Construindo seu próprio storage usando ZFS

27/11/2014

Fernando Frediani

nic.br
Núcleo de Informação e Coordenação do Ponto BR

cgi.br
Comitê Gestor da Internet no Brasil

Disclaimer

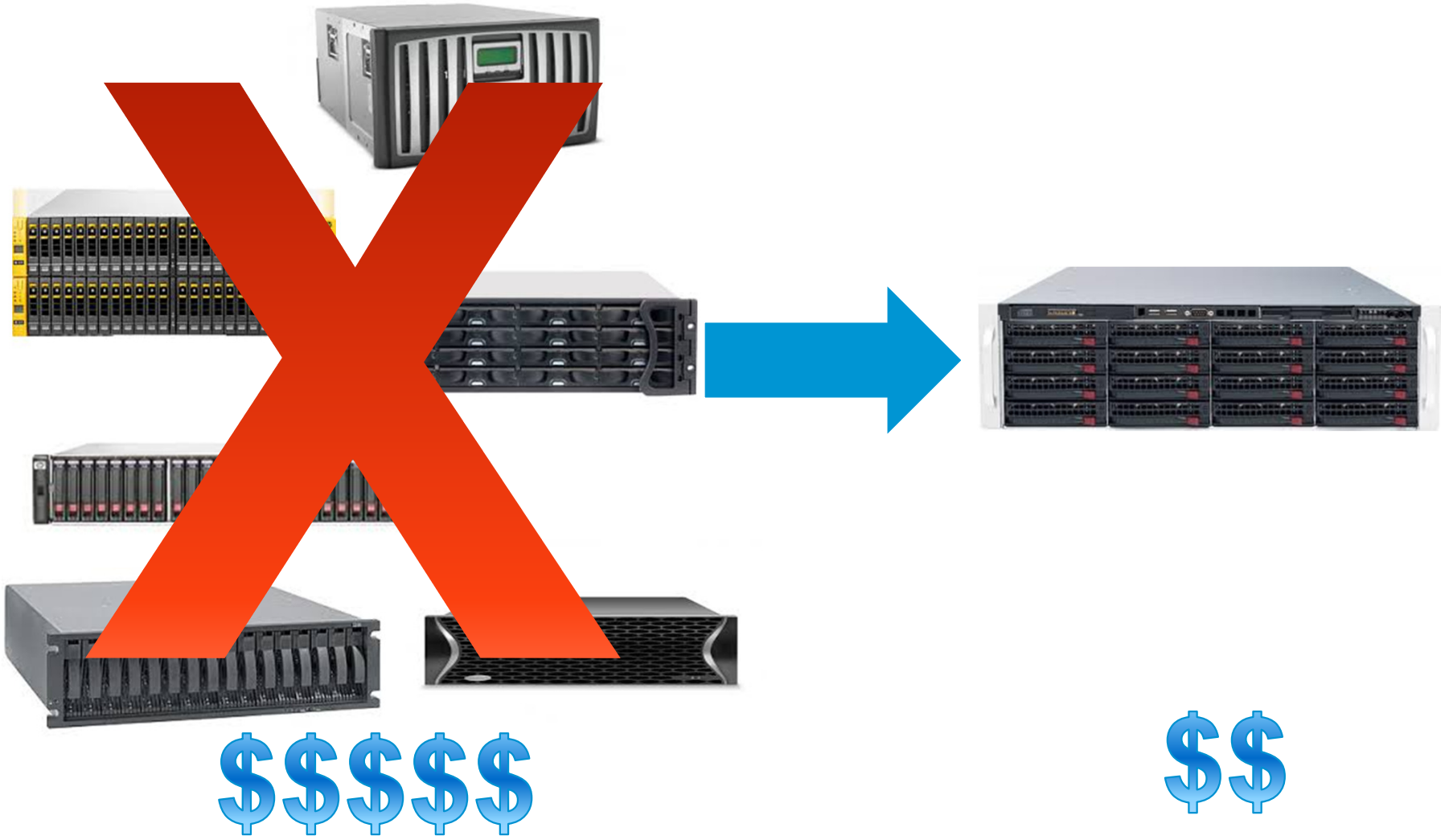
- Eu não tenho qualquer relação com qualquer um dos fabricantes mencionados e não recebo quaisquer benefícios deles.

- O conteúdo mencionado nesta apresentação é baseado na minha visão pessoal adquirida durante pesquisa, desenvolvimento e uso dessas tecnologias ao longo dos anos.

- Eu não garanto que as informações apresentadas aqui são livres de erros e não posso me responsabilizar por perdas e danos causados pelo seu uso. Portanto eu fortemente recomendo não utiliza-las em ambientes importantes para a empresa e/ou sem cópia de segurança dos dados.

Introdução

Por que construir seu próprio storage ?



Introdução

- Storage hoje em dia, em muitos casos, por ser responsável por até 50% dos custos de uma plataforma.

- Nem todo dado precisa ser armazenado em um storage “Enterprise” Super-Rápido e de Altíssima Disponibilidade.

- Você pode construir seu próprio Storage compartilhado usando componentes de baixo custo, tecnologia livre e gratuita, ainda sim mantendo um ótimo nível de confiabilidade e performance.

- O objetivo desta apresentação **não é** desencorajar o uso de Storage Enterprise ou tradicionalmente usado, porém de mostrar uma alternativa para se balancear aonde cada tipo de dado pode ser armazenado.

Resumo

A receita

- Especificações do Servidor
- Hard Drives
- Solid State Drives (SSDs)
- Controladores RAID
- Placas de Rede
- Cabeamento
- Sistemas Operacionais
- Sistema de Arquivos (ZFS)
- Sistema de Gerenciamento
- Extras (JBOD para Expansão , High-Availbiliy, etc)

Resumo

- **Possíveis casos de uso**
  - ISO / Repositório de Templates
  - Datastore para Maquinas Virtuais em ambientes LAB / Testing.
  - Storage para Sistemas de Backup.
  - Armazenar grande quantidades de dados que não necessitam estar online 24/7.
  - Para sistemas de alta performance (centenas de milhares de IOPS)

- **Pontos fracos**
  - “Single controlled” – Downtime é necessário para atualização do sistema.
  - Downtime prolongado se o hardware falhar.
  - Necessita de conhecimento mais especializado.
  - Suporte de hardware mais restrito.
  - Dificuldade em obter suporte se o problema for relacionado a storage (ex: VMware).

Servidor

- Servidores 3U ou 4U with 16 ou 24 x 3.5” disk slots.
- Placa mãe (1-2 sockets)
- 1-2 Quad-core CPU dependendo da quantidade de “live data”.*
- Mínimo de 24-32 GB de Memória um bom cache**(quanto mais melhor).
- 2 x 2.5” discos internos para o Sistema Operacional.
- 2 x Fontes de Alimentação

* Para funcionalidades como NFS e/ou compressão de grandes quantidades de dados é recomendado utilizar 2 CPUs.
** Se utilizar deduplication, memória extra deve ser considerada.

Hard drives

- **Enteprise class**
  - SAS or SATA
    - Seagate Constell ES.2, Hitachi Ultrastar, etc
  - 7200 RPM, 10.000 RPM, 15.000 RPM
  - NÃO UTILIZAR Green Drives.
  - Single or Dual port
  - Evite misturar SAS e SATA
- **SSDs**
  - MLC – Cache de Leitura (ZFS L2ARC)
  - SLC e eMLC– Para Escrita (ZFS ZIL)
  - DRAM based (STEC ZeusRAM, DDR Drive) (ZFS ZIL)

Controladores RAID

- Os controladores RAID devem ser capazes de operar em modo JBOD e permitir ao ZFS o controle direto dos discos – IT-mode (Sem hardware RAID).

- Ser bem suportado pelo sistema operacional.

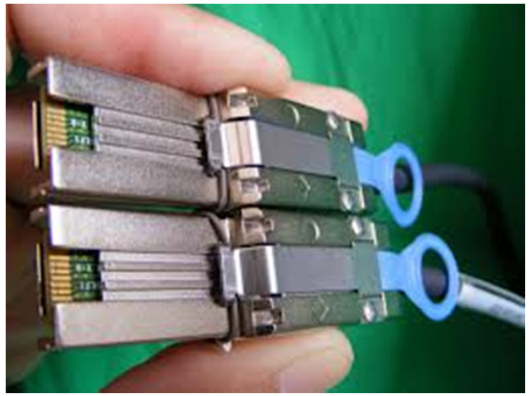
- Não é necessário utilizar controladores com bateria.

Controladores RAID

- **LSI 9211-8i**
  - 2 conectores Mini-SAS (1 para os discos internos de 2.5”, 1 para o backplane).
  - Fusion MPT 2.0 IO Controller
  - 600 MB/s bandwidth per lane
  - Até 256 SAS/SATA hard drives
  - Mais de 320,000 IOPS

Controladores RAID

- **LSI 9200-8e**
  - 2 conectores Mini-SAS (SFF8088) (Para expansão utilizando JBODs adicionais).
  - Fusion MPT 2.0 IO Controller
  - 600 MB/s bandwidth per lane
  - Até 512 SAS/SATA hard drives
  - Mais de 320,000 IOPS



Placas de Rede

- Placas de rede adicionais são recomendadas para banda extra e separação de rede.
- Configuração típica (2 portas onboard para Gerencimento + 2 portas adicionais para o tráfego de Storage)
- Placas Intel são conhecidas por terem ótimo suporte e estabilidade (82540EM, 82574L, 82575EB, 82576)
- Portas 10 Gigabit - Intel and Emulex funcionam bem.
- **Redundância de Switch**
  - Usar port-channel ou IPMP (Similar ao Bonding do Linux).

Cabeamento

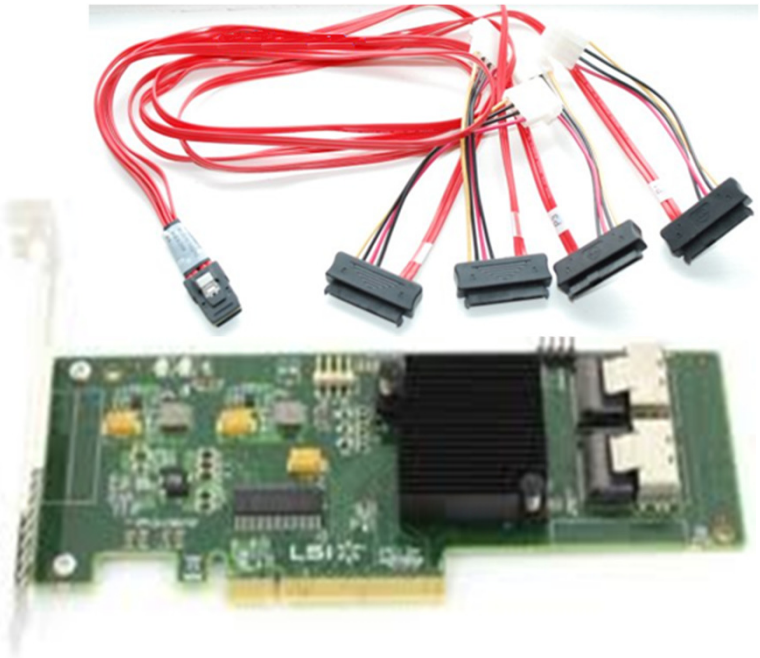
Discos Internos – 2.5”

Sistema Operacional

L2ARC

ZIL

SFF-8087 to 4 x SFF-8482



*** Caso utilizar as portas SATA Onboard, configura-las em modo AHCI.

Cabeamento

Interno - Backplane

Sistemas Operacionais

3 escolhas de Sistemas Operacionais

- OpenIndiana (Open Source, CLI ou GUI com Time Slider) – Baseado em IllumOS.
- OmniOs (instalacao mínima servidor) – Baseado em IllumOS.
- Oracle Solaris Express 11 (Sem uso comercial).

Sistema de Arquivos - ZFS

- **Built in RAID**
- **Ótima integridade dos dados**
- **Online scrub (“similar ao fsck”)**
- **Snapshots ilimitados**
  - Send/Receive – Async remote replication
- **ARC, L2ARC and ZIL**
- **Modelo Copy-on-write**
- **Deduplication**
- **Encryption**
- **Compressão**
- **Expansão Online**
- **Limite impossível de ser atingido (128 bit filesystem)**
- **Etc…**

Sistema de Arquivos - ZFS

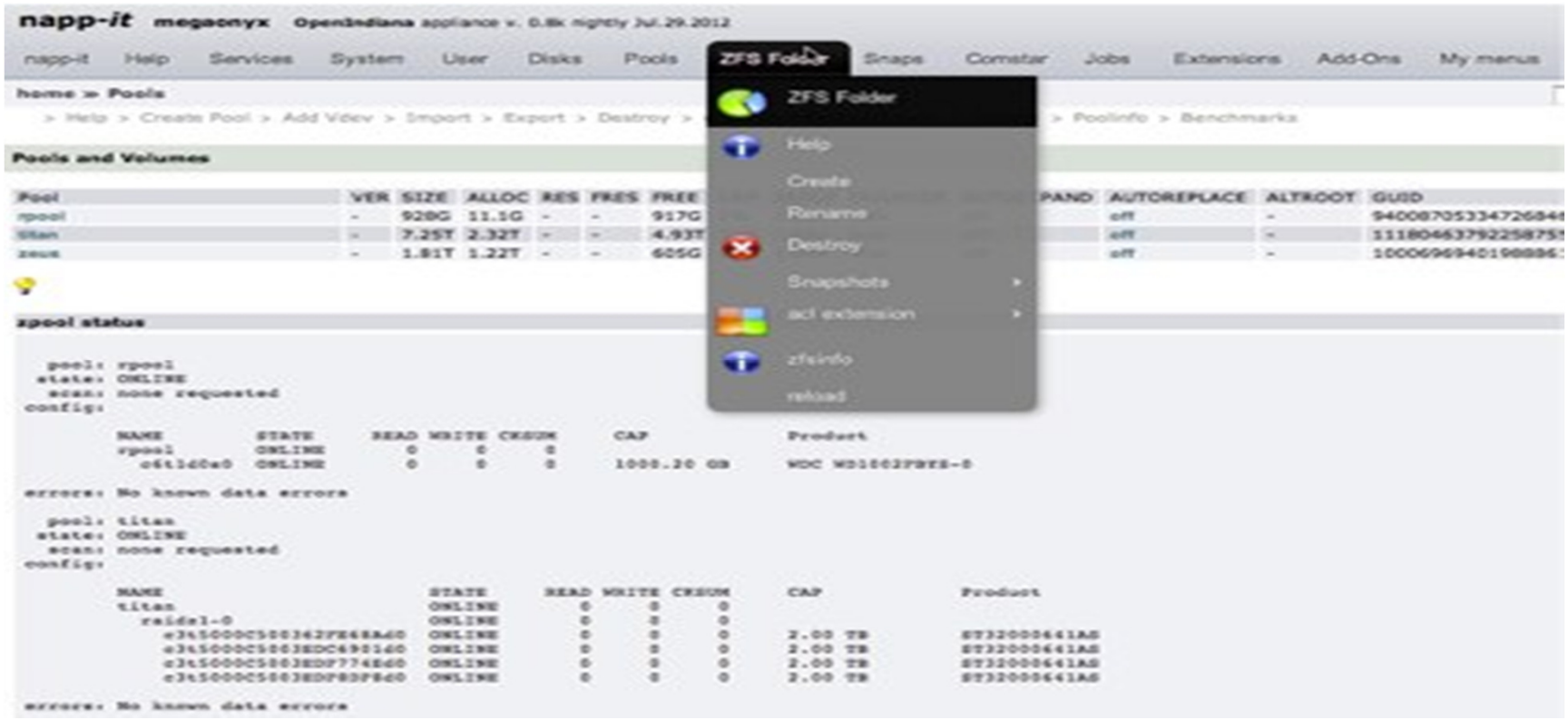
- **Uma pool ZFS (zpool) é constituída de um ou mais "vdev's".**
- **Um "vdev" tradicionalmente é constituído de vários hard drives em RAID ou em Mirrors (RAID 1).**
- **Modos ZFS RAID**
  - Non-redundant (similar ao RAID 0)
  - Mirrors (Similar ao RAID1 – Vários vdevs juntos similar ao RAID 10)
  - RAIDZ – (Similar ao RAID 5) (Não recomendado para mais de 4 discos)
  - RAIDZ2 – (Similar ao RAID 6)
  - RAIDZ3 – (Similar ao RAID 6 porém com 3 discos usados para paridade)
- **"zvols"(iSCSI, FibreChannel) e Shares (CIFS, NFS, WebDAV, Rsync) são criados em cima da pool ZFS.**

Sistema de Arquivos - ZFS

- **ZVOLs**
  - Um “block device” virtual que pode ser exportado via iSCSI ou Fibre Channel usando COMSTAR.
  - Pode ser Thin-provisioned.
  - Tamanho do bloco pode ser variável.
  - Compressão (LZJB or LZ4) pode melhorar performance.
  - Deduplicacão (Usar com cuidado)
- **Shares**
  - Um diretório em cima da pool ZFS que pode ser exportado via NFS, CIFS, WebDAV, Rsync, etc
  - Funcionalidades similares ao ZVOL

Sistema de Gerenciamento

- **Napp-it - http://www.napp-it.org**
  - Interface Web bastante funcional para gerenciar ZFS.
  - Extensões Non-free disponíveis (Remote Replication, ACL e Gerenciamento Avançado de Usuários, Monitoramento) – Custo pequeno para o desenvolvimento do projeto.



Extras

- **JBOD para Expansao**
  - Não necessita de CPU, Motherboard, Memória, etc.
  - Mais barato do que um chassis convencional.
  - Consumo menor de energia.
  - Fácil de expandir (expansão online através de cascading ou stacking).

- **High availability**
  - Não é fácil. Possível utilizando PaceMaker/Heartbeat.
  - Necessita utilizar ‘dual port’ hard drives (sem SATA).
  - Chassis 1U ou 2U podem ser utilizados como controladores.
  - Sem discos para pool ZFS nos nodes controladores. Somente para o SO.

Extras

- **Sistema All-in-one**
  - ESXi + ZFS no mesmo servidor físico.
  - Appliance virtual ZFS Storage em cima do ESXi
  - Exporta via NFS or iSCSI

  - **Pré-requisitos**
  - Servidor capaz de VT-d.
  - PCI-passthrough para o controlador RAID ao Appliance virtual.
  - VMXNET3 para melhor performance interna.

  - http://www.napp-it.org/doc/downloads/all-in-one.pdf

Extras

- **Monitoramento**
  - SNMP - Estatísticas
  - Nagios - nrpe plugin
  - IPMI – para monitoramento extra de hardware.
  - S.M.A.R.T

- **Supporte**
  - Opeindiana é suportado pela comunidade de usuários e desenvolvedores.
    - http://openindiana.org/support/
    - http://wiki.openindiana.org/oi/OpenIndiana+Wiki+Home
  - OmniOS possui suporte comercial da OmniTI
    - http://omniti.com/does/omnios
  - Solaris 11.1 possui suporte comercial da Oracle
    - Conferir se o hardware é compatível com a HCL

Extras

- **ZFS on Linux -** http://zfsonlinux.org/
  - ZFS está sendo portado para Linux pelo Lawrence Livermore National Laboratory (LLNL)
  - Disponível na maioria das distribuições mais utilizadas.
  - Ainda não recomendado para produção.
  - Ainda não otimizado para performance
  - Vale a pena dar uma olhada e seguir.

- **ZFS no FreeBSD**
  - Disponível e estável
  - FreeNAS é o projeto mais popular relacionado - http://www.freenas.org/
    - Roda de um USB stick
  - Performance é boa o suficiente, porém não como nos Sistemas Operacionais derivados de Solaris.

Exemplos de Configuração

- Configuração 1 - Pequeno

Chassis	Supermicro 3U
CPUs	1 Quad-core
Memory	24GB
Disks	16 x 1TB SATA
SSDs	2 x 40GB Intel SATA (OS), 1 x 240GB Intel MLC (L2ARC), 1 x SLC or eMLC (ZIL)
ZFS Pool	2 x RAIZ2
Network	1 x Dual port NIC
RAID Controller	1 x LSI 9211-8i
Extras	None
Usable space	12TB
Total cost	**~£4100 - (R$ 25.000)**

Exemplos de Configuração

- Configuração 2 - Médio

Chassis	Supermicro 4U
CPUs	1 Hex-core
Memory	96GB
Disks	22 x 2TB SAS
SSDs	2 x 300GB SAS (OS), 1 x Samsung SM1625 400GB(L2ARC), 1 x 8GB STEC ZeusRAM (ZIL)
ZFS Pool	11 x Mirrors
Network	1 x Quad port NIC
RAID Controller	1 x LSI 9211-8i
Extras	1 x LSI 9200-8e (for future expansion)
Usable space	22TB
Total cost	**~£9000 – (R$ 50.000)**

Exemplos de Configuração

- Configuração 3 - Grande

Chassis	Supermicro 4U
CPUs	2 Hex-core
Memory	192GB
Disks	66 x 2TB SAS
SSDs	2 x 300GB SAS (OS), 1 x Intel 910 800GB PCI-e (L2ARC), 2 x 8GB STEC ZeusRAM (ZIL)
ZFS Pool	8 x RAIDZ2
Network	1 x Dual 10Gb Nic
RAID Controller	1 x LSI 9211-8i
Extras	1 x LSI 9200-8e, 2 x 4U Expansion JBOD, 2 Hot spare disks
Usable space	96TB
Total cost	**~£24750 – (R$ 100.000)**

Recomendações

- Checar Hardware Compatibility List para o SO escolhido.
  - http://wiki.openindiana.org/display/oi/Community+HCL
  - http://illumos.org/hcl/
  - http://www.oracle.com/webfolder/technetwork/hcl/data/s11ga/index.html
- Mantenha sempre componentes reservas no local – É mais barato do que suporte de hardware premium.
  - CPUs e Motherboard
  - Controlador RAID
  - Discos
  - Cabos
- Coloque o máximo possível de RAM.
- Utilize um número razoável de ‘vdevs’ para uma maior performance.
- Se a zpool chegar a perto de 80% da capacidade, aumente-a para evitar problemas de performance.

Perguntas ?

Obrigado

Contato: fhfrediani@gmail.com