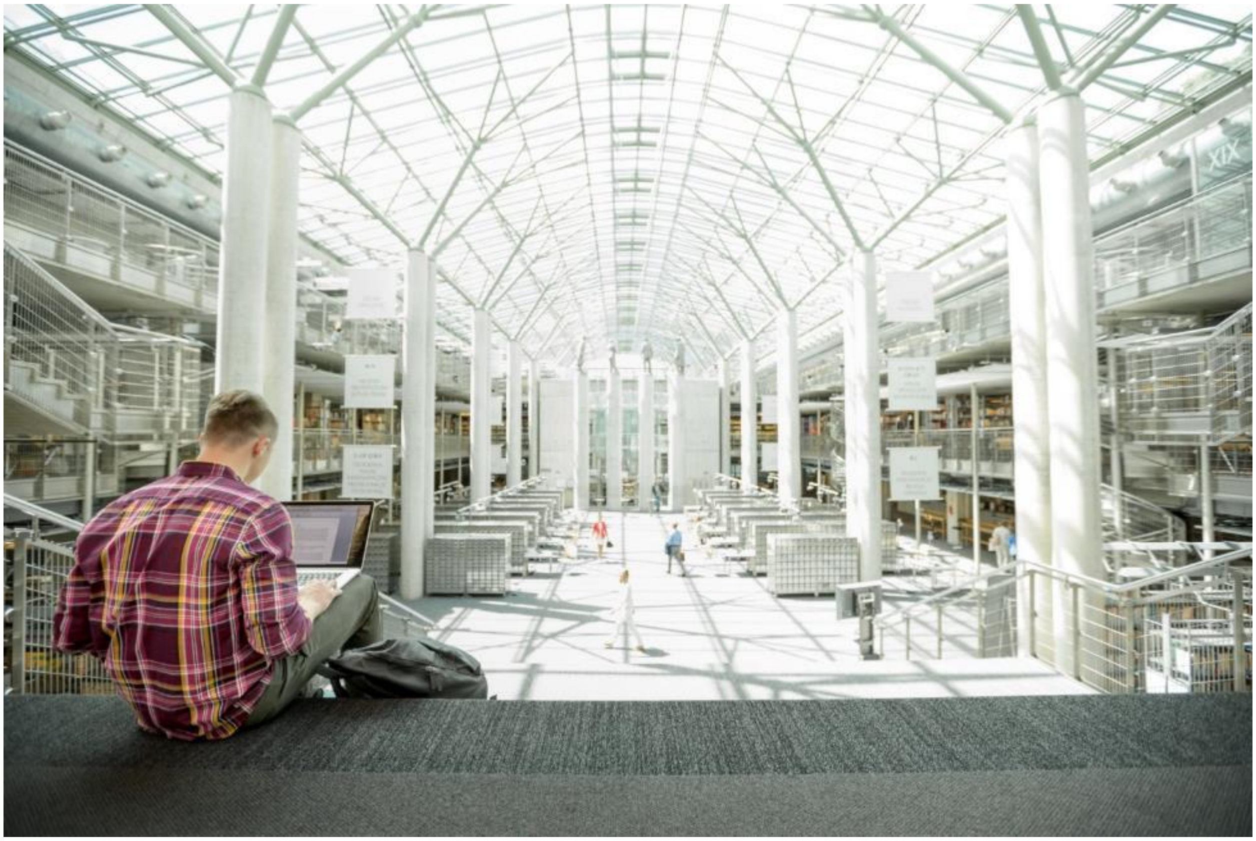


Cisco UCS C240 M5 with Scality Ring Design Guide

Last Updated: Mai 14, 2019

About the Cisco Validated Design Program

The Cisco Validated Design (CVD) program consists of systems and solutions designed, tested, and documented to facilitate faster, more reliable, and more predictable customer deployments. For more information, go to:

http://www.cisco.com/go/designzone.

ALL DESIGNS, SPECIFICATIONS, STATEMENTS, INFORMATION, AND RECOMMENDATIONS (COLLECTIVELY, "DESIGNS") IN THIS MANUAL ARE PRESENTED "AS IS," WITH ALL FAULTS. CISCO AND ITS SUPPLIERS DISCLAIM ALL WARRANTIES, INCLUDING, WITHOUT LIMITATION, THE WARRANTY OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND NONINFRINGEMENT OR ARISING FROM A COURSE OF DEALING, USAGE, OR TRADE PRACTICE. IN NO EVENT SHALL CISCO OR ITS SUPPLIERS BE LIABLE FOR ANY INDIRECT, SPECIAL, CONSEQUENTIAL, OR INCIDENTAL DAMAGES, INCLUDING, WITHOUT LIMITATION, LOST PROFITS OR LOSS OR DAMAGE TO DATA ARISING OUT OF THE USE OR INABILITY TO USE THE DESIGNS, EVEN IF CISCO OR ITS SUPPLIERS HAVE BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES.

THE DESIGNS ARE SUBJECT TO CHANGE WITHOUT NOTICE. USERS ARE SOLELY RESPONSIBLE FOR THEIR APPLICATION OF THE DESIGNS. THE DESIGNS DO NOT CONSTITUTE THE TECHNICAL OR OTHER PROFESSIONAL ADVICE OF CISCO, ITS SUPPLIERS OR PARTNERS. USERS SHOULD CONSULT THEIR OWN TECHNICAL ADVISORS BEFORE IMPLEMENTING THE DESIGNS. RESULTS MAY VARY DEPENDING ON FACTORS NOT TESTED BY CISCO.

Table of Contents

Executive Summary 5
Solution Overview 6
Introduction 6
Audience 6
Purpose of this Document 6
Solution Summary 6
Technology Overview 8
Cisco Unified Computing System 8
Cisco UCS C240 Rack Server 8
Cisco UCS Virtual Interface Card 1387 10
Cisco UCS 6300 Series Fabric Interconnect 11
Cisco Nexus 9336C-FX2 Switch 12
Cisco UCS Manager 12
Scality RING Overview 14
Scality RING Architecture 15
RING Connectors 16
Storage Nodes and IO Daemons 17
RING Systems Management 17
S3 Connector: AWS S3 Storage with Identity and Access Management (IAM) 20
Scale-Out-File-System (SOFS) 20
Intelligent Data Durability and Self-Healing 21
Replication Class of Service (COS) 22
Flexible Erasure Coding 22
Self-healing 23
Scality RING Multi-Site Deployments 23
File System (SOFS) Multi-Site Geo-Distribution 23
S3 Object Multi-Site Geo-Distribution 25
Solution Design 27
Solution Overview 27
General Hardware Requirements 27
Compute Layer Design 28
Cisco UCS Server Connectivity to Unified Fabric 28
Software Distributions and Versions 31
High Availability 31
Design Criteria 33
Requirements 34

Physical Infrastructure Considerations 34
Replication versus Erasure Coding 34
Flash Storage 34
JBOD versus RAID Disks 34
Memory Sizing 34
Network Considerations 34
Network Uplinks 34
Multi-Site Deployments 35
Connectors 35
Expansion of the Cluster 35
Summary 36
About the Authors 37
Acknowledgements 37

Executive Summary

Cisco Validated Designs consist of systems and solutions that are designed, tested, and documented to facilitate and improve customer deployments. These designs incorporate a wide range of technologies and products into a portfolio of solutions that have been developed to address the business needs of our customers.

The purpose of this document is to describe the design of Scality RING on Red Hat Enterprise Linux and on latest generation of Cisco UCS C240 Rack Servers. This validated design provides the framework of designing Scality SDS software on Cisco UCS C240 Rack Servers. The Cisco Unified Computing System provides the storage, network, and storage access components for Scality RING, deployed as a single cohesive system.

Cisco Validated design describes how the Cisco Unified Computing System can be used in conjunction with Scality RING 7.4. With the continuous evolution of Software Defined Storage (SDS), there has been increased demand to have small Scality RING solutions validated on Cisco UCS servers. The Cisco UCS C240 Rack Server, originally designed for the data center, together with Scality RING is optimized for such object storage solutions, making it an excellent fit for unstructured data workloads such as active archive, backup, and cloud data. The Cisco UCS C240 Rack Server delivers a complete infrastructure with exceptional scalability for computing and storage resources together with 40 Gigabit Ethernet networking.

Cisco and Scality are collaborating to offer customers a scalable object storage solution for unstructured data that is integrated with Scality RING. With the power of the Cisco UCS management framework, the solution is cost effective to deploy and manage and will enable the next-generation cloud deployments that drive business agility, lower operational costs and avoid vendor lock-in.

Solution Overview

Introduction

Traditional storage systems are limited in their ability to easily and cost-effectively scale to support large amounts of unstructured data. With about 80 percent of data being unstructured, new approaches using x86 servers are proving to be more cost effective, providing storage that can be expanded as easily as your data grows. Software Defined Storage is a scalable and cost-effective approach for handling large amounts of data.

But more and more there are requirements to store unstructured data even in smaller quantities as object storage. The advantage of identifying the data by metadata and not taking over management of the location is very attractive even for smaller quantities. As a result, new technologies need to be developed to provide similar levels of availability and reliability as large scale-out object storage solutions.

Scality RING is a storage platform that is ideal for holding large amounts of colder production data, such as backups and archives, and very large individual files, such as video files, image files, and genomic data and can also include support of warm or even hot data, by increasing CPU performance and/or memory capacity. Scality RING is highly reliable, durable, and resilient object storage for that is designed for scale and security.

Together with Cisco UCS, Scality RING can deliver a fully enterprise-ready solution that can manage different workloads and still remain flexible. The Cisco UCS C240 Rack Server is an excellent platform to use with object and file workloads, such as capacity-optimized and performance-optimized workloads. It is best suited for sequential access, as opposed to random, to unstructured data, and to any data size. It is designed for applications, not direct end-users.

This document describes the architecture, design procedures of Scality storage on Cisco UCS C240 M5 servers.

Audience

The audience for this document includes, but is not limited to, sales engineers, field consultants, professional services, IT managers, partner engineers, IT architects, and customers who want to take advantage of an infrastructure that is built to deliver IT efficiency and enable IT innovation. The reader of this document is expected to have the necessary training and background to install and configure Red Hat Enterprise Linux, Cisco Unified Computing System (Cisco UCS), Cisco Nexus, and Cisco UCS Manager (UCSM) as well as a high-level understanding of Scality RING Software and its components. External references are provided where applicable and it is recommended that the reader be familiar with these documents.

Readers are also expected to be familiar with the infrastructure, network and security policies of the customer installation.

Purpose of this Document

This document describes the architecture and design of a Scality RING solution on Cisco UCS. It shows the simplicity of installing and configuring the shared infrastructure platform and illustrates the need of a well-conceived network architecture for low-latency, high-bandwidth.

Solution Summary

This Cisco Validated Design (CVD) is a simple and linearly scalable architecture that provides Software Defined Storage for object and file on Scality RING 7.4 and Cisco UCS C240 rack server. This CVD describes in detail the design of Scality RING on Cisco UCS C240 rack server. The solution includes the following features:

- Infrastructure for scale-out storage

- Design of a Scality RING solution together with Cisco UCS C240 Rack Server
- Simplified infrastructure management with Cisco UCS Manager (UCSM)

The configuration uses the following architecture for the deployment:

- 3 x Cisco UCS C240 Rack Servers
- 2 x Cisco UCS 6332 Fabric Interconnect
- 1 x Cisco UCS Manager
- 2 x Cisco Nexus 9336C-FX2 Switches
- Scality RING 7.4
- Red Hat Enterprise Linux 7.6

The solution has various options to scale capacity. The tested configuration uses ARC (Advanced Resiliency Configuration) 7+5 and COS 3 replication for small objects. A base capacity summary for the tested solution is listed in Table 1 .

Table 1 Usable Capacity Options for Tested Cisco Validated Design

HDD Type	Number of Disks	Usable Capacity
4 TB 7200-rpm LFF SAS drives	36	79 TB
6 TB 7200-rpm LFF SAS drives	36	118 TB
8 TB 7200-rpm LFF SAS drives	36	158 TB
10 TB 7200-rpm LFF SAS drives*	36	197 TB
12 TB 7200-rpm LFF SAS drives	36	237 TB

* Tested configuration

Technology Overview

Cisco Unified Computing System

The Cisco Unified Computing System™ (Cisco UCS) is a state-of-the-art data center platform that unites computing, network, storage access, and virtualization into a single cohesive system.

The main components of Cisco Unified Computing System are:

- Computing - The system is based on an entirely new class of computing system that incorporates rack-mount and blade servers based on Intel® Xeon® Scalable processors. The Cisco UCS servers offer the patented Cisco Extended Memory Technology to support applications with large datasets and allow more virtual machines (VM) per server.
- Network - The system is integrated onto a low-latency, lossless, 10/25/40-Gbps unified network fabric. This network foundation consolidates LANs, SANs, and high-performance computing networks which are separate networks today. The unified fabric lowers costs by reducing the number of network adapters, switches, and cables, and by decreasing the power and cooling requirements.
- Virtualization - The system unleashes the full potential of virtualization by enhancing the scalability, performance, and operational control of virtual environments. Cisco security, policy enforcement, and diagnostic features are now extended into virtualized environments to better support changing business and IT requirements.
- Storage access - The system provides consolidated access to both SAN storage and Network Attached Storage (NAS) over the unified fabric. By unifying the storage access, the Cisco Unified Computing System can access storage over Ethernet (NFS or iSCSI), Fibre Channel, and Fibre Channel over Ethernet (FCoE). This provides customers with choice for storage access and investment protection. In addition, the server administrators can pre-assign storage-access policies for system connectivity to storage resources, simplifying storage connectivity, and management for increased productivity.

The Cisco Unified Computing System is designed to deliver:

- A reduced Total Cost of Ownership (TCO) and increased business agility.
- Increased IT staff productivity through just-in-time provisioning and mobility support.
- A cohesive, integrated system, which unifies the technology in the data center.
- Industry standards supported by a partner ecosystem of industry leaders.

Cisco UCS C240 Rack Server

The Cisco UCS C240 Rack Server is a 2-socket, 2-Rack-Unit (2RU) rack server offering industry-leading performance and expandability. It supports a wide range of storage and I/O-intensive infrastructure workloads, from big data and analytics to collaboration. Cisco UCS C-Series Rack Servers can be deployed as standalone servers or as part of a Cisco UCS managed environment to take advantage of Cisco's standards-based unified computing innovations that help reduce customers' TCO and increase their business agility.

Figure 1 Cisco UCS C240 Rack Server

In response to ever-increasing computing and data-intensive real-time workloads, the enterprise-class Cisco UCS C240 server extends the capabilities of the Cisco UCS portfolio in a 2RU form factor. It incorporates the Intel® Xeon® Scalable processors, supporting up to 20 percent more cores per socket, twice the memory capacity, and five times more Non-Volatile Memory Express (NVMe) PCI Express (PCIe) Solid-State Disks (SSDs) compared to the previous generation of servers. These improvements deliver significant performance and efficiency gains that will improve your application performance. The Cisco UCS C240 M5 delivers outstanding levels of storage expandability with exceptional performance, comprised of the following:

- Latest Intel Xeon Scalable CPUs with up to 28 cores per socket
- Up to 24 DDR4 DIMMs for improved performance
- Up to 26 hot-swappable Small-Form-Factor (SFF) 2.5-inch drives, including 2 rear hot-swappable SFF drives (up to 10 support NVMe PCIe SSDs on the NVMe-optimized chassis version), or 12 Large-Form-Factor (LFF) 3.5-inch drives plus 2 rear hot-swappable SFF drives
- Support for 12-Gbps SAS modular RAID controller in a dedicated slot, leaving the remaining PCIe Generation 3.0 slots available for other expansion cards
- Modular LAN-On-Motherboard (mLOM) slot that can be used to install a Cisco UCS Virtual Interface Card (VIC) without consuming a PCIe slot, supporting dual 10-, 25- or 40-Gbps network connectivity
- Dual embedded Intel x550 10GBASE-T LAN-On-Motherboard (LOM) ports
- Modular M.2 or Secure Digital (SD) cards that can be used for boot

The Cisco UCS C240 rack server is well suited for a wide range of enterprise workloads, including:

- Object Storage
- Big Data and analytics
- Collaboration
- Small and medium-sized business databases
- Virtualization and consolidation
- Storage servers
- High-performance appliances

Cisco UCS C240 rack servers can be deployed as standalone servers or in a Cisco UCS managed environment. When used in combination with Cisco UCS Manager, the Cisco UCS C240 brings the power and automation of unified computing to enterprise applications, including Cisco® SingleConnect technology, drastically reducing switching and cabling requirements.

Cisco UCS Manager uses service profiles, templates, and policy-based management to enable rapid deployment and help ensure deployment consistency. If also enables end-to-end server visibility, management, and control in both virtualized and bare-metal environments.

The Cisco Integrated Management Controller (IMC) delivers comprehensive out-of-band server management with support for many industry standards, including:

- Redfish Version 1.01 (v1.01)
- Intelligent Platform Management Interface (IPMI) v2.0

- Simple Network Management Protocol (SNMP) v2 and v3
- Syslog
- Simple Mail Transfer Protocol (SMTP)
- Key Management Interoperability Protocol (KMIP)
- HTML5 GUI
- HTML5 virtual Keyboard, Video, and Mouse (vKVM)
- Command-Line Interface (CLI)
- XML API

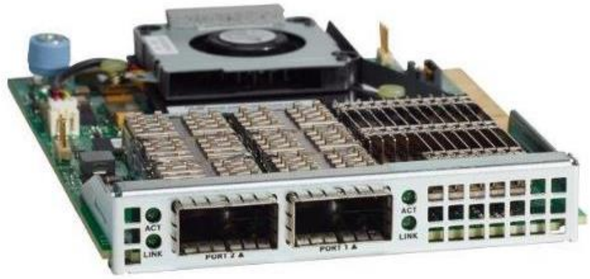
Management Software Development Kits (SDKs) and DevOps integrations exist for Python, Microsoft PowerShell, Ansible, Puppet, Chef, and more. For more information about integrations, see Cisco DevNet (https://developer.cisco.com/site/ucs-dev-center/).

The Cisco UCS C240 is Cisco Intersight™ ready. Cisco Intersight is a new cloud-based management platform that uses analytics to deliver proactive automation and support. By combining intelligence with automated actions, you can reduce costs dramatically and resolve issues more quickly.

Cisco UCS Virtual Interface Card 1387

The Cisco UCS Virtual Interface Card 1387 offers dual-port Enhanced Quad Small Form-Factor Pluggable (QSFP+) 40 Gigabit Ethernet and Fibre Channel over Ethernet (FCoE) in a modular-LAN-on-motherboard (mLOM) form factor. The mLOM slot can be used to install a Cisco VIC without consuming a PCIe slot providing greater I/O expandability.

Figure 2 Cisco UCS Virtual Interface Card 1387



The Cisco UCS VIC 1387 provides high network performance and low latency for the most demanding applications, including:

- Big data, high-performance computing (HPC), and high-performance trading (HPT)
- Large-scale virtual machine deployments
- High-bandwidth storage targets and archives

The card is designed for the M5 generation of Cisco UCS C-Series Rack Servers and Cisco UCS S3260 dense storage servers. It includes Cisco's next-generation converged network adapter technology and offers a comprehensive feature set, so you gain investment protection for future feature software releases.

The card can present more than 256 PCIe standards-compliant interfaces to its host. These can be dynamically configured as either network interface cards (NICs) or host bus adapters (HBAs).

This engine provides support for advanced data center requirements, including stateless network offloads for:

- Network Virtualization Using Generic Routing Encapsulation (NVGRE)
- Virtual extensible LAN (VXLAN)
- Remote direct memory access (RDMA)

The engine also offers support for performance optimization applications such as:

- Server Message Block (SMB) Direct
- Virtual Machine Queue (VMQ)
- Data Plane Development Kit (DPDK)
- Cisco NetFlow

Cisco UCS 6300 Series Fabric Interconnect

The Cisco UCS 6300 Series Fabric Interconnects are a core part of Cisco UCS, providing both network connectivity and management capabilities for the system (Figure 3). The Cisco UCS 6300 Series offers line-rate, low-latency, lossless 10 and 40 Gigabit Ethernet, Fibre Channel over Ethernet (FCoE), and Fibre Channel functions.

Figure 3 Cisco UCS 6300 Series Fabric Interconnect

The Cisco UCS 6300 Series provides the management and communication backbone for the Cisco UCS B-Series Blade Servers, 5100 Series Blade Server Chassis, Cisco UCS C-Series Rack Servers, and Cisco UCS S-Series Storage Dense Server managed by Cisco UCS. All servers attached to the fabric interconnects become part of a single, highly available management domain. In addition, by supporting unified fabric, the Cisco UCS 6300 Series provides both LAN and SAN connectivity for all servers within its domain.

From a networking perspective, the Cisco UCS 6300 Series uses a cut-through architecture, supporting deterministic, low-latency, line-rate 10 and 40 Gigabit Ethernet ports, switching capacity of 2.56 terabits per second (Tbps), and 320 Gbps of bandwidth per chassis, independent of packet size and enabled services. The product family supports Cisco® low-latency, lossless 10 and 40 Gigabit Ethernet unified network fabric capabilities, which increase the reliability, efficiency, and scalability of Ethernet networks. The fabric interconnect supports multiple traffic classes over a lossless Ethernet fabric from the server through the fabric interconnect. Significant TCO savings can be achieved with an FCoE optimized server design in which network interface cards (NICs), host bus adapters (HBAs), cables, and switches can be consolidated.

The Cisco UCS 6332 32-Port Fabric Interconnect is a 1-rack-unit (1RU) Gigabit Ethernet, and FCoE switch offering up to 2.56 Tbps throughput and up to 32 ports. The switch has 32 fixed 40-Gbps Ethernet and FCoE ports.

Both the Cisco UCS 6332UP 32-Port Fabric Interconnect and the Cisco UCS 6332 16-UP 40-Port Fabric Interconnect have ports that can be configured for the breakout feature that supports connectivity between 40 Gigabit Ethernet ports and 10 Gigabit Ethernet ports. This feature provides backward compatibility to existing hardware that supports 10 Gigabit Ethernet. A 40 Gigabit Ethernet port can be used as four 10 Gigabit Ethernet ports. Using a 40 Gigabit Ethernet SFP, these ports on a Cisco UCS 6300 Series Fabric Interconnect can connect to another fabric interconnect that has four 10 Gigabit

Ethernet SFPs. The breakout feature can be configured on ports 1 to 12 and ports 15 to 26 on the Cisco UCS 6332UP fabric interconnect. Ports 17 to 34 on the Cisco UCS 6332 16-UP fabric interconnect support the breakout feature.

Cisco Nexus 9336C-FX2 Switch

Based on Cisco Cloud Scale technology, the Cisco Nexus® 9300-EX and 9300-FX platforms are the next generation of fixed Cisco Nexus 9000 Series Switches. The new platforms support cost-effective cloud-scale deployments, an increased number of endpoints, and cloud services with wire-rate security and telemetry. The platforms are built on modern system architecture designed to provide high performance and meet the evolving needs of highly scalable data centers and growing enterprises.

Figure 4 Cisco Nexus 9336C-FX2 Switch

Cisco Nexus 9300-EX and 9300-FX platform switches offer a variety of interface options to transparently migrate existing data centers from 100-Mbps, 1-Gbps, and 10-Gbps speeds to 25 Gbps at the server, and from 10- and 40-Gbps speeds to 50 and 100 Gbps at the aggregation layer. The platforms provide investment protection for customers, delivering large buffers, highly flexible Layer 2 and Layer 3 scalability, and performance to meet the changing needs of virtualized data centers and automated cloud environments.

The Cisco Nexus 9336C-FX2 Switch is a 1RU switch that supports 7.2 Tbps of bandwidth and over 2.8 bpps. The switch can be configured to work as 1/10/25/40/100-Gbps offering flexible options in a compact form factor. All ports support wire-rate MACsec encryption. Breakout is supported on all ports.

The platform hardware is capable of collecting comprehensive Cisco Tetration Analytics™ telemetry information at line rate across all the ports without adding any latency to the packets or negatively affecting switch performance. This telemetry information is exported every 100 milliseconds by default directly from the switch's Application-Specific Integrated Circuit (ASIC). This information consists of three types of data:

- **Flow information**: This information contains information about endpoints, protocols, ports, when the flow started, how long the flow was active, and so on.
- **Interpacket variation**: This information captures any interpacket variations within the flow. Examples include variation in Time To Live (TTL), IP and TCP flags, payload length, and so on.
- **Context details**: Context information is derived outside the packet header, including variation in buffer utilization, packet drops within a flow, association with tunnel endpoints, and so on.

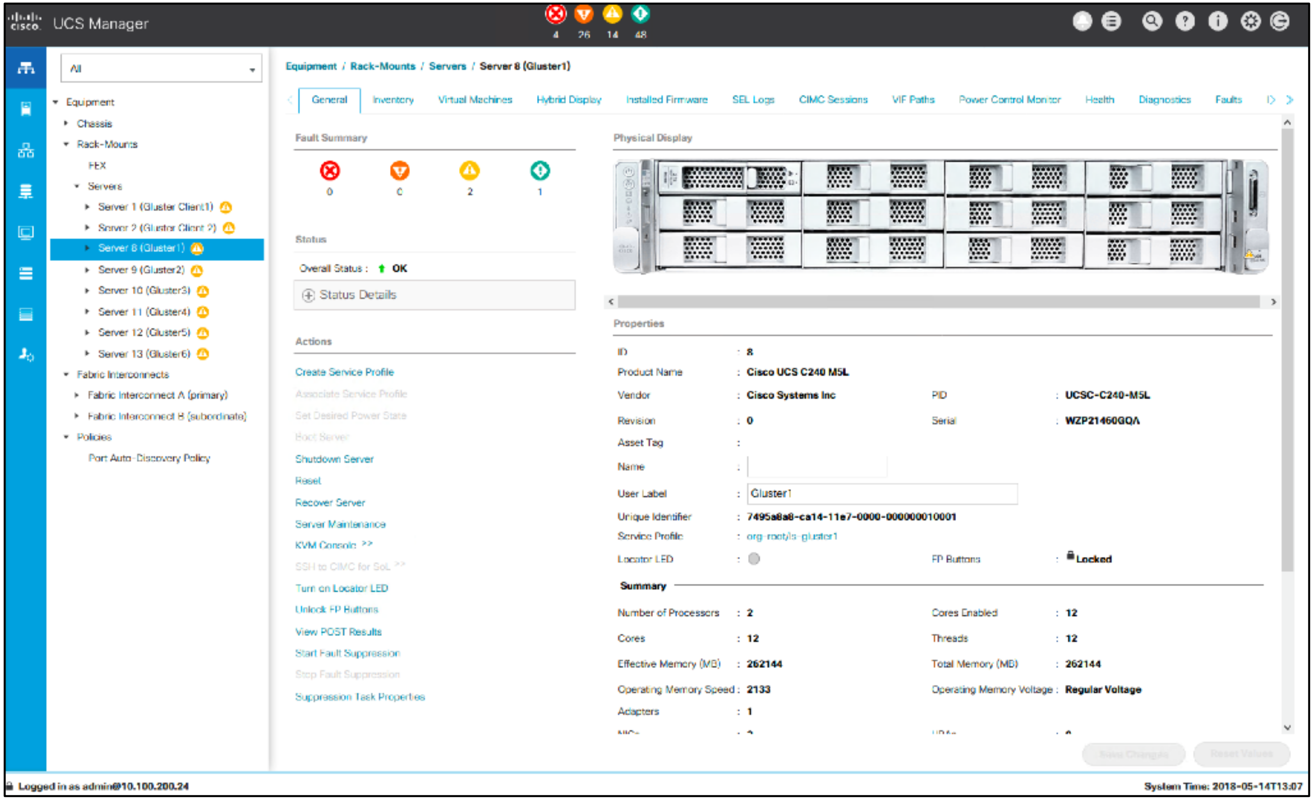
The Cisco Tetration Analytics platform consumes this telemetry data, and by using unsupervised machine learning and behavior analysis it can provide outstanding pervasive visibility across everything in your data center in real time. By using algorithmic approaches, the Cisco Tetration Analytics platform provides a deep application insights and interactions, enabling dramatically simplified operations, a zero-trust model, and migration of applications to any programmable infrastructure. To learn more, go to https://www.cisco.com/go/tetration.

Cisco provides two modes of operation for Cisco Nexus 9000 Series Switches. Organizations can use Cisco NX OS Software to deploy the switches in standard Cisco Nexus switch environments (NX-OS mode). Organizations also can use a hardware infrastructure that is ready to support the Cisco Application Centric Infrastructure (Cisco ACI™) platform to take full advantage of an automated, policy-based, systems-management approach (ACI mode).

Cisco UCS Manager

Cisco UCS® Manager (UCSM) (Figure 5) provides unified, embedded management of all software and hardware components of the Cisco Unified Computing System™ (Cisco UCS) across multiple chassis, rack servers, and thousands of virtual machines. It supports all Cisco UCS product models, including Cisco UCS B-Series Blade Servers, C-Series Rack Servers, and S- and M-Series composable infrastructure and Cisco UCS Mini, as well as the associated storage resources and networks. Cisco UCS Manager is embedded on a pair of Cisco UCS 6300 or 6200 Series Fabric Interconnects using a clustered, active-standby configuration for high availability. The manager participates in server provisioning, device discovery, inventory, configuration, diagnostics, monitoring, fault detection, auditing, and statistics collection.

Figure 5 Cisco UCS Manager



Overall Status : OK

ID : 8
Product Name : Cisco UCS C240 M5L
Vendor : Cisco Systems Inc
PID : UCSC-C240-M5L
Revision : 0
Serial : WZP21460GQA
Asset Tag :
Name :
User Label : Gluster1
Unique Identifier : 7495a8a8-ca14-11e7-0000-000000010001
Service Profile : org-root/ls-gluster1
Locator LED :
FP Buttons : Locked

Number of Processors : 2
Cores Enabled : 12
Cores : 12
Threads : 12
Effective Memory (MB) : 262144
Total Memory (MB) : 262144
Operating Memory Speed : 2133
Operating Memory Voltage : Regular Voltage
Adapters : 1

Logged in as admin@10.100.200.24

System Time: 2018-05-14T13:07

An instance of Cisco UCS Manager with all Cisco UCS components managed by it forms a Cisco UCS domain, which can include up to 160 servers. In addition to provisioning Cisco UCS resources, this infrastructure management software provides a model-based foundation for streamlining the day-to-day processes of updating, monitoring, and managing computing resources, local storage, storage connections, and network connections. By enabling better automation of processes, Cisco UCS Manager allows IT organizations to achieve greater agility and scale in their infrastructure operations while reducing complexity and risk. The manager provides flexible role- and policy-based management using service profiles and templates.

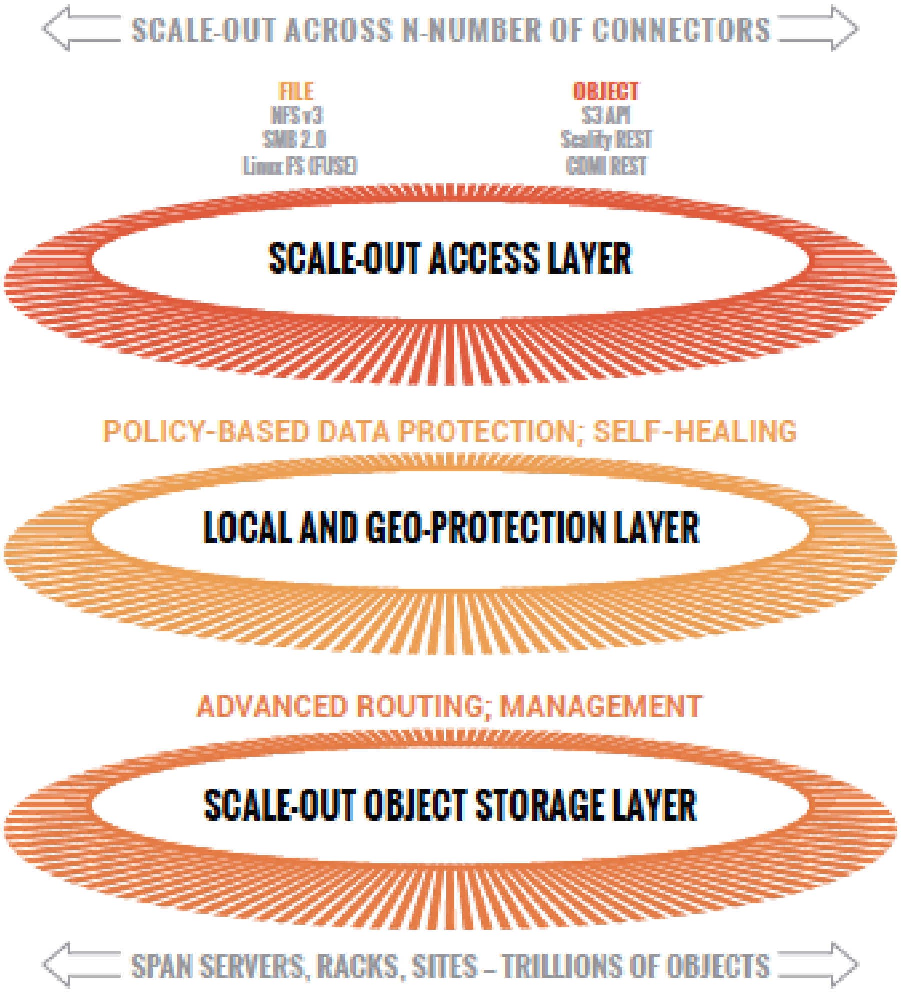
Cisco UCS Manager manages Cisco UCS systems through an intuitive HTML 5 or Java user interface and a CLI. It can register with Cisco UCS Central Software in a multi-domain Cisco UCS environment, enabling centralized management of distributed systems scaling to thousands of servers. Cisco UCS Manager can be integrated with Cisco UCS Director to facilitate orchestration and to provide support for converged infrastructure and Infrastructure as a Service (IaaS).

The Cisco UCS XML API provides comprehensive access to all Cisco UCS Manager functions. The API provides Cisco UCS system visibility to higher-level systems management tools from independent software vendors (ISVs) such as VMware, Microsoft, and Splunk as well as tools from BMC, CA, HP, IBM, and others. ISVs and in-house developers can use the XML API to enhance the value of the Cisco UCS platform according to their unique requirements. Cisco UCS PowerTool for Cisco UCS Manager and the Python Software Development Kit (SDK) help automate and manage configurations within Cisco UCS Manager.

Scality RING Overview

RING is a cloud-scale, distributed software solution for petabyte-scale unstructured data storage. It is designed to create unbounded scale-out storage systems for the many petabyte-scale applications and use cases, both object and file, that are deployed in today’s enterprise data centers. RING is a fully distributed system deployed on industry standard hardware, starting with a minimum of three (3) storage servers. The system can be seamlessly scaled-out to thousands of servers with 100’s of petabytes of storage capacity. RING has no single points of failure, and requires no downtime during any upgrades, scaling, planned maintenance or unplanned system events. With self-healing capabilities, it continues operating normally throughout these events. To match performance to increasing capacity, RING can also independently scale-out its access layer of protocol “Connectors”, to enable an even match of aggregate performance to the application load. RING provides data protection and resiliency through local or geo-distributed erasure-coding and replication, with services for continuous self-healing to resolve expected failures in platform components such as servers and disk drives. RING is fundamentally built on a scale-out object-storage layer that employs a second-generation peer-to-peer architecture. This approach uniquely distributes both the user data and the associated metadata across the underlying nodes to eliminate the typical central metadata database bottleneck. To enable file and object data in the same system, the RING integrates a virtual file system layer through an internal NoSQL scale-out database system, which provides POSIX-based access semantics using standard NFS, SMB and FUSE protocols with shared access to the files as objects using the REST protocol.

Figure 6 Scality RING Diagram


SCALE-OUT ACROSS N-NUMBER OF CONNECTORS
FILE
NFS v3
SMB 2.0
Linux FS (FUSE)
OBJECT
S3 API
Scality REST
CDMI REST
SCALE-OUT ACCESS LAYER
POLICY-BASED DATA PROTECTION; SELF-HEALING
LOCAL AND GEO-PROTECTION LAYER
ADVANCED ROUTING; MANAGEMENT
SCALE-OUT OBJECT STORAGE LAYER
SPAN SERVERS, RACKS, SITES – TRILLIONS OF OBJECTS

Scality has designed RING along the design criteria spearheaded by the leading cloud-scale service providers, such as Google, Facebook, and Amazon. RING leverages loosely-coupled, distributed systems designs that leverage commodity, mainstream hardware along the following key tenets:

- 100 percent parallel design for metadata and data - to enable scaling of capacity and performance to unbounded numbers of objects, with no single points of failures, service disruptions, or forklift upgrades as the system grows.
- Multi-protocol data access – to enable the widest variety of object, file and host-based applications to leverage RING storage.
- Flexible data protection mechanisms - to efficiently and durably protect a wide range of data types and sizes.
- Self-healing from component failures – to provide high-levels of data durability, the system expects and tolerates failures and automatically resolves them.
- Hardware freedom – to provide optimal platform flexibility, eliminate lock-in and reduce TCO.

RING incorporates these design principles at multiple levels, to deliver the highest levels of data durability, at the highest levels of scale, for most optimal economics.

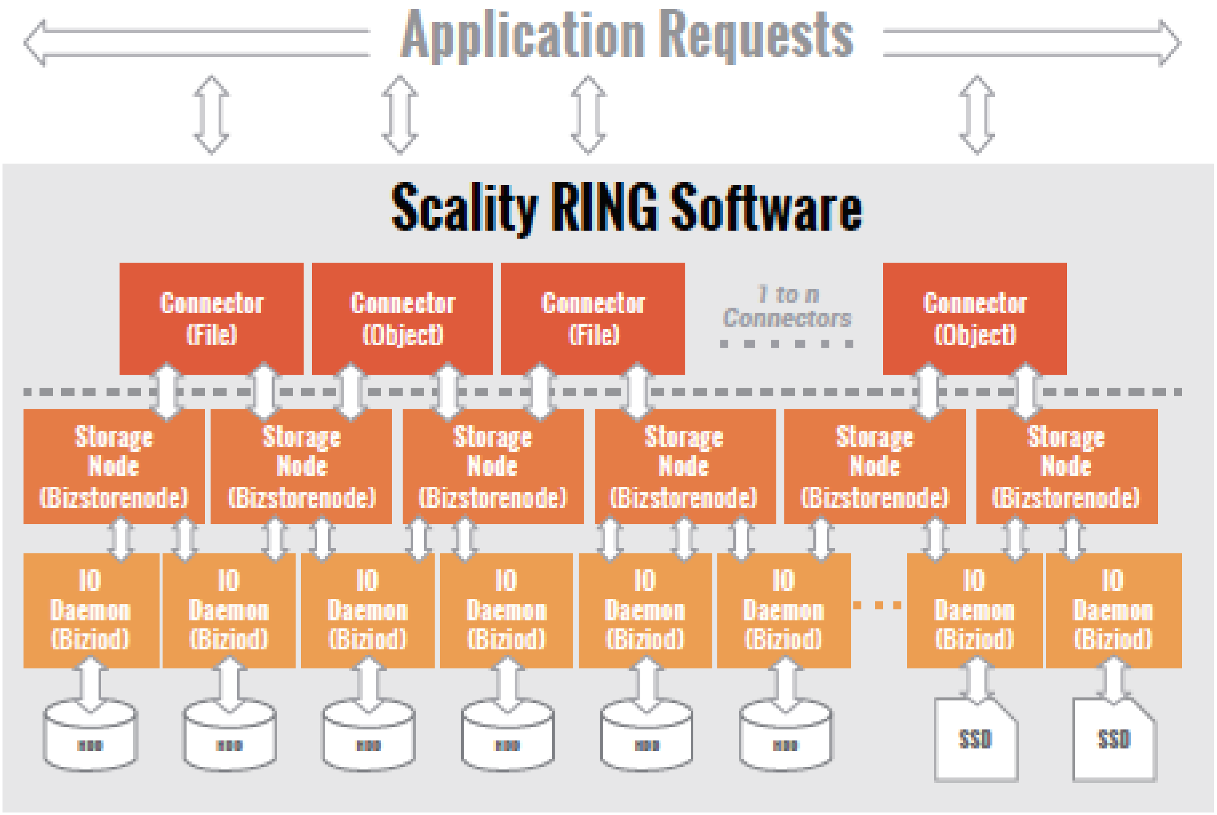
Scality RING Architecture

To scale both storage capacity and performance to massive levels, the Scality RING software is designed as a distributed, parallel, scale-out architecture with a set of intelligent services for data access and presentation, data protection and systems management. To implement these capabilities, RING provides a set of fully abstracted software services including a top-layer of scalable access services (Connectors) that provide storage protocols for applications. The middle layers are comprised of a distributed virtual file system layer, a set of data protection mechanisms to ensure data durability and integrity, self-healing processes and a set of systems management and monitoring services. At the bottom of the stack, the system is built on a distributed storage layer comprised of virtual storage nodes and underlying IO daemons that abstract the physical storage servers and disk drive interfaces.

At the heart of the storage layer is a scalable, distributed object key/value store based on a second-generation peer-to-peer routing protocol. This routing protocol ensures that store and lookup operations scale efficiently to very high numbers of nodes.

RING software is comprised of the following main components: RING Connectors, a distributed internal NoSQL database called MESA, RING Storage Nodes and IO daemons, and the Supervisor web-based management portal. The MESA database is used to provide object indexing, as well as the integral Scale-Out-File-System (SOFS) file system abstraction layer, and the underlying core routing protocol and Keyspace mechanisms are described later in this paper.

Figure 7 Scality Scale-out Architecture



RING Connectors

The Connectors provide the data access endpoints and protocol services for applications that use RING for data storage. As a scale-out system, RING supports any number of Connectors and endpoints to support large and growing application workloads. The RING 7 release provides a family of object and file interfaces:

- AWS S3 API - a comprehensive implementation of the AWS S3 REST API, with support for the Bucket and Object data model, AWS style Signature v4/v2 authentication, and the AWS model of Identity and Access Management (IAM)
- http/REST (sproxyd) - the RING's native key/value REST API, provides a flat object storage namespace with direct access to RING objects
- NFS v3 - SOFS volumes presented as a standard NFSv3 mount points
- SMB 2.0 - SOFS volumes presented as SMB Shares to Microsoft Windows clients. Several SMB 3.0 features are currently supported, a later release will provide full SMB 3.0 support
- FUSE - SOFS volumes presented as a local Linux file system
- CDMI/REST - support for the SNIA CDMI REST interface, with full compatibility to SOFS file data
- S3 on SOFS - SOFS volumes may be accessed in read-only mode over the S3 protocol, for namespace and data sharing between objects and files
- NFS v4/v3 on S3 - S3 buckets may be exported as NFS v4/v3 mount points

Connectors provide storage services for read, write, delete and lookup for objects or files stored into the RING based on either object or POSIX (file) semantics. Applications can make use of multiple connectors in parallel to scale out the number of operations per second, or the aggregate throughput of the RING. A RING deployment may be designed to provide a mix of file access and object access (over NFS and S3 for example), simultaneously—to support multiple application use cases.

Storage Nodes and IO Daemons

The heart of the ring are the Storage Nodes, that are virtual processes that own and store a range of objects associated with its portion of the RING's keyspace. A complete description of RING's keyspace mechanism is provided in the next section. Each physical storage server (host) is typically configured with six (6) storage nodes (termed bizstorenode). Under the storage nodes are the storage daemons (termed biziod), which are responsible for persistence of the data on disk, in an underlying local standard disk file system. Each biziod instance is a low-level software process that manages the IO operations to a particular physical disk drive and maintains the mapping of object keys to the actual object locations on disk. Biziod processes are local to a given server, managing only local, direct-attached storage and communicating only with Storage Nodes on the same server. The typical configuration is one biziod per physical disk drive, with support for up to hundreds of daemons2 per server, so the system can support very large, high-density storage servers.

Each biziod stores object payloads and metadata in a set of fixed size container files on the disk it is managing. With such containerization the system can maintain high-performance access even to small files, without any storage overhead. The bizoid deamons typically leverage low-latency flash (SSD or NVMe) devices to store the index files for faster lookup performance. The system provides data integrity assurance and validation through the use of stored checksums on the index and data container files, which are validated upon read access to the data. The use of a standard file system underneath biziod ensures that administrators can use normal operating system utilities and tools to copy, migrate, repair and maintain the disk files if required.

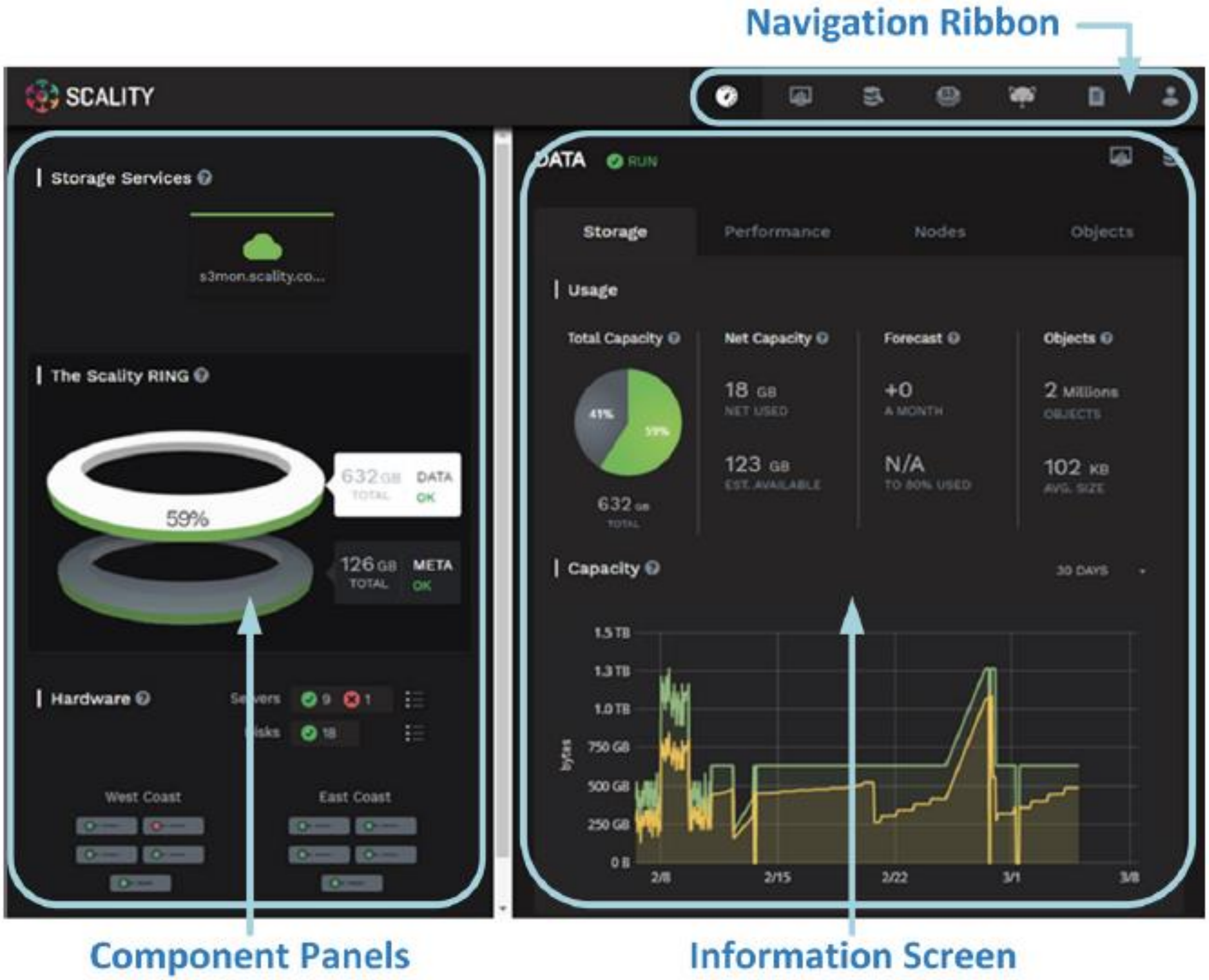
The recommended deployment for systems that have both HDD and SSD media on the storage servers is to deploy a data RING on HDD, and the associated metadata in a separate RING on SSD. Typically, the requirements for metadata are approximately 2 percent of the storage capacity of the actual data, so the sizing of SSD should follow that percentage for best effect. Scality can provide specific sizing recommendations based on the expected average file sizes, and number of files for a given application.

RING Systems Management

Managing and monitoring the RING is enabled through a cohesive suite of user interfaces, built on top of a family of RESTful interfaces termed the Supervisor API ("SupAPI"). The SupAPI provides an API based method that may be accessed from scripts, tools and frameworks for gathering statistics, metrics, health check probes and alerts, and for provisioning new services on the RING. The SupAPI is also enabled with Role Based Access Control (RBAC), by supporting an administrator identity to provide access control privileges for Super-Admin and Monitor admin user Roles.

RING provides a family of tools that use the SupAPI for accessing the same information and services. RING 7 includes the new "Scality Supervisor", a browser-based portal for both systems monitoring and management of Scality components. In RING 7, the Supervisor now provides capabilities across object (S3) and file (NFS, SMB, FUSE) Connectors including integrated dashboards including Key Performance Indicators (KPIs) with trending information such as "Global Health", "Performance", "Availability" and "Forecast". The Supervisor also includes provisioning capabilities to add new servers in the system and a new zone management module to handle customer failure domains for multi-site deployments.

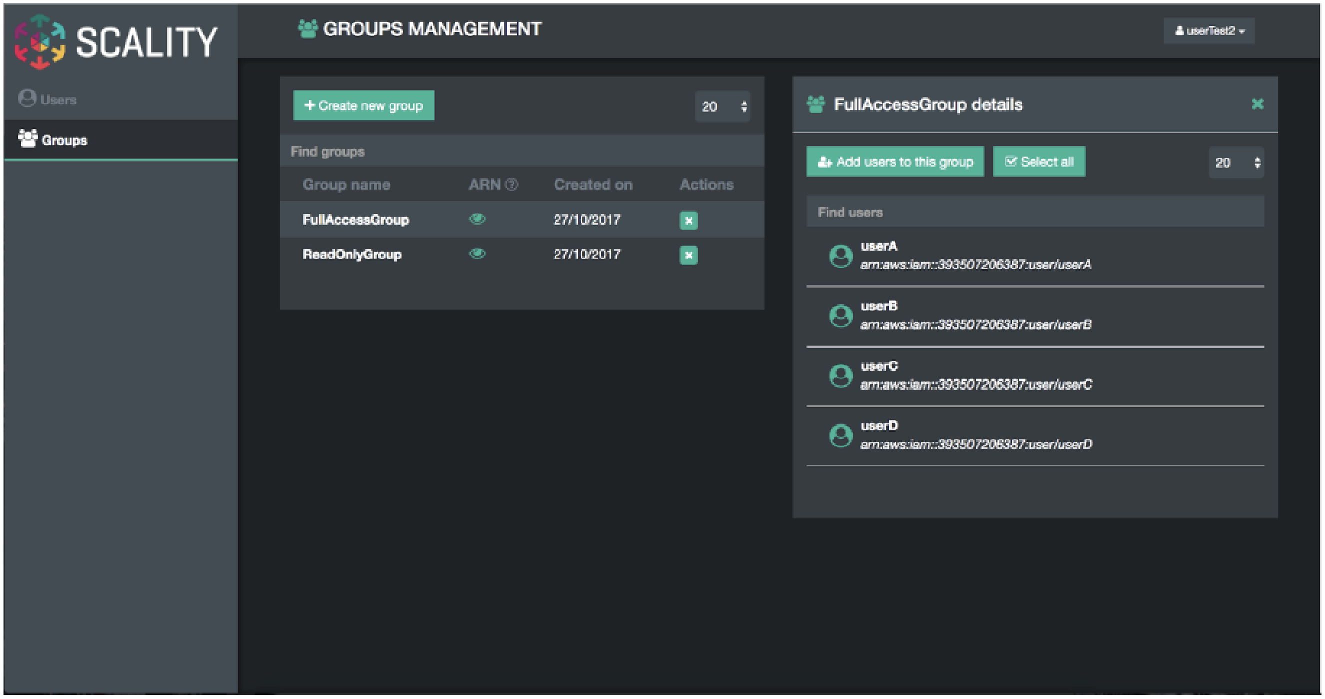
Figure 8 Supervisor Web GUI



RING Supervisor also includes an "Advanced Monitoring" dashboard where all collected metrics can be graphed and analyzed component per-component and per-server. This is based on a very powerful graphing engine that has access to thousands of metrics.

A new "S3 Service Management console" portal is provided to manage the integrated AWS Identity and Access Management (IAM) model of S3 multi-tenancy in the RING. This provides two-level management of Accounts, Users/Groups and IAM access control policies. The S3 Console may also be easily customized for white-labeling purposes.

Figure 9 S3 Service Management Console



GROUPS MANAGEMENT

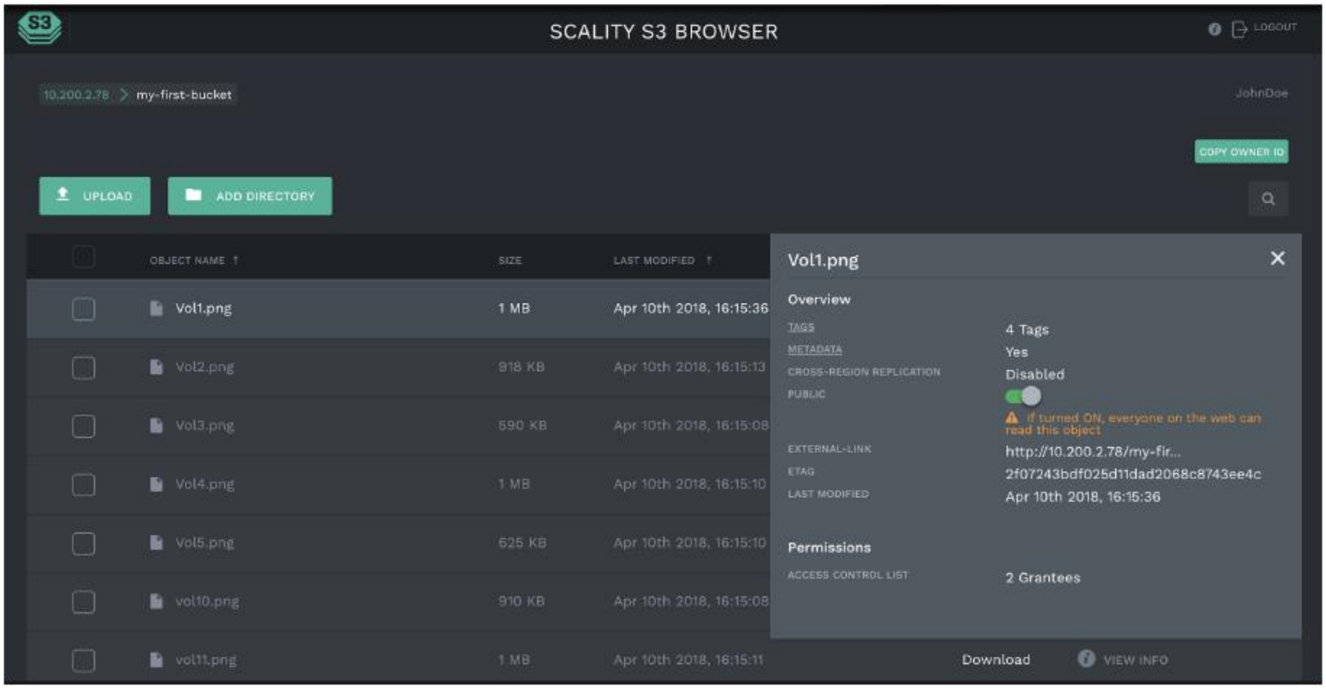
Group name	ARN	Created on	Actions
FullAccessGroup	👁	27/10/2017	✖
ReadOnlyGroup	👁	27/10/2017	✖

FullAccessGroup details

- userA arn:aws:iam::393507206387:user/userA
- userB arn:aws:iam::393507206387:user/userB
- userC arn:aws:iam::393507206387:user/userC
- userD arn:aws:iam::393507206387:user/userD

A new **"Scality S3 Browser"** is also provided to browse S3 buckets, upload and download object data, and for managing key S3 features such as bucket versioning, CORS, editing of metadata attributes and tagging. The S3 Browser is an S3 API client that runs on the S3 user browser and is accessible to both the Storage administrator and also to the S3 end-user.

Figure 10 Scality S3 Browser



	OBJECT NAME ↑	SIZE	LAST MODIFIED ↑
□	Vol1.png	1 MB	Apr 10th 2018, 16:15:36
□	Vol2.png	918 KB	Apr 10th 2018, 16:15:13
□	Vol3.png	590 KB	Apr 10th 2018, 16:15:08
□	Vol4.png	1 MB	Apr 10th 2018, 16:15:10
□	Vol5.png	625 KB	Apr 10th 2018, 16:15:10
□	vol10.png	910 KB	Apr 10th 2018, 16:15:08
□	vol11.png	1 MB	Apr 10th 2018, 16:15:11

Vol1.png

Overview

TAGS	4 Tags
METADATA	Yes
CROSS-REGION REPLICATION	Disabled
PUBLIC	⚠ If turned ON, everyone on the web can read this object
EXTERNAL-LINK	http://10.200.2.78/my-fir...
ETAG	2f07243bdf025d11dad2068c8743ee4c
LAST MODIFIED	Apr 10th 2018, 16:15:36

Permissions

ACCESS CONTROL LIST	2 Grantees

A scriptable Command Line Interface (CLI) called RingSH is also provided, as well as an SNMP compatible MIB and traps interface for monitoring from standard SNMP consoles. RING is designed to be self-managing and autonomous to free administrators to work on other value-added tasks, and not worry about the component level management tasks common with traditional array-based storage solutions.

S3 Connector: AWS S3 Storage with Identity and Access Management (IAM)

The Scality S3 Connector provides a modern S3 compatible application interface to the Scality RING. The AWS S3 API has now become the industry's default cloud storage API and has furthermore emerged as the standard RESTful dialect for object storage as NFS was for the NAS generation. The S3 Connector is built on a distributed scale-out architecture to support very high levels of application workloads and concurrent user access. This is based on a highly-available, high-performance metadata engine that can also be scaled-out for increased performance. Deployments can be geo-replicated deployments to enable highly-available disaster recovery solutions, for both Metro-Area Network environments ("stretched" deployments), as well as Cross Region Replication (CRR) asynchronous replication of individual S3 buckets or a full site.

The Scality S3 Connector provides also provides a full implementation of the AWS multi-tenancy and identity management (AWS IAM) model with federated authentication to LDAP and Active Directory to integrate into enterprise deployment environments. In addition to the RING Supervisor management UI, the S3 Service Provider UI is a web-based user interface to manage multi-tenancy accounts, users, group and policies. To support enterprise security, development and operational methodologies, the S3 Connector on RING supports:

- Integration with Enterprise directory/security servers: most commonly Microsoft Active Directory or LDAP servers. Federated authentication integration is supported through a SAML 2.0-compatible Identity Provider such as Microsoft ADFS, and many other SAML compatible products, to enable a complete Single Sign-On (SSO) solution.

- Secure Multi-tenancy support: through IAM Accounts, secure access keys, Users, Groups, access control policies and v4 authentication per-tenant, bucket encryption (integrated with corporate KMS solutions) and auditing

- Utilization reporting to enable chargeback: the S3 Connector Utilization API provides an extended API for reporting on comprehensive usage metrics including capacity, #objects, bandwidth and S3 operations (per unit time). This provides all of the metrics required for consumption into corporate tools for chargeback calculations.

- Cloud-based S3 monitoring: integration through Scality's upcoming Cloud Monitoring console, and load-balancer driven health checks

- High-performance, scale-out access: to support many corporate applications and workloads simultaneously reading and writing to the S3 service

- Highly-available disaster-recovery solutions: enabled deployments through multi-data center deployments to provide availability in the event of site failure

In RING 7, the feature set for the S3 Connector now supports Bucket Versioning via the S3 API, and for Cross Region Replication (CRR) of Buckets through the S3 API, this provides bucket-level asynchronous replication to another S3/RING deployment.

Scale-Out-File-System (SOFS)

RING supports native file system access to RING storage through the integrated Scale-Out-File-System (SOFS) with NFS, SMB and FUSE Connectors for access over these well-known file protocols. SOFS is a POSIX compatible, parallel file system that provides file storage services on the RING without the need for external gateways.

SOFS is more precisely a virtual file system, which is based on an internal distributed database termed MESA ("table" in Spanish) on top of the RING's storage services. MESA is a distributed, semi-structured database that is used to store the file system directories and file inode structures. This provides the virtual file system hierarchical view, with the consistency required for file system data, by ensuring that file system updates are always atomic. This means that updates are either committed or rolled back entirely—which guarantees the file system is never left in an intermediate or inconsistent state. A key advantage for scaling is that MESA is itself is distributed as a set of objects across all of the RING's storage node in a shared nothing manner to eliminate any bottlenecks or limitations.

File system lookups are performed using the RING's standard peer-to-peer routing protocol. For fast access performance, SOFS metadata is recommended to be stored on flash storage, typically on its own dedicated SSD drives in the storage servers, with the SOFS file payloads stored in the "data RING" on hard disk drives (HDDs). SOFS works directly with the data protection and durability mechanisms present in the RING, including replication and configurable Erasure Coding schemas.

SOFS can be provisioned into one or more "volumes" and can be scaled in capacity as needed to support application requirements. Each volume can be accessed by any number of Connectors to support the incoming load workload, even with mixed protocols (NFS, SMB or FUSE). RING can support an enormous number of volumes (up to 232) and can grow to billions of files per volume. There is no need to pre-configure volumes for capacity (the RING effectively supports thin-provisioning of volumes). Volumes will utilize the RING's storage pool to expand as needed when files are created and updated. For efficient storage of very large files, the RING supports the concept of "sparse files", effectively files combined from multiple individual data-stripes.

While multiple Connectors may be used to simultaneously access a volume, the RING currently supports scale-out access for multiple concurrent readers, and a new "File Access Coordination" mode that allows multiple readers on a file while it is being written from another Connector. This is useful in use-cases such as video streaming where very large video files are written over the course of minutes or hours, but the file must be accessed for content distribution before the write is complete. Multiple Connectors attempt to write to the same directory or one per file within a directory, SOFS maintains view consistency across multiple connectors. By supporting scale-out across any number of Connectors, SOFS throughput can be scaled out to support increasing workload demands. When performance saturation is reached, it is always possible to add more connectors or storage nodes (and disk spindles) to the RING to further increase throughput into the system. The system can achieve 10's of Gigabytes per second of aggregate throughput for parallel workloads through this architecture.

SOFS provides volume-level utilization metering and quota support, in addition to User and Group (uid/gid) quotas. This enables administrators to effectively use the concept of volumes to meter, report and limit space (capacity) usage at the volume level. This is useful in a multi-tenant environment where multiple applications or use cases are sharing the same RING, but accessing data stored in their own volume.

SOFS also provides integrated failover and load balancer services for the NFS and SMB Connectors. The load balancer uses an integrated DNS service to expose one or more service names (e.g., sofs1.companyname. com) on Virtual IP addresses (VIPs), which can be mounted as NFS mount points or SMB shares. The load balancer can be configured with multiple underlying NFS or SMB connector real IP addresses, and provides load balancing of file traffic across these SOFS connectors. In combination with the RING 6.0 Folder Scale-out feature, this also provides transparent multi-connector access to a single folder, as well as enabling failover. In the event one of the underlying NFS or SMB Connectors becomes non-responsive, the load balancer can select another Connector IP address as the access point for the request.

Intelligent Data Durability and Self-Healing

RING is designed to expect and manage a wide range of component failures including disks, server networks and even across multiple data centers, while ensuring that data remains durable and available during these conditions. RING provides data durability through a set of flexible data protection mechanisms optimized for distributed systems, including replication, erasure coding and geo-replication capabilities that allow applications to select the best data protection strategies for their data. These flexible data protection mechanisms implement Scality's design principle to address a wide spectrum (80 percent) of storage workloads and data sizes. A full description of multi-site data protection is provided in the next section, Multi-Site Geo-Distribution.

Replication Class of Service (COS)

To optimize data durability in a distributed system, the RING employs local replication, or the storage of multiple copies of an object within the RING. RING will attempt to spread these replicas across multiple storage nodes, and across multiple disk drives, in order to separate them from common failures (assuming sufficient numbers of servers and disks are available). RING supports six Class-of-Service levels for replication (0-5), indicating that the system can maintain between 0 to 5 replicas (or 1-6 copies) of an object. This allows the system to tolerate up to 5 simultaneous disk failures, while still preserving access and storage of the original object. Note that any failure will cause the system to self-heal the lost replica, to automatically bring the object back up to its original Class-of-Service, as fast as possible.

While replication is optimal for many use cases where the objects are small, and access performance is critical, it does impose a high storage overhead penalty compared to the original data. For example, a 100KB object being stored with a Class-of-Service=2 (2 extra copies so 3 total), will therefore consume 3 x 100KB = 300KB of actual physical capacity on the RING, in order to maintain its 3 replicas. This overhead is acceptable in many cases for small objects but can become a costly burden for megabyte or gigabyte level video and image objects. In this case, paying a penalty of 200% to store a 1GB object since it will require 3GB of underlying raw storage capacity for its 3 replicas. When measured across petabytes of objects, this becomes a significant cost burden for many businesses, requiring a more efficient data protection mechanism.

Flexible Erasure Coding

Scality's erasure coding (EC) provides an alternative data protection mechanism to replication that is optimized for large objects and files. RING implements Reed-Solomon erasure coding6 techniques, to store large objects with an extended set of parity "chunks", instead of multiple copies of the original object. The basic idea with erasure coding is to break an object into multiple chunks (m) and apply a mathematical encoding to produce an additional set of parity chunks (k). A description of the mathematical encoding is beyond the scope of this paper, but they can be simply understood as an extension of the XOR parity calculations used in traditional RAID. The resulting set of chunks, (m+k) are then distributed across the RING nodes, providing the ability to access the original object as long as any subset of m data or parity chunks are available. Stated another way, this provides a way to store an object with protection against k failures, with only k/m overhead in storage space.

Many commercial storage solutions impose a performance penalty on reading objects stored through erasure coding, due to the fact that all of the chunks, including the original data, are encoded before they are stored. This requires mandatory decoding on all access to the objects, even when there are no failure conditions on the main data chunks. With Scality's EC, the data chunks are stored in the clear, without any encoding, so that this performance penalty is not present during normal read accesses. This means that EC data can be accessed as fast as other data, unless a data chunk is missing which would require a parity chunk to be accessed and decoded. In summary, for single-site data protection, Scality's replication and EC data protection mechanisms can provide very high-levels of data durability, with the ability to trade-off performance and space characteristics for different data types.

Note that replication and EC may be combined, even on a single Connector, by configuring a policy for the connector to store objects below a certain size threshold with a replication CoS, but files above the file size limit with a specific EC schema. This allows the application to simply store objects without worrying about the optimal storage strategy per object, with the system managing that automatically.

Note that RING does not employ traditional RAID based data protection techniques. While RAID has served the industry well in legacy NAS and SAN systems, industry experts have written at large about the inadequacies of classical RAID technologies when employed on high-density disk drives, in capacity-optimized and distributed storage systems. These deficiencies include higher probabilities of data loss due to long RAID rebuild times, and the ability to protect against only a limited set of failure conditions (for example, only two simultaneous disk failures per RAID6 group). Further information and reading on the limitations of RAID as a data protection mechanism on high-capacity disk drives is widely available

Self-healing

RING provides self-healing processes that monitor and automatically resolve component failures. This includes the ability to rebuild missing data chunks due to disk drive or server failures, rebalance data when nodes leave and join the RING, and to proxy requests around component failures. In the event a disk drive or even a full server fails, background rebuild operations are spawned to restore the missing object data from its surviving replicas or EC chunks. The rebuild process completes when it has restored the original Class of Service - either the full number of replicas or the original number of EC data and parity chunks. A local disk failure can also be repaired quickly on a node (distinct from a full distributed rebuild), through the use of an in-memory key map maintained on each node. Nodes are also responsible for automatically detecting mismatches in their own Keyspace, rebalancing keys and for establishing and removing proxies during node addition and departure operations. Self-healing provides the RING with the resiliency required to maintain data availability and durability in the face of the expected wide set of failure conditions, including multiple simultaneous component failures at the hardware and software process levels.

To optimize rebuilds as well as mainline IO performance during rebuilds, RING utilizes the distributed power of the entire storage pool. The parallelism of the underlying architecture pays dividends by eliminating any central bottlenecks that might otherwise limit performance or cause contention between servicing application requests, and normal background operations such as rebuilds, especially when the system is under load. To further optimize rebuild operations, the system will only repair the affected object data, not the entire set of disk blocks, as is commonly the case in RAID arrays. Rebuilds are distributed across multiple servers and disks in the system, to utilize the aggregate processing power and available IO of multiple resources in parallel, rather than serializing the rebuilds onto a single disk drive.

By leveraging the entire pool, the impact of rebuilding data stored either with replication or EC is minimized since there will be relatively small degrees of overlap between disks involved in servicing data requests, and those involved in the rebuilds.

Scality RING Multi-Site Deployments

To support multi datacenter deployments with site protection and complete data consistency between all sites, the RING natively supports a stretched (synchronous) deployment mode across sites. In this mode, a single logical RING is deployed across multiple data centers, with all nodes participating in the standard RING protocols as if they were local to one site.

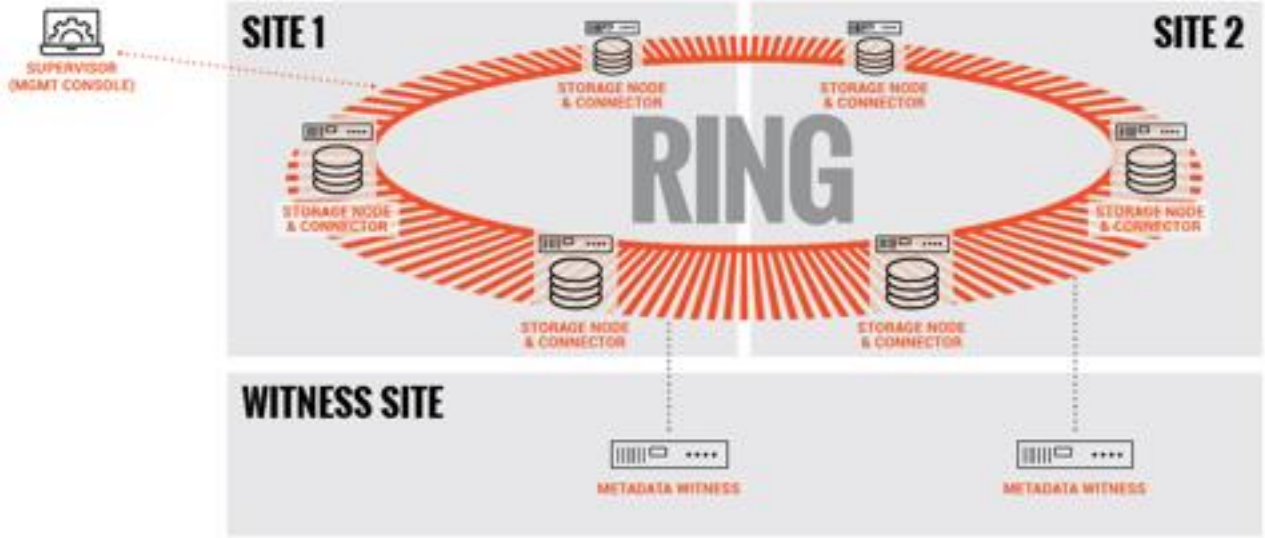
When a stretched RING is deployed with EC, it provides multiple benefits including full site-level failure protection, active/active access from both data centers, and dramatically reduced storage overhead compared to mirrored RINGs. An EC schema for a three-site stretched RING of EC (7,5) would provide protection against one complete site failure, or up to four disk/server failures per site, plus one additional disk/server failure in another site, with approximately 70 percent space overhead. This compares favorably to a replication policy that might require 300-400 percent space overhead, for similar levels of protection across these sites.

File System (SOFS) Multi-Site Geo-Distribution

The Scality RING can be stretched across 2 to 3 sites within a Metro-Area Network (MAN) to provide full site failover, should the latency between the several sites go above 10ms. The stretched architecture provides zero RTO and RPO since the failover is automatized. This is the same for the failback procedure since when the lost site is recovered, the system will recover automatically the data. For the two-site stretched architecture only and to manage the mitigation between the 2 sites, 2 witness servers will be needed.

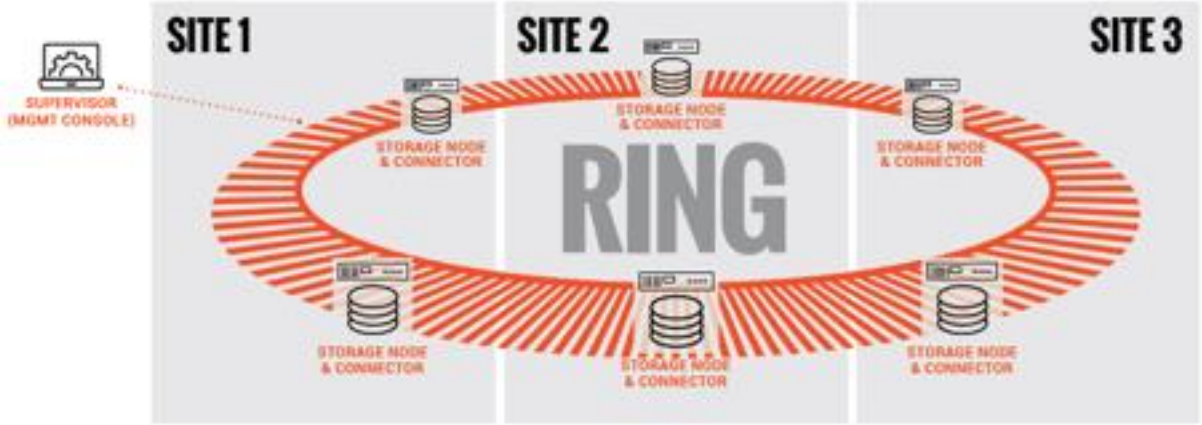
The 2 stretched sites + witness is an Active / Active replication system based on a synchronous replication.

Figure 11 SOFS – Two-site Stretched



The 3 stretched sites is an Active / Active replication system based on a synchronous replication.

Figure 12 SOFS – Three-Site Stretched

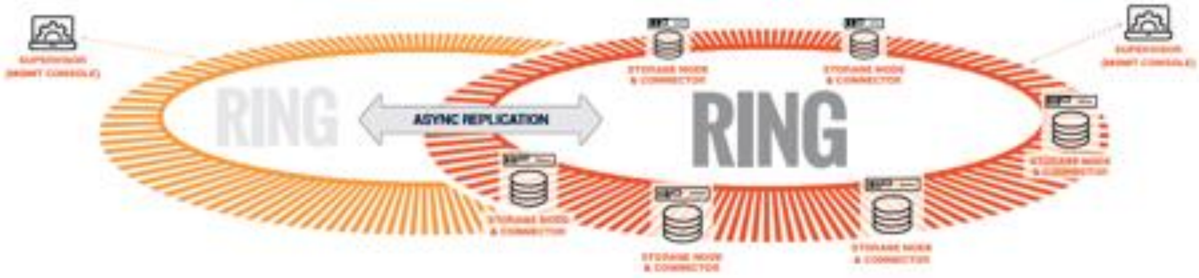


For high latency between sites, Scality supports SOFS 2 Sites Full Asynchronous replication mechanism at Scale to enable the replication of massive amount of file data across the 2 sites. Scality also supports a Full Diff mechanism that can compare at scale the content of the 2 sites to ensure the data are effectively replicated. Should one site be fully lost, Scality provides a mechanism to fully reconstruct the lost site.

To manage replication burst, Scality integrates a back-pressure system to be sure your production network link won't be overloaded by the replication itself and in the same time will respect the RPO defined during the setup of the installation. This feature enables the Disaster Recovery (DR) feature by providing Failover and Failback system to recover in case of partial or full loss.

The 2 sites with high latency between them is an Active / Passive replication system based on an asynchronous replication.

Figure 13 SOFS – Two-Site Asynchronous Replication

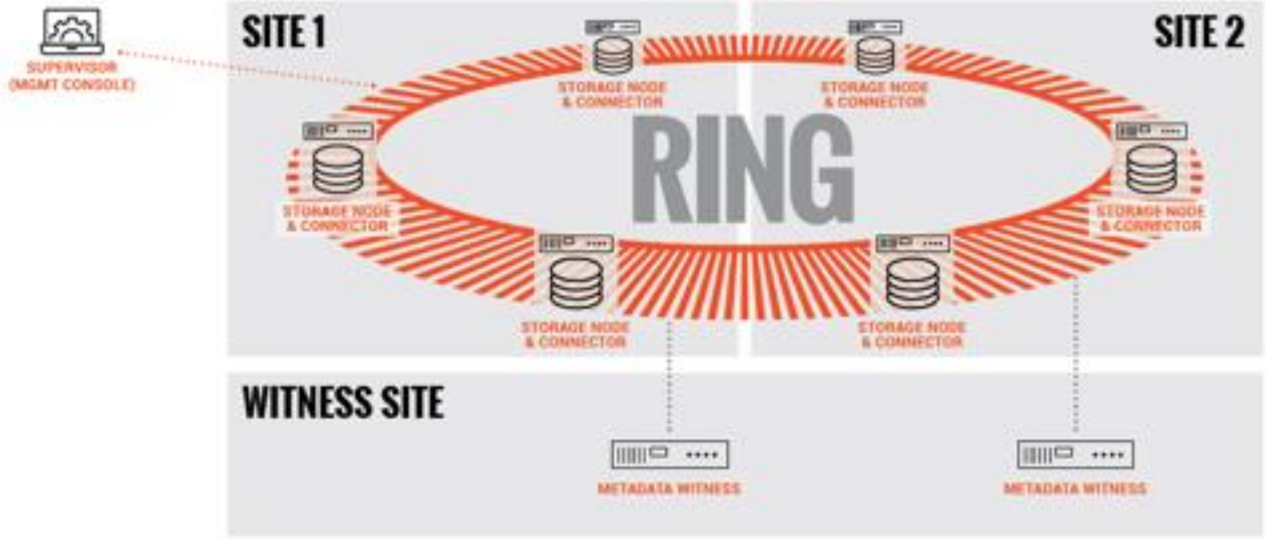


S3 Object Multi-Site Geo-Distribution

The same multi-site architectures are supported for S3 as with SOFS, both synchronous & asynchronous. The first one with a stretched solution on two and three sites with no RPO and no RTO. As for SOFS, a stretched architecture is available within a MAN to provide full site failover. Should the latency between the several sites goes above 10ms. For the two-site stretched architecture only and to manage the mitigation between the 2 sites, 2 witness servers will be needed.

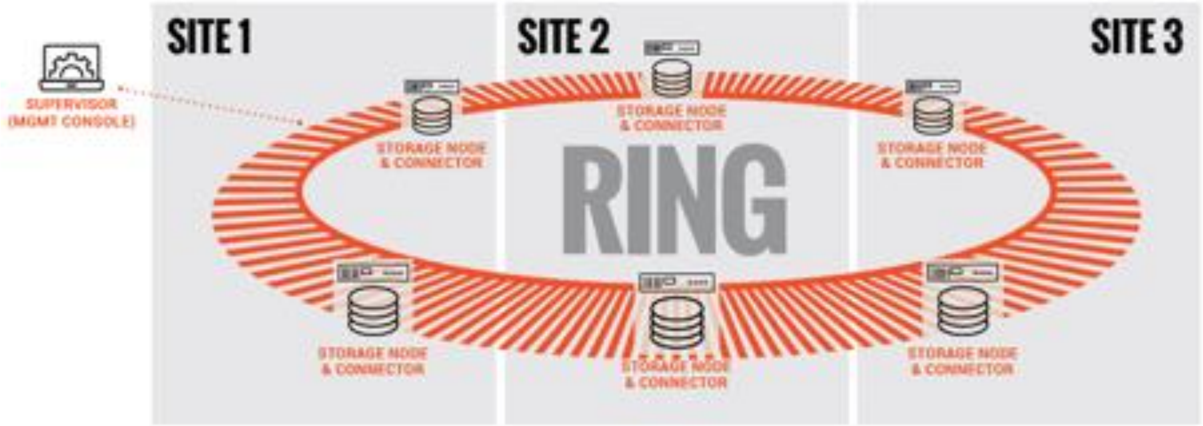
The 2 stretched sites + witness is an Active / Active replication system based on a synchronous replication.

Figure 14 S3 – Two-site Stretched



The 3 stretched sites is an Active / Active replication system based on a synchronous replication.

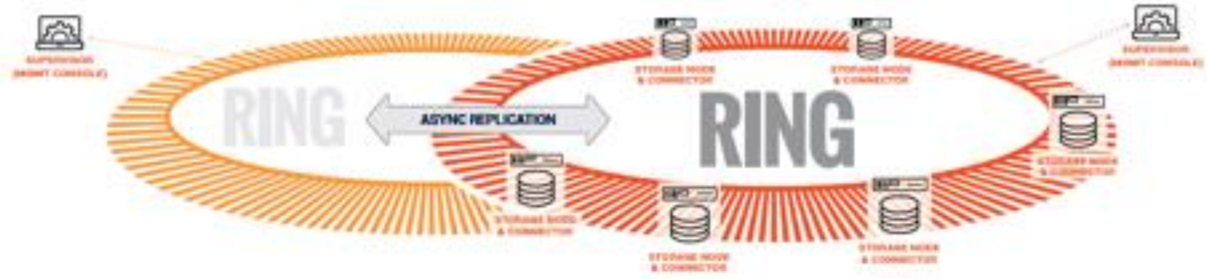
Figure 15 S3 – Three-Site Stretched



For high latency between sites (such as on a Wide Area Network - WAN), Scality supports the S3 2 Sites Full Asynchronous replication mechanism at Scale to enable the replication of massive amount of data across the 2 sites. This system is based on the S3 CRR design to replicate a bucket between 2 sites. For site replication, Scality developed its own system to support site replication instead of just bucket. This feature enables the Disaster Recovery (DR) feature by providing Failover and Failback system to recover in case of partial or fully (flooding, fire, and so on) lost.

The 2 sites with high latency between them is an Active / Passive replication system based on an asynchronous replication.

Figure 16 S3 – Two-Site Asynchronous Replication



Solution Design

Solution Overview

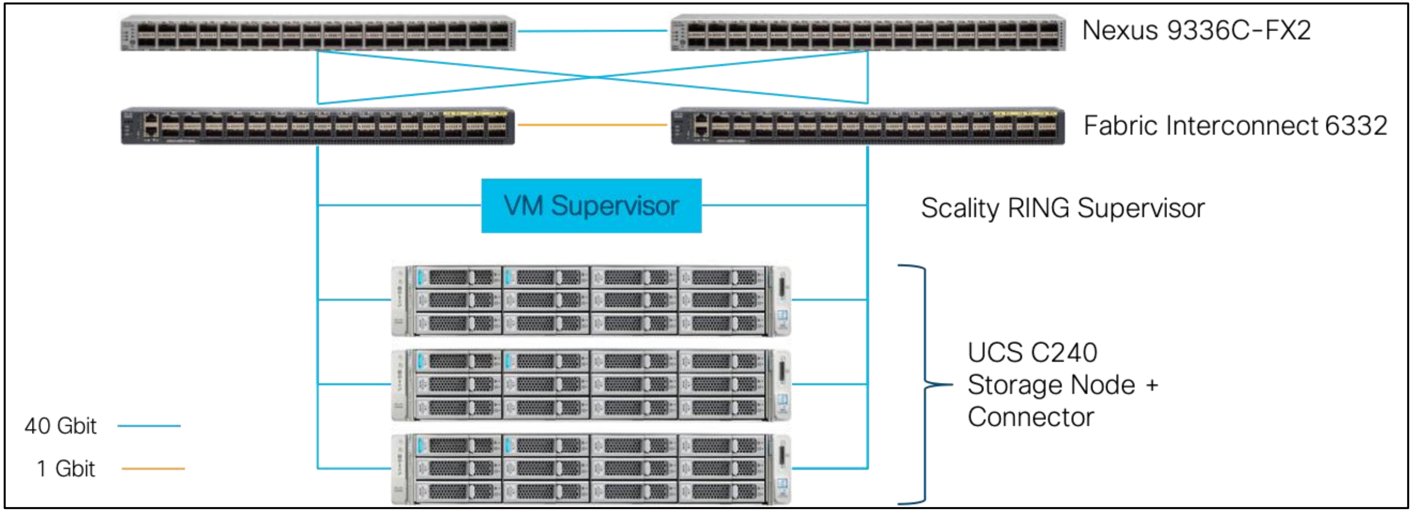
This Cisco Validated Design provides a comprehensive, end-to-end guide for designing Scality RING on Cisco UCS C240 within infrastructure made possible by Cisco UCS Manager and the Cisco UCS 6332 Fabric Interconnects.

One of the key design goals of this scale out architecture was to deploy all elements on 40GbE networking end to end within a single Cisco UCS domain and start small with Scality RING. Both Scality components – Storage Node and Connector – utilize the robust throughput and low latency only provided by the Cisco UCS 6332 Fabric Interconnect. Additionally, both components take advantage of the flexibility provided by the stateless nature of Cisco UCS service profiles and service profile templates.

This design uses the Cisco Nexus 9000 series data center switches in NX-OS standalone mode but provides investment protection to migrate to ACI or higher network bandwidths (1/10/25/40/50/100Gbps) while enabling innovative analytics and visibility using Tetration and automation that support in-box and off-box Python scripting and Open NX-OS that support dev-ops tools (Chef, Puppet, Ansible).

The key design for Scality RING on Cisco UCS C240 is shown in Figure 17.

Figure 17 Topology of Scality RING on Cisco UCS C240 M5L



- Supervisor instance deployed as virtual machine
- Connector node deployed on Storage node
- Storage node deployed on Cisco UCS C240
- Cisco UCS C240 connected to UCS 6332 Fabric Interconnect with 40Gbps line speed
- Cisco UCS 6332 Fabric Interconnect connected to Nexus 9336C-FX2 with 40Gbps line speed

General Hardware Requirements

Table 2 List of Components

Component	Model	Quantity	Comments

Component	Model	Quantity	Comments
Scality RING Storage/Connector Node	Cisco UCS C240	3	Per server node: • 2 x Intel Xeon Silver 4110 • 256 GB Memory • 1 x VIC 1380 • 12 Gbit SAS RAID Controller • Disks ○ 2 x SSD/HDD RAID 1 – Boot ○ 12 x NL-SAS HDD RAID 0 – Data ○ 2 x M.2 SSD JBOD - Metadata
Scality RING Supervisor	Virtual Machine	1	• 4 vCPU • 16 GB Memory • 800 GB Disk • 1 x Network
Cisco UCS Fabric Interconnects	Cisco UCS 6332 Fabric Interconnects	2	
Switches	Cisco Nexus 9336C-FX2	2	

Compute Layer Design

Each Cisco UCS C240 rackmount server is equipped with a Cisco UCS Virtual Interface Card (VIC) supporting dual 40-Gbps fabric connectivity. The Cisco UCS VICs eliminate the need for separate physical interface cards on each server for data and management connectivity. For this solution with Scality RING the VIC is configured with two virtual NICs, one on each physical VIC interface.

Cisco UCS Server Connectivity to Unified Fabric

Cisco UCS servers are typically deployed with a single VIC card for unified network and storage access. The Cisco VIC connects into a redundant unified fabric provided by a pair of Cisco UCS Fabric Interconnects. Fabric Interconnects are an integral part of the Cisco Unified Computing System, providing unified management and connectivity to all attached blades, chassis and rack servers. Fabric Interconnects provide a lossless and deterministic FCoE fabric. For the servers connected to it, the Fabric Interconnects provide LAN, SAN and management connectivity to the rest of the network.

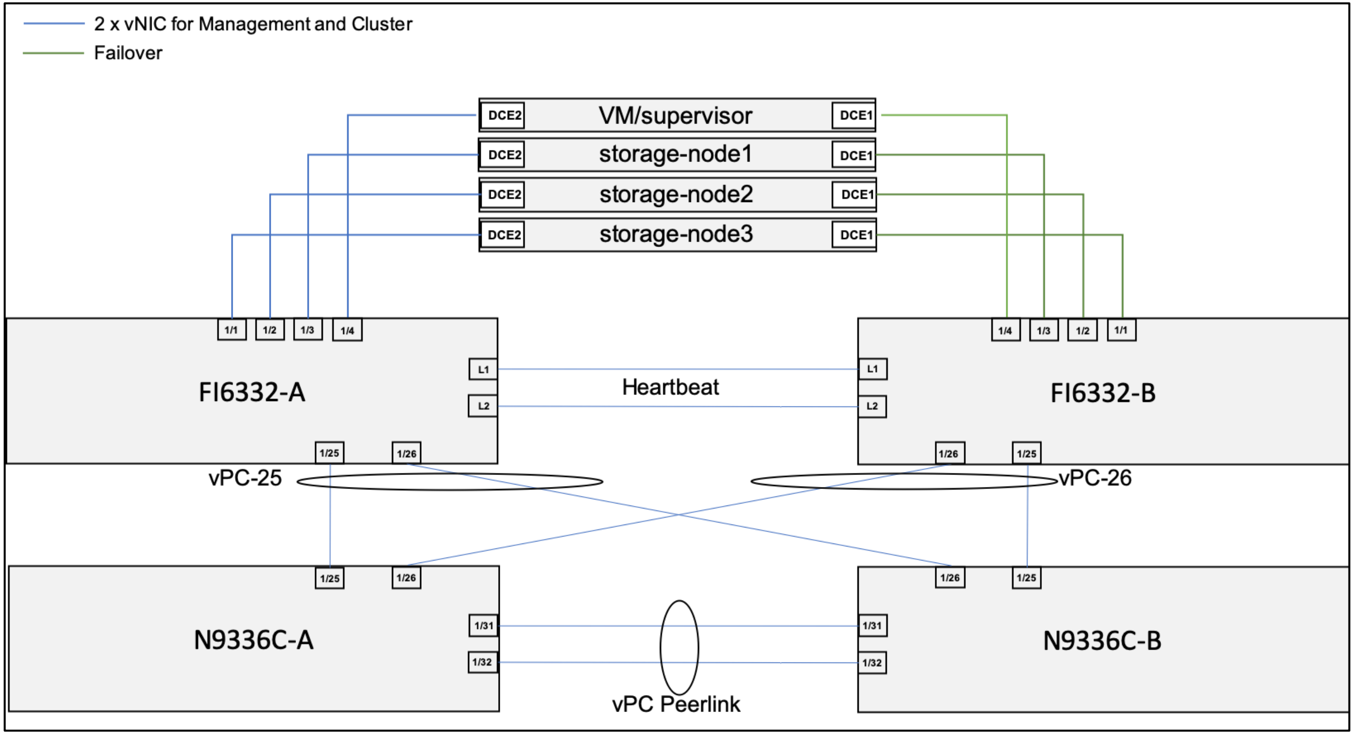
Validated Compute Design

The connectivity of the solution is based on 40 Gbps. All components are connected together via 40 Gbips QSFP cables. Between both Cisco Nexus 9336C-FX2 switches are 2 x 40 Gbit cabling. Each Cisco UCS 6332 Fabric Interconnect is connected via 1 x 40 Gbps to each Cisco UCS Nexus 9336C-FX2 switch. And each Cisco UCS C240 M5L rack server is connected with a single 40 Gbit cable to each Fabric Interconnect.

The exact cabling for the Scality RING solution is illustrated in Figure 18.

The virtual Scality RING Supervisor node is connected to both Fabric Interconnects as well and has access to the Storage/Connector nodes.

Figure 18 Scality RING Cabling Diagram



For a better reading and overview, the exact physical connectivity between the Cisco UCS 6332 Fabric Interconnects and the Cisco UCS C-Class server is listed in Table 3 .

Table 3 Physical Connectivity between FI 6332 and Cisco UCS C240 M5L

Port	Role	FI6332-A	FI6332-B
Eth1/1	Server	storage-node1, DCE2	storage-node1, DCE1
Eth1/2	Server	storage-node2, DCE2	storage-node2, DCE1
Eth1/3	Server	storage-node3, DCE2	storage-node3, DCE1
Eth 1/4	VM	supervisor, DCE2	supervisor, DCE1
Eth1/25	Network	N9336C-A, Eth1/25	N9336C-B, Eth1/25
Eth1/26	Network	N9336C-B, Eth1/26	N9336C-A, Eth1/26

High Availability

The Cisco and Scality solution was designed for maximum availability of the complete infrastructure (compute, network, storage) with no single points of failure.

Compute

- Cisco UCS system provides redundancy at the component and link level and end-to-end path redundancy to the LAN network.
- Cisco UCS C240 rack server is highly redundant with redundant power supplies and fans.

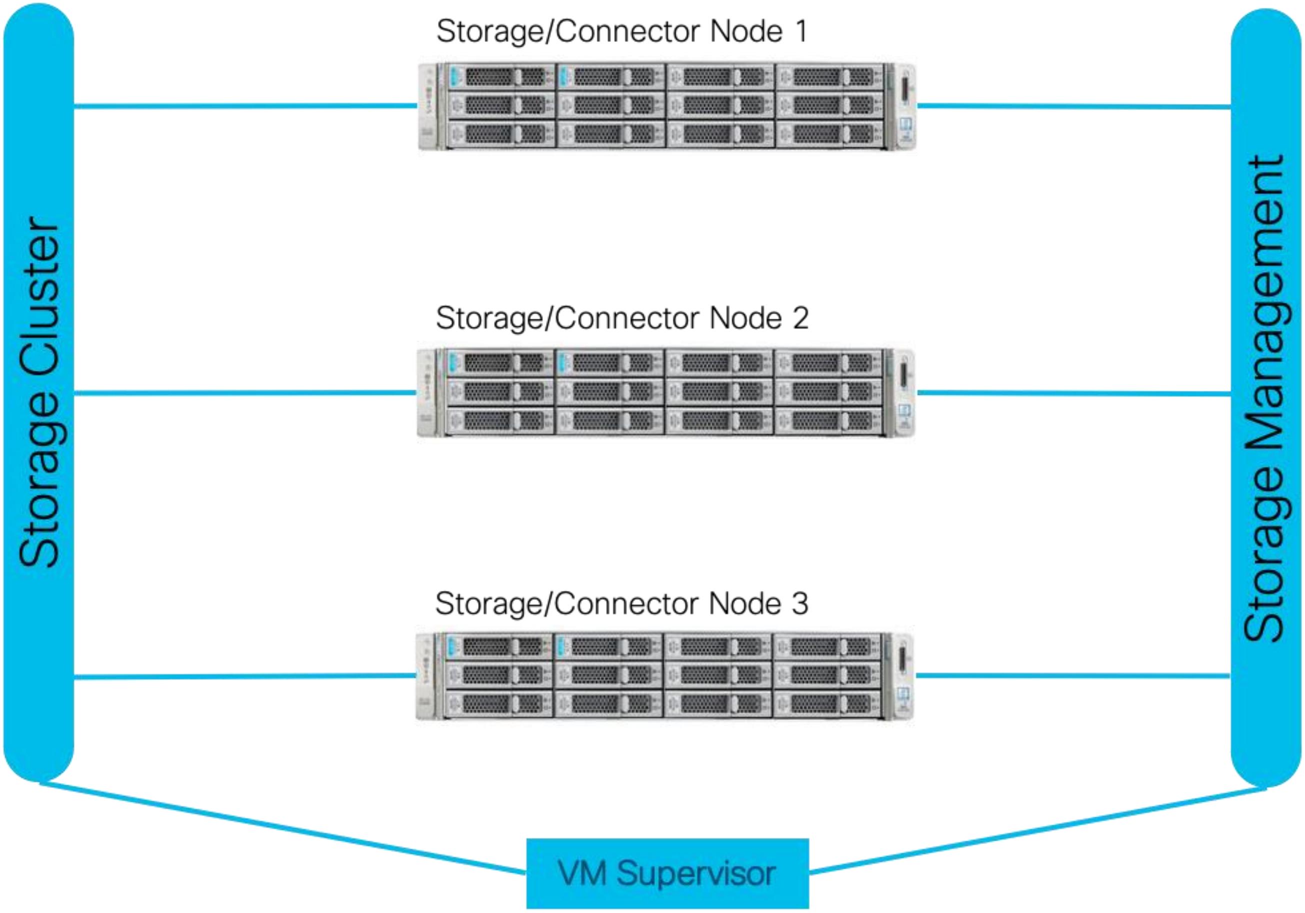
- Each server is deployed using vNICs that provide redundant connectivity to the unified fabric. NIC failover is enabled between Cisco UCS Fabric Interconnects using Cisco UCS Manager. This is done for all Storage/Connector node vNICs.

Network

- Link aggregation using port channels and virtual port channels can be used throughout the design for higher bandwidth and availability, if the optional Cisco UCS Nexus 9336C-FX2 is deployed. Between each Cisco UCS 6332 Fabric Interconnect and both Cisco Nexus 9336C-FX2 is one virtual Port Channel (vPC) configured. vPCs allow links that are physically connected to two different Cisco Nexus 9000 switches to appear to the Fabric Interconnect as coming from a single device and as part of a single port channel.

Figure 19 illustrates the logical configuration of the network for the Scality RING solution.

Figure 19 Logical View of the Network Configuration used in the CVD



QoS and Jumbo Frames

Cisco UCS, Cisco Nexus, and Scality RING nodes in this solution provide QoS policies and features for handling congestion and traffic spikes. The network-based QoS capabilities in these components can alleviate and provide the priority that the different traffic types require.

This design also recommends end-to-end jumbo frames with an MTU of 9000 Bytes across the LAN and Unified Fabric links. Jumbo frames increase the throughput between devices by enabling larger sized frames to be sent and received on the wire while reducing the CPU resources necessary to process them. Jumbo frames were enabled during validation on the LAN network links in the Cisco Nexus switching layer and on the Unified Fabric links.

Software Distributions and Versions

The required software distribution versions are listed below in Table 4 .

Table 4 Software Versions

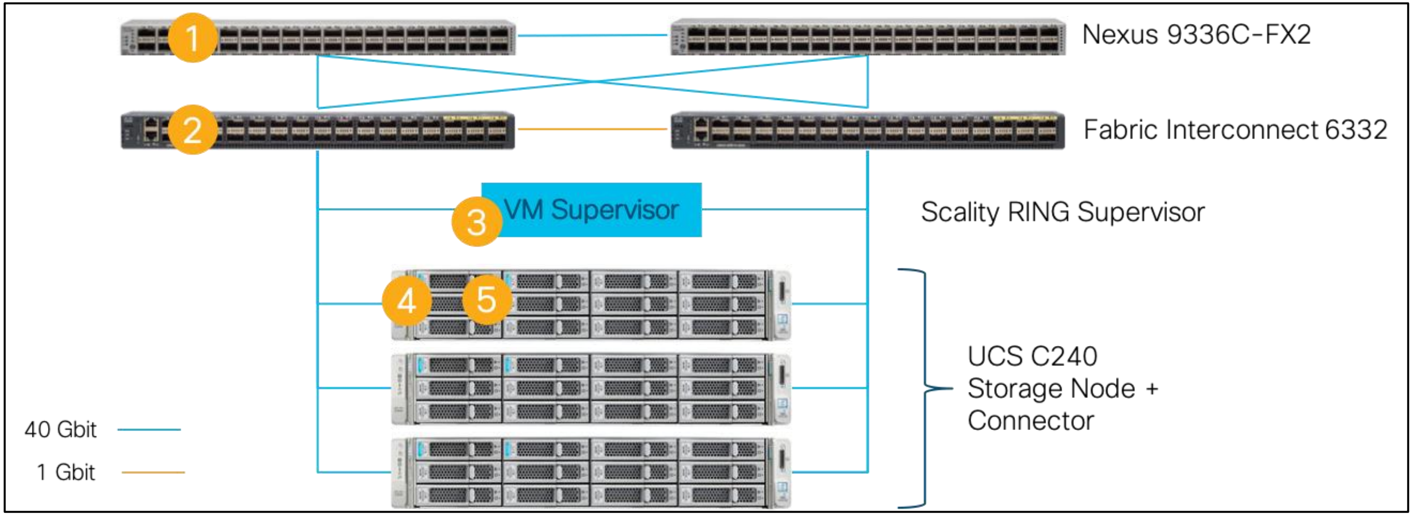
Layer	Component	Version or Release
Cisco UCS C240	Adapter	4.3(2b)
	BIOS	C240M5.4.0.2e
	Board Controller	40.0
	CIMC Controller	4.0(2f)
	Storage Controller SAS 2	50.6.0-1952
Network 6332 Fabric Interconnect	UCS Manager	4.0(2d)
	Kernel	5.0(3)N2(4.02c)
	System	5.0(3)N2(4.02c)
Network Nexus 9336C-FX2	BIOS	05.33
	NXOS	9.2(3)
Operating System	Red Hat Enterprise Linux	7.6
Software	Scality RING	7.4.2.8

High Availability

As part of hardware and software resiliency, the following tests are in the process of being conducted on the test bed. The results of the tests will be included in the Deployment Guide which will be published at a later date.

1. Nexus failure
2. Fabric Interconnect failure
3. Supervisor failure
4. Disk failure
5. Storage/Connector node failure

Figure 20 High Availability Testing



Design Criteria

There are several use cases and target industries where you can use Cisco UCS and Scality's SDS RING solution. The use cases and industries are many, but not limited to the following:

Table 5 Use Cases

Priority	Use Case
Primary	Backup and Archive
	Private and Hybrid Cloud
	Video/Content Distribution (VOD/Origin Server)
	Media Near-line Archive
	Medical Imaging
	Public Cloud – Email
	Public Cloud – Consumer Services
Secondary	Video Surveillance
	Enterprise File Sync and Share
	Hadoop Datalake
	Deep Learning

Table 6 Target Industries

Priority	Industry
Target	Telco, Mobile Operator, and Cable Operator
	SaaS and other Cloud Services
	Financial Services
	Police and Intelligence Agencies
	Hospitals and Medical Imaging Vendor
	Transportation
	Other Global 2000 (non XaaS, FIN, M&E, Transp, Hosp)

The following sections describe some items that you might consider for the design of the infrastructure and the Scality RING.

Requirements

The requirements for the storage have to be understood for the design. These may include the total usable space, future expansion and organic growth for the capacity of the cluster, the performance of the cluster in terms of throughput and bandwidth, the average block size of IO, single site, multi-domain or Multi-site requirements, and so on.

Physical Infrastructure Considerations

Replication versus Erasure Coding

Scality offers the best of both worlds with replication and erasure coding. There is a single standard protection model for a single site platform with 3 nodes. The capacity overhead is based on the distribution of small to large files (above and below 60KB by default). Below 60KB the object is replicated 3 times for a total of 4 copies, the original plus 3. Below 60KB the object is erasure coded using Scality's Advanced Resiliency Configuration (ARC) with 7data parts and 5 coding parts.

Flash Storage

Flash Storage with SAS SSD's or NVME's are used to store metadata for performance. The standard capacity requirement for flash are less than 1 percent of the total data capacity. Standard design also calls for having a ratio of 1 SSD for 16 HDD. When using NVMe, one PCIe card is sufficient for all the drives in the C240 systems.

JBOD versus RAID Disks

While Scality as a SDS solution works with either JBODs or RAID0 disks, it is recommended to use RAID0 for the solution. The 12G SAS RAID controller in C240 can provide up to 4G of cache which can be used for writes. Each disk on the system is setup in one-disk RAID0 array, just so that it can benefit from the RAID Controller cache.

Memory Sizing

Memory sizing is based on the number of objects stored on each storage server, which is related to the average file size and the data protection scheme. Standard designs call for 256 GB for the C240 M5.

Network Considerations

Scality Network requirements are standard Ethernet only; refer to the Network layout in Figure 15. While Scality software can work on a single network interface, it is recommended to carve out different virtual interfaces in Cisco UCS and segregate them. A management network and Cluster network are required for the operation of the RING. Cisco UCS C240 has two physical ports of 40G each and the VIC allows you to carve out many Virtual interfaces on each physical port.

Cisco UCS C240 comes with 2 x 40Gb physical ports where multiple Virtual Interfaces can be configured. It is recommended to have Storage Cluster network on one port and Storage Management network on another port. The Storage Management bandwidth requirements are minimal and hence this port can be shared with other networks like Client/Application network. This will provide 40Gb bandwidth for each of these networks.

Network Uplinks

The Network uplinks from Fabric Interconnects provide upstream connectivity to the Nexus switches. A reboot for instance may be needed during a non-disruptive firmware upgrade. While there is a complete high availability built in the infrastructure, the performance may drop depending on the uplink connections from each FI to the Nexus switches, in the case of such failures or reboots, increase the uplink connections as well.

Multi-Site Deployments

Multisite deployments are available in three reference architectures from Scality:

- Stretched 2 site with Witness
- Stretched 3 sites
- 2 Single Sites asynchronously replicated

Designing a multisite is beyond the scope of this document and, for simplicity only a single site deployment test bed is setup. Please contact Cisco and Scality in case you have a multisite requirements.

Connectors

Connectors both for S3 and SOFS are installed on the storage servers, collocated with the storage processes themselves. As the RING expands for more capacity, the connector layer can also be expanded. Most customer workload requirements are supported by this configuration. Connectors can also be deployed on dedicated servers or as virtual machines, but that architecture is beyond the scope of this document.

If a customer's workload and use case requirements do not fit the assumptions made while building these standard configurations, Cisco and Scality can work together on building a custom hardware sizing to support the customer's workload.

Expansion of the Cluster

Cisco UCS hardware along with Scality RING offers exceptional flexibility in order to scale-out as your requirements change.

Cisco UCS 6332 Fabric Interconnects have 32 ports each. Each server is connected to either of the FI's. Leaving the Network uplinks and any other servers connected to the Cisco UCS Fabric interconnects, 27 Storage server nodes can be easily configured with Cisco UCS FI pairs. If more servers are needed you should plan for a multi-domain system.

Cisco UCS offers server (KVM) management both in-band and out-of-band. If out-of-band management is planned, you may have to increase/reserve as many free OOB IPmanaging the servers. This planning while designing the cluster makes expansion very simple.

Cisco UCS provides IP pool management, MAC pool management along with policies that can be defined once for the UCS domain. Any future expansion for adding nodes etc., is just a matter of expanding the above pools.

Cisco UCS is a template and policy-based infrastructure management tool. All the identity of the servers is stored through Service Profiles that are cloned from templates. Once a template is created a new Service Profile for the additional server can be created and easily applied on the newly added hardware. Cisco UCS makes Infrastructure readiness, extremely simple, for any newly added storage nodes. Rack the nodes, connect the cables and then clone and apply the Service Profile is what needed.

Once the nodes are ready, you may have to follow the node addition procedure per Scality documentation.

The simplified management of the infrastructure with Cisco UCS and the well tested node addition from Scality makes the expansion of the cluster very simple.

Summary

Object storage is an increasingly popular form of distributing data in a scale-out system. The entry size sinks to more and more smaller units. Scality with Scality RING is leading the pack with their technology when it comes to storing data as an object or file with high availability and reliability on small entry-level solutions.

The entry-level solution in this design guide provides customers and partners with everything necessary to store object or file data easily and securely. Cisco's leading technology of centralized management and advanced networking technology helps to easily deploy, manage and operate the Scality RING solution.

The Cisco and Scality solution provides customers new solutions to reliably implement object and file storage.

About the Authors

Oliver Walsdorf, Technical Marketing Engineer for Software Defined Storage, Computer Systems Product Group, Cisco Systems, Inc.

Oliver has more than 20 years of storage experience, working in different roles at different storage vendors, and is now an expert for software-defined storage at Cisco. For the past four years Oliver was focused on developing storage solutions at Cisco. He now works on Scality RING, develops Co-Solutions with Scality for the overall storage market and published several Cisco documents. With his focus on SDS he drives the overall attention in the market for new technologies. In his leisure time, Oliver enjoys hiking with his dog and motorcycling.

William Kettler, Customer Solution Engineer, Scality

William Kettler is a Customer Solution Engineer Partner within Scality's Technical Services group. His current role includes helping customers deploy their petabyte-scale storage solutions, certifying strategic ISVs, and being a technical resource for Scality partners like Cisco.

Acknowledgements

For their support and contribution to the design, validation, and creation of this Cisco Validated Design, we would like to acknowledge the following for the significant contribution and expertise that resulted in developing this document:

- Chris O'Brien, Cisco Systems, Inc.
- Jawwad Memon, Cisco Systems, Inc.
- Maziar Tamadon, Scality