Noch eine "ODA"? Nein - viel besser...

Stefan Hinker
Oracle Deutschland B.V. & Co. KG
Düsseldorf

Schlüsselworte
MiniCluster, Datenbank, Sicherheit, PCI-DSS, FIPS, Software in Silicon, InMemory, Flash, Performance

Einleitung
Noch ein Vortrag mit Produktwerbung? Eher am Rande. Viel mehr soll es hier um all die Features gehen, die man sich bisher in Engineered Systems – und auch in herkömmlichen Systemen – mühsam selbst zusammen konfigurieren musste, die man nun mit wenig Aufwand direkt nach dem Auspacken des Systems nutzen kann. Welches andere System bietet bei der Erstkonfiguration nur die Wahl zwischen „PCI-DSS"-Konform und „DISA STIG"-Konform und macht das dann einfach, ohne dass man sich um die vielen Details kümmern müsste?

Was ist in der Packung?

MiniCluster ist als „kleiner Bruder" des SuperCluster konzipiert. Die Einsatz-Szenarien sehen daher, grob skizziert, den Betrieb von Datenbanken und optional Anwendungs-Umgebungen vor. Entsprechend der Hardware-Ausstattung ist MiniCluster dabei eher für den Einsatz in kleineren Umgebungen gedacht. Gerade dort, also in Betrieben mit eher kleinen Betriebsmannschaften, ist Zeit und KnowHow ein teures und knappes Gut. Das Design des MiniCluster hat zum Ziel, diese beiden Ressourcen zu schonen, ohne dabei Kompromisse bei Sicherheit oder Leistungsfähigkeit zu machen.

Was ist also in der Packung? Kurz gesagt: Zwei SPARC S7-2 Server und ein Plattengehäuse. Die Server sind mit jeweils 16 SPARC Kernen und 512GB RAM ausgestattet. Das Plattengehäuse enthält SSDs und SAS-Platten für eine Netto-Kapazität von ca. 7.2 TB Datenbankspeicher (mit Normal Redundancy) und ca. 18TB Platz auf einem hochverfügbaren NFS. Die genauen Details stehen im Datenblatt[1].

Auf den ersten Blick sieht das natürlich ganz ähnlich aus wie eine ODA. Aber wer genauer hinsieht erkennt Unterschiede aus denen sich verschiedene Einsatzzwecke ableiten lassen:

Feature	MiniCluster	ODA X5-2	ODA X6-2M
CPU Kerne	2x16	2x36	2x20
Memory pro Kern	32 GB	7 GB bis 21 GB	12 GB bis 38 GB
RAC Datenbank möglich	Ja	Ja	Nein

1 https://www.oracle.com/us/products/servers-storage/servers/sparc/minicluster-ds-3050222.pdf

Feature	MiniCluster	ODA X5-2	ODA X6-2M
Oracle SE2 support	in Vorbereitung	Nein	Ja
Datenbank und Redo auf Flash	Ja	Nein	Ja
Flash Kapazität (Normal Redundancy)	7.2TB	-	3.2TB bis 6.4TB
Vorbereitet für PCI-DSS oder CIS/STIG Zertifizierung	Ja	Nein	Nein
Software in Silicon (DAX/Silicon Secured Memory)	Ja	Nein	Nein
Hardware Crypto Support	Ja	nur AES[2]	nur AES[2]

So geht MiniCluster beispielsweise, anders als ODA X5 und X6, keine Kompromisse zwischen Performance und Hochverfügbarkeit ein – die Datenbank wird auf redundant angeschlossenem Flash gespeichert und somit ein Betrieb mit RAC möglich. Trotzdem bietet er, anders als ODA X5 auch die Möglichkeit, nur nach Standard Edition 2 zu lizenzieren, schränkt hier also nicht bereits durch die Wahl der Hardware die Flexibilität der Lizenzierung ein[3].

Das Betriebskonzept

Eines der Designziele des MiniCluster war es, einen Betrieb vollständig über den Browser zu ermöglichen. Dies ist, mit der Ausnahme eines einzelnen Schrittes während der Erstinstallation, gelungen. Die Installation verschiedener Datenbank-VMs sowie der Datenbank-Software (DB Home) und einzelner Datenbanken ist ebenso im Browser möglich wie die Verwaltung der System-Benutzer oder das Patchen des gesamten Systems. Damit wird MiniCluster u.A. für diejenigen interessant, die sich nicht mit den Details der manuellen und oft mühsamen oder gar fehlerträchtigen Installation von RAC-Datenbanken befassen wollen. Der Preis dafür ist natürlich eine standardisierte Installation mit nur minimaler Auswahl an Optionen. Dafür bekommt man jedoch innerhalb kürzester Zeit eine laufende RAC-Datenbank, ohne je eine Kommandozeile gesehen zu haben.

Entsprechend wenig Kenntnisse setzt MiniCluster für den Betrieb voraus. Natürlich muss man weiterhin wissen, was SCAN-Namen sind und die dafür notwendigen IP-Adressen zur Verfügung stellen. Auch einige weitere, essentielle Datenbank-Parameter wie SGA und PGA Größen und Zeichensätze werden natürlich abgefragt. Und wer eine RAC-Datenbank betreiben möchte, sollte sich auch damit auskennen. Es entfällt jedoch bspw. die Auseinandersetzung mit den Voraussetzungen des Betriebssystems und den notwendigen Einstellungen für Shared Memory oder dem RAC Interconnect. All das wird zuverlässig von der Automatik, dem „MiniCluster Virtual Assistant“, übernommen – das Betriebssystem selbst bleibt vollständig im Hintergrund. MiniCluster ist ein SPARC System, das die

2 AES-NI auf Intel CPUs ist deutlich weniger leistungsfähig als die SPARC Crypto-Beschleunigung. Siehe hierzu bspw. https://blogs.oracle.com/BestPerf/entry/20160315_tde_t7_1

3 In Vorbereitung, ein genaues Datum für die Verfügbarkeit ist noch nicht bekannt.

neuen Features wie Hardware-Support für Dekompression und InMemory Operationen der Datenbank voll unterstützt. Dennoch ist MiniCluster durch das konsequent umgesetzte Betriebskonzept vermutlich das erste System, das die Möglichkeiten von SPARC und Solaris einem Datenbank-Administrator zugänglich macht, der bisher stets unter Windows gearbeitet hat.

Bei all dieser Orientierung an Bedienbarkeit im Browser wurde jedoch nicht auf ein gleichwertiges Kommandozeilen-Interface verzichtet. Alle Funktionen des Browser-Interfaces sind auch mit einem speziellen „mcmu“ Kommando (**M**ini**C**luster **M**aintenance **U**tility) verfügbar.

Sicherheit – Eingebaut, nicht angeflanscht

Wie beim Betrieb geht MiniCluster auch beim Thema Sicherheit einige Schritte weiter als bisherige Systeme. Ziel der Entwicklung war es, das System so auszuliefern dass es nach der Installation ohne wesentlichen Zusatzaufwand einen Audit für PCI-DSS oder CIS[4] besteht. Viel dieser Arbeit ist für den Benutzer nicht sichtbar. So werden z.B. die meisten Dateisysteme verschlüsselt, Rechte an die verschiedenen System-Benutzern nur nach dem Least Privilege Prinzip vergeben und ein systemweites Auditing aller Aktionen vorkonfiguriert. Dabei geht die Konfiguration teilweise weit über das hinaus, was ein normales Solaris „Out of the Box“ liefert, indem es z.B. strenge Passwort-Regeln umsetzt und das Anlegen neuer System-Benutzer automatisch durch einen System-Supervisor genehmigen lässt. Der Benutzer hat zu Beginn der Installation die Wahl, ob das System nach PCI-DSS oder nach CIS auditierbar sein soll. Die gleiche Auswahl kann man später für die einzelnen Anwendungs- bzw. Datenbank VMs treffen. Die Option „nicht sicher“ gibt es dagegen nicht. Je nach Auswahl wird das System dann entsprechend abgesichert. Für den Audit kann man diese Absicherung dokumentieren lassen – im Browser-Interface werden Compliance Berichte für beide Security Benchmarks angeboten. Diese kann man auf Wunsch auch regelmäßig wiederholen lassen um so kontinuierliche Compliance zu dokumentieren.

In einem „MiniCluster Security Guide“[5] werden einerseits diese Einstellungen beschrieben, andererseits auch gezeigt, wie man darüber hinausgehende Absicherungen implementieren kann. Hier ist z.B. die Konfiguration der Plattform oder der VMs als „Immutable Zones“ oder die Einstellungen für FIPS 140-2 Compliance beschrieben. Selbstverständlich sind insbesondere im Bereich Datenbank-Sicherheit weitere Features wie bspw. Oracle TDE oder Database Vault möglich und unterstützt. Allerdings gehen diese über die Funktion der integrierten Benutzeroberfläche hinaus.

Storage-Infrastruktur

Das Plattengehäuse des MiniCluster ist redundant mit beiden Rechenknoten verbunden und liefert so die Grundlage für die Hochverfügbarkeit des Speichers. Die vorhandenen „Festplatten“ lassen sich in drei Gruppen einteilen:

1. 14x 1.6 TB Flash-Speicher zur Verwendung als Datenbankspeicher
2. 4x 200 GB Flash-Speicher für die Redo-Logs der Datenbanken
3. 6x 8 TB Festplatten für ein hochverfügbares NFS

Die gesamte Konfiguration dieses Speichers wird während der Erstinstallation vorgenommen, der Benutzer muss sich darum also nicht kümmern. Für die Datenbanken steht anschließend eine

---

4 https://benchmarks.cisecurity.org/

5 https://docs.oracle.com/cd/E69469_01/html/E69475/index.html

Clusterware-Umgebung zur Verfügung, die mittels ASM den Flash-Speicher für die Datenbank verwaltet. Hier hat man bei der Erstinstallation die Wahl zwischen Normal oder High Redundancy. Dabei ist es unerheblich, ob die Datenbanken RAC-, RAC-OneNode oder Single-Instance Datenbanken sind.

Das NFS wird ebenfalls vorkonfiguriert. Dazu werden pro Rechenknoten zwei Kerne für je eine „Infrastruktur-VM“ reserviert, die die Platten als hochverfügbares NFS zur Verfügung stellen. Dabei kommt aus Sicherheitsgründen ausschließlich NFSv4 zum Einsatz. Der Zugang zu diesem NFS kann den VMs vom Administrator entsprechend zugewiesen oder auch wieder entzogen werden.

Damit bietet MiniCluster funktional eine ähnliche Speicherarchitektur wie SuperCluster, allerdings entsprechend der sehr viel einfacheren Hardware natürlich auf weit einfacherem Niveau. Was im SuperCluster die Storage Server unter Verwaltung von ASM sind, ist im MiniCluster ein einfaches ASM auf Shared Storage. Das NFS des MiniCluster bietet ein überall verfügbares NFS, wenn auch nicht so flexibel wie die ZFS Appliance des SuperCluster.

Performance

In Sachen Performance muss sich der MiniCluster nicht verstecken. Zwar ist die Ausstattung mit CPU-Kernen und Memory nicht so üppig wie beim großen Bruder SuperCluster. Aber diese Ausstattung ist sowohl im Verhältnis CPU-Kerne zu Memory als auch bei IO-Bandbreite und Latenz gut ausbalanciert und verspricht so hohe Leistung und Effizienz.

Für jeden Datenbank-Server ist natürlich die IO-Performance ein wesentlicher Faktor. Da MiniCluster hier ausschließlich auf Flash setzt ist mit entsprechend hoher Leistung zu rechnen. Ein kurzer, sicher nicht alle Aspekte betrachtender Test mit der Datenbank-Funktion „calibrate IO“ bestätigt diese Annahme:

```
SET SERVEROUTPUT ON
DECLARE
lat INTEGER;
iops INTEGER;
mbps INTEGER;
BEGIN
DBMS_RESOURCE_MANAGER.CALIBRATE_IO (14, 10, iops, mbps, lat);
DBMS_OUTPUT.PUT_LINE ('max_iops = ' || iops);
DBMS_OUTPUT.PUT_LINE ('latency = ' || lat);
dbms_output.put_line('max_mbps = ' || mbps);
end;
/

max_iops = 647985
latency = 0
max_mbps = 8694
```

Beispiel für Calibrate IO auf MiniCluster

Die unter realen Bedingungen erreichbare Performance weicht natürlich je nach Anwendung hiervon ab. Mit den hier gemessenen ca. 650k IOPS und 8.5 GB/sec sollte das IO-Subsystem jedoch ausreichend Leistung mitbringen, um die vorhandenen CPUs in den allermeisten Fällen ohne Engpässe mit Daten zu versorgen. Und genau das ist ja das Ziel eines ausbalancierten Designs.

Besonders interessant ist der Blick auf die CPUs und hier insbesondere zwei Aspekte. Einerseits handelt es sich um die neuesten SPARC Kerne mit immerhin 4.27 GHz Taktrate und 16 MB L3 Cache. Hier lohnt sich ein Blick auf den Vergleich mit aktuellen Intel-CPUs. Andererseits bringt diese CPU, genau wie die M7 CPU, die neuen Software-in-Silicon Features mit, die einige Datenbank-Operationen signifikant beschleunigen können.

Allgemeine CPU-Leistung

Häufig werden zum Performance-Vergleich von CPUs kleine Micro-Benchmarks verwendet. Diese sind schnell und einfach auszuführen, beschreiben jedoch leider meist nur einen einzelnen Aspekt der CPU – z.B. das Erzeugen von Zufallszahlen, das Ziehen von Wurzeln oder das Sortieren eines Arrays im Speicher. Entsprechend schwach ist die Aussagekraft und Übertragbarkeit der Ergebnisse. Mehr hierzu bspw. auf meinem Blog[6]. Sinnvoller, aber entsprechend aufwändiger ist der Vergleich mittels komplexer Anwendungs-Benchmarks. Das Problem hierbei ist oft die Verfügbarkeit von Vergleichsmessungen, da nicht alle Hardware-Hersteller immer alle Benchmarks mit jeder CPU veröffentlichen. Im Fall der M7 und S7 CPU hat Oracle jedoch selbst Vergleiche mit x86 CPUs veröffentlicht, auf die man zurück greifen kann.

Allgemeine Anwendungsleistung wird dabei mit den Benchmarks SPECjbb2015 und SPECjEnterprise2010 verglichen. Während SPECjbb2015 ein reiner Java-Benchmark ist, hat SPECjEnterprise2010 eine Datenbank-Komponente, so dass hier eine 2-Tier Architektur simuliert wird. Die genauen Ergebnisse für diese Benchmarks hat Oracle auf seiner Webseite veröffentlicht[7]. Hier nur die für den Vergleich wesentlichen Daten:

Benchmark	SPARC Ergebnis	x86 Ergebnis	SPARC Vorteil pro Kern
SPECjEnterprise2010 (S7-2 vs. E5-2699 v4)	882 EjOPS verschlüsselt (55 pro Kern)	631 EjOPS klartext (14.3 pro Kern)	3.8x
SPECjbb2015 (S7-2 vs. E5-2699 v4)	36922 critical-jOPS (2308 pro Kern)	55858 critical-jOPS (1270 pro Kern)	1.8x

Diese Ergebnisse zeigen einen klaren Performance-Vorteil für die SPARC CPU – pro Kern, wohlgemerkt. Bemerkenswert ist im Ergebnis des SPECjEnterprise2010, da das SPARC Ergebnis mit Verschlüsselung, das der x86 CPU ohne Verschlüsselung ermittelt wurde. Hier zeigt die SPARC CPU, welchen Vorteil die vollständige Crypto-Unterstützung in der CPU im Härtetest bringt.

InMemory Performance

Während diese beiden Benchmarks die „klassische“ CPU-Leistung vergleichen, zeigen die beiden folgenden den Vorteil, den die SPARC Hardware-Unterstützung für die InMemory-Operationen der Oracle Datenbank bringt. Hierfür gibt es bisher keine geeigneten Standard-Benchmarks, weswegen Oracle auf eigene Vergleiche zurückgreifen musste. Dennoch sind diese nicht weniger beachtlich. Wieder sind die Details auf den Webseiten von Oracle nachzulesen[8].

---

6 https://blogs.oracle.com/cmt/entry/ein_paar_gedanken_zu_single

7 SPECjEnterprise2010: https://blogs.oracle.com/BestPerf/entry/20160629_jent_sparc_s7_2
SPECjbb2015: https://blogs.oracle.com/BestPerf/entry/20160629_jbb_sparc_s7_2

Benchmark	SPARC Ergebnis	x86 Ergebnis	SPARC Vorteil pro Kern
RCDB In-Memory (S7-2 vs. E5-2699 v4)	205 qpm (12.8 pro Kern)	73 qpm (1.66 pro Kern)	7.7x
Real-Time Enterprise (S7-2 vs. E5-2699 v3)	195,790 TPS (12237 pro Kern)	216,302 TPS (6008 pro Kern)	2x

Die Ergebnisse sind durchaus beeindruckend, insb. mit der sogenannten „Real Cardinality Database“, einer Test-Datenbank deren synthetisch erzeugte Werte innerhalb einer Spalte so verteilt sind, dass diese Verteilung denen von realen Datenbanken nahe kommt. Bei diesen Tests kommen die beiden Hardware-Features „Dekompression“ und „Suchen im Memory“ bzw. „Suchen in komprimiertem Memory“ vermehrt zum Tragen, was den erheblichen Performance-Vorteil gegenüber der x86-CPU erklärt. Die entsprechenden Hardware-Einheiten können hierbei den Hauptspeicher mit bis zu 120 GB/sec durchsuchen. So geben diese beiden Benchmarks einen Eindruck der möglichen Leistung der S7 CPU in einem (kleinen) Datawarehouse, wie es auch auf dem MiniCluster mit immerhin 512GB RAM pro Knoten realisierbar ist.

Betrachtet man diese Leistungswerte gemeinsam, ergibt sich durchgehend ein Performance-Vorteil der SPARC-Kerne von 1.5x oder besser gegenüber aktuellen x86-CPUs. Dieser Performance-Vorteil alleine ist jedoch nicht ausschlaggebend für einen System-Vergleich, bspw. mit der ODA. Für eine Betrachtung der Gesamtkosten, bspw. für ein kleines, hochverfügbares DWH, muss man die notwendigen Lizenzen für die Enterprise Edition und RAC hinzurechnen. Würden, unter Vernachlässigung der IO-Leistung bei der ODA 2x22 Kerne (also nur ein Teil der maximal 2x36 Kerne) für die zu erwartende Last benötigt, wären dies (mit dem Faktor 1.5 gerechnet) nur etwas weniger als 2x15 Kerne bei MiniCluster. Man braucht also (bei Core Faktor von jeweils 0.5) mindestens 6 EE/RAC Lizenzen weniger, wenn man das DWH auf MiniCluster betreibt – womit die Hardware-Kosten des Systems allein durch Einsparungen in der Lizenzierung schon weitestgehend gedeckt sein dürften. Weitere Einsparungen durch DAX, das stärkere IO-System sowie den sehr einfachen Betrieb kommen eventuell noch dazu.

Eine Demonstration der MiniCluster Benutzeroberfläche wird es im Demo-Kino am 17.11. um 12:00 Uhr geben. Einige Demo-Videos finden sich außerdem auf YouTube[9].

Kontaktadresse:
Stefan Hinker
Oracle Deutschland B.V. & Co. KG
Hamborner Str. 51
D-40472 Düsseldorf

Telefon: +49 211 7483 9848
E-Mail Stefan.Hinker@Oracle.com
Internet: https://blogs.oracle.com/cmt

---

8 RCDB In-Memory: https://blogs.oracle.com/BestPerf/entry/20160629_imdb_sparc_s7_2
Real-Time Enterprise: https://blogs.oracle.com/BestPerf/entry/20160629_rte_sparc_s7_2

9 https://www.youtube.com/user/michaelpalmeter