Understanding Cisco Unified Computing System Service Profiles

What You Will Learn

Cisco Unified Computing System™ technology has redefined the enterprise computing environment. By breaking the traditional model of older data centers and redefining data center infrastructure as pools of virtualized server, storage, and network resources, the Cisco Unified Computing System has enabled a new computing model that both delivers advantages in capital and operational cost, and does so in a manner that improves flexibility, availability, and reduces the amount of time needed for IT to respond to business changes.

To understand how the Cisco Unified Computing System delivers these benefits, it is necessary to understand the concept of a “service profile,” the fundamental mechanism by which the Cisco Unified Computing System models the necessary abstractions of server, storage, and networking. This document explores the underlying structure of Cisco Unified Computing System service profiles, explains how they are used to enable Cisco Unified Computing System capabilities, and shows the benefits that accrue from their use.

Redefining the Server and the Data Center

The Old View: Separate Stacks, Lots of Glue

Today’s data centers are migrating from the client-server, distributed model of the past toward the more virtualized model of the future. This steady migration is fueled by the need to conserve space and energy[1], as well as a desire to overcome the myriad problems that arise from supporting a heterogeneous data center environment:

- **Complexity:** Too many software and hardware vendors in the data center leads to too many conflicting components. These incompatibilities hinder new product deployments and upgrades.
- **Archaic Server Architectures:** While server performance has increased dramatically, the basic architecture of the server has not changed in decades. Each server is essentially its own management island, with local management and settings, some of which, such as MAC addresses, are assumed to be immutable. Multiple generations of management tools have attempted to provide a single management interface to multiple servers to simplify the operating experience for users, but these fundamentally are still layers of tools stacked on top of disjointed management domains. Blade servers as a product class have improved this situation, but they still lack true management integration beyond the chassis level.
- **Storage:** As business applications increase in complexity, the need for larger and more reliable storage solutions becomes a data center imperative. Storage was the first significant server resource to be shared, in the form of Network-Attached Storage (NAS) and SANs, and today networked, shared storage is the norm in most organizations.
- **Disjointed Network, Server, and Management Domains:** Even more crippling than older server architectures from the perspective of streamlining data center operations has been the evolution of server, storage, and networking solutions into separate technology and management silos.
- **Disaster Recovery and Business Continuity:** Data centers must maintain business processes (with limited or no downtime) for the overall business to remain competitive.
- **Inefficiency:** Application slowdowns result from server and application incompatibilities, inconsistent server policies, and poor integration of third-party hardware and software.

[1] Cisco estimates that over the next 3 years, 50 percent of large organizations will face an annual energy bill that is higher than their annual server budget.

- **Cooling:** Good ventilation saves power and reduces costs. However, dense server racks make it very difficult to keep data centers cool and hold down costs.
- **Cabling:** Many of today's data centers have evolved into a complex mass of interconnected cables, further increasing rack density and further reducing data center ventilation.

Virtual Servers: The First Abstraction

The first major breakthrough was the development of enterprise-class hypervisors for x86 systems and the common Linux and Microsoft Windows operating systems[2]. These hypervisors and their associated virtual servers delivered the first building block of a virtualized data center: a server abstraction that was not bound to a physical hardware element. With virtual machines, users could now define a server in software and deploy it anywhere in the data center without altering any of the underlying hardware.

To understand virtual machines - and service profiles - consider how a virtual machine abstracts a server definition.

Instead of thinking of a server as a box in a rack, consider that a server is defined by the combination of the following sets of physical resources:

- **Processing Resources:** CPU and memory
- **Storage Bindings:** Connections to external storage, which may contain the boot images, applications, etc. Storage may also include the local disk, but as will be discussed later, this is a less effective choice except in environments in which the boot image is almost invariant.
- **Network Interfaces:** Available network connections and Network Interface Cards (NICs), Host Bus Adapters (HBAs), Converged Network Adapters (CNAs), etc.

These elements provide a model of a physical server before it is deployed: with resource entitlements and connections to the rest of the data center, but without any personalization such as server name, MAC address, boot parameters, Fibre Channel Worldwide Name (WWN) settings for the HBAs, etc. that change the definition from a generic server to a specific server. This separation of definition and instantiation, and the capability to create generic virtual machine templates, is one of the great strengths of all virtualization environments. Generalized templates can be created, and then new servers can be created by applying a set of Unique IDs (UIDs) such as MAC addresses and server names to the template as the server is created in the runtime environment, enabling one-to-many deployment and drastically reducing the number of touchpoints needed to create and manage the server population.

Data Center Integration: The Missing Element

While virtual server hosts and their associated virtual machines were a major breakthrough and extremely valuable in their own right, their success highlighted the fact that the data center still remained a world of islands, with separate network, storage, and server technology and administration stacks. Early attempts to integrate these separate operational spheres were coarse-grained attempts to join dissimilar management consoles together under a set of menus with a common look and feel. Tools were usually from different code bases and often from multiple original manufacturers, and the differences were readily apparent to most users. Little real integration was achieved, and despite more than a decade of progress, promises, and products, integrated management remains an elusive goal.

[2] IBM actually developed the first commercial hypervisors more than 30 years ago on its mainframe products, but it took more than two decades for commercial hypervisors to emerge on general-purpose systems such as x86 servers.

Despite this relative lack of success, the vendor community as a whole has not ignored this combined problem and opportunity, and users today have at least 200 products to choose from that purport to offer benefits in managing some aspects of the combined storage, server, and networking environments. Some of them are very capable within their domains, but none are truly integrated, with an integrated data abstraction and a common code base of management software that spans the entire server, storage, and network fabric. Further, integration of these tools often remains an exercise left to the user. Additionally, true hardware independence is difficult to achieve, and collections of loosely coupled tools are subject to errors such as race conditions when data center operators attempt to perform large numbers of asynchronous operations. However, the technology exists today to merge servers, networks, and storage into a single integrated environment with an integrated management platform - but it requires a new converged fabric architecture that removes the old separations between these domains.

To meet this challenge, Cisco introduced the Cisco Unified Computing System, the first commercially available system that delivers an integrated management model that spans server, storage, and networking as well as hardware designed to take advantage of its innovations. The following sections provide an overview of the basic functions of the Cisco Unified Computing System along with the details of the Cisco Unified Computing System service profiles, the underlying abstraction model that is central to Cisco Unified Computing System.

Cisco Unified Computing System Overview

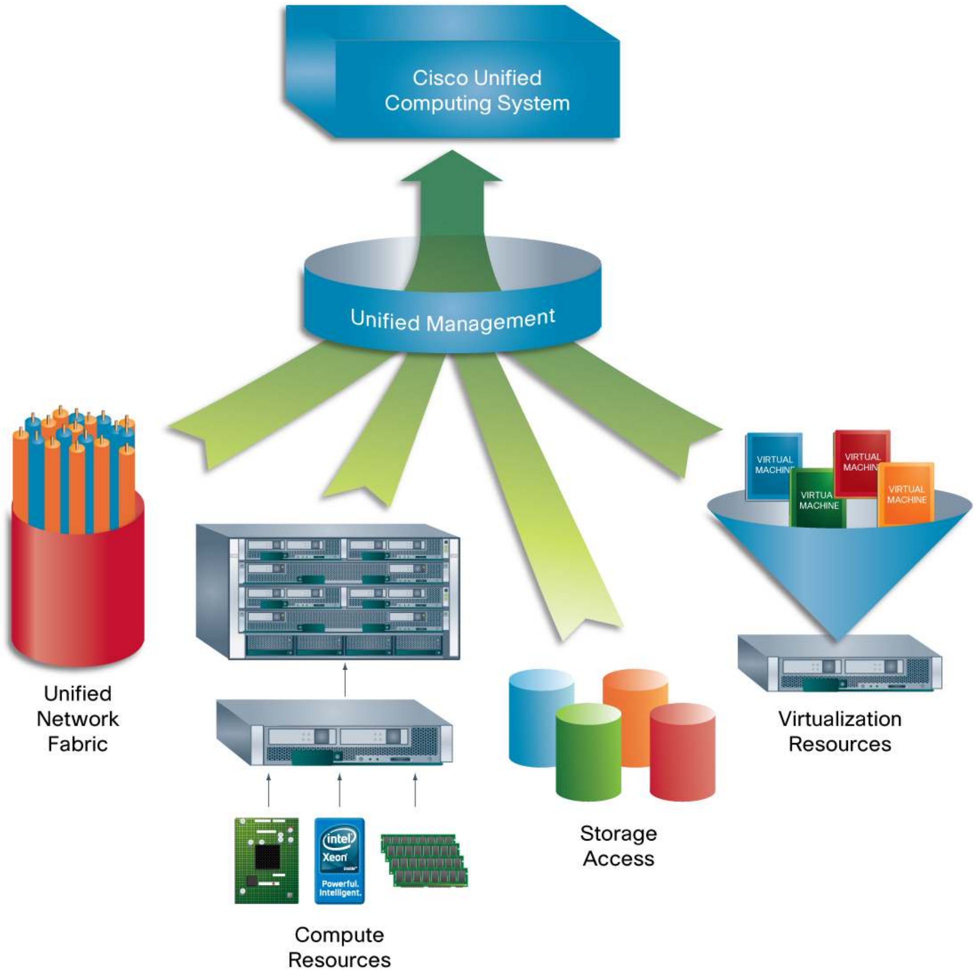
The Cisco Unified Computing System integrates low-latency, unified network fabric with enterprise-class, x86-based servers, creating an integrated, scalable, multichassis platform in which all resources participate in a unified management domain. A single system can scale across multiple chassis and thousands of virtual machines[3] (Figure 1).

The approach of the Cisco Unified Computing System enables data center servers to become stateless and fungible, where the server's identity (using MAC or WWN addressing or UIDs) as well as build and operational policy information such as firmware and BIOS revisions and network and storage connectivity profiles can be dynamically provisioned or migrated to any physical server in the system. The Cisco Unified Computing System integrates server management with network and storage resources to meet the rapidly changing needs in today's data centers. New computing resources can be deployed "just in time." Traditional physical and virtual workloads can be easily migrated between servers through remote management, regardless of physical connectivity. The Cisco Unified Computing System directly improves capital utilization and operational cost and enables gains in availability, security, agility, and performance through an integrated architecture.

The Cisco Unified Computing System currently supports older dedicated OS environments (Microsoft Windows and multiple versions of Linux) as well as both VMware and Microsoft Hyper-V hypervisors and maintains a high level of integration with their respective vendors' management stacks. It is also designed to readily integrate with third-party software and user management tools. In addition to its own Command-Line Interface (CLI) and internal XML interfaces, the Cisco Unified Computing System supports a wide variety of interfaces to external systems, including Common Information Model (CIM) XML; Systems Management Architecture for Server Hardware (SMASH) Command Line Protocol (CLP); Simple Network Management Protocol (SNMP); Intelligent Platform Management Interface (IPMI); remote Keyboard, Video, and Mouse (KVM); and Serial over LAN (SoL).

[3] Note that the Cisco Unified Computing System does not attempt to manage the deployment or lifecycle of the virtual machines that are deployed on the physical Cisco Unified Computing System server. Cisco customers were almost unanimous in their position that they wanted to use their current virtual machine management stack within the Cisco Unified Computing System infrastructure. As result, Cisco determined that virtualization of the server fabric is the purview of the Cisco Unified Computing System, not virtual machine management. Where the domains intersect, as in the implementation of the Cisco Nexus® 1000V Series Switches or the Cisco UCS M81KR/P81E Virtual Interface Cards, Cisco is fully integrated with VMware management.

Figure 1. The Cisco Unified Computing System Integrates Network, Compute, Storage Access and Virtualization Resources into a Single Cohesive System



Service Profiles: Cisco Unified Computing System Foundation Technology

What Is a Service Profile?

Conceptually, a service profile is an extension of the virtual machine abstraction applied to physical servers. The definition has been expanded to include elements of the environment that span the entire data center, encapsulating the server identity (LAN and SAN addressing, I/O configurations, firmware versions, boot order, network VLAN, physical port, and quality-of-service [QoS] policies) in logical "service profiles" that can be dynamically created and associated with any physical server in the system within minutes rather than hours or days. The association of service profiles with physical servers is performed as a simple, single operation. It enables migration of identities between servers in the environment without requiring any physical configuration changes and facilitates rapid bare-metal provisioning of replacements for failed servers.

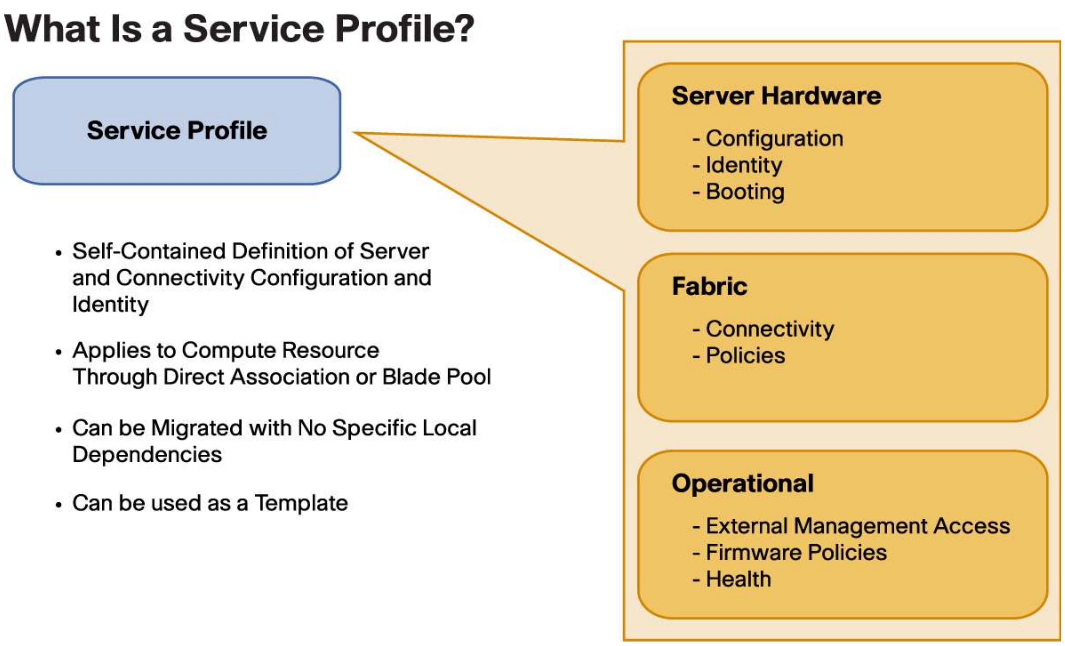
Service profiles also include operational policy information, such as information about firmware versions.

This highly dynamic environment can be adapted to meet rapidly changing needs in today's data centers with just-in-time deployment of new computing resources and reliable movement of traditional and virtual workloads. Data center administrators can now focus on addressing business policies and data access on the basis of application and service requirements, rather than physical server connectivity and configurations.

Service profiles can be abstracted from the specifics of a given server to create a service profile template, which defines policies that can be applied any number of times to provision any number of servers. Service profile templates help enable large-scale operations in which many servers are provisioned as easily as a single server.

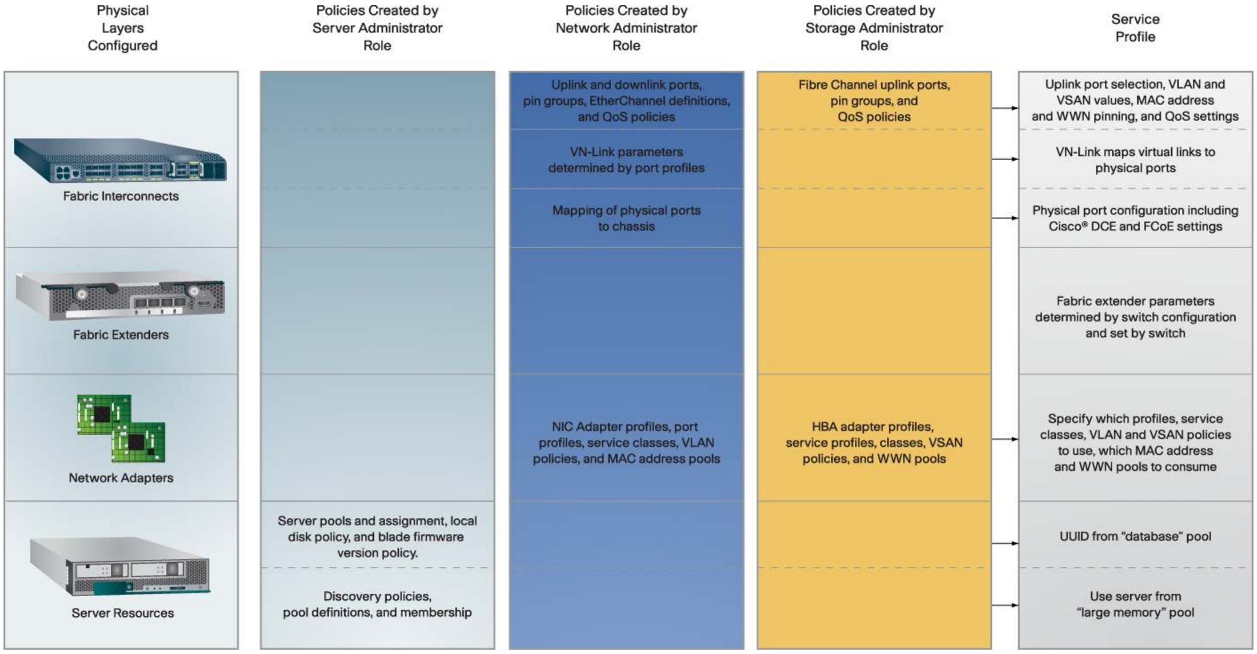
In addition, using service profiles, Cisco® UCS Manager provides logical grouping capabilities for both physical servers and service profiles and their associated templates. This pooling or grouping, combined with fine-grained role-based access, allows businesses to treat a farm of compute blades as a flexible resource pool that can be reallocated in real time to meet their changing needs, while maintaining any organizational overlay on the environment that they want. Figure 2 shows the major elements of a service profile.

Figure 2. The Cisco Unified Computing System Service Profile Incorporates a Complete Metadata Description of the Information Required to Provision a Server in a Networked Data Center, Including Storage, Network, and Operational Policies



Another way to understand the concept of service profiles is by looking at their configuration points: the aspects of the Cisco Unified Computing System that they control. Figure 3 shows the most significant configuration points for Cisco Unified Computing System service profiles.

Figure 3. The Cisco Unified Computing System Service Profile Includes All the Settings Needed by Server, Storage, and Network Administrators to Define and Provision a Server



The service profile components define the server environment, including the local server settings plus storage (SAN) settings such as VSAN specifications and complete network settings such as uplink, VLAN and QoS settings.

Service profiles themselves are built on policies: administrator-defined sets of rules and operating characteristics. These policies are low-level objects in the Cisco Unified Computing System management model, and after they have been defined, they can be reused multiple times to build service profiles. Examples of policies include:

- **SoL:** Defines the behavior of a virtualized serial line over the LAN
- **Firmware:** Enables the specification of firmware updates, including a backup firmware version as well as a current version; one of the most powerful of the Cisco Unified Computing System abstractions. (To learn more, see Achieve Automated, End-to-End Firmware Management with Cisco UCS Manager)
- **IPMI:** Defines the IPMI behavior of the system
- **Stats:** Controls the way system data is collected
- **BIOS:** Controls the BIOS versions and parameters
- **Boot Policy:** Establishes boot paths and order
- **Local Disk Configuration:** Defines the configuration of any local storage

Policies are an underlying mechanism for propagating changes across the system. When a policy is changed, this change propagates to all service policies that use it, and any servers requiring a reboot are flagged. The major exception to immediate policy execution is firmware policy, which stages the new firmware and applies it either at reboot or upon administrator command. The ability of Cisco UCS Manager to apply these policy updates at reboot has significant operational implications. With the Cisco Unified Computing System, servers with incorrect revisions of firmware, for example, will be forced into compliance automatically at boot, or a new policy specifying a new firmware release can be selectively applied to groups of servers under operator control. In all cases, the firmware update process has transactional semantics and will either be completed correctly or rolled back to a previous state, avoiding messy partial completion scenarios.

Service Profiles and Templates

Cisco Unified Computing System service profiles are a powerful tool for streamlining the management of modern data centers. They provide a mechanism for rapidly provisioning servers and their associated network and storage connections with consistency in all details of the environment, and they can be set up in advance of the physical installation of the servers, which is extremely useful in most organizations. They also enable new operational policies for high availability, since a service profile can be applied to a spare server in another rack or blade slot in the event of a server failure, or the profile can be applied to a bare-metal replacement in the failed unit physical slot. However, these basic service profile use cases show only a small fraction of the power and utility of service profiles and primarily provide a transition methodology as users transition from older operating processes to new processes that fully utilize the capabilities of the Cisco Unified Computing System.

The real power of the service profile becomes evident in templates. A service profile template parameterizes the UIDs that differentiate one instance of an otherwise identical server from another. Templates can be categorized into two types: initial and updating.

- **Initial Template:** The initial template is used to create a new server from a service profile with UIDs, but after the server is deployed, there is no linkage between the server and the template, so changes to the template will not propagate to the server, and all changes to items defined by the template must be made individually to each server deployed with the initial template.
- **Updating Template:** An updating template maintains a link between the template and the deployed servers, and changes to the template (most likely to be firmware revisions) cascade to the servers deployed with that template on a schedule determined by the administrator.

Service profiles, templates, and other management data is stored in high-speed persistent storage on the Cisco Unified Computing System fabric interconnects, with mirroring between fault-tolerant pairs of fabric interconnects.

Service Profile Lifecycle

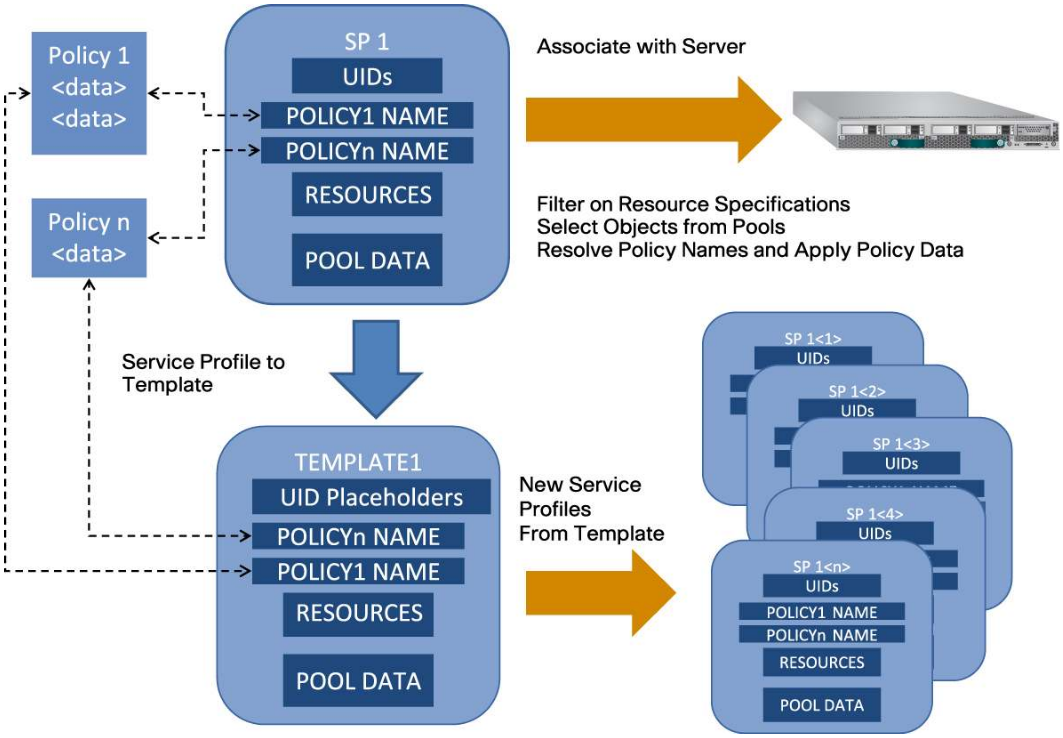
Since service profiles can be used to create templates, and templates can be used to create service profiles, which are then associated with a set of resources, including physical servers, to create a running server, their lifecycle is complicated. The lifecycle of a service profile starts with its creation[4]. Service profiles can be created in several ways:

- **Manually:** Create a new service profile using the Cisco UCS Manager GUI.
- **From a Template:** Create a service policy from a template.
- **By Cloning:** Cloning a service profile creates a replica of a service profile. Cloning is equivalent to creating a template from the service policy and then creating a service policy from that template to associate with a server.

Figure 4 shows the basic cycle of service profile and template creation. Note that the policies are reusable objects, referenced by names from the profiles.

[4] At a low level, a service profile can be considered a collection of policy objects, resources from the various resource pools, and system-specific UID information. As noted earlier, a template is created from the service policy by parameterizing the UIDs and resources. To create a service policy from the template, values and resource names are substituted for the parameters, and then the service profile is associated with the physical resources.

Figure 4. The Cisco Unified Computing System Service Profile Is the Basis for Templates, Which in Turn Can Be the Basis for Multiple Service Profiles



Service Profiles and System Deployment

The most common use of a template after it has been defined is to deploy new applications on a fabric of underlying physical servers, storage solutions, and networks.

A template, either initial or updating, can be used to generate either a single instance or multiple instances of its underlying server definition, each with a different UID. This capability is one of the most powerful features of service profiles and is the underlying mechanism that enables organizations to use the Cisco Unified Computing System to scale a workload to meet demand, deploy large numbers of servers in a fabric, deploy an instance of a complex application that spans multiple servers in a distributed environment, and deploy other similarly powerful applications.

Whether using simple service profiles or service profile templates, Cisco Unified Computing System users can deploy service profiles and templates for servers and applications in several ways:

- **Manually:** Uses UIDs supplied by the server administrator at deployment time; this is the approach generally used when service profiles are used in a standalone manner rather than as templates
- **Automatically:** Uses UIDs generated by the template at runtime; names are in the form <Name_NN>, where NN is a sequence number generated by the deployment engine, and MAC address, WWN, etc. is chosen from the available resource pools according to defined policy
- **Programmatically:** Controlled by external software though the available Cisco Unified Computing System interfaces

Service Profiles and Cost Savings

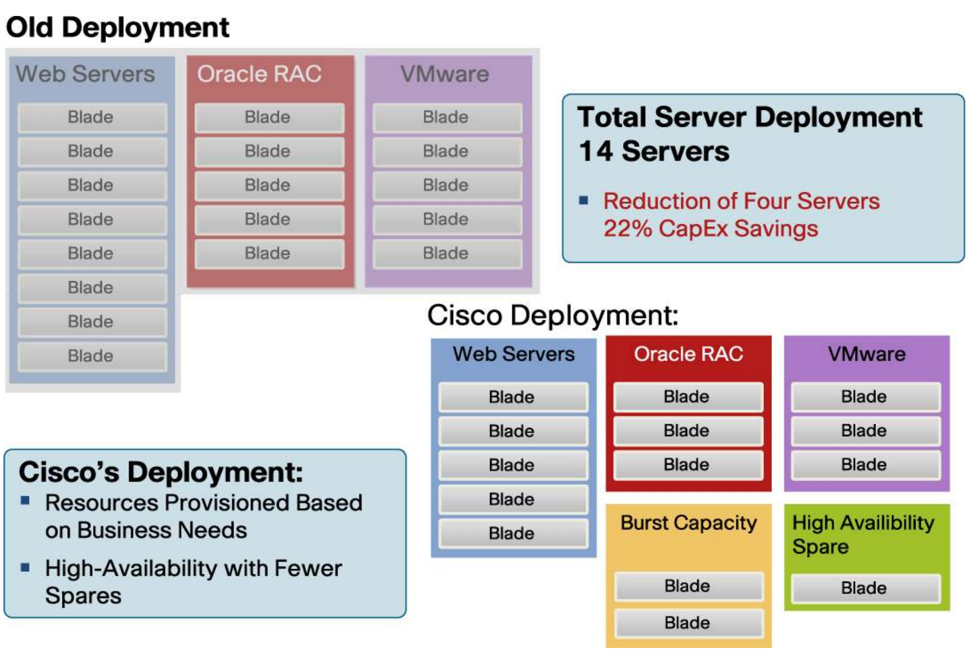
Service profiles, particularly when set up as updating templates, have powerful capabilities that yield immediate and measurable benefits to data center operations. The capability to stage firmware updates to an updating template and then trigger a bulk update of all systems associated with that template provides fast updates at nearly no labor cost, and because the entire update cycle is transactional on a per-server basis, the updates will either be completed correctly or rolled back to the initial state - a feature that prevents the extremely challenging scenario of only partially successful updates.

The financial effect on operations is straightforward, valuable, and easy to measure:

- **Reduced Labor Cost:** Most operations can be performed more quickly by staff with lower skill levels. Cisco UCS Manager provides granular, role-based privileges to allow delegated authority to, for example, deploy but not change a template. Delegation can be implemented along organization boundaries as well as by privilege.
- **Reduced Error Rate:** Most organizations are hesitant to discuss error rates and their consequences, both for proprietary reasons and to avoid exposure to potential liability, but the limited available data as well as a rich store of anecdotal data suggests that the error rate for CLI operations is somewhere in the range of 1 to 3 percent, with a wide range of recovery times and costs. Cisco Unified Computing System service policies and templates should be able to reduce the error rate effectively to zero, decreasing costs, reducing disruptions, and freeing time for more valuable activities.
- **Simplification of Tool Portfolio:** The combined capabilities of Cisco UCS Manager and templates may allow the elimination of third-party firmware management tools.

Service profiles, in combination with the stateless nature of Cisco Unified Computing System servers, provide the underlying mechanism that allows the use of a common pool of spare servers that can be quickly repurposed for nearly any requirement. For most organization and applications, this feature can result in an immediate reduction in capital expenditures (CapEx) because required spare and overflow capacity can be shared among multiple departments and applications, as shown in Figure 5. Users can tailor the cost and acceptable risk by varying the size of these shared resource pools.

Figure 5. The Cisco Unified Computing System Service Profiles and Resource Pools Allow the Sharing of Resources That Are Not Expected to Be Used on a Normal Duty Cycle, Such as Burst Capacity and High-Availability Spares



Additional CapEx benefits are inherent in the converged fabric architecture of the Cisco Unified Computing System.

While service profiles provide definite CapEx advantages, it is in the ongoing management of complex data centers that the cost benefits of the Cisco Unified Computing System and service profiles are most evident. Ongoing operating expenses (OpEx) for previously deployed applications can be roughly categorized into three types of costs: maintenance (primarily firmware updates and application and OS patches); moves, additions, and requests; and exception conditions such as system failure and the need to add capacity.

In addition, Cisco UCS Manager, through its use of service profiles, enables the streamlining of many routine maintenance operations.

While each customer environment is unique, customers are reporting CapEx reductions of 20 percent or more, and OpEx reductions in a wider range, depending on how aggressively they move to take advantage of advanced features of the Cisco Unified Computing System, including templates[5].

Service Profiles and Compliance

Service profiles aid in demonstrating operational controls to meet any compliance reporting requirements. Combined with the rich suite of Cisco network compliance and reporting tools and third-party tools from vendors such as BMC BladeLogic Server Automation Suite, CA Spectrum Infrastructure Manager and others, service profiles support a comprehensive reporting and compliance program.

In addition, the Cisco Unified Computing System eliminates configuration drift, one of the most pernicious problems in large-enterprise management, since noncompliant changes to configurations cannot be made. Today, many companies use a combination of third-party discovery and configuration management reporting tools to monitor configuration drift. With Cisco Unified Computing System, this expense and process is eliminated.

Cisco Unified Computing System Compared to Competitors

When a new technology category is emerging, users often have difficulty determining the validity of seemingly conflicting claims from competing vendors.

Multiple vendors claim capabilities similar to those of the Cisco Unified Computing System: solutions that purport to offer an integrated environment that presents a virtualized pool of resources to the data center with capabilities for flexible deployment. While many of these tools are quite capable within limited subdomains of the data center infrastructure, all suffer from one or more major shortcomings, including:

- **Multiple Products:** Most of these solutions are in fact a collection of products, with the appearance of integration provided by a common console from which the tools for the various products are invoked.
- **Old Code Bases:** A solution may offer a visually appealing interface with a good presentation of required functions, but the worst problems with some of these tools are not readily visible to the user at first glance, since they have to do with the provenance of the software code base. Some products on the market today are composed of major subsystems from multiple lineages and different code bases, some more than a decade old and most developed by different teams at different times, and sometimes in different companies. This development history has significant repercussions for the solution’s reliability and the speed with which the vendor can add new features to the solution.

[5] The major variable in OpEx reduction seems to be the degree to which users invest in developing templates and using the full automation capabilities of Cisco UCS Manager. For customers who need to routinely deploy a high number of servers, reconfigure servers, or deploy replicates of applications for their customers, Return on Investment (ROI) can be very high and very rapid.

- **Poor Integration:** Reflecting the varied origins of the technology base, integration of the various functional modules is often incomplete, resulting in such complexities as the need to switch between multiple subconsoles to perform specific management tasks, often those involving storage and networks and deployment. Perhaps the strongest indication of the unintegrated nature of some products is that some components still have separate licensing. Another common practice is to extend the management scope beyond the core infrastructure and to include additional features that overlap the major hypervisor features. Cisco Unified Computing System customers providing feedback about virtual machine management have unanimously favored retaining their vendor-specific management tools to manage their virtual machines, and they have no desire to add a less-functional, lower-performance management tool from another vendor to their portfolio.
- **Weak Service Profile Equivalent:** Some vendors can manage WWN and MAC addresses, but no product on the market today offers the equivalent of Cisco Unified Computing System service profiles, with their capability to span server, storage, and network configuration elements, and with their advanced policy features. In almost all cases, the scope of the management domain stops at the server edge and does not extend to the surrounding network.
- **No Converged Fabric Storage Integration:** Cisco is unique in its management of converged fabric environments. Other vendors offer converged fabric capabilities in their system, but have gaps in the integration of these capabilities into their management stacks.

Cisco Unified Computing System and Cost Benefits

Technology for technology's sake is not Cisco's business, and when we designed Cisco Unified Computing System, it was with a clear set of customer benefits in mind. By providing a virtualized environment and an integrated management and operations environment, our customers are able to reduce IT costs and improve profitability, as well as the agility of the overall business as it responds to competitive forces and new market opportunities.

Reduced Costs

In most organizations and deployments, Cisco has observed total CapEx saving of more than 20 percent from the combination of reduced server count and efficiencies inherent in the Cisco Unified Computing System solution's converged fabric architecture.

Likewise, ongoing OpEx is reduced by the advanced one-to-many deployment and management capabilities of the Cisco Unified Computing System, lowering labor expenses, allowing cleaner separation of management roles, and reducing errors.

The money saved in both CapEx and OpEx can be applied as corporate policy dictates. It can be applied to other IT initiatives or allocated to other corporate functions, or it can flow to profits. Corporate management is in charge, instead of being bound by old-technology cost curves.

Agility and Opportunity

With rare exceptions, business today depends on IT, and in some industries, IT is the business. By providing an environment that allows rapid deployment and reconfiguration and enforces standardization, Cisco is providing a substrate upon which the business can rapidly implement new solutions to address both competitive pressures and new opportunities. The multiple-month process for deploying a new application on older data center infrastructure can be reduced to hours or minutes with Cisco Unified Computing System, removing a major impediment to innovation.

Conclusion

Cisco Unified Computing System™ technology has redefined the enterprise computing environment. By breaking the traditional data center model and redefining data center infrastructure as pools of virtualized server, storage, and network resources, the Cisco Unified Computing System has delivered a new computing model with advantages in capital and operational cost, improving flexibility and availability, and reducing the amount of time needed for IT to respond to business changes.

A key element in the power of Cisco Unified Computing System is the service profile, the fundamental mechanism by which the Cisco Unified Computing System models the necessary abstractions of server, storage, and networking. This document has explained the underlying structure of Cisco Unified Computing System service profiles, how they are used to enable Cisco Unified Computing System capabilities, and the benefits that accrue from their use.

For More Information

For more information, go to http://www.cisco.com/en/US/products/ps10265/index.html or contact your local Cisco office.

Americas Headquarters
Cisco Systems, Inc.
San Jose, CA

Asia Pacific Headquarters
Cisco Systems (USA) Pte. Ltd.
Singapore

Europe Headquarters
Cisco Systems International BV
Amsterdam, The Netherlands

Cisco has more than 200 offices worldwide. Addresses, phone numbers, and fax numbers are listed on the Cisco Website at www.cisco.com/go/offices.

Printed in USA C11-590518-01 04/10

