Oracle Elite
Engineering Exchange

An Oracle White Paper
October 2013

Oracle's SPARC M5-32 and SPARC M6-32 Servers: Domaining Best Practices

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 the SPARC M6-32 into Oracle's server product line, price/performance scales linearly without the cost penalty for “Big Iron” or its enhanced features. What this means is that 16 SPARC T5-2s with 32 total CPUs are priced similarly to a SPARC M6-32 with 32 CPUs. This effectively removes the large price premium traditionally associated with this class of system, giving users 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 M5-32 and SPARC M6-32 servers deliver on all these requirements and are ideal solutions for server consolidation. With the SPARC M5-32 and SPARC M6-32 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, applications have been deployed on a single server for each application instance. 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 as excess capacity is reduced significantly.

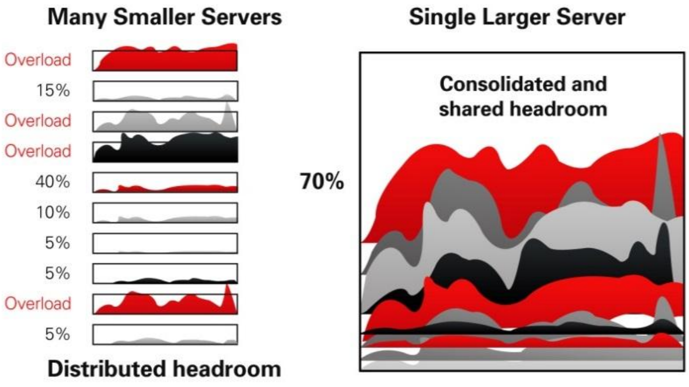


Figure 1. Consolidating and sharing headroom in large symmetric multiprocessing servers.

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 and newer servers provides many benefits beyond improved utilization. The newer servers will have more capacity, better performance, better energy and space efficiencies, improved availability features, and 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 reuse of existing applications and have effective virtualization and resource management tools. Since 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 the performance requirements of each application are met.

Consolidation on Large, Vertically Scalable SMP Servers

Large symmetric multiprocessing (SMP) servers, such as Oracle's SPARC M6-32, 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 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.

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 M6-32 server, also can be partitioned using virtualization tools to create multiple instances of the OS using subsets of the 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 sub-assemblies. 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 like Oracle Enterprise Manager Ops Center.

SPARC M5-32 and SPARC M6-32 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. On the following page, Figure 2 shows the various levels of virtualization technologies available, at no cost, on the current SPARC Enterprise M-Series servers from Oracle.

At the lower tier of the virtualization stack is the SPARC platform. The SPARC platform provides the first level of virtualization, the Dynamic Domains feature of the SPARC Enterprise M-Series (also known as physical domains or PDoms), which 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 Servers for SPARC partitions (also known as 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. Oracle VM is also recognized as an Oracle hard partition for software licensing purposes.

The third level of virtualization is Oracle Solaris Zones, the finest grained level of virtualization, and a feature of Oracle Solaris. Each zone in Oracle Solaris Zones shares a common Oracle Solaris kernel and patch level. They have significant advantages of flexibility when it comes to creation and reboot, and are extremely fast and lightweight. Each of these instances of Oracle Solaris can use Oracle Solaris Resource Manager to limit CPU or memory 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.

For the purpose of running performance-critical workloads, it is possible to configure the Oracle VM Server for SPARC domains, so that each domain is directly attached to its own PCIe slots. While this limits the total number of Oracle VM Server for SPARC domains per PDom, it provides measurable performance and isolation benefits. The Oracle SuperCluster M6-32 utilizes this type of Oracle VM Server for SPARC domains, but this type of configuration can equally be applied to both of the SPARC M5-32 and SPARC M6-32 platforms.

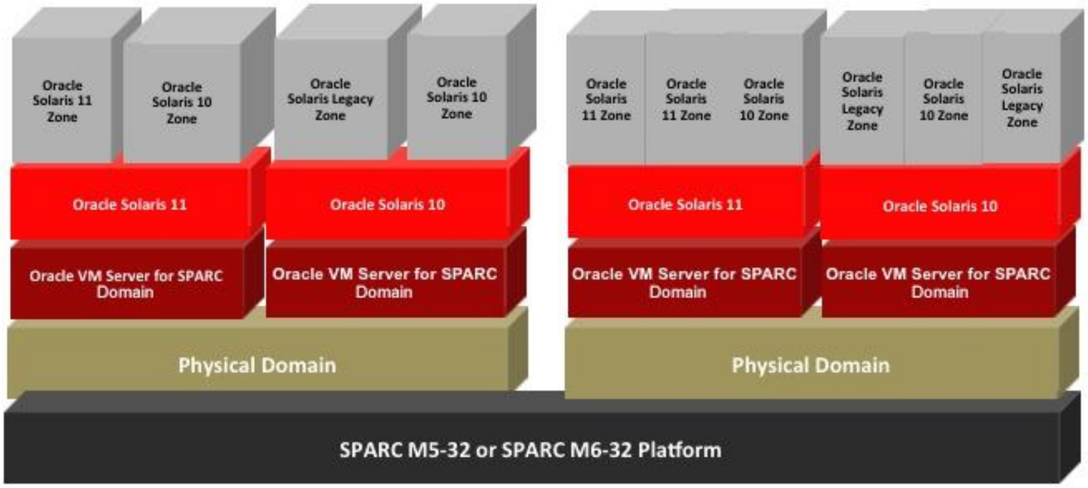


Figure 2. Virtualization technology stack on the SPARC M5-32 or SPARC M6-32 servers.

Dynamic Domains

As mentioned above, Dynamic Domains (also known as physical domains or PDoms), is a feature that enables electronically isolated partitions. PDoms make it possible to isolate multiple applications and multiple copies of the Oracle Solaris OS on a single server. The Dynamic Domains 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 Dynamic Domains feature is now in its sixth 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, PDoms can be further virtualized by running Oracle VM Server for SPARC, which allows multiple independent Oracle Solaris instances to coexist within the same physical domain.

With Dynamic Domains, software and hardware errors and failures do not propagate themselves beyond the domain in which the fault occurred. Complete fault isolation between Dynamic Domains 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 Dynamic Domains 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 provides full virtual machines that run independent instances of the operating system and are available on all of Oracle's SPARC T-Series and new SPARC M-Series based platforms. Called LDoms or Logical Domains, each operating system instance contains dedicated CPU, memory, storage, console, and cryptographic devices. LDoms are unique in the fact that many of the virtualization functions are provided natively by the underlying hardware, and that both CPU and memory are directly assigned to domains without incurring any virtualization overhead. I/O can be directly assigned either to domains, with the benefit of higher performance, or can be virtualized with 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 domains in the system, although there is an upper limit of 128 domains per server or PDoms in the case of the SPARC M5-32 or SPARC M6-32. If the root domain model is being deployed as in Oracle SuperCluster M6-32, where domains are assigned exclusive ownership of complete PCIe slots, then the number of this type of domain is actually limited by the quantity of I/O cards available in the PDom. The use of Oracle Solaris Zones within the LDom domain creates a third layer of virtualization, mitigating the effect of this reduction in Oracle VM Server for SPARC domain count.

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 a logical domain to be migrated to another PDom on the same server or a different server. While the use of live migration in Oracle VM Server for SPARC implementations is typical, the expected primary workloads on the SPARC M5-32 or SPARC M6-32 platforms are likely to require the most performant use of I/O, which will preclude the use of live migration. However, there may be a number of secondary workloads that could be placed on the SPARC M5-32 or SPARC M6-32 platform for which live migration would be ideal.

By layering Logical Domains on top of Dynamic Domains, organizations gain the flexibility to deploy multiple operating systems simultaneously onto 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, and 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 servers 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 the Oracle Solaris OS to balance workloads can reduce the processing resource requirements, resulting in fewer processors, smaller 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 may 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 OS 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 Resource Manager

Oracle Solaris Resource Manager is a group of techniques that allow the consumption of CPU, memory, and I/O resources to be allocated and shared among applications within an Oracle Solaris instance including 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, Oracle Solaris now greatly increases its offerings of network services virtualization 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.

Management of the Consolidation Technologies

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 high-end servers such as the SPARC M5-32, SPARC M6-32, and Oracle SuperCluster M6-32, 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 with Oracle Premier Support contracts.

Layered Consolidation with SPARC M5-32 and SPARC M6-32 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 choice of 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.

SPARC M6-32 and SPARC M5-32 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. Dynamic Domains: Partitioning of a server into independent isolated servers.

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. The Dynamic Domains feature provides the greatest amount of isolation, but provides 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, Dynamic Domains, Oracle VM Server for SPARC, and Oracle Solaris Zones all are considered to be hard partitions for licensing software.[1]

A Consolidation Philosophy

When faced with multiple options for consolidation, it is useful to remember the original 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.
  - By consolidating as high up the stack as possible, users will naturally reduce the total number of managed objects, as well as create as much standardization as possible.
- Maximize Workload Efficiency
  - One of the trade-offs of increased isolation is a potential increase in the virtualization overhead. Users should bear this in mind, and only create additional isolation where necessary.
  - Some workloads are quite small in comparison to the footprint of a modern OS instance. Users should try to co-locate multiple workloads per OS instance where possible.

[1] 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

We can consider the two extremes of the spectrum and the options in between:

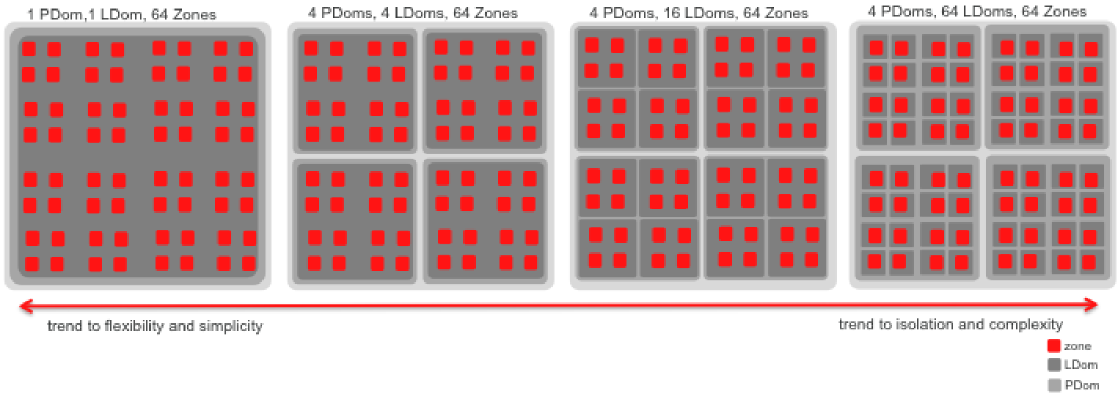

1 PDom,1 LDom, 64 Zones
4 PDoms, 4 LDoms, 64 Zones
4 PDoms, 16 LDoms, 64 Zones
4 PDoms, 64 LDoms, 64 Zones
trend to flexibility and simplicity
trend to isolation and complexity
zone
LDom
PDom

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 by having four independent Dynamic Domains (PDoms), with each PDom running 16 Oracle VM Server for SPARC logical domains (LDoms), with an OS instance per workload.

This layout allows the highest level of isolation, but comes at a cost of much higher complexity, since 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, one could choose 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[2].

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 as it creates a single failure domain, which will have a high impact for both planned and unplanned outages. One should aim to keep as far to the left of the diagram above as possible, but move 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.

[2] Sixty-four zones is not the limit; the theoretical limit is more than 8,000 zones per Oracle Solaris instance.

Dynamic System Domains (PDoms)

The SPARC M5-32 and SPARC M6-32 servers feature a balanced, highly scalable SMP design that utilizes the SPARC processors connected to memory and I/O using a high-speed, low-latency system interconnect.

Domain Configuration Units (DCUs) are the hardware building blocks of PDoms inside the system, being composed by one, two, or four CPU/memory unit (CMU) boards, one I/O unit, and one service processor proxy (SPP) board.

The system can be physically divided into up to four fault-isolated partitions called PDoms, each running independent instances of Oracle Solaris or Oracle VM Server for SPARC.

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. That is, if one uses an 8-socket PDom for a production database, you pay for 8 sockets times the number of cores/socket as a license fee.

PDoms can contain 1, 2, 3, or 4 DCUs. In the case of two or more DCUs, the scalability switch board (SSB) is used to allow memory access and cache coherency throughout the system. When the PDom contains only a single DCU, the use of the SSB is not required, and it is optionally possible to further isolate the PDom by configuring it as a “bound” PDom, and this has the effect of disabling access to the SSB.

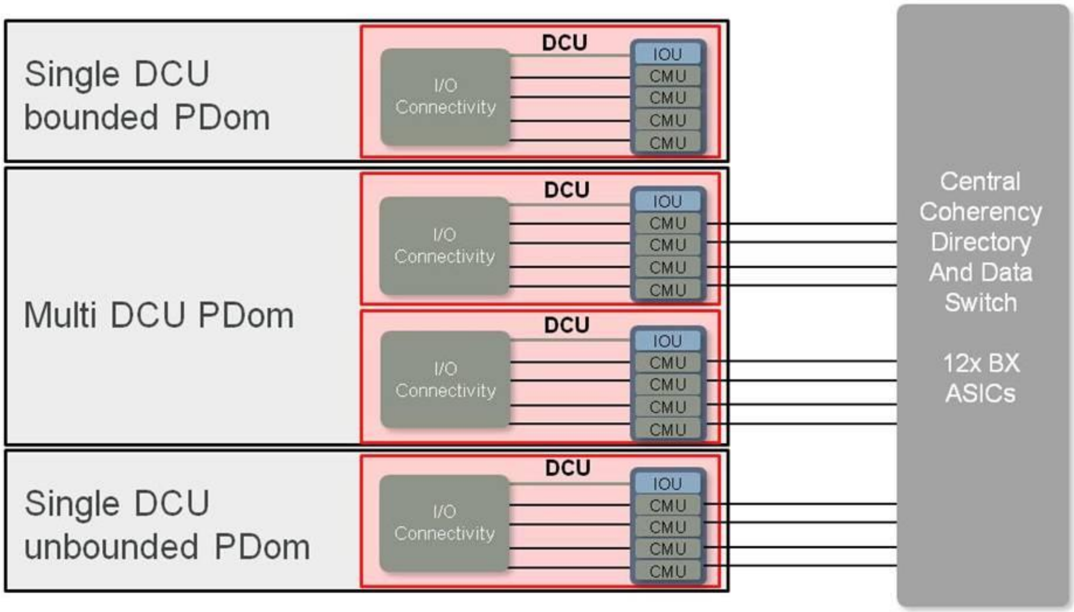


Figure 4. Single and multiple DCU physical domains.

PDom Sizing: Implications on Performance, Availability, and Flexibility

Single Domain Configuration Unit PDoms

These PDoms have the best performance and the best availability since all traffic is isolated to the 8-socket local interconnect.

All communications can be delivered inside the local coherency plane at maximum bandwidth with minimum latency, exploiting the dedicated 7x 153.6 Gb/second local coherency links per single processor.

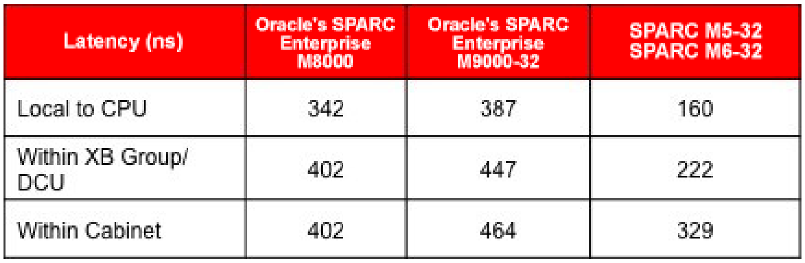
It should be noted that an 8-socket, 8 TB domain is sufficiently large to accommodate the majority of the workloads currently running on systems today, and this domain type/size is expected to be the typical configuration for SPARC M5-32 or SPARC M6-32 domain deployments.

PDoms of this type can be either bound or unbound. Bound PDoms are isolated from the scalability switch board, and should be considered the default deployment option.

Multi-Domain Configuration Unit PDoms

This configuration allows much larger compute, memory, and I/O in a single system image. When the PDom is configured in this mode, all memory caching information has to be passed to the scalability switch board (SSB), enabling inter-DCU memory traffic.

It must be considered, though, that the memory latency between two or more different DCUs is higher than within a single DCU, and this can have a performance impact. Take, for example, two independent 8-socket DCU PDoms and a single 16-socket DCU PDom running a number of workloads. While the 16-socket PDom gives better flexibility and dynamic reallocation of resources within that PDom, there is a dependence on memory and thread placement optimization of Oracle Solaris to work very well for it to deliver the same performance for those workloads spread between two 8-socket PDoms. This is because the two 8-socket PDoms are guaranteed a maximum 222 ns latency for all memory accesses, while in the case of the 16-socket PDom, while one would expect Oracle Solaris to ensure memory and thread placement, it is still possible to experience 329 ns for some memory accesses in the worst case. The actual difference in performance is a factor of how often a cross-DCU memory call needs to be made.



Latency (ns)	Oracle's SPARC Enterprise M8000	Oracle's SPARC Enterprise M9000-32	SPARC M5-32 SPARC M6-32
Local to CPU	342	387	160
Within XB Group/ DCU	402	447	222
Within Cabinet	402	464	329

If LDoms are used, the latency effect can be minimized (almost reduced to zero) by defining the LDom resources so that they do not cross a DCU boundary.

From a reliability and availability point of view, the difference between a single DCU domain and a multiple DCU domain from a hardware perspective is the reliance on the SSB. It should be noted that the mean time between failure (MTBF) of the SSB is higher than the expected lifecycle of the server, and the possibility of an unrecoverable fault leading to a reboot, is very rare.

From a serviceability point of view, having larger domains means having to deal with relocating or interrupting more workloads when planned maintenance is required on that domain.

Oracle SuperCluster M6-32 PDoms

The Oracle SuperCluster M6-32 is available as an engineered system. Oracle SuperCluster M6-32 has a predefined set of fixed hardware configurations, with the ability to support a number of specific PDom configurations. It allows the ability to create either 2 or 4 PDoms, and these PDoms can be created from either single or dual DCUs. This is outlined in more detail in Appendix: Oracle SuperCluster M6-32 Configuration Rules.

Oracle VM Server for SPARC

An Oracle VM Server for SPARC domain (also referred to as a logical domain or LDom) is a virtual machine comprised 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 only permitted to observe and interact 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. The following diagram shows the Hypervisor supporting four logical domains. It also shows the following layers that make up the logical domains' functionality:

- User/services, or applications
- Kernel, or operating systems
- Firmware, or Hypervisor
- Hardware, including CPU, memory, and I/O

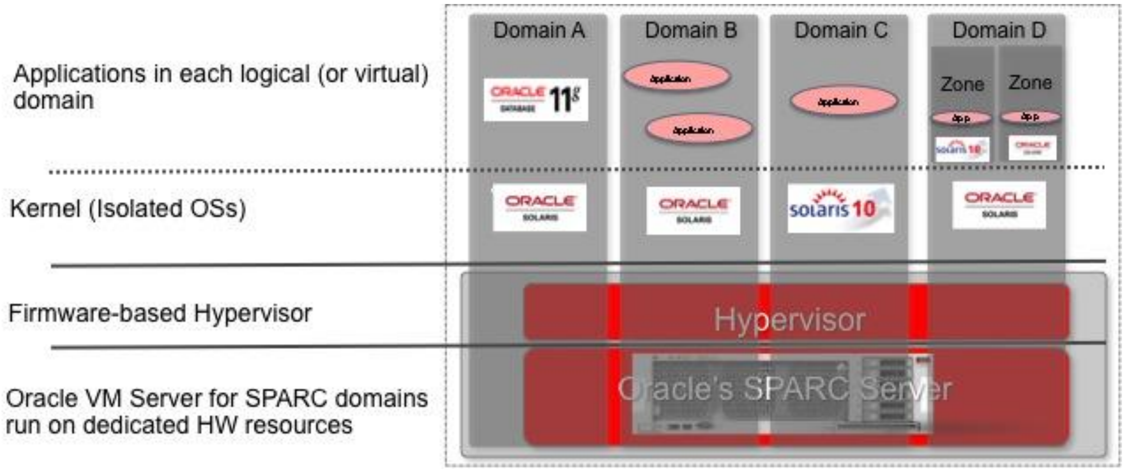


Figure 5. Oracle VM Server for SPARC virtualization.

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 support of multiple operating systems simultaneously, 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:

- 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 each other without requiring users to perform a power cycle of the server.

LDoms Inside Dynamic Domains

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 using Dynamic Domains alone.

Given that the SPARC M5-32 and SPARC M6-32 servers support a maximum of four Dynamic Domains, it is expected that many of the deployments will make use of the Oracle VM Server for SPARC technology to create additional workload isolation at the logical domain level if required.

The Control Domain

When PDoms are first created, what is known as the primary or control domain is created. On a SPARC M5-32 and SPARC M6-32 system, this initial domain MUST run Oracle Solaris 11.1 or later. This control domain initially owns all the hardware available in the PDom, including all CPUs, all memory, and all I/O resources.

If only a single domain running Oracle Solaris 11.1 is required, then there is no further work to do, as 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 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 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, 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 only be 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, as it must own physical I/O resources in order to virtualize them for guest domains. In most cases, these service domains have PCIe root complexes assigned to them, and could be termed a root domain in this case.

Guest domain—a domain whose devices are all virtual rather than physical: 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 is used to run applications within the domain, rather than provide services like the service domain above. Physically there is no difference between these service domains and guest root domains other than their usage, and they often will be simply referred to as root domains.

The configuration of Oracle VM Server for SPARC within Dynamic Domains on SPARC M5-32 and SPARC M6-32 servers is no different from the way it 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 which are 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 Domain	With guest root domains, root complexes are directly assigned to the guest domains, and they have direct ownership of their I/O.	Simplest model as there is no need to create multiple virtual disk and network services, but also the least flexible, as one can have only as many domains as there are root complexes. However, these guests run at 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 http://www.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 as the expected workloads on SPARC M5-32 and SPARC M6-32 based systems are likely to be a good fit for this particular operating model, and more specifically, the Oracle SuperCluster M6-32 is configured in this way.

A guest root domain is the concept of 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 in the Oracle VM Server for SPARC technology, and in particular, 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 (i.e., 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 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 (i.e., NIC or HBA) impacts only that domain. This in contrast to the service domain, direct I/O, or SR-IOV models, where all domains that share those components are impacted by their failure.
- Improved security: There are fewer shared components or management points.

A physical domain is composed of one to four DCUs (DCU0, DCU1, DCU2, DCU3). Each DCU has an associated I/O unit (IOU). Each IOU supports up to 16 PCIe slots, eight 10 GbE ports, and eight HDD/SSDs, on four EMS modules.

In total, 64 root complexes are available in both SPARC M5-32 and SPARC M6-32 servers, 16 root complexes per DCU, and up to four DCUs. These root complexes are named pci_0 to pci_63. Each root complex is associated with one PCIe slot, and four root complexes per DCU will have access to an EMS module with access to local boot and 10 GbE ports.

A typical configuration is 16 guest root domains with four PCIe slots per domain, each with access to local disk for booting. Oracle SuperCluster M6-32 has specific domain configuration rules which define optimized sizes and layouts for domains within a PDom, and these are outlined in more detail in Appendix 1. These rules could also be applied to SPARC M5-32 and SPARC M6-32 domain layouts, as it represents Oracle preferred practice for the layout of guest root domains on this platform.

It is important to note that for cases that do not involve Oracle SuperCluster M6-32, guest root domains are not the only option. For the SPARC M5-32 and SPARC M6-32 servers, a solution that comprises some systems with service domains as well as some systems with guest root domains may be appropriate. In fact, the same Dynamic Domains could consist of two root domains running applications, and two service domains providing services to a number of fully virtualized guest domains, or different Dynamic Domains could be configured completely differently.

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 provides OS-level virtualization, which gives the appearance of multiple OS instances rather than multiple physical machines. Oracle Solaris Zones 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, Oracle Solaris Zones configured as capped or dedicated CPUs are considered to be hard partitions.

The nature of OS virtualization means that Oracle Solaris Zones provides very low-overhead, low-latency environments. This makes it possible to create hundreds, 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. Oracle Solaris Zones enables 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 nets, 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 also allows the ability to run older Oracle Solaris versions within zones. These are 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.

Oracle Solaris Zones are also integrated with Oracle Solaris DTrace, the 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 above 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.

Some typical examples of how these technologies could be combined are provided below.

Configuration Examples

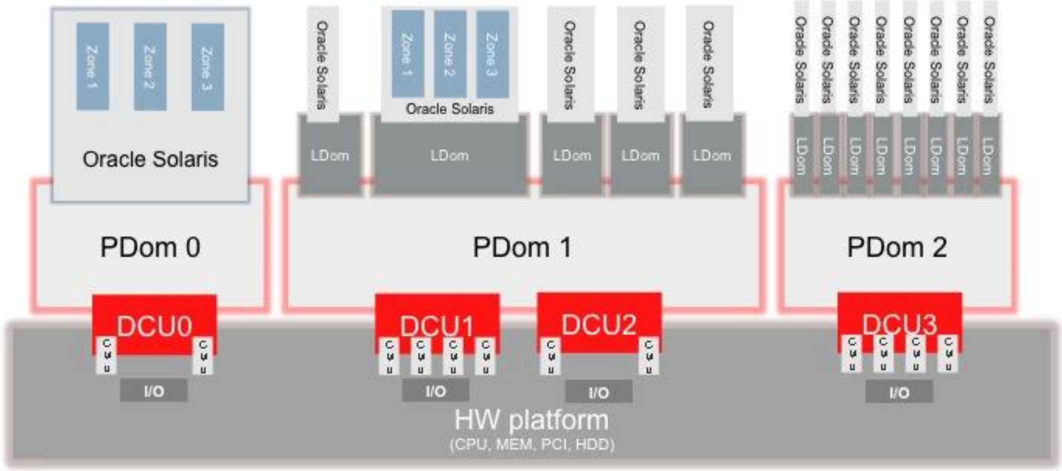


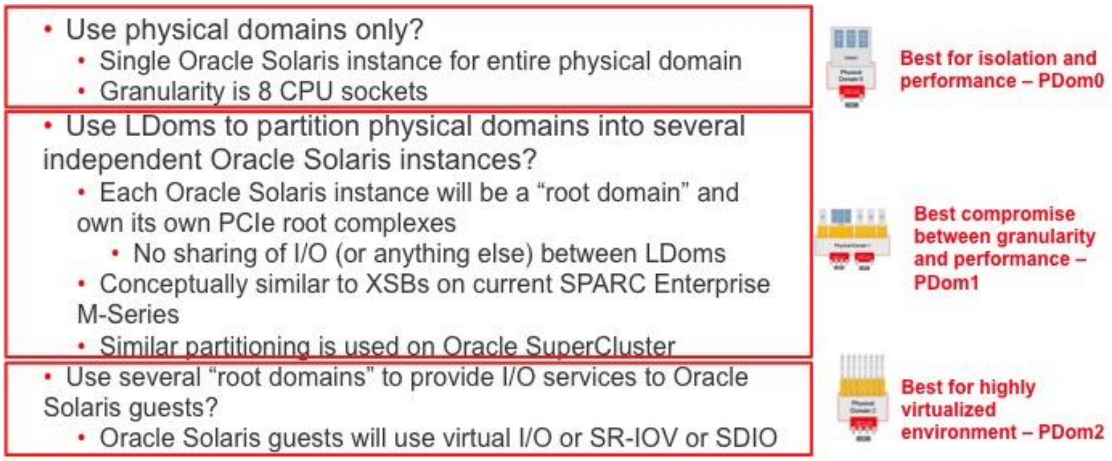
Figure 6. Multiple physical domain configuration options.

In the above example, PDom 0 has Oracle Solaris 11 running directly on bare metal, giving the maximum performance with no virtualization overhead. Zones can be used to isolate workloads without impacting performance. There will be one single Oracle Solaris OS image. This implementation allows consolidation at the highest level of the stack.

In PDom 1, LDoms are introduced as a server virtualization layer. To maintain the best compromise between virtualization and performance, LDoms are configured as root domains, so each LDom has direct access to its own PCIe slots. With LDoms, workloads can be further isolated down to the OS level. Each LDom requires its own OS instance, so manageability is a bit more complex than in the previous scenario. Inside each LDom, Oracle Solaris Zones can be used to further isolate workloads. This implementation creates hardware isolated domains, and workloads can be consolidated on the same hardware without sacrificing any performance.

In PDom 2, LDoms are used with virtual I/O, in order to achieve the best flexibility. There will be two or more I/O service domains that will export the virtual I/O to all the other LDoms. With such an arrangement, an LDom per workload or environment can be created. This implementation is a fully virtualized environment, with large numbers of independent, isolated domains with unique OS instances, at the cost of lower performance due to the virtualization overhead. A similar level of virtualization granularity could be achieved using fewer LDoms and consolidating higher in the stack using multiple zones created by Oracle Solaris Zones.

Of course the three examples are not mutually exclusive, and any combination among the three scenarios is possible.



- Use physical domains only?
  - Single Oracle Solaris instance for entire physical domain
  - Granularity is 8 CPU sockets

Best for isolation and performance – PDom0

- Use LDoms to partition physical domains into several independent Oracle Solaris instances?
  - Each Oracle Solaris instance will be a "root domain" and own its own PCIe root complexes
    - No sharing of I/O (or anything else) between LDoms
  - Conceptually similar to XSBs on current SPARC Enterprise M-Series
  - Similar partitioning is used on Oracle SuperCluster

Best compromise between granularity and performance – PDom1

- Use several "root domains" to provide I/O services to Oracle Solaris guests?
  - Oracle Solaris guests will use virtual I/O or SR-IOV or SDIO

Best for highly virtualized environment – PDom2

Figure 7. Suggested PDom configuration use cases.

While virtualized I/O provides the highest degree of flexibility and dynamic allocation and usage of I/O resources, in some cases this may create higher latency or lower throughout than native I/O. In cases where native I/O characteristics are important, the use of root domains is appropriate. The possible number of root domains is defined by the number of root complexes, and further by the number of PCIe slots available for I/O cards. In these situations the use of Oracle Virtual Networking for SPARC M5-32 and SPARC M6-32 servers to aggregate network and FC traffic onto a single InfiniBand card may be a useful solution by halving the number of PCIe slots required to provide both SAN and Ethernet network connectivity.

Conclusion

At this point it should be very apparent that the SPARC M5-32 and SPARC M6-32 servers combine the dynamic system domaining features of the previous generation SPARC Enterprise M-Series with the Oracle VM Server for SPARC (LDom) features of the current SPARC T-Series systems. This delivers a layered virtualization solution with Oracle Solaris Zones that can meet a wide and varied set of requirements. The sheer size of the SPARC M5-32 and SPARC M6-32 servers with their 32 TB memory footprint provides the most advanced capabilities spread over thousands of active threads. The SPARC M5-32 and SPARC M6-32 servers offer both availability and serviceability, designed from the very core of the processor up, delivering new levels of performance, availability, and ease of use to enterprise level applications. The sophisticated resource control provided by Dynamic Domains, Oracle VM Server for SPARC, and Oracle Solaris

Zones further increases the value of this server by helping enterprises to optimize the use of their hardware assets. By deploying fast, scalable SPARC M6-32 servers from Oracle, organizations will gain extraordinary per-thread performance and flexibility, a strategic asset in the quest to gain a competitive business advantage.

It is expected that the decision to procure a SPARC M6-32 is based on the need to run highly intensive workloads that are not appropriate for the smaller SPARC-based systems for either availability or vertical performance-related reasons. In many cases there may be single or multiple large workloads that require a large PDom, but there are also likely to be a large number of additional workloads that may be optimally placed on smaller PDoms on multiple LDoms. This flexibility makes the SPARC M6-32 an ideal consolidation platform.

There are a large number of use cases to consider for the deployment of workloads on a SPARC M6-32 class machine. In general, if all the workloads under consideration fit easily into an 8-socket building block, then creating four 8-socket bound PDoms is a straightforward choice—because it delivers the best performance and isolation but with large enough domains to be able to flexibly allocate resource within it, but not too large as to make serviceability a challenge.

If the reason that the SPARC M6-32 is purchased is to run large single OS instance images requiring more than 8 sockets and 8 TB of RAM, then it is necessary to create multi-DCU PDoms sized appropriately for the workloads.

If additional granularity of workload is required, this can be provided by either creating zones directly on top of Dynamic Domains, or by inserting Oracle VM Server for SPARC domains for additional isolation.

When using Oracle VM Server for SPARC domains, 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, zones created with Oracle Solaris Zones can still be layered on top of the Oracle VM Server for SPARC domains.

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

Best Practices for Availability

This paper has not discussed high availability (HA) in any great detail. However, it is an extremely important facet to defining the architecture of a SPARC M5-32 or SPARC M6-32 deployment.

Best practices for high availability such as clustering across compute nodes and remote replication for disaster recovery should always be applied to the degree the business requirements warrant. For example, HA should not be implemented with both nodes of the cluster located within the same PDom. However, multiple tiers (i.e., web, app, 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 Maximum Availability Architecture and Optimized Solutions sections of the Oracle Technology Network (OTN).

Summary

In the simplest possible terms, the following high-level guidelines should be appropriate for SPARC M5-32 and SPARC M6-32 based deployments:

- Use 8-socket Dynamic Domains by default, unless there is a specific requirement for a larger than 8-socket domain. Up to four 8-socket domains are possible.
- Use Oracle VM Server for SPARC domains within Dynamic Domains if further isolation is required. Use the root domain model for small numbers of large domains, or the virtualized I/O guest model for large numbers of smaller domains.
- 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 the Elite Engineering Exchange

The Elite Engineering Exchange (EEE) is a cross-functional global organization consisting of Oracle's elite sales consultants (SCs) and systems engineers (Product Engineering). The EEE connects Product Engineering directly to the top experts in the field through joint collaboration, bidirectional communication of customer and market trends and deep insight in the technology directions of future generation products. The EEE brings real-world customer experiences directly to Engineering and Engineering technical details and insights to SCs; both enable better solutions to meet the changing demands of Oracle's customers.

Appendix: Oracle SuperCluster M6-32 Configuration Rules

The Oracle SuperCluster M6-32 is an engineered system, which combines the SPARC M6-32 server with an Oracle Exadata Storage Expansion Half Rack consisting of nine Exadata Storage Servers and three Infiniband switches. This is supplemented by an Oracle ZFS Storage Appliance, installed within the rack.

A detailed explanation of all the nuances of the SuperCluster configuration is out of the scope of this white paper, but this section aims to highlight the domain (PDom and LDom) configuration options for the Oracle SuperCluster M6-32.

As an engineered system, the Oracle SuperCluster M6-32 is designed to provide a limited number of fixed configurations, which represent best practices in terms of performance, scalability, and availability. In other words, while the SPARC M6-32 server can be configured in a wide variety of different ways, the SuperCluster variant is restricted to a smaller number of possibilities. This allows Oracle to ensure consistency of more stringently tested configurations throughout the install base, as well as quick reproduction and analysis of issues should they occur.

Oracle SuperCluster M6-32 Domain Building Blocks

The SuperCluster is composed of two or four PDoms, which are created by combining fixed-configuration DCUs in a number of predetermined combinations. Oracle VM Server for SPARC technology is then used to create up to four root domain LDoms per PDom.

Each DCU consists of a number of fixed components:

- 4 Base I/O cards, each with 2 x 10 GbE ports
- 8 HDDs (900 GB each)
- 4 dual-port InfiniBand HCAs
- 1 quad-port GbE NIC
- 11 available PCIe slots

The variable component of the DCU is the number of CMUs within the DCU and can be one of:

- 2 CMUs (with four SPARC M6 processors), or
- 4 CMUs (with eight SPARC M6 processors)

Finally, the DCUs must be fully populated with 32 dual inline memory modules (DIMMs) per SPARC M6 processor, which can be either 16 or 32 GB capacity. This equates to either 512 GB RAM per processor or 1 TB RAM per processor. It should be noted that the DIMM capacity choice applies to all DCUs within an Oracle SuperCluster M6-32.

PDom Configuration: Base or Extended

When there are only two DCUs present, the only option is for the system to be composed of two single DCU PDoms. Single-DCU PDoms are considered the normal configuration approach and are referred to as “base configuration” PDoms.

However, when there are four DCUs, there is an option to configure either two PDoms consisting of two DCUs each, or four single DCU PDoms. The rationale for an “extended configuration” PDom is simple: Either there is a requirement for an LDom with more than 8 sockets of CPU assigned to it, or the LDoms require more I/O capability than a single DCU will provide.

In most cases, it is expected that single-DCU PDoms will be used.

Finally, it is possible to split the DCUs between two SPARC M6-32 racks, rather than having them coexist within the same rack. This option is provided solely for cases where extreme availability and additional isolation is required, as it provides a slightly improved RAS over the single-rack solution.

LDom Configuration: One to Four LDoms

The domain configuration rules, which govern how the LDoms are allocated to PDoms, define four configurations for each of the normal or extended PDom layouts. In all cases, the LDoms are configured as root domains, with each domain given exclusive ownership of one or more root complexes. This ensures bare metal performance for each of the LDoms.

For the normal PDoms, the four configurations are as follows:

One LDom with ALL resources allocated to it (one large)

Two LDoms with I/O resources split evenly between them (two medium)

Three LDoms with one large and two small LDoms (one medium, two small)

Four LDoms with I/O resources split evenly amongst them (four small)

For the extended PDoms, the assumption is made that one of the LDoms will be huge. The configurations are identical to above, except that the first LDom in the above configuration will also have the entire first DCU’s I/O allocated to it.

Oracle SuperCluster M6-32 Conclusion

As can be seen, the Oracle SuperCluster M6-32 configuration rules are an excellent example of the best practice guidelines for domain configurations using the root domain model, and could also be used as a blueprint for configurations that do not involve Oracle SuperCluster M6-32 or M5-32 configurations.

Oracle is committed to developing practices and products that help protect the environment

Oracle's SPARC M5-32 and M6-32 Servers:
Domaining Best Practices

October 2013
Authors: Michael Ramchand, Tom Atwood, Michele Lombardi, Henning Henningsen, Martien Ouwens, Ray Urciuoli, Roman Zajcew

Oracle Corporation
World Headquarters
500 Oracle Parkway
Redwood Shores, CA 94065
U.S.A.

Worldwide Inquiries:
Phone: +1.650.506.7000
Fax: +1.650.506.7200

oracle.com

This document is provided for information purposes only, and the contents hereof are subject to change without notice. This document is not warranted to be error-free, nor subject to any other warranties or conditions, whether expressed orally or implied in law, including implied warranties and conditions of merchantability or fitness for a particular purpose. We specifically disclaim any liability with respect to this document, and no contractual obligations are formed either directly or indirectly by this document. This document may not be reproduced or transmitted in any form or by any means, electronic or mechanical, for any purpose, without our prior written permission.

Hardware and Software, Engineered to Work Together