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＠金沢
2010年8月2日

T2Kシステム5年後

建部修見（筑波大）

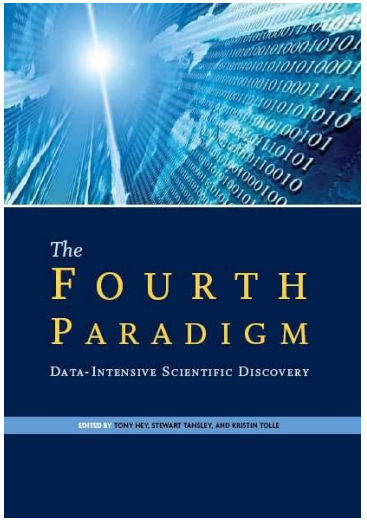
T2Kシステムについて

- H20年6月導入
  - 筑波95TFlops，東大計140TFlops，京大61TFlops+SMP
- 同一基本仕様
  - X86_64, 16CPUコア/ノード，メモリバンド幅40GB/s
  - 250GB RAID-1/ノード，IPMI2.0
  - ノード間ネットワークバンド幅5GB/s
  - MPI1.2, MPI遅延8.5μ秒，MPIバンド幅4GB/s
  - 64bit Linux, 自動並列化コンパイラ，OpenMP, Java
  - BLAS, LAPACK, ScaLAPACK
- 調達型から開発型へ

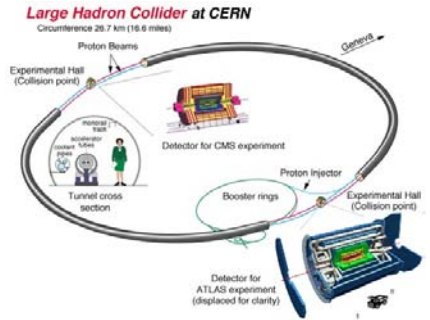
次期システム

- 同一基本仕様？
- エクサフロップマシンにむけて
- 計算科学とeサイエンス
- （日本）ベンダによる競争

The 4th paradigm: data intensive scientific discovery (e-Science)



- 実験、理論、計算に次ぐ第4の科学
- FLOPSだけではなくBytesとBytes/sec
- 研究コミュニティによる大規模科学
- データセントリックサイエンス
  - 大規模データによりすすめられる科学
  - 大規模データの研究コミュニティでの共有
    - CERN LHCコラボレーション、KEK Belleコラボレーション
    - International Lattice Data Grid (ILDG)
    - International Virtual Observatory (IVOA)



データセンタにおけるシステム開発

Sun Data Center, jp.sun.com

- 超巨大データセンタへ資源の集約
- スケールアウトするシステム開発
  - 数千台～数十万台（数百万コア）
  - Google FS、Dynamo、BigTable、Gfarm FS
- 並列言語処理系の開発
  - MapReduce、Sawzall、PIG
  - GXP Make、Pwrake
- MPI、MPI-IO、並列ファイルシステムとは全く異なるアプローチ

三つのシステム

- サブエクサフロップスA
- サブエクサフロップスB
  - 100ギガバイト/秒
- サブエクサバイトC
  - 10テラバイト/秒
- エクサバイト広域ファイルシステム
  - センタ間10ギガバイト/秒

サブエクサバイトC

- 目標：1000ノード並列アクセス，100 PB，10 TB/sのファイルシステム
- 計算ノードのローカルディスクを有効利用
  - ローカルディスクの仮想化と局所性利用の両立
- (2TB SSD x 8, SAS/SATA 12Gbps) x 16
  - 256 TB, 19.2 GB/s
- 1000ノード
  - 256 PB, 19.2 TB/s

エクサバイト広域ファイルシステム

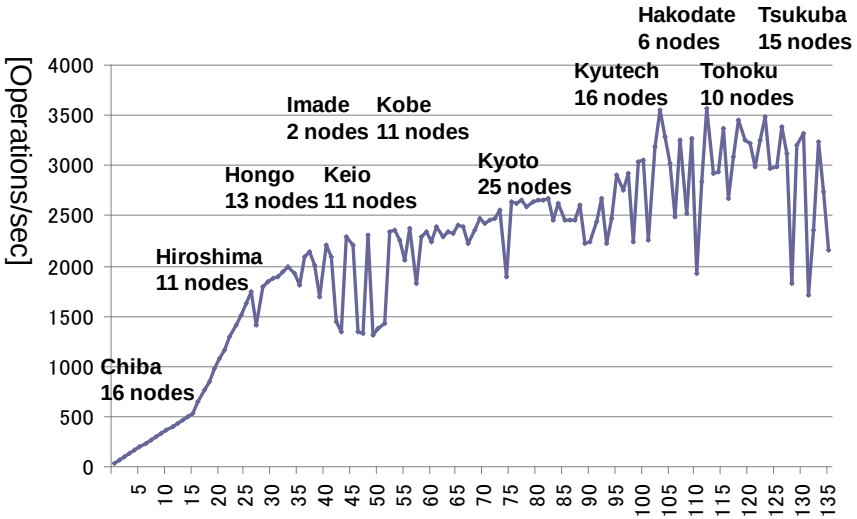
- センタ間のデータ共有以上のもの
  - ログインノードでコピー・ステージングは×
  - ファイル複製10ギガバイト/秒
- 計算ノードからの高速アクセス
  - 計算ノードでマウント
  - 100ギガバイト/秒～10テラバイト/秒
- アプリの性能，利用負荷に応じたサイト選び
- 大規模広域分散eサイエンス
  - エクサバイトデータインテンシブコンピューティング

Shameless plug: InTriggerでのGfarm 広域ファイルシステムの性能評価

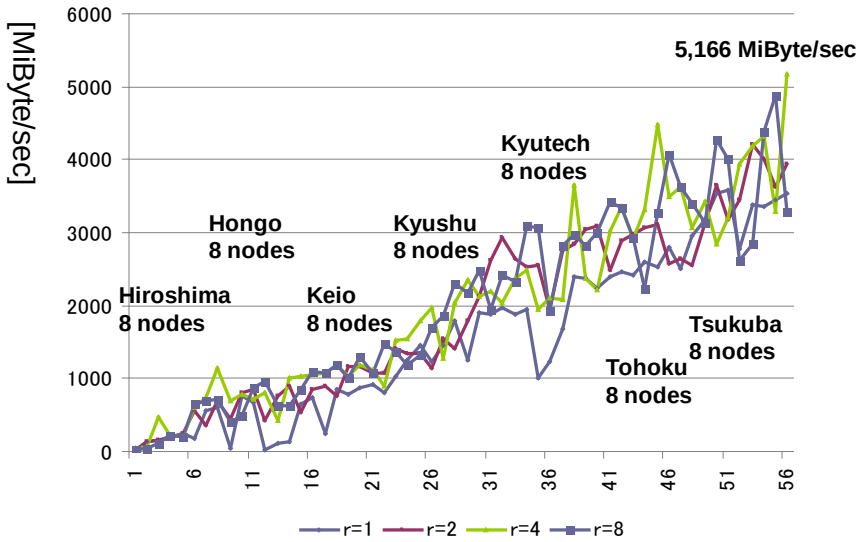
- >150TByte（>200ノード，>14拠点）の広域ファイルシステム

(New Generation Computing, 2010, to appear)

メタデータ操作 3,570 ops/sec



共有データ読込 5,133 MB/sec



プログラミング言語処理系

- MPI-IO, HDF5, PnetCDF
- MapReduce, Sawzall, PIG
- メニーコアにむけた新しいプログラミング言語
  - より関数型: Fortress, Scala
  - 近代CSP: go
- 大規模ワークフロー処理系
  - 複数プログラムの組合せ
  - データアクセス局所性，データ移動最小化
  - Pwrake

まとめとして

- サブエクサバイトCの構築
  - 10テラバイト/秒
- エクサバイト広域ファイルシステム
  - サイト内100ギガバイト/秒～10テラバイト/秒
  - サイト間10ギガバイト/秒
- 第3，第4の科学の推進
- システムを見据えたプログラミング言語処理系の研究開発