WHITE PAPER

Building Simple, Scalable, and Seamless IP Clos Networks with ArcOS Switches and Routers

ARRCUS
NETWORK DIFFERENT™

Contents

Introduction

5G, edge computing deployments continue to ramp up, driven by explosive growth in rich media content, hyper-connected users/devices, and mobile traffic. The pressure to deliver high resiliency and control costs is unrelenting. The 5G scale and performance requirements are driving a re-think of the entire underlying network infrastructure, whether it be for on-prem, edge, or multi-cloud deployments. In fact, compute (applications) and data have moved closer to the end-consumer leading to the notion of "micro-data centers". So, businesses are looking at cost-effective ways to build out modern network infrastructure without compromising on their requirements.

A key building block of this transformation is the popular "IP Clos" network design for the data center. This design was named after Charles Clos (pronounced "cloh") who introduced the concept of a non-blocking, multi-stage circuit switching architecture for telecommunication systems in which there are multiple paths between each layer of devices without the inefficiencies of n-squared laws. The architecture was incorporated into network designs initially in the data centers (particularly at the hyperscalers and global enterprises) driven by the need to address modern application needs around traffic patterns ("East-West" traffic) and scale. From a networking technology point of view, that meant moving to routing (Layer 3) and away from proprietary technologies such as Layer 2. Further, these designs have now extended to every part of the network (Central Office, CDN PoPs, Edge, Backbone, etc.)

Merchant silicon advancements, especially in the area of high-performance switching and routing, have given companies the ability to deploy open, inexpensive hardware platforms. This allows them to break free from proprietary silicon-based switches and routers without compromising on the performance and scale they require for their ideal network. From a software perspective, merchant silicon-based switches have enabled open, routing-centric designs to become the norm in place of the legacy, siloed L2-centric designs. Arrcus's ArcOS enables customers to have the freedom to choose from a wide variety of open hardware platforms to achieve the IP Clos design they want while also providing them the flexibility to meet the ever-changing needs of data centers. ArcOS was built, from first principles, as a 64-bit internet-scale network operating system with a focus on routing.

In this document, we will discuss some of the factors that go into IP Clos designs including variants such as the "Fat-Tree" variant, calculating the number of devices, routing, automation and telemetry. We will also highlight how first principles-based commercial-grade networking software is essential to building an IP Clos that is simple, scalable, secure, and seamless for solving modern day challenges.

The Evolution of the IP Clos Fabric

Network architectures are becoming increasingly routing-centric, moving from rigid and brittle to agile and elastic. Prior to 2015, most of the datacenter architectures were L2-heavy i.e. siloed and rigid. However, starting first with the big cloud providers and then many of the big enterprises, the datacenter architectures became spine-leaf IP Clos designs with routing to the Top of Rack (ToR) whether that was IP-based or overlay-based fabrics. At the same time, merchant silicon-based switches became the leading part of the data center.

Clos networks started with 1G servers and 10G spine-leaf links and over time have transitioned to 100G servers with 400G spine-leaf links. Arrcus has enabled customers with flexible solutions that span media speeds from 1G to 400G. As an example, if we were to build a Broadcom-based data center, the merchant silicon switches would consist of high-density Tomahawk 3-based switches for spines and super spines along with Trident 3- and Trident 4-based switches for the Top-of-Rack layers. ArcOS supports all of these options.

Figure 1: Broadcom switching and routing lineup

The scale-out design of IP Clos networks has advantages. Refer to Figure 2 below which depicts a two-tier Clos design. Instead of investing in a large chassis upfront and filling it with line cards later when capacity is needed, operators can purchase and deploy switches horizontally as their needs change without the higher upfront investment. Additionally, based on the scale needed, if proper devices are chosen for spine and leaf layers, then a single SKU can be chosen for each layer. This allows for simplified troubleshooting, spares for replacement, and additional nodes for growth. Compared to the large reduction in bandwidth when a chassis goes down, a spine in an IP Clos fabric going down will have less of an impact. Lastly, using fixed format devices and just a few SKUs in your fabric allows you to take advantage of economies of scale.

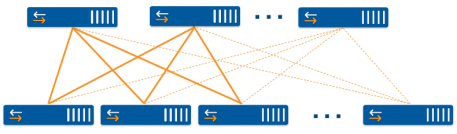

Figure 2: Two-tiered spine-leaf design

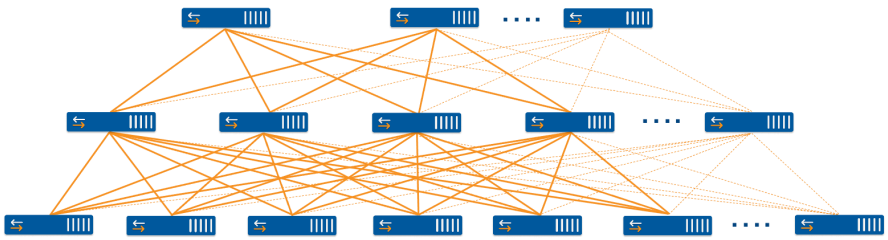

Figure 3: Three-tiered spine-leaf design

The oversubscription rate in the design will dictate how wide the fan-out is as well as the maximum limit of scaling horizontally at each layer. For example, if the leaf node has 32 ports of 100G with 8 ports facing the fabric and 24 ports facing the workloads (1:3 oversubscription), then, at most, you can scale out to 8 spines (each leaf connects one port to each spine). If those spines have 32 ports of 100G, then the fabric can have a maximum of 32 leafs. This, for example, will yield a fabric with 8 spines that supports 32 (leafs) x 24 (workload ports/leaf) = 768 workload ports of 100G.

To determine the fabric design:

1. Start with the maximum number of workload ports the fabric will need when fully scaled out
2. Determine the number of leaves needed for the fully scaled-out fabric
3. Define your oversubscription ratio
4. Determine the number of spines needed

ArcOS supports breakout interfaces (4x10G, 4x25G, 2x50G depending on the particular hardware type), which allow flexibility, especially in the Top-of-Rack/leaf layer, in terms of investing in a higher-capacity device and using breakout cables to handle various speeds of today. As the speeds of workload interfaces increase, higher capacity (lower fan-out count) breakout cables can be used until the workload can connect directly at the native speed of the interface.

Depending on the scale needed, some data centers deploy a three-tiered design as shown in Figure 3, which depicts the super-spine, spine, and leaf layers.

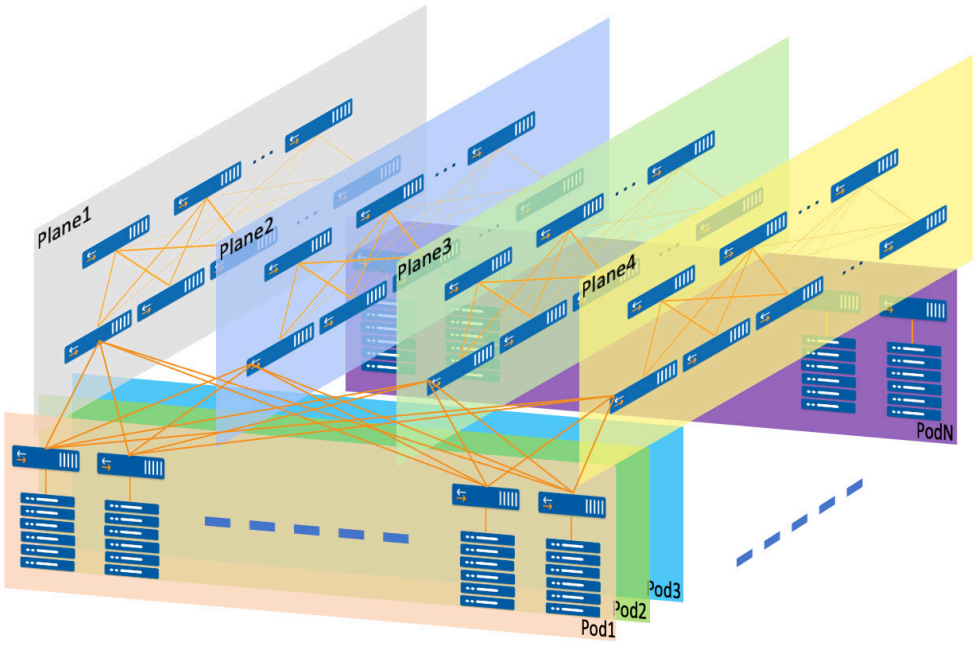

Figure 4: 3-layer Fat-Tree server pod design

Some companies (such as the hyperscalers, OTT) employ a variation of Clos called Fat-Tree that drives massive scale at lower costs. Multi-plane Fat-Tree Clos enables seamless workload communication and availability across the entire fabric. Refer to Figure 4.

This approach offers the following key benefits:

- Network units of "server pods" with highly available fabric-wide connectivity
- Ability to scale compute capacity by simply adding server pods
- Ability to scale intra-fabric network capacity by adding spine switches to spine planes
- Means to adjust the oversubscription rate based on the number of spine switches on the spine planes

To give an example of the massive scale of this design:

- The top tier in the spine plane could have 32 32x100G devices with each port split into 2x50G connecting to 64 32x100G devices on the bottom tier of the spine plane. This would allow 64 server pods to connect to the corresponding spine plane devices via the 32 ToR devices in each pod.
- Each device on the bottom tier of the spine plane has 16x100G ports used (32x50G ports) for connecting to the top layer of the spine plane and 16x100G ports used (32x50G ports) for connecting to the 32 ToR devices in the corresponding server pod.
- The links between the ToR devices and the servers could be 10G, 25G, 50G, or 100G depending on the needs. Assuming a 32x100G device is used for the ToR device in a server pod with each device having 4 50G uplinks, that leaves 30 ports of 100G on each ToR to connect to servers.
- There are 32 ToRs in each server pod, which means, in total, there would be 32x30 = 960 server-facing 100G ports within each server pod. Across the whole fabric of 64 server pods there would then be 64x960 = 61,440 server-facing 100G ports.

Depending on scale, performance, and availability requirements of the business any one of the above design choices can be implemented and ArcOS supports all of them.

Deep-buffer Router IP Clos

Deep-buffer routers are also being deployed as edge devices in conjunction with an IP Clos fabric or even as part of the fabric. This is especially useful in specialized environments such as storage centric application PoDs, bursty big data application PoDs, or MPLS/SR/SRv6 edge-cloud PoPs where Clos designs have permeated. ArcOS supports state of the art deep-buffer routers (based on Jericho+/Jericho2) with port speeds ranging from 10G to 400G allowing customers to mix and match any merchant silicon platform to suit their needs with ArcOS as the common software across switching and routing environments.

There are also new alternatives to huge, siloed modular chassis platforms that leverage the IP Clos design. These are called Virtualized Distributed Routers (VDR), some call them disaggregated distributed chassis, which are effectively specialized deep-buffer merchant silicon-based IP Clos fabrics where the control plane is de-coupled from the Clos and run from virtualized external compute. Arrcus's VDR enables a single logical management & control plane, low cost per port, high availability, control plane hardware at consumer PC pricing, and adaptability to run on any silicon/hardware while retaining cell-based forwarding capability which are typically used in chassis-based router platforms. You can learn more about Arrcus's uniquely differentiated and massively scalable VDR solution **here**.

IP Clos Routing Control Plane Considerations

The control plane is the brains of any distributed system. The key control plane attributes that drive modern IP Clos environments are that they need to be simple, scalable, secure, and seamless.

- Simple to deploy and operate with support for standards-based protocols (e.g. BGP, IS-IS, etc.) with a hook to standards-based configuration/provisioning data models (e.g., OpenConfig/YANG)
- Scalable to hyperscale requirements both from a control-plane perspective and from a fabric-wide policy and automation perspective
- Secure from a system-wide perspective along with deep visibility into routing datastore and analytics that enables operational insights and real-time system alerts
- Seamless from a routing connectivity point of view across physical (switch/router), virtual (VM/container), and cloud environments

ArcOS supports routing protocols that are popular in the IP Clos deployments including OSPF, IS-IS, and BGP. While multi-tenant architectures such as EVPN Fabrics typically tend to separate underlay (IGP) and overlay (BGP/EVPN) protocols, eBGP is now the most common design choice for hyperscale IP Clos networks due to its simplicity in operations and troubleshooting, massive session scalability, incremental updates, no flooding, reliable transport and per-hop TE unequal cost multipath (UCMP). Taking this further, Arrcus is leading a standards-based initiative to further enhance IP Clos scale, convergence, and simplicity with BGP Link State Vector Routing (LSVR).

For customers that need layer 2 domains to stretch across their data center, ArcOS supports EVPN. There is flexibility for the choice of the underlay between ISIS, OSPF, eBGP, and iBGP. To learn more about EVPN support in ArcOS, refer **here**.

Extending IP Clos to Hybrid Cloud Environments with Arrcus Solutions

Public cloud providers are extending the cloud infrastructure to on-premises data centers and to edge environments with products such as AWS outposts, Google Anthos and Azure Sonic/Arc while promoting the ability to deploy and consume application workloads with ease. However, the majority of these solutions do not integrate with the data center IP Clos posing yet another disjointed solution for IT to manage. With Arrcus's Multi-cloud Networking solution as an extension of the core IP Clos architecture, companies can deploy a deeply integrated hybrid cloud with a node in the IP Clos as border leaf bookended with a node in the virtual private cloud (VPC) of the cloud providers. You can learn more about Arrcus's MCN solution **here**.

Driving Operational Simplicity & Consistency through Automation

As data centers scale out the only way to continue to operate at scale is through automation of provisioning, policy, and process. For Day-0, ArcOS supports Zero Touch Provisioning (ZTP) which provides the ability to run a boot script on the first boot of the device. For Day-1 and Day-2, ArcOS provides NETCONF, RESTCONF, ArcAPI (python APIs), SNMP, and Ansible support. The software is OpenConfig compliant allowing operators to use vendor-neutral YANG data models to program devices. Based on Debian Linux, ArcOS is an open system allowing operators the flexibility of installing third-party applications using Debian packages. To learn more about automation with ArcOS refer **here**.

Deep Visibility & Advanced Analytics with ArcIQ

ArcOS has been designed with streaming telemetry in mind. Data is modeled as JSON schemas and sent over Kafka as well as gNMI. Using ArcIQ, Arrcus's deep visibility and analytics platform, ArcOS devices can be monitored with ease for health and event correlation. To learn more about how ArcIQ provides enhanced visibility, refer **here**.

ArcOS IP Clos Differentiators

ArcOS is the industry-leading software for IP Clos use-cases providing the following benefits:

- Architected from first principles as a microservices-based networking software with scale, performance, and availability as key pillars
- Supports IP Clos on both switching and deep-buffer routing merchant silicon hardware
- Supports a highly available infrastructure with Clos-wide resiliency, process restartability, and rapid software upgradability
- Seamless integration with modern automation frameworks such as Ansible, Terraform etc.
- Seamless connectivity between on-prem and cloud environments to form a hybrid cloud environment using Arrcus's MCN solution

There are many factors to consider in building a Clos design, as covered in this paper. ArcOS's superior capabilities allow enterprises and service providers to seamlessly build massively scalable, cloud-ready data center infrastructure while offering operational simplicity and deployment flexibility.

Learn more

Visit **www.arrcus.com** to learn how Arrcus can meet your organization's evolving business needs with its IP Clos solutions.

Network Different – with Arrcus

About Arrcus

Arrcus was founded to enrich human experiences by interconnecting people, machines, and data. Our mission is to democratize the networking industry by providing best-in-class software, the most flexible consumption model, and the lowest total cost of ownership (TCO). The Arrcus team consists of world-class technologists who have an unparalleled record in shipping industry-leading networking products, complemented by industry thought leaders, operating executives, and strategic company builders.

The company is headquartered in San Jose, California.

For more information, go to www.arrcus.com or follow @arrcusinc.

www.arrcus.com **2077 Gateway Place, Suite 400, San Jose, CA**