

Monday, August 27. 2007

Kris Empfehlungen für mysql auf Solaris umgesetzt

Ich habe mir mal die Anmerkungen von Kris zur Installation von mysql angeguckt und in OpenSolaris Speak umgesetzt: zpool create mysqlpool mirror mirror ... log zfs set atime=off mysqlpoolzfs set blocksize=8k mysqlpoolroot kommt in ein UFS. ZFS ist ja leider noch nicht bootfähig. Kommt noch. Über die Aufteilung der Partitionen muss man sich keine Gedanken machen. Wer das allerletzte an Performane rauskitzeln will kann noch data und log in unterschiedliche pools trennen, aber das seperated log macht das weitestgehend unnötig.

Der Hinweis mit dem Thinwidestripping ist auch für ZFS gültig. Datenbanken mögen das wirklich. Deswegen hassen DBA auch 500 GB Platten. Die sind aber garnicht so schlimm, wenn man viele davon hat. Eine X4500 "Thumper" ist daher eine sehr nette Datenbankmaschine.

Ein Hinweis für mysql und Solaris: Meiner Erfahrung nach ist libumem für mysql eine verdammt gute Idee ... also ins Startup setenv LD_PRELOAD_64 /usr/lib/sparcv9/libumem.so eintragen.

Posted by Joerg Moellenkamp at 07:22

Was ist mit Copy-on-Write? Das sollte man bei ner Datenbank doch dringend abschalten, oder? Wie geht das?
Anonymous on Aug 27 2007, 10:12

Wenn die Blocksize der Datenbank identisch groesser der Blocksize des Filesystems ist, dann ist genau das kein problem. Du willst sogar genau das, weil dadurch auch das entgültige Schreiben auf die Platte linearisiert werden und nicht nur die Schreibzugriffe auf das Mysql-Log
Anonymous on Aug 27 2007, 11:03

also der Kollege hier (http://dimitrik.free.fr/db_STRESS_BMK_Part2_ZFS.html) erreicht nach vielen Optimierungen gerade mal 100% der UFS Performance auf einem ZFS (für seine Tests) - gibts da nen aktuelleren Stand? Da es auch noch ne Menge Bugs gibt und gab (stichwort aggressive ZFS Cache allocation eating up memory) ist ja die Frage ob man das schon "produktiv" einsetzen sollte.
Anonymous on Aug 27 2007, 11:34

Ich beantworte es mal so: Eine ganze Reihe unserer Kunden hat es bereits im Einsatz.

Zu dem Benchmark: Zum einen beruecksichtigen die Benchmarks noch nicht das separated ZIL, zum anderen sind die Werte im vergleich zu direct i/o zu sehen. Da aggressive Cache allocation kann man sehr gut unter kontrolle bringen. Siehe dazu auch solarisinternals.com wiki.
Anonymous on Aug 27 2007, 13:26

Ich gehe mal davon aus das es sich bei den 500GB Platten um SATA handelt. Viel hilft zwar bekannt viel, aber SATA ist IMHO für den Bereich nicht wirklich ausgelegt. SATA Platten hassen kleine random I/Os, wie sie ja nun mal gerade bei DBs vorkommen, wie die Pest. Dafür sind die Dinger auch nicht ausgelegt (Arme, Köpfe, Lager, Drehzahlen usw.). SATA macht Spaß bei langen sequentiellen I/Os, wie z.B. bei Backup-2-Disk, aber bei viel random I/O würde ich immer noch SCSI bzw. SAS Platten vorziehen, vorzugsweise 2,5", 15k UPM: Dreht schneller, kleinere Platten, kurze Seektime als 3,5" und von der Lebensdauer her stehen die den großen auch ins nichts nach - vor allem keiner SATA.

Just my 2 Cent.
Anonymous on Aug 29 2007, 12:03

Also, es gibt auch SATA-Platten in vernünftiger Qualität. Die Platten in der Thumper sind nicht Consumerqualität. Wir verwenden die Maschine durchaus für Datenbanken. Vornehmlich im Datawarehousing. Aber du hast recht, schwerlast/random-i/o wuerde ich da auch nicht draufpacken
Anonymous on Aug 29 2007, 18:35