Oracle VM: Der Aschenputtel Hypervisor?

Mag. Dr. Thomas Petrik
Sphinx IT Consulting
Wien

Schlüsselworte

Oracle VM, Hypervisor, Virtualisierung, High Availability, Cluster, Serverpool, Lizenz

Einleitung

Die Verbreitung von Oracle VM steigt seit einigen Jahren deutlich. Ein wesentlicher Grund dafür ist Oracles Lizenzpolitik, gemäß der das Hard Partitioning unter Oracle VM voll als solches im lizenzrechtlichen Sinn für alle Oracle Produkte anerkannt wird, während dies beim Marktführer VMware nicht der Fall ist. Aber auch die technische Entwicklung des auf XEN basierenden Hypervisors hat rasant Fahrt aufgenommen, nicht zuletzt deswegen, weil Oracle selbst in allen Appliances (Exadata, ODA, Exalogic, etc.) OVM als Hypervisor einsetzt, um den Kunden die Möglichkeit der Virtualisierung und Partitionierung auch auf diesen Systemen zu bieten.

Bei näherer Beschäftigung mit dem Produkt zeigt sich, dass alle grundlegenden Enterprise Features vorhanden sind und durchwegs zuverlässig funktionieren. Zudem sticht das eine oder andere Architekturkonzept heraus, das in dieser Form beim Mitbewerb gar nicht zu finden ist. Im produktiven Betrieb stellt sich Oracle VM als sehr stabile Alternative heraus.

Der Vortrag behandelt ausschließlich Oracle VM für x86, die meisten Aussagen bez. Architektur und Funktionalität haben aber auch in der SPARC Variante Gültigkeit.

Basiskomponenten

Oracle VM ist ein Hypervisor Type 1 (läuft also direkt auf der Hardware ohne Träger-OS) und basiert in der aktuellen Version 3.3 auf XEN 4.3 mit einem UEK 3 (entsprechend Oracle Enterprise Linux oder RedHat 6). XEN verwendet eine Dom0, die einen I/O-Proxy für die Kommunikation (Netzwerk und Disk I/O) beherbergt. Die Dom0 ist nicht auf eine fixe Zahl von CPUs beschränkt, wird aber durch Settings in der /etc/grub.conf bez. Memory limitiert. Der Memory-Bedarf der Dom0 steigt mit der Zahl der I/O-Threads und der Zahl der VMs (in Abhängigkeit des zugewiesenen RAM).

Auf Oracle VM lassen sich im Grunde alle x86-lauffähigen Betriebssysteme betreiben, insbesondere RedHat kompatible Varianten ab der Version 4, Solaris x86 ab Version 10, SUSE Linux ab Version 11 und Microsoft Windows Server 2003 R2 bis 2012 R2.

Die theoretischen Limits pro Server liegen jenseits dessen, was derzeit Hardware-seitig auf x86 verfügbar ist und mit bis zu 32 Knoten pro HA-Cluster und 2560 VMs sollten die meisten Installationen das Auslangen finden.

Genaue Daten sind den jeweiligen Release Notes zu entnehmen.

Eine wichtige Komponente stellen die 3 zur Verfügung stehenden Virtualisierungsvarianten dar: Hardware-Virtualisierung (HVM), Paravirtualisierung (PV) und Hardware-Virtualisierung mit PV-Treibern. Je nach eingesetzter Variante stehen eine unterschiedliche Anzahl von Disken und Netzwerkkarten in der VM zur Verfügung. Speziell im Windows Umfeld ist der Einsatz der SVVP zertifizierten Treiber dringend zu empfehlen, um die I/O-Performance drastisch zu verbessern.

Architektur

Es werden grundsätzlich 2 Varianten unterschieden: Single Server und Cluster, wobei im letzteren Fall zwischen HA und non-HA unterschieden werden kann. Die gesamte Server-Farm wird von einem OVM Manager verwaltet (eigene RedHat 6 basierte Maschine, physisch oder als VM). Cluster (Serverpools) benötigen ein Shared Storage (NFS, iSCSI oder FC), ein zusätzliches NFS-Repository für Backup- und Deployment-Zwecke ist empfehlenswert. Die Trennung der Netzwerkinfrastruktur in ein VM-LAN und ein Management-LAN (Traffic zwischen Manager und Knoten) ist optional aber sinnvoll. IPMI bzw. ILO Interfaces werden für alle Features benötigt, die der unmittelbaren Server-Steuerung bedürfen, z.B. Start Server, DPM oder Fencing im HA-Cluster.

Lokale Disken, Partitionen und LVM Volumes können nur vom jeweiligen Server eingebunden werden. Oracle VM bietet volle Unterstützung für Bonding und VLANs, die Konfigurationen erfolgt grundsätzlich über das Manager-GUI.

Der Manager selbst ist nicht erforderlich für den laufenden Betrieb sondern dient ausschließlich der Konfiguration und Überwachung. Die Kommunikation der einzelnen Knoten zum Manager erfolgt über eigene Agents in der Dom0, der Manager selbst basiert auf Weblogic mit einer MySQL Datenbank.

Enterprise Features

Mittels Live Migration können VMs im laufenden Betrieb zwischen 2 Clusterknoten verschoben werden. Es kann (und sollte) der Live Migration ein dediziertes Netzwerk zugeordnet werden, am besten mit 10 GBit Bandbreite. Voraussetzung sind ein Serverpool mit Shared Storage (nicht notwendigerweise ein HA-Pool) sowie das gleiche Instruction Set auf den beteiligten Knoten (zu diesem Zweck gibt es eigene Server Processor Compatibility Groups).

Achtung: Der Einsatz von Live Migration schließt CPU Pinning lizenzrechtlich (nicht technisch) aus.

DRS (Dynamic Resource Scheduling) nutzt Live Migration, um VMs automatisch zwischen den Knoten eines Serverpools zu verschieben, sobald ein definierbarer Threshold in Bezug auf CPU Last oder Network Utilization überschritten wird. Es können auch nur bestimmte Server eines Pools Teil dieser Policy sein.

DPM (Dynamic Power Management) nutzt Live Migration, um VMs von unterforderten Knoten wegzumigrieren und diese Server abzuschalten. Wird der eingestellte CPU oder Netzwerk-Threshold überschritten, werden Standby-Server wieder hochgefahren und die VMs zurück migriert.

Anti-Affinity Groups stellen sicher, dass bestimmte VMs nie auf dem gleichen Serverknoten gestartet werden.

Disken, Snapshots, Clones

In einem HA-Serverpool werden vorzugsweise iSCSI oder FC LUNs eingesetzt, während NFS-Anbindungen eher für shared Repositories zur Ablage von ISO-Images, Templates oder für Backup-Zwecke zum Einsatz kommen. Normalerweise legt OVM über diese LUNs ein OCFS2 Filesystem, auf dem schließlich die Images der Virtual Disks abgelegt werden. Alternativ können die LUNs aber auch direkt als Physical Disks zur VM durchgereicht werden.

Benchmarks zeigen allerdings, dass der Performancegewinn durch Verwendung von Physical Disks (Verzicht auf den OCFS2-Layer) minimal und in vielen Fällen gar nicht vorhanden ist (ein performantes Storage Subsystem vorausgesetzt). Zusätzlich entfällt die Möglichkeit Copy-On-Write (COW) Snapshots der VMs im laufenden Betrieb zu erzeugen. Der Erstellung solcher Thin Clones liegt im Sekundenbereich. Derart erzeugte Snapshot Clones sind vollwertige und eigenständige Maschinen, die sofort gestartet werden können.

Achtung: Aufgrund des COW Mechanismus kann es bei einer hohen Änderungsrate in der geklonten VM zu massivem I/O kommen, da alle geänderten 1 MB Cluster des OCFS2 neu geschrieben werden müssen.

Der Snapshot-Mechanismus kann auch für Backups eingesetzt werden. Im laufenden Betrieb wird ein Clone erzeugt auf ein NFS-Repository verschoben. Alternativ besteht die Möglichkeit, das OCFS2 Filesystem via NFS-Export für einen bestimmten NFS-Client freizugeben und den Clone mittels Backupsoftware wegzusichern.

Besonders interessant ist die Snapshot-Methode für Oracle DBAs. Von einer VM mit laufender Oracle DB wird ein Snapshot erzeugt. Der Clone wird gestartet, die darauf befindliche DB macht ein Crash Recovery und kann – sofern es sich um eine EE handelt und Flashback aktiviert war – mittels "flashback database" zurückgesetzt werden, ohne die Originaldatenbank zu beeinträchtigen. Aufgrund der zugrunde liegenden COW-Technologie beansprucht der Snapshot keinen zusätzlichen Platz. Erst durch Änderung in der weiter laufenden Originaldatenbank werden neue Blöcke (bzw. Cluster) geschrieben.

Lizenz und Support

Oracle VM Server ist Open Source und daher kostenlos. Der OVM Manager ist Closed Source aber kostenfrei. Optional kann ein kostenpflichtiger Support gekauft werden. Im Unterschied zu VMware sind alle Oracle Produkte zertifiziert und supported, Hard Partitioning (CPU Pinning, Server Pinning) wird lizenzrechtlich anerkannt. Problematischer sieht es bei anderen Herstellern speziell im SW-Appliance Bereich aus: da ist OVM meist unbekannt und steht zumindest nicht von Haus aus auf den Zertifizierungslisten.

Beim CPU-Pinning ist jedenfalls darauf zu achten, dass unterschiedliche Cores und nicht unterschiedliche Threads der gleichen Cores zugewiesen werden. Benchmarks zeigen, dass das Pinnen auf Cores unterschiedlicher CPU-Sockel weitgehend ohne Einfluss bleibt (minimale NUMA Effekte).

Performance in virtuellen Umgebungen

Benchmark-Vergleiche zwischen unterschiedlichen Hypervisoren (auch unter Verwendung unterschiedlicher Betreibssysteme) und Bare Metal Installationen belegen, dass Oracle VM den Vergleich mit dem kommerziellen Mitbewerb nicht scheuen muss. Die Einbußen der Virtualisierung gegenüber einer Bare Metal Installation sind klar messbar aber nicht dramatisch. Real World Messungen zeigen mitunter sogar drastische Beschleunigungen durch Virtualisierung.

DR-Lösungen

Der klassische Ansatz einer Desaster Recovery Lösung über 2 Standorte sieht den Einsatz zweier Serverpools vor, wobei der eine im Standby Modus betrieben wird. Die Replikation der Daten (VM Disks) übernimmt in diesem Szenario das Storage.

Alternativ besteht die Möglichkeit, mittels DRBD einen active/active Cluster aufzubauen, sodass beide Knoten mit lokalen Disken arbeiten können, diese aber zu OVM als shared Disk durchgereicht werden. DRBD Volumes werden out-of-the-box nicht als shared Device erkannt, dazu ist einiges an Customizing und geschickter Assemblierung erforderlich.

Marktposition

Ein Vergleich der Gartner Quadranten zeigt, dass sich Oracle VM – obwohl weit abgeschlagen hinter VMware und Microsoft Hyper-V – an vorderster Front der Verfolger liegt. Die leichte Rückstufung aus dem linken oberen Quadranten (im Vergleich zu 2014) wird mit der zu wenig offenen Ausrichtung

des Produkts begründet. Gartner hält ebenfalls fest, dass als häufigstes Argument für Oracle VM die Möglichkeit des Hard Partitioning gebracht wird.

Zusammenfassung

Oracle VM – in einer durchdachten Infrastruktur eingesetzt – stellt eine stabile und technisch durchaus interessante Alternative zu kommerziellen Produkten dar, wobei betont werden muss, dass sich sich um ein Multipurpose System handelt, das nicht auf die Oracle Produktwelt beschränkt ist. Verbesserungspotential ist vorhanden, speziell beim GUI Design oder bei der Storage-Verwaltung (Migrationen von einem Storage auf ein anderes sind in VMware deutlich komfortabler), aber die derzeit halbjährlichen Release-Zyklen zeigen die Konstanz in der Weiterentwicklung. Die weitere Integration in Open Stack wird mit Spannung erwartet und birgt enormes Potential für zukünftige Einsatzgebiete.

Kontaktadresse:

Mag. Dr. Thomas Petrik
Sphinx IT Consulting
Aspernbrückengasse 2
A-1020 Wien

Telefon: +43 664 155 8304
Fax: +43 (1) 599 31-99
E-Mail Thomas.Petrik@sphinx.at
Internet: www.sphinx.at