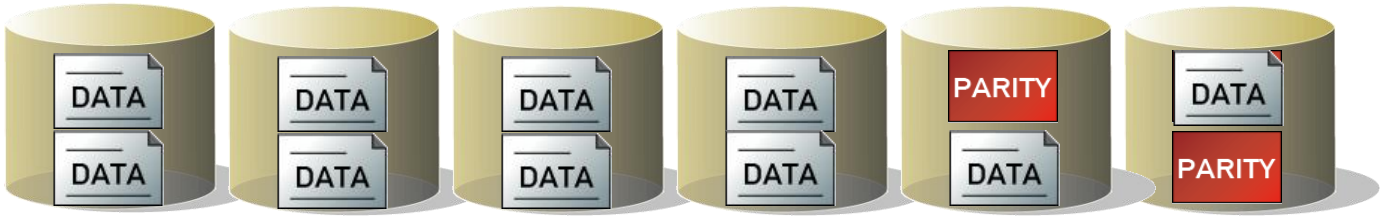
- ファイルは、別々のStorageBlade上にコンポーネントオブジェクトとして分散して配置
- ファイル属性の情報は2つのコンポーネントオブジェクトで2重に保持
- RAID処理は、ランダムに分散して処理

2-shelf BladeSet

ディスクミラー又は9-OSDパリティストライプ



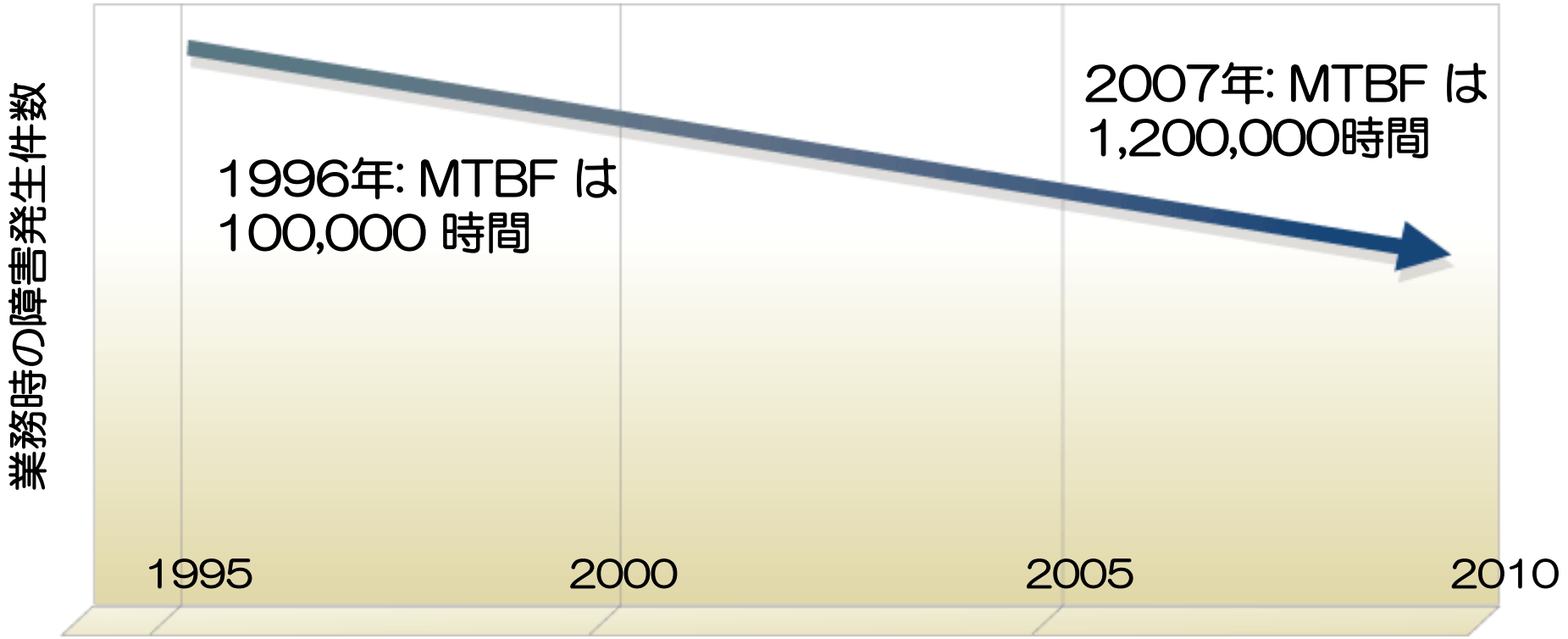
1996年当時のRAIDシステムの状況



RAIDはストレージシステムの性能と信頼性の向上のために利用されている

- RAIDは、ディスク障害時にデータを失くことを防ぐために、ディスクのセットを利用
- 1996年時点であ、RAIDでのデータ回復に失敗しても、~50GB程度をテープなどから復元するとしても、数時間でその作業を終了することが可能

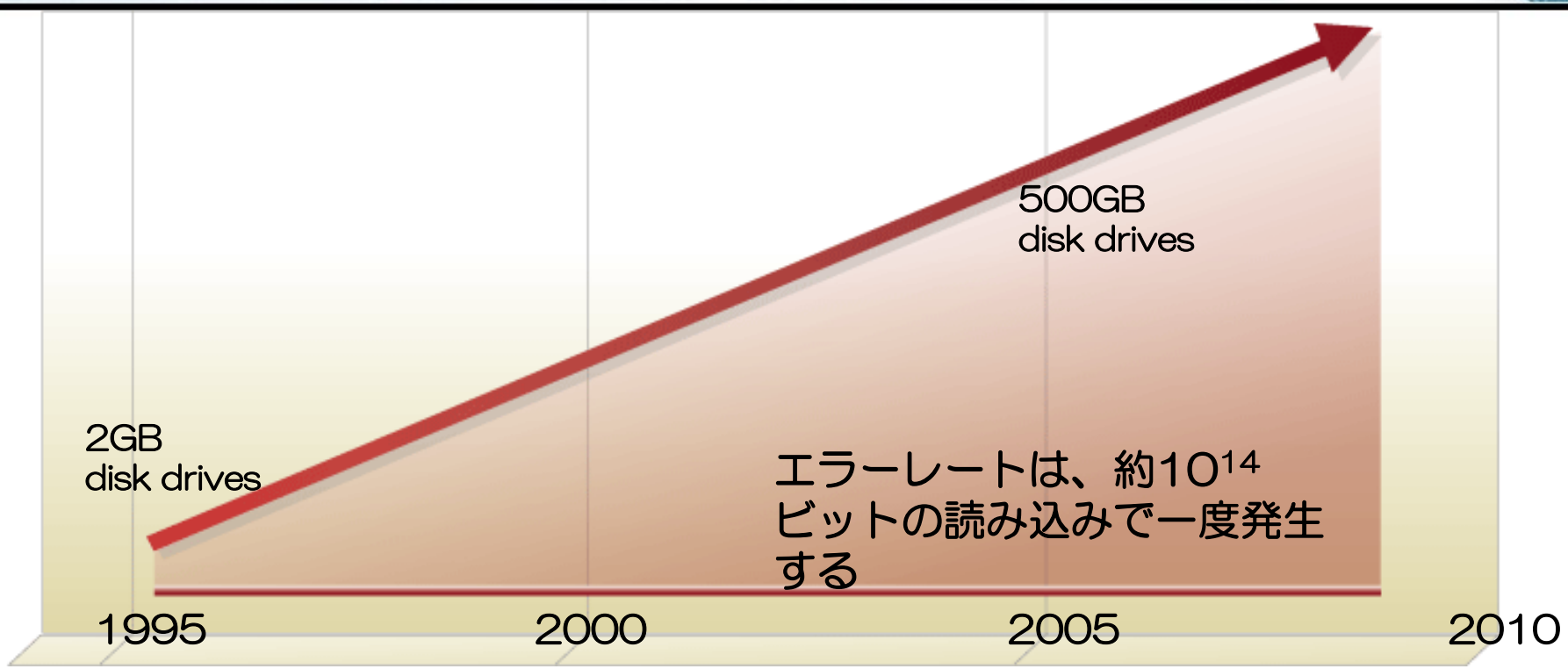
ディスクの信頼性の継続的な向上



- RAID 5 によるデータプロテクションは、MTBFが100,000時間であれば、十分
- 現在のディスクは、1996年当時と比較して、10倍以上も信頼性が高い

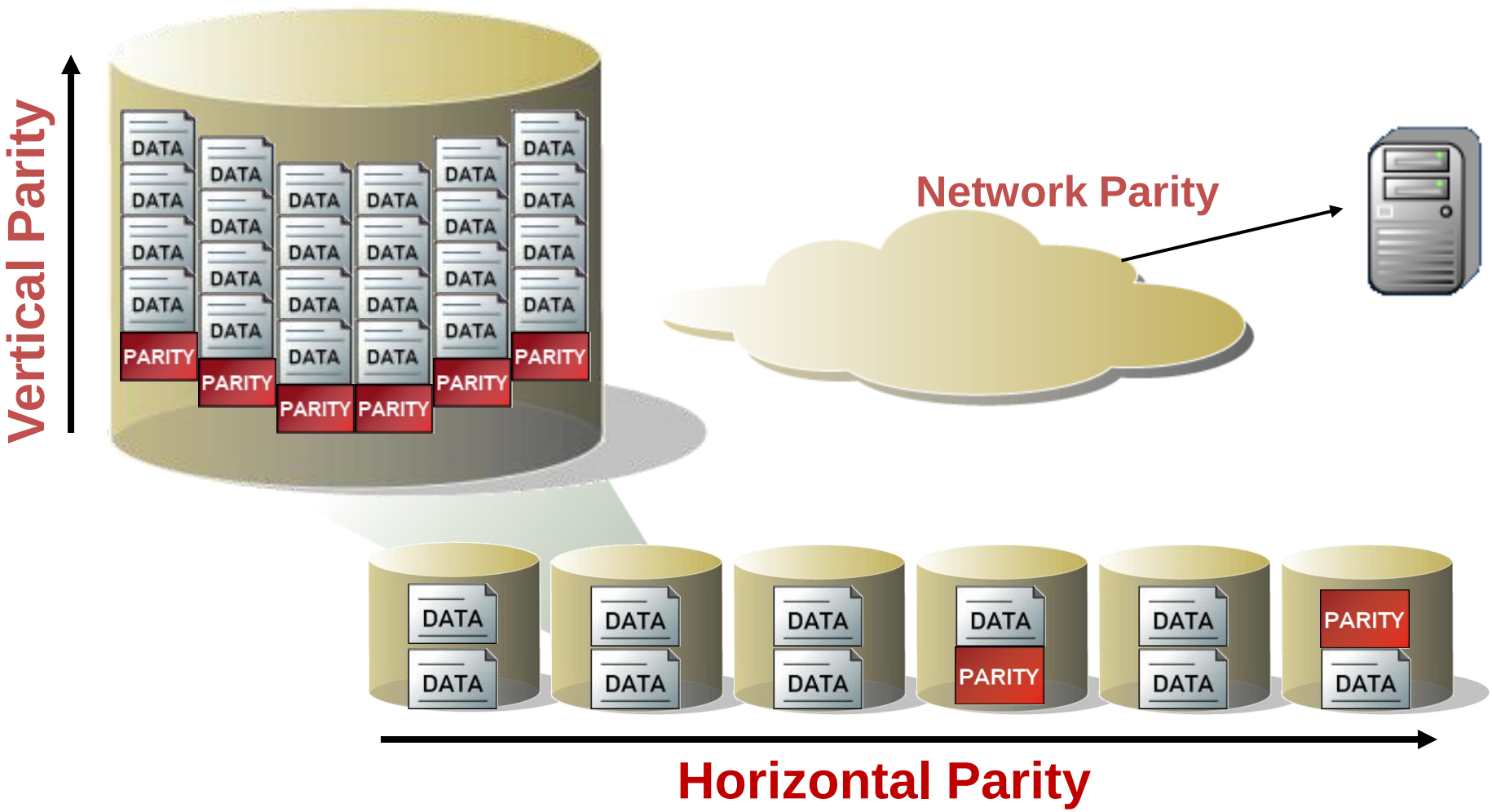
RAIDの信頼性は、10倍以上高いはず

ディスクの密度の向上



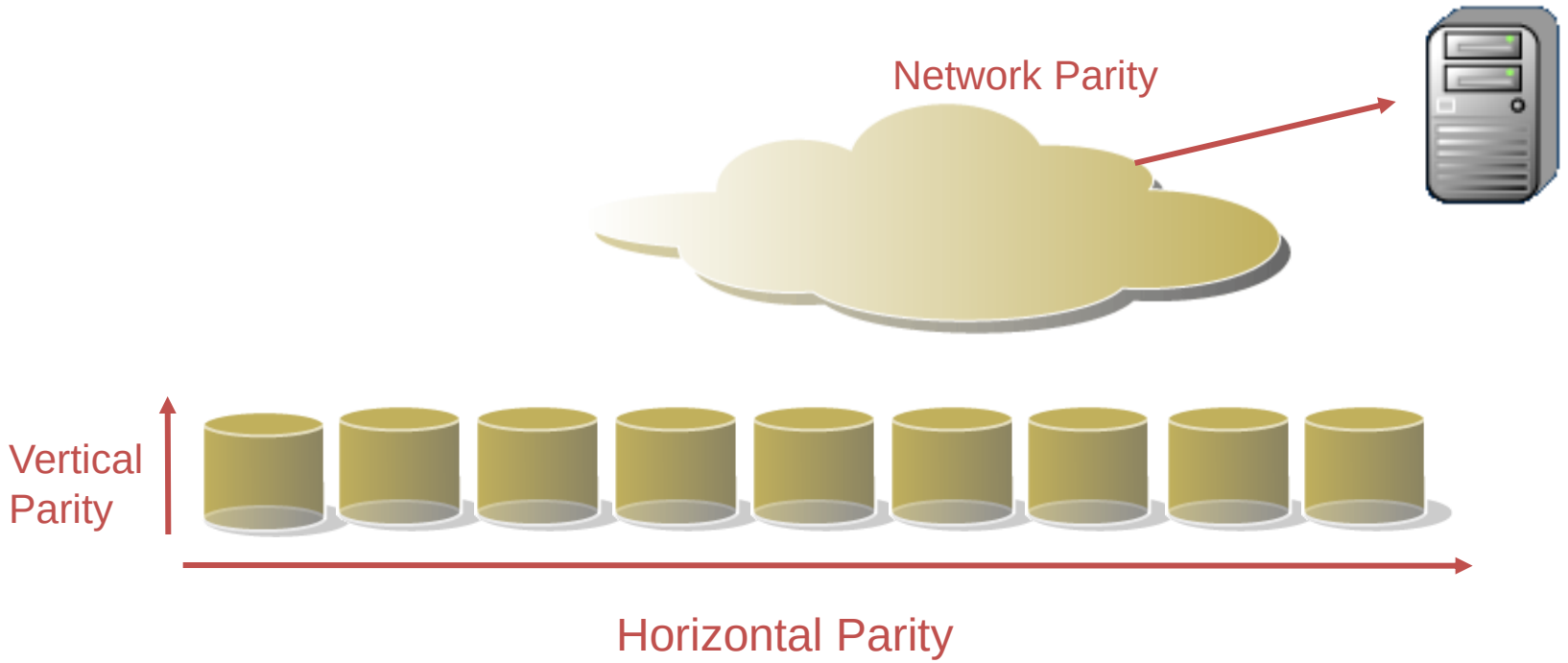
- ディスクの密度は10年前を比較して、250倍以上高密度になっている。しかし、メディアエラーの発生頻度は、ほぼ、一定。
- メディアエラーが発生するとRAIDの再構成に失敗し、データが失われる。
- 再構成時のメディアエラーの発生確率は、非常に高い可能性を持つ。

Panasas Tiered Parity



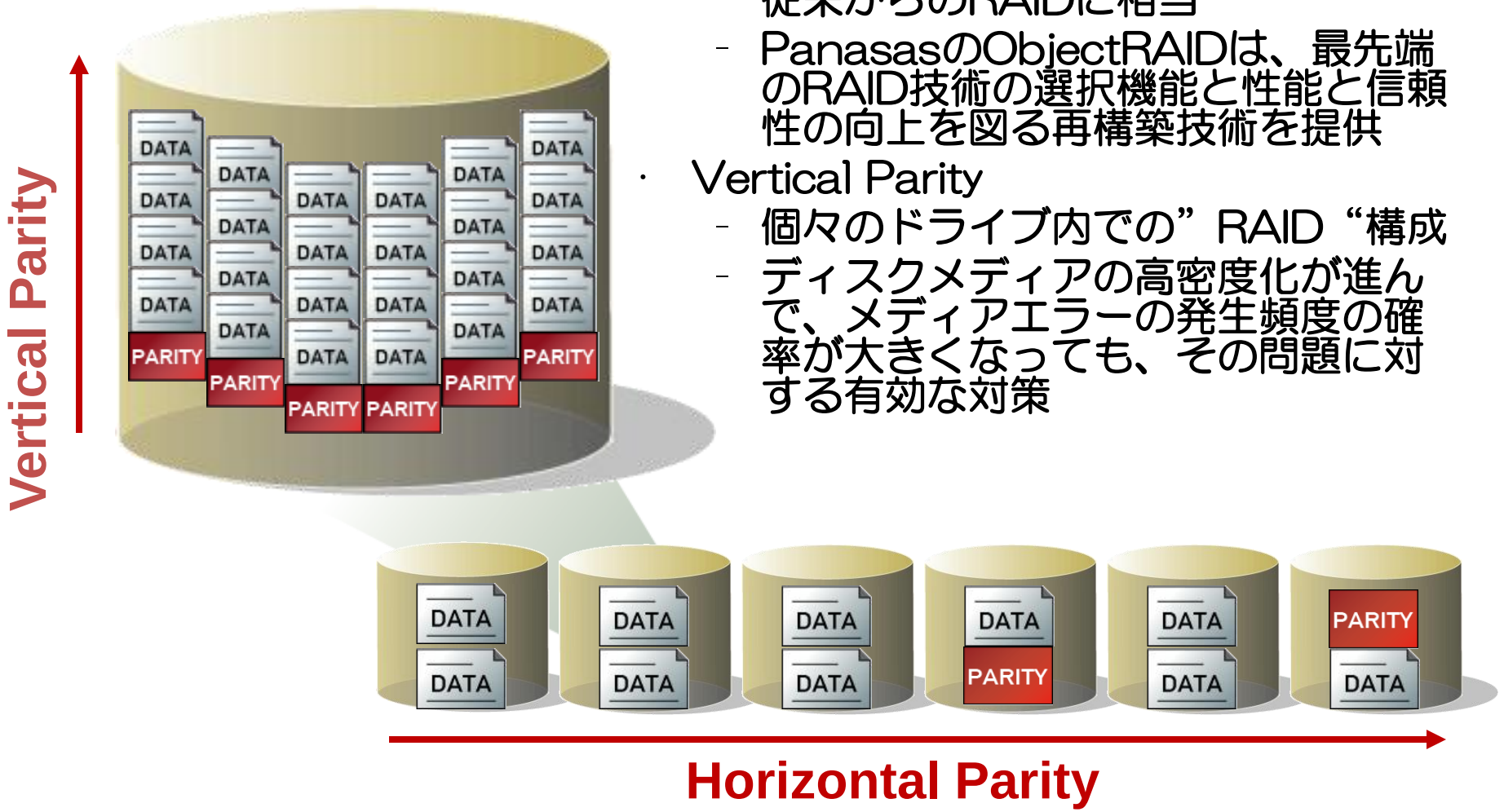
Panasas Tiered Parity

- 各Tierオペレーションは、独立したパリティの処理を行うことが可能であり、エラー検知とデータ修正を行う
- PanasasのTiered Parityが提供する3つのパリティ処理は、互いに相互補完



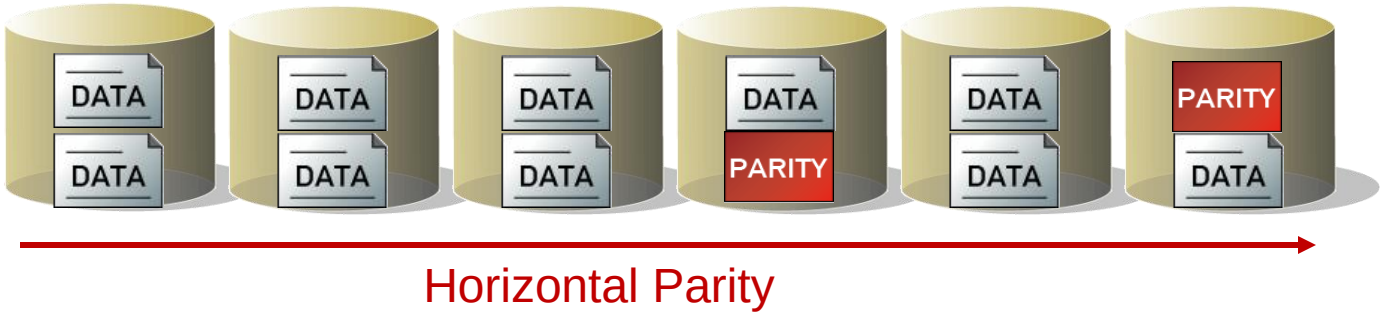
Panasas Tiered Parity

- Horizontal Parity
  - 従来からのRAIDに相当
  - PanasasのObjectRAIDは、最先端のRAID技術の選択機能と性能と信頼性の向上を図る再構築技術を提供
- Vertical Parity
  - 個々のドライブ内での”RAID“構成
  - ディスクメディアの高密度化が進んで、メディアエラーの発生頻度の確率が大きくなっても、その問題に対する有効な対策



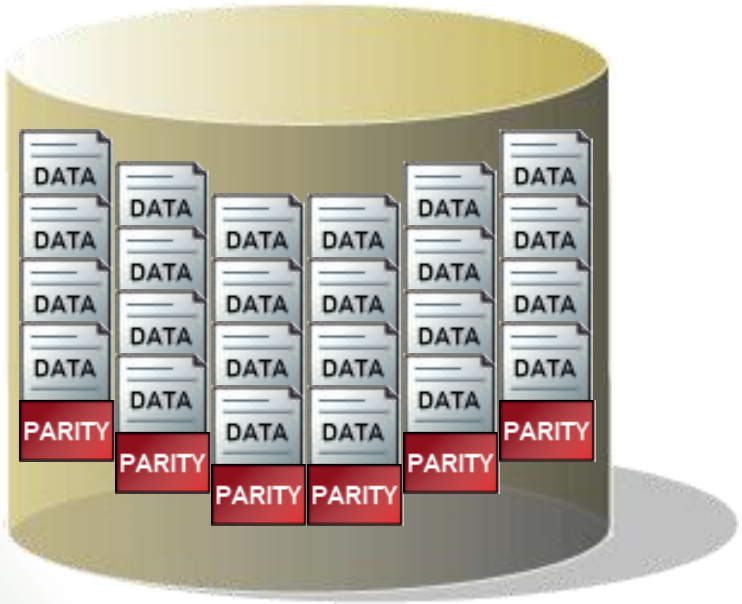
Horizontal Parity

- 従来からのRAIDに相当
- PanasasのObjectRAIDは、最先端のRAID技術の選択機能と性能と信頼性の向上を図る再構築技術を提供
- ObjectRAIDは、典型的なケースで、従来のRAIDに対して、10倍以上　高速での再構築が可能
- ObjectRAIDは、再構築をパラレルに実行し、また、再構築で必要となるスペースを最小限にすることが可能であり、時間とスペースの双方を大幅に減らすことが可能



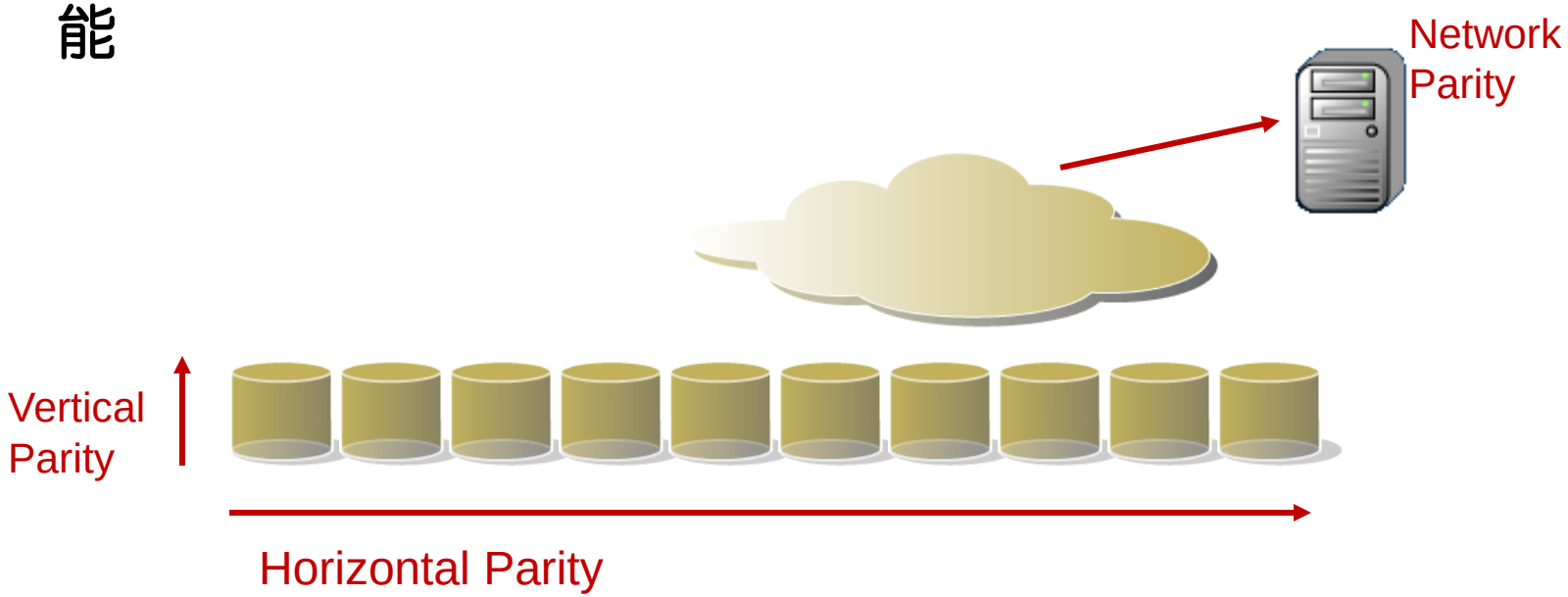
Vertical Parity

- 個々のドライブ内での”RAID“構成
- 今後、更にディスクメディアの高密度化が進んで、メディアエラーの発生頻度の確率が大きくなっても、その問題に対する有効な対策
- ダブルやトリプルでのHorizontal Parity の設定を不用



Network Parity

- ディスクドライブのサイズがより大きく、又、ドライブ数が増えた場合、ネットワークを介して格納され、管理されるデータのボリュームは増加
- ネットワーク間でのデータに関して、クライアント自身がそのデータを読み込む場合にそのデータの検証を行うことを可能



ディスクドライブの高密度化に対する対応・対策

問題点と課題	他社の提案	Panasasの提案
メディアエラー発生頻度の上昇 - RAIDの障害とRAID再構成時の再構成の失敗の可能性	パリティの数を増やす	Vertical Parity は、ディスクドライブの信頼性の向上を図ります。これは、メディアエラーの発生に際して、そのデータエラーの排除を修復を可能とします。RAID Array として利用されるディスク単体の信頼性とエラー回復を図ることを可能とします。
RAID再構成に要する時間の増大とRAID再構成に失敗した場合のデータ復元に要する作業負荷	RAID arrayのサイズを小さくし、同時にパリティの数を増やす	Horizontal Parity は、通常のRAIDと同じように複数のディスクドライブ間でのRAIDグループのデータの信頼性を提供します。Panasas社のObject RAIDは、より高速に、効率よくシステムの再構築を可能とします。
データ破損はメモリ、スイッチ、ネットワークインフラを通過するデータ量の増加によって、ストレージシステム以外の部分で発生する可能性が高い	なし	Network Parity は、ストレージシステムとクライアント間でのデータ統合を行います。ネットワークインフラが引き起こすデータの破損をクライアント自身がデータ検証を行うことで防ぐことが出来ます。

Panasas Tiered ParityとRAID 6比較

	RAID 5	RAID 6	Panasas Tiered Parity
シングルHDD障害	Yes	Yes	Yes
シングルHDD障害+メディアエラー	No	Yes	Yes
ダブルHDD障害	No	No*	No
検知出来ないデータ破損	No	No	Yes

* RAID再構成時にメディアエラーが発生した場合

スケーラブルシステムズ株式会社

まとめとして

Panasa ActiveStorの特徴

- 圧倒的な性能（スループットとバンド幅の双方）
- 容易に導入可能で、利用も簡単
- システムの増設も自由
- 既に、多くの導入実績を持つ
- NFS/CIFSでの利用とDirectFlowの双方を同時に利用可能
- 新技術への対応：pNFSやTiered Parity

Panasa ActiveStorの利点

	テクノロジー	利点
性能	DirectFLOW（高いバンド幅をもつクラスタ構成）	並列にダイレクトアクセスが可能
	クラスタ化したNASシステム	N多重での同一ファイルのエクスポートが可能
	クラスタ化した大規模キャッシュ	大規模データセットに対応
信頼性	リカバリー機能を付加したクラスタ	システムの再構成を高速に実行
	高速でのアーカイブ機能	バックアップ/リストアを高速に実行
	Panasas Tiered Parity	信頼性を損なうことなく性能とスケーラビリティの向上
運用管理	クラスタの利用効率の向上	システムのロードバランスが容易
	プロビジョニング	仮想化されたストレージ
	高機能なクラスタマネージメント	統合されたH/WとS/W
	pNFS	パラレルストレージの標準化

Panasas ActiveScale ストレージクラスタ

クラスタコンピューティングのために設計されたシステム

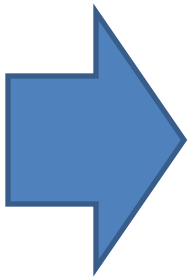
機能とその利点	Panasas ActiveStor	NAS サーバ (NetApp, EMC, start-ups)	SAN ファイルシステム (Lustre, GPFS)
ターゲットとするアプリケーション	Batch + Interactive	Interactive	Batch
高いバンド幅	○		○
クライアント数のスケーラビリティ	○		○
ストレージ容量のスケーラビリティ	○		○
NFSとCIFSのサポート	○	○	
統合システム	○	○	
可用性	○	○	
高いランダムIO性能	○	○	

継続的な技術革新

- ストレージ技術に関する継続的な技術革新
- pNFS:パラレルNFS
  - Network File System v4 プロトコル規格の拡張
  - パラレルかつダイレクトでのデータアクセスが可能
  - ストレージデバイスは、複数のストレージプロトコルをサポート
  - NFSサーバはデータパスに直接介在しない
- Panasas Tiered Parityアーキテクチャ
  - 信頼性に関する問題を解決するエラー検知とデータ修正のためのアーキテクチャ
  - つのパリティ処理を提供（相互補完）

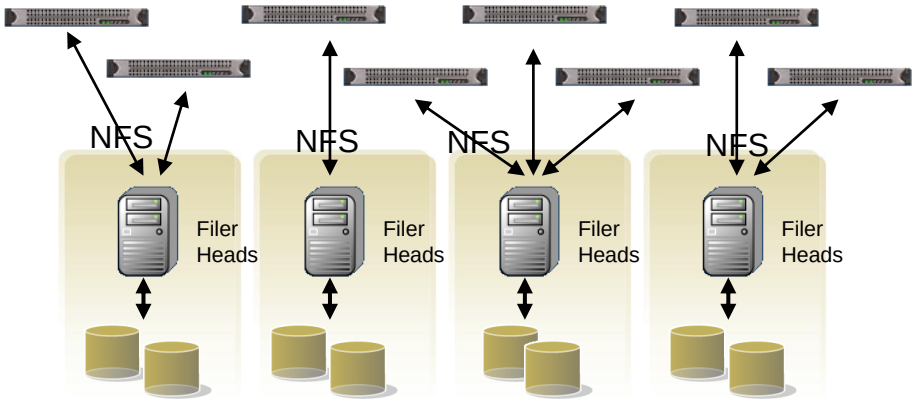
パラレルI/Oの標準化問題

- パラレルストレージを提供するベンダー間で互換性の欠如
  - Panasas PanFS
  - IBM GPFS
  - EMC MPFSi (High Road)
  - IBRIX Fusion
- オープンソースの活動も独自に進展
  - Red Hat GFS
  - PVFS
  - Lustre
  - オープンソースの製品間でも、相互に互換性がない



パラレルI/Oの標準化は、コスト削減やユーザの選択肢の広がりなどの点で多くの利点がある

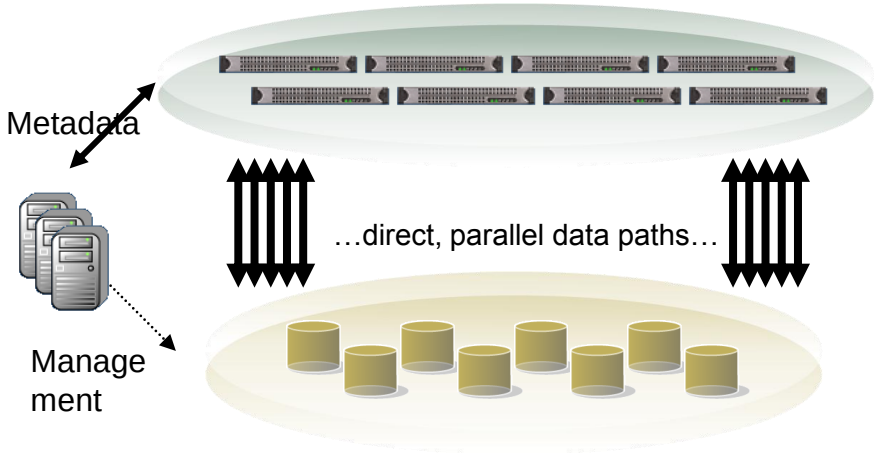
パラレルストレージ



“ストレージは独自に点在”

Filerヘッドが、I/O 性能のボトルネックとなる

複数のストレージの運用管理は容易ではない

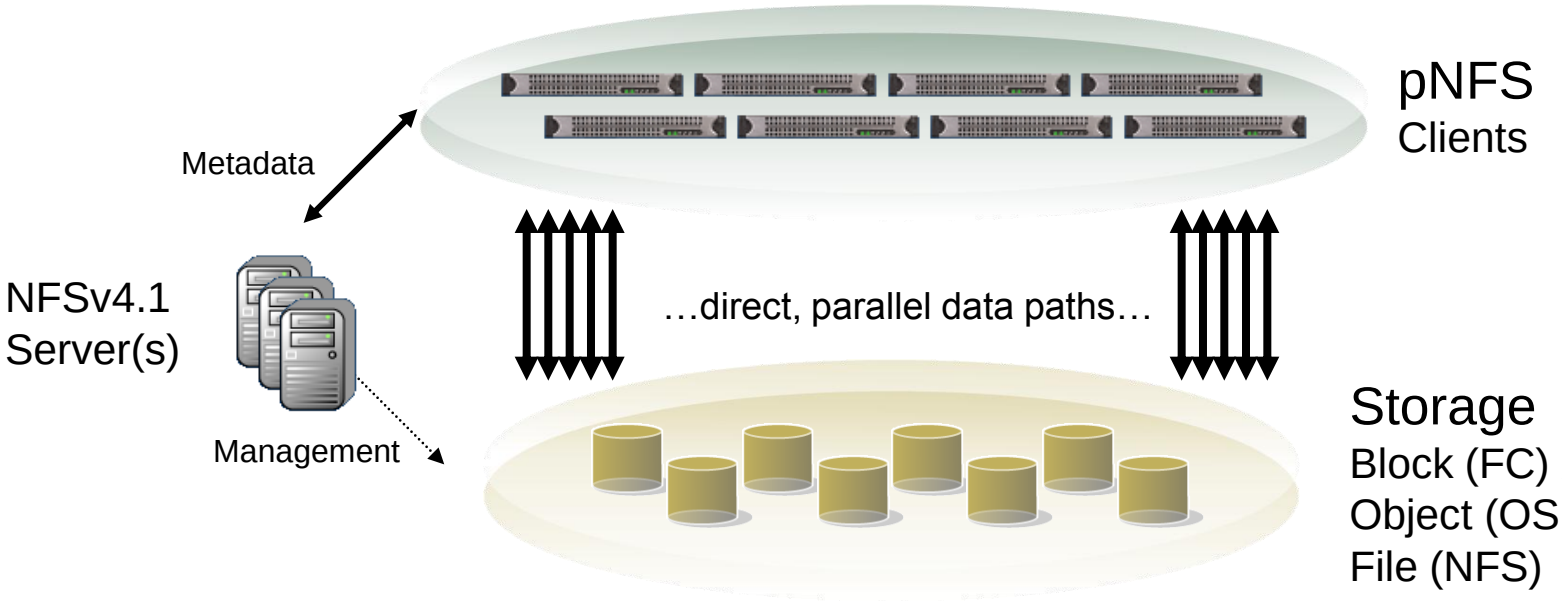


“パラレルクラスタストレージのプール”

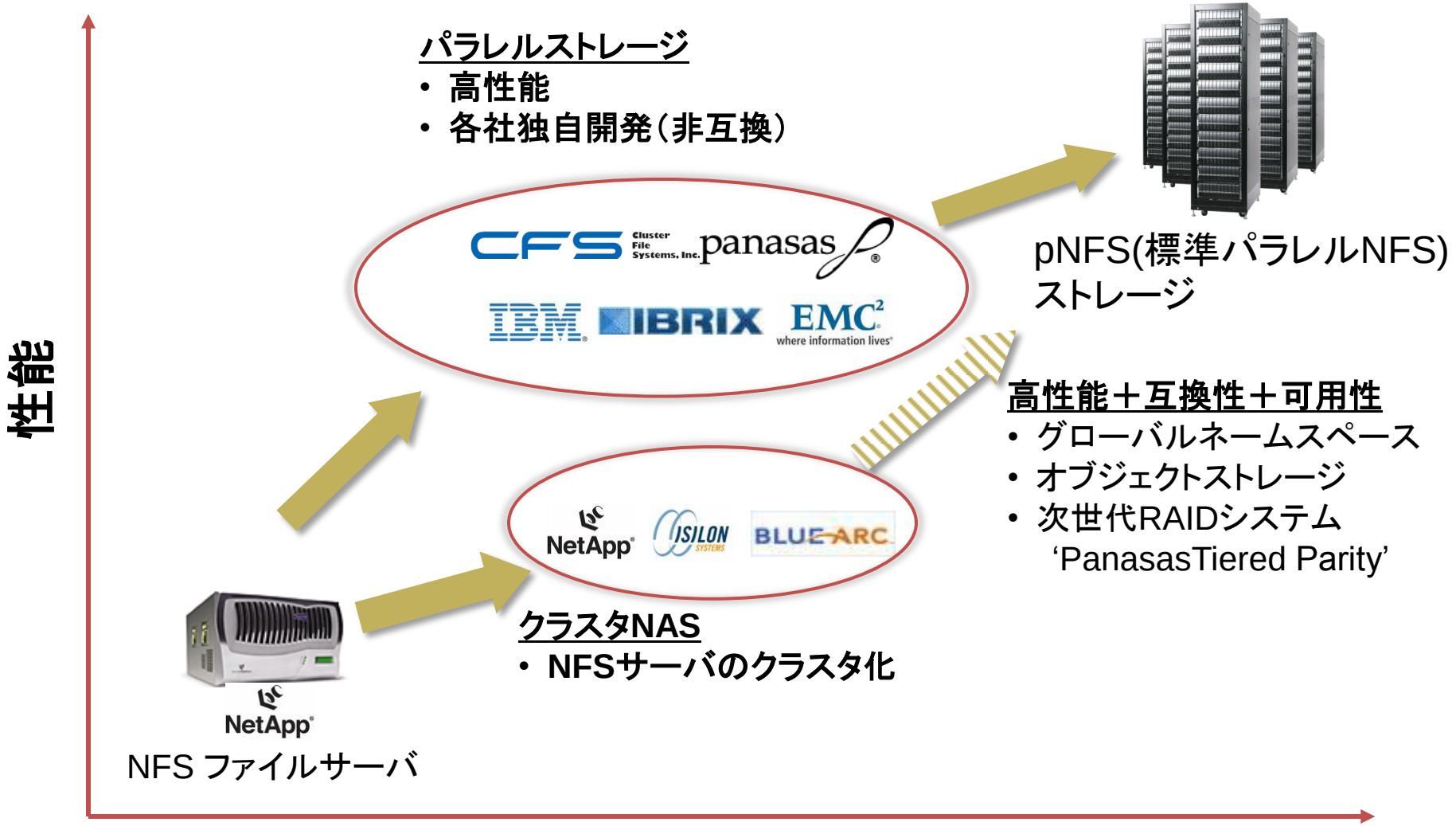
Filerをデータパスから排除することで、I/O性能のボトルネックと運用管理の問題を解決

pNFS: 標準パラレルNAS

- pNFS は、Network File System v4 プロトコル規格の拡張
  - パラレルかつダイレクトでのデータアクセスが可能
  - ストレージデバイスは、複数のストレージプロトコルをサポート
  - NFSサーバはデータパスに直接介在しない



高性能ストレージシステムの将来



お問い合わせ

0120-090715
携帯電話・PHSからは（有料）
03-5875-4718
9:00-18:00（土日・祝日を除く）

WEBでのお問い合わせ
www.sstc.co.jp/contact