Panasasストレージクラスタソリューション

グローバルネームスペースの活用

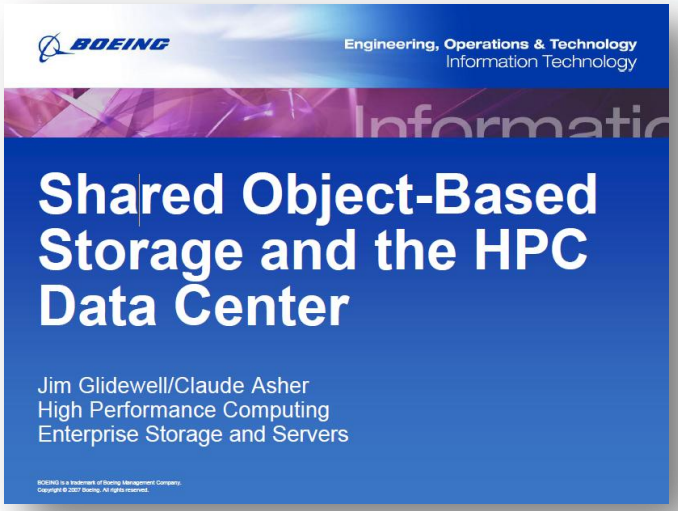
スケーラブルシステムズ株式会社

データセンタ導入事例

データセンター事例
ボーイング社データセンター

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



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

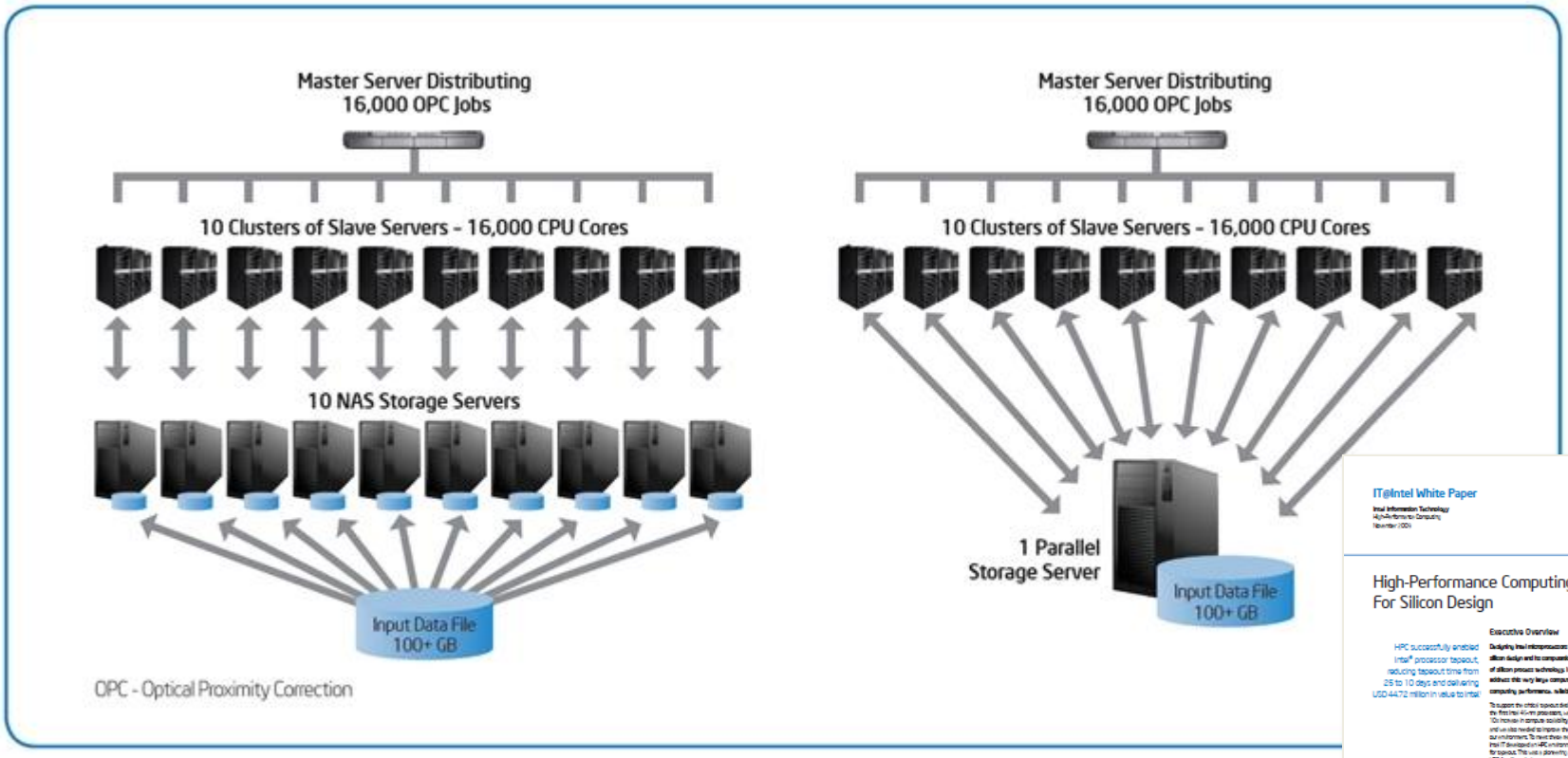
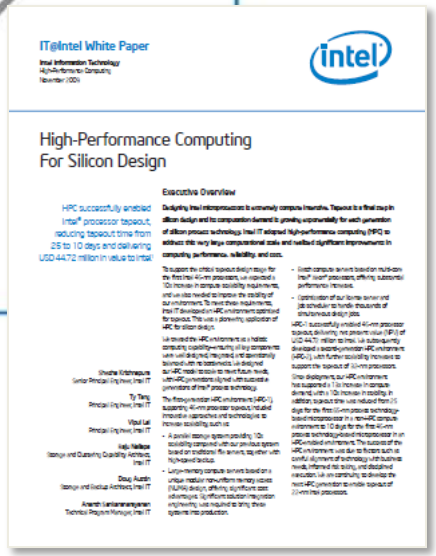


Figure 4. Consolidation with the high-performance computing (HPC) parallel storage environment.


IT@Intel White Paper
intel
High-Performance Computing
For Silicon Design

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 Intel Information Technology High-Performance Computing November 2009
High-Performance Computing For Silicon Design

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

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.

48計算ノード
InfiniBand
Panasas
ネットワークスイッチ

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

ストレージに関する課題

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

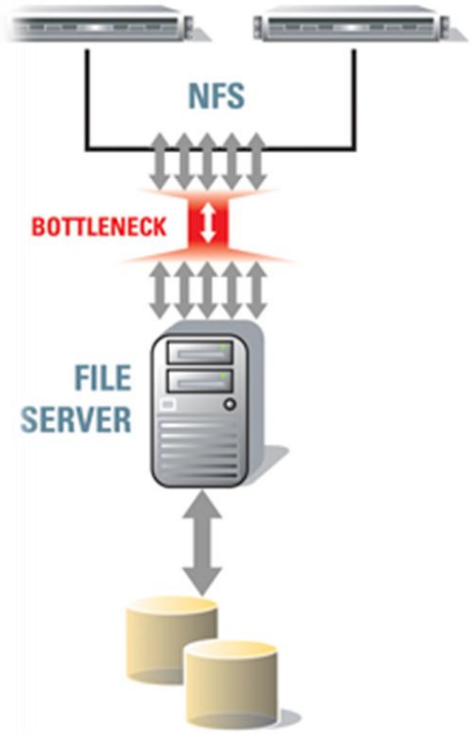
従来のネットワークストレージ

BOTTLENECK

バックアップ/リストア

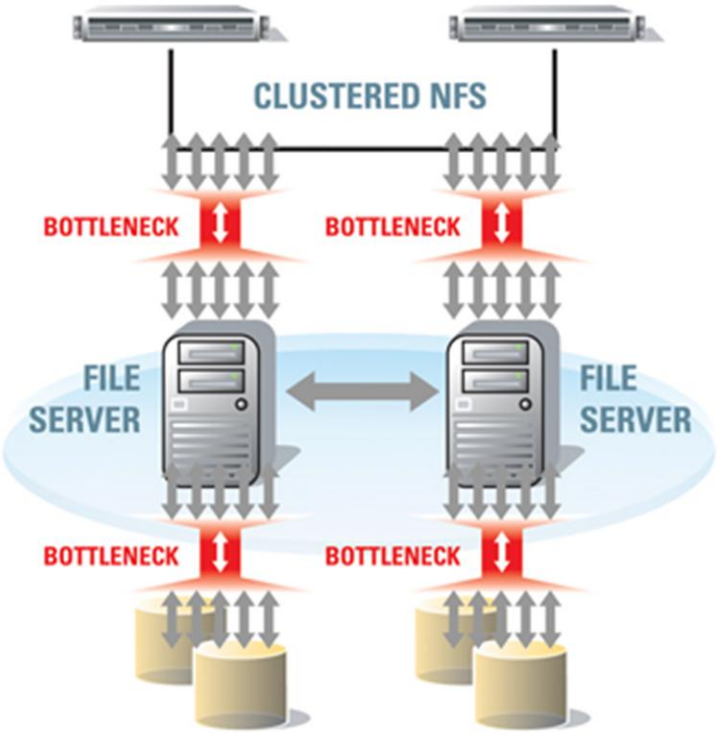
バックアップ/リストア
- バックアップ処理のためのストレージシステムの負担
- バックアップ実施のタイミング
- 高速でのバックアップの問題

ストレージアーキテクチャ



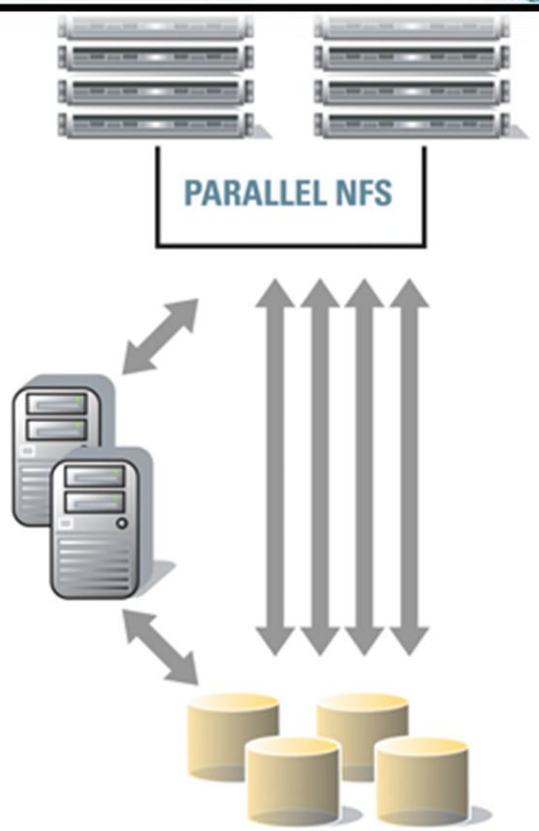
NAS

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

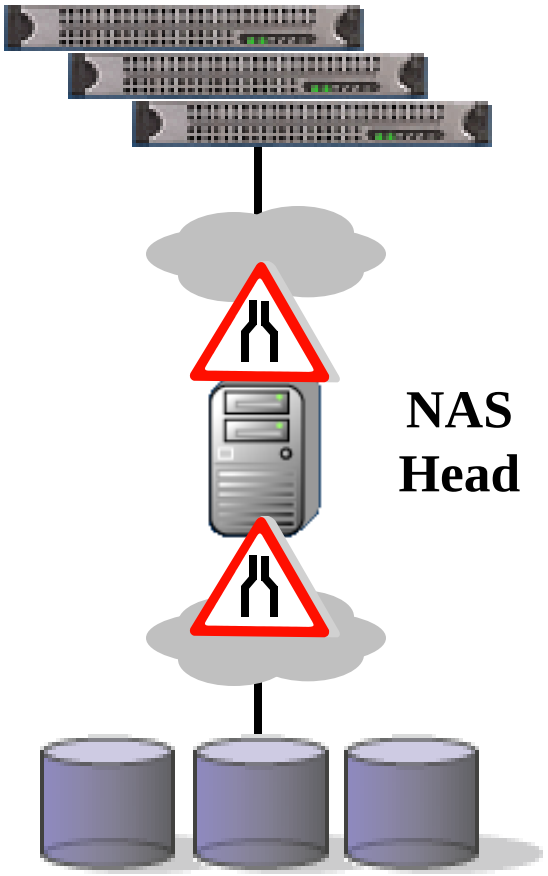


PARALLEL STORAGE

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

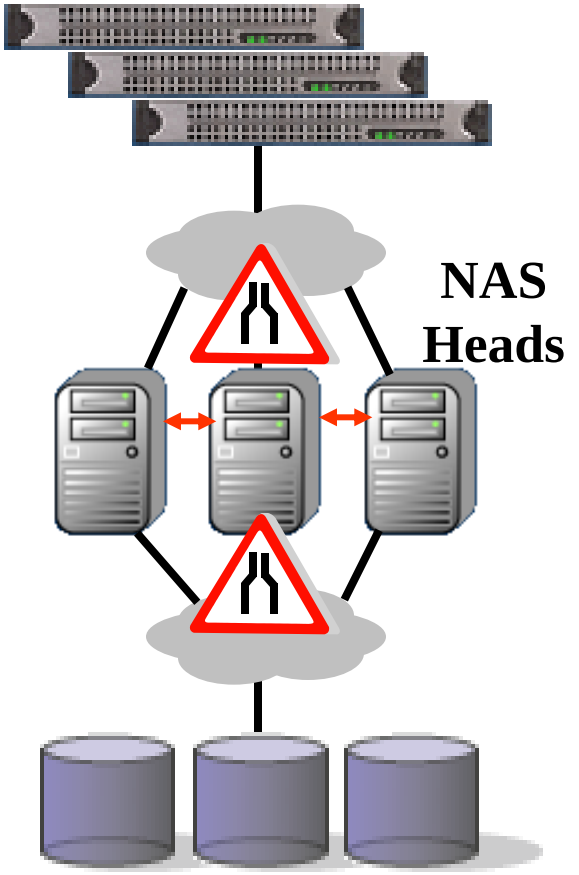
Network-Attached Storage：NAS
ネットワークアタッチトストレージ

- ファイルサーバは、ファイルレベルで、ストレージをエクスポート
  - NFS/CIFは広く利用されている
  - NFSは、唯一の業界標準のファイルシステム
- スケーラビリティは、サーバのハードウエアによって制限される
  - 中規模のクライアント数（数十から100程度）
  - 中規模のストレージ容量（数テラバイト）
- データ転送能力の限界以内であれば、非常に優れたモデル
  - 複数のストレージは個別に配置
  - サーバの能力によって、ファイルアクセスのバンド幅が制限される
- NetApp (ONTAP 7.x), Sun/HP/IBM NAS, SnapServer, EMC Celerra, whitebox Linux



クラスタNAS

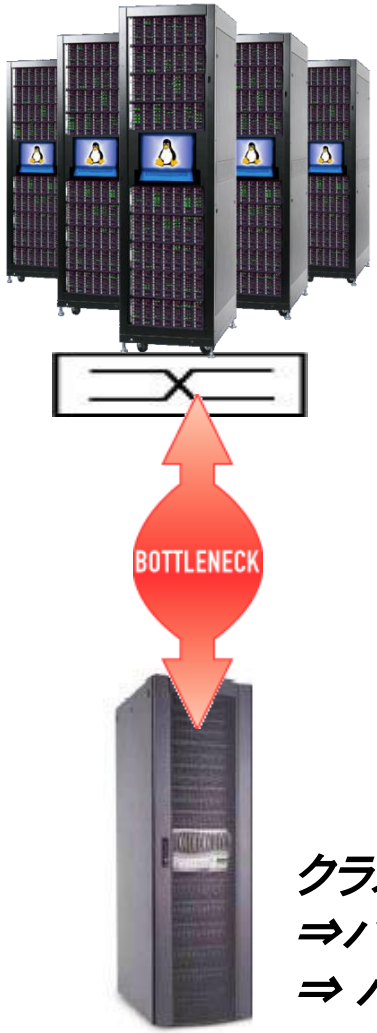
- シングルヘッドのNASよりもよりスケーラブル
  - 複数のNASヘッドがバックエンドのストレージを共有
  - 個々のNASヘッドの性能がボトルネックとなる。また、NASヘッドを追加することでコストは高くなる。
- 2つの主要アーキテクチャ
  - データを‘オーナー’ヘッドへ転送
  - クラスタSANファイルシステムからNASとしてエクスポート
- NFSは動的なロードバランス機構を提供出来ない
  - クライアントは、いずれかのノードに恒久的にマウントされる
- GPFS、Isilon OneFS、IBRIX、Polyserve、NetApp GX、BlueArc、Exanet ExaStore、ONStor、Pillar Data、IBM/Transarc AFS、IBM DFS



クラスタ利用時のボトルネック

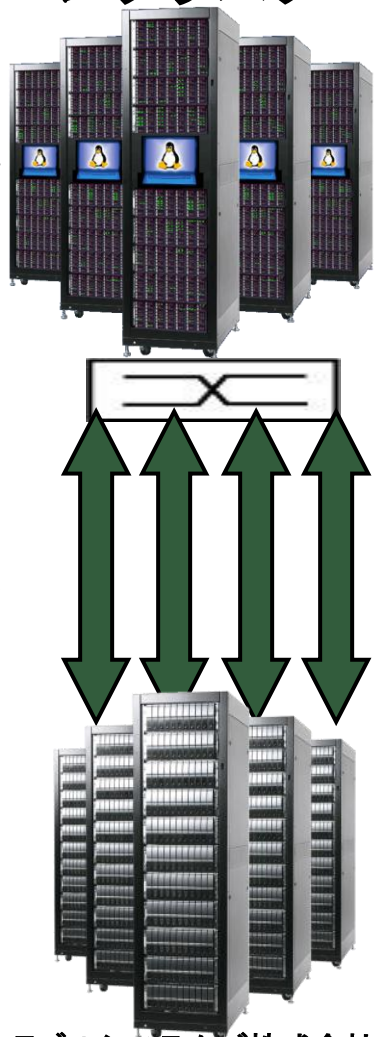
一般のNFSサーバ

- ボトルネック
  - ストレージに対する単一のデータパス
- 問題点
  - スケーリング
  - 限定されたIOバンド幅
  - システム拡張の限界
  - 柔軟性の欠如
  - 高価なシステム構成



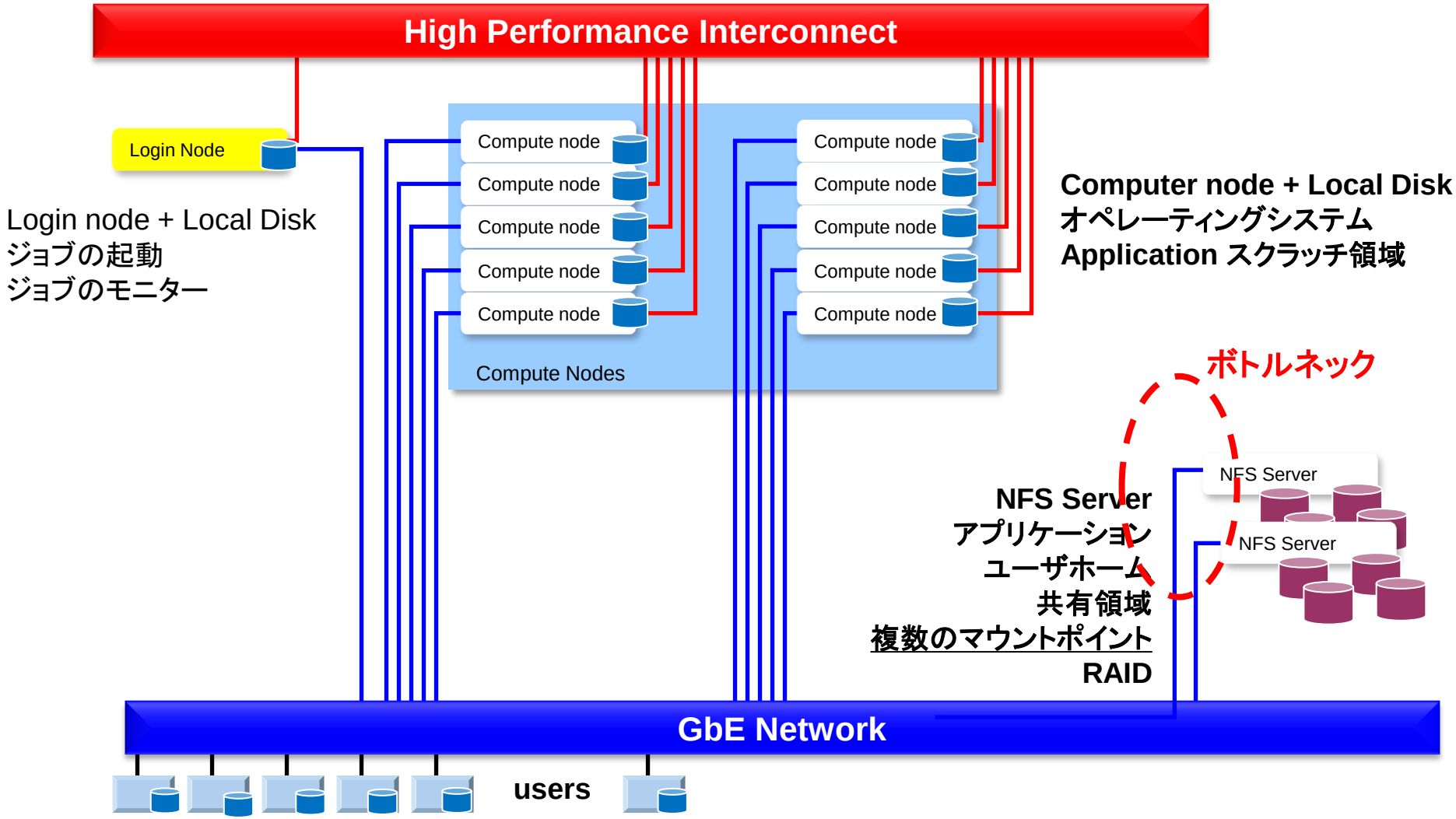
Panasas ストレージクラスタ

- ボトルネックの解消
  - ストレージに対するパラレルなデータパス
- 利点
  - スケーラビリティ
  - 高いIOバンド幅
  - グローバルネームスペース
  - 容易な運用管理
  - 低価格



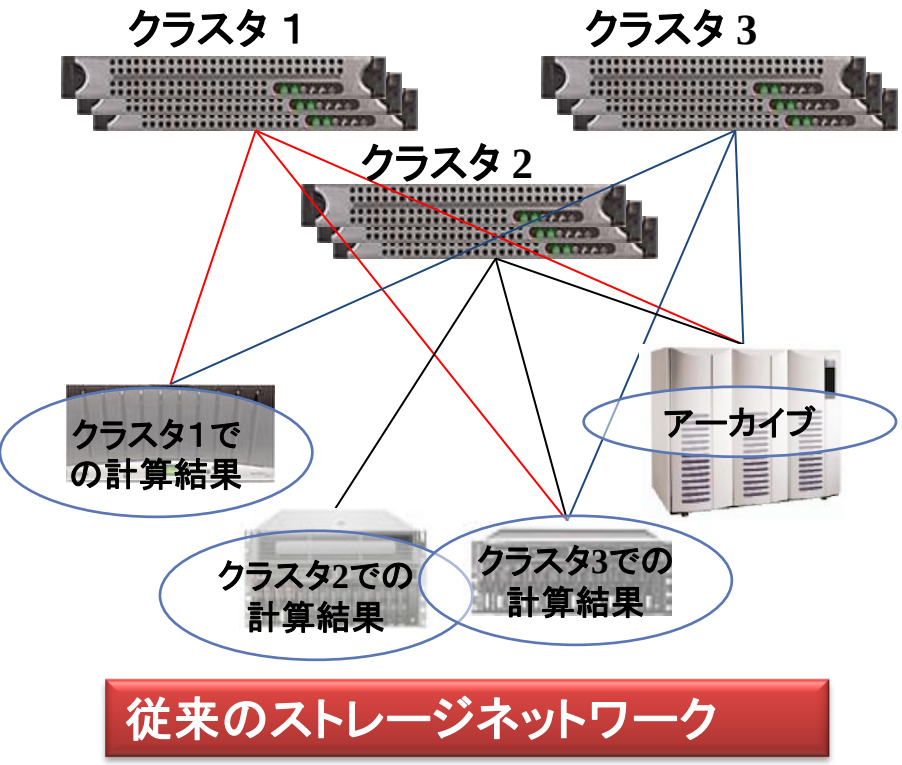
クラスタ
⇒パラレルコンピューティング
⇒ パラレルIOが必要

データセンターでのボトルネック



NFSによる共有ストレージの実現

- クラスタやSMPシステムによる計算システムの構築に際して、NFSによる共有ストレージ構築の構築



問題点と限界

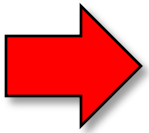
- シングルファイルシステムの限界
- 複数ボリュームとマウントポイント
- 負荷分散（容量&アクセス負荷）
- アップグレード

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

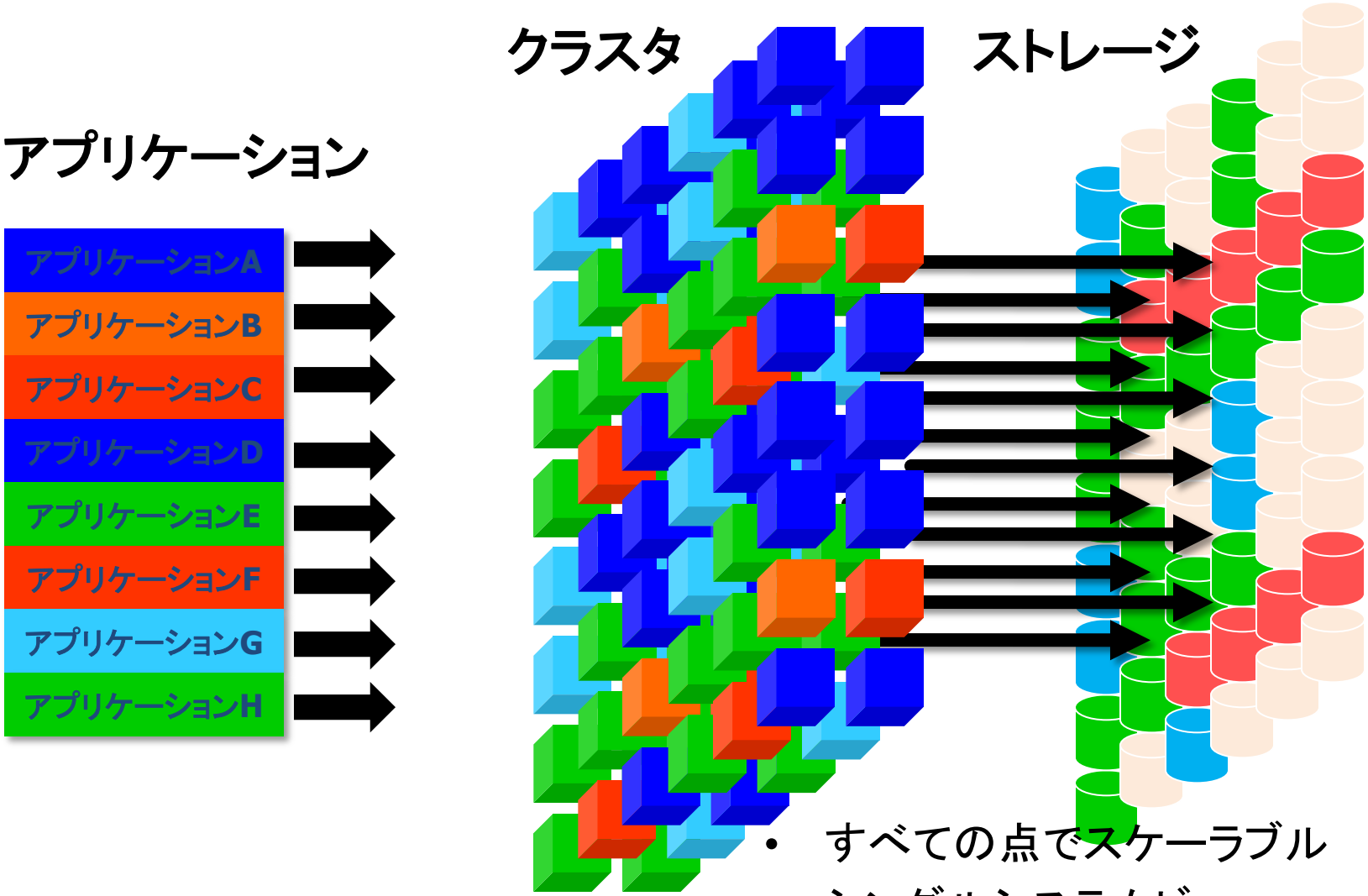
ストレージシステムの課題

- 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



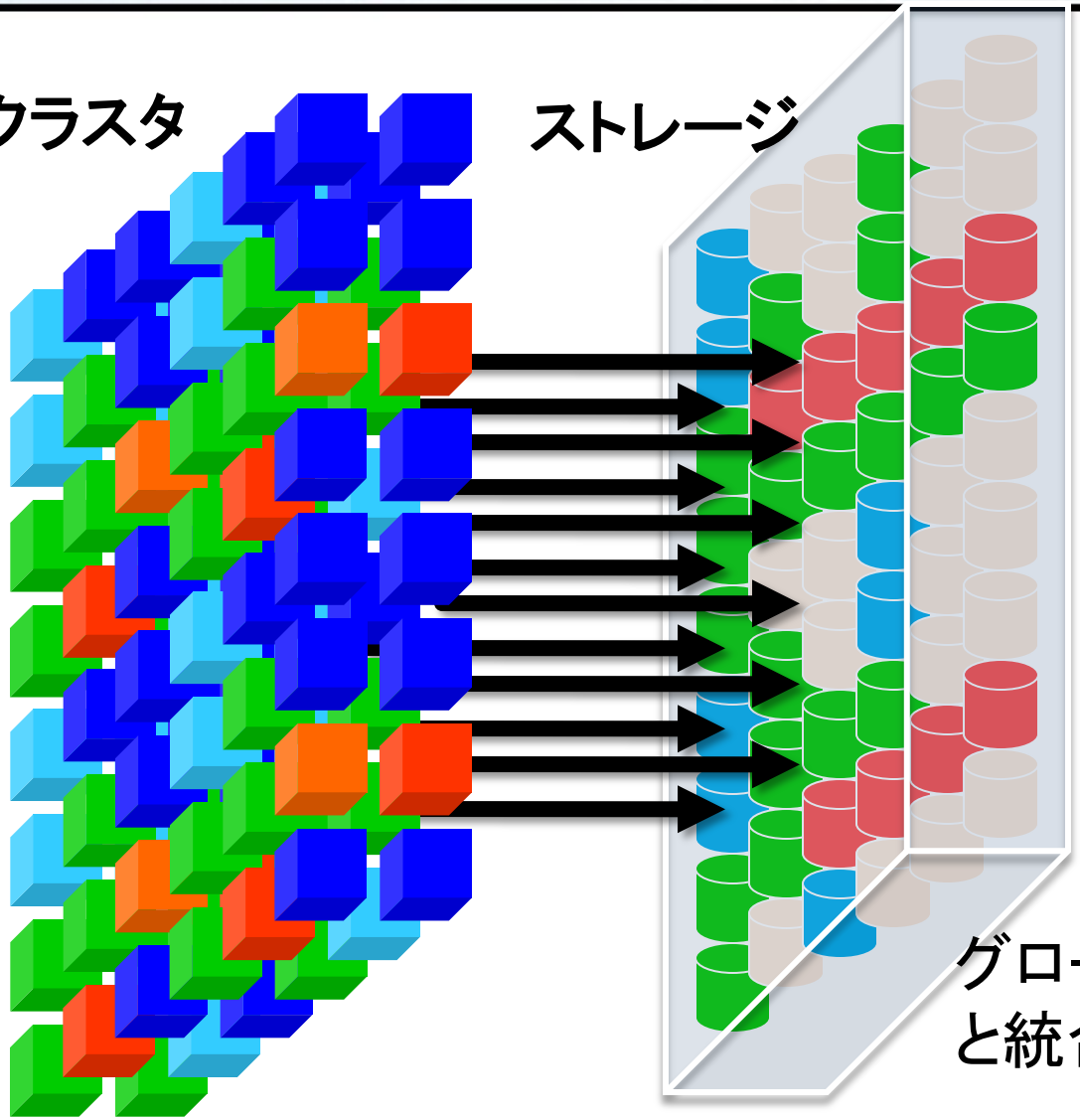
グローバルネームスペースによるソリューション

データセンター



- すべての点でスケーラブル
- シングルシステムビュー

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

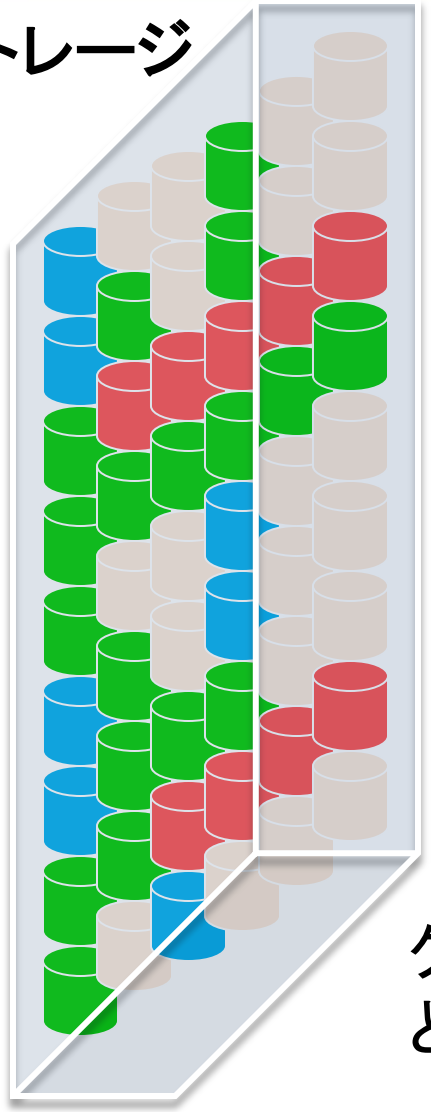


ストレージプール
- 単一の仮想リソース
- 透過的なデータアクセス
- システムの再構築が容易
- 様々なデータの格納が可能
- 可用性

グローバルネームスペースと統合された運用管理環境

グローバルネームスペース

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

グローバルネームスペースの利点

ハイパフォーマンス

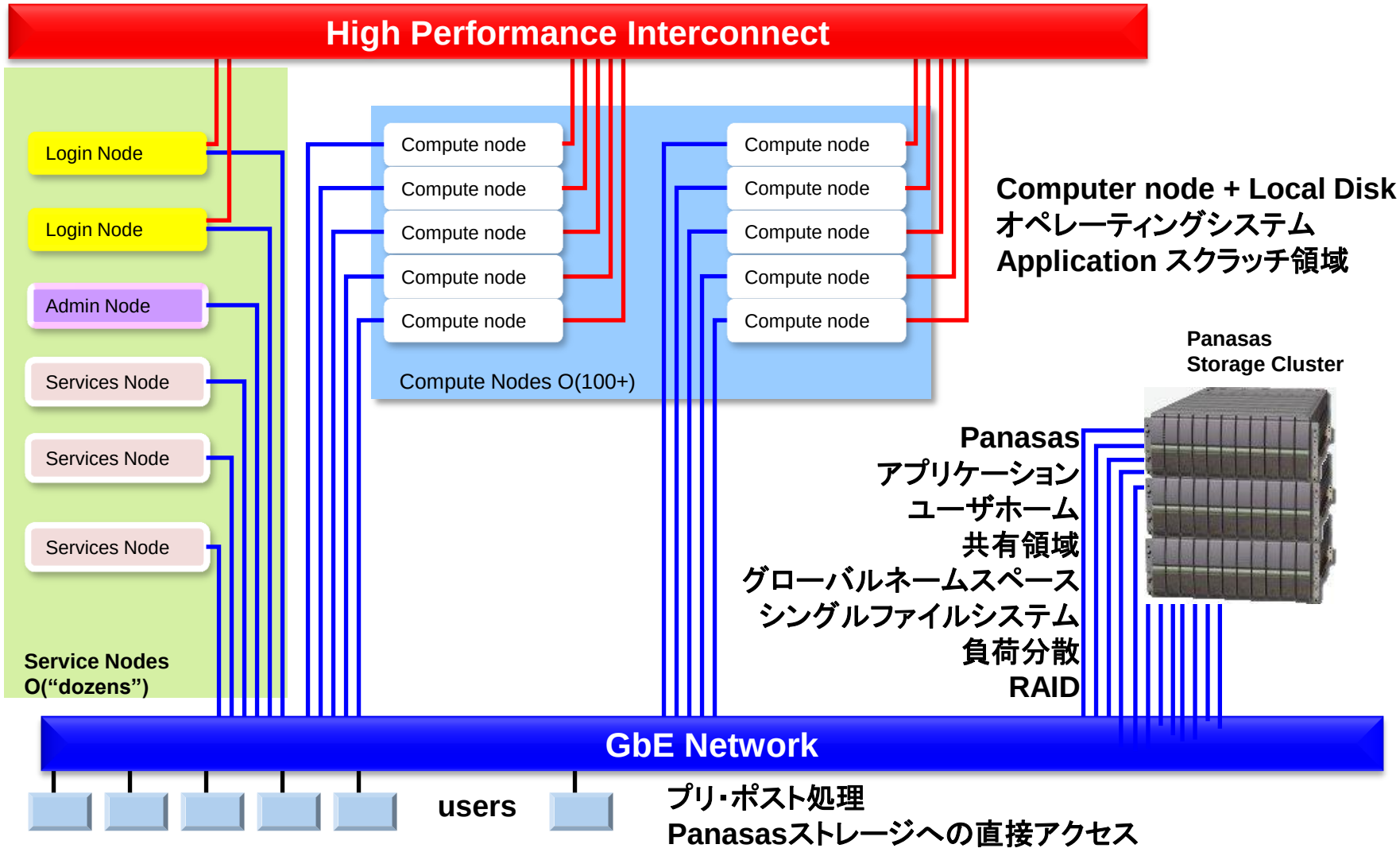
容易な運用管理

スケーラビリティと拡張性

可用性

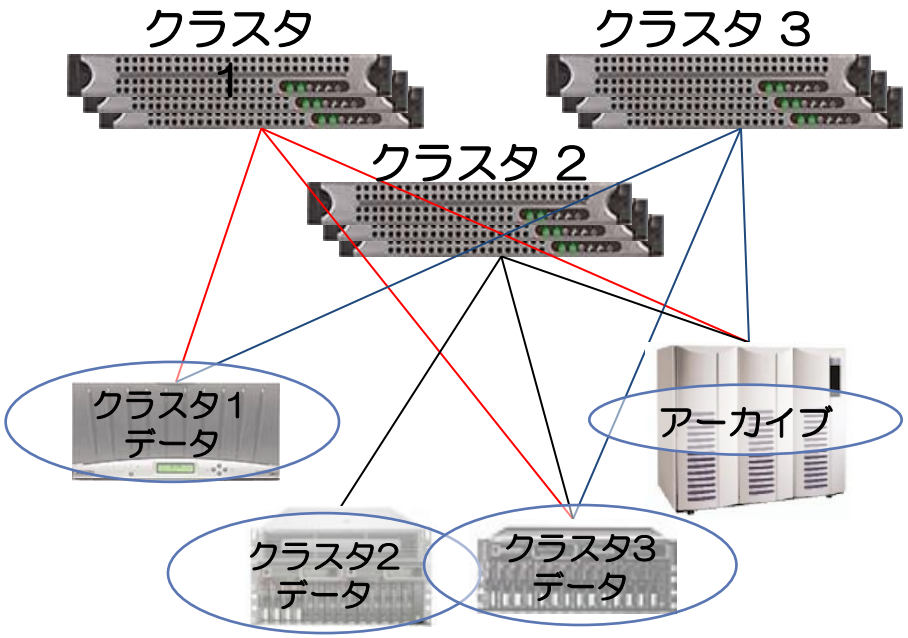
Panasasグローバルネームスペース

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



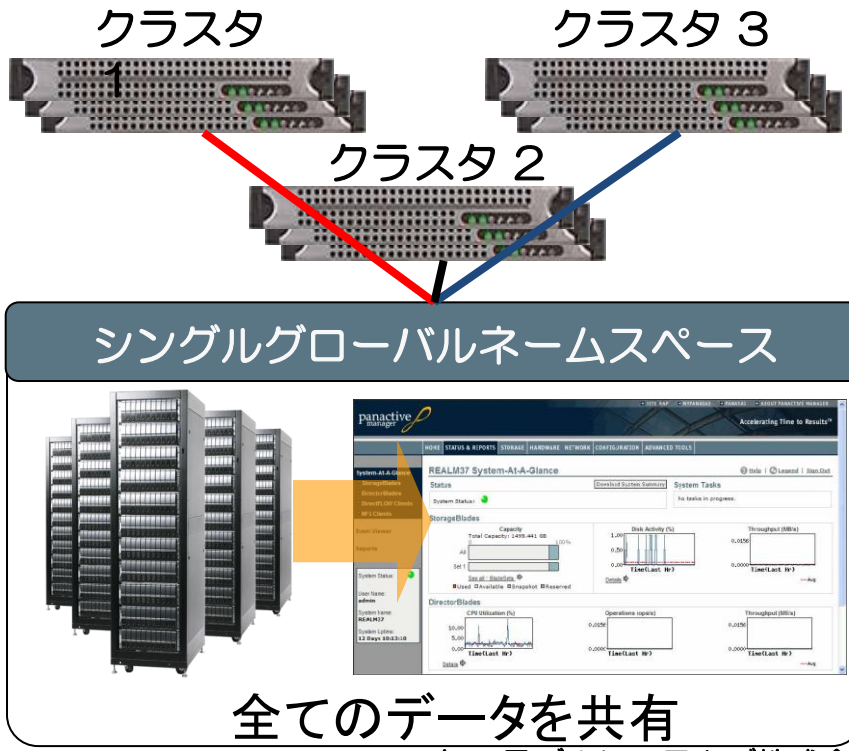
シングルグローバルネームスペース

- 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



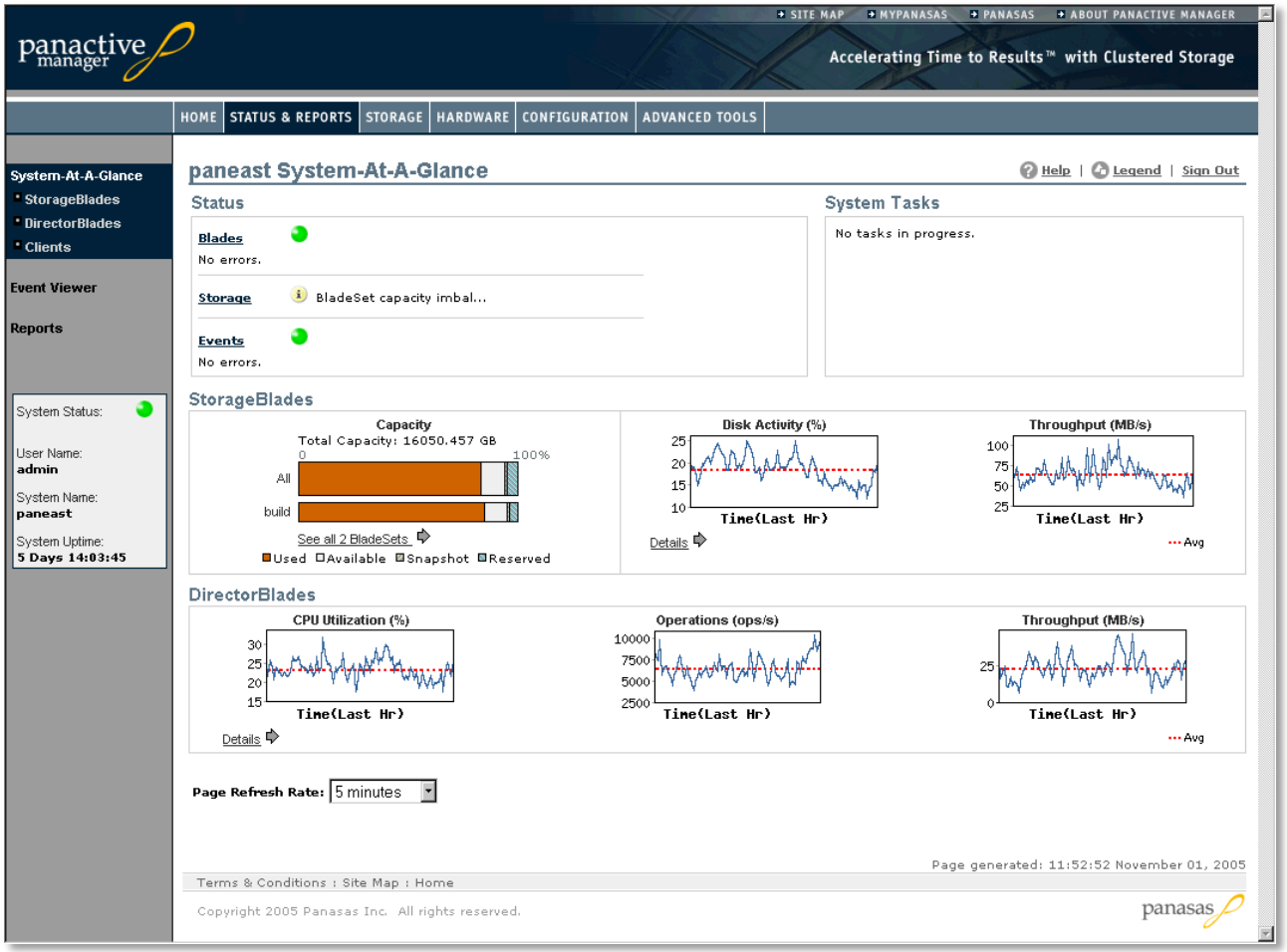
全てのデータを共有

容易な導入と運用管理

- 容易な導入
  - 10分以内でのセットアップ(ESG Lab Test)
  - シェルフ増設時の自動的なシステム構成
- 容易に利用可能
  - 全てのクライアントから一つのネームスペースで利用可能
  - 自動的なファイルシステムでのロードバランスの実現
- 容易な管理・運用
  - シングル管理画面
  - スナップショット、ユーザクォータなどのデータ、ユーザ管理

PanActive Webインターフェイス

- 管理を一元化
- 直ぐ利用可能なインターフェイス
- 増設やシステム構成変更に容易に対応可能
- 豊富なレポート機能
- リアルタイムモニター
- CLI(コマンドラインインインターフェイス）でも利用可能



グローバルネームスペースの利点

- パフォーマンス
  - ストレージに対するパラレルなデータパス
- 容易な運用管理
  - WEBとコマンドインターフェイスからの運用管理
  - 管理運用を一つのネームスペースに集約し、包括的に実行
  - ネームスペースとファイルシステムの双方を同時に管理
- スケーラビリティ
  - 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
  - ネームスペースの変更なしでのシステム拡張
- 可用性
  - システム全体のデータ分析とそのレポートが容易
  - データの保護、バックアップ

グローバルネームスペースの利点

- 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、また、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ルが可能

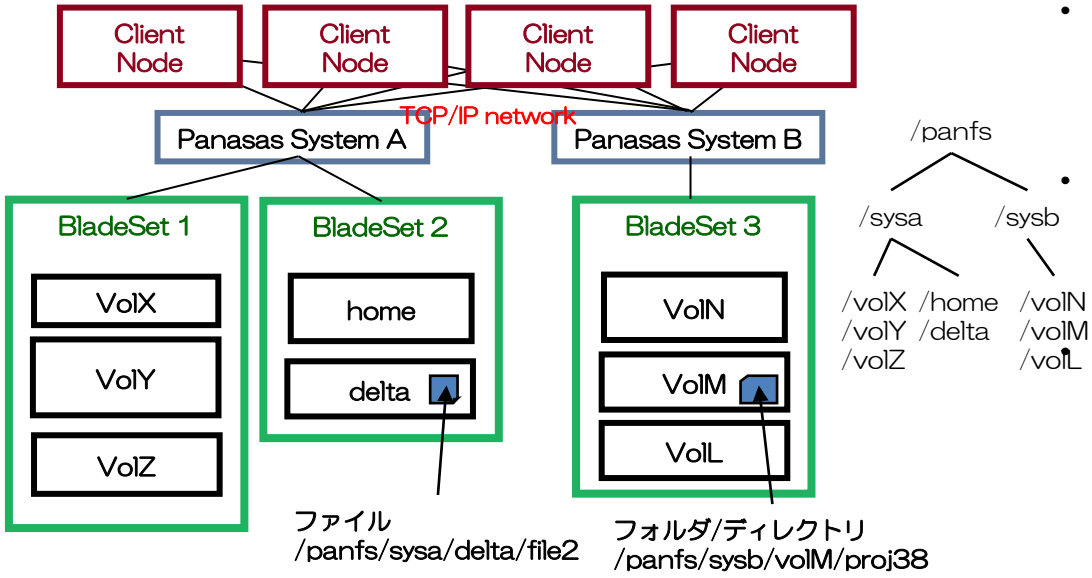
グローバルネームスペース

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

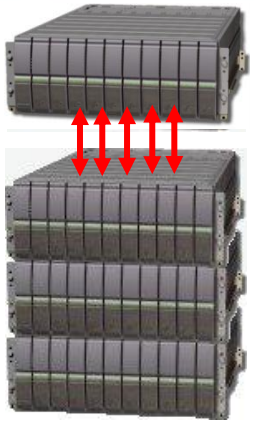
データ管理や移動はユーザに対して、透過的に行われ、データの保管場所などを気にすることなくデータへのアクセスが可能

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

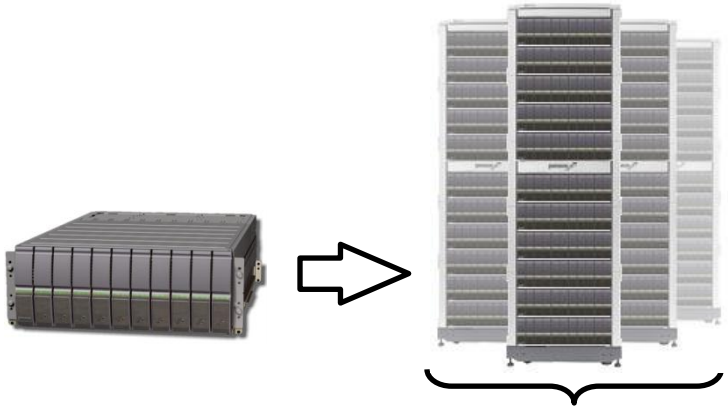
自動プロビジョニングによる容易な拡張

- 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



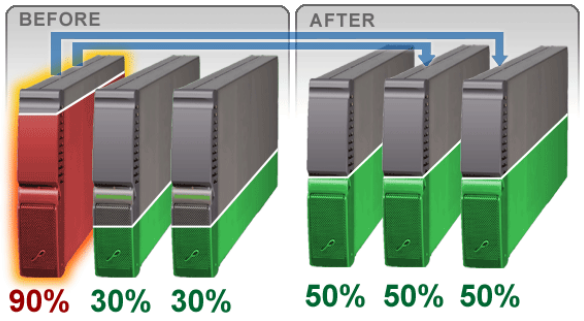
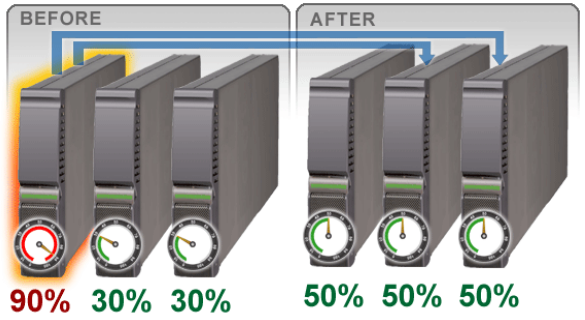
プライベートポート上でのDHCP構成

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	StorageBlade性能	DirectorBlade性能
		
• 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	• 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	• 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ブレードに配置する • 必要な場合には、再配置する

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

スナップショット

- ユーザのデータのリカバリとオンラインバックアップ
- “Copy On Write”によるデータ重複なしでのスナップショット

オンライン中で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

ActiveStor 可用性オプション

クォーラム(Quorum)ベースでのクラスタマネージメント

- 3台もしくは5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

ファイルシステム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

シームレス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
- 仮想 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

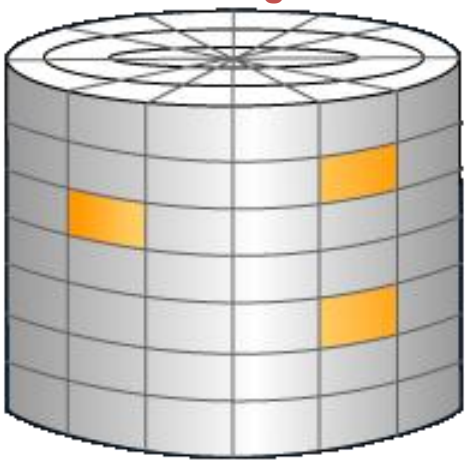
アドレッシング:

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

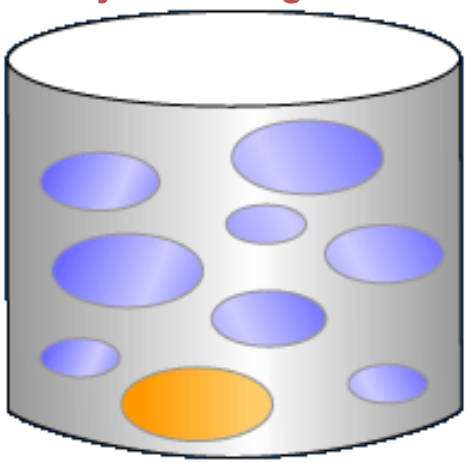
アドレッシング:

[object, byte range]

割り当て:

Internal

Object Storage Device

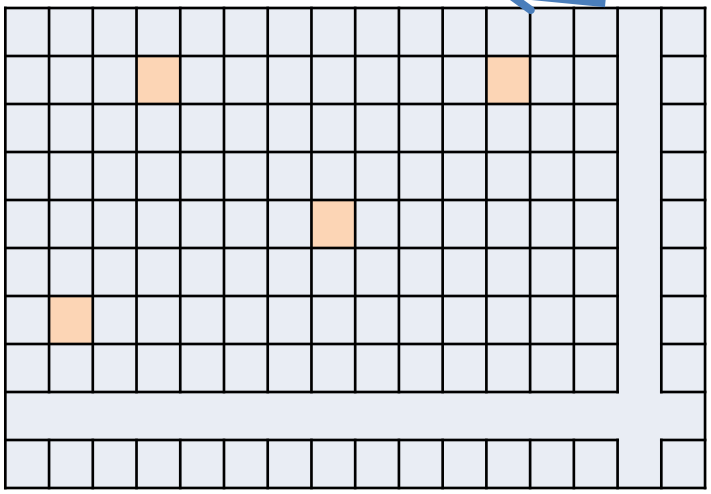


ブロック・ベースと
オブジェクトベースの違い

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

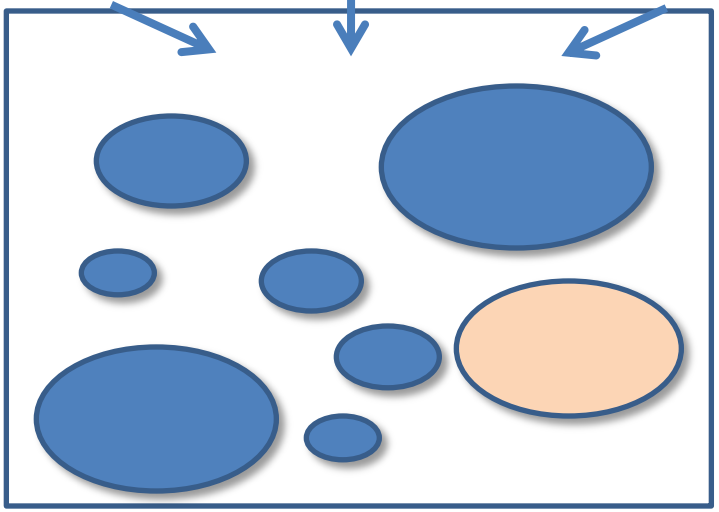


ブロックベースストレージシステム

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

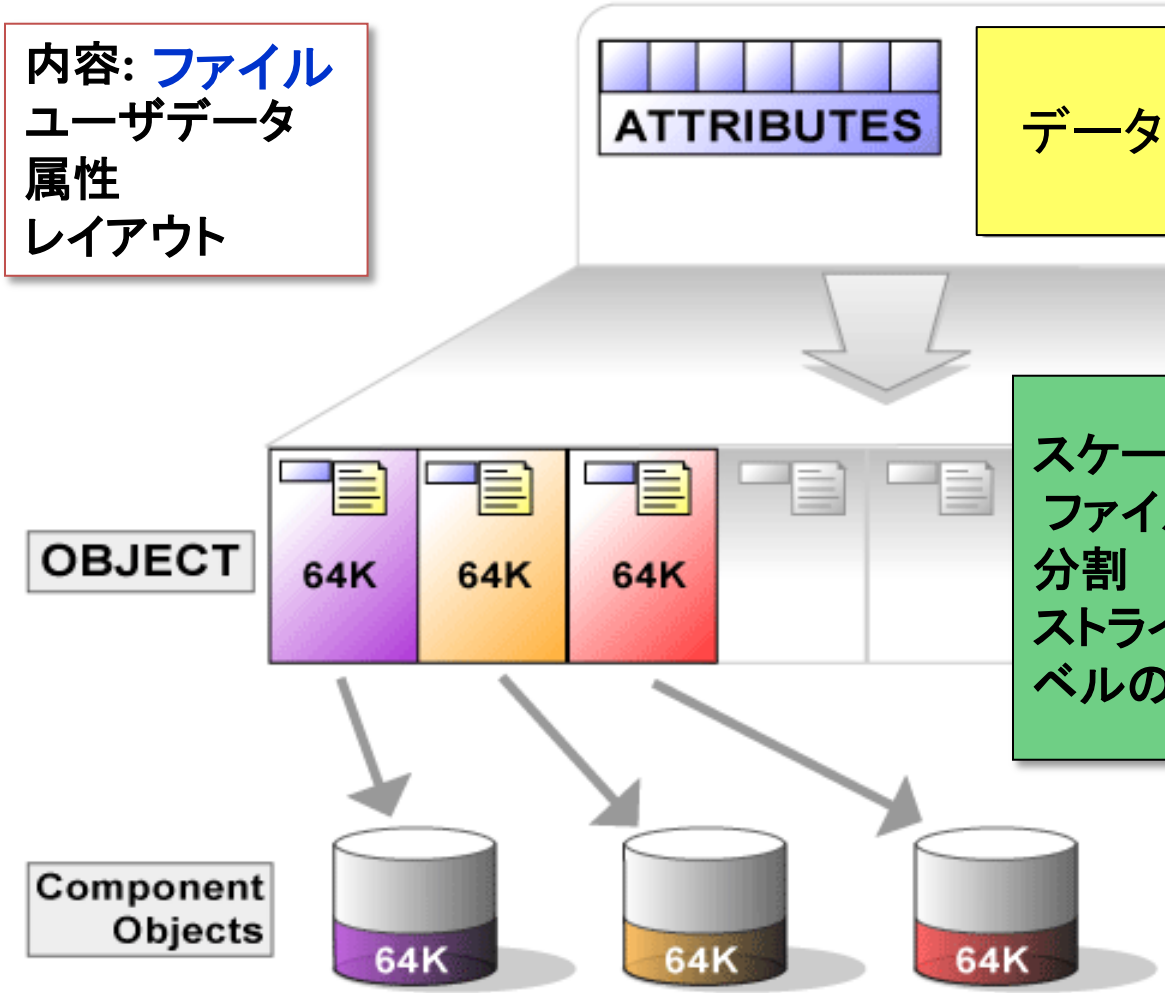


オブジェクトベースストレージシステム

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

Panasas 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

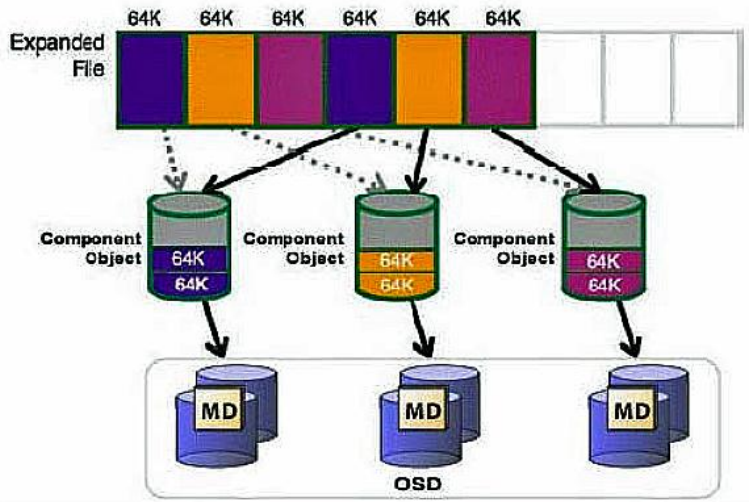
- ホットスタンバイされたスペアを利用してのファイルシステムの再構成が必要
- 残ったディスクからデータを読み込み、（ホット/コールド）スタンバイのスペアにデータを書く込む必要がある
- したがって、システム内の全ドライブを利用しての再構築となるため、システムに大きな負荷をかけることになる
- 再構成に要する時間は、交換したディスクへのデータの書き込みの要する時間によって決まる

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
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



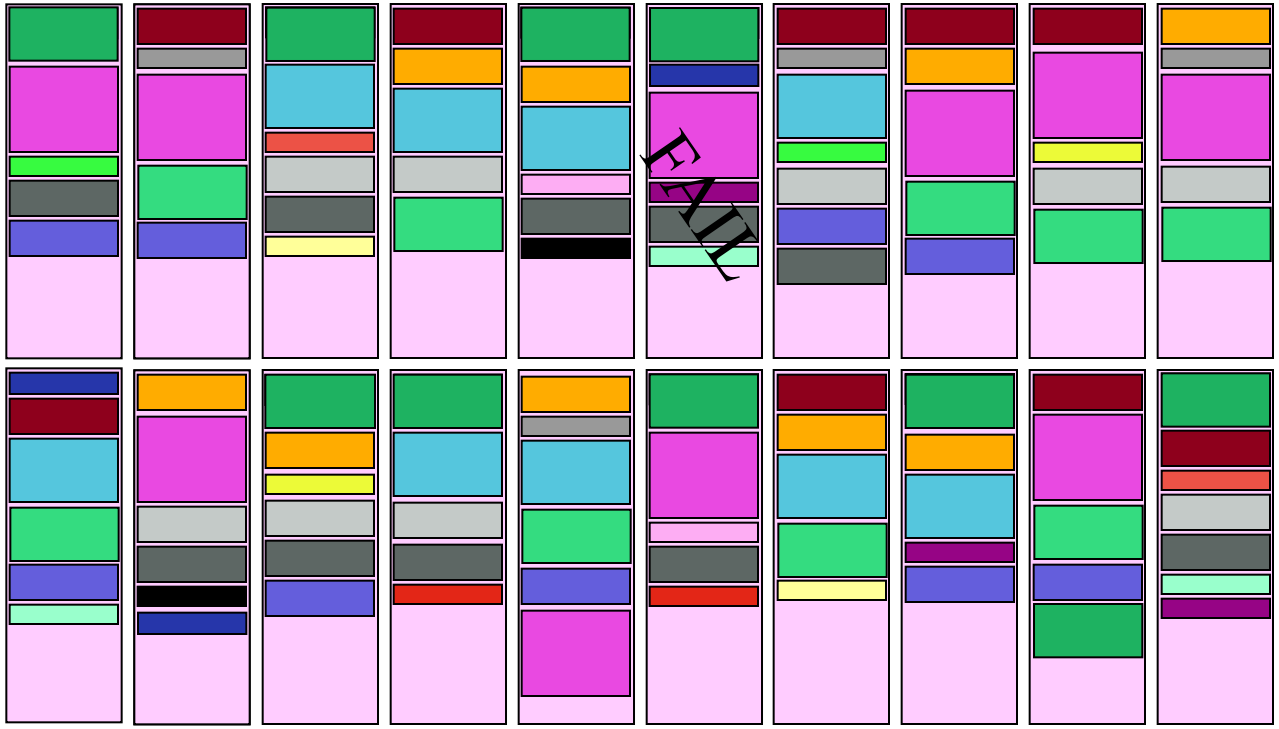
- ストライプユニット
  - 一つのOSDにアサイン（64Kがデフォルト）
- RAIDレベル（0/1/5)
- データ分割幅
  - ストライピングされるOSDの数
  - ファイルの最大の転送速度（バンド幅）が得られるように設定
- パリティストライプ幅（RAID 5設定）
  - パリティの値は、クライアントがデータの書き出し時に計算

Panasas RAID
ファイルシステム再構成

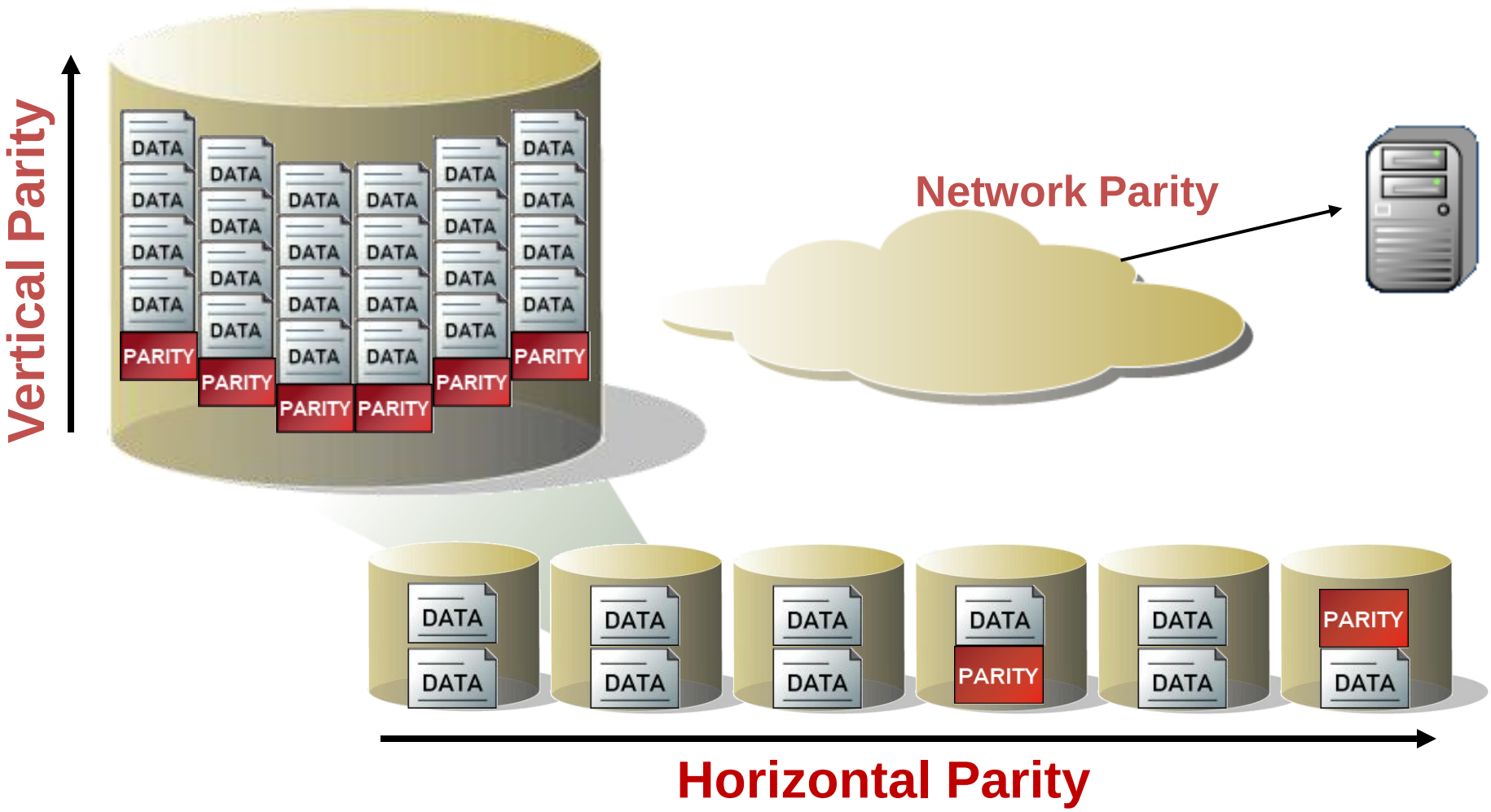
- ファイルは、別々のStorageBlade上にコンポーネントオブジェクトとして分散して配置
- ファイル属性の情報は2つのコンポーネントオブジェクトで2重に保持
- RAID処理は、ランダムに分散して処理

2-shelf BladeSet

ディスクミラー又は9-OSDパリティストライプ

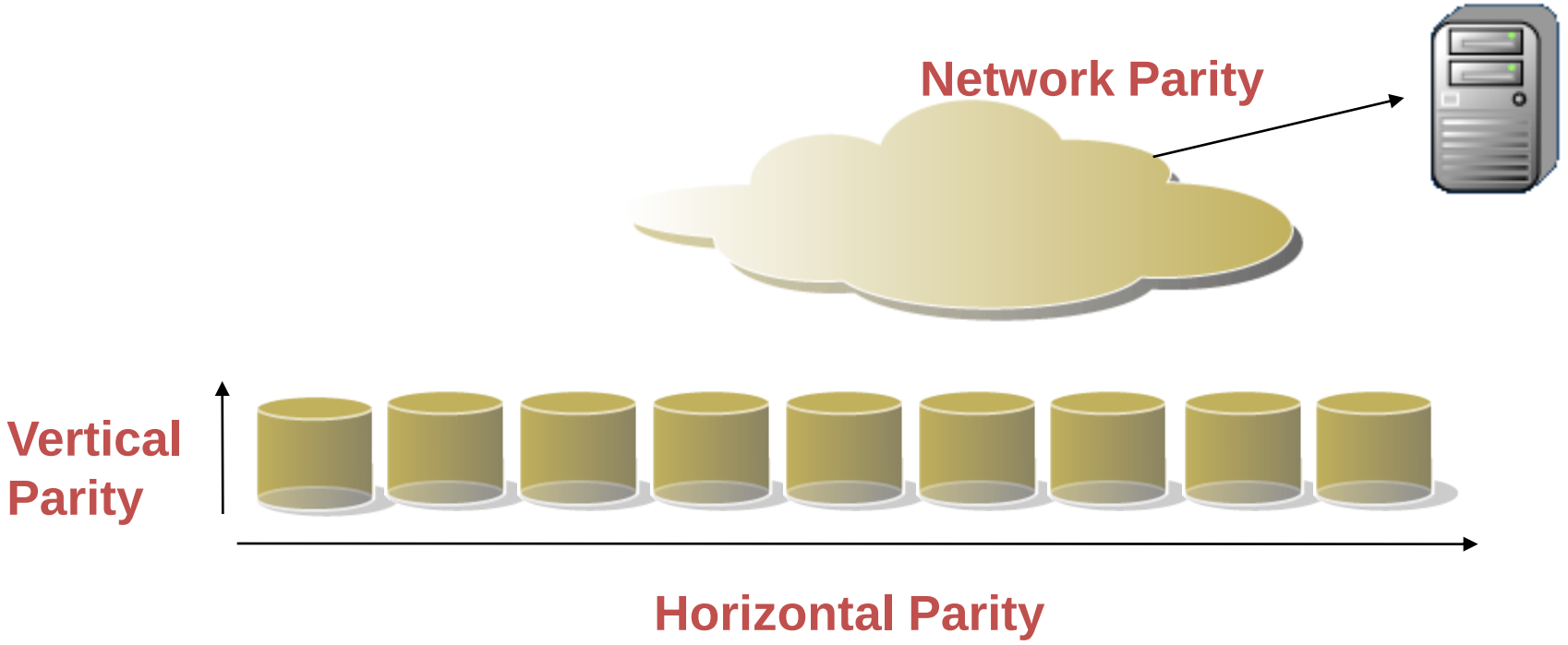


Panasas Tiered Parity

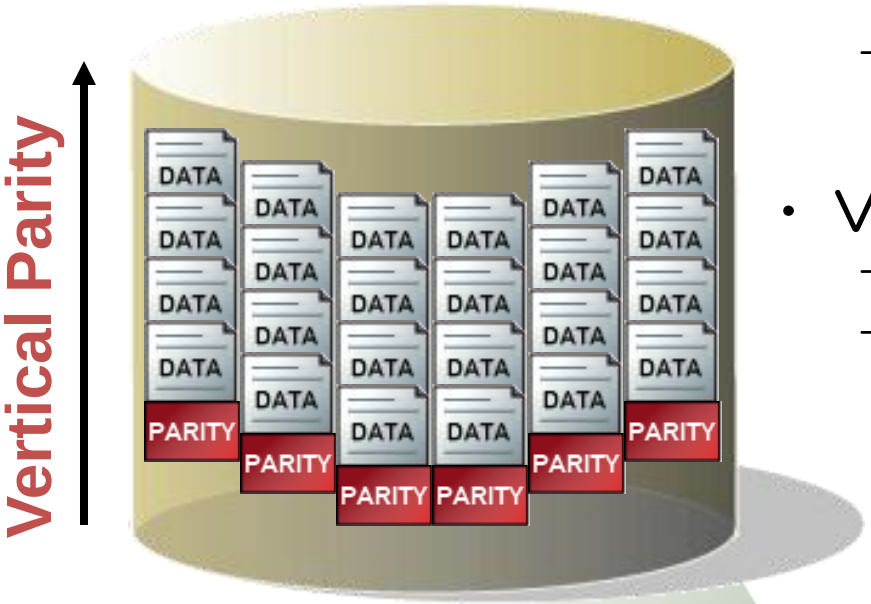


Panasas Tiered Parity

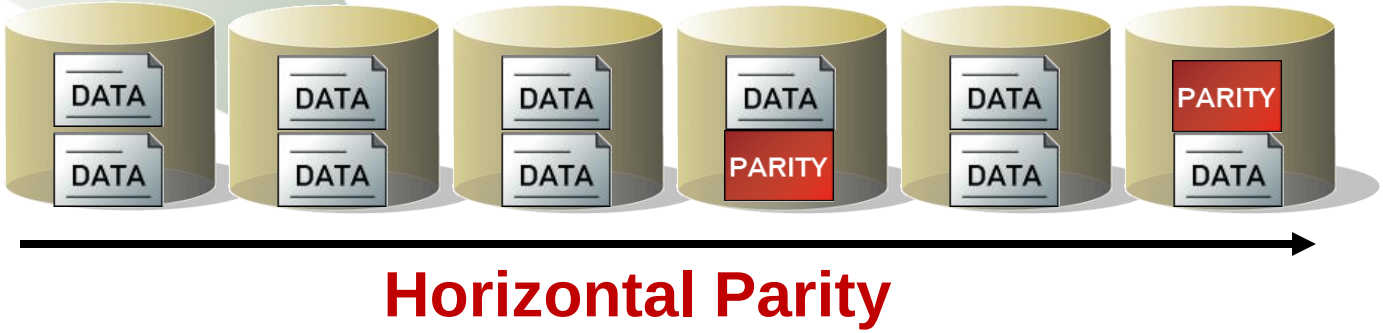
- 各Tierオペレーションは、独立したパリティの処理を行うことが可能であり、エラー検知とデータ修正を行う
- PanasasのTiered Parityが提供する3つのパリティ処理は、互いに相互補完



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



ディスクドライブの高密度化に対する対応・対策

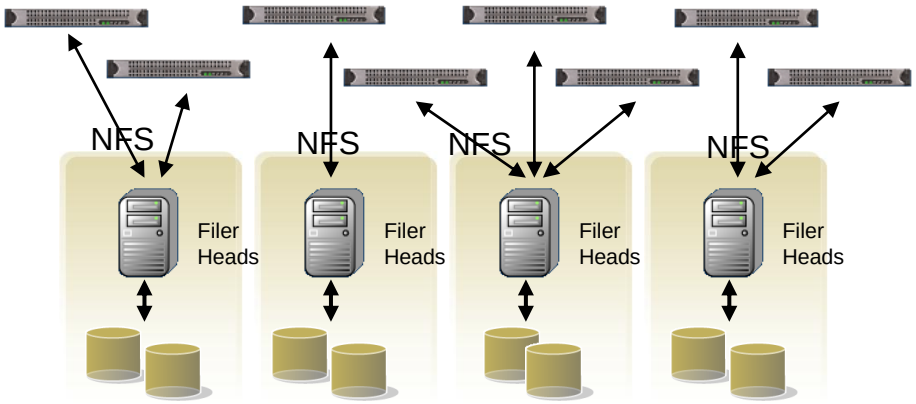
問題点と課題	他社の提案	Panasasの提案
メディアエラー発生頻度の上昇 – RAIDの障害とRAID再構成時の再構成の失敗の可能性	パリティの数を増やす	Vertical Parity はディスクドライブの信頼性の向上を図ります。これはメディアエラーの発生に際して、そのデータエラーの排除を修復を可能とします。RAID Array として利用されるディスク単体の信頼性とエラー回復を図ることを可能とします。
RAID再構成に要する時間の増大とRAID再構成に失敗した場合のデータ復元に要する作業負荷	RAID arrayのサイズを小さくし、同時にパリティの数を増やす	Horizontal Parity 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の比較

	RAID 5	RAID 6	Panasas Tiered Parity
Single Disk Failure	**Yes**	**Yes**	**Yes**
Single Disk Failure + media error	No	**Yes**	**Yes**
Double Disk Failure	No	No*	No
Silent Data Corruption	No	No	**Yes**

＊RAID再構成時にメディアエラーが発生した場合

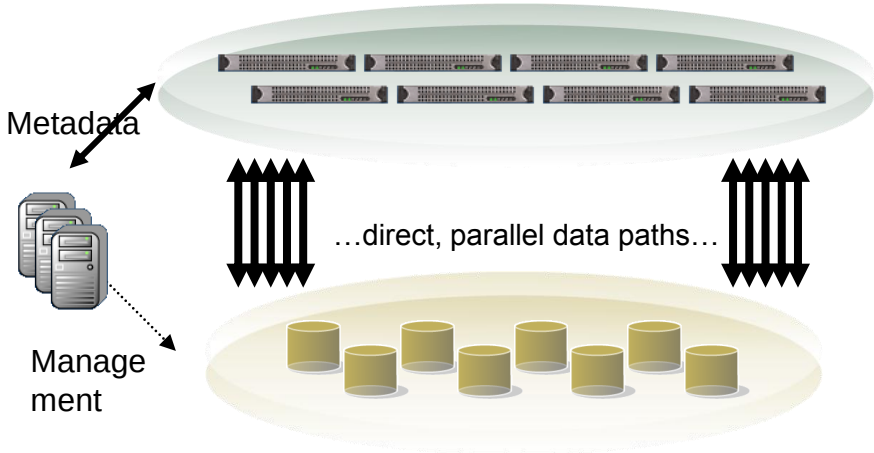
パラレルストレージ



“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

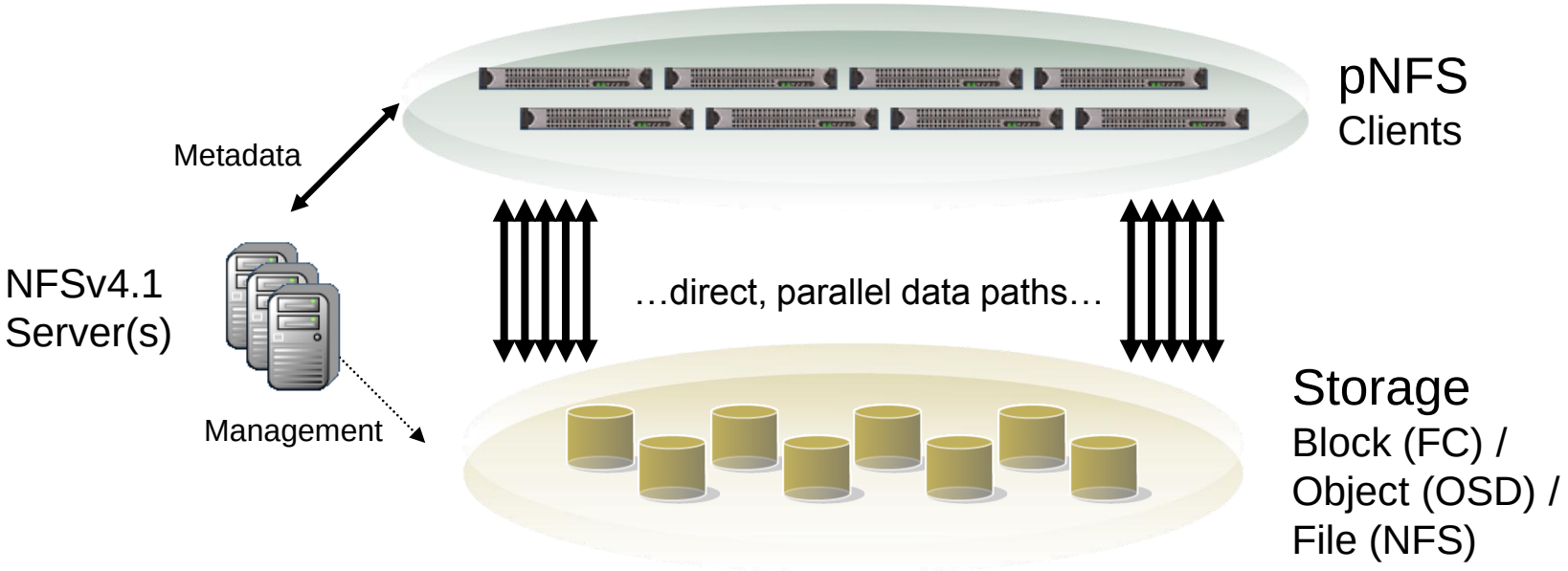


“パラレルクラスタストレージのプール”

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

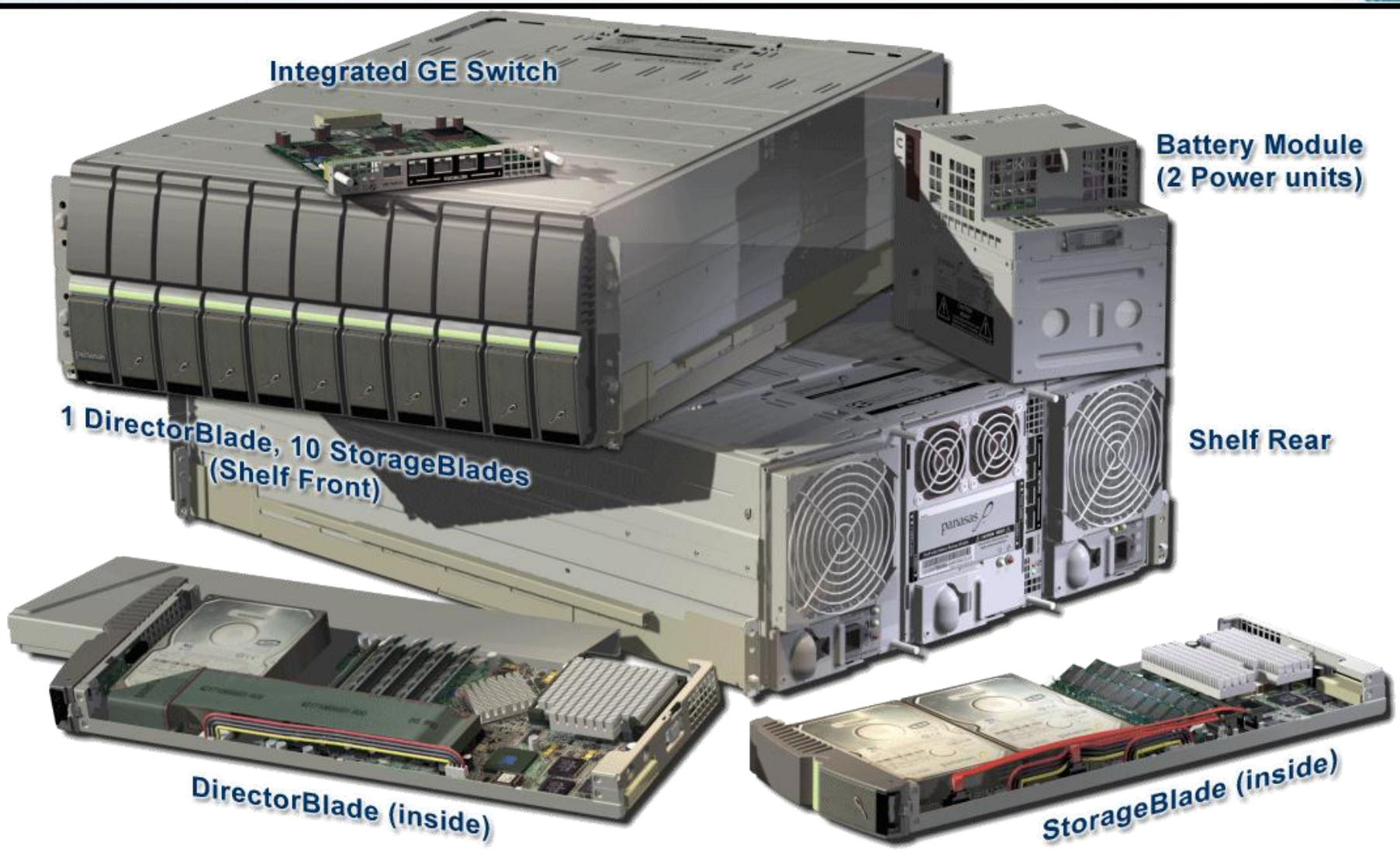
pNFS: 標準パラレルNAS

- 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



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



スケーラブルシステムズ株式会社

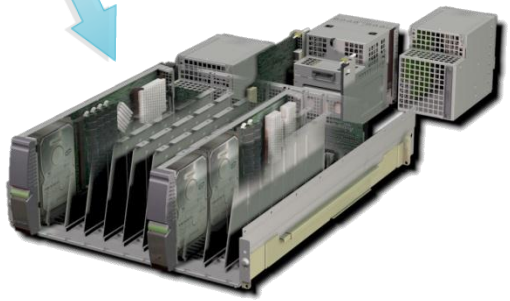
アプライアン的にデザイン

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

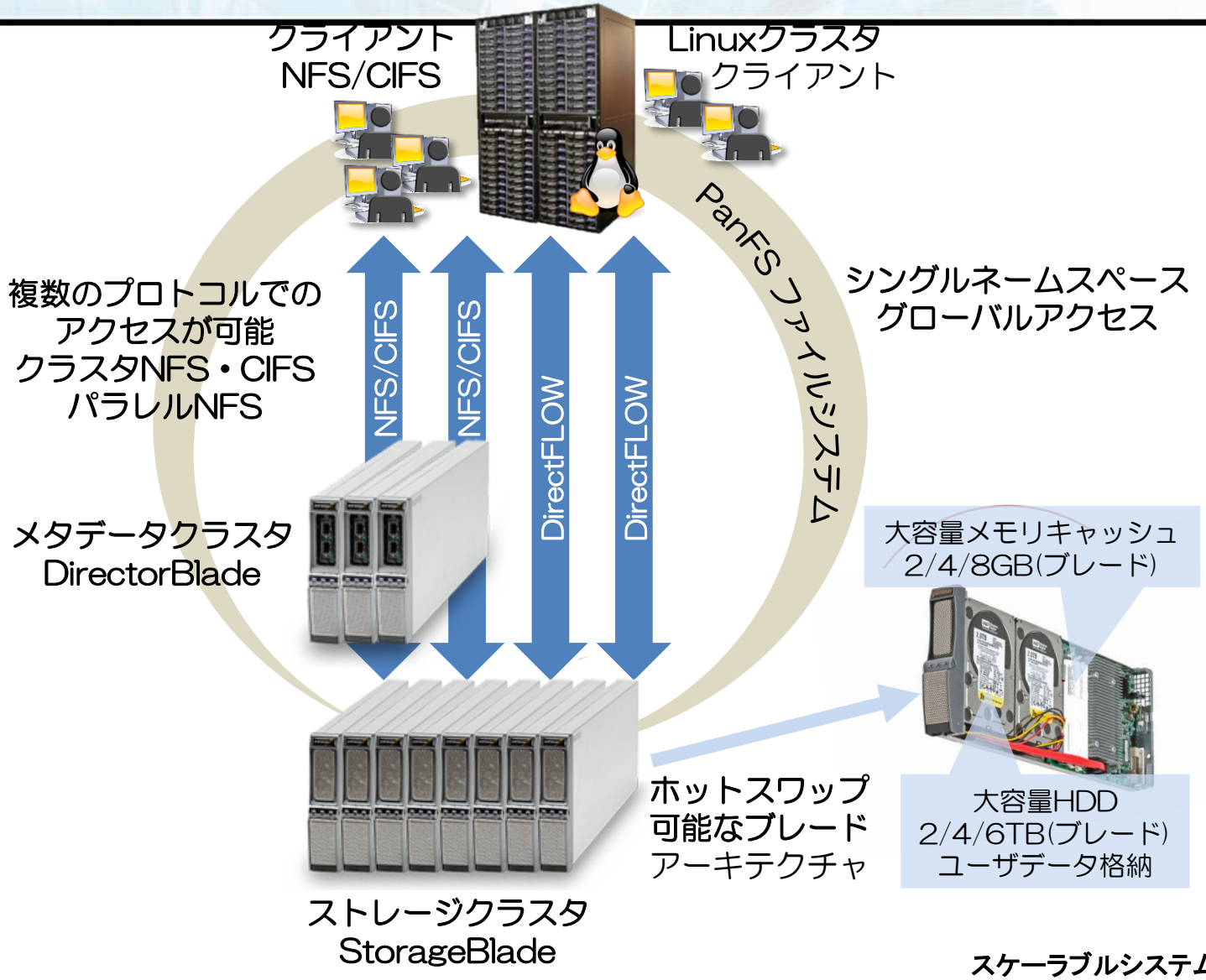


- Shelf あたり10から40TB搭載可能
- ラックあたり100 から400 TB搭載可能

- 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 ActiveStor



スケーラブルシステムズ株式会社

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（10-40 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

スケーラブルシステムズ株式会社

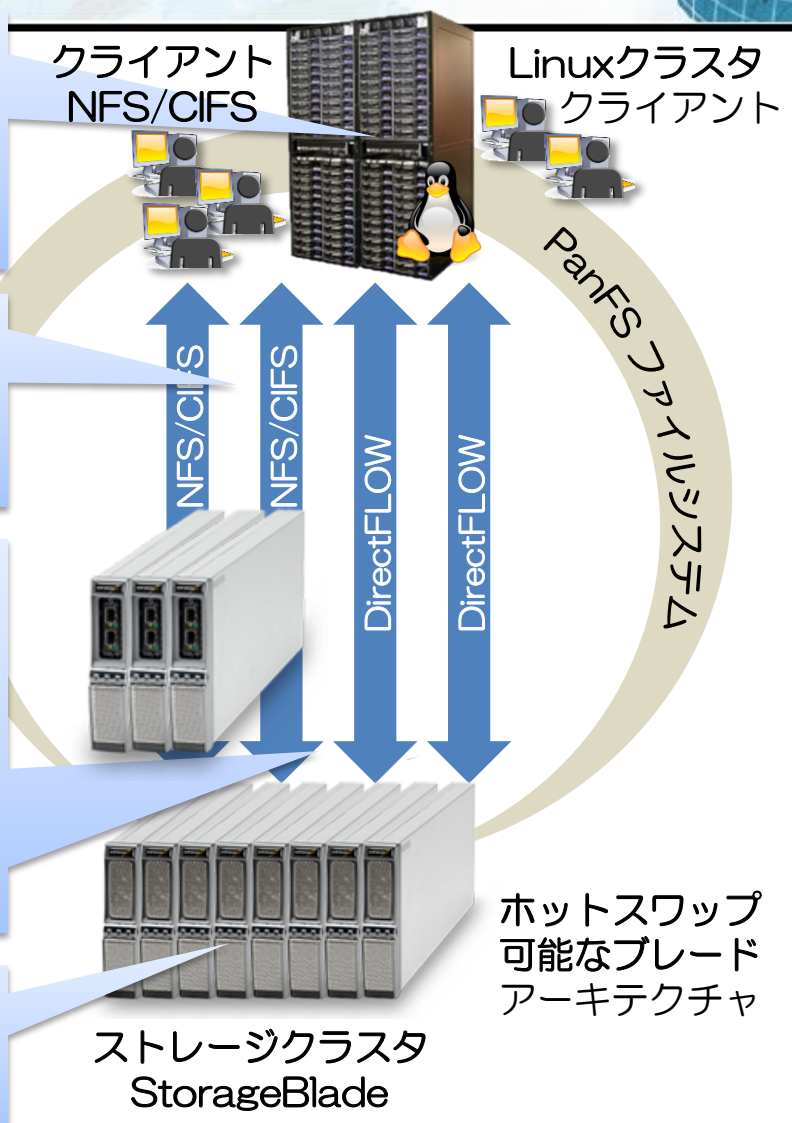
Panasasストレージクラスタ

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

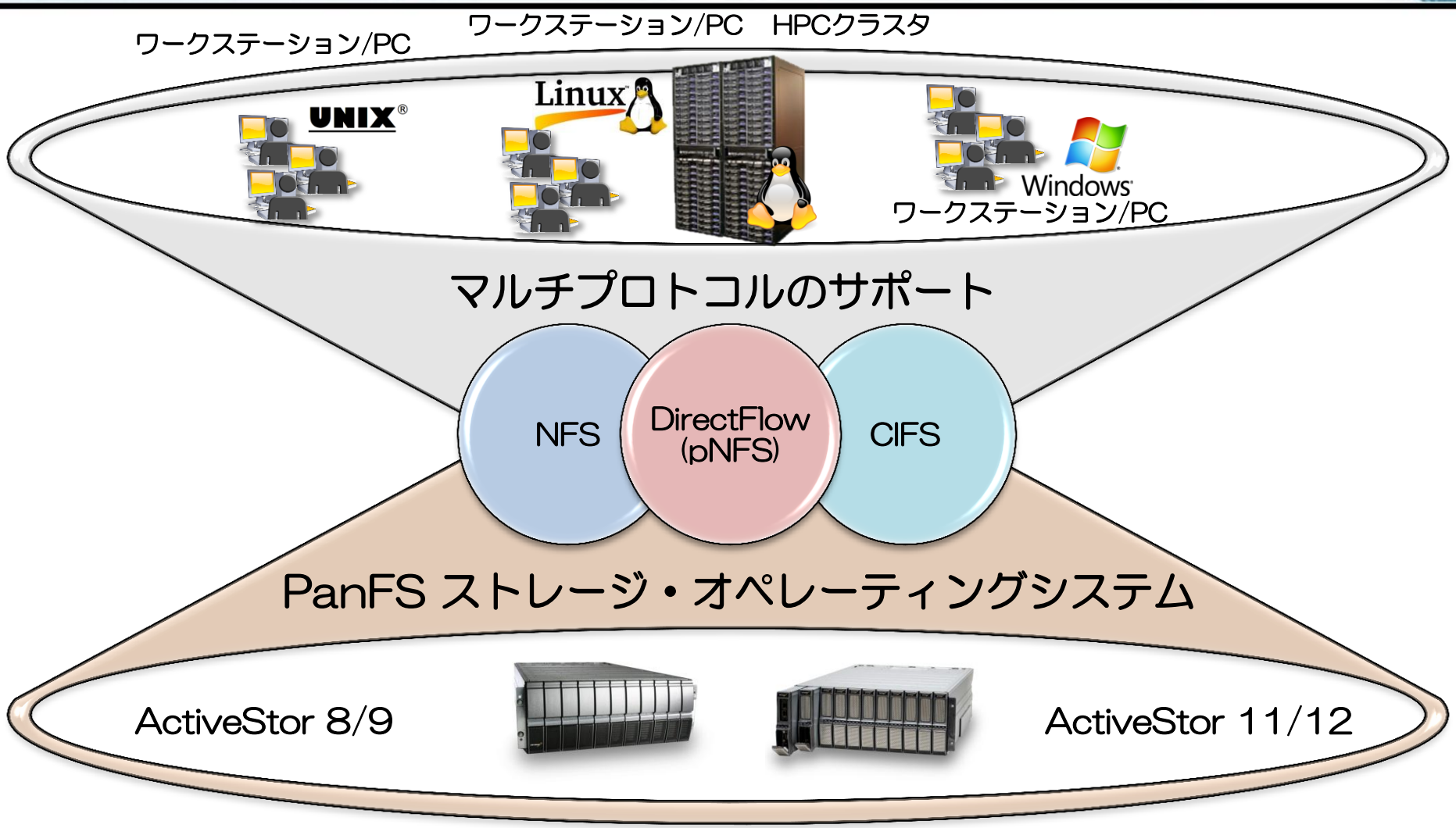
スケーラブルな NFS/CIFS/NDMPサーバ
- 負荷を自動的にストレージクラスタ全体に分散
- クライアント数の増加に合わせてスケーラブルな性能拡張
- 全てのDirectorBladeが全てのファイルにアクセス可能

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

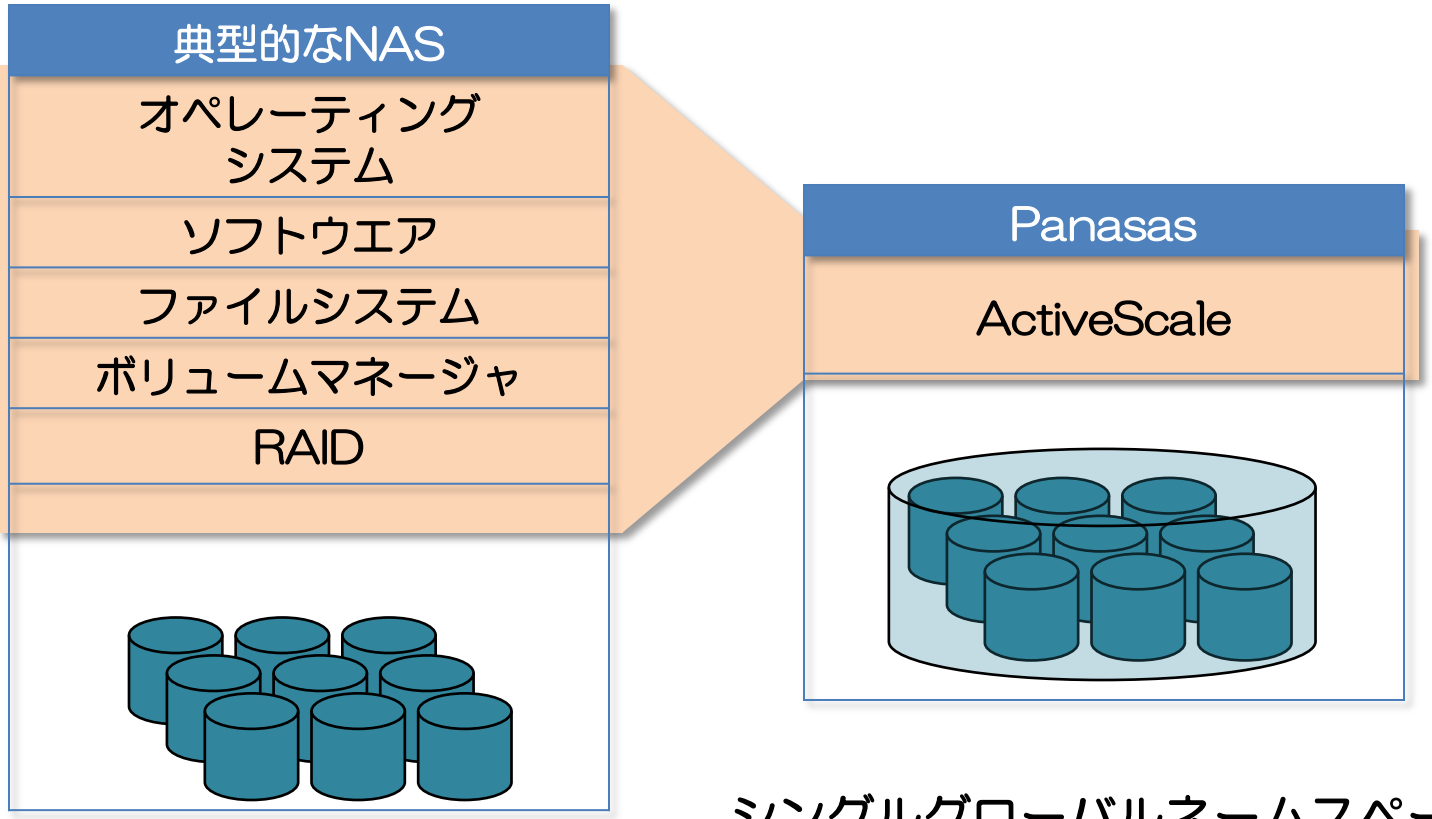
オブジェクトベース
- 優れたスケーラビリティ、信頼性、運用管理
- Panasas Tiered Parityによるデータ保護の強化



Panasasストレージクラスタ



ActiveScaleオペレーティング環境

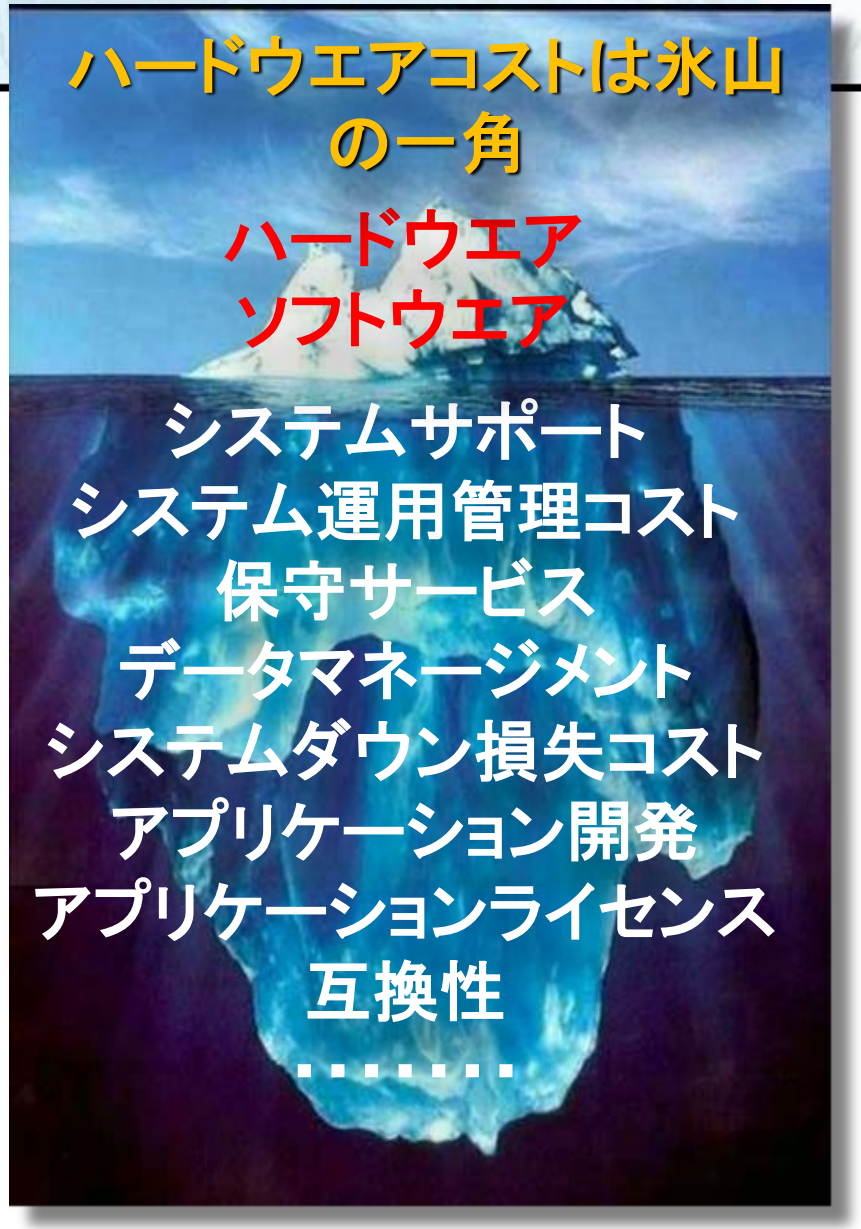


シングルグローバルネームスペース
PanFS/Object RAID/Tiered Parity
NFS/CIFS/DirectFlow プロトコル
ActiveImage/ActiveGuard

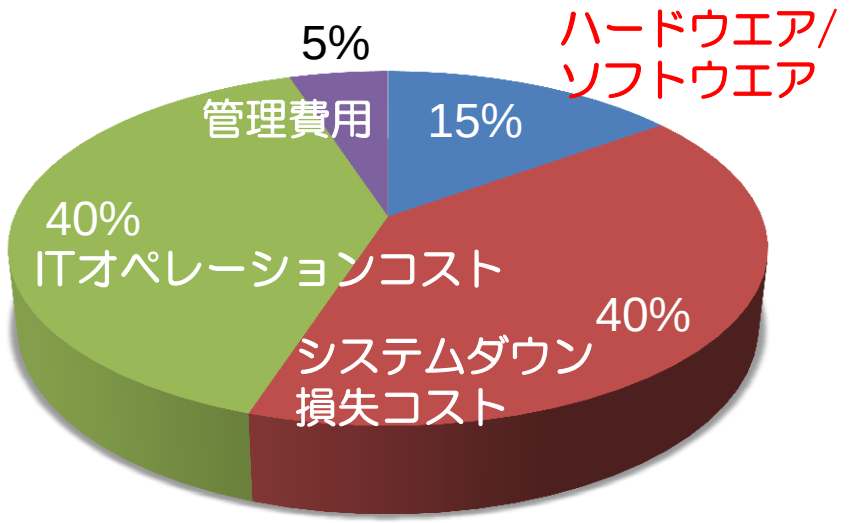
スケーラブルシステムズ株式会社

まとめとして

TCO : Total Cost of Ownership



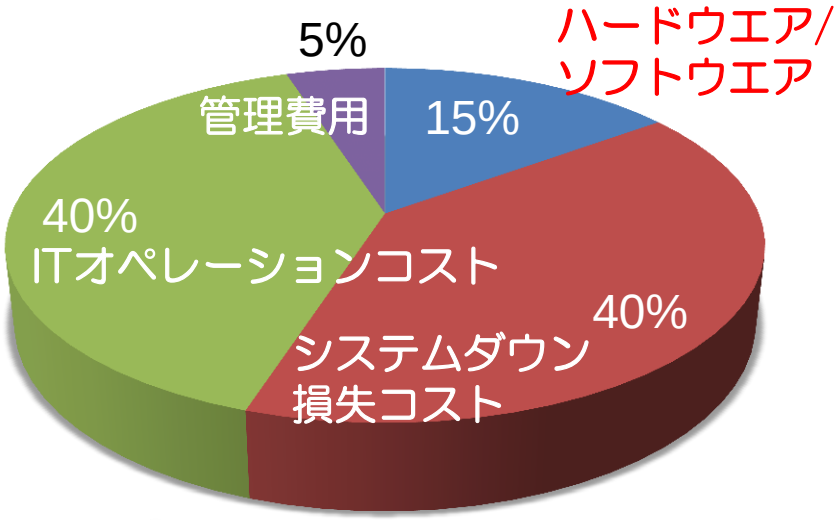
TCO構成分析



Source Gartner Research

TCO削減による
高い対費用効果の実現

TCO構成分析 Source Gartner Research



TCO削減

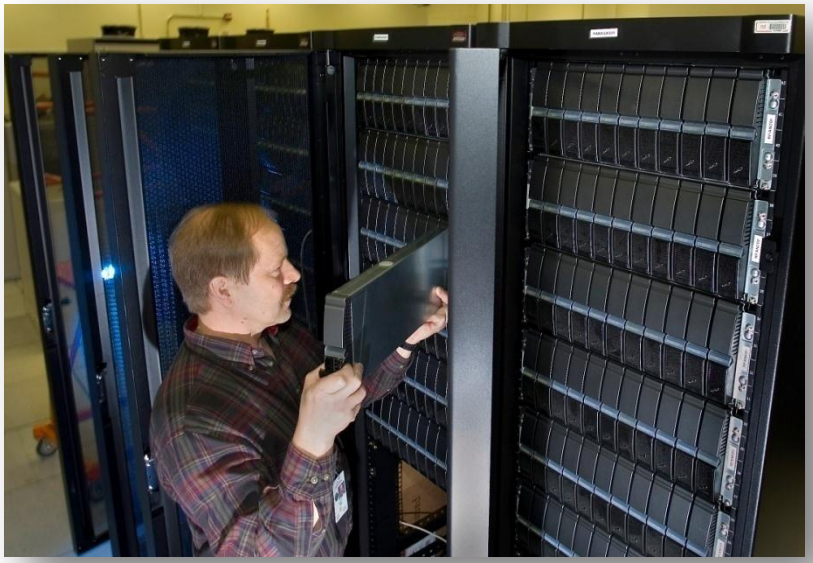
管理・運用の自動化
容易なオペレーション
可用性オプション
ボトルネックの解消
・・・・・・・

Panasas ActiveStor
ストレージクラスタ

高い自己管理機能を持つ
ブレード型ストレージクラスタ

スケーラブルシステムズ株式会社

システム管理と高可用性機能

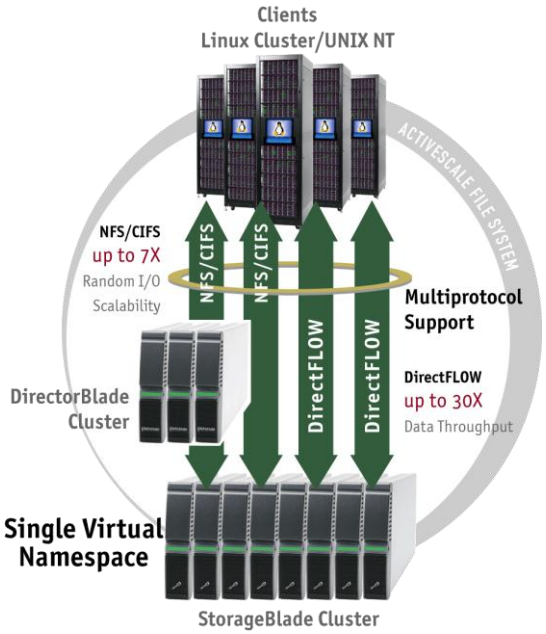


- 予防的システムマネージメント
  - データとディスクのスキャンを継続的にバックグランドで実施
  - 問題発生の可能性のあるブレードのシステムからの切り離し

- 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

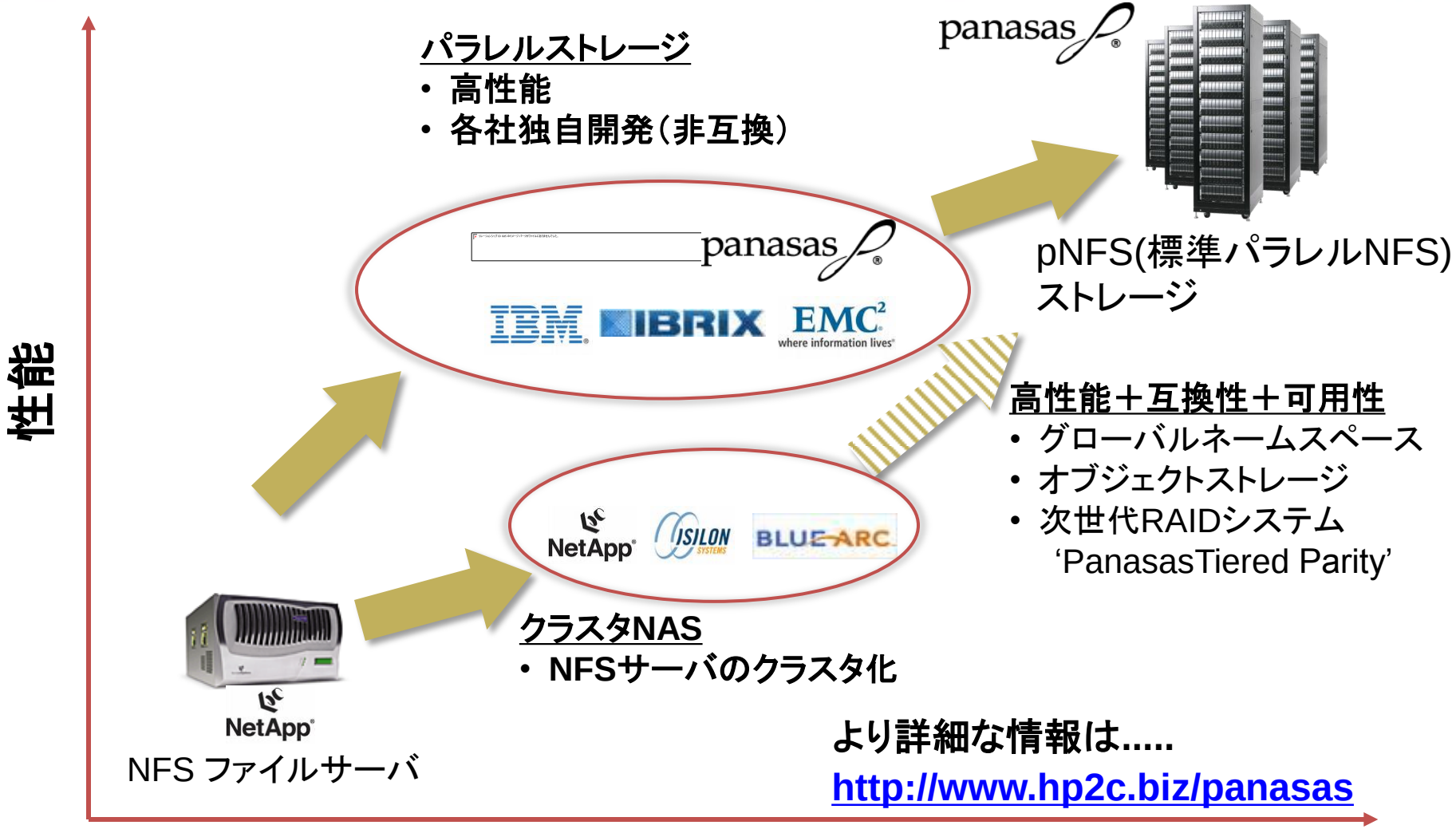
Panasa ActiveStorの特徴

- 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



Panasas社はストレージクラスタのリーディングカンパニー
(NFS, CIFS, DirectFlow)

高性能ストレージシステムの将来



システムへの要求とPanasasの利点

	課題と要求	Panasasの利点
ユーザ	•システムを容易に利用可能（ジョブ実行、データ管理） •高い実行性能とスケーラビリティ •豊富なアプリケーション	•ジョブの実行に際してFTPなどの作業が不要 •ファイルサイズを気にしない •プリ・ポストと解析でのファイル共有 •Windows/Linux/Unix間でのファイル共有
運用管理者	•容易な導入とセットアップ •クラスタシステムでの利用 •スケーラビリティ •増設などが容易	•シングルグローバルネームスペース •マネージメント •ディスクエラー、メディアエラーへの対応（Tired Parity RAID） •動的負荷分散
開発者	•標準化インターフェイス •互換性	•スケーラブルな標準API (MPI-IO) •容易なボトルネックの把握

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	>10GB/s
	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	>100 サーバ
	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	無制限
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	>10x高速
	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	>600 MB/s
	Panasas Tiered Parity	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	ファイルレベル
	プロビジョニング	仮想化されたストレージ	自動
	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	ペタスケール
	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	○	○	

テクノロジー比較

オブジェクトベースのパラレルファイルシステムと従来のストレージソリューションの比較（優位性）

ストレージ	Panasasの優位性
DAS	制御とデータパスの分離。メタデータとデータ処理は分散処理可能 複数のアクセスポイントによる冗長性とスケーラビリティ アプリケーションとストレージアクセスに関するサーバリソース利用に関するバランスを取る必要がない
NAS	大規模システムでのスケーラビリティと運用管理（複数のマウントポイントの管理）
SAN	クライアントは、ストレージに直接アクセス可能で、中間のゲートウェイは不要 IPベースのコミュニケーションが可能で、インフラ構築が廉価。GbE,10GbE,InfiniBandなどの選択が可能

PANASASパラレルストレージを選択する10の理由

アプリケーション性能

・DirectFLOWプロトコルによって、NFSを利用したストレージソリューションに対して、データ処理の速度と効率において優位性があります。スケーラブルなI/O性能はシミュレーションやモデリングアプリケーションをより効率良く、高速に実行することを可能とします。

エンジニアリングやシュミレーションでの実績

・Panasasは非常に厳しいビジネスを勝ち抜くために常に最先端のITテクノロジーを求める多くの企業で採用されています。企業は常に競合他社との製品開発競争に直面しており、よる少ない開発期間でより優れた製品の開発を行う必要があります。そのような会社が、Panasasを採用しています。

TCO低減を実現する容易な運用管理とグローバルネームスペース

・Panasasの運用管理はシステムの規模が大きくなっても複雑さを増すことなく管理運用することが可能です。同時に、PanFSパラレルファイルシステムのグローバルネームスペースは、管理をより簡便に利用者の透過的なデータアクセスを実現しTCOの削減を可能とします。

優れた信頼性

・ハイエンドのネットワークストレージに大規模な信頼性と有効性を提供できる多くの機能を取り入れたシステムです。高可用性ソフトウエアによるフェイルオーバー機能やPanasasのユニークなTiered Parityアーキテクチャによるデータ保護によって、高い信頼性と可用性を実現します。

主要アプリケーションに最適化されたストレージ

・エンジニアリングやシュミレーションで使われる主要なアプリケーションソフトウェアに対して、その高速実行と高い実行効率を実現するために様々な活動を行っています。このような活動によって、高い実行性能を実証し、また、アプリケーションの開発自身もサポートし、より効率の良いアプリケーションの開発を可能としています。

PANASASパラレルストレージを選択する10の理由

パラレルNFS (pNFS)への最も簡単な経路

・pNFS (parallel NFS)は、パラレルI/Oの新しい基準であり、NFS基準の次世代の主要な拡張です。pNFS基準は、Panasas DirectFLOWアーキテクチャに強く影響されているため、pNFSが製品化された場合Panasas DirectFLOWプロトコルからのアップグレードは非常に容易です。

性能と容量においては無制限のスケーラビリティ

・ストレージ容量に関しては、30分以内で迅速かつ容易に追加や再構成が可能です。ストレージ容量が増加すると、テラバイトからペタバイトまでスケーラブルに提供可能となり、システムバンド幅は100GB/s以上まで増大します。

迅速な導入が迅速な結果を生み出す

・Panasasパラレルストレージクラスタは、優れたパフォーマンスと容易な運用を提供することにより、ビジネス上の決断をより迅速に導くことができます。

ストレージ基盤の統一

・Panasasパラレルストレージは、単一または、共有ストレージプールへのNFSやCIFSアクセスのサポートを提供することによりストレージインフラの統一を可能にします。これはストレージ管理を単純化して、データの複製を減少させかつ、エンドユーザの生産性を増大します。

世界クラスのサービスとサポート

・365日24時間のWEBや電話、メールでのサポートを提供するPanSelectサポートサービスプログラムなどが利用可能です。

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