

Sparks and White Noise

The SPARC Enterprise T5140/T5240 and Virtualization

By:
Octave Orgeron

<http://unixconsole.blogspot.com>



Introduction:

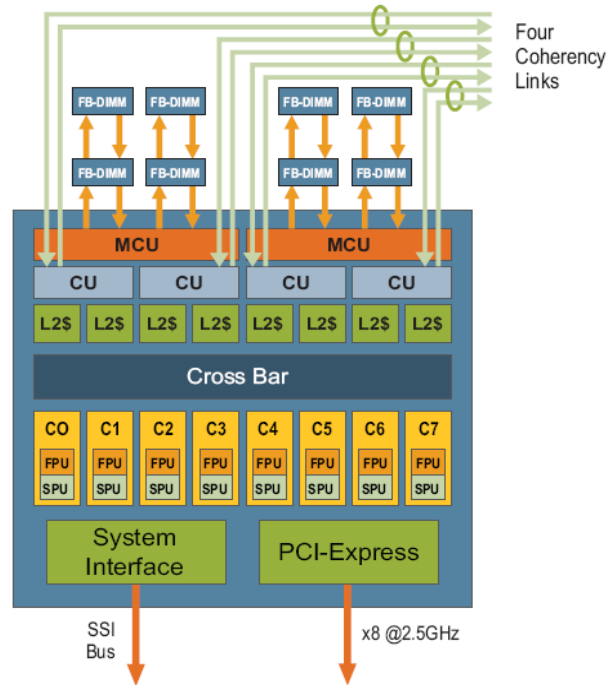
The new Sun SPARC Enterprise T5140/T5240 servers are the next major leap in CMT technology. These are the first multi-socket CMT servers on the market. In this product review, I will introduce you to the features and benefits of these two amazing servers with emphasis on how they can be applied to virtualization. This will be beneficial to architects and engineers looking to consolidate legacy servers and increase utilization.

Hardware:

Feature:	T5140:	T5240:
Chassis	1 x RU	2 x RU
Processor	2 x UltraSPARC-T2 Plus (4, 6, or 8 core @ 1.2Ghz)	2 x UltraSPARC-T2 Plus (4, 6, or 8 core @ 1.2Ghz or 8 core @ 1.4Ghz)
Memory	16 x FB-DIMM Slots. Supports up to 64GB (4GB FB-DIMMS)	32 x FB-DIMM Slots. Supports up to 128GB (4GB FB-DIMMS)
Disks	Up to 4 x 73GB or 146GB SAS HDDs (Hot-Swappable). LSI RAID controller.	Upto 8 or 16 SAS slots. Supports 73GB or 146GB SAS HDDs (Hot-Swappable). LSI RAID controller.
Networking	4 x 1GbE, 2 x 10GbE XAUI	4 x 1GbE, 2 x 10GbE XAUI
PCI	3 x PCI-E (x8 lane) or 1 x PCI-E (x8 lane) and 2 x XAUI 10GbE.	6 x PCI-E (x8 lane) or 4 x PCI-E (x8 lane) and 2 x XAUI 10GbE.
Removable Media	1 x DVD +/- RW Slimline	1 x DVD +/- RW Slimline
I/O	4 x USB 2.0, 1 x DB9 Serial	4 x USB 2.0, 1 x DB9 Serial
Management	ILOM with Network and Serial Ports	ILOM with Network and Serial Ports
PSU	2 x 720W AC, N+1 Redundancy, Hot-Swappable	2 x 1100W AC, N+1 Redundancy, Hot-Swappable

Both of these servers are based off of the same system board. This helps to reduce complexity and manufacturing costs. The chassis for both models are identical to the previous generation T5120/T5220 servers, which can make them difficult to identify in a rack without looking at the label closely. Some of the key differences between the previous generation and these new servers are:

- 2 x UltraSPARC-T2 Plus Processors
- Double the amount of RAM on the 2U version
- Double the amount of SAS drives on the 2U version
- All PCI-E slots are 8 lane slots.
- Power capacity to handle additional hardware.



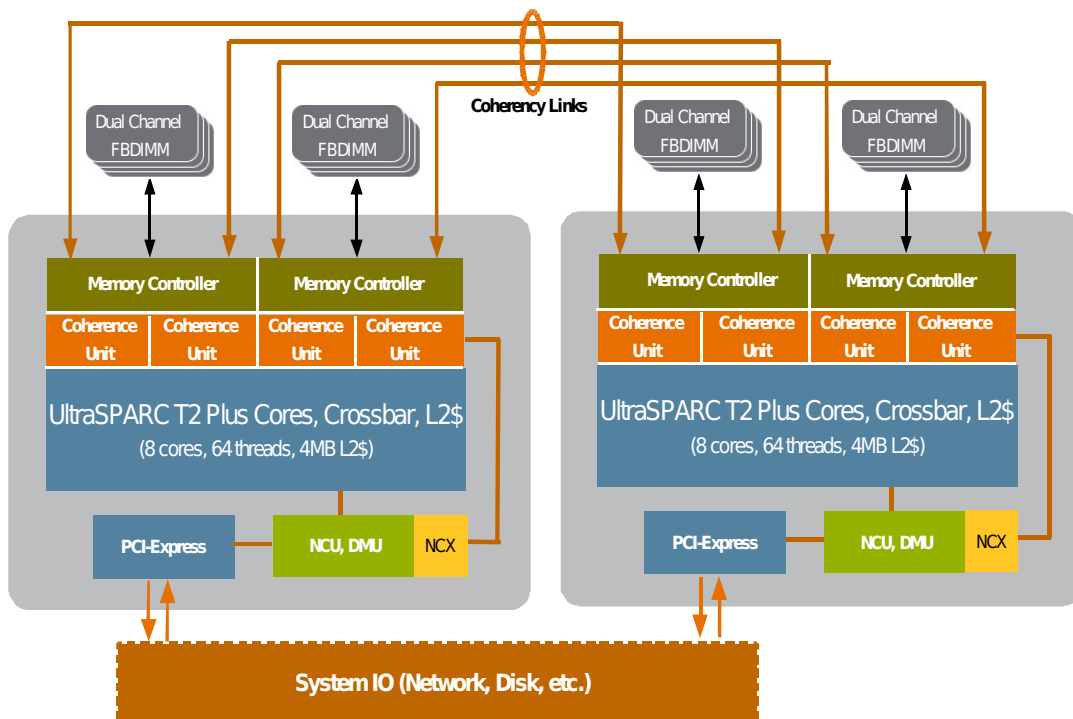
The UltraSPARC-T2 Plus processor is the third generation of the CMT processor design from Sun. The processor still has the capacity for 8 cores that 8 threads, a floating point unit, and a cryptographic unit each. The processor even has the same 4MBs of level 2 cache and an integrated PCI-E controller. The major difference is the SMP capability. This allows an UltraSPARC-T2 Plus processor to be connected into a multi-socket configuration, thusly allowing the CMT design to scale beyond a single piece of silicon to 16 cores and 128 threads. This ability is achieved through the use of built-in coherency logic and communication links that ensure coherency between CMT processors.

In traditional processor designs, SMP capabilities are handled via external ASICs to coordinate the coherency between the processors. Such traditional designs came at a cost of extra latency, power, and hot-spots on the system board. By moving this capability into the processor itself, the costs, latency, and power consumption can be significantly reduced.

The coherency units (CU) sit between the level 2 cache and the memory controller units (MCU) that manage the FB-DIMMs. The CUs examine cache and memory requests. If the requested data is located on the other CMT processor, it can be transmitted over the coherency links to the requesting thread. However, this ability did come at a cost as it requires specialized circuitry to fit onto the CPU die to keep latency levels low. The following items were sacrificed to supply enough room:

- 10GbE Controller
- 2 x Memory Controller Units

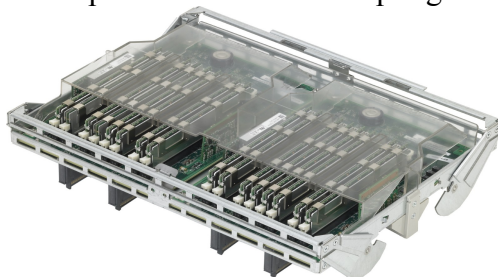
The removal of these components is due to the limited amount of space available on the CPU die. Hopefully the next generation CMT processor will be shrunk down to 32nm to enable these components to be integrated again.



The coherency links are based on serializer/deserializer or SERDES high speed links. These were already being used in the UltraSPARC-T2 between the MCUs and the FB-DIMMs. By removing two of the MCUs, there are four SERDES channels for inter-socket communications. Each channel has 14 transmit and 14 receive links. Each of these links run at 4Gbps for the 1.2Ghz processors and 4.8Gbps for the 1.4Ghz processors. For the 1.4Ghz processors, the aggregate bandwidth is 8.4GB/s per channel for a total of 33.6GB/s bandwidth across all four channels. This creates an extremely wide and fast low-latency transport for processor coherency.

The latency for memory access is determined by the locality of the memory itself. If the requested data is stored on an FB-DIMM behind an MCU on the same CMT processor, the access is local. If the requested data is stored on an FB-DIMM behind an MCU on the other CMT processor, the access is remote. The differences between the two are 15ns vs. 72ns, respectively.

With the bandwidth and speed of the coherency links taken into consideration, the maximum theoretical memory bandwidth is ~32GB/s per processor. This means that the total maximum memory bandwidth is ~64GB/s for the T5140/T5240. This bandwidth though is impacted by the number of populated FB-DIMM slots. The more slots that are populated, the greater the bandwidth. One of the advantages of the T5240 is the ability to install an FB-DIMM riser board to double the amount of memory to 128GB's. This should be kept in mind when attempting to leverage the memory capabilities.



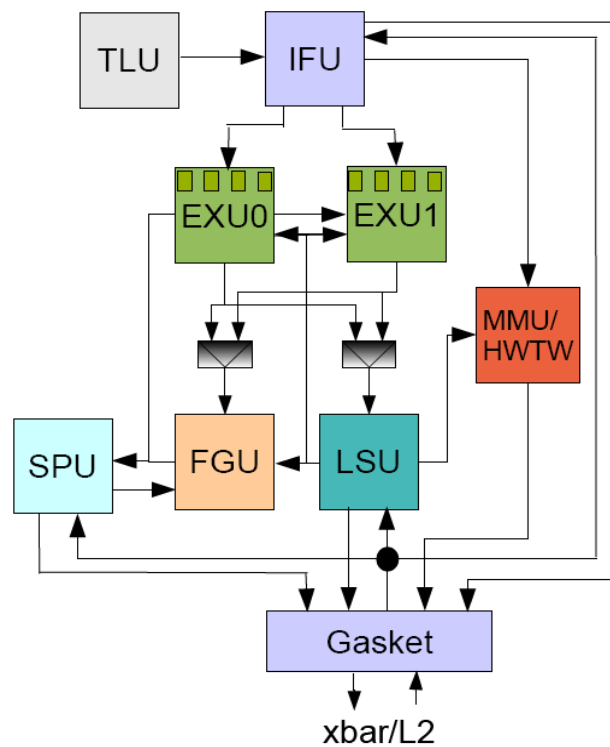
At the heart of the UltraSPARC-T2 Plus processor is the integrated cache crossbar switch (CCX) that provides about 300GB/s of bandwidth for all of the processor components to communicate across. This is an extremely large amount of bandwidth that is squeezed into single chip!

The processor has a total of 4MBs of L2 cache that is broken up into 8 banks. The L2 cache banks are 16 way associative and are shared with all of the cores.

Each UltraSPARC-T2 Plus core implements the SPARC V9 instruction set and includes the following components:

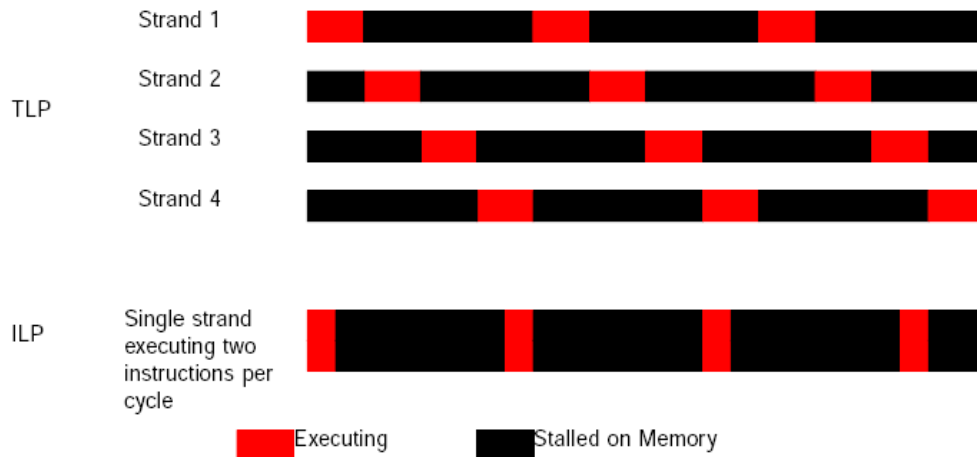
- 2 x Execution Units (EXU), with 4 threads each.
- 1 x Floating Point and Graphics Unit (FGU).
- 1 x Stream Processing Unit (SPU) for cryptographic acceleration.
- 1 x Memory Management Unit (MMU).
- 1 x Load Store Unit (LSU)
- 1 x Instruction Fetch Unit (IFU)
- 1 x Trap Logic Unit (TLU)

Below is a block diagram showing how these components interface:



The processor supports full hardware multi-threading with each core managing 8 threads. Each core contains two EXUs that manage 4 threads each. These threads run simultaneously, but are at different phases of execution. At most only two threads are executing in the core per cycle, while the other threads are waiting on cache or their slot in the next execution cycle. When a thread encounters a cache miss, it is marked unavailable and goes into a wait state until the requested data is available in

the cache. During that time, the other threads will continue to be available for execution. Once the data is available in cache, it is marked as available and ready for execution. This ability to quickly switch between threads enables the processor to continue running applications and overcome the cache latency limitations of instruction level parallelism (ILP) processor designs. The below diagram demonstrates the differences between thread level parallelism (TLP) and instruction level parallelism (ILP) at the EXU level:



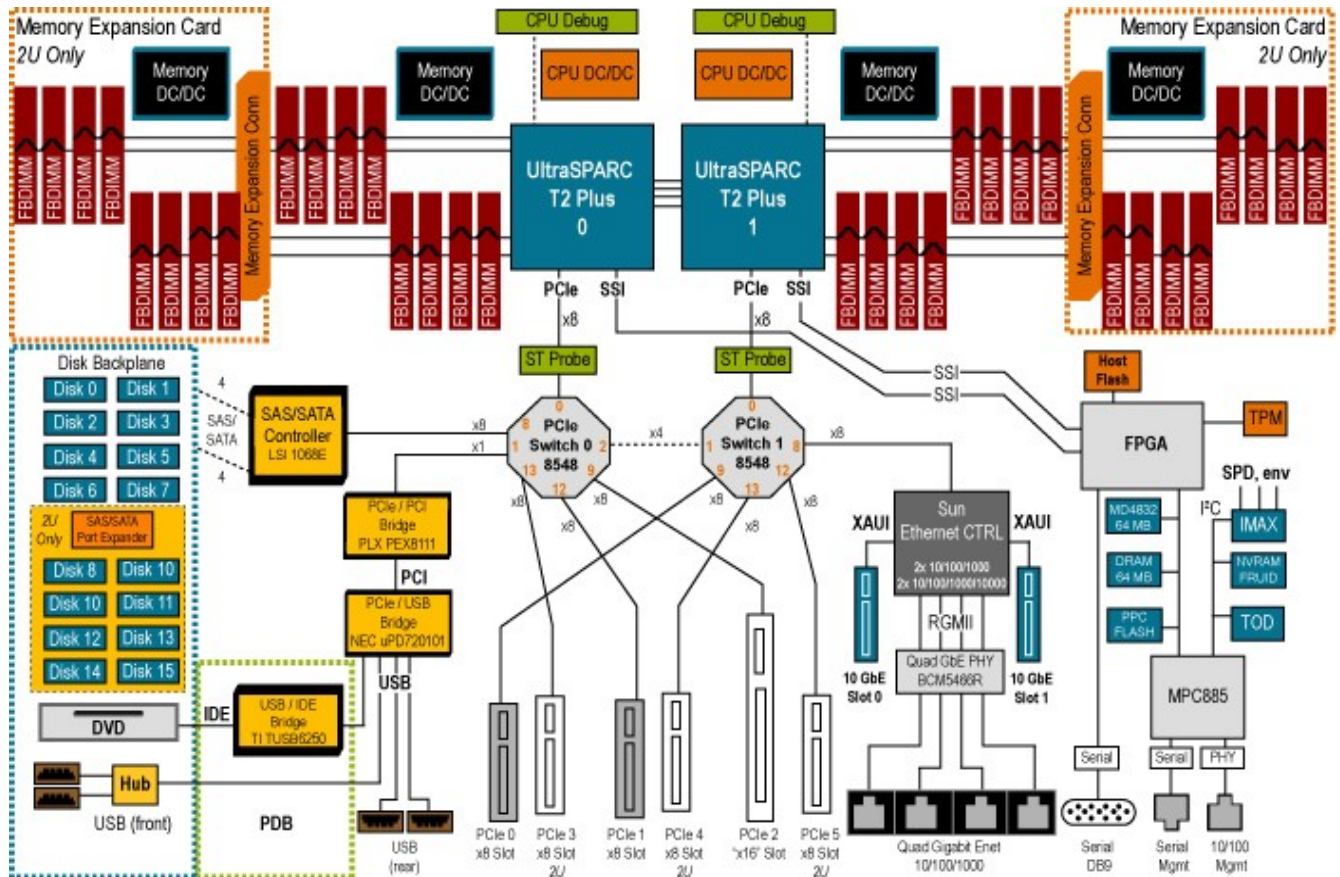
Inside each core, there is a floating point and graphics unit (FGU) that allows the UltraSPARC-T2 Plus to overcome the limitations of the UltraSPARC-T1 processor, which only had a single FPU for the entire processor. The difference with the UltraSPARC-T2 Plus implementation is that the FGU implements a full floating point unit and VIS 2.0 multimedia extensions. This enables the UltraSPARC-T2 Plus to take on HPC workloads that can leverage multi-threaded floating point calculations. Through the VIS multimedia extensions, graphics processing and streaming can be accelerated.

In conjunction, each core includes a stream processing unit (SPU) for handling cryptographic functions. The SPU is capable of natively handling the following encryption algorithms:

- DES and 3DES
- AES
- RC4
- SHA1 and SHA256
- MD5
- RSA up to 2048 key length.
- ECC (Elliptic Curve Cryptography)
- CRC32

The SPUs can be dynamically accessed by applications through the Solaris cryptographic framework, which is a set of libraries and kernel hooks. This enables applications to leverage this built in cryptographic acceleration. This can greatly increase the security performance of web applications, network communications, file systems, databases, and other applications that require encryption. This also saves businesses from having to buy costly cryptographic accelerator cards since the functionality is built right into the processor. One has to also consider that the processor has 8 SPUs available for such applications and services.

Another interesting component that is integrated into the UltraSPARC-T2 Plus processor is the PCI-E controller. Each processor has its own PCI-E controller with 8 lanes running at 2.5Ghz. This provides 4GB/s of bandwidth bidirectionally for each processor. These lanes are connected to PCI-E switches on the system board. All of the I/O components are thusly connected to these PCI-E switches:



The T5140/T5240 have two PCI-E switches that all of the PCI-E I/O connects to. The PCI-E switch on the left is known as switch 0 and the PCI-E switch on the right is known as switch 1. Switch 0 has the SAS controller for the internal storage, a PCI-E to PCI bridge for the USB and DVD peripherals, and three PCI-E slots. Switch 1 has the onboard 1/10GbE controller and three PCI-E slots. This creates a well balanced configuration as both switches have three PCI-E slots and a major I/O controller (SAS or 1/10GbE). Below is a reference chart for understanding which components are managed by each PCI-E switch:

Device Type:	PCI-E Switch 0:	PCI-E Switch 1:
PCI-E Slot	T5140 = 1 and 2. T5240 = 1, 2, and 3.	T5140 = 0. T5240 = 0, 4, and 5
Storage Controller	On-Board LSI SAS Controller	N/A
Network Controller	N/A	Sun Neptune 1/10GbE. 2 x XAUI 10GbE Slots, 4 x 1GbE Ports
Misc. Controllers	PCI-E to PCI Bridge, PCI-E to USB Bridge	N/A
Removable Media	DVD+/-RW Drive, USB Ports 0-3	N/A

All of the PCI-E slots have x8 lanes, where each lane is capable of 250MB/s of data transfer bandwidth. This means that each slot is capable of ~2GB/s of bandwidth. This supplies ample capacity for networking and storage adapters. It is important to note that slot 2 is mechanically a x16 lane slot to allow the installation of such cards, but is electrically an x8 lane slot. This allows a x16 lane PCI-E cards to be physically installed, but they will only function at half speed. PCI-E slots 0 and 1 also share the same physical space with the 10GbE XAUI slots. If a 10GbE XAUI transceiver card is installed in one of these slots, it can not be used for PCI-E adapters.

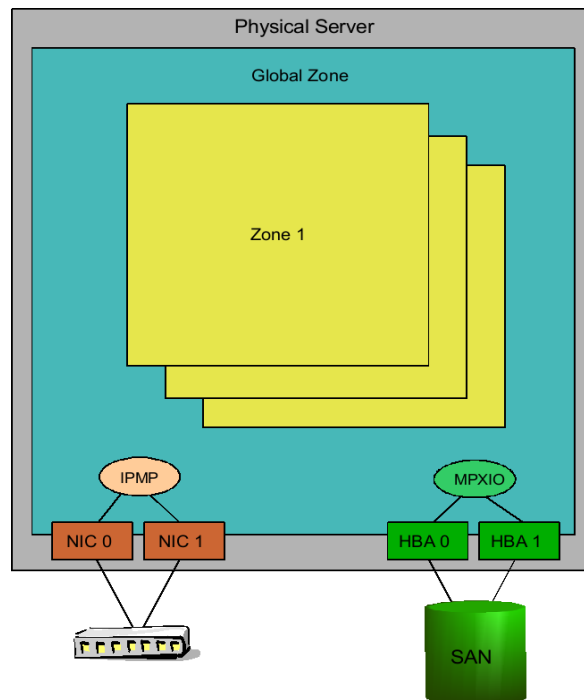
Since the 10GbE capability is now on-board and not integrated directly into the processors, there are some trade-offs for using the Sun Neptune ASIC. The on-board 10GbE ASIC is responsible for not only supplying 10GbE capabilities, it also responsible for supplying the 1GbE ports. If a XAUI transceiver is installed, an on-board 1GbE port will be disabled. If XAUI slot 0 is populated, then the on-board NIC port 0 will be disabled. If XAUI slot 1 is populated, then on-board NIC port 1 is disabled. This is an interesting compromise that may complicate server planning. From a performance and redundancy stand-point, the Sun Neptune ASIC is on a single PCI-E switch, which may create a bottleneck. It is important to keep this in mind when enabling the 10GbE feature. Ideally in the future, the 10GbE capability will be moved back into the processors and the 1GbE capability will be handled by dual ASICS on different PCI-E switches.

The integration of the 10GbE Neptune ASIC enables the platform to be applied in HPC, multimedia streaming, high throughput network facing applications, high speed backups, and network based storage. This opens the door to many possibilities when combined with the features in Solaris 10 such as VLAN tagging, IPQoS, iSCSI, NFSv4, and in the future Crossbow. This wide range of applications can also lead to reductions in cabling infrastructure for data centers that today must account for multiple connectivity points.

To the far right of the block diagram are the SSI ports that go to the ASICs that make up the service processor (SP). The SP contains the integrated lights out management (iLOM) interface, NVRAM, FRU-ID tags, TOD/RTC clock, serial interface, and network interface for remote access. The SP controls the platform in a number of interesting ways. It is responsible for running the iLOM software, POST, OBP, and the sun4v hypervisor.

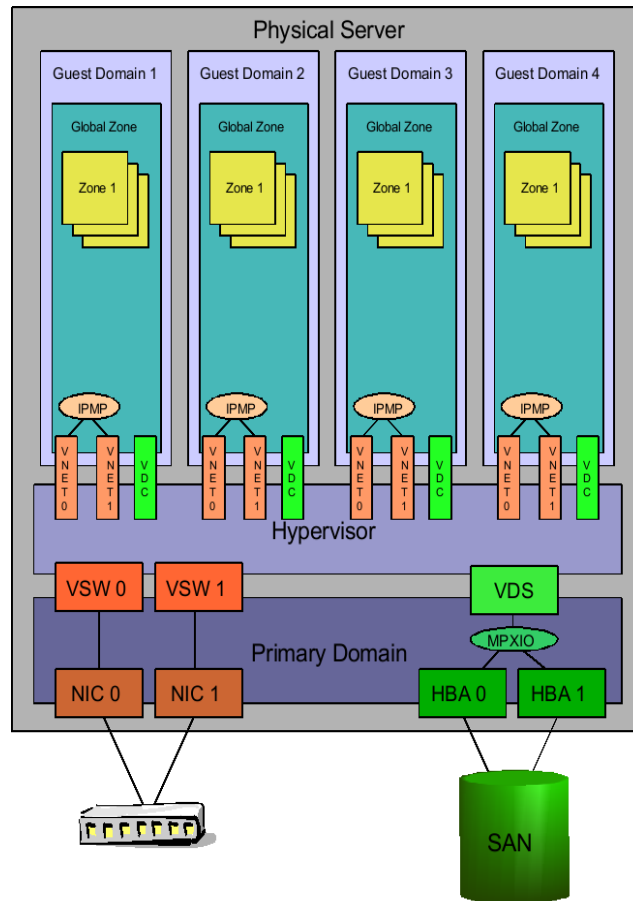
Virtualization:

One of the most compelling aspects of the T5140 and T5240 are the virtualization capabilities. By leveraging Solaris Containers and Logical Domains it is possible to consolidate legacy servers and application infrastructures. Many traditional servers are heavily under utilized and consume expensive data center resources, which are negatively impacting businesses today.



Solaris containers were introduced with Solaris 10 as a native virtualization solution that creates secure run-time environments with resource controls. In this model, the kernel virtualizes a complete run-time environment that enables applications and users to be secured from each other. Each container or zone as it is also known, has its own operating environment, SMF services, network connectivity, and storage to run applications. From a user or application perspective, each container is a separate entity incapable of seeing other containers without standard network methodologies. When the resource management features in Solaris 10 are applied to a zone, it becomes a container. Resource management allows one to slice and dice the CPU, memory, and network resources that are available to each container. As a result, a single physical server can be divided into hundreds or thousands of virtual machines with granular resource controls. This functionality is available on any server that can run Solaris 10 or above.

The T5140 and T5240 have an amazing amount of CPU and memory capacity for consolidating applications with Solaris containers. By having 128 CPU threads and up to 128GB's of RAM on the T5240, there is a lot of headroom for applications. In typical servers, it is difficult to create multiple processor sets and resource pools due to the small amount of CPUs or threads. However, with 128 threads, it is possible to create a wide array of resource pools to handle different workloads for consolidation. When combined with the memory, networking, and I/O capabilities, these servers definitely stand out as consolidation workhorses.



One of the unique features of the UltraSPARC-T1, UltraSPARC-T2, and UltraSPARC-T2 Plus servers is the ability to virtualize the server into discreet virtual machines called logical domains through an integrated hypervisor.

Hypervisors provide a virtualization platform for running multiple operating system instances. Hypervisors have been around since the 1960's, thanks to IBM's CP/CMS which is the ancestor of IBM's current z/VM solution. Until recently, such technology was only found on such proprietary platforms. However, with the advent of Xen and VMware ESX, hypervisors are becoming more common place. The hypervisor found in Sun's UltraSPARC-T1, UltraSPARC-T2, and UltraSPARC-T2 Plus architecture, known as the UltraSPARC Hypervisor, is a new addition to this growing virtualization methodology.

The UltraSPARC hypervisor is a thin layer of software stored within the ALOM or iLOM CMT firmware. It creates a layer of abstraction between the operating system and the physical hardware. Traditionally, operating systems have the concept of non-privileged and privileged access to the underlying hardware. The hypervisor introduces an additional layer of privileged access, known as hyper-privileged access. Hyper-privileged access enables the hypervisor to either expose or hide resources from an instance of an operating system. This allows resources to be grouped into logical partitions or domains. This is similar to Sun's Dynamic System Domains, with the main difference being that the resources are not electrically partitioned, but virtualized.

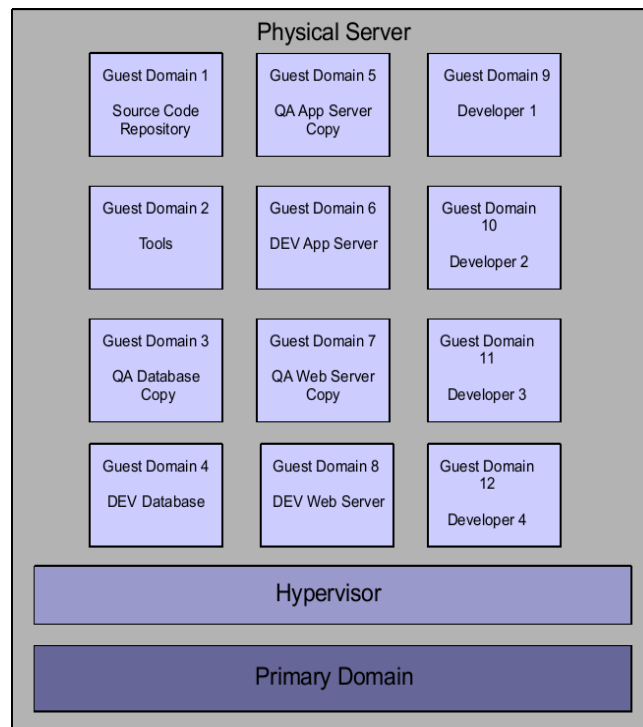
Resources such as CPU threads, cryptographic threads, and memory are partitioned into a logical domain. While other resources are virtualized and serviced through the use of Logical Domain Channels, or LDCs. LDCs provide secure communication and data pathways between LDOMs and the hypervisor. This allows an operating system in one LDom to proxy the I/O request to another LDom that has hyper-privileged access to the physical hardware. Virtualizing the network and storage I/O allows the utilization of the physical hardware to increase and reduces the number of physical I/O cards required to support each LDom.

In order to utilize this feature on the T5140 or T5240, at least version 1.0.3 of the Logical Domain Manager (LDM) is required. This is a separate piece of software that must be installed in addition to Solaris 10 07/08. One must also ensure that the firmware is at version 7.1.3d or higher. Once these base requirements are met, this feature can be fully leveraged.

Let us explore some of the possible use cases for these technologies. Containers provide an excellent mechanism for isolating and controlling the resources of a given application. Logical domains are an excellent mechanism for consolidating servers and increasing utilization. It is possible to combine these two technologies and maximize their strengths.

Development Environment in a Box:

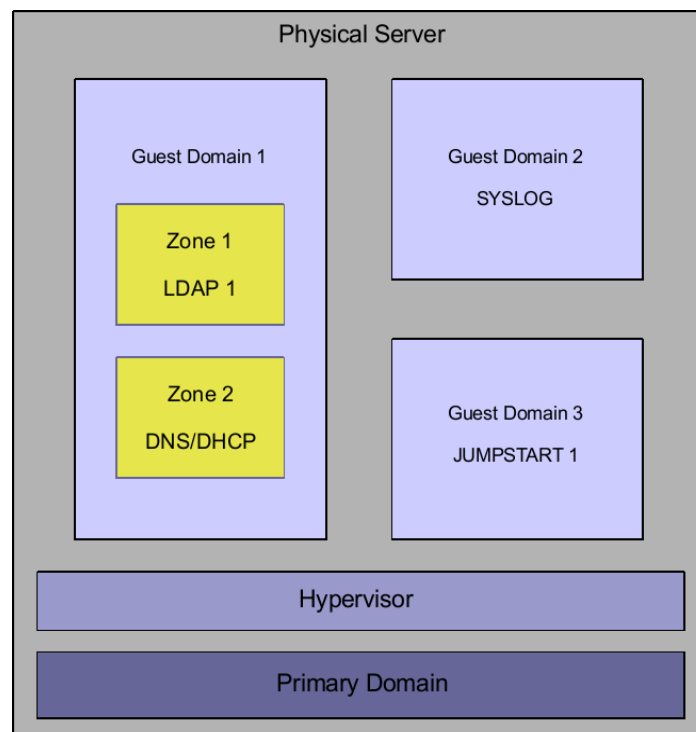
One possible exciting use is to consolidate development environments into a single server by making use of logical domains. Each developer can have an LDom with containers for compiling, debugging, and testing their applications. These can quickly be provisioned through the use of ZFS volumes with snapshots and clones on the back-end. This would enable developers to rapidly perform regression testing. It is even possible to create guest domains that mimic the production environment for QA testing.



By leveraging Logical Domains, one can consolidate an entire development environment onto a single server or small handful of servers. The impact this can have in the data center is very significant. In the above diagram, twelve LDom's are running on a single server. In a traditional configuration, one would have needed twelve physical servers to accomplish the same desired result. This would have consumed more space, power, cooling, network ports, and SAN ports. The benefits of using virtualization become clearer as one considers the impact on the bottom line.

Infrastructure Services in a Box:

Another interesting use of virtualization is to consolidate low-utilization services that typically make up the infrastructure back-bone. These include services such as LDAP, DNS, DHCP, SYSLOG, and Jumpstart. Typically, these kinds of services are configured as standalone instances, don't require clustering, and use very little server resources. These are excellent candidates for virtualization. Some of these services can be virtualized differently depending on the requirements. For example, LDAP can easily be configured into a Solaris container. DNS and DHCP can be grouped together into a Solaris container as well. Syslog servers are typically audited and it may not make sense to mix it with other services on the same OS instance. Jumpstart servers typically act as NFS servers to share out the OS media and related materials, which can not be done with a Solaris container. These requirements can lead to both Solaris containers and LDom's being leveraged on same physical server:



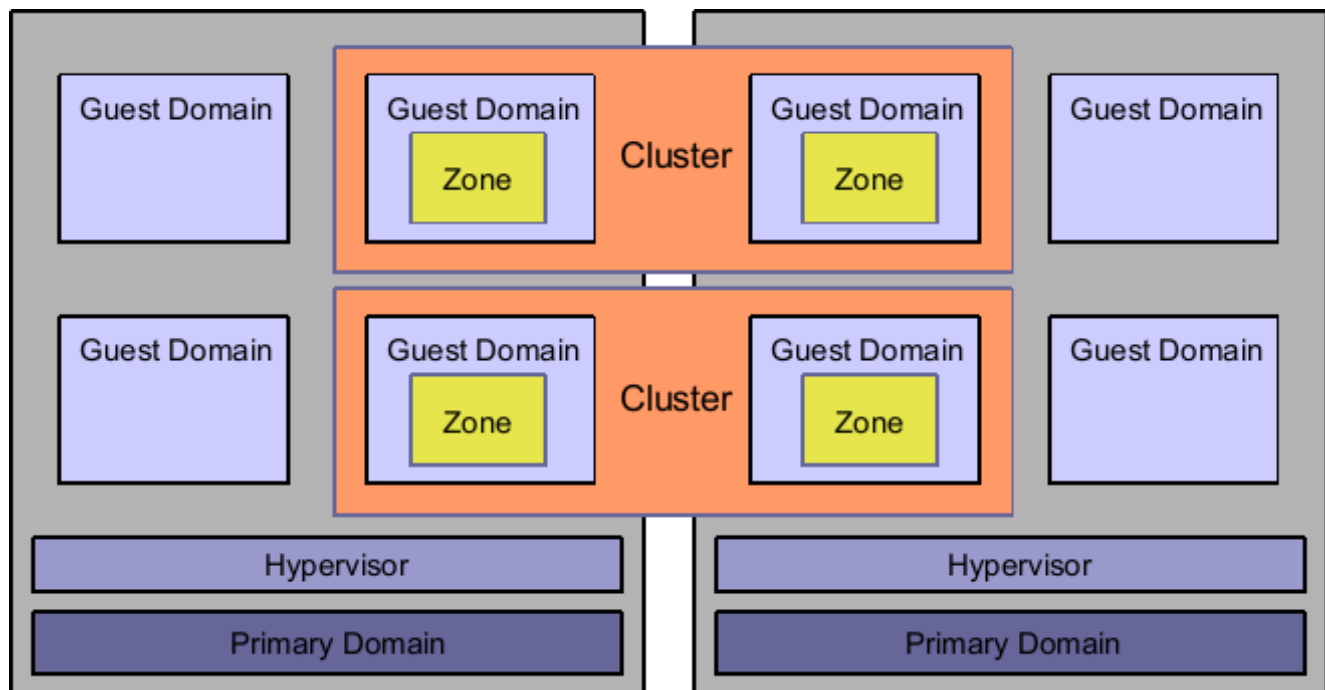
Infrastructure services could be consolidated with these methods and help reduce the number of physical servers in a data center. These types of services typically can be deployed with multiple instances and do not require the use of clustering. For example, LDAP servers can be replicated and most companies have multiple DNS or DHCP servers. As a result, the above configuration could be replicated onto multiple servers for redundancy. It also means that if a new data center has to be constructed, one could deploy such a server to get everything else up and running quickly.

Scalable Production Environment:

These virtualization features can also be leveraged for production environments that require highly available services and applications. This can be accomplished via the following methodologies:

- Leverage applications that are distributed or cluster at the application layer.
- Leverage clustering software.
- Enable the mobility of the application.

These methodologies can be combined depending on the application or service capabilities. For example, some commercial application servers have clustering capabilities built-in. This means that each instance of the application server is able to maintain coherency at the application layer without the assistance of another product or expensive infrastructure. This quickly become the goal for many commercial and in-house applications where having tightly coupled clusters or fail-over services is too costly and inefficient. However, there are many applications that still need traditional clustering capabilities. Solaris Cluster can be leveraged with virtualization to provide HA capabilities for applications. Another methodology that is become common place is the ability to move a virtual machine from one host to another when there is hardware failure or planned maintenance.



In the above example, two physical servers are running LDOMs and Solaris Clustering. Some of the guest domains have Solaris containers and are clustered. This allows applications to be monitored and restarted automatically. If there is a failure, the containers can fail-over between the guest domains that are on different physical servers. Meanwhile, there are other guest domains that are not clustered, but can easily be brought up on another LDom capable server in the data center. All of this is possible by virtualizing the servers, networking, and storage.

Pros and Cons:

Both of these new servers have new features and benefits. However, as with all technology there are pros and cons that must be carefully reviewed when selecting a product for any given deployment.

Here are some of the pros for using the T5140 or T5240:

- 1U and 2U form factors available
- Two UltraSPARC-T2 Plus processors supplying 128 threads total
- 128GB's of RAM on the T5240
- Six PCI-E x8 lane slots
- Integrated 10GbE
- 4-16 SAS drives depending on model
- Excellent platform for virtualization and consolidation

Here are some of the cons for using the T5140 or T5240:

- Only one integrated SAS controller, which creates a bottleneck and prevents a full split PCI-E configuration for LDOMs.
- 10GbE is no longer integrated into the processors and sits on a single PCI-E switch.
- Higher power requirements when using the T5240 with 16 SAS drives.
- No remote console through the iLOM web interface.
- No support for the I/O Expansion Unit that is available on the M-Series
- Single threaded performance can still be an issue when migrating legacy applications.

While there are a few cons against the platform, they do not prevent it from hitting the mark and delivering a robust and powerful platform. Hopefully, the next generation of T-Series servers will address some of these issues.

Summary:

The T5140 and the T5240 offer many unique and exciting features that set it apart from the competition. The UltraSPARC-T2 Plus processor with 8 cores, 64 threads, SMP, PCI-E, and cryptographic features are revolutionary in the computing industry. By combining these features with the ability to virtualize the server with Solaris Containers and Logical Domains, it is possible to consolidate legacy equipment and increase the efficiencies in the data center today. It also opens the door to creating utility computing based on farms of virtualized servers and infrastructure. This enables greater flexibility for deploying services and confronting the challenges faced in traditional data centers.

References:

Sun web site for the CoolThreads servers:

<http://www.sun.com/servers/coolthreads/overview/index.jsp>

Sun Documentation on the T5140 and T5240:

<http://docs.sun.com/app/docs/coll/t5140>

Sun SPARC Enterprise T5120, T5220, T5140, and T5240 Server Architecture:

<http://www.sun.com/servers/coolthreads/t5140/wp.pdf>

OpenSPARC.net site for the UltraSPARC-T2 Architecture and Specifications:

<http://www.opensparc.net/opensparc-t2/index.html>

OpenSolaris LDoms Community Site:

<http://www.opensolaris.org/os/community/ldoms>

Sun Cluster 3.2 02/08 Release Notes Wiki:

<http://wikis.sun.com/display/SunCluster/Sun+Cluster+3.2+2-08+Release+Notes#SunCluster3.22-08ReleaseNotes-optguestdomain>