Virtualização do Armazenamento

Paulo Pires
Maj Cav, Eng Info
Centro de Informação Geoespacial do Exército
ppires@igeoe.pt

Artigo integrado no seguinte conjunto:
Virtualização
2.1 Evolução histórica (publicado)
2.2 Virtualização de Servidores (publicado)
2.2.1 VMware (VMware, Inc.)
2.2.2 Xen (Citrix Systems, Inc.)
2.2.3 Integrity VM (HP)
2.2.4 Hyper-V (Microsoft)
2.3 Virtualização do Armazenamento
2.3.1 Armazenamento em ambientes virtualizados
2.4 Virtualização da Rede

Resumo

O HP CloudSystem Matrix (CSM) faz parte de uma pilha de software HP para computação na cloud que cobre todos os níveis de serviço considerados relevantes: IaaS (Infra-estrutura como Serviço), PaaS (Plataforma como Serviço) e SaaS (Software como Serviço). Apesar de ser a base desta pilha, i.e., oferecer o nível IaaS, é um produto extremamente complexo pois interage com todas as infra-estruturas: as computacionais (i.e., servidores físicos ou virtuais), as de armazenamento (do disco interno aos discos em servidores de armazenamento), e as de interligação (redes Ethernet e FC).

Apesar de toda a complexidade da infra-estrutura, real e virtual, que gere, o CSM torna conceptualmente simples a entrega aos consumidores de infra-estruturas para suporte a aplicações: 1) o administrador define que recursos da infra-estrutura estão disponíveis para integrar a "oferta cloud"; 2) o arquitecto define templates para as arquitecturas que considera adequadas para necessidades dos consumidores (e.g., arquitectura 3-tier para uma solução ERP - Enterprise Resource Planning); e 3) o consumidor escolhe o template que melhor se ajusta às suas necessidades e efectua um pedido de aprovisionamento da infra-estrutura.

A interacção entre os diferentes interlocutores (1), (2), (3) e o CSM é fundamentalmente realizada sobre portais; contudo, especialmente no caso do consumidor, o portal disponibilizado pelo produto tem sido considerado como "complexo", por apresentar informação demasiado técnica, "rígido", por não poder ser customizado (por exemplo para suprimir a "informação demasiado técnica"), e "grosseiro" por não permitir a especificação mais fina das características da infra-estrutura que se quer aprovisionar (por exemplo, permite variar o número de CPUs e a quantidade de memória de um servidor, mas não permite escolher a tecnologia dos discos que se pretendem aprovisionar, e.g., SSD em vez de FC, 15K em vez de 10K rpm). Assim, torna-se fundamental a virtualização, em especial, a virtualização do armazenamento, com base num conjunto (extensível e configurável) de opções pré-definidas e em layouts customizáveis, definindo portais que se integram com o HP CloudSystem Matrix e que permitam aos utilizadores (consumidores) uma interacção não só mais simples, mas também mais versátil.

1. Virtualização do Armazenamento

Há muito que o armazenamento passou a ser um tópico de "primeira classe" tal como "a rede". Para tal contribuiu significativamente a introdução de tecnologias tais como o RAID (Redundant Array of Independent Disks) (Patterson, 1989) que promoveram o desempenho e/ou a tolerância a faltas a (comparativamente) baixo custo – note-se que no artigo original de Patterson o I significava Inexpensive. Com a introdução do RAID, rapidamente apareceram os armários de discos (disk arrays), externos ao "sistema" e, uma vez que o protocolo SCSI suportava a interligação de discos (Logical Units) a múltiplos adaptadores e, daí, a múltiplos sistemas, o conceito de rede de armazenamento aparece naturalmente.
Como se mostra na figura 1, o disco lógico aparece

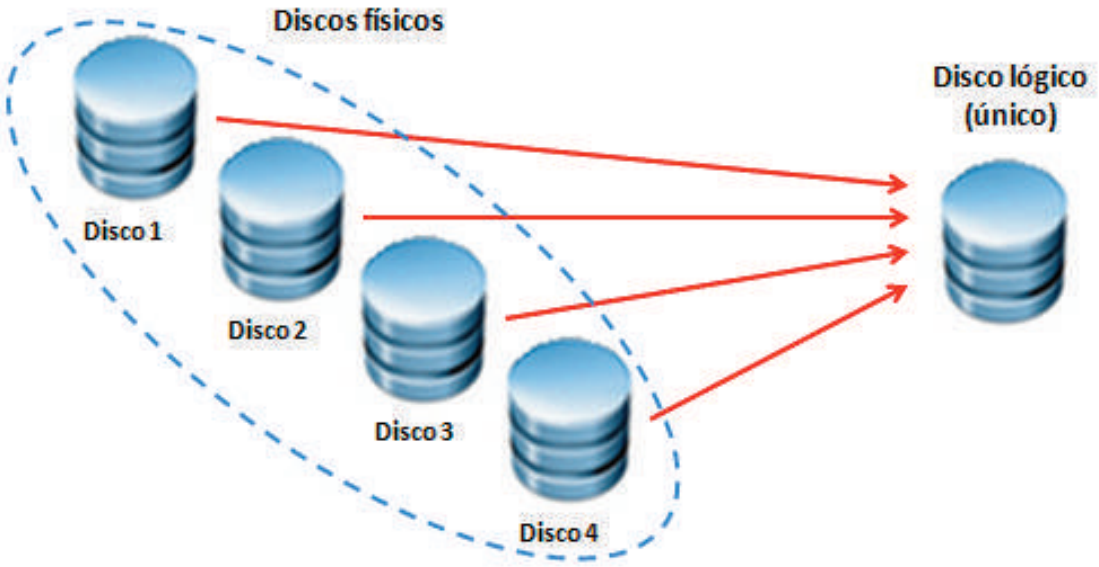

Figura 1: Redundant Array Independent Disks (RAID)

como uma forma elementar de virtualização, já que "aglutina" elementos de diferentes discos físicos mas é apresentado ao sistema como um único disco – ao ponto de ser impossível ao próprio sistema de operação distingui-lo de um disco físico.

Numa rede de armazenamento que interliga vários armários de discos e servidores (hosts), os primeiros oferecem volumes (ou discos lógicos) acessíveis por identificadores únicos e os segundos tomam posse desses volumes (tipicamente cada host tem uso exclusivo de um ou mais volumes) e formatam-nos, neles instalando sistemas de ficheiros (e.g., ext3, XFS, NTFS), que depois usam da forma mais conveniente. Este conceito de rede de armazenamento é conhecido como SAN – Storage Area Network, figura 2.

A forma mais comum de SAN usa uma infra-estrutura Fibre Channel (FC), o que significa que tanto hosts como disk arrays possuem interfaces FC, e na rede existem comutadores (switches) FC aos quais hosts e arrays (e eventualmente outros switches) se interligam. Uma outra tecnologia que pode ser usada em SANs é a Ethernet: neste caso hosts, disk arrays e switches têm interfaces Ethernet, e o protocolo de transporte usado é FCoE (Fibre Channel over Ethernet). Na figura seguinte mostra-se um ambiente que inclui simultaneamente uma outra tecnologia de "transporte", ISCSI e uma "zona" FC.

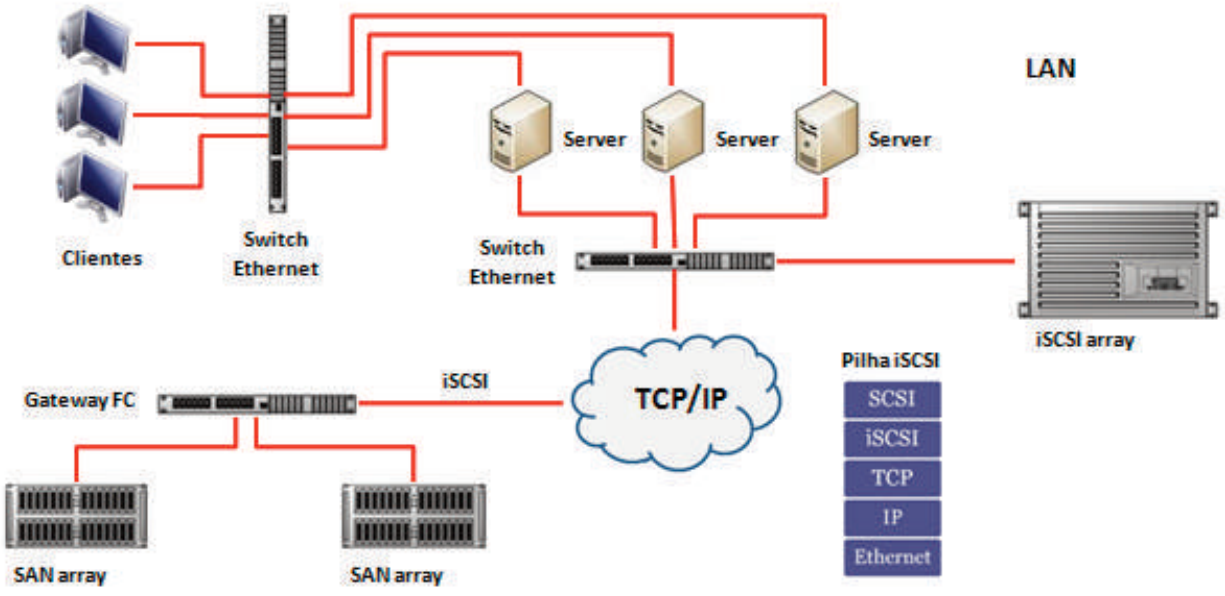

Figura 2: Storage Area Network (SAN)

Se bem que a relação volume/host de um-para-um seja de longe a mais frequente, também é possível estabelecer uma relação de um-para-muitos, na qual um volume é detido, de forma partilhada, por múltiplos hosts; tal situação implica necessariamente o uso de sistemas de ficheiros (SF) especializados, conhecidos como shared-disk file systems. Como exemplo indica-se o GFS (Whitehouse, 2007), GPFS (Schmuk, 2002), e VMFS (Vaghani, 2010), este último desenvolvido especificamente para armazenar VMs. Num SF para discos partilhados todos os hosts que partilham um dado volume têm uma visão coerente do estado do volume. Em situações de tolerância a faltas esta arquitectura de discos partilhados é muito apetecível pois em caso de falha de um host um dos restantes inicia uma recuperação (ao estilo de um sistema transacional) das últimas operações e rapidamente o SF regressa a um estado coerente.

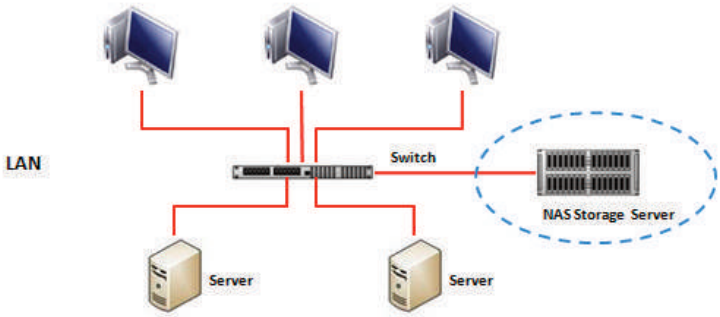

Figura 3: Network Attached Storage (NAS)

Uma outra forma, completamente distinta, de armazenar informação é usar um SF distribuído; neste caso, não há partilha de discos físicos, mas sim de ficheiros, sendo que um host pode deter apenas parte da árvore do SF (e.g., AFS) ou, nos SFs distribuídos com arquitectura cliente/servidor, um host, dito cliente dispõe de um módulo que lhe permite aceder ao SF remoto de um outro host, dito servidor (e.g., NFS e CIFS), como se o SF remoto fosse local.

2. Armazenamento em ambientes virtualizados

Quando se aborda o armazenamento em ambientes virtualizados, são de considerar os seguintes aspectos: a) como representar a arquitectura de uma VM propriamente dita (quanta memória e CPUs tem, qual o chipset utilizado, que adaptadores – de LAN, SAN, gráficos – possui, etc.); b) como representar um disco acessível à VM; e, c) como armazenar conteúdos (voláteis ou não) de componentes da VM, tais como a memória RAM e a memória não volátil, NVRAM, que contém o BIOS.

A solução adoptada pelo hipervisor VMware ESXi[1] é a seguinte: a) a representação da arquitectura é efectuada em XML que possui as informações de configuração; b) um disco acessível à VM é representado ou por um ficheiro que virtualiza o próprio disco, ou por um ficheiro que, ao estilo de uma ligação simbólica, descreve o caminho para um disco real; e finalmente, c) tais conteúdos são armazenados em ficheiros. Isto é, tudo é representado por, ou via, ficheiros.

A pergunta que se segue é óbvia: onde armazenar esses ficheiros? E a resposta, evidente: num repositório (datastore, na terminologia da VMware) gerido pelo hipervisor, onde uma VM aparece como uma pasta, e os seus componentes como ficheiros no interior dessa pasta, como se mostra na figura 4.

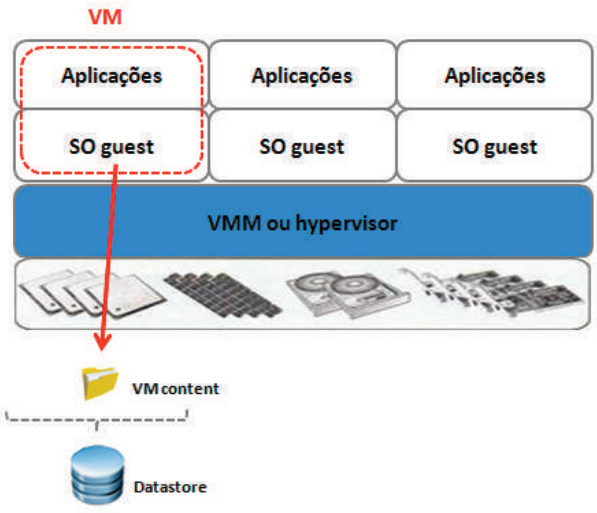

Figura 4: Datastore e VMs

O VMware ESXi suporta armazenamento em sistema de ficheiros VMFS (sobre discos) ou NFS; no caso do VMFS, os discos podem ser internos ao servidor, ou podem estar numa SAN.

[1] Para uma comparação entre o VMware ESXi e o Citrix Xen ver Anexo H

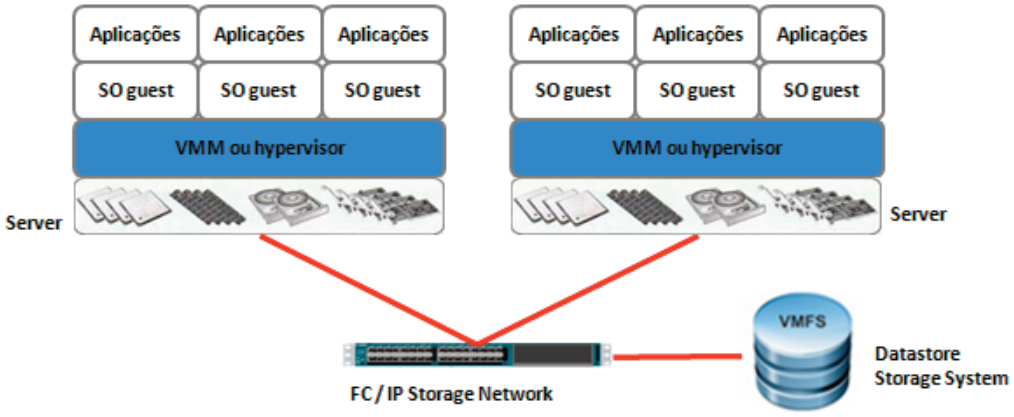

Figura 5: Sistema de Ficheiros VMFS

3. Sistema de Ficheiros VMFS

O VMFS (Virtual Machine File System) é um sistema de ficheiros em cluster, e foi optimizado para armazenar grandes ficheiros e realizar operações de entrada e saída (I/O) com que movimentam grandes volumes de dados. Os dois servidores, figura 5, têm, portanto, uma visão coerente do datastore, e das VMs nele existentes; por isso é, quando comparamos esta arquitectura com uma na qual cada servidor tem a sua datastore, muito mais simples e eficiente mover uma VM activa de um servidor para o outro, já que apenas as páginas modificadas residentes em memória têm de ser copiadas (o resto está no SF partilhado).

4. Network File System

Nos casos em que os datastores residem num sistema de ficheiros NFS (Pawlowski, 1994) os servidores ESXi podem também partilhar o mesmo datastore, como se vê na figura 6; assim, a movimentação de VMs entre servidores, por razões de equilíbrio de cargas ou de manutenção de um ou outro servidor, também são muito eficientes.

Para concluir, note-se que se abrem, nos ambientes virtualizados, duas oportunidades muito importantes quando um disco virtual é realizado sob a forma de ficheiro: a) a potencial poupança de espaço e b) a possibilidade de implementar com facilidade uma técnica de snapshots. No primeiro caso, conhecido como thin provisioning, referimo-nos à possibilidade de usar um ficheiro esparso para atribuir um espaço lógico de endereçamento ao disco (ficheiro) muito superior ao espaço de facto "consumido" pelo ficheiro no SF. No segundo, usando um mecanismo de versões baseado em técnicas de copy-on-write, armazenar snapshots da VM em determinados instantes e para usar como mecanismo de recuperação do estado, em caso de "falha".

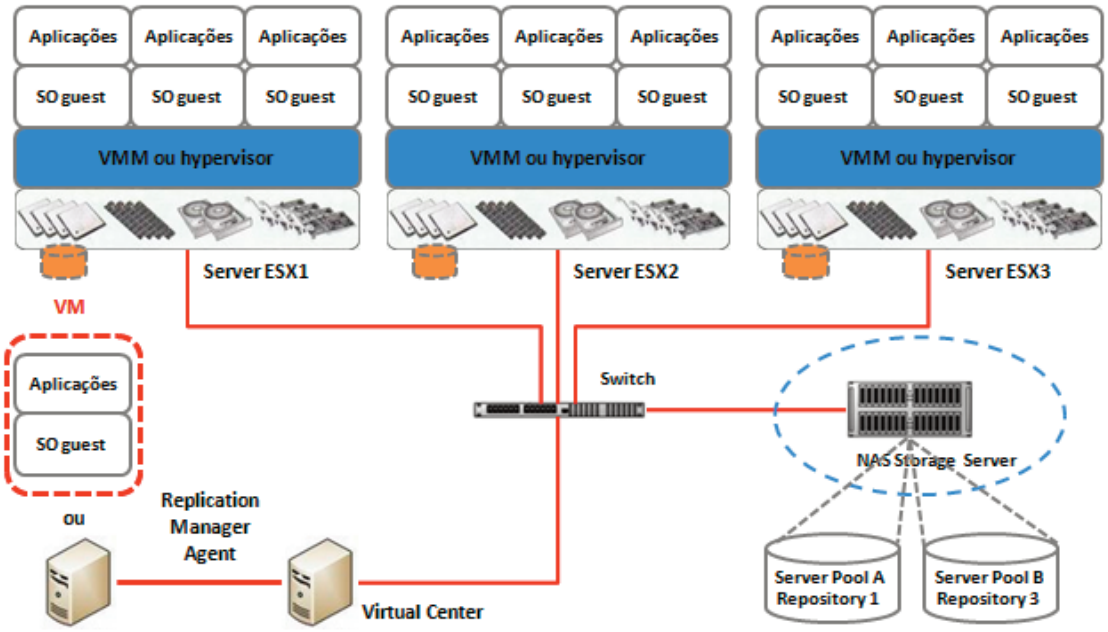

Figura 6: Network File System (NFS)

5. Referências bibliográficas

- Patterson, D. et al. A Case for Redundant Arrays of Inexpensive Disks (RAID). Proceedings of the 1989 ACM-SIGMOD International Conference on the Management of Data, ACM, 1989, pp. 109-116.
- Whitehouse, S. The GFS2 Filesystem. Proceedings of the Linux Symposium, June 27th-30th 2007, Ottawa, Canada.
- Schmuk, F. and Haskin, R. GPFS: A Shared-Disk File System for Large Computing Clusters. Proceedings of the Conference on File and Storage Technologies (FAST'02), 28–30 January 2002, Monterey, CA, pp. 231–244.
- Vaghani, S. Virtual Machine File System. ACM Operating Systems Review, Vol. 44, Number 4, December 2010, pp. 57-70.
- Pawlowski, B. et al. NFS version 3 design and implementation. Proceedings of the Summer USENIX Conference, June 1994, pp 137-152.