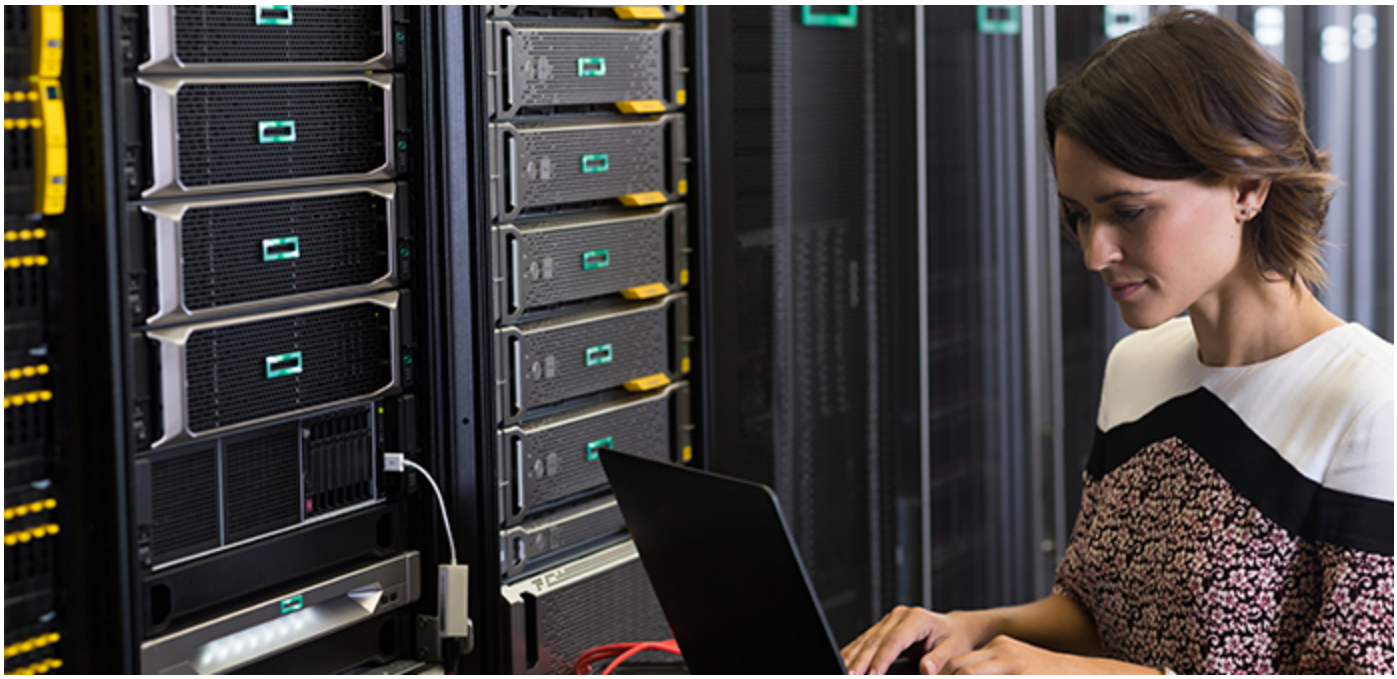
Hewlett Packard Enterprise

Technical white paper

HPE Composable Fabric

Data center fabric explained



Contents

Introduction

Today's business and government organizations are focused on digital transformation. As a result, DevOps and business agility are at the forefront of most current strategic IT discussions. To deliver both, enterprise organizations have increasingly leveraged public cloud service providers and are now building on-premises clouds with infrastructure that delivers on three basic requirements:

1. Friction-free agility of physical resources
2. Control systems that maximize physical resource utilization and provide maximum ROI
3. Integration of the various infrastructure components for automated provisioning and resource management

While these requirements exist across the spectrum of infrastructure components (compute, storage, and network), the network plays a foundational role. The network fabric acts as the glue between compute and storage, so the agility, control, and integrality of the network (or lack thereof) directly impacts an organization's ability to maximize their compute and storage resources. Any bottlenecks in the data center fabric will negatively impact the organization's ability to run applications efficiently and optimally distribute data. The more the network is in your way, the less business/mission agility can be attained.

Hewlett Packard Enterprise has created a composable data center fabric operating environment that leverages commodity off-the-shelf Ethernet hardware and innovative software that forms a single distributed network without the limitations of traditional approaches—HPE Composable Fabric. Within the HPE Composable Fabric solution, the physical Ethernet devices that make up the fabric are referred to as rack connectivity modules.

For enterprise customers as well as cloud providers, this modern cloud fabric can be leveraged as a fully integrated data center networking solution (switches, controllers, and such) for ease of deployment. For larger cloud providers, the constituent technologies are designed to be leveraged as a set of tools woven into the overall cloud orchestration strategy. This paper focuses on describing those required capabilities and the way the HPE Composable Fabric solution addresses them.

The problems

Traditional networks have many obstacles including, but not limited to, full resource agility, network control, and utilization. Many enterprises are forced to impose network-specified boundaries that do not correspond to business or workload needs. They also leverage tightly coupled control planes that cannot easily be overridden to achieve specific objectives. In addition, the network contains no inherent, simple mechanism to manipulate the overall state or specific aspects of the fabric. Instead, these enterprises rely on a port/box configuration paradigm that, at best, can only provide scripted automation. Let's look at these in more detail.

Network design-imposed bottlenecks

Traditional data center networks are typically wired in a leaf-and-spine configuration, whereby a set of top-of- rack (ToR) leaf nodes (switches) aggregate rack-based servers and storage. This ToR leaf switch is then connected to a cross-rack spine layer to provide broader connectivity to other hosts (see Figure 1).

The rack-to-rack bandwidth available in the leaf-and-spine architecture depends on the number of fabric-dedicated ports in each ToR switch and their bandwidth. The remaining ports on the switch can be used for host connectivity. For example, if a given ToR switch has 32 ports with the same link rate but only four of these are dedicated to the fabric, 28 ports are available for node connectivity. When all 28-node ports are in use and the traffic pattern from the attached nodes desires to leave the ToR switch, such a port economy has a contention ratio of 28:4 or 7:1.

This rate of oversubscription favors workloads that can be contained locally in the rack with traffic patterns only traversing the ToR switch. When rack-to-rack traffic is encountered at full-node bandwidth, the first four nodes have ample bandwidth, but the fifth node has no choice but to wait for the available bandwidth. The result of such bandwidth crowding increases transfer times, higher latency, and potential link timeouts. In fact, at this point, the network has stopped functioning as the network and applications begin to suffer.

Less contention at the ToR enables a more agile data center fabric. It is, therefore, highly desirable to design balanced networks that have just as many ports dedicated to nodes as is dedicated to the fabric itself. Again, using a 32-port switch as the example, it would mean using 16 ports for nodes and 16 ports for the fabric. This ToR configuration is considered non-blocking.

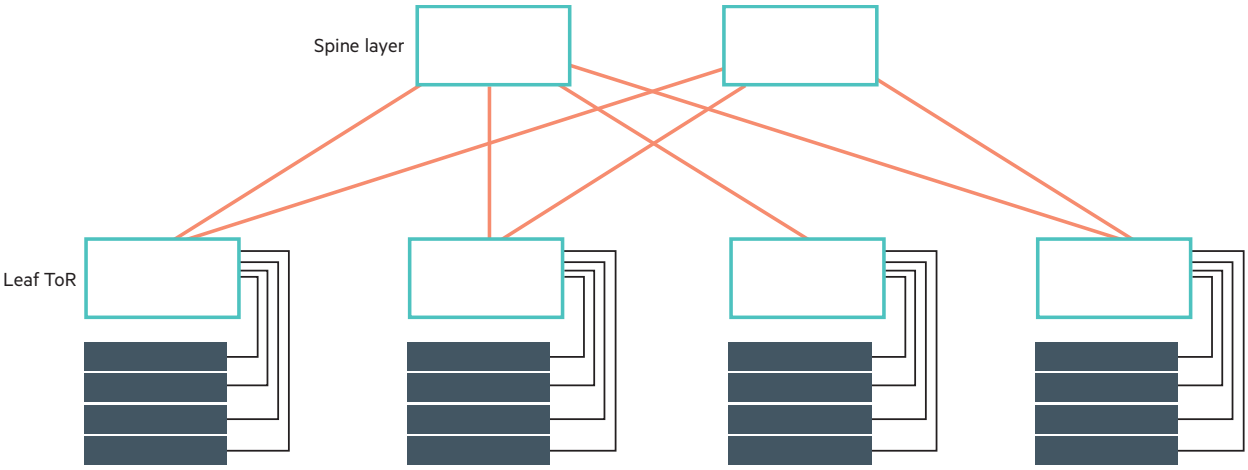


Figure 1. Traditional leaf-and-spine fabric

Within limitations, it is possible to build a non-blocking leaf-and-spine fabric by dedicating half the bandwidth in the ToR switch to the fabric but in many cases this strategy is deemed too expensive. This is especially true for modern data center fabrics where the same 32-port switch is intended for both the leaf-and-spine duty. The challenge building high-bandwidth leaf-and-spine fabrics is the cost of the optical transceivers. Using a modern fully configured 32-port 100GbE switch, the cost of the optics can easily constitute two-thirds of the overall cost. It is, therefore, easy to understand why network designers are willing to live with some ToR contention. Even when the decision to do so impedes the goal of maximum business agility.

Load balancing-imposed bottlenecks

Another of the desired goals in traditional network design is to achieve maximum usage of the available "fabric" links (shown as the red lines in Figure 1) while accounting for failure scenarios, without creating network loops, which can create havoc in any network. In general, this is known as loop-free active:active load distribution. In most data centers, this is achieved by confining Layer 2 (L2) communications to the leaf nodes and leveraging Layer 3 (L3) addressing for inter-rack traffic.

In this design, potential L2 loops can be contained to a single rack and a simple load distribution protocol—Equal-Cost Multi-Path (ECMP)—can be used to randomly place traffic across a set of uplinks from the leaf switch to each spine switch. These protocols create mobility boundaries that negatively impact the application and compute/storage infrastructure utilization. In addition, the intended goal of evenly distributing storage and application bandwidth across these leaf-and-spine architectures is not the actual achieved outcome, as discussed later in the document. So, while this may seem like a very elegant solution that achieves the objectives of maximal network resource utilization, let's look at the actual impact this architecture has on an agile data center.

Protocol boundaries

Creating an L2/L3 boundary at the leaf node helps from the perspective of leaf-to-spine traffic distribution since all available links can be used in a load-balanced multipath approach. However, this type of model creates a **network-imposed** addressing boundary that may complicate application deployments and mobility. Certain applications expect to have the same IP address regardless of their physical location and may need to have clusters of resources on the same network address. Newer applications may be implemented such that they are agnostic of network addressing schemes but many organizations have just begun the multiyear process to modernize such applications. To work around such issues, some vendors have promoted the use of software-based overlays that create host-to-host tunnels through the IP fabric to give applications or virtual machines the appearance of being on the same L2 network. These types of overlays, however, create an additional layer of overhead and complexity that needs to be managed/administered. It also creates new network and application conditions that need to be understood.

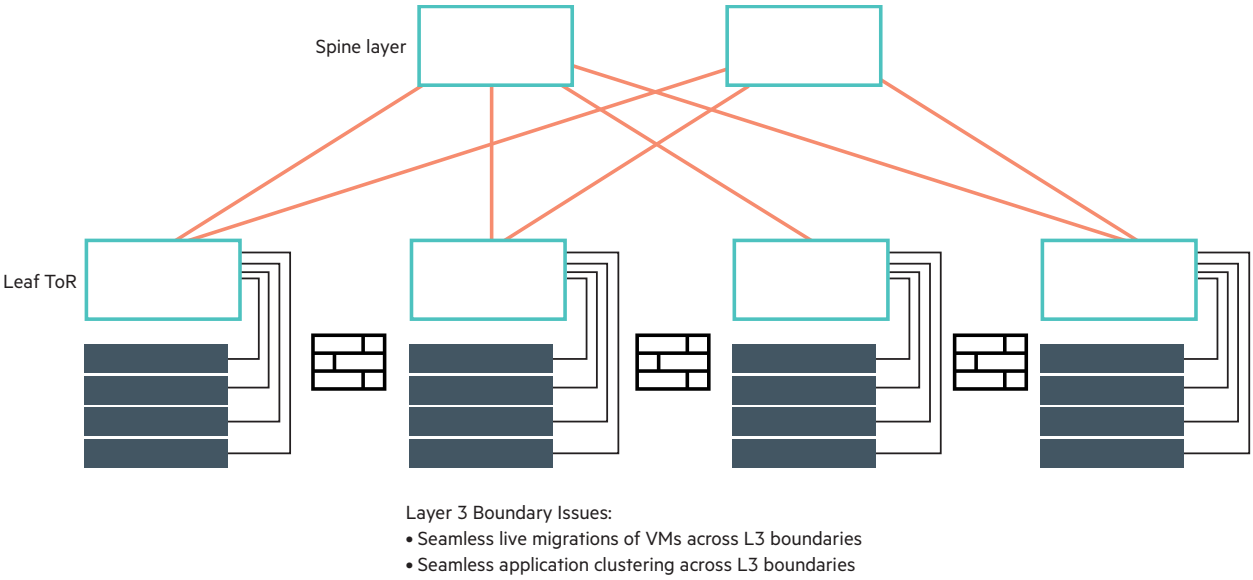


Figure 2. L2/L3 Rack boundaries

Geographic boundaries

Another issue to consider with this type of design is the ease of extensibility to multiple data center locations. While supporting disaster recovery types of scenarios is relatively simple from an application perspective, many organizations now want to leverage multiple data centers for load distribution of active workloads. Again, many applications are written to require L2 connectivity for cluster members, and in this case, the organization needs to employ some type of tunneling mechanism between data centers. Also, it's very likely that the inter-data center bandwidth or connectivity properties (such as latency) are different than the intra-data center properties. Care must be taken in designing how applications will be extended to account for these limitations.

Bandwidth distribution

One of the selling points of an L2/L3 boundary in this type of architecture is the ability to evenly distribute bandwidth across the fabric. However, this is a major downside in most data centers where workloads are not evenly distributed and do not evenly consume bandwidth. In fact, the distribution model is not even at all. This distribution model is random. These networks leverage some type of equal cost multipath technology, such as open shortest path first (OSPF) ECMP, which aim to simulate evenness by selecting a random link for a given new flow based on a hash of the flow's header data. Since most data centers have an uneven distribution of workloads (in size, scope, and temporal usage of network resources) instead of resulting in an even distribution of traffic, this leads to hot spots and congestion problems that are very difficult to troubleshoot. The only solution is to overbuild or overprovision the network. This leads to significantly higher CAPEX than is necessary and it does not solve the root problem.

Latency

While bandwidth should be variable based on the specific needs of the workload at any given time, latency is the opposite. For a given workload, latency should be deterministic; however, in leaf-and-spine architectures they are not. Since links are selected at random and any given end-to-end path could traverse through a multistage spine, there is a large variance in latency characteristics.

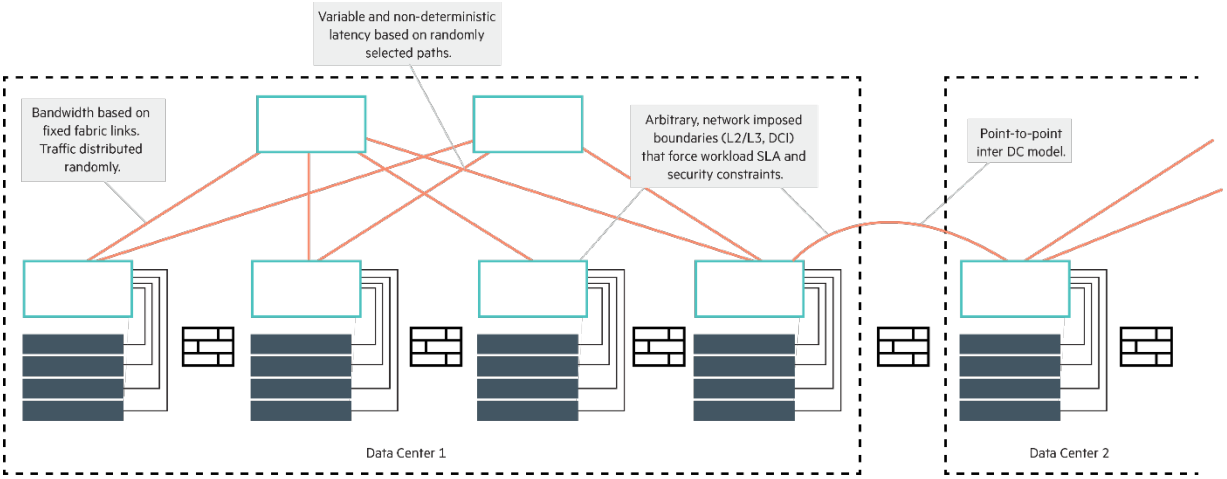


Figure 3. Multi-data center concerns

Control plane limitations

Traditional data center networks typically employ a tightly coupled control plane model. This means that traffic distribution, link failover, route selection, and other control plane functions are handled by a set of limited protocols that are not designed to accept dynamic user input. For example, most data center networks leverage OSPF ECMP to achieve or attempt to achieve even load distribution. OSPF is a very simplistic protocol that attempts to solve for the shortest path. However, a given workload or set of workloads may have a more complicated set of constraints to solve for. For example, any good GPS will let a user solve for shortest path, cheapest path (no tolls), quickest path, or some combination thereof. It may also allow the user to identify specific waypoints.

The notion of software-defined networking (SDN) has been introduced relatively recently into the conversation on data center networks to try to address some of these limitations. SDN aims to decouple the control plane from the network in a way that makes it more extensible. However, early attempts at SDN solutions have always assumed that the underlying physical infrastructure (such as the roads) were fixed (since very few startups have the ambition to solve the whole problem, and the incumbents do not have the motivation).

This assumption means that there is very little that can be customized dynamically in the network. In most cases, SDN boils down to the ability to override the embedded control plane's route selection to a specific path that better meets the user's criteria. However, this requires a user to mentally understand individual flows and individual paths through the network. This is certainly not a scalable model for a data center of any modest size.

Data center fabric capability	Impact of traditional approaches	Desired strategy
Bandwidth	• Static • ToR contention tradeoffs made at deployment • Path randomization attempts even traffic distribution	• Dynamic • Distributed based on workload need
Latency	Variable and non-deterministic	Deterministic and based on workload demands
Boundaries	• Determined by network wiring topology • Irrelevant or harmful to workloads	Defined completely by workload segmentation needs
Data center interconnect and extensions	• Limited to point-to-point model • Unable to create seamless and elastic pool of resources	• Multipoint • Seamless and elastic pool of resources; able to factor link qualities into resource allocation

If the assumption about a fixed set of paths is lifted, the control plane becomes more powerful. In this case, the control plane can now not only understand user inputs or specific constraints to solve but can also manipulate the physical resources of the network. In this scenario, the controller becomes an incredibly powerful tool that understands workload constraints and can leverage dynamic and flexible network resources to best meet those constraints.

Workflow automation plane limitations

Finally, when all these capabilities are put together, the need to automate the allocation of network resources quickly arises. This automated allocation must support workloads based on a model that dynamically understands the events that occur in the data center and has pre-programmed logic to react to those events. Typical integration environments focus only on the ability to script what was once manually entered cryptic command-line interfaces or graphical user interfaces (GUIs). Embedding complex logic into this type of scripted environment is fraught with peril, as small errors can explode in their impact, and the logic must be adjusted and applied to individual network elements. In addition, as different scenarios or needs are encountered, the entire logic base becomes increasingly complex to the point of unmanageability.

The new model is to treat automation as an event processing system. Events are what they appear to be—things that happen in the data center. Simplicity is gained from an automation model where known events are decomposed and the system dynamically detects the occurrence of those events, known as triggers. Based on the occurring trigger conditions, it becomes much easier to enact specific policies that specify the resulting automated actions. As the type of events increase, it becomes much easier to visualize and account for the various interactions that come into play.

HPE Composable Fabric

HPE has built a comprehensive, composable data center fabric that addresses these issues at the data plane, the control plane, and the integration plane. Each of these planes is meant to evolve independently. In some large cloud environments, the individual technologies within each plane can be leveraged and mated to technologies that the cloud operator might already be implementing in the other planes. However, the goals of the three planes are very clear and represent what HPE believes to be required capabilities for cloud builders.

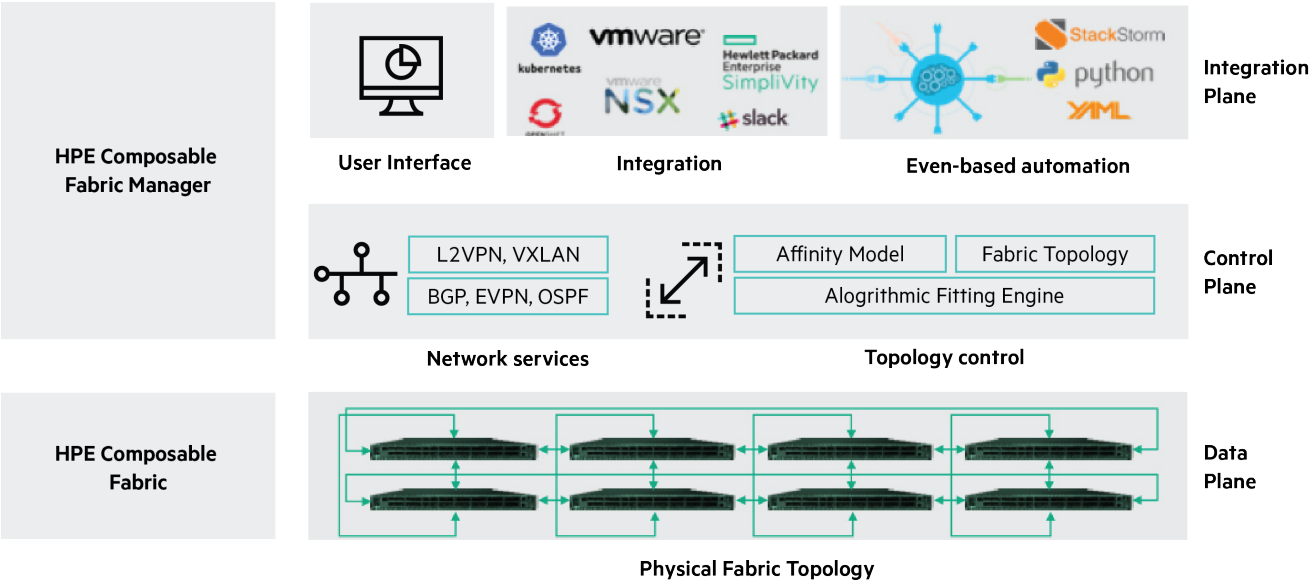


Figure 4. HPE Composable Fabric—SDN building blocks

The data plane

The HPE Composable Fabric's rack connectivity modules form a composable network fabric or data plane that provides physical connectivity, topology, and data/packet forwarding. Central to the HPE Composable Fabric is the collapsing of connectivity and routing into a single building block. Due to advanced distributed software, there is no longer a need for dedicated spine switches. HPE Composable Fabric Manager performs distributed intelligent routing between rack connectivity modules. This enables a variety of new, more efficient, and highly diverse data center fabric topologies.

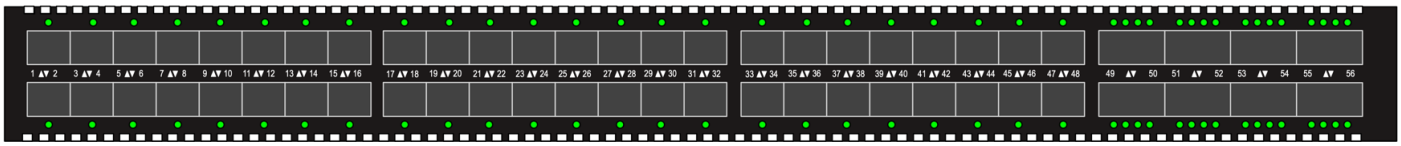


Figure 5. HPE Fabric Module 3180 48 QSFP28 (25GbE)/8 QSFP28 (100GbE) ports

By serving as a high-diversity fabric, HPE Composable Fabric data plane provides many ways for workloads to interconnect (between member resources as well as other workloads). In contrast to aggregation-based (or low-diversity fabrics) such as traditional multi-tier, leaf-spine, Clos, and such, a high-diversity fabric allows intelligent software algorithms. These selectively place workload traffic on the corresponding fabric paths that meet specific SLA or security parameters for that workload.

These advanced fabrics have been used successfully in closed systems such as high-performance computing clusters with proprietary interconnects such as Torus, Hypercube, and more. But have not been generally available, until now, for Ethernet/IP-based networks due to the limitations of a traditional legacy protocol-based control plane. The HPE Composable Fabric solution allows these advanced topologies to be used with Ethernet/IP traffic by supplementing the embedded protocol-based control plane with a highly advanced algorithm-based control plane.

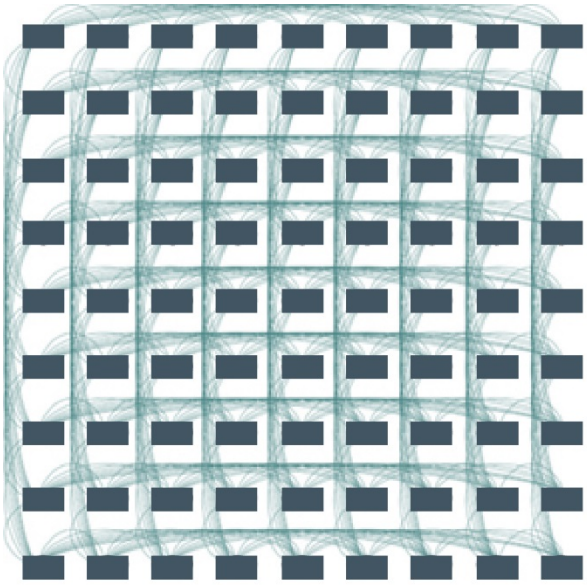


Figure 6. Highly diverse HPE Composable Fabric

HPE Composable Fabric is created by using rack connectivity module interconnects to form a fully connected mesh. While fully connected meshes are not new, the ability to make use of these in multiple dimensions simultaneously for use in general purpose is what cloud computing is.

The easiest way to understand this advanced and powerful data center topology is to start with a single centrally placed rack. Because modern workloads are multidirectional (east-west and north-south traffic patterns), direct connections from the central rack connectivity module is laid out in two dimensions—4-left, 4-right, 4-north, and 4-south. The same is true for all rack connectivity modules in the fabric.

To make cabling easier, the ToR rack connectivity modules are interconnected directly (using simplified cabling/fiber solutions that HPE provides) to form data center-wide fabric without the use of an expensive, power hungry, and latency- inducing spine or aggregation tiers. Smaller data center implementations can be wired in a single dimension, essentially leveraging the standard ToR uplink ports as east-west ports to neighboring connectivity modules.

HPE Composable Fabric FM 1006 is a passive optical module that handles distributing individual fibers to other rack connectivity modules in the mesh to provide direct connectivity from any given module to a number of other modules (generally around 10). These rack connectivity modules can also act as transit points to allow any given module to connect to the rest of the fabric with minimal module hops.

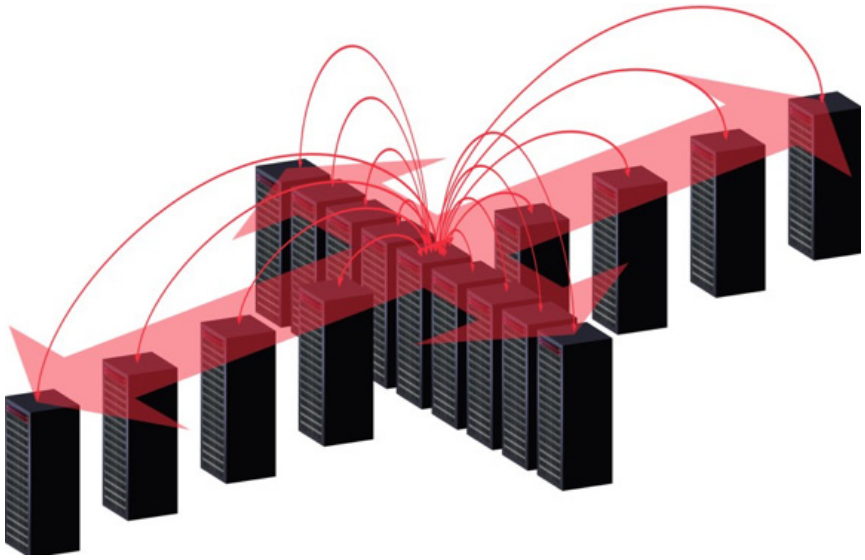


Figure 7. HPE Composable Fabric—fully connected rack-to-rack fabric topology

For larger or more bandwidth- and latency-sensitive designs, a second dimension is added, which allows double the number of direct connectivity modules reachable from any given module, as well as drastically increasing the 2-hop reachability. Within the resulting matrix appears a myriad of logical fabrics that are available to the control plane for selective workload placement depending on need.

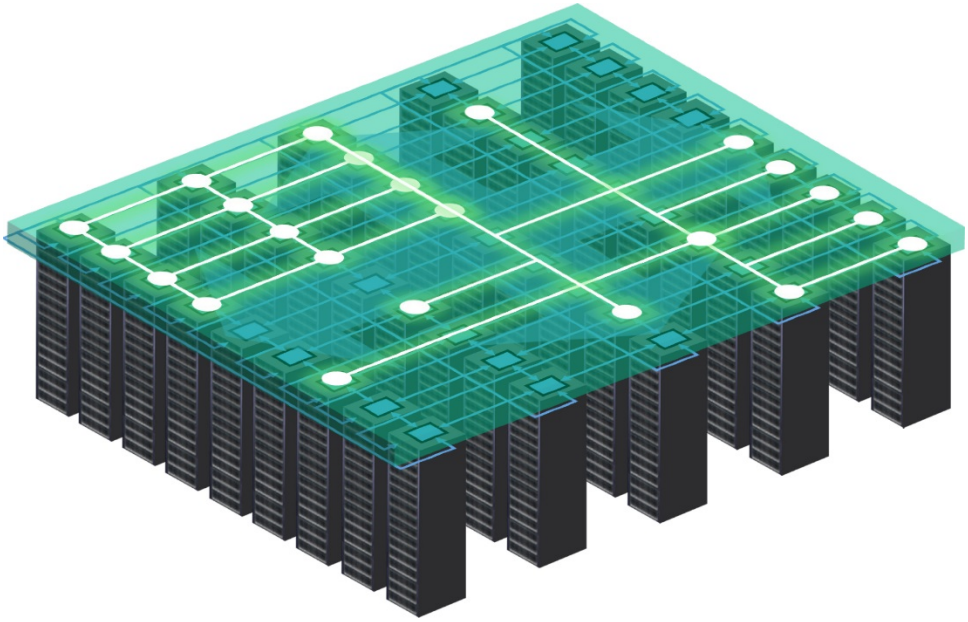


Figure 8. HPE Composable Fabric—dynamic fabric underlay

Control and management plane

The control and management plane, which is part of the HPE Composable Fabric Manager, supplements the embedded protocols used in the data plane and provides a single point of management for the network. It also provides APIs to directly manipulate the network state and objects from external systems.

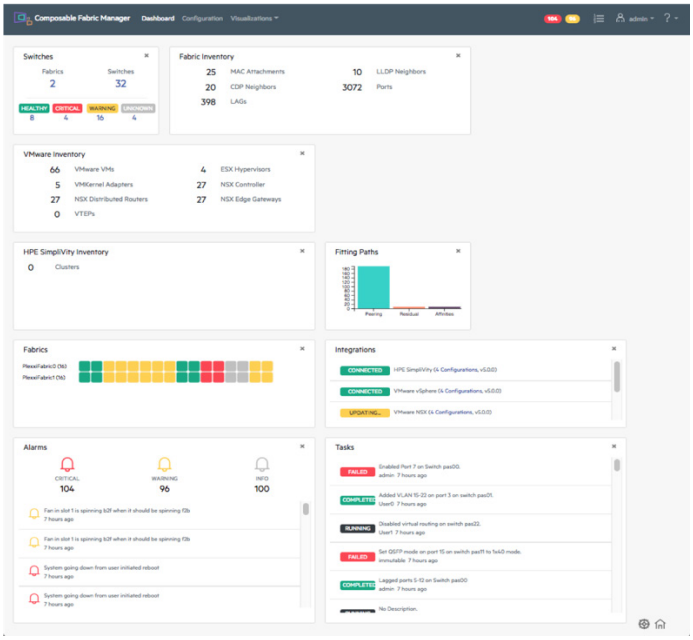


Figure 9. HPE Composable Fabric Manager

What makes HPE Composable Fabric Manager different from a basic SDN controller is the ability to define workloads through the affinity data model and associated APIs that are used to define workload and their relationships. This is done with the algorithms that are used to manipulate how and what type of connectivity these workloads receive from the data plane (a process called fitting). Fitting allows users to add new workloads that have specific performance (bandwidth, latency, isolation) or security/fate sharing (keep apart / keep together) requirements explicitly. As against the traditional network approach that is letting the embedded protocols decide based on equal-cost algorithms that have no workload awareness.

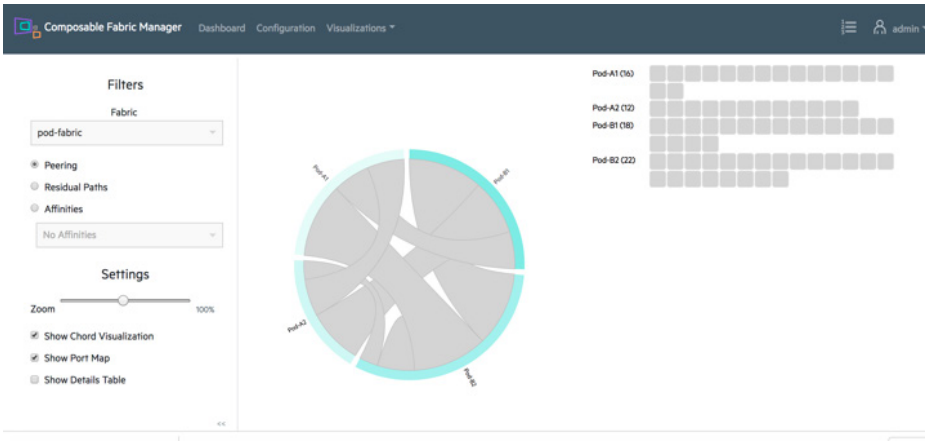


Figure 10. Fabric visualizer

Integration plane

The integration plane creates an easy-to-use, event-based automation platform that allows the user to define very simple conditions (trigger events) that result in specific actions. This type of event-based automation is powerful as it allows users to make the network fully dynamic in the face of network-intensive events (such as data ingest or movement, time-of-day events, and more). Additionally, HPE provides a number of built-in integrations into popular cloud orchestration and software-defined storage systems (SDSS). HPE Composable Fabric Manager is built on top of a popular open source project called StackStorm that has a vibrant community that also builds integrations into popular DevOps and ChatOps automation tools.

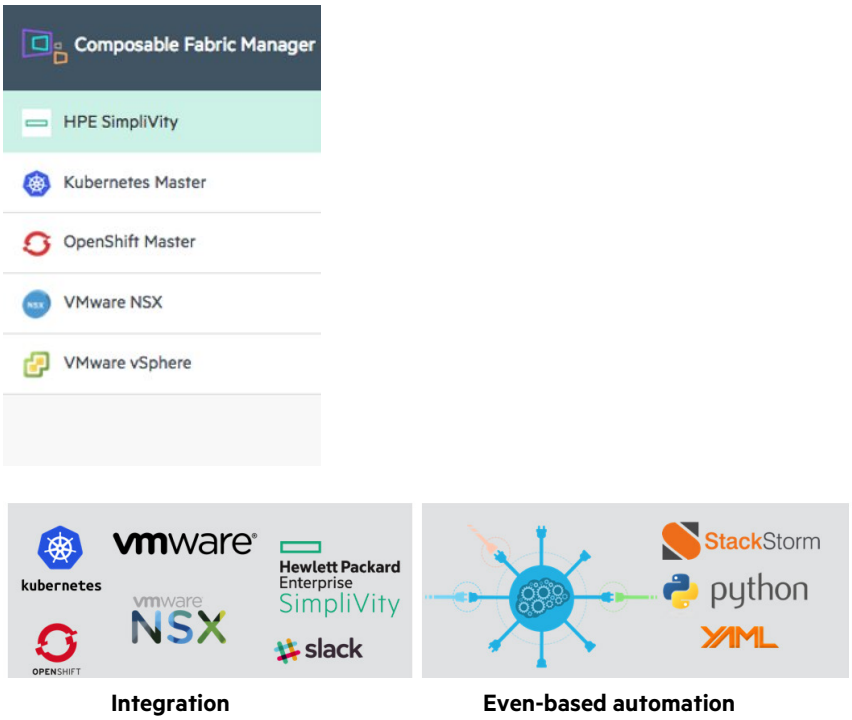


Figure 11. HPE Composable Fabric—API and event-driven application integration

Summary

The next era of IT is guided by the need for business and mission agility. Users and developers alike have come to expect a friction-free cloud experience from IT. As new applications are delivered and deployed, the data center now has emerged as the common transport for getting things done. New and highly diverse IT architectures are now needed to meet requirements for agility, integration, and simplicity.

New cloud-driven IT consumption models have emerged, such as IT as a service in the public cloud and IT as a converged offering in the private cloud. As a result, today's data center decision-makers need a better and more responsive data center network strategy.

HPE offers this new data center fabric today. **HPE rack connectivity modules** are scalable, dynamic, and responsive. **HPE Composable Fabric Manager** software is intelligent, automated, and application-driven delivering integration layer that provides processes and tools that modern IT can leverage to efficiently transition business to the digital era.

Learn more at

hpe.com/us/en/integrated-systems/composable-fabric.html

a00060583ENW, December 2018