Using IKAROS to form Scalable Storage Platforms

Christos Filippidis, Christos Markou
Institute of Nuclear and Particle Physics
NCSR "Demokritos"
Agia Paraskevi, Greece
Christos.Filippidis@cern.ch

Christos Filippidis, Yiannis Cotronis
Department of Informatics and Telecommunications
National and Kapodistrian University of Athens
Panepistimiopolis Ilisia, Greece

***Abstract*—We present IKAROS as a utility that permit us to form virtual scalable storage platforms. IKAROS is a data centric, energy aware platform seeking synergies between wider communities, permits ad-hoc nearby storage formations and is able to use a huge number of low spec, low power consumption I/O nodes in order to increase the available bandwidth and decrease the overall power consumption. It unifies remote and local access in the overall data flow, by permitting direct access to each I/O node. This approach enable users to virtually connect the several different computing facilities (Grids, Clouds, HPCs, Data Centers, Local computing Clusters and personal storage devices), being utilized by an international collaborative experiment, on-demand based on their needs.**

Keywords—scientific computing; resilient storage formations; metadata utility; data management; storage; DaaS; Overall Data Flow;

I. Introduction

International collaborative scientific experiments are generating datasets which are increasing exponentially in both complexity and volume, making their analysis, archival, and sharing one of the grand challenges of the 21st century. These experiments, in their majority, adopts computing models consisting of different Tiers (each Tier is made up of several computer centres and provides a specific set of services) and for the different steps of data processing (simulation, filtering, calibration, reconstruction and analysis) several software packages are utilized. The computing requirements are extremely demanding and, usually, spans from serial to multi-parallel or GPU-optimized jobs. The collaborative nature of these experiments demands very frequent WAN data transfers and data sharing among individuals and groups. Typically, such a computing model utilizes several different computing infrastructures like: Grids, Clouds, HPCs, Data Centers and Local computing Clusters.

More specifically, The KM3NeT experiment (both an ESFRI project and a CERN recognized experiment) will introduce a distributed network of neutrino telescopes with a total volume of several cubic kilometers at the bottom of the Mediterranean Sea. A researcher member of the experiment, in his/her every day working process will have to move datasets from Data Centers to a Local computing Cluster by using secure copy tools (SCP). Then s/he will conduct a short local analysis and upload the output to a grid User Interface (UI) in order to start more coherent analysis, using the European Grid Infrastructure (EGI). Finally s/he will have to copy the output back to the local clusters and probably replicate data to a Data Center, by using several interfaces and tools.

The above mentioned working process is dictated by many several aspects. Within a collaboration, a user has to deal with several policies and techniques. Most of the times the computing resources, utilized by a collaboration, implement different technical approaches, have different owners and totally different goals, so they cannot follow the same philosophy.

II. Related Work

Existing Grid Computing tools and frameworks, like PhEDEx [1], cannot provide the flexibility that the next generation collaborative projects demands. The main criticism against these tools and frameworks, being developed the past decade, is that at the Virtual Organization (VO) level they transform the "Grid" to a centralized infrastructure. Even though they provide a distributed implementation there is a centralized control (e.g: a central database). This approach violates the decentralized concept of a Grid Infrastructure, which is the real advantage of the Grid computing model. Current, Grid implementations provides a decentralized resource management in terms of CPU slots and Storage but then, at the application level, most VOs handle data in a more traditional centralized way. On the other hand, current Data as a Service (DaaS) solutions present a lack of flexibility in terms of allowing users to customize the underlying data models by including new concepts or functionalities (lack of data model *enrichment*). The platform providers usually offers a set of global APIs to enable users to access, download, or upload the data based on Create, Read, Update, Delete (CRUD) commands, or "sell" and transfer data to clients so they can do whatever they want with them. Thereby, collaboration becomes limited and sometimes is not even feasible [7]. IKAROS can provide a more sophisticated and flexible approach, based on tools and utilities, that can fulfill the demands of the next generation international collaborative experiments.

III. IKAROS Approach

To avoid the above limitations, we deploy IKAROS as a utility that enable us to create virtual scalable storage formations, on-demand. IKAROS unifies remote and local access in the overall data flow, by permitting direct access to each I/O node. In this way we can handle the overall data flow at the network layer, which is much faster, while the other combined systems (local-remote access) demand a more active involvement of the operating system layer, due to the isolation between the local and remote access and the lack of proper coordination. IKAROS is a data centric, energy aware

Part of this research has been funded by the Thales programme financed by EU-ESF and the Greek NSRF.

platform seeking synergies between wider communities, permits ad-hoc nearby storage formations and is able to use a huge number of low spec, low power consumption I/O nodes in order to increase the available bandwidth and decrease the overall power consumption [3,4]. IKAROS permit us to create more user-driven computing facilities with application users and owners playing a decisive role in governance and focusing on placing computer science and the harvesting of 'big data' at the center of scientific discovery. This approach enable us to virtually connect, at the users level, the several different computing facilities (Grids, Clouds, HPCs, Data Centers, Local computing Clusters and personal storage devices), being utilized, on-demand based on the needs, by using well know standards and protocols, like HTTP. Additionally, this approach enables users to experience and utilize, seamlessly, techniques like parallel data transfers and striping servers outside the Grid, without using complicated tools like GridFTP and without focusing to IT issues. Local cluster or even Data Center administrators, most of the times, are not willing to install and maintain tools like globus/GridFTP [7] if the infrastructure is not participating in a Grid related project.

The IKAROS metadata management service (iMDS) holds a key role in the IKAROS architecture. The iMDS permit us to handle the metadata sub-systems differently based on the needs. We may respond to a client/application request with three different ways. The client may find the answer: 1) within his own "cache", 2) locally at a nearby iMDS utility or 3) at an external infrastructure/utility. This approach, focusing on flexibility, can scale both up and down and so can provide more cost effective infrastructures for both large scale and smaller size systems, unlike the competition. We are using existing cloud infrastructures (acting as the external MDS utility), such as Facebook and Gmail, in order to dynamical manage, share and publish metadata. In this way we do not have to build our own utilities for searching, sharing and publishing, additionally we are enabling users to dynamically use the infrastructure, by creating on demand storage formations [6].

By further analyzing the overall data flow we can easily understand that there is a lack of coordination between the several different software layers. Parallel I/O systems commonly treat access to distributed data as a collective operation, and collective communication operations seek to optimize data transformations and transfers by coordinating related activities, In two-phase I/O and in Remote I/O (3-phase I/O), data is read and then reorganized via interprocess communication prior to transfer [2].

More specific, in a practical level, in order to access remote data we have to combine a local parallel File System and a WAN protocol (e.g: GridFTP+PVFS2). In table I we can clearly show that IKAROS outperforms the combined GridFTP+PVFS2 system. The data transfers were carried out between the ZEUS Computer Cluster at NCSR Demokritos located at Athens Greece and the CERN computing facilities located at Geneva Switzerland. Each data point represents the average of 10 runs of the same experiment. In table I we highlight the limitations that currently exists in accessing data remotely. This happens because GridFTP (or whatever protocol we are using for remote access) is not aware of the underlying parallel file system and because the storage servers consisting the parallel file system, most of the times, are "isolated" by using private IPs and by implementing several different policies. pNFS concept could probably address this problem but the pNFS implementation, also, demands an underlying parallel/distributed file system. IKAROS reduces the required I/O phases, regardless of the Tier, in the computing model, we are acting at and the underlying storage implementation. At the same time we are preserving the separation of each level.

TABLE I.

Overall I/0 operations (Local-Remote Access)	**1 GB Shared WAN Connection**	
	IKAROS (MB/s)	***GridFTP+PVFS2 (MB/s)***
Data collection of 2000GB, using 10 parallel data channels	68,92	36,4

CONCLUSION

IKAROS permit us to create virtual scalable storage formations, concluding to more user-driven computing facilities with application users and owners playing a decisive role in governance and focusing on placing computer science and the harvesting of 'big data' at the center of scientific discovery. This approach enable users to virtually connect the several different computing facilities (Grids, Clouds, HPCs, Data Centers, Local computing Clusters and personal storage devices), being utilized by an international collaborative experiment, on-demand, based on the needs. As a future work, we intent to develop IKAROS further in order to add value to current DaaS solutions by sharing not only the data (between the different computing facilities) but also the data models themselves (data model *enrichment*). We argue that in order to succeed that we must first deploy a storage infrastructure that follows the IKAROS approach.

REFERENCES

[1] TUURA, Lassi et al, PhEDEx high-throughput data transfer management system. International Conference on Computing in High Energy and Nuclear Physics (CHEP2006).

[2] W. Allcock, J. Bresnahan, R. Kettimuthu, M. Link, C. Dumitrescu, I. Raicu, and I. Foster, "The globus striped gridftp framework and server," in *SC '05: Proceedings of the 2005 ACM/IEEE conference on Supercomputing*. Washington, DC, USA: IEEE Computer Society, 2005, p. 54.

[3] Filippidis C., Cotronis Y.,Markou C.,(2013) IKAROS: an HTTP-based distributed File System, for low consumption and low specification devices, Journal of Grid Computing springer.

[4] Filippidis C., Cotronis Y.,Markou C.,(2012) Design and Implementation of the Mobile Grid Resource Management System. Computer Science, 13 (1). pp. 17-24. ISSN 1508-2806.

[5] Bryce Allen and John Bresnahan and Lisa Childers and Ian Foster and Gopi Kandaswamy and Raj Kettimuthu and Jack Kordas and Mike Link and Stuart Martin and Karl Pickett and Steven Tuecke, Software as a Service for Data Scientists, Communications of the ACM, pp. 81-88, volume: 55, 2012.

[6] Filippidis, C., Cotronis, Y., Markou, C. (2014). Forming an ad-hoc nearby storage, based on the IKAROS and social networking services, *J. Phys.: Conf. Ser.* **513** 042018.

[7] Martí J, Queralt A, Gasull D, Cortes T. Living Objects: Towards Flexible Big Data Sharing. Journal of Computer Science & Technology. 2013 ;13:56-63. Available from: http://hpc.ac.upc.edu/PDFs/dir20/file004283.pdf