opensolaris

© CZOSUG & GUUG

developer conference

October 28 - 30, 2009, Dresden

High Availability with a minimal Cluster

29. October 2009

Thorsten Früauf
Availability Engineering
Sun Microsystems GmbH

Agenda

- Motivation
- Open HA Cluster 2009.06
- Minimal HA Configuration
  - Weak Membership
  - COMSTAR / iSCSI / ZFS
  - Crossbow
  - IPS
- Live Demo
- References

Why care about High Availability?

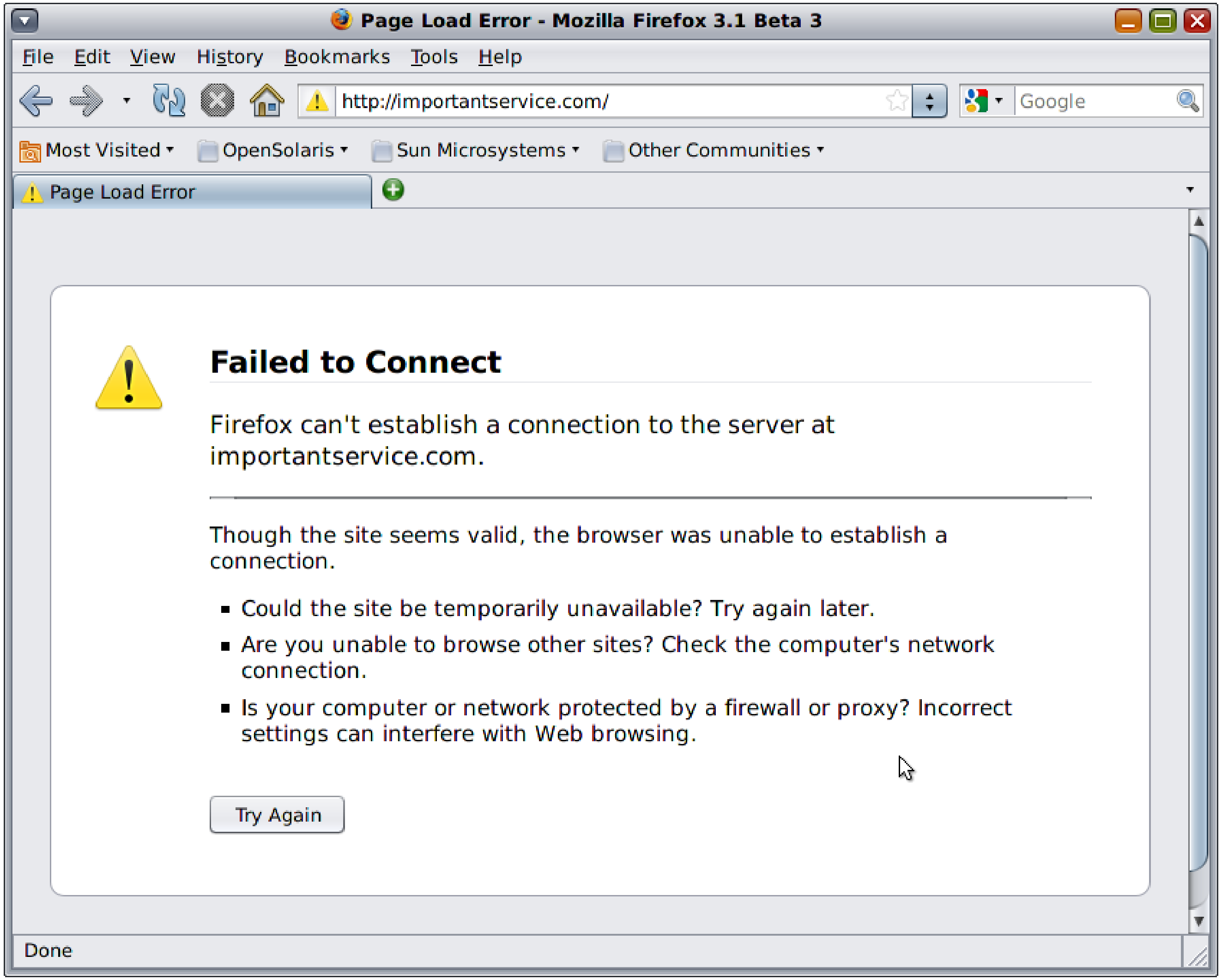
- Computer systems provide services:
  - Web Services, Databases, Business Logic, File Systems, etc.
- Downtime is costly
  - Services should be available as close as possible to 100% of the time

- Failures are inevitable:
  - Software Bugs
  - Hardware components
  - People and Processes
  - Natural Disaster
  - Terrorism

The Goal of High Availability

HA Cluster automate the recovery process from inevitable failures to minimize downtime and cost.

You don't want your users to see this...


Page Load Error - Mozilla Firefox 3.1 Beta 3
File Edit View History Bookmarks Tools Help
http://importantservice.com/
Google
Most Visited
OpenSolaris
Sun Microsystems
Other Communities
Page Load Error

Failed to Connect

Firefox can't establish a connection to the server at importantservice.com.

Though the site seems valid, the browser was unable to establish a connection.

- Could the site be temporarily unavailable? Try again later.
- Are you unable to browse other sites? Check the computer's network connection.
- Is your computer or network protected by a firewall or proxy? Incorrect settings can interfere with Web browsing.

Methods to implement HA

- Redundant hardware
  - physical nodes, network adapters, network paths, storage, storage paths, etc.
- Software monitoring
  - physical nodes, applications, network paths, storage paths, etc.
- Failover to secondary hardware when problems detected

Perceptions of HA Clusters

- complex
- complicated
- heavyweight
- difficult to install
- difficult to use
- requires special hardware
- expensive

Perceptions not completely unfounded...

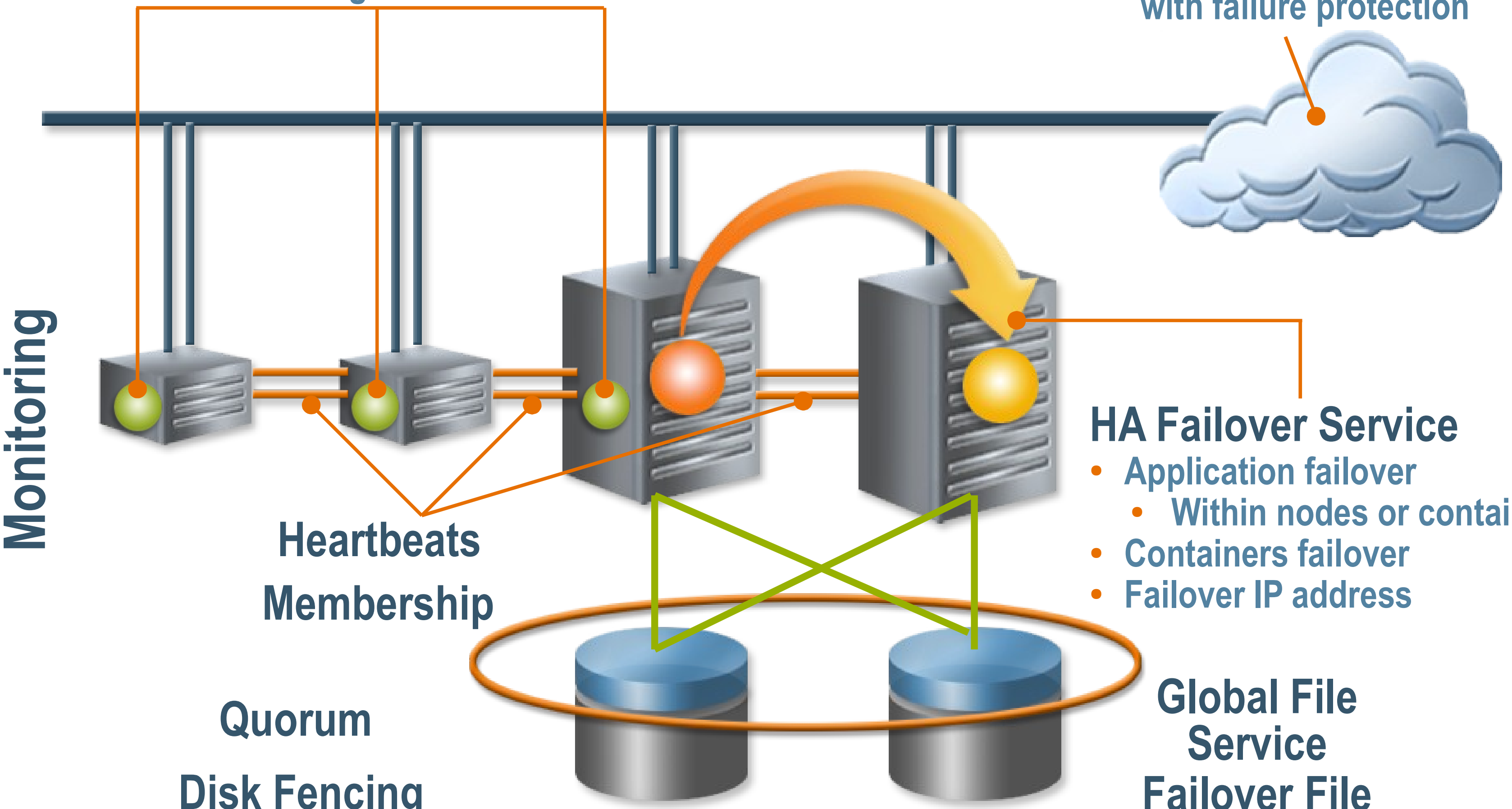
Typical HA Cluster Hardware config

- Two or more physical machines
- Four or more network adapters on each machine
- Dedicated interconnects between nodes
- Shared disk storage
  - multi-homed disks or network-attached storage
- Redundant storage paths from each node
- Quorum arbitration device
- etc.

Typical HA Cluster software components

- Heartbeats
- Membership
- Distributed configuration repository
- Service management
- Cluster-private networking layer
- Global file system
- Network load-balancing
- etc.

Solaris Cluster Architecture



Scalable Service

Software Load Balancing

Global Network Service

Provides Global IP address with failure protection

Monitoring

HA Failover Service

- Application failover
  - Within nodes or container
- Containers failover
- Failover IP address

Heartbeats

Membership

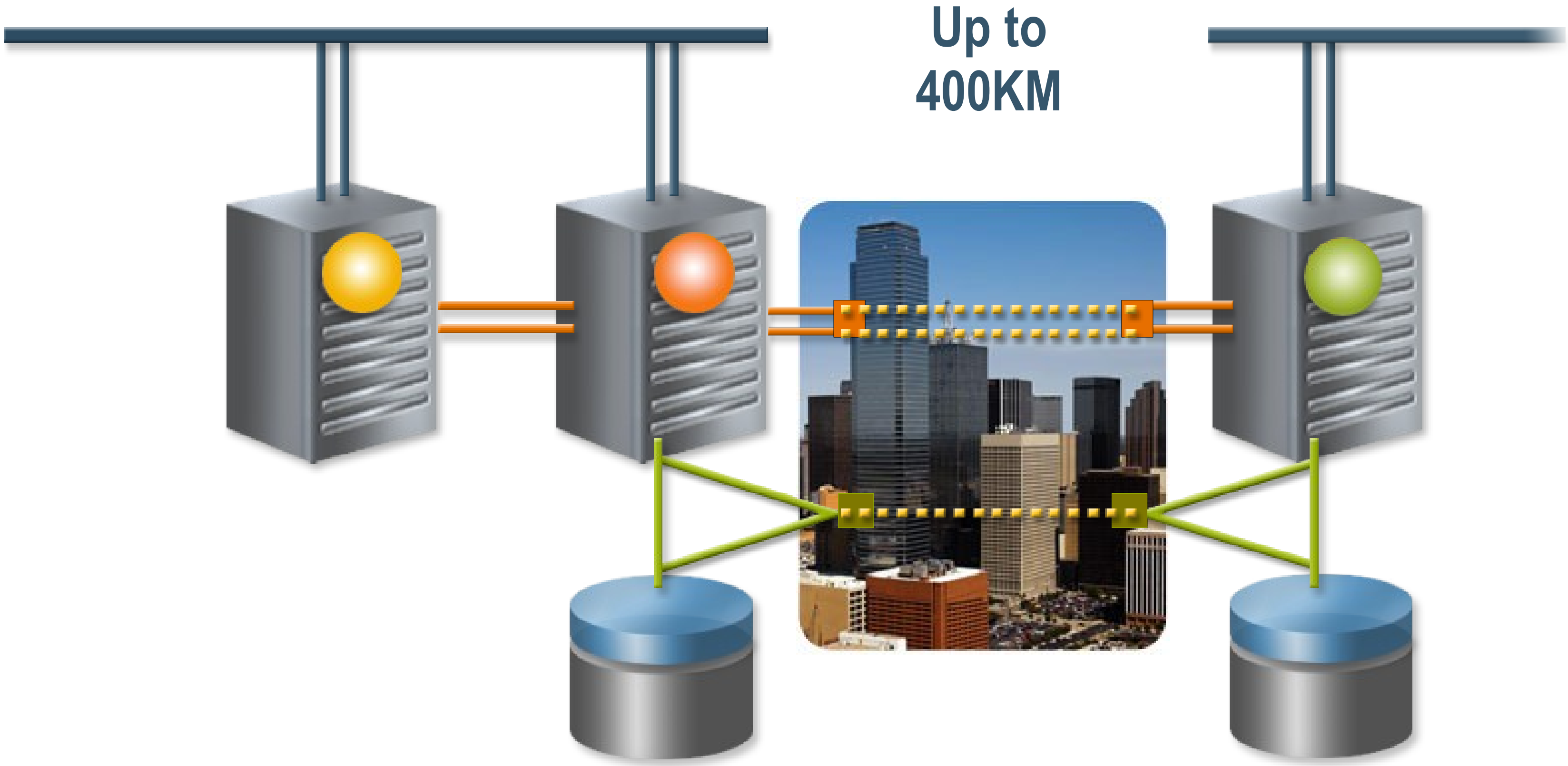
Quorum

Disk Fencing

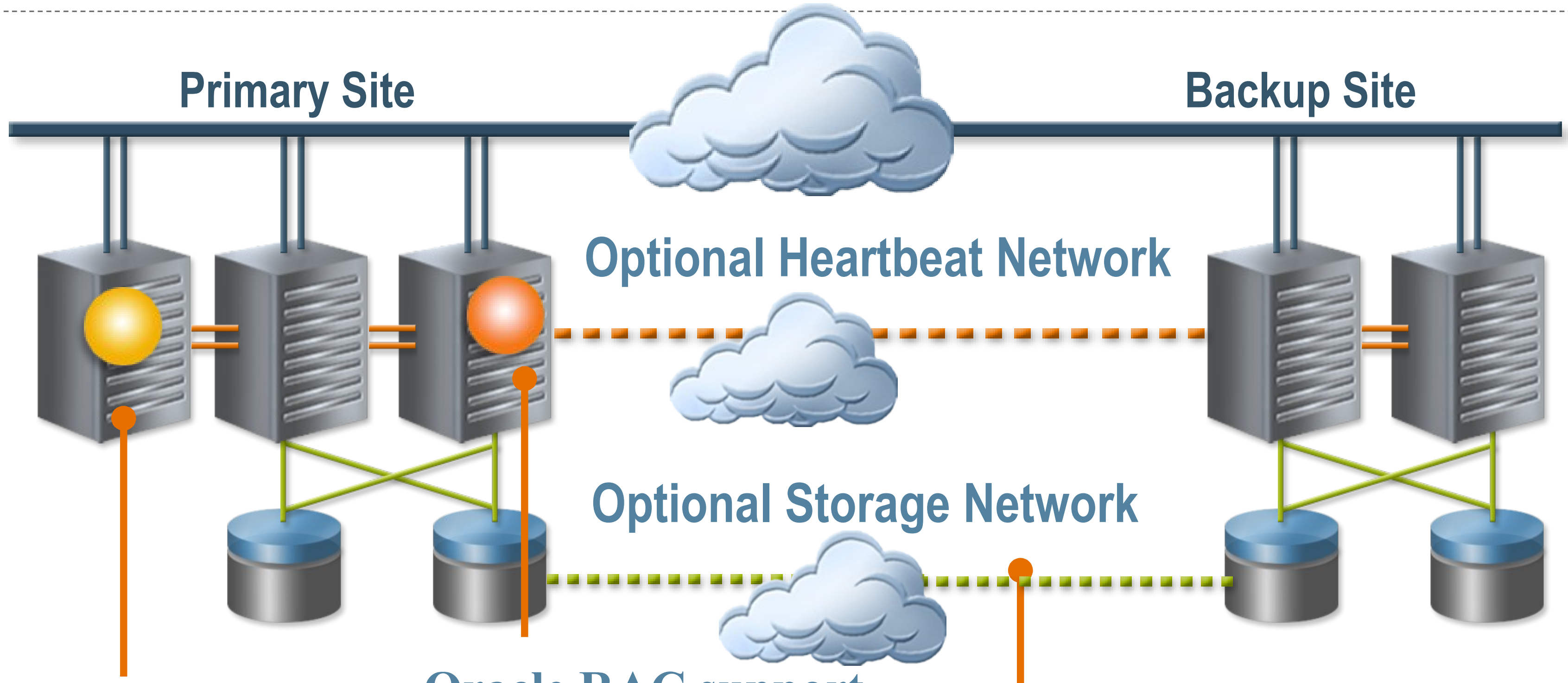
Global File Service

Failover File service

Campus / Metro Cluster



Solaris Cluster Geographic Edition



Primary Site

Backup Site

Optional Heartbeat Network

Optional Storage Network

Oracle RAC support

solaris 9 & 10

opensolaris

intel

ULTRASPARC

Replication

Sun StorEdge Availability Suite 4.0
EMC SRDF
HDS Truecopy
Dataguard for Oracle RAC
Script based plugin for MySQL

Re-evaluate HA Cluster Complexity

- Many use cases (incl. SLA) require all the hardware and software in traditional HA Clusters
- … but not everything... approach: “good enough” is sufficient as well!
- Configure, install, and use only the hardware and software components you actually need

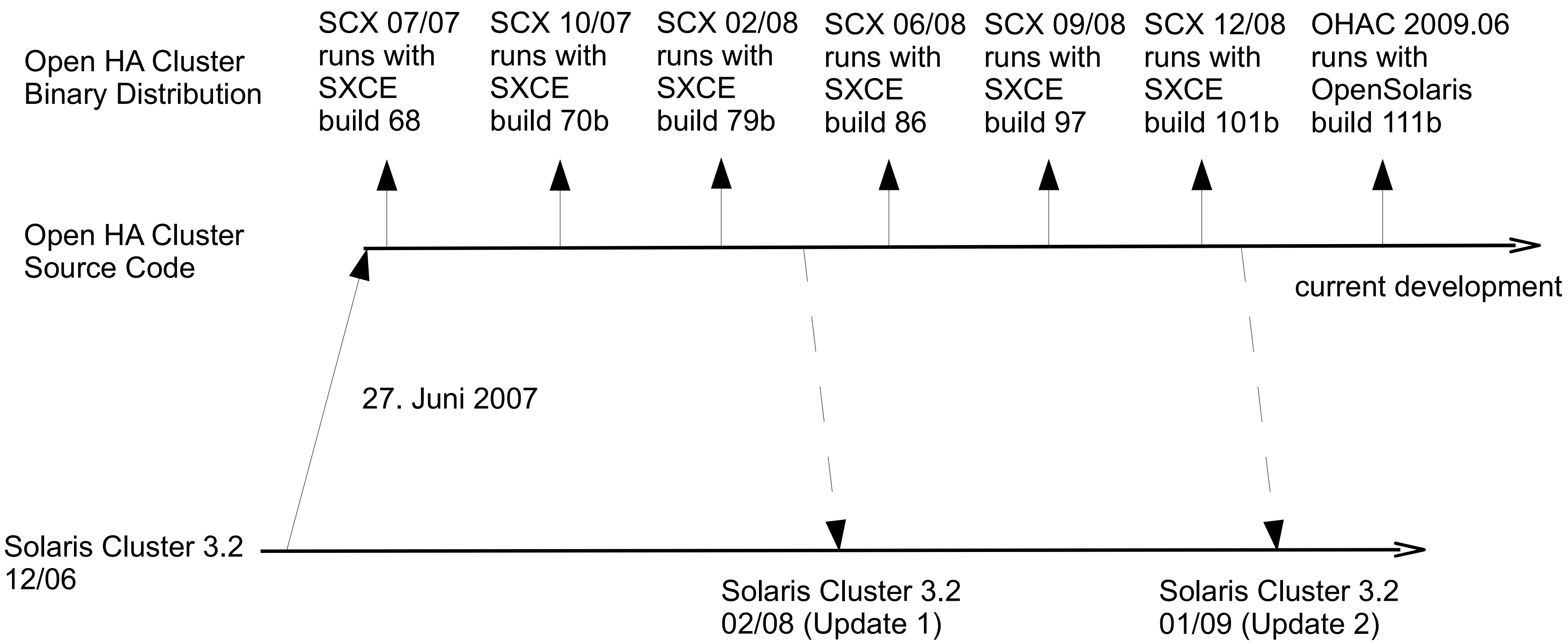
Goals of Project Colorado

- Provide a lightweight, modular, cluster that can run on minimized hardware configurations
- What has been possible before should still be possible to configure
  - as much as OpenSolaris allows

How to Get There

- Port Open HA Cluster source to work with OpenSolaris
- Add hardware minimization features
- Leverage OpenSolaris Image Packaging System (IPS) for software modularity and extensibility
  - Analyze all package dependencies

Development context



Open HA Cluster Binary Distribution

SCX 07/07 runs with SXCE build 68

SCX 10/07 runs with SXCE build 70b

SCX 02/08 runs with SXCE build 79b

SCX 06/08 runs with SXCE build 86

SCX 09/08 runs with SXCE build 97

SCX 12/08 runs with SXCE build 101b

OHAC 2009.06 runs with OpenSolaris build 111b

Open HA Cluster Source Code

current development

27. Juni 2007

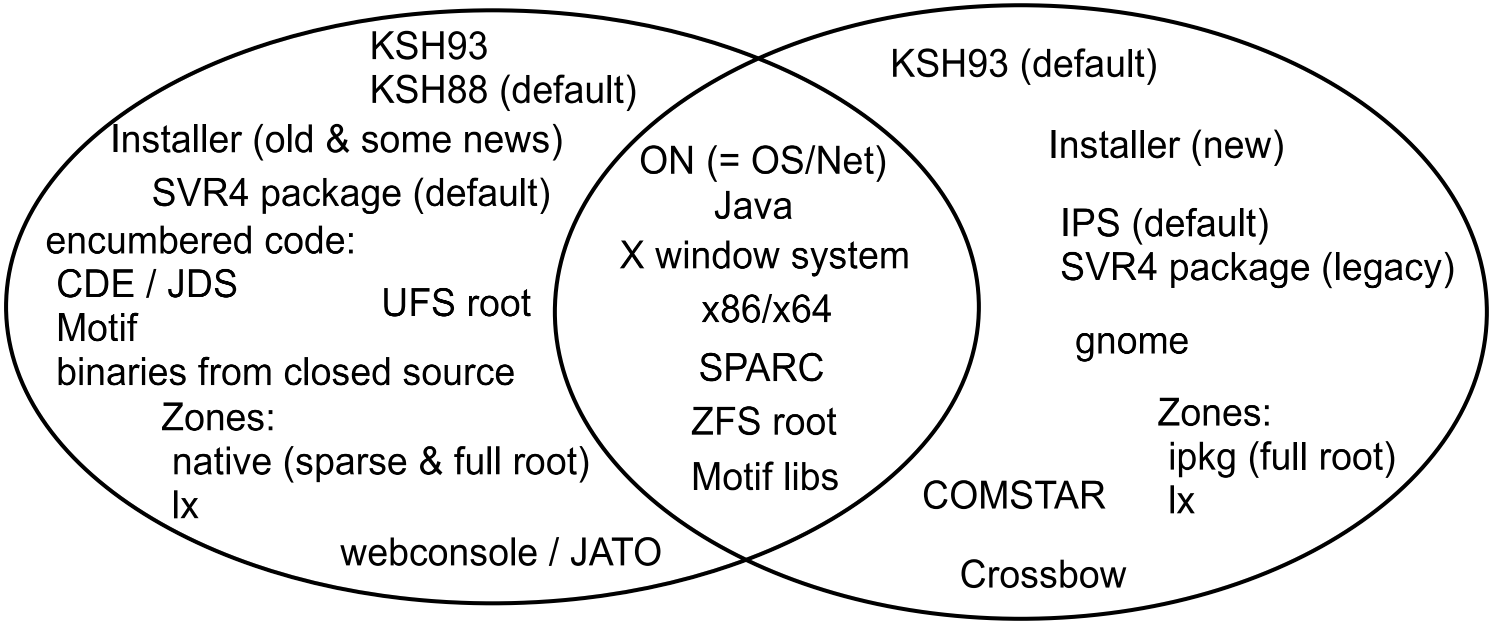
Solaris Cluster 3.2 12/06

Solaris Cluster 3.2 02/08 (Update 1)

Solaris Cluster 3.2 01/09 (Update 2)

SCX = Solaris Cluster Express
OHAC = Open HA Cluster
SXCE = Solaris Express Community Edition

Solaris Express vs. OpenSolaris



KSH93
KSH88 (default)
Installer (old & some news)
SVR4 package (default)
encumbered code:
CDE / JDS
Motif
binaries from closed source
UFS root
Zones:
native (sparse & full root)
lx
webconsole / JATO

ON (= OS/Net)
Java
X window system
x86/x64
SPARC
ZFS root
Motif libs

KSH93 (default)
Installer (new)
IPS (default)
SVR4 package (legacy)
gnome
Zones:
ipkg (full root)
lx
COMSTAR
Crossbow

Solaris Express (Nevada)

Binary distribution of
usr/src und usr/closed
not freely redistributable

OpenSolaris 200X.Y

Binary distribution on LiveCD
freely redistributable packages
(pkg.opensolaris.org Repo)

not freely redistributable packages
(pkg.sun.com Repo)

Open HA Cluster 2009.06 (Colorado-I)

- Runs on OpenSolaris 2009.06 (SPARC & x86/x64)
- Many features from Solaris Cluster 3.2 available
- Free to use (without support)
  - Support subscriptions available
- Installation from IPS package repository https://pkg.sun.com/opensolaris/ha-cluster
- Source is open and freely available at http://www.opensolaris.org/os/community/ha-clusters/

Open HA Cluster 2009.06 Agents

- Apache Webserver
- Apache Tomcat
- MySQL
- GlassFish
- NFS
- DHCP
- DNS
- Kerberos
- Samba
- HA Containers
  - ipkg Zones

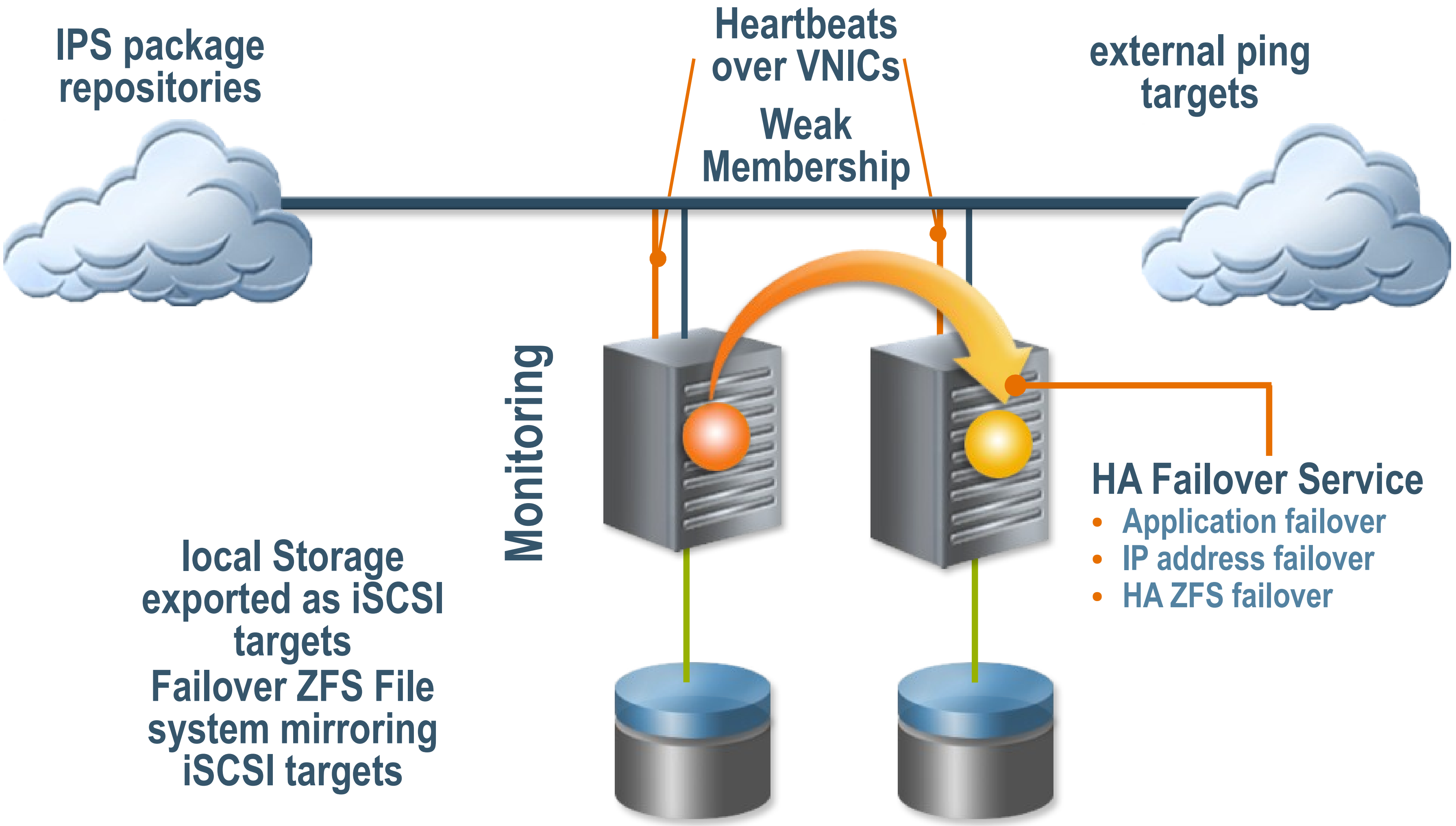
- Generic Data Service (GDS)

Hardware Minimization

- Using local disks as “Poor man's shared storage” with COMSTAR iSCSI and ZFS
- Using Crossbow VNICs for private cluster traffic over public network
- “Weak membership” (preview-only feature)

Taken together, allow any two-nodes on the same IP subnet to form a functional cluster.

Minimale HA Konfiguration



local Storage exported as iSCSI targets

Failover ZFS File system mirroring iSCSI targets

HA Failover Service

- Application failover
- IP address failover
- HA ZFS failover

Technologies useable for Minimization

- Weak Membership
- Software Quorum
- Quorum Server
- Optional Fencing
- HA ZFS

- COMSTAR / iSCSI
- IPsec
- Crossbow
- IPS
- VirtualBox
  - for training and development

HA Cluster „Strong Membership“

- Use concept of quorum to ensure cluster consistency in the presence of partitions in space and time
  - Partition in space (network partition) can cause split-brain
  - Partition in time can cause amnesia
- Two-node cluster requires third arbitration device in case of partitions
  - Typically hardware disk or software quorum server

Weak Membership (preview feature)

- Run a two-node cluster without a quorum device
- External “ping target” used as “health check” to arbitrate in case of split-brain
  - Worst-case, both nodes stay up and provide service
  - OpenSolaris Duplicate Address Detection (DAD) can mitigate somewhat
- Places importance of availability above data integrity
  - Can lead to data loss

Why use Weak Membership?

- Read-only or read-mostly applications
- Availability is more important than data integrety
  - the SLA matches (solution is “good enough”)
- Test Cluster with limited resources
- Demos
- Development
- Training

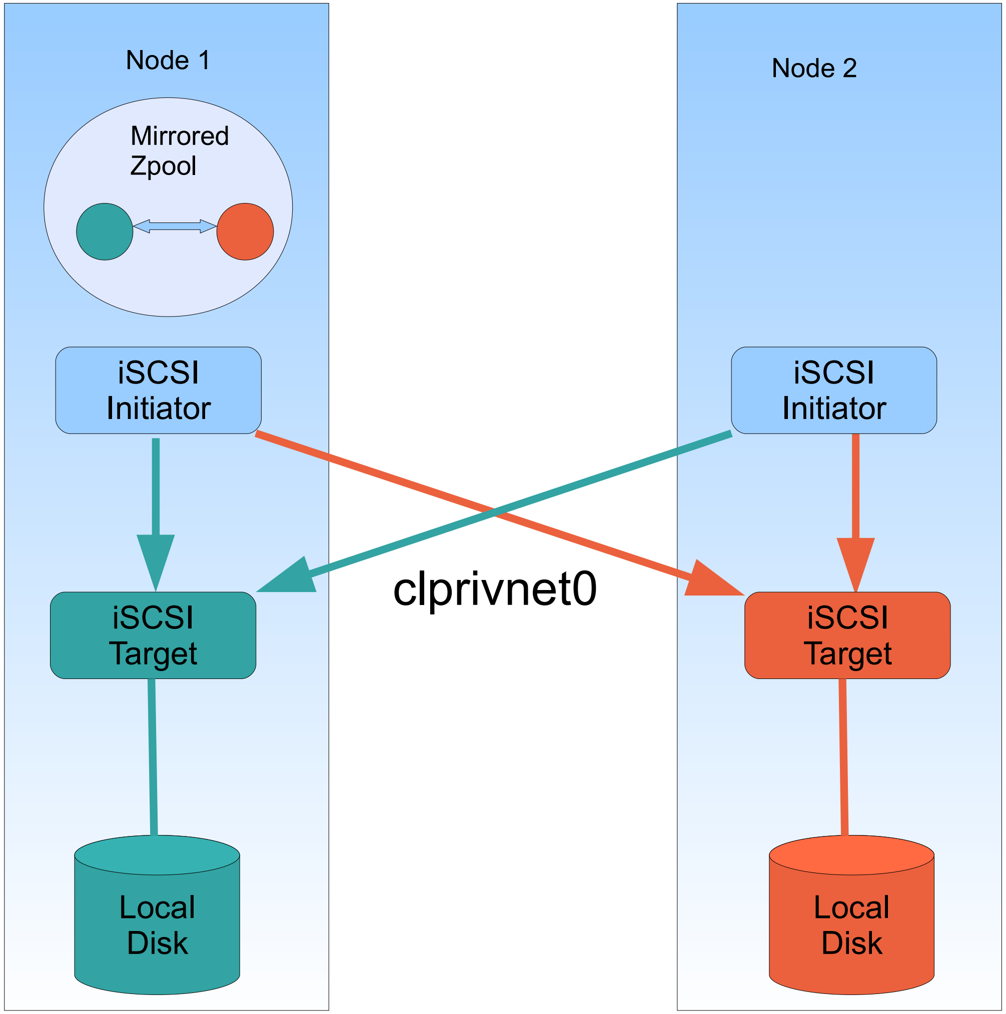
iSCSI Storage

- IP-based storage networking standard
- Initiators (clients) send SCSI commands to targets (storage devices) over regular IP networks
- Alternative to NAS, SAN and DAS
- The OpenSolaris Common Multiprotocol iSCSI Target (COMSTAR) implements the iSCSI protocol

COMSTAR iSCSI for OHAC 2009.06

- Each node exports directly-attached disk as iSCSI target
- Nodes access both disks through iSCSI initiators
- Mirrored zpool built on top of the two disks
- HAStoragePlus imports zpool on node hosting the services that need it
- If one node goes down, local half of mirror still available and accessible from other node

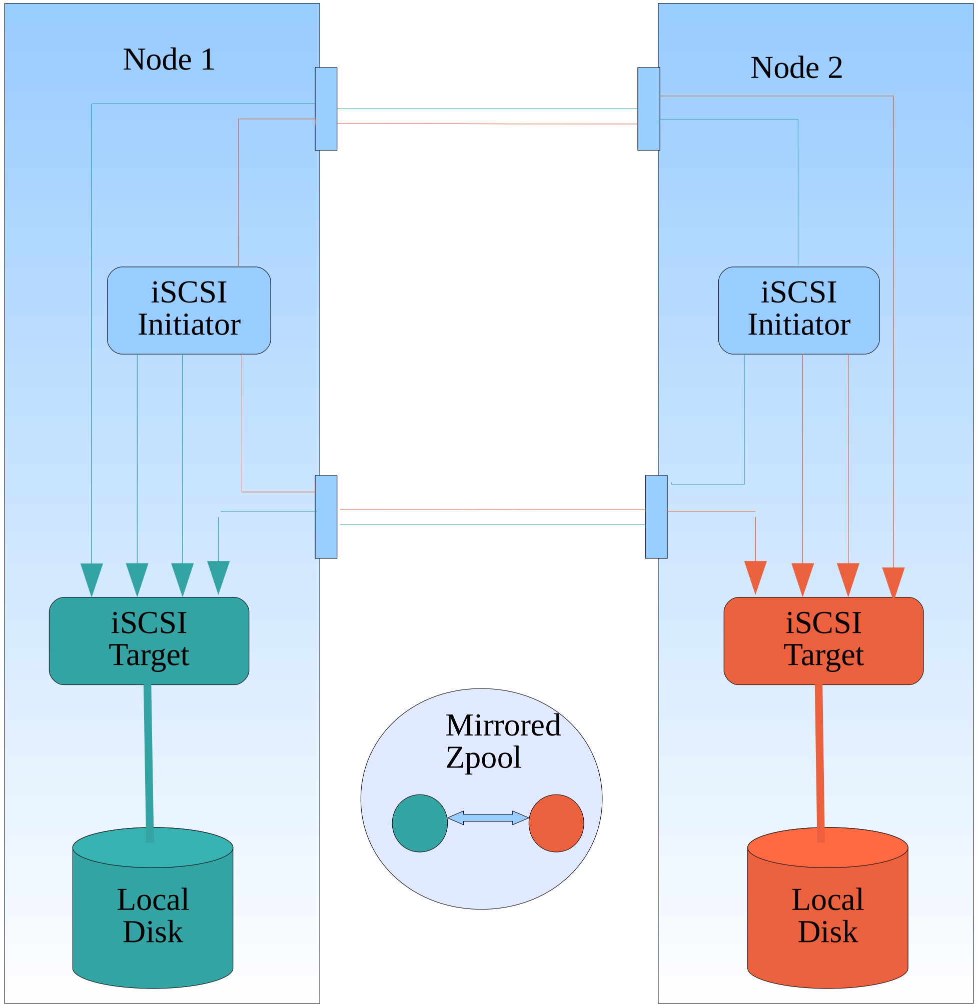
COMSTAR iSCSI Configuration



iSCSI with MPxIO and OHAC 2009.06

- OpenSolaris Storage Multipathing
  - Multiple redundant paths to storage
  - Operates above transport layer
- Configure MPxIO over redundant private interconnects to provide multiple paths to iSCSI targets (needs strong membership)
- Benefits of MPxIO with iSCSI in OHAC
  - Supports RDMA (infiniband)
  - Round-robin load balancing for increased throughput

iSCSI with MPxIO Configuration



Crossbow VNICs

- Virtual Network Interface Card (VNIC)
- Pseudo-network interface
- VNICs configured on a physical interface

```
# dladm create-vnic -l e1000g0 vnic1
```

Crossbow VNICs with OHAC 2009.06

- Cluster private interconnect can use VNICs as endpoints instead of physical adapters
- Work over dedicated physical adapter or public adapter
- Use IPsec to protect cluster-private traffic
  - Though DLPI heartbeats not protected
- Resource consolidation: Share physical adapters
- Easier setup: No dedicated private physical adapters and cabling required

OHAC 2009.06 Software Modularization

- ha-cluster-full group package
  - Contains core framework, wizards, agents, man pages, l10n, … (everything)
- ha-cluster-minimal group package
  - Only core framework
  - Add agents, wizards, l10n, man pages, telemetry, etc. individually as needed
- Install quorum server and agent builder without core framework

Why Minimal Installation is Useful

- Minimizing resources (you don't pay for what you don't need)
  - Disk space
  - Network download bandwidth
  - etc.
- Security minimization
- Minimizing administrative overhead
  - Both initial and ongoing

Installing Open HA Cluster 2009.06

- Accept terms of use at pkg.sun.com and download key and certificate to /var/pkg/ssl
- Set ha-cluster publisher (on all nodes):

```
# pkg set-publisher  \
  -k /var/pkg/ssl/Open_HA_Cluster_2009.06.key.pem \
  -c /var/pkg/ssl/Open_HA_Cluster_2009.06.certificate.pem \
  -O https://pkg.sun.com/opensolaris/ha-cluster/ ha-cluster
```

Installing OHAC 2009.06 (cont)

- Install the cluster software (on all nodes):

```
# pkg install ha-cluster-full
```

- Configure the cluster (on one node):

```
# /usr/cluster/bin/scinstall
```

Live Demo

- Toshiba M10
  - 4 GB main memory
  - 160 GB hard disk
  - OpenSolaris 2009.06
  - Open HA Cluster 2009.06
  - VirtualBox 3.0.8

References (1)

- Open HA Cluster 2009.06 Documentation
  - http://www.opensolaris.com/learn/features/availability/
  - http://docs.sun.com/app/docs/prod/open.ha.cluster~2509.1#hic
- Solaris Cluster Blog
  - http://blogs.sun.com/SC
- White Paper: Running Open HA Cluster on OpenSolaris with VirtualBox
  - http://opensolaris.org/os/project/colorado/files/Whitepaper-OpenHAClusterOnOpenSolaris-external.pdf

References (2)

- HA Clusters Community Group
  - http://opensolaris.org/os/community/ha-clusters/
- Project Colorado
  - http://opensolaris.org/os/project/colorado/
- Project Image Packaging System (IPS)
  - http://opensolaris.org/os/project/pkg/
- Project Crossbow (VNICs)
  - http://opensolaris.org/os/project/crossbow/
- Project COMSTAR (iSCSI)
  - http://opensolaris.org/os/project/comstar/

opensolaris™

Open HA Cluster

Thank You! Questions?

thorsten.frueauf@sun.com
http://blogs.sun.com/tf

Thorsten Früauf
Availability Engineering
Sun Microsystems GmbH