Oracle Elite
Engineering Exchange

Oracle's SPARC T7 and SPARC M7 Servers: Domaining Best Practices

ORACLE TECHNICAL WHITE PAPER | OCTOBER 2015

Table of Contents

Introduction

The benefits of enterprise consolidation are well understood. By consolidating workloads, applications, databases, operating system instances, and servers, it is possible to reduce the number of resources under management, resulting in improved system utilization rates and lower costs. With higher utilization rates, the need to make additional hardware purchases is reduced. If consolidation also can be combined with simplification of the overall IT infrastructure, considerable savings can be made in the operational costs of running the data center.

Consolidation also contributes to strategic goals, such as improving security, delivering more predictable service levels, and increasing application deployment flexibility. With the addition of Oracle's latest SPARC T7 and SPARC M7 servers—the SPARC T7-1, T7-2, T7-4, M7-8, and M7-16 servers, all of which are based on the SPARC M7 processor—, price/performance scales linearly without the cost penalty for “big iron” or its enhanced features. What this means is that 8 of Oracle’s SPARC T7-1 servers with 8 total CPUs are similar in price to a 8-CPU SPARC M7-8 server. This effectively removes the large price premium traditionally associated with this class of system, providing an additional reason to use bigger servers: namely, that for the SPARC platform, it is no longer cheaper to procure a number of smaller servers instead of a single larger one.

For successful consolidation deployments, it is necessary to select a server platform that has the scalability to support many application instances. Additionally, the server platform must have the high availability needed for mission-critical applications, the resource management and virtualization capabilities to simplify managing numerous applications, and the tools to manage the consolidated environment.

Oracle’s SPARC M7 processor-based servers deliver on all these requirements and are ideal platforms for server consolidation. With the SPARC M7 processor-based servers, IT managers can create pools of compute resources that can be rapidly and dynamically allocated to meet new and changing workloads.

Why Server and Application Consolidation?

Traditionally, each instance of an application has been deployed on a single server. In the case of complex enterprise applications, this style of deployment means that data centers require many servers for a single application, with separate servers for the web tier, application tier, and database tier.

Furthermore, many enterprise applications require test and development servers in addition to the production servers. Commonly, the production servers, when initially deployed, have enough headroom to support spikes in the workload, but as the applications grow, the only way to add more capacity is to add more servers, thereby increasing complexity. As the number of servers increases, the number of operating system (OS) instances that need to be managed also grows, adding further layers of complexity and reducing IT flexibility.

Server utilization is normally very low—between 10 percent and 30 percent—in the one application-per-server deployment model, which is a very inefficient use of server resources. Each server needs to be large enough to handle spikes in workload, but normally will need only a small part of the server capacity.

Figure 1 illustrates this point, showing many small servers running a single application instance. Each one of these servers needs to have enough headroom to meet peak capacity requirements and cannot "share" headroom with other servers that need more capacity or have excess capacity.

If these servers could share headroom, loaning it out or borrowing it as needed, they would have higher utilization rates. By consolidating multiple applications on a single larger server, where resources shift dynamically from application to application, the workload peaks and troughs tend to even out, and the total compute requirement is less variable. The more applications that are consolidated, the more even the server usage. Applications that are consolidated on a larger server benefit from shared headroom, so consolidating applications can lead to much higher server utilization because excess capacity is reduced significantly.

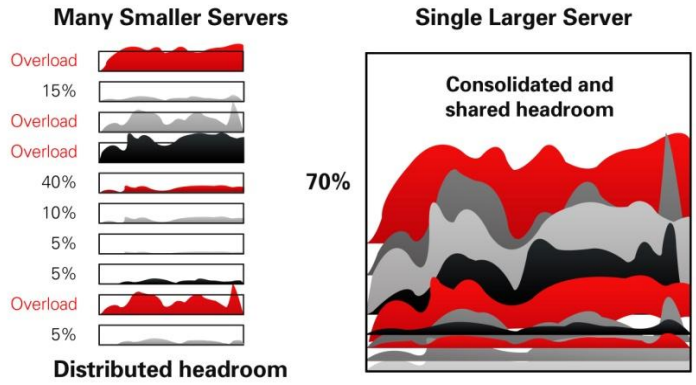

Figure 1. Consolidating and sharing headroom in large symmetric multiprocessing servers.

Improved server utilization means more efficient use of server resources, which improves ROI and reduces the total server hardware required to meet workload requirements.

Consolidating many older and smaller servers onto fewer larger, newer servers provides many benefits beyond improved utilization. The newer servers will have more capacity, better performance, better energy and space efficiencies, and improved availability features, and they will be easier to manage.

Requirements for Consolidation

Servers used for consolidation must provide scalability and high capacity, high availability, and simple upgrade paths. They also must enable the reuse of existing applications and have effective virtualization and resource management tools. Because applications are combined on consolidated servers, these servers need the capacity to handle dozens of workloads of all types. The performance of each application, when consolidated with other applications, must match or exceed its performance when deployed by itself on its own server.

Consolidation, by definition, means putting "more eggs in one basket," so a system failure will have a greater effect on application availability than if each application were deployed on its own server. Servers used for consolidation must have high-availability features, both in hardware and software, to reduce both planned and unplanned downtime. Consolidation servers must be extremely reliable so that they rarely go down. They also need to have advanced serviceability features so they can be reconfigured, upgraded, and repaired with minimal or no downtime.

Consolidation servers are mainly used to run older applications in a newer environment, so they must be able to run legacy applications as well as new applications.

A consolidation environment will have many workloads of different types, and these various workloads all will have specific patch, resource, security, and performance requirements. In many cases, the operating system will have enough tools to manage multiple applications, but in other cases, applications will require separate environments to run effectively. Virtualization and resource management tools are required so that the pool of resources in a consolidation server can be partitioned and deployed as needed for multiple applications. Virtualization enforces application separation, and resource management guarantees that the performance requirements of each application are met.

Consolidation on Large, Vertically Scalable High-End SMP Servers

All servers consist of the same essential components, but different server architectures combine, connect, and utilize these components in different ways.

Large, vertically scalable symmetric multiprocessing (SMP) servers, such as Oracle's SPARC M7-8 and M7-16 servers, have dozens of processors and I/O slots and terabytes of RAM, all housed in a single cabinet that can be flexibly deployed in a single massive OS instance or separated into resource managed domains.

In essence, vertically scalable servers are large pools of resources that can support dozens of workloads of various sizes and types to simplify consolidation and application deployment. New applications can be deployed on a large SMP server, eliminating the need to install a server for each new application. Existing applications can grow by taking advantage of the extra headroom available.

Vertically scalable servers—generally larger SMP servers hosting eight or more processors—have a single instance of the OS to manage multiple processors, memory subsystems, and I/O components, which are contained within a single chassis. Most vertically scalable servers, such as Oracle's SPARC M7 processor-based servers, also can be partitioned using virtualization tools to create multiple instances of the OS using subsets of each server's resources. Virtualization tools are used to share or separate resources, as required, based on the workload and the security and availability requirements.

In a vertically scalable design, the system interconnect is commonly implemented as a tightly coupled centerplane or backplane that provides both low latency and high bandwidth. In vertical or SMP systems, memory is shared and appears to the user as a single entity. All processors and all I/O connections have equal access to all memory,

eliminating data placement concerns. Oracle's high-end SPARC SMP servers have provided linear scalability since 1993, demonstrating the value of tight, high-speed and low-latency interconnects.

The cache-coherent interconnect maintains information on the location of all data, regardless of its cache or memory location. There are no cluster managers or network interconnects in SMP servers, because the internal interconnect handles all data movement automatically and transparently. Resources are added to the chassis by inserting system boards with additional processors, memory, and I/O subassemblies. Vertical architectures also can include clusters of large SMP servers that can be used for a single, large application.

High-end SMP servers greatly simplify application deployment and consolidation. Large SMP servers have a huge pool of easily partitioned processor, memory, and I/O resources. This pool of resources can be assigned dynamically to applications using Oracle Solaris Resource Manager and manipulated using standard systems management tools such as Oracle Enterprise Manager Ops Center.

SPARC M7-8 and M7-16 Server Consolidation Technologies

The following sections examine the consolidation technologies that enable the deployment of many applications together to improve system utilization, optimize the use of computing resources, and deliver greater ROI from IT investments. Figure 2 shows the various levels of virtualization technologies available, at no cost, on Oracle's SPARC M7-8 and M7-16 servers.

- At the lower tier of the virtualization stack is the SPARC platform itself. The SPARC platform provides the first level of virtualization: the PDom[1] feature (also known as physical domains or Dynamic Domains), which first appeared in Oracle's SPARC Enterprise M-Series servers. PDoms are electrically isolated hardware partitions, meaning they can be completely powered up or down and manipulated without affecting any other PDoms.
- At the second level of virtualization, each PDom can be further split into hypervisor-based Oracle VM Server for SPARC partitions (also known as logical domains or LDoms). These partitions can run their own Oracle Solaris kernel and manage their own I/O resources. It's not uncommon to have different versions of Oracle Solaris running different patch levels under Oracle VM Server for SPARC. Oracle VM Server for SPARC is also recognized as an Oracle hard partition for software licensing purposes[2].
- The third level of virtualization is Oracle Solaris Zones technology, the finest-grained level of virtualization, and a feature of Oracle Solaris. Oracle Solaris Zones share a common Oracle Solaris kernel and patch level. They have significant advantages of flexibility when it comes to creation and reboot, and they are extremely fast and lightweight. Each of these instances of Oracle Solaris can use Oracle Solaris Resource Manager to limit the CPU or memory resources that an application can consume, usually managed with Oracle Enterprise Manager Ops Center.

All of these virtualization techniques are very useful for consolidating many applications onto a single server. The next few sections describe these virtualization and resource management technologies in more detail.

1 Oracle has deemed certain technologies, possibly modified by configuration constraints, as hard partitioning, and no other technologies or configurations qualify. Approved hard partitioning technologies include Physical Domains (also known as PDoms, PDomains, Dynamic Domains, or Dynamic System Domains), Oracle Solaris Zones (also known as Oracle Solaris Containers, capped Zones/Containers only).

2 As of the Oracle VM Server for SPARC 2.0 release, hard partitioning is enforced by using CPU whole-core configurations and specifying the maximum number of cores that can be assigned to the domain.

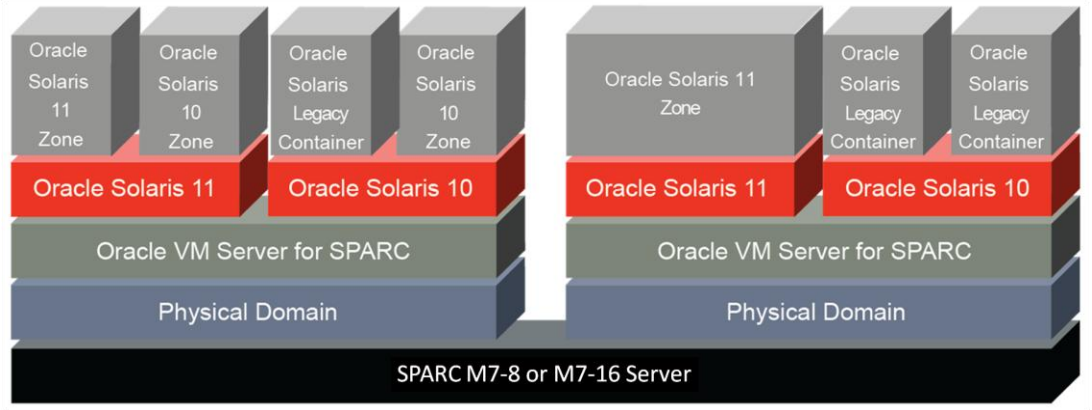

Figure 2. Virtualization technology stack on the SPARC M7-8 and M7-16 servers.

Physical Domains (PDoms)

As mentioned above, on M7-8 and M7-16 servers, PDoms enable electronically isolated partitions. PDoms make it possible to isolate multiple applications and multiple copies of the Oracle Solaris OS on a single server. The PDoms feature enables administrators to isolate hardware or security faults and constrain their exposure to each domain. The result is a superior level of system availability and security. The PDoms feature is now in its seventh generation, having previously been available in Oracle's SPARC Enterprise M-Series servers, making it the most mature and established partitioning option in the UNIX server market. As discussed below, by running Oracle VM Server for SPARC, PDoms can be further virtualized. This allows multiple independent Oracle Solaris instances to coexist within the same PDom.

With PDoms, software and hardware errors and failures do not propagate beyond the domain in which the fault occurred. Complete fault isolation between PDoms limits the effect on applications of any hardware or software errors. This helps to maintain a high level of availability in these servers, which is necessary when consolidating many applications. The PDoms feature separates the administration of each domain, so a security breach in one domain does not affect any other domain.

Oracle VM Server for SPARC

Oracle VM Server for SPARC, also called LDoms or logical domains, provides full virtual machines (VMs) that run independent instances of the operating system and are available on all of Oracle's SPARC T-Series servers; SPARC M5 processor–based servers; SPARC M6 processor–based servers; and SPARC M7 processor–based servers. Each operating system instance contains dedicated CPU, memory, storage, and console devices. LDoms are unique in that many of the virtualization functions are provided natively by the underlying hardware, and both CPU and memory are directly assigned to LDoms without incurring any virtualization overhead. I/O can be directly assigned to LDoms, providing the benefit of higher performance, or it can be virtualized, providing the benefit of increased utilization of hardware resources and the ability to use live migration. The number of physical threads limits the number of possible LDoms in the system, although there is an upper limit of 128 LDoms per server or per PDom.

For the purpose of running performance-critical workloads, it is possible to configure the LDoms so that each domain is directly attached to its own PCIe slots by assigning root complexes to domains. Domains of this type are called *root domains*. This allows these domains to be completely independent from each other within a single server

or PDom, and they operate at bare-metal performance levels. The SPARC M7 processor-based servers have been designed with considerably more root complexes, allowing a much finer granularity of PCIe slot allocation. This in turn allows more root domains to be created per platform than with previous SPARC processor generations.

In many respects, a PDom and a root domain are very similar. They both provide a fully independent domain within a physical server with zero virtualization overhead. The PDom provides more isolation, but it is less flexible, because the PDom granularity is higher, and it is impossible to dynamically reallocate CPU, memory, and I/O resources between PDoms. A root domain, while maintaining the same performance characteristics of a PDom, provides slightly less isolation, but this allows the dynamic reallocation of CPU, memory, and I/O resources, as required.

Oracle VM Server for SPARC has the ability to perform live migration of a domain from one system to another. As the name implies, the source domain and application do not need to be halted or stopped. This allows an LDom to be migrated to another PDom on the same server or on a different server. Live migration is possible only on domains that are using virtual I/O, and is, therefore, not possible on domains using the root domain model, or single root I/O virtualization (SR-IOV).

By layering LDoms on top of PDoms, organizations gain the flexibility to deploy multiple operating systems simultaneously on multiple electrically isolated domains. These domains all run Oracle Solaris, which can additionally host Oracle Solaris Zones to create yet another layer of virtualization.

Oracle Solaris

The Oracle Solaris OS is very efficient at scheduling large numbers of application processes among all the processors in a given server or domain, dynamically migrating processes from one processor to the next based on workload. For example, many enterprises run more than 100 instances of Oracle Database on single SPARC server using no virtualization tools. Oracle Solaris is able to effectively manage and schedule all the database processes across all the SPARC cores and threads.

With this approach, a large vertically scalable server can assign resources as needed to the many users and application instances that reside on the server. Using Oracle Solaris to balance workloads can reduce the processing resource requirements, resulting in fewer processors, less memory, and lower acquisition costs. Oracle Solaris increases flexibility, isolates workload processing, and improves the potential for maximum server utilization.

Oracle Solaris Zones

In a consolidated environment, it is sometimes necessary to maintain the ability to manage each application independently. Some applications might have strict security requirements or might not coexist well with other applications, so organizations need the capability to control IT resource utilization, isolate applications from each other, and efficiently manage multiple applications on the same server.

Oracle Solaris Zones technology (formerly called Oracle Solaris Containers), available on all servers running Oracle Solaris, is a software-based approach that provides virtualization of compute resources by enabling the creation of multiple secure, fault-isolated partitions (or zones) within a single Oracle Solaris instance. By running multiple zones, it is possible for many different applications to coexist in a single OS instance.

The zones environment also includes enhanced resource usage accounting. This highly granular and extensive resource tracking capability can support the advanced client billing models required in some consolidation environments.

Oracle Solaris Kernel Zones

Oracles Solaris Kernel Zones are the latest addition to Oracle Solaris Zones technology. Native Oracle Solaris Zones are extremely powerful and efficient due, in part, to the underlying shared Oracle Solaris kernel. However, this also means that all native zones need to run the same OS version and patch levels, which makes it impossible to run applications that need a specific kernel version or setting. This same shared-kernel technology makes it difficult to live-migrate these zones. Kernel zones are simply zones that allow different, independent OS versions and patch levels, which in turn creates more isolation between workloads.

Oracle Solaris Resource Manager

Oracle Solaris Resource Manager is a group of techniques that allows the consumption of CPU, memory, and I/O resources to be allocated and shared among applications within an Oracle Solaris instance, including within Oracle Solaris Zones. Oracle Solaris Resource Manager uses resource pools to control system resources. Each resource pool may contain a collection of resources, known as *resource sets*, which may include processors, physical memory, or swap space. Resources can be dynamically moved between resource pools as needed. Also, with Oracle Solaris 11, network services virtualization is greatly increased as well.

Fair Share Scheduler

Oracle Solaris Resource Manager incorporates an enhanced fair share scheduler, which may be used within a resource pool. When using the fair share scheduler, an administrator assigns processor shares to a workload that may comprise one or more processes.

These shares enable the administrator to specify the relative importance of one workload to another, and the fair share scheduler translates that into the ratio of processor resources reserved for a workload. If the workload does not request processor resources, those resources may be used by other workloads. The assignment of shares to a workload effectively establishes a minimum reservation of processor resources, guaranteeing that critical applications get their required server resources.

Managing Consolidation Technologies Using Oracle Enterprise Manager Ops Center

One of the key goals of server consolidation is to simplify server management by reducing the number of servers and OS instances that need to be managed. Oracle Enterprise Manager Ops Center 12*c* achieves this by merging the management of systems infrastructure assets into a unified management console.

Through its advanced server lifecycle management capabilities, Oracle Enterprise Manager Ops Center 12*c* provides a converged hardware management approach that integrates the management of servers, storage, and network fabrics, including firmware, operating systems, and virtual machines. Oracle Enterprise Manager Ops Center 12*c* provides asset discovery, asset provisioning, monitoring, patching, and automated workflows. It can also discover and manage virtual servers as well as physical servers, simplifying the management of SPARC M7 processor-based servers and Oracle SuperCluster M7 as well as all other Oracle servers in a data center. Oracle Enterprise Manager Ops Center 12*c* is available free of charge to all Oracle server customers who have Oracle Premier Support contracts.

Layered Consolidation with SPARC M7 Processor–Based Servers

The most important aspect of vertically scaled systems is the flexibility of deployment models. In a horizontally scaled environment, there is usually only one virtualization technique: the use of VMs. Vertically scaled systems offer consolidation opportunities at a number of layers, and this helps drive higher utilization and greater simplicity.

The SPARC M7-8 and M7-16 servers offer three main types of layered virtualization at the infrastructure level:

1. Oracle Solaris Zones: Allowing multiple application coexistence and resource management within a single OS instance.

2. Oracle VM Server for SPARC: Allowing multiple OS instances to coexist on the same physical infrastructure with dynamic reallocation of hardware resources.

3. PDoms: Partitioning of a server into independent isolated servers.

The SPARC T7-1, T7-2, and T7-4 servers offer two main types of layered virtualization at the infrastructure level:

1. Oracle Solaris Zones: Allowing multiple application coexistence and resource management within a single OS instance.

2. Oracle VM Server for SPARC: Allowing multiple OS instances to coexist on the same physical infrastructure with dynamic reallocation of hardware resources.

Each of the virtualization techniques provides different benefits. In general, zones provide the highest flexibility and dynamic usage of resources, but the lowest isolation and less-granular serviceability. PDoms provide the greatest amount of isolation, but provide much less flexibility. The most appropriate deployment model is likely to be a blended approach of all three of the technologies above. For Oracle software licensing purposes, PDoms, Oracle VM Server for SPARC, and Oracle Solaris Zones are all considered to be hard partitions for licensing software[3].

A Consolidation Philosophy

When faced with multiple options for consolidation, it is useful to remember the reasons for consolidating in the first place, and use those initial requirements to derive the most appropriate solution.

- Maximize operational efficiency
  - The benefits of consolidation are not purely from a reduction in hardware cost. The majority of the consolidation benefits are derived from the simplicity that accrues from standardization of an operating model, and the reduction in the number of managed objects.
  - Consolidating as high up the stack as possible naturally reduces the total number of managed objects and creates as much standardization as possible.
- Maximize workload efficiency
  - One of the trade-offs of increased isolation is a potential increase in the virtualization overhead. Bear this in mind, and create additional isolation only where necessary.
  - Some workloads are quite small in comparison to the footprint of a modern OS instance. Try to co-locate multiple workloads per OS instance where possible.

3 Please refer to the Oracle Partitioning Policy for the most up-to-date rules concerning the use of these technologies as hard partition boundaries: http://www.oracle.com/us/corporate/pricing/partitioning-070609.pdf

Figure 3 illustrates the two extremes of the spectrum and the options in between:

Figure 3. Options for deploying workloads with varying levels of isolation and flexibility.

It is possible to deploy 64 workloads with the highest possible isolation possible by having four independent PDoms, with each PDom running 16 LDoms, with an OS instance per workload.

This layout allows the highest level of isolation, but comes at a cost of much higher complexity, because supporting 64 LDoms will require having 128 virtual boot disks configured. Additional service domains will be needed to provide services for each domain and the 64 unique OS instances.

At the other end of the spectrum, would be to have a single PDom spanning the whole system, with a single Oracle Solaris instance running in it, with 64 zones, each running one of the workloads[4].

This option is the most efficient in terms of resource utilization and operational simplicity. However, it opens up a number of serviceability and manageability challenges, because it creates a single failure domain, which will have a high impact for both planned and unplanned outages. The aim should be to keep as far to the left of Figure 3 as possible, while moving to the right as isolation and serviceability requirements demand.

The reality is that the optimum solution based on the characteristics of the workloads in question is somewhere between these two extremes, and the intent of this white paper is to discuss the three layers of virtualization technologies in sufficient detail so as to allow organizations to make an educated choice.

Physical Domains on SPARC M7-8 and SPARC M7-16 Servers

The SPARC M7-8 and SPARC M7-16 servers feature a balanced, highly scalable SMP design that connects their SPARC M7 processors to memory and I/O through a high-speed, low-latency system interconnect.

The SPARC M7-8 server consists of a single CPU, memory, and I/O unit (CMIOU) chassis. The SPARC M7-16 server consists of a pair of CMIOU chassis, which are connected together using a switch chassis.

Each CMIOU chassis can contain up to eight CMIOU boards, which consist of one SPARC M7 processor, 16 memory slots, and three PCIe I/O slots. The chassis can be configured as either a single PDom or two PDoms. Each PDom must have at least two CMIOU boards. This allows the following configurations:

» SPARC M7-8 Server

4 You aren't limited to 64 zones, The theoretical limit is more than 8,000 zones per Oracle Solaris instance.

- Single PDom with 2–8 CMIOU boards
- Two PDoms each with 2–4 CMIOU boards

The choice of one or two PDoms is a factory configuration choice, although the PDom sizing can be altered in the field by populating it with CMIOU boards.

- SPARC M7-16 Server
  - Single PDom with 2–16 CMIOU boards
  - Two PDoms each with 2–8 CMIOU boards
  - Three PDoms, one with 2–8 CMIOU boards, and two with 2–4 CMIOU boards
  - Four PDoms, each with 2–4 CMIOU boards

These PDoms are field-configurable, and provide a wide range of flexible configurations in terms of both the number of PDoms and the sizing of those PDoms.

A PDom operates like an independent server that has full hardware isolation from any other PDoms in the chassis. A hardware or software failure within one PDom will not affect any other PDom in the chassis. For Oracle software licensing purposes, a PDom is considered a hard partition.

For more details about the physical configuration of the SPARC M7-8 and SPARC M7-16 servers, please refer to the “Oracle’s SPARC T7 and SPARC M7 Server Architecture” white paper.

Oracle VM Server for SPARC: Logical Domains

An Oracle VM Server for SPARC domain (also referred to as a logical domain or LDom) is a virtual machine composed of a discrete logical grouping of resources. A logical domain has its own operating system and identity within a single computer system. Each logical domain can be created, destroyed, reconfigured, and rebooted independently, without requiring a power cycle of the server. A variety of application software can be run in different logical domains and kept independent for performance and security purposes. For Oracle software licensing, an LDom is considered a hard partition.

Each logical domain is permitted to observe and interact only with those server resources that are made available to it by the hypervisor. The logical domains manager enables users to manipulate the hypervisor via the control domain. Thus, the hypervisor enforces the partitioning of the server's resources and provides limited subsets to multiple operating system environments. This partitioning and provisioning is the fundamental mechanism for creating logical domains. Figure 4 shows the hypervisor supporting four logical domains. It also shows the following layers that make up the logical domains' functionality:

- User/services (applications)
- Kernel (operating systems)
- Firmware (hypervisor)
- Hardware, including CPU, memory, and I/O

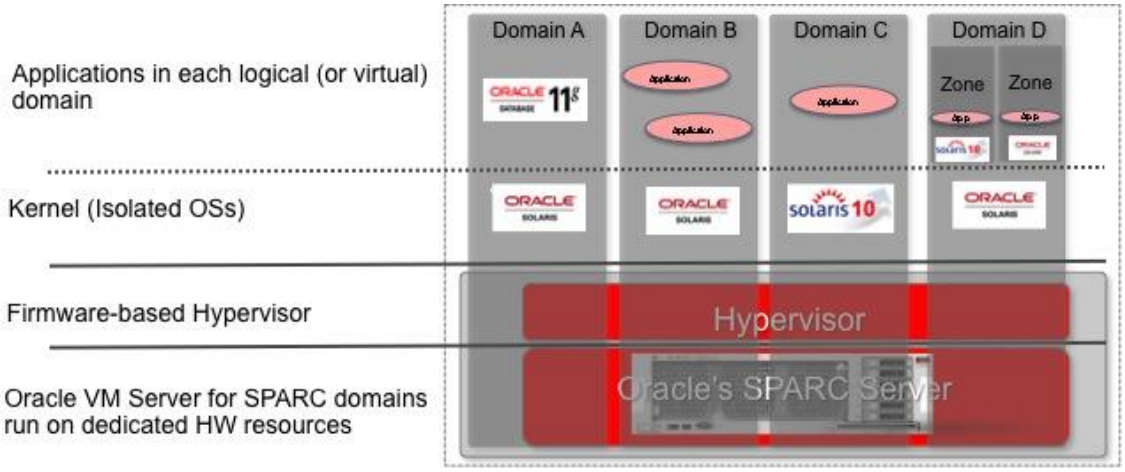

Figure 4. Oracle VM Server for SPARC virtualization.

The number and capabilities of each logical domain that a specific SPARC hypervisor supports are server-dependent features. The hypervisor can allocate subsets of the overall CPU, memory, and I/O resources of a server to a given logical domain. This enables simultaneous support of multiple operating systems, each within its own logical domain. Resources can be rearranged between separate logical domains with an arbitrary granularity. For example, CPUs are assignable to a logical domain with the granularity of a CPU thread.

Each logical domain can be managed as an entirely independent machine with its own resources, such as the following:

- Kernel, patches, and tuning parameters
- User accounts and administrators

- Disks
- Network interfaces, Media Access Control (MAC) addresses, and IP addresses

Each logical domain can be stopped, started, and rebooted independently of the others without requiring users to perform a power cycle of the server.

LDoms Inside PDoms

Oracle VM Server for SPARC provides the flexibility to further carve up the physically isolated domains into more domains and allows substantially more-independent OS instances than possible when using PDoms alone.

The SPARC M7-8 and SPARC M7-16 servers support a maximum of two or four PDoms, respectively. It is expected that many deployments will make use of the Oracle VM Server for SPARC technology to create additional workload isolation at the logical domain level, if required.

Although the SPARC T7-1, T7-2 and T7-4 servers do not support PDoms, and it is expected that most deployments will make use of the Oracle VM Server for SPARC technology directly on the server.

The Control Domain

When PDoms or servers are first installed, what is known as the primary or control domain is created. On a SPARC M7 processor-based system, this initial domain *must* run Oracle Solaris 11.3 or later. This control domain initially owns all the hardware available in the PDom/server, including all CPUs, all memory, and all I/O resources.

If only a single domain running Oracle Solaris 11.3 is required, then there is no further work to do, because this configuration does not require the use of Oracle VM Server for SPARC functionality. This type of usage is expected for configurations with very large vertically scaled workloads requiring large numbers of CPU and memory resources. In all other cases, however, Oracle VM Server for SPARC needs to be configured to create the additional domains and to assign I/O ownership to the domains, as required.

I/O, Root, Service, and Guest Domains

A number of different names are used for the various types of domains that can exist in an Oracle VM Server for SPARC deployment. This is complicated by the fact that a domain can be more than one type simultaneously. For example, a control domain is always an I/O domain, and it is usually a service domain. For the purposes of this paper, the following terminology is used to describe the different Oracle VM Server for SPARC domain types:

- **Control domain—**Management control point for virtualization of the server, which is used to configure domains and manage resources. It is the first domain to boot on a power-up, is an I/O domain, and is usually a service domain as well. There can be only one control domain.
- **I/O domain—**Has been assigned physical I/O devices: a PCIe root complex, a PCIe device, or a single-root I/O virtualization (SR-IOV) function. It has native performance and functionality for the devices it owns, unmediated by any virtualization layer. There can be multiple I/O domains.
- **Service domain—**Provides virtual network and disk devices to guest domains. There can be multiple service domains. A service domain is always an I/O domain, because it must own physical I/O resources in order to virtualize them for guest domains. In most cases, service domains have PCIe root complexes assigned to them, and they could be called a root domain in this case.
- **Guest domain—**A domain whose devices are all virtual rather than physical, such as virtual network and disk devices provided by one or more service domains. In common practice, this is where applications are run. There usually are multiple guest domains in a single system.
- **Guest root domain—**A domain that has one or more PCIe root complexes assigned to it, but it is used to run applications within the domain, rather than to provide services the way a service domain does. Physically there is no difference between service domains and guest root domains other than their usage, and they often will be referred to simply as root domains.

The configuration of Oracle VM Server for SPARC LDoms within PDoms on SPARC M7-8 and M7-16 servers is no different from the way LDoms would be configured on a traditional server. The control domain is used to create additional domains and assign CPU, memory and I/O to those domains. The allocation of CPU and memory is relatively straightforward, but the intended purpose of the domains, and the way in which I/O is assigned to the domains, varies widely depending on the use case.

There are, broadly speaking, three models that are typically used when running Oracle VM Server for SPARC, as shown in Table 1.

TABLE 1. THREE MODELS TYPICALLY USED WHEN RUNNING ORACLE VM SERVER FOR SPARC.

Model	Description	Characteristics	Typical Use Cases
Single control/service domain	In this model, the control domain owns ALL the root complexes and creates virtual devices for all the guest domains.	Most flexible model, suits cases where there are larger numbers of relatively small domains, with low impact of failure. All guest domains are affected by an outage of the control domain. Live migration is possible for these guest domains.	Ideal for test and development environments. Useful also for lightweight production environments where availability is provided by horizontal scaling.
Multiple service domains	One or more service domains are created where root complexes are assigned to those service domains. This also allows redundant I/O for the guest domains.	Similar to the above, except that the guest domains are not majorly affected by a control domain or service domain failure.	Good for production environments where higher availability is required.
Guest root domains	With guest root domains, root complexes are directly assigned to the guest domains, and they have direct ownership of their I/O.	Simplest model because there is no need to create multiple virtual disk and network services, but also the least flexible, because there can be only as many domains as there are root complexes. However, these guests have bare-metal performance and are independent of each other.	Ideal for environments where a small number of highly performant and independent domains is required.

These different deployment models are described in much more detail in numerous white papers and webcasts located at oracle.com/us/technologies/virtualization/oracle-vm-server-for-sparc/resources/index.html.

Guest Root Domains

Guest root domains are discussed here in more detail because the expected workloads on SPARC M7 processor-based servers, combined with the additional PCIe root complex availability on these platforms, is likely to be a good fit for this particular operating model, which is also used extensively on the Oracle SuperCluster platform.

A guest root domain is the concept of a domain hosting one or more applications directly, without relying on a service domain. Specifically, domain I/O boundaries are defined exactly by the scope of one or more root PCIe complexes.

This offers a number of key differences over all of the other models available with the Oracle VM Server for SPARC technology and, in particular, it provides a distinct advantage over all other hypervisors using the traditional “thick” model of providing all services to guest VMs through software-based virtualization.

» Performance: All I/O is native (that is, bare metal) with no virtualization overhead.

- Simplicity: The guest domain and associated guest operating system own the entire PCIe root complex. There is no need to virtualize any I/O. Configuring this type of domain is significantly simpler than configuring the service domain model.
- I/O fault isolation: A guest root domain does not share I/O with any other domain. Therefore, the failure of a PCIe card (for example, a NIC or HBA) impacts only that domain. This in contrast to the service domain, direct I/O, or SR-IOV models in which all domains that share those components are impacted by their failure.
- Improved security: There are fewer shared components or management points.

For more details, please refer to "Implementing Root Domains with Oracle VM Server for SPARC."

It is important to note that for many cases, guest root domains are not the only option. A solution that comprises some systems with service domains as well as some systems with guest root domains might be appropriate. In fact, the same PDoms or servers could consist of two root domains running applications and two resilient I/O domains providing SR-IOV services to a number of guest domains.

Oracle Solaris Zones

Oracle Solaris includes a built-in virtualization capability called Oracle Solaris Zones, which allows users to isolate software applications and services using flexible, software-defined boundaries. Unlike hypervisor-based virtualization, Oracle Solaris Zones technology provides OS-level virtualization, which gives the appearance of multiple OS instances rather than multiple physical machines. Oracle Solaris Zones technology enables the creation of many private execution environments from a single instance of the operating system, with full resource management of the overall environment and the individual zones. For Oracle software licensing purposes, zones configured as Capped or Dedicated CPUs are considered to be hard partitions.

The nature of OS virtualization means that zones provide very low-overhead, low-latency environments. This makes it possible to create hundreds, or even thousands, of zones on a single system. Full integration with Oracle Solaris ZFS and network virtualization provides low execution and storage overhead for those areas as well, which can be a problem area for other virtual machine implementations. Zones enable close to bare-metal performance for I/O, making these software components an excellent match for outstanding I/O performance.

Oracle Solaris 11 provides a fully virtualized networking layer. An entire data center network topology can be created within a single OS instance using virtualized networks, routers, firewalls, and NICs. These virtualized network components come with high observability, security, flexibility, and resource management. This provides great flexibility while also reducing costs by eliminating the need for some physical networking hardware. The networking virtualization software supports quality of service, which means that appropriate bandwidth can be reserved for key applications.

Oracle Solaris Zones technology also provides the ability to run older Oracle Solaris versions within zones. This is called *branded zones*. When running an Oracle Solaris 10 global zone, it is possible to run Oracle Solaris 8 and Oracle Solaris 9 zones within it. This allows legacy applications to be easily consolidated onto a more modern platform. Additionally, Oracle Solaris 10 workloads can take advantage of the network virtualization features of Oracle Solaris 11 by running Oracle Solaris 10 zones on top of an Oracle Solaris 11 global zone.

Zones are also integrated with Oracle Solaris DTrace, an Oracle Solaris feature that provides dynamic instrumentation and tracing for both application and kernel activities. Administrators can use DTrace to examine Java application performance throughout the software stack. It provides visibility both within Oracle Solaris Zones and in the global zone, making it easy for administrators to identify and eliminate bottlenecks and optimize performance.

Use Cases

As can be seen from the previous sections, the three layers of virtualization each have different capabilities and can be combined in different ways to deliver the best combination of flexibility and isolation based on the specific requirements of the application workloads.

PDoms deliver the highest amount of isolation, but they have less flexibility in terms of the dynamic reallocation of resources and features such as live migration. Oracle VM Server for SPARC offers a number of models, ranging from the highly isolated but more resource-flexible root domain model, to the less isolated but highly flexible and agile guest domain model, with a number of additional choices in between.

Selecting the most appropriate combination of virtualization technologies is largely dependent on understanding the workloads to be deployed in terms of their resource characteristics, service levels, and variability, as well as optimizing for the most efficient operating model to deliver the required levels of serviceability and availability.

For more information on a methodology of evaluating the different virtualization technologies' characteristics and matching them to workload and operational requirements, refer to the "Consolidation Using Oracle's SPARC Virtualization Technologies" white paper.

Conclusion

At this point, it should be very apparent that while Oracle's SPARC M7-8 and M7-16 servers contain the PDom features of Oracle's previous-generation SPARC Enterprise M-Series servers, all the SPARC M7 processor-based servers can also make full use of the Oracle VM Server for SPARC (LDom) features that have been available on all of Oracle's previous SPARC T-Series systems. This capability provides a layered virtualization solution with Oracle Solaris Zones, which can meet a wide and varied set of requirements. The sheer size of the SPARC M7-16 server with its 8 TB memory footprint provides the most-advanced capabilities spread over thousands of active threads. The SPARC M7 processor-based systems offer both availability and serviceability—designed from the very core of the processor up—to deliver new levels of performance, availability, and ease of use to enterprise-level applications. The sophisticated resource control provided by PDoms, Oracle VM Server for SPARC (LDoms), and Oracle Solaris Zones further increases the value of these servers by helping enterprises to optimize the use of their hardware assets. By deploying fast, scalable SPARC M7 processor-based servers from Oracle, organizations will gain extraordinary per-thread performance and flexibility, which is a strategic asset in the quest to gain a competitive business advantage.

Oracle's SPARC T7-1, T7-2, and T7-4 servers provide a simple capability to procure server resources of the required size with limited variability in configurations. The SPARC M7-8 and M7-16 servers provide a significantly larger number of configuration choices, allowing the platform to be sized for existing needs with the plan to scale up over time as workloads change. It is expected that the decision to procure a SPARC M7-16 server will be based on the need to run highly intensive workloads that are not appropriate for the smaller SPARC-based systems for reasons related to either availability or vertical performance. In many cases, there might be single or multiple large workloads that require a large PDom, but it is also likely that there will be a large number of additional workloads that might be optimally placed on smaller PDoms on multiple LDoms. This flexibility makes the SPARC M7-16 server an ideal consolidation platform.

There are a large number of use cases to consider for the deployment of workloads on a SPARC M7-16 server, and they apply in part to the SPARC M7-8 server. In general, if all the workloads under consideration fit easily into a 4-socket building block, then creating four 4-socket PDoms is a straightforward choice—because it delivers the best

performance and isolation with large enough domains to be able to flexibly allocate resources, while not being large enough to make serviceability a challenge.

Alternatively, if the reason that the SPARC M7-16 server is purchased is to run large, single-OS-instance images requiring more than 4 sockets, the server can be configured to scale upwards all the way up to 16 sockets.

If additional granularity of workload is required, it can be provided by creating LDoms and then creating zones within those LDoms.

When using LDoms, there is the option of deploying a smaller number of large highly performant domains using the guest root domain model, or a larger number of smaller domains using the standard guest model. In both cases, Oracle Solaris Zones can still be layered on top of the LDoms.

In all cases, the model that delivers the required levels of isolation and serviceability with the most simplicity should be chosen.

Best Practices for High Availability

This paper has not discussed high availability (HA) in any great detail. However, HA is an extremely important facet when defining the architecture of a SPARC M7 processor-based system deployment.

Best practices for HA, such as clustering across compute nodes and using remote replication for disaster recovery, should always be applied to the degree that the business requirements warrant. For example, HA should not be implemented with both nodes of a cluster located within the same PDom, although the isolation of the PDoms within a system make it perfectly acceptable to cluster domains in different PDoms within the same physical server. Multiple tiers (for example, web, application, and database) can be placed in different domains and then replicated on other nodes using common clustering technology or horizontal redundancy.

Oracle has published a number of papers that discuss these concepts further and specific to individual workloads. Refer to the Oracle Maximum Availability Architecture and Optimized Solutions sections of the Oracle Technology Network.

Summary of Guidelines

In the simplest possible terms, the following high-level guidelines should be appropriate for SPARC M7 processor based system deployments:

- Select the number of PDoms/servers required based on the workloads to be deployed and the isolation requirements.
- Use Oracle VM Server for SPARC domains (LDoms) within PDoms if further isolation is required. Use the root domain model for a small number of large domains, use domains based on SR-IOV in cases where a larger number of domains with I/O–sensitive workloads is required, and use the virtualized I/O guest model for the largest number of smaller domains and additional migration flexibility.
- In all cases, use Oracle Solaris Zones to encapsulate applications within the domains. Use zones for flexible and dynamic resource control and security isolation.
- Create high availability by using application-level horizontal scaling or application-based clustering, or by using a clustering product to cluster workloads at the zone or domain level.

About Oracle Elite Engineering Exchange

Oracle Elite Engineering Exchange (Oracle EEE) is a cross-functional global organization consisting of Oracle's elite sales consultants (SCs) and systems engineers (Product Engineering). Oracle EEE connects system engineers directly to the top experts in the field through joint collaboration, which enables bidirectional communication about

customer and market trends and deep insight into the technology directions of future-generation products. Oracle EEE brings real-world customer experiences directly to system engineers and engineering technical details and insights to sales consultants, both of which enable better solutions to meet the changing demands of Oracle's customers.

Oracle Corporation, World Headquarters
500 Oracle Parkway
Redwood Shores, CA 94065, USA

Worldwide Inquiries
Phone: +1.650.506.7000
Fax: +1.650.506.7200

ORACLE

CONNECT WITH US

blogs.oracle.com/oracle
facebook.com/oracle
twitter.com/oracle
oracle.com

Integrated Cloud Applications & Platform Services

Oracle's SPARC T7 and SPARC M7 Servers: Domaining Best Practices
October 2015
Authors: Michael Ramchand, Michele Lombardi, Henning Henningsen, Martien Ouwens, Jeff Savit, Roman Zajcew