ハードウェアによる様々な構造型ストレージの高速化

松谷 宏紀
慶應義塾大学 理工学部、
JSTさきがけ、NII

自己紹介: 松谷 宏紀

学部@慶應大学SFC
- インターネットの研究
- IPv6 / Mobile IPv6
- 組込み機器向け IPv6 スタック

ポスドク(SPD)@東京大学
- Network-on-Chip(NoC)
- 3次元積層
- 65nm チップ試作
- 徐々に回路寄りになる

2000 2003 2004 2009 2011 2014

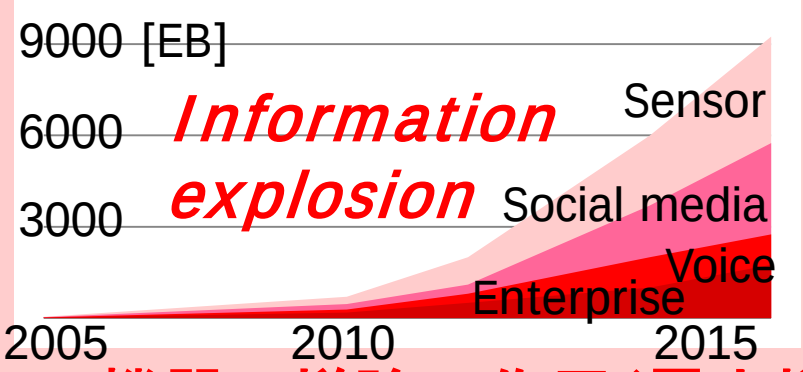
修士・博士@慶應大学
- 計算機 HW の研究
- Network-on-Chip(NoC)
- トポロジ・ルーティング
- やや理論的だった

教員@慶應大学
- フルカスタム回路設計
- 反動(?)でビッグデータに興味を持つ
- 全国大会に呼ばれる!

ICT におけるトレンド: ビッグデータとグリーン化

Big data: the next oil

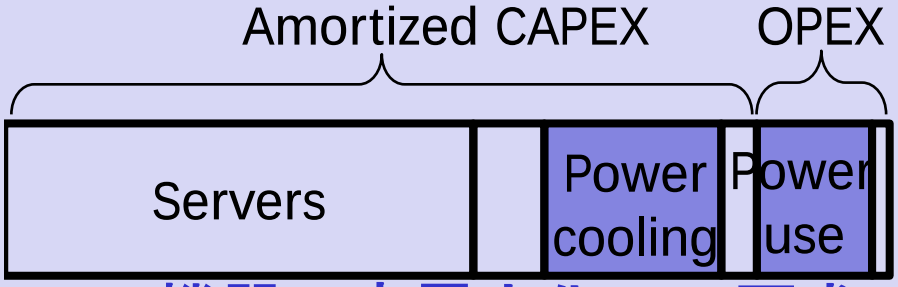
データの蓄積・利活用によって
さまざまなイノベーションが期待



→ IT 機器の増強へ作用(電力増)

Green datacenters

地球温暖化防止の観点、経済面(データセンター運用コスト)から消費電力の削減は必須



→ IT 機器の省電力化への要求

- IT 機器の省電力化をこれまで以上に推し進めなければ、電力がビッグデータ利活用の**大きな足かせ**になる
- 制限: IT 機器の省電力化はすでにやり尽くされている
  - データセンターでは、コモディティ機(コスト効率重視)が多用
  - そもそも回路の電源電圧はもう下げられない

今こそ、計算機アーキテクチャのレベルからの再考が必要と言える

本発表の概要: 計算強度とI/O強度の観点から

- ビッグデータ向け計算機アーキテクチャの研究例

ストレージマイグレーション・仮想マシンマイグレーション

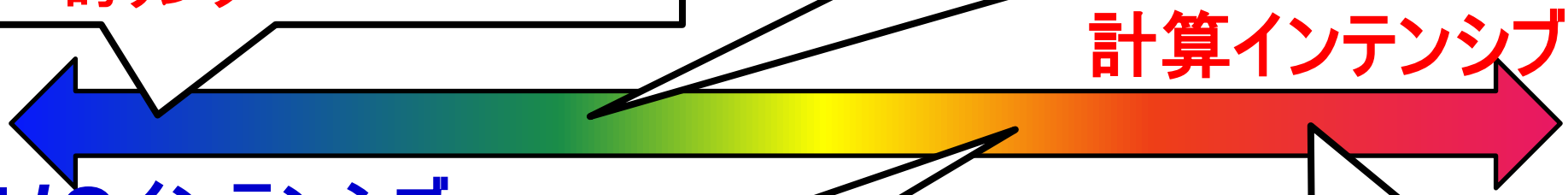
- ビッグデータをサーバからサーバへ移動させる
- サイズは数GBからTB級

→ 光無線による 40GbE 動的リンク

構造型ストレージ(NOSQL)

- 用途特化型でスケーラビリティの高いデータベース
- 大量のデータ転送を扱う

→ 40GbE FPGAボードを用いた DB キャッシュ HW



メニーコアプロセッサ

- リクエストレベル並列性
- 多数のプロセッサを積層

→ ワイヤレス3次元CMP

グラフ型 DB の探索処理

- ソーシャルグラフの探索
- ノード数は10万以上

→ GPU を用いた並列処理

高速化の基本: ルーフラインモデル(HPC の例)

- アプリケーション性能を律速する要因
  - 計算能力
  - メモリ I/O 性能

  } どちらを改善すればよいかは、対象アプリにおける計算と I/O の割合に依存

- アプリの算術強度(FP演算/メモリアクセス)の例

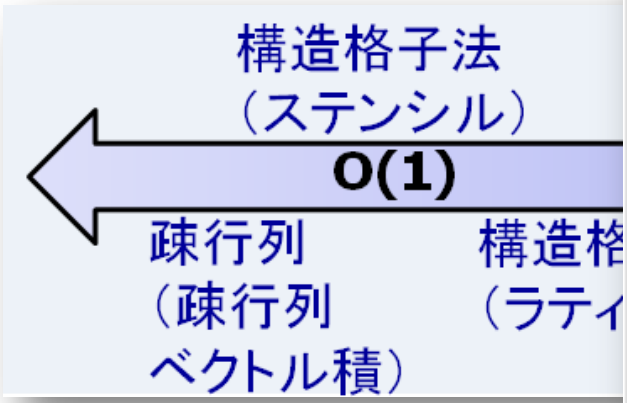
- 達成可能な GFLOPs/sec
  - ピークのメモリ帯域 × 算術強度
  - ピークのFP演算性能

  } どちらか小さいほうに律速

※図はD.Patterson & J.Hennessy, "Computer Organization and Design (4/E)"より

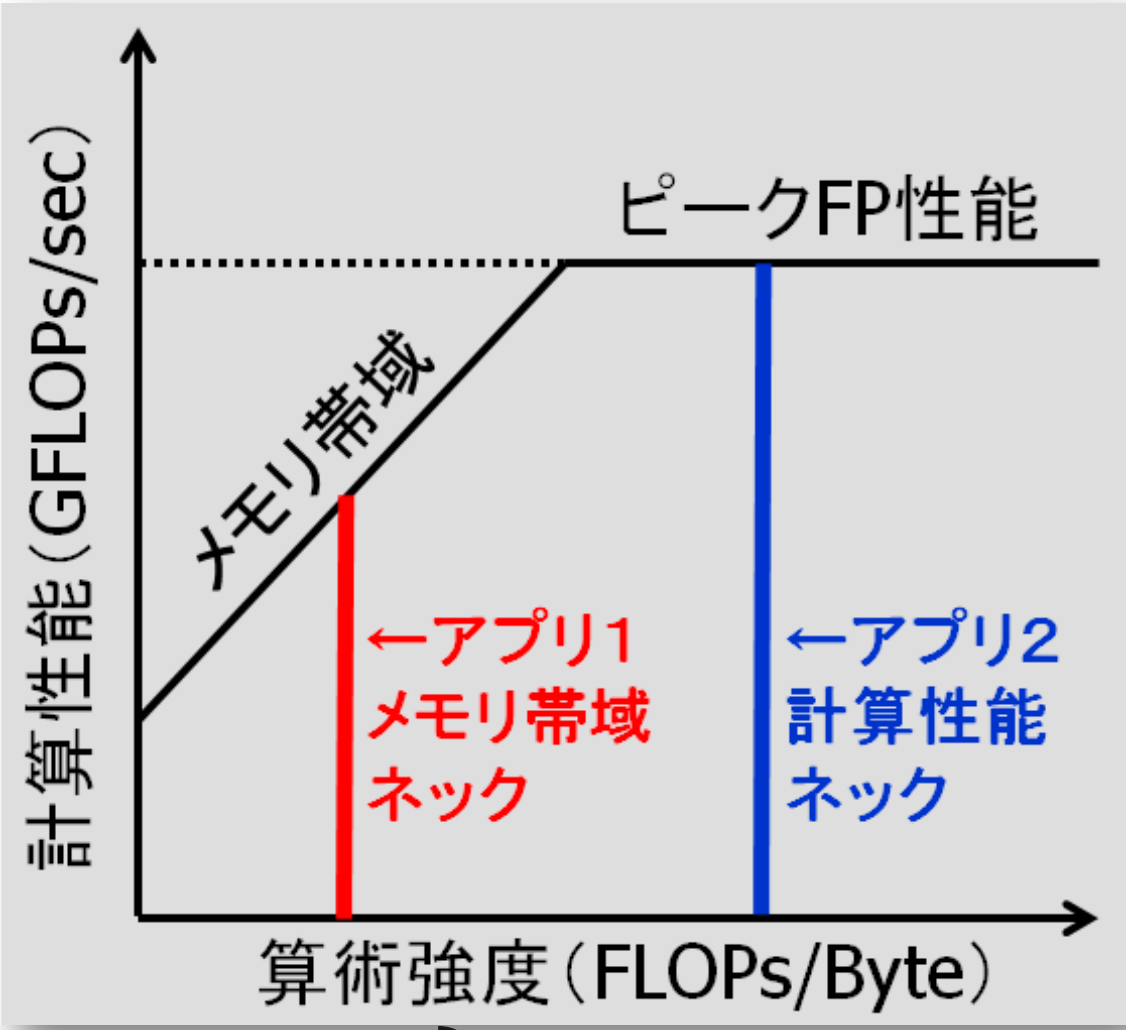
高速化の基本: ルーフラインモデル(HPC の例)

- アプリケーション性
  - 計算能力
  - メモリ I/O 性能
- アプリの算術強度(



- 達成可能な GFLOP
  - ピークのメモリ帯域 × 算術強度
  - ピークのFP演算性能

どちらか小さい
ほうに律速



※図はD.Patterson & J.Hennessy, "Computer Organization and Design (4/E)"より

本発表の概要: 計算強度とI/O強度の観点から

- ビッグデータ向け計算機アーキテクチャの研究例

ストレージマイグレーション・仮想マシンマイグレーション

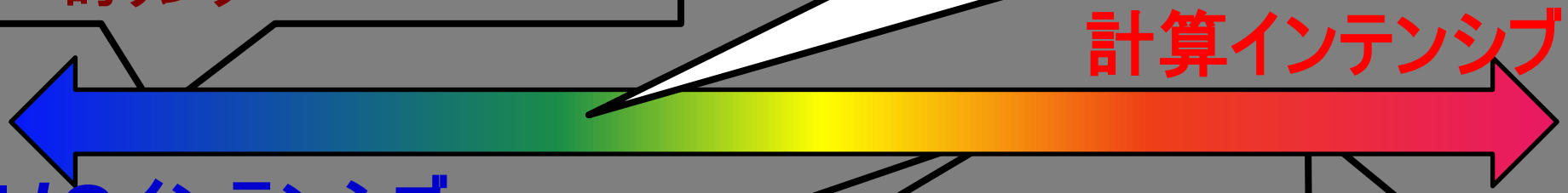
- ビッグデータをサーバからサーバへ移動させる
- サイズは数GBからTB級

→ **光無線による 40GbE 動的リンク**

構造型ストレージ(NOSQL)

- 用途特化型でスケーラビリティの高いデータベース
- 大量のデータ転送を扱う

→ **40GbE FPGAボードを用いた DB キャッシュ HW**

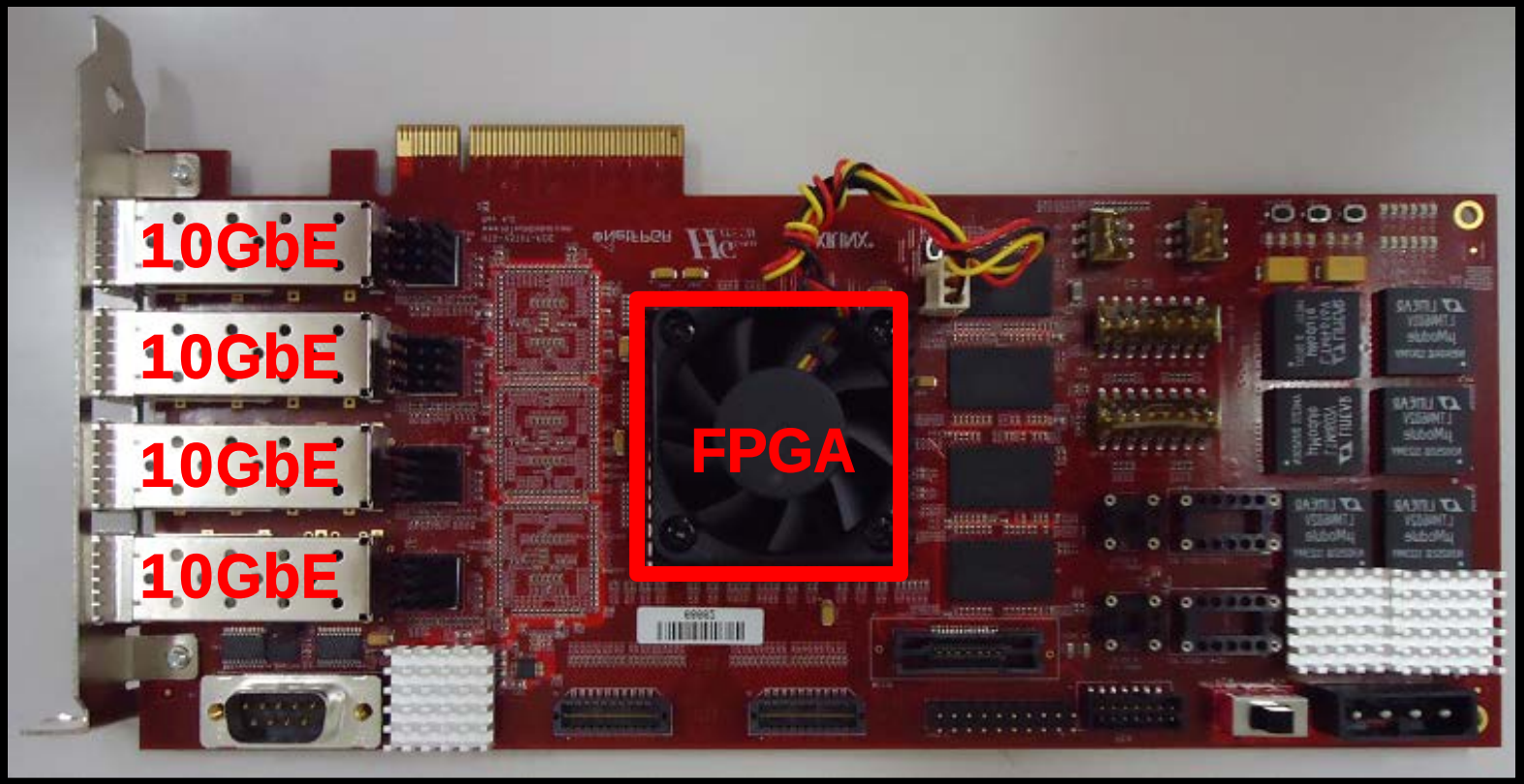


メニーコアプロセッサ

グラフ型 DB の探索処理

RDBMSに比べると処理はシンプル(例: Key-value store)
→ I/O ネックになるので、通信と計算の「密結合」が有利

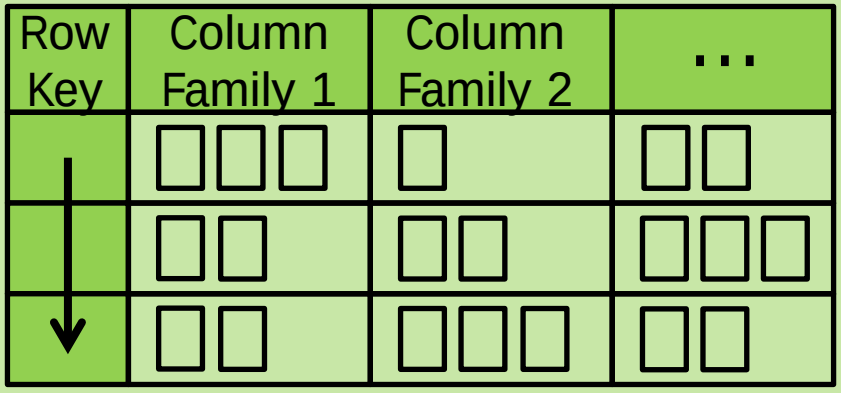
本発表の概要: 計算強度とI/O強度の観点から



RDBMSに比べると処理はシンプル（例: Key-value store）
→ I/O ネックになるので、通信と計算の「密結合」が有利

構造型ストレージ: データ構造の点から分類

構造型ストレージは、水平スケーラビリティに優れるが得手不得手がある(特定用途特化型)



HBase, BigTable

カラム指向型

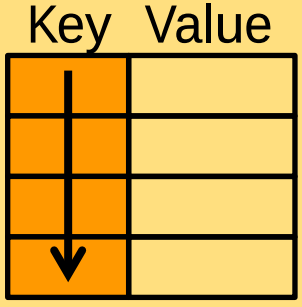
MongoDB

ドキュメント指向型

```
{ _id : ObjectId(0),
  name : Risa,
  tel : 1234 }
{ _id : ObjectID(1)
  name : Shinpei,
  mail : kato@x.jp}
```

Schema-less DB

Memcached

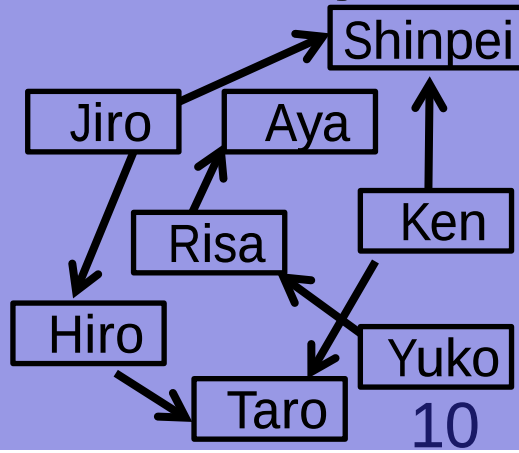


キーバリューストア型

Shopping cart, User profile, Session, etc

Neo4j

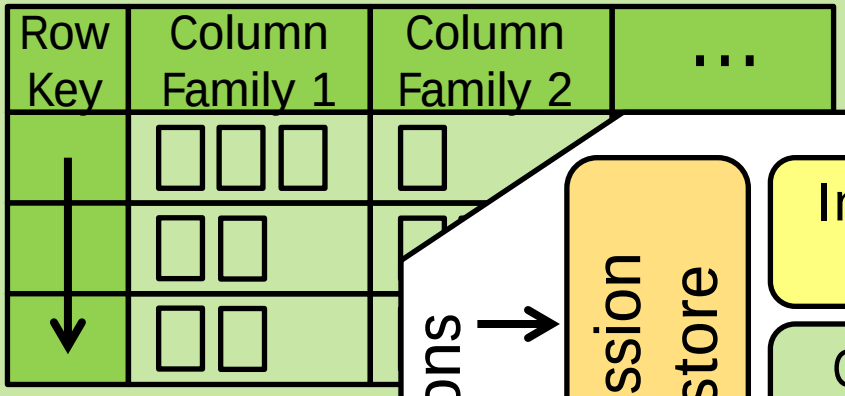
グラフ型



Customer social graph

ポリグロット永続化: 複数 DB を相補的に利用

ポリグロット＝多言語 → 特定用途に特化した構造型ストレージを組み合わせれば、複雑なサービスも実現できる



HBase, BigTable

MongoDB

```
{ _id : ObjectId(0),
  name : Risa,
  tel : 1234 }
{ _id : ObjectID(1)
  name : Shinpei,
  mail : kato@x.jp}
```

Schema-less DB

Applications

Cache & session Key-value store

Inventory & price RDBMS

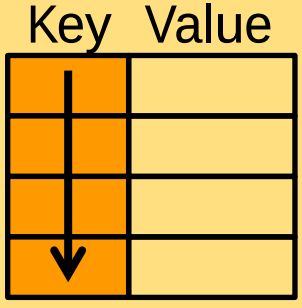
CMS & blogging Column-oriented

Questionnaire Document DB

Recommendation Graph DB

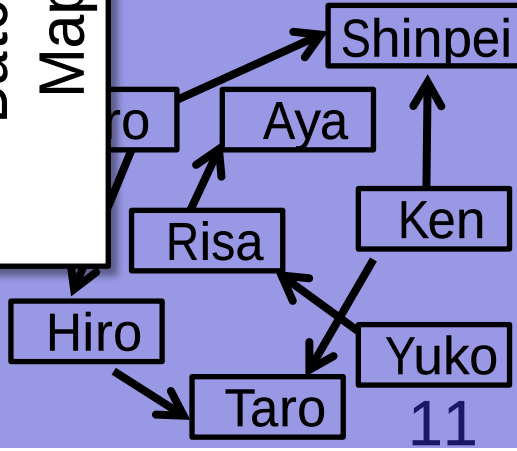
Batch processing MapReduce, BI

Memcached



Shopping cart, User profile, Session, etc

Neo4j



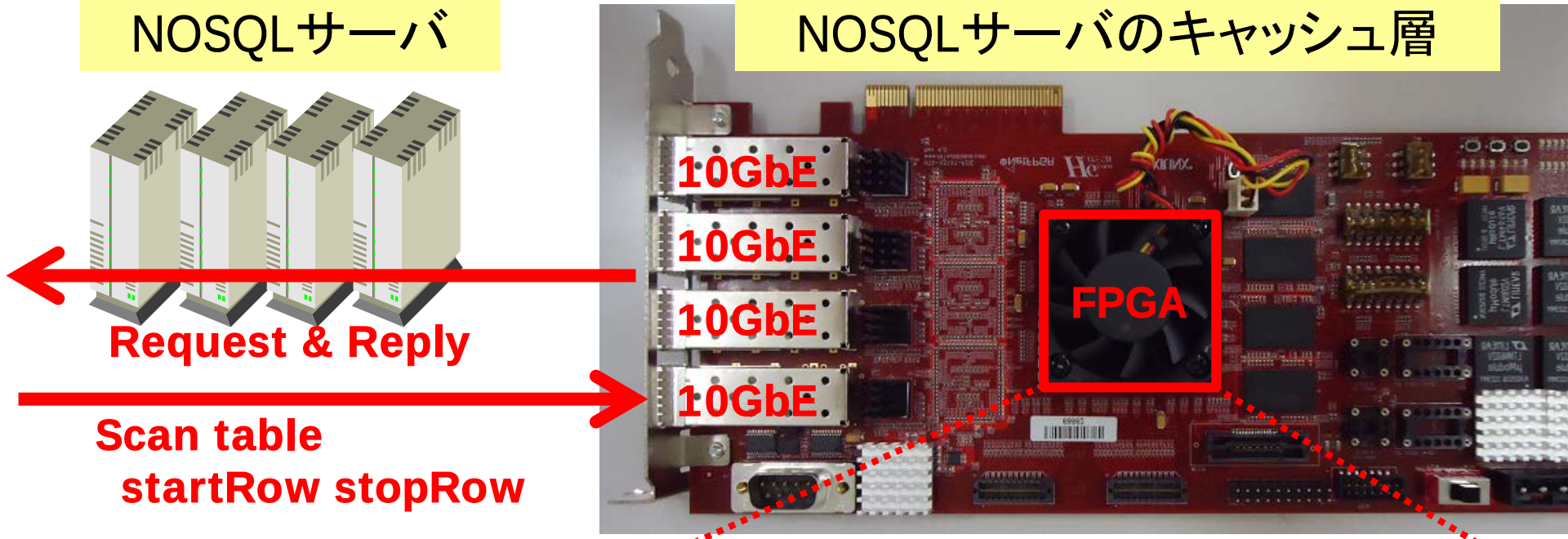
Customer social graph

40GbEとFPGAを用いたNOSQLキャッシュ

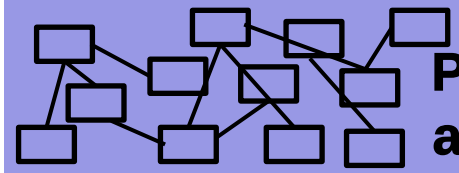
- 各種NOSQLのCRUD操作をFPGA上にハード化
- 40GbEネットワークとDB HWを直結（I/Oネック）

NOSQLサーバ

NOSQLサーバのキャッシュ層

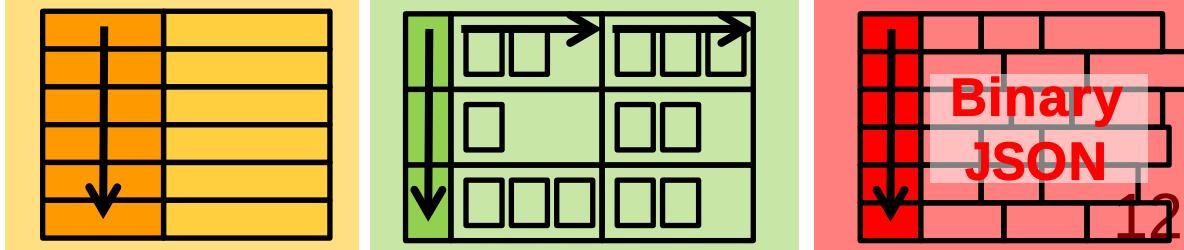


Graph processing using
Many cores or GPUs



Parallel algorithm

Hardware-based table management



40GbEとFPGAを用いたNOSQLキャッシュ

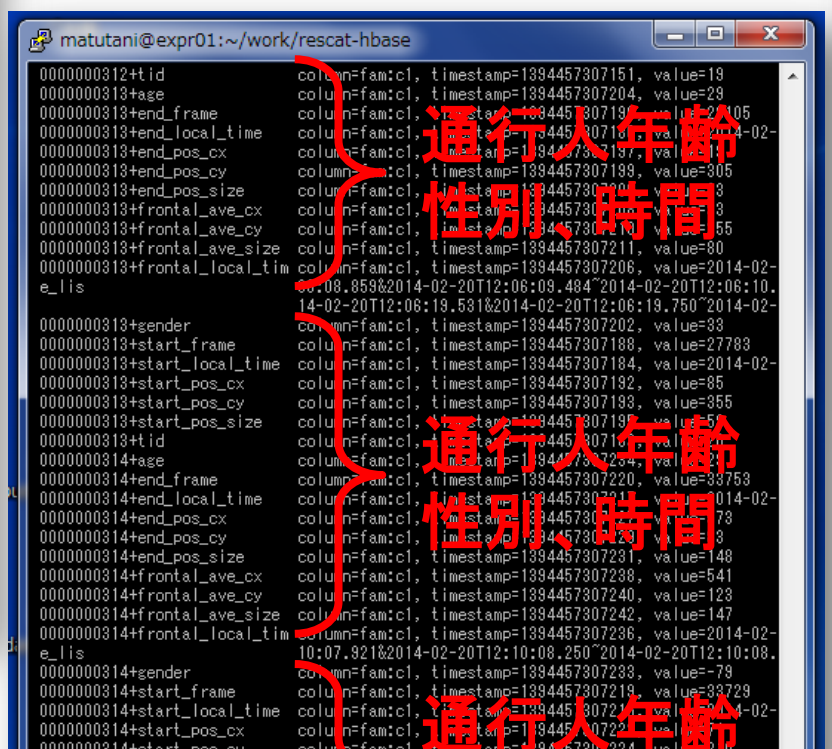
- 各種NOSQLのCRUD操作をFPGA上にハード化
- 40GbEネットワークとDB HWを直結（I/Oネック）

HBase サーバ群

Put table 0101+age 28
Put table 0101+gender M
…

カメラ画像の
リアルタイム解析

沖電気RESCAT

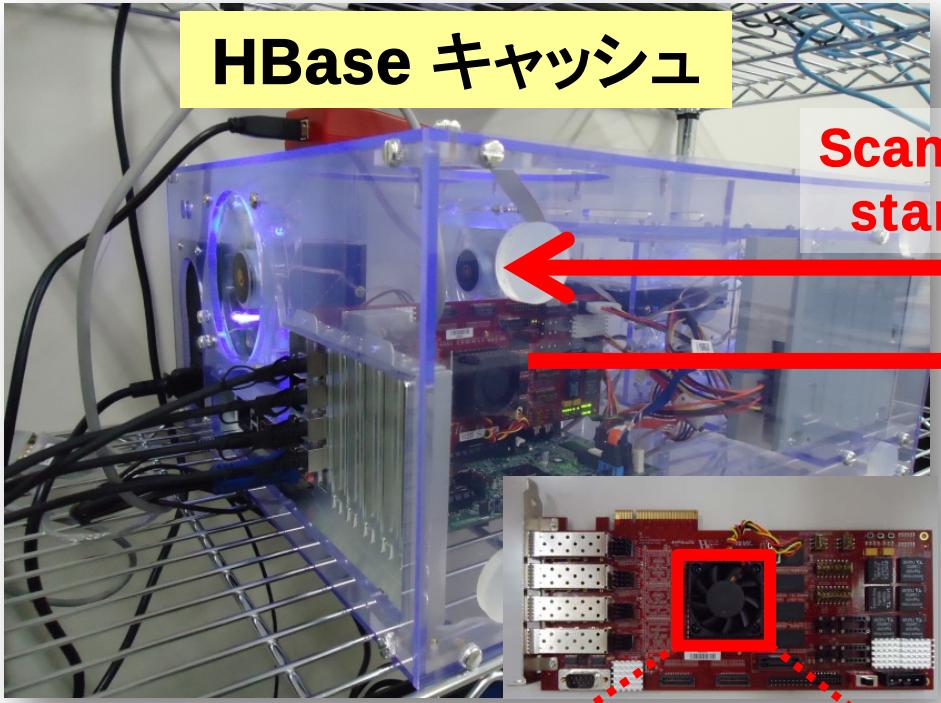


40GbEとFPGAを用いたNOSQLキャッシュ

- 各種NOSQLのCRUD操作をFPGA上にハード化
- 40GbEネットワークとDB HWを直結（I/Oネック）

HBase サーバ群

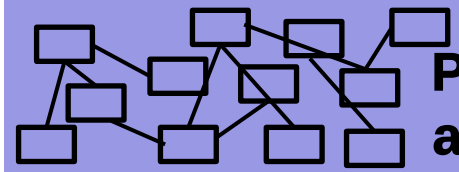
HBase キャッシュ



Scan table
startRow stopRow

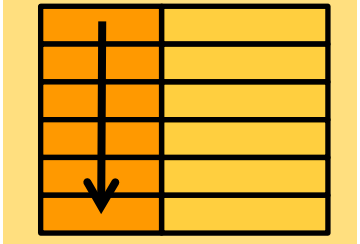
Cached Results

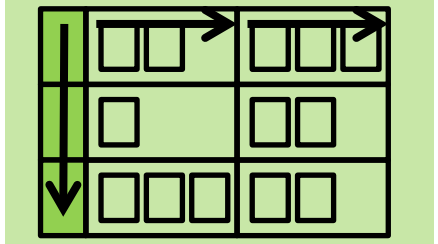
Graph processing using
Many cores or GPUs

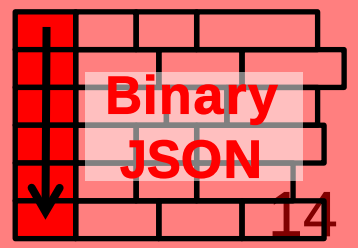


Parallel algorithm

Hardware-based table management







Binary JSON

本発表の概要: 計算強度とI/O強度の観点から

- ビッグデータ向け計算機アーキテクチャの研究例

グラフ計算量は問題サイズ(ノード数、次数)によって増加
ソーシャルグラフはノード数が非常に大きく、次数も大きい(Facebookの平均次数は197！)
→GPUやメニーコアなど計算インテンシブなデバイス有利

的リンク

計算インテンシブ

I/Oインテンシブ

メニーコアプロセッサ

- リクエストレベル並列性
- 多数のプロセッサを積層

→ワイヤレス3次元CMP

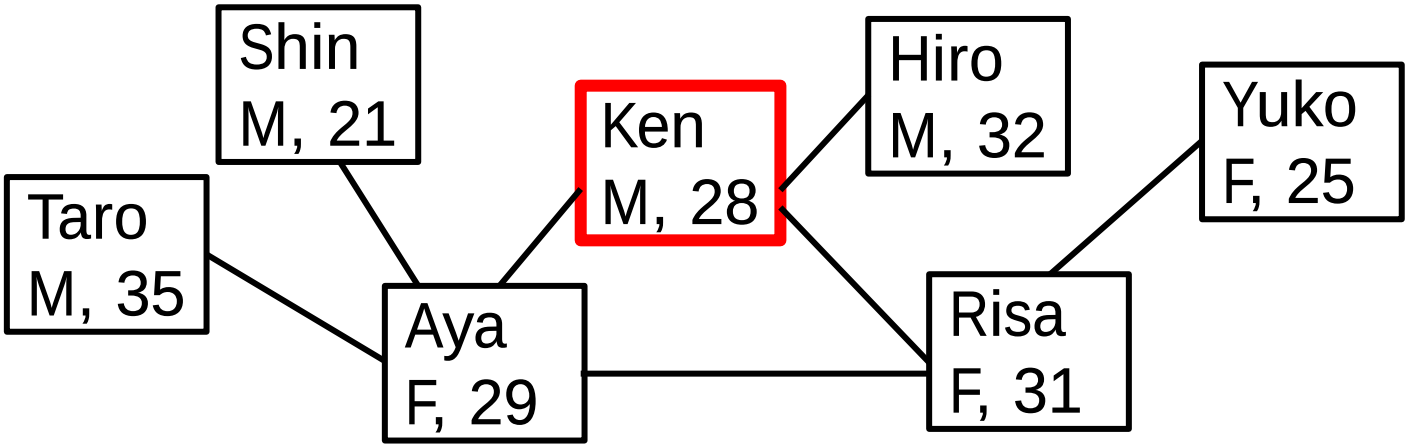
グラフ型 DB の探索処理

- ソーシャルグラフの探索
- ノード数は10万以上

→ GPU を用いた並列処理

SNS とグラフ型データベース

- ソーシャルネットワーキングサービス（SNS）
  - 膨大な数の会員数（例: ノード数＝億単位、次数＝197）
  - インターネット上の交流、リコメンデーションエンジン
- 検索クエリの例
  - 「Ken」の2ホップ以内の友人で、25歳以上の男性は？



- グラフ型データベース
  - ノード（特徴）、エッジ（ノード間の関係性）の操作に特化
  - Dijkstra、A*、Shortest Path、All Path、All Simple Path

GPU によるグラフ型 DB 高速化

GeForce GTX 780Ti
2,880 cores

数十万ノードのグラフ探索を GPU を用いて高速化
30～40倍の高速化

Dijkstra法、A*法を実装

Neo4j（Java）からjcudaを用いて CUDA 呼び出し

本発表の概要: 計算強度とI/O強度の観点から

- ビッグデータ向け計算機アーキテクチャの研究例

グラフ計算量は問題サイズ(ノード数、次数)によって増加
ソーシャルグラフはノード数が非常に大きく、次数も大きい(Facebookの平均次数は197！)
→GPUやメニーコアなど計算インテンシブなデバイス有利

的リンク

計算インテンシブ

I/Oインテンシブ

メニーコアプロセッサ

- リクエストレベル並列性
- 多数のプロセッサを積層

→ ワイヤレス3次元CMP

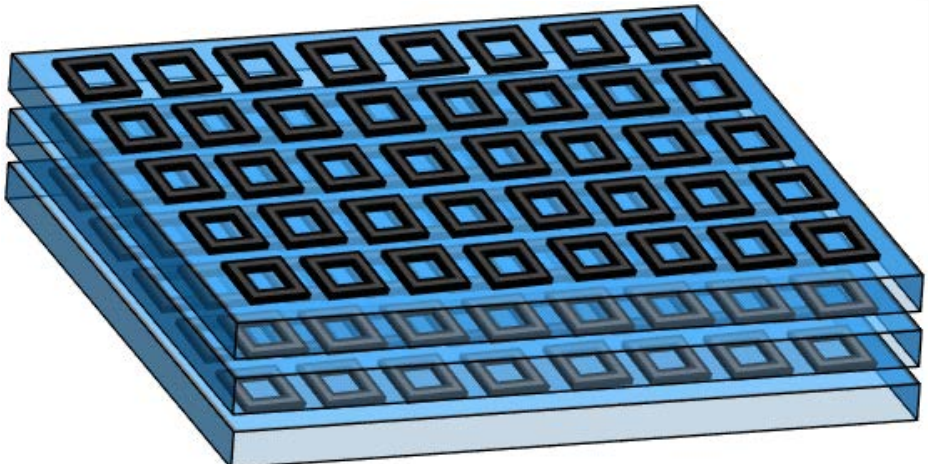
グラフ型 DB の探索処理

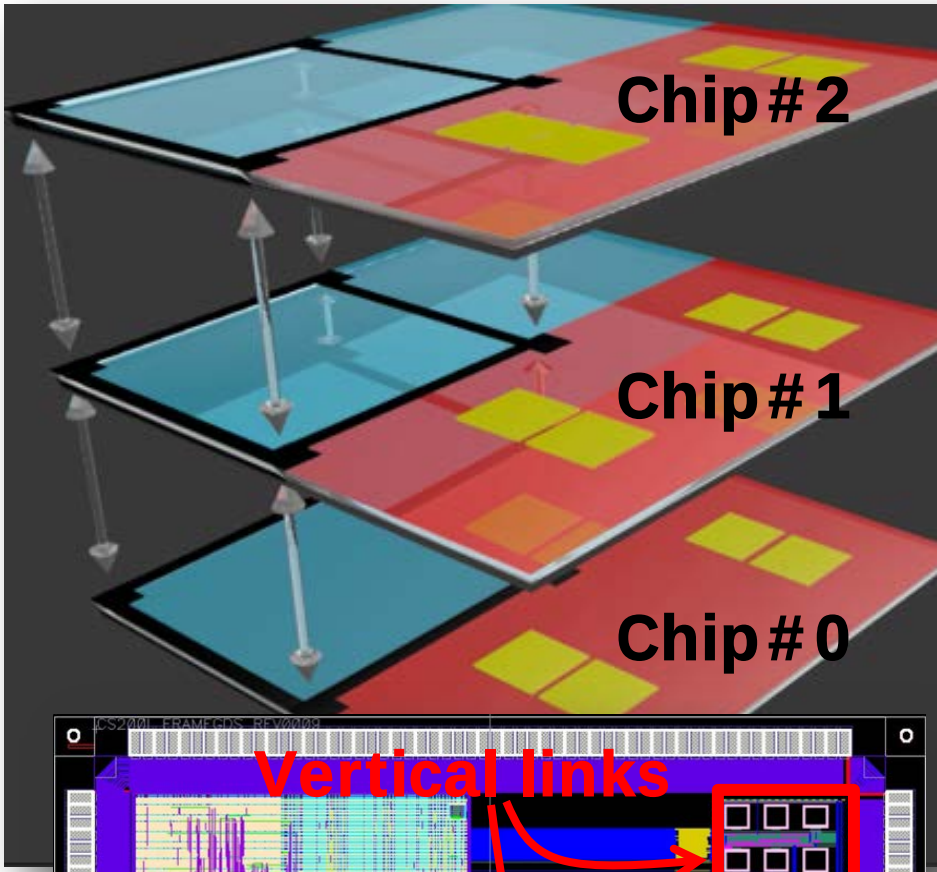
- ソーシャルグラフの探索
- ノード数は10万以上

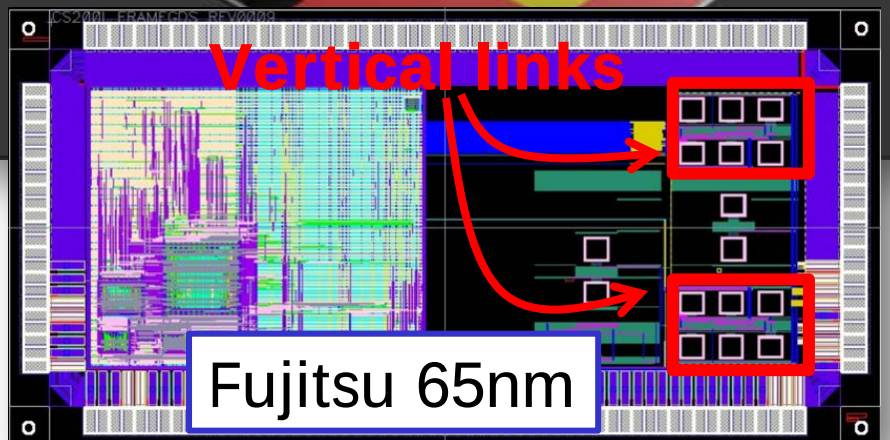
→ GPU を用いた並列処理

ワイヤレス3次元チップマルチプロセッサ

- 共有メモリ型マルチプロセッサ（CMP）
  - プロセッサ（L1キャッシュ）
  - 共有L2キャッシュバンク
  - オンチップルータで結合
- 3次元CMP
  - CMP チップを複数積層
  - 積層チップ間は無線通信
    （例: 誘導結合）







本発表の概要: 計算強度とI/O強度の観点から

- ビッグデータ向け計算機アーキテクチャの研究例

ストレージマイグレーション・仮想マシンマイグレーション

- ビッグデータをサーバからサーバへ移動させる
- サイズは数GBからTB級

→ **光無線による 40GbE 動的リンク**

構造型ストレージ(NOSQL)

- 用途特化型でスケーラビリティの高いデータベース
- 大量のデータ転送を扱う

→ **40GbE FPGAボードを用いた DB キャッシュ HW**

計算インテンシブ

I/Oインテンシブ

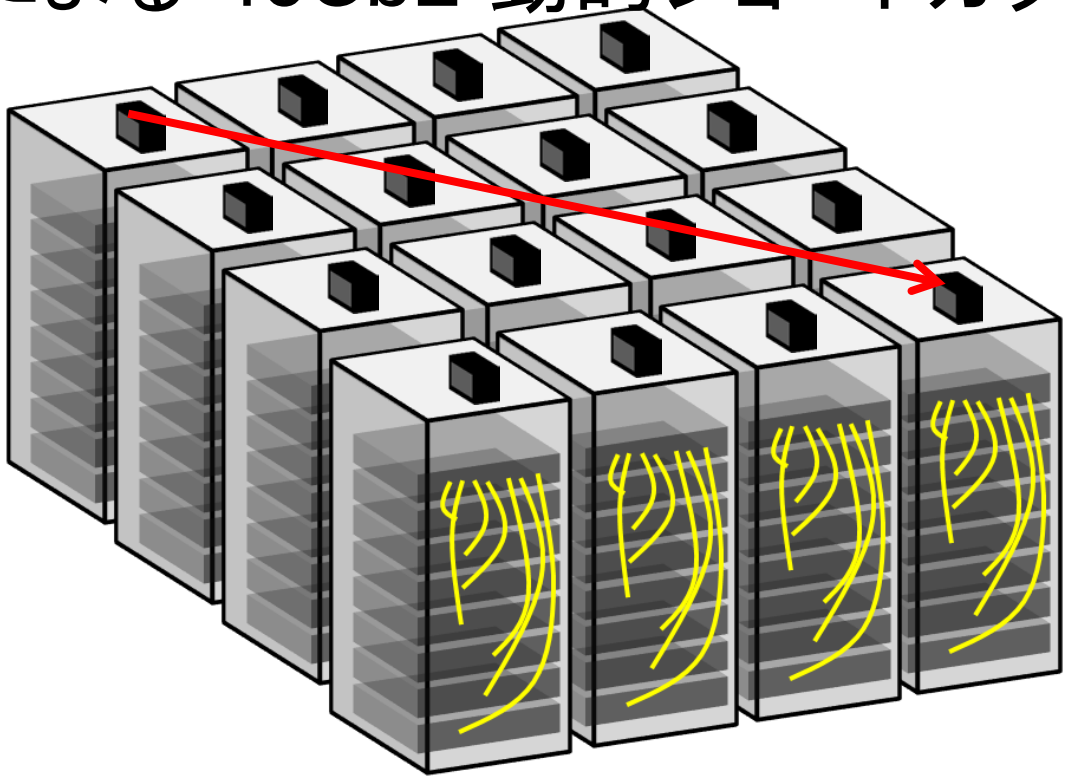
メニーコアプロセッサ

グラフ型 DB の探索処理

データ転送なので計算量は少なく、完全に I/Oネックとなる
→ ネットワークを増強すべき

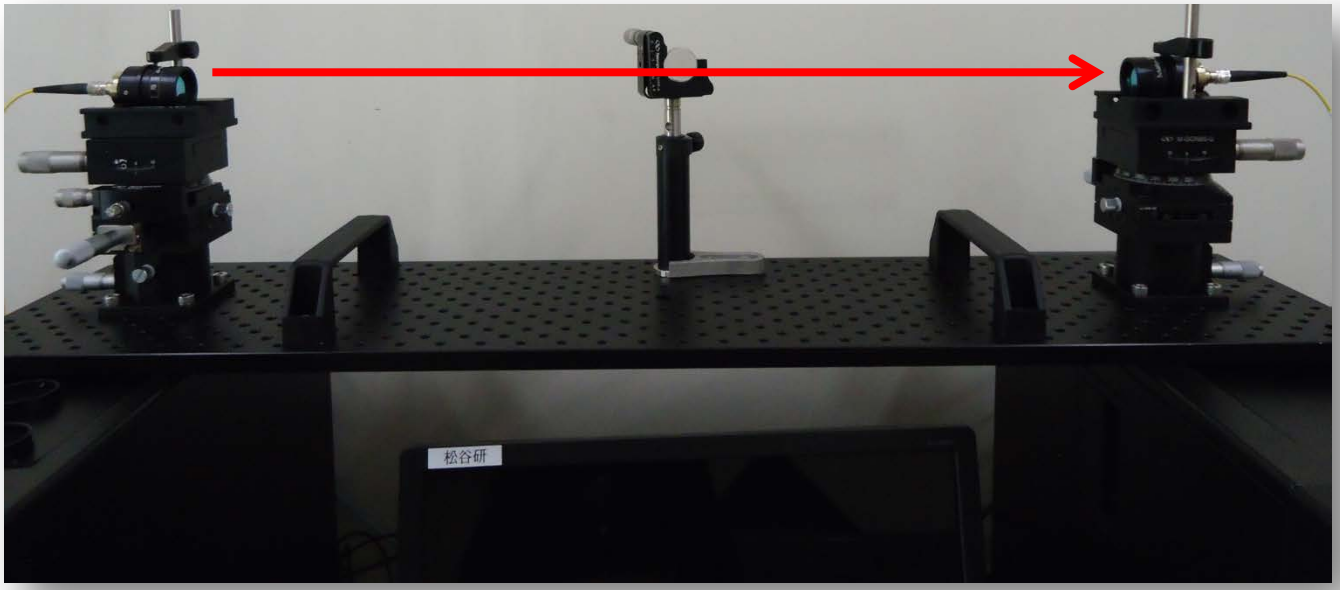
光無線による40GbE動的ショートカットリンク

- データセンターにおける突発的データ転送
  - 仮想マシンの移送（数GBオーダ）
  - ストレージマイグレーション、分散DBのストリーミング
  - 1GbE リンクでは、VM 移送に数分かかることもある
- 光無線による 40GbE 動的ショートカットリンク



光無線による40GbE動的ショートカットリンク

- 光無線による 40GbE 動的ショートカットリンク
  - 40GbE LR4（波長1300nm）をコリメータレンズに直結
  - レンズの向きを調整することで動的にリンクを形成



- VM（仮想マシン）ハイウェイ
  - 負荷分散やメンテナンスのためのVM移送など
  - 大容量のデータ転送の前に追加リンクを動的に準備

本発表のまとめ: ビッグデータ処理高速化の指針

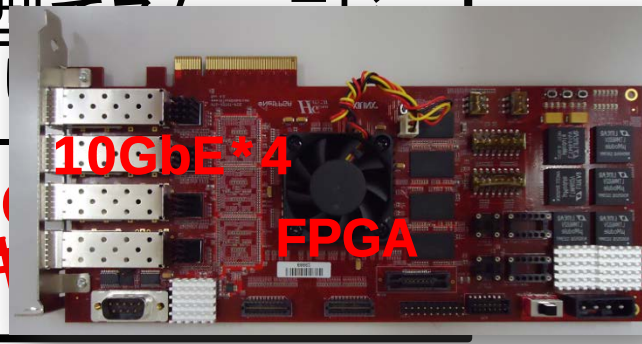
- アプリケーションごとに計算強度、I/O強度は異なる
- ルーフラインモデルを用いて、どこを改善すべきか判断

ストレージマイグレーション・仮想マシンマイグレーション

→ 光無線による 40GbE 動的リンク

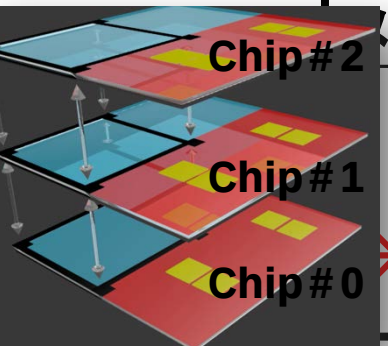
構造型ストレージ(NOSQL)

- 用途特化型
  リティの高
- 大量のデー
→ 40GbE FP
  いた DB キ



計算インテンシブ

I/Oインテンシブ



メニーコアプロセッサ

リクエストレベル並列性
多数のプロセッサを積層
→ ワイヤレス3次元CMP

グラフ型 DB の探索処理

-
-
→

Acknowledgement:
本研究の一部はJSTさきがけ、総務省SCOPEの支援を受けています