HITACHI
Inspire the Next

5年後のファイルシステム

戦略的高性能計算システム開発に関する ワークショップ

2010/08/02

株式会社 日立製作所 中央研究所
プラットフォームシステム研究部

松葉　浩也

日立グループ100周年
100th
確かな技術でつぎの100年へ

HITACHI
Inspire the Next

Contents

1. ファイルシステムに関する考察
2. 専門の方への質問

HITACHI
Inspire the Next

1.ファイルシステムに関する考察

1-1　前提(システム規模見積もり)

HITACHI
Inspire the Next

- 現在のマシンの例 (T2K東大)
  - 演算性能:140TFLOPS
  - メモリ量:32TB
  - ストレージ量:1PB(メモリの30倍)
  - ストレージ性能:2GB/s(書き込み・センター広報掲載の推定値)
    - メモリをすべて書き出すのに16000秒

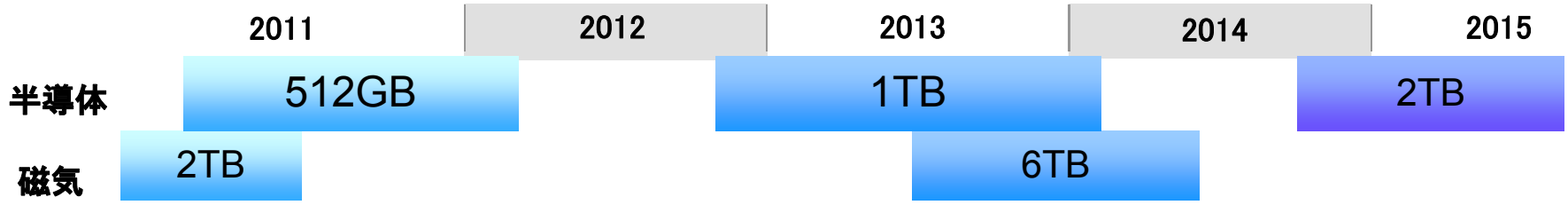
利用形態に大きな変化がないとして単純に規模拡大

- 想定マシン(演算性能:100PFLOPS)
  - (何とかバランスする)最低のメモリ量:20PB
  - (何とかバランスする)最低のストレージ容量:300PB
    - メモリの15倍
  - (何とかバランスする)最低のアクセス性能:15TB/s
    - メモリをすべて書き出すのに約20分

1-2 ストレージデバイスの検討

HITACHI
Inspire the Next

ストレージ技術の未来予測(NEDO技術戦略マップ2009より)



- 300PB
  - SSDなら150k個 → 2Uに256個入れて1200U → 24ラック
  - HDDなら35k個 → 2Uに192個入れて364U → 12ラック
    - ※RAIDを考えるなら単純にRAIDのオーバヘッドを掛け算
- 15TB/s
  - SSD:150k個で15TB/s → 100MB/s/ディスク
    - **十分可能**
  - HDD:35k個で15TB/s → 420MB/s/ディスク
    - ディスク数を2倍にしてSSDと同数のラックまで許しても210MB/s/ディスク
    - **磁気ディスクに何かしらのブレークスルーがない限り不可能**

1-3 ファイルシステムの検討

HITACHI
Inspire the Next

- 現在の方式をそのまま拡張する場合
  - 15TB/s のストレージシステムを支えるファイルサーバ
    - Infiniband HDR x4 が 25GB/s (2014年)
    - IBを2本使用し、ファイルサーバー台あたり30GB/sとするとファイルサーバは500台（10ラック超）
  - ディスク24ラック ＋ ファイルサーバ10ラック ＋ コントローラ ＝ 50ラック程度のストレージシステム

- 各ノードにSSDを入れる場合
  - 1TFlops/node とすると100Pシステムは100kノード
  - 先ほどまでの計算ではSSDは150k個
    - だいたい comparable な数なので、各ノードにSSDを入れるのでも容量、性能のバランスは悪くない
  - Single directory tree を提供する場合、計算中のノードに他のユーザーのI/O処理が紛れ込む
    - 50ラックのストレージシステムとのトレードオフ

1-4 ファイルシステムに関する課題

HITACHI
Inspire the Next

- 信頼性
  - failover機能を持つのは当然だが、本当に確実に動くのか？
  - SSDになるとディスク自体の故障の心配は減る？
- 新たなインタフェースの策定
  - 「UNIXのインタフェース ＋ NFSのような動作」を求められると効率的な実装はかなり困難
    - ほとんどの人には不要なレベルの一貫性維持のために多大なオーバヘッドがかかる
- ステージングとの併用
  - 並列プログラムの実行可能ファイルなど、全ノードが同じファイルを必要とする場合などは、アプリケーション側で効率的な配布をして欲しい

HITACHI
Inspire the Next

2.専門の方への質問

2-1　専門の方への質問

HITACHI
Inspire the Next

- 懇親会のときに議論させてください

- アプリケーション分野の方への質問
  - 夢のある世界を教えてください
    - 「実効××FLOPSあればこんなブレークスルーがある」という話があればぜひ教えてください
      - 例えば、震源から建物の揺れまでを全部シミュレーションするために10PFLOPS必要という話は聞いたことがあります
  - よくない機械・ソフトの例を教えてください
    - 「計算機屋が好きなことをしているだけで使い物にならない」と感じた経験など

- 計算機の専門家の皆様に質問（提言）
  - 「××FLOPS」に代わる指標を使いませんか？
    - FLOPSよりも、メモリ、ネットワークがネックになる時代
    - ボトルネックとなる部分の数字を比較するのが技術的には正しい

HITACHI
Inspire the Next

END

2010/00/00

株式会社 日立製作所　中央研究所
プラットフォームシステム研究部

松葉　浩也

HITACHI

Inspire the Next