Firewalling beyond 10Gbps

Firewall design, deployment and management for a very large site

Authors

Edoardo Martelli: CERN, IT department. Email: edoardo.martelli@cern.ch
Nils Høimyr: CERN, IT department. Email: Nils.Hoimyr@cern.ch
David Gutierrez Rueda:CERN, IT department,. Email: David.Gutierrez@cern.ch
Denise Heagerty: CERN, IT department. Email: Denise.Heagerty@cern.ch
Jean-Michel Jouanigot: CERN, IT department. Email: Jean-Michel.Jouanigot@cern.ch
Lionel Cons: CERN, IT department. Email: Lionel.Cons@cern.ch
Nicholas Garfield: CERN, IT department. Email: Nicholas.Garfield@cern.ch
Nuno Ricardo Cervaens Costa: CERN, IT department. Email: Nuno.Cervaens@cern.ch
Zbigniew Stanecki: CERN, IT department. Email: Zbigniew.Stanecki@cern.ch
CERN address: 1211 Geneva 23, Switzerland

Keywords

Firewall, network, security, management framework

Abstract

CERN is building the LHC, one of the most complex scientific instruments ever built. The whole system relies on the LHC Computing Grid (LCG) to analyze all the data that will be produced by the experiments. The LCG computing model requires high speed connectivity between CERN and the remote institutes that will help with analyzing the produced data. Thus, CERN had to re-design its whole network infrastructure, from the detector pits to the connections to remote sites scattered around the world.
An aggregate WAN bandwidth that already exceed 20Gbps, dozens of servers accessible world wide, data to be pushed to remote sites at the highest possible throughput, hundreds of freshly deployed applications, these are factors that have posed new challenges to the computer security at CERN and that require a powerful but still flexible firewall infrastructure.
This paper describes how this security challenge has been tackled and how the designed solution was deployed.

Terms definitions

In this paper the term "firewall" refers to a generic device (or set of devices) that can perform different actions on the flowing through traffic, from basic stateless packet filtering to deep packet inspection.

The CERN network for the LHC

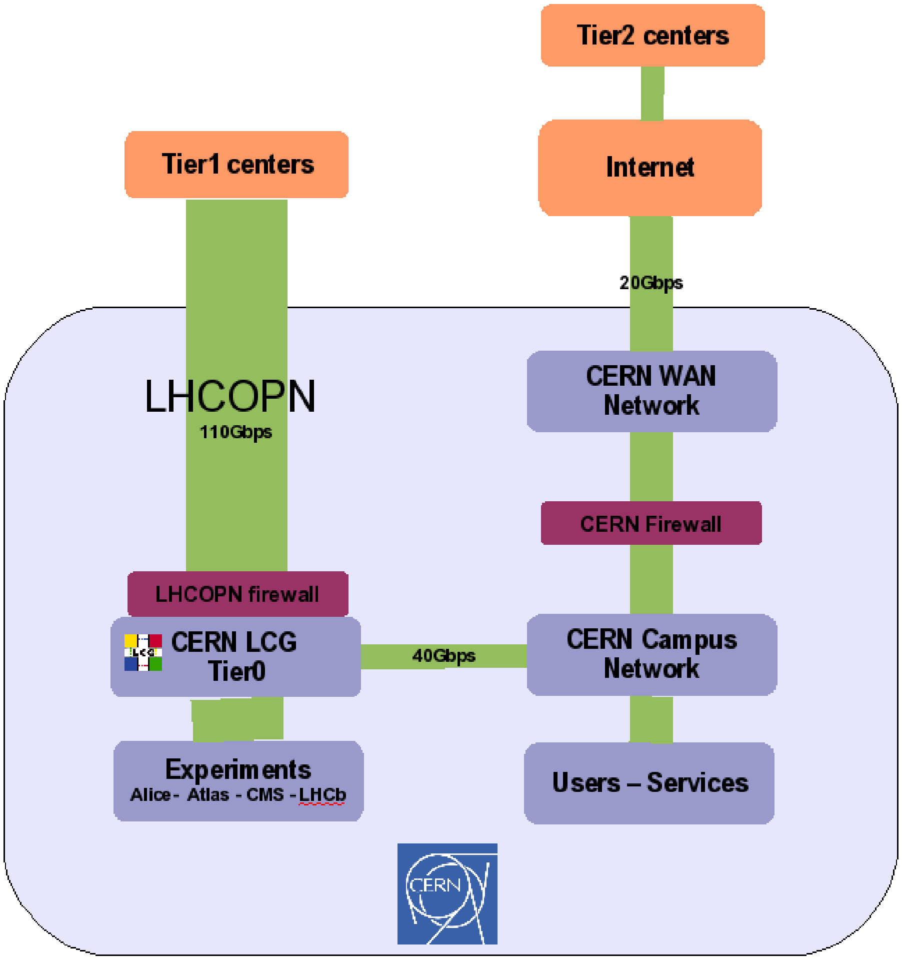
CERN is the centre of the LCG and the source of all the data coming from the LHC experiments. The LCG is connected to the eleven Tier1 computer centres by mean of the LHCOPN. Every Tier1 is required to be connected to CERN with a 10Gbps link, due to the high volume of traffic expected. Still at the time of writing, the market cannot offer any stateful firewall with such capabilities, therefore CERN decided to connect the LHCOPN links directly to the so called "LCG Backbone" [link??], and to protect its infrastructure using stateless ACLs on the boundary routers. Thus, the firewall described in this paper is not filtering the Tier0-Tier1 traffic.

However, due to the routing architecture of CERN, the general purpose Internet is used to provide last resort backup connectivity to the Tier1s, in case their LHCOPN links are down. Furthermore, CERN is both Tier0 and Tier1, so it has to transfer data also to the Tier2 centres, and this happens via the general purpose Internet.

Furthermore, as soon as the LHC will start, hundreds of scientists are expected to be hosted daily at CERN, everyone with his/her own network connectivity needs.

These were among the main reasons that pushed the upgrade of the main CERN firewall.

The CERN network and its connection to the Internet and to the Tier1/2 centres is depicted in this picture:



Security challenges

The CERN computer infrastructure serves the scientific community that builds the LHC accelerator and detectors and will use the LHC for their physics experiments. To better accomplish this task, access to data, applications, documents from outside to CERN and must be simple and straightforward for everyone. Thus the CERN network is built in a way that every computer connected to the campus might be accessible from anywhere.

At the same time, security threats are increasing both in quantity and devastating effects. The common practice to reduce the risk is to hide a network as much as possible, but this clashes with the goal of providing easy access.
CERN had to develop an ad-hoc strategy to conciliate these two requirements: the accessibility is obtained using public IP addresses for every machine connected; security is provided using a fine grained filtering on the firewall, allowing connectivity from outside only to the strictly necessary ports in every single host. This means a big effort in collecting and deploying all the users' requests, thus the need of automate this process as much as possible.

The size of the community using the CERN network posed another issue: the great amount of required bandwidth to the Internet. The generic Internet is by far the major source of attacks, so most of the security checks have to be done one the traffic coming from there. For this purpose, CERN has been having a stateful firewall that interconnects the campus network with the Internet, but the traffic load has always been greater than the capacity of any commercial product could provide. CERN's strategy to solve this problem is to offload some well know and well defined traffic from the stateful firewall, with the double advantage to reduce the load on it and also to avoid to introduce any sort of throttling to data transfer that requires high bandwidth. In order to have this bypass granted, users must define their traffic in term of source and destination IP addresses and application ports.

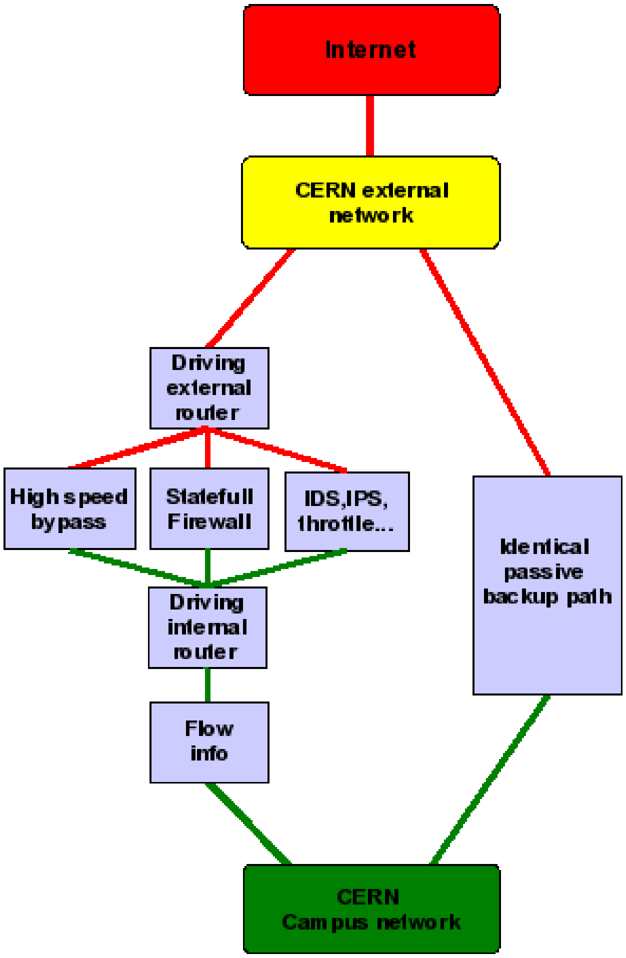
CERN Firewall upgrade

With the major challenges posed by the LCG and future LHC operation, the CERN IT Department launched a project early 2006 for a major upgrade of the CERN main firewall. A team of network engineers, software developers and security experts were dedicated to the design and deployment of a new system that could tackle the challenges posed by the LHC.

Requirements for the hardware

The requirements for the new CERN main firewall were:
- Redundancy: two symmetrical paths must exist, one active and one passive, both with the same capabilities and bandwidth capacity.
- Stateful firewalling: generic traffic must go through a stateful firewall for deep inspection. It must handle at least 2Gbps full-duplex without service degradation.
- High speed traffic offload: it must be possible to offload very well defined high speed data flows from the state-full inspection path, using policy based routing. The new system must provide at least 40Gbps.
- Bandwidth throttle: it must be possible to throttle the bandwidth used by potentially harmful and policy defined traffic.
- Possibility to add additional alternative paths and traffic mirroring capabilities, in order to send part or all of the traffic to other screening devices, like IDS and IPS.
- Flows information: the system must provide information about all the traffic flows.
- Scalability: the system must scale in order to handle the traffic growth.
- Modularity: single components should be easily upgraded as they become obsolete.
- Manageability: it must be possible to configure the system with an automatic software framework

The desired architecture is depicted in this diagram:

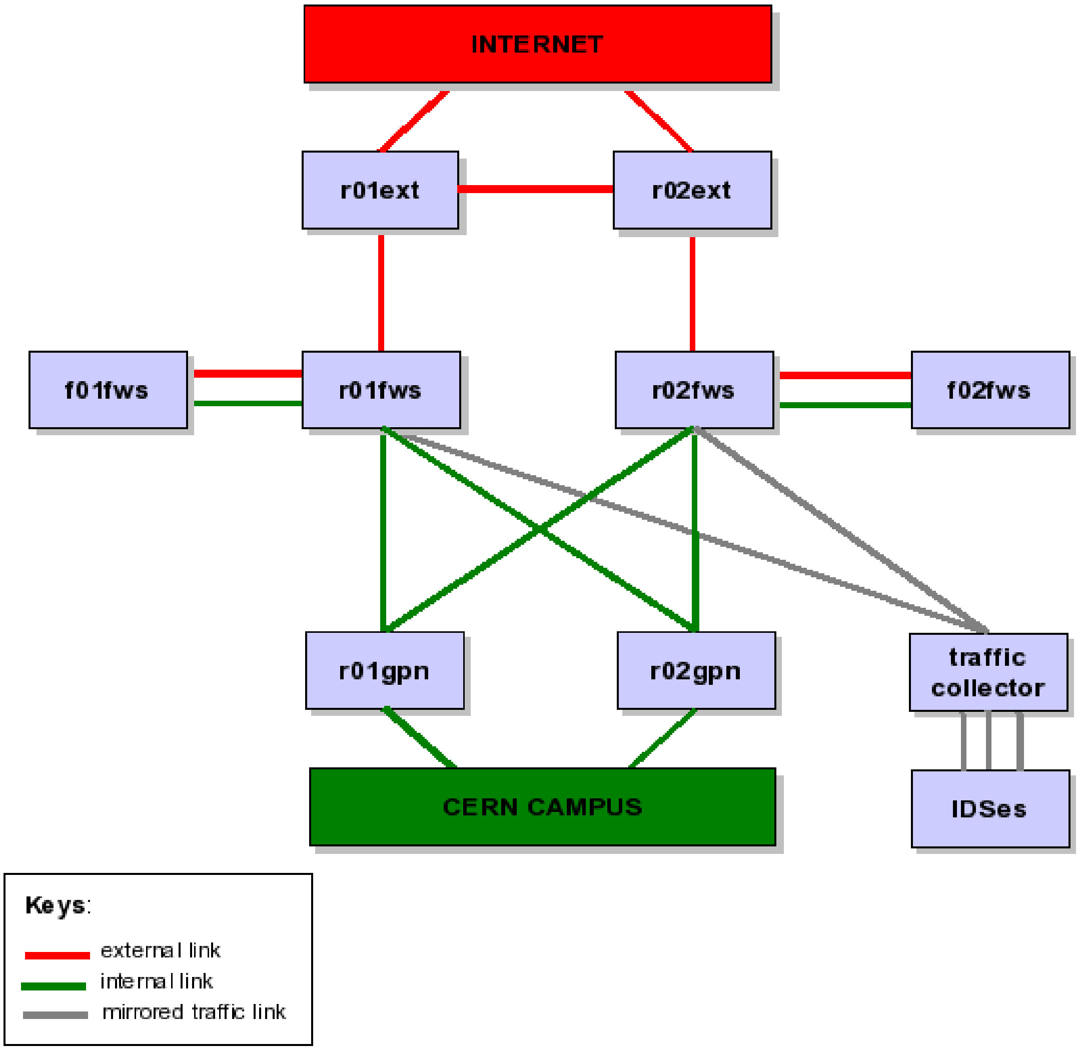


Requirements for the management framework

- allow the complete configuration of the system
- re-use of all the information regarding hosts and IP address assignments already stored in the CERN's Network Database
- vendor independent (able to configure any kind of equipment)
- architecture independent (to be re-used with any firewall)
- provide a web interface to the end-users to request firewall openings

CERN firewall architecture

The architecture designed and the equipment selected met all the requirement. The following picture shows what has been implemented:

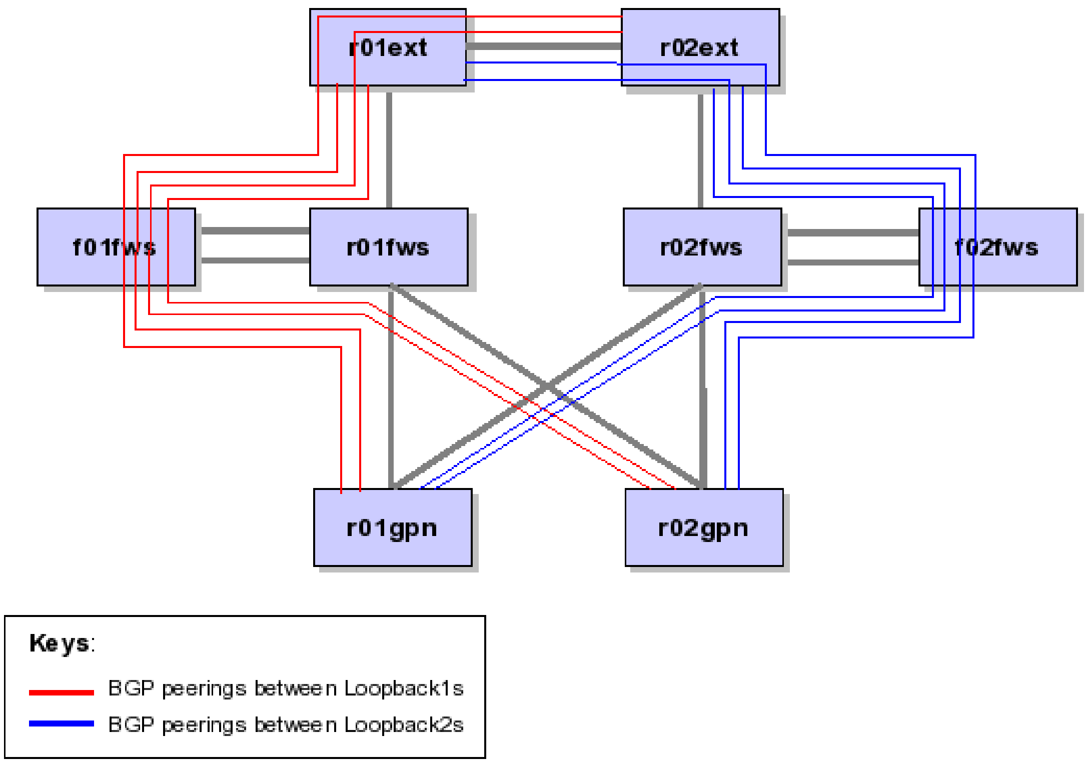


Routing through the Firewall

The Main CERN Firewall connects the Campus network to the External Network.
The CERN External Network provides the CERN community with the connectivity to the generic Internet. All WAN links, excepted the ones that belongs to the LHCOPN, are terminated to the External Network routers: Geant2-IP, Esnet, Abilene and three commercial Internet Service Providers guarantees reachability to any Internet destination.

Dynamic routing protocols
The External Network uses OSPF as IGP; this OSPF instance is completely independent from the one used in the campus network.
The external network provide the default route to the CERN campus network by means of iBGP peerings established through the primary and secondary gates. There is double full iBGP mesh among the two main campus backbone routers and the two main External Network routers. The primary mesh peerings are established using the loopback interfaces number 1 and all go via the primary gate; the secondary mesh peerings are established using the loopback interfaces number 2 and all go via the secondary firewall. The external routers assign a MED of 100 (preferred) to the prefixes received and announced into the primary mesh; they assign a MED of 50 to the prefixes received and announced into the secondary mesh.
In case of a failure of one of the two gates, all the peerings of a mesh will go down, and so the survived ones will be preferred. This mechanism is necessary because it was decided to configure the secondary gate in hot-standby and not load-share, so to have the two firewalls completely independent one from

the other.


r01ext
r02ext
f01fws
r01fws
r02fws
f02fws
r01gpn
r02gpn
Keys:
BGP peerings between Loopback1s
BGP peerings between Loopback2s

R01EXT and R02EXT also peers, and they are Route Reflectors in a cluster. R01GPN and R02GPN are route reflector clients, thus they don;t need to peer one with the other.

Static routes
Static routes redistributed into OSPF are used to force each mesh to go via two different paths. R01EXT has static routes to the Loopback1s of R01GPN and R02GPN pointing to R01FWS, and it redistribute them into the external OSPF instance. R02EXT has statics to Loopback2 of R01GPN and R02GPN pointing to R02FWS, and it redistribute them into the external OSPF instance. R01FWS has static routes to loopback1 of R01EXT and R02EXT pointing to R01EXT, and it redistributes them into the internal OSPF instance. R02FWS has static routes to loopback2 of R01EXT and R02EXT pointing to R02EXT, and it redistributes them into the internal OSPF instance. Thh redistribution is used to ensure that alternative paths will be inside the two networks in case of link failures.

Default route
In the campus network, the two main routers generate a default route for all the campus and distribute it with OSPF. The default route learnt by BGP is fact not redistributed, but used locally to send the traffic to the active gate. The path is chosen using the different MED value set by the External routers.

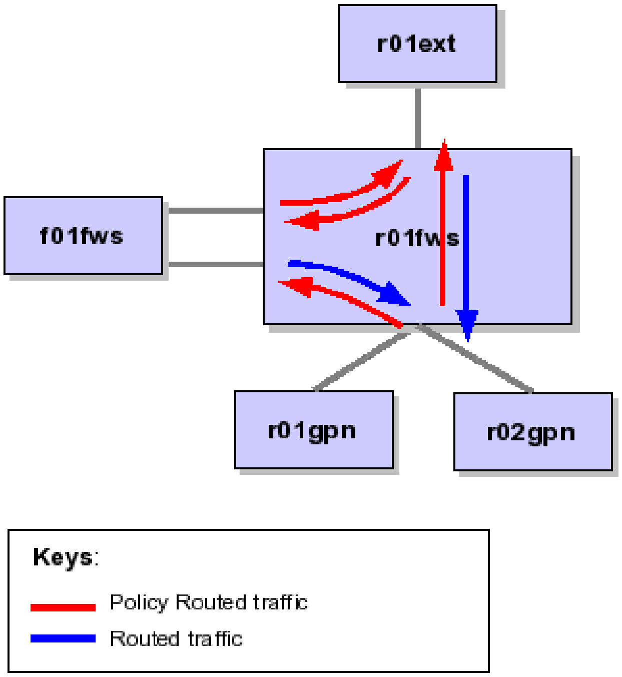
Internal prefixes.
The two main Campus Network routers announce five public CERN prefixes to the external network via iBGP. These prefixes are necessary to the External network to decide which firewall path to use. Also, they are used to announce them into the eBGP peerings.

Gate routing
The two firewalls, primary and secondary, behave in the same way, so only one will be described here. Due to the double connection to F01FWS (the Stateful Firewall), the packets in R01FWS have to be routed accordingly to their source and destination addresses, and not only considering the destination as usually happen. This technique of routing packets not accordingly to the routing table is called Policy Based Routing (PBR) and is realized using ACLs.
R01EXT route the packets in this way:
- all the packets coming from R01GPN and R02GPN are policy routed to F01FWS, except the packets that has to go via the fast path that are policy routed to R01EXT.

- the packets coming from the external interface of F01FWS are policy routed to R01EXT

- the packets coming from the internal interface of F01FWS are routed accordingly to the OSPF routing table and sent to R01GPN or R02GPN.
- the packets received from R01EXT are policy routed to F01FWS, except the one the path that has to go via the fast path that are routed accordingly to the OSPF routing table and sent to R01GPN or R02GPN.

R01FWS doesn't accept the default route it receives in OSPF and has a static default route pointing to NULL. There are two reason for this: the first reason is that there must be no default route to the external network to avoid that any packet might be sent out if not coming from the firewall or decided by the HTAR policy. The second reason is to avoid to send packets coming from the external network and directed to CERN subnets not in use (the OSPF routing table contains in fact only the sub-nets in use).



Management Framework

A research institute like CERN has a diverse user community running many applications using different protocols. The environment is highly dynamic, notably Grid and physics application software require the management of higher of port ranges that are rarely used on other sites. Thanks to the database driven network management and dynamic work-flow system, the CERN Computer Security team is now able to quickly open specific ports for dedicated hosts upon user requests.

Transforming the generic ACLs into instructions for routers and firewall modules from multiple vendors has been another major challenge. This is achieved by means of modular software components, where the ACL-generation from firewall rules is done by an ACL-generation module and then feed to a router-compiler with vendor-specific modules. The system has been deployed step-wise onto different types of network equipment that is used for gateways and firewalls within the CERN corporate network as well as the external firewall.

Following the deployment of the high speed firewall, IDS and packet inspection also require far more computing resources.

A suite of web-based management applications for the definition of the firewall filters, gates and rules has been implemented. In addition, software has been developed that extracts the rules from the database and translates them in the configuration commands for network equipment of different vendors. The main software components of the framework deployed are:

- a *Gate Model* with a *Database Schema* that implements all the components needed to represent the firewall
- a web based *Management Framework* that allows for the definition and the correlation of all the components of a gate as well as the rules to be applied in the various interfaces of a gate.
- a web based work-flow application that allows for *End-user requests* for changes to the firewall configuration, and to easily and securely implement them.
- a *Configuration Manager* that understands all the information gathered in the database and builds the configuration for all the network devices that compose the gate.

Gate Model

The combination of a highly dynamic environment and granular security policies have meant that changes in the firewall configuration are requested on a regular daily basis. The core of the CERN network is a database, which contains a complete description of the network topology as well as all devices connected to the network. The database provides a framework for network configuration management, ranging from end-user requests to connect new devices to management of routers and switches. A flexible generic network interconnection model has been designed and implemented within the network database to manage the interconnection of different network domains and sites. The model can be applied to external firewalls as well as internal firewalls and screening routers within a network site. The model supports filtering of port and IP-based access control, as well as macros that create ad-hoc rules for specific attacks and special configurations for high throughput computing.
The main components of the model are:
- Domain = A representation of a well defined network area.
- Gate = a system that interconnects two Domains by mean of Interfaces and capable of filtering the traffic exchanged.
- Interface filter = relationship between a Filter and the Interface to which is applied.
- Filter = a set of Rules applied to one or more Interface Filters of a Gate.
- Rule = an unidirectional communication relationship between two IP prefixes belonging to two Domains connected by a Gate .
- Interface = An IP interface belonging to a network equipment.

Domain

It represents a well defined network domain, i.e. a set of network devices and IP network prefixes. IN the built system, the domain is just a name to help in the definition of the gate.

Gate

It is the set of devices that interconnect two domains and that filter the network traffic they exchange. In order to simplify the creation of the configuration for the devices, the two domains are identified in Left Domain and Right Domain
The gate provides the list of Interface Filters used in the gate.

Interface filter

It is the definition of an action that a device has to do on the packet traversing an interface. The action can be:
Input Filter, i.e. an ACL applied to the packets entering the interface,
Output Filter, i.e. an ACL applied to the packets leaving the interface
PBR, i.e. a routing policy applied to the packet entering the interface
The Interface filter links to the interface where it has to be applied and to a set of Filters that build the ACL. In order to allow the re-utilization of Filters in different directions, the Interface Filters include the information of which domain is faced by the interface (Left or Right)

Filter
It is a set of Rules.

Rule
it is a single Access List Entry, i.e. a set of characteristics that can match an IP packet. The characteristics considered are: Layer3 or Layer4 protocol, source and destination IP addresses, source and destination Layer4 protocol port.

Database schema

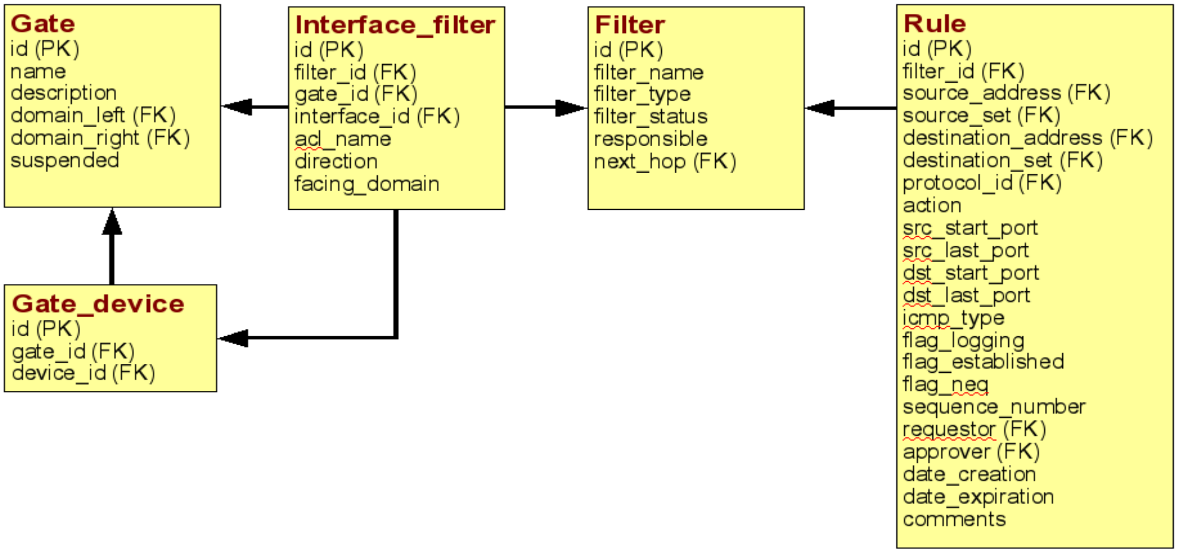
The model had to be represented with a database schema. This schema was designed with two main goals in mind: to use as much information as possible about networks and hosts already stored in the CERN network database; and also to provide a schema capable of representing not only the planned CERN main firewall structure, but any firewall in general.

The first goal was achieved by using database ids for referencing firewalled CERN entities: networks, services and hosts. This way, when a machine with some firewall privileges changes its IP address, nothing changes in terms of accessing it through the firewall.

The second goal was achieved with the definition of the Filter concept. A Filter, logically, is a set of rules that can be applied to any interface of any machine that is a part of any firewall. For greater flexibility filter can consist not only of a set of rules, but can call database or perl procedures which, in turn, will dynamically generate the filter content each time the compiler runs.

Firewall privileges for large number of machines that require the same type of access through the firewall (like the large clusters used for the LCG, or the pool of servers providing public services) can be easily managed with the use of Sets, a database table that can group a list of machines. Firewall rules can be defined using Sets as destination address.

The model also keeps track of the changes made to the rules during a Gate lifetime, by storing historical information in dedicated tables.



Management interface

The framework provides a web interface for network and security experts to manage the firewalls. The management interface was developed extending an already existing platform in use at CERN for the management of the network infrastructure.
The interface allows the definition of all the components needed for a gate configuration.

Configuration manager

The framework provides a tool called “cfmgr-gate” that extract from the Network database all the information related to any gate and use them to build the final configuration of all the devices involved. cfmgr-gate is part of the software tool already used at CERN to configure and manage the Network devices.
cfmgr-gate, before applying the computed configurations, checks that they can be supported by the destination devices. This check is especially needed to avoid to exhaust the memory resources of router and switches that have to host the ACLs. Router and switches, in order to apply access list at wire speed without overloading their CPUs, have to store the ACLs in a special memory used by the network processors. This memory, due to its characteristics of fast speed and access, is very expensive, so limited in quantity. If an ACL exceeds the available space, its process happen in the device's CPU, degrading in this way the overall performance.
Since the software framework uses a database with virtually unlimited resources that would allow the creation of ACLs of any size, special checks have to be performed in order to avoid the performance degradation.

End-users requests

The framework provides a web interface that allow end-users to request any kind of firewall opening infor the hosts they own or manage .
The web interface was developed extending an already existing platform provided to the CERN users to manage the network connectivity of their equipment.
Users who needs a firewall opening must fill the form provided by the framework. As a consequence a request is sent to the Computer Security Team for approval. The Computer Security team can accept or reject the request following the evaluation of the request and a secuirty scan of the involved equipment. The approval is made via anoter web interface that automatically creates the correct rule in the Network database and allow the configuration manager to apply it to the relevant firewall.

Implementation experience

The statefull firewall bypass has been implemented using a single switch-router with multiple 10Gbps ethernet interfaces and capable of policy based routing of traffic at wire-speed. The switch-router is also capable of pre-filtering the traffic directed to the statefull firewall by means of access control lists executed by the hardware. This is done in order to offload some tasks from the statefull firewall. Also, it can export flow information without negative impact on the performance.Several tests have been carried out in order to measure the number of ACL entries supported by every device. The results have been used to enable the configuration generator software to simulate resource depletion, thereby avoiding harmful configuration.

Failover and BGP

The redundancy has been implemented using two identical sets of equipment; one set is the active path, the other set is the hot-standby path. The failover mechanism is implemented using iBGP routing among the two main campus backbone routers and the two main external routers, as explained beforehand.

The reason of the double mesh is because iBGP peers need to be fully meshed to avoid routing inconsistency. The use of eBGP would have reduced the number of peerings, but it would have implied the use of a private ASN for the Campus network. Unfortunately, the stripping of the private ASN and the inclusion of the Campus prefixes in the public CERN's ASN would have required many configuration changes in the External Network routers, so it was abandoned.

Another option would have been the use of BGP confederations, but that's would have also required several changes in the running configurations of the External Network routers, so it was also abandoned.

TCAM exhaustion

Multilayer security versus high performance and availability has been an important point of discussion with our colleagues of the Computer Security team at CERN. Can we implement a policy in every device participating in the network so that a breach in one of the links won't be enough to put the whole chain at risk? Concerning this project, the demand from the Security Team was clear: apart from the stateful inspection and advanced filtering provided by the firewall, implement in the routers access control policies which were granular enough to filter based on application port numbers.

The Gate model defined above is wide enough to support application filtering on routers, but because of the abstraction provided by the user interface, the gate administrators will not be aware of the impact of the filters on the real hardware, in fact, they probably won't know what hardware is implementing the gate.

As responsible for the firewall system our concern is to guarantee that only configurations that fit in the hardware resources will be configured on the routers. This sounds straightforward, but unfortunately router manufacturers don't provide information on how the different elements in an access list entry impact the resources available. For example, an access list entry containing layer 4 operators will consume more resources than another without them, and even the resources consumed vary depending on the combination of these operators. To accomplish this objective several scripts were developed to populate access lists with all possible combinations of operators and to read the resources consumed. The data extracted was used to build the equations that matched the behavior in the resource consumption for every particular case, and these equations were implemented in algorithms for every different router model.

Glossary

ACE = ACL Entry
ACL = Access Control List
ASN = Autonomous System Number
BGP = Border Gateway Protocol, inter domain routing protocol
CPU = Central Processing Unit
HTAR = High Throughput Application Route
Layer 4 operator = Operator used in an access list to select application ports, like http or ntp. Common layer 4 operators are equal, greater/lower than, range and not equal
LCG = LHC Computer Grid
LHC = Large Hadron Collider
LHCOPN = LHC Optical Private Network
OSPF = Open Shortest Path First, intra domain routing protocol
MED = Multi Exit Discriminator, BGP parameter used in the route calculation algorithm
WAN = Wide Area Network

Reference

Integrated Site Security for Grids (ISSeG) EU FP6 Project no: 026745 http://cern.ch/isseg/

Worldwide LHC Computing Grid (LCG), http://cern.ch/lcg

Acknowledgments

The security related aspects of this work have been co-funded by the European Commission ISSeG project, http://www.isseg.eu/

Author Biographies

Edoardo Martelli works at CERN as network expert in the Communication Systems group. He received an MSc degree in Computer Science from University of Bologna, Italy in 1994. He started is career in Italy as network engineer for Cineca and then for Nextra. He joined CERN for the DataTAG project in 2002.
Zbigniew Stanecki works at CERN as software developer in the Communication Systems group. He received an MSc degree in Physics from University of Science and Technology, Cracow, Poland in 2001. He started his career in Poland as software engineer for Matrix.pl (www.matrix.pl). He joined CERN for the ISSeG project in 2006.
Nuno Cervaens works at CERN as a Network/Software Engineer in the Communication Systems group. He received a 5 year Degree in Electronics and Telecommunications Engineering from Oporto University, Portugal in 2002. He started his career at Novis Telecom (an ISP) and then joined CERN in Feb 2003.
Nils Høimyr got his MSc in Engineering at the Norwegian Institute of Technology, Trondheim in 1991 and has pursued a career at CERN on engineering projects and later information systems. He joined the Communication Systems group in 2005 for the implementation of network and process control security.
David Gutiérrez works at CERN as network expert in the Communications Systems group. He received the degrees of Engineer and Technical Engineer in Computer Science in 2003 and 1997 respectively. He started his career in 1997 as a Network Engineer at the University Carlos III in Madrid and joined CERN in 2005.