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

System Software for Post Petascale Data Intensive Science

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/

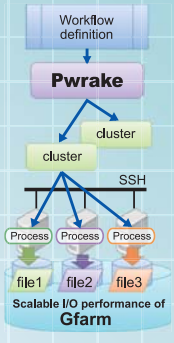
Research Topics

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

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

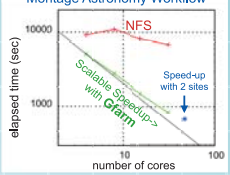
Runtime System

Pwrake: Scalable Workflow System



- Workflow Language : ***Rake***
  - Allows users to write complex scientific workflows
- Scalable I/O Performance by :
  - Adaptive scheduling
  - File locality
  - Graph partitioning
- For the use of 1 million cores :
  - Hierarchical management
  - Autonomous execution

Scalability of Pwrake
Montage Astronomy Workflow



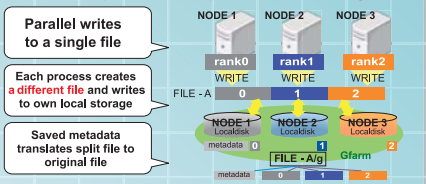
MPI-IO / Gfarm

MPI-IO/Gfarm is an MPI-IO implementation for the Gfarm file system designed to achieve scalable parallel I/O performance.
Problem : Poor performance of parallel writes to a single file

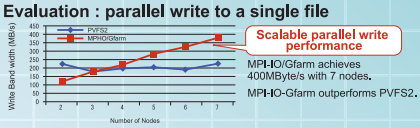
Parallel writes to a single file

Each process creates a different file and writes to own local storage

Saved metadata translates split file to original file



Evaluation : parallel write to a single file

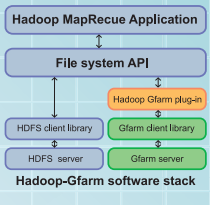


Scalable parallel write performance

MPI-IO/Gfarm achieves 400MByte/s with 7 nodes.
MPI-IO-Gfarm outperforms PVFS2.

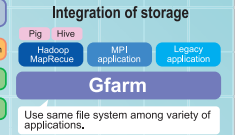
Hadoop / Gfarm

We have developed a Hadoop-Gfarm plugin which enables Hadoop Mapreduce applications to access a Gfarm file system.



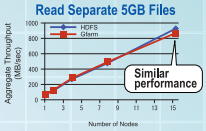
Problem : HDFS cannot be used from applications other than mapreduce

Integration of storage

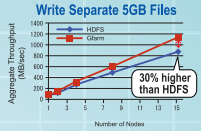


Use same file system among variety of applications.

Read Separate 5GB Files

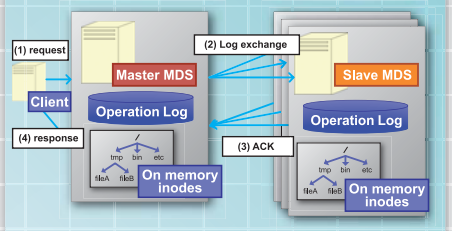


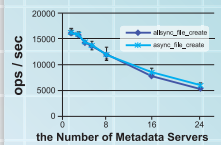
Write Separate 5GB Files



Distributed File System

Reliable Distributed Metadata Servers



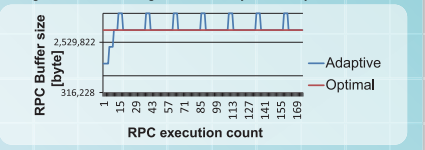


We implemented and evaluated a log-exchange based Master-slave Metadata server that evaluates the relationship between various trade-offs such as performance, consistency, durability, reliability, and availability.

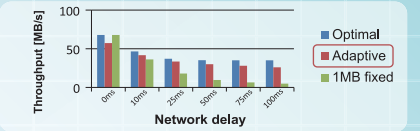
Optimization of Remote File Access

Dynamic optimization method for remote file access systems under the high network latency environment.

Change the transferring block size dynamically:

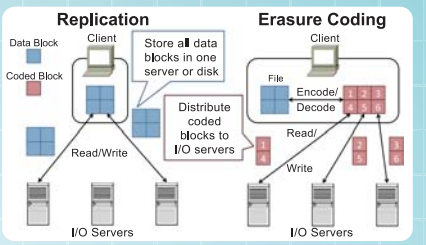


Performance improvement: (Adaptive vs. 1MB fixed)



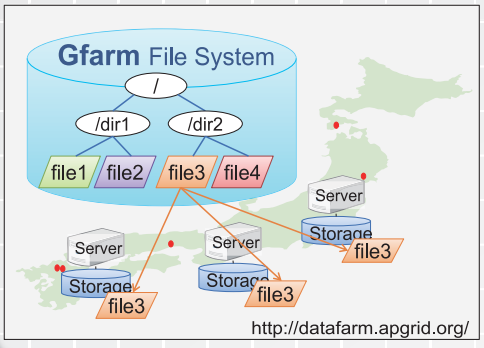
Redundant Data Storage Method

We plan to propose a redundant data storage approach through the use of erasure coding scheme, while keeping high I/O performance

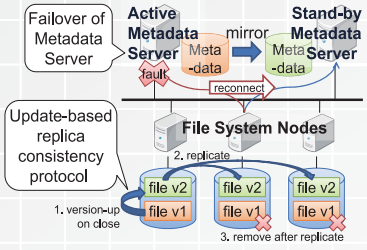


Both of schemes can tolerate one failure of data access, although capacity overhead of replication is high (at least 2 copy of whole data blocks). On the other hand, if we adopt erasure coding,
- capacity overhead is smaller than replication.
- computation costs of encoding/decoding are needed.

Gfarm: Wide-area Distributed File System



Gfarm Technology for Failover and File Replication



This research is supported by the **RENKEI (Resources linkage for e-Science)** project sponsored by MEXT of Japan.
RENKEI aims to federate (= renkei) e-Science communities through research and development of middleware technologies.
RENKEI website: http://www.e-sciren.org/

METI Next Generation Green IT infrastructure: Accountability for Cloud Computing

As part of a 3 year effort to create highly reliable and accountable cloud storage platforms (PI : NTT Communications), The Ministry of Economy, Trade and Industry(METI) has awarded The University of Tsukuba a grant to research exascale cloud storage infrastructure technologies capable of federating thousands of individual clouds.