

# open



USE



IMPROVE



EVANGELIZE

## Open HA Cluster

Piotr Jasiukajtis  
2008

開  
放  
的  
열린  
مفتوح  
libre  
मुक्त  
ಮುಕ್ತ  
livre  
libero  
ముక్త  
开放的  
açık  
open  
nyílt  
⋮⋮⋮  
πππ  
オープン  
livre  
ανοικτό  
offen  
otevřený  
öppen  
открытый  
வெளிப்படை



# Na podstawie dokumentacji:

- <http://www.opensolaris.org/os/community/ha-clusters/ohac/Documentation>
- <http://docs.sun.com/app/docs/prod/sun.cluster32#hic>
- <http://www.opensolaris.org/os/project/colorado/>



## Solaris Cluster:

- Komercyjny produkt
- Wysoka dostępność działających usług
- Mniejsze ryzyko strat (SLA)
- Disaster Recovery (GEO)



## Solaris Cluster, SCX, Colorado, OHAC?

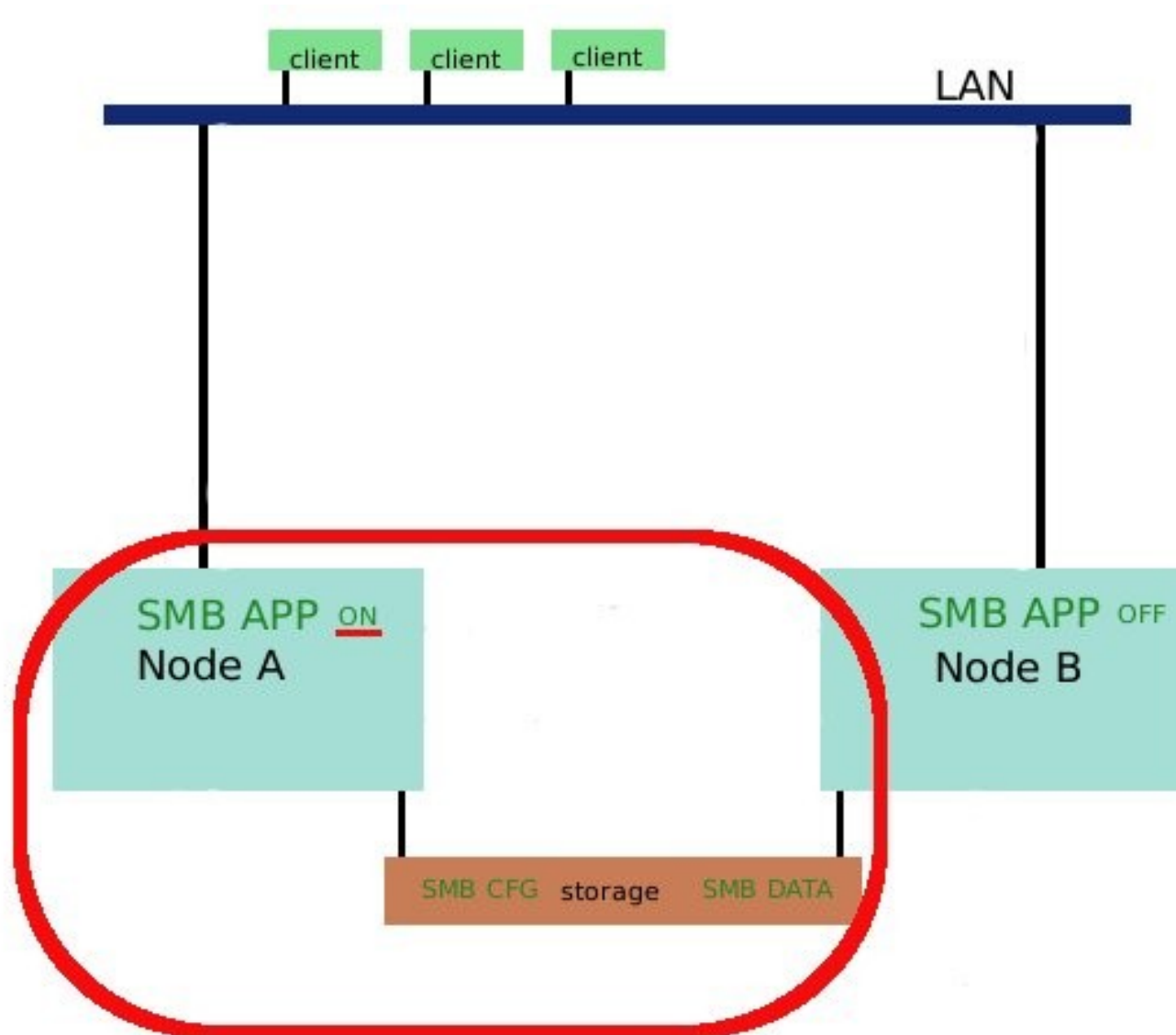
- Solaris Cluster → Solaris
- OHAC → kod
- Solaris Cluster Express → SXCE
- Colorado → OpenSolaris 2008.x



## Zasada działania Solaris Cluster:

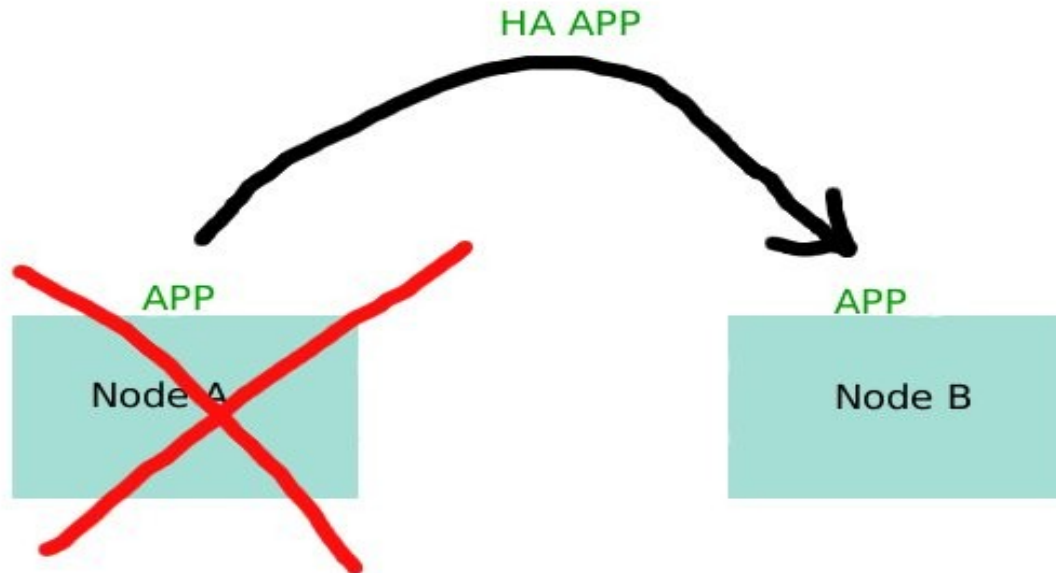
- Rozszerzenie systemu operacyjnego
- Dodatkowe warstwy usług/aplikacji/jądra zapewniające łączenie maszyn w logiczną całość
- Zmieniamy podejście – nie zarządzamy już pojedynczym systemem
- Klastrujemy aplikacje/usługi, a nie system

# Jak działa aplikacja HA w Sun Cluster:



# W jaki sposób następuje przełączanie?

- Aplikacja działa na maszynie A
- Gdy aplikacja nie jest dostępna na maszynie A, jest ona uruchamiana na maszynie B





## Co jest monitorowane?

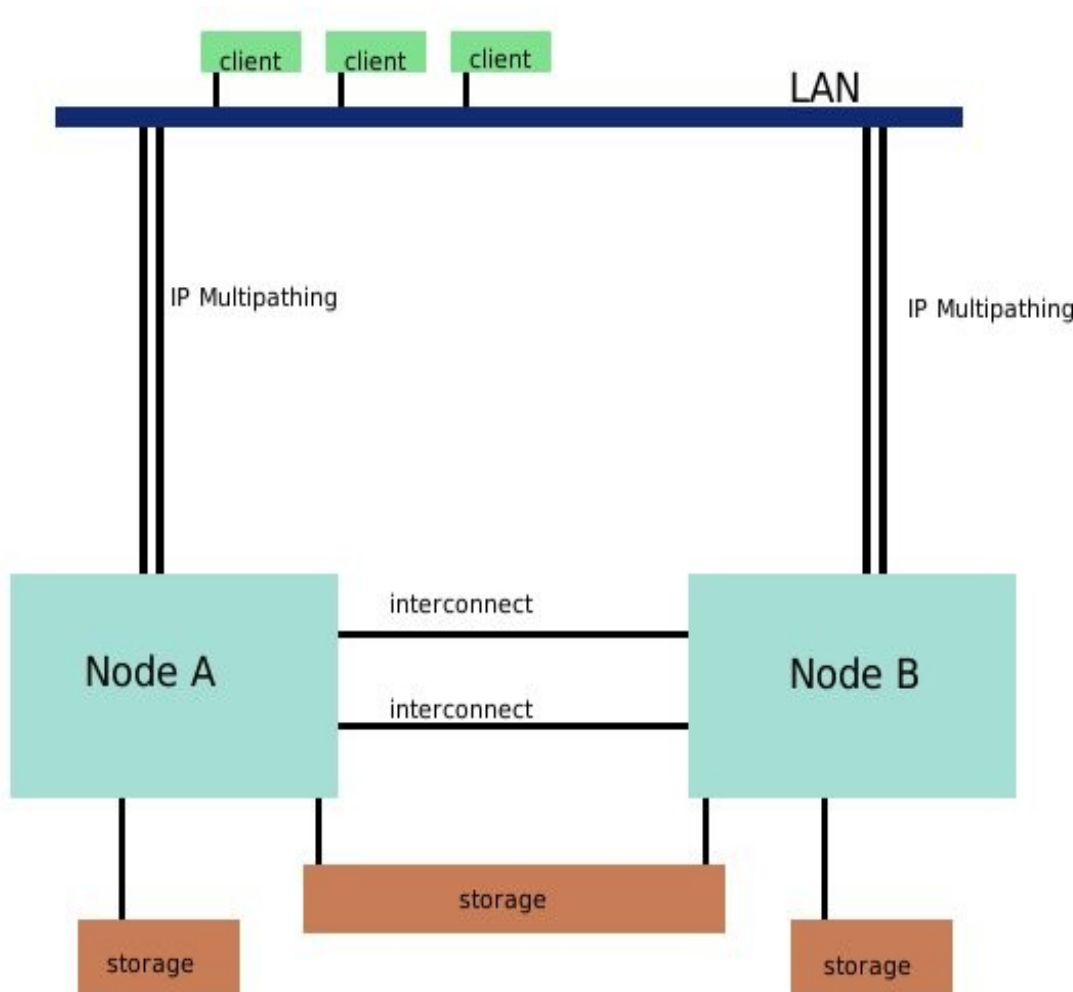
- Problemy sprzętowe
  - System operacyjny
  - Sieć
  - Aplikacje
- 
- Niektóre problemy błyskawicznie wykryte – dzięki integracji z jądrem systemowym (problemy ze sprzętem/siecią).



## Ile węzłów?

- 1 węzłowy klaster :)
- 2 węzłowy klaster
- 3 i więcej węzłów
- Zony jako wirtualne węzły

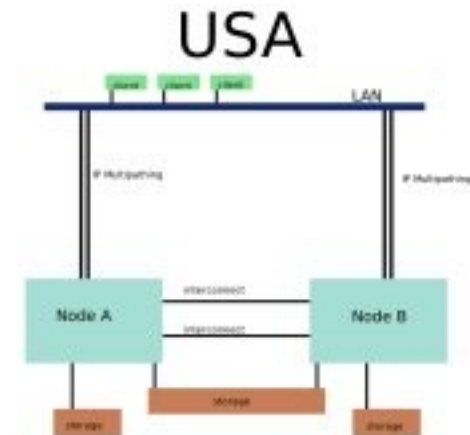
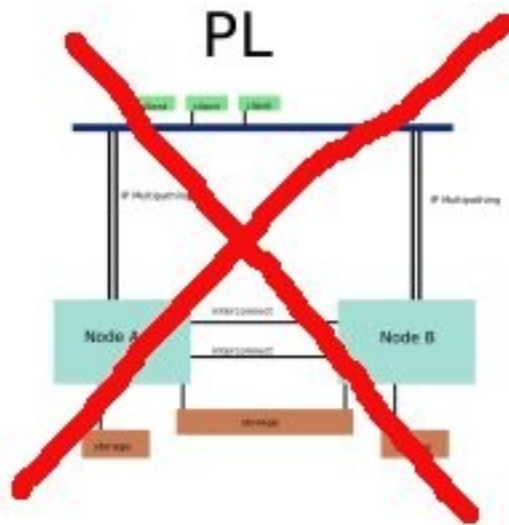
# Prosty przykład 2-węzłowego klastra:



- Nodes - węzły klastra
- Public network
- Cluster Interconnect – prywatna podsieć
- Quorum Devices
- Devices
- Shared storage
- Local storage

# Sun Cluster Geographic Edition:

GEO





## Project Colorado:

- Binarna dystrybucja OHAC dla Indiany
- Dostępna poprzez IPS oraz źródła
- Podstawowy framework
- Uproszczona konfiguracja – klaster w 15min
- Duże zmiany w kodzie

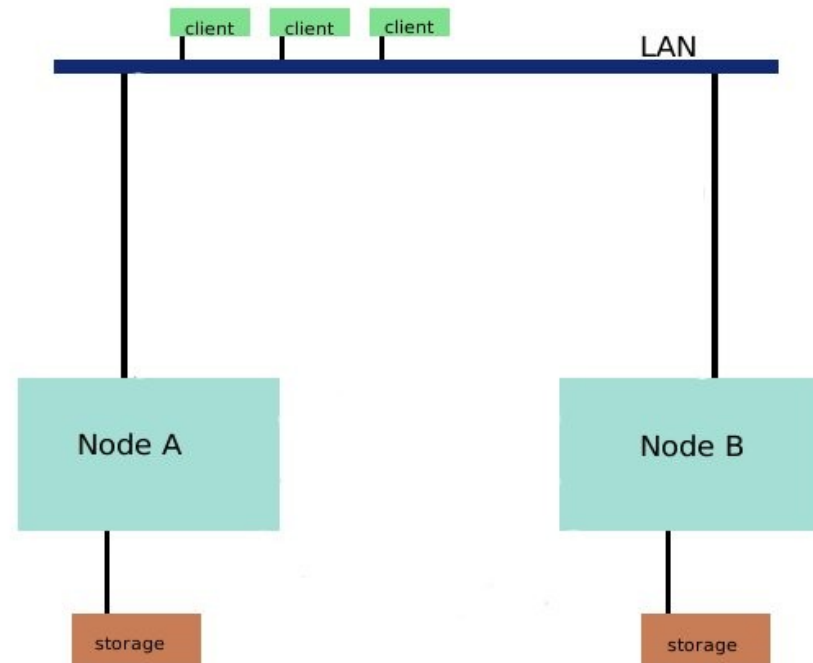


## Nowości w Colorado:

- iSCSI jako shared storage
- Strong Membership / Weak Membership
- Redukcja dedykowanych kart sieciowych
- Project Clearview (IPMP)
- Project Crossbow (VNIC)
- 2-węzłowy klaster bez Quorum Device

# Minimalna wersja klastra w Colorado:

- Weak Membership
- Non Shared Storage – AVS
- Brak Quorum Device
- Crossbow (VNIC)





## Założenia projektu Colorado:

- Sun Cluster na platformie OpenSolaris
- HA dostępne dla wszystkich
- Wykorzystanie ogólnodostępnego sprzętu
- Szybka konfiguracja – np. SAMP, SAMBA
  
- Mniej paczek instalacyjnych
- Redukcja użycia pamięci i CPU



## SC i mniejsza podatność na awarie:

- Monitoring aplikacji + SMF
- Solaris Volume Manager
- ZFS
- IPMP – interconnect
- Link Aggregation – public network
- Quorum Device / Quorum Server
- Klaster Geograficzny + replikacja AVS
- Integracja z jądrem systemu



## Integracja w jądrze systemu:

- Szybka i skuteczna reakcja systemu na niektóre problemy

```
# /usr/sbin/modinfo | grep cl_
 2 ffffffff83f000  6e0 - 1 cl_bootstrap (DCS bootstrap module)
 3 ffffffff83f5f0  4e78 - 1 cl_runtime (cluster runtime support)
172 ffffffff83d9000 d8410 - 1 cl_comm (Sun Cluster communication
  suppo)
173 ffffffff846d000  3bb8 - 1 cl_load (Sun Cluster cladadmin support)
174 ffffffff8470000 14fdd8 - 1 cl_orb (CORBA runtime support)
175 ffffffff8552000 1a7870 - 1 cl_haci (Sun Cluster Cluster Membership )
176 ffffffff867f000 1a3f0 - 1 cl_quorum (Sun Cluster Quorum support)
284 ffffffff8c2b000 168288 - 1 cl_dcs (Major-Minor server code)
```



# Elementy instalacyjne SC / OHAC:

- **Core**

Rdzeń klastra. Silnie integruje się z jądrem systemu.

- **Agents**

Wsparcie dla dodatkowych aplikacji np. Samba, Apache

- **Geo**

Klaster geograficzny.

Zdalna replikacja danych (np AVS).



## Obsługiwane typy aplikacji:

### Failover services

- Aplikacje nie wiedzą nic o klastrze.
- Aplikacja działa tylko na jednym węźle.

### Scalable services

- Aplikacja działa na wielu węzłach.

### Parallel Applications

- Oracle Real Application Cluster (RAC)



# Obsługiwane aplikacje:

Oracle Database, Sybase ASE, Oracle's Siebel CRM (server and gateway)

BEA WebLogic Server, DNS, NFS Server, Kerberos, Apache Web Server

Sun Java ES Web Server,

Sun Java ES Application Server

Sun Java ES Message Queue Broker

Agfa IMPAX, Solaris DHCP, Oracle E-Business Suite

Oracle Application Server, Apache Tomcat, MySQL, SWIFTAlliance Access

SWIFTAlliance Gateway, IBM WebSphere MQ, Sun N1 Grid Engine

Sun N1 Service Provisioning system, PostgreSQL,

SAP Web Application Server, SAP liveCache,

MaxDB (previously called SAP DB)

Samba



# Moja aplikacja nie jest obsługiwana?

- Generic Data Service template :)

<http://opensolaris.org/os/community/ha-clusters/ohac/GDS-template/>



## Generic Data Service (GDS):

- Framework dla własnych aplikacji
- resource type SUNW.gds
- Uruchamia oraz zamyka aplikacje
- Monitoring usług



# Aplikacje to nie wszystko!

- HA Solaris Containers
- Zone Cluster (SCX 9/08)
- HA xVM domU Live Migration
- AVS



## Co zyskujemy dzięki Open HA Cluster?

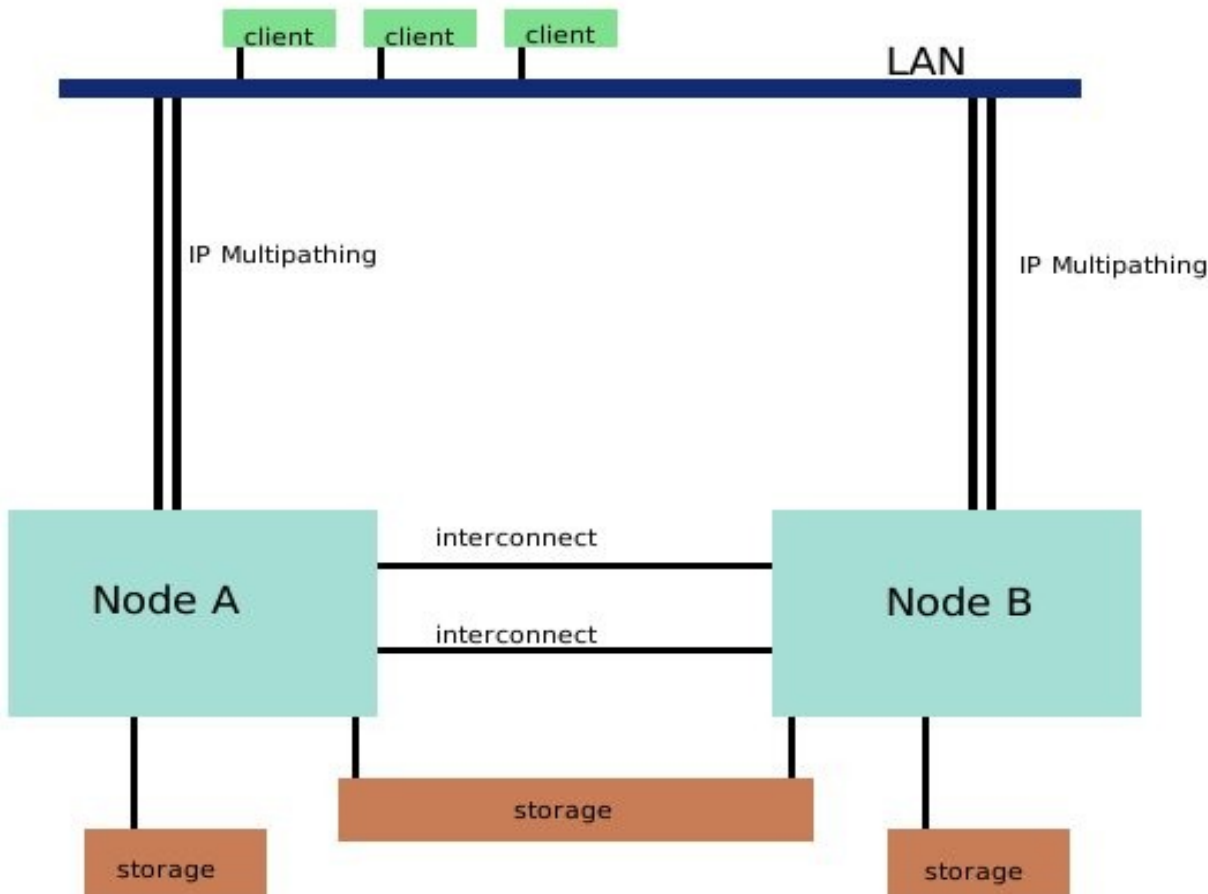
- Otwarty kod stabilnej i przetestowanej implementacji klastra wysokiej dostępności
- Gotowy, spójny produkt, bez błądzenia w półśrodkach i niepełnych rozwiązaniach
- Wysoka skuteczność dzięki integracji z jądrem systemowym
- Oparty na stabilnych technologiach Solarisa
- HA Community
- Dużo zabawy :)



## Nowe projekty:

- Project Colorado:
  - OHAC w OpenSolaris 2008.x
- HA xVM:
  - Wysokodostępne domeny XEN
- MRSL.NONsharedDevice:
  - replikacja bez konieczności dzielonego storage

# Opis budowy 2-węzłowego klastra:



- Nodes - węzły klastra.
- Cluster Interconnect – prywatna podsieć.
- Cluster Membership Monitor (CMM).
- Cluster Configuration Repository (CCR).
- Fault Monitors
- Quorum Devices
- Devices
- Data Services (RGM)



## Elementy klastra:

- Nodes - węzły klastra.
- Cluster Interconnect – prywatna podsieć.
- Cluster Membership Monitor (CMM).
- Cluster Configuration Repository (CCR).
- Fault Monitors
- Quorum Devices
- Devices
- Data Services (RGM)



## Cluster Interconnect:

- Prywatna podsieć klastra.
- Oddzielny VLAN lub bezpośrednie połączenie w przypadku 2-węzłowego klastra.
- Zalecane minimum 2 połączenia między klastrami.



## Cluster Interconnect w Colorado:

- VNIC poprzez sieć publiczną (Crossbow)
- Weak Membership



## Cluster Membership Monitor (CMM):

- Zarządza członkostwem węzłów w klastrze.
- Zmienia konfiguracje klastra w przypadku awarii.
- Otrzymuje informacje od węzłów przez prywatną podsieć.
- Gwarantuje, że tylko jeden węzeł korzysta z dzielonych danych.



## CMM - przykład:

NOTICE: CMM: Node node2 (nodeid = 2) is dead.

NOTICE: CMM: Node node2 (nodeid = 2) is down.

NOTICE: CMM: Cluster members: node1.

NOTICE: CMM: node reconfiguration #134 completed.

CMM: Node node2 (nodeid = 2) is down.

## Quorum:

- Stan niezbędny w celu utworzenia klastra.
- Głosują węzły oraz Quorum Device - dzielony storage (SCSI).
- Quorum Device można zastąpić poprzez Quorum Server.
- Każdy węzeł domyślnie ma głos 1 gdy dołącza do klastra.
- Quorum Device otrzymuje głos  $N-1$ , gdzie  $N$  jest liczbę połączeń do QD.  
(QD w 2-węzłowym klastrze ma  $\text{vote}=1$ )



## Quorum - przykład:

NOTICE: CMM: Node node1: attempting to join cluster.

NOTICE: CMM: Node node2 (nodeid: 2, incarnation #: 1218801478) has become reachable.

NOTICE: CMM: Cluster has reached quorum.

NOTICE: CMM: Node node1 (nodeid = 1) is up; new incarnation number = 1218804108.

NOTICE: CMM: Node node2 (nodeid = 2) is up; new incarnation number = 1218801478.

NOTICE: CMM: Cluster members: node1 node2.

NOTICE: CMM: node reconfiguration #138 completed.

NOTICE: CMM: Node node1: joined cluster.



## Quorum i spójność klastra:

### Split brain:

- Gdy prywatne połączenia klastra są przerwane.
- Klaster dzieli się na subklastry.
- Subklaster nie wie nic o pozostałych subklastrach.
- Może powodować błędy w dzielonym storage, duplikację adresów IP itp.

## Quorum i spójność klastra: cd.

Amnezja (Amnesia):

1. Gdy węzeł A się wyłączy (awaria), konfiguracja klastra (CCR) jest zmieniona i zapisana tylko na węźle B.
2. Gdy węzeł B się wyłączy i jednocześnie uruchomi się ponownie węzeł A, węzeł A będzie mieć starą (niepoprawną) konfigurację klastra (CCR).



## Quorum i Colorado:

- Możliwość wymuszenia włączenia Resource Group przez administratora
- Wymuszenie działania klastra bez quorum
- Awaryjne uruchomienie 'klastra'



## Cluster Configuration Repository (CCR):

- Opiera się na CMM, aby zagwarantować, że klaster jest uruchomiony gdy jest osiągnięte quorum.
- CCR odpowiada za poprawną i spójną konfigurację klastra.
- Aktualizuje konfigurację w razie awarii.



## Devices:

### Global Devices – Global File System

- Dyski, napędy optyczne i taśmy dostępne z każdego węzła.
- Jedynie dyski zapewniają HA.

### Local Devices

- Szybsze – nie ma potrzeby przechowywania informacji o replikacji na wszystkich węzłach.

### Device Groups

- Zarządzane przez volume manager.



## Data Services:

- Zarządzane przez Resource Group Manager (RGM).
- Data Service zapewnia wsparcie dla aplikacji, aby działała bez modyfikacji w Sun Cluster.
- Uruchamia i zamyka aplikacje.
- Monitoruje stan usługi i zapewnia failover.



## Data Services: cd.

### Resource Types

- Definiuje typ zasobu w klastrze.

### Resources

- Podstawowy element konfiguracji klastrowanych usług.

### Resource Groups

- Spójne zarządzanie grupami zasobów.



## Monitorowane elementy klastra:

- Aplikacje – data services
- Dyski – disk-path monitorinng (DPM)
- Sieć – IP multipath



## Reprezentacja aplikacji HA w klastrze:

- Aplikacja
- Resource
- Resource Group



# Resource Types - przykład:

```
# clresourcetype list  
SUNW.LogicalHostname:2  
SUNW.SharedAddress:2  
SUNW.gds:6  
SUNW.HAStoragePlus:5  
SUNW.dns:3.2
```



# Resource Types - przykład:

```
# clresource show SUNW.LogicalHostname
```

```
=== Registered Resource Types ===
```

```
Resource Type:                SUNW.LogicalHostname:2
RT_description:                Logical Hostname Resource Type
RT_version:                    2
API_version:                   2
RT_basedir:                    /usr/cluster/lib/rgm/rt/hafoip
Single_instance:              False
Proxy:                         False
Init_nodes:                    All potential masters
Installed_nodes:               <All>
Failover:                      True
Pkglist:                       SUNWscu
RT_system:                     True
Global_zone:                   True
```



# Resources - przykład:

```
# clresource show mail1-lh-r
```

```
=== Resources ===
```

```
Resource: mail1-lh-r
Type: SUNW.LogicalHostname:2
Type_version: 2
Group: mail1-rg
R_description:
Resource_project_name: default
Enabled{node1}: True
Enabled{node2}: True
Monitored{node1}: True
Monitored{node2}: True
```



# Resource Groups – przykład:

```
# clrg list
smb1-rg
mail1-rg
# clrg status smb1-rg
```

=== Cluster Resource Groups ===

Group Name	Node Name	Suspended	Status
-----	-----	-----	-----
smb1-rg	node1:smb1-n1	No	Online
	node2:smb1-n2	No	Offline



# Resource Groups - przykład:

```
# clrg status mail1-rg
```

```
=== Cluster Resource Groups ===
```

Group Name	Node Name	Suspended	Status
-----	-----	-----	-----
mail1-rg	node1	No	Online
	node2	No	Offline



# Podstawowe komendy administracyjne:

- clquorum
- clresourcetype
- clresourcegroup
- clresource
- clreslogicalhostname
- cldevicegroup
- cldevice
- clnode
- clinterconnect
- cluster



## Kompilacja OHAC:

- Wymagany Solaris Express jako platforma.
- Sun Studio 11, trwają prace nad SS12.
- Źródła ONNV (snv97 dla edycji z sierpnia).
- Java Dynamic Management Kit
- ON Specific Build Tools
- OHAC Specific Build Tools
- OHAC Core Closed Binaries
- OHAC Core External Packages



## Kompilacja OHAC: cd.

- /opt/scbld/bin/nbuild
  - debug build
  - non-debug build
  - dvd image
  - lint, tests
- Na Core 2 Duo 2.4Ghz kompilacja OHAC Core (debug) z paczkami oraz instalatorem trwa < 1h (bez testów kodu).

<http://opensolaris.org/os/community/ha-clusters/ohac/Documentation/Core>



## Kompilacja Colorado:

- Potrzebny OpenSolaris 2008.x
- Sun Studio Express
- ONNV b97
- Pierwsza wersja SC na Indiana i SS12

[http://www.genunix.org/wiki/index.php/Compiling\\_OHAC](http://www.genunix.org/wiki/index.php/Compiling_OHAC)



## Więcej informacji:

OpenSolaris Community: HA Clusters

<http://opensolaris.org/os/community/ha-clusters/>

Open High Availability Cluster

<http://opensolaris.org/os/community/ha-clusters/ohac/>

Solaris Cluster docs:

<http://docs.sun.com/app/docs/prod/sun.cluster32#hic>

Mailing list:

<http://opensolaris.org/os/community/ha-clusters/discussions/>

IRC: #OHAC w sieci freenode

# open



USE



IMPROVE



EVANGELIZE

Dziękuję!

Piotr Jasiukajtis  
estseg@gmail.com  
estibi@opensolaris.com.pl  
estseg.blogspot.com

“open” artwork and icons by chandan:  
<http://blogs.sun.com/chandan>

開  
放  
的  
열린  
مفتوح  
libre  
मुक्त  
ಮುಕ್ತ  
livre  
libero  
ముక్త  
开放的  
açık  
open  
nyílt  
•••••  
πικρ  
オープン  
livre  
ανοικτό  
offen  
otevřený  
öppen  
ОТКРЫТЫЙ  
வெளிப்படை