Contribution ID: **90** Type: **Oral contribution**

Encrypted File System on the EGEE grid applied to Protein Sequence Analysis

Wednesday 1 March 2006 14:15 (15 minutes)

Introduction
Biomedical applications are pilot ones in the EGEE project [1][2] and have their own virtual organization: the "biomed"VO. Indeed, they have common security requirements such as electronic certificate system, authentication, secured transfer; but they have also specific ones such as fine grain access to data, encrypted storage of data and anonymity. Certificate system provides biomedical entities (like users, services or Web portals) with a secure and individual electronic certificate for authentication and authorization management. One key quality of such a system is the capacity to renew and revoke these certificates across the whole grid. Biomedical applications also need fine grain access (with Access Control Lists, ACLs) to the data stored on the
grid: biologists and biochemists can then, for example, share data with colleagues working on the same project in other places. Thus, biomedical data need to be gridified with a high level of confidentiality because they can concern patients or sensitive scientific/industrial experiments. The solution is then to encrypt the data on the Grid storage resources, but to provide authorized users and applications with transparent and unencrypted access.

Biological data and protein sequence analysis applications
Biological data and bioinformatics programs have both special formats and behaviors, especially highlighted when they are used into a distributed computing platform such as grid [2].
Biological data represent very large datasets of different nature, from different sources, with heterogeneous models: protein three-dimensional structures, functional signatures, expression arrays, etc. Bioinformatics experiences use numerous methods and algorithms to analyze whole biological data which are available to the
community [3]. For each domain of Bioinformatics, they are several different high-quality
programs that are available for computing the same dataset in as many ways. But most bioinformatics programs are not adapted to distributed platform. One important disadvantage is that they are only accessing data with local file system interface to get the input data and to store their results, an other one being that these data must be unencrypted.

The European EGEE grid
The Enabling Grids for E-sciencE project (EGEE) [4], funded by the European Commission, aimed to build on recent advances in grid technology and to develop a service grid infrastructure such as described by Foster et al. at the end of 1990s [5].
The EGEE middleware provides grid users with a "user interface"(UI) to launch a job. Among the components of the EGEE grid: the "workload management system"(WMS) is responsible of job scheduling. The central piece is the scheduler (or "resource broker") that determines where and when to send a job on the "computing elements"(CE) and get data from the "storage elements"(SE). The "data management system"(DMS) is a key service for our bioinformatics applications. Having efficient usage of DMS will be synonymous of good distribution of our protein sequence analysis applications. Inside the DMS, the "replica manager system" (RMS)
provides users with data replica functionalities. But there is no available encryption service onto the production grid of EGEE, built upon the LCG2 middleware.

"EncFile"encrypted file manager
We have developed the EncFile, encrypted file management system, to provide our bioinformatics applications

with facilities for computing sensitive data on the EGEE grid. The cipher algorithm AES (Advanced Encryption Standard) is used with 256 bits keys. And to bring fault tolerance properties to the platform, we have also applied a M-of-N technique described by Shamir for secret sharing [6]. We split a key into N shares, each stored in a different server. To rebuild a key, exactly M of the N shares are needed. With less than M shares, it is impossible to deduce several bits or even one of them.
The "EncFile"system is composed of these N key servers and one client. The client is doing the decryption of the file for the legacy application, and is the only component able to rebuild the keys, securing their confidentiality. The transfer of the keys between the M servers and the client is secured with encryption and mutual authentication. In order to determine user authorization, the EncFile client send the user proxy to authenticate itself. Nonetheless, to avoid that a malicious person creates a fake EncFile client (e.g. to retrieve key shares), a second authentication is required with a specific certificate of the EncFile system.
As seen before, most bioinformatics programs are only able to access their data through local file system interface, and also not encrypted. To answer to these 2 strong issues, we have combined the EncFile client and the Parrot software [7]. The resultant client (called Perroquet in Figure 1) acts as a launcher for applications, catching all their standard IO calls and replacing them with equivalent remote calls to remote files. Perroquet understands the logical file name (LFN) locators of our biological resources onto the EGEE grid,
and do on-the-fly decryption. This has mainly two consequences: (i) higher security level because decrypted file copies could endanger data, (ii) better performances because files aren't read twice to locally copy and to decrypt.
Thus, the EncFile client permits any applications to transparently read and write remote files, encrypted or not, as if they were local and plain-text files. We are using EncFile system to secure sensitive biological data on the EGEE production platform and to analyze them with world-famous legacy bioinformatics applications such as BLAST, SSearch or ClustalW.

Conclusion
We have developed the EncFile system for encrypted files management, and deployed it on the production platform of the EGEE project. Thus, we provided grid users with a user-friendly component that doesn't require any user privileges, and is fault-tolerant because of the M-of-N technique, used to deploy key shares on several key servers. The EncFile client provides legacy bioinformatics applications with remote data access, such as the ones used daily for genomes analyses.

Acknowledgement
This works was supported by the European Union (EGEE project, ref. INFSO-508833). Authors express thanks to Douglas Thain for the interesting discussions about the Parrot tool.

References
[1] Jacq, N., Blanchet, C., Combet, C., Cornillot, E., Duret, L., Kurata, K., Nakamura, H., Silvestre, T., Breton, V. : Grid as a bioinformatics tool. , Parallel Computing, special issue: High-performance parallel bio-computing, Vol. 30, (2004).
[2] Breton, V., Blanchet, C., Legré, Y., Maigne, L. and Montagnat, J.: Grid Technology for Biomedical Applications. M. Daydé et al. (Eds.): VECPAR 2004, Lecture Notes in Computer Science 3402, pp. 204–218, 2005.
[3] Combet, C., Blanchet, C., Geourjon, C. et Deléage, G. : NPS@: Network Protein Sequence Analysis. Tibs, 25 (2000) 147-150.
[4] Enabling Grid for E-sciencE (EGEE). Online: www.eu-egee.org
[5] Foster, I. And Kesselman, C. (eds.) : The Grid 2 : Blueprint for a New Computing Infrastructure, (2004).
[6] Shamir., A. "How to share a secret". Communications of the ACM , 22(11):612–613, Nov. 1979.
[7] Thain, D. and Livny, M.: Parrot: an application environment for data-intensive computing. Scalable Computing: Practice and Experience 6 (2005) 9-18

Summary

Biomedical applications are pilot ones in the EGEE project and have their own virtual organization: the "biomed"VO. Indeed, they have common security requirements such as electronic certificate system, authentication, secured transfer; but they have also specific ones such as fine grain access to data, encrypted storage of data and anonymity. Thus, biomedical data need to be gridified with a high level of confidentiality because they can concern patients or sensitive scientific/industrial experiments. The solution is then to encrypt the data on the Grid storage resources, but to provide authorized users and applications with transparent and unencrypted access.
We have developed the EncFile system for encrypted files management, and deployed it on the production platform of the EGEE project. Thus, we provided grid users with a user-friendly component that doesn't require any user privileges, and is fault-tolerant because of the M-of-N technique, used to deploy key shares

on several key servers. The EncFile client provides legacy bioinformatics applications with remote data access, such as the ones used daily for genomes analyses.

Author: Dr BLANCHET, Christophe (CNRS IBCP)

Co-authors: Prof. DELEAGE, Gilbert (CNRS IBCP); Mr MOLLON, Rémi (CNRS IBCP)

Presenters: Dr BLANCHET, Christophe (CNRS IBCP); Mr MOLLON, Rémi (CNRS IBCP)

Session Classification: 1a: Life Sciences

Track Classification: Life Science