



# 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
- 組み込み機器向け 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

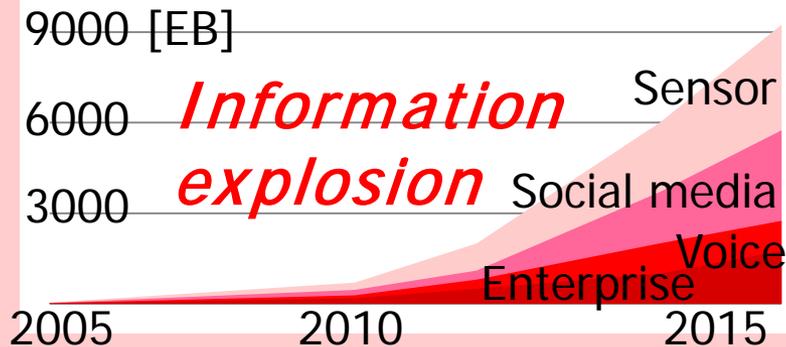
## 教員@慶應大学

- 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

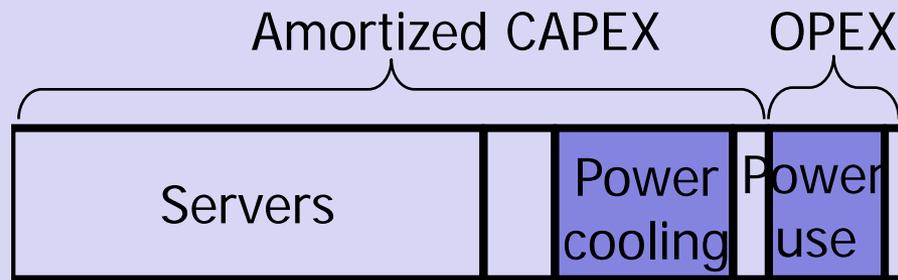
データの蓄積・利活用によって  
さまざまなイノベーションが期待



→ 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**

**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**

# 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

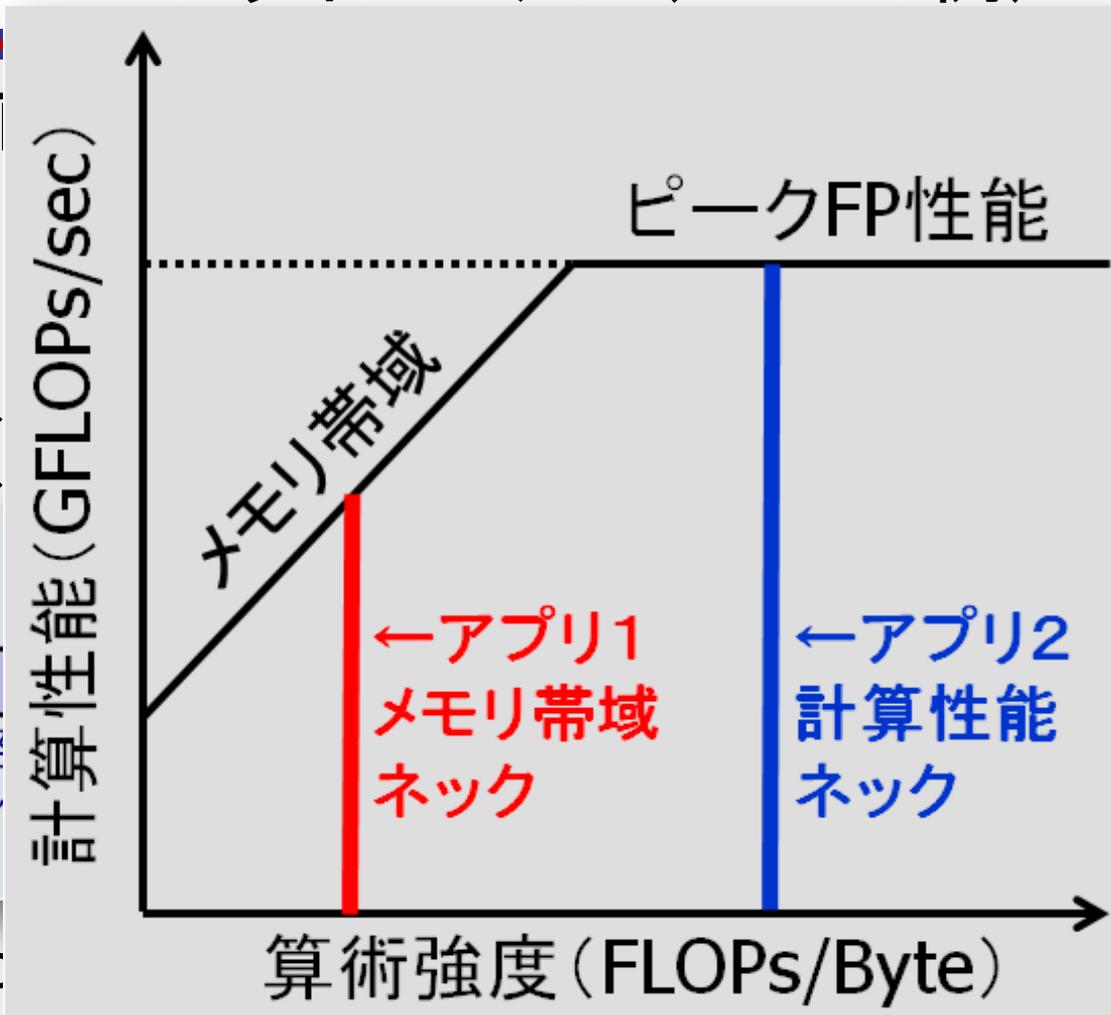
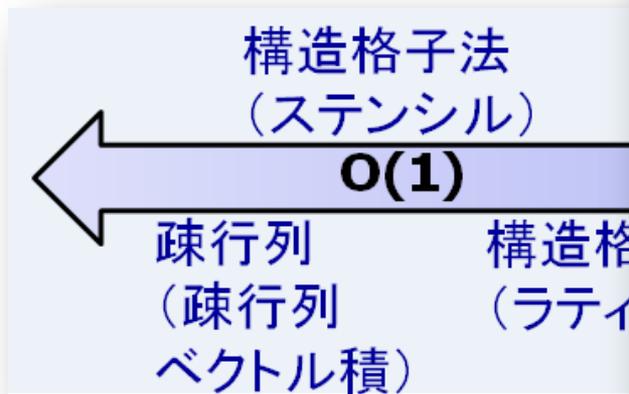
- ピークのメモリ帯域 × 算術強度

- ピークのFP演算性能

どちらか小さいほうに律速

# 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
  - 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

# 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

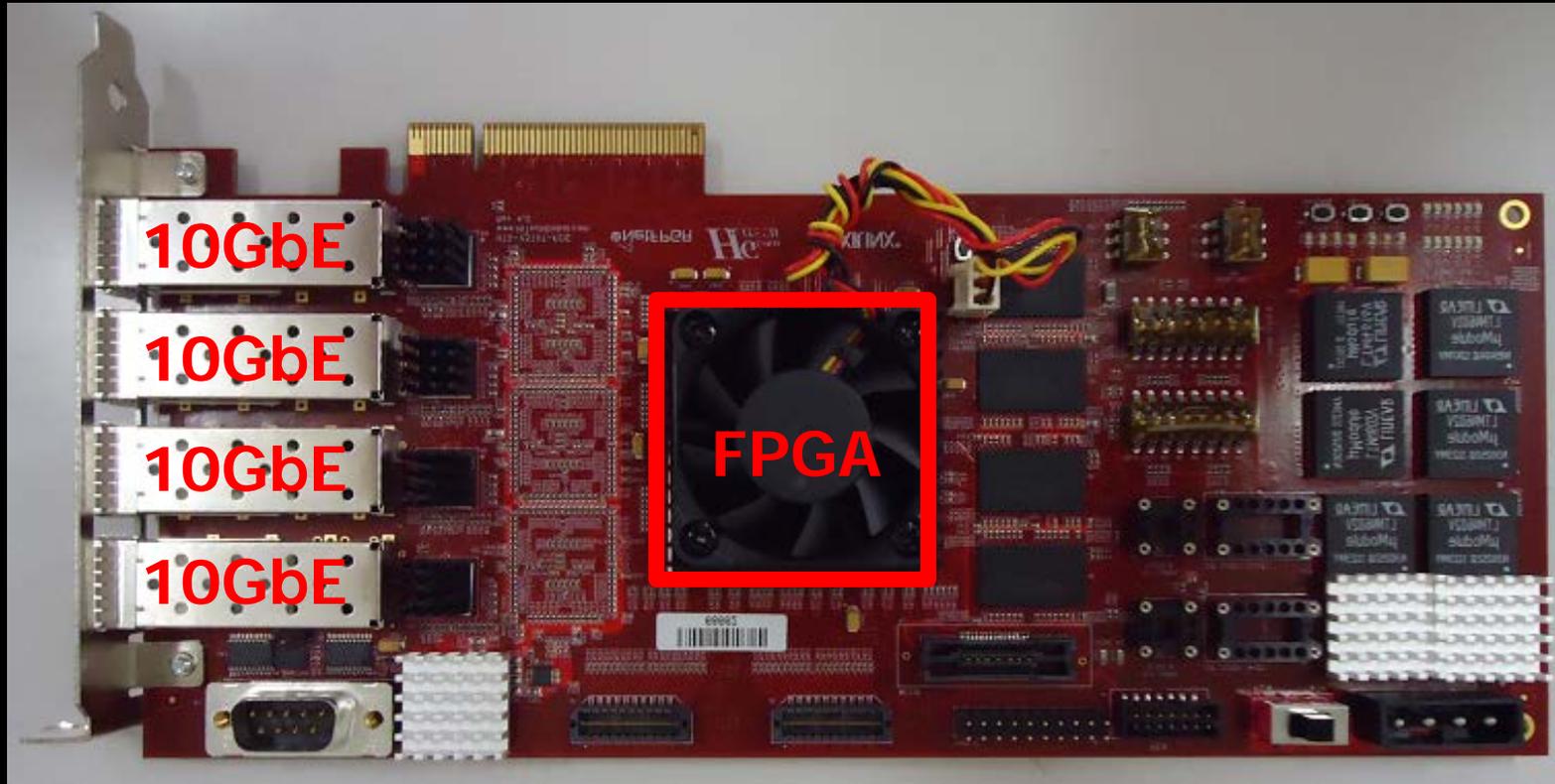
I/Oインテンシブ

メニーコアプロセッサ

グラフ型 DB の探索処理

RDBMSに比べると処理はシンプル(例: Key-value store)  
→ I/O ネットになるので、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 ネットになるので、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)

Row Key	Column Family 1	Column Family 2	...
↓	□□□	□	□□
↓	□□	□□	□□□
↓	□□	□□□	□□

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

Key	Value
↓	
↓	
↓	
↓	

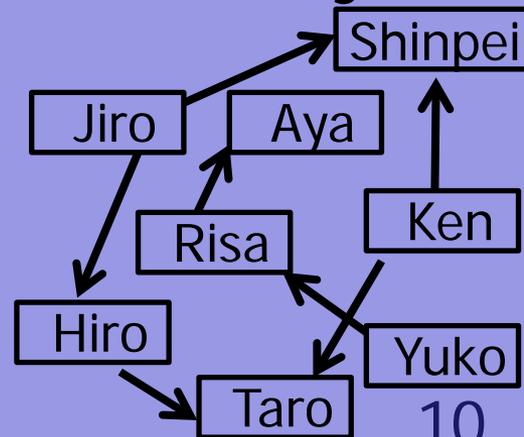
キーバリュー  
ストア型

Shopping cart, User  
profile, Session, etc

グラフ型

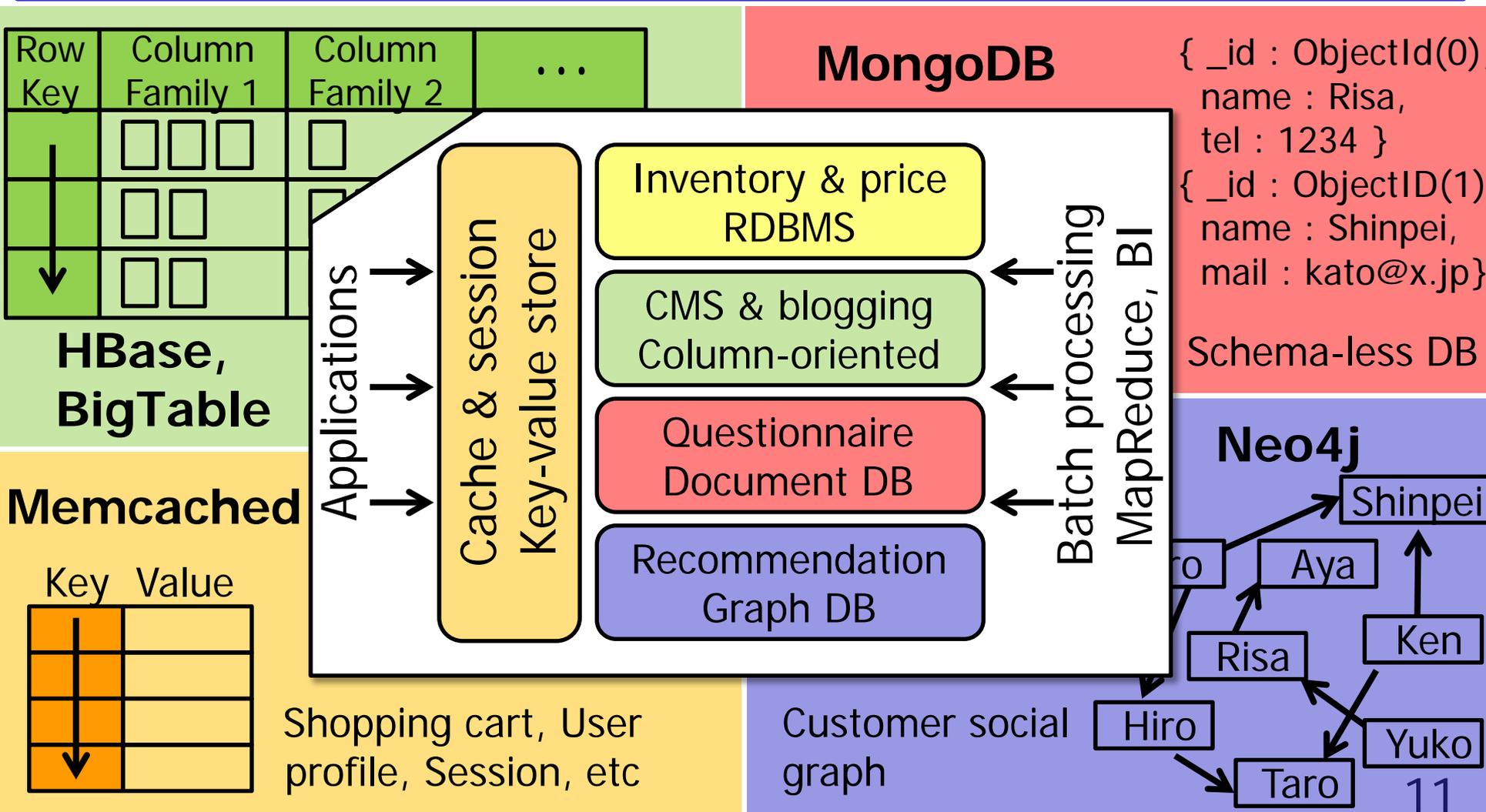
Customer social  
graph

Neo4j



# 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

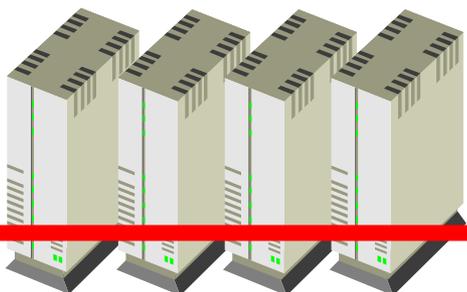


# 40GbEとFPGAを用いたNOSQLキャッシュ

- 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)

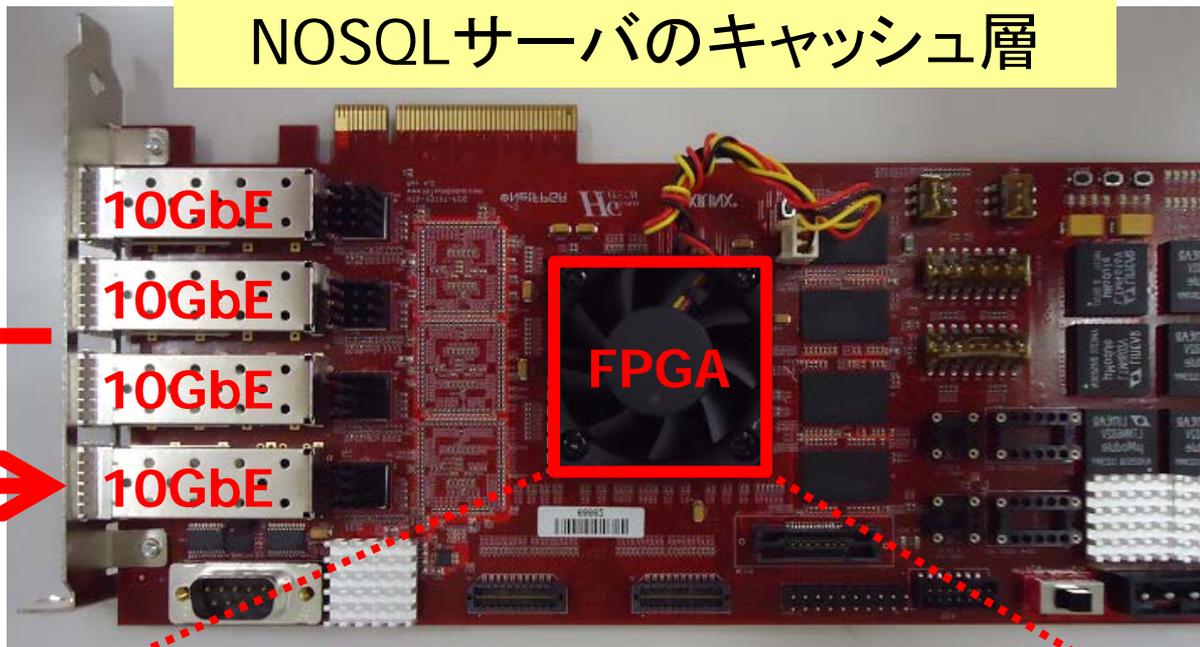
NOSQLサーバ

NOSQLサーバのキャッシュ層



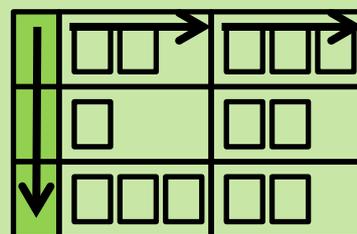
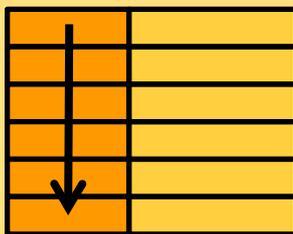
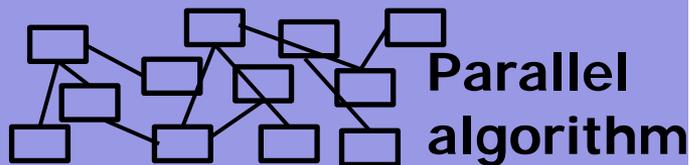
Request & Reply

Scan table  
startRow stopRow



Graph processing using  
Many cores or GPUs

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  
...

```
matutani@expr01:~/work/rescat-hbase
0000000312+id column=fam:c1, timestamp=1394457307151, value=19
0000000313+age column=fam:c1, timestamp=1394457307204, value=29
0000000313+end_frame column=fam:c1, timestamp=1394457307193, value=2905
0000000313+end_local_time column=fam:c1, timestamp=1394457307193, value=2014-02-
0000000313+end_pos_cx column=fam:c1, timestamp=1394457307193, value=8
0000000313+end_pos_cy column=fam:c1, timestamp=1394457307193, value=305
0000000313+end_pos_size column=fam:c1, timestamp=1394457307206, value=3
0000000313+frontal_ave_cx column=fam:c1, timestamp=1394457307206, value=8
0000000313+frontal_ave_cy column=fam:c1, timestamp=1394457307210, value=55
0000000313+frontal_ave_size column=fam:c1, timestamp=1394457307211, value=80
0000000313+frontal_local_time column=fam:c1, timestamp=1394457307206, value=2014-02-
0000000313+frontal_local_time_e_lis 00.08.859&2014-02-20T12:06:09.484~2014-02-20T12:06:10.
14-02-20T12:06:19.531&2014-02-20T12:06:19.750~2014-02-
column=fam:c1, timestamp=1394457307202, value=33
0000000313+gender column=fam:c1, timestamp=1394457307188, value=27783
0000000313+start_frame column=fam:c1, timestamp=1394457307184, value=2014-02-
0000000313+start_local_time column=fam:c1, timestamp=1394457307192, value=85
0000000313+start_pos_cx column=fam:c1, timestamp=1394457307193, value=355
0000000313+start_pos_cy column=fam:c1, timestamp=1394457307193, value=5
0000000313+start_pos_size column=fam:c1, timestamp=1394457307193, value=2
0000000313+id column=fam:c1, timestamp=1394457307193, value=19
0000000314+age column=fam:c1, timestamp=1394457307220, value=22
0000000314+end_frame column=fam:c1, timestamp=1394457307220, value=33753
0000000314+end_local_time column=fam:c1, timestamp=1394457307220, value=2014-02-
0000000314+end_pos_cx column=fam:c1, timestamp=1394457307220, value=73
0000000314+end_pos_cy column=fam:c1, timestamp=1394457307220, value=8
0000000314+end_pos_size column=fam:c1, timestamp=1394457307231, value=148
0000000314+frontal_ave_cx column=fam:c1, timestamp=1394457307238, value=541
0000000314+frontal_ave_cy column=fam:c1, timestamp=1394457307240, value=123
0000000314+frontal_ave_size column=fam:c1, timestamp=1394457307242, value=147
0000000314+frontal_local_time column=fam:c1, timestamp=1394457307236, value=2014-02-
0000000314+frontal_local_time_e_lis 10:07.321&2014-02-20T12:10:08.250~2014-02-20T12:10:08.
column=fam:c1, timestamp=1394457307233, value=79
0000000314+gender column=fam:c1, timestamp=1394457307219, value=33729
0000000314+start_frame column=fam:c1, timestamp=1394457307222, value=2014-02-
0000000314+start_local_time column=fam:c1, timestamp=1394457307222, value=8
0000000314+start_pos_cx column=fam:c1, timestamp=1394457307222, value=3
0000000314+start_pos_cy column=fam:c1, timestamp=1394457307222, value=3
```

通行人年齢  
性別  
時間

通行人年齢  
性別  
時間

通行人年齢

カメラ画像の  
リアルタイム解析



沖電気RESCAT

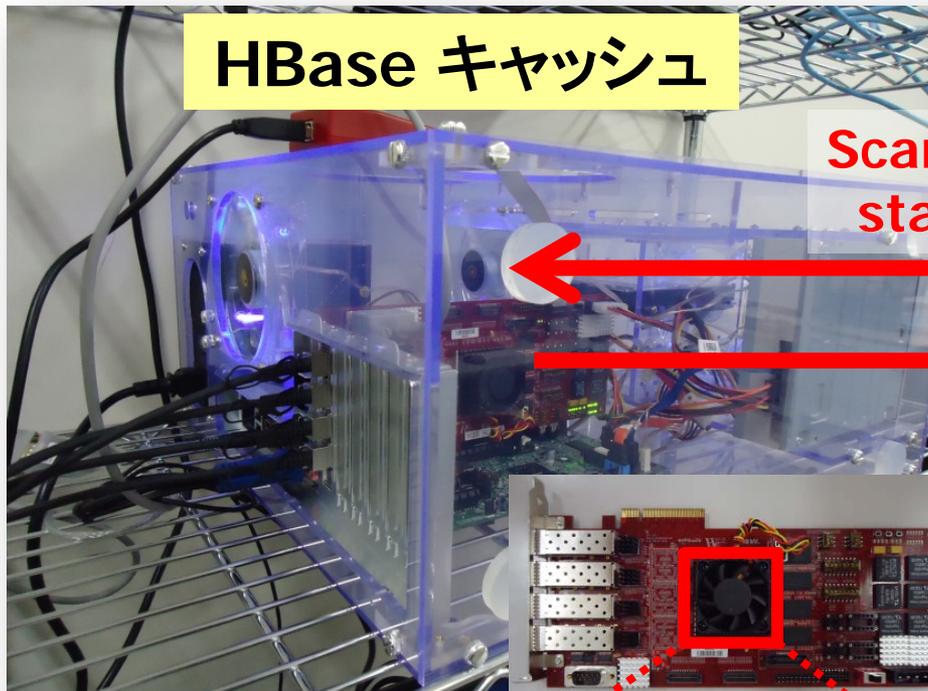
# 40GbEとFPGAを用いたNOSQLキャッシュ

- 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

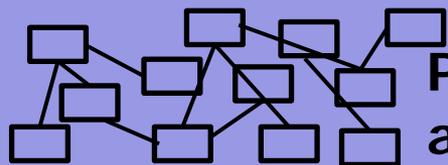


Scan table

startRow stopRow

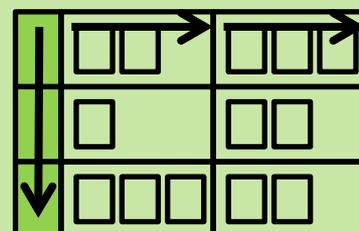
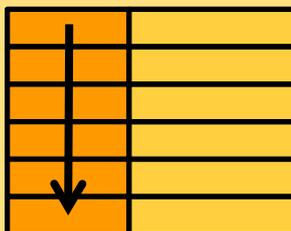
Cached Results

Graph processing using  
Many cores or GPUs



Parallel  
algorithm

Hardware-based table management



# 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

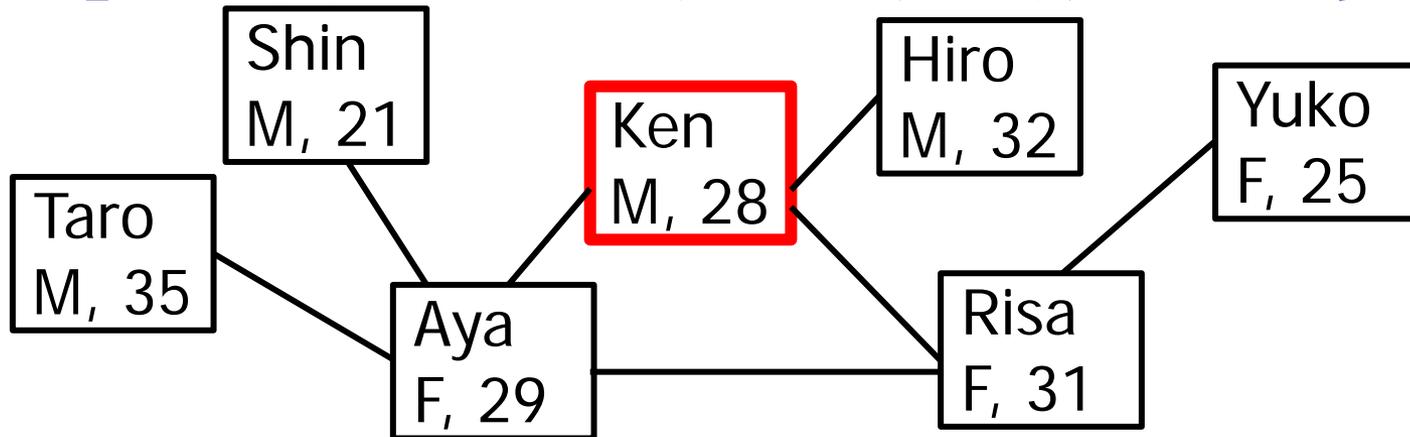
- リクエストレベル並列性
  - 多数のプロセッサを積層
- ワイヤレス3次元CMP

## 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ク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

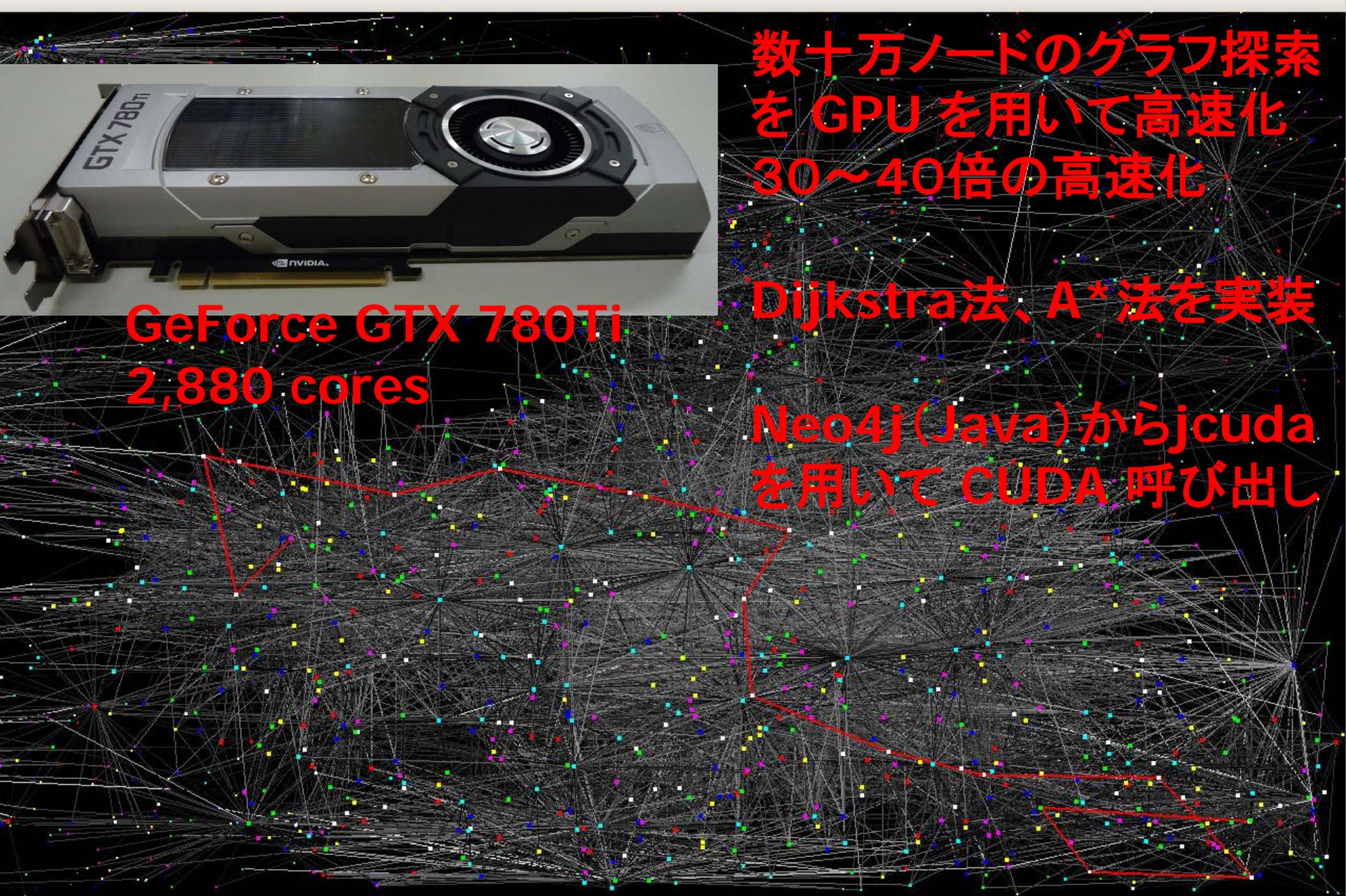


GeForce GTX 780Ti  
2,880 cores

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

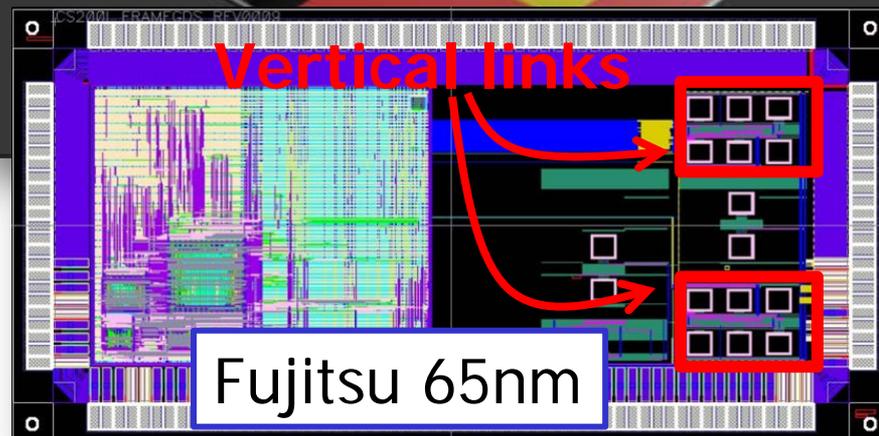
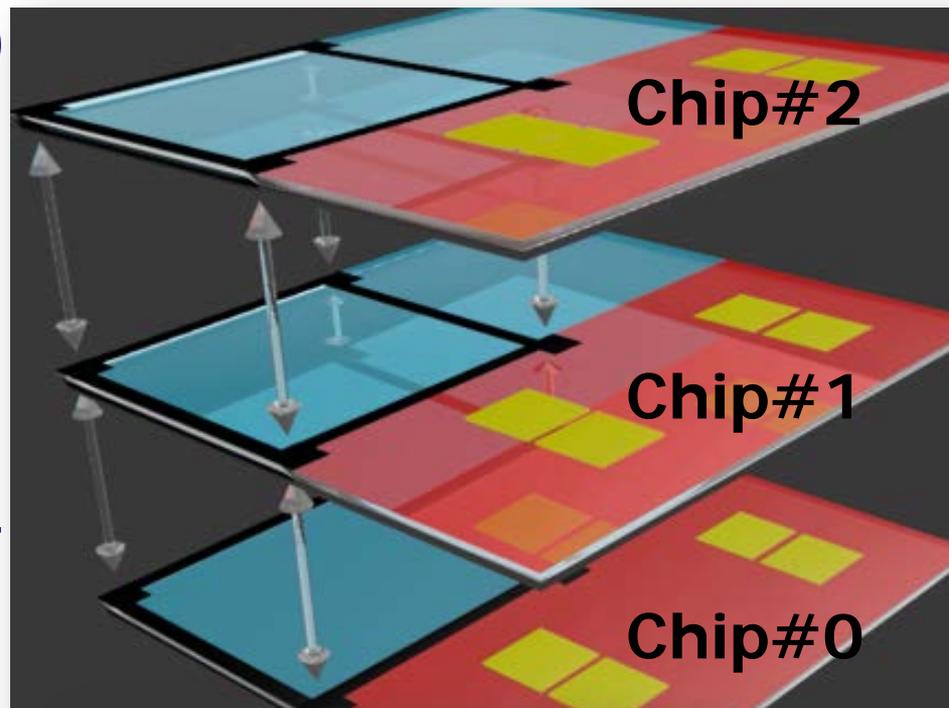
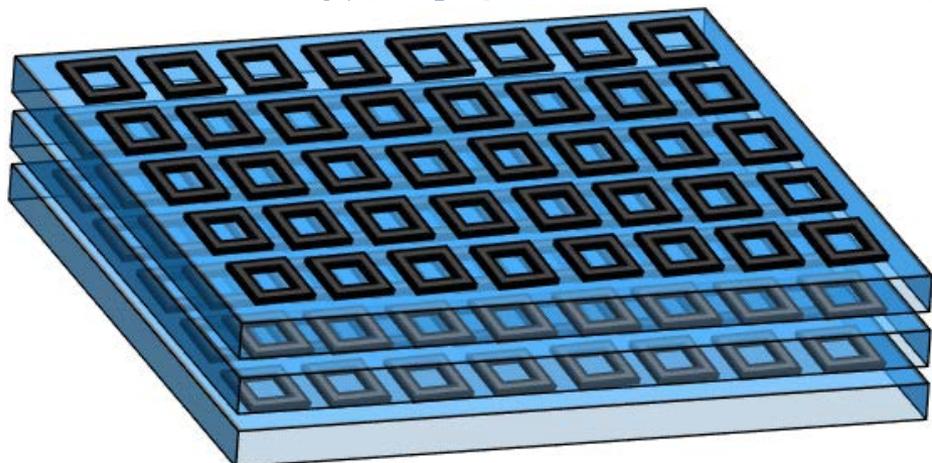
# ワイヤレス3次元チップマルチプロセッサ

- 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

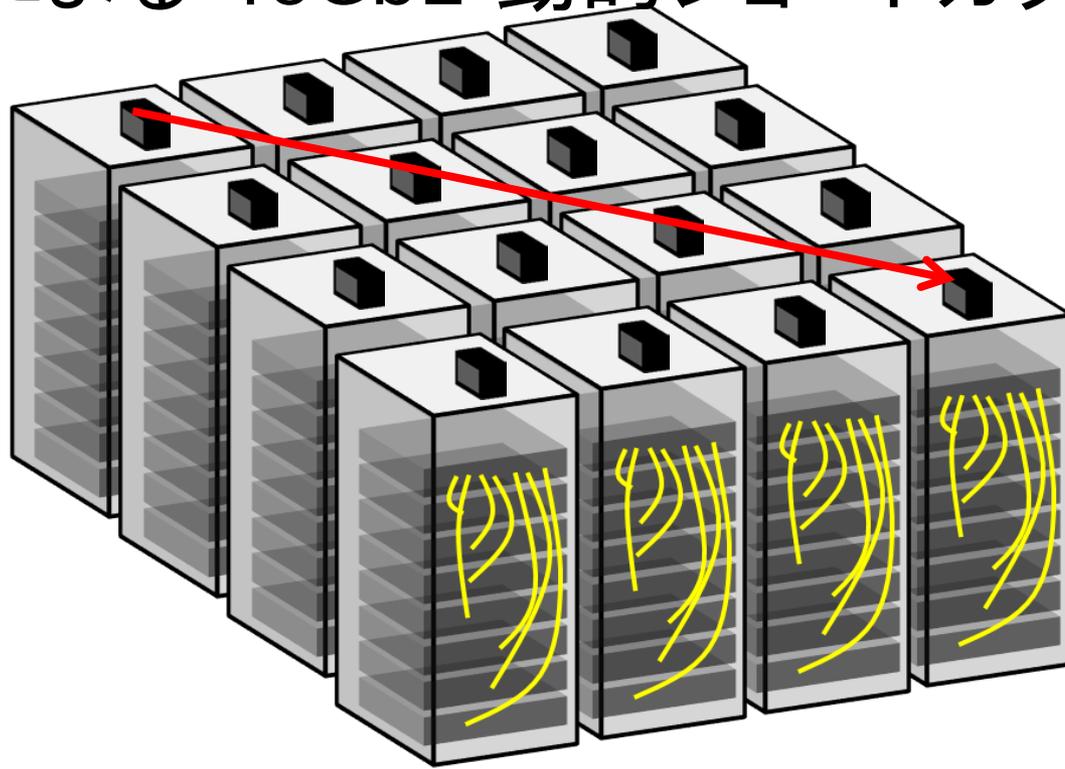
データ転送なので計算量は少なく、完全に I/Oネックとなる  
→ ネットワークを増強すべき

サーバに搭載された 40GbE

サーバに搭載された並列処理

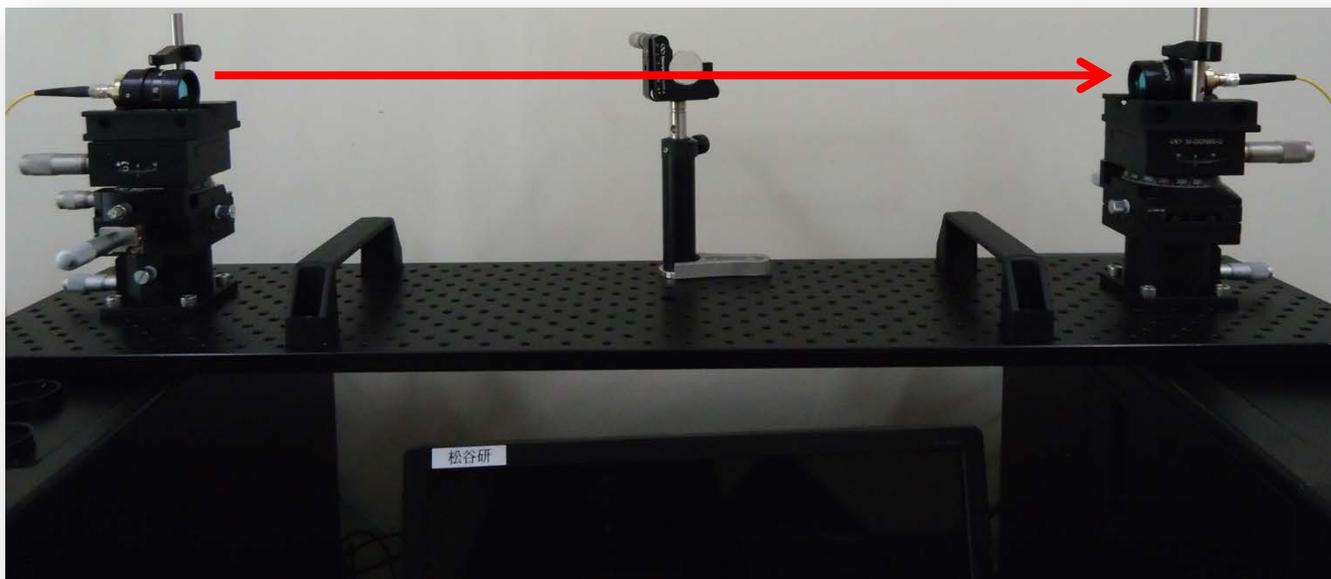
# 光無線による40GbE動的ショートカットリンク

- 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オーダー)
  - 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

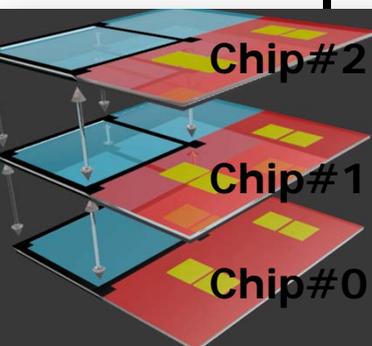
## 構造型ストレージ(NOSQL)

- 用途特化型データベース
  - リティの高
  - 大量のデータ
- 40GbE FPC  
いた DB



計算インテンシブ

I/Oインテンシブ



## 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

- -
- GeForce 780Ti GPU  
2,880 cores





***Any Questions?***

**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