Expectations for optical network from the viewpoint of system software research

Ryousei Takano

National Institute of Advanced Industrial Science and Technology (AIST)

Special session on challenges and opportunities of integrated photonics in future datacenters

ACSI 2015@Tsukuba, 27 Jan. 2015

Outline

- Trends in datacenter research and development
- AIST IMPULSE Project
- Workload analysis
- Proposed architecture: Dataflow-centric computing

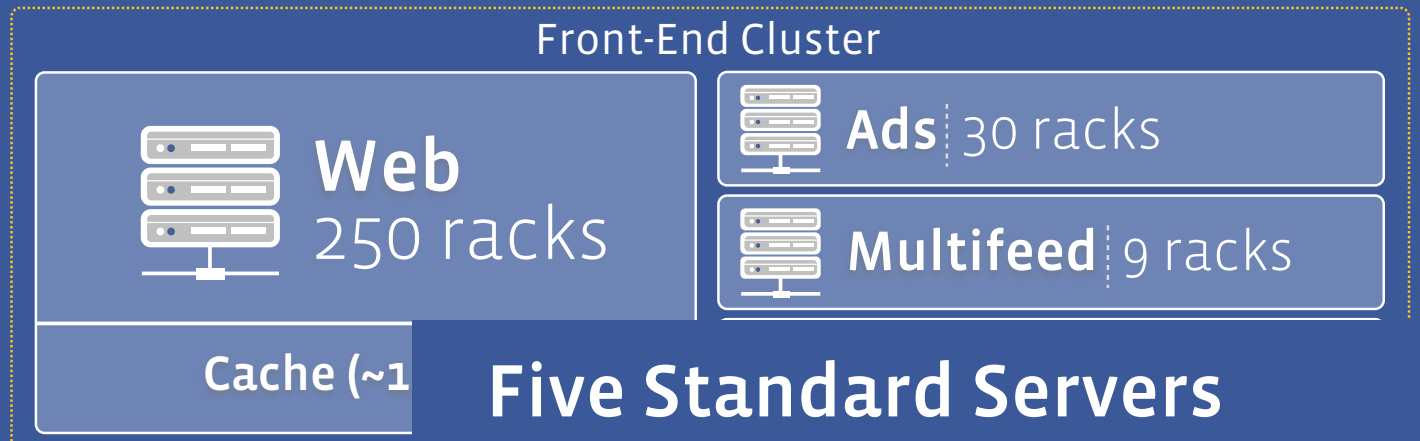
Introduction

- BigData is a killer app in datacenters, it requires a clean slate architecture like **“disaggregation”**, **“datacenter in a box”**.
- **Optical network is key** to making them.
- Optical path network (all optical path between end-to-end) in a datacenter
  - Pros: huge bandwidth, energy efficiency
  - Cons: path switching latency, utilization
- To take advantage of optical path network, a new datacenter OS is essential.
  - Key idea: control/data plane separation

Optical Network in DCs

- Similar concept (***"disaggregation"*** or ***"datacenter in a box"***) projects are launched recently.
  - Open Compute Project (Facebook)
  - Rack Scale Computing (Intel)
  - Extremely Shrinking Computing (IBM)
  - The Machine (HP)
  - FireBox (UCB)
  - CTR Consortium (MIT)
- Optical network, including photonic-electronic convergence and short (<1km) reach inter-connection, is key to drive innovation in future datacenters.

Architecture



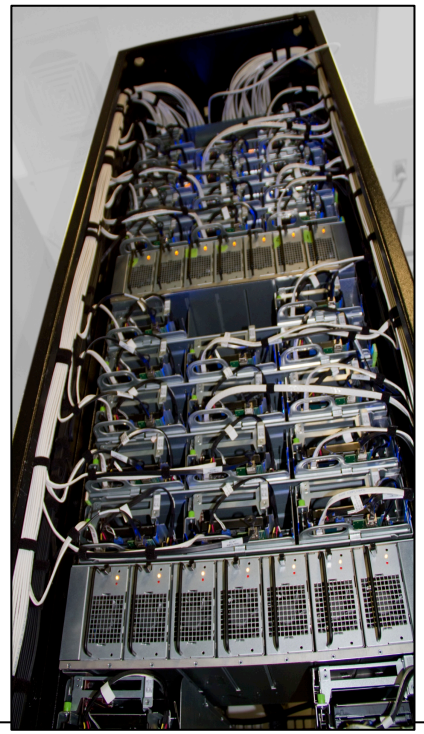
Five Standard Servers

Standard Systems	I Web	III Database	IV Hadoop	V Photos	VI Feed
CPU	High 2 x E5-2670	High 2 x E5-2660	High 2 x E5-2660	Low	High 2 x E5-2660
Memory	Low	High 144GB	Medium 64GB	Low	High 144GB
Disk	Low	High IOPS 3.2 TB Flash	High 15 x 4TB SATA	High 15 x 4TB SATA	Medium
Services	Web, Chat	Database	Hadoop (big data)	Photos, Video	Multifeed, Search, Ads

"Flash at Facebook", Flash Summit 2013

Open Compute Project

- OCP was founded to openly share designs of datacenter products by Facebook in April 2011.
- Shift from commodity products to **user-driven design** to improve the energy efficiency of large scale datacenters
  - Industry Standard: 1.9 PUE
  - Open Compute Project: 1.07 PUE
- Specifications: server, storage, rack, network switch, etc.
- Products: Quanta Rackgo X, GIGABYTE DataCenter Solution



Open Compute Rack v2

Evolution of Rack Scale Infrastructure

Today

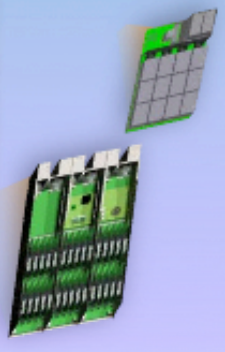
Infrastructure Disaggregation

- Shared Power
- Shared Cooling
- Rack Management

Next

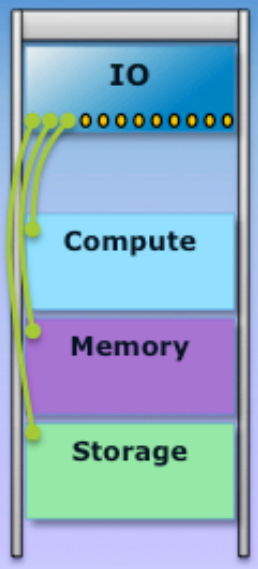
Fabric Integration

- Rack Fabric
- Optical Interconnects
- Modular refresh



Future

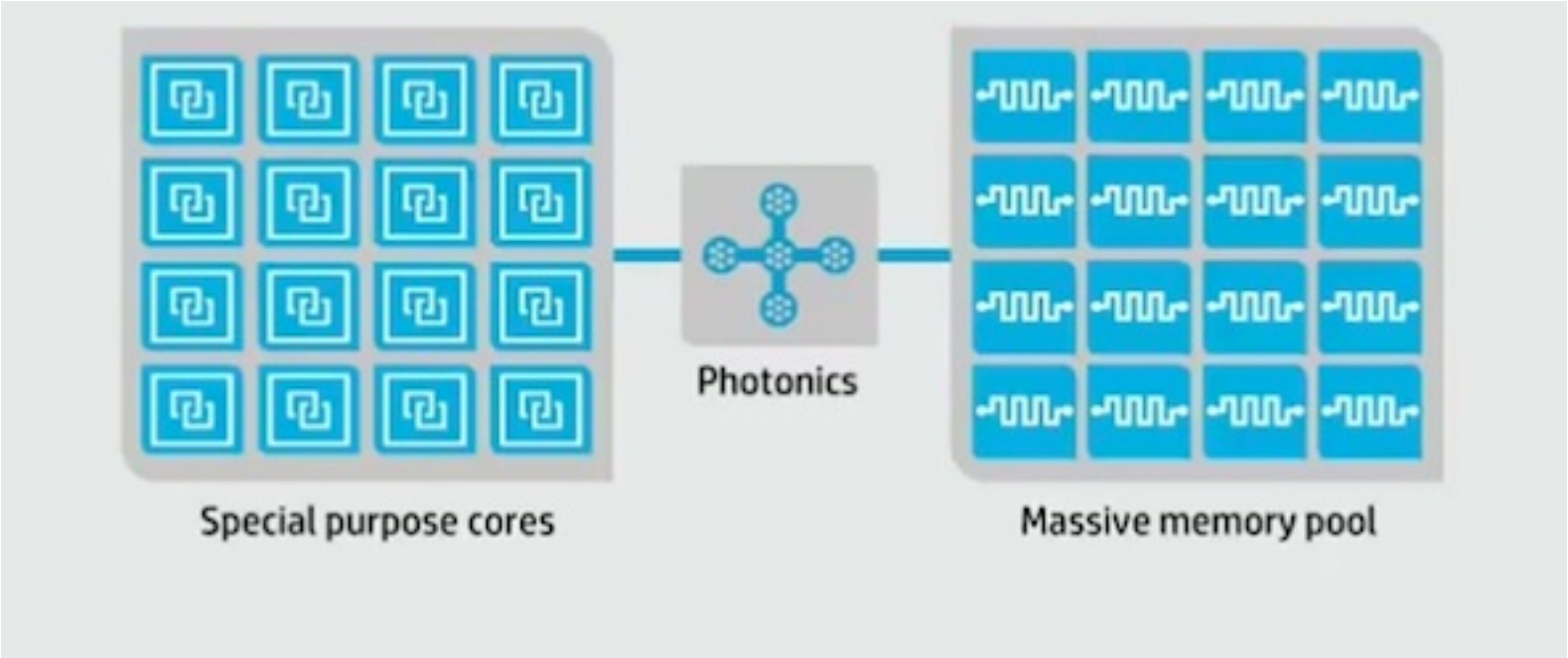
Fully Modular Resources



- Pooled compute
- Pooled Memory
- Pooled Storage

Enable flexible and efficient data centers through disaggregation of resources

HP “The Machine”

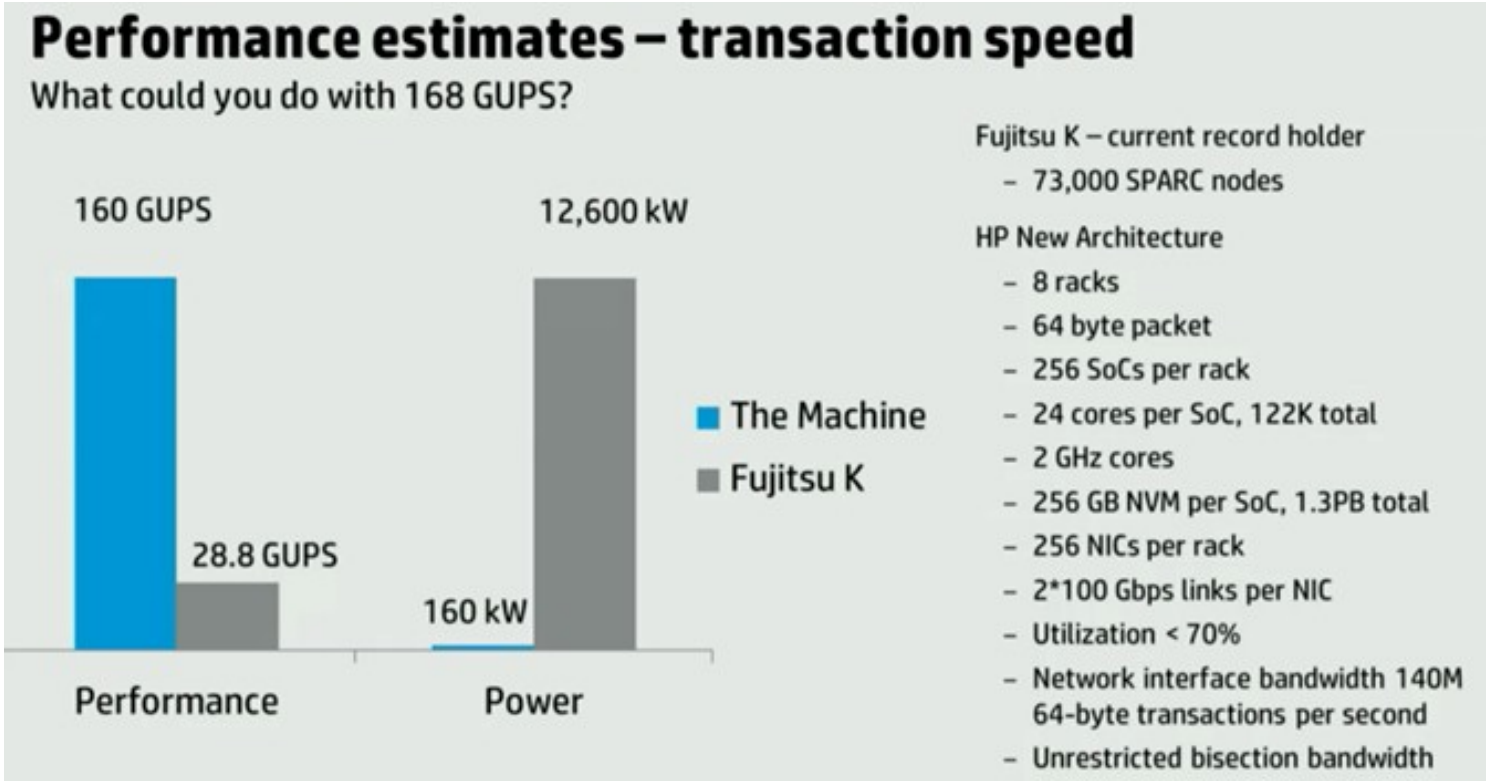


The Machine could be **six times more powerful** than an equivalent conventional design, while using just **1.25 percent of the energy** and being around **1/100 the size**.

http://www.hpl.hp.com/research/systems-research/themachine/

Datacenter in a Box

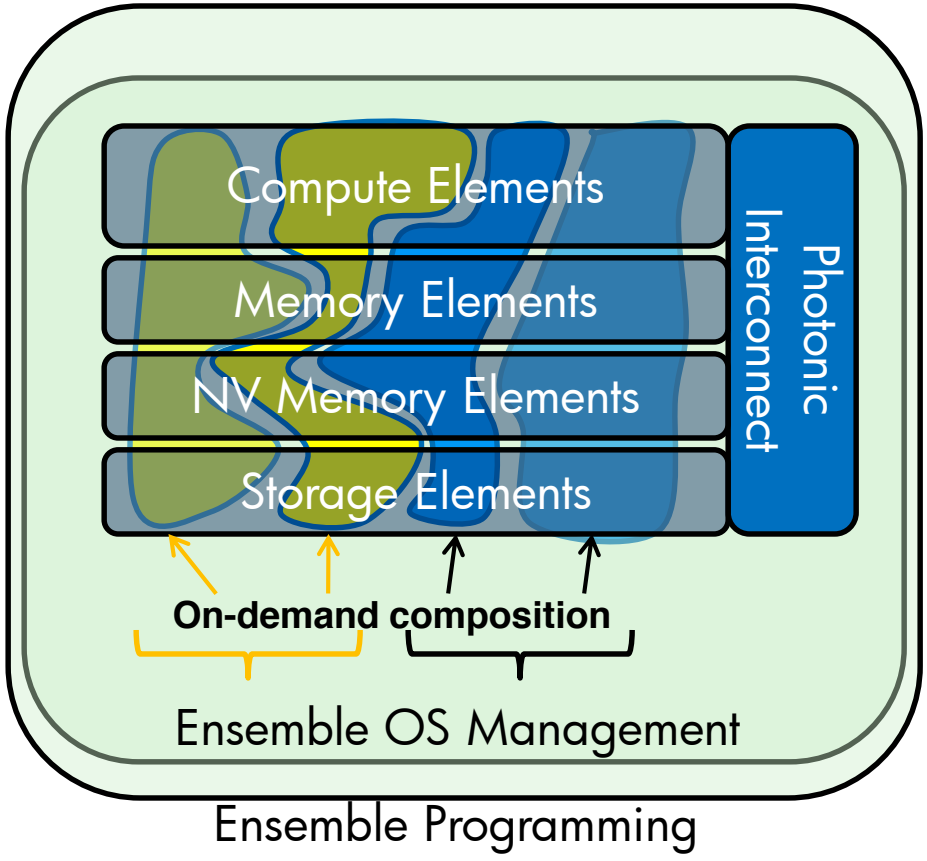
- The machine is six times faster with 1.25 percent of the energy comparing with K.



HPC Challenge's RandomAccess benchmark

The Machine: Architecture

Architecture evolution/revolution



"Computing Ensemble": bigger than a server, smaller than a datacenter, built-in system software

- Disaggregated pools of uncommitted compute, memory, and storage elements
- Optical interconnects enable dynamic, on-demand composition
- Ensemble OS software using virtualization for composition and management
- Management and programming virtual appliances add value for IT and application developers

Machine OS

- Linux++: Linux-based OS for The Machine
  - A new concept of memory management
  - An emulator to make a conventional computer behave like The Machine
  - A developer's preview released in June 2015?
- Carbon
  - HP will replace Linux++ with Carbon.

FireBox Overview

1 Terabit/sec optical fibers

Up to 1000 SoCs +
High-BW Mem
(100,000 core total)

Inter-Box
Network

High Radix
Switches

Many Short Paths
Thru High-Radix Switches

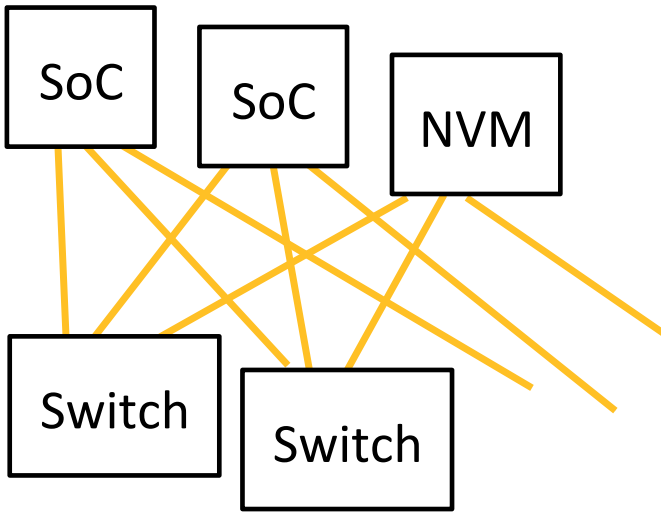
A similar concept of
The Machine

Up to 1000 NonVolatile
Memory Modules (100PB total)

ASPIRE UC Berkeley

Photonic Switches

- Monolithically integrated silicon photonics with Wave-Division Multiplexing (WDM)
  - A fiber carries 32 wavelengths, each 32Gb/s, in each direction
  - Off-chip laser optical supply, on-chip modulators and detectors
- Multiple radix-1000 photonic switch chips arranged as middle stage of Clos network (first and last Clos stage inside sockets)
- 2K endpoints can be configured as either SoC or NVM modules
- In Box, all paths are two fiber hops:
  - Electrical-photonic at socket
  - One fiber hop socket-to-switch
  - Photonic-electrical at switch
  - Electrical packet routing in switch
  - Electrical-photonic at socket
  - One fiber hop switch-to-socket
  - Photonic-electrical at socket



Electrical packet switching

IMPULSE: Initiative for Most Power-efficient Ultra-Large-Scale data Exploration

AIST IMPULSE

Non-Volatile Memory

- Voltage-controlled, magnetic RAM mainly for cache and work memories

Optical Network

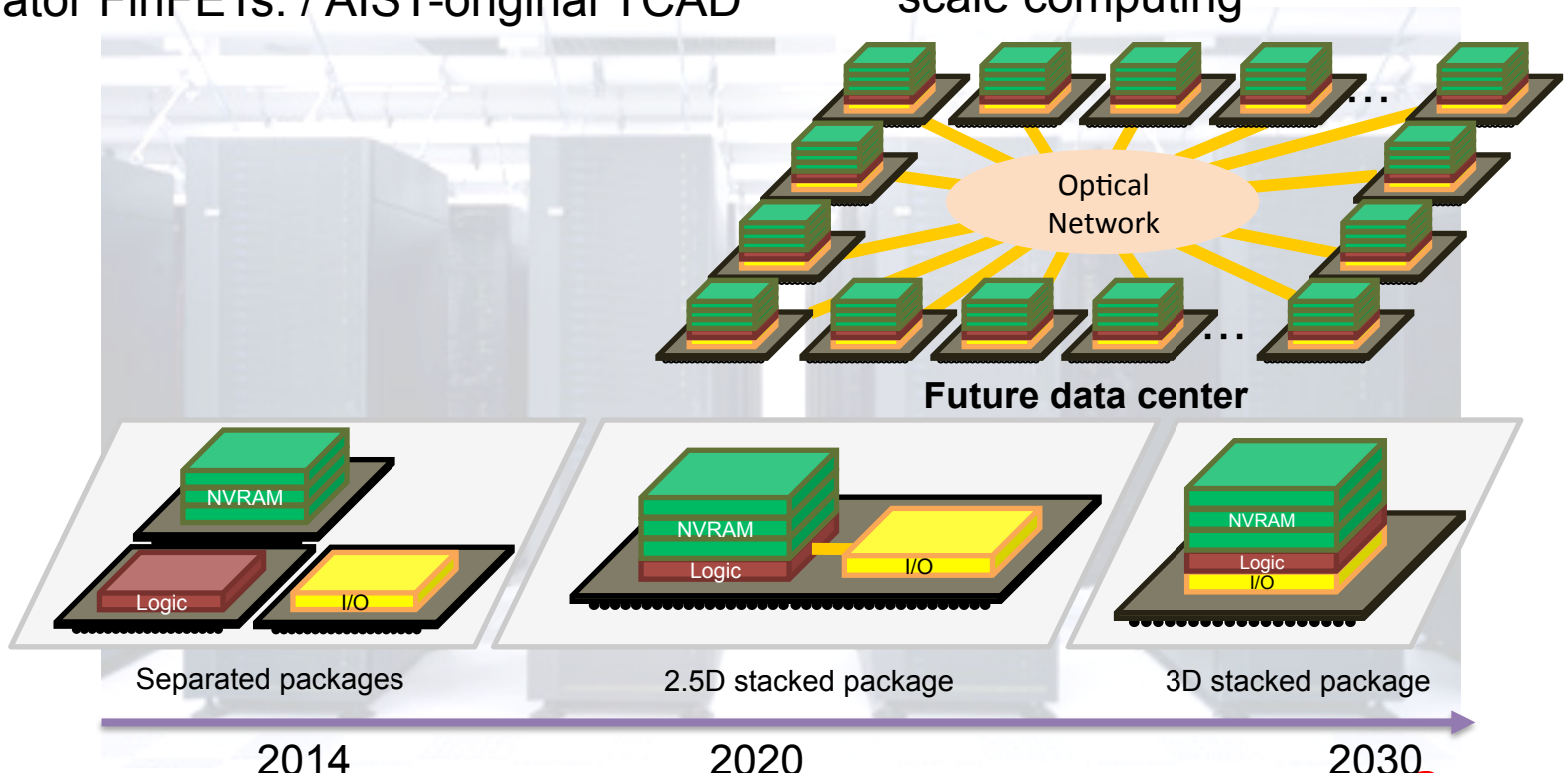
- Silicon photonics cluster SW
- Optical interconnect technologies

High-Performance Logic

- 3D build-up integration of the front-end circuits including high-mobility Ge-on-insulator FinFETs. / AIST-original TCAD

Architecture

- Future data center architecture design / Dataflow-centric warehouse-scale computing

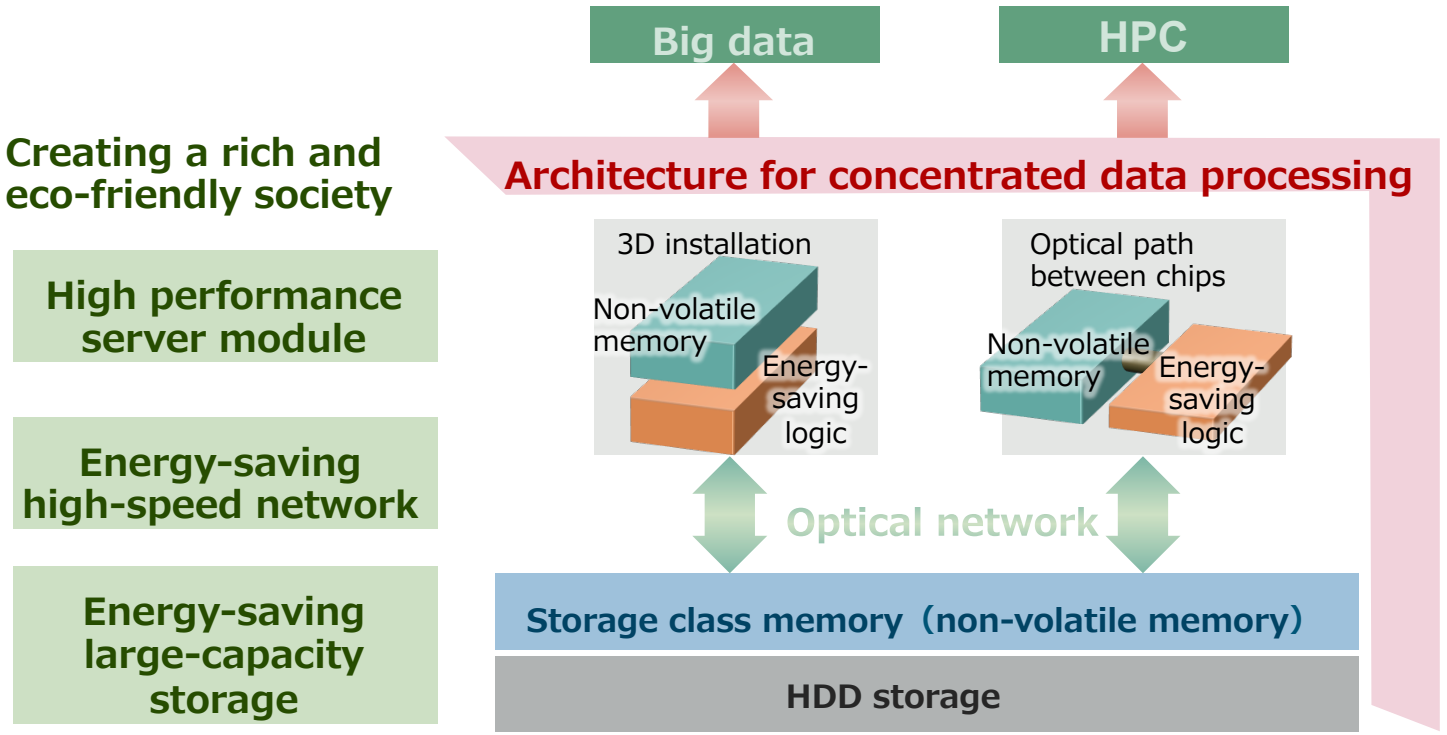


AIST's IMPULSE Program

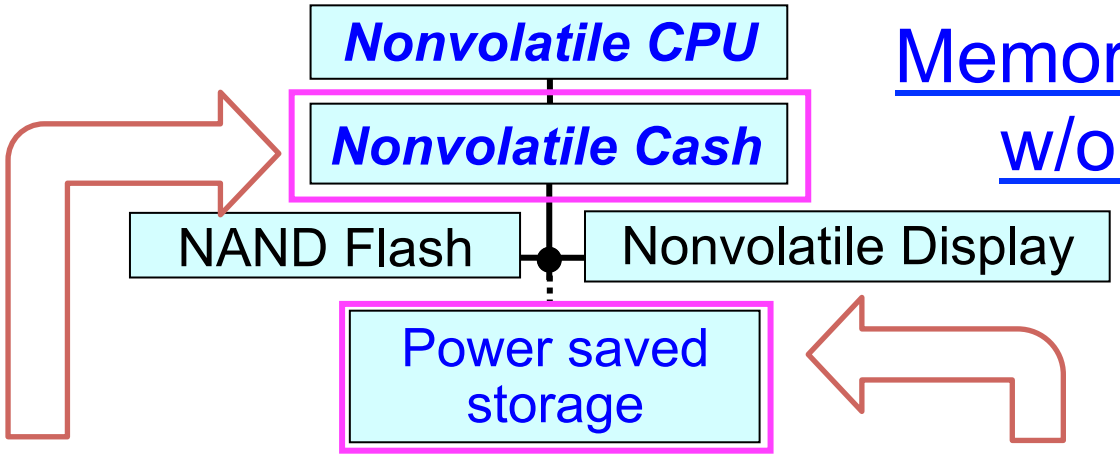
IMPULSE | STrategic AIST integrated R&D (STAR) program

＊STAR program is AIST research that will produce a large outcome in the future.

Initiative for Most Power-efficient Ultra-Large-Scale data Exploration

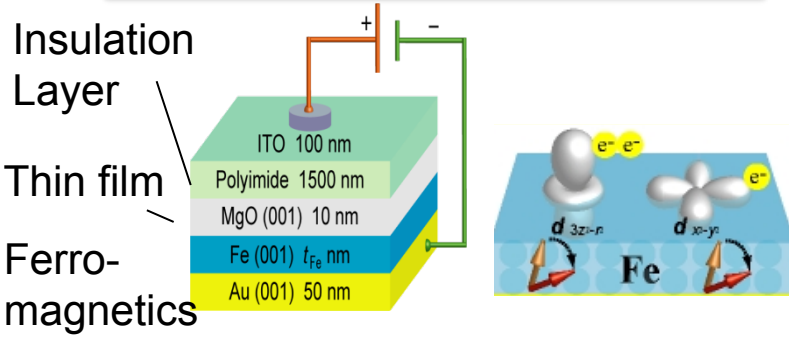


Voltage-controlled Nonvolatile Magnetic RAM



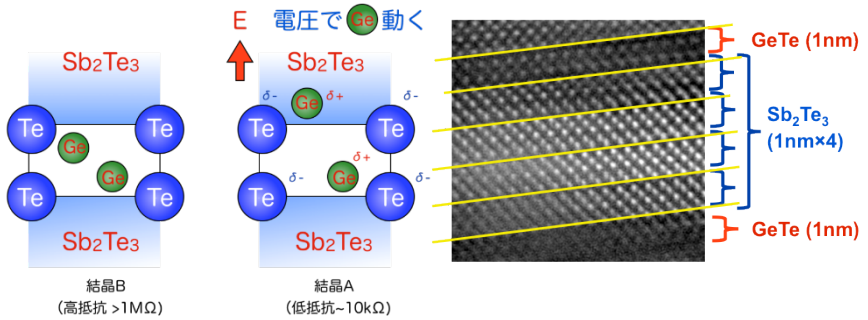
Memory keeping w/o power

Voltage Controlled Spin RAM



- voltage-induced magnetic anisotropy change
- Less than 1/100 rewriting power

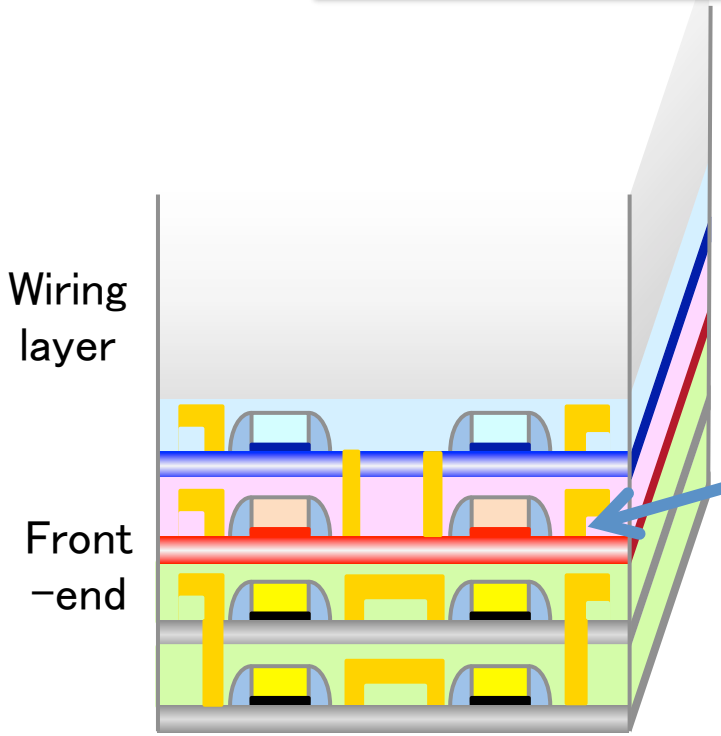
Voltage Controlled Topological RAM



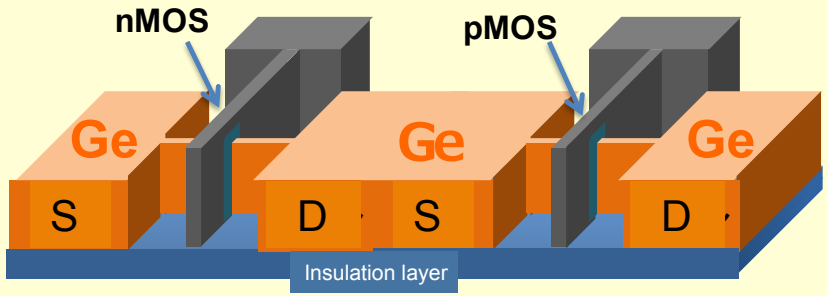
- Resistance change by the Ge displacement
- Loss by entropy: < 1/100

Low Power High-performance Logic

Front-end 3D integration



Ge Fin CMOS Tech.



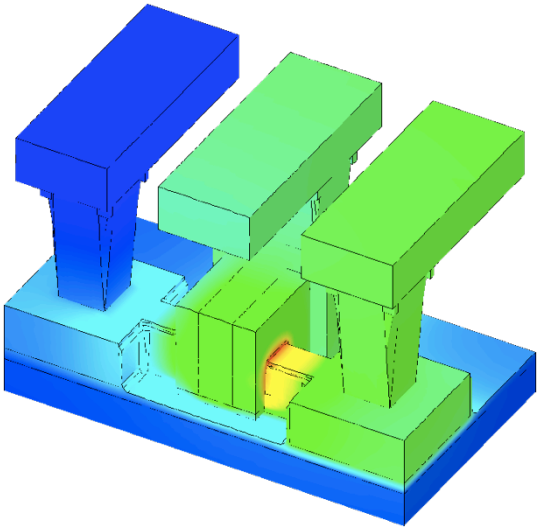
- Low−power/high−speed by Ge
- Toward 0.4V − Ge Fin CMOS

- Dense integration w/o miniaturization
- Reduction of the wiring length for power saving
- Introduction of Ge and III-V channels by simple stacking process
- Innovative circuit by using Z direction

Simulations

TCAD

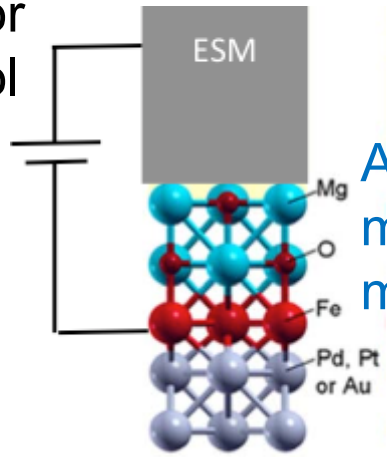
- Large-scale simulation for 3D structure, novel devices, and latest material



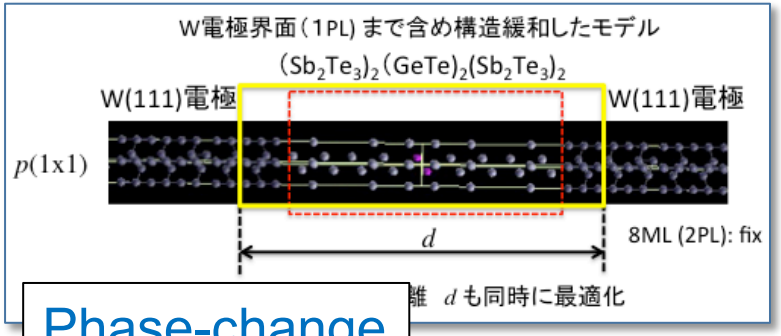
Simulation for temperature distribution

First-principle sim.

- Clarification for voltage-control
- Parameter setting to the TCAD



Anisotropic magnetic material



Phase-change

Optical Network Technology for Future Datacenters

- Large-scale silicon photonics based cluster switches
- DWDM, multi-level modulation, highly integrated “elastic” optical interconnects
- Ultra-low energy consumption network by making use of optical switches

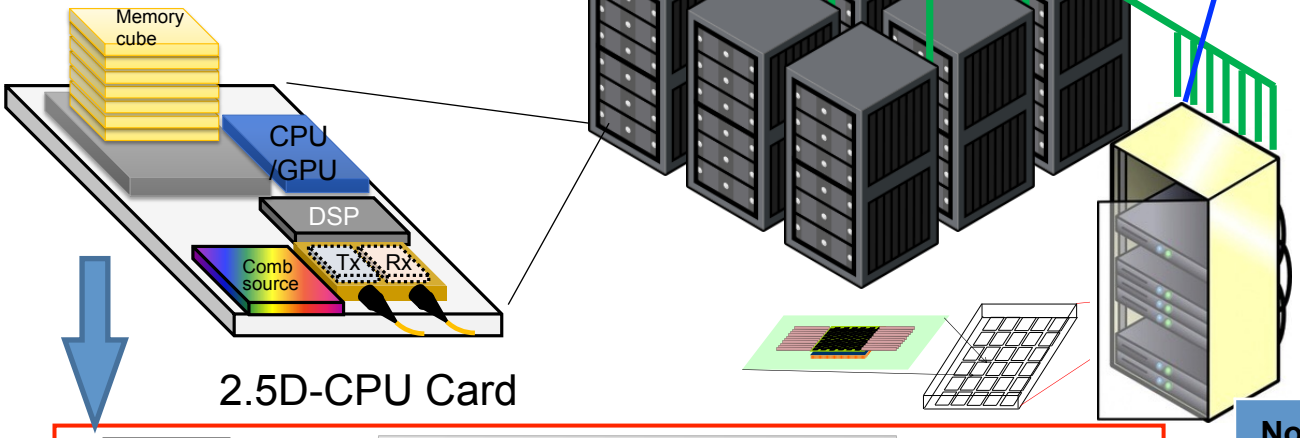
DWDM, multi-level modulation optical interconnects

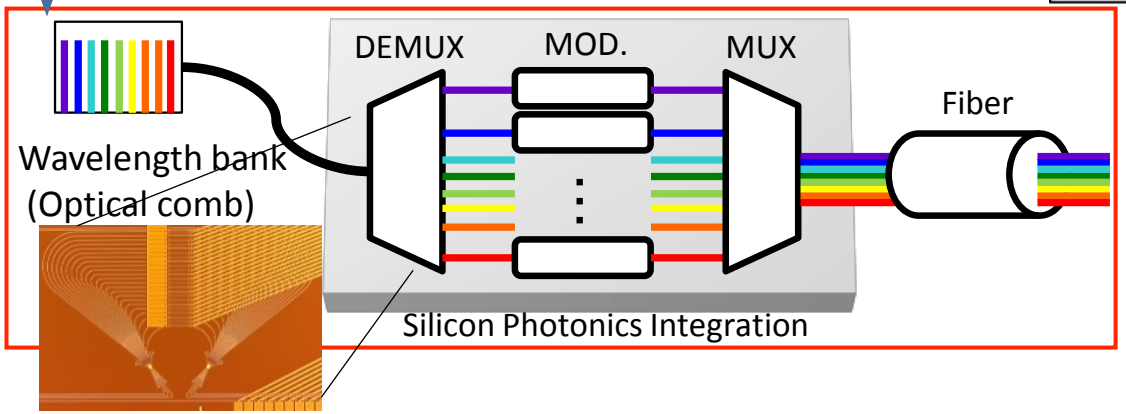
Datacenter server racks

Silicon photonics cluster switches

- Ultra-compact switches based on silicon photonics
- 3D integration by amorphous silicon
- A new server architecture

Current electrical switches:
～130Tbps → ～500Pbps





No of λs	Order of mod.	Bit rate
1	1	20 Gbps
4	8	640 Gbps
32	8	5.12 Tbps

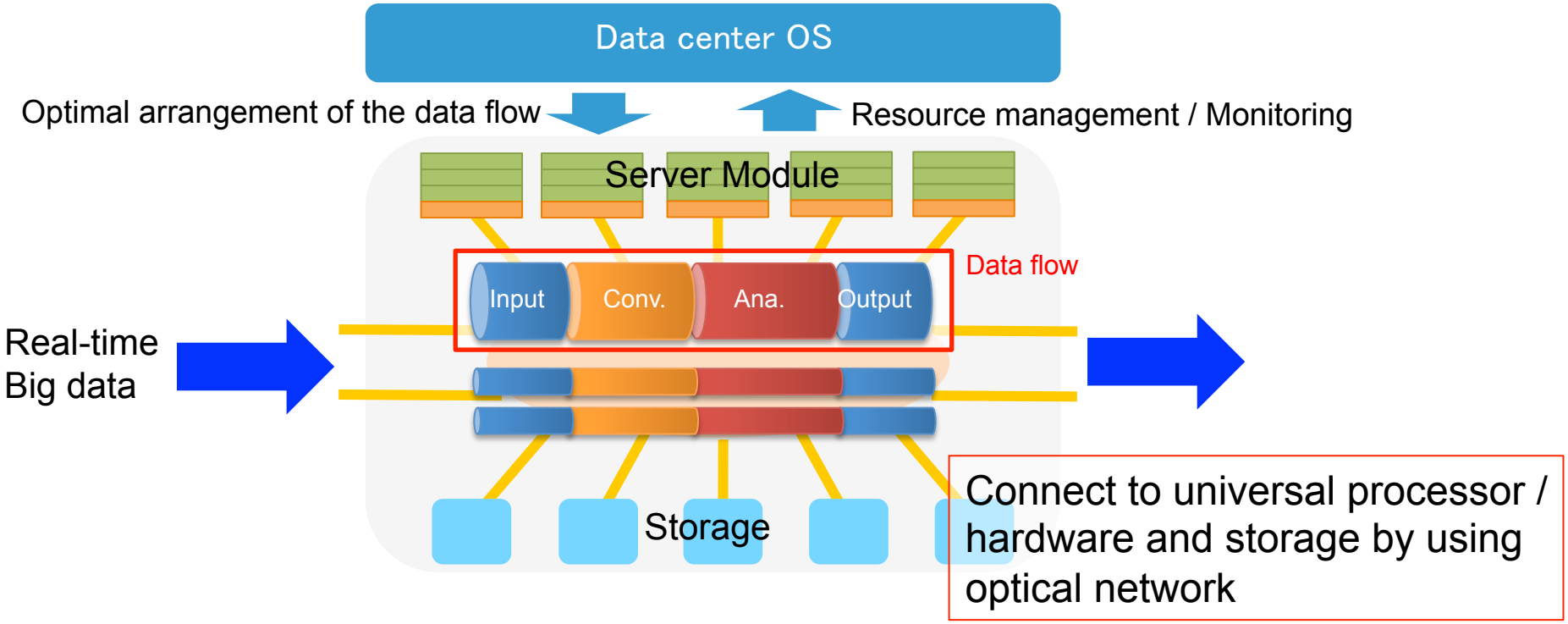
Current state-of-the-art Tx
100Gbps → ～ 5.12Tbps

Architecture for Big Data and Extreme-scale Computing

Data flow centric warehouse scale computing

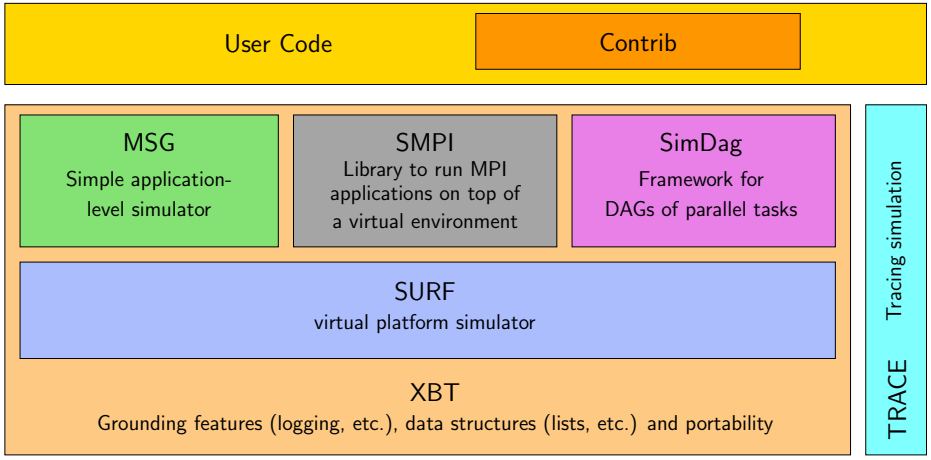
1 - Single OS controls entire data center

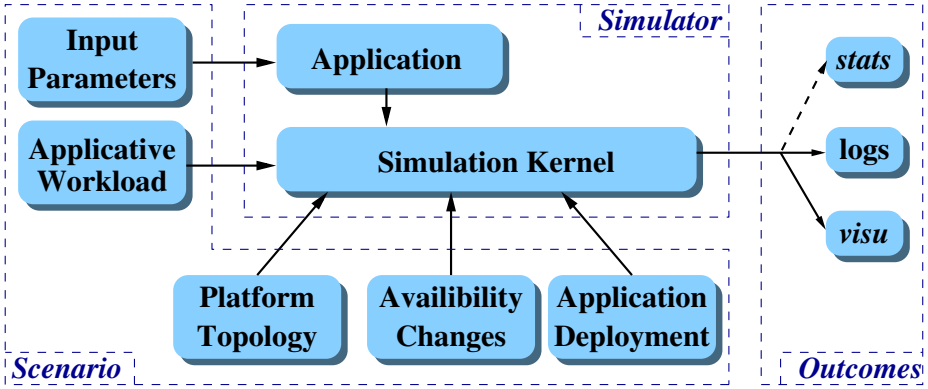
2 - Split a data center OS into the data plane and the control plane to guarantee real-time data processing



Performance Estimation

- Estimate the performance of typical both HPC and BigData workloads on a future datacenter system
- **SimGrid simulator**
  - Simulator of large-scale distributed Systems, such as Grids, Clouds, HPC, and P2P.



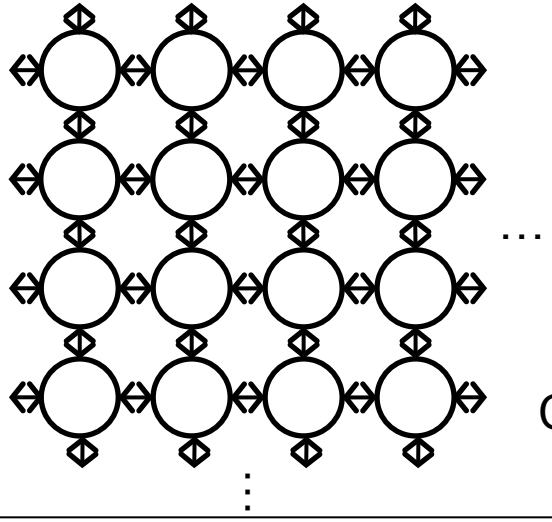


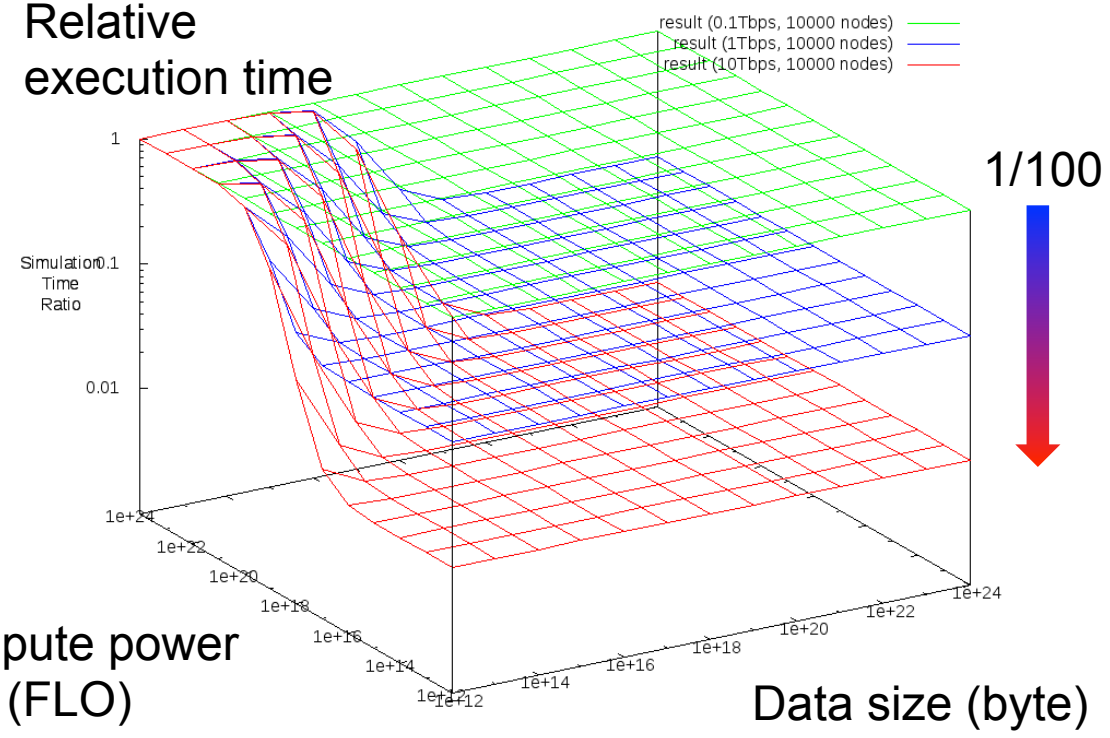
http://simgrid.gforge.inria.fr

Workload 1: Simple Message Passing

- Iteration of neighbor communication (bottom left)
- Big impact of increasing link bandwidth if an application is network intensive.

#node: 10000
Link bandwidth: 0.1, 1, 10Tbps
Link latency 100ns
CPU power: 10TFLOPS
Data size : 10^{12}～10^{24}B

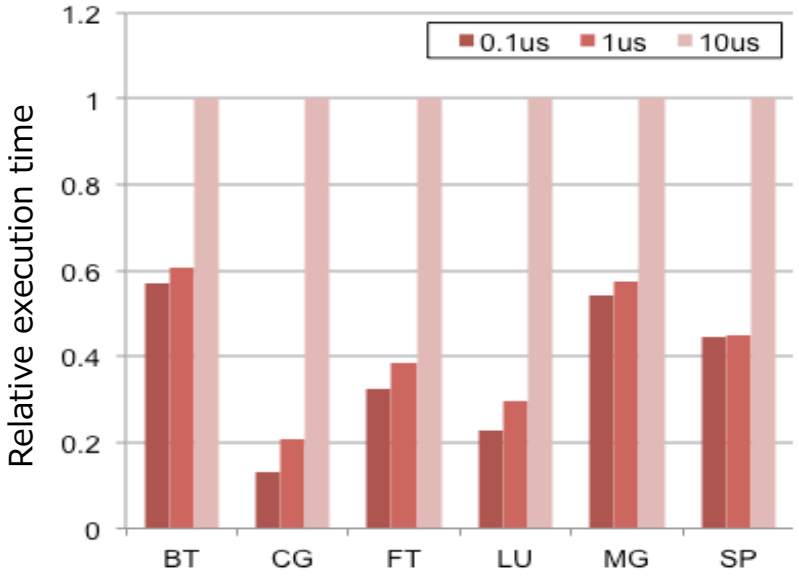




Workload 2: HPC Application

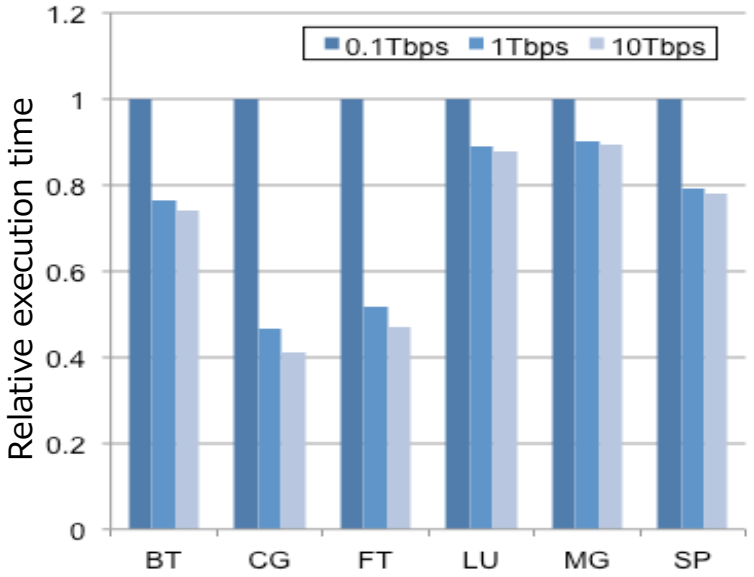
- NAS Parallel Benchmark (256 procs, class C)
  - Low latency is more important than huge bandwidth.
  - The problem size is too small to utilize huge bandwidth.

Effect of reducing the link latency



CPU power：1TFLOPS
Link bandwidth：1Tbps

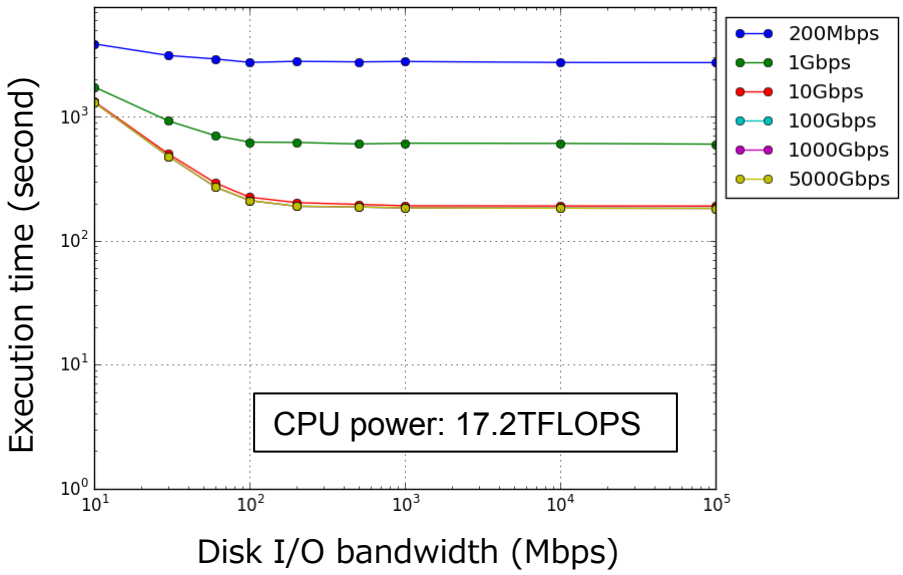
Effect of increasing the link bandwidth

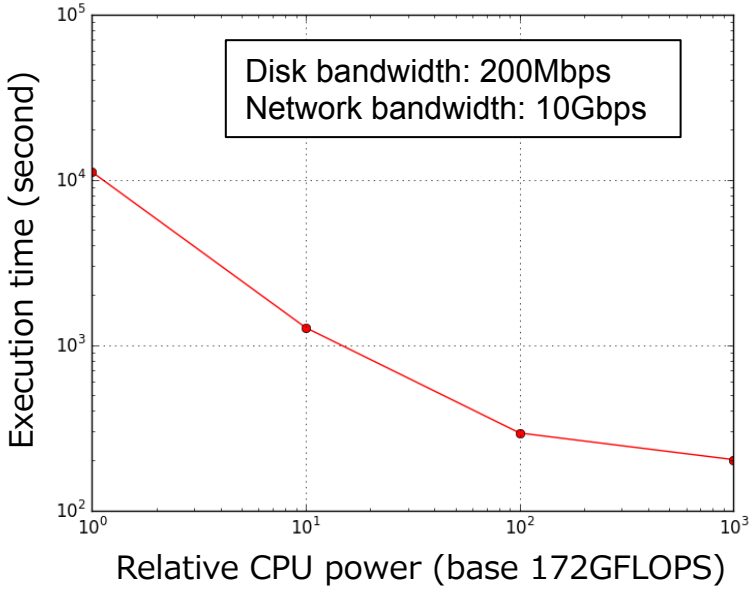


CPU power：1TFLOPS
Link latency：0.1us

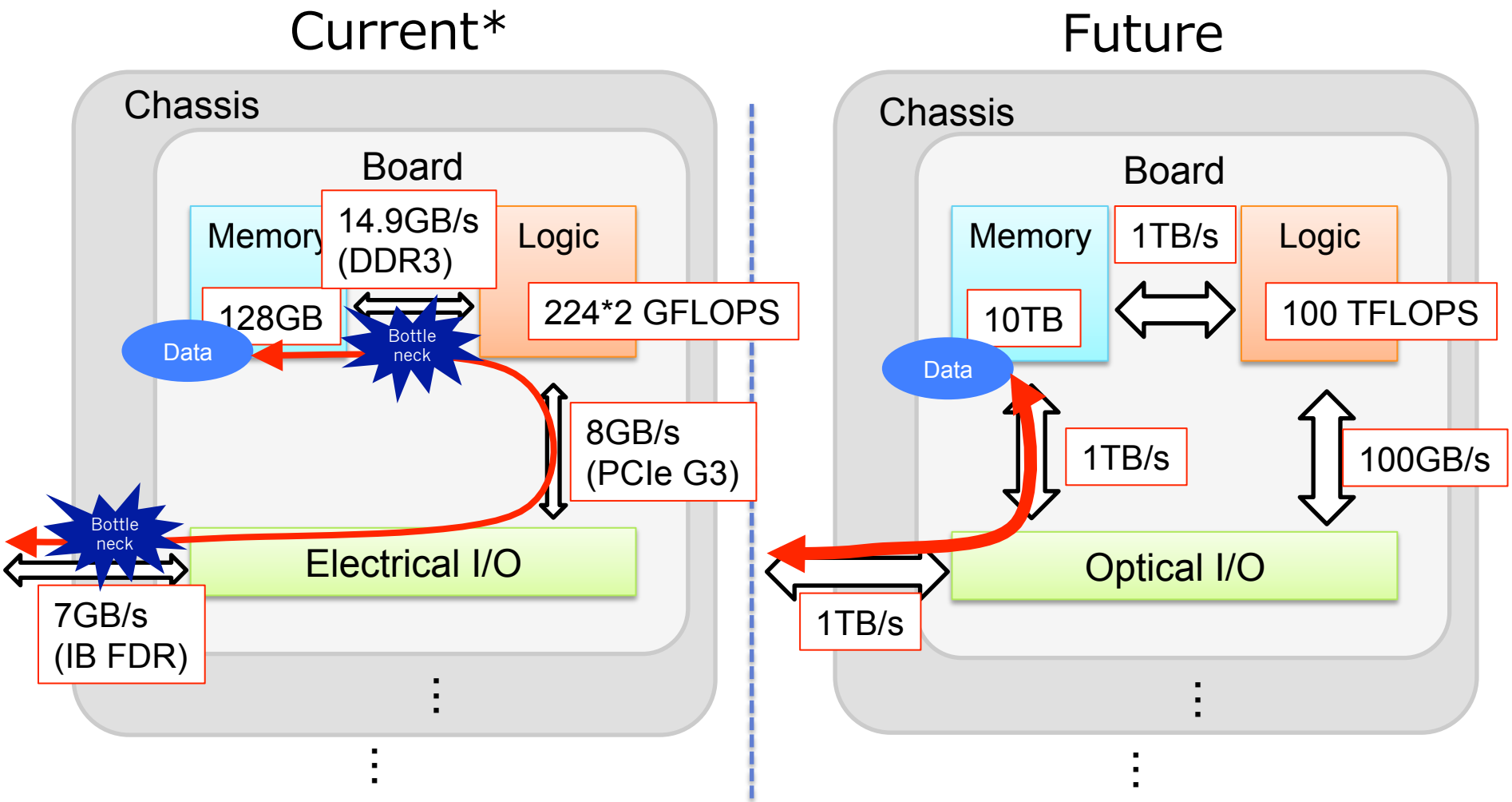
Workload 3: MapReduce

- KDD Cup 2012, Track 2: predict the click-through rate of ads (using Hadoop and Hivemall)
  - Machine learning is CPU intensive.
  - The effect of huge bandwidth is limited, because...
    - The concurrency of the used model is not enough.
    - Hadoop is optimized to make jobs run faster on the current I/O devices.



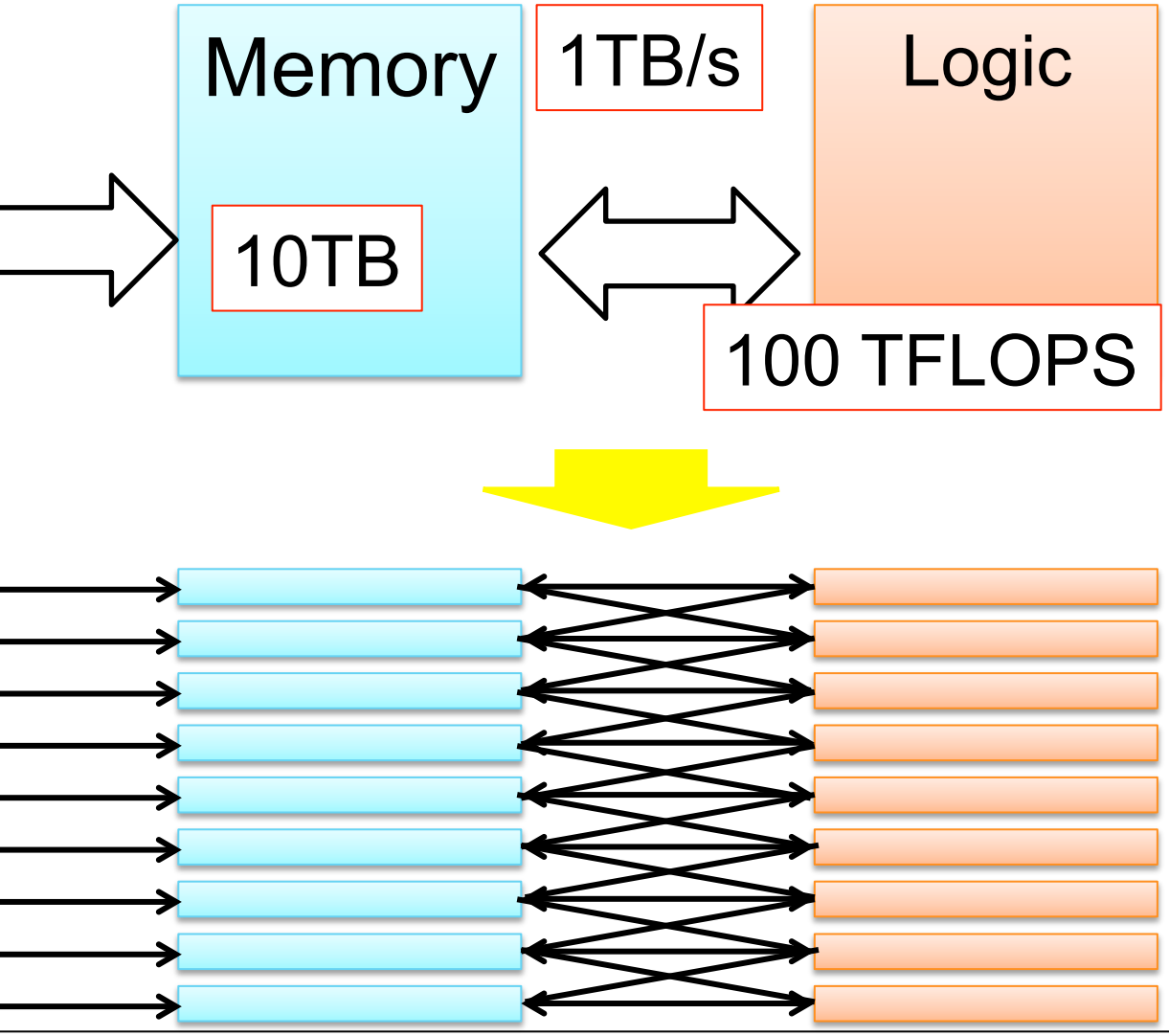


Data Movement Problems



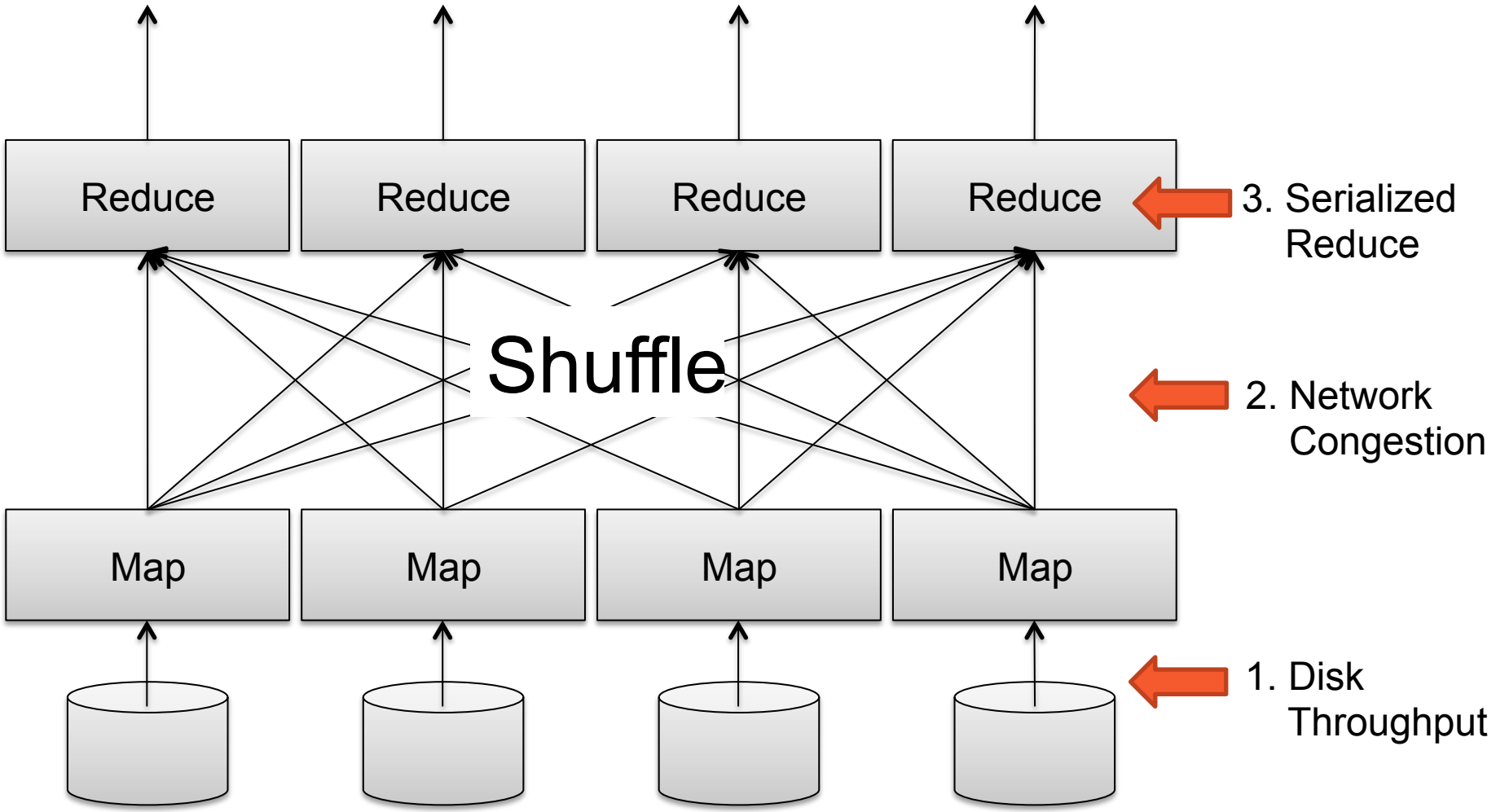
* AIST Super Green Cloud (ASGC)

Parallelization inside Package

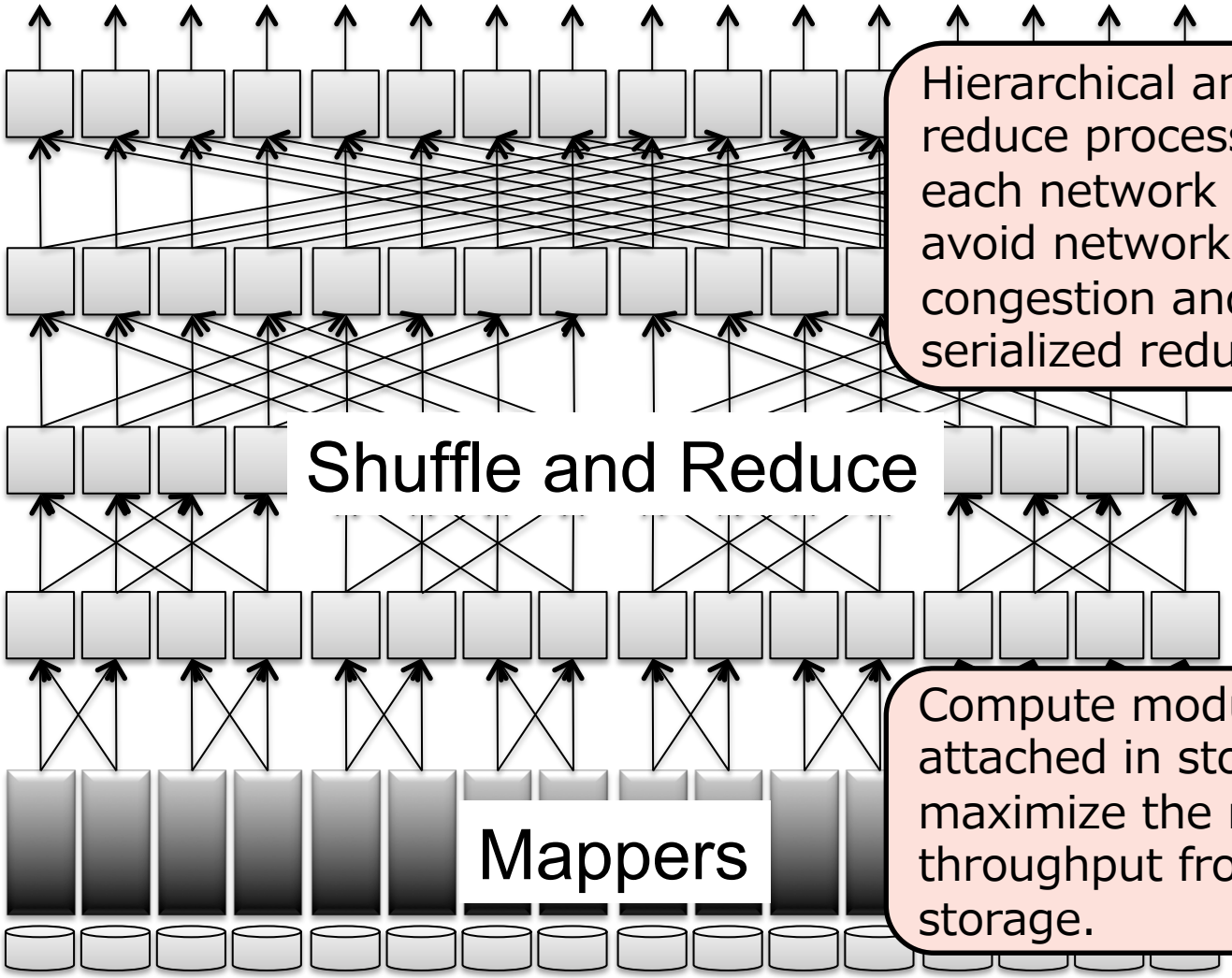


- Need efficient parallelized structure
  - On-chip network
  - Interconnection

Bottlenecks in MapReduce



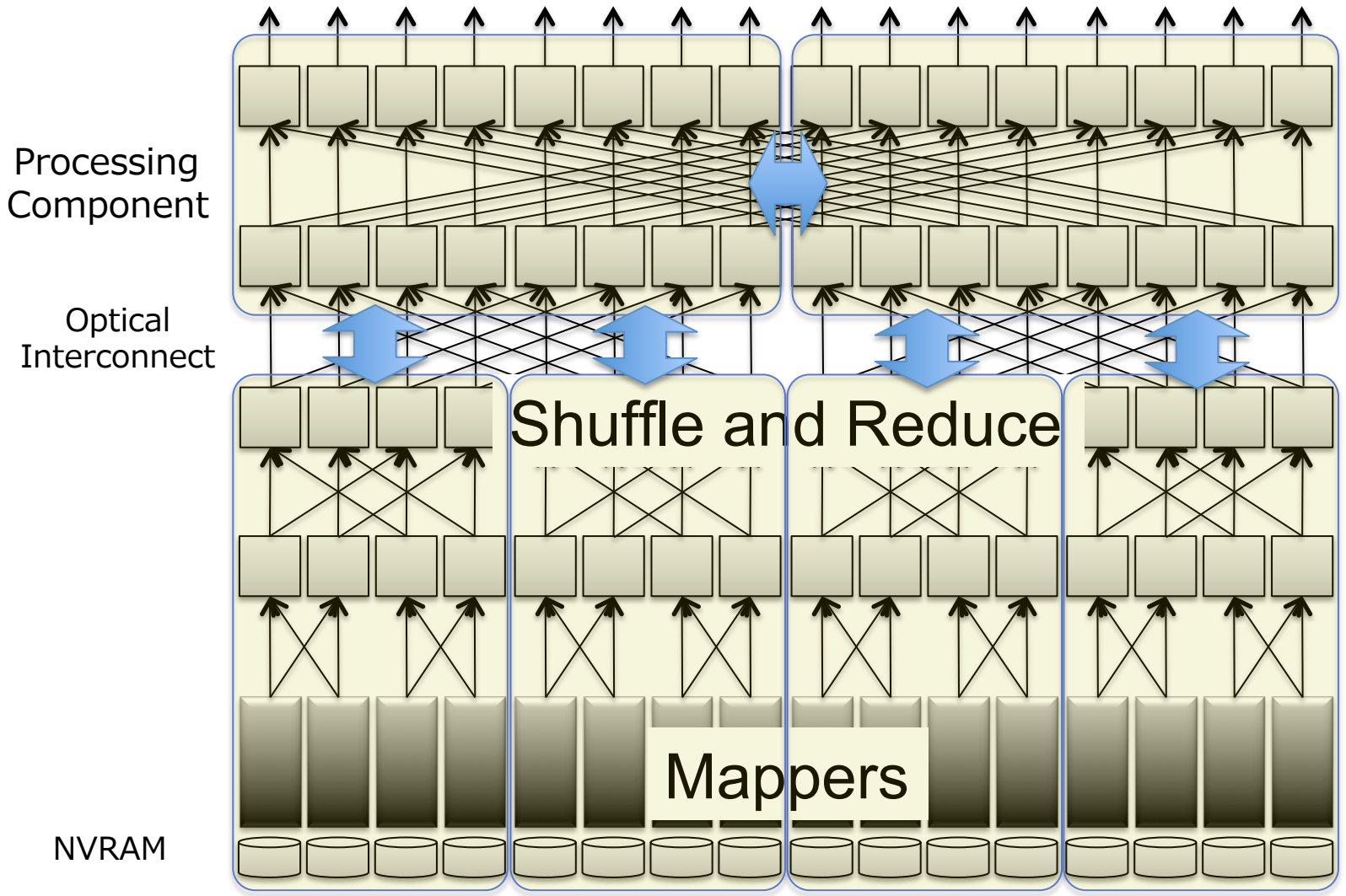
In-storage, network processing



Hierarchical and partial reduce processing in each network node to avoid network congestion and serialized reduce.

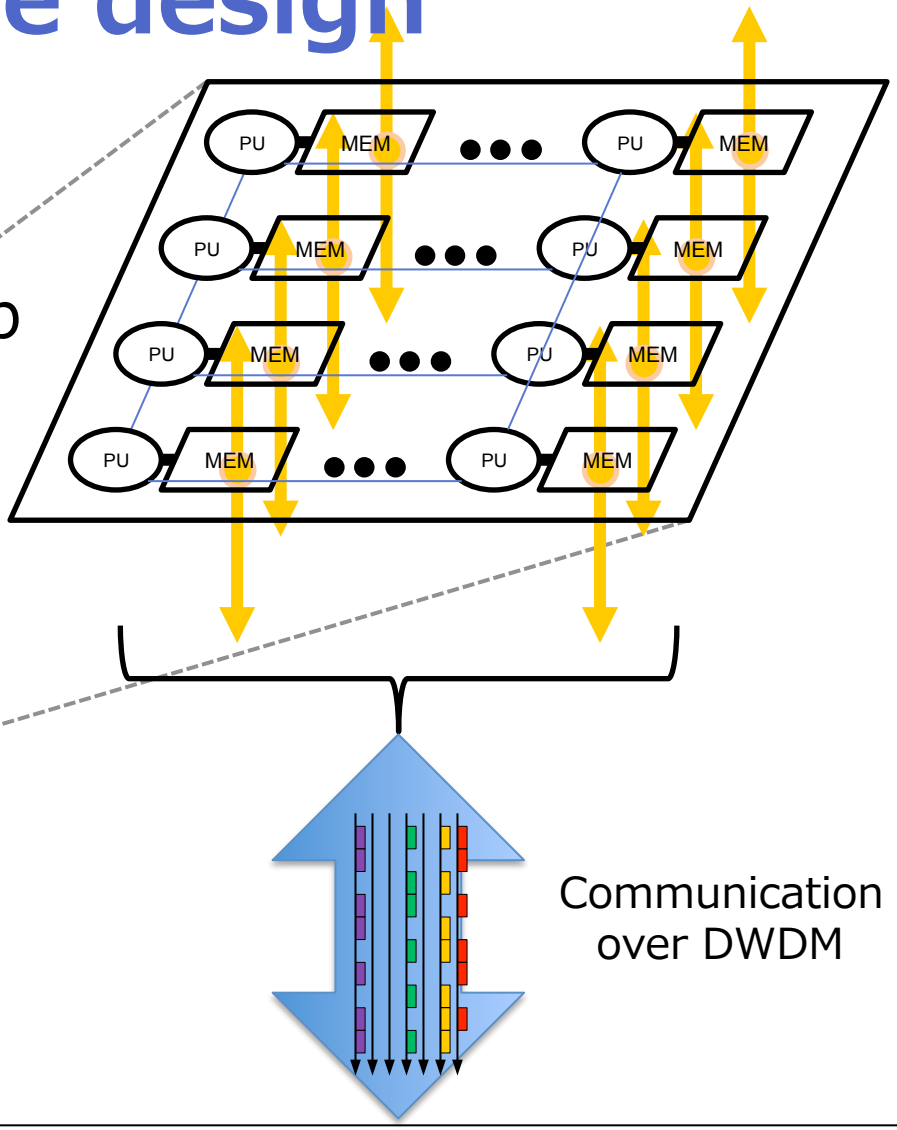
Compute modules are attached in storage to maximize the read throughput from storage.

In-storage, network processing



In-storage, network processing: hardware design

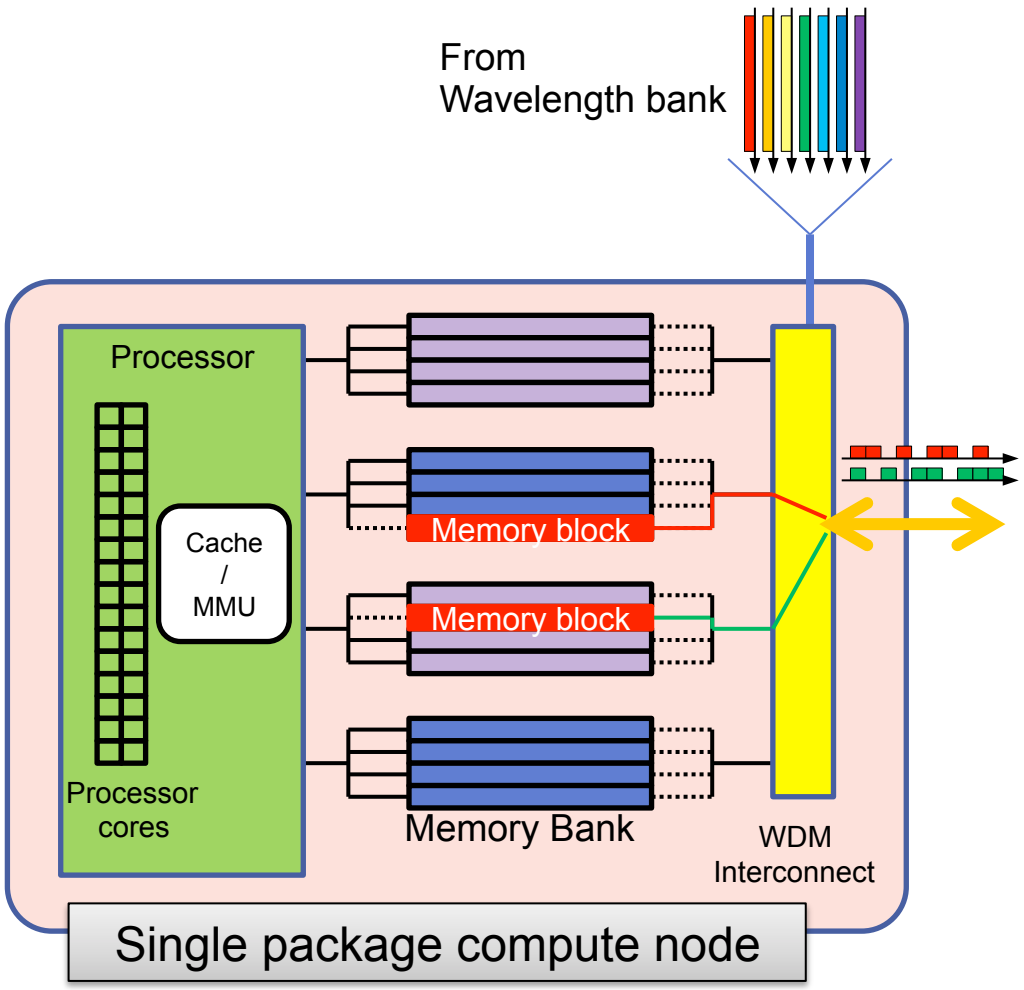
Direct optical I/O connection to non-volatile memory modules distributed on a chip



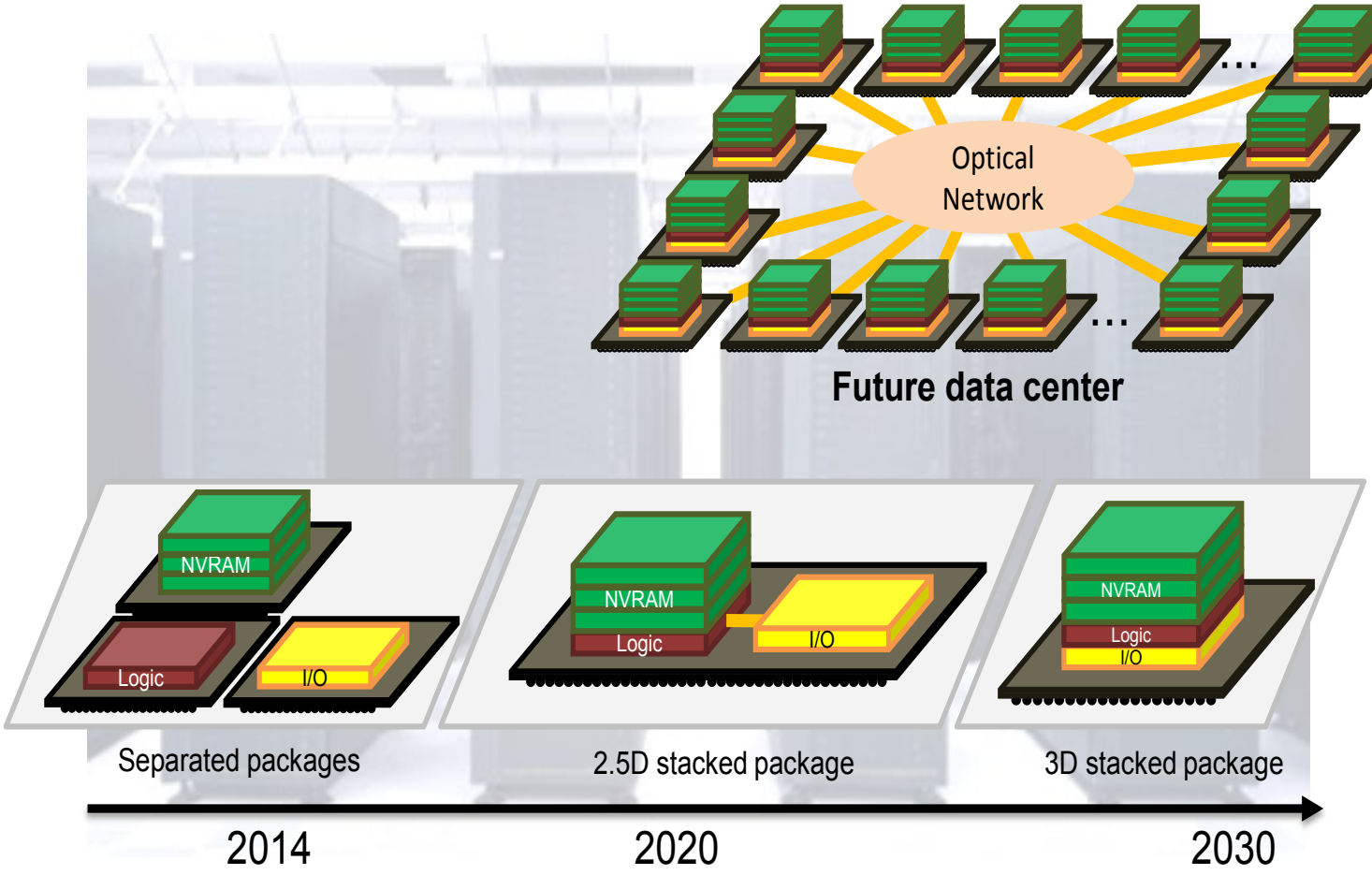
Communication over DWDM

Direct Memory Copy over DWDM

- Assume processor-memory embedded package with WDM interconnect.
- To fully utilize the huge I/O bandwidth realized by DWDM.
- Multiple memory blocks can be sent/received simultaneously using multiple wavelengths.
- Memory-centric network is a similar idea [PACT13]

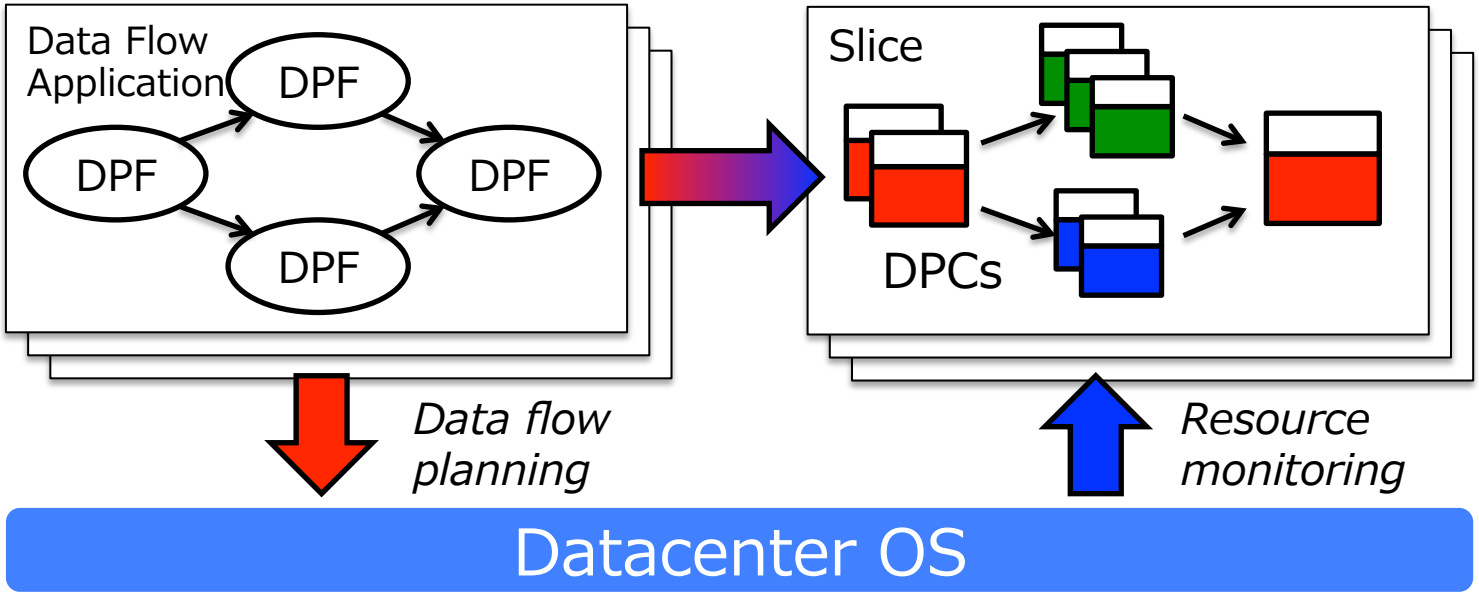


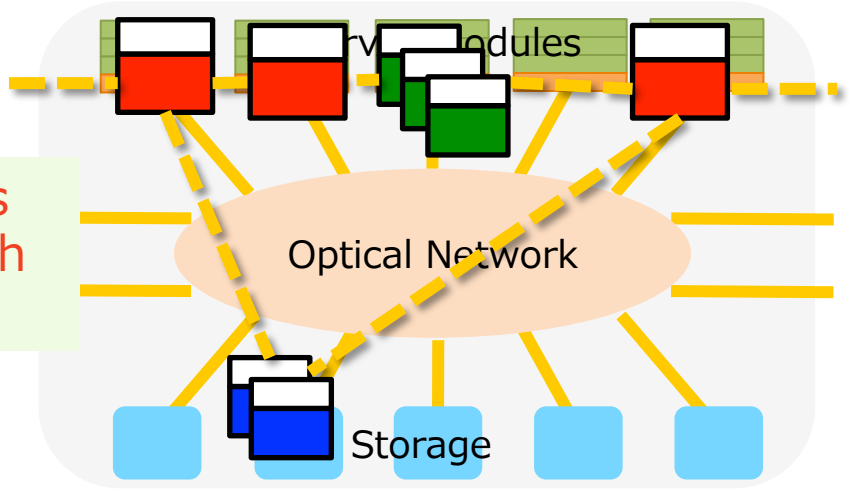
Our Vision of Future Datacenter



Goal: 100x energy efficiency of data processing

Dataflow Processing System



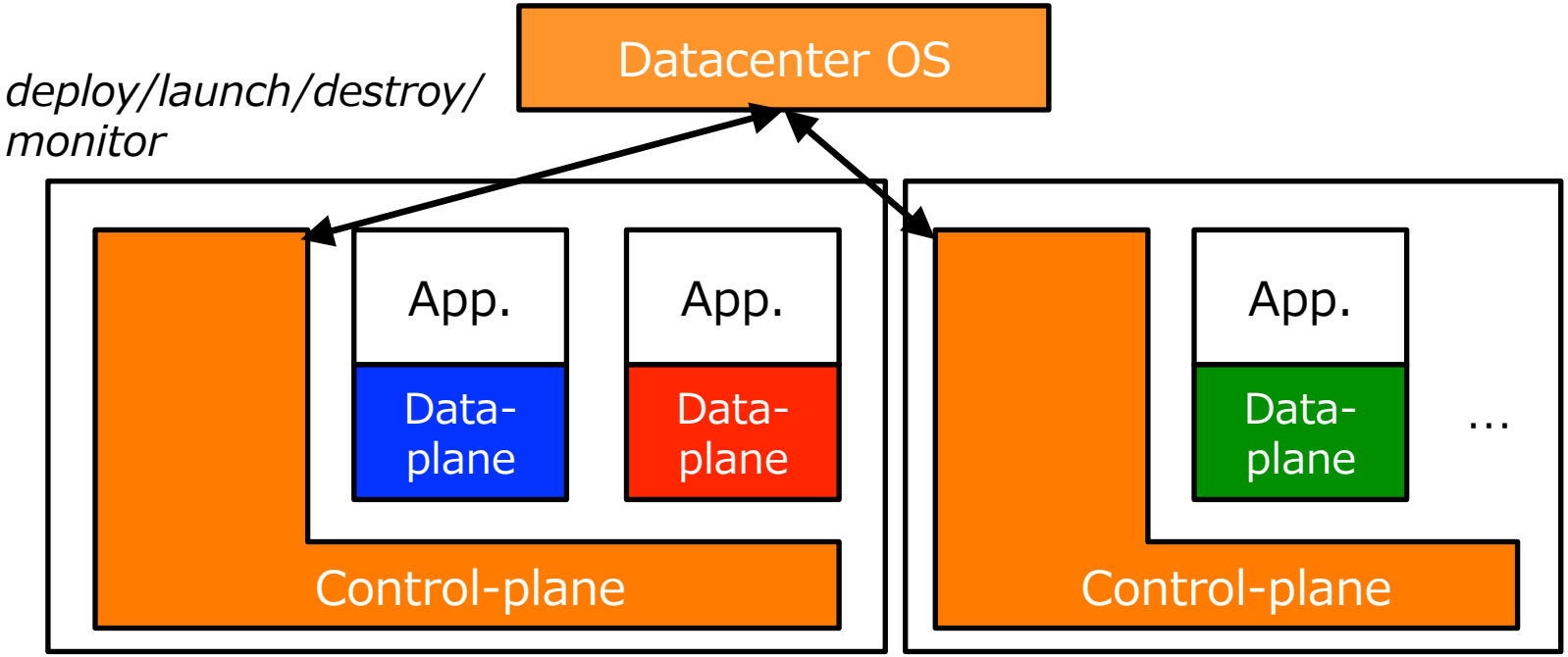


Co-allocate DPCs and network path between them.

DPF: Data Processing Function
DPC: Data Processing Component

IMPULSE Datacenter OS

- A single OS for datacenter-wide optimization of the energy efficiency and the performance
- Separation of data plane and control plane:
  - Data plane is an application specific library OS.
  - Control plane manages resources (server, network, etc).



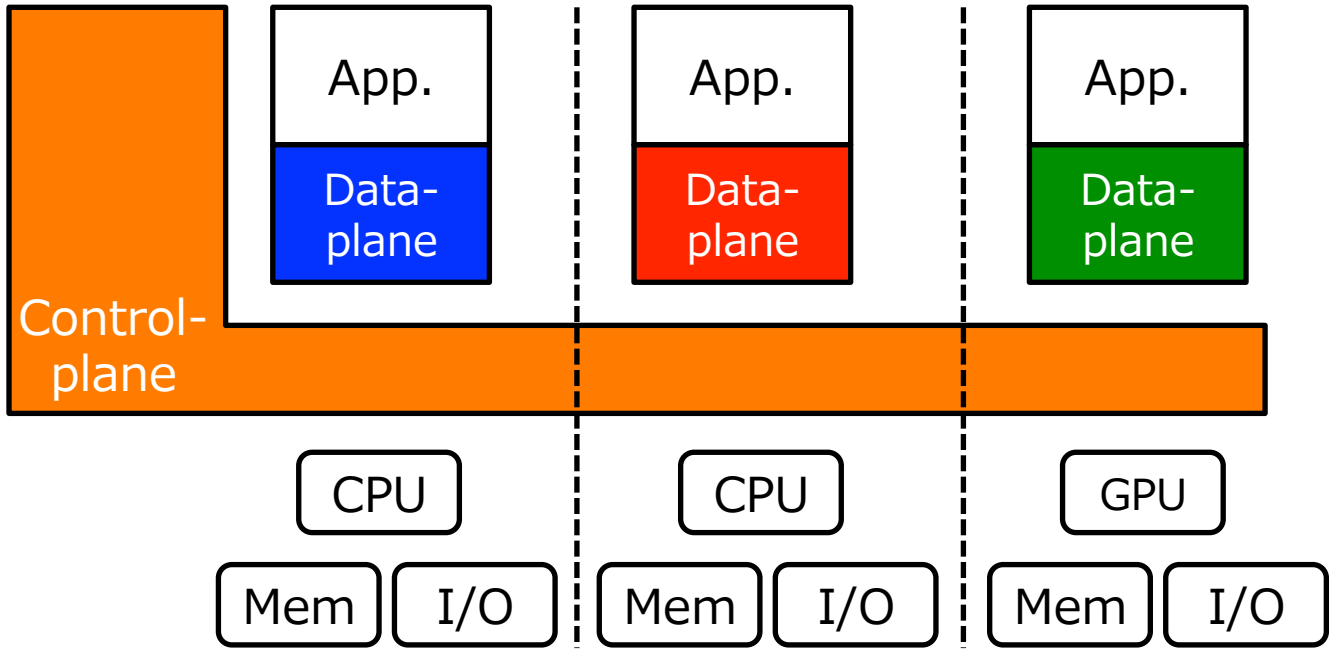
IMPULSE Datacenter OS

Control plane

- Resource management
- Logical and secure resource partitioning for data planes
- Running on the firmware

Data plane

- Application specific library OS
  - E.g., machine learning, data store, etc.
- Mitigate the OS overhead to fully utilize high performance devices.



Related Work

- Datacenter-wide resource management
  - OpenStack, Apache CloudStack, Kubernetes
  - Hadoop YARN, Apache Mesos
- Dataflow processing engine
  - Google Cloud Dataflow
  - Lambda architecture
- Control and data planes separated design in OS
  - Arrakis (U. Washington)
  - IX (Stanford)

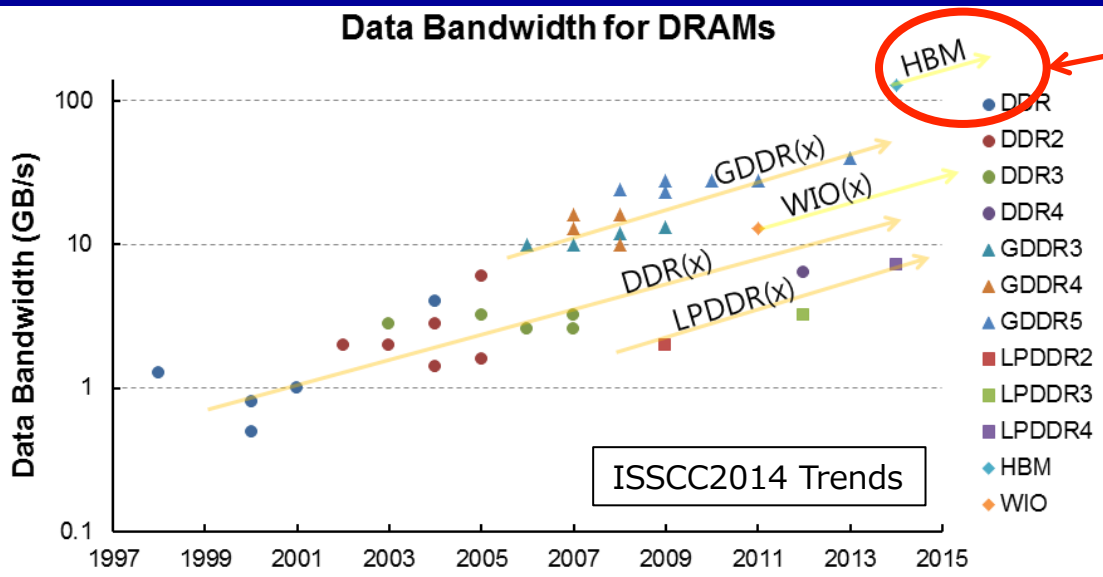
Summary

- New visions of future datacenters: “disaggregation” and “datacenter in a box”
- **Optical network is key** to making them.
- Hardware and software **co-design** is critical.
- Optical path network encourages **C/D separation** in a datacenter OS.
  - **Control plane** manages resources and establishes a path between data processing components.
  - **Data plane** fully utilizes the huge bandwidth.

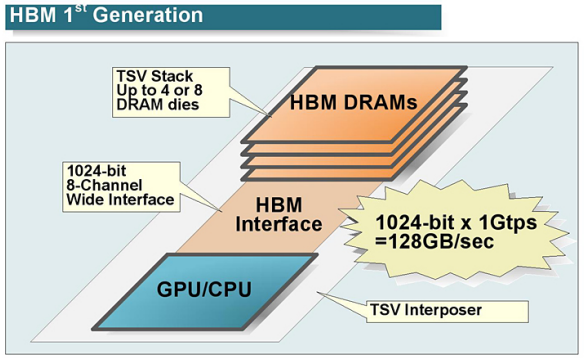
Join us!
Thanks for your attention

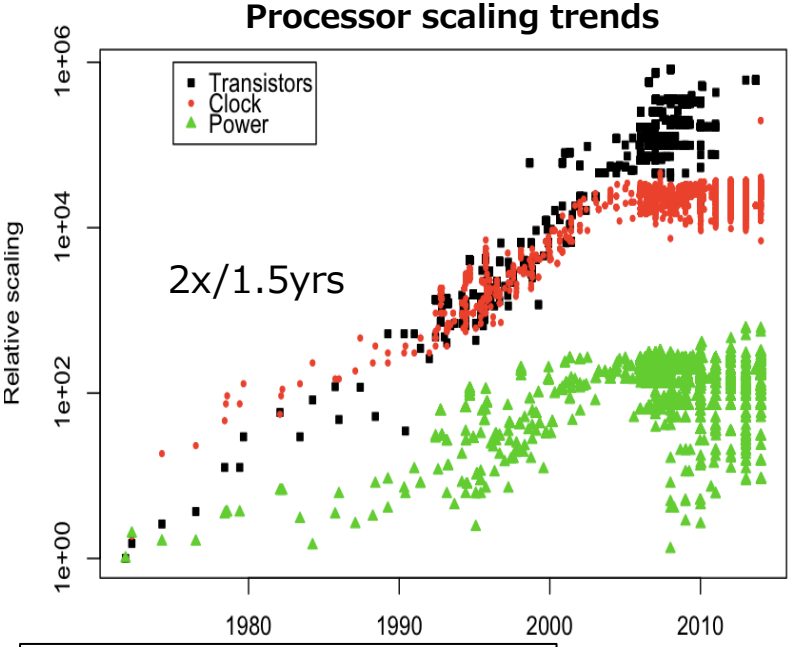
Reference

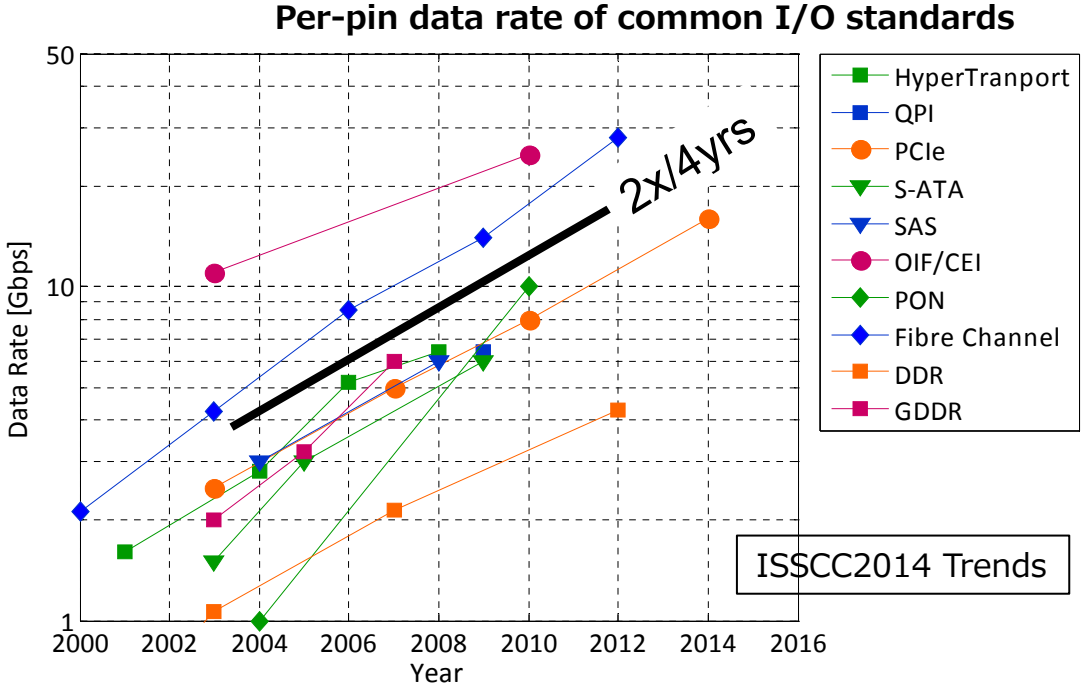
- Rack scale architecture for Cloud, IDC2013
  - https://intel.activeevents.com/sf13/connect/fileDownload/session/6DE5FDFBF0D0854E73D2A3908D58E1E2/SF13_CLDS001_100.pdf
- Intel rack scale architecture overview, Interop2013
  - http://presentations.interop.com/events/las-vegas/2013/free-sessions---keynote-presentations/download/463
- New technologies that disrupt our complete ecosystem and their limits in the race to Zettascale, HPC2014
  - http://www.hpcc.unical.it/hpc2014/pdfs/demichel.pdf
- HPが「Tech Power Club」で見せた“未来のサーバー技術”, ASCII.jp
  - http://ascii.jp/elem/000/000/915/915508/



High Bandwidth Memory

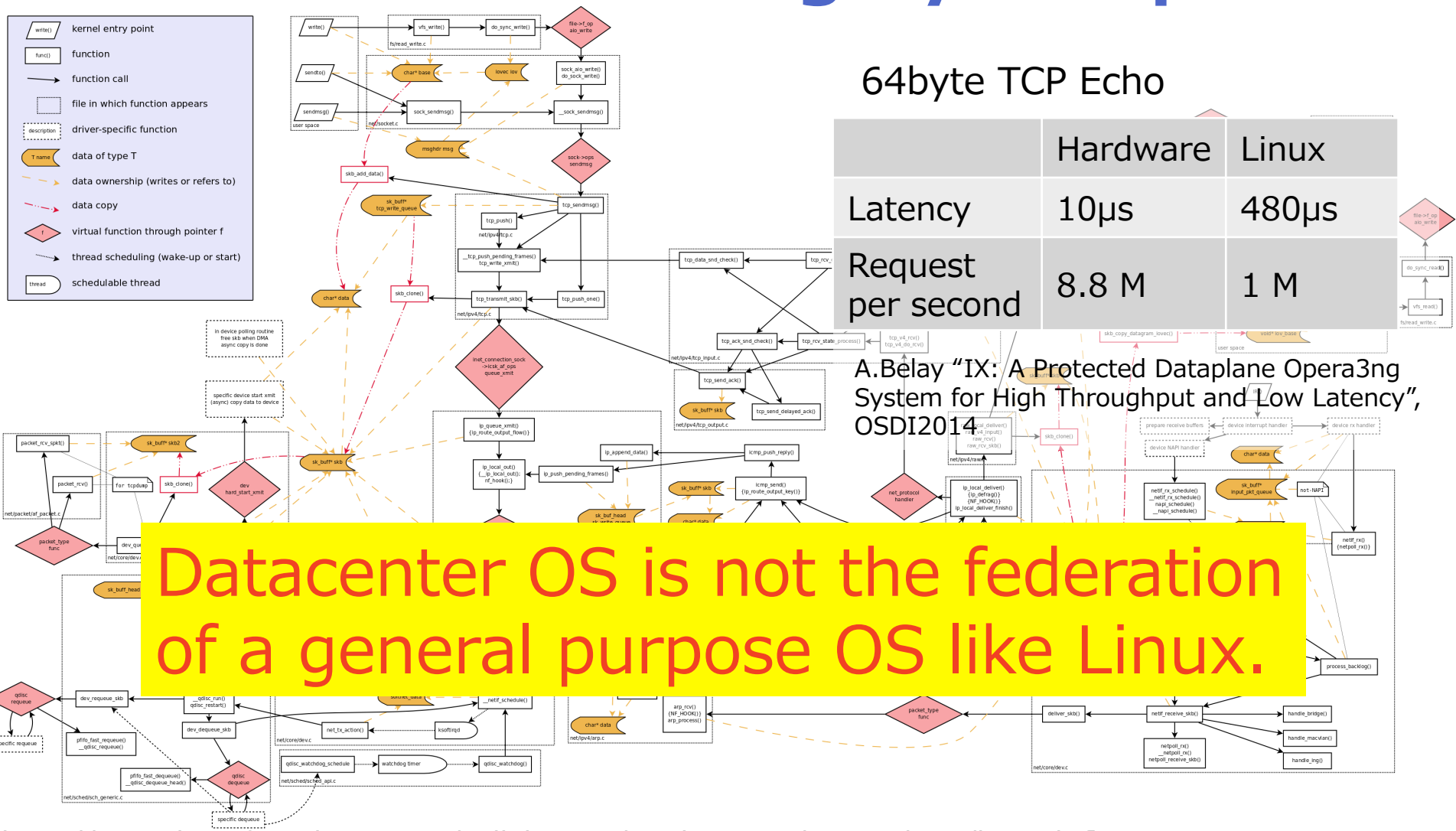






Data source:
http://cpudb.stanford.edu/

Linux kernel is highly complex



64byte TCP Echo

	Hardware	Linux
Latency	10µs	480µs
Request per second	8.8 M	1 M

A.Belay “IX: A Protected Dataplane Opera3ng System for High Throughput and Low Latency”, OSDI2014

Datacenter OS is not the federation of a general purpose OS like Linux.

http://www.linuxfoundation.org/collaborate/workgroups/networking/kernel_flow