Keio University

Accelerator Design for Various NOSQL Databases

Hiroki Matsutani

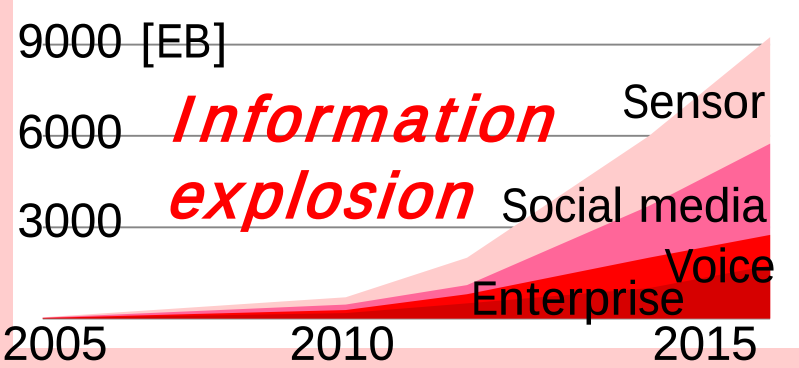
Dept. of ICS, Keio University

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

Two competing trends in ICT

Big data: the next oil

Data reuse & repurposing make innovations

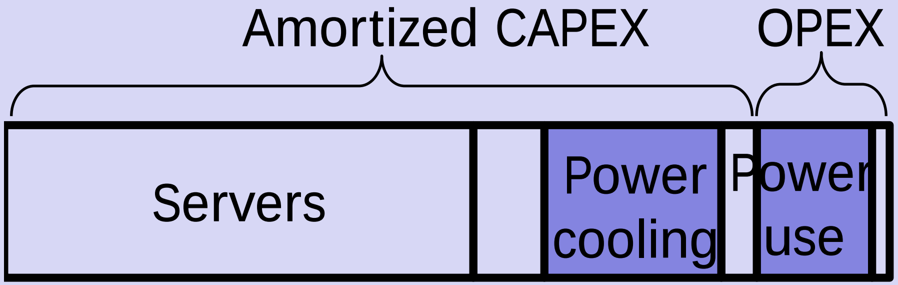


→Augmenting IT equipments

Green datacenters

Prevent global warming

Power & cooling are major sources of datacenter cost



→Promoting energy-savings

Observation: Without more energy-efficient solutions, augmenting more computers for Big data becomes harder

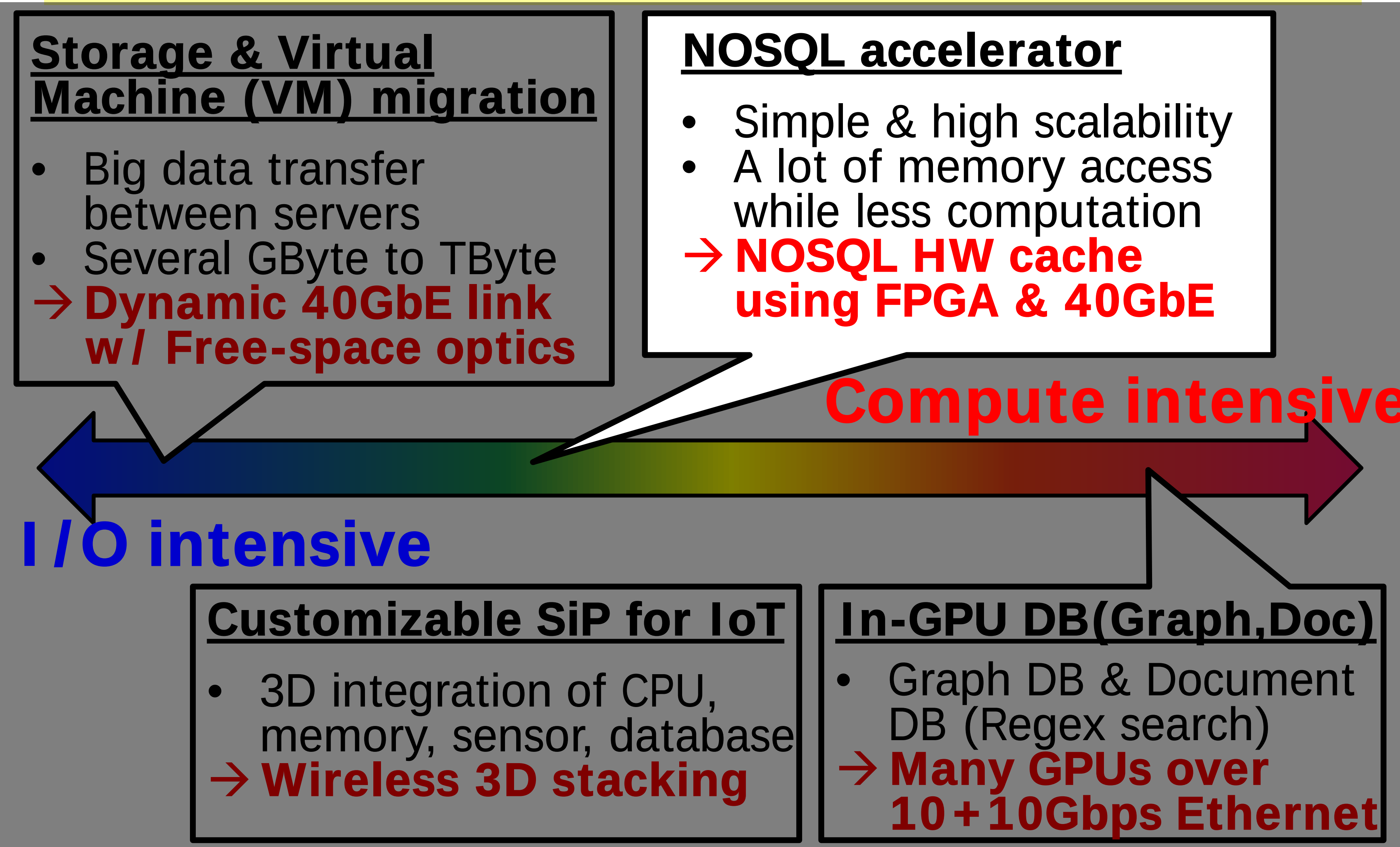
Limitations: Computers are already very efficient

Thousands of low-end commodity servers optimized for cost-performance and energy efficiency

We need Architectural Innovations (not rely on Moore's law)

Our introduction: Today's talk

The best solution changes depending on I/O intensity



Storage & Virtual Machine (VM) migration

- Big data transfer between servers
- Several GByte to TByte

→ **Dynamic 40GbE link w/ Free-space optics**

NOSQL accelerator

- Simple & high scalability
- A lot of memory access while less computation

→ **NOSQL HW cache using FPGA & 40GbE**

Compute intensive

I/O intensive

Customizable SiP for IoT

- 3D integration of CPU, memory, sensor, database

→ **Wireless 3D stacking**

In-GPU DB(Graph,Doc)

- Graph DB & Document DB (Regex search)

→ **Many GPUs over 10+10Gbps Ethernet**

Structured storages (NOSQLs)

Structured storages (NOSQLs) have good horizontal scalability, while they are specialized for some specific purposes

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

HBase, BigTable

Column-oriented

MongoDB

Document-oriented

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

Schema-less DB

Memcached

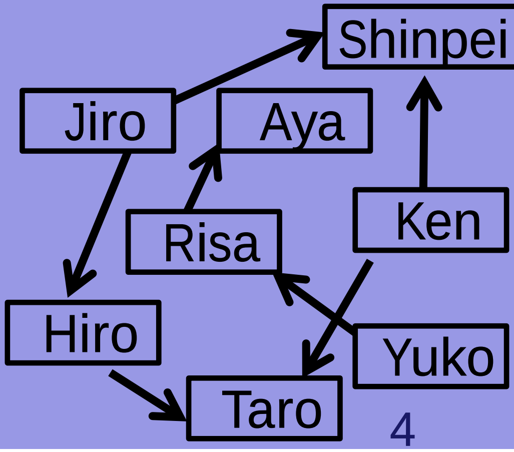
Key	Value
↓	

Key value store

Shopping cart, User profile, Session, etc

Neo4j

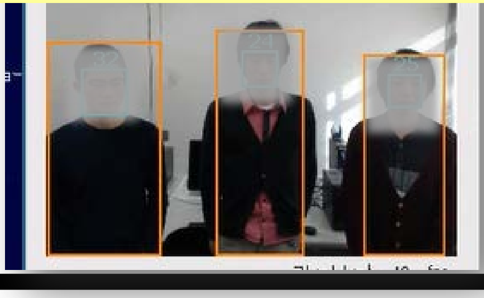
Graph DB

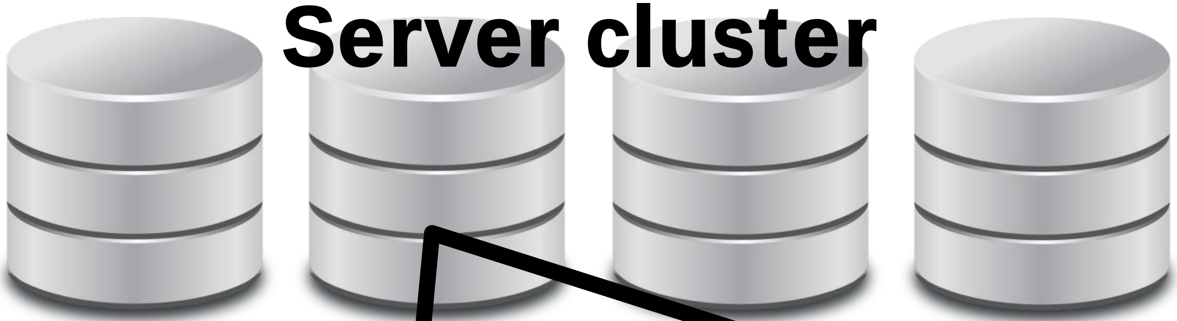


Customer social graph

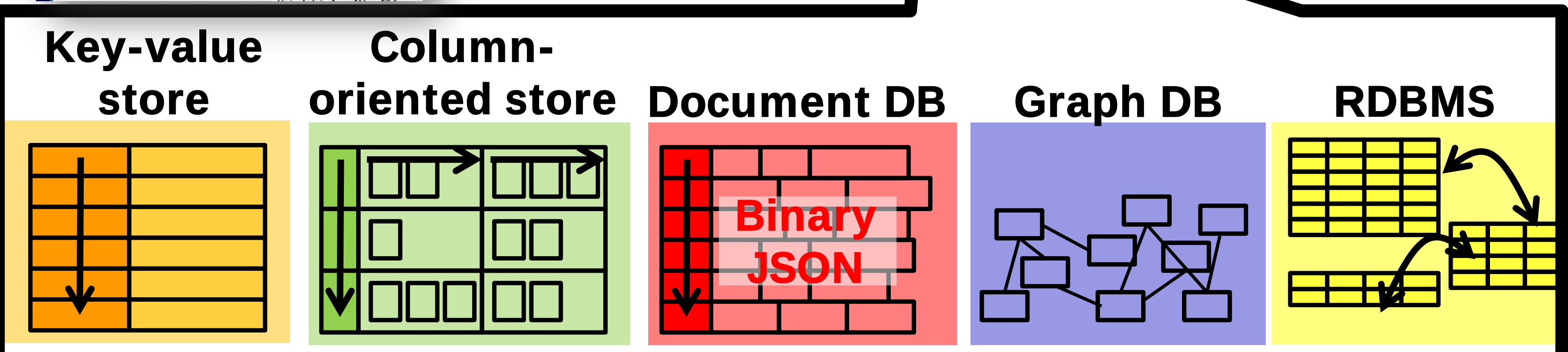
Polyglot Persistence: Mixture of NOSQLs

Real-time gender & age detection





Server cluster



Key-value store

Column-oriented store

Document DB

Graph DB

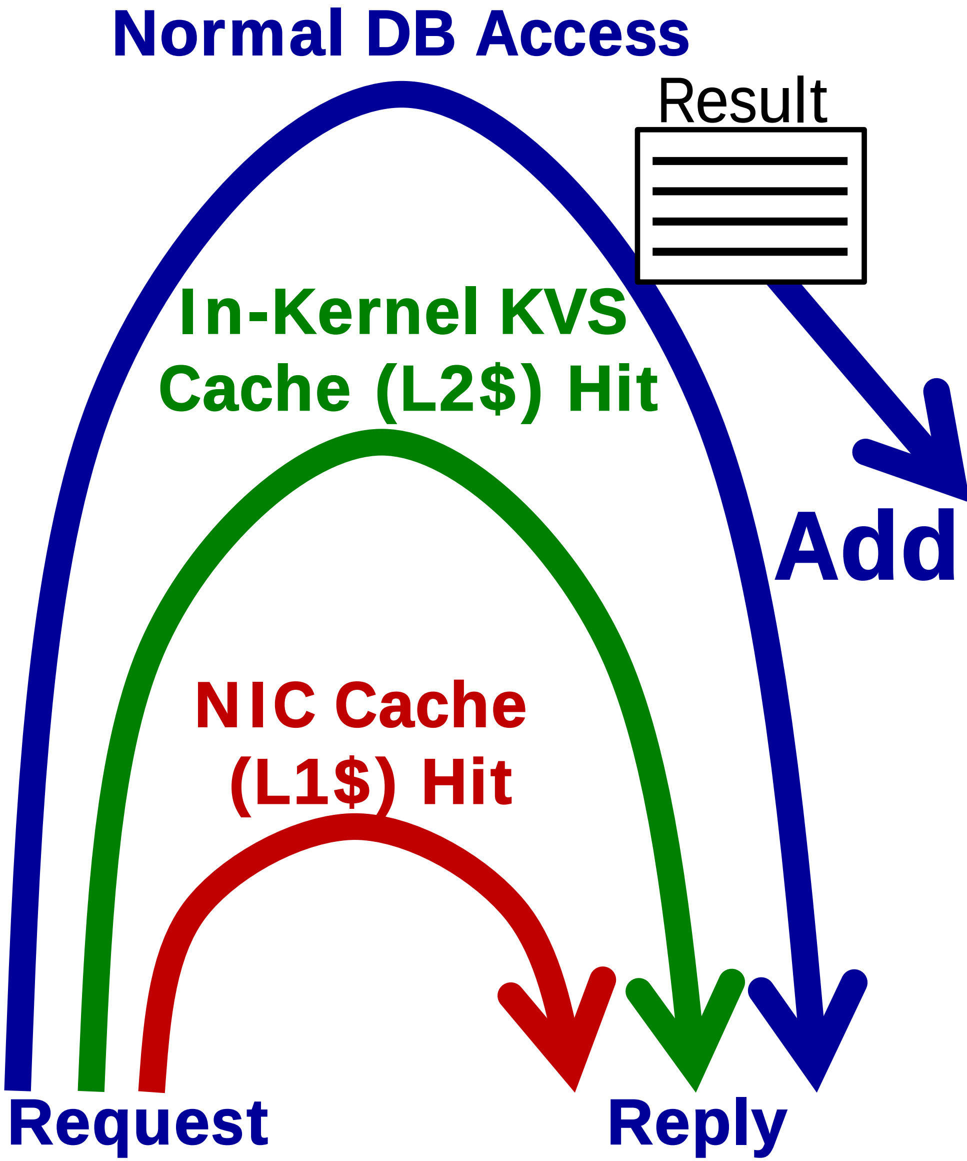
RDBMS

Data analysis framework

hadoop Spark Streaming

Our Target: Mixture of structured storages to take advantage of the fact that different structures are suitable for tackling different problems

Multilevel NOSQL cache: FPGA NIC



User Space

mongoDB

APACHE HBASE

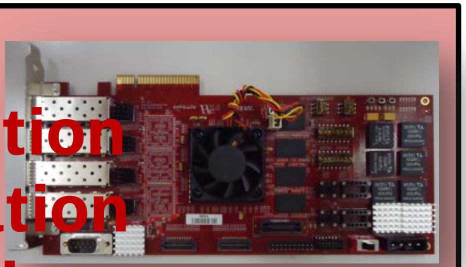
Neo4j

NetFPGA10G Device Driver

In-Kernel KVS
(Key-Value Store)
Cache 32GB - 128GB

NetFPGA 10G

Hash Computation
Flow Classification
Machine Learning



DRAM 288MB

10GbE x4

Multilevel NOSQL cache: FPGA NIC

Multilevel NOSQL cache:

FPGA-based hardware cache as L1 NOSQL cache

In-kernel software cache as L2 NOSQL cache

Good balance between speed and capacity:

L1 NOSQL cache ... Very fast but small

L2 NOSQL cache ... Fast and large

Design space explanation → [IEEE HoTI'16]

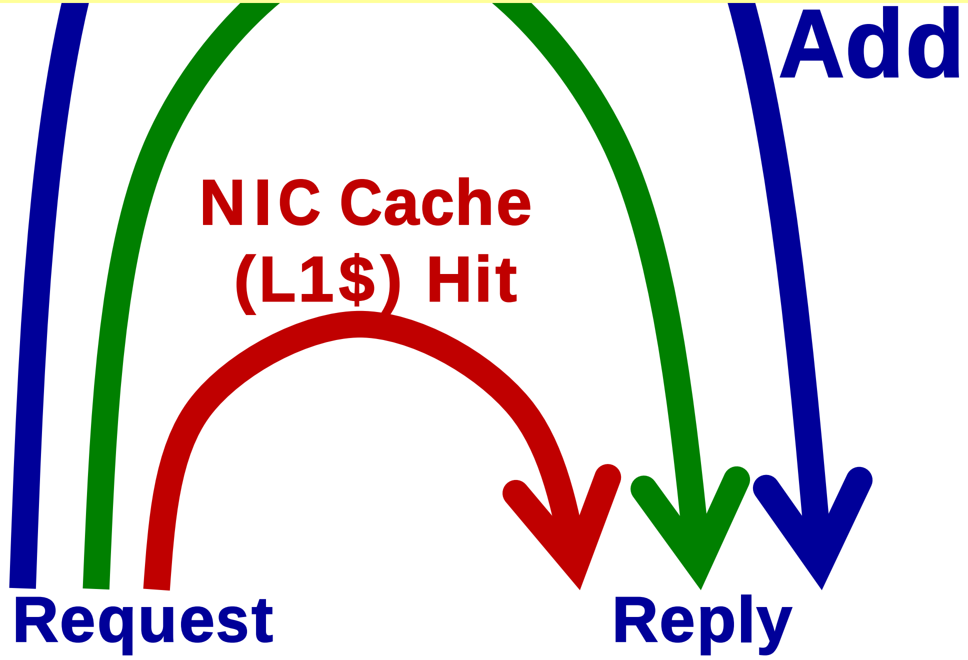
oDB

eo4j

iver

(Key-Value Store)

Cache 32GB - 128GB

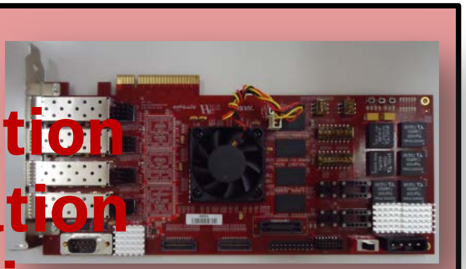


NetFPGA 10G

Hash Computation

Flow Classification

Machine Learning



DRAM 288MB

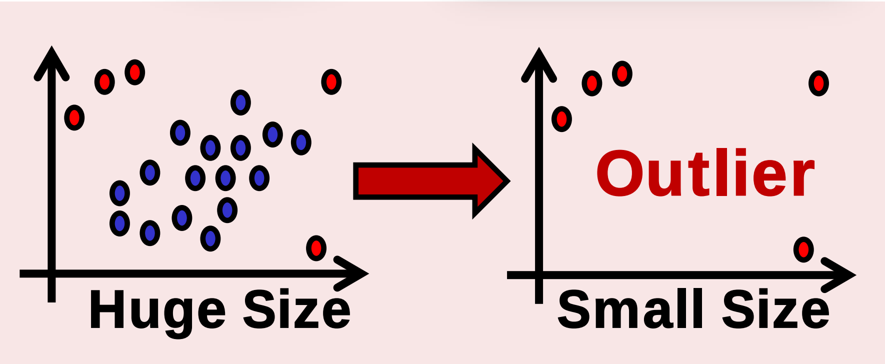
10GbE x4

10GbE outlier filtering FPGA NIC

Sensor Data Explosion

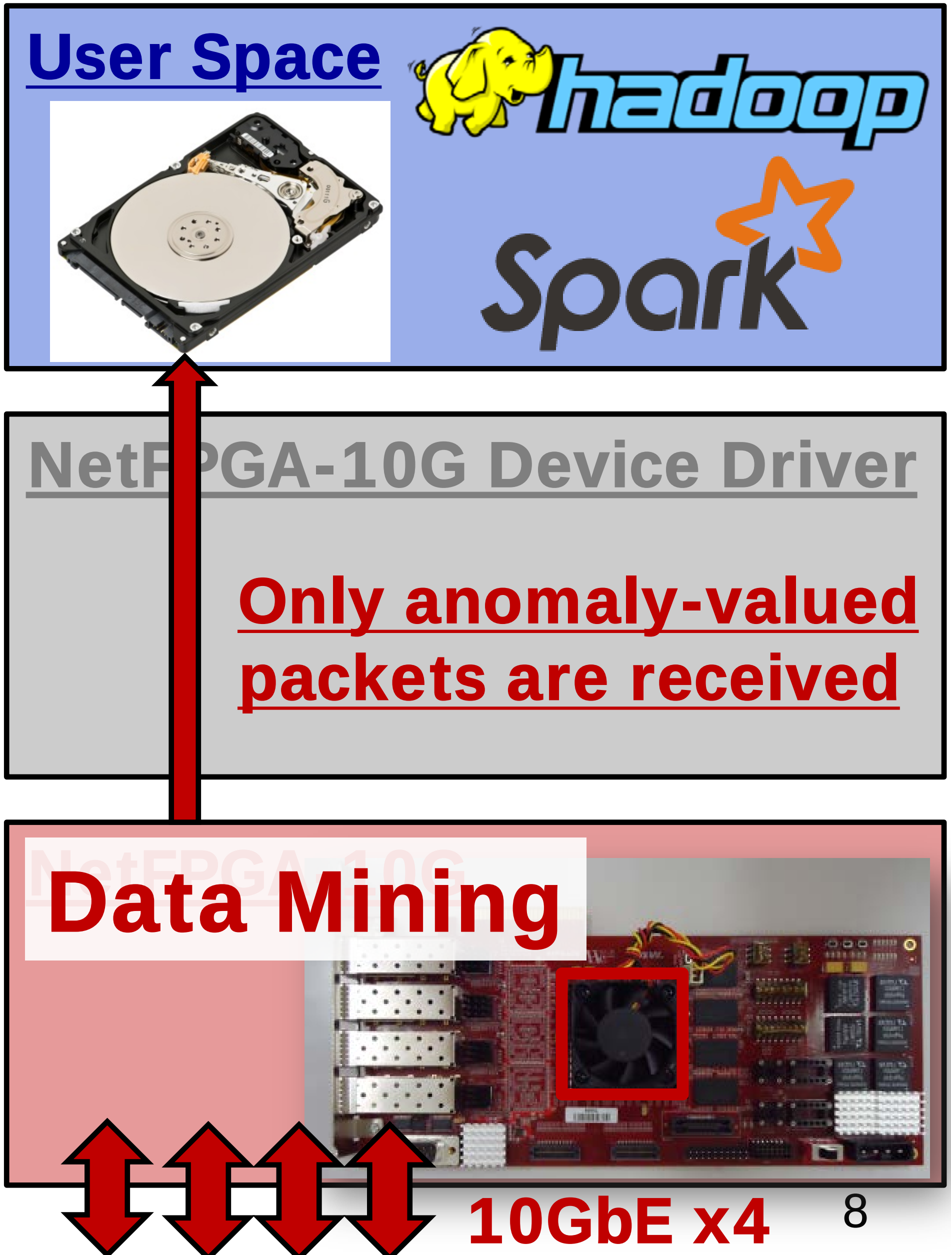
RESCAT端末管理

画像サイズ 640×480



Machine learning algorithms

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



User Space

NetFPGA-10G Device Driver

Only anomaly-valued packets are received

NetFPGA-10G

Data Mining

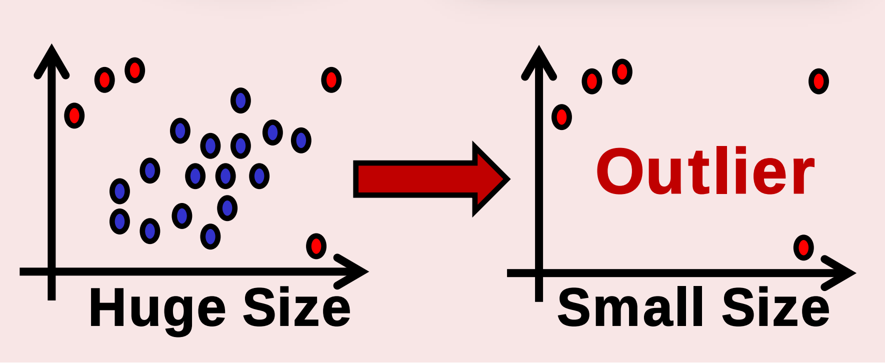
10GbE x4

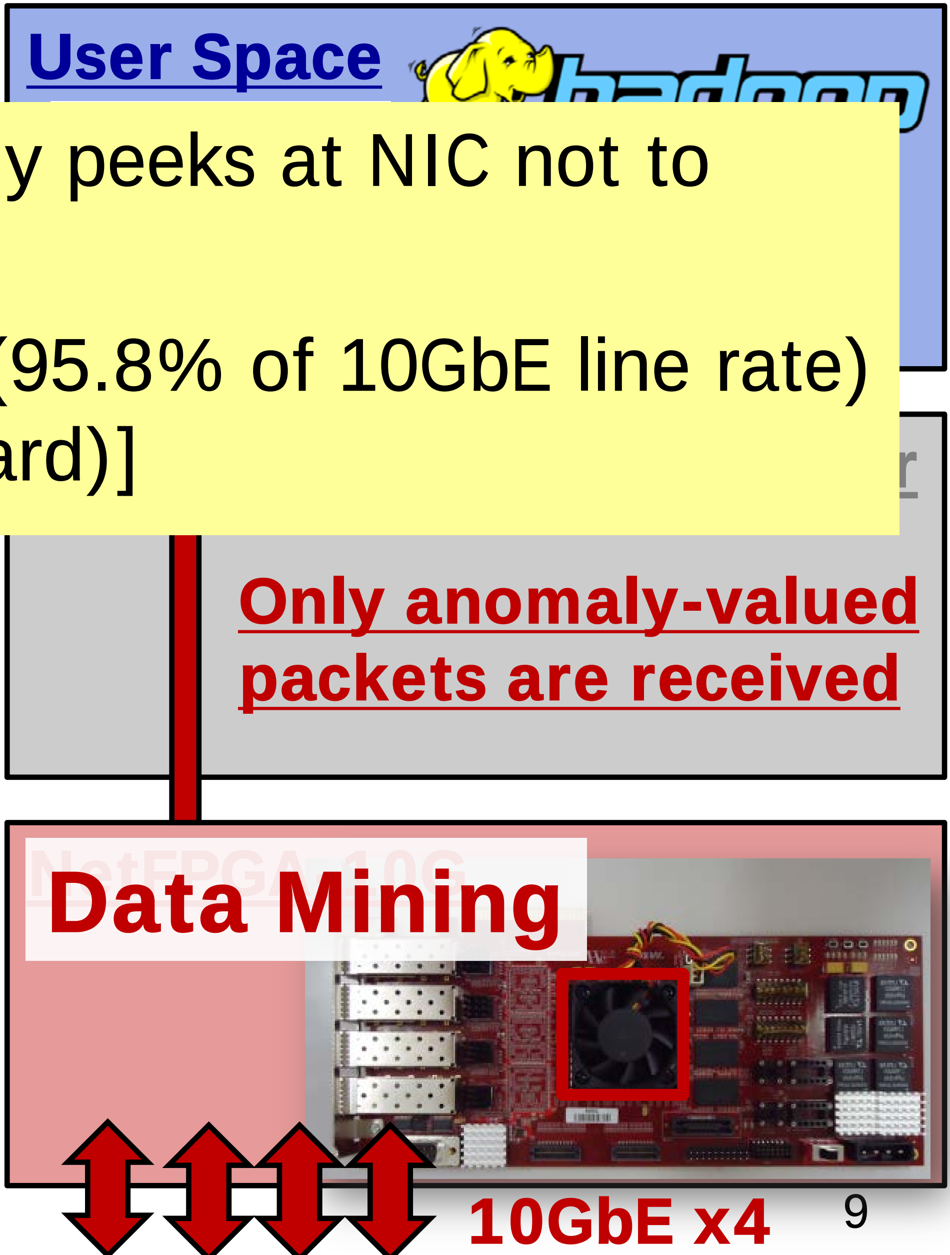
10GbE outlier filtering FPGA NIC

Sensor Data Explosion

Issue: Software periodically peeks at NIC not to forget what is “normal”

Result: 14M samples/sec (95.8% of 10GbE line rate) [HEART’15 (Best paper award)]



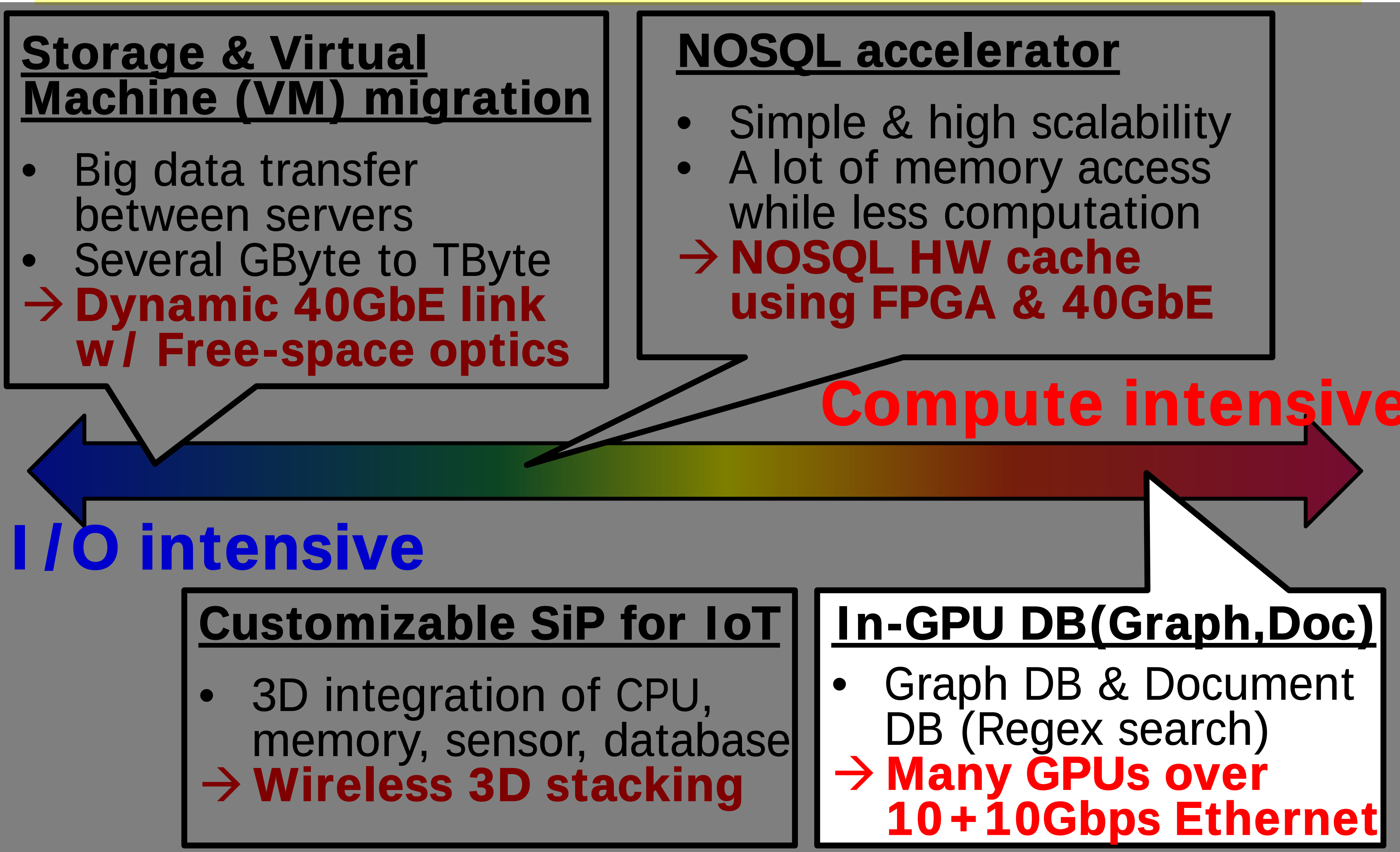


Machine learning algorithms

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

Our introduction: Today’s talk

The best solution changes depending on I/O intensity



Storage & Virtual Machine (VM) migration

- Big data transfer between servers
- Several GByte to TByte

→ **Dynamic 40GbE link w/ Free-space optics**

NOSQL accelerator

- Simple & high scalability
- A lot of memory access while less computation

→ **NOSQL HW cache using FPGA & 40GbE**

Compute intensive

I/O intensive

Customizable SiP for IoT

- 3D integration of CPU, memory, sensor, database

→ **Wireless 3D stacking**

In-GPU DB(Graph,Doc)

- Graph DB & Document DB (Regex search)

→ **Many GPUs over 10+10Gbps Ethernet**

NOSQL cache with GPUs

Compute intensive tasks are offloaded to GPUs

User Space

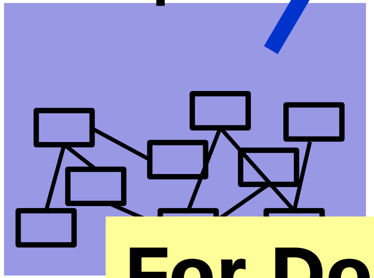
SLOW

NetFPGA10G Device Driver

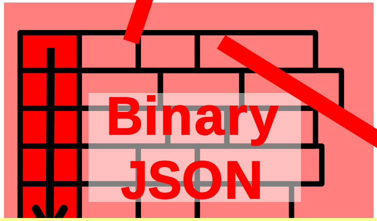
In-Kernel KVS

(Key-Valu FAST e)

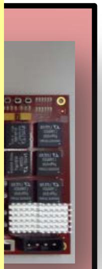
Graph DB



Document DB



Ke



MB

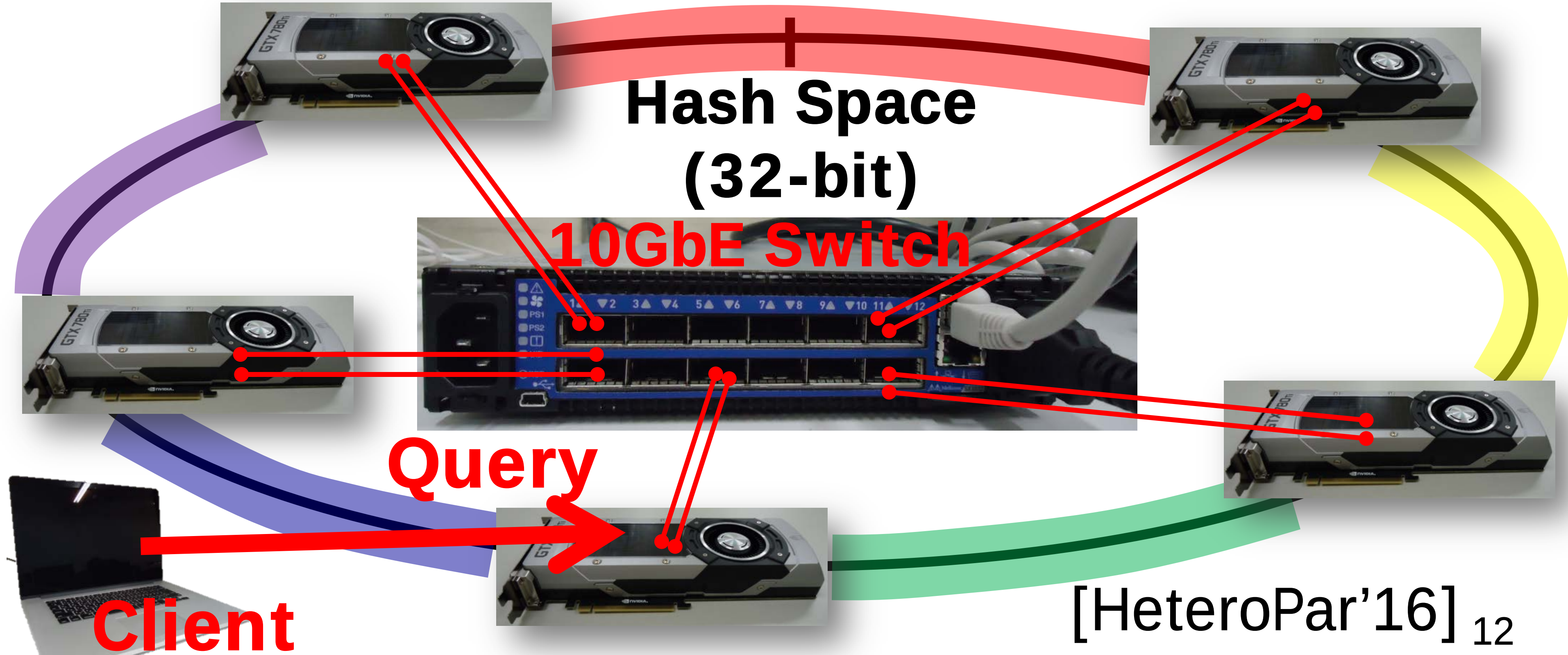
For Document DBs (e.g., MongoDB):
Regular expression based text search is offloaded to GPUs [ISPA’15]

For Graph DBs (e.g., Neo4j):
Graph search (e.g., SSSP) is offloaded to GPUs [ACM Comp Arch News (2014)]

In-GPU distributed DB w/ ExpEther

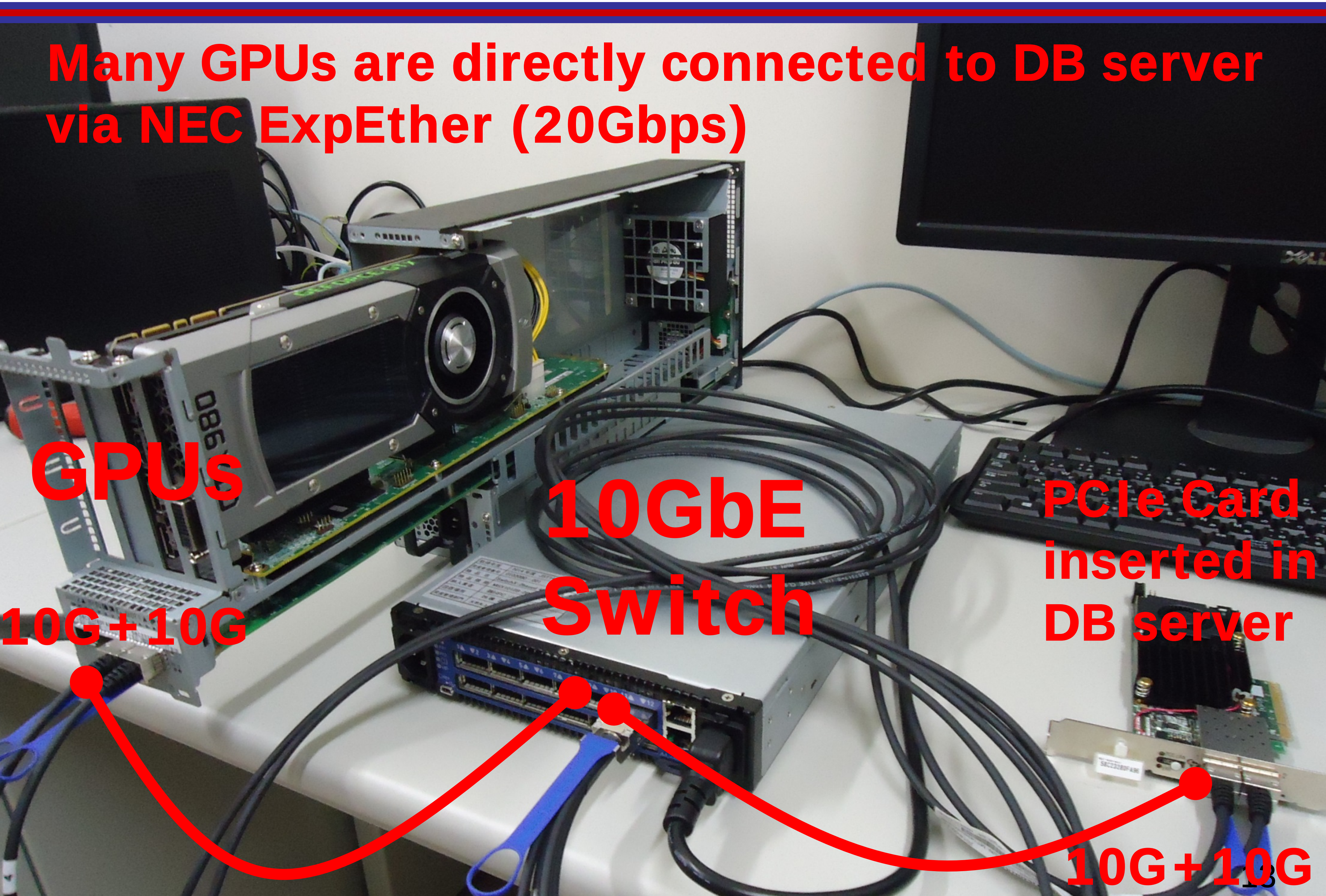
To exploit more GPUs → In-GPU databases

- In-GPU distributed DBs over NEC ExpEther
  - GPU's device memory is used as a **cache** of the DB
  - Many GPUs are directly connected via 10GbE switch



[HeteroPar'16]

In-GPU distributed DB w/ ExpEther



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

GPUs

10G+10G

10GbE Switch

PCIe Card inserted in DB server

10G+10G

Our introduction: Today’s talk

The best solution changes depending on I/O intensity

Storage & Virtual Machine (VM) migration

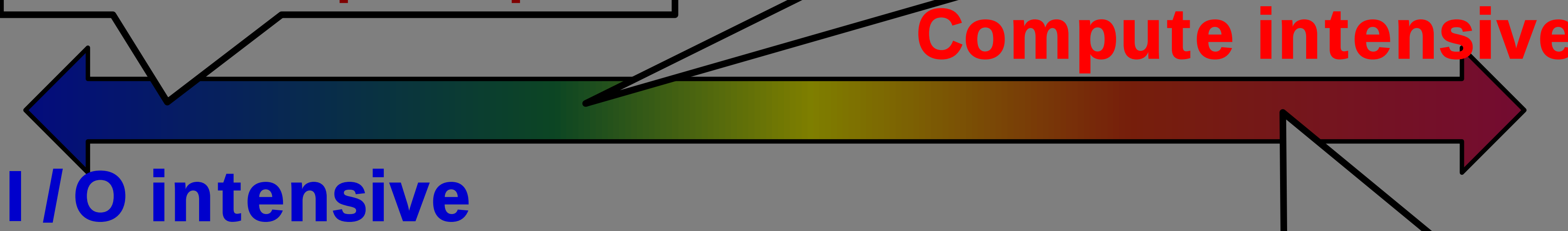
- Big data transfer between servers
- Several GByte to TByte

→ **Dynamic 40GbE link w/ Free-space optics**

NOSQL accelerator

- Simple & high scalability
- A lot of memory access while less computation

→ **NOSQL HW cache using FPGA & 40GbE**



Customizable SiP for IoT

- 3D integration of CPU, memory, sensor, database

→ **Wireless 3D stacking**

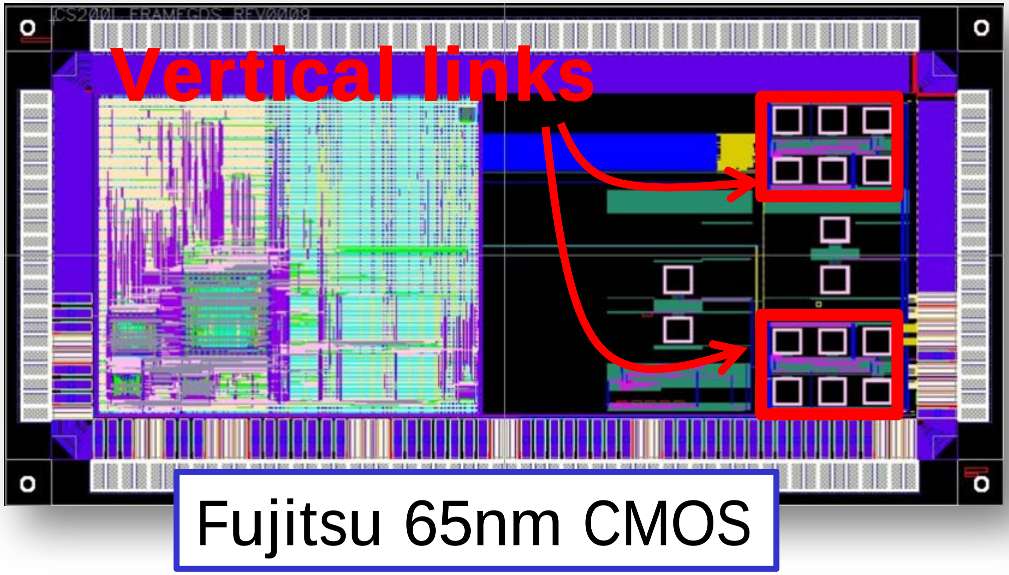
In-GPU DB(Graph,Doc)

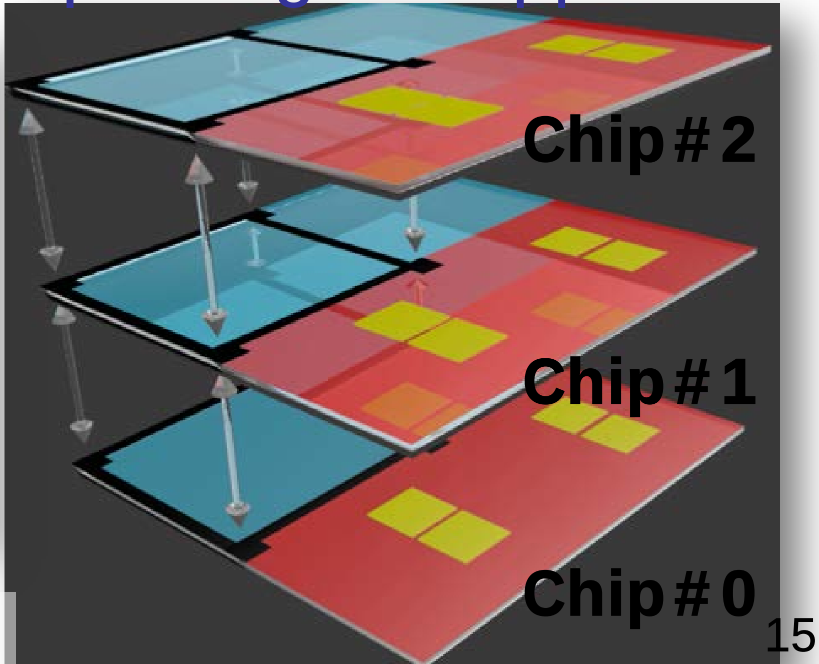
- Graph DB & Document DB (Regex search)

→ **Many GPUs over 10+10Gbps Ethernet**

Wireless 3D chip stacking for IoT

- System-in-Package (SiP) for sensor nodes
  - – Required chips are selected and stacked in package
  - – E.g., CPU chip, Memory chip, Sensor chip, ...
- Wireless inductive-coupling for vertical links
  - – Not electrically-connected
  - – Add, remove, and swap chips for given applications

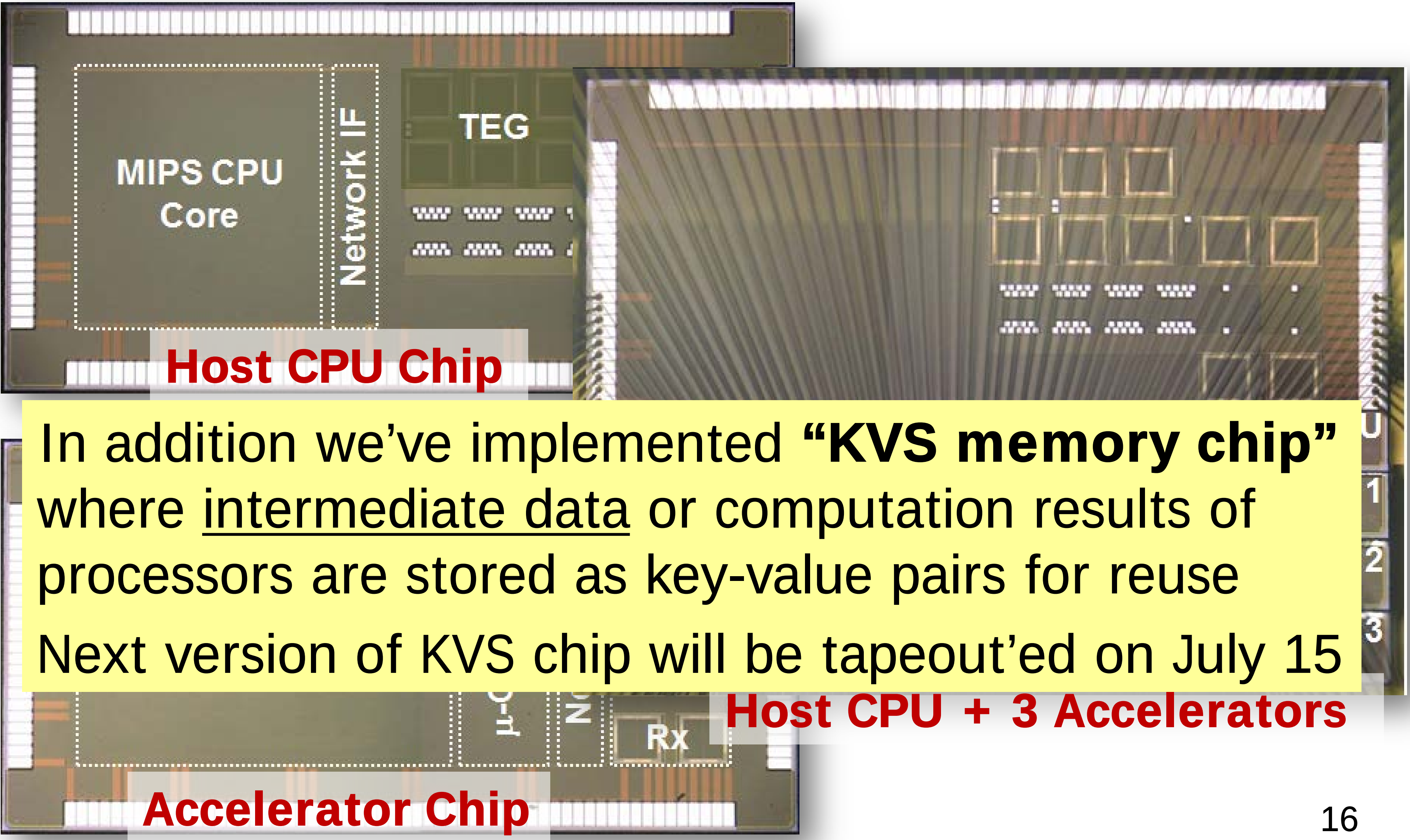




Joint research with Prof. Kuroda @ Keio U

Wireless 3D chip stacking for IoT

We can change the number & types of chips in a package



In addition we’ve implemented **“KVS memory chip”** where intermediate data or computation results of processors are stored as key-value pairs for reuse

Next version of KVS chip will be tapeout’ed on July 15

Our introduction: Today’s talk

The best solution changes depending on I/O intensity

Storage & Virtual Machine (VM) migration

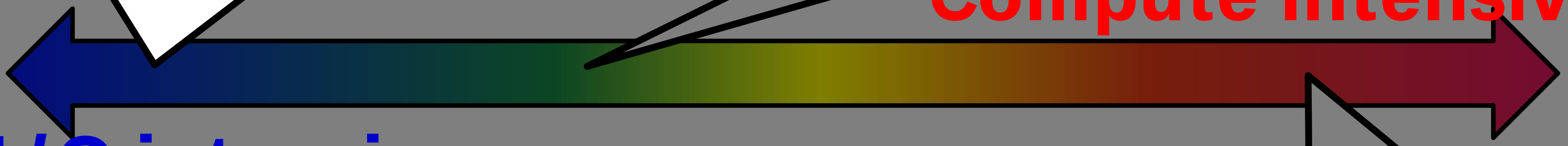
- Big data transfer between servers
- Several GByte to TByte

→ **Dynamic 40GbE link w/ Free-space optics**

NOSQL accelerator

- Simple & high scalability
- A lot of memory access while less computation

→ **NOSQL HW cache using FPGA & 40GbE**



Compute intensive

I/O intensive

Customizable SiP for IoT

- 3D integration of CPU, memory, sensor, database

→ **Wireless 3D stacking**

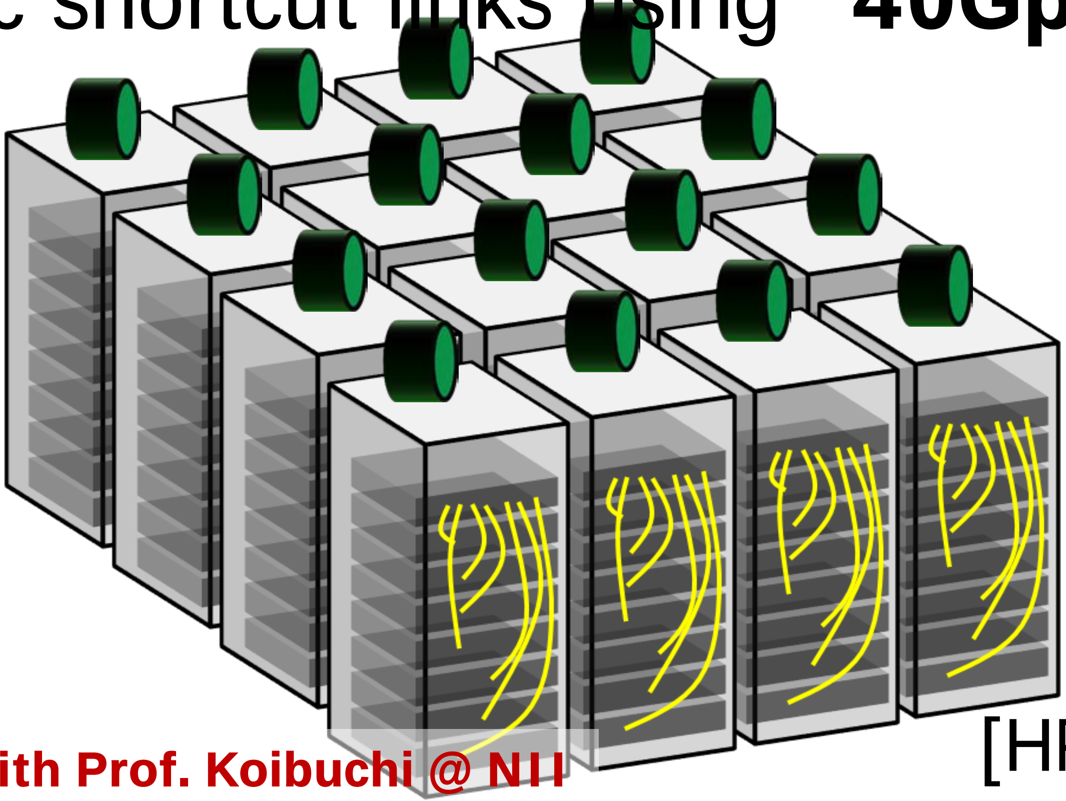
In-GPU DB(Graph,Doc)

- Graph DB & Document DB (Regex search)

→ **Many GPUs over 10+10Gbps Ethernet**

Dynamic 40G shortcut links w/ FSO

- Emergent big data transfers in Datacenter NW
  - Virtual machine (VM) migration
  - Storage migration and DB streaming
  - E.g., Several minutes for VM migration w/ 1GbE
- Dynamic shortcut links using **“40Gps beam”**

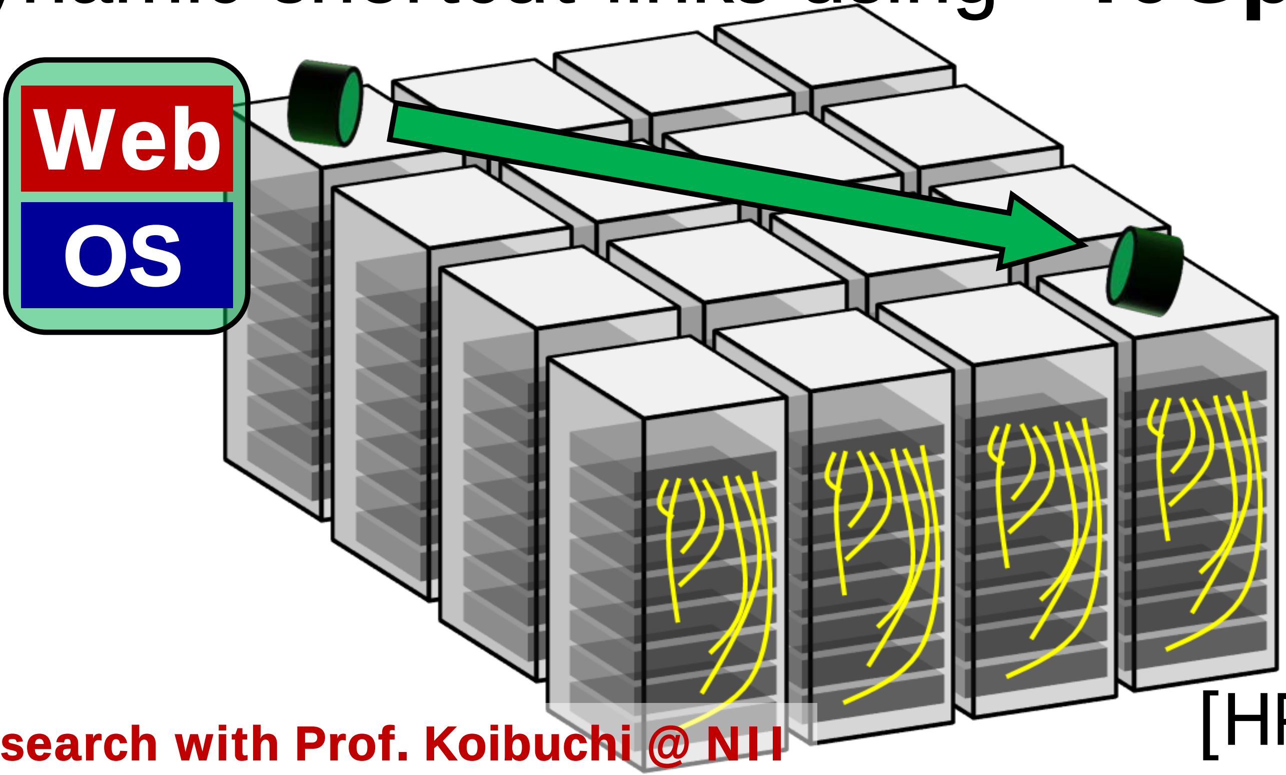


Joint research with Prof. Koibuchi @ NII

[HPCA’15]

Dynamic 40G shortcut links w/ FSO

- Emergent big data transfers in Datacenter NW
  - Virtual machine (VM) migration
  - Storage migration and DB streaming
  - E.g., Several minutes for VM migration w/ 1GbE
- Dynamic shortcut links using **“40Gps beam”**



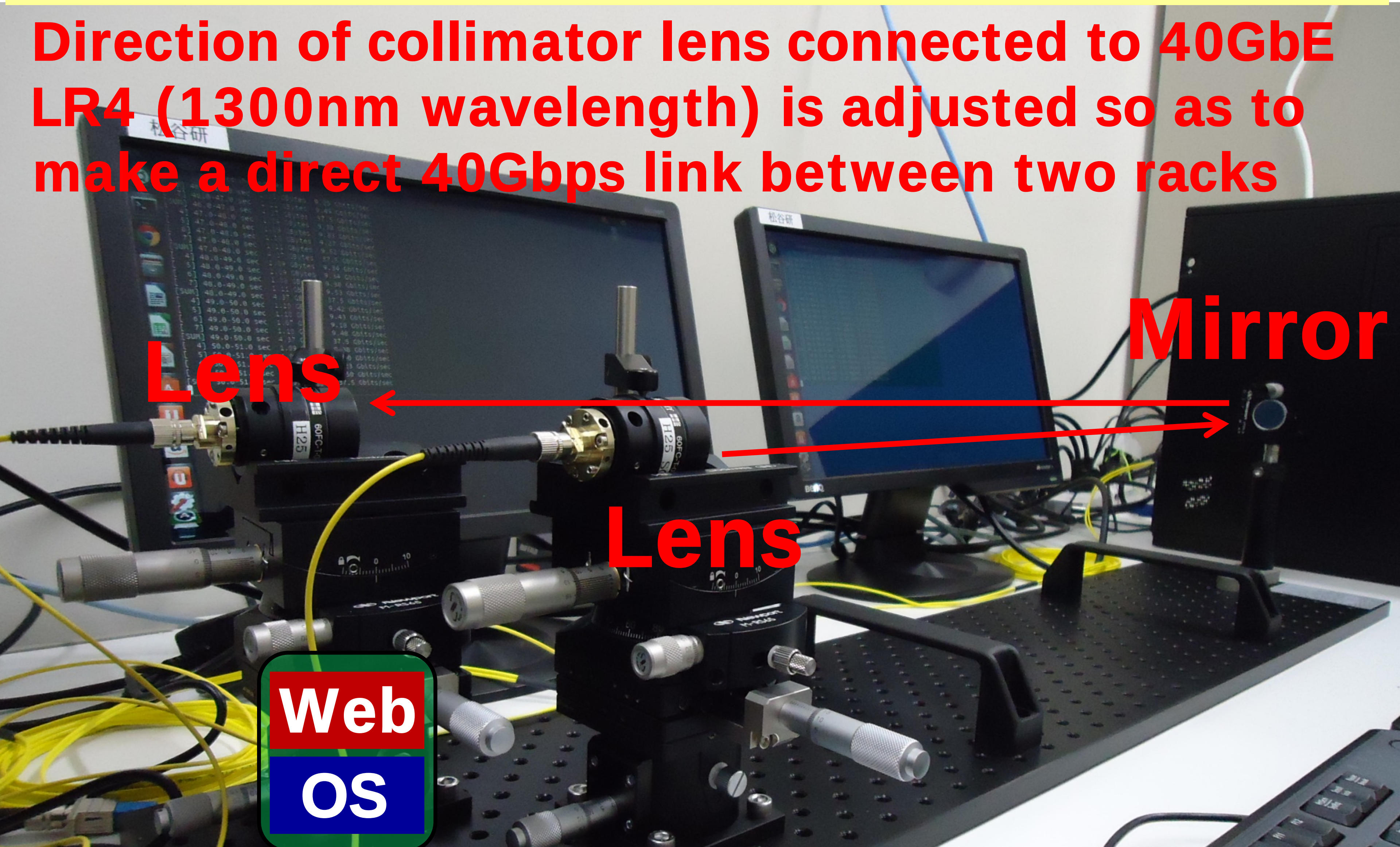
[HPCA’15]

Joint research with Prof. Koibuchi @ NII

“VM Highway” using 40G FSO

Dynamic 40GbE links for VM (virtual machine) migration

Direction of collimator lens connected to 40GbE LR4 (1300nm wavelength) is adjusted so as to make a direct 40Gbps link between two racks



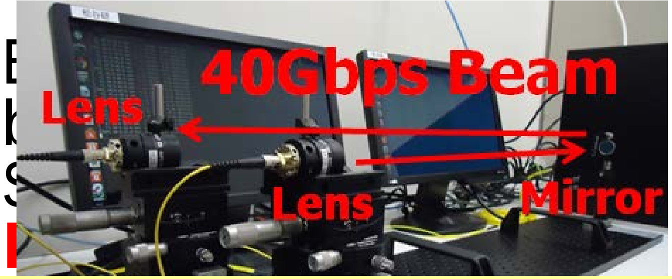
Our introduction: Today's talk

The best solution changes depending on I/O intensity

Storage & Virtual Machine (VM) migration

- E
  b
- S

→ I



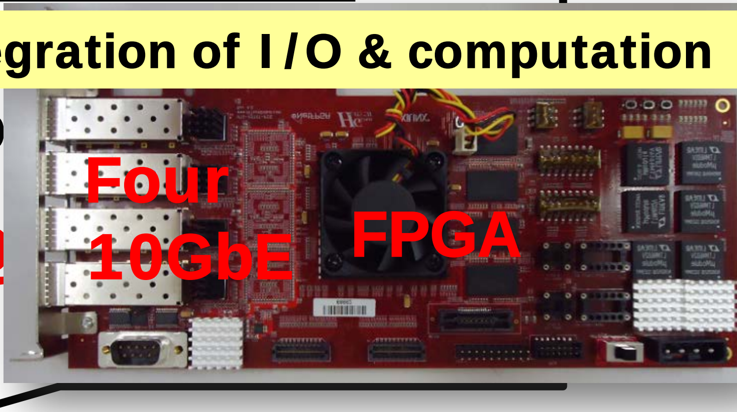
Enhancement of network

NOSQL accelerator

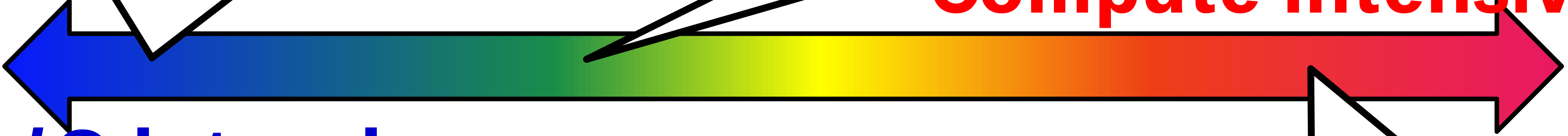
Tight integration of I/O & computation

- A lot o
  while I

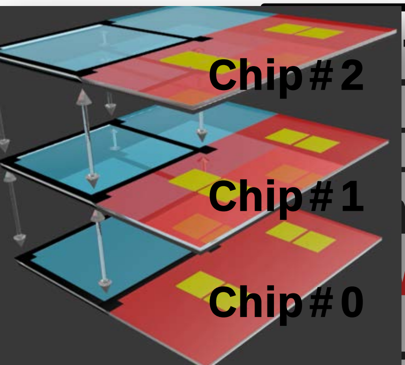
→ **NOSQ
using**



Compute intensive



I/O intensive



tomizable SiP for IoT

D integration of CPU,
emory, sensor, database

/ireless 3D stacking

In-GPU DB(Graph,Doc)

- Gra
  DB

→ **Ma
1**

Massive parallelism

References (1/3)

- Key-value store accelerators
  - 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.
- Machine learning accelerator
  - Ami Hayashi, et.al., "A Line Rate Outlier Filtering FPGA NIC using 10GbE Interface", ACM SIGARCH CAN (2015).

References (2/3)

- GPU-based accelerations of NOSQLs
  - 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).
- Free-space optics (FSO) for data centers
  - Ikki Fujiwara, et.al., "Augmenting Low-latency HPC Network with Free-space Optical Links", HPCA 2015.

References (3/3)

- Wireless inductive-coupling 3D stacking
  - 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