ORDIX AG

Linux Memory Management für MySQL

SIG MySQL - Performance
13.03.2012

Marius Dorlöchter
mdo@ordix.de
www.ordix.de

Vorstellung

ORDIX AG

- Marius Dorlöchter
- Consultant bei ORDIX seit 2006

- Gruppe Systemintegration
  - Betriebssysteme: Linux, Solaris, AIX
  - Monitoring
  - Apache / Tomcat

Agenda

ORDIX AG

- Speichervarianten
- RAM, Swap und Virtueller Speicher
- Virtual vs. Resident
- Buffer und Cache
- Paging / Swapping Verhalten
- Memory Overcommitment
- „Out of Memory“
- PageTables und Translation Lookaside Buffer (TLB)
- Hugepages
- Relevante Limits
- Tools zur Performance Analyse

Speichervarianten auf einem Linux System

ORDIX AG

Wie teil sich der Speicher eines Systems auf?

- CPU Kern / Register (Von der Hardware verwaltet)
- L1-3 Cache (Von der Hardware verwaltet)
- RAM (Vom Betriebssystem verwaltet)
- Festplatte (Vom Betriebssystem verwaltet)

RAM

ORDiX AG

- ca. 100 x langsamer als der CPU Kern
- ca. 100.000 x schneller als Festplattenzugriffe
- Wofür wird der Arbeitsspeicher (RAM) bei Linux genutzt?
  - Kernel Memory: kann nicht ausgelagert werden (erscheint nicht als Teil des verfügbaren Speichers)
  - Program Memory: Prozessdaten (Code, Heap, Stack)
  - Shared Memory: Gemeinsam genutzter Speicher
  - Buffers: Puffer für Zugriffe auf Blockdevices, FS-Metadaten
  - Cache: Puffer für Dateiinhalte
  - Free Memory: Ungenutzter Speicher

SWAP

ORDIX AG

- Speicherbereich auf der Festplatte zur „Erweiterung des RAM“
- Realisiert als Partition oder Datei („Swapfile“ oder „Pagefile“)
- „RAM + SWAP = Maximal nutzbarer Speicher“
- Wird tatsächlich mehr Speicher genutzt als RAM vorhanden, werden inaktive Seiten auf den SWAP Bereich ausgelagert (pageout)
- Werden die ausgelagerten Seiten wieder benötigt, wird beim Zugriff ein *pagefault* erzeugt und der Kernel schiebt die Seiten wieder in den RAM (pagein)
- Eingelagerte Seiten bleiben im Swapspace solange sie readonly genutzt werden (Swapcache)
- Zuständiger Kernel-Prozess: kswapd

Virtual vs. Resident

ORDIX AG

- Virtueller Speicher kann größer sein als RAM
  (in der Regel fordern Prozesse mehr Speicher an als sie tatsächlich brauchen)

Virtual Set Size (VSZ):	Die Speichermenge, die einem Prozess insgesamt zugeordnet bzw. für ihn reserviert ist (Code + Libs + Heap/Stack)
Resident Set Size (RSS):	Die Speichermenge, die ein Prozess tatsächlich gerade nutzt
Private RSS:	Der Teil des Speichers eines Prozesses, der aktuell genutzt und nicht mit anderen geteilt wird
Shared RSS:	Der Teil des Speichers eines Prozesses, der aktuell genutzt und mit anderen geteilt wird

Buffer und Cache

ORDIX AG

Warum ist der Speicher bei Linux nach einer Weile immer „voll“?

- „Ungenutzter Speicher ist verschwendeter Speicher“
- Ungenutzter Speicher wird daher für Buffer und Cache verwendet
  - Buffer: Block-I/O (z.B. Inode-Table, etc)
  - Cache: Filesystem (Dateiinhalte)
- Buffer und Cache sind potentieller freier Speicher und stehen sofort zur Verfügung, falls Speicher von Prozessen benötigt wird
- Eine kleine Menge freier Speicher reicht normalerweise aus, um akuten Anforderungen der Prozesse sofort nachzukommen

Paging / Swapping Verhalten

ORDIX AG

Was passiert, wenn RAM zu 100% belegt ist und ein Prozess zusätzlichen Speicher braucht?

- Alternativen
  - Buffer / Cache freigeben
  - Inaktive Seiten auslagern (pageout)
- Konfigurierbar über Kernel-Parameter „vm.swappiness“:

  (Werte zwischen 0 und 100, default: 60)
  - 100: pageout immer bevorzugen
  - 0: cache/buffer immer freigeben

Memory Overcommitment

ORDIX AG

- Prozesse fordern in der Regel deutlich mehr Speicher an, als sie wirklich benutzen
- Standardmäßig keine Begrenzung des Virtuellen Speichers!
- Konfigurierbar über Kernelparameter:
  - vm.overcommit_memory
  - vm.overcommit_ratio
- Werte für overcommit.memory:
  - 0: „heuristic Overcommitment“
  - 2: Overcommitment bis „vm.overcommit_ratio/100 * RAM + SWAP“
  - 1: Overcommitment ohne Begrenzung

Out of Memory (OOM)

ORDIX AG

Was tun, wenn Overcommitment schief gegangen ist?

→ Einen Prozess beenden…aber welchen? Den „bösesten“!

Kommentar im Kernel-Quellcode zur Funktion badness() bzgl. der Kriterien, nach denen der zu beendende Prozess ausgewählt wird:

1. we lose the minimum amount of work done
2. we recover a large amount of memory
3. we don't kill anything innocent of eating tons of memory
4. we want to kill the minimum amount of processes (one)
5. we try to kill the process the user expects us to kill, this algorithm has been meticulously tuned to meet the principle of least surprise ... (be careful when you change it)

PageTables und Translation Lookaside Buffer

ORDIX AG

- Speicher ist aufgeteilt in Pages der Größe 4k
- Zuordnung Virtual <-> Physical Speicheradresse erfolgt in PageTable
- Translation Lookaside Buffer (TLB) = Cache für PageTable Einträge
- Jeder Prozess erhält eigene virtuelle Zuordnungstabelle

Hugepages

ORDIX AG

- Standardgröße einer Speicherseite (Page) : 4k
- Größe einer Hugepage: 2048k
- Zur Reduzierung der Pagetable-Größe und Optimierung des TLB
- Zusammenhängender Speicherbereich
- Reservierung erfolgt über Kernelparameter `vm.nr_hugepages`
- MySQL-Parameter (InnoDB): `large-pages`

Ulimit

ORDIX AG

Mit ulimit können Limits auf User-Ebene gesetzt werden:

- file size
- max locked memory
- open files
- cpu time
- max user processes
- virtual memory

Tools zur Performance-Analyse

ORDIX AG

Analyse-Tools: `free [ -m ]`

- free zeigt eine Zusammenfassung der Speicherauslastung des Systems an

Hinweise:

- Spalte „shared“ ist immer 0 (nicht implementiert)
- Wichtig sind vor allem die Werte **„+/- buffers / cache**“ - sie definieren die **reale** Speicherauslastung durch die Prozesse
- Shared Memory Segmente „verstecken“ sich in „**cached**“

Tools zur Performance-Analyse

ORDIX AG

Analyse-Tools: `sar <Option> <interval> <# messungen>`

- Für fast alle Performance-Messungen zu gebrauchen - muss evtl. nachinstalliert werden (Paket: sysstat)

- Option `-r`: Speicherauslastung
- Option `-W`: Pagein / Pageout Aktivitäten
- Option `-S`: Swapspace Nutzung

Hinweise:

- Es ist nicht schlimm, wenn Swap genutzt wird, solange nicht permanent ein- und ausgelagert wird
- Ist der Pagein/Pageout Wert groß, besteht RAM-Engpass (es wird ständig geswappt)

Tools zur Performance-Analyse

ORDIX AG

Analyse-Tools: `vmstat <interval> <# messungen>`

- Komprimierte Anzeige von Speicherauslastung, Paging-Aktivitäten, I/O und CPU-Auslastung

Hinweise:

- Die Werte für **si/so** zeigen die Swapping-Aktivität an (idealerweise -> 0, kleine Werte (<100) vertretbar)

Tools zur Performance-Analyse

ORDIX AG

Analyse-Tools: `ps aux`

- Speicherbelegung der Prozesse auslesen

Hinweise:

- Beachte die Spalten RSS (Resident Set Size) und VSZ (Virtual Set Size)
- ACHTUNG: VSZ und RSS beinhalten i.d.R. große Teile geteilten Speichers (Summe aller RSS != reale Speicherauslastung!)

Tools zur Performance-Analyse

ORDIX AG

Analyse-Tools: `pmap <PID>`

- Detaillierte Speicherbelegung einzelner Prozesse auslesen

Hinweise:

- nur die Speicherbereiche, die als „rw“ gekennzeichnet sind und nicht auf Shared Memory verweisen, belegt der Prozess wirklich exklusiv!

ORDIX AG

Zentrale Paderborn
Westernmauer 12 - 16
33098 Paderborn
Tel.: 05251 1063-0

Trainingscenter Wiesbaden
Kreuzberger Ring 13
65205 Wiesbaden
Tel.: 0611 77840-00

Zentrales Fax:
0180 1 67349 0
0180 1 ORDIX 0

Weitere Geschäftsstellen
in Köln, Münster und Neu-Ulm

E-Mail: info@ordix.de
Internet: http://www.ordix.de

Vielen Dank für Ihre Aufmerksamkeit!