CALIENT
OPTICAL SWITCHES

VPod: Boundary -Less Data Center

Whitepaper | CALIENT Solution Brief

Poor data center compute resource-utilization is one of the primary sources of waste and inefficiency in computing today. Disparity in utilization contributes to building data centers for peak capacity based on aggregate of average utilization. Data Centers often only use less than 40 percent of their available computing cycles, according to general industry estimates, and yet compute resources (CPUs, GPUs, and storage) represent 85% of the electronics cost in a data center.

A new virtual pod (vPod) data center architecture, enabled by Optical Circuit Switching can deliver significant improvements in the utilization of server and storage resources resulting in improved computational performance and operational efficiency.

How Current Architectures Limit Resource Utilization

Typical large modern data centers are organized into physical pods - the pod architecture can be likened to having multiple mini data centers. These pods are usually the increment of upgrades and typically have common failure modes like power or common cooling. Within the pod, servers and storage are organized into modular racks and rows of racks.

In data centers today, architects have to over-build each resource pod with compute resources to handle peak demand for four main reasons:

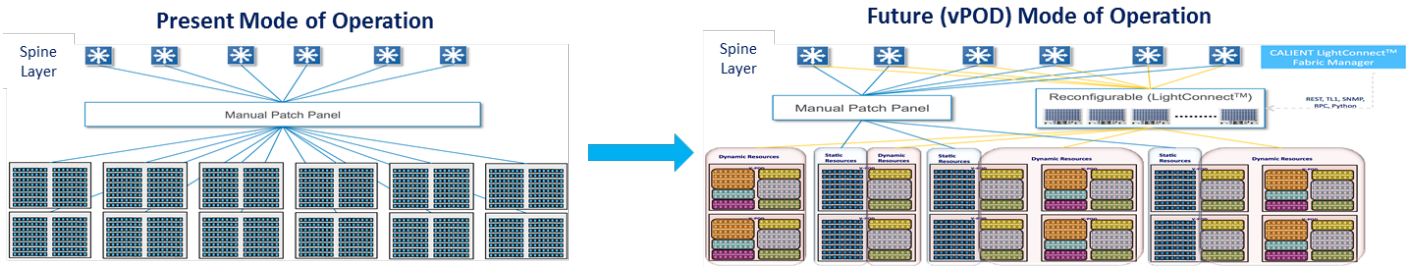
- Variability in resource demands by different services and applications. Estimated demand, actual demand, and allocated resources are not perfectly matched.
- Capacity planning & resource allocation are based upon estimation with built-in over-dimensioning and long provisioning cycles to mitigate risks.
- Virtualization limitations across networks, for example, Layer 2 and 3 domains can cause resource fragmentation, resulting in degraded performance. The increased latency through multiple L2/3 network hops alone degrades application performance.
- Compute pods typically require tightly coupled individual nodes. A single compute workload requires frequent communication among nodes. This implies that the pod shares a dedicated network, is densely located, and probably has homogeneous nodes.

One current practice to improve compute utilization in data centers is to use load-balancing techniques to pick up a resource-constrained workload and move (typically generates very large east-west “elephant” flows between pods) it to a different pod where resources are available, but this requires policies and significant planning, and it’s definitely not available “on-demand”.

The Calient LightConnect™ Fabric VPod Architecture

How can we achieve significant improvements in compute resource utilization in a practical, manageable way?

Imagine if compute resources could be easily reassigned between pods at Layer 1, allowing workloads to grow without the need to massively overbuild?



The application workloads in most Cloud data centers contain the analytics to enable pre-emptive reassignment of resources so that part exists. But to make it happen we need a way to reconfigure the optical layer connectivity in the data center to physically reconnect compute resources.

This is precisely the function of CALIENT's LightConnect™ Fabric, which allows pod resources to be reassigned at the optical layer in response to the needs of workloads. It does so with a new virtual pod or vPod architecture.

The LightConnect™ Fabric is a flat optical-circuit-switched network layer that sits between TORs or end-of-row switches and spine switches above the pod. It layers across multiple pods to allow sharing of resources between pods across the data center.

Some of the resources in each pod can be static (static resources are permanently assigned to each physical pod) while other resources are dynamic (dynamic resources can either be used within the physical pod or allocated to other pods) – connected through optical circuit switched paths to the spine switches rather than through manual patch panels. So, the dynamic resources can be considered to be resource pool available to the whole data center.

As compute demand grows within a pod it can be reconfigured to borrow resources from other pods. Depending on the granularity of the LightConnect™ Fabric, it can borrow rows of racks or individual racks. vPods can be either partially or fully re-configurable. In the partially re-configurable case a percentage of the compute resources are directly connected to spine switches while the remaining resources are connected to the spine switches through an intermediate LightConnect™ Fabric layer. In the fully-reconfigurable case all compute resources are connected through the LightConnect™ Fabric.

Because this is all happening at Layer 1, we're creating a larger compute capability within the pod and allowing workloads to expand without moving them or splitting them across domain boundaries. Conversely a pod can "give up" compute resources to another pod if it has available capacity. This is the vPod concept.

The power of this approach is that it is no longer necessary to over-build every pod with resources to handle worst case anticipated demands. Instead, we can overbuild by an average and lower amount across the entire data center and reallocate this on demand between pods. This results in significant improvements in compute resource utilization - data center can be operated at much greater capacity and operational efficiency (which represent the majority of the electronics cost in a large data center).

Calient Core Technology

CALIENT's Optical Circuit Switch is a large port count all-optical (OOO) switch that establishes, monitors and switches physical layer connections between single-mode optical fibers using Micro-Electro-Mechanical Systems (MEMS) based optical switching. Connections are made between fibers carrying signals with any data rate or protocol. Any input fiber on the S-Series OCS can be connected to any output fiber making a fully non-blocking switch fabric.

Light is directed from the input fibers to the output fibers using arrays of tiny silicon mirrors that are fabricated using the proven CALIENT MEMS process. An optical signal transmitted through the OCS passes through three sections of the switch core: the input collimator array, which directs the light from each input fiber to its input mirror; the mirror matrix, an array of MEMS input mirrors and an array of MEMS output mirrors; and the output collimator array, which couples light from each output mirror back into its output fiber. High-quality mirrors and collimators and precise electrostatic control of the position of each mirror, enable typical switch times of less than 50 ms and optical loss that is less than 3.0 dB for CALIENTs complete line of optical circuit switches.