IO-Performance
Planung, Messung, Optimierung

Ulrich Gräf
Oracle B.V. & Co. KG

Schlüsselworte:

IO, Performance, Planung, Messung, Optimierung, Systembetrieb

Abstract

Viele aktuelle einfache IT Projekte bleiben bei den Anforderungen weit unter den Möglichkeiten die die Hardware bietet. Wenn jedoch im Rahmen von Kosteneinsparungen Konsolidierungen im Bereich vom Plattenspeicher vorgenommen werden, kann es sein dass man an Limits der eingesetzten Hardware stößt.
Der Vortrag zeigt aktuelle Trends auf und erläutert Funktionsweisen der zur Zeit verfügbaren Hardware. Es wird aufgezeigt, wie man den Leistungsbedarf einer Applikation misst, die Leistungsfähigkeit einer Platte einschätzen kann und dann das Speichersubsystem entsprechend planen kann.

Plattentechnologien

Für aktuellen Rechenzentrumsbetrieb stehen zur Zeit lokal angeschlossene Platten, Plattensubsysteme mit Cache (LSI, HDS, EMC, …) und SSDs zur Verfügung.

Lokale Platten werden direkt an die Rechner angeschlossen, oder über Kabel in einem sogenannten JBOD (Just a Bunch Of Disks = ist kein Controller drin) betrieben.
Die Antwortzeiten dieser Platten sind im einstelligen Millisekunden-Bereich.

Bei Plattensubsystemen wird ein Controller benutzt, der in der Regel einen auf schnellem RAM basierenden Cache betreibt. Dieser wird als Lese oder Schreib-Cache betrieben.
In der Regel wird der Cache gespiegelt und/oder mit Batterien gepuffert, so dass das Subsystem eine Erfolgsmeldung des Schreibens zum Host melden kann, ohne dass auf die Platte gewartet werden muss. Dadurch haben Plattensubsysteme mit Cache sehr kurze Schreibzeiten.
Die Funktion als Lese-Cache wirkt im Hochlast-Betrieb kaum, da der Cache in der Regel kleiner ist, als die Gesamt-Datenmenge. Wird wie bei Datenbanken auf Host-Seite nochmal gecacht, dann stimmt die Statistik des Caches nicht mehr – im Prinzip hat er zuwenig Informationen, was die Datenbank braucht.
Schreibe und Lesen aus dem Cache ist meist in unter einer Millisekunde möglich. Lesen von der Platte (wenn die Daten nicht im Cache sind), braucht einige Millisekunden.

SSDs auf Basis von DRAMs sind seit einigen Jahren im Einsatz. Die Daten werden durch Batterien geschützt. Leider sind die Kosten durch den Einsatz von DRAMs extrem hoch. Für bestimmte Anwendungen sind sie jedoch die einzige Möglichkeit.

Im Einsatz in rauhen Umgebungen, bei denen rotierende Platten mechanische Probleme bekommen, werden diese SSDs zu Zeit von SSDs auf Basis von Flash ersetzt.
Die Antwortzeiten sind im einstelligen Mikrosekunden Bereich. Die Datenübertragung zu einer solchen Platte hat einen wesentlichen Anteil an der Antwortzeit.

SSDs auf Basis von Flash basieren auf Halbleiter Chips, die ihre Daten auch behalten, nachdem der Strom abgeschaltet ist. Viele Jahre waren die Speicher nur in USB-Sticks, portablen Geräten usw. im Einsatz, jedoch mittlerweile werden sie wie Platten eingesetzt.
Die Antwortzeiten sind im 2-3 stelligen Mikrosekunden-Bereich, also besser als Platten aber schlechter als SSDs auf Basis von DRAM.

Leistung von Platten

Platten werden meist nach der Datenmenge eingekauft. Im Einsatz im Rechenzentrum, speziell wenn es um Konsolidierung oder um gemeinsame Benutzung von Plattensubsystemen geht, sind jedoch auch noch die Parameter Datenrate und Anzahl von Ios/Sekunde wichtig.

Die Größe einer Platte ist durch die Hardware bestimmt. Unter Umständen kann es aber sinnvoll sein, auf Kapazität der Platte zu verzichten um mehr Leistung zu erzielen (short stroked disk, siehe unten). Reale Platten haben zur Zeit eine Kapazität von bis zu 2 TB.

Die Datenrate einer Platte wird durch die Datendichte und die Geschwindigkeit des Kopfes bestimmt. Obwohl die heutigen Platten dem Host normalerweise eine regelmäßige Struktur aus Plattenoberfläche, Spur und Sektor präsentieren, werden die Daten anders abgelegt (zone bit recording). Auf den äußeren Spuren ist mehr *Fläche*, so dass hier mehr Sektoren gespeichert werden. Außerdem gibt es aufgrund der hohen Speicherdichte immer defekte Sektoren, die durch Reserve-Sektoren ersetzt werden. Daher werden die Platten abhängig davon in verschiedenen Qualitäten vermarktet (*quality*, *grade*, *selectio, ...*). Daher ist die Datenrate dort höher. Die Datenraten aus den Datenblättern müssen nach diesem Aspekt interpretiert werden.
In der Regel erreicht man diese Datenrate nur mit sequentiellem Schreiben oder Lesen. Die Datenraten liegen heute bei bis zu 250 MByte pro Sekunde.

Die Anzahl IOs/Sekunde bei nicht sequentiellem Zugriff ist der wesentliche limitierende Faktor bei Konsolidierungen. Sobald man random liest und schreibt muß man auf der Platte positionieren. Dies wird auch dann notwendig, wenn ein Plattensubsystem von mehreren Hosts benutzt wird und die LUNs auf alle Platten verteilt werden. Diese Policy ist im Sinne einer gleichmäßigen Auslastung des Plattensubsystems sinnvoll, allerdings wird dadurch die Höchstleistung bei sequentiellem Zugriff eines Hosts verhindert (der Plattenkopf wird immer wieder weggezogen).
Um einen beliebigen Zugriff zu implementieren, muss sich der Kopf auf die richtige Spur bewegen und die Platte muss sich weit genug drehen, damit der richtige Sektor unter dem Kopf liegt. Die aktuellen Platten drehen mit 15000 RPM = 250 mal pro Sekunde, das entspricht 1 x pro 4 Millisekunden. Der mittlere Zugriff auf einen Sektor erfolgt somit in 2 Millisekunden. Die Bewegung des Kopfes ist meist so angepasst, dass die richtige Spur ebenfalls in dieser Zeit erreicht ist. Leider gilt diese Zeit nur dann, wenn genau ein Auftrag an eine Platte gesandt wurde. Heutige Platten können mehrere Befehle (bis zu 256) zum Lesen und Schreiben entgegennehmen (tags, tagged-token-queuing), die Abarbeitung optimieren und die Ergebnisse ggf in anderer Reihenfolge zurückliefern. Da die Positionierung für den 2. Befehl erst nach dem 1. Befehl erfolgen kann, usw steigt die mittlere Antwortzeit von Platten wenn viele Requests von einer Platte zu beantworten sind.
(1 Befehl: 2 ms, 2 Befehle: (2+4)/2= 3 ms, 10 Befehle: 11 ms, …).

In der Praxis zeigt sich weiterhin, dass man zu der reinen Positionierzeit, die man aus der Umdrehungszeit in RPM berechnen kann, noch Datenübertragungszeiten zwischen Platte und Controller in der Platte, im Plattensubsystem, auf dem Fibre-Channel usw. dazu rechnen muss.
Eine 15000 RPM-Platte schafft daher in der Regel gerade mal 200-250 IOs pro Sekunde.

Die Misere mit den Platten

Die Platten sind in den letzten Jahren immer langsamer geworden. Wirklich?

Die Datenblätter sind dank Internet immer noch erreichbar.
Im Jahr 1990 gab es die brandaktuelle 200 Mbyte-Platte, die eine mittlere Positionierzeit von 12 ms und eine Datenrate von ca. 500 Kbyte/Sekunde hatte.
Heute im Jahr 2010 ist die 2Tbyte-Platte aktuell, die eine mittlere Positionierzeit von 4 ms hat und 200 Mbyte pro Sekunde Datenrate beherrscht.

Sicher, wenn ich eine Applikation habe, die 200 Mbyte braucht, und ich nutze eine neue Platte, dann habe ich 3 mal mehr IOs/Sekunde und 400 mal mehr Datenrate: sieht gut aus!
Diese Sicht ist auch die, welche die meisten User aus der Industrie haben, weil sie arbeiten an einem PC (o.ä.) und haben die Platte für sich. Die sichtbare Leistung ist extrem gestiegen.
Im Rechenzentrum will man allerdings konsolidieren, und im Extremfall will man die 2 TB voll ausnutzen, dann gehen 10000 dieser Applikationen auf die neue Platte (2TB / 200 MB).
Allerdings hätte man dann pro Applikation einen extremen Leistungseinbruch, weil man hat nur noch 3/10000 der IOs/Sekunde und 400/10000 = 1/25 der Datenrate, die mit der alten Platte der Applikation zur Verfügung gestanden hätte.

Man kann auch relative Messwerte berechnen, die einem die Verlangsamung der Platten begreiflicher machen:

- IOs/Sekunde pro GB (wieviele Ios stehen jedem Gbyte zu):
  200 MB Platte: 416 io/s 2 TB Platte: 0.125 io/s

- MB/Sekunde pro GB (welche Datenrate ist für jedes Gbyte übrig):
  200 MB Platte: 2.5 MB/s 2 TB Platte: 0.1 MB/s

- Zeitdauer Plattensicherung (wie lange dauert es eine Platte zu sichern):
  200 MB Platte: 400 Sekunden 2 TB Platte: 10000 Sekunden

Man sieht aus diesen Werten, dass die Speicherkapazität weit mehr als die anderen Parameter gestiegen ist. Leider ist es aus mechanischen Gründen extrem schwierig die Umdrehungszahl zu erhöhen (Zentrifugalkraft, Luftgeschwindigkeit vs. Schallmauer).

Plattensubsysteme mit batteriegepuffertem Cache führen weitere Engstellen ein, die in der Konfiguration berücksichtigt werden müssen. So haben bei den meisten Herstellern die Anschlusspunkte für die SAN-Anschlüsse und die internen Controller eine limitierte Leistung bei Datenrate und IOs. Meist sind weitere Bussysteme zwischen den SAN-Anschlüssen, den Controllern den Platten-Trays und den Platten vorhanden, die bei der Konfiguration auf Engpässe geprüft werden müssen, wenn es um Spitzenleistungen geht (in Zusammenarbeit mit dem Hersteller).

Bei hohen Datenraten kann in einem Plattensubsystem mit batteriegepuffertem Cache ein fataler Betriebszustand entstehen: Der Überlauf des Schreibcache.

Die Ursache ist in der Regel die, dass zuviele Random-Writes aus der Applikation erfolgen oder dass die Platten von mehreren Hosts über verteilte LUNs genutzt werden. Wenn dann zuwenig Platten konfiguriert sind, schaffen es die Controller des Plattensubsystems nicht mehr, die Daten wegzuschreiben. Zu erkennen ist das daran, dass der Füllgrad des Write-Caches im Plattensubsystem andauernd steigt. Wenn der Cache voll ist, dann muss für jeden Schreib-Vorgang vom Host auch auf die Platte geschrieben werden. Die Antwortzeit steigt schlagartig an, wodurch mehr parallele Schreibvorgänge durchgeführt werden müssen (weil sie ja länger aktiv sind). Die Folge ist ein extremer Leistungseinbruch auf allen angeschlossenen Hosts.

Ein Beispiel: Problem durch Modernisierung

In einem aktuellen Fall hat ein Kunde eine Applikation auf einen neuen Rechner mit einem neuen Plattensubsystem umgezogen. Die alte Konfiguration nutzte 72 GB FC-Platten, die neue Konfiguration setzte neue Technologie ein: 300 GB FC Platten. Die Zugriffszeit verbesserte sich um den Faktor 1.5, da die neuen Platten jetzt mit 15000 RPM drehen (die alten mit 10000 RPM).
Allerdings waren nur noch ca. 1/5 der Platten für den Datenbestand notwendig und auch so konfiguriert. Die Platten kamen mit den 5 fach höheren IOs nicht zurecht und die Applikation lief zu langsam.
Die Symptomatik konnte entschärft werden, indem die LUNs auf alle Platten verteilt wurden.

Allerdings wurde ebenfalls festgestellt, dass bei dem geplanten Wachstum die Applikation zwar nur 1/8 des Platzes des neuen Plattensubsystems braucht, allerdings 60 % der möglichen IOs. Es können daher nur noch Applikationen mit geringen IO-Anforderungen parallel auf das gleiche Plattensubsystem geschaltet werden.

Ermittlung des Leistungsbedarfs

Die Datenmenge ist in der Regel planbar aus Anzahl der Nutzer, Größe der Webseiten, Anzahl Artikel, usw. Die Datenraten und die IOs/Sekunde lassen sich mit Faustregeln der Applikationen abschätzen. Dazu bitte den Hersteller der Applikationen kontaktieren.

Eine bessere Abschätzung erhält man wenn man eine Vorversion der Applikation ausmessen kann. Auf einem Unix (Solaris!) System kann mit iostat die IOs/Sekunde und die Datenrate ermitteln. Für Linux, BSD, Windows gibt es ebenfalls Kommandos, die diese Werte liefern.

Wird eine Datenbank eingesetzt gibt es in der Regel die Möglichkeit die benötigten Raten auszulesen und dann auf neue Userzahlen etc. hochzurechnen. Bei Oracle zB. liefert Statspak diese Informationen.

Wird ein Oracle ZFS Storage Appliance (auch S7000) eingesetzt, kann man mit den mitgelieferten Analytics diese Werte ermitteln und dann die Zahl der notwendigen Platten planen.

Techniken zur Leistungssteigerung

Die Nutzung von SSDs zur Leistungssteigerung springt ins Auge. Diese Variante zur Zeit jedenfalls (2010) noch sehr teuer. Eine Variante ist, die am meisten abgefragten Tabellen oder die Indizes auf

SSD Speicher zu legen, wenn sie nicht sowieso permanent im Hauptspeicher gehalten werden können. Leider erfordert das einen erhöhten Administrationsaufwand.
Eine andere Variante ist die Nutzung von Flash Storage mit Oracle DB ab Version 11gR2. Hier wird die Flash SSD von der Datenbank direkt gesteuert und der Aufwand zum Tunen entfällt oder ist geringer. Diese Technik wird auch in den Exadata Systemen eingesetzt.

Handelt es sich um Filesysteme hat Oracle mit dem ZFS im Oracle Solaris optimale Möglichkeiten die Leistung zu optimieren.

- ZFS schreibt automatisch parallel auf alle konfigurierten Platten
- ZFS fasst Random Writes zusammen, Resultat sind weniger IOs auf den Platten
- Synchrone Schreibvorgänge werden sequentiell im Log abgelegt
- Das Log kann punktuell beschleunigt werden, indem man dafür eine SSD (oder SSD-mirror) einsetzt (*Logzilla*). Es gibt spezielle SSDs mit hoher Transaktionsrate die hier besonders geeignet sind (im ZFS Storage Appliance / S7000 verwendet). Die gewünschte Schreibgeschwindigkeit kann über die Anpassung der Anzahl der Log-SSDs erzielt werden.
- SSDs können immer wieder gelesene Daten vorhalten (*Readzilla*), die nicht mehr in den Hauptspeicher des Systems passen. Diese Variante ist günstiger als den Hauptspeicher des Hosts zu erweitern.
  Die Variante ist auch günstiger als den Cache des Plattensubsystems zu erweitern, weil dieser ist so teuer wie Hauptspeicher im Host. Die Lese SSDs sind günstiger, weil sie auf Flash basieren.
- Diese SSDs (*Logzilla* und *Readzilla*) konfigurieren sich automatisch weil sie vom Host gesteuert werden, währendessen ein Plattensubsystem mit Cache falsche Statistiken erhält und daher den Cache nicht richtig steuern kann.
- ZFS kann auf Latency- oder Throughput-Optimierung eingstellt werden, je nach Bedarf.

Diese Optimierungen sind bei anderen Filesysteme oder Plattensubsystemen nicht möglich.

Die Technologien des ZFS stehen auch den Nutzern eine ZFS Storage Appliance (S7000) zur Verfügung.

Ausblick

Ein vollständiger Ersatz von Platten durch SSDs ist zur Zeit (2010) noch zu teuer. Es ist auch abzusehen, dass SSDs in Zukunft anders als Platten angeschlossen werden, da die heutigen Anschlussmöglichkeiten (IDE, SCSI, FC, SATA, SAS) auf die limitierten Fähigkeiten von mechanischen Platten abgestimmt sind.
Die Adoption der Flash SSDs wird durch diese Anpassungen wahrscheinlich verzögert.

Zur Zeit ist die beste und kostengünstigste Möglichkeit das Beste aus der augenblicklichen Situation bei Platten herauszuholen, der punktuelle Einsatz von Flash SSDs. Dies kann durch die Nutzung von Flash Storage in der Oracle DB oder mit ZFS in Oracle Solaris erfolgen.

Kontaktadresse:

Ulrich Gräf	Telefon:	+49 (0) 6103 753 359
Oracle B.V. & Co. KG	E-Mail	ulrich.graef@oracle.com
Amperestr. 6	Internet:	www.oracle.com
D-62225 Langen		