



SDC¹⁸

September 24-27, 2018
Santa Clara, CA

www.storagedeveloper.org

Experiments in Storing Scientific Research Data in Cloud Cold Storage Services

Hiroshi Yoshida
Center for Cloud Research and Development
National Institute of Informatics

Agenda

- ❑ Overview of cloud cold storage services
- ❑ Overview of the experiments
- ❑ Results of the experiments
 - ❑ Basic benchmark
 - ❑ Case study (1): High-energy physics data
 - ❑ Case study (2): Astronomy data
- ❑ Conclusion

Overview of Cloud Cold Storage Services



What is Cold Storage?

- ❑ Definition by SNIA [SNIA Dictionary <https://www.snia.org/education/dictionary>]

Data storage device, system, or service used to store cold data at a cost that is at least an order of magnitude less than the cost of primary storage. Cold Storage features large capacity, energy saving, and long-term data preservation, in order to achieve low cost rather than performance.

- ❑ Cold data: Data that are accessed infrequently.

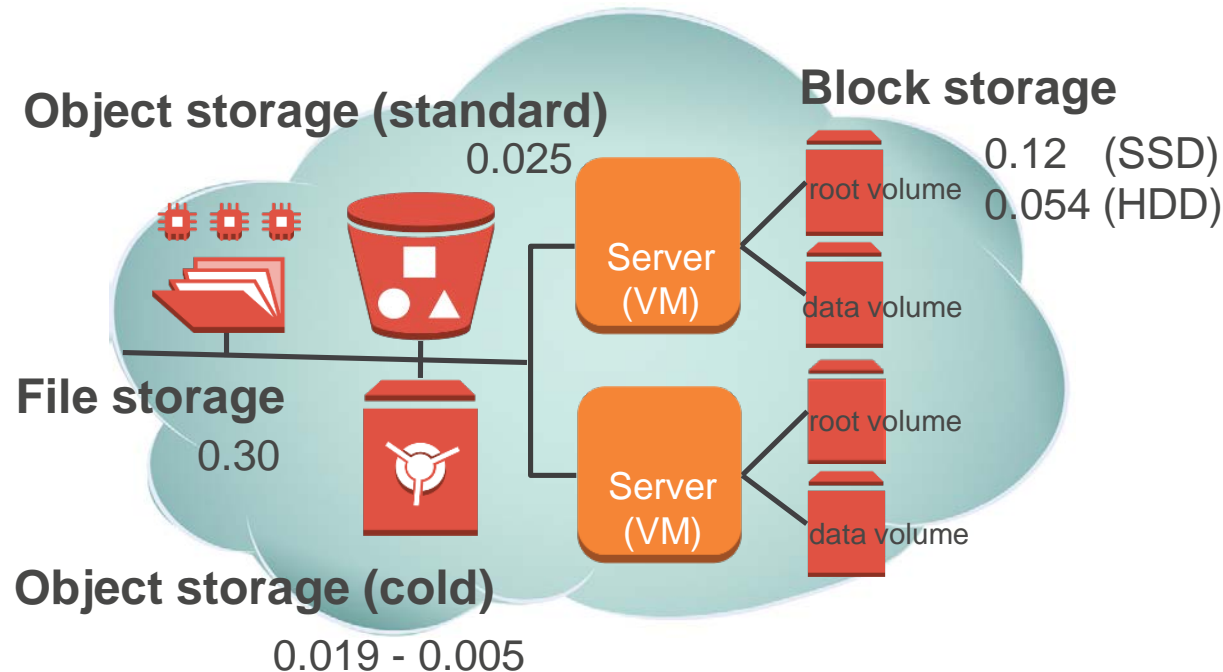
- ❑ Technology

- ❑ Device: HDD, tape, optical disk, etc.
 - ❑ System: distributed storage, software-defined storage, etc.
 - ❑ Service: cloud service

Storage Services of IaaS

□ Typically an IaaS provides 3 types of storage

- Block storage
- Object storage (standard/cold)
- File storage



Price unit: USD/(GB*month)
[as of July 1, 2018]
AWS **Japan region**

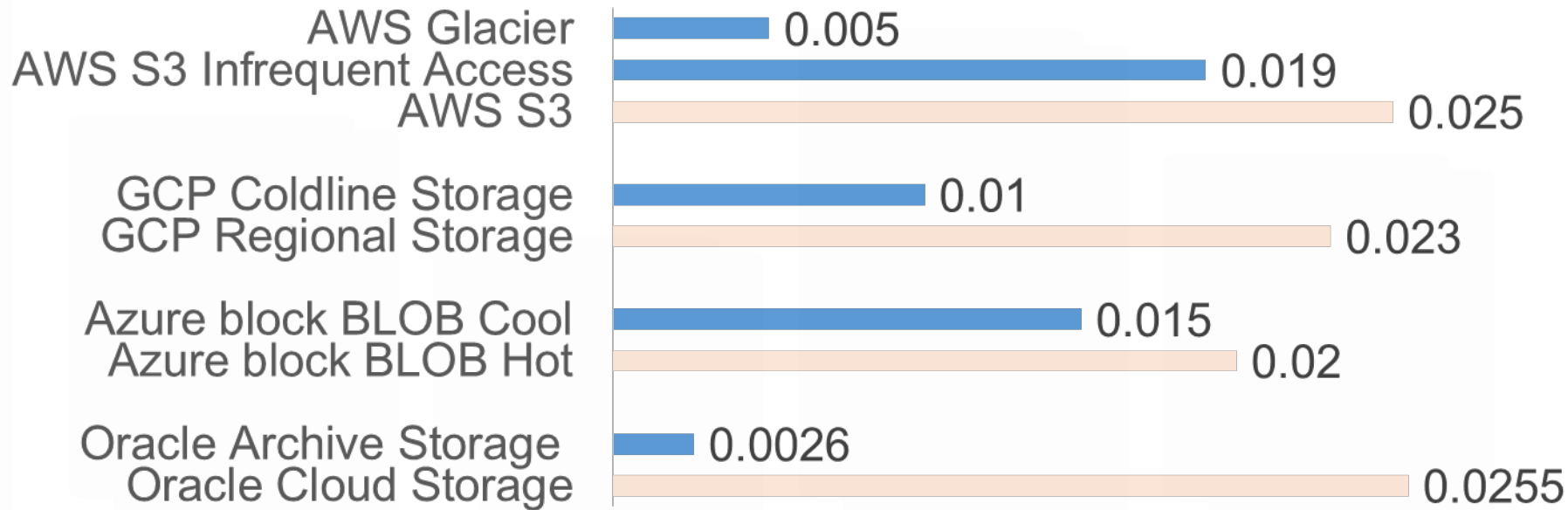
Cold Storage Services of IaaS (1)

- ❑ Object storage services
- ❑ Designed mainly for long-term preservation of infrequently accessed data
- ❑ History
 - ❑ AWS took a lead position after launching “Glacier” in 2012.
 - ❑ Other major providers entered the market and by 2015 a highly competitive market had been established.

Cold Storage Services of IaaS (2)

- ❑ Storage charge per GB*month is relatively inexpensive compared to standard object storage services.
- ❑ Drawbacks in exchange for inexpensive storage charge
 - ❑ Restoration process is time consuming (requires hours)
 - ❑ Extra charge for data retrieval
 - ❑ Minimal retention period
 - ❑ Limited performance, or extra charge for additional performance.
 - ❑ Reduced availability

Pricing of Cold Storage Services



Price unit: USD/(GB*month) [as of July 1, 2018, **Japan region** except Oracle]

- ❑ 2/3 - 1/10 less expensive compared to standard object storage services

Taxonomy of Cold Storage Services

Standard object storage services		AWS Oracle GCP Azure	S3 Cloud Storage Regional Storage Block BLOB Hot
Cold storage service	Separate services from standard services	AWS Glacier <ul style="list-style-type: none"> Restore processing required (time depends on charge) Extra read charge Minimal retention period: 90 days Undefined availability 	
	Optional features of standard storage services	AWS S3 Infrequent Access (IA) <ul style="list-style-type: none"> Extra read charge Minimal retention period: 30 days Lower availability than S3 Minimum data length is rounded up to 128KB 	
		GCP Coldline <ul style="list-style-type: none"> Extra read charge Minimal retention period: 90 days 	
		Azure block BLOB Cool <ul style="list-style-type: none"> Extra read charge Minimal retention period: 30 days Extra write charge (based on data quantity) charge 	
	Intermediate	Oracle Archive Storage <ul style="list-style-type: none"> Restore processing required (up to 4 hours) Extra read charge Minimal retention period: 90 days Some incompatibility with the standard service (e.g., handling of large objects) 	

Overview of the Experiments



Possible Cold Storage Use Cases in Academic Field

- ❑ Research areas which process a large amount of experimental or observational data
 - ❑ Genome
 - ❑ Experimental physics (high-energy physics)
 - ❑ Astronomy
 - ❑ Seismology etc.
- ❑ Open research data
 - ❑ Publish not only papers but also related research data
- ❑ Compliance
 - ❑ Preserve data used for experiments for a long time period
 - ❑ Assure the repeatability of experiments

Target of Experiment

- ❑ Objective

“Is it possible to adopt cloud cold storage to store a large amount of scientific research data for a long time?”

- ❑ Reduction of storage management labor and TCO is expected.
- ❑ As very few precedents exist, feasibility in terms of performance, manageability, and cost remains unknown.

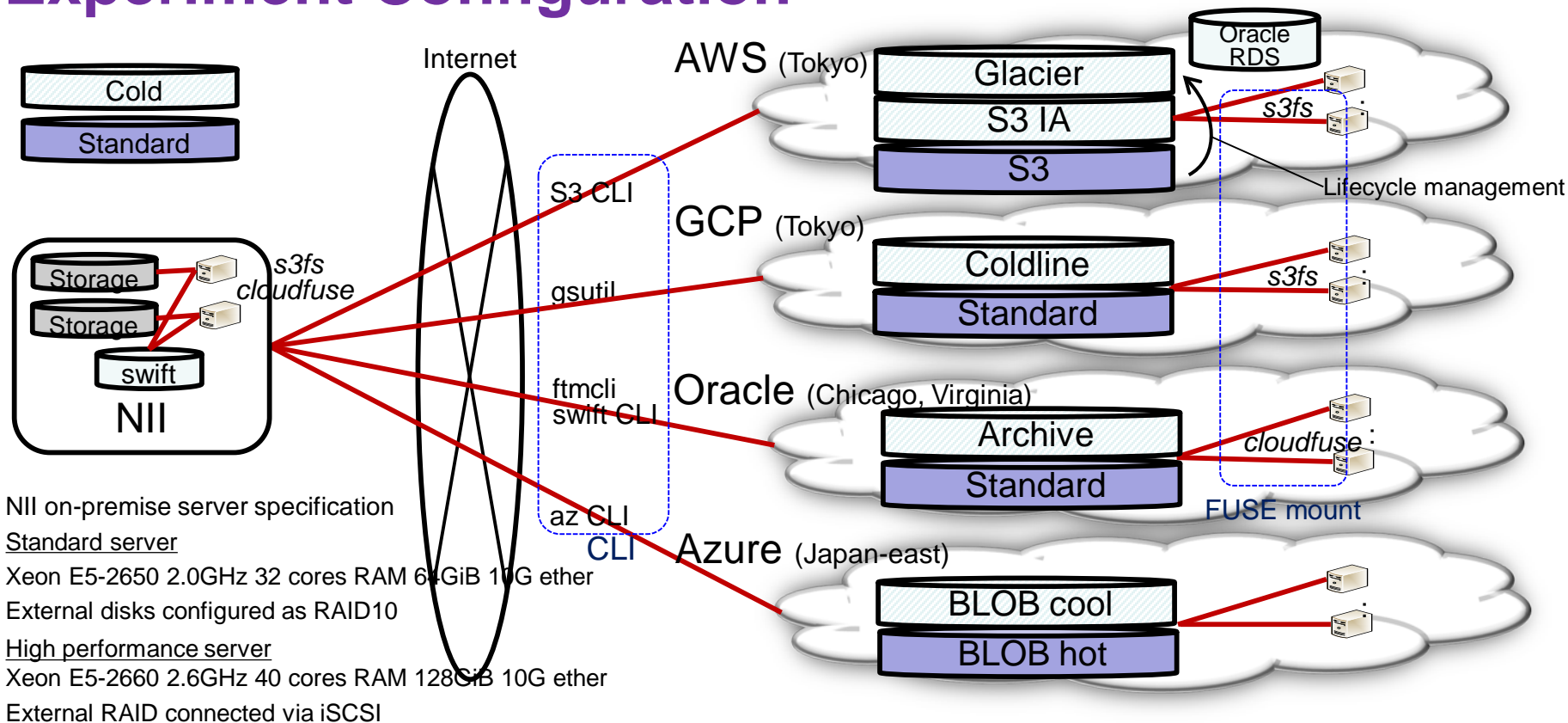
- ❑ Aim of experiments

To acquire practical information to determine the suitability of storing research data in cloud cold storage, and to design an overall data storage architecture.

➡ Experiments in cold storage services of multiple commercial public clouds

- ❑ Basic benchmark tests including storing up to 1PB data.
- ❑ Case study analyses using actual research data and applications.

Experiment Configuration



Case Study Scenario

- ❑ Store actual research data
- ❑ Access the data by actual research applications in the close manner of actual research activities
- ❑ Analyze results in terms of performance, cost, and manageability
- ❑ Experimented research area
 - ❑ High-energy physics
 - ❑ Astronomy
- ❑ Fixed-point observations considering the fast evolution of cloud services
 - ❑ Comparison of research data upload performance in 2017 and 2018
- ❑ Trial to store large amount of research data up to 1PiB

Experiment Policy

- ❑ Try to perform tests according to what an average user would do.
 - ❑ Adopt the usages which cloud service providers provide and/or recommend as far as possible.
- ❑ Adopt CLIs, GUIs, and libraries provided by the service providers as much as possible rather than developing original programs using API
 - ❑ Parallel object upload CLIs are heavily used.
 - ❑ Default settings of CLI tuning parameters such as number of parallel threads are used at first, and tuned only in the case of trouble.

Parallel Object Upload CLI

AWS S3 IA	S3 CLI	Provided by AWS
GCP Coldline	gsutil	Provided by Google
Azure Cool	az CLI	Provided by Microsoft
Oracle Archive	ftmcli	Provided by Oracle
	Swift CLI	OSS (Compatible with Swift)

- ❑ Cloud service providers usually provide proprietary CLIs to enable users to manage their resources in the cloud services.
- ❑ Some cloud providers provide compatibility with other cloud platforms such as S3 or OpenStack.

File Access API for Cold Storage Services

- ❑ Features of object storage and cold storage
 - ❑ 2-tier architecture: containers (or buckets) and objects in a container
 - ❑ Accessed through REST API
- ❑ Many existing research applications use the POSIX-compatible file API.
- ❑ To avoid modifying existing research applications, open-source programs which enable mounting a container as a file system are used in the experiment.

AWS S3/S3 IA	A bucket is mounted as a FUSE* through OSS “s3fs”.
GCP Coldline	ditto (because of compatibility with S3)
Azure Cool	N/A
Oracle Archive	A container is mounted as a FUSE* through OSS “cloudfuse”.

* Filesystem in user space

Restore Operation of Cold Storage Services

- ❑ Restore operations must be executed before stored data can be read for certain cold storage services.

Cold storage		Restore
AWS	S3 IA	
	Glacier	❑
GCP	Coldline	
Oracle	Archive	❑
Azure	BLOB cool	

- ❑ An example of restore processing (AWS):
Restore S3 objects migrated to Glacier through lifecycle management function

- ❑ Issue restore request

```
s3cmd restore --recursive --restore-days=1 --restore-priority=bulk s3://bucket-name/
```

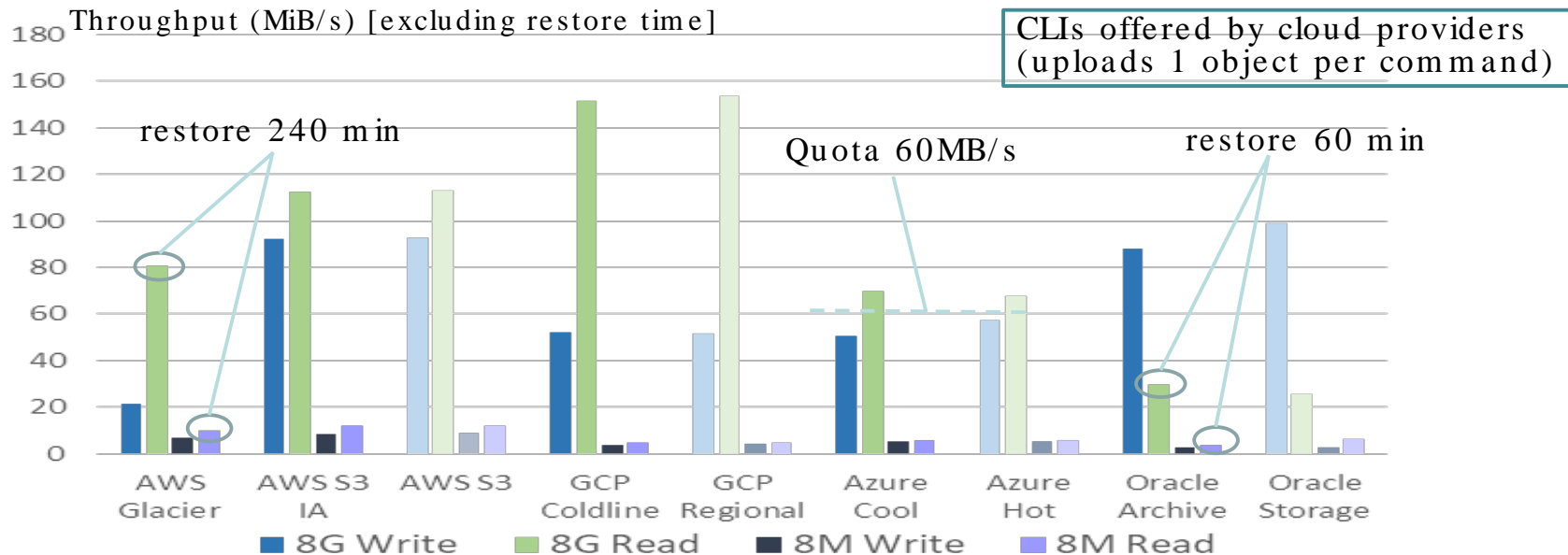
- ❑ Poll for completion

```
for key in `aws s3api list-objects --bucket bucket-name --output json |
jq -r '.Contents[].Key'`; do
    status=`aws s3api head-object --bucket bucket-name
--key $key --output json | jq -r .Restore | awk '{print $1}'`
    if ["$status"="$true"] then continue
    else ... /*escape because the restore operation is still in progress*/
fi
done
```

Results of the Experiments



Basic Benchmark: Object Access Performance in 2017



- ❑ Various performance characteristics (e.g., difference between write and read)
- ❑ Better throughput for large objects
- ❑ No difference in performance if cold storage is an option for standard storage.
- ❑ Restore operations are usually faster compared to specified maximum values.

Alternatives in Performance Measurement

- ❑ In some cloud services, no performance difference is observed between standard object storage and cold storage.
- ❑ In those cloud services, the performances of standard object storage are sometimes measured instead of cold storage in the case studies to save the experiment cost.

❑ AWS	S3 IA	➡ S3
❑ GCP	Coldline	➡ Regional
❑ Azure	BLOB cool	➡ BLOB hot

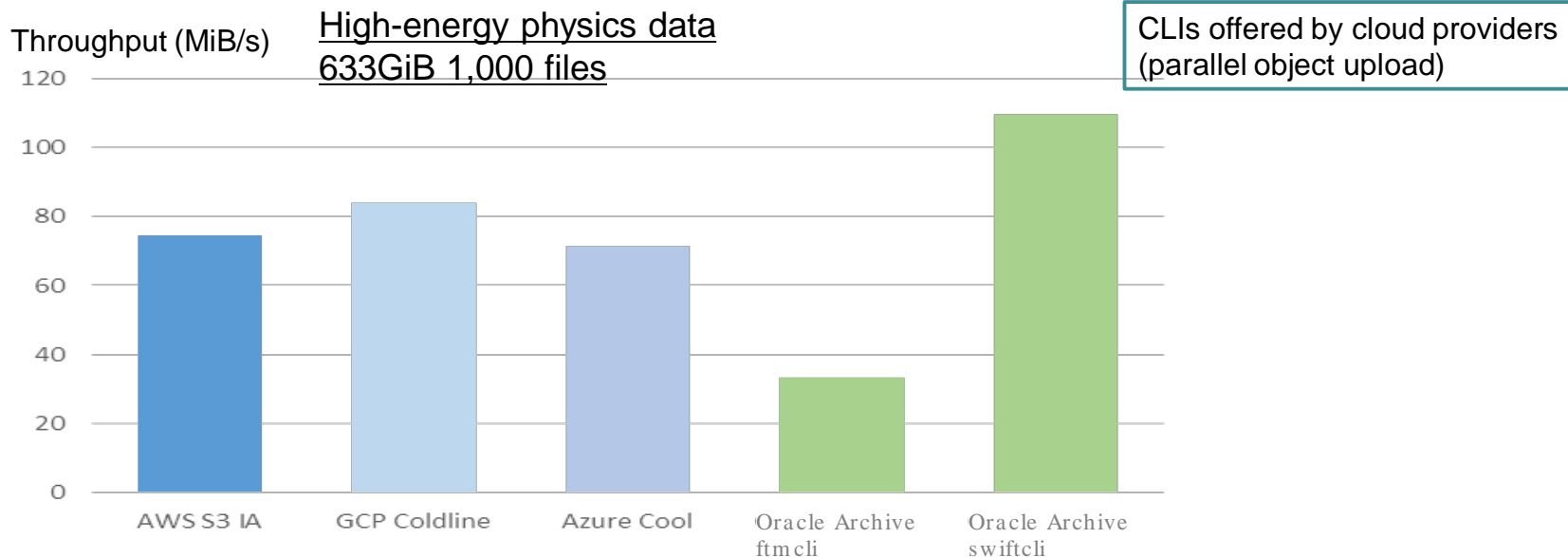
Research Data and Application

Research area	High-energy physics	Astronomy
Data contents	Belle II experiment physical simulation data	ALMA radio telescope Observation/analysis data
Quantity	633GiB, 1,000 files	58.5TiB, 1,380,000 files
Size	Almost the same (600–700MiB)	Falls between smaller than 1MiB and larger than 100GiB (average: 44MiB)
Application	Read files through analysis support software environment “BASF2” (Belle II Analysis Framework 2)	Data management and retrieval by archive system “NGAS” (Next Generation Archive System)

Case Study (1): High-energy Physics Data

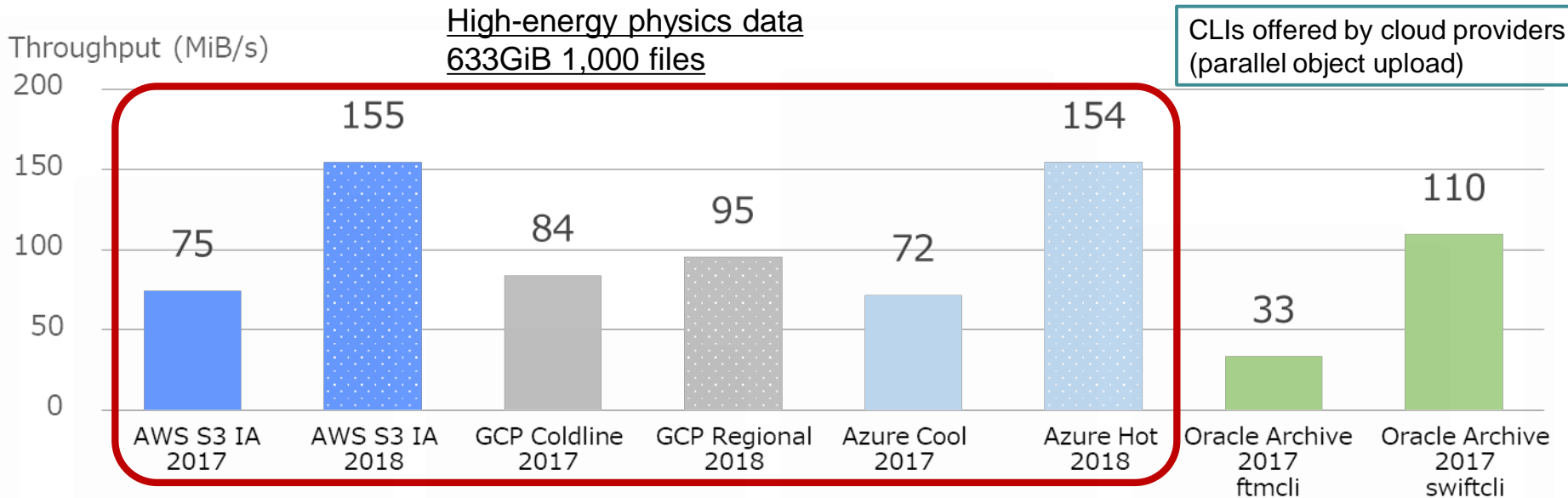
Research area	High-energy physics	Astronomy
Data contents	Belle II experiment physical simulation data	ALMA radio telescope Observation/analysis data
Quantity	633GiB, 1,000 files	58.5TiB, 1,380,000 files
Size	Almost the same (600–700MiB)	Falls between smaller than 1MiB and larger than 100GiB (average: 44MiB)
Application	Read files through analysis support software environment “BASF2” (Belle II Analysis Framework 2)	Data management and retrieval by archive system “NGAS” (Next Generation Archive System)

Upload Performance: Experiments in 2017



- ❑ Performance depended on the used CLI even in the same cloud.
- ❑ The data center location may affect performance.
 - ❑ AWS, GCP, and Azure: Japan
 - ❑ Oracle: US

Upload Performance: Experiments in 2018



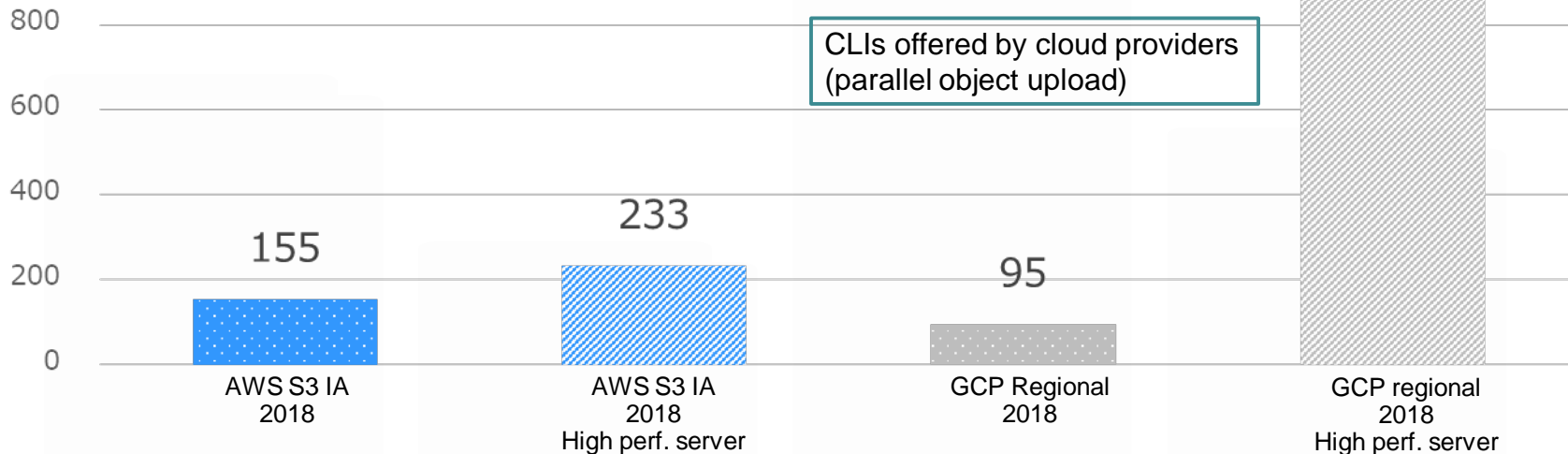
- For all cloud services in which the same measurements were taken (both in 2017 and 2018), performance improvements were achieved.
 - Possible reason: Improvement of providers' infrastructure and improvement of upload CLI

Upload Performance: Upgraded Upload Server

Throughput (MiB/s)

High-energy physics data
633GiB 1,000 files

CLIs offered by cloud providers
(parallel object upload)



- ❑ The services that show performance improvement in 2018 achieve greater performance improvement when data are uploaded from the higher performance server.

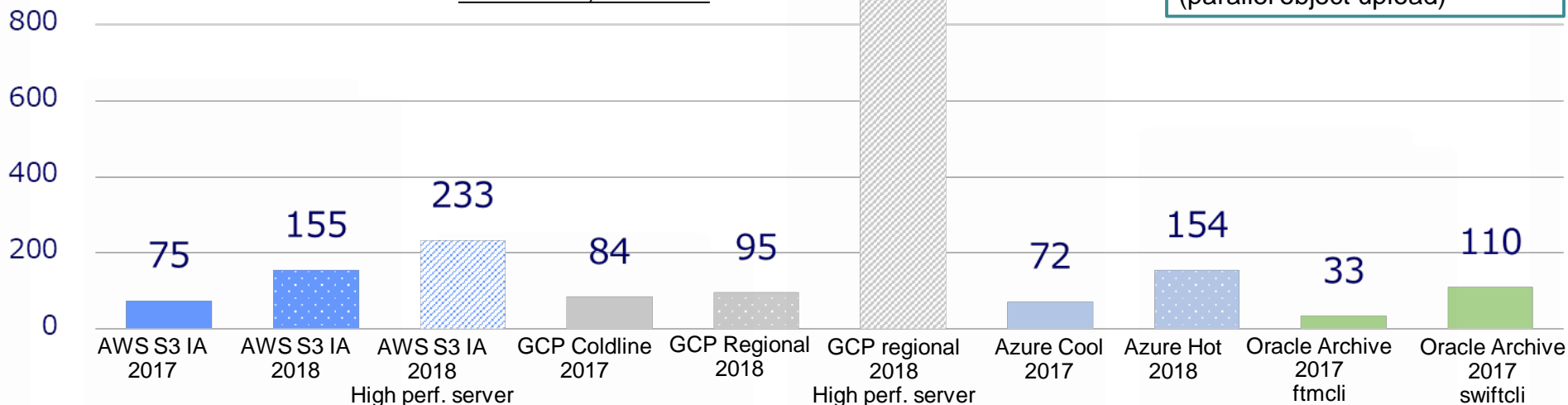
Upload Performance: Summary

Throughput(MiB/s)

High-energy physics data
633GiB 1,000 files

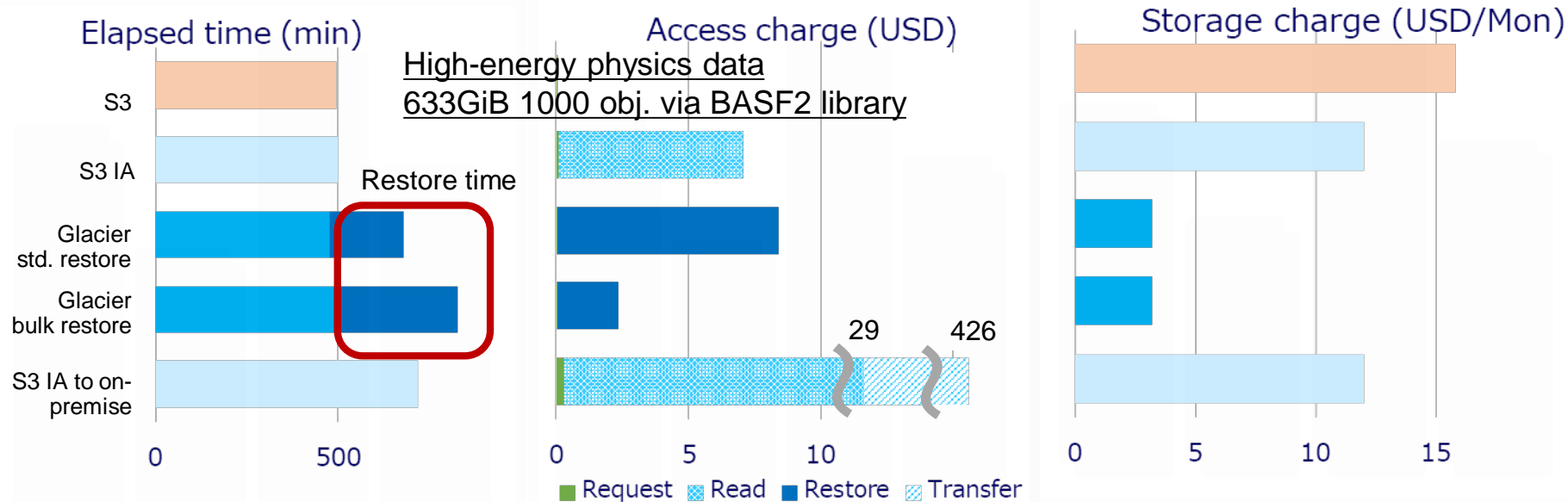
893

CLIs offered by cloud providers
(parallel object upload)



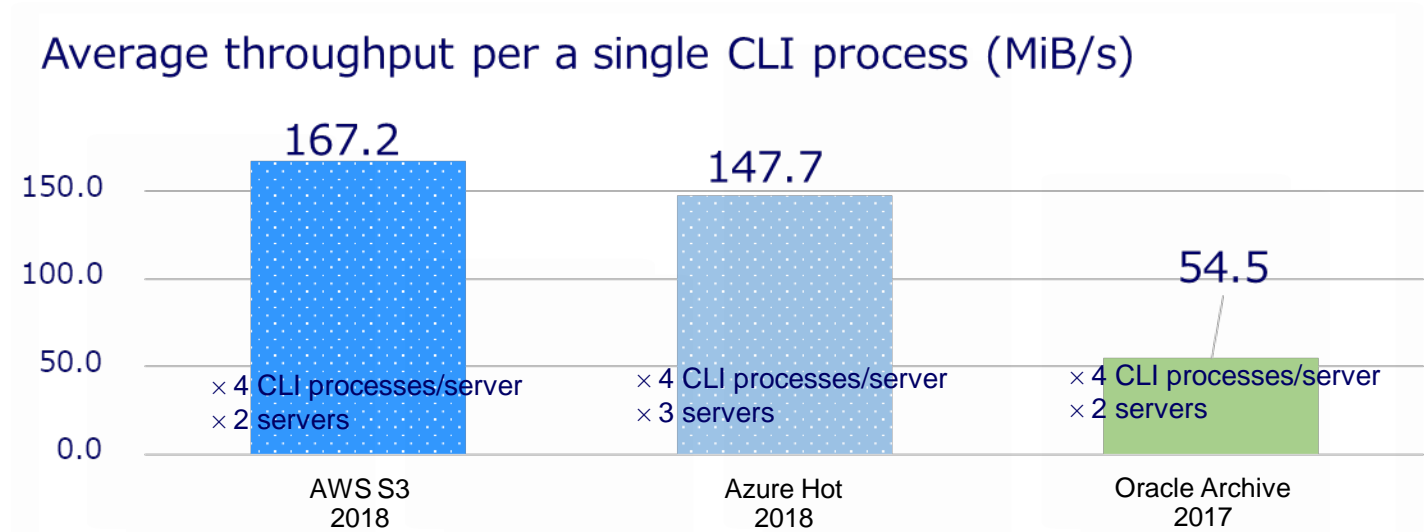
- ❑ For all cloud services in which the same measurements were taken both in 2017 and 2018, performance improvements were achieved.
- ❑ More performance improvement is achieved when data are uploaded from the higher performance server.

Read Performance vs. Cost



- ❑ A retrieval from the cold storage requires read and restore cost.
 - ❑ S3 IA: A little lower storage cost and additional read cost
 - ❑ Glacier: Lower storage cost and additional restore cost with long restore time
- ❑ Excluding Glacier's restore time, access performance results for cloud are almost identical.
- ❑ Transfer cost to send data outside the cloud is high (further investigation required).

Storing and Manipulating 1PiB Data (1)



- ❑ High-energy physics data files are written repeatedly until the total amount of data reaches 1PiB.
- ❑ Although the write throughput is limited, the total elapsed time can be reduced by using multiple CLI processes in parallel.

Storing and Manipulating 1PiB Data (2)

- ❑ Elapsed time to list all objects is too large for interactive processing.

Storage	Number of stored objects	Elapsed time (sec)
AWS S3	1,677,000	860
Azure BLOB Hot	1,665,000	1712

- ❑ Read performance seems to be almost independent of the number of total objects in a bucket/container. ➡ Scalability of access performance

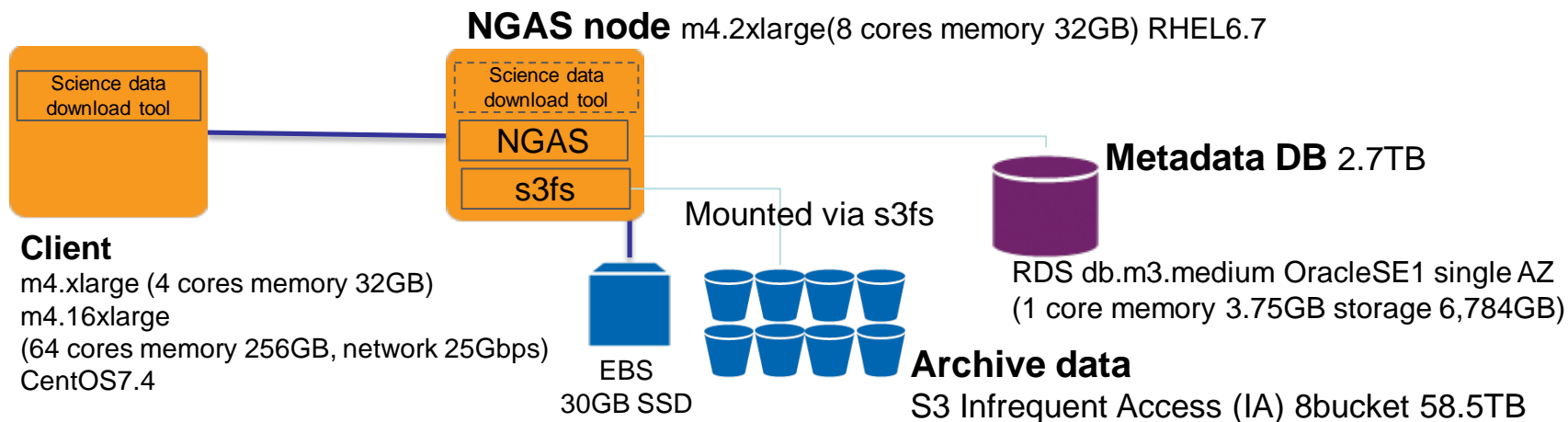
Storage	Read method	Number of stored objects	Number of read objects	Elapsed time (min)
AWS S3	BASF2 library	228,000	1,000	501
		1,000	1,000	493
Azure BLOB Hot	Download CLI	106,835	1,000	198
		1,000	1,000	184

Case Study (2): Astronomy Data

Research area	High-energy physics	Astronomy
Data contents	Belle II experiment physical simulation data	ALMA radio telescope Observation/analysis data
Quantity	633GiB, 1,000 files	58.5TiB, 1,380,000 files
Size	Almost the same (600–700MiB)	Falls between smaller than 1MiB and larger than 100GiB (average: 44MiB)
Application	Read files through analysis support software environment “BASF2” (Belle II Analysis Framework 2)	Data management and retrieval by archive system “NGAS” (Next Generation Archive System)

Configuration of NGAS on AWS

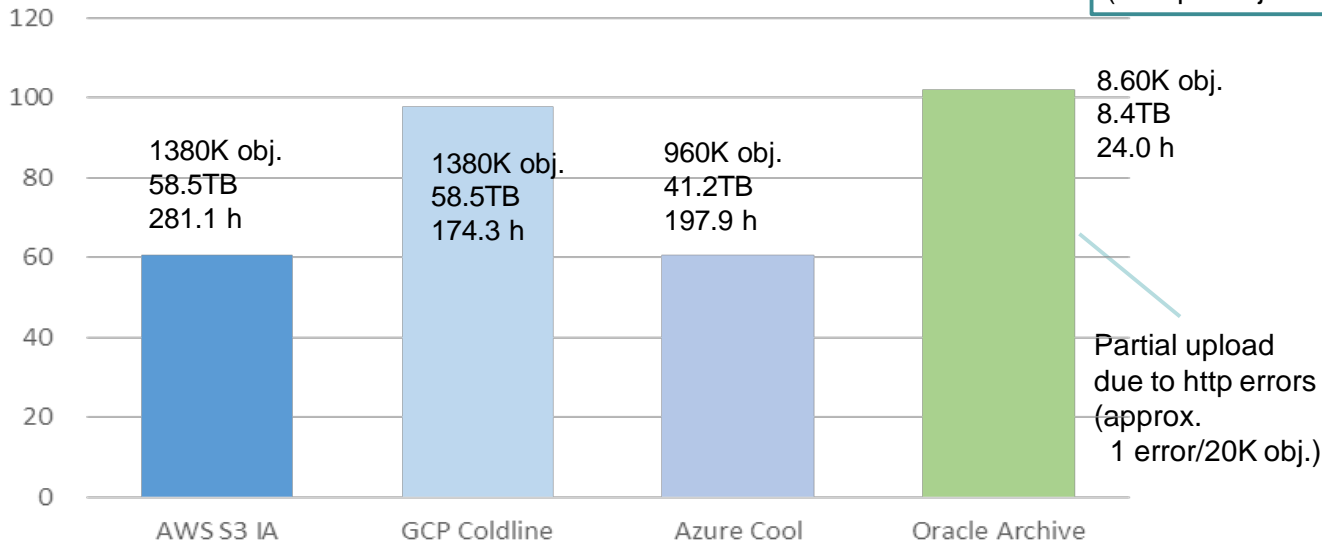
- ❑ S3 buckets storing archive data are mounted as file systems through s3fs-fuse so that objects are accessed as files.
- ❑ Oracle RDS is used to maintain the metadata of archive data.
- ❑ The “science data download tool” on an on-premise server or on a separate EC2 client instance (in the same VPC) retrieves archive data.



Upload Performance: Experiments in 2017

Throughput(MiB/s) Astronomy data of up to 1380K files (58.5TB)

CLIs offered by cloud providers
(multiple object upload)

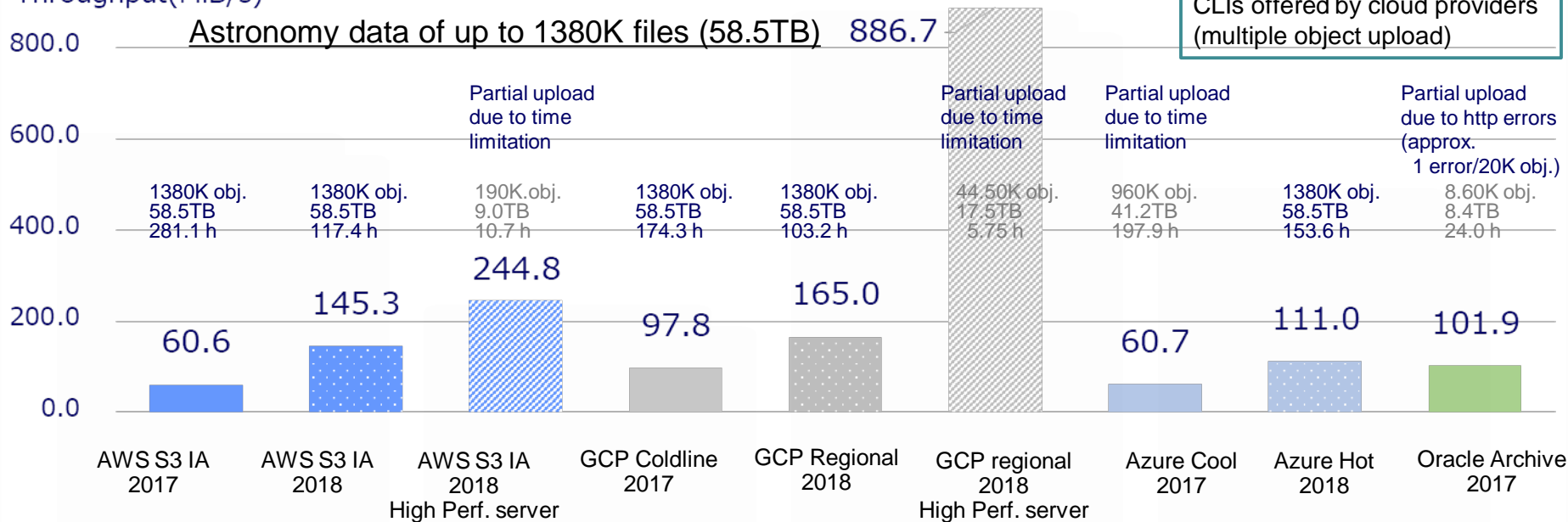


- ❑ Throughput values were less than those for high-energy physics data uploads because of smaller average object size (44MiB).
- ❑ In the Oracle case, data center location (US) may affect network stability.

Upload performance: Summary

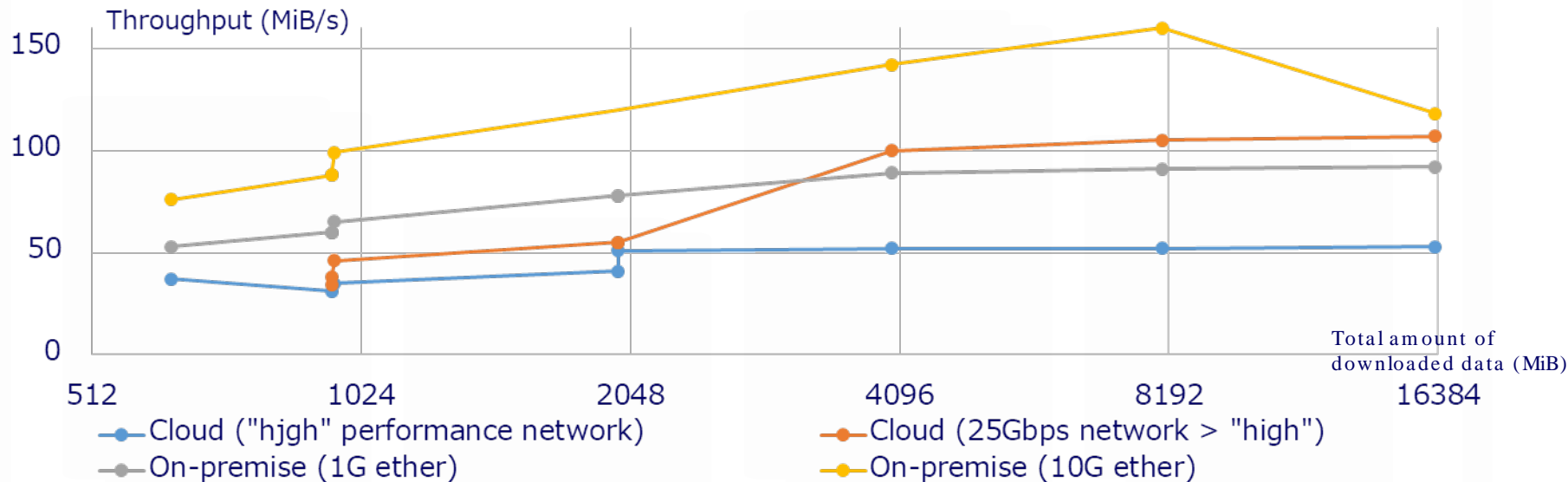
Throughput(MiB/s)

Astronomy data of up to 1380K files (58.5TB)



- ❑ The results show the similar tendency to high-energy physics data.
 - ❑ Performance improvement between 2017 and 2018
 - ❑ More performance improvement is achieved in the higher performance server cases

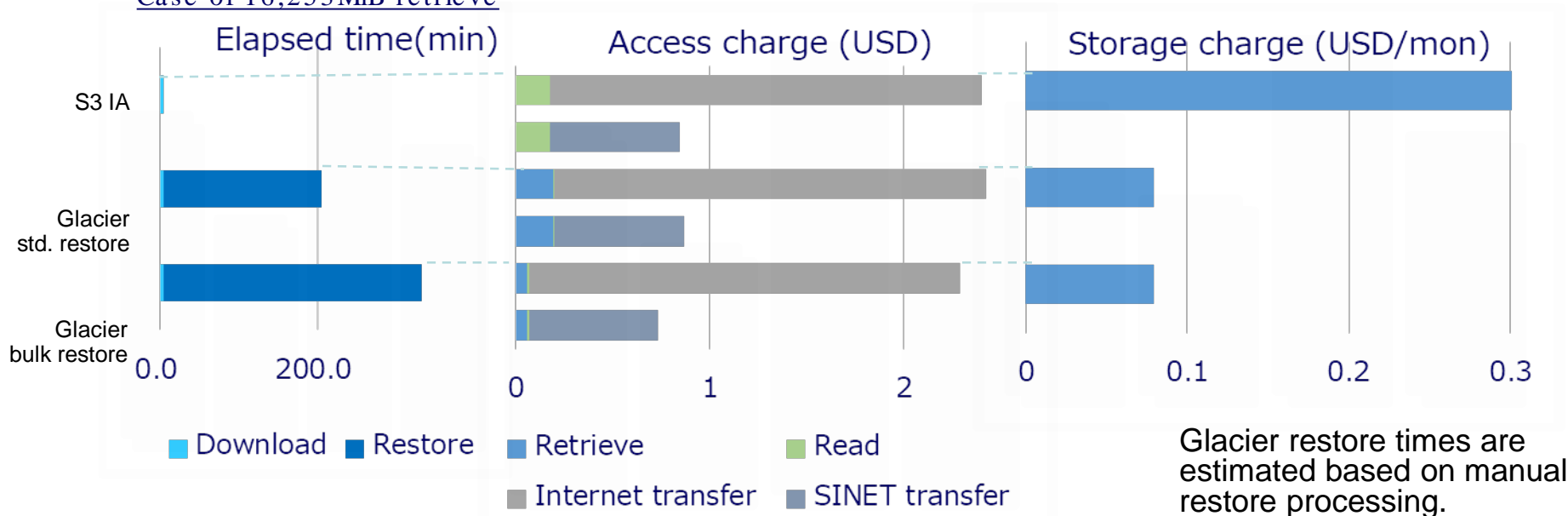
Data Download Performance of NGAS



- ❑ The performance of cloud cases is lower than that of on-premise cases.
 - ❑ Possible reasons: difference of performance between S3 IA and on-premise storage, overhead of s3fs, performance limitation of 1 core RDS instance (4 cores for on-premise)
- ❑ Throughput values are improved along with the improvement of network performance.
 - ➡ Practical performance for production use will be achieved by appropriate sizing.

Retrieve Performance and Cost of NGAS

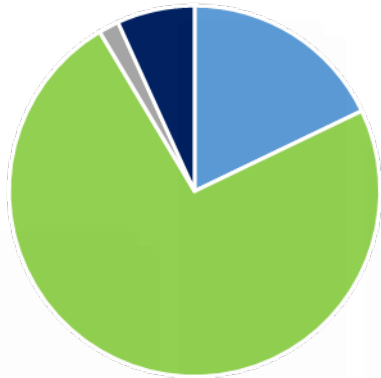
Case of 16,253MiB retrieve



- ❑ A retrieval from the cold storage required read and restore cost.
- ❑ In the Glacier case, restore cost and time is required in spite of low storage cost.
- ❑ Transfer cost to send data outside the cloud is comparatively high.

An Example of Cost Estimation of NGAS on AWS

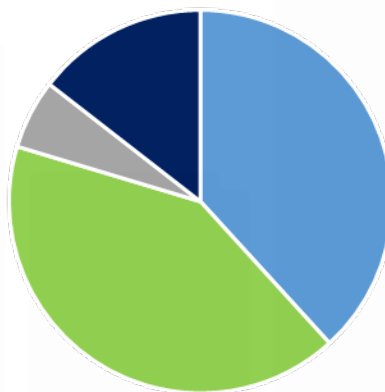
- Store 500TiB archive data in S3 IA and retrieve 250TiB per year



■ NGAS instance and DB ■ S3 IA storage
■ S3 IA read ■ Data transfer

- Yearly operation cost: 158,666 USD
- Storage cost is 2/3 of the total cost.

- Store 500TiB archive data in Glacier and retrieve 250TiB per year

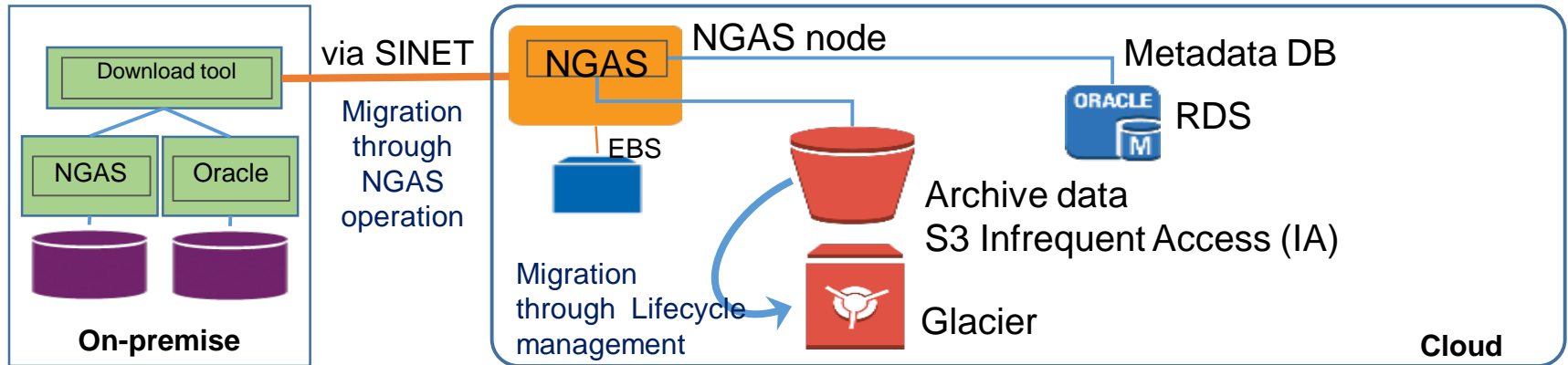


■ NGAS Instance and DB ■ Glacier storage
■ Glacier standard restore & read ■ Data transfer

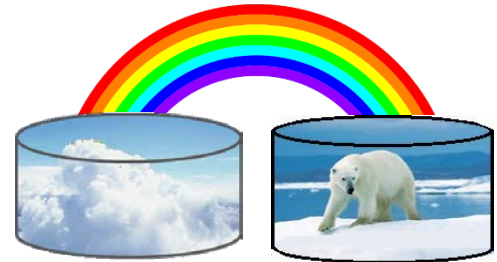
- Yearly operation cost: 74,191 USD
- Restore time (standard: 3.3 hours) is always required.

Hybrid Configuration

- ❑ Hybrid configuration of the on-premise NGAS system and NGAS system on cloud
 - ❑ The on-premise NGAS system is already in use.
 - ❑ Migrate data to cloud after the predefined period (e.g., 3 years) because the access frequency of old data is low..
- ❑ Investment in on-premise NGAS system can be kept constant.
- ❑ Service levels for frequently accessed new data is assured.
- ❑ The access and data transfer costs from the cloud are reduced.



Conclusion



Conclusions

- ❑ Experiments using actual scientific research data in multiple commercial cold storage services were conducted.
- ❑ Cloud services are evolving very quickly.
 - ❑ 2x – 10x improvements of data upload performance in a year were achieved.
- ❑ Information which helps estimating cost for storing data in cloud was acquired.
 - ❑ When storing and accessing a large amount of research data, the major part of the cost is that for data storage and data transfer outward cloud.
 - ❑ Data read cost of cold storage has a small effect on the total cost.
- ❑ Cold storage services which require rather long (hourly) restore operations before data access are viable solutions in terms of cost.
 - ❑ Cost performance of the whole system can be optimized with appropriate storage tiering.

Next Steps

- ❑ Optimize applications and usages of services considering the characteristics of cloud cold storage
 - ❑ Optimize mapping between files and objects
 - e.g., 1 file to 1 object ➡ multiple files accessed at a time to 1 object
 - ❑ Reduction of time and cost of restore processing
 - ❑ Improved handling of multiple objects
 - ❑ Adopt cloud-native object storage API
 - ❑ Improved performance and stability
 - ❑ Predictability of charge
- ❑ Continuously collect and share helpful information and tools for cloud cold storage
 - ❑ Benchmark results, experiment results, and fixed-point observation results
 - ❑ Scripts to handle cold storage services

Thank you!



**Inter-University Research Institute Corporation /
Research Organization of Information and Systems**

National Institute of Informatics