Service to Encrypt Biological Data on Grid

Mollon R., Blanchet C. and Deleage G.

Pôle Bioinformatique de Lyon – PBIL

Institut de Biologie et de Chimie des Protéines

IBCP CNRS UMR 5086

Lyon – Gerland, France

R.Mollon@ibcp.fr

C.Blanchet@ibcp.fr

Information Society and Media

Institute of Biology and Chemistry of Proteins

- **French CNRS Institute, associated to Univ. Lyon 1**
  - Life Science
  - About 160 people
  - http://www.ibcp.fr
  - Located in Lyon, France

- **Study of proteins in their biological context**
  - Approaches used include : integrative cellular (cell culture, various types of microscopies) and molecular techniques, both experimental (including biocrystallography, and nuclear magnetic resonance) and theoretical (structural bioinformatics)
- **Three main departments, bringing together 13 groups**
  - Topics such as cancer, extracellular matrix, tissue engineering, membranes, cell transport and signalling, bioinformatics and structural biology

Outlines

- **Security Bioinformatics Requirements**
  - Data confidentiality : patient, industrial
- **Encrypted File Management System**
  - Key sharing between several servers
  - Encryption / Decryption client
  - Transparent access to remote file : Perroquet

Bioinformatics requirements

- **Certificate management**
  - For all entities (like users, services, Web portals, ...)
  - Renew and revoke mechanisms

DONE

- **Fine grain access to data**
  - Access Control Lists (ACL) support
  - The owner can do modifications

In Progress

- **Data encryption**
  - Long-term storage of encrypted data
  - Transparent (unencrypted) access for authorized users

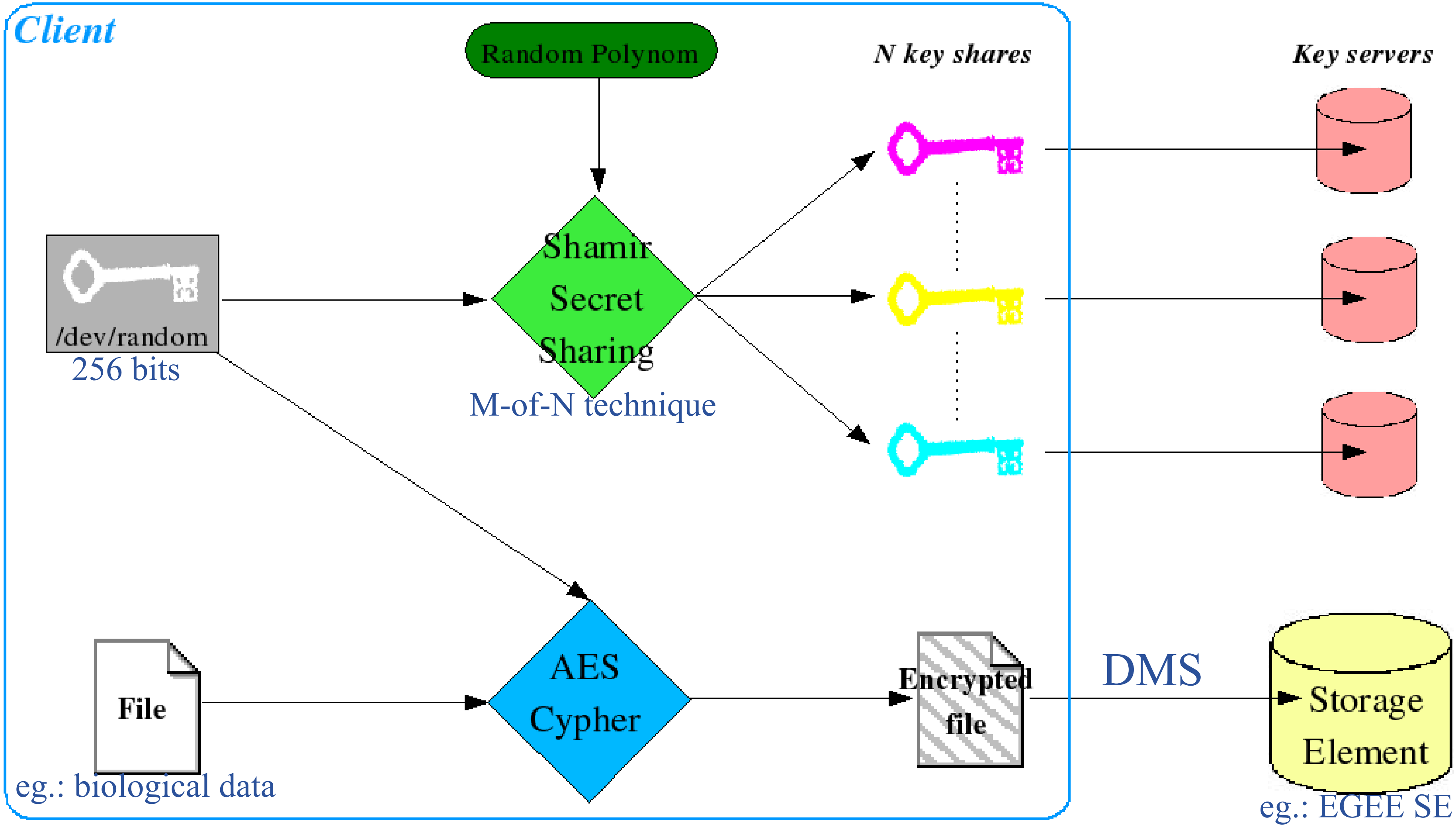
In Progress

*According to EGEE **requirement database**,*

Presentation of EncFile

- **Encrypted File Management System on a Grid**
  - Files are encrypted with AES algorithm and 256-bit keys
    - Fast encryption/decryption (~30 Mbits/s) and good security properties
  - Encrypted files are stored with the Grid Data Management
    - Authorizations are ensured by the grid mechanism (LFC)
      - *Possible thanks to proxy delegation to key servers*
  - Mutual authentication between clients and servers
  - Communications are secured thanks to OpenSSL
  - Secure and survivable cryptographic key storage
    - Several keys servers which can be on different sites
    - M-of-N technique : Shamir's Secret Sharing Algorithm
      - *Keys are split into N shares*
      - *Key reconstruction needs M of these N shares*
      - *With less than M shares, no information can be deduced*

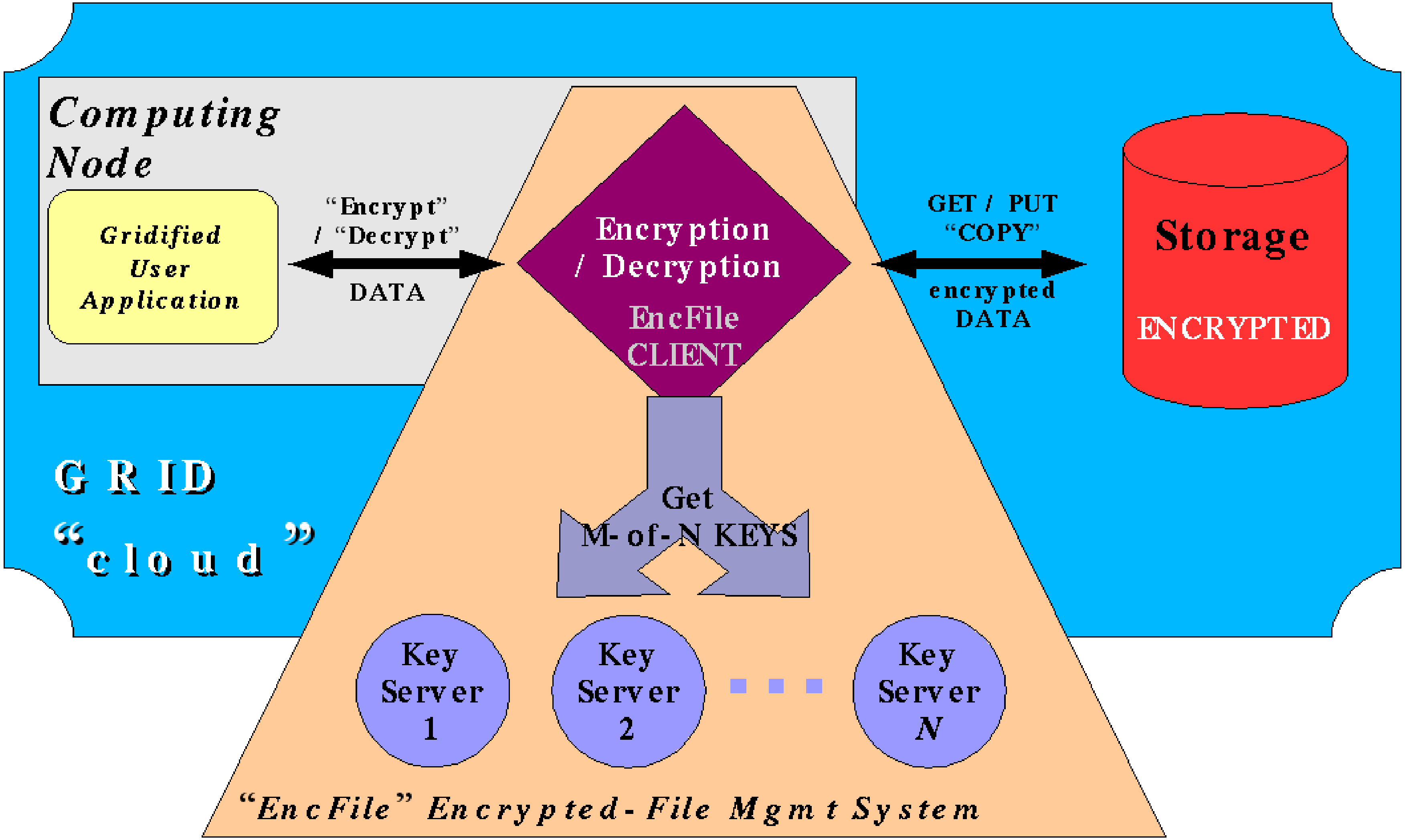
EncFile: Encryption Scenario



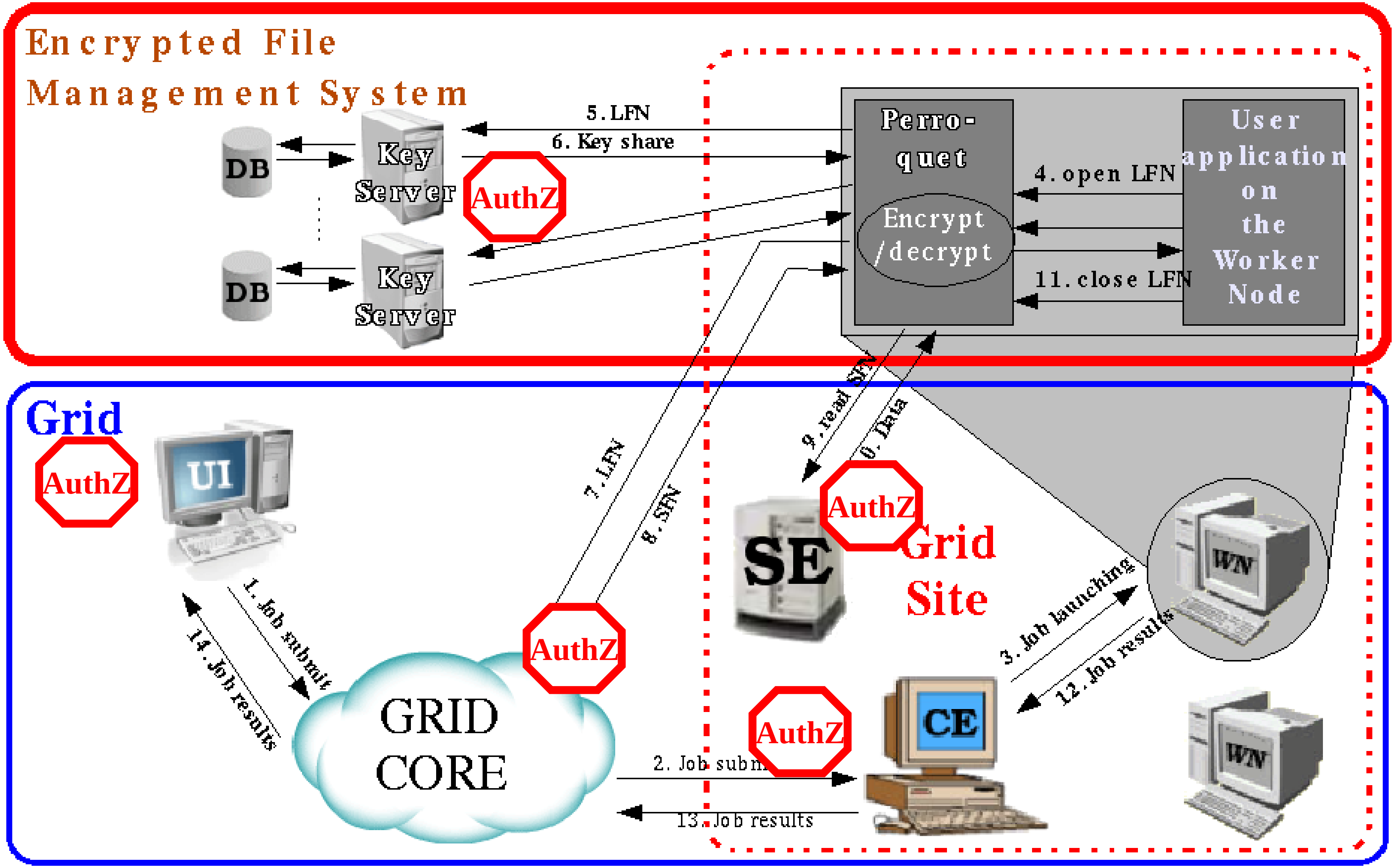
Perroquet : File Access Forwarder

- **Based on Parrot Software**
  - Collaboration with D. L. Thain, Univ. Notre Dame, USA
  - Permits to applications to transparently access to remote files
  - Supports several protocols : chirp, http, ftp, gsiftp, dcap, rfio, glite, nest
- **Added functionalities**
  - LFN namespace
    - Checks authorizations with LFC server
    - Resolve LFN into a Gsi-FTP url
    - Read, write, create are supported
    - No GFAL support, because of local-site-only limitations of RFIO
  - On-the-fly encryption and decryption
    - Integration of the EncFile client in Perroquet

EncFile: Decryption Scenario

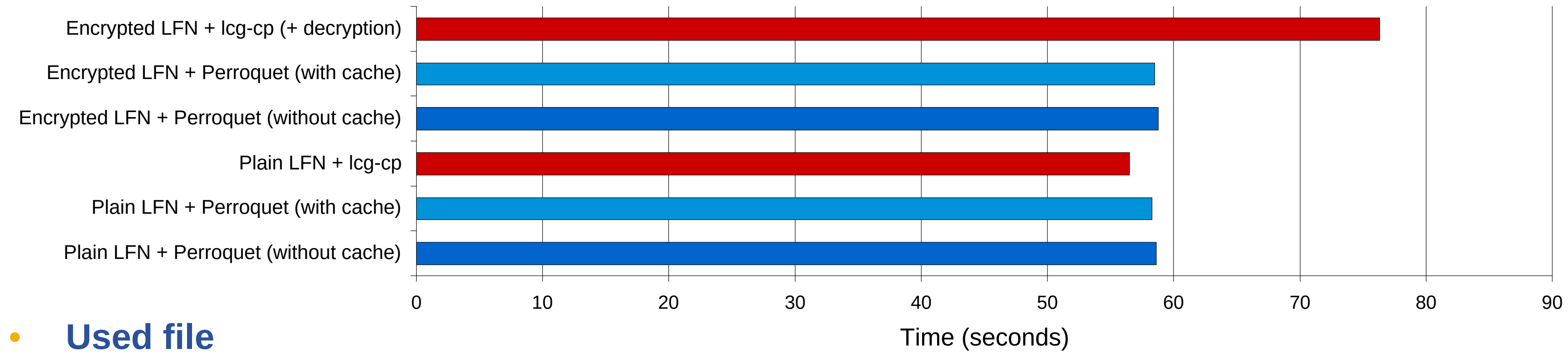


EncFile: Architecture



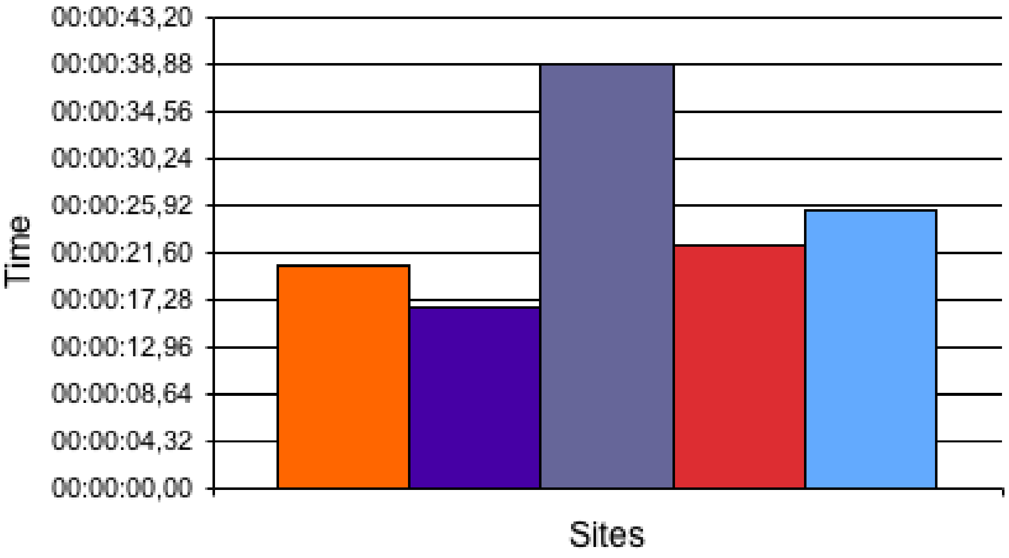
EncFile: Performances

Time to download a 205-MB gridified file



- **Used file**
  - 500,000 protein sequences
  - ~200 MB
  - replicated to several SEs
- **Experiment**
  - Copying this file locally on the used worker node

Near-SE download time on different EGEE grid sites



C. Blanchet, R. Mollon and G. Deleage: Building an Encrypted File System on the EGEE grid: Application to Protein Sequence Analysis. IEEE Proceedings of ARES 2006, Vienna, 20-22 April

Perroquet Usage

- **Put a file on the Grid with encrypting it**
  - perroquet -e cp /local/path/to/my/file lfn:/grid/path/to/my/file
- **Get a encrypted file from the Grid with decrypting it**
  - perroquet cp lfn:/grid/path/to/my/file /local/path/to/my/file
- **Run a blast on swissprot**
  - perroquet blastall -i my_sequence.fas -d lfn:/grid/biomed/db/swissprot/last/sprot.fas -o blast.out -p blastp

Conclusion

EncFile

Availability	Compatible with the EGEE production platform since Aug 2005 (LCG2) Not specific to a given middleware: port to other grid
Integration	Transparent access for legacy applications (Perroquet) C++ API, Command line interface (EncFile client)
Authorization	Compatible with the used middleware: *e.g.* LFC ACL database
Encryption cipher	AES algorithm, 256-bit keys
Encryption	Explicit
Decryption	Implicit
Encryption/decrypt. location	In local memory of the computing node
Key storage	M-of-N technique (Shamir *et al.*)
Encryption flag	Comptaible with the used middleware: *e.g.* stored in LFC metadata
Deployment	Used in GPS@ Web Portal and its applications http://gpsa-pbil.ibcp.fr
Application Specificity	No

Perspectives

- **Key redistribution**
  - Counter a key server compromising
  - Redistribution protocol
    - Key reconstruction isn't needed
    - Decentralized protocol
- **Verifiable secret sharing protocol**
  - Verify the correctness of (re)distribution
  - Deal with fault server during (re)distribution
- **Interface to CHIRP storage component**
  - Integrate EncFile in Parrot system
  - In Collaboration with D.L. Thain (Univ. Notre Dame)

Acknowledgement

o **Science collaborators**

- D.G. Thain (Univ. ND, US)
- Y. Denneulin (IMAG, Fr)
- Members of the EU-FP6 EGEE project

o **Team collaborators**

- C. Blanchet
- R. Mollon (EGEE fellow)
- V. Daric (EMBRACE fellow)
- C. Combet
- G. Deléage (Team Leader)

[Gri]-P-{PS}
Grid Protein Pattern Scanning

Work supported in part by projects:
French ACI GriPPS (FR-GRID-PPL02-05),
EU-FP6 EGEE-II (INFSO-RI-031688)
EU-FP6 EMBRACE (LHSG-CT-2004-512092)

EMBRACE Grid
Network of excellence

egee
Enabling Grids for E-sciencE