
TOMORROW
starts here.
CISCO
Cisco live!

Next Generation Computing Architectures for Cloud Scale Applications

BRKCOM-2602

Steve McQuerry, CCIE #6108, Manager Technical Marketing

#clmel

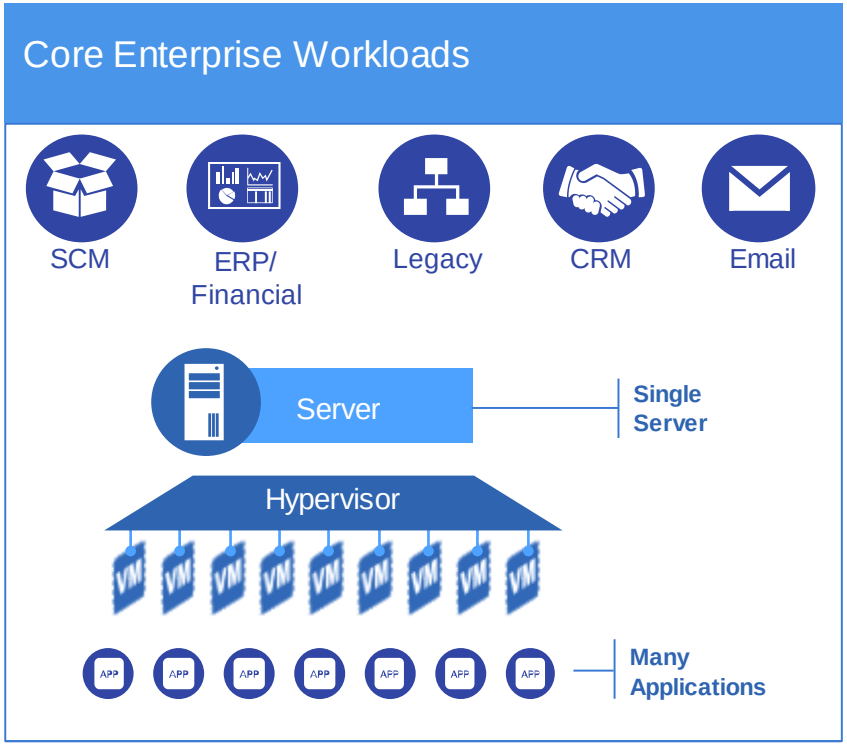
Cisco live!

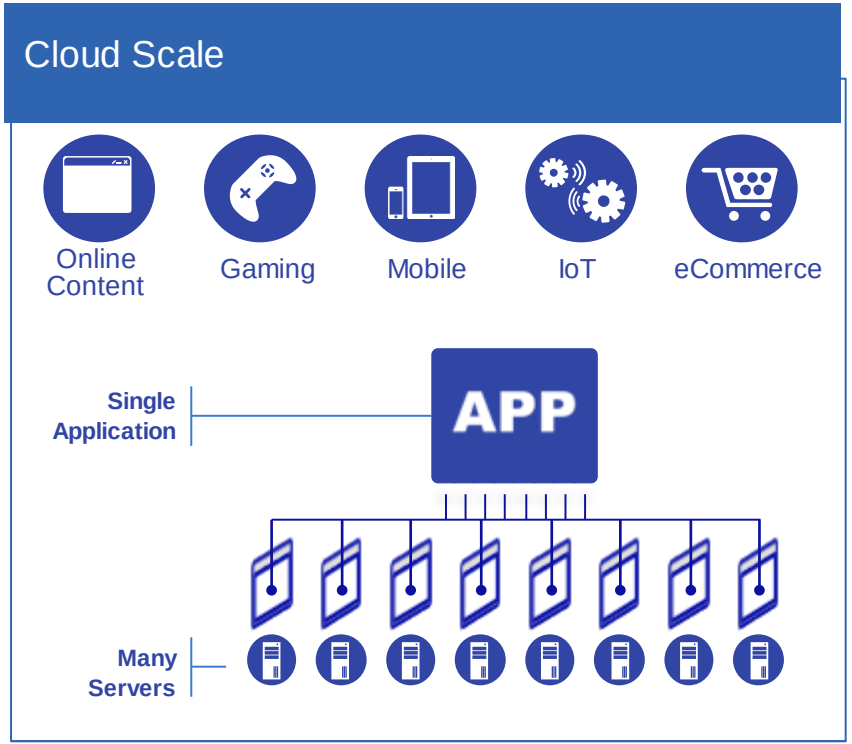
Agenda

- Introduction
- Cloud Scale Architectures
- System Link Technology
- Mapping Application Architecture to Infrastructure
- Scaling and Maintaining the infrastructure
- Incorporating Storage into the Infrastructure
- Conclusion

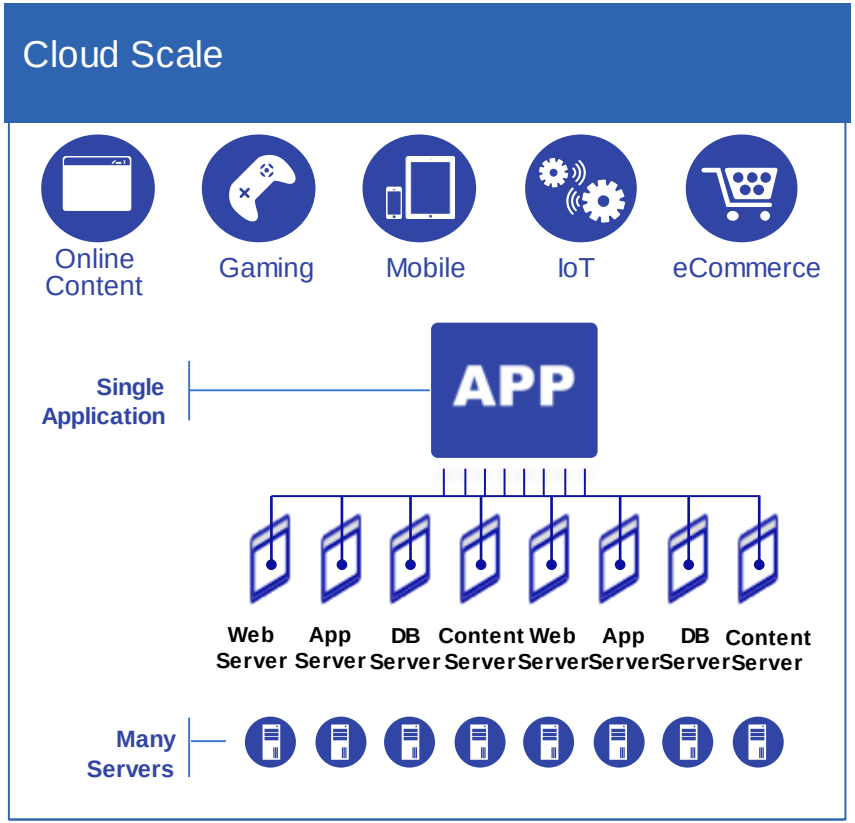

Cloud Scale Architectures
Cisco live!

Cloud-Scale Inverts Computing Architecture



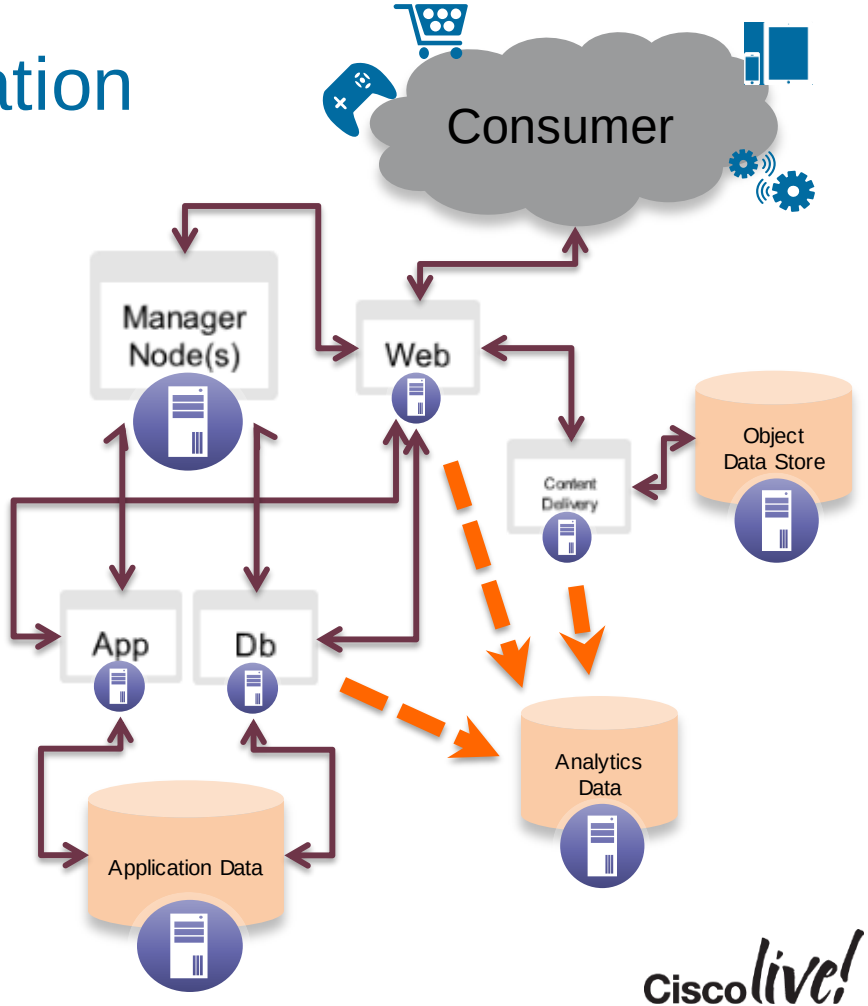


Cloud-Scale Application Components



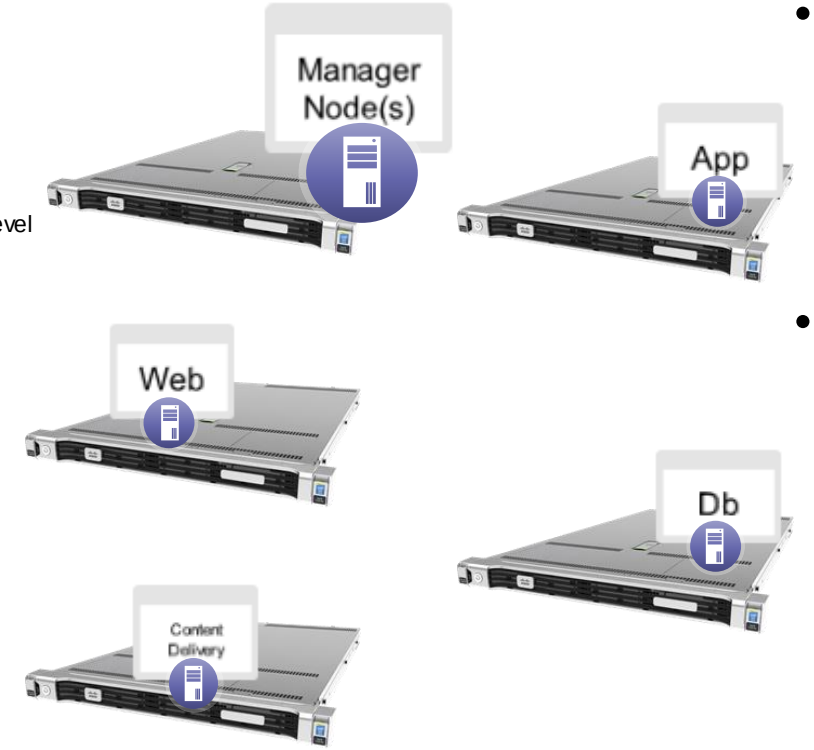
Sample Cloud Scale Application

- Cloud scale applications distribute the workload across multiple component nodes
- These nodes have various system requirements
- Distributed Components report into manager nodes
- Manager nodes note availability, farm out workloads and may receive data from worker nodes
- Worker nodes provide the bulk of cloud scale applications



Compute Infrastructure Requirements

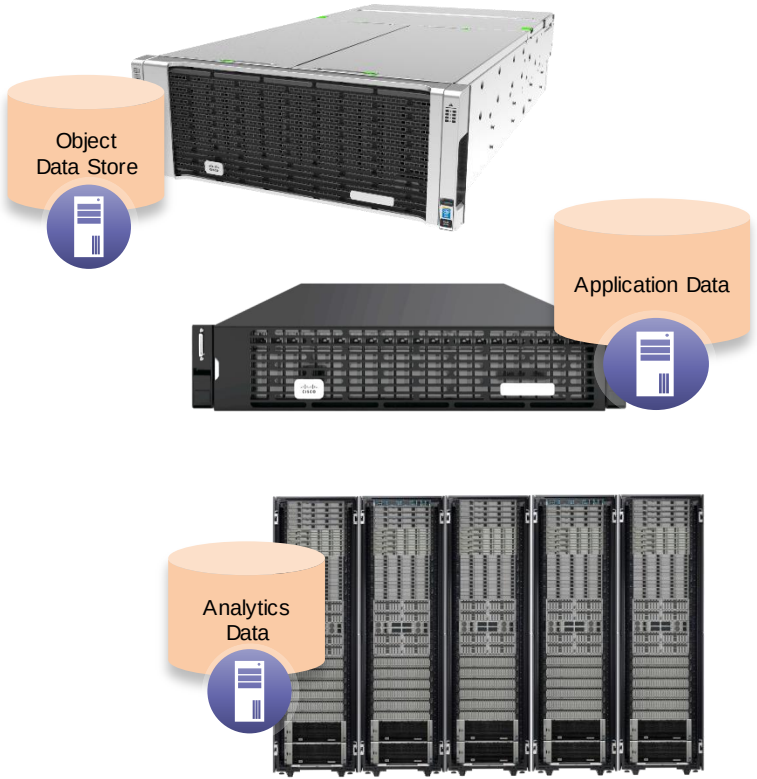
- **Manager Node**
  - Dual-Socket/8-16 core
  - 2.5Ghz or better
  - 128-512GB Memory
  - 1/10Gbps Ethernet
  - 300GB-4TB HDD (RAID)
  - Redundancy at HW & app level
- **Web Node**
  - Single Socket/2-4 cores
  - 1.0-2.0Ghz
  - 8-16GB Memory
  - 1Gbps Ethernet
  - 20-100GB HDD
  - Redundancy at app level
- **Content Node**
  - Single Socket/2-4 Core
  - 2.0-3.7 Ghz
  - 16-32GB Memory
  - 1/10Gbps Ethernet
  - 50-200GB HDD
  - Redundancy at app level



- **App Node**
  - Single or Dual Socket/4-18 Core
  - 2.0-2.5Ghz
  - 16-128GB Memory
  - 1Gbps Ethernet
  - 50-100GB HDD
  - Redundancy handled at app level
- **Db Node**
  - Single or Dual Socket/4-24 Core
  - 2.0-3.0Ghz
  - 32-256GB Memory
  - 1Gbps Ethernet
  - 100-250GB HDD
  - Redundancy handled at app level

Storage Infrastructure Requirements

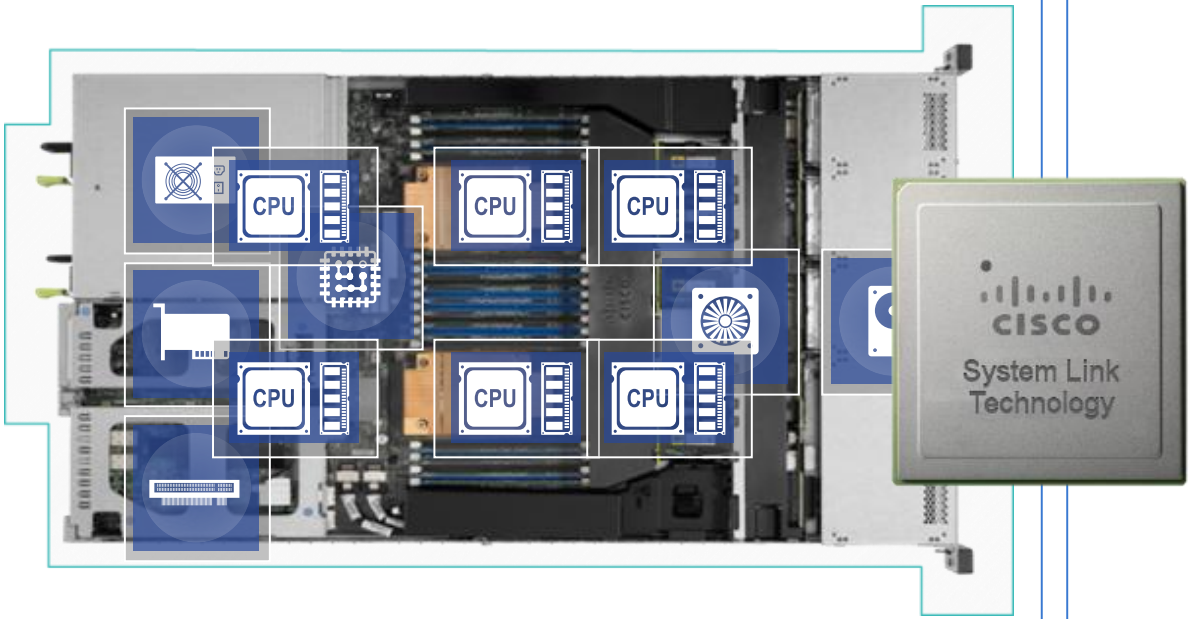
- **Object Store**
  - 1-500TB Storage
  - SSD Options
  - JBOD/RAID capabilities
  - 1-40Gbps Network BW
  - FC/FCoE initiator capabilities
  - Dual Socket/24-48 Cores
  - 2.0-2.5Ghz
  - Redundancy at HW level
- **Application Data**
  - High Performance I/O – Application Acceleration
  - Data Optimization
  - Various Workloads
  - High Availability
  - Scalability
  - FC or iSCSI connectivity
- **Analytics Data**
  - Typically a combination of HDFS, Analytics SW, and Database SW running on various rack servers.
  - See Big Data Reference architectures for more information.



Cisco System Link Technology

Extending the UCS Fabric inside the server

Compute	Shared Infrastructure



M-Series Design Considerations

- UCS M-Series was designed to complement the compute infrastructure requirements in the data centre
- The goal of the M-Series is to offer smaller compute nodes to meet the needs of scale out applications, while taking advantage of the management infrastructure and converged innovation of UCS
- By disaggregating the server components, UCS M-Series helps provide a component life cycle management strategy as opposed to a server-by-server strategy
- UCS M-Series provides a platform that will provide flexible and rapid deployment of compute, storage, and networking resources

UCS M-Series Modular Servers

Compact Chassis

Lightweight
Compute Cartridge

Shared Local Resources

UCS M-Series

True Server Disaggregation

Based on Cisco System Link Technology

3rd Gen VIC extends UCS fabric to within the server

Shared Local Resources

Four shared SSDs in the chassis

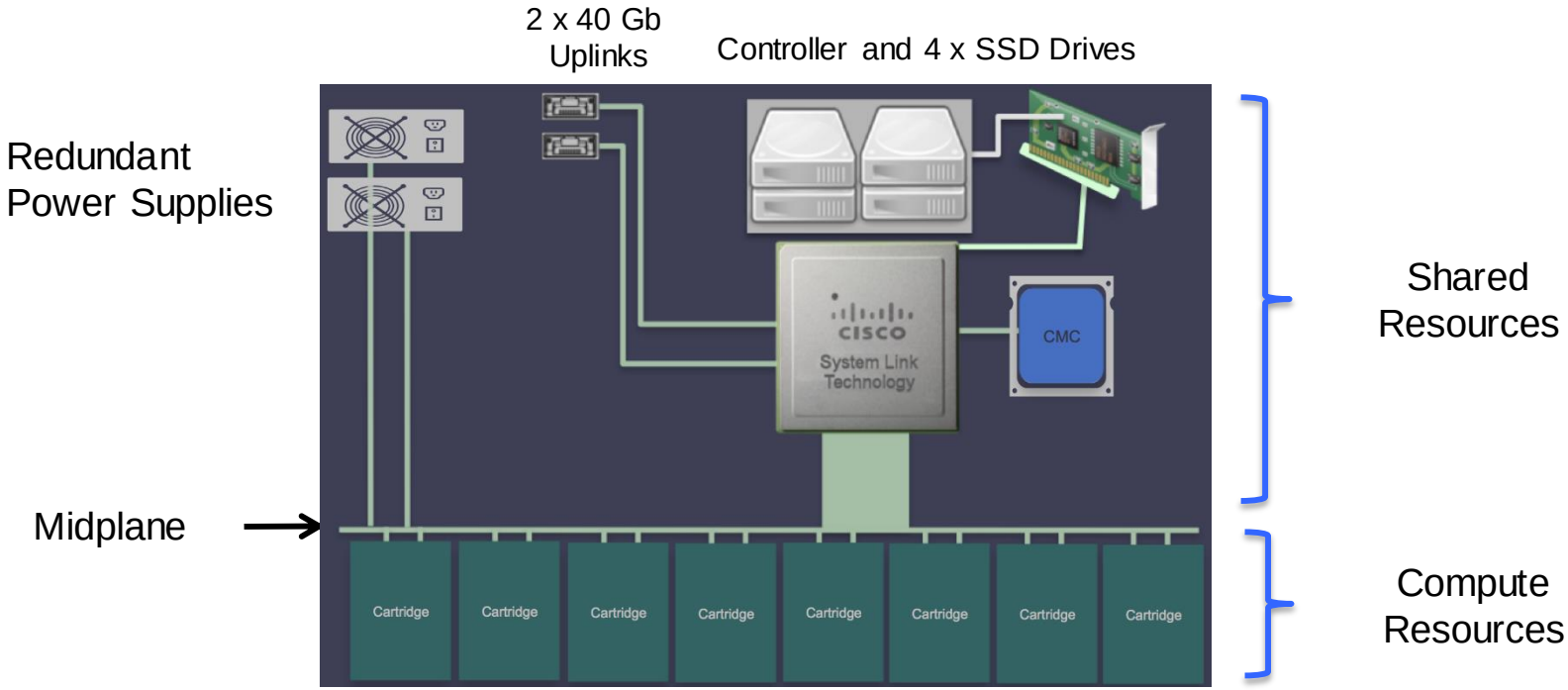
Shared dual 40Gb connectivity

Compute Density

Initial offering of16 Intel Xeon E3 Compute nodes in 2RU chassis

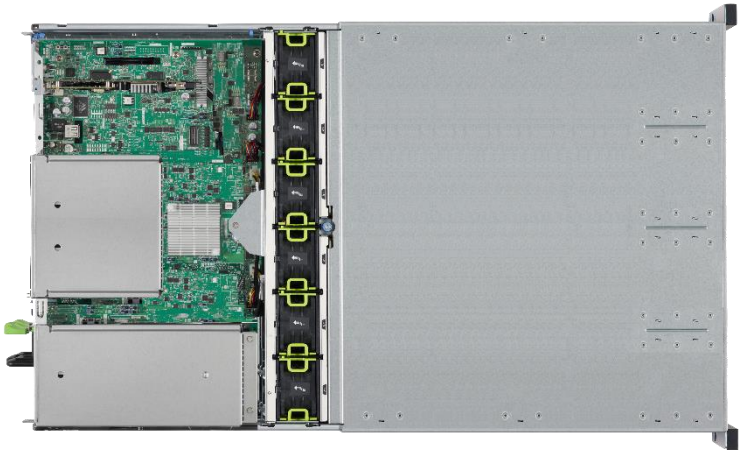
Each cartridge holds two independent compute nodes

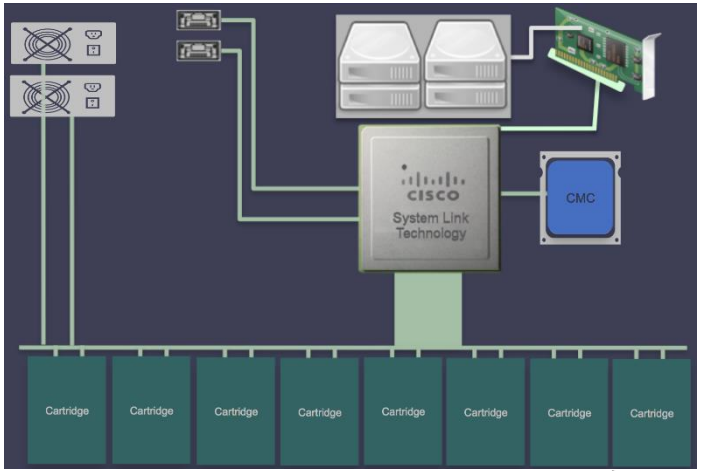
System Architecture Diagram



UCS M4308 Chassis

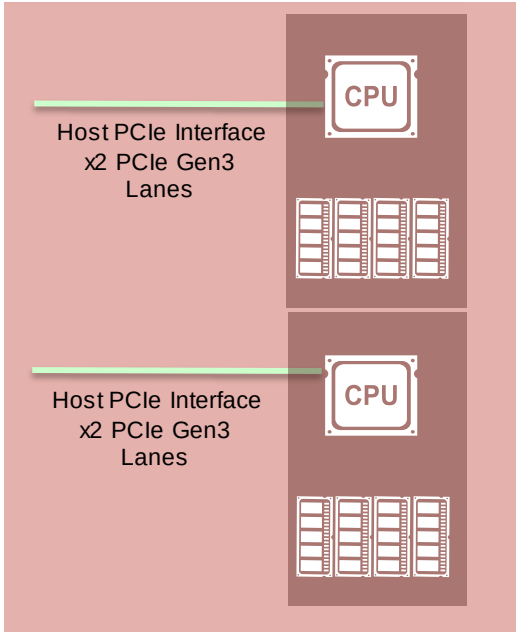
- 2U Rack-Mount Chassis (3.5” H x 30.5” L x 17.5” W)
- 3rd Generation UCS VIC ASIC (System Link Technology)
- 1 Chassis Management Controller - CMC (Manages Chassis resources)
- 8 cartridge slots (x4 PCIe Gen3 lanes per slot)
  Slots and ASIC adaptable for future use
- Four 2.5” SFF Drive Bays (SSD Only)
- 1 Internal x 8 PCIe Gen3 connection to Cisco 12 SAS RAID card
- 1 Internal x8 PCIe Gen3 slot, ½ height, ½ Width (future use)
- 6 hot swappable fans, accessed by top rear cover removal
- 2 external Ethernet management ports, 1 external serial console port (connected to CMC for out of band troubleshooting)
- 2 AC Power Module bays - (1+1 redundant, 220V only)
- 2 QSFP 40Gbps ports (data and management - server and chassis)





UCS M142 Cartridge

- UCS M142 Cartridge contains 2 distinct E3 servers
- Each Server independently manageable and has it's own memory, CPU, management controller (CIMC)
- Cartridge connects to Mid-plane for access to power, network, storage, management.
- x2 PCIe Gen3 Lanes Connect each server to System Link Technology for access to storage & network
- ~15.76 Gbps I/O Bandwidth per server

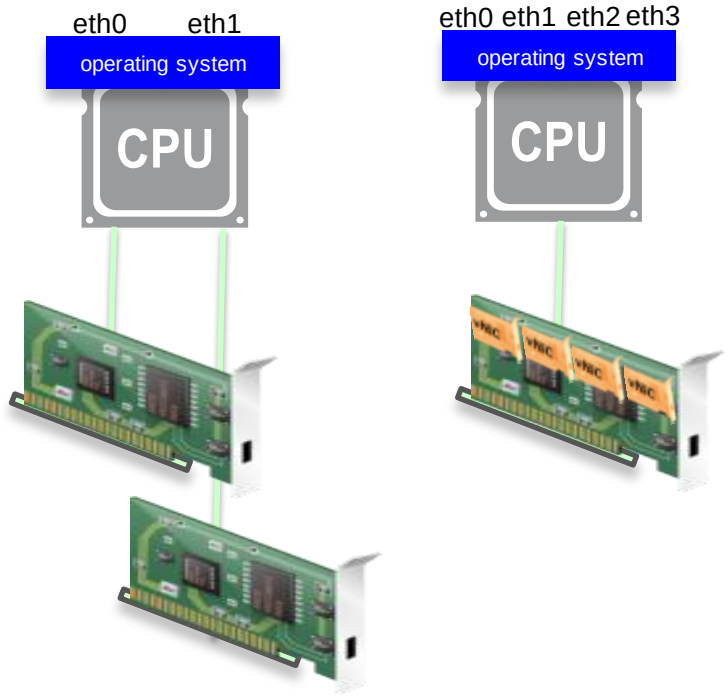


System Link Technology

Cisco live!

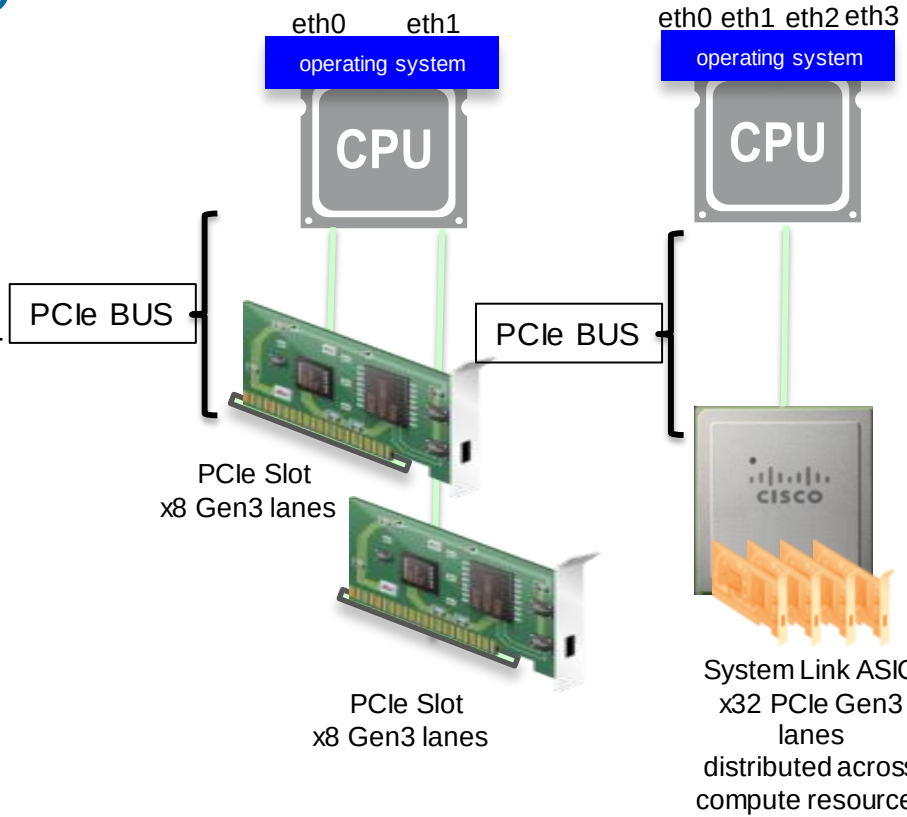
System Link Technology Overview

- System Link Technology is built on proven Cisco Virtual Interface Card (VIC) technologies
- VIC technologies use standard PCIe architecture to present an endpoint device to the compute resources
- VIC technology is a key component to the UCS converged infrastructure
- In the M-Series platform this technology has been extended to provide access to PCIe resources local to the chassis like storage



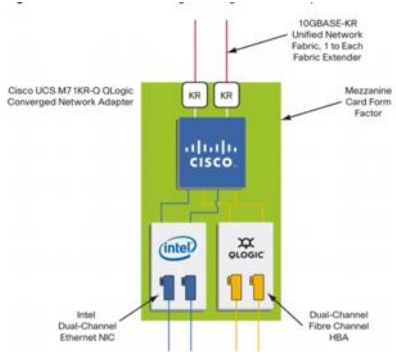
System Link Technology Overview

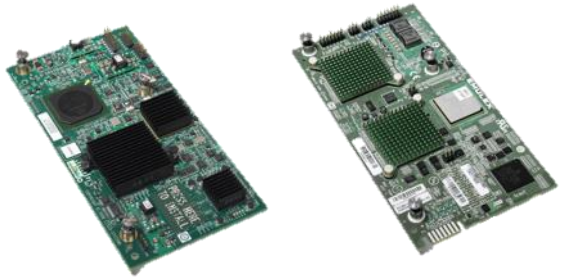
- Traditional PCIe Cards require additional PCIe slots for additional resources
- System Link Technology provides a mechanism to connect to the PCIe architecture of the host
- This presents unique PCIe devices to each server
- Each device created on the ASIC is a unique PCIe physical function



Cisco Innovation -Converged Network Adapter (M71KR-Q/E)

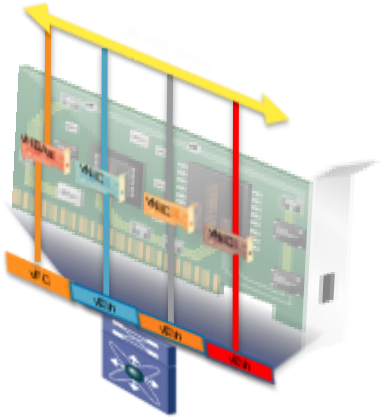
- The Initial Converged Network adapter from Cisco combined two physical PCIe devices on one card.
  - Intel/Broadcom Ethernet Adapter
  - Qlogic / Emulex HBA
- The ASIC provided a **simulated** PCIe switch for each device to be connected to the OS so they used native drivers and PCIe communications.
- The ASIC also provided the mechanism to transport the FC traffic in an FCoE frame to the Fabric Interconnect.
- The number of PCIe devices is limited to that of the hardware installed behind the ASIC





Cisco Innovation – The Cisco VIC (M81KR, 1280, & 1380)

- The Cisco VIC was an extension of the first Converged networking adapters but created 2 new PCIe devices.
  - vNIC
  - vHBA
- Each device is a PCIe physical function and is presented to the operating system as a unique PCIe endpoint.
- Drivers were created for each device.
  - eNIC driver for vNIC
  - fNIC driver for vHBA
- The ASIC allows for the creation of multiple devices (e.g. 128, 256, 1024), limited only by the ASIC capabilities

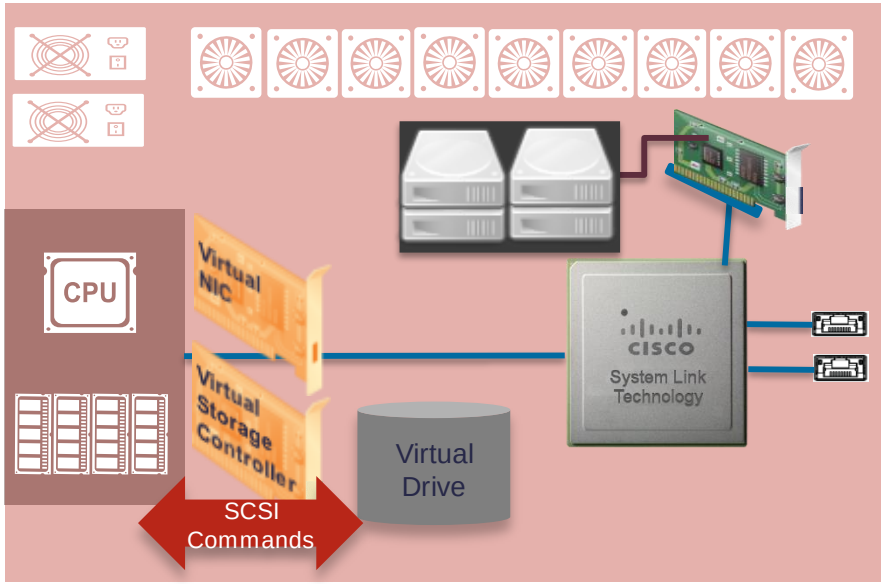


Cisco VIC Technology vs. SR-IOV

- A point that is often confused is that of SR-IOV and the operation of the VIC.
- SR-IOV allows for the creation of virtual functions on a physical PCIe card.
- The main difference is that a virtual function does not allow for direct configuration and must use the configuration of the primary physical card they are created on.
- In addition SR-IOV devices require that the operating system be SR-IOV aware to communicate with the virtual endpoints.
- VIC technology differs because each vNIC or vHBA is a PCIe physical function with full independent configuration options for each device, and requiring no OS dependencies.
- It is important to note that VIC technology is SR-IOV capable. For operating systems that can use and/or require SR-IOV support the capability does exist and is supported on the card.

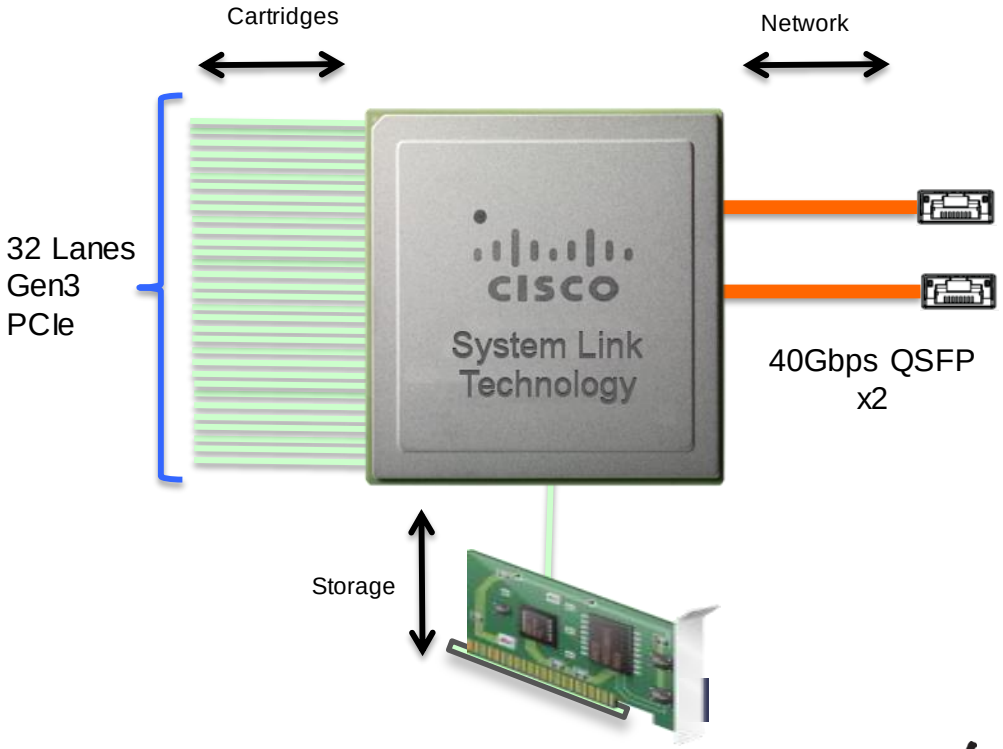
System Link Technology – 3rd Generation of Innovation

- System Link technology provides the same capabilities as a VIC to configure PCIe devices for use by the server
- The difference with System Link is that it is an ASIC within the chassis and not a PCIe card
- The ASIC is core to the M-Series platform and provides access to I/O resources
- The ASIC connects devices to the compute resource through the system mid plane
- System Link provides the ability to access network and storage shared resources



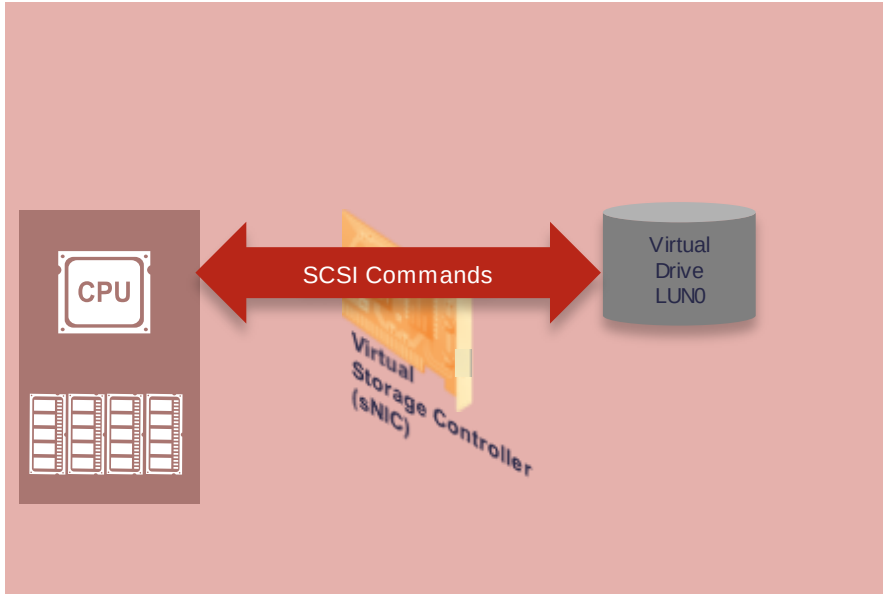
System Link Technology – 3rd Generation of Innovation

- Same ASIC used in the 3rd Generation VIC
- M-Series takes advantage of additional features which include:
  - Gen3 PCIe root complex for connectivity to Chassis PCIe cards (e.g Storage)
  - 32 Gen3 PCIe lanes connected to cartridges CPUs
  - 2 x 40Gbps uplinks
  - Scale to 1024 PCIe devices created on ASIC (e.g. vNIC)



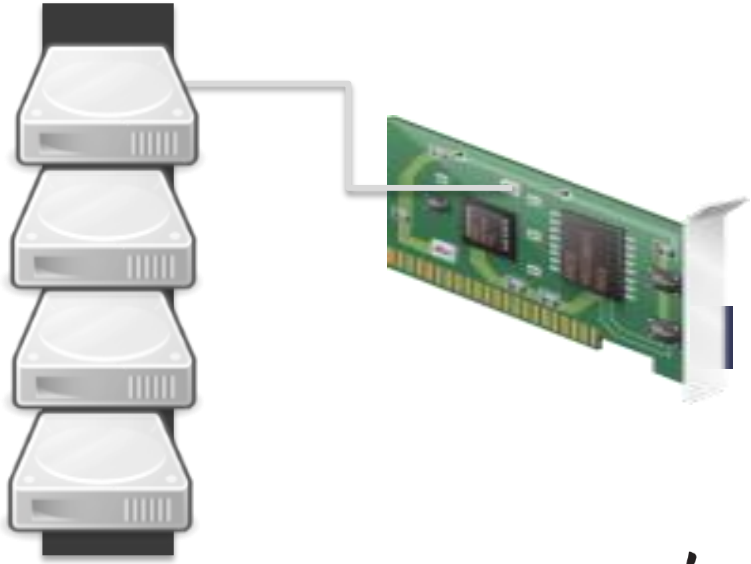
Introduction of the sNIC

- The SCSI NIC (sNIC) is the PCIe device that provides access to the storage components of the UCS M-Series Chassis
- The sNIC presents to the operating system as a PCIe connected local storage controller
- The communication between the operating system to the drive is via standard SCSI commands



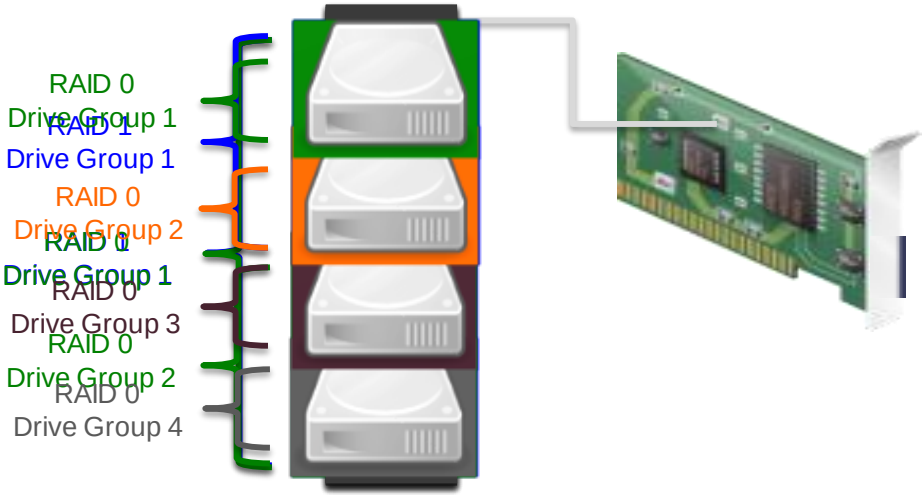
Chassis Storage Components

- Chassis Storage Consist of:
  - Cisco Modular 12Gb SAS RAID controller with 2GB Flash
  - Drive Mid-Plane
  - SSD Drives
- Controller supports RAID 0,1,5,6,10,50, & 60
- Support for 6Gb and 12Gb SAS or SATA Drives
- No support for Spinning Media SSD only (power, heat, performance)
- All drives are hot Swappable
- RAID rebuild functionality



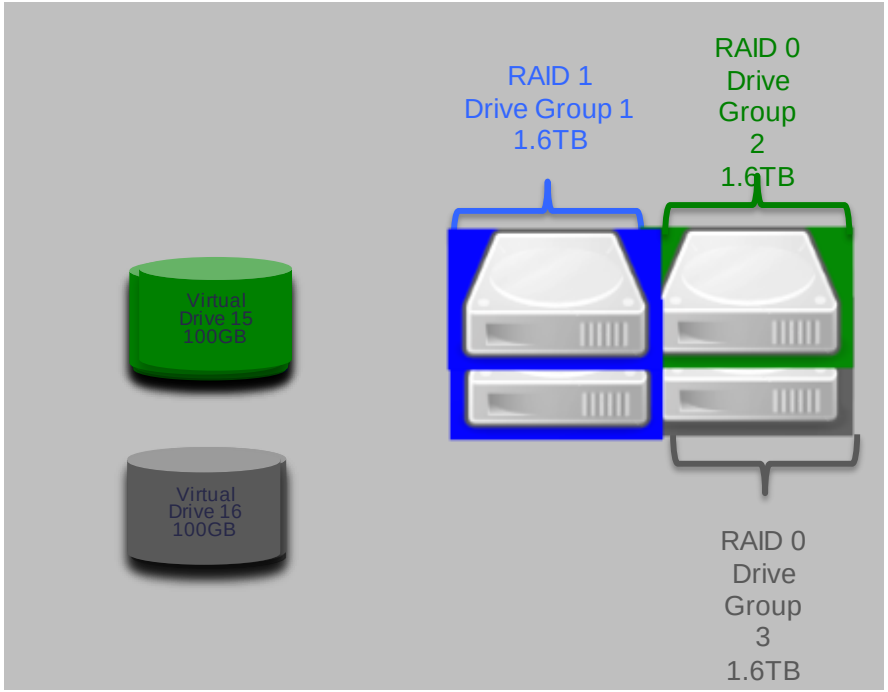
Storage Controller - Drive Groups

- RAID configuration groups drives together to form a RAID volume
- Drive Groups define the operation of the physical derives (RAID level, write back, stripe size)
- Drive Groups can bas as small as 1 1drive (R0) or 4 drives (R0, R1, R5, R10)
- M-Series chassis supports multiple drive groups
- Drive groups are configured through new policy setting in UCS Manager



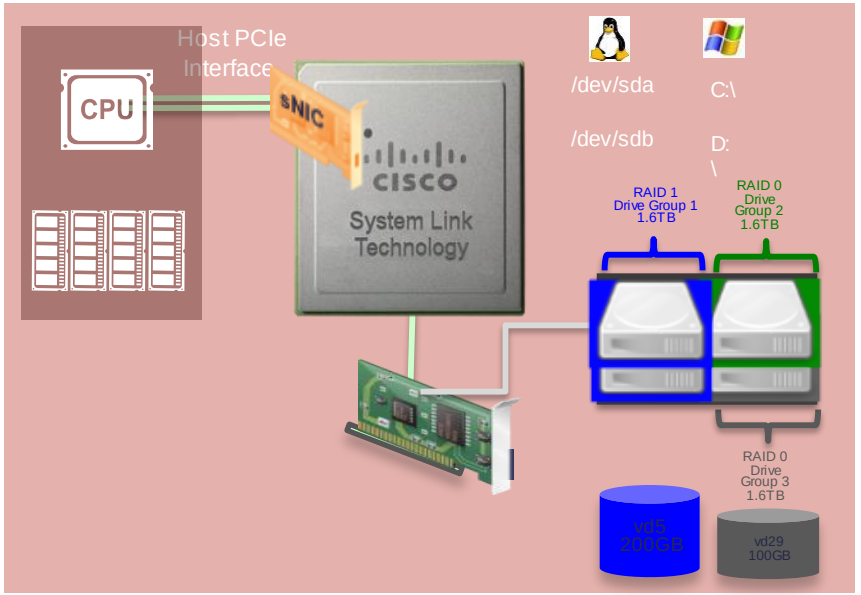
Storage Controller - Virtual Drives

- After creating a drive group on a controller a virtual drive is created to be presented as a LUN (drive) to the operating system
- It is common to use the entire amount of space available on the drive group
- If you do not use all the available space on the drive group it becomes possible to create more virtual drives on that drive group to be presented as LUNs to the device
- For UCS M-Series the virtual drives are created on a drive group to be assigned to a specific server within the chassis for local storage

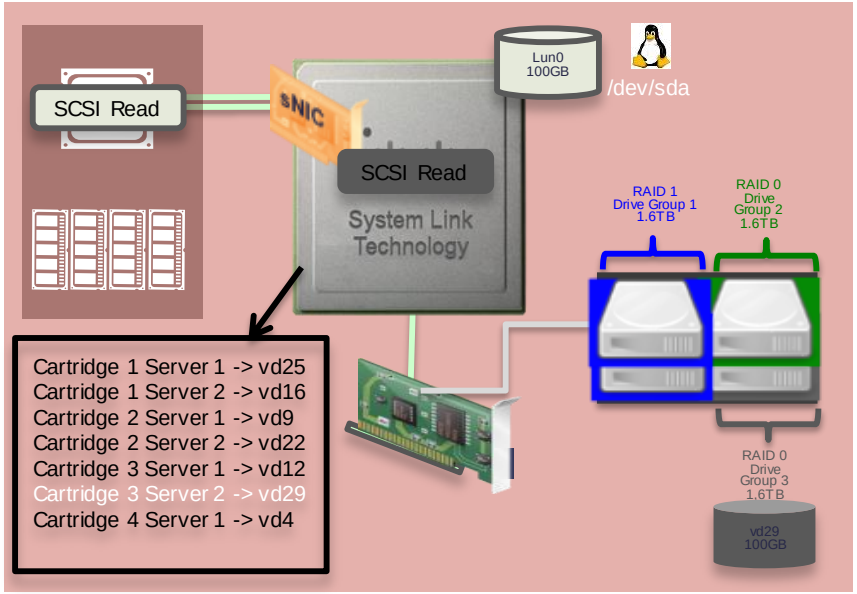


Mapping Disk Resources to M-Series Servers

- The server sees the storage controller and the virtual drive(s) as local resources.
- Through the use of policies and service profiles, UCS Manager create the sNIC in the System Link Technology as a PCIe endpoint for the server
- Through the use of policies and service profiles, UCS Manager maps the virtual drive from the RAID disk group to the server through the configured sNIC
- Each virtual drive configured in the system is applied to only 1 server and appears as a local drive to the server resource
- The operating system and sNIC send SCSI commands directly to the virtual drive through the PCIe architecture in the System Link Technology
- The end result are SCSI drives attached to the servers as LOCAL storage

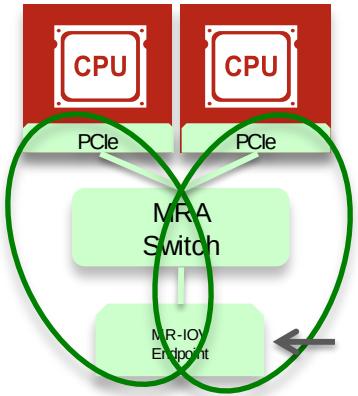


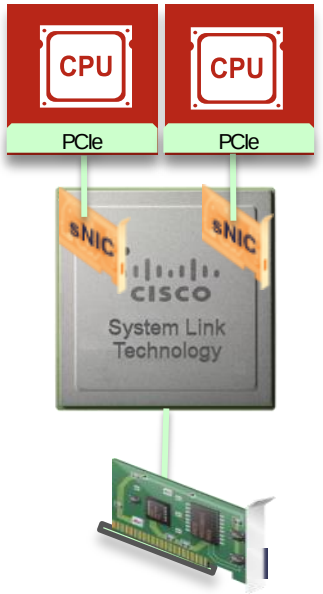
SCSI Packet Flow



System Link vs. MR-IOV

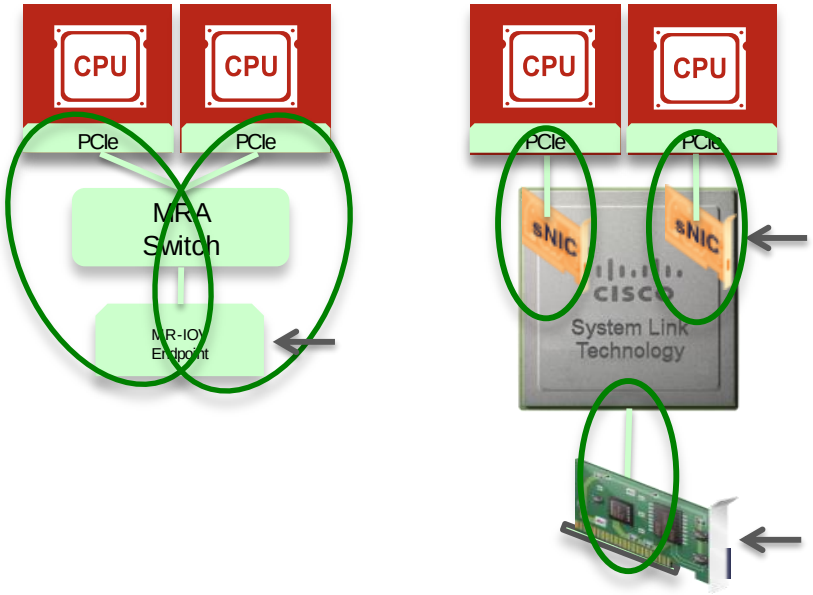
- MR-IOV is a PCIe specification that allows Multiple end host CPU to access a PCIe endpoint connected through a multi-root aware (MRA) PCIe Switch
- The same endpoint is shared between the hosts and must be able to identify and communicate with each host directly.
- MR-IOV Protocol changes are introduced to PCIe to support MR-IOV support
- Operating systems must understand and support these protocol change to be MR-IOV aware





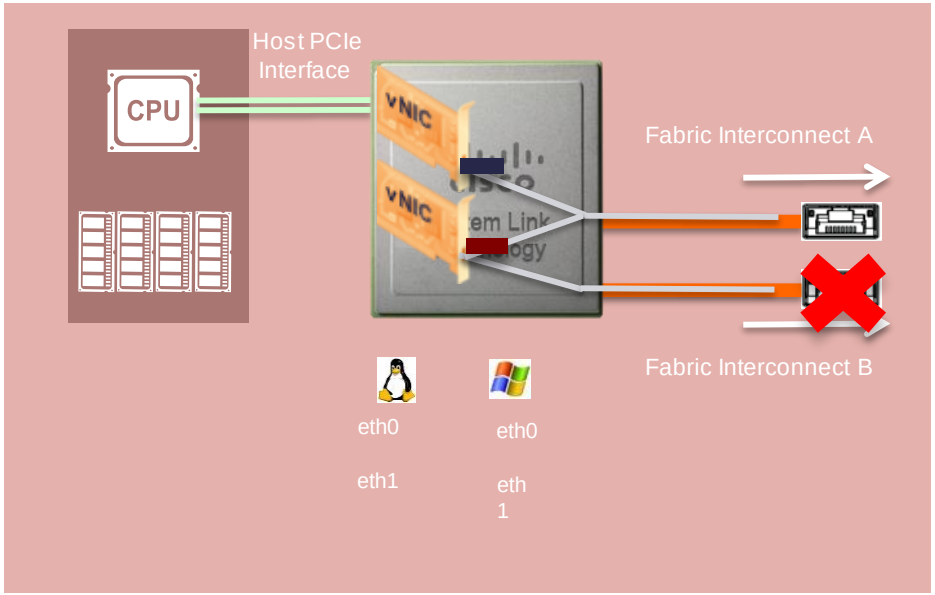
System Link vs. MR-IOV

- System Link DOES NOT require MR-IOV to support multi host access to storage devices
- System link technology provides a single root port that connects the ASIC to the storage subsystem
- The sNIC and SCSI commands from the host are translated by System Link directly to the controller so that from the controller perspective it is only communicating with one host (the System Link Technology ASIC)
- For the M-Series chassis the PCIe endpoint is the sNIC adapter and the storage device is the virtual drive provided through the mapping in the System Link Technology ASIC



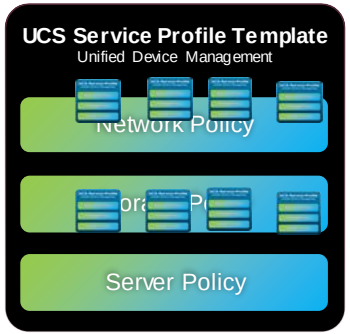
Mapping Network Resources to the M-Series Servers

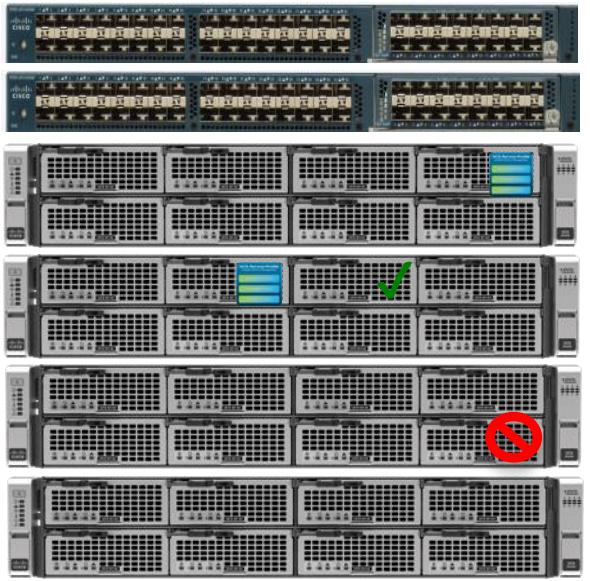
- The System Link Technology provides the network interface connectivity for all of the servers
- Virtual NICs (vNIC) are created for each server and are mapped to the appropriate fabric through the service profile on UCS Manager
- Servers can have up to 4 vNICs.
- The operating system sees each vNIC as a 10Gbps Ethernet but they can be rate limited and provide hardware QoS marking.
- Interfaces are 802.1Q capable
- Fabric Failover is supported, so in the event of a failure traffic is automatically moved to the second fabric



Service Profile Mobility

- When a virtual drive (Local LUN) is created for a service profile, it is physically located on the hard drives for the chassis where it was first applied
- Service profiles can be moved to any system in the domain, BUT the Local LUN and the data remains on the chassis where it was first associated
- Service profile mobility between chassis will NOT move the data LUN to the new chassis



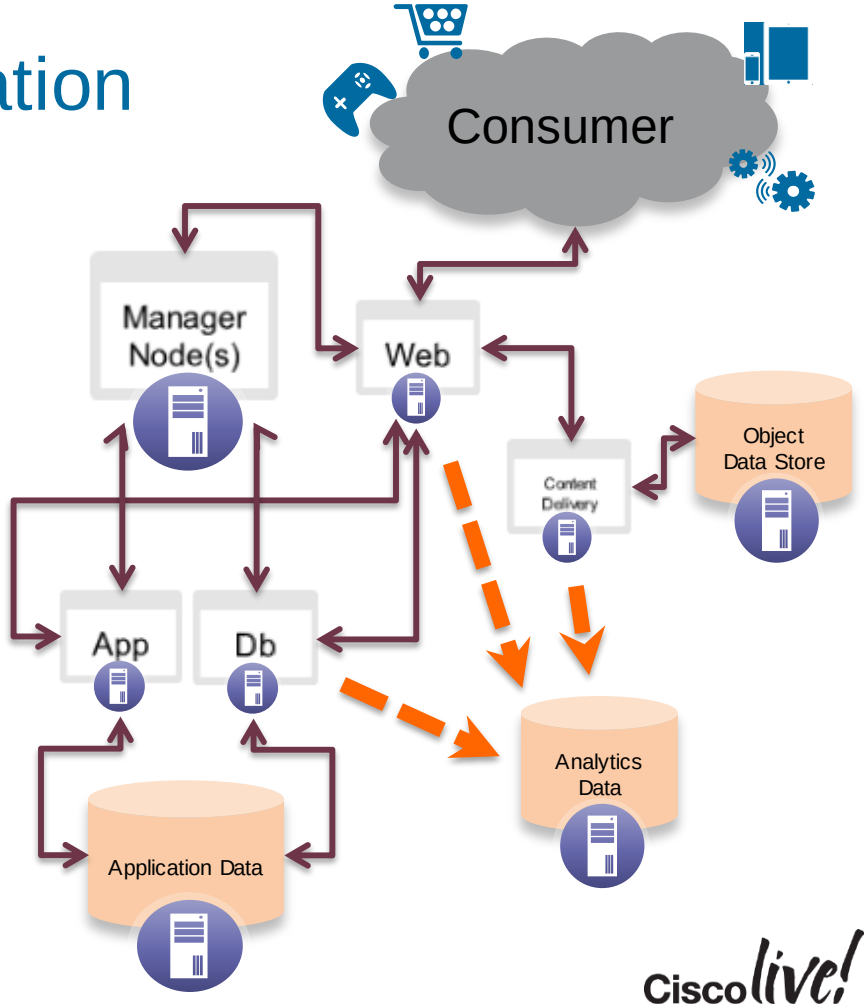


Mapping Application Architecture to Infrastructure

Cisco live!

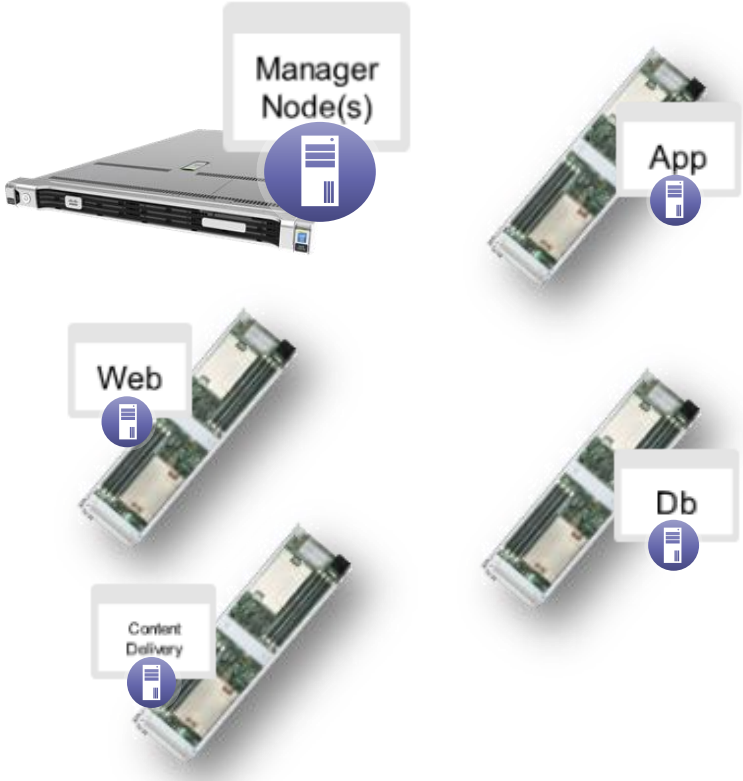
Sample Cloud Scale Application

- Cloud scale applications distribute the workload across multiple component nodes
- These nodes have various system requirements
- Distributed Components report into manager nodes
- Manager nodes note availability, farm out workloads and may receive data from worker nodes
- Worker nodes provide the bulk of cloud scale applications



Compute Infrastructure Requirements

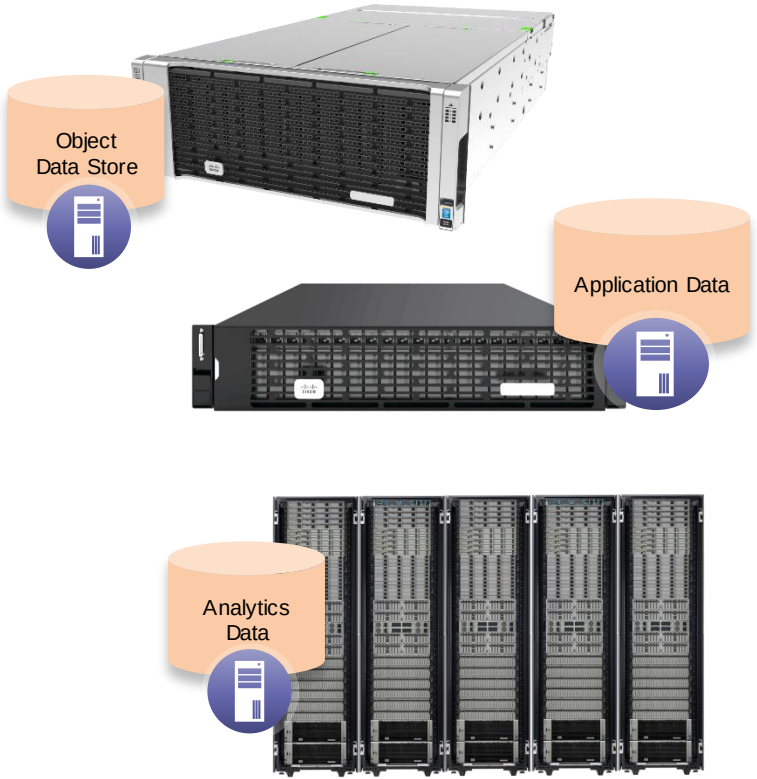
- **Manager Node**
  - Dual-Socket/8-16 core
  - 2.5Ghz or better
  - 128-512GB Memory
  - 1/10Gbps Ethernet
  - 300GB-4TB HDD (RAID)
  - Redundancy at HW & app level
- **Web Node**
  - Single Socket/2-4 cores
  - 1.0-2.0Ghz
  - 8-16GB Memory
  - 1Gbps Ethernet
  - 20-100GB HDD
  - Redundancy at app level
- **Content Node**
  - Single Socket/2-4 Core
  - 2.0-3.7 Ghz
  - 16-32GB Memory
  - 1/10Gbps Ethernet
  - 50-200GB HDD
  - Redundancy at app level



- **App Node**
  - Single or Dual Socket/4-18 Core
  - 2.0-2.5Ghz
  - 16-128GB Memory
  - 1Gbps Ethernet
  - 50-100GB HDD
  - Redundancy handled at app level
- **Db Node**
  - Single or Dual Socket/4-24 Core
  - 2.0-3.0Ghz
  - 32-256GB Memory
  - 1Gbps Ethernet
  - 100-250GB HDD
  - Redundancy handled at app level

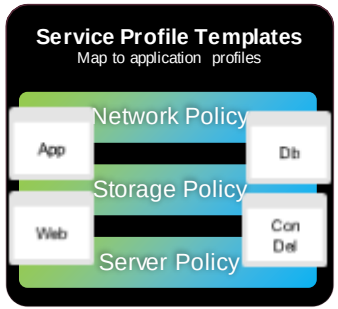
Storage Infrastructure Requirements

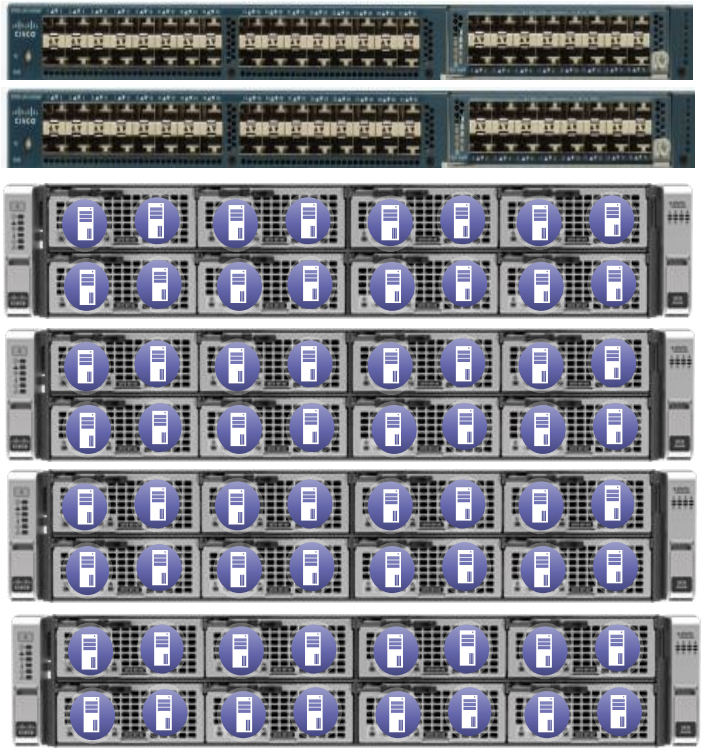
- **Object Store**
  - 1-500TB Storage
  - SSD Options
  - JBOD/RAID capabilities
  - 1-40Gbps Network BW
  - FC/FCoE initiator capabilities
  - Dual Socket/24-48 Cores
  - 2.0-2.5Ghz
  - Redundancy at HW level
- **Application Data**
  - High Performance I/O – Application Acceleration
  - Data Optimisation
  - Various Workloads
  - High Availability
  - Scalability
  - FC or iSCSI connectivity
- **Analytics Data**
  - Typically a combination of HDFS, Analytics SW, and Database SW running on various rack servers.
  - See Big Data Reference architectures for more information.



Application Profiles

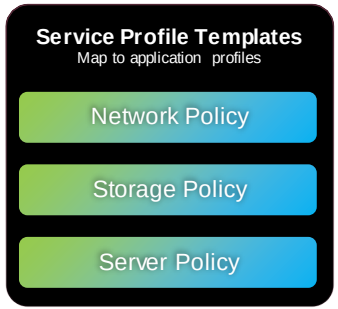
- Service Profiles allow you to build an application profile for each type of node within the application stack
- The service profile defines the node type, network connectivity, and storage configuration for each node
- These can easily be associated with compute resources across the domain

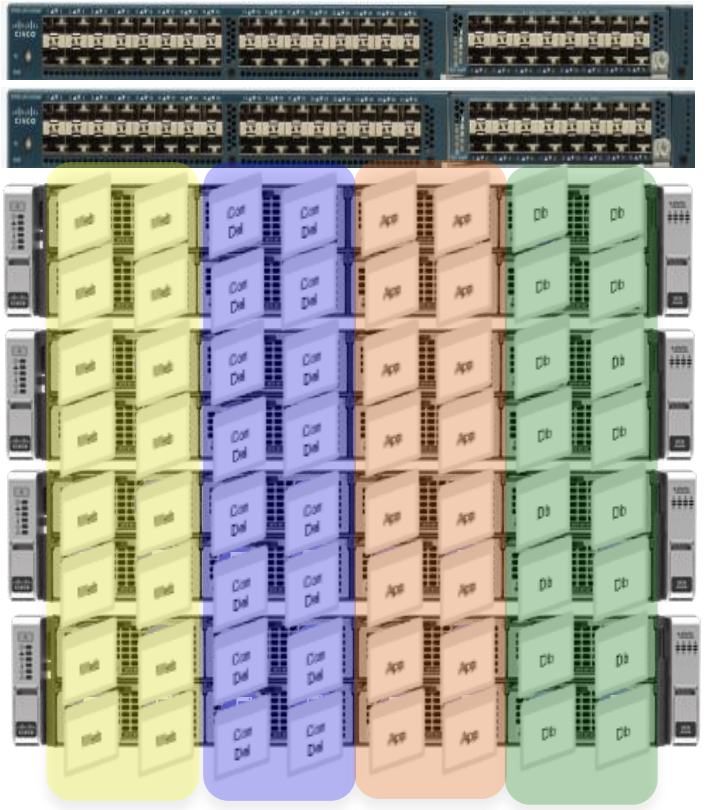




Application Profiles

- Cloud applications are built to withstand loss of multiple nodes, but the individual nodes should be striped across the chassis
- Striping the nodes also allows for better distribution of the network and storage resources





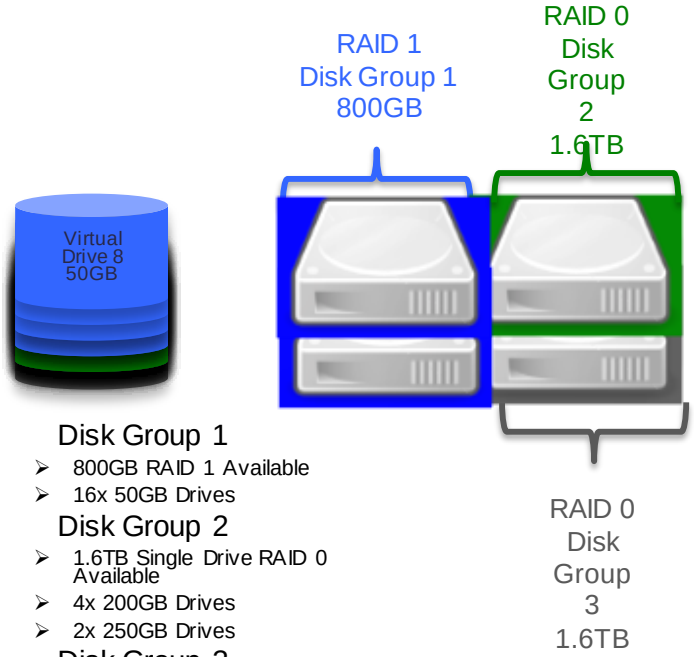
Mapping Applications to Shared Resources

- Shared Resources Example
  - 2x800GB SAS SDD
  - 2x1.6TB SAS SDD
  - 2x40Gbps Network Connections

- Web Node
  - 50GB Drive RAID1
  - 2x1Gbps Ethernet
- Content Node
  - 50GB Drive RAID 1
  - 200GB Drive
  - 1x1Gbps & 1x10Gbps Ethernet
- App Node
  - 50GB Drive RAID 1
  - 200GB Drive
  - 2x1Gbps & 1x10Gbps
- Db Node
  - 50GB Drive RAID 1
  - 250GB Drive
  - 2x1Gbps & 1x10Gbps

Storage Resource Configuration

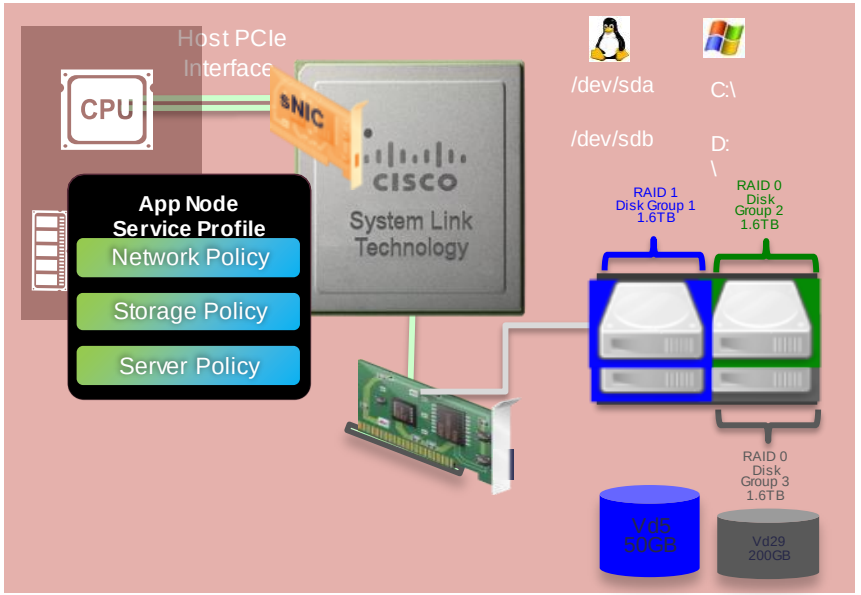
- Disk groups will be used by storage profile configurations to specify how the resources are consumed by the server nodes
- Create a RAID 1 disk group with the 2x800GB drives to host the 50GB RAID 1 required by each server node
- Create 2 separate RAID 0 disk groups to accommodate the 200GB and 250 GB drive requirements of the other server nodes
- Use specific drive numbers for the RAID 0 groups so that you can control the mapping of specific applications to specific drives



- Disk Group 1
  - 800GB RAID 1 Available
  - 16x 50GB Drives
- Disk Group 2
  - 1.6TB Single Drive RAID 0 Available
  - 4x 200GB Drives
  - 2x 250GB Drives
- Disk Group 3
  - 1.6TB Single Drive RAID 0 Available
  - 4x 200GB Drives
  - 2x 250GB Drives

Mapping Disk Resources to M-Series Servers

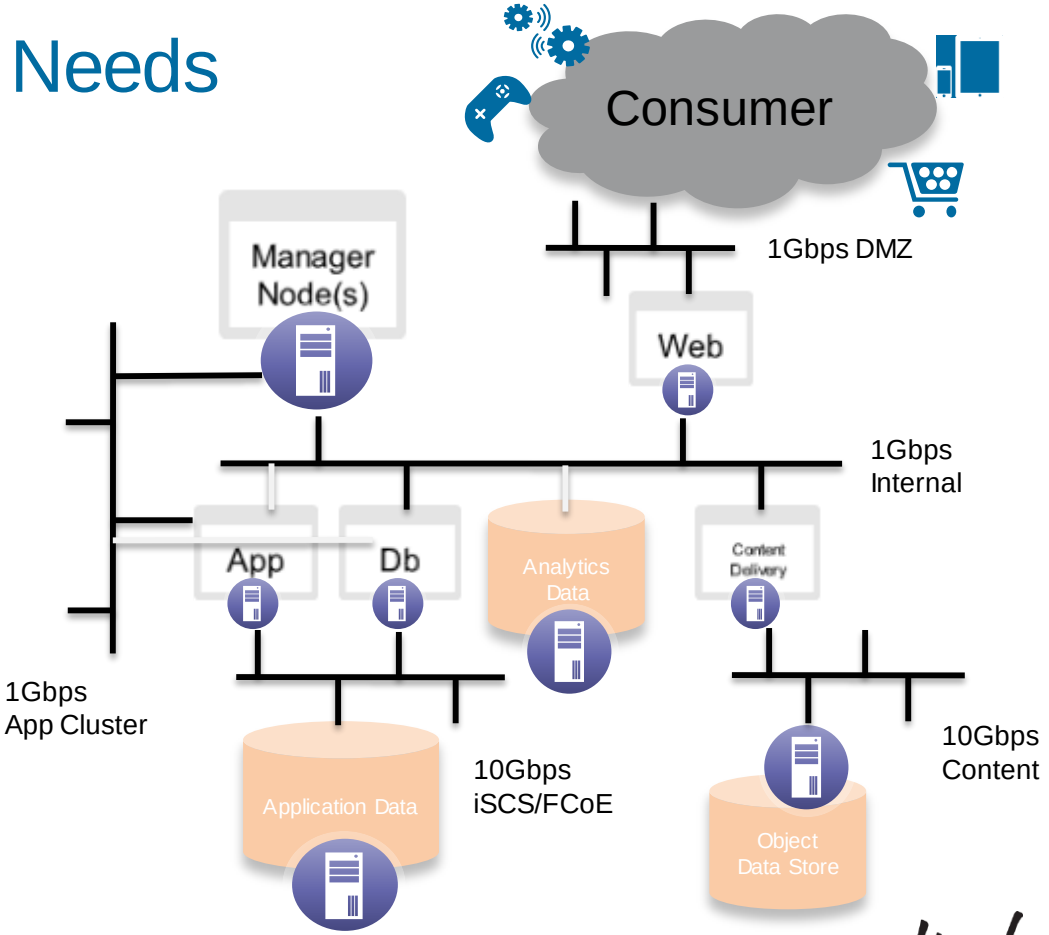
- Create a storage profile for each application type that defines the number of drives, the size of the drives and which disk group to use
- The storage profile is then consumed by the service profile to specify which resources are available for the compute node
- These resources belong to the service profile and not the physical compute node



Demo
Creating Storage Policies

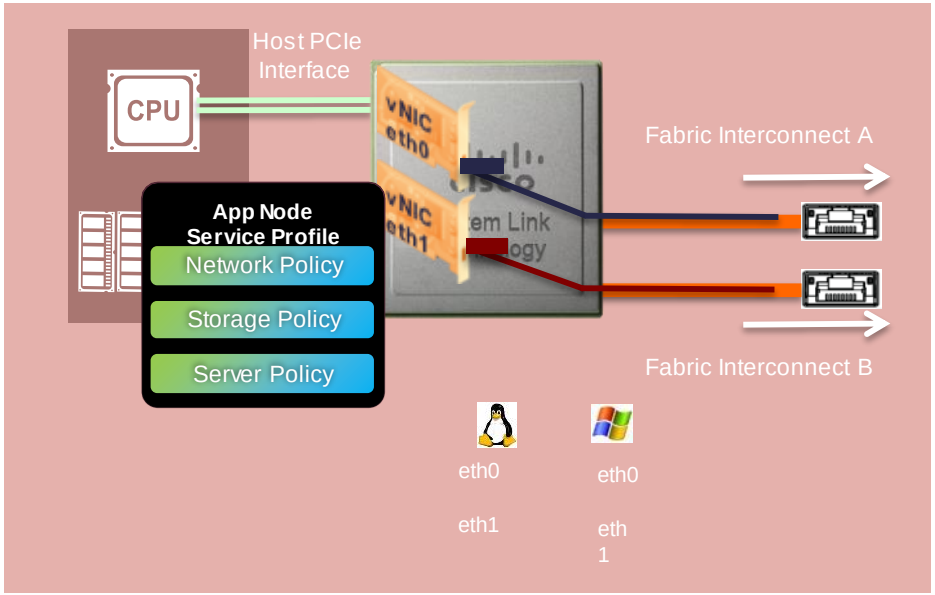
Application Networking Needs

- A cloud scale application will typically have many network segments and requirements.
- Within the construct of the service profile we can connect the server node to the required networks
- Networks may need to be added or upgraded over the life of the application.



Mapping Network Resources to the M-Series Servers

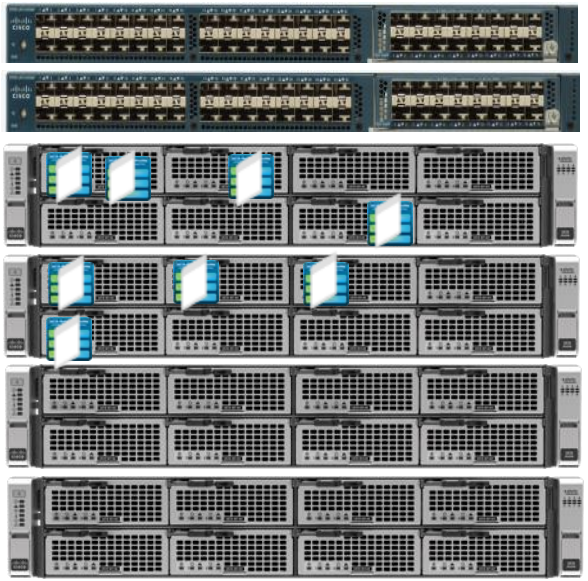
- Interfaces are created within the service profile depending on the needs of the application and mapped to a specific fabric or configured for redundancy.
- VLANs are mapped to an interface based on the network connectivity needs.
- Each interface can has full access to the 40Gbps uplink, but QoS policies and BW throttling can be configured on a per adapter basis.
- Fabric Failover provides redundancy without the need for configuration within the Operating System
- Interfaces and networks can be added without entering the data centre to change any cabling structure or add adapters.

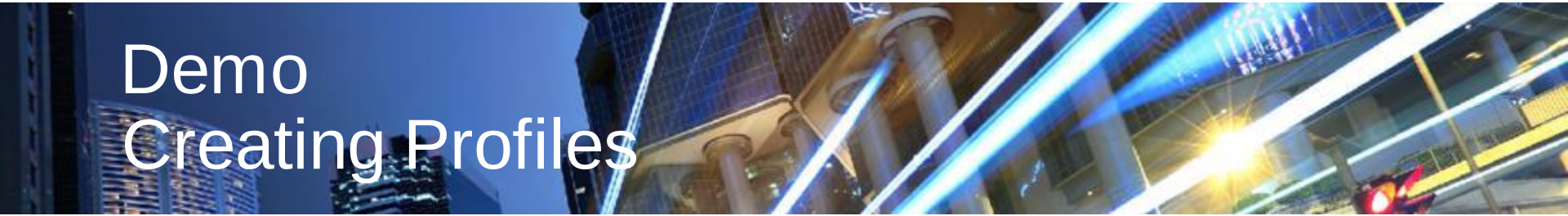


Demo
Creating Network Policies

Application Profiles

- The service profile is used to combine; Network, Storage, and Server policies to map to the needs of a specific application
- The profile defines the server and it is applied to a compute resource
- Once the service profile is defined other service profiles can easily be created through cloning or templates
- Through templates changes can be made to multiple compute nodes at one time



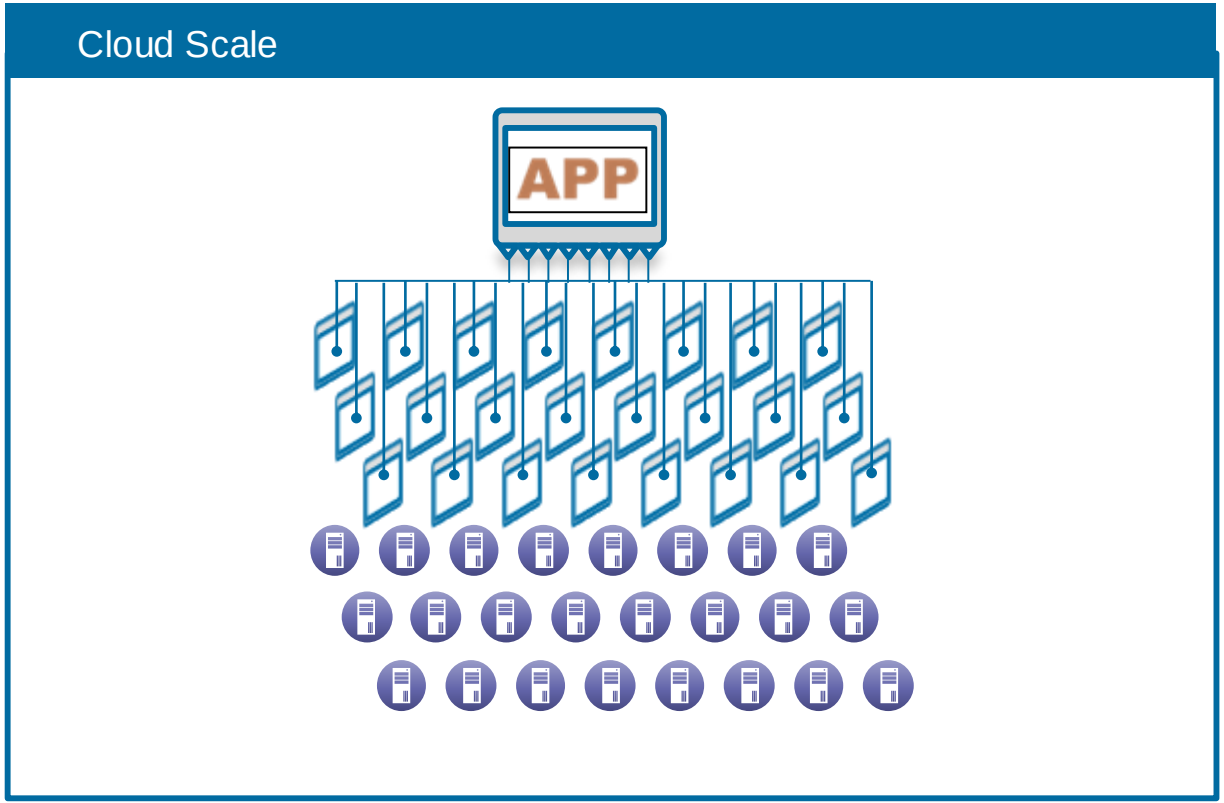


Demo
Creating Profiles

Scaling and Maintaining The Infrastructure

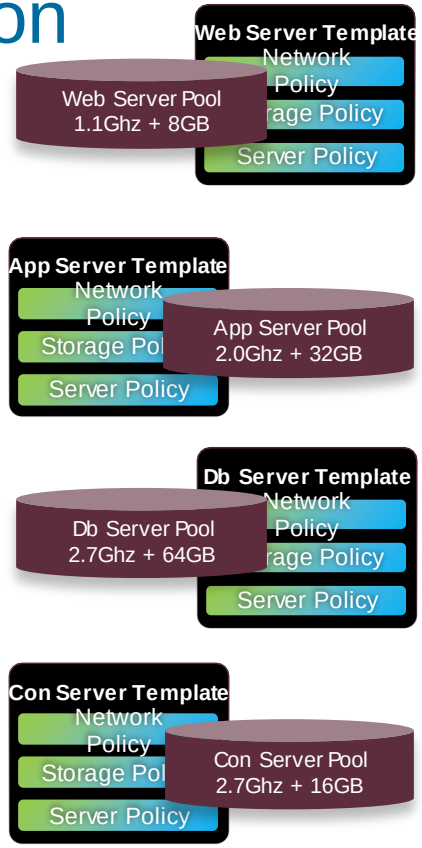
Cisco live!

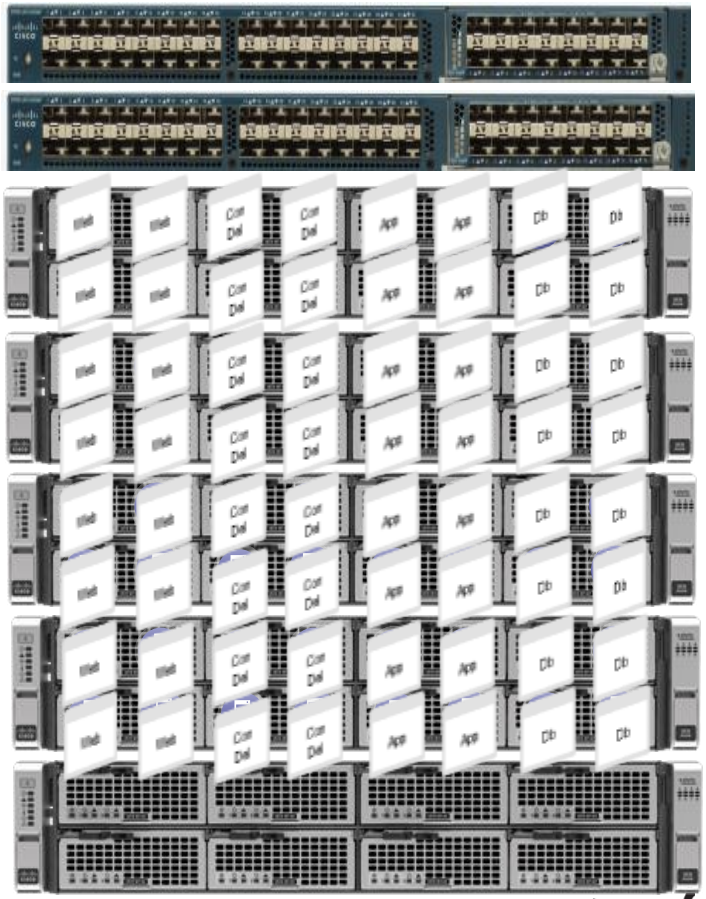
The “Scale” in Cloud-Scale



Scale Out Automation

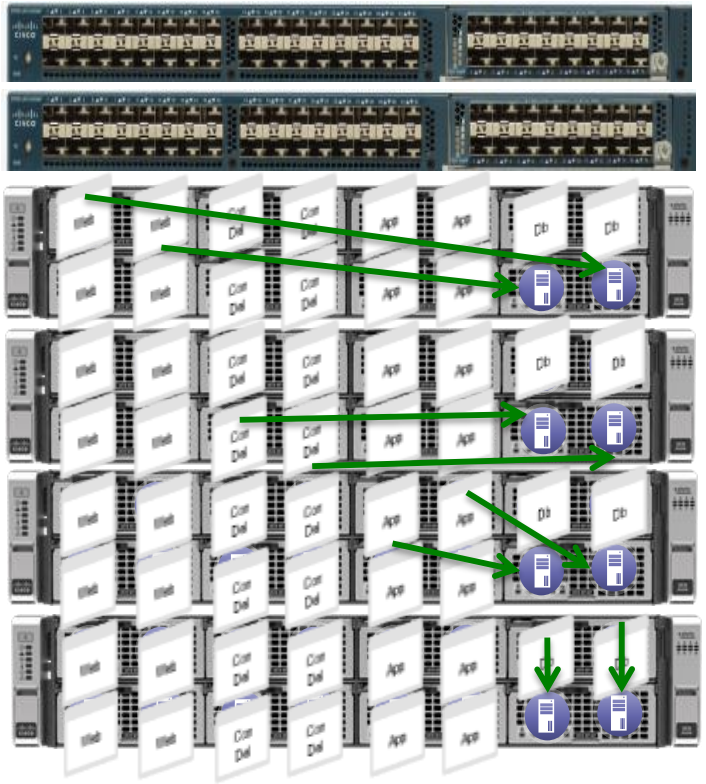
- UCS Manager Allows servers to be qualified into resource pools depending on available resources
- Service Profile templates can be assigned to consume resources in pools
- When new chassis are added to a system the resources can immediately be added to pools and have service profiles assigned
- Expansion requires no additional configuration or cabling other than installing the chassis in the rack, connecting to the Fabric, and applying power





Maintenance and Elasticity

- Storage and Data is specific to the chassis and not the server node.
- If there is a failure of a critical component you can move the application to a new server node if one is available
- It is also possible to repurpose a server node during peek times



Component Life-Cycle Management

- Server lifecycle is separate from storage, power, networking
- You can upgrade memory, processors and cartridges without the need to change network cabling or rebuild/recover drive data
- As new cartridges are released they can be installed in existing chassis providing a longer life-cycle for the platform

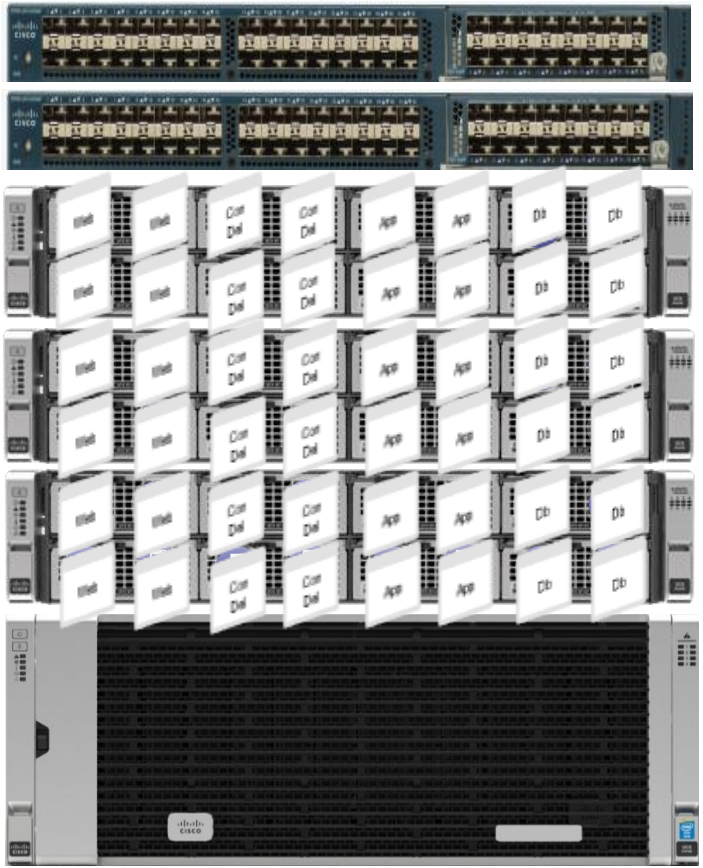
- As the upstream network is upgraded from 10Gbps to 40Gbps the chassis platform remains the same
- Drive failures for RAID protected volumes do not require downtime on the server node for repair
- Drive and controller life-cycle are separated from the individual server nodes
- If a chassis replacement were required the data from the drives would be recoverable in the replacement chassis

Incorporating Storage Into The Infrastructure

Cisco live!

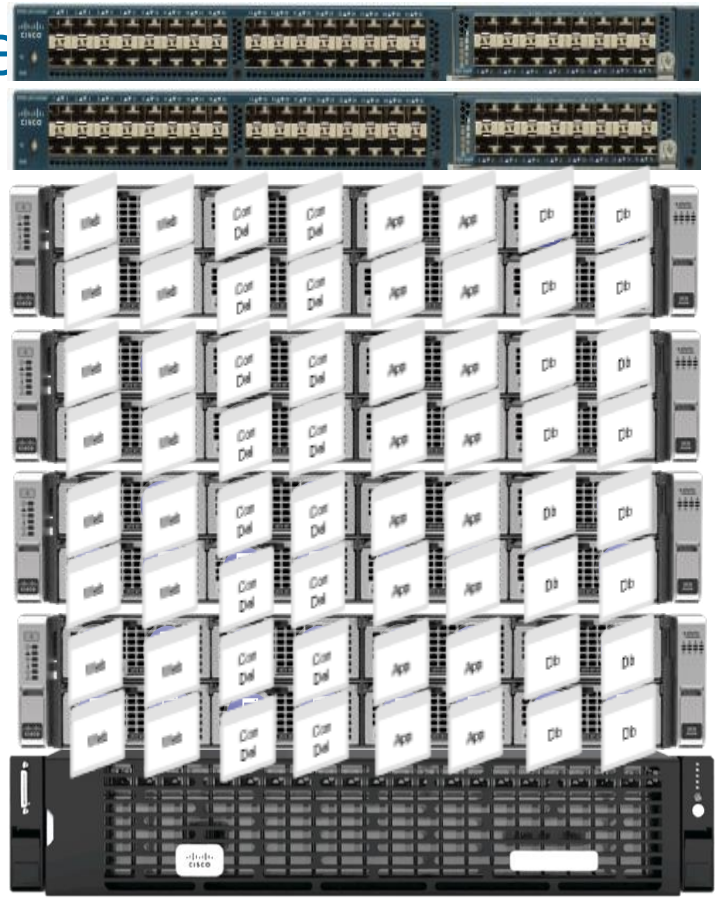
Object Store

- Within cloud scale application the emergence of object based network file systems are becoming an important part of the architecture
- The Cisco C3160 dense storage server can be loaded with a cloud scale file system (e.g. Ceph, Swift, Gluster, etc.) to provide this component of the architecture
- This can be added to a UCS infrastructure as an appliance device on the fabric interconnect
- Longer term the storage server platform will become integrated into UCS Manager



Performance Application Storage

- Many applications are starting to take advantage of high speed application storage.
- The Invicta appliance, storage blade, and scale out system provide this type of performance storage infrastructure
- Today the Invicta appliance can be added to a M-Series infrastructure as a a network appliance
- Roadmap provides the opportunities to connect to a storage blade, Appliance or scale out system in the future
- The system also provides the possibility for the storage to these systems to be included in a service profile providing more application automation and flexibility



Demo
Inserting a C3160 into the architecture

Summary

- The UCS M-Series platform provides a dense scale-out platform for cloud scale applications
- Shared resources like networking and local storage make it easy to map applications to specific resources within the infrastructure
- Service profiles provide a means to map the shared resources and abstract the application from the physical infrastructure
- The disaggregation of the server components provide an infrastructure that separates component life-cycles and maintenance and provides system elasticity
- As advanced storage capabilities becomes an important part of the UCS infrastructure, these components can be utilised within the infrastructure to build complete cloud architecture.


Q & A

Complete Your Online Session Evaluation

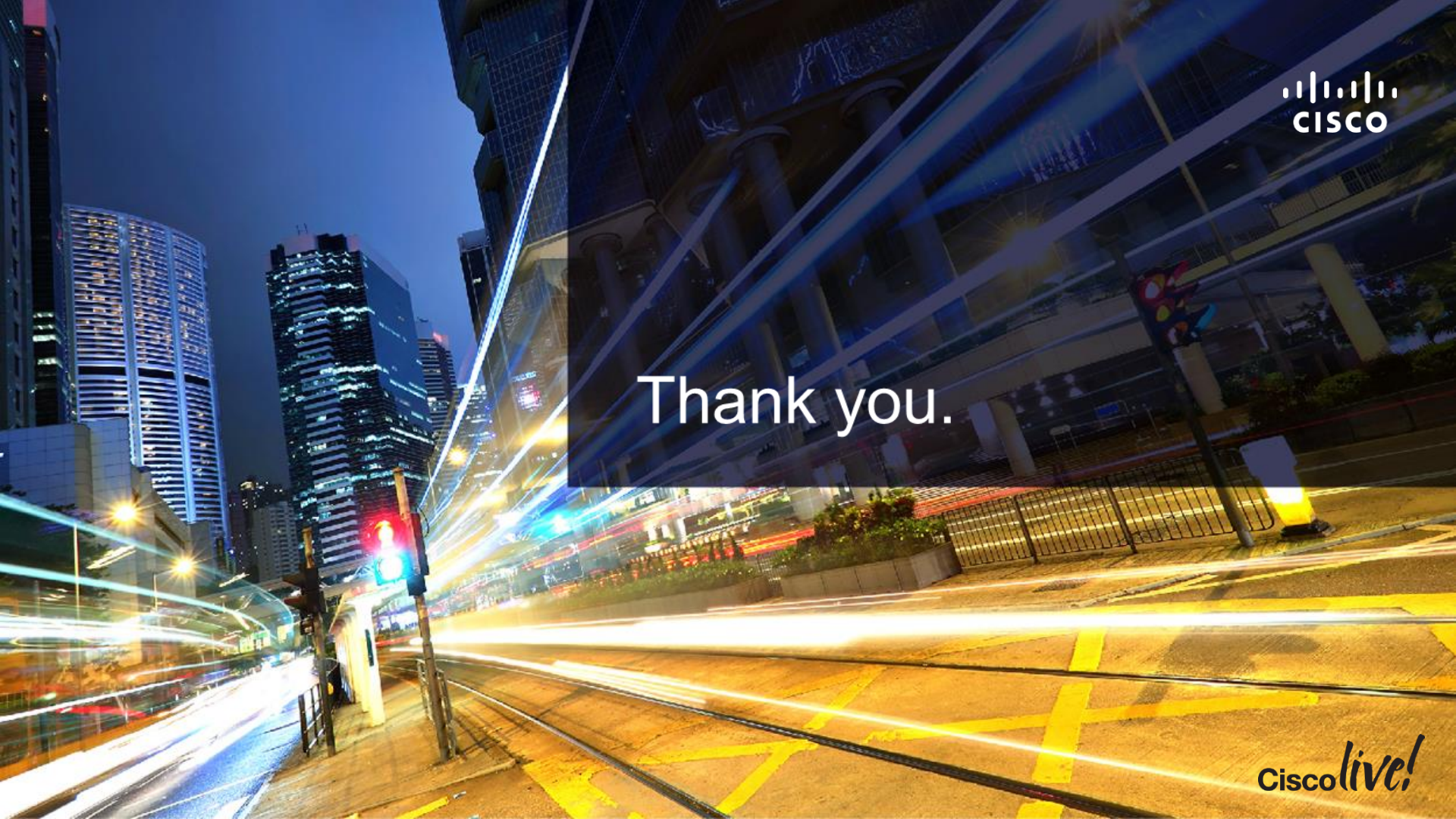
Give us your feedback and receive a Cisco Live 2015 T-Shirt!

Complete your Overall Event Survey and 5 Session Evaluations.

- Directly from your mobile device on the Cisco Live Mobile App
- By visiting the Cisco Live Mobile Site http://showcase.genie-connect.com/clmelbourne2015
- Visit any Cisco Live Internet Station located throughout the venue

T-Shirts can be collected in the World of Solutions on Friday 20 March 12:00pm - 2:00pm

Learn online with Cisco Live!
Visit us online after the conference for full access to session videos and presentations. www.CiscoLiveAPAC.com


Thank you.


CISCO