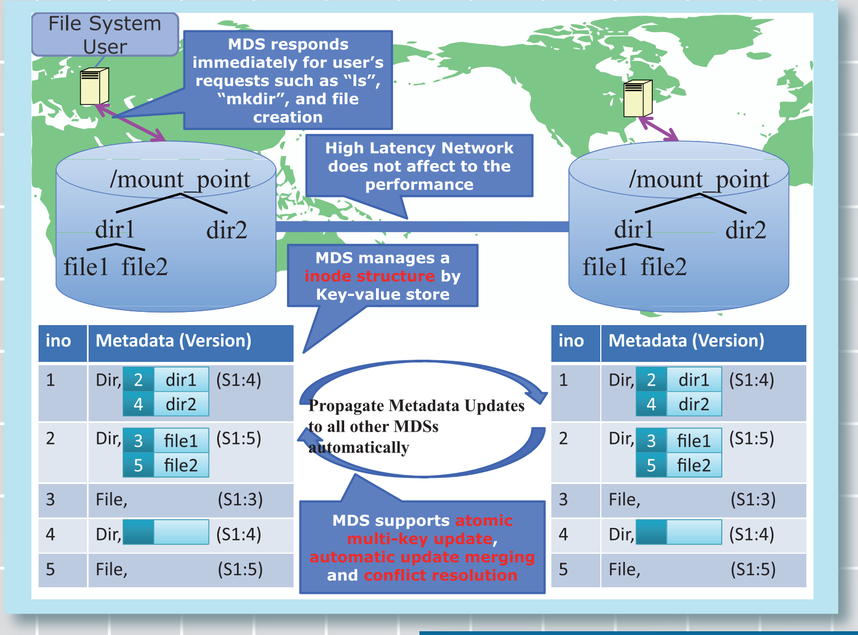
Center for Computational Sciences, University of Tsukuba
www.ccs.tsukuba.ac.jp

High Performance Computing Research

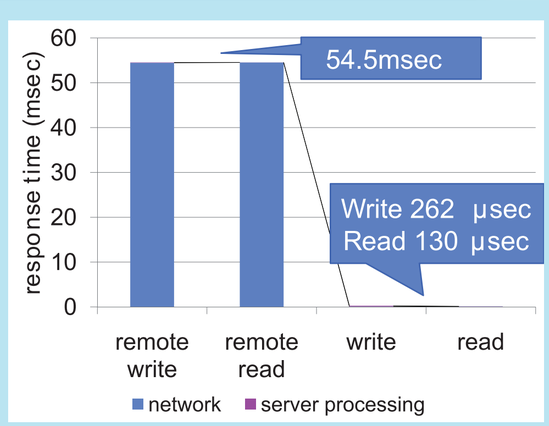
Distributed MDS for Wide-area File System

Multimaster key-value store that is enhanced to manage inode structure in eventual consistency



- Each **MDS responds immediately**
- Updates are propagated soon
- Update **conflicts** are **automatically resolved**
- Metadata **sync operations** ensure the update

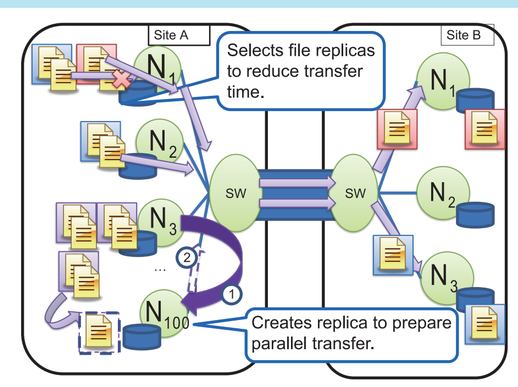
Evaluation Result



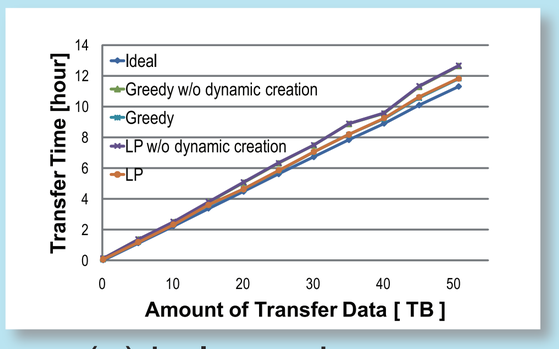
N-to-M Data Replication Scheduling

Data transfer scheduling for large set of data spread across cluster nodes to another cluster

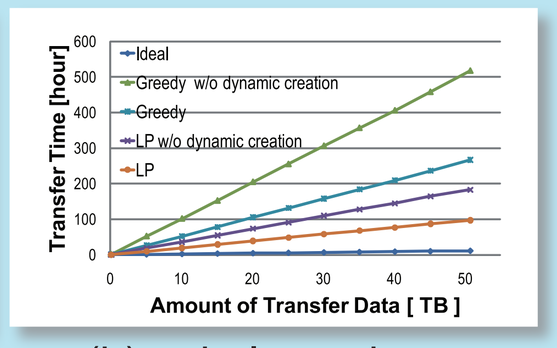
- Stable transfer rate to achieve maximum bandwidth of Long Fat Networks
- File replica selection by LP or greedy algorithm
- Dynamic data replication at sender side
- Do not access to the same storage at the same time



Evaluation Result



(a) balanced case



(b) unbalanced case

- Ideal transfers all data at the rate of 10Gbps
- LP provides good scheduling even in unbalanced case
- Greedy algorithm is good enough in balanced case

Dependable, Power-aware, and High-performance Network for Small-scale Cluster using PCI Express: PEARL

Concepts

To provide dependability, although redundancy on parallel and distributed systems is available, communication in such systems must be also dependable.

We proposed dependable network with redundancy, high-performance, and also power-awareness for embedded system: PEARL (PCI Express Adaptive and Reliable Link), and PEACH (PCI Express Adaptive Communication Hub) chip is under development.

- High bandwidth and low-power using PCI Express technology without modification of PCI-E standard
- Flexible control by embedded processor for dependability and power-awareness

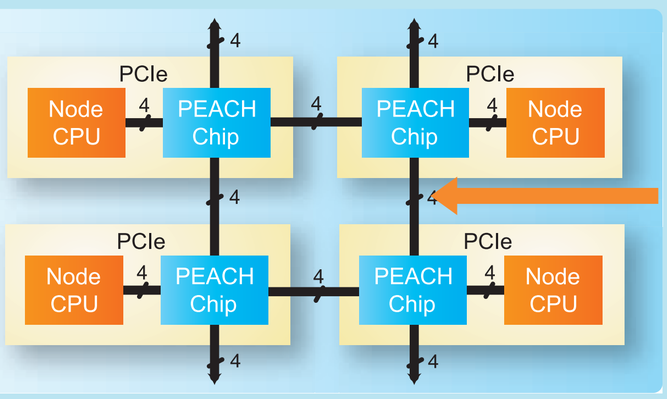
PEACH chip & board

Overview of PEACH chip

- Embedded CPU: M32R (Renesas Tech., 4core, SMP)
- PCI Express Gen 2, x4 lanes (20Gbps) x 4 ports
- DMA engine for each PCI Express port
- Process: 45nm rule

Overview of PEACH board

- PCI Express x4 host adapter board
- 3 ports for PCI Express external cable
- Self-independent operation from host system
- API: MCAPI(Multicore Communications API) with multi-node extention, Socket for TCP/IP

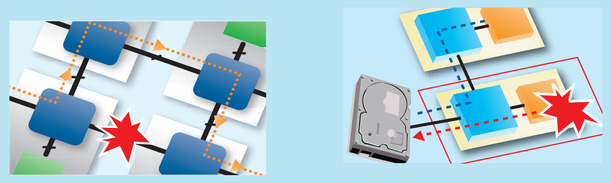


Overview of PEARL system

PCI Express external cable

Selection of the number of lanes and lane speed (power consumption ratio on PHY)

# of lanes / Lane speed	x1	x2	x4
Gen1	2.5Gbps (21)	5Gbps (38)	10Gbps (75)
Gen2	5Gbps (28)	10Gbps (50)	20Gbps (100)



(a) Fault link recovery (b) Device fail-over

Dependablity facilities of PEARL

This project is supported by JST/CREST program entitled "Computation Platform for Power-aware and Releiable Embedded Parallel Processing System" in the research area of "Dependable Embedded Operating Systems for Practical Use" (Oct. 2006 - Nov. 2011).