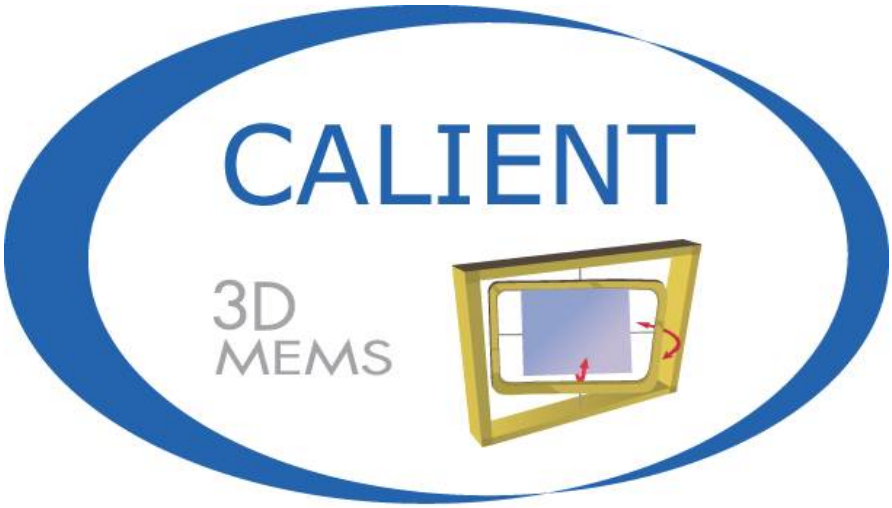


Move the light, not the fiber

The Software Defined Hybrid Packet Optical Datacenter Network

SDN AT LIGHT SPEED™

INTRODUCTION

In datacenter networks, video, mobile data, and big data are driving an explosion in network growth and server deployments. According to Cisco [(1)], Global datacenter traffic will grow fourfold and reach a total of 6.6 zettabytes annually by 2016. Global cloud traffic, the fastest-growing component of datacenter traffic, will grow six-fold, representing a huge 31% CAGR. Importantly, 76% of this traffic is forecast to remain within the datacenter, representing high capacity east-west interactions between clusters.

This growth is already resulting in networking bottlenecks and performance degradation and the result is that the performance of new and expensive server resources is being constrained by the traditional datacenter network architectures and networking equipment. The network interconnecting racks and clusters has become a point of constriction.

Optical circuit switching can be deployed together with packet-based switching solutions to provide a hybrid network to dramatically improve performance and to scale to support this rapid growth so that the full value of new server resources can be realized.

This paper describes a high-level hybrid packet-circuit based datacenter network architecture that equally supports both short bursty traffic and high-capacity, high-persistence data flows across the datacenter.

THE DATACENTER CHALLENGE

Big Data, together with video, server virtualization, cloud applications, mobile data, and the need to store and replicate vast amounts of data has led to a situation where there are large dynamically changing data traffic patterns and flows across modern datacenters. This is particularly the case in large cloud datacenters where a relatively small number of applications can consume vast amounts of server resources and where virtual machine migrations and software frameworks like Hadoop create very large persistent “elephant” data flows. The net result is that traditional inter-cluster networks suffer from bandwidth constriction resulting in degraded server and application performance and increased latency.

One solution contemplated by network architects is to design the datacenter network for worst case traffic - essentially this means designing for low or zero oversubscription. However this causes a significant increase in capital cost at a time when revenue from customers is flat, and can also add considerable latency due to the size of the switch fabric.

What is needed is the capability to reconfigure the network on demand to deliver the available capacity when and where it is needed most to support large traffic flows as they arise and subside. This results in a high performance network that makes the best use of available resources with lowest capital investment.

To illustrate, what is found in most Datacenter networks today are fixed capacities with everything to everything connectivity and significant levels of oversubscription as shown in Figure 1. The problem is that the connectivity doesn't scale to support large data flows as they arise.

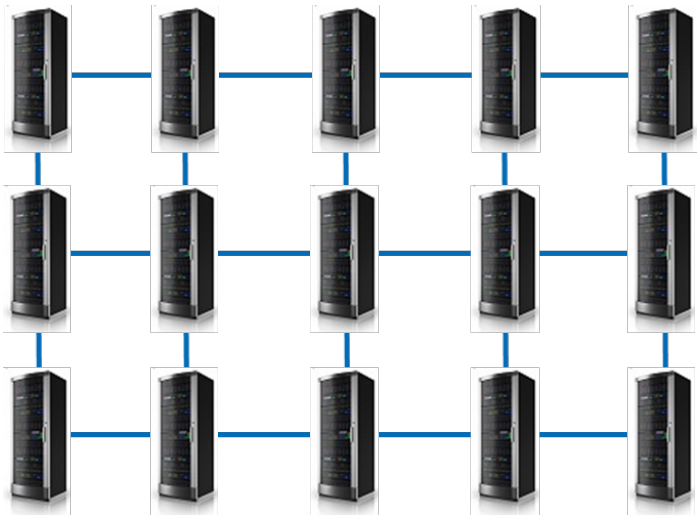


Figure 1: Typical Datacenter Network with Fixed Capacities

What is actually needed is shown in Figure 2: Connectivity and capacity where it's required, reconfigurable and scalable on demand. This provides the capability to deliver the bandwidth and server resources needed by applications precisely when and where they are needed.

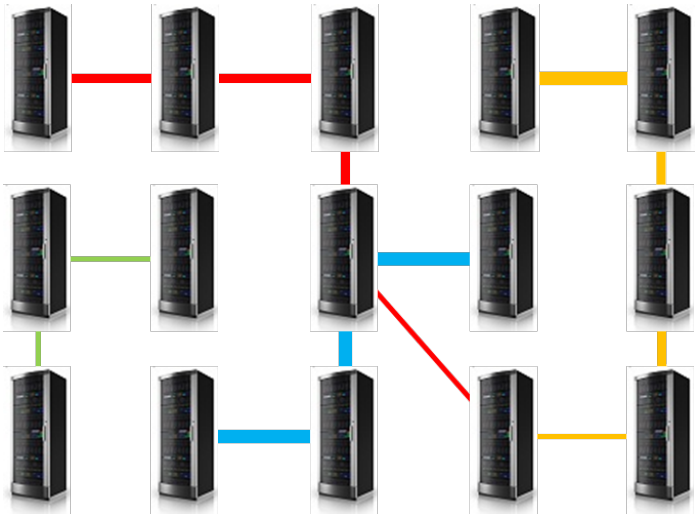


Figure 2: Dynamic Flexible Connectivity and Capacity Where Needed

This network reconfiguration can only be partially realized with traditional switching and routing solutions however. While typically optimized for short-bursty traffic these solutions can't economically support the switching of large data flows such as those generated by VM migrations.

In such scenarios Optical Circuit Switching can augment packet-based switching to create a hybrid solution with the much needed capability to economically reconfigure the network to support massive data flows with low latency and to scale as needed to 100 Gbit/s or beyond.

TRADITIONAL DATACENTER NETWORKS

Before introducing the hybrid datacenter network concept, let's briefly review a typical modern datacenter network architecture and the trends now seen in these networks. Figure 3 depicts such an example.

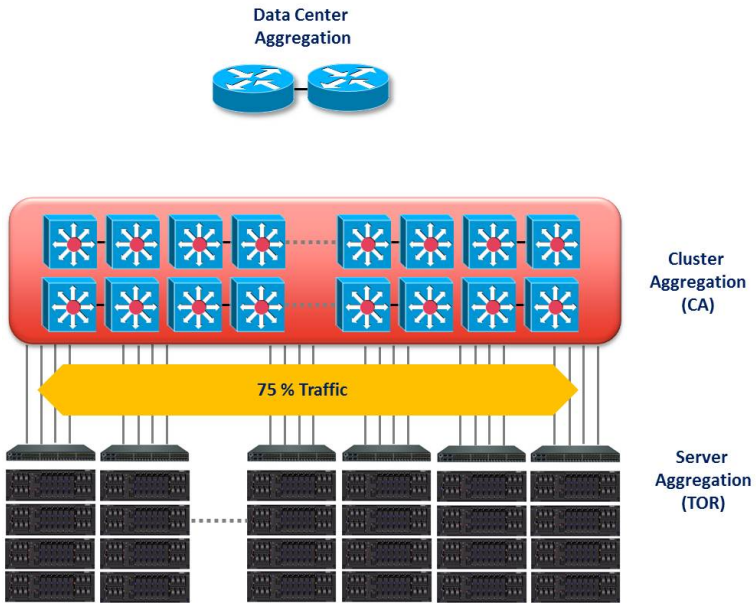


Figure 3: Typical Modern Datacenter Network Architecture

At the lowest level this architecture depicts a layer of server racks (typically blade servers in the latest datacenters), each with a Top of Rack switch (TOR). Typically groups of racks are organized as clusters and each cluster has a centralized Access TOR (or End of Row TOR). Interface rates vary however a recent datacenter might have multiple 10G links connecting TORs down to servers, and multiple 10G or 40G uplinks from the TORs to the datacenter network.

Above the Cluster/TOR layer is a packet-based Cluster aggregation network that provides east-west connectivity between clusters and also north-south bandwidth to the datacenter aggregation routers that connect to the wide area network. However, as revealed in the opening paragraphs, 75% of the traffic is east-west and this is where most problems arise.

Current trends seen in these networks are:

- Bandwidth demand is increasing 2x-4x per year, 75% of which is east-west.
- Top of Rack Switch (TOR) interfaces are scaling to 40 and 100 Gbit/s (and eventually beyond).
- With the growth of virtualization, the number of hosts any given network has to serve is increasing exponentially, creating a need for better network performance. With more virtual machines (VMs) running on servers, networking traffic is growing at a fast pace as each VM is competing for available bandwidth, and hence there is a great need to optimize datacenter structure to permit scalable and affordable solutions for east-west traffic flows.
- Network traffic consists not just of short bursty traffic, but also very high bandwidth, high-persistence data flows that can continue for several seconds or minutes. Typical causes for this highly persistent traffic are virtual machine migrations, data migrations, and the MapReduce function within Hadoop software frameworks.
- Many applications are sensitive to latency.

The packet-based Cluster Aggregation layer doesn't scale well in the face of these challenges. Ideally a network designed with zero oversubscription could provide the capabilities needed, however as we explore this further, serious limitations become evident:

- The cost of a zero-oversubscribed packet-based network is prohibitive. At 40 and 100G interface rates the optical transceivers alone add a large cost element, and the electronics add further cost and high power consumption to the picture.
- A large complex packet based network has inherent latency limitations because of the multiple stages of switches required to create a large fabric. This results in higher network latencies when exactly the opposite is required by applications. It also consumes large amounts of power.
- As the network grows and expands, expansion of the packet based network is expensive in dollars, space, and power consumption. At a minimum, as interface rates scale from 40G to 100G new optical transceivers are required, and the switch fabric itself may also require expansion.

There is a solution to this dilemma.

THE HYBRID PACKET-CIRCUIT DATACENTER NETWORK

Since 1995, super-computing researchers have predicted the need for a hybrid datacenter network consisting of both packet and optical circuit based elements. [(2)] Until now however the bandwidth demand was not sufficient to justify large-scale deployments. Huge growth in cloud computing datacenters is now driving deployments of these hybrid networks and the vision of these groundbreaking researchers is starting to be realized.

A depiction of a hybrid packet-circuit datacenter network is shown in Figure 4.

In this hybrid solution the packet network continues to exist with any-any connectivity between clusters. Its focus shifts however to handle relatively short "front end" bursty data flows.

In parallel with the packet network an Optical Circuit Switch (OCS) Trunk network consisting of a fabric of optical circuit switch elements is also deployed. The role of this circuit-switched network is to switch in as needed to support large persistent flows, thereby freeing up the packet-based network to remove any points of constriction.

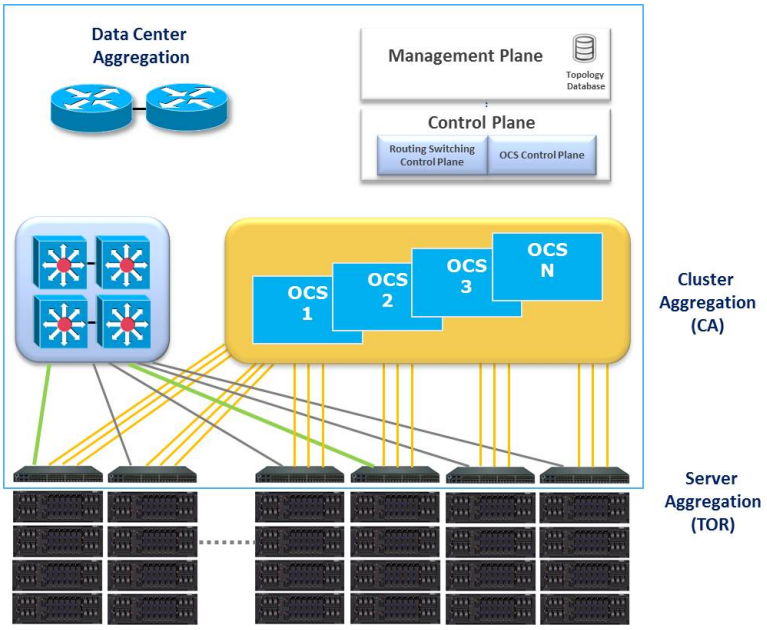


Figure 4: Hybrid Packet-OCS Datacenter Network Architecture

The OCS fabric provides essentially unlimited bandwidth that will scale without upgrade as network speeds increase to 100G and beyond. This is because pure photonic 3D MEMS based OCS solutions such as CALIENT's 320 port S320 are completely transparent to data rate and protocol and use no optical transceivers.

In addition the OCS fabric offers extremely low latency paths (less than 60ns) between TORS providing excellent support for latency sensitive applications.

The setup time of an optical circuit switch such as CALIENT's S320 is typically 25ms (50ms max). This is dictated by the fact that electrostatic re-positioning of micro-mirrors is required to achieve the switching and the laws of physics impose limitations. In the packet world 25ms seems quite high however in a hybrid architecture such setup times are quite acceptable. The reason is that most large flows persist for minutes or more and so the OCS setup time is essentially irrelevant. In the interim period before an OCS connection is made, the packet network will continue to transport the traffic flow and no data is lost.

Overall, this hybrid network provides a scalable, low-cost, high capacity, low-latency, future-proof solution that solves the most significant network challenge in modern datacenters.

A HYBRID NETWORK EXAMPLE

Let's review a high level example of how the hybrid packet-OCS network would handle a high persistence data flow. Figure 5 shows a generic hybrid network architecture

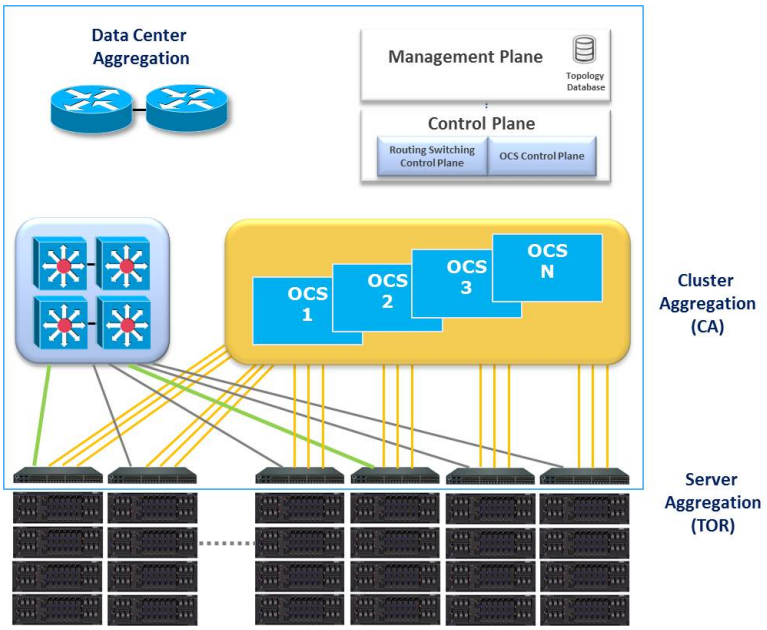


Figure 5 – Starting State

In the starting state all TORS are connected to both packet and optical circuit fabrics. The green lines indicate an active low-level data flow between racks 1 and 4 through the packet network.

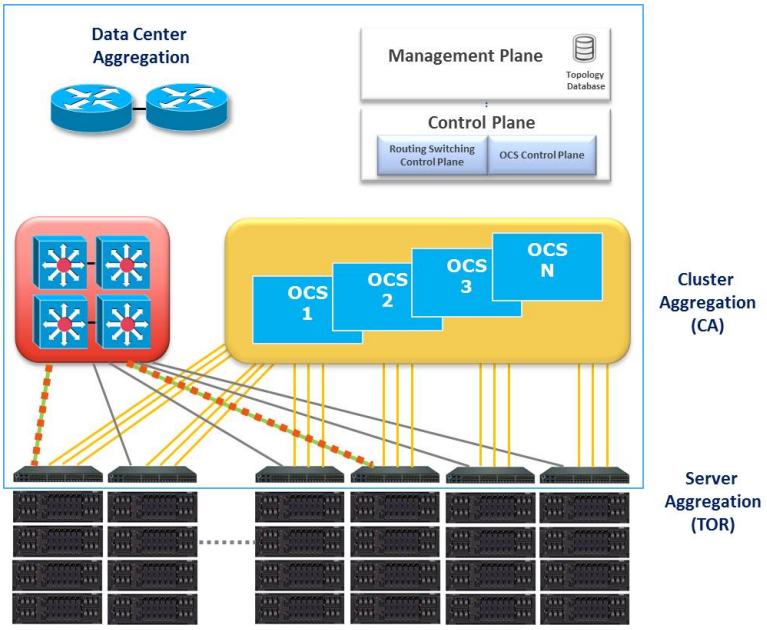


Figure 6 – High Persistence Traffic Flow – Network Constriction

In figure 6, the data flow has now increased to a persistent high-bandwidth stream (e.g VM migration) that is limited by the packet network capacity. Switch buffers are now filling and a higher capacity path is needed in order to allow the data to pass without constriction.

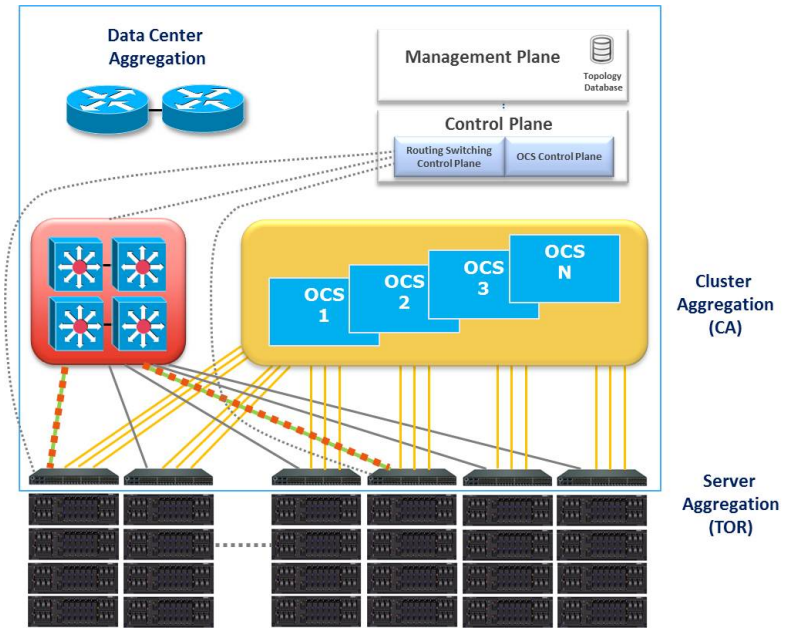


Figure 7 – Packet Network Notifies Management/Control Plane

In figure 7, the packet network notifies the management/control plane that a network constriction exists.

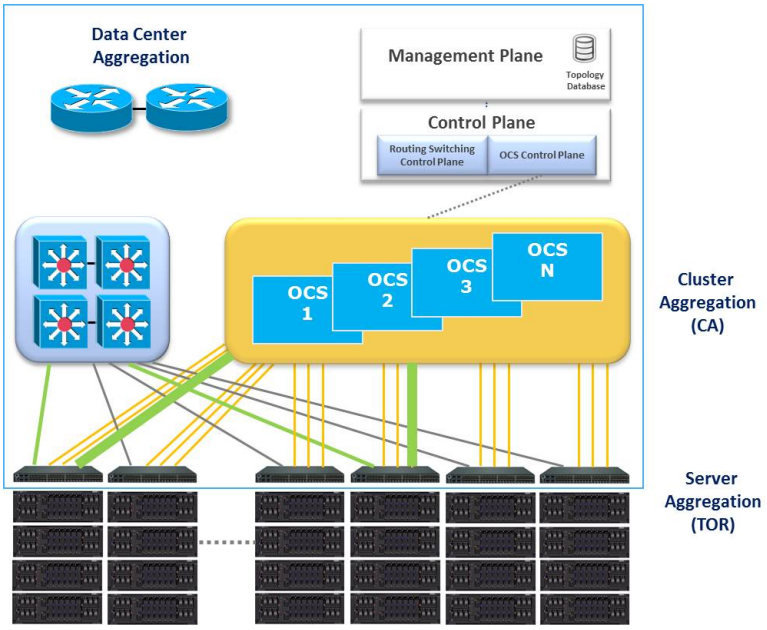


Figure 8 – OCS Fabric opens high capacity, low latency trunk and relieves constriction

In figure 8, the management/control plane responds by issuing a command to the OCS fabric to open a direct optical trunk between racks 1 and 4. The data flow now has access to a very high bandwidth, low latency connection and the packet network constriction is relieved.

Similar activity can me envisaged to be happening between other TORS and the OCS trunk connections can be setup, taken down, and reallocated on demand wherever high bandwidth paths are needed.

The management /control plane can initiate setup and take down of optical paths based on a number of criteria, including response to real-time network demand, time of day, or potentially predictive traffic algorithms.

THE ROLE OF SDN

To complete the solution the hybrid packet-OCS network needs a means to control it. This can be based on a range of solutions from simple scripts to full SDN implementations with high levels of network intelligence.

Figure 9 shows the well-known multi-layer SDN model with application, control, and infrastructure layers.

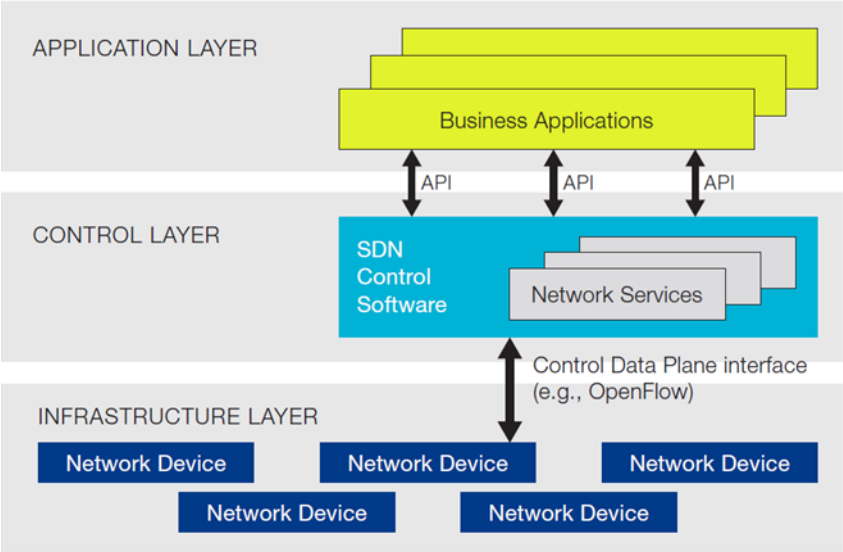


Figure 9 – SDN Model Layers

The important feature is that the packet and optical circuit switches can all coexist together in the infrastructure layer with coordinated control from the upper layers.

How we see this coming together today in the large datacenters is a management plane and a control plane containing both packet and circuit elements as shown in Figure 10.

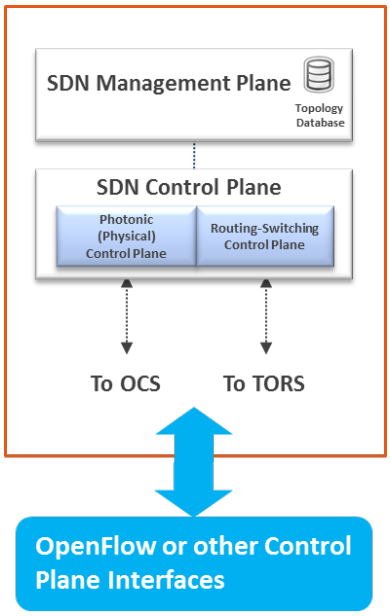


Figure 10 – Datacenter SDN Model Implementation

The Management Plane creates and manages topologies and related configurations, and analyzes the various flows within the network during runtime in coordination with Photonic and Routing-Switching Control Planes.

Based upon operational needs (such as run time triggers, scheduled and cyclic patterns, maintenance activities etc.) it creates new topologies and related configurations and propagates the configurations to the respective control planes for asynchronous execution.

The control planes orchestrate the topology related configuration changes: The photonic engine manages the topology changes across the optical circuit switch fabric while the routing-switching engine further manages the topology changes across various packet-based routing and switching fabric elements.

Both engines maintain the status of execution of each step in the configuration flow and communicate the run time statuses and replies for command processing for topology changes with the management plane.

CONCLUSION

The hybrid datacenter network was first envisioned by researchers in the 1990s and is now being realized in commercial datacenters where huge traffic loads and variable flow patterns are demonstrating the limitations of traditional all-packet based network architectures. The hybrid packet-OCS network philosophy offers significant advantages notably:

- The capability to handle large persistent data flows with unlimited bandwidth at low cost, thereby freeing up the packet network and removing bandwidth constriction.
- Ultra low latency (<60ns) which is very important to modern latency-sensitive applications. In contrast, low-oversubscription all-packet-based networks can have high levels of latency due to the size of the fabric.
- The ability to scale beyond 100G without upgrade as network interfaces are upgraded on TORS and Core Routers. This represents a significant capital cost savings compared with multiple stages of upgrade over time of an all packet based network.

CALIENT is focused on the development and manufacture of 3D MEMS optical circuit switches so our primary expertise area is building the OCS network element hardware and control systems. Our S320 OCS which boasts the highest port density in the industry (320 x 320 Ports) was engineered with a focus on datacenter deployments.

We have collaborated with Datacenter operators and Large Datacenter OEMs on this hybrid network strategy and are currently pursuing options with partners to implement control and application layer functions to manage hybrid packet-OCS networks.

ABOUT CALIENT TECHNOLOGIES

Headquartered in Santa Barbara, California, CALIENT Technologies is the global leader in adaptive Optical Circuit Switching with systems that enable dynamic optical layer optimization in next generation datacenters and software defined networks. CALIENT's 3D MEMS switches have demonstrated years of reliability, with eight years of successful continuous operation. With more than 80,000 optical terminations shipped, CALIENT has one of the largest installed bases of Optical Circuit Switches worldwide. The company designs and fabricates its systems using the state of art MEMS equipment in its own facility located at its corporate headquarters.

CONTACT INFORMATION

CALIENT Corporate Headquarters
CALIENT Technologies
27 Castilian Drive
Goleta, CA 93117
Phone: 805.562.5500
Fax: 805.562.1901

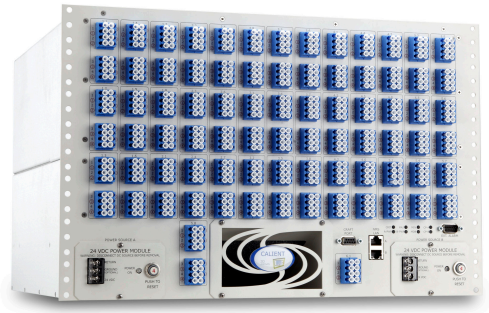
Email: sales@calient.net

Website: http://www.calient.net

REFERENCES

1. Source: Cisco Global Cloud Index (2011 – 2016)
2. Hybrid OCS Datacenter Architecture Technical Papers: http://www.calient.net/resources/hybrid-ocs-datacenter-architecture-papers/

CALIENT'S S320 OPTICAL CIRCUIT SWITCH



The explosion of video, mobile data, and server virtualization is driving the demand for flexible, scalable, high-bandwidth networks. CALIENT's S320 Optical Circuit Switch is a reliable and cost-effective solution for these networks because the technology is transparent to data speed, and is protocol agnostic, thus it offers very high bandwidth and configuration flexibility as networks grow in speed from 10Gbps to 40Gbps and 100Gbps.

Based on field proven 3D Optical MEMS technology that CALIENT has deployed in more than 80,000 optical connections globally, the new S320 Optical Circuit Switch delivers a sweet-spot of high reliability, small form factor, low power consumption & cost, and ease of use that allows the benefits of true all-optical switching to be realized for the first time in a wide range of service provider and datacenter applications.

Applications

The S320 provides the scalable and protocol independent automated fiber interconnect and management infrastructure for a wide range of Datacenter, Service Provider, and Government applications including:

- Flexible, scalable on-demand resource optimization in enterprise and cloud computing datacenters
- Rapid disaster-recovery from multiple network failure scenarios in any optical network
- High port count colorless, directionless and contention-less (CDC) ROADMs in fiber-optic service provider networks
- Physical fiber network virtualization in metro SDN Service Provider networks
- Fiber To The Home (FTTH/FTTP) network automation – automated service activation & testing
- Sharing of high-value testing resources in lab automation & Cyber-range applications

Features & Benefits

- Small Size: 320 Ports (Tx/Rx pairs) in 7RU Chassis (LC Connectors)
- Low Power Operation: 45 Watts typical
- Low Cost: Supports deployment in datacenter, service provider, and government networks
- Ultra-low Latency: All-optical connectivity adds no latency.
- Scalable: Supports all data rates to 100 Gbps and beyond
- Reliable: Based on proven 3D MEMS design deployed in over 80,000 fiber terminations globally
- Simple to use and integrate: GUI-driven, EMS-ready, supports TL1, SNMP, CORBA, OpenFlow
- Low loss: 2 dB typical insertion loss
- Built-in power monitoring: Every in/out fiber is monitored providing powerful network diagnostic capabilities.