

eVIP Internetworking

Evolved Virtual IP

INTERWORK DESCRIPTION

Copyright

© Ericsson AB 2017. All rights reserved. No part of this document may be reproduced in any form without the written permission of the copyright owner.

Disclaimer

The contents of this document are subject to revision without notice due to continued progress in methodology, design and manufacturing. Ericsson shall have no liability for any error or damage of any kind resulting from the use of this document.

Trademark List

All trademarks mentioned herein are the property of their respective owners. These are shown in the document Trademark Information.



Contents

1	Introduction	1
1.1	Prerequisites	1
2	Connecting Cluster to External Data Communication Networks	3
2.1	Containment of VIP Addresses	3
2.2	Front End	4
2.3	Interlinking Networks	8
3	Configuring eVIP with OSPF	11
3.1	OSPF between Cluster and eVIP Gateway Router	11
3.2	OSPF Supervision	12
3.3	OSPF Areas to Cluster	13
3.4	FEE Interface	14
3.5	FEE and OSPF Router ID	16
4	Configuring eVIP with Static Routing and BFD	17
5	Configuring eVIP with Static Routing	19
6	Geographic Redundancy	21
6.1	Active-Active and Active-Standby Switchover	21
6.2	System Developer Specifics	22
7	Path Diversity	25
7.1	Scope and Purpose	25
7.2	Network Resiliency	25
7.3	Functional Overview	26
7.4	Basic Principle	27
7.5	Deployment Considerations in Virtualized Environments	32
8	Interworking Rules, Recommendations, and Limitations	33
8.1	Rules	33
8.2	Recommendations	34
8.3	Limitations	37
9	eVIP Gateway Router to DCN	39



10	Examples of Networks with L2 Switches	41
-----------	--	-----------



1 Introduction

Evolved Virtual IP (eVIP) is used to connect a cluster to one or more external Data Communication Networks (DCNs). A concept for collective addressing with Virtual IP (VIP) addresses is used. With eVIP, a shared IP address is used to address distributed functions in a multi-processing cluster. The shared IP addresses are called VIP addresses.

This document describes the interfaces and basic configuration concepts. It is intended to be used for network engineering and must be read as a complement to *eVIP Management Guide*.

1.1 Prerequisites

Ensure that the manufacturer documents for the eVIP gateway routers are available.





2 Connecting Cluster to External Data Communication Networks

A cluster is connected to DCNs through eVIP gateway routers, see Figure 1.

At least two eVIP gateway routers are used to connect the cluster to external DCNs to protect against a situation of router failure. The routers forward incoming traffic with destination IP addresses, corresponding to VIP addresses, to the cluster. The VIP addresses are in the cluster configured by eVIP. The interfaces from the cluster to the routers are in the cluster configured by eVIP Front-End Elements (FEEs).

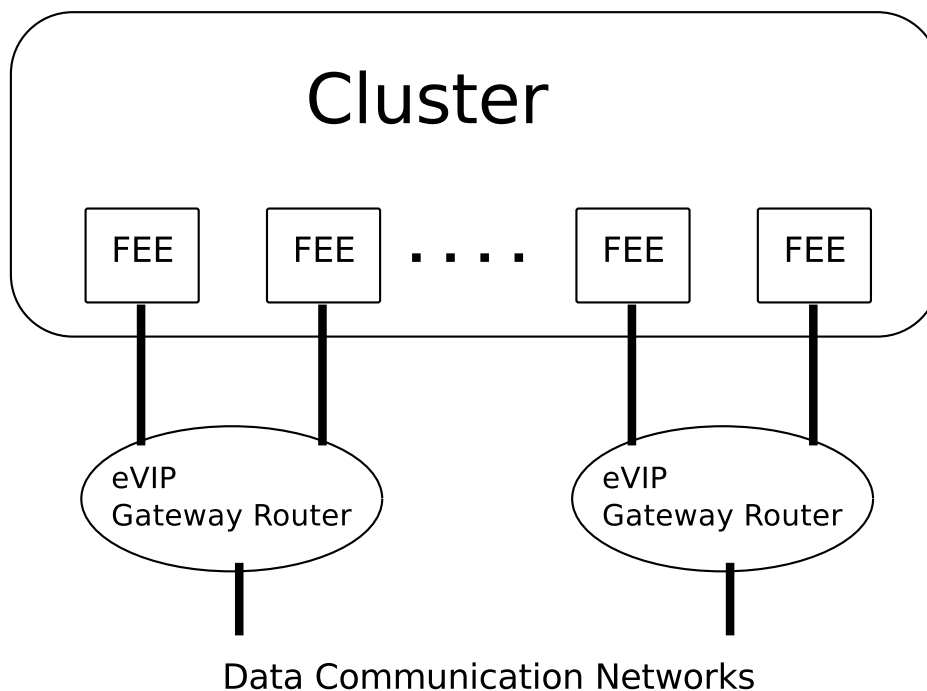


Figure 1 Cluster Connected to DCNs through eVIP Gateway Routers

2.1 Containment of VIP Addresses

VIP addresses are in eVIP configured to an Abstract Load Balancer (ALB), which is a technical term in eVIP used for configuring VIP addresses, external interfaces, and other eVIP resources. An ALB is a logical container used to scope configuration data and facilitate network separation. External network interfaces to a DCN are configured to an ALB.

Typically, separate DCNs are configured to separate ALBs, see Figure 2 An ALB is given a name when configured in runtime to identify the ALB, for example, O&M_Network_LB and Traffic_Network_LB.

Up to 8 ALBs can be configured in a cluster and each ALB must be given a unique name.

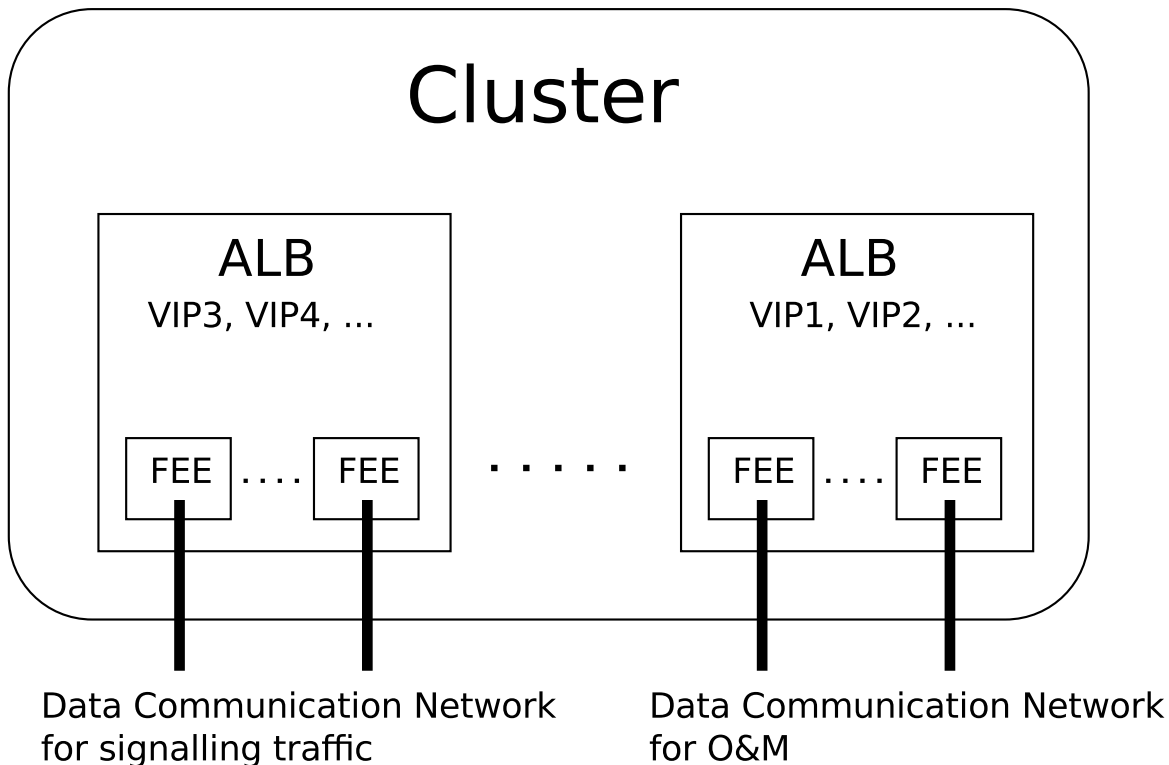


Figure 2 Cluster Connected to Two Separate DCNs

An ALB is configured with its own set of VIP addresses. The VIP addresses are virtual, that is, they are not configured to any particular physical interface or Virtual Local Area Network (VLAN). The VIP addresses can be IPv4 or IPv6 addresses, which are known to the external network, for example, public IP addresses. Each ALB can have a collection of IPv4 and IPv6 addresses configured as VIP addresses, for example, a set of non-contiguous addresses.

Note: VIP addresses from different non-private ALBs must not overlap, that is, the same VIP addresses configured to one non-private ALB must not be reused for another non-private ALB. VIP addresses of private ALBs can overlap. Moreover private ALBs might re-use VIP addresses configured in non-private ALBs as well.

2.2 Front End

The processing units in the cluster with external eVIP interfaces are called "front-end processing units". The Front-End Element (FEE) instances are on the front-end blades. The mapping between the external Layer 3 interfaces and the FEE instances is one-to-one. The external Layer 3 interfaces are used to interlink the FEEs with an eVIP gateway router. Each external Layer 3 interface of an FEE has an interface IP address, but these interface addresses are not



VIP addresses. The external Layer 3 interfaces can reside in a VLAN or an untagged Ethernet network.

The FEEs are contained in an ALB. An ALB contains a set of FEE instances. Different FEE instances on the same blade can belong to different ALBs and then announce to the eVIP gateway routers the VIP addresses pertaining to their respective ALB, see Figure 3.

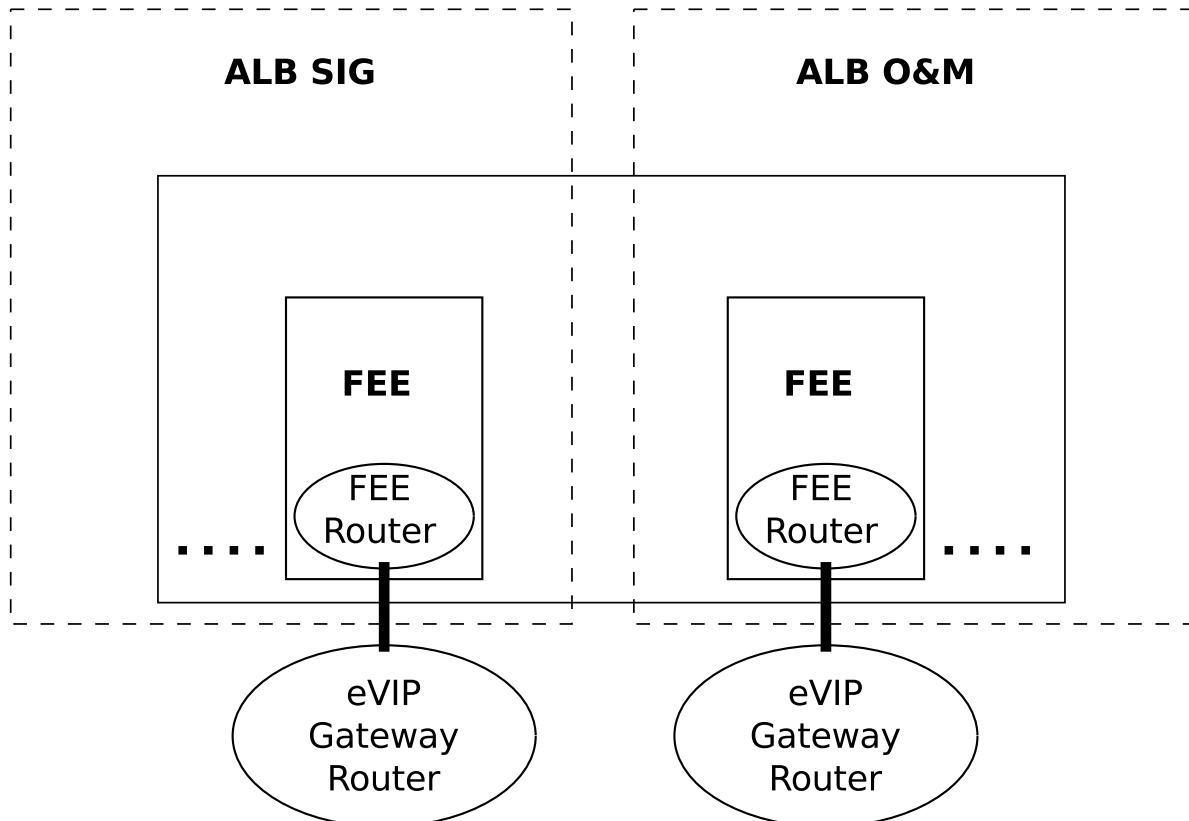


Figure 3 Blade with Two Interfaces Belonging to Different ALBs

At least two physical eVIP gateway routers are required to achieve redundancy. Dynamic routing is normally used in the internetworking between FEEs and eVIP gateway routers. The VIP addresses of an ALB are then announced to one or more eVIP gateway routers by a dynamic routing protocol, Open Shortest Path First (OSPF). The OSPFv2 protocol is used for IPv4, and OSPFv3 is used for IPv6 VIP address announcements.

When FEEs are configured with router instances using OSPF, the eVIP gateway routers regard the cluster as a collection of routers using the OSPF routing protocol. The eVIP gateway router therefore regards each connected FEE instance as a router.

The OSPF hello protocol takes care of the next hop supervision between FEEs and VIP gateway routers in both directions. The OSPF supervision can optionally (by configuration settings) be aided by the Bidirectional Forwarding

Detection (BFD) protocol for more rapid link failure detection, see Section 3.2.1 on page 12.

The use of OSPF between FEEs and eVIP gateway routers is recommended, because the protocol prevents accidental configuration of routing loops and provides dynamic management of VIP addresses from the cluster.

The use of OSPF between FEEs and eVIP gateway routers is typically recommended, because of succinct configuration, the protocol prevents accidental configuration of routing loops and provides dynamic management of VIP addresses from the cluster. The use of OSPF is necessary in conjunction with eVIP and Anycast based geographical redundancy.

If the solution that eVIP is being deployed on does not allow for the use of OSPF, then the option exists to use static routing with or without BFD. More consideration must be given to network engineering when using static routes, especially without BFD for supervision. The capabilities of the deployment infrastructure that eVIP is internetworking with must be taken into consideration to avoid traffic black holes.

2.2.1 Deployment of FEEs

A physical, bare-metal environment has normally hardware constraints, such as physical interfaces and cabling dictating where the FEEs can be deployed in the cluster.

In a virtualized environment, such as cloud, deployment constraints normally do not exist as the external network interfaces are also virtualized. In the elastic and scaling cloud environment, it is often desired to have a set of uniform processing units (Virtual Machines (VMs)) making it possible to deploy FEEs on any of the processing units provided to the clustered application.

If the FEEs are on processing units that potentially can disappear during a scaling-in, there is a risk that eVIP can lose all external connectivity, causing a complete traffic outage.

To circumvent this, it is possible to configure FEEs as floating. This means that an FEE is automatically relocated to another processing unit if the current utilized processing unit disappears.

When relocating an FEE, the Layer 3 configuration is brought along. This means that each processing unit that can accommodate a floating FEE must have an external network interface connected to the Layer 3 network providing access to the eVIP gateway routers. This allows the FEE to reestablish connectivity with the eVIP gateway router when relocated.

During FEE relocation the Layer 2 link breaks. If BFD link supervision is used, the disturbances are only minor as traffic quickly is relocated to other redundant links and FEEs. When the Layer 2 link is reestablished, the BFD session is also reestablished allowing the relocated FEE to carry traffic again.



2.2.2 Resilient FEE IP Addresses

In deployments where the intermediary connectivity infrastructure between the FEE to the eVIP gateway router, end-to-end, is highly available to such a degree that there is no technical need for any Layer 3 supervision protocol (for example, BFD), eVIP offers a resiliency mode of operation based on the ability of eVIP to automatically relocate the external IP address of an FEE to a "Host FEE". On a "Host FEE" multiple relocated external IP addresses can be "stacked".

This mode of operation is referred to as "Resilient FEE IP Addresses" or "IP stacking". "Resilient FEE IP Addresses" can only be used for configurations with static routing without Layer 3 supervision, where it is automatically applied for floating nodes, see Section 2.2.1 on page 6, and optionally for non-floating nodes, that is "fixed nodes".

It is required that the FEEs of an ALB, which are configured for resilient FEE IP addresses, share the same subnet for external connectivity toward eVIP gateway routers. The automatic relocation and stacking of resilient FEE IP addresses is controlled by eVIP and can take place in response to failure detected by eVIP or can occur in response to recovery-related handling and system reconfiguration.

Stacking can be used to manage cases where no processing units are available for relocating a floating FEE to or, optionally, if a static FEE becomes unavailable. In the floating case such a situation would occur if an ALB has more FEEs to float than there are available processing units or, if an FEE on a lower priority processing unit has been displaced because of the floating of a higher priority node.

The stacking capability is only supported in certain configurations where no dynamic or supervised static routing is provided and when the deployment infrastructure meets the necessary requirements, see Section 8.2 on page 34. When the conditions have been met to stack an IP address, then only the FEE's external IP address is relocated and hosted on another FEE under the same ALB. This move is typically temporary, once the "Home FEE" (the FEE that the IP address is provisioned upon) is available again; the IP address will be enabled again on this FEE and removed from the "Host FEE".

The traffic from the gateway routers follows the movements of the resilient FEE IP addresses. This is achieved by sending gratuitous ARP messages from the resilient FEE IP address that results in traffic being attracted to the said resilient FEE IP address.

To achieve an even distribution of traffic across FEEs when IP stacking occurs, the following measures have been implemented by eVIP:

- Per ALB, the least "stacked" FEE (the one with the fewest external IP addresses assigned to it) is always chosen to host an IP. If there is a tie situation, the highest priority interface is chosen first. In order to stack on fixed nodes the configuration option for stacking fixed node must be set for each intended ALB.

- The distribution of IP addresses on FEEs is checked periodically and will be rebalanced automatically as required.

With all the movement of IP addresses between FEEs, it is possible that the layer 2 (MAC) addresses associated with the FEE external IP addresses on the underlying switch or next hop router that are connected to the FEE could get stale. To prevent this situation each FEE will send a periodic gratuitous ARP message for each IP address that is provisioned on it.

2.3 Interlinking Networks

The eVIP gateway routers are connected to the cluster through interlinking IP networks.

The interlinking networks can exist in directly connected cables or fiber between front-end interfaces and the eVIP gateway routers. That is, they are carried over dedicated physical Ethernet links or the interlinking networks can be connected through an intermediary switch or switches.

An example of front-end blades interconnected with physical cables and one subnet per cable is shown in Figure 4.

For example configurations with switches, see Section 10 on page 41.

The interfaces on the cluster side front end, which are used for communication with the eVIP gateway routers, are called "external FEE interfaces". These are "Layer 3 interfaces" and each external interface has an IP interface address.

The interlinking networks are configured in the FEEs and in an eVIP gateway router. In an FEE this is done by configuring an IP subnet to a named bridged interface provided by Linux®, for example, as in a Linux Open Telecom Cluster (LOTIC) bundling.

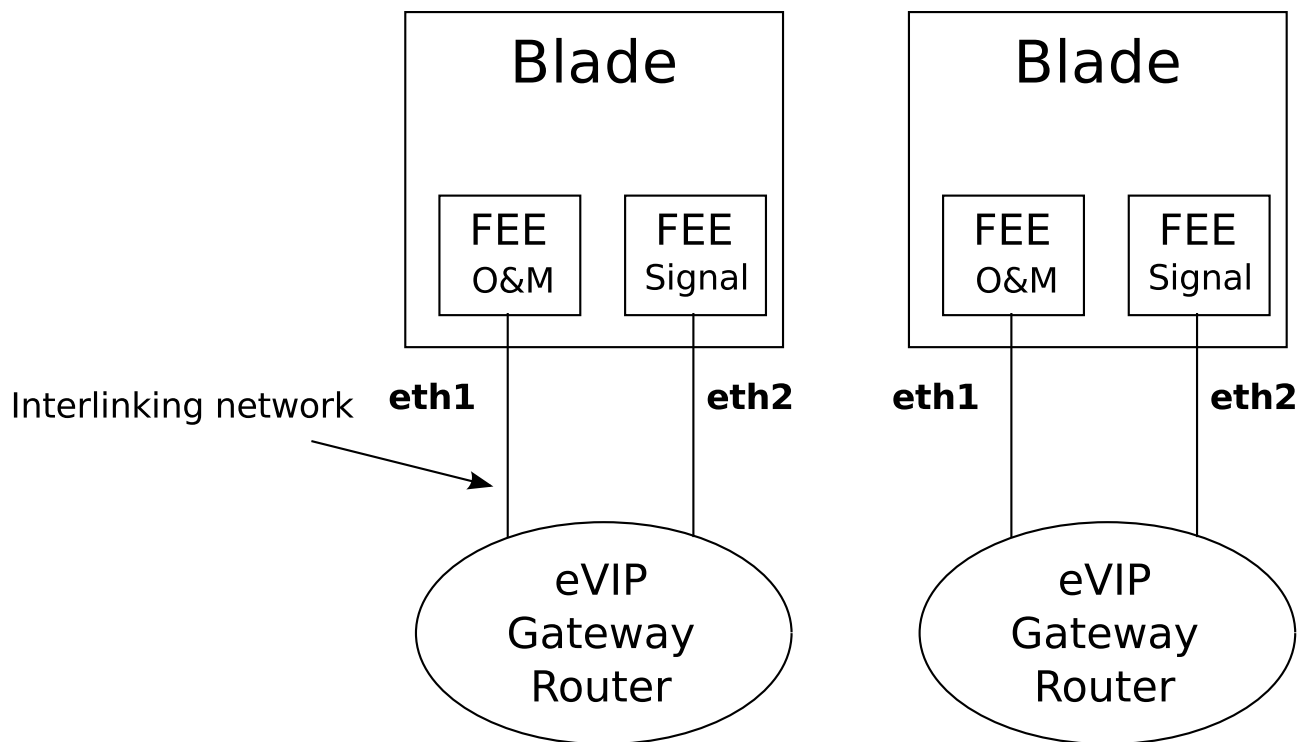


Figure 4 Front-End Blades Interconnected with Cables

An external eVIP interface can be part of a LAN or VLAN. The LANs or VLANs for the interlinking networks are mapped one-to-one to interlinking subnets. An external FEE interface is part of an interlinking subnet. The mapping between external eVIP interfaces and FEEs is one-to-one, that is, for each FEE in eVIP only one external interface is configured. A blade with four VLAN interfaces over two physical interfaces is shown in Figure 5.

Note: The VIP addresses configured to the ALBs are not part of the interlinking subnets. For example, the interlinking subnets are private IP addresses whereas the VIP addresses can be public IP addresses.

The external Layer 3 interfaces are typically, for IPv4, part of small subnets used to interlink an eVIP gateway router with the corresponding FEE. For IPv6, the interlinking is done at the link layer using IPv6 "link local addresses". Conceptually the OSPFv3 model (which is used for IPv6) is slightly different from OSPFv2. For OSPFv3 the router is said to connect to the "link" and not to the "subnet".

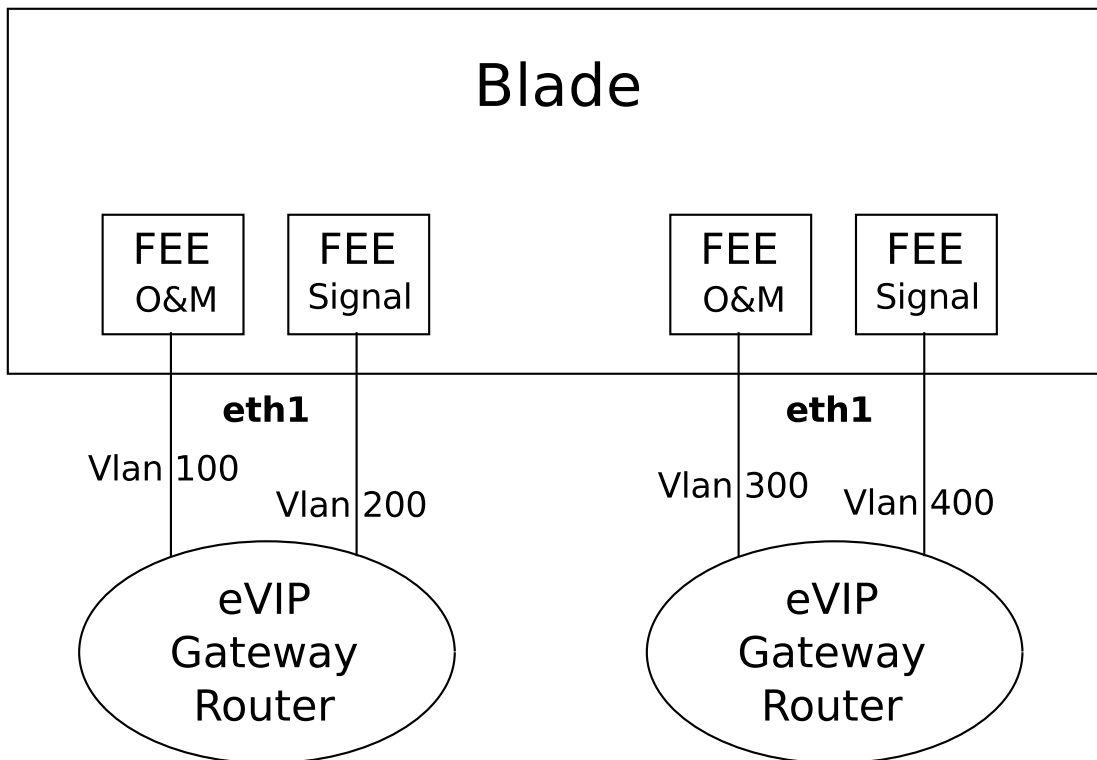


Figure 5 Blade with Four VLAN Interfaces over Two Physical Interfaces



3 Configuring eVIP with OSPF

An eVIP gateway router is configured according to instructions provided by the manufacturer. When the eVIP gateway router is connected to the cluster, the eVIP function must be made aware of the gateway. This is done by configuring an FEE in eVIP for each Layer 3 interface on the cluster side that is used for interlinking the cluster to the gateway router, refer to *eVIP Management Guide*.

When configuring an FEE instance in eVIP, the following values must be known:

- Front-end node attached to the eVIP gateway router
- The name of the external eVIP interface on the front-end node
- The IP address of the eVIP gateway router interface
- OSPF parameters
- The ALB to which the FEE instance belongs

The front-end nodes and interfaces are identified by names provided by the used middleware or operating system, for example, LOTC.

The OSPF parameters include timer intervals, router ID, area ID, and a priority value. The OSPF parameters must correspond to the settings in the eVIP gateway router. Optionally, the BFD can be used with OSPF where the BFD parameters must correspond to the settings in the eVIP gateway router for stable behavior. For more information about the OSPF parameters, see Section 3.1 on page 11.

3.1 OSPF between Cluster and eVIP Gateway Router

The OSPF routing protocol is used between the cluster and the eVIP gateway router. For a description of the OSPF protocol, refer to OSPF Version 2.

OSPF is used for the following tasks:

- Supervision of links and routers
- Redundancy switch-over for links and routers
- Establishment of equal cost paths to a VIP address destination in the cluster

The eVIP gateway router regards the cluster as a collection of OSPF neighbor routers. By eVIP the cluster starts a number of OSPF router instances, called FEE router instances, on the front-end nodes. The eVIP gateway router automatically establishes OSPF adjacencies with the neighboring FEE router instances, which in this way become OSPF peers.

The FEE router instances use OSPF functionality for link supervision and to announce VIP addresses of an ALB to the eVIP gateway router. The announcement of VIP addresses to the eVIP gateway router attracts packet traffic to the announcing FEE router instances. In this way, packets with destination IP addresses matching the VIP addresses are attracted to interfaces of an ALB.

3.2 OSPF Supervision

For supervision, both the FEE router instances in the cluster and the eVIP gateway router periodically send out hello packets. Timer values are configurable. If a link failure occurs, the eVIP gateway router "loses contact" with the corresponding FEE router inside the cluster and hello is not answered. OSPF updates the routing information so that incoming packets can be forwarded over other available links.

For outgoing packets, the FEE router instances in the cluster work in a similar way. The FEE router instances detect that contact is lost with the eVIP gateway router, and eVIP ensures that the packets are forwarded over other available links belonging to the same ALB.

The timer values for `hello_interval` and the `dead_interval` must be the same in the eVIP gateway router and the FEE router instance.

Table 1 Default Values for OSPF Supervision

OSPF Parameter	OSPFv2 Default Value	OSPFv3 Default Value
hello_interval	10 seconds	10 seconds
dead_interval	40 seconds	40 seconds
retransmit_interval	5 seconds	5 seconds
transmit_delay	1 second	1 second
router_priority	0 (zero)	0 (zero)
spf_delay	0.5 second	0.5 second
spf_interval	1 second	1 second

3.2.1 OSPF Supervision with BFD

Fast supervision of OSPF peers can be done through the BFD protocol. This requires that the router model used as eVIP gateway router at least supports BFD asynchronous mode with OSPF.

OSPF with BFD supervision works as follows:

- When an OSPF adjacency is established with a neighbor, OSPF activates the BFD software in the router, which starts to check reachability to the neighbor addresses.



- A BFD session is initiated and the sending of hello packets starts.
- When the BFD failure detection interval expires without receiving a hello, the OSPF software is informed so that the corresponding OSPF adjacency can be taken down and routes can be recalculated.

With the BFD, the failover time can be shortened compared to OSPF without the BFD.

A feature of the BFD is called "echo". The BFD echo can optionally be configured in the FEE router. For an eVIP gateway router with many BFD sessions to answer, the processing load can be overwhelming for some router architectures. In this situation, the use of the echo function can reduce the processing load on the eVIP gateway router. When the echo function is enabled, the normal BFD hello packets are sent at a slower rate. Fast detection is still achieved by echo packets, which generally cost less to process but this depends on the router architecture in the eVIP gateway router. The BFD echo packets are transmitted so that they are echoed back from the interface and data plane of the opposite side, for example, the Network Interface Card (NIC).

Note: The timer values for the BFD and FEE router instance must be coordinated with settings in the eVIP gateway router.

For the BFD parameters and recommended settings, see Section 8.2 on page 34.

3.2.2 SPF Exponential Hold Time Backoff

Shortest Path First (SPF) runs when there is a topology change. The SPF algorithm is a very CPU intensive process. To prevent frequent SPF calculation, the system waits for a period, called the hold time, and delays SPF calculations during network instability. SPF exponential backoff makes it possible to make the hold time variable change based on the frequency of topology changes. The hold time minimum and maximum periods can be configured using the `spf_delay` and `spf_interval` parameters. The hold time changes exponentially when a topology change occurs.

3.3 OSPF Areas to Cluster

OSPF has a concept of routing areas. Between the cluster and gateway routers, the type of OSPF routing area must be either a stub area or a Not-So-Stubby-Area (NSSA). Stub area and NSSA are technical terms defining two specific area types in OSPF and both supported by eVIP.

Note: All routers connected to a stub area must be configured for area type equal to stub area. This includes FEE routers and eVIP gateway routers. All routers connected to an NSSA area must be configured with the area type equal to NSSA. This includes FEE routers and eVIP gateway routers.

A single OSPF stub area or NSSA area between FEEs and gateway routers is sufficient for the eVIP gateway routing purpose. However, several stub areas (or NSSA areas) can be used, see Figure 6. For example, separate stub areas can be used for the ALBs serving separate DCNs.

An OSPF area has an ID that is given in the decimal dot notation format of an IP address. The area ID corresponds to an existing IP address or can be fictitious. However, within the stub area or NSSA area the router ID must be unique.

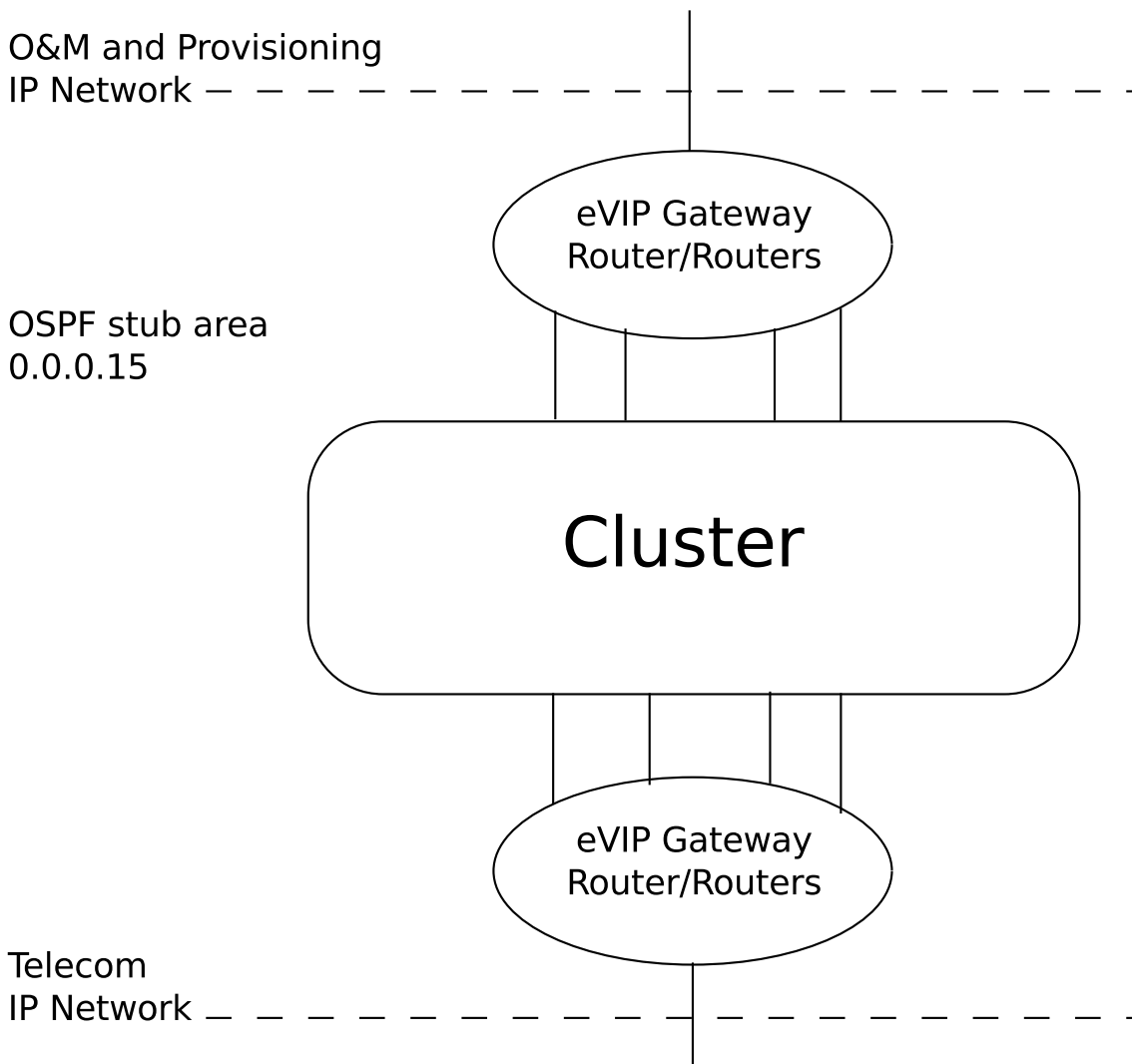


Figure 6 Cluster Connected to Different Networks with Different OSPF Stub Areas

3.4 FEE Interface

External FEE Layer 3 interfaces have a local interface IP address on networks interlinking FEEs to an eVIP gateway router. Only one such Layer 3 interface can belong to an FEE instance. For example, a blade with two external Layer 3 interfaces connected to an eVIP gateway router requires two FEEs, one for



each Layer 3 interface. The Layer 3 interfaces can in this example either be two VLANs or two (untagged) physically separate Ethernet interfaces.

A software router inside FEE is configured to each FEE instance for interworking with the eVIP gateway router. For each eVIP gateway router, at least one FEE router instance is configured in eVIP. Each FEE router instance has its own router ID, see Figure 7.

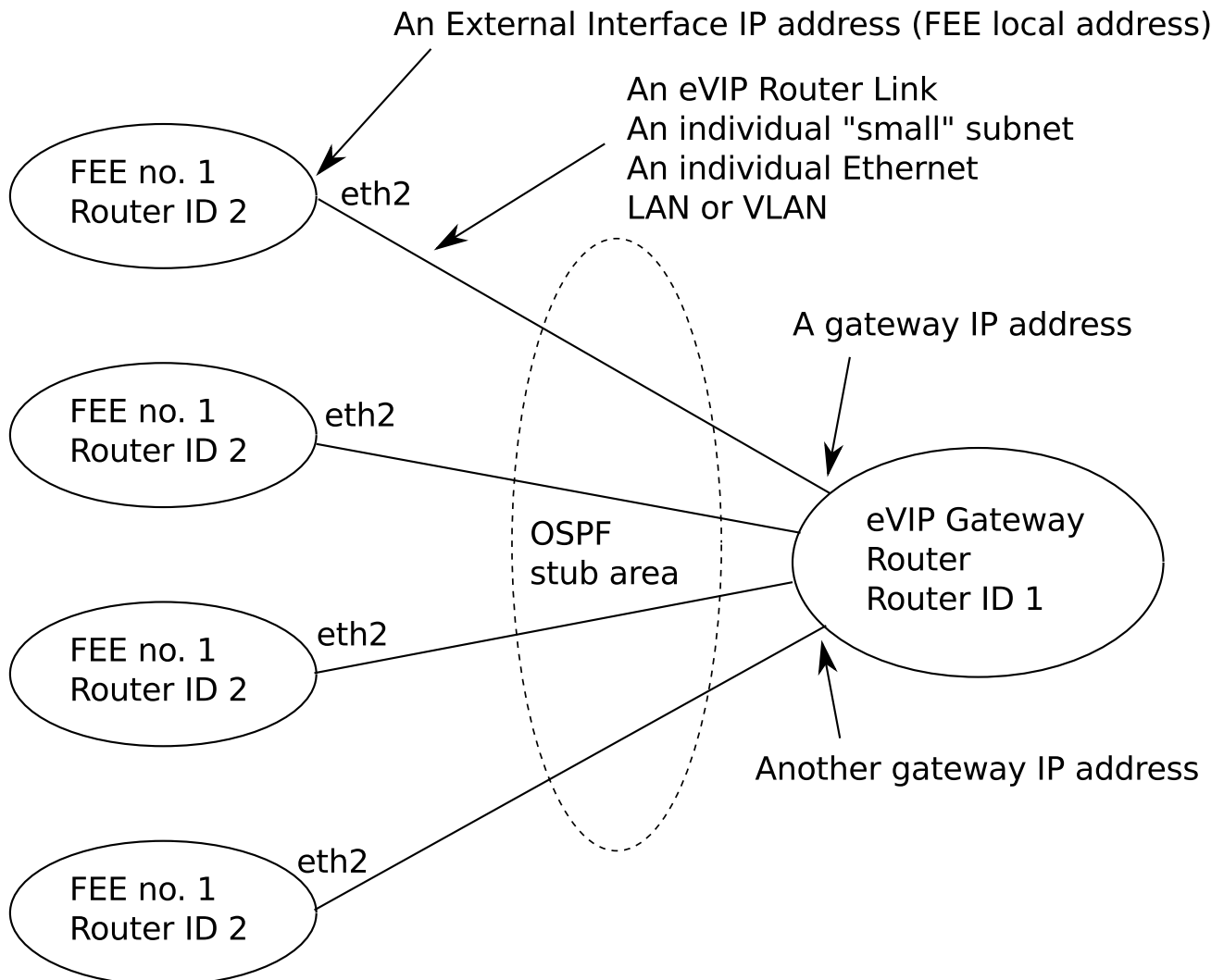


Figure 7 Typical Configuration with One Gateway Interface per eVIP Router Link

The VIP addresses are configured to an ALB, which is associated to one or more FEE instances. The VIP addresses are announced to the eVIP gateway router by OSPF. Packets arriving to the eVIP gateway router, with destination IP addresses corresponding to one of these VIP addresses, are forwarded by the eVIP gateway router to the cluster.

Usually an ALB is configured with several FEE instances, which form a pool of alternative paths from the eVIP gateway router; this corresponds to so called Equal-Cost Multipaths (ECMPs) routing.

For an OSPF stub area, a default route is automatically injected by an OSPF Area Border Router (ABR) into the stub area. The eVIP gateway routers are typically connected to an OSPF back-bone area (area 0) and would therefore act as ABRs. Hence, the outgoing traffic from FEEs is attracted to eVIP gateway routers by the automatically injected default routes.

For NSSA however, the eVIP gateway routers do not automatically inject a default route, but instead the gateway routers must be explicitly instructed by a router configuration command to inject a default gateway into the NSSA. The syntax of this configuration command typically varies for routers from different manufactures.

3.5 FEE and OSPF Router ID

By convention, the router ID for an FEE instance in eVIP is specified to be identical to the IP address of the external eVIP interface connected to the eVIP gateway router. However, this is not strictly necessary. The router ID for an FEE instance in eVIP can, for example, be a fictitious IP address given in the decimal dot notation format.



4 Configuring eVIP with Static Routing and BFD

Static routing with BFD supervision can be used between FEEs and the eVIP gateway router.

When static routing is used, static routes in the eVIP gateway router must be configured with VIP addresses of an ALB as route destination and the "interface addresses" on the FEE side as next hop. If a connection failure between an FEE and an eVIP gateway occurs, it is detected by BFD and then another available FEE is automatically selected.

When static routing is used, eVIP can select any eVIP gateway router with working connectivity to FEEs in the outgoing traffic case.

When static routing is used, as opposed to the case with the dynamic routing protocol OSPF, the selection of eVIP gateway router has no means to consider the network connectivity situation beyond the first hop, that is, from the FEE to the eVIP gateway router.





5 Configuring eVIP with Static Routing

Static routing can be used between FEEs and the eVIP gateway router.

When static routing is used, static routes in the eVIP gateway router must be configured with VIP addresses of an ALB as route destination and the "interface addresses" on the FEE side as next hop.

When static routing is used without BFD supervision, no detection is in place if the eVIP gateway router becomes unavailable or there is a problem between the router and the FEE.

The eVIP gateway router is assumed to always be reachable, otherwise the traffic is black-holed.

For ideal conditions on how to use static routes, see Section 8.2 on page 34.





6 Geographic Redundancy

In telecommunication networks, geographical redundancy is a method used to ensure availability of a service when the provided services rely on a geographically centralized function deployed on a particular node in the network. That is, in the event of a local node failure situation, the centralized function becomes available from another node located elsewhere in the geographically distributed network.

The method is intended to protect against a local node failure that cannot be prevented or mitigated by foresight or local technical arrangement. For example, an earthquake can incapacitate a complete building and all its electrical equipment. The basic idea is to allocate two (or more) geographically separated telecommunication nodes for the centralized functions and to ensure the following:

- Provide network arrangements to support both O&M controlled switchover between the geographically separated nodes under normal operational conditions.
- Have mechanisms implemented that in the event of a disaster situation would effectuate a corresponding automatic failover in the network.

When geographical redundancy is implemented in an application product by eVIP, the OSPF protocol must always be used between FEEs and gateway routers. However, between gateway routers on the DCN side to which all geographical sites are connected other routing protocols can be used, for example, the Border Gateway Protocol (BGP). This example requires that the gateway routers support BGP and can redistribute OSPF into BGP.

For system developers that would implement cluster-based telecom nodes using eVIP, eVIP provides Application Programming Interfaces (APIs) for the application developers to exercise application software control of the announcement (or withdrawal) of VIP addresses into an OSPF stub area or NSSA. This is the fundamental mechanism upon which the eVIP geographical redundancy failover scheme is based.

6.1 Active-Active and Active-Standby Switchover

Geographical redundancy can be implemented so that under normal operational circumstances all geographically separated nodes with the protected centralized function share the traffic load generated in the network. For example, by the "IP Anycast method", where the traffic load from different IP addresses in the network is distributed to the geographically redundant nodes based on geographical proximity routing metrics. However, geographical redundancy can also be implemented so that under normal operational circumstances only one

of the geographically separated nodes with a protected centralized function is active while another node is held in a stand-by state.

When eVIP is used in a product setup providing geographical redundancy, the same VIP address is to be published from several geographically separated sites. If the application system developers choose a solution with an active-standby redundancy scheme, only the active site is allowed to announce the VIP address into the routing network and thus attracting traffic to FEEs.

If a complete outage of the active site occurs, the announced VIP route automatically disappears. However, to cover the failure situation of partial node degradation, the application system developers need to implement fencing logic and threshold logic to determine the fault gravity level upon which the fencing logic is to be activated. When activated, the faulty node is fenced off by withdrawing VIP address routes.

For an IP Anycast arrangement, the network routing protocols automatically take care of failover to the other node (nodes) still announcing VIP address routes. After repair of the faulty node, the threshold and fencing logic would deactivate fencing of the repaired node.

However, for an active-passive (standby) solution, additional failover logic must be implemented by application system developers so that a passive node (standby node), which is not deemed to be faulty, can determine when to become active. When the passive site becomes active, it starts to announce the VIP route from the new network location and thus attracts traffic to this site instead.

6.2 System Developer Specifics

The information in this section is primarily intended for application system developers.

Note: eVIP only supports an API for programmatic control of the activation and deactivation of VIP addresses. Design work by application developers is required for building any complete geographical redundancy solution.

The eVIP API in question is a system internal interface available to application programmers and is referred to in eVIP as the Dynamic Traffic Management (DTM) interface.

If the geographical redundancy solution requires any form of data replication between the geographically separated nodes, this must be implemented by the application design as no support for data replication between sites is provided by eVIP.

From the eVIP perspective, activating a VIP address only means announcing it through OSPF. Deactivation hinders the inbound traffic to the VIP address in question by removing it from the routing network so that incoming traffic with the deactivated VIP address (as destination IP address) stops arriving at FEEs. It has no effect on the existing server sockets. Outbound traffic can still



leave the cluster with packets originating from a deactivated VIP address. The activation or deactivation of the VIP address only decides if the route for the VIP address is announced into the network or not, but is not similar to a firewall barring function. If an application effectively wants to programmatically block traffic on the FEE side, for example, traffic to/from a VIP address, the DTM API can also be used for adding packet filter policies (iptables rules) into the FEEs of eVIP. This would be equivalent to firewall blocking.





7 Path Diversity

Path diversity is a networking method for multi-hop resiliency failover between end-to-end supervised, disjoint multi-hop paths.

In real world network deployments, the number of path diversity paths for a destination is typically limited to two paths for practical reasons to avoid network configuration complexity. Therefore, two diverse paths will henceforth always be assumed to be the offered configuration, unless otherwise explicitly stated.

7.1 Scope and Purpose

Path diversity with eVIP is primarily intended for applications using the Signalling System No. 7 (SS7) Common Application Feature (CAF) with eVIP in networks where multihomed Stream Control Transmission Protocol (SCTP) sessions are used for SCTP communication resiliency.

The IP-based applications designed for end-to-end network path diversity can be able to open several IP sockets so that traffic originating from different sockets can be carried across two distinct network paths to a remote peer socket. For example, traffic originating from one socket is carried over an upper path, whereas traffic originating from a second socket is carried over a second lower path.

Two entirely different methods for achieving path diversity with eVIP and SS7 CAF exist. The two methods are in this document denoted as follows:

- Cluster external path bisection
- Cluster internal path bisection

The specific prerequisites for the use of these methods are different regarding application software design requirements and gateway router configuration requirements.

Some considerations regarding specific deployments of path diversity in virtualized cloud environments are also covered.

7.2 Network Resiliency

Resiliency is a desirable property in many network interconnected systems. This can generally be achieved by redundancy, for example, the case of an IP router, routing around a faulty link by choosing an alternative link. Special aspects of network resiliency and terminology regarding path diversity is explained in the following sections.

7.2.1 Path Diversity versus ECMP

When a router spreads traffic between several available equal links to next-hop routers within an autonomous routing domain, this is referred to as Equal-Cost Multipath (ECMP). For example, the term ECMP is used to denote this type of arrangement when OSPF or static routing is used. Although it is true in a graph-theoretic sense that the ECMP type of single-hop multi-path choice can be seen as a special case of path diversity, the term path diversity is however not used to denote ECMP type of arrangements.

Path diversity means ensuring that a packet can transit between two points in a network, for example, a client and a server, using disjoint redundant paths, where each separate path is assumed, end-to-end, to span multiple routing hops. The primary purpose is to achieve resiliency by avoiding a "shared fate" failure situation. That is, the possibility of sharing any point of failure between the diverse paths along the network, end-to-end. Path diversity implies the existence of an end-to-end supervision between the communicating parties that can trigger an end-to-end path diversity failover.

7.2.2 Diverse Paths and Shared Fate

Ideally, shared fate is avoided at all network and system levels. When different IP routes are separated through logical IP networks they can still share fate in the form of a common underlying resource, for example, the same Layer 2 link, physical equipment, or cable, and thus still be vulnerable to a Single Point Of Failure (SPOF) event. In practical networks however, it is often considered acceptable to allow for some limited shared fate locally, if the equipment (power, Layer 2 link, and so on) at the potential failure points in a network segment is protected by adequate local equipment redundancy methods.

The specific path diversity arrangements on the DCN side of the network are highly customer-specific and can vary within sections of the network. The diverse paths would typically be carried across separated infrastructure, for example representing different Virtual Private Networks (VPNs) or different BGP autonomous systems. If a DCN network is organized so that all physical network resources are bisected into separate VPNs, whereby the traffic flow between remote peers (for example, a client and a server) can traverse any of the two VPN-separated paths, then if a traffic blocking fault situation occurs in one VPN, this triggers a failover to the other VPN.

7.3 Functional Overview

In relation to eVIP, path diversity can be implemented by either of the two following methods:

- Cluster external path bisection
- Cluster internal path bisection



An application based on the SS7 CAF and eVIP must at design time make provisions for which method or methods to support.

The second method requires specific software implementation by the application using the SS7 CAF APIs.

An SCTP-based application based on the SS7 CAF with eVIP can have chosen to implement support for only one of the methods or for both methods. All legacy implementations with SS7 CAF and eVIP have used the first method for path diversity.

7.3.1 Cluster External Path Bisection

In the connectivity segment between eVIP FEEs and gateway routers, no path diversity is implemented. Hence, a shared situation exists here. However, in this segment full redundancy support is implemented. High ability can thus still be upheld without any need for path diversity in this local segment. Towards the DCN side however, the gateway routers separate the traffic into diverse paths in the DCN network. As the bisection of path is done externally on the gateway routers, and not inside the cluster, this method is referred to as cluster external path bisection.

7.3.1.1 Prerequisites and Impact

This method requires a single ALB. The bisection that is the separation of traffic into two diverse paths in the DCN network is achieved entirely by features available in the gateway router. In the network segment between FEEs and gateway routers, there is no path diversity between FEEs and gateway router, but redundancy-wise connectivity is still well protected by OSPF and BFD, or static routing and BFD.

This method has no impact on application software design. For example, legacy application software can be used and this method has no particular impact on the eVIP configuration done on the cluster side. However, this method requires special configuration on the gateway router side.

7.4 Basic Principle

The path bisection is achieved in the gateway routers as follows. Towards the DCN network side, each gateway router is only connected to network resources of a single path diversity path. The two gateway routers can transfer packets between each other.

On a gateway router, once packets arrive from the eVIP FEEs, the gateway router splits the traffic into the two diverse paths. This is done by first determining which alternative to use:

- To send the packet onto the DCN side network of this router, thus selecting a path diversity path



- To transfer the packet to the other gateway router, which would then forward the packet on its DCN network which corresponds to the other diverse path

On the gateway routers, the policies that govern which packets to send to DCN network and which packets to transfer to the other gateway router are typically configured by "route maps" and Policy Based Routing (PBR) rules. For example, for the packets originating from the cluster, the gateway router can inspect the source IP address (originating VIP address) and thereby select a desired path diversity path.

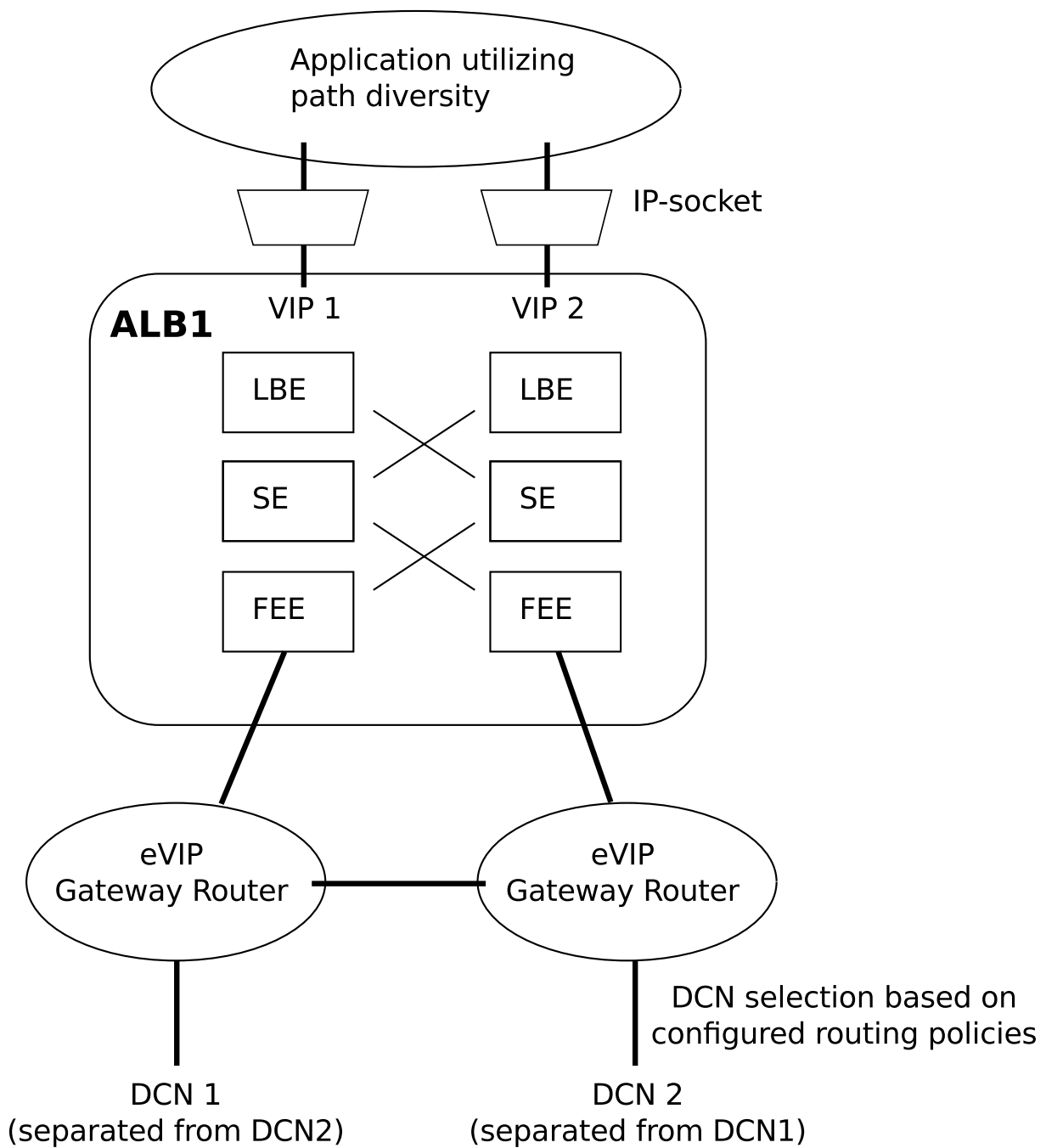


Figure 8 Cluster External Path Bisection

7.4.1 Cluster Internal Path Bisection

The internal path bisection method can be used if the available type of gateway routers would not support the feature set required for cluster external path bisection, or for other reasons a gateway-centric method is not preferred. With the internal path bisection method, the path bisection can be achieved internally

in the cluster already at the socket layer, thus establishing the end-to-end path diversity already at this point. With the cluster internal path bisection method, traffic is logically separated inside the eVIP cluster from the socket layer all the way to the eVIP gateway router.

Note: eVIP can normally only ensure a logical traffic separation inside the cluster. A situation of shared fate with respect to a common hardware unit inside the cluster can still occur, for example, the diverse paths can share the physical backplane. Here backplane resiliency is secured by means of internal redundancy mechanisms.

7.4.1.1 Prerequisites and Impact

Path diversity is achieved by using two ALBs whereby each ALB is used for a distinct path-diversity path. For example, each of the two ALBs mapped to a separate VPN. For example, a distinct DCN-side VPN on a gateway router is extended to the FEEs of a particular ALB. This method requires that special software design toward SS7 CAF is implemented by the application.

7.4.1.2 Basic Principle

For each path diversity path one ALB is connected to one gateway router. That is, all the FEEs of such an ALB must be connected to only one of the two gateway routers.

The eVIP configuration must in this setup include two (or more) ALBs, with each ALB hosting one VIP address. When using this method of path diversity, the application must open two or more IP sockets and relate each of these to a separate ALB by binding the socket to the ALB.

In a deployment using path diversity, the application, for example the SS7 CAF, normally supervises the connectivity and quality of different paths available. Based on the supervision results the application can choose to move the traffic load back and forth between the different paths. The SCTP support for multi-homing is an example of this.

If path diversity is fully secured by the underlying network and the application failover mechanism reacts fast enough, other failover mechanisms at lower layers can become redundant or even superfluous. In principle, this removes the objective need for cluster internal redundancy within the eVIP ALBs. Hence, each ALB would in theory only need to include one of each element type (Load Balancer Element (LBE), Security Element (SE), FEE) and connect to a single eVIP gateway router. However, the number of configured eVIP elements can still in a situation need to be higher to handle the traffic load or provide for an even better level of resiliency. All this depends on the specific application, the protocols used, and its deployment configuration.

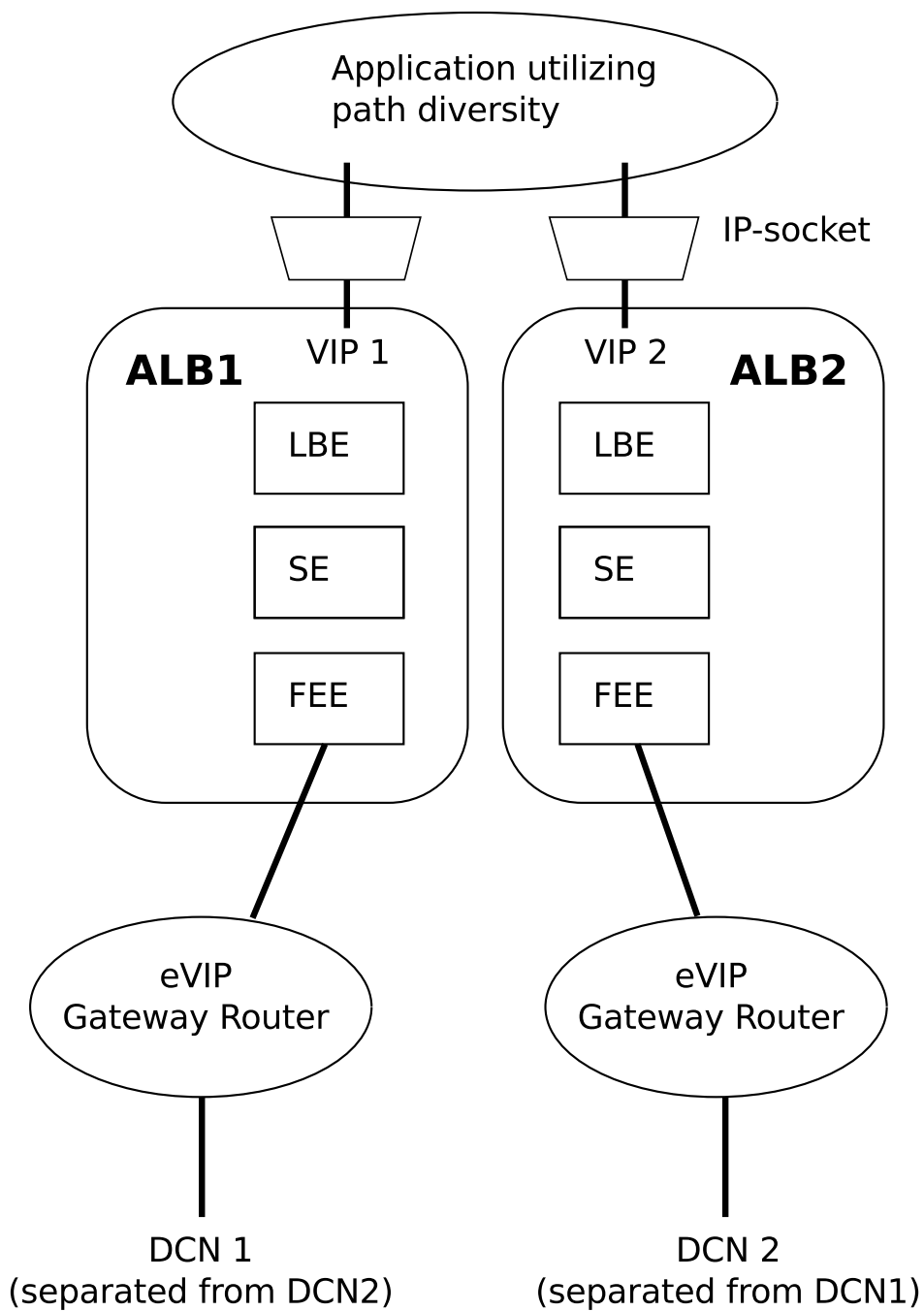


Figure 9 Cluster Internal Path Bisection

7.5 Deployment Considerations in Virtualized Environments

7.5.1 Affinity Regarding Hardware

To secure end-to-end path diversity without shared fate, it must be ensured that the eVIP elements belonging to different ALBs are not collocated on the same physical hardware. In a bare-metal installation this can easily be secured through eVIP configuration, by having the different element pin-pointed to different physical blades in the cluster.

In some virtualized environments, such as an auto-scaling cloud environment, it can be difficult to secure path diversity because this new paradigm has not yet resolved some of its practical limitations. The Virtual Machines (VMs) provided by the environment can be collocated on the same physical hardware. This is often deemed as desirable. Normally the VM user is not even aware of the distribution of VMs. If VMs are to be spread across separated physical hardware, it requires some explicit affinity rules applied into the environment by configuration. The affinity issue becomes even more complex if the eVIP elements are configured as "floating". The floating configuration secures that VIP elements being lost because of a VM failure or a scale-in is reinstantiated on another available VM.

This means that all VMs potentially could host eVIP elements, and thus none of the VMs must be collocated if internal path diversity is to be upheld. Nevertheless, a floating configuration also provides some added degree of resiliency as traffic connectivity is retained when lost eVIP elements are reinstantiated.

Recommendation

Unless the affinity of VMs can be secured in a virtualized environment, it is recommended to use eVIP in a floating element configured mode.



8 Interworking Rules, Recommendations, and Limitations

8.1 Rules

The interworking rules for eVIP are as follows:

- | | |
|---------------|---|
| Rule 1 | eVIP gateway routers connected to traffic networks must be duplicated for redundancy. |
| Rule 2 | Each ALB must be configured with an individual name. |
| Rule 3 | If VLANs are used in networks interlinking an eVIP gateway router and FEEs, there must be a one-to-one correspondence between the VLANs and subnets. |
| Rule 4 | IPv6 routing must be configured between the FEE and eVIP gateway router when IPv6 VIP addresses are used to external networks. |
| Rule 5 | <p>VIP addresses from different non-private ALBs must not overlap, that is, the same VIP addresses configured to one ALB must not be reused for another ALB.</p> <p>Note: For outgoing traffic leaving the ALB, that is, with source addresses being VIP addresses, the destination addresses can overlap in traffic from different ALBs. For example, destination addresses to remote parties in different VPNs can overlap; sockets must then be bound to the corresponding ALB. This is done by arrangements in the application. It follows that any default route must not be configured on the Payload Nodes (PNs) where this application software is deployed.</p> |
| Rule 6 | The VIP addresses configured to the ALBs must not be part of the interlinking subnets configured in the FEEs. |
| Rule 7 | eVIP gateway routers must not use OSPF authentication to the cluster. Specify no authentication in the eVIP gateway routers. |
| Rule 8 | Priority 0 (zero) must be configured to the OSPF instances of the FEE routers of the cluster. |



- Rule 9** eVIP gateway routers must not use a priority value equal to 0 (zero). Specify a positive integer greater than 0 (zero).
- Rule 10** The OSPF router ID in an eVIP gateway router must not be equal to the IP address of any internal interface in the cluster.
- Rule 11** The timer values for the hello_interval and dead_interval must be the same in the eVIP gateway router and the FEE router instance.
- Rule 12** The OSPF hello_interval must be set to a value greater than, or equal to, 3 seconds. If the interval is set to less than 3 seconds, false alarms can occur. Defaults to 10 seconds.
- Rule 13** The router dead_interval must be set to a value greater than twice the hello_interval. Recommended value for dead_interval is three times the hello_interval. Defaults to 40 seconds.
- Rule 14** The type of OSPF routing area type between the cluster and eVIP gateway routers must be configured to be either a stub area or NSSA.
- Rule 15** In a stub area, all routers belonging to the stub area must be configured so that the OSPF area_type parameter in each router is set to stub. This includes FEE routers and gateway routers.
- Rule 16** In an NSSA area, all routers belonging to the NSSA area must be configured so that the OSPF area_type parameter in each router is set to nssa. This includes FEE routers and gateway routers.
- Rule 17** If geographical redundancy is implemented in an application product by eVIP functions, and this application product is deployed in a geographical redundant configuration, then the OSPF protocol must always be used between corresponding FEEs and the gateway routers.

8.2 Recommendations

The interworking recommendations for eVIP are as follows:

**Recommendation 1**

In general, use OSPF between FEEs and eVIP gateway routers, because of the following:

- The protocol prevents accidental configuration of routing loops.
- The protocol provides dynamic management of VIP addresses.
- The protocol is inherently a stable robust protocol.

BFD can further be used with OSPF for rapid link failure detection.

However, static routing with BFD (in both directions between FEEs and an eVIP gateway router, without OSPF) is a permissible configuration if the requirements on network design do not require dynamic routing end-to-end. In these scenarios, it is assumed that the network engineering of the network beyond the eVIP gateway router is robust enough, so it can avoid situations of undesirable black-holing of packets.

Recommendation 2

In general, for the sake of simplicity, it is recommended to use OSPF area_type equal to stub as a default choice. However, use nssa when an identified networking case motivates the choice of NSSA, or if the networking plan for the targeted deployment is already prepared to accommodate NSSA.

Recommendation 3

The BFD parameters and the recommended settings are shown in Table 2.

Recommendation 4

Utilizing the FEE floating scheme, see Section 2.2.1 on page 6, can disrupt dynamic routing protocols such as OSPF. The relocation of an FEE is seen as a routing instance disappearing for a short period and thus triggering routing updates to be sent by the peers. Such updates can spread across the entire external network causing route-flapping. + It is therefore recommended that in network scenarios where a set of FEEs are required to use OSPF, this particular set of FEEs is configured as "fixed element" FEEs.

Recommendation 5

All FEEs under the same ALB must have the same routing protocol (for example, ospfv2/ospfv3, bfd_static/bfd_static6).

Recommendation 6

To expect carrier grade availability for systems deployed with the stackable resilient FEE IP address feature, the following conditions must be met:

- The deployment infrastructure (virtual or bare metal) that eVIP is deployed upon must provide automatic end-to-end protection against any bi-directional or uni-directional fault in the intermediary Layer 2 connectivity path between FEEs and gateway router.
- The highly available gateway IP and the external IP addresses of all FEEs under the same ALB must all be in the same subnet.
- The external gateway router function provided by the deployment infrastructure must support static routing with ECMP routing towards the Virtualized Network Function (VNF) (FEEs).
- A static route must be provisioned on the external gateway router for each next hop, which corresponds to an FEE IP address belonging to an ALB, per VIP address on the ALB in question.
- The FEEs under a given ALB must be configured with a static route, which uses a highly available IP address, provided by the deployment infrastructure, as the gateway address.
- For a cloud environment where port security is enabled, it is assumed that the range of IP addresses assigned to the FEEs for external network connectivity are permitted on all VNICs.
- If eVIP is configured on fixed nodes and IP stacking is required on these nodes, the ALB option to stack on fixed nodes must be enabled.

Recommendation 7

Static routing without BFD supervision is only to be used in the deployment scenario mentioned in Recommendation 6 and possibly also in conjunction with SCTP, which is a protocol that detects and avoids unreachable paths.



Table 2

BFD Parameter	Recommended BFD Setting
echo	off
bfd_interval	200 milliseconds
minrx	200 milliseconds
multiplier	5

8.3 Limitations

The interworking limitations for eVIP are as follows:

- A system limit of maximum 8 ALBs can be configured in a cluster.
- The number of internetworking links an eVIP system can support is limited by the number of FEEs that can be provisioned. For information on the number of supported FEEs, refer to Section Constraints in *eVIP System Architecture Description*.





9 eVIP Gateway Router to DCN

From the eVIP gateway router to the DCNs, the interworking options are many and depend on the network design and the capabilities of the router used as eVIP gateway router. It is a wide field beyond the scope of this document to be covered in detail. Typically the following methods are used:

- BGP routing

BGP routing requires that eVIP addresses (routes) are redistributed from the OSPF stub area or NSSA into the BGP. Access Lists (ACLs) or "route maps" are used to filter out routes that are not to be seen in the BGP autonomous system, for example, the interlinking networks between an eVIP gateway router and the cluster.

- OSPF routing

OSPF routing in the external network can be done either over the area 0, the backbone area, or, if the router used as eVIP gateway router is capable thereof, redistribute eVIP addresses from the stub area or NSSA to a non-backbone area of a separate OSPF autonomous system. Redistribution must only be done in one direction, that is, from the stub area or NSSA to the external autonomous system.

- Static routing

Static routing with alternative routes configured with different metrics or ECMPs used.

- In general (this is beyond the scope of eVIP), regardless of the routing protocols or configuration, it is recommended that BFD is used to supervise connections between nodes in the network end-to-end. This increases the robustness of the overall network and provides faster fault detection.





10 Examples of Networks with L2 Switches

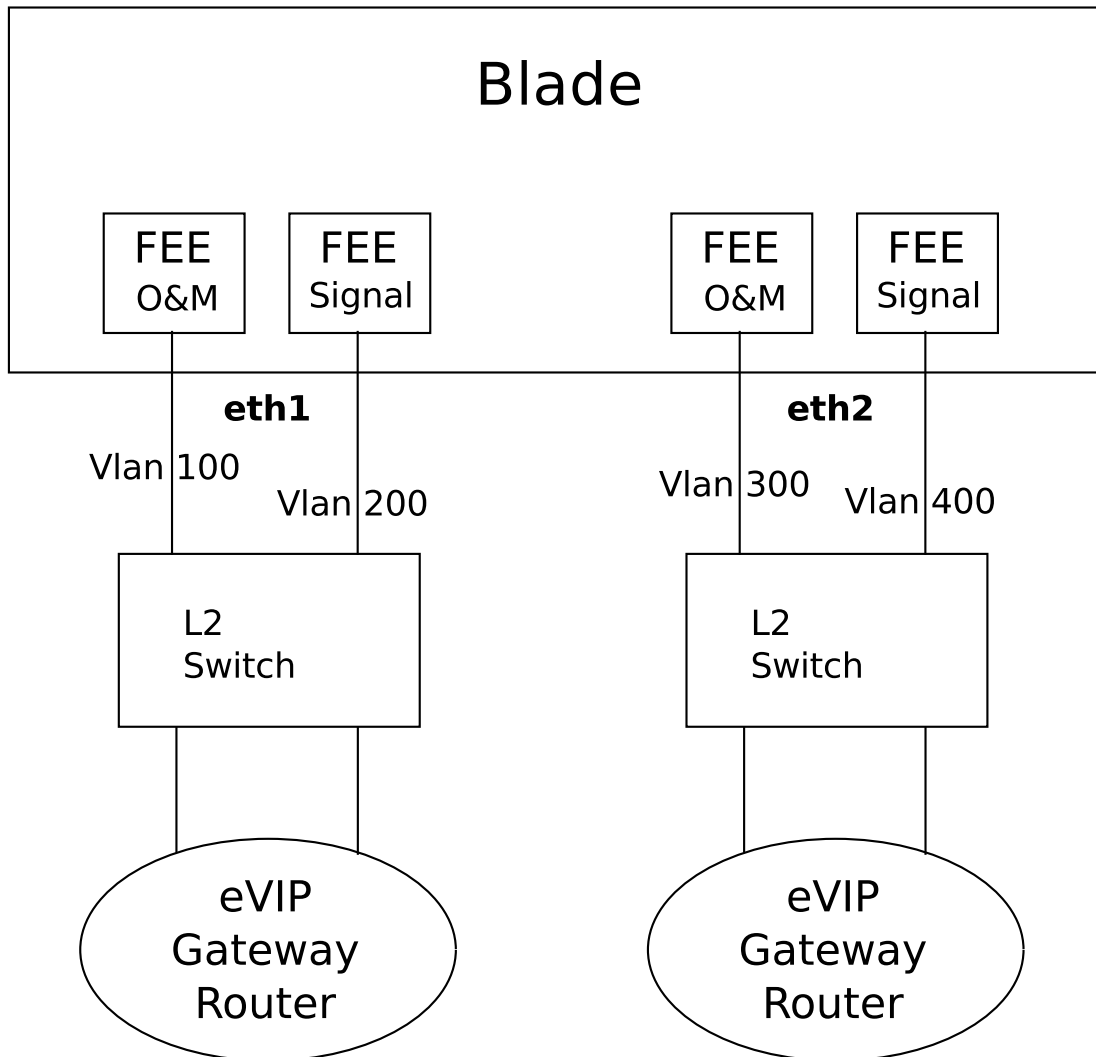


Figure 10 Configuration with L2 Switches, Four VLANs, and Two Subnets

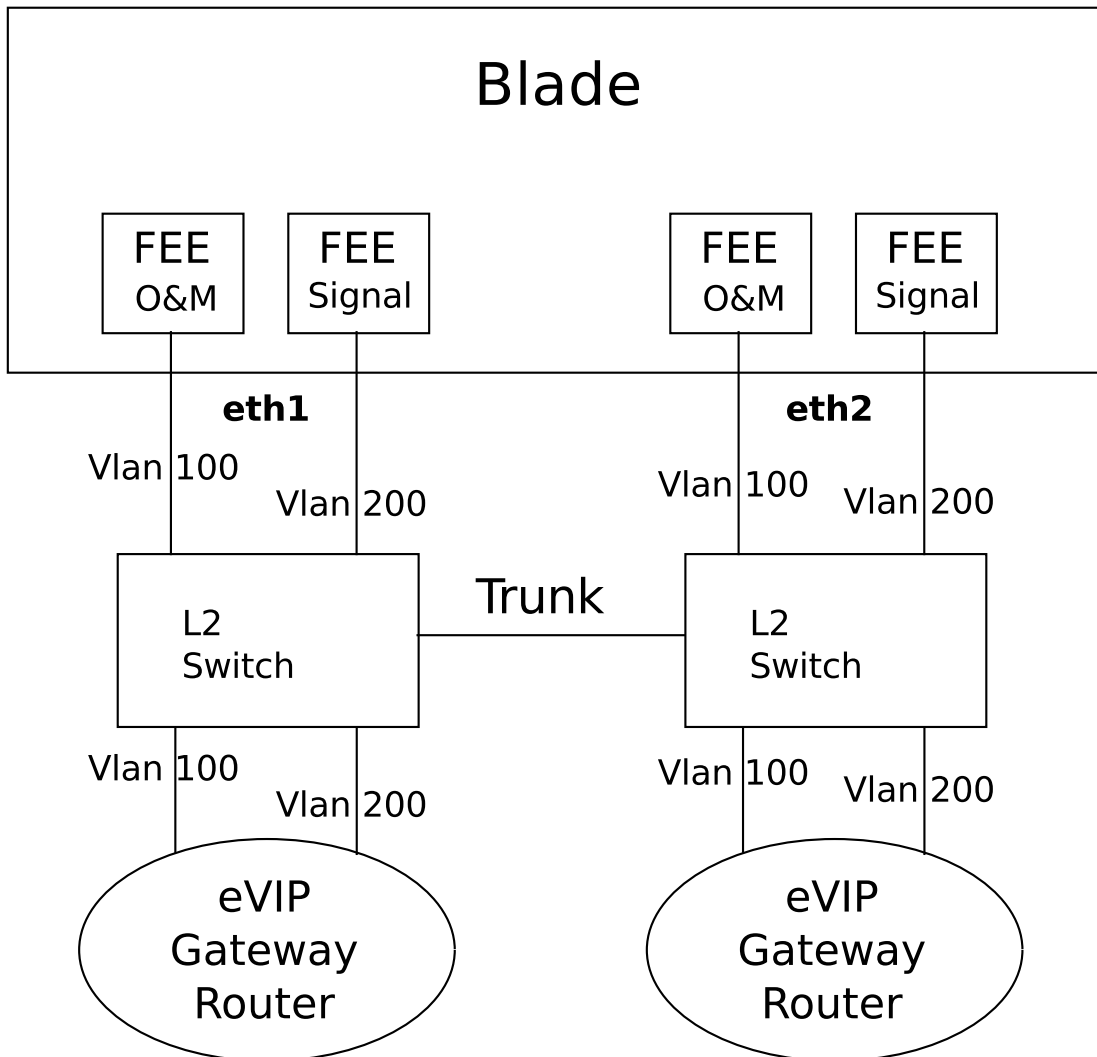


Figure 11 Configuration with L2 Switches, Two VLANs, and Two Subnets

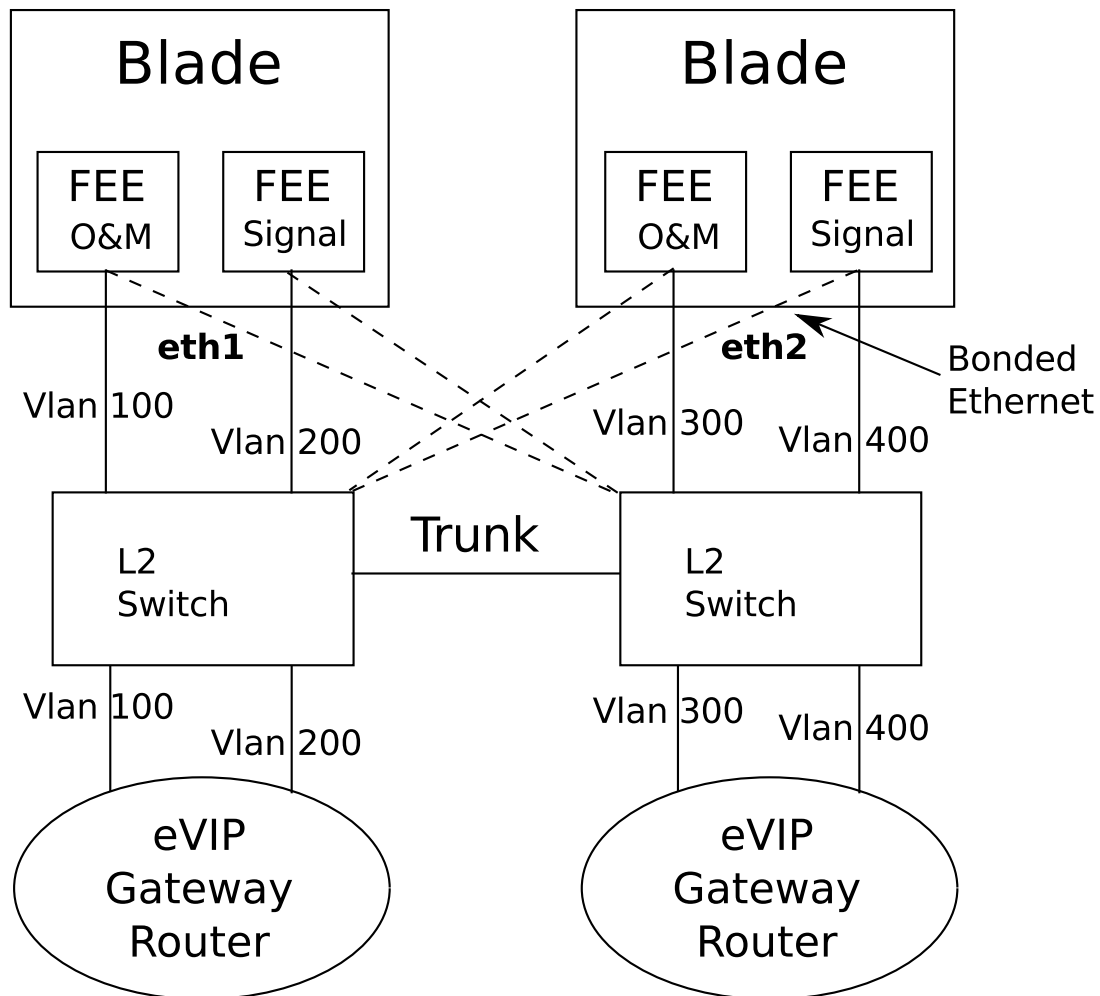


Figure 12 Configuration with Bonded Ethernet External Interfaces

Note: An FEE can only have a single external Layer 3 interface. However, a Layer 3 interface can have more than one underlying Layer 2 interface, for example, bonded Layer 2 interfaces.