HRG Assessment:
Cisco's Unified Computing System (UCS)

Business and IT professionals are challenged by increasing business environment complexity and the drive to do more with less as IT budgets are increasingly consumed by management, administration, and indirect non-technology costs. Converged systems - an innovation in packaging - bring together network, compute, storage, and system level software in a single easy to manage IT infrastructure solution. Cisco Unified Computing System (UCS) meets customer requirements for a cost effective solution that is easy to buy, install, implement, manage and scale.

Cisco UCS

Cisco's UCS solution includes: Cisco UCS Manager, Cisco UCS Central Software, Cisco B-Series blades, Cisco M-Series servers, Cisco C-Series servers, Cisco Fabric Interconnects, Cisco Fabric Extenders, and Cisco partner ecosystems solutions. Using a modular "fabric centric" approach customers can easily scale corporate IT infrastructure to 1000s of blades managed through a single management interface.

Cisco's UCS architecture is modular, "fabric centric" and built on top of Intel Xeon processors and XML API technologies combined with Cisco Networking expertise and Cisco's 10GB converged fabric. The UCS Manager software and the UCS Fabric Interconnect hardware are sold as an integrated package. Cisco UCS Manager only manages Cisco Blades, Rack Mount Servers, M-Series servers, and UCS certified components. All Cisco Blades and Rack mount servers are Intel based. One Cisco UCS manager domain can scale out to manage up to 160 UCS B-Series Blade servers, or 320 UCS M-Series servers. By year end 2015 HRG expects to see Cisco enable the management of a combination of B-Series, C-Series, and M-Series Cisco UCS servers within a single UCS Manager domain.

Cisco provides an XML API for developers and system level software, tools, and utilities providers to integrate their offerings with Cisco UCS Manager. Microsoft, BMC, and CA use this API to work with Cisco UCS Manager to stand up, provision, and manage UCS Blade and Rack mount servers as well as the Virtual Machines on those servers. Cisco UCS Manager manages Cisco blade servers, rack servers, VICs, and Fabric Interconnects. Since the introduction of Cisco UCS (2009) Cisco and its partners have continually evolved their XML API based integrations to enhance the manageability of the UCS platform.

West Coast Children's Hospital   *Customer Perspective*

The hospital replaced HP with Cisco. They feel that Cisco UCS is a platform that they can grow with. The management aspect (simplicity – single management console, flexibility – the ability to flexibly allocate and reallocate compute resources, and resilience) of UCS was important to the hospital in their selection process. The hospital expects reduced and more manageable system administration workloads with Cisco UCS as compared to HP. Today the hospital runs 2 UCS domains consisting of 22 B-Series Blade servers (B200 M3 with 256 GB RAM) each. They are in the process of expanding the total UCS footprint to 60 B-Series servers.

Cisco UCS is being used to host the Virtual Desktop Infrastructure consisting of VMware for virtualization and Citrix for remote application presentation. Remote Virtualized Desktops are used in a clinical care setting by nurses and other care providers in the hospital to deliver actionable information and guidance at the point of care. The VDI clinical application is being used to share and capture Electronic Medical Record (EMR) data for individual patients. Nurses and other clinicians use mobile wireless tablets and wireless cart mounted compute devices to capture and access clinical care information on individual patients in order to track treatment and share information with the attending physician. HRG expects the hospital to realize measurable improvements in terms of the quality of care and the cost of care.

Each nurse or clinician has their own unique Citrix instance which is hosted on the Cisco UCS Infrastructure. The hospital currently runs two nursing shifts with between 700 and 800 concurrent users for each shift resulting in as many concurrent wireless connections from mobile in hospital clinical care compute devices which are all served up by the UCS hosted Virtual Desktop Infrastructure.

In addition to the VDI workload the hospital is supporting a Phillips medical telemetry application on the same UCS installation. This telemetry application will ideally support both in hospital wireless patient monitoring and data collection as well as some remote off campus monitoring in support of in home care for some chronic and other less critical care patients. It is HRG's expectation that this telemetric data will be integrated with individual EMR information to enhance patient care and streamline clinical care pathways so as to eliminate unnecessary procedures and reduce the cost of treatment.

The hospital plans to integrate genomic and other relevant data in order to eventually be used in ongoing in hospital genomic research and to deliver individual personal and integrated medical information to clinicians, and other care givers. Currently these workloads are running on a Cisco UCS test bed system and there is a plan to move forward with a fuller implementation once the proof of concept trail has successfully concluded. This workload will ultimately consist of personal clinical, genomic, and patient personal history data integration and the presentation of summarized and distilled data via data analytics. In HRG's opinion this type of workload is a good fit for Cisco's M-Series Servers. Cisco M-Series Servers are purpose built to host single application deployment across multiple servers for performance and resiliency.

Global Financial Services Firm   *Customer Perspective*

The Global Technology Infrastructure division of the bank operates as a utility (a generic shared pool of compute resources) within the bank and the head of that group is working toward delivering IT as a service on a global basis across all parts of the bank. The strategy is to create a single pool of compute resources that is then offered as a service to everybody in every line of business. 70% of all new workloads will be hosted on this this shared Infrastructure as a Service (IaaS) utility. The remaining 30% of the banks workloads are unique, requiring a special configuration for a specific project, due either to the nature of the project or the specific workload requirements. The bank is working to standardize around a defined set of SLAs and in this way avoid as much as possible, unique requirements. There is a "my compute portal" which is an end user portal where the user can self-subscribe to use a pre-established configuration to run their workloads.

The bank has set UCS up to run either as Physical Server Instances (PSI) or Virtual Server Instances (VSI). In the case of the PSI set up each workload is dedicated to its own hardware but collectively the PSIs are all part of a managed domain allowing the bank to take advantage of shared I/O and shared standardization. VSI - a Pod of servers divided into virtual machines used to run various different workloads in a dynamic fashion such that utilization of those compute resources is 60 to 70% as compared to 25 to 30% utilization for the PSI instances.

Currently the bank has Multiple UCS Management domains spanning 110 pods consisting of between 12 to 14 thousand Cisco B-Series servers. Each domain is comprised of 2 Fabric Interconnects and 2 or 3 racks of UCS servers. The banks IT architecture is SAN oriented (storage centric) and the UCS Blade solution is a good fit. The bank currently does not use UCS central although they see the benefit of using it. With regard to Cisco M-Series servers there is a good probability that the bank will include these systems as part of their IaaS utility in the future.

The bank believes that moving to an integrated converged fabric will result in a significant amount of port reduction, cable reduction, and adapter reduction. This will further result in simplification, less expense, easier management, reduced power, reduced cooling and a reduction in support requirements. The bank estimates that it will realize up to a 30% reduction in power and cooling. With UCS and UCS manager the time to market (time for server build and provisioning) has been significantly reduced and time to productive first use of a new service or application has been reduced from 6 months down to 45 days or less.

Federal Agency   *Customer Perspective* (some specifics redacted at the agency's request)

Cisco UCS is being used to support the knowledge workers at this federal agency. The IT team manages the virtual desktop infrastructure (VDI). They serve up the VDI using VMware Horizon View for the client side users. The users have access to a number of databases through a web interface.

They are currently working on a continuity of operations project and are using VMware Sight Recovery Manager (SRM) for disaster recovery and fail over. They are in the process of setting up a remotely located environment for disaster recovery. The remote facility will have a separate NetApp SAN and a separate UCS domain they will use VMware for SAN to SAN data replication. The plan is to have teleworkers use the remote site so that it is kept active and up to date. If there is an outage the remote site can support roughly 75% of the users.

They are upgrading their legacy voice and IM application to Cisco Jabber. Supporting Jabber there will be two components a main office voice component and a remote Cisco UCS site where they will host the capability for presence, IM, cell phone, and video as necessary so they can escalate conversations as needed from IMs up to video conferencing.

With Cisco UCS the customer can do firmware upgrades live on production systems with no disruption of service and no end user impact by failing over from the primary FI, upgrading its firmware, failing over from the secondary back to the primary and then upgrading the firmware on the secondary. They have done a number of firm ware upgrades and had no problems. They rely on VMware to provide high availability functionality for their workloads.

Prior to installing Cisco UCS they were using HP ProLiant servers. They moved to Cisco UCS because of the FlexPod architecture and the UCS ability to boot from SAN. They can replace any failed components with no disruption of service thanks to Cisco's recommended dual FI architecture. Service profiles provide flexibility in terms of workload specific configurations. Cisco's use of stateless hardware and service profiles really provides good benefit.

They first installed Cisco UCS in 2012 and currently have 2 Cisco UCS Manager Domains set up - one UCS domain comprised of B200 M3 blades is for classified workloads and the second domain is for unclassified workloads. The maximum number of concurrent users to date has been around 700. In their building they have roughly 850 people at any given time. They are running at roughly 50% capacity utilization. They are planning for an increase in the number of users from current levels to approximately 1200. The recently upgraded their UCS B200 M3 Series server cores and maxed out the memory on each.

Basic Cisco UCS Architecture

The Cisco UCS 5108 Blade Server Chassis is six rack units (6RU) high and can mount in any industry-standard 42 U rack. Each 42 U rack can fit 2 Fabric Interconnect (FI) switches and six Cisco UCS Blade chassis. Each Fabric Interconnect can handle up to 20 chassis per FI per – or 20 pairs of Fabric Extenders (FEX) mixing blade and rack servers. A Cisco 5108 chassis accommodates 2 FEX or Fabric Extenders with up to eight half-width Cisco UCS B-Series Blade Servers or a mix of larger form factor blade servers. Cisco servers are only available with Intel processors and UCS Manager only manages Cisco UCS servers.

Each of the 2 FEX in a Cisco UCS Blade chassis is connected to the south side server ports on the Fabric Interconnect. Two Fabric Extenders (FEX) are required for each UCS 5108 chassis for availability and fail over purposes. The FEX are connected to the south side ports on one of the Fabric Interconnects. Each FEX has either 4 or 8 north bound 10 GB uplink ports depending on the model. HRG expects to see Cisco move to 40 GB uplinks for all of their next gen UCS solutions some time later this year (2015) based on the observation that the new M-Series systems all have 40 GB uplinks as a standard.

Each Fabric Extender connects to a different top of rack Fabric Interconnect (FI) to ensure that there are redundant 10 GB Cisco converged fabric paths in the event that either one of the FEX or one of the Fabric Interconnect switches fails. Cisco UCS provides multiprotocol support for FCoE and iSCSI on the Cisco 10GB converged fabric. Under normal conditions UCS manager runs primarily on one FI, with the other FI mirroring all the activity via a heartbeat connection. If the primary FI goes down, UCS Manager takes over from the secondary FI with no major disruption of service. It is only when one of the interconnects is taken out of service for maintenance or goes down that all the traffic in the system flows through one Fabric Interconnect creating a temporary potential for increased latency at the switch.

The Cisco UCS Blade System UCS 5108 chassis based solution requires 2 identical Top of Rack Fabric Interconnect switches in order to provide redundancy and availability at the Layer 2 Fabric Interconnect Switch level. The same number of uplink ports on each FEX must be cable connected to the south side ports on each of the Fabric Interconnects such that FEX 1 will connect only to FI 1 and FEX 2 will only connect to FI 2 in order to preserve system availability and fail over capability throughout the Fabric path.

For the UCS Blade system to deliver its maximum throughput, all uplink ports on each of the Fabric Extenders (FEX) in the UCS 5108 chassis must be connected to one or the other of the top of rack Fabric Interconnect switches. When considering an IT infrastructure that is heterogeneous, and based on open standards, customers need to understand ahead of time how a converged system like Cisco UCS will be integrated and managed as part of that environment. Many cluster oriented workloads can be run effectively on UCS.

A single UCS domain can span several racks (160 B-Series Servers, 320 M-Series Servers, or B-Series, C-Series, and M-Series Cisco UCS servers)all connected to one pair of FI's. Within a single Cisco UCS domain it takes a predictable amount of time (latency) for each server to be accessed through the Fabric Interconnect. In those cases where customers are running latency sensitive HPC type workloads that require extremely low levels of messaging latency HRG recommends real-world benchmarking of customer workloads on a Cisco UCS system to ensure that the system will perform as expected.

Cisco UCS Manager

Cisco UCS Manager is a common platform for management of Cisco UCS B-Series, C-Series, and M-Series servers. Cisco UCS Manager's embedded device management software manages the software and hardware components of the Cisco Unified Computing System ™ across multiple chassis and virtual machines through a Java based GUI, a CLI (command-line interface), an XML (Extensible Markup Language) Application Programming Interface (API), or HTML5 on the UCS Mini platform. By year end 2015 HRG expects to see UCS Manger transition to HTML5 for all platforms thereby removing any Java dependency. Service Profiles in the UCS Manager application can be used to set up and configure stateless Intel Xeon based Cisco blade and rack mount servers and virtual machines. UCS Manager by itself does not configure VM's. However, when UCS Manager is integrated using the available XML API with either VMware, Microsoft, or Openstack, VM configuration can be accomplished from within those platforms. VMs can be deployed from within UCS Director when deploying Service Profile settings. These Service Profiles can be 'moved' with a virtual machine when it is moved using VMware's VMotion as in the case of a server failover or when reallocating capacity to satisfy changing workload requirements on the fly.

Cisco UCS Central Software

Cisco UCS Central software, a virtual machine image, manages numerous Cisco UCS Manager domains, can be run on a hypervisor and can manage up to 6,000 Cisco UCS servers from a single unified UCS Central console. Cisco UCS Central integrates and consolidates multiple Cisco UCS Manager domains to provide global configuration capabilities for resource pools, policies, and firmware thereby simplifying policy compliance across distributed data centers giving administrators a "define once, deploy many times" workflow capability for the entire infrastructure.

Cisco UCS Central brings inventory and fault information across multiple Cisco UCS domains together in one place facilitating service assurance for the entire infrastructure. UCS Central exposes a rich XML API, just like Cisco UCS Manager, for the integration of customer and partner system management software solutions.

Service Profiles

Cisco UCS Manager Service Profiles are created by server, network, and storage administrators and stored on the UCS Fabric Interconnect in an object based data store. Cisco UCS Manager discovers UCS devices that are added, moved, or removed from the UCS system. This information, added to the UCS Manager's inventory (a light weight CMDB), is saved on the Fabric Interconnects. Cisco reports that a number of their customers are using the XML API to connect UCS Manager to their Configuration Management Data Base (CMDB); BMC provides an enterprise class CMDB, in order to facilitate change tracking and management. UCS Manager uses this information when deploying Service Profiles to newly discovered resources. When a Service Profile is deployed UCS Manager configures the server, VICs, vNICs, and vHBAs. Service Profiles can also be used to enable Virtual Network Link (VN-Link) capabilities for VN-Link supported hypervisors.

SingleConnect Technology

SingleConnect Technology, an umbrella term, is Cisco's name for VN-Link and all VM FEX related functionality. Cisco uses the term SingleConnect Technology to describe various ways of providing physical network type visibility to non-physical or non-directly attached devices. For example, virtual machines, virtual interfaces on a remote interface card, or physical interfaces on a non-switching remote device. SingleConnect benefits include Unified Fabric / Ethernet, multiprotocol storage (FCoE, iSCSI) and improved support for mixed rack and blade form factors.

Partner Focused

Cisco UCS Manager when used in conjunction with partner provided software solutions can stand up, provision, and manage UCS Blade and Rack mount servers as well as the Virtual Machines on those servers. Cisco UCS supports the VMware ESX, ESXi, Microsoft Hyper-V, KVM, OVM, and Xen hypervisors. Cisco does not sell system level management and monitoring software instead Microsoft, BMC, EMC, CA, IBM and others provide this functionality. Cisco UCS Manager by itself does not do bare metal operating system installations or application software installations. Cisco UCS provides this functionality through close integration with additional stand-alone software like Symantec, BMC BladeLogic, and Cisco UCS Director.

Storage Connectivity

Cisco UCS Manager preserves traditional Ethernet and Fibre Channel connectivity to LAN and SAN environments north of the Fabric Interconnect. Cisco relies on other vendors like Nimble Storage, NetApp, EMC, and IBM Storwize for storage and storage management. Cisco provides highly customizable XML APIs that developers and system level software, tools, and utilities providers use to integrate their offerings with Cisco UCS Manager. Storage tools can export the definitions of pools of storage resources through these APIs to the UCS Manager for use in allocating storage resources to UCS servers using the UCS Service Profiling capability.

Native Fibre Channel (FC) connectivity is available on the north or upstream side of the Fabric Interconnect. Cisco's current generation FI comes with unified port technology so that individual upstream ports can be configured as either native FC or Ethernet, providing customers with enhanced flexibility in meeting changing workload requirements. However, there is no native Fibre Channel connectivity available within the Cisco UCS 5108. Customers currently running native Fibre Channel for SAN connectivity from individual rack mount or blade servers can do the same connection from the Fabric Interconnect to their SAN environment.
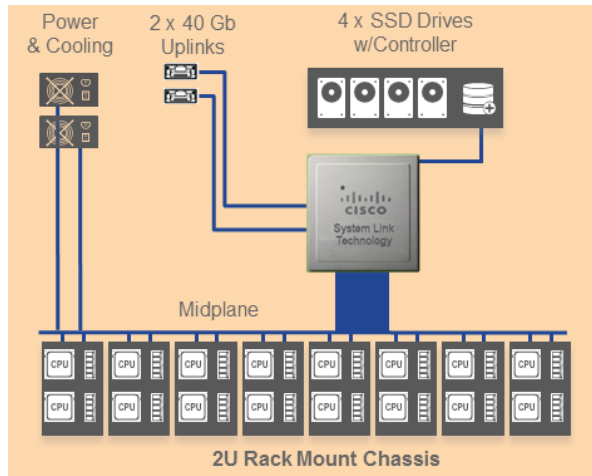
UCS Mini

Cisco UCS Mini is a good choice for deployments of 2 to 15 servers (8 Cisco UCS B200 M3/M4 Blade Servers and up to 7 direct-connect rack servers) for branch-office, remote-office, point-of-sale, and smaller IT environments to handle IT and web infrastructure from distributed database to ERP, and CRM applications. UCS Mini delivers servers, storage, and 10 Gigabit networking. Expandable to 15 servers, Cisco UCS Mini provides a scalable solution. Cisco UCS Central Software provides multisite remote management ensuring server consistency across sites. Dual-voltage AC (100-120V and 200-240V) DC power supplies make Cisco UCS Mini ready for global deployment. The Cisco UCS 6324 Fabric Interconnect, mounted inside the US Mini 5108 chassis, provides unified server and networking capabilities.

UCS M-Series Composite Machine Architecture

Power & Cooling
2 x 40 Gb Uplinks
4 x SSD Drives w/Controller
System Link Technology
Midplane
2U Rack Mount Chassis

Today Converged and Hyper Converged Systems deliver more compute power in a smaller energy efficient footprint than previous generation clusters and stand-alone servers. A recent addition to the UCS family, Cisco's M-Series Server is based on Cisco's System Link Technology that enables shared Local Disk and I/O and allows the disaggregation of 16 processors and their associated RAM to deliver up to 16 fully independent servers per each 2 U Rack Mount M-Series Chassis. Cisco UCS M-Series addresses next generation workload requirements eliminates unnecessary infrastructure, reduces cable count, and enables unified management across severs.

The Cisco UCS M-Series application-specific integrated circuit (ASIC) and Cisco System Link Technology combined present discrete flexibly configurable resources via the Gen 3 PCIe mid-plane to each server in the chassis. Virtual network interface cards (vNICs) and virtual host bus adapters (vHBAs) are seen by the Cisco UCS Manager software as PCIe end devices and the OS communicates as if they are physical PCIe devices. In the M-Series Cisco decoupled the server management controller so that individually managed cards are not needed. As shown in the diagram each M-Series server consisting of CPU and RAM is connected by the PCIe Gen 3 chassis mid-plane to the common resources of power, management, cooling, storage, and networking. A RAID controller in the chassis provides advanced RAID functionality and caching. System Link Technology provides virtualized shared local storage and the storage controller provides a SCSI NIC PCIe device (sNIC) and presents it to OS that sees these sNICs as locally attached SCSI devices. The drives can be virtualized and distributed across the M-Series servers as needed.

Cisco UCS Manager manages the Cisco UCS M-Series servers and chassis providing policy-based management thru an XML API and a growing ecosystem of strategic partnerships for the configuration and management of 20 M-Series chassis (320 M-Series servers) in a single Cisco UCS Manager domain.

Conclusion and Recommendation

It is HRG's opinion that Cisco UCS is a good fit for many general business workloads. In those cases involving mission critical workloads where reduced transactional latency is a requirement HRG recommends real world benchmarking of customer workloads on a Cisco UCS system prior to purchase. Cisco's Intel centric approach to the Blade market is a simplified, easy to understand, and easy to scale solution.

Cisco customers consistently report on the IT and overall business benefits they are seeing. These benefits include: reduced time to market, more manageable system admin workloads, ease of management, and more (see the *Customer Perspective* sections earlier in this paper). With the introduction of Cisco UCS Mini and Cisco M-Series solutions HRG expects to hear more from customers regarding the benefits and cost savings they are experiencing.

Harvard Research Group is an information technology market research and consulting company. The company provides highly focused market research and consulting services to vendors and users of computer hardware, software, and services. For more information please contact Harvard Research Group:

Harvard Research Group
Harvard, MA 01451 USA

Tel. (978) 456-3939

e-mail: hrg@hrgresearch.com

http://www.hrgresearch.com