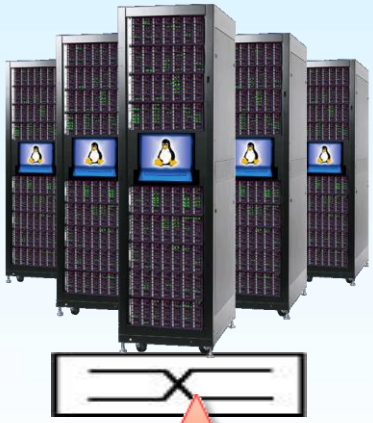
並列処理におけるIO処理の問題

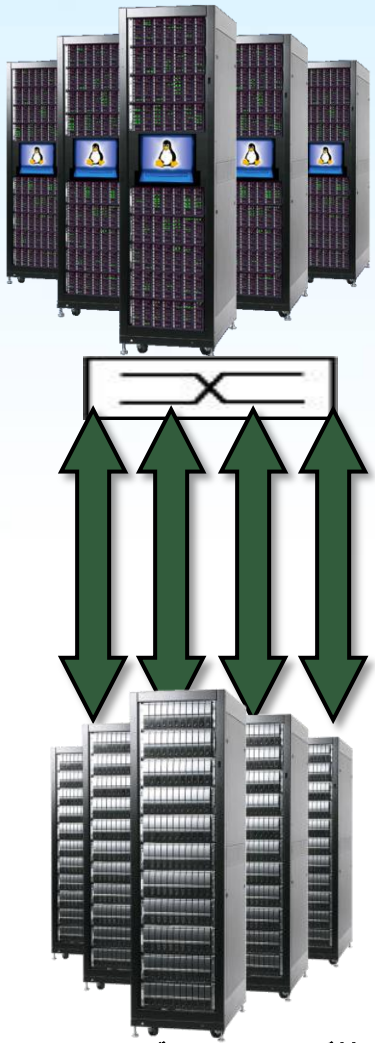
補足資料

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



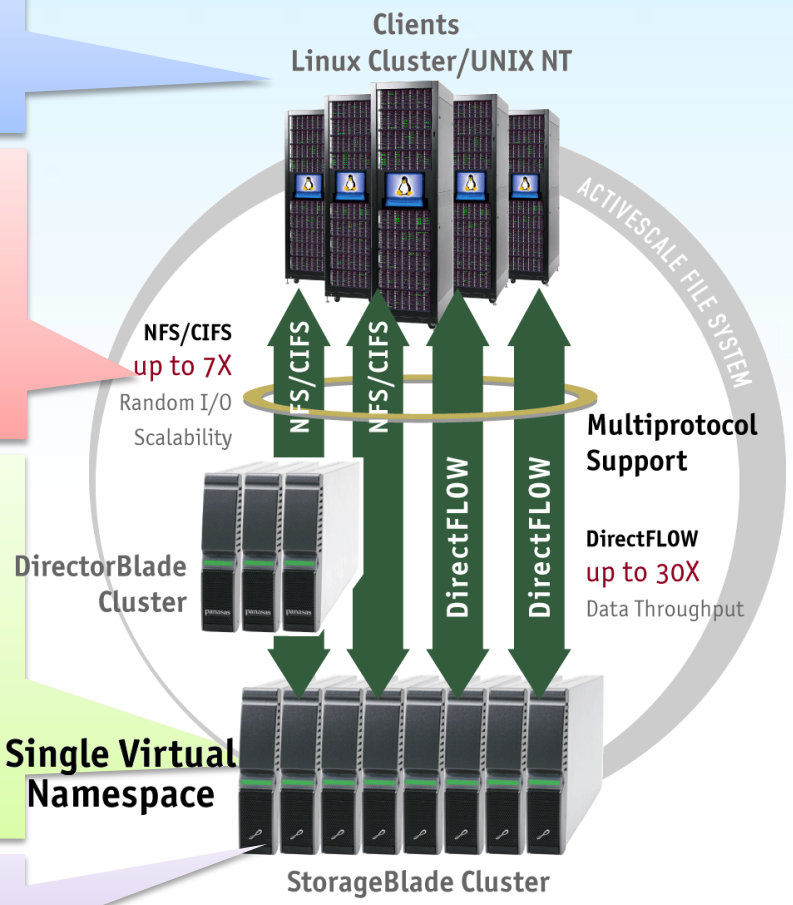
Panasasストレージクラスタ

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

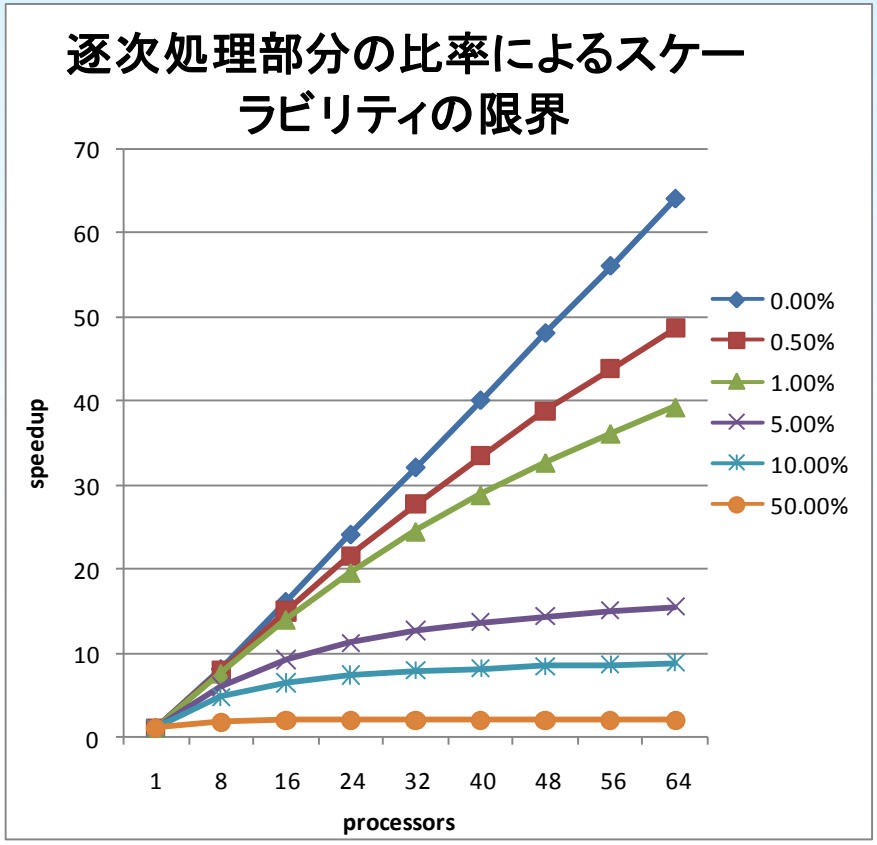
10 min　26 min　35 min　*63% I/O*

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

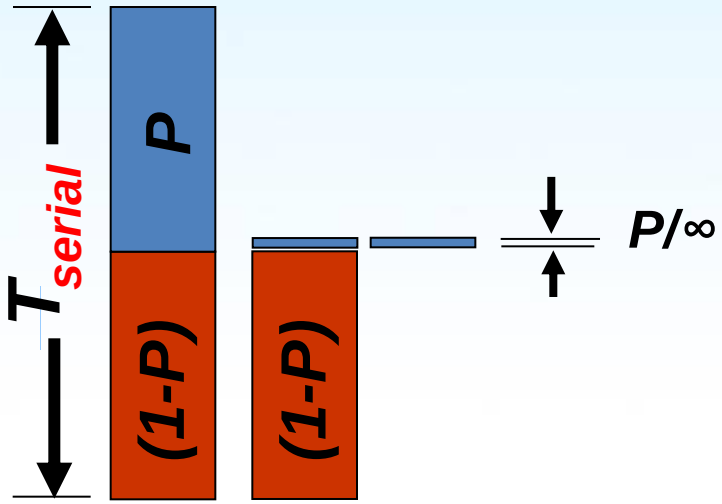
- 実行時間 = 逐次処理 + 並列処理時/P

- 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

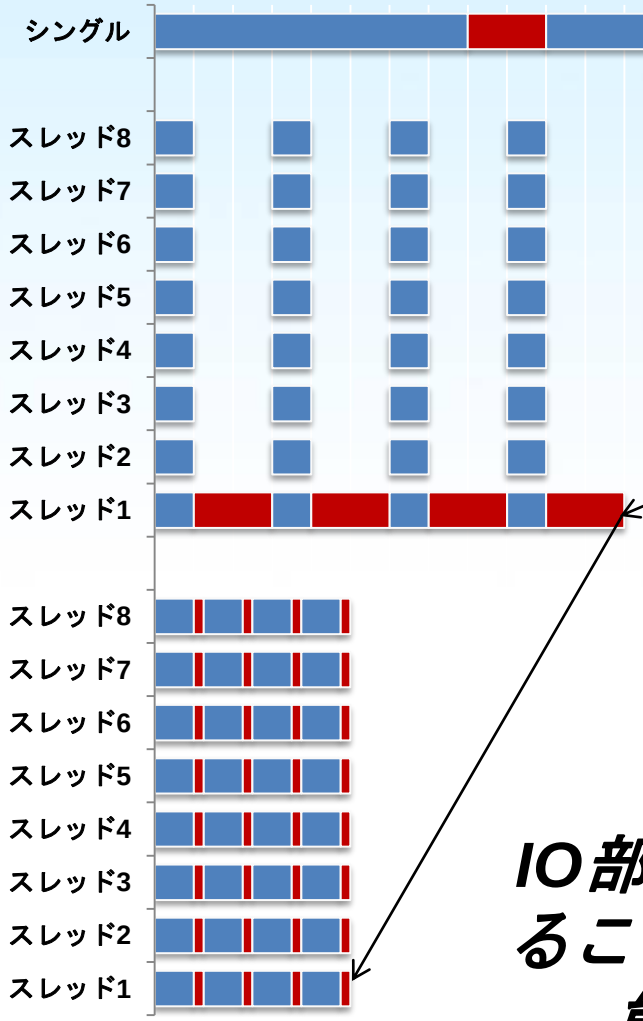
$$Scaling = T_{serial} / T_{parallel}$$

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



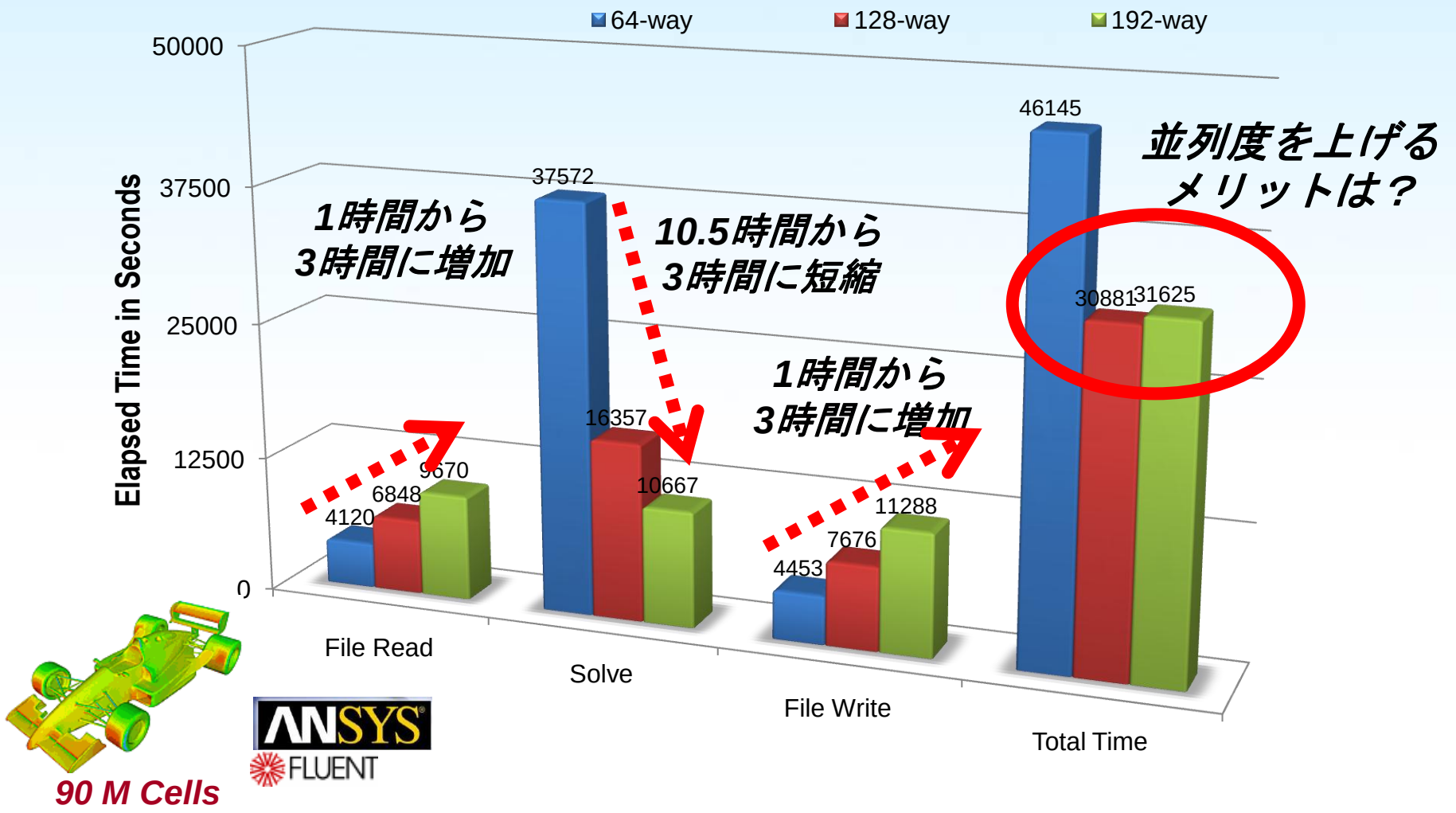
例えば、n=8, P = 0.8 の場合
Scaling = 1.0 / (0.2 + 0.1) = 3.3

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

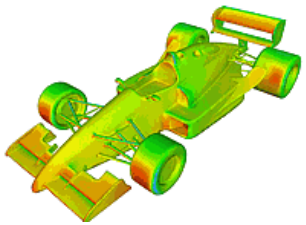
より多くのIOにも対応可能

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



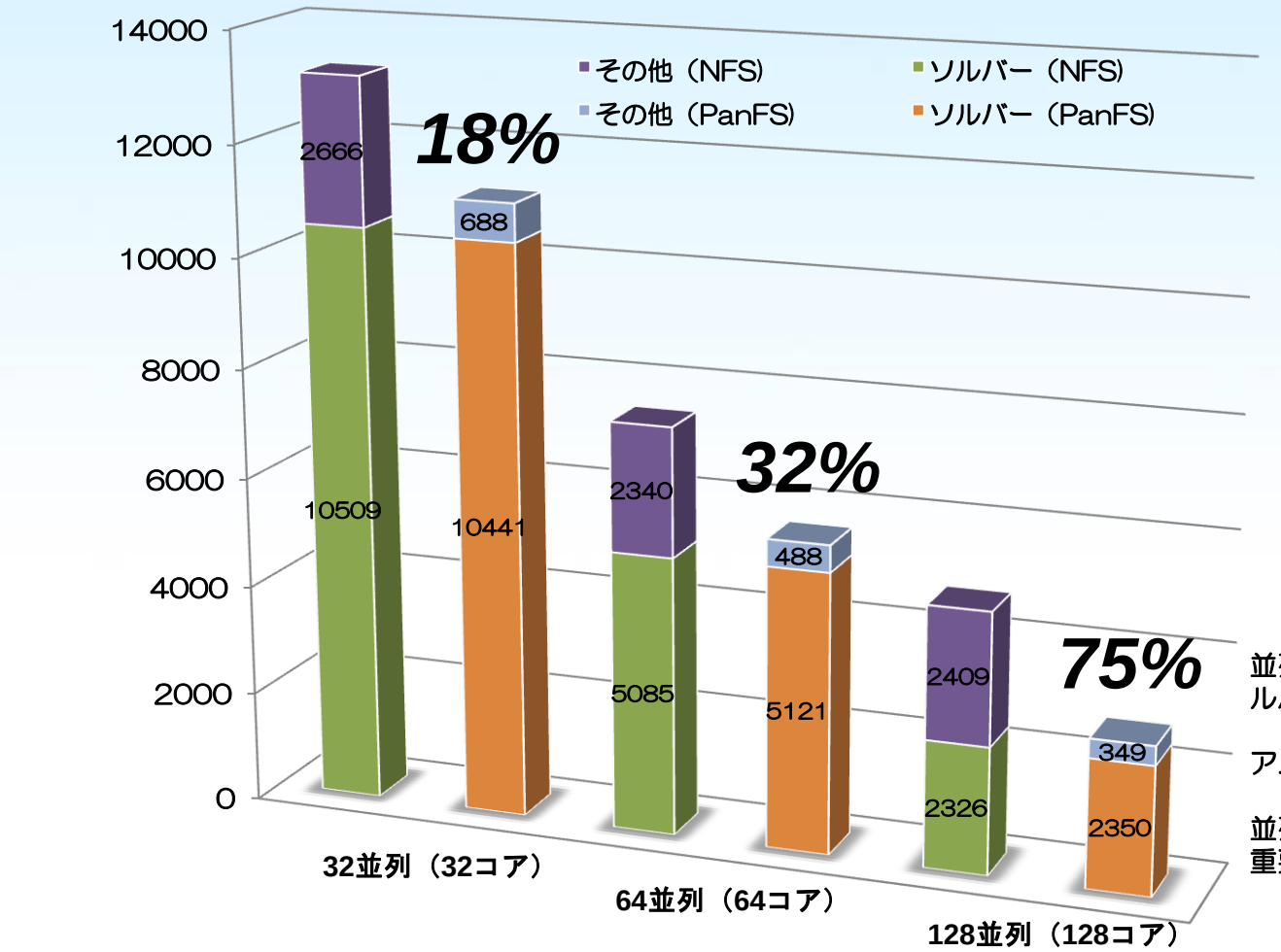
90 M Cells

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

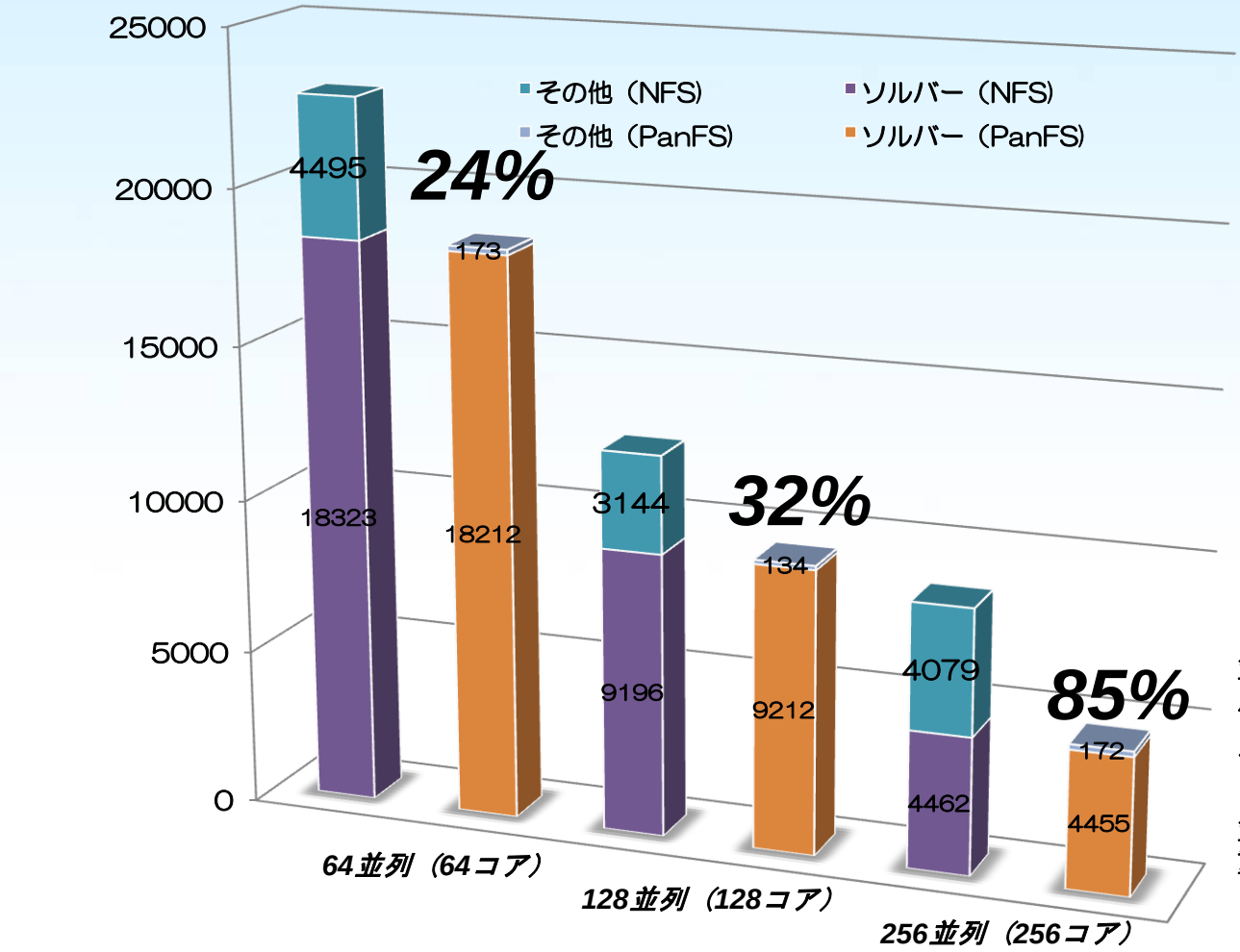


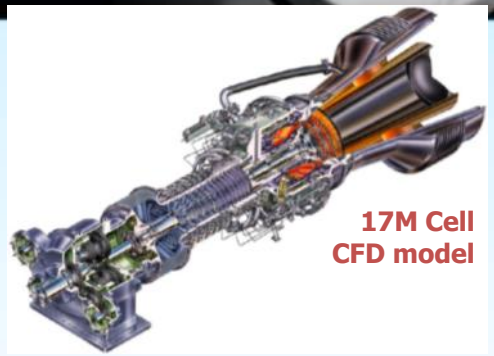
Number of cells
19,921,786
Solver
CGS, Steady
Iterations
500 total iterations - data save after every 10 iters
Each solution output (50 total)
~1,500 MB

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





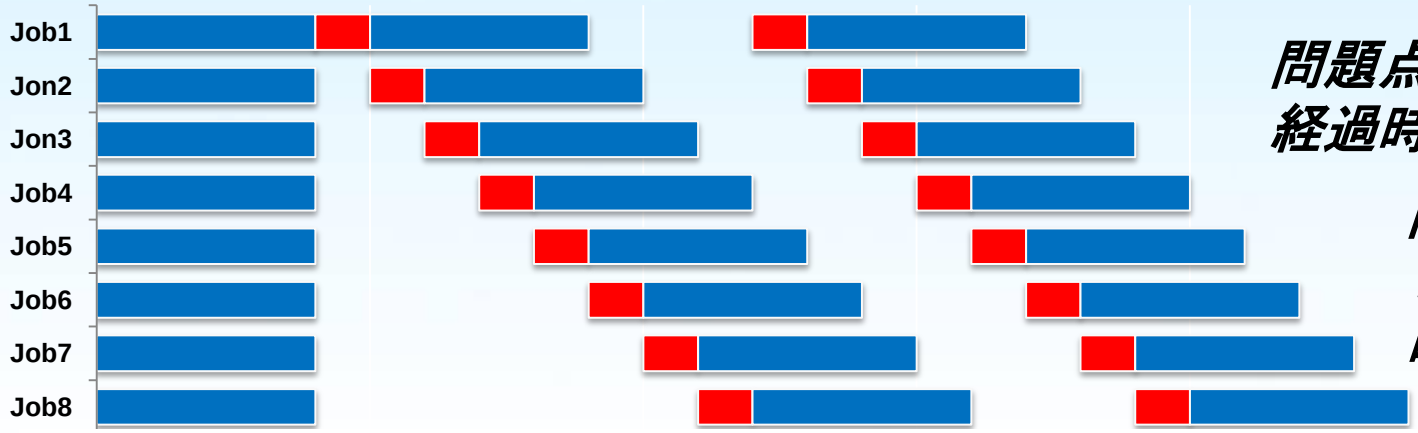
Number of cells
16,930,109
Solver
CGS, Single Precision
Iterations
300 total iterations -
data save after every 100 iters
Total solution output
~48 GB

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

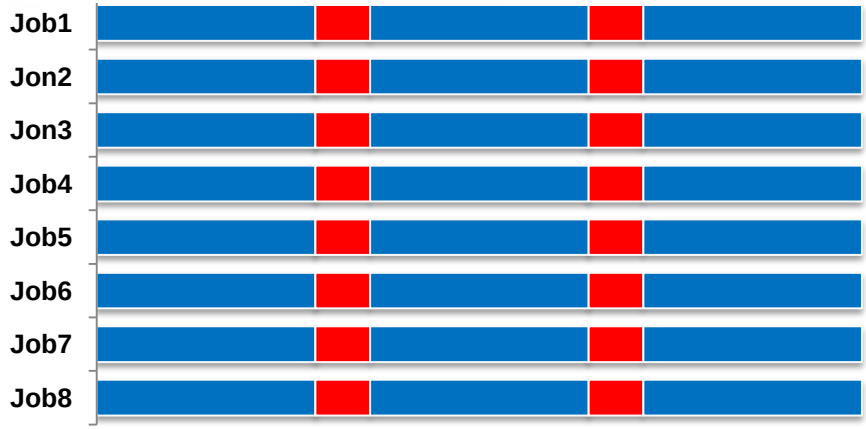
IO処理が逐次的に実行され、ジョブのIO処理時は他のジョブは処理の終了を待つ



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

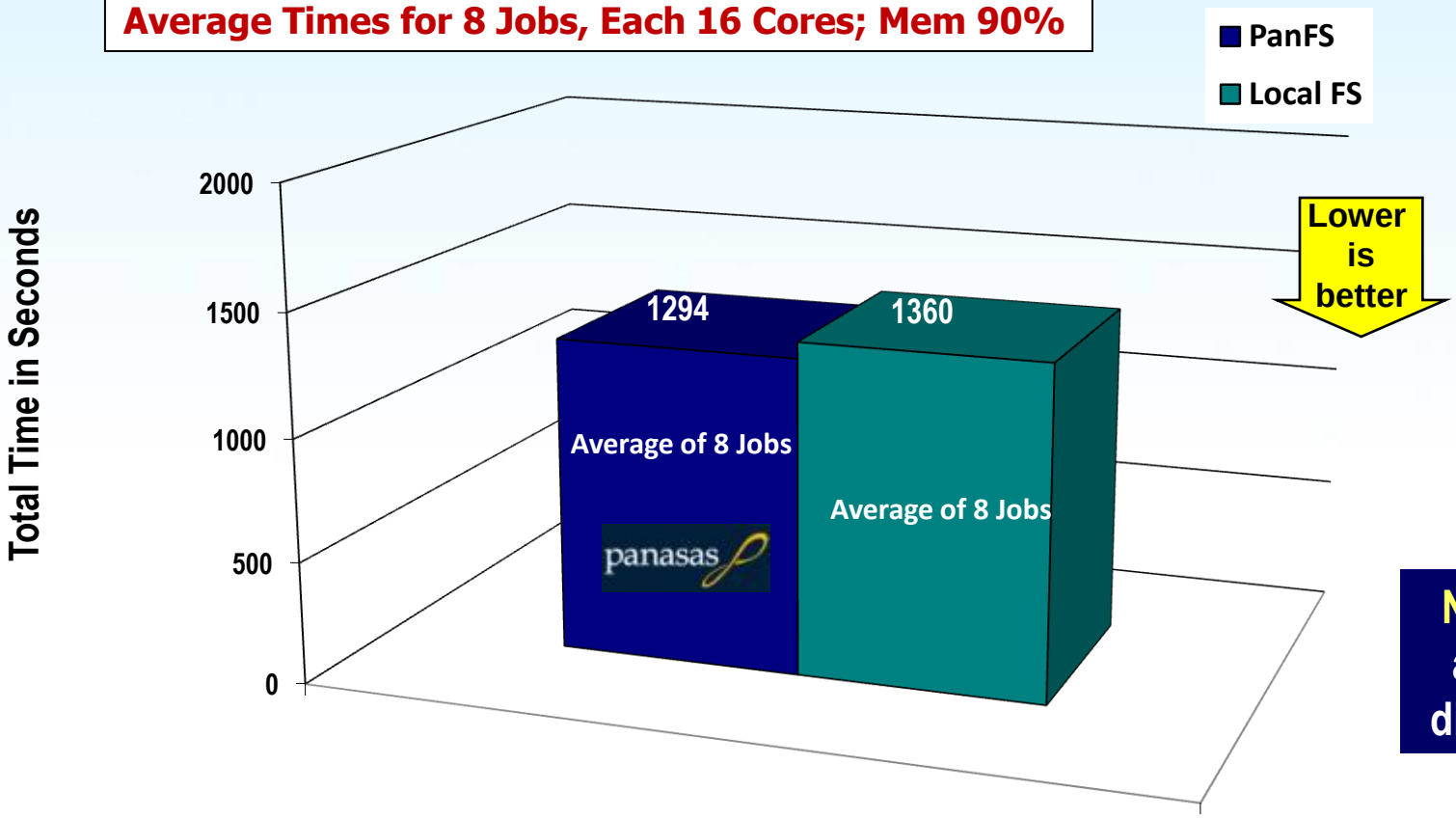
計算処理部分

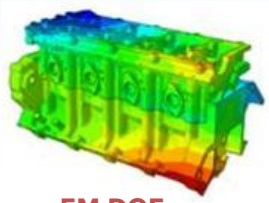
IO処理部分

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%



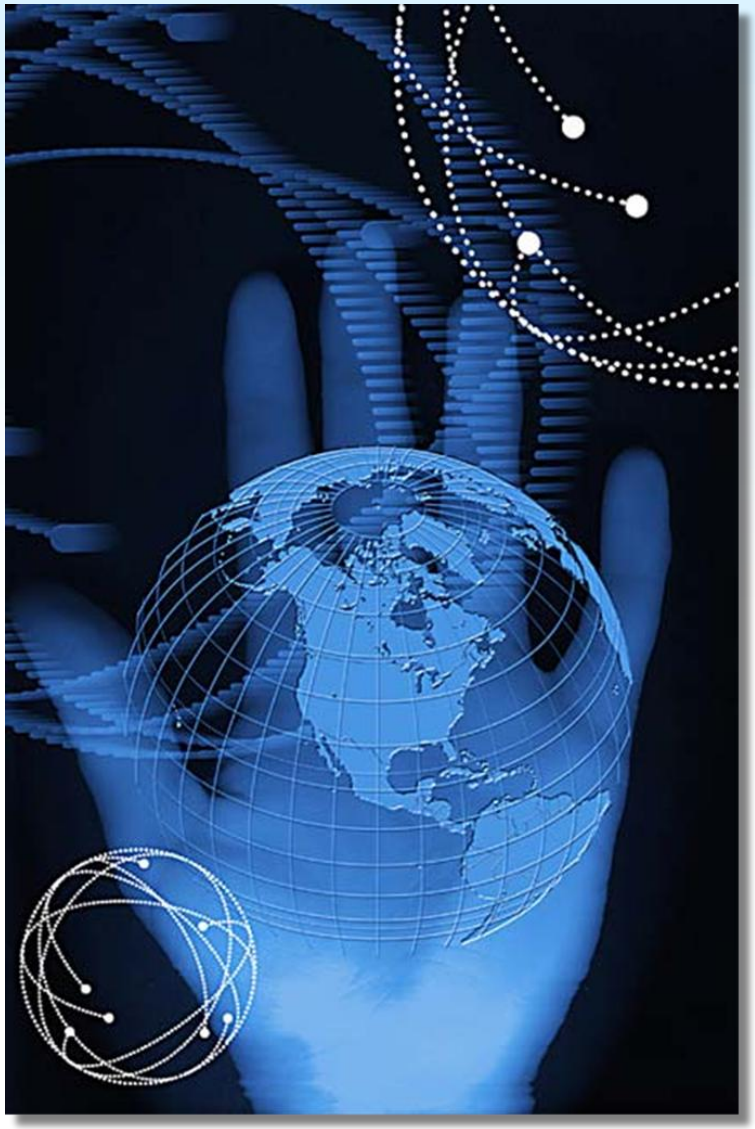


5M DOF
Engine Block

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

12/10/2009

スケーラブルシステムズ株式会社