東京大学情報基盤センター
大規模データ解析基盤運用に向けて

東京大学
石川　裕

© Yutaka Ishikawa

2010/10/05

再配布･部分転載不可

はじめに

- 東京大学グローバルCOE「ゲノム情報ビッグバンから読み解く生命圏」公開シンポジウムにおける石川の発表「大規模データ解析のための高性能並列システム環境構築に向けて」において、

おわりに

- I/Oネットワーク、ストレージ性能向上、信頼性向上のために物量を投入するのはコストがかかる
  - システムソフトウェア＆アプリケーションによる最適I/O処理を実現していかなければならない
- 大規模データ解析のための学際研究
  - 学際大規模情報基盤共同利用共同研究拠点で一緒にやりましょう
  - T2K東大版ファイル性能にまだご不満があるかと思いますが、将来の大規模データ解析のための高性能並列システム開発に向けて協力していきたい

2010/04/10 The University of Tokyo 21

© Yutaka Ishikawa 再配布･部分転載不可

Outline

- 学際大規模情報基盤共同利用共同研究拠点の紹介
- 情報基盤センタースパコンの紹介
- 「e-サイエンス実現のためのシステム統合・連携ソフトウェアの高度利用促進」の紹介
- ゲノム情報解析系性能向上の取り組み紹介

© Yutaka Ishikawa
再配布・部分転載不可

学際大規模情報基盤共同利用共同研究拠点

詳しくは、http://jhpcn-kyoten.itc.u-tokyo.ac.jp/

- 北海道大学，東北大学，東京大学，東京工業大学，名古屋大学，京都大学，大阪大学，九州大学ににそれぞれ附置するスーパーコンピュータを持つ8つの共同利用の施設を構成拠点とし，東京大学情報基盤センターがその中核拠点として機能する「ネットワーク型」共同利用・共同研究拠点（大学の研究ポテンシャルを活用し、研究者が共同で研究を行う）
- 目的
  - これまでに解決や解明が極めて困難とされてきた，いわゆるグランドチャレンジ的な問題について，学際的な共同利用・共同研究を公募し実施する
    - 超大規模数値計算系応用分野
      地球環境，エネルギー，物質材料等の科学・工学分野における科学技術シミュレーション，および，それを支えるモデリング，数値解析アルゴリズム，可視化手法，情報基盤等
    - 超大規模データ処理系応用分野
      - ゲノム情報，Webデータ（Wikipedia，ニュース，ブログなどを含む），学術情報コンテンツ，センサーネットワークからの時系列データ，映像を始めとするストリームデータなどに対する高度なメディア情報処理，プログラム解析，アクセスや検索，情報抽出，統計的・意味的分析，データマイニング等

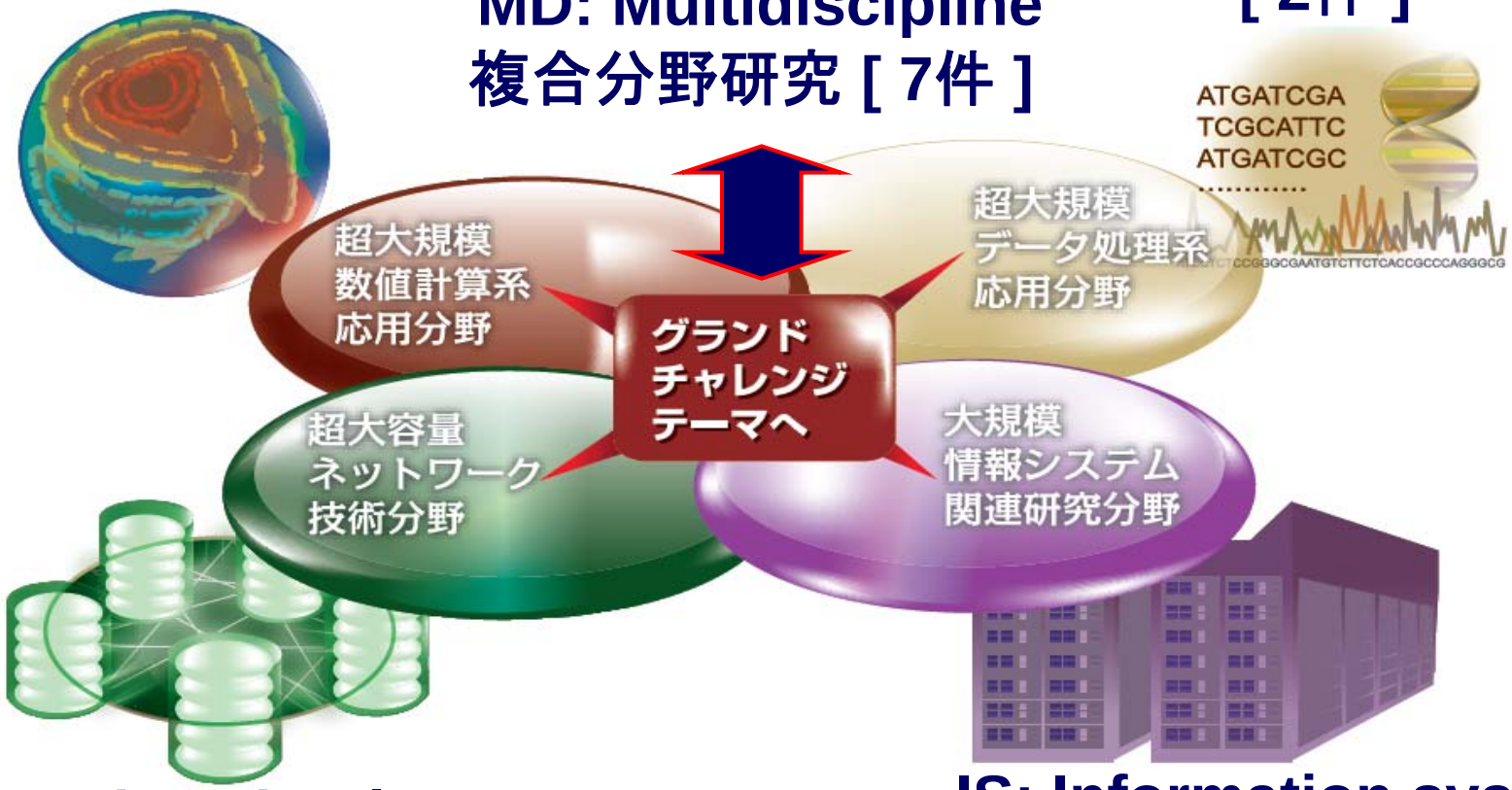
© Yutaka Ishikawa
再配布・部分転載不可

平成22年度採択状況

NA: Numerical analysis
[23件]

DA: Data analysis
[2件]

MD: Multidiscipline
複合分野研究 [7件]



NW: Network technology
[0件]

IS: Information system
[5件]

© Yutaka Ishikawa
再配布･部分転載不可

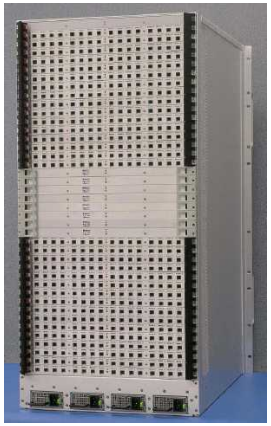
東京大学情報基盤センター　スパコン

HITACHI SR11000 model J2

Total Peak performance	: 18.8 TFLOPS
Total number of nodes	: 128
Total memory	: 16384 GB
Peak performance per node	: 147.2 GFLOPS
Main memory per node	: 128 GB
Disk capacity	: 94.2 TB

T2K東大　　HA8000-tc/RS425 x 952

Total Peak performance	: 140 TFLOPS
Total number of nodes	: 952
Total memory	: 32000 GB
Peak performance per node	: 147.2 GFLOPS
Main memory per node	: 32 GB, 128 GB
Disk capacity	: 1 PB



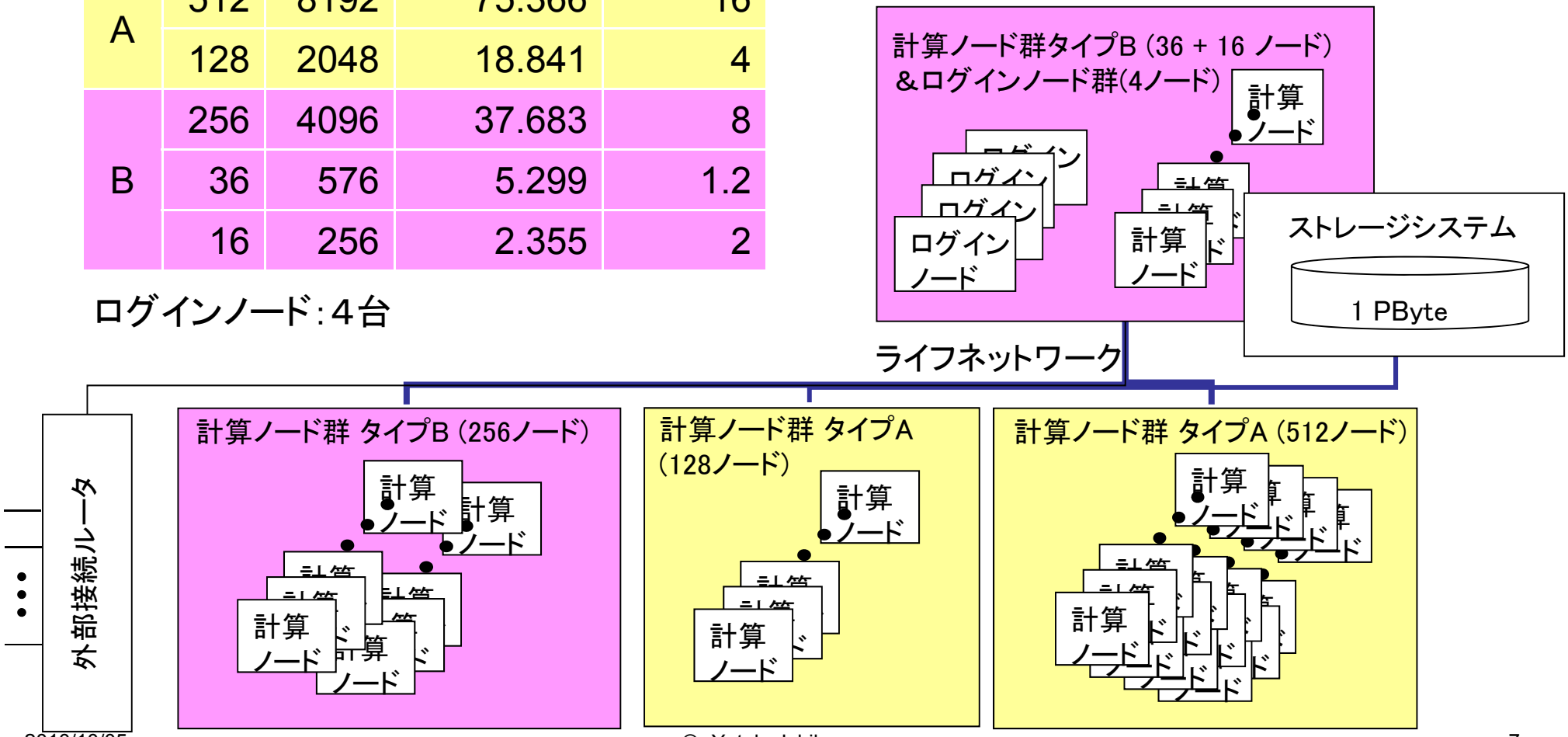
© Yutaka Ishikawa　再配布･部分転載不可

T2Kオープンスパコン東大版 HA8000-tc/RS425クラスタ群

タイプ	ノード数	CPU数	理論演算性能 (TFlops)	総主記憶 (TB)
A	512	8192	75.366	16
	128	2048	18.841	4
B	256	4096	37.683	8
	36	576	5.299	1.2
	16	256	2.355	2

ログインノード：4台

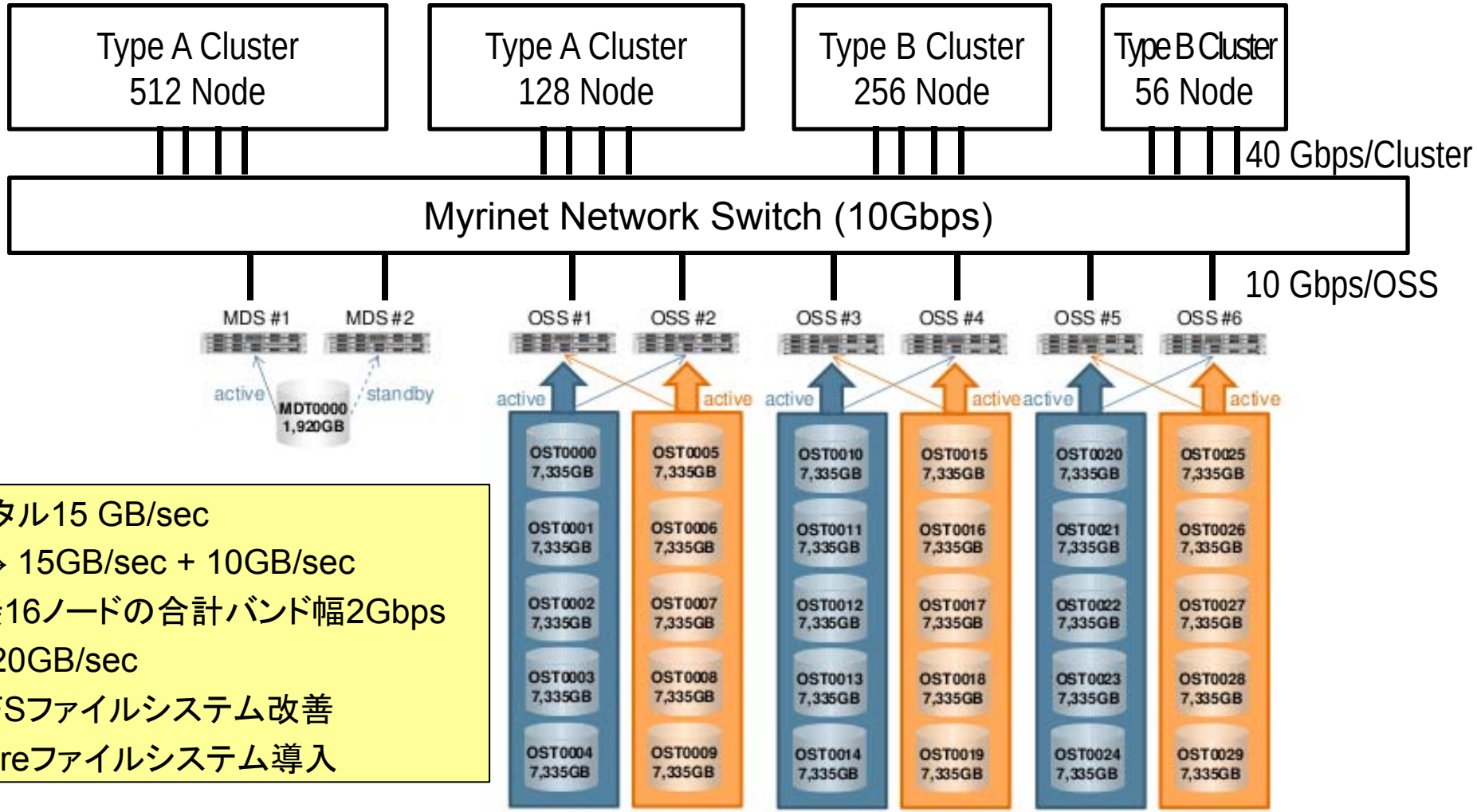
計算ノード間ネットワーク性能
タイプA　5GB/sec (片方向通信)
タイプB　2.5 GB/sec (片方向通信)



© Yutaka Ishikawa
再配布･部分転載不可

T2K 東大版における現在のファイルシステム構成

ネットワーク増強＆追加ファイルシステム (Lustreファイルシステム)



- トータル15 GB/sec
  → 15GB/sec + 10GB/sec
- 隣接16ノードの合計バンド幅2Gbps
  → 20GB/sec
- HSFSファイルシステム改善
- Lustreファイルシステム導入

再配布･部分転載不可

e−サイエンス実現のためのシステム統合・連携ソフトウェアの高度利用促進

初年度所要額：30億円

出典：文部科学省http://www.mext.go.jp/b_menu/houdou/22/06/1295019.htm

※次年度以降は初年度の実績を踏まえ継続の可否を判断

・我が国の科学技術をより強化し、研究分野や国・地域を越えた連携を推進するためには、『e-サイエンス』（インターネットを活用して計算、実験、観測結果等の巨大データや計算資源を共用するという新しい科学の方法論）へのパラダイムシフトを図ることが必要であり、e-サイエンスを支える世界トップレベルの最先端研究情報基盤の整備充実が不可欠。

・そこで、ユーザーニーズに応じて、PCクラスタから基盤センターに設置されているスーパーコンピュータ、さらには次世代スーパコンピュータをシームレスかつオープンに利用できるコンピューティング環境を実現し、国際的に競争環境にある大規模コンピューティング環境分野において、世界に先駆けて革新的な大規模並列計算および大規模データ解析の研究拠点を構築する。

支援対象機関：

・東京大学 （東京大学を中心とした大学の情報基盤センター等の機関がネットワーク型の拠点を形成。理化学研究所や国立情報学研究所も連携）

・東大情報基盤センターは、全国8大学の情報基盤センターから構成される『学際大規模情報基盤共同利用・共同研究拠点』の中核拠点

・これまでも全国の大学研究者等と共同研究の実績

既存事業の具体的内容

（『e-サイエンス実現のためのシステム統合・連携ソフトウェアの研究開発』：H20~23年）

PCクラスタから大学情報基盤センター等に設置されているスパコンまでシームレスに接続するコンピューティング環境の構築

①大量データ高速転送機構

②分散環境におけるファイルシステム

③並列プログラミング言語機能

新たな課題

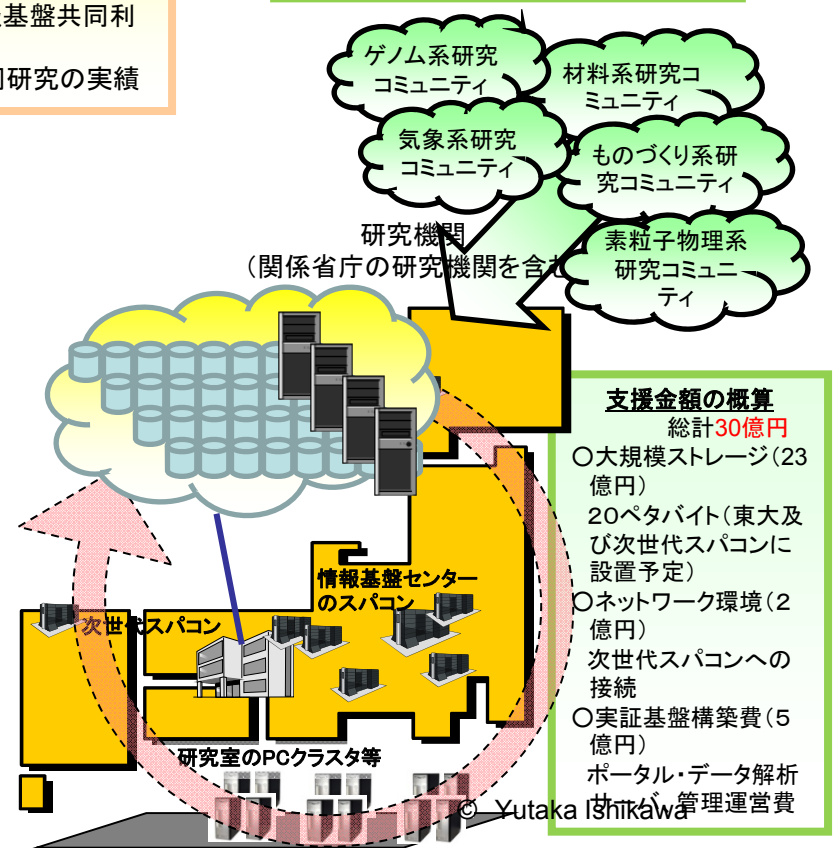
①増大する大容量データへの対応（現行4PB程度）

②次世代スパコンなど新たなコンピューティング環境への応用

本施策の内容

① 20ペタバイトクラスの大容量ストレージを設置

②次世代スパコン等との接続



支援金額の概算

総計30億円

○大規模ストレージ（23億円）

20ペタバイト（東大及び次世代スパコンに設置予定）

○ネットワーク環境（2億円）

次世代スパコンへの接続

○実証基盤構築費（5億円）

ポータル・データ解析サーバ、管理運営費

拠点機能の強化

超膨大量の科学計算が可能な大規模オープンコンピューティング環境を次世代スパコン完成の平成24年度までに整備。世界トップクラスの研究者が集積する国際的頭脳循環拠点が形成。

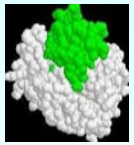
ネットワーク接続環境下での計算が可能

次世代スパコンを含めシームレスに接続

創出される研究成果等

・分散していた地球環境分野の超大容量データ（地質データ、気象データ、衛星画像データ等）を同一ストレージ上で有機的に結合することで超高精細かつ高精度な地球温暖化シミュレーションが実現。（平成27年度まで）

・複数研究機関が保有するゲノム情報を一括参照することにより、超大規模ゲノム解析を可能にし、医学上の新たな発見や革新的新薬の開発が加速。（平成27年度まで）



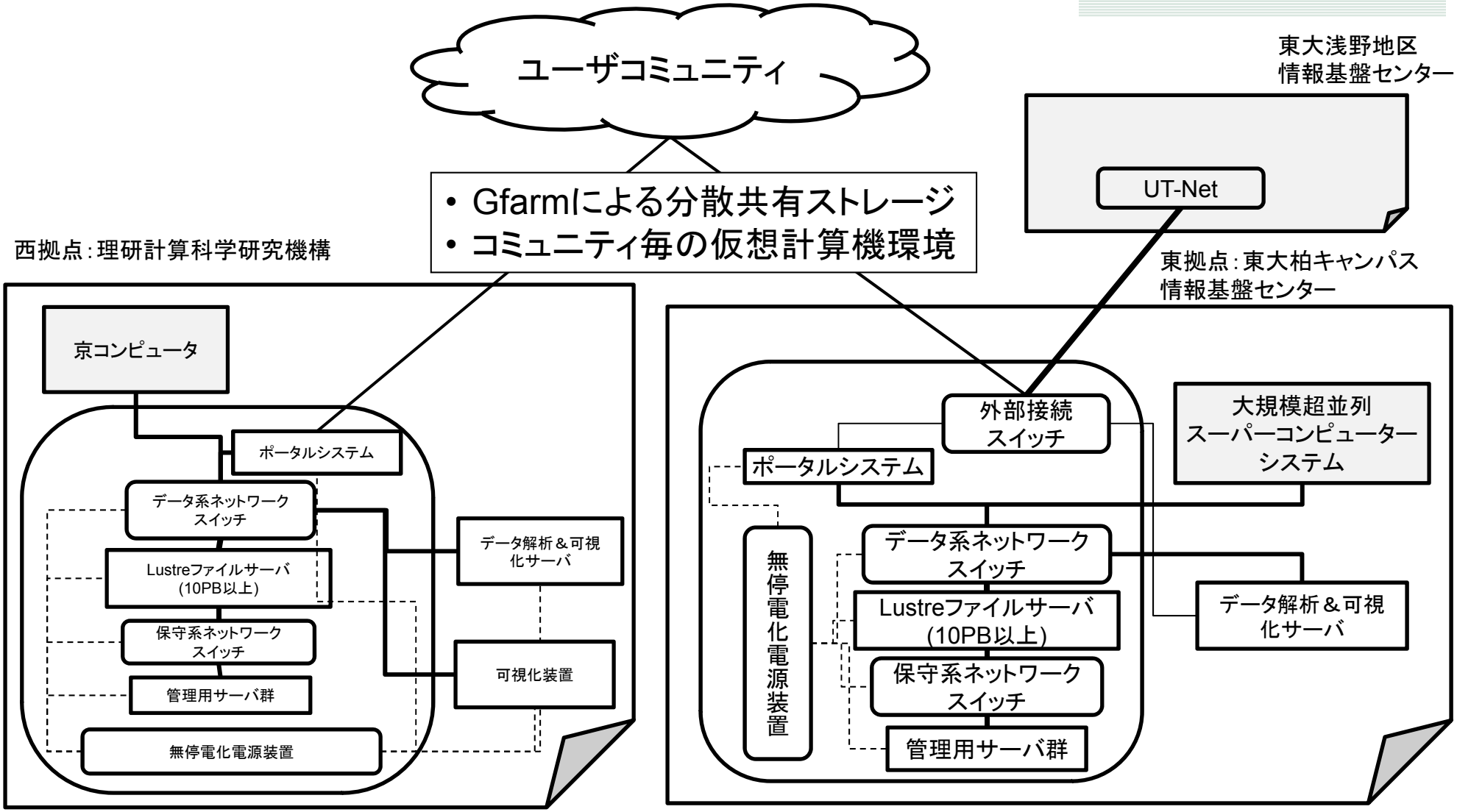
・平成22年度から推進する革新的ハイパフォーマンス・コンピューティング・インフラ（HPCI）構築への寄与。

・計算機資源の優先的利用枠の創設や、インターネット経由でのアクセス環境構築により、若手・女性研究者の研究環境が改善。

・全国各地において科学データの高度利活用が可能となり、地域性を活かした研究開発が促進。

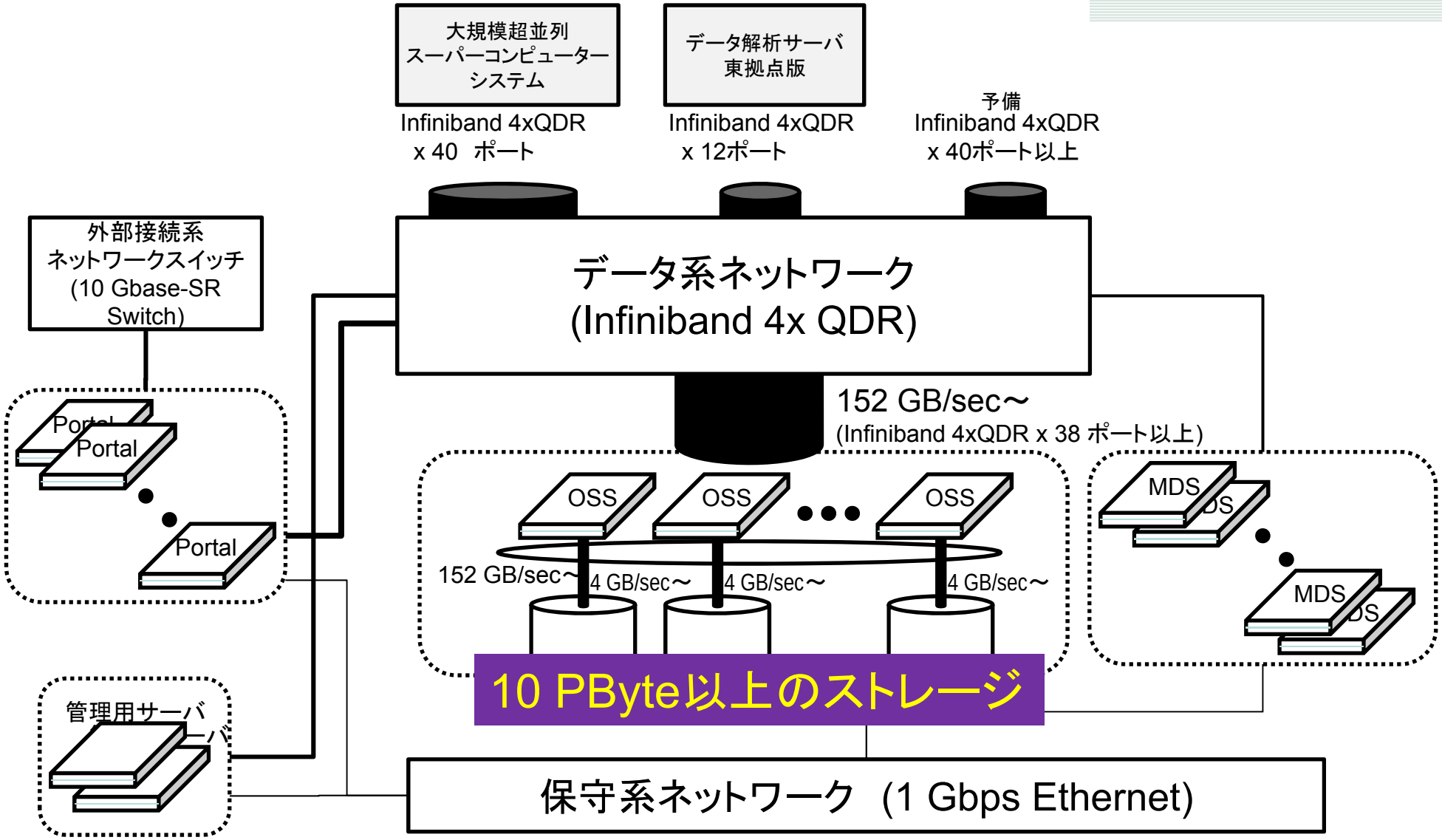
© Yutaka Ishikawa
再配布・部分転載不可

全体概念図



© Yutaka Ishikawa
再配布･部分転載不可

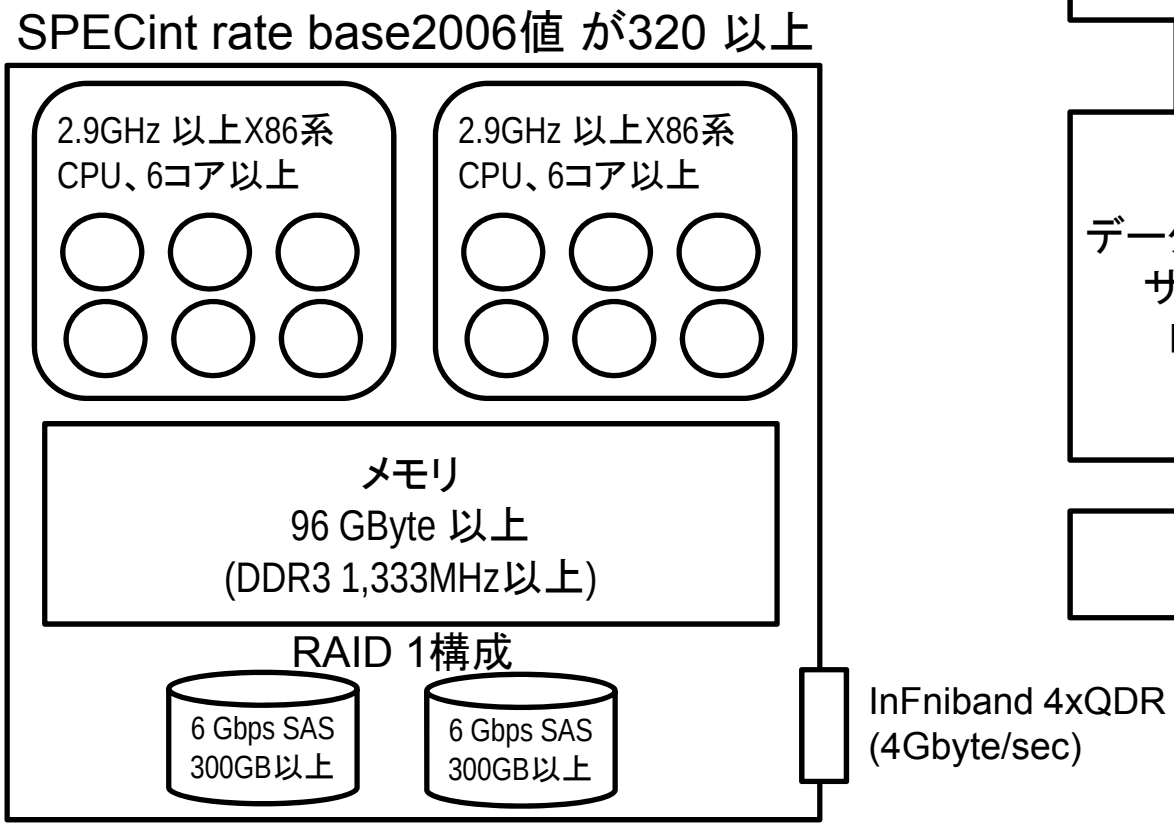
ストレージ＆ポータルシステム東拠点版の概要

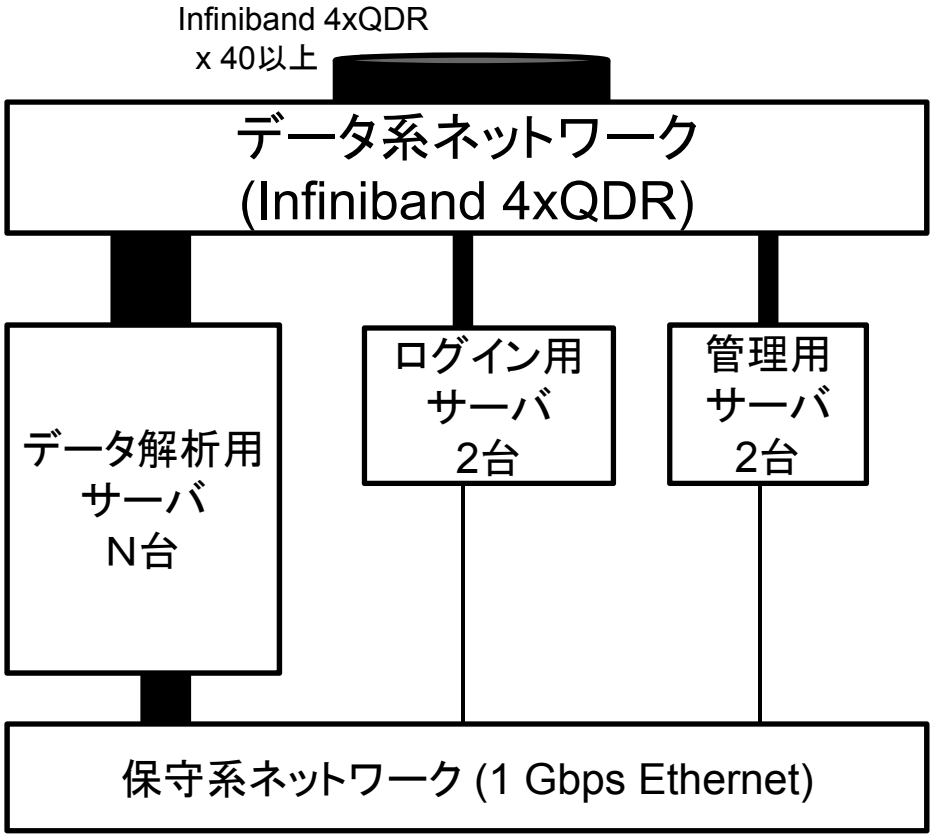


© Yutaka Ishikawa
再配布･部分転載不可

データ解析＆可視化システム東拠点版の概要

- 全体でSPECint rate base2006 値が最低25,600
- 各ノード性能が320であれば、80ノード





© Yutaka Ishikawa
再配布･部分転載不可

運用

- 運用ポリシー
  - データ解析ジョブを優先
  - 負担金は、電気代＆保守費用（詳細は年内に決定予定）
  - 若手・女性研究者に対し、負担金はとらない
  - 試験運用中は、負担金を取らない

- スケジュール
  - 2011年1月中旬　データ解析＆可視化システム東拠点版納入
  - 2011年2月データ解析＆可視化システム東拠点版試験運用
  - 2011年3月中旬　ストレージ＆ポータルシステム東拠点版納入
  - 2011年3月中旬～ストレージ＆ポータルシステム東拠点版試験運用
  - 2011年4月～本格運用

© Yutaka Ishikawa
再配布･部分転載不可

Burrows-Wheeler Alignment (BWA) Toolの性能

- BWAに含まれる処理
  - ゲノムアラインメント処理 (並列処理)
  - フォーマット変換
- 実験環境
  - Intel Xeon X5670 2.93GHz, 48GB Memory
  - HDD
    - FUJITSU MBD2300RC 300GB SAS 16M 10025rpm
    - Read: 104 MB/s Write: 88.3 MB/s
      (RAID1 PERC6/i 256MB)
  - SSD OCZ
    - OCZSSD2-2VTXP200G 200GB
    - Read: 242 MB/s Write: 208 MB/s
  - SSD GS
    - FM-25S2S-240GBP2 240GB
    - Read: 133 MB/s Write: 132 MB/s
  - 解析データ: 約20GB (SRR029655_1.filt.fastq)
    - 読み込み200秒 in HDD

アラインメント処理 (15 threads) Sec

	HDD	SSD OCZ	SSD GS
Real	639	572	594
User	723	731	709
Sys	4663	4522	4443

フォーマット変換 (1 thread) Sec

	HDD	SSD OCZ	SSD GS
Real	1593	1367	1515
User	287	259	266
Sys	1111	1054	1127

Real : 実際の時間
User : ユーザレベルCPU時間 (全スレッド合計)
Sys: OS処理時間 (全スレッド合計)

By 情報基盤センター 特任助教 山本啓二

© Yutaka Ishikawa
再配布･部分転載不可

Burrows-Wheeler Alignment (BWA) Toolの性能改善

1. 無駄なメモリアクセス、I/Oの削減
   - 同じファイルから同じデータを読み込み
     - 1GByteのファイルを約1000回読み込む
   - メモリマップドファイルI/Oにより読み込み削除
2. I/Oボトルネック
   - ファイル読み込み時のバッファサイズが小さく、readシステムコールが多数発行されている
     - I/Oバッファサイズを大きくする
3. 同期処理
   - 頻繁にクリティカルセクションが存在する
   - 並列度を増やしてもロックの取り合いで性能が上がっていない
     - 逆に、ロックの取り合いで性能低下する場合もあった
   - 前処理できる部分を処理してクリティカルセクションを削減

アラインメント処理 (15 threads) Sec

	HDD	SSD OCZ	SSD GS
Real	639	572	594
User	723	731	709
Sys	4663	4522	4443

フォーマット変換 Sec

	HDD	SSD OCZ	SSD GS
Real	1593	1367	1515
User	287	259	266
Sys	1111	1054	1127

Real : 実際の時間
User : ユーザレベルCPU時間 (全スレッド合計)
Sys: OS処理時間 (全スレッド合計)

By 情報基盤センター 特任助教 山本啓二

© Yutaka Ishikawa
再配布･部分転載不可

Burrows-Wheeler Alignment (BWA) Toolの性能改善

アラインメント処理 (15 threads)　　Sec

	HDD	SSD OCZ	SSD GS
Real	639	572	594
User	723	731	709
Sys	4663	4522	4443

頻繁なロックの取り合い

2～2.3倍の高速化

	HDD	SSD OCZ	SSD GS
Real	305	243	279
User	436	390	422
Sys	16	14	15

10分から5分に短縮

ディスクの性能が2倍速くても全体性能は10%～20%しか高速化しない
- HDD (Read: 104 MB/s Write: 88.3 MB/s)
- SSD OCZ (Read: 242 MB/s Write: 208 MB/s)
- SSD GS (Read: 133 MB/s Write: 132 MB/s

フォーマット変換 (1 thread)　　Sec

	HDD	SSD OCZ	SSD GS
Real	1593	1367	1515
User	287	259	266
Sys	1111	1054	1127

無駄なファイル読み込み(約1TB)

2.2～2.7倍の高速化

	HDD	SSD OCZ	SSD GS
Real	699	505	616
User	330	311	356
Sys	118	115	120

26分から11分に短縮

By 情報基盤センター　特任助教　山本啓二

© Yutaka Ishikawa
再配布･部分転載不可

おわりに

- データ解析サーバ＆大規模ストレージは、若手・女性研究者支援として優先利用＆負担金免除としていく予定です
- 今後の情報は、
  - http://www.itc.u-tokyo.ac.jp/

  をご覧ください
- 「10年先のデータベースを考える」というテーマに関して
  - アプリケーションプログラム開発者との共同研究を通して、プログラムの高度化を進めていかなければならない
  - レガシコードも書き直さないと。。。
  - プログラム開発者の情報教育が重要 （単なるプログラミングではなく計算機システムの本質を理解すること）

© Yutaka Ishikawa
再配布･部分転載不可