Universidade Federal do Rio de Janeiro

Instituto Tércio Pacitti de Aplicações e Pesquisas Computacionais

Alan Pacheco Estrella

ANÁLISE DAS REDES DE ARMAZENAMENTO: Suporte para Alta Disponibilidade de Servidores Virtuais

Rio de Janeiro

2013

Alan Pacheco Estrella

ANÁLISE DAS REDES DE ARMAZENAMENTO:

Suporte para Alta Disponibilidade de Servidores Virtuais

Monografia apresentada para obtenção do título de Especialista em Gerência de Redes de Computadores no Curso de Pós-Graduação Lato Sensu em Gerência de Redes de Computadores e Tecnologia Internet do Instituto Tércio Pacitti de Aplicações e Pesquisas Computacionais da Universidade Federal do Rio de Janeiro – NCE/UFRJ.

Orientador:

Moacyr Henrique Cruz de Azevedo, M.Sc., UFRJ, Brasil

Rio de Janeiro

2013

Alan Pacheco Estrella

ANÁLISE DAS REDES DE ARMAZENAMENTO:

Suporte para Alta Disponibilidade de Servidores Virtuais

Monografia apresentada para obtenção do título de Especialista em Gerência de Redes de Computadores no Curso de Pós-Graduação Lato Sensu em Gerência de Redes de Computadores e Tecnologia Internet do Instituto Tércio Pacitti de Aplicações e Pesquisas Computacionais da Universidade Federal do Rio de Janeiro – NCE/UFRJ.

Aprovada em março de 2013.

Moacyr Henrique Cruz de Azevedo, M.Sc., UFRJ, Brasil

AGRADECIMENTOS

Gostaria de agradecer aos meus pais que dedicaram suas vidas a minha criação e educação, a minha esposa que de forma carinhosa me deu força e coragem, ao meu filho que embora não tivesse conhecimento disto, iluminou de maneira especial os meus pensamentos me levando a buscar mais conhecimento. E queria agradecer também a todos aqueles que fizeram parte da minha vida e de alguma maneira me influenciaram para que eu chegasse a este momento, familiares, amigos e professores.

RESUMO

ESTRELLA, Alan Pacheco. **ANÁLISE DAS REDES DE ARMAZENAMENTO: Suporte para Alta Disponibilidade de Servidores Virtuais**. Monografia (Especialização em Gerência de Redes e Tecnologia Internet). Núcleo de Computação Eletrônica, Universidade Federal do Rio de Janeiro. Rio de Janeiro, 2013.

O presente trabalho tem por objetivo desenvolver uma visão das redes de armazenamento descrevendo as principais características das arquiteturas *Network Attached Storage* (NAS), *Storage Area Network* (SAN) e *Storage Area Network Internet Protocol* (SAN IP), realizando uma breve comparação entre elas a fim de auxiliar na definição da arquitetura da rede de armazenamento a ser implementada ao projetar um ambiente de servidores virtuais com alta disponibilidade.

ABSTRACT

ESTRELLA, Alan Pacheco. **ANÁLISE DAS REDES DE ARMAZENAMENTO: Suporte para Alta Disponibilidade de Servidores Virtuais**. Monografia (Especialização em Gerência de Redes e Tecnologia Internet). Núcleo de Computação Eletrônica, Universidade Federal do Rio de Janeiro. Rio de Janeiro, 2013.

This study aims to develop a view of storage networking by describing the main characteristics of architectures Network Attached Storage (NAS), Storage Area Network (SAN) and Storage Area Network Internet Protocol (SAN IP), doing a brief comparison between them to assist in architecture definition of storage networking that will be implemented to design virtual servers with high availability..

LISTA DE FIGURAS

Figura 1 – Camada de Virtualização [VMWARE, 2007] 14
Figura 2 – Duas Formas de Implementar a Camada de Virtualização 19
Figura 3 – Arquitetura típica de um processador X86/x64 [VMWARE1, 2007] . 21
Figura 4 – Arquitetura do Processador na Virtualização Total [VMWARE1, 2007] 22
Figura 5 – Arquitetura do Processador na Paravirtualização [VMWARE1, 2007] 24
Figura 6 – Arquitetura do Processador na Virtualização Assistida por Hardware [VMWARE1, 2007] 25
Figura 7 – Virtualização de Memória [VMWARE, 2007] 26
Figura 8 – NIC Teaming na Camada de Virtualização [VMWARE, 2003] 27
Figura 9 – Ponto único de falha 29
Figura 10 – Implementação de tolerância à falha 29
Figura 11 – Recurso de Failover do Cluster de Virtualização [VMWARE3, 2007] 32
Figura 12 – Recurso de Balancemanto de Carga do Cluster de Virtualização [VMWARE2, 2007] 32
Figura 13 – Uso da rede de armazenamento para acesso aos dados pelos hosts [Info-Tech, 2008] 33
Figura 14 – Componentes do sistema de armazenamento [NETO, ANTÔNIO, 2004] 35
Figura 15 – Configurações dos RAID 0 e RAID 1 [ARMBRUST, VICTOR, 2012]. 37
Figura 16 – Diferenças de configuração entre RAID 0+1 e RAID 1+0 [ARMBRUST, VICTOR, 2012] 38
Figura 17 – Diferenças de configuração entre RAID 3 e RAID 5 [ARMBRUST, VICTOR, 2012] 39
Figura 18 – Modelo de comunicação SCSI [NETO, ANTÔNIO, 2004] 42
Figura 19 – Arquitetura de DAS externo 45
Figura 20 – vSphere Storage Applicance [MANFRIN, ALEXANDER, 2011] 47
Figura 21 - Componentes do NAS [SOMASUNDARAM, G., 2011] 49
Figura 22 – Fibras monomodo e multimodo 57
Figura 23 – Topologia Fibre Channel Arbitrated Loop [DUARTE, OTTO, 2008] 58
Figura 24 – Topologia Fibre channel switched fabric (FC-SW) [DUARTE, OTTO, 2008] 59
Figura 25 – Topologia Core-edge 59
Figura 26 – Topologia mesk parcial e mesk integral [SOMASUNDARAM, G., 2011] 60
Figura 27 – Portais Fibre Channel [NETO, ANTÔNIO, 2004] 62
Figura 28 – Pilha de protocolos Fibre channel [JAVVIN, 2012] 63
Figura 29 - Exemplo de zoneamento [DUARTE, OTTO, 2008] 64
Figura 30 – Tipos de zoneamento [SOMASUNDARAM, G., 2011] 65
Figura 31 – Implementações de comunicação iSCSI 67
Figura 32 – Topologias de conectividade iSCSI Nativa e com Ponte 69
Figura 33 – Pilha de protocolos da arquitetura iSCSI [DUARTE, OTTO, 2008] . 70
Figura 34 – Comparação das topologias de Redes de Armazenamento 74

LISTA DE ABREVIATURAS E SIGLAS

ANSI	*Asynchronous Transfer Mode*
ATA	*Advanced Techonology Attachment*
ATM	*Asynchronous Transfer Mode*
CIFS	*Common Internet File System*
CPL	*Current Privilege Level*
CPU	*Central Processing Unit*
CRC	*Cyclic Redundancy Check*
DAS	*Direct Atached Storage*
DRS	*Distributed Resource Scheduler*
E/S	Entrada e Saída
EUI	*Extended Unique Identifier*
FC	*Fibre Channel*
FC-AL	*Fibre Channel Arbitrated Loop*
FCP	*Fibre Channel Protocol*
FC-SW	*Fibre Channel Switched Fabric*
Gbps	Giga Bits por Segundo
HA	*High Availability*
HBA	*Host Bus Adapter*
HIPPI	*Hlgh Performance Parallel Interface*
IDC	*Internacional Data Corporation*
IDE	*Intergrated Device Electronics*
IEEE	*Institute of Electrical and Electronics Engineers*
IP	*Internet Protocol*
IQN	*iSCSI Qualifield Name*
ISA	*Instruction Set Architecture*
iSCSI	*Internet Small Computer System Interface*
ISID	*Initiator Session Identification*
ISL	*Inter-Switch Link*
iSNS	*Internet Storage Name Service*
LB	*Load Balance*
LDAP	*Lightweight Directory Acess Protocol*
LUN	*Logical Unit Number*
MAC	*Media Acess Control*
MMF	*Milti-Mode Fiber*
MMU	*Memory Management Unit*
MMV	*Monitor Máquina Virtual*
MTU	*Maximum Transmission Unit*
NAS	*Network Attached Storage*
NFS	*Network File System*
NIC	*Network Interface Card*
NIS	*Network Information Service*
OSI	*International Organization for Standardization*
RAID	*Redundant Array of Independent*
RDM	*Raw Device Mapping*
DRS	*Distributed Resource Scheduler*
RPC	*Remote Procedure Call*
SAN	*Storage Area Network*

SAS	*Serial Attached SCSI*
SATA	*Serial Advanced Techonology Attachment*
SCSI	*Small Computer System Interface*
SMB	*Server Massage Block*
SMF	*Single-Mode Fiber*
SO	Sistema Operacional
TCP	*Transmission Control Protocol*
TI	Tecnologia da informação
TLB	*Translation Lookaside Buffer*
TOE	*TCP Offload Engine*
TSID	*Target Session Identification*
UDP	*User datagram Protocol*
VLAN	*Virtual Local Area Network*
VM	*Virtual Machine*
WWN	*World Wide Names*

SUMÁRIO

1 **INTRODUÇÃO**.....12
2 **VIRTUALIZAÇÃO DE SERVIDORES**.....16
2.1 CONCEITOS BÁSICOS DE VIRTUALIZAÇÃO.....16
2.1.1 **Máquina Virtual** 16
2.1.2 **Camada de Virtualização**.....18
2.1.3 **Virtualiza ção de CPU** 20
2.1.3.1 Virtualização Total Usando Tradução Binária 21
2.1.3.2 Paravirtualização.....23
2.1.3.3 Virtualização Assistida por Hardware.....24
2.1.4 **Virtualização de Memória** 25
2.1.5 **Virtualização dos Dispositivos de Entrada e Saída**.....26
3 **ALTA DISPONIBILIDADE**.....28
3.1 TOLERÂNCIA À FALHA 28
3.2 ALTA DISPONIBILIDADE PARA SERVIDORES VIRTUAIS.....30
3.2.1 **Cluster na Camada de Virtualização**.....31
3.2.2 **Importância da Rede de Armazenamento**.....32
4 **SISTEMA DE ARMAZENAMENTO**.....34
4.1 COMPONENTES DO SISTEMA DE ARMAZENAMENTO 34
4.2 MATRIZ DE DISCOS E PROTEÇÃO DE DADOS: RAID.....35
4.2.1 **RAID 0** 36
4.2.2 **RAID 1** 36
4.2.3 **RAID 0+1 e RAID 1+0**.....37
4.2.4 **RAID 3** 38
4.2.5 **RAID 5** 39
4.3 PROTOCOLOS DE COMUNICAÇÃO.....40
4.3.1 ***Integrated Device Electronics* / *Advanced Technology Attachment* (IDE/ATA) e Serial ATA (SATA)** 40
4.3.2***Small Computer System Interface* (SCSI) e *Serial Attached* SCSI (SAS)** 40
4.3.3 ***Fibre Channel Protocol* – FCP**.....42
5 **ARMAZENAMENTO DIRETAMENTE CONECTADO e ARMAZENAMENTO CONECTADO À REDE**.....44
5.1 ARMAZENAMENTO CONECTADO DIRETO – *DIRECT-ATTACHED STORAGE* (DAS).....44
5.1.1 **Vantagens e Limitações** 45
5.1.2 **Alternativas do DAS para Virtualização** 46
5.1.2.1 vSphere Storage Applicance.....46
5.1.2.2 VM6 Manage Cloud Platform 47
5.2 ARMAZENAMENTO CONECTADO À REDE – *NETWORK-ATTACHED STORAGE* (NAS).....48
5.2.1 **Componentes do NAS** 48
5.2.2 **Implantação do NAS** 49
5.2.3 **Protocolos de Compartilhamento**.....50
5.2.3.1 *Network File System* (NFS).....50
5.2.3.2 *Common Internet File System* (CIFS) 51
5.2.4 **Vantagens e Limitações do NAS**.....51
5.2.5 **Utilizando NAS para Virtualização de Servidores** 52

6 **REDES DE ÁREA DE ARMAZENAMENTO – *STORAGE AREA NETWORK* (SAN)** 55
6.1 REDES DE ÁREA DE ARMAZENAMENTO SOBRE FC – *STORAGE AREA NETWORK* (SAN FC) 55
6.1.1 **Componentes da SAN** 56
6.1.2 **Cabeamento *Fibre Channel*** 56
6.1.3 **Topologia *Fibre Channel*** 57
6.1.3.1 Ponto-a-ponto 57
6.1.3.2 Laço Arbitrado – Fibre Channel Arbitrated Loop (FC-AL) 57
6.1.3.3 Fibre Channel Switched Fabric (FC-SW) 58
6.1.3.4 *Fabric Core-Edge* 59
6.1.3.5 *Mesk* 60
6.1.4 **Portas *Fibre Channel*** 61
6.1.5 ***Fibre Channel Protocol* (FCP)** 62
6.1.5.1 Pilhas de Protocolos Fibre Channel 63
6.1.6 **Zoneamento** 64
6.2 REDES DE ÁREA DE ARMAZENAMENTO SOBRE IP – *STORAGE AREA NETWORK* IP (SAN IP) 65
6.2.1 **Protocolo iSCSI** 66
6.2.1.1 Conectividade iSCSI 66
6.2.1.2 Topologias para Conectividade iSCSI 68
6.2.1.3 Pilha de Protocolos da Arquitetura iSCSI 69
6.2.1.4 Sessão iSCSI 70
6.2.1.5 Manipulação de Erros e Segurança iSCSI 72
7 **ANÁLISE COMPARATIVA DAS REDES DE ARMAZENAMENTO** 74
7.1 COMPARAÇÃO DOS PROTOCOLOS iSCSI, NFS e FC PARA VIRTUALIZAÇÃO 75
8 **CONCLUSÃO** 84

1 INTRODUÇÃO

Atualmente os sistemas computacionais ganharam grande importância para as organizações, tornando cada vez mais rápido e eficiente o atendimento de seus serviços. Para prover toda a demanda dos negócios são utilizados sistemas como correio eletrônico, aplicações web, sistemas de banco de dados, que por sua vez geram aumento na quantidade de equipamentos e nos custos em relação à administração e suporte destes equipamentos e sistemas.

Outro ponto relevante para as organizações é a alta disponibilidade dos sistemas e serviços, que está diretamente relacionada à continuidade do negócio da empresa e pode ser o diferencial na escolha do cliente. Imagine uma empresa bancária ou um site de conteúdo web ter seus serviços interrompidos, isso prejudicaria a imagem da empresa e acarretaria prejuízos financeiros.

Os setores de tecnologia da informação (TI) têm buscado utilizar a virtualização como solução para a redução de custos com a consolidação de servidores e se beneficiando de recursos como a alta disponibilidade proporcionada por esta tecnologia. Recentemente a empresa de pesquisa *Internacional Data Corporation* (IDC) divulgou que quase 40% dos gerentes de TI escolheram a virtualização e consolidação de servidores como principal prioridade de TI para 2012 [MULLINS, ROBERT, 2012].

A virtualização consiste basicamente em criar uma versão virtual de algo real, como um sistema operacional, *system operation* (SO), servidor, dispositivo de armazenamento e recursos de rede. A virtualização de servidores permite que um único servidor físico ou host possa executar vários servidores virtuais ou máquinas virtuais, *virtual machine* (VM).

Cada máquina virtual pode executar diferentes sistemas operacionais e aplicações isoladamente. A máquina virtual simula um computador físico, emitindo requisições aos dispositivos de hardware como processadores, memória, disco rígido, interface de rede e outros dispositivos que são geridos por uma camada de virtualização que traduz essas solicitações para o hardware [DESAI, ANIL, 2011].

A camada de virtualização é uma camada de software que pode ser instalada e executada como um aplicativo sobre o sistema operacional ou diretamente sobre o hardware do servidor. Estes conceitos serão melhor entendidos no próximo capítulo. Essa camada realiza o particionamento e partilha dinâmica dos recursos de hardware disponíveis, abstraindo o sistema operacional e aplicações do hardware e encapsulando em máquinas virtuais portáteis, permitindo flexibilidade e independência de hardware. A figura 1 ilustra este conceito. A máquina virtual pode ser migrada do servidor físico sem a necessidade de alterações de drives, sistema operacional ou aplicativos. Isso possibilita que máquinas virtuais funcionem com tolerância a falha, sem tempo de indisponibilidade quanto à necessidade de mudanças de configuração de infraestrutura e manutenção de hardware.

Para realizar a migração da máquina virtual entre servidores físicos, estes servidores precisam compartilhar uma área de dados comum, isso só será possível se for utilizada uma área de armazenamento remoto em um dispositivo de armazenamento.

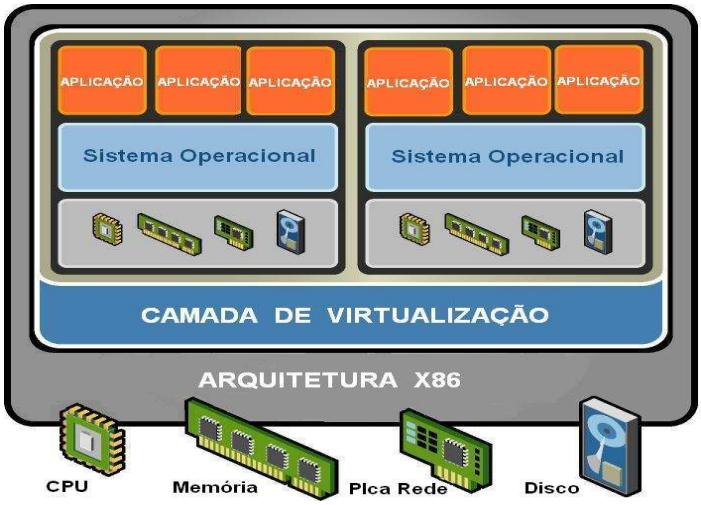

Figura 1 – Camada de Virtualização [VMWARE, 2007]

O dispositivo de armazenamento ou s*torage*, termo mais utilizado por profissionais de TI, é o dispositivo que armazena os dados de forma magnética ou óptica para acesso de servidores, que pode ser diretamente conectado ao servidor ou através da rede [SOMASUNDARAM, G., 2011].

O acesso à área de armazenamento remoto pelos servidores ocorre através de uma rede de armazenamento. Existem vários tipos de rede de armazenamento suportados pelos softwares de virtualização, cada uma delas pode oferecer vantagens e desvantagens dependendo do cenário a ser analisado. As principais são: Armazenamento Conectado à Rede, *Network Attached Storage* (NAS); Redes de Área de Armazenamento, *Storage Area Network* (SAN) e Redes de Área de Armazenamento sobre Protocolo Internet, *Internet Protocol* SAN (SAN IP).

As empresas se mostram entusiasmadas com a implementação da virtualização dos servidores por causa dos benefícios que proporciona, e a rede de armazenamento desempenha um papel importante nesta implementação. Um

planejamento adequado desta implementação é fundamental para garantir alta disponibilidade e tolerância à falha.

Sobre este aspecto um entendimento claro das tecnologias disponíveis para rede de armazenamento se faz necessário. Este trabalho tem como objetivo discutir os principais tipos de redes de armazenamento utilizadas pela camada de virtuaização, de forma a descrever as vantagens e desvantagens de cada uma delas e sua melhor aplicação para determinadas situações, contribuindo no planejamento e na implantação de servidores virtuais, visando garantir alta disponibilidade, levando-se em consideração aspectos como desempenho, capacidade, segurança e custo.

O trabalho a ser realizado será teórico, baseado na pesquisa literária, onde há uma introdução dos conceitos básicos de virtualização e alta disponibilidade, uma discussão de forma mais ampla, das principais tecnologias de redes de armazenamento para garantir alta disponibilidade de servidores virtuais e uma comparação entre as tecnologias apresentadas. Esta comparação será realizada com base no software de virtualização vSphere da VMware, que conforme pesquisa recente, realizada pela Gartner, é o software de virtualização mais utilizado no mercado [GARTNET, INC, 2011].

Outro aspecto importante para garantir alta disponibilidade dos servidores virtuais é a replicação das áreas de armazenamento, que não serão abordadas neste trabalho, ficando em aberto para trabalhos futuros.

2 VIRTUALIZAÇÃO DE SERVIDORES

A virtualização de servidores está se tornando rapidamente parte fundamental dos centros de processamento de dados. São muitas as motivações para implantar um projeto de virtualização. A redução dos custos é o grande fator motivacional. Na maioria dos servidores físicos o nível de utilização dos recursos é baixo durante a maior parte do tempo, causando desperdício. A ideia básica de executar vários servidores virtuais em um único servidor físico possibilita uma melhor utilização dos recursos computacionais.

A redução de servidores físicos é eminente, o que possibilita uma redução nos custos operacionais como manutenção de hardware, consumo de energia, refrigeração e alocação de espaço no centro de processamento de dados. Além disso, a virtualização de servidores traz outros benefícios como: gerenciamento centralizado, flexibilidade, isolamento de ambiente e alta disponibilidade.

2.1 CONCEITOS BÁSICOS DE VIRTUALIZAÇÃO

Entende-se melhor como é possível o processo de virtualização, através do compartilhamento dos recursos de hardware. Mas antes precisa-se conhecer os dois principais componentes da virtualização: a máquina virtual e a camada de virtualização.

2.1.1 **Máquina Virtual**

A máquina virtual é uma implementação de software totalmente isolada onde um sistema operacional e aplicativos podem ser instalados e executados se comportando exatamente como um computador físico. Os sistemas operacionais e aplicativos não são capazes de diferenciar uma máquina virtual de um

computador real. A máquina virtual contém todos os dispositivos de hardware virtuais: processador, memória, disco rígido, interface de rede e outros dispositivos, que são geridos por uma camada de virtualização. Esta camada traduz as solicitações para o hardware, trazendo muitas vantagens em relação ao computador real [VMWARE, 2012]:

- Compatibilidade: As máquinas virtuais são compatíveis com os sistemas operacionais, aplicativos e drivers disponíveis para padrão x86/x64;
- Isolamento: Apesar de compartilhar recursos de hardware com outras máquinas virtuais, elas são isoladas uma das outras. Se ocorrer um problema com uma das máquinas virtuais de um mesmo host, as demais não são afetadas;
- Encapsulamento: A máquina virtual é como um contêiner de software que contém um conjunto de recursos de hardware virtuais, sistema operacional e aplicativos, tornando as máquinas virtuais portáteis e de fácil gerenciamento. Basicamente é um conjunto de arquivos que pode-se mover ou copiar para qualquer dispositivo de armazenamento de dados onde o software de virtualização tenha acesso para sua execução;
- Independência de Hardware: As máquinas virtuais são independentes do hardware e têm seus componentes de hardware virtuais que podem ser diferentes dos componentes do hardware real e podem executar sistemas operacionais diferentes.

2.1.2 **Camada de Virtualização**

Através da camada de virtualização é possível gerenciar e implementar as máquinas virtuais, que podem ser implementada de duas formas: através das máquinas virtuais de processo ou do sistema conhecido como monitor de maquina virtual (MMV) ou hipervisor.

A máquina virtual de processo é um programa de aplicação, instalado e executado sobre o sistema operacional que oferece suporte para uma variedade de hardware x86/x64, podendo executar com outros sistemas operacionais. Um exemplo é a execução do VirtualBox[1] executando sobre um computador Windows e carregando uma maquina virtual Linux.

O monitor de máquina virtual, ou hipervisor, é o software da camada de virtualização que é instalado diretamente sobre o computador, tendo acesso direto aos recursos de hardware, não passando por um sistema operacional, o que o torna mais eficiente do que as máquinas virtuais de processo, tendo maior desempenho, escalabilidade e robustez. Através do hipervisor é possível executar vários sistemas operacionais diferentes, simultaneamente no mesmo hardware. A figura 2 ilustra as duas formas de implementar a camada de virtualização.

Neste trabalho será usado como referência da camada de virtualização o hipervisor por se tratar de um ambiente mais robusto para a virtualização de servidores e a implementação de alta disponibilidade.

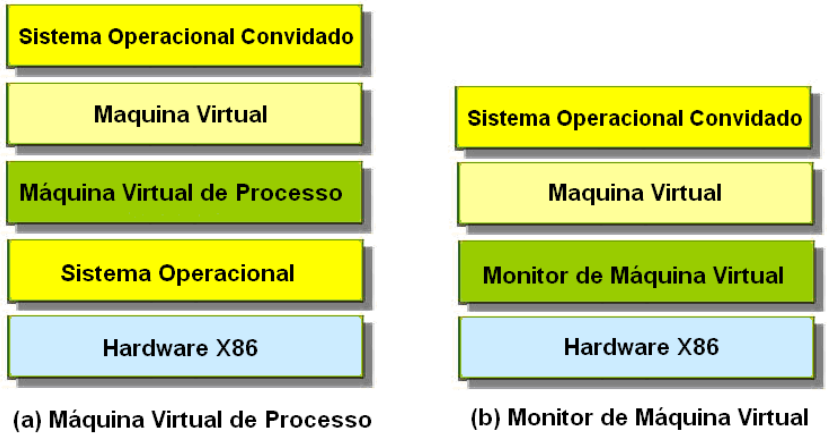

Figura 2 – Duas Formas de Implementar a Camada de Virtualização

Vantagens do uso do hipervisor:

- Gerenciamento centralizado: o Hipervisor passa a ser o elemento central na administração dos servidores virtuais;
- Flexibilidade: a criação de novas máquinas virtuais e a inclusão de novos recursos para as máquinas virtuais já existentes levam poucos minutos;
- Isolamento de ambiente: mesmo compartilhando recursos cada máquina virtual é executada isoladamente, de forma que uma máquina virtual não interfere em outra. Um exemplo é que se uma máquina virtual apresentar problema as demais continuam sua execução normalmente;
- Alta disponibilidade: provê disponibilidade do serviço sem interrupção, através de redundância dos hosts virtuais. O hipervisor permite a criação

[1] VirtualBox é um software de virtualização que visa criar máquinas virtuais para instalação de sistemas operacionais distintos, executado sobre outro sistema operacional como um aplicativo.

de *clusters*[2] e a integração com o S*torage* que possibilita a migração de servidores virtuais de *hosts* que fazem parte do *cluster*, caso um destes *hosts* falhe.

2.1.3 **Virtualização de CPU**

A virtualização não é uma tarefa simples. Além da preocupação com desempenho existe a questão de como compartilhar os recursos de hardware entre o sistema operacional nativo e convidado, sem que ocorram conflitos, começando pelo processador que é o principal componente do computador. Para entender precisa-se conhecer alguns conceitos da arquitetura x86/x64.

O processador contém um conjunto de instruções de máquina, *Instruction Set Architecture* (ISA), que é a interface entre o hardware e o software. Estas instruções são divididas em quatro modos de operação denominados de anéis de proteção, *Rings* ou nível de privilégio corrente, *Current Privilege Level* (CPL), identificados de 0 a 3. Os sistemas operacionais convencionais, como Windows e Unix, utilizam apenas dois níveis de privilégios, o *Ring* 0 que é usado pelo sistema operacional e permite a execução de instruções de maior privilégio, como manipulação de recursos de hardware, e o *Ring* 3 instruções de menor privilégio que são executadas por aplicações de usuário. Uma alternativa para aplicações de usuários são as chamadas de sistemas que permitem que programas de usuários, após autenticação e validação da operação, acessem de forma indireta e controlada recursos de hardware. Isso era a grande dificuldade

[2] *Cluster* é formado por um conjunto de computadores, que utiliza um tipo especial de sistema operacional classificado como sistema distribuído. Ligados em rede e comunicam-se através do sistema, trabalhando como se fossem uma única máquina de grande porte.

para a virtualização na arquitetura x86/x64. A figura 3 demonstra a arquitetura típica de um processador x86/x64.

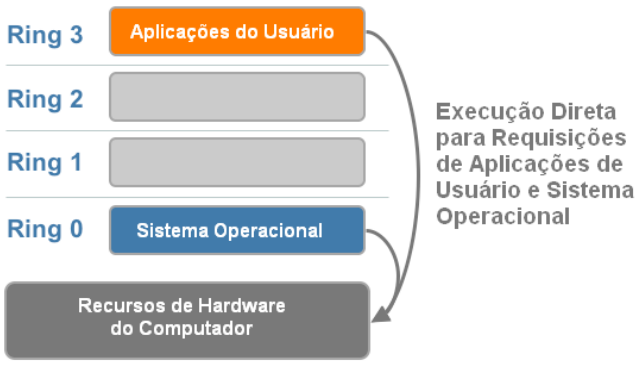

Figura 3 – Arquitetura típica de um processador x86/x64 [VMWARE1, 2007]

Este desafio foi vencido pela VMware em 1998, através da técnica de translação binária, que possibilita ao hipervisor executar com privilégio *Ring* 0 isolado, movendo o sistema operacional para um nível de privilégio maior que as aplicações de usuário, *Ring* 3, e menor que o *Ring* 0. Na prática a virtualização de CPU na arquitetura x86/x64 funciona através da alteração de privilégios de execução das instruções de máquina dos processadores. Existem três técnicas alternativas para virtualizar CPU: virtualização total usando tradução binária, paravirtualização e virtualização assistida por hardware.

2.1.3.1 *Virtualização Total Usando Tradução Binária*

Virtualização Total usando tradução Binária consiste em criar uma réplica dos componentes de hardware para a máquina virtual de forma que o sistema operacional e aplicações possam ser executados como se estivessem sobre o hardware original. Nesta abordagem as instruções privilegiadas emitidas pelo sistema operacional convidado são traduzidas pelo hipervisor para uma nova sequência de instruções que têm o efeito pretendido no hardware. Já as

instruções de nível de usuário são executadas diretamente no processador. A figura 4 representa esta abordagem. Os produtos de virtualização da VMware e Microsoft são exemplos de virtualização total.

Figura 4 – Arquitetura do Processador na Virtualização Total [VMWARE1, 2007]

Esta combinação de execução direta e tradução binária fornece a virtualização total para o sistema operacional convidado de forma a abstrair totalmente o hardware pela camada de virtualização. A grande vantagem é que não há necessidade de modificações no sistema operacional convidado para executar sobre o hipervisor, o que facilita a migração de máquinas virtuais entre servidores físicos, pois existe total independência dos recursos de hardware. Porém esta abordagem tem algumas desvantagens.

O hipervisor controla todos os processos e todas as chamadas ao hardware, realizando a tradução binária para as instruções privilegiadas, podendo levar a uma perda de desempenho. Outra desvantagem é em relação à diversidade de componentes de hardware que compõem um computador: imitar o comportamento exato destes componentes é uma tarefa difícil e a solução adotada é a utilização de um conjunto genérico de dispositivos. Com isso pode ocorrer perda de desempenho, podendo haver substituição de recursos quando

comparado ao hardware real. Existe outro problema técnico em relação à implementação de memória virtual através de paginação pelos sistemas operacionais convidados, que é converter o espaço de endereçamento do sistema operacional convidado para um endereçamento real, disputando recursos com outros sistemas convidados. Esse tratamento também representa uma perda de desempenho.

2.1.3.2 *Paravirtualização*

A Paravirtualização surgiu como uma abordagem alternativa para contornar as desvantagens da virtualização total referente a processamento. Esta abordagem implica em alterar o *Kernel*[3] do sistema operacional convidado para substituir instruções privilegiadas que se comunicam diretamente com o hipervisor, melhorando seu desempenho. As instruções de usuário são executadas diretamente no processador como na virtualização total. A figura 5 ilustra esta abordagem. O projeto Xen de código aberto, *Xen Open Source*, é um exemplo de Paravirtualização.

Os dispositivos da máquina virtual são uma abstração do hardware que não são idênticos ao hardware físico. Estes dispositivos são acessados através de drivers do próprio hipervisor, o que otimiza o desempenho. Sua desvantagem é justamente em relação à modificação necessária no sistema operacional convidado, que dificulta sua adoção [VERAS, MANUEL, 2011].

---

[3] *Kernel* é o núcleo do sistema operacional, responsável pela comunicação entre as aplicações e os componentes de hardware, realizando o gerenciamento dos componentes de hardware como: processador, memória e dispositivos de entrada e saída.

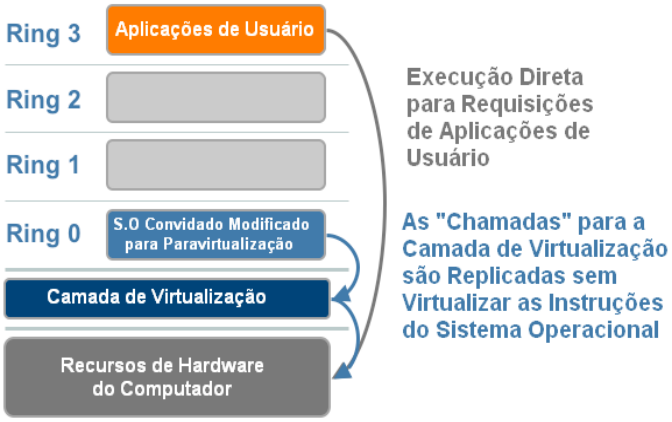

Figura 5 – Arquitetura do Processador na Paravirtualização [VMWARE1, 2007]

2.1.3.3 *Virtualização Assistida por Hardware*

Virtualização assistida por hardware é na verdade o desenvolvimento de novos recursos nos processadores, pelos fornecedores, para simplificar as técnicas de virtualização. A Intel e AMD desenvolveram respectivamente as tecnologias de virtualização VT-x e AMD-V, alterando o modo de operação dos processadores, de forma que o hipervisor passou a executar em um *Ring* abaixo do *Ring* 0, ou seja, com maior prioridade que o sistema operacional, como demonstra a figura 6. As instruções privilegiadas são automaticamente interceptadas pelo hipervisor de forma a eliminar a necessidade da tradução binária ou Paravirtualização.

O problema é que estes novos processadores foram lançados em 2006 e apenas novos sistemas ganham vantagens dos recursos da virtualização assistida por hardware. A primeira geração destes processadores perdia em desempenho para a virtualização total, em função da rigidez da programação. Apenas os processadores x64 permitiam que sistemas desenvolvidos para virtualização total fizessem também a virtualização assistida por hardware, ganhando desempenho [VMWARE1, 2007].

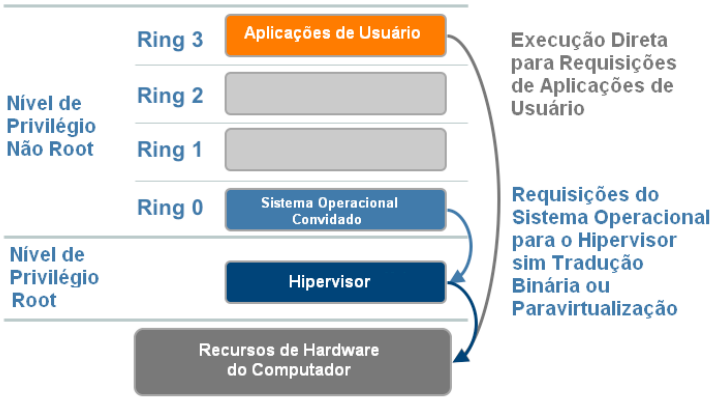

Figura 6 – Arquitetura do Processador na Virtualização Assistida por Hardware [VMWARE1, 2007]

2.1.4 **Virtualização de Memória**

A memória é o próximo componente crítico no processo de virtualização que nada mais é do que o compartilhamento da memória física e atribuição dinâmica à máquina virtual. Funciona parecido com o processo de memória virtual dos atuais sistemas operacionais. Os sistemas operacionais realizam a tradução do número da página virtual para o número da página física armazenando em uma tabela de paginação. Os atuais processadores contêm uma unidade de gerenciamento de memória, *Memory Management Unit* (MMU), que realiza esta tradução e utiliza um cache associado chamado *Translation Lookaside Buffer* (TLB) para otimizar o desempenho da memória virtual.

Basicamente é preciso virtualizar a MMU. O sistema operacional convidado continua realizando o mapeamento entre o endereço virtual e o físico, mas sem acesso direto à memória real. O hipervisor realiza o mapeamento da memória física da máquina virtual para a memória real através do TLB da máquina virtual, para evitar duas camadas de tradução, conforme demonstrado na figura 7. A virtualização da MMU gera uma sobrecarga na virtualização, mas a segunda

geração da virtualização assistida por hardware oferece ganhos de desempenho [VMWARE1, 2007].

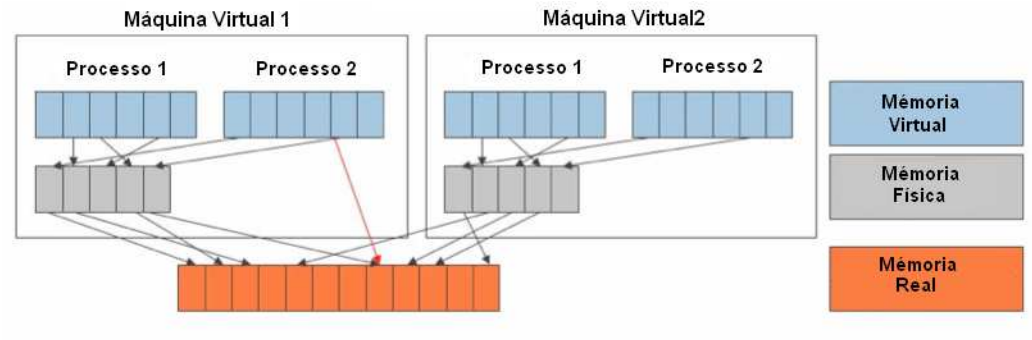

Figura 7 – Virtualização de Memória [VMWARE, 2007]

2.1.5 **Virtualização dos Dispositivos de Entrada e Saída**

É o compartilhamento dos componentes físicos de entrada e saída para a máquina virtual. O hipervisor virtualiza os componentes de hardware para as máquinas virtuais como um conjunto padronizado de dispositivos virtuais, utilizando drivers de dispositivos padronizados, traduzindo e gerenciando as solicitações dos dispositivos virtuais para os componentes de hardware. Isso possibilita uma portabilidade com todas as máquinas virtuais configuradas para execução dos mesmos dispositivos virtuais, independente dos componentes de hardware.

O software de gerenciamento e virtualização de entrada e saída permite um rico conjunto de recursos, como a criação de switches virtuais, placas de redes virtuais e redes virtuais entre máquinas virtuais sem tráfego na rede física. Também permite o agrupamento de múltiplas placas de rede físicas, *Network Interface Card* (*NIC) Teaming*, de forma transparente para a máquina virtual. Caso ocorra uma falha em uma das placas de rede as demais mantêm a conexão física sem a percepção de falha para a máquina virtual, provento

tolerância à falha. Este conceito será melhor entendido com o próximo capítulo. A figura 8 ilustra a configuração da *NIC Teaming*.

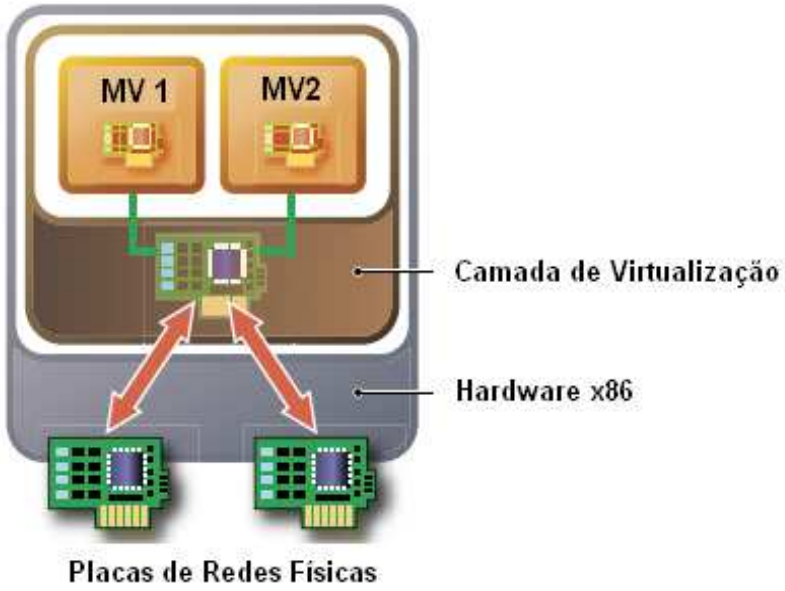

Figura 8 – NIC Teaming na Camada de Virtualização [VMWARE, 2003]

A criação de NIC *Teaming* e a padronização dos dispositivos de entrada e saída virtuais, em conjuntos com outros recursos da virtualização, como a criação de cluster de servidores, permite a alta disponibilidade para as maquinas virtuais.

3 ALTA DISPONIBILIDADE

Alta disponibilidade, *High Availability* (HA), é a capacidade de prover disponibilidade de recursos e serviços com o mínimo de interrupções. É uma combinação de soluções de redundância de hardware, software e um conjunto de bons processos [MICROSOFT, 2008].

O acesso contínuo às informações é essencial para a continuidade dos negócios das empresas. A interrupção do acesso pode acarretar em grandes prejuízos financeiros, por este motivo as empresas buscam cada vez mais soluções de alta disponibilidade.

Pela definição percebemos que a alta disponibilidade tem por objetivo a busca por uma forma de manter os serviços disponíveis o máximo possível, mesmo que ocorra alguma falha, aí está implícito o conceito de tolerância à falha.

3.1 TOLERÂNCIA À FALHA

Primeiro vamos entender o que é um ponto único de falha. O ponto único de falha se refere à falha de um único componente que causará a indisponibilidade do sistema ou serviço. A figura 9 ilustra possíveis pontos únicos de falha em um sistema complexo que contém vários componentes: servidor, rede, switch e storage. Neste cenário o serviço executado no servidor está disponível ao cliente através de uma rede IP e grava e lê dados no *storage*, através da conexão de fibra óptica, *fibrel channel* (FC), entre o servidor e o *storage*. Esta conexão é possível através de uma placa de fibra óptica, conhecida como *host bus adapter* (HBA), instalada no servidor e conectada ao switch FC, que por sua vez é conectado à placa HBA do *storage*.

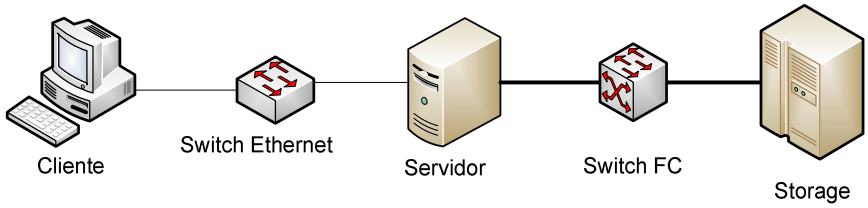

Figura 9 – Ponto único de falha

Com esta configuração uma simples falha em um dos componentes envolvidos causará a indisponibilidade do serviço. Neste exemplo podemos indicar diversos pontos únicos de falha: servidor, única placa HBA no servidor, rede IP, switch de fibra óptica, porta da placa HBA no *storage* e o próprio *storage*. Para minimizar as causas do ponto único de falha é essencial à implementação dos mecanismos de tolerância à falha.

A tolerância à falha é o conceito que, através dos mecanismos de redundância e replicação, procura atenuar um ponto único de falha, de forma que o serviço só ficará indisponível caso todos os componentes do grupo de redundância falhem, e não apenas um [SOMASUNDARAM, G., 2011]. A figura 10 ilustra a implementação de tolerância à falha.

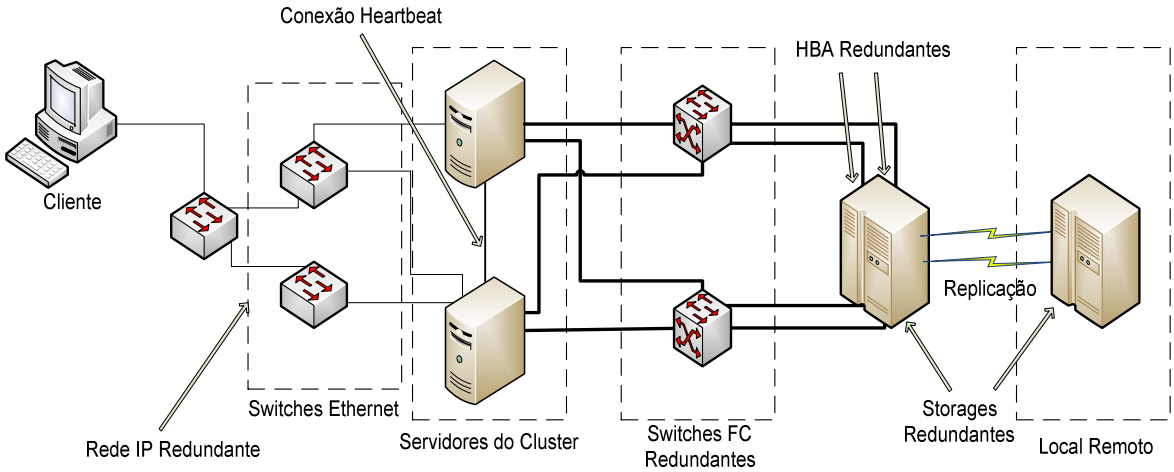

Figura 10 – Implementação de tolerância à falha

Neste cenário os pontos únicos de falha são eliminados através de melhorias na infraestrutura:

- Configuração de várias placas HBA’s no servidor para atenuar uma falha em uma única placa HBA;
- Configuração de vários switches de FC para minimizar uma falha em um dos switches;
- Configuração de várias portas do *storage* para aumentar a disponibilidade de acesso aos dados no *storage*;
- Implementação de replicação de dados entre *storage* local e o *storage* em localidade remota para atenuar falha no *storage*;
- Implementação de *cluster* de servidores que garante a redistribuição de carga de processamento entre os demais servidores do *cluster* caso ocorra uma falha em algum deles.

3.2 ALTA DISPONIBILIDADE PARA SERVIDORES VIRTUAIS

Um dos principais recursos da alta disponibilidade é o *cluster* de servidores. O *cluster* de servidores é um mecanismo de redundância que permite a criação de um *pool* de recursos entre dois ou mais servidores, cujo objetivo é distribuir carga de processamento entre eles. Com isso é possível realizar processamentos que só seria possível em um servidor de alto desempenho.

Os servidores que compõem o *cluster* são denominados *host* ou nó. Todos os nós devem ser interconectados através de uma rede que permita o acréscimo ou a retirada de um nó sem interromper a execução do *cluster*.

Através da conexão *heartbeat*[4] os servidores do *cluster* trocam informações para identificar seus estados. Caso um dos servidores falhe a carga de processamento é distribuída entre os demais.

Por uma questão conceitual passarão a ser chamadas máquinas virtuais de servidores virtuais e servidores físicos de *hosts*.

3.2.1 **Cluster na Camada de Virtualização**

O *cluster* é configurado no hipervisor, a versão do hipervisor deve ser a mesma para todos os nós do *cluster*. Existem duas opções de configurações disponíveis: balanceamento de carga, *load balance* (LB) e f*ailover*.

- *Failover* – É a técnica no qual um host assume os serviços de outro quando este apresenta falha. Neste caso os servidores virtuais que estavam em execução no host que falhou são reiniciados nos *hosts* sem problemas*.* Cada *host* monitora os demais através da conexão *hearbeat* para identificar a falha. O hipervisor garante que haja no *pool* recurso suficiente para que os servidores virtuais sejam redistribuídos. A figura 11 ilustra este recurso.
- *Load Balance* (LB) – O hipervisor monitora os recursos disponíveis em cada *host* e caso um *host* tenha pouco recurso disponível é possível mover uma instância em execução de um servidor virtual de *host* para outro do *pool* de recursos, de forma a redistribuir carga de

[4] *Heartbeat* é o processo responsável por monitorar o status dos servidores do *cluster* através de testes periódicos, permitindo através do status, detectar falha em um dos nós de forma que o outro nó sem problema assuma as funções do nó que falhou.

processamento entre os *hosts* sem que haja indisponibilidade do servidor virtual durante este processo. A figura 12 ilustra este recurso.

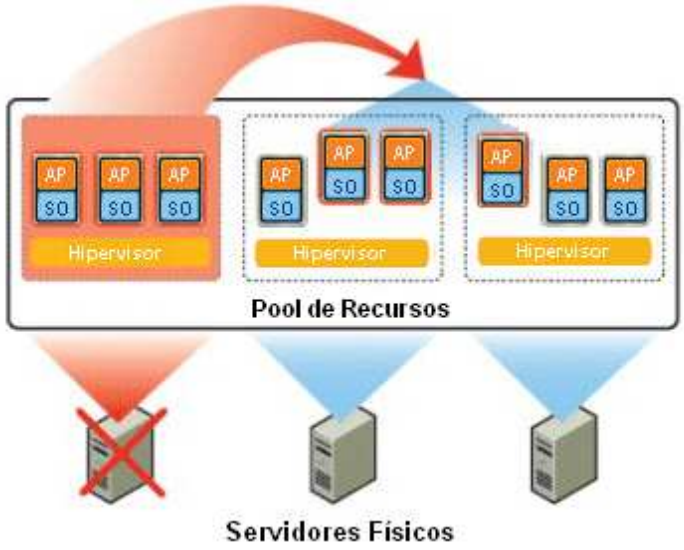

Figura 11 – Recurso de Failover do Cluster de Virtualização [VMWARE3, 2007]

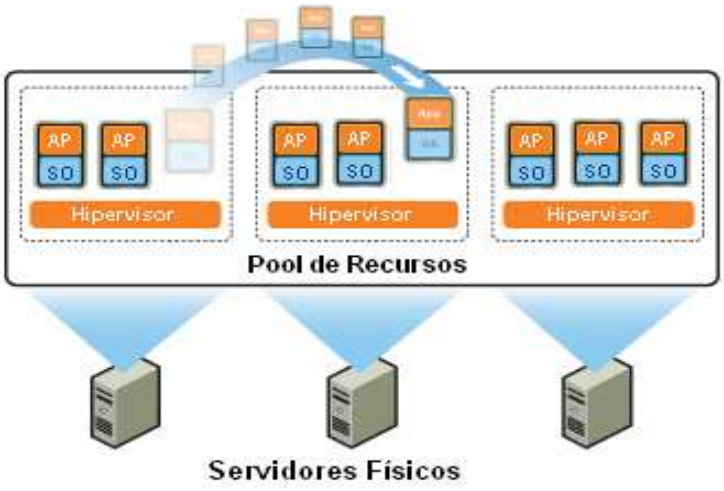

Figura 12 – Recurso de Balancemanto de Carga do Cluster de Virtualização [VMWARE2, 2007]

3.2.2 **Importância da Rede de Armazenamento**

O *cluster* na camada de virtualização é possível devido à flexibilidade proporcionada pelos servidores virtuais. O servidor virtual é basicamente um

conjunto de arquivos interpretados pelo hipervisor. Para configuração do *cluster* na camada de virtualização todos os *hosts* que farão parte do *cluster* devem executar a mesma versão do hipervisor e ter acesso de leitura e gravação ao conjunto de arquivos dos servidores virtuais.

Através da rede de armazenamento é possível que uma área de armazenamento seja compartilhada pelos *hosts* de forma que o conjunto de arquivos dos servidores virtuais seja acessado por todos os *hosts*. A figura 13 ilustra esta configuração.

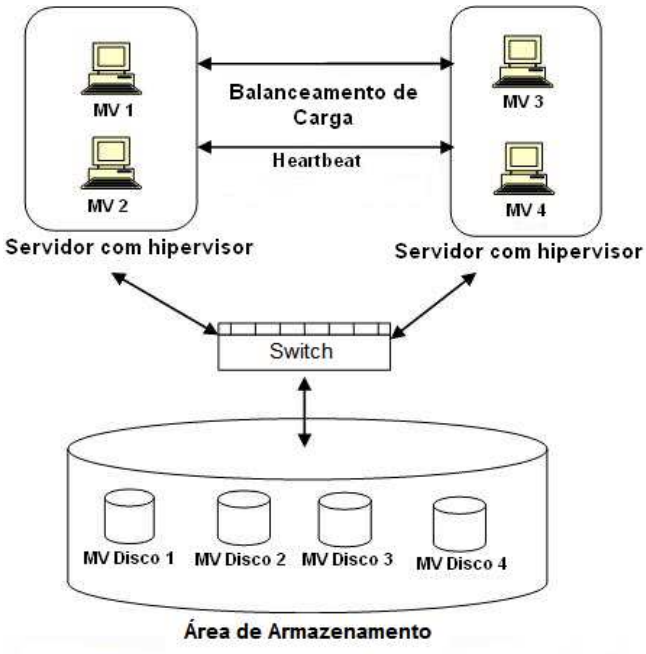

Figura 13 – Uso da rede de armazenamento para acesso aos dados pelos *hosts*

[Info-Tech, 2008]

Nos próximos capítulos serão vistos conceitos sobre as redes de armazenamento e discutida sua aplicação para a alta disponibilidade dos servidores virtuais.

4 SISTEMA DE ARMAZENAMENTO

O armazenamento de dados é um dos principais elementos para os sistemas e aplicativos. Os usuários utilizam aplicativos para armazenar e recuperar dados. Os dados tratados pelos aplicativos são gravados no meio de armazenamento passando durante este processo por vários componentes que coletivamente são chamados de sistema de armazenamento.

4.1 COMPONENTES DO SISTEMA DE ARMAZENAMENTO

Os principais componentes do sistema de armazenamento são:

- Servidor, *Host*: provê recursos de hardware e software para as aplicações. Os *hosts* podem ser desktops simples ou complexos *clusters* de servidores;
- Aplicação: fornece ambiente de entrada e saída (E/S) de dados referentes aos processos executados. O hipervisor é um exemplo de aplicação;
- Meio de Armazenamento: provê a infraestrutura necessária para o armazenamento e recuperação dos dados provenientes das aplicações. São alguns exemplos de meio de armazenamento: discos rígidos, discos ópticos e unidades de fitas. Os discos rígidos são mais utilizados nos sistemas de armazenamento;
- Dispositivos de Armazenamento, *Storage*: são equipamentos específicos de armazenamento de dados com características de alto desempenho, flexibilidade, escalabilidade e alta disponibilidade;

- Meio de Transporte: provê conectividade física entre servidores e dispositivos de armazenamento. Exemplo cabeamento;
- Protocolo de Comunicação: realiza a comunicação lógica entre servidores e dispositivos de armazenamento. Os protocolos *Small Computer System Interface* (SCSI) e *Fibre Channel* (FC) são alguns exemplos.

A figura 14 ilustra os componentes do sistema de armazenamento.

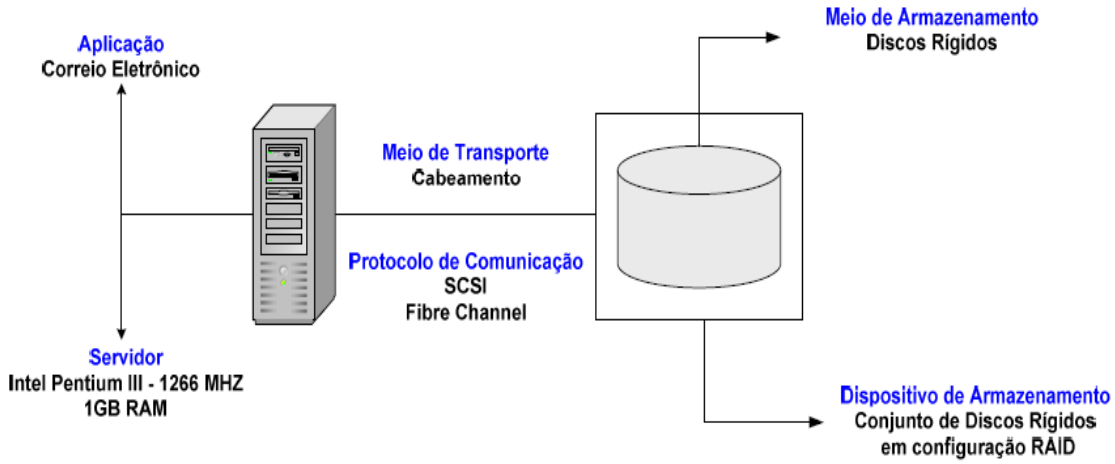

Figura 14 – Componentes do sistema de armazenamento [NETO, ANTÔNIO, 2004]

4.2 MATRIZ DE DISCOS E PROTEÇÃO DE DADOS: RAID

Quando se fala em sistema de armazenamento, um importante mecanismo deve ser levado em consideração em relação ao desempenho de E/S dos dados: proteção contra falha ou *redundant array of independent* (RAID).

RAID é a tecnologia que permite o uso de vários discos em conjunto, fornecendo proteção aos dados contra falhas nos discos e otimizando desempenho de E/S de dados. Pode ser implementado por software, em nível de sistema operacional, que afeta o desempenho devido à carga adicional de

processamento, sendo pouco usado. Pode também ser implementado por hardware, através do uso de controladores de discos que tenham esta função.

Existem vários tipos de configuração de RAID baseadas em técnicas de distribuição, espelhamento e paridade. Essas técnicas influenciam no desempenho de E/S dos dados e determinam sua disponibilidade. As diferenças entre os tipos devem ser consideradas na definição de qual usar. Este tópico detalha os tipos de RAID.

4.2.1 **RAID 0**

A configuração RAID 0 oferece unicamente distribuição de dados. Os dados são distribuídos pelos discos da matriz do RAID, permitindo que um arquivo seja fracionado por vários discos da matriz. Esta configuração aumenta o desempenho de acesso aos dados uma vez que os controladores de discos podem trabalhar paralelamente. Todavia, não fornece proteção e disponibilidade de dados em caso de falha em algum disco da matriz. A figura 15(a) ilustra esta configuração.

4.2.2 **RAID 1**

Nesta configuração é aplicada a técnica de espelhamento, onde os dados são gravados simultaneamente em dois discos, produzindo duas cópias dos dados para proteção e redundância à falha. Caso um disco falhe, os dados permanecem íntegros no outro e a controladora de discos passa a atender as solicitações a partir do disco bom. Esta é a configuração de RAID mais segura e de menor impacto na recuperação de falha, tendo em vista que todos os dados são duplicados. Porém, o desempenho de gravação é prejudicado e seu custo é

elevado devido à necessidade de armazenamento ser duas vezes maior do que a de dados. A figura 15(b) ilustra esta configuração.

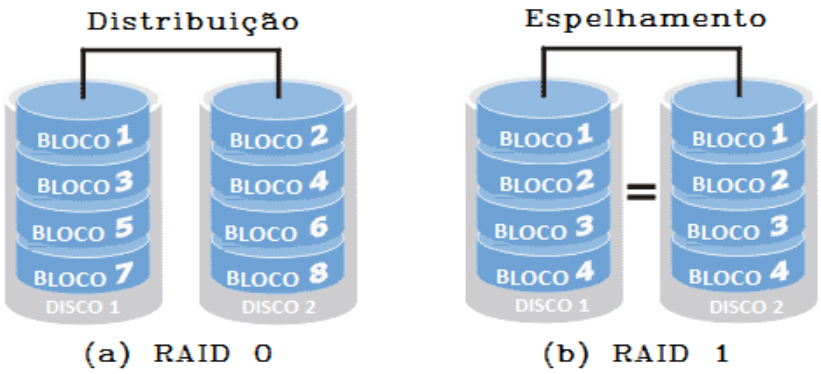

Figura 15 – Configurações dos RAID 0 e RAID 1 [ARMBRUST, VICTOR, 2012]

4.2.3 **RAID 0+1 e RAID 1+0**

Na maioria dos casos é necessário o desempenho oferecido no RAID 0 e a redundância de dados do RAID 1. Essas duas configurações combinam técnicas de distribuição e espelhamento, porém são necessários um número maior de discos, no mínimo quatro. Elas oferecem às mesmas vantagens, entretanto, a operação de reconstrução do RAID após uma falha se difere entre as duas. A figura 16 ilustra as diferenças de configuração.

No RAID 0+1 os dados inicialmente são gravados divididos entre os discos para melhorar o desempenho e depois são duplicados utilizando outros discos. No caso de falha de um disco o par de distribuição inteiro falhará, passando a funcionar como um RAID 0. Na reconstrução do RAID, o par de distribuição inteiro é reproduzido a partir do par saudável, gerando um aumento na carga de leitura dos discos bons. O que não acontece no RAID 1+0.

No RAID 1+0 ou RAID 10 o elemento base é o espelhamento. Primeiro os dados são espelhados e depois distribuídos entre os demais discos da matriz.

Quando um disco falha apenas o espelho é reconstruído e a controladora de disco replica apenas o disco bom do par espelhado.

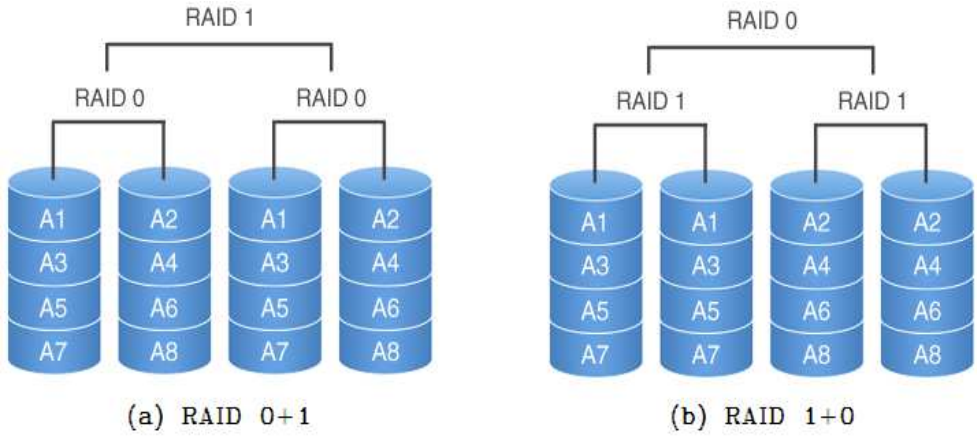

Figura 16 – Diferenças de configuração entre RAID 0+1 e RAID 1+0 [ARMBRUST, VICTOR, 2012]

4.2.4 **RAID 3**

Nesta configuração é usado paridade para melhorar a tolerância à falha e os dados são distribuídos nos discos da matriz para obter alto desempenho. Paridade é a técnica que, através de cálculos matemáticos, é possível recuperar dados perdidos caso ocorra uma falha. É uma verificação de redundância que garante a integridade, sem o custo de duplicar todos os dados. As informações de paridade são armazenadas em um disco dedicado da matriz de discos. Tem um bom desempenho na leitura dos dados, mas tem um custo na gravação, devido à paridade e pelos blocos lógicos distribuídos nos discos serem acessados consecutivamente. Já na recuperação de falha tem um desempenho inferior comparado ao espelhamento, mas tem um menor custo de espaço para garantir à tolerância a falha. A figura 17(a) ilustra esta configuração.

- **RAID 5**

Esta configuração é parecida com o RAID 3 com algumas diferenças. No RAID 5 assim como no RAID 3 é usada paridade para melhorar a tolerância à falha, porém diferencia-se porque assim como os dados a paridade é distribuída nos discos da matriz, o que não ocorre no RAID 3, onde apenas os dados são distribuídos. Outra diferença é o acesso independente aos discos, onde para os dados específicos serem acessados em um único disco não há necessidade de ler todo o bloco lógico distribuído, o que ocorre no RAID 3. Estas diferenças melhoram o desempenho de gravação em relação ao RAID 3. Esta configuração de RAID é muito versátil e esta ilustrada na figura 17(b).

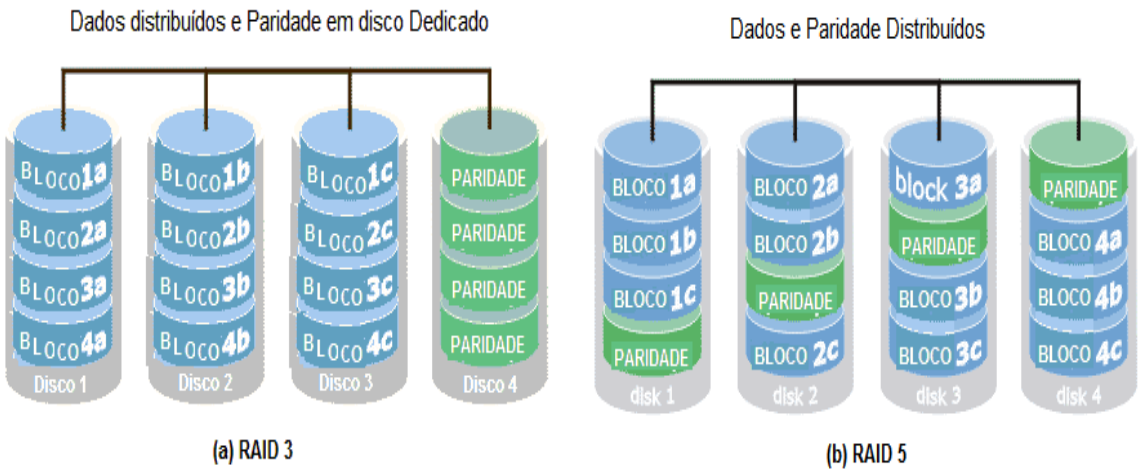

Figura 17 – Diferenças de configuração entre RAID 3 e RAID 5 [ARMBRUST, VICTOR, 2012]

Outras configurações de RAID como RAID 2, RAD 4 e RAID 6 também utilizam paridade. Porém estas configurações são pouco utilizadas devido à boa eficiência do RAID 5.

4.3 PROTOCOLOS DE COMUNICAÇÃO

O servidor se comunica com os discos, através de protocolos predefinidos, como *Integrated Device Electronics* / *Advanced Technology Attachment* (IDE/ATA), serial ATA(SATA), *Small Computer System Interface* (SCSI), *Serial Attached SCSI* (SAS) e FC. Estes protocolos são implementados na controladora de disco do servidor, e normalmente o disco é conhecido pelo nome do protocolo que suporta.

- ***Integrated Device Electronics / Advanced Technology Attachment* (IDE/ATA) e Serial ATA (SATA)**

  O IDE/ATA exprime duas especificações, o IDE que é a especificação da comunicação entre controladores de discos e placa-mãe, e o ATA que é a especificação para conexão das controladoras de discos e o meio de armazenamento.

  SATA é a especificação serial para o padrão IDE/ATA, desenvolvida para substituir a versão ATA paralela. Os discos SATA são atualmente utilizados como padrão de mercado para computadores pessoais.

- ***Small Computer System Interface* (SCSI) e *Serial Attached* SCSI (SAS)**

  O SCSI foi desenvolvido para criar um protocolo de alto desempenho, que logo virou padrão de mercado com o reconhecimento pela *American National Standards Institution* (ANSI[5]). Ele sofreu modificações, passando pelos padrões SCSI-1, SCSI-2 e SCSI-3 e evoluindo para um padrão robusto, oferecendo melhor desempenho e opções de expansão e compatibilidade em comparação

[5] ANSI é uma organização americana com a função de estabelecer quais normas desenvolvidas devem virar padrões.

com o IDE/ATA. O SCSI fornece um eficiente barramento de E/S de dados ponto-a-ponto com suporte a múltiplos dispositivos de armazenamento, o que o torna apropriado para computadores de grande desempenho como o *storage*.

A arquitetura SCSI-3 é utilizada no desenvolvimento e projetos de hardware com base nos três componentes principais.

- Protocolo de comando SCSI-3: consiste nos comandos básicos comuns a todos os dispositivos e comandos específicos exclusivos a determinada classe;
- Protocolo de camada de transporte: Conjunto de regras pela quais os dispositivos comunicam e compartilham informações;
- Interconexões da camada física: Padrão de interface, como métodos de sinalização elétrica e modos de transferência de dados.

A arquitetura SCSI-3 utiliza um conceito iniciador-destino, *initiator-target*, representando um modelo cliente servidor, onde cada dispositivo executa as seguintes funções:

- Dispositivo *initiator* SCSI: Envia comando para o dispositivo destino executar. Uma controladora SCSI é um exemplo;
- Dispositivo *target* SCSI: Executa os comandos enviados pelo *initiator*. Um dispositivo de armazenamento é um exemplo, porém em determinadas implementações a controladora SCSI pode ser um dispositivo *target*.

No modelo de comunicação *initiator-target* SCSI os dispositivos comunicam-se através de comandos SCSI, definidos no componente protocolo de comando e são independentes do tipo de interface, podendo ser SCSI paralelo, SAS e FC.

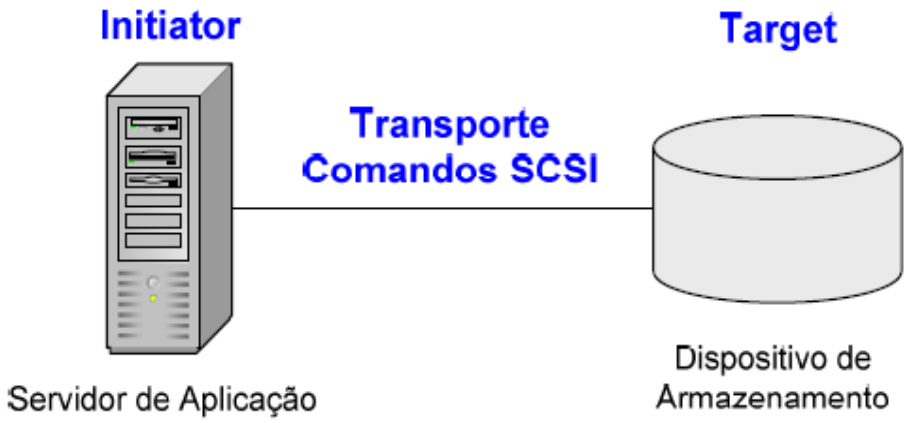

Figura 18 – Modelo de comunicação SCSI [NETO, ANTÔNIO, 2004]

SAS é a evolução do SCSI paralelo, que implementa o padrão SCSI com comunicação serial. Aborda a robustez necessária para os computadores de alto desempenho, suportando taxa de transferência de 3 Gbps e conectividade com múltiplos dispositivos.

- ***Fibre Channel Protocol* – FCP**

  É o protocolo que implementa o padrão SCSI-3 para as redes *fibre channel* (FC). FC é a rede de armazenamento de alta velocidade que utiliza como meio físico a fibra óptica, com velocidade de acesso 8,5 Gbps. É amplamente utilizado na arquitetura SAN.

4.4 REDES DE ARMAZENAMENTO

As redes de armazenamento são redes projetadas para permitir que os *hosts* tenham acesso remoto às áreas de dados no dispositivo de armazenamento. As

tecnologias existentes se diferenciam pelo meio de transporte e protocolo de comunicação.

Uma importante classificação para o entendimento dos protocolos de transporte, que serão apresentados nos próximos capítulos, é o nível de acesso aos dados, que são de dois tipos:

- Acesso em nível de bloco: é o mecanismo pelo qual os dados são acessados nos discos em baixo nível. Os dados são armazenados e recuperados dos discos, através da identificação do endereço lógico do bloco. O bloco é a unidade básica de armazenamento e é a localização dos dados no disco físico. O endereço do bloco é derivado da configuração geométrica do disco;
- Acesso em nível de arquivo: abstrai o acesso em nível de bloco. A solicitação é enviada por meio da especificação do nome e caminho do arquivo. O sistema de arquivo gerenciado pelo sistema operacional é que usa o acesso em nível de bloco correspondente, abstraindo as complexidades do endereçamento de bloco para as aplicações.

A necessidade de armazenamento atende as necessidades das aplicações, neste trabalho o *cluster* do hipervisor. Por isso é ideal o entendimento de cada tecnologia disponível para as redes de armazenamento.

5 ARMAZENAMENTO DIRETAMENTE CONECTADO E ARMAZENAMENTO CONECTADO À REDE

5.1 ARMAZENAMENTO CONECTADO DIRETO – *DIRECT-ATTACHED STORAGE* (DAS)

A arquitetura DAS tem o modelo mais simples de interconexão em um sistema de armazenamento. É a tecnologia que conecta diretamente o servidor ao dispositivo de armazenamento, utilizando protocolos de acesso em nível de bloco.

O DAS é classificado como interno ou externo, conforme a localidade do dispositivo de armazenamento em relação ao servidor.

- DAS interno: o dispositivo de armazenamento é conectado internamento ao *host,* através do barramento e fica limitado a distância, devido a curta conectividade do barramento para alta velocidade, e a quantidade de dispositivos suportadas pelo barramento. Usa protocolos IDE / ATA, SATA, SAS ou SCSI.

- DAS externo: o *host* se conecta diretamente ao dispositivo de armazenamento externo, podendo usar conexão física por fibra óptica ou cabo par trançado, usando respectivamente os protocolos FC ou *Small Computer System Interface* (SCSI). A figura 19 ilustra o DAS externo.

As principais diferença entre o DAS interno e externo é que o externo supera as limitações de distância, quantidade de dispositivos e fornece gerenciamento centralizado.

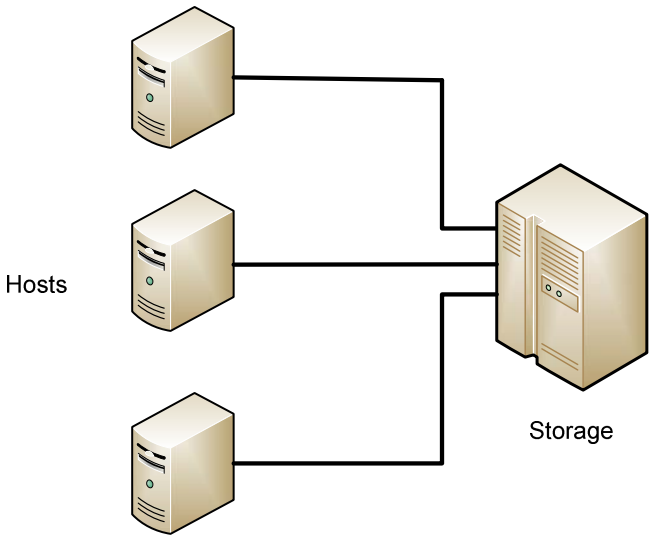

Figura 19 – Arquitetura de DAS externo

- **Vantagens e Limitações**

O DAS requer um investimento menor em relação às redes de armazenamento compartilhadas como SAN e NAS, e se mostra uma boa alternativa para a expansão de área de armazenamento para poucos servidores, sendo uma solução viável para pequenas e médias empresas. A configuração é simples e é gerenciada por ferramentas baseadas no servidor como o sistema operacional, o que torna o gerenciamento fácil com menos tarefas e elementos de hardware e software em relação às redes SAN e NAS.

Entretanto, o DAS demonstra algumas limitações como escalabilidade, falta de componentes de redundância, quantidade de servidores conectados devido ao número de portas, largura de banda limitada que restringe a capacidade de processamento de E/S de dados e limitações de distância devido à conexão direta. Porém, sua maior limitação é a incapacidade de distribuir área de armazenamento entre servidores com capacidade de armazenamento sobrando

e servidores com falta de área de armazenamento [SOMASUNDARAM, G., 2011].

Algumas destas limitações inviabilizam a utilização do DAS para a alta disponibilidade de servidores virtuais, porém o custo de uma solução de armazenamento compartilhado, SAN e NAS, muitas vezes não se justifica para pequenas e médias empresas. Pensando nisso muitas empresas têm investido em soluções alternativas de hardware e software para o uso do DAS no suporte a virtualização.

- **Alternativas do DAS para Virtualização**

Existem várias soluções alternativas de utilizar o DAS para a virtualização. Duas delas serão apresentadas, sendo uma para plataforma VMware e outra para a plataforma Microsoft.

5.1.2.1 *vSphere Storage Applicance*

A Vmware oferece na versão do vSphere 5 o vSphere *Storage Appliance* (VSA), uma solução que possibilita a criação de um *storage* virtual usando discos locais de cada *host*, e assim oferece um ambiente com alta disponibilidade e balanceamento de carga. Esta solução visa atender pequenos ambientes com até três *hosts*, onde não justificaria o investimento em soluções como NAS e SAN.

A solução implementa um *cluster* VSA, onde é criado um *storage* virtual entre os *hosts*. Estes *hosts* pertencentes ao *cluster* passam a ter acesso ao volume VSA pelo protocolo *Network File System* (NFS). Este protocolo será descrito no próximo tópico. Estes volumes são replicados entre os *hosts* para oferecer

proteção à falha [MANFRIN, ALEXANDER, 2011]. A figura 20 ilustra esta solução.

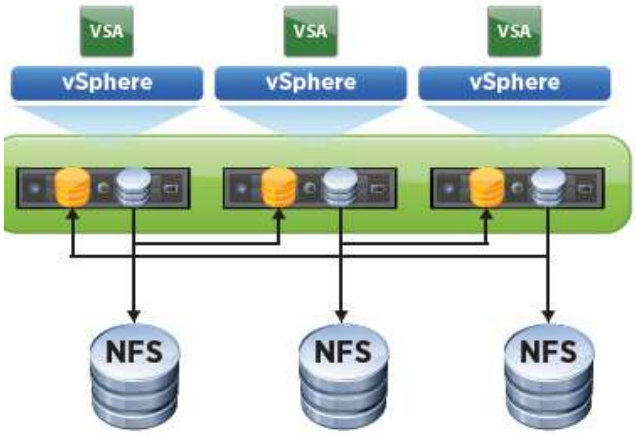

Figura 20 – vSphere *Storage Applicance* [MANFRIN, ALEXANDER, 2011]

5.1.2.2 *VM6 Manage Cloud Platform*

A solução VM6 Software é compatível com o Hyper-V da Microsoft. O VM6 deve ser instalado em servidores Windows Server 2008 com Hyper-V. Ele irá criar uma nuvem entre os servidores e assumirá o controle do armazenamento não utilizado, criando um armazenamento compartilhado virtual, virtual SAN. Nesta virtual SAN pode-se criar partições virtuais que serão vistas por todos os nós da nuvem, permitindo que máquinas virtuais possam ser migradas entre os nós [VM6 SOFTWARE, 2011].

Existem outras soluções de hardware e software com o objetivo de atender a necessidade das pequenas e médias empresas em utilizar o DAS para virtualização. As redes de armazenamento compartilhadas, NAS e SAN, são mais adequadas a grandes empresas com grande parque tecnológico e alto investimento na aquisição e custo operacional.

5.2 ARMAZENAMENTO CONECTADO À REDE – *NETWORK-ATTACHED STORAGE* (NAS)

O NAS é um dispositivo de compartilhamento de dados, em nível de arquivo, conectado a rede local IP. Seu principal objetivo é a eliminação de múltiplos servidores de arquivo, permitindo de forma rápida e direta o compartilhamento de arquivos com baixo custo de gerenciamento.

O dispositivo NAS é um sistema de armazenamento e serviço de arquivo dedicado de alto desempenho, e tem seus componentes de hardware e software integrados para melhor atender as especificações do serviço de arquivo. Tem um sistema operacional próprio de tempo real dedicado ao serviço de arquivo, que é otimizado para executar com maior eficiência as tarefas de E/S de dados. Ele serve a vários clientes através da rede IP e a maioria dos dispositivos NAS suportam múltiplas placas de redes. Alguns fornecedores de NAS suportam recursos como *cluster* para obter alta disponibilidade.

5.2.1 **Componentes do NAS**

O dispositivo NAS é composto pelos seguintes componentes de hardware e software:

- Hardware: NAS *head* compostos de CPU e memória; matriz de armazenamento que consiste em uma quantidade de discos e configurações de RAID; placas de redes que são interfaces que fornecem conectividade com a rede IP;
- Software: sistema operacional otimizado para o gerenciamento e funcionalidades do compartilhamento de arquivos; protocolos de compartilhamento de arquivos, normalmente *Network File System*

(NFS) para sistemas da família Unix e *Common Internet File System* (CIFS) para sistemas Windows; protocolos de armazenamento que conecta e gerencia os recursos de discos físicos.

O NAS *head* provê conectividade aos clientes e atende as solicitações de E/S de dados através dos protocolos NFS ou CIFS. O cliente realiza o acesso ao ambiente NAS por meio da rede IP. A figura 21 ilustra os componentes NAS e o acesso pelos clientes.

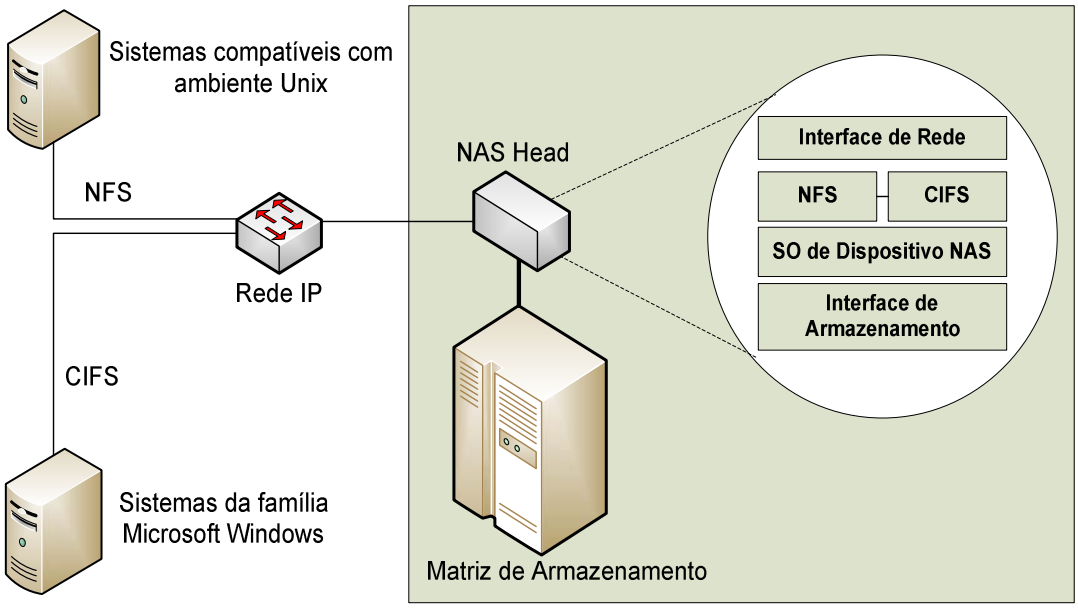

Figura 21 - Componentes do NAS [SOMASUNDARAM, G., 2011]

5.2.2 **Implantação do NAS**

O dispositivo NAS pode ser implementado de duas formas: integrado ou *gateway*. NAS integrado tem todos os componentes e sistema de armazenamento em um único compartimento. Esta solução pode permitir redimensionamento da capacidade de armazenamento ou NAS *head*, porém há um limite para esta expansão. Já o NAS *gateway* consiste em um NAS *head*

independente conectado a uma ou mais matrizes de armazenamento. Esta solução é mais redimensionável, permitindo que NAS *heads* e matrizes de armazenamento sejam expandidas independentes quando necessárias.

5.2.3 **Protocolos de Compartilhamento**

Para lidar com as solicitações de leitura e gravação de arquivos para um sistema de arquivo remoto o NAS suporta dois protocolos de compartilhamento: o NFS utilizado predominantemente em ambiente Unix e CIFS utilizado em ambiente Microsoft.

5.2.3.1 *Network File System* (NFS)

O NFS é um protocolo baseado no modelo cliente/servidor para o compartilhamento de arquivos. Utiliza o *Remote Procedure Call* (RPC[6]) para comunicação interprocessos entre os computadores cliente e servidor. Através de um conjunto de RPCs o NFS oferece acesso ao sistema de arquivo remoto, disponibilizando as seguintes operações de execução: pesquisa de arquivos e diretório; abrir, ler, gravar e fechar arquivos; alterar atributos de arquivos; modificar conexões de arquivos e diretórios.

Existem três versões de NFS em uso atualmente conforme [SOMASUNDARAM, G, 2011]:

- NFS versão 2 (NFSv2): Usa o protocolo de transporte *User Datagram Protocol* (UDP), que fornece conexão entre cliente e servidor sem estado, ou seja, não mantem nenhuma tabela para manter informações

sobre ponteiros e arquivos abertos durante a conexão. Operações como bloqueio são executadas fora do protocolo;

- NFS versão 3 (NFSv3): Mais usado, utiliza os protocolos UDP ou *Transmission Control Protocol* (TCP) e é baseado em um projeto sem estado de conexão. Inclui novos recursos como tamanho de arquivos da arquitetura x64, gravação assíncrona e atributos adicionais para arquivos;

- NFS versão 4 (NFSv4): Utiliza o protocolo de transporte TCP com estado de conexão. Oferece maior segurança.

5.2.3.2 *Common Internet File System* (CIFS)

O CIFS é um protocolo de aplicação baseado no modelo cliente/servidor, uma variação aberta do protocolo *Server Massage Block* (SMB). Através do protocolo CIFS clientes podem ter acesso a arquivos remotos e compartilhar esses arquivos com outros clientes por meio de bloqueios especiais. Os usuários referem-se ao sistema de arquivo remoto através de um esquema de nomenclatura \\servidor\compartilhamento.

5.2.4 **Vantagens e Limitações do NAS**

O NAS traz alguns benefícios: acesso abrangente às informações, onde muitos clientes podem ser servidos por um NAS e um cliente pode ter acesso a muitos NAS; armazenamento centralizado, eliminado múltiplos servidores de

---

[6] Remote Procedure Call (RPC) é uma tecnologia de comunicação de processos que permite um programa de computador chamar um processo em outro computador conectado à rede. É implementado por um modelo cliente / servidor onde o cliente envia uma mensagem para o servidor remoto para executar um procedimento especifico.

arquivos; gerenciamento simplificado, através de console central de gerenciamento do sistema de arquivo; capacidade de expansão, possibilitando a expansão, conforme necessidade do tipo de aplicação; infraestrutura simplificada, utilizando a rede local IP existente para troca de dados; alta disponibilidade, oferecendo recursos de replicação, recuperação de dados, redundância dos componentes de rede e pode utilizar tecnologia de cluster para *failover*.

A largura de banda e a latência são as principais limitações que afetam o desempenho e a disponibilidade do NAS, uma vez que utiliza a rede IP. Com isso, é necessário ter uma rede local bem planejada para atender as necessidades do uso do NAS. Outros fatores podem influenciar o desempenho: retransmissão de dados; uso saturado dos equipamentos intermediários da rede; tempo de autenticação nos serviços de diretórios como LDAP e NIS; alto nível de utilização dos dispositivos NAS. Algumas ações podem ser realizadas para melhorar o desempenho como a criação de rede local virtual*, Virtual Local Area Network* (VLAN); estabelecer unidade máxima de transmissão, *Maximum Transmission Unit* (MTU); e estabelecer o tamanho da janela de transmissão TCP. Assim como agregação de links e redundância de rede para garantir alta disponibilidade.

5.2.5 **Utilizando NAS para Virtualização de Servidores**

O NAS pode ser utilizado como alternativa de armazenamento para a alta disponibilidade de servidores virtuais através do protocolo NFS. Porém, nem todos os softwares de virtualização oferecem suporte a este protocolo, é o caso da hyper-V da Microsoft. As considerações para escolha do NAS como

alternativa de armazenamento passam por questões como custo relativamente baixo, desempenho, disponibilidade e facilidade de gerenciamento.

A facilidade proporcionada pela utilização da infraestrutura de rede IP já existente aliada ao alto custo das redes SAN fazem com que o NAS seja, na maioria das vezes, a escolha de armazenamento por pequenas e médias empresas que implementam virtualização.

Atributos adicionais do armazenamento NAS podem ser levados em consideração na escolha como: provisionamento fino, *thin provisioning*, eliminação de duplicação e facilidade de *backup* e restauração dos servidores virtuais.

Thin provisioning é o recurso que oferece melhor utilização da capacidade de armazenamento, somente alocando espaço no volume quando dados são gravados. Na maioria dos casos a alocação do espaço ocorre na criação do volume, o que pode significar desperdício quando este espaço alocado não é utilizado. O protocolo NFS oferece este recurso por padrão.

Uma solução de *backup* e restauração possível é através da matriz de armazenamento baseado na tecnologia de *snapshot*[7].

Uma consideração importante é referente à segurança em uma área de armazenamento NFS, onde o tráfego é transmitido como texto não criptografado através da rede local. Por este motivo é recomendado usar NFS em redes confiáveis, de preferência em uma rede dedicada ou em uma VLAN isolada.

---

[7] *Snapshot* é a tecnologia que permite realizar uma copia instantânea do sistema de arquivos proporcionando proteção e recuperação de dados.

Em relação ao desempenho para a maioria dos ambientes virtuais os protocolos NFS e iSCSI, que utilizam rede IP, fornecem desempenho de leitura e gravação apropriado. Conforme testes realizados pela VMware o desempenho é semelhante, com ligeiro aumento de processamento no host para as transações NFS [VMWARE5, 2008]. Este assunto será melhor discutido no capítulo 7.

6 REDES DE ÁREA DE ARMAZENAMENTO – *STORAGE AREA NETWORK* (SAN)

A SAN é uma rede dedicada de alto desempenho para a conexão de servidores e dispositivos de armazenamento compartilhado, onde existem vários caminhos disponíveis para o transporte de dados entre dois pontos. Tradicionalmente a conexão é realizada através da rede *fibre channel* (FC), que utiliza cabeamento de fibra óptica e dispositivos de interconexão como *hubs* FC e *switches* FC, proporcionando maior disponibilidade, flexibilidade e segurança.

A combinação efetiva das vantagens da SAN junto com o FC se tornou uma solução de armazenamento muito utilizada por grandes empresas. Porém, as organizações precisam do desempenho e a capacidade de expansão da SAN tradicional, com a facilidade de uso e o menor custo da solução NAS, baseada no protocolo IP.

O surgimento da tecnologia IP com suporte a acesso a dados em nível de blocos, iSCSI, permitiu o uso da SAN através da rede IP. Estas duas tecnologias SAN FC, ou simplesmente SAN, e SAN IP são as mais utilizadas para alta disponibilidade de servidores virtuais atualmente.

6.1 REDES DE ÁREA DE ARMAZENAMENTO SOBRE FC – *STORAGE AREA NETWORK* (SAN FC)

A SAN FC proporciona alto desempenho no acesso aos dados, podendo chegar à taxa de transferência de 10 Gbps, alta disponibilidade e confiabilidade com caminhos redundantes e espelhamento de dados, escalabilidade provendo capacidade de adicionar mais recursos à infraestrutura e gerenciamento

centralizado dos recursos. Porém, para prover todos esses benefícios é necessário uma complexa infraestrutura.

- **Componentes da SAN**

  A arquitetura SAN é composta de três componentes básicos:

  - Componentes de interconexão: responsável pela conectividade física entre os servidores e dispositivos de armazenamento;
  - Protocolo de Transporte: responsável pelo transporte das mensagens SCSI. O protocolo utilizado na SAN tradicional é o protocolo fibre channel (FCP);
  - Protocolo de comunicação: responsável pela comunicação entre os elementos servidores, *initiator,* e dispositivos de armazenamento, *target*. O protocolo utilizado é o SCSI.

6.1.2 Cabeamento *Fibre Channel*

A SAN utiliza cabeamento de fibra óptica como componente de interconexão. Existem dois tipos de fibras ópticas: monomodo e multimodo. A figura 22 ilustra os dois tipos.

- Multimodo: *Multi-Mode Fiber* (MMF) emite múltiplos sinais de luz simultaneamente, projetados no núcleo do cabo em diferentes ângulos. Estes sinais tendem a dispersar e colidir, o que enfraquece a força do sinal. Este processo é conhecido como Dispersão Modal. Devido a esta atenuação, os cabos MMF são utilizados normalmente em distâncias de até 500 metros;
- Monomodo: Single-Mode Fiber (SMF) emite apenas um único raio de luz projetado através do pequeno núcleo. Este sinal de luz viaja através

do núcleo em linha reta, limitando a dispersão modal. O monomodo é utilizado em distâncias maiores, máxima de até 10 KM, devido a baixa atenuação de sinal, limitado apenas pela potência do laser no transmissor e pela sensibilidade do receptor.

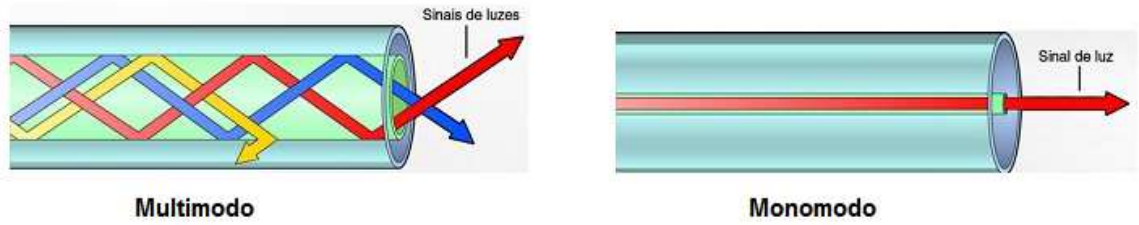

Figura 22 – Fibras monomodo e multimodo

6.1.3 **Topologia *Fibre Channel***

A arquitetura *Fibre Channel* pode ser implementada através das topologias: ponto-a-ponto, laço arbitrado, *fabric*, *fabric core-edge* e *mesh*.

6.1.3.1 *Ponto-a-ponto*

Na topologia ponto-a-ponto a conexão é direta entre duas portas, ligando diretamente dois dispositivos, através de uma conexão dedicada.

6.1.3.2 *Laço Arbitrado – Fibre Channel Arbitrated Loop (FC-AL)*

A topologia FC-AL utiliza o hub FC como dispositivo de conexão, onde os nós[8] são conectados a um loop compartilhado, onde os dados são transmitidos de nó para nó até chegar ao destino. Por este motivo o acesso ao meio é ordenado e tem características de topologia lógica em anel e física em estrela. Os nós compartilham a largura de banda e apenas um nó pode executar operações de leitura e gravação de cada vez. Os dispositivos no loop devem

[8] Os nós podem ser entendidos como servidores, dispositivos de interconexão e dispositivos de armazenamento.

arbitrar para obter o controle sobre o meio. Pouco usada. A figura 23 ilustra esta topologia.

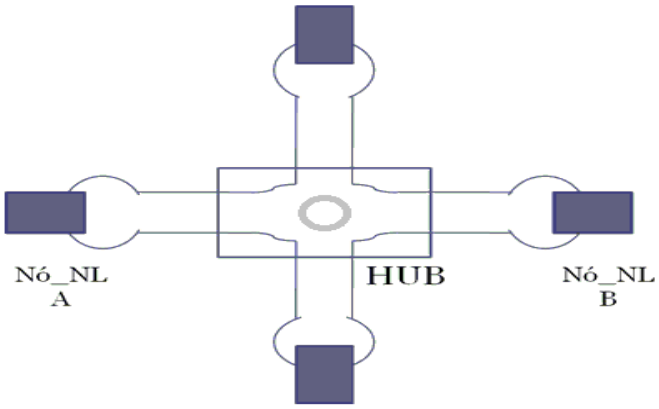

Figura 23 – Topologia *Fibre Channel Arbitrated Loop* [DUARTE, OTTO, 2008]

6.1.3.3 *Fibre Channel Switched Fabric (FC-SW)*

A topologia FC-SW oferece dispositivos interconectados, largura de banda dedicada e capacidade de expansão. Utiliza um ou mais *fabric switches*, onde os dados são direcionados diretamente entre as portas físicas. Desta forma os nós não compartilham a largura de banda, proporcionando melhor capacidade de expansão, uma vez que a adição e remoção de um nó não afeta o tráfego em andamento.

O FC-SW também é conhecido como conexão de *fabric.* O *fabric* é um espaço virtual no qual todos os nós se comunicam em uma rede, podendo ser criado com um ou mais switches e cada switch da *fabric* contém um identificador de domínio único. A figura 24 ilustra esta topologia.

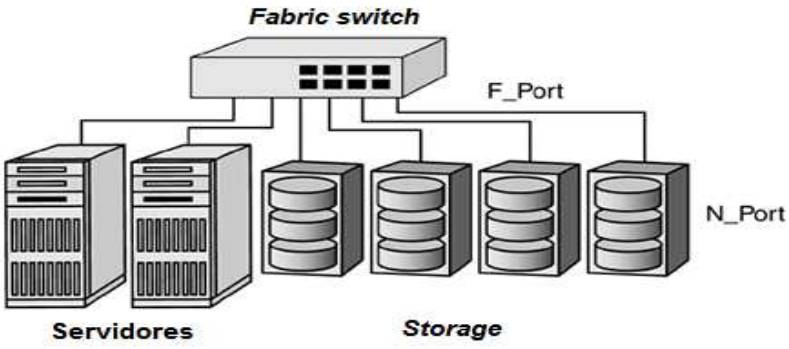

Figura 24 – Topologia *Fibre channel switched fabric* (FC-SW) [DUARTE, OTTO, 2008]

6.1.3.4 *Fabric Core-Edge*

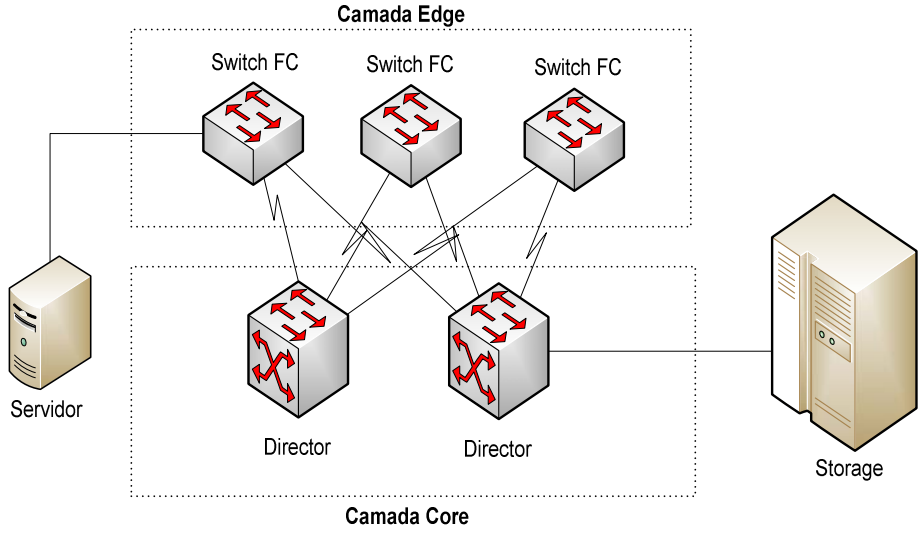

Figura 25 – Topologia Core-Edge

Esta topologia oferece alta disponibilidade e aumenta a conectividade da *fabric*. Há duas camadas de switches: camada *edge*, onde ficam os switches de ponta e são conectados os hosts, e a camada *core*, onde são utilizados os *directors* e são conectados os dispositivos de armazenamento. Os switches de

ponta estão ligados aos *directors*[9] através da ligação inter-switch, *inter-switch link* (ISL). A figura 25 ilustra esta configuração.

6.1.3.5 *Mesk*

Na topologia *mesk* cada switch é conectado diretamente ao outro, através da ISL. Existem dois tipos de topologias *mesk*: integral ou parcial que são ilustradas na figura 26. Na topologia *mesk* integral todos os switches da rede são interligados servindo tráfego *host-storage*, de forma que no máximo uma ILS ou um salto é necessário para o tráfego *host-storage*.

Na topologia *mesk* parcial diversos saltos ou ILSs são necessários para o tráfego *host-storag.* Isso porque os switches são interligados de forma a ter caminhos redundantes a todos os nós, porém, não são todos interligados. A figura 26 ilustra os dois tipos de topologia *mesk*.

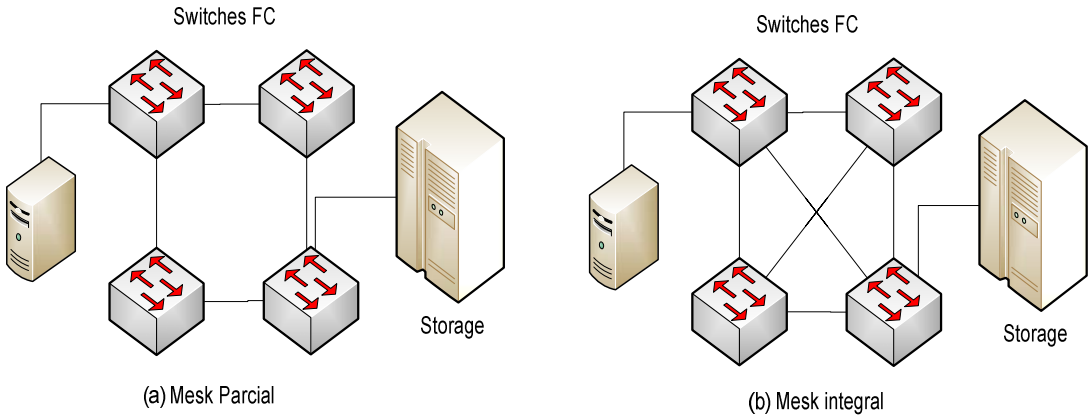

Figura 26 – Topologia *mesk* parcial e *mesk* integral [SOMASUNDARAM, G., 2011]

[9] *Directors* são semelhantes aos switches FC, porem com maior capacidade de portas e tolerância a falha, usados em implementações *Fabric core-edge*.

6.1.4 **Portas *Fibre Channel***

A infraestrutura Fibre Channel utiliza um modelo cliente/servidor com base no modelo SCSI, onde cada ponto de interconexão é denominado de porta *Fibre Channel,* que possui a finalidade de estabelecer conexão entre os elementos do sistema de armazenamento.

A figura 27 ilustra os tipos de portas suportadas na arquitetura Fibre Channel, que são:

- *Node Port* – N_Port: é utilizada para estabelecer uma conexão com um dispositivo de interconexão como um *fabric* switch. Geralmente é uma porta do servidor ou do dispositivo de armazenamento, uma HBA;
- *Fabric Port* – F_ports: porta existente no *fabric* switch que apenas pode se conectar a N_port;
- *Loop Port* – L_port: são utilizadas para o compartilhamento de largura de banda em uma topologia FC-AL, que utiliza o hub *fibre channel* como dispositivo de interconexão. Pode ter 2 designação: quando o servidor se conecta ao hub FC, através da *Node Loop Port*, é designada de NL_Port, ou quando tem-se uma configuração onde o switch FC se conecta a um hub FC, é designada *Fabric Loop Port*, FL_Port;
- *Expansion Port* – E_Port: portas de expansão da topologia *Fibric,* utilizadas para interligar switches FC. Apenas são conectadas em portas do mesmo tipo E_port;
- *Generic Port* – G_Port: porta genérica que pode funcionar como uma E_Port ou F_Port. Determina automaticamente sua função ao ser inicializada.

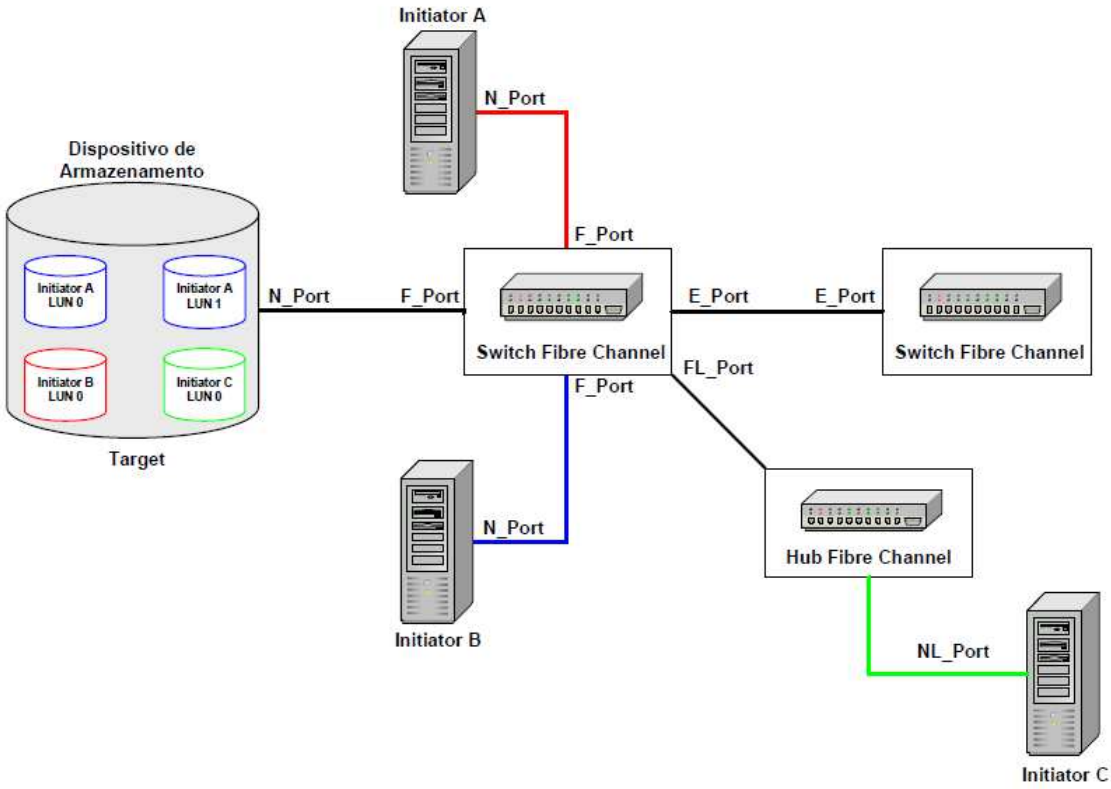

Figura 27 – Portas *Fibre Channel* [NETO, ANTÔNIO, 2004]

6.1.5 ***Fibre Channel Protocol* (FCP)**

O protocolo *Fibre Channel* implementa o padrão SCSI-3 serial em uma rede FC. Na arquitetura FCP os *hosts* identificam os dispositivos de armazenamento externos ou remotos conectados à rede SAN como dispositivos locais.

As principais vantagens do FCP são:

- Largura de banda contínua em distâncias longas; e
- Suporta um número maior de dispositivos endereçáveis em uma rede. Teoricamente pode suportar mais de 15 milhões de endereços em uma rede.

Possui características de transporte de canal e fornece velocidade de até 8.5 Gbps.

6.1.5.1 *Pilhas de Protocolos Fibre Channel*

Protocolo de comunicação é melhor entendido como uma estrutura de camadas independentes, onde cada camada se comunica com a camada subsequente através de protocolos definidos. O FCP é definido em cinco camadas: FC-0 até FC-4. A figura 28 ilustra a pilha de protocolos *fibre channel*.

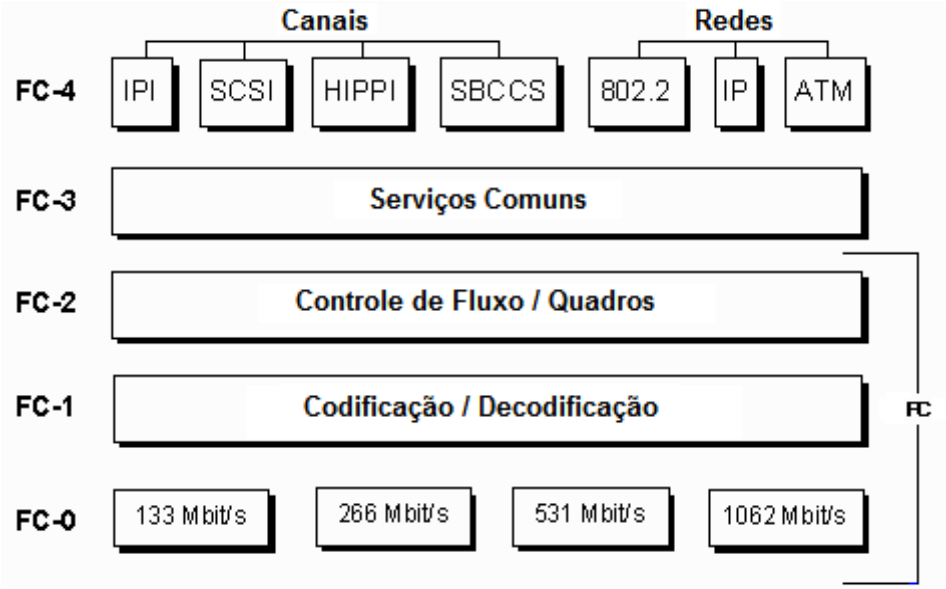

Figura 28 – Pilha de protocolos *Fibre channel* [JAVVIN, 2012]

FC-4 é o protocolo de camada superior que define as interfaces com a aplicação, através de diversos protocolos como SCSI, HIPPI, ATM e IP.

FC-3 é a camada usada para serviços comuns exigidos para recursos avançados, como grupo de busca e *multicast*.

FC-2 é a camada de transporte que contém a carga útil, endereçamento de portas de origem e destino e informações de controle de conexão.

FC-1 é a camada de transmissão, onde são definidas as regras de codificação e decodificação serial, caracteres especiais usados e controle de erro.

FC-0 é a camada de interface física, mais baixa da pilha, define a interface física, mídia e a transmissão de bits brutos.

6.1.6 **Zoneamento**

O zoneamento é uma função do switch FC que permite segmentar logicamente em grupos os dispositivos conectados no *fabric*, de forma que apenas os membros da mesma zona possam estabelecer conexão. A figura 29 ilustra esta função.

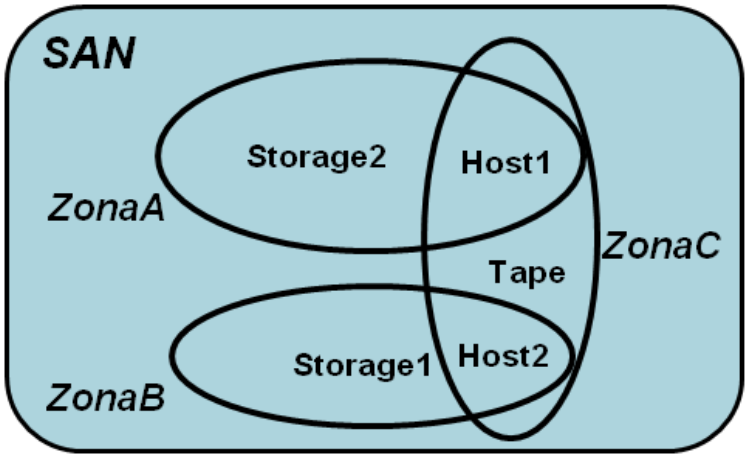

Figura 29 - Exemplo de zoneamento [DUARTE, OTTO, 2008]

A figura 30 ilustra os tipos de zoneamento, que podem ser categorizado de três tipos:

- Zoneamento de portas: Usa os endereços FC das portas físicas para determinar a zona. O acesso aos dados é determinado pela porta física do switch a qual o nó está conectado. Este endereço é atribuído dinamicamente quando o dispositivo se conecta à *fabric*. Este tipo de zoneamento é mais seguro, porém uma alteração na configuração do *fabric* requer uma reconfiguração da zona;

- Zoneamento WWN: Utiliza o *World Wide Names*[10] (WWN) para determinar a zona. Uma vantagem importante é sua flexibilidade, permitindo que a SAN seja “recabeada” sem reconfigurar a zona;
- Zoneamento Misto: Combina o zoneamento de portas com o WWN, permitindo que determinada porta seja associada ao WWN de um nó.

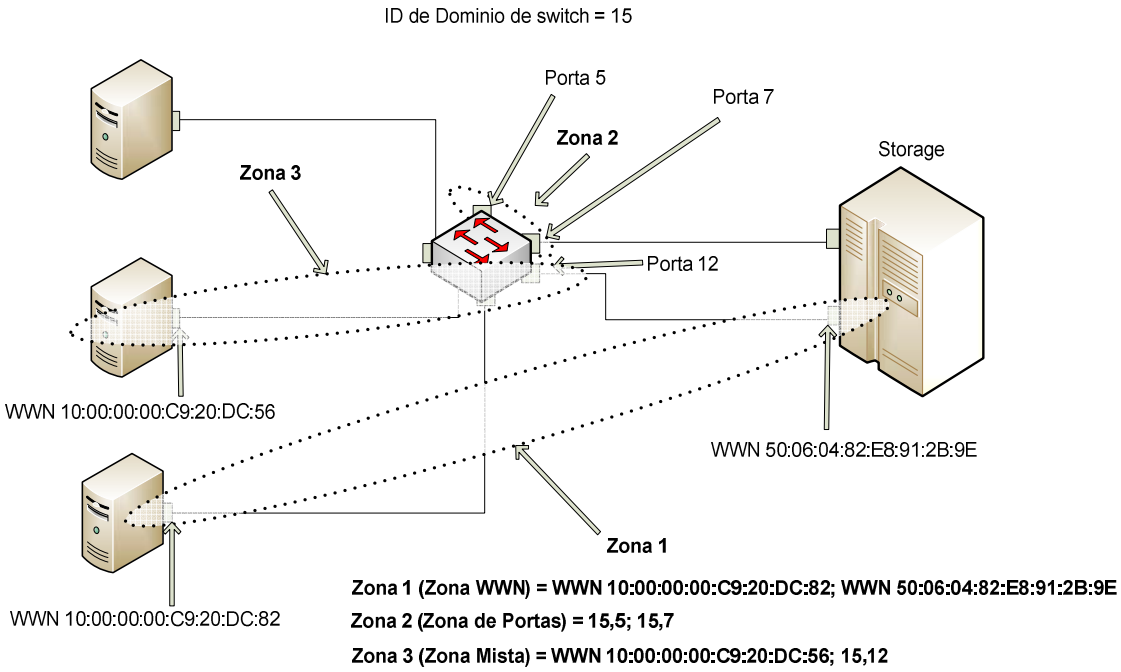

Figura 30 – Tipos de zoneamento [SOMASUNDARAM, G., 2011]

6.2 REDES DE ÁREA DE ARMAZENAMENTO SOBRE IP – *STORAGE AREA NETWORK* IP (SAN IP)

A SAN tradicional apresenta inúmeras vantagens e se tornou a principal tecnologia de rede de armazenamento. No entanto, o surgimento do protocolo *Internet Small Computer System Interface* (iSCSI) proporcionou a utilização da arquitetura IP para redes SAN, surgindo assim a SAN IP.

[10] WWN é um identificador único usado para identificar HBAs. O identificador é “queimado” no hardware, semelhante aos endereços *Media Acess Control* (MAC) das redes Ethernet.

A SAN IP surge como uma opção atrativa para as empresas, tendo em vista que possibilita a utilização da infraestrutura ethernet já existente, o que reduz custo de implementação e manutenção, uma vez que equipamentos *fibre channel* são mais caros e a arquitetura ethernet está bem difundida no mercado tecnológico.

Outro beneficio é em relação à taxa de transferência, que com a padronização dos 10 Gbps Ethernet não há perda de desempenho para as redes *fibre channel*, e o protocolo iSCSI tende a ganhar mais espaço nas redes de armazenamento.

6.2.1 **Protocolo iSCSI**

O protocolo iSCSI possuí o mesmo objetivo do protocolo *fiber channel,* que é prover o transporte do protocolo SCSI entre os servidores de aplicação e os dispositivos de armazenamento, tendo como principal diferença o meio físico utilizado.

6.2.1.1 *Conectividade iSCSI*

O modelo de comunicação é o mesmo do protocolo SCSI, onde normalmente o servidor é *initiator* e o dispositivo de armazenamento *target*. O protocolo iSCSI foi desenvolvido com objetivo principal de disponibilizar acesso a dispositivos padrão SCSI-3 conectados em uma rede TCP/IP, podendo ser implementado através de software ou hardware. A figura 31 ilustra as implementações de comunicação iSCSI.

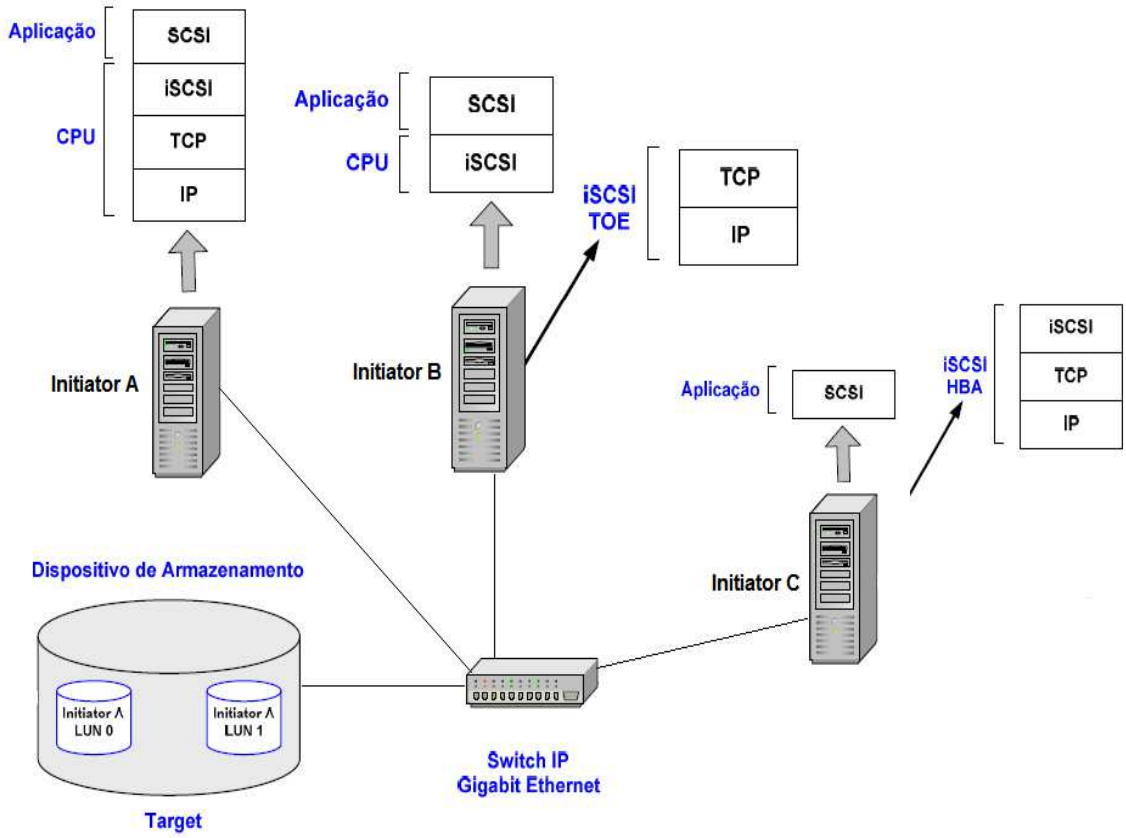

Figura 31 – Implementações de comunicação iSCSI [NETO, ANTÔNIO, 2004]

O iSCSI implementado por software, o software iSCSI, é instalado no sistema operacional do servidor e utiliza a interface de rede ethernet para comunicação com o dispositivo de armazenamento. Nesta implementação todo o processamento de encapsulamento do protocolo SCSI na pilha TCP/IP é realizado pela CPU do servidor. Um exemplo é o software Microsoft iSCSI *Initiator.*

Já a implementação por hardware pode ser realizada de duas formas: iSCSI TOE e iSCSI HBA.

O iSCSI TOE é uma interface de rede com a finalidade de realizar o processamento da pilha TCP/IP, ajudando a aliviar a carga de processamento do servidor que continua com o processamento do protocolo iSCSI entregando o

pacote à interface iSCSI TOE, que realiza o encaminhamento para o destino por meio do protocolo TCP/IP. Esta solução melhora o desempenho mas a funcionalidade iSCSI ainda é realizada por um software *initiator*.

O iSCSI HBA realiza toda a carga de processamento da pilha de protocolos TCP/IP e iSCSI na interface, fornecendo melhor desempenho. O uso de uma interface iSCSI HBA é mais simples, uma vez que não há necessidade de instalação de software no sistema operacional, porém o custo é maior em comparação às outras interfaces de rede apresentadas.

6.2.1.2 *Topologias para Conectividade iSCSI*

As topologias utilizadas para a implementação do protocolo iSCSI são categorizadas por dois tipos: nativa e com ponte. A figura 32 ilustra estas topologias.

- Topologia Nativa: Nesta topologia não existem componentes *fibre channel,* a comunicação é toda realizada através do protocolo TCP/IP;
- Topologia com ponte: permite a existência de comunicação *fibre channel* e IP, realizando a ponte entre os protocolos iSCSI-FC. Um exemplo é o servidor *initiator* que se comunica através da rede IP e o dispositivo de armazenamento, *target,* que se comunica através de uma conexão FC.

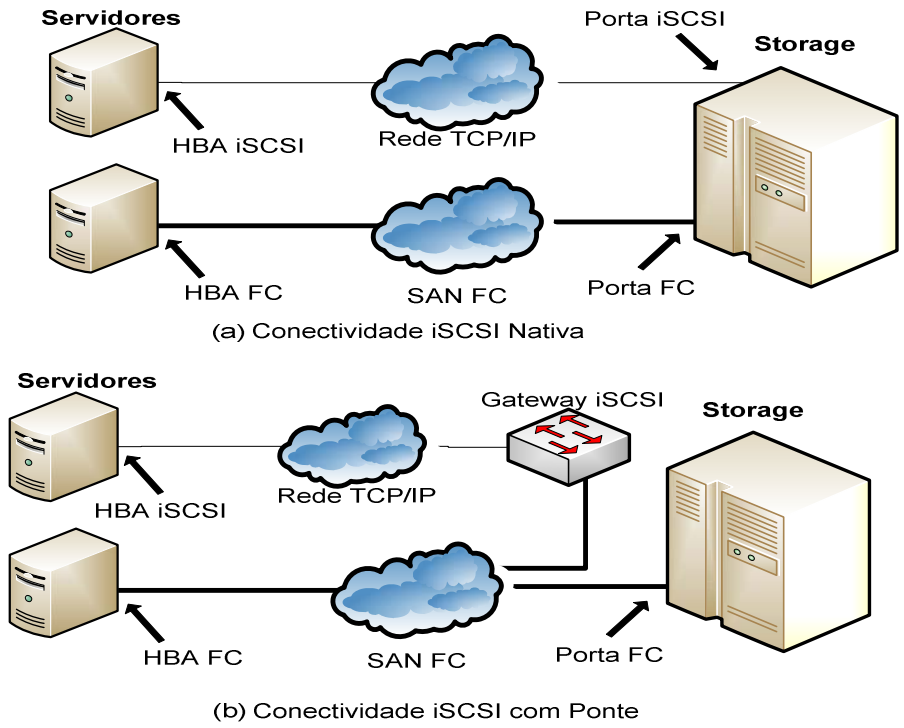

Figura 32 – Topologias de conectividade iSCSI Nativa e com Ponte

6.2.1.3 *Pilha de Protocolos da Arquitetura iSCSI*

O protocolo SCSI é baseado no modelo cliente/servidor e trabalha no nível da camada de aplicação do modelo OSI.

O protocolo iSCSI trabalha na camada de sessão e é responsável por iniciar uma sessão confiável entre dispositivos iSCSI realizando o encapsulamento das mensagens SCSI.

O protocolo TCP é usado para garantir um serviço de entrega confiável, atuando na camada de transporte: é responsável pelo controle de fluxo das mensagens, janela de transmissão, recuperação de erros e retransmissão.

O IP atua na camada de rede e provê endereçamento e conectividade global. Já o protocolo da camada de enlace permite comunicação nó-a-nó para cada salto, através da rede.

A figura 33 ilustra a ordem de encapsulamento dos protocolos e suas respectivas camadas do modelo OSI.

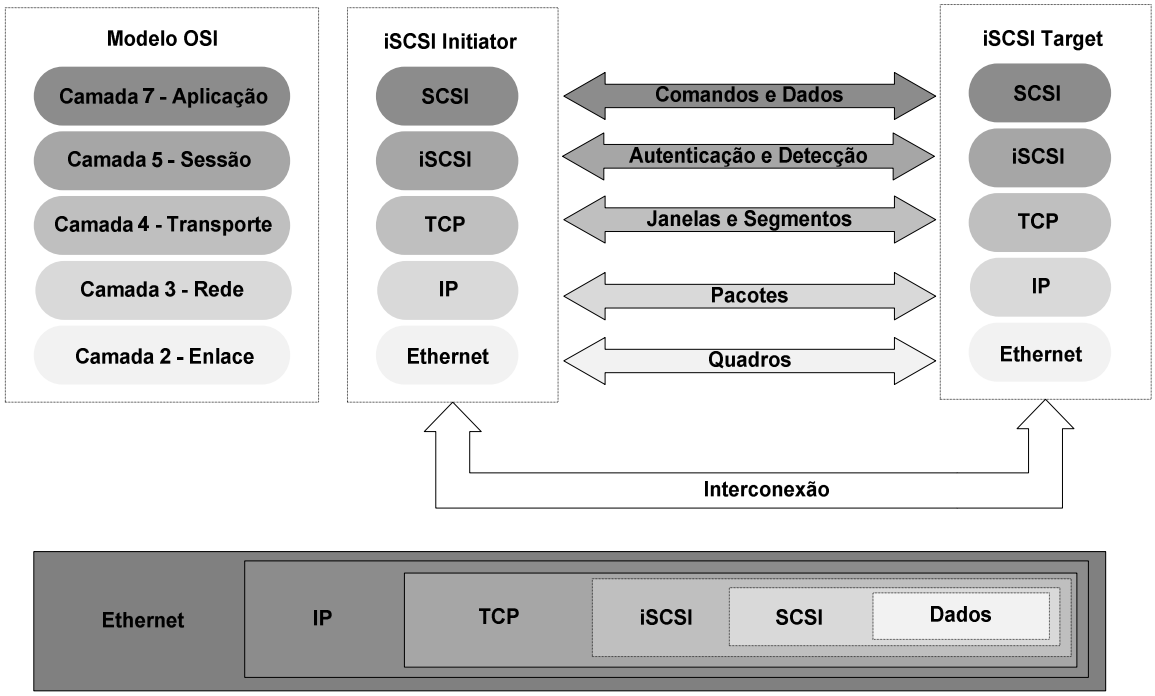

Figura 33 – Pilha de protocolos da arquitetura iSCSI [NETO, ANTÔNIO, 2004]

6.2.1.4 *Sessão iSCSI*

Uma sessão iSCSI pode ser definida como um enlace lógico entre as entidades *initiator* e *target* para a troca de mensagens e dados. A sessão iSCSI é responsável pela manipulação do *login*, autenticação, detecção de destinos e gerenciamento da sessão. Ela é constituída de pelo menos uma conexão TCP/IP entre as entidades. Quando a sessão iSCSI é estabelecida é criado um ID de sessão que inclui um ID de sessão *initiator* (ISID) e um ID de sessão *target* (TSID).

A sessão iSCSI pode servir a dois motivos: realizar a detecção de dispositivos disponíveis para o *initiator,* através do serviço *Internet Storage Name Service* (iSNS) e realizar a operação normal de troca de dados entre as entidades.

A detecção de dispositivos pode ser realizada de duas formas: detecção por *SendTargets*, onde a configuração é manual no *initiator,* sendo configurado a porta, endereço IP do *target* e nome iSCSI; ou iSNS, onde a detecção é automática.

O iSNS é o serviço onde as entidades *initiator* e *target* podem se registrar, e sempre que um *initiator* precisa conhecer um destino que tenha permissão de acesso realiza a consulta ao serviço iSNS. Não é muito usado.

A configuração manual depende de três informações: nome de domínio ou endereço IP do dispositivo de armazenamento; a porta de conexão TCP, por exemplo, 3260; e o nome iSCSI, um identificador mundial único que nomeia as entidades *initiator* e *target*.

Nomes iSCSI são fornecidos em dois formatos:

- iSCSI *Qualifield Name* (IQN): é um nome gerado automaticamente formado do prefixo “iqn” mais uma data e o nome de domínio do dispositivo. Exemplo: iqn.2008-02.com.exemplo. O nome IQN pode ser configurado manualmente, porém precisará ser mantido como um nome único;
- *Extended Unique Identifier* (EUI): é um identificador global único baseado no padrão de nome IEEE[11] EUI-64. O ID EUI compreende o prefixo “eui” seguido por um numero hexadecimal de 16 caracteres, onde 24 bits identificam a empresa, atribuídos pelo IEEE, e 40 bits para

[11] *Institute of Electrical and Electronics Engineers* (IEEE) é uma associação profissional com sede em Nova York que se dedica a promover a inovação tecnológica e excelência. Um de seus mais importantes papéis é o estabelecimento de padrões para a área de tecnologia e telecomunicações.

um Identificador único, como um número de série. Exemplo: eui.0300732A32598D26.

6.2.1.5 *Manipulação de Erros e Segurança iSCSI*

O protocolo iSCSI define mecanismos de recuperação de erros, considerando os possíveis problemas na entrega de dados decorrentes das redes IP. Conexões TCP ocasionalmente podem falhar, o *checksum*[12] de 16 bits utilizado pelo protocolo como mecanismo de detecção de erros pode não ser suficiente.

O protocolo iSCSI define mecanismos para tratamento desses erros, sem a necessidade de falha para o comando SCSI correspondente. O sequenciamento de comandos é um mecanismo utilizado para o controle de fluxo. Os comandos que falham são detectados a partir dos números de sequência.

A detecção e recuperação de erros no protocolo iSCSI pode ser classificado de três tipos:

- *Session Recovery*: é o tipo mais básico de recuperação, quando um erro é detectado a sessão iSCSI é terminada. Todas as conexões TCP são fechadas e todos os comandos SCSI pendentes são completados com *status* de erro apropriado. Em seguida uma nova sessão é estabelecida e o protocolo SCSI pode então refazer os comandos que falharam na execução. O protocolo iSCSI passa a recuperação dos comandos para o protocolo superior;
- *Digest Failure Recovery*: O protocolo iSCSI define um checksum de 32 bits, o CRC digest para seguimentos iSCSI, com a finalidade de

[12] *Checksum* ou soma de verificação é um código usado para verificar a integridade de dados transmitidos através de um canal com ruídos.

detecção de erros. Caso seja identificada a chegada de dados com o CRC inválido o seguimento é rejeitado e o *initiator* é notificado para o reenvio do mesmo;

- *Connection Recovery*: No caso de falha na conexão TCP, o comando SCSI fica pendente. Novas conexões TCP são abertas para substituir a que falhou, e a nova conexão continua a partir do ponto que ocorreu a falha.

Para a maioria das aplicações que executam requisições de E/S de dados utilizando o protocolo iSCSI o mecanismo *Session Recovery* é suficiente. Em enlaces de longas distâncias a ocorrência de problemas na comunicação é mais comum. Para estes casos um nível mais complexo de recuperação torna-se adequado como *Digest Failure Recovery* ou *Connection Recovery*.

Em relação a segurança, como o protocolo iSCSI utiliza protocolos desenvolvidos para a pilha TCP/IP, o iSCSI pode se beneficiar da tecnologia IPSec, proporcionando um nível de segurança maior na sua comunicação.

7 ANÁLISE COMPARATIVA DAS REDES DE ARMAZENAMENTO

Tendo em consideração as características das redes de armazenamento apresentadas, bem como suas principais diferenças, este tópico apresentará uma análise comparativa com o objetivo de fornecer o conhecimento necessário para a escolha da melhor arquitetura de rede de armazenamento, conforme a solução a ser implementada para virtualização de servidores.

A figura 34 ilustra as topologias de redes de armazenamento apresentadas, DAS, NAS, SAN e SAN IP, com foco nos protocolos de transporte e todo o processo de encapsulamento, demostrando todas as camadas do processo, do sistema operacional do servidor até o acesso aos dados no dispositivo de armazenamento.

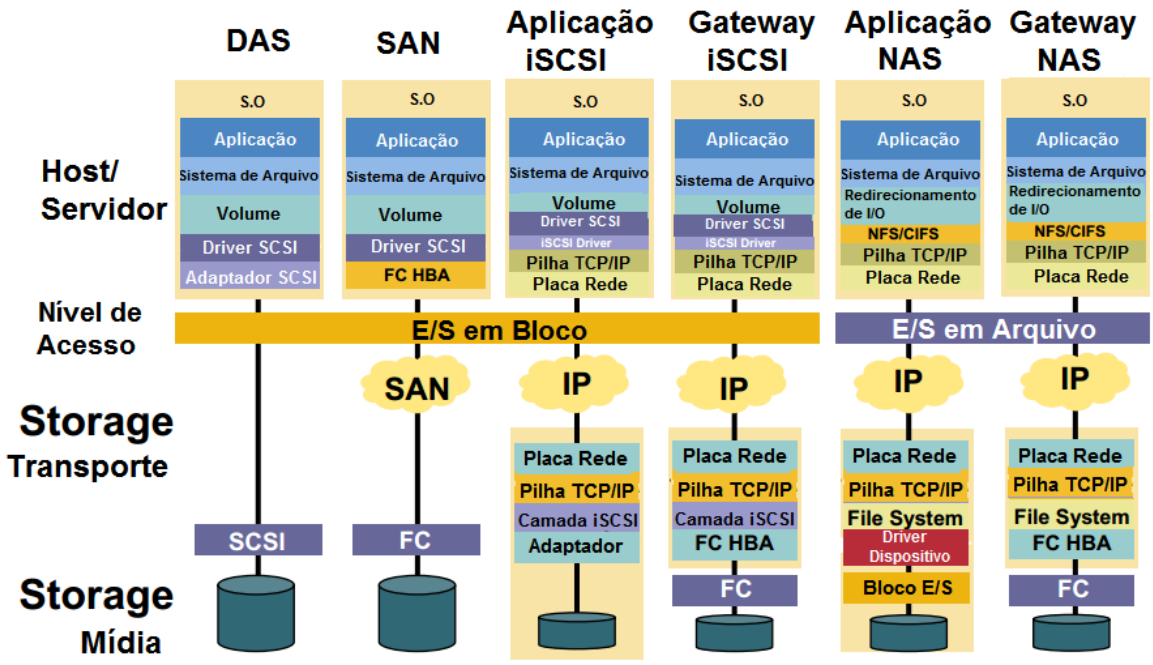

Figura 34 – Comparação das topologias de Redes de Armazenamento [DUARTE, OTTO, 2008]

Os principais itens que devem ser observados são: o meio físico e o protocolo de transporte. Basicamente são eles que definem todos os parâmetros de comparação.

A arquitetura DAS é a mais simples e limitada e sua configuração original não atende às necessidades de alta disponibilidade para os servidores virtuais. Sendo possível sua utilização para este fim através de softwares adicionais, apresentados no capítulo 5, que são limitados a uma quantidade pequena de servidores virtuais. Por este motivo a arquitetura DAS não será discutida mais amplamente, ampliando esta discussão para as arquiteturas NAS, SAN e SAN IP e seus respetivos protocolos NFS, FC e iSCSI.

7.1 COMPARAÇÃO DOS PROTOCOLOS iSCSI, NFS e FC PARA VIRTUALIZAÇÃO

Este tópico apresenta uma comparação dos principais parâmetros dos protocolos, proporcionando uma visão geral que ajudará em uma escolha mais acertada ao definir a arquitetura de armazenamento para uma implementação de alta disponibilidade de servidores virtuais.

Como mencionado no capitulo 5, o NFS não é suportado por todos os softwares de virtualização do mercado, um exemplo é o HyperV da Microsoft. Para esta comparação foi utilizado como base o software vSphere da VMware.

- **Descrição**
  - iSCSI – O iSCSI possibilita que o servidor de virtualização em vez de acessar blocos de um disco local, realize operações e/s de dados através de uma rede de armazenamento, utilizando um protocolo de acesso em nível de bloco. No caso do iSCSI, blocos remotos são

acessados através do encapsulamento de comandos SCSI e dados em pacotes TCP/IP.

- NFS – O NFS possibilita que o servidor de virtualização tenha acesso à dispositivos de armazenamento em nível de arquivo através da rede TCP/IP. O dispositivo de armazenamento NFS disponibiliza seus sistemas de arquivo local para montagem nos servidores de virtualização.
- FC – O *Fibre Channel* (FC) é semelhante ao iSCSI, permitindo acesso aos dispositivos de armazenamento em nível de bloco. Novamente, as operações e/s de dados são realizadas através de uma rede local, usando um protocolo de acesso de bloco. Na FC blocos remotos são acessados através do encapsulamento de comandos SCSI e dados em quadros de FC.

- **Opções de Implementação**
  - iSCSI – Através de adaptador de rede com recursos de iSCSI ou usando software iniciador iSCSI.
  - NFS – Através de adaptador de rede padrão.
  - FC – Requer um adaptador, *Host Bus Adapter* (HBA). Normalmente dois adaptadores, para redundância de caminhos.
- **Considerações Sobre o Desempenho**
  - iSCSI – O iSCSI pode executar em redes de 1Gbps, com links agregados ou uma rede TCP/IP de 10 Gbps. Várias conexões podem ser multiplexadas em uma única sessão, estabelecida entre o iniciador iSCSI e o destino. O software vSphere da VMware oferece suporte a

quadros jumbo para tráfego iSCSI, o que pode melhorar o desempenho. Quadros Jumbo enviam cargas maiores que 1.500 bytes.

- NFS – O NFS pode executar em redes de 1Gbps, com links agregados ou uma rede TCP/IP de 10 Gbps, pode introduzir sobrecarga na CPU do servidor de virtualização. Suporta UDP, mas na implementação do VMware requer TCP. O software vSphere da VMware oferece suporte a quadros jumbo para tráfego NFS, o que pode melhorar o desempenho.
- FC – O FC pode ser executado em 1 Gbps, 2 Gbps, 4 Gbps, 8 Gb ou 16 Gbps. Este protocolo afeta menos o consumo de CPU do servidor de virtualização porque HBAs, necessários para FC, realizam a maioria do processamento como encapsulamento de dados SCSI em quadros de FC. Uma consideração para a versão 5 do software vShepe e que as HBAs de 16 Gbps devem ser configuradas para executar a 8 Gbps no recurso “Buffer de créditos acelerador de soluções de problemas” para garantir uma rede sem perdas.

- **Balanceamento de Carga**
  - iSCSI – A arquitetura de armazenamento conectável, *Pluggable Storage Architecture* (PSA) da VMware, oferece um rodizio Round Robin (RR) através da política de seleção de caminho, *Path Selection Policy* (PSP), que distribui a carga entre vários caminhos para um destino iSCSI. Melhor distribuição de carga com PSP_RR é alcançada quando vários armazenamentos de dados são acessados simultaneamente.

- NFS – Não há nenhum balanceamento de carga na atual implementação do NFS porque trabalha apenas com uma única sessão. A largura de banda agregada pode ser configurada através da criação de vários caminhos para dispositivos de armazenamento NAS, acessando alguns armazenamentos de dados através de um caminho e outros armazenamentos de dados através de outro caminho.
- FC – A arquitetura de armazenamento conectável, *Pluggable Storage Architecture* (PSA) da VMware, oferece um rodízio Round Robin (RR) através da política de seleção de caminho, *Path Selection Policy* (PSP), que distribui a carga entre vários caminhos para um destino FC. Melhor distribuição de carga com PSP_RR é alcançada quando vários armazenamentos de dados são acessados simultaneamente.

- **Recuperação de Falha**
  - iSCSI – A arquitetura de armazenamento conectável, *Pluggable Storage Architecture* (PSA) da VMware, implementa *failover* através do *Storage Array Type Plugin* (SATP) para todos as matrizes iSCSI com suporte. O método preferido para fazer isso é através da implementação do software iSCSI, mas pode ser conseguido adicionando múltiplos destinos em diferentes sub-redes mapeados pelo iniciador iSCSI.
  - NFS – Através do agrupamento de adaptador de rede, que pode ser configurado de modo que se uma interface falhar, outra pode assumir seu lugar. No entanto, isso depende de uma falha de rede e pode não ser capaz de manipular condições de erro que ocorrem do lado do dispositivo de armazenamento NFS.

- FC – A arquitetura de armazenamento conectável, *Pluggable Storage Architecture* (PSA) da VMware, implementa *failover* através da *Storage Array Type Plugin* (SATP) para todas as matrizes suportadas pela arquitetura FC.

- **Verificação de Erros**
  - iSCSI – O iSCSI usa TCP, que reenvia pacotes descartados.
  - NFS – O NFS usa TCP, que reenvia pacotes descartados.
  - FC – O FC é implementado como uma rede sem perdas. Isto é conseguido através da limitação de *throughput* em tempos de congestionamento.
- **Segurança**
  - iSCSI – O iSCSI implementa o *Challenge Handshake Authentication Protocol* (CHAP) para garantir que os iniciadores e destinos confiem uns nos outros. VLANs ou redes privadas são altamente recomendadas para isolar o tráfego iSCSI de outros tipos de tráfego.
  - NFS – VLANs ou redes privadas são altamente recomendadas para isolar o tráfego NFS de outros tipos de tráfego.
  - FC – Alguns switches FC suportam os conceitos de Virtual SAN (VSAN), que isola partes da infraestrutura de armazenamento. VSANs são conceitualmente semelhantes a VLANs. Zoneamento entre *hosts* e destinos FC também oferece um grau de isolamento.
- **Tamanho e Numero Máximo de Dispositivo**
  - iSCSI – O iSCSI suporta um número máximo de 256 dispositivos com tamanho máximo de 64 terabytes.

- NFS – O NFS suporta por padrão 8 dispositivos, podendo suportar 256. Para o tamanho máximo do dispositivo deve-se consultar o fornecedor do dispositivo de armazenamento NAS para informações. Tamanho teórico é muito maior do que 64 terabytes.
- FC – O FC suporta um número máximo de 256 dispositivos com tamanho máximo de 64 terabytes.

- **Protocolo direto à Maquina Virtual**
  - iSCSI – O iSCSI da suporte ao acesso direto a máquina virtual por meio do iniciador iSCSI.
  - NFS – O NFS da suporte ao acesso direto a máquina virtual através do cliente NFS.
  - FC – O FC não da suporte ao acesso direto a máquina virtual. Porem a VMware permite que os dispositivos FC sejam mapeados diretamente para a máquina virtual com o NPIV. Isso ainda exige prévio mapeamento de RDM para a máquina virtual e o hardware deve oferecer suporte a NPIV (switch FC, HBA).
- **Configurações Especificas do VMware**
  - *Raw Device Mapping* (RDM) é uma opção no servidor de virtualização do VMware que permite que uma máquina virtual se conecte diretamente a uma área de armazenamento de dados através da arquitetura SAN. Suportado pelos protocolos iSCSI e FC, o NFS não dá suporte ao RDM.
  - VMware *VMotion* move em tempo real máquinas virtuais em execução de um servidor de virtualização para outro e, ao mesmo tempo,

mantém a disponibilidade contínua dos serviços. Os protocolos iSCSI, NFS e FC dão suporte a esta configuração.

- VMware *Distributed Resource Scheduler* (DRS) realiza automaticamente a migração de máquinas virtuais entre os servidores de virtualização do *pool* de recursos de forma a redistribuir com inteligência os recursos disponíveis. Os protocolos iSCSI, NFS e FC dão suporte a esta configuração.

- **Facilidade de Configuração**
  - iSCSI – Médio, configurar o iniciador iSCSI exige conhecimento do FQDN ou endereço IP de destino, além de algumas configurações para mapeamento do iniciador e apresentação da unidade logica, *Logical Unit Number* (LUN) do lado do dispositivo de armazenamento. Depois que o destino foi descoberto através de uma verificação da SAN, LUNs estão disponíveis para armazenamentos de dados ou RDMs.
  - NFS – Fácil, requer apenas o IP ou o FQDN do destino, mais o ponto de montagem. Armazenamentos de dados aparecem imediatamente após o acesso ser concedido do lado do dispositivo de armazenamento NFS.
  - FC – Difícil, envolve zoneamento em nível de switch FC e máscara de LUN no nível do dispositivo de armazenamento, após o zoneamento é completo. É mais complexo para configurar do que o armazenamento IP. Depois que o destino foi descoberto através de uma verificação da SAN, as LUNs estão disponíveis para armazenamentos de dados ou RDMs.

- **Vantagens**
  - iSCSI – Nenhum hardware adicional é necessário. Pode usar componentes de hardware de rede existentes e *driver* iSCSI, assim tem baixo custo para implementar. Bastante maduro, administradores com competências de rede devem ser capazes de implementar. Soluções de problemas podem ser realizadas com ferramentas de rede genéricas como o Wireshark.
  - NFS – Nenhum hardware adicional é necessário. Pode usar componentes de hardware existente na rede, por isso tem baixo custo de implementação. Protocolo conhecido, também é muito maduro. Administradores com competência de rede devem ser capazes de implementar. Soluções de problemas podem ser realizadas com ferramentas de rede genéricas como o Wireshark.
  - FC – Protocolo bem conhecido e bem compreendido. Muito maduro e confiável. Encontrado na maioria dos ambientes de missão critica.
- **Desvantagens**
  - iSCSI – Incapacidade para rotear com ligação iSCSI implementada. Problema de segurança possível, porque não há nenhuma criptografia interna, então o cuidado deve ser tomado para isolar o tráfego, por exemplo, VLANs. Software iSCSI pode causar sobrecarga adicional na CPU do servidor de virtualização. TCP pode apresentar latência para iSCSI.
  - NFS – Há apenas uma única sessão por conexão. Mesmas preocupações de segurança que o iSCSI, porque tudo é transferido

não criptografado, então deve ter cuidado para isolar o tráfego, por exemplo, VLANs. Pode causar sobrecarga adicional na CPU do servidor de virtualização. TCP pode apresentar latência para NFS. No VMware a implementação NFS ainda é a versão 3, que não tem as características de vários caminhos ou segurança de NFS v4 ou NFS v 4.1.

- FC – Requer adaptador HBA dedicado, switch FC e interface FC no dispositivo de armazenamento, que faz com que a implementação FC tenha um alto custo. Sobrecarga de gerenciamento adicional, por exemplo, mudar o zoneamento. Pode revelar-se mais difícil de solucionar do que os outros protocolos. No VMware ainda funciona somente a 8 Gbps, que é mais lento do que outras redes.

8 CONCLUSÃO

A arquitetura *Storage Area Network Fibre channel* (SAN FC) é hoje uma arquitetura consolidada no mercado e amplamente utilizada pelas grandes corporações quando se pensa em alta disponibilidade de servidores virtuais. Diversos fatores influenciam para esta escolha favorável à utilização da arquitetura SAN FC: quadro sem *overhead*[13] adicional ao seu enquadramento; possibilidade de utilização de sequenciamento de quadros e de classes de serviços; implementação de enlaces de alto desempenho; e ser a arquitetura de rede de armazenamento implementada nas primeiras versões dos principais softwares de virtualização.

De fato, a arquitetura SAN FC apresentou ao longo do tempo um amadurecimento que proporciona a confiança necessária para sua implementação em alta disponibilidade de servidores virtuais. Porém sua implantação pode não ser viável financeiramente dependendo do cenário.

O surgimento do protocolo iSCSI e sua padronização possibilitou uma nova arquitetura de rede de armazenamento, a *Storage Area Network Internet Protocol* (SAN IP). Por ser baseada na arquitetura IP depende dos protocolos das camadas superiores para prover a entrega das mensagens SCSI ao dispositivo de armazenamento, introduzindo assim um *overhead* pela pilha TCP/IP e enquadramento Ethernet. Porém, com o aumento de banda passante das redes ethernet para 1Gbps e 10Gbps, utilização de links agregados, VLANs e até mesmo o isolamento da rede para uso exclusivo do iSCSI, são soluções que

[13] *Overhead* é a carga excessiva de processamento ou armazenamento, seja de tempo de CPU, memória ou largura de banda.

fazem com que este *overhead* não atrapalhe seu desempenho. Por utilizar rede IP seu custo é atraente para muitas empresas e passou a ser uma opção de rede de armazenamento interessante para os softwares de virtualização.

A arquitetura *Network Area Storage* (NAS) através do protocolo NFS passou a ser uma opção de rede de armazenamento na versão 4 do vSphere, software de virtualização da VMware, empresa que possui a maior representação no mercado de virtualização. Porém, algumas limitações foram apresentadas e por ser uma opção relativamente nova para o ambiente de virtualização, onde vários softwares de virtualização não têm suporte, é preciso alcançar uma maior maturidade. Mesmo assim, pode ser uma opção para pequenas e médias empresas por ser uma arquitetura de baixo custo de implementação e manutenção, e com desempenho próximo ao alcançado pela arquitetura SAN IP.

Seguindo a necessidade de atender a um público de pequenas e médias empresas, algumas soluções de alta disponibilidade para Virtualização utilizando a arquitetura *Direct Area Storage* (DAS) foram desenvolvidas. Porém, estas soluções se mostram bem limitadas, atendendo a um número pequeno de servidores virtuais, podendo ser uma opção para determinados cenários com poucos servidores virtuais.

Com isso, pode-se concluir que não há uma solução de arquitetura de rede de armazenamento padrão para atender a alta disponibilidade de servidores virtuais. Tudo vai depender do cenário que se deseja implementar esta solução e quais premissas são importantes para este cenário. Este trabalho apresentou uma visão das características das redes de armazenamento que podem ser adotadas, para que seja realizada a melhor escolha para os cenários que se apresentem.

REFERENCIAS BIBLIOGRAFICAS

VERAS, MANUEL – Virtualização: Componente Central do Datacenter, Rio de Janeiro, Editora Brasport, 2011.

SOMASUNDARAM, G., SHRIVASTAVA, ALOK, EMC Education Services – Armazenamento e Gerenciamento de Informações: Como armazenar, gerenciar e proteger informações digitais, Tradução: Acauan Pereira Fernandes; Revisão Técnica: EMC Brasil – Porto Alegre; Editora: Bookman, 2011.

DESAI, ANIL – Definition Virtual Machine, 2011. Disponível em http://searchservervirtualization.techtarget.com/definition/virtual-machine. Ultimo acesso em 6 de junho de 2012.

MANFRIN, ALEXANDER – Cenários para Virtualização, 2010. Disponível em http://www.vmworld.com.br/br/index.php/component/content/article/50-virtualizacao/81-cenarios-para-virtualizacao.html. Ultimo acesso em 7 de junho de 2012.

MULLINS, ROBERT – Virtualização é a prioridade top dos CIOs em 2012, diz IDC. Disponível emhttp://informationweek.itweb.com.br/6517/virtualizacao-e-a-prioridade-top-dos-cios-em-2012/. Ultimo acesso em 7 de junho de 2012.

GERTNER, INC - Magic Quadrant for x86 Server Virtualization Infrastructure. Disponível em http://www.gartner.com/technology/media-products/reprints/vmware/213635.html. Ultimo acesso em 9 de junho de 2012.

VMWARE – Conceitos básicos da virtualização, 2012. Disponível em http://www.vmware.com/br/virtualization/virtualization-basics/what-is-virtualization.html. Último acesso em 26 de setembro de 2012.

VMWARE1 – *White Paper – Understanding Full Virtualization, Paravirtualization, and Hardware Assist*, 2007. Disponível em http://www.vmware.com/files/pdf/VMware_paravirtualization.pdf. Ultimo acesso em 26 de setembro de 2012.

CARISSIMI, ALEXANDRE – Virtualização: Princípios Básicos e Aplicações, 2009. Disponível em http://www.lbd.dcc.ufmg.br/colecoes/erad/2009/004.pdf. Ultimo acesso em 26 de setembro de 2012.

MICROSOFT – Alta Disponibilidade (Analysis Services – Dados Multidimensionais), 2008. Disponível em http://msdn.microsoft.com/pt-br/library/bb500217(v=sql.105).aspx. Ultimo acesso em 14 de setembro de 2012.

INFO-TECH – *The Bottom Line on Virtualization*, 2008. Disponível em http://www.vmware.com/files/pdf/InfoTech_SMB_DR_Whitepaper-Sept_1808.pdf. Ultimo acesso em 16 de setembro de 2012.

VMWARE2 – Vmware DRS – Balanceamento e Alocação Dinâmicos de Recursos para Máquinas Virtuais, 2007. Disponível em http://www.vmware.com/files/br/pdf/products/07Q3_VM_DRS_DS_BR_A4.pdf. Ultimo acesso em 16 de setembro de 2012.

VMWARE3 – Vmware *High Availability*, 2007. Disponível em http://www.vmware.com/pdf/ha_datasheet.pdf. Ultimo acesso em 16 de setembro de 2012.

VMWARE4 – *White Paper - NIC Teaming*, 2003. Disponível em http://www.vmware.com/pdf/esx2_NIC_Teaming.pdf. Ultimo acesso em 26 de setembro de 2012.

VMWARE5 – *Comparison of Storage Protocol Performance*, 2008. Disponivel em http://www.vmware.com/files/pdf/storage_protocol_perf.pdf. Ultimo acesso em 21 de novembro de 2012.

ARMBRUST, VICTOR – Oracle x RAID – Melhores Práticas, 2012. Disponível em http://victor-dba.blogspot.com.br/2012/03/oracle-x-raid-melhores-praticas.html. Ultimo acesso em 5 de novembro de 2012.

VM6 SOFTWARE – *Microsoft Hyper V Virtualization Product*, 2011. Disponível em http://www.vm6software.com/microsoft-hyper-v-virtualization-product. Ultimo acesso em 5 de novembro de 2012.

NUTANIX – 10 *Design considerations while bulding a storage system for virtualization*, 2011. Disponível em http://www.nutanix.com/blog/2012/06/21/10-design-considerations-while-building-a-storage-system-for-virtualization/. Ultimo acesso em 5 de novembro de 2012.

MANFRIN, ALEXANDER – vSphere 5 e as novas funcionalidades: VSA – vSphere *Storage Appliance*, 2011. Disponível em http://www.vmworld.com.br/br/index.php/component/content/article/41-virtualizacao/114-vsphere-5-e-as-novas-funcionalidades-vsa-vsphere-storage-appliance.html. Ultimo acesso em 7 de novembro de 2012.

JAVVIN – FC & FCP: *Fibre Channel and Fibre channel Protocol*, 2012. Disponível em http://www.javvin.com/protocolFCP.html. Ultimo acesso em 6 de dezembro de 2012.

NETO, ANTÔNIO – Um Estudo do Desempenho dos Protocolos iSCSI e *Fibre Channel*, 2004. Disponível em

http://www.bibliotecadigital.unicamp.br/document/?code=vtls000375890. Ultimo acesso 11 de dezembro de 2012.

VMWARE6 – *Storage Protocol Comparison White Paper*, 2012. Disponível em http://www.vmware.com/files/pdf/techpaper/Storage_Protocol_Comparison.pdf Ultimo acesso em 19 de fevereiro de 2013.

DUARTE, Otto – *Fibre Channel* e *Storage Area Network*, 2008. Disponível em http://www.gta.ufrj.br/grad/08_1/san/index.html. Ultimo acesso em 19 de fevereiro de 2013.