
ORACLE®

ORACLE®

IO Performance -
Planung Messung, Optimierung

Ulrich Gräf
Principal Sales Consultant
Oracle Deutschland B.V. und Co. KG

The following is intended to outline our general product direction. It is intended for information purposes only, and may not be incorporated into any contract. It is not a commitment to deliver any material, code, or functionality, and should not be relied upon in making purchasing decisions.
The development, release, and timing of any features or functionality described for Oracle's products remain at the sole discretion of Oracle.

Platten heute

- Platten sind schnell geworden... 15000 RPM
- Platten haben hohe Kapazität... 1 ... 2 TB
- Platten haben hohe Übertragungsrate
  - On disk: 250 MByte / Sekunde
  - SATA: 600 MByte / Sekunde

Wahr oder falsch ?

Platten heute

- Platten sind schnell geworden... 15000 RPM
- Platten haben hohe Kapazität... 1 ... 2 TB
- Platten haben hohe Übertragungsrate
  - On disk: 250 MByte / Sekunde
  - SATA: 600 MByte / Sekunde

Wahr oder falsch ?

Platten sind schnell im Zugriff

- 15000 RPM = Umdrehungen pro Minute
  = 15000 / 60 = 250 Umdrehungen pro Sekunde
  = 4 ms pro Umdrehung

- Random IO (aktuelle SAS Platten):
  - Kopf muss auf die richtige Spur (0.2 ... 1.2 ms)
  - Platte muss sich an den Sektor drehen (0 ... 4 ms)
  - => ~ 2 ms mittlere Zugriffszeit (peak, (= nie erreichbar))
  - Realistisch sind 50 % der möglichen IOs erreichbar
- Sequential IO: kein Problem
  - Daten kommen mit Lesegeschwindigkeit
    (realistisch ~ 200 MB/Sekunde)

Wie schnell sind Platten unter Last?

- Tagged Token Queueing / Command Tags / ...
  - Platte bearbeitet mehrere Befehle
    (sichtbar in: `iostat -xzn`, Spalte: actv)
  - Platte definiert die Reihenfolge selbst
  - Antworten werden durch die Tags zugeordnet

Wie schnell sind Platten unter Last?

- Antwortzeiten (random read):
  - 1 Befehl: 2 ms (siehe oben)
  - 2 Befehle: der erste: 2 ms, der zweite: 2 ms (danach) Mittelwert: (2 + 4) / 2 = 3 ms
  - 10 Befehle: 2 ms ... 20 ms Mittelwert: 11 ms Antwortzeit
  - moderne Platten: 128 Tags: 129 ms Antwortzeit
- Meist eine Mischung random read mit sequential read
  - 16 ms Antwortzeit => 2 ... 30 ms => 15 random reads pro Sekunde
  - bei 30 reads pro Sekunde => 50% sequential

Wesentlich ist die Antwortzeit!

- Neue Platten können auch einen lokalen Flash Cache haben
- ZFS schaltet den Write Cache der Platte ein
  - Write-Commands sind schnell abgearbeitet
  - Intern braucht die Ausführung noch Zeit und Positionierungen
- Bestimmte iostat Werte sind dann irrelevant
  - %busy
  - asvc_t wird verfälscht (weil write zu schnell)
    zeigt aber den wahren Zustand der Platte

Neues Projekt beim Kunden

- Aktuelles Projekt
  - 200 Platten a 73 GB ersetzt durch 50 x 300 GB Platten
  - Alte Platten 10k RPM, neue Platten 15k RPM
- Random IOs
  - Kunde erwartet 15000 Random IOs / Sekunde
- Situation:
  - Erst nur ein Teil der Platten im Einsatz (20)
    20 x 250 pro Platte = 5000 => Performance mangelhaft
  - Dann Nutzung von 20% der Kapazität und 80% der IOs

Die Platten werden immer schneller

- Ein Vergleich
  - 1990: 424 MB Platte, 12 ms Zugriffszeit, 500 KB / Sekunde
  - 2010: 2 TB SATA Platte, 4 ms Zugriffszeit, 200 MB / Sekunde

200 MB/Sek / 500 KB/Sek = 400 x Schneller? NEIN!

- Sicherung der ganzen Platte:
  - 424 MB Platte: ~ 1000 Sekunden ... 20 Min
  - 2 TB Platte: ~ 10000 Sekunden ... 3 Stunden
- IOs für 2 TB
  - 424 MB Platte: 80 pro Platte, 320 000 (für 4000 Platten)
  - 2 TB Platte: 250 pro Platte und für 2 TB

Platten Technologie geht immer weiter

- Rechenbeispiel: 3.5 in Platte
  - 3.5 in Durchmesser = ~ 8.75 cm Durchmesser
  - ~ 28 cm Umfang
  - Bei 250 Umdrehungen (15k RPM): 250 x 28 cm Geschwindigkeit
  - 250 x 28 cm/Sek = 7000 cm/Sek = 70 m/Sekunde
  - ~ 250 Km/h
- Geht noch schneller? Nein!
  - Heutige Platten sind Winchester Platten
  - Kopf wird angedrückt und "fliegt" auf einem Luftpolster
  - Überschreitet man die Schallgeschwindigkeit an einer Stelle: Stoßwelle (Überschallknall) -> Headcrash!

IO Performance - ist es das wirklich?

Sicherstellen dass IO das Limit ist

- Ausschliessen CPU Überlastung
  - vmstat: idle >> 0%
  - mpstat: keine Einzel-CPU überlastet
  - genug Threads in der Applikation
- Ausschliessen Memory Überlastung
  - vmstat: memory-free > 1/32 des Hauptspeichers
  - vmstat: sr = 0 (ausser bei Spitzen)
- Nachweis schlechter Antwortzeiten
  - iostat -xzn 5: asvc_t > 20 ms (Antwortzeit der Platte)

Planung

- Hinweise aus der Applikation
  - Testläufe mit iostat -xzn 5
  - Oracle: Statspak, ...
  - Angaben des Kunden über Wachstum
- Bestimmung der Parameter:
  - Größe des benötigten Storage
  - Benötigte Datenrate
  - Anzahl von IOs

Planung: Sizing

- Bestimmung der Zahl der Platten
  - Alle Parameter müssen erfüllt sein!
    - Bestimmend sind heute meist die IOs / Sekunde
  - Kunden / der Einkauf des Kunden kaufen jedoch nach Kapazität
  - Viele Random IOs können die mögliche Datenrate behindern
  - Mit Mitteln der Applikation die sequentiellen Anteile entkoppeln
    (full table scans auf eigene Platten)

IO Performance Optimierung

- Datenbank: höchste Performance mit ASM
- Datenbank auf Filesystem
  - Directio einschalten
    - UFS: forcedirectio mount-Option
    - ZFS: zfs set primarycache=metadata ...
  - Last access time update abschalten
    - UFS: noatime mount-Option
    - ZFS: zfs set atime=off ...

IO Performance Metadaten

- UFS: 1/2000 (1/1000 auf x64)
  - Block Pointer ist 32 Bit = 4 Byte
  - Block ist 8kb / 4kb
- ZFS: 1/1000
  - Block Pointer ist 128 Byte (checksums...)
  - Block ist 128kb

- Platz im Hauptspeicher notwendig
  - Bei UFS passt es meist (max 16 TB)
  - Bei grossen ZFS: genug Hauptspeicher oder mehr IOs (aktueller Fall: 300 TB Filesystem auf 64 GB ZFS Appliance)

Cache im SAN

- Cache im SAN
  - Hilft bei Schreib-Operationen
  - Hilft nicht beim Lesen
    - Datenbanken haben eigenen Cache (SGA)
    - Wichtige Sachen sind lange in der SGA
      dann aber nicht mehr im Cache des SAN
- Lösung: gemeinsame Kontrolle aller Caches
  - wie bei CPUs

IO Performance mit Flash SSDs

- Flash ist die kommende Technologie
  - Noch zu teuer für generellen Ersatz von Platten
  - Schon besser als mechanische Platten bei
    - IOs / Sekunde pro GB
    - IOs / Sekunde pro $
- Einsatz wo?
  - bei einzelnen Dateien / DB Tabellen
  - Intelligenter Einsatz
    - ZFS
    - Oracle DB 11g R2
    - Exadata

ZFS und Flash SSDs

ZFS ist ein Filesystem einer neuen Generation
- Integration Volume Manager und Filesystem
- ... und vieles mehr

- ZFS benutzt ein Log für synchrone Schreiboperationen
  - Kann auf separate Disk gelegt werden
  - Performance mit spezieller Flash SSD
  - SAN Device (separate RAID-Gruppe)
- ZFS kann Flash als Read Cache einsetzen
  - Bestimmte Filesysteme (konfigurierbar)
  - Zentrale Kontrolle des Caches vom ZFS aus

Oracle DB und Flash SSDs

- Oracle 11gR2 kann Flash disk einsetzen
  - Automatischer Read Cache
  - Auch spezielle Tabellen / Indizes
  - Limitiert auf den einzelnen Rechner
- Exadata und Flash SSDs
  - RAC Cluster + Storage Knoten
  - Teil des ASM Codes auf dem Storage Knoten
  - Flash Disk liegt im Storage Knoten
  - Cache ist für alle Knoten des Clusters verfügbar