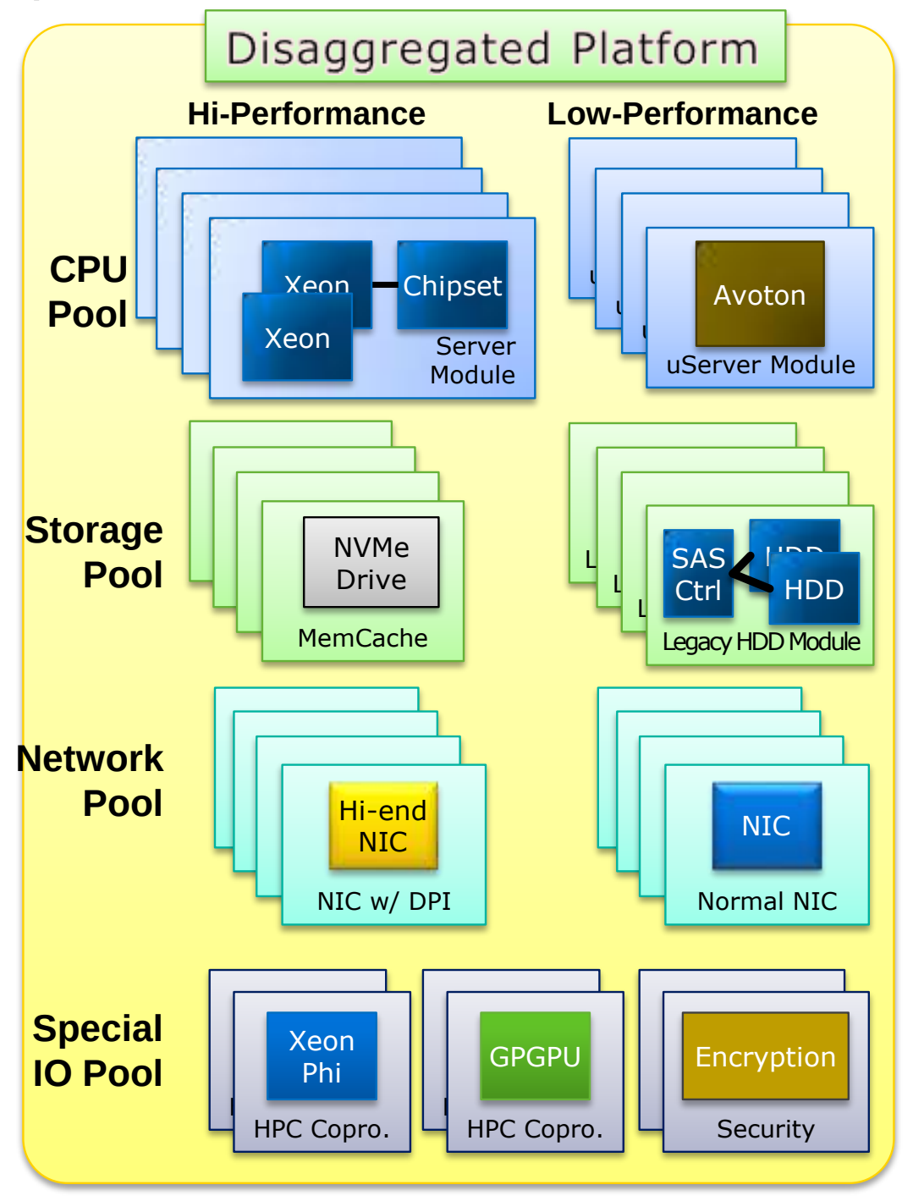
Bridge Chip Composing a PCIe Switch over Ethernet to Make a Seamless Disaggregated Computer in Data-Center Scale

Takashi Yoshikawa[1], Jun Suzuki[1], Yoichi Hidaka[2], Junichi Higuchi[2], and Shinji Abe[3] [1]Green Platform Res. Labs, [2]System Device Division, [3]IT Platform Division, NEC

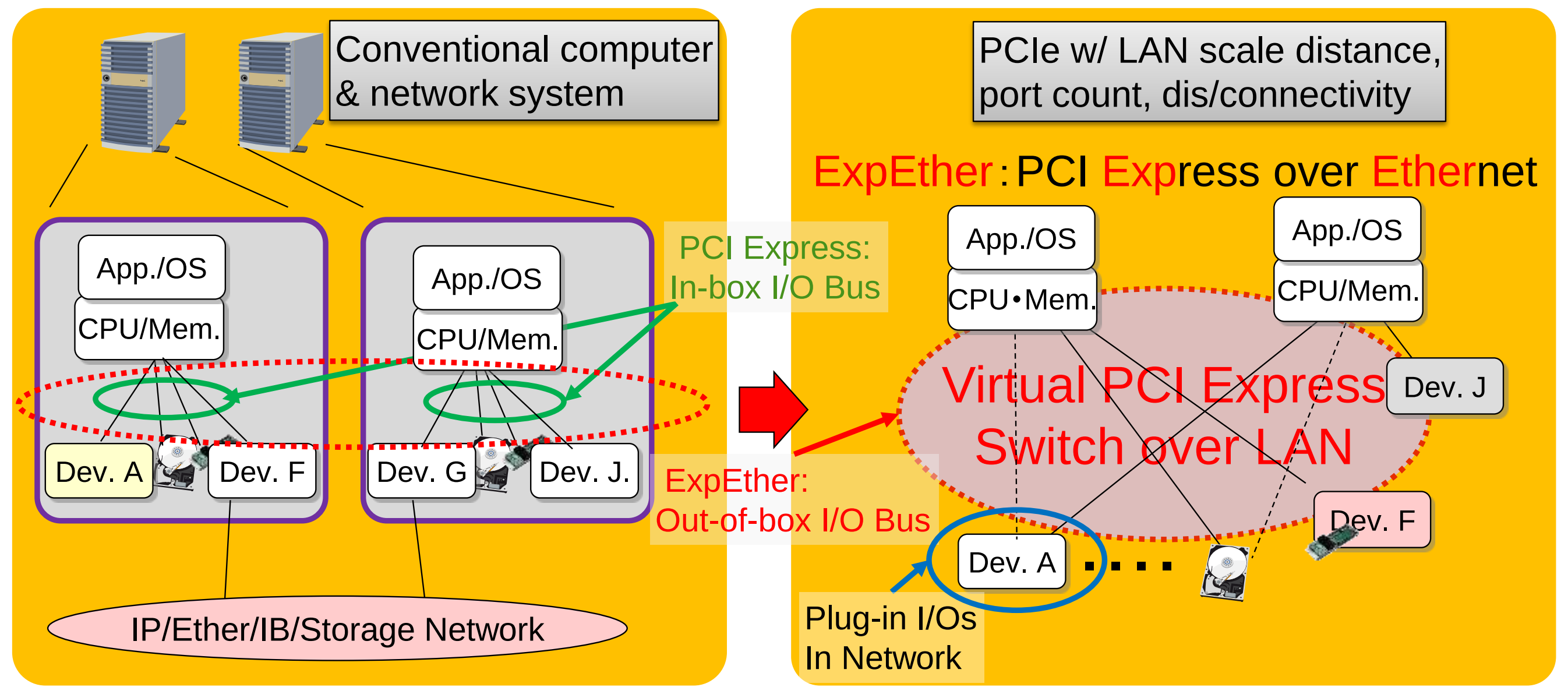
Problem Statement

- Cloud data center must provide computer platforms along with diversified user requests quickly, in reasonable price
  - One of the solutions is a “disaggregated platform” which has resource pools of CPUs, storages, GPUs, and other devices.
- PCIe is appropriate interconnection
  - Simple protocol, vast hard/soft assets.
- However, the transport layer of PCIe doesn’t reach DC-scale.
  - Connection distance and port count are limited to in-box scale.
  - Coupling of host and I/O devices are too tight to disaggregate them into shared resource pool.



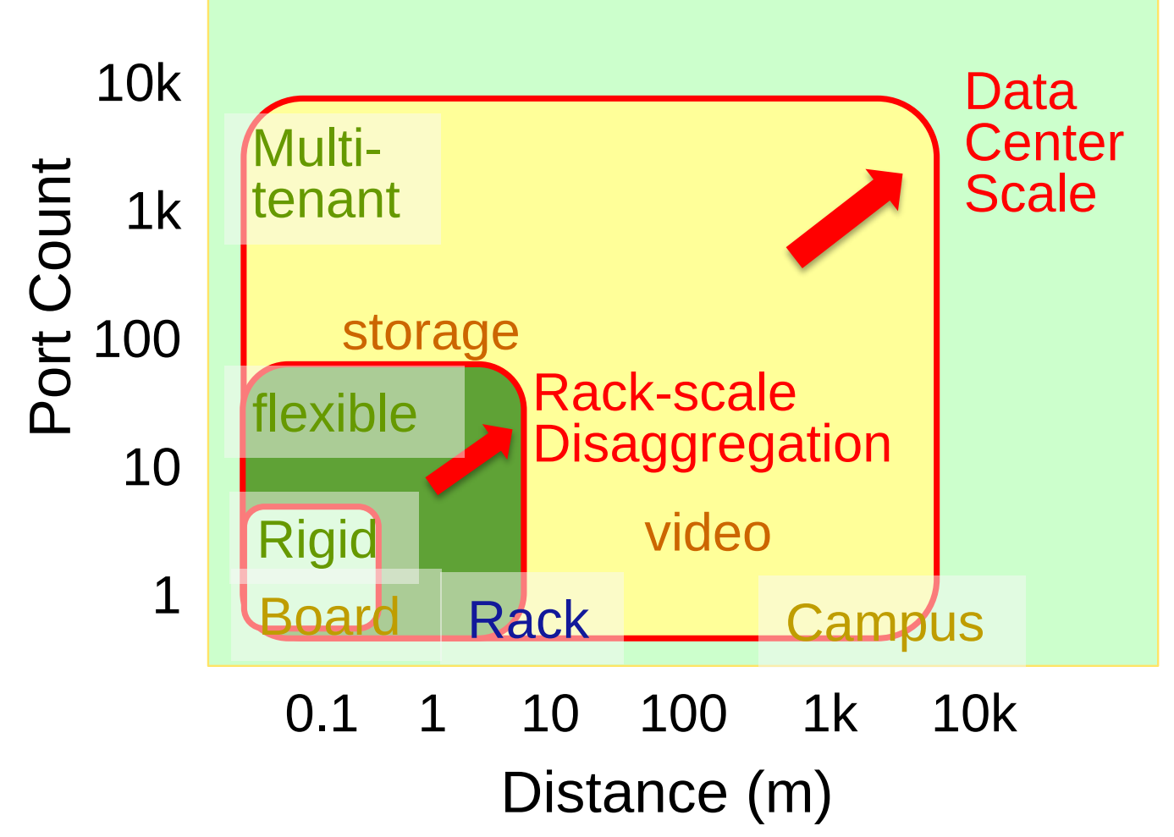
Proposal

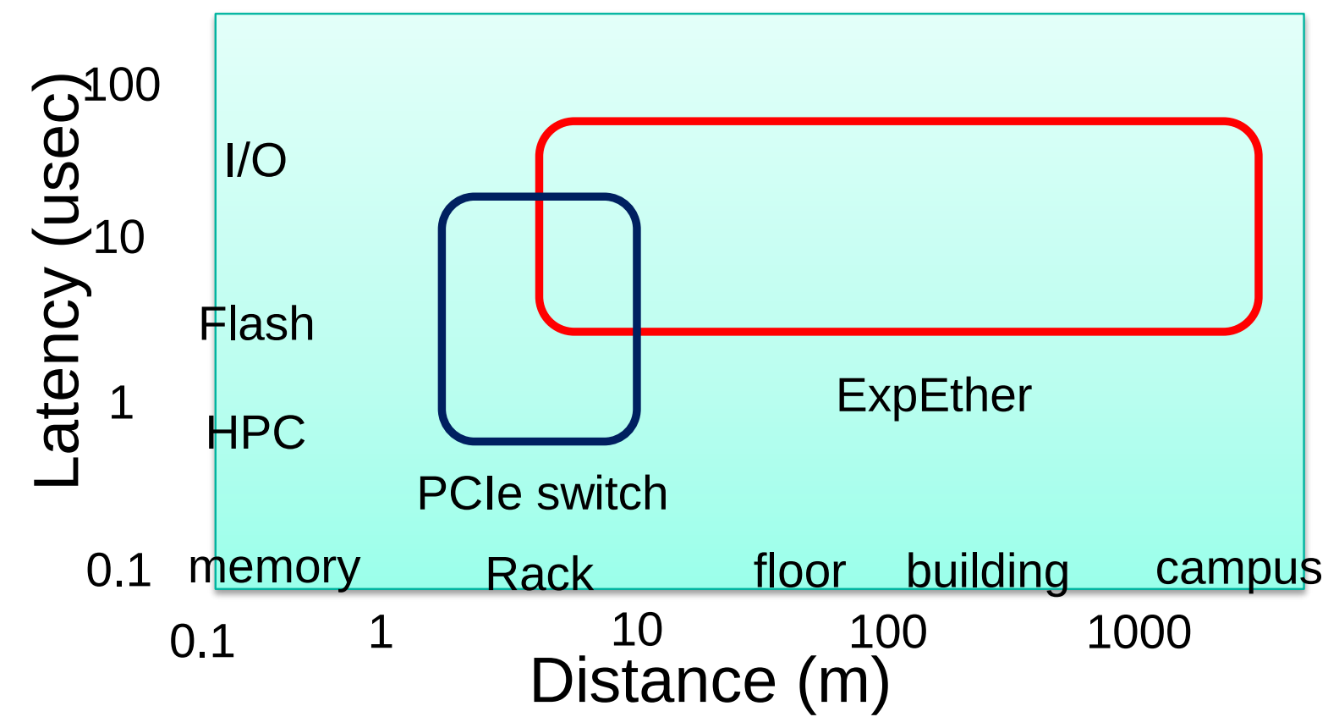
- Using Ethernet as a transport of PCI Express to realize 100-times port-count and connection-distance
  - Seamless expansion of PCIe to realize a big single computer with many CPU/IO resources on a single network.
  - Software defined configuration to realize function/performance along with user’s diversified requirements.


Conventional computer & network system
App./OS
CPU/Mem.
Dev. A
Dev. F
App./OS
CPU/Mem.
Dev. G
Dev. J.
PCI Express: In-box I/O Bus
IP/Ether/IB/Storage Network
PCIe w/ LAN scale distance, port count, dis/connectivity
ExpEther : PCI Express over Ethernet
App./OS
CPU•Mem.
App./OS
CPU/Mem.
Virtual PCI Express Switch over LAN
Dev. J
Dev. F
Dev. A
ExpEther: Out-of-box I/O Bus
Plug-in I/Os In Network

Conclusion

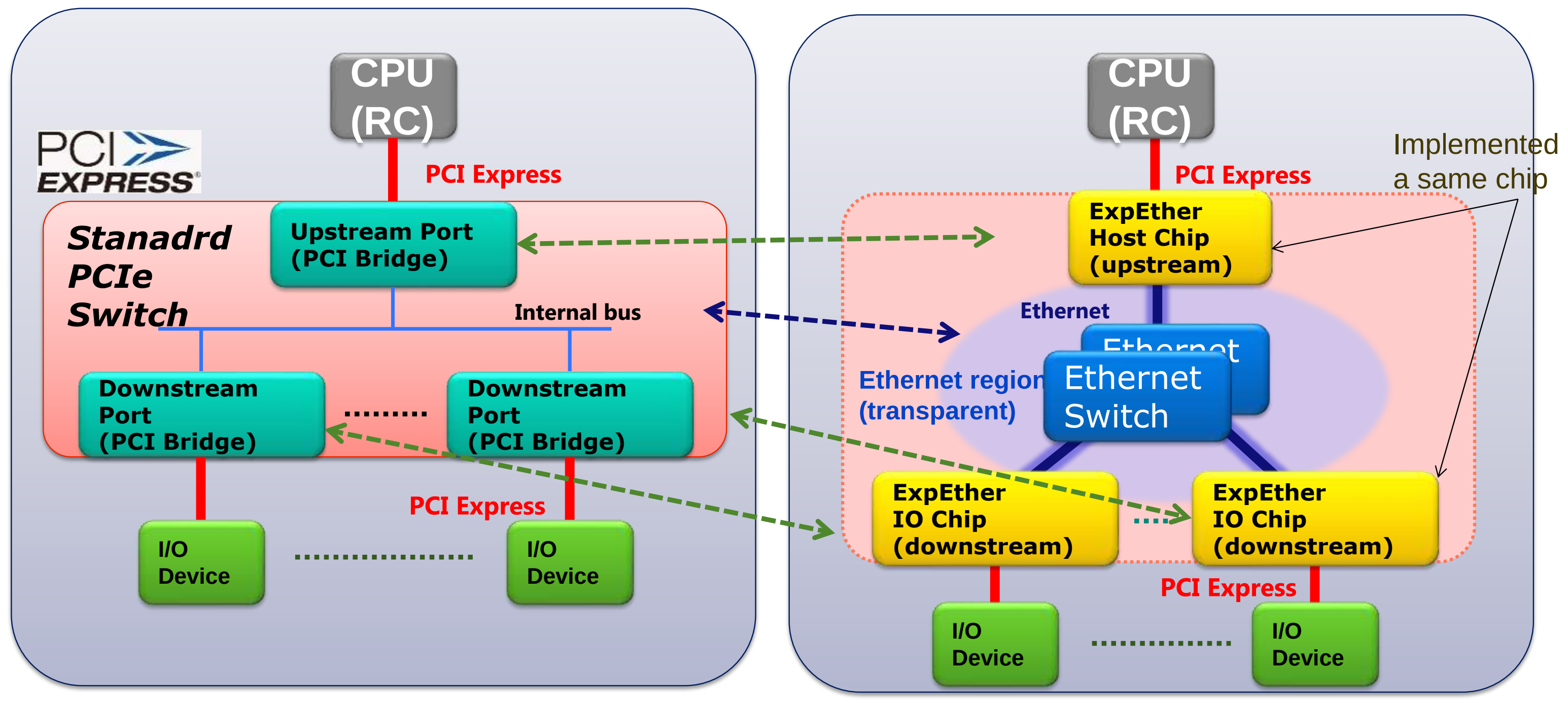
- Distributed PCIe switch architecture and reliable Ethernet realize LAN-scale disaggregated computer w/ resource pool
  - PCIe compliant : Utilize low-cost commodity soft/hardware.
  - Comparable performance with that of direct attached device.
  - Software defined configuration achieving required performance.
  - Technology Proven : Already in service.





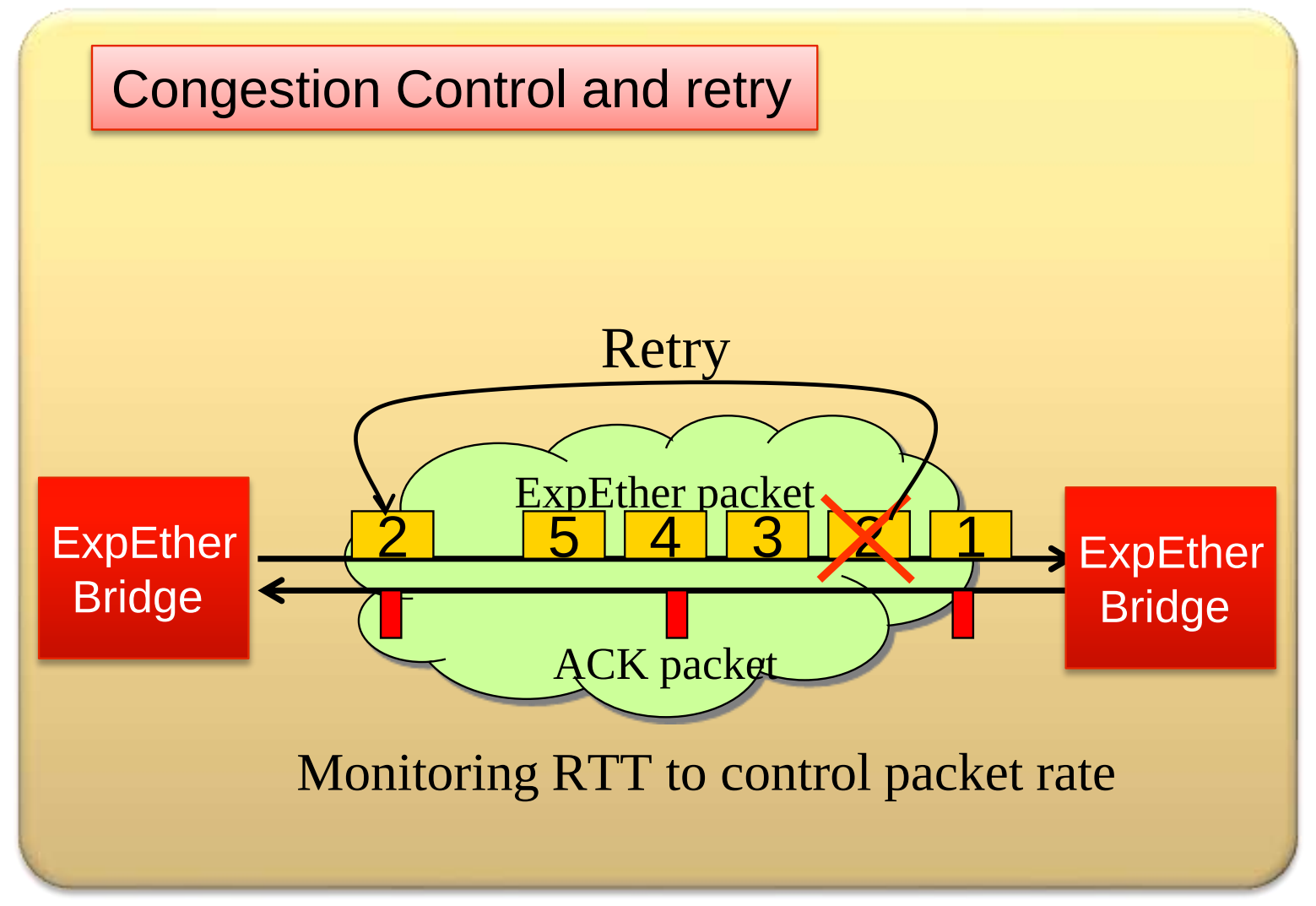
Architecture $^1/_2$: Distributed PCIe switch

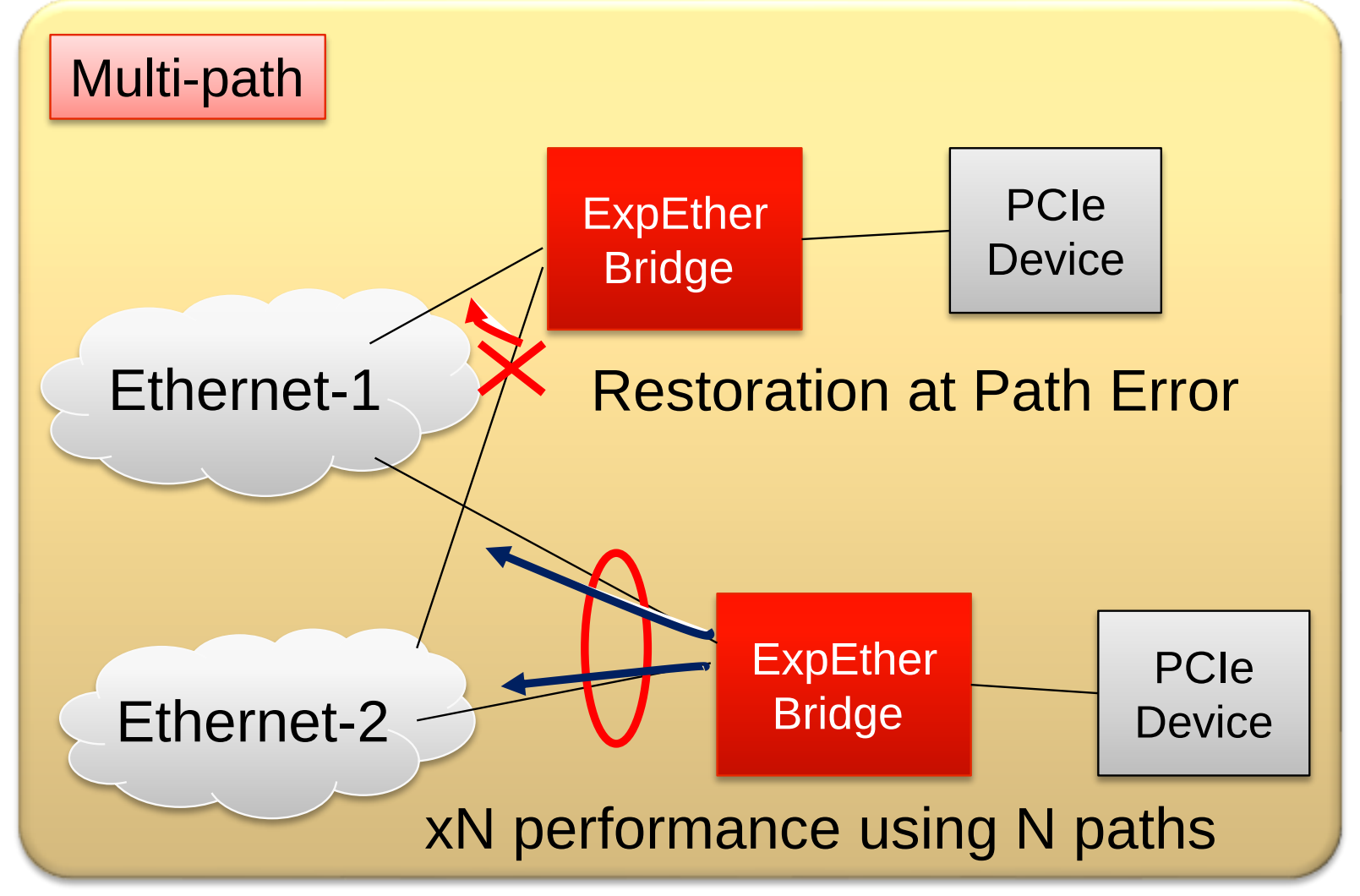
- To be PCIe compliant, Ethernet must be transparent.
  - Internal bus of PCIe switch chip is extended by Ethernet.
  - ExpEther appears as a single hop PCIe switch for OS/software.
  - Utilize commodity device, OS, device driver w/o modification.


PCI EXPRESS
CPU (RC)
PCI Express
Standard PCIe Switch
Upstream Port (PCI Bridge)
Internal bus
Downstream Port (PCI Bridge)
Downstream Port (PCI Bridge)
PCI Express
I/O Device
I/O Device
CPU (RC)
PCI Express
Implemented in a same chip
ExpEther Host Chip (upstream)
Ethernet
Ethernet region (transparent)
Ethernet Switch
ExpEther IO Chip (downstream)
ExpEther IO Chip (downstream)
PCI Express
I/O Device
I/O Device

Architecture $^2/_2$: Reliable Ethernet

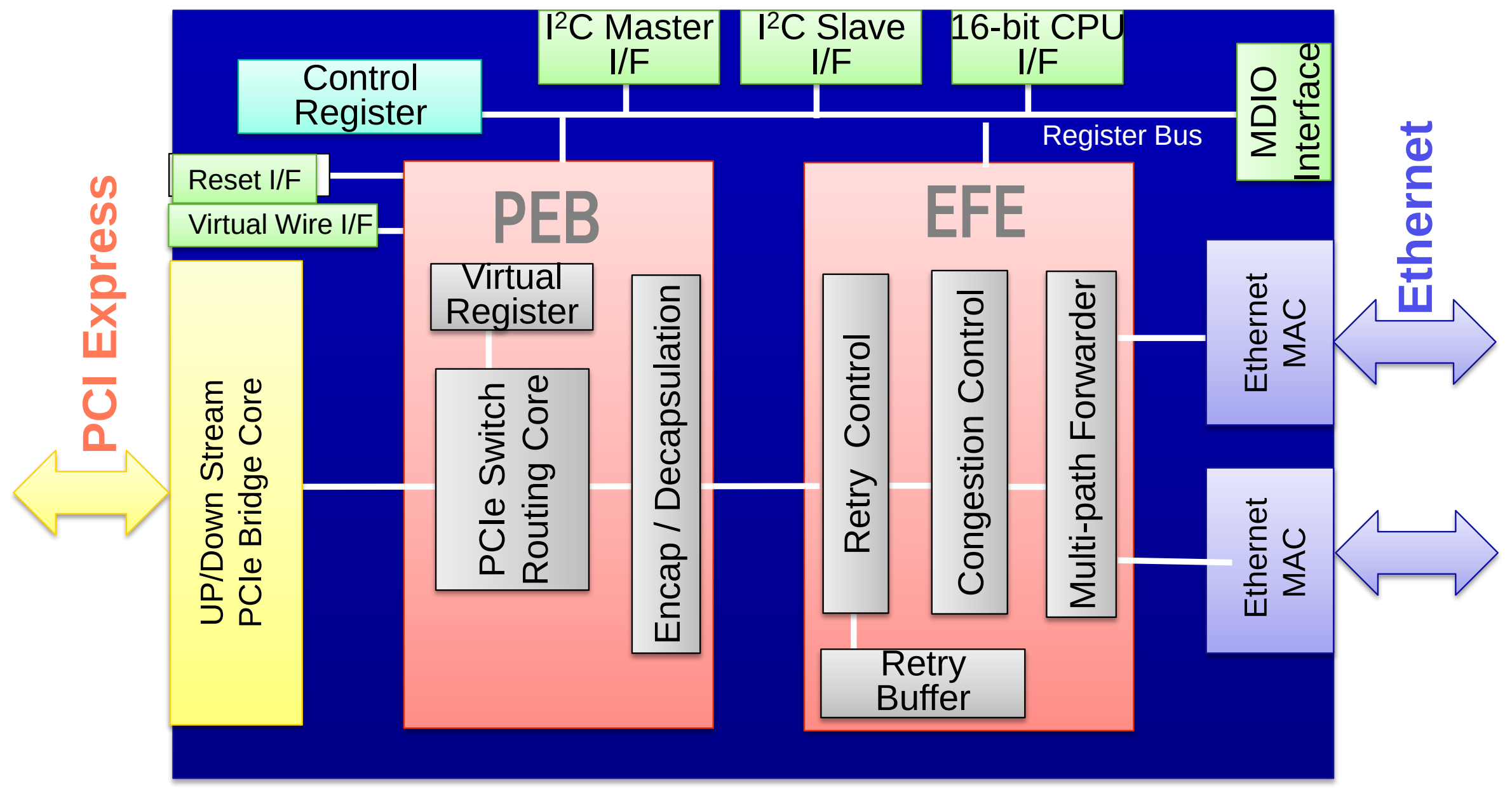
- To maintain PCIe connection, Ethernet must be reliable.
  - Prevent packet loss by congestion control and retry
  - xN bandwidth and redundancy by multipath transport
  - Utilize conventional Ethernet switch w/o modification





Implementation

- To have comparable performance with PCIe switch, all functions are implemented in a chip w/o S/W stack.
  - Main functional blocks are PCIe Ethernet Bridge (PEB) and Ethernet Forwarding Engine (EFE).
  - External interface are standard; PCI Express and Ethernet.



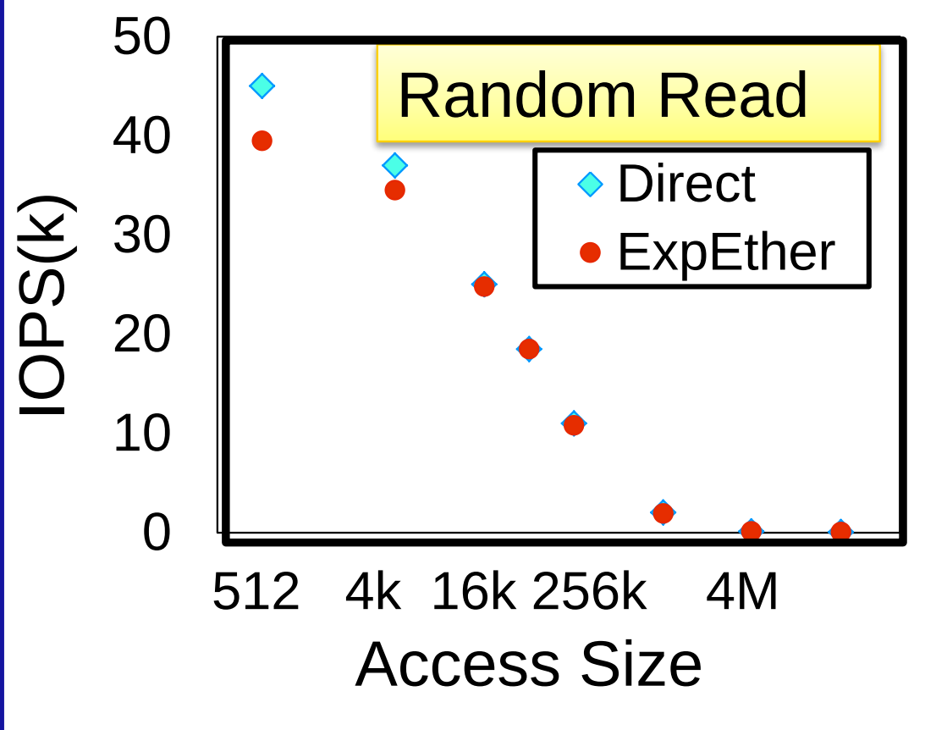
Detailed Specification of chips

chip	ASIC	FPGA	Structured ASIC
PCIe	Gen2.1 1 lane @ 2 Gbps	Gen2.1 8 lane @2 Gbps	Gen2.1 8 lane @2 Gbps
Ethernet	2 x GbE	2 x 10GE	2 x 10GE
Power(RMS)	1.0W	6.7 W	3.5 W
Dimension	19 x 19mm	33 x 33 mm	29 x 29 mm
Pin Count w/ Power/GND	388 pin	780 pin	780 pin
Process	130 nm	40 nm	40 nm
Latency	1050 ns	700 ns	700 ns
Internal Clock	125 MHz	125/156 MHz	125/156 MHz

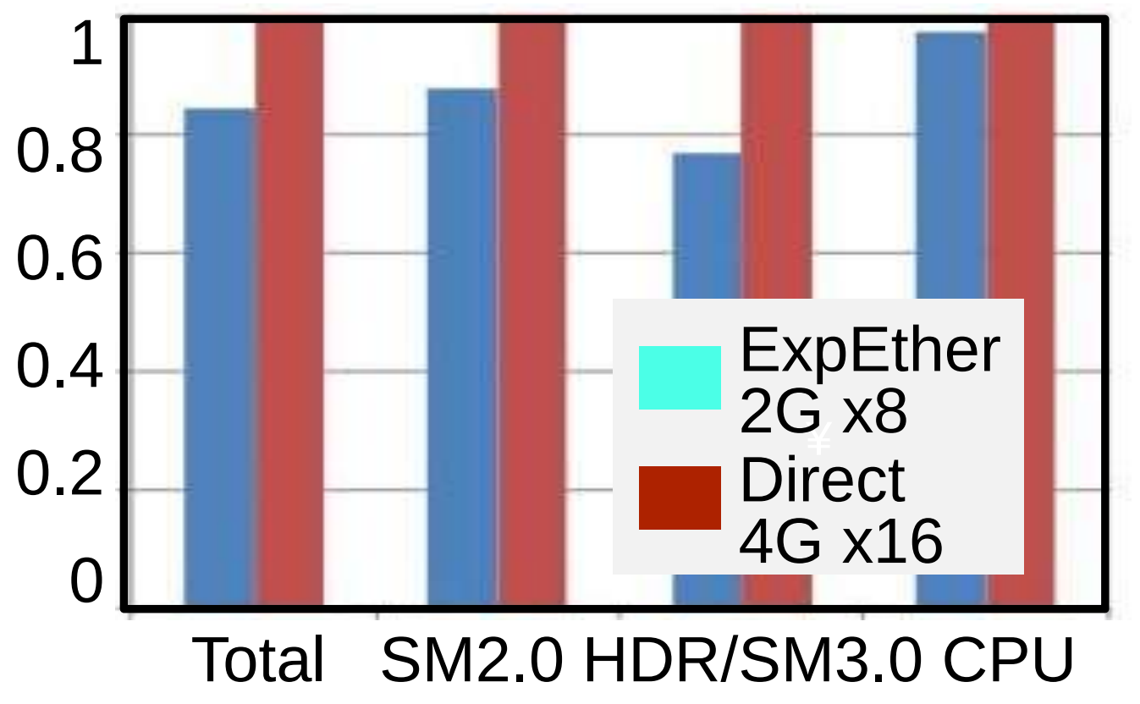
Performance

- Storage : Block I/O read/write to PCIe SSD shows 92/74% of direct-attached device’s performance @4KB size
  - x7 and x2 performance of those of iSCSI, and iSCSI/ToE
- GPU Application : 3D graphic benchmark7 performance is not so much degraded, marking about 80% to 97% score of those direct attached case.

Storage



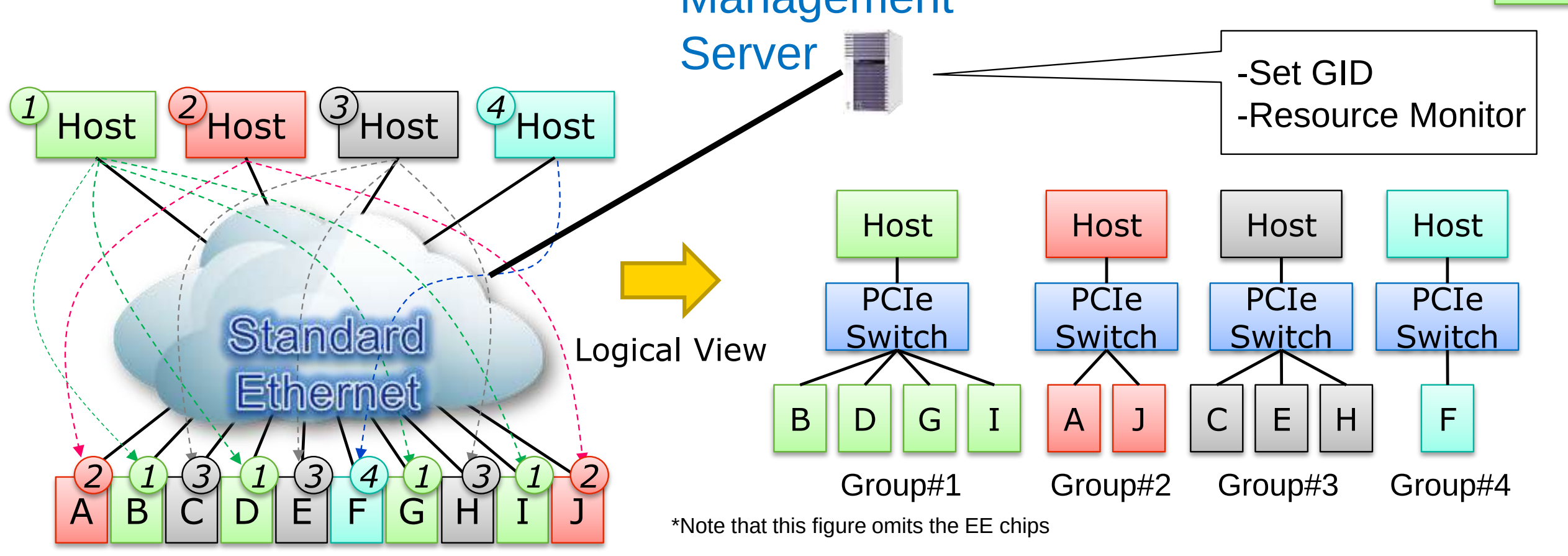
GPU Application



Latency Estimation
(for 10GE-FPGA)
-EE chip ~ 0.7us
-Ether SW ~ 0.2us
-Cable 5ns/m

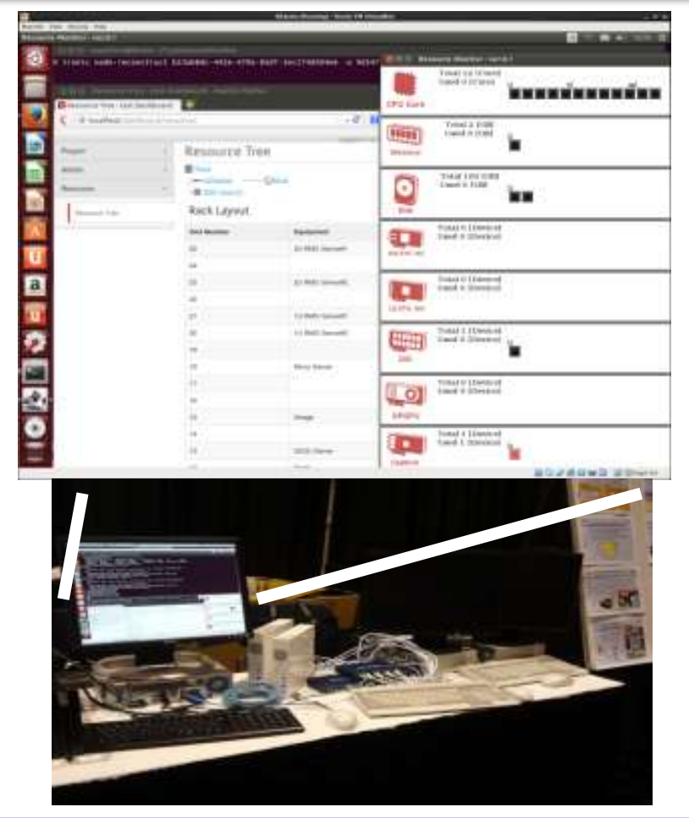
Disaggregated Resource Management

- To make PCIe connection, ExpEther chip has group ID and MAC address
  - I/O devices are attached to host which has the same GID.
  - Packet routing is autonomously performed by Ethernet layer.
- GID can be set remotely by management software.
  - Software defined configuration by OpenStack based software.



*Note that this figure omits the EE chips

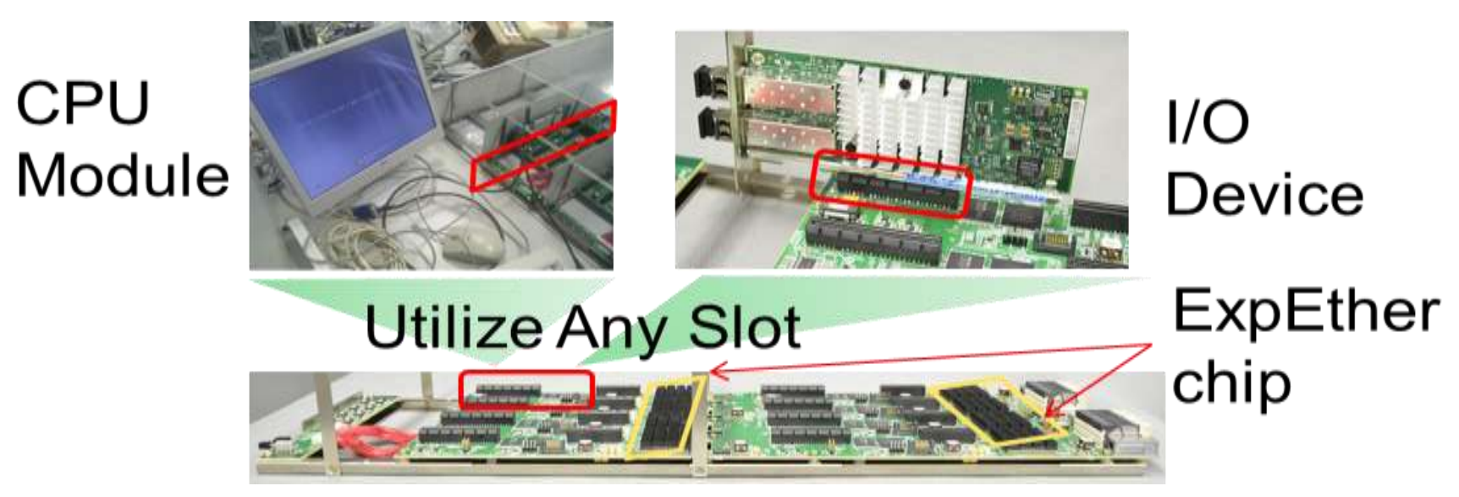
CPU pool system w/ OpenStack Control



Seamless Disaggregated Computer

- To make a computer providing user-required performance and function, necessary devices can be selected and attached from resource pool.
- ExpEther provides a singe-hop PCIe connection for all CPU / devices in the resource pool over Ethernet.

Resource pool proto-type w/ ExpEther on mother board.
Each PCIe slot can be used for both CPU module and IO devices



Multi-rack resource pool system w/
More than100 server / IO devices
(GPU, Storage (HDD, PCIe Flash),
VDI GRID, PCoIP) Service in Apr. 2014.