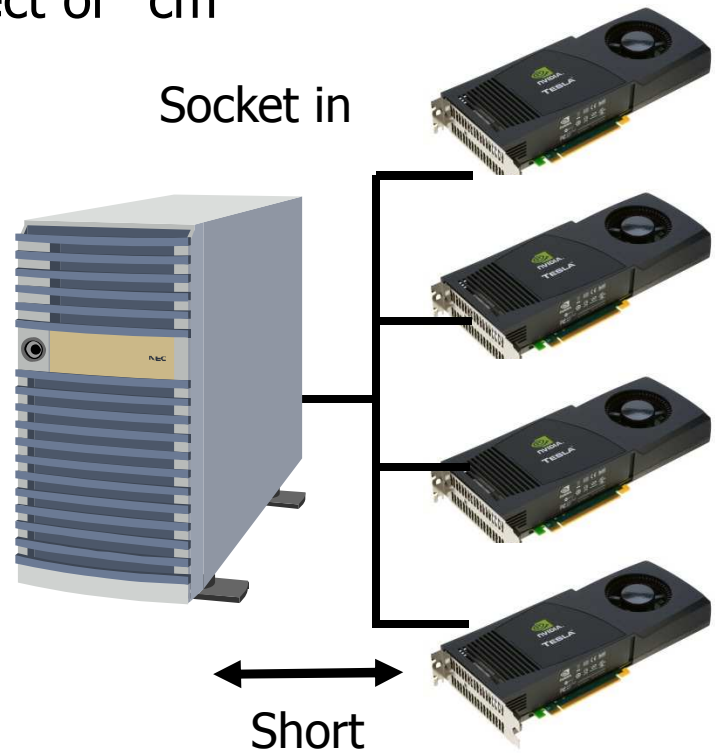
Empowered by Innovation NEC

A scalable connection method beyond processor cores

Yuichi Nakamura
General Manager
Green Platform Research Labs., NEC Corporation

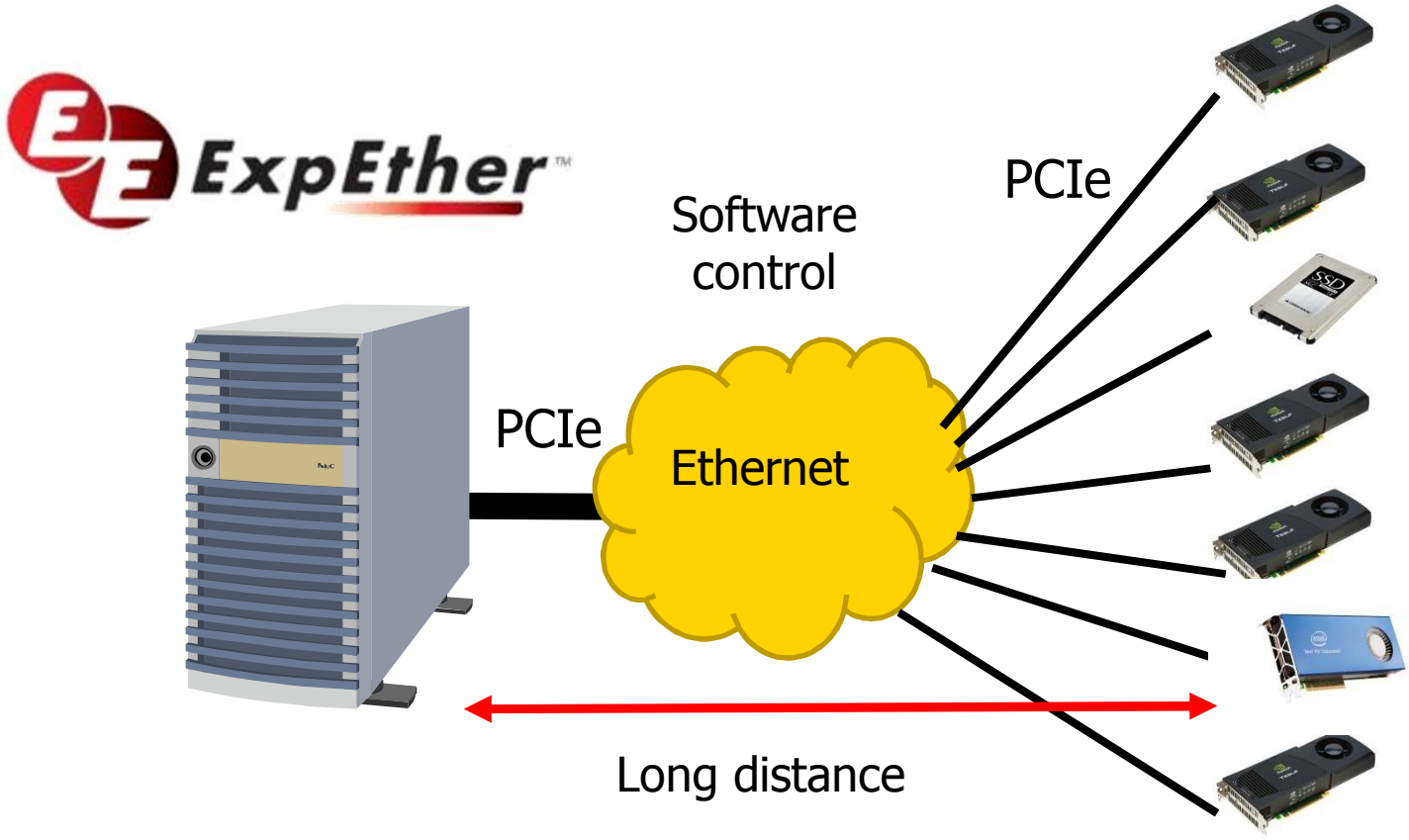
How many PCI devices do you have?

- How many PCI devices your computer have?
  - PCI devices : Interfaces(Display, Graphic), accelerators (GPGPU, etc.)
  - Max 4, Typically 1 or 2.
- How long between your computer and PCI devices?
  - Socket direct or “cm”



Today's talk

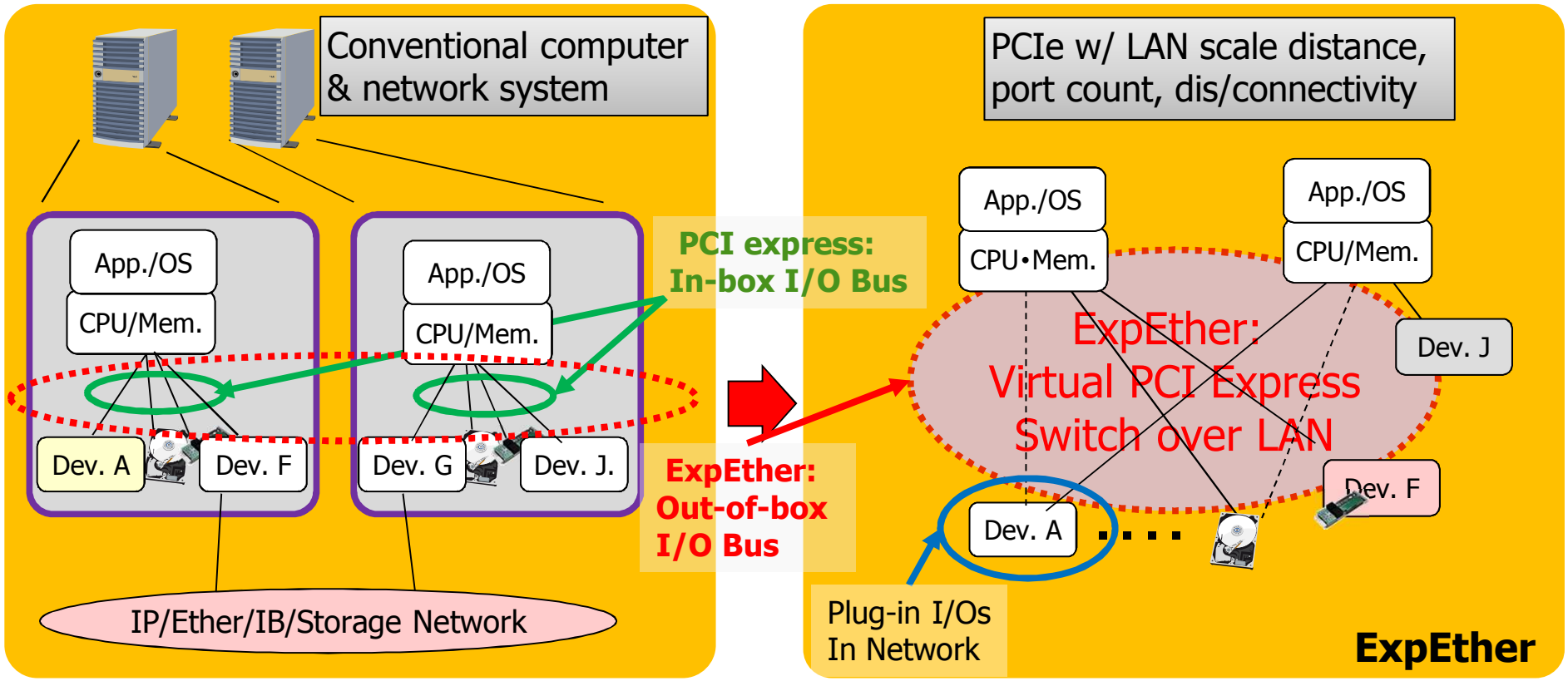
- Many PCI devices can be connected to your computer.
- Long distance connections from your computer to PCI devices.



PCIExpress Protocol on Ethernet= ExpressEther=ExpEthter

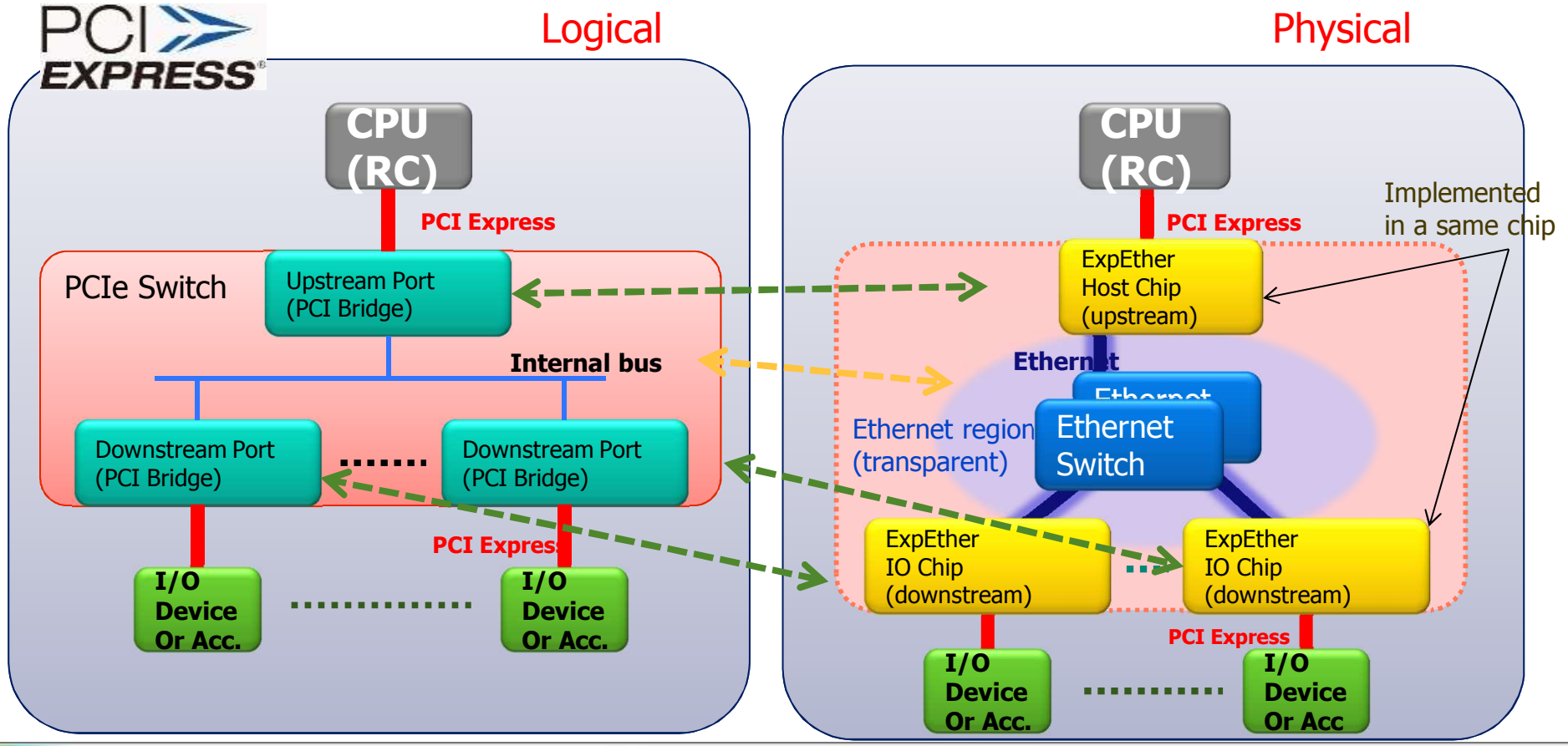
Extend PCIe System Scale

- Seamless extension realizes a big PCIe system with many resources on a single network.
  - function/performance along with requirements



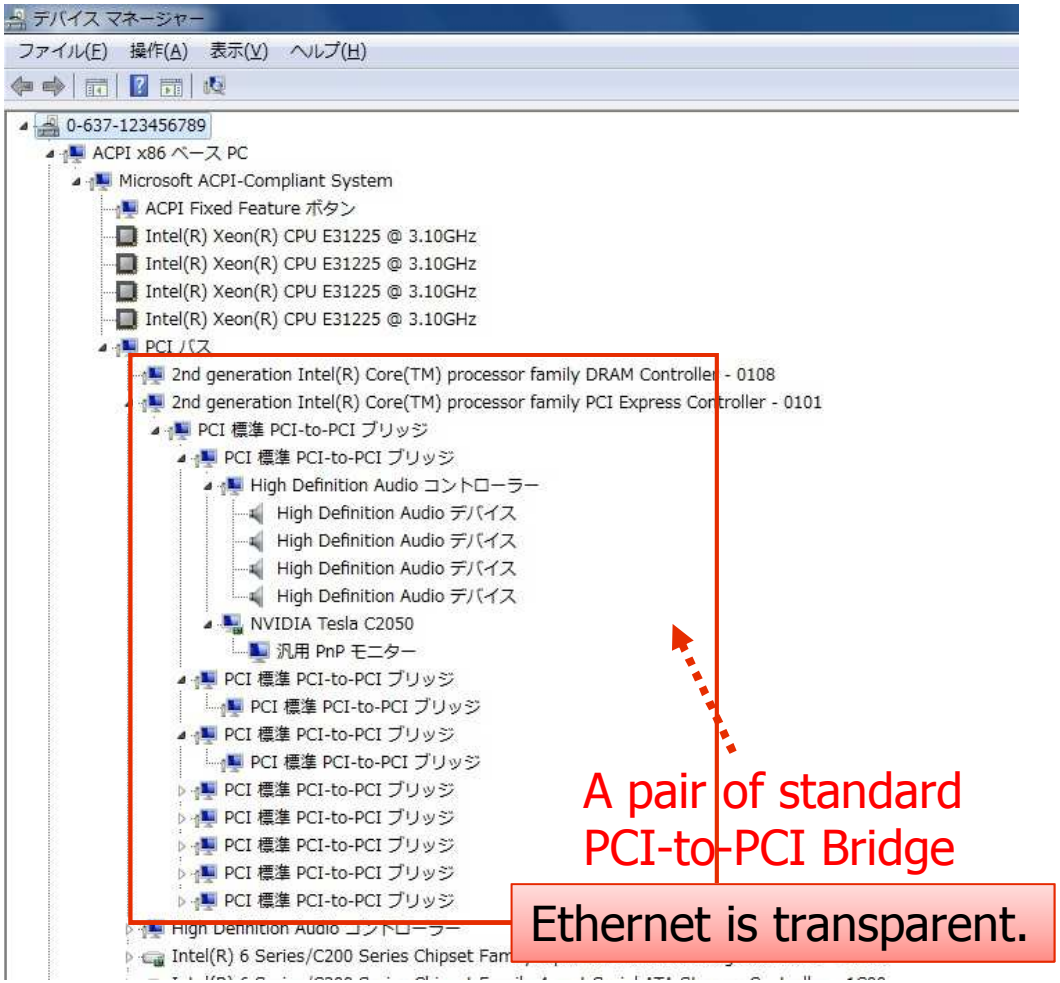
Architecture 1/2 : Distributed PCIe Switch

- Single-hop PCI Express switch over Ethernet.
  - ✓ combination of up/down bridge and Ethernet transport.



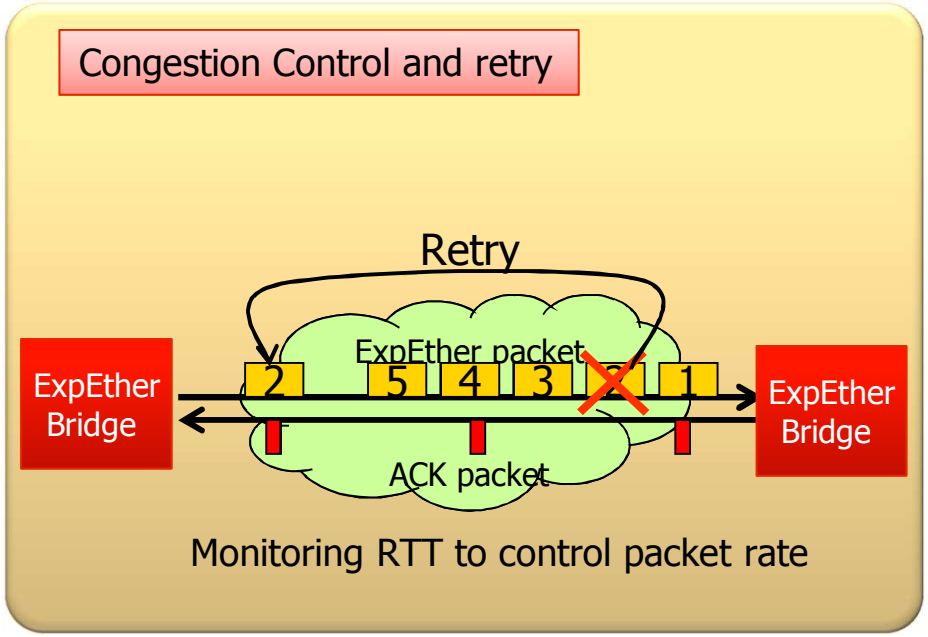
Logical view is a single-hop PCI Express Switch

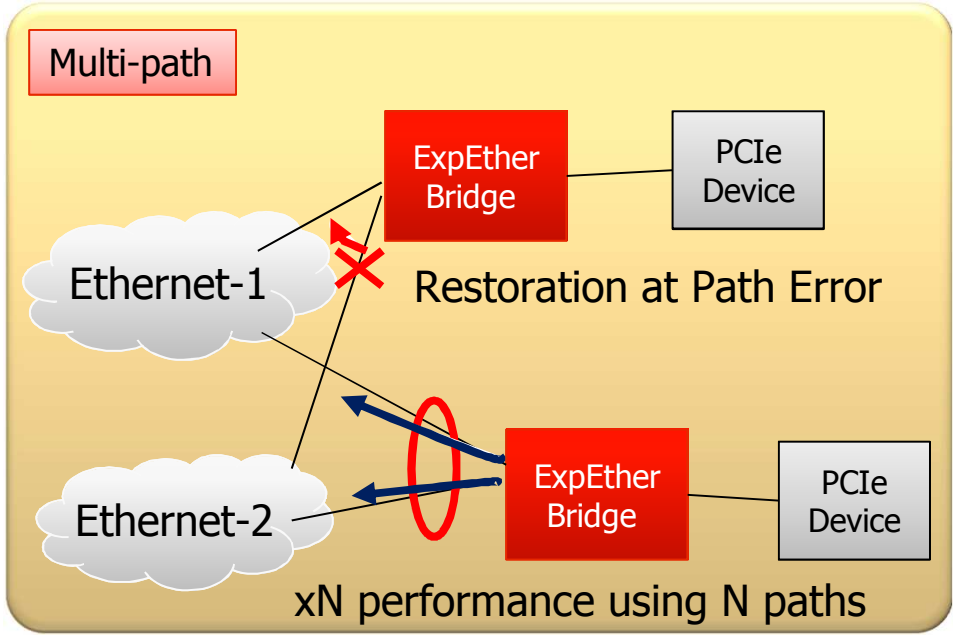
- ExpEther appears as a PCI Express switch from OS/software
- Utilize commodity device, OS, device driver w/o modification



Architecture 2/2 : Reliable Ethernet

- Reliable transport on Ethernet by congestion control and retry.
- xN bandwidth, redundancy by multipath.

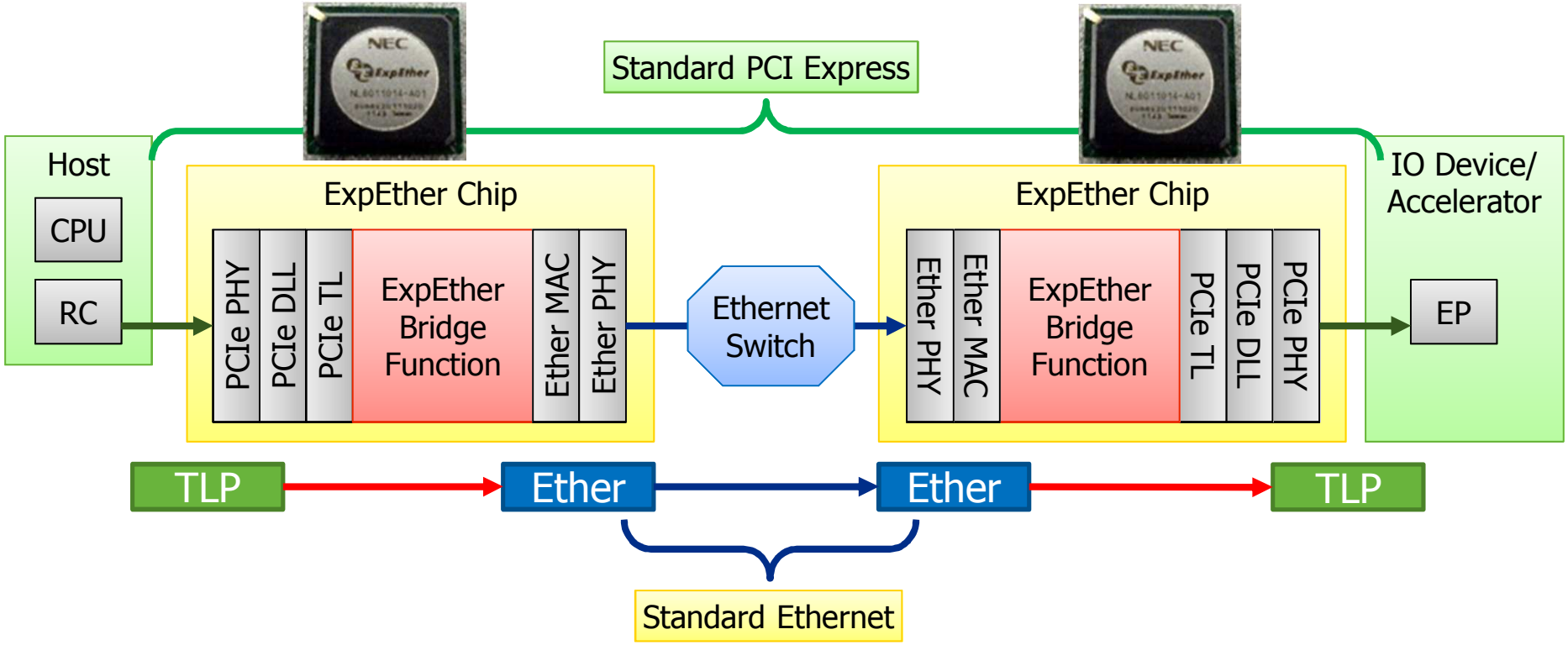




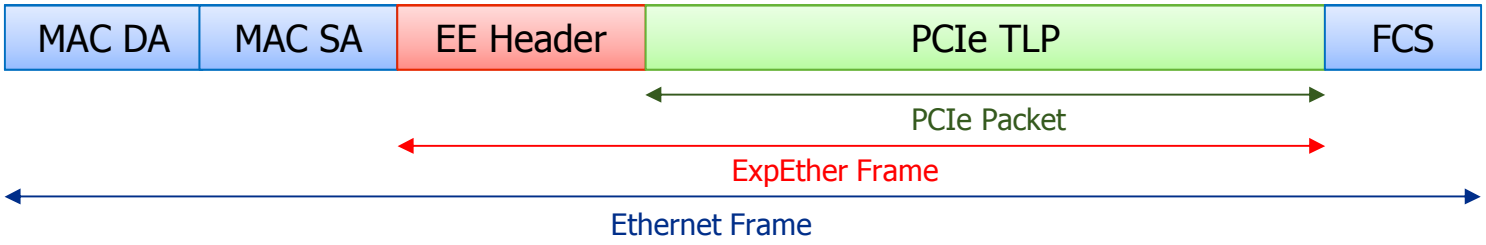
Utilize standard Ethernet cable and switch.

+

Utilize Std. PCIe/Ethernet? Protocol Stack, Frame Format

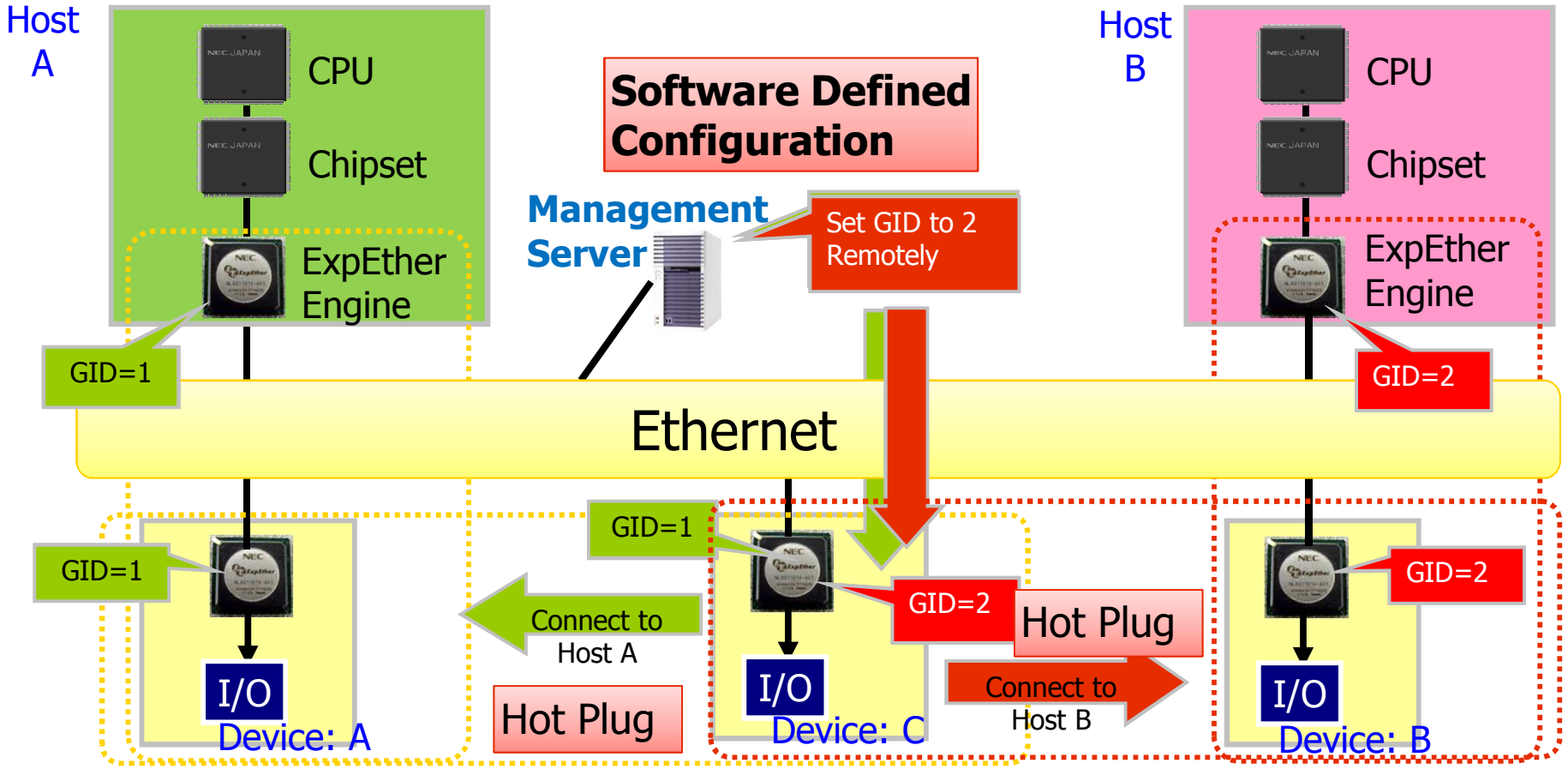


- PCI over Ethernet : ExpEther Frame Format



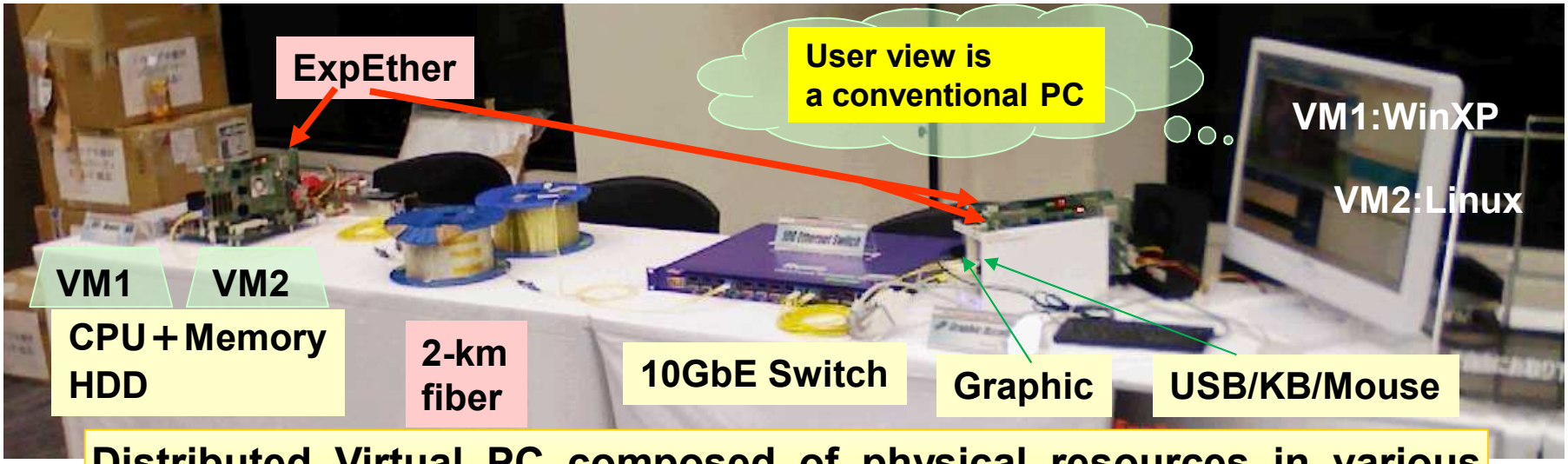
PCIe Hierarchy among Many up/down ports

- Automatic Grouping along with Group ID set in ExpEther chip.



PCIe compliant logical view ~ Distributed PC over 2-km fiber

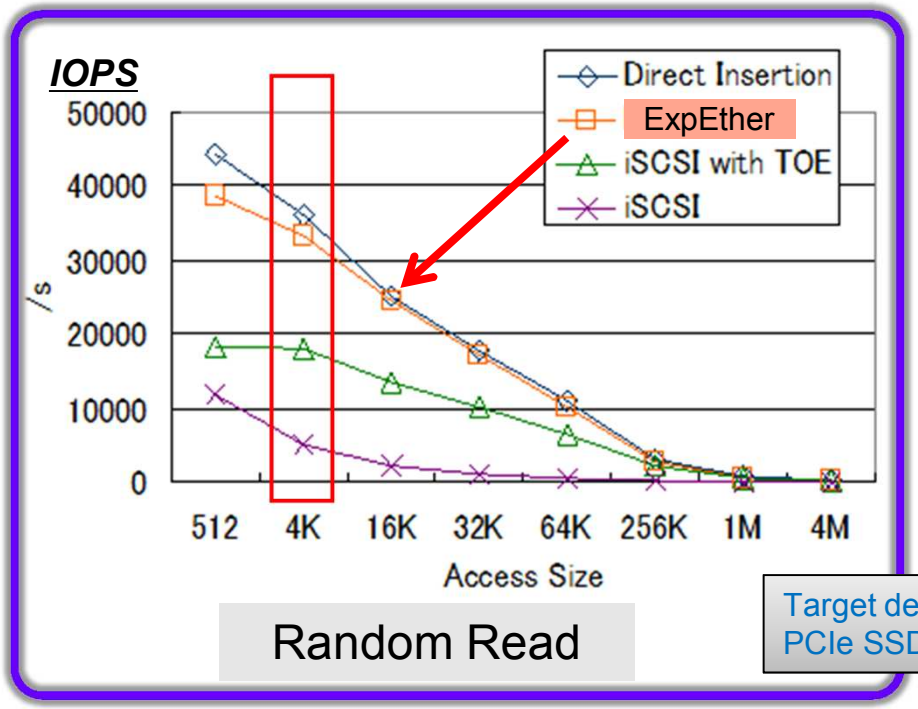
- Mother board at the left side only has CPU/Memory/HDD and ExpEther card.
- Graphic card and USB KB/M at the right side are connected to the mother board by ExpEther through 2-km fiber and Ethernet switch.
- Whole system performed conventional operation without any modification for OS/driver/application.



Distributed Virtual PC composed of physical resources in various location on the Network w/o modification for OS/driver/application.

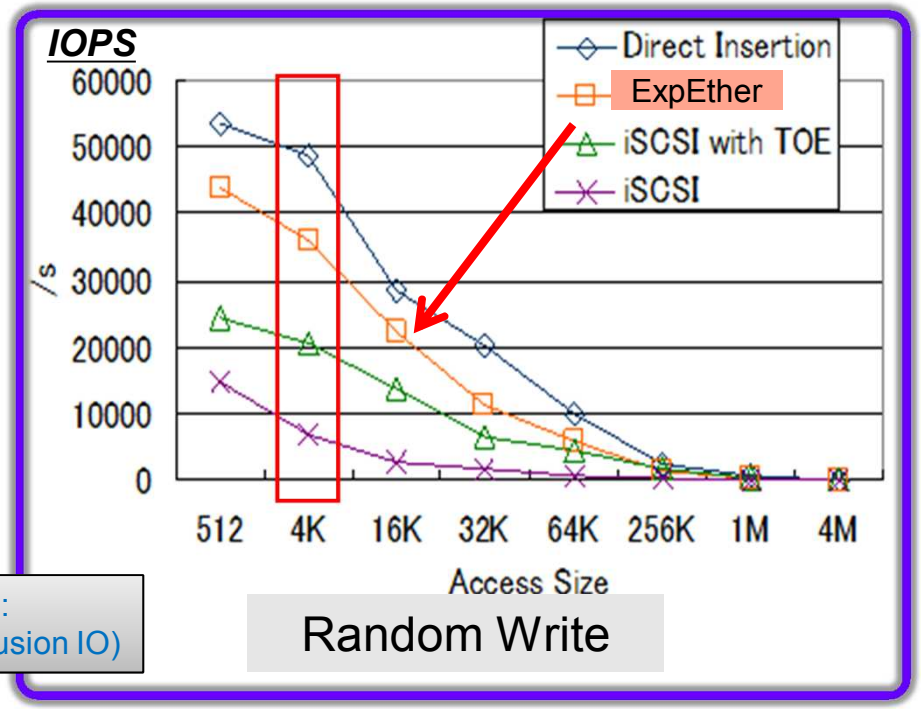
Performance comparable to direct PCIe

- Block I/O read/write to PCIe SSD shows 92/74% performance of those local device @4KB size
  - All function are implemented in a chip w/o S/W stack, TCP slow start window.
  - x2 performance of iSCSI/ToE



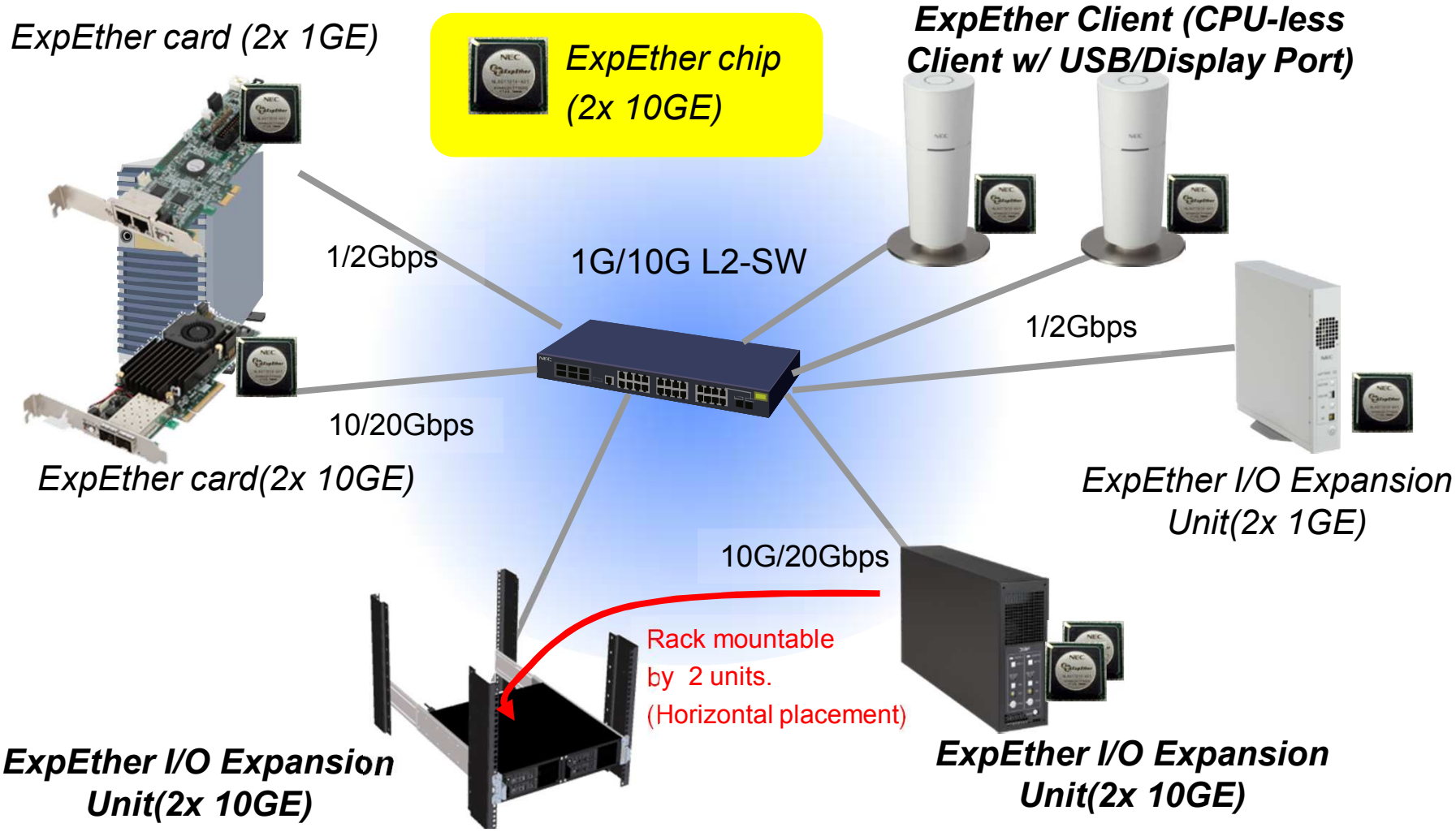
Random Read

Target device:
PCIe SSD (Fusion IO)



Random Write

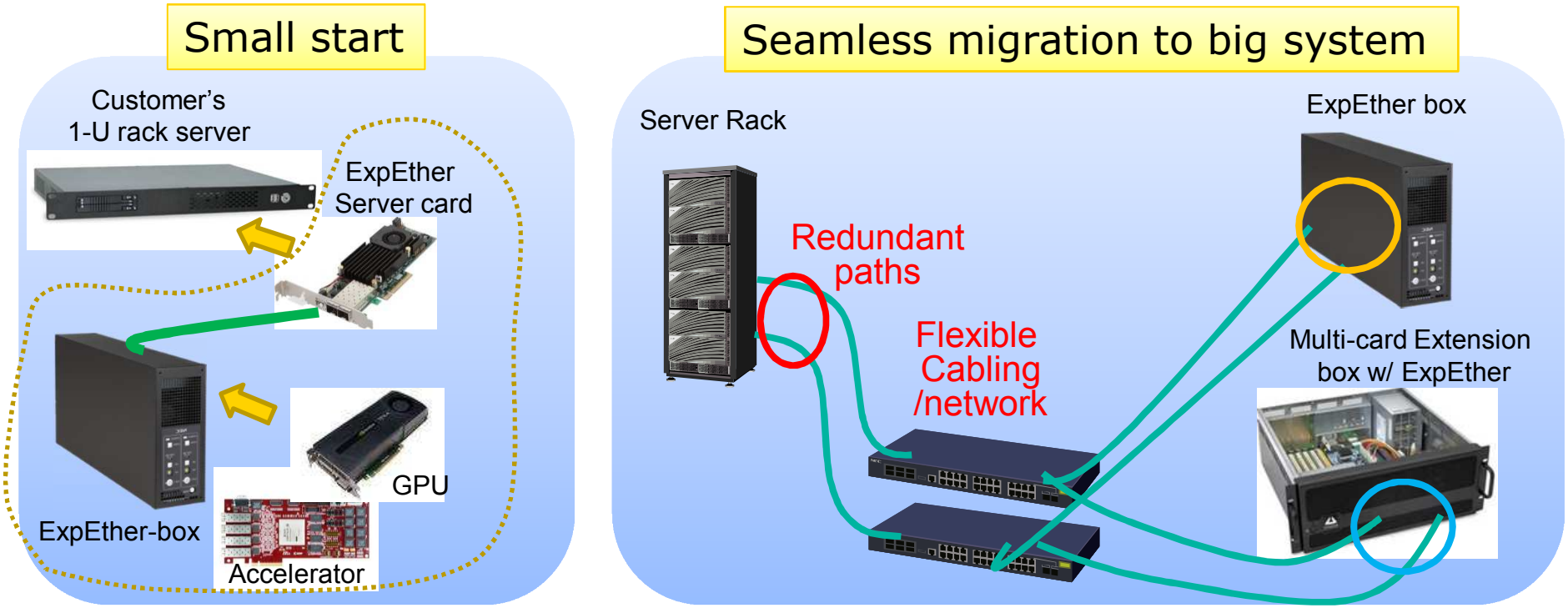
Implementation Example



* An optional rack mount kit enables this box to be mounted within any standard rack environment while only taking up 2U of vertical space.

Application 1 :Small Start, Seamless Scale-up to Data Center Scale

- Host machine only installs ExpEther card.
  - Utilize small machine like 1-u rack-mount server.
- Seamless Migration to big system via Ethernet.

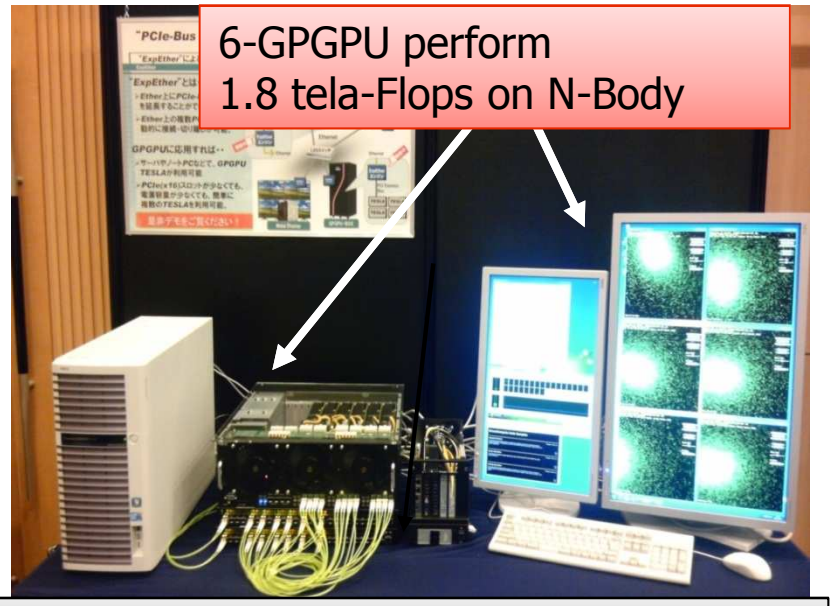


Application 2: net GPU (Graphical Processor Unit)

- Conventional PC can install only 1-2 GPGPUs.
  - Always consume power dedicated to one machine.
- Conventional PC w/ 6 GPGPU performs 1.8 Tera-Flops processing.
  - hot-plug GPU from pool.
  - save power, share device.

8 slots w/ 8-ExpEther chip each has 2x10G Ethernet ports.

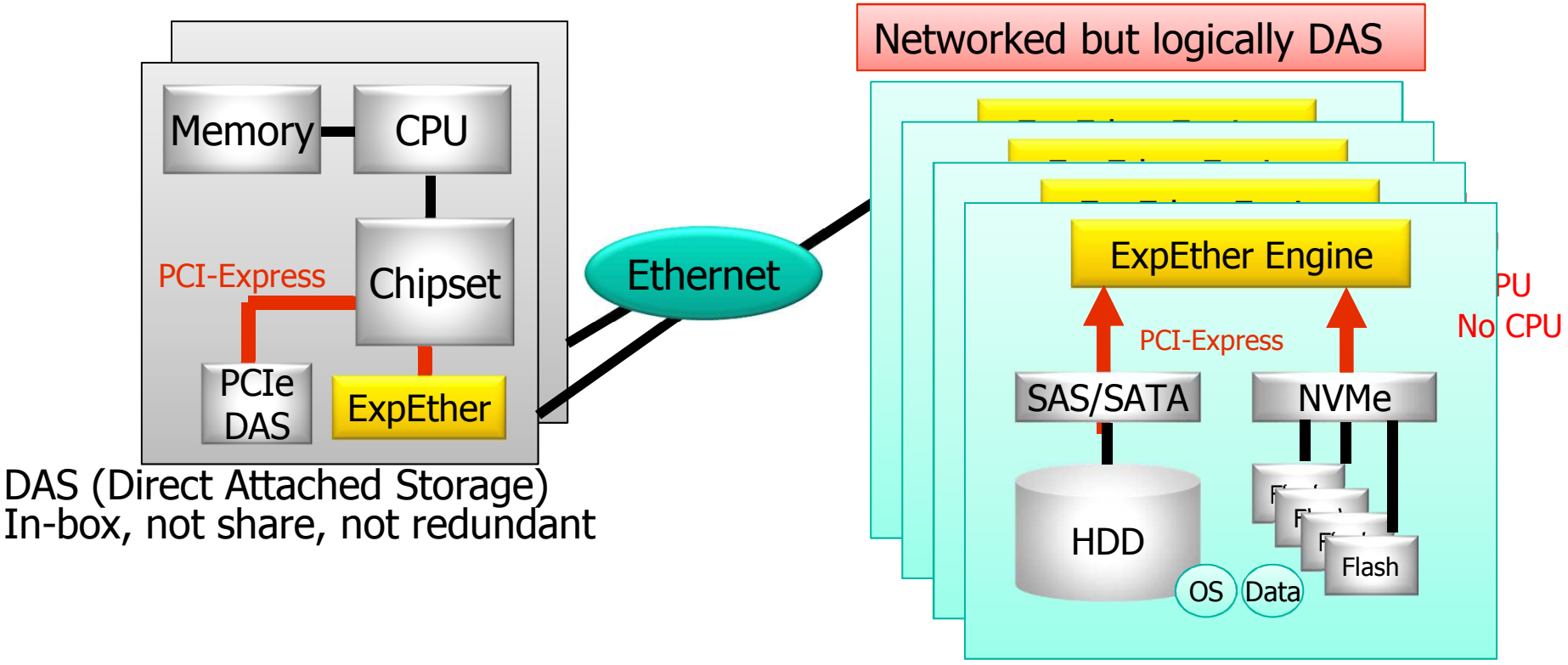
8 GPGPU BOX (Proto Type)



Demo in NVIDIA GTC Japan 2011

Application 3: net DAS (Direct Attached Storage)

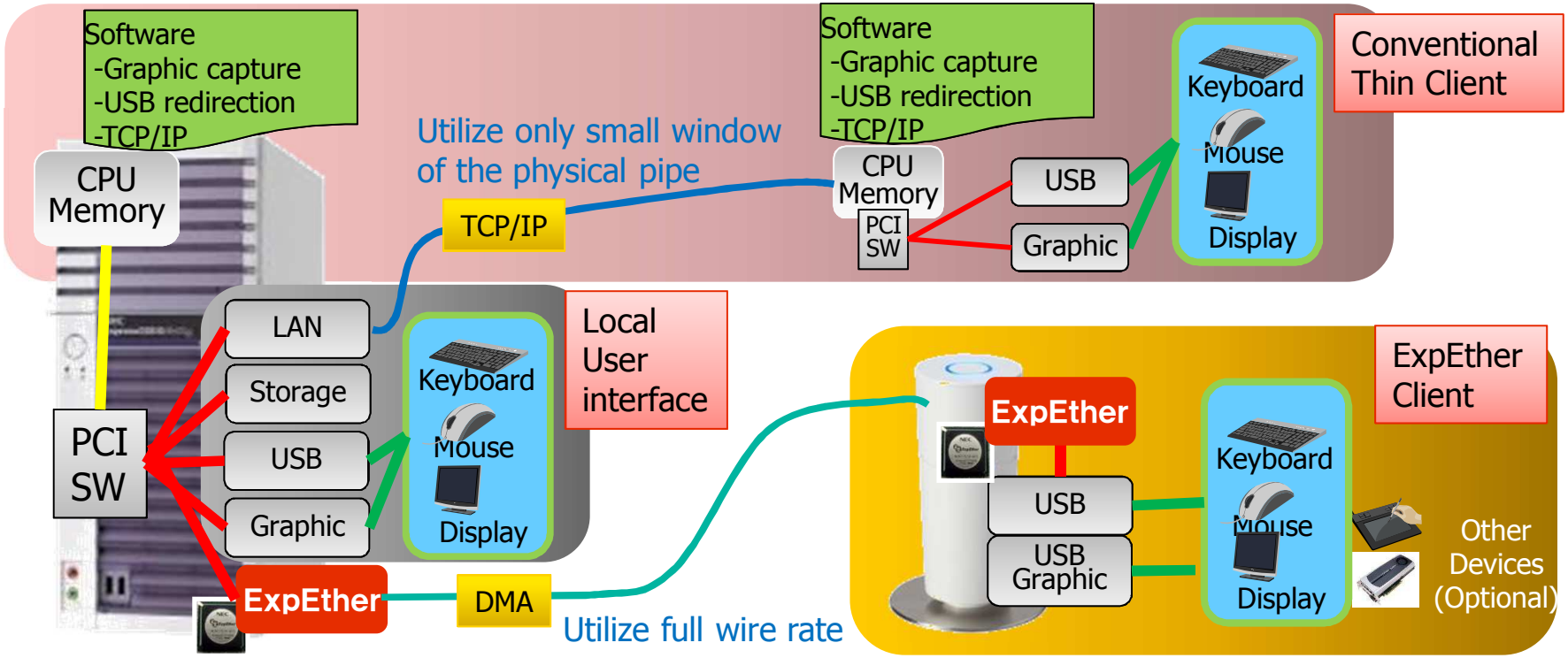
- Place many PCIe DASs into network.
  - Functions (Boot, RAID, HBA) of storage controller can be used over the network.



DAS (Direct Attached Storage)
In-box, not share, not redundant

Application 4: CPU-less High-Performance Thin Client

- Re-locate USB/Graphic to "remote".
  - Logical view, performance are equivalent to "local".
    - w/o CPU, S/W
    - Quick response. Hi-spec graphic, USB (Full HD, USB3.0).



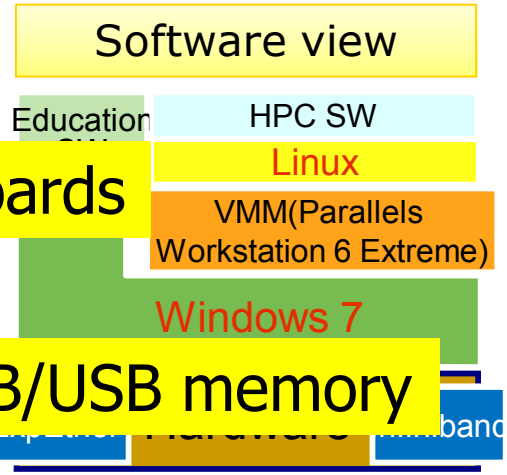
Use case 1: Campus-Scale Single Computer(Osaka Univ.)

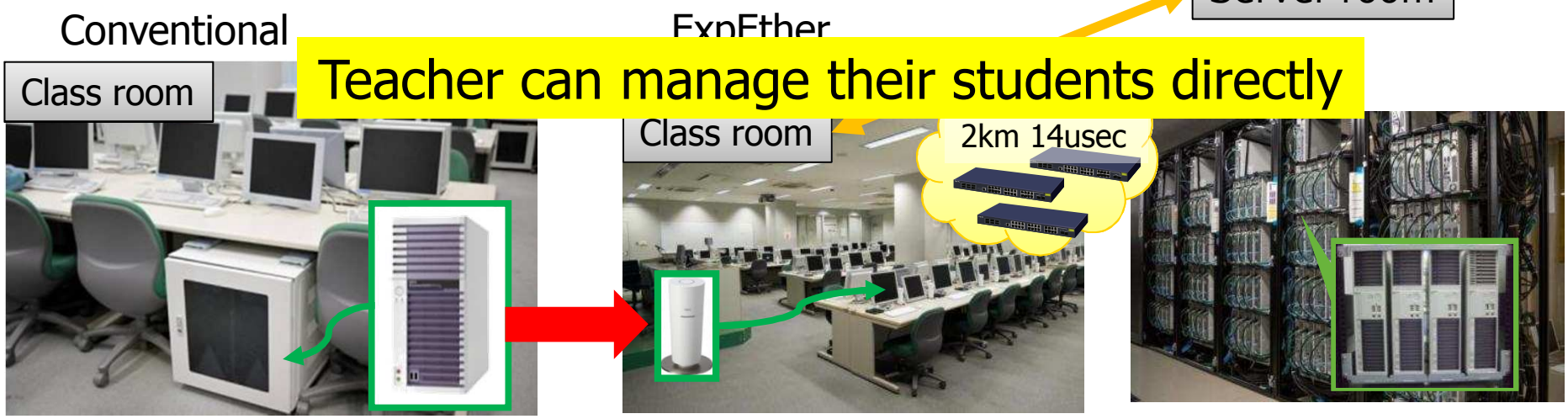
- Conventional: Each desk has a high class PC for student.
  - l
- **(ExpEther) Client w/ USB/Display only.**
  - **1/20 power, 1/10 area, fan/noise less**
  - **Flexible resource allocation**
    - **Day: for students Night: for HPC**

A computer has 600 displays and keyboards

Students can submit their home work via USB/USB memory

Teacher can manage their students directly

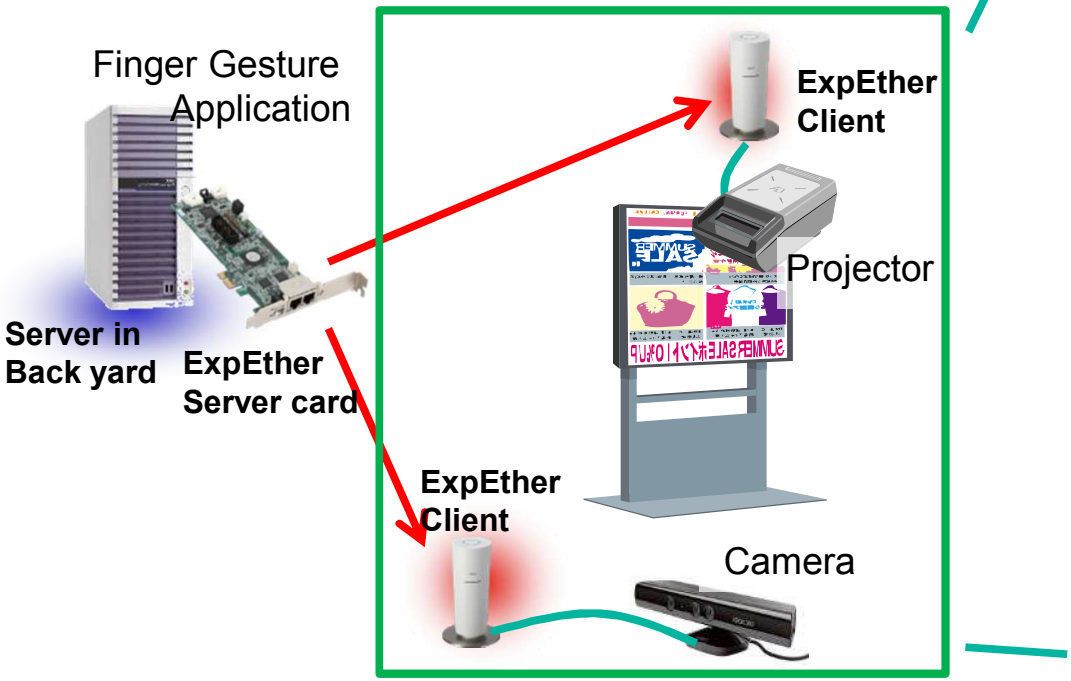




Total 600 units have been working in the campuses since Oct. 2012.

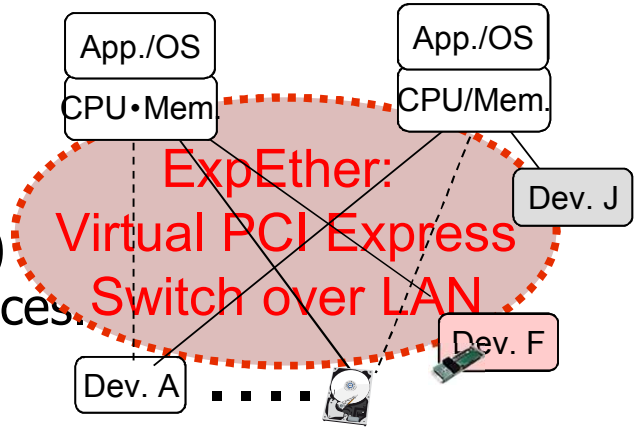
Use case 2: Interactive Motion-Display in Hospital Reception

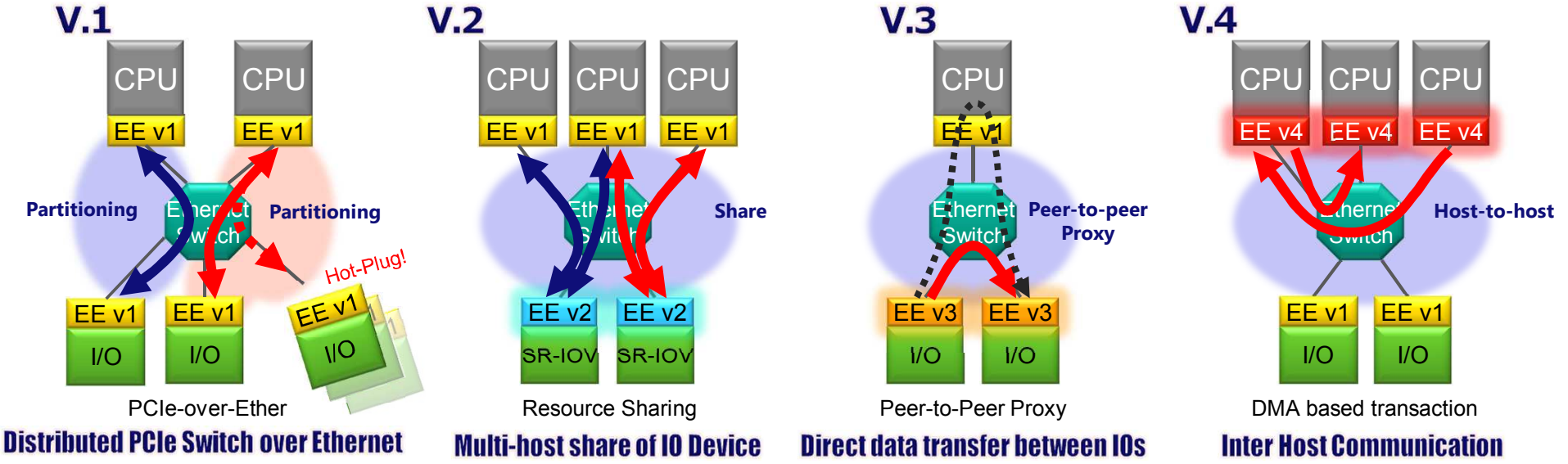
- prevent infection, theft.
- compose motion-display system w/o remote-operational software.



Technology Roadmap

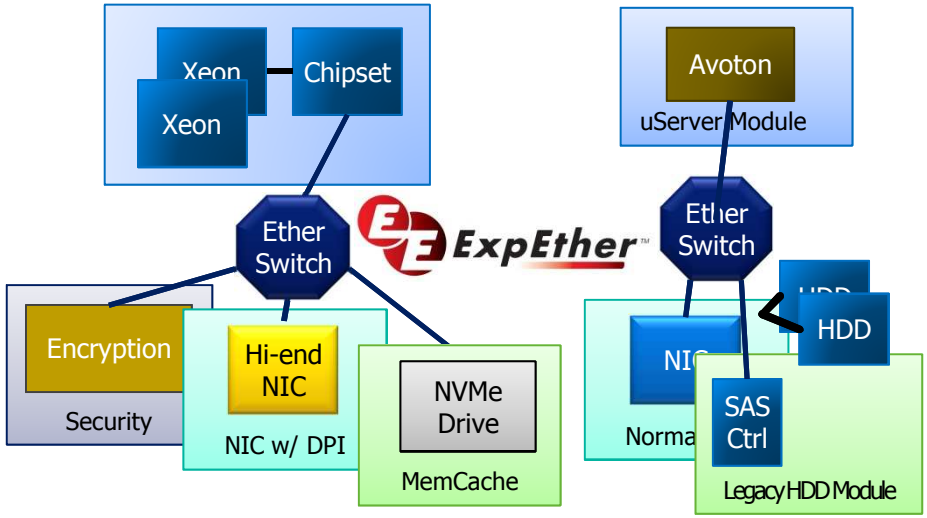
- Ver.1 : PCIe-over-Ethernet (product)
  - Distributed PCIe switch over reliable Ethernet
- Ver.2 : I/O share (product)
  - Multi-host share of SR-IOV device
- Ver.3 : I/O direct connection (Labo. sample)
  - Proxy of peer-to-peer transfer between I/O devices.
- Ver.4 : Inter host communication (Plan)
  - Hi-speed data transfer between hosts

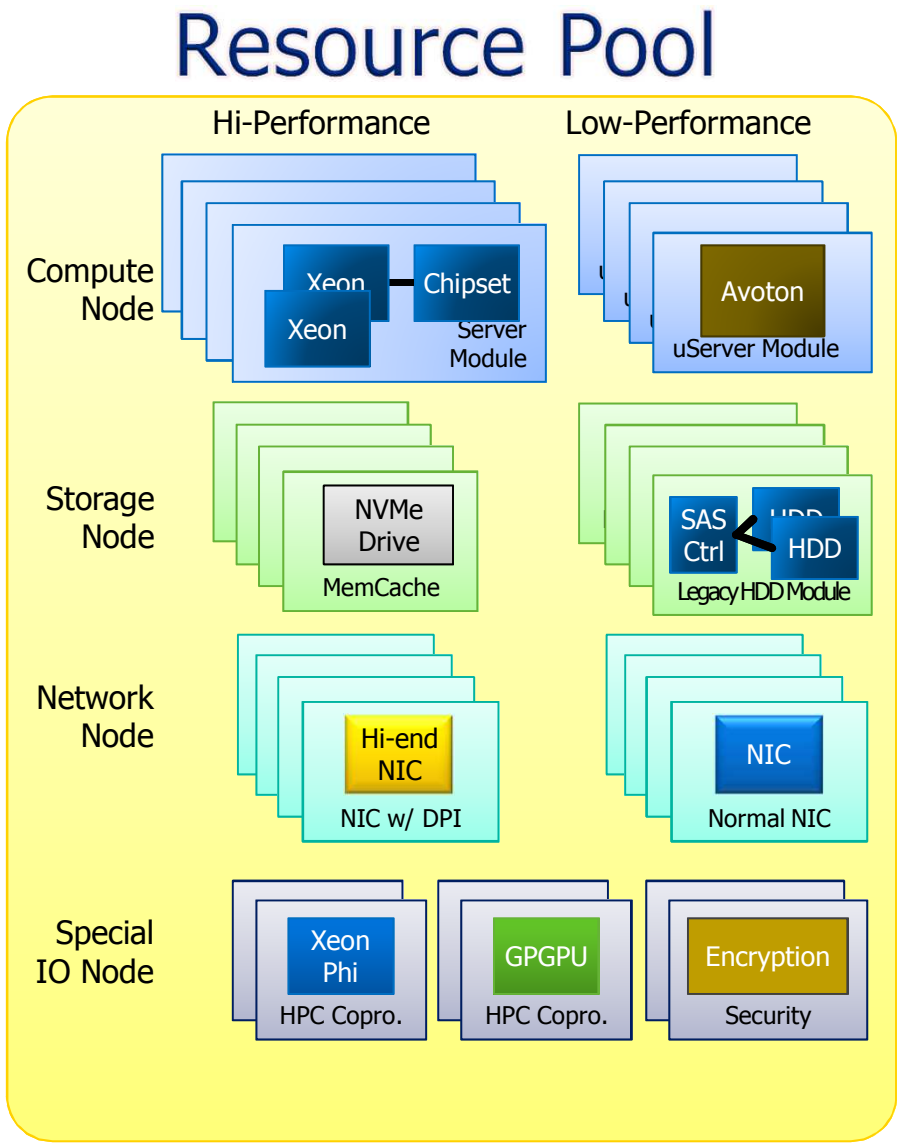




Disaggregated and Adaptive Computer for Various Requirements

- Customer can compose the most appropriate server by combining special or high-performance devices along with various requirements, from resource pool, interactively.



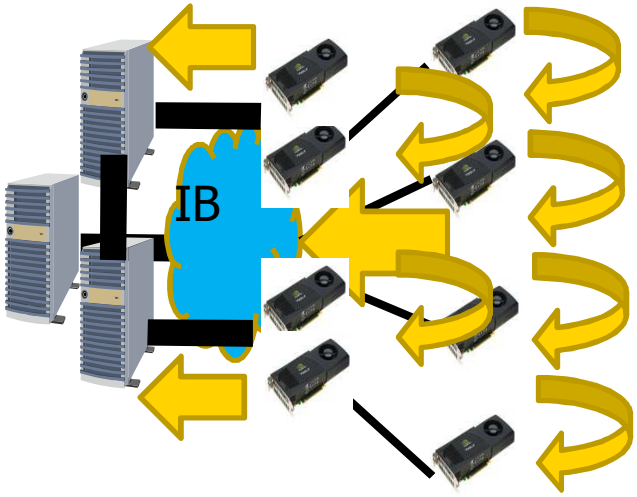


Disadvantage and Future plan

- Bandwidth and speed is limited by ExpEther switch
  - For frequent access among PCI devices and host computer,
    ExpEther may cause serious degradation.
  - New PCIe Switch for large bandwidth is on going
  - Application level design is the most important.
    - Which kind of application ?
    - How to implement for ExpEther
- Software control is important
  - Flexible connection according to traffic condition
- Security

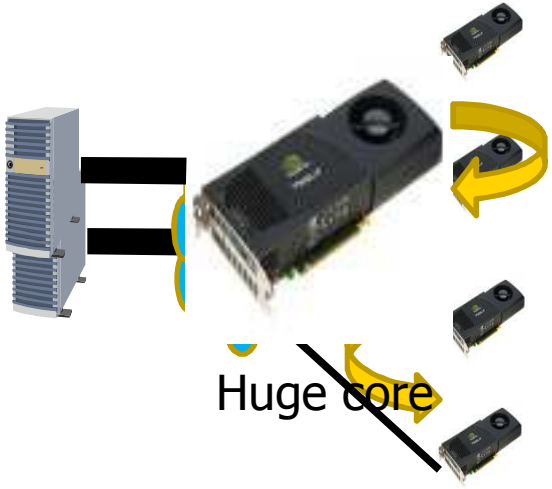
Potential Applications

- Each accelerator works independently.
- Sometimes, huge communication is occurred.



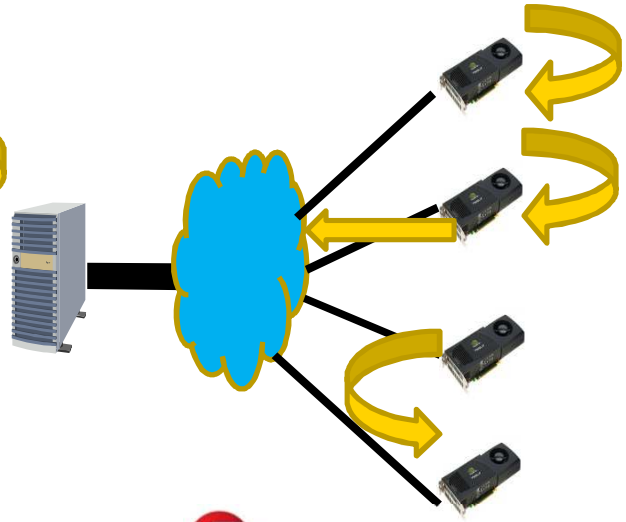
Sever level integration is the best (Infiniband)

- Frequent large communications are exist.



Chip scale integration is necessary

- Intermittent small communications are occurred



ExpEther is the best

We are stating a scheduling and allocation method for ExpEther based computers

Summary

- Using Ethernet as a transport, PCIe usage can be extended.
- Distributed PCIe switch architecture and reliable Ethernet function realize Lan-scale PCIe system w/o software modification.
  - Utilize open commodity software, hardware.
- The technology is already standard, in service.
  - Osaka Univercity and Kurashiki Citizens Hospital
- Implementation Issues exist for better performance, interoperability, and reliability.
  - Bandwidth and software control
- Adaptive computing to compose from a resource pool will be realized in near future.