Designing monitoring system for HPCI shared storage and statistics collected

Hidetomo Kaneyama
Riken Center for Computational Science
Japan
hidetomo.Kaneyama@riken.jp

Hiroshi Harada
Riken Center for Computational Science
Japan
hiroshi.harada@riken.jp

Asuka Ohta
PARARS
Japan
ask@parars.riken.jp

Hiroyuki Nakatsuka
Metahack Japan
Japan
hiroyu-n@metehack.jp

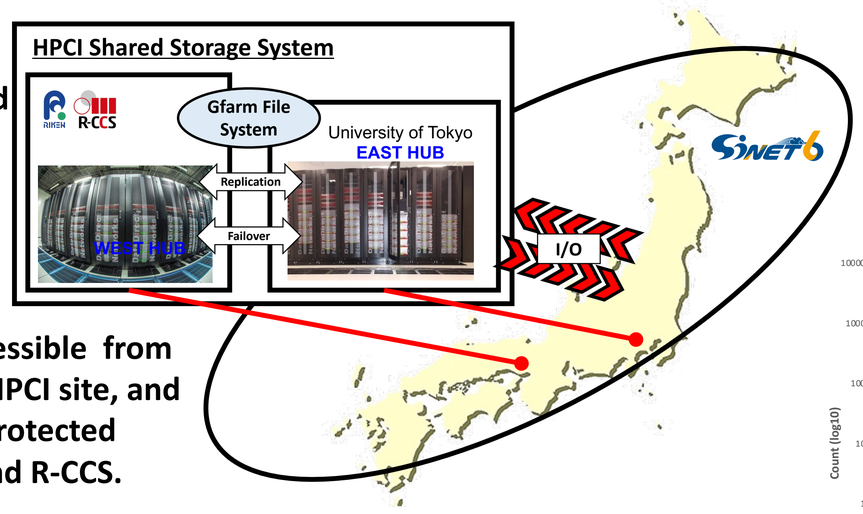
INTRODUCTION

Designed and built a monitoring environment used in the operation of HPCI Shared Storage. This has improved the acquisition of statistical information, enabling more detailed storage utilization information to be acquired in real time and used to provide it to users and notify them of failures.
The acquired statistical information shows that there is a lot of SMALL MULTIPLE research data and old archive data. These data may be straining the meta-access load and capacity that could be provided and utilized by other users.
To remedy these problems, we are currently considering compression functions and provision of detailed information.

What is HPCI Shared Storage ?

HPCI Shared Storage is a storage infrastructure for HPCI operated by R-CCS and the University of Tokyo.
This storage operates a network file system called "Gfarm" developed by Prof. Tatebe of the University of Tsukuba.
I/O to this storage can be performed from the global network. It is accessible from the Login nodes of each HPCI site, and the data is redundantly protected by University of Tokyo and R-CCS.
Project Site:

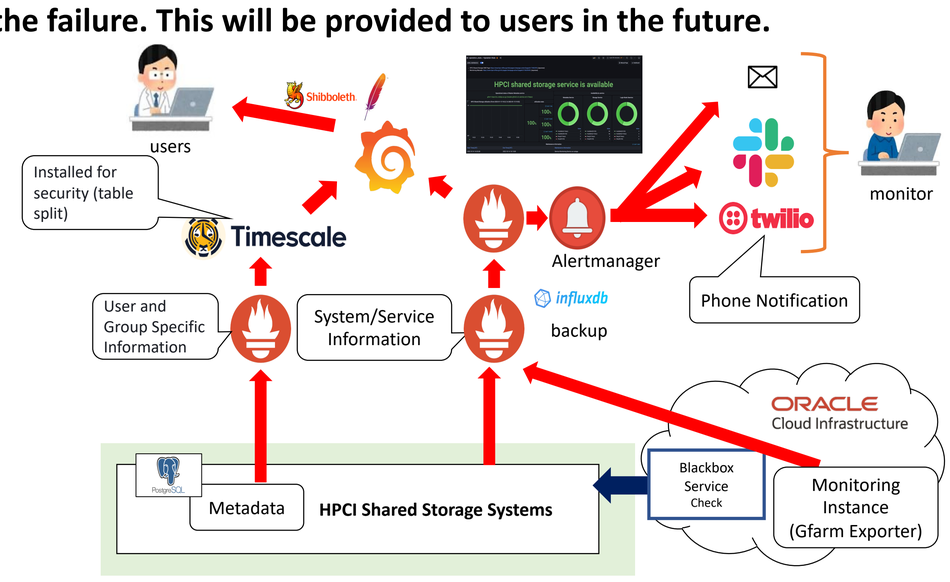
- https://www.hpci-office.jp/info/pages/viewpage.action?pageId=111380786



Designing Monitoring System for HPCI Shared Storage

The monitoring environment is mainly built using Prometheus and Grafana.

- ☐ The User and Group statistics is obtained directly from Gfarm metadata information, and these stats are provided to users via Timescale DB for security.
- ☐ Perform black box testing using OCI and get availability and transfer delays from the test results.
- ☐ Failure notifications are sent in several ways, depending on the severity of the failure. This will be provided to users in the future.



Data Statics collected by HPCI Shared Storage

The statistical information obtained from the monitoring system will be used to identify future issues and to add and improve functions of the HPCI shared storage.
In Gfarm, since meta-information is stored in PostgreSQL and information can be obtained in SQL language, it is easy to obtain these information on a regular basis.
As shown in the table (1), the number of files less than 10 MB accounts for 67% of the data stored in the HPCI shared storage. The transfer rate of these data tends to be slow due to meta-access bottlenecks.

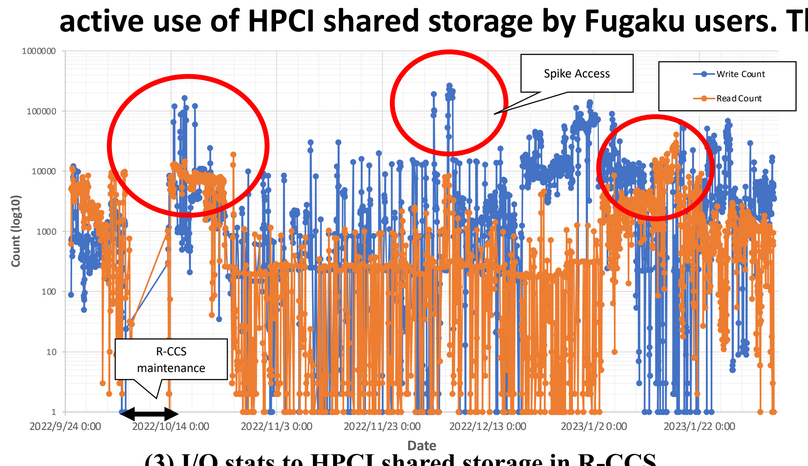
FileStats	File Size		File Number	
	Value	Rate	Value	Rate
- 10MB	0.21PB	1.00%	119 M	67.55%
10MB - 100MB	0.87PB	4.16%	26 M	14.94%
100MB - 1GB	6.25PB	29.89%	26 M	14.91%
1GB - 10GB	7.25PB	34.68%	4 M	2.49%
10GB - 100GB	3.31PB	15.83%	177 K	0.10 %
100GB -	3.02PB	14.43%	8579	0.00 %

(1) Capacity and number of files for storage by access period

Access Period	File Size		File Number	
	Value	Rate	Value	Rate
- 3month	4.08PB	19.74%	29M	15.68%
3m - 1year	8.08PB	39.09%	76M	41.72%
1y - 3y	4.27PB	20.66%	44M	24.23%
3y - 5y	3.19PB	15.43%	18M	9.82%
5y -	1.05PB	5.08%	16M	8.56%

(2) Capacity for storage and number of files by file size

As the table in (2) shows, user usage as of 01/01/2022 was robust, with more than 50% of all files accessed between 2021 and 2022. We assume that this is due to the active use of HPCI shared storage by Fugaku users. The access graph in (3) shows that access information has rapidly increased after the maintenance of R-CCS. In addition, data older than 3 years accounts for about 20% of the total. If these file are left without being accessed, they consume resources needlessly.

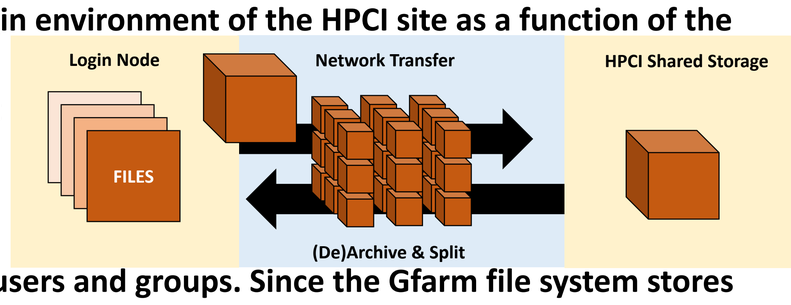


(3) I/O stats to HPCI shared storage in R-CCS

Future GOAL

- ☐ Measures for Small Many data

  We are planning to develop a mechanism for compression, data archiving (Tar), and splitting from the login environment of the HPCI site as a function of the Gfarm file system.



- ☐ Countermeasures against dark files

  Detailed usage statistics will be added to the information provided to users and groups. Since the Gfarm file system stores the md5 of each file, we plan to use this information to provide users with data information on the same md5 and the same file size.

- ☐ Measures for I/O Spike

  I/O spikes occur after maintenance on supercomputers such as Fugaku. To resolve this, transfers need to be balanced. We are considering and investigating providing workflow and buffer storage to users for this purpose.

REFERENCE

Osamu T. Gfarm Grid File System. 2010.7 NEW GENERATION COMPUTING SPRINGER