Hoch verfügbare LDOMs mit Oracle Solaris Cluster

Marco Stadler, JomaSoft GmbH

Dieser Artikel beschreibt das Thema „Hochverfügbarkeit von Oracle VM Server for SPARC (LDOMs)“ mittels Oracle Solaris Cluster. Dabei wird gezeigt, welche Möglichkeiten und Limitation die Cluster-Software hat und wie sie sich durch die VDCF-Software besser integrieren und um weitere Probes erweitern lässt.

Voraussetzung für die Technologie ist ein Oracle SPARC Server der T-Serie (CMT-System), denn nur in dieser Server-Hardware ist der notwendige Hypervisor integriert *(siehe Abbildung 1)*. 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 sein. Somit lassen sich verschiedene Solaris-Releases gleichzeitig auf derselben Hardware betreiben – eine ideale Möglichkeit, parallel zu Solaris 10 neue Solaris-11-Umgebungen aufzubauen.

Control Domain verwaltet die LDOMs (oder Guest Domains). Sie stellt virtuelle Devices und Services bereit, die den LDOMs zugeteilt werden können und somit den Zugriff auf Disks und Netzwerk ermöglichen. Ressourcen wie CPU und Memory werden den LDOMs fix zugewiesen, lassen sich aber später auch zur Laufzeit verändern. Eine LDOM kann ohne Unterbrechung von einem Server auf einen anderen migriert werden, wenn die Daten auf einem zentralen Storage abgelegt sind (Live Migration).

LDOM ist eine kostenlose Technologie, um die Virtualisierung und Konsolidierung im Solaris-Rechenzentrum zu unterstützen. Mit den von Oracle angebotenen „physical to virtual“-Tools (P2V) lassen sich alte und nicht mehr unterstützte Systeme einfach auf neue Hardware migrieren, ohne dabei an der eigentlichen Server-Installation etwas verändern zu müssen.

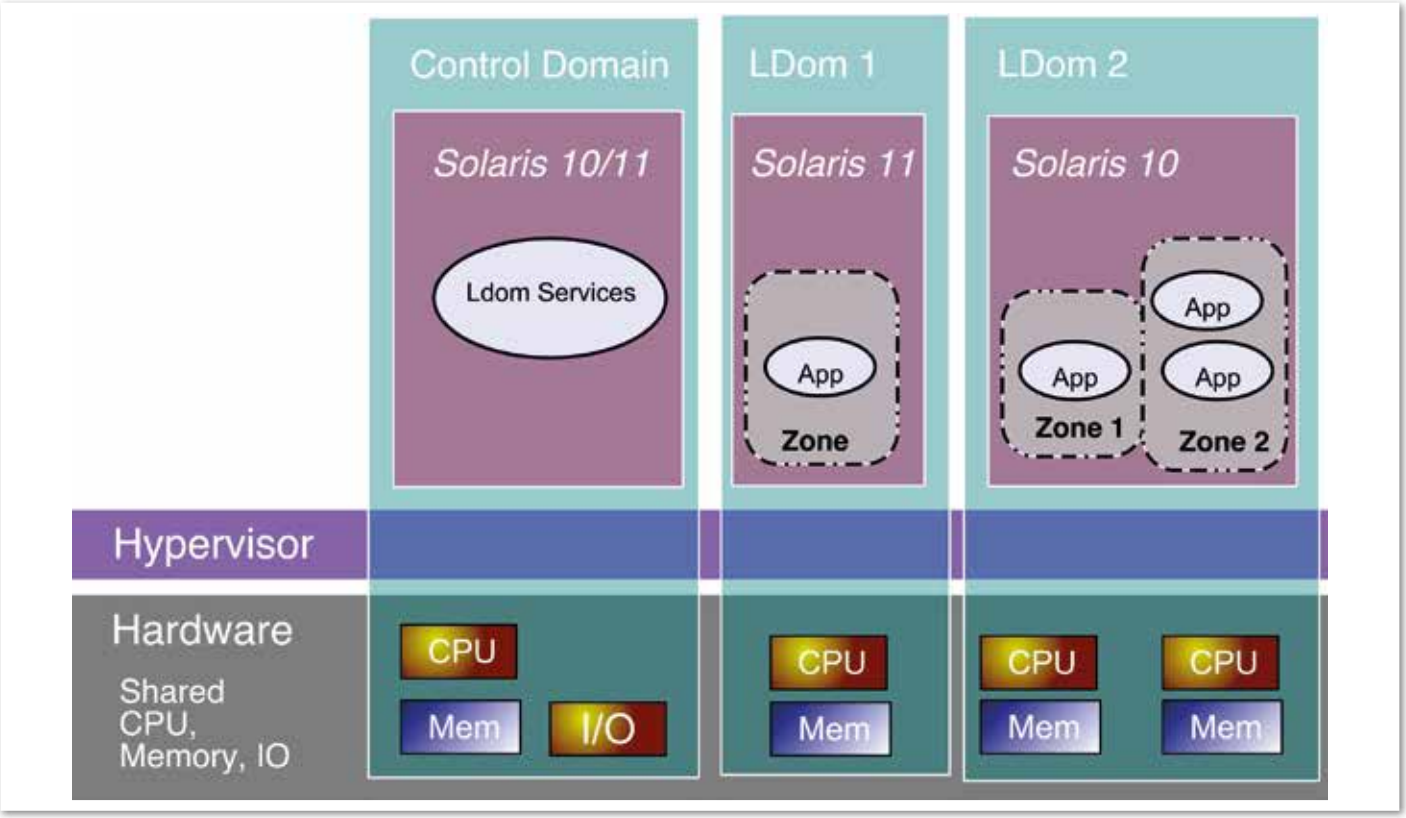

Abbildung 1: Oracle VM Server for SPARC

Dank der Migrations-Funktionen können die LDOMs bei Bedarf zwischen Systemen verschoben werden. Dabei lassen sich Kosten einsparen, da die bestehende Hardware besser ausgelastet ist. Die reduzierte Anzahl physischer Server führt zu weniger Bedarf an Platz, Strom und Kühlung.

Für die Oracle-Software sind LDOMs als „Partitionen" akzeptiert, wodurch Software-Lizenz-Optimierungen/-Einsparungen entstehen. Die LDOM-Technologie bietet von sich aus keine Features, um solche Instanzen hochverfügbar zu machen. Oracle Solaris Cluster kann die Verfügbarkeit der LDOM vor Hardware-Ausfällen schützen.

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; sie können also von einer LDOM zur anderen verschoben werden, wenn dies gewünscht ist. Durch die darunterliegende LDOM ist es auch möglich, unterbrechungsfrei von einer Hardware auf eine andere zu gelangen, was mit den Zonen allein 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 hochverfügbare 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, die zum Cluster gehören sollen, installiert sein. Voraussetzung für die Version 4.x ist dabei eine Solaris-11-Version als Control-Domain-Betriebssystem (primary). Für die Guest Domains kann nach Belieben Solaris 10 oder 11 zum Einsatz kommen. Der Cluster in der Control Domain benötigt das Feature „HA LDOM", das als Paket unter der Bezeichnung „ha-cluster/data-service/ha-LDOM" im Solaris 11 Repository (IPS) zu finden ist.

Darüber hinaus benötigt der Cluster zwei freie Netzwerk-Ports, um für die Cluster Privat Interconnects nutzen zu können. Auf diesem Netzwerk kommunizieren die Cluster Nodes untereinander. Die LDOM wird auf Shared Storage im SAN abgelegt. Dieser muss natürlich mit allen Cluster Nodes verbunden sein. Die LDOM-Konfiguration wird vom Cluster in das Cluster Configuration Repository (CCR) gespei-

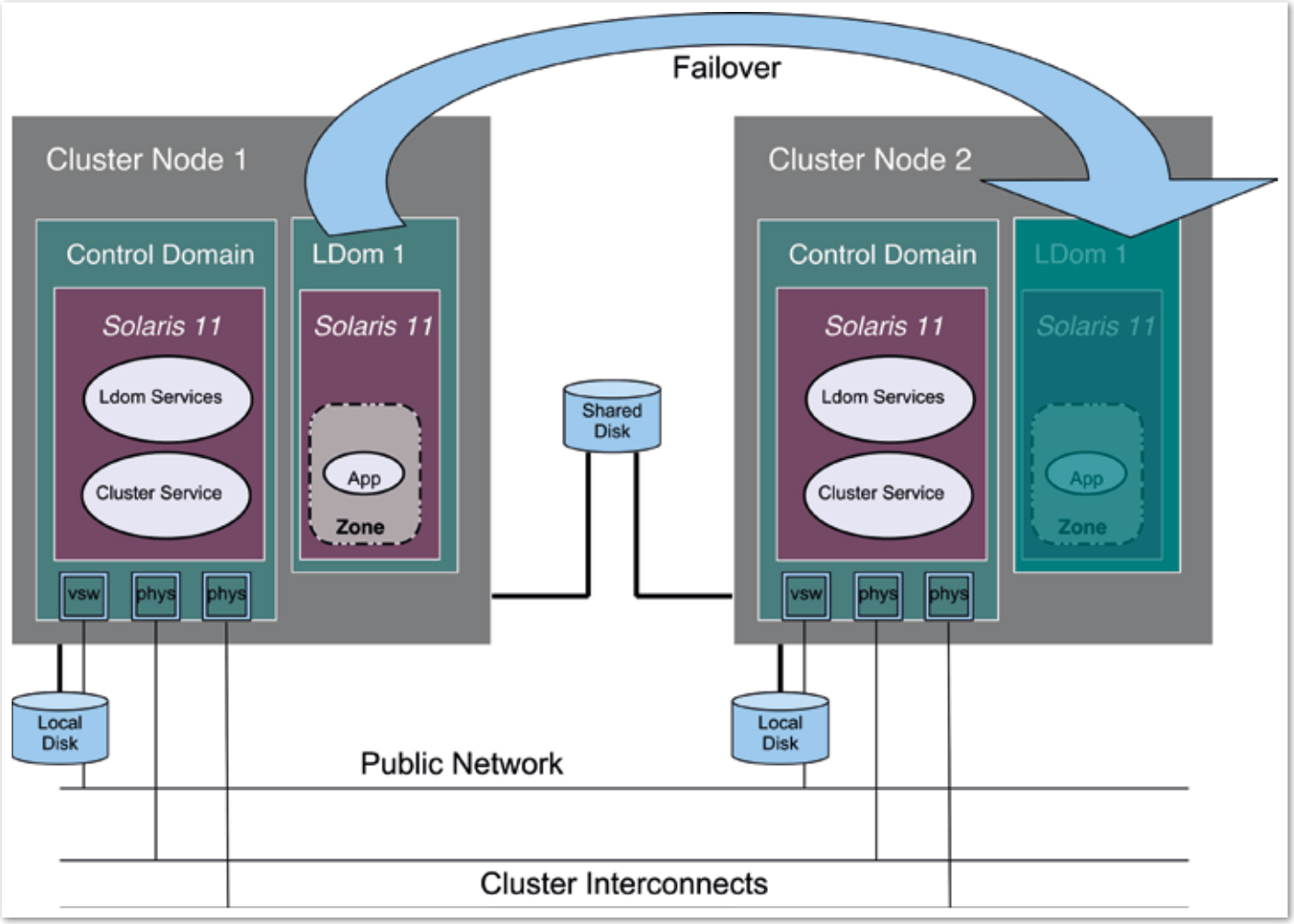

Abbildung 2: Oracle-Solaris-Cluster-Konfiguration

chert, auf das alle Cluster-Member 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 Netzwerk-Konfigurationen, sind auf allen Nodes gleich eingestellt. *Abbildung 2* zeigt das Setup von einem Zwei-Node-Cluster.

Das Cluster überwachen und eine Migration auslösen

Die HA LDOM Probe im Cluster prüft für die Überwachung der Guest Domain alle sechzig Sekunden den Domain-Status. Es erfolgen also keine Tests innerhalb der LDOM und deren Betriebssystem. Als funktionierende LDOM werden folgende Status 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 Betriebssystem 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 sinnvoll ist, Oracle-Cluster einzusetzen und, wie im Unternehmen des Autors, eigene Erweiterungen für eine bessere Überwachung zu implementieren, wird später beschrieben.

LDOMs im Oracle Solaris Cluster integrieren

Um eine Guest Domain vom Cluster überwachen zu können, muss diese zuerst manuell oder, falls vorhanden, mit einem Framework erstellt werden. Dabei ist es wichtig, dass alle virtuellen Services von Netzwerk- und Disk-Komponenten 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 in einen fehlerhaften Zustand kommen, muss die Master-Ausfall-Policy auf „reset“ gesetzt sein. Wenn dieser Fall eintritt, werden dadurch alle Slave Domains sofort beendet. Diese Konfiguration ist ebenfalls auf allen Cluster Nodes gleichzusetzen *(siehe Listing 1)*.

Im Gegenzug muss bei allen Guest Domains mit „#ldm set-domain master=primary g0078“ definiert werden, wer ihr Master ist, sonst wirkt die oben genannte Konfiguration nicht. Danach wird die LDOM unter Cluster-Kontrolle gebracht. Um die LDOM im Cluster nun hochverfügbar zu machen, muss zuerst der Cluster-Resource-Type „SUNW.LDOM“ im Cluster registriert werden. Anschließend wird die LDOM als „resource“-Gruppe erfasst, wie das Beispiel in *Listing 2* zeigt. In *Listing 3* sind die Cluster-Ressourcen zu sehen.

Hochverfügbarkeit

Die Firma JomaSoft in der Schweiz beschäftigt sich seit fünfzehn Jahren mit Oracle Solaris und SPARC. Eine selbst entwickelte Data-Center-Management-Software (VDCF) integriert und automatisiert jeweils die neuesten Features und Möglichkeiten, die diese Technologien anbieten. Für den größten ICT-Anbieter in der Schweiz wurde die LDOM-Technologie zusammen mit Oracle Solaris Cluster hochverfügbar gemacht, um damit den hohen Anforderungen der Kunden gerecht zu werden.

VDCF installiert vollautomatisiert physikalische Server als Cluster Nodes. Danach

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

Listing 1

```
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
```

Listing 2

```
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)
```

Listing 3

ist man in der Lage, mit wenigen Befehlen über das Framework LDOMs auf den Cluster Nodes zu erzeugen. Diese werden automatisch im Cluster registriert und sind somit sofort unter Cluster-Kontrolle, damit sie bei einem Ausfall 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 man auf einem Cluster-Mitglied eine LDOM erzeugt, integriert diese das Framework automatisch im Cluster und führt alle nötigen Konfigurationsschritte durch, damit alles richtig eingestellt ist und nichts vergessen wird. Auch bei einer Migration der LDOM kann das Framework unterscheiden, ob es die Kontrolle an den Cluster übergeben oder ob die LDOM vom Framework-internen Mechanismus transferiert werden muss.

Damit die Verfügbarkeit der LDOMs im Oracle Solaris Cluster noch verbessert werden konnte, ist das Monitoring erweitert. Zusätzlich kann man mit VDCF die LDOMs per Ping überwachen. Falls die Guest Domain damit nicht mehr erreichbar ist, wird geprüft, ob die LDOM Console noch verfügbar ist und erreicht werden kann. Wenn auch dies nicht mehr möglich ist, wird der Cluster angewiesen, die LDOM auf einem anderen Node neu zu starten.

Zudem lässt sich im VDCF konfigurieren, dass der ZPOOL-Failmode für Cluster LDOMs auf „panic“ gesetzt wird. Dies garantiert, dass die LDOM bei einem Fehler vom Zpool ebenfalls auf einem anderen Node neu gestartet wird. Das kann zum Beispiel der Fall sein, wenn ein Storage-Device im SAN nicht mehr erreichbar ist.

Als dritte Erweiterung wurde ein IPMP-Monitor implementiert, der die IP-Multi-Path-Gruppen in der LDOM überwacht. Sollten etwa beide Pfade einer Gruppe ausfallen, wird vom Monitor ein Komplettausfall registriert und ein Cluster Switch ausgelöst. Damit wird die LDOM auf einem anderen, funktionsfähigen Node neu gestartet und das Netzwerk steht wieder ordnungsgemäß zur Verfügung.

Marco Stadler
stadler@jomasoft.ch