Flash und Disk Storage – die Mischung macht‘s

Franz Haberhauer
Oracle Deutschland B.V. & Co KG
Stuttgart

Schlüsselworte

Flash Memory, SSD, HDD, Hybrid Storage Pool, Smart Flash Cache, Oracle FS1 Flash Storage System, Storage Quality of Service.

Einleitung

In den letzten Jahren haben sich Flash-basierte Speicher als neue Komponente in der Speicherhierarchie etabliert, mit der sich hohe IO-Raten zu attraktiven Kosten realisieren lassen. Dabei unterscheiden sich Flash-basierte Speicher in ihren Eigenschaften signifikanter voneinander als man es von Festplatten gewohnt ist. Hohe Speicherkapazitäten mit geringen Leistungsanforderungen lassen sich aber weiterhin am preisgünstigsten mit Festplatten realisieren.

Spezielle Features in Betriebssystemen, Datenbanken und Middleware ermöglichen eine effiziente Kombination dieser Vorteile, etwa in den hybriden Storage Pools von ZFS oder dem Smart Flash Cache in der Oracle Datenbank.

Bei der Oracle OpenWorld 2014 wurde das Oracle FS1 Flash Storge System vorgestellt, in dem als Speichermedien unterschiedliche Technologien kombiniert werden: Performance- bzw. kapazitäts-orientierte Flash-Speicher sowie Performance- bzw. kapazitätsorientierte Festplatten. Über ein fein granuliertes Auto-Tiering können Daten automatisch auf die jeweils geeignetste Speichertechnologie verschoben werden. Spezifisch für Anwendungen und sogar für Datenbereiche innerhalb einzelner Anwendungen (etwa Datafiles, Redologs, Controlfiles in der Oracle DB) kann ein Storage Quality of Service definiert werden. Dabei können „wichtige“ Anwendungen priorisiert und so eine hohe Auslastungen des Speichersystemen bei dennoch guter Performance erreicht werden.

Der Vortrag gibt einen aktuellen Überblick über Flash im Hardware- und Software-Portfolio von Oracle, zeigt die praktische Nutzung in konkreten Architekturen und vergleicht alternative Ansätze mit unterschiedlichen Produkten.

Flash Memory als Speichertechnologie

Flash Memory sind Halbleiter-Speicher, die eine nichtflüchtige Speicherung bei niedrigem Energieverbrauch bieten sowie einen wahlfreien Zugriff mit niedriger Latenz bei gleichzeitig hohem Durchsatz. Die traditionellen Schwächen der Flash-Komponenten der Frühzeit - langsames Überschreiben von Speicherbereichen sowie eine beschränkte Anzahl von Löschzyklen je Speicherzelle (je nach Technologie zwischen 1k und 100k), woraus bei Hotspots eine begrenzte Lebensdauer von Speicherelementen resultieren konnte - sind inzwischen in der Architektur der Controller, die heute jeder Flashspeicher mitbringt, adressiert. Flash wird mit schnellen DRAM-basierten Schreibpuffern kombiniert, wobei durch Kondensatoren sichergestellt wird, dass deren Speicherinhalte auch bei einem Stromausfall noch in Flash ausgeschrieben werden kann. Durch Wearleveling im Controller werden Hotspots vermieden, zudem gibt es eine höhere interne Kapazität an Speicherzellen, so daß Blöcke, in die nicht mehr geschrieben werden kann, ersetzt werden können. Es gibt unterschiedliche Flash-Technologien, die sich hinsichtlich ihrer „Endurance“ unterscheiden. Als Metrik zur Spezifikation von Anforderungen zur Charakterisierung von Produkten haben sich

„Drive Writes per Day“ (DWPD) über einen bestimmten Zeitraum etabliert – im Konsumerbereich von 0.1 DWPD über 3 Jahre hin zu 5-50 DWPD über 5 Jahre im Unternehmenseinsatz für Server, wobei Einsatzbereiche mit hoher Schreiblast wie Caching oder Logging höhere Anforderungen stellen als solche, die primär lesen. Gelegentlich wird als auch eine Metrik auch „Total Bytes Written“ (TBW) verwendet, die allerdings wegen der Korrelation zur Kapazität zwischen Produkten nicht so offensichtlich vergleichbar ist.

Flash-Speicher gibt es in Form von Solid State Disks (SSD) mit SATA- oder SAS-Schnittstellen, die einfach in Platteneinschüben verbaut werden können. Das ist insbesondere in Speichersystemen interessant. In Servern werden angesichts der Leistungsdaten von Flash die Platten-Controller schnell zum Engpass, so dass dort Flash häufig in Form von PCIe-Karten verbaut werden.

Die Leistungsdaten einer Flash-Karte liegen in der Größenordnung von über 100.000 IOPS bei 100-200 µs Zugriffszeit, 1-2 GB/s Durchsatz – die Kapazitäten allerdings in der Regel unter einem TB.

Festplatten kommen gerade einmal auf 300 IOPS pro Laufwerk, haben dafür aber eine Kapazität von 4TB.

Betrachtet man die Kosten, so liegen die Kosten für 1 IOPS für Flash über eine Größenordnung niedriger als für Festplatten – allerdings liegen die Kosten pro GB Speicherkapazität dafür immer noch ein bis zwei Größenordnung höher.

Flash-Komponenten im Produktportfolio von Oracle

Oracle hat im Portfolio einerseits PCIe-Karten der Sun Flash Accelerator F-Familie, aktuell die F80 mit 800GB Kapazität, die optional in den meisten Servern von Oracle verbaut werden kann und auch in der Exadata zum Einsatz kommt. Zum anderen gibt es SSDs, die in den ZFS Storage Appliances für spezielle Zwecke zum Einsatz kommen –einerseits SSDs, die insbesondere für das Schreiben optimiert sind, als „Writezillas“ für den ZFS Intent Log, wo sie durch kurze Latenzen das Ausschreiben synchroner Schreiboperationen auf persistenten Speicher beschleunigen, andererseits SSDs, die als Readzillas den Second-Level-Cache für Dateien in den sogenannten Hybriden Storage Pools realisieren. Ganz neu im Portfolio ist das zur Oracle OpenWorld angekündigte Oracle FS1-2 Flash Storage System, das mit bis zu 912 TB Flash Storage geliefert werden kann.

Flash in der Speicherhierarchie

Wird nur eine kleine Speicherkapazität benötigt, kann man durchaus Flash direkt als Speichermedium nutzen. Ansonsten ist es effizienter, Flash mit Festplatten zu kombinieren, wobei Speicherung von Daten auf und gegebenenfalls die Verlagerung zwischen den unterschiedlichen Speichertechnologien automatisiert erfolgen sollte.

Hierzu gibt es je nach Einsatzbereich unterschiedliche Mechanismen: In den **ZFS Storage Appliances** wird Flash-Speicher als schneller Puffer verwendet: Zum einen, um synchrone Schreiboperationen schnell in nicht-flüchtigen Speicher zu schreiben während das Ausschreiben auf die Datenfestplatten dann ohne Performance-Einbußen asynchon erfolgen kann. Dazu werden Daten in den sogenannten ZFS Intent Log (ZIL) geschrieben, wo sie nur für einen kurzen Zeitraum (weniger als eine Minute) gehalten werden müssen, bis die Daten auch in die eigentlichen Datenpools auf Festplatten geschrieben sind. Dabei gibt es zwei Modi („Logbias“): Beim Logbias Latency wird die Latenz optimiert, indem die geänderten Daten in den ZIL geschrieben werden – damit wird eine minimale Latenz der Schreiboperationen erzielt, allerdings zu Lasten einer Verdoppelung der benötigten Bandbreite, da ja jede Schreib-I/O eine I/O sowohl in den ZIL wie auf die Datenplatten auslöst. Beim Beim Logbias Throughput werden die Daten deshalb direkt auf die Datenplatten geschrieben und

lediglich Metadaten in den ZIL, womit im ZFS schnelle synchrone Schreiboperationen außerhalb der Transaktionsgruppen erlaubt. Im Normalbetrieb wird der ZIL nie gelesen – er wird nur nach einem Crash benötigt. Daher ist die für den ZIL benötigte Kapazität recht klein (das maximale Schreibvolumen an synchronen Operationen über drei Transaktionsgruppen - voreingestellt je 5s). In Hochverfügbarkeitskonfigurationen mit zwei geclusterten Köpfen muss der Log vom überlebenden Knoten eingespielt werden. Daher können keine PCIe-Karten in den Servern verwendet werden, es werden vielmehr dualhomed SSDs verwendet.

Gelesene Daten werden zunächst im ZFS-Puffer (ARC) im Hauptspeicher (DRAM) der ZFS Storage Appliances gepuffert, so dass Rereferenzierungen sehr schnell sind. Je nach Modell sind 512 GB bis 2 TB DRAM verbaut. ZFS bietet die Möglichkeit den ARC um einen Second-Level-Cache, den L2ARC, zu erweitern – in den aktuellen ZFS SA mit SSDs um bis zu 12,8 TB. In diesen Hybrid Storage Pools der Kosten-Vorteil von Festplatten bei der Kapazität mit der Performance von Flash kombiniert..

Ein ähnlicher Ansatz wie beim L2ARC für Dateien im ZFS wird mit dem **Smart Flash Cache** für die Oracle Datenbank verfolgt, der seit 11gR2 für Oracle Solaris und Oracle Linux verfügbar ist. Der Smart Flash Cache ist letztlich ein Level 2 Cache für die SGA.

Auch in der **Exadata** wird der Flash-Speicher intensiv genutzt: Jeder X4 Storage Node enthält vier F80-PCIe-Karten mit je 800GB Kapazität. In einem Full Rack sind das ingesamt 44,8TB. Die effektive Größe kann durch Kompression (Advanced Compression Option) noch gesteigert werden. Dabei werden Daten beim Schreiben in den Exadata Smart Flash Cache komprimiert und beim Lesen wieder dekomprimiert. Ein weiteres Performance-Feature, das mit der Einführung der Exadata-Generation X3 kam, war die Option den Exadata Smart Flash Cache nicht mehr als Write-Through-Cache zu betreiben sondern als Write-Back-Cache. Im Write-Through-Modus müssen beim Schreiben Daten tatsächlich synchron auf Platte ausgeschrieben werden, während beim Write-Back der Cache das Ausschreiben eigenständig asynchron übernimmt, was Lastspitzen abfedern kann.

Lastspitzen werden auch beim Exadata Smart Flash Logging adressiert. Hierbei werden IOs zeitgleich gegen die plattenbasierten Redo-Logs wie gehgen einen kleinen Bereich im Flache Cache gestartet, wobei der erste erfolgreiche Abschluß der I/Os für die Datenbank ausreicht, was Performance-Ausreiser reduziert.

Bei dem zur OpenWorld 2014 angekündigten **Oracle FS1 Flash Storage System** wird Flash-Speicher nicht wie in den oben vorgestellen Systemen als Cache verwendet sondern als Primärer Storage – bis zu 228TB Performance Flash bzw. 912 TB Capacity Flash in einer FS1-2. In einer FS1-2 können neben Flash aber auch Festplatten bis zu einer Gesamtkapazität von 2,9 PB installiert werden. Daten werden dann über Autotiering entsprechend ihrer anwendungsabhängigen Priorität auf Speicherklassen – 300GB Performance SSDs, 1,6TB Capacity SSDs, 900GB 10krmp-Performance Disks oder 4TB 7.2krmp Capacity Disks - abgelegt bzw. zwischen Speicherklassen verschoben. Dabei wird eine im Vergleich zu Wettbewerbern sehr feine Granularität von 640 kB in den Heatmaps genutzt, die dem Autotierung zugrunde liegen. Dazu kommt das bereits von den Pillar Axiom Systemen her bekannte Konzept des anwendungsspezifischen Storage Quality of Service, über den nicht nur Speicherklassen definiert werden können, sondern auch die I/Os selbst priorisiert werden. Dabei kann sogar für Datenbereiche innerhalb einzelner Anwendungen (etwa Datafiles, Redologs, Controlfiles in der

Oracle DB) ein unterschiedlicher Storage Quality of Service definiert werden. „Wichtige“ Anwendungen können priorisiert und so eine hohe Auslastungen des Speichersystemen bei dennoch guter Performance erreicht werden. Für die Oracle Datenbank und ein Spektrum an Anwendungen sind vordefinierte Storage Quality of Service Profile verfügbar.

Das FS1-2 Flash Storage System hat eine flexible Scale-Out-Architektur mit einem aktiv/aktiv-konfigurierten Paar von Kontrolleinheiten, an die bis zu 30 Gehäuse mit Laufwerken jeweils einer Storage-Klasse angeschlossen werden können. Dazu kommt ein sogenannter Oracle FS Pilot als Managementsystem. Innerhalb eines Systems können sogenannte Storage Domains angelegt werden - virtualisierte, voneinander strikt isolierte Datencontainer, in denen auch unterschiedliche QoS-Einstellungen haben möglich. Storage Domains bestehen jeweils aus einer Reihe von RAID-Gruppen, die in verschiedenen Laufwerksgehäusen liegen können.

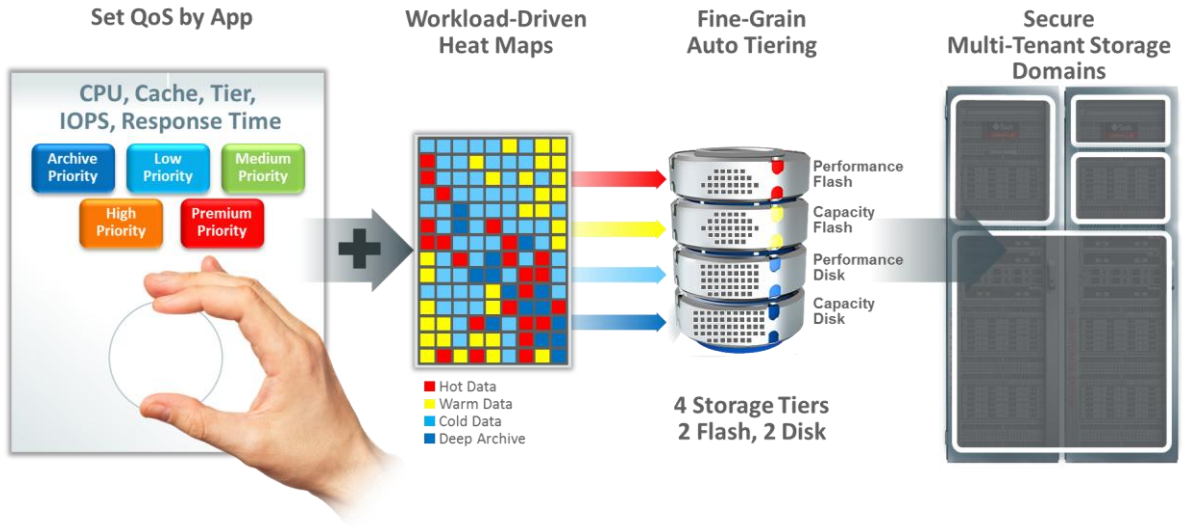

Abbildung 1: Storage Quality of Service im Oracle FS1 Flash Storage System

Eine zunehmende Bedeutung hat die Möglichkeit einer End-to-End-Integritäts-Prüfung der Daten. Bei den ZFS Storage Arrays ist ein solcher End-to-End-Mechanismus zur Integritätsprüfung elementarer Bestandteil der Architektur des zugrundeliegenden Filesystems, beim FS1 Flash Storage System wurde hierfür auf die T10 Protection Information Spezifikation (T10-PI) implementiert, die Oracle mit verschiedenen Anbietern von Storage-Komponenten entwickelt hatte.

Zusammenfassung

Flash-Speicher sind schnell und bezogen auf die Kosten je IOPS auch preisgünstig - über eine Größenordnung günstiger als für Festplatten – allerdings liegen die Kosten pro GB Speicherkapazität dafür immer noch ein bis zwei Größenordnung höher. Daher ist es besonders effizient, Flash mit Festplatten zu kombinieren, wobei die Speicherung von Daten auf und gegebenenfalls die Verlagerung zwischen den unterschiedlichen Speichertechnologien automatisiert erfolgen sollte.

Kontaktadresse:

Franz Haberhauer

Oracle Deutschland B.V. & Co. KG

Liebknechtstr. 35

70565 Stuttgart

Telefon: +49 (0) 711-72840-295

E-Mail franz.haberhauer@oracle.com

Internet: http://blogs.oracle.com/FranzHaberhauer