



Logical Domains (LDoms) on the Sun4v architecture

Presenter: Nathan Kroenert
N2 FanBoy and consumer of steak

Based loosely on the Customer Ready presentation of Logical
Domains. :)
Diagrams © Sun Microsystems



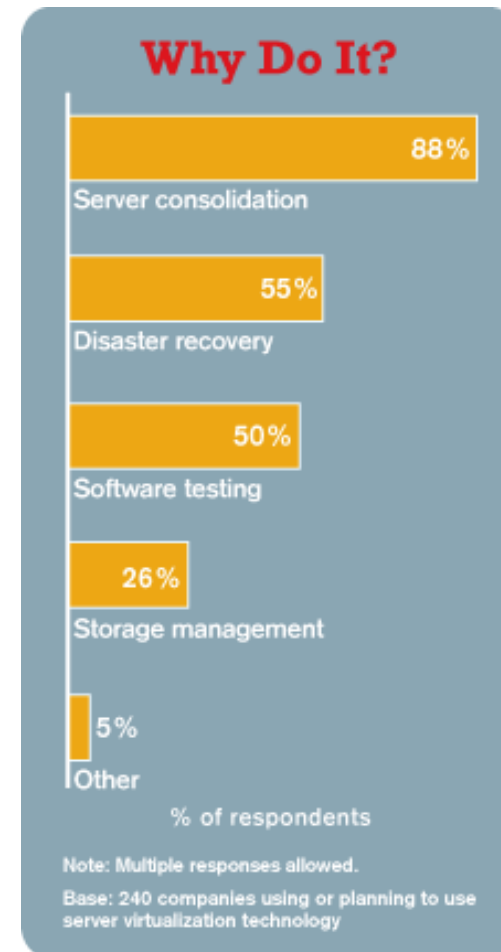
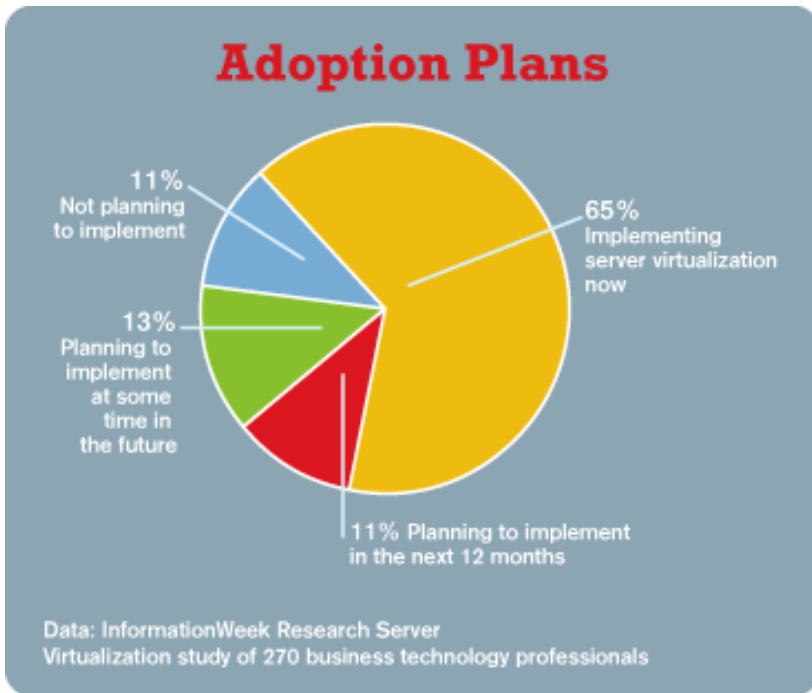
Caveats / disclaimer / get out of jail free...

- LDOMs are gaining market acceptance quickly, but are still not always the right choice! I'm not here to sell 'em. Just tell you stuff about 'em.
- What is presented here is from the publicly available information
- I and the information I depend on might be wrong!
- Sun definitely makes no warranties of correctness...
- I continue to eat Steak. It's too tasty not to. If you live close to Frankston, visit Brancatisano's and get a Rib-Eye steak.

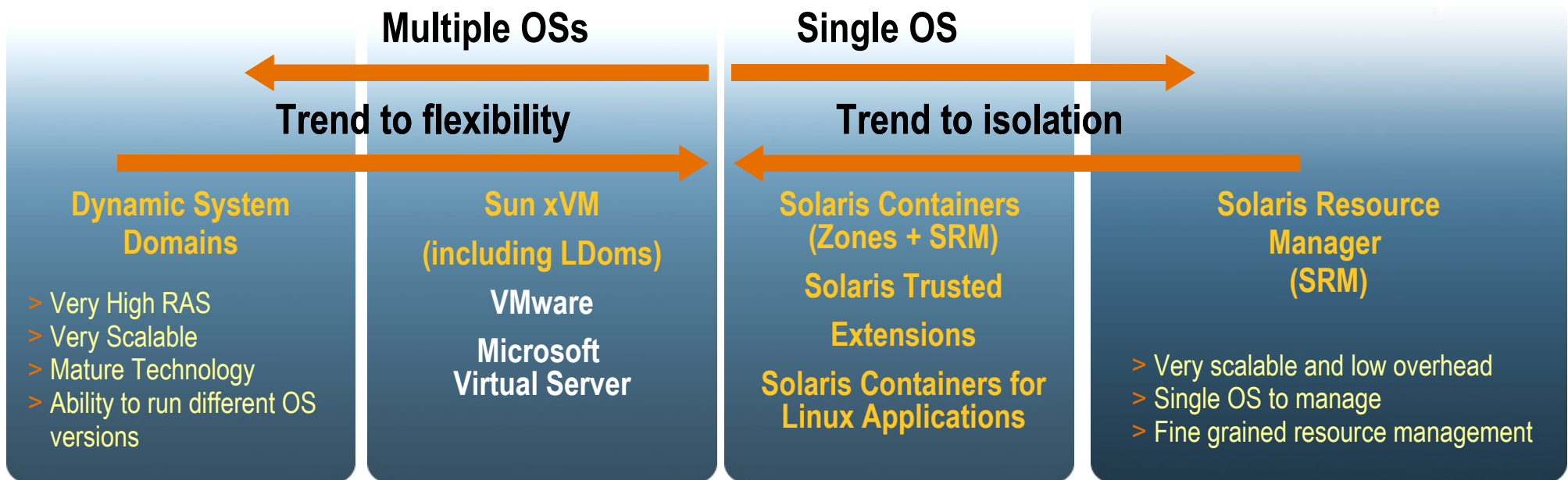
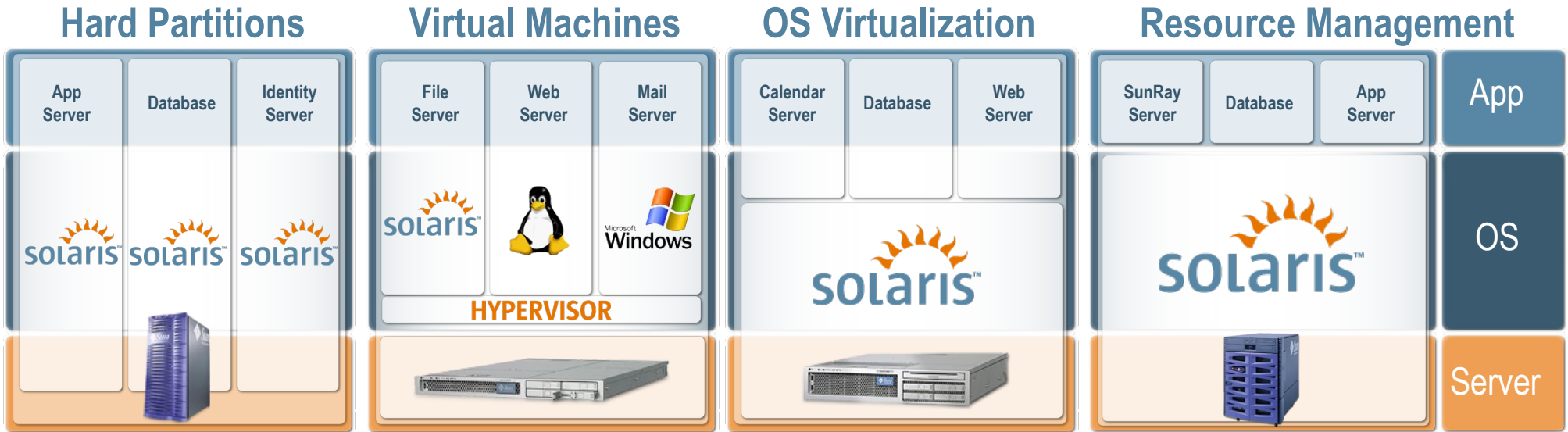
Agenda

- LDoms Overview (quick)
- LDoms Technical Details (Not so quick)
- LDom example on a real box scattered throughout
- Pizza?

Virtualization Takes Off



Server Virtualization@Sun

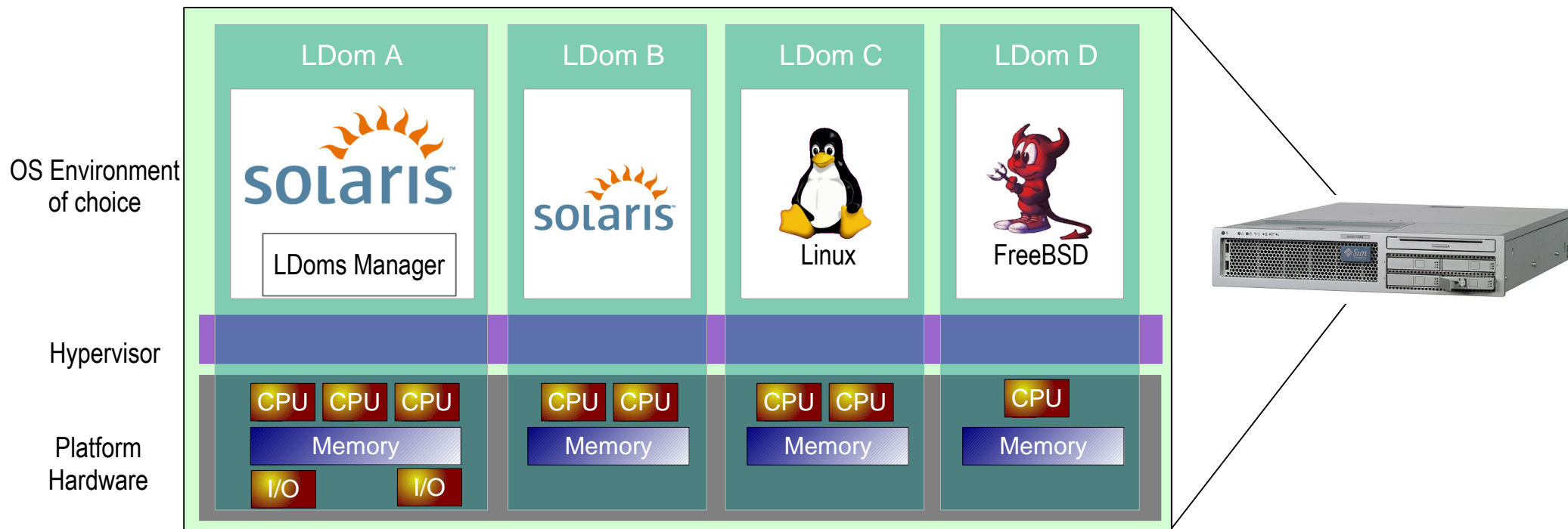


Review of Sun Server Virtualization Offerings

- Midrange and High End SPARC - Dynamic System Domains
 - > Electrically isolated, no single point of failure that will prevent the system from restarting
 - > Resource granularity at board level
 - > Improved granularity to processor level with SPARC M-series Enterprise servers, though that has it's own baggage
- x64 Virtual machines
 - > Solaris unique among Unix systems for running under VMware
 - > Solaris support for open-source Xen in development. Available NOW in Solaris Express (Note: Not Solaris 10, and currently no support)
 - > Sun x64 servers excellent for VM-based consolidation
- Solaris Containers on both SPARC and x64
 - > Low overhead, single OS instance, scalable across platforms
- Logical Domains on Sun SPARC CMT
 - > Low overhead, multi-OS flexibility, easy of operation, full OS isolation

Logical Domains Technology on Sun4v

- Virtualization and partitioning of machine resources
 - > Each domain is a full virtual machine, with a dynamically reconfigurable sub-set of machine resources, and its own independent OS
 - > Protection & isolation via SPARC hardware and hypervisor firmware. Many fault classes impact only one domain.



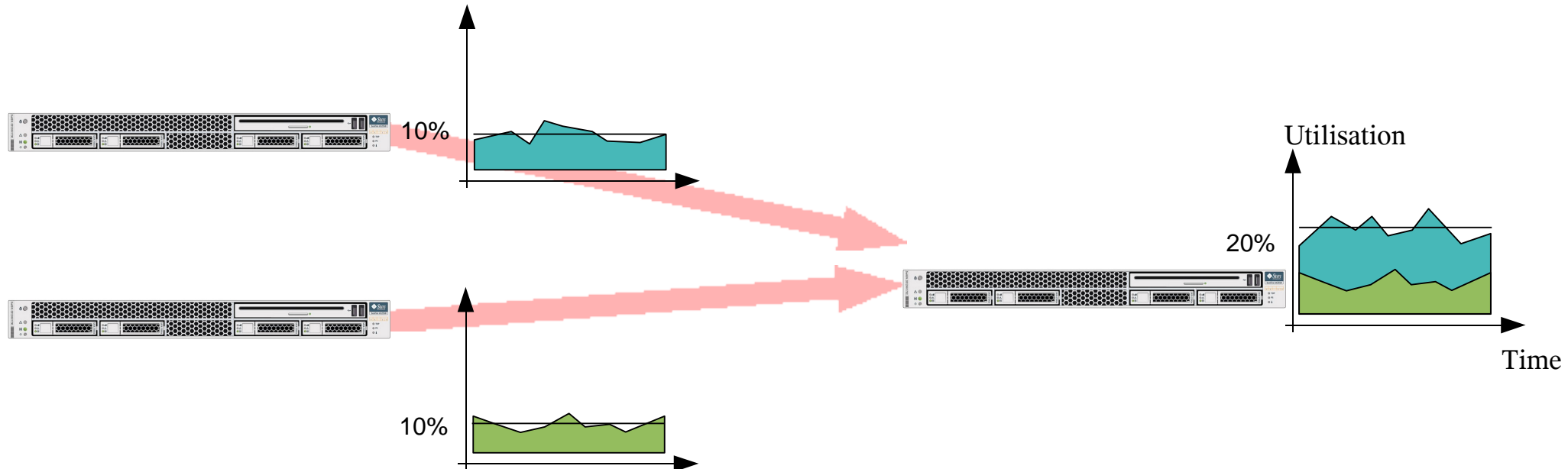
LDoms Fundamentals

- Each virtual machine appears as an entirely independent machine
 - > own kernel, patches, tuning parameters
 - > own user accounts, administrators
 - > own disks (or at least virtual devices)
 - > own console and OBP to boot from
 - > own network interfaces, MAC & IP addresses
 - > each domain can start, stop and reboot independently of each other*
 - > We have unprivileged, privileged and hyperprivileged 'modes' to mediate access where required

*Assuming it's not the Service Domain, or that there is only one service domain... If it is, then you pause, and wait for it to come back.

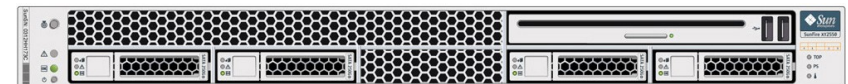
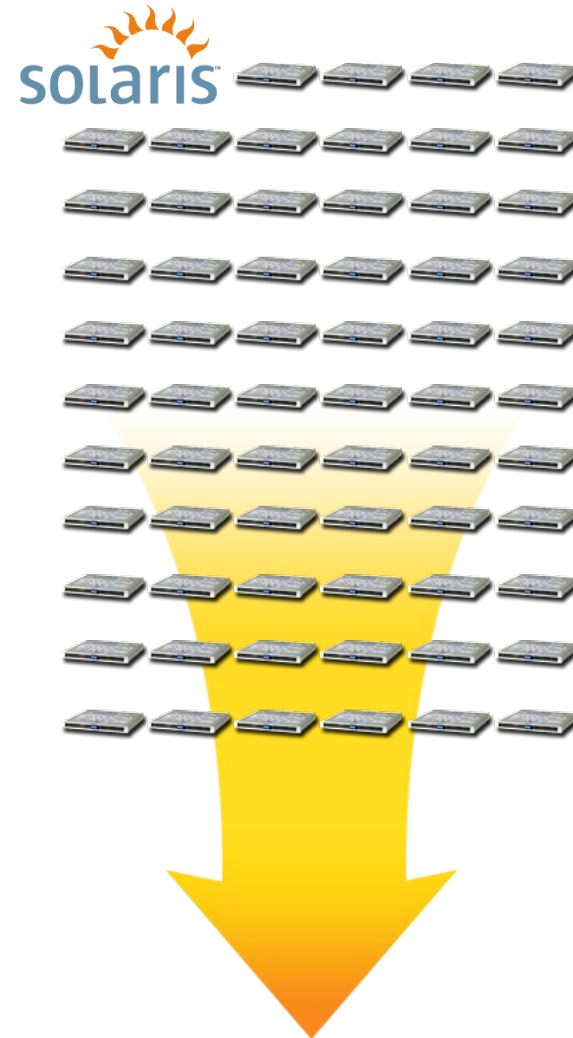
Just some of the problems LDOMS can help to solve:

- Problem: Low server utilization rate in the data center
 - > Consolidation is the major driver to reduce server footprints and better utilize the server capacity
 - > Goal: cut IT hardware spending and ongoing support and administration cost while streamlining the deployment of new IT services and increasing availability of the business services



Consolidation Saves

- Scale vertically and horizontally at the same time with LDomS
 - > Up to 64 domains per system (T2)
 - > 2,560 domains per rack (URK!)
- Granularity down to a single thread if required, usually only to core in reality.
- Memory is also sliced and diced as required



Logical Domains 1.0 Features

Released on 4/26/2007

- Up to 32 LDoms per UltraSPARC T1 based systems
 - > Fine-grained resource control at the CPU thread level
- Guest domains can be configured, started and stopped independently
 - > Without requiring a power-cycle of machine
- Ability to dynamically add and remove vCPUs while OS is running
- Predictive self healing capability for each logical domain
- Control domain hardening

LDoms 1.0 Supported Platforms

- Supported on UltraSPARC T1 (Niagara) based systems
 - > Sun Fire and SPARC Enterprise T1000 Servers
 - > Sun Fire and SPARC Enterprise T2000 Servers
 - > Netra T2000 Server
 - > Netra CP3060 Blade
 - > Sun Blade T6300 Server Module
- Requiring Solaris 10 11/06 at a minimum

LDoms improved:

- LDoms 1.0 released on 4/26 to support UltraSPARC T1 platforms, running Solaris 10 11/06
- LDoms 1.0.1 released on 10/11 to support UltraSPARC T2 platforms , running Solaris 10 8/07
- Warning – Marketing below!
- Within 1st 3 months of 1.0 release
 - > 500+ different customers started evaluation
 - > 66% of Sun's top global customers started LDoms proof of concept
 - > Customers started deploy LDoms in production
 - > Telco design wins – Solaris/CMT/LDoms/NDPS

New Benefits in LDoms 1.0.1



- Preinstalled on all Sun SPARC Enterprise T5120/T5220 systems and Sun Blade T6320 Server Modules
- Start or stop primary and guest logical domains independently and *without* a system reboot
- Includes full support for UltraSPARC T1 systems
- LDoms Management Information Base (MIB) helps third-party system management programs remotely manage and start/stop LDoms via Simple Network Management Protocol (SNMP)

LDoms 1.0.1 Supported Platforms

- In addition to the UltraSPARC T1 platforms, LDoms 1.0.1 adds support for UltraSPARC T2 platforms
 - > Sun SPARC Enterprise T5120 Server
 - > Sun SPARC Enterprise T5220 Server
 - > Sun Blade T6320 Server Module
- Requiring Solaris 10 8/07 OS for the primary domain; however, Solaris 10 11/06 plus required patches can run on the guest domains (but, don't... Use update 4 if at all possible. Performance fixes, patching improvements, live-upgrade and ZFS fixes alone make it worthwhile. :)

Why I think it's cool:

- > Partitioning of a single physical machine into more than one virtual machines. Even on small boxes like the T2000 / T5120
- > Different OSs or different versions of the OS (& apps) to run in different domains on the same physical machine – Different patch levels too
- > Same box could be used for Dev, Test, Q/A and Production.
- > Makes Disaster Recovery simple with SANs...
- > Cool tech. :)
- > Greatly increased utilization and flexibility for admins catering to business unit requirements
- > Lower total cost of ownership (TCO)
- > Simplify and accelerate creation of new domains for changing requirements. Someone wants a proof of concept 'slice' of a box? No worries. You are only about 10 commands away from having one.

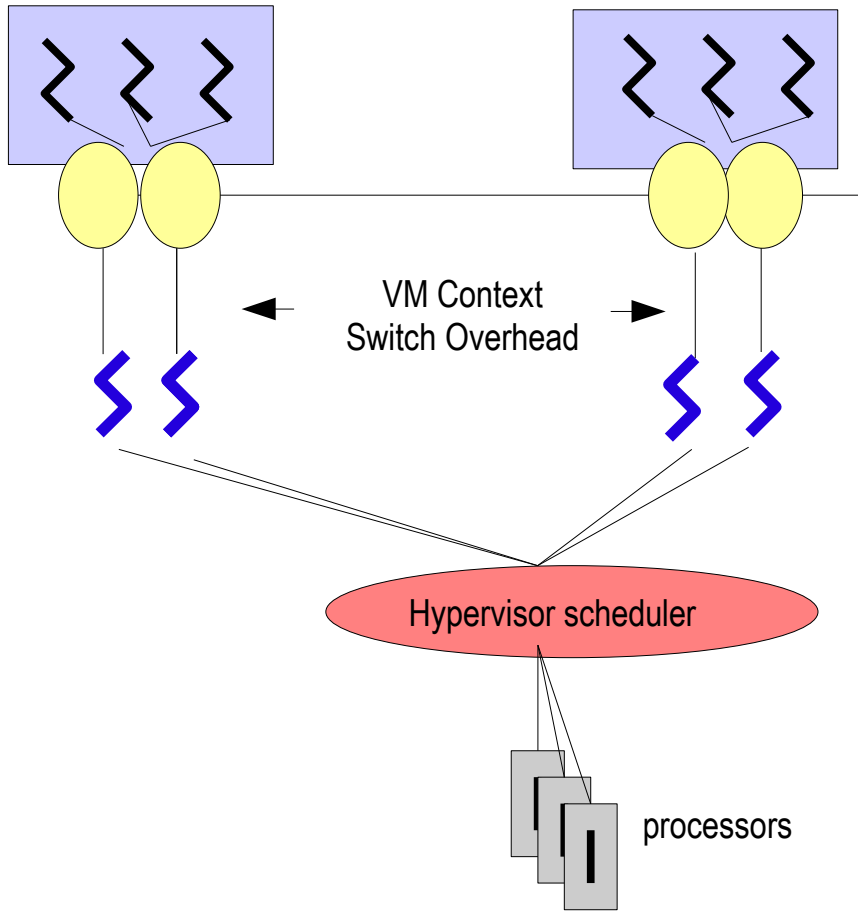
Logical Domain Technical Details

Hypervisor Support

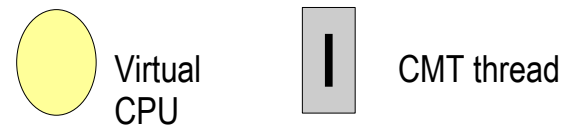
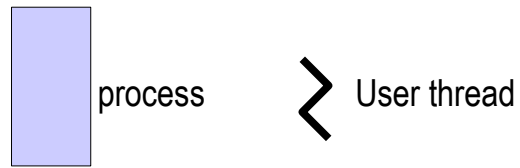
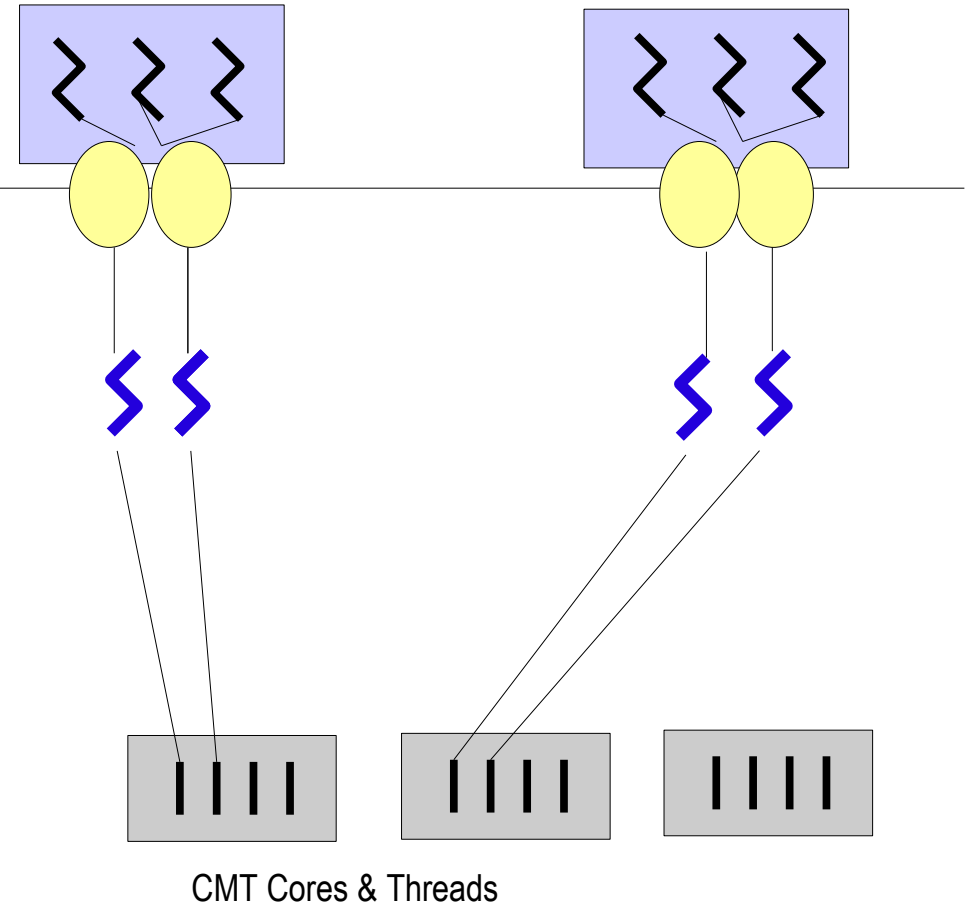
- For low level details on the hypervisor, see
 - > http://opensparc-t1.sunsource.net/specs/Hypervisor_api-26-v6.pdf
- Hypervisor is responsible for maintaining separation between domains
 - > Using extensions built into a sun4v CPU
 - > Think of it as the 'Fat Controller' in the chip. (Not to be confused with Fat Agnes....:)
- Also provides Logical Domain Channels (LDCs) so that domains can communicate with each other
 - > Mechanism by which domains can be virtually networked with each other, or provide services to each other such as virtual disk devices / domain console

CPU Virtualization

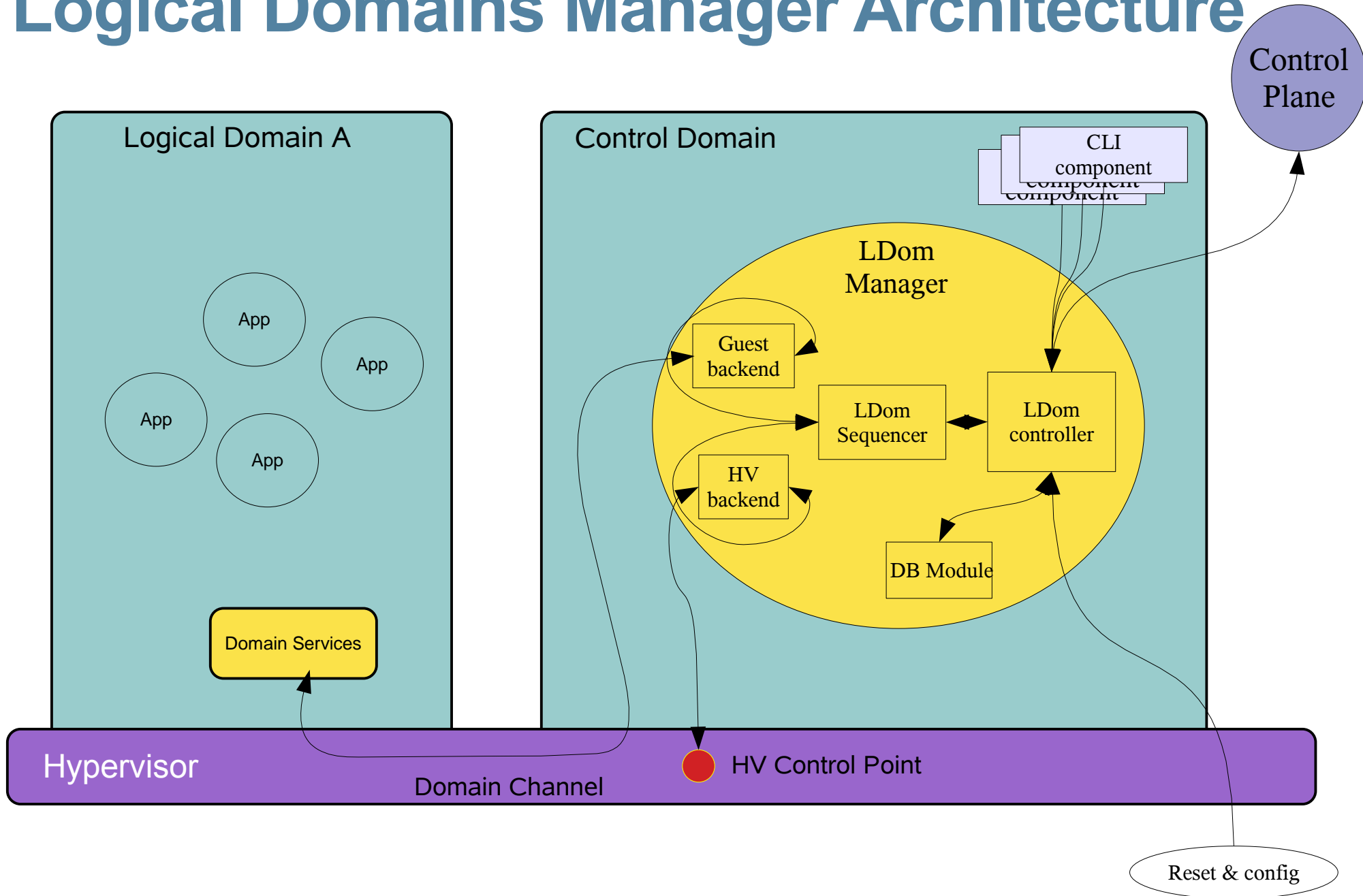
Scheduler Based (IBM, HP, VMware, Xen)



CMT Based: Hardware Scheduled



Logical Domains Manager Architecture



LDoms Manager

- Concept of a “service domain”, which services the requests of the non-service LDOM's
- One Manager per machine
 - > Can be run in any domain, but only 1 domain at a time
 - > The “Control Domain”
 - > Used to create the other LDOMS
 - > Controls Hypervisor and all its LDoms
- Exposes CLI to administrator
- Maps Logical Domains to physical resources
 - > Constraint engine
 - > Heuristic binding of LDoms to resources
 - > Assists with performance optimization
 - > Assists in event of failures / blacklisting

LDoms Manager & Service Processor

- The service processor (ALOM or ILOM) monitors and runs the physical machine
 - > Does not know about virtual machines, but stores the LDoms configuration so that it's persistent. Can only invoke pre-created configurations setup in the Control Domain.
- The LDoms manager runs the virtual machines
 - > Idmd is the 'gateway' to do all this control
 - > Manages the mapping of physical resources
 - > Responsible for providing new rule sets to the Hypervisor

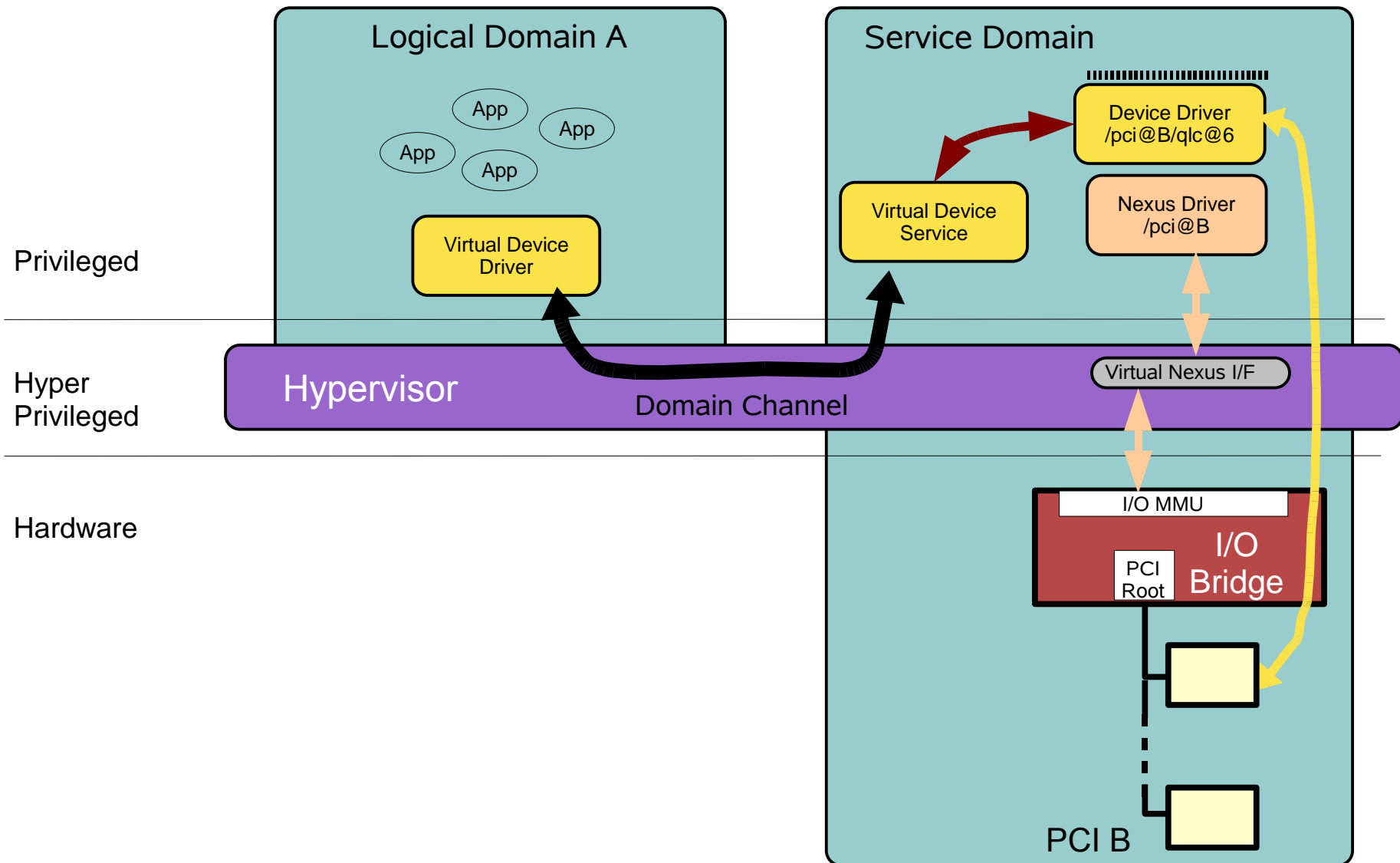
Logical Domain Roles

- **Control domain** - Creates and manages other logical domains and services by communicating with the hypervisor.
 - > Control domain is usually acting as a service and I/O domain and is the accepted best practice. Simplifies configuration.
- **Service domain** - Provides services, such as a virtual network switch or a virtual disk service, to other logical domains.
- **I/O domain** - Has direct ownership of and direct access to physical input/output devices, such as a PCI Express card or a network device. Shares those devices to other domains in the form of virtual I/O devices.
- **Guest domain** - Uses services from the service and I/O domains and is managed by the control domain.

LDoms Virtual IO

- Logical Domains allows assignment of resource on single chassis by abstracting the underlying compute and IO resources
- Not always possible to provide domains direct access to the IOMMU or the device and its registers, so hence service domains
- LDoms VIO infrastructure provides indirect access to domains via virtualized devices that communicate with the 'service domain'. Service domain completely owns a device along with its driver, and functions as a proxy to the device.
- Implemented as a client-server model where, client virtual devices communicate with their service counterpart via inter-domain LDoms Channels (LDC).

Virtualized I/O



Split PCI configuration

- PCI-E buses on T1000 and T2000 platforms have two ports that can be assigned to separate domains
 - > Provides direct I/O access to two domains in the system
- I/O Domains can function as service domains and export virtual disks or provide network service to other domains.
- Referred to as “bus_a” and “bus_b” by the LDom Manager
 - > Ensure following a split, each bus has appropriate disk and network available
- T5120 and T5220 systems have only one bus. Cannot be split.
- Watch this space for future systems with more stuff

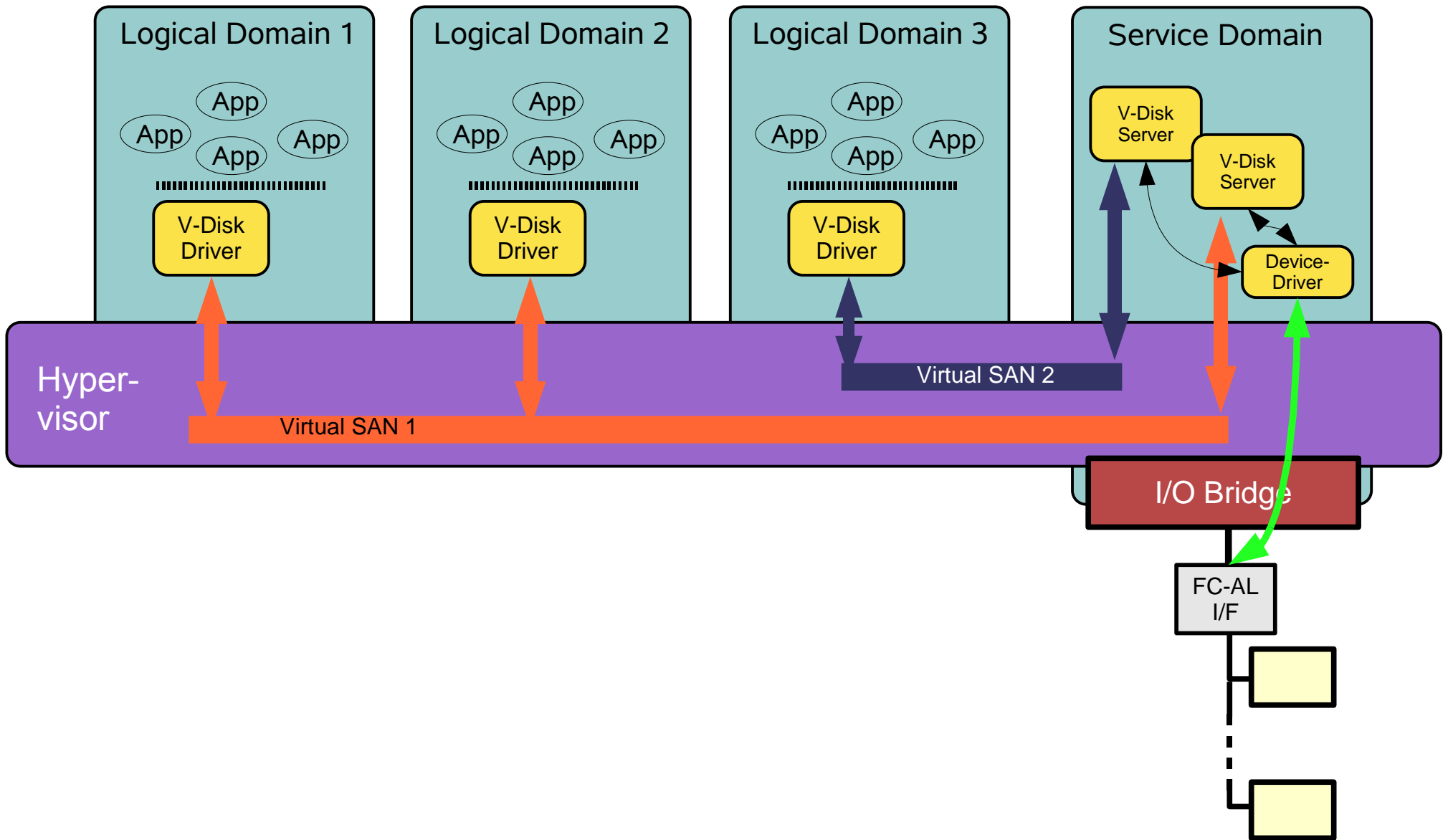
Split PCI configuration

- T5120 and T5220 systems have only one bus. Cannot be split.
- Watch this space for future systems with more stuff
 - > <http://www.sun.com/emrkt/innercircle/newsletter/0407feature.html>
Is innacurate. It has some interesting hints, but is wrong. :)
- theregister.co.uk is closer...

Virtual IO componets

- Channel Nexus
- Virtual Devices
 - > Virtual Switch and network device
 - > Virtual Disk Server (VDS) and client
 - > Virtual console concentrator
 - > Virtual Network Terminal Server (NTS) daemon
- Core framework
 - > Layered over LDoms Channels (LDC)
 - > Common protocol for client-server, and client-client communication
 - > Shared memory and descriptor ring based data transfer

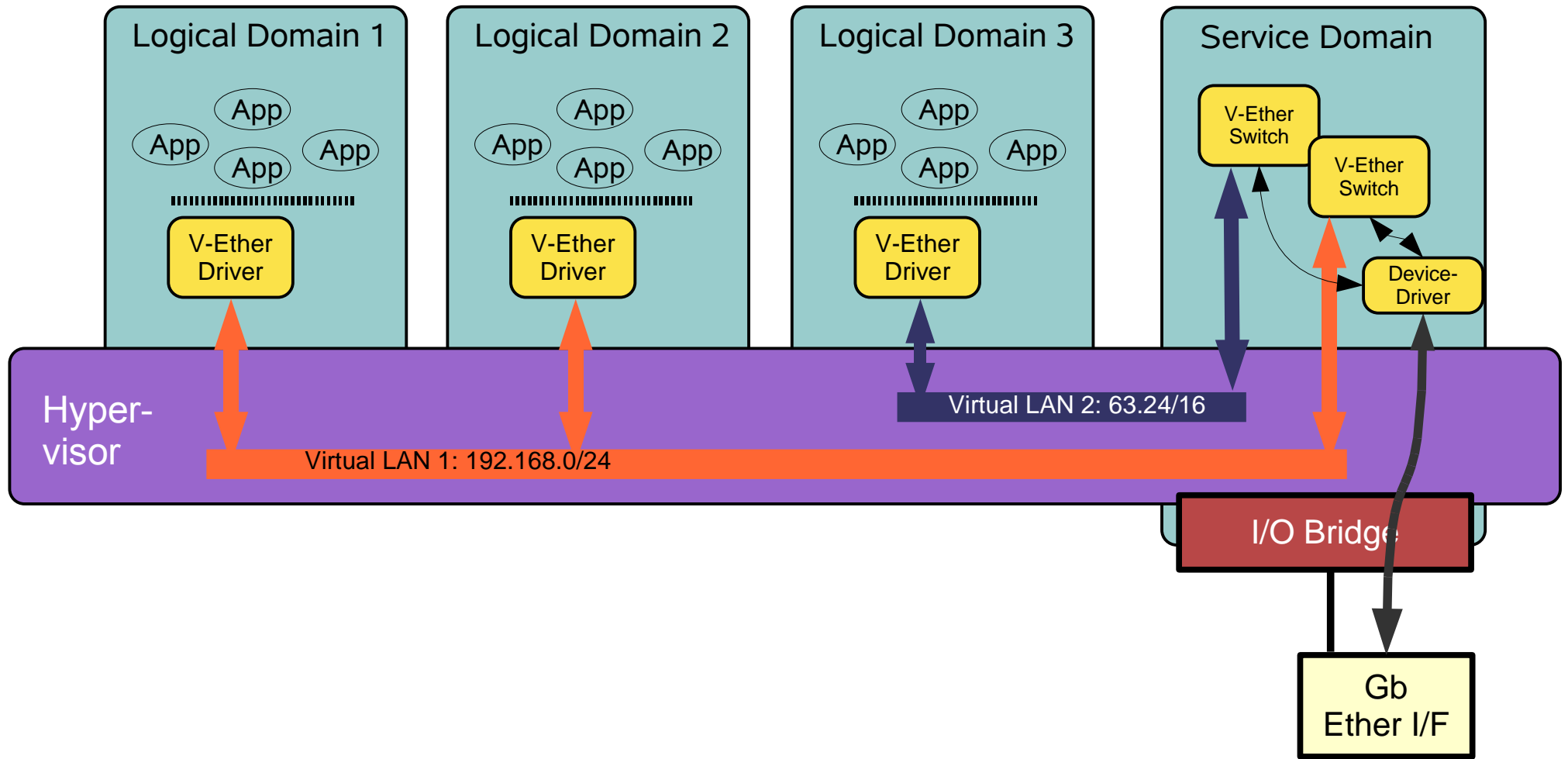
Virtual (Block) Disk device



Virtual Disk Server

- Exports a virtual disk to a guest domain
- The Virtual disk can be based upon several device types:
 - > An entire physical disk, which could also be a storage partition presented by a SAN device, sometimes referred to as a logical unit number (LUN)
 - > Single slice of a disk or LUN
 - > Disk image file on a file system (such as UFS or ZFS)
 - > Disk volumes (ZFS, SVM, VxVM)
 - > Flat files **might** work. YMMV. Don't seem to survive a reboot of the control / I/O domain.

Virtual Ethernet device



Virtual network

- Virtual network device (“vnet”)
 - > Simple virtual Ethernet device
 - > Exports GLDv3 compliant mac interface
 - > Vnets connect directly for domain to domain traffic
- Virtual switch (“vsw”)
 - > Switches unicast and broadcast packets
 - > Connects to a GLDv3 compliant device driver for external network connection
 - > Provides a GLDv3 compliant driver interface
 - > Use service-domain's kernel for Layer-3
 - > Routing, iptable filtering, NAT, firewalling ???
 - > Need explanation of this, as it's not how I remember it!

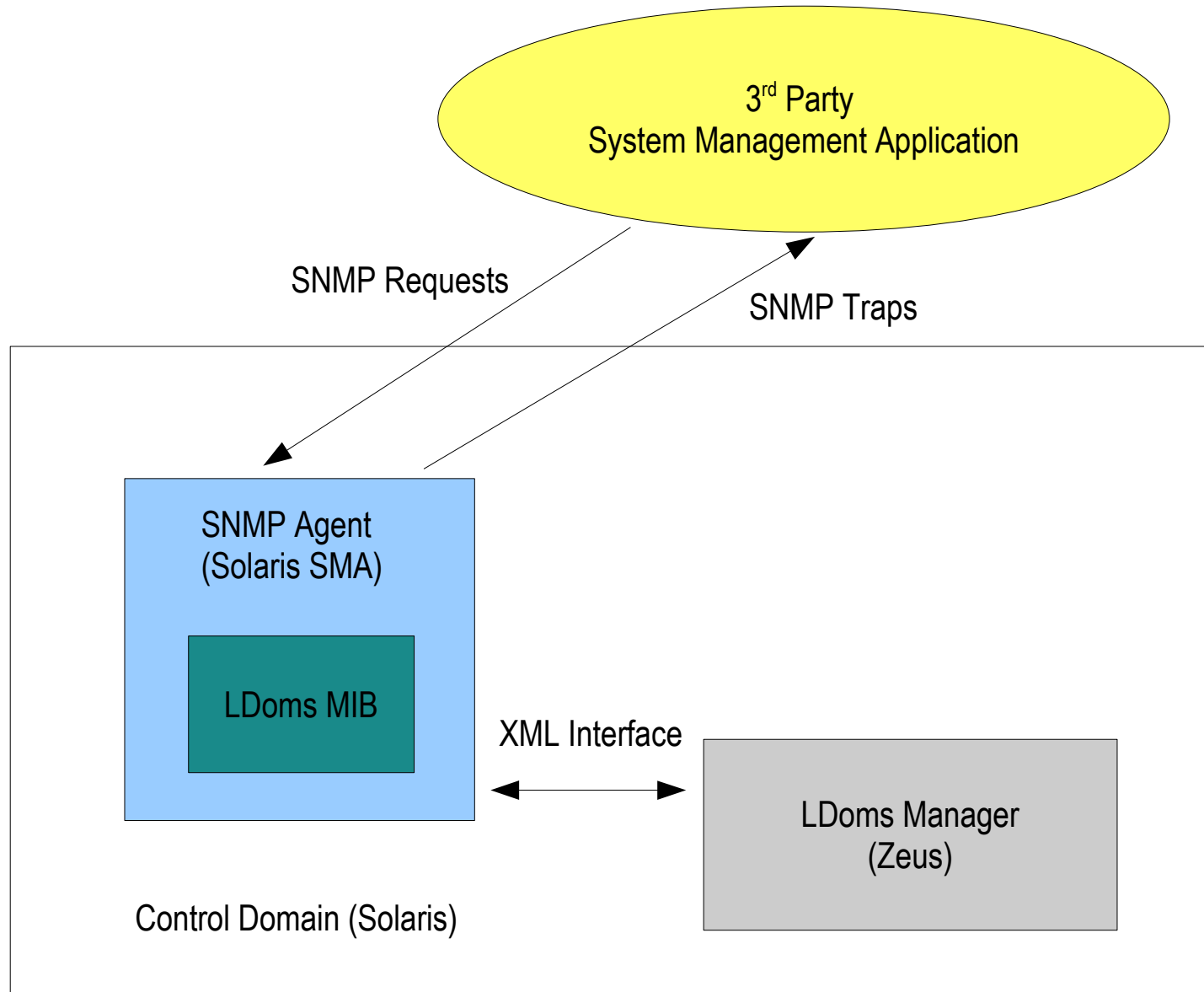
Virtual Consoles

- The virtual console concentrator (“vcc”) driver in the service domain aggregates LDoms consoles and exports access via TTY devices.
- Virtual NTS daemon (“vntsd”) layered on top of vcc exports access via Unix sockets for connecting via telnet within control domain. No – No SSH at this stage! :)
- Control domain console still accessible from the service processor.
- No changes to the Solaris console driver (qcn)

Dynamic Reconfiguration (DR)

- Ability to dynamically grow or shrink compute capacity of a logical domain on demand.
- No need to re-boot Solaris
- Simply add/remove vCPUs
 - > Memory and I/O in future releases. Looking like LDOMs next...
- Improve utilization by balancing resources between Ldoms. Can be fabulous to move resources for the changing balance between OLTP / Batch for example.

LDoms MIB Architecture



LDoms MIB

- SNMP (Simple Network Management Protocol) support for LDoms
 - > monitoring and basic active management (start/stop domains)
- Modeled after industry standard DMTF-CIM
- LDoms MIB (management information base) is delivered as an extension module to the System Management Agent (SMA – Net-SNMP 5.0.9, part of Solaris 10) running in the control domain.

LDoms BUI (Browser User Interface)

- Uses the standard Lockhart Interface (Sun Java Web Console)
 - > Launched in the same way as the ZFS BUI.
 - > Provides an overview of all domain configurations in a single system
- Provides a simplified 'point-and-click' interface
- Allows basic management and reconfiguration
 - > Does not provide all the functionality of the CLI
- To be released with LDoms 1.0.1 as a freeware (unsupported)
- Moving forward with Virt-Manager / Libvirt as the converged GUI tool for single node management

Putting it all together

Bottom line of getting LDOM's running:

Follow the blueprint!! <http://www.sun.com/blueprints/0207/820-0832.html>

The simplest summary is this:

- > Get a Sun4v box
- > Install the *latest* firmware, whatever it is.
- > Install S10U4
- > Create the control/service domain, which is essentially 'restricting' the 'primary' domain's config to a smallish number of CPU's and memory so the other resources are available and creating some of the virtual server resources, like the virtual switch, disk server and console server.
 - > `huron:/root # /opt/SUNWldm/bin/ldm add-vds primary-vds0 primary`
 - > `huron:/root # /opt/SUNWldm/bin/ldm add-vcc port-range=5000-5100 primary-vcc0 primary`
 - > `huron:/root # /opt/SUNWldm/bin/ldm add-vsw net-dev=e1000g0 primary-vsw0 primary`
 - > `/opt/SUNWldm/bin/ldm set-mau 1 primary`
 - > `/opt/SUNWldm/bin/ldm set-vcpu 4 primary`
 - > `/opt/SUNWldm/bin/ldm set-memory 1024m primary`
 - > `huron:/root # /opt/SUNWldm/bin/ldm add-spconfig initial`

Putting it all together

You can now reboot and get into a 'minimized' control/service domain.

Start the vnts server service – `svcadm enable svc:/ldoms/vntsd:default`

From there, for each LDOM you want to create

```
ldm add-domain myldom1
```

```
ldm add-vcpu 12 myldom1
```

```
ldm add-memory 1G myldom1
```

```
ldm add-vnet vnet1 primary-vsw0 myldom1
```

```
ldm add-vdsdev /s3root/ldom/myldom1 vol1@primary-vds0
```

Note: The disk image above was simply created using `mkfile`. I use 'file based' here, but you can use `filebased`, `disk`, `slice` etc.)

```
ldm add-vdisk vdisk1 vol1@primary-vds0 myldom1
```

```
ldm set-variable auto-boot\?=false myldom1
```

```
ldm set-variable boot-device=/virtual-devices@100/channel-devices@200/disk@0 myldom1
```

```
ldm bind-domain myldom1
```

```
ldm start myldom1
```

And you are ready to 'ldm list', check the console port of your LDOM and `telnet 0 <domain port#>`

Key Resources

- Logical Domains
 - > <http://www.sun.com/ldoms>
 - > BigAdmin: <http://www.sun.com/bigadmin/hubs/ldoms/>
 - > OpenSolaris: <http://www.opensolaris.org/os/community/ldoms/>
- Sun Virtualization Solutions
 - > <http://www.sun.com/datacenter/consolidation/virtualization/>
- Virtualization Learning Center
 - > <http://www.sun.com/solaris/virtualization>
- Creating LDOMS blueprint
 - > <http://www.sun.com/blueprints/0207/820-0832.html>



Thank you!

<http://sun.com/ldoms>