debian

Debian GNU/Linux kernel tuning para cargas extremas de I/O de disco e rede para SGBD e e-mail

debian

*Cenário deste caso de uso:
  *Virtualização XenServer
  *Debian GNU/Linux stable
  *SAN e Data storage server **muito** alto desempenho IOPS.
  *Multipath, LVM, sistema de arquivos tunados para arquivos pequenos, acessos paralelos, escritas síncronas, acesso aleatório.
  *IPv4
  *Servidores com suporte a virtualização no hw.
  *Produção de missão crítica é conservadora.

debian

*Como usuários percebem os sintomas:
  *Grande demora nas respostas a cliques/comandos em certos horários.
  *Aparente baixa carga de CPU.
  *"Faísca atrasada"; "reloginho" mesmo para respostas pequenas e simples.
  *"Depois que começa, vai".
  *"Baixar uma ISO leva 15 segundos. Baixar 1 KB também leva 15 segundos."

debian

*O que ganho resolvendo o problema de solução composta:
  *Respostas **até 84 vezes** mais rápidas no cenário.
  *Sistema mais "**rápido**" e "**ágil**".
  *Sistema **resiliente**.
  *Mas pode **reduzir** velocidade de transferência.
  ***Eficiência** da infraestrutura e menores custos.
  ***Maximização** de desempenho.
  *Boa experiência de usuário.
  *Usuários **satisfeitos**.

*Identificando o problema real.
  *Camada de virtualização
  *Sistema operacional convidado

debian

*Identificando o problema real.
  *Camada de virtualização
    *__top__ mostra elevado software interrupts (__%si__) : __spinlocks__
    *__sar -q__ mostra elevado __runq-sz__ : enfileirando processos
    *__xentop__ mostra pacotes de rede perdidos
    *__ifconfig__ no dom0 mostra pacotes de rede perdidos nas vif
    */__proc/interrupts__ mostra quase todas interrupções executadas num dos núcleos.

debian

*Identificando o problema real.
  *Sistema operacional convidado
    *__top__ mostra elevadíssimo __%wa__ (iowait)
    *__iostat -dmxthN__ mostra elevadíssimo __%util__
    *__tcptrace__ sobre dados do __tcpdump__ mostram conexões muito demoradas, vários segundos para transmitir poucos bytes, velocidade baixa, idle time alto.

debian

*Conceitos da solução composta.
  *Pouco importa a velocidade. O foco é na **latência**.
  *Latência.
  *Latência.
  *Latência.
  *Latência.
  *Latência.

debian

*Camada de virtualização

*Todo I/O (rede e disco) das VM é feito por sw, visível no dom0.

*Existe um desbalanceamento de sw irq e 1 núcleo não é tão poderoso para dar conta das cargas de i/o.

*Muitos chaveamentos de contexto também.

debian

*Sistema operacional convidado

*Deixar o problema de organizar e agrupar blocos para o ZFS data storage server.

*Enviar os blocos a serem escritos rapidamente ao ZFS data storage server.

*flush de block buffers o mais rápido viável

*reduzir a probabilidade de usar swap

*reaproveitar sockets tcp rapidamente, pois o tempo default TIME_WAIT é alto.

debian

*Escolhendo parâmetros para este cenário.
*Implantação da solução.
  *Camada de virtualização
    *__Nunca__ configurar número total de vcpu de todas VM maior que número de __cpu – 1__.
      *__Esqueça__ HT ou equivalente. __Conte núcleos REAIS__.
    *Instalar no dom0 e domU um daemon __irqbalance__ para reequilibrar dinamicamente as sw irq.
    *Não vale a pena cpu pinning. Perde flexibilidade gestão.
    *Discos virtuais para dados com alto IOPS em tipo **RAW LVM**.
      *só linha de comando.
    *Xen: esqueça snapshots, fast clones, clones: apenas **full copy**.
      *implemente esses no lado do __ZFS data storage server__.
    *Discos virtuais para baixo IOPS podem continuar em tipo VHD.

debian

*Adicionar ao /etc/sysfs.conf (sysfsutils)

```
#AFM 20120523

block/xvdb/queue/nr_requests = 4

block/xvdb/queue/scheduler = deadline

block/xvdb/queue/iosched/front_merges = 1

block/xvdb/queue/iosched/fifo_batch = 1
```

debian

*Adicionar ao /etc/sysctl.conf

```
vm.swappiness = 10
vm.dirty_background_ratio = 1
vm.dirty_expire_centisecs = 500
vm.dirty_ratio = 15
vm.dirty_writeback_centisecs = 100
fs.file-max = 2048000
kernel.shmmax = 68747619736
kernel.shmall = 4294967296

net.ipv4.tcp_tw_reuse = 1
net.ipv4.tcp_low_latency = 1
net.ipv4.tcp_keepalive_time = 1800
net.ipv4.tcp_max_syn_backlog = 1024
net.ipv4.ip_local_port_range = 15000 61000
net.ipv4.tcp_fin_timeout = 10
```

incluir no /etc/security/limits.conf

```
#*       soft  nofile 5000000
#*       hard  nofile 5000000
@adm soft nofile 1024000
@adm hard nofile 1024000
* soft nofile 1024000
* hard nofile 1024000
```

incluir no /etc/pam.d/common-account

```
#AFM 20110915
session required pam_limits.so
```

incluir no /etc/profile

```
#AFM 20110916
ulimit -n 1024000
```

http://wiki.debian.org/WhyDebianForDevelopers
http://wiki.debian.org/DebianForNonCoderContributors
http://wiki.debian.org/PkgSplit
http://www.debian-rs.org
http://www.debianbrasil.org

André Felipe Machado <andremachado@techforce.com.br>
http://www.techforce.com.br

20130716


debian