Ops Center 12c R2

OVM Server for SPARC Datacenter-Ready

Stefan Hinker & Elke Freymann
Oracle Deutschland B.V. & Co. KG
München

Schlüsselworte

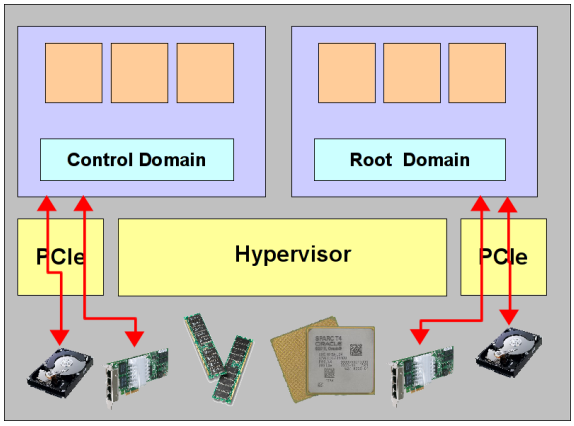
Enterprise Cloud Infrastructure, Oracle Enterprise Manager Ops Center, SPARC Virtualisierung

Einleitung

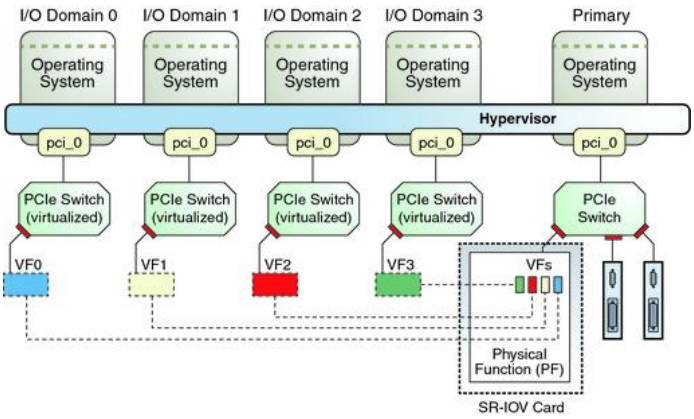
OVM Server for SPARC, besser bekannt als LDoms, stellt als effizienter Hypervisor die technische Grundlage für die Virtualisierung aktueller SPARC Systeme bereit. Damit ist es jedoch nicht getan, denn in welchem Rechenzentrum steht nur ein einzelner Server? Oracle Enterprise Manager Ops Center 12c Release 2 (im folgenden kurz Ops Center) stellt das Bindeglied zwischen RZ-Administrator und SPARC-Virtualisierung dar und ermöglicht die effiziente Verwaltung großer Serverpools.
Auf der Grundlage von Projekt- und Kundenerfahrungen beschreibt dieser Vortrag Best Practices zum Einsatz in mittleren und großen Umgebungen. Dabei werden die Neuerungen des Release 2 von Ops Center und OVM Server for SPARC 3.1.1 besonders beleuchtet.

Der SPARC Hypervisor – Effiziente Hardware-Partitionierung

Der SPARC Hypervisor wird, anders als die meisten anderen Lösungen, nicht als Software installiert, sondern ist ein in die Firmware integrierter Bestandteil der Hardware. Er ist immer aktiv, also auch bei einem nicht partitionierten (oder virtualisierten) System und kann nicht etwa ein- und ausgeschaltet werden. In einer schichtweisen Betrachtung des Systems befindet sich der Hypervisor zwischen der Hardware und dem OBP – oder besser den OBPs. Beim Start des Systems lädt der Hypervisor eine „Machine Description“ – eine Beschreibung der Hardware-Ausstattungen der verschiedenen Teilsysteme. Diese werden konfiguriert und für jede eine Kopie des OBP gestartet. Auf diese Weise entstehen ein oder mehrere Teilsysteme, die jeweils aus disjunkten Teilmengen der gesamten zur Verfügung stehenden Hardware bestehen. Im einfachsten Fall entstehen so einzelne Domains, die jeweils einen PCIe Root Complex (PCI Controller) sowie ein wenig CPU und Memory besitzen. Diese Domains, Root-Domains genannt, sind dann jeweils für sich lauffähig und voneinander vollständig unabhängig. Der Hypervisor hat in diesem Fall nach dem Starten der Konfiguration nichts weiter zu tun. Der Zugriff der Domains auf ihre jeweilige Hardware ist direkt und ohne Verluste durch die Virtualisierung. Auf diese Weise können so viele Domains erzeugt werden, wie es PCIe RootComplexe im System gibt.

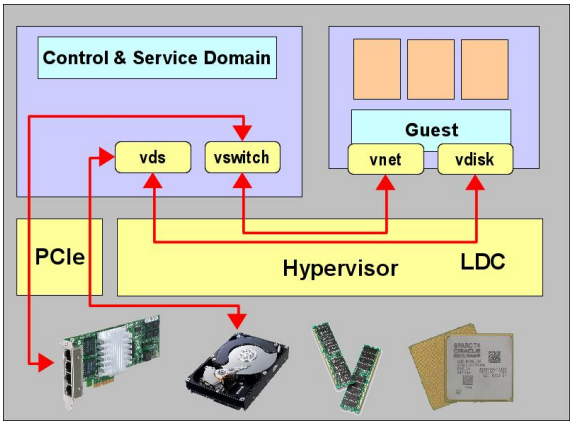


Um eine feinere IO-Granularität zu erreichen, als mittels Root-Domains realisierbar ist, gibt es zwei weitere Methoden, die jeweils PCIe-Virtualisierung verwenden: DirectIO und SR-IOV. Bei DirectIO wird einer Domain statt eines vollständigen RootComplexes nur ein (oder mehrere) PCIe Slots zugewiesen. Die notwendige PCIe Controller-Infrastruktur wird dabei von der Root-Domain betrieben und im Gast virtualisiert zur Verfügung gestellt. Der Gast kann nun die im Slot befindliche Karte exklusiv und ohne Verluste nutzen. DirectIO wurde ursprünglich eingeführt, um in Systemen mit relativ wenigen PCIe RootComplexen mehr Flexibilität zu erreichen. In heutigen Systemen mit idR. 1-3 Slots pro RootComplex und aktuell 2 RootComplexen pro CPU ist der Bedarf für diese Lösung stark zurück gegangen.



Die weitaus granularer arbeitende Lösung SR-IOV setzt auf der Virtualisierungsfähigkeit der PCIe-Karten selbst. Dabei wird in einem ersten Schritt die Karte angewiesen, eine bestimmte Anzahl virtueller Karten zu erzeugen. Diese erscheinen (nach einer Rekonfiguration des PCIe-Busses) auf dem Bus und können in einem zweiten Schritt einer Domain zugewiesen werden. Diese kann die virtuelle Karte dann ähnlich wie eine physische Karte verwenden, allerdings ist, je nach Kartentyp, evtl. nicht der volle Funktionsumfang verfügbar. Performance-Einschränkungen dieser Lösung hängen ebenfalls von der verwendeten Karte ab, sind aber idR. vernachlässigbar. Seit der Version 3.1.1 von Oracle VM Server for SPARC werden Karten für Infiniband, Ethernet und FibreChannel unterstützt.

Beide Varianten, DirectIO und SR-IOV, sind insofern eingeschränkt, als es eine Abhängigkeit von der Root-Domain gibt, die den physischen RootComplex besitzt und betreibt. Muss diese Domain booten oder aus anderen Gründen den PCIe-Bus resetten, pflanzt sich dieser Reset bis in die Karten fort und betrifft damit auch die Domains, die die Slots bzw. virtuellen Karten verwenden. Diese Abhängigkeit und die Reaktion auf einen Reset ist im LDom-System konfigurierbar, muss aber in jedem Fall beachtet werden.



Eine weitere Alternative zur direkten Hardware-Zuweisung ist voll-virtuelles IO, welches dann auch eine Live-Migration der Gastsysteme möglich macht. In diesem Fall werden vom Hypervisor bereitgestellte „Logical Domain Channels“ als Verbindungen zwischen virtuellen Disk- und Netzwerk-Treibern in den Gästen und entsprechenden Services in speziellen IO-Domains verwendet. Diese IO-Domains dienen als Vermittler zwischen physischer Hardware und den virtuellen Geräten, die den Gästen zur Verfügung gestellt werden. Unter Anderem kann dadurch redundante Hardware einfach einer Vielzahl von Gästen zur Verfügung gestellt werden. Anders als bei direktem Hardwarezugriff ist bei virtuellem IO, bedingt durch den „Umweg“ durch die IO-Domain, mit geringen Aufschlägen bei der Latenz eines

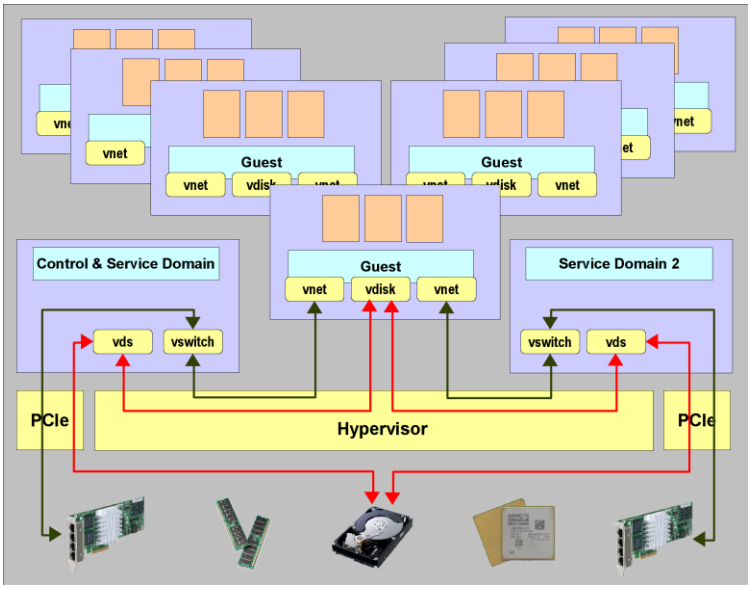
einzelnen Datenpakets zu rechnen. Neuere Verbesserungen im Hypervisor und bei den Treibern reduzieren diese Latenzen jedoch so deutlich, dass sie in den wenigsten Fällen eine praktische Rolle spielen. Der Durchsatz auch eines einzelnen LDC bzw. des damit realisierten Netzwerk- oder Disk-Geräts ist dagegen keinen praktischen Limitierungen unterworfen, da gemessene Grenzen höher liegen als der mögliche Durchsatz der physischen Geräte. Werden zwei IO-Domains redundant konfiguriert, ist ein hoch verfügbarer und weitestgehend unterbrechungsfreier Betrieb der Gäste möglich.

Best Practices für zwei Einsatz-Szenarien

LDoms werden häufig in einem von zwei Szenarien eingesetzt, wobei natürlich auch Mischformen möglich sind. Die am weitesten verbreitete Form ist das typische Konsolidierungs-Szenario mit einzelnen virtuellen Maschinen. Hierbei werden, aufbauend auf einer redundanten Infrastruktur, eine große Anzahl von Gästen auf einer Hardware betrieben. Diese Gäste sind dabei idR. eher klein bzw. anspruchslos in ihren Anforderungen an CPU, Memory und IO-Leistung. Dieses Szenario kommt häufig in Server-Refresh Projekten zum Einsatz um größere Mengen älterer Server abzulösen.

Die Empfehlungen für dieses Szenario sind:

- Zwei Redundante IO-Domains
- In den IO-Domains jeweils MPxIO für redundanten Anschluss an das SAN und LACP oder DLMP Aggregationen für redundanten Netzwerk-Zugang.
- In den IO-Domains Solaris 11.2
- Einen bis zwei CPU-Kerne in jeder IO-Domain.
- Die Gäste sind nach dem jeweiligen Bedarf auszustatten. Hier kann es keine festen Empfehlungen geben, da die Anforderungen sehr unterschiedlich sind.



Im zweiten Szenario, ebenfalls häufig für Konsolidierungszwecke verwendet, wird ein größerer Server, typischerweise mit 4 oder mehr CPUs, in wenige, gut ausgestattet Root-Domains aufgeteilt. In diesen wird dann entweder eine einzelne, anspruchsvollere Anwendung betrieben, oder sie dient ihrerseits als Konsoliderungsplattform für eine größere Anzahl von Solaris-Zonen. Prominentes Beispiel für diese Variante ist der Oracle SuperCluster, der genau nach diesem Model konfiguriert wird.

Die Empfehlungen für dieses Szenario sind:

- CPUs möglichst in ganzen Sockeln der Domain zuweisen.
- Mindestens einen, besser zwei Root Complexe pro Domain. Dabei diktieren die Anzahl der benötigten IO-Slots evtl. die Anzahl der Root Complexe. Nach Möglichkeit die CPUs verwenden, die die entsprechenden PCIe Root Complexe steuern.
- Memory möglichst lokal zuweisen, d.h. diejenigen Memory-Bereiche verwenden, die von der jeweiligen CPU gesteuert werden.

Mit diesen Empfehlungen werden die Latenzen beim Zugriff auf Memory und IO ideal reduziert. Da diese symetrische Partitionierung jedoch nicht immer die optimalen Bedingungen für die Anwendungen bietet (hoher CPU-Bedarf bei kleinem Memory-Footprint) sind andere Konfigurationen nicht zwingend schlechter. Rein funktional ist jede Kombination von Komponenten möglich. Insbesondere CPU und Memory können auch in einem Root-Domain Szenario jederzeit und während der Laufzeit des Betriebsystems zwischen einzelnen Domains verschoben werden.

Innerhalb der einzelnen Domains sind alle weiteren Funktionen von Solaris, also insbesondere Solaris Zonen und auch Kernel Zones verfügbar. Damit stellen Root-Domains ein ideales Mittel zur Aufteilung größerer Systeme dar, um die einzelnen Partitionen anschliessend zur Konsolidierung mit Zonen zu nutzen.

OpsCenter – One Tool to Bind Them All

Ops Center ist ein umfassendes, grafisches Managementtool, das den kompletten Lifecycle im Rechenzentrum unterstützt:

- Discovery und Inventarisierung bereits installierter Infrastruktur bzw. neuer Systeme
- Firmwareaktualisierung und Konfiguration
- Installation und Parametrisierung eines Betriebssystems (OS Provisionierung) bzw. eines Hypervisors und von Gast-Betriebssystemen (Virtualization Management)
- Patch-Management
- Überwachung der Systeme hinsichtlich Hardwarefehlermeldungen sowie OS-Monitoring.

Ops Center ist Teil der Oracle Enterprise Manager Produkt-Suite. Es kann dabei sowohl eigenständig installiert werden als auch mit Enterprise Manager Cloud Control gekoppelt werden.

Unter dem Gesichtspunkt „OVM Server for SPARC Datacenter-Ready“ sind natürlich die Funktionen, die Ops Center in der Sparte Vitualization Management bietet von besonderem Interesse. Ganz häufig wird zum Beispiel die Frage gestellt „auf welchem physischen Server läuft mein virtualisiertes Gastsystem nun eigentlich“. Aufgabe von Ops Center ist u.a. über die graphische Benutzerfläche einen klaren Überblick über die Verhältnisse zu geben, Diagramme zur Auslastung der Ressourcen und auch eine graphische Schnittstelle für Konfigurationsanpassungen zur Verfügung zu stellen.

Zielstellung für Ops Center ist, den kompletten Satz der Virtualisierungtechnologien, die Oracle im Umfeld Server bietet, möglichst vollständig zur unterstützen und diese Technologien auch als Schnittstelle für die Nutzung durch ein Cloud Management, das übergeordnet angesiedelt ist, bereit zu stellen.

Ein zentrales Konzept das Ops Center dabei on top zu den Technologien OVM Server for SPARC bzw. Solaris Zonen implementiert sind die so genannten Server Pools. In einem Server Pool werden mehrere Systeme, die hinsichtlich ihrer virtualisierten Gäste eine Einheit bilden sollen, zusammen gefasst.

Server Pools sind dann die Einheiten, gegen die sie so genannten Placement Policies, wie beispielsweise „wenn ein neuer Gast ausgerollt wird, platziere ihn auf dem Server mit der aktuell geringsten Auslastung“ definiert werden. Für Server Pools ist es auch möglich „Automatic Recovery“ zu nutzen: sollte beispielsweise die Hardware ausfallen, auf der eine Gast-VM gerade läuft, so startet Ops Center diesen Gast auf einem anderen Server des gleichen Pools automatisch gemäß der definierten Placement Policies neu und schaltet auch per ILOM die ausgefallene Maschine stromlos um Split-Brain-Szenarien zu vermeiden.

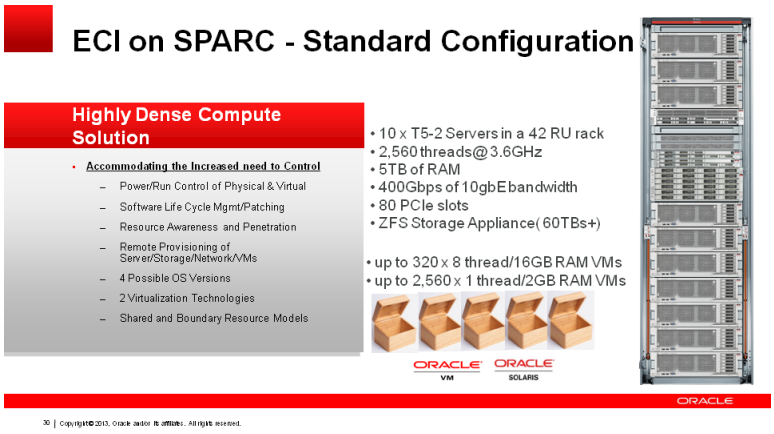
Wollte man bis dato dieses schöne graphische Managementwerkzeug Ops Center mit seinem Virtualization Management und den Server Pools für die Technologie OVM Server for SPARC voll umfänglich nutzen, so war man mit Einschränkungen hinsichtlich der Konfigurationsmöglichkeiten für die Domains konfrontiert. Insbesondere war es nicht möglich Systeme, die redundante IO Domainen betreiben voll ins Ops Center zu integrieren. Ebenso war es nicht möglich mit Ops Center Root Domain Systeme zu konfigurieren oder in die Verwaltung mit Ops Center auf zu nehmen. Das gleiche galt für Konfigurationen mit Direct IO und SR-IOV. Mit dem aktuellen Release von Ops Center sind diese Limitierungen entfallen. Ops Center kann nun so konfigurierte Systeme per discovery erkennen und nach ein paar zusätzlichen Verwaltungschritten auch managen. Außerdem ist es möglich, mit Ops Center auch solche Konfigurationen zu erstellen.

Enterprise Cloud Infrastructure – Wie aus den Teilen ein Ganzes wird

Auf der Grundlage von SPARC Servern, Oracle VM Server for SPARC, Oracle Solaris und Oracle Enterprise Manager Cloud Control und Ops Center ist es relativ einfach möglich, eine eigene Cloud-Umgebung im Sinne von Infrastructure as a Server (IaaS) aufzubauen. Das ist nichts neues und wird bereits in vielen Umgebungen in unterschiedlicher Ausprägung genutzt. Der Nachteil hierbei ist jedoch, dass ein nicht unerheblicher Aufwand betrieben werden muss, um die einzelnen Komponenten aufeinander abzustimmen und die gesamte Umgebung zu entwerfen und aufzubauen. Hier bietet Oracle mit der „Optimized Solution for Enterprise Cloud Infrastructure“ Abhilfe.

Die „Oracle Optimized Solutions“ sind eine Art Blaupause für mögliche Implementierungen mit Oracle Technologien. Das Portfolio deckt ein breites Spektrum von Enterprise Applications bis System Infrastructure ab. Dabei füllen die Optimized Solutions die Lücke zwischen den Oracle Engineered Systems als hochintegrierte Gesamtsysteme auf der einen, und Best of Breed Einzelkomponenten auf der anderen Seite. Bei „Enterprise Cloud Infrastructure“ handelt es sich um eine skalierbare Lösung zum Aufbau eine SPARC Cloud. Die verwendeten Komponenten sind u.A.:

- SPARC T5-2 Server mit Solaris und OVM Server for SPARC
- ZFS Storage Appliance
- 10GBit Ethernet oder optional Oracle Virtual Networking
- Enterprise Manager Cloud Control und Ops Center



Anders als bei Engineered Systems kann man Optimized Solutions nicht als Produkt bestellen. Sie dienen jedoch in Zusammenarbeit mit Oracle als Grundlage für eine direkte Umsetzung im Projekt. Die jeweilige Lösung ist dafür detailiert dokumentiert, u.A. mit vollständigen Teile-Listen und einem sehr ausführlichen Implementation Guide. Erfahrungen mit der Optimized Solution for Enterprise Cloud Infrastructure bei Kunden erbrachten einen Konsolidierungs-Faktor von ca. 15:1 und bis zu 2x bessere TCO und 5x schnelleres Deployment als bei selbstgebauten Lösungen. Damit erfüllt diese Lösung Oracles Anspruch, mit Entwicklung bei Oracle Entwicklungsaufwand und damit Kosten beim Kunden zu sparen und gleichzeitig optimal konfigurierte Umgebungen zu liefern.

Kontaktadresse:

Stefan Hinker & Elke Freymann

Oracle Deutschland B.V. & Co. KG

Riesstr. 25

D-80992 München

	Stefan Hinker	Elke Freymann
Telefon:	+49 211 7483-9848	+49 89 1430-2037
E-Mail	stefan.hinker@oracle.com	elke.freymann@oracle.com
Internet:	https://blogs.oracle.com/cmt	http://oracle.com/