Resource Control & Virtualization on Oracle Solaris

A TECHNOLOGY OVERVIEW

Berlin 2013

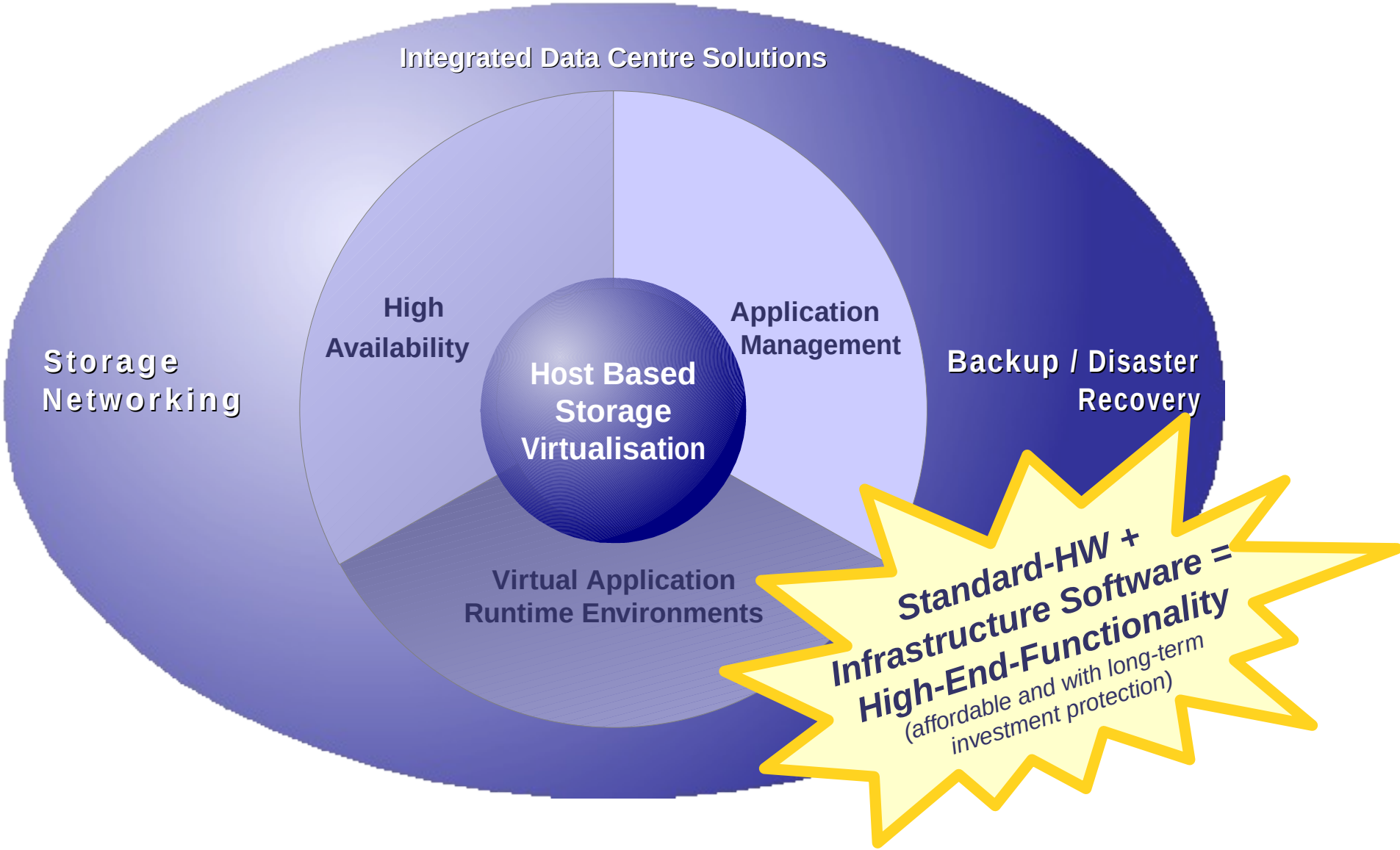
bert.miemietz@osl.eu

Who is OSL?

Author: Bert Miemietz, Rev. 11/2013

OSL Is Developing Infrastructure Software

Storage Virtualization • Volume Management • Converged Networking
Virtual Machines • Clustering • High Availability • Disaster Protection • Consulting



Author: Bert Miemietz, Rev. 11/2013

OSL Storage Cluster 4.0

An Integrated Data Center Solution in a Modular Design

application awareness

i/o-bandwidth control

user management

Application Control Option

clusterwide control of applications

virtualized (hardware abstracted) runtime environments

high availability

resource based self-management

Application Resource Description

mirror entire applications

clone entire applications

B2D / Tape-BU / DR toolset

Clustered Storage Virtualization

global storage pools (across all hosts)

global disk inventory

globa devices / global namespace

cluster volume manager / auto-allocation

disk access management

i/o-multipathing

RSIO

data center block-i/o e.g. via ethernet

integrated with storage virtualization

multipathing + trunking

cross-platform

Extended Data Management

integration of RAID-based data copies & snapshots

host base mirroring

online data migration

cloning of data

Author: Bert Miemietz, Rev. 11/2013

OSL RSIO - Remote Storage I/O

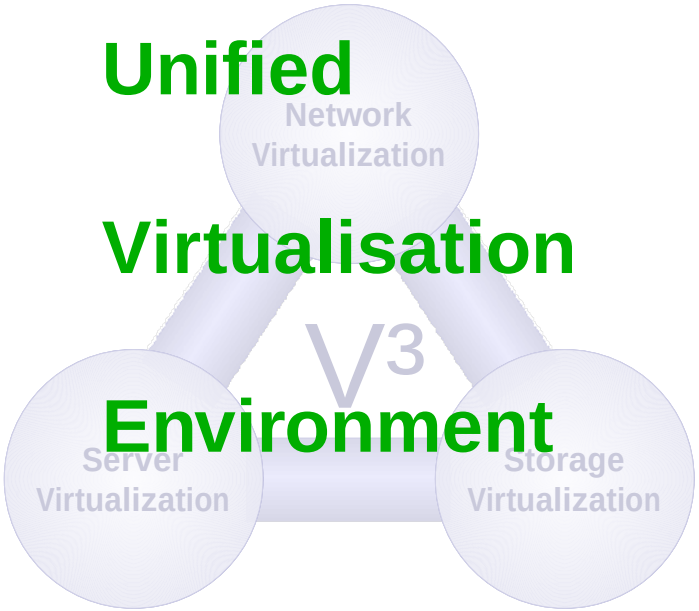
Highlights of the New OSL-Technology for LAN-Attached (Shared) Block Devices

- New block-i/o protocol developed by OSL
- Direct transport of all relevant i/o calls (read, write, ioctl)
- Relies on own frames ⇒ design and operation not limited to ethernet and IP
- Integates connection setup, monitoring, path multiplexing and trunking
- Capable of self configuration and error recovery
- Can handle all modern storage szenarios:
  - simple server and multiple clients including multipathing
  - clusters of storage servers (targets)
  - clusters of storage clients (initiators)
  - integrated clusters of servers and clients
  - storage server farms
  - cloud concepts
- Designed for seamless integration with storage virtualization
  - easy-to-use device names
  - fdisk (partitioning) no longer needed for clients
  - on-demand allocation and online reconfiguration
  - many useful functions
  - enables client-side administration
- In combination with OSL SC RSIO boasts LAN-free backup capabilities

Author: Bert Miemietz, Rev. 11/2013

OSL Unified Virtualisation Environment

Leading Edge Integrated Solution for the Software Defined Data Centre



Unified Virtualisation Environment

(SDDC-Solution)

Unified Virtualisation Server

Converged Networking

Unified Virtualisation Client

Author: Bert Miemietz, Rev. 11/2013

Server & OS-Virtualization

- Overview -

Author: Bert Miemietz, Rev. 11/2013

Classification of Virtualization

Virtualization Is Neither New nor Limited to Certain Fields of Information Technology

- Unix systems (and to a certain extent most other modern OS) will always provide a virtualized runtime environment for applications:
  - Virtual filesystem
  - Virtual address space
  - Hardware abstracted device handling
  - Processes, threads etc.
  - Isolated user and group environments

  This feature is sometimes (misleadingly) called "partial virtualization".
- There are many fields of virtualization:
  - Storage
  - Network
  - Hardware / Server
  - OS
  - Application Runtime Virtualization
  - Desktop ...
- We will focus on virtualization for compute resources
  (OS virtualization, server virtualization, virtual runtime rnvironments)

Author: Bert Miemietz, Rev. 11/2013

Basics of Virtualization

Hardware Abstraction, Improved Operation and Resource Utilization

- The main issue is to provide an additional abstraction layer to:
  - get a better system resource utilization (consolidation & over-provisioning)
  - facilitate operation / reduce portability issues
  - isolate several application runtime environments
  - isolate faults and improper workloads
  - create more powerful runtime environments (rare cases)

- Hardware Partitioning
  - IBM LPAR
  - SPARC Dynamic System Domains (Fujitsu M-Series)
- Server Virtualization (Hypervisor)
  - Full Virtualization — KVM, Xen, VBox, OVM, Hyper-V, VMWare
  - Partial Virtualization — MVS, BS2000, Unix/Linux
  - Paravirtualization — Xen, LDOMs
- OS Virtualization — Solaris Zones, AIX Workload Partitions, OpenVZ, Linux-VServer, LXC
- Virtual Runtime Environments — OSL Storage Cluster

Author: Bert Miemietz, Rev. 11/2013

Basics of Virtualization

Hardware Abstraction, Improved Operation and Resource Utilization

- The main issue is to provide an additional abstraction layer to:
  - get a better system resource utilization (consolidation & over-provisioning)
  - facilitate operation / reduce portability issues
  - isolate several application runtime environments
  - isolate faults and improper workloads
  - create more powerful runtime environments (rare cases)

Hardware

- Hardware Partitioning
  - IBM LPAR
  - SPARC Dynamic System Domains (Fujitsu M-Series)

Software

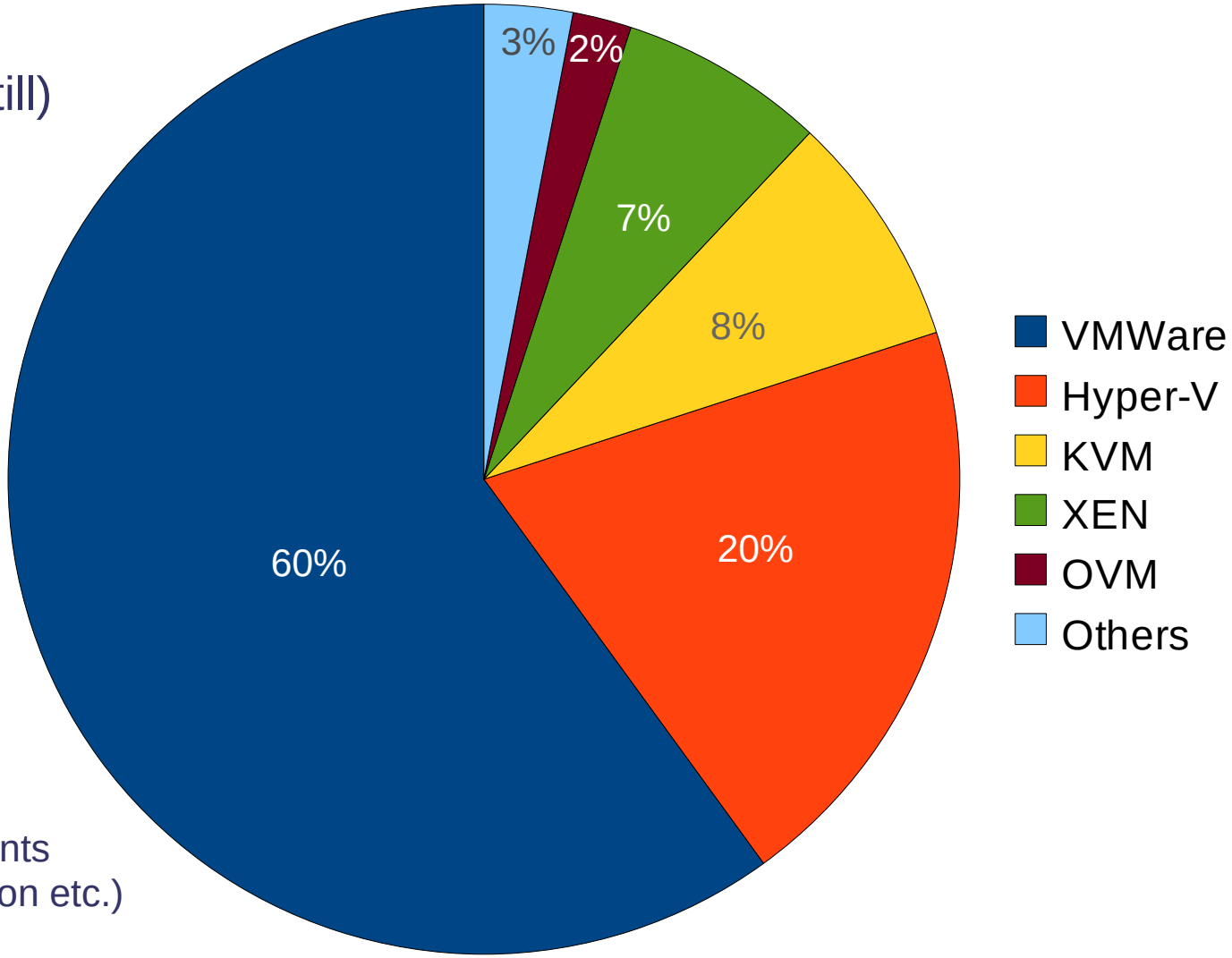
- Server Virtualization (Hypervisor)
  - Full Virtualization — KVM, Xen, VBox, OVM, Hyper-V, VMWare
  - Partial Virtualization — MVS
  - Paravirtualization — Xen, LDOMs
- OS Virtualization — Solaris Zones, AIX Workload Partitions, OpenVZ, Linux-VServer, LXC
- Virtual Runtime Environments — OSL Storage Cluster

Author: Bert Miemietz, Rev. 11/2013

Market Situation

There Seems To Be a Clear Situation

- Server Virtualization (still) dominated by VMWare
- KVM growing fastest (2012 - 2013: +50%)
- Things are changing at high speed
- Main differentiators:
  - **not** hypervisor features
  - integration into frameworks
  - adaption to new developments (hardware, storage integration etc.)
  - cloud capabilities
  - price



Estimations based on IDC 2013 (Al Gillen) and others:
http://readwrite.com/2013/05/02/idc-virtualizations-march-to-cloud-threatens-vmware
http://enterprisesystemsmedia.com/article/move-over-vmware-kvm-has-arrived
http://blogs.aberdeen.com/it-infrastructure/is-the-hypervisor-market-expanding-or-contracting/
http://wikibon.org/wiki/v/VMware_Dominant_in_Multi-Hypervisor_Data_Centers

The Cloud Effect

Cloud Concepts are Shifting Motivations

- First virtualization wave has been driven by the idea of **consolidation**

- Cloud concepts are driven by:
  (based on definition of the "National Institute of Standards and Technology"):

  - **scalability and rapid elasticity**
  - resource pooling (flexible pools with multi-tenant models)
  - reliability and fault tolerance
  - optimization and consolidation
  - measured service / QoS
  - On-demand self-service (self-provisioning, service on demand)
  - broad network access

- Server virtualization is more and more considered to be
  a component of cloud infrastructures

Author: Bert Miemietz, Rev. 11/2013

Why Should I Care?

Big Unix Servers Are Not the Only Possible Solution

- UNIX RISC vendors market share is constantly dropping
- Low-end tasks have been moving to x86
- Linux is becoming more and more an enterprise computing platform (SAP)
- x86-virtualization is nowadays used even for business critical databases
- awful design of most applications requires dedicated systems
- Current UNIX servers are too big by far for most single application workloads:

	CPU Type	Clock (MHz)	nCPU	Cores	Threads	RIPmix*
Intel Pentium Generic	Intel Pentium	100	1	1	1	1
Intel Primergy RX100S5	Intel Xeon E3110	2992	1	2	2	89
FSC Primepower 650	SPARC64 V	2159	8	8	8	137
HP ProLiant DL580 G7	Intel Xeon E7-8837	2667	4	32	32	246
SPARC Enterprise M4000	SPARC64 VII	2400	4	16	32	347
Fujitsu RX350 S7	Intel Xeon E5-2630L	2000	2	12	24	419
Fujitsu M10-1	SPARC 64 X	2800	1	16	32	615
SPARC Enterprise M8000	SPARC64 VII	2520	8	32	64	670
SPARC T4-2	SPARC T4	2848	2	16	128	762
SPARC T4-4	SPARC T4	3000	4	32	256	1591
Fujitsu M10-4	SPARC 64 X	2800	4	64	128	(?)

* RIP – relative integer performance is used as an indicator for proper workload assignment by the OSL Storage Cluster engine, mix is the geometric average of 32 and 64 bit integer performance

Author: Bert Miemietz, Rev. 11/2013

Why Should I Care?

Big Unix Servers Are Not the Only Possible Solution

- UNIX RISC vendors market share is constantly dropping
- Low-end tasks have been moving to x86
- Linux is becoming more and more an enterprise computing platform (SAP)
- x86-virtualization is nowadays used even for business critical databases
- awful design of most applications requires dedicated systems
- Current UNIX servers are too big by far for most single application workloads:

typical single system int. performance for SAP/DB in Germany

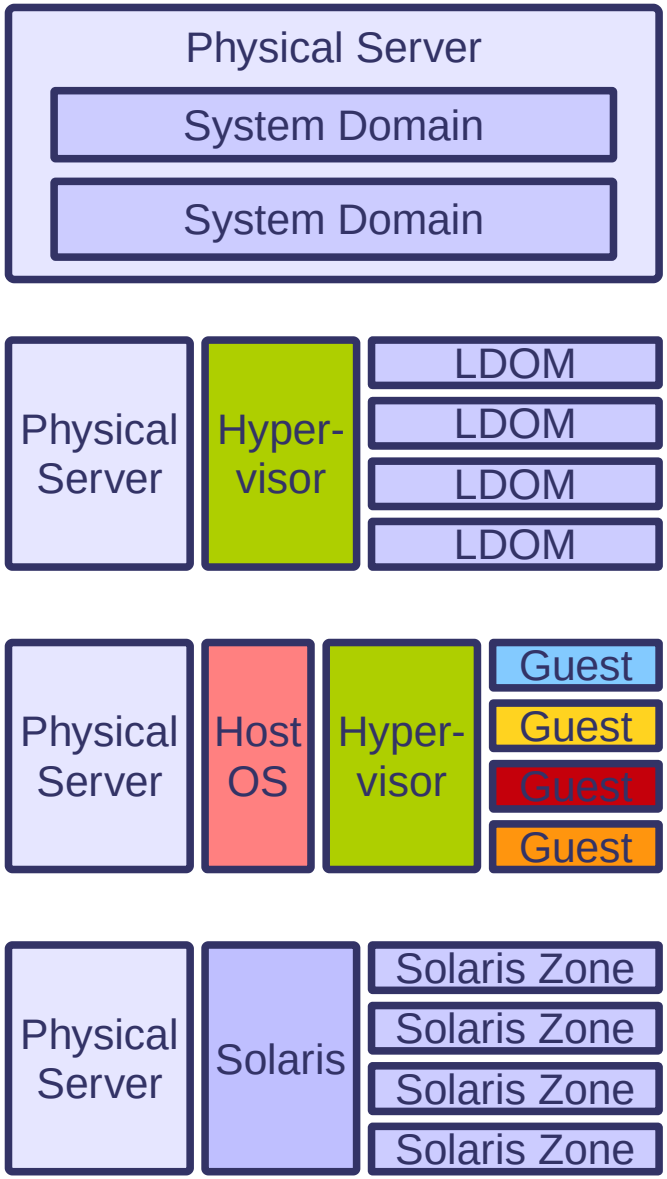
	CPU Type	Clock (MHz)	nCPU	Cores	Threads	RIPmix*
Intel Pentium Generic	Intel Pentium	100	1	1	1	1
Intel Primergy RX100S5	Intel Xeon E3110	2992	1	2	2	89
FSC Primepower 650	SPARC64 V	2159	8	8	8	137
HP ProLiant DL580 G7	Intel Xeon E7-8837	2667	4	32	32	246
SPARC Enterprise M4000	SPARC64 VII	2400	4	16	32	347
Fujitsu RX350 S7	Intel Xeon E5-2630L	2000	2	12	24	419
Fujitsu M10-1	SPARC 64 X	2800	1	16	32	615
SPARC Enterprise M8000	SPARC64 VII	2520	8	32	64	670
SPARC T4-2	SPARC T4	2848	2	16	128	762
SPARC T4-4	SPARC T4	3000	4	32	256	1591
Fujitsu M10-4	SPARC 64 X	2800	4	64	128	(?)

* RIP – relative integer performance is used as an indicator for proper workload assignment by the OSL Storage Cluster engine, mix is the geometric average of 32 and 64 bit integer performance

Author: Bert Miemietz, Rev. 11/2013

Oracle Portfolio for OS- & Server-Virtualization

Several Technologies

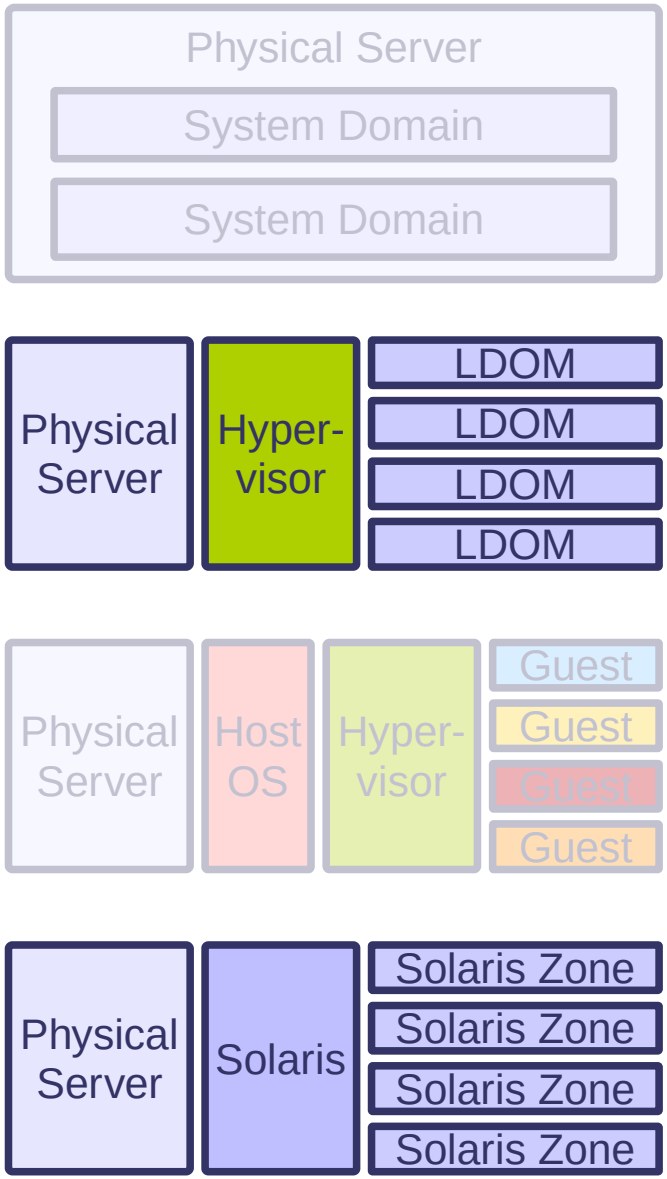


Technology	Platform
Physical Partitioning / Dynamic System Domains	SPARC M-Series
Oracle VM Server for SPARC	SPARC T-Series + M10
Oracle VM Server for x86 and Oracle VM VirtualBox	x86
Oracle Solaris Zones	all Solaris platforms SPARC + x86

Author: Bert Miemietz, Rev. 11/2013

Oracle Portfolio for OS- & Server-Virtualization

Several Technologies and Our Focus



Technology	Platform
Physical Partitioning / Dynamic System Domains	SPARC M-Series
Oracle VM Server for SPARC	SPARC T-Series + M10
Oracle VM Server for x86 and Oracle VM VirtualBox	x86
Oracle Solaris Zones	all Solaris platforms SPARC + x86

Author: Bert Miemietz, Rev. 11/2013

Standalone Resource Control in Solaris

Author: Bert Miemietz, Rev. 11/2013

Projects and Tasks

Some More Complexity Beyond Standard Unix

- Projects and tasks are the most SVR4-like mechanisms for resource control as compared to ulimit() - yet, they are not portable:
  - extended attributes to processes (and their hierarchy), inherited by fork()
  - hence with a meaning to process groups and sessions
  - user and group dimension
  - processes can be manipulated with Solaris commands depending on project or task membership
- Projects
  - are assigned to users and groups (default projects)
  - can be managed across the network (DNS, NIS, LDAP)
  - cmds: login, setproject
- Tasks
  - group processes belonging to a project into manageable entities representing a certain workload component
  - cmds: login, setproject, newtask

Resource Controls

A Means To Handle Some More Limits

- Resources controlled are e.g.:
  - standard Unix rlimits
  - processes and LWPs
  - IPC objects
  - CPU usage (shares, time, pools)
  - memory usage
  - zone related attributes
- Controls / Constraints can apply to:
  - zones
  - projects
  - tasks
  - processes
- Resource controls are enforced in-time by the kernel (synchronous)
- Concepts, commands and administration tend to be somewhat complex
- Resource controls in combination with projects were mainly used in early Solaris 10 installations to facilitate some environment settings that had to be made via /etc/system in former releases

Author: Bert Miemietz, Rev. 11/2013

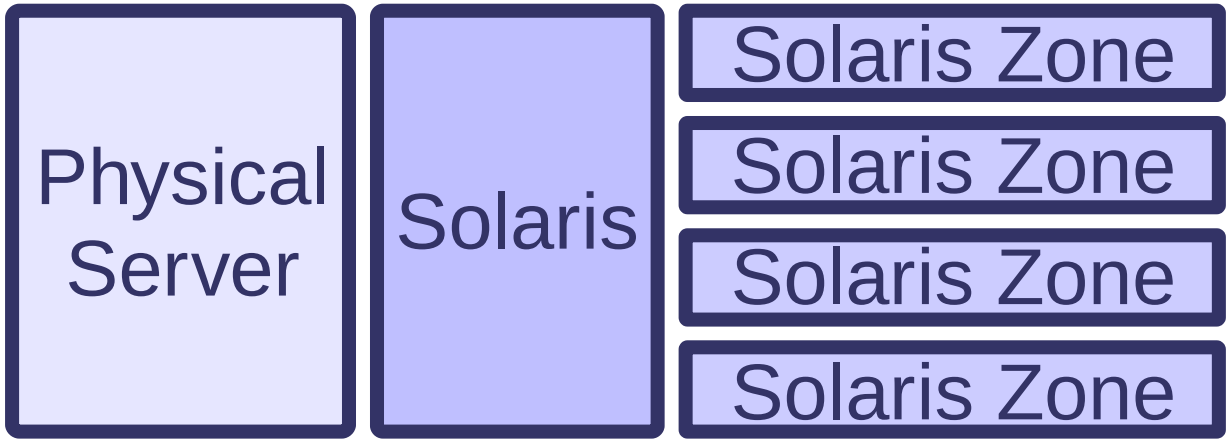
Control Mechanisms for CPU and RAM

If Things Are Not Complicated Enough

- Processor Sets
  - a means of partitioning available number of VCPUs into almost independent sets of VCPUs
- FSS – Fair Share Scheduler
  - developed as SUN's alternative to the standard Unix timesharing mechanism
  - tries to assign CPU slices according to the importance of certain workloads
  - can be combined with processor sets
  - increases system complexity – in our experience not suitable for smaller systems
- Resource Capping Daemon (rcapd)
  - additional limitations with the "resource caps" mechanism
  - resource caps are enforced by rcapd at user level with some delay (asynchronous)
  - controls physical memory usage / RSS (resident set size) of zones or projects
- Resource Pools
  - mainly targeted at CPU resources
  - used in combination with processor sets
  - can be used with floating or pinned CPUs
  - controlled by poold
- All this stuff tends to turn out as very complicated in daily routine
  ⇒ most customers today prefer easier mechanisms provided by zones

Author: Bert Miemietz, Rev. 11/2013

Solaris Zones



Author: Bert Miemietz, Rev. 11/2013

Solaris Zones - Overview

A Slim OS Virtualization Technology

- OS virtualization technology introduced in 2005 with Solaris 10
  - main target: isolated and secure environment for running applications
  - non-global zones provide virtualized OS environments within a global Solaris instance
  - limited to Solaris
- Available on SPARC and x86 platforms
- Zones share the same kernel
- A mixture of chroot and other process attributes maintained by the kernel with special extensions of the operating system
- Little overhead, in theory more than 8000 zones per OS instance
- In combination with resource controls and resource caps the illusion of a VM becomes even more complete
- Zone concepts get more and more bound to the ZFS filesystem services
- Solaris zones have become the preferred (Solaris) virtualization technology

Author: Bert Miemietz, Rev. 11/2013

Solaris Zones – What They Can Do

Enablement of Different Features

- Combine several applications in isolated environments on the same server
- Encapsulation of vulnerable (security!) or instable applications
- Delegated administration
- Different versions of the same program on the same machine
- Hide physical host environment
- Granular assignment of resources
- Virtualization with almost no performance impact on running applications

Solaris Zones – And What Problems Might Occur

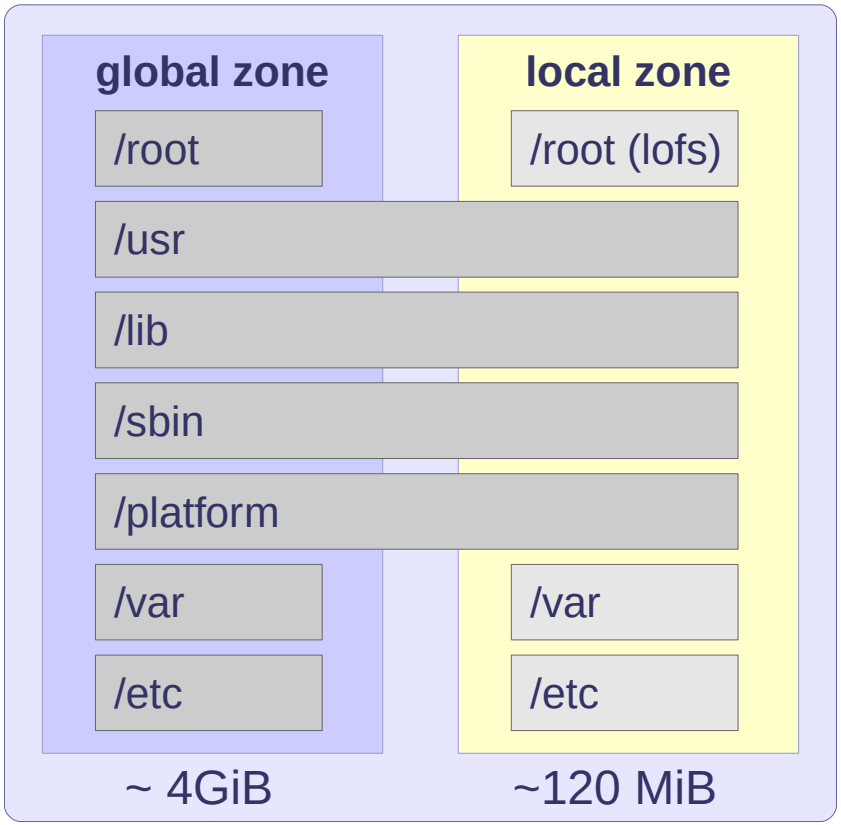
- Can't load another OS, individual drivers and/or individual devices
- Cannot prevent panics or hung systems
- In our experience still some (rare) problems with overload situations
- Global, non-global and other non-global zones have some links

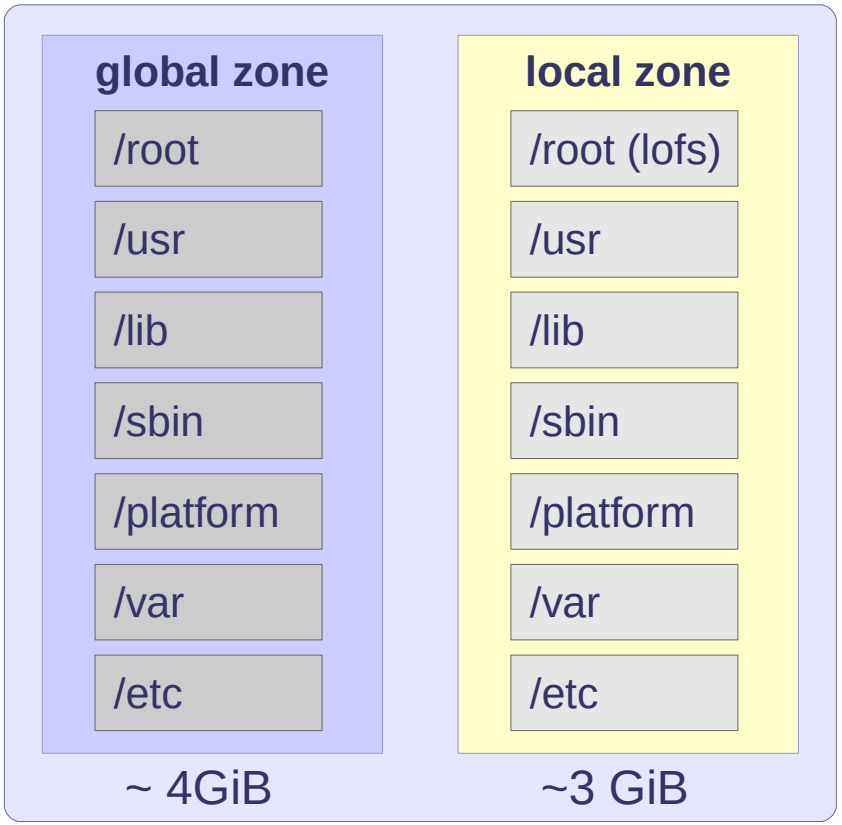
Author: Bert Miemietz, Rev. 11/2013

Types of Zones (I)

File System Layout - Depending on Your Choice and on OS Version

- Differentiate by scale of filesystem sharing with global zone (Solaris 10):
  - sparse root zone (left picture)
  - whole root zones -> enable unlimited package installation





- The sparse zone concept has been abandoned in Solaris 11
  ⇒ expect issues when migrating from Solaris 10 to Solaris 11!

Author: Bert Miemietz, Rev. 11/2013

Types of Zones (II)

Several Ideas and Concepts - Choose a Proper Setup

- Global <-> non-global (local) zones
- Shared IP <-> Exclusive IP
- Readonly Zones (immutable zones)
- Branded Zones
  - can run another OS version than the global zone
  - mainly interesting for running Solaris 10 environments on Solaris 11 Host
  - expect problems with statically linked libraries
- Trusted / Labeled Zones
  - special security enhancements (Solaris Trusted Extensions environment)
- Zones on Shared Storage (ZOSS)
  - require a dedicated zpool
  - facilitate moving zones between hosts
- Zoneroot on ufs <-> zoneroot on zfs / zpool

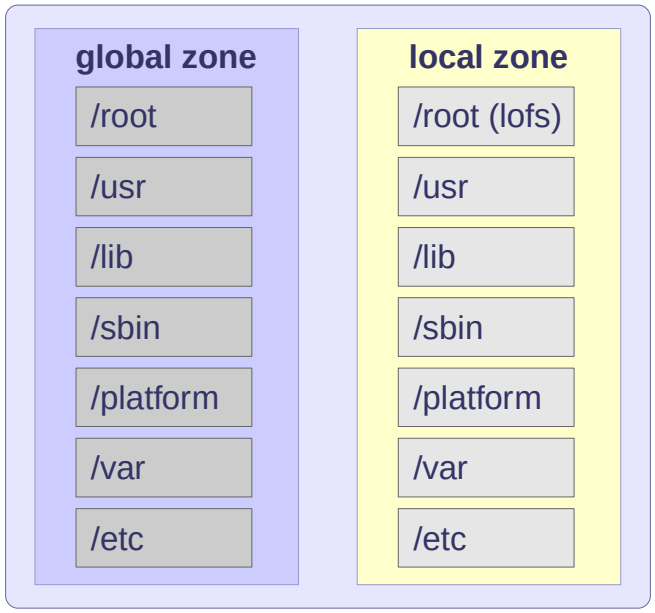
Zoneroot on ufs is no longer supported in Solaris 11!

Author: Bert Miemietz, Rev. 11/2013

Zones in Solaris 11

The Picture Becomes Clearer

- only whole-root zones are supported
  ⇒ consider readonly zones for migration of sparse zones
- zones use IPS instead of SVR4-packages for setup
  ⇒ all repositories must be accessible during zone creation
  ⇒ minimized package set for starting zones
- zones must be located on a zfs dataset
- zones use "boot environments" (see beadm(1m))
- Default brand changed from "native" to "solaris"
- Zones can run CIFS and NFS
- enhanced functionality for ZOSS
  (rootzpool resource / suri / suriadm)
- Several more changes in details



Whole-root is mandatory in Solaris 11

Author: Bert Miemietz, Rev. 11/2013

Zones and Resource Control

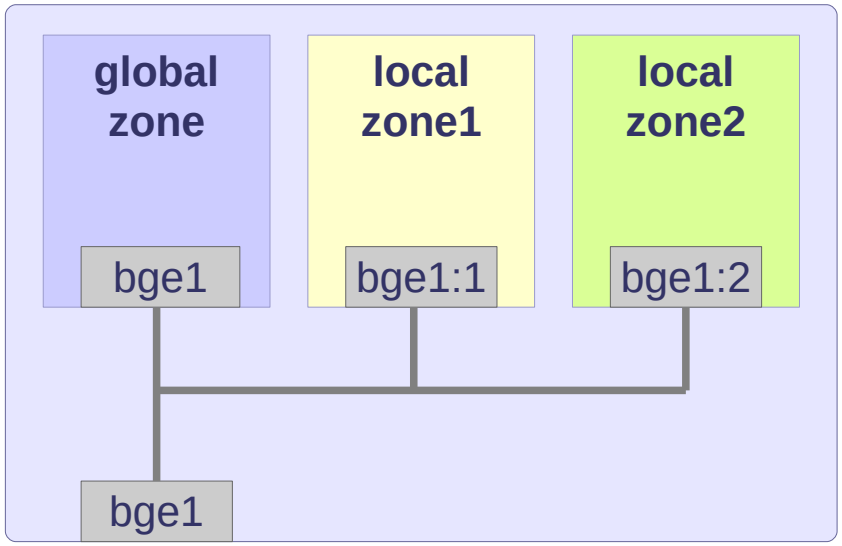
Almost Everything Possible – From KISS to Complicated

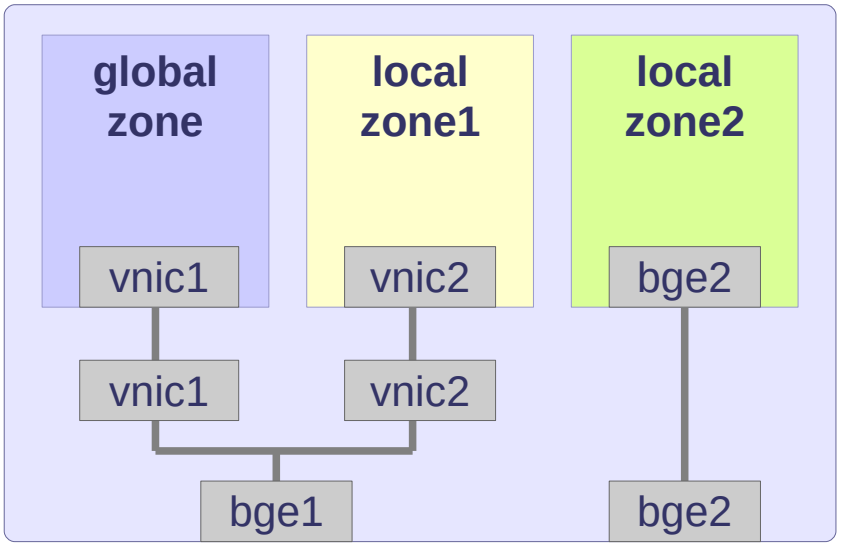
- resource pools / cpu pools
- capped cpu
- dedicated cpu
  - almost vm behaviour
  - easiest way of partitioning cpu power
  - can solve some licence problems (limited number of CPUs visible)
- control memory: - physical memory (max-rss)
  - swap
  - locked memory
- control IPC mechanisms via projects

Author: Bert Miemietz, Rev. 11/2013

Zones and Networking

Shared and Exclusive IP



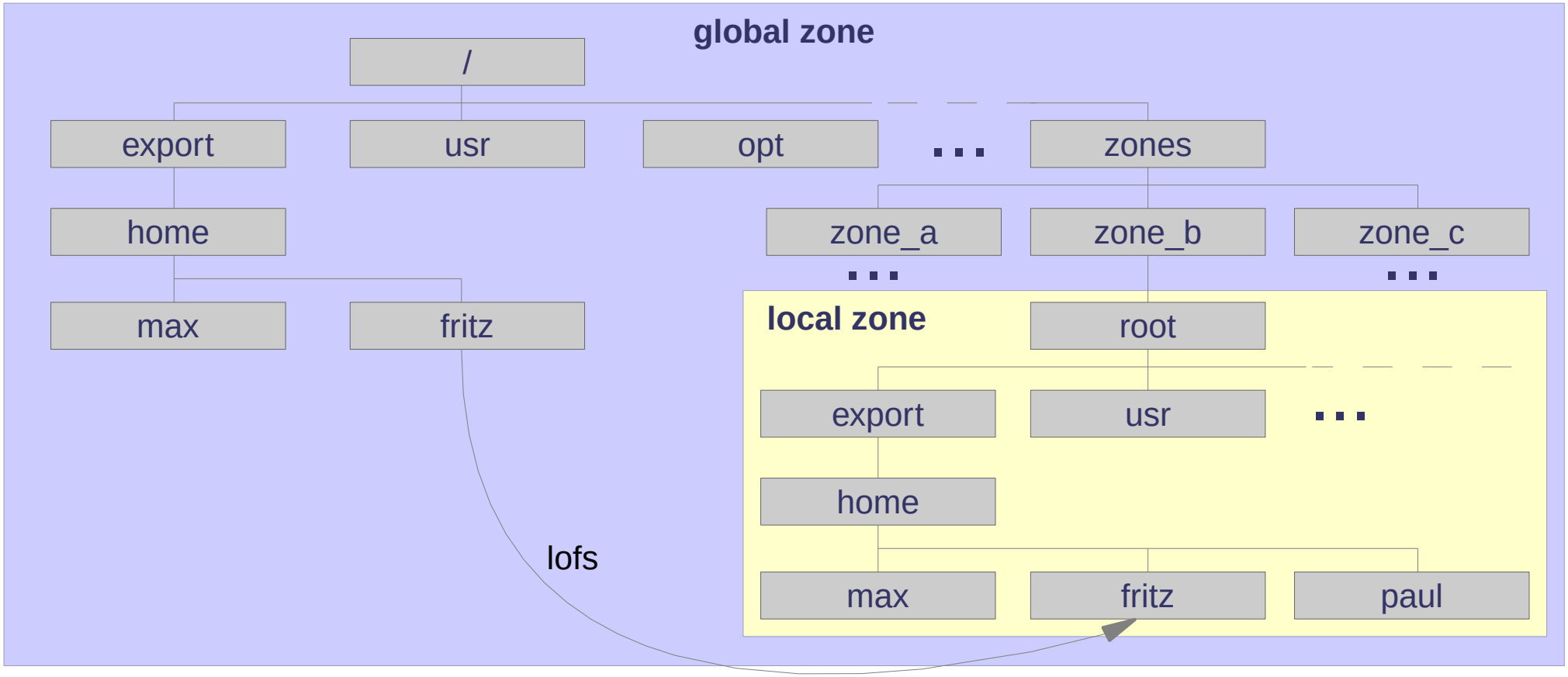


- Shared IP shares IP layer with the global zone and its interfaces
- Exclusive IP have an own instance of the IP layer
  - either on a dedicated physical interface
  - or on a dedicated VNIC (Solaris 11)
- Additional features with exclusive IP:
  - IP routing
  - ipfilter and NAT
  - IPMP
  - DHCP / IPv6 autoconfiguration

Author: Bert Miemietz, Rev. 11/2013

Zones and Storage
Concept Differs From Other VM Types

- Hypervisor VMs run their own fs-instances on devices
- Zones *do not* run their own driver instances ⇒ they depend on the “hypervisor's” / global zones filesystems

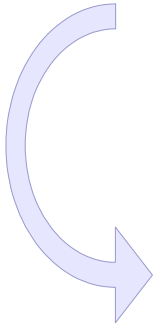


- Raw devices can be passed into zones / each device needs configuration

Zones and ZFS

Merged Layers Instead of Multi-FS Modular and Layered Design

- ZFS has had an increasing influence on many operational interfaces of SunOS
- Whereas early versions of SunOS 5.10 did not support zfs as root fs
  Solaris 11 now enforces zfs as root fs
- Advantages of zfs:
  - availability of all zfs features including snapshots
  - allows easy cloning of zones
  - support of boot environments
  - fine grained permissions
  - delegation
- ZFS datasets can be delegated to administration by non-global zone root
- Issues with ZFS:
  - no cluster awareness
  - shared devices can result in panics (not a cluster fs)
  - externally created copies cannot be imported on same host
  - “magic” replaces well-known Unix interfaces / no external config



Expect problems with snapshot / data copy RAIDs
(Eternus, EMC, Netapp, IBM DS ...)

Configuration of Zones
Interfaces

- Configuration is done via cli-utility zonecfg(1M):
  - create / destroy zone configurations
  - add / remove resources and set resource properties
  - query configurations
  - roll back to a previous configuration
  - rename zones
- There are many script solutions around zones
- Difficulties: - complex tasks
  - moving zones between nodes
  - zones in clustered environments
  - device handling
  - any dynamic configuration changes

Author: Bert Miemietz, Rev. 11/2013

Migrations and Zones

The Right Path Can Make Things Much Easier

- Important paths:
  - Move Solaris 10 zone from Solaris 10 to Solaris 11
  - Convert Solaris 10 system to Solaris 10 zone on Solaris 11
  - Convert Solaris 11 system to Solaris 11 zone on Solaris 11
- There is a **zonep2vchk** (1M) utility to assist possible migrations
- Moving zones from Solaris 10 to Solaris 11 is a quick path to combine both:
  - minimum changes in OS runtime environment
  - use of Solaris 11 improvements on modern machines (performance, memory ...)

Author: Bert Miemietz, Rev. 11/2013

Zones and Applications

Zones Turned Out as the Standard Virtualization Technology

- Today almost all customers use zones as a VM-like virtualization
- Resources can be controlled
- “Dedicated CPU” is preferred way of partitioning (licence issues)
- Entire environment for an application can be created (IPC etc.)
- “Full machine control” is most comfortable for many application administrators
- Standard conforming applications can be run in zones,
  few exceptions e.g. with branded zones and statically linked libraries
- Major challenges:
  - Applications with massive raw device access
  - Clustered environments / clusterwide application control
  - OS-dependent applications
  - Lifecycle management

Author: Bert Miemietz, Rev. 11/2013

Why OSL SC in Solaris Environments?

Advantages With the Storage Virtualization of OSL Storage Clusters i. a.:

- Storage virtualization and cluster in a single integrated product
  ⇒ sophisticated design but easy to use
- Global storage pool – enterprise storage directory
- Very much simplified device handling
  - global devices / global namespace / arbitrary device names
  - all storage connectivity solutions from SCSI / iSCSI to FC and Infiniband in the same simple scheme
  - integrated easy-to-use multipathing, dynamic hardware reconfiguration capabilities
  - identical administration from Solaris 7 to Solaris 11, Sparc, x86 and even Linux
  - **also available for zones and LDOMs**
- Automated disk access management
  - in general tremendous security increase in clustered environments at no admin costs
  - zfs can be used safely in shared storage and cluster environments
- Application and VM awareness
  - application specific automated operations (mirroring, backup-to-disk, backup-to-tape, DR)
  - global views and reports on storage pool usage grouped by applications / VMs
  - allocation und bandwidth control by applications
- Leading edge performance and bandwidth control
  - no appliance - no bottleneck / at-will scalability of throughput and availability
  - capable of bandwith control (per volume and per application/VM)

Author: Bert Miemietz, Rev. 11/2013

Zones in OSL SC – Very Much a Simple VM

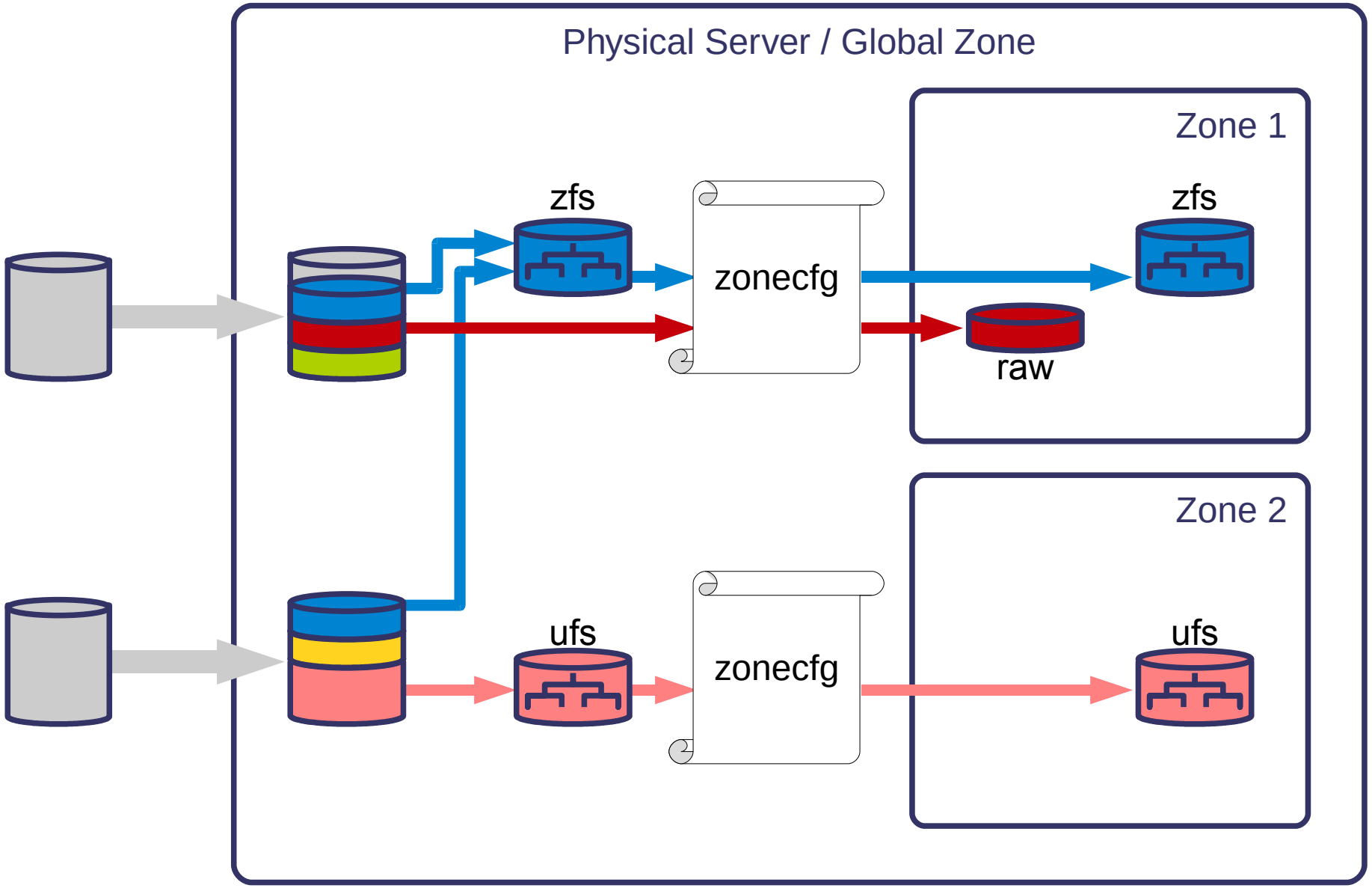
Zones in an Integrated Concept

- Creation and resource control very simple with vmadmin
  - # vmadmin -c vm_name -F {lkvm | lfxen | **szone** | ...}
  - assign cpu / memory / storage
- Start / stop / failover by standard Storage Cluster mechanisms
- Automated creation / installation via menu system or cli utility zone_install
- Additional details available with standard Solaris interfaces (zonecfg)
- Solaris zones in OSL Storage Cluster for all the time since 2006 have been “zones on shared storage” with many additional protection mechanisms and a compelling failover concept
- Backup to disk / tape integrated (dvam-tools)
- Of course excellent failover features
- Automated creation of all needed network configurations (hardware abstracted)
- Can be used for migration of zones from Solaris 10 to Solaris 11

Author: Bert Miemietz, Rev. 11/2013

Zones in OSL SC as a Standard VM

A Closer Look at Disk Devices and File Systems



Zones in OSL SC as Virtual Node

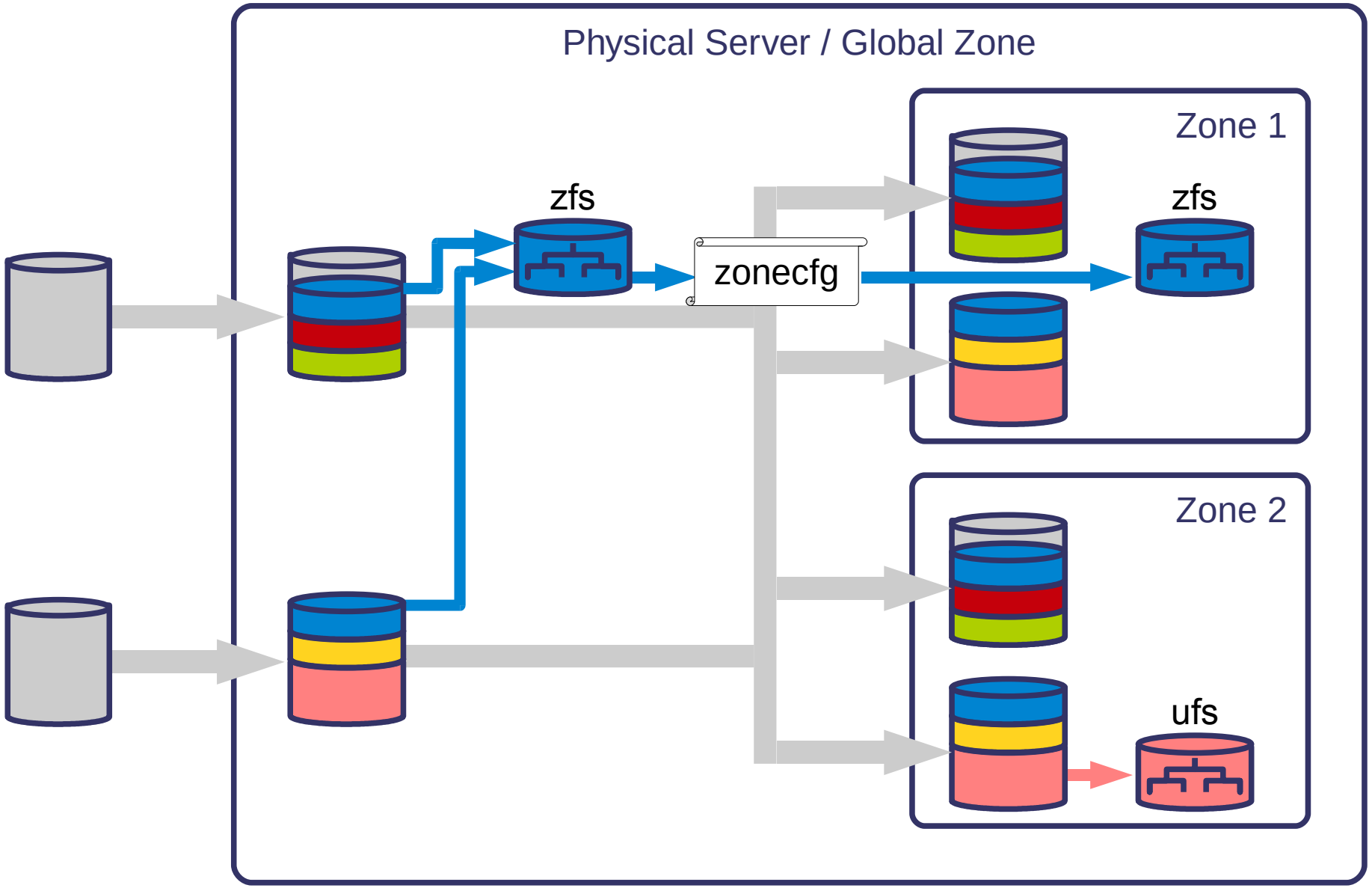
Zones as Cluster Nodes

- OSL Storage Cluster as of version 4.0 can be installed and run inside zones
- Zone is becoming a cluster node
- This permits application control into zones
- At-will migration of applications between hosts and zones (p2v, v2v, v2p)
- Zone failover on the same host, between domains and/or hosts
- Full access to storage virtualization from the inside of zones
- Application specific devices + disk access management for zones (ASM!)
- Applications running in zones can control mirrors, administer multipathing etc.
- Zone can mirror itself
- Zone can (almost) be administered like a physical host (vfstab ...)

Author: Bert Miemietz, Rev. 11/2013

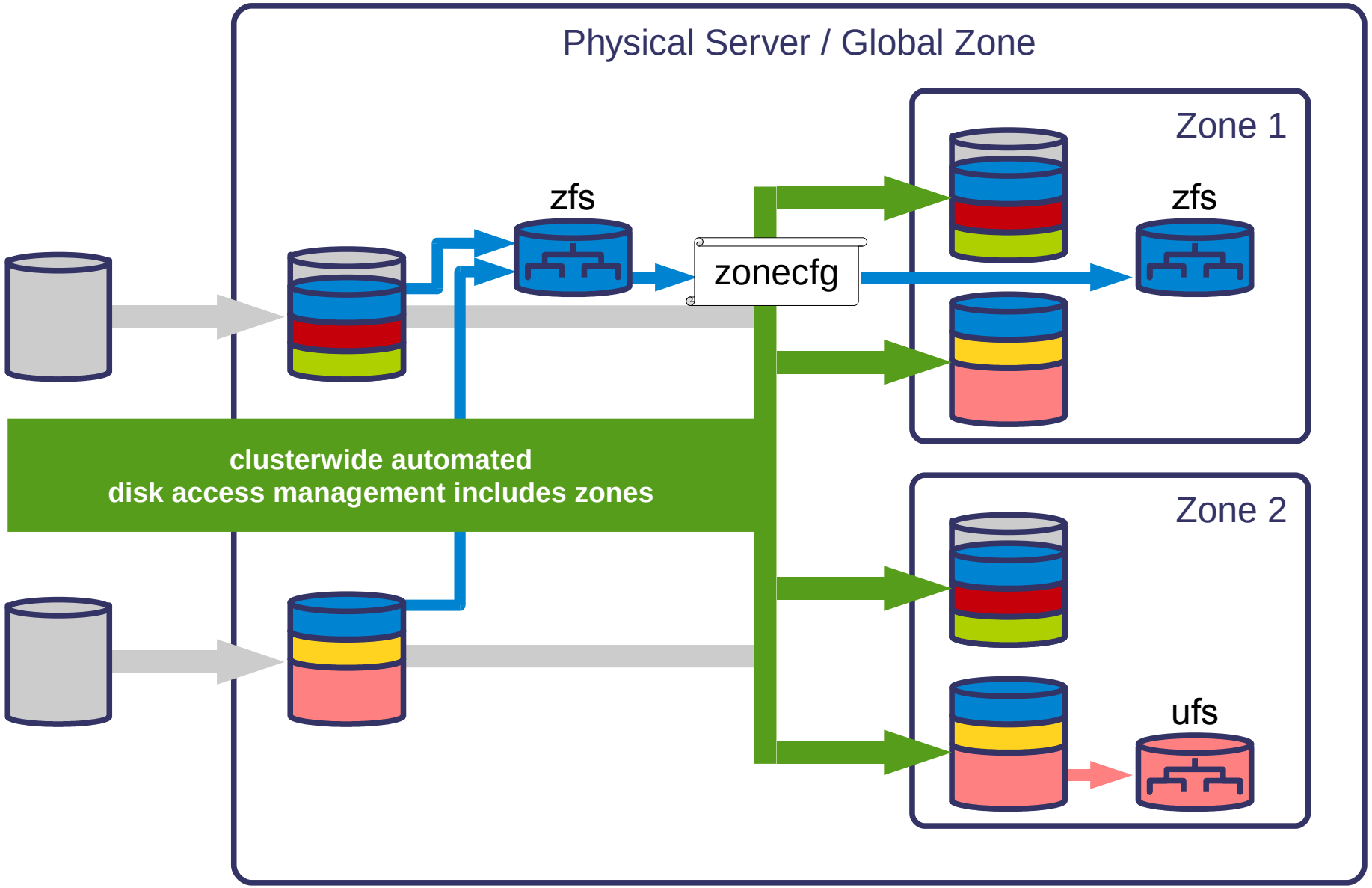
Zones in OSL SC as Virtual Node

A Closer Look at Disk Devices and File Systems



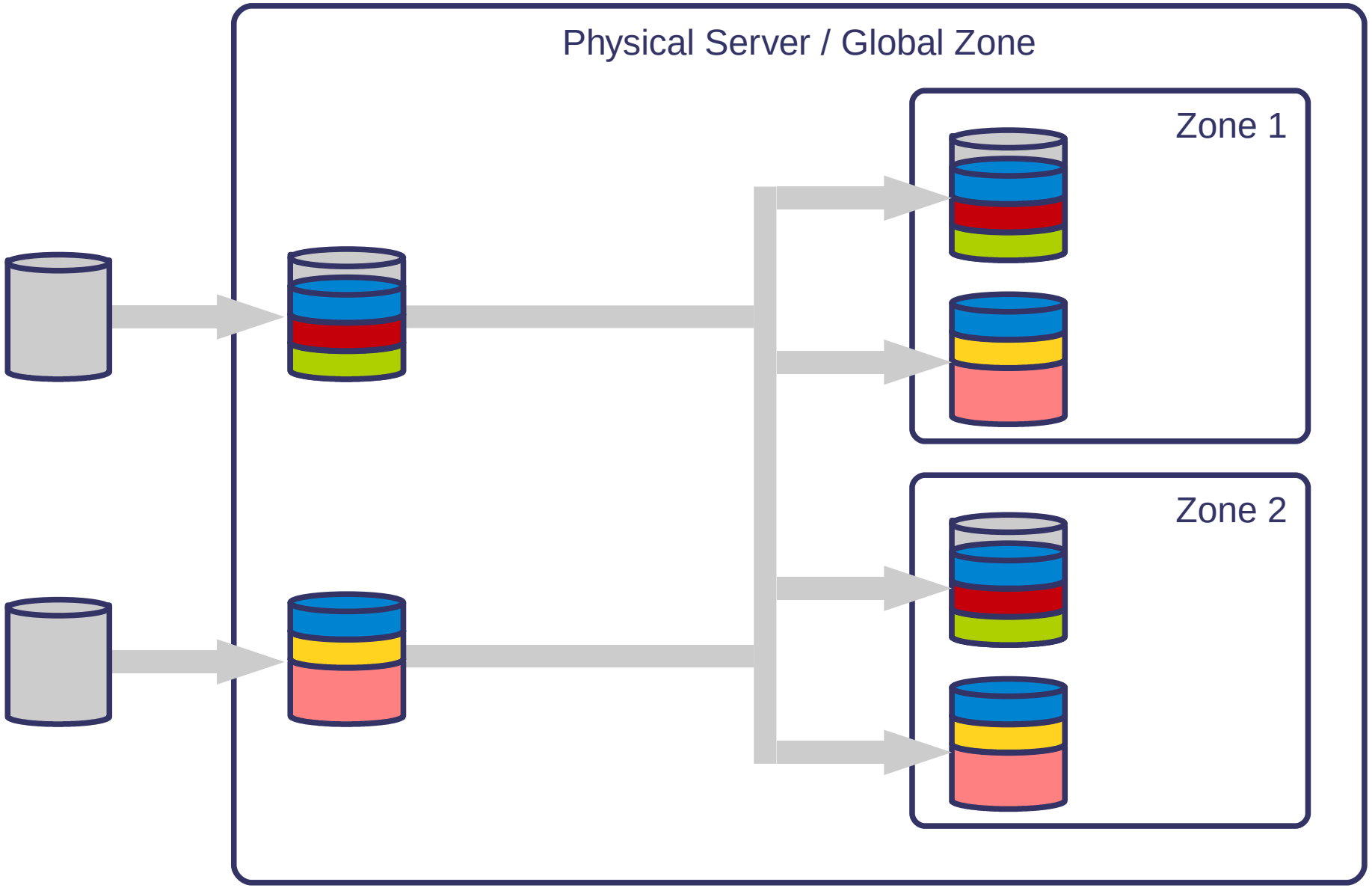
Zonen in OSL SC as Virtual Node

Ease of Use and Improved Security not only with Raw Devices



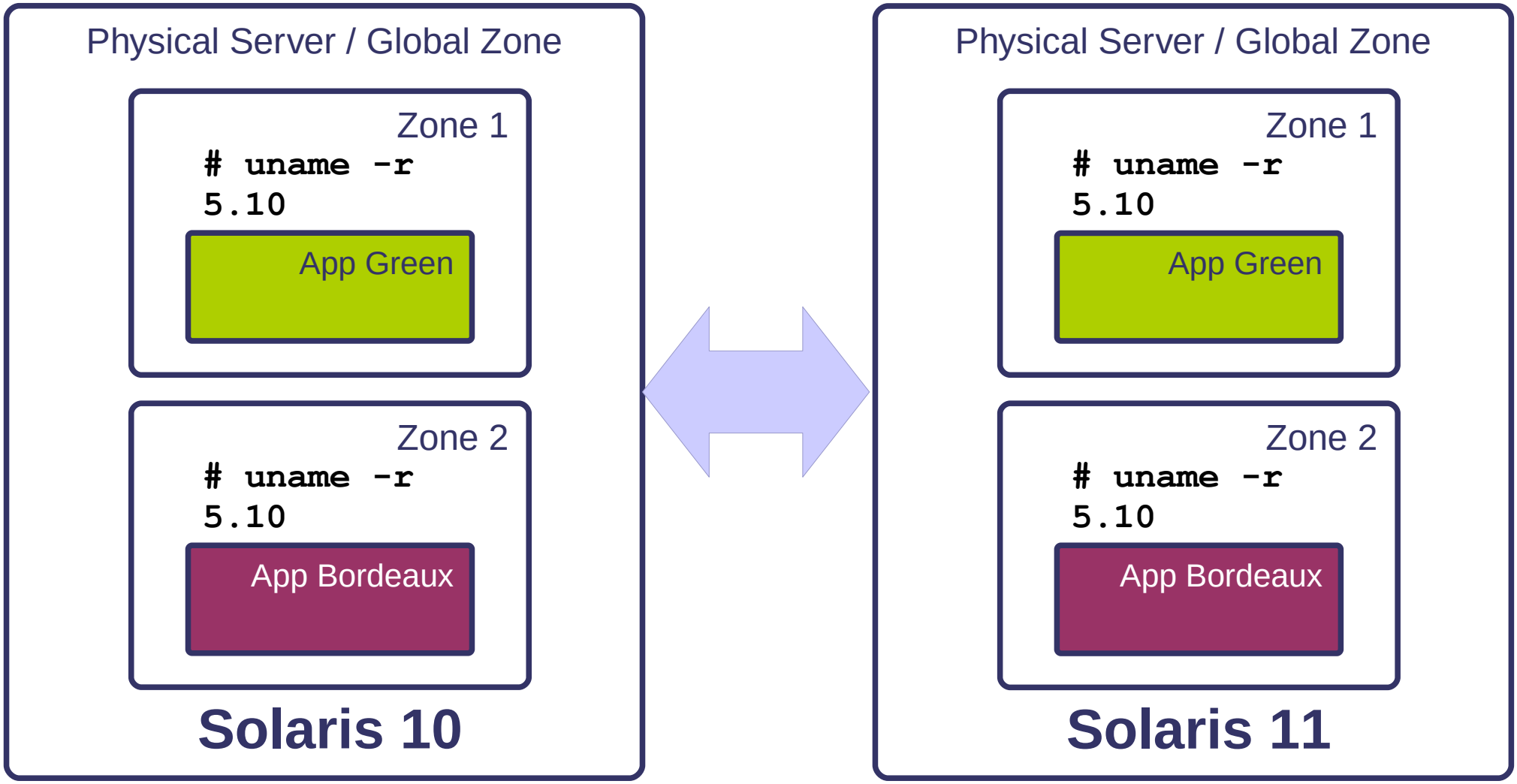
Zones in OSL SC as Virtual Node

Unified Device Representation on All Layers



Zones in OSL SC - Migration to Solaris 11

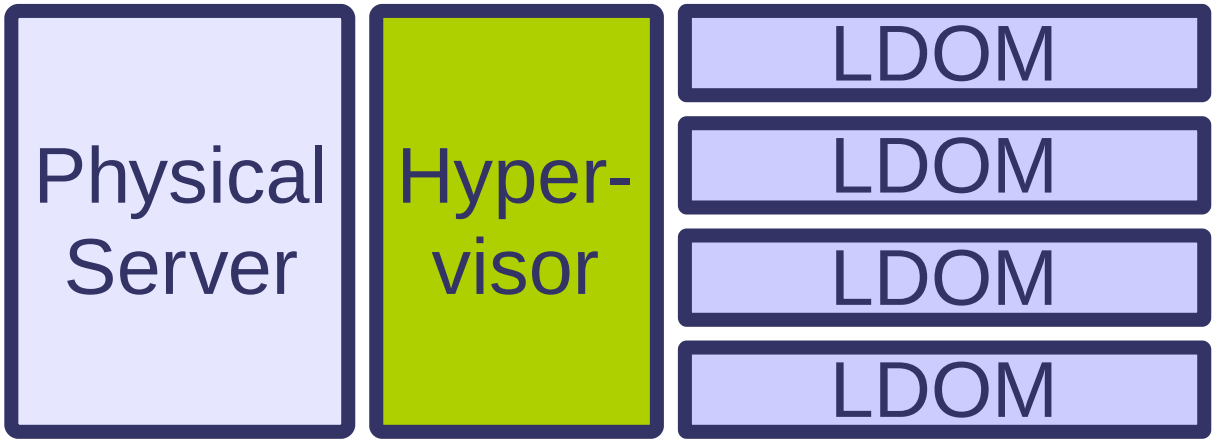
Branded Zones for a Start-Smart-Migration in Small Steps



- Use advantages of Solaris 11
- No or almost no changes in application runtime environment

Author: Bert Miemietz, Rev. 11/2013

Oracle VM Server for SPARC (LDOMs)



Author: Bert Miemietz, Rev. 11/2013

Oracle VM Server for SPARC / LDOMs - Overview

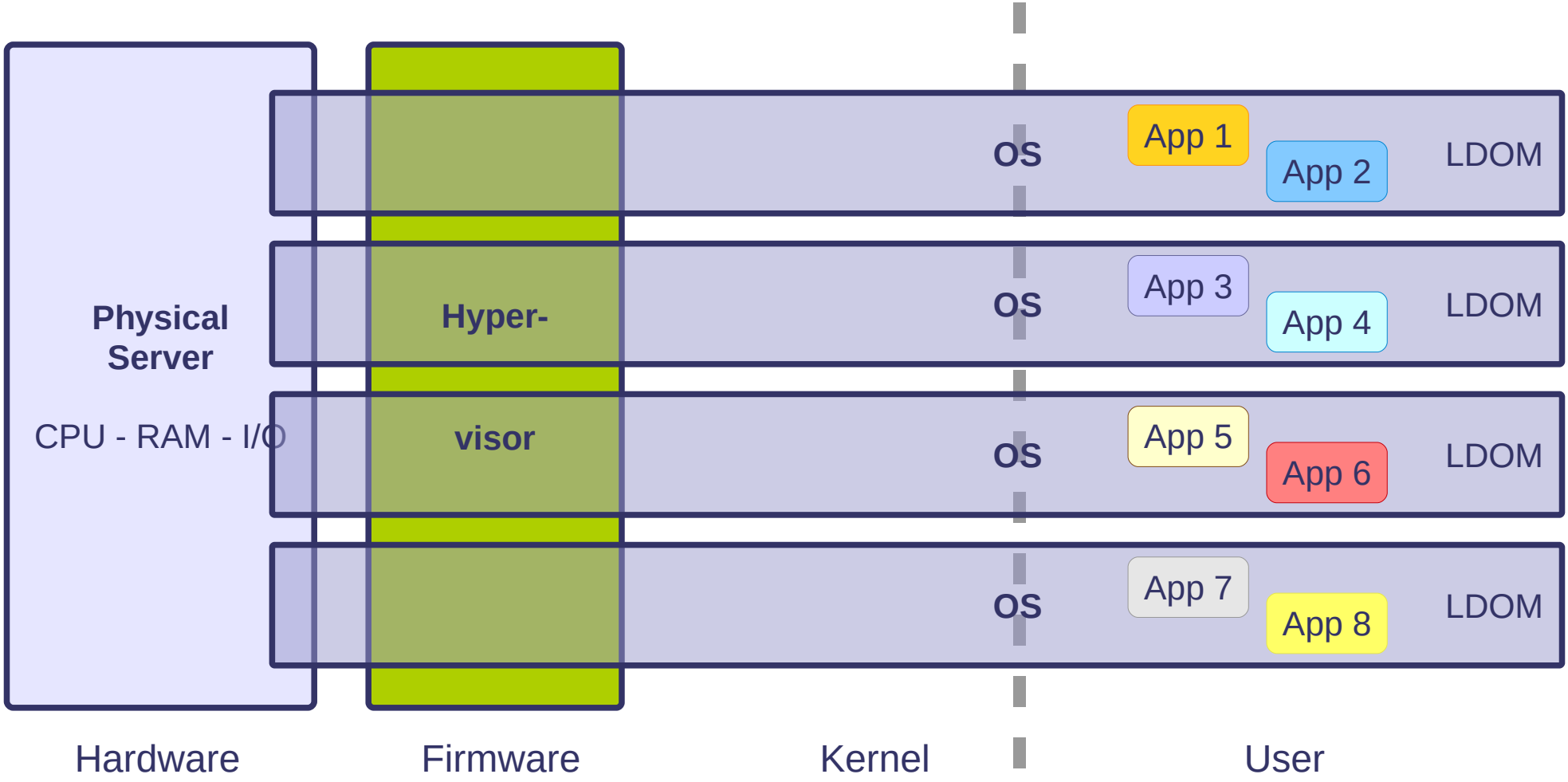
Key Facts

- Not really a new concept
  - first became visible in Germany with T2000 systems (~ 2007)
  - has been ignored by customers in production environments
  - got a strong push with T4 systems
- Slim overhead paravirtualization
- Comes at no additional costs
- Much finer granularity as compared to M-Series physical partitioning
- Dynamic reconfiguration
- Optionally redundant virtual I/O
- Static direct I/O
- Live-migration capabilities
- Can be used to optimize application licences

Author: Bert Miemietz, Rev. 11/2013

LDOMs - A Closer Look

Architecture



LDOMs - Definitions

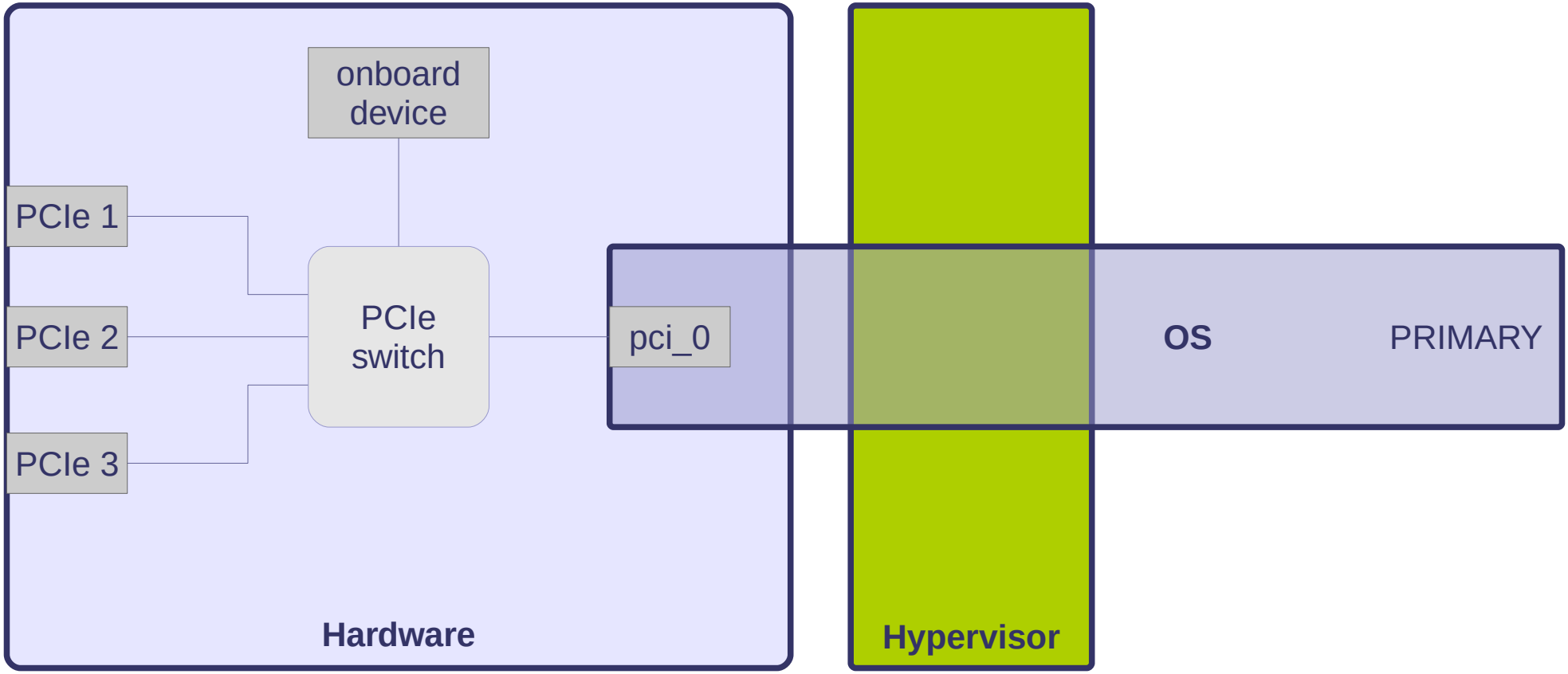
Basic Terminology

- (Primary) control domain
  - first domain in the system, cannot be removed, only one (!) control domain
  - used for creation and management of other logical domains
  - assigns physical resources
  - runs the Logical Domain Manager (LDM)
- Service domain
  - provides virtual I/O-services for guest domains
- I/O-domain
  - has direct physical access to i/o-devices:
    - PCIe root complex ⇒ "**root domain**"
    - PCIe slot or onboard device with direct I/O (DIO)
    - PCIe SR-IOV virtual function
- Guest domain
  - provides compute power
  - consumes virtual device services provided by a service domain
  - might have live migration capability

Author: Bert Miemietz, Rev. 11/2013

Control Domain
The Management Hub

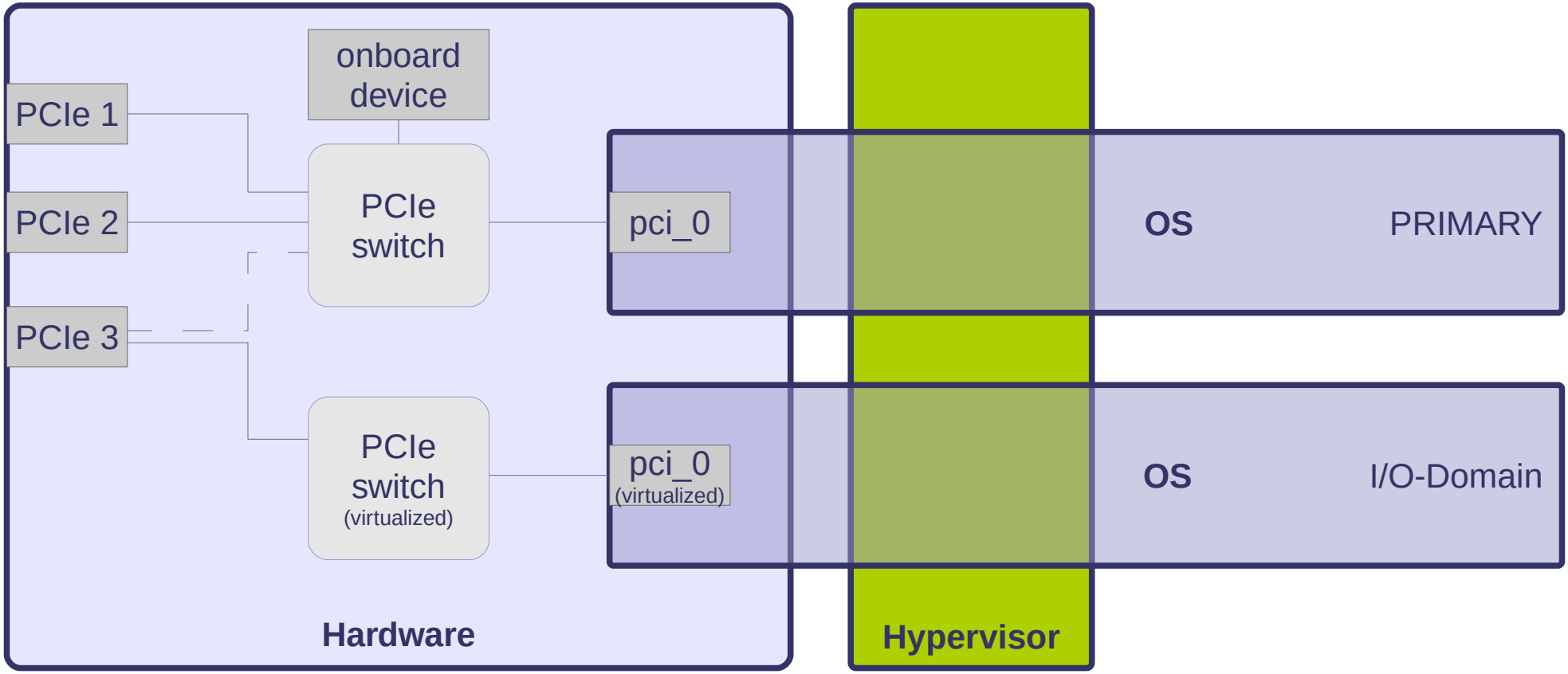
- (Primary) control domain
  - first domain in the system, cannot be removed, only one control domain
  - used for creation and management of other logical domains
  - assigns physical resources
  - runs the Logical Domain Manager (LDM)



I/O and Service Domains

Have Direct Access to I/O-Devices

- I/O-domain
  - has direct physical access to i/o-devices:
    - PCIe root complex ⇒ "**root domain**"
    - PCIe slot or onboard device with direct I/O (DIO)
    - PCIe SR-IOV virtual function
- Service domain
  - provides virtual I/O-services for guest domains

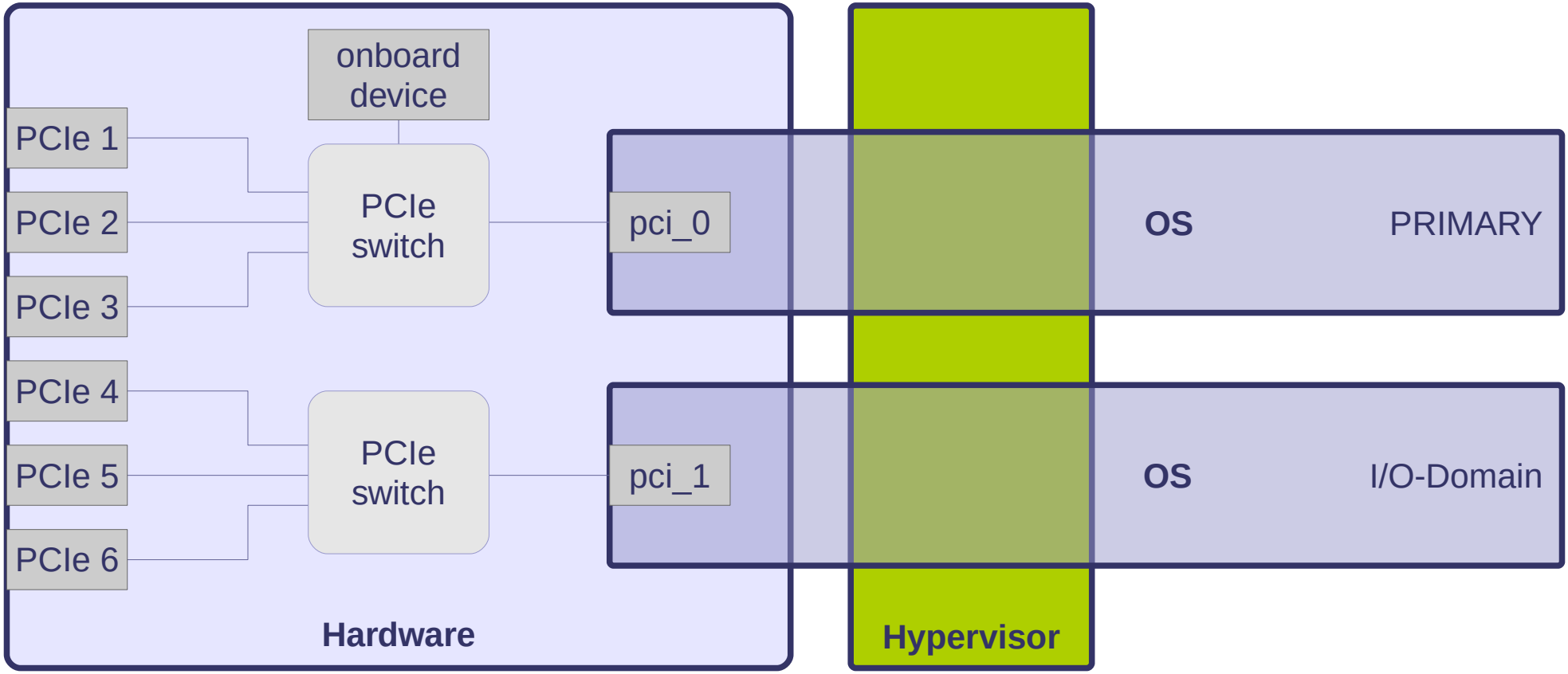


Author: Bert Miemietz, Rev. 11/2013

Root Domains

Have Direct Access to an Entire PCIe Root Complex

- A root domain is an I/O domain that owns an entire PCIe root complex:
  - owns a PCIe fabric
  - provides all fabric-related services incl. fabric error handling
- Maximum number of root domains depends on the platform:
  - T4-4 ⇒ up to 4 root domains
  - M10-1 ⇒ 2 PCIe switches | M10-4 ⇒ 4 PCIe switches (expansion units!)

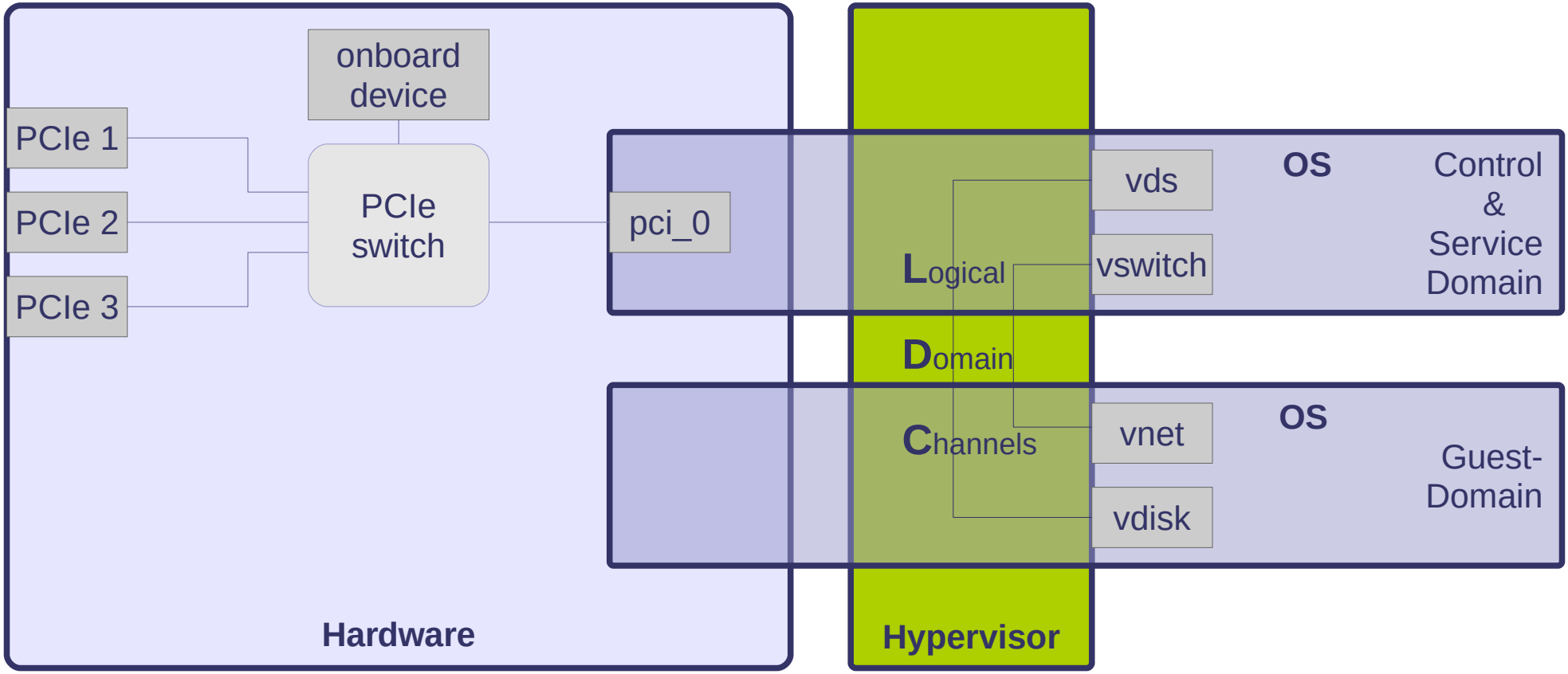


Author: Bert Miemietz, Rev. 11/2013

Guest Domains - The Entire Picture

How the Different Domain Types Co-operate

- A guest consumes the device services provided by a service domain:
  - i/o resources / virtual disk services
  - network resources (vswitch)
  - access is accomplished by logical domain channels
- By using virtual device services the guest domain gets live migration capabilities



Author: Bert Miemietz, Rev. 11/2013

Storage Backends for Virtual Disk Service

Several Possibilities

- LUNs (local or SAN and iSCSI)
- Flat files
- Logical volumes including ZFS
- Files on NFS (??? careful!!!)

Note that all devices consume LDCs:

(quoted from Oracle release notes for OVM/Sparc 3.1)

1. The control domain allocates approximately 15 LDCs for various communication purposes with the hypervisor, FaultManagement Architecture (FMA), and the system controller (SC), independent of the number of other logical domains configured. The number of LDC channels that is allocated by the control domain depends on the platform and on the version of the software that is used.
2. The control domain allocates one LDC to every logical domain, including itself, for control traffic.
3. Each virtual I/O service on the control domain consumes one LDC for every connected client of that service.

Bypass limits (T2-family 512 LDCs, all newer incl. Fujitsu M-series 768 LDCs) with:

- OSL Storage Cluster Devices
- OSL RSIO-Devices
- Run RSIO-Client inside LDOM (saves LDCs)

Author: Bert Miemietz, Rev. 11/2013

What Matters
Choosing the Right Approach

- Device setup requires a decision between
  - maximum performance
  and
  - flexibility and live migration feature
- In theory you can run:
  M10-1 32 domains
  M10-4 128 domains
  M10-4S 256 domains
- Good practice is:
  - 2 Cores (in most cases) for control & service domain
  - no less than 1 to 2 core(s) for guest domain
  which results in:
  M10-1 ~ 6 domains
  M10-4 ~ 32 domains (irregardless of PCIe)
- There is little experience with virtual devices but unless performance is a great issue I'd prefer to use virtual devices
- Test all desired operation including live migration !

Author: Bert Miemietz, Rev. 11/2013

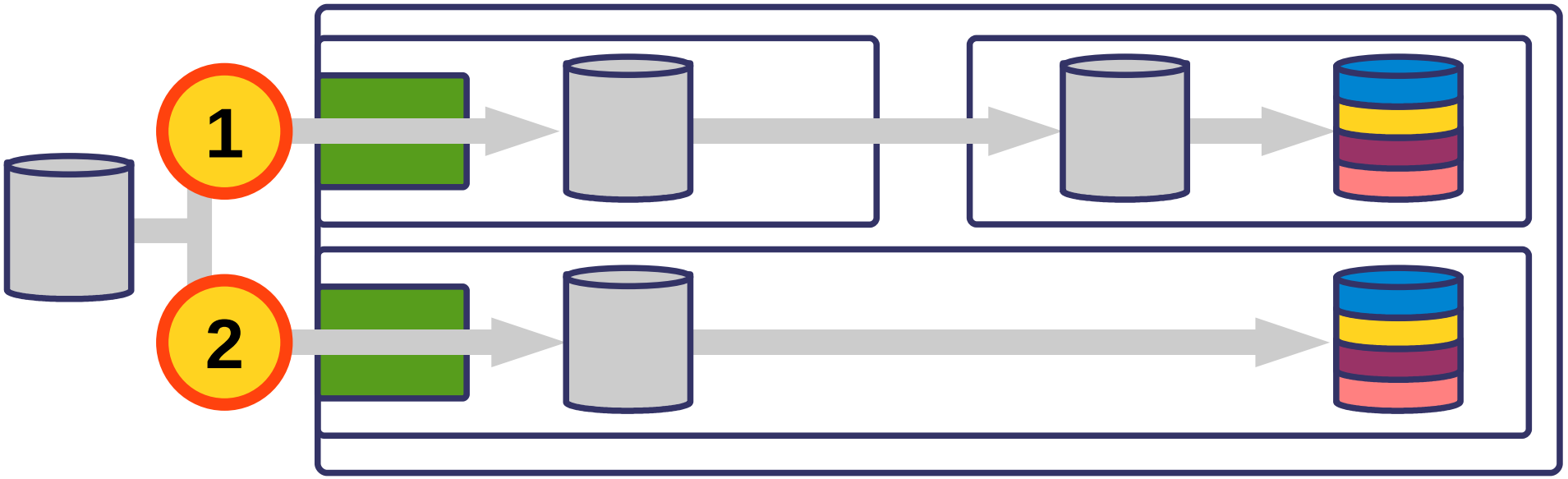
LDOMs
&
OSL Storage Cluster

Author: Bert Miemietz, Rev. 11/2013

For Those in a Hurry

LDOM as Physical Node

- LUNs provided via: 1. service domain -> guest domain
  2. native I/O-domain
- OSL Storage Cluster is being installed in the desired domain, Virtualization is using these LUNs
- Only method 1 allows live migration but has some limits
  - error-prone device configuration in LDM framework
  - limited by LDCs
- Method 2 has a tight limit of possible domains (mainly because of i/o slots)

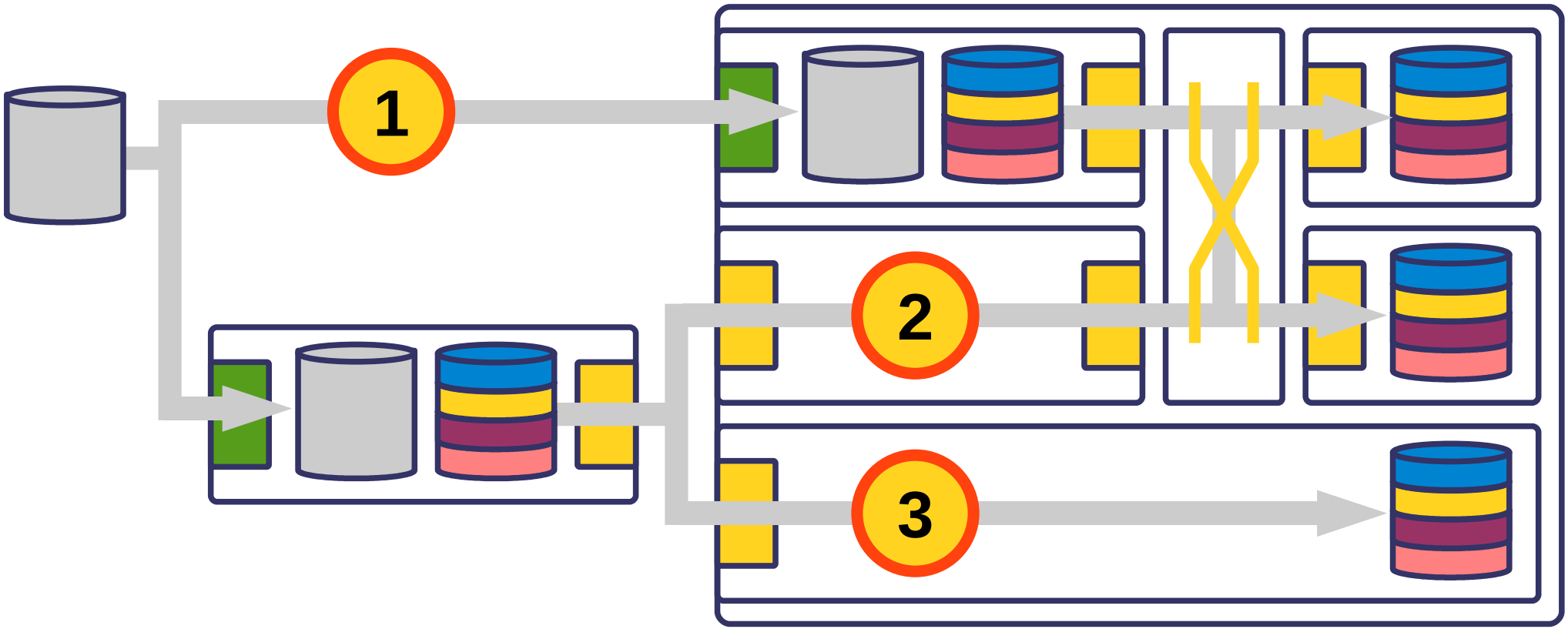


Author: Bert Miemietz, Rev. 11/2013

For Clever People

LDOM as RSIO-Client / Virtual Node

- V-Storage via RSIO:
  1. FC-to-Ethernet in Service-Domain
  2. via external RSIO-Server and Service-Domain
  3. via external RSIO-Server and Ethernet-I/O-Domain
- Live-Migration always possible
- Number of LDOMs without additional limits
- flexible and slim device handling even for thousands of devices



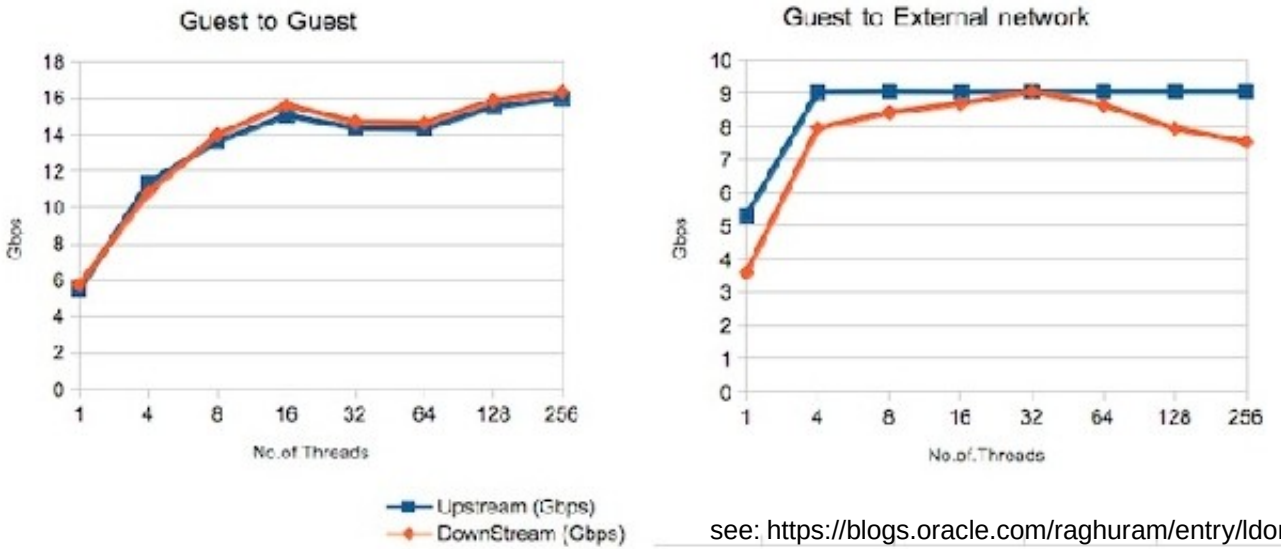
And What About Performance?

Ease of Use and Increased Flexibility Make the Difference

- Over a single connection / v-switch we get > 300 MiB/s with TCP/IP
- Today possible: multiple connections since RSIO scales linearly (in reasonable limits)
  ⇒ today via 4 connections > 1 GByte/s
- Further improvements seem possible:
  ⇒ RSIO over raw-Sockets (no TCP/IP)
  ⇒ significant improvements in Virtual Networking for LDOMS Sol11.1/SRU9

Example: Oracle improved network performance for T5-2 with 2 cores control and guest each

Raghuram Kothakota's Weblog

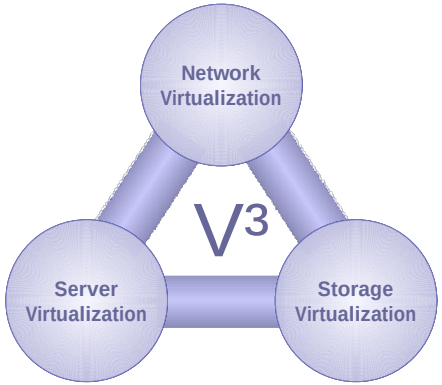


see: https://blogs.oracle.com/raghuram/entry/ldoms_virtual_network_performance_greatly1

And Even More – SPARC as UVC

Project Integration as Unified Virtualisation Client

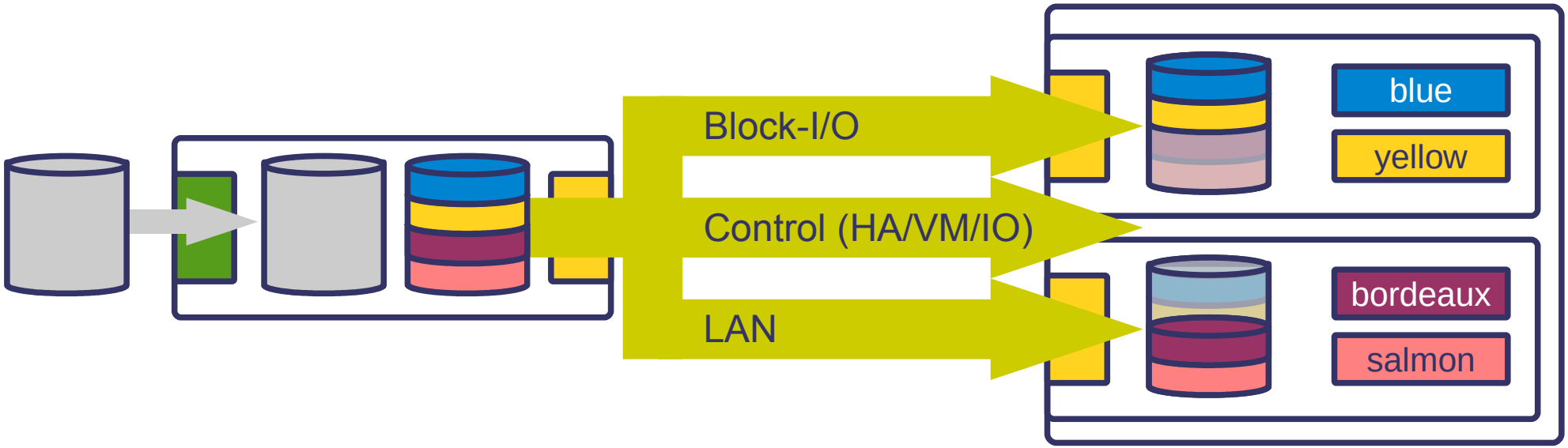
- LDOM considered just another Hypervisor in the Unified Virtualisation Environment



Unified Virtualisation Server
- Control-Engine
- HA
- Disk-Access-Mgmt.

Converged Network
- Hypervisor
- VM Execution

Unified Virtualisation Client
- Hypervisor
- VM Execution



Author: Bert Miemietz, Rev. 11/2013

Summary

Keep in Mind

Things Have Changed Since SunFire / M4000 and Solaris 10

- Due to increased hardware performance you will need virtualization
- Physical Partitions are very unlikely to be sufficient as sole technology
- Can combine LDOMs and Solaris Zones
- LDOMs provide cool new features
- Most installations today are Solaris 10 with massive zone usage
- You will need a migration concept for machines and update to Solaris 11
- Branded Zones can facilitate fast migration
- Very high performance requirements need special inspection (devices, latency ...)
- Request experts help in system planning, deployment and migration
- Look for integrated concepts that comprise:
  - Server virtualization
  - Storage virtualization
  - Network virtualization
  - Global Management
  - High availability
- Have a closer look at OSL's integrated solutions ☺

Author: Bert Miemietz, Rev. 11/2013

virtualization and clustering – made simple