PRESENTED BY RED HAT

LEARN. NETWORK.
EXPERIENCE OPEN SOURCE.

www.theredhatsummit.com

TUNING RED HAT ENTERPRISE LINUX FOR ORACLE & ORACLE RAC

Scott Croft
Senior Consultant, Red Hat, Inc.
June 24th, 2010

Agenda

- OS installation and kickstart
- Network configuration (with bonding)
- Kernel tuning

OS installation and kickstart

- Required packages
  - Base X
  - Binutils, compat-db, compat-gcc-34, compat-gcc-34-c++, compat-libstdc++-33, elfutils-libelf-devel, gdd, gdd-c++, gdb, gdbm, glibc, glibc-common, glibc-devel, ksh, libXp, libXtst, libaio, libaio-devel, libgcc, libgnome, libstdc++, libstdc++-devel, make, setarch, sysstat, unixODBC, unixODBC-devel, util-linux, xorg-x11-xinit, compat-libstdc++-296
  - 32 bit packages for 64 bit installation: glibc-devel, libaio, glibc, libgcc, compat-libstdc++, openssl, libXp, libXtst

OS installation and kickstart: filesystem layout

- OS filesystem layout
  - Compartmented or one large /
  - Swap

OS installation and kickstart: %post

- Automatic network configuration
  - HOSTNAME=$(grep HOSTNAME /etc/sysconfig/network|awk -F= '{print $2}')
  - hostname $HOSTNAME
  - GATEWAY=$(ip route list |grep default |awk '{print $3}')
  - echo "GATEWAY=$GATEWAY" >> /etc/sysconfig/network
  - DEFDEV=$(ip route list|grep default|awk '{FS=" "; print $5}')
  - IPADDR=$(ip addr show $DEFDEV |grep inet |grep -v inet6|awk '{print $2}'|awk -F/ '{print $1}')
  - echo "IPADDR=$IPADDR" >> /etc/sysconfig/network-scripts/ifcfg-$DEFDEV
  - sed -i 's/dhcp/static/' /etc/sysconfig/network-scripts/ifcfg-$DEFDEV
  - BCAST=$(ip addr show eth0 |grep inet |grep -v inet6|awk '{print $4}')
  - echo "BROADCAST=$BCAST" >> /etc/sysconfig/network-scripts/ifcfg-$DEFDEV
  - echo "NETMASK=255.255.255.0" >> /etc/sysconfig/network-scripts/ifcfg-$DEFDEV

OS installation and kickstart: %post

- Memory calculation – no core file generation
  - mem=$(free|grep Mem|awk '{print$2}')
  - totmem=$(echo "$mem*1024"|bc)
  - huge=$(grep Hugepagesize /proc/meminfo|awk '{print $2}')
  - max=$(echo "$totmem*75/100"|bc)
  - all=$(echo "$max/$huge"|bc)
  - echo "kernel.shmmax = $max" >> /etc/sysctl.conf
  - echo "kernel.shmall = $all" >> /etc/sysctl.conf

OS installation and kickstart: %post

- Memory calculation – proper core file generation
  - shmmax is set to 4GB – 1 byte or 4294967295
  - shmall is set to shmmax / (getconf PAGE_SIZE) or 1048575

OS installation and kickstart: %post

- Oracle user parameters: /etc/security/limits.conf
  - Open file descriptors for the Oracle user
    - oracle soft nproc 131072
    - oracle hard nproc 131072
  - Per process locked memory for the Oracle process: specified for the use of hugepages. Stated in KB, but matches the number of hugepages allocated in sysctl.conf
    - oracle soft memlock 5000000
    - oracle hard memlock 5000000

Network configuration and bonding

- Public vs private networks
  - Bonding
    - Mode 0 – active/active – load balance, round-robin
    - Mode 1 – active/passive
    - Mode 4 – 802.3ad link aggregation
    - Mode 6 – balance-alb (adaptive load balancing)

Kernel tuning

- Network tuning
  - Receive socket buffer size
    - net.core.rmem_default=262144
    - net.core.rmem_max=4194304
  - Send socket buffer size
    - net.core.wmem_default=262144
    - net.core.wmem_max=4194304
  - TCP socket buffer size
    - net.ipv4.tcp_rmem=4096 262144 4194304
    - net.ipv4.tcp_wmem=4096 262144 4194304
  - Network port range
    - net.ipv4.ip_local_port_range=1024 65000

Kernel tuning

- Performance for Oracle
  - How often to send keep alive packets when a connection is unused
    - net.ipv4.tcp_keepalive_time=30
  - How long the kernel waits in between probes
    - net.ipv4.tcp_keepalive_intvl=60
  - How many probes are sent before a connection is considered broken
    - net.ipv4.tcp_keepalive_probes=9
  - How many times to retry before killing the connection
    - net.ipv4.tcp_retries2=3
  - How many times to retry transmitting the syn packet
    - net.ipv4.tcp_syn_retries=2

Kernel tuning

- Memory settings
  - Swapping for Oracle is bad.....
    - vm.swappiness=0
  - Maximum percentage of active memory that can have dirty pages
    - vm.dirty_background_ratio=3
  - Maximum percentage of total memory that can have dirty pages
    - vm.dirty_ratio=15
  - How long data can be in page cache before being expired (hundreths of a second)
    - vm.dirty_expire_centisecs=500
  - How often pdflush is activated to clean dirty pages (hundreths of a second)
    - vm.dirty_writeback_centisecs=100

Kernel tuning: Hugepages

- Huge pages
  - Obtain the Hugepagesize from /proc/meminfo
  - Setup the oracle user to be able to use hugepages in /etc/sysctl.conf
    - vm.hugetlb_shm_group=`id -g oracle`
  - Memlock is calculated by the number of huge pages allocated * 1024 * 2
  - Huge pages are used automagically when set correctly and can be calculated 2 ways. Before the database is setup:
    - (SGA+PGA+(20k * # of Oracle processes running)) / 2MB
      - 20GB SGA + 10GB PGA + 1k of oracle processes / 2MB
      - vm.nr_hugepages=15369
      -
    - After the database is running, by executing the script in appendix A

Kernel tuning: shared memory

- Shared memory – no core file generation
  - Obtain the total memory from the system
    - mem=$(free|grep Mem|awk '{print$2}')
  - Convert the value of $mem to bytes
    - totmem=$(echo "$mem*1024"|bc)
  - Get the Hugepagesize from /proc/meminfo
    - huge=$(grep Hugepagesize /proc/meminfo|awk '{print $2}')
  - Calculate what 75% of the total memory on the system for SHMMAX
    - max=$(echo "$totmem*75/100"|bc)
  - Divide the SHMMAX value by the Hugepagesize to get SHMALL
    - all=$(echo "$max/$huge"|bc)
  - Set the SHMMAX value in the /etc/sysctl.conf file
    - echo "kernel.shmmax = $max" >> /etc/sysctl.conf
  - Set the SHMALL value in the /etc/sysctl.conf file
    - echo "kernel.shmall = $all" >> /etc/sysctl.conf

Kernel tuning: shared memory

- Setting the maximum number of shared memory segments with SHMMNI.
  - SHMMNI=4096 in /etc/sysctl.conf, kernel.shmmni=4096

Kernel tuning: semaphores

- SEMMSL is the maximum number of semaphores per semaphore set, 250
- SEMMNI defines the number of entire semaphore sets for the system, 142
- SEMMNS defines the total number of semaphores for the system, 32000
- SEMOPM defines the maximum number of semaphore operations per semaphore call, 100
- In /etc/sysctl.conf the entry for semphores is kernel.sem="250 32000 100 142"

References

- Oracle 10g Server on Red Hat Enterprise LinuxÂ® 5: Deployment Recommendations
- Tuning and Optimizing Red Hat Enterprise Linux for Oracle 9i and 10g Databases
- Ipsysctl tutorial 1.0.4

Appendix A

```
#!/bin/bash

KERN=`uname -r | awk -F. '{ printf("%d.%d\n",$1,$2); }'`

# Find out the HugePage size

HPG_SZ=`grep Hugepagesize /proc/meminfo | awk '{print $2}'`

# Start from 1 pages to be on the safe side and guarantee 1 free HugePage

NUM_PG=1

# Cumulative number of pages required to handle the running shared memory segments

for SEG_BYTES in `ipcs -m | awk '{print $5}' | grep "[0-9][0-9]*"`

do

  MIN_PG=`echo "$SEG_BYTES/($HPG_SZ*1024)" | bc -q`

  if [ $MIN_PG -gt 0 ]; then

    NUM_PG=`echo "$NUM_PG+$MIN_PG+1" | bc -q`

  fi

done

# Finish with results

case $KERN in

  '2.4') HUGETLB_POOL=`echo "$NUM_PG*$HPG_SZ/1024" | bc -q`;

       echo "Recommended setting: vm.hugetlb_pool = $HUGETLB_POOL" ;;

  '2.6') echo "Recommended setting: vm.nr_hugepages = $NUM_PG" ;;

   *) echo "Unrecognized kernel version $KERN. Exiting." ;;

esac
```


- Questions?

FOLLOW US ON TWITTER

www.twitter.com/redhatsummit

TWEET ABOUT IT

#summitjbw

READ THE BLOG

http://summitblog.redhat.com/