ポストムーア時代の通信と計算のバランスとアクセラレータクラウド構想

工藤知宏

東京大学 情報基盤センター

産業技術総合研究所 情報技術研究部門

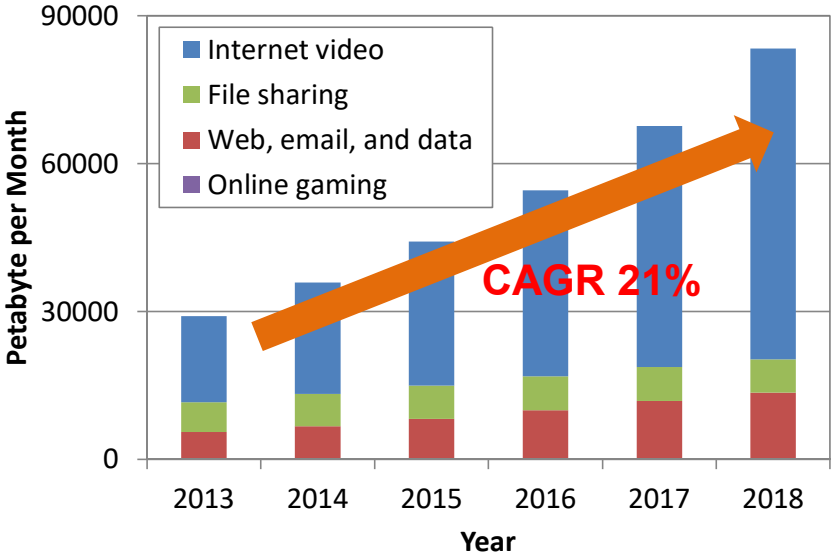
本発表で紹介する研究の一部はJSPS科研費16H02793の助成を受けたものです

Computation and Communication

- Computation and communication have been two key components of IT infrastructure for a long time
  - Integration of C and C is becoming more important
    - Data Intensive / AI / IoT applications
    - Heterogeneous computing for post-Moore era
- C&C (Computer and Communication) has been a slogan (or symbol) of NEC corporation since 1977
  - At that time, the Internet was at the very early stage. For example, TCP/IP did not exist. The word “Communication” was used for broadcast and online computers.
  - Now the term C&C usually stands for Command and Control servers, which are used attackers to control malware affected computers…

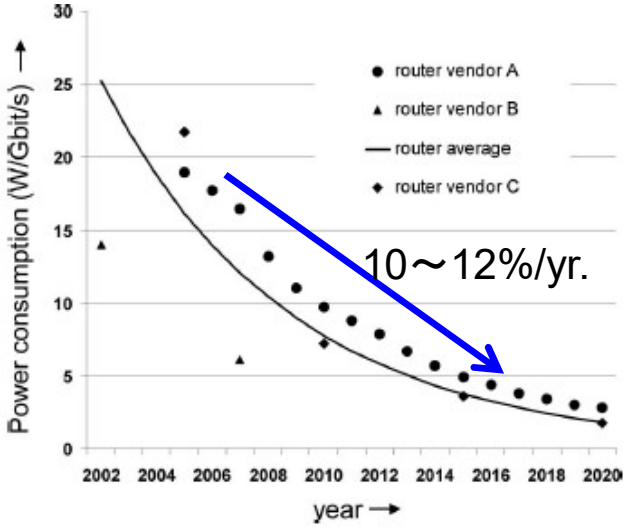
Wide area network trend

Global Consumer Internet Traffic



Source: Cisco VNI Global Forecast 2013-2018

Energy Efficiency Trend in IP routers



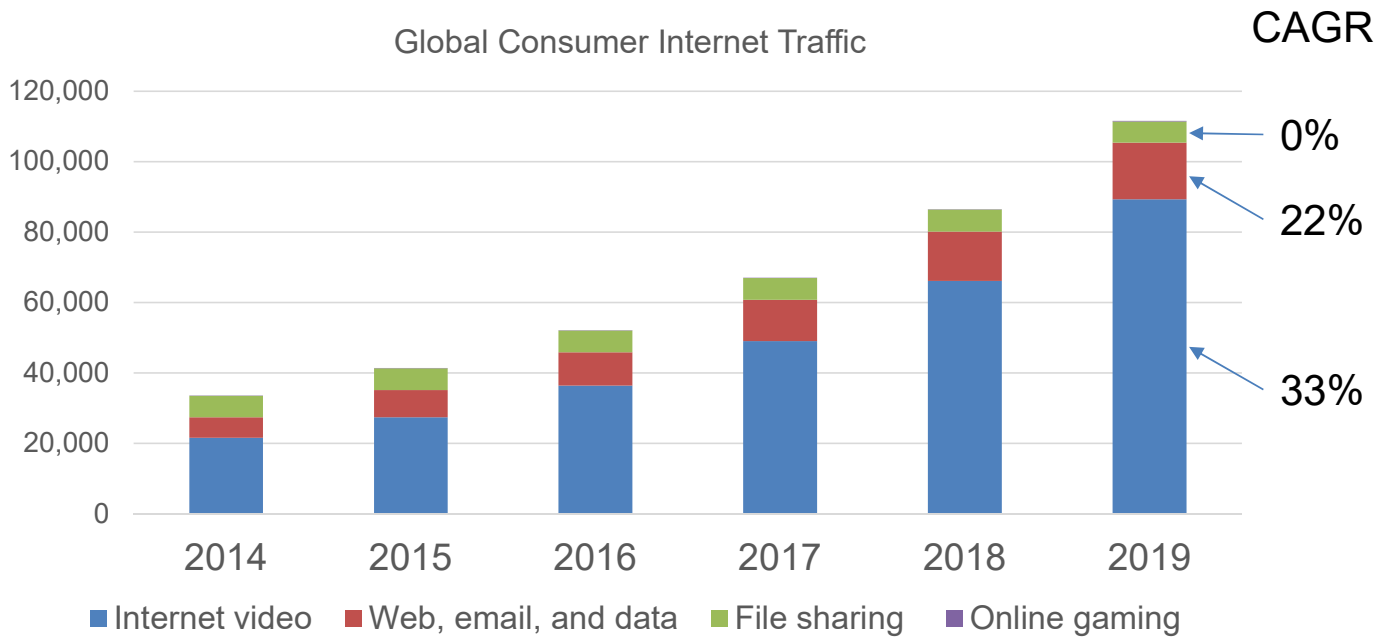
Source: C. Lange, et. al., IEEE JSTQE, vol. 17, no.2, pp. 285-295, March/April, 2011, Fig. 10

Traffic Growth Rate > Energy Efficiency Improvement Rate

Concern about energy consumption in communication networks is increasing

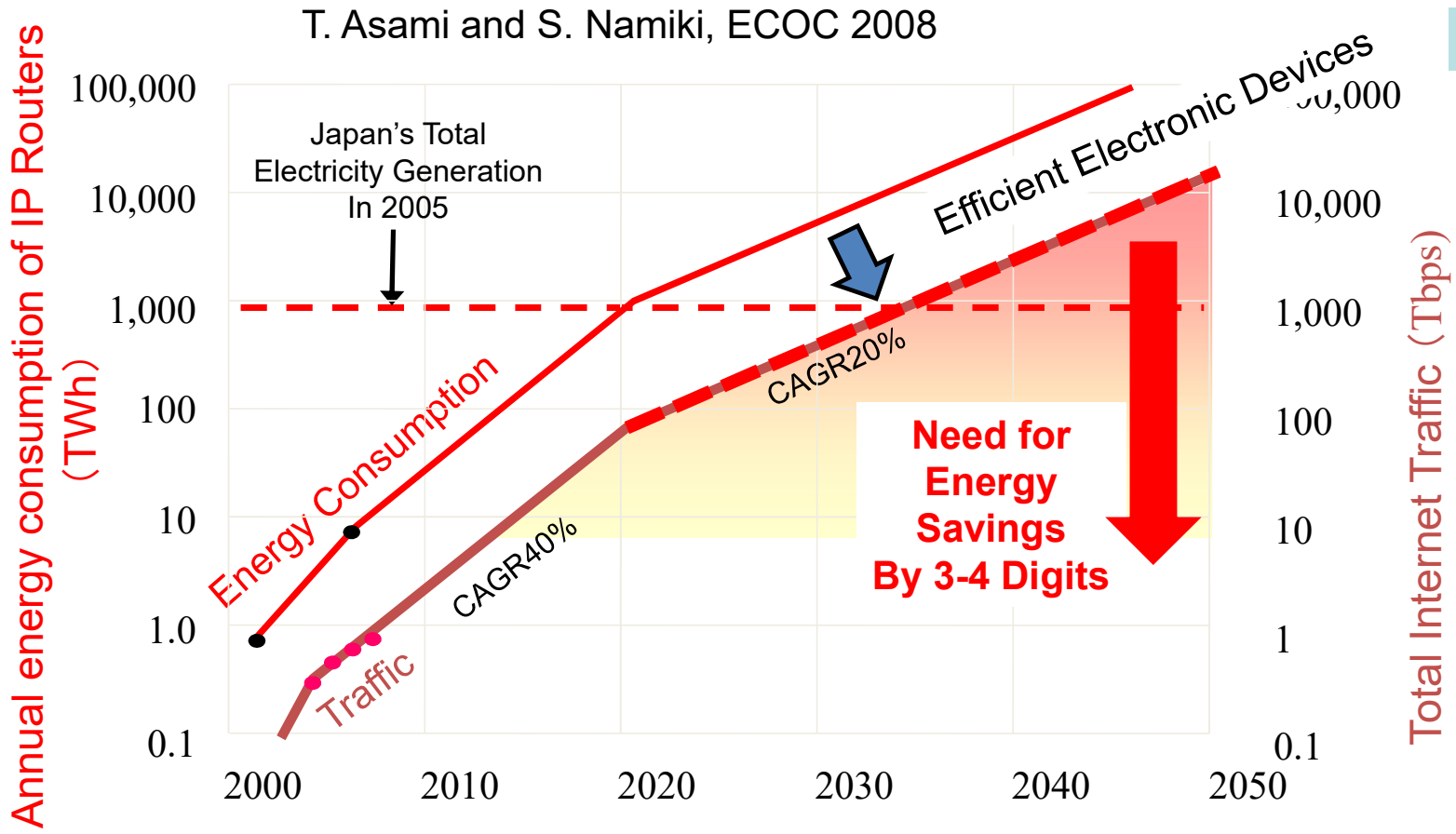
Ever-increasing video related traffic

- Increasing users due to YouTube + SmartPhone
- Higher definition: HD→4K→8K
- Demands to real-time share of video→uncompressed



Source: Cisco Visual Networking Index:Forecast and Methodology, 2014–2019

The Fundamental Problem of the Internet



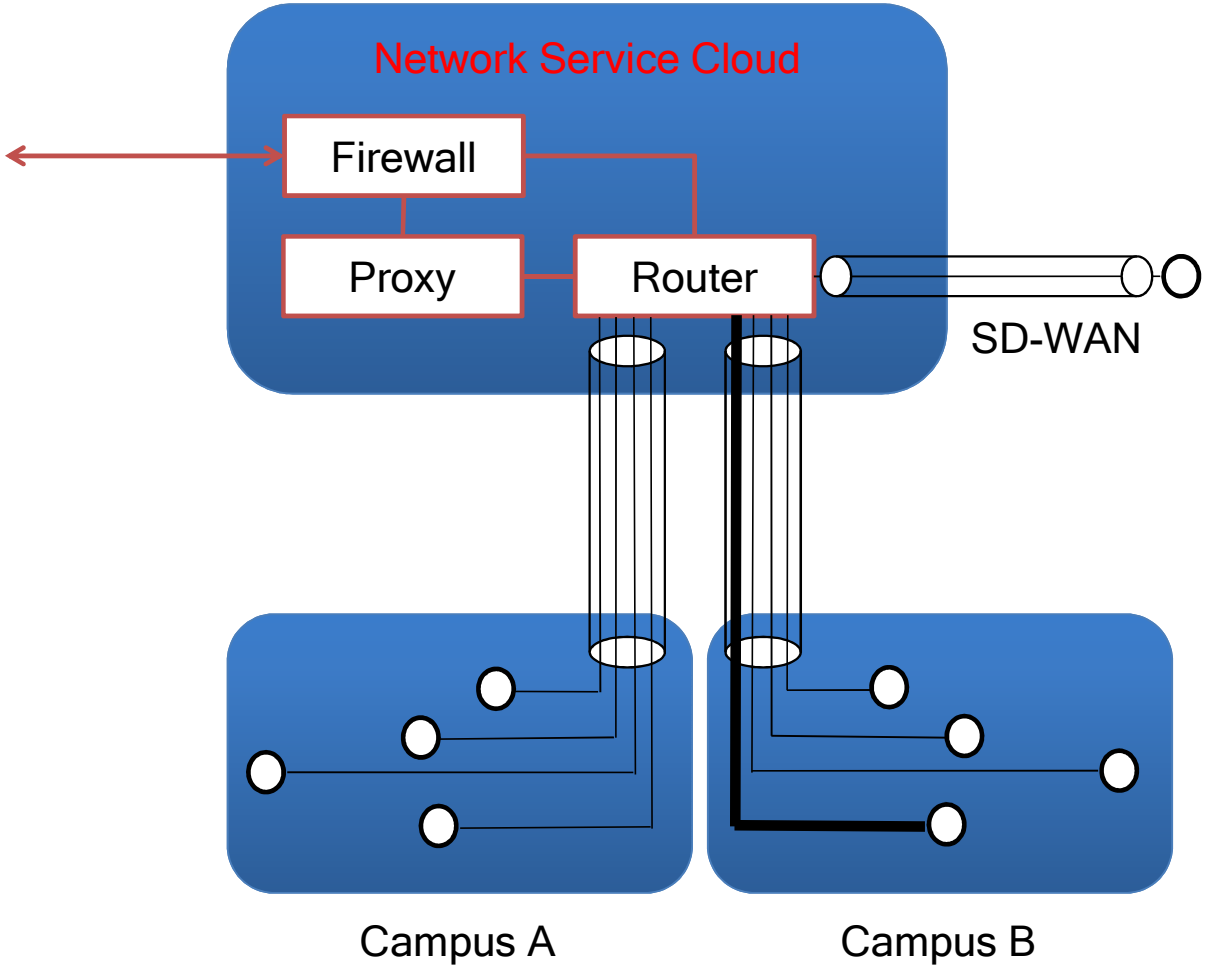
- The current technologies can't scale to the increasing traffic in future.
- 3-4 digit energy saving is necessary, which means we need a new paradigm.

Network extends to Cloud

- Networks (especially organizational networks such as companies and universities campus networks) provides many functions
  - Routing, Firewall, Access control, Redundant services etc.
- As the available bandwidth of the internet increases, functions can be moved to cloud
  - Use of network function virtualization (NFV)

Network Service Cloud (NSC)

- Network functions (router, firewall etc.) implemented as VNF (Virtual Network Function) and built up in Cloud
- Can be connected from anywhere (incl. outside of campuses)
- Campus network can be very simple
  - Just provide physical network connectivity



Advantages of NSC

- Cost efficient
  - Cloud infrastructure shared by many organizations
  - Allocate resources according to demands
  - Centralized operation reduces cost for 7D/24H
- Reliable
  - Failed resources can be replaced with new one dynamically
- Position independent
  - Can be accessed from anywhere
  - Disaster prevention by using data centers at distant places

Challenges for NSC

- Cloud for “network”
  - Current clouds are not designed for network
  - Bandwidth hungry applications require stable guaranteed performance
  - Coordination of “physical” and “virtual” is an issue
    - Including connections between clouds and campuses
- Interoperation with AAA/policy
  - Need to support different kind of members who have different rights/duties
- Manageability
  - Network managers of different organizations should be able to manage NFV functions according to their needs/policies

Cloud extends to network

- Recently, IoT is gathering wide attention
  - Computing is done at the edge, middle of the network and clouds
- Network is not just pipes anymore. It is a manageable resource
  - Software Defined Network (SDN)
- Cloud (esp. IaaS) extends to network and edge

5G, IoT, Cloud

http://www.chinadaily.com.cn/business/tech/2016-02/23/content_23608976.htm

IoT: Edge, Cloud, Fog

https://leanbi.ch/wp-content/uploads/2018/02/Fog_Achitecture.png

5G impact: requirements of 5G

http://www.3glteinfo.com/when-5g-coming/

https://ars.els-cdn.com/content/image/1-s2.0-S2352864817301335-gr3.jpg

Wide area network challenges

- Coordination of many different kinds of resources
  - Network (SDN)
  - Edge devices, Fog devices (cf. MEC (Mobile Edge Computing) Servers)
  - Cloud resources, VNF
  - Services
  - IoT support
    - Huge number of edge devices
    - In and outside campuses
- Computing and communication should be more tightly coupled

Data Centers : Current

- General purpose CPUs and storages are connected by a network such as Ethernet
  - Almost homogeneous.
  - Recently, GPUs and FPGAs are being introduced as co-processors of servers.
- Expect to stay in the main stream for the next 5 years
- Cost of data movement between servers is high
  - Near data processing: avoid moving data as much as possible

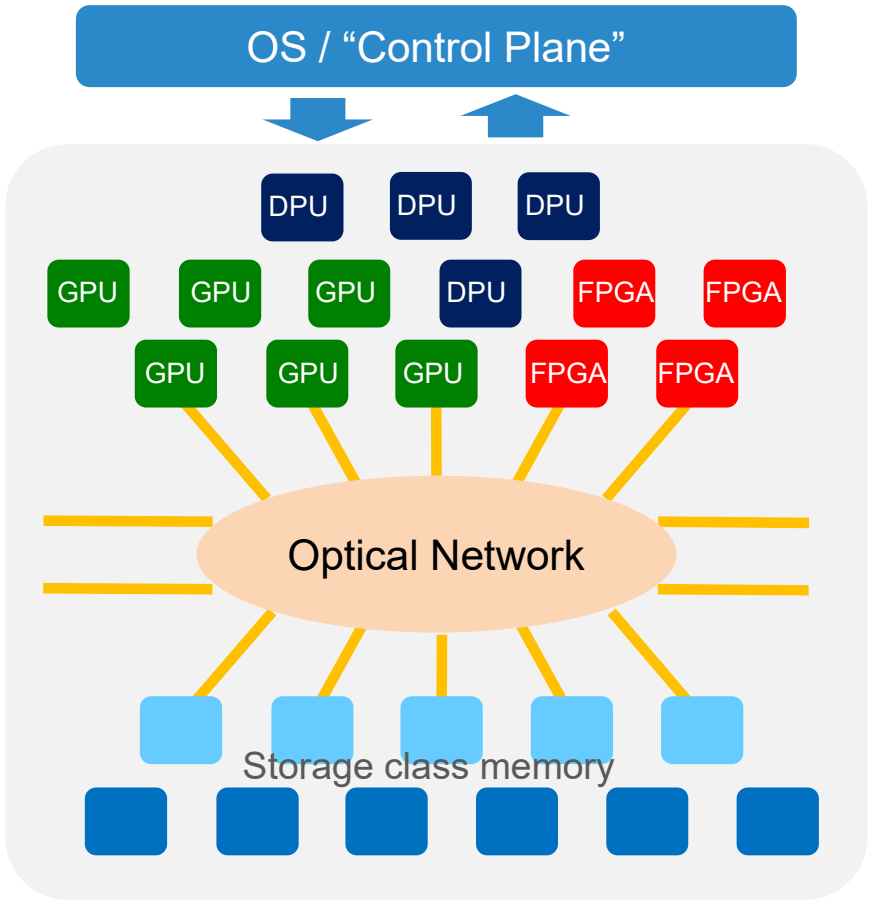
Data Center Challenges after 2023

- End of Moore's law
  - No more performance advances by shrinking transistors
  - Need new architecture for data centers
- Emerging new applications
  - Data processing, esp. machine learning
  - More operations per clock/transistor
    - Neuromorphic, analog, quantum computing etc.
    - Need to move more data to/from processing engines

Control flow to Data flow centric

- Data flow of emerging applications (machine learning, big data) are rather simple
- Circuit switched network will be sufficient for such applications

Ideal Heterogeneous System



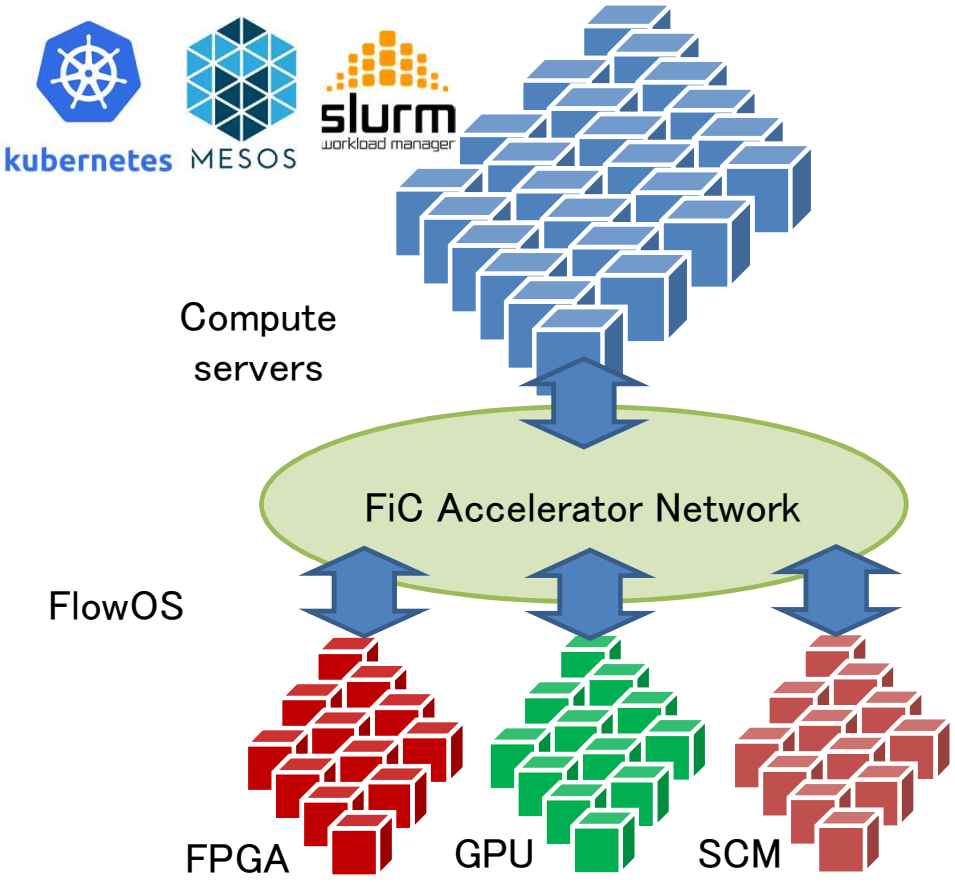
1. For resource utilization efficiency, cloud (shared pool of resources) is advantageous for heterogeneous systems

2. To provide performance guaranteed slices to users, bare-metal type slice is required

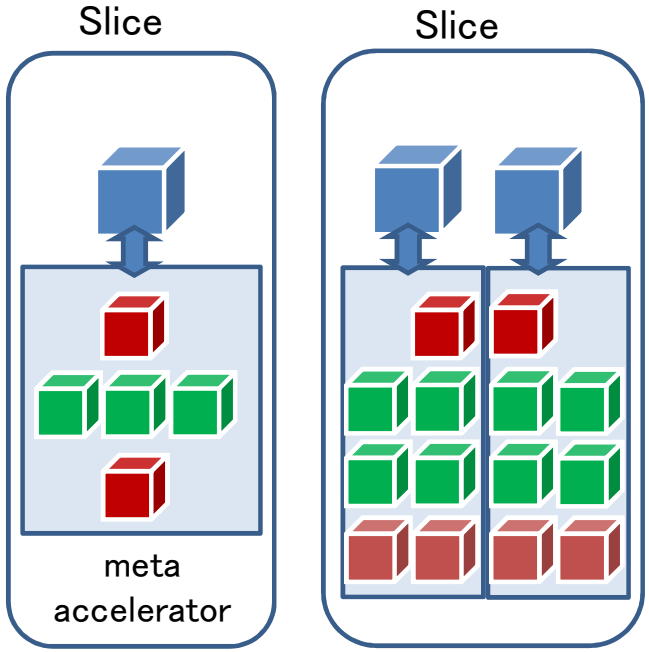
3. Make use of ultra-wide bandwidth interconnection

Curtesy of R. Takano (AIST)

Flow-in-Cloud (FiC): Disaggregated accelerator cloud



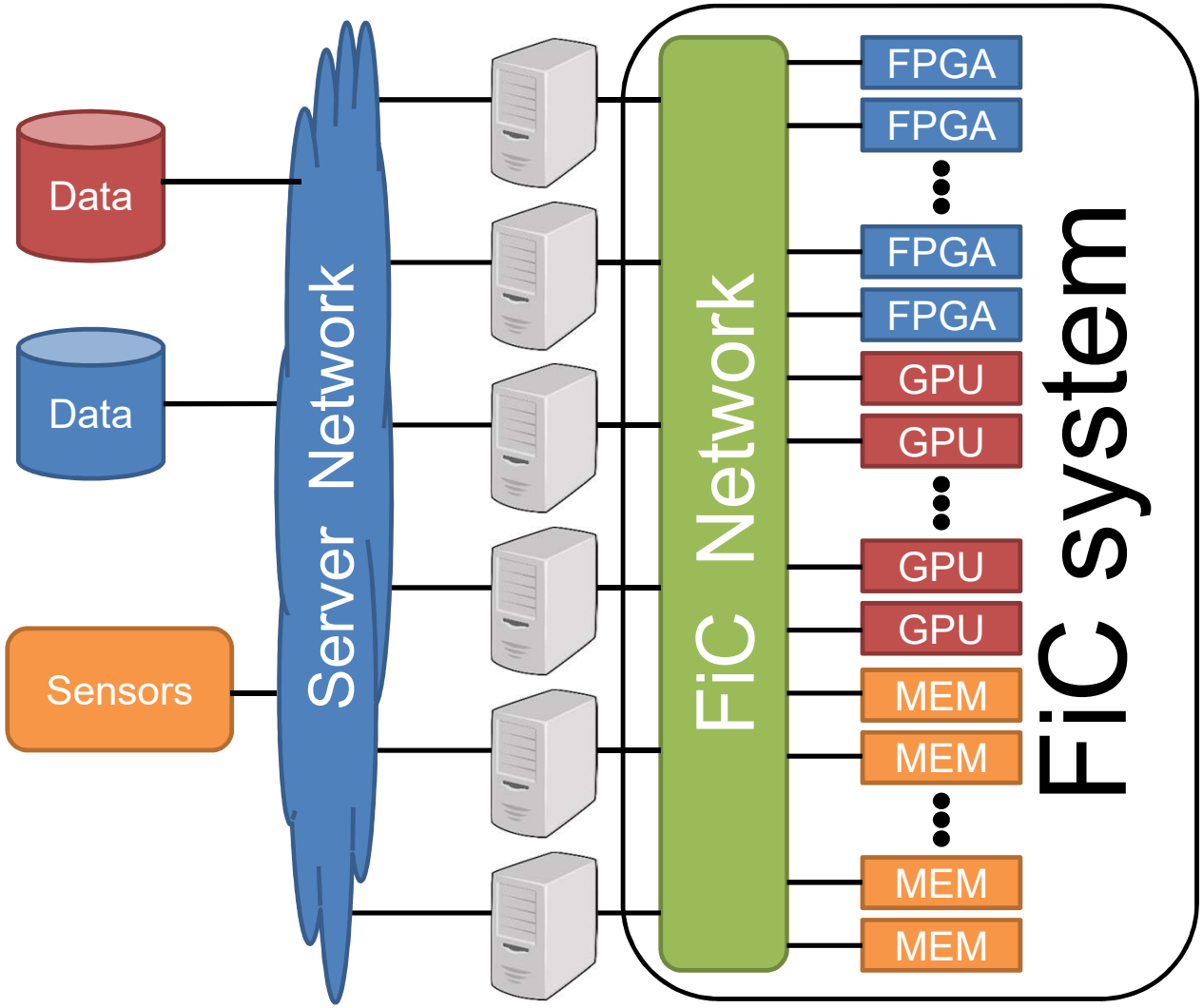
Resource pool of "engines"



Configuring *"meta-accelerators"*
according to the application requirement

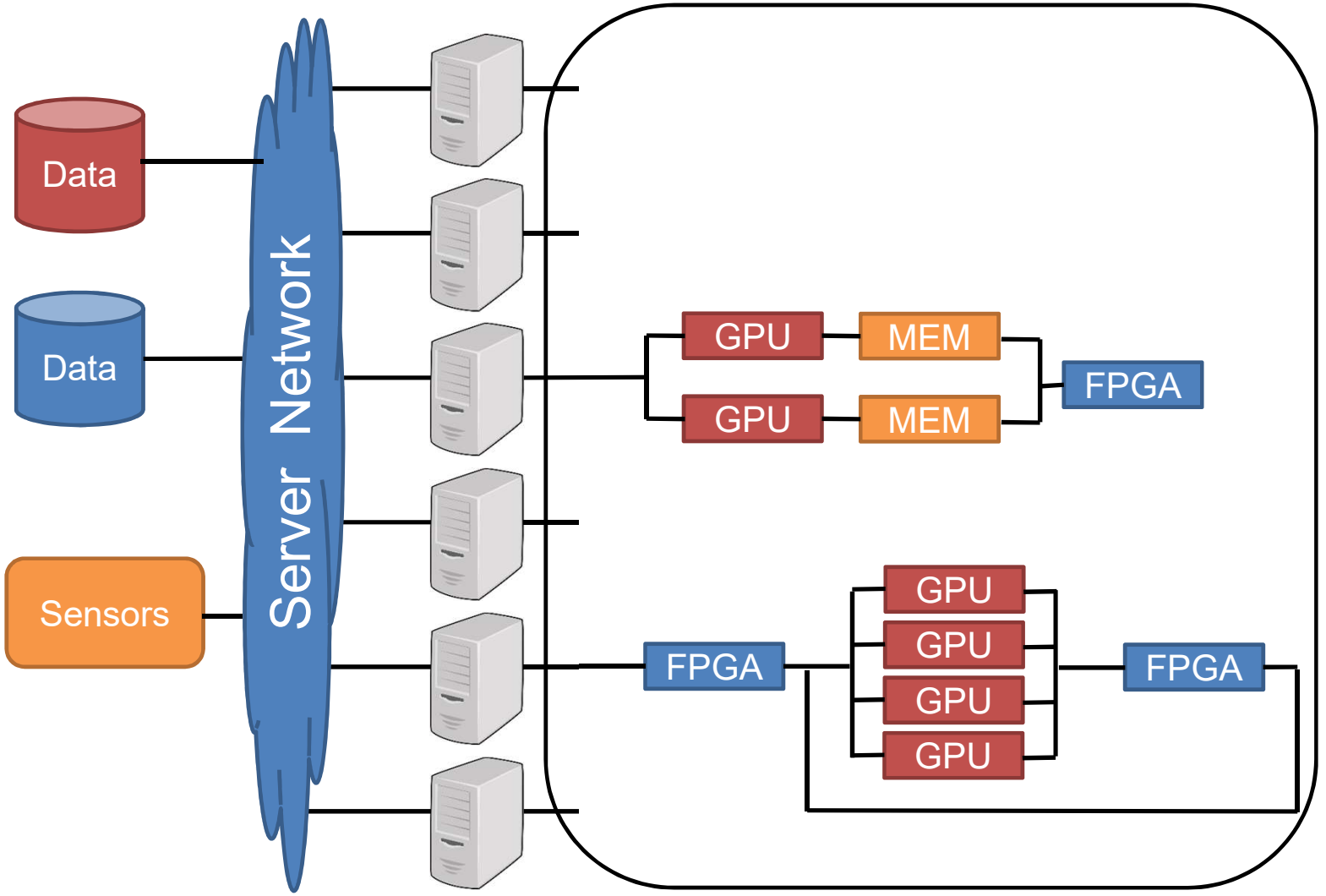
Curtesy of R. Takano (AIST)

Accelerator bare-metal cloud



- For resource utilization efficiency, shared pool of resources is advantageous for heterogeneous systems
- To provide performance guaranteed slices to users, bare-metal type slice is required

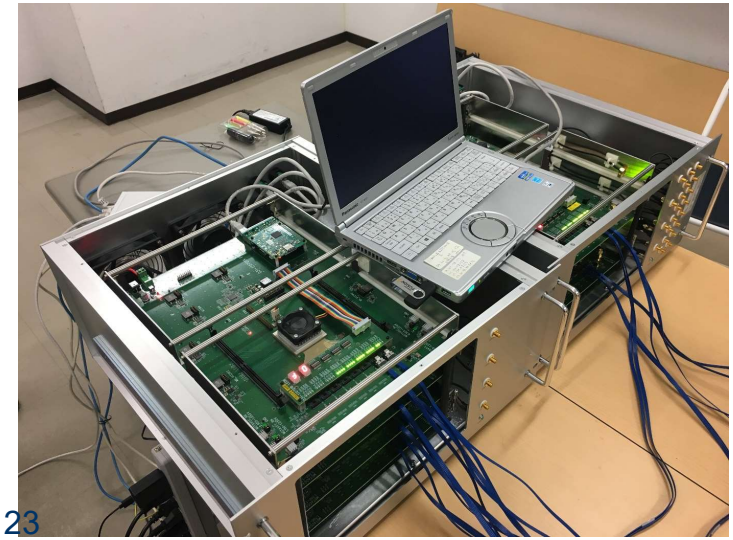
Accelerator bare-metal cloud



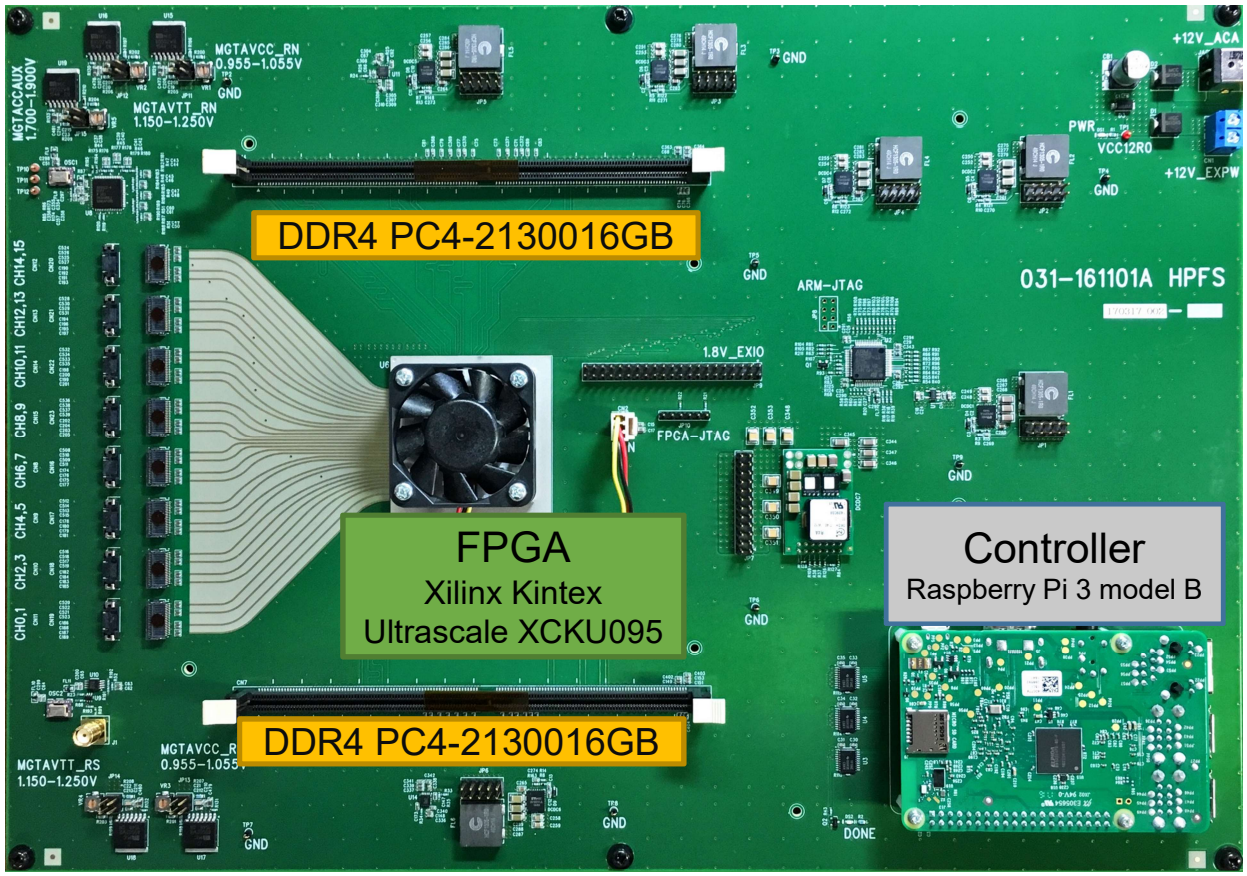
- For resource utilization efficiency, shared pool of resources is advantageous for heterogeneous systems
- To provide performance guaranteed slices to users, bare-metal type slice is required

FiC Switch and FPGA Board

- A prototype of FiC switch and FPGA board is working.
- FiC switch provides circuit switching with time-division multiplexing.



8Gbps full dup x 8



Data Movement is Critical

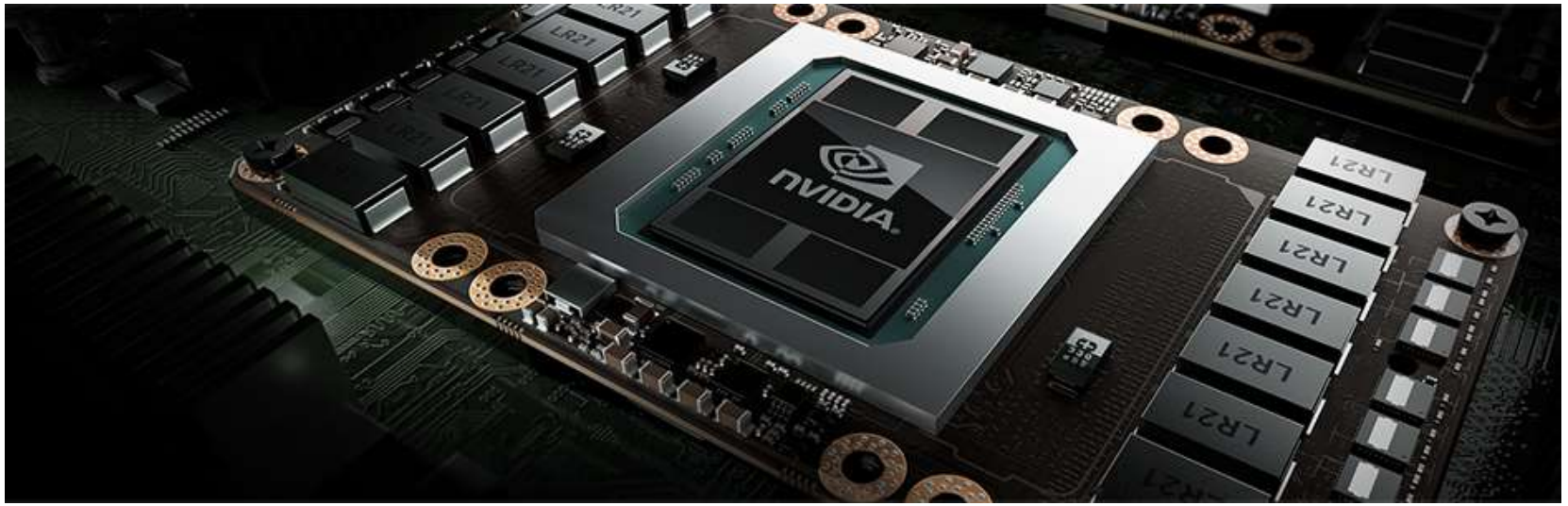
- Data movement is unavoidable on a heterogeneous system and the bottleneck to achieve higher performance
- E.g., ABCI@AIST
  - 12 GFlops/W ⇔ 83 pJ/Flop
    - GFlops/W = GFlop/second / Joule/second = GFlop/Joule
  - Energy per bit budget: 0.42 pJ/bit
    - 200 bits/Flop
- Scaling performance is getting harder under tight energy budget

Data Movement Energy

Access to SRAM	O(10 fJ/bit)
Access to DRAM	O(1 pJ/bit)
Movement to HBM	O(10 pJ/bit)
Movement to DDR3 off-chip	O(100 pJ/bit)

Curtesy of R. Takano (AIST)

System in a Package (SiP)



NVIDIA Tesla P100 https://www.nvidia.co.jp/object/tesla-p100-jp.html

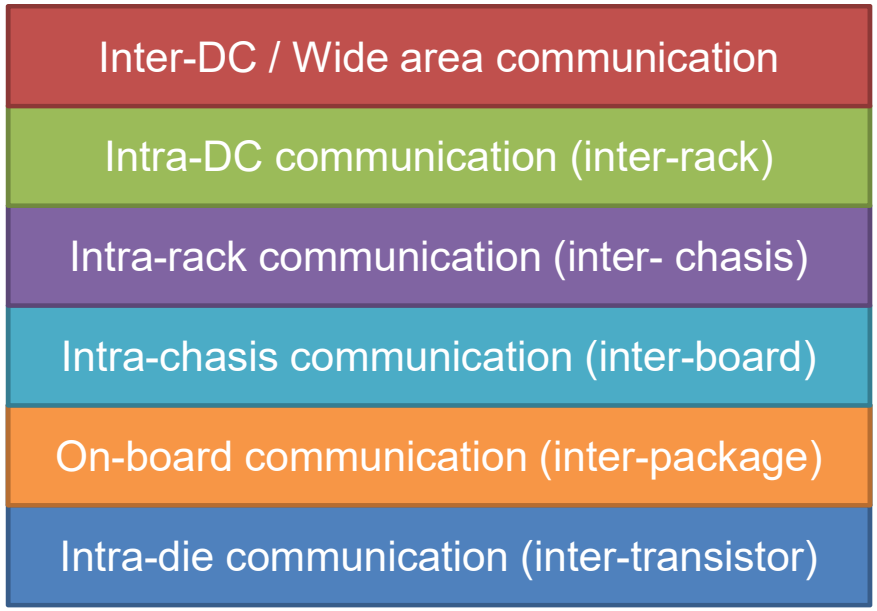
- SoC and HBM2 memory are put on top of a silicon interposer
- Avoiding pin bottleneck of the package

HBM and GPU integration on an interposer

- Silicon interposer is used to mitigate the bin-bottleneck problem of packages.
- Intel announced EMIB which uses smaller silicon to interconnect dies

https://japan.cnet.com/article/35069635/

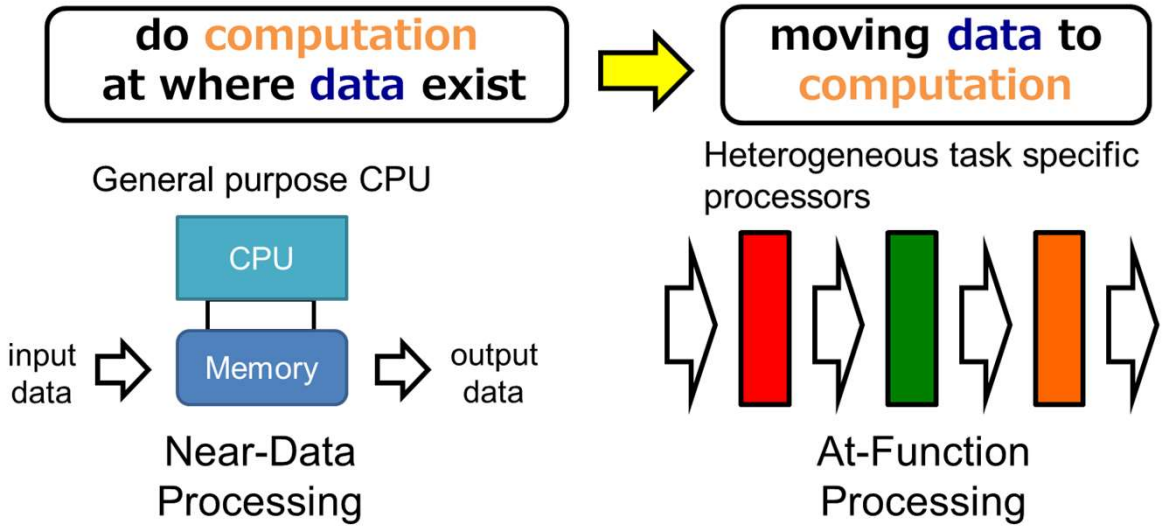
SiP introduced new “layer” in communication



- Realizable bandwidth and power consumption differs layer by layer

Extremely wide bandwidth can be a breakthrough

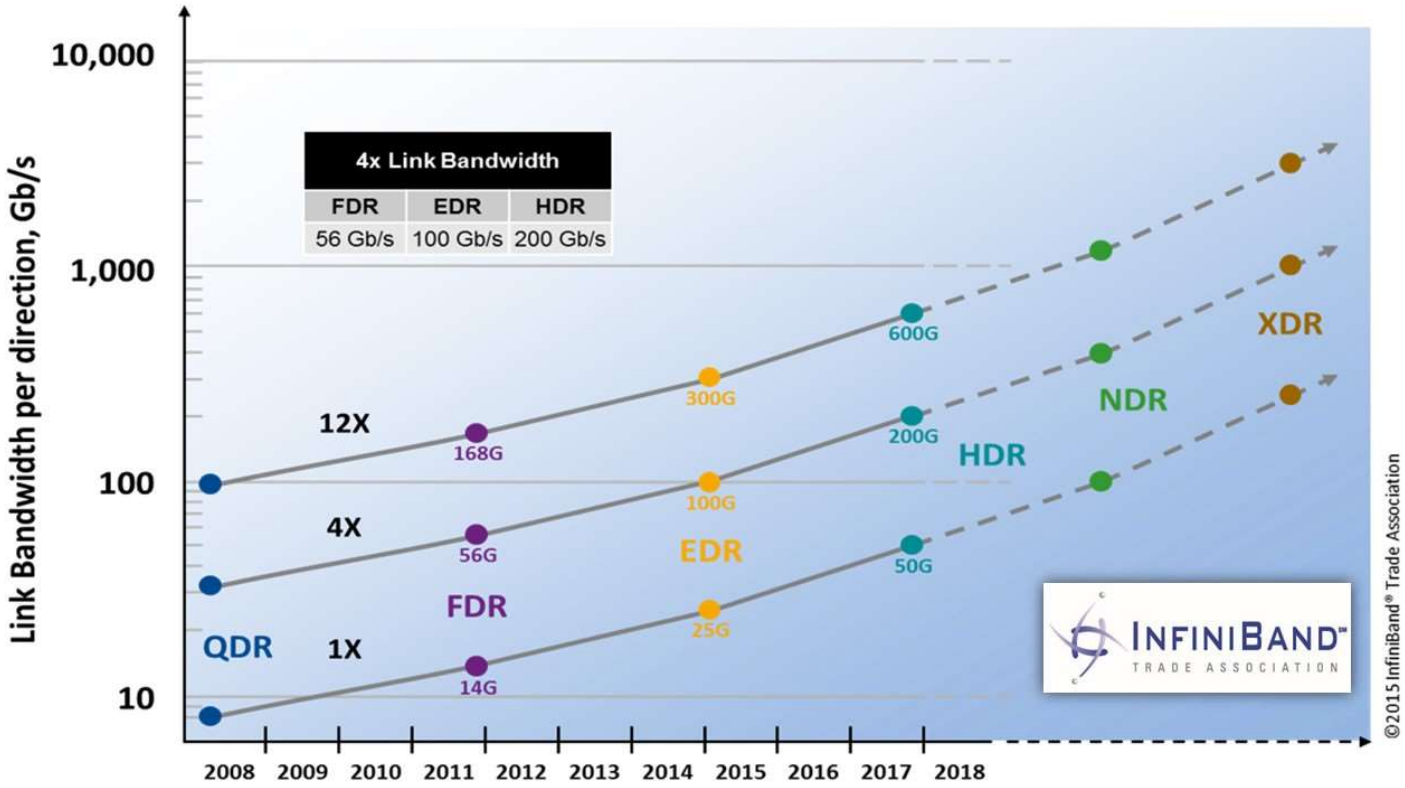
- Take advantage of heterogeneous computing to improve performance after the end of the Moore's law
  - Use of "engines" suitable for special purposes
- At function processing
  - Data have to be moved between engines
- Bandwidth is the key



Communication bandwidth larger than DRAM's will change the whole scenario of computing

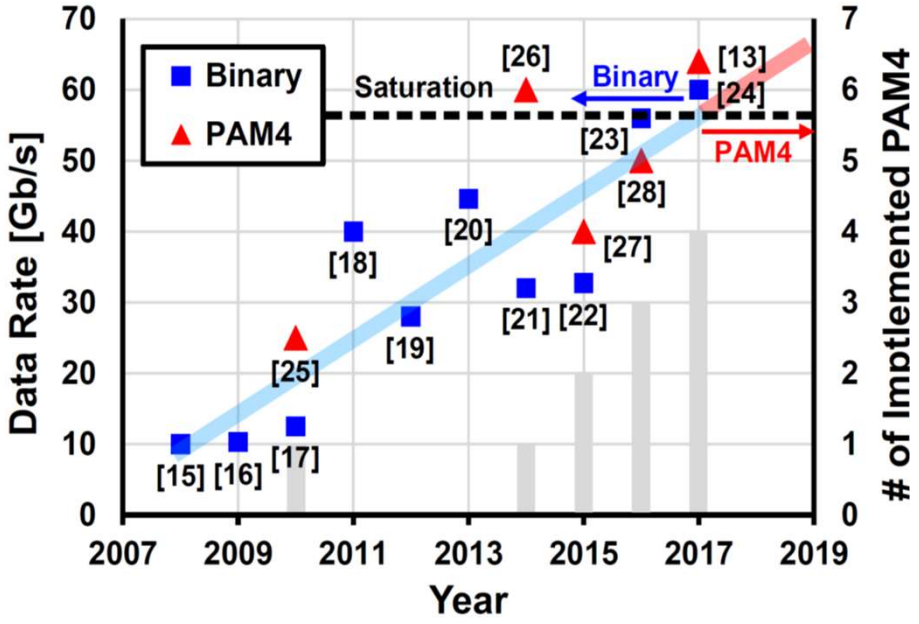
Inter-Server interconnect BW trends

- Ethernet is widely used
- InifiniBand for HPC
  - Mellanox Oligopoly
- Intel OmniPath



InfiniBand Roadmap (http://www.inifnibandta.org)

Trends of electrical links



Trends of copper-based electrical links from recent 10-year ISSCC papers.

Jeong, G.-S.; Bae, W.; Jeong, D.-K. Review of CMOS Integrated Circuit Technologies for High-Speed Photo-Detection. Sensors 2017, 17, 1962. より引用

- 50Gbpsis used in InfiniBand(HDR)
- 6~7times /10 years improvement
- Short range only for wide bandwidth
  - Mellanox's InfiniBand HDR copper AOC: up to 3 meters)

Limit of electrical I/O and possibility of Optical I/O

- Limits of bandwidth of electrical I/O
  - Currently, 50Gbps/ch, slow performance improvement
  - Not all pins of a package can be used for high speed I/O
- New packaging technologies :MCP/2.5D/3D
  - SiP (System in Package)
    - Multiple chips are integrated to a single package
    - Bandwidth between chips can be improved
    - Limitation in size. For example, integratable memory size is limited.

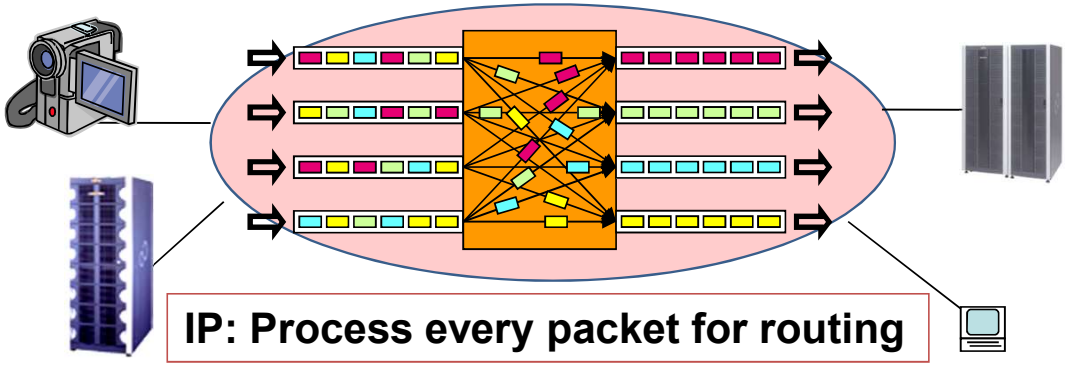
Requirements for future DC networks

- Moving bulk-data is important
  - In the “At-Function Processing”, data should be moved between processing “engines”
- Data movement performance and cost comparable to DRAM access should be realized
  - >5Tbps end-to-end bandwidth for bulk data transfer
    - DRAM bandwidth :HBM2 (3D stack DRAM): 1TBytes/s (4 stacks)
  - <10pJ/bit end-to-end power consumption
    - DRAM power consumption : DRAM: 5pJ/bit or more

More bandwidth for inter-acc. network

- OCS (Optical Circuit Switching)
  - Electrical switching requires O/E/O conversion of all the data → Switching of multiple 10s of Tbps links is unrealistic
- Hybrid use of EPS and F-OCS (Fast-OCS)
  - EPS for short and low bandwidth flows
  - F-OCS for long lasting wide bandwidth data flows
- Take advantage of future data centers' traffic profile
  - Machine learning dominates the data center load
  - Control flow is combined with data flow
    - Relatively static bulk data communication patterns

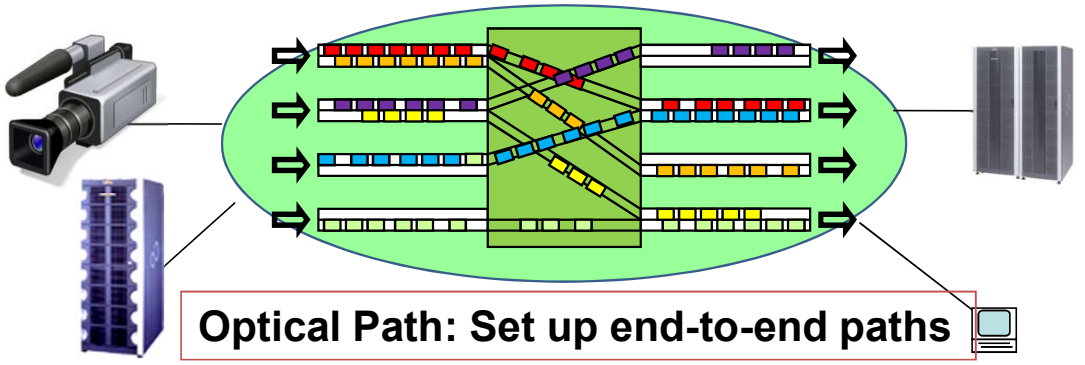
Electrical packet switching and Optical circuit switching



~ 10 nJ/bit

Suitable for small granularity data for different destinations

~ W/port



Suitable for bulk-data transfer with QoS guarantee

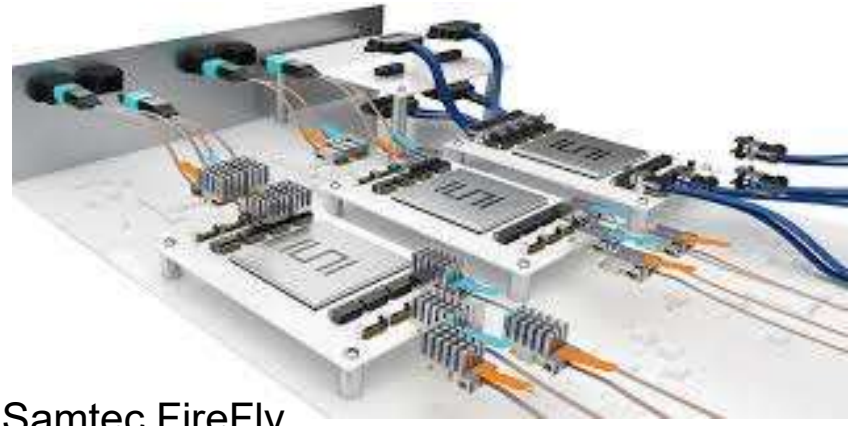
Complementary aspects → The hybrid use!

- EPS
  - Low energy efficiency
  - Good affinity with computing
  - Good for mice flows
- F-OCS
  - High energy efficiency
  - Control plane needed
  - Good for elephant flows

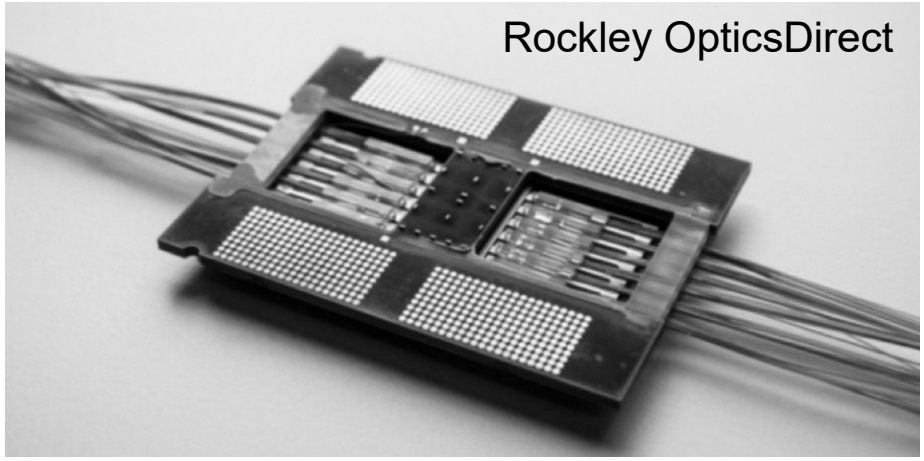
Hybrid network for highly energy efficient, flexible switching for all granularity

Current status of optics for interconnects

- OBO(On-Board Optics)
  - COBO (Consortium for OBO): Microsoft, Juniper, Cisco, Broadcom etc.
- Inter Package I/O
  - Integrate OEO devices in a SiP
  - Remove pin BW bottleneck of a package

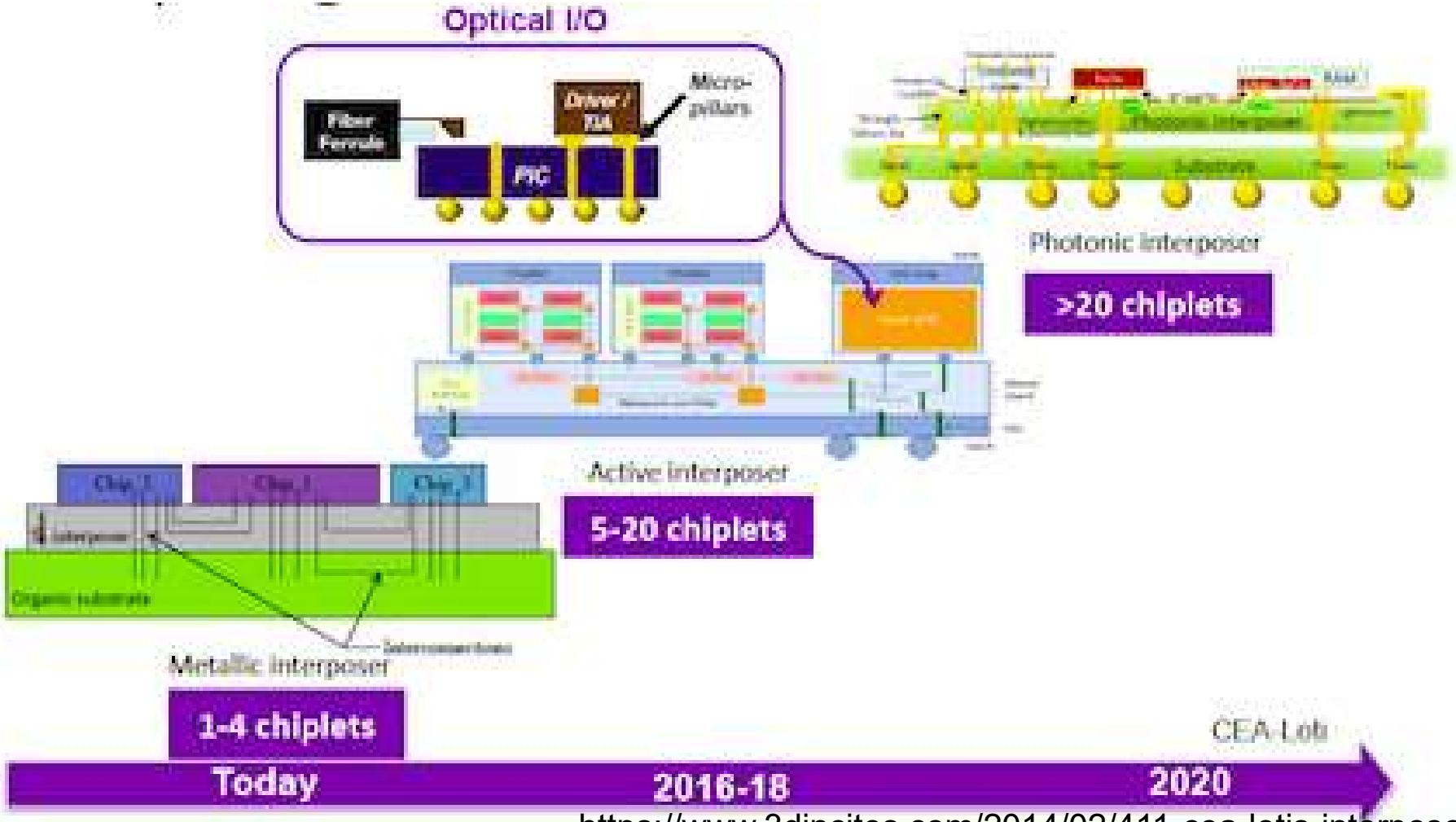


Samtec FireFly



Rockley OpticsDirect

Optical integration in a package



https://www.3dincites.com/2014/02/411-cea-letis-interposer-roadmap/

Optical Switch Technologies

Space optics

MEMS or Piezoelectric beam-steering

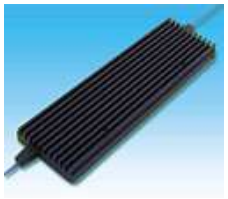
http://www.glimmerglass.com/

http://www.polatis.com/index.asp

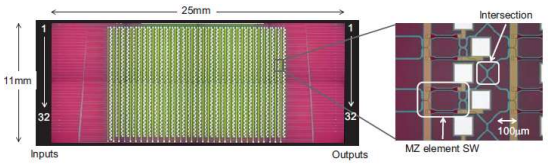
- ☺ Port count: ~ 384x384
- ☹ Physical size: can be large
- ☺ Low insertion loss and crosstalk
- ☹ Switching speed: 10 ms ~

Waveguide

PLC or Silicon Photonics



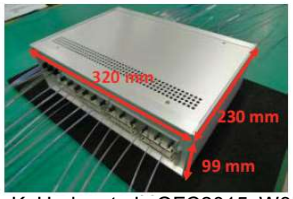
http://www.ntt-electronics.com/index.htm



K. Tanizawa, et al., OFC2015, M2B.5

- 😐 Port count: ~ 32x32
- ☺ Physical size: can be very small
- ☹ Insertion loss, crosstalk, and PDL
- ☺ Switching speed: 3 ms, 30 us

AWG router



K. Ueda, et al., OFC2015, W3D.1

- ☺ Port count: ~ 2000x2000
- ☹ Tunable laser required
- 😐 Tradeoff between the port count and the per port bandwidth

WSS

LCOS or MEMS

https://jp.finisar.com/

- 😐 Port count: ~ 1x20
- ☹ Switching speed: 10s ~ 100s ms
- 😐 Insertion loss: ~6dB
- ☺ Wavelength routing

Fast switch

PLZT, SOA, LN, DLP WSS…

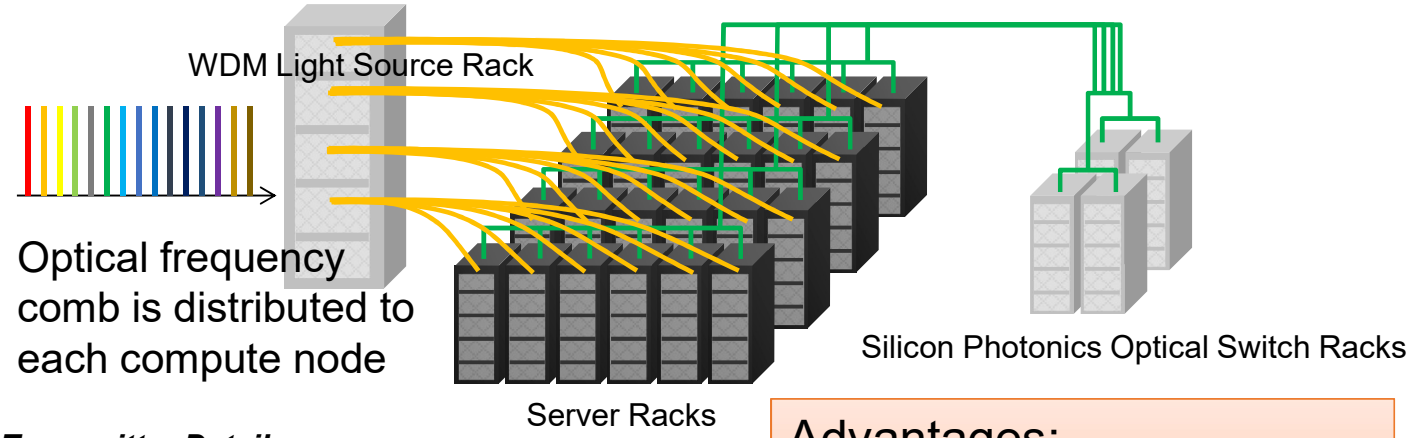
- ☺ Switching speed: 10 ns ~
- ☹ Port count: ~ 4x4
- ☹ Insertion loss, crosstalk, and PDL

Curtesy of K. Ishii (AIST)

Ultra-wide Bandwidth Optical Datacenter Network

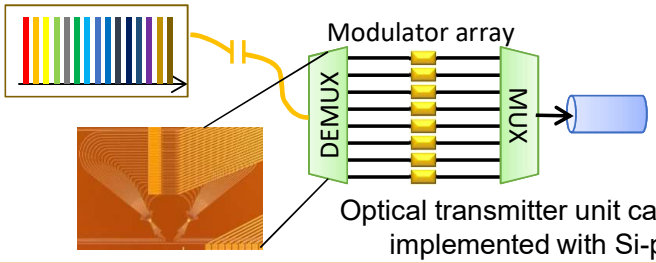
- **Wavelength Bank (WB):**
  - ✓ Single DWDM light source in a system: Distributed to computing nodes via optical amplifiers
  - ✓ No light source is required at each computing node: low cost, low power

- **Silicon photonics optical circuit at each node**
  - ✓ De-multiplex, modulate, multiplex and transmit
  - ✓ Enables hybrid implementation with electronics



Target in 2025
Bandwidth: 10 Tbps / fiber
Energy: 1 ~ 10pJ/bit

Transmitter Details



Optical transmitter unit can be monolithically implemented with Si-ph technologies

Advantages:
- No hermetic seal
- No thermoelectric cooler
- High frequency accuracy

# of wavelength	# of levels / ch	Bandwidth / fiber
1	1	20 Gbps
4	8	640 Gbps
32	8	5.12 Tbps

Courtesy of S. Namiki (AIST)

Conclusion

- Integration of computing and communication is becoming more and more important
- We have to consider power consumption and performance balance of computing and communication
- There are many layers in communication, and performance and power consumption differ layer by layer
- Optical communication can bring a breakthrough in IT infrastructure