Variable: Secure Cloud of Clouds System for Big Data

D. Saritha Reddy

Assistant Professor, Department of MCA, Narayana Engineering College-Gudur, Nellore, AP

Sk. Chand Basha

PG Student, Department of MCA, Narayana Engineering College-Gudur, Nellore, AP

Abstract—The cloud-backup storage system capable of storing andsharing big data in deficient, reliable and using multiple cloud providers and storage repositories to comply with the legal requirements of sensitive personal data. CHARON is having the 3 features: and (1)any single entity it does not require trust, (2) it efficiently deals with large files over a set of geo-dispersed storage services.(3) they can run without client-managed server, to avoid write-write conflicts between clients accessing shared repositories to avoid write-write conflicts between clients accessing shared repositories by using novel Byzantine-resilient data-centric leasing protocols we can developed. The CHARON usingapplication-basedand microbenchmarks simulating representative workflows from required bioinformatics and life sciences organization, a prominent big data domain. The results show that our unique design is not only feasible but also presents an end-to-end performance of up to 2.5×better than other cloud-backed solutions.

Index Terms—Cloud storage, Byzantine fault tolerance, Big-data storage.

1 INTRODUCTION

The main motivation for building this system is to support the management of genomic data required by bioinformatics and life science organization. We present CHARON, a cloud-backed storage system capable of storing and sharing big data in a secure, reliable, and efficient way using multiple cloud providers and storage repositories.

CHARON implements three distinguishing features:

1. It does not require trust on any single entity.
2. It does not require any client-managed server, and
3. It efficiently deals with large files over a set of geo-dispersed storage services.

Unfortunately, many organizations are still reticent to adopt public cloud services. First, few tools are already integrated with clouds, introducing difficulties to non-technical users. We present CHARON, a near-POSIX cloud-backed storage system capable of storing and sharing big data with minimal managementand node dictated infrastructure. The main motivation for building this system was to support the management of genomic data, as required by bioinformatics and life sciences organizations.

As an example, consider the case of biobanks. These institutions were originally designed to keep physical samples that could be later retrieved for research purposes. More recently, they are becoming responsible also for storing and analysing the data related to such samples. A sequenced human genome can reach up to 300GB, and each individual may have his genome sequenced many times during his life. The problem is that biobanks lack the scalable infrastructure for storing and managing this potentially vast data volume. Public clouds have plenty of resources for that.

CHARON employs a set of Byzantine-resilient data-centric algorithms, including a novel leasing protocol to avoid write-write conflicts on shared files. For instance, DepSky’s mutual exclusion and Metadyne’s Paxo’s rely on the strong consistency of such services. However, this property does not hold when listing the objects in a container in popular services like Amazon S3 and Rack space Cloud Files, leading to potential safety violations in these algorithms if such services are used.

In the project we are using the big data to store in to the cloud. Big data is a field that treats way to analyze, systematically extract information from, or otherwise deal with data sets that are too large or complex to be deal with by traditional data-processing application software. Data with many cases (rows) offer greater statistical power, while data with heifer complexity may lead to a higher false discovery rate. Big data challenges include capturing data, data storage, data analysis, quiring, updating, information privacy and the source.

Big Dara current usage of the term bigdata tends to refer to the use pf predicates analytics, user's behaviour analysis, or certain other advances data analysis methods that extract value from Dara, and seldom to a particular size of data set. "there is little doughty that the quantities of the data now available are indeed large, but that's not the most relevant characteristic of this new data ecosystem analysis of data sets can find new correlation to "spot business trends, prevent diseases, combat crime and so on.

2. RELATED WORK

Like in previous works, we consider an unconstrained set of clients and a group of cloud providers. Each client has a unique id, an account for each cloud, and limited local storage. Every cloud provider offers one or more services, which implement access control mechanisms to ensure that only authorized accounts can access them.

We summarize the related work in the following categories.

Distributed File Systems CHARON is an intrusion-tolerant file system that maintains dataconfidentiality,integrity, andtheexistence of compromised components. Our design adopts some ideas from existing file systems, such as the separation of data and metadata from NASD, volume leases from AFS, and background updates from several peer-to-peer file systems. In particular, far site has some equalities with the system, but is critically different in its use of complex Fault-Tolerant replica groups for allocating leases and uphold metadata steadily. Another related system is xFS, a datacentric network file system in which all data and metadata are stored at the client side. A distributed file system for cloud is a file system that allows many clients to have access to data and supports operations (create, delete, modify, read, write) on that data. Each data file may be partitioned into several parts called chunks. Each chunk may be stored on different remote machines, facilitating the parallel execution of applications. Typically, data is stored in files in a hierarchical tree, where the nodes represent directories. There are several ways to share files in a distributed architecture: each solution must be suitable for a certain type of application, depending on how complex the application is. Meanwhile, the security of the system must be ensured. confidentiality, availability and integrity are the main keys for a secure system.

Data-centric coordination key feature of CHARON is the use of Byzantine-resilient datacentric algorithms for implementing storage and coordination. There are some works that propose the use of this kind of algorithms for implementing dependable systems. These algorithms could be used to implement mutual exclusion satisfying deadlock-freedom (a stronger liveness guarantee than obstruction-freedom). However, these solutions would require a much larger number of cloud accesses. Our lease protocol, on the other hand, requires only two to four cloud accesses for acquiring a lease. A data-centric enterprise is one where all application functionality is based on a single, simple, extensible data model.

Multi-cloud storage in the last years, many works have been proposing the use of multiple cloud providers to improve the integrity and availability of cloud-of-clouds, single cloud, or private repository. A problem in some of them is the fact they only provide object storage (i.e., read/write registers), which hardens their integration with existing applications.

Multi-cloud is the use of multiple cloud computing and storage services in a single heterogeneous architecture. This also refers to the distribution of cloud assets, software, applications, etc. across several cloud-hosting environments. With a typical multi-cloud architecture utilizing two or more public clouds as well as multiple private clouds, a multi-cloud environment aims to eliminate the reliance on any single cloud provider. It differs from hybrid cloud in that it refers to multiple cloud services rather than multiple deployment modes (public, private, legacy). Also, in a multi-cloud environment, synchronization between different vendors is not essential to complete a computation process, unlike parallel computing or distributed computing environments.

CHARON implements a security model where the owner of the file pays for its storage and defines its permissions. This is enforced by mapping the file system permissions (POSIX ACLs) to cloud services access control mechanisms (see details in§4.3). Therefore, a malicious client can only see, modify, and delete its own files and the files shared with him.

The expected size of the files and the envisioned users justify this decision. More specifically, (1) solving conflicts manually in big files can be hard and time-consuming; (2) users are likely to be non-experts, normally unaware of how to repair such conflicts; and (3) the cost of maintaining duplicate copies of big file scan be significant. For instance, collaborative repositories, such as the Google Genomics, require such control since they allow users to read data about available samples, process them, and aggregate novel knowledge on them by sharing the resulting derived data into the bucket containing the sample of interest.

3.Proposed work

CHARON separates file data and metadata in different objects stored in diverse locations and manages them using different strategies, as illustrated in Figure 1. File data locations are of three kinds in CHARON: cloud-of-clouds, single (public) storage cloud, and private repository (e.g., a private cloud). These alternatives explore various cost-dependability trade-offs and address all placement requirements we have encountered with life sciences and big data applications. For example, the cloud-of-clouds can store critical data (CHARON'S namespace and file B) that needs the availability and confidentiality assured by the multi-cloud scenario (provider-fault tolerance). A single cloud can store non- critical public studies and anonymized datasets (file D) (provider- dependent and potentially less expensive.

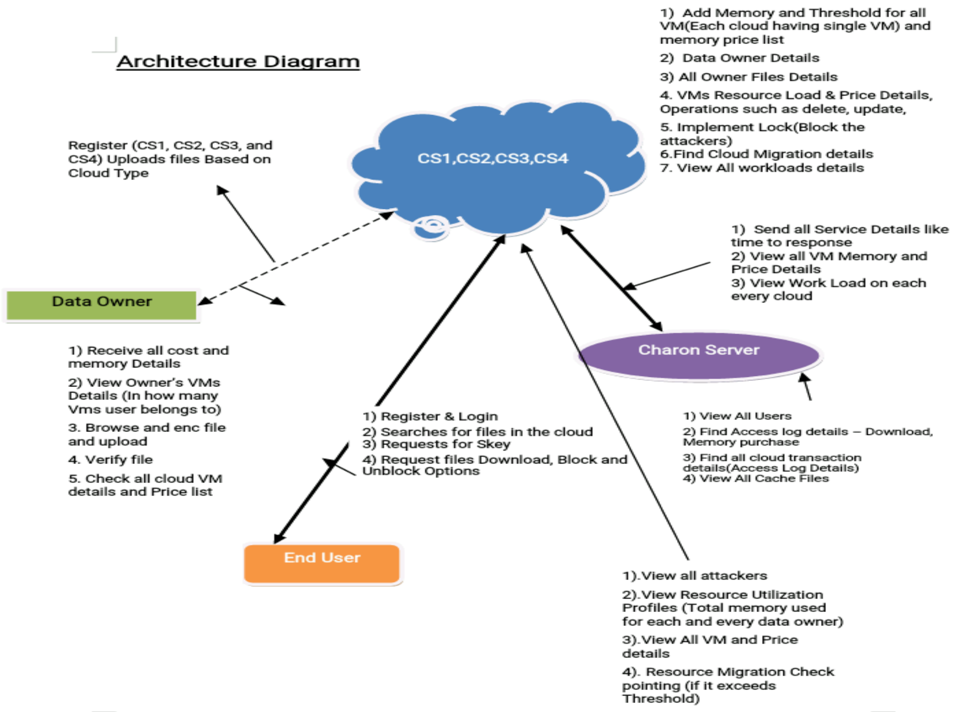

Architecture Diagram
1) Add Memory and Threshold for all VM(Each cloud having single VM) and memory price list
2) Data Owner Details
3) All Owner Files Details
4. VMs Resource Load & Price Details, Operations such as delete, update,
5. Implement Lock(Block the attackers)
6.Find Cloud Migration details
7. View All workloads details
Register (CS1, CS2, CS3, and CS4) Uploads files Based on Cloud Type
CS1,CS2,CS3,CS4
1) Send all Service Details like time to response
2) View all VM Memory and Price Details
3) View Work Load on each every cloud
Data Owner
Charon Server
1) Receive all cost and memory Details
2) View Owner's VMs Details (In how many Vms user belongs to)
3. Browse and enc file and upload
4. Verify file
5. Check all cloud VM details and Price list
1) Register & Login
2) Searches for files in the cloud
3) Requests for Skey
4) Request files Download, Block and Unblock Options
1) View All Users
2) Find Access log details – Download, Memory purchase
3) Find all cloud transaction details(Access Log Details)
4) View All Cache Files
End User
1).View all attackers
2).View Resource Utilization Profiles (Total memory used for each and every data owner)
3).View All VM and Price details
4). Resource Migration Check pointing (if it exceeds Threshold)

Figure 1. CHARON architecture.

Finally, private repositories must be used to keep clinical data from human samples that cannot leave the boundaries of a particular institution (file A) or country (file C) (subject to local infrastructure restrictions).

The system has following implantation modules

DATA OWNERA data owner is an individual who is accountable for a data asset. This is typically an executive role that goes to the department, team or business unit that owns a data asset.

- In this module, initially the data owner has to get register to the cloud server (CS1, CS2, CS3, CS4)
- Data owner will login to the corresponding cloud server he got registered.

- Data owner encrypt will upload file to the cloud server (CS1, CS2, CS3, CS4) Data owner verifies the file he uploaded either it is safe or not.

Charon Server:

- The cloud server manages a cloud to provide data storage service.
- Data owners encrypt their data files and store them in the cloud for sharing with cloud consumer.

CLOUD CONSUMER: The cloud consumer is the principal stakeholder for the cloud computing service. A cloud consumer represents a person or organization that maintains a business relationship with, and uses the service from a cloud provider

- Cloud consumer first has to register to the cloud server (CS1, CS2, CS3, CS4) which particular cloud he has to use.
- Cloud consumer has to login to the cloud he got registered.

This proposed work follows the following algorithm

- Previous data-centric fault-tolerant mutual exclusion algorithms were designed to work directly on top of storage services.
- In this paper, we propose a more modular approach in which we build non-fault-tolerant base lease objects, each on top of a specific cloud-provided service, and 3 f +1 of these services are combined in an f -fault-tolerant composite lease object.
- This approach allows the design of more efficient base lease objects on top of any cloud-provided service (e.g., queues) instead of relying on fault-tolerant register constructions as in previous works.
- We still provide the design of base lease objects on top of storage services because they are the only abstraction available in certain cloud providers.

Efficiently storing large data sets of human genomes is a long-term ambition from both the research and clinical life sciences communities. Cloud computing is a natural economic alternative to provide infostructure's, but it is not as good an alternative to provide infrastructure in terms of security and privacy. These mechanisms encompass.

4.CHARON IMPLEMENTATION

Metadata Organization Metadata is the set of attributes assigned to a file/directory (e.g., name, permissions). Independently of the location of the data chunks, CHARON stores all metadata in the cloud-of-clouds using single-writer multi-reader registers to improve their accessibility and availability guarantees. More specifically, we redesigned and optimized the SWMR register implementation of DepSky to improve the performance and concurrency as described in the remaining of this section. Metadata are small pieces of information that can be attached to files and other objects in a computing environut.

Data Management This section describes the most important techniques CHARON uses to manage big files efficiently. Data management is the practice of collecting, keeping, and using data securely, efficiently, and cost-effectively. The goal of data management is to help people, organizations, and connected things optimize the use of data within the bounds of policy and regulation so that they can make decisions and take actions that maximize the benefit to the organization. Cloud data management is a way to manage data across cloud platforms, either with or instead of on-premises storage. The cloud is useful as a data storage tier for disaster recovery, backup, and long-term archiving. ... Data stored in the cloud has its own rules for data integrity and security

Multi-level cache CHARON uses the local disk to cache the most recent files used by clients. Moreover, it also keeps a fixed small main-memory cache to improve data accesses over open files. Both of these caches implement least recently used (LRU) policies. Multilevel Cache Organisation. Cache is a random-access memory used by the CPU to reduce the average time taken to access memory. Multilevel Caches is one of the techniques to improve Cache Performance by reducing the "MISS PENALTY".

Cloud-backed Access Control CHARON clients are not required to be trusted since access control is performed by the cloud providers, which enforce the permissions for each objectto many security dealers and integrators, cloud-based access control is something that utilizes a cloud-first or cloud-only approach. Many security dealers and integrators, cloud-based access control is something that utilizes a cloud-first or cloud-only approach. Typically, it is used for hosted or managed access control, and is primarily aimed at smaller-sized end users that don't have their own security or IT departments. When configuring the system, users define the API credentials of the 3f +1 clouds CHARON will use. In this way, when a CHARON client starts, it authenticates in each one of the providers, and after that each file or directory a user creates results in the creation of one or more objects associated with user cloud accounts.

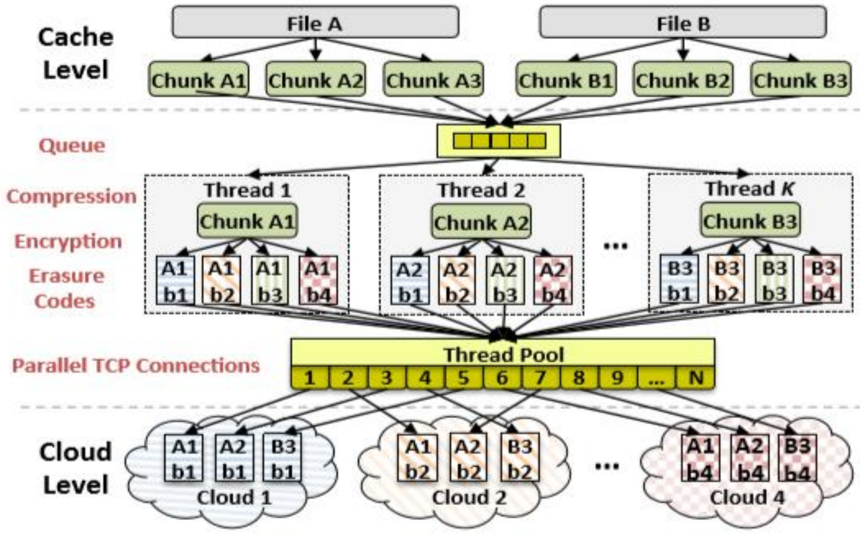


Figure 2. Data chunks management

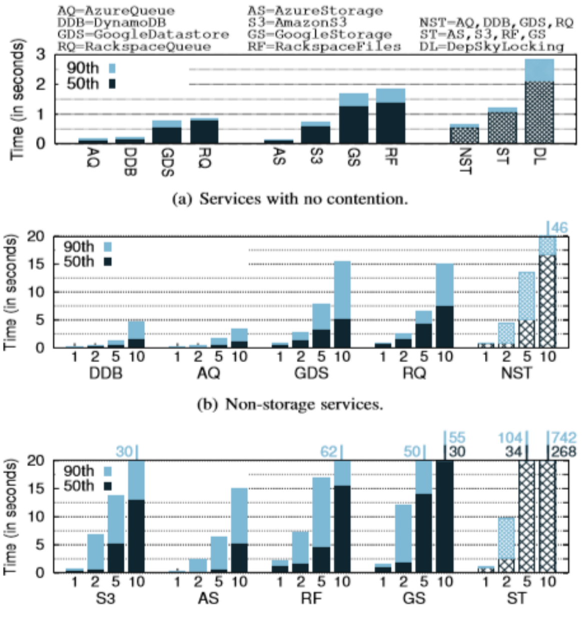
Dependable and secure storage in a cloud of clouds: In this work we present DepSky, a system that improves the availability, integrity, and confidentiality of information stored in the could trough the encryption, encoding, and replication of the data on diverse could that form a cloud-of-clouds. In this work we present DepSky, a system that improves the availability, integrity, and confidentiality of information stored in the cloud through the encryption, encoding, and replication of the data on diverse clouds that form a cloud-of-clouds.

5.Experiment Rustles

ExperimentalEnvironmentAccordingly, an experimental learning environment can be defined as a place where individuals can learn by generating or testing hypotheses in a controlled way.We evaluate CHARON and compare it with other systems. The experiments present results of (1) the latency of the leasing algorithms, (2) several microbenchmarks of metadata and data intensive operations, and (3) a bioinformatics benchmark. we used four machines (Intel Xeon E5520, 32 GB RAM, 15kRPM HDD, SSD) connected through a gigabit network located in Portugal.

The three storage locations for CHARON were configured as follows. The cloud-of-clouds storage uses Amazon S3 (US), Windows Azure Storage (UK), Rack space Cloud Files (UK), and Google Cloud Storage (US). For the single cloud storage, we use only Amazon S3 (US). The private repository was either located in the client's machine disk or in a different machine in the same LAN. For the composite lease, we use additional cloud services: Azure Queue, Rack Space Queue, Amazon DynamoDB,andGoogleDatastore.Therefore,allcloudof-clouds configurations consider f = 1.

ExecutionsUnder contention an important aspect of a lease algorithm is how its performance degrades with contention. We perform experiments with a varying number of clients (1, 2, 5, and 10) trying to acquire a lease on the same SNS (and releasing it right after), and measure the time for a client to acquire the lease. For obstruction-free lease algorithms, we use a random back off time of up to one second.



(a) Services with no contention.

(b) Non-storage services.

(c) Storage services.

Figure 3. Latency of lease acquisition algorithms without contention and under contention of up to 10 clients.

Composite Leasing composite cloud is an approach to not get locked in to a hybrid cloud that restricts the add/remove of cloud services over time. ... It is ok for some parts of enterprise IT to not make it into the cloud, but with use of cloud proxies and APIs those old legacy systems could get a new lease of life

File System Microbenchmarks This section compares CHARON with other file systems using the File bench microbenchmark suite. The first two experiments are focused on evaluating the performance of isolated system calls, while the third one measures the latency associated with interacting with the clouds for downloading and uploading data. The presented results do not consider the lease acquisition time.

Metadata-intensive operations Our first experiment focuses on how well the system deals with metadata intensive operations when compared with other systems. Metadata is data that describes other data. Meta is a prefix that -- in most information technology usages -- means "an underlying definition or description." Metadata summarizes basic information about data, which can make finding and working with particular instances of data easier. Table 2 presents the number of operations per second for ext4 (on SSD), NFS, S3QL, SCFS (discussed in §6), and CHARON for different operations on 0-byte files.

Data-intensive operations Table 3 presents the results for similar microbenchmarks, but now focusing on data-intensive operations with files of 256MBUnsurprisingly, ext4 offers the best-read throughput both for sequential and random workloads. S3QL and NFS provide a slightly lower read throughput than ext4.Data-intensive computing is a class of parallel computing applications which use a data parallel approach to process large volumes of data typically terabytes or petabytes in size and typically referred to as big data. Despite presenting a lower performance, SCFS and CHARON are still competitive for read workloads. When considering write throughput, ext4 and NFS present the best performance for sequential workloads.

This happens because previous multi-cloud data replication algorithms are not optimized to deal with big files: they do not break the file in chunks (or use chunks too small), neither use our upload strategy (see Figure 4), nor use techniques such as prefetching.Furthermore,CYRUSround-robindistributionofdata chunks among all clouds (trying to balance storage usage) makes it quite slow, as some clouds are noticeably slower than others. On the other hand, our approach on data chunks management allows CHARON to perform as fast as the f +1th (resp. 2f +1th) fastest clouds when reading (resp. writing) chunks.

The difference between the latency of CHARON using Amazon S3 or the cloud of-clouds is quite small for both reading and writing results. For writing, the additional latency presented by the cloud-of-clouds comes from the fact that we need to write the data in three clouds to finish it. Thus, the end-to-end latency will be dictated by the third fastest cloud.

Bioinformatics Workworn last set of experiments aims to compare the performance of different configurations of CHARON and alternative systems using FS-Biobank, a novel storage benchmark in the domain of bioinformatics (described in the Supplemental Material). Cloud-based bioinformatics workflow platform. The computing nodes as well as Cloud storage are then automatically configured on AWS. ... The shared file system provides a common storage accessible to Galaxy, Globus Transfer, and the Amazon EC2 nodes used for genome analysis.

The FS-biobank emulates specifically the I/O operations of eight representative bioinformatics workloads, summarized in Table 4, and is independent of any external tool. Most workflows use sequential reads and writes, and they differ in the number of accessed files, their size and structure, and the execution pattern

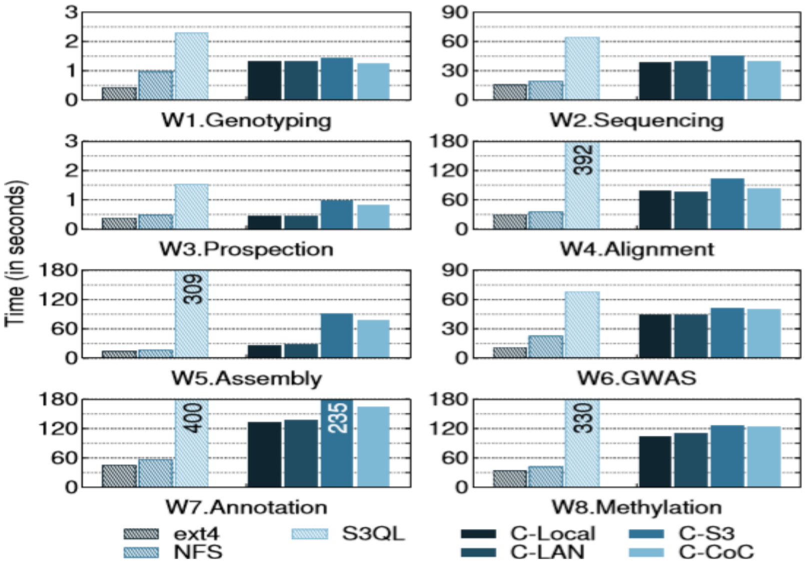


Figure 4. FS-biobank execution for different configurations.

(e.g., reading an entire file before writing anything, interposing reads and writes, reading more than one file in parallel). These workflows include complex pipelines to achieve concrete results in bioinformatics and were selected from the workflows analysed and implemented in the Biobank Cloud project.

Figure 6 presents the duration of FS-biobank workflows for ext4 on SSD, NFS with the client and server in the same LAN, S3QL,and CHARON usingarepositoryindifferentlocations: SSD in the same machine (C-Local), disk in a server in the same LAN (C-LAN), AWS S3 (C-S3), and cloud-of-clouds (C-CoC). SCFS is not evaluated because it does not support big files. We execute every workflow ten times on each scenario and report average values. CHARON's and S3QL's caches are cleaned after each workflow execution, as all results would be similar to Cloacal if the files were cached.

In conclusion, CHARON runs the workflows up to 2.5× (W4) faster than the other (single) cloud-backed file system (S3QL). Furthermore, oursystemusingthecloud-of-cloudsis30% to 200% slower than NFS in all workflows but W5 (which is read intensive). This is an excellent result as the latency of accessing the cloud is 100×higher than accessing a LAN-based server.

6. CONCLUSIONS

CHARON is a cloud-backed file system for storing and sharing big data. Its design relies on two important principles: files metadata and data are stored in multiple clouds, without requiring trust on any of them individually, and the system is completely datacentric. This design has led us to develop a novel Byzantineresilientleasingprotocoltoavoidwrite-writeconflictswithoutany Cust server.Our result showthatthisdesignisfeasibleandcan be employed in real-world institutions that need to store and share large critical datasets in a controlled way.

REFERENCES

1. Cloud Harmony, "Service Status," https://cloudharmony.com/ status-of-storage-group-by-regions, 2019.

2. CloudSecurity Alliance, “Top Threats,” https://cloudsecurityalliance.org/ group/top-threats/, 2016.
3. H.S.Genital,“Whydoesthecloudstopcomputing: Lessonsfrom hundreds of service outages,” in Proc. of the SoCC, 2016.
4. European Commission, “Data protection,” https://ec.europa.eu/info/law/ law-topic/data-protection en, 2018.
5. G. Gaskell and M. W. Bauer, Genomics and Society: Legal, Ethical and Social Dimensions. Routledge, 2013.
6. A. Bessani et al., “Biobank Cloud: a platform for the secure storage, sharing, and processing of large biomedical data sets,” in DMAH, 2015.
7. H. Gottweis et al., “Biobanks for Europe: A challenge for governance,” EuropeanCommission,Directorate-GeneralforResearchandInnovation, Tech. Rep., 2012.
8. C. Basescu et al., “Robust data sharing with key-value stores,” in Proc. of the DSN, 2012.
9. Amazon, “AmazonS3dataconsistencymodel,” https://docs.aws.amazon.
10. [10] Google, “Google Genomics,” https://cloud.google.com/genomics/, 2019.

authors profiles:

D. Saritha Reddy has received herM.Techdegree in *computer science* from *Acharya Nagarjuna University*, Guntur in 2010 and pursuing Ph. D with area computer Networks. At present she is working as *Assistant professor in Narayana Engineering College*, Gudur, Andhra Pradesh, India

Sk. Chand Basha has received his B.Sc. degree in *computer science* from *Dr. CRR degree college*, sydapuram affiliated to *VikramaSimhapuri University*, Nellore in 2017 and pursuing PG degree in *Master of Computer Applications* (MCA) from *Narayana Engineering College*, Gudur affiliated to *JNTU*, *Anantapur*, Andhra Pradesh, India.