



Panasas クラスタストレージ のご紹介



住商情報システム株式会社
Sumisho Computer Systems Corporation

IT基盤ソリューション事業部
先端技術システム部



はじめに

- マーケットの動向とプラットフォーム
- Panasas社 会社概要

Panasasストレージクラスタ

- アーキテクチャ
- ハードウェア
- ソフトウェア



Panasas社概要



設立	1999年4月にGarth Gibson氏 (RAIDの開発者) によって設立
本社及びオフィス	本社: 米国カリフォルニア州フリーモント 開発センター ピッツバーグ州ピッツバーグ ミネソタ州ミネアポリス イスラエル テルアビブ EMEA: イギリス、ドイツ APAC: 中国、韓国、オーストラリア (パートナー)
顧客	2003年10月に最初の製品を出荷し、約7 PBを納入
主要マーケット	ハイパフォーマンスなストレージ製品を必要とするユーザ エネルギー 自動車 航空宇宙 政府系研究機関 ライフサイエンス 大学 ファイナンシャル メディアなど
キーとなる出資者	The Carlyle Group Centennial Ventures Mohr Davidow Ventures

スケーラブルシステムズ株式会社

2

Panasas社のアーキテクチャの価値



Panasas社のアーキテクチャの特徴	効果と価値
スケーラブルなバンド幅 クラスタ環境でのスケーラブルなI/O性能の実現とシングルクライアントでの高いI/O性能	アプリケーションの実行効率の向上
グローバルネームスペースによる共有ファイルシステム データ移動が不要 データをより容易に見つけることが可能	ユーザと管理者の生産性の向上
標準製品を利用したハードウェア構成 一般に利用されるGbE/10GbE/InfiniBandなどにシームレスに統合可能 COTS/商用製品を利用することで、導入コストの低減を図る	導入コストの低減
標準ベースのソフトウェア SNIA/T10 で進められている次世代NFS/pNFSに対応	投資の保護
“アプライアンス” 真義による完全なソリューションの提供 数時間でのシステムのインストールと構築が可能 エンタープライズ用途にも対応したハードウェアとソフトウェア機能 (様々な高可用性機能)	運用管理が容易

ミッション: 高性能で運用管理が容易なストレージソリューションをユーザに提供すること

2007/10/16

スケーラブルシステムズ株式会社

4

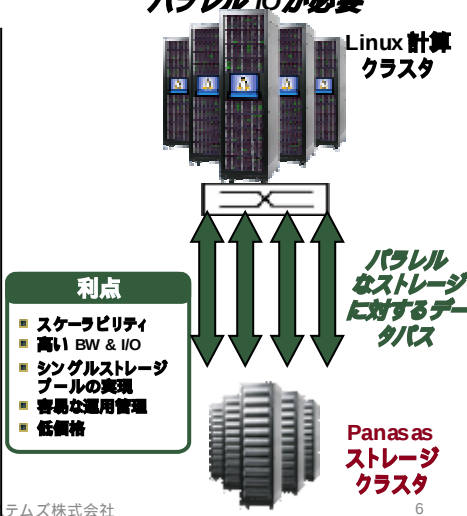
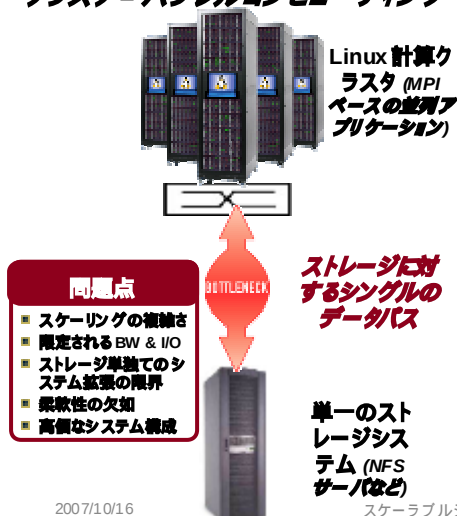
多くの先進的なユーザが採用

Oil & Gas 	Life Sciences 					
HPC Simulations 	<table border="1"> <tr> <td data-bbox="826 645 1008 824"> Finance  </td> <td data-bbox="1024 645 1248 824"> Fluid Dynamics  </td> </tr> <tr> <td data-bbox="826 840 944 974">  </td> <td data-bbox="965 840 1082 974">  </td> <td data-bbox="1104 840 1220 974">  </td> </tr> </table>	Finance 	Fluid Dynamics 			
Finance 	Fluid Dynamics 					
						

クラスタ利用時のボトルネックの問題

クラスタ = パラレルコンピューティング

パラレルコンピューティングでは、**パラレルIOが必要**

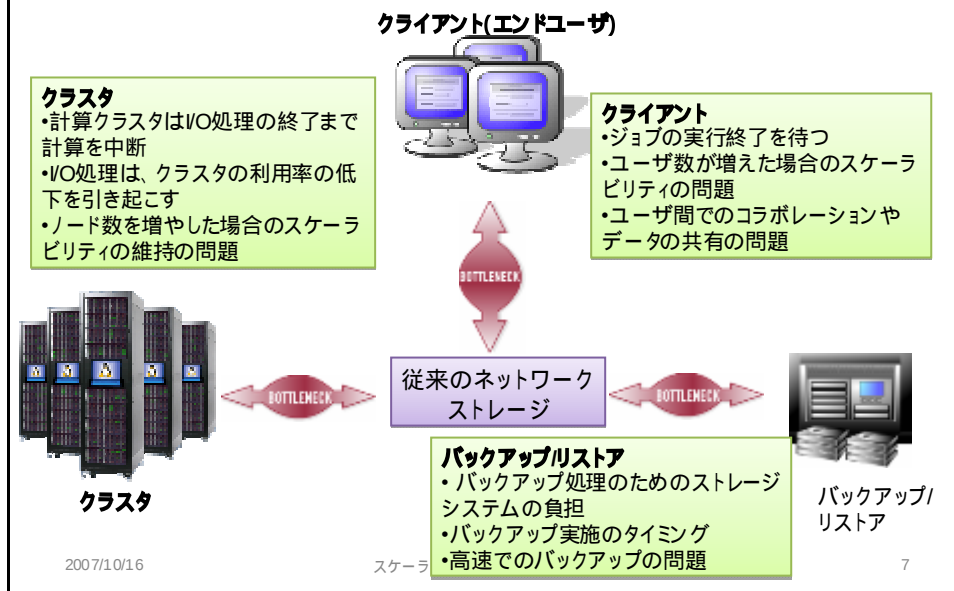


2007/10/16

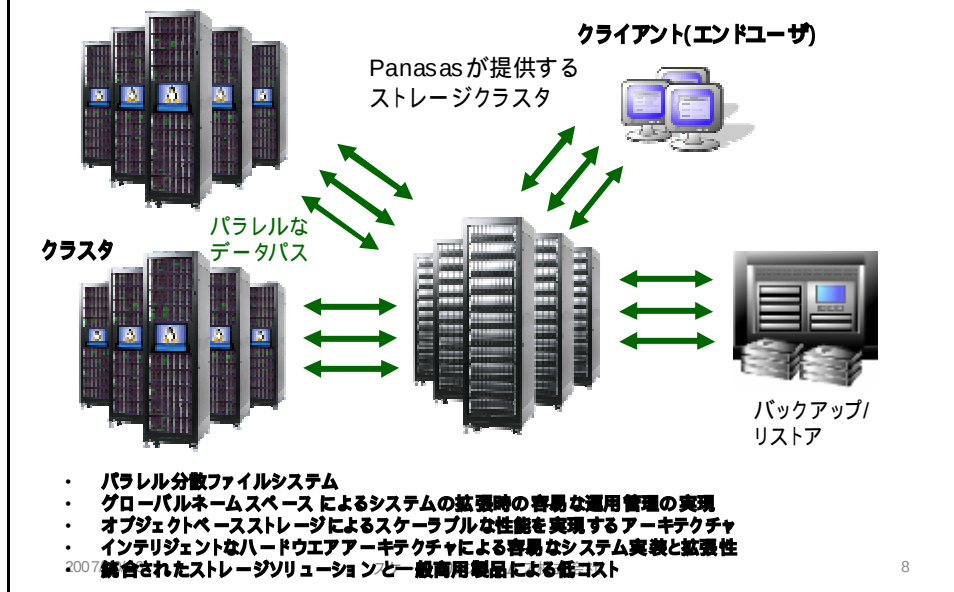
スケラブルシステムズ株式会社

6

ストレージに関する課題



クラスタでのボトルネックの解決



ストレージに対する要求



Interactive

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

“Run, Evaluate,
Re-Run”



“Run & Done”



Batch

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I/Oバンド幅
一言した可用性
シンプルなSW構成

Oil & Gas Seismic Interpretation

Oil & Gas Seismic Processing

EDA Design/Analysis

Chip Simulation & Tape Out

Auto/Aero Design

Air Flow and Crash Simulation

Animation SFX Creation

Rendering

Trading/Portfolio Mgmt

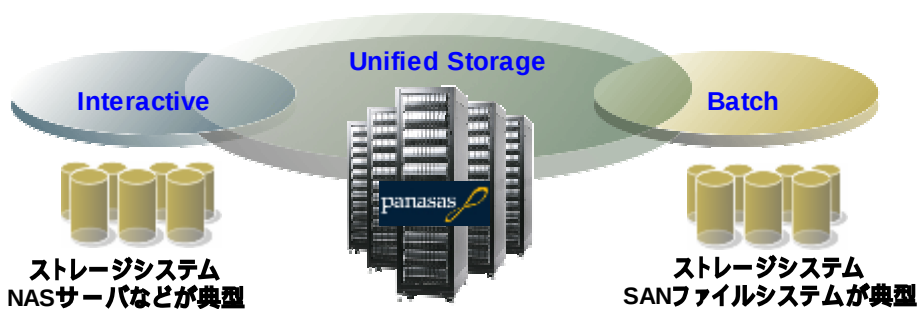
Risk Analysis

2007/10/16

スケーラブルシステムズ株式会社

9

‘Unified HPC Storage’の提案



完全に統合されたシングルポイントでアクセス可能なストレージプール

共有データへの高速で、容易なアクセスを可能とする

結果が得られるまでの時間を短縮(売り上げの増加)

データの多重保持が不要(コストの削減)

リソースのシングルプールとしての管理による容易な運用

2007/10/16

スケーラブルシステムズ株式会社

10

Panasas社が提供する構築ブロック



- オブジェクト
 - 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
- 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
- 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後で動作
 - 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

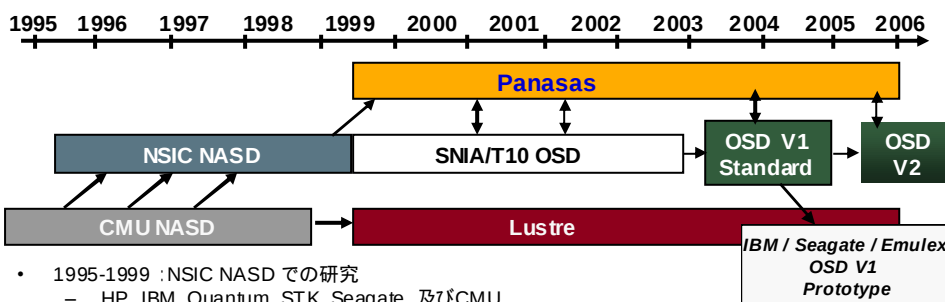


2007/10/16

スケーラブルシステムズ株式会社

11

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



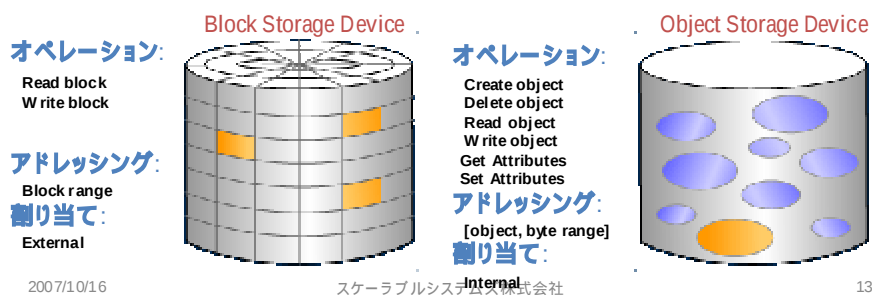
2007/10/16

スケーラブルシステムズ株式会社

12

オブジェクトストレージアーキテクチャ SCS

- 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



ActiveScale アーキテクチャ SCS

- クラスタテクノロジーによりスケーラブルなディスク容量と性能向上を実現: ディスク容量を増加させるとプロセッサ、キャッシュ、ネットワークバンド幅の向上となる
- クラスタ構築を一般商用品で行なうことで、スケーラブルな性能を廉価なHWで構築可能であり、高い価格/性能比を示す
- オブジェクトベースストレージは、ブロックベースのSANファイルシステムと比較して、より高いスケーラビリティと強固なセキュリティ・ポリシーの適用を可能とする
- クラスタ全体でストレージリソースのロードバランスの管理が可能
- 共有ファイルシステム (POSIX) としてのNASの利点とDASやSANのストレージ直結での性能の利点の双方を兼ね備えたシステム
- 業界標準に準拠したシステム構成:
 - iSCSI/OSD
 - (将来的には) pNFS
 - (将来的には) POSIX HECEWG: HPC IOのためのAPI拡張

ActiveScale 3.0: 大規模システムでの信頼性と運用管理の容易さを実現

- 予測的自己マネージメント
- ActiveScan – 予測的メディアマネージメント
 - 連続的にStorageBladeのデータとディスクメディアのチェック
 - 不具合が見つかった場合、システムは率先してメディア欠陥を修正
- 予測的ディスクマネージメント
 - 自動化した障害解析によりディスクに発生するであろう問題を予測
 - もし、ブレードの障害が予測される場合には、他のブレードへのデータ移動を行ない、システムの再構成を避ける
- 障害ブレードに対するパラレル再構築
 - システム中のDirectorBladeは、再構築を高速で実行するため、協調して動作
- ActiveUpgrade – オンラインでのクラスタストレージのアップグレード
 - アプリケーション動作中でもシステムのアップグレードが可能
- クライアントからのロード負荷をリアルタイムでモニター
 - ストレージユーザに対する性能のボトルネックを検知

ストレージクラス構成要素

- 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ェアとソフトウェアによるソリューション
 - 4Uのシェルフに11のブレード(5-10 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 クラスタファイルシステム
スマートに商用製品を活用したハードウェア構成

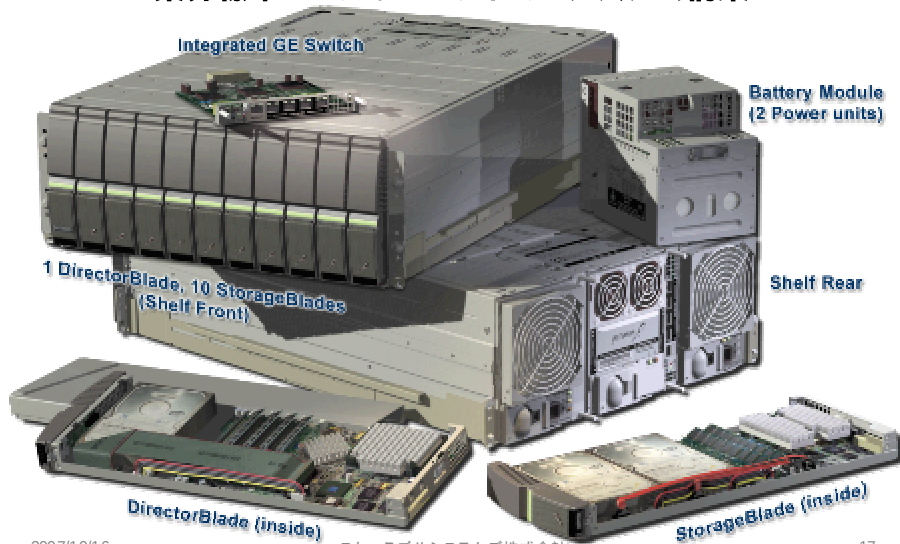


Panasonic ActiveScale
ストレージクラス

Panasas ストレージクラス



業界標準のコンポーネントでのシステム構築



2007/10/16

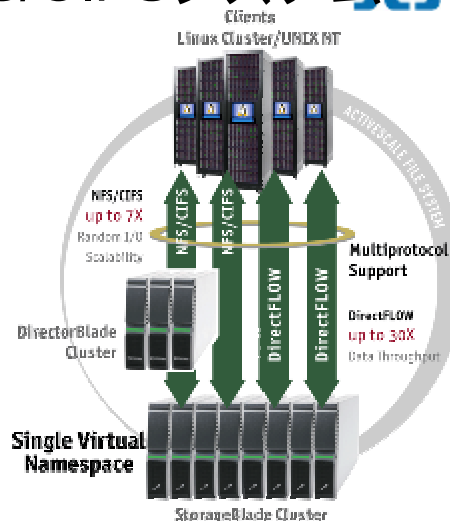
スケーラブルシステムズ株式会社

17

スケーラブルNFS/CIFSシステム



- スケーラブルな NFS/CIFS サーバ
 - 負荷を自動的にスケーラブルな DirectorBlade 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
 - いずれの DirectorBlade も、どのファイルにもアクセス可能
- 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
 - シングルファイルイメージ
 - DirectFLOW/NFS/CIFS 間で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



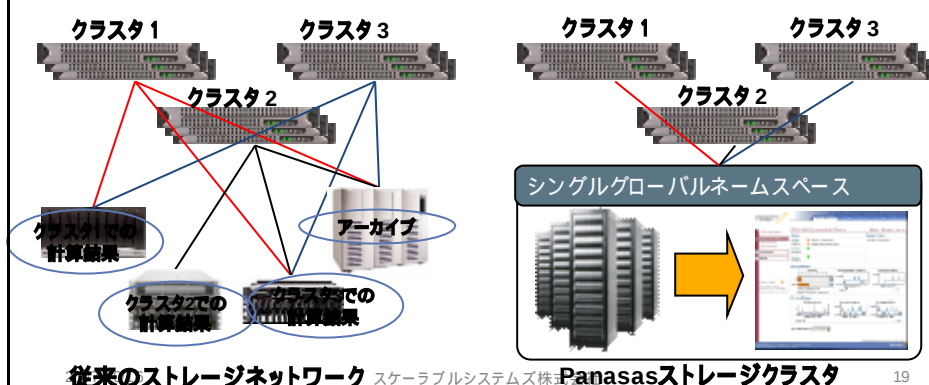
2007/10/16

スケーラブルシステムズ株式会社

18

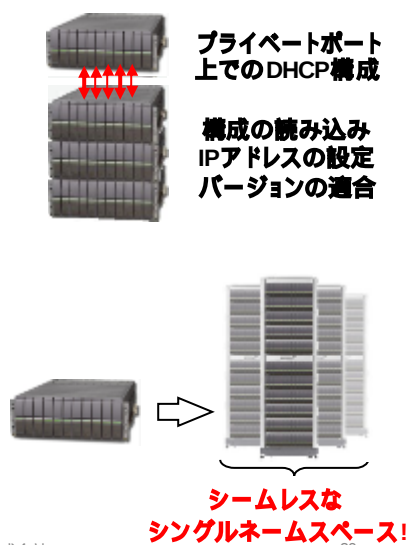
シングルグローバルネームスペース

- Panasasシステムでは、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



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Director Blade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



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

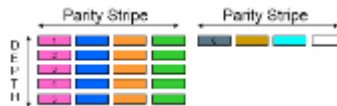
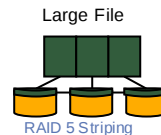
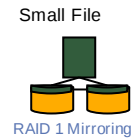
RAIDスペアと再構成の取り扱い

- 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
- スペアの分散
 - 一つのスペアに対して再構成を行うのではなく、Panasasのストレージクラスは、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

自動的なRAID構成



- ファイル単位でのRAID構成
 - RAID の構成レイアウトは、オブジェクト中に属性として格納される
- システムは、RAIDレベルをファイルサイズに応じて、自動的にアサインする
 - < 64 KB RAID 1 構成 (効率的なスペース配置)
 - > 64 KB RAID 5 構成 (最大性能の実現)
- 自動的にRAID 1 から RAID 5に再構成
 - ストライピングの再実行は不要
- RAID マップの構成は、二重構造(ストライプ幅と深さ)
 - ストライプサイズは自動的に最適化



システムの規模に応じた最適なデータ配置と障害時のファイル再構成を高速に(並列に)実行する

スケラブルシステムズ株式会社

23

データの保護と可用性



- Panasasのストレージクラスターのハードウェア設計の基本
 - 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
 - UPS内蔵
 - ECCメモリ
 - Shelf内にアックアップのネットワークを内蔵
- Panasasソフトウェアでの信頼性向上のための機能
 - 高速な再構成が可能なRAID 1 及び 5 を提供 (RAID 6もサポート予定)
 - 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ウンドで実行されるため、サービスの停止は不要
 - 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
 - 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
 - システムのサービスのフェイルオーバー、ファイルサービスのフェイルオーバー、メタデータ処理のフェイルオーバーなど
 - スケラブルな高速バックアップのサポート
 - 信頼性の高いオペレーティングシステム

2007/10/16

スケラブルシステムズ株式会社

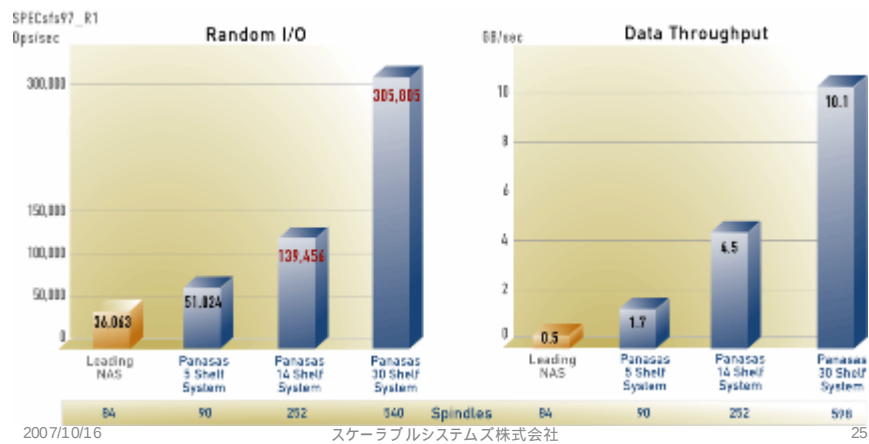
24

業界最高性能を実証



データスループットとランダムI/Oでの卓越した性能

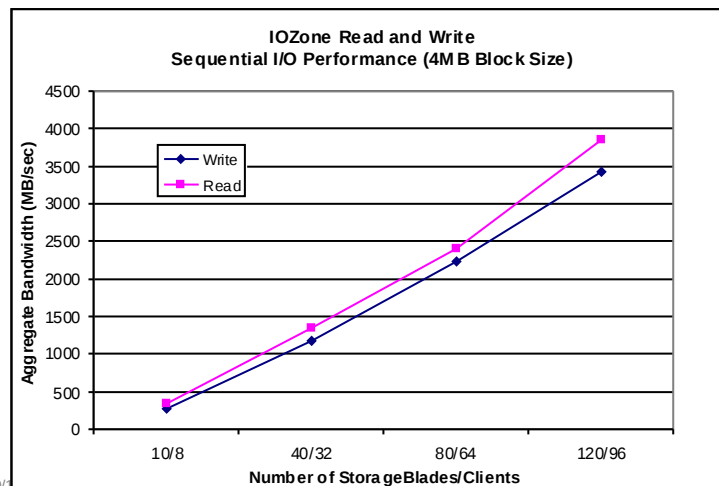
全てのワークロードに対して、高い性能とスケーラビリティを提供



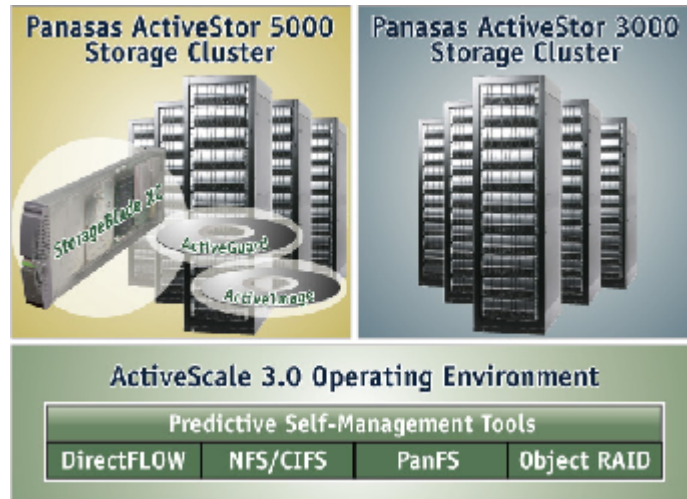
システムのスケールアップと性能向上



クライアント数とシステム規模双方のスケーラビリティ (N-to-N Iozoneベンチマーク)



Panasas 製品ファミリー



2007/10/16

スケラブルシステムズ株式会社

27

ActiveStor 5000: ‘Unified HPC’ クラスタストレージ



ActiveStor 5000 Storage Cluster



- 最大 11 StorageBlades / シェルフ
- 最大 3 DirectorBlades / シェルフ
- 500GB 又は 1TB StorageBlades
- シェルフ当たり10TB , 42Uサイズのラックに最大 100 TB搭載

ActiveScale 3.0 オペレーティングシステム

予測的自己管理ツール	
DirectFLOW	NFS/CIFS
PanFS	Object RAID

新機能:

- ActiveScale 3.0オペレーティングシステム
- StorageBlade XC (2GB cache)
 - 20GB Cache / シェルフ
- ActiveGuard
 - 高可用性の実現
- ActiveImage
 - ボリューム毎のスナップショットの実現

2007/10/16

スケラブルシステムズ株式会社

28

ActiveStor 5000: システム拡張時の可用性



- 全てのストレージシステムにおいて、‘起動時間 (Uptime)’は重要
 - ‘Unified HPC’ストレージにおいては、起動時間は最も重要
- ActiveGuard
 - 高可用性-ノンストップオペレーションを可能とする
 - ストレージネットワークのフェイルオーバー
 - DirectorBladeのメタデータのフェイルオーバー
- ActiveImage
 - 仮想ボリュームの選択 – イメージの定期的な自動作成と削除
 - 前回のイメージとの差分だけを保持することで、容量を最小とする
 - システムあたり最大32,000 ActiveImages (スナップショット)

2007/10/16

スケーラブルシステムズ株式会社

29

ActiveStor 3000:



大規模クラスタ向けスケーラブルストレージ

ActiveStor 3000 ストレージクラスタ
(Batch HPC マーケット)

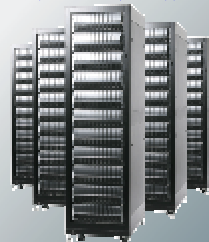


- 最大11 StorageBlades / シェルフ
- 最大3 DirectorBlades / シェルフ
- ブレード当たり512MBキャッシュ

新機能:

- ActiveScale 3.0オペレーティングシステム
- 1TB StorageBlades
- 42Uサイズのラックに最大 100 TB搭載
- プレインストールラック構成
- RoHS指令 準拠

Panasonic ActiveStor 3000
ストレージクラスタ



ActiveScale 3.0
オペレーティングシステム

予測的自己管理ツール

DirectFLOW

NFS/CIFS

PanFS

Object RAID

2007/10/16

スケーラブルシステムズ株式会社

30

ユーザ事例



- 国公立研究機関、大学、エネルギー、製造業、ファイナンス、メディアなどの分野でのシミュレーションとモデリングのためのアプリケーションを利用しているHPC機関とユーザ(CFD, FEA, EDA, Weather, 計算化学など)
- 32ノードを越えるLinuxクラスタシステムのストレージや多くのクライアントワークステーションを利用しているエンジニアリンググループ
- 最低でも10TBのストレージ容量の要求
- 大規模なファイルへのアクセスや非常に多数のファイルへのランダムなアクセスを行なう場合
- 典型的なNFSサーバの用途



2007/10/16

スケーラブルシステムズ株式会社

31

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



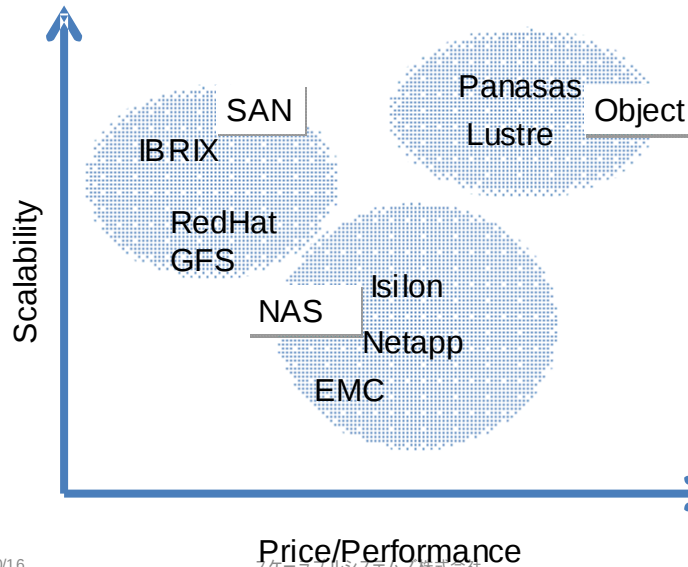
機能とその利点	Panasas ActiveStor	NAS サーバ (NetApp, EMC, start-ups)	SAN ファイルシステム (Lustre, GPFS)
ターゲットとするアプリケーション	Batch + Interactive	Interactive	Batch
高いバンド幅	✓		✓
クライアント数のスケーラビリティ	✓		✓
ストレージ容量のスケーラビリティ	✓		✓
NFSとCIFSのサポート	✓	✓	
統合システム	✓	✓	
可用性	✓	✓	
高いランダムIO性能	✓	✓	

2007/10/16

スケーラブルシステムズ株式会社

32

他のファイルシステムとの比較 SCS



2007/10/16

スケーラブルシステムズ株式会社

33

他のファイルシステムとの比較 SCS

- Lustre
 - 規格外のオブジェクトストレージインターフェイス (T10 OSDに非準拠)
 - システムの安定性を無視した性能重視の設計
 - SANマネジメントの問題と‘オープンソース’の問題
- GPFS
 - ブロックベースのアーキテクチャによる限界
 - ファイルシステムあたり、シングルトークンでのマネジメント
 - 特定システム (IBM SP2) をターゲットとして、設計されたシステム
- Isilon, IBRIX, Polyserve
 - NFSクラスタとしての限界
- NetApp GX (Spinnaker)

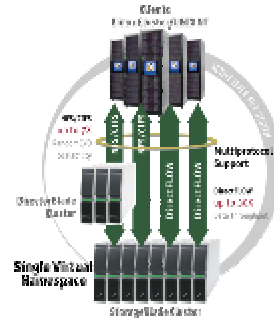
2007/10/16

スケーラブルシステムズ株式会社

34

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



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

2007/10/16

スケーラブルシステムズ株式会社

35

まとめとして

	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 faster
	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
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	ベタスケール

2007/10/16

スケーラブルシステムズ株式会社

36