Hochverfügbare LDoms mit Oracle Solaris Cluster

Marco Stadler
JomaSoft GmbH
St. Gallen / Schweiz

Schlüsselworte
Oracle Solaris, Virtualisierung, Solaris Zonen,
LDom, Oracle VM Server for SPARC, Oracle Solaris Cluster, VDCF

Einleitung
Die modernen Oracle SPARC T5-Server sind ideale Plattformen um Applikationen unter Solaris 11 zu konsolidieren, und alte Server abzulösen. Diese T-Systeme verfügen über die, von der Hardware unterstützte, 'Logical Domains' (LDom) Technologie, mit welcher man mehrere, voneinander unabhängige Solaris Instanzen, auf demselben Server betreiben kann. LDoms können von einem physikalischen Server auf einen anderen migriert werden, falls die Hardware ausfällt. Wenn die Verfügbarkeit der Applikation verlangt, dass dies möglichst schnell passiert, kann dies mit Hilfe von Oracle Solaris Cluster automatisch erledigt werden, und somit eine Hochverfügbarkeit des Betriebssystems erreicht werden. Dieser Vortrag beschreibt, wie man LDoms mit dem Cluster verbindet, und wie diese vom Cluster überwacht werden. JomaSoft hat mit seinem 'Virtual Datacenter Control Framework' diesen Prozess für einen Kunden vollautomatisiert und zusätzliche Monitoring Möglichkeiten hinzugefügt. Dies erlaubt dem Kunden in kürzester Zeit neue Umgebungen bereit zu stellen, welche alle gleich aussehen und damit einfach zu verwalten sind.

JomaSoft GmbH
Die JomaSoft wurde als Software und Beratungs-Unternehmen im Jahr 2000 gegründet. Als Oracle Gold Partner sind wir insbesondere auf Oracle Solaris 11, SPARC T4 und T5 Server spezialisiert. Wir bieten Software-Entwicklung, Consulting, Implementation und Administration im Bereich Solaris. Abgerundet wird unser Angebot durch das Produkt VDCF. Ein Framework, welches die Installation, das Management und Desaster Recovery von Solaris Servern, Solaris Zonen und LDoms vereinfacht und automatisiert. Unterstützt sind die Betriebssystem Versionen Solaris 10 und Solaris 11 auf den Plattformen SPARC und x86. Dieses Framework wird bei zahlreichen Kunden in Europa seit mehr als 8 Jahren produktiv eingesetzt.

LDoms (Oracle VM Server for SPARC)
'Logical Domains' sind, wie die Bezeichnung andeutet, logische Domänen, welche vom Hypervisor der SPARC-T-Systeme unterstützt werden. Voraussetzung für diese Technologie ist ein Oracle SPARC Server der T-Serie (CMT System). Nur in dieser Server Hardware ist der notwendige Hypervisor integriert. Da der Hypervisor in der Hardware/Firmware enthalten ist, wird der Virtualisierungs-Overhead auf ein Minimum reduziert. Die LDom Manager Software ist Bestandteil von Solaris 11 und kann für Solaris 10 kostenlos von Oracle bezogen werden. In jede logische Domäne (LDom) kann eine unabhängige Solaris Betriebssystem Instanz installiert werden. Somit können verschiedene Solaris Releases gleichzeitig auf derselben Hardware betrieben werden. Dies ergibt eine ideale Möglichkeit parallel zu Solaris 10 neue Solaris 11 Umgebungen aufzubauen. Via Control Domain werden die LDoms (oder Guest Domains) verwaltet. Die Control Domain stellt virtuelle Devices und Services bereit, welche den LDoms zugeteilt werden können und somit den Zugriff auf Disks und Netzwerk ermöglichen. Ressourcen wie CPU und Memory werden den LDoms fix zugewiesen, können aber später auch zur Laufzeit verändert werden. Eine LDom kann ohne Unterbrechung von einem Server auf einen anderen migriert werden, wenn die Daten auf einem zentralen Storage abgelegt sind (Live Migration).

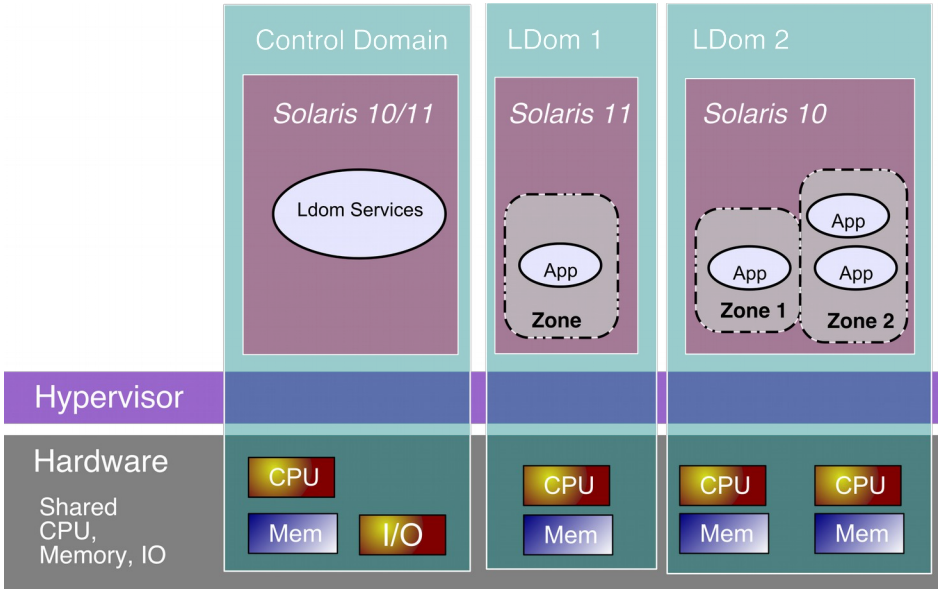

Abbildung 1: Oracle VM Server for SPARC

LDom ist eine kostenlose Technologie, welche die Virtualisierung und Konsolidierung im Solaris Rechenzentrum unterstützt. Mit den von Oracle angebotenen physical-to-virtual (P2V) Tools lassen sich alte, nicht mehr supportete Systeme einfach, auf neue Hardware migrieren, ohne dabei an der eigentlichen Server-Installation etwas verändern zu müssen. Dank den Migrations-Funktionen können die LDoms bei Bedarf zwischen Systemen verschoben werden. Es können Kosteneinsparungen erzielt werden, da die bestehende Hardware besser ausgelastet werden kann. Die reduzierte Anzahl physischer Server führt zu weniger Bedarf an Platz, Strom und Kühlung. Für Oracle Software sind LDoms als "Partitionen" akzeptiert, womit sich Software Lizenz Optimierung/Einsparungen erzielen lassen. Die LDom Technologie bietet von sich aus keine Features, um solche Instanzen hoch verfügbar zu machen. Um die Verfügbarkeit der LDom vor Hardwareausfällen zu schützen, kann Oracle Solaris Cluster eingesetzt werden.

Mit LDoms können neue Applikationsumgebungen in wenigen Minuten bereitgestellt werden. Aus organisatorischen Gründen empfiehlt es sich pro Kunde/Mandant mindestens eine LDom zu erstellen und in der LDom mehrere Solaris Zonen für die einzelnen Applikationen/Umgebungen. Damit gewinnt man maximale Flexibilität, weil die Zonen unter den LDoms auch transportierbar (migrierbar) sind. D.h. sie können von einer LDom zu anderen verschoben werden, wenn dies gewünscht ist. Durch die darunter liegende LDom ist es auch möglich, unterbruchsfrei von einer Hardware auf eine andere zu gelangen, was mit den Zonen alleine noch nicht möglich ist.
Weil die LDom Technologie eine Zunahme von Technologien, Komplexität und Flexibilität im Rechenzentrum bedeutet, sind bei einem Ausfall eines physischen Servers zahlreiche Solaris Instanzen und Applikationen betroffen. Darum ist es wichtig, ein geeignetes Management Werkzeug einzusetzen, damit man möglichst viel standardisieren und automatisieren kann.

Voraussetzungen für hoch verfügbare LDoms
Oracle Solaris Cluster bietet seit Version 4.0 (Solaris 11) oder Version 3.3 (Solaris 10) die Möglichkeit LDoms zu überwachen, und diese bei einem Hardware Ausfall auf eine andere Hardware zu migrieren. Die Cluster Software muss dafür in allen Control Domains, welche zum Cluster gehören sollen, installiert werden. Voraussetzung für die Version 4.x ist dabei eine Solaris 11 Version als Control Domain (primary) Betriebssystem. Für die Guest Domains kann nach belieben Solaris 10 oder 11 benutzt werden. Der Cluster in der Control Domain benötigt das Feature 'HA LDOM', welches als Paket unter der Bezeichnung 'ha-cluster/data-service/ha-ldom' im Solaris 11 Repository (IPS) zu finden ist.
Des weiteren benötigt der Cluster zwei freie Netzwerk Ports, welche für die Cluster Privat Interconnects benutzt werden. Auf diesem Netzwerk kommunizieren die Cluster Nodes untereinander. Die LDom wird auf shared Storage im SAN abgelegt. Dieser muss mit allen Cluster Nodes verbunden sein. Die LDom Konfiguration wird vom Cluster in das Cluster Configuration Repository (CCR) gespeichert, auf welches alle Cluster Members über das private Netzwerk Zugriff haben. So kann bei Bedarf auf jedem anderen Cluster Node die LDom wieder erzeugt werden, vorausgesetzt die virtuellen Services, wie Disk- und Netzwerkkonfigurationen, sind auf allen Nodes gleich konfiguriert. Das Setup von einem Zwei-Node-Cluster entspricht in etwa dem Inhalt der folgenden Grafik.

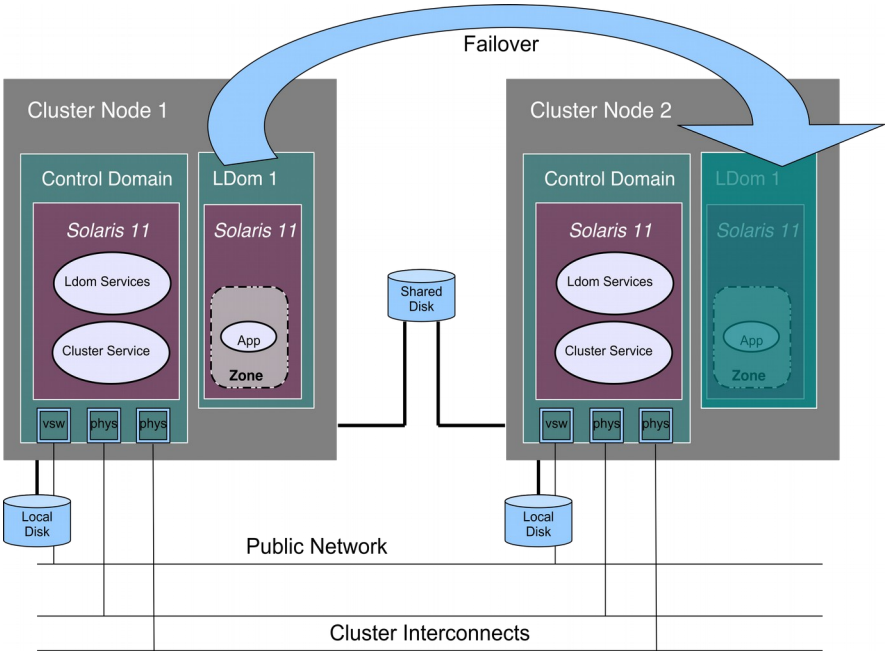

Abbildung 2: Oracle Solaris Cluster Konfiguration

Was wird im Cluster überwacht, wann wird eine Migration ausgelöst?
Die HA LDom Probe im Cluster prüft für die Überwachung der Guest Domain alle 60 Sekunden den Domain Status. D.h. es werden keine Tests innerhalb der LDom und dessen Betriebssystem durchgeführt. Als funktionierende LDom werden folgende Stati angeschaut: active, suspending, resuming, suspended und starting. Bei einem anderen Status wird die LDom neu gestartet oder auf einen anderen Cluster Node verschoben, sollte sie auf dem bestehenden Node nicht einen akzeptablen Zustand erreichen.
Oracle Solaris Cluster lässt sich damit gut für den Schutz gegen Hardware Ausfälle benutzen. Wenn das OS in der LDom aus irgendeinem Grund auf den OK Prompt fällt, wird das vom Cluster nicht bemerkt, und es findet kein Failover auf einen anderen Cluster Node statt.
Warum es trotzdem Sinn macht, Oracle Cluster einzusetzen, und wie JomaSoft eigene Erweiterungen für eine bessere Überwachung implementiert hat, wird weiter unten im Artikel beschrieben.

Wie werden LDoms im Oracle Solaris Cluster integriert?
Um eine Guest Domain vom Cluster überwachen zu können, muss diese zuerst manuell, oder wenn vorhanden mit einem Framework, erstellt werden. Dabei ist es wichtig, dass alle virtuellen Services von Netzwerk- und Diskkomponenten auf allen Cluster Nodes genau gleich verfügbar sind. Die LDom wird nur auf einem Cluster Node konfiguriert und ist zur Laufzeit auch nur auf einem Node aktiv. Damit bei einem Ausfall der Control Domain die LDoms nicht weiter ausgeführt werden, muss die Master Ausfall Policy auf reset gesetzt werden. Wenn dieser Fall eintritt, werden dadurch alle Slave Domains sofort runter gerissen. Diese Konfiguration ist ebenfalls auf allen Cluster Nodes gleich zu setzen:

```
# ldm set-domain failure-policy=reset primary
# ldm list -o domain primary
```

Im Gegenzug muss bei allen Guest Domains definiert werden, wer sein Master ist, sonst wirkt die oben genannte Konfiguration nicht:

```
# ldm set-domain master=primary g0078
```

Danach wird die LDom wie folgt unter Cluster Kontrolle gebracht. Um die LDom im Cluster nun hoch verfügbar zu machen, muss zuerst der Cluster Resource Type SUNW.ldom im Cluster registriert werden. Anschließend wird die LDom als resource Gruppe erfasst, wie das folgende Beispiel zeigt:

```
sc-node1# clresourcetype register SUNW.ldom
sc-node1# clresourcegroup create g0078_rg
sc-node1# clresource create -g g0078_rg -t SUNW.ldom \
-p Migration_type=NORMAL \
-p Domain_name=g0078 g0078_LDOM
```

Nun sehen die Cluster Ressourcen folgendermaßen aus:

```
sc-node1# clrg status g0078_rg

=== Cluster Resource Groups ===

Group Name       Node Name       Suspended      Status
----------       ---------       ---------      ------
g0078_rg         s0028           No             Offline
                 s0009           No             Online
```

```
sc-node1# # clrs status g0078_LDOM

=== Cluster Resources ===

Resource Name       Node Name      State        Status Message
-------------       ---------      -----        --------------
g0078_LDOM          s0028          Offline      Offline - Successfully
stopped g0078
                    s0009          Online       Online - g0078 is active
(normal)
```

Wie arbeitet VDCF mit HA LDom zusammen?
Die Firma JomaSoft in der Schweiz beschäftigt sich seit 15 Jahren mit Oracle Solaris und SPARC. Eine selbst entwickelte Data Center Management Software (VDCF) integriert und automatisiert jeweils die neusten Features und Möglichkeiten, welche diese Technologien anbieten. Für den grössten ICT-Anbieter in der Schweiz wurde die LDom Technologie zusammen mit Oracle Solaris Cluster hoch verfügbar gemacht, um damit den hohen Anforderungen ihrer Kunden gerecht zu werden.
Mit VDCF können vollautomatisiert physikalische Server als Cluster Nodes installiert werden. Danach ist man in der Lage, mit wenigen Befehlen mittels dem Framework LDoms auf den Cluster Nodes erzeugen zu lassen. Diese werden automatisch im Cluster registriert und sind somit sofort unter Cluster Kontrolle, damit bei einem Ausfall, diese auf einem anderen Node neu gestartet werden. Für den Benutzer ist es völlig transparent, ob er mit Cluster Nodes oder normalen Control Domains arbeitet. Wenn auf einem Cluster Mitglied eine LDom erzeugt wird, integriert diese das Framework automatisch im Cluster und führt alle nötigen Konfigurationsschritte durch, damit alles richtig eingestellt ist und nichts vergessen geht. Auch bei einer Migration der LDom kann das Framework unterscheiden, ob es die Kontrolle an den Cluster übergeben muss, oder ob die LDom vom Framework internen Mechanismus transferiert werden muss.

Welche Erweiterungen bietet VDCF für HA LDOMs?
Damit die Verfügbarkeit der LDoms im Oracle Solaris Cluster noch verbessert werden konnte, hat JomaSoft das Monitoring erweitert. Zusätzlich kann man mit VDCF die LDoms per Ping überwachen. Falls die Guest Domain nicht mehr per Ping erreichbar ist, wird geprüft, ob die LDom Console noch verfügbar ist und erreicht werden kann. Wenn dies auch nicht mehr möglich ist, wird der Cluster angewiesen, die LDom auf einem anderen Node neu zu starten.
Des weiteren kann man im VDCF konfigurieren, dass der ZPOOL Failmode für Cluster LDoms auf panic gesetzt wird. Damit wird garantiert, dass die LDom bei einem Fehler vom Zpool ebenfalls auf einem anderen Node neu gestartet wird. Das kann z.B. der Fall sein, wenn ein Storage Device im SAN nicht mehr erreichbar ist.
Als dritte Erweiterung wurde ein IPMP Monitor implementiert. Dieser überwacht die IP-Multi-Path Gruppen in der LDom. Sollten z.B. beide Pfade einer Gruppe ausfallen, wird vom Monitor ein komplett Ausfall registriert, und ein Cluster Switch wird ausgelöst. Damit wird die LDom auf einem anderen, funktionsfähigen Node neu gestartet, und das Netzwerk steht wieder ordnungsgemäß zur Verfügung.

Die gesamte VDCF Produkt Dokumentation ist öffentlich. Eine frei verfügbare Test-Version "VDCF Free Edition" ist auf unserer Website ebenfalls zu finden: **http://www.jomasoft.ch/vdcf**

Kontaktadresse:
Marco Stadler
JomaSoft GmbH
Falkensteinstrasse 54a
CH-9000 St. Gallen

Telefon:	+41 (0)71-288 92 11
Fax:	+41 (0)71-288 92 12
E-Mail	stadler@jomasoft.ch
Internet:	www.jomasoft.ch