Hewlett Packard Enterprise

HPE Cloud-First Reference Architecture Guide – 2000 Servers

Enabling the hybrid infrastructure

Contents

Introduction 2
Value proposition 2
Objectives 2
Blade server 1-Tier design guidelines 3
  Architecture 3
Blade server 1-tier design using HPE FlexFabric 12908E Series Switches 4
  Key Features 6
  Specifications 7
Rack server spine and leaf designs 8
  Architecture 9
Spine and leaf design using HPE FlexFabric 12908E spines and 5930 leaf switches 11
  Key Features 12
  Specifications 14
Legacy 3-Tier network design 15
L3 routed deployments 16
Glossary 18
Additional links 21

Architecture Guide

Introduction

We live and work in the Idea Economy. It has never been easier to turn ideas into new products, services, applications, and industries. To compete in this Idea Economy todays Enterprises need to be able to create and deliver new value instantly and continuously from all of their applications. This requires a hybrid infrastructure that is able to maximize performance and cost.

Businesses must change along four axes in order to survive and thrive in the Idea Economy. They must transform to a hybrid infrastructure; protect the digital enterprise; empower the data driven organization; and enable workplace productivity.

This Architecture Guide provides a reference architecture design guideline for a 2,000 physical server data center that can take advantage of a Hybrid Infrastructure that can allow them to compete in this Idea Economy. This guide is intended for technology decision-makers, solution architects and other experts tasked with improving data center networking. It can serve as a baseline for network planning and design projects.

Value proposition

This reference architecture focuses on optimizing a ground up data center deployment which consists of 2,000 physical servers supporting 30 to 50 VMs on each server. These architectures include critical features such as agility and flexibility, high bandwidth, high availability, resiliency, exceptional performance, and the ability to scale well past 2,000 servers.

Objectives

Whether you are building a 1-tier solution or a spine and leaf solution, the objectives in each case can be summed up as follows:

- Public, private, hybrid cloud support:
  - It is critical when building out a new Hybrid Infrastructure data center that the infrastructure chosen provides developers with the on demand infrastructure they require to go from idea to revenue.
  - Enable DevOps orchestration infrastructure which can accelerate IT operations, app delivery, and quality
- Create flexibility:
  - Open and standard-based systems and environments reduce risk and increase flexibility for your applications, infrastructure, and data needs.
- Flatter network with increased frame forwarding and packet forwarding:
  - Traditional legacy 3-tier data centers were not able to keep providing the performance needed in the modern virtualized data centers that now see immense amounts of east-west traffic flows. Data center networks can see great benefits from eliminating these unneeded hops needed to get from rack to rack.
  - 1-tier networks are able to truly optimize the traffic flows between racks by ensuring each rack can reach another rack with one hop.
  - 2-tier spine and leaf deployments have become one of the most common data center network architectures used by the industry today. These types or designs offer greater flexibility and scalability than 1-tier solutions while still providing for exceptional performance, predictably, efficiency, and resiliency with a maximum of 2 hops from rack to rack.
  - HPE IRF can be used as a switch virtualization feature, helping to create larger scalable devices which provide many benefits including adding redundancy but also allowing extra layers like aggregation layers to be eliminated, helping to reduce the number of hops and increasing performance.
- L2 vMotion flexibility:
  - L2 connectivity always needs to be considered. Data centers can deploy a single L2 domain, leveraging TRILL or SPB, or by using L3 with VXLAN overlay technology, or vSphere 6.0.

- Reduced management complexity:
  - Flattening the network with large core switches and the elimination of aggregations switches combined with leveraging IRF in the various layers of the network simplifies what was typically a complex management scheme.
- Zero need for STP/RSTP:
  - HPE IRF can eliminate the need for STP, by presenting many physical switches as a single logical switch. This allows active/active connections between the layers, instead of relying of a loop prevention technology like STP, which results in better performance and faster fail-over.

Blade server 1-Tier design guidelines

When looking at flattening a network and providing substantial support for virtualization and east-west traffic flows, blade server 1-tier topologies provide many advantages. Server blade enclosures allow for substantial compute density per rack, row, and HPE has optimized the HPE C-Class BladeSystem Server portfolio to support the vision and reality of virtualization. This type of blade server 1-tier network topology optimizes and combines the reality of high-performance networking with simplicity. It allows flexibility in networking by supporting Intelligent Resilient Fabric (IRF) and a variety of interconnect options, including HPE Comware-based HP 6127/6125XLG Switches as well as HPE Virtual Connect (VC) modules with server optimized features such as server profiles.

Architecture

Typical blade server 1-tier deployments will usually consist of two to four core switches utilizing IRF, which then connect to blade-servers housed in C-Class enclosures. These types of deployments are able to extend VLANs across the entire data center and are optimized for virtualized environments where VMs may be moving from rack to rack or data center to data center.

Uplink speeds in data centers has moved to 40GbE and even 100GbE. These faster uplink speeds are providing great benefits to data centers with regards to performance and scalability. However, when moving to 40/100GbE fibre uplinks a customer needs to rethink the cabling infrastructure. Current multi-mode optic standards for 40GbE and 100GbE use multiple 10Gbps lasers or multiple 25Gbps lasers simultaneously transmitting across multiple fiber strands to achieve the high data rates. Because of the multi-lane nature of these optics, they use a different style of fiber cabling, known as MPO or MTP cabling. For 40GbE, we can use 8 fibers or 12 fibers MPO/MTP cables. These MPO/MTP type cabling solutions can be used in existing data centers that are making incremental upgrades to the uplinks. However, upgrading an entire cabling plant can be cost prohibitive.

A preferred option for existing data centers is to leverage 40GbE BiDi transceivers which allows the use of the same two MMF fiber strands with duplex LC connectors, currently used by 10GbE connections, to be used by the new 40GbE connections. In many cases the 40GbE BiDi optics are less costly than MPO/MTP optics and cabling combined. Using 40GbE BiDi optics in the existing data center means that migrating from 10GbE to 40GbE will be a smooth, cost-effective transition.

When building out new data centers from the ground up, however, customers will need to consider building out this new data center using single mode fibre rather than multi-mode fibre. It has been standard practice to build a data center using MMF which has, until recently, been appealing because lower cost solutions which are able to meet most distance requirements. However, new data centers are starting to see that moving to SMF will provide many benefits, including cost and performance. SMF solutions are able to leverage dual strands of fibre for 10/40/100GbE connections, but they will also see advantages with:

- Network Taps: Many customers will tap optical fibers which means that there is one tap (2 splitters for 2 fiber duplex communications) per connection. For new data centers that leverage SMF 40/100GE, this does not change. However, if this new data center was built using MMF, then for 40/100GbE a customer must plan for an entirely new cabling infrastructure because there are 8-20 fibers per connection, which means a single bidirectional, passive optical tap will require 8-20 splitters. In addition, a 40/100GbE data center will also have to ensure that all of the patch panels are MTP connectorized patch panels to terminate the associated 8-20 fiber cables, further increasing costs.
- Attenuation: Due to the higher bandwidth more attenuation is incurred resulting in smaller optical power budgets to accommodate splitter, splice, and fiber losses. With 40/100GE loss budgets around 1.5dB, it is difficult to insert passive optics to monitor these links. If SMF is used the

link loss budget is substantially higher (distance dependent of course), so tapping these optical signals for monitoring is virtually the same as it is for 10GE today.

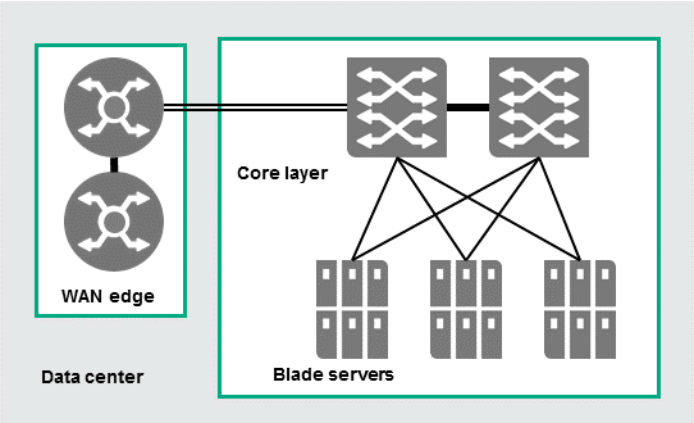

Figure 1. Blade server 1-tier architecture view

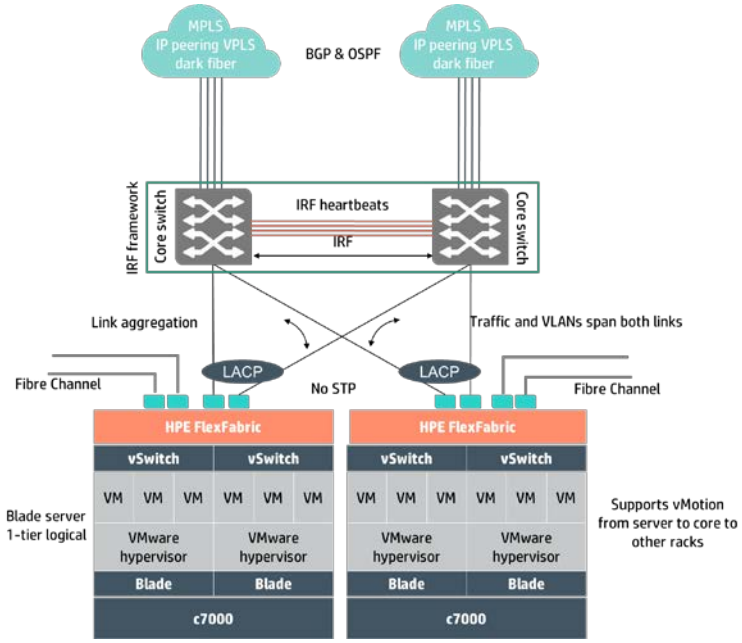

Figure 2. Blade server 1-tier logical view

Blade server 1-tier design using HPE FlexFabric 12908E Series Switches

The solution shown below consists of two HPE FlexFabric 12908E Switches at the core. These switches are IRF'd together to provide a highly resilient core that allows Link Aggregation Groups (LAGs) from the HP c7000 BladeSystem enclosures to different modules in different chassis. The HPE FlexFabric 12908E solution shown offers the throughput and buffering required to scale to 2,000 servers, and beyond. It also provides added redundancy with dual management modules and redundant fabrics within each chassis.

The most common blade enclosure scenario involves using the HPE ProLiant BL460c Gen8/9 half height server blades with 10GbE or 20GbE sever NICs. This configuration allows for 16 physical servers per enclosure bringing the total number of enclosures, in this example, to 125 enclosures for a total of 2000 physical blade servers.

In this deployment the core HPE FF 12908Es will act as the gateways for the servers in the enclosures.

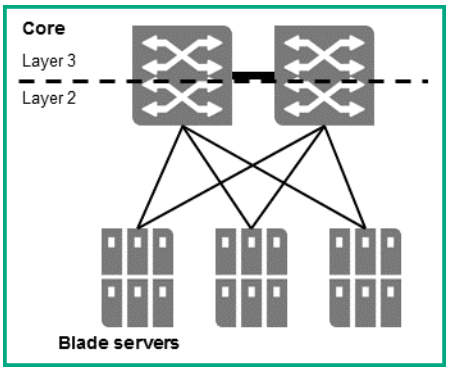

Figure 3. L3 boundary

A typical 2000 Blade Server deployment using 10GbE active/passive server NIC connections may require 20,000 Gb of uplink bandwidth (160 Gb from each enclosure). If using 10GbE active/active server NIC configurations, this solution would require 40,000 Gb of uplink bandwidth (320 Gb from each enclosure).

If using 20GbE server NICs, this 2000 Blade Server deployment using active/passive server NIC connections would require 40,000 Gb of uplink bandwidth (320 Gb from each enclosure). If using 20GbE active/active server NIC configurations, this same solution would require 80,000 Gb of uplink bandwidth (640 Gb from each enclosure).

The configuration below utilizes 125 HPE c7000 BladeSystem enclosures each with 16 blade servers for a total of 2,000 physical servers. Connectivity from the HPE c7000 BladeSystem would use either Comware based HPE 6125/6127XLG interconnect modules or HPE Virtual Connect modules. These modules provide 10GbE and 40GbE uplink flexibility to increase bandwidth if needed.

Deployments vary depending on oversubscription requirements, however, this specific soluiton provides for a 4:1 oversubscription ratio when using 10GbE server NICs in an active/active fashion. In this scenario, the core HPE FlexFabric 12908E switches are equipped with 250 40GbE ports (125 x 40GbE in each 12908E) for server enclosure connections (2 from each enclosure). Oversubscription ratios could be reduced by increasing the number of 40GbE ports to each enclosure. A similar HPE FlexFabric 12908E solution using the same oversubscription ratios can scale to up to 4,608 physical blade servers, and an HPE FlexFabric 12916 solution could scale to 9,216 physical blade servers.

This solution utilizes 40GbE SMF LR4 optics to connect the HPE FF 12908E switches to the uplink ports on the c7000 enclosures. This SMF solution future proofs the data center so that it is ready for easy migration from 40GbE to 100GbE.

Power, cooling, and weight are the three major considerations when positioning equipment in racks. A fully-populated HPE c7000 BladeSystem enclosure can weigh 480 pounds so while four 10RU enclosures can technically fit in a single rack, we typically see two enclosures per rack.

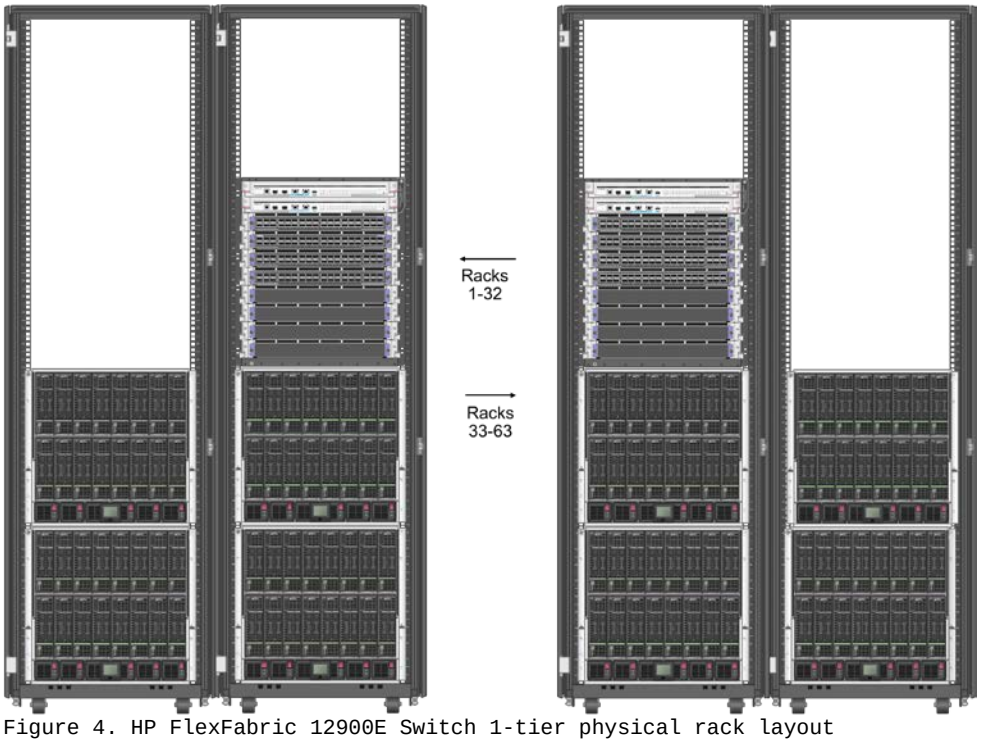

Figure 4. HP FlexFabric 12900E Switch 1-tier physical rack layout

Key Features

- VXLAN L2/L3 support
  - While scalability and multi-tenancy isn't necessarily a concern with a 100 server solution, being able to adapt your network to the latest networking protocols extends the initial deployment of networking hardware.
  - Enables L2 connectivity between rack/servers even when using a modern L3 infrastructure.
  - Hybrid cloud environments can take utilize VXLAN to support the on-demand networking requirements so critical for success.
- Chassis based IRF
  - IRF allows the two HPE FlexFabric 12900E switches to look like a single logical switch. This allows active/active connections from the server blade enclosures as well as to an upstream device. IRF fail-over is measured in the millisecond range so if a failure does occur, the network will fail-over seamlessly.
  - Combining LACP and IRF can provide high-speed link aggregation with re-convergence times at sub-50 ms in the event of a link failure. It also allows links to be aggregated and utilized for higher bandwidth from the converged network adapters across all switches to forward traffic.
- DC performance and feature set
  - The HPE FlexFabric 12908E delivers unprecedented levels of performance, buffering, scale, low latency, and availability with high density 10GbE, 40GbE and 100GbE connectivity. It also has a mature DC feature set which includes support for TRILL, SPB, Data Center Bridging (DCB)/Fibre Channel over Ethernet (FCoE), Virtual Ethernet Port Aggregator (VEPA), HPE Ethernet Virtual Interconnect (EVI), Multitenant Device (MDC), OSPF and BGP for IPv4 and IPv6, Multiprotocol Label Switching (MPLS), and L2 and L3 VXLAN support for demanding cloud overlay solutions.
- 10GbE, 40GbE, or 100GbE
  - The HPE FlexFabric 12908E supports up to 384 line rate 10GbE ports, 288 line rate 40GbE ports, or 64 line rate 100GbE ports. The 16 slot HPE FlexFabric 12916E supports up to 768 line rate 10GbE ports, 576 line rate 40GbE ports, or 128 line rate 100GbE ports. Customers can have confidence that if they connect the blades at 10GbE today they already have a 40GbE/100GbE strategy for the future.

- Modular components
  - The HPE FlexFabric 12900E is a modular chassis that, as mentioned above, can provide the necessary bandwidth for today as well as future scaling needs. As an example, a customer could buy the 40GbE modules today and using direct attach cables (DAC) split each port into 10GbE ports. When the need for 40GbE arrives the existing cable just needs to be swapped for either a 40GbE optic or 40GbE DAC. It also supports redundant and hot-swappable power supplies, fabrics, management modules, and fan trays.
  - The HPE c7000 BladeSystem has the same modularity as the HPE FlexFabric 12900E switch series. The server blades can be swapped as newer processing becomes available. The HP 6125/6127XLG blade interconnect and HP Virtual Connect FlexFabric-20/40 both support 40GbE uplinks as well. The agility for a company to increase bandwidth or the agility to redeploy a network is unparalleled.
- Optimized for VMs
  - 1-tier designs such as this eliminate unneeded hops and provide good foundations for L2 domains ideal for VMs.
- FC/FCoE/iSCSI support
  - The VC blade interconnect solution provides support for native Fibre Channel where FC ports can be directly connected to an existing SAN or an HPE 3PAR device. Additionally, the VC Modules, HPE 6127/6125XLG, and HPE FlexFabric 12900E switch series support DCB for end to end FCoE/iSCSI deployments.
- VXLAN L2/L3 support
  - VXLAN L2/L3 features, supported by the HPE FF 12900E switch provides the L2 extensions, scalability, and multi-tenancy that is needed in large data centers.
- Orchestration support including Puppet, Chef, and Ansible
  - Network automation tools like Chef, Puppet, and Ansible are able to provide the controls necessary to manage and perform everyday management functions, which might range from just device discovery all the way to complete network configuration management and the provisioning of virtual network resources. Ansible is an agentless tool which communicates with the CLI interface over a Secure Shell, while Puppet and Chef rely on agents installed on the device.
- Management
  - This network configuration supports the full feature complement of HPE Intelligent Management Center (IMC) from the core switches down to the rack. This provides a single-pane-of-glass management platform to support the servers, VMs, virtual switches, IRF frameworks, and IP routing.
  - HPE IMC VAN Fabric Manager (VFM) Software is an IMC module which simplifies the management of data center and Fibre Channel Storage Area Networks (SAN) fabrics. It provides a unified view of all of the network and storage devices in the data center fabric alongside fabric health to enable quick troubleshooting and pro-active management. VFM allows you to easily configure SPB or TRILL (Transparent Interconnect of Lots of Links) through the same graphical user interface used to automate, monitor and manage your entire network.

Specifications

Core switch

The following devices are recommended as the core switch in the above design scenario.

- HPE FlexFabric 12900E Switch Series

BOM

Here is a sample 12908E BOM which can be used in this architecture.

Table 1. Core switch example BOM

PART NUMBER	DESCRIPTION	QUANTITY	COMMENTS
JH255A	HP FF 12908E Switch Chassis	2	
JH108A	HP FF 12900E 2400W AC PSU	16	

JH258A	HP FF 12908E Fan Tray Assembly	4	
JH107A	HP FF 12900E LPU Adapter	8	
JC665A	HP X421 Chassis Universal Rck Mntg Kit	2	
JH257A	HP FF 12908E 5.0Tbps Type F Fabric Module	12	
JH104A	HP FF 12900E Main Processing Unit	4	
JH045A	HP FF 12900 36p 40GbE QSFP+ FX Module	8	
JG328A	HP X240 40G QSFP+ QSFP+ 5m DAC Cable	3	Used for IRF & BFD MAD connectivity
JG661A	HP X140 40G QSFP+ LC LR4 SM Transceiver	250	Used for server connectivity, add more for WAN edge connectivity

Core switch scalability table

This is the scalability table for 12908E discussed in this solution:

Table 2. Core switch scalability

FEATURE	12908E FX MODULE
MAC Address	Up to 256K *
ARP (host)	16K/128K (Uni mode) *
Link aggregation: ports/group	64/1024
IPV4 LPM	32K
IPV6 LPM	8K **
IPv4/IPv6 MC	8K/1K **
Ingress/Egress ACL	18K/9K **
VXLAN	Yes
Native FC	No
DCB	Yes
SPB	Yes

* Shared Resource - up to 256K table entries

** Shared Resource of 24K

Rack server spine and leaf designs

Similar to the 1-tier design, spine and leaf designs, commonly referred to as 2-tier architectures, can provide a balanced network that can be optimized for virtualization while also providing for great scale and flexibility so the data center can adapt to changing application requirements. Spine and leaf topologies are also used when the physical cabling plant cannot support a 1-tier design.

Although these topologies may also utilize rack servers, enhancements to the HPE C-Class BladeSystem server portfolio allow these types of topologies to combine the reality of high-performance networking with simplicity. The enclosures allow flexibility in networking supporting a variety of interconnect options, including Comware based HPE 6125/6127XLG Switches with IRF solutions as well as HPE Virtual Connect modules with server optimized features such as server profiles.

Key benefits of this design:

- The ability to use 10G Base-T copper design or DACs due to the shorter distance
- Allows for greater scale and flexibility
- Being able to provide ToR redundancy and resiliency using IRF
- Issue isolation ensures that each ToR/IRF group can be treated as an independent module in which issues and outages are isolated to that specific rack
- Lays the framework for implementing L2 spine and Leaf fabric technologies as well as implementing L3 from the ToR

Architecture

Many spine and leaf deployments will utilize two to four spine (core) switches, which then connect to various leaf switches (ToR). These leaf switches, which may or may not be utilizing IRF, will then connect to servers usually within the same rack. Leaf switches can also be blade switches which are installed directly into HPE C-Class BladeSystem enclosures.

Large scale L2 spine and leaf topologies can be built using TRILL or SPB as the technology used to extend L2 networks across the backbone data center network. Devices running SPB or TRILL can still leverage IRF. However, large scale spine and leaf solutions usually use L3 solutions (either BGP or OSPF) coupled with overlay solutions like VXLAN to enable L2 connectivity between leafs.

Spine and leaf solutions are easily able to scale well past the common dual spine solution. In these solutions that use more than two spine switches the WAN edge is usually connected to a pair of leaf switches, known as border leafs. This allows for dual redundant conections to the WAN edge without requiring the WAN edge to be directly connected to each and every spine switch.

The physical connections between the spine and leaf in these topologies will be either LAGs or routed links of 10GbE, 40GbE, and even 100GbE uplinks, providing for very high performance that can provide line rate performance between the spine and leaf.

Today's 40GbE and 100GbE uplink speeds are providing a great benefit to data centers with regards to performance and scalability. However, when moving to 40/100GbE fibre uplinks a customer needs to rethink the cabling infrastructure. Current multi-mode optic standards for 40GbE and 100GbE use multiple 10Gbps lasers or multiple 25Gbps lasers simultaneously transmitting across multiple fiber strands to achieve the high data rates. Because of the multi-lane nature of these optics, they use a different style of fiber cabling, known as MPO or MTP cabling. For 40GbE, we can use 8 fibers or 12 fibers MPO/MTP cables. These MPO/MTP type cabling solutions can be used in existing data centers that are making incremental upgrades to the uplinks. However, upgrading an entire cabling plant can be cost prohibitive.

A preferred option for these existing data centers is to leverage 40GbE BiDi transceivers which allows the use of the same two MMF fiber strands with duplex LC connectors, currently used by 10GbE connections, to be used by the new 40GbE connections. In many cases the 40GbE BiDi optics are less costly than MPO/MTP optics and cabling combined. Using 40GbE BiDi optics in the existing data center means that migrating from 10GbE to 40GbE will be a smooth, cost-effective transition.

When building out new data centers from the ground up, however, customers will need to consider building out this new data center using single mode fibre rather than multi-mode fibre. It has been standard practice to build a data center using MMF which has, until recently, been appealing because lower cost solutions which are able to meet most distance requirements. However, new data centers are starting to see that moving to SMF will provide many benefits, including cost and performance. SMF solutions are able to leverage dual strands of fibre for 10/40/100GbE connections, but they will also see advantages with:

- Network Taps: Many customers will tap optical fibers which means that there is one tap (2 splitters for 2 fiber duplex communications) per connection. For new data centers that leverage SMF 40/100GE, this does not change. However, if this new data center was built using MMF, then for 40/100GbE a customer must plan for an entirely new cabling infrastructure because there are 8-20 fibers per connection, which means a single bidirectional, passive optical tap will require 8-20 splitters. In addition, a 40/100GbE data center will also have to ensure that all of the patch panels are MTP connectorized patch panels to terminate the associated 8-20 fiber cables, further increasing costs.
- Attenuation: Due to the higher bandwidth more attenuation is incurred resulting in smaller optical power budgets to accommodate splitter, splice, and fiber losses. With 40/100GE loss budgets around 1.5dB, it is difficult to insert passive optics to monitor these links. If SMF is used the

link loss budget is substantially higher (distance dependent of course), so tapping these optical signals for monitoring is virtually the same as it is for 10GE today.

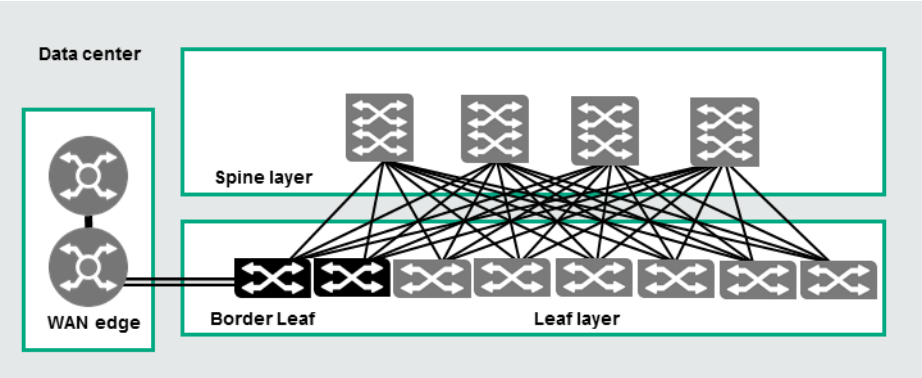

Figure 5. Typical spine/leaf (with border leaf) topology view

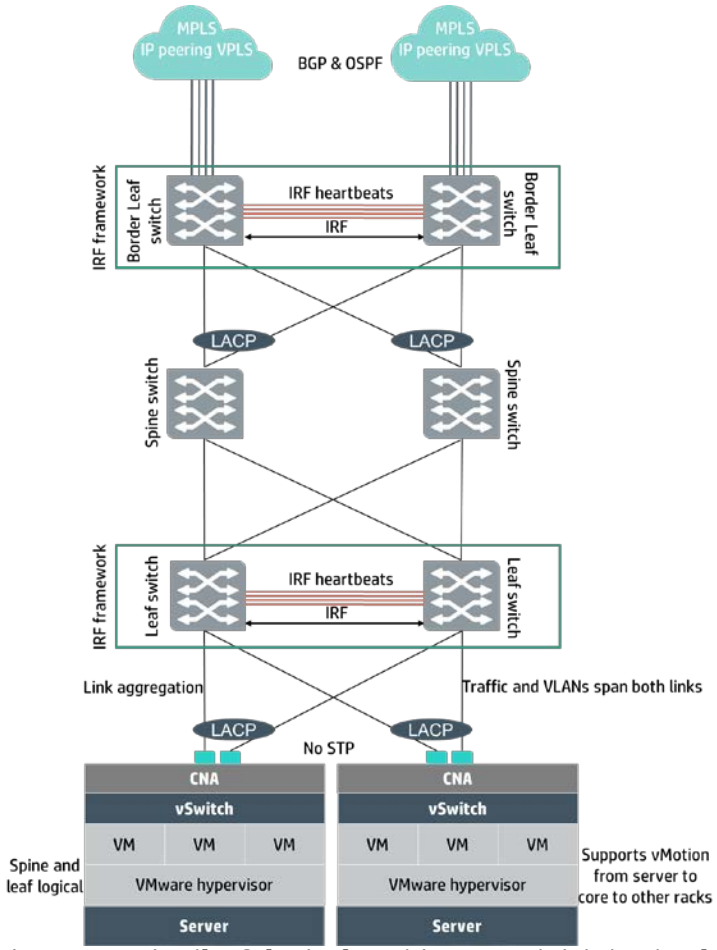

Figure 6. Spine/leaf logical architecture (with border leafs)

Spine and leaf design using HPE FlexFabric 12908E spines and 5930 leaf switches

The solution below utilizes HPE FlexFabric 12908E spine switches, which are better suited for larger deployments requiring more buffering and larger table scalability. The solution also utilizes the modular HPE FlexFabric 5930 leaf switches, which are able to provide 10GbE and 40GbE interface options. The HPE FF ToR portfolio includes the 5900, 5930, 5940, and 5950 switch series. All could be used in this example depending on the scalability and feature set required.

This example solution consists of two HPE FlexFabric 12908E spine switches and one-hundred HPE FlexFabric 5930 leaf switches. The HPE FlexFabric 12908E spine switches connect to all leaf switches. This configuration positions 20 rack servers in 100 separate racks for a total of 2,000 rack servers. Each rack has a single HPE FF 5930 ToR leaf switch and the racks are paired up so that a pair of racks would have the HPE FlexFabric 5930 ToR switches combined using IRF, which in turn connect to each of the 40 servers in those two racks. This design provides for added resiliency and bandwidth to each server.

The HPE FF 5930s used are the 2-slot modular versions, and each leaf switch has been configured with 48 1/10GbE Base-T ports for connections to the servers using Cat6a cables. 10G Base-T support on rack servers is gaining in popularity as it allows customers to reuse existing Ethernet cables. 10GBaseT works great in most solutions, however, in environments where lower latency is a primary factor then it is best to use a pure optical solution. 10GBaseT ports also does consume a bit more power per port, so that also needs to be considered.

Each of the leaf switches is also fitted with six 40GbE QSFP+ ports. Two of the QSFP+ ports will be used for IRF links between the pairs of HP FF 5930 switches, and the remaining four 40GbE QSFP+ ports are used to connect back to the spine. 40GbE DACs are used for the IRF connections. 40GbE SMF LR4 optics are used to connect the HPE FF 5930 leaf switches to the HPE FF 12908E spine switches. This SMF solution future proofs the data center so that it is ready for easy migration from 40GbE to 100GbE.

A typical 2000 server deployment using 10GbE active/passive server NIC connections may require 20,000 Gb of uplink bandwidth (400 Gb from each pair of 50 racks). If using 10GbE active/active server NIC configurations, this solution would require 40,000 Gb of uplink bandwidth (800 Gb from each pair of 50 racks).

Deployments vary depending on oversubscription requirements, however, this specific 2,000 server deployment is using 400 40GbE uplink ports (8 from each pair of ToR leaf switches) to the spine layer which provides for a 2.5:1 oversubscription ratio when using 10GbE server NICs in an active/active fashion. In this deployment each HPE FlexFabric 12908E is equipped with 200 40GbE ports for the ToR leaf switch connections.

A similar HPE FlexFabric 12908E solution using the same oversubscription ratios can scale to up to 72 pairs of ToR leaf switches and 2,880 physical servers.

An HPE FlexFabric 12916 solution using the same oversubscription ratios could scale to up to 144 pairs of ToR leaf switches and 5,760 physical servers.

In this deployment the ToR Leaf HPE FF 5930 switches which will act as the gateways for the servers. This L3 deployment scenario positions the gateways as close to the server as possible and uses OSPF as the dynamic routing protocol. L2 extension between the racks can be accomplished by leveraging VXLAN solutions.

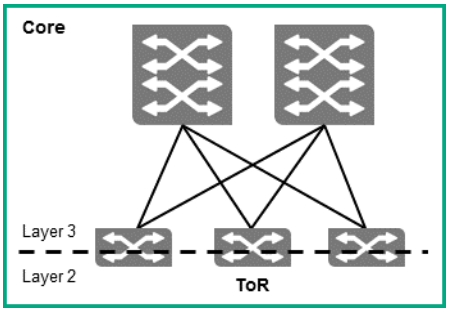

Figure 7. L3 boundary

Power, cooling, and weight should be a major consideration when filling out the racks. The diagrams below have limited the number of physical rack servers per rack to twenty, although this number will vary depending on the situation.

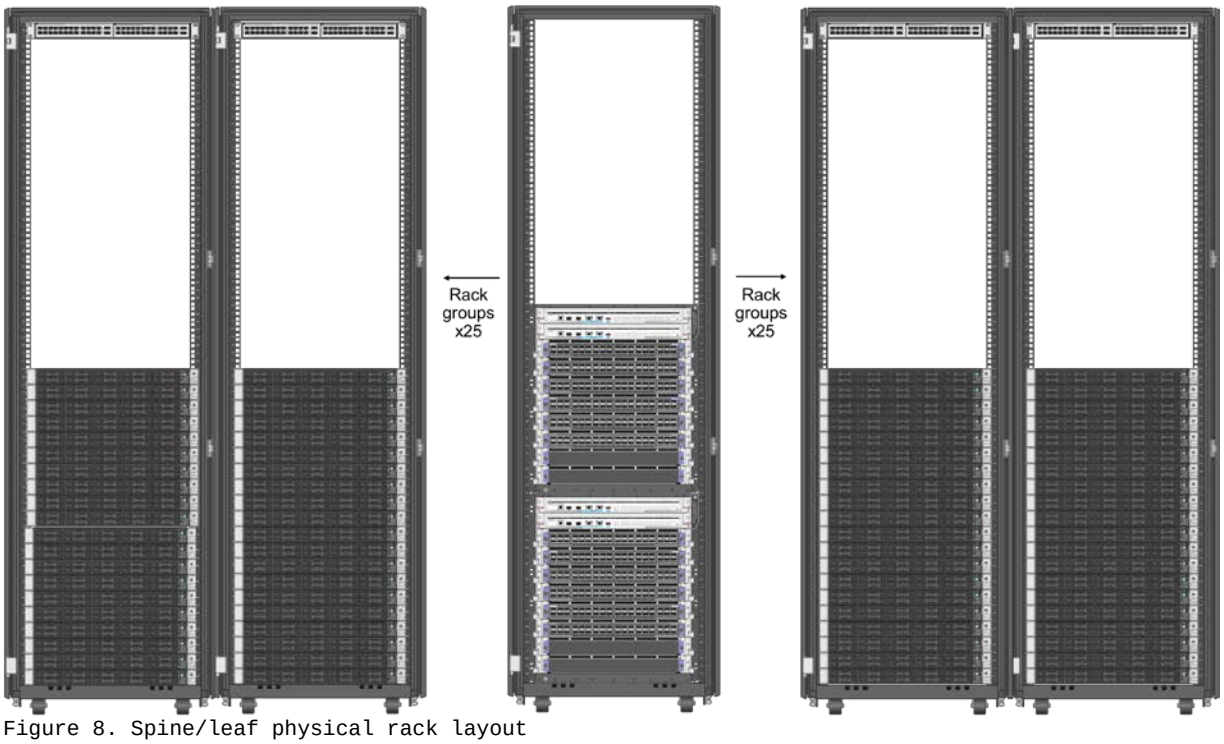

Figure 8. Spine/leaf physical rack layout

Key Features

- Exponential growth with spine and leaf architecture allows for immense growth. Adding capacity consists of adding a rack with servers and a ToR switch. When ready, connect the uplinks to the spine switch(es)
- VXLAN L2/L3 support
  - Enables L2 connectivity between rack/servers even when using a modern L3 infrastructure.
  - Hybrid cloud environments can take utilize VXLAN to support the on-demand networking requirements so critical for success.
  - HPE FF 5930 supports Open vSwitch Database Management Protocol (OVSDB) which can be integrated with the various HPE Network Virtualization solutions to provide automation and bridge virtual networks with physical networks
- IRF:
  - IRF allows multiple HPE FlexFabric Switches to look like a single logical switch. This allows active/active connections from the servers to the leaf as well as to an upstream network switch. IRF fail-over is measured in the millisecond range so if a failure does occur, the network will fail-over seamlessly.
  - Combining LACP and IRF can provide high-speed link aggregation with re-convergence times at sub-50 ms in the event of a link failure. It also allows links to be aggregated and utilized for higher bandwidth from the converged network adapters across all switches to forward traffic.
- DC feature set:
  - The HPE FlexFabric 12900E and 5930 are pure DC switch with a proven mature feature set.
  - In addition to advanced routing protocols like OSPF and BGP for IPv4 and IPv6 the switches are able to support MPLS, TRILL, MDC, EVI, FCoE, and L2/L3 VXLAN.
  - The HPE FlexFabric 12900E is also able to support added DC features including MDC and EVI.

- Large buffers and table sizes:
  - The HPE FlexFabric 12900E series switch is ideal for those customers that need large ARP/MAC/RIP/FIB table sizes and larger buffer capabilities for bursty network conditions.
- Scalability:
  - A spine and leaf design allows growth well beyond what a single chassis can support. It can also reduce the total port count in the chassis provided it meets the customer's oversubscription ratio. Adding capacity consists of adding a rack with servers and a ToR switch. When ready, connect the uplinks to the spine switch(es).
- Design flexibility:
  - The traditional design has ToR leaf using a LAG at L2 to connect to the spine. For those that are looking to simplify the management of devices, HPE's IRF technology can be used in the spine as well as in the ToR allowing for fewer configurations to exist. Using IRF on the ToR also allows active/active connections from a server to a different ToR(s). In the event that a ToR fails, all components are still operational since a LAG from the server is terminated on different switches.
- Future-proofing:
  - For customers that don't have 10GbE connections to servers, this design can start as a 1GbE ToR solution with 10GbE uplinks and migrate or co-exist with ToR deployments that might use 10GbE or 40GbE connections. Over time this solution can keep pace with the changing protocols and technologies as they can co-exist on the same hardware.
- Calculated fault zones:
  - Customers can calculate the impact on applications when a component fails. If connectivity of a ToR is lost, customers can calculate how much of the compute aspects will be lost.
- FC/FCoE support:
  - The HPE FlexFabric 59x0 Switch Series has models with the ability to support native FC interfaces which can connect directly to FC initiators and targets.
- 10GbE, 40GbE, or 100GbE:
  - The HPE FlexFabric 12908 supports up to 384 native line rate 10GbE ports, 288 line rate 40GbE ports, or 64 line rate 100GbE ports. The 16 slot HPE FlexFabric 12916 supports up to 768 native line rate 10GbE ports, 576 line rate 40GbE ports, or 128 line rate 100GbE ports.
- Modular components:
  - The HPE FlexFabric 12900E switch is a modular chassis that was released in 2013. It can provide the necessary bandwidth and buffering for today as well as future scaling needs. It also supports redundant and hot swappable power supplies, fabrics, management modules and fan trays.
- L2 optimized for VMs:
  - 2-tier spine/leaf designs such as this are able to scale to large size L2 domains using IRF and TRILL or SPB.
- Orchestration support including Puppet, Chef, and Ansible
  - Network automation tools like Chef, Puppet, and Ansible are able to provide the controls necessary to manage and perform everyday management functions, which might range from just device discovery all the way to complete network configuration management and the provisioning of virtual network resources. Ansible is an agentless tool which communicates with the CLI interface over a Secure Shell, while Puppet and Chef rely on agents installed on the device.
- Management
  - This network configuration supports the full feature complement of HPE Intelligent Management Center (IMC) from the core switches down to the rack. This provides a single-pane-of-glass management platform to support the servers, VMs, virtual switches, IRF frameworks, and IP routing.
  - HPE IMC VAN Fabric Manager (VFM) Software is an IMC module which simplifies the management of data center and Fibre Channel Storage Area Networks (SAN) fabrics. It provides a unified view of all of the network and storage devices in the data center fabric alongside

fabric health to enable quick troubleshooting and pro-active management. VFM allows you to easily configure SPB or TRILL (Transparent Interconnect of Lots of Links) through the same graphical user interface used to automate, monitor and manage your entire network.

Specifications

Spine switch

The following devices are recommended as the core switch in the above design scenario.

- HPE FlexFabric 12900E Switch Series

BOM

Here is a sample 12908E BOM which can be used in this architecture.

Table 3. Spine example BOM

PART NUMBER	DESCRIPTION	QUANTITY	COMMENTS
JH255A	HP FF 12908E Switch Chassis	2	
JH108A	HP FF 12900E 2400W AC PSU	16	
JH258A	HP FF 12908E Fan Tray Assembly	4	
JH107A	HP FF 12900E LPU Adapter	8	
JC665A	HP X421 Chassis Universal Rck Mntg Kit	2	
JH257A	HP FF 12908E 5.0Tbps Type F Fabric Module	12	
JH104A	HP FF 12900E Main Processing Unit	4	
JH045A	HP FF 12900 36p 40GbE QSFP+ FX Module	12	
JG661A	HP X140 40G QSFP+ LC LR4 SM Transceiver	400	Used for leaf switch connectivity

Spine switch scalability table

This is the scalability table for 12908E discussed in this solution:

Table 4. Spine switch scalability

FEATURE	12908E FX MODULE
MAC Address	Up to 256K *
ARP (host)	16K/128K (Uni mode) *
Link aggregation: ports/group	64/1024
IPV4 LPM	32K
IPV6 LPM	8K **
IPv4/IPv6 MC	8K/1K **
Ingress/Egress ACL	18K/9K **
VXLAN	Yes
Native FC	No
DCB	Yes
SPB	Yes

* Shared Resource - up to 256K table entries

** Shared Resource of 24K

ToR/leaf switch

The following devices are recommended as a ToR/leaf switch with any of the above design scenarios.

- HPE FlexFabric 5900 switch series if less than four 40GbE ports are required or if VXLAN or integration with any of the HPE Network Virtualization solutions are not required
- HPE FlexFabric 5930 switch series if more than four 40GbE ports are required or if VXLAN or integration with any of the HPE Network Virtualization solutions are required
- HPE FlexFabric 5930 switch series are recommended for L2 VXLAN deployments
- HPE FlexFabric 5940 switch series are recommended for L2/L3 VXLAN deployments

Table 5. Leaf example BOM

Part Number	Description	Quantity	Comment
JH178A	HPE 5930 2-slot 2QSFP+ Switch	100	
JC680A	HPE 58x0AF 650W AC Power Supply	200	
JG552A	HPE X711 Frt(prt) Bck(pwr) HV Fan Tray	200	
JH182A	HPE 5930 24p 10GBASE-T/2p MCsc QSFP+ Mod	200	Used for WAN edge, BFD MAD and server connectivity
JG326A	HPE X240 40G QSFP+ QSFP+ 1m DAC Cable	100	Used for IRF connections
JG661A	HP X140 40G QSFP+ LC LR4 SM Transceiver	400	Used for spine switch connectivity

ToR/leaf switch scalability table

This is the scalability table for 5930 discussed in this solution:

Table 6. HPE FF 5930 table scalability and support

Feature	Support
MAC Address	288K
ARP (host)	120K
Link aggregation: ports/group	32/512
IPV4 LPM	120K
IPV6 LPM	64K
IPv4/IPv6 MC	4K/4K
Ingress/Egress ACL	3K/1K
VXLAN	Yes
Native FC	Yes – CP modules
DCB	Yes
SPB	Yes

Legacy 3-Tier network design

While it is possible to create a legacy 3-tier network for 2,000 physical servers, the added latency and complexity outweigh the benefits of this design. For those who wish to pursue this type of architecture, view the 3-tier optimized section in the HPE CFRA Guide.

L3 routed deployments

The advent of virtualization drove wide scale adoption of large L2 underlays, where typically L3 routing was only done in the spine switches. Many Enterprises have also worked to extend L2 deployments to other data centers so they can take advantage of long distance VM migration and disaster recovery protection. However, modern protocols can allow for L2 VM migration even over a L3 underlay, providing IT professionals with flexibility and choice of underlays.

The following are a few design scenarios that favor L3 deployments at the ToR switch.

- Foundation for Overlays and cloud solutions:
  - Overlay networks typically use a routed underlay network as the foundation for the solution. Private clouds can also leverage this architecture that allows for multi-tenancy and service insertion.
- Table scalability:
  - 2,000 physical server L2 deployments can very quickly reach table size scalability limits. L3 deployments are typically better able to handle the scalability required for larger deployments.
- Exponential growth with pods:
  - A pod approach can be used to scale to very large sizes. In this scenario, a chassis product would provide the L2 and L3 connectivity for a set of racks that would communicate via the L3 network to other pods.
- Deterministic path selection:
  - L3 deployments are able to engineer traffic to use a particular link and in the event of a failure know exactly what link it will take. In a L2 design, we rely on hash algorithms to distribute the packets across all links. Hashing is imperfect, so we don't actually evenly share across all links. A routing protocol calculates the best path based on different parameters. Adjusting these parameters, like path cost in OSPF or local preference in BGP, will determine what path is chosen every time in a routed network as well as the path chosen every time in a failure. In addition, LAG groups are limited to hardware and software restrictions with the 5900 currently supporting 16 ports per LAG. Depending on the product, ECMP can support 16, 32, or 64 links, which can increase the performance between the ToR and core.
- Network segmentation:
  - L3 also provides a way to segment the network. By using ACLs, access to a rack or certain servers within a rack can be easily controlled. While this can be done using MAC based ACLs, writing L3 ACLs allows for more hosts to be blocked rather than writing a specific ACL for each host and potentially exhausting ACL resources.
- Restrict VM sprawl:
  - L3 deployments can mitigate VM sprawl within a data center.
  - There are currently two prevailing protocols for layer 3 underlays.
    - Open Shortest Path First (OSPF): A link-state interior gateway routing protocol (IGP), OSPF is a protocol which has knowledge of the complete topology, allowing it to be able to provide for good traffic engineering so that routes can be manipulated based on requirements.

      Allows for good scaling but as networks get larger, the size and frequency of topology updates can disrupt stability and delay route calculation while topologies converge. However, OSPF offers summarization and area isolation techniques that can help to overcome these types of issues.

      ECMP provides deterministic load-balancing across multiple paths.
    - Border Gateway Protocol (BGP): Widely associated as an exterior routing protocol that runs the internet, BGP can also be used as an interior protocol or both

      Designed to exchange reachability information between separate autonomous systems. BGP is able to make routing decisions based on network policies and paths available

      Proven to scale very large, BGP sends updates only when a topology change has occurred and only the affected part of the table is sent

BGP devices contain two sets of routing tables. One is for Internal BGP (iBGP) routes within the same autonomous system, and the other table is reserved for routes between autonomous systems (eBGP)

Glossary

ACL: A network access control list (ACL) is an optional layer of security for your VPC that acts as a firewall for controlling traffic in and out of one or more subnets.

ARP: The address resolution protocol (arp) is a protocol used by the Internet Protocol (IP) [RFC826], specifically IPv4, to map IP network addresses to the hardware addresses used by a data link protocol. The protocol operates below the network layer as a part of the interface between the OSI network and OSI link layer.

BGP: Border Gateway Protocol (BGP) is a standardized exterior gateway protocol designed to exchange routing and reachability information among autonomous systems (AS) on the Internet. The protocol is often classified as a path vector protocol but is sometimes also classed as a distance-vector routing protocol.

DCB: Data center bridging (DCB) is a series of enhancements to the IEEE 802.1 standard to provide extensions to Ethernet for support for converged technologies such as Fiber Channel over Ethernet (FCoE).

EVI: Ethernet Virtual Interface (EVI) runs over Internet Protocol (IP) transport and extends layer 2 domains across a WAN network, typically between data centers. By virtualizing and automating the link-layer domain across data centers, EVI delivers the elements necessary to enable Software-defined Networking (SDN) data center infrastructure. It enables several data centers to work as one that is more responsive, with higher efficiency and solid high availability for business resiliency.

FCoE: Fiber Channel over Ethernet (FCoE) is an encapsulation of Fiber Channel frames over Ethernet networks. This allows Fiber Channel to use 10 Gigabit Ethernet networks (or higher speeds) while preserving the Fiber Channel protocol.

FC: Fiber Channel, a Gigabit-speed network technology primarily used for storage networking.

HPE IMC: HPE Intelligent Management Center (IMC) delivers next-generation, integrated and modular network management capabilities that efficiently meet the end-to-end management needs of advanced, heterogeneous enterprise networks.

IRF: Intelligent Resilient Framework (IRF) is a software virtualization technology developed by H3C (3COM). Its core idea is to connect multiple devices through physical IRF ports and perform necessary configurations, and then these devices are virtualized into a distributed device.

iSCSI: The Internet small computer system interface (iSCSI) is a TCP/IP-based protocol for establishing and managing connections between IP-based storage devices, hosts, and clients, called the storage area network (SAN).

Jumbo Frames: Jumbo frames often mean 9,216 bytes for Gigabit Ethernet, but can refer to anything over 1,500 bytes.

LACP: Link aggregation control protocol (LACP) is part of the IEEE specification 802.3ad that allows you to bundle several physical ports to form a single logical channel.

LAG: Link aggregation (LAG) is used to describe various methods for using multiple parallel network connections to increase throughput beyond the limit that one link (one connection) can achieve.

MAC: A media access control address (MAC address), also called physical address, is a unique identifier assigned to network interfaces for communications on the physical network segment. MAC addresses are used as a network address for most IEEE 802 network technologies, including Ethernet and WiFi.

MDC: Multitenant Device Context (MDC) is an HPE feature for multi-tenancy which gives you the ability to virtualize a physical switch into multiple logical devices; each logical switch has its own tenants.

MMF: Multi-mode optical fiber is a type of optical fiber mostly used for communication over short distances, such as within a building or on a campus.

MPLS: Multiprotocol Label Switching (MPLS) is a type of data-carrying technique for high-performance telecommunications networks that directs data from one network node to the next based on short path labels rather than long network addresses, avoiding complex lookups in a routing table.

MPO/MTO: MPO (Multi-fiber Push On) is a connector for ribbon cables with four to twenty-four fibers. MTP is a brand name for a version of the MPO connector with improved specifications.

MSTP: The multiple spanning tree (MST) protocol carries the concept of the IEEE 802.1w rapid spanning tree protocol (RSTP) a leap forward by allowing the user to group and associate VLANs to multiple spanning tree instances (forwarding paths) over link aggregation groups (LAGs).

NIC: Network interface cards (NIC) are adapters attached to a computer (or other network device such as a printer) to provide the connection between the computer and the network.

NMS: The network management system (NMS) is a combination of hardware and software used to monitor and administer a network.

OSPF: Open Shortest Path First (OSPF) is a routing protocol for Internet Protocol (IP) networks. It uses a link state routing algorithm and falls into the group of interior routing protocols, operating within a single autonomous system (AS).

OVSDB: The Open vSwitch Database Management Protocol (OVSDB) is an OpenFlow configuration protocol that is designed to manage Open vSwitch implementations.

RIB: In computer networking a routing table, or routing information base (RIB), is a data table stored in a router or a networked computer that lists the routes to particular network destinations, and in some cases, metrics (distances) associated with those routes.

RSTP: The rapid spanning tree protocol (RSTP IEEE 802.1w) can be seen as an evolution of the IEEE 802.1d standard more than as a revolution. IEEE 802.1w is also capable of reverting back to IEEE 802.1d in order to interoperate with legacy bridges (thus dropping the benefits it introduces) on a per-port basis.

SAN: A storage area network (SAN) is a high-speed special-purpose network (or subnetwork) that interconnects different kinds of data storage devices with associated data servers on behalf of a larger network of users.

SCSI: The small computer system interface (SCSI), an ANSI standard, is a parallel interface standard used by Apple Macintosh computers, PCs, and many UNIX systems for attaching peripheral devices to computers.

SFP: The small form-factor pluggable (SFP) is a compact, hot-pluggable transceiver used for both telecommunication and data communications applications. The form factor and electrical interface are specified by a multi-source agreement (MSA).

SMF: Single Mode fiber optic cable has a small diametral core that allows only one mode of light to propagate. This application is typically used in long distance, higher bandwidth runs.

SNMP: The simple network management protocol (SNMP) is used by network management systems to communicate with network elements.

SPB: Shortest path bridging (SPB) provides logical Ethernet networks on native Ethernet infrastructures using a link state protocol to advertise both topology and logical network membership.

STP: The spanning tree protocol (STP) is an L2 protocol designed to run on bridges and switches. The main purpose of the spanning tree is to prevent loops from forming in a bridged network.

QSFP: The Quad Small Form-factor Pluggable (QSFP) is a compact, hot-pluggable transceiver used for data communications applications. It interfaces networking hardware to a fiber optic cable or active or passive electrical copper connection.

ToR: Top-of-rack utilizes a switch at the top of each rack (or close to it).

TRILL: TRILL (Transparent Interconnection of Lots of Links) is an IETF Standard implemented by devices called RBridges (routing bridges) or TRILL Switches. TRILL combines techniques from bridging and routing and is the application of link state routing to the VLAN-aware customer-bridging problem.

VEPA: A standard being led by HP for providing consistent network control and monitoring for virtual machines (of any type).

VM: A virtual machine is a system that enables multiple operating systems to concurrently run on a single physical server, providing much more effective utilization of the underlying hardware.

VLANs: Virtual LANs (VLANs) provide the capability to overlay the physical network with multiple virtual networks. VLANs allow you to isolate network traffic between virtual networks and reduce the size of administrative and broadcast domains.

VTEP: A VXLAN (Virtual Extensible Local Area Network) Tunnel End Point is a VTEP is a host interface which forwards Ethernet frames from a virtual network via VXLAN or vice-versa. All hosts with the same VNI configured must be able to retrieve and synchronize data (ARP and MAC tables for example).

VXLAN: Virtual Extensible LAN (VXLAN) is a network virtualization technology that attempts to improve the scalability problems associated with large cloud computing deployments. It uses a VLAN-like encapsulation technique to encapsulate MAC-based OSI layer 2 Ethernet frames within layer 4 UDP packets.

Additional links

HP Cloud-First Reference Architecture

HP Cloud-First Reference Architecture guide-100 Server

HP Cloud-First Reference Architecture guide-500 Server

Product support manuals and software

Learn more about HPE transformational areas

Learn more at
HPE.com/networking

Sign up for updates

★ Rate this document

4AA5-7340ENW, Apr 2016