EĞİTİM AMAÇLI

Hazırlayan: Asiye YİĞİT

Cluster Tanımı

Cluster, iki veya daha fazla farklı sunucunun uyumlu bir birim olarak gruplanması şeklinde tanımlanabilir

Cluster Karakteristikleri

- Her biri shared olmayan (kendine ait) işletim sisteminden açılan farklı node lardır
- Aynı cluster node ları arasında özel haberleşmeyi sağlayan ve bu amaç için ayrılmış HW interconnect ler vardır
- Multi-ported storage olmalıdır. En azından cluster içerisindeki iki node her bir fiziksel storage a ulaşır
- Cluster yazılımı framework olmalıdır. Cluster a özel bilgininin node lara ulaşımını sağlar. Bu sayede HW in sağlamlığı, her bir node un sağlamlığı cluster ı oluşturan node lar tarafından bilinir
- Genel amacı, üzerinde çalışan uygulamalar için HA platform sağlamasıdır.
- Pek çok Cluster-aware ve cluster-unaware uygulamaları destekler

HA

- Market’te uygulamalar için HA sağlayan tek yol cluster olarak tanımlanmış durumdadır. HA, hiç down time olmaması olarak değil, down time süresinin minimize edilmesi olarak tanımlanır.
- HA standartlar, yılda sadece yaklaşık 5 dakikalık bir kesintiyi gerektirir. Bu da 5 tane 9 olarak ifade edilir, 99.999. Açıktır ki temiz bir sunucu reboot işlemi bile, 5 dakikadan fazladır.

Cluster Nasıl HA Sağlıyor

- Cluster da oluşan donanım veya yazılımsal hata durumunda uygulama servisleri ve data otomatik olarak, insan müdahelesine gerek kalmaksızın, geri alınır. Bu, cluster içerisinde fazla node olması ve fazla storage path olması sayesinde olur.
- HA özelliğini planlı kesintilerde de kullanabilirsin. Tek node üzerinde planlı bir bakım yapma ihtiyacı olabilir, bu durumda bu node üzerinde çalışan servisleri diğer node üzerine alabilirsin.
- Fault-tolerant sistemler, HA e alternatif değildir. Fault-tolerant sistemler, HW arızasını gizlerken, SW hatalarını geri almak için oluşturulmamışlardır. Mesela, OS kernel panic veya uygulama hatasını geri almak amaçlı dizayn edilmemişlerdir.

Scalability Platform

- Cluster,scalability servisleri için de HW ve SW ortamı sağlar. Uygulamaya ait bir den fazla instance farklı node larda çalışır. Bu instance lar genellikle aynı data ya ulaşır.
- Scalable uygulamalarda, hatalı uygulamayı başka bir node a yerleştirmeye gerek olmayabilir çünkü diğer instance lar zaten diğer node larda çalışır durumdadır.

Sun Cluster 3.1 HW and SW Environment

- 2 den 8 node a kadar destekler
- Storage, en az iki farklı node a path ler aracılığıyla bağlı olmak zorunda olsa bile, cluster üzerinde olan tüm device lar lojik olarak tüm node lar tarafından görülür. Yani fiziksel olarak bağlı olmasa bile, node device ları görebilir(global device implementation).
- Storage topolojisinden bağımsız olarak aynı dosyalar, cluster da bulunan tüm node lar tarafından erişilebilir durumdadır (global file system implementation)
- Cluster framework, kernel in içerisine gömülü durumdadır. Node ların monitör edilmesi, transport un monitör edilmesi, global device ve global file system implementasyonu gibi pek çok özellik kernel içerisine işlenmiş durumdadır. Bu sayede güvenilirlik ve performans sağlanır.
- Pek çok uygulama sun cluster tarafından desteklenir.
- Global interface özelliği ile scalable uygulamaları destekler. Mesela apache web server, sun one web server gibi scalable servislerde, cluster dışında olan client lar servisin multiple-node instance larını, tek bir IP den tek bir servis gibi görürler.

Sun Cluster 3.1 e Giriş

- Cluster node ları (Solaris 8 Up 7 ve sonrası) Heterojen sistemler destekleniyor
- Her node üzerinde farklı boot diskleri (her node üzerinde mirror lı boot diskleri)
- Sistem ve subnet başına bir veya birden fazla public network interface leri (tercien en az iki public network interface)
- Private cluster transport interface
- Dual-hosted, mirrored disk storage

Cluster Transport Interface

- Cluster içinde olan tüm node lar private cluster tarnsport ile birbirine bağlıdır. Cluster-wide monitoring&recovery, global data access, cluster-aware uygulamaya özel transport için kullanılırlar
- Cluster minimum iki ayrı private network e sahip olmalıdır. İki den fazla olabilir ve daha sonra da eklenebilir. Global data access trafiği private hat lar üzerinden stripe yapıldığından performans kazancı sağanlıyor.
- İki node lu cluster da genellikle çapraz kablolar kullanılır. İki node lu cluster için switch ler opsiyoneldir. İki node dan fazla ise switch kullanımı şarttır.

Public Network İnterface

- Her node IPMP kontrolünde public network interface lere sahip olmalıdır. Şiddetle önerilen, kullanılan her subnet için her node un IPMP içerisinde en az iki interface e sahip olmasıdır.
- Birden fazla subnet e bağlantı sağlayabilirsin. Sun cluster hw in router olarak davranılmasına izin verilmez.

Cluster Disk Storage

- Storage kontroller üzerinden SVM veya VxVM ile mirror lanmalıdır.

Cluster Boot Disk

Her node için boot diski node üzerinde , lokal olmalıdır. Multi-ported storage üzerinde olmamalıdır.

Node başına iki tane lokal diske sahip olunmalıdır. Önerilen bu iki diskin SVM veya VxVM ile mirror lanmasıdır.

Sun Cluster Hardware

- Redundant Server nodes are REQUIRED
- Redundant transport is REQUIRED
- Redundant storage arrays are REQUIRED
- HW RAID storage arrays are OPTIONAL
- Software mirroring across data controllers is REQUIRED
- Redundant public network interfaces per subnet are RECOMMENDED
- Redundant boot disks are RECOMMENDED

Uygulama Türleri

- Cluster-unaware Uygulamalar
  - Failover uygulamalar
  - Scalable uygulamalar

Bu uygulamaların ortak bileşenleri:

- RGM (cluster resource group manager) uygulamaya ilişkin tüm başlantma ve sonlandırma işlemlerini yapar.
- Uygulama için geliştirilen data service agent , uygulamanın cluster içerisinde doğru çalışmasını sağlar. Bu, uygulamanın cluster içerisinde uygun şekilde kapatılıp açılmasına ilişkin metodları, uygulamaya ilişkin hata izleme metodlarını da kapsar.

Failover Uygulamalar (Cluster-Unaware uygulamalar)

- Failover uygulamalar bir t anında sadece bir node üzerinde çalışır. Cluster, uygulamanın aynı node üzerinde veya farklı node üzerinde yeniden otomatik başlatılmasını sağlar.
- Fialover servisleri genel olarak bir uygulama IP si ile eşleştirilirler. Bu IP, uygulama ile node dan node geçiş yapar. Cluster dışında olan istemciler, logical host name i görürler. Servisin hangi node da çalıştığından bağımsız olarak, logical hostname üzerinden servis alırlar.
- Aynı resource group da olan çoklu failover uygulamaları aynı IP adresi paylaşırlar. Bir resource group da olan çoklu uygulamalar aynı anda diğer node üzerine geçerler.

Scalable Uygulamalar (Cluster-Unaware Uygulamalar)

- Uygulamanın birden fazla instance ı aynı cluster içerisinde çalışır. Dışarıdan tek bir servis gibi görünür. Tek bir servis gibi görünmesini mümkün kılan, global interface özelliğidir. Global interface özelliği, tek bir IP ve load-balancing özelliği sağlar. Kilit mekanizması türünde yazılmayan uygulamalar failover olarak çalışabilir fakat scalable uygulamalar olarak çalışmaz.

Cluster-Aware Uygulamalar

- Cluster bilgisi uygulamaya gömülü durumdadır.
- Uygulama instance ları farklı node lar üzerinde çalışır ve birbirlerinden haberdardır ve private transport üzerinden haberleşirler.
- RGM ile start/stop edilmeye gerek yoktur çünkü bu uygulamalar cluster dan haberdardırlar ve kendi başlarına script lerle veya el ile başlatılabilirler.

Sun Cluster SW HA Framework

- Sun Cluster SW framework bir dizi daemon ve kernel modüllerden oluşur. Framework ün büyük bir kısmı, kernel a işlenmiş olduğundan daha hızlı ve güvenilir çözüm sunar.
- Cluster membership monitor (CMM) her node üzerinde kernel üzerinde konumlanır. Cluster üzerinde satus değişikliklerini algılar. Mesela node lar arasındaki iletişim kaybını algılar. CMM, transport kernel modüle güvenir. Transport kernel modül, transport medium boyunca hareket eden heartbeat leri üretir. Tanımlı bir time-period içerisinde heartbeat sinyali alınmaz ise cluster node un problemli olduğu düşünülür.
- Public network interface ve cluster transport interface potansiyel hatalar için monitör edilir.
- Genel cluster konfigürasyon bilgisi, global konfigürayon dosyalarında depolanır. Global konfigürasyon dosyaları, cluster configuration repository (CCR) olarak ifade edilir.

Cluster Configuration Repository

- Cluster ve node isimleri
- Cluster transport konfigürasyonu
- Veritas disk group isimleri, svm diskset isimleri
- Her disk grubun ait olduğu node ların listesi
- Data servis operasyonel parametre değerleri
- Disk ID (DID) device konfigürasyonu
- Aktif cluster statusu

Global Naming, Device, and System Services

- Disk ID Devices (DID)
  - Cluster da olan her disk drive, cd-rom drive, tape drive için tek bir device ismi sağlar. Multi-ported diskler, farklı node lar üzerinde farklı logical isimlere sahip olabilir. Bu disklere cluster üzerinde tek bir DID instance numarası verilir.
- Global Devices
  - Storage ın fiziksel olarak nereye bağlı olduğuna bakılmaksızın tüm node lardan storage a aynı anda erişim sağlar. Erişim, DID device lar, CD-ROM ve tape cihazları, Veritas volume leri, SVM metadevice ları için geçerlidir. Tüm node lar aynı device adresini kullanır fakat sadece primary node disk device ile storage medium u kullanarak konuşur. Diğer tüm node lar device a primary node ile cluster tarnsport üzerinden haberleşerek erişir.
- Device Files for Global Devices
  - Her node üzerinde özel bir dosya sistemi vardır. Global device lar için device dosyalarını depolar. Bu dosya sistemi için mount noktası, /global/.devices/node@nodeID şeklindedir. nodeID, node u ifade eden bir tamsayıdır. Bu dosya sistemi, genellikle boot diskinde bu dosya sistemine adanmış bir bölümlemeye ihtiyaç duyar.
  - /dev/did/dsk, /dev/did/rdsk, ve /dev/did/rmt global erişim path leri değildir. Sun Cluster SW, /dev/global altında alternate path isimleri yaratır. Bunlar, global device uzayına linklidir.

Global Dosya Sistemleri

- Global dosya sistemi özelliği, lokasyonlarından bağımsız olarak dosya sistemlerini tüm node lara kullanılabilir hale getirir. Ufs, hsfs, vxfs desteklenen dosya sistemleridir. Global mount opsiyonuyla global hale getirilirler. Komut satırından, #mount –o global,logging /dev/vx/dsk/nfs-dg/vol-01 /global/nfs; /etc/vfstab dosyasında, /dev/vx/dsk/nfs-dg/vol-01 /dev/vx/rdsk/nfs-dg/vol-01 /global/nfs ufs 2 yes global,logging
- Global dosya sistemi, global device özelliği ile aynı ilkelerle çalışır. Bir t anında sadece tek node primary olur. Ve bu node bu dosya sistemiyle konuşur. Diğer node lar primary e cluster transportu kullanarak erişirler.

Local Failover Dosya Sistemleri

- Dosya sistemi tüm node lar tarafından aynı anda erişilebilir değildir. Bir t anında sadece tek node üzerinden erişilebilirlerdir. Bu node da servislerin çalıştığı ve storage a fiziksel bağlantının olduğu node dur.
- Local failover dosya sistemi, failover servisler için uygundur. Scalable servisler için uygun değildir. Bu servisler global dosya sistemi erişimine ihtiyaç duyarlar.
- Lokal failover dosya sistemi erişimi, uygun kullanıldığında, global dosya sistemi erişimine göre faydalara sahiptir. Multiple node lar üzerinde replicated state bilgisi eş zamanlı olarak tutulduğu için bir gidere sahiptir.

Quorum Devices

- Boot Device Kısıtlamaları
  - Shared storage device, boot device olarak kullanılamaz. Eğer storage device, birden fazla host a bağlıysa shared device dır.
  - DMP desteklenmiyor. Yine de DMP için gereken driver lar sistem üzerinde olmalıdır
  - /globaldevices için en az 512MB lık bir alan boot diskinde olmalıdır

Cluster Topolojileri

- Clustered pairs topology
  - İki veya daha fazla çiftli node lardan oluşur. Her çift, storage a bağlıdır.
  - Failover data servsileri için uygundur
- Pair+N topology
  - Shared storage a bağlı bir çift ve shared storage a bağlı olmayan node lar içerir. Shared storage a fiziksel olarak bağlı olmayan node lar cluster interconnect leri kullanarak storage a erişirler.
  - Scalable data servisleri için uygundur
- N+1 Topology
  - Tek bir sistem, diğer her sistem için storage backup gibi davranır. Storage device lara giden secondary path ler bu sisteme bağlıdır.
  - Failover data servisleri için uygundur
- N*N Topology
  - İkiden fazla node, fiziksel olarak aynı storage a bağlanabilir. İkiden fazla node üzerinde oracle parallel server çalışacak ise bu konfigürasyon gereklidir.
- Non-storage topology
  - İkiden fazla node içeren cluster ların shared storage e ihtiyaçları yoktur. Dikkat, iki node lu cluster storage a ihtiyaç duyar çünkü quorum device a ihtiyaç vardır.

Quorum Votes ve Quorum Devices

- Cluster, voting sistemine göre çalışır
  - Her node un bir oyu vardır
  - Belirli diskler quorum disk olarak tanımlanır ve bu device lara vote atanır.
  - Mümkün olan oyların %50 den fazlası var olmalıdır ki, cluster oluşturulabilsin veya cluster içerisinde kalınabilsin.
- Eğer iki-node lu cluster da sadece node vote ları olursa ne olur? Cluster ın çalışabilmesi için her iki node un da boot etmesi gerekirdi. Oysa ki bu, cluster ın en önemli hedeflerine ters. Çünkü, node un birinin fail etmesi durumunda, cluster ın ayakta kalması gerekir.

Failure Fencing

- Eğer node lar arasında interconnect iletişimi durursa ki bu interconnect hatası veya node crash olayı olabilir, her node diğerinin fonksiyonel olduğunu düşünür. Bu olay split-brain olarak isimlendirilir. İki ayrı cluster ın çalışmasına izin verilemez çünkü bu durum data kaybına neden olur. Her node, quorum oyu elde etmeye çalışarak cluster oluşturmayı deneyecektir. Her iki node quorum u elde etmeye çalışacaktır. Quorum u da alan ilk node, çoğunluğu oluşturacak ve cluster üyesi olarak kalacaktır. Quorum device ı kaybeden node, sun cluster yazılımını abort edecektir çünkü oyların çoğunluğunu alamamış durumdadır.

Amnesia Prevention

- İki node lu bir cluster düşünelim, Node 1 ve Node 2. Node 2 nin bakım için durdurulduğunu veya crash aldığını düşünelim.
- Cluster konfigürasyon değişimi, Node 1 üzerinde yapılır
- Node 1 kapatılsın
- Daha sonra Node2 açılarak cluster oluşturulmaya çalışılsın. Bu durumda eski cluster bilgisiyle sun cluster oluşturulmaya çalışılacaktır. Node 2 de cluster konfigürasyonunun doğru kopyasına sahip değildir.

Persistent Reservations and Reservation Keys

- Persistent reservation, quorum device larda olan reservasyon bilgisi anlamına gelir.
  - Device a bağlı olan tüm node lar resetlense bile hayatta kalır
  - Quorum device ın kendisi power on/off olsa bile hayatta kalır
- Belli türden datanın diskin üzerine yazılması reservasyon key olarak isimlendirilir.
  - Her node a tek 64-bit reservasyon key atanır
  - Fiziksel olarak quorum device a bağlı her node, quorum device a yazılan reservasyon key e sahiptir.
- Quorum device a bağlı iki node düşünelim. Qurum device da Node 1 ve Node 2 ye ait reservasyon key leri mevcuttur. Node 2 herhangi bir nedenle cluster ı terk ederse, Node 1, Node 2 ye ait key ı quorum device dan alacaktır. Eğer split-brain olsaydı, Node 1, quorum device a olan yarışı kazanırdı.
  - Boot aşamasında node quorum device ı eğer üzerinde bir reservasyon key yok ise oy olarak sayamaz. Dolayısıyla amnesia senaryosunda, Node 2 önce boot etmeye çalışırsa, quorum vote u sayamayacaktır ve Node 1 in boot etmesini beklemek zorunda kalacaktır. Node 1 cluster a katıldıktan sonra, Node 1, interconnect lerden Node 2 yi görecek ve reservasyon key ini tekrar quorum device a yazacaktır. Yani reservasyon key, key i hala quorum device da olan başka bir node tarafından quorum device a yazılır.

Cluster Interconnect

- İki türde cluster interconnect mevcuttur: point-to-point ve junction-based
  - Point-to-point konfigürasyonda, interface ler çapraz kablolarla doğrudan birbirine bağlanır. Sun cluster 3.1 kurulumunda her kablo için end-point interface isimlerini vermek zorundasın.
  - Eğer iki node dan fazla ise, interconnect interface leri switch leri kullanarak birleştirmelisin. Switch ler kullanılacak ise, junction-based interface seçilmelidir ve switch isimleri verilmelidir.
- IP adresleri, 172.16.0.0 baz alınarak otomatik olarak verilir. Değiştirilmesi önerilmez. Cluster interconnect IP leri B-Class olmalıdır.

Installmode Flag Ne Demek?

- İlk node u, diğer node ları cluster a almadan önce reboot etmelisin. İlk node u reboot ettikten sonra, installmode flag otomatik olarak set edilir bu provisional mode olduğumuzu gösterir. Bu konumda ek cluster node ları ekleyebilirsin. Her node üzerinde kurulum tamamlandıkça, her node reboot edilir ve quorum vote olmadan cluster a dahil olurlar. Bu konumda eğer ilk node u da reboot edersen, diğer node lar panic verecektir çünkü bir quorum vote elde edemeyeceklerdir. Ortada henüz bir quorum device yok. İlk node reboot edildiği zaman, ilk node a quorum vote olarak 1 verilir. Diğer tüm node ların quorum vote u 0 dır. Installmode flag ın anlamı budur. Cluster node ları, scsetup dan installmode flag ı resetleyene kadar install mode da kalır.

Quorum Mathematics and Consequences

- Cluster çalıştığında aşağıdakilerin farkındadır:
  - Total possible quorum votes (number of nodes + the number of disk quorum votes defined in the cluster)
  - Total present quorum votes (number of nodes booted in the cluster + the number of disk quorum votes physically accessible by those nodes)
  - Total needed quorum votes equal grater than 50 percent of the possible votes
- Bu durumda,
  - Node, needed vote sayısını boot aşamasında bulamaz ise donar ve başka bir node un katılımını bekler
  - Cluster da boot etmiş bir node artık needed votes sayısını bulamaz ise kernel panic verir.

Genel Cluster Status

- /usr/cluster/bin/scstat –q komutuyla cluster üyeliğini ve quorum vote bilgisini görebilirsin.
- Scconf –p ile, CCR da genel cluster bilgisini görebilirsin.
- scdidadm –L ile DID device ları görebilirsin.

Genel Cluster Yönetimi

- Cluster Daemons
  - Cluster – KILL sinyali gönderilirse, sistem kernel panik alır
  - Clexecd – remote cluster komutlarını çalıştırmak için de kullanılır. Eğer daemon öldürülürse ve 30 saniye içinde başlatılmaz ise failfast driver, kernel ı paniğe sürükler
  - Cl_eventd – cluster event larını kaydeder ve yönlendirir. Öldürülürse, rpc.pmfd tarafından yeniden başlatılır
  - Rgmd – cluster-unaware uygulamaların durumunu yönetmek için kullanılır. Eğer daemon öldürülürse ve 30 saniye içerisinde başlatılmaz ise failfast driver, kernel ı paniğe sürükler
  - Rpc.fed – rgmd den gelen istekleri idare eder. Eğer daemon öldürülürse ve 30 saniye içerisinde başlatılmaz ise failfast driver, kernel ı paniğe sürükler
  - Scguieventd – sun plex manager için cluster event lerini işler. Böylece GUI, gerçek zamanlı olarak güncellenir
  - Rpc.pmfd – process leri monitör eder. Bazı cluster framework daemon larını yeniden başlatır. Eğer daemon öldürülürse ve 30 saniye içerisinde başlatılmaz ise failfast driver, kernel ı paniğe sürükler
  - Pnmd – lokal network yönetimi daemon ı. Her node da çalışan local IPMP den alınan network durumlarını yönetir. Eğer node lar üzerinde tam public network hatası oluşursa, uygulama geçişlerini sağlar. Eğer bu daemon durursa rpc.pmfd tarafından yeniden başlatılır.

Cluster Status Komutları

- Scstat –w – cluster transport statüsünü gösterir
- Sccheck – bellek, swap, kontrol eder. Ek olarak, vfstab dosyasını, tüm node lar için /global/.devices dosya sistemine ve diğer global dosya sistemlerine göre kontrol eder
- Scinstall –pv (lokal node üzerinde) - Sun cluster SW revizyonu, yüklü paketleri gösterir
- /etc/cluster/release dosyası, sun cluster sw framework ile ilgili bilgi verir
- Tek node u kapatmak için, init 0 kullanılır. Node u kapatmadan önce, resource gruplar diğer node üzerine alınmalıdır
- Tüm cluster ı kapatmak için, scshutdown komutu kullanılabilir. Scshutdown –y –g 30 kullanılabilir bir komut.
- Node un Cluster operasyonlarına katılmaması için ok prompt da, boot –x komutunu kullanabilirsin.

Cluster Yönetimi Araçları

- Scconf – HW ve device group konfigürasyonu için
- Scrgadm – resource grup konfigürasyonu için
- Scswitch – device group ve resource grup ların geçişi için
- Scstat – cluster statusu için
- Scsetup – menu-driven utility ---- scconf a göre çok daha kolay. Komut opsiyonlarını ezberlemene gerek kalmıyor, tüm işi scsetup yapıyor
- https://nodename:3000 ile SunPlex Manager a bağlanıyorsun
  - Detay için lütfen komutların manual sayfalarına bakınız

VxVM

- VxVM disk gruplarını register yaptığınız zaman, Sun Cluster SW bu disk grupların bilgilerine sahip olur. Register işleminden sonra, disk grup cluster tarafından yönetilir ve cluster import/deport işlemlerini yerine getirir. Açıktır ki, cluster, bu işlem için VxVM komutlarını kullanır. Disk grubu import eden node, primary device grup server olur.
- Sun cluster SW de sadece tek node, disk grup larını import eder. Buna rağmen daha önce de değindiğimiz sun cluster sw global device alt yapısı, disk grup daki device lara tüm node ların erişimini mümkün kılar (node fiziksel olarak bu device lara bağlı olmasa bile). Bu disk grup için primary olmayan hatta disk lere fiziksel olarak bağlı olmayan node lar, cluster transport u kullanarak data ya ulaşır.
- CVM lisanslı bir özelliktir. Birden fazla node un aynı anda disk gruplarına sahip olmasını sağlar. Bu özellik, Oracle Parallel Server (OPS) ve ORACLE Real Application Cluster uygulamaları için kullanılır.

DMP

- Sun Cluster DMP yi DESTEKLEMEZ. Multipathing, Sun StorEdge Traffic Manager SW ile desteklenmektedir. Volume Manager, Sun StorEdge Traffic Manager SW i algılar ve bu path ler için DMP yi kullanmaz.

Registering VxVM Disk Groups

- Yeni disk grupları ve volume leri yarattıktan sonra disk grubu, scsetup veya scconf ile register etmelisin. VxVM disk grup, register edildikten sonra, sun cluster ortamında device grup olarak ifade edilir. VxVM disk grup register edilmediği müddetçe, cluster disk grupları algılamayacaktır. Disk grupları, vxdg list ile görsen bile, scstat –D ile göremezsin.
- Scsetup ana menüden, option 4 (Device Groups and Volumes) seçilerek gelen alt menüden ilgili işlem yapılır.
- UYARI – VxVM disk grupları sun cluster sw e register edildikten sonra, vxdg import / vxdg deport komutlarını disk sahipliğini kontrol etmek için kullanmayınız. Bu cluster ın, device grubunu failed olarak görmesine neden olur. Bunun yerine, scswitch –z –D nfsdg –h node_to_switch_to komutunu kullanın.
- VxVM device grup cluster a register edildikten sonra, yeni bir volume yaratıldığında veya silindiğinde, device grup senkronize edilmelidir. Aşağıdaki komutla Sun cluster disk grupları scan eder ve uygun device dosyalarını yaratır ve siler. Scconf –c –D name=nfsdg,sync komut kullanılabilir. Scsetup dan da menü 4 Device groups and volumes den synchronize volume information seçerek de senkronize işlemini yapabilirsin.

Global and Local File Systems on VxVM Volumes

- Global File Systems – storage a bağlı olmasa bile, tüm node lar tarafından aynı anda ulaşılabilirdir.
- Local failover filesystem – storage a fiziksel olarak bağlı olmalıdır. Data servisi çalıştıran node tarafından mount durumdadır.
- Global File System
  - /dev/vx/dsk/nfsdg/nfsvol /dev/vx/rdsk/nfsdg/nfsvol /global/nfs ufs 2 yes global,logging
- Local failover filesystem
  - /dev/vx/dsk/nfsdg/nfsvol /dev/vx/rdsk/nfsdg/nfsvol /localnfs ufs 2 no logging

SVM

- Sadece aynı diskset içerisinde olan diskler bir birim olarak düşünülür. Yani mirror volume ler oluşturulur, bir bütün olarak node a geçişi sağlanır.
- Shared diskset yaratmak için, SVM shared olmayan diskler üzerinde olmayan lokal diskset e ihtiyaç duyar. Aslında ihtiyaç sadece, lokal diskset metadb varlığıdır.
- Local Replica Mathematics
  - Tanımlı metadb replikaların %50 den azı kullanılabilir ise, SVM çalışmayacaktır
  - Boot aşamasında tanımlı metadb replikaların %50 si veya %50 den azı kullanılabilir ise, node boot etmeyecektir. Fakat tek-kullanıcı modda sistemi açarak kullanılamaz olanları silebilirsin.
- Shared diskset replika yönetimi
  - Her diskset için ayrı metadb replikaları vardır
  - Metadb replikalar, disk leri diskset lere eklediğimizde otomatik olarak slice 7 de oluşur
- Shared diskset replika quorum matematiği
  - Diskset için tanımlı replikaların %50 den azı varsa, diskset çalışmayacaktır
  - Diskset için tanımlı replikaların %50 si veya daha azı var ise, diskset switch-over yapılamaz ve de ownership liğide alınamaz

Shared Diskset Mediators

- Sun cluster 3.1, SVM için özel bir add-on içerir ki bu add-on mediators olarak isimlendirilir. Mediator ler syaesinde, node un kendisi, diskset metadb replikaların tam olarak %50 si gittiği zaman, “tie-breaking votes” olarak tsnıtır. Her node üzerinde, mediator datası memory de tutulur. Eğer array ı keybedersen, node mediator leri, golden statüse geçecektir ve shared diskset quorum matematiği için ek iki oy getirecektir. Bu, metadb replikaların %50 si kalsa bile, diskset operasyonlarının yapılmasına olanak sağlar. Bu noktada node un tekini de kaybedebilirsin. Bu durumda hala tek node, golden mediator olarak görev yapacaktır.

Shared Diskset and Mediatots

- metaset –s nfsds –a –h proto192 proto193 → boş bir diskset oluşturuyor. –h den sonraki ilk host, disksetin sahibidir.
- metaset –s nfsds –a –m proto192 proto193 → mediator ekliyorsun.
- metaset –s nfsds –a /dev/did/rdsk/d9 /dev/did/rdsk/d17 → diskset e disk ekliyorsun
- metaset → disksetin içeriğini görüntüler
- metadb –s nfsds → diskset içerisinde olan replikaları gösterir
- medstat –s nfsds → mediator lerin statüsünü gösterir

Shared Disksets

- Bir disk, disksete eklendiği zaman 0. silindirden başlayan ufak bir alan s7 de replikalar için oluşturulur. Geri kalan alan s0 a alınır ve s2 silinir.
- Did isimlerini kullanarak, stripe/concate ler yaratabilir, yarattığın stripe/concate leri mirror layabilir ve oluşturduğun mirror metadevice dan soft bölümlemeler yaratabilirsin.
- Mesela,
  - /dev/did/rdsk/d9s0 dan d101 (stripe/concat)
  - /dev/did/rdsk/d17s0 dan d102 /stripe/concat)
  - d101 ve d102 den d100 mirror
  - D100 mirror dan da d10, d11 soft partition lar yaratabilirsin
- #metainit –s nfsds d101 1 1 /dev/did/rdsk/d9s0
- #metainit –s nfsds d102 1 1 /dev/did/rdsk/d17s0
- #metainit –s nfsds d100 –m d101
- #metattach –s nfsds d100 d102
- #metainit –s nfsds d10 –p d100 200m
- #metainit –s nfsds d11 –p d100 200m
- #metastat –s nfsds → volume leri kontrol ediyorsun

SVM shared DiskSets

- SVM ile shared diskset yarattığında, diskset ler otomatik olarak cluster ın yönettiği device grup olur. Yani otomatik olarak register olurlar, VxVM disk grupları gibi manual olarak register etmen gerekmiyor. Diskset içerisinde yeni bir metadevice yarattığında veya diskset içerisinden metadevice sildiğinde senkronize işlemine gerek yok, otomatik olarak senkronize oluyor. Device gruplarını, scstat –D ve/veya scconf –pv | grep nfsds komutlarıyla görebilirsin.
- Eğer diskset i switch-over yaparken mirror işlemi varsa, sync işlemi diğer node a geçiş olduğunda yeniden başlar. Device grubu karşı node üzerine almak için

#scswitch –z –D nfsds –h proto192

Global and Local FileSystems on Shared Diskset Devices

- Global ve local-failover arasındaki fark, /etc/vfstab dosyasındaki mount-at-boot ve options kolonlarıdır.
- Global dosya sistemi,
  - /dev/md/nfsds/dsk/d10 /dev/md/nfsds/rdsk/d10 /global/nfs ufs 2 yes global,logging
- Local fail-over dosya sistemi,
  - /dev/md/nfsds/dsk/d10 /dev/md/nfsds/rdsk/d10 /global/nfs ufs 2 no logging

Data Servisleri

- Sun Cluster SW agent lar, data servislerin cluster ortamında düzgün çalışmasını sağlar. Data servis agent lar aşağıdakileri içerir,
  - Servis için fault monitor
  - Cluster içerisindeki servisi başlatmak ve sonlandırmak için metodlar
  - Fault monitörü başlatmak ve sonlandırmak için metodlar
  - Cluster içerisinde ki servisin konfigürasyonunu doğrulamak için metodlar
  - Sayılan tüm metodlar hakkında tüm bilgileri tutmaya olanak sağlayan registration information dosyası. Bu sayede agent içerisindeki tüm bileşenleri işaret eden resource type ı referans olarak göstermen yeterli olacaktır
- Data servis agent fault monitor bileşenleri,
  - Fault monitör bileşenleri, data servisine özeldir ve data servisin çalıştığı node da çalışır. Fault monitör bileşenleri uygulama hatalarını algılamak için yazılmışlardır ve uygulamanın yeniden başlatılmasını veya diğer node a geçişini önerebilir. Genel anlamda fault monitör,
    - Daemon ların sağlığını kontrol eder bunu process monitoring facility (rpc.pmfd) kontrolü altına alarak sağlar
    - Client komutlarını kullanarak servislerin sağlıklı olup olmadığını kontrol eder

Data Servisler

- Scinstall aracı, pkgadd yerine menu-driven bir arabirim ile data servis agent larının yüklenmesine olanak sağlar
- Data servis agent, agent hakkındaki tüm bilgileri resource type olarak bilir. Cluster yazılımına, resource type kaydedildiği zaman, agent bileşenlerinin ismini veya lokasyonunu bilmek zorunda değilsin. Yapman gereken tek şey, uygulama için resource type ını referans olarak göstermendir. Bu sayede, bileşenler için doğru fault monitör ve methodları garantilemiş oluyorsun.
- Dikkat, paketlerle, resource type lar farklıdır. Mesela SUNWschtt paketi için resource type, SUNW.iws dir.

Resources, Resource Groups, Resource Group Manager

- Data servislerini cluster kontrolüne vermek için, servisler, resource gruplar içerisinde resource olarak konfigüre edilir. Rgmd, resource group manager dır. Resource gruplar ve resource larla ilgili tüm aktiviteyi kontrol eder. Rgmd daemon, cluster içerisindeki tüm data servisleri kontrol eder.
- Bir resource, cluster framework katmanı üzerinde çalışan bir element olarak düşünülebilir. Öyleki bu element açılıp kapatılabilir ve monitör edilebilir. Her resource, spesifik bir türe sahiptir. Mesela apache web server için tür, SUNW.apache dir. IP yi ve storage ı ifade eden resource larda cluster için gereklidir.
- Bir resource, türü, tek olması gereken simi ile, bir dizi özelliğiyle anılır.
- Resource groups, resource ların toplamıdır. Failover veya scalable olabilir. Failover uygulamalr için, resource group bir birimdir ve bir t anında sadece tek node üzerinde çalışır ve diğer bir node a resource group içindeki resource larla birlikte geçiş yapar. Birden fazla data servis, aynı resource grup içerisinde olabileceği gibi farklı resource grup içerisindede olabilir.

FailOver Resource Gruplar

- Bir resource ismi global olarak tek olmalıdır. Yani sadece resource grup içerisinde tek olması yetmez.
- Resource lar resource grup içerisinde olmalıdır
- Her resource, resource type a sahiptir. Mesela SUNW.nfs, NFS resource içindir.
- Özel resource türleri
  - SUNW.LogicalHostname resource type – belli bir subnet de olan IP adresini ifade eder. İlgili servis için logicalIP dir. İstemciler bu IP yi kullanarak servise erişirler. SUNW.LogicalHostname resource ile tanımlanan her IP adresi servislerle birlikte node dan node a geçiş yapar
  - SUNW.SharedAddress resource type – scalbale servisler için ihtiyaç duyulan özel bir IP adresidir.
  - SUNW.HAStorage resource type – resource grubun online olacağı node üzerinden global device veya global dosya sistemlerinin ulaşılıp ulaşılmadığını kontrol eden START metodu vardır – Resource_dependencies standard özelliği ile real data servisini SUNW.HAStorage resource type bağlarsın, böylece rgmd, servisin bağlı olduğu storage kullanılabilir durumda değil ise servisi başlatmayı denemez- AffinityOn resource özelliğini True ya set edersen, SUNW.HAStorage resource türü resource gruplarını ve disk device gruplarını aynı node üzerinde konumlandırmaya çalışır. Bu sayede disk bağımlı servisler için performans artmış olur. Eğer device grubun, resource grupla birlikte taşınması imkansız ise (servis, storage a bağlı olmayan node üzerine taşınıyorsa mesela), bu özellik hala kullanılabilir. AffinityOn=true, device grubu nereye alabiliyorsa, oraya taşır. SUNW.HAStorage resource type, sadece global device ve global file sistemleri destekler. Bu resource type, dosya sistemlerini bir node dan unmount edemez, başka bir node a mount da edemez. Sadece bu resource type ın START metodu, global device veya file systemlerin kullanılabilir olup olmadığını kontrol eder.

SUNW.HAStoragePlus

- SUNW.HAStoragePlus, SUNW.HAStorage ı kapsar. Local failover dosya sistemi desteği ek olarak da gelmiş durumdadır. Local fail over dosya sistemi, servisin failover olduğu node üzerine failover olmalıdır. Failover, storage ın fiziksel olarak bağlı olduğu node a olmalıdır. Local failover dosya sisteminin performans avantajı vardır. Scalable servisler için kullanılamzlar. Yine, storage ın bağlı olmadığı node üzerine failover yapılacaksa da kullanılamazlar. Global device ve global file sistemi hala desteklemektedir.
- SUNW.HAStoragePlus resource türü, global ve lokal file system için FilesystemMountpoints özelliğini kullanır. Bu resource türü, vfstab dosyasına bakarak global ve lokal file sistem arasındaki farkı anlar. Global dosya sistemi, mount at boot kolonunda yes, mount options kolonunda global,logging i içerir. Bu durumda, SUNW.HAStoragePlus, SUNW.HAStorage olarak davranır. Yani START metodu sadece kontrol yapar, STOP method ise hiçbir şey yapmaz. Lokal file sistem, mount at boot kısmında no ya sahiptir, ve mount options da global opsiyonu yoktur. Bu durumda STOP metodu, bir node dan dosya sistemini unmount eder; START method da diğer node a file sistemi mount eder.

SUNW.HAStorage & SUNW.HAStoragePlus

- SUNW.HAStorage eskidi. Yani Cluster ın yeni sürümlerinde bu olmayacak. Zaten, SUNW.HAStoragePlus, SUNW.HAStorage ın sahip olduğu özellikleri de kapsıyor. Eee o zaman ne gerek var SUNW.HAStorage ı kullanmaya? Değil mi?

Global File System & Local File System

- Global File system – eğer scalable servis kullanacaksan, servisi, storage ın fiziksel olarak bağlı olmadığı node üzerine fail over yapacaksan kullanmanda fayda var. Hala AffinityOn=true kullanabilirsin. Storage ı servis ile birlikte geçirmeye çalışacak. Performans kazancı sağlar.
- Local File System – file system sadece failover servisler içinse, Nodelist, storage a fiziksel olarak bağlı node ları içeriyorsa, yani resource grup için Nodelist ile device için Nodelist aynı ise kullanılabilir. Bu koşullar altında, lokal dosya sistemi daha fazla performans kazancı getirecektir.

Resource Dependencies

- Aynı grup içerisindeki resource lar arasında bağımlılık oluşturabilirsin. Eğer Resource A, Resource B ye bağlıysa;
  - Resource B gruba ilk eklenmelidir
  - Resource B ilk başlatılmalıdır
  - Resource A ilk önce sonlandırılmalıdır
  - Resource A gruptan ilk önce silinmelidir
  - Rgmd daemon, eğer Resource A hatalıysa, Resource B yi başlatmayı denemeyecektir
- Eğer weak dependency var ise, yukardakilerin sonuncusu hariç geçerlidir.

Resource ve Resource Group Özellikleri

- Resource ve resource grup özellikleri, name=value şeklindedir.
- Standard resource properties leri, herhangi türde olan resource lar için kullanılabilir. man r_properties ile standard resource özelliklerini ve anlamlarını görebilirsin.
- Extension properties ler ise, resource türlerine spesifiktir. Bunun için de spesifik resource type ın manual sayfasına bakabilirsin. man SUNW.apache, man SUNW.HAStoragePlus.
- Resource grup properties – tüm resource gruba uygulanır. man rg_properties ile bilgi alabilirsin.
- scrgadm komutuyla resource lara ilşikin, resource türlerine ilişkin, resource gruplara ilişkin konfigürasyon bilgilerini görebilir ve set edebilirsin.

Scswitch Komut

- #scswitch –F –g nfs-rg → resource grubu kapatmak için
- #scswitch –Z –g nfs-rg → resource grubu açmak için
- #scswitch –z –g nfs-rg –h node → resource grubu başka bir node üzerine almak için
- #scswitch –R -g nfs-rg –h node → resource grubu restart yapmak için
- #scswitch –S –h node → resource ları ve resource grupları bir node üzerinden boşaltmak için
- #scswitch –n –j nfs-res → resource u ve fault monitor ü disable yapmak için
- #scswitch –e –j nfs-res → resource u ve fault monitor ü enable yapmak için
- #scswitch –c –j nfs-res –h node –f STOP_FAILED → STOP_FAILED flagı temizlemek için
- #scswitch –n –M –j nfs-res → fault monitor ü disable yapmak için
- #scswitch –e –M –j nfs-res → fault monitörü enable yapmak için
- #scstat –g → resource ve resource grupların statüsünü gösterir