Empowered by Innovation NEC

Adaptive Memory System over Ethernet

Jun Suzuki, Teruyuki Baba, Yoichi Hidaka†, Junichi Higuchi, Nobuharu Kami, Satoshi Uchida††, Masahiko Takahashi, Tomoyoshi Sugawara, and Takashi Yoshikawa

System Platforms Research Laboratories, NEC Corporation

†IP Network Division, NEC Corporation

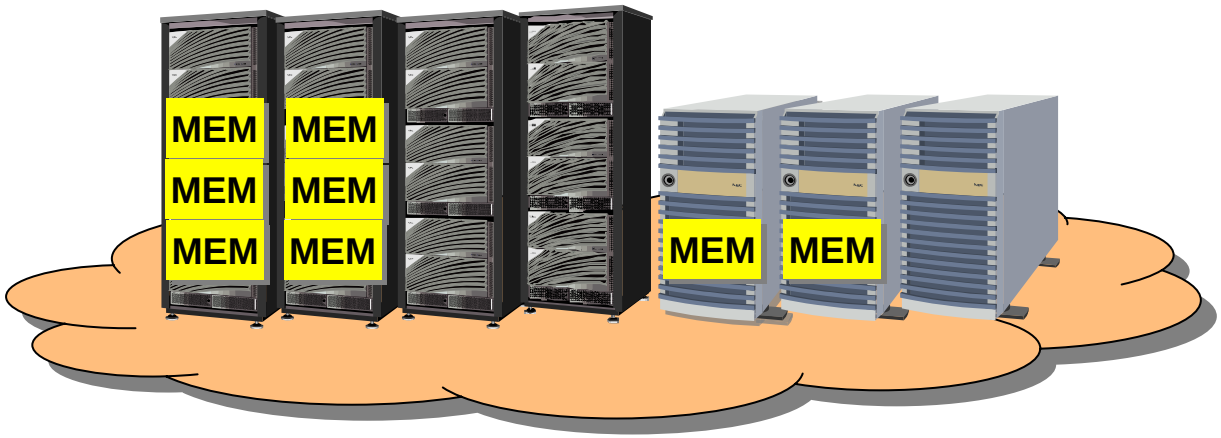
††2nd Computer Software Division, NEC Corporation

Memory Scalability in Cloud Computing

Computer memory is limited by individually loaded resources

- Cannot scale depending on service requirements
- Service performance limited by memory
- Slow block I/O devices

➡ *Needs for scaling memory beyond individually loaded amount*



Resource Should Simultaneously be High-Performance and Networked

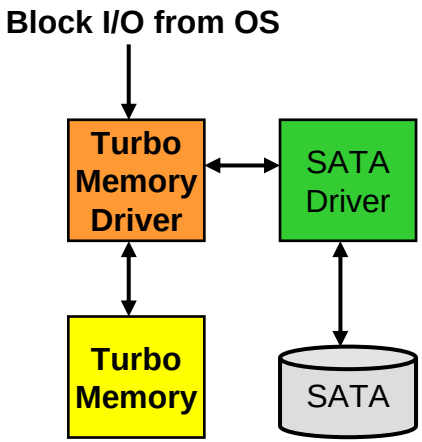
High-Performance

- Large throughput, low latency
- Avoid firmware process and memory copy to transfer data

Networked

- Resource share among multiple computers
- Ease of management

Related Works

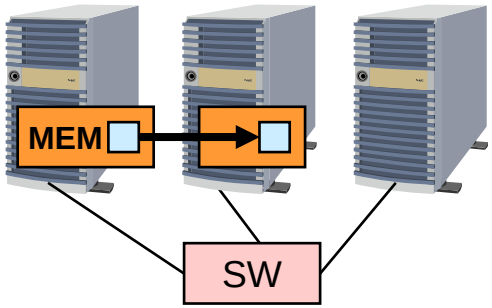


(A) Intel Turbo Memory

High Performance

- PCIe flash device for disk cache
- Device driver between OS and disk driver

J. Matthews *et al.*, “Intel Turbo Memory: Nonvolatile Disk Caches in the Storage Hierarchy of Mainstream Computer Systems”, ACM Trans. on Storage, vol.4, no. 2, article 4, 2008.



(B) Remote Page Swap

Resource Share by Network

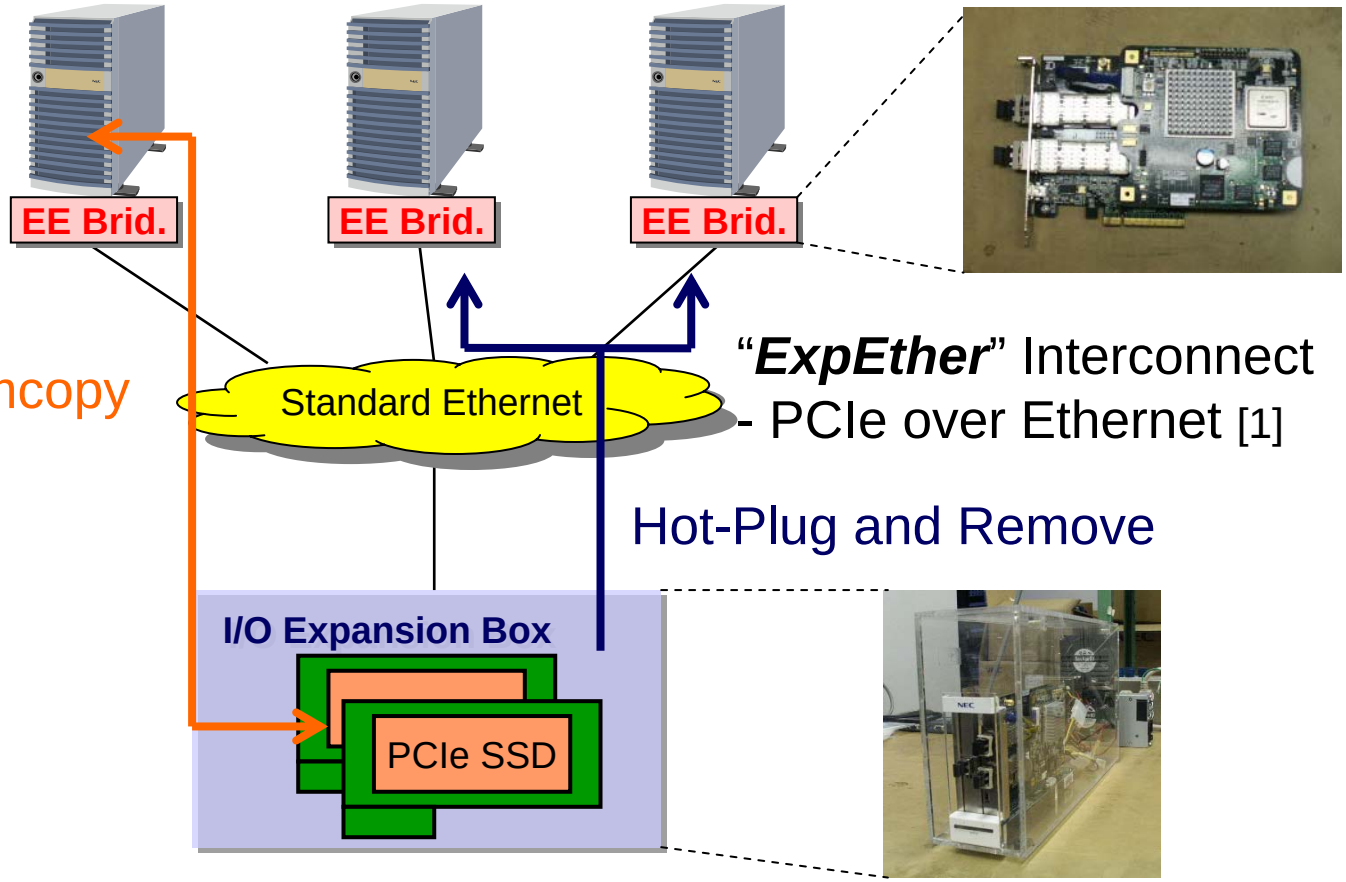
- Using memory of next machine with swapping
- Standard interconnection, e.g., Ethernet

E. P. Markatos and G. Dramitinos, “Implementation of a Reliable Remote Memory Pager”, USENIX 1996 Annual Technical Conference, 1996.

Our Method: Ethernet-Attached SSD as High-Speed Swap Device

High-Performance AND Resource Share

- Standard Ethernet, PCIe SSD



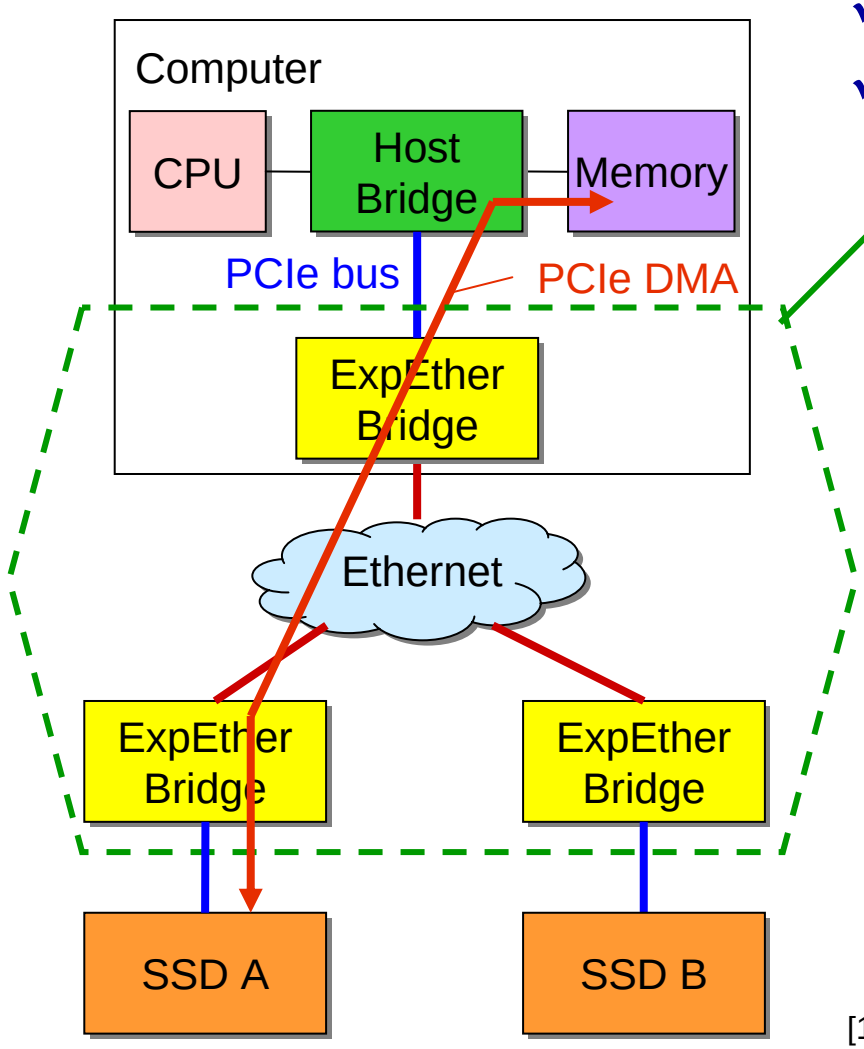
PCIe DMA
w/o Firmware, Memcopy

"***ExpEther***" Interconnect
- PCIe over Ethernet [1]

Hot-Plug and Remove

[1] J. Suzuki *et al.*, "ExpressEther – Ethernet-Based Virtualization Technology for Reconfigurable Hardware Platform", 14th IEEE Symposium on High-Performance Interconnects, pages 45-51, 2006.

PCIe DMA over Ethernet



✓*No Firmware Process*

✓*No Memory Copy*

☐ Extending PCIe Tree over Ethernet

- PCIe packet encapsulation into Ethernet frames
- Ethernet region is PCIe switch

No Driver

☐ High-Speed Ethernet Transport [1]

- Delay-based congestion control
- < 8.5% of TCP-based delay

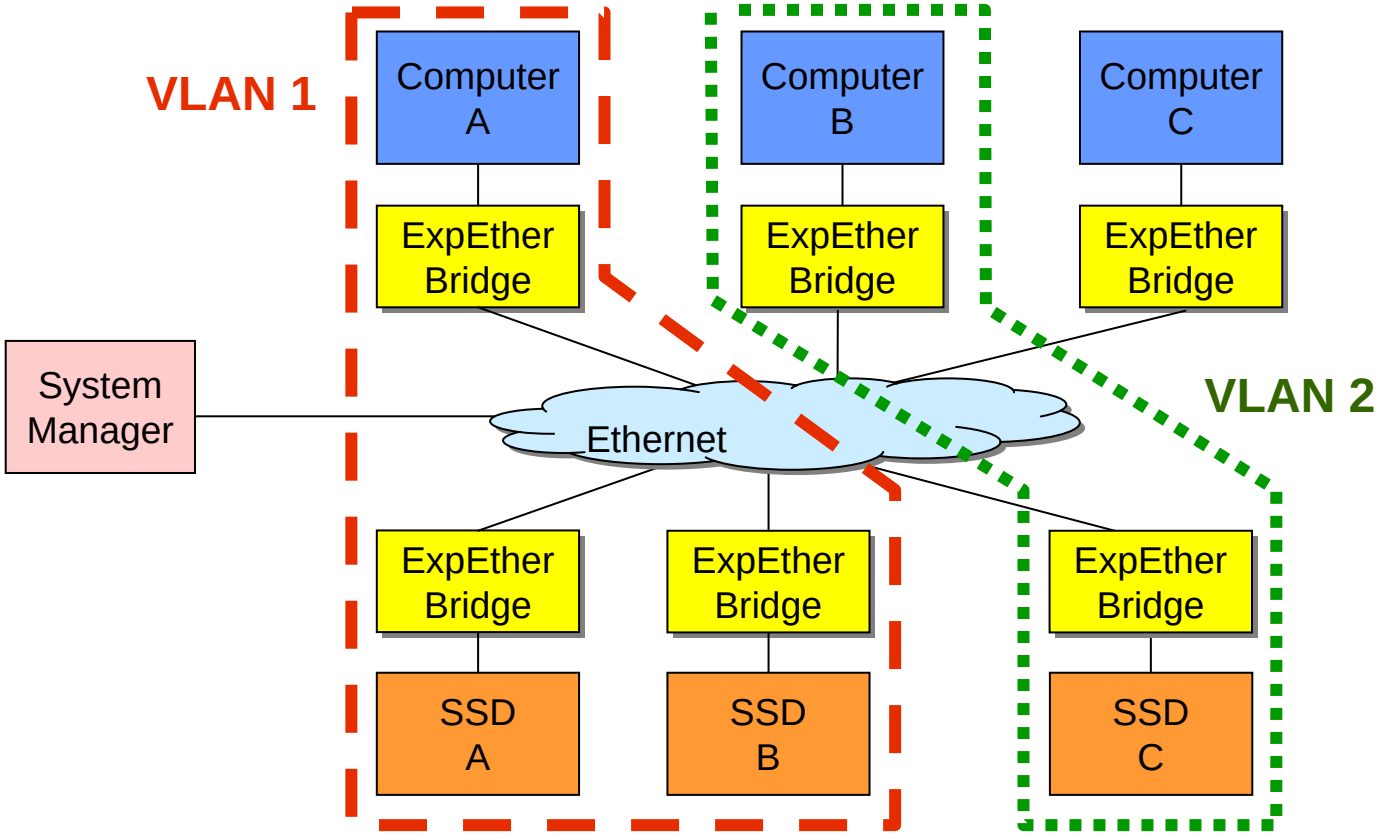
Standard Ethernet

[1] H. Shimonishi *et al.*, "A Congestion Control Algorithm for Data Center Area Communications", 2008 International CQR Workshop, 2008.

Hot-Plug and Remove

SSDs Assigned to Computer with VLAN Grouping

- Adaptive assignment using system manager
- PCIe-standard hot-plug and remove

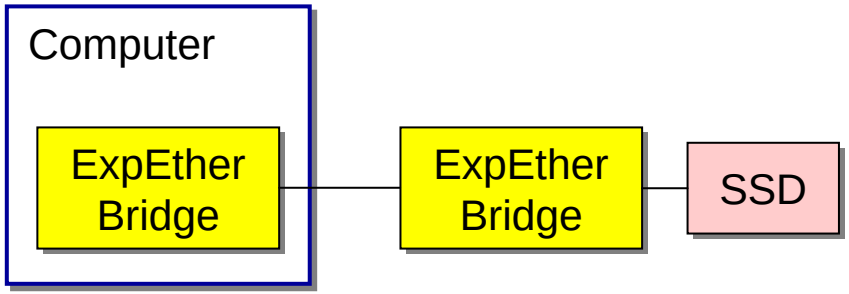


Evaluations

- Block I/O Performance of Ethernet-Attached SSD
- System Evaluation: In-Memory DB

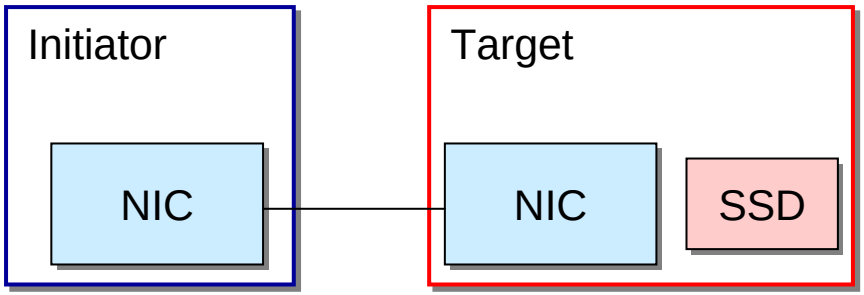
Evaluation Setups

Proposal

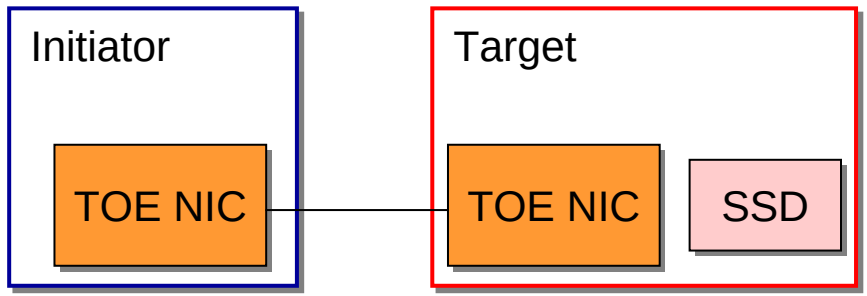


(a) ExpEther

Conventional



(b) iSCSI

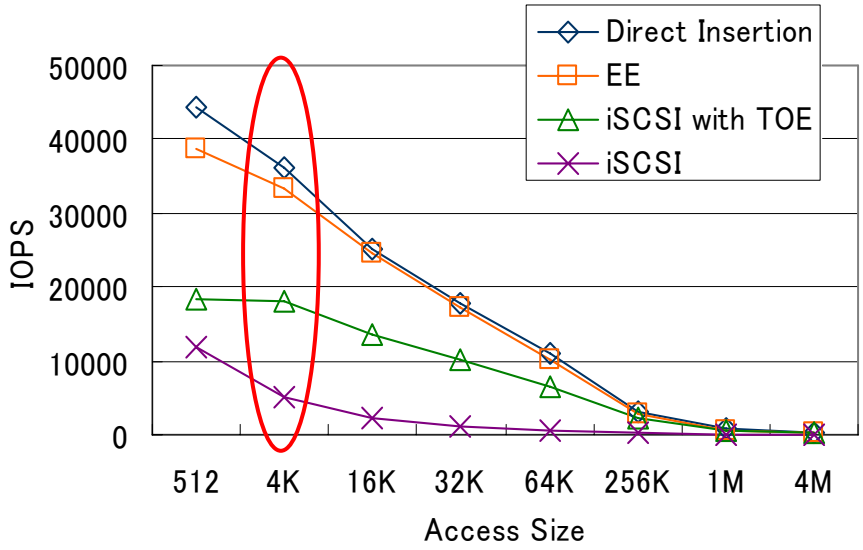


(c) iSCSI with TOE

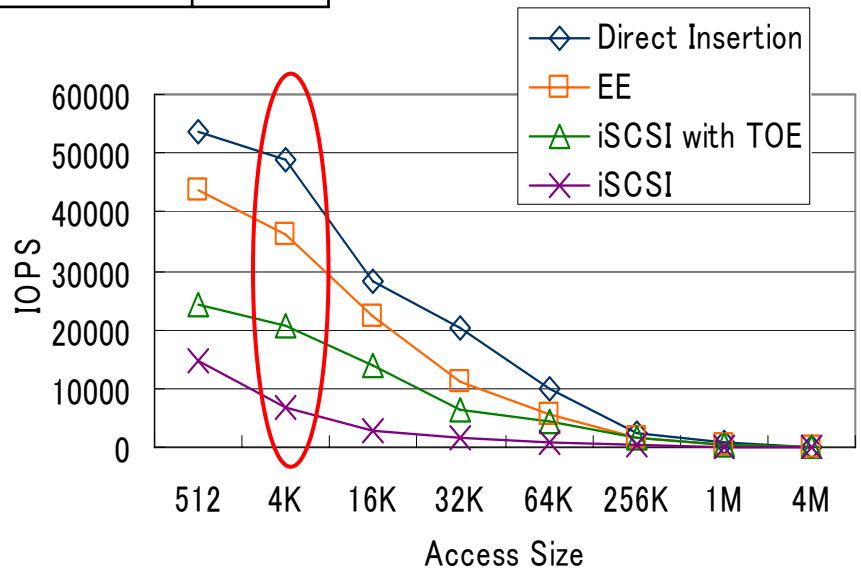
Block I/O Performance (IOPS) of Ethernet-Attached SSD

Read Close to Host I/O Slot, Write Twice of TOE iSCSI

	Host I/O Slot	**ExpEther**	iSCSI w/ TOE	iSCSI
Ran. Read	100	**92**	50	14
Ran. Write	100	**74**	42	14
Ran. Read w/ Switch	100	**91**	46	14
Ran. Write w/ Switch	100	**68**	39	14



(a) Random Read IOPS

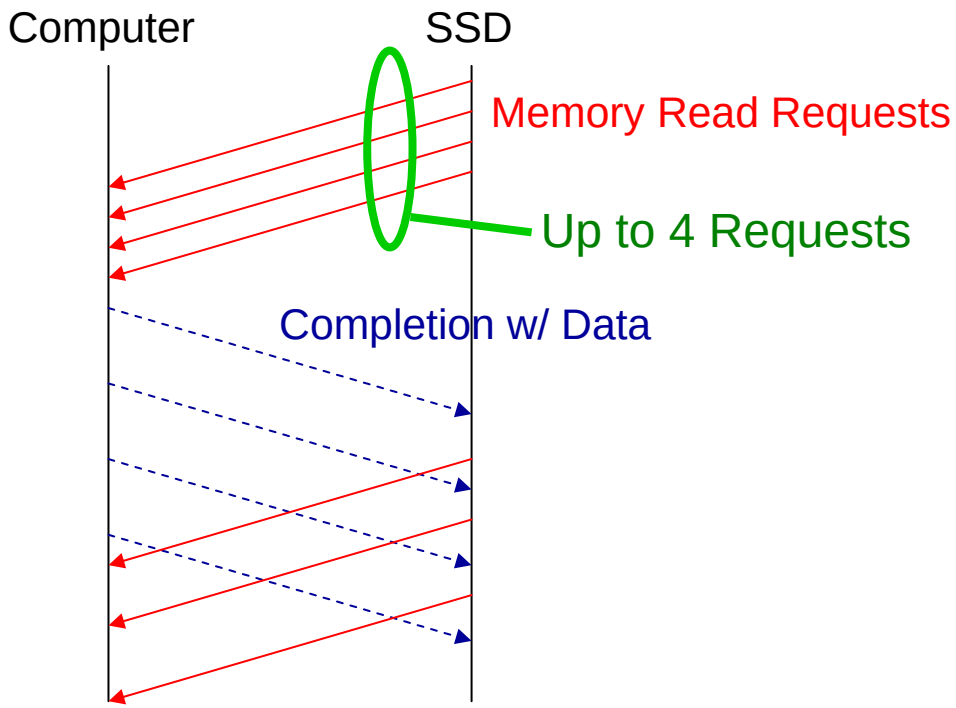


(b) Random Write IOPS

Write IOPS Overhead and Its Solution

Number of SSD's outstanding read request limited by its implementation

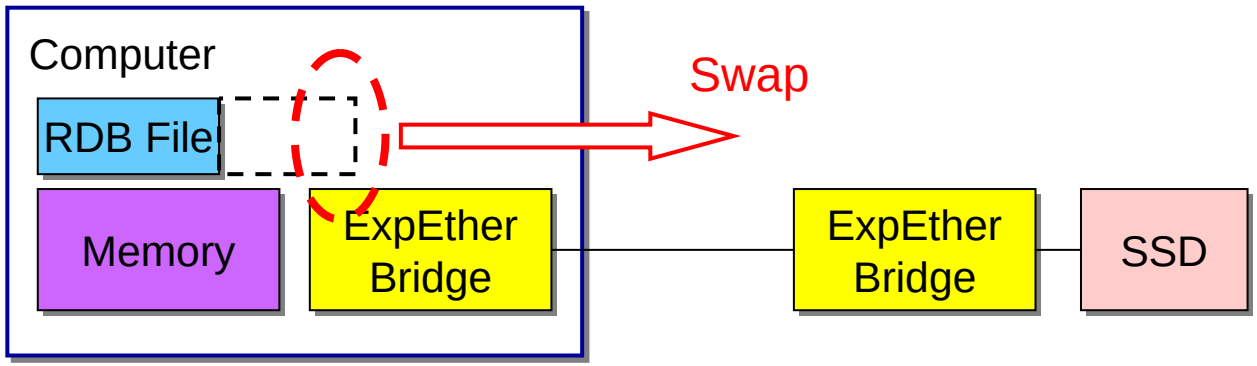
➡ *Increasing number of requests enhances performance close to host I/O slot*



System Evaluation: In-Memory Database

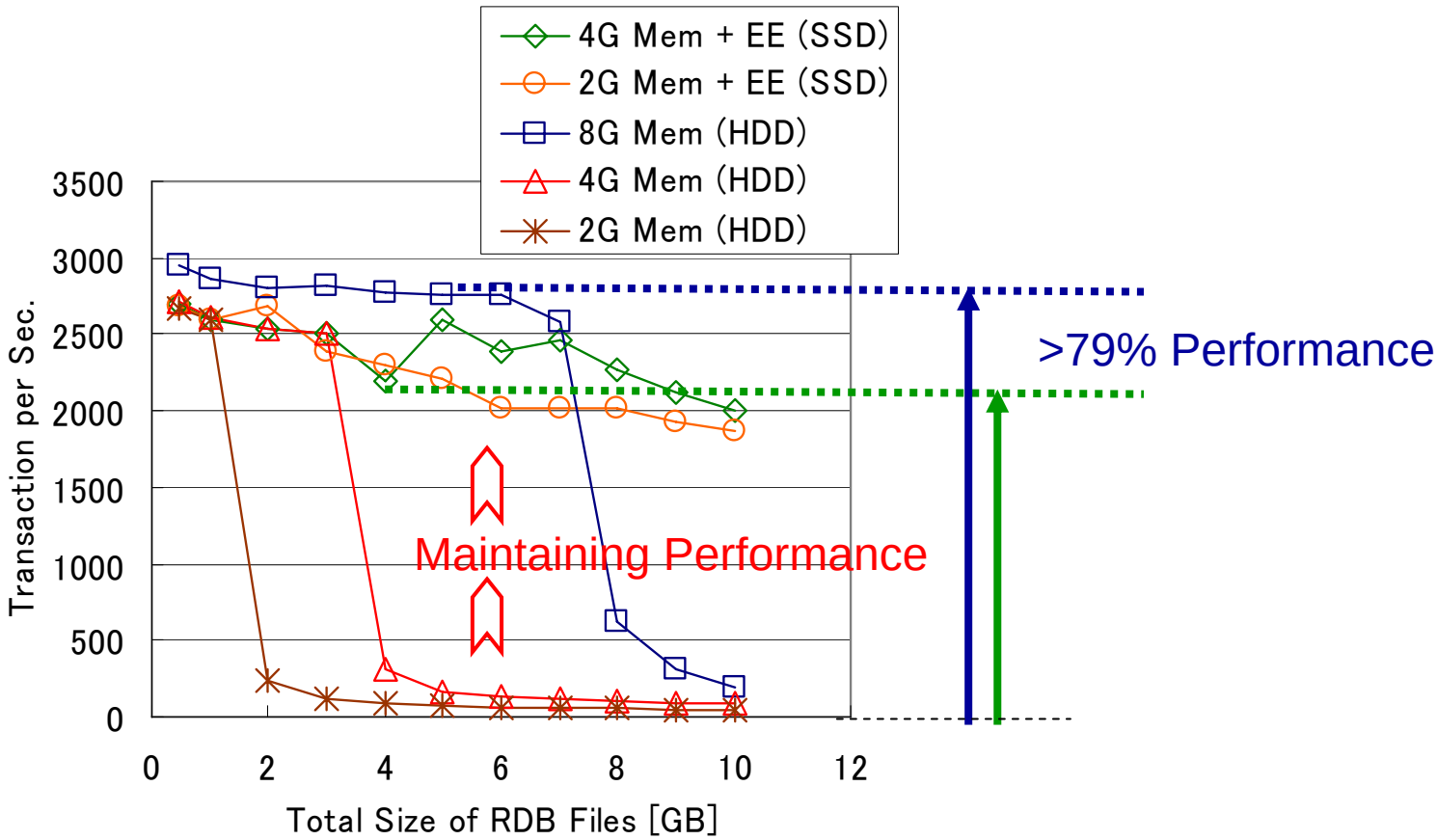
Placing RDB File on Ramdisk

RDB: postgresql 8.1
Bench: pgbench (TPC-B-like)
CPU: Intel Core 2 Quad
OS: CentOS 5.3 (Linux 2.6.18)
Ethernet: 10GbE
SSD: 16-GB Partition of Fusion IO 160 GB (Write Improve Mode)
#Client: 100
Transaction per Client: 1000



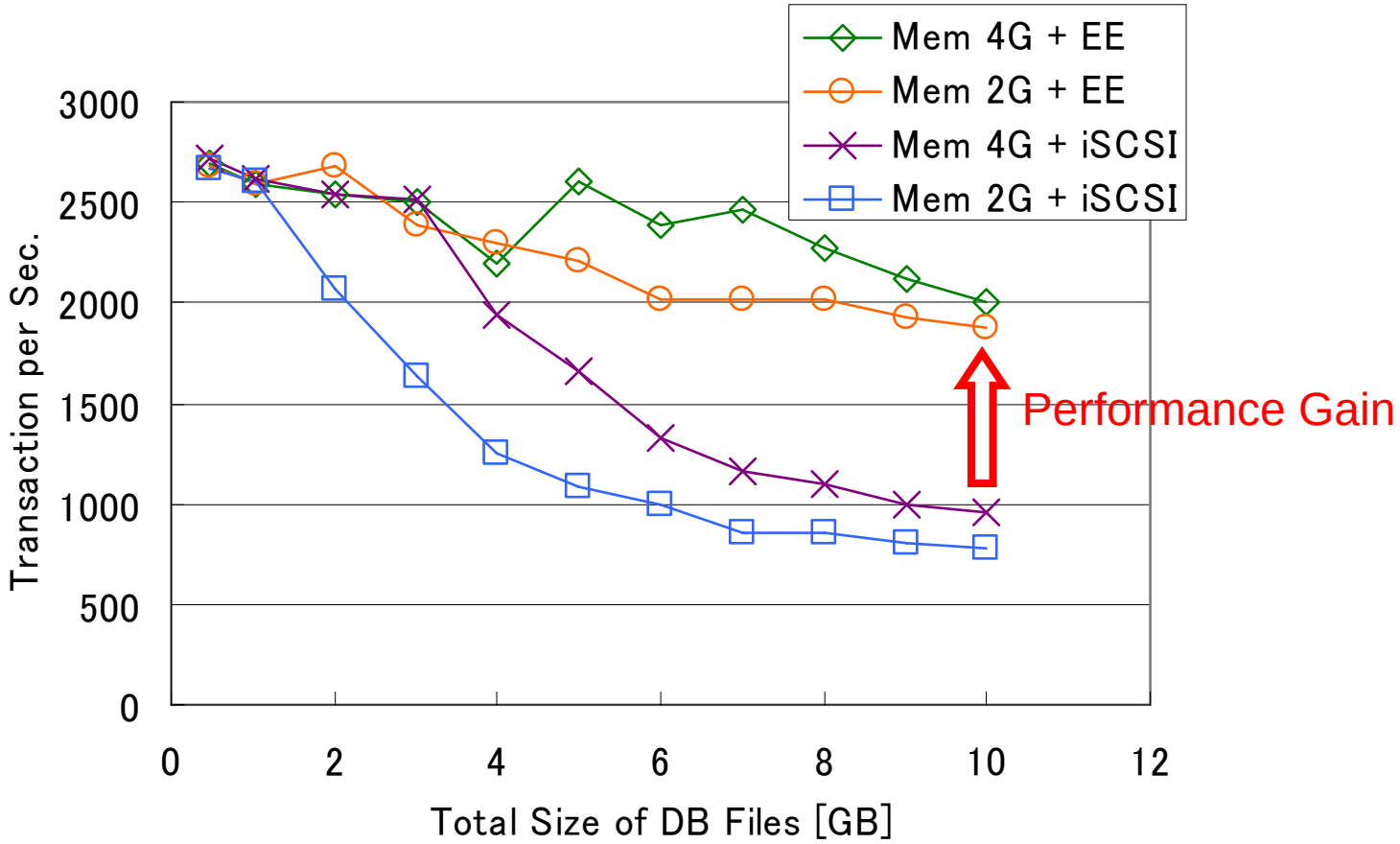
Scaling-Up beyond Main Memory

- Maintaining performance when DB files enlarged beyond system memory
- >79% performance of all-in-memory at 4G Mem + ExpEther case



Comparison with Conventional Protocol

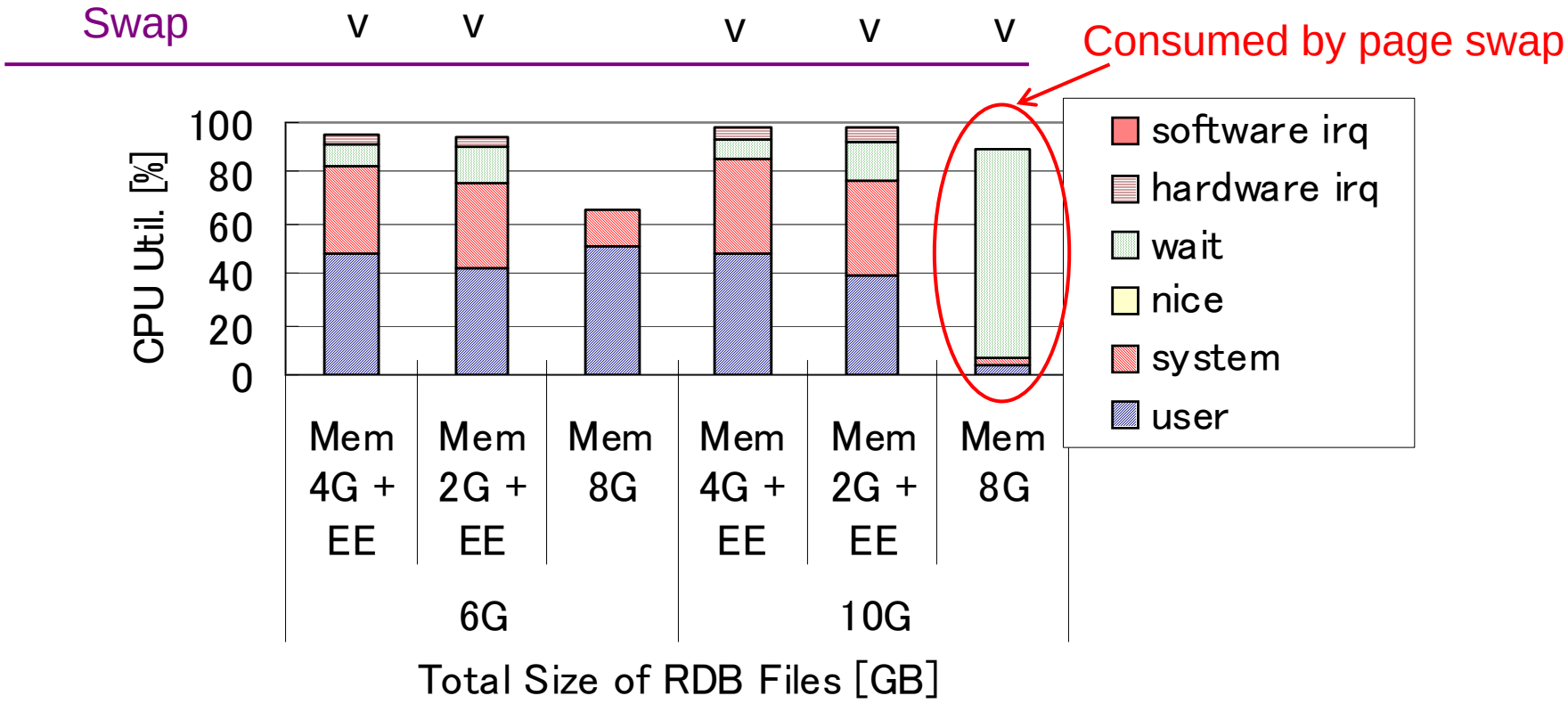
Proposal outperforms iSCSI by 139% at best case



[Note] iSCSI with TOE could not be evaluated by software bug.
Calculation indicates proposal outperforms it by 21%

Saving CPU Resource for Transaction Processing

High-speed swap saves CPU for user process



Conclusion

Adaptive Memory Expansion
with Ethernet-Attached SSD as High-Speed Swap Device

- ✓ Standard Components
  - Standard Ethernet and PCIe SSD
  - No software driver for Ethernet expansion
- ✓ High-Performance and Resource Share
  - PCIe DMA over Ethernet
  - Superior block-io performance than conventional protocol
  - PCIe hot-plug and remove
- ✓ Proven System Merits
  - Maintains database performance beyond system memory

Future Works

Simultaneous Share of SSD among multiple computers

- PCIe I/O virtualization emerges
- Efficient resource utilization
- High-speed data share

Solve Performance Bottleneck of Storage and Database System

- Network storage for system availability
- Performance bottleneck by network storage