Cisco's Massively Scalable Data Center

Overview

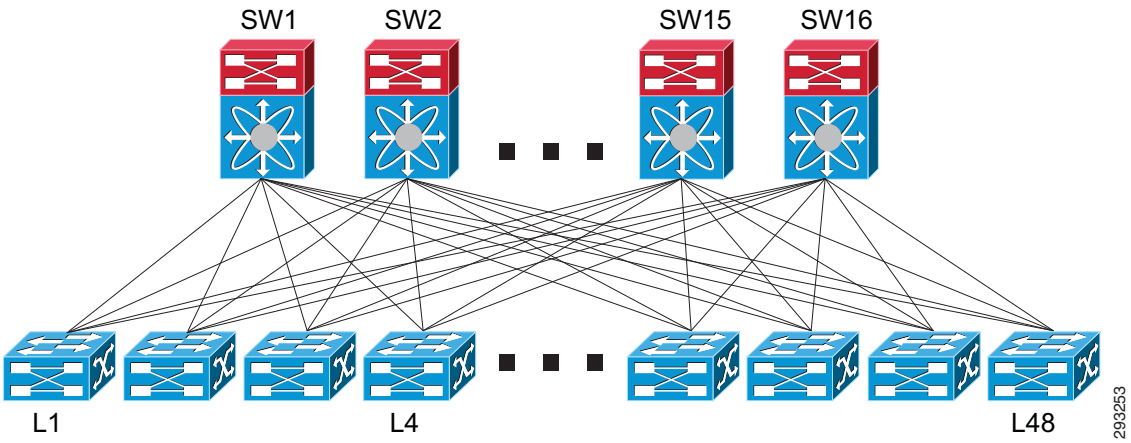
Network Fabric for Warehouse Scale Computer

Cisco Massively Scalable Data Center Reference Architecture

The data center network is arguably the most challenging design problem for a network architect. It is in the data center which runs mission-critical applications for the business. The network fabric that enables the communication between the server nodes that run these applications are therefore mission critical. With advances in distributed computing, the data center bottleneck has once again shifted to the network and modern design methodologies employed to overcome these challenges. Cisco's Massively Scalable Data Center Reference Architecture is an implementation based on Cisco's Massively Scalable Data Center Framework of a data center network fabric that can cost effectively accommodate 12, 288 at 10Gbps with a 1:1 oversubscription ratio. The total switching capacity at scale is 160Tbps.

Figure 1 illustrates the conceptual view of the reference architecture. The reference architecture implements a spine-leaf class multi-root topology three-stage Clos topology.

Figure 1 Cisco MSDC Reference Architecture—Conceptual View



Infrastructure Components

The Network

The network topology is commonly referred to as a spine-leaf topology based on Clos network architecture. This architecture has proven to deliver the high bandwidth, low-latency, non-blocking server-to-server connectivity that is required for massively scalable data center.

The system was built using the Cisco Nexus 7000 series as spines and Cisco Nexus 3000 as leafs. The control plane of the network fabric is a hotly debated topic. Cisco MSDC reference architecture uses a set of Layer 3 protocols for the control plane. The network components are described in Table 1, Table 2, and Table 3.

Spine

Table 1 Spine

Spine Configuration/Device Count	2-way Spine	4-way Spine	8-way Spine	16-way Spine
Number of Nexus 70xx	1x 7018, 1x 7010 1x 7010, 1x 7009	2X 7009, 7010	8, 7018, 7010, 7009	16, 7018
Number of F2 Blades/Chassis	21X N7K-F248XP-25	7/8, N7K-F248XP-25	7/8/16, N7K-F248XP-25	16, N7K-F248XP-25
Number of Ports used for leaf per chassis	384 (7010), 336 (7009)	384 (7010), 336 (7009)	768 (7018), 384 (7010), 336 (7009)	768 (7018), 384 (7010), 336 (7009)

Leaf

Table 2 Leaf

Spine Configuration/Device Count	2-way Spine	4-way Spine	8-way Spine	16-way Spine
Number of Nexus 3064	8	48	96	192
Number of Port facing Fabric	2	32	32	32
Number of Ports Facing Servers	1 + 5	32	32	32

Note 7010= 384 ports. 1:1 oversubscription. Ports*# spines / ports per leaf= 384*2/32= 24 (2 way), 384*4/32= 48, 8 way = 96.

Control Plane

Table 3 Control Plane

Layer	Functionality
Layer 2	Local MAC learning at the Leaf VM's assigned MACs via the Hypervisor Each VM in its own VLAN (/30) to a respective SVI on the upstream switch Physical link trunked to upstream switch Changed MAC aging timer to be: 1500s
Layer 3	ARP caches fully resolve BGP and OSPF used in conjunction with ECMP for load distribution Intra-Leaf: 1 L3 lookup Inter-Leaf: 3 L3 Lookups {Max number of hops in a 3-stage folded Clos architecture} Default route at Leaf for all the VMs Any Static Routes? No. All connected routes and protocol learned routes EBGP Setup: 1 ASN/Leaf, 1 ASN for all Spines (additionally running BFD w/ N3K & N7K default timers), only redistribute connecteds via normal route-maps OSPF Setup: Spine is Area 0, Each Leaf was ABR, no redistribution.
TCP	Standard MTU Cwin set to 10 Delayed ACKS disabled RTO set to 200ms.

Server

Cisco MSDC reference architecture used a 512 virtual node cluster hosted on 40 physical nodes and Ixia emulation. The server hardware is shown in Table 4.

Table 4 Server

Server Path	Specification
Chassis	C200-M2
Memory	48GB
Processor	E5620 2.5GHz dual quad-core
Disk	1 x 500GB SATA 3 x 1TB SATA
Network	Intel X520 NICs with dual port 10GE cards
Server OS Image	Hadoop Nodes: CentOS 6.2 Control: Ubuntu 12.04 LTS Hypervisor KVM

Management

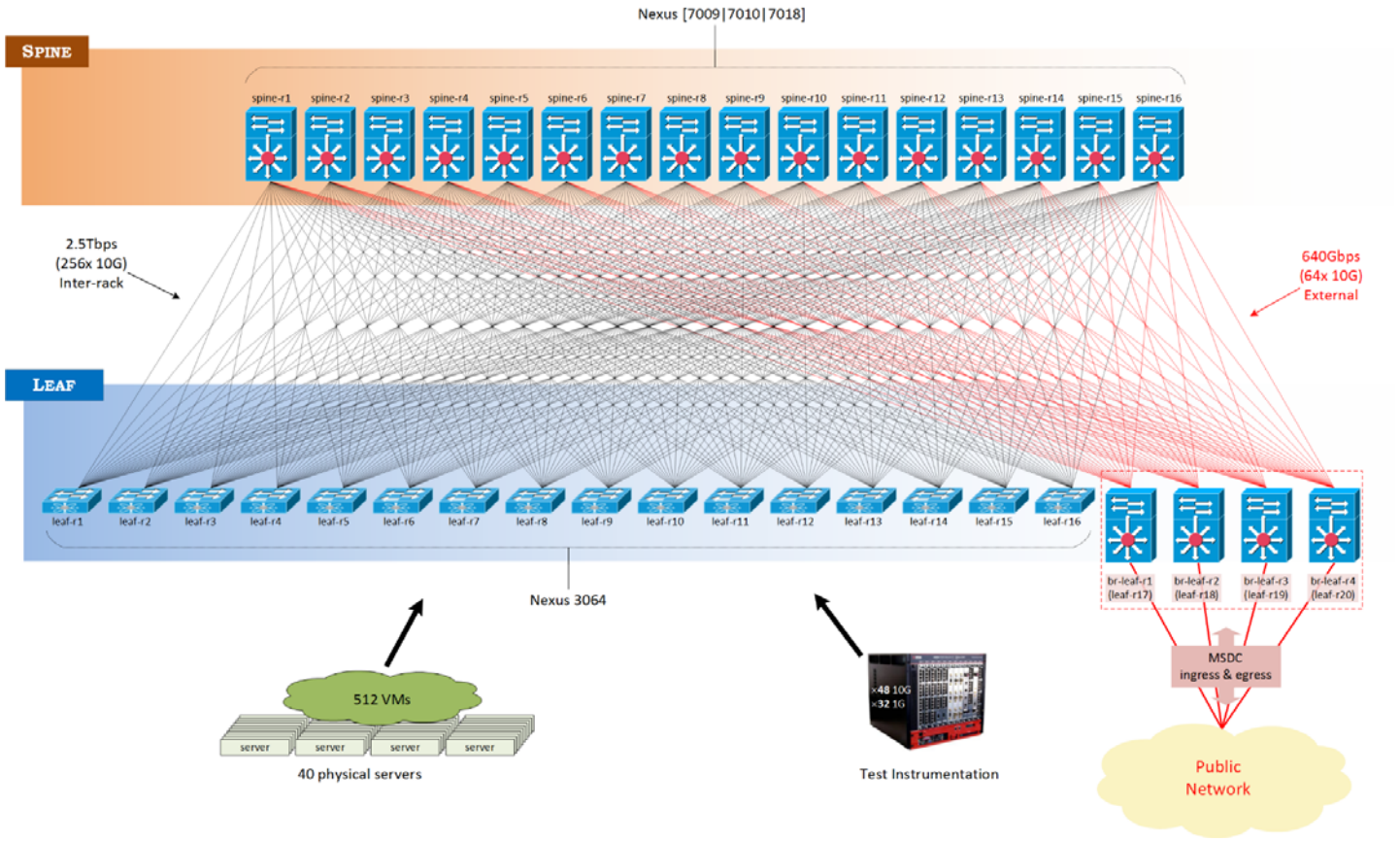
Cisco MSDC architecture uses multiple open source toolsets combined with built-in management features of the Cisco platform to manage and monitor the network fabric. The entire management framework is described in Table 5.

Table 5 Management

Package	Primary Purpose
Nagios v3.2.3	Pull-based host and switch monitoring, graphing v3.2.3 with CentOS with EPEL packages v1.4.15 plugins and Cisco custom plugins gearmand .025 and mod_gearman 1.3.6 for distributing checks to worker nodes nrpe 2.12 for host-side checks pnp4nagios 0.6.16 for graphing check data
Ganglia v3.1.7	cluster monitoring and app-specific (e.g. Hadoop and Java/JMX) monitoring
Cobler v2.2.2	Initial provisioning of baremetal hosts, DHCP for kickstart and POAP
Graphite 9.10 and Collected v5.1.0	Push based monitoring of host and VM statistics. Used for feeding data into graphite **Graphite used for real-time graphing from host and Cisco Nexus switches. Cisco Nexus switches use custom python daemons.**
Puppet v2.7.11 on master and management node Puppet v2.6.12 on CentOS	Configuration management, software deployment, VM spinup
Nexus 3048/64 On Devices	POAP Native Python interpreter support in NX-OS

Figure 2 shows the conceptual view of the reference architecture. The reference architecture implements a spine-leaf class multi-root topology three-stage Clos topology.

Figure 2 Cisco MSDC Reference Architecture—Physical View



Bill of Material

Table 6 Cisco MSDC Reference Architecture Bill of Material

Part Number	Description	Quantity
N3K-C3048TP-1GE	Nexus 3048	16
N3K-C3048TP-1GE	Nexus 3048	16
N3K-C3064PQ-10GE	Nexus 3064	22
N5K-C5548P-FA	Nexus 5548P	2
N7K-C7009-BUN2	Nexus 7009	7
N7K-C7009-FAB-2	Nexus 7009 Fabric Module	35
N7K-C7010-BUN	Nexus 7010	11
N7K-C7010-FAB-2	Nexus 7010 Fabric Module	32
N7K-C7018	Nexus 7018	2
N7K-C7018-FAB-2	Nexus 7018 Fabric Module	8
N7K-F248XP-25	Nexus 7000 F2 Series 48 Port 10GbE	36
N7K-SUP1-BUN	Nexus 7000 Supervisor1	21
R200-1120402W	UCS C200 Series server	40
OPTIXIAXM12-02	Ixia Chassis	1
LSM10XM8-01	Ixia Line Cards	4