



# Solaris Cluster

High Availability in Solaris

Tomáš Dzik

Solaris Cluster Revenue Product Engineering

Sun Microsystems, Czech

---

# Goals of this presentation

---

- Explain what is Solaris Cluster (SC)
- To show where SC could be used
- Explain concepts and architecture of SC
- Show how to make your application Highly Available (HA)

# Definition

---

**“Cluster is a collection of loosely coupled computing nodes that provides a single client view of network services or application.”**

# Solaris Cluster (SC)

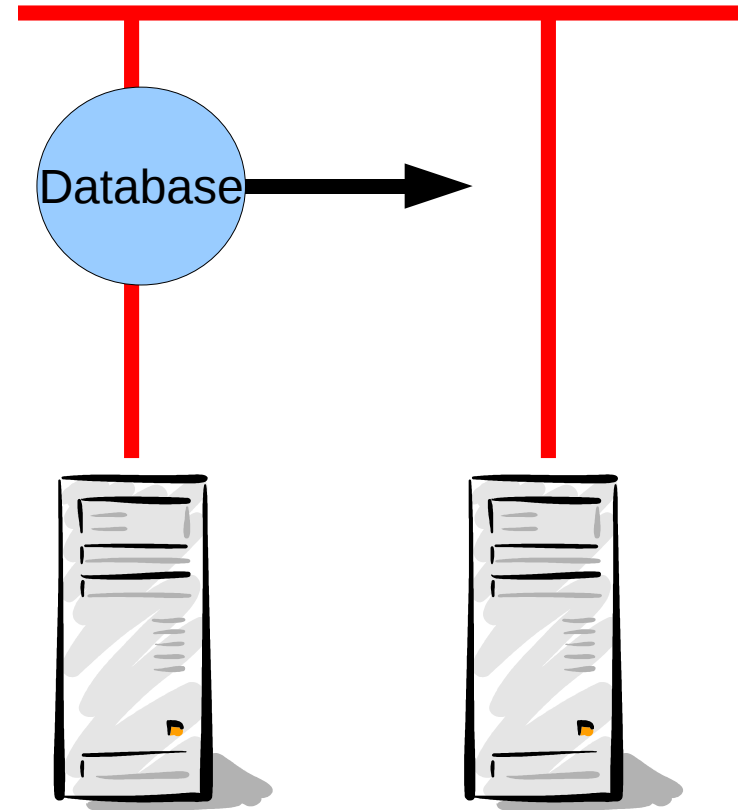
---

- Is High Availability solution from Sun
- Supports also scalable services
- Runs on Solaris
- Open High Availability Cluster (OHAC) runs on OpenSolaris and it is open sourced
- Most of stock application running on Solaris could be made HA using SC without changing an application code

# Basic Idea – What Administrators Do

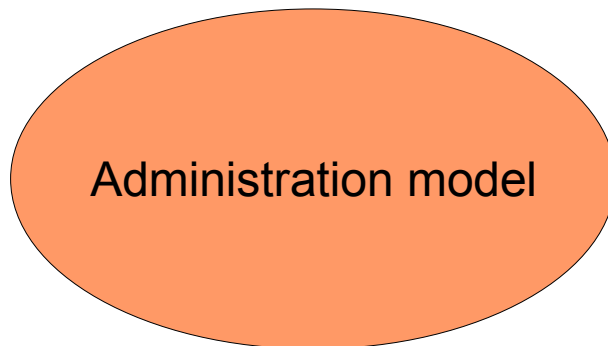
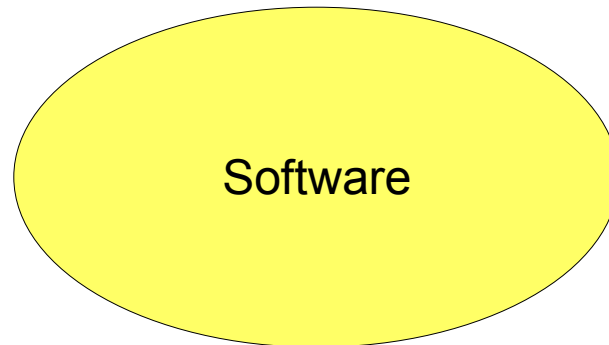
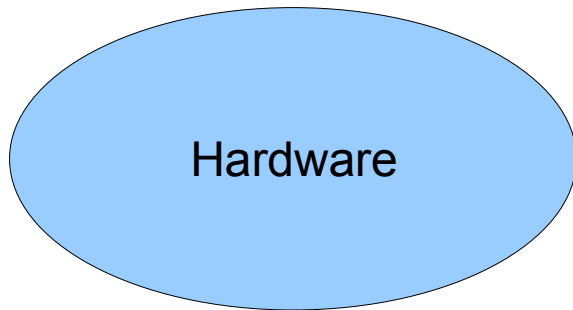
---

- Monitor your application
- If application fails, try to restart it.
- If application fails again, move disk with data to another computer and start it again



# Views of Solaris Cluster

---



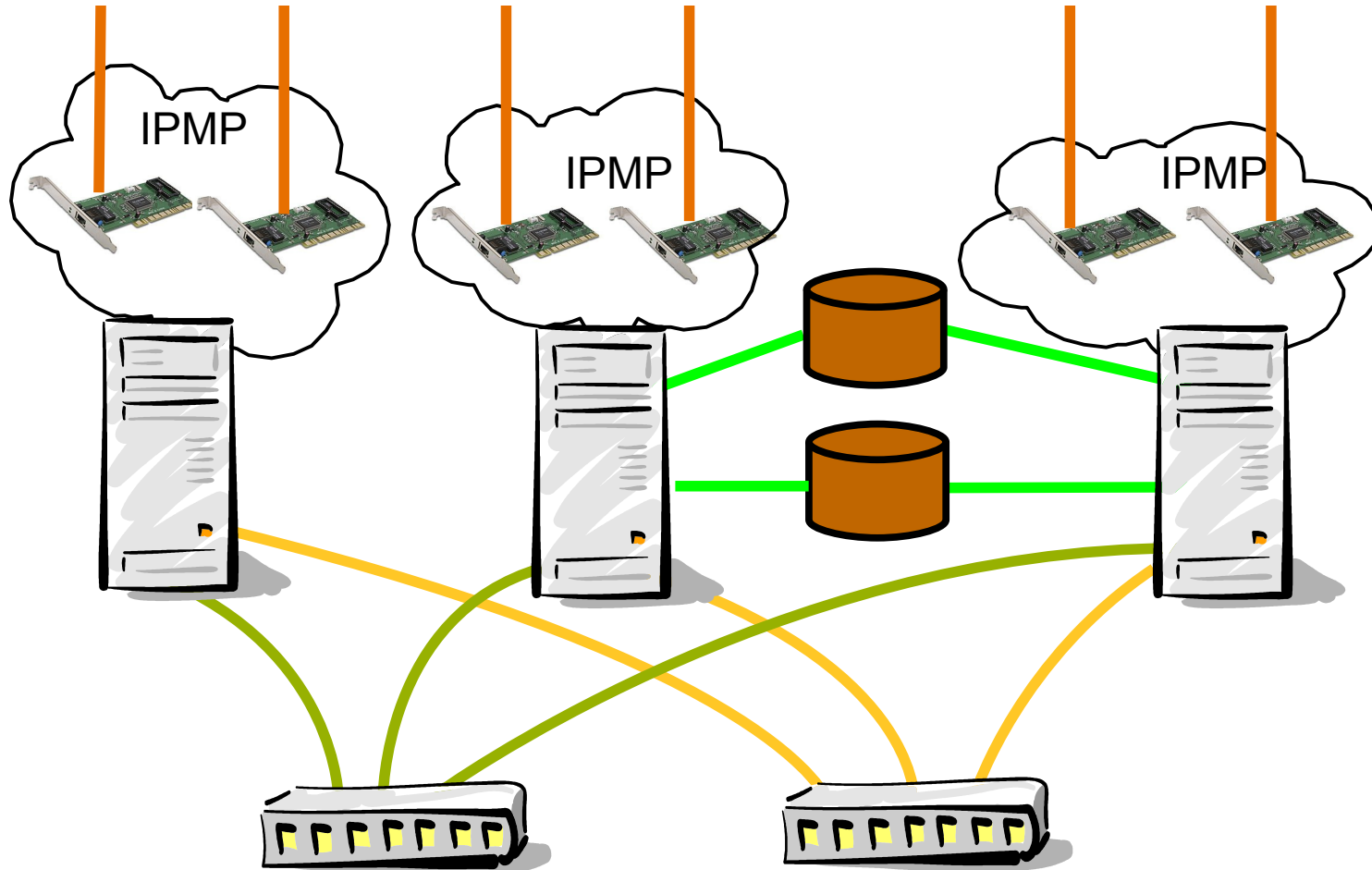
# Three Views of Solaris Cluster

---

- Hardware
- Software stack
  - Data services
  - Services provided by SC
- Abstraction for administrators
  - Resource groups, resources, dependencies, ...

# Hardware

---



# Hardware Components

---

- Cluster nodes
  - Computers with their own local boot disk
- Multihost storage (SCSI, FC or SATA)
  - Not necessary for OHAC (could use iSCSI)
  - I would recommend using Fibre Channel
- Cluster interconnect
  - Either ethernet or Infiniband
- Public NIC (it is usually ethernet cards)
- Clients

# Multihost Storage

---

- Uses SCSI 2 or SCSI 3 reservations
  - Helps to protect data integrity in case of failure
- Uses multi-initiator SCSI for shared SCSI bus
- Since SC 3.2u2 can use SATA disks
  - Uses software quorum
- Storage could be accessed
  - Through primary node (most of applications)
  - From more nodes simultaneously if application manages locking (for example Oracle RAC)

# Cluster interconnect

---

- Usually two 1 Gb ethernet cards
  - Requirement was relaxed recently – even 1 interconnect could be enough
  - For OHAC virtual NICs over public interface can replace proper interconnects
  - It is possible to configure IPsec on interconnect in OHAC
- Junctions
  - Ethernet switches
  - Crossed cables

# Public Network Interface Cards

---

- Uses IPMP – IP Multi Pathing for failure detection
  - Standard Solaris feature used in cluster
  - TCP connections failing to another adapter are **not** disconnected
  - Ethernet NICs must have unique MAC address
  - Health could be checked by pinging
    - Default router
    - Other routers
    - `ping -s 224.0.0.1`

# Software stack

---

- Solaris or OpenSolaris
- Volume Management (optional)
  - Solaris Volume Manager
  - ZFS
  - Veritas VxVM
- Solaris Cluster Software
- Data service for clustered application
- Clustered application

# Data Service

---

- Code which attach application to SC
- Tells SC (namely to Resource Group Manager) how to:
  - Start application
  - Stop application
  - Monitor application
  - Validate configuration
  - ...
- It is implemented using callbacks

# Data Service

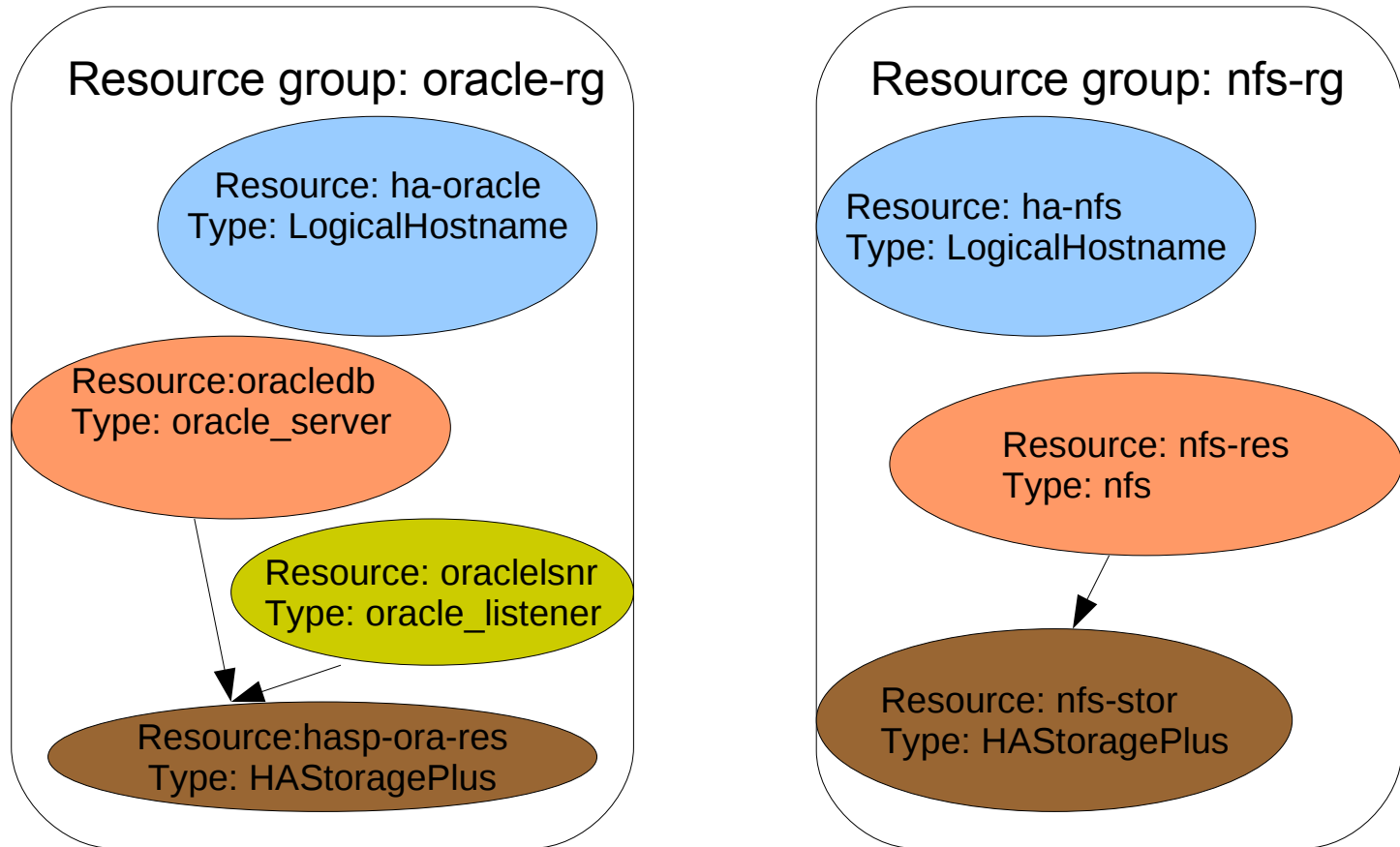
---

An example is worth a thousand words

<http://src.opensolaris.org/source/xref/ohac/ohacds/usr/src/cm>

# Data Services Management Model

---



# Data Services Management Model

---

- Resource group
  - Group of resources which are managed together
- Resource
  - For example highly available IP address, highly available storage or application
  - Resource has always type
- Resources and Resource groups can have dependencies

# Resource Types

---

- Described by resource type registration (RTR) file
- Consists of:
  - Implementation of callback methods
  - A set of properties – see `rt_reg(4)`

# Under the Hood of SC

---

## ○ Global devices

- Each device (even local) could be accessed from any node of cluster
- Path: `/dev/global`
- Device could be disk, CD-ROM or tape

## ○ Global namespace

- Each device could be accessed using the same name on all nodes
- Simplifies administration
- Path: `/dev/did/`

# Device groups

---

- Entity used for managing global devices
  - For each disk SC automatically generates device group
  - Device group exists also for each SVM disk set or VxVM disk group
  - Only 1 node has physical access do device group
    - Disk group is imported on that node
    - Other nodes have access using cluster interconnect
  - Oracle RAC uses shared disk groups

# Cluster filesystem - pxfs

---

- Transparent shared filesystem with POSIX semantics
- PxFS – proxy filesystem
  - Act's as proxy – send operations to appropriate node
  - Usefull mainly with scalable applications
- Could be managed by HAStoragePlus resource type

# Quorum

---

- 2 common problems

- Split brain
- Amnesia

- Solution

- Every node has 1 vote. Cluster is operational, only if it has majority (more than half) of votes
- For 2 node clusters, we need another vote – we give the vote to device
- Quorum device has  $N-1$  votes.  $N$  is number of “votes connected to this device”

# Quorum

---

- Quorum device could be
  - Multihosted ( $\geq 2$ ) shared disk supporting SCSI-3 PGR
  - Dual-hosted ( $=2$ ) shared disk supporting SCSI-2 reservations
  - Quorum server process running on the quorum server machine – doesn't need to be dedicated machine
  - Since SC 3.2u2 software quorum on any supported multihosted disk

# Quorum – Weak Membership In OHAC

---

- <http://opensolaris.org/os/project/colorado/De>
- Another type of membership
- Split-brain is possible
- Data loss is possible
- Use only in applications, where availability is more important than data integrity

File Edit View Go Bookmarks Tools Window Help

Back Forward Reload Stop <https://c4u-v240x.czech:6789/SunClusterManager/index/Index> Search Print

Home Bookmarks Java Desktop System Sun Microsystems

Sun SunWeb: SMI Internal Po... CLUSTNEWINFO.pdf (app... Dorfl > OurLab http://clust...CDevTOI.html TOI < Tech < TWiki Sun Cluster Manager

APPLICATIONS VERSION REFRESH LOG OUT HELP

User: root Server: c4u-v240x

# Sun™ Cluster Manager

Sun™ Microsystems, Inc.

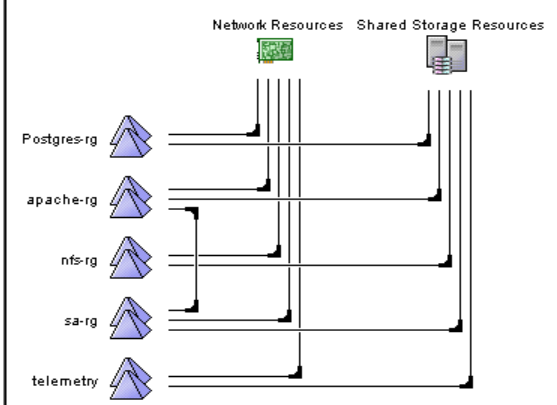
- ha-v240
  - Tasks
  - Nodes
  - Resource Groups**
    - Postgres-rg
    - apache-rg
    - nfs-rg
    - sa-rg
    - telemetry
  - Storage
  - Private Interconnects
  - IPMP Groups
  - Quorum
  - SNMP Modules

**Status** **Topology** **Utilization**

RG to Nodes RG Dependencies RG Affinities Resource Dependencies

## Resource Group Dependencies Topology

This page illustrates the dependencies between resource groups. For an explanation of the icons, see [Icons and Conventions](#).



User: root Server: c4u-v240x

# Sun™ Cluster Manager



Sun™ Microsystems, Inc.

## ha-v240

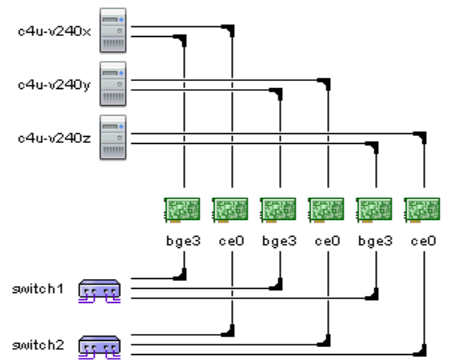
- Tasks
- Nodes
  - c4u-v240x
  - c4u-v240y
  - c4u-v240z
- Resource Groups
  - Postgres-rg
  - apache-rg
  - nfs-rg
  - sa-rg
  - telemetry
- Storage
  - Device Groups
  - Nas Devices
  - Disks
- Private Interconnects
- IPMP Groups
- Quorum
- SNMP Modules

Status
Properties
Topology
Utilization

Private Interconnect
Resource Groups
Resource Dependencies
Device Groups

### Private Interconnect Topology

This page illustrates the private network connections between cluster nodes. For an explanation of the icons, see [Icons and Conventions](#). For information about private cluster interconnects, see [Private Cluster Interconnects](#).



# Solaris Cluster

---

Questions ???



# Solaris Cluster

High Availability in Solaris

Tomas.Dzik@sun.com

---