Available online at www.prace-ri.eu

Partnership for Advanced Computing in Europe

Security in HPC Centres

Michał Nowak, Gerard Frankowski, Norbert Meyer
Poznań Supercomputing and Networking Center, Poznań, Poland (PSNC)

Erhan Yilmaz, Okan Erdogan
National Center for High Performance Computing of Turkey (UHEM)

Contributors:
Jean-Philippe Nominé, François Robin
CEA/DIF/DSSI - CEA/DAM Ile-de-France, Bruyères-le-Châtel 91297 Arpajon Cedex France

Abstract

Securing the HPC infrastructure is an important task. The level of awareness regarding the importance of this topic is high, but the level of investments and skills required to organise a proper protection make it a difficult task, with contrasted levels of solutions and practices observed. There are a huge number of security threats coming from both the Internet and internal networks and, despite the fact that it may seem as a high cost, it is crucial to introduce an adequate level of security to the infrastructure, because the costs of losing data are usually much higher.

Based on a survey of the PRACE community, this paper describes security technologies used in data centres and especially its subset, i.e. HPC centres. It gives a set of general recommendations concerning how to enhance security of the HPC infrastructure.

Table of content

Introduction

IT security is a complex and multifaceted problem. The attack vectors may vary from a malicious employee trying to sabotage the network from the inside to a distributed denial of service attack coming from thousands, or even millions of hosts. Additionally, one of the best known IT security experts, Bruce Schneier[a], once said that security is a process, not a product. It means that the HPC infrastructure will never be completely secure. It may be resilient to the known attacks at a given time but, since new attacks are being invented every day, its security level degrades with time, if not continuously improved. Fortunately new protection methods are invented, too.

Due to the big number of resources each HPC centre has, they are interesting and tempting targets for the attackers. Breaking security of one of HPC centres gives the attacker a set of resources for further use. Moreover, numerous HPC centres participate in national and international initiatives (like PRACE) or projects that join parts of their infrastructures in dedicated networks, the internal traffic of which may be considered as more trusted than the traffic incoming from the Internet. Another factor that must be taken into account is the heterogeneity of the IT infrastructure of the HPC centres, especially those engaged with a significant number of R&D projects. They have different servers, clusters, services and applications. Using different technologies extends the attack vectors, complicates the securing and hardening process, and finally makes it much more difficult to secure and harden all systems. In addition, the used applications may often be immature (in the market sense), not created within a fully formal software development lifecycle and therefore contain more security bugs as well.

A perfect solution for security problems obviously does not, and will never exist. One should not rely on security provided by a single product. The key to having an adequate level of security is called defence-in-depth. According to it, multiple layers of security controls are placed throughout the IT system. Even if one or two layers fail (e.g. a badly written Web application allows to perform an SQL injection attack and read sensitive data), other still exist and may stop the attack (e.g. an intrusion detection system recognizes changes in the common HTTP request pattern to the vulnerable URL and causes the blocking of the source IP address on the firewall). Defence-in-depth also does not guarantee the absolute security, but provides means to make a successful attack so difficult that it becomes cost-ineffective for the attacker.

This White Paper presents the most important technologies used to secure the data centres (chapter 1), short information about a survey sent to PRACE partners (chapter 2) and general security recommendations (chapter 3) and sums-up the conclusions (chapter 4). The aspect of physical security of HPC data centres are outside the scope of this document.

1. Technologies used to secure data centres

1.1. Local network firewalls

A firewall is the main solution for protection against network threats coming both from inside and outside of the HPC network. In the easiest example it is a device installed on the network perimeter and examining the traffic between the internal and external networks.

Guarding the entire network belongs to the tasks of the network firewall and therefore it is usually placed as close to the external network as possible. That prevents unnecessary network flow from occupying resources of LAN and therefore enhances performance. One can, however, imagine the severe consequences of a network firewall failure – no connection between the two sides of the firewall can be established. This is the reason to maintain a high availability (HA) feature. One of the easiest ways of achieving this is redundancy. Instead of having just one firewall guarding the network, it is a common practice to use two or more firewall devices cooperating with each other. Redundancy may be additionally used together with load balancing (both firewalls work in parallel, each processing some part of the network traffic, and when one of them fails, the other starts to process the whole traffic volume) or not (normally only one of the firewalls processes the traffic while the other is inactive, starting its work only after the first fails). The load balancing approach may require a more sophisticated (and therefore more expensive) hardware.

[a] http://www.schneier.com/crypto-gram-0005.html

It is advised to use not only the network firewalls but also support them with the local firewalls. Their task is to guard a single system and therefore they are placed, in contrast to network firewalls, as close as possible to the guarded system. That approach is consistent with the defence-in-depth principle.

The simplest firewall is a packet filter. It operates on the network and transport ISO/OSI layers and has the knowledge about the source and destination of IP addresses and TCP/UDP ports. More advanced firewalls (beside the capability of packet filters) can also keep track of the state of network connections, i.e. perform stateful packet inspection.

Application firewalls are the state-of-the-art among firewalls. They control all seven layers of the ISO/OSI model. Being able to inspect the application layer allows them to control applications or services specifically, which significantly enhances security.

1.2. Antivirus software

Security threats may also come in a form of malicious software (shortly: malware), such as software used by attackers to disrupt computer operation, gather sensitive information, or gain access to computer systems. That group contains: computer viruses, adware, backdoors, malicious BHOs, dialers, fraudtools, hijackers, keyloggers, rootkits, spyware, Trojan horses, worms, etc.

A software combating malware is popularly called anti-virus software, despite the fact that it detects and neutralizes all kinds of malicious software. Usually, it detects the malevolent software basing on signatures, i.e. known patterns of data within the executable code. It is therefore crucial to keep the signatures database up-to-date. However, it is possible for a computer to be infected with a new malware for which signatures are not known yet. To counter the so-called zero-day threats, heuristics can be used, based on differences in software behaviour patterns. This is a much more complicated (but also more powerful) approach and produces much more false positives. Real-time protection mode, as the name suggests, can also block malware threats in real-time.

1.3. Local Intrusion Detection/Intrusion Prevention Systems

Due to the fact that firewalls and antivirus systems are not sufficient to reliably secure the IT infrastructure and prevent attackers from gaining unauthorized access to companies' internal networks, the implementation of additional security measures is necessary. One of the solutions that are designed to help and support Security Administrators are Intrusion Detection Systems (IDS). An IDS is dedicated to monitor critical parts of the infrastructure in order to discover intrusions and intrusion attempts and immediately inform administrators about it. To reduce the negative effects and minimize the impact of a successful breach it is crucial to minimize the window of opportunity when undergoing attacks are unnoticed. Therefore the aim of Intrusion Detection Systems is to discover and notify about attacks and intrusion attempts in near real-time.

The two main categories of Intrusion Detection Systems are:

- Host-based intrusion detection system (HIDS) – an application installed on a specified machine. Its main goal is to monitor certain operating system components as well as applications and network interfaces in order to discover suspicious activity that may be a sign of a break-in attempt.
- Network-based intrusion detection system (NIDS) – monitors network traffic and attempts to discover known attack patterns (signature-based approach) or unusual network activity (anomaly detection approach).

There are other IDS categories (for example Distributed Intrusion Detection System) and approaches to intrusion detection but most of them are based on the two main categories presented above. Intrusion Detection Systems with the active attack prevention mechanism and functionalities like blocking certain ports, resetting suspicious connection, etc. are called Intrusion Prevention Systems.

1.4. Distributed Denial of Service protection

A specific group of network attacks that is worth mentioning are Denial of Service (DoS) attacks. Their concept bases on rejecting a legitimate user access to a certain service. A Distributed Denial of Service (DDoS) attack is a DoS attack that is additionally carried out from numerous different locations. It introduces two main factors: the attack volume is much higher and it is much more difficult (or even impossible) to define the list of attacking IP addresses.

In the 1990s, conducting DDoS attacks required a broad technical knowledge plus much time and effort but this is no longer valid today mostly due to the wide availability of penetration test tools. The motivation has also changed: while

the first DDoS attacks were mostly a way of proving hacking skills or simply a proof of concept, today they have evolved and become actually a commercialized black market industry – a threat to be recognized in the future, especially when dealing with the SaaS marketing model. It is possible to purchase a certain DDoS attack period on the black market as for any other botnet-related activities, like sending SPAM.

There are several types of DDoS attacks: from a very simple and outdated (as Ping of Death) to the more complex and sophisticated ones, which combine a number of attack methods and use amplification techniques. It is important to notice that although some attacks (e.g. flooding attacks) are technically simple, it certainly does not mean they are ineffective. Sometimes it is just the opposite – they are more dangerous than the complex ones. It is therefore important to develop and implement effective ways of dealing with them.

Protecting against DDoS attacks is not an easy task. While a small scale attack can be mitigated by using software solutions such as precisely selected well adopted iptables rules or even dedicated hardware solutions (e.g. firewall with 10Gb/s network interfaces) the problem becomes much more complex with the increase in attack volume. At first one needs to establish procedures specifying actions to be taken in case of detection of a DDoS attack. Only then technical solutions can be meaningful and effectively applied. Also, in case of attack one should involve the most powerful network devices in the defence actions in order to protect the internal infrastructure. The main target is to stop the attack at the network perimeter without wasting the internal network's resources. If it is impossible to stop it one should try to involve geographically distributed redundancy. HPCs could benefit from the fact that they are interconnected and geographically separated which can help to successfully defend their resources.

1.5. Honeypots

Knowing only that the attack happened often is not enough, one wants to know how exactly it proceeded. The technology which helps to achieve this is a honeypot. It is a trap set to detect, deflect, or in some manner counteract attempts of unauthorized use of the system. It also allows gathering information about it for further analysis. Generally it consists of a host that appears to be part of a network, but is actually isolated and monitored. To attract the attackers it should also seem to contain valuable information or resources.

1.6. Data Loss Prevention / Data Leakage Prevention software

Information in an HPC centre is of great value (e.g. scientific results). Solutions for controlling the traffic looking for information leaks are therefore used more and more often. The mechanism performing such information leakage checks is a DLP (Data Loss Prevention / Data Leakage Prevention) system.

DLP is the common name for a mechanism designed to control data transfers from a protected system to the external (public) one. It is especially dedicated to detect and, in some cases, prevent potential data leakage. There are various types of DLP systems. Some of them work on the server level and others on the network level. Network level DLPs are based on analysis of the data traffic: all data transferred through the DLP system are compared to the data patterns recorded in the DLP system to discover the sensitive data inside the transferred message. There are of course many potential actions to be taken by the DLP system, depending on the DLP design and its configuration. In the least restrictive configuration DLP can only record (e.g. to a log file) the fact of a sensitive data transfer attempt, in the most restrictive configuration the DLP can block the data transfer. Some other potential actions may be taken like notification to the data administrator about the sensitive data transfer, request of confirmation of the data transfer from the user who initiated the transfer or from the data administrator.

One of the most important steps of setting up a DLP system is a process of defining sensitive data patterns. Some patterns can be very complicated and can depend on the specifics of the organization and its database content. Therefore it is very difficult to define a good pattern set. It should be defined very carefully, and well tested. Otherwise, a DLP system can produce many false positives and/or false negatives. The other important step is the placement of the DLP system in the network structure. It should be placed at a point where a risk of sensitive data transfer from the inside of the organization to the public network exists. Usually it is an internal network edge router or firewall. There are, of course, some techniques to cheat DLP systems by an experienced, malicious user. Most of them are based on cryptographic techniques, but on the other hand there are still some methods on the organizational or technical level to neutralize them. The DLP technology, however, should be currently considered as still relatively immature and therefore should not be treated as the main pillar of the data security but rather as an auxiliary solution. Some hardware vendors equip firewalls or security appliances with DLP modules that may be optionally activated.

1.7. Network segmentations – Demilitarized Zone, Virtual LANs

Proper network segmentation must not be omitted. It is recommended to place public services in a separate network segment called Demilitarized Zone (DMZ), without access to the internal LAN. It can be achieved by either physical separation or using VLANs (Virtual LANs), which enable the administrators to divide the physical network into logical sub-networks. Except the easier network segmentation VLANs help, for example, to separate the user traffic from the administration traffic, which obviously increases the overall security level. How the final configuration looks like it depends on the detailed requirements of the data centre's services.

1.8. Authentication

Authentication based on a username and a static password is the most popular form of authenticating users. It is not, however, the most secure one. There are plenty of other, safer methods. The main idea of increasing authentication security is usually two-factor authentication. In that approach having a password, "something you know", is not enough for a user to be authenticated. A second factor is required. Mostly it is the possession factor, "something you have", like a private and public key.

One of the relatively recent technologies that have been getting more attention lately is a one-time password method. OTP is a password that is valid for only one login session, which helps avoid some of the shortcoming of static passwords such as vulnerability to replay attacks, i.e. even if the attacker obtains the password, it is not valid and cannot be used anymore. It requires a special device to generate passwords, but the associated cost is not high.

A better known technique is a use of asymmetric cryptography. A user needs a public and a private key for a successful authentication. The public key is placed on the server which authenticates the user before his/her first authentication. Then the user proves his or her identity by being able to decipher the challenge encrypted with his or her public key. Possession of the private key and knowledge of the passphrase used for securing it is required for that.

1.9. Incident response procedure

What is considered an "incident" in the first place? What happens when a security incident is discovered? What is done when the attacker invokes an attack? Who gets called and when? It is useful to test the procedure with a sort of incident-response procedure drill. People to consider calling may include officers of the company, the marketing manager (for press relations), system and network administrative staff, and the police. When you call them, and in what order, must be part of the procedure. Calling too many people too soon risks letting the cat out of the bag, so to speak, or a crying wolf scenario. Calling too few people, too late, risks lawsuits.

Although this process does not require any particular technical expertise, it does require a lot of thoughts. Senior managers should carefully take into the consideration an incident response procedure after receiving a briefing based on the vulnerability assessment.

2. Current practices in European HPC centres

There are many security-related technologies used in HPC centres. Those of the greatest importance have been presented in the previous chapter. The current state of implementation of the security means has been assessed basing on an electronic survey ran within PRACE consortium. All of the HPCs involved in the PRACE project received a list of questions about their security infrastructure and practices, and were asked to assess certain security aspects. The authors received 16reponses from European HPC sites and based this document on them.

The details of security needs and requirements differ between centres. On the other hand, certain part of HPC centres activity, especially if they have the similar profile like providing services for the scientific community, will also undergo similar threats that can be addressed in a specific manner - optimally in the best recognized and recommended, coherent way in all centres. An HPC centre has to securely store its data. The data must be appropriately protected when being transferred to and from the centre. Suitable authorization methods in data access are necessary.

The survey scope has been focused on the common profile of an HPC centres activity. The detailed results are not relevant in a public white paper since considerations specific to each site may lead to different implementations of security measures.. However, the survey results were used for preparing the general recommendations hereafter.

3. Recommendations

This chapter presents a set of recommendations based on the experience of PRACE partners, analysis of the current state of the art and good security practices. We have to say here that the real level of security depends on the site requirements.

Recommendation 1: Perform security audits periodically

- Create or outsource a security department or team dealing with security issues.
- HPC infrastructure should undergo a security test periodically, for example every 6-12 months.
- Perform a security test of every new device or system before introducing it to the HPC infrastructure.
- Perform a security test whenever a security breach has been detected.
- If a new attack technique or a critical software vulnerability have been found, perform a security audit in the involved area.
- Security tests should cover various scenarios:
    - Penetration tests from the Internet (white box and black box),
    - Penetration tests from the internal network (white box and black box),
    - Configuration reviews,
    - Application code review (if applications are developed),
- All security tests should be carried out by IT security professionals who are not directly involved in administering the infrastructure or system being audited.

Recommendation 2: Perform formal audits of the organization

- Create or outsource an auditing department capable of performing formal (non-technical) security audits.
- Create and introduce an Information Security Management System (ISMS) based on a known standard/norm (e.g. ISO27001), covering the whole organization.
- Periodically, e.g. once a year, perform formal audits checking consistency with ISMS.

Recommendation 3: Network security

- Ensure a proper network segmentation by introducing DMZ and VLANs
- Introduce stateful network firewalls as the main network protection means.
- For better and deeper traffic inspection, an introduction of application firewalls is advised.
- Firewalls should work in the High Availability mode.
- Consider introducing a DLP system.
- Attack detection, mitigation and analysis:
    - Introduce network IDS or IPS,
    - Create honeypots for better understanding of attacks against the network,
    - Implement and introduce an anti-DDoS system.

Recommendation 4: Host security

- Introduce a local, host-based firewall on every system working in the infrastructure.
- Install antivirus software and keep it updated.
- If there is no network solution, it is recommended to install DLP software.
- If there is no network solution, it is recommended to install host-based IPS/IDS.
- Allow remote management connections only to low privilege accounts.

Authentication should not rely only on the username and password. Introduce a higher level of security by using 2-factor authentication with, for example, X.509 certificates or one-time passwords.

4. Conclusions

Although all of the surveyed institutions were High Performance Computing Centres, actively involved in research activities, they differ in many aspects, mainly due to the kind of projects they work on. This is the first reason why the number and quality of security means introduced to their infrastructures differ, the second reason being local considerations and constraints. However, it is always of interest to study whether additional measures can be implemented.

One of the major differences among sites is the existence of a separate security department or team, which is highly recommended if possible in the context of the site. A person whose daily duties include administering the system or network infrastructure is never the one who should audit those systems or infrastructure. They simply cannot do it well, even if they possess extensive knowledge about security. The situation resembles book editing by the author– one does not see his or her own mistakes.

Another important point is that security is a process, not a product. Security audits have to be performed regularly by, again, people not directly involved in maintaining of the infrastructure. Not all sites do perform configuration review and penetration tests on a regular basis; improving this situation is highly desirable.

Firewalls are necessary to the HPC centre's infrastructure, as the main means of network security. Technologies such as application firewalls, both network- and host-based IPS and IDS systems along with DLP software and honeypots, are not ubiquitous. Implementation of application firewalls and IPS/IDS systems is particularly advised.

Another aspect of security that needs attention is the antivirus software both on desktops and servers.

Distributed Denial of Service attacks evolve in time. New types of DDoS attacks are invented and the average volume of such attacks continually grows. The volume of DDoS attacks is expected to continue increasing significantly in the following years. The centres are highly encouraged to acknowledge this fact.

Last but not least, a security area that is also worth considering is the Information Security Policy that should be formalised and updated on a regular basis, so the non-technical aspects of maintaining a decent level of security. One might say it is the Achilles heel of most companies or organisations, including the surveyed HPC centres. The Information Security Policy, if it exists, is based on internal procedures rather than known standards or norms such as, for example, ISO27001.

As can be seen, High-Performance Computing Centres present different levels of security. Some of them pay more attention to it while others seem to be more focused on performance. Although security always comes with costs, both material and in the use of resources, it is highly advised to maintain at least the minimum level because performance of a compromised system may tend to zero.

Acknowledgements

This work was financially supported by the PRACE-2IP project, funded in part by the EU 7th Framework Programme (FP7/2007-2013) under grant agreement no. FP7-283493.

Terms and abbreviations

BHO	Browser Helper Object
DC	Data Centre
DDoS	Distributed Denial of Service
DLP	Data Leak/Loss Prevention
DMZ	Demilitarized Zone
DoS	Denial of Service
HA	High Availability
HPC	High-Performance Computing
IDS	Intrusion Detection System
IPS	Intrusion Prevention System
ISO/OSI	International Organization for Standardization/Open Systems Interconnection
LAN	Local Area Network
OTP	One-Time Password
SaaS	Software as a service
TCP	Transmission Control Protocol
SSH	Secure Shell
UDP	User Datagram Protocol
VLAN	Virtual LAN