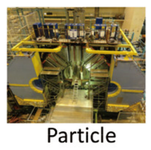
System Software for Post Petascale Data Intensive Science

http://www.hpcs.cs.tsukuba.ac.jp/project/crest-ppfs/en/

Background

Data –intensive Science

Next-generation DNA Sequencer



Particle Accelerator

Telescope

Post Petascale Computer

x1,000,000

Exabyte Data

Objective

Development of **System Software** for Data-intensive Computing to promote Data-intensive Science

This research is supported by **JST/CREST**
http://postpeta.jst.go.jp/en/

JST CREST Core Research for Evolutional Science and Technology
Japan Science and Technology Agency

Research Topics

- Runtime System
- File System Kernel Driver
- Distributed File System

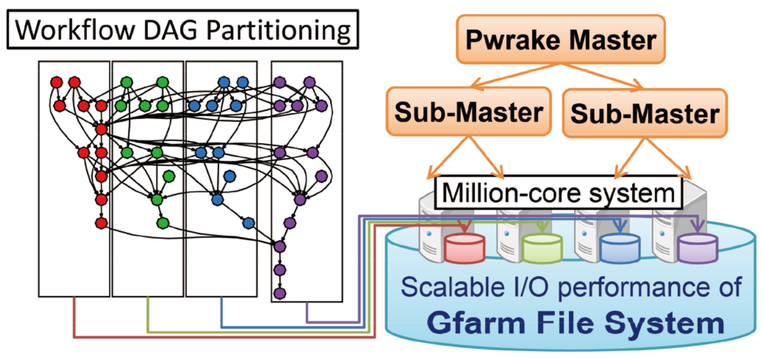
File System Kernel Driver is developed at The University of Electro-Communications UEC TOKYO

Runtime System

Pwrake: Scalable Workflow System

Pwrake is a workflow engine for data-intensive sciences with the following powerful features.

- ***Rake*** as a **Workflow Language** [HPDC 2010]
  - ◇ *Rake* is a DSL for a powerful build tool in *Rake* , which enables to describe complex scientific workflows
- **Scalable I/O Performance** based on **File Locality.**
  - ◇ Workflow scheduling to minimize data transfer based on **Multi-Constraint Graph Partitioning** [CCGrid 2012].
- **Post Petascale system** is the next target of Pwrake.
  - ◇ **Hierarchical Pwrake System** manages **100M tasks** executed on **one million cores.**



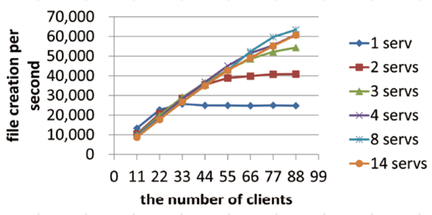
Distributed File System

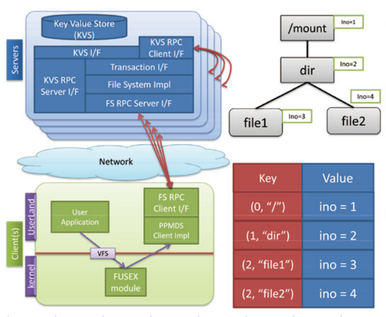
PPMDS: A Distributed Metadata Server

PPMDS is a distributed metadata management system for a distributed file system of post-peta scale super computers.

Fine-grained parallelization, the system manages directory namespace **in parallel** by ordered key-value stores. The key consists of **a pair of a parent inode number and a file name.** The value is the metadata. The largest granularity of locking is a key-value pair.

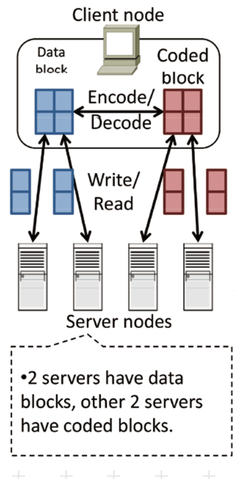
Nonblocking transactions across multiple key-value servers based on Dynamic STM to update multi key-value pairs transactionaly.



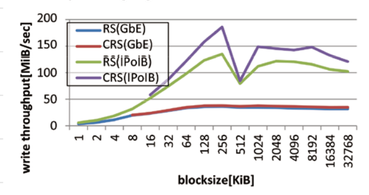


Redundant data store across nodes

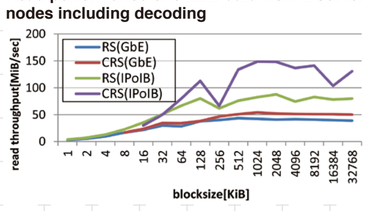
We investigate the performance of reliable write and read operations using erasure coding across multiple nodes



Write performance of 64MiB data to 4 server nodes by changing the block size



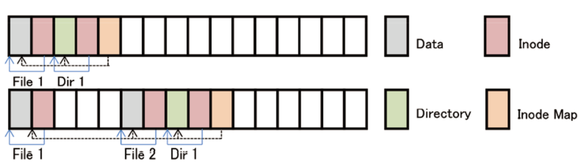
Read performance of 64MiB data from 2 server nodes including decoding

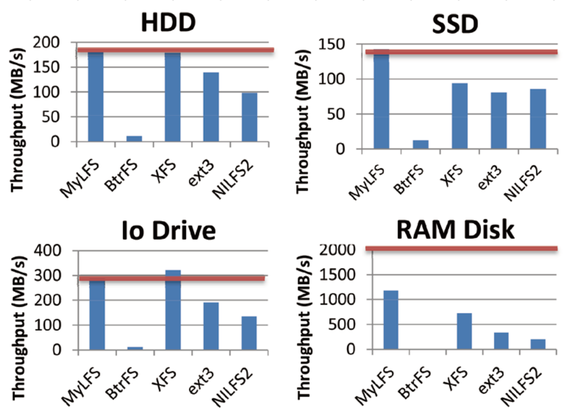


File System for next-generation storage

We are developing a prototype file system based on log structured file system to achieve maximum performance

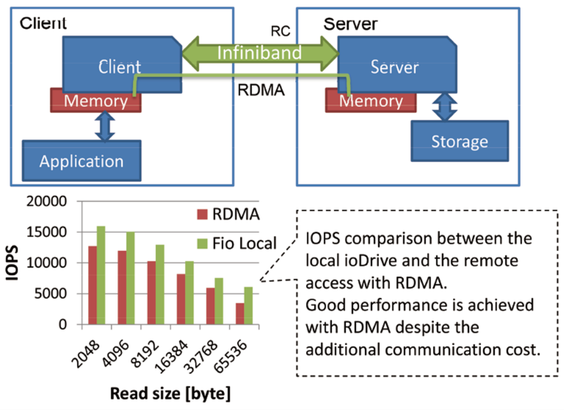
Initial Performance result of MyLFS for the sequential write





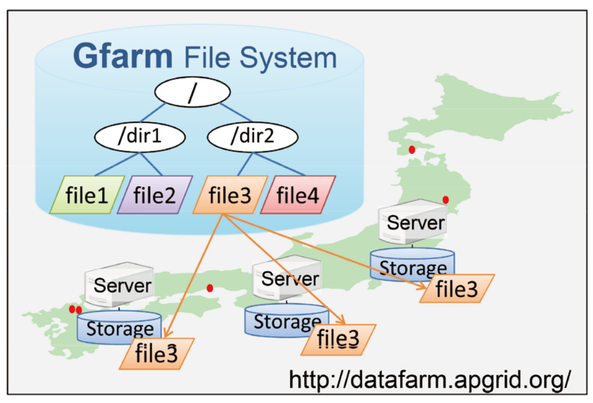
File Access with Infiniband RDMA

Infiniband is a high throughput and low latency network. The objective of this research is to improve the performance of distributed file systems by using RDMA (Remote Direct Memory Access). This function has a significant effect to reduce the overhead of communication.

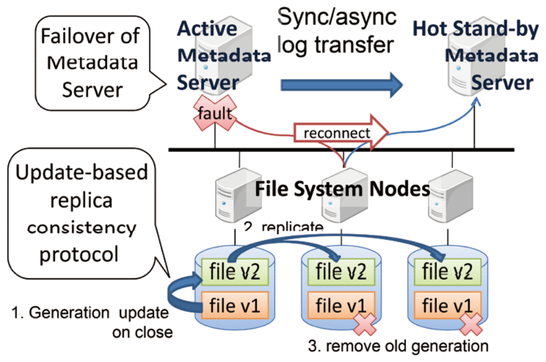


Gfarm: Wide-area Distributed File System

http://sf.net/projects/gfarm/



Failover and File Replication Mechanism



HPCI Storage – Nation-wide file system

Gfarm File System realizes 21.2 Petabyte Nation-wide File System (HPCI Storage), which can be initially accessed from nine national universities and the "K" computer site. The operation began from September 2012.