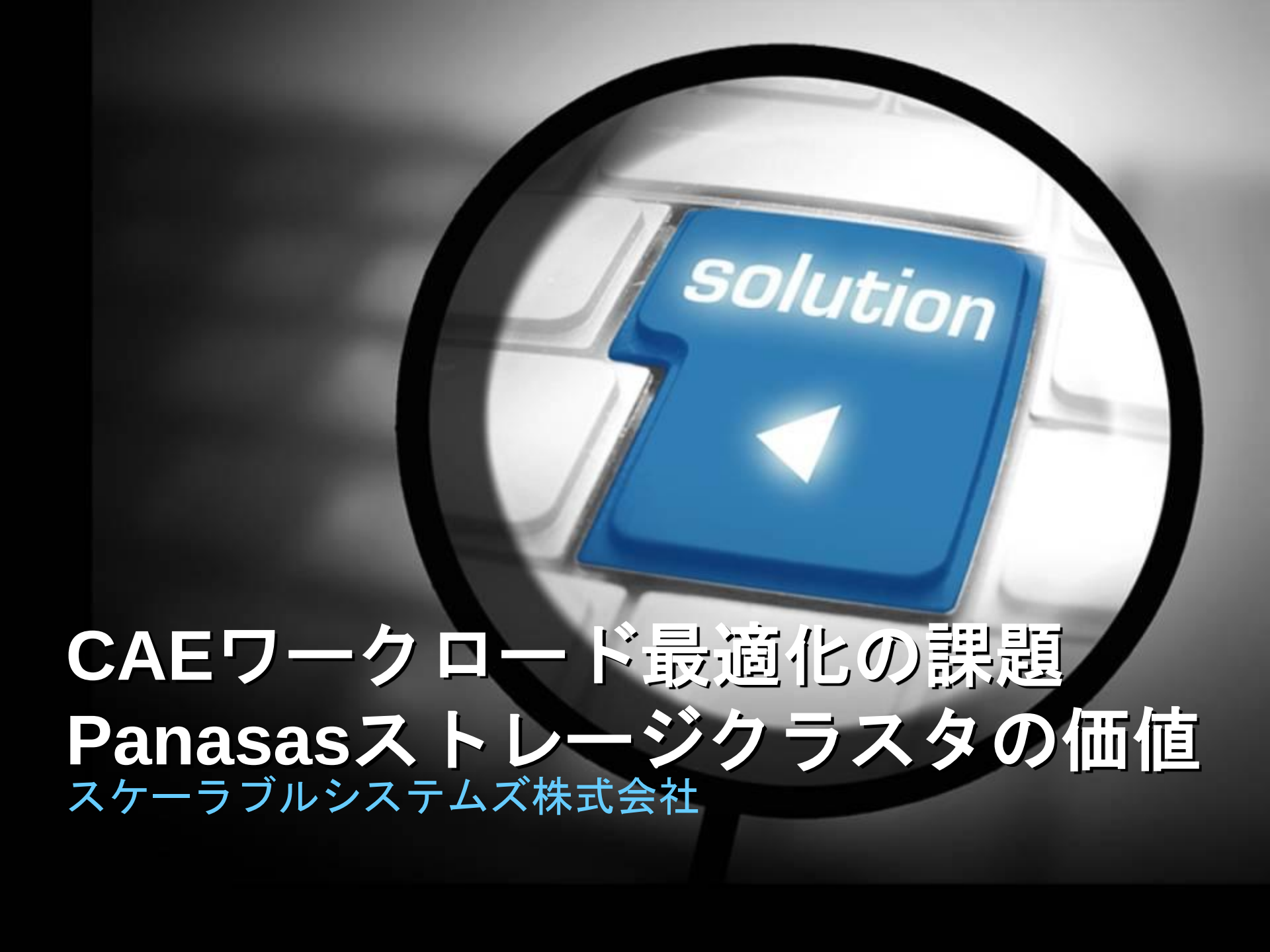


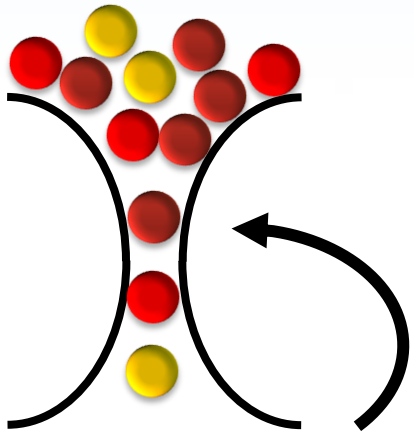
CAEワークロード最適化の課題 Panasasストレージクラスタの価値

スケーラブルシステムズ株式会社

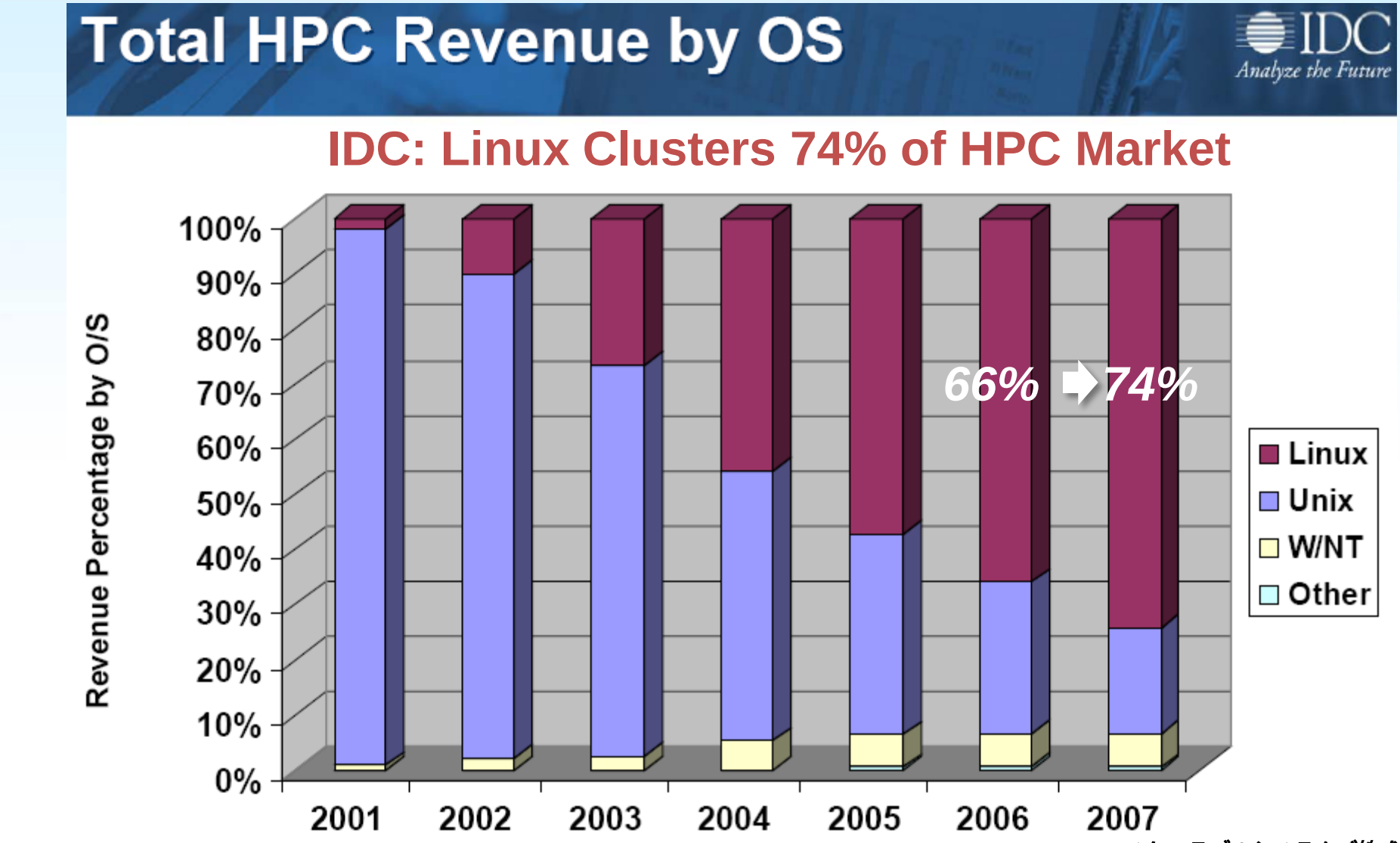
CAEワークロード最適化の課題
Panasasストレージクラスタの価値

- はじめに
  - Panasasストレージクラスタ導入事例

    （この部分は、弊社にお尋ねください）
- Panasasストレージクラスタの価値
  - ボトルネックの解消
  - CAEワークフローの改善
  - CAEワークロードに対する
    柔軟なシステム構築
- まとめとして



HPCクラスタの導入動向

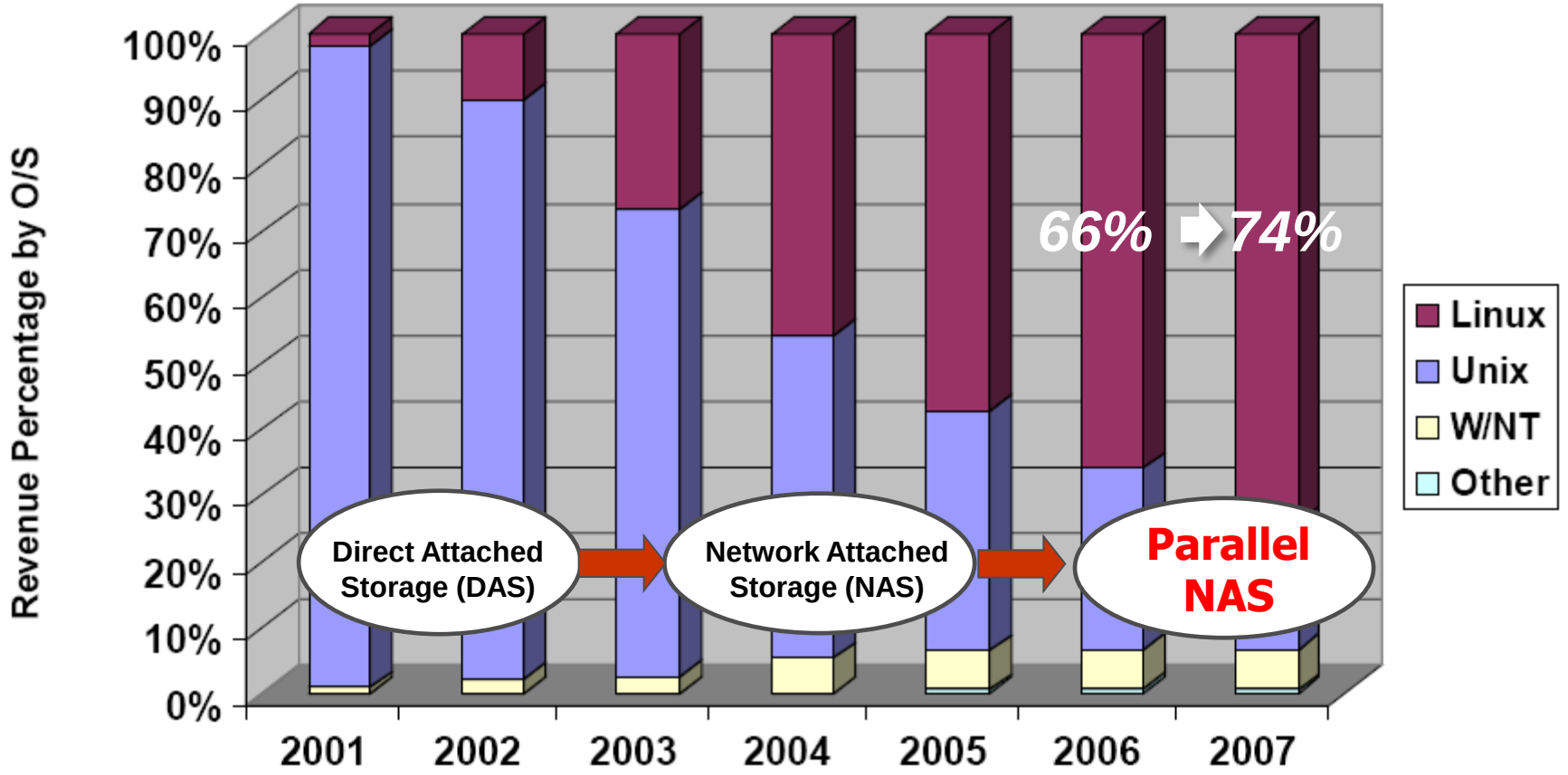


HPCクラスタの導入動向

Total HPC Revenue by OS

IDC
Analyze the Future

IDC: Linux Clusters 74% of HPC Market



ユーザにとっての現状の課題

クラスタを利用すること、また、その管理運用を行うことは現在でも多くの課題がある

- クラスタの規模が大きくなることに伴いシステムがより複雑になる
- 電力、冷却、設置スペースは、大きな問題
- ソフトウエアコストの問題
- 全てのレベルでのインターコネクトの性能の問題
- アプリケーション及び自作プログラムの並列化 – 複数ノードを利用しての高い性能の実現
- RAS (信頼性、可用性、サービス性）はより重要
- ストレージとデータマネージメントでのボトルネックが顕在化
- ヘテロな計算環境に対する取り組み

ユーザにとっての現状の課題

ソフトウエアの問題がシステム導入に際しての第一の課題となりつつある

- より優れた管理ツールが必要
- HPCクラスタのセットアップと運用をより容易にすること
- より広範囲なユーザに対して – 新規導入に際して、「より簡便で容易に」が求められる
- パラレルソフトウエアの課題
  - 多くのアプリケーションはアプリケーション自身の再設計が必要
  - マイクロプロセッサのマルチコア化に際しては、性能や機能などの面で新しい課題をもたらす

ITデータセンターでの新たな課題

- プロセッサコア数（プロセッサ数 x マルチコア x サーバ数）の増加はITに大きな課題をもたらす：
- 運用管理の複雑さ
  - 複雑なクラスタシステムをより効率良くマネージメントするか？
  - 「個々のコンポーネント」大量に購入することなくクラスタをインストールし、セットアップを行うか？
- 電力/冷却/設置スペース
- アプリケーションのスケーラビリティとハードウエアの利用率
  - 現在利用中のアプリケーションを「変更なし」で如何に効率良く、また、スケーラビリティの向上を図るのか？
  - 新しいマイクロプロセッサやシステム設計をどのように利用するのか？

典型的なCAEワークフローとボトルネック

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

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

BOTTLENECK

BOTTLENECK

従来のネットワークストレージ

BOTTLENECK

クラスタ

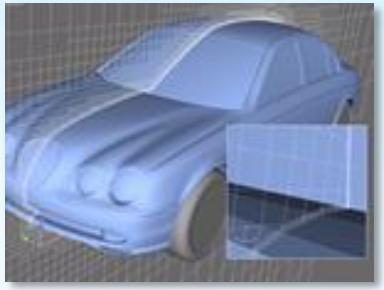
バックアップ/リストア
- バックアップ処理のためのストレージシステムの負担
- バックアップ実施のタイミング
- 高速でのバックアップの問題

バックアップ/リストア

CAEにおけるI/O処理での課題

CAEシュミレーションでのリアリズムの追及への高い要望

- より大きなモデル＋より多くの機能の追加要求=高精細なCAEシュミレーション



高精細なCAEシュミレーションの実現のための分散並列処理

- 高い並列度でのシュミレーション処理の一般化
- X86プロセッサのマルチコアによるコストの削減

高い並列度での処理では、I/Oストリームとファイルサイズの増大が問題

- 計算は高い並列度で分散し、I/Oが逐次処理の場合のボトルネックの問題

CAEにおけるI/O処理での課題

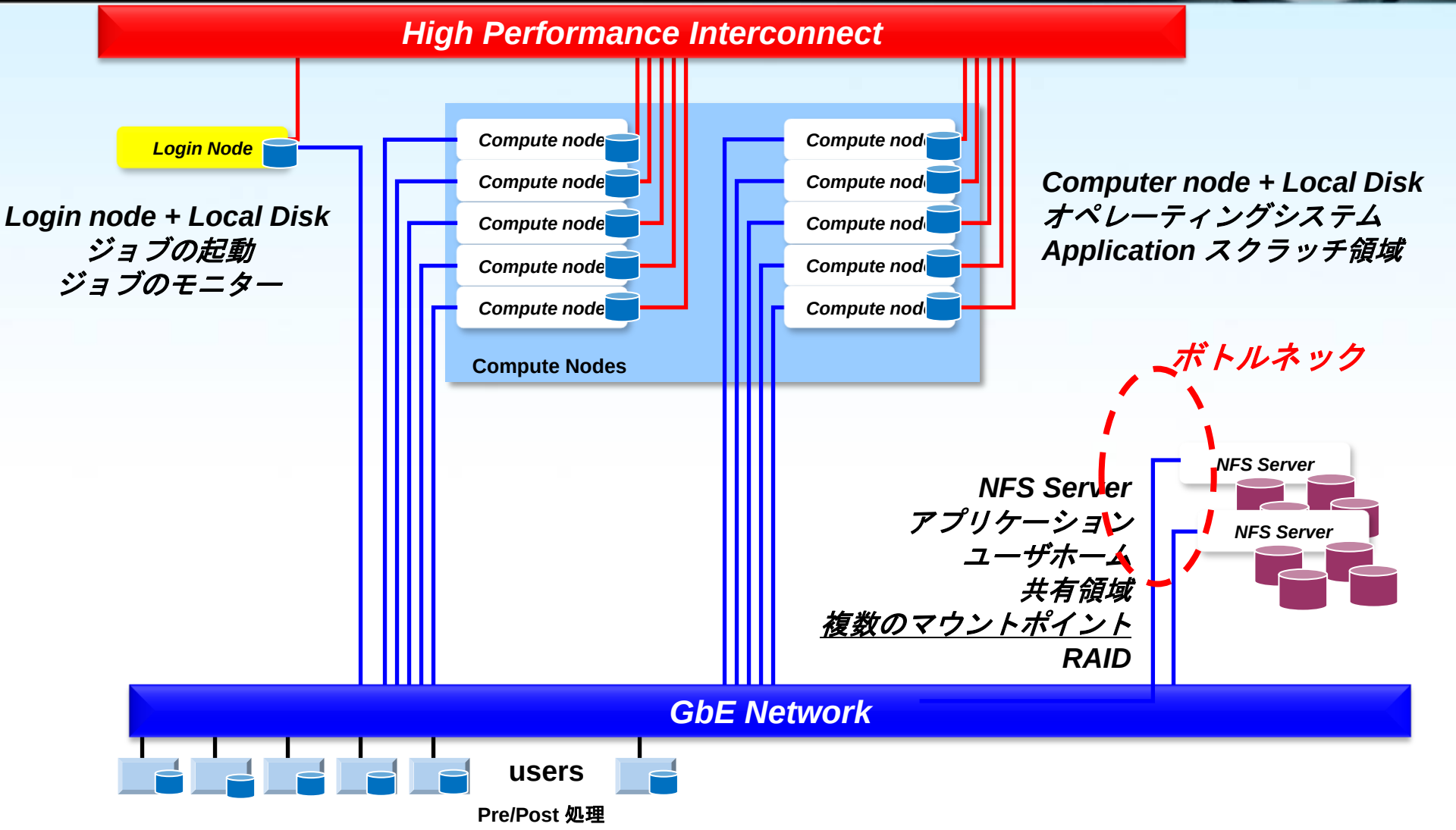
ワークフロー(ユーザのコラボレーションタスク)

- CAEソルバーでのメッシュ分割などのプリ処理の時間
- ネットワークでのファイル転送の負荷
- CAEシュミレーションの解析モデルや計算結果の管理

ワークロード(クラスタでの並列処理)

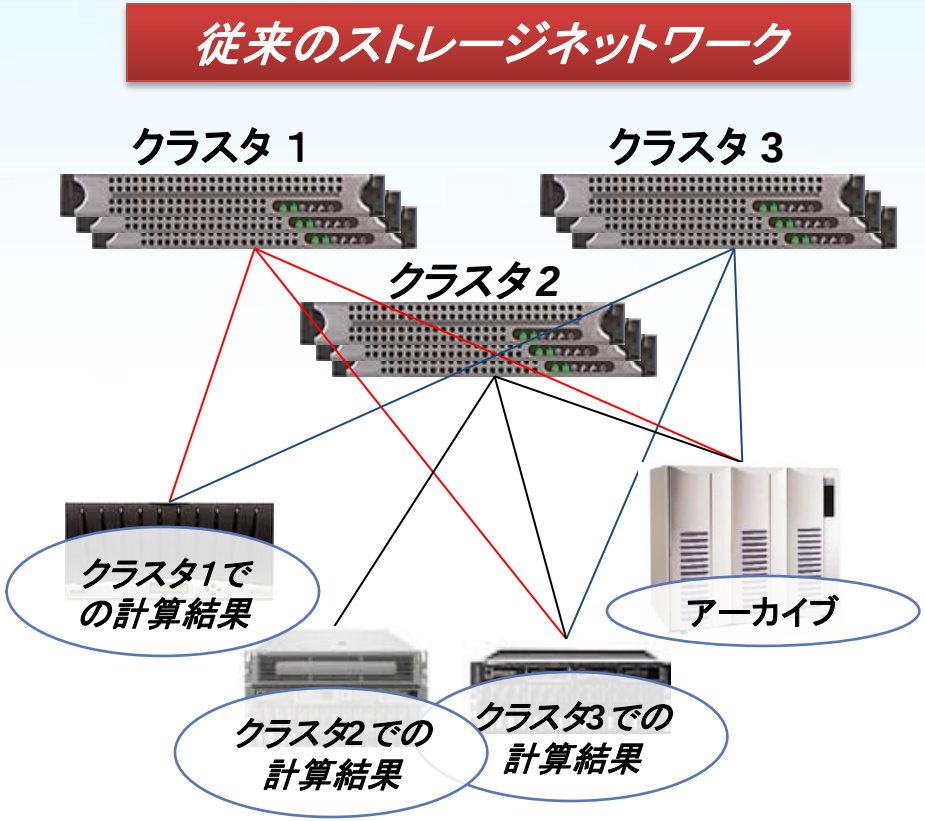
- I/Oリソースに対して様々なワークロード処理
- 非定常計算などでのより頻度の高いデータ書き出し要求
- より自由度の大きな計算処理における‘Out-of-Core’ ソルバーの利用
- 最適化計算などにおけるモデリングの自動化とパラメータ解析
- マルチスケール、多変量、統合モデリングシュミレーションへの対応

CAEワークロードでのボトルネック



NFSによる共有ストレージの実現

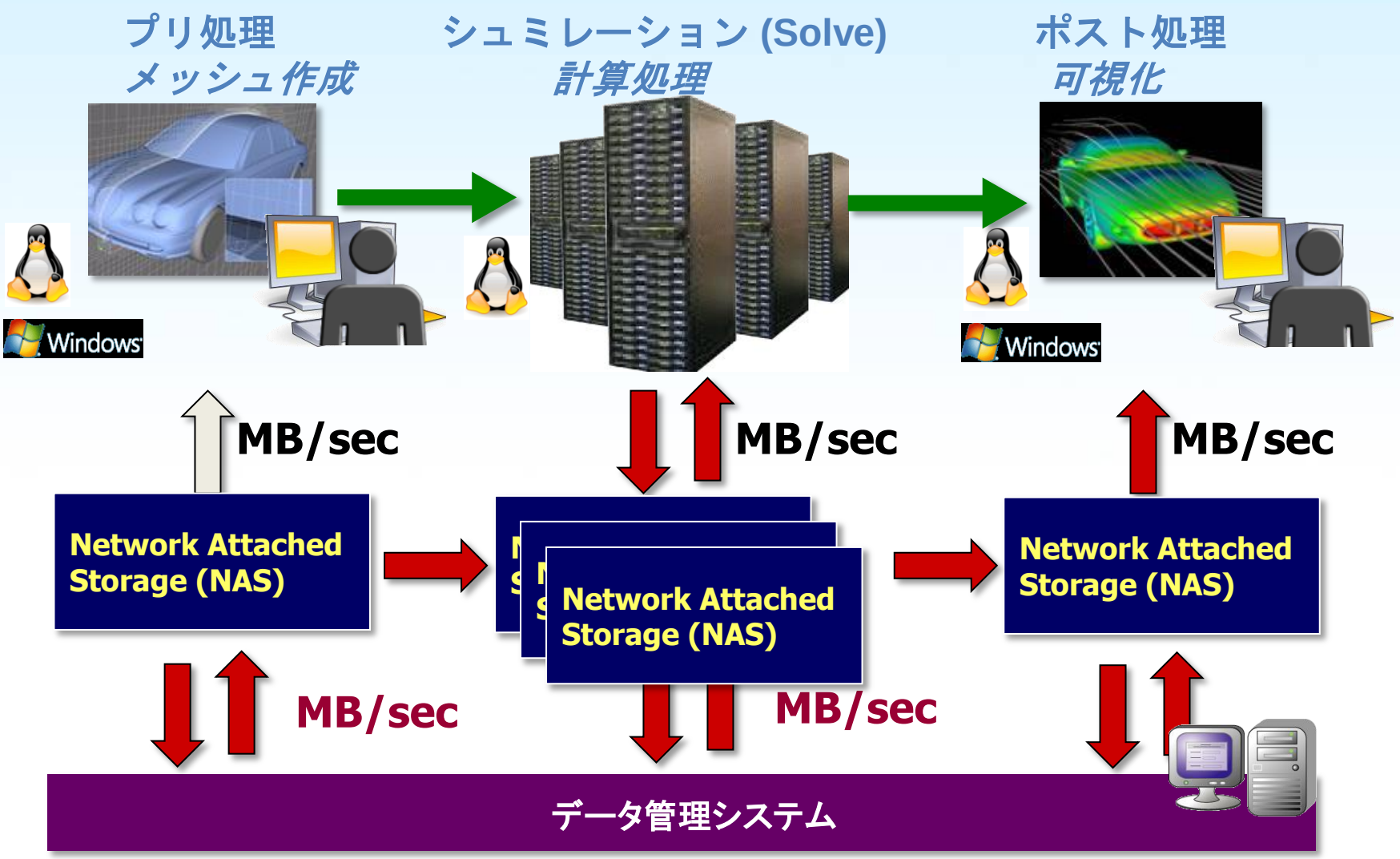
- クラスタやSMPシステムによる計算システムの構築に際して、NFSによる共有ストレージ構築の構築



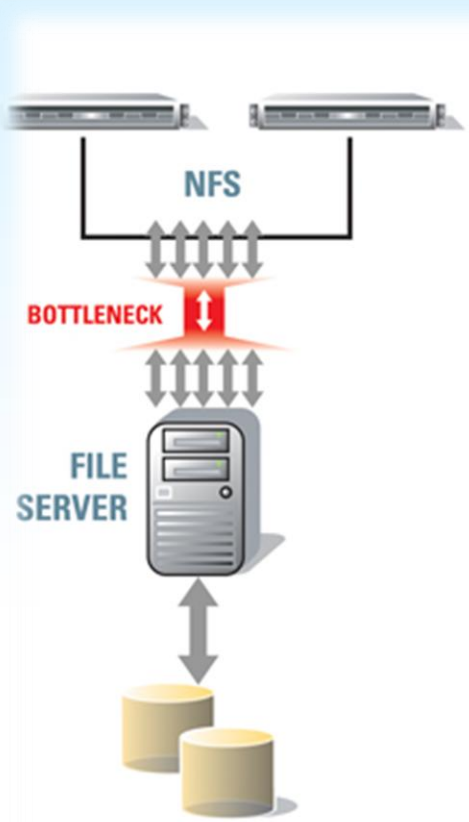
問題点と限界

- 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

CAEにおけるI/O処理での課題
ボトルネックの解消

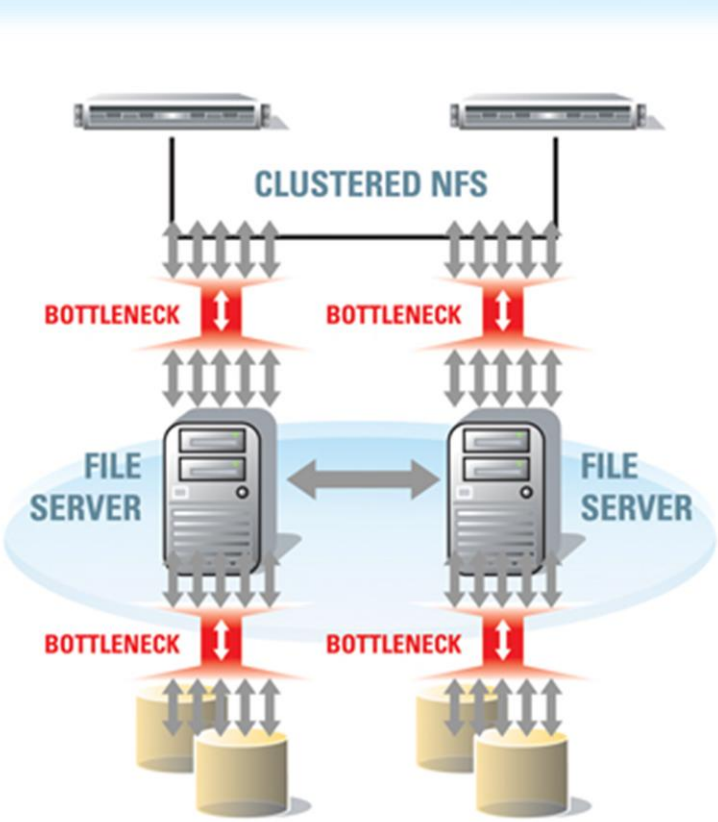


ストレージアーキテクチャ
パラレルストレージの価値の検証



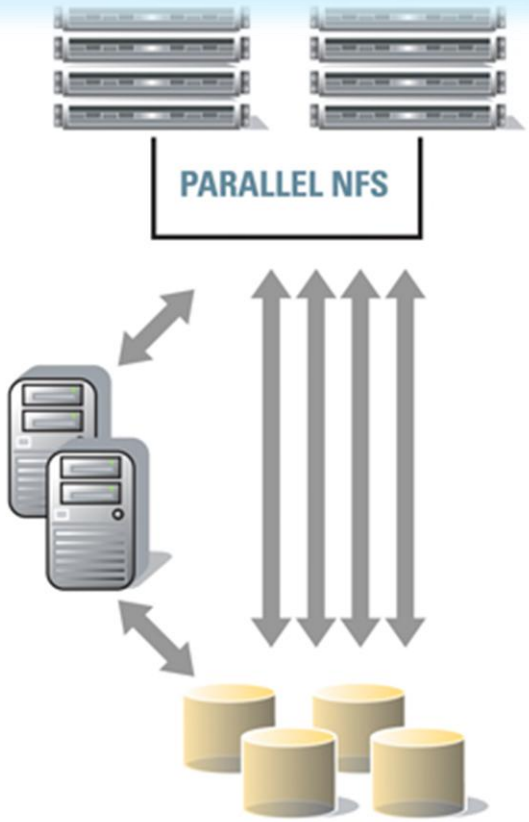
NAS

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

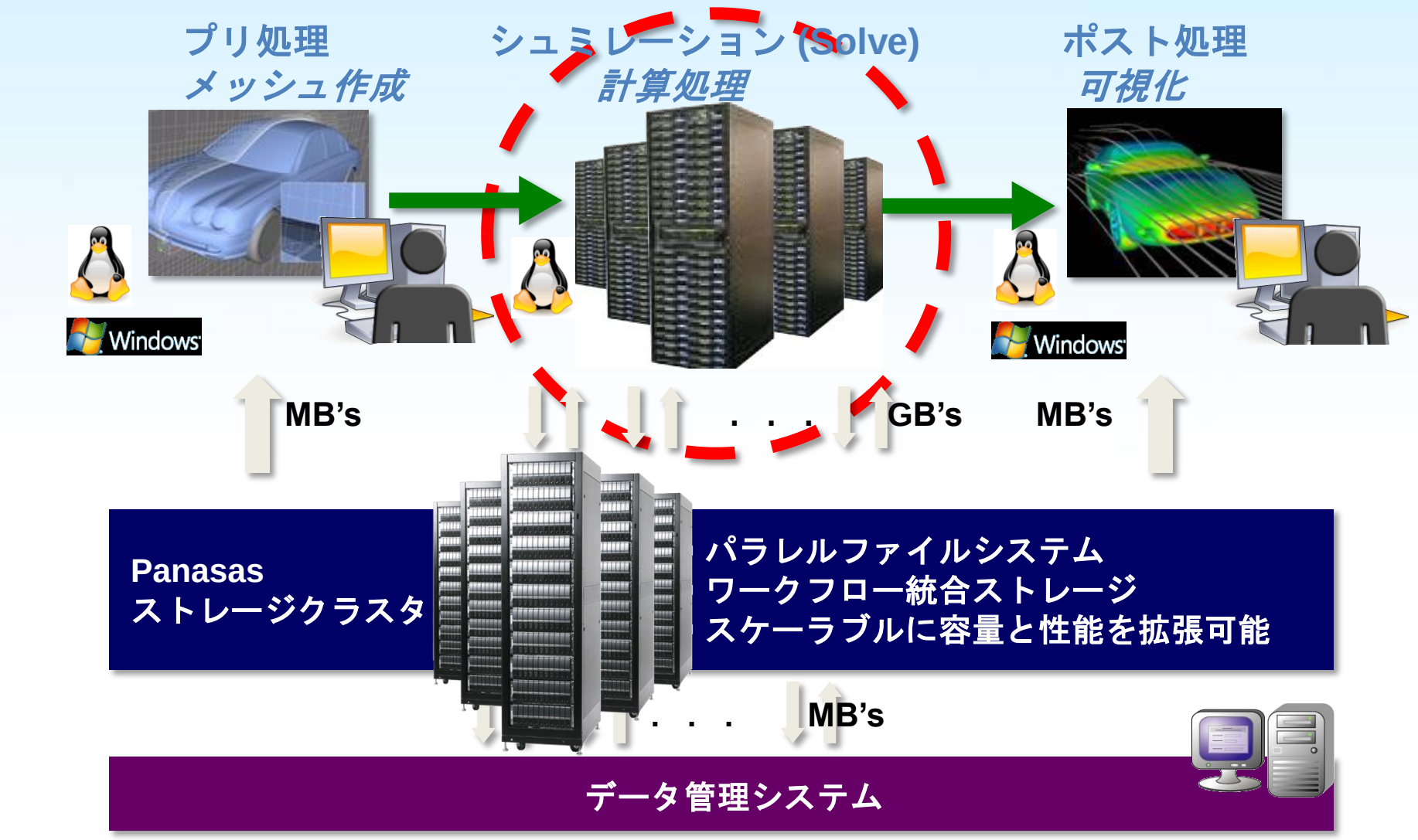
ファイルサーバを経由しないデータ
転送パス
シリアルI/Oのボトルネックの解消と
容易なシステム全体の運用管理

従来型ストレージの課題

従来のシステム	問題点
システムの導入前に‘利用時のシナリオ’を想定して、システム設定を行う	•実際に運用を開始した場合、想定シナリオ以外の利用方法や要望が発生した場合の対応が必要 •システムの再設計及び再構築の必要性が発生しても対応が難しい
WindowsとLinuxでのファイル共有	•複数のファイルサーバをOS毎に用意して利用した場合、各ファイルサーバ間でのファイル転送が必要 •計算ノードとクライアントからの同時データアクセス
NFSでのボトルネック	•クラスタの各ノードからの同時I/Oリクエストに対応するには、NASヘッドの性能とバンド幅が必要
ディスクドライブの高密度化によるメディアエラー	•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る
並列アプリケーションと複数ジョブの同時実行の双方の同時実行	•アプリケーション毎に要求されるI/O性能が異なり、また、複数ジョブの同時実行に際して、リソース競合による各ジョブの実行時間が一定しない

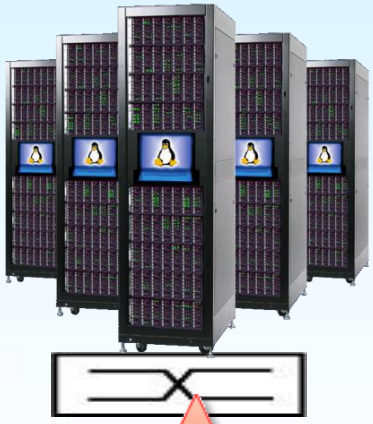
Panasasによるボトルネックの解消とCAEワークフローの改善

ボトルネックの解消
①スケーラブルなI/O処理の実現

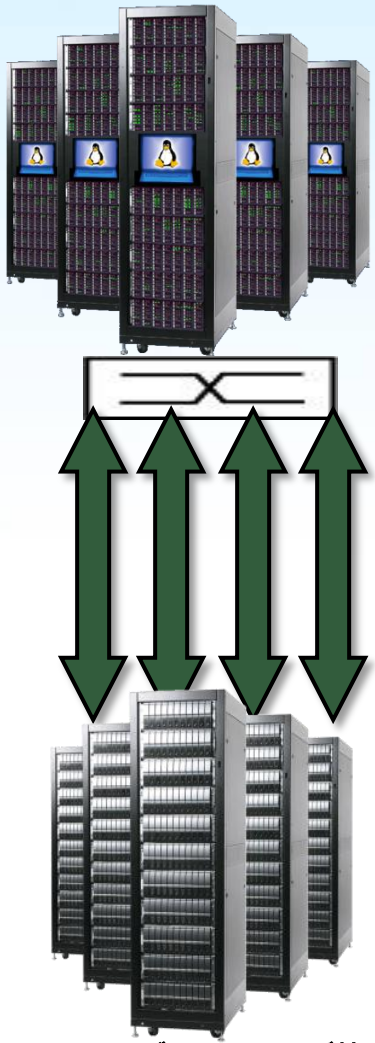


クラスタ利用時のボトルネック

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



CAEにおけるI/Oボトルネック

CAEでのシングルジョブのI/O処理の比重

1999: Desktops

Typical overnight job – single thread

read = 10 min　compute = 14 hrs　write = 35 min　5% I/O

2004: SMP Servers

注意: 以下の推定ではジョブのI/O時間は一定で、並列処理によって、非I/O処理部分の時間は短縮されると仮定しています.....

8-way　I/O significant

10 min　1.75 hrs　35 min　30% I/O

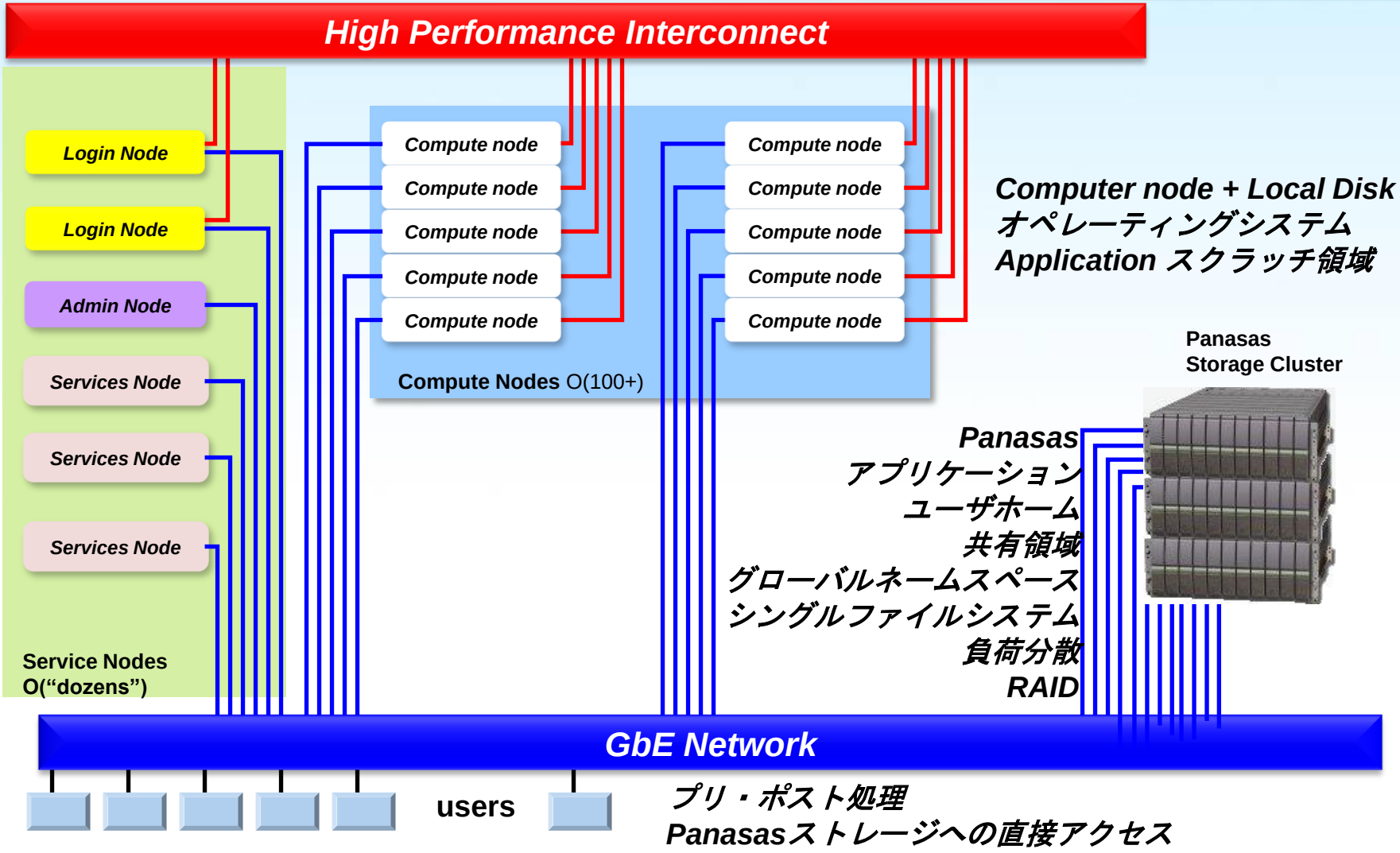
2009: HPC Clusters

32-way　I/O-bound – growing bottleneck

10 min　26 min　35 min　*63% I/O*

注意: I/O処理部分に関して、性能向上や並列化などの改善がないという極端な仮定での推定であり、実際のCAEでのシングルジョブのI/O処理を完全にシュミレーションした結果ではありません。

Panasasストレージクラスタ



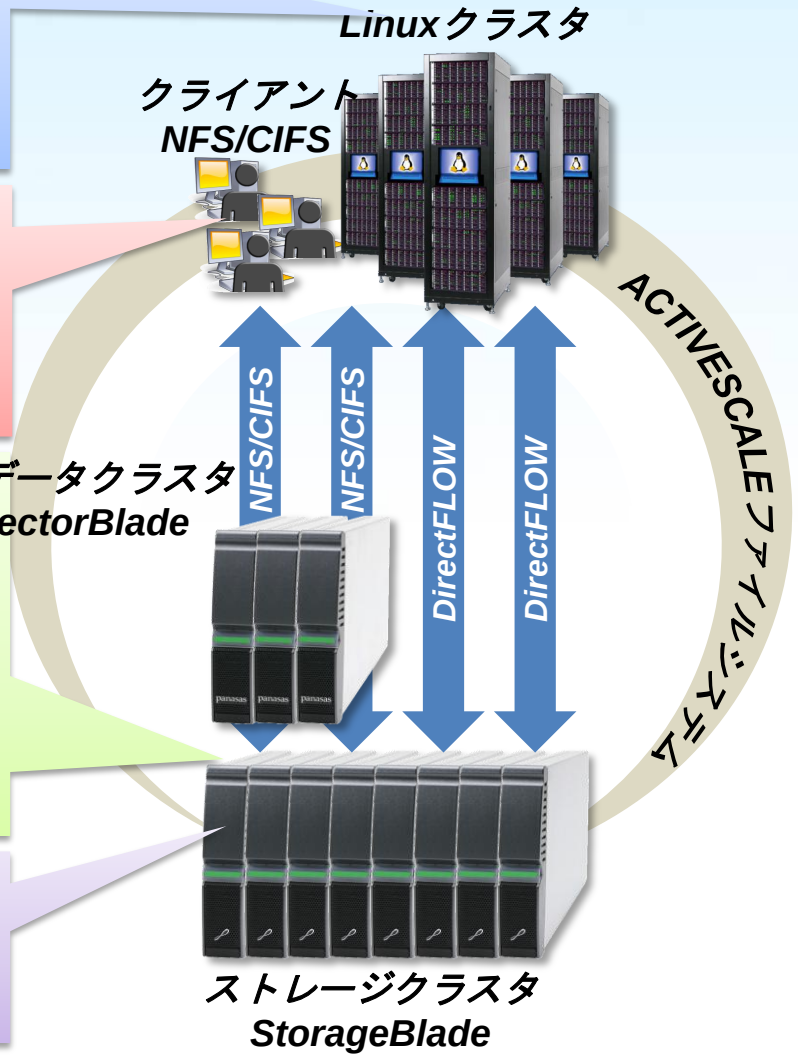
Panasasストレージクラスタ

DirectFLOW クライアントS/W
- *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にも対応可能

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*

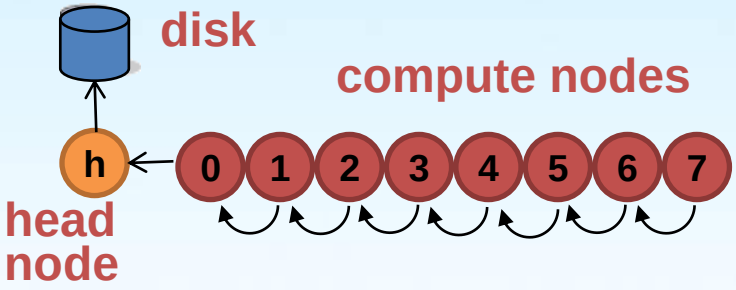
オブジェクトベース
- 優れたスケーラビリティ、信頼性、運用管理
- *Panasas Tiered Parityによるデータ保護の強化*



NFS、PanFS、DASのI/O処理

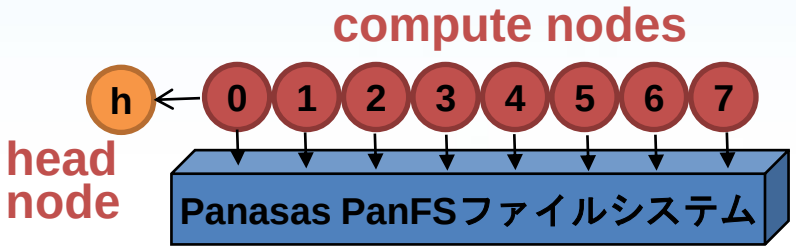
NFS

- 一般のネットワークファイルシステムの場合、ヘッドノードが各パーティション上のデータを集めて、シリアルにI/Oを行う



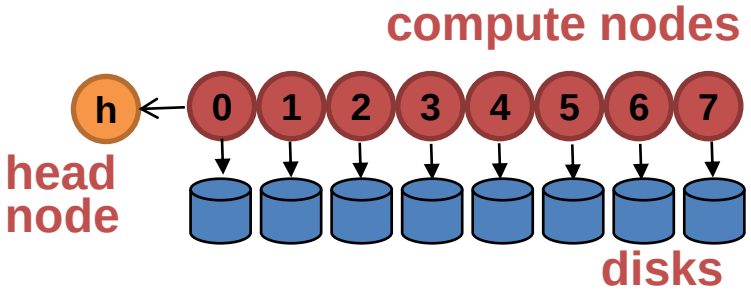
PanFS

- 並列処理が可能なネットワークファイルシステムの場合、各ノードは、共有されたストレージに対するそれぞれのデータパスを持つ

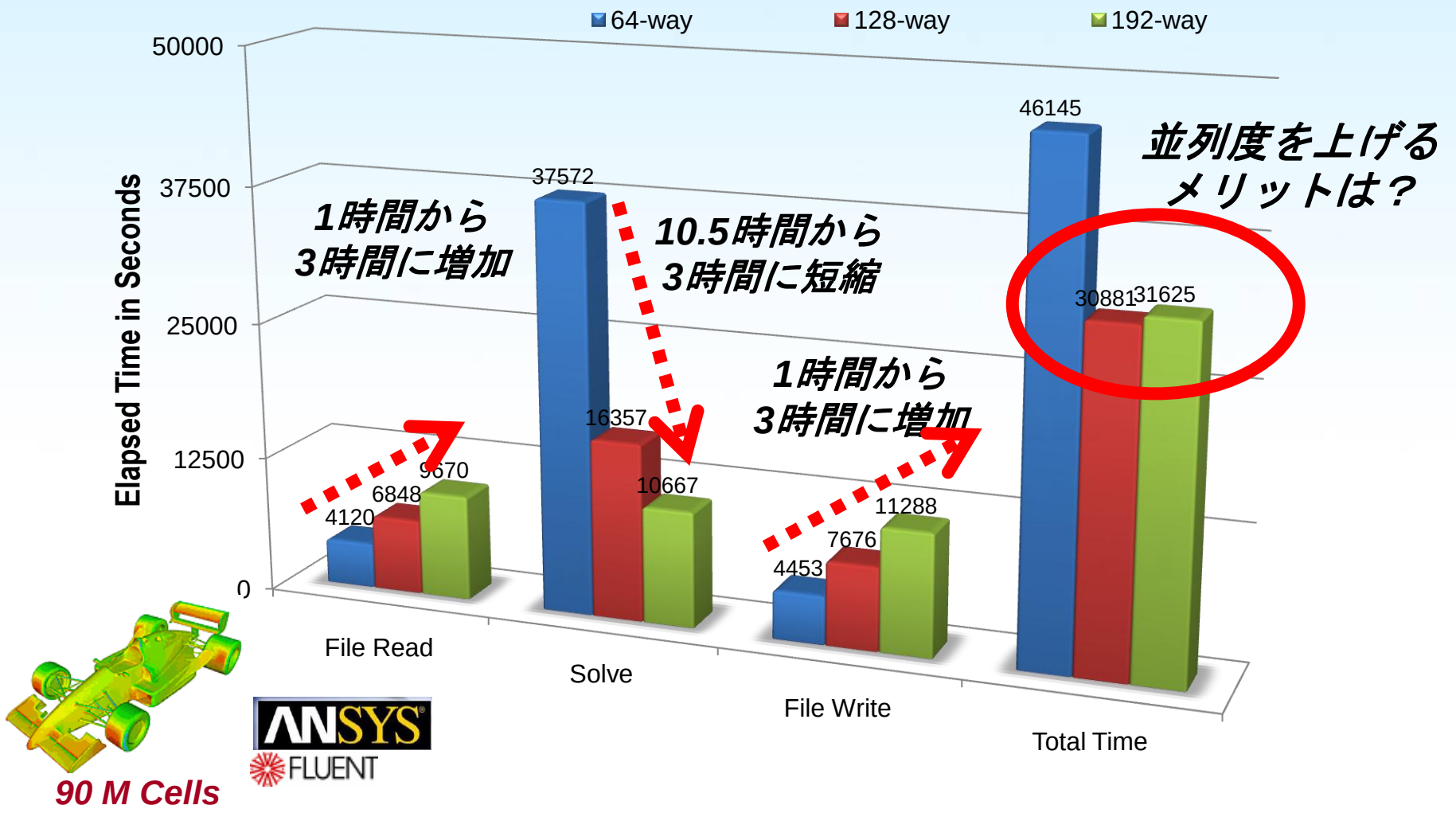


DAS

- ダイレクト・アタッチド・ストレージでは、個々のローカルファイルシステムに並列にI/Oを行う
- 個々のファイルシステムの管理が必要

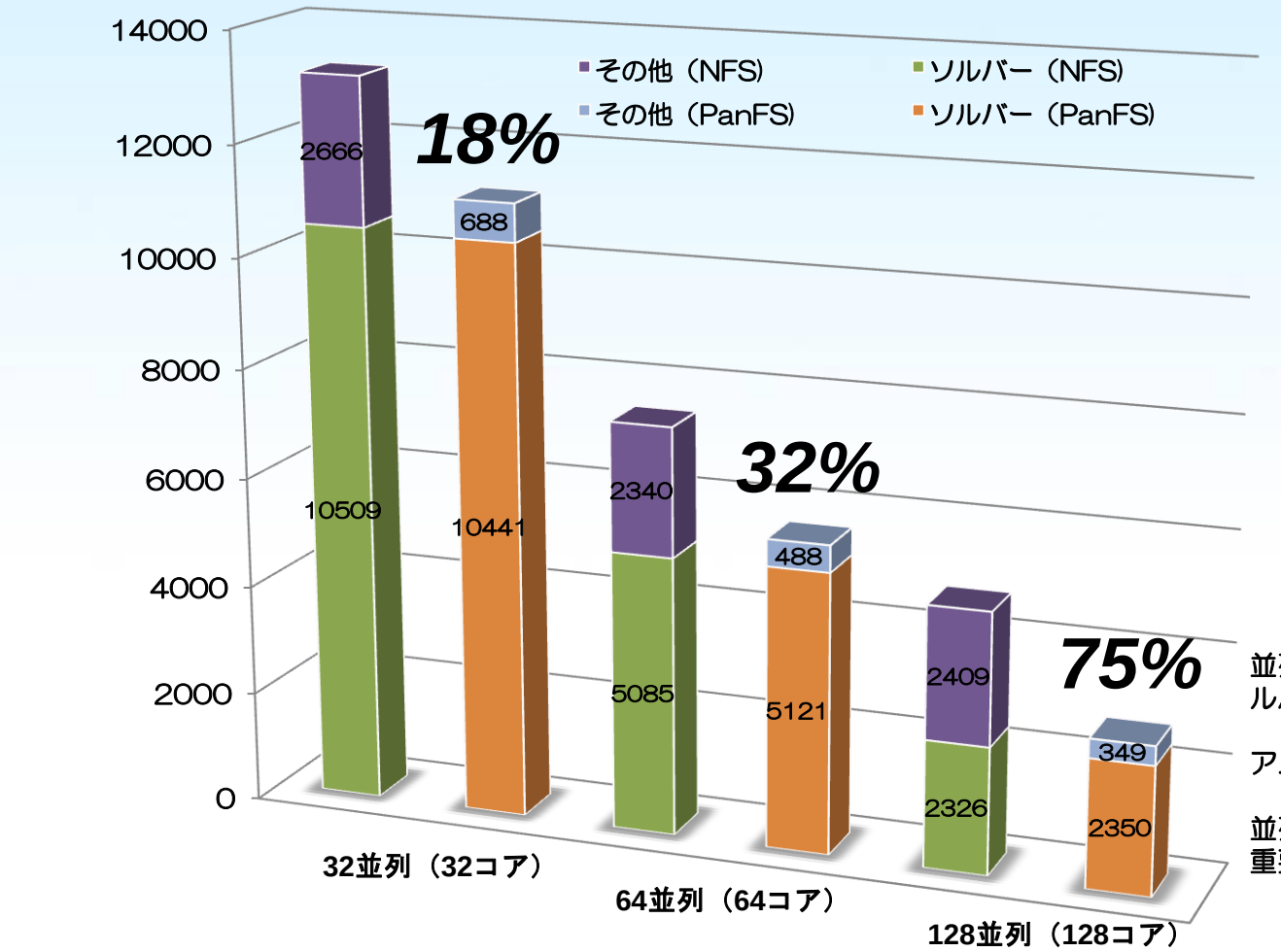


FLUENT: Serial I/O (6.2)



Source: Barb Hutchings Presentation at SC06, Nov 2006, Tampa, FL

STAR-CD v4 性能評価



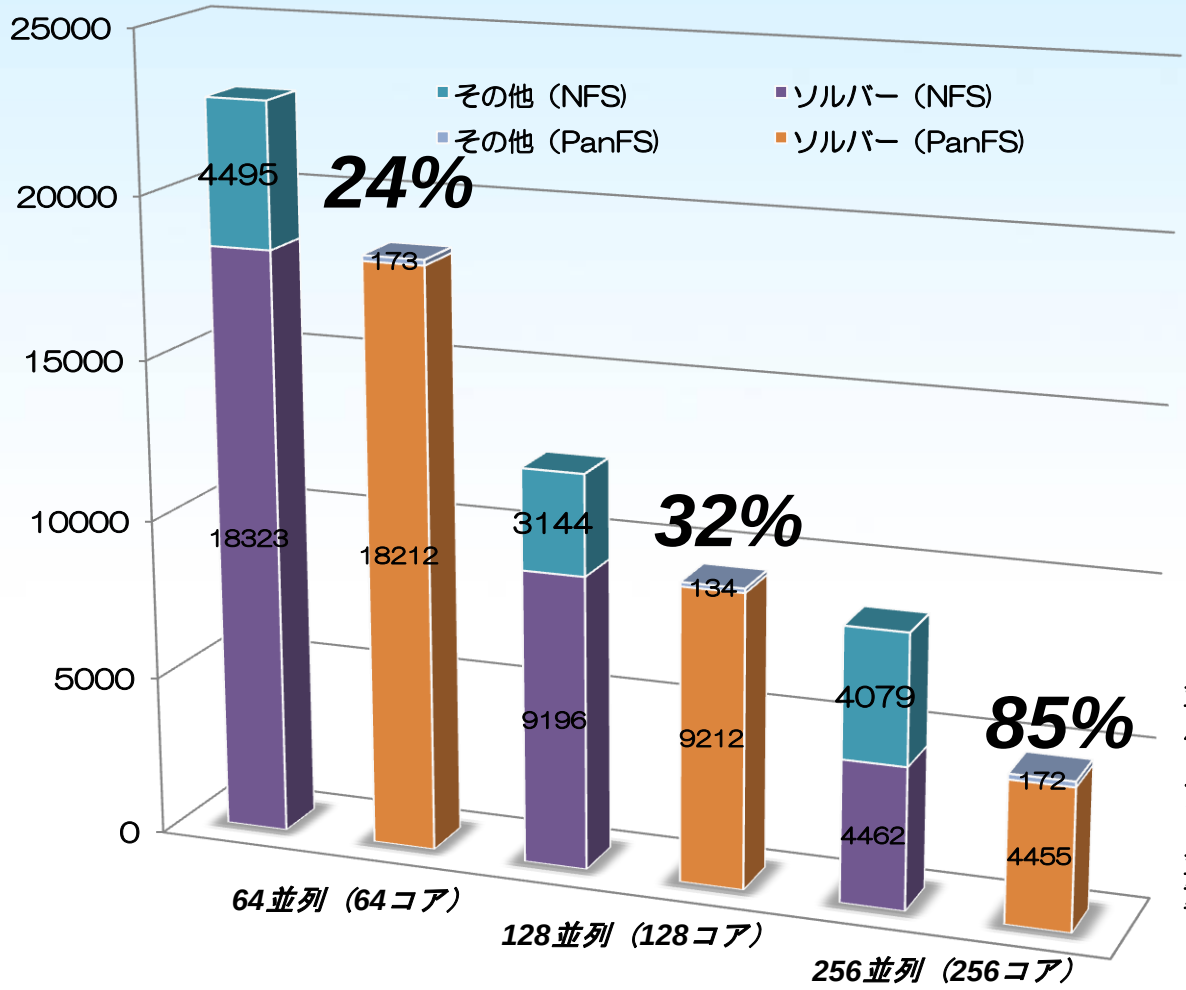
Number of cells
19,921,786
Solver
CGS, Steady
Iterations
500 total iterations - data save after every 10 iters
Each solution output (50 total)
~1,500 MB

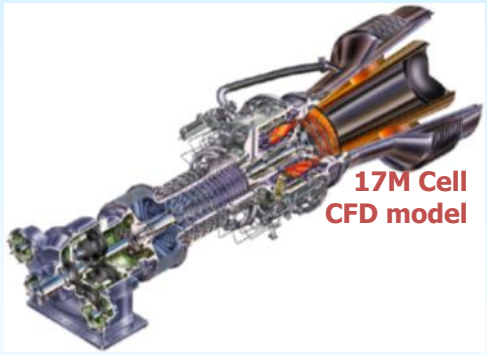
並列度（コア数）が大きくなるに伴って、非ソルバー部分の比重が大きくなる
↓
アムダールの法則（非並列部分が性能を左右）
↓
並列IO処理などによる非並列計算部分の削減が重要

この性能評価はPanasas社とインテル社が、インテル社のクラスタシステム（2048コア）を利用して計測した性能です。
File Systems -- Panasas,: 7 shelves, 35 TB storage;（各シェルフは、4xGbE接続でトータル2.8GB/sec のバンド幅）
NFS: Dell 2850 File Server, 6 x 146 GB SCSI drives, RAID 5

スケーラブルシステムズ株式会社

STAR-CD v4 性能評価





Number of cells
16,930,109
Solver
CGS, Single Precision
Iterations
300 total iterations -
data save after every 100 iters
Total solution output
~48 GB

並列度（コア数）が大きくなるに伴って、非ソルバー部分の比重が大きくなる
↓
アムダールの法則（非並列部分が性能を左右）
↓
並列IO処理などによる非並列計算部分の削減が重要

この性能評価はPanasas社とインテル社が、インテル社のクラスタシステム（2048コア）を利用して計測した性能です。
File Systems -- Panasas,: 7 shelves, 35 TB storage;（各シェルフは、4xGbE接続でトータル2.8GB/sec のバンド幅）
NFS: Dell 2850 File Server, 6 x 146 GB SCSI drives, RAID 5

スケーラブルシステムズ株式会社

FLUENT 12 における 750M Cell モデルの解析

Source: Dr. Dipankar Choudhury Technical Keynote of the European Automotive CFD Conference, 05 July 2007, Frankfurt Germany

ANSYS FLUENT

ANSYS CFD 12.0:　Core Area Advances (I)

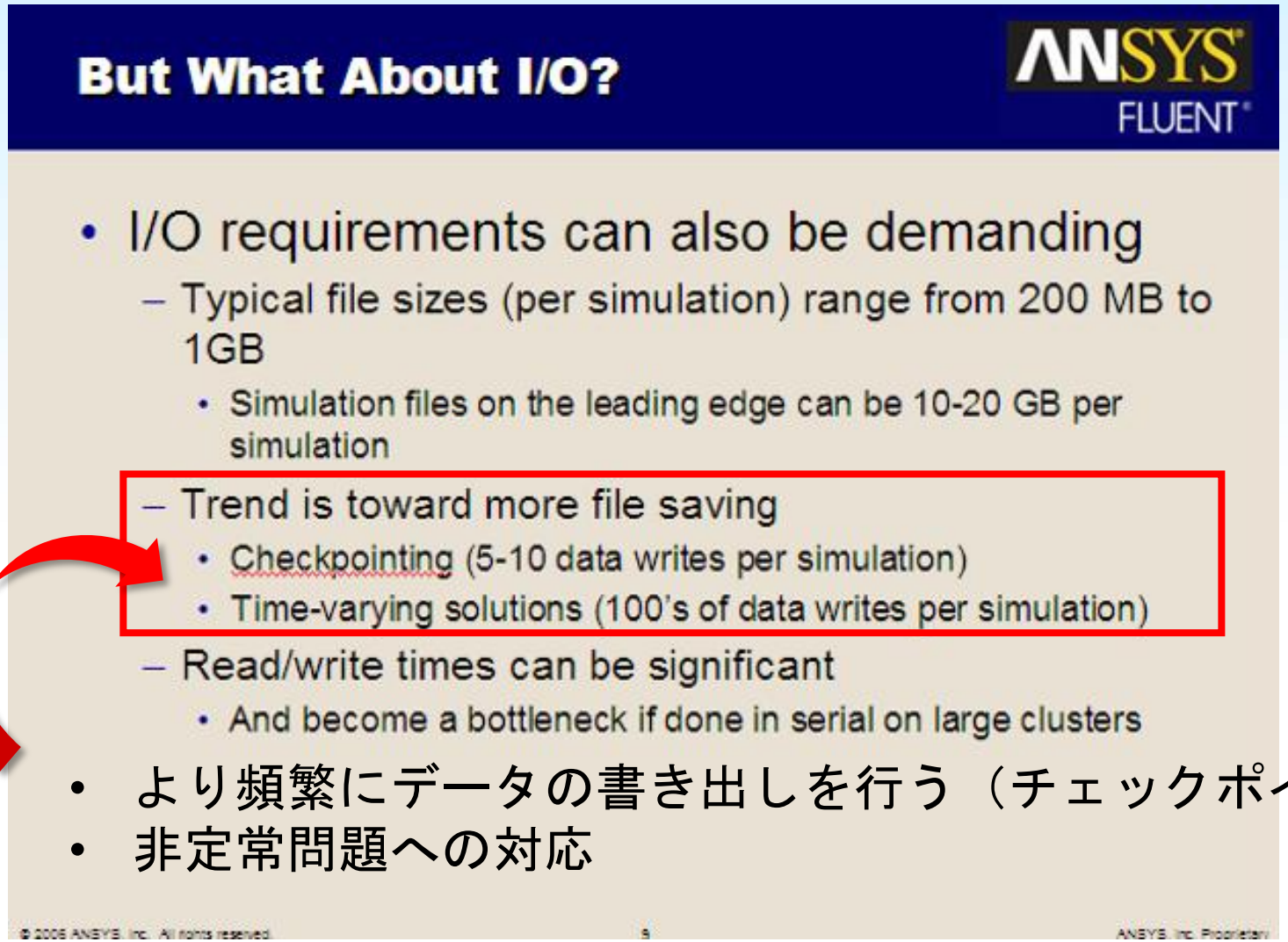
- 並列処理
  - 真の並列処理の実現
  - 逐次処理I/Oからの大幅な性能向上
  - スケーラビリティの向上

750 million cell FLUENT 12 case
(80GB pdat file)
Intel IB cluster, Panasas FS

512 cores	Serial I/O	Parallel I/O
Read	1708s	157s
Write	4255s	335s

11 x →
13 x →

CAEアプリケーションでのIO要求



- より頻繁にデータの書き出しを行う（チェックポイント）
- 非定常問題への対応

Source: Barb Hutchings Presentation at SC06, Nov 2006, Tampa, FL

FLUENT: Serial I/O (6.2) vs. Parallel I/O (6.4/12-beta)

64-way 128-way 192-way

Elapsed Time in Seconds

50000
37500
25000
12500
0

46145
30881
31625
38656
17442
11759
44479
23266
17582

クラスタ導入のメリットは？

10.5時間から3時間に短縮

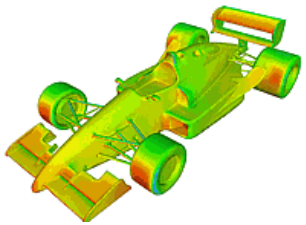
より多くのIOにも対応可能

PanFSは、Serial I/O に対して3倍以上の性能（192コア）
より多くのI/Oでもスケーラブルな性能を実現

6.2 - Serial I/O

PanFS - 1 Write

PanFS - 100 Writes



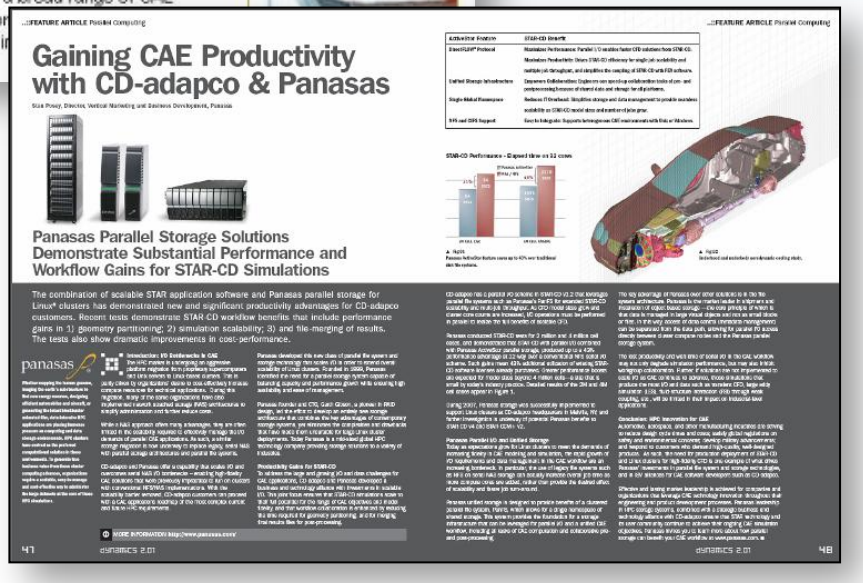
90 M Cells

ANSYS FLUENT

Panasas導入による並列処理の劇的な向上

Source: Barb Hutchings Presentation at SC06, Nov 2006, Tampa, FL

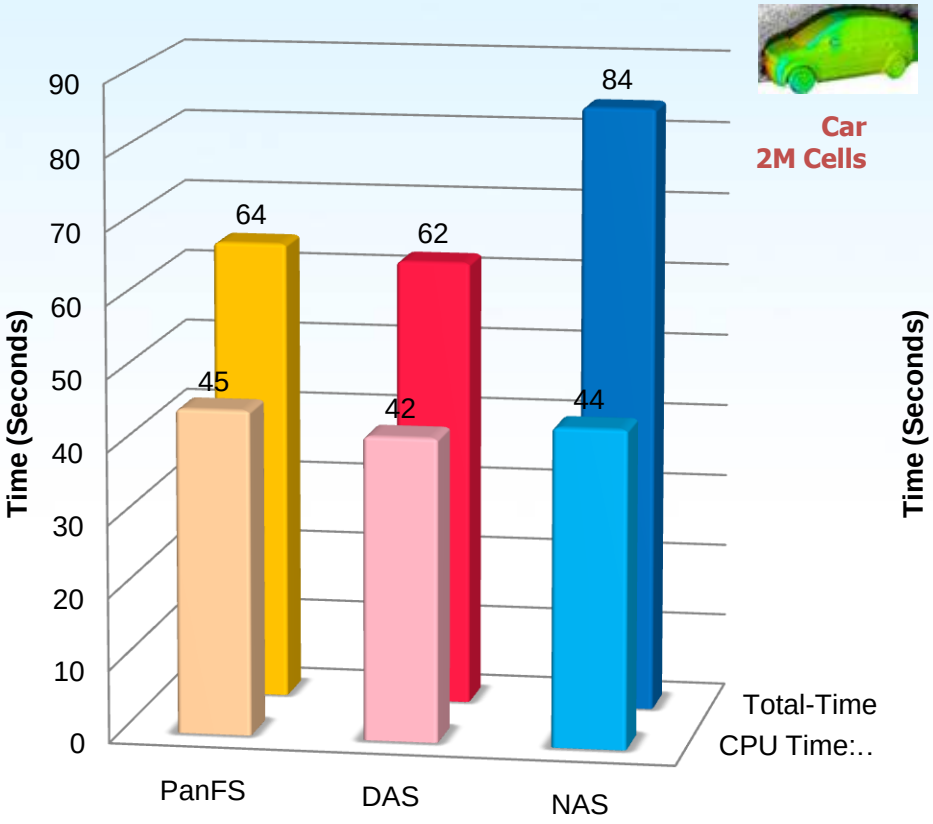
CD-adapco



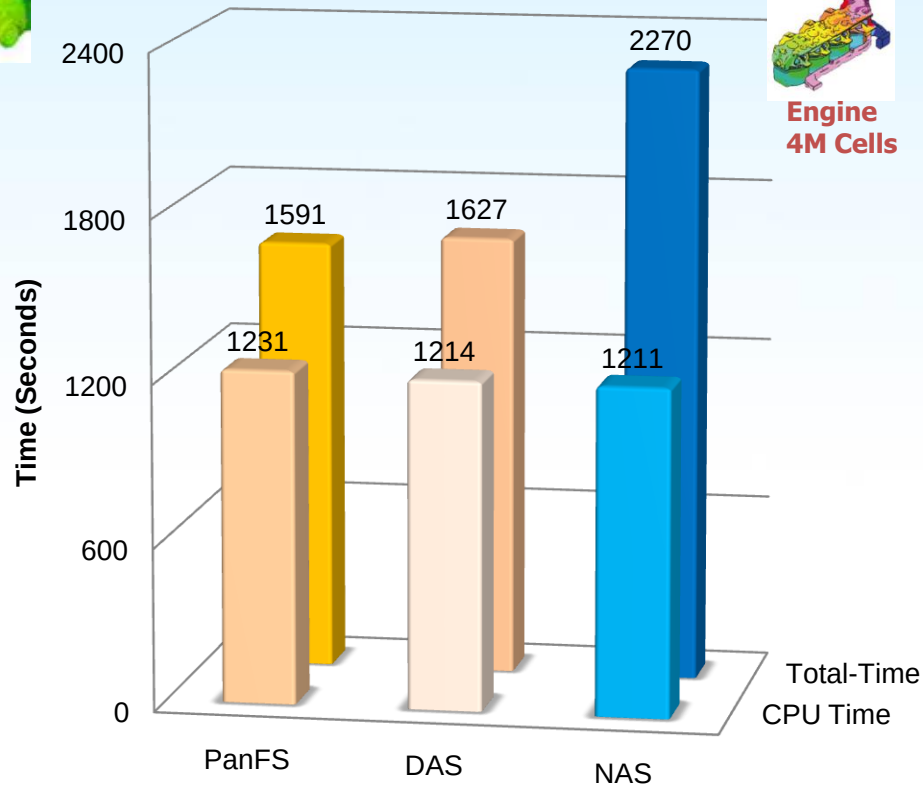
スケーラブルシステムズ株式会社

STAR-CD 非定常計算テストケース

STAR-CD 3.2.6:　Panasas　PanFS とNFSとの性能比較（32-way）



Car
2M Cells



Engine
4M Cells

Panasasのネットワークストレージは、ダイレクト・アタッチド・ストレージと同等の性能を示している
PanFSのNFSに対する優位性は、モデルサイズが大きくなるとより顕著になる

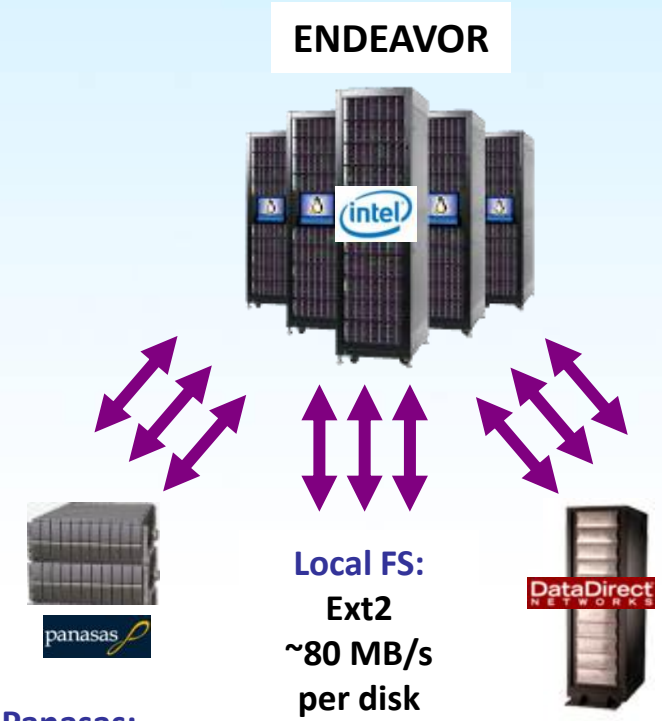
CAEソルバーとI/Oスケーラビリティ

Discipline	ISV: Software	Scalability (cores-per-job) 2008	2009	2010	Parallel I/O Status
CFD	ANSYS: FLUENT	24 – 48	32 – 64	64 – 96	Y – v12 in q3 08
	CD-adapco: STAR-CD	24 – 48	32 – 64	64 – 96	Y – v3.26, v4.2 in q2 08
	CD-adapco: STAR-CCM+	32 – 64	48 – 64	64 – 96	Y – v2.x in q4 08
	Metacomp: CFD++	32 – 64	48 – 64	64 – 128	Y – v6.x in q1 08
	Acusim: AcuSolve	32 – 64	48 – 64	64 – 128	Y – v5.x, need to test
	ANSYS: CFX	32 – 64	48 – 64	64 – 96	N – plan to follow FLUENT
CSM Explicit	LSTC: LS-DYNA	32 – 64	48 – 64	64 – 96	N – no plans announced
- Impact	ABAQUS: ABAQUS/Explicit	32 – 64	48 – 64	64 – 96	N – no plans announced
	ESI: PAM-CRASH	32 – 64	48 – 64	64 – 96	N – no plans announced
	Altair: RADIOSS	32 – 64	48 – 64	64 – 96	N – no plans announced
CSM Implicit	ANSYS: ANSYS	04 – 06	04 – 06	06 – 12	Y & N – scratch, not results
- Structures	MSC.Software: MD Nastran	04 – 06	04 – 06	04 – 08	Y & N – scratch, not results
	ABAQUS: ABAQUS/Standard	04 – 06	06 – 12	08 – 24	Y & N – scratch, not results

Panasasとインテル社による Abaqus S4b での性能検証

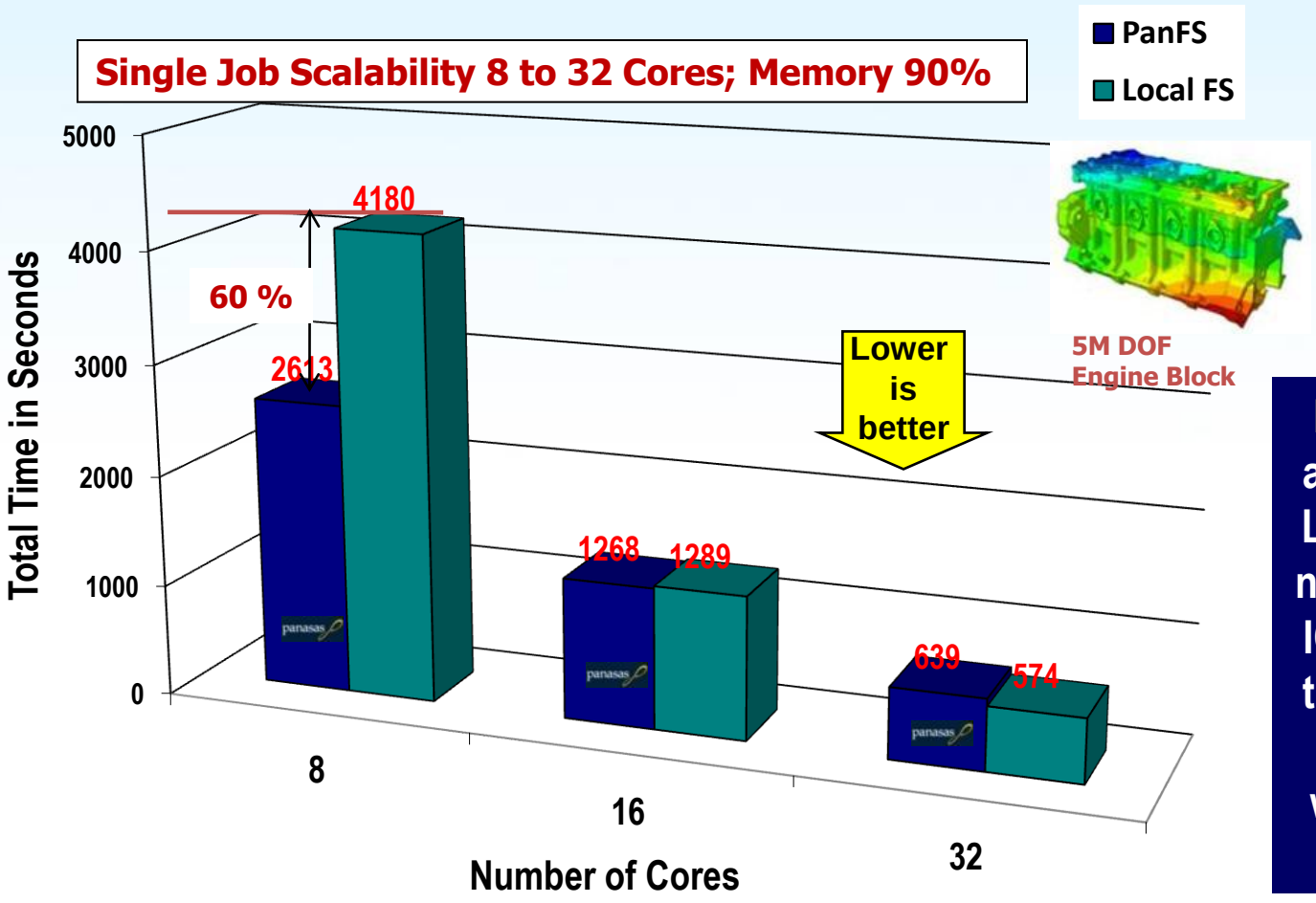
Intel ENDEAVOR Xeon Cluster
Location: Intel HPC Customer Enabling Center, Dupont ,WA
Vendor: Intel; 80 nodes; 640 c ores; 18 GB memory per node
CPU: Intel Xeon (Nehalem) QC, 2.8 GHz, 8 cores per node
Interconnect: Infiniband
File Systems: Panasas PanFS; Lustre on DDN; Local disk
Operating System: **RHEL** Linux v5.2

- ENDEAVOR File Systems and Storage
  - PanFS: 2 Shelves AS6000 (1+10 and 2+9), 38 TB FS; network connected through 10GigE switches and IB router, ~ 1.2 GB/s
  - Lustre: DDN storage, 100 TB FS, ~ 5 GB/s
  - Local FS: Ext2 FS, 370 GB SATA drive, 80 MB/s per disk



ABAQUS標準ベンチマーク （S4b）シングルジョブ性能

Abaqus/Standard 6.8-3: Comparison of PanFS vs. Local FS Ext2

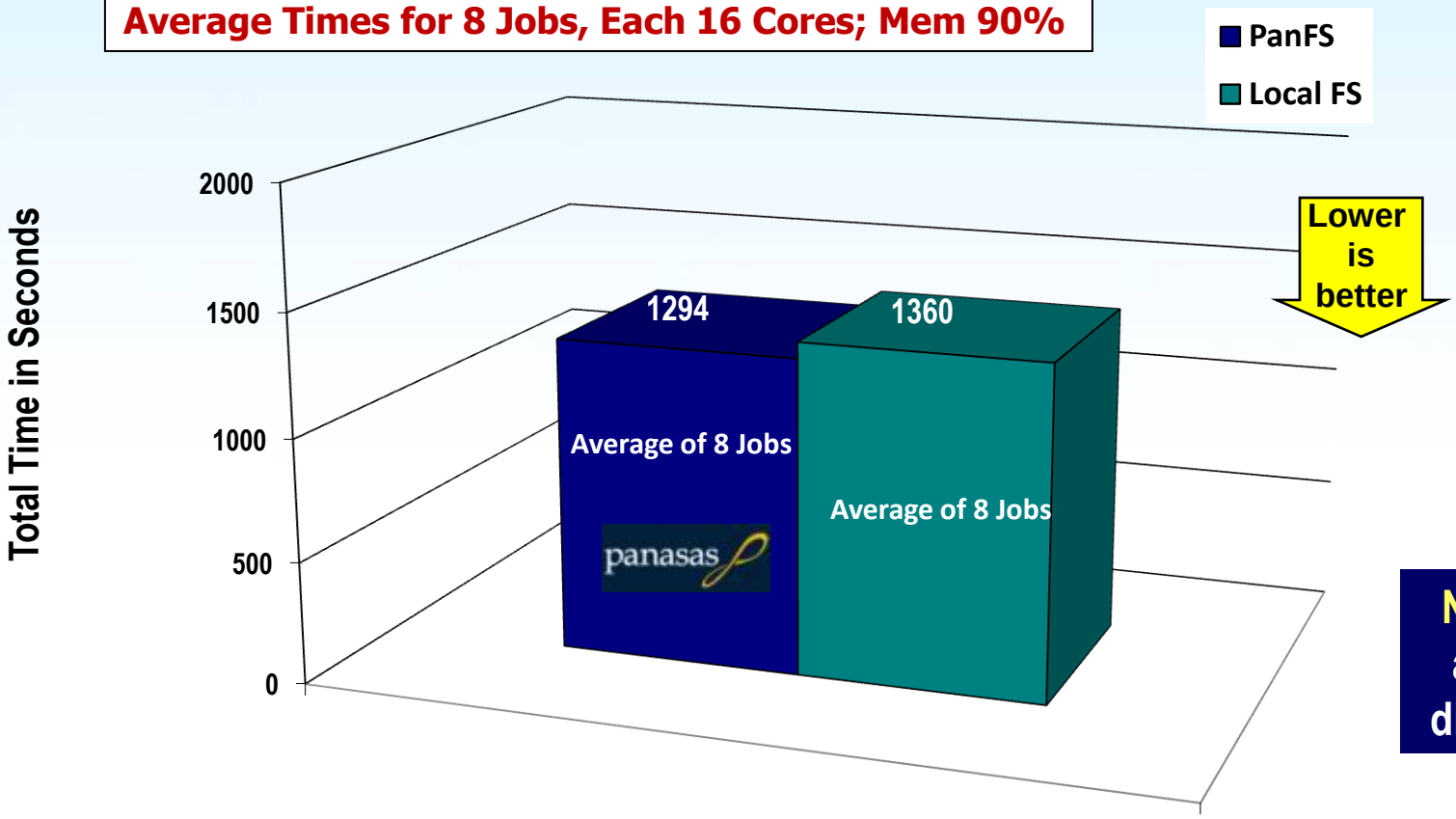


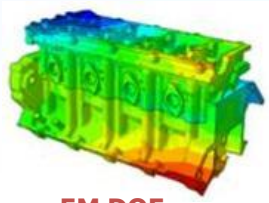
NOTE: PanFS advantage over Local for single node case when IO is heavy – in the same range for 2-4 nodes when job goes in-memory

ABAQUS標準ベンチマーク （S4b） マルチジョブ性能

Abaqus/Standard 6.8-3: Comparison of PanFS vs. Local FS Ext2

Average Times for 8 Jobs, Each 16 Cores; Mem 90%





5M DOF Engine Block

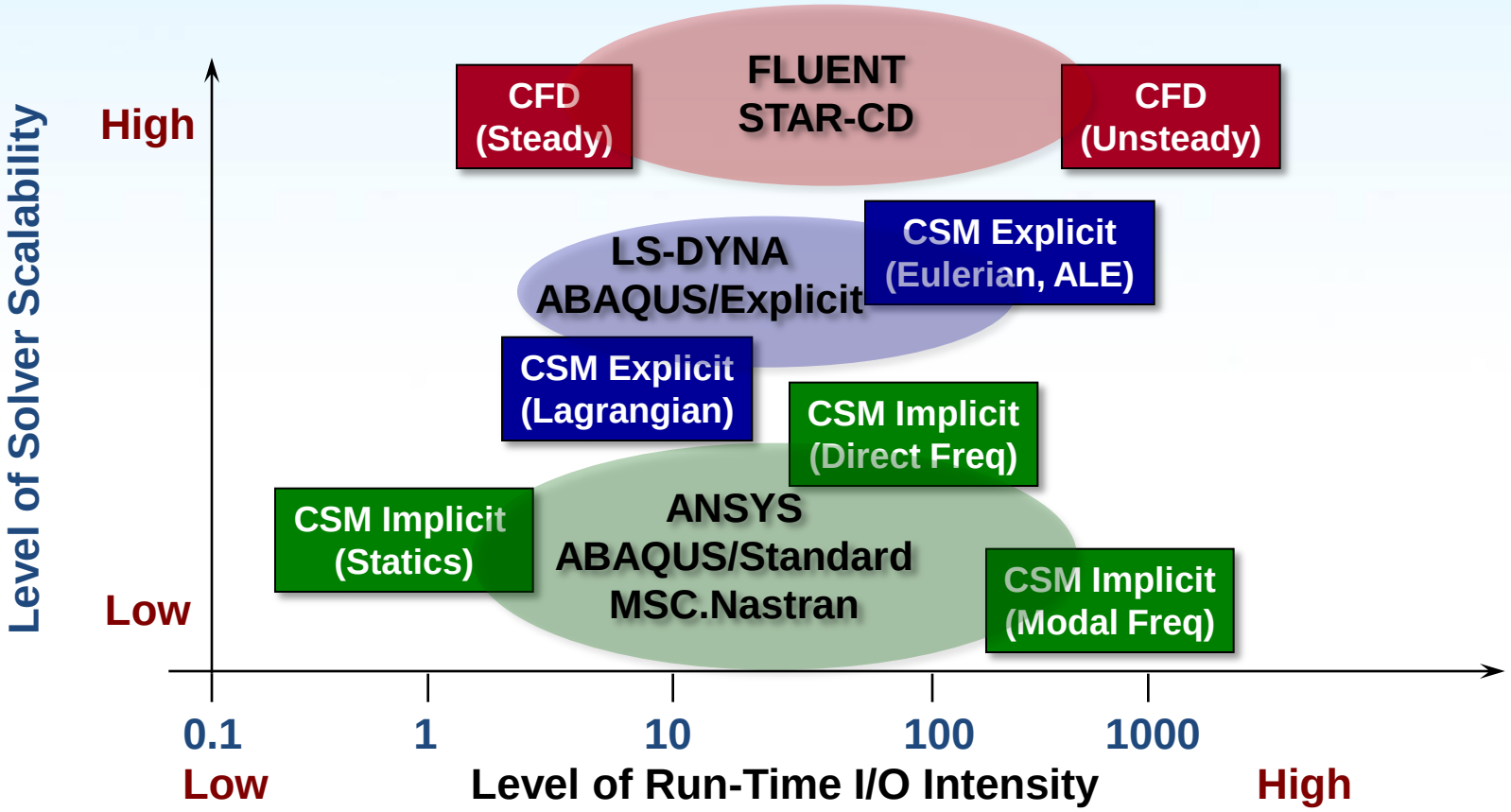
NOTE: PanFS and Local FS difference ~ 5%

8 Jobs x 16 Nodes x 128 Cores

Average Times for 8 Jobs | Each Job on 2 Nodes | Each Job on 16 Cores | Total 128 Cores

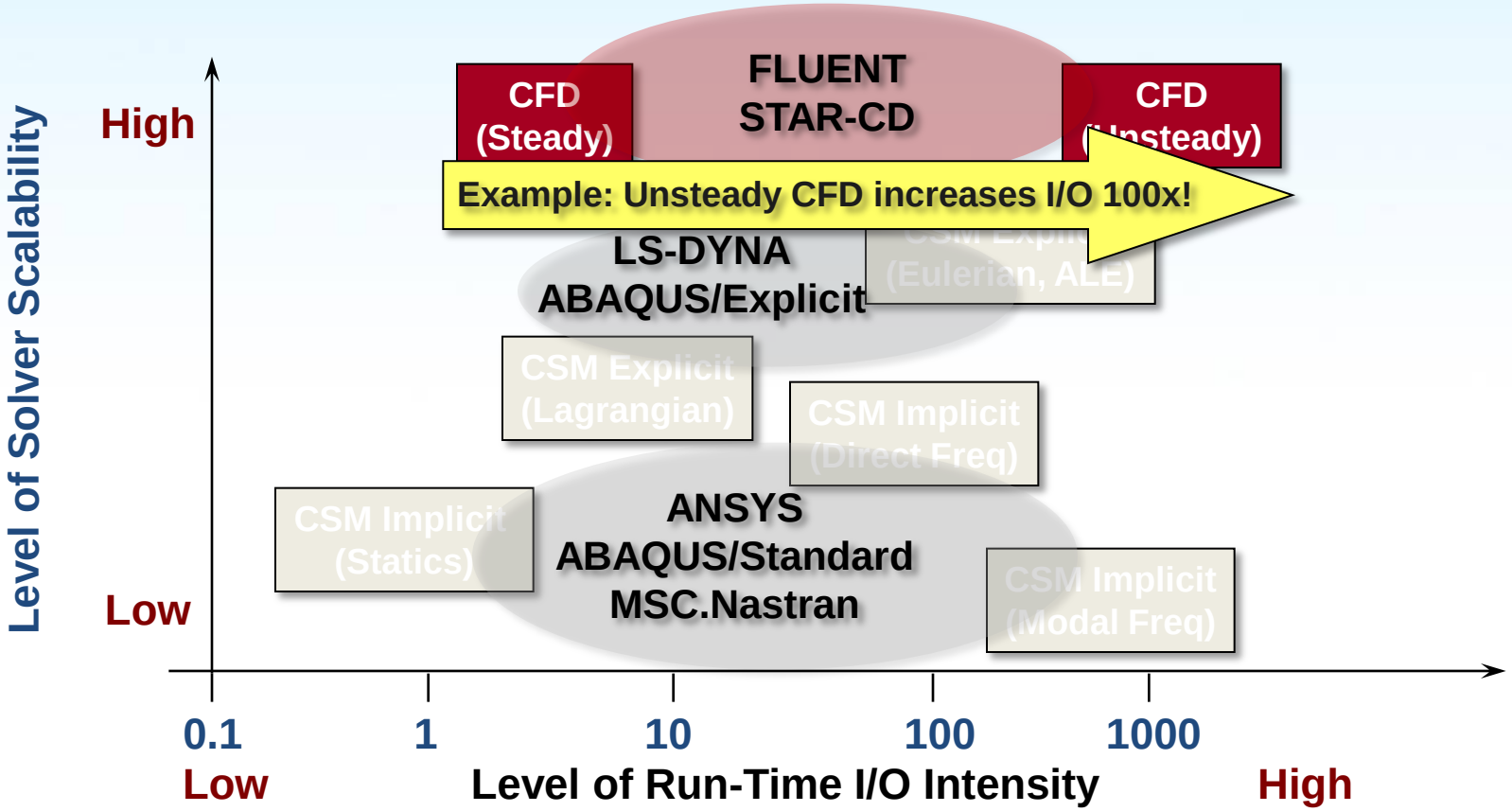
CAEアプリケーションの実行特性分析

CAEセグメントでのシングルジョブの実行特性（スケーラビリティとIO）

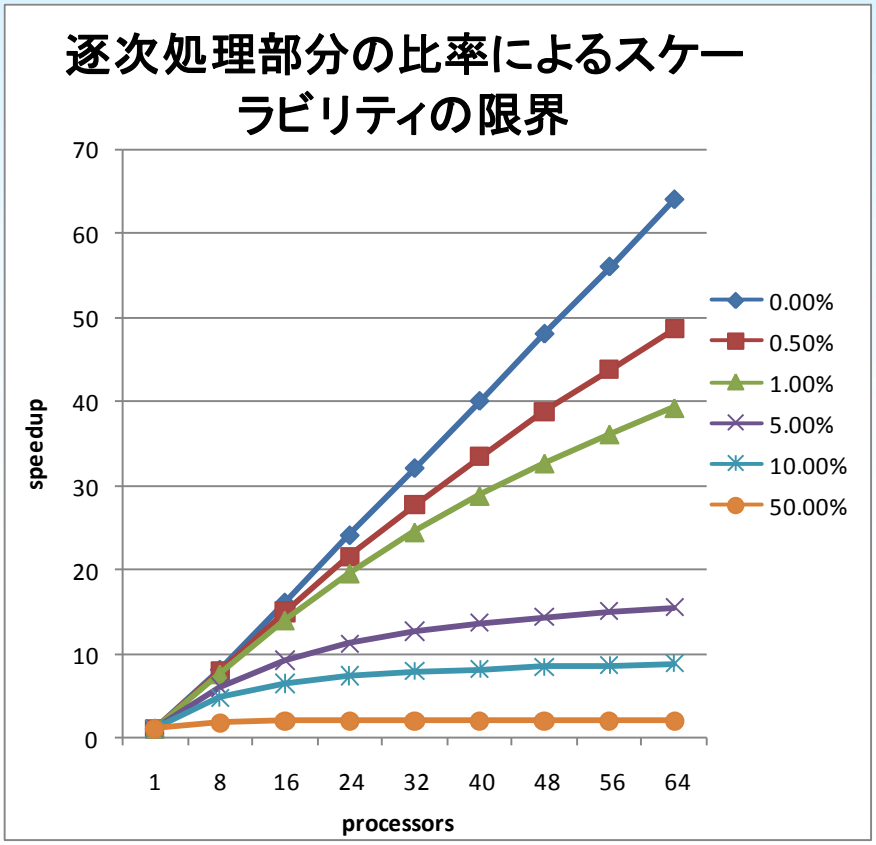


CAEアプリケーションの実行特性分析

CAEセグメントでのシングルジョブの実行特性（スケーラビリティとIO）



アムダールの法則



- 実行時間 = 逐次処理 + 並列処理

理論的な性能向上の限界

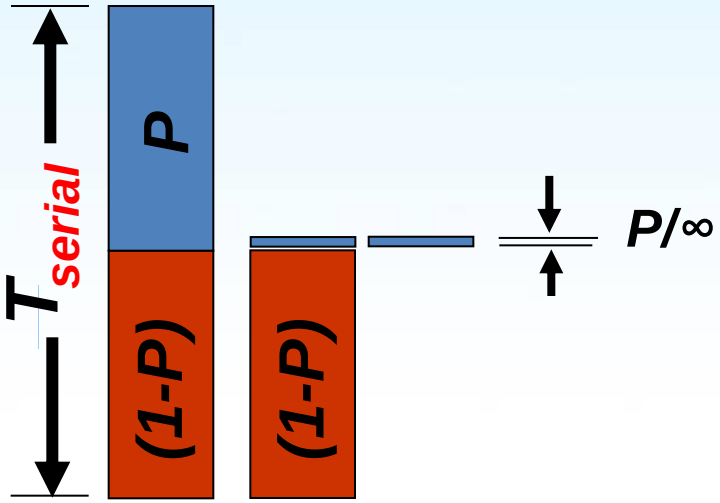
- 実行時間 = 逐次処理 + 並列処理時/P

- 64プロセッサで50倍の性能向上を得るには、逐次処理部分を0.5% 以下にする必要がある

I/O処理は逐次処理の典型であり、I/O処理自身を並列に処理することが高いスケーラビリティの実現のためには必須である

アムダールの法則

並列処理での性能向上の上限値（スケーリング）



$$T_{parallel} = \{(1-P) + P/n\}\, T_{serial} + O$$

n = number of processors

$$Scaling = T_{serial} / T_{parallel}$$

プログラムの逐次処理部分（非並列処理）部分の排除が必要

例えば、n=∞, P = 0.5 の場合

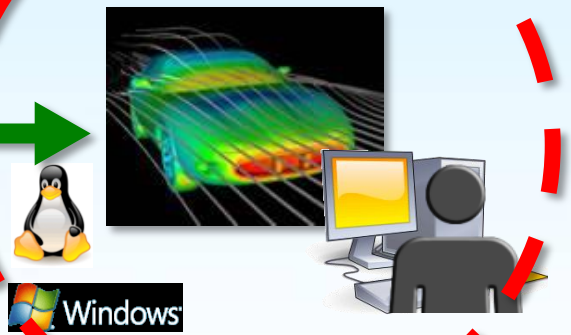
$$Scaling = 1.0 / (0.5 + 0) = 2.0$$

ボトルネックの解消
①スケーラブルなI/O処理の実現

プリ処理
メッシュ作成

シュミレーション (Solve)
計算処理

ポスト処理
可視化



MB's . . . GB's MB's

Panasas
ストレージクラスタ

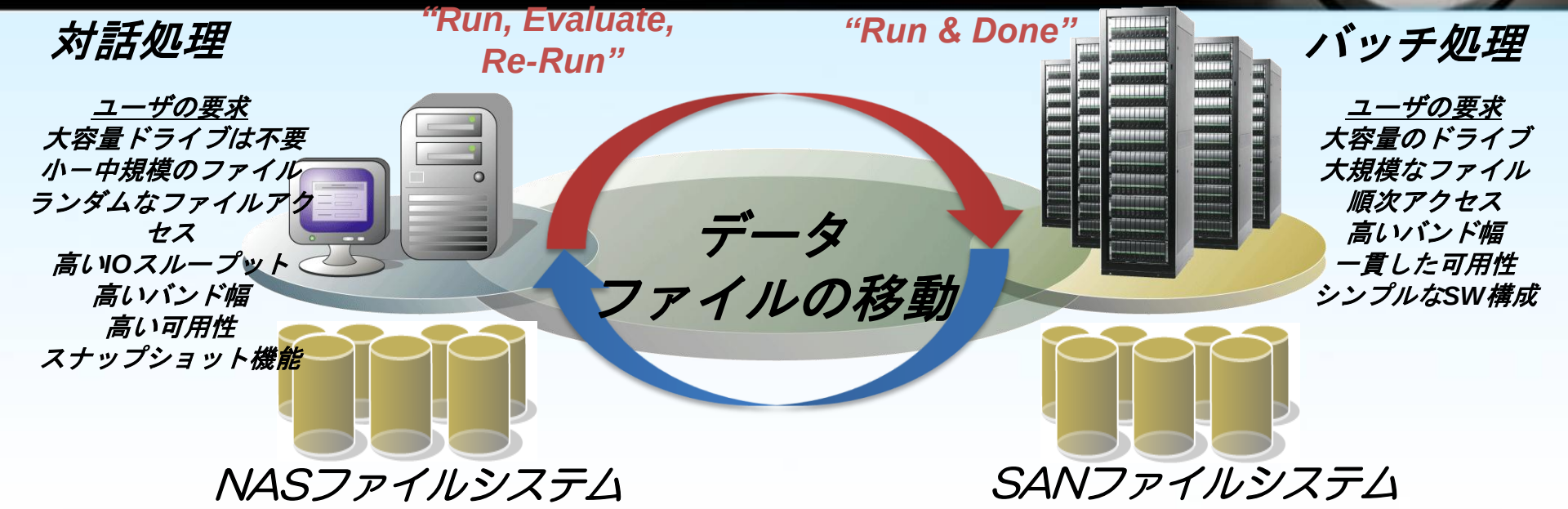
パラレルファイルシステム
ワークフロー統合ストレージ
スケーラブルに容量と性能を拡張可能

MB's

データ管理システム

スケーラブルシステムズ株式会社

ストレージに対する要求

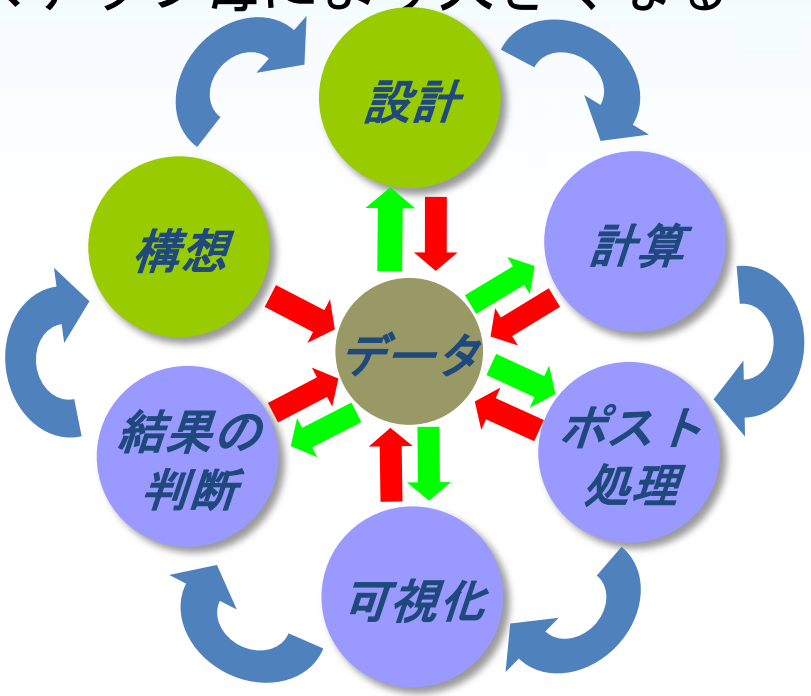


対話処理	バッチ処理
地下資源探査結果の評価	地質探査解析（大規模データ処理）
EDA デザイン	チップシュミレーション&Tapeout
モデル化、解析結果評価	空力解析、衝突解析
アニメーション処理	レンダリング
トレーディング/ポートフォリオ	リスクマネージメント

理想的なデータの流れの提案

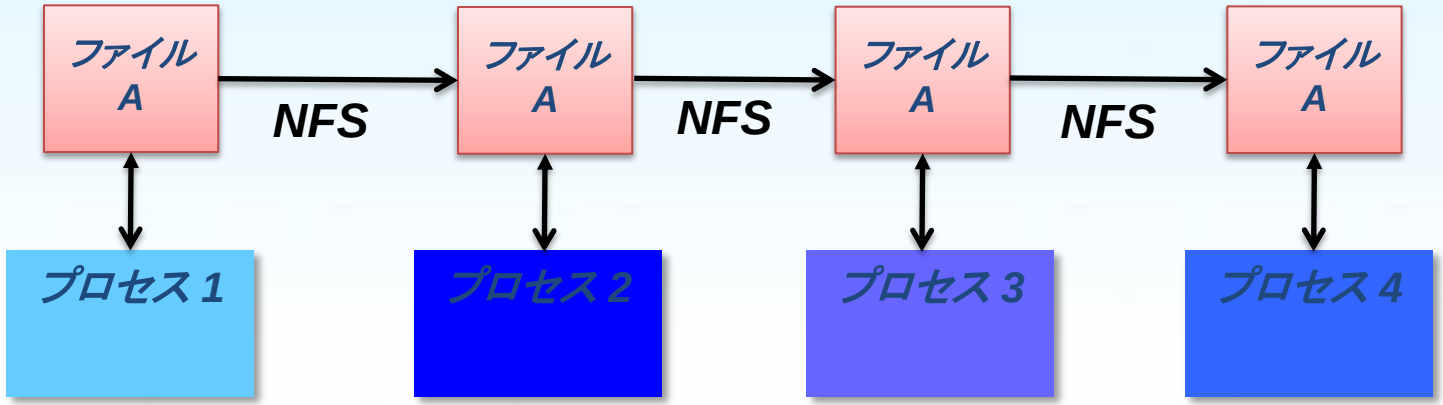
- 一般的な業務の流れを考えてみると..
  - コンセプトの段階から、最終的な製品化の段階まで、データは、業務の流れの中心に位置する
  - 情報は、多くのグループで共有され、データは、ホストコンピュータ間で移動する
  - データセットのサイズは、各ステップ毎により大きくなる

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

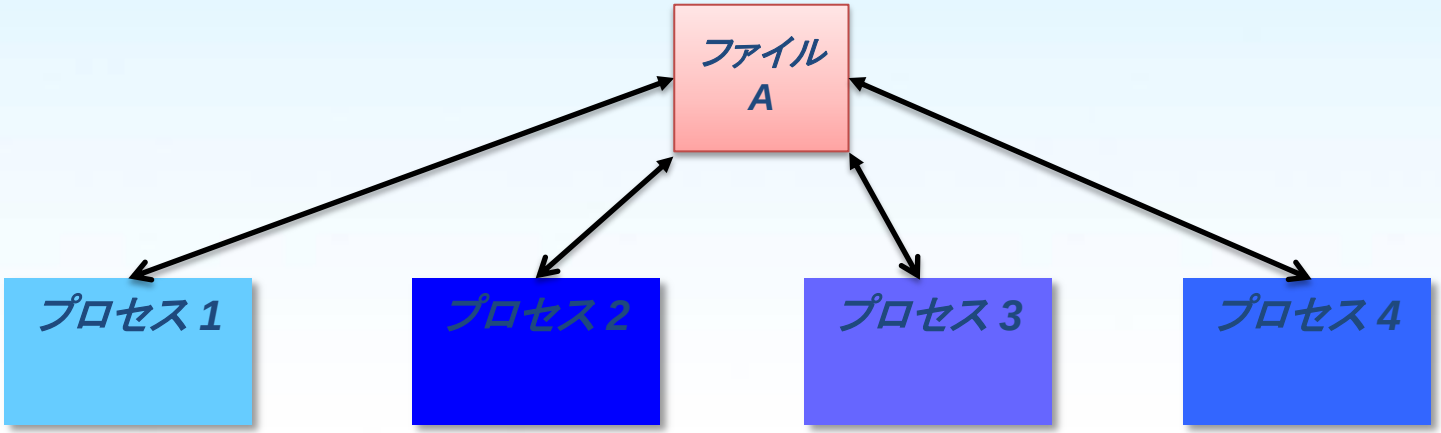
ファイル共有が無い場合には、ネットワーク介してデータの移動が必要
データの保管場所や移動作業などのオーバヘッド
NFSサーバのボトルネックがワークフロー全体に影響する

Panasasのストレージクラスタ

- SANの高い拡張性とパフォーマンスとネットワーク・アタッチド・ストレージ（NAS）の接続性・共有性の双方を提供することが可能です。
  - 従来SANとNASを利用してきた、「対話処理」と「バッチ処理」の双方を一つストレージシステムで最適に実行可能となり、大規模な情報へのスムーズなアクセスを可能にしています。
- ワークフローで必要となるファイルについて、そのファイルを必要とする処理プロセスにおいての共有を可能にし、ファイルの複製や大きなファイルの移動といったデータ処理を解消します。
  - システム・ワークフローの改善と運用コストの削減を実現します。
  - 全てのシステムから全てのデータへ直接アクセスすることができるようになり、システムはPanasasストレージクラスタが提供する高いIOバンド幅を有効活用し、直接データの読み書きを行うことが可能になります。
- ネットワークの混雑やファイルサーバのオーバーロードによるボトルネックに悩まされる必要はなります。また、同時に複数のストレージシステム間で不用なデータのコピーが重複したデータの保存などのオーバヘッドも解消します。

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

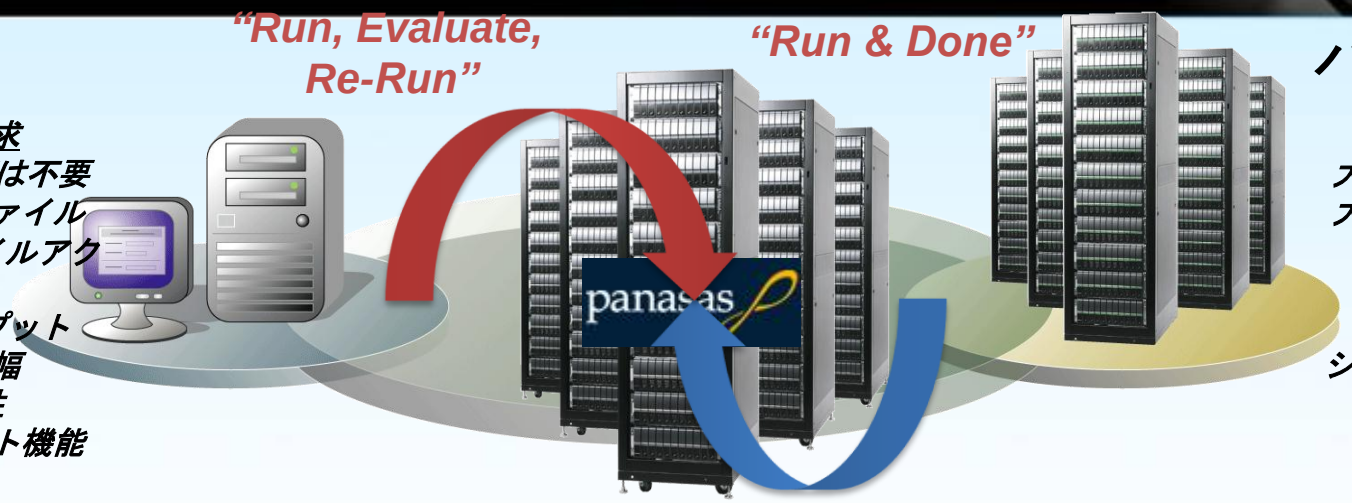
ワークフロー統合ストレージ

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"

"Run & Done"



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

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

Panasas ActiveStorの特徴

機能とその利点	Panasas ActiveStor	NAS	SAN
ターゲットとするアプリケーション	バッチ処理 +対話処理	対話処理	バッチ処理
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	◎	◎	

プリ・ポストと計算処理を統合

ポスト処理での利点
Collaboration

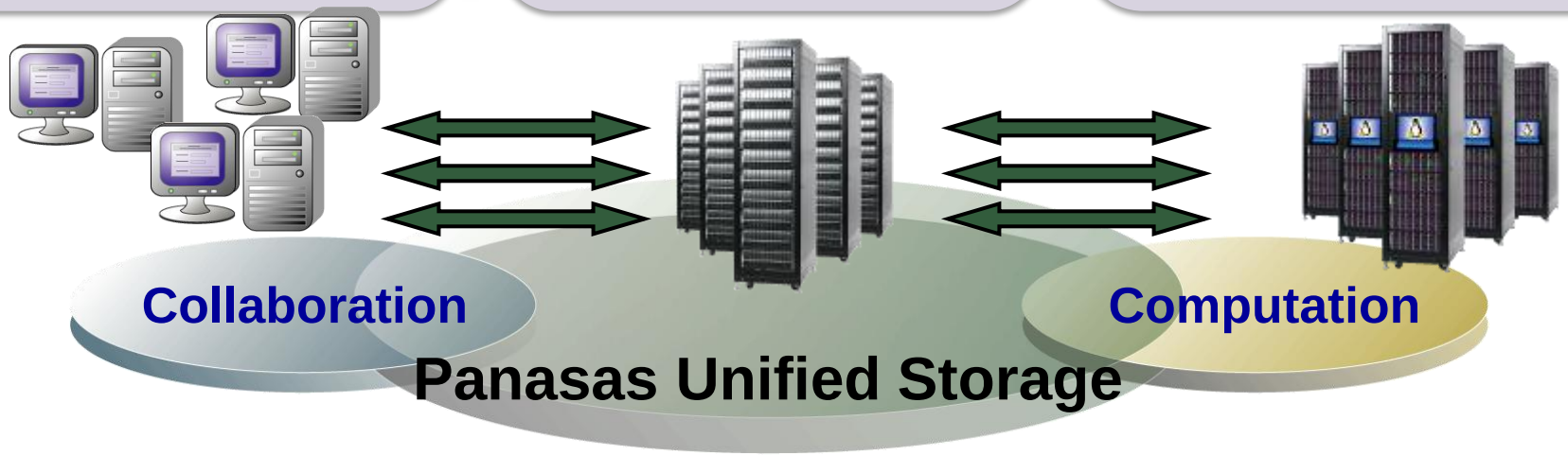
- 計算結果の不要なコピーの排除
- ポスト処理の高速化とコラボレーションの改善

Panasasでの更なる利点

- 複数ジョブの同時実行
- より多くの処理を効率的に処理可能

計算処理での利点
Computation

- シングルジョブの処理時間の短縮
- ワークロードの改善とアプリケーション費用の低減

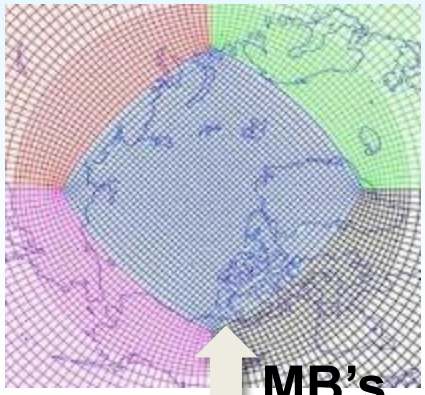


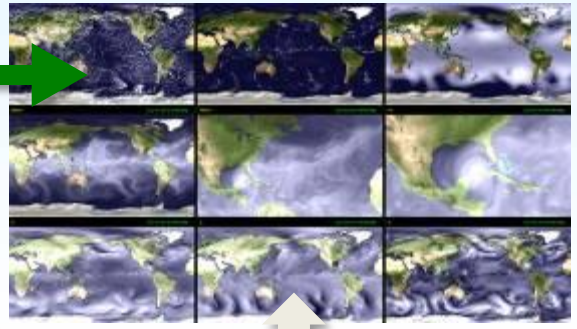
大規模モデルでの
プリ・ポスト処理&解析処理

プリ処理
メッシュ作成

シュミレーション (Solve)
計算処理

ポスト処理
可視化





MB's

. . .

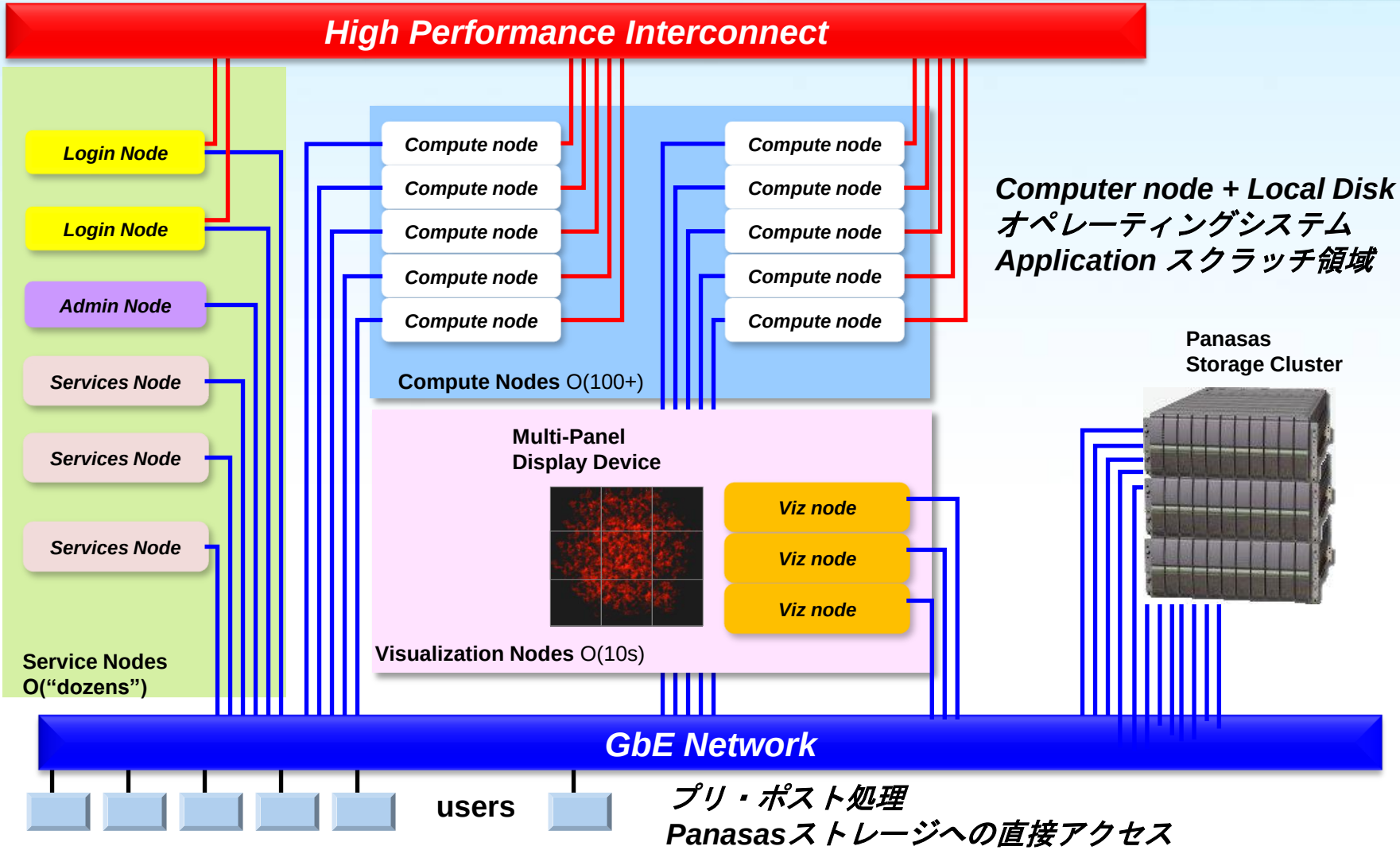
GB's-TB's

GB's-TB's

Panasas
ストレージクラスタ

- パラレルファイルシステム
- Unified (共有)ストレージ
- スケーラブルに容量と性能を拡張可能

可視化サーバ（マルチパネル）



可視化とI/O要求の双方に対応可能

HPC wire

Panasas and Rocks Visualization Cluster

Collaborators: SU – Panasas – Dell – Cisco – UCSD/Clustercorp

The Leading Source for Global News and Information Covering the Ecosystem of High Productivity Computing / November 15, 2006

Location: HPCC at Stanford, managed by Dr. Steve Jones in support of Flow Physics and Computational Engineering Group

STANFORD UNIVERSITY

clustercorp.com

The Rocks Rolls used for the project:

Visualization:

Viz Roll (from EVL and the Rocks Cluster Group)
EnSight DR Roll (from CEI, Roll by Clustercorp)
ParaView Roll (from ParaView.org, Roll by Clustercorp)

Storage:

Panasas Roll (from Panasas, Roll by Clustercorp)

Networking:

Topspin IB Roll (from Cisco, Roll by Clustercorp)

General:

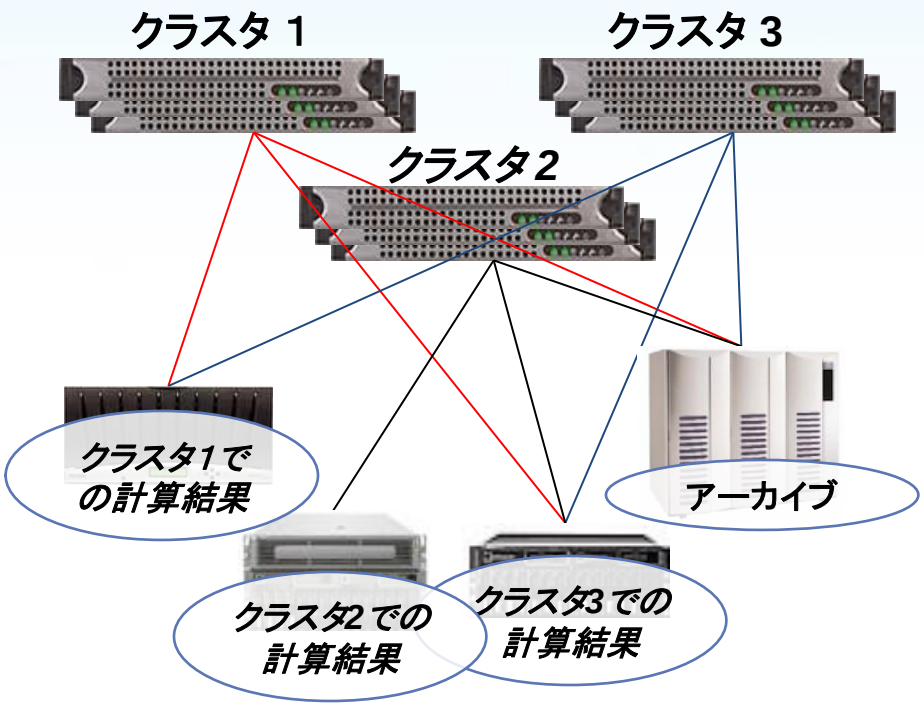
Kernel, Base, HPC, OS, Web-Server, Ganglia, Java and Service Pack (from the Rocks Cluster Group)
PBS Roll (from the University of Tromso)

Full Article: http://www.hpcwire.com/hpc/1098852.html

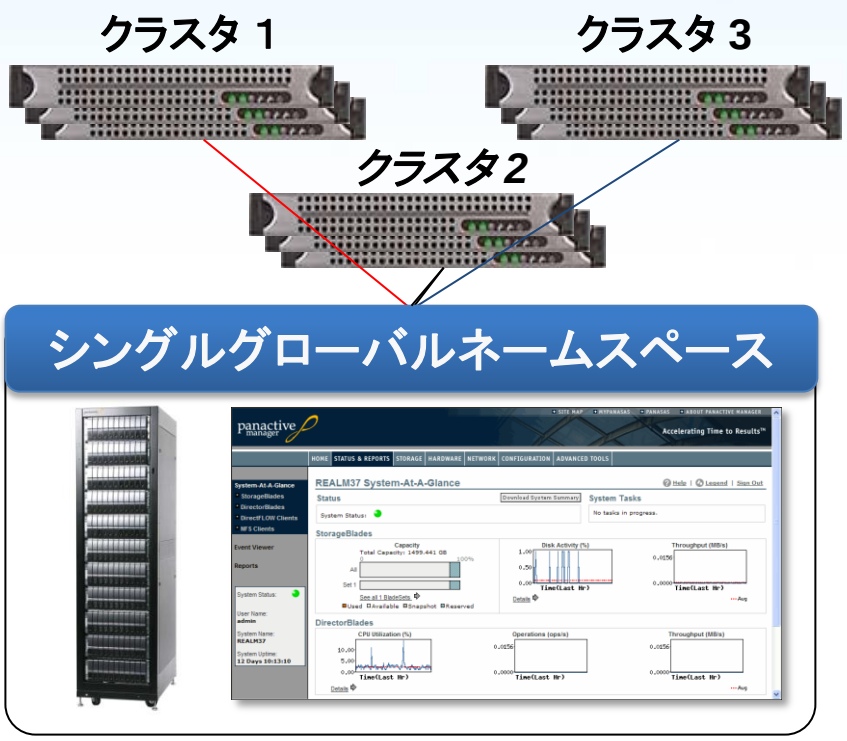
シングルグローバルネームスペース

- 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

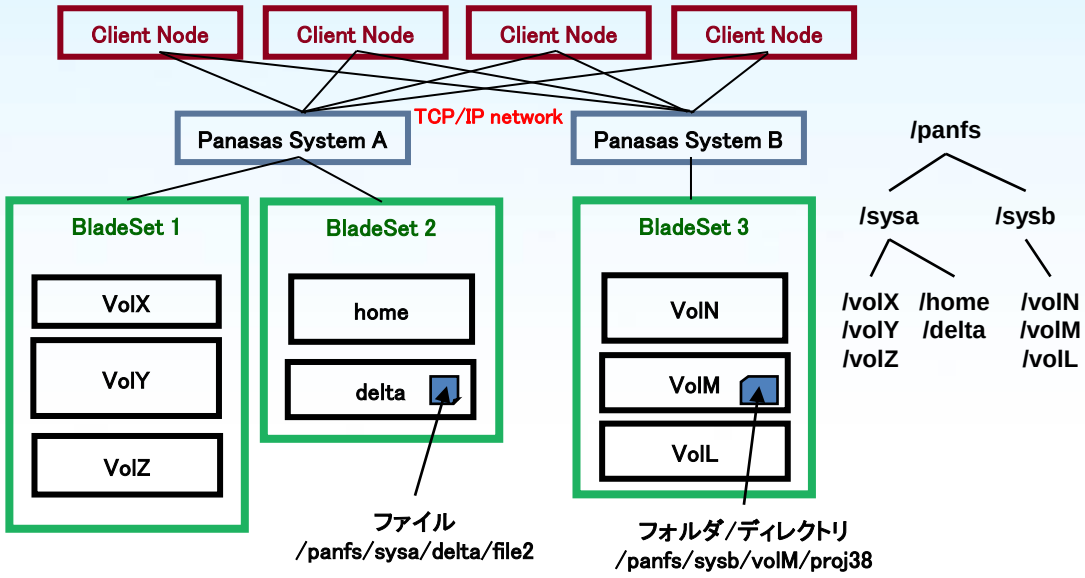


Panasasストレージクラスタ



シングルグローバルネームスペース

すべてのクライアントから同じパス名でファイル/panfs/sysa/delta/file2、フォルダ/ディレクトリ/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、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スが可能

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

まとめ：CAEワークロードでの利点

ワークフロー統合ストレージ:CAEワークフローでのコラボレーション

- CAEシュミレーションとプリ・ポスト処理のデータ共有の効率化
- 複数プロトコルサポートによるプラットフォーム非依存での共有データへのアクセス

パラレルI/O:高いI/O性能とボトルネックの解消

- CAEシュミレーションでの生産性の向上
- 高いシングルジョブ性能(スケーラビリティ）
- 複数ジョブでのスループット

NFSとCIFSサポート:システムインテグレーション

- 異機種混在のCAE環境における複数プロトコルのサポートによる容易なシステム導入と運用

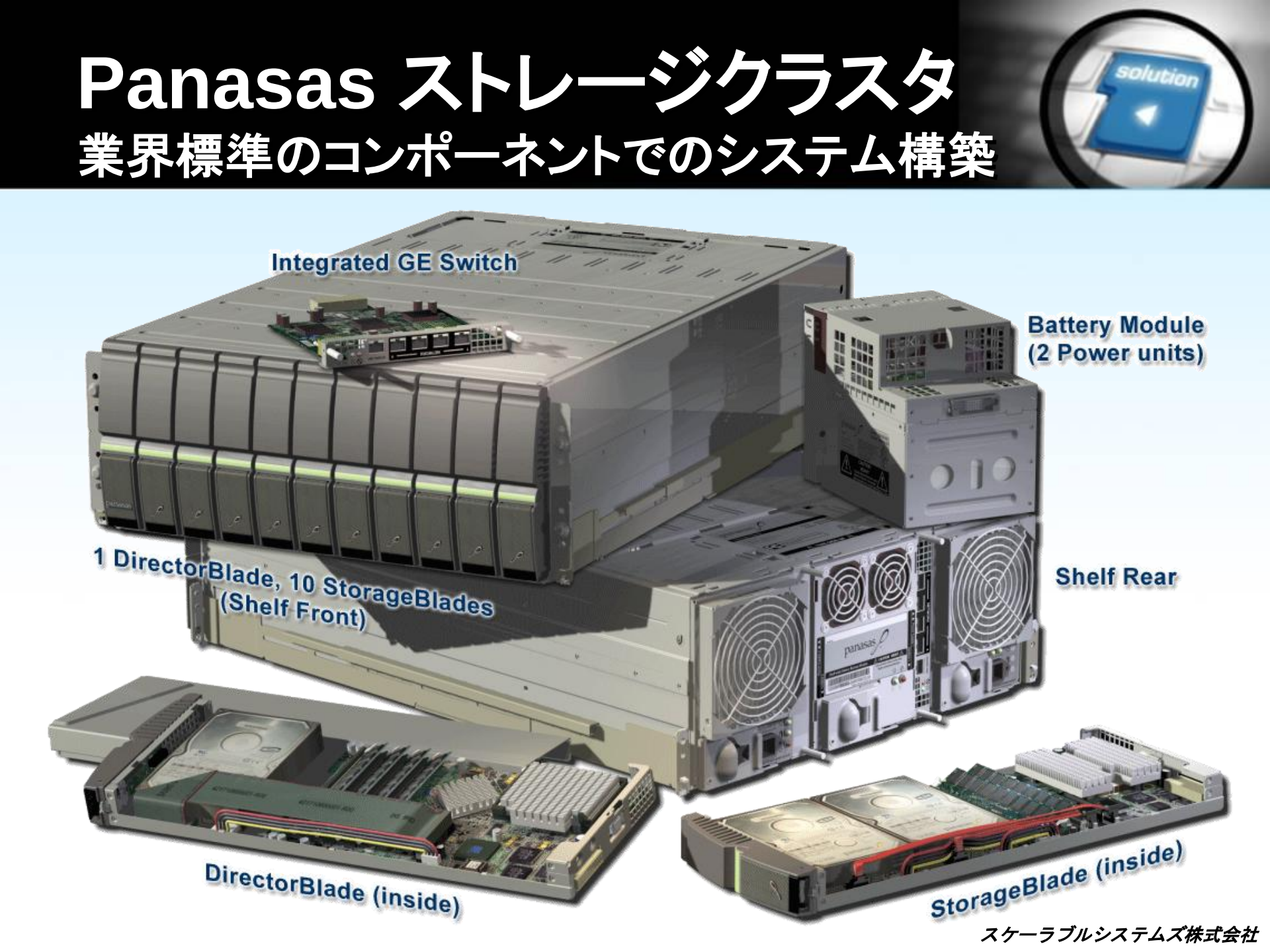
シングルネームスペース:ITマネージメントのオーバヘッドの低減

- ストレージマネージメントとデータ管理をシンプルに実行可能な運用管理機能と増設時の容易なオペレーション

Panasas ストレージクラスタ

製品ライン

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

Panasas ActiveStor 仕様一覧

	Series 7	Series 8	Series 9	AS200 [4]	AS4000	AS6000
最大バンド幅 MB/sec	350MB/sec	600MB/sec	600MB/sec	350MB/sec	600MB/sec	600MB/sec
最大IOPS[1]	12K	12K	21K		12K	12K
物理ストレージ容量	16TB-44TB	16TB-44TB	8TB-10TB	104TB	8TB-22TB	8TB-22TB
ストレージ キャッシュ容量	5GB-22GB	16GB- 44GB	32GB- 40GB	25GB	4GB-5.5GB	16GB-22GB
平均レスポンス時間	7-12msec	5-9msec	0.5-4msec	7-12msec	5-9msec	5-9msec
接続オプション、データマネージメント、高可用性オプション						
シングル ネットワーク スイッチ	1 GbE x 4	1 GbE x 4 10 GbE x 1 (Copper 又は Optical)	1 GbE x 4 10 GbE x 1 (Copper 又は Optical)	1 GbE x 5	1 GbE x 4 10 GbE x 1 (Copper 又は Optical)	1 GbE x 4 10 GbE x 1 (Copper 又は Optical)
追加ネットワーク スイッチ	オプション	オプション	オプション	N/A	オプション	オプション
InfiniBand接続	N/A	オプション	オプション	N/A	オプション	オプション
スナップショット [2]	オプション	標準サポート	標準サポート	N/A	オプション	標準サポート
可用性オプション [3]	オプション	標準サポート	標準サポート	N/A	オプション	標準サポート

1) 性能データはネットワーク構成やクライアント構成に依存します。
2) Panasas Activelmage Snapshotはボリュームスナップショットの機能を提供します。
3) Panasas ActiveStor Network and Volume Failoverによるボリュームとネットワークのフェイルオーバーの提供を行います。
4) AS200には、一年間のHW/SWサービスとNDMPバックアップオプションが標準で付随します。

スケーラブルシステムズ株式会社

Panasas ActiveStor SERIES 製品ライン

性能/スケーラビリティ

SERIES 7

全機能搭載
エントリ
システム

・優れた性能
・低コスト
・GbE接続

SERIES 8

高性能
エンタープライズ
機能

・バンド幅の大幅な向上
・シングルクライアント
での性能向上
・10GbEとInfiniBand
接続

SERIES 9

テクノロジー
リーダシップ
大幅な性能向上

・IOPS性能の大幅な向上
・仮想階層ストレージ
・10GbEとInfiniBand
接続
・高い可用性

可用性/運用・管理機能

Panasas ActiveStor SERIES 製品ライン

Panasas ActiveStor製品仕様

	Series 7		Series 8		Series 9	
性能データ [1] 及び容量	Performance Module (単体)	ラック [4] あたりの性能と容量	Performance Module (単体)	ラック [4] あたりの性能と容量	Performance Module (単体)	ラック [4] あたりの性能と容量
最大バンド幅 MB/sec	350MB/sec	3.5GB/sec	600MB/sec	6GB/sec	600MB/sec	6GB/sec
最大IOPS[1]	12K	120K	12K	120K	21K	210K
物理ストレージ容量	16TB-44TB	160TB-440TB	16TB-44TB	160TB-440TB	8TB-10TB	80TB-100TB
ストレージキャッシュ容量	5GB-22GB	50GB-220GB	16GB- 44GB	160GB-440GB	32GB- 40GB	320GB-400GB
平均レスポンス時間	7-12msec	7-12msec	5-9msec	5-9msec	0.5-4msec	0.5-4msec
接続オプション、データマネージメント、高可用性オプション						
シングルネットワークスイッチ	1 GbE x 4		1 GbE x 4 10 GbE x 1 (Copper 又はOptical)		1 GbE x 4 10 GbE x 1 (Copper 又はOptical)	
追加ネットワークスイッチ	オプション		オプション		オプション	
InfiniBand(DDR/QDR)接続	N/A		オプション		オプション	
スナップショット [2]	オプション		標準サポート		標準サポート	
可用性オプション [3]	オプション		標準サポート		標準サポート	

1) 性能データはネットワーク構成やクライアント構成に依存します。
2) Panasas ActiveImage Snapshotはボリュームスナップショットの機能を提供します。
3) Panasas ActiveStor Network and Volume Failoverによるボリュームとネットワークのフェイルオーバーの提供を行います。
4) ラックは標準の42Uサイズの19インチラックを想定しています。

スケーラブルシステムズ株式会社

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 - 40TB / Module

DirectorBlade
メタデータ処理

セカンドネットワーク
スイッチ（オプション）

バッテリモジュール
電源バックアップ

StorageBlade
オブジェクトデータ処理
2TB - 4TB / Blade

GbE、10GbE
ネットワークスイッチ

冗長化電源
ホットスワップ可能

Panasas ActiveStor SERIES 9

ホットスワップ可能なストレージブレード

大容量メモリ
キャッシュ
40GB/モジュール

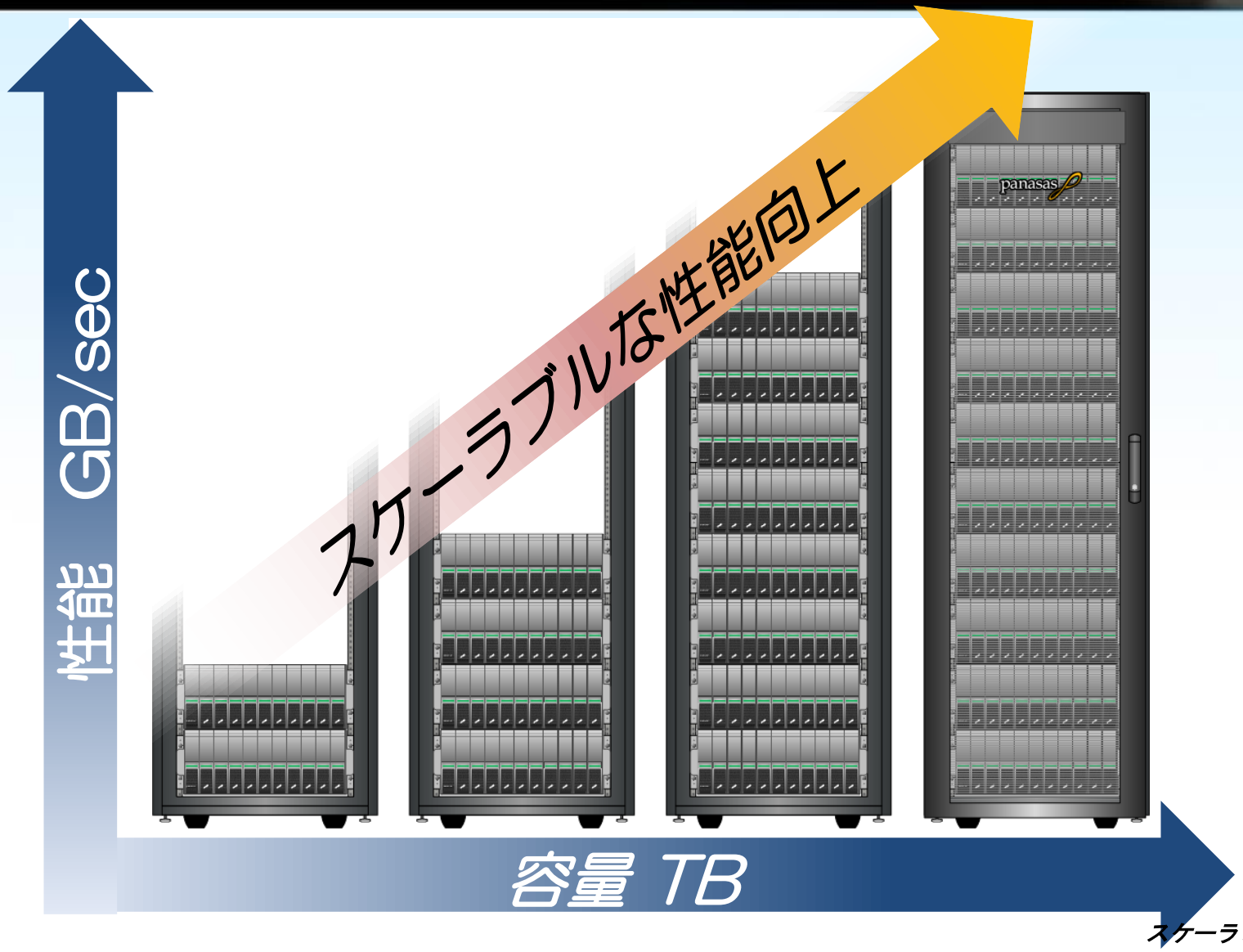
大容量HDD
ユーザデータ格納

SSD ドライブ
メタデータ処理と小さな
ユーザデータの処理の効率化

- SSD搭載StorageBlade
  - **4 GB RAM(バッテリバックアップ)**
  - **32 GB SLC Solid State Disk**
  - **1 TB SATA drive**
- 特徴
  - **データサイズによって、SSDとHDDへの自動的なデータ転送**
  - **IOPS 性能の大幅な向上**
  - **DRAM + SSD + HDD によるコストパフォーマンスの最適化**
  - **高い信頼性**

Panasas ActiveStor

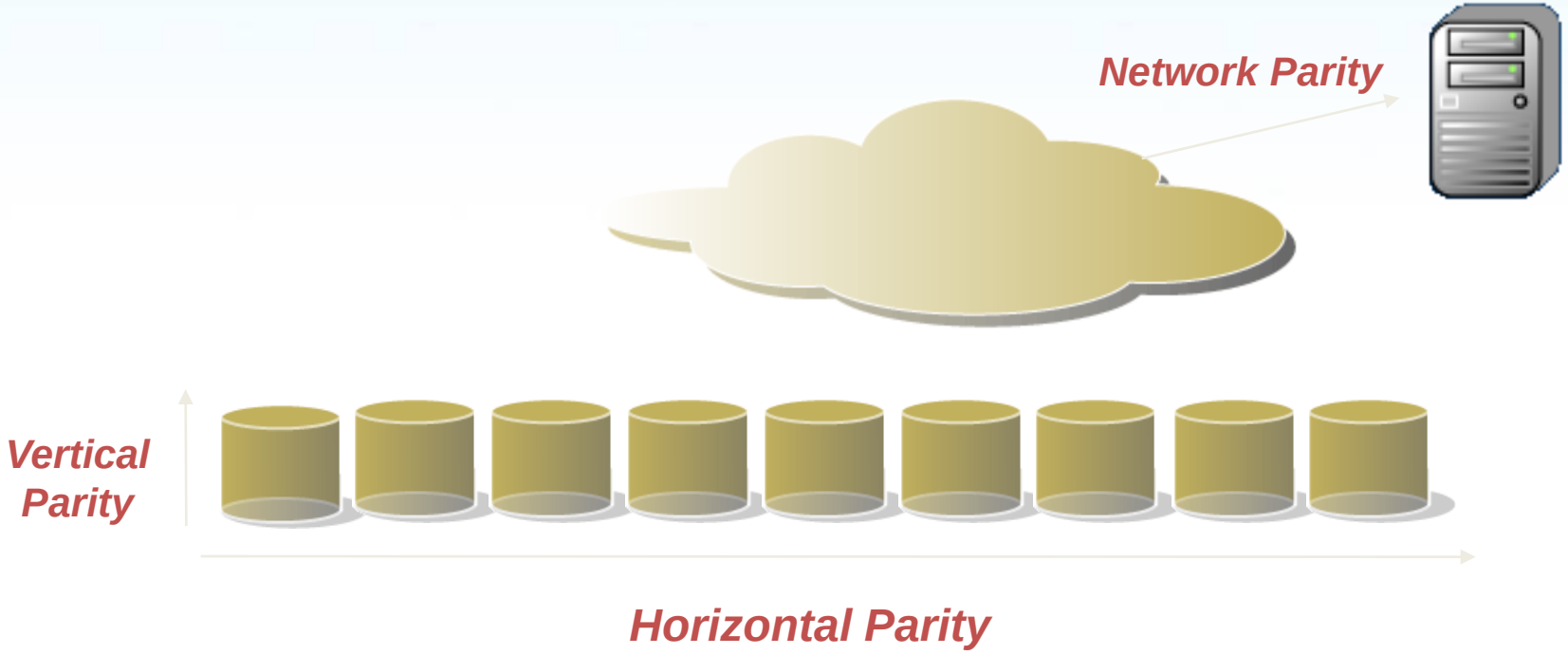
スケーラブルストレージアーキテクチャ



スケーラブルシステムズ株式会社

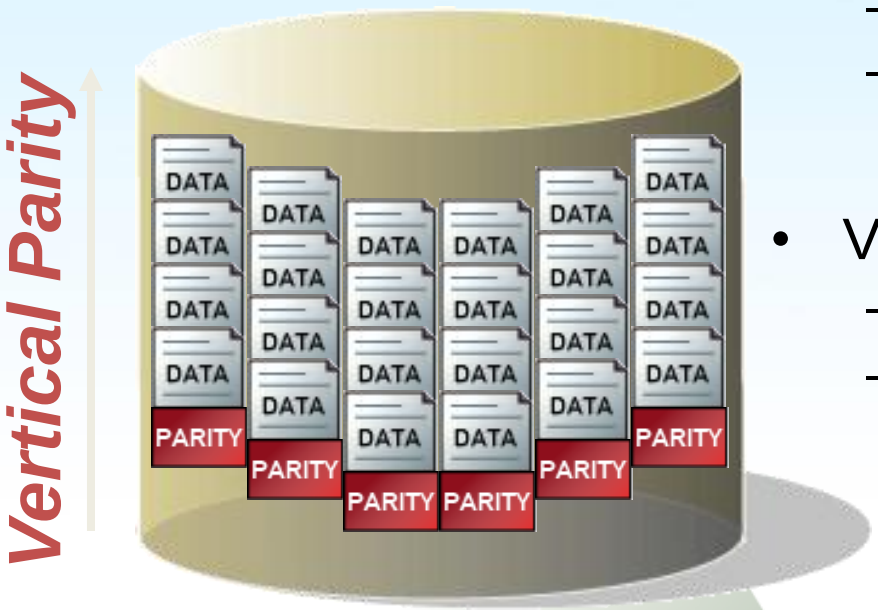
Panasas Tiered Parity

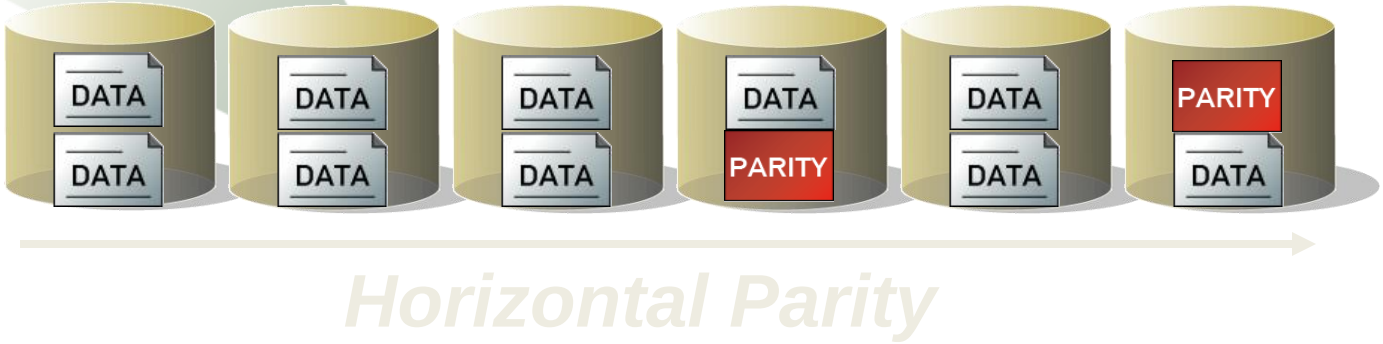
- 各Tierオペレーションは、独立したパリティの処理を行うことが可能であり、エラー検知とデータ修正を行う
- PanasasのTiered Parityが提供する3つのパリティ処理は、互いに相互補完



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 比較一覧

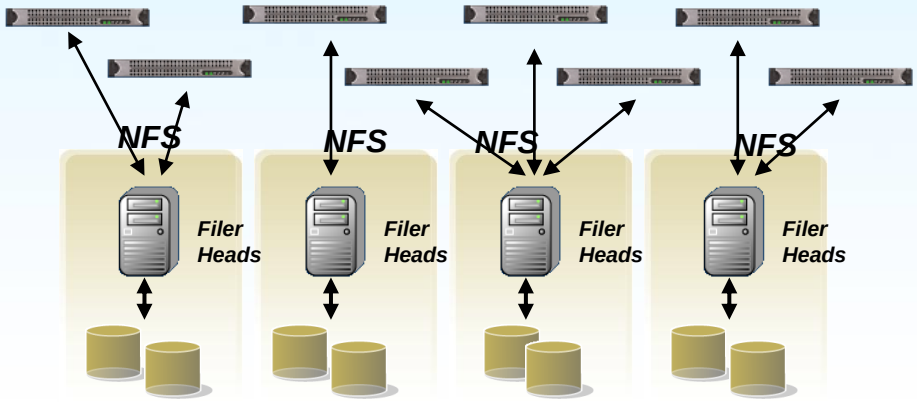
	RAID 5	RAID 6	Panasas Tiered Parity
ディスク1台が故障	Yes	Yes	Yes
ディスク1台が故障＋メディアエラー	No	Yes	Yes
ディスク2台が故障	No	No*	No
データ破損（メッセージなし）	No	No	Yes

** RAID再構成時にメディアエラーが発生した場合*

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

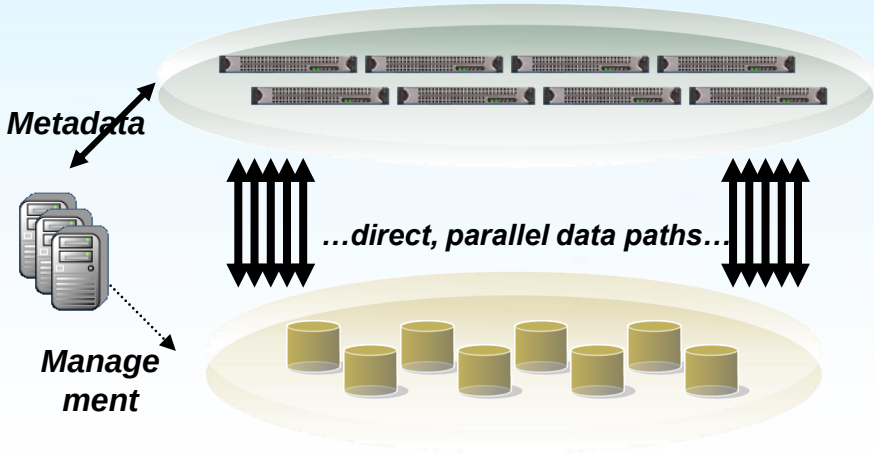
パラレルストレージ



“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

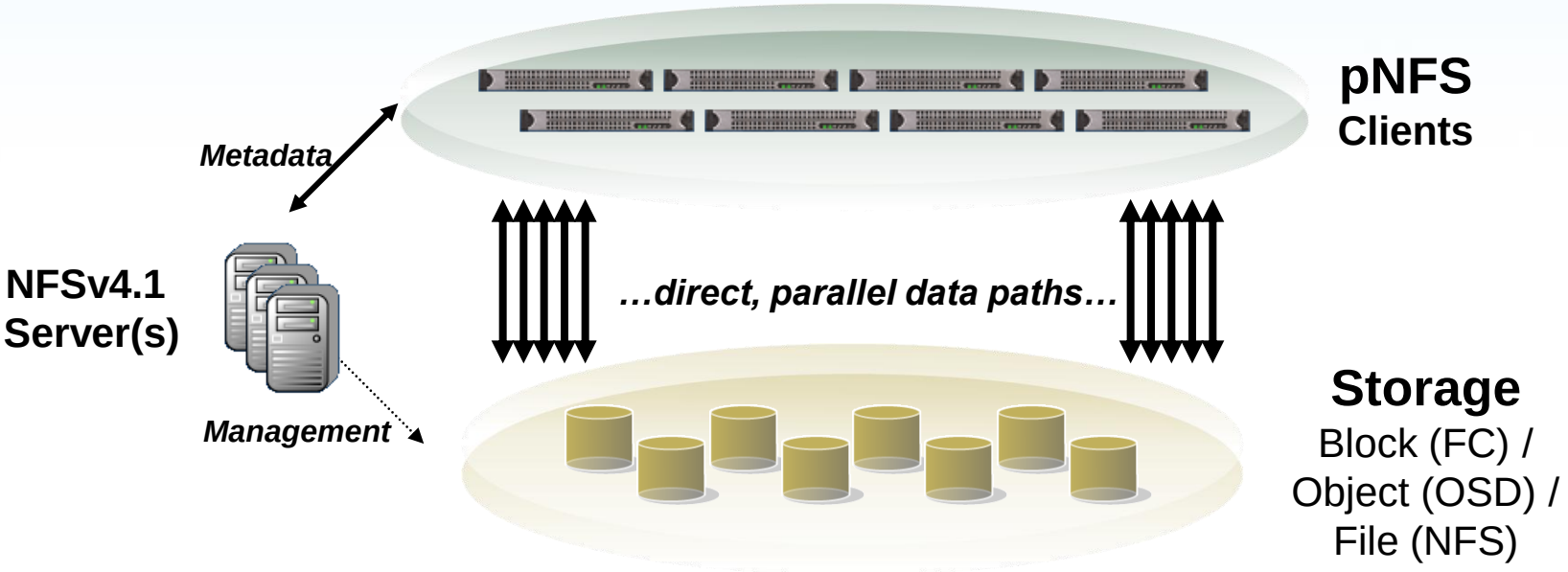


“パラレルクラスタストレージのプール”

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

性能

パラレルストレージ
- *高性能*
- *各社独自開発（非互換）*

CFS Cluster File Systems, Inc. panasas IBM IBRIX EMC² where information lives

pNFS(標準パラレルNFS)ストレージ

高性能＋互換性＋可用性
- *グローバルネームスペース*
- *オブジェクトストレージ*
- *次世代RAIDシステム*
  ‘PanasasTiered Parity’

NetApp ISILON BLUE ARC

クラスタNAS
- *NFSサーバのクラスタ化*

NetApp

NFS ファイルサーバ

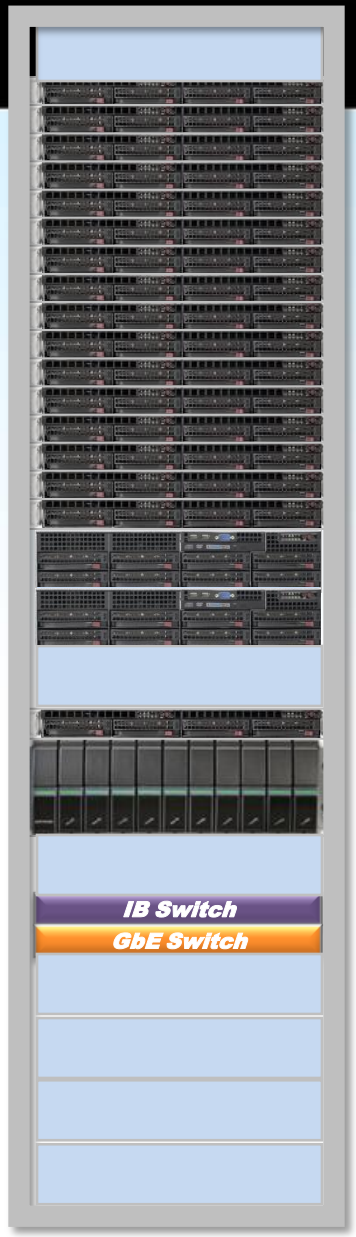
より詳細な情報は.....

http://www.hp2c.biz/panasas

Late 1980s *~2000* *2007+*

All-In-One HP^{2}C システム

All-In-One
HP²C システム



クラスタ（VXPRO R1400)
- 32ノード、256コア
- 3.0TFLOPSピーク性能
- InfiniBand (QDR)

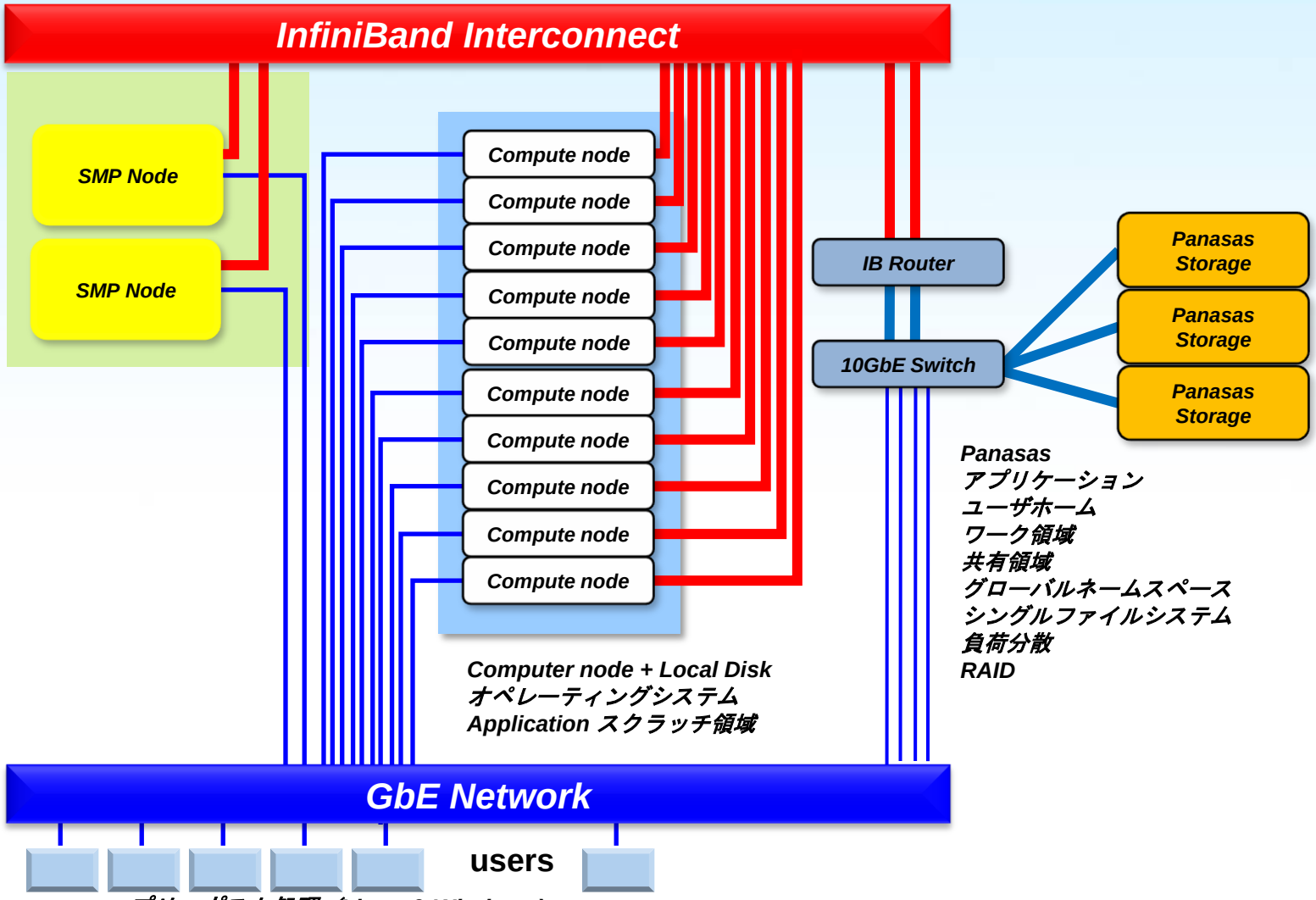
SMPシステム（VXSMP 2280)
- 16コア、144GB
- 共有メモリ＋標準Linux

ストレージ（Panasas Series8）
- 20TB共有ストレージ
- スケーラブル共有ストレージ
- Panasas IB Router

インターコネクト
- InfiniBandスイッチ (QDR)
- GbEスイッチ

42ラック標準ラック

All-In-One HP²C システム構成



All-In-One HP²Cシステム構成システム

VXPROサーバ

VXPROは、1U/2U/3Uサイズで提供するサーバ製品です。HPCのワークロードに最適な筐体サイズ、チップセット、プロセッサ、IO、ネットワークの構成を選択して構築可能な製品です。製品はすべて標準ラックに搭載可能です。VXPRO R1440は、1Uサイズに最新のインテルXeonプロセッサを4台搭載可能なサーバ製品です。クァッドコアプロセッサを搭載した場合、このサーバは1Uサイズに16コアまで搭載可能であり、メモリも192GBまで拡張可能です。また、このサーバは、オンボードにQDR InfiniBandインターフェイスを搭載しているため、より容易にInfiniBandクラスタとしての構築が可能となります。VXPRO R2800 は、このVXPRO R1440の2台分以上の計算リソースを2Uサイズに搭載したもので、更に可用性も強化されています。

Panaceas ActiveStor パラレルストレージクラスタ

ActiveStorは、最近注目を集めているオブジェクト・ベース・ストレージです。この技術により、ActiveStorは保存したデータに対して強固なセキュリティ・ポリシーを適用したり、保存期間を設定し自動的に削除させたりといった高度なデータ管理が可能になったり、I/O処理を高度に並列化することが可能となります。グローバルなシングル ネームスペースによって、システムの運用管理におけるマウントポイントの管理を非常にシンプルにします。利用するクラスタのサイズとデータが様々なジョブの実行に際して、そのワークロードの処理に非常に柔軟に対応可能です。ActiveStorは、GbE、10GbE、InfiniBand(IB Router)に接続可能です。

FUSION 1200 x86 SMPシステム

FUSION 1200は、デュアルコア/クアッドコアのインテル Xeonプロセッサを採用し、8プロセッサから最大48プロセッサまでの拡張性を備えているSMPシステムです。筐体あたり570GFLOPS以上の計算能力と最大768GBのメモリ容量、最大6TBの内蔵ストレージを備え、デスクサイドとしての設置とラック搭載の両方が可能 です。また、標準のLinuxディストリビューションをサポートし、100%のバイナリ互換であり、容易な運用とアプリケーションの高い実行性能を提供します。

InfiniBandスイッチ

スケーラブルシステムズは、All-In-Oneシステムに対して、InfiniBandスイッチによるインターコネクトを提案しています。InfiniBandは、全２重の通信方式を採用した入出力インターフェース技術で、QDR （40Gビット/秒）のリンク速度でノード間の接続が可能となります。

プリセットアップ

- オペレーティングシステム
  - 標準Linuxディストリビューション
  - マイクロソフト　Windows HPC Server 2008
- Panasasストレージクラスタ
  - ActiveScale 3.2オペレーティングシステム
- IB スイッチ　（32ポート 8x QDR)
  - IB ソフトウエアスタック
- 開発環境
  - インテルソフトウエア

All-In-One "*Look before you leap*"

Q:ヘッドノードをNFSサーバとして構築した方が、廉価なシステムが構築できると思いますが、なぜ、Panasasストレージクラスタを提案するのですか？

A:クラスタシステムの導入後、I/O性能が問題になることが多々あります。これは、クラスタシステムは、非常に高い処理性能を持ち、様々なワークロード処理を行うに際して、I/O性能がCPU処理性能に対して、相対的に弱いことが、導入後、顕在化することに起因します。導入前の簡単な性能評価やクラスタのカタログ性能では、このようなCPUとI/Oのバランスをワークロードに対して、評価するのは、非常に困難です。導入後、問題が発生した後で、その対応を検討することは、生産性の低下、より大きなコスト、ユーザ利用環境の再構築など様々な問題を引き起こします。

Mellanox InfiniScale IV 40Gb/s InfiniBandスイッチシステム

- 主なスイッチ仕様
  - 20&40Gb/s InfiniBand スイッチシステム
  - 36 InfiniBandポート
  - 最大40Gb/s/ポート（2.88Tb/sスイッチング能力）
  - QSFP コネクタ
  - 2重化電源
  - １Ｕサイズ（ラックマウント）
  - アダプティブ・ルーティング/最大6仮想サブネット
- 高バンド幅と低レイテンシによりハイパフォーマンスコンピューティング環境へ最高性能のファブリックソリューションを提供

ソフトウエア環境
プリインストール

- オペレーティングシステム
  - 標準Linuxディストリビューション
- Panasasストレージクラスタ
  - ActiveScale 3.2オペレーティングシステム
- IB スイッチ（36ポートQDR)
  - IB ソフトウエアスタック
- 開発環境
  - インテルソフトウエア
- クラスタ管理
  - オープンソースクラスタソフトウエア（Rocksなど）

Mellanox BridgeX BX4000 ゲートウェイシステム

- ハードウエア
  - 簡単な設定と管理用の内蔵CPU
  - アップリンク:最大4ポートの10/20/40 Gb/s InfiniBand、または最大12ポートの10GigE
  - ダウンリンク:最大16ポートの2/4/8Gb FCまたは、12ポートの10GigE
  - 冗長化用デュアルパワーサプライ
  - 高可用性とフェイルオーバーモード
- ソフトウエア
  - 設置と管理 ：VLAN, NPIV, MAC, WWN
  - 統合されたゲートウェイ管理
- インフィニバンドやイーサネットを使ったVirtual Protocol Interconnect(VPI)ゲートウェイ

RackSwitch G8000 / G8100 / G8124
GbE/10GbE スイッチ

G8000

- 高さ：1U、48 x 1G + 最大4 x 10Gアップリンク
  - 44 x 1G RJ45 / 4 x 1G SFP fiber / 4 x 10G （CX4 or SFP+）
- 高性能
  - 100%ラインレート性能
  - 176Gbpsスループット
- 低消費電力：124W
- エアフロー：Rear-Front / Front-Rear選択可

G8100
G8124

- 高さ：1U、24 x 10G
  - 20 x CX-4、4 x SFP+*1
  - 24 x SFP+ （Direct Attach可）*2
- 高性能
  - 100%ラインレート性能
  - 480Gbpsスループット
  - レイテンシ：平均約360ns*1, 700ns*2
- 次世代Ethernet対応（CEE/FCoE）
- 低消費電力：120W*1, 200W*2
- エアフロー：Rear-Front / Front-Rear選択可

**1：G8100, *2：G8124*

BLADE NETWORK TECHNOLOGIES

All-In-One HP²C システム構成例

- InfiniBand & Panasas IB Router
  - InfiniBandでの高速のインターコネクトと共有ファイルシステムの構築
  - Panasas IB Routerによる柔軟な運用
- InfiniBand & 10GbE Gateway
  - InfiniBandでの高速のインターコネクトと共有ファイルシステムの構築
  - 10GbE Gateway によるシステム構築
- GbE & 10GbEスイッチ
  - 高速なインターコネクトを必要としない場合に、高速な共有ファイルを利用可能とするシステム構築

All-In-One HP^{2}C システム構成
InfiniBand & Panasas IB Router

InfiniBand Interconnect

SMP Node

SMP Node

Compute node

Compute node

Compute node

Compute node

Compute node

Compute node

IB Router

10GbE Switch

Panasas
アプリケーション
ユーザホーム
ワーク領域
共有領域
グローバルネームスペース
シングルファイルシステム
負荷分散
RAID

Computer node + Local Disk
オペレーティングシステム
Application スクラッチ領域

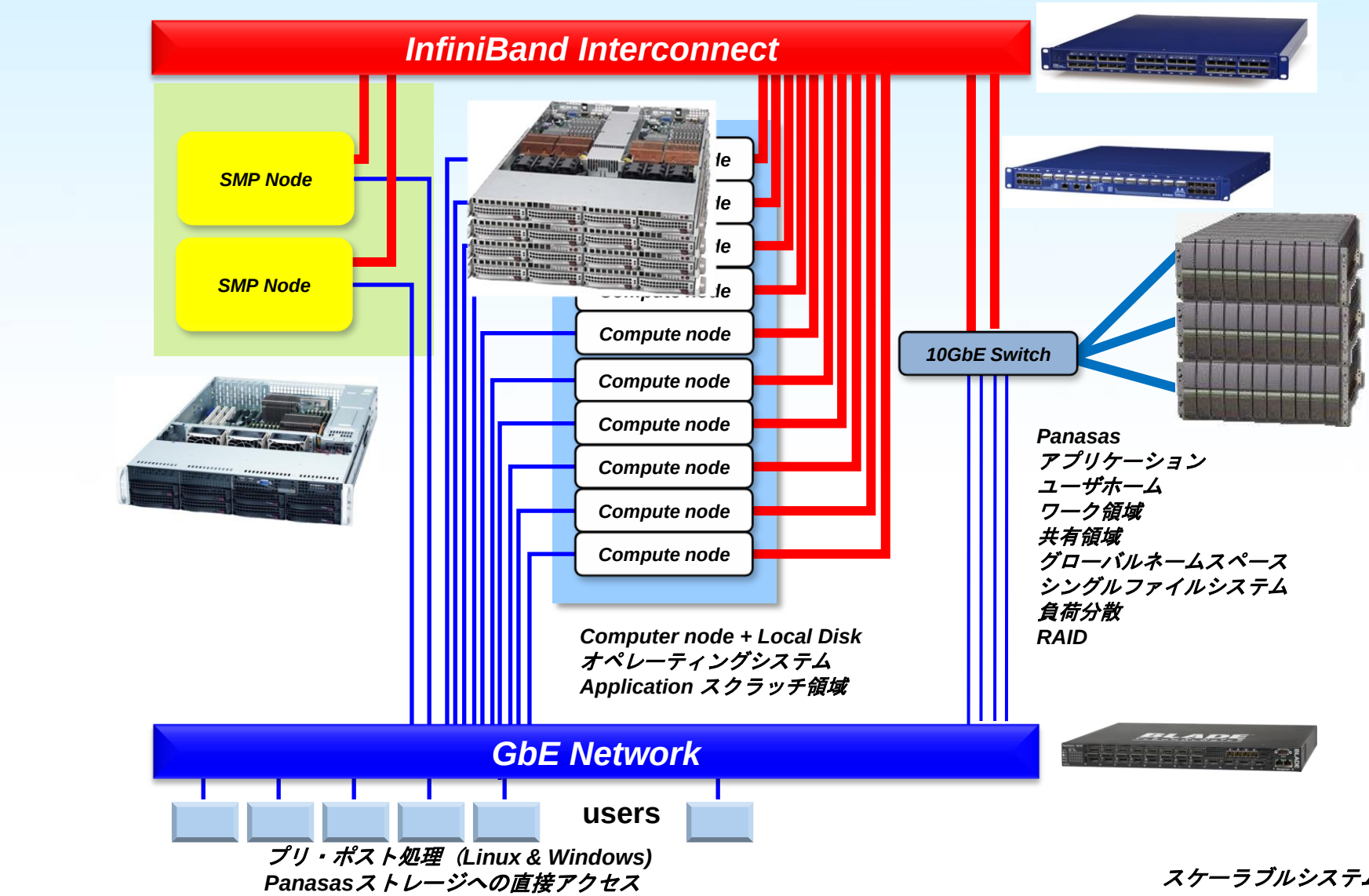
GbE Network

users

プリ・ポスト処理（Linux & Windows)
Panasasストレージへの直接アクセス

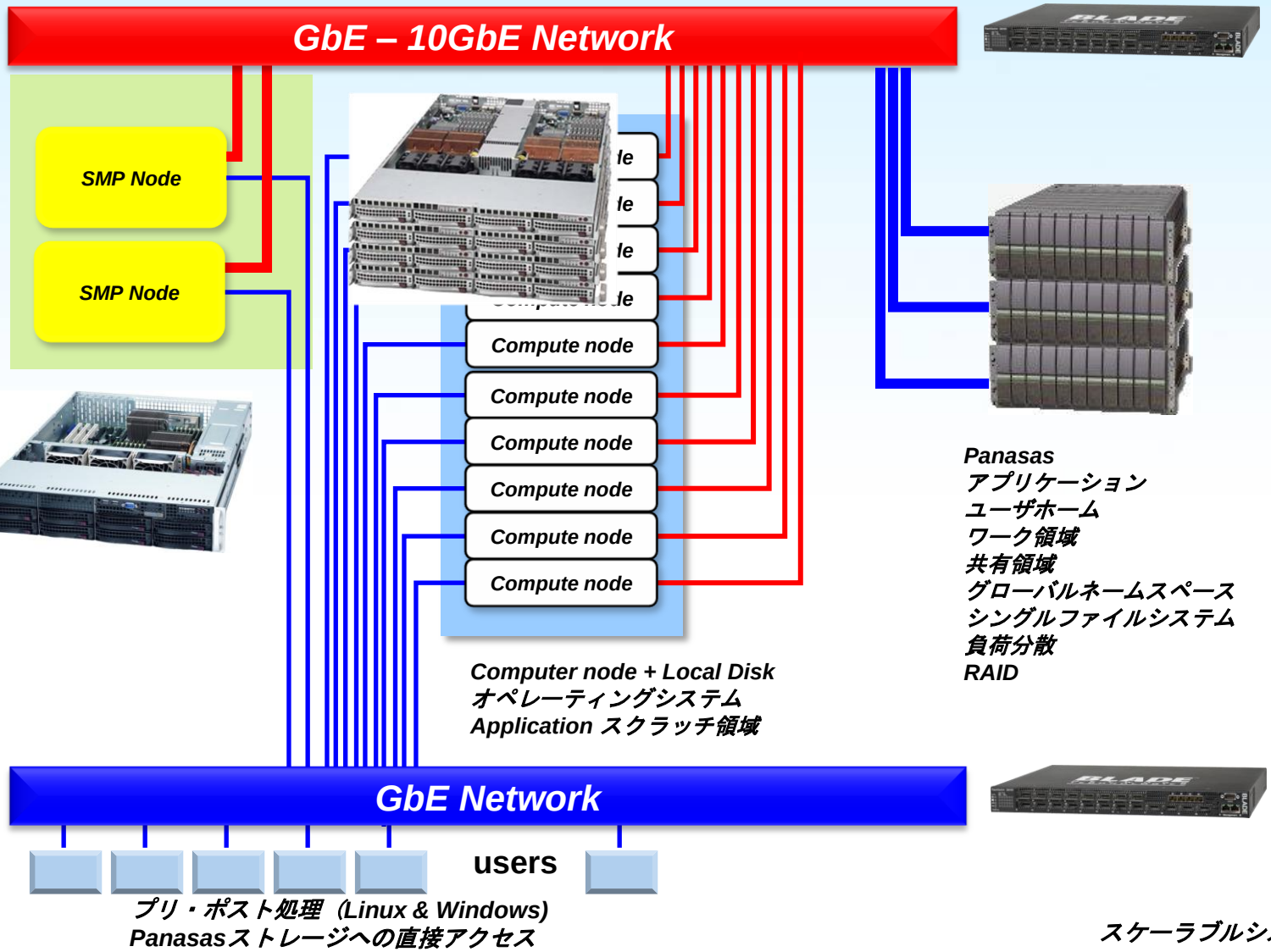
スケーラブルシステムズ株式会社

All-In-One HP^{2}C システム構成
InfiniBand & 10GbE Gateway



スケーラブルシステムズ株式会社

All-In-One HP^{2}C システム構成
GbE & 10GbEスイッチ



HP²Cコンサルテーション

システム構築

- 現有システムの評価とボトルネックの明確化
- 次期システム導入のための性能評価（ベンチマーク）
- 性能評価の結果などをベースに次期システムでのワークロード処理のシュミレーション
- 次期システムの構築とシステムの動作検証などの支援

ベンチマーク

- ベンチマークの内容についてのアドバイス
- ベンチマークの実施のためのリソースの提供
- 実際の性能評価試験の実施
- プログラムについての、最適化支援や並列化に関するコンサルテーション

導入後サポート

- プログラミングトレーニング
- 利用技術に関するWEBページなどの提供
- アプリケーションに関する技術コンサルテーション（トラブル時の対応サポートや問題の切り分けやISV様との技術的対応支援など）
- システムの増強や構成変更

スケーラブルシステムズ株式会社

コンサルテーション

アプリケーションの最適な実行
を支援するテクノロジー

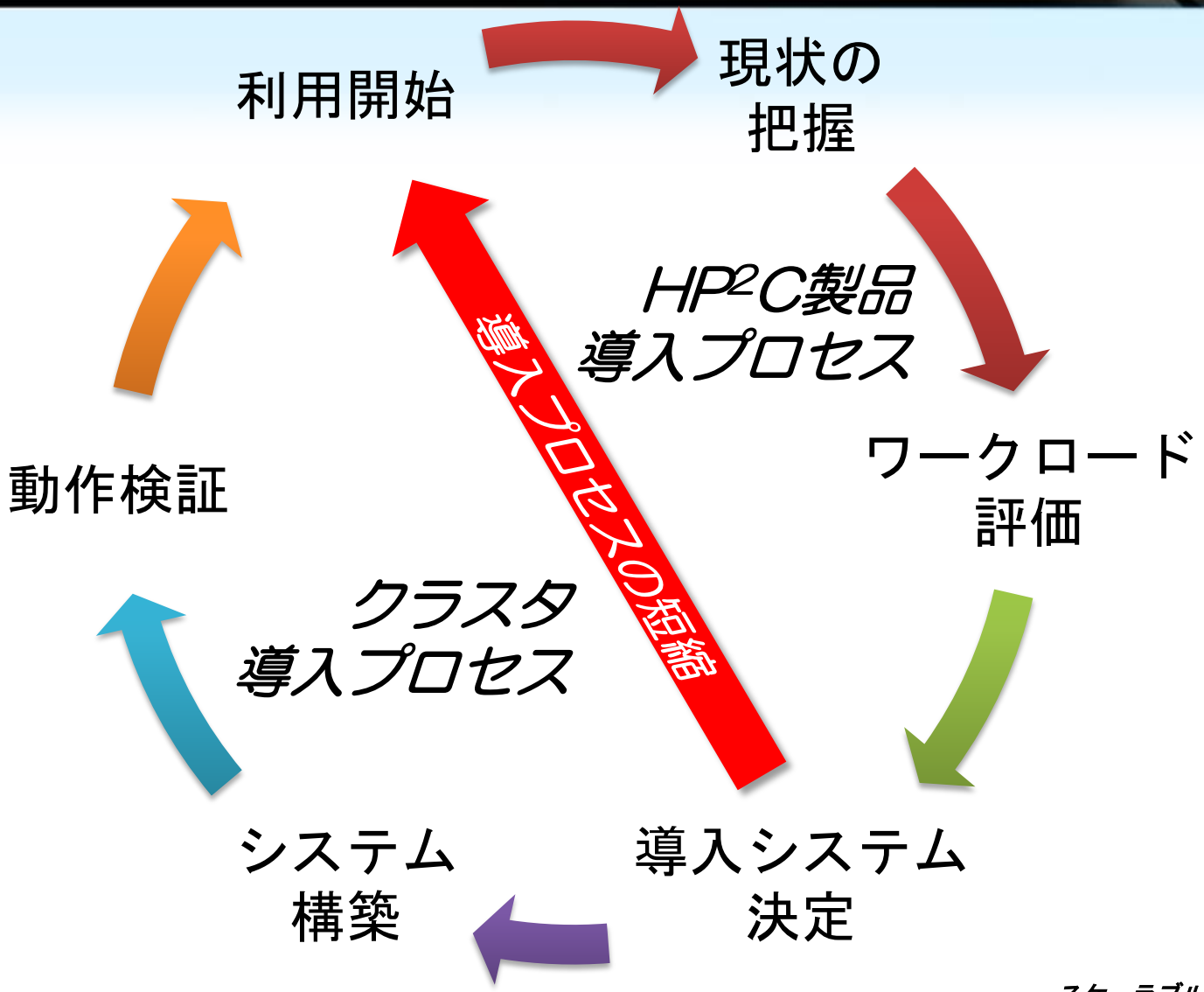
HP²C製品

アプリケーション実行のための
最適なプラットフォーム

HP²Cコンサルテーション

クラスタの運用管理費用の大幅な低減
ワークステーションやSMPサーバとのギャップの解消
クラスタでの面倒な利用環境の問題の解消

HP²C製品システム導入プロセス



スケーラブルシステムズ株式会社

HP²C製品ライン

FASTER

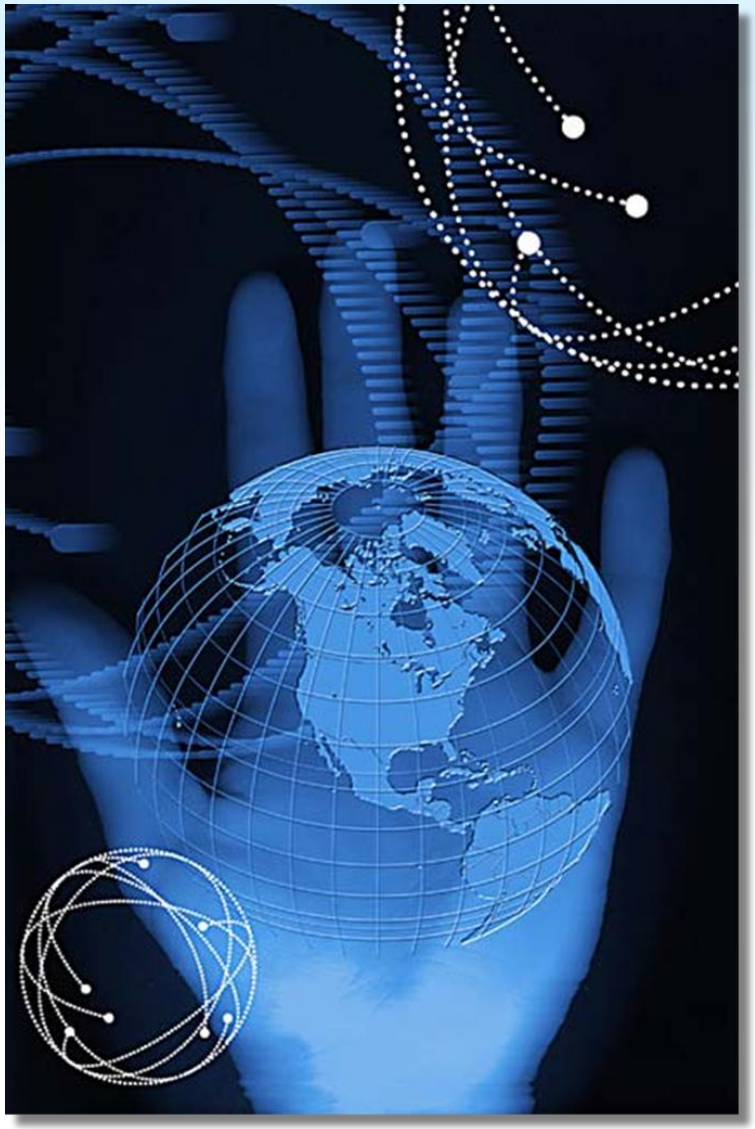
- 高い性能とスケーラビリティ
- 最新テクノロジーによるシステム提案

BETTER

- 運用管理が容易で利用し易い
- 容易なプログラミングと互換性

AT LOWER COST

- 導入コストの低減
- 運用管理コストの削減



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

2/18/2010

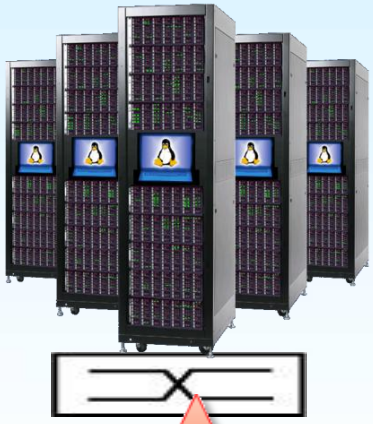
スケーラブルシステムズ株式会社

並列処理におけるIO処理の問題

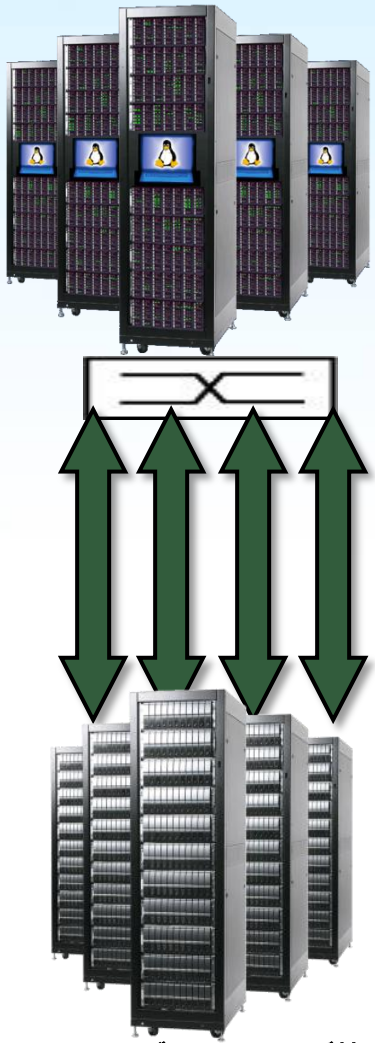
補足資料

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



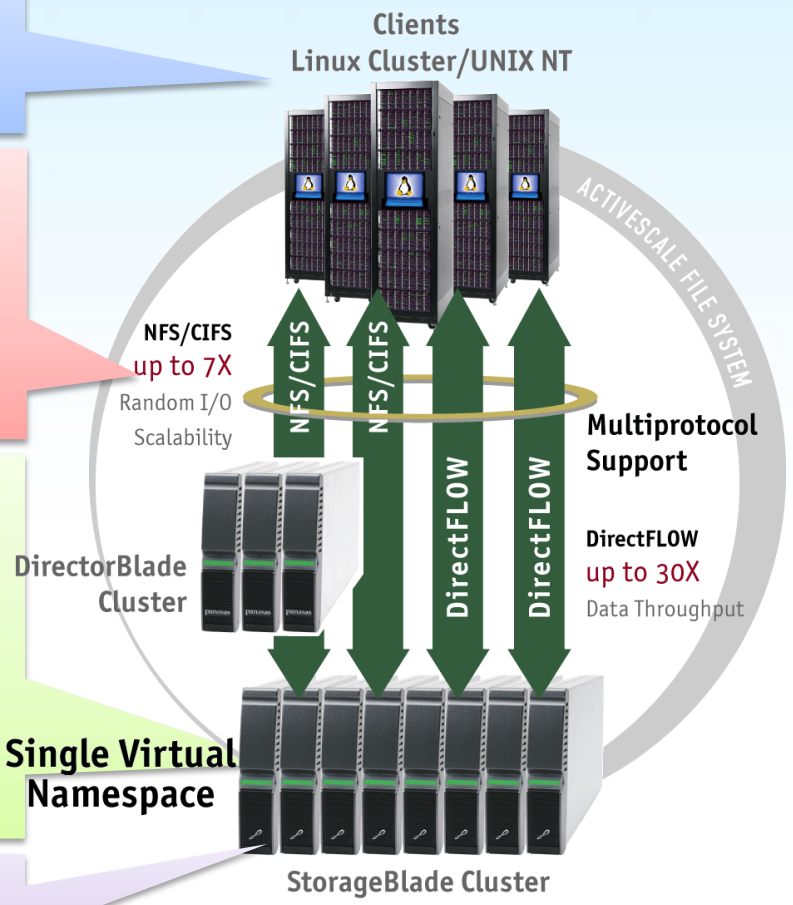
Panasasストレージクラスタ

DirectFLOW クライアントS/W
- *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にも対応可能

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増強が可能
- 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
- *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*



CAEにおけるI/Oボトルネック

CAEでのシングルジョブのI/O処理の比重

1999: Desktops

Typical overnight job – single thread

read = 10 min　compute = 14 hrs　write = 35 min　5% I/O

2004: SMP Servers

注意: 以下の推定ではジョブのI/O時間は一定で、並列処理によって、非I/O処理部分の時間は短縮されると仮定しています.....

8-way　I/O significant

10 min　1.75 hrs　35 min　30% I/O

2009: HPC Clusters

32-way　I/O-bound – growing bottleneck

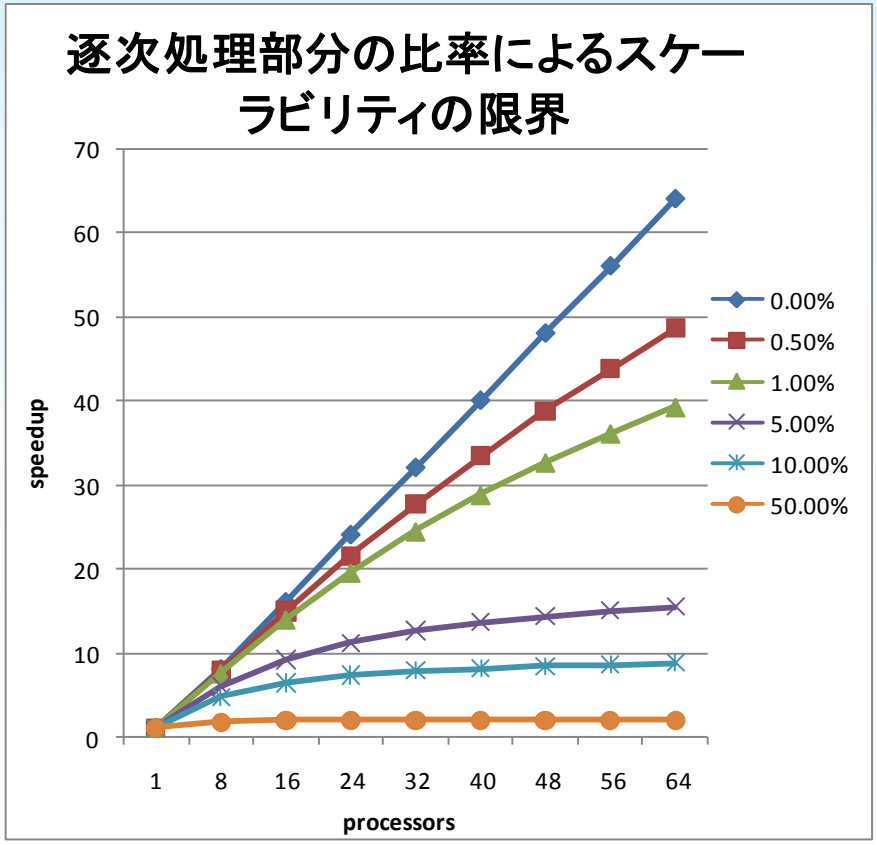
10 min　26 min　35 min　*63% I/O*

注意: I/O処理部分に関して、性能向上や並列化などの改善がないという極端な仮定での推定であり、実際のCAEでのシングルジョブのI/O処理を完全にシュミレーションした結果ではありません。

並列処理でのI/O処理の課題

- IO処理
  - 遂次処理の典型であり、I/O処理自身を並列に処理することが高いスケーラビリティの実現のためには必須
- 並列処理でのI/O処理の課題（問題点）
  - マルチスレッド（マルチプロセッサ）を利用する並列アプリケーションの実行時の課題
  - 複数ジョブの同時実行における課題

アムダールの法則



- 実行時間 = 逐次処理 + 並列処理

理論的な性能向上の限界

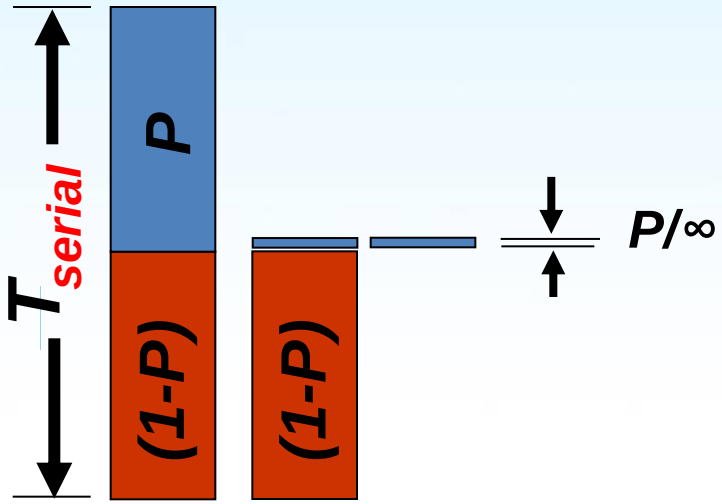
- 実行時間 = 逐次処理 + 並列処理時/P

- 64プロセッサで50倍の性能向上を得るには、逐次処理部分を0.5% 以下にする必要がある

I/O処理は逐次処理の典型であり、I/O処理自身を並列に処理することが高いスケーラビリティの実現のためには必須である

アムダールの法則

並列処理での性能向上の上限値（スケーリング）



$$T_{parallel} = \{(1-P) + P/n\}\, T_{serial} + O$$

n = number of processors

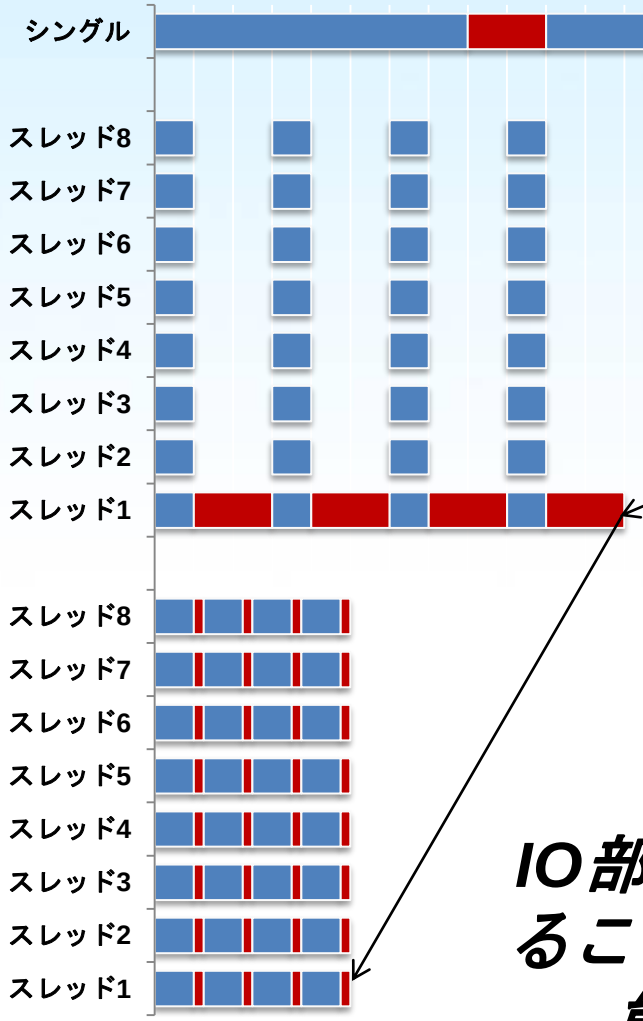
$$Scaling = T_{serial} / T_{parallel}$$

プログラムの逐次処理部分（非並列処理）部分の排除が必要

例えば、n=8, P = 0.8 の場合

$$Scaling = 1.0 / (0.2 + 0.1) = 3.3$$

アプリケーションの並列実行



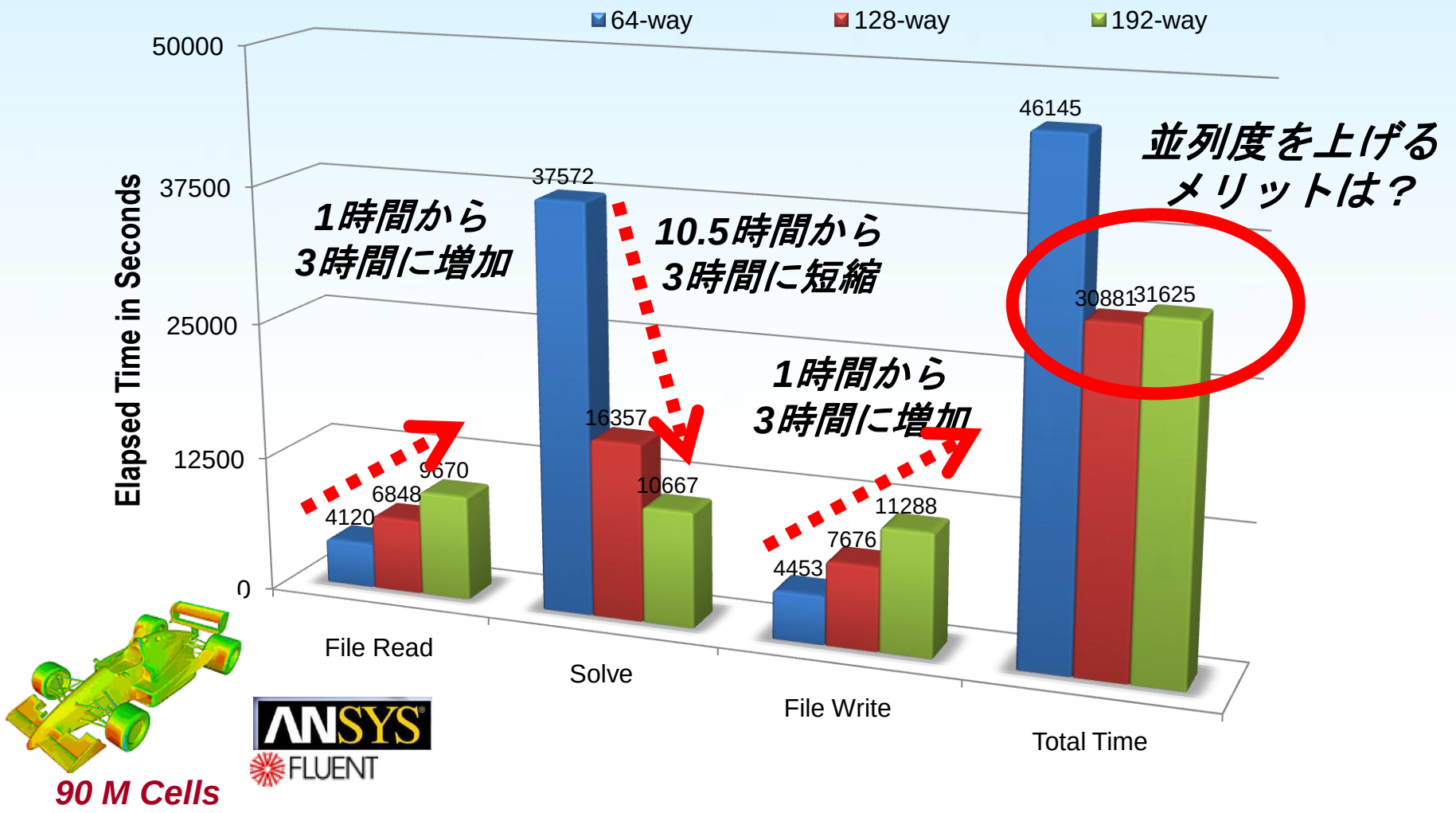
例えば、n=8, P = 0.8 の場合
Scaling = 1.0 / (0.2 + 0.1) = 3.3

プログラムの逐次処理部分（非並列処理）部分の排除が必要

IO部分を並列処理することで、2.4倍の性能向上が可能

計算処理部分
IO処理部分

FLUENT: Serial I/O (6.2)



Source: Barb Hutchings Presentation at SC06, Nov 2006, Tampa, FL

FLUENT: Serial I/O (6.2) vs. Parallel I/O (6.4/12-beta)

64-way 128-way 192-way

Elapsed Time in Seconds

50000
37500
25000
12500
0

46145
30881
31625
38656
17442
11759
44479
23266
17582

クラスタ導入のメリットは？

10.5時間から3時間に短縮

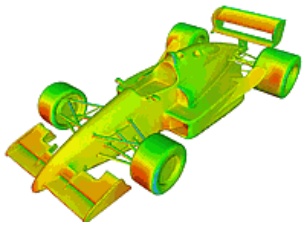
より多くのIOにも対応可能

PanFSは、Serial I/O に対して3倍以上の性能（192コア）
より多くのI/Oでもスケーラブルな性能を実現

6.2 - Serial I/O

PanFS - 1 Write

PanFS - 100 Writes



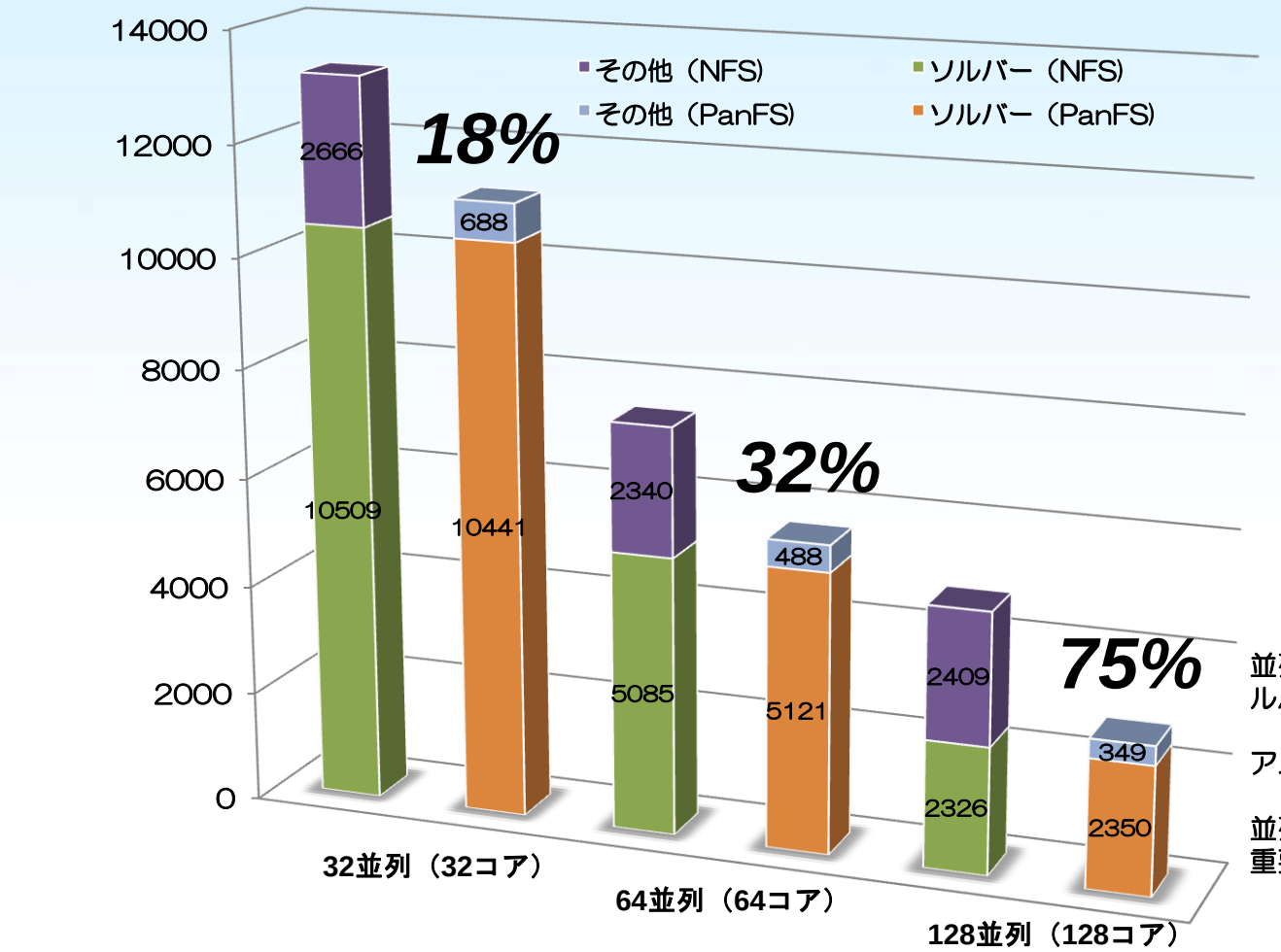
90 M Cells

ANSYS FLUENT

Panasas導入による並列処理の劇的な向上

Source: Barb Hutchings Presentation at SC06, Nov 2006, Tampa, FL

STAR-CD v4 性能評価

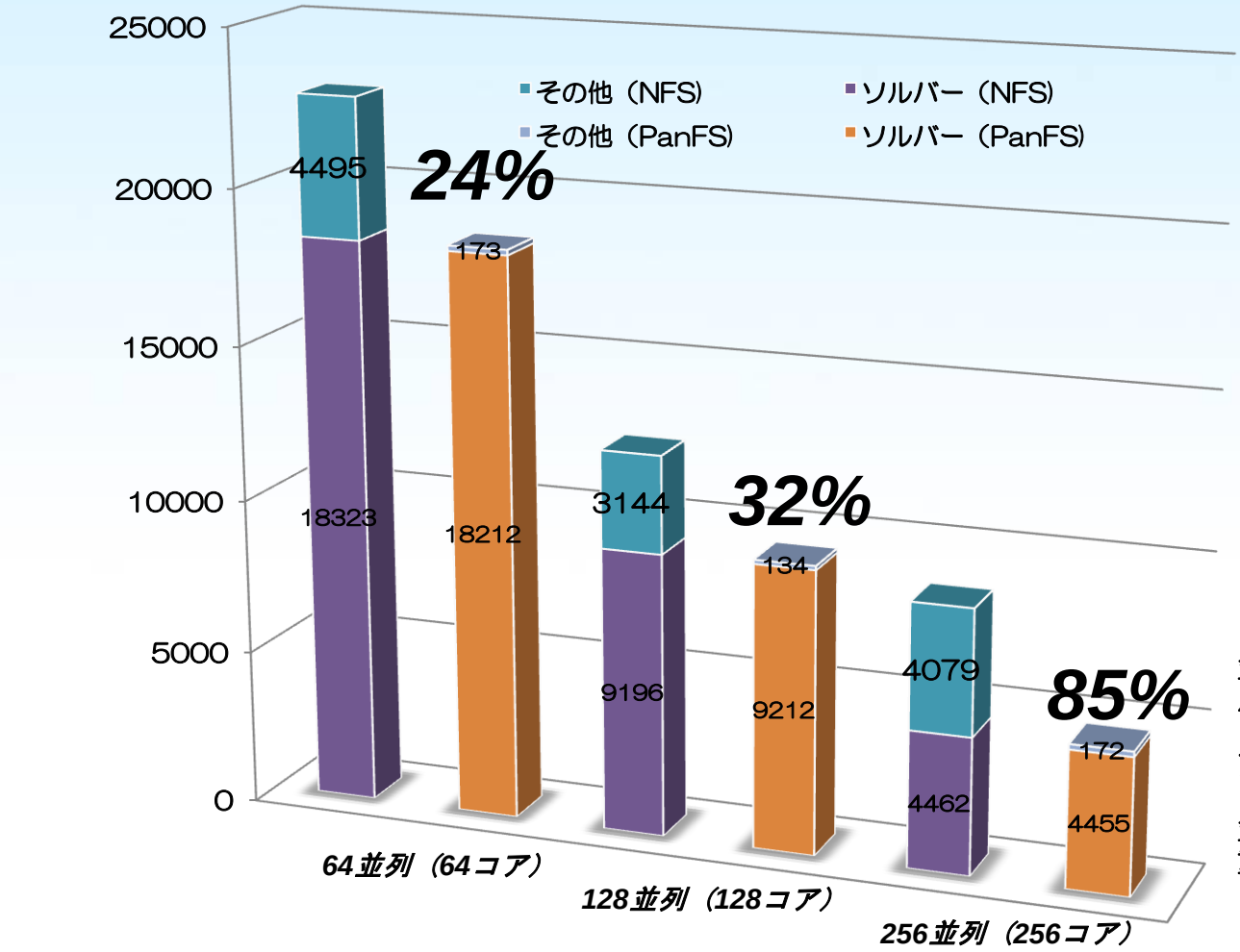


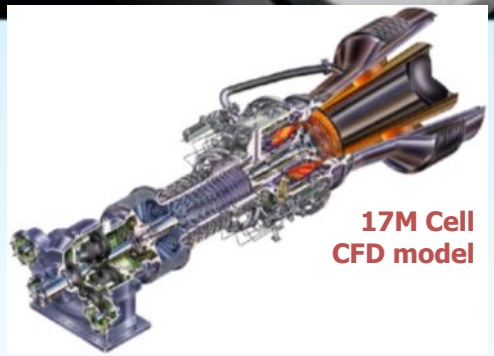
Number of cells
19,921,786
Solver
CGS, Steady
Iterations
500 total iterations - data save after every 10 iters
Each solution output (50 total)
~1,500 MB

並列度（コア数）が大きくなるに伴って、非ソルバー部分の比重が大きくなる
↓
アムダールの法則（非並列部分が性能を左右）
↓
並列IO処理などによる非並列計算部分の削減が重要

この性能評価はPanasas社とインテル社が、インテル社のクラスタシステム（2048コア）を利用して計測した性能です。
File Systems -- Panasas,: 7 shelves, 35 TB storage;（各シェルフは、4xGbE接続でトータル2.8GB/sec のバンド幅）
NFS: Dell 2850 File Server, 6 x 146 GB SCSI drives, RAID 5

STAR-CD v4 性能評価





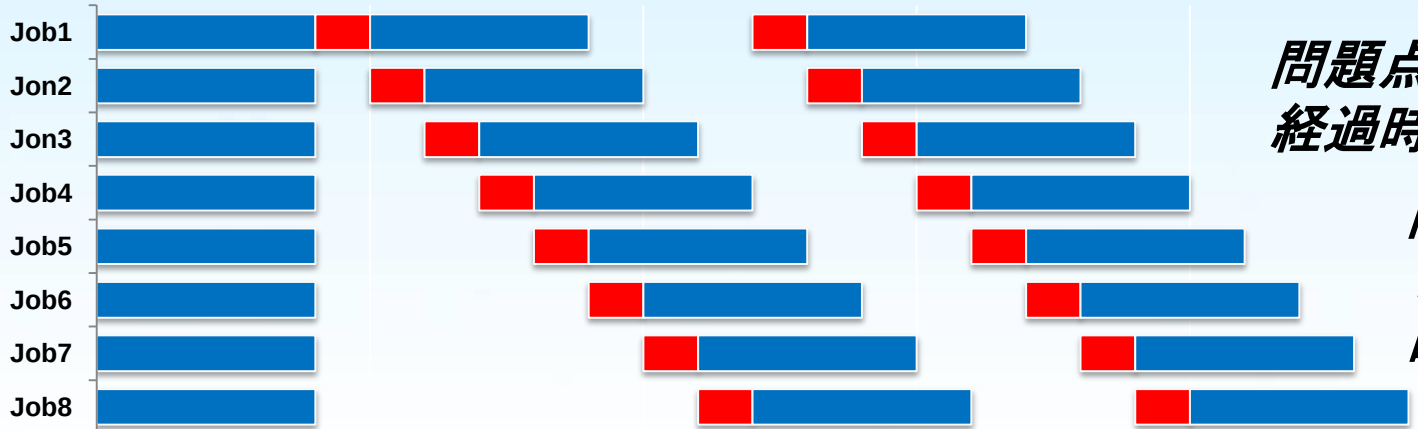
Number of cells
16,930,109
Solver
CGS, Single Precision
Iterations
300 total iterations -
data save after every 100 iters
Total solution output
~48 GB

並列度（コア数）が大きくなるに伴って、非ソルバー部分の比重が大きくなる
↓
アムダールの法則（非並列部分が性能を左右）
↓
並列IO処理などによる非並列計算部分の削減が重要

この性能評価はPanasas社とインテル社が、インテル社のクラスタシステム（2048コア）を利用して計測した性能です。
File Systems -- Panasas,: 7 shelves, 35 TB storage;（各シェルフは、4xGbE接続でトータル2.8GB/sec のバンド幅）
NFS: Dell 2850 File Server, 6 x 146 GB SCSI drives, RAID 5

マルチジョブでのIO処理

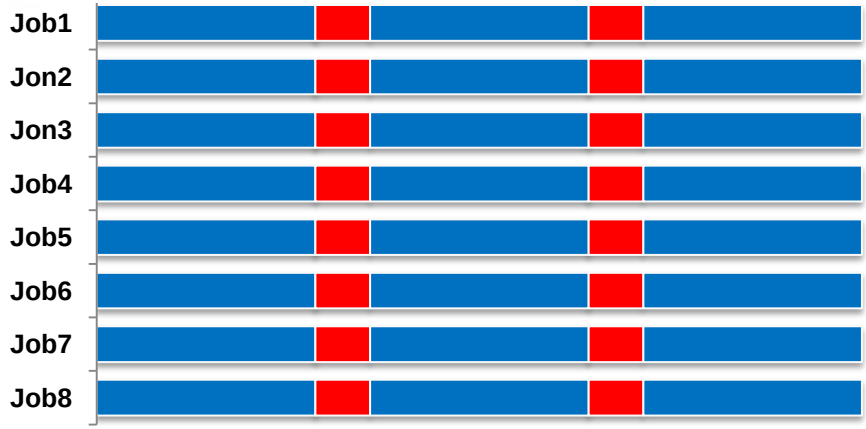
IO処理が逐次的に実行され、ジョブのIO処理時は他のジョブは処理の終了を待つ



問題点①
経過時間が伸びる

問題点②
ジョブ毎に処理時間が異なる

各ジョブが同時にIO処理を行うことが可能な場合には、IO待ちによる遅延は発生しない



複数ジョブの同時IO処理に対応可能なシステムでのIO処理

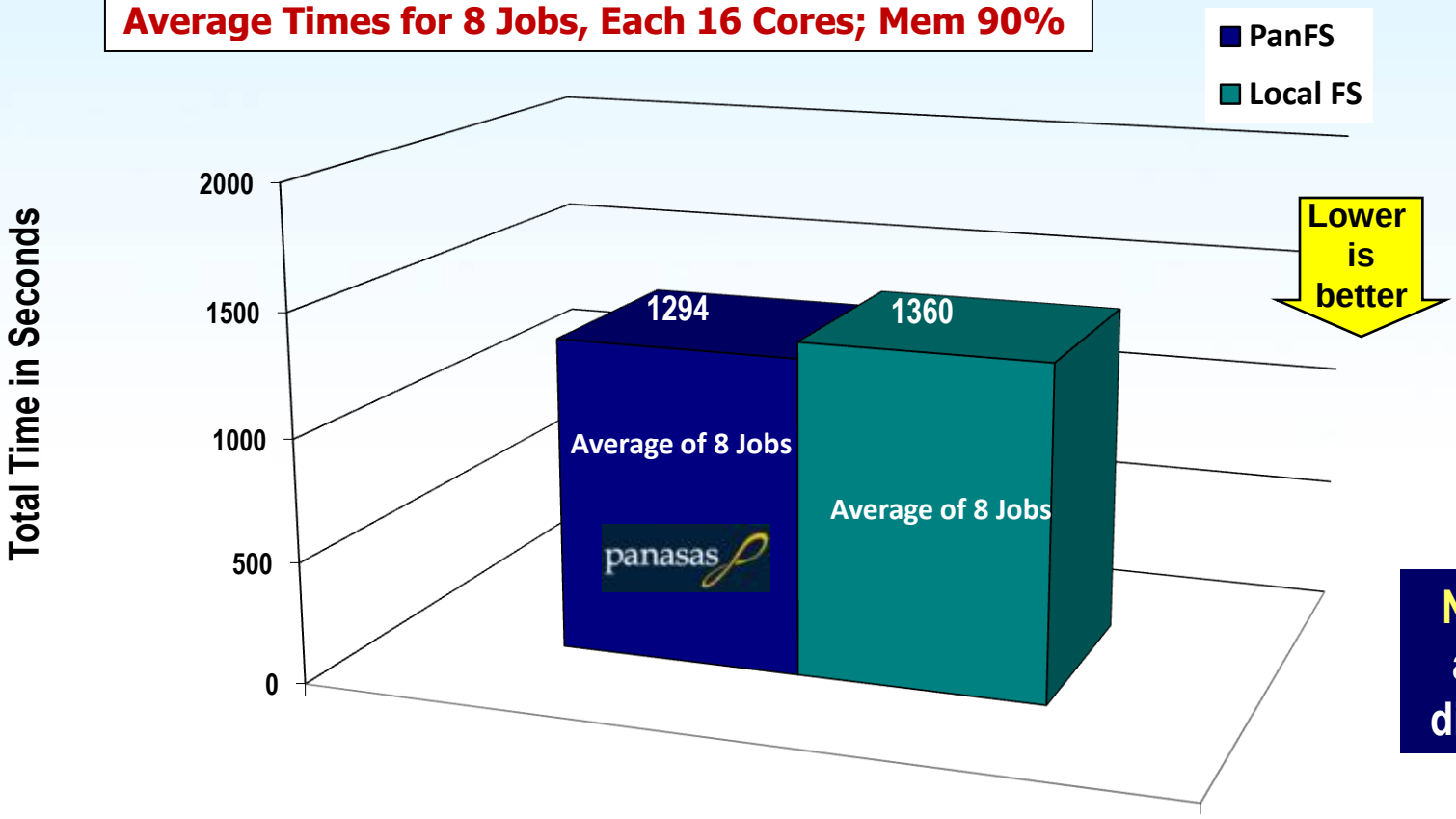
計算処理部分

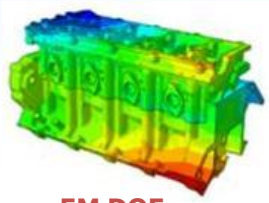
IO処理部分

Abaqusマルチジョブ性能

Abaqus/Standard 6.8-3: Comparison of PanFS vs. Local FS Ext2

Average Times for 8 Jobs, Each 16 Cores; Mem 90%



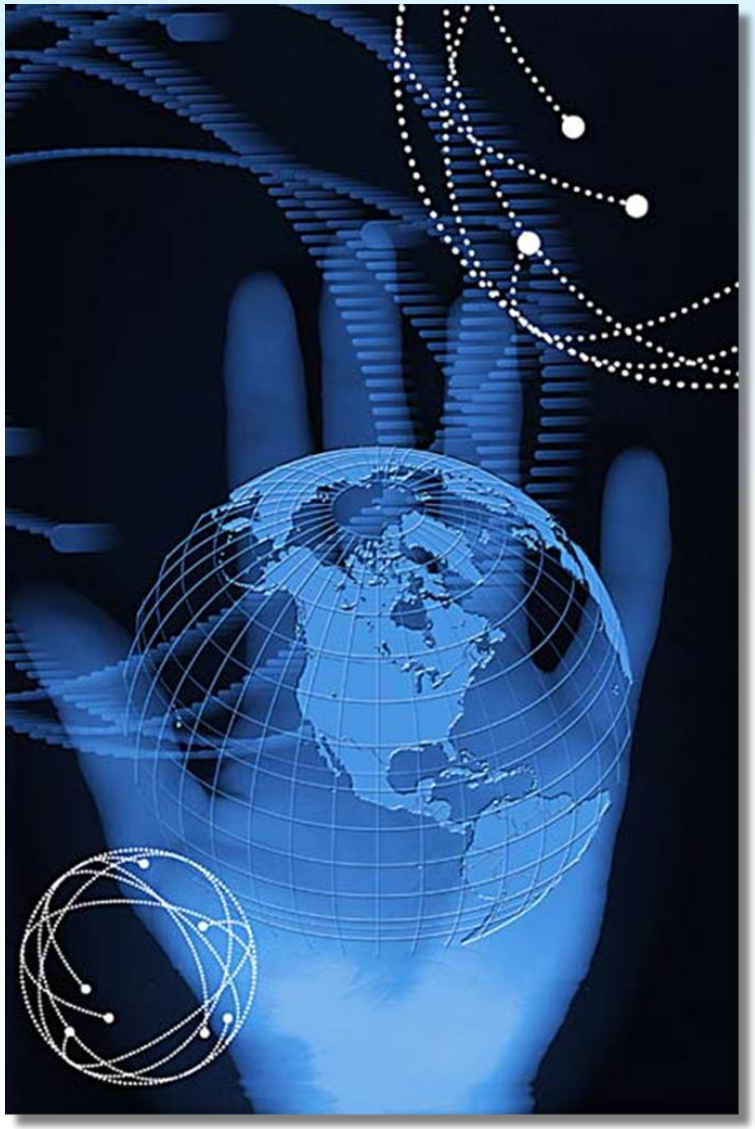


5M DOF
Engine Block

NOTE: PanFS and Local FS difference ~ 5%

8 Jobs x 16 Nodes x 128 Cores

Average Times for 8 Jobs | Each Job on 2 Nodes | Each Job on 16 Cores | Total 128 Cores



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

2/18/2010

スケーラブルシステムズ株式会社