



ビッグデータ利活用のための 計算基盤

慶應義塾大学 理工学部

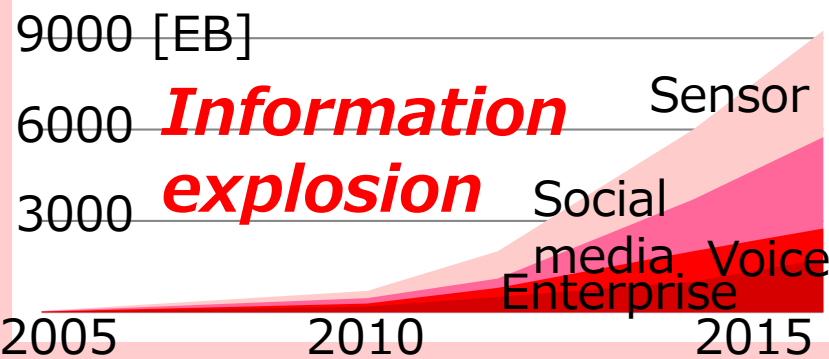
松谷宏紀

<http://www.arc.ics.keio.ac.jp/~matutani>

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



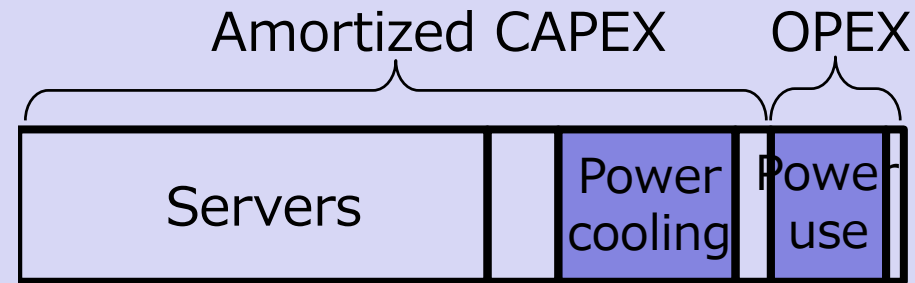
→ IT 機器の増強へ作用 (電力増)

- 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

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

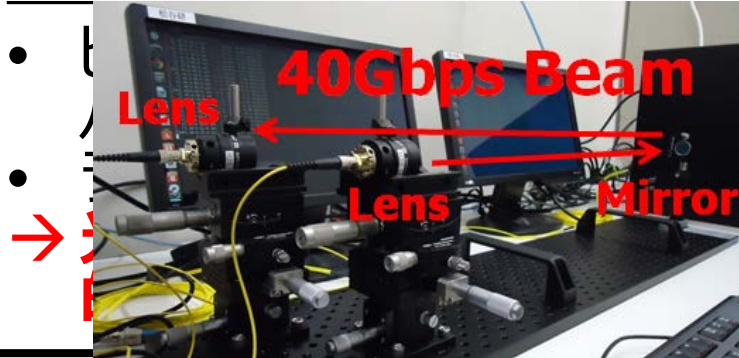


→ IT 機器の省電力化への要求

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

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



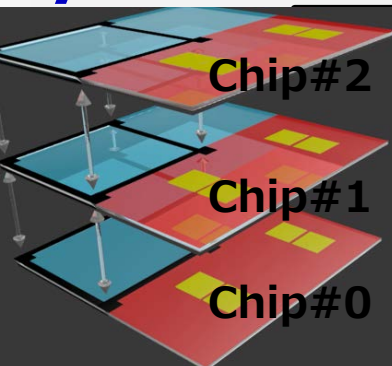
FPGA NICによるNOSQL処理

- キーバリ
→ HW/S
- 10GbE 機
- 莫大な
れ値以外



計算インテンシブ

I/Oインテンシブ



GPU 向け 3D 積層システム

プロセッサ、アクセラレータチップを3次元積層
チップ間通信はワイヤレス
KVSMメモリチップを開発

In-GPU 分散データベース

- GPU
 - ドキ
 - 10 GP
- Image of a GPU system with a Host and GPU. The text "GPUs 20GbE Host" is overlaid in red.

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

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
- ラック天井部から光ビーム
→ **光無線による 40GbE 動的リンク**

FPGA NICによるNOSQL処理

- キーバリューストアDB
→ **HW/SW 階層DBキャッシュ**
- 10GbE 機械学習 NIC**
- 莫大な受信データの中から外れ値以外をNICで捨てる

計算インテンシブ

I/Oインテンシブ

本研究では、ビッグデータ利活用の要であるデータベース（構造型ストレージ）をハードウェア化することによって、スループット性能を維持しつつ、多数のサーバを専用ハードウェアに置き換え、コストと電力効率の向上を狙う

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

Memcached

Key	Value
↓	
↓	
↓	
↓	

キーバリューストア型

Shopping cart, User profile, Session, etc

MongoDB

ドキュメント
指向型

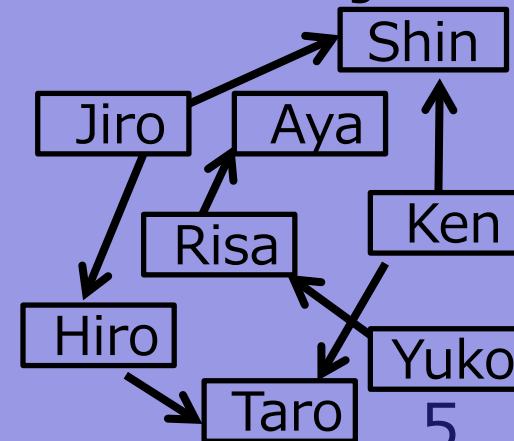
```
{ _id : ObjectId(0),  
  name : Risa,  
  tel : 1234 }  
{ _id : ObjectId(1),  
  name : Shin,  
  mail : shin@x.jp }
```

Schema-less DB

Neo4j

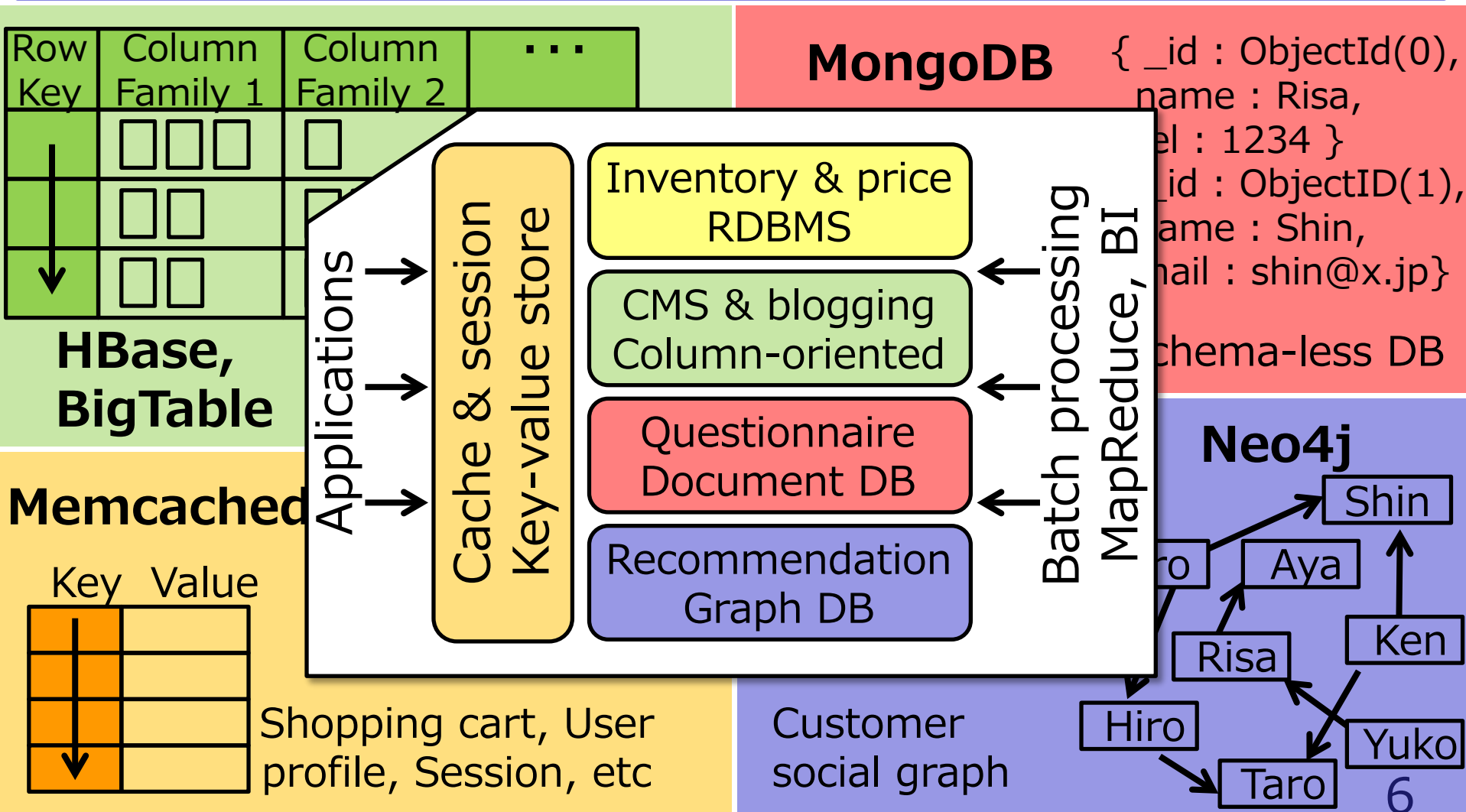
グラフ型

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

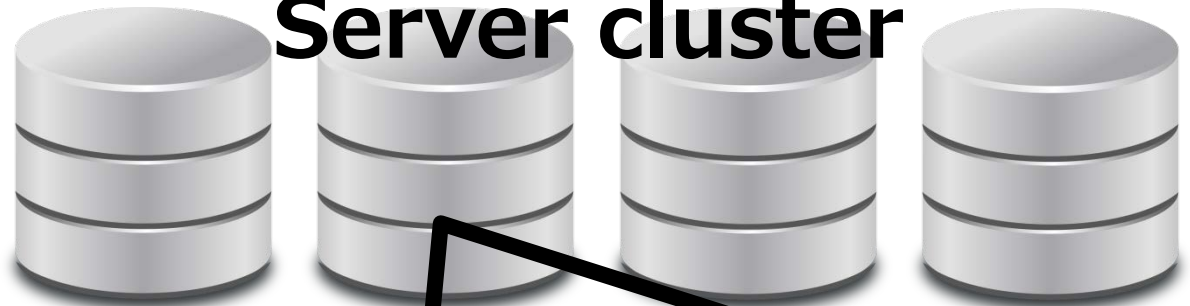


想定システム（2013年の夏から…）

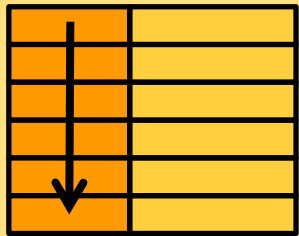
客層分析システム



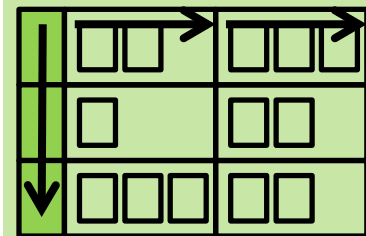
Server cluster



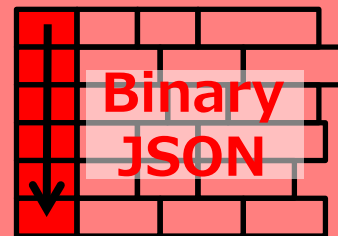
Key-value store



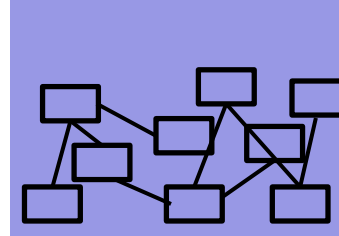
Column-oriented store



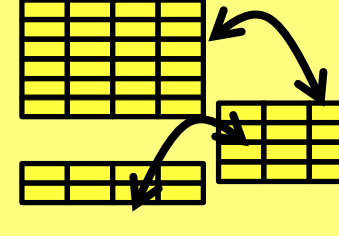
Document DB



Graph DB



RDBMS



Data analysis framework



hadoop

Spark

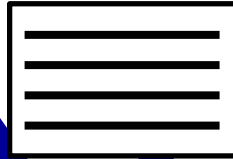
Streaming

(1) 各種 NOSQL 及び Spark Streaming が密結合されたシステムを想定し、(2) それらを FPGA/GPU/In-Kernel キャッシュなどを使って高効率化する

階層 NOSQL キャッシュ: FPGA NIC

通常のデータベースアクセス

Result



In-Kernel KVS
Cache (L2\$) Hit

Add

NIC Cache
(L1\$) Hit

Request

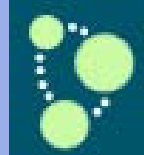
Reply

User Space



mongoDB

APACHE
HBASE



Neo4j

NetFPGA-10G Device Driver

In-Kernel KVS
(Key-Value Store)
Cache 512GB

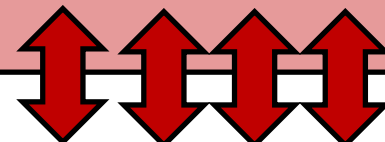


NetFPGA-10G

NIC HW Cache
Machine Learning



DRAM 288MB



10GbE x4 9

階層 NOSQL キャッシュ: FPGA NIC

階層 NOSQL キャッシュ:

L1 NOSQL cache ... FPGA NIC による HW キャッシュ

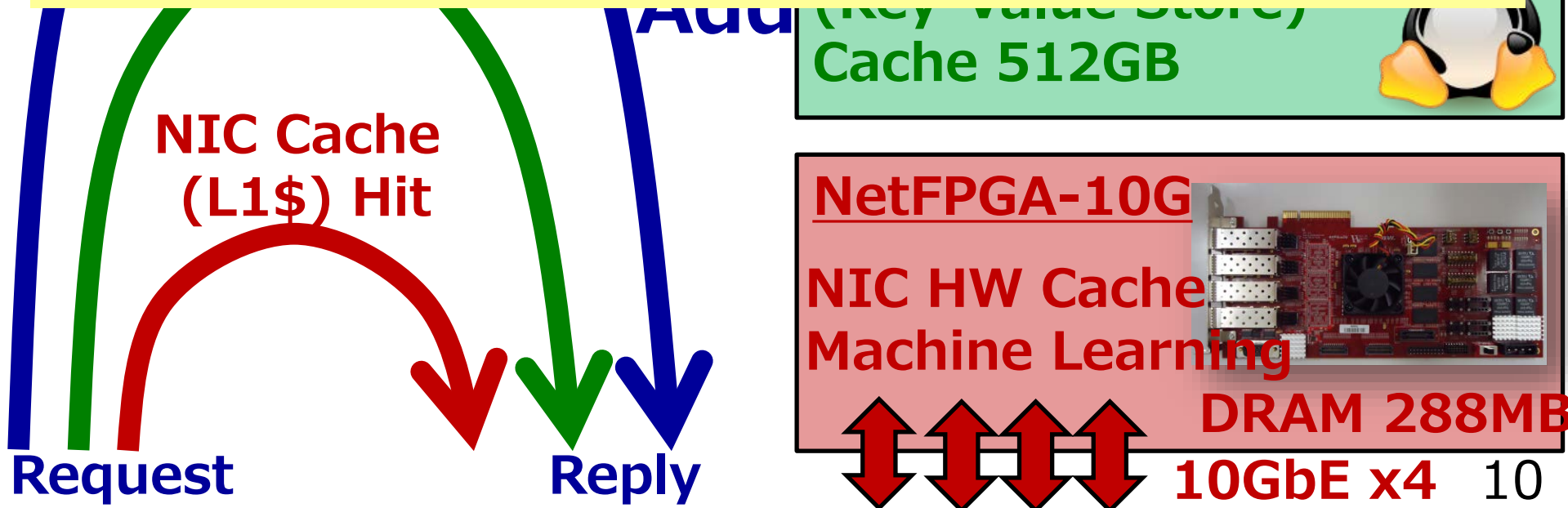
L2 NOSQL cache ... Linux カーネル内の SW キャッシュ

キャッシュ容量と速度のトレードオフ:

L1 NOSQL cache ... 高速・高効率だが小容量

L2 NOSQL cache ... 高速で大容量

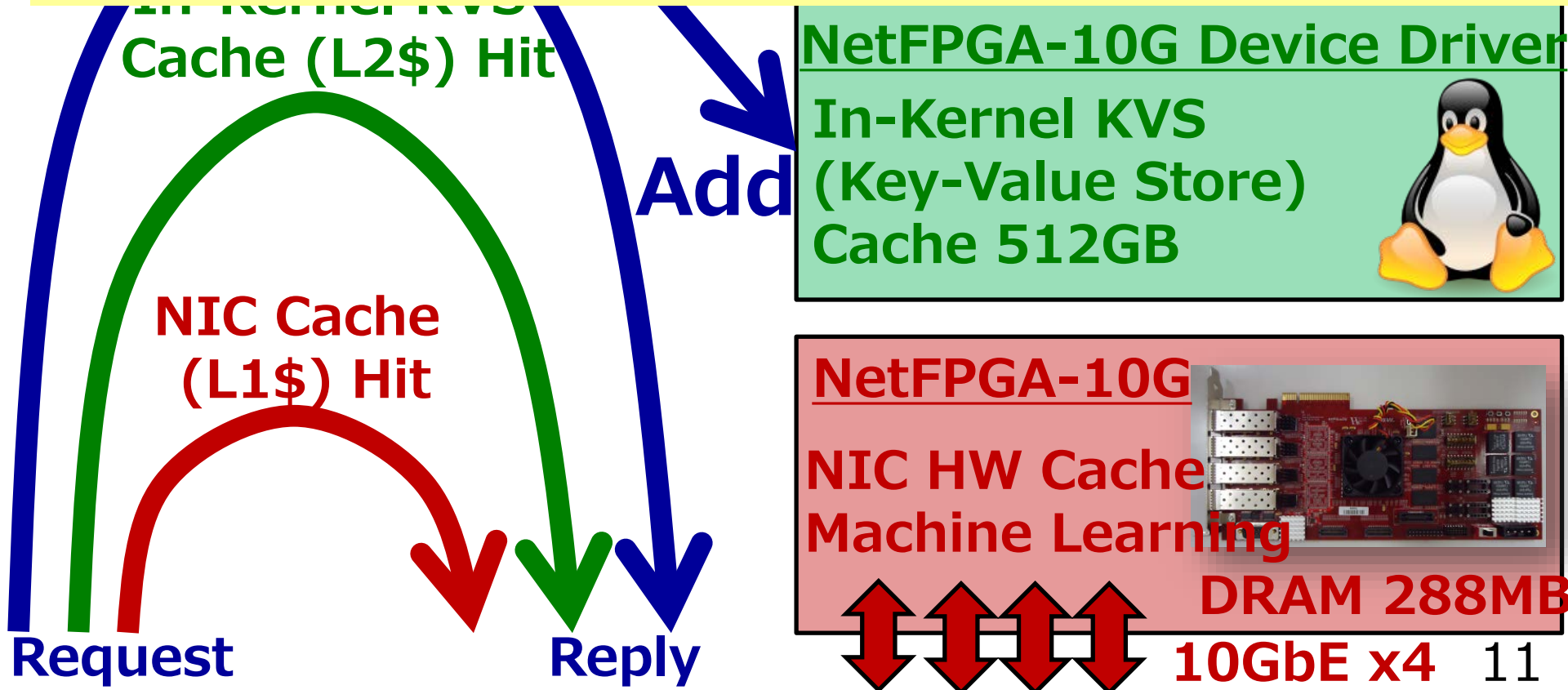
階層キャッシュ設計空間の探索 → [IEEE HoTI'16]



階層 NOSQL キャッシュ: FPGA NIC

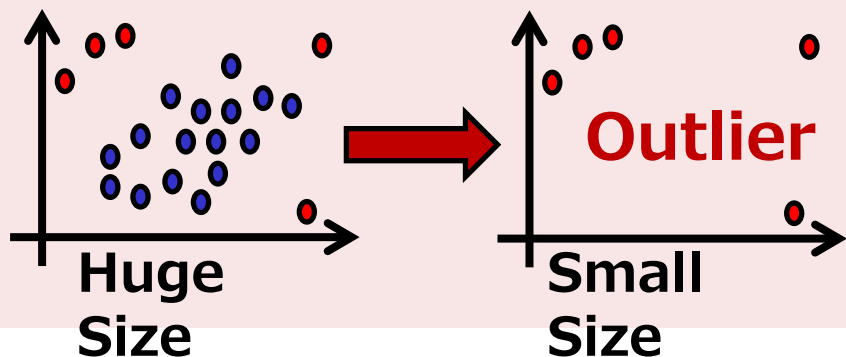
アプリケーション:

- データベースキャッシュ (KVS、カラム、ドキュメント、グラフ)
- ビットコイン (Blockchain) 公開取引台帳のキャッシング
- 深層学習パラメータサーバ (確率的勾配降下法の集約)
- DDoS Mitigation Box



FPGA NIC による 10Gbps 外れ値検出

Sensor Data Explosion



機械学習アルゴリズム

- ✓ Mahalanobis Distance
- ✓ Local Outlier Factor (LOF)
- ✓ K-Nearest Neighbor (KNN)

User Space



NetFPGA-10G Device Driver

Only anomaly-valued packets are received

Data Mining

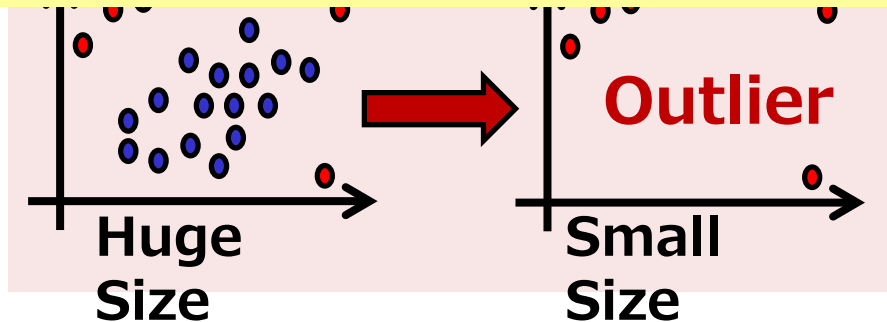


10GbE x4 12

FPGA NIC による 10Gbps 外れ値検出

マハラノビス距離版:

- 次元ごとに分散を考慮した距離（例：分散が小さい次元ではユークリッド距離よりも遠く扱われる）
- 教師なし外れ値検出（古典的手法）
- 14M samples/sec（10GbEラインレートの95.8%）
[HEART'15（Best Paper Award）]

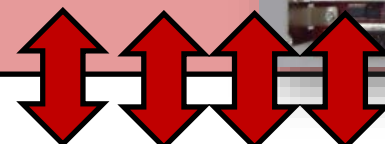


機械学習アルゴリズム

- ✓ Mahalanobis Distance
- ✓ Local Outlier Factor (LOF)
- ✓ K-Nearest Neighbor (KNN)

valued packets are received

Data Mining



10GbE x4 13

FPGA NIC による 10Gbps 外れ値検出

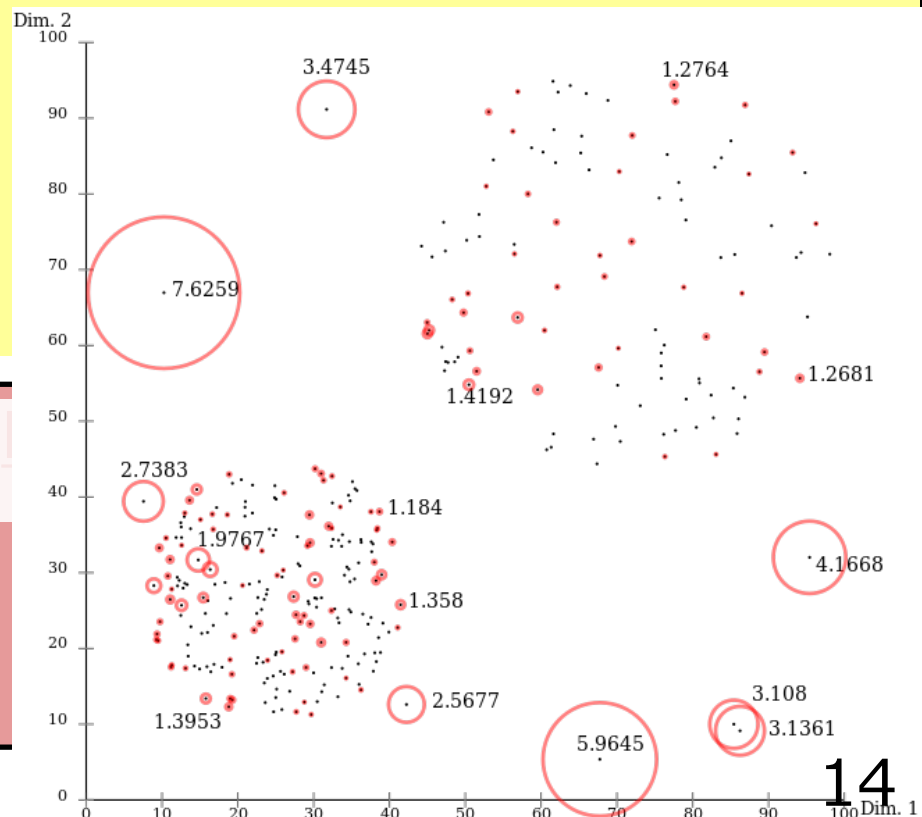
LOF版（およびKNN版）：

- 判定には密度を利用（密度 = K近傍の距離）
- 密度が低い → LOF 値が高い
- 密度計算には過去データが必要だが FPGA NIC には全部は置けない
- 代表的クラスタ（過去データ）を FPGA NIC でキャッシュ [PDP'17]



機械学習アルゴリズム

- ✓ Mahalanobis Distance
- ✓ Local Outlier Factor (LOF)
- ✓ K-Nearest Neighbor (KNN)



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

ストレージマイグレーション・

FPGA NICによるNOSQL 処理

グラフ型 DB: グラフ探索（ソーシャルグラフの解析）

ドキュメント指向 DB: 文字列探索（正規表現による探索）

→ GPU や NEC ExpEther を活用したデータベース処理

ACM CAN (2014)、ISPA'15、HeteroPar'16、ICPADS'16

計算インテンシブ

I/Oインテンシブ

IoT 向け 3D 積層システム

- プロセッサ、アクセラレータチップを3次元積層
 - チップ間通信はワイヤレス
- **KVSメモリチップを開発**

In-GPU 分散データベース

- グラフ型DB（グラフ探索）
 - ドキュメントDB（文字列）
- **10GbEネットワーク接続 GPUクラスタ**

GPU による NOSQL の高速化

User Space



mongoDB

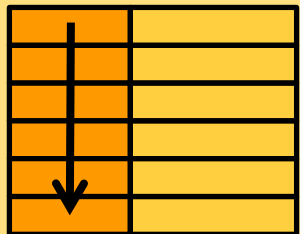
ドキュメント指向 DB (MongoDB) :

正規表現ベースの文字列探索は CUDA を呼び出す
[ISPA'15]

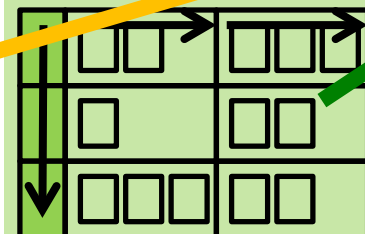
グラフ型 DB (Neo4j) :

グラフ探索 (ソーシャルグラフの解析) は CUDA を呼び出す
[ACM CAN(2014)]

Key-value
store



Column-
oriented store



NetFPGA-10G

NIC HW Cache
Machine

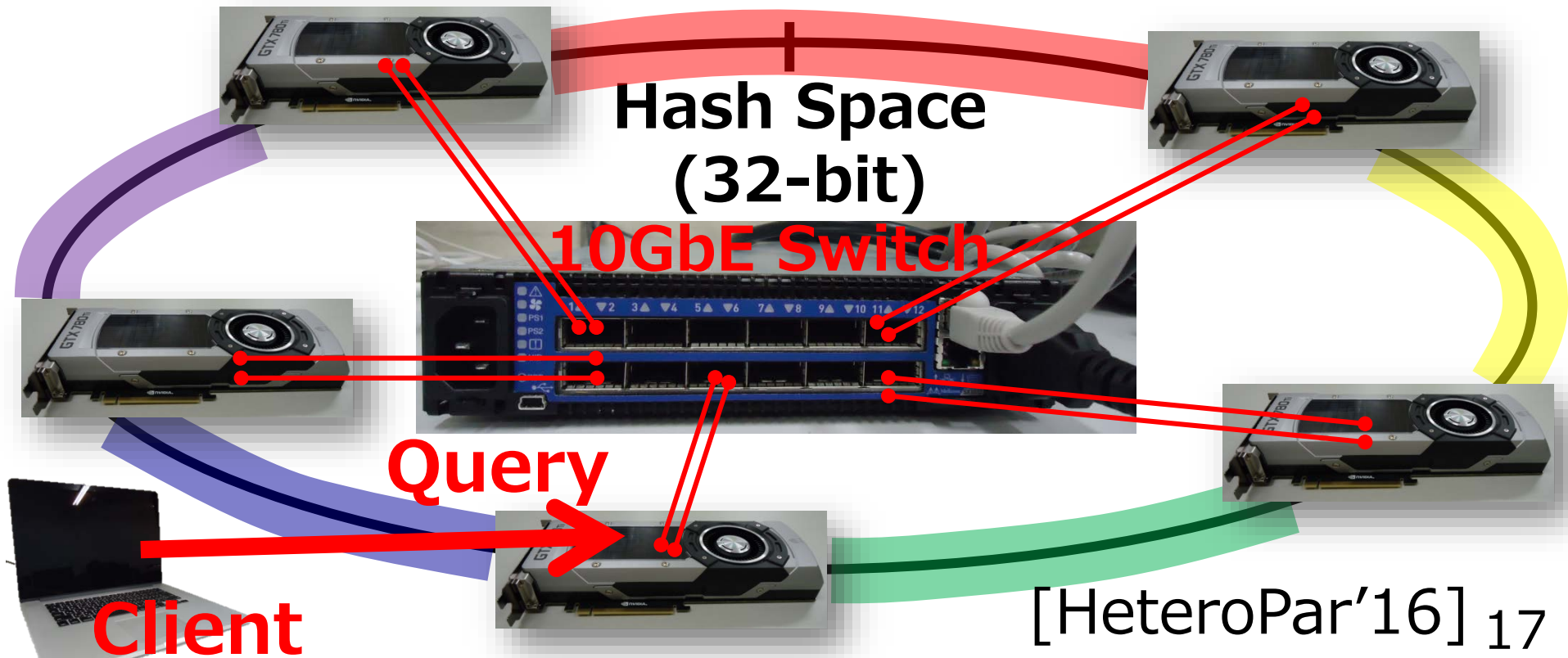
VERY FAST

10GbE x4 16

分散 In-GPU データベース

さらに多くのGPUを活用したい → 分散 In-GPU DB

- 分散 In-GPU データベース over NEC ExpEther
 - GPU のデバイスメモリを「**DB キャッシュ**」として利用
 - 多数の GPU デバイスを 10GbE スイッチで接続



分散 In-GPU データベース

Many GPUs are directly connected to DB server via NEC ExpEther (20Gbps)

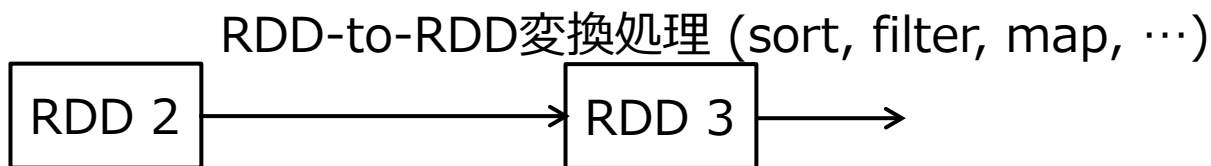
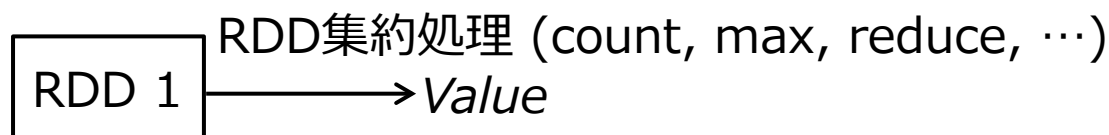
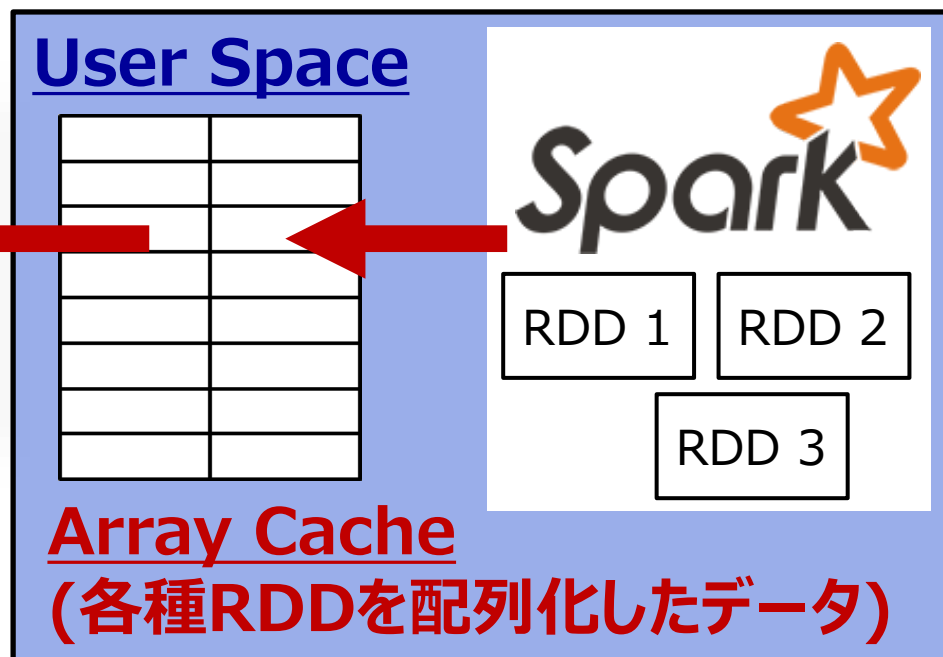


GPU + 分散処理フレームワーク Spark



Array Cache (RDDs)

**RDDの集約処理とRDD-to-RDD
変換処理を GPU にオフロード**

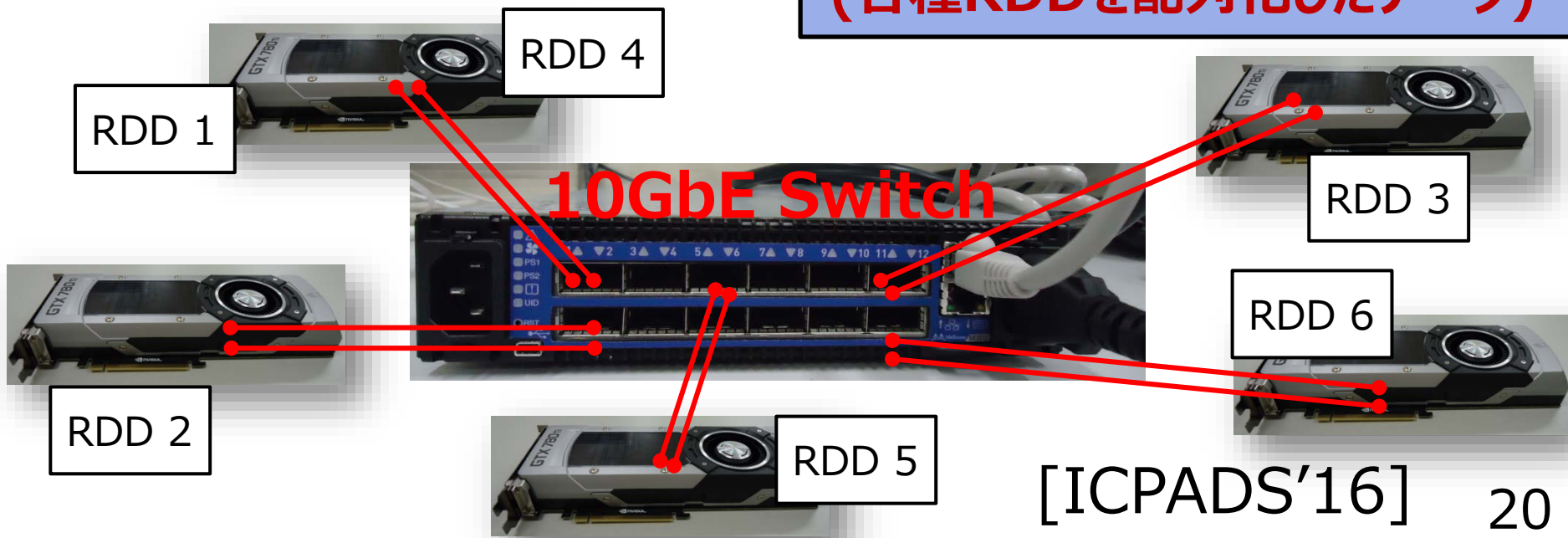
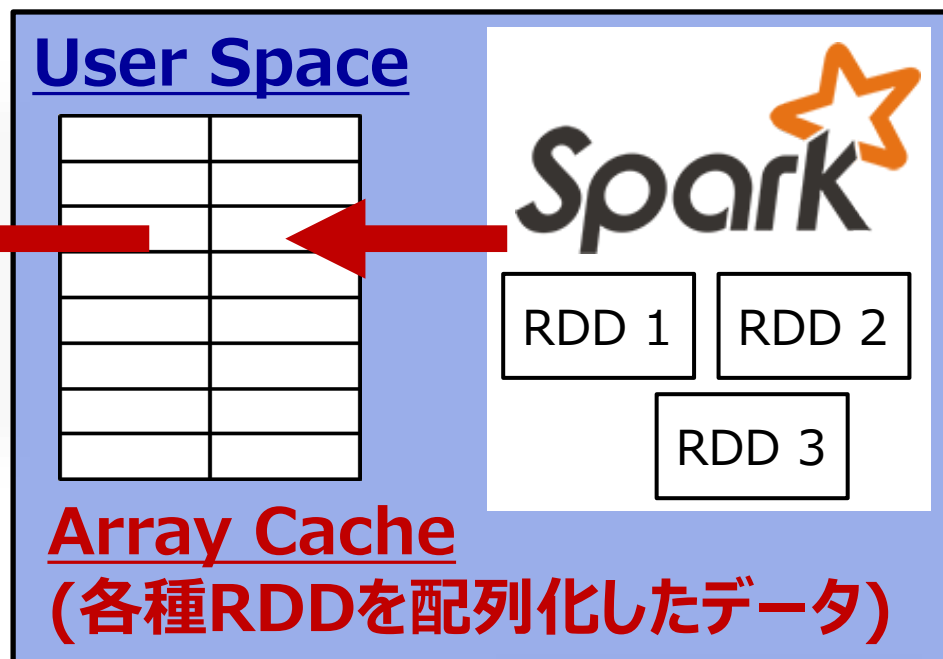


Spark ではデータは RDD と呼ばれる分散共有メモリに格納
RDD を GPU 処理に適した配列 (Array Cache) としてGPUへ転送

GPU + 分散処理フレームワーク Spark



**RDDの集約処理とRDD-to-RDD
変換処理を GPU にオフロード**



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

今までと一転、ビッグデータを集める側のデバイスに着目

IoT 向け 3次元チップ積層 (System-in-Package)

低消費電力、および、機能（チップ）の組み換え可能性

→FPGA アクセラレータの成果を「小さいほう」にも展開

的リンク

い恒以外をNICで括する

計算インテンシブ

I/Oインテンシブ

IoT 向け 3D 積層システム

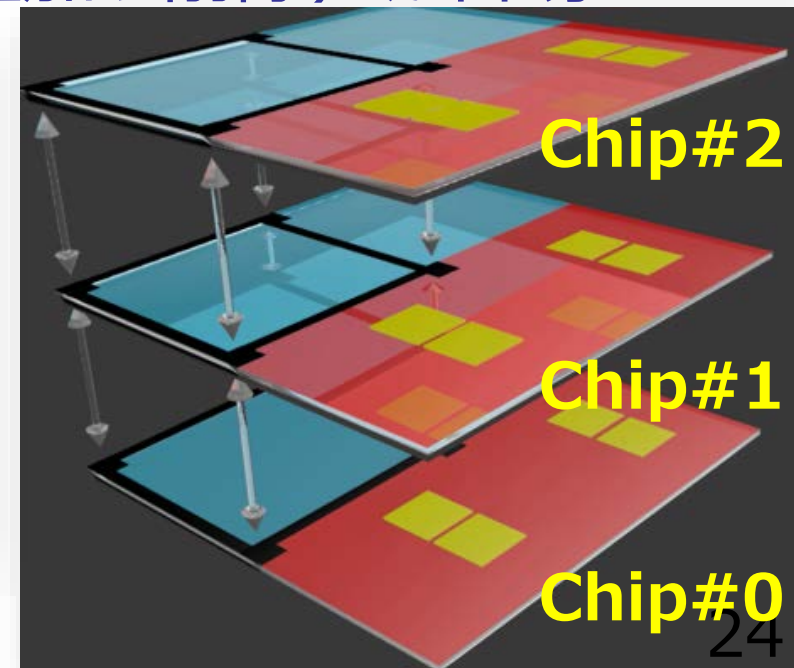
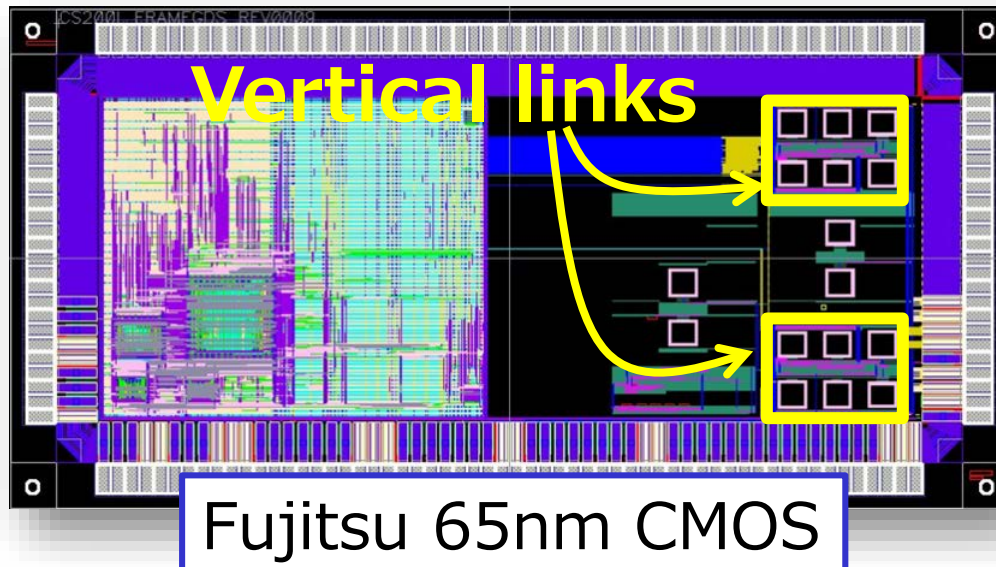
- プロセッサ、アクセラレータチップを3次元積層
 - チップ間通信はワイヤレス
- **KVSメモリチップを開発**

In-GPU 分散データベース

- グラフ型DB（グラフ探索）
 - ドキュメントDB（文字列）
- **10GbEネットワーク接続 GPUクラスタ**

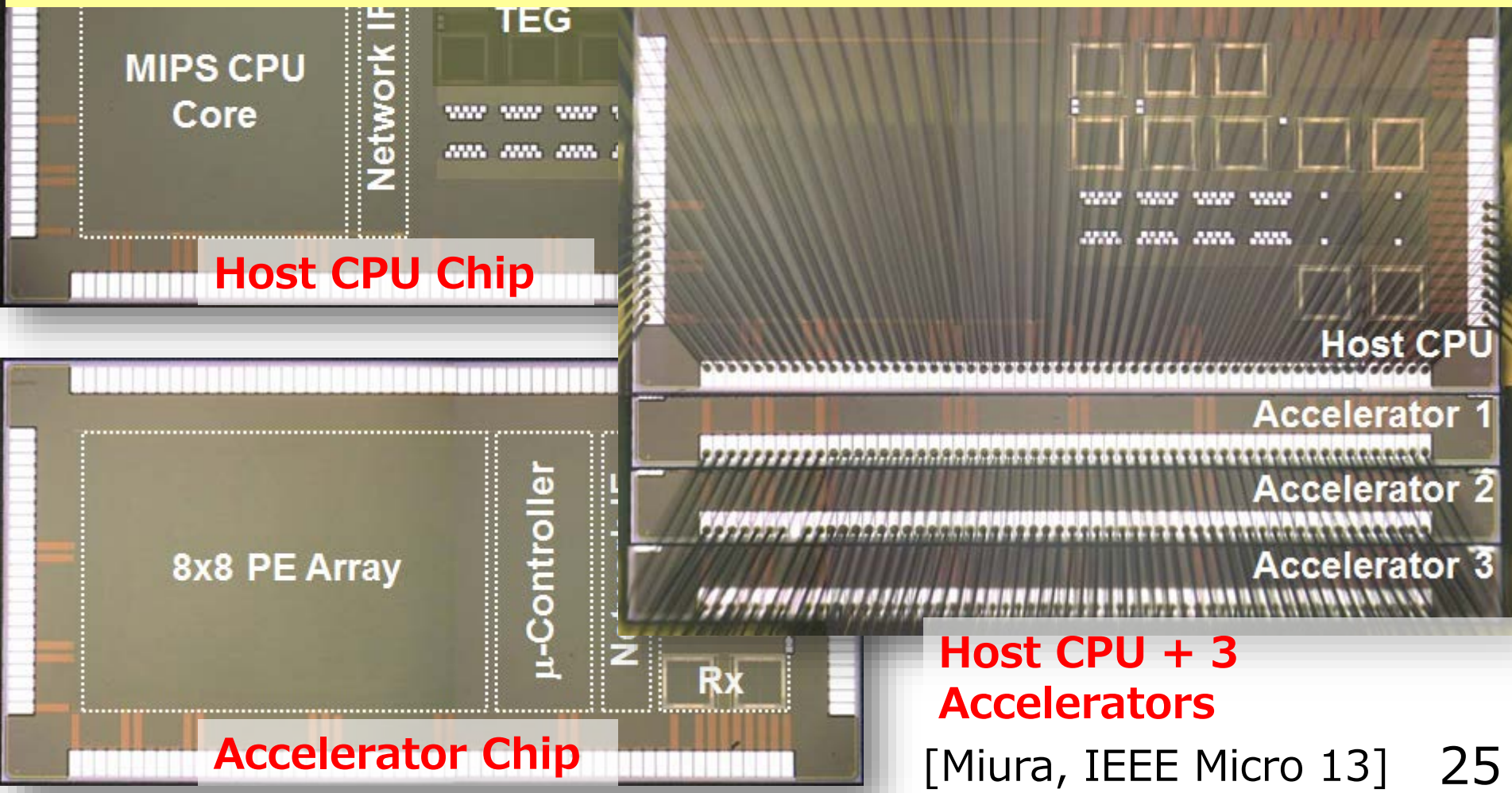
IoT 向けワイヤレス 3 次元チップ積層

- 誘導結合を用いた System-in-Package (SiP)
 - 必要なチップを積み重ねてシステムを構築
 - CPU チップ、アクセラレータチップ、メモリチップ、...
- 誘導結合の利点
 - チップ間は電氣的に非接触
 - SiP 内のチップの入れ替え（追加や削除）が容易



Cube-1 (実用的プロトタイプ実装)

チップ間ワイヤレス接続 → CPU チップ、メモリチップ、アクセラレータチップを必要に応じて組み合わせることができる
IEEE Trans Comp (2014)、IEEE Trans VLSI (2016)



**Host CPU + 3
Accelerators**

IoT 向け 3 次元積層システム (SiP)

- 各チップは送受信機を装備
 - Geyser チップ (CPU)
 - CMA チップ (アクセラレータ)
 - Neuro チップ (推論)
 - KVS チップ (SRAM ストレージ)
- KVS メモリチップの詳細

Processor Chip

Amano (Keio U)

Data are stored
as KVS pairs

Accelerator Chip

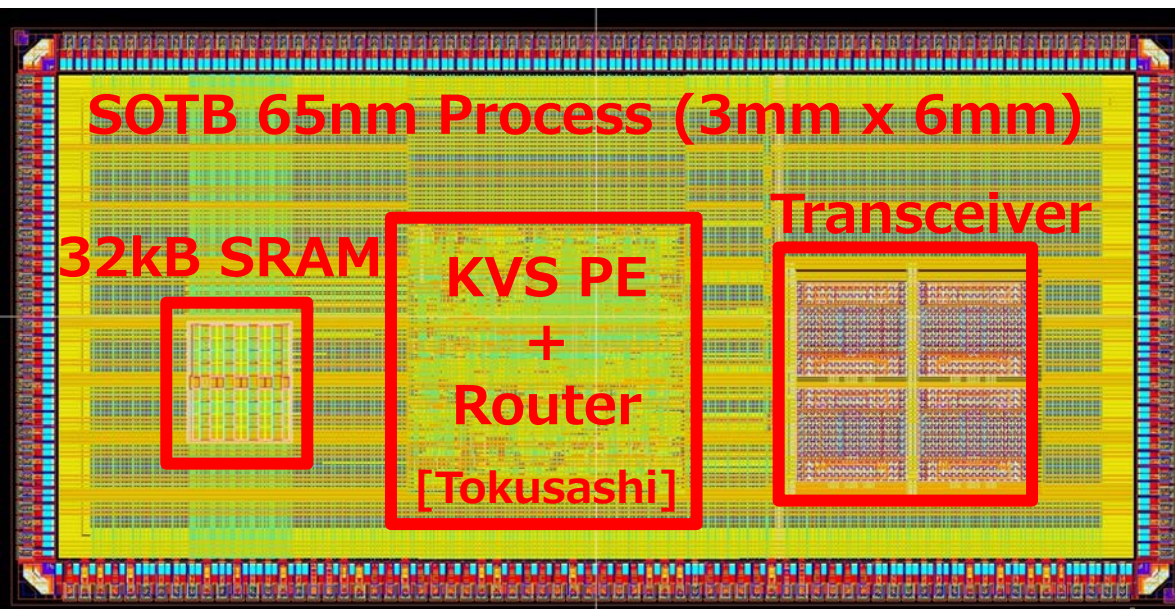
Amano (Keio U)

KVS Memory Chip

Matsutani (Keio U)

Neuro Chip

Kondo (U Tokyo)



IoT 向け 3 次元積層システム (SiP)

命令は 2 つだけ:

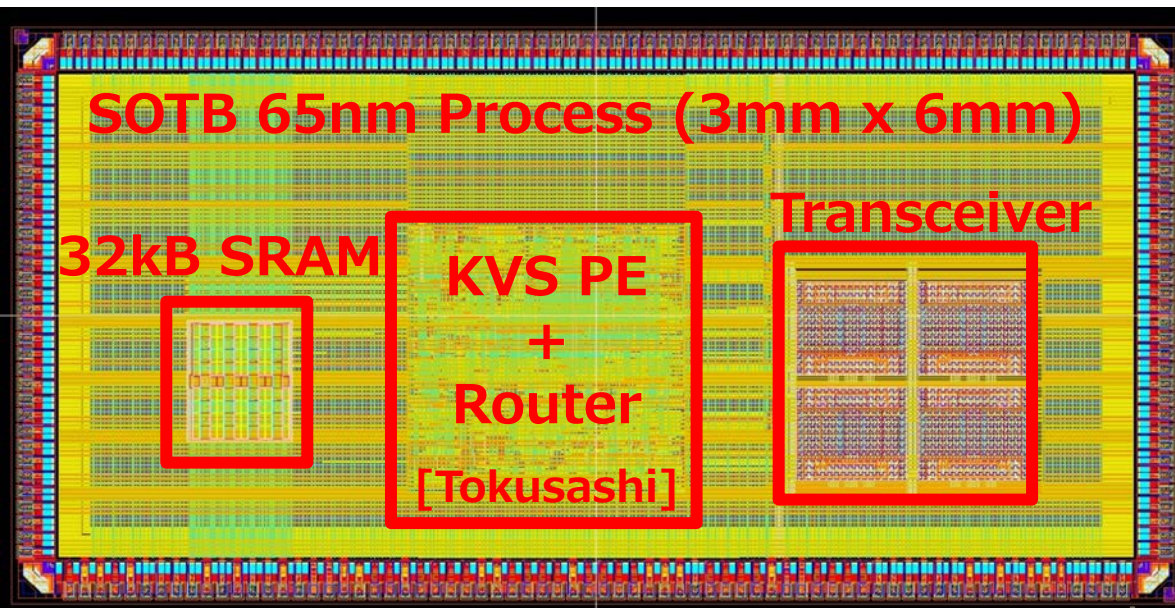
- **Set(Key, Value)**

TCI 経由でやってきた Key と Value の組をメモリに保存

- **Get(Key)**

TCI 経由でやってきた Key に対応する Value をメモリから読みだして送信元に返信

- KVS メモリチップの詳細



KVS Memory Chip
Matsutani (Keio U)

Neuro Chip
Kondo (U Tokyo)

IoT 向け 3 次元積層システム (SiP)

アプリケーション:

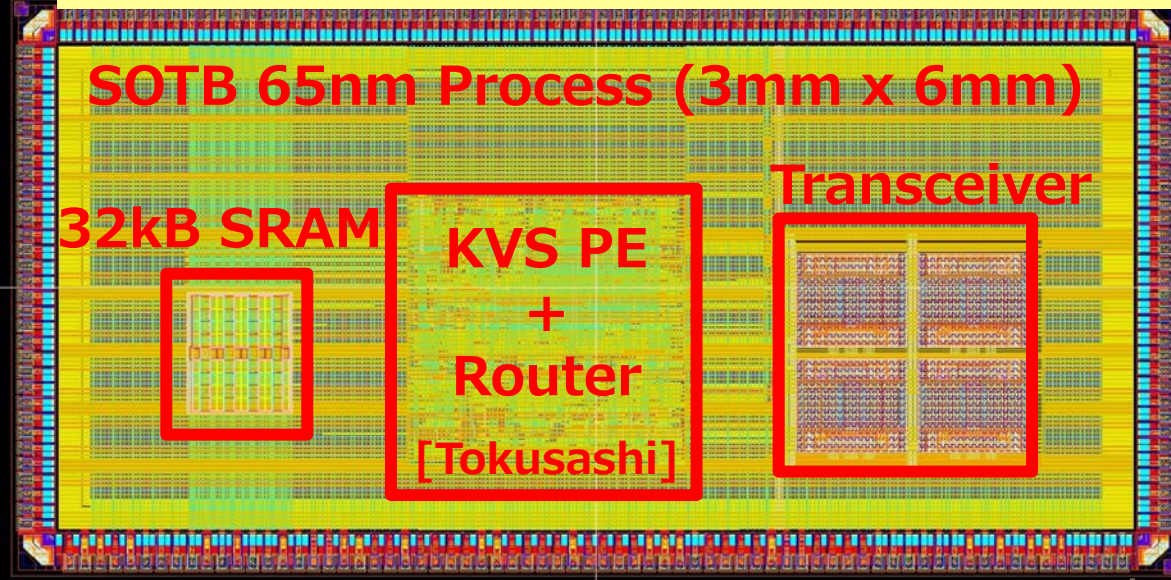
類似の入力値に対し得られたプロセッサの計算結果を KVS メモリに記憶し、計算結果を再利用 (計算をサボる)

Key: ハッシュ関数 (関数名 + 入力値の上位ビット)

例: GPS位置情報 + ジャイロセンサ (角度)
入力画像

Value: その関数の計算結果

例: VGAへの出力画像
Neuro Chipの推論結果



KVS Memory Chip
Matsutani (Keio U)

Neuro Chip
Kondo (U Tokyo)

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

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
- ラック天井部から光ビーム
→ **光無線による 40GbE 動的リンク**

FPGA NICによるNOSQL処理

- キーバリューストアDB
→ **HW/SW 階層DBキャッシュ**
- 10GbE 機械学習 NIC
- 莫大な受信データの中から外れ値以外をNICで捨てる

計算インテンシブ

I/Oインテンシブ

IoT 向け 3D 積層システム

In-GPU 分散データベース

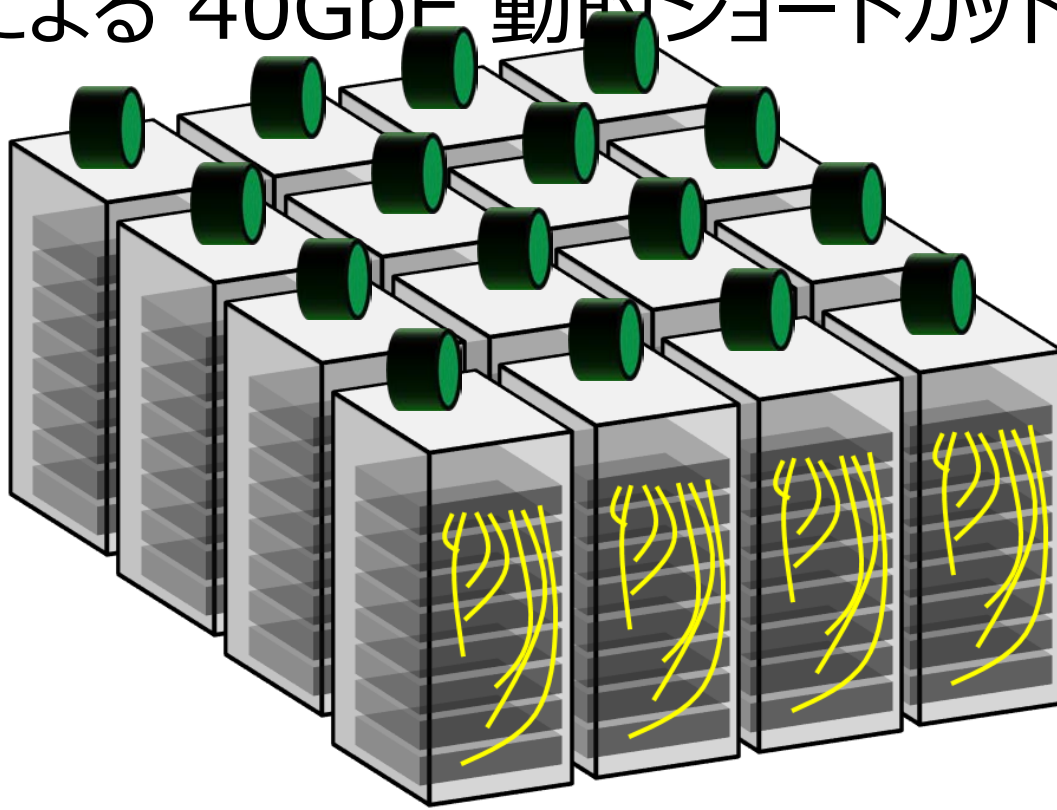
データ転送なので計算量は少なく、完全に I/Oネックとなる
→ ネットワークを増強すべき

→ **KVSメモリチップを開発**

GPUクラスタ

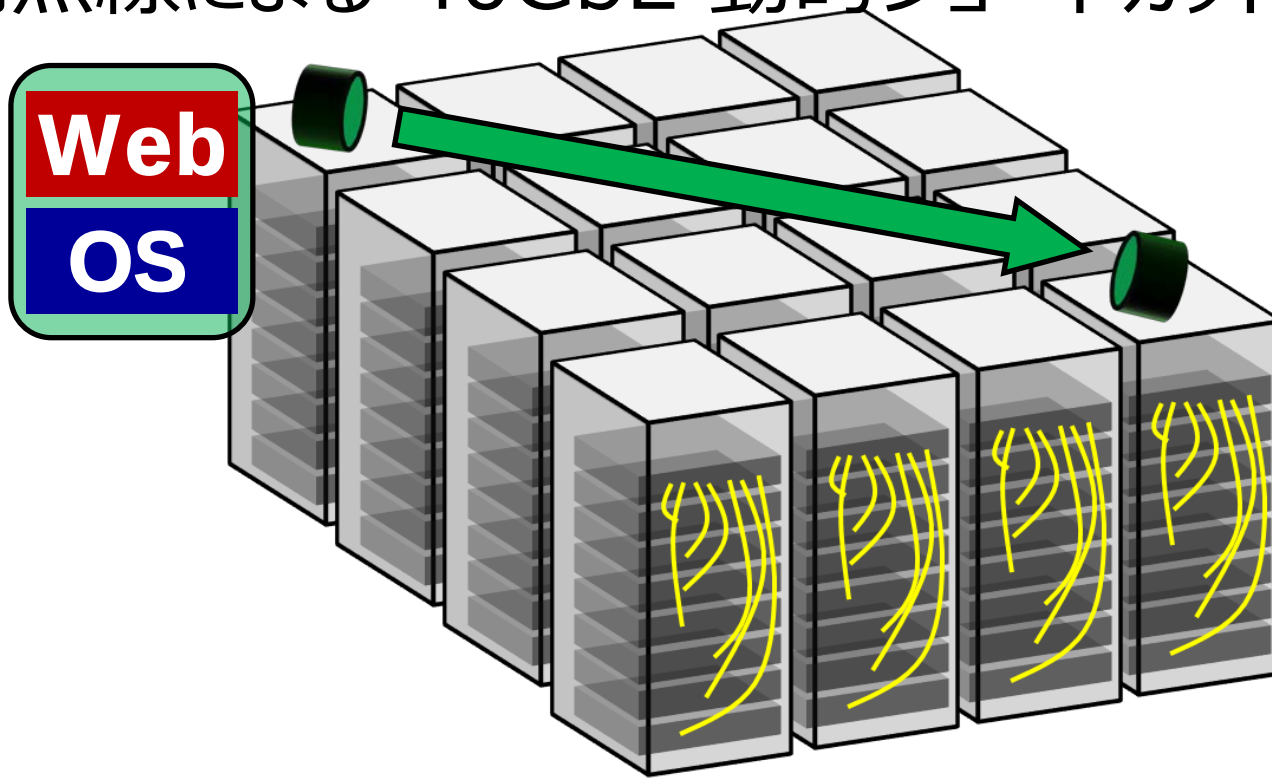
40GbE 光無線による仮想マシンの移送

- 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）
 - ストレージも一緒に移送したい（Web、データベース、…）
 - 1GbE リンクでは移送に時間が掛かったり、収束しない
- 光無線による 40GbE 動的ショートカットリンク



40GbE 光無線による仮想マシンの移送

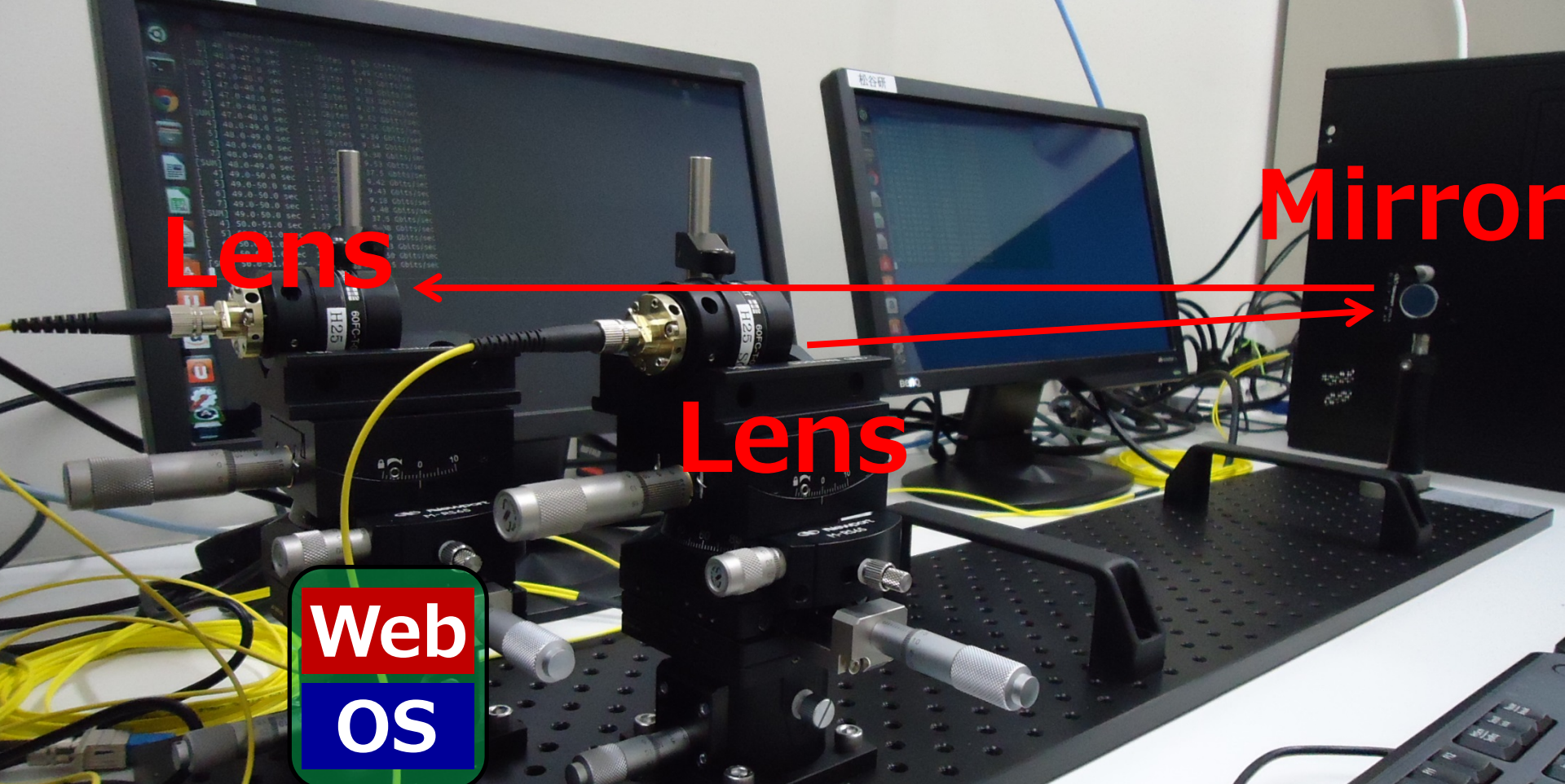
- データセンターにおける突発的データ転送
 - 仮想マシンの移送（数GBオーダ）
 - ストレージも一緒に移送したい（Web、データベース、…）
 - 1GbE リンクでは移送に時間が掛かったり、収束しない
- 光無線による 40GbE 動的ショートカットリンク



40GbE 光無線による仮想マシンの移送

- 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し、
レンズの向きを調整することで動的にリンクを形成

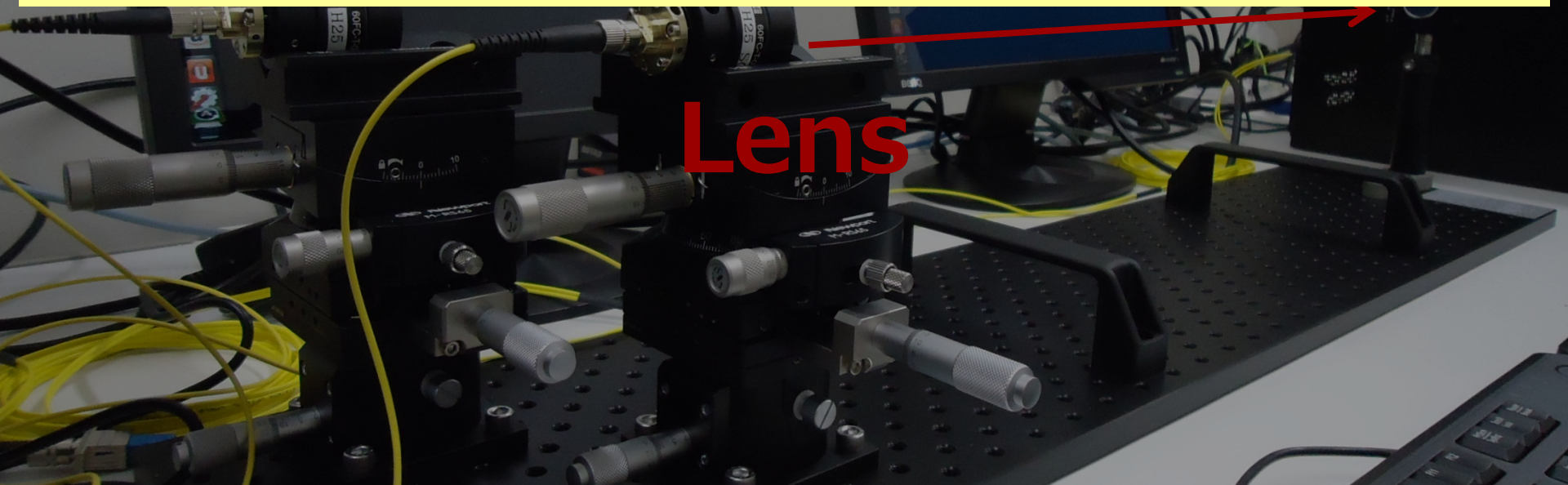


40GbE 光無線による仮想マシンの移送

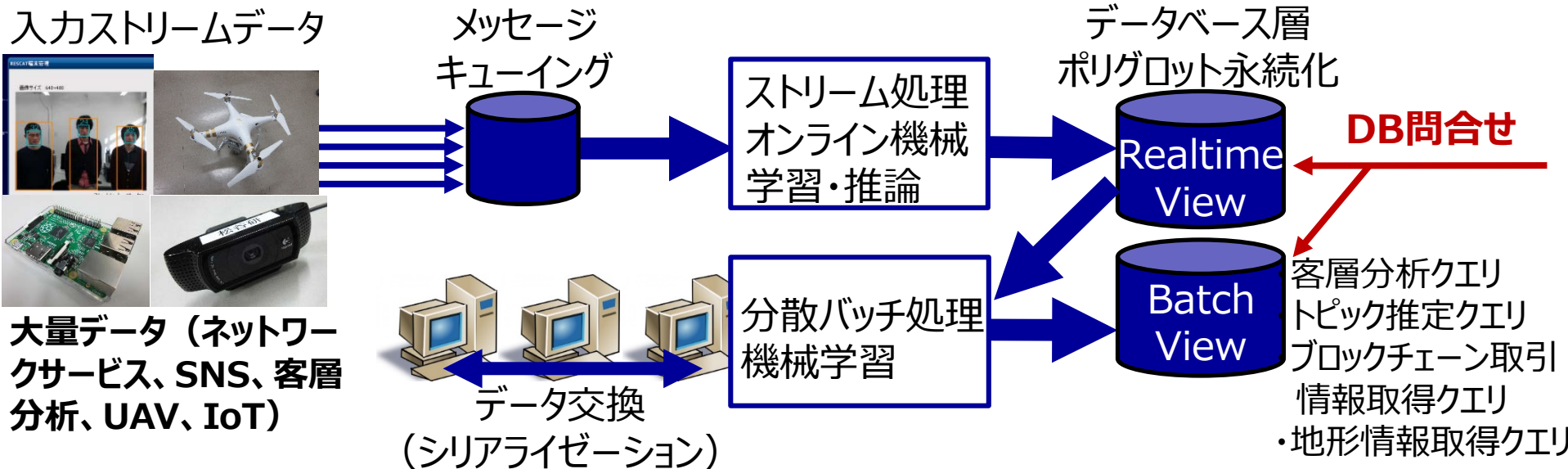
VM Highway: 仮想マシンの移送、ストレージ移送などの大容量データ転送に合わせてラック間コリメータレンズを調整し、40GbE 動的ショートカットリンクを形成

省電力化: FSO によるショートカットリンクによる性能向上→低貢献リンクの停止による省電力化

HPCA'15 (NICT藤原先生／NII鯉渕先生との共同研究)

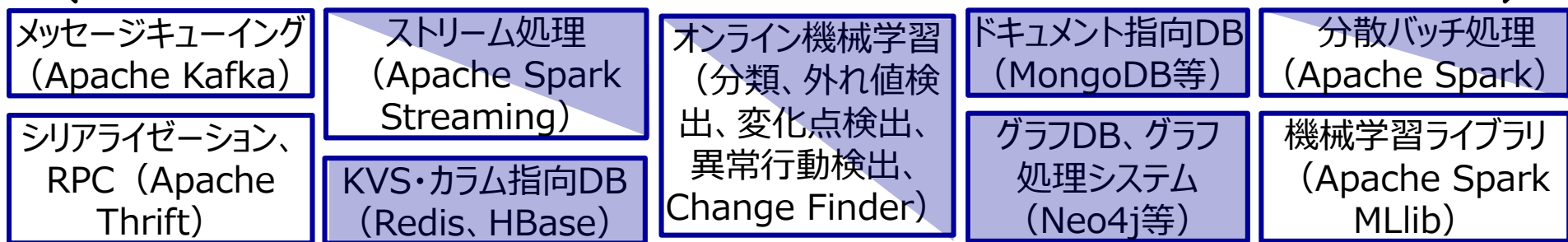


ビッグデータ基盤システムと今後の展望



I/Oインテンシブ

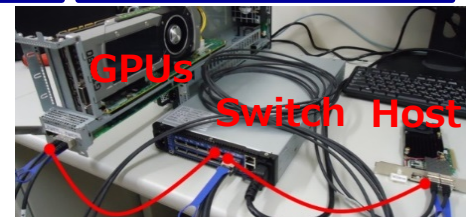
計算インテンシブ



I/Oと計算の密結合
→FPGA NIC

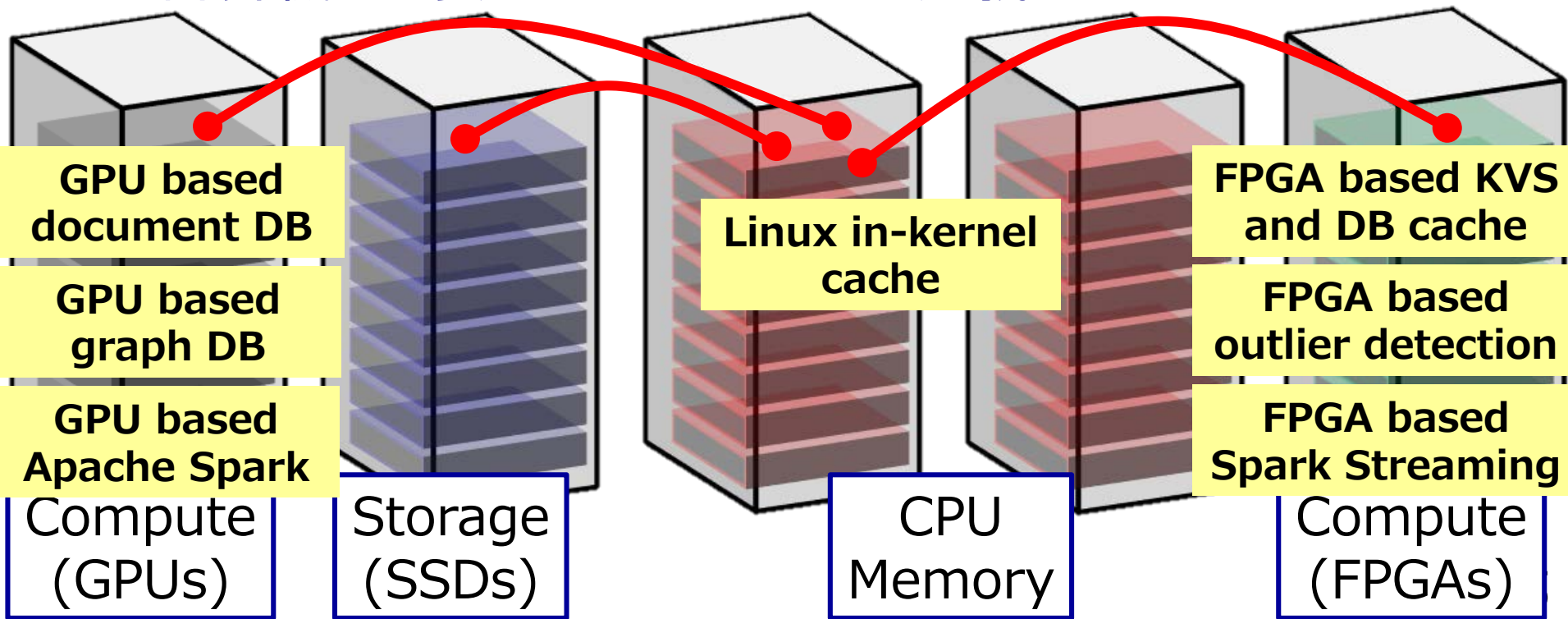


高い計算強度
→ネットワーク
直結GPUクラスター



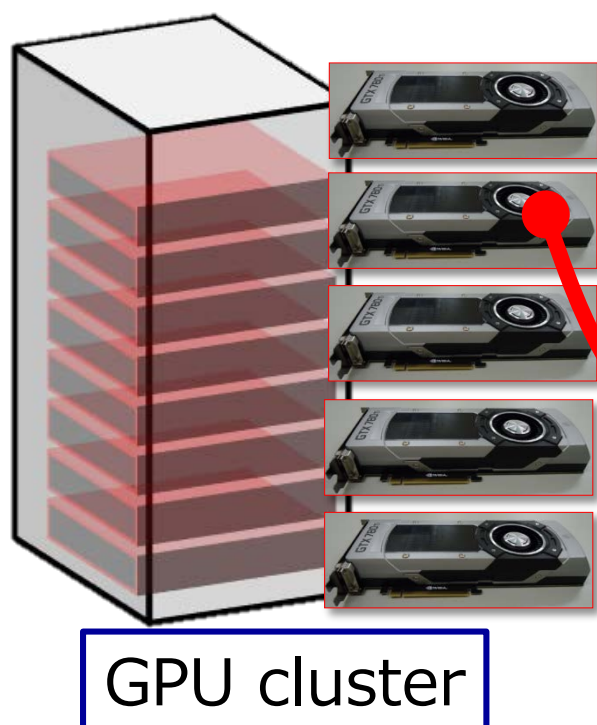
ラックスケールアーキテクチャへの展開

- 多数の計算リソースをラック内に収容
 - 例: CPU、ストレージ (RAM、SSD) 、計算 (GPU、FPGA)
- Software-defined infrastructure
 - ラック内、ラック間の計算リソースをネットワーク結合
 - 計算機を必要に応じてソフトウェア定義

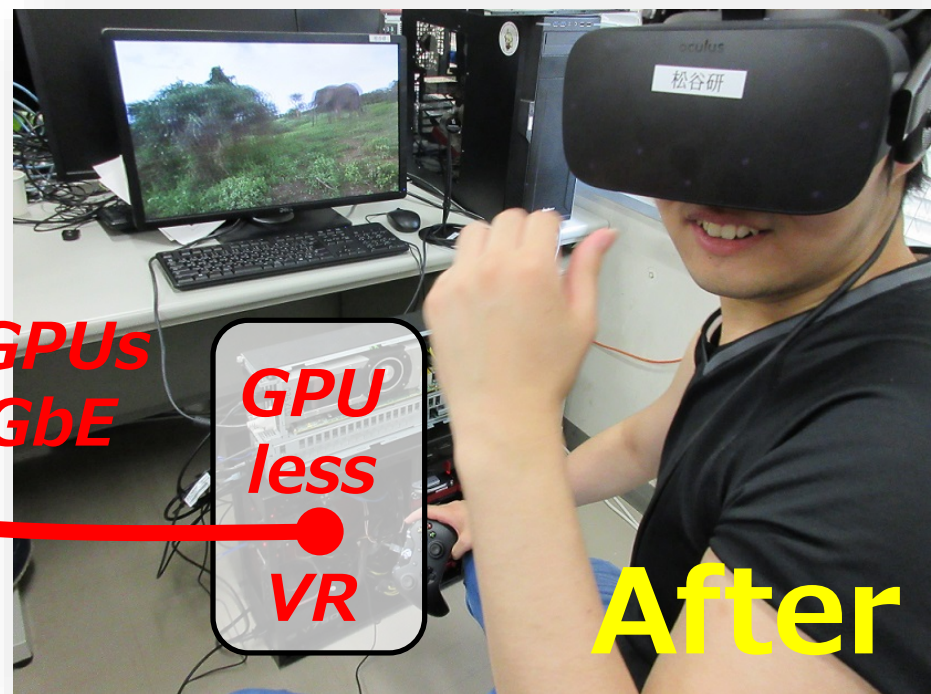


ラックスケールアーキテクチャへの応用

- 多数の計算リソースをラック内に収容
 - 例: CPU、ストレージ (RAM、SSD) 、計算 (GPU、FPGA)
- Software-defined infrastructure
 - ラック内、ラック間の計算リソースをネットワーク結合
 - 計算機を必要に応じてソフトウェア定義



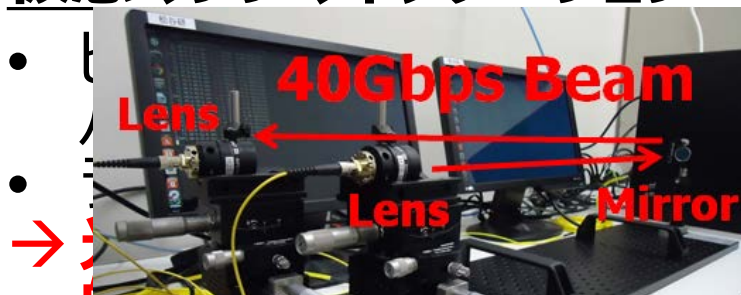
*Remote GPUs
over 10GbE*



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

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



Enhancement of network

FPGA NICによるNOSQL処理

Tight integration of I/O & computation

→ HW/SW

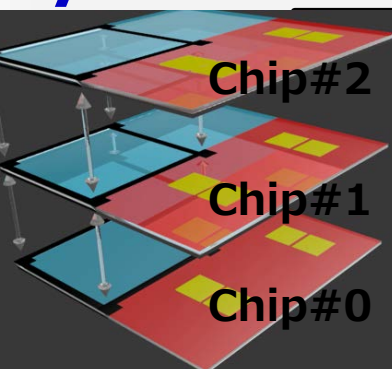
10GbE 機

- 莫大な
れ値以外



計算インテンシブ

I/Oインテンシブ

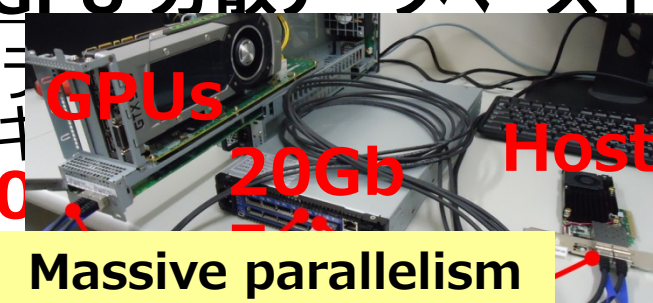


GPU 向け 3D 積層システム

プロセッサ、アクセラレータチップを3次元積層
チップ間通信はワイヤレス
KVSMメモリチップを開発

In-GPU 分散データベース

- GPU
 - ドキュメント
- 100 Gbps
- Massive parallelism



参考文献 (1/4)

- Key-Value Store アクセラレータ
 - Yuta Tokusashi, et.al., "A Multilevel NOSQL Cache Design Combining In-NIC and In-Kernel Caches", Hot Interconnects 2016.
 - Yuta Tokusashi, et.al., "NOSQL Hardware Appliance with Multiple Data Structures", Hot Chips 2016 (Poster).
 - Korechika Tamura, et.al., "An In-Kernel NOSQL Cache for Range Queries Using FPGA NIC", FPGA4GPC 2016.
 - Akihiko Hamada, et.al., "Design and Implementation of Hardware Cache Mechanism and NIC for Column-Oriented Databases ", ReConFig 2016.

参考文献 (2/4)

- 機械学習アクセラレータ
 - Ami Hayashi, et.al., "A Line Rate Outlier Filtering FPGA NIC using 10GbE Interface", ACM SIGARCH CAN (2015).
 - Ami Hayashi, et.al., "An FPGA-Based In-NIC Cache Approach for Lazy Learning Outlier Filtering", PDP 2017.
- 分散処理フレームワーク Spark 高速化
 - Yasuhiro Ohno, et.al., "Accelerating Spark RDD Operations with Local and Remote GPU Devices", ICPADS 2016.
 - Kohei Nakamura, et.al., "An FPGA-Based Low-Latency Network Processing for Spark Streaming", IEEE BigData Workshops 2016.

参考文献 (3/4)

- GPU による NOSQL データベース高速化
 - Shin Morishima, et.al., "Distributed In-GPU Data Cache for Document-Oriented Data Store via PCIe over 10Gbit Ethernet", HeteroPar 2016.
 - Shin Morishima, et.al., "Performance Evaluations of Document-Oriented Databases using GPU and Cache Structure", ISPA 2015.
 - Shin Morishima, et.al., "Performance Evaluations of Graph Database using CUDA and OpenMP-Compatible Libraries", ACM SIGARCH CAN (2014).

参考文献 (4/4)

- 誘導結合を用いたワイヤレス 3 次元積層
 - Takahiro Kagami, et.al., "Efficient 3-D Bus Architectures for Inductive-Coupling ThruChip Interfaces", IEEE TVLSI (2016).
 - Hiroki Matsutani, et.al, "Low-Latency Wireless 3D NoCs via Randomized Shortcut Chips", DATE 2014.
 - Yasuhiro Take, et.al., "3D NoC with Inductive-Coupling Links for Building-Block SiPs", IEEE TC (2014).
 - Hiroki Matsutani, et.al., "A Case for Wireless 3D NoCs for CMPs", ASP-DAC 2013. (Best Paper Award)



***Thank you for
listening!***

Acknowledgement:

**A part of this work is supported by JST PRESTO,
MIC SCOPE, NEDO IoT, and JSPS Kakenhi**