Zone Clusters

Oracle RAC in Solaris Containers

Paul Mitchell

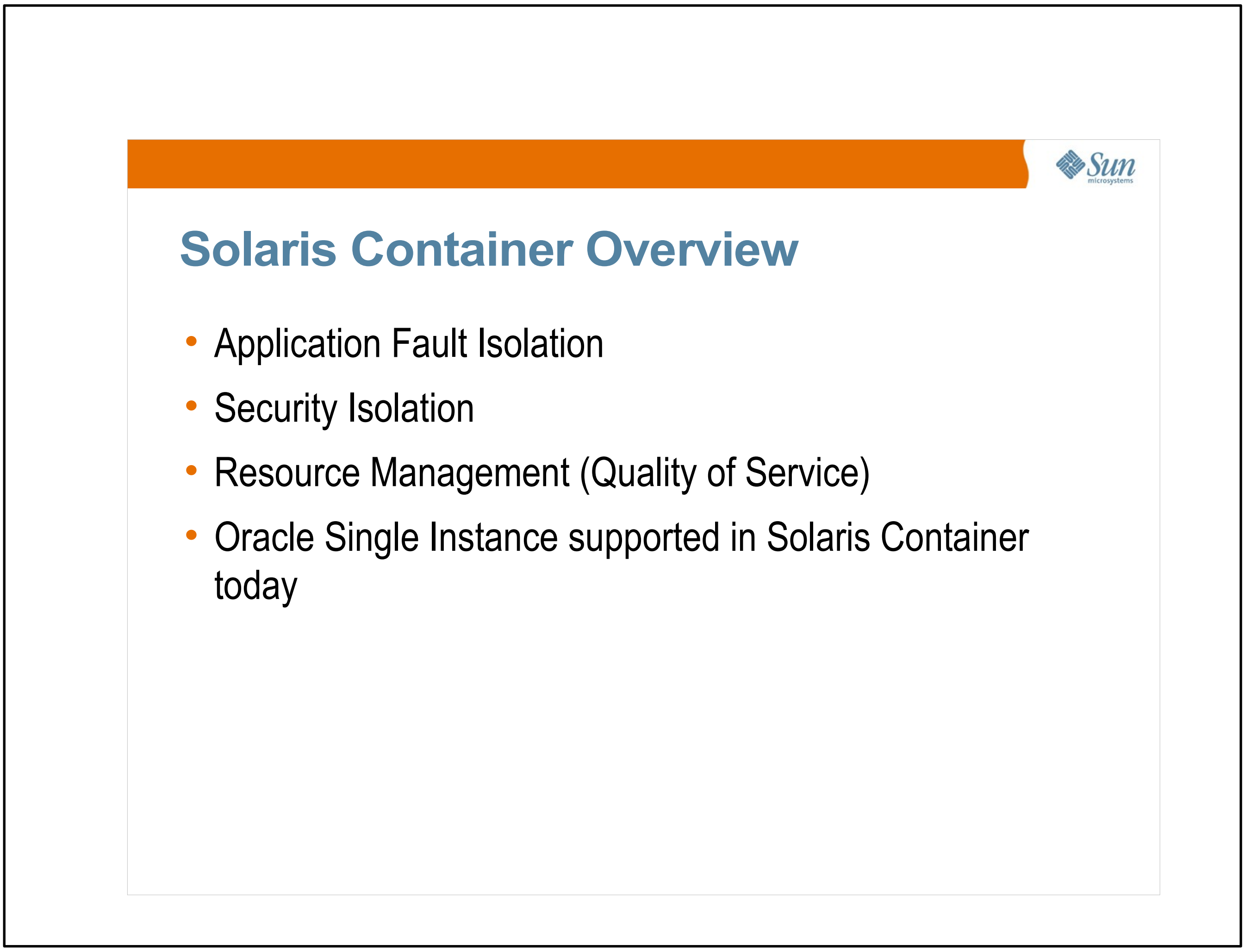
Sun Microsystems, Inc.

This evening I'm going to talk about the latest feature to be released in the Sun Cluster product.

This feature is Zone cluster and it's basically a new level of infrastructure to provide a Virtual cluster based upon the Solaris container feature also know as a zone

In this presentation I will cover a number of concepts.

1. Why somebody would want to use zones.
2. Explain what zone cluster is
3. Explain the major use cases
4. The major features that have been added



Solaris Container Overview

- Application Fault Isolation
- Security Isolation
- Resource Management (Quality of Service)
- Oracle Single Instance supported in Solaris Container today

First I'll talk a little about Solaris containers and what they are:

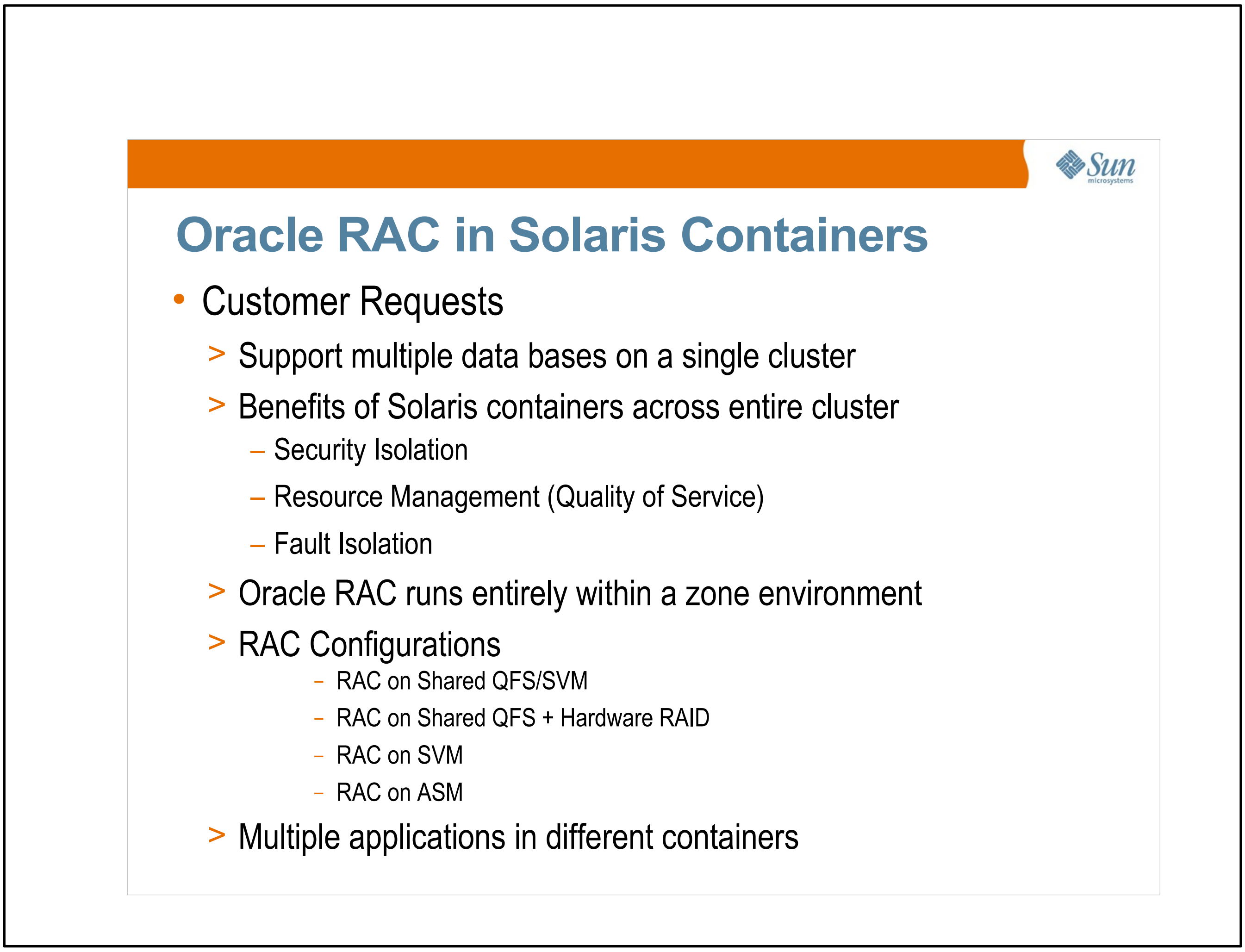
The first one is Application fault isolation – the OS provides some level of application fault isolation, for example the panic of one application instance does not cause failures in all other applications. However, there are actions that an application can take that will cause failures of other applications. Oracle RAC is one of those applications which can order a node to reboot, which obviously affects all other applications on that node. A Solaris zone has been designed to reduce the possibility that a misbehaving app can negatively affect other applications. So if we have our RAC application within a zone and it issues a reboot, it's not going to affect applications out side of that zone.

The next thing that happens is that virtual clusters are used primarily for consolidation, and one of the things you have with consolidation is that you have to start supporting multiple organizations. So security is one of the key features within zones. Specifically they follow the principle that the only resources and things that you can see or affect are those things that are explicitly configured to be present in that zone.

One of the main problems when you start doing consolidation is that each organization does not want their schedules or deliverables to impact any of the others.

So what has been done with zones is that they have resource controls that you can establish, you can grant specific devices, you can grant specific shares of resources, for example you could allocate a certain number of CPU's. This applies a wide range of resources and I don't have time to cover all of them but the main ones are CPU, memory, devices and file systems

So why did Sun Cluster team start to think of using zones? it's for the same reason that people are consolidating single machines. Applications from multiple machines onto a single machine, also applies to Clusters. The machines are getting faster and the storage is getting bigger and bigger as well as the bandwidth of the network is increasing. So what is happening now a days is that systems have plenty of surplus capacity , so to get a more cost effective delivery mechanism, customers are now placing multiple applications including Cluster applications on the same set of physical machines. So we have to have a way to do this safely.



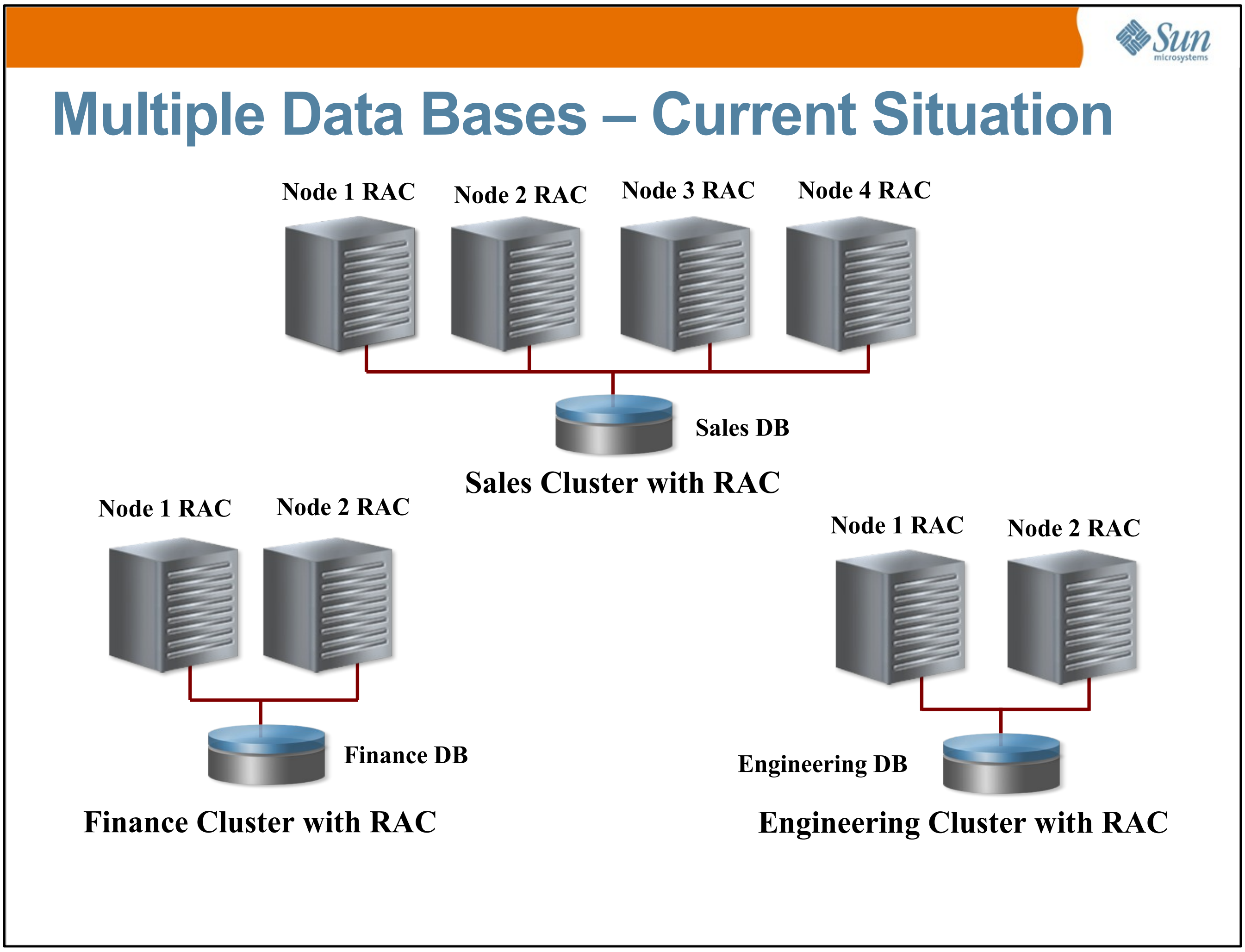
Oracle RAC in Solaris Containers

- Customer Requests
  - Support multiple data bases on a single cluster
  - Benefits of Solaris containers across entire cluster
    - Security Isolation
    - Resource Management (Quality of Service)
    - Fault Isolation
  - Oracle RAC runs entirely within a zone environment
  - RAC Configurations
    - RAC on Shared QFS/SVM
    - RAC on Shared QFS + Hardware RAID
    - RAC on SVM
    - RAC on ASM
  - Multiple applications in different containers

one of clusters biggest set of customers is the ones that run Oracle RAC databases, and just being able to support a single zone isn't really good enough for that application. Oracle RAC runs multiple instances on multiple machines all at the same time. It also needs a notion of membership, and this notion of membership is quite private and its only interested in those machines that are hosting its instances. So we have to create a virtual cluster that has a notion of membership and isolate that database from other databases. We want to be able to use zones and all its features like application isolation, security and the resource management. If you want to keep multiple RAC databases on the same system and run them with no interference, you've got to make sure that one RAC DB doesn't take out any of the other RAC databases. So by placing Oracle RAC inside zones you can achieve that.

Customers don't use just one type of storage topology. Zone Cclusters support a wide range of storage topologies. It supports the ability to run Oracle RAC on shared QFS distributed FS, running on top of the Solaris Volume manager. It also supports the ability to run Oracle RAC on shared QFS directly on top of HW Raid, you can also run it on top of a volume manager without a file system or on top of the ASM product from Oracle. And on the same system but in different zone clusters you can run each and every one of those concurrently for different databases - it's entirely up to you.

We also want to be able to run other applications, almost all customers use other applications that either feed data into the database or extract it out.



Multiple Data Bases – Current Situation

Node 1 RAC

Node 2 RAC

Node 3 RAC

Node 4 RAC

Sales DB

Sales Cluster with RAC

Node 1 RAC

Node 2 RAC

Finance DB

Finance Cluster with RAC

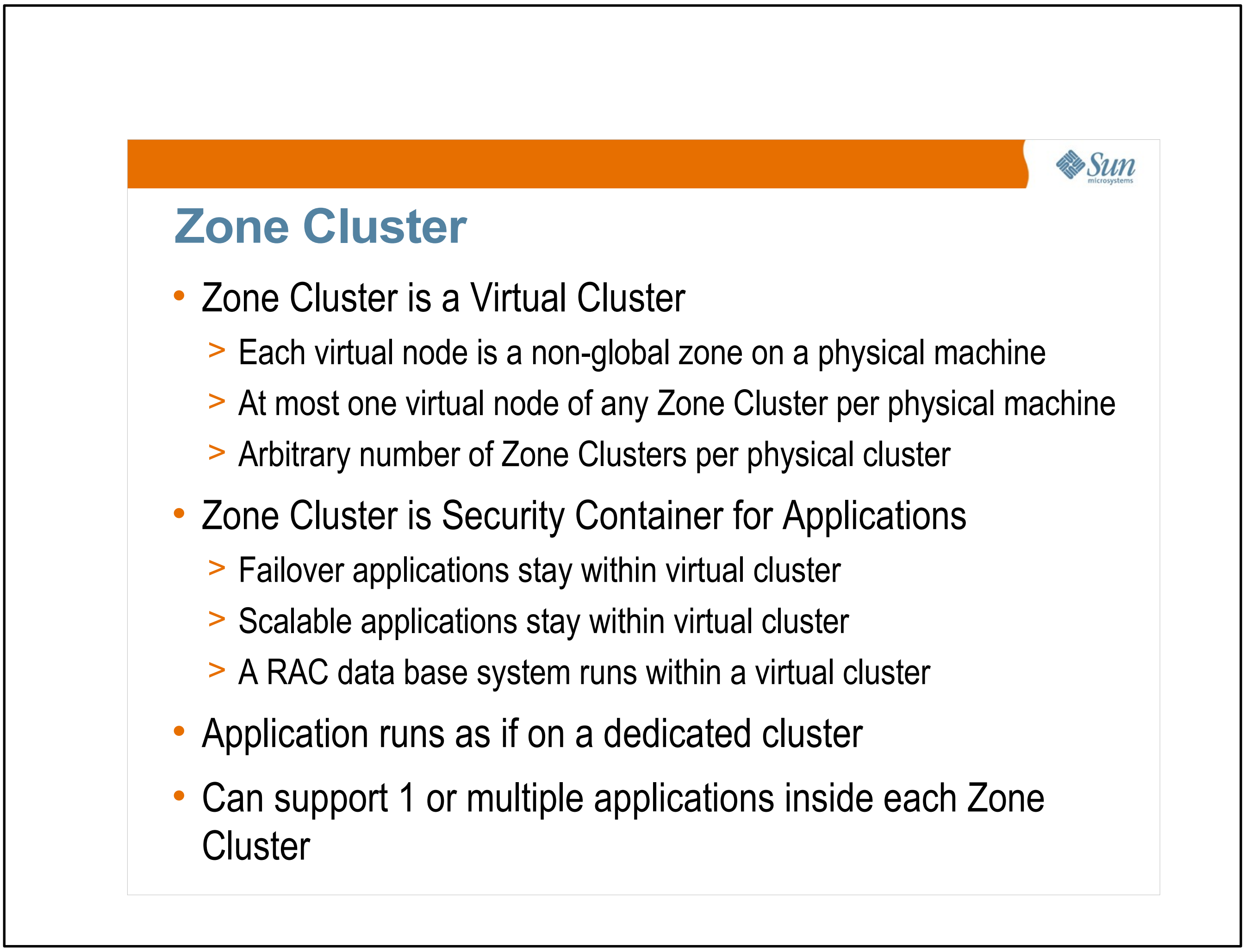
Node 1 RAC

Node 2 RAC

Engineering DB

Engineering Cluster with RAC

This slide simply shows you how we support multiple databases today, in particular you will see that have one cluster per database.



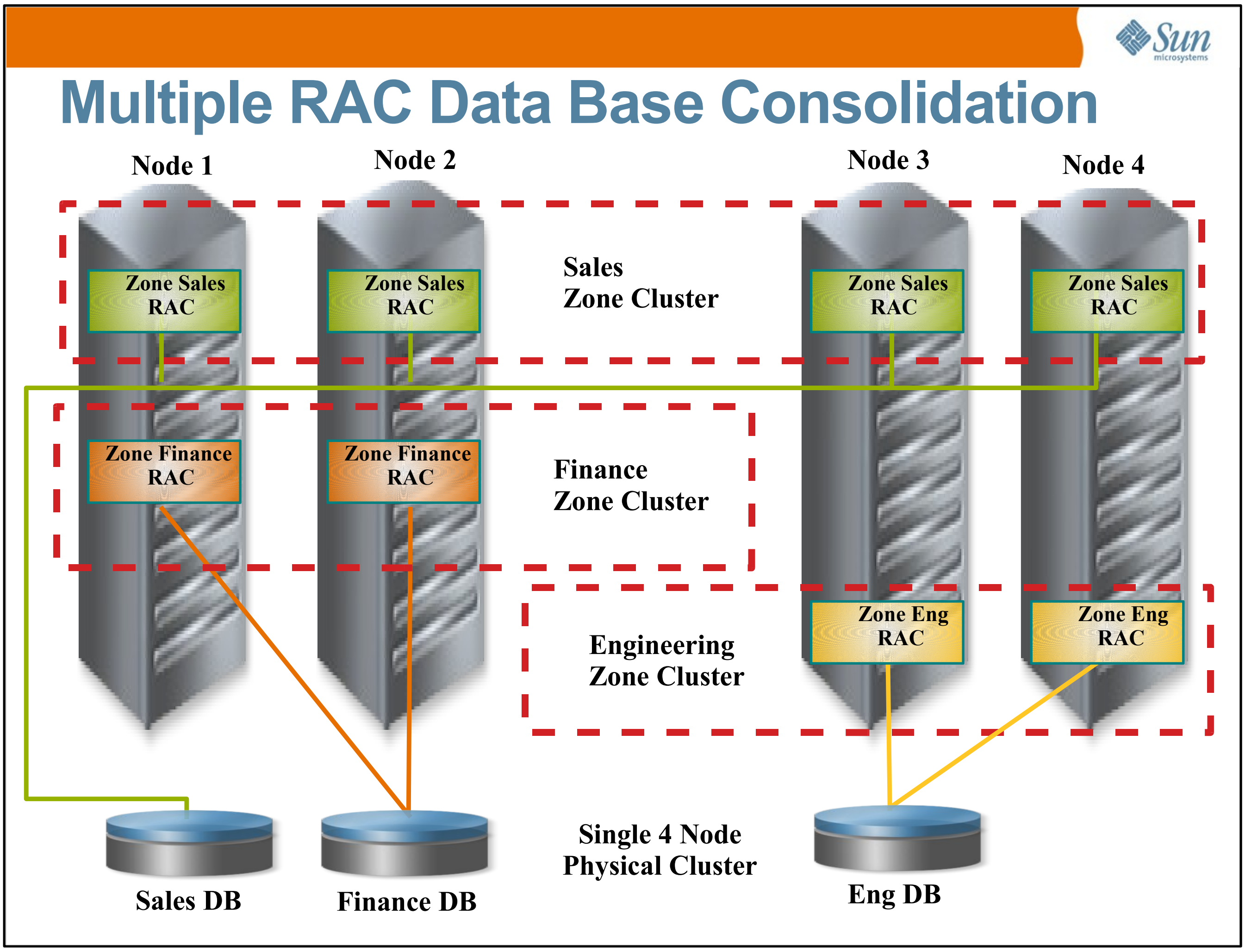
Zone Cluster

- Zone Cluster is a Virtual Cluster
  - > Each virtual node is a non-global zone on a physical machine
  - > At most one virtual node of any Zone Cluster per physical machine
  - > Arbitrary number of Zone Clusters per physical cluster
- Zone Cluster is Security Container for Applications
  - > Failover applications stay within virtual cluster
  - > Scalable applications stay within virtual cluster
  - > A RAC data base system runs within a virtual cluster
- Application runs as if on a dedicated cluster
- Can support 1 or multiple applications inside each Zone Cluster

So what is a Zone cluster – its's a virtual cluster, where each virtual node is a zone of a specific type and its brand is a clustered zone. This is basically a standard zone that has some hooks for clustering. We run each virtual node on a separate machine, and we do that for availability, we don't want the failure of one machine to cause multiple nodes in the virtual cluster to die. There can be an arbitrary number of virtual clusters on a given system. The limit is actually thousands but the real limit is the number of cpu's and memory you have to support the applications you plan to run.

The next aspect that's important about zone clusters is a security container for the application. Sun Cluster supports two kinds of application. One is the failover application, where an application runs on one machine at a time and if that machines fails or based on loading we want to move the application, the application can be moved to another machine.

The other class of application is a group that we call scalable applications. A scalable application actually runs software components on multiple machines, at the same time and there can be an arbitrary number of machines. With Zone Cluster that scalable application will have all of its instances and components all residing within the same zone cluster. So basically an application within a Zone Cluster cannot get outside of that Zone Cluster.



Multiple RAC Data Base Consolidation

Node 1

Node 2

Node 3

Node 4

Zone Sales RAC

Zone Sales RAC

Sales Zone Cluster

Zone Sales RAC

Zone Sales RAC

Zone Finance RAC

Zone Finance RAC

Finance Zone Cluster

Engineering Zone Cluster

Zone Eng RAC

Zone Eng RAC

Sales DB

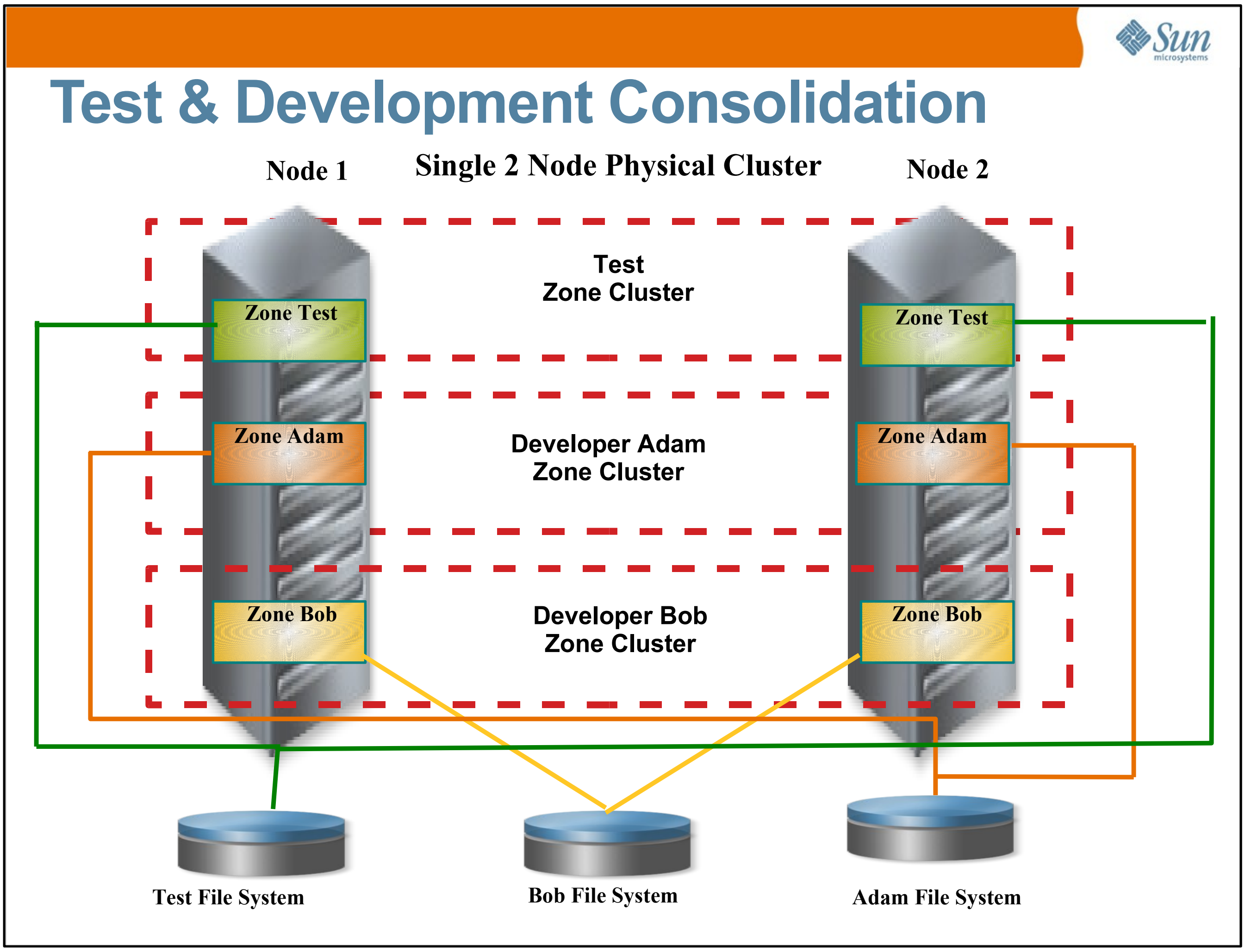
Finance DB

Single 4 Node Physical Cluster

Eng DB

In this slide you can see multiple RAC D.B.'s in a Zone Cluster. It shows four machines where one particular Zone cluster spans all four nodes and the other two D.B.'s span just two nodes each. Here you can see that you can have any kind of combination of nodes in zone cluster. The only requirement that we have, is that each Zone Cluster must span some subset of these zones.

I'd like to point out with that you can run multiple versions of RAC, not just the latest version. Oracle RAC 9i, RAC 10G and RAC 11G have been run in separate Zone Cluster's on the same physical cluster at the same time, and they are completely isolated and that's very useful for consolidation. It also gives you the advantage that if you want to do upgrades where your currently running 9i, you could also put a new DB running 11g in another zone cluster at the same time your running that other DB on 9i. So it gives you an upgrade path you don't have with other kinds of systems.



Test & Development Consolidation

Node 1

Single 2 Node Physical Cluster

Node 2

Test
Zone Cluster

Developer Adam
Zone Cluster

Developer Bob
Zone Cluster

Zone Test

Zone Adam

Zone Bob

Zone Test

Zone Adam

Zone Bob

Test File System

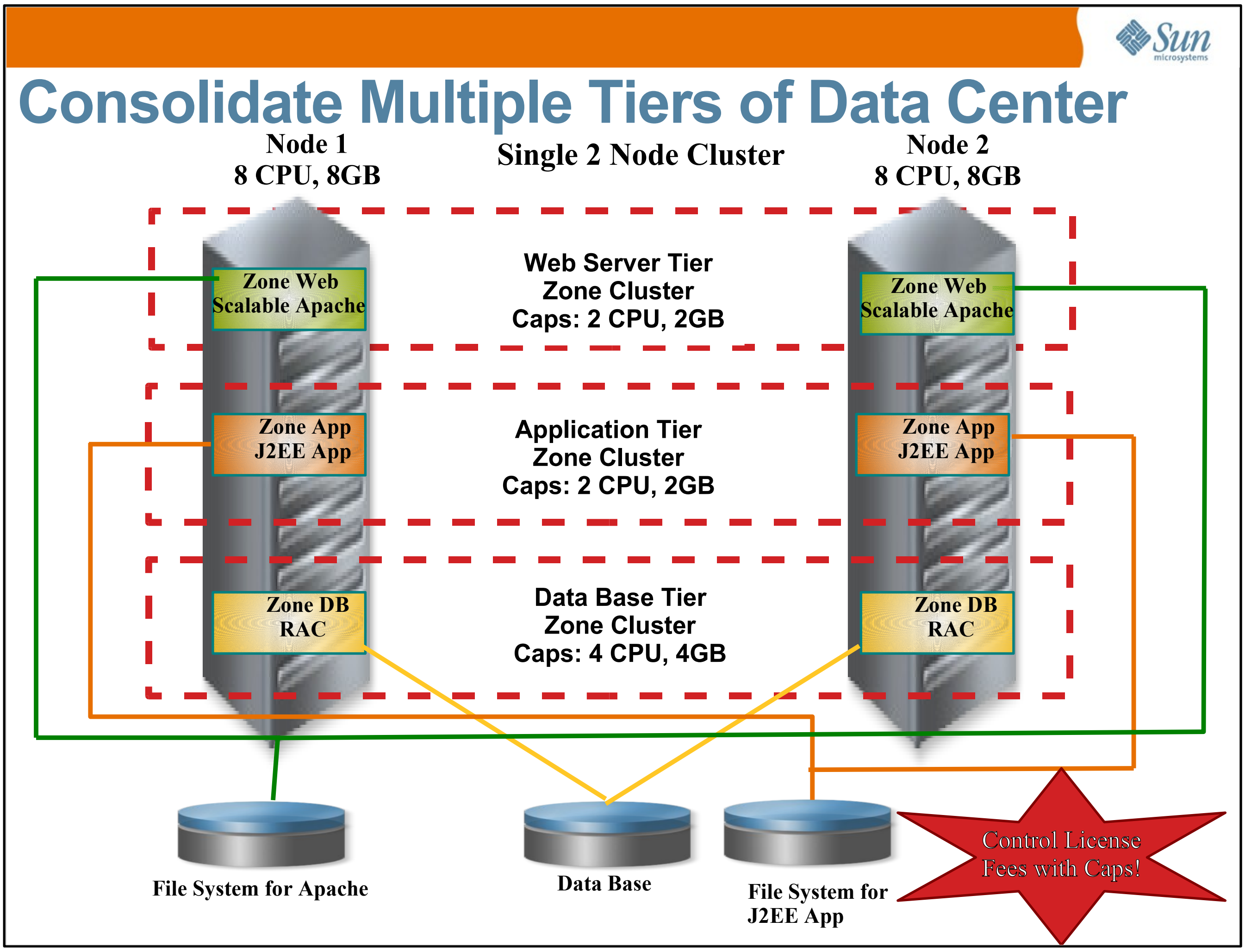
Bob File System

Adam File System

lets look at another kind of use case that we have with Zone Clusters.

This is to do with test and development consolidation. Many customers will run 3 different clusters, one for the production environment, another cluster for the test organization and the third cluster for the developers.

There will also be many customers that never wish to put anything on their production environment. With Zone Cluster the application being tested is isolated from the others so a reboot, failure testing and whatever don't impact with each other. You can also control what resources the developers get. There is no need for the developers to sign out for a slot at 2 in the morning for four hours, they can all work during the same time during the day. You can also dynamically create and destroy Zone Clusters without affecting other Zone Clusters. So if you have a new developer start, you create a new Zone Cluster by allocating space off of your disk, allocate some network bandwidth , allocate some CPUs and they are all set to go.



Consolidate Multiple Tiers of Data Center

Node 1
8 CPU, 8GB

Single 2 Node Cluster

Node 2
8 CPU, 8GB

Web Server Tier
Zone Cluster
Caps: 2 CPU, 2GB

Application Tier
Zone Cluster
Caps: 2 CPU, 2GB

Data Base Tier
Zone Cluster
Caps: 4 CPU, 4GB

Zone Web
Scalable Apache

Zone App
J2EE App

Zone DB
RAC

File System for Apache

Data Base

File System for
J2EE App

Control License
Fees with Caps!

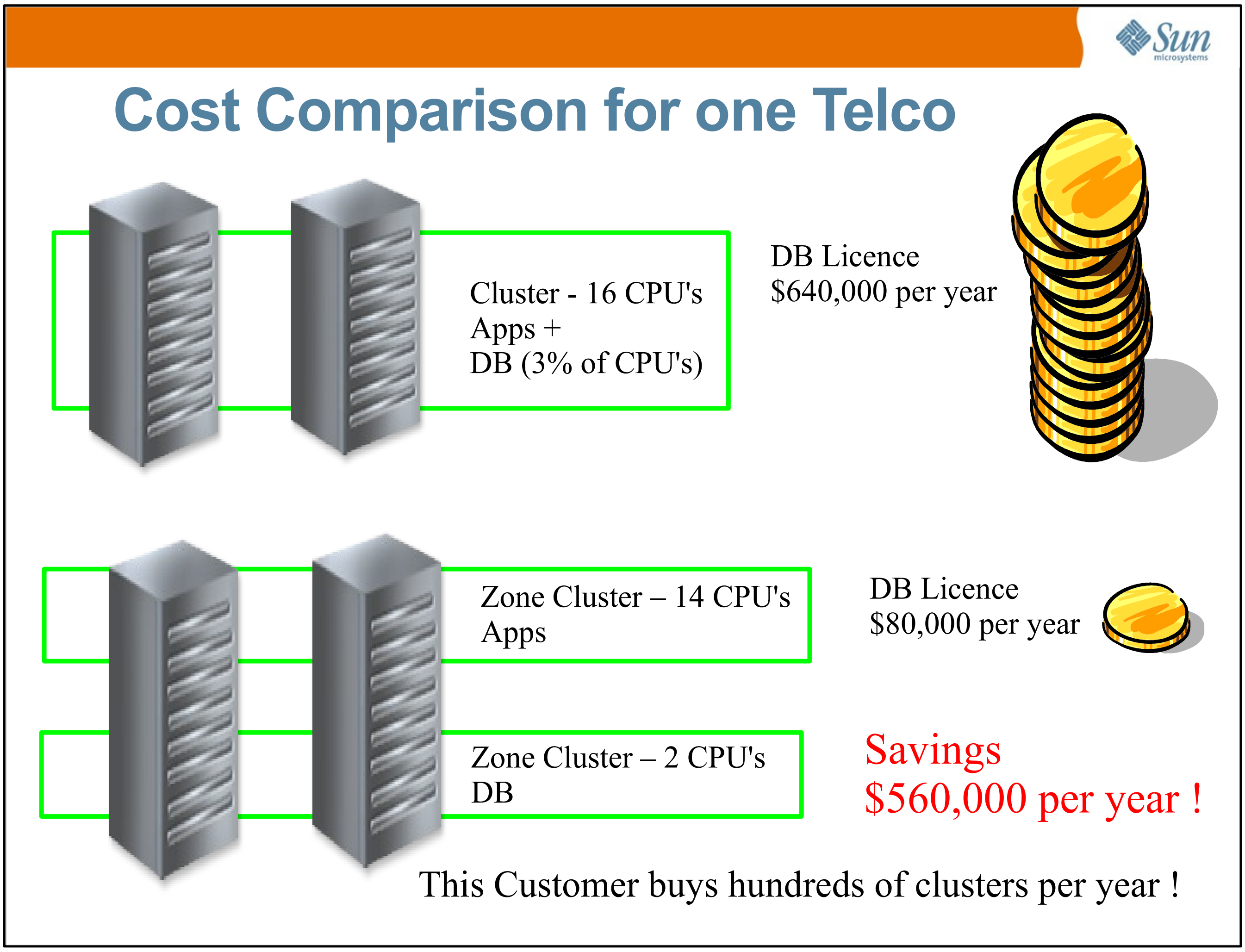
Here's another use case for Zone cluster.

A lot of people are going to use this in the data center and they typically follow a 3 tier model. In this slide I show the 3 tiers of the data center. One tier is for the web server front end, one for the application tier and one for the database tier. In this example I show that you can run multiple applications and you can specify some sort of resource control. I have placed two cpus in the web tier, two for the application tier and four for the database tier. Similarly you can control the amount of memory being used by each tier.

Another thing that I'd like to point out with this slide is that with our existing cluster product we can control the order in which the applications and resources are brought online and initialized. If I start putting them in different zone clusters, how does that work if you want to keep them isolated.

Well, what we do is allow the administrator in the global zone to control and establish dependency and affinity relationships between different zone clusters. It still supports security so that somebody that's inside one zone cluster can't go and grab or allocate a resource dependency with something in a different Zone Cluster.

Now you can see that we can do resource controls, let me show you in the next slide an example for a telco company where this can be a very financial consideration for their organization.



Cost Comparison for one Telco

Cluster - 16 CPU's
Apps +
DB (3% of CPU's)

DB Licence
$640,000 per year

Zone Cluster – 14 CPU's
Apps

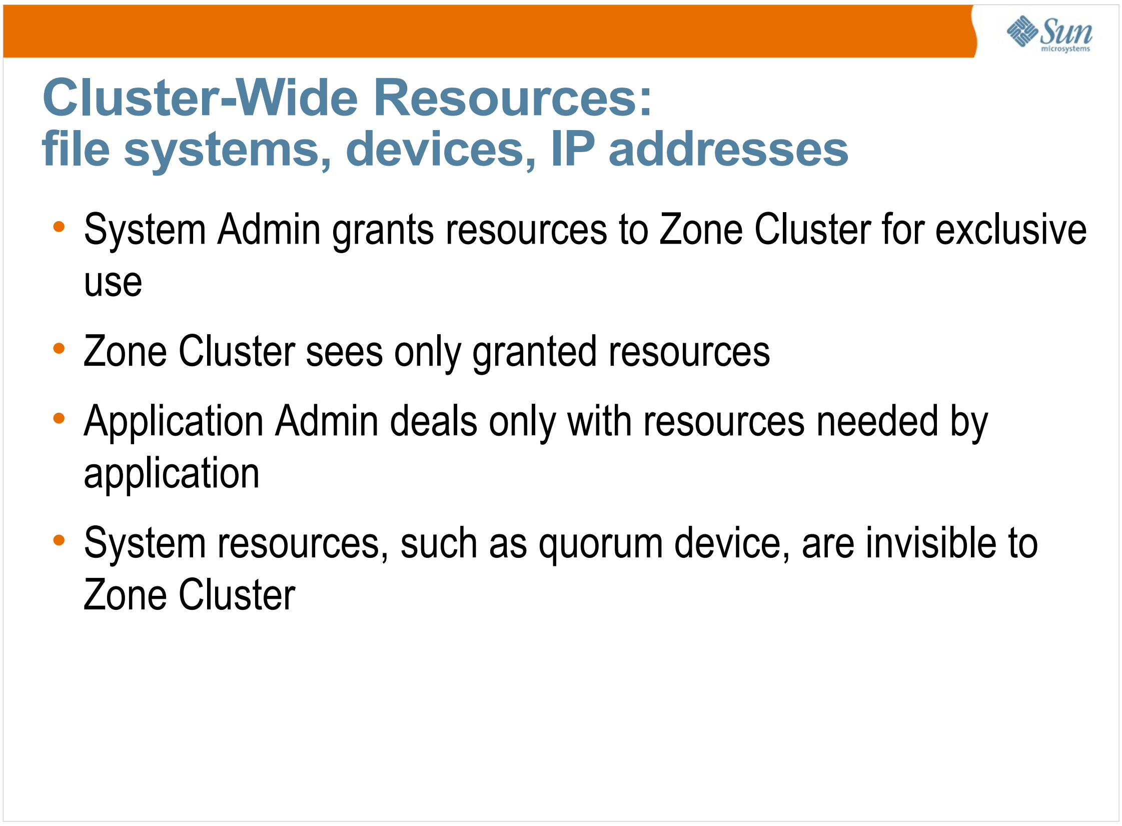
DB Licence
$80,000 per year

Zone Cluster – 2 CPU's
DB

Savings
$560,000 per year !

This Customer buys hundreds of clusters per year !

In this particular example we will have a telco company that's name I'll keep quiet for privacy purposes. They are going to come up with their next generation of product which is a two node system where each node will have 8 cpus, with a total of 16 cpus. One of the major database providers out there, charges a license fee of 40000 dollars per year per cpu, so the net license cost would be 640,000 dollars per year per cluster. In this particular scenario they run multiple applications and it turns out that the database is a small part of their configuration, and they only need 3% of the cpu resources. So what they can do is place the Database inside a zone cluster and allocate one cpu per machine. They can now lower the license cost fees from 640,000 dollars down to 80,000 dollars a year for a net cost of 540,000 dollars of savings per year, and they buy hundreds of clusters per year. So the financial savings are quite significant. This doesn't just apply to Database vendors, there are many vendors that charge a per license fee. So as you can see that's a fairly good example for this feature and hopefully most people would find significant cost reductions using this feature



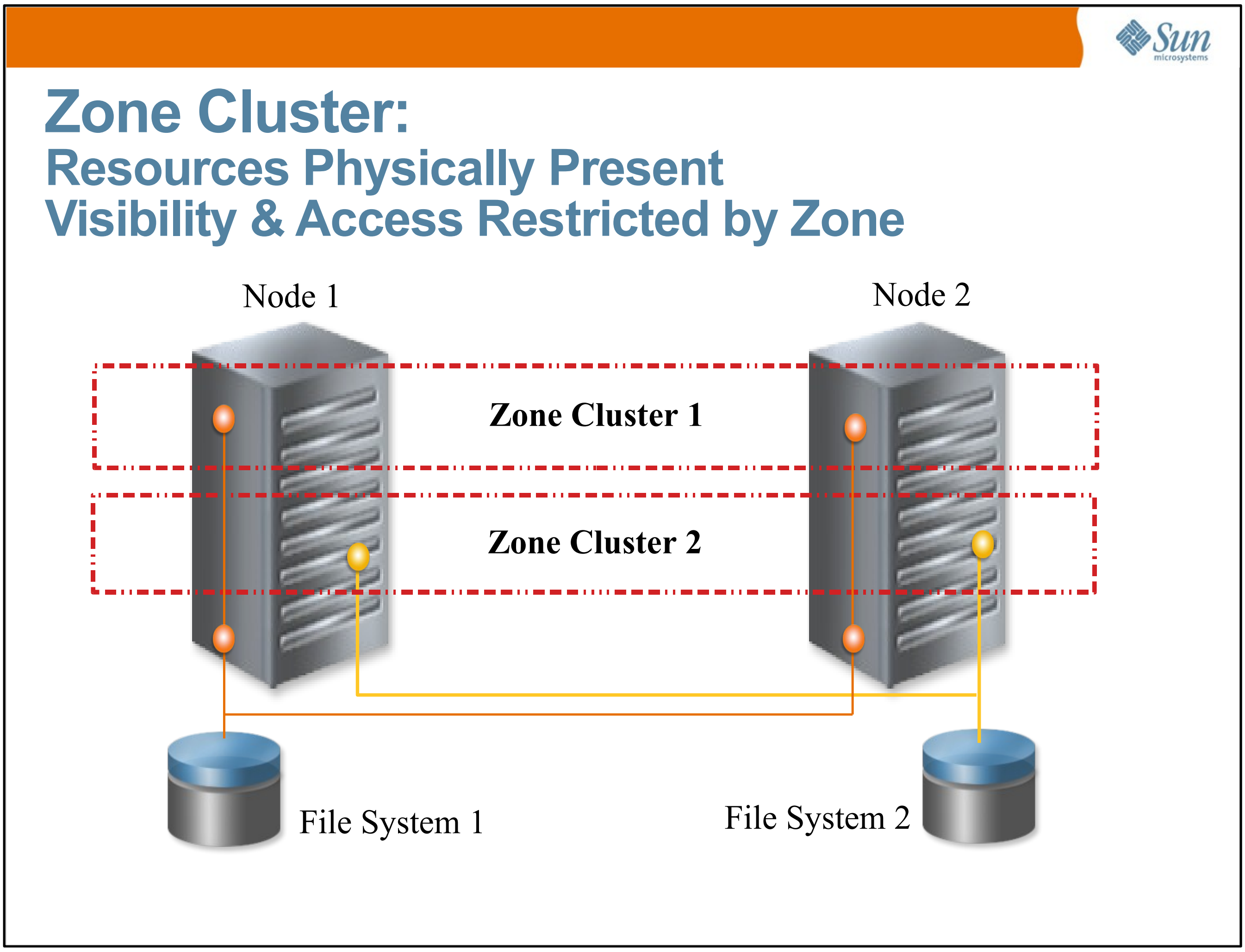
Cluster-Wide Resources: file systems, devices, IP addresses

- System Admin grants resources to Zone Cluster for exclusive use
- Zone Cluster sees only granted resources
- Application Admin deals only with resources needed by application
- System resources, such as quorum device, are invisible to Zone Cluster

So how are resources on a cluster made available to the zone cluster. What really happens is that the system administrator grants what resources are actually allocated to the Zone Cluster.

A Zone Cluster can only see what has been explicitly allocated to that particular zone cluster. It has simplified the world for people that use clusters. Clusters have a reputation of having a lot of things to configure and deal with, well, in zone cluster that has changed to limit what is actually required. So your not going to find quorum devices, have to configure heartbeats, at least not in the initial product,

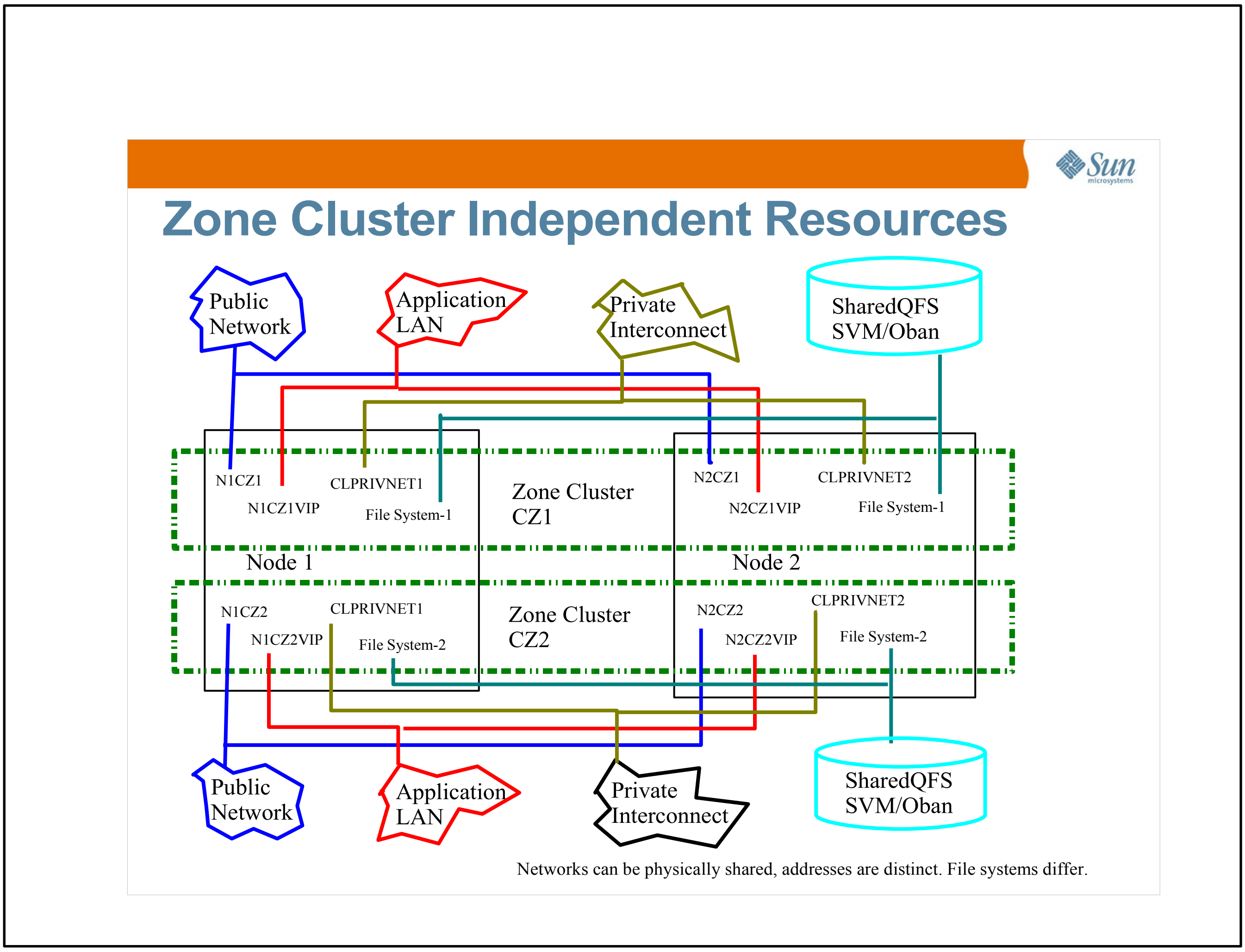
The only things that are going to be there are the things directly used by the application.



Zone Cluster:

Resources Physically Present
Visibility & Access Restricted by Zone

Here, I show two file system resources which are present and accessible on the physical machine, but its the global zone admin who is the one that determines which one of those resources is available for that particular Zone Cluster.



Zone Cluster Independent Resources

Public Network

Application LAN

Private Interconnect

SharedQFS SVM/Oban

N1CZ1

N1CZ1VIP

CLPRIVNET1

File System-1

Zone Cluster CZ1

N2CZ1

N2CZ1VIP

CLPRIVNET2

File System-1

Node 1

Node 2

N1CZ2

N1CZ2VIP

CLPRIVNET1

File System-2

Zone Cluster CZ2

N2CZ2

N2CZ2VIP

CLPRIVNET2

File System-2

Public Network

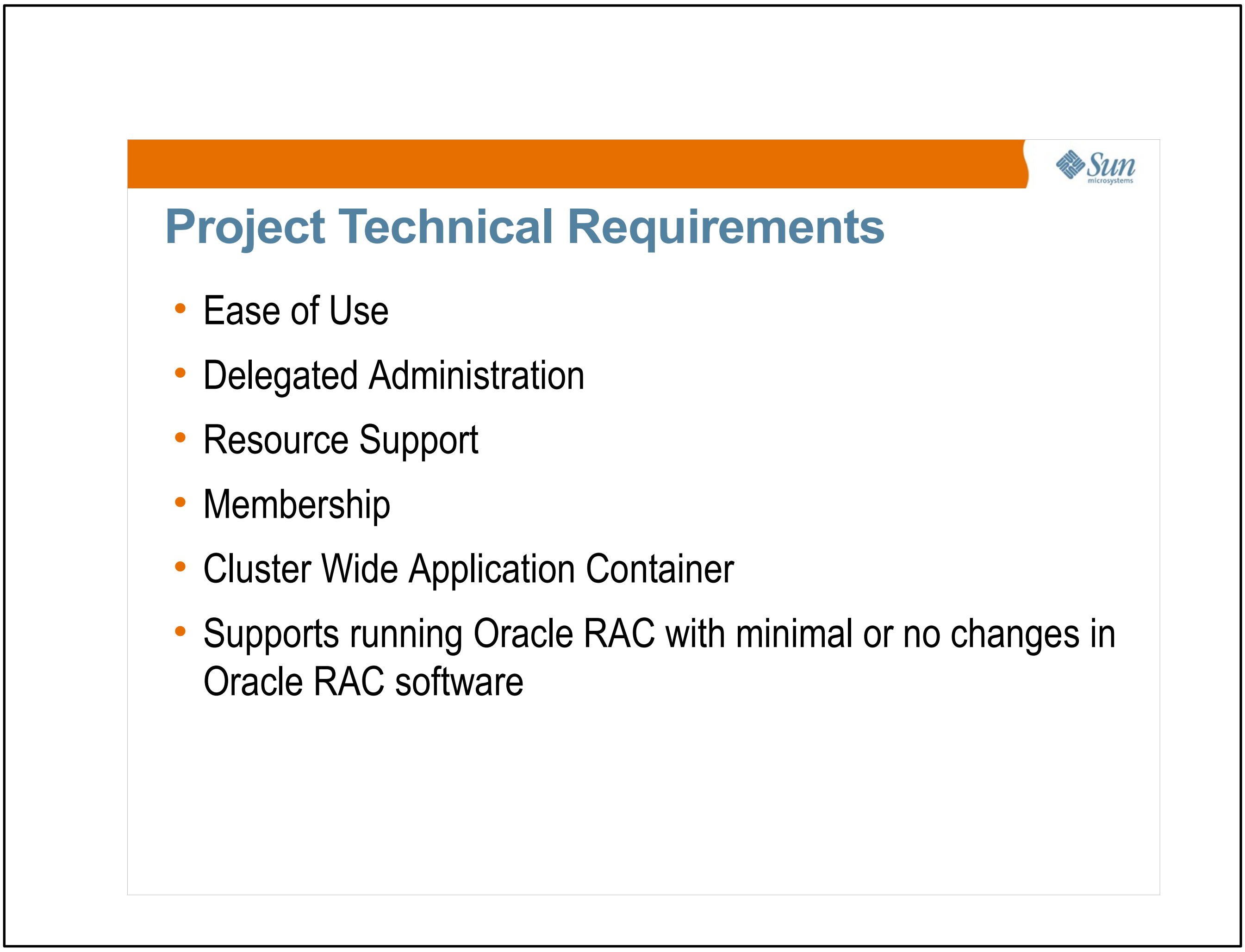
Application LAN

Private Interconnect

SharedQFS SVM/Oban

Networks can be physically shared, addresses are distinct. File systems differ.

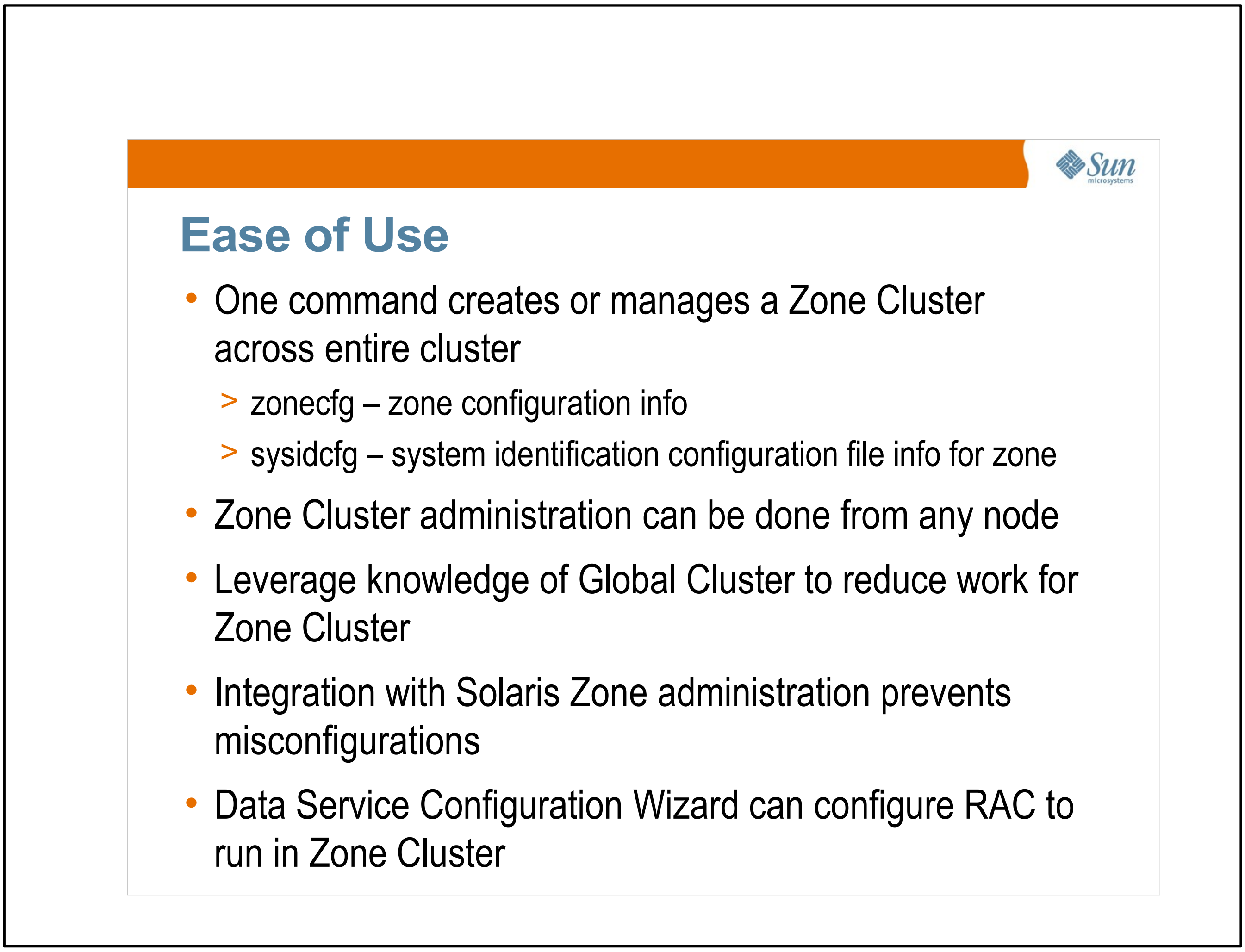
This slide shows the resource configurations for Oracle RAC. Oracle RAC has four major kinds or resources that it needs. It needs a resource for the public network, one for the application LAN, basically that's for the application on your cluster that talks across the LAN to your RAC database. One for the private interconnect so that the various RAC components can talk to each other. The last main resource is the data storage that can be a file system or device. This particular slide shows how you can keep them all separate for different DB installations on the same physical system



Project Technical Requirements

- Ease of Use
- Delegated Administration
- Resource Support
- Membership
- Cluster Wide Application Container
- Supports running Oracle RAC with minimal or no changes in Oracle RAC software

The next major area I want to go through is a list of things that have been done specifically in terms of features for zone cluster in order to make the zone cluster a practical solution. One of the first and most important areas is ease of use. In this day and age, people want the computer to do the work so they've looked at how they could support a virtual cluster and tried to reduce the amount of work that the administrator has to do.



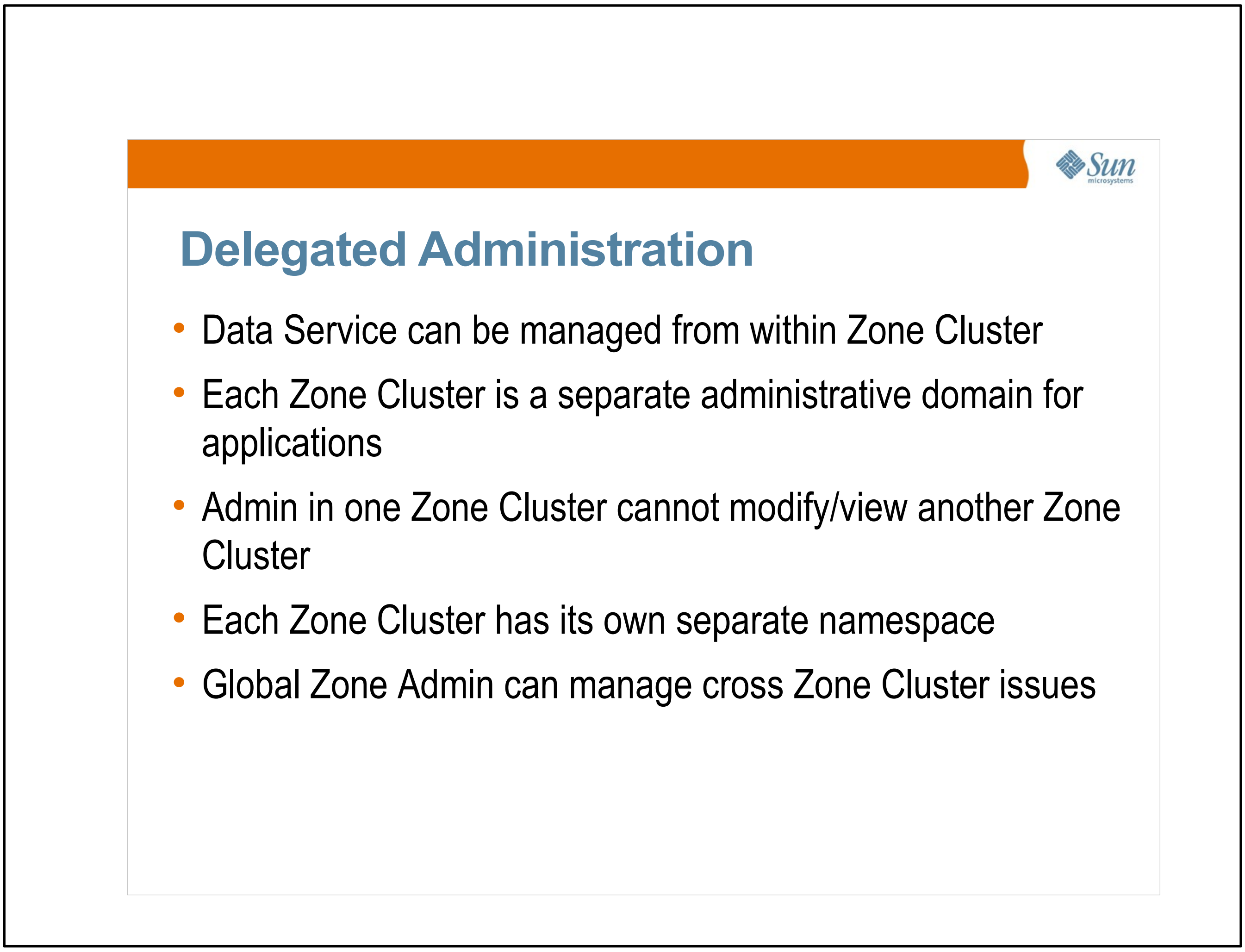
Ease of Use

- One command creates or manages a Zone Cluster across entire cluster
  - zonecfg – zone configuration info
  - sysidcfg – system identification configuration file info for zone
- Zone Cluster administration can be done from any node
- Leverage knowledge of Global Cluster to reduce work for Zone Cluster
- Integration with Solaris Zone administration prevents misconfigurations
- Data Service Configuration Wizard can configure RAC to run in Zone Cluster

A single command line interface has been created, so running one command from any machine in the cluster will create the entire zone cluster. Also where Solaris has two commands, with Zone cluster they have been combined to reduce that to one command. The next thing that has been done is the zone cluster is created after you've created the global cluster or the cluster of the global zones. We can take a look at that configuration and in many cases use that information, for example the timezone on a zone cluster is likely to be the same as the global cluster. So you don't have to input that data.

One of the biggest areas where you can have problems is where you have commands for Solaris, commands for clustering and sometimes the customer does one without doing the other and that leads to all sorts of bad things. Care has been taken to get the contracts and interfaces with Solaris so that you can't change the configuration for a zone cluster using the Solaris commands without the sun cluster software knowing about it, and making the appropriate checks and warnings so misconfigurations can be prevented.

The next thing in the ease of use, is the initial data service that we are going to support, which is Oracle RAC. The data service configuration wizard can actually do a lot of the leg work for configuring RAC within zone clusters. So the data service configuration wizard that was running on the global zone cluster has been modified so that it is aware of zone cluster and does the appropriate thing.



Delegated Administration

- Data Service can be managed from within Zone Cluster
- Each Zone Cluster is a separate administrative domain for applications
- Admin in one Zone Cluster cannot modify/view another Zone Cluster
- Each Zone Cluster has its own separate namespace
- Global Zone Admin can manage cross Zone Cluster issues

The world that we view for administration with zone clusters is really split into two parts.

The first part is what I call system administration or platform administration - the admin in the global zone configures the global cluster and creates the zone cluster. The admin creates the filesystems and does all that kind of work and then adds the final resources in to the zone cluster.

Then in the zone cluster you have an application admin, that with Oracle RAC would be your database admin. That admin inside his zone cluster can operate completely independently from any database administrators in other zone clusters. The name spaces for managing applications are completely independent, so they don't have to go and coordinate with each other. The database admins in one zone cluster cannot see or modify what goes on in another zone clusters.

Resource Support - Data

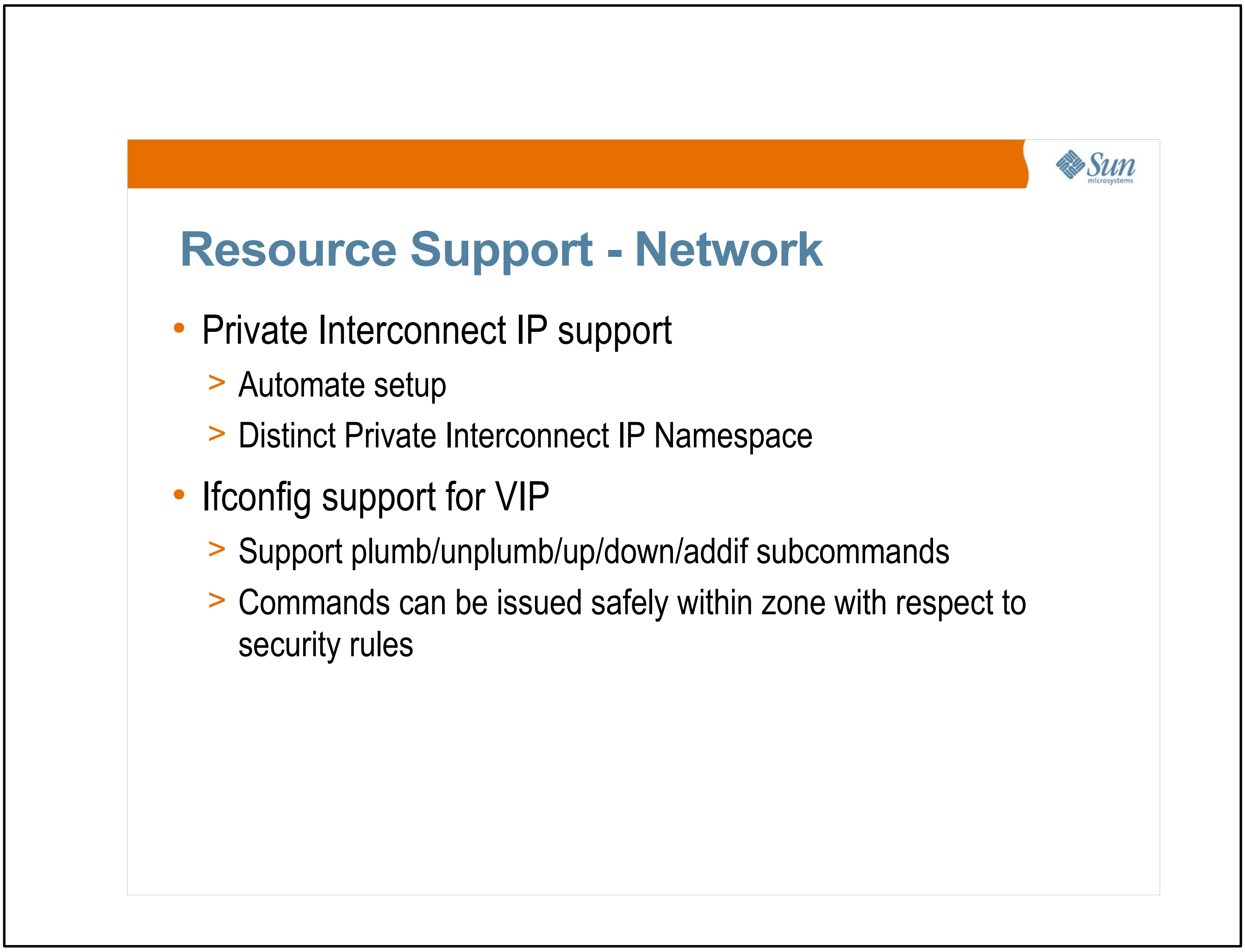
- Add Global File System Support
  - > Shared QFS Meta Data Server support
  - > Volume support hidden in global zone
- Add support for Storage Devices
  - > Direct Zone Support
    - – Oracle RAC on raw devices
    - – Mount capability
  - > Indirect Zone Support
    - – Volume administration in global zone
    - – Volume reconfiguration done in global zone

A number of features have been added to support zone clusters and the first one I'd like to talk about is the global filesystem. Zone cluster supports shared QFS for database purposes. Shared QFS allows each node to directly read and write to the actual storage. It also has something called a metadata server and there's only one of these and it can go from node to node. A new feature has been implemented to move around the metadata server when zones come up and go down, or when the whole global zone cluster goes up and down.

When you place a file system into a zone cluster, you typically do not place the devices in the zone cluster and there's a reason for that. Raw devices have ioctls that can be issued by applications and some of those ioctls can be mishandled and this may result in machine failures. So if you take away the devices and just give the application a FS there's no access to those dangerous ioctls. Just because you place the FS in to the zone cluster it doesn't mean the volume doesn't go away, so support has been put in place so you can run shared QFS on top of volume managers and the software takes care of the appropriate work to manage the volumes in the global zone.

The other thing that's been added is to do with storage devices. You can actually place storage devices inside the zone cluster so you can directly mount file systems with regular mount commands. You can also place the storage devices in to the zone cluster so that the Oracle ASM product can run directly on top of raw devices.

Another thing I'd like to talk about in terms of storage is that all of the volume manager administrator work i.e. creating a volume, changing the number of disks in a volume and that kind of work is all done in the global zone. Its just the resultant volume that you create which is placed in to the zone cluster.



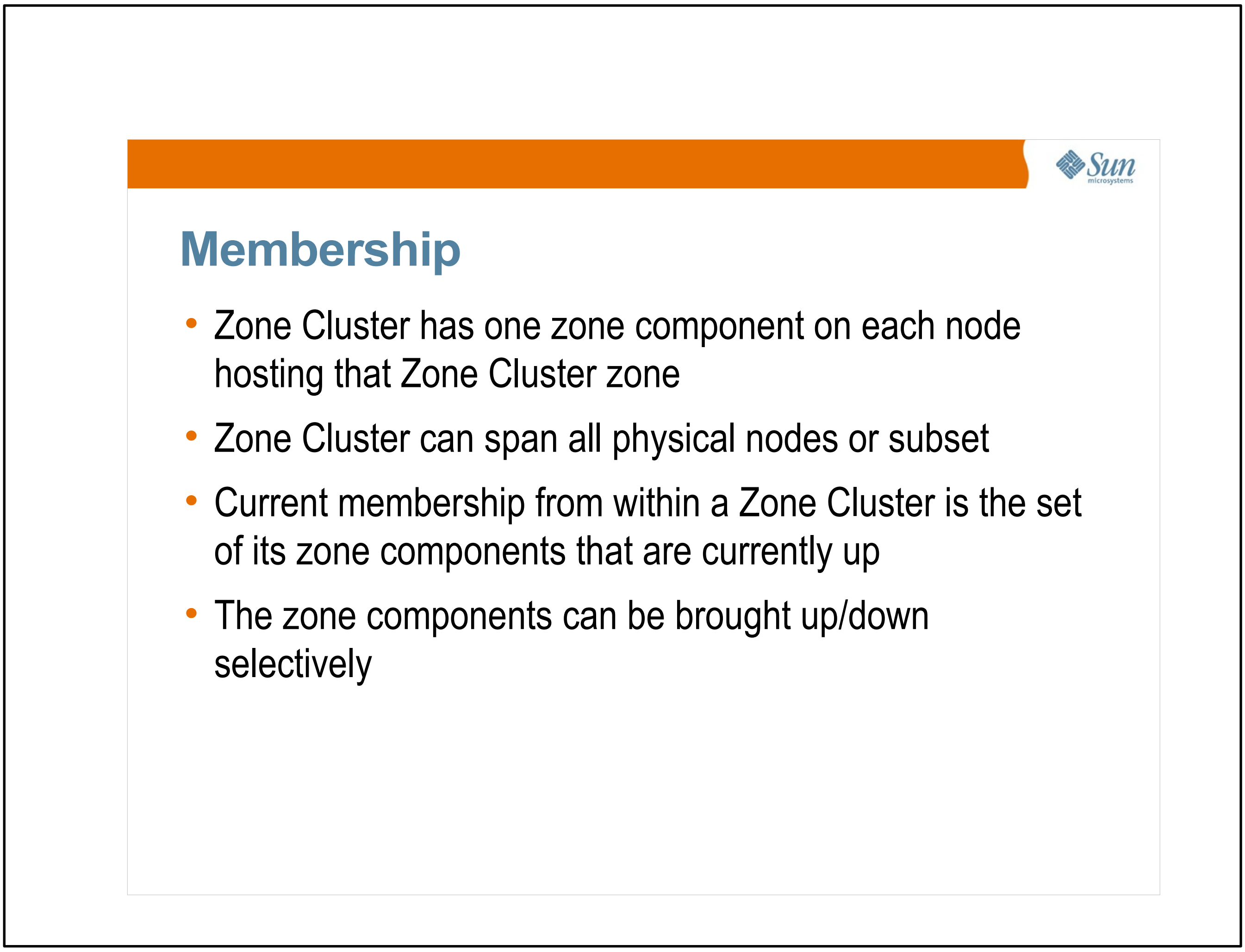
Resource Support - Network

- Private Interconnect IP support
  - Automate setup
  - Distinct Private Interconnect IP Namespace
- Ifconfig support for VIP
  - Support plumb/unplumb/up/down/addif subcommands
  - Commands can be issued safely within zone with respect to security rules

By the time that we create the zone cluster, the global zone cluster has already been setup, which means we have already setup and initialized the private interconnect. So when we setup the private interconnect for the zone cluster we already have all the information we need. All we require is that there be sufficient IP addresses to establish that. The next thing is that there is a namespace that we use to look up the addresses, the namespaces for each of the zone clusters is independent so you can use the same names on all the different zone clusters without having any interference.

The next thing is ifconfig support, I have to diverge a little bit first and talk about two options that have been added to zones. One of them is the feature of IP-type=shared and the other is IP-type=exclusive. IP-type=exclusive is a feature where by an entire NIC and its entire stack is assigned exclusively to a particular zone. The other type of resource is the one that's called IP-Type=shared and in that configuration the NIC is shared between zones and the way that we provide security is that zone is restricted to a specific IP address and NIC combination, so we can give a logical interface to that particular zone cluster. With our initial product we are going to support iptype=shared and the reason for that is in solaris 10 we do not have virtual NIC support for all physical NIC's. Also if you take a look at a cluster you'll find that you'll want two NICs to have high availability for the private cluster interconnects and have two nics for the public networks. That's a total of 4 NICs. Now take some of the machines that are available today , I require four nics for the global cluster and I create my first zone cluster, that's eight NICS. You can see we are quickly exhausting the number of physical NICs that you can place on most machines.

OK, so we've chosen the network feature and used the iptype=shared, in a zone there are no privileges to allow you to plumb up new logical interfaces. Well oracle RAC requires that, and there are a number of sub commands to ifconfig that they need to use as well. We want to be able to support RAC without changing it, so what has been done is that ifconfig has been modified, well actually a proxy has been created where it looks at the request and forwards it to the global zone. A security check is also done, to make sure that the particular zone cluster is authorized to work with that particular IP address and NIC relationship. If it is authorized then the request is passed off to the real ifconfig in the global zone to go off and perform the actual operation. This means that Oracle RAC can issue their ifconfig commands to support their VIP resources which are IP address resources. At the same time we provide the capability in a way that respects all the security restraints, so you have the capability while retaining the security.

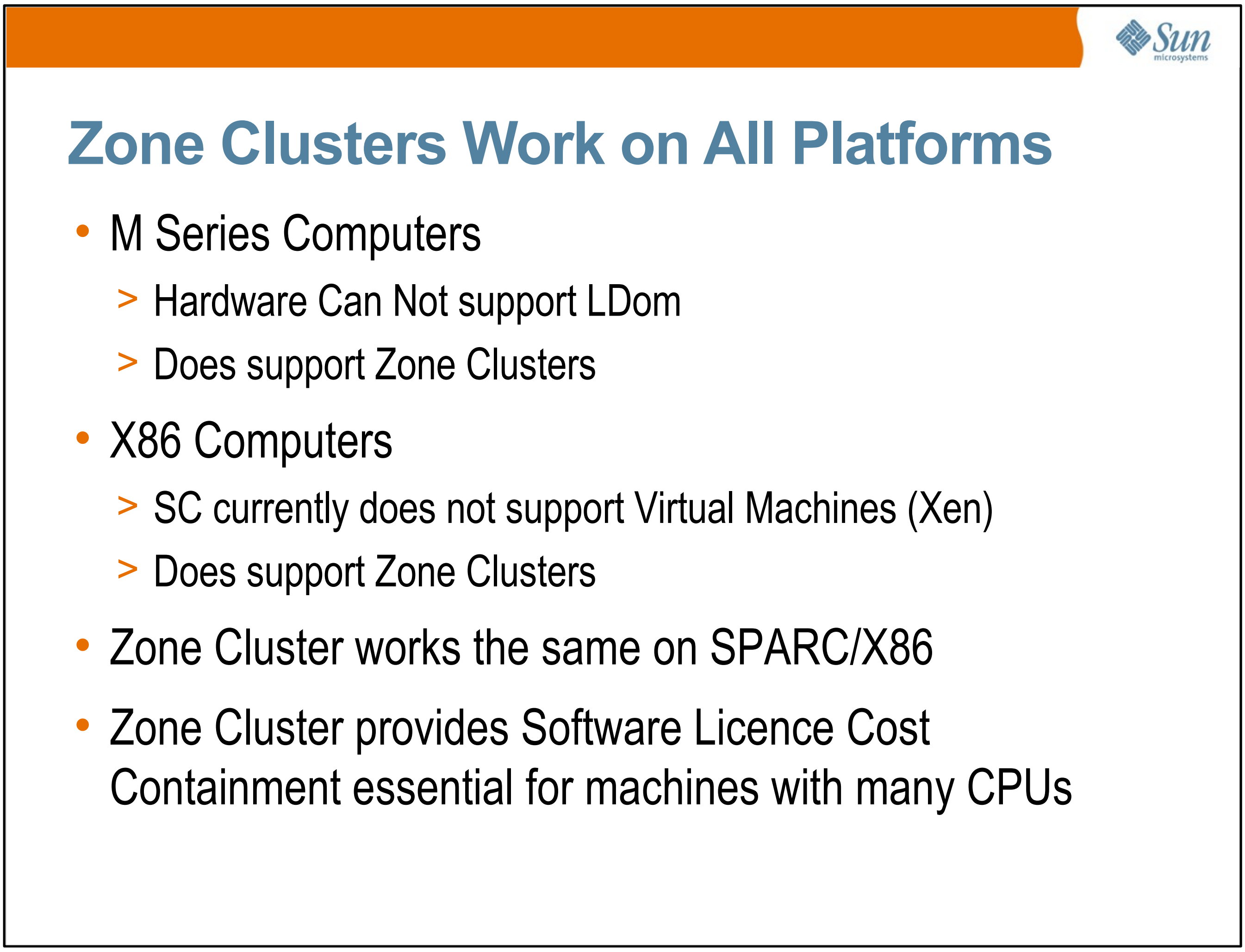


Membership

- Zone Cluster has one zone component on each node hosting that Zone Cluster zone
- Zone Cluster can span all physical nodes or subset
- Current membership from within a Zone Cluster is the set of its zone components that are currently up
- The zone components can be brought up/down selectively

The area of membership functionality has been added to provide a separate membership for each zone cluster. In our virtual cluster which is our zone cluster the virtual nodes, which are zones go up and down, more technically they can halt or boot independently, so the membership of one zone cluster can be different for all the other memberships of the other zone clusters. In fact it can be different from the global zone cluster. So membership information is provided to those applications that need it and most specifically that includes Oracle RAC. They need a membership that tells them which nodes are up or down.

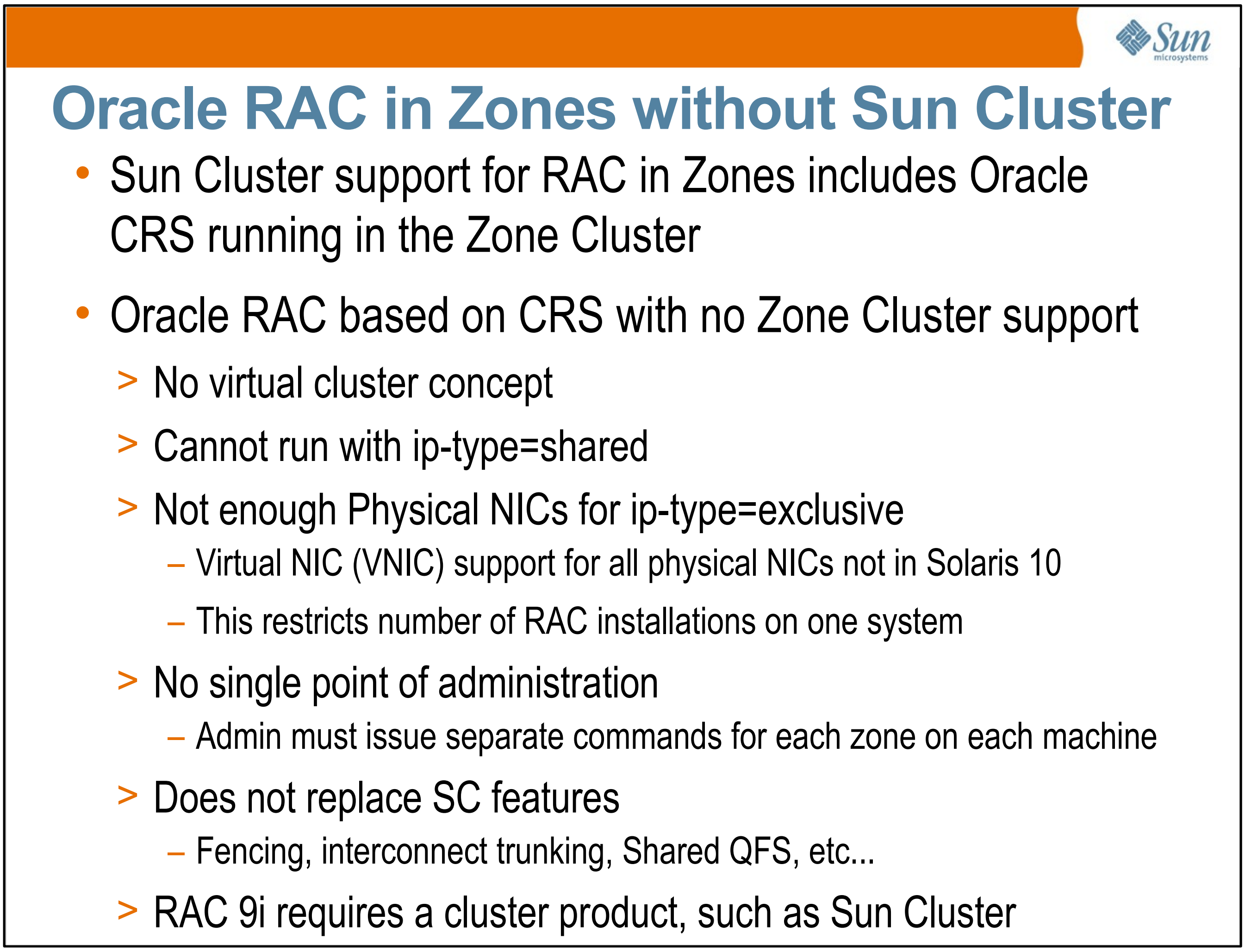
A command has been provided where you can boot an entire cluster at one time or select to boot or halt nodes individually.



Zone Clusters Work on All Platforms

- M Series Computers
  - > Hardware Can Not support LDom
  - > Does support Zone Clusters
- X86 Computers
  - > SC currently does not support Virtual Machines (Xen)
  - > Does support Zone Clusters
- Zone Cluster works the same on SPARC/X86
- Zone Cluster provides Software Licence Cost Containment essential for machines with many CPUs

Sun produces a series of computers called the M series and that hardware is not capable of supporting LDOMS, so zones is the only virtualization option available on that hardware. The other area is that of X86 computers, currently Sun cluster does not support the virtual machine solution (Xvm Server), but zone cluster works on x86. So today zone clusters works on all the platforms that Sun sells and it works the same on both sparc and x86. I'd like to point out that some of the boxes coming out of Sun are going to have large number of cpu's so the cost containment feature of sun cluster is very important.

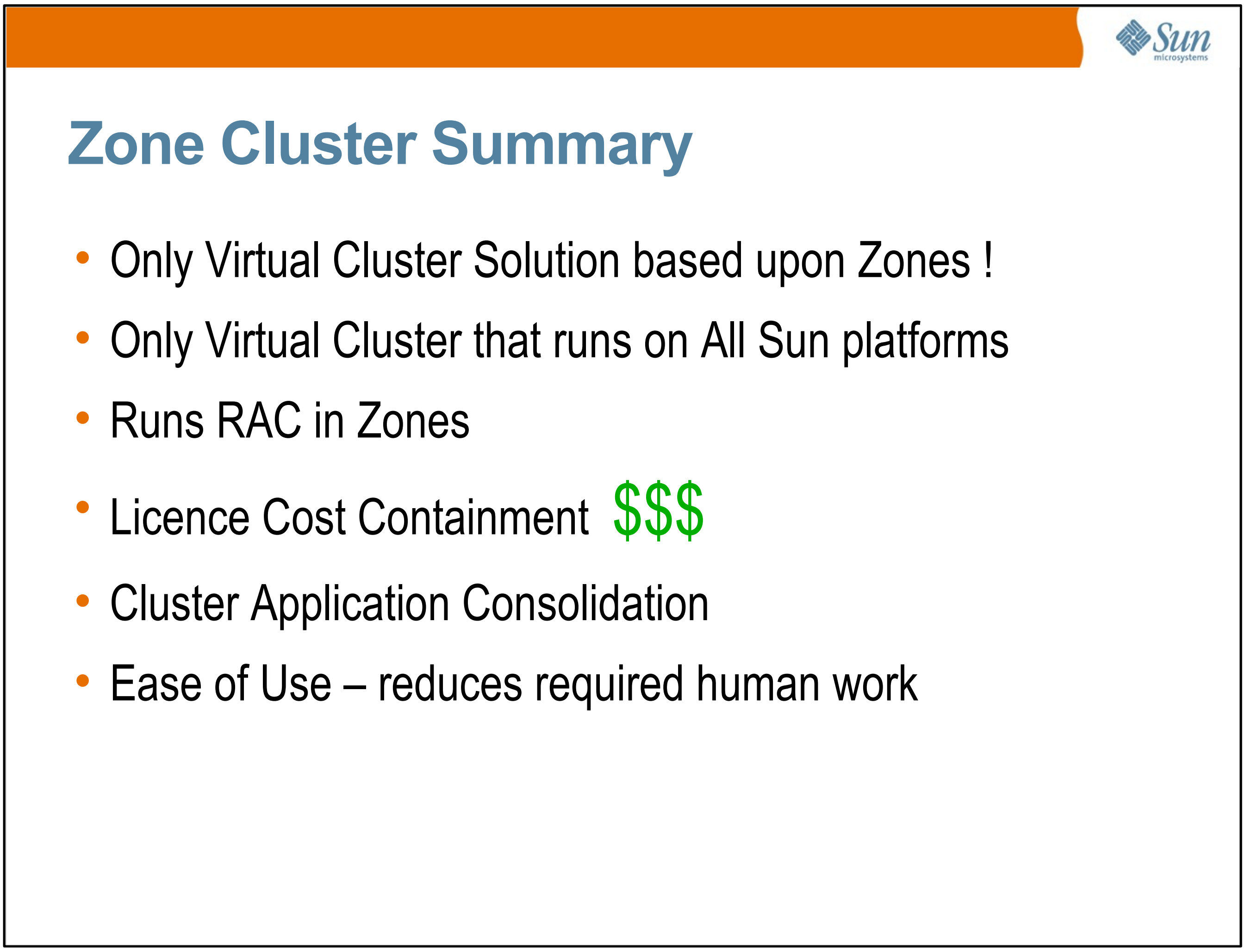


Oracle RAC in Zones without Sun Cluster

- Sun Cluster support for RAC in Zones includes Oracle CRS running in the Zone Cluster
- Oracle RAC based on CRS with no Zone Cluster support
  - No virtual cluster concept
  - Cannot run with ip-type=shared
  - Not enough Physical NICs for ip-type=exclusive
    - Virtual NIC (VNIC) support for all physical NICs not in Solaris 10
    - This restricts number of RAC installations on one system
  - No single point of administration
    - Admin must issue separate commands for each zone on each machine
  - Does not replace SC features
    - Fencing, interconnect trunking, Shared QFS, etc...
  - RAC 9i requires a cluster product, such as Sun Cluster

If you want to run Oracle RAC without zone cluster, what do you have to look forward too. Well it is possible, however RAC does not have a virtual cluster concept based upon zones. The next thing is that it cannot run in a configuration where the feature iptype=shared because they don't have the ability to be able to plumb the IP resources. That means they have to use iptype=exclusive and depending on what type of NICs you have, you may be severely limited to the number of zone clusters that you can create because of the restriction of the number of physical nics that are present. The next thing that sun cluster provides is a single point administration - one command that can be run to administer the whole cluster. With Oracle RAC you have to create each zone independently and then also work independently on the different machines. Another point is that there are a whole load of features that you get with Sun Cluster that you don't get with RAC, for example strong fencing support, we have interconnect trunking , we have shared QFS filesystems and a variety of other features.

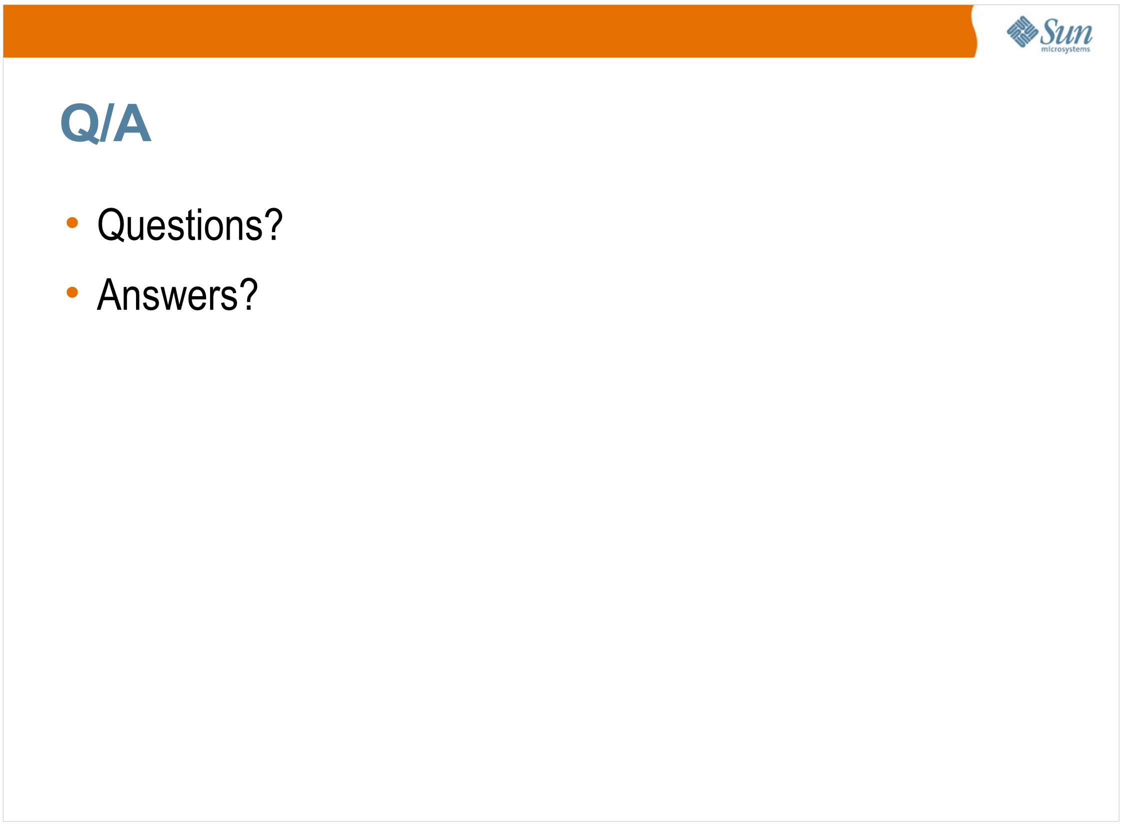
The last thing I'd like to mention is that Oracle RAC 9i doesn't come with CRS, so you'll need a cluster product to support Oracle RAC 9i



Zone Cluster Summary

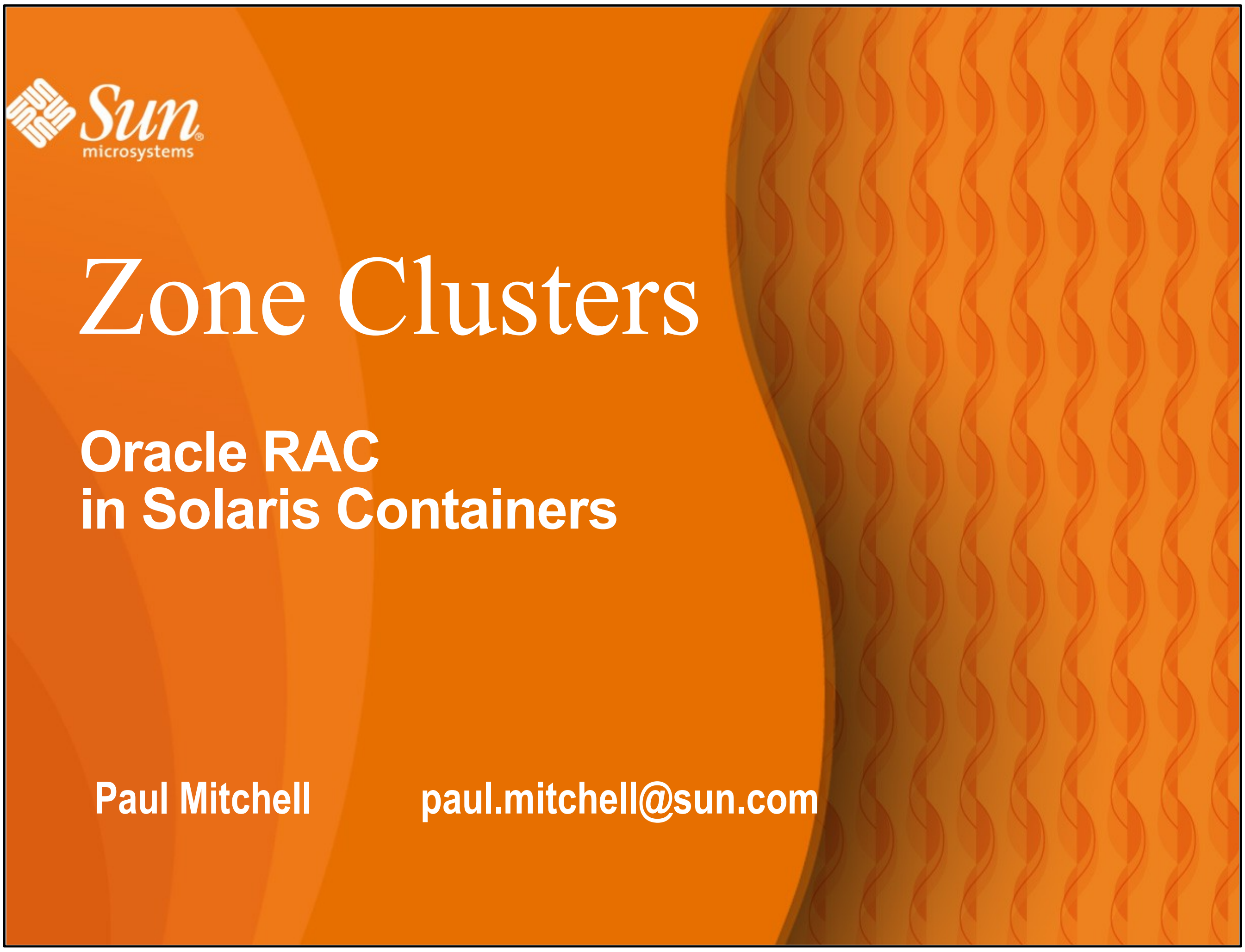
- Only Virtual Cluster Solution based upon Zones !
- Only Virtual Cluster that runs on All Sun platforms
- Runs RAC in Zones
- Licence Cost Containment $$$
- Cluster Application Consolidation
- Ease of Use – reduces required human work

In summary, we have the only virtual cluster solution that I'm aware of that is based upon zones. It runs on all sun platforms and it runs RAC in zones where RAC is unmodified. It has the strong license cost containment feature and you can consolidate all applications. I will point out that not all of the Cluster data services have been qualified at this time. Finally we have a lot of “ease of use” features to manage virtual clusters



Q/A

- Questions?
- Answers?



Zone Clusters

Oracle RAC in Solaris Containers

Paul Mitchell paul.mitchell@sun.com