Hochverfügbarkeit und virtualisierte Umgebungen

Hartmut Streppel
Oracle Deutschland B.V. & Co. KG
München

Schlüsselworte
Virtualisierung, Hochverfügbarkeit, Solaris Container, Solaris Zonen, Logical Domains, Oracle VM Server for SPARC, Oracle Solaris, Oracle Solaris Cluster

Einleitung
Virtualisierung ist en vogue. Lange wurde Virtualisierung als Allheilmittel der IT-Industrie beschrieben. Leider ist es so einfach nicht. Sie ist kein Selbstzweck und muss mit Bedacht eingesetzt werden. Die Frage, wie in virtualisierten Umgebungen Hochverfügbarkeit implementiert werden kann, wird häufig gar nicht oder zu spät gestellt. Oder es wird sogar die Möglichkeit, eine virtuelle Maschine im Betrieb, d.h. „live“ zu migrieren, als die Wunderwaffe für Hochverfügbarkeit betrachtet.
Dieses kurze Papier soll einen Überblick geben, wie im Oracle Solaris Umfeld Hochverfügbarkeit mit Solaris virtuellen Maschinen (Logical Domains) und mit Solaris Containern, bzw. Zonen implementiert werden kann.

Solaris Container/Zonen und Logical Domains
Solaris Container wurden in Solaris 10 eingeführt. In Solaris 11 werden sie nur noch als Zonen bezeichnet. Im Folgenden wird der Begriff Zone sowohl für Solaris 10 Container als auch für Solaris 11 Zonen verwendet. Zonen stellen eine sichere, in sich gekapselte Ablaufumgebung zur Verfügung, die einer Anwendung, aber auch einem Administrator den Eindruck vermittelt, eine vollständige Solaris Umgebung zur Verfügung zu haben. Mit Hilfe eines ausgefeilten Ressource Managements ist es möglich, die für den Betrieb einer Zone notwendigen Ressourcen exakt zuzuweisen. Damit ist nicht nur der saubere Betrieb der Zone gewährleistet, sondern auch die Sicherheit, dass aus dem Ruder laufende Anwendungen in einer Zone nicht die Ressourcen anderer Zonen beeinflussen können. Da Zonen direkten Zugriff auf Betriebssystem Ressourcen haben, vor allem auch auf das Netzwerk und Storage Systeme, und damit keinerlei Performancenachteile auftreten, ist es sinnvoll, Anwendungen grundsätzlich in Zonen zu kapseln.

Logical Domains (Ldom), d.h. virtuelle Maschinen, die mit Hilfe von Oracle VM Server for SPARC implementiert werden, stellen im Gegensatz zu den Containern eine vollständige Betriebssystemumgebungen zur Verfügung. Diese nutzen je nach Konfiguration virtuelle Devices für Netz- und Storage-Zugriff.
Innerhalb von Ldoms können nun wiederum Solaris Zonen konfiguriert werden. Solche kombinierten Umgebungen sind z.B. das Standard Deployment Modell im Oracle SuperCluster. Beide Technologien sind komplementär und schließen sich nicht gegenseitig aus.

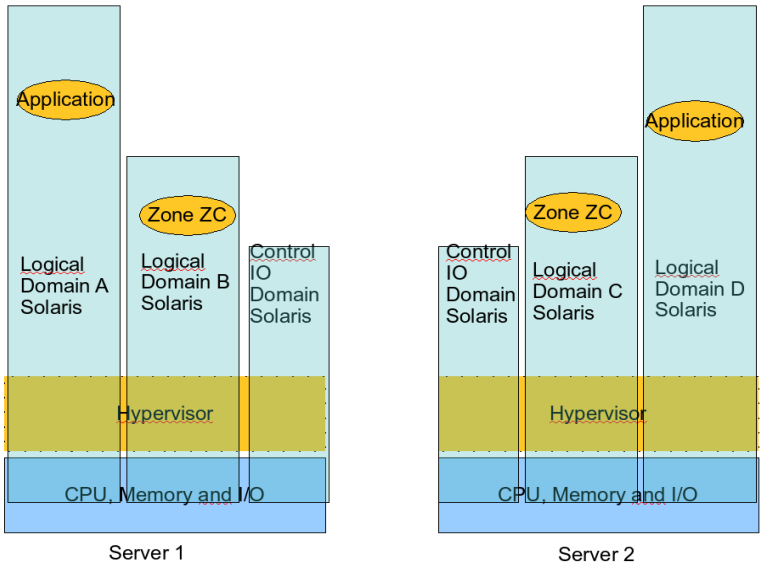

Abb. 1: Solaris Zonen und Logical Domains unter Oracle VM for SPARC

Hochverfügbarkeit

Die Verfügbarkeit eines Systems oder eines Dienstes wird als prozentualer Anteil an der möglichen Gesamtverfügbarkeit gemessen. Hochverfügbare Systeme haben typischerweise eine Verfügbarkeit von mindestens 99,9% pro Jahr, d.h. solche Systeme sind maximal 9h pro Jahr nicht verfügbar. Da SLAs in der Regel Verfügbarkeiten pro Monat festlegen, sprechen wir hier von einer maximalen Nichtverfügbarkeit von weniger als einer Stunde im Monat. Dies kann man nur mit voll redundanten Systemen und dem Einsatz von Cluster Software erreichen.

Virtuelle Umgebungen als Cluster-knoten

Oracle Solaris Cluster kann sowohl in Solaris Zonen als auch in Ldoms installiert werden. Damit werden virtuelle Umgebungen zu Cluster-knoten, die dann zu vollwertigen Clustern konfiguriert werden können. Aus Solaris Zonen werden dann Zonencluster und aus Ldoms „normale“ Cluster, deren Nutzung sich nicht von einem Cluster auf einer nicht virtualisierten Plattform unterscheidet.

Innerhalb der Cluster werden hochverfügbare Dienste mit den Standard Solaris Cluster Agenten verwaltet. Der Betrieb eines solchen Clusters, das aus virtuellen Umgebungen besteht, unterscheidet sich also kaum vom Betrieb eines Clusters, das direkt auf der Hardware läuft.

Ein grundsätzliches Problem, das den meisten Virtualisierungslösungen inhärent ist, ist die Nichtlinearität der Zeit in einer virtuellen Umgebung. D.h. die Systemzeit kann durchaus für mehrere Sekunden stillstehen, was für Cluster, die von einer konstant laufenden Systemuhr ausgehen, in der Regel ein Problem ist. Dieses besteht für Ldoms, die Oracle Solaris Cluster Knoten sind, auf Grund der Implementierung von Oracle VM Server for SPARC nicht. Bei Zonen als Cluster-knoten existiert das Problem grundsätzlich nicht, da sie dieselbe Uhr verwenden, die auch in der globalen Zone sichtbar ist.

Zonencluster

Ein Alleinstellungsmerkmal von Oracle Solaris Cluster sind Zonencluster – Cluster, die aus mehreren

Solaris Zonen bestehen. Diese brauchen ein Cluster in der globalen Zone, d.h. man kann nicht allein Oracle Solaris Cluster in Zonen installieren. Zonencluster sind virtuelle Cluster, die Dienste des „globalen“ Clusters nutzen. So werden der Cluster Interconnect, die Cluster Membership und das Quorum-Device ausschließlich im globalen Cluster verwaltet. Der Cluster Interconnect kann allerdings auch innerhalb eines Zonenclusters als Netz für Anwendungskommunikation mit Hilfe einer sog. *clprivnet* Adresse genutzt werden.

Konfiguriert werden Zonencluster mit Hilfe des *clzonecluster* Kommandos. Dieses installiert und konfiguriert die Zonen und die notwendigen Pakete. D.h. mit einem einzigen Kommando kann ein Zonencluster auf mehreren Cluster-knoten konfiguriert und installiert werden.

Devices und Resource Controls werden wie bei anderen Zonen auch konfiguriert.

Cluster mit Logical Domains

Logical Domains können ebenfalls als Cluster-knoten agieren. Dazu wird ganz normal die Oracle Solaris Cluster Software in einer LDOM installiert und konfiguriert. Innerhalb der Ldoms ist als einziger Unterschied zu bemerken, dass statt direkter Devices virtuelle genutzt werden. Bei der Konfiguration des Cluster-Interconnects, der typischerweise auch für mehrere Cluster-knoten auf demselben physischen System gemeinsam genutzt wird, ist dafür Sorge zu tragen, dass die vom Cluster verwendeten und von ihm selbst konfigurierten IP-Adressbereiche sich nicht überlappen oder alternativ durch VLANs sauber voneinander getrennt sind.

Um eine Priorisierung der Heartbeat Pakete, die das Cluster nutzt, um den Status anderer Cluster-knoten zu überprüfen, zu gewährleisten, muss in der IO-Domain die Property „*mode*“ für das virtuelle Netz für den Cluster Interconnect auf den Wert „*sc*“ gesetzt werden.

Virtuelle Umgebungen als Cluster Ressource

Eine Alternative zur Nutzung virtueller Umgebungen als Cluster-knoten ist deren Verwendung als eine Cluster Ressource. Eine virtuelle Umgebung wird als Cluster Ressource angelegt und dann von einem speziellen Cluster Agenten gestartet, gestoppt und überwacht. Dieser Agent betrachtet nur die virtuelle Umgebung als Ganzes, nicht aber Anwendungen oder Komponenten innerhalb der virtuellen Umgebung. Der Solaris Cluster Agent, der Zonen überwacht, ist allerdings in der Lage, mittels Shell oder SMF (Service Management Facility) auch direkt in die Zone zu greifen und dort Skripte auszuführen oder SMF Dienste zu starten, zu überwachen oder zu stoppen. Dieses technische Ansatz wird auch als „black box“ bezeichnet, da das Innere der virtuellen Umgebung dem Cluster-agenten verborgen bleibt.

Failover Zonen

Failover Zonen, auch als „Flying Container“ bekannt, sind Zonen, die als ganze Einheit zwischen Cluster-knoten geschwenkt werden können. Innerhalb der Zonen läuft keine Cluster-software. Anwendungen, die in den Zonen laufen, müssen deshalb mit anderen Mitteln hochverfügbar gemacht werden. Dies geschieht entweder über die Integration in SMF, wobei SMF auch eine Prozessüberwachung durchführt oder mit Hilfe der Skript- oder SMF Erweiterung des HA Zone/Container Agenten von Oracle Solaris Cluster.

Es ist wichtig zu verstehen, dass eine Failover Zone ein „single point of failure“ (SPOF) ist. Sollte eine Zone z.B. auf Grund einer fehlerhaften Patchaktion nicht mehr booten können, hilft es auch nicht, dass sie mit Hilfe des Clusters auf einen anderen Cluster-knoten verlagert werden kann. Da das Problem in der Zone selbst liegt, wird sie auf dem anderen Knoten sicherlich auch nicht booten.

Der wesentliche Vorteil in diesem Konzept, den viele Kunden sehen, ist, dass nur eine OS- und Anwendungsumgebung innerhalb der Zone verwaltet werden muss. Bei einem Cluster zwischen Zonen müssen alle auf demselben Stand gehalten werden, damit garantiert ist, dass Anwendungen auch auf allen Knoten problemlos laufen. Dies erfordert natürlich ein exzellentes Change Management.

Failover Ldoms

Genau wie bei Failover Zonen können auch Ldoms als Cluster Ressource konfiguriert werden. Um sie in Oracle Solaris Cluster zu integrieren, steht der Cluster Agent für Oracle VM for SPARC zur Verfügung. Dieser Agent startet, stoppt und überwacht Ldoms. Aber genau wie der Agent für die Zonen betrachtet er die LDOM als eine „black box“ und schaut nicht auf das, was in der virtuellen Maschine passiert. Hier muss also wieder eine zusätzliche Integration einer hochverfügbaren Anwendung in SMF durchgeführt werden, um automatisches Starten und Stoppen und eine Prozessüberwachung zu implementieren.

Wie bei den Failover Zonen ist zu bedenken, dass eine Failover LDOM ein SPOF ist. Kann eine LDOM auf Grund eines Fehlers in der LDOM nicht mehr gebootet werden, ist es unwahrscheinlich, dass sie auf einem anderen Cluster-knoten gebootet werden kann.

Um eine Failover LDOM in Oracle Solaris Cluster zu integrieren, muss das Cluster in den Control Domains installiert und konfiguriert sein. Da der Cluster Agent für die Failover Domains diese verwalten muss, benötigt er Zugriff auf den ldmd (Logical Domain Manager Daemon), der nur in der Control Domain läuft. Der ldmd würde auch in einer Konfiguration mit redundanten IO Domains nur in der Control Domain laufen.

Die Implementierung von Oracle VM Server for SPARC hat eine Besonderheit: der Ausfall der IO-Domain führt nicht automatisch dazu, dass Gast-Domains, die virtuelle Devices dieser IO-Domain nutzen, ebenfalls ausfallen. Stattdessen laufen diese weiter und warten darauf, dass die Devices wieder verfügbar werden. Dieses Verhalten ist in einer Cluster-Umgebung nicht gewünscht, da ja für den Fall der Nichtverfügbarkeit der Control Domain als Cluster-knoten ein Neustart der Failover Domains auf einem anderen Cluster-knoten erwartet wird. Dieser Neustart kann aber nur dann sicher durchgeführt werden, wenn die Failover Domain auf dem anderen System sicher nicht mehr läuft. Um dies zu garantieren, muss die LDOM Property *failure-policy* auf den Wert *reset* gesetzt werden. Damit wird sichergestellt, dass ein Ausfall der Control-Domain auch zum Ausfall der von ihr abhängigen Gastsysteme führt.

Live Migration virtueller Maschinen

Zum Schluss soll noch kurz darauf hingewiesen werden, dass Live Migration für virtuelle Maschinen nicht direkt als Hochverfügbarkeitsmechanismus betrachtet werden kann. Bei einer Live Migration wird – stark vereinfacht - im laufenden Betrieb der von der Domain verwendete Hauptspeicherbereich auf den Zielknoten kopiert, dann die laufende Domain gestoppt (suspended) und auf dem Zielknoten wieder aktiviert (resumed). Dies führt für den Anwender und die Anwendungen dazu, dass fast keine Serviceunterbrechung zu bemerken ist.

Mit Hilfe dieser Technologie können geplante Serviceunterbrechungen minimiert, aber nicht ganz beseitigt werden. Z.B. wird nach dem Einspielen von Updates die Domain fast immer neu gestartet werden müssen. Bei ungeplanten Service-Unterbrechungen hilft Live Migration überhaupt nicht. Der Ausfall eines Systems z.B. durch einen Hardwarefehler führt auch zu einem Ausfall der Domain, die dann vom Cluster auf einem anderen Cluster-knoten neu gestartet werden muss.

Es sollte an dieser Stelle noch darauf hingewiesen werden, dass es auf Grund der von Oracle Solaris Cluster (und auch anderen Cluster Frameworks) erwarteten Linearität der Zeit nicht supportet ist, eine Domain, die ein Solaris Cluster Knoten ist, mittels Live Migration zu verschieben.

Zusammenfassung

Hochverfügbarkeit im Umfeld von Virtualisierung im Solaris Umfeld wird realisiert mit Oracle Solaris Cluster. Es stehen zwei sehr unterschiedliche Modelle zur Verfügung, die sowohl für virtuelle Maschinen als auch für Zonen genutzt werden können: die Behandlung einer virtuellen Umgebung

- als sog. „Black Box“, d.h. die virtuelle Umgebung ist eine Cluster-Ressource, deren Status überwacht wird, oder
- als Cluster-knoten, d.h. innerhalb der virtuellen Umgebung läuft Oracle Solaris Cluster und überwacht hochverfügbare Anwendungen.

Damit steht eine vollständige Hochverfügbarkeitslösung für Virtualisierung im Solaris Umfeld zur Verfügung.

Mehr Informationen

- Oracle Solaris Cluster 3.x Dokumentation: http://docs.oracle.com/cd/E37745_01/
  - HA Oracle Solaris Containers: http://docs.oracle.com/cd/E18728_01/html/821-2677/
  - HA Oracle VM Server for SPARC: http://docs.oracle.com/cd/E18728_01/html/821-2904/
- Oracle Solaris Cluster 4.x Dokumentation: http://docs.oracle.com/cd/E29086_01/
  - HA Oracle Solaris Zones: http://docs.oracle.com/cd/E23623_01/html/E26828/
  - HA Oracle VM Server for SPARC: http://docs.oracle.com/cd/E23623_01/html/E25230/

Kontaktadresse:
Hartmut Streppel
Oracle Deutschland B.V. & Co. KG
Riesstr. 25
D-80992 München

Telefon: +49 (0) 89-1430 2588
Fax: +49 (0) 89-1430 1150
E-Mail Hartmut.Streppel@oracle.com
Internet: www.oracle.com