GÉANT
Networks • Services • People

Providing high network availability at data centres

Best Practice Document

Produced by the CESNET-led IPv6 working group

Authors: Martin Pustka (CESNET)

March 2016

Document No: GN4P1-NA3-T2-C2.1
Version / date: V1
Original language : Czech
Original title: Zajištění vysoké dostupnosti sítě v datových centrech.
Original version / date: 28.3.2016
Contact: Martin Pustka

The work has been carried out by a CESNET led working group on Infrastructure as part of a joint-venture project within the HE sector in Czech Republic.

The research leading to these results has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement No. 691567 (GN4-1).

Best Practice Document:
Providing high network availability at data centres

Table of Contents

1 Introduction 3

2 Basic terms 4

3 Data centre layered network infrastructure 5

4 High availability 7
- 4.1 Other characteristics of high availability 7

5 Distributed data centre 9

6 L2 design 12
- 6.1 Converged networks 12
- 6.2 Virtual switches 12
- 6.3 Main L2 network points 13
- 6.4 Physical servers and L2 infrastructure 14
- 6.5 Tips for the L2 network 15

7 L3 design – linking network routers 16
- 7.1.1 Virtual multi-chassis solution 16
- 7.1.2 Standalone routers 16
- 7.2 Dynamic routing protocols 17
- 7.3 First-hop redundancy on L3 18
- 7.4 Configuration 20
  - 7.4.1 VRRP 20
  - 7.4.2 VRRP3 22
  - 7.4.3 HSRP 23
  - 7.4.4 GLBP 24

8 Problems and their solutions 27
- 8.1 Split brain 27
- 8.2 Problems with network asymmetries 27

Table of Figures

Fig. 1: DC layered architecture 5
Fig. 2:Linkage of two data centres 10
Fig. 3: Data centre from the L2 perspective 13
Fig. 4: Connection of a physical server. 14
Fig. 5: Routers and use of router protocols. 17
Fig. 6: Principle of firsta hop redundancy 19

1 Introduction

The massive virtualisation, technical development of data centres and the growing functions and demands on centralised IT services are accompanied by growth in the importance of technicians ensuring the stability and availability of network services provided.

The increased availability of services operated at data centres requires redundancy, best on each infrastructure layer of the data centre.

This document is focuses on and describes computer IP network design and technology in the area of data centres, which help ensure the increased availability of virtualised systems.

2 Basic terms

For these purposes we need to define some basic terms used in this document.

Virtualisation infrastructures are infrastructures creating an environment for virtualisation. Most commonly this involves the products VMWARE vSphere, KVM, Hyper-V and XEN.

Virtual system (virtual machine) – sometimes the term **virtualised system** is also used; this is usually represented by a virtual server, virtual station or virtual router operated within the virtualisation infrastructure.

Virtual infrastructures or, alternatively, **virtualised infrastructures** are infrastructures which are created by virtual machines, form a functional whole and are installed in virtualisation infrastructures.

Virtual routers are virtual systems fulfilling the function of routers in virtual infrastructures.

3 Data centre layered network infrastructure

Layered architecture defines infrastructure layers and their functions within the data centre. In terms of the computer network, we mean the following layers:

- routers
- L2 switches
- virtualised switches (in the case of virtualised infrastructures)
- servers

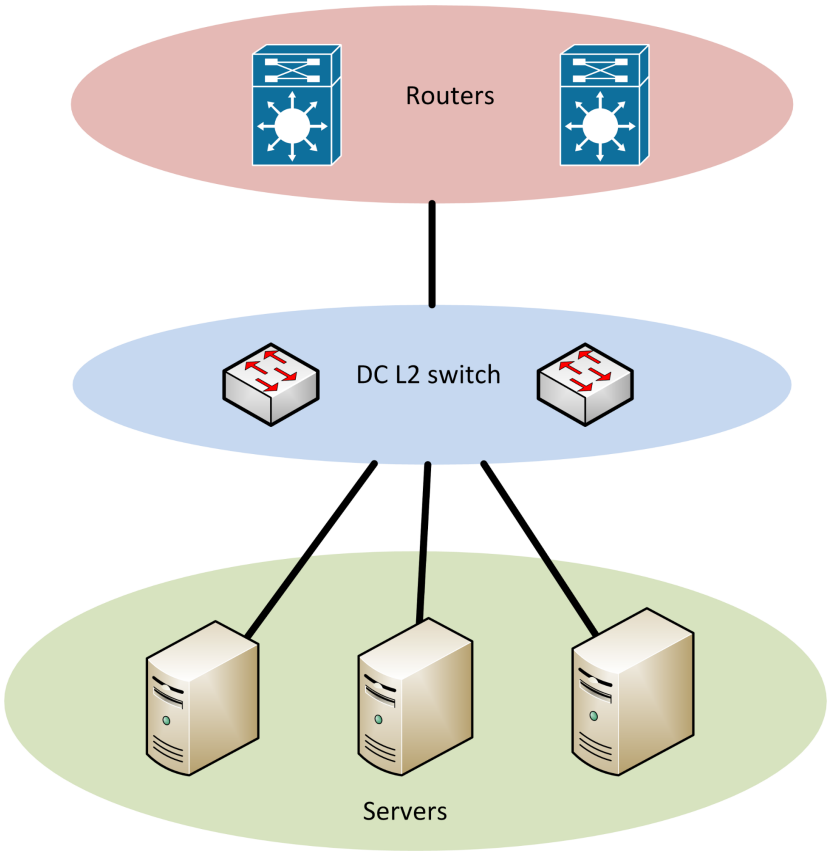

Fig. 1: DC layered architecture

Physical servers ensure the operation of operating systems or virtualisation solutions, in which individual virtual servers are operated. Servers are connected to the network infrastructure of data centre switches.

To increase stability the server may be connected through several data links to several elements, which may form one virtual data centre element.

Data centre switches provide server connections to the computer network. Switches can support modern converged network technology (LAN/FCoE/FC).

Routers ensure the routing of IP traffic. Routers of several IP networks may be linked to L2 network infrastructure, ensuring the routing of individual virtual L2 networks.

4 High availability

When designing infrastructure it is necessary to define what high availability means. Different environments with different requirements may have different definitions of availability. This means that the solution will be different in various cases as well.

High availability is a term that expresses the requirement for greater reliability of systems when used by users. High availability means the level or ability to operate IT services with minimum defined interruption.

When declaring values of high availability we must statistically evaluate and define the maximum times during which the service will be unavailable for a certain time period.

Whatever arbitrary values we arrive it, in the case of high availability, it is necessary to ensure service operation at the very least, even in the event of failure of one element. In the case of layered architecture, then each layer in the event of failure of one element.

This document is devoted to the provision of computer network high availability and therefore procedures, proposals and settings that ensure that network high availability are described here.

4.1 Other characteristics of high availability

In addition to the classical and already mentioned resistance and stability of operation in the event of failure of an element, this design also provides other advantages:

- operational stability
- scalability of technical facilities
- easy maintenance and updating
- easy HW/SW modification
- flexible network infrastructure

The chief added value is the possibility of relatively easy HW/SW modification, reconfiguration of parts of the whole conscious of the fact that operation is temporarily secured with a smaller or even no level of redundancy (in that case it depends again on requirements placed on systems operated).

The replacement of physical cabling, program upgrades, replacement or breakdowns of physical HW thus need not normally mean shutting down running services or services operated on those virtualisation infrastructures, because one of the great benefits of virtualisation infrastructures are also live migrations of virtual systems to other HW without their outage.

Since not all parts of the infrastructure need be actively used, it is necessary, from the perspective of the computer network, to consider quality monitoring of the situation and unused parts. It is good, for example, to monitor the states of the interfaces of passive backup routes, used only in the event of failure of the primary ones.

5 Distributed data centre

Increasing requirements for availability give rise to the need to ensure operation even in the event of failure of a locality, which may be at risk particularly from non-IT influences, for example, failure of electricity supply, flooding, fires or other.

In comparison to technologies used in the past, virtualisation relatively significantly reduces requirements for the quantity of HW and construction of distributed data centres is thus more attainable.

Distributed data centres are usually in several localities and are usually connected in such a way as to form one whole. In the event of failure of one locality, they are able to provide full or partial functionality of the systems and applications operated.

In terms of the computer network, we are usually dealing with the issue of interconnection on the 1st - 3rd layer of the OSI model. We need to consider the design of the storage network (SAN), Ethernet L2 networks and IPv4/IPv6 networks on the 3rd layer. This is an advantage if convergence technology networks are used in the network infrastructures. In such cases it is unnecessary to technologically separate the LAN and SAN networks and there are financial savings, but also easier configurations and integrations.

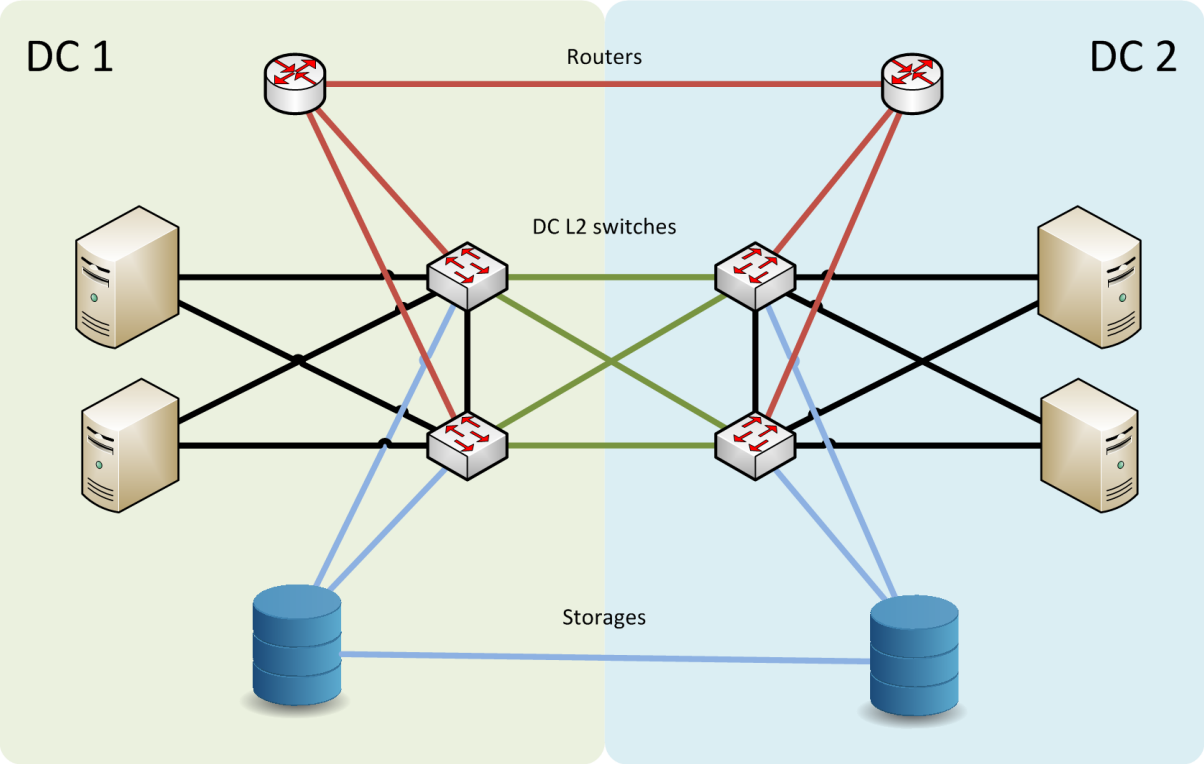

Fig. 2: Linkage of two data centres

You will notice that the foundation of the network and communication consists of several data centre switches, which all systems are connected to redundantly in every locality.

Dedicated interconnection between both routers and both disk stores is of special importance. These interconnections are not used in operations, they just increase redundancy and are there in particular to eliminate split-brain problems, i.e. in the event of collapse of the whole data centre network or for direct, internal communication between the data stores, if the solution used so requires.

When using converged networks it is possible to make use of the physical network infrastructure and its redundant integration for SAN networks as well. Other parallel SAN network infrastructure, which is represented in the existing infrastructure by dedicated virtual networks (VLAN), is not necessary to build. In addition, usually only relatively simple integration for direct internal communication of the disk store itself is implemented.

- scalability of technical facilities
- easy maintenance and updating
- easy HW/SW modification
- flexible network infrastructure

The chief added value is the possibility of relatively easy HW/SW modification, reconfiguration of parts of the whole conscious of the fact that operation is temporarily secured with a smaller or even no level of redundancy (in that case it depends again on requirements placed on systems operated).

The replacement of physical cabling, program upgrades, replacement or breakdowns of physical HW thus need not normally mean shutting down running services or services operated on those virtualisation infrastructures, because one of the great benefits of virtualisation infrastructures are also live migrations of virtual systems to other HW without their outage.

Since not all parts of the infrastructure need be actively used, it is necessary, from the perspective of the computer network, to consider quality monitoring of the situation and unused parts. It is good, for example, to monitor the states of the interfaces of passive backup routes, used only in the event of failure of the primary ones.

6 L2 design

Classical Ethernet is still predominant on the second layer of the OSI model, with all its advantages and disadvantages. If we are talking about redundant designs of L2 network topology, then classical Ethernet has quite a few disadvantages, stemming mainly from its historical characteristics, which result from the time period when it was used.

Many manufacturers are trying more or less successfully to eliminate the limitations of this protocol, whether using other protocols (e.g. TRILL, 801.1aq) or technologies adopting Ethernet in modern data centre networks (e.g. VPC).

In any case, it is advisable to have functional and redundant L2 infrastructure in the data centre that will minimise the number of passive links and also allow expansion and enhancement of network capacities during operation and without outage.

6.1 Converged networks

Although opinions on the use of converged network infrastructure differ, it is advisable (but not required) for data centre network infrastructure to support converged network protocols. An indisputable advantage is better scalability, better redundancy, but also financial savings, because one redundant network infrastructure is built and not two separate ones for LAN and SAN.

6.2 Virtual switches

The virtual switches of different manufacturers used today work on several basic principles. First of all, from the perspective of network infrastructure, this involves access switches with virtual ports connected to individual virtual systems.

In the case of simple separate switches, when migrating VM to other HW, classical updating of the VM MAC address on superordinate physical switches applies. This approach can cause minor disruptions during the time between VM migration and updating of the switching tables.

More modern distributed virtual switches can then be administered as conventional switches, where aggregate control logic and data exchange logic are separate. By this it can be relatively easy to implement virtualised infrastructure among tools for managing physical networks.

Other concepts then make network terminal VM interfaces accessible to superordinate physical switches (e.g. VM-FEX). In these cases very fast convergence is ensured, even when migrating VMs between individual VI nodes or between data centres.

6.3 Main L2 network points

As can be seen from the diagrams, centralised physical data centre switches will always be the main network nodes. Therefore, for rapid convergence it is advisable to set up a primary root bridge for each virtual network for the switch cluster of the primary data centre and a secondary root bridge for the switch cluster of the secondary data centre.

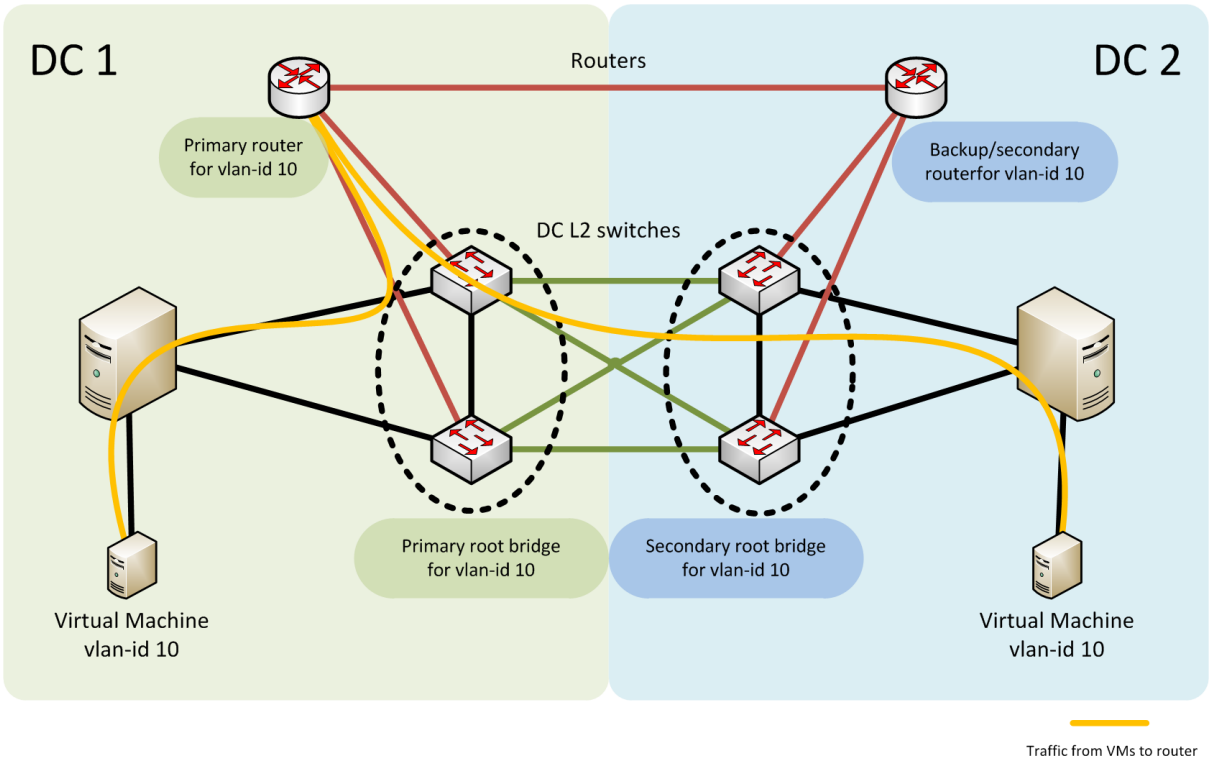

Fig. 3: Data centre from the L2 perspective

6.4 Physical servers and L2 infrastructure

In the case of physical non-virtualisation servers, it is advisable for those servers to connect to the network through two coupled network interfaces. This coupled (aggregated) connection has different names according to manufacturer, e.g. port-channel or NIC teaming. The IEEE LACP protocol is standardised for this type of connection.

This integration can solve connectivity redundancy and increases in capacity, because all aggregated interconnections are actively used. The following apply to this connection:

- links of the same capacity are aggregated
- the hash function is used for routing traffic along both links, whose parameters are most often IP or MAC addresses
- as a result, the usual communication between the two systems goes through one of the links (this can be a problem, for example, when backing up)
- the hash type function should be the same on both sides of the interconnection

Network access can be configured above the aggregated link as above a physical port. The link can be assigned for the application server to one defined VLAN network or a 802.1Q trunk can be configured above that link with multiple networks, which is a typical solution for virtualisation servers.

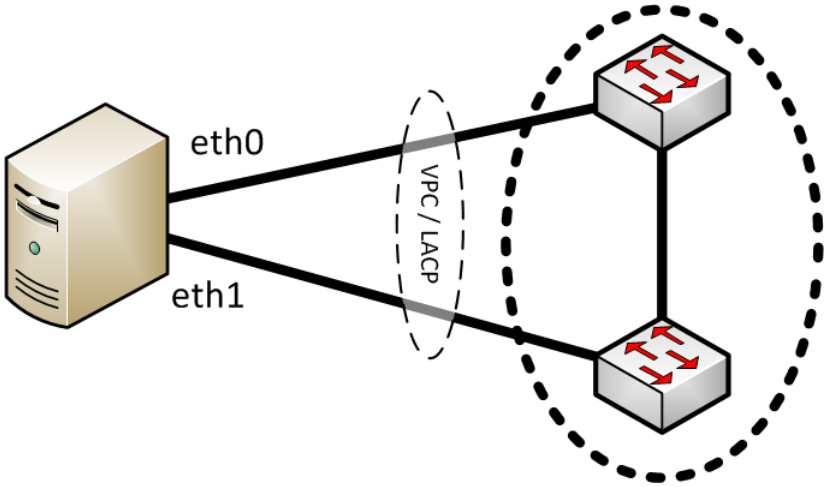

Fig. 4: Connection of a physical server.

6.5 Tips for the L2 network

Since redundantly operated physical data centre switches become the centre of the entire network, it is necessary to ensure rapid convergence even in cases of failure or changes in the network topology. Topology in the Ethernet networks is managed by the family of STP protocols, which are deployed in various versions and implementations (e.g. MST, RSTP and PVRSTP). In any case, it is advisable for central data centre switches to be selected and configured as the primary or secondary root bridge. As already mentioned, these are usually the focal points of the network and they also provide the high computing power needed for fast network convergence in the event of failure.

7 L3 design – linking network routers

Routers, whether in one or more locations, can be combined into a virtual multi-chassis solution or two separate routers can be operated that reciprocally back each other up on the second or third layer of the OSI model.

7.1.1 Virtual multi-chassis solution

Advantages

- integrated management
- better use of data links (elimination of passive and unused connections)
- minimisation of downtime in the event of physical breakdowns

Disadvantages

- reduced resistance to configuration or functional errors
- split brain – devices are maximised in the network to the same extent, which obviously causes great problems
- generally technically congruent devices from the same manufacturer must be used

7.1.2 Standalone routers

Advantages

- separate management – increased resistance to configuration errors
- split brain – the device appears on the network as various devices
- various devices and devices of various manufacturers can be used

Disadvantages

- separate management – increased administrative demands
- backup data links are not used in the topology

7.2 Dynamic routing protocols

IP networks and routers are usually integrated into other network infrastructures. Particularly in larger networks, it is advisable to propose solutions to reflect the internal network routing of the data centre network, where we require fast convergence.

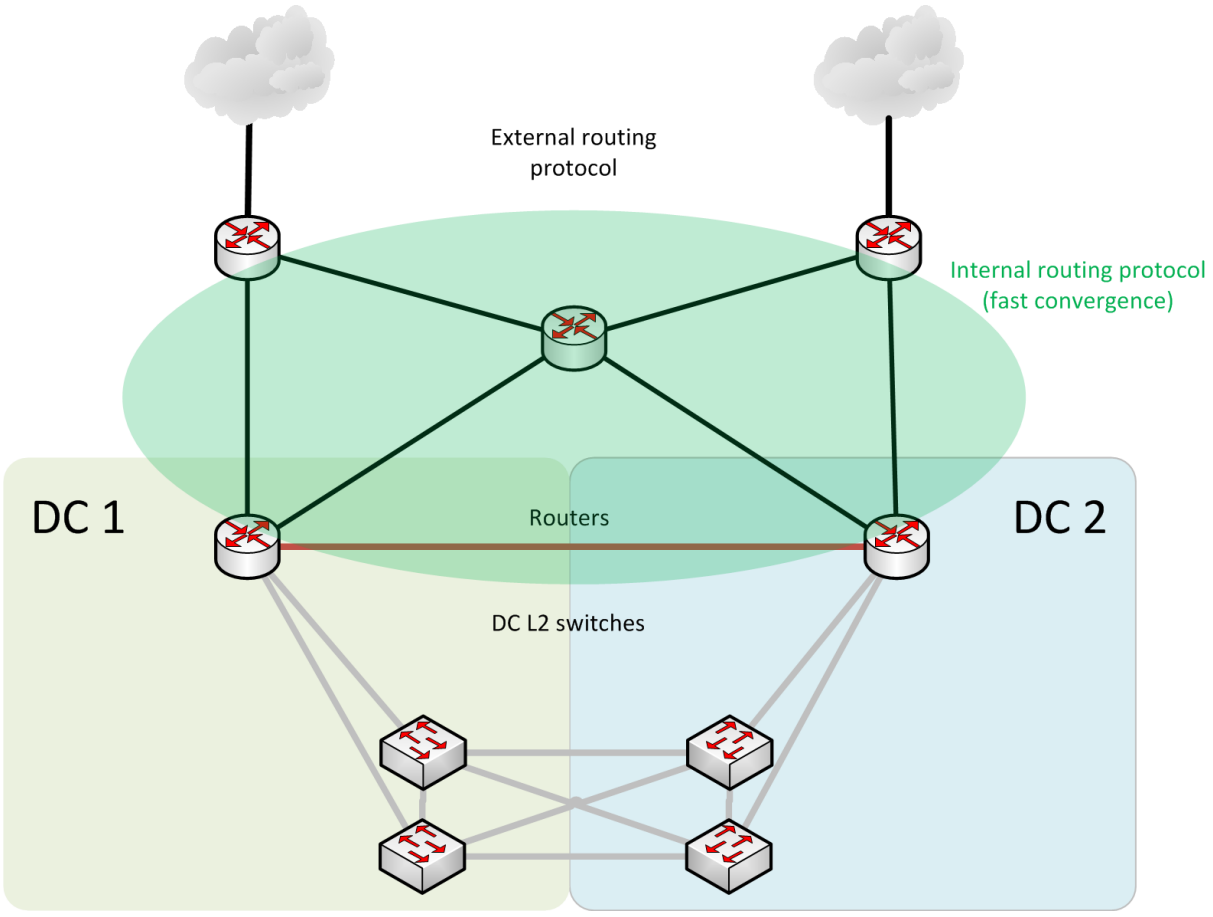

Fig. 5: Routers and use of router protocols.

The rule that it is necessary to connect the element to at least two other independent elements for higher availability applies across all network layers. Thus the data centre routers are connected to two data switches centre, in the same way as they are connected in the direction of the backbone network to two backbone routers.

For fast convergence in case of failure it is advisable to use one of the rapidly converging internal routing protocols. For example, the open protocols OSPF and IS-IS or proprietary IGRP and EIGRP.

It is relatively good to use the load-balancing function (see load-balancing) and define the number of paths to the target at the same value (maximum paths) for internal routing protocols. Routing protocols also have, according to implementation, backup path functions that precalculate the backup path which can be used immediately upon failure detection.

External connectivity is usually addressed by one of the external routing protocols, most commonly the BGP protocol.

7.3 First-hop redundancy on L3

To ensure "first-hop redundancy" we use two or more routers that provide the services of a default gateway for terminal systems in a computer network. The role of the default gateway for a single IP network is played by one of the routers. The backup router starts to provide the role of the default gateway only when it detects the unavailability of the primary router.

Only the GLBP protocol supports load-balancing of the operation of terminal systems on a single IP network between several routers. The other protocols do not support load-balancing, though it is possible to define different routers in the role of the default gateway on a single interface for different IP networks.

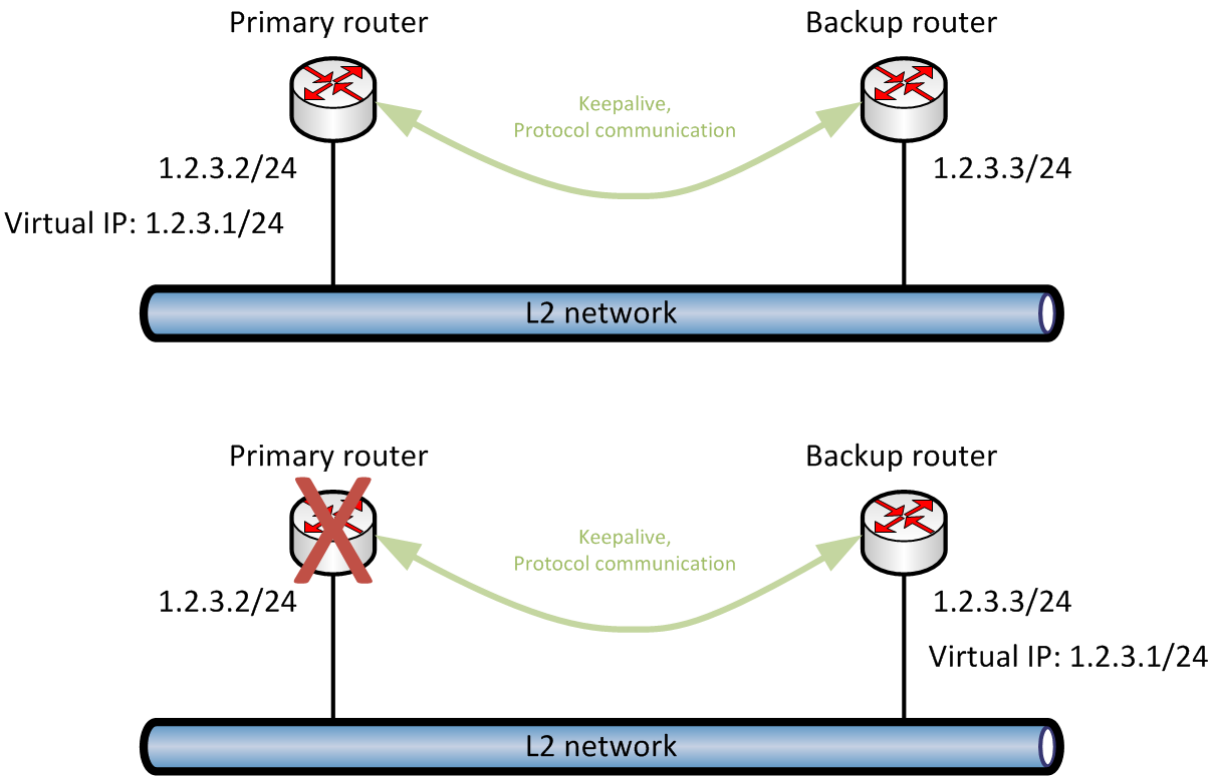

Fig. 6: Principle of first-hop redundancy

	VRRP	VRRP3	HSRP	GLBP
Specifications	RFC 3768	RFC 5798	RFC 2281	Cisco proprietary
IPv6 support	No	Yes	Yes	Yes
Multicast IPv4 address	224.0.0.18	224.0.0.18	224.0.0.102	224.0.0.102
Multicast IPv6 address	-	FF02:::12	FF02::66	FF02::224.0.0.102
Transport protocol	UDP/112	UDP/112	UDP/1985	UDP/3222
Timers	Advertisement 1s down time=3*A	Advertisement 1s down time=3*A	Hello 3s, Hold 10s	Hello 3s, Hold 10s

Number of routers	2, one active router	2, one active router	2, one active router	Up to 4 active routers can be used
Load-balancing function	Several VRRP groups are supported by the interface.	It supports load-balancing, several active routers can be set up and several VRRP groups are supported by the interface.	Several HSRP groups are supported by the interface.	It is oriented and supports load-balancing.
Object tracking	No	Yes	Yes	Yes
Interoperability	Standard	Standard	Cisco proprietary	Cisco proprietary

7.4 Configuration

These given configurations are intended for Cisco routers. Non-proprietary protocols can be similarly deployed on routers made by other manufacturers, even in combination with routers from different manufacturers.

7.4.1 VRRP

VRRP is a protocol that supports the IPv4 protocol only, but it is still often encountered and still used today. The main reason is its long history and, if there is no need to support the IPv6 protocol, it is sufficient and its configuration is very simple. The protocol is listed here particular for comparison and for information, and possible migration to any of the more modern protocols.

However, when planning HA for new installations, it is now advisable to use more modern protocols, which can be seamlessly extended to support IPv6.

A router with higher priority primarily retains the stand-by IP/MAC address.

Router A:

```
interface vlan 10
      ip address 1.2.3.2 255.255.255.0
      vrrp 10 ip 1.2.3.1
      vrrp 10 timers advertise 3
      vrrp 10 preempt delay minimum 10
      vrrp 10 priority 150
```

Router B:

```
interface vlan 10
      ip address 1.2.3.3 255.255.255.0
      vrrp 10 ip 1.2.3.1
      vrrp 10 timers advertise 3
      vrrp 10 preempt delay minimum 10
      vrrp 10 priority 100
```

Output of the VRRP state on one of the routers:

```
Vlan10 - Group 10
  State is Master
  Virtual IP address is 1.2.3.1
  Virtual MAC address is 0000.5e00.0118
  Advertisement interval is 3.000 sec
  Preemption enabled, delay min 10 secs
  Priority is 150
  Master Router is 1.2.3.2 (local), priority is 150
  Master Advertisement interval is 3.000 sec
```

```
      Master Down interval is 9.414 sec
```

7.4.2 VRRP3

In terms of interoperability the use of this protocol is advisable, because support for this open protocol can be found among almost all network router manufacturers. In terms of characteristics it is similar to the HSRP protocol, but it has lower initial timer values. It does not support load-balancing. Version 3 also supports the IPv6 protocol, in contrast to the earlier versions.

Router A:

```
fhrp version vrrp v3
interface vlan 10
   ip address 1.2.3.2
   ipv6 address 2001:db8:1001:123::2/64
 vrrp 10 address-family ipv4
   address 1.2.3.1
 vrrp 10 address-family ipv6
   address 2001:db8:1001:123::1
```

Router B:

```
fhrp version vrrp v3
interface vlan 10
   ip address 1.2.3.3
   ipv6 address 2001:db8:1001:123::3/64
 vrrp 10 address-family ipv4
   address 1.2.3.1
 vrrp 10 address-family ipv6
```

```
address 2001:db8:1001:123::1
```

7.4.3 HSRP

HRSP is a Cisco proprietary protocol, but it is deployed in operations quite often. Since the protocol is proprietary, its implementation is only available for Cisco devices. It supports IPv4 and IPv6 protocols and its functionality is basically identical to the VRRP v3 protocol.

Router A:

```
interface vlan 10
     ip address 1.2.3.2 255.255.255.0
     standby 10 ip 1.2.3.1
     standby 10 priority 200
     standby 10 preempt delay minimum 60
     standby 10 authentication authpass
     standby 10 name vlan10
```

Router B:

```
interface vlan 10
     ip address 1.2.3.3 255.255.255.0
     standby 10 ip 1.2.3.1
     standby 10 priority 100
     standby 10 preempt delay minimum 60
     standby 10 authentication authpass
     standby 10 name vlan10
```

Output of the HSRP state on one of the routers:

```
Vlan10 - Group 10
  State is Active
     2 state changes, last state change 23w0d
  Virtual IP address is 1.2.3.1
  Active virtual MAC address is 0000.0c07.ac24
     Local virtual MAC address is 0000.0c07.ac24 (v1 default)
  Hello time 3 sec, hold time 10 sec
     Next hello sent in 2.224 secs
  Authentication text, string "authpass"
  Preemption enabled, delay min 60 secs
  Active router is local
  Standby router is 1.2.3.3, priority 100 (expires in 9.648
sec)
  Priority 200 (configured 200)
  Group name is "vlan10" (cfgd)
```

7.4.4 GLBP

GLBP (Gateway load-balancing Protocol) supports both IPv4 and IPv6 protocols. In contrast to the other protocols, it offers the option of balancing the load of the network terminal systems. ARP / IPv6 ND provides one dedicated router known as AVG (*Active Virtual Gateway)*, which ensures that terminals are transparently updated with the router's virtual MAC address. This ensures that each router ensures operation of a group of terminal devices. Load is balanced by the Round-Robin system, therefore network operation load or volume are not considered.

Configuration of router A:

```
interface vlan 10
   ip address 1.2.3.2
   ipv6 address 2001:db8:1001:123::2/64
   glbp 10 authentication text authpass
```

```
    glbp 10 ip 1.2.3.1
    glbp 10 ipv6 2001:db8:1001:123::1
    glbp 10 priority 100
```

Configuration of router B:

```
 interface vlan 10
    ip address 1.2.3.3
    ipv6 address 2001:db8:1001:123::3/64
    glbp 10 authentication text authpass
    glbp 10 ip 1.2.3.1
    glbp 10 ipv6 2001:db8:1001:123::1
    glbp 10 priority 200
```

The state output looks like this:

```
Vlan10 - Group 10
  State is Active
     1 state change, last state change 00:00:33
  Virtual IP address is 1.2.3.1
  Hello time 3 sec, hold time 10 sec
     Next hello sent in 0.192 secs
  Redirect time 600 sec, forwarder time-out 14400 sec
  Authentication text, string "authpass"
  Preemption disabled
  Active is local
  Standby is unknown
  Priority 100 (default)
  Weighting 100 (default 100), thresholds: lower 1, upper 100
  Load balancing: round-robin
  Group members:
```

```
    0026.cb39.a5c0 (1.2.3.2) local
There is 1 forwarder (1 active)
Forwarder 1
    State is Active
    1 state change, last state change 00:00:22
    MAC address is 0007.b400.0a01 (default)
    Owner ID is 0026.cb39.a5c0
    Redirection enabled
    Preemption enabled, min delay 30 sec
    Active is local, weighting 100
```

8 Problems and their solutions

8.1 Split brain

This scenario occurs at the moment when there is a breakdown in communication between the routers ensuring high availability. In such a situation there are at least two routers with master router status and they primarily handle inbound and outbound traffic and it is likely there will be asymmetries and loss of part of the network traffic.

This scenario may occur, for example, when interrupting routes between routers or locations in the event of interface breakdown.

A solution is to provide two independent paths (sufficient for L2) in order to minimise this risk. Routes should be terminated at other interfaces or modules in order not to cause even a single breakdown.

8.2 Problems with network asymmetries

Network asymmetry can occur very easily with redundant integrations. This means that traffic from point A to point B does not go through the same network path as from point B to point A. The asymmetry may be permanent or transient.

It often occurs permanently with complex configurations, non-consolidated path values in network protocols or preferences for one of the two equivalent paths when the other party chooses a different path.

If possible, it is advisable to allow routers to use multiple paths, when those paths have the same value in terms of optimal path selection. Firstly, in this way we can ensure load-balancing between multiple paths and need not be unpleasantly surprised if there is a change in the path and we have an error in the network that manifests at the least appropriate moment – thus, when active use of the unused path must occur.

Asymmetries also have disadvantages. The main disadvantage is the need to partially disable filtering of interpolated traffic – uRPF (Unicast Reverse-Path Forwarding).

URPF technologies are deployed at router interfaces, usually at the interfaces of terminal networks and their task is to drop interpolated IP addresses. The router checks the source IP address for each packet. If routing is not set for this address (i.e. a return route) at the interface from which the packet arrived, it drops the packet.