

NASA/CR-2003-212168
NIA Report No. 2003-02



Additional Security Considerations for Grid Management

Thomas M. Eidson
National Institute of Aerospace, Hampton, Virginia

The NASA STI Program Office . . . in Profile

Since its founding, NASA has been dedicated to the advancement of aeronautics and space science. The NASA Scientific and Technical Information (STI) Program Office plays a key part in helping NASA maintain this important role.

The NASA STI Program Office is operated by Langley Research Center, the lead center for NASA's scientific and technical information. The NASA STI Program Office provides access to the NASA STI Database, the largest collection of aeronautical and space science STI in the world. The Program Office is also NASA's institutional mechanism for disseminating the results of its research and development activities. These results are published by NASA in the NASA STI Report Series, which includes the following report types:

- **TECHNICAL PUBLICATION.** Reports of completed research or a major significant phase of research that present the results of NASA programs and include extensive data or theoretical analysis. Includes compilations of significant scientific and technical data and information deemed to be of continuing reference value. NASA counterpart of peer-reviewed formal professional papers, but having less stringent limitations on manuscript length and extent of graphic presentations.
- **TECHNICAL MEMORANDUM.** Scientific and technical findings that are preliminary or of specialized interest, e.g., quick release reports, working papers, and bibliographies that contain minimal annotation. Does not contain extensive analysis.
- **CONTRACTOR REPORT.** Scientific and technical findings by NASA-sponsored contractors and grantees.
- **CONFERENCE PUBLICATION.** Collected papers from scientific and technical conferences, symposia, seminars, or other meetings sponsored or co-sponsored by NASA.
- **SPECIAL PUBLICATION.** Scientific, technical, or historical information from NASA programs, projects, and missions, often concerned with subjects having substantial public interest.
- **TECHNICAL TRANSLATION.** English-language translations of foreign scientific and technical material pertinent to NASA's mission.

Specialized services that complement the STI Program Office's diverse offerings include creating custom thesauri, building customized databases, organizing and publishing research results . . . even providing videos.

For more information about the NASA STI Program Office, see the following:

- Access the NASA STI Program Home Page at <http://www.sti.nasa.gov>
- Email your question via the Internet to help@sti.nasa.gov
- Fax your question to the NASA STI Help Desk at (301) 621-0134
- Telephone the NASA STI Help Desk at (301) 621-0390
- Write to:
NASA STI Help Desk
NASA Center for Aerospace Information
7121 Standard Drive
Hanover, MD 21076-1320

NASA/CR-2003-212168
NIA Report No. 2003-02



Additional Security Considerations for Grid Management

Thomas M. Eidson
National Institute of Aerospace, Hampton, Virginia

National Aeronautics and
Space Administration

Langley Research Center
Hampton, Virginia 23681-2199

Prepared for Langley Research Center
under Contract NAS1-02117

September 2003

Available from the following:

NASA Center for AeroSpace Information (CASI)
7121 Standard Drive
Hanover, MD 21076-1320
(301) 621-0390

National Technical Information Service (NTIS)
5285 Port Royal Road
Springfield, VA 22161-2171
(703) 487-4650

ADDITIONAL SECURITY CONSIDERATIONS FOR GRID MANAGEMENT

Thomas M. Eidson¹

ABSTRACT

The use of Grid computing environments is growing in popularity. A Grid computing environment is primarily a wide area network that encompasses multiple local area networks, where some of the local area networks are managed by different organizations. A Grid computing environment also includes common interfaces for distributed computing software so that the heterogeneous set of machines that make up the Grid can be used more easily. The other key feature of a Grid is that the distributed computing software includes appropriate security technology. The focus of most Grid software is on the security involved with application execution, file transfers, and other remote computing procedures. However, there are other important security issues related to the management of a Grid and the users who use that Grid. This note discusses these additional security issues and makes several suggestions as how they can be managed.

1. INTRODUCTION

The term, Grid, is used to denote a distributed software and hardware environment for executing scientific and engineering applications over wide area networks. Much of the focus is on the protocol architecture for interoperability; specifically, a set of standards that support the straight-forward and efficient execution of distributed applications on a collection of heterogeneous computers that are located at a number of different sites. An important aspect of the Grid is that these different sites will generally be managed by different organizations. This means that the management of users, machines, and networks is generally more difficult than managing just a local area network (LAN). These difficulties revolve around security issues related to the network access of computers, but also involve the communication between the various members of a virtual organization that includes users and managers from a number of actual organizations that have agreed to create a Grid.

The Globus Project has developed core software for such an environment, called the Globus Toolkit.[1] The toolkit includes the Globus Security Infrastructure (GSI), a library that implements a set of core protocols and procedures to support a good security model for a Grid. The GSI security model has proved to be an excellent model and has been widely adopted, both as part of the Globus Toolkit and as the core for other Grid software environments. However, GSI does not solve all the security issues related to managing a Grid. In this note, additional issues are presented and possible solutions are discussed.

The ideas presented herein were developed as part of a Grid project, the Tidewater Regional Grid Partnership (TRGP), that was created by two organizations: ICASE at NASA Langley Research

¹National Institute of Aerospace (NIA), 144 Research Dr., Hampton, VA 23666.

Center and the Computational Sciences Cluster laboratory at the College of William and Mary. TRGP used the Globus Toolkit as its core software. Additional security procedures and software were also developed to manage users and user communication. During the project several security issues related to the design of LANs and the management and use of the computers within a LAN were observed. This experience is the primary source of the ideas presented herein.

2. USER COMMUNICATION

A Grid virtual organization is generally formed from several member organizations that each manage, either or both, a local area network and a group of users. Each member organization needs to provide one or more people to administer the following functionality.

- ⑩ A voting administrator is authorized to make formal agreements for the member organization within the Grid organization.
- ⑩ A user representative validates the identity of users sponsored by the member organization.
- ⑩ A system administrator manages the Grid software on any machines made available for Grid use.

The Grid organization will need at least one person to manage a certificate authority (CA) unless that functionality is done by some independent organization. The Grid organization could need other personnel who perform various organizational and support tasks.

The GSI as implemented in the Globus Toolkit provides good security for Grid communications within a distributed application. However, additional security needs are associated with communication among the various members of the Grid organization. In general, each member organization will be located at a different physical site. Additionally, the systems and users of each member organization could be at different sites. This means that e-mail or phone calls will be used to communicate between people. For many communications this is not a problem. However, some communications will require formal authorization. Traditionally, this has been done with paper documents that are signed by an authorizing agent. In a Grid organization, this may not be sufficiently efficient.

Some of the most critical communications are between a user at one member organization and the Grid CA. For example, the user must create a certificate that will be used as part of any remote Grid request to authenticate that the request comes from the user. The CA must electronically sign the user certificate, using public key technology, to validate the certificate. The Grid software on each computer is configured to trust the signature of the CA; therefore, the Grid software will trust any request associated with a CA-signed user certificate. [Note, this does not mean that the request will be completed as a separate local authorization process is also required, which will be discussed later.] Since the CA and many users will be located at different sites, the user certificate will need to be sent to the CA and returned, usually by e-mail. If this procedure is not done securely, the result is a serious security hole in the Grid operation. The CA needs to insure that only certificates of valid users are signed and that any signed certificate is sent and received by the correct valid user.

While the above example is a very important need for secure communication, users and administrators will also need to send other secure information at times. One solution is to use a public-private key system with e-mail.[2] Each member of the Grid organization will need to create a public-private key pair and make the public key available to others. [In this note, this public-private key pair will be assumed independent of that used as part of GSI. A system such as Pretty Good Privacy, PGP, is one example as PGP is easy to integrate with e-mail and other document signing or encryption needs. [3] It would be preferred that the same public-private key system be used for both GSI and these additional communications needs. GSI uses SSL[4] to provide a public-private key pair.] Then, important e-mails can be signed with the private key of the sender. The receiver can use the public key of the sender to verify that the message is valid. If the contents of the message are sensitive, then the sender can also encrypt the message using the public key of the receiver. The private key of the receiver can be used to decrypt the message if necessary.

All of the above is good except that the security hole has still not been filled. How does the user of a public key know if that key really belongs to the right person? The best approach would be to obtain the public key as part of a physical meeting with the owner of that key. One could even require that formal identification, such as a passport, be made available. [For the very paranoid, fingerprinting or DNA testing could be included.] But a meeting between every pair of people who need to communicate with each other is not generally reasonable. This is where the role of a user representative for each member organization can provide a solution. Rather than direct meetings between everybody, a trust chain can be used to provide faith in each public key. The voting representatives of the Grid organization need to choose a Grid manager who is trusted by all as the center of the trust chain. The Grid manager and each user representative need to meet and exchange public keys. As part of the key exchange, key fingerprints should also be exchanged. [A fingerprint is the message digest of a key, in this case the public key. A message digest is a number with significantly fewer bits than the key. It is produced by a function that uses the key as input and produces the message digest as output. The message digest is not unique as more than one key can produce the same output. However, it is computationally improbable to find two keys that produce the same fingerprint. If two copies of a key have the same fingerprint, then it is probable that the two copies are the same key.[2]] Verification of the fingerprint adds a check on the validity and correctness of the public keys being exchanged. The Grid manager would sign the public key of each user representative. Each user representative would then be responsible for having a physical meeting with each user at the member site that they represent. The user representative would exchange public keys with the user as well as sign the public key of the user. Now all public keys can be put in some public location so that anyone can get one. User A at site X can trust the public key of user B who is remotely located at site Y. The public key of user B is signed by the user administrator at site Y, the public key of the site Y administrator is signed by the Grid manager, user B has a version of the public key of the Grid manager that was signed by the site X administrator, and user A trusts the site X administrator.

It is reasonable for the Grid manager to also have the role of CA. If not, a public key for the CA could be signed by the Grid manager and distributed. Alternately, the Grid manager could receive any requests that are sent to the CA. This would be particularly useful if the CA was managed by an independent party.

The above model can be extended to allow each member organization to have multiple sites with one user representative at each site. Even then there will be situations where a user is not located at a member site. If the Grid member organizations agree, a phone verification can be substituted. In this case the procedure for identifying the person being called needs to be consistent with the level of desired security. Possible procedures include:

- ⑩ calling the person at a home phone number,
- ⑩ having a third party make the phone introduction, and
- ⑩ exchanging ID documentation via the mail that includes a phrase to initiate the phone conversation.

With good ingenuity, a phone conversation can meet many desired levels of security requirements. The fingerprint verification can be used as part of a phone-call key verification. The public key is long and reciting the complete key for verification will be prone to mistakes.

3. USER AUTHORIZATION

The above discussion deals with the authentication of a user. The CA-signed Grid certificate can be used as part of a remote computer request to describe and to verify the identity of a member of the Grid organization. It does not authorize that person to use any computer resources of any actual organization belonging to the Grid, even those resources at the local site. The user or some representative of a group of users must request that a user be given authorization on specific machines at any site where actually computing is to be done. This usually means that the user needs to have a local account at this remote site. [The concept of a general, possibly temporary account that can be used by more than one remote user is a current topic of research within the Grid development community.] Hopefully using secure user communications, such a request is made and the remote site system administrator will take the necessary actions to authorize the user. In the Globus Toolkit, the primary way this is done is to add the user to a mapfile. In the mapfile, the Grid identity that is part of the Grid certificate is associated with (or mapped to) a user name for a valid local account. Any remote access by the user must first be sent to a gatekeeper daemon at the remote site. The gatekeeper verifies that the Grid certificate of the user is correctly signed by the Grid CA and thus the Grid identity in the certificate can be trusted. The gatekeeper then uses the mapfile to determine if there is an authorized local account for the request from the user. If so, the request is executed under this local account. [More sophisticated authorization procedures where a user is given restricted access are possible.]

The request from a user that is sent to a remotely located system administrator to obtain authorization at that remote site is another example where secure communication between Grid members is important. In many cases the system administrator at a remote site will not know anything about the person making the authorization request. One possibility is that another user at the remote site is working with the requester and will vouch for the user. In other cases, the user is making the request under some general agreement between the two sites. In either case, the system administrator will need to contact an appropriate person and obtain permission to authorize the requester. This communication may or may not be efficient depending on the number of users that are making such requests.

One possibility is to create a user information database. The database would contain the user Grid identity and contact information. [The contact information should include both slow and fast methods of contact. For example, a remote user job could be causing a problem at a remote site. The user will need to be contacted quickly; otherwise, the job may have to be canceled.] The database could also contain information related to the authorization process. For example, site X could include Grid members who were both U.S. citizens and foreign nationals. Site Y could have an agreement with site X that only U.S. citizens will be given automatic authorization. Therefore, the citizenship of Grid members could be part of the database. Any database request could be signed in a similar manner to the public key trust chain procedure discussed above. This would make it possible for the authorization process to be both efficient and secure. The user could make an authorization request solely by e-mail. The system administrator would use the information in the secure database to determine if authorization was appropriate.

4. LOCAL AREA NETWORK SECURITY ISSUES

On the surface, the two primary topics of this note, distributed communication and security, are in direct conflict. Some organizations who would like to be part of a virtual Grid organization are also implementing stronger firewall configurations which prevent some of the communication functionality needed by many distributed applications. This can lead to debates over which functionality is more important. The problem stems from the decision to design most LANs as a collection of machines configured for minimal security behind a single firewall configured for high security. In most organizations, the security requirements for different machines, applications, and data are not the same. It is suggested that a hierarchical LAN design can be used to satisfy the requirements of security and distributed communication. For example, the entire LAN would be protected by a primary gateway configured for good, but minimal security. The LAN would then have multiple subnets, each protected by a secondary gateway configured with the appropriate security needed for the machines on the associated subnet. The following discussion will define several classes of computers found within LANs along with their security needs. These are only examples as each organization will have different needs. The main goal of these examples is to suggest how a LAN can be configured to support the security and open communication needs of Grid computing.

In the early days, many LANs were formed as a set of servers, many of which were used as desktop machines. Client-only, desktop machines, many with little or no security capability, were gradually added to the networks and security was focused on a primary gateway that attempted to protect the entire LAN. Client-only, desktop machines are distinguished by the fact that most or all network activity is originated at that machine. Often these machines are maintained by the user, possibly with minimal help from a system administration staff. For these and other reasons, security is generally weaker on such machines. These machines could be configured on a separate internal network with a gateway configured as a medium- to high-security firewall. They can even use private network addresses which both improve security and reduce the number of public addresses that need to be owned and maintained. In general, the firewall would only allow remote network communication that originated from inside the subnet. [For remote access needs such as remote system administration, the firewall could also be configured to allow traffic from specific authorized machines.] This configuration will be referred to as a private subnet in the following discussion, even if public network addresses are

used.

A private subnet does not have to be limited to client-only, desktop uses. They can be used to manage information and machine access that have critical security requirements. The network architecture would be the same, just the firewall configuration would be more restrictive. For example, no general outside access could be allowed. Even e-mail and web access could be prohibited. Any information that needed to be exchanged might have to be copied to or from a machine, a security gateway, outside the private subnet but within the organization's LAN. These transfers could use a special set of file transfer software that was designed with appropriate encryption and authentication technology. The point is that protecting sensitive data usually defines the most critical security needs. Having restrictive, less convenient procedures to access that data is often reasonable. By separating a data subnet from a client-only subnet, the high-security network needs do not have to restrict the more routine network activities of users of client-only machines. [For extreme security needs, outbound transfers could be further restricted or prevented. A machine in a physically secure office could even be setup that was the only way to access the data on this subnet.]

A private subnet could include servers. These could be limited for use within a subnet or they could be used to create a secure server environment. The subnet firewall would be configured to limit requests to the servers so that only intended operations can be preformed. A secure server subnet would be the most difficult to design. All the applications and communication software on these machines would have to be carefully designed to meet appropriate security requirements. Writing such software has historically proved to be more costly. The problem is that it is difficult to foresee every possible intrusion strategy. It is not the purpose of this note to provide guidelines for writing such applications. In the remainder of this section an alternate approach to LAN management is suggested that will be a more cost effective solution.

With the high security needs met, the remainder of a LAN can be configured to meet the needs of open communication. However, even this part of a network can be configured into subnets with different security requirements. At the low security end are public information providers. The most common examples are http (web) and ftp servers. Other examples might be a machine for remote users to be able to access with a wide range of protocols, some of which might be considered as security risks (e.g. Telnet). The information and even the operating system on these machines would be considered as temporary. While they should be configured with good, minimal security features, the basic approach would be to maintain copies of the information and operating system configuration in a separate secure location. Periodically, the operating system and all other data would be replaced. Modern operating systems can be setup so that this procedure can be done automatically during low-usage periods. The frequency of these updates would depend on the nature of the organization and the information being stored on the machines. Since these machines are used to store publicly available information, the security concern is to prevent someone from gaining access to the machines for malicious purposes. The operating system update would delete any Trojan horses, viruses, or other malicious access. One could even change the system passwords during each update. Monitoring the size and checksums of key files and directories can be used to check for other types of invasions. The security strategy on these machines would focus on quickly recovering from a hacking attack, not on preventing the attack.

The remaining category would have security needs that fall between the cases mentioned above. It is suggested that many scientific and engineering servers would fall in this category. In many cases the information being computed will be published at some point. And, often the developer of the applications will give the codes to others along with support. Much of science is about open information. However, even in these cases the application owner generally wants to know and possibly to control how information is being shared. Also, an application user does not want an intruder to tamper with an executing application, even accidentally. Machines in this class might be managed similar to the above low-security machines, but with more emphasis on logging system integrity. Random checksum and file size logs of a limited number of system files could be kept. These logs could then be inspected to determine if anyone is tampering with the system. The frequency of the writing of these log records and the inspections would need to be selected to match the desired security level. While not a failsafe procedure, the cost effectiveness and the resulting security level should be a good match.

Such a machine environment would be compatible with Grid software such as the Globus Toolkit. The Grid gatekeepers and other access daemons could be the only remote access allowed to initiate a session on such machines. Then ports could be left open for applications to setup efficient communications. It is this need for user applications to open communication ports without restriction that is a major basis for separating Grid computers from servers requiring high security.

With such a configuration, it is probably the case that most data and applications will need to be stored in a more secure area when not part of an active application. This means that significant data will need to be moved as part of most application executions. This can be problematic for several reasons. One, the network performance has to be designed so that any transfers have acceptable performance. This is generally only a problem for very large data transfers. Standard local area network technology has proved to be satisfactory for most application needs. However, site to site performance where data is transferred over wide area networks can be costly. The bigger issue relates to programming. Data and application files will need to be moved more frequently in such an environment. In simple cases, it is satisfactory to use scripts to move files before and after an execution. However, scripts may not be sufficiently flexible, particularly for more complex needs. Using a hierarchical network design will increase the need for portable applications. Security needs may be different for development and production execution. Porting applications to different machines is considered a burden. These problems are generally not fundamentally difficult. The problem is that hardware and software have evolved rapidly and users would rather focus on the science being computed rather than the managing complex distributed computing requirements. Therefore, it is important to include a good programming environment as part of the design of such a network.

Clearly, some distributed applications would need the added security of a more secure firewall. A private subnet can still be used to maintain a high-security Grid, web server, or data server. To reiterate the main point, a local area network needs to be configured to match the security, accessibility, and performance needs of the applications and data being used. A LAN protected by a single firewall will probably not meet all those needs.

5. CERTIFICATE AUTHORITY ISSUES

A practical security system is generally always based on trust. Both the GSI model and the ideas presented above focus this trust on a central person (or set of persons), the certificate authority. This means that the CA for a Grid must be chosen so that all members of the Grid organization have sufficient trust in this person or persons. This may require that the CA be run in a manner where the actions of any one person can be monitored or reviewed. But there is another criteria that is often overlooked. This is that the CA must be run so an individual who knows critical information can resign from the CA without invalidating the Grid security system.

For example, the GSI system is based on public-private key technology. A CA must generate a public-private key set and use the private key to sign certificates and member public keys. One person in the CA group might have direct access to the private key and know the passphrase that accesses that key. When this target person leaves the CA group the passphrase can be changed, but this does not solve the security problem. If the target person keeps a copy of the old private key, then the old passphrase can still be used for this old copy to sign certificates even though the new passphrase is needed to access the new official private key. This means that all certificates created before the target person left cannot be distinguished from certificates signed by the target person after leaving the group. One could create a new public-private key pair for the CA, then recreate and redistribute a new set of certificates. This would solve the security problem, but would generally be inefficient.

There are ways to solve the problem such as modifying the security software to use additional information but this will generally result in a more complicated security scheme. Also, a Grid organization generally has to use available software and the simple operation of the GSI system is otherwise satisfactory. The following suggestions provide ideas on how to avoid the problem at the Grid organization level. The example given below is targeted for organizations such as government laboratories, university departments, and small businesses who run a certificate authority; not for professional security companies. The example is meant to illustrate a possible approach, not to guarantee a foolproof system.

The core of the problem is that access to the private key and knowledge of the passphrase by the same person is a security risk. For this and other reasons, it is probably a good idea if the computer, on which the private key resides and is used, should not have general access over a network. For example, one person can maintain the CA computer and a different person would be the certificate signer who knows the passphrase. The signer would be provided scripts or other software on some remote system that would communicate with software on the secure computer. This software would transmit a request from the signer (i.e., a certificate to be signed) along with the passphrase (appropriately encrypted with the public key associated with the software on the secure computer). The secure computer could be configured to only receive remote requests from the signer's computer and then only on ports used by the secure communication software. These ports could even be changed at some appropriate frequency to further enhance security.

Even though the signer would not have direct access to the private key of the CA, the security problem is not completely resolved. The secure computer manager has access to the private key and control of the computer. If the communication software is designed incorrectly, the

computer manager could intercept the passphrase. The communication software must be designed to preclude this possibility. Then both the computer manager and the certificate signer can leave without causing a serious problem.

Since every CA procedure performed by the signer must go through the secure communication software, every request can be logged. This will allow the work of the signer to be monitored, which provides additional security.

6. SUMMARY

The security plan for a Grid must consist of a set of appropriately balanced procedures. These procedures should not be too severe or they will get ignored and security will be compromised. They also should all be at the same general security level or a weak procedure will undermine the extra work used to implement a more severe security procedure.

REFERENCES

- [1] Globus Project webpage, *www.globus.org*, Argonne National Laboratory, February, 2003.
- [2] C. Kaufman, R. Perlman, and M. Speciner, **Network Security**, Prentice Hall, Eaglewood Cliffs, NJ, 1995.
- [3] PGP webpage, *www.pgp.com*, PGP Corp., Palo Alto, CA, February, 2003.
- [4] OpenSSL Project webpage, *www.openssl.org*, OpenSSL Project, February, 2003.

REPORT DOCUMENTATION PAGE

*Form Approved
OMB No. 0704-0188*

The public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.

PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

1. REPORT DATE (DD-MM-YYYY)			2. REPORT TYPE			3. DATES COVERED (From - To)		
4. TITLE AND SUBTITLE					5a. CONTRACT NUMBER			
					5b. GRANT NUMBER			
					5c. PROGRAM ELEMENT NUMBER			
6. AUTHOR(S)					5d. PROJECT NUMBER			
					5e. TASK NUMBER			
					5f. WORK UNIT NUMBER			
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)					8. PERFORMING ORGANIZATION REPORT NUMBER			
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)					10. SPONSORING/MONITOR'S ACRONYM(S)			
					11. SPONSORING/MONITORING REPORT NUMBER			
12. DISTRIBUTION/AVAILABILITY STATEMENT								
13. SUPPLEMENTARY NOTES								
14. ABSTRACT								
15. SUBJECT TERMS								
16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON			
a. REPORT	b. ABSTRACT	c. THIS PAGE			19b. TELEPHONE NUMBER (Include area code)			