INSTITUTO SUPERIOR DE ENGENHARIA DE LISBOA

ÁREA DEPARTAMENTAL DE ENGENHARIA DE ELETRÓNICA E TELECOMUNICAÇÕES E DE COMPUTADORES

Estudo e Planeamento de uma Infraestrutura Computacional

HENRIQUE MANUEL ALJUSTREL LOPES
(Licenciado)

Projeto de mestrado para obtenção do grau de Mestre em Engenharia de Eletrónica e Telecomunicações

Orientadores:
Professor Adjunto Paulo Alexandre Medeiros de Araújo
Professor Adjunto Porfírio Pena Filipe

Júri:
Presidente: Professor Adjunto João Miguel Duarte Ascenso
Vogais: Professor Adjunto João Beleza Teixeira Seixas de Sousa
Orientador: Professor Adjunto Paulo Alexandre Medeiros de Araújo
Co-Orientador: Professor Adjunto Porfírio Pena Filipe

Outubro de 2012

Agradecimentos

À minha esposa Sónia o meu eterno obrigado pelo apoio, compreensão e ajuda ao longo destes anos. A ela e ao meu filho recém-nascido Duarte, dedico esta dissertação.

A toda a minha família um forte agradecimento, pois sem ela não teria chegado onde cheguei.

Um agradecimento aos Professores Porfírio Filipe e Paulo Araújo pela orientação, disponibilidade e apoio na realização desta dissertação. À unidade de informática do caso em estudo pelo apoio na análise e enquadramento do ambiente de estudo.

À VMWare Portugal, nomeadamente, ao Celso Capão pelas ideias discutidas e apoio em documentação VMWare assim como conhecimentos técnicos.

À EMC Portugal e a todos os que ajudaram na disponibilização de documentos de testes práticos realizados em laboratório quer em dúvidas ao longo desta dissertação.

Agradeço ao ISEL – Instituto Superior de Engenharia de Lisboa e em especial à ADEETC – Área Departamental de Engenharia de Eletrónica e Telecomunicações e de Computadores.

Aos meus colegas do ISEL, pelo apoio e companheirismo demonstrado ao longo de todo o percurso académico e pelas longas horas passadas a discutir ideias que sem dúvida me ajudaram a chegar onde cheguei.

A todos que de forma direta ou indireta contribuíram para a concretização desta dissertação e me ajudaram nos momentos mais difíceis ao longo de todo o meu percurso académico, o meu bem-haja.

Resumo

As tecnologias de informação representam um pilar fundamental nas organizações como sustento do negócio através de infraestruturas dedicadas sendo que com o evoluir do crescimento no centro de dados surgem desafios relativamente a escalabilidade, tolerância à falha, desempenho, alocação de recursos, segurança nos acessos, reposição de grandes quantidades de informação e eficiência energética. Com a adoção de tecnologias baseadas em *cloud computing* aplica-se um modelo de recursos partilhados de modo a consolidar a infraestrutura e endereçar os desafios anteriormente descritos.

As tecnologias de virtualização têm como objetivo reduzir a infraestrutura levantando novas considerações ao nível das redes locais e de dados, segurança, *backup* e reposição da informação devido á dinâmica de um ambiente virtualizado. Em centros de dados esta abordagem pode representar um nível de consolidação elevado, permitindo reduzir servidores físicos, portas de rede, cablagem, armazenamento, espaço, energia e custo, assegurando os níveis de desempenho.

Este trabalho permite definir uma estratégia de consolidação do centro de dados em estudo que permita a tolerância a falhas, provisionamento de novos serviços com tempo reduzido, escalabilidade para mais serviços, segurança nas redes *Delimitarized Zone* (DMZ), e *backup* e reposição de dados com impacto reduzido nos recursos, permitindo altos débitos e rácios de consolidação do armazenamento. A arquitetura proposta visa implementar a estratégia com tecnologias otimizadas para o *cloud computing*.

Foi realizado um estudo tendo como base a análise de um centro de dados através da aplicação VMWare Capacity Planner que permitiu a análise do ambiente por um período de 8 meses com registo de métricas de acessos, utilizadas para dimensionar a arquitetura proposta. Na implementação da abordagem em *cloud* valida-se a redução de 85% de infraestrutura de servidores, a latência de comunicação, taxas de transferência de dados, latências de serviços, impacto de protocolos na transferência de dados, *overhead* da virtualização, migração de serviços na infraestrutura física, tempos de *backup* e restauro de informação e a segurança na DMZ.

Palavras-chave: *cloud computing*, virtualização, consolidação de sistemas, alta disponibilidade, centro de dados

Abstract

Information technologies represent a major pillar on the organization business with dedicated infrastructure, although with the growing demands on the datacenter, challenges rise on scalability, fault tolerance, performance, resource allocation, access security, restore big amounts of data and power efficiency.
With the adoption of cloud computing technologies a new model based on shared resources consolidates the datacenter infrastructure.

The virtualization technologies have as the main objective to reduce physical infrastructure with some considerations on local networks, storage networks, security, backups and restore of data due to the nature of the dynamic environment. In the datacenter this can represent a huge degree of consolidation, taking organizations to reduce physical servers, network ports, cabling, storage, space, power consumption and cost ensuring performance.

This work allows to define a strategy of consolidation on the studied datacenter, that can tolerate fault conditions, provisioning of new services more quickly, scalability for more applications, security in the Delimitarized Zone (DMZ), and backup and restore of the data with less impact on the physical resources, allowing high throughput rates and high consolidation of storage equipments. The proposed architecture intends to implement the strategy with optimized technologies for the cloud computing paradigm.

A study was made on a case study having as its basis an analysis of a datacenter with the application VMWare Capacity Planner that allowed the monitoring of the environment for 8 months with inventory and performance data collected for modeling to size a new proposed infrastructure.
For the cloud implementation approach it is validated the consolidation of the server infrastructure in 85%, communication latency, data throughput rates , service latency, impact of storage protocols, virtualization overhead, migration of services within the physical infrastructure, backup and restore windows and security on the DMZ network.

Keywords: *cloud computing*, virtualization, consolidation, high availability, datacenter

Índice

Índice de Figuras

Índice de Tabelas

Acrónimos e Abreviaturas

CBT – Change Block Tracking

CNA – Converged NetWork Adapter

DAS – Direct-Attached Storage

DMZ – Demilitarized Zone

DR – Disaster Recovery

FC – Fibre Channel

HBA – Host Bus Adapter

I/O – Input/Output

IaaS – Infrastructure-as-a-Service

ICMP – Internet Control Message Protocol

IP – Internet Protocol

(i)SCSI – (Internet) Small Computer System Interface

LAN – Local Area Network

LTO – Linear Tape Open

LUN – Logical Unit Number

MAC – Media Access Control

NAS – Network Attached Storage

NIC – NetWork Interface Card

OLTP – OnLine Transaction Processing

P2V – Physical-to-Virtual

PaaS – Platform-as-a-Service

PCIe – Peripheral Component Interconnect express

PPS – Pacotes por Segundo

PVLAN – Private Virtual Local Area NetWork

RAID – Redundant Array Independent Disks

RARP – Reverse Address Resolution Protocol

RPO – Recover Point Objective

RTO – Recover Time Objective

SaaS – Software-as-a-Service

SAN – Storage Area NetWork

SAS – Serial Attached SCSI

SATA – Serial Advanced Technology Attachment

SO – Sistemas Operativos

TCP – Transmission Control Protocol

TI – Tecnologias de Informação

ToE – TCP/IP Offload Engine

VACL – Virtual Access Lists

VLAN – Virtual Local Area NetWork

VPN – Virtual Private Network

VTL – Virtual Tape Library

WAN – Wide Area NetWork

WWPN – World Wide Port Name

Convenções tipográficas

Ao longo do texto desta dissertação surgem termos em inglês em situações em que a sua tradução para português não reflete na realidade todo o seu significado, ou por serem termos que já são universalmente aplicados. Tal situação acontece devido à documentação existente sobre este tema ser, na sua maioria, publicada em língua inglesa e, sempre que possível, são utilizadas traduções que se considerem apropriadas. Estes termos são apresentados em caracteres itálicos.

Para evitar a repetição de expressões técnicas longas, que possam tornar a leitura desta dissertação repetitiva, são utilizados acrónimos ao longo do texto com a respetiva tradução numa área distinta no início neste documento. Todas as referências bibliográficas utilizadas ao longo da dissertação são evocadas entre parêntesis retos e são apresentadas no final desta dissertação.

De modo a se destacar determinada temática ao longo da dissertação recorre-se ao *bold*.

1. Introdução

A maior parte das organizações suportam o seu negócio nas Tecnologias de Informação (TI) e dependem cada vez mais da disponibilidade e fiabilidade destas para se manterem sustentáveis. Com o crescimento do negócio, aumentam as exigências de uma infraestrutura computacional dinâmica com capacidade de tolerância a falhas, modular e escalável para acompanhamento dos requisitos de desempenho e capacidade. A evolução tecnológica dos últimos anos levou as organizações a adotarem cada vez mais modelos de *cloud computing* [1] de modo a reduzir custos capitais e operacionais assegurando a alta disponibilidade das suas aplicações críticas consolidando desta forma todo o *hardware* e aumentando a eficiência dos recursos disponíveis.
Com o crescimento das organizações e consequentemente das suas aplicações de TI surgem necessidades de consolidar as aplicações e servidores, redes, armazenamento e *backups*. No entanto, para garantir que não existem erros de dimensionamento da infraestrutura consolidada, existem ferramentas que permitem recolher dados estatísticos de modo a obtermos informação concreta para aplicar numa nova infraestrutura a dimensionar.

Neste capítulo serão identificados os problemas da abordagem convencional no centro de dados, bem como a estratégia adotada para os resolver através de tecnologias associadas ao *cloud computing*.

1.1. Problema

Grande parte das empresas possuem infraestruturas dedicadas para cada aplicação, nomeadamente servidores, sendo que existem produtos para *clustering* de servidores para assegurarem a alta disponibilidade, no entanto, com algumas limitações como a complexidade de implementação, o suporte para Sistemas Operativos (SO) e o custo associado. Esta abordagem implica que no caso de avaria física de um servidor, a aplicação terá quebra de serviço associado e consequentemente impacto financeiro numa organização, sendo que a recuperação da mesma está limitada quanto à disponibilidade de um novo servidor, configuração da aplicação e recuperação do último *backup* disponível para entrar novamente em produção. Este processo descrito poderá demorar entre dias a semanas, comprometendo a organização numa perspetiva de custos e credibilidade no mercado. Outra problemática associada é a própria manutenção de infraestrutura que poderá implicar uma extensa paragem de serviço sem qualquer alternativa de

mobilidade aplicacional devendo a intervenção ser feita de forma calculada e prevista numa janela temporal para não prejudicar o negócio em período laboral. Com o crescimento exponencial de informação nas empresas [2], a utilização de *Direct-Attached Storage* (DAS) torna praticamente insustentável a gestão da informação devido a limitações físicas dos servidores, nomeadamente, endereçamento de discos físicos. Além do mais o aumento de acessos *Input/Output* (I/O) às aplicações misturado com o restante tráfego IP não é recomendado por fabricantes [3] devido a problemas de desempenho pois muitas aplicações, como ambientes *OnLine Transaction Processing* (OLTP) são sensíveis à latência, prejudicando a funcionalidade da mesma, assim como a configuração do armazenamento em servidores físicos é limitada quanto a técnicas de proteção e de alta disponibilidade.

Existem ainda requisitos em determinadas organizações quanto a ambientes de *Disaster Recovery* (DR) para garantir a operação em caso de falha total no centro de dados, sendo um exemplo disso a lei norte-americana em que toda a empresa cotada em bolsa é obrigada a ter um plano de contingência para garantir a continuidade das operações num sítio alternativo. Os *backups* dos dados tipicamente têm uma janela de tempo associada sendo maioritariamente feitas em períodos de menor atividade, já que despoletam uma elevada leitura de dados em disco. Com o crescimento da informação, o volume de *backups* é ainda mais acentuado devido à quantidade de informação a reter levando á extensão da janela de *backup* até períodos de produção, interferindo no desempenho das aplicações já que os recursos vão ser solicitados quer nos acessos tradicionais laborais de produção quer no próprio *backup*.

1.2. Motivação

O modelo implementado atualmente no exemplo prático com infraestrutura dedicada, implica um conjunto de desafios tais como:

- Gestão morosa e ineficaz,
- Maior consumo de recursos físicos,
- Menor eficiência energética, arrefecimento e espaço no centro de dados,
- Complexidade das arquiteturas,
- Fraca implementação de mecanismos de alta disponibilidade,
- Incapacidade de alocação de mais recursos para uma aplicação de forma dinâmica,
- Limitação quanto a janelas de *backup*,
- Custo geral elevado de toda a solução implementada.

Associado a todas estas problemáticas, o fabricante VMWare pioneiro em tecnologias de virtualização, elaborou um estudo [4] no qual verificou que a maior parte das organizações utilizam uma fatia média muito reduzida nos seus servidores, entre os 5 a 10%, ou seja, o investimento feito num servidor traduz-se na maior parte das vezes numa utilização reduzida face ao investimento já feito. Este estudo conclui ainda que a integração de um novo serviço ou servidor e respetivas aplicações na rede variam entre as 960 a 240 horas dependendo do ramo de negócio.

1.3. Estratégia

De modo a endereçar todas estas problemáticas propõe-se inicialmente a aplicação de um modelo baseado em c*loud computing* sustentado por uma infraestrutura tecnológica capaz de endereçar todos os pontos identificados. Um projeto de c*loud computing* pode conhecer duas realidades:

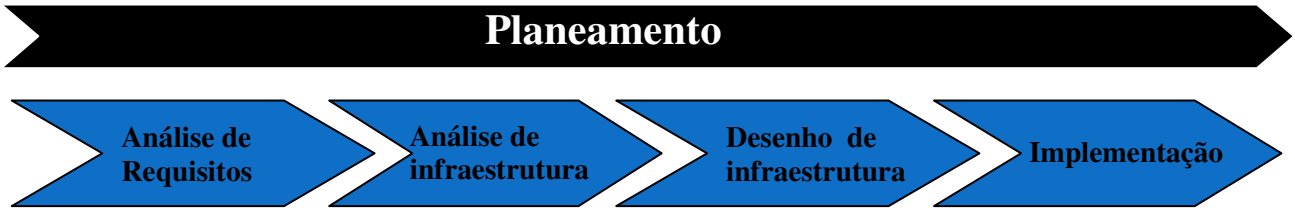
a) Renovação, no qual, existem tecnologias implementadas numa abordagem física,

b) Projecto de raiz como se se tratasse de uma nova empresa onde os dados para planeamento do centro de dados são baseados em pressupostos ou previsões.

O âmbito deste projeto, nomeadamente o ambiente específico, será focado para uma renovação tecnológica no qual se enquadra a componente de análise para colecionar dados estatísticos reais e dimensionar uma solução em todas as vertentes à medida do atual e respetiva margem de crescimento numa janela temporal. A aplicação deste modelo pode disponibilizar três tipos de serviços [5] como o *Software-as-a-Service* (SaaS), *Platform-as-a-Service* (PaaS) ou *Infrastructure-as-a-Service* (IaaS) sendo a sua aplicação baseada em quatro modelos distintos: Privado, Público, Comunidade ou Híbrido. O caminho optado a seguir prende-se com a realidade da infraestrutura redimensionada, ou seja, através do modelo IaaS com aplicação de *cloud* privada através de um conjunto de tecnologias identificadas neste documento.

A virtualização de servidores é fundamental para se atingir uma consolidação dos sistemas físicos e aplicações numa *cloud* privada, assim como proporcionar serviços de alta disponibilidade através da passagem transparente de servidores virtuais entre servidores físicos sempre com acesso aos dados críticos. As tecnologias de armazenamento otimizadas para escritas e leituras,

permitem escalar até TBytes sendo imprescindíveis para a alta disponibilidade das aplicações e gerem toda a componente de acesso a discos. Estas tecnologias possibilitam a replicação de dados, local ou remota, para soluções de proteção de dados para rápidas recuperações e DR, respectivamente. As redes *Storage Area Networks* (SAN) ou *Network Attached Storage* (NAS) proporcionam alta disponibilidade e performance às aplicações e são responsáveis pela ligação de servidores aos dispositivos de armazenamento. Com a consolidação de armazenamento e servidores surgem necessidades nas redes *Local Area Networks* (LAN) que têm que ser endereçadas para suportarem o tráfego das aplicações que irão estar a correr no mesmo servidor físico. De igual forma a solução de *backups* terá que ser dimensionada para uma janela de tempo disponível, com o mínimo impacto possível nas redes e aplicações, sendo capaz de escalar conforme o crescimento de dados. A segurança transversal a todos estes pontos traduz-se em técnicas específicas de cada área que permitem restringir acessos e proteger um ambiente. O desenvolvimento deste projeto passa por uma série de considerações de análise, desenho e implementação que percorrem quatro fases distintas [6]:



Análise de Requisitos	Análise Infraestrutura	Desenho Infraestrutura	Implementação
Reunião TI	Virtualização	Virtualização	Validação
Requisitos operação	Armazenamento	Armazenamento	Técnica
Requisitos gestão	*Backups*	*Backups*	Gestão
Requisitos manutenção	Redes	Redes	Migração
Outros Requisitos	Segurança	Segurança	Formação

Figura 1 - Planeamento da infraestrutura computacional

A aproximação ao projeto será analisar os requisitos iniciais como conhecer a organização, modo de operação, gestão, manutenção ou outros requisitos essenciais para o dimensionamento. Nesta fase é igualmente importante pré-dimensionar os requisitos de crescimento que muitas vezes pode

ser complexo de prever, como por exemplo, uma empresa comprar outra e absorver todo o negócio a ser suportado numa infraestrutura.

De seguida a análise de infraestrutura, como o levantamento ou utilização de *software* específico tal como o VMWare *Capacity Planner* que permite obter dados reais de uma infraestrutura para posicionar uma solução de virtualização de servidores e respetivos rácios de consolidação. Por fim o desenho da nova solução tendo em conta os passos anteriores e migração e implementação da nova solução.

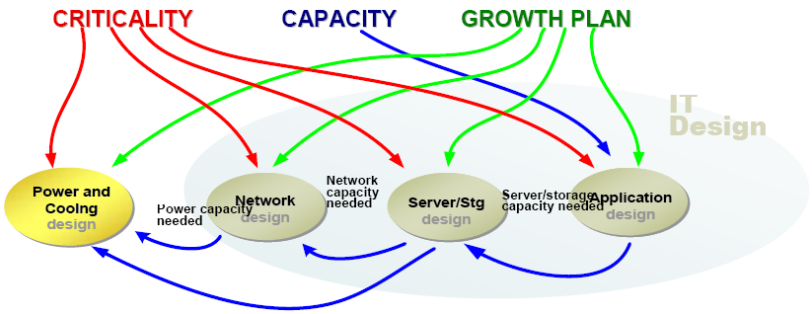


Figura 2 - Relação no dimensionamento [6]

Todas estas componentes relacionam-se para posicionar uma infraestrutura desenhada à medida com as tecnologias alinhadas numa perspetiva de criticidade, capacidade e respetivo crescimento suportado tal como descrito na figura 2.

1.4. Estruturação da Dissertação

Esta dissertação inicia-se com a análise do Estado da Arte, sendo que neste capítulo 2 é efetuado o enquadramento do problema e são analisadas as tecnologias emergentes mais relevantes para a resolução dos problemas identificados. No capítulo 3 é feito o levantamento da arquitetura atual assim como os resultados da análise de consumo de recursos da infraestrutura do caso em estudo. O capítulo 4 descreve a proposta de infraestrutura preparada em *cloud* baseado nos consumos analisados e por fim no capítulo 5 é feita a validação da transição do modelo físico para o modelo de recursos partilhados com os consumos verificados e respectivas margens de proteção e crescimento.

2. Estado da Arte

Neste capítulo são reunidos os modelos e componentes relevantes de inovação na área de infraestrutura dinâmica, escalável e robusta para o *cloud computing*, sendo que, começa-se inicialmente por descrever os modelos existentes e respetivas tecnologias para endereçar cada ponto, ou seja, virtualização, redes locais e de dados, armazenamento partilhado, *backups* e segurança, identificando-se os desenvolvimentos tecnológicos para cada área.

2.1. Enquadramento do Problema

Relativamente ao contexto desta dissertação, os modelos e tecnologias em estudo serão aplicados à infraestrutura do caso em estudo, sendo composto por um conjunto de tecnologias sem aplicação do modelo de *cloud computing* e com pouca penetração de virtualização nos seus sistemas para otimizar e reagir de forma rápida a falhas.
Para se tirar partido destas tecnologias é necessária a análise de infraestrutura para determinar a quantidade de acessos e taxas de utilização de modo a dimensionar a quantidade de recursos partilhados para a *cloud*. As tecnologias apresentadas e posicionadas terão um alinhamento quer tecnológico, quer financeiro, de modo a justificar a aplicação do modelo como benéfica para o caso em estudo.

2.1.1. Cloud Computing

O *cloud computing* é um paradigma nas TI que se traduz numa abordagem de recursos partilhados e consumos baseados na utilização real de uma infraestrutura [5], ou seja, os recursos físicos tais como, rede, armazenamento, computação, aplicações e serviços possuem uma camada de abstração de modo a ser projetado um conjunto de recursos partilhados e disponíveis numa infraestrutura para rentabilizar ao máximo os investimentos nas TI.

O modelo convencional implica que a disponibilização de uma aplicação, por exemplo, para o negócio seja feita através da encomenda de um novo servidor, configurações e parametrizações adicionais, configuração de rede e respetiva disponibilidade de portas, armazenamento, entre

outras sendo que o tempo medido desde a fase de decisão até á fase de implementação pode durar dias a semanas sendo insustentável para muitas organizações.

O modelo proposto baseado em *cloud computing* permite a disponibilização de recursos *on demand,* proporcionando desta forma eficiência e rapidez de resposta face a oportunidades de negócio, assim como aumenta a competitividade das empresas.

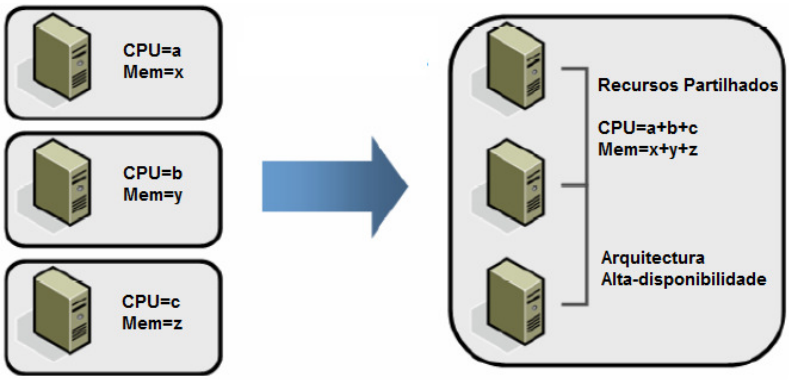


Figura 3 - Passagem do modelo convencional para o modelo de *cloud computing*

Estudos efetuados a nível mundial [1] demonstram que as vendas geradas pelo negócio do *cloud computing* nos principais fornecedores de TI entre 2010 e 2015 representem uma taxa de crescimento anual de 27,6% sendo quatro vezes superior ao projetado pelo crescimento anual do mercado das TI, ou seja, a maior parte das organizações encontram-se de momento a migrar e implementar serviços de *cloud computing*. Este paradigma traduz-se essencialmente por um conjunto de características determinantes, tais como:

- Serviços *on-demand*,
- Acesso de rede alargado,
- Recursos partilhados,
- Elasticidade,
- Medição de utilização;

No primeiro ponto um utilizador de *cloud computing* deverá ser capaz de provisionar computação, capacidade de rede e armazenamento *on-demand*, ou seja, de forma simples, rápida, eficiente e automática. O acesso de rede alargado determina a capacidade de acesso e mobilidade de qualquer dispositivo na rede considerando um ambiente heterogéneo em tecnologias. Uma *cloud* deve proporcionar recursos partilhados ou a capacidade de agregação de recursos visíveis

de forma unificada para servir múltiplos utilizadores com recursos físicos e virtuais aplicados e reaplicados de acordo com o consumo efetuado.

Neste ponto tem-se ainda em consideração o serviço sempre disponível onde quer que esteja num ambiente físico já que o serviço é sempre disponibilizado por recursos partilhados (garantia de qualidade de serviço). A elasticidade é fundamental para garantir a rápida escalabilidade de uma estrutura computacional assim como garantir disponibilidade para novos serviços a qualquer momento. Por fim deverá existir a capacidade de medir a utilização dos recursos para o utilizador final validar o consumo que está a ser feito por uma aplicação em todas as componentes que a constituem na infraestrutura, nomeadamente, rede, armazenamento, computação e *backup*.

2.1.2. Modelos de Serviço e Implementação

Para implementação do paradigma do *cloud computing* existem duas considerações a ter em conta, nomeadamente, modelos de serviço e implementação [5] com o objetivo de proporcionar utilização ao consumidor final tendo como premissas as características mencionadas anteriormente.

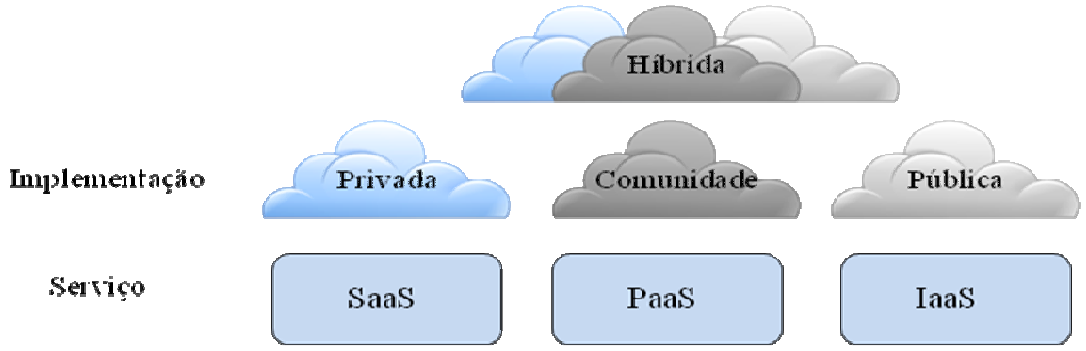


Figura 4 - Modelos de implementação e serviço

A implementação deste paradigma pode assentar num modelo privado, ou seja, a infraestrutura é dedicada à organização e pode ser gerida pela mesma ou por um prestador de serviço com a infraestrutura nas instalações ou fora destas. O modelo em comunidade implica que a infraestrutura seja partilhada por várias organizações suportando um conjunto de características partilhadas entre as mesmas, podendo estar nas instalações ou fora, e como exemplo os serviços centrais de uma universidade disponibilizando serviços a pólos universitários. O modelo público

traduz-se num modelo generalizado onde os recursos de infraestrutura são disponibilizados ao público, ou seja, a todas as organizações que pretendam serviço com recursos partilhados num prestador de serviço. No híbrido a arquitetura pode aplicar dois ou mais modelos que são agregados numa infraestrutura lógica através de meios estandardizados de mercado de modo a permitir o suporte tecnológico de ambientes e comunicação entre os dois modelos, sendo exemplo claro, um modelo privado, com infraestrutura dedicada agregado a um modelo público para efeitos de redundância física numa situação de contingência.

Na camada de serviço, o SaaS permite a disponibilização de aplicações que correm numa infraestrutura em *cloud*, sendo a mesma acedida através da internet por um *browser*. O utilizador não controla a infraestrutura onde assenta a aplicação, no entanto, utiliza serviço e parametriza a aplicação conforme o seu negócio, sendo um exemplo serviços de *mail* como o Google.
O PaaS tem a capacidade do utilizador criar e correr as suas aplicações numa infraestrutura em *cloud* através de módulos específicos disponibilizados baseados em linguagens de programação e ferramentas suportadas pelo prestador de serviço. O utilizador final não controla a componente física da infraestrutura mas controla as configurações e gestão das suas aplicações integradas neste modelo de serviço. Um exemplo deste tipo de serviço é a CloudFoundry que integra com um conjunto de aplicações e linguagens implementadas no mercado (VMWare, PHP, JavaScript).

O IaaS permite a disponibilização de um conjunto de recursos para uma organização, ou seja, disponibiliza recursos tais como rede, armazenamento, computação, entre outros. O utilizador final já tem a capacidade de gerir de forma limitada as componentes de rede, computação e armazenamento disponibilizadas pelo fornecedor de serviço. Um exemplo claro deste modelo é o vCloud Director da VMWare que permite criar centros de dados virtualizados podendo, por exemplo, aproveitar partilhar as mesmas gamas de IPs na mesma infraestrutura física.

2.2. Tecnologias

Para se endereçar a aplicabilidade de um modelo em c*loud computing* adotam-se um conjunto de tecnologias para garantir todas as funcionalidades inerentes ao modelo. Seguidamente serão abordadas as mais relevantes, assim como o seu endereçamento às problemáticas na transição do modelo convencional para a abordagem de modelo partilhado com foco no SO otimizado para o *cloud computing*.

2.2.1. Virtualização

A virtualização tem um papel fundamental no *cloud computing* já que permite agregar recursos da camada física de forma a partilhá-los, para se traduzir num modelo de consolidação de utilização da infraestrutura. A virtualização tem como objetivos consolidar recursos de forma eficiente e endereçar uma problemática que consiste na maior parte das organizações utilizarem uma fatia bastante reduzida da capacidade de processamento dos sistemas que possuem desperdiçando recursos e investimento. Desta forma recorrendo à virtualização de servidores é possível consolidar várias aplicações em sistemas físicos, reduzindo o número de sistemas para gerir e rentabilizando a infraestrutura [4].

Existem 3 tipos de tecnologias de virtualização [7] baseado em: virtualização ao nível do SO, servidores virtuais e para-virtualização ao nível da máquina virtual. O primeiro modelo é conseguido com o servidor a correr um único *kernel* e através deste é feito o controlo dos SO dos servidores virtuais. Basicamente são criados contentores isolados ou partições num único servidor físico, sendo que as instâncias de SO correm independentemente das outras partições. Um exemplo claro deste modelo é o SUN Solaris Zones.

O segundo modelo (servidores virtuais) baseado em servidores x86 assenta o *hyper-visor*, ou seja, a camada que coordena as operações dos recursos físicos (processamento, interfaces de rede, disco) com os servidores virtuais, sendo que existe uma máscara da camada física proporcionando uma abstracção dos SO que correm em cada servidor virtual. O exemplo mais conhecido dentro deste modelo é o VMWare ESX, que foi a tecnologia escolhida para o estudo da infraestrutura computacional devido á sua aceitação de mercado (líder mundial na virtualização de servidores). O último modelo tem semelhanças com o modelo de servidores virtuais, no entanto, existe uma modificação no SO de cada servidor virtual para permitir que estes corram correctamente. Um exemplo no mercado é o Citrix Xen.

A abordagem de virtualização de servidores num centro de dados levanta novas considerações quanto à arquitetura de sistemas, ou seja, o fato de se consolidar servidores num único servidor físico levanta necessidades como por exemplo a concentração de I/O na interface de rede, conetividade de rede virtualizada através de *switches* virtuais, assim como a partilha de outros recursos por parte dos servidores virtuais no mesmo servidor físico.

Os *switches* virtuais são uma peça fundamental para a segmentação de tráfego e otimização dos recursos de rede físicos num ambiente virtualizado. A figura 5 demonstra as componentes físicas e lógicas onde uma máquina virtual possui interfaces de rede virtuais (interface de redes emuladas ou paravirtualizadas) conectadas a *switches* virtuais que por si contêm *uplinks* para as interfaces físicas de rede. Uma grande vantagem desta arquitetura é a própria segmentação de tráfego para servidores que estejam no mesmo servidor físico, podendo um pedido entre servidores ser resolvido no servidor físico sem impactar a rede LAN, sendo que a comunicação entre servidores virtuais é suportada através dos mesmos protocolos que são utilizados em *switches* físicos. Outra aplicabilidade é criar as *Virtual Local Area NetWorks* (VLAN) e associar os respectivos servidores virtuais a cada. Tomemos como exemplo as redes de: vMotion ou a rede associada à passagem de servidores virtuais entre nós físicos, *backups*, dados, como por exemplo uma SAN para aceder aos dispositivos de armazenamento, monitorização e gestão.

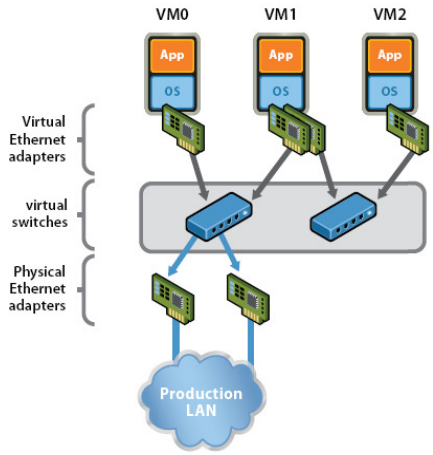


Figura 5 - Arquitetura de uma infraestrutura virtualizada [8]

A configuração de acesso à rede física deve possuir *Network Interface Card* (*NIC) Teaming* (agregação de interfaces) na ligação ao *switch* virtual de modo a ter-se redundância na infraestrutura virtualizada assim como as interfaces de rede *TCP/IP Offload Engine* (ToE) ajudam a aumentar o desempenho e libertar *Central Processing Unit* (CPU), já que o (des)encapsulamento *Transmission Control Protocol / Internet Protocol* (TCP/IP) é feito na interface. Um servidor virtual pode ser configurada com um ou mais adaptadores de rede, sendo que cada terá um IP associado assim como endereço *Media Access Control* (MAC), ou seja, cada servidor virtual tem as mesmas propriedades que uma máquina física de um ponto de vista da rede. Os *switches* virtuais proporcionam segurança à infraestrutura já que são imunes a tipos de ataque que envolvam funcionalidades VLAN. Caso o pretendido não seja as VLANs por questões

de maturidade do ambiente físico de redes ou não seja o método adotado, podem-se combinar as redes de gestão em um ou mais *switches* virtuais, no entanto, deve-se ter em conta a separação dos servidores virtuais em redes separadas da de gestão através dos *switches* virtuais com *uplinks* separados.

Se considerarmos na figura 5 como sendo um servidor físico correndo um conjunto de servidores virtuais, a velocidade à qual está ligado na rede física não é relevante para tráfego na rede virtual (interna) já que todo o processo de transferência de dados ocorre na memória RAM do sistema tendo a vantagem de não ocorrerem colisões ou outros erros que são comuns às redes físicas. Outra característica dos *switches* virtuais é o facto da sua carga no *hyper-visor* correr o apenas necessário para o seu funcionamento, ou seja, o sistema aplica a menor complexidade proporcionando mais desempenho ao sistema.

Relativamente á componente física de redes, os *switches* virtuais proporcionam igualmente a tabela interna de MAC *port-forwarding* e realizam tarefas tais como: validação destino MAC na chegada da trama, encaminhamento da trama para um ou mais portos e evita entrega desnecessária das tramas (não se comporta como um hub). Neste tipo de infraestrutura não é possível fazer interligação de *switches* virtuais, sendo o mesmo definido numa única camada e desta forma não é necessária a implementação do protocolo *Spanning-Tree Protocol*.

Uma das grandes vantagens da virtualização é a mobilidade na alocação de recursos, nomeadamente, a passagem de servidores virtuais entre nós físicos permitindo a continuidade de serviço em caso de falha no *hardware* assim como a flexibilidade na operação dos servidores, no entanto, deve-se acautelar os recursos disponíveis para suportar todas estas funcionalidades, ou seja, ter uma rede física bem desenhada, armazenamento partilhado entre outros recursos partilhados na infraestrutura [9].
A migração em tempo real de servidores virtuais entre físicos implica transferir todo o estado de execução da máquina do *hyper-visor* fonte para o de destino através de uma rede que os liga, sendo que esta transferência é representada por:

a) Estado de execução: CPU, configuração de rede, adaptadores adaptador *Small Computer Systems Interface* (SCSI) e informação do SO,
b) Ligações externas: componente de redes e dispositivos SCSI,
c) Memória física da máquina virtual,

O estado de execução refere-se à serialização do dispositivo virtual que é tipicamente inferior a 8 MBytes de dimensão, podendo em alguns casos dependendo da carga e configuração estender-se aos 128 MBytes [9]. A configuração das redes, é relativamente simplificada através do conceito de *switches* virtuais e interfaces virtuais pois estas últimas possuem os seus endereços MAC que são independentes do endereço da interface física desde que os *hyper-visors* fonte e destino estejam na mesma sub-rede. O processo de migração é feito de forma a não existir *time-out*, ou seja, assume rede dedicada ou largura de banda suficiente numa VLAN de modo a que a transferência seja realizada.

Depois da máquina virtual ser migrada, o *hyper-visor* destino envia um pacote *Reverse Address Resolution Protocol* (RARP) para o *switch* de rede físico de modo a assegurar que o mesmo atualiza as suas tabelas internas com as novas portas para endereçar a localização da máquina migrada sendo que este processo é completamente transparente para os clientes ligados a esta máquina virtual. Havendo lugar a armazenamento centralizado quer seja SAN ou NAS, torna todo este processo simples e rápido já que as imagens das servidores virtuais estão centralizadas e disponíveis para o servidor fonte e destino.

A memória física de uma máquina virtual é a componente de maior peso a ser transferida durante uma migração, sendo que, quanto maior este valor e taxa de alteração mais tempo demorará a transferência da máquina virtual entre servidores físicos. De modo a permitir que uma máquina virtual continue ativa durante o processo de migração existem três fases distintas para manter o estado da memória [10]:

a) Mapeamento da máquina virtual: o mapeamento é colocado nas páginas de memória da máquina virtual de modo a verificar qualquer modificação durante o processo da migração. Esta verificação pode causar uma quebra no tráfego de I/O sendo que este impacto é proporcional ao tamanho da memória a transferir,

b) Pré-Cópia: como o servidor virtual continua ativo durante a migração e com operações de alteração na memória do servidor fonte, é feita a cópia do *hyper-visor* fonte para o de destino num processo iterativo. A primeira cópia é da memória total, posteriormente, são feitos os mapeamentos das páginas de memória que sofrem as últimas modificações, sendo que o número de iterações e o número de páginas de memória entre cada, depende da taxa de alteração no *hyper-visor* fonte devido às operações de I/O geradas pela(s) aplicação(ões) residente(s) no servidor virtual. A maior fatia de migração de dados na rede LAN acontece nesta fase, ocupando uma curta fatia de CPU no entanto, como esta

fase implica criar um mapeamento na memória, vamos obter uma menor quantidade útil de memória para as aplicações na máquina virtual, logo maior impacto no desempenho das aplicações.

c) Comutação: durante esta fase a máquina virtual é colocada num modo *quiesced** no *hyper-visor* fonte e as últimas alterações de memória (antes da migração ficar terminada) são copiadas para o *hyper-visor* de destino sendo que a máquina virtual transitada fica temporariamente parada neste passo. Apesar desta duração ser geralmente inferior a um segundo é o ponto em que se denota uma maior latência num curto espaço de tempo, sendo que este impacto depende de vários fatores entre os quais: infraestrutura de rede, configuração do armazenamento partilhado, estrutura física, versão do *hyper-visor* e o I/O no servidor virtual.

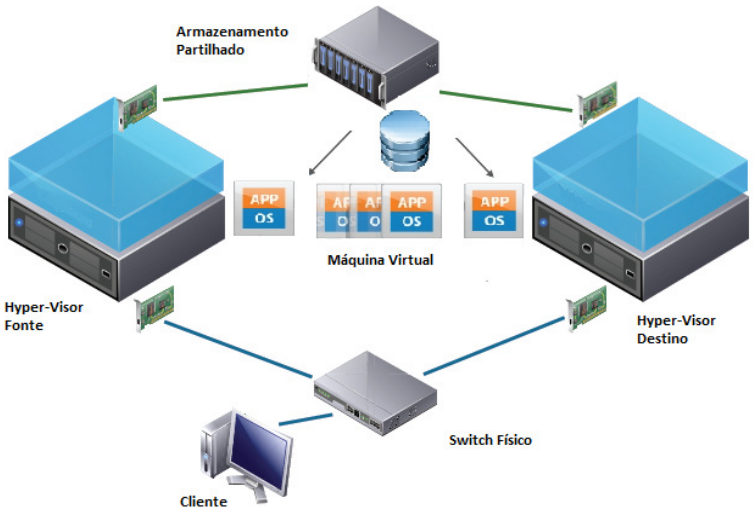


Figura 6 - Alta disponibilidade e mobilidade de servidores [9]

A configuração de rede de migração de servidores, nomeadamente largura de banda, depende essencialmente da aplicação que irá gerar mais carga e utilização de memória, sendo que teoricamente quanto maior a largura de banda desta rede mais rápido será o processo acima descrito. De acordo com a figura 6, uma máquina virtual a migrar entre servidores físicos, gera tráfego de migração (as três fases anteriormente referidas) sendo que o acesso aos dados é feito através do armazenamento partilhado (SAN ou NAS) de forma transparente para o cliente.

**quiesced* – termo que identifica que um computador, aplicação ou processo é colocado num estado temporariamente inativo e que pode ser ativo de forma rápida num dado ponto no tempo.

Para se tirar maior desempenho da infraestrutura é recomendada a separação física destes dois tipos de tráfego e configuração dos acessos dos servidores físicos ao armazenamento centralizado. A configuração de rede para suportar esta funcionalidade está dependente da carga das aplicações na infraestrutura e do plano de crescimento, no entanto, existem algumas considerações a ter em conta tais como [9]: utilização de rede dedicada 10 Gigabit Ethernet e/ou agregação de interfaces aquando de servidores virtuais com bastante carga (utilização igual ou superior a 64 GB de memória) e infraestruturas onde existam uma quantidade elevada de servidores virtuais pois há maior probabilidade de ocorrerem migrações de servidores quer por avarias ou balanceamento automático na infraestrutura.
Deve-se ainda ter uma margem nunca inferior a 30% de CPU alocado aos servidores pois caso a fatia de CPU seja reduzida a migração ocorre sendo o desempenho afetado, espaço em SAN ou NAS dimensionado quer em capacidade como desempenho e caso exista necessidade de misturar tráfego da migração e dados recorrer a tecnologia de qualidade de serviço do *hyper-visor* para particionar o consumo de largura de banda.

Em situações de alta disponibilidade num *cluster* de *hyper-visors* caso um servidor físico falhe, existe um mecanismo [10] capaz de reiniciar os servidores virtuais de forma automático noutro servidor físico, traduzindo-se desta forma numa arquitetura flexível e tolerante a falhas.

Para isto ser assegurado no *cluster* de *hyper-visors* existe um servidor físico que será o *Master* segundo uma eleição entre os nós (tipicamente o que tem mais volumes montados) e os restantes no *cluster* serão os *Slaves*, ficando o *Master* com o objetivo de: monitorizar o estado dos *Slaves* e caso um destes falhe são identificadas quais os servidores virtuais que necessitam de ser reiniciados, monitorizar o estado dos servidores virtuais, ou seja, caso um falhe, garanta que este é imediatamente reiniciado num servidor de forma automática e ainda gerir os *clusters* e proteger os servidores virtuais.

A metodologia utilizada para a alta disponibilidade funciona baseado nos seguintes passos:

a) O *Master* monitoriza os *Slaves* através de *heartbeats* a cada segundo através da rede LAN,
b) Se o *Master* não receber o *heartbeat* de um dos *Slave* procura um outro recetor antes de o declarar como falhado, e valida que o *Slave* possui *heartbeats* com volumes montados em armazenamento partilhado. Neste ponto, ainda valida se o *Slave* responde a pacotes *Internet Control Message Protocol* (ICMP) enviados para o seu IP de gestão,

c) Postas as condições acima, o *Slave* é declarado como falhado e os servidores virtuais que nele estavam contidos são agora iniciados em outros disponíveis baseado em políticas pré-configuradas,

Desta forma existem três situações distintas que podem levar á detecção de erro [11]: falha física no servidor, servidor ficar isolado numa perspetiva de rede e o servidor ficar sem conetividade para o *Master*.

A referência a um servidor ficar isolado numa perspetiva de rede, indica que um servidor não consegue verificar o tráfego dos *heartbeats* (rede de gestão) dos servidores de alta disponibilidade no *cluster* devendo-se por exemplo a uma falha física num *switch* de rede ou NIC físico levando ao particionamento da rede e consequentemente à degradação do desempenho dos servidores virtuais.
Se o servidor não conseguir receber estes pacotes, tenta enviar uma mensagem ICMP ping para os endereços IP de isolamento do *cluster* e se mesmo assim não conseguir, o servidor físico declara-se como isolado da rede. Como o *Master* faz toda a monitorização dos servidores virtuais que correm num servidor isolado, se verificar que estes estão desligadas e sendo este o responsável pelos mesmas, fará a reinicialização dos servidores.

A melhor forma de evitar esta situação é a configuração de um ambiente totalmente redundante de acordo com as componentes abaixo descritas:

a) Interfaces dos servidores:
    i. NICs,
    ii. *Host Bus Adapter* (HBA) caso seja SAN *Fibre Channel* (FC),
    iii. *Converged Network Adapter* (CNA) para redes unificadas *Fibre Channel over Ethernet* (FCoE),
b) Servidores,
c) Redes LAN,
d) Armazenamento e redes SAN ou NAS,

A configuração de redes é extremamente crítica para permitir uma boa resiliência de modo a assegurar o acesso ininterrupto dos utilizadores às aplicações que correm na *cloud* privada. Os *switches* de rede físicos deverão suportar a opção "*PortFast*" ou equivalente para prevenir que um servidor determine que a rede está isolada durante a execução do algoritmo *Spanning-Tree*

aquando da reinicialização do *switch*, utilizar a resolução de DNS, utilizar nomes consistentes para as VLANs em todos os *hyper-visors* do *cluster* e por fim garantir que a *firewall* no ambiente não compromete a comunicação entre portos específicos para os *heartbeat*. A redundância é outro fator chave quer em termos de *switches* físicos como na agregação de interfaces de rede para garantir a alta disponibilidade em caso de falha de uma das interfaces ou *switch*, ligações de cada interface a *switch* de rede distintos para se obter caminhos independentes para envio/recepção de *heartbeats* com a ressalva de que uma interface está ativa de cada vez pois o *heartbeat* apenas é transmitido/recebido por um único caminho.
Para além da configuração de caminhos redundantes entre servidores deve-se reduzir ao máximo os caminhos de rede no *cluster* pois quantos mais *hops* na rede maior a latência entre os *heartbeats* e especificar os endereços IP de isolamento caso existam vários *clusters*, sendo que por defeito é utilizado o *default-gateway* como endereço de isolamento de rede.

O acesso ao armazenamento em SANs FC ou iSCSI possuem requisitos como o *multi-pathing* para garantir a alta disponibilidade e balanceamento de tráfego entre o *cluster* de *hyper-visor* e o dispositivo de armazenamento externo. Caso exista uma falha ao nível da interface de acesso á rede de dados no servidor, um cabo de rede ou fibra ótica, um dos *switches* ou um controlador do dispositivo de armazenamento externo assegurar que a máquina virtual continua a operar sem qualquer paragem.

2.2.2. Redes locais

A introdução da virtualização numa infraestrutura implica várias considerações ao nível de redes, quer seja a nível de balanceamento, tolerância á falha, redundância, serviço e integração com a componente de virtualização. Como tal abordam-se técnicas como:

a) Agregação de interfaces de rede
b) Balanceamento de tráfego
c) Alta disponibilidade
d) Segmentação de tráfego

A agregação de interfaces de rede tem extrema importância num ambiente virtualizado, não só pela alta disponibilidade assim como a largura de banda disponibilizada aos servidores virtuais e consequentemente aplicações que nelas geram tráfego. Neste tipo de infraestrutura pode-se ligar

um único *switch* virtual a múltiplas interfaces de rede físicas usando a tecnologia de agregação, que suporta a carga entre ambientes físicos e virtuais. Numa situação de falha quer da interface física quer do lado da rede, esta configuração permite *failover* passivo, ou seja, á primeira falha ocorrida encaminha o tráfego para as interfaces disponíveis na agregação.
O balanceamento de tráfego numa infraestrutura virtualizada implica alguns desafios que não se verificam nas redes físicas, ou seja, um único servidor físico engloba vários virtuais assim como toda a carga de I/O que flui nas interfaces de rede. Desta forma os desafios prendem-se com questões de nível de serviço que pretendemos aplicar a cada aplicação e consequentemente a cada servidor virtual. O NIC *Teaming* é essencial neste tipo de infraestrutura no entanto existem um conjunto de políticas [8] que podem ser aplicadas para o balanceamento dos dados:

a) *Routing* baseado na identificação da porta do *switch* virtual,
b) *Routing* baseado no *hash* endereço MAC na fonte,
c) *Routing* baseado no endereço IP,

O primeiro método implica que o tráfego de uma dada interface de rede virtual seja enviado de forma consistente para uma interface física (com exceção de uma situação de *fail-over*). A resposta será recebida pela mesma interface física pois o *switch* físico ficou com o registo da associação da porta. O segundo método é baseado num *hash* do endereço MAC da trama Ethernet onde o tráfego de uma interface de rede virtual é enviado consistentemente para um físico sendo as respostas recebidas na mesma interface de rede (*switch* físico aprende a associação). Neste método uma única máquina virtual não pode utilizar mais do que uma interface física, exceto, quando recorre a múltiplos endereços MAC. Ambas as técnicas permitem uma boa distribuição de carga se o número de interfaces virtuais for muito superior ao de interfaces físicas.
O terceiro método baseia-se no *hash* do endereço IP (fonte e destino) de cada pacote transmitido e caso exista algum tipo de pacote não IP o *hash* é calculado mediante o *offset*. A distribuição de tráfego depende do número de ligações TCP/IP sendo que se pretender ter um caminho mais rápido para um único servidor virtual pode-se recorrer a agregação de interfaces que ajuda a prevenir reflexão de pacotes pois não retransmite o tráfego *multicast* ou *broadcast*.

A alta disponibilidade na rede física é assegurada pela redundância de ligações para *switches* de rede físicas e com a agregação de interfaces de rede no servidor físico. No entanto, se existir uma falha no troço de rede ou mesmo uma avaria de um *switch* físico existem técnicas adjacentes à virtualização para garantir a continuidade de operação. Para ir de encontro a esta problemática existem dois métodos para deteção de *failover*, sendo estes [8]: estado do troço e *beacons*.

O estado do troço baseia-se essencialmente no estado do troço reportado pelo adaptador de rede permitindo identificar situações tais como desligar cabo, falha de energia no *switch*, falhas nas configurações do *switch* (erro configuração VLANs, portas bloqueados pelo protocolo *Spanning-Tree*) ou mesmo cabos desligados em outra ligação do *switch*.

O método de *beacons* baseia-se no envio de pacotes (*beacons*) na rede Ethernet através de *broadcasts* de modo a detetar falhas de conectividade na rede em todas as ligações configurado de acordo com as figuras:

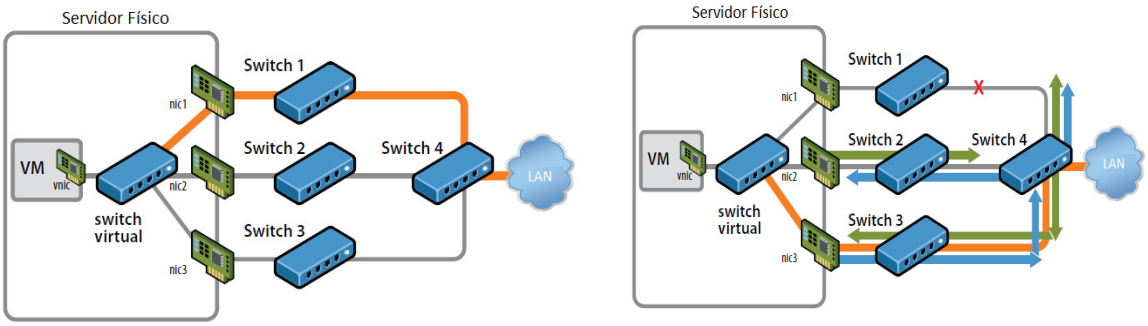


Figura 7 - *Routing* de tráfego em situação de falha [8]

Este método baseia-se igualmente no estado do troço para determinar falha na primeira ligação, endereçando as restantes problemáticas anteriormente identificadas de acordo com o exemplo da figura 7, onde a falha de um troço de rede faz com que cada adaptador Ethernet envie um *beacon* na rede, sendo a receção feita apenas pelos adaptadores 2 e 3 que serão os disponíveis para o *routing* de dados.

Por defeito a agregação de portas aplica política de *fail-back*, ou seja, um adaptador de rede que falhou mas que entretanto ficou disponível vai voltar a desempenhar a sua função despromovendo o adaptador que entrou em funcionamento aquando do evento. Caso um adaptador físico possua falhas num servidor físico, o *switch* de rede vai verificar mudanças de endereço MAC sendo que poderá não aceitar tráfego assim que um adaptador fique disponível. Para minimizar o impacto que este processo pode gerar devem-se ter em consideração desabilitar o uso do protocolo *Spanning-Tree* no *switch* de rede físico ligado ao servidor físico. Caso o mesmo seja Cisco deve-se recorrer à interface configurada em *portfast mode* ou *portfast trunk*, diminuindo o tempo de inicialização da porta do *switch* e ainda desabilitar a negociação de *trunking* (se aplicável). O VMWare VSphere permite configurar as políticas de *failover* de modo a definir a distribuição da

carga para os adaptadores físicos de rede no servidor físico devendo ter-se em consideração que um servidor virtual não pode utilizar mais que um adaptador físico exceto se configurados múltiplos adaptadores de rede.

Com o aumento de servidores virtuais na infraestrutura surgem desafios relativamente á consolidação de servidores em diferentes zonas (*Trusted Zones*). Uma *Trusted Zone* é definida como um segmento de rede no qual os dados que nela fluem são relativamente livres, sendo que todo o tráfego que entra e sai está sujeito a restrições superiores. Um claro exemplo desta segmentação são as DMZs ou uma segmentação lógica para um determinado conjunto de aplicações num dado ambiente, ou seja, uma aplicação *web* com serviços de bases de dados no seu *back-end*. Uma abordagem de infraestrutura virtualizada não implica necessariamente uma alteração à arquitetura de redes já que é possível consolidar servidores sem misturar as *Trusted Zones*. De acordo com as melhores práticas VMWare existem três abordagens [12] que permitem segmentar o tráfego de rede e isolar zonas do ambiente: zonas físicas, virtualizada ou interna.

A primeira abordagem implica separação física das zonas tal como se pode verificar na figura 8, onde em cada zona existem *hyper-visors* dedicados.

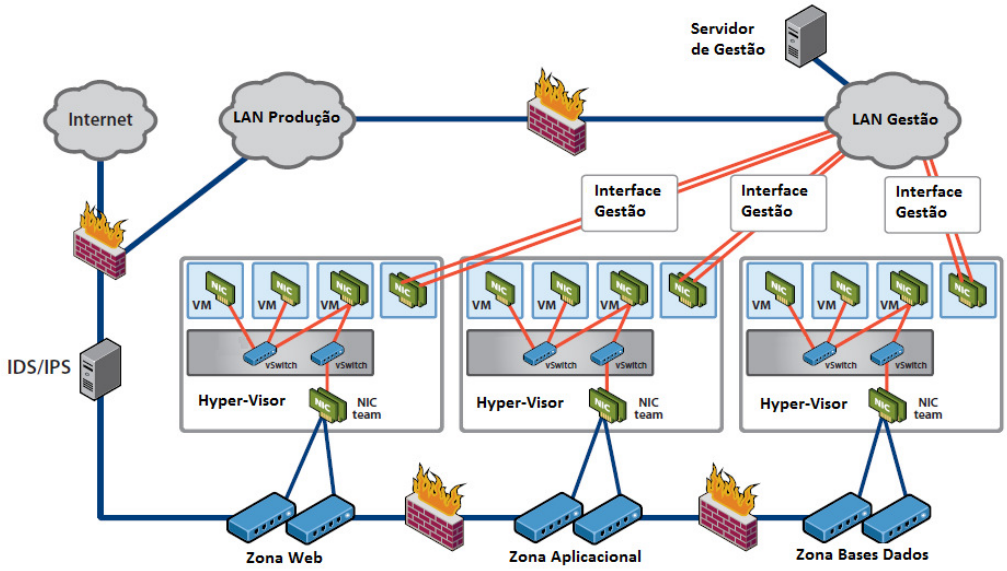


Figura 8 - Segmentação por zonas físicas [12]

O isolamento das zonas é baseado através de dispositivos de rede físicos e a principal diferença para uma infraestrutura física são apenas os servidores de cada zona estarem virtualizados. Esta configuração limita a utilização de recursos assim como o nível de consolidação, no entanto, o impacto num ambiente tradicional (físico) é mínimo já que não implica configurações específicas,

riscos associados de má configuração, não existe a necessidade de se configurar VLANs e todo o isolamento é feito no *switch* físico.

Na segunda abordagem recorre-se à virtualização onde se podem colocar servidores virtuais no mesmo *hyper-visor,* não obstante, com segmentação das aplicações para *switches* virtuais dedicados e/ou interfaces físicas dedicadas. Como nesta abordagem a consolidação de servidores físicos é uma realidade, necessita-se de menos infraestrutura sendo que o acesso entre zonas ocorre quer na camada virtual (definição dos servidores virtuais ligados á respetiva zona) como na física (aplicação de políticas de segurança).

Esta situação obriga o administrador da plataforma a ter configurações mais precisas e qualquer falha na segmentação pode introduzir risco na infraestrutura, mesmo assim, existe a componente física com a definição de controlo de acessos que permite trazer segurança complementar à da camada virtual. Apesar de na figura 9 se ter *switches* virtuais separados é possível aplicar VLANs para consolidar as redes, no entanto, deve-se dedicar pelo menos duas interfaces de rede física (redundância) à rede de gestão da virtualização.

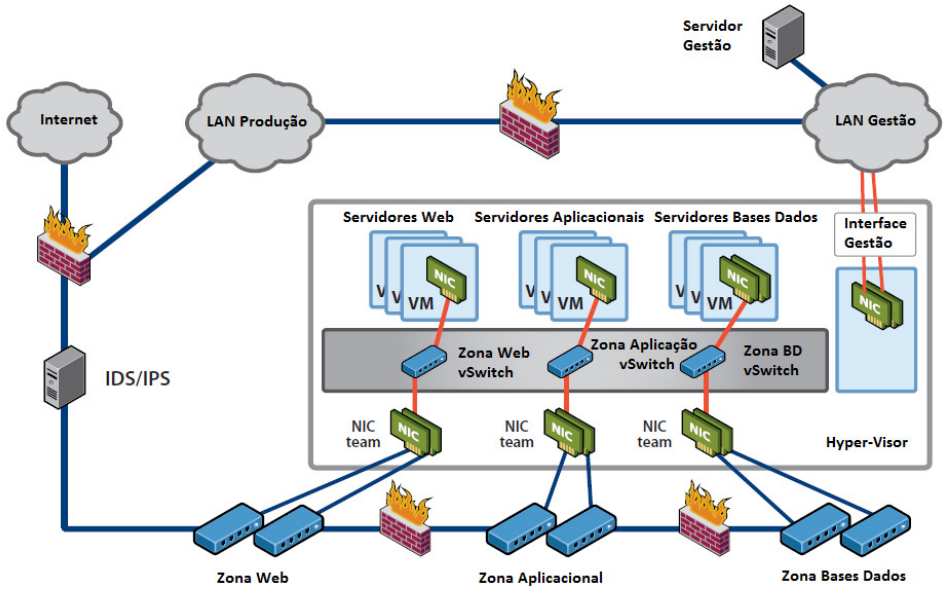


Figura 9 - Segmentação virtualizada [12]

Na terceira abordagem o grau de consolidação, gestão e virtualização são ainda superiores recorrendo a *Firewalls* virtuais otimizadas que aplicam a segurança entre zonas. É possível aplicar níveis de segurança distintos no(s) mesmo(s) servidor(es) físico(s), no entanto se *Trusted Zones* estiverem em diferentes segmentos de rede, o *routing* é feito a nível físico se não existir nenhum router virtual tal como descrito na figura 10. Das três abordagens apresentadas esta reflete a mais complexa pois introduz algum risco associado às más parametrizações que poderão

ser feitas na infraestrutura e a melhor forma de reduzir o risco é a separação de tarefas com permissões e funções distintas para cada gestão específica.

Quando é desenhada a arquitetura de sistemas deve-se ter em conta a comunicação entre partes de modo a garantir a segmentação correcta entre servidores virtuais em redes separadas sendo que neste último o *routing* é assegurado através de *firewalls* virtuais ou outro mecanismo de segurança em utilização. Deve-se igualmente validar todas as configurações e a sua consistência no ambiente virtual pois basta a política de segurança ser distinta entre *Firewall* virtuais ou *switches* virtuais para ter ocorrências como quebra de ligação (exemplo: insucesso de uma migração de um servidor virtual entre servidores físicos).

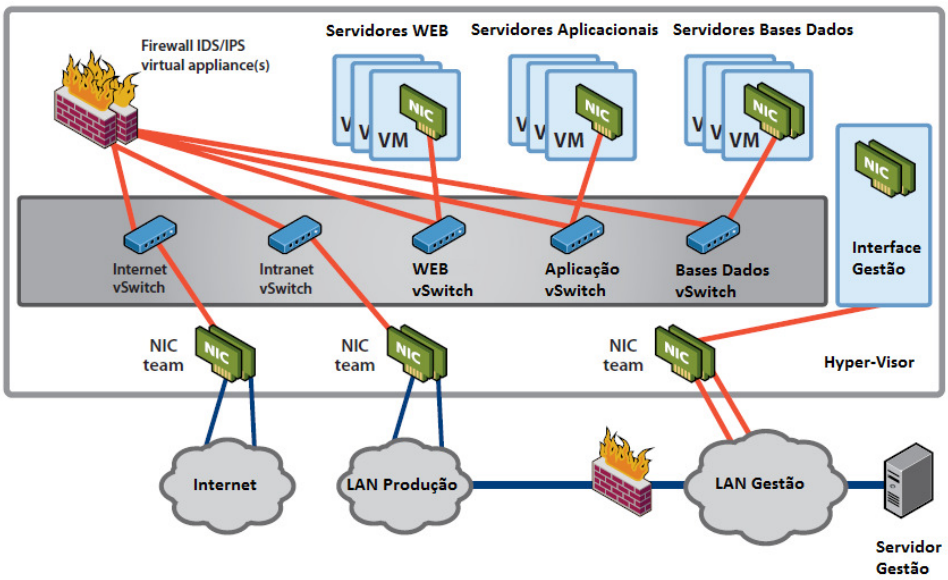


Figura 10 - Segmentação interna [12]

Numa perspetiva de implementação esta abordagem implica maior risco, maior complexidade, maior investimento de tempo na operação da infraestrutura em benefício do custo e da gestão da mesma.

2.2.3. Redes de dados

Com a evolução e expansão das TI, estudos [2] revelam o fenómeno relativo ao crescimento exponencial da informação nas organizações, passando o armazenamento de dados a ter uma relevância bastante alta não só para proteger os dados mas como endereçar o desempenho. As tecnologias FC, iSCSI (também conhecido por IP SAN), FCoE baseadas em leituras e escritas de

blocos e NAS baseado em leituras e escritas de ficheiros são as principais tecnologias implementadas para as redes de dados, sendo o FCoE a tecnologia mais recente e a mostrar sinais de forte expansão no mercado. O intuito das redes de dados é otimizar o acesso e transferências de dados em redes dedicadas de altos débitos, facilitar a integração do armazenamento. É importante referir que todas as tecnologias referidas anteriormente recorrem ao protocolo SCSI para troca de operações de leitura ou escrita em dispositivos de armazenamento.

O SCSI é um protocolo estandardizado utilizado por aplicações para enviar comandos ao armazenamento, incluindo operações de leitura e escrita designada por I/O que é também referido como comandos SCSI. Os comandos SCSI podem conter dados de controlo ou de informação e são enviados pelas aplicações em modo embebido através de protocolos que os encapsulam em redes Ethernet (iSCSI) ou *Fibre Channel (FC).*

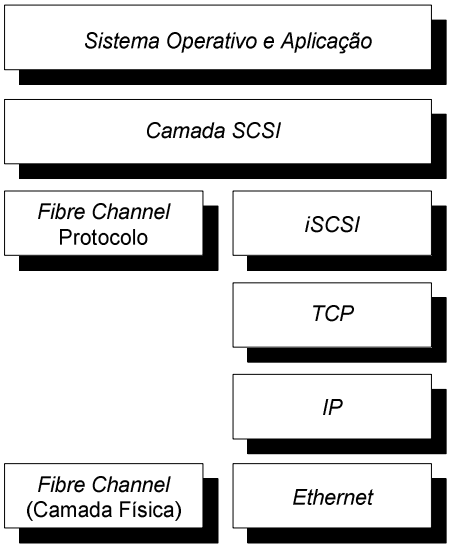


Figura 11 - Camadas SCSI em redes FC e iSCSI

A principal diferença entre as duas abordagens prende-se com o facto de em iSCSI utilizar-se recursos já existentes como *switches* Ethernet e interfaces de rede enquanto com a adoção do FC necessita-se de adaptadores de fibra, denominados HBAs e *switches* FC. Relativamente às camadas ilustradas na figura 11, o iSCSI recorre para transporte e rede aos protocolos TCP/IP enquanto o FC incorpora estas duas componentes no seu *standard* com *overhead* minimizado comparativamente ao iSCSI. Apesar de hoje em dia a maior parte das interfaces de rede implementarem o encapsulamento TCP/IP (interfaces de rede TOE) o *overhead* ainda é superior o

que denota que o FC seja o protocolo recomendado para ambientes altamente transacionais, ou seja, com números elevados de I/O e sensíveis a latência.

As redes tradicionais operam num ambiente aberto e com comportamentos imprevisíveis, sendo que qualquer dispositivo numa rede pode comunicar com qualquer outro, levando a que exista um esforço de implementação de métodos de verificação de acessos, o estabelecimento de sessões e *routing*. Com o aumento da quantidade de dados e exigência de altas taxas de transferência, surge a necessidade de um protocolo capaz de endereçar esta problemática nas infraestruturas computacionais. A tecnologia FC é a que possui maior implementação de mercado para a área de redes de dados com acessos baseados em blocos de dados e permite altos débitos conjugados com uma baixa latência, sendo ideal para ambientes transaccionais como bases de dados ou OLTP. As redes de dados FC detêm um conjunto de características que permitem flexibilidade, controlo de acessos, redundância e alta disponibilidade assim como métodos de autenticação dos seus clientes.

As redes FC podem ser configuradas em diferentes tipologias, tais como [13]:

- Ponto a ponto – dois dispositivos ligados diretamente,
- *Arbitrary-loop* – ligação em anel entre dispositivos sendo implementado, por exemplo, em *back-ends* de sistema de armazenamento,
- *Fabric-switched* – através de SAN *switches* FC.

As SANs convencionais são implementadas através de *fabric-switch*, no qual, um *fabric* pode corresponder a um ou mais *switches* e a ligação é representada de acordo com a figura 12:

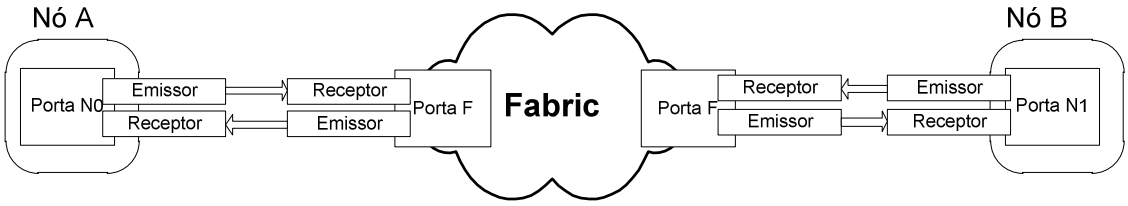


Figura 12 - Fabric-Switched

Um dispositivo na rede possui uma porta FC, no qual liga por exemplo, um cabo OM3 que tem 2 fibras sendo uma para receção e outra para emissão ligada a um *switch* FC. Cada nó ou dispositivo na rede possui portas N ou NL enquanto os *fabrics* possuem portas F ou FL, sendo o "L" caso a topologia seja em *loop* e ainda as portas "E" que permitem a expansão do *fabric* para realizar *inter-switch links* entre SAN *switches*. A ligação entres dispositivos (nó A e B)

denomina-se por zona e podem ser definidas mais zonas entre dispositivos para assegurar um maior conjunto de caminhos ativos para balanceamento de dados. Uma zona é constituída por membros, ou seja, por identificadores únicos de 64bit *World Wide Port Name* (WWPN) neste caso do nó A e B, sendo que, quando é feito o processo de zona no *fabric*, o SAN *switch* vai ainda atribuir no processo de *login* um identificador (FCiD) a cada membro da zona e a partir desse momento dá-se o processo de descoberta. As zonas têm relevância num contexto de isolamento para prevenir que as portas do *switch* vejam outras que não autorizadas, no entanto, existem dois métodos de implementar segurança na zona: *soft* e *hard zoning*.

O primeiro consiste em permitir que um dispositivo pesquise no *fabric* um conjunto de WWPN que esteja autorizado para se ligar, enquanto o segundo método assegura a zona à porta e permite que um dispositivo veja apenas um único WWPN que será o da porta definida sendo o mais seguro. A integração de uma rede de dados com um sistema de armazenamento levanta requisitos para assegurar redundância, alta disponibilidade e gestão de *fail-over*, ou seja, considerando um nó (servidor) ligado a uma rede de dados temos vários pontos de falha a endereçar, tais como: HBA, ligações ao *fabric*, quantidade de *fabrics*, controladores do armazenamento e respetivas ligações aos *fabrics*.

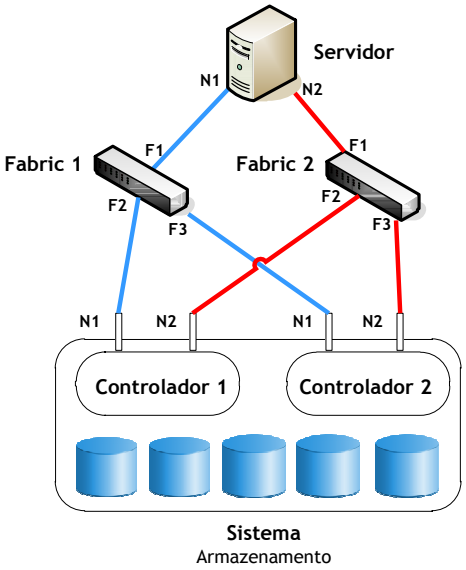


Figura 13 - Alta disponibilidade de uma rede de dados

Para se tirar partido de alta disponibilidade da infraestrutura, deve-se considerar, duas HBA, cada uma com uma porta, obrigando a ter disponível entradas *Peripheral Component Interconnect Express* (PCIe) no servidor, mas em caso de falha de uma placa a outra assegura a continuidade de operação. Dois *switches* FC garantem a alta disponibilidade do servidor assim como o sistema

de armazenamento possuir dois controladores com portas FC para ligações redundantes. Desta forma, a alta disponibilidade do servidor é garantida através de quatro zonas, com os respetivos membros:

a) Zona1: Servidor_N1 - Fabric1_F1 - Fabric1_F2 – Controlador1_N1,
b) Zona2: Servidor_N1 - Fabric1_F1 - Fabric1_F3 – Controlador2_N1,
c) Zona3: Servidor_N2 - Fabric2_F1 – Fabric2_F2 – Controlador1_N2,
d) Zona4: Servidor_N2 – Fabric2_F1 – Fabric2_F3 – Controlador2_N2,

Toda a gestão dos caminhos, balanceamento de dados, *fail-over* é gerida pelo *software* de *multi-pathing* presente no servidor que é responsável nas camadas do I/O por injetar os dados pelos canais disponíveis assim como em casa de falha de uma interface, fazer o *fail-over* para os canais disponíveis. Existem diversos *softwares* capazes de realizar estas tarefas, quer sejam nativos aos SO, quer sejam desenvolvidos por outros fabricantes.
As redes iSCSI recorrem a tecnologia já presente nomeadamente a *switches* de LAN tradicionais e apresentam como vantagem, face à rede FC, o suporte de altas distâncias sem recorrer a *hardware* adicional, servindo-se apenas da conectividade de rede. Para ser assegurada a alta disponibilidade deve-se igualmente recorrer a dois *switches* Ethernet sendo que é uma boa prática [3] dedicar *switches* para um SAN iSCSI, no entanto, se não for possível e o ambiente não for tão exigente numa perspetiva de carga, é possível recorrer às VLANs.

Existem um conjunto de fatores [3] que influenciam o desempenho de uma rede de dados iSCSI, tais como, contenção de rede, *routing* ineficiente e erros de configuração na LAN/VLAN.. O *routing* de tráfego é possível numa rede iSCSI, no entanto, introduz latência pelo que, o servidor (*initiator*) e o sistema de armazenamento (*target*) deverão estar na mesma sub-rede sem *gateways* definidas na porta iSCSI. Para se maximizar o desempenho de uma rede de dados iSCSI existem considerações determinantes para rentabilizar a infraestrutura, sendo estes:

a) Jumbro Frames,
b) Pause Frames,
c) TCP Delayed Ack

O recurso às *Jumbro Frames* pode aumentar o desempenho de uma rede até 50% [3] para determinados ambientes pois permitem mais comandos SCSI e um *payload* superior comparativamente com pacotes SCSI tradicionais, minimizando a fragmentação. Para recorrer a este método deve ser garantido que todos os elementos no troço de rede suportam as *Jumbo*

Frames tendo como valor máximo os 9000 Bytes fazendo com que a largura de banda de rede seja maximizada.

As *Pause Frames* permitem que um servidor temporariamente pare todo o tráfego vindo do sistema de armazenamento com o objetivo, em conjunto com o *switch,* de controlar as taxas de transferência. Devido à natureza e aos perfis de I/O esta funcionalidade não deverá ser ativa numa rede iSCSI devido á introdução de latência na comunicação entre servidor e armazenamento.

Em ambientes Windows e VMWare ESX, o TCP *Delayed Ack* permite atrasar um *acknowledge* para um pacote específico recebido num servidor e quando ativo este atraso chega até aos 500 ms ou até receber dois pacotes. Como durante o período de espera não existe comunicação entre servidor e armazenamento, o servidor inicia comandos SCSI *inquiry* para todas as *Logical Unit Number* (LUNs) a fim de obter informação sobre as mesmas o que causa mais tráfego e em situações de congestão pode diminuir o desempenho da rede de dados.

Existem diversas opções quanto à interface física, nomeadamente, NIC, NIC TOE e iSCSI HBA, sendo que todas recorrem ao *driver* iSCSI e diferem a nível de processamento no *stack* do SCSI *initiator*. As redes de dados iSCSI associam *targets* para um determinado servidor (*initiator*) conseguir "ver" discos de um sistema de armazenamento partilhado e recorrem, quanto a autenticação ao protocolo *Challenge-Handshake Authentication Protocol* (CHAP) largamente adotado em redes LAN.

O protocolo FCoE impulsionado pela Cisco tem como objetivo atingir outro grau de consolidação num centro de dados, ou seja, unificar comunicações quer LAN como SAN numa rede dedicada, reduzindo a gestão de elementos de rede, cablagem, interfaces nos servidores e reduzindo custos.

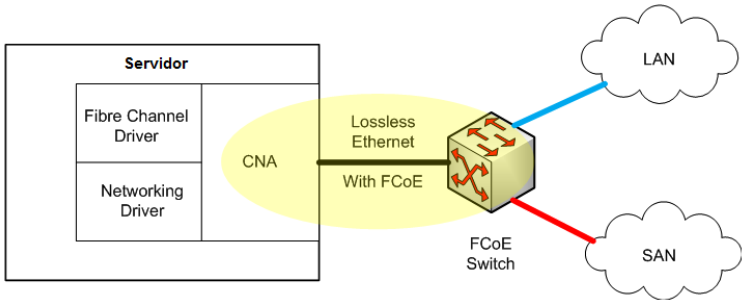


Figura 14 - Rede FCoE [3]

O protocolo FCoE encapsula o protocolo FC em redes Ethernet de 10 Gigabit havendo uma clara evolução nos próximos anos para o aumento da largura de banda desta tecnologia. De acordo com a figura 14, um servidor deixa de ter dois tipos distintos de interface, ou seja, rede LAN (NIC) e

SAN (HBA), consolidando numa única placa denominada CNA onde num único cabo passam dados LAN e SAN ligando a um *switch* que suporte o protocolo FCoE (sendo exemplo o Cisco Nexus) que permite interligar dispositivos de uma rede tradicional e sistemas de armazenamento.

O tipo de tráfego que flui nas redes de dados possui características distintas de uma rede LAN tradicional, pois nela flui tráfego de comandos SCSI e dados. O bom dimensionamento de uma rede de dados passa ainda por conhecer o tipo de aplicação que nela vai gerar tráfego assim como os dispositivos de armazenamento que estão registados, deste modo, estes dispositivos devem ter igualmente dimensionamento em conformidade para garantir a redundância e tolerância a falhas mas também o desempenho estipulado para a aplicação. Neste contexto enquadram-se os tipos de tráfego existentes numa rede de dados:

a) Escritas ou Leituras,
b) Blocos superiores ou inferiores,
c) Estável ou *Burst,*
d) Múltipla ou mono tarefa,
e) Sequencial ou aleatório,

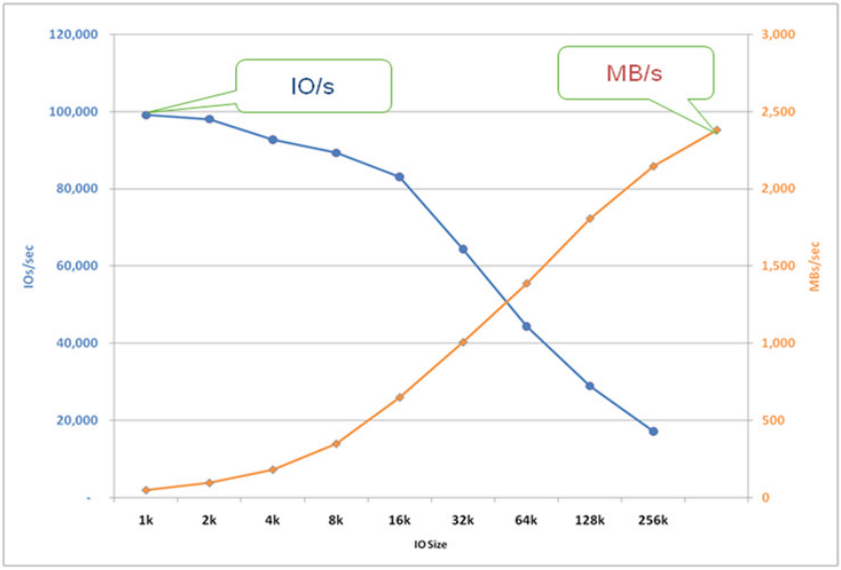


Figura 15 - Perfil tráfego de dados [3]

Estas considerações de tráfego levantam igualmente requisitos na tecnologia de discos que se enquadra para suportar o armazenamento da aplicação, assim como a proteção associada. É ainda tida em conta não só a capacidade de armazenamento com respetiva margem de crescimento mas também o tipo de desempenho que se pretende, ou seja, I/O.

Associado aos perfis já referidos, o tipo de tráfego designado por aleatório ou sequencial tem impacto distinto no desempenho do armazenamento assim como no seu dimensionamento. Quando se analisa o tráfego de dados em que o requisito fundamental é I/O, trata-se essencialmente de aplicações que geram tráfego com tamanhos de blocos pequenos como se pode validar no gráfico, ou seja, este tipo de perfil está essencialmente implícito em aplicações de bases de dados e sistemas OLTP.

Assim que escalamos em tamanho de blocos passamos para outros perfis no qual o I/O diminui e passamos a necessitar de largura de banda para altos débitos de transferência de dados de acordo com a figura 15, sendo exemplo claro deste perfil aplicações de *streaming* de dados, *data warehouse*, áreas de ficheiros e *backups* de dados. Não só a escolha de rede de dados depende do propósito de utilização assim como o tipo de dispositivo de armazenamento deve estar alinhado com esta necessidade.

2.2.4. Armazenamento Partilhado

O sistema de armazenamento permite consolidar toda a informação numa estrutura endereçada para o balanceamento e proteção de dados de forma eficiente, escalável e de fácil gestão. Existem diversos fabricantes no mercado que desenvolveram *hardware* específico para alto desempenho das aplicações suportadas num repositório centralizado que permite flexibilidade para replicação de dados, alta disponibilidade de aplicações combinado com a virtualização de servidores assim como proporcionar o desempenho para os ambientes mais exigentes.

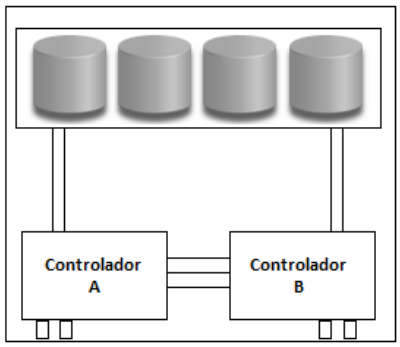


Figura 16 - Arquitetura de um sistema de armazenamento

Para se compreender melhor estes sistemas e o seu enquadramento numa infraestrutura computacional dinâmica apresenta-se a arquitetura tradicional destes sistemas na figura 16.

Essencialmente estes sistemas de armazenamento externo são compostos pelos seguintes componentes [3]:

a) Conetividade (ou *Front-Ends*) – portas de ligação da rede de dados ou servidores (DAS) aos controladores, podendo os mesmos suportar conectividade FC, Ethernet (para redes de dados iSCSI e efeito NAS), *Serial Attach SCSI* (SAS), FCoE entre outros;
b) Controladores – módulos de processamento de todos o I/O e operações internas composto por processadores, cache de leitura e escritas, memória e canais internos PCIe para comunicação entre componentes (por exemplo, comunicação entre controladores para efeitos de *fail-over*). Podem ter duas arquiteturas distintas: ativo/ativo ou ativo/passivo;
c) *Back-End* – conetividade entre módulos de processamento e as baías de discos, ou seja, gavetas onde são colocados os discos. A conetividade ao *backend* é feita através de portos de *backend* e canais de dados internos PCIe.

Os sistemas de armazenamento permitem apresentar LUNs aos servidores através de uma rede de dados (SAN), sendo que o SO vê um disco como se este fosse local, formatando-o para seu uso. Este método traduz-se em flexibilidade e centralização dos dados num repositório dedicado para este efeito permitindo ainda aplicar proteção tais como: *Redundant Array Independent Disks* (RAID), discos de *spare,* formatação de blocos com *checksums* em disco e alta disponibilidade assegurada por controladores redundantes e técnicas de *failover.*
Uma componente essencial num sistema de armazenamento é a sua cache que otimiza o I/O reduzindo o tráfego de *backend* e consequentemente processamento de paridades RAID, assim como diminui os tempos de resposta dos pedidos. A cache de um sistema de armazenamento pode ser vista como uma área de *staging* do I/O e desdobra-se em dois tipos com propósitos e algoritmos de otimização distintos: cache de leitura e escritas [3].

A cache de leitura aplica um algoritmo de *prefetch* que basicamente joga com o facto de carregar na cache blocos de dados se ocorrerem leituras de um conjunto de blocos de forma sequencial e é sempre a primeira localização quando existe um pedido de leitura, no entanto, se o bloco a ler não estiver em cache o sistema devolve o mesmo indo a disco (maior tempo de resposta). A cache de escritas combina blocos de escrita para nivelar o I/O de forma a eliminar a necessidade de rescrever dados (o armazenamento fornece o *acknowledge* à aplicação).
A cache é organizada em segmentos únicos (páginas) que possuem um tamanho de bloco fixo, ou seja, caso um determinado I/O seja menor que a dimensão da página, existe perda de eficiência da

cache já que vários I/O são realizados para partilhar a mesma página quando os blocos são contíguos. Enquanto a cache é utilizada para o I/O a memória do sistema é utilizada para processos internos tal como cálculos de paridade, replicações entre outras funcionalidades internas ao dispositivo de armazenamento [7].

Os sistemas de armazenamento implementam um conjunto de conectividade de *front-end* descritos no capítulo anterior mas também implementam tecnologias ao nível da camada física no seu *back-end* sendo os protocolos mais usuais o FC e o SAS [3]. Comparativamente as duas tecnologias foram desenhadas para suportar elevadas taxas de transferência de dados sendo no entanto a tecnologia SAS a que apresenta uma capacidade evolutiva superior com um custo tendencialmente mais baixo que o FC. A tecnologia SAS apresenta hoje em dia uma capacidade de 6 Gb/s de largura de banda, sendo implementada em sistemas de armazenamento da seguinte forma:

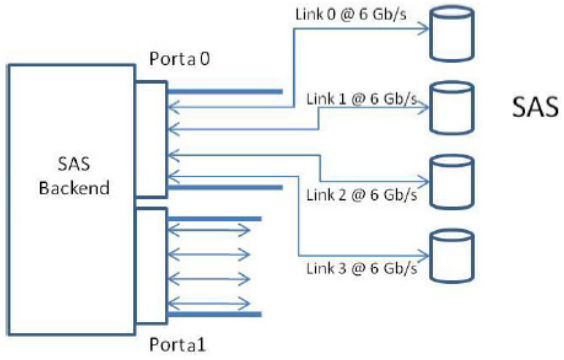


Figura 17 - Porta SAS de *backend* de 4 vias [3]

De acordo com a figura 17 em que cada porta apresenta 4 vias, cada uma, a 6 Gb/s, obtemos um total de 24 Gb/s e idealmente 3 GByte/s por porta sendo que o *framing*, *delays* entre outros pontos que adicionam *overhead,* totalizam uma largura de banda útil de 2,2 GByte/s por porta. Considerando que o tráfego do armazenamento flui sempre no bus interno PCIe, o limite por via é imposto no máximo suportado pela tecnologia de 500 MByte/s (PCIe 2.0), sendo o tráfego por porta máximo do PCIe multiplicado pelo número de vias, ou seja, 2 GByte/s teóricos por controlador (considerando 4 vias). As tecnologias de armazenamento tradicionalmente conjugam três tipos de discos distintos com características diferentes e igualmente com variadas aplicabilidades, estando agrupados da seguinte forma [3]:

a) SSD: Discos com tecnologia *flash* podendo ser *Single Level Cell,* com um bit por célula ou *Multi Level Cell*, com mais bits por célula, possuindo o primeiro menor latência que o

seguinte em detrimento da capacidade que consegue armazenar tendo como premissa de desempenho os 3500 IOPS,

b) SAS ou FC: Discos a 10 ou 15 mil rotações por minuto com interface FC ou SAS dependendo da conectividade de *backend* do armazenamento tendo como premissa os 150 IOPS(10krpm) e 180 IOPS(15krpm),

c) *Serial Advanced Technology Attachment* (SATA): Discos a 5,4 e 7,2 mil rotações por minuto, com menor custo por TB e maior capacidade, tendo como premissa os 90 IOPS.

Os sistemas de armazenamento partilhado possuem diferentes mecanismos de proteção [3] para a salvaguarda da informação, sendo estas implementadas através de *checksums* em disco e discos dedicados para funções de *sparing*, ou seja, dedicados para substituir automaticamente um disco sempre que exista uma falha. Um grupo RAID implica um conjunto determinado de discos agrupado com uma determinada proteção no qual vão assentar os volumes lógicos, sendo que a constituição destes volumes possui elementos lógicos *Logic Block Address* por disco de modo a espalhar todo o volume e respetiva informação por todos os discos. O RAID em espelho assegura uma maior proteção e também desempenho em determinados perfis de I/O com a penalidade de necessitar do dobro de discos para a mesma capacidade útil. É implementado em configurações tais como RAID1 (dois discos) e RAID1+0 (N discos em espelho e *striping*).

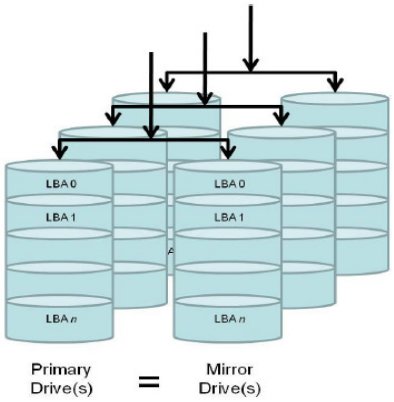


Figura 18 - Grupo RAID em espelho [3]

O RAID em paridade assegura uma proteção intermédia e também desempenho em determinados perfis com a penalização de necessitar de discos de paridade mas mais vantajoso em capacidade útil comparativamente ao RAID em espelho. O RAID em paridade é implementado de diversas formas através de configurações distintas tais como RAID 3(N+1), 5(N+1) e 6(N+2). O método em paridade implica um algoritmo que calcula paridade para guardar em disco a informação das aplicações e informação de paridade para permitir reconstruir a informação em caso de falha de

um ou dois discos conforme o RAID aplicado. Comparativamente a outras técnicas de RAID, em paridade implica carga adicional de CPU para cálculo de dados de paridade para serem armazenados em disco.

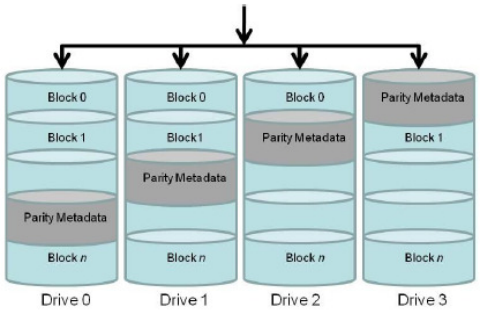


Figura 19 - Grupo RAID em paridade [3]

O RAID em *striping* traduz-se na associação dos discos físicos como se fossem um único disco lógico no qual um volume (LUN) está espalhado por todos de modo a todos colaborarem no desempenho. Este método é combinado com o método de espelho para uma variante do RAID1, ou seja, RAID10. O *striping* pode ser ainda implementado com método de paridade permitindo associar RAID a uma pilha de recursos disponíveis, onde um volume vai ser criado em cima de todos os discos disponíveis de modo a ter uma quantidade máxima de discos e maior desempenho.

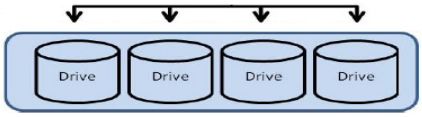


Figura 20 - Grupo RAID em *striping* [3]

Em suma, as configurações RAID existentes traduzem-se numa abordagem onde é importante ter em conta o tipo de tráfego de dados, proteção e custo associado:

RAID	Custo	Proteção	Desempenho (IOs)	Método
0	Baixo	Inexistente	Alta	*Striping*
1	Alto	Excelente	Limitada a 2 discos	Espelho
1+0	Alto	Excelente	Excelente	
3	Médio	Média	Média	Paridade
5	Médio	Média	Alta	
6	Médio	Alta	Alta	

Tabela 1 - Variantes de grupos RAID

De modo a se calcular e dimensionar uma infraestrutura de armazenamento quer para capacidade ou desempenho (I/O ou largura de banda) é necessário ter dados aplicacionais ou

comportamentais de escrita e leituras de uma aplicação, sendo que os dados são: percentagem de leituras e de escritas, total de I/O do servidor, escolha do RAID, formatação do bloco aplicacional, tipo de disco para utilização e capacidade útil. A percentagem de leitura e escritas reflete-se no padrão aplicacional e no rácio de escrita e leitura da aplicação, sendo relevante este ponto com o tipo de RAID escolhido já que os RAID em paridade elaboram mais I/O de escrita devido á escrita da informação e do bloco de paridade, sendo que este valor (pe) varia de acordo com o tipo de RAID:

$$pe_{(RAID\_1)} = 2$$
$$pe_{(RAID\_5)} = 4$$
$$pe_{(RAID\_6)} = 6$$

Considerando o valor de (pe) para RAID6 por cada escrita que é feita pelo servidor, no *backend* do armazenamento são realizadas 6 operações de escrita aumentando desta forma a quantidade de IOPS em disco. Desta forma estão englobados todas as variáveis para o cálculo de IOPS e largura de banda, respetivamente, para desenho de uma infraestrutura de armazenamento:

IOPS:

$$TotalIOs = hIOPSx[\%L] + pexhIOPSx[\%E]$$

$hIOPS \rightarrow IOPS\,do\quad servidor$
$\%L \rightarrow Percentagem\quad Leituras$
$pe \rightarrow Penalidade\quad escrita$
$\%E \rightarrow Percentagem\quad Escritas$

Largura de Banda:

$$RAID5_{(MB/s)} = L_{(MB/s)} + E_{(MB/s)}x\left(1 + \frac{1}{duRG}\right)$$
$$RAID6_{(MB/s)} = L_{(MB/s)} + E_{(MB/s)}x\left(1 + \frac{2}{duRG}\right)$$
$$RAID10_{(MB/s)} = L_{(MB/s)} + E_{(MB/s)}x2$$

$ddRG \rightarrow Discos\quad Úteis\qquad RaidGroup$

Para finalizar o desenho deve-se ter em consideração os seguintes passos [7]:

a) Análise da carga (padrão aplicacional),
b) Determinação dos I/O ou largura de banda nos discos,
c) Determinação do tipo e quantidade de discos,
d) Determinação da capacidade,
e) Determinação do sistema de armazenamento,

O *Thin Provisioning* é uma técnica de eficiência [14] que permite reclamar espaço que não está a ser utilizado pela aplicação, ou seja, se for criada uma LUN com um determinado espaço este será apresentado pelo armazenamento à aplicação, no entanto, a aplicação (dependendo do SO)

preenche o armazenamento com zeros sendo que mesmo que não utilize todo o espaço, este fica alocado e não pode ser reclamado. Esta técnica de eficiência permite que o armazenamento mascare o valor e apresente um valor mais baixo, por exemplo, alocar 100 GBytes para um serviço, mas como inicialmente apenas são colocados 10 GBytes de dados, o armazenamento reclama o espaço e apresenta á mesma os 100 GBytes, mas aloca apenas 10 GBytes mais uma margem consoante o fabricante, fazendo com que não exista desperdício de espaço em disco e reduza as necessidades de armazenamento. Apesar de esta técnica traduzir-se em eficiência, alguns fabricantes apresentam uma degradação no desempenho face a um volume totalmente aprovisionado (*Thick*) com diferença de que em *Thin* o armazenamento aloca blocos de informação à medida que a aplicação necessita.

A introdução de um sistema de armazenamento permite flexibilidade para replicação de dados (já que a mesma não impacta o servidor, a aplicação ou a rede local) através de tecnologias de replicação local, *snapshot* e *clones,* e replicação remota através dos *mirrors*. O *snapshot* de dados permite obter uma fotografia no tempo tal e qual como eram os dados, ou seja, se tirar um *snapshot* a uma hora, é possível ter um volume de dados da aplicação que posso apresentar a outro servidor para elaborar testes, *backup* ou outra função.

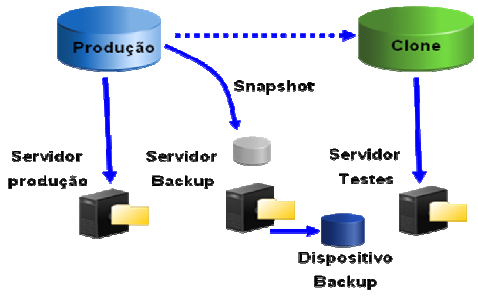


Figura 21 - *Snapshot* e clone de um volume

Um *snapshot* é nada mais do que um volume que contém apontadores para todos os blocos do volume de informação mais os blocos que entretanto foram modificados para garantir que o volume está consistente com o ponto no tempo a que foi criado. Esta técnica permite que tradicionalmente se recorra a um curto espaço em disco adicional para efetuar o *snapshot* já que para além dos apontadores que representam alguns *Bytes* tem que ser somado os blocos que entretanto foram alterados.

O clone permite criar uma cópia integral dos dados com a desvantagem de necessitar do espaço em disco, o equivalente à LUN de produção, no entanto, pode ser utilizado para efeitos de cópia

integral noutro grupo RAID (para redundância da informação) ou para efectuar *backup* sem afetar a respetiva aplicação.

A maior vantagem desta metodologia é a minimização do impacto no ambiente, nomeadamente, nas redes locais e de dados, pois se tiver de realizar *backup* em tempo útil de um volume através de uma rede local gigabit Ethernet necessitaríamos de ler todos os blocos de informação, enviá-los via rede local para o servidor de *backups* e por fim escrever os dados no dispositivo. Nesta abordagem, e de acordo com a figura 21 não se efetuaria a cópia via rede local mas apenas o último passo de escrever os dados, poupando tempo e aumentando a eficiência no *backup* da informação. Se se pretender um *backup* o mais transparente possível, pode-se apresentar um clone ao servidor de *backups* e ler de outra área de disco de modo a não despoletar I/O de leitura para não impactar a produção (exemplo: uma aplicação com desempenho constante).

A replicação remota visa obter a proteção em caso de disrupção no centro de dados principal, quer seja por falha energética, desastres naturais ou mesmo terrorismo. Desta forma as organizações continuam com os seus sistemas em produção, contando com algumas perdas de tempo de reposição de serviço ou *Recovery Time Objective* (RTO) e de informação ou *Recovery Point Objective* (RPO), através do redireccionamento dos seus utilizadores para o centro de dados secundário. O centro de dados secundário poderá estar implementado em qualquer tipo de modelo em *cloud* e a replicação de dados entre sítios é feito através da rede *Wide Area Network* (WAN) sobre IP, FC (limitado na distância através das comunicações) ou usando IP mas com conversores FCIP em caso de utilização de uma rede de dados FC para encapsulamento de pacotes FC em redes IP.

2.2.5. Backups

Os *backups* são uma parte crítica de uma operação em TI e servem para proteger dados críticos em caso de falhas como corrupção de dados, erro humano e falha de componentes de *hardware*. O objetivo do *backup* tipicamente divide-se em quatro áreas distintas [15]: recuperação de um sistema, como por exemplo recuperar um servidor, também designado por *bare-metal*, recuperação de dados, DR para em situação de desastre permitir ter cópias da informação de *backup* retida, e por fim requisitos legais. Uma estratégia de *backups* passa por definir e enquadrar as necessidades específicas para se proteger a informação: razão pela qual é feito o

backup, tipologia, quais as aplicações a fazer *backup* e considerações quer ao nível de dimensionamento de rede, disco, processamento, memória e janela de tempo disponível.

Para se compreender a arquitetura de uma solução de *backups*, existe o conceito de *software* de *backup* para gestão de toda a infraestrutura. Tradicionalmente na indústria um *software* gestor de *backups* tem um conjunto de componentes na sua arquitetura sendo estas [15]:

a) Servidor *Backups* e Clientes,
b) Catálogo e Índice,
c) *Backup online* e *offline*,
d) Servidor *Proxy*,
e) Metadados,
f) Retenção e Rotação,
g) *Staging* e *Cloning*,

O servidor de *backups* é a interface de gestão de todo o ambiente de *backups* sendo responsável pela operação da sua zona de *backups*, ou seja, responsável por um conjunto de clientes. O catálogo ou índice reflecte os *backups* que estão associados ao dispositivo onde é realizado, ou seja, fazem o registo e mapeamento dos *backups* para a área de armazenamento onde são salvaguardados. Para além deste mapeamento temos ainda dados tais como os tempos, nomes e extensão de ficheiros, sendo conhecidos como os metadados ou a informação da informação. Os clientes tradicionalmente são as aplicações ás quais se faz a salvaguarda da informação havendo módulos próprios de cada fabricante para integração com as aplicações para *backups online* sem paragem aplicacional. O servidor de *proxy* é um servidor que num ambiente de *backups* tem uma funcionalidade semelhante ao servidor de *backups* no cliente ao qual pertence, ou seja, é configurado num cliente e permite que o fluxo de dados seja direto para o destino.

A retenção e rotação dos *backups* definem um período no qual os dados são protegidos e renovados sendo que existem retenções associadas aos *backups* diários, semanais, mensais e anuais. O *staging* de dados são operações que podem dinamizar o ambiente de modo a tirar partido de uma maior flexibilidade na recuperação e salvaguarda de informação com camadas adicionais de armazenamento, enquanto o *cloning* permite através da própria consola de gestão dos *backups* realizar cópias integrais num dispositivo para outro (ex: disco para fita magnética).
A operação de início dos *backups* é despoletada por um *scheduler* definido no *software*, no entanto, e para se dar início é necessário verificar no catálogo que dados deverão ser protegidos.

De seguida o servidor de *backups* instrui o servidor *proxy* para preparar o dispositivo de armazenamento e posteriormente o servidor de *backups* pede os metadados a todos os seus clientes envolvidos na operação, assim como, indica para enviarem os seus dados para o servidor de *proxy*. Por conseguinte, os dados são escritos no dispositivo de armazenamento e o servidor de *proxy* avisa o servidor de *backups* do sucesso dos mesmos de modo a este poder fazer a atualização do catálogo.

A operação de recuperação de um ficheiro implica a pesquisa do mesmo no catálogo do servidor de *backups*, para saber a origem deste, dimensão assim como última data e local onde foi armazenado. O servidor de *backups* instrui o servidor de *proxy* para carregar o dispositivo de armazenamento, os dados são lidos e enviados para o cliente (aplicação) e o servidor de *proxy* envia os metadados da operação de recuperação para o servidor de *backups* atualizar o seu catálogo. De acordo com um dado ambiente aplicacional, quantidade de informação e janela de tempo existem três topologias distintas e uma variante que integra duas topologias para assegurar os *backup* de dados, sendo estas:

a) Diretamente conetada
b) Baseada em LAN
c) Baseada em SAN

A topologia diretamente conetada é tipicamente aplicada em ambientes com poucas aplicações, pois conforme se cresce em servidores a gestão vai sendo mais complexa.

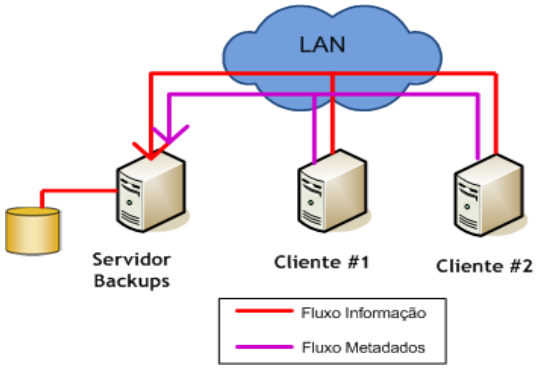


Figura 22 - Fluxo de dados com *backup* via rede LAN

O servidor de *backup* instrui o cliente (servidor de *proxy*), para este escrever diretamente os seus dados para o dispositivo de armazenamento. As vantagens nesta tipologia são: minimização de impacto na rede, pois *backups* passam diretamente para o dispositivo de escrita sendo obviamente mais rápidos, no entanto, tem como desvantagens a escalabilidade (conforme crescemos em

aplicações maior número de dispositivos de armazenamento têm que ser geridos), custo (devido aos dispositivos de armazenamento praticamente dedicados) e ainda a gestão que é agravada devido ao aumento do número de dispositivos. A topologia baseada em LAN implica que fluxo de dados passe na rede local até um servidor de *backups* ou *proxy* para serem posteriormente escritos no dispositivo de armazenamento.

Na figura 22 o servidor de *backups* instrui os seus clientes para enviarem dados para este escrevê-los num dispositivo de armazenamento. As principais vantagens são: maior escalabilidade com menor dispositivos de escrita associados, possibilidade de segmentação de tráfego na LAN ou rede dedicada de *backups*, simplicidade na adição de novos clientes à solução e ainda maior centralização da gestão comparativamente com a anterior. A grande desvantagem associada é o impacto na largura de banda da rede LAN onde grandes quantidades de dados podem congestionar a rede.

Esta tipologia pode ser associada com a anterior como por exemplo, realizar *backup* a um servidor com mais dados ligado a um segundo dispositivo de armazenamento o que implica mais gestão e custo. Na figura 23 o servidor de *backups* instrui os seus clientes (servidores *proxy ou media*) para enviarem dados diretamente para o dispositivo de armazenamento.

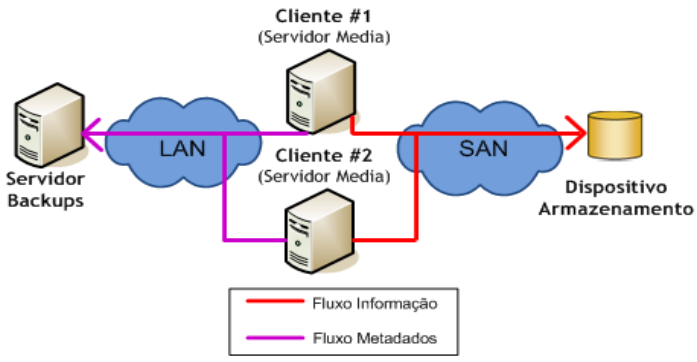


Figura 23 - Fluxo de dados com *backup* via rede SAN

As principais vantagens são: maior escalabilidade da solução com menor dispositivos de escrita associados, possibilidade de segmentação tráfego na SAN, *zonning* de dispositivos apresentados na SAN para os clientes configurados como servidor *proxy*, alto desempenho dos *backups*, flexibilidade na apresentação de dispositivos, como por exemplo, um sistema *Virtual Tape Library* (VTL) que é capaz de emular dispositivos de armazenamento e apresentá-los via SAN para múltiplos servidores consolidando a gestão e custos associados a *hardware* e minimização

do impacto nas redes LAN. As desvantagens associadas são: custo da solução a nível de *hardware* e *software*, infraestrutura adicional (SAN) e parametrização de redes de dados.

Dependendo da criticidade das aplicações, existem diferentes políticas de *backup* a aplicar conforme nível de serviço definido para cada aplicação diariamente, ou seja, com um RPO de 24 horas. Assim sendo existem três políticas distintas associadas aos *backups* aplicacionais, nomeadamente a total, cumulativo ou diferencial e a incremental.

O *backup* total é referente à totalidade de dados em disco de uma determinada aplicação, sendo o *backup* mais simples e mais rápido (RTO) em caso de recuperação de dados, no entanto, é o que consome mais recursos desde o processamento de todos os clientes para envio de informação ao servidor de *backups*, memória, impacto na largura de banda e maior tempo de *backup*. Caso a aplicação tenha uma política de *backups* que obrigue sempre ao *backup* total, em caso de perda de informação ou outro fator, é possível recuperar a última imagem dos dados do dia anterior num único passo.

O *backup* incremental é o *backup* com os dados alterados face ao último *backup* total ou incremental, tendo um RTO mais alto, pois a recuperação obriga á recuperação do total e restantes incrementais até à data pretendida. Por exemplo, caso uma aplicação tenha um *backup* total aos sábados e incrementais durante os dias úteis da semana e se acidentalmente ocorrer corrupção de informação ou erro humano que cause perda de dados numa quinta-feira, teria que se fazer recuperação do *backup* total de sábado e dos incrementais de domingo até quarta-feira para se ter a última imagem dados salvaguardada. Esta política tem menor impacto nos recursos e consequentemente nas aplicações, largura de banda e espaço em disco, no entanto, o RTO é o mais elevado. Uma vertente do incremental é o cumulativo ou diferencial que contém todas as alterações face ao último *backup* total, sendo os passos de recuperação reduzidos para apenas uma recuperação total e o último cumulativo ou diferencial.
O RTO é mais baixo que a política de incremental no entanto quanto mais perto se está do próximo *backup* total, maior as alterações de dados da aplicação, maior consumo de recursos no processamento de dados, armazenamento e maior impacto na largura de banda da rede.

Estas políticas apresentadas são dimensionadas em necessidades de negócio distintas para cada aplicação, sendo que se o *backup* for realizado numa primeira instância para disco a recuperação é um processo mais simplificado pois requer um passo para recuperar os dados dessa área de

disco para a aplicação, ao invés das fitas magnéticas que obrigam a dois passos, sendo estes a recuperação para uma área de disco e posteriormente para a aplicação.

Com o acentuado crescimento de informação nas organizações e de forma a flexibilizar e reduzir investimento foram introduzidas no mercado tecnologias de armazenamento de dados para *backups* baseadas em fitas magnéticas e disco. Diversas tecnologias baseadas em fitas magnéticas surgiram ao longo dos tempos, tais como, DLT, ADR e *Linear Tape Open* (LTO), sendo que a LTO ganhou mais expressão e adesão devido ao seu desenvolvimento tecnológico, estandardização na versão Ultrium e custo reduzido. Esta tecnologia teve diversas gerações desde o ano de 2000 sendo que novas gerações continuam em previsão para os próximos anos:

Tecnologia	Capacidade nativa [GB]	Taxa Transferência [MB/s]	Ano
LTO-1	100	20	2000
LTO-2	200	40	2003
LTO-3	400	80	2005
LTO-4	800	120	2007
LTO-5	1500	140	2010
LTO-6	3200	270	2011
LTO-7*	6400	315	2013
LTO-8*	12800	472	2015

Tabela 2 - Variantes da tecnologia LTO [16]
(*) previsto para os próximos anos e em desenvolvimento em laboratório

Os dispositivos baseados em fitas magnéticas possuem três componentes essenciais para o seu desempenho e dimensionamento para ir de encontro à performance necessária dos *backups* e ainda assegurar a retenção de dados exigida pela organização.

Na arquitetura de um dispositivo de armazenamento baseado em fita magnética temos um conjunto de componentes essenciais para gestão dos *backups*, sendo estes [15]: braço mecânico, *slots*, *drives*, controlador, etiquetas e respetivo leitor e o tipo de conetividade das *drives*. O controlador SCSI permite que um servidor envie pedidos de escrita/leitura para as fitas magnéticas e para receção de informação de configuração do dispositivo.
O braço mecânico é fundamental para a mobilidade das fitas magnéticas para as *drives,* para se poder escrever/ler informação das mesmas, sendo que a leitura e associação de dados a fitas é assegurada pelas etiquetas únicas a cada fita e um leitor ótico capaz de ler a etiqueta de modo a

otimizar a velocidade de criação ou atualização das fitas no inventário. Os *slots* são o espaço físico onde residem as fitas magnéticas, sendo que existem diversos equipamentos que podem suportar desde alguns *slots* a milhares. As *drives* são os dispositivos SCSI que são apresentados ao SO para se poder ler/escrever dados, podendo ter diversos tipos de conetividade associada, sendo as mais expressivas as ligações SCSI, iSCSI, SAS, FC, FCoE, FICON e ESCON, estas últimas associadas aos *mainframes*.

Apesar de estes dispositivos trazerem benefícios quanto a custo por TByte, flexibilidade de deslocalização, conectividade, compressão de dados, etc… os dispositivos baseados em disco apresentam flexibilidade superior, menor gestão e maiores taxas de transferência de dados. Quando se pretender recuperar informação de uma determinada fita magnética, o administrador tem de colocá-la no dispositivo, carregá-la manualmente, rebobinar, até chegar à informação pretendida e realizar a pesquisa dos ficheiros até chegar ao pretendido, ou seja, são processados vários ficheiros para se aceder ao que se pretende.

De forma a se otimizar o *backup*, mas acima de tudo a recuperação dos mesmos, as tecnologias baseadas em disco são vitais para os RTO mais exigentes, sendo que existem diversas formas de o realizar:

a) Disco local ao servidor de *backups* ou servidor de *proxy*,
b) Área de rede (NAS),
c) Disco apresentado via SAN FC ou iSCSI,
d) Dispositivo VTL,

Os dispositivos VTL baseados em disco, possuem uma camada de *software* capaz de emular fitas magnéticas físicas assim como *drives*, sendo apresentadas através de tecnologia (i)SCSI, SAS ou FC aos servidores de *backup* ou *proxy* para se escreverem os dados, ou seja, quer o *software* de gestão de *backups*, quer o SO vêem uma drive física, no entanto, esta é virtualizada e apresentada por uma HBA (no caso de FC) tendo um *initiator* por cada dispositivo apresentado.
As principais vantagens destes dispositivos é permitir criar *slots* e *drives*, recorrer a discos de alta capacidade e custo reduzido como os discos SATA, capacidade de exportar fitas virtuais para fitas físicas e ter utilização eficiente de espaço em disco recorrendo a tecnologias de deduplicação de informação. Outra característica destes sistemas é o facto de se aplicar a tecnologia de RAID para prevenir perda de dados enquanto numa fita magnética um erro mecânico pode originar corrupção de dados.

Na figura 24 podemos verificar a flexibilidade de um dispositivo de armazenamento baseado em VTL, sendo que neste exemplo cada servidor *proxy* faz *backups* diretos via SAN para o dispositivo VTL. Esta configuração implica parametrização a nível da SAN para definição de *zonning* para cada cliente, sendo que havendo caminhos redundantes deve ser efetuado para efeitos de alta disponibilidade e balanceamento de carga.

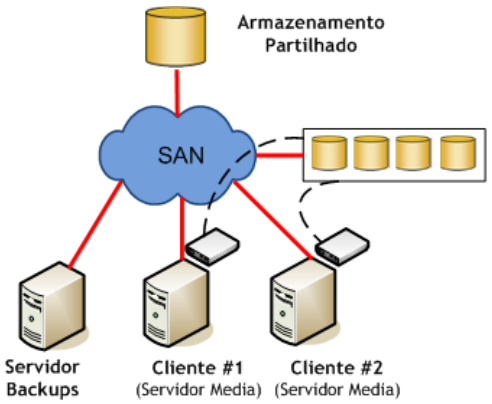


Figura 24 - Dispositivo de armazenamento baseado em disco VTL

Uma grande vantagem desta configuração é o impacto mínimo na rede LAN pois apenas são transmitidos metadados sobre a informação a fazer *backup*, assim como a flexibilidade na apresentação de dispositivos aos clientes, ou seja, caso seja necessário no futuro adicionar um terceiro cliente, basta criar uma drive apresentá-la ao servidor através do *zonning* e criar *slots* conforme a capacidade pretendida. Conjugando esta solução que permite recuperações mais rápidas de informação, é possível fazer a exportação da fita magnética virtual para física, para um dispositivo físico ligado à SAN.

A compressão é uma técnica que permite procurar bits redundantes de modo a eliminá-los e reduzir a utilização de espaço baseados numa redundância estatística, a deduplicação analisa os dados de forma mais inteligente e com diferentes patamares de granularidade de modo a procurar ficheiros ou blocos redundantes e guardar apenas uma única instância. A compressão de dados pode ser feita por *hardware* ou *software* consumindo recursos de processamento, memória para verificar os bits redundantes, no entanto, caso existam dois ficheiros semelhantes ambos serão guardados com a aplicação do algoritmo de compressão. Tipicamente os algoritmos de compressão conseguem atingir taxas de compressão entre os 1:2 até 1:3 [18], sendo que o rácio de compressão depende do tipo de informação. Documentos como folhas de cálculo,

apresentações, texto apresentam taxas altas enquanto imagens, vídeo, ficheiros encriptados não representam quaisquer taxas vantajosas de compressão.

Quando é aplicada a deduplicação de dados temos vantagens consideráveis [18] visto que determinados processos de *backup* envolvem taxas de retenção de dados e mesmo que um dado ficheiro não tenha ganhos de redução de espaço, como já existe uma instância, pode ser referenciado por método de apontadores e consegue-se ter ganhos a longo termo. As tecnologias de deduplicação proporcionam outros benefícios face aos algoritmos de compressão devido à forma como analisam as semelhanças entre pedaços de informação, tendo granularidade ao nível do ficheiro, bloco e bloco variável. A deduplicação ao nível de ficheiro permite que apenas seja guardada uma instância duplicada em disco, mesmo que tenha metadados distintos desde que o corpo do ficheiro seja igual (as restantes serão substituídos por apontadores). Apenas a deduplicação ao nível de blocos tem a inteligência de perceber que pedaços de informação poderão ser novos e salvaguardar apenas as diferenças em disco baseado em técnicas de *hashing*.

Figura 25 - Deduplicação em um *stream* de dados [17]

De acordo com a figura 25 temos cerca de 334 blocos de dados com padrões semelhantes identificados pelas cores associadas, sendo que o *stream* será reconstruído através de um conjunto de apontadores para estes blocos. Enquanto numa fase inicial tínhamos 334 blocos, passamos a ter 6 padrões com um total de 21 blocos mais os apontadores para referenciar o *stream* de dados.

Se assumirmos que cada bloco tem 4 KBytes de dimensão e cada apontador tem dimensão de 20 Bytes, tínhamos espaço inicial necessário de 334*4KBytes, totalizando 1336 KBytes, sendo que com a aplicação da deduplicação poderíamos alcançar 21*4Bytes, ou seja, 84 KBytes, somando 100 apontadores de 20 Bytes, finalizando em 86 KBytes.
Com este tipo de algoritmo podemos ter um rácio de deduplicação de 1:15,5, ou seja, por cada KByte no sistema são salvaguardados 15,5 KBytes. Sempre que a um dado ficheiro seja realizado o *backup*, o sistema de deduplicação partirá o ficheiro em blocos e calculará o seu *hash*

comparando no seu índice local se este pedaço de informação já existe e caso este exista apenas será atribuído à referência do ficheiro um apontador, caso seja um bloco novo será adicionado ao índice com o respetivo *hash* para referenciar o ficheiro.

Com a desmaterialização do servidor e com a consolidação de servidores virtuais em físicos, existe uma alta probabilidade de redundância de blocos de dados relativamente às imagens dos servidores virtuais, como por exemplo, ambiente Microsoft Windows no qual vários blocos de dados do SO são semelhantes entre servidores virtuais. O *hyper-visor* tem a capacidade de monitorizar os blocos de dados dos servidores virtuais assim como os blocos alterados *Change Block Tracking* (CBT) fazendo com que integre com *softwares* de *backup* para realização de *backups* aos servidores virtuais.
Como o CBT permite saber quais os blocos alterados de informação [19], este notifica o *software* de *backups* para realizar o *scan* de dados apenas aos blocos alterados em vez de realizar o *scan* à imagem integral poupando desta forma recursos de processamento, memória, rede e consequentemente armazenamento conjugado com um *backup* de dados baseado em *snapshot.*

2.2.6. Segurança

Para dimensionar uma infraestrutura em *cloud* tendo em conta a segurança, é fundamental validar as redes internas, assim como os isolamentos e serviços para a rede externa de modo a enquadrar possíveis ataques quer de fonte interna como externa. Tradicionalmente a comunicação com redes externas implica uma maior atenção sobre os aspetos de segurança para uma sub-rede isolada denominada DMZ que permitirá acessos externos à organização para interagir com serviços, nomeadamente, serviços Web, VPN, FTP, entre outros.

A segurança no centro de dados num modelo físico é implementada tipicamente através de *appliances* em localizações estratégicas sendo que numa abordagem em *cloud* com a tecnologia de virtualização surgem desafios, nomeadamente, o *routing* de tráfego que flui dos servidores virtuais feito por *appliances* físicas pelo que a arquitetura virtualizada deve ser concebida para todas as ligações de rede do servidor físico passarem por uma quantidade reduzida de dispositivos. Em termos de capacidade as *appliances* físicas são fixas, não sendo o ideal [20] para um ambiente virtualizado que escala em termos dinâmicos, não permitem a visibilidade entre

servidores virtuais nem possuem a capacidade de resposta em ambientes dinâmicos, contrariamente ao modelo físico com redes estáticas.

Tipicamente as DMZ processam pedidos da internet e iniciam ligações para serviços de *back-end* (noutra sub-rede ou segmento de uma DMZ) sendo que, ao mesmo tempo não é suposto que os serviços falem entre si ou iniciem ligações externas.

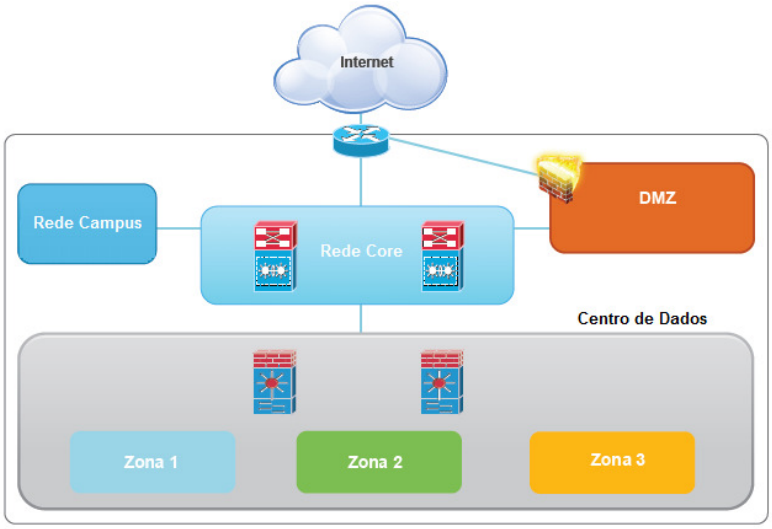


Figura 26 - Redes numa organização [20]

No centro de dados a segurança nos acessos enquadra problemáticas como as sub-redes DMZ e a sua ligação a redes externas ou públicas, no qual, é necessário ter segurança e proporcionar serviços de DHCP, VPN, NAT e balanceadores (exemplo: aplicações *Web*).

Com a introdução da virtualização no modelo do *cloud computing*, a segurança estática será endereçada por um conjunto de soluções dinâmicas [21] que sejam compatíveis com o novo paradigma de redes virtualizadas (que não são restringidas pela localização física) e com maior capacidade de adaptação à mudança, no entanto, capazes de escalar à medida que mais aplicações são provisionadas e com capacidade de resposta face às alterações de configurações.

O *switch* virtualizado distribuído permite ter um único *switch* lógico distribuído por todos os *hyper-visors* de modo a reduzir gestão e acompanhar o dinamismo no *cloud computing*, pois qualquer configuração elaborada num *hyper-visor* é replicada a todos os presentes no *cluster*. A rede segura e/ou isolada será a sub-rede que possuirá os serviços externos para a rede pública proporcionando desta forma uma camada adicional de segurança visto que será o único ponto na rede no qual um atacante terá acesso. Os requisitos desta sub-rede passam por ter balanceamento

de tráfego entre servidores *web*, acesso à internet para *download* de *patches*, acesso SSH remoto (manutenção remota), túneis VPN, isolamento completo do resto das redes para que em caso de ataque não se espalhe às restantes redes, regras de *firewall* aplicado ao tráfego que entra e sai da sub-rede e entre servidores virtuais e ainda o endereçamento automático na DMZ incluindo IPs fixos para servidores específicos.

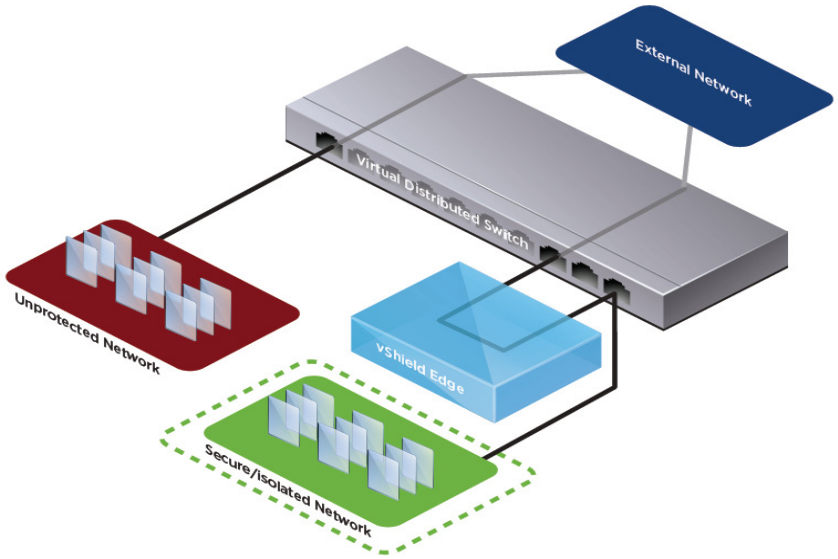


Figura 27 - Arquitetura de segurança [21]

Como tal deverá ter a capacidade de permitir aos administradores monitorizar e controlar redes de servidores virtuais com *logging* compreensivo de todos os eventos de segurança ao nível do centro de dados virtualizado e ter a capacidade de perceção da mudança, como por exemplo, um servidor virtual ser migrado entre servidores físicos assegurando que as políticas e configurações são igualmente migradas. As necessidades a nível de segurança no centro de dados são as mesmas que num modelo físico tradicional sendo uma camada virtual que reside no servidor físico: a aplicação de regras de *firewall* baseado em parâmetros tais como endereço IP, portos e protocolo, recorrer a NAT para tradução de e para ambiente virtual assim como mascarar IPs do centro de dados virtualizado para localizações que não tenham relação de confiança, definição de regras de DHCP para atribuição de IPs aos servidores virtuais tais como dedicação de IPs fixos, conjuntos de IPs, DHCP *lease-time* e comunicações seguras VPN.

Em situações de falha de servidor físico em que reside a *appliance* virtual de segurança, a funcionalidade de alta disponibilidade irá reiniciar a *appliance* num servidor físico disponível no *cluster* desde que existam recursos disponíveis de modo a assegurar a proteção do ambiente. Se o servidor virtual tiver falhas a nível de SO, o *heartbeat* de monitorização deteta e reinicia

automaticamente sendo que o tráfego que flui dos servidores virtuais protegidos será interrompido até que a *appliance* virtual termine de reiniciar. As *appliances* virtuais podem ser definidas para proteção ao nível do centro de dados, do *cluster* e conjunto de recursos no centro de dados, como por exemplo, armazenamento, rede e servidores virtuais.

Tal como se se tratasse de um ambiente físico os servidores que estão numa VLAN podem ser acedidos diretamente por outro na mesma, o que compromete a segurança do ambiente, podendo o ataque ser despoletado do servidor comprometido ou do domínio de *broadcast* através de técnicas de ataques via aplicacional, *man-in-the-middle* e ao nível de rede (captura de pacotes).
É possível ainda restringir comunicações através de *Private* VLANs (PVLANs) para isolamento de servidores na mesma VLAN, havendo dois tipos distintos: a primária e a secundária, sendo a primária a VLAN de acesso e a mapeada na infraestrutura e a secundária a VLAN apenas conhecida pelo *switch* físico ou virtual no qual está configurada. Cada VLAN secundária está mapeada com uma VLAN primária, múltiplas VLANs secundárias podem ser associadas a uma única primária e cada VLAN primária possui uma única sub-rede sendo possível associar várias secundárias à mesma sub-rede.

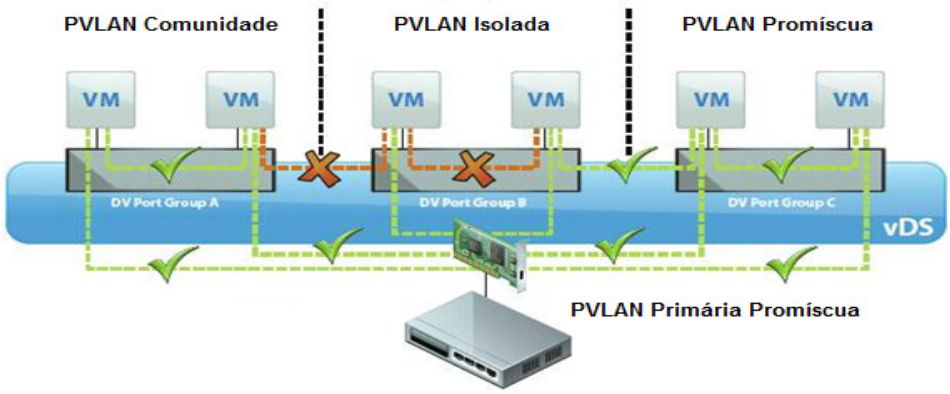


Figura 28 - Tipologia de portas com PVLANs [22]

Existem três tipos de portas disponíveis para configuração de PVLANs [22] para permitir isolamento na segunda camada e no mesmo domínio de *broadcast*:

a) agregação para acesso de e para VLANs secundárias (promíscuo) sendo implementado tipicamente para *uplinks* que transportam a PVLAN primária,
b) porta isolada para comunicação apenas entre promíscuo e não poder comunicar com portas isolados no *switch* sendo caso típico o isolamento entre servidores virtuais numa infraestrutura em *cloud,*

c) comunidade que pode comunicar entre comunidades e promíscuas ideais para serviços em *clustering* e comunicação direta entre servidores virtuais.

De acordo com a figura 28 encontram-se as possíveis configurações descritas sendo que todos os servidores virtuais estão na mesma sub-rede e possuem o *uplink* na PVLAN primária promíscua que liga o *switch* virtual distribuído no *cluster* á interface física de rede que por si liga ao *switch* físico.

As *Virtual Access-Lists* (VACL) são igualmente outro método de segurança que permite filtrar acessos baseado em porto, protocolo e endereço e possibilita outra granularidade face a outras soluções de segurança distintas no centro de dados. As VACL conjugadas com as PVLAN permitem endereçar problemáticas na área da segurança assim como se complementam para minimizar diversos tipos de ataque possíveis à rede.

A forma como se complementam prende-se com o fato das PVLANs endereçarem somente a segunda camada, ou seja, se um atacante realizar ataque noutra camada, tal como a de rede, como por exemplo ter acesso a um *router* pode injetar facilmente tráfego na rede ultrapassando a segurança das PVLANs. Como tal as VACL vão permitir endereçar outras camadas para restringir este tipo de situações.

3. Análise de infraestrutura

Enquadrando o caso de estudo para desenho de uma infraestrutura dinâmica, tolerante a falhas e com capacidade de computação, rede, armazenamento na *cloud* privada foi elaborada uma análise á infraestrutura recorrendo ao *software* VMWare Capacity Planner [23] que permite obter dados estatísticos da infraestrutura para a consolidação dos recursos num ambiente virtualizado. Foi tida ainda em consideração através de outra análise (NetApp Auto-Support [24]) para avaliar as métricas de acessos a disco relativos ao serviço de *mail* do caso de estudo.

O VMWare Capacity Planner permite obter dados de coleção de infraestrutura, ou seja, modelos físicos, configurações de *hardware* (memória, rede, processamento e armazenamento) e informação de performance com dados médios e picos de utilização. A arquitetura desta aplicação ilustra-se na figura 29:

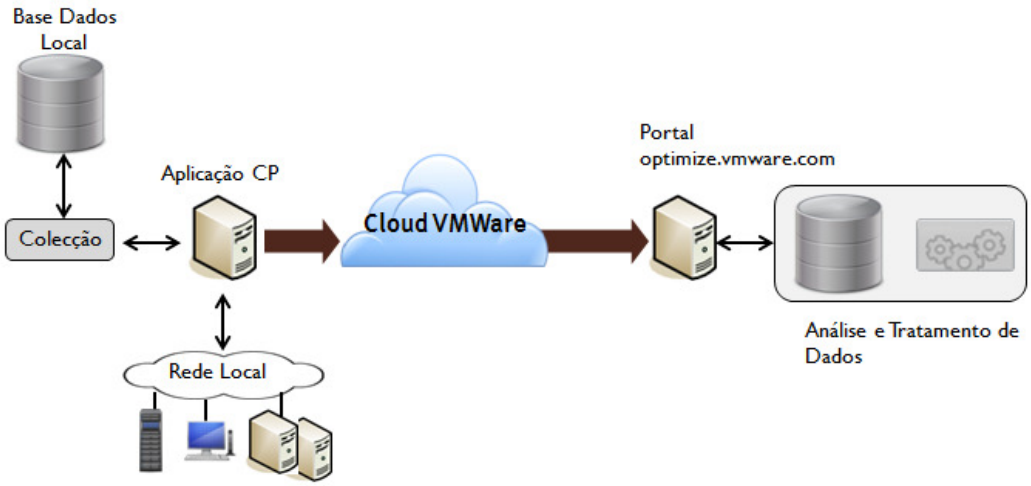


Figura 29 - Arquitetura do VMWare Capacity Planner

A recolha de dados é baseado em processos de SO, ou seja, a aplicação instalada numa máquina virtual Windows 2008 na rede em análise, recolhe dados somente estatísticos (tarefa de colecção) através de processos nativos dos clientes (perfmon, WMI em Windows e comandos em Linux como iostat, vmstat, etc..) que são invocados de dez em dez minutos tendo impacto mínimo na rede com o armazenamento dos dados numa base de dados embebida na aplicação. Posteriormente ao sincronismo para a *cloud* a cada 24 horas, existe o acesso *Web* ao Portal onde é permitido analisar os dados e gerar relatórios de consolidação de infraestrutura assim como tabelas com estatísticas de utilização. A análise efetuada decorreu entre Setembro de 2011 a Maio de 2012 com uma janela temporal de 8 meses sensivelmente.

3.1. Levantamento de infraestrutura

Atualmente a entidade do caso em estudo, enquadra-se na área de negócio no ramo da educação sendo sustentada por uma infraestrutura computacional na qual assentam um conjunto de aplicações distintas. A infraestrutura existente no centro de dados é composta por:

a) 21 Servidores físicos enquadrados na análise com configurações distintas,
b) Solução de *mail* suportada por um *cluster* ativo/passivo e armazenamento de dados totalmente redundante NetApp FAS2020 com acesso iSCSI,
c) Armazenamento SUN StorEdge iSCSI para algumas aplicações,
d) Maioritariamente os servidores possuem armazenamento interno (DAS),
e) Infraestrutura de rede com *switches* de rede a 1 Gb/s,
f) Maioritariamente ambiente com SO Windows no entanto existem algumas versões de Linux,
g) *Backups* efetuados através de *scripts* e *snapshots* de dados em algumas aplicações,
h) 2 bastidores com equipamentos a consumir energia, espaço e ar condicionado,
i) Índice reduzido de virtualização na infraestrutura com algumas aplicações menos críticas representando apenas 4 serviços;

A organização em estudo possui um conjunto de servidores físicos que poderão ser consolidados através da virtualização de servidores, sendo que grande parte destes servidores usam armazenamento DAS para as aplicações. Muitos servidores no mercado são limitados quanto a crescimento suportando um número limitado de discos sem qualquer otimização de I/O, sendo que a gestão do espaço atualmente é feita servidor a servidor. Caso a organização queira implementar uma solução de DR para um centro de dados secundário ou mesmo a continuidade de negócio com replicação de dados, sem armazenamento partilhado teria que recorrer a ferramentas aplicacionais que consumiriam processamento dos servidores existentes podendo impactar os recursos de um servidor crítico.

A gestão seria ineficaz assim como a adição de servidores e respetivas aplicações implicaria gestão adicional de configuração de replicação de dados. A adição de servidores acarretaria adjudicar um equipamento, obter licenciamento, efetuar as devidas manutenções, tempo de entrega, tornando-se menos eficiente caso seja necessário ter em pouco tempo uma aplicação disponível.

Outra componente importante prende-se com *backups* e recuperação dos dados, sendo que o crescimento de informação traduz-se num crescimento ainda mais acentuado no volume de dados de *backup* pois existem políticas, períodos de retenção de dados que se traduzem numa quantidade de dados bastante elevada e janelas de *backup* a cumprir para não impactar a aplicação em período laboral.

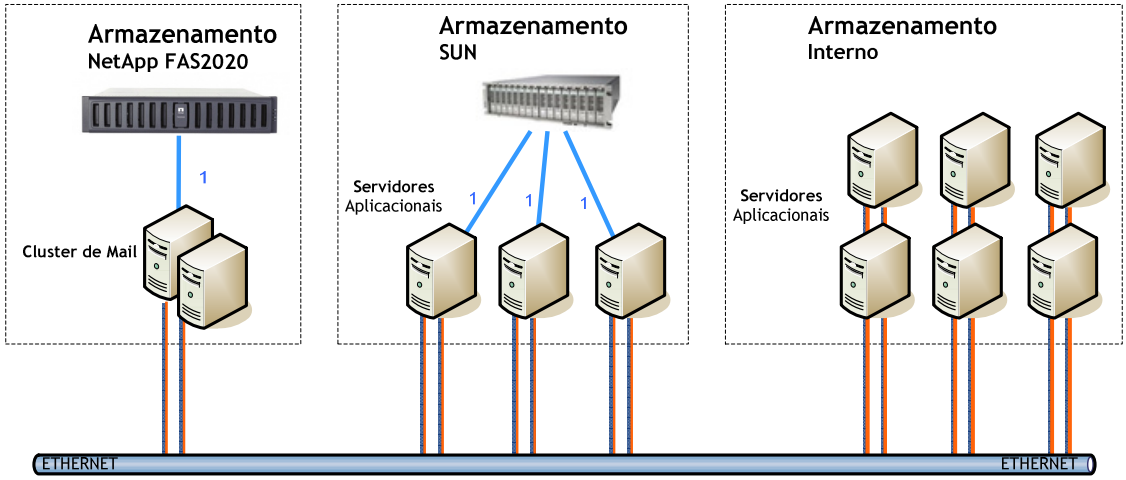


Figura 30 - Arquitetura da infraestrutura do caso em estudo

A atual configuração contém o serviço de *mail* composto por um *cluster* ativo/passivo de 2 nós, ligados diretamente via iSCSI a um controlador de um sistema NetApp Fas2020, sendo que como a ligação é directa apenas existe possibilidade de *failover* no caso de avaria de um nó ou um NIC, não havendo possibilidade de balanceamento de dados, ou seja, a largura de banda entre serviço de *mail* e armazenamento é de 1 Gb/s.

O sistema SUN disponibiliza armazenamento a alguns servidores aplicacionais e de bases de dados via iSCSI ligado a um *switch* de rede e ainda os restantes servidores com armazenamento DAS.

O levantamento feito na infraestrutura da organização para enquadrar no modelo de *cloud* privada a propor é demonstrado na tabela 3 com a capacidade de memória, processamento, rede e armazenamento.

Na tabela 3 consta todo o levantamento, sendo que apenas foi possível a monitorização no período descrito de 16 servidores físicos.

	RAM (GB)	CPU		Portas de rede	Quantidade Disco DAS	Capacidade Bruta [GBytes]
		GHz	Cores			
Servidor1	14,336	2,833	2	2x1 Gb	2	881
Servidor2	28,672	2,833	4	2x1 Gb	3	734
Servidor3	8,704	1,867	4	2x1 Gb	2	734
Servidor4	8,192	2,833	4	2x1 Gb	2	734
Servidor5	8,192	2,822	2	2x1 Gb	3	734
Servidor6	4,096	1,867	2	1x1 Gb	1	160
Servidor7	12,288	2,833	2	2x1 Gb	1	147
Servidor8	20,48	1,867	4	1x1 Gb	1	147
Servidor9	15,104	2,833	8	2x1 Gb	2	292
Servidor10	8,192	3	1	2x1 Gb	2	292
Servidor11	1,024	1,86	2	1x100 Mb + 1x1 Gb	2	320
Servidor12	2,048	1,86	2	1x1 Gb	1	160
Servidor13	7,935	3,056	2	1 x 1 Gb	2	184
Servidor14	1,536	3	1	2x100 Mb	2	122
Servidor15	4,096	3	1	2 x 1 Gb	1	292
Servidor16	4,096	3	1	2 x 1 Gb	2	292
Servidor17	4,096	3	1	2 x 1 Gb	2	292
Servidor18	4,786	1,83	1	2 x 1 Gb	2	292
Servidor19	1,024	2,86	2	2 x 1 Gb	2	292
Servidor20	8,192	3	1	2 x 1 Gb	2	292
Servidor21	8,192	3	1	2 x 1 Gb	2	292
TOTAL	**175,281**	**123,076**	**48**	**35x1Gb+3x100Mb**	**39**	**7685**

Tabela 3 - Levantamento de infraestrutura

Na tabela destaca-se o servidor2 com os recursos do serviço de *mail* mais exigente e a capacidade de armazenamento apresentada é relativa ao armazenamento DAS e não via SAN.

A maior parte dos sistemas foram monitorizados em termos de utilização para aplicação da sua migração para um modelo em *cloud* privada.

A configuração do sistema de armazenamento do serviço de *mail* encontra-se configurado de acordo com as melhores práticas do fabricante e balanceando o desempenho pelos discos e portas disponíveis de acordo com a figura 31 [25].

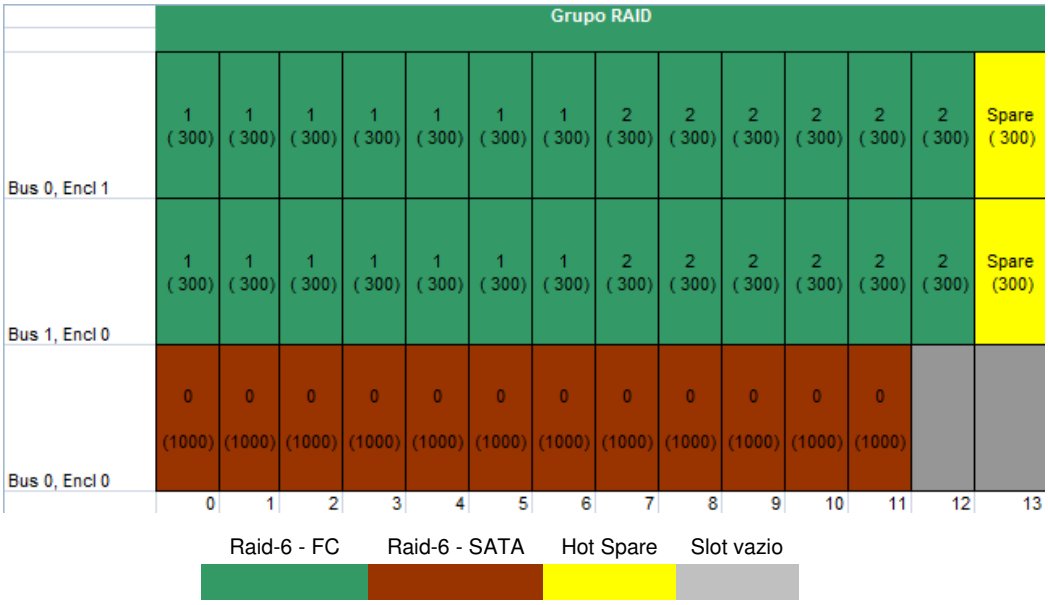


Figura 31 - Arquitetura dos grupos RAID

Presentemente existem 28 discos FC de 300 GB a 15Krpm configurados com dois Raid6, o grupo identificado como 1 com raid6(12+2), ou seja, 12 para dados e 2 de paridade e o grupo identificado como 2 com raid6(10+2), ou seja, 10 para dados e 2 de paridade sendo que ambos têm a proteção adicional de 2 discos para funções de *spare* e ainda 12 discos em Raid6(10+2) com discos SATA. Os discos são endereçados pelos canais de comunicação internos SAS (Bus) e posicionados nas gavetas de discos ou *enclosures* (Encl).

Como podemos verificar as gavetas estão balanceadas pelas portas de *backend* do armazenamento assim como os RAID para garantir a máxima disponibilidade e largura de banda de *backend* para situações de reconstrução de dados em situação de falha de disco.

3.2. Análise de processamento e memória

Foram analisados 16 servidores num universo de 21 (representando 76% do universo e com maior fatia de consumo de recursos no centro de dados), sendo os resultados de utilização de percentagem de CPU na janela temporal da análise o presente na figura 32:

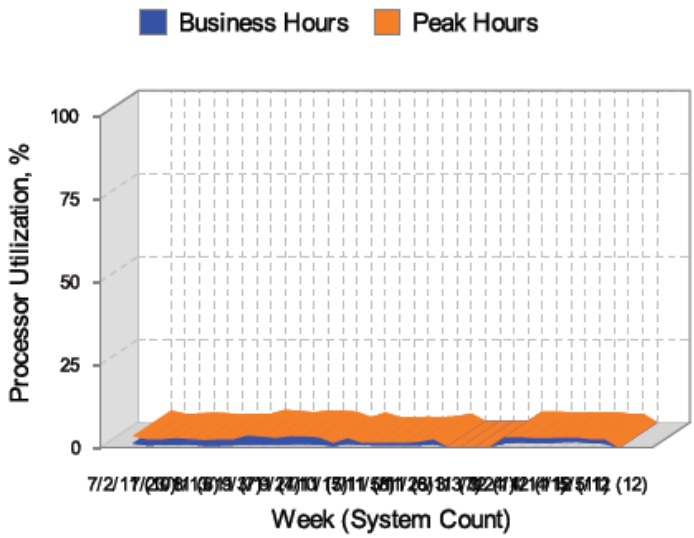


Figura 32 - Taxa de utilização de computação

A figura 32 ilustra a taxa de ocupação dos recursos de computação na infraestrutura sendo a taxa média inferior a 2% de utilização com um valor de pico inferior a 8% em toda a janela temporal de análise. Pode-se aferir que da capacidade total de computação do centro de dados do caso em estudo que dos 127 GHz totais nos diversos *cores* existentes apenas são consumidos em valor médio 2,54 GHz sendo o valor de pico de 10 GHz. Com os valores reais de consumo pode-se dimensionar um servidor para a carga sendo que os valores apresentados demonstram que poucos servidores conseguem gerir toda a carga numa perspetiva de processamento. Foram considerados dois servidores para formar o *cluster* de virtualização com dois CPUs de 4 *cores*, ou seja, 8 *cores* por servidor num total de computação de 45,76 GHz sendo largamente o necessário.

A escolha deve-se ao facto de hoje em dia a maior parte dos servidores possuírem estas características mínimas com elevada capacidade de computação associada e o consumo estimado ser baixo. Este recurso tem a capacidade de escalar em mais serviços e encontra-se ainda com margem para os 50%, ou seja, em termos de processamento em caso de falha do primeiro nó do *cluster* o segundo consegue gerir todo o processamento ficando ainda nos 22% da sua capacidade

total. A análise demonstrou ainda que nenhum dos serviços apresenta *queues* elevadas em CPU, ou seja, nenhuma aplicação em estudo tem pontos de contenção no processamento. A memória de um sistema tem um papel fundamental no desempenho das aplicações e utilizadores que suporta e o seu dimensionamento adequado implica menos I/O em disco e menor tráfego na rede de dados, operações de rede que ocorrem dentro do *hyper-visor* e ainda ajustar a carga e respetivas margens para salvaguardar falhas de outros nós.

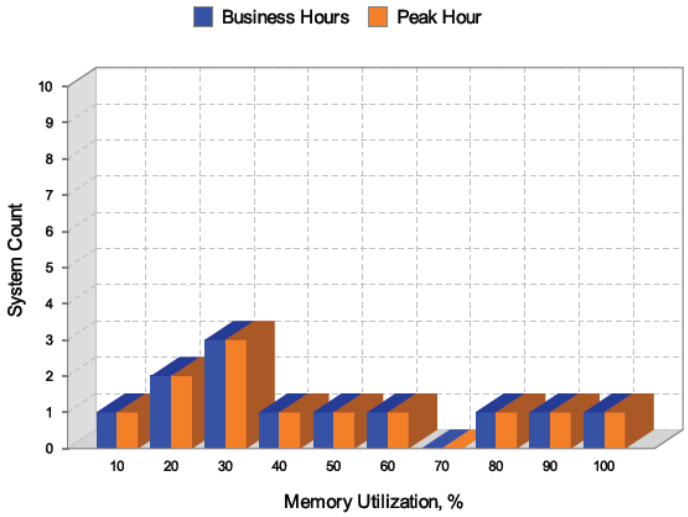


Figura 33 - Memória utilizada por sistemas

A figura 33 é ilustrativa da distribuição de utilização por sistemas sendo que os sistemas que apresentam taxas elevadas de consumo de memória, são maioritariamente sistemas que possuem pouca memória para a carga que têm e mais uma vez o serviço de *mail* que consome em média 90% da memória disponível.

O centro de dados possui 175 GBytes de memória total sendo o consumo total de aproximadamente 37% deste valor, sendo a aplicação de *mail* a que se destaca no centro de dados recorrendo a 55% de toda a memória consumida. A aplicação de *mail* possui servidores de *front-end* para ligação e de *back-end* (base dados) que possui os 28 GB de memória.

A memória disponível é bastante mais do que o real consumido, no entanto, algumas aplicações encontram-se limitadas por terem memória com níveis de utilização elevados o que pode futuramente condicionar o desempenho da infraestrutura.

A abordagem em *cloud* e de recursos partilhados vai permitir racionar a memória disponível assim como aplicar técnicas de eficiência como a deduplicação de memória para permitir reduzir espaço e dar maior escalabilidade à infraestrutura [26].

Os servidores que não foram possíveis de analisar foram enquadrados na análise com extrapolação do rácio do consumido e alocado do tipo aplicacional dos serviços analisados.

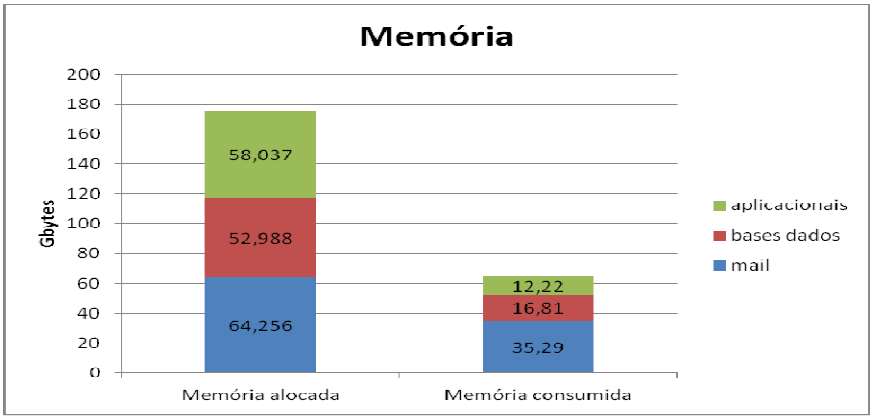


Figura 34 - Consumo de memória por tipo aplicacional

A análise efectuada demonstra ainda uma utilização constante da cache na janela temporal em termos da sua capacidade utilizada pelas aplicações não havendo picos consideráveis devido às aplicações que maior consumo fazem estarem com taxas de utilização já bastante elevadas.

3.3. Análise de armazenamento

O armazenamento numa *cloud* privada deve ser desenhado quer em capacidade como em desempenho, para tal, valida-se o desempenho do sistema de armazenamento com o serviço que possui maior carga, o serviço de *mail*.

A estrutura de *mail* contém um conjunto de LUNs que no período decorrente da análise (análise armazenamento NetApp durante uma semana) produz de pico aproximadamente 4000 I/O agregado sendo um valor despoletado no início de cada dia entre as 00h e 03h da manhã devido a processos de *backup* de dados, sendo que a média ronda os 555 I/O.

Para se ter tempo de resposta aceitável, deve-se ter em conta a quantificação de disco, tecnologia e valor de referência médio inferior a 15 milisegundos para a aplicação obter o melhor desempenho possível [3].

O sistema encontra-se abaixo da referência média inclusive de picos, havendo um pico de 15 milisegundos nas leituras durante a noite relativamente ao *snapshot* que está programado.

É notório que o sistema realiza mais leituras do que escritas e na hora do *snapshot* o tempo de resposta nas escritas aumenta devido à funcionalidade ativa que se trata de esmagar o último *snapshot* para o novo que fica no sistema, sendo que, o tempo está sempre abaixo da referência e o desempenho é aceitável.

Figura 35 - Desempenho do sistema de armazenamento

O serviço de *mail* representa o serviço mais exigente podendo-se validar que se trata de um ambiente de maior índice de transações do que largura de banda [3] (com I/O inferior a 64 KBytes) através do pedido de blocos pequenos, como é o caso da aplicação:

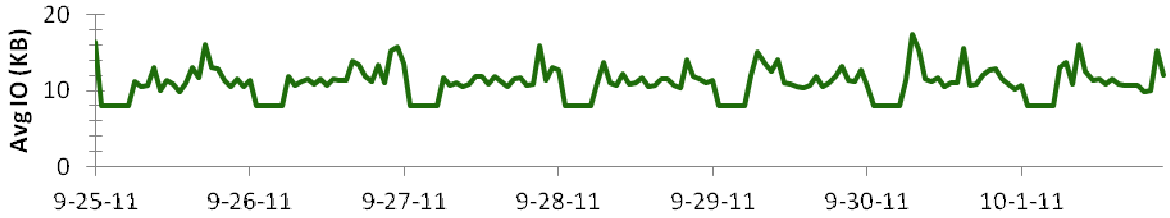


Figura 36 - Dimensão do tamanho do bloco

A arquitetura aplicacional recorre a blocos de 8 KBytes, sendo que a figura 36 comprova esta tendência, no entanto o fato do tamanho do bloco ser variável no tempo deve-se ao fator desalinhamento no qual um bloco pode ter que ser dividido no armazenamento ocupando em vez

de um, dois ou mais blocos. Neste caso em estudo apenas se valida para a janela temporal que na pior das situações um bloco pode ocupar dois (16 KBytes).

Relativamente ao armazenamento verificaram-se os seguintes resultados decorrentes do período de análise com valores de capacidade alocada, taxa de ocupação, crescimento e função aplicacional de acordo com a tabela 4.

Servidor	Capacidade alocada [GB]	Set-11	Mai-12	Ocupação[%] Mai-12	Crescimento [%] 8m	Estrapulado Anual [%]	Funcao
Servidor1	880	141,1	141,8	16	0,5	0,75	Domínio
Servidor2	3000	889,8	1102,2	29,7	23,9	35,85	*Mail*
Servidor3	734	590,3	590,3	80,4	0	0	Base Dados
Servidor4	734	164,9	165,6	22,5	0,4	0,6	Base Dados
Servidor5	734	56,8	62,1	7,7	9,3	13,95	Domínio
Servidor6	160	56,2	58,7	35,1	4,4	6,6	Aplicacional
Servidor7	146	15,3	17,3	10,5	13,1	19,65	Domínio
Servidor8	146	101	117	69,2	15,8	23,7	*Mail*
Servidor9	136	40	47,2	29,4	18	27	*Mail*
Servidor10	136	11,1	11,1	8,2	0	0	Base Dados
Servidor11	149	6,5	6,5	4,4	0	0	Domínio
Servidor12	160	9,3	9,3	5,8	0	0	Domínio
Servidor13	292	204,4	236,3	70	15,6	23,4	Aplicacional
Servidor14	183	21	24,7	11,5	17,6	26,4	Base Dados
Servidor15	115	13,5	13,5	11,7	0	0	Domínio
Servidor16	63,99	15,9	18,1	24,8	13,8	20,7	Aplicacional
Total	7768,99	2337,1	2621,7	-	12,1	18,3	-

Tabela 4 - Detalhes do armazenamento

Dados resultantes da análise presentes na tabela 5 contêm variáveis importantes para o dimensionamento da nova infraestrutura, nomeadamente a taxa de crescimento anual em termos de espaço assim como as aplicações que maior crescimento têm e que necessitam de maior atenção. O serviço de *mail* apresenta a maior taxa de crescimento anual (servidor2) na ordem dos 35%, englobado num universo total de um crescimento anual de 18,3%.

Outro aspeto relevante do armazenamento é o facto de existir um total de capacidade alocada na ordem dos 7,8 TBytes sendo que á data da 2ª amostragem apenas 33% desta capacidade é a real consumida.

3.4. Recursos do centro de dados

A capacidade total do centro de dados do caso em estudo é a seguinte:

a) Processamento: 123 GHz (através de 48 *cores*),
b) Memória: 175 GB,
c) Capacidade de rede (NIC): 35,3 Gbit/s (agregado de portas em servidores),
d) Espaço em bastidor 52U (dois bastidores):
    i. 21 x 2 U em servidores,
    ii. 8 U sistema NetApp,
    iii. 2 U sistema SUN StorEdge,
e) Armazenamento:
    i. 39 discos locais (imagem de SO) configurados em Raid1 e Raid0 com capacidade bruta sem proteção de 7,6 TB,
    ii. 52 discos em armazenamento partilhado disponibilizando a capacidade bruta de 24,2 TB,
    iii. Após aplicação de raid, *spares*, temos capacidade líquida de 3,35 TB em armazenamento interno e 19,8 TB em armazenamento partilhado,
f) Consumo energético de sistemas 24,54 kVA:
    i. 1,04 kVA x 21 (servidores),
    ii. 1,7 kVA (NetApp),
    iii. 1 kVA (SUN),

3.5. Conclusões da análise

Relativamente ao modelo atual com a arquitetura ilustrada existem os seguintes desafios que implicam gestão adicional, quebra de serviço e risco associado ao negócio:

a) Gestão de servidores, *patches*, *firmwares* nos quais são necessárias muitas vezes reiniciar os SO o que implica a existência de quebra de serviço e janelas de tempo de operação muitas vezes fora de horas. Esta gestão é igualmente feita máquina a máquina havendo configurações semelhantes feitas vezes sem conta e ocupando um maior tempo de operação,
b) O lançamento de um novo serviço com uma aplicação específica, implica a existência de um novo servidor (que em caso de aquisição implica o tempo de

entrega de material que poderá ser de dias a semanas), licenciamento ou não de SO (dependendo do SO),

c) Em caso de falha de um servidor físico, existe quebra no serviço (exceção é apenas o serviço de *mail*). Um serviço poderá ter uma quebra de horas a dias e ficará dependente da manutenção associada para troca de peças, não existindo técnicas de alta disponibilidade,
d) *Backups* efetuados através de *scripts* implicam que cada modificação na aplicação leva a reescrever o *script* de *backups* que está sujeito a erro humano e consequentes falhas para salvaguarda dos dados. A reposição é um processo moroso, complexo, tem impacto na rede e é feito por várias consolas de gestão,
e) *Backups* através de *snapshots* do servidor de *mail*, são guardados no *filesystem* de produção (sistema NetApp) que no caso de quebra do RAID ou sistema de armazenamento, incorre no risco de perder os *backups* e consequentemente a informação,
f) Migrações de dados e aplicações em ambientes de DAS,
g) Flexibilidade limitada em termos de operação do armazenamento (exemplo: migração de volumes internamente para tipo de disco diferente, habilidade de *snapshots* para testes, escalabilidade, replicação de dados, entre outros),
h) Eficiência energética dos sistemas em questão,
i) Custo de manutenção e operação associado a todas as plataformas descritas,
j) Solução limitada em termos de escalabilidade e desempenho para a maioria das aplicações,

4. Proposta de arquitetura

A abordagem aplicada à problemática identificada é baseada num modelo de *cloud* privada, ou seja, infraestrutura dedicada nas instalações do caso em estudo assim como suportar os seus serviços fornecidos como IaaS, no entanto, para uma entidade apenas, com disponibilização dos recursos totais de infraestrutura de modo a renovar a existente. Para que tal seja exequível, será necessária a conversão *Physical-to-Virtual* (P2V) [27] dos servidores, ou seja, a desmaterialização dos servidores físicos para um conjunto de ficheiros assim como a migração de dados para o modelo proposto.

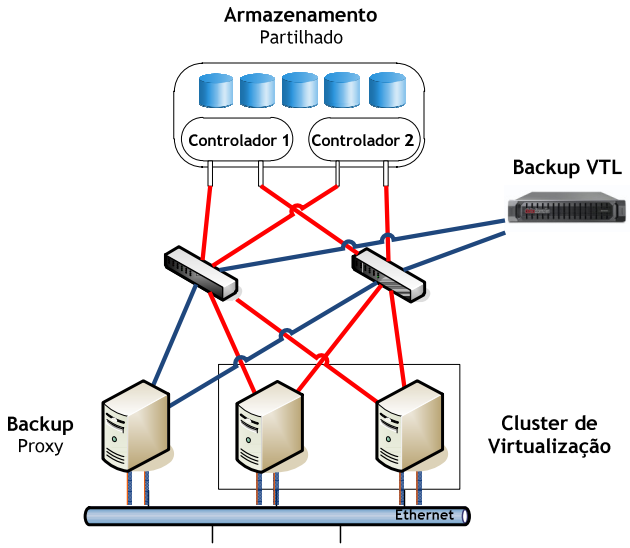


Figura 37 - Arquitetura proposta para o centro de dados

O desenho computacional apresentado assenta sobre três servidores, um *cluster* de dois nós para virtualização e um servidor *proxy* de *backup* consolidando numa infraestrutura em *cloud* privada, tratando-se de uma redução de 85% de infraestrutura de servidores. Os detalhes das configurações podem ser revistos no Anexo II.

4.1. Cluster de Virtualização

De acordo com o levantamento efetuado no que concerne às métricas relativas à infraestrutura existente, a proposta passa por consolidar os 21 servidores em apenas 2 servidores físicos, cada um com 2 CPU *quad-core* 2,86 GHz, 96 GB de memória e 3 NIC de 4 portas cada. O tipo de conetividade do armazenamento proposto é baseado em FC que requer *hardware* adicional e com

custo relativamente superior à abordagem atual, no entanto com latências inferiores comparativamente com a primeira opção entre outros benefícios, como o balanceamento nas interfaces disponíveis [3]. Para além desta abordagem temos a implementação em ambas as arquiteturas da *cloud* privada através do *software* de vitualização VMWare ESX de modo a consolidar a infraestrutura existente e utilizar os recursos disponíveis de forma partilhada reduzindo a quantidade de servidores físicos.

De acordo com o proposto e aplicando as métricas de consumo estimadas em pico obtemos a seguinte relação de consumo estimado face ao padrão de utilização do caso em estudo na janela temporal de análise gerado pelo *software* de análise VMWare Capacity Planner:

	Capacidade			Estimativa situação de pico				
	CPU [GHz]	Memória GB	Consumo kVA	CPU	Memória	Disco I/O	Disco MB/s	Rede Mb/s
Cluster Virtualização	45,7	187	7	22%	36%	4919	38,7	11,74

Tabela 5 - Consumo estimado em pico

A arquitetura proposta baseada em recursos partilhados permite obter consumo máximo de 22% de CPU, ou seja, mesmo assumindo uma falha num nó existe margem em processamento para dar continuidade à operação. A memória é um fator crítico e nesta configuração em pico apresenta o valor de 36%, ou seja, neste ponto é importante salientar que em caso de falha de um nó existe margem para assegurar a continuidade de serviço mas igualmente existe margem de crescimento para adicionar mais serviços com tolerância a falhas.

4.2. Armazenamento partilhado

Relativamente ao armazenamento foi possível retirar da análise pelas taxas de transferência da infraestrutura, quer em leituras como escritas com 35,7 MB/s e 3 MB/s, respetivamente, sendo o valor percentual estimado de escritas na ordem de 8,4%, enquanto o serviço de *mail* apresenta uma taxa de escrita na ordem dos 4% sendo um ambiente maioritariamente baseado em leituras.

Assim sendo, valida-se numa perspetiva de desempenho o necessário em termos de discos, recorrendo às fórmulas, sendo o factor penalidade de escritas igual a 6 associado ao Raid6

utilizado pelo sistema proposto NetApp. Como o serviço mais crítico de *mail* já se encontra isolado numa ótica de grupo RAID com discos dedicados a 15Krpm ou seja, com grupo RAID dedicado mantém-se essa abordagem e restantes dados de aplicações serão colocados noutros:

$$TotalIOs = hIOPSx[\%L] + pexhIOPSx[\%E]$$
$$= 4177x96\% + 6x4177x4\% = 5013IOs$$
$$Q_{discos15k} = \frac{TotalIOs}{TotalIO_{disco15K}} = \frac{5013}{180} = 27{,}85 = 28discos$$

De seguida validam-se as necessidades de armazenamento com total de *mail* de 3,083 TBytes úteis que deve ser somado ao espaço de *snapshot* recomendado pelo fabricante assim como *overhead* de SO, paridade e *spare.* A configuração de disco será igual à existente já que com 24 discos de 300 GB a 15Krpm atinge-se o máximo de 3,5 TBytes úteis de acordo com o *overhead* do fabricante indicado na figura 38:

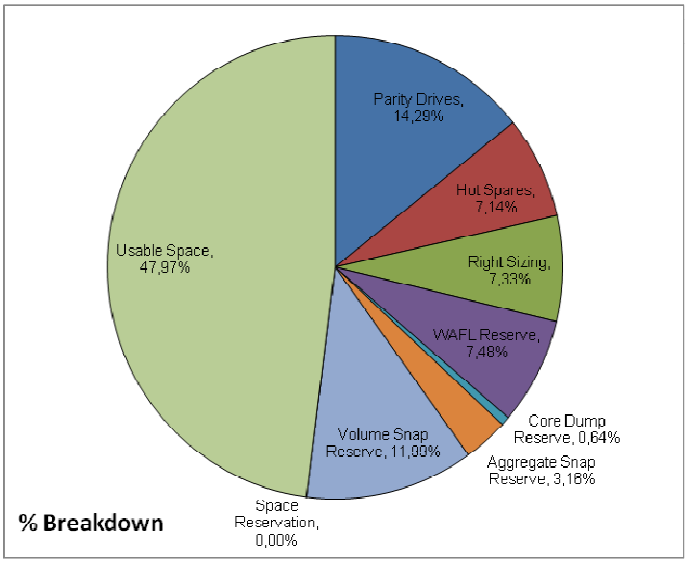


Figura 38 - *Overhead* de um sistema NetApp aplicada á capacidade bruta

Considerando o restante das aplicações temos aproximadamente 742 IOPS de pico e taxas de transferência de leitura e escrita médias de 1,9 e 1,7 MByte/s, respetivamente, ou seja, 53% e 47% pelo que calcula-se a quantidade de disco necessária:

$$TotalIOs = hIOPSx[\%L] + pexhIOPSx[\%E]$$
$$= 742x53\% + 6x742x47\% = 2486IOs$$
$$Q_{dis\cos 15k} = \frac{TotalIOs}{TotalIO_{disco15K}} = \frac{2486}{180} = 13{,}8 = 14dis\cos$$

Como é necessário o espaço dimensionado a 3 anos, temos quer em bases de dados, serviços aplicacionais e imagens de servidores virtuais um total de 2,23 TBytes úteis, neste caso contemplando discos de 450 GBytes seria suficiente obtendo-se um total útil de 2,75 TBytes. Com toda estas considerações obtém-se o sistema configurado, baseado em desempenho e balanceamento dos grupos RAID pelas baías de discos na figura 39:

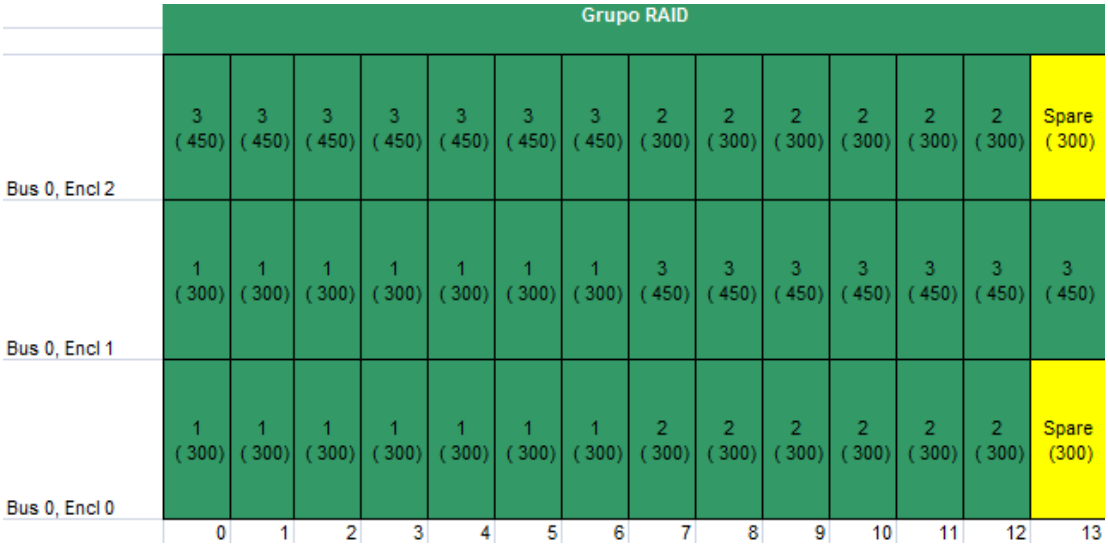


Figura 39 - Arquitetura dos grupos RAID proposta

Algumas aplicações apresentam taxas de crescimento reduzidas ou nulas com taxas de ocupação muito baixas, sendo ideais para utilização da tecnologia de *thin provisioning* de modo a rentabilizar os recursos de armazenamento partilhado.

A tecnologia de *thin provisioning* tem impacto nos tempos de resposta comparativamente a um volume *full provisioning* sendo que não foi contemplado no serviço mais crítico em termos de desempenho, ou seja, o *mail*.

Os discos são endereçados pelos canais de comunicação internos SAS (Bus) e posicionados nas gavetas de discos ou *enclosures* (Encl).

Se contemplar o uso desta tecnologia nos restantes volumes, aplicar a taxa de crescimento anual para uma janela de três anos e associar ainda 10% de margem de volume, obtemos um rácio de poupança de consumos de armazenamento de acordo com a tabela 6.

Servidor	capacidade [GB] alocada	Thin Provisioning [margem 10%]
Servidor1	880	725
Servidor2	3000	0
Servidor3	734	84
Servidor4	734	554
Servidor5	734	633
Servidor6	160	84
Servidor7	146	113
Servidor8	146	0
Servidor9	136	31
Servidor10	136	124
Servidor11	149	141
Servidor12	160	150
Servidor13	292	0
Servidor14	183	128
Servidor15	115	100
Servidor16	63,99	29
Total	7768,99	2896

Tabela 6 - Aplicação de *Thin Provisioning*

Em suma, a aplicação da tecnologia identificada vai permitir poupar cerca de 37,3% da capacidade identificada assim como fazer uma utilização eficiente dos recursos de armazenamento disponíveis permitindo poupar aproximadamente 3 TBytes úteis de espaço em disco. O sistema configurado apresenta dois controladores para efeito de redundância, cada um com 2 portos FC para ligação em alta disponibilidade á SAN.

4.3. Backup

Na figura 37 as ligações a azul são igualmente FC mas retratam as componentes de *backup*, quer a VTL como o *proxy.* Esta arquitetura de *backups* permite, em conjunto com o armazenamento partilhado apresentar um *snapshot* ao *proxy* tendo como benefício: o *offload* de recursos (minimizando o impacto no ambiente de produção), maior controlo dos dados de *backup* sem

impactar produção, servidor *proxy* atua como um servidor para escrever na VTL diretamente via FC (não impactando a rede LAN) e múltiplas cópias de dados de diferentes servidores podem ser geridas no *proxy* para *backup*.

O *software* de *backups* proposto é o EMC NetWorker [15] que permite gerir o ambiente de *backups* e disponibiliza agentes online para recuperações granulares. O *software* do armazenamento NetApp SnapManager [28] permite integrar com a aplicação de modo a coordenar a replicação de ambientes de *mail* Microsoft e proporcionará o *snapshot* ao EMC NetWorker através de um módulo aplicacional. Como o serviço de *mail* possui maior quantidade de informação, maior desempenho e criticidade aplicou-se esta técnica para minimizar o impacto.

A abordagem ao *backup* passa por endereçar as aplicações por tipo de serviço sendo que com especial atenção ao serviço de *mail* com maior volume e complexidade, pois dimensionando a três anos exige um total de 3 TBytes de dados úteis que têm que ser considerados a passar na rede numa janela temporal que idealmente não afete o período laboral para não impactar a produção.

Grupos de Backup	**Totais**
Mail	3083
Bases Dados	819,4
Aplicacionais	836,5
Servidores Virtuais	575
Backup Total	**5313,9**

Tabela 7 - Grupos de Backup com valores a 3 anos

A política de *backups* proposta por grupo considerada é a descrita na tabela 8:

Grupos de Backup	**Política Semanal**	**Retenção**		**Volume Total (GB)**
		Total	**Incremental**	
Mail	1 Total + 5 Incremental	6(1,5meses)	20 (1mês)	21581
Bases Dados	1 Total + 5 Incremental	6(1,5meses)	20 (1mês)	5735
Aplicacionais	1 Total + 5 Incremental	6(1,5meses)	20 (1mês)	5855
Servidores Virtuais	1 Total	4(1 mês)	-	2300
Backup Total	**-**	**-**	**-**	**35472**

Tabela 8 - Política de *backup* proposta e volume retido

De acordo com a análise efectuada podemos inferir três conjuntos aplicacionais sendo o serviço de *mail*, sistemas de bases de dados diversas e ainda ambientes aplicações para realizar *backup* ao nível do *filesystem*. Como a abordagem em *cloud* através da virtualização permite a

desmaterialização de um servidor físico contempla-se o *backup* de imagens de servidores, ou seja, o ficheiro do servidor virtual de modo a poder repor uma imagem de um servidor.

O quadro acima demonstra a relação de políticas de *backup* assim como totais retidos para permitir aceder á informação numa janela temporal de 2 meses sendo que as políticas mais exigentes estão centradas nas aplicações e uma política menos exigente para as imagens dos servidores virtuais.

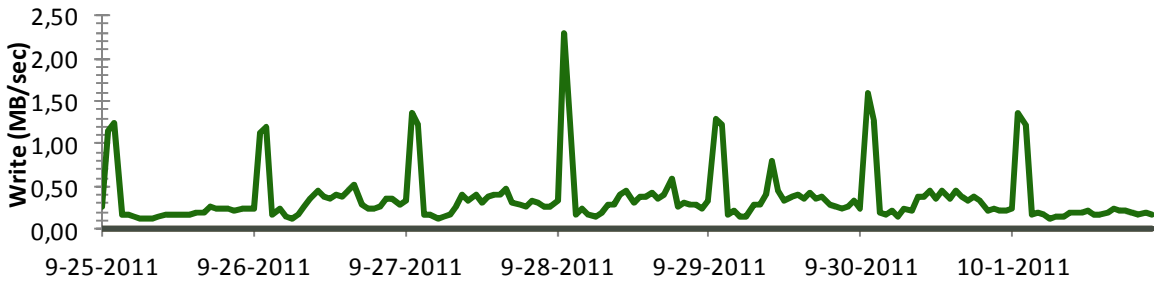


Figura 40 - Taxa de escritas no serviço de mail

Nos cálculos foram assumidos para os *backups* incrementais 10% de taxa alteração pois extrapolou-se um valor médio, baseado em pico e restante valor de alteração do serviço de *mail*, sendo considerado de 1,23 MByte/s, ou seja, 106 GBytes/dia sendo á data a taxa de alteração inferior a 9%. O valor considerado foi arredondado para os 10% e extrapolado para as restantes aplicações para o cálculo nos *backups* incrementais.

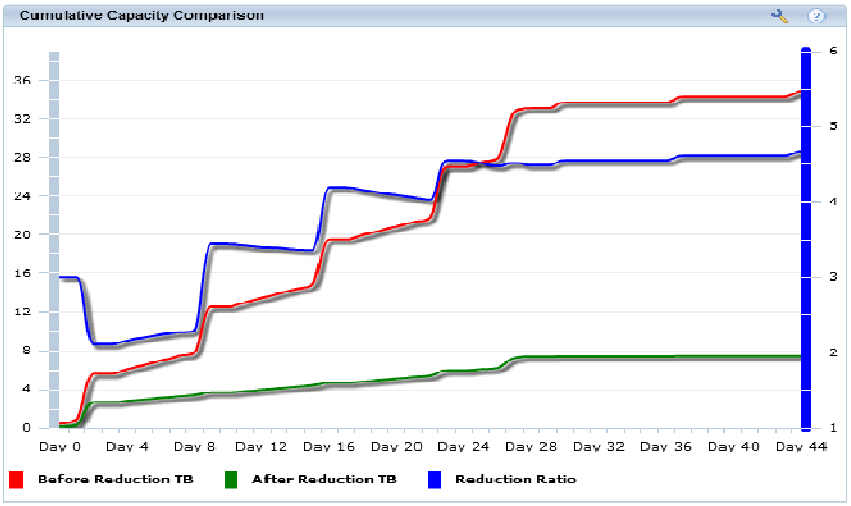


Figura 41 – Rácio de deduplicação

Relativamente ao rácio de deduplicação proposto pela aplicação EMC *Backup System Sizer* do fabricante, consegue-se atingir rácio teórico de 1:4,6, ou seja, armazenar 35 TBytes úteis em sensivelmente 7,4 TBytes com mês e meio de retenção (aproximadamente 45 dias).

4.4. Redes locais e de dados

Os servidores vão ligar via FC a *switches* SAN Brocade DS-300B [29] que suportam como velocidade de 8 Gbit/s por porta, assim sendo assumindo o serviço de *mail* num único controlador (devido à arquitetura do armazenamento) temos uma largura de banda teórica de duas portas FC, ou seja, 16 Gbit/s, sendo 16x superior ao modelo atual já que neste modelo não só se suporta o *failover* na falha da interface como existe balanceamento de dados pelas portas disponíveis. Todos os servidores estão ligados em alta disponibilidade à SAN para maior tolerância a falha assim como na ligação ao armazenamento partilhado.

Como numa *cloud* muitas operações vão se realizar por *switches* virtuais, existem características de *switches* físicos que são implementadas ao nível do *hyper-visor*, no entanto, nem todas são suportadas pelos fabricantes, pelo que, proporcionam plataformas de integração para desenvolvimento de novas funcionalidades de acordo com a área de cada fabricante. Para se poder ter funcionalidades como *access-lists*, existe uma componente Cisco, designada Nexus 1000v [30] que permite estender funcionalidades de *switching* Cisco para a *cloud*, ou seja, a implementação de funcionalidades de rede a nível de *software* integrada com o SO em *cloud* tal como apresentado na figura 42.

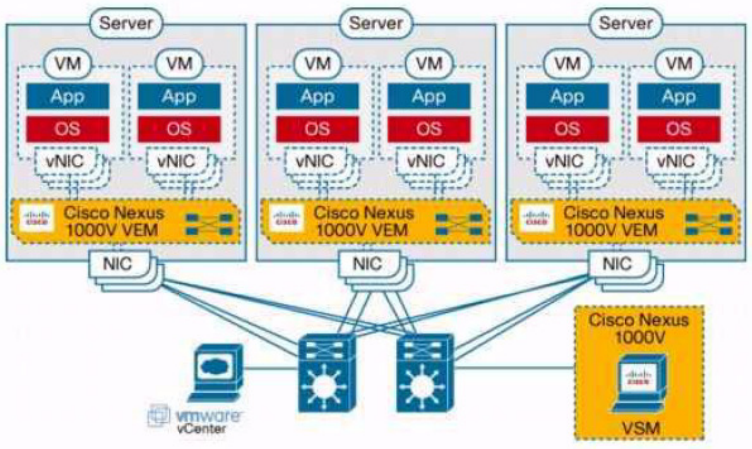


Figura 42 - Cisco Nexus 1000v para ambientes *cloud* [30]

A implementação deste *switch* virtual é baseada em duas componentes distintas, o VSM que é uma máquina virtual que implementa o SO NX-OS e o VEM que representa um agente que é instalado no *kernel* do *hyper-visor* que é responsável por processar tráfego entre interfaces de rede virtuais e físicas.

Estão representadas na figura43 todas as redes a configurar sendo que a rede de vMotion (migração de recursos) ficará na rede de gestão e monitorização para não impactar outras redes como a intranet, visto que eventos de migração de recursos não ocorreram tão frequentemente devido a serem poucos nós no *cluster* e devido à rede de gestão e monitorização ter taxas de utilização mais reduzidas. Neste último *uplink* considera-se VLAN para cada uma das redes mencionadas, sendo a rede de dados associada às HBAs com *zonning* definido para o sistema de armazenamento e VTL em SAN *switches* redundantes para garantir a continuidade em caso de falha com quatro caminhos para o armazenamento e dois caminhos para a VTL.

A rede virtualizada será assegurada pelo *switch* virtual distribuído Cisco Nexus 1000v que permite estender funcionalidades de *switching* ao *hyper-visor* e ser visto como um único *switch* virtual em todos os nós do *cluster*, ou seja, qualquer alteração de configuração será atualizada imediatamente em todos os nós.

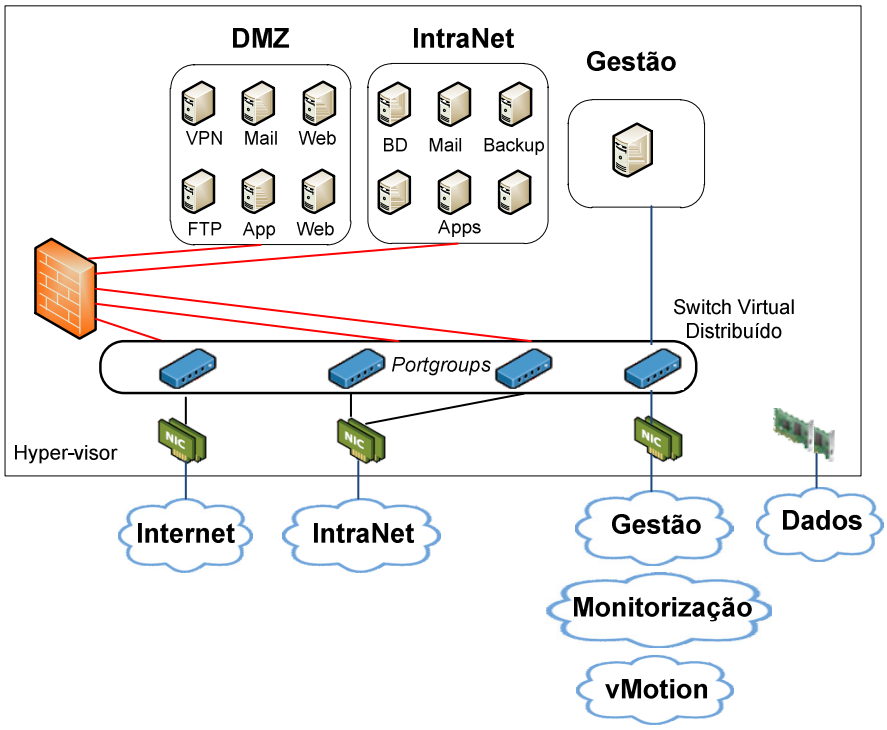


Figura 43 - Diagrama de rede virtualizada

Os *uplinks* considerados apresentam para cada sub-rede 4 portas, com um agregado de 4 Gbit/s, sendo que cada sub-rede tem um par de portas sempre em NIC distintos para em caso de falha de NIC continuar em operação a 2 Gbit/s ligados em *switches* físicos distintos. Os servidores virtuais são ainda agrupados logicamente no *hyper-visor* a um *portgroup* que permite a um servidor virtual ter um determinado tipo de conetividade em cada servidor físico que possa estar.

4.5. Segurança

De acordo com a figura 43 o fato de existir separação física entre redes (*Layer 1*), proporciona maior nível de segurança, no entanto, a proposta passa ainda por adicionar *appliances* de *Firewall* virtuais e funcionalidades Cisco com o *switch* virtual distribuído Nexus 1000v sendo estas as PVLANs e VACL para complemento adicional de segurança. As PVLAN farão a componente de segurança (*Layer 2*) na infraestrutura enquanto as VACL asseguram a configuração para restantes camadas (*Layer 3* e acima). Por motivos de confidencialidade não são aqui revelados os planos de segurança do caso em estudo, no entanto, são enquadradas estas tecnologias para segurança no ambiente proposto em *cloud*.

4.6. Migração para modelo *cloud*

O processo de migração de uma infraestrutura física para o modelo de *cloud* implica um conjunto de melhores práticas e planeamento [27] para conversão de servidores físicos para virtuais no qual a máquina origem é duplicada para um *cluster* de virtualização como destino através de ferramentas tais como o VMWare Converter.

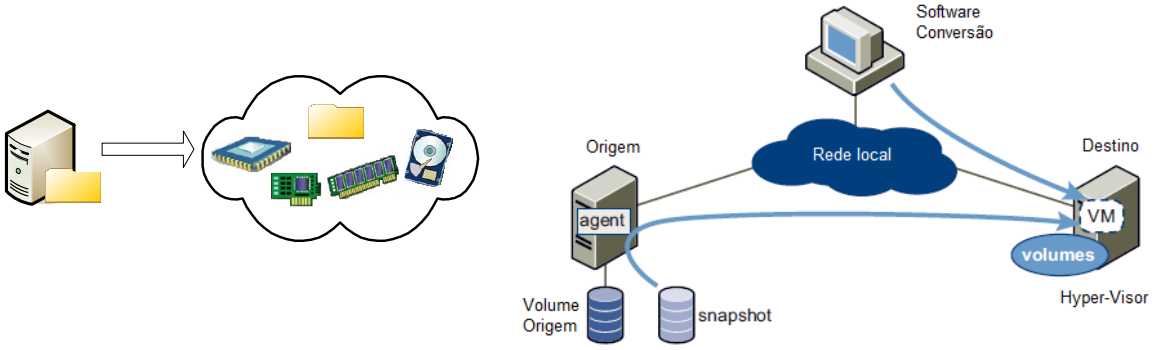


Figura 44 - Migração modelo físico para *cloud* privada [27]

As migrações são tradicionalmente processos morosos e complexos em ambientes físicos, no entanto, através de *software* de conversão é possível fazer o P2V dos servidores onde são definidas as imagens de disco com SO, *boot*, configurações e aplicações que serão desmaterializadas para um ficheiro de extensão .vmdk. Depois de implementada a solução de virtualização, a conversão dos servidores pode ser feita de duas formas distintas, a quente e sem paragem ou a frio com paragem aplicacional. Numa situação de migração o agente faz o *snapshot*

do volume de origem, enquanto o *software* de conversão cria uma máquina virtual no servidor destino e o agente copia os dados para essa máquina sendo que os dados de configuração são igualmente instalados na máquina destino para permitir o arranque do SO na máquina virtual. Durante a configuração o agente personaliza a máquina virtual como por exemplo, informação de IP e for fim faz a limpeza da informação que ficou do lado da máquina fonte. Este processo pode correr em paralelo na conversão de diversos servidores, no entanto, a conversão implica a migração de alguns GBytes até centenas impactando a rede local, no entanto, a utilização existente de armazenamento partilhado evita cópia da informação já que a nova máquina comuta e aponta para o sistema de armazenamento onde reside a informação relacionada com a aplicação.

4.7. Benefícios de implementação do modelo

Apresenta-se de seguida de forma sumária um comparativo entre modelo de *cloud computing* e modelo físico e atual no centro de dados do caso em estudo:

Funcionalidade	**Modelo Físico**	**Modelo *Cloud Computing***
Alta disponibilidade	Apenas o serviço de *mail*	Total
Recursos Partilhados	Não aplicável	Sim
Provisionamento de um novo serviço	Dias a semanas	Minutos a horas
Adaptação de carga a picos de utilização	Estático	Dinâmico
RTO serviços	RTO=0 para serviço de *mail* RTO de dias a semanas para restantes	RTO é igual ao reiniciar do servidor virtual (minutos)
Largura de banda acesso a dados	1 Gb/s no *mail*	16 Gb/s
Balanceamento de dados e alta disponibilidade no acesso aos dados	Apenas *fail-over*	*Fail-over* e balanceamento
Backup retido	Maioritariamente logs e *snapshot* diário do serviço de *mail*	Protecção de toda a informação inclusive imagens de servidores
Impacto *backup* na aplicação	Sim	Não, via *proxy*
Migração de dados	Complexo e moroso	Simples via vMotion
Ocupação de sistemas	52U	16U
Manutenção de servidores	21	3
Consumo energético (kVA)	24,54	7

Tabela 9 – Comparativo de modelos

5. Detalhes de implementação

Na componente experimental pretende-se entender o impacto da introdução da virtualização no centro de dados através de testes comparativos, quer nas latências e taxas de transferência entre ambientes virtuais e físicos, impacto do consumo de CPU na latência de rede e tempos de migração estimada entre os nós do *cluster* de virtualização. Valida-se ainda a escolha do protocolo de comunicação para a rede de dados, *backups* na infraestrutura com janelas de tempo estimadas e rácios de deduplicação. Por fim implementa-se a segurança na rede DMZ para minimizar os possíveis ataques á solução proposta assim como se configura a segurança para acessos externos via serviço VPN.

5.1. Rede Local

Para se perceber o impacto de uma abordagem em *cloud* face à convencional baseada num modelo físico, existem um conjunto de fatores que se traduzem em latência tais como o *overhead* da virtualização, processamento de pacotes (*switches* virtuais), *scheduler* do *hyper-visor* com execução de tarefas de receção e envio de pacotes, *coalescing* virtual associado à espera da receção ou transmissão, a contenção de largura de banda devido ao fato de ter vários vNICs a partilhar a mesma interface física.

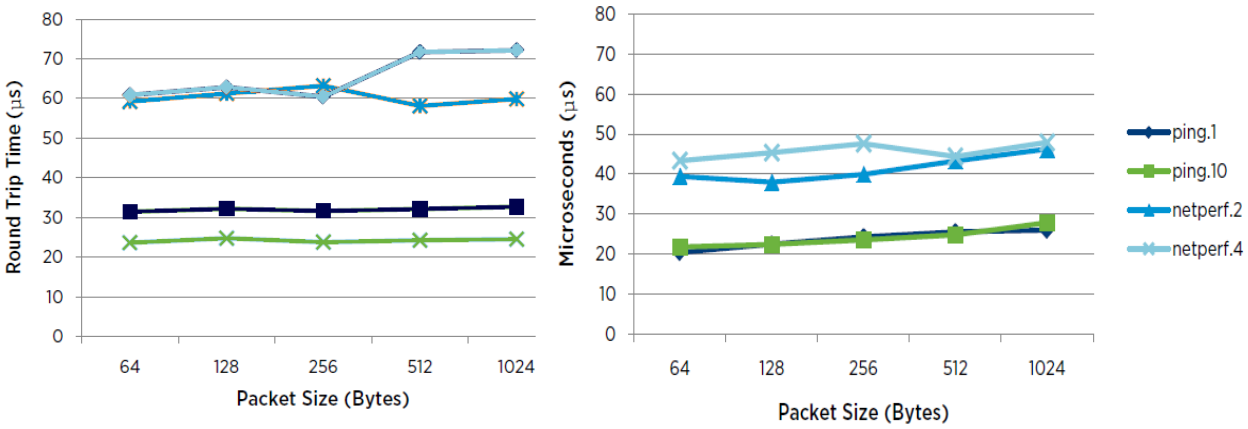


Figura 45 - Latência de rede em ambiente virtual e físico [31]

Um teste [3] que permite validar estes pontos compara o desempenho de rede através de ferramentas como *ping* configurado para 1.000 e 10.000 pacotes por segundo (PPS) com intervalos temporais de 1 milisegundo e ainda o netperf configurado com dimensões de rajada de

tráfego distintas e variação entre 40.000 e 80.000 PPS para simular o mais aproximado possível o tráfego real de aplicações numa rede.

Na figura 45, temos respetivamente, o teste de máquina virtual para máquina virtual e servidor físico para servidor físico com as variações de PPS no teste de ping e dimensão do *burst*, tendo o valor mais elevado a representação de uma situação de *bursts* mais acentuados na rede apresentado com valores de variação de tamanho do pacote.

Na comunicação máquina virtual para máquina virtual a latência é semelhante à comunicação de servidor físico para físico sendo que no teste de *netperf* a variação face ao *ping* deve-se ao facto de as mensagens ping serem servidas pela camada de rede enquanto o *netperf* envia mensagens TCP que percorrem a camada de transporte até á aplicacional. Apesar da comunicação físico para físico possuir latência do *switch* e cabos, a comunicação de virtual para virtual substitui o cabo e *switch* por cópia de memória para memória sendo que as latências de virtualização já descritas continuam a ser uma fonte adicional de latência.

Para ambos os testes validaram-se ainda os valores de qualidade de serviço em percentil 99, com picos de latência de 1ms inferior a 0,009%, picos de latência inferior a 5ms inferior a 0,001% e tempo máximo de resposta inferior a 10 ms traduzindo-se desta forma a variação da latência quer em ambientes virtuais ou físicos como um valor baixo comparativamente com o modelo físico.

O fato de no modelo proposto consolidar-se múltiplos servidores num único, implica concorrência pelos recursos disponíveis, sendo que, a partilha de recursos de rede influencia a utilização de CPU no *hyper-visor*, nomeadamente a latência. Baseado no teste anterior alterou-se o número de PPS gerado por máquina virtual para 5000, dando uma utilização média de 20% de CPU virtual atribuído ao servidor, com tamanho de pacote de 256 Bytes de modo a simular concorrência de aplicações no servidor físico.

Os testes [31] demonstram o aumento da latência à medida que se ligam servidores virtuais e o aumento da latência comparada com a inexistência de carga não ultrapassando os 100 microsegundos desde que o CPU não esteja acima dos 90% (verificou-se até aos 32 servidores virtuais por *hyper-visor* com rácio de consolidação de 4, ou seja, 4 servidores virtuais por *core*). Assim que o valor de utilização do CPU atingiu os 60% existiu um crescimento exponencial de latência devida à contenção no *hyper-visor*.

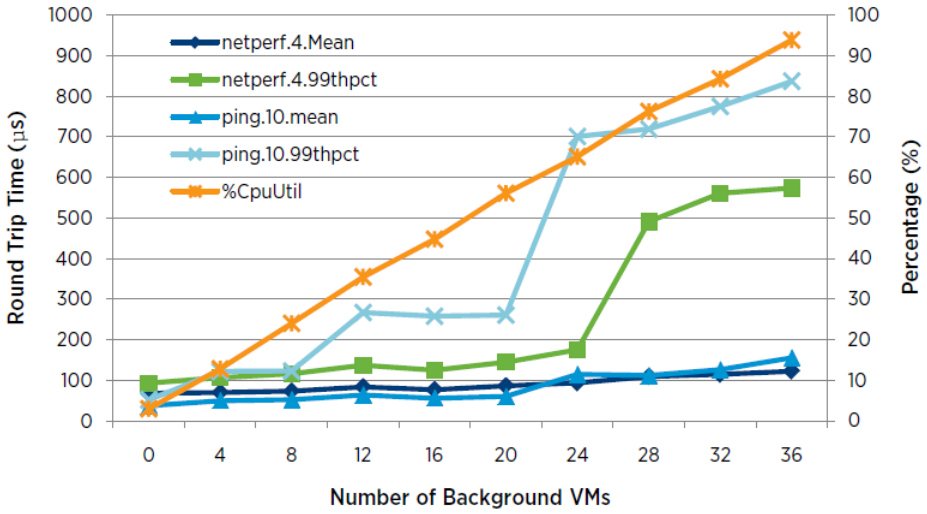


Figura 46 - Impacto do CPU na latência de rede [31]

Deve-se ter em conta que por questões de alta disponibilidade no *cluster* torna-se necessário reservar uma percentagem de CPU para tolerar falhas físicas, sendo este teste apenas demonstrativo da capacidade computacional de um único *hyper-visor*. Relativamente a picos de utilização verificou-se ainda que picos de latência superior a 5 ms são praticamente inexistentes até aos 36 servidores virtuais, picos na ordem do 1 ms tornam-se superiores a 0,1% para 24 servidores virtuais e 65% de utilização de CPU, sendo que não existindo *over-subscription* o número de picos na ordem do 1 milisegundos é inferior a 0,1% desde que o rácio de consolidação seja de três. Extrapolando para o caso de estudo com 21 servidores virtuais com 8 *cores* em cada nó do *cluster* de virtualização (total de 16) obtemos um rácio de 1,3 servidores virtuais por *core* muito abaixo de 3, garantindo uma latência mínima. Apesar de no teste se assumir uma carga de CPU de 20% por servidor virtual, podemos verificar que até aos 11 servidores por nó a taxa de processamento é de 30%. Os servidores físicos analisados no caso em estudo demonstram taxas muito reduzidas de utilização de CPU pelo que o valor final da solução proposta será inferior ao do teste prático.

Á medida que uma *cloud* privada cresce, a probabilidade de comunicação entre servidores virtuais dentro do *hyper-visor* aumenta, devido ao fato de existirem *switches* virtuais a ligarem os servidores o que se traduz em vantagens, ou seja, a transferência de pacotes sem colisões e impacto minimizado na rede física. Existem diversos factores [32] que variam os valores e que podem ser customizados para aumentar o desempenho e escalabilidade de uma rede virtualizada, já que os valores não dependem da velocidade de rede mas sim do CPU e NIC virtual.

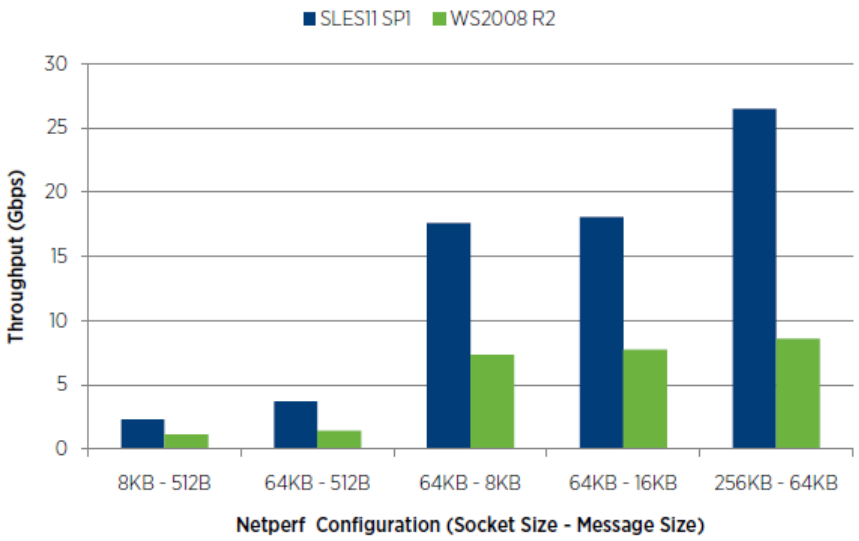


Figura 47 - Taxa de transferência num ambiente virtualizado [32]

A figura 47 demonstra a comunicação entre dois servidores virtuais com a variação do tamanho do pacote através da ferramenta *netperf* sendo que em ambiente Unix/Linux (com SO SLES 11) atinge-se uma taxa de aproximadamente 27 Gbit/s, ou seja, três vezes mais que a velocidade de rede LAN de teste a 10 Gb/s.

Estas taxas são maioritariamente limitadas pelo servidor virtual que está a receber os pacotes, uma vez que é mais intensivo de CPU receber do que transmitir os pacotes e no caso prático a diferença entre Unix/Linux e Windows deve-se à implementação de *Large Receive Offload* que agrega múltiplos pacotes numa *stream* única para um *buffer* maior, antes de ser passado à camada de rede.

Verificou-se que fazendo passar um pacote do servidor virtual para um servidor físico introduz-se latência adicional devido ao *overhead* da virtualização.

Na figura 48 demonstra-se o impacto nas taxas de transferência sendo que como se pode validar para pacotes de dimensão mais elevada praticamente se esgota o *link* do teste prático (10 Gbit/s).

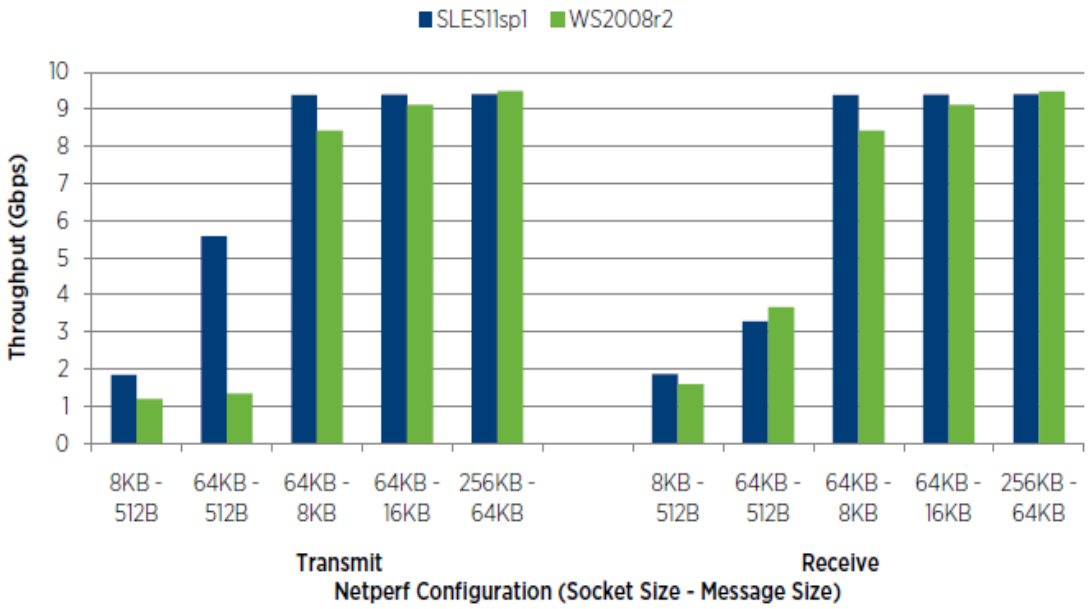


Figura 48 - Taxa de transferência em ambiente misto [32]

A diferença de desempenho entre pacotes mais pequenos e maiores deve-se a uma otimização [32] na ferramenta netperf que permite rentabilizar os valores de transferência, sendo o objetivo para os pacotes mais pequenos validar a capacidade de processamento dos mesmos.

Qualquer que seja o SO em teste nos servidores virtuais, cada um tem a capacidade de processar até 800.000 pacotes de 512 Bytes e se se aumentar a quantidade de servidores virtuais que irão concorrer em CPU, a taxa de transferência praticamente mantém-se até aos 32 servidores virtuais havendo em algumas situações em pacotes de dimensão superior, uma quebra de 20%.

5.2. Rede de dados

Relativamente à componente de rede de dados, existem três formas de disponibilizar acesso às aplicações, num ambiente virtualizado em VMWare [33], ou seja, através de FC, iSCSI e NFS de modo a validar as possíveis diferenças de desempenho. No ambiente de teste foi reduzida a quantidade de cache no servidor de *mail* de modo a forçar mais IOPS na rede de dados para validar os consumos de recursos associados a cada tecnologia. Os testes de desempenho demonstram que o protocolo FC proposto permite obter latências mais baixas comparativamente às restantes tecnologias em estudo quer com um perfil de acessos exigente e outro mais exigente com o dobro das transações.

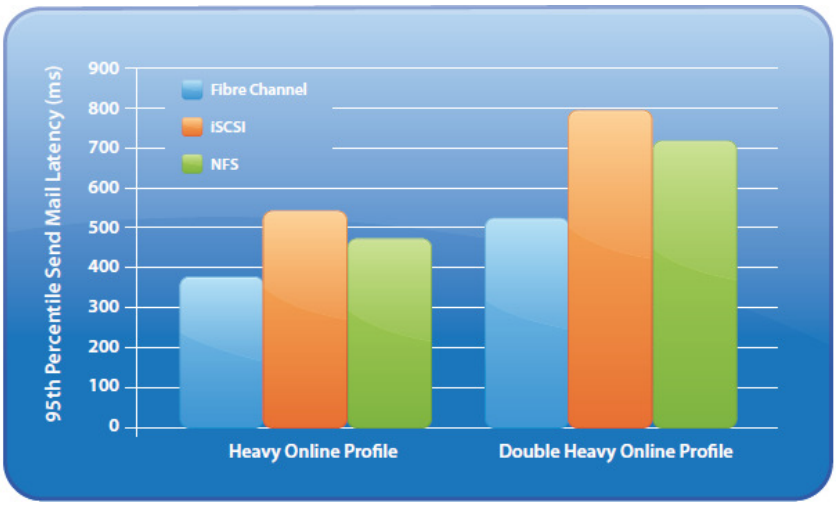


Figura 49 - Latência do serviço de mail em Percentil95 [33]

Face à natureza da aplicação possuir alturas de pico e *burst* de I/O apresenta-se a figura 49 em percentil 95 para representar a latência o mais aproximadamente possível da realidade em toda a janela temporal do teste. O teste acima demonstra dois perfis distintos, sendo o primeiro para um conjunto de utilizadores com perfil de 47 tarefas diárias e um segundo perfil para utilizadores mais intensivo com 94 tarefas diárias, sendo uma simulação o mais aproximada possível através de uma ferramenta nativa para gerar carga no serviço de *mail*.

No primeiro teste o FC destaca-se sendo 26% e 40% mais eficiente que os protocolos NFS e iSCSI respetivamente, quanto ao segundo teste obtém-se 35% e 42% de benefício na latência do serviço.

O impacto de processamento nos recursos disponíveis podem ser validados de acordo a figura 50 em ambiente SAN (FC e iSCSI) e NAS (NFS). O FC é o protocolo que possui impacto inferior devido ao *overhead* dos pacotes ser mínimo e ainda grande parte do processamento ser feito ao nível das HBAs, sendo que os restantes protocolos de rede possuem impacto superior devido ao processamento adicional de dados, nomeadamente a componente TCP/IP em redes Ethernet.

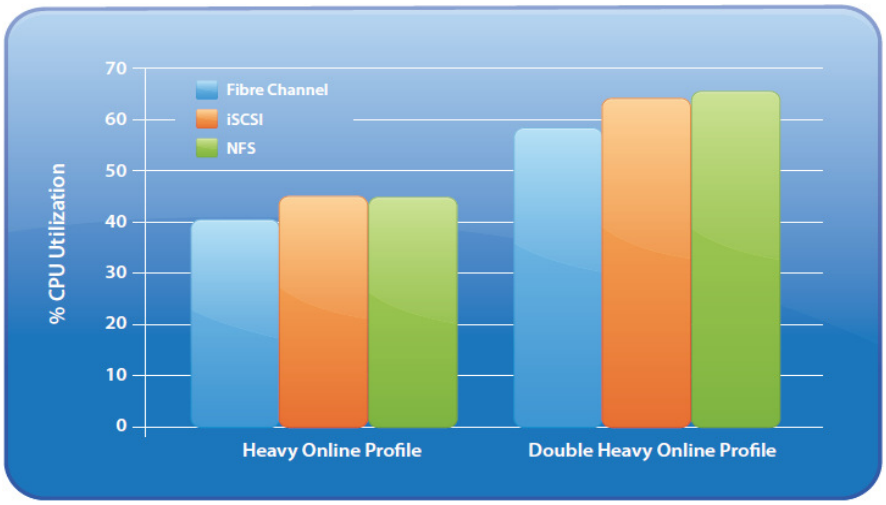


Figura 50 - Impacto no processamento dos protocolos [33]

No primeiro teste tanto o iSCSI como o NFS apresentam 12,5% de processamento adicional face ao FC e no segundo teste apesar de em NFS ter ligeiramente um impacto adicional ao iSCSI, o *overhead* será de aproximadamente 12% face ao FC.

A performance do serviço de *mail* pode ser validada na figura 51 para ambos os protocolos sendo o teste feito numa janela de 8 horas para atingir um estado estável com valor médio de 3700 IOPS que servirá de base para extrapolar para a situação de pico validada no serviço de mail do caso de estudo. Neste teste recorrendo a uma única ligação gigabit e devido á natureza aplicacional com *bursts* de I/O facilmente se esgota a capacidade de uma única ligação.

A quantidade de discos contemplada é semelhante á proposta e permite endereçar a carga em situação de pico de tráfego. O teste realizado implica uma simulação de carga no serviço para ambos os protocolos, sendo o valor semelhante em todos, apenas varia a latência e impacto de computação nos sistemas.

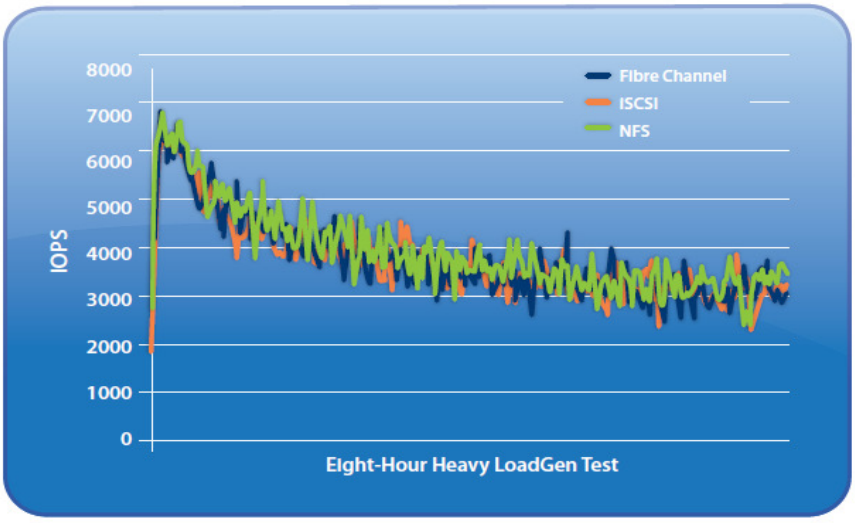


Figura 51 - Desempenho do armazenamento [33]

Quanto ao protocolo existente para acesso aos dados, valida-se a taxa de transferência do serviço de *mail* aos dias de hoje medida nas portas dos controladores de armazenamento na figura 52:

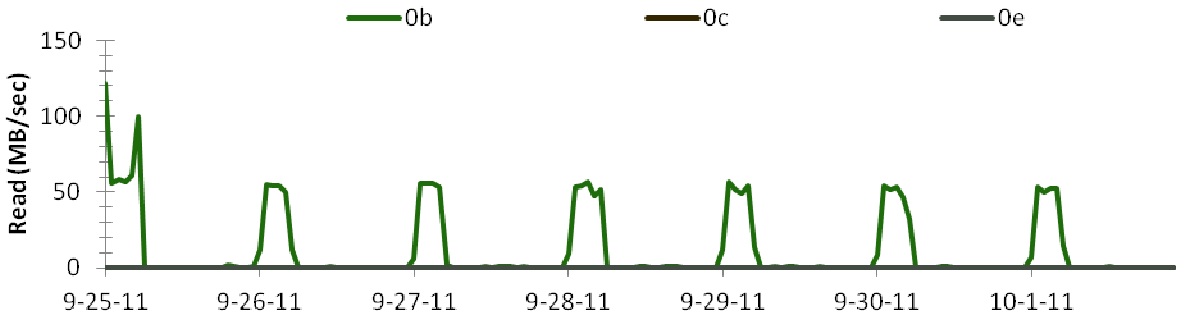


Figura 52 - Largura de banda no acesso a dados

De acordo com a janela disponível podemos verificar as taxas de transferências no caso em estudo com valores mais elevados nas leituras, sendo que no primeiro dia do gráfico atinge-se o pico de 120 MByte/s associado aos *backups*, ou seja, muito próximo de esgotar o *link* sendo 1 Gbit/s iSCSI numa altura considerada de pico, sendo ainda o valor médio decorrido da análise de 33,9 MByte/s na leitura e 1,23 MByte/s na escrita.

5.3. Virtualização

De acordo com o teste elaborado, o mais próximo possível da realidade [34], chegaram-se a um conjunto de resultados comparativos com o modelo atual e o modelo em *cloud* recorrendo à tecnologia de virtualização.

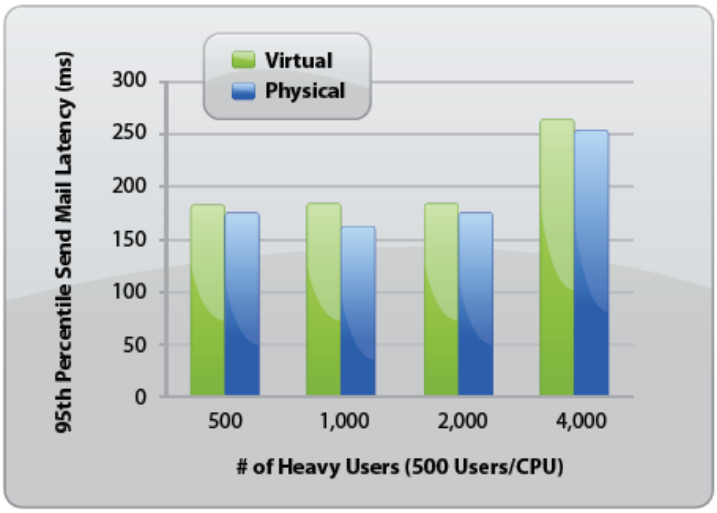


Figura 53 - Serviço de *mail* em modelo em *cloud* e físico [34]

A figura 53 traduz o impacto no serviço de *mail* relativamente ao tempo que as mensagens permanecem em espera até serem enviadas recorrendo a um padrão de utilização do serviço elevada. O impacto da introdução da virtualização é mínimo comparativamente ao modelo físico e aumentando a carga de utilização no serviço a latência permanece ligeiramente acima do modelo físico com a penalização média de 5% de todos os testes efetuados.

Para se garantir a alta disponibilidade e a mobilidade das aplicações entre servidores físicos recorreu-se a um teste de carga [35] em ambiente do *mail* e aplicação *web* e registaram-se as três fases distintas abordadas, nomeadamente: mapeamento, pré-cópia e comutação [10].

Verificou-se que o tempo de migração de servidor virtual em ambiente de *mail* com 28 GBytes de cache, semelhante ao perfil no caso em estudo que utiliza em média 90%, em utilização para um ambiente de 4.000 utilizadores concorrenciais de perfil exigente demorou 47 segundos (rede dedicada de 10 Gb/s).

As fases distintas da migração podem ser validadas no gráfico para um teste de carga em ambiente *web* abaixo onde se denotam as fases aplicadas, com um impacto aproximado de 30% nas transações para o mapeamento da cache numa reduzida fração de tempo e a comutação que possui o maior impacto na aplicação levando á perda de transacções.

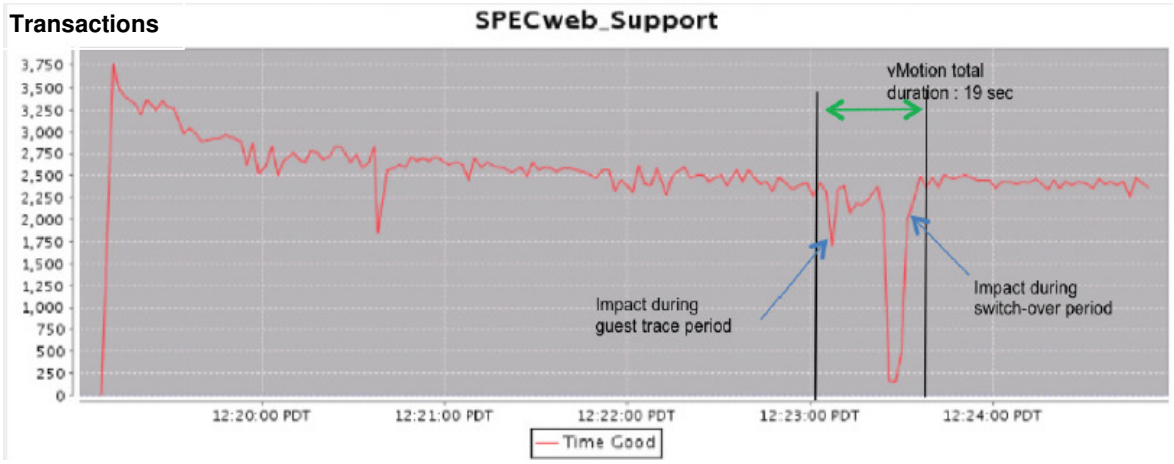


Figura 54 - vMotion de um servidor *Web*[35]

O serviço de *mail* apresenta a situação mais crítica para realizar uma migração entre servidores físicos, considerando o exemplo prático de 47 segundos numa única ligação a 10 Gbit/s para a transferência de 28 GBytes de dados, tem-se uma taxa de transferência média de 596 MByte/s. Considerando a proposta com quatro portas configuradas em *teaming* e com a VLAN de vMotion, obtemos uma largura de banda teórica de 4 Gbit/s sendo o valor prático expetável traduzido por extrapolar para rede Gigabit e assumindo tráfego próximo de zero na rede, obtendo-se assim um tempo estimado de migração de 117 segundos.

Assumindo a tolerância a falha no sistema se um NIC tiver falha física ou ficar *offline* devido a manutenção ou outro fator, implica um aumento de migração para o dobro do valor. Considerando dois NICs de duas portas no servidor de *backups* e os doze portos por nó do *cluster* de virtualização obtemos um total de 28 portos na infraestrutura em *cloud* face aos 43 atuais, obtendo-se uma redução de 35% em portas podendo escalar se considerar em consolidar mais aplicações.

5.4. *Backups*

Com a aplicação da arquitetura proposta de *backups* é possível atingir os rácios de deduplicação de dados com retenção de 6 semanas de informação em disco de acordo com a tabela 9 [36]:

Serviço de Mail		GBytes	SUB-TOTAL	TOTAL
1º Full	1,46 : 1		2111,643836	
Subsequent Full	37,0 : 1		333,2972973	
Incr [10% alteracao)	37,0 : 1		166,6486486	2611,5898
Serviços de Bases de Dados			SUB-TOTAL	TOTAL
1º Full	1,5 : 1		546,2666667	
Subsequent Full	6,0 : 1		655,52	
Incr [10% alteracao]	6,0 : 1		273,1333333	1474,92
Serviços Aplicacionais			SUB-TOTAL	TOTAL
1º Full	1,55 : 1		557,6666667	
Subsequent Full	10,5 : 1		318,6666667	
Incr [10% alteracao]	10,5 : 1		159,3333333	1035,6667
Servidores Virtuais			SUB-TOTAL	TOTAL
1º Full	4,0 : 1		143,75	
Subsequent Full	17,0 : 1		101,4705882	
				245,22059
			TOTAL	5367,397

Tabela 10 - Rácios de deduplicação de dados de *backup*

Partindo de um valor de 35,4 TBytes úteis de volume de *backup* podem ser retidos em apenas 5,36 TBytes, permitindo uma poupança de 85% de espaço em disco com um rácio de 1:6,6 e quanto maior a retenção maior a eficiência já que devido à deduplicação de dados apenas são guardados os blocos distintos ao invés de todo o *backup*.

A nível de rede consegue-se igualmente reduzir o tráfego do primeiro *backup* total em 38% e restantes em 94% assim como o processamento é reduzido devido ao valor diminuto de blocos a escrever em disco.

Para além do armazenamento necessário para os diferentes tipos de aplicações, extrapolou-se o seguinte gráfico com tempos estimados (RTO) de recuperação de *backups* e restauro dos dados para serviços de *mail*, bases de dados e aplicacionais [36]:

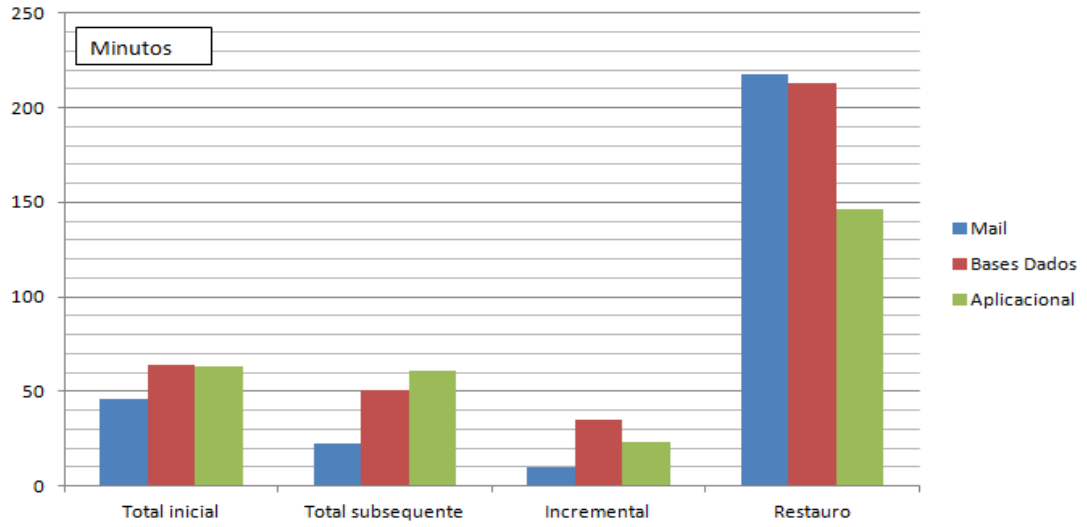


Figura 55 - Tempo estimado associado ao restauro de dados

A figura 55 demonstra os tempos de *backup* iniciais totais, subsequentes e incrementais assim como o restauro, em minutos, relativos a um volume de 537 GBytes, sendo a premissa escolhida devido a uma das maiores LUNs do armazenamento associado ao serviço de *mail*. O RTO (restauro) associado ao *mail* para o volume referido é de 218 minutos para qualquer período da retenção, no entanto, o sistema de armazenamento possui um *snapshot* com a imagem até ao dia anterior que permite o restauro em 8 minutos feito pelo sistema de armazenamento.

Durante o restauro da informação da VTL o consumo de recursos de processamento atinge os 24% assim como a utilização de rede os 41% sendo o serviço mais exigente de recursos para recuperação. Os totais subsequentes são na sua maioria inferiores aos iniciais devido à deduplicação de dados quantificada na tabela 9 o que permite diminuir igualmente o tempo de *backup* assim como o incremental será sempre inferior já que os blocos de dados apenas reflectem os alterados.

O fato da quantidade de dados a realizar *backup* ser elevada e a janela fixa, tipicamente no período da noite, implica ainda processamento adicional devido ao próprio *backup*. Desta forma aborda-se um servidor *proxy* de *backup* para realizar todo o processamento e realizar o próprio *backup* via SAN para não impactar a rede LAN, disponibilizar a maior largura de banda possível para o fluxo da informação e não impactar o servidor de produção.

5.5. Segurança

A implementação de segurança recorrendo a PVLANs e VACLs pode ser elaborada num ambiente virtualizado em *cloud* [37] de forma semelhante como se ratasse de uma rede física, sendo que os servidores físicos serão substituídos por servidores virtuais e as VACLs usadas para comunicação entre os mesmos e para isolar a rede de produção e gestão.

A arquitetura virtualizada representa os servidores virtuais ligados ao *switch* virtual Cisco Nexus 1000v através de NICs virtuais, que por fim possuem os *uplinks* do *switch* para os NICs físicos e consequentemente as redes LAN físicas. Se, por exemplo, for necessário isolar tráfego entre servidores virtuais (VM2 e VM3), uma VACL pode ser aplicada à porta virtual de VM2 com IP 10.10.10.10 e igualmente bloquear o tráfego de VM3 com IP 10.10.20.20. Como o tráfego entre os dois servidores não sai do servidor físico, as VACLs permitem filtrar através dos seguintes comandos:

```
n1000v(config)# ip Access-list deny-vm-to-vm-traffic
n1000v(config)# deny ip host 10.10.10.10 host 10.10.20.20
n1000v(config)# permit ip any any
n1000v(config)# deny ip host 10.10.20.20 host 10.10.10.10
n1000v(config)# permit ip any any
```

Se se pretender isolar tráfego de um servidor virtual para a rede de gestão através da *service-console* (componente da virtualização para aceder à gestão) as PVLANs e as VACLs podem ser combinadas para traduzir maior segurança nos acessos, bloqueando tráfego da VM 10.10.10.10 para prevenir acesso á gestão na rede 192.168.20.0:

```
n1000v(config)# ip Access-list deny-vm-traffic-to-service-console
n1000v(config)# deny ip 10.10.10.10 192.168.20.0
n1000v(config)# permit ip any any
```

5.5.1. Sub-rede DMZ

O primeiro passo [38] consiste em garantir o isolamento *Layer2* através das PVLANs e garantir que servidores na DMZ não falam entre si enquanto clientes internos e externos conseguem acedê-los. Esta implementação é feita colocando os servidores numa PVLAN secundária com portos isolados, enquanto a *firewall* deve ser definida na PVLAN primária com porto promíscuo.

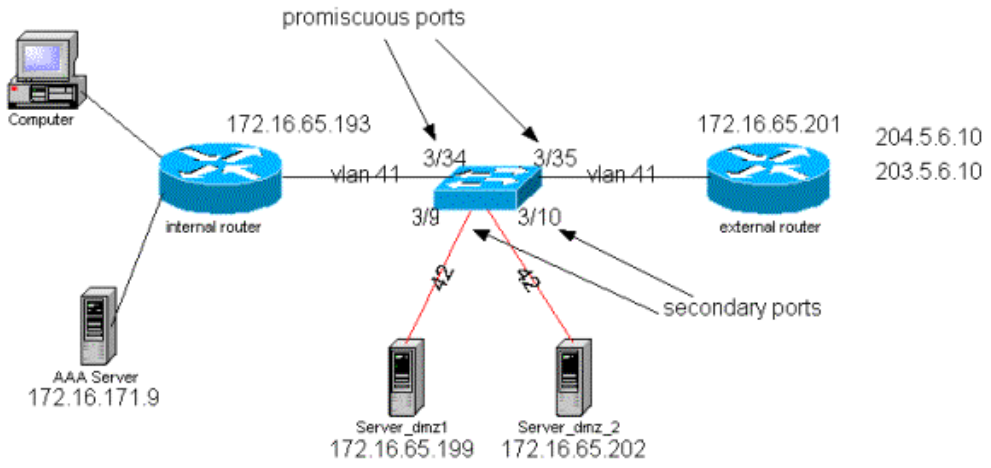


Figura 56 – Aplicação de PVLANs na DMZ [38]

De acordo com a figura 56 temos serviços na DMZ ligados aos portos 3/9 e 3/10, com rede interna ligada ao porto 3/34 e rede externa ligada ao porto 3/35 e as PVLANs configuradas são a 41 como primária e a 42 como secundária:

```
n1000v (enable) set vlan 41 pvlan primary
n1000v (enable) set vlan 42 pvlan isolated
n1000v (enable) set pvlan 41 42 3/9-10
Successfully set the following ports to Private Vlan 41,42: 3/9-10

n1000v (enable) set pvlan mapping 41 42 3/35
n1000v (enable) set pvlan mapping 41 42 3/34

Port  Name                 Status     Vlan       Duplex Speed Type
----- ------------------ ---------- ---------- ------ ----- --------
----
 3/9  server_dmz1       connected  41,42      a-half  a-10
10/100BaseTX
 3/10 server_dmz2       connected  41,42      a-half  a-10
10/100BaseTX
  3/34 to_6500_1         connected   41         auto  auto
10/100BaseTX
 3/35 external_router_dm connected  41         a-half  a-10
10/100BaseTX
```

Seguidamente são efetuados vários testes para validação da implementação entre comunicação externa, interna e intra-DMZ:

a) **Teste 1** – Comunicação entre rede externa, interna e DMZ:

```
external_router#ping 172.16.65.193
Type escape sequence to abort.
Sending 5, 100-byte ICMP Echos to 172.16.65.193, timeout is 2
seconds:
!!!!

external_router#ping 172.16.65.202
Type escape sequence to abort.
Sending 5, 100-byte ICMP Echos to 172.16.65.202, timeout is 2
seconds:
!!!!!
Success rate is 100 percent (5/5), round-trip min/avg/max = 1/2/4 ms
external_router#ping 172.16.65.199
Type escape sequence to abort.
Sending 5, 100-byte ICMP Echos to 172.16.65.199, timeout is 2
seconds:
!!!!!
Success rate is 100 percent (5/5), round-trip min/avg/max = 1/1/4 ms
```

Apenas os servidores na DMZ foram possível atingir por pacotes ICMP, sendo que á rede interna não foi possível chegar.

b) **Teste 2** – Comunicação DMZ para rede externa e interna:

```
server_dmz1#ping 203.5.6.10

Type escape sequence to abort.
Sending 5, 100-byte ICMP Echos to 203.5.6.10, timeout is 2 seconds:
!!!!!
Success rate is 100 percent (5/5), round-trip min/avg/max = 1/2/4 ms

Type escape sequence to abort.
Sending 5, 100-byte ICMP Echos to 172.16.65.193, timeout is 2
seconds:
!!!!!
Success rate is 100 percent (5/5), round-trip min/avg/max = 4/4/4 ms

server_dmz1#ping 172.16.65.202
Type escape sequence to abort.
Sending 5, 100-byte ICMP Echos to 172.16.65.202, timeout is 2
seconds:
.....
Success rate is 0 percent (0/5)
```

Os servidores na DMZ conseguem pingar a rede externa, *default-gateway* mas não chegam aos servidores que estão na mesma VLAN secundária. De modo a melhorar a segurança na sub-rede as VACLs são cruciais pois mesmo que os servidores pertençam a diferentes VLANs secundárias ou à mesma VLAN isolada, existe sempre a possibilidade de um atacante poder utilizá-las para comunicar entre elas. Se o servidor tentar comunicar diretamente, não o fará ao nível da *Layer2* devido às PVLANs, no entanto, se o servidor estiver comprometido e configurado pelo atacante na forma em que o tráfego para a mesma sub-rede seja enviado pelo *router* (*Layer3*), este irá reenviar o tráfego para a mesma sub-rede, ultrapassando a segurança das PVLANs. Desta forma as VACL têm que ser configuradas na VLAN primária (VLAN que transporta tráfego dos *routers*) com as seguintes políticas [38]:

i. Permitir tráfego com endereço IP fonte igual ao IP do router,
ii. Negar o tráfego com endereço IP fonte e destino na sub-rede DMZ,
iii. Permitir restante tráfego

Sendo implementado da seguinte forma:

```
n1000v (enable) sh sec acl info protect_pvlan
set security acl ip protect_pvlan
------------------------------------------------
1. permit ip host 172.16.65.193 any
2. permit ip host 172.16.65.201 any
3. deny ip 172.16.65.192 0.0.0.15 172.16.65.192 0.0.0.15
4. permit ip any any

n1000v (enable) sh sec acl
ACL                              Type VLANS
-------------------------------  ---- -----
protect_pvlan                    IP   41
```

c) **Teste 3** – Comunicação externa (via router) para sub-rede DMZ:

```
external_router#ping 172.16.65.199
Type escape sequence to abort.
Sending 5, 100-byte ICMP Echos to 172.16.65.199, timeout is 2
seconds:
...
Success rate is 0 percent (0/5)
```

Desta forma mesmo que exista controlo de um servidor por parte do atacante e o tráfego seja configurado para ser enviado para a sub-rede via *router*, as VACLs não permitirão esta comunicação.

Quando se recorrem a VACLs o tráfego é descartado no *hardware*, não afetando os recursos de processamento do *router* ou *switch*, mesmo que se trate de um ataque Distributed Denial of Service (DDoS) o *switch* descartará todo o tráfego não permitido à velocidade do *link* sem penalidades de desempenho.

5.5.2. Acessos VPN

Depois de validada a segurança em redes DMZ existe outro desafio [38] associado ao acesso externo autorizado por túneis VPN de modo a proporcionar a experiência de acesso à *cloud* privada no caso em estudo para um utilizador externo ligado em qualquer ponto na *internet*. Uma implementação usual é a abordagem em paralelo a qual é mais simples de implementar, sem impacto na infraestrutura sendo o concentrador de acessos ligado a segmentos internos como externos sem passar por *firewalls*.

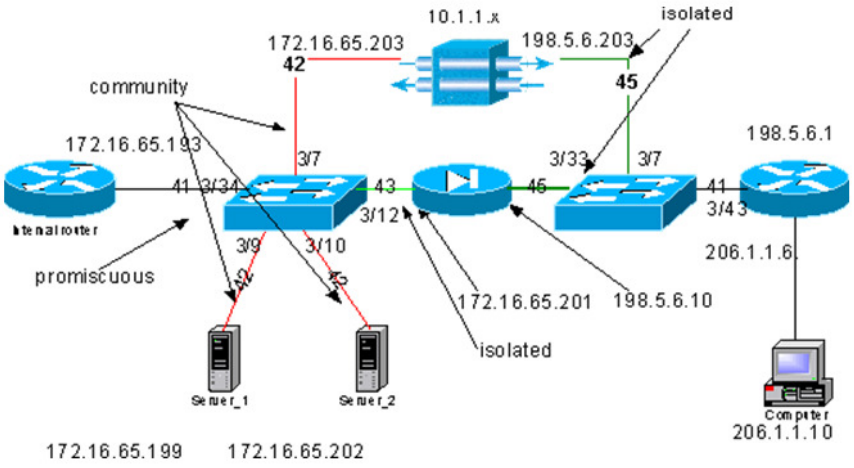


Figura 57 – Aplicação de PVLANs para acessos VPN [38]

A implementação desta arquitetura tem como premissas:

a) Para o *switch* interno:
   i. Clientes VPN terem total acesso a um conjunto de servidores,
   ii. Clientes internos podem igualmente aceder aos servidores,
   iii. Clientes internos terem total acesso à *internet*,
   iv. Tráfego proveniente do concentrador VPN isolado da *firewall*,

b) Para o *switch* externo:
   i. Tráfego do *router* deverá ir tanto para o concentrador como para a *firewall*,
   ii. Tráfego da *firewall* deverá estar isolado do tráfego da VPN,

Adicionalmente contempla-se evitar o tráfego da rede interna para o concentrador de acessos da VPN através das VACL. Como o objetivo principal é segregar o tráfego da *firewall* do tráfego dos servidores e VPN, configura-se a *firewall* (*appliance* virtual) numa PVLAN diferente da dos servidores do concentrador de acessos VPN. O tráfego que flui da rede interna tem que aceder aos servidores assim como o concentrador e *firewall,* tendo que ser configurados como porta promíscua. Os servidores e o concentrador pertencem à mesma VLAN secundária pois têm de comunicar entre eles, o *switch* externo que ligará ao *router* que dará acesso à internet é ligado num porto promíscuo enquanto o concentrador e a *firewall* pertencem à mesma PVLAN com portas isoladas (não podem trocar tráfego).

```
Switch Interno:
n1000v (enable) set vlan 41 pvlan primary
n1000v (enable) set vlan 42 pvlan community
n1000v (enable) set vlan 43 pvlan isolated
n1000v (enable) set pvlan 41 42 3/7
n1000v (enable) set pvlan 41 42 3/9-10
n1000v (enable) set pvlan 41 43 3/12
n1000v (enable) set pvlan mapping 41 42 3/34
n1000v (enable) set pvlan mapping 41 43 3/34
n1000v (enable) sh port (id)
Port  Name                 Status     Vlan       Duplex Speed Type
----- ------------------ ---------- ---------- ------ ----- --------
 3/7  to_vpn_conc         connected  41,42      a-half  a-10
10/100BaseTX
 3/9  server_1            connected  41,42      a-half  a-10
10/100BaseTX
 3/10 server_2            connected  41,42      a-half  a-10
10/100BaseTX
 3/12 to_pix_intf1        connected  41,43      a-full a-100
10/100BaseTX
3/34 to_int_router       connected  41         a-full a-100
10/100BaseTX

Switch Externo:
ecomm-6500-1  (enable) set vlan 41 pvlan primary
ecomm-6500-1  (enable) set vlan 45 pvlan isolated
ecomm-6500-1  (enable) set pvlan 41 45 3/7
ecomm-6500-1  (enable) set pvlan 41 45 3/33
ecomm-6500-1  (enable) set pvlan mapping 41 45 3/43
ecomm-6500-1  (enable) sh port (id)
ecomm-6500-1 (enable) sh port 3/7
Port  Name                 Status     Vlan       Duplex Speed Type
----- ------------------ ---------- ---------- ------ ----- --------
 3/7  from_vpn            connected  41,45      a-half  a-10
10/100BaseTX
 3/33 to_pix_intf0        connected  41,45      a-full a-100
10/100BaseTX
 3/43 to_external_router connected  41         a-half  a-10
10/100BaseTX
```

De seguida são feitos vários testes para validação da implementação entre comunicação externa, interna, VPN e *firewall*:

a) **Teste 1** – *Router* interno consegue chegar ao *router* externo:

```
ping 198.5.6.1
Type escape sequence to abort
Sending 5, 100-byte ICMP Echos to 198.5.6.1, timeout is 2 seconds:
Success rate is 100 percent (5/5), round-trip min/avg/max = 1/1/1
ms
```

b) **Teste 2** – Servidor1 para *router* interno, VPN, *firewall* e *router* externo:

```
server_1#ping 172.16.65.193

Type escape sequence to abort.
Sending 5, 100-byte ICMP Echos to 172.16.65.193, timeout is 2
seconds:
Success rate is 100 percent (5/5), round-trip min/avg/max = 1/2/4
ms

server_1#ping 172.16.65.203

Type escape sequence to abort.
Sending 5, 100-byte ICMP Echos to 172.16.65.203, timeout is 2
seconds:
Success rate is 100 percent (5/5), round-trip min/avg/max = 1/2/4
ms

server_1#ping 172.16.65.201

Type escape sequence to abort.
Sending 5, 100-byte ICMP Echos to 172.16.65.201, timeout is 2
seconds:
Success rate is 0 percent (0/5)

server_1#ping 198.5.6.1

Type escape sequence to abort.
Sending 5, 100-byte ICMP Echos to 198.5.6.1, timeout is 2 seconds:
Success rate is 0 percent (0/5)
```

Mesmo que os servidores e *firewall* pertençam a duas VLAN secundárias existe sempre a possibilidade de um atacante gerar comunicação, apesar de as duas secundárias não conseguirem, é sempre possível através do servidor enviar tráfego para o *router* que irá remeter o tráfego para a mesma sub-rede.

Com o recurso á implementação de VACL na PVLAN primária que permita tráfego IP da fonte IP do *router*, não permita tráfego com IPs destino e fonte da sub-rede dos servidores e permita restante tráfego:

```
ecomm-6500-2 (enable) sh sec acl info protect_pvlan
set security acl ip protect_pvlan
------------------------------------------------
1. permit ip host 172.16.65.193 any
2. deny ip 172.16.65.192 0.0.0.15 172.16.65.192 0.0.0.15
3. permit ip any any
```

c) **Teste 3** – *Router* interno chegar ao concentrador (10.1.1.1)

```
Sem VACL:

Type escape sequence to abort.
Sending 5, 100-byte ICMP Echos to 10.1.1.1, timeout is 2 seconds:
!!!!!
Success rate is 100 percent (5/5), round-trip min/avg/max = 4/4/4
ms

Com VACL:

Type escape sequence to abort.
Sending 5, 100-byte ICMP Echos to 10.1.1.1, timeout is 2 seconds:
Success rate is 0 percent (0/5)

Ping para router externo:

Type escape sequence to abort.
Sending 5, 100-byte ICMP Echos to 198.5.6.1, timeout is 2 seconds:
Success rate is 100 percent (5/5), round-trip min/avg/max =
100/171/192 ms
```

A VACL não afeta o tráfego gerado entre servidores e *firewall*, apenas previne os *routers* de enviarem tráfego dos servidores para a mesma VLAN e negar tráfego da rede interna para os utilizadores de VPN.

6. Conclusões e Trabalho Futuro

A presente dissertação de Mestrado tinha como objetivo o de estudar uma abordagem em *cloud computing* para o centro de dados do caso em estudo, recorrendo a tecnologias de virtualização, armazenamento partilhado, redes locais e de dados, segurança assim como *backups.* O trabalho desenvolvido passou pelo Estado da Arte onde foram estudadas as tecnologias nesta área. Foi abordado o SO em *cloud* e analisado o impacto de um modelo de infraestrutura partilhada, nomeadamente, tolerância a falhas, técnicas de alta disponibilidade, arquiteturas virtualizadas, segurança na DMZ virtualizada e escalabilidade numa infraestrutura dinâmica.

Foi efetuada uma análise ao ambiente de estudo de modo a validar as métricas com estatísticas dos serviços, nomeadamente, consumos de processamento, memória, armazenamento e rede de modo a enquadrar numa abordagem partilhada. A análise foi elaborada através do *software* cedido pela VMWare, o VMWare Capacity Planner que recolheu dados durante 8 meses permitindo calcular taxas de crescimento, taxas de utilização e desempenho dos sistemas. Posteriormente foi proposta uma arquitetura em *cloud* privada disponibilizando IaaS ao caso prático recorrendo a tecnologias identificadas para este novo paradigma.

Pôde-se validar que o serviço de *mail* do caso em estudo era o serviço mais exigente de recursos no qual teve um foco maior nos testes elaborados.
Depois de validado o ambiente, foi possível confirmar que o consumo de processamento médio era inferior a 2% com picos de 8% e que seria possível consolidar o centro de dados com 21 servidores físicos em apenas 2, com características adequadas à carga e com tolerância a falhas na infraestrutura. Verificou-se ainda que a abordagem de *cloud computing* e IaaS permite uma poupança energética na ordem dos 71% (o equivalente a poupar 4,4 toneladas de CO2 por ano) e reduz em 73% o espaço ocupado com infraestrutura.

Numa perspetiva financeira foi estimado que a aplicação da proposta face à manutenção dos equipamentos actuais permite um retorno de investimento a 3 anos de 35% e que incluindo uma solução global de *backups* permitia igualmente a 3 anos um retorno de investimento de 15%. Técnicas de consolidação e eficiência permitem ainda reduzir o armazenamento em 60% e reduzir portas de rede em 35%, sendo que esta nova abordagem apresenta margem de escalabilidade em todos os recursos disponíveis.

Verificou-se igualmente que a virtualização introduz *overhead* estimado de 5% no serviço de *mail* assim como a memória e processamento dos sistemas é fundamental visto que a virtualização introduz componentes que anteriormente eram físicas e passaram a ser lógicas (ex: *switches* virtuais como o Cisco Nexus 1000v) e que a margem de CPU e memória são determinantes para que não existam problemas de desempenho no centro de dados virtualizado. A latência de rede subiu exponencialmente a partir de 60% de utilização dos recursos de processamento. Em algumas situações validou-se que a taxa de transferência entre servidores virtuais ultrapassa o valor da interface física e que a comunicação entre físico e virtual consegue esgotar a capacidade da ligação de rede. A mobilidade de servidores virtuais possui maior impacto na rede LAN devido ao mapeamento da memória em utilização, tendo-se atingindo aproximadamente 2 minutos para migrar a aplicação com utilização mais intensiva.

A abordagem de *backups* com deduplicação permite proteger todo o ambiente reduzindo em 85% o volume de dados a salvaguardar diminuindo o consumo de recursos. A introdução do protocolo FC permite não só reduzir as latências das transacções mas como proporcionar maior largura de banda para o tráfego de *backups* ser efetuado na janela de tempo que não impacte período laboral. Verificou-se que o impacto na rede de dados diminui 38% no primeiro *backup*, e mais de 90% nos restantes.

A nível de segurança implementou-se uma arquitetura com segmentação interna para reforçar a segurança na camada física e para minimizar possíveis ataques á DMZ implementou-se segurança através das PVLANs e VACLs definidas no *switch* virtual Cisco Nexus 1000v.

Relativamente a proposta de trabalho futuro, esta dissertação poderia ser melhorada através da inclusão de:

a) Abordagem de uma *cloud* com unificação de redes, recorrendo ao protocolo FCoE com métricas de desempenho,
b) Replicação de dados para site remoto via WAN com técnicas de optimização de rede, orquestração e sincronismo entre centros de dados remotos e integração da *cloud* privada com *cloud* públicas ou híbridas,
c) Virtualização de *desktops* recorrendo a protocolos de optimização de rede como o PCoIP dentro de um centro de dados virtualizado,
d) Migração *online* de servidores virtuais entre centros de dados remotos e respectivos impactos na rede WAN,

Referências

[1] IDC Cloud Research [Online] [Fevereiro de 2012], www.idc.com/prodserv/idc_cloud.jsp

[2] Gantz, John F. (2008). *An updated forecast of WorldWide Information Growth Through 2011*. White Paper, IDC sponsored by EMC, Boston.

[3] EMC Corporation, EMC Unified Storage Fundamentals for Performance and Availability, Applied Best-Practices, 2011

[4] VMWare, Reducing Server total cost of ownership with VMWare Virtualization Software White Paper, 2006

[5] Formal Definition of Cloud Computing by NIST [Online] [Julho de 2009], http://thecloudtutorial.com/nistcloudcomputingdefinition.html

[6] Rasmussen, Neil e Niles, Suzanne (2007). *DataCenter Projects: System Planning*. White Paper #142, APC, America.

[7] Types of server virtualization [Online] [Julho de 2008], http://it.toolbox.com/wiki/index.php/Server_Virtualization

[8] VMWare, Virtual NetWorking Concepts, Information guide, 2007

[9] VMWare Corporation, vSphere vMotion Architecture, Performance and Best Practices, White Paper, 2011

[10] VMWare Corporation, VMWare High Availability Concepts, Implementation and Best Practices, White Paper, 2007

[11] VMWare Corporation, vSphere Availability, Technical guide, 2011

[12] VMWare, Network Segmentation in Virtualized Environments Bes-Practices With-Paper, 2009

[13] Fibre Channel SAN Topologies V2.1 [Online] [2011], http://www.emc.com/collateral/hardware/technical-documentation/h8074-fibre-channel-san-tb.pdf

[14] NetApp Thin Provisioning – Increase storage utilization (2011). Technical report, Carlos Alvarez, 2009.

[15] EMC Education Services (2008). EMC NetWorker Administration for Unix and Windows Student Guide, EMC Corporation, Boston, America.

[16] LTO Ultrium Datasheet (2010), ULTRIUM LTO, Hewlett-Packard, IBM and Quantum, America

[17] Storage Engeneering Team (2009), Data De-duplication and its Benefits WhitePaper, MindTech, Bangalore, India

[18] Data Reduction: Realizing the benefits of deduplication and compression, Storage Strategies Now, Patrick Corrigan and Deni Connor, 2011, America

[19] VMWare Corporation, vSphere Data Protection, Technical White Paper, 2012

[20] VMWare Corporation, vShield App Design Guide, Technical White Paper, 2011

[21] VMWare Corporation, vShield Edge and vShield App Reference Design Guide, Technical White Paper, 2010

[22] VMWare Corporation, What´s New in VSphere 4, White Paper, 2009

[23] VMWare Corporation, VMWare Capacity Planner – Optimize Business and capacity planning, White Paper, 2009

[24] NetApp Corporation, NetApp Support global Services, Manual, 2009

[25] NetApp Corporation, Storage Subsytem Resiliency Guide, Technical Report, Jay White, 2011

[26] VMWare Corporation, Understanding memory resource management, VMWare Corp, Technical White Paper, 2011

[27] VMWare Corporation, VMWare vCenter Converter, Users Guide, 2011

[28] NetApp Corporation, NetApp SnapManager for Microsoft Exchange, Manual, 2010

[29] Brocade Corporation, Brocade 300 Switch, Specification Sheet, 2008

[30] Cisco Corporation, Cisco Nexus 1000v Series Switches, DataSheet, 2012

[31] VMWare Corporation, Network I/O Latency – Performance Study, Technical White Paper, 2012

[32] VMWare Corporation, Networking Performance, Technical White Paper, 2011

[33] VMWare Corporation, VSphere: Exchange Server on NFS, iSCSI and Fibre Channel, White Paper, 2009

[34] VMWare Corporation, Microsoft Exchange Server 2007 – Performance on VMWare , White Paper, 2009

[35] VMWare Corporation, vMotion architecture, performance and best-practices – Performance Study, Technical White Paper, 2009

[36] EMC Corporation, EMC Backup and Recovery for Microsoft Applications Deduplication enabled by EMC Clariion and Data Domain, White Paper, 2010

[37] Cisco Corporation, DMZ virtualization using VMWare vSphere and Cisco Nexus 1000v, White Paper, 2009

[38] Securing networks with Private VLANs and VLAN Access Control Lists [Maio de 2008], Cisco Systems, http://www.cisco.com/en/US/products/hw/switches/ps700/products_tech_note09186a008013565f.shtml

Anexo I – Requisitos VMWare Capacity Planner

Neste anexo abordam-se os requisitos necessários de modo a desenvolver o *assessment* á infraestrutura atual. Na metodologia a abordar neste projecto de mestrado, para além da prévia análise de requisitos, o *assessment* é fundamental para proporcionar dados reais para elaboração de uma arquitetura para as necessidades actuais e futuras.

O VMWare Capacity Planner permite uma análise ao ambiente de modo a reunir dados estatísticos de performance, utilização cache, processamento, disco, etc.. de modo a se poder estipular a quantificação da consolidação dos servidores físicos em virtuais garantindo níveis de performance atuais e futuros assim como assegurar uma infraestrutura que possa escalar no futuro.

Pré-requisitos

- Ficheiro XLS com *hostname* dos servidores a analisar (mapeado com outros dados dos servidores (ie.: função do servidor, *hardware*, SO., etc)).
- Credenciais: preferencialmente recorrer a uma conta global que seja *Domain Admin* (caso não seja possível, recorrer a um *Administrator* local, servidor por servidor). Será necessário introduzir as *passwords* das contas em causa, devendo os mesmos ser inseridos por um administrador de sistemas do ambiente em análise que acompanhe a instalação do VMWare CP.
- 1 ou 2 servidores Windows 2008 ou Windows 2003 SP2 *server English Edition* (instalação da aplicação VMWare Capacity Planner).
    - 2 GB RAM,
    - 2 CPU,
    - 4 GB espaço livre em disco (D-drive),
    - 100 Mb/s NIC,
    - Porto 80 aberto para a Internet (para o Collector),
- Acessos à *Internet*, pois dados irão sendo actualizados para optimize.vmware.com no porto 443 TCP (permissão *firewall*),
- Relativamente a regras de *firewall*, é importante que o tráfego de rede não seja bloqueado entre o servidor de Capacity Planner e os servidores a analisar; sendo a sugestão habilitar tráfego bidireccional sem restrições, entre o servidor de Capacity Planner e os servidores em análise,

- Não é necessário instalar agentes; os serviços consultados pelo Capacity Planner são: - *Windows Management Instrumentation* (WMI) - *Remote Registry* - *Performance Monitor* (perfmon) - *File and print services* - Windows NT (deverá ter o diskperf ativo para recolher estatísticas de disco). Nos Linux, será usado SSH e a aplicação irá executar comandos típicos do sysstat/SAR (top, mpstat, etc). Estes processos deverão estar ativos e a correr em todos os servidores a serem analisados,
- Existindo servidores na DMZ a serem monitorizados, será necessário um *Collector* em separado a ser instalado,
- Mínimo ligação á Internet de 20 KByte/s, podendo ser via *Proxy*.

Anexo II – Resumo das tecnologias propostas

A proposta de infraestrutura assenta quer em *hardware* como em *software* nos seguintes componentes:

- Armazenamento partilhado e rede de dados:
    - NetApp FAS2240:
        - 4 portos FC a 8 Gb/s,
        - 28 discos SAS de 300 GB@15Krpm,
        - 14 discos SAS de 450 GB@15Krpm,
        - SnapManager+Snapshots+SnapRestore,
    - Switches SAN:
        - 2xBrocade DS300B
        - 8 portas activas a 8 Gb/s,
        - 12 fibras ópticas LC-LC,
- Infraestrutura de virtualização:
    - Cluster de virtualização:
        - 2 x CPU Quad-Core 2,86 GHz,
        - 96 GB memória,
        - 3 NIC ToE de quatro portas 1 Gb,
        - 2 HBAs com uma porta 8 Gb,
        - 2 discos de 146 GB@15k,
    - Licenciamento VMWare ESX:
        - vCenter Server Foundation,
        - 4 x Enterprise Plus,
    - Switch virtual:
        - Cisco Nexus 1000v,
- Infraestrutura de *backup*:
    - VTL:
        - EMC Data Domain DD640,
        - 12 TB capacidade bruta,
        - 2 portas FC a 8 Gbit,
    - *Software* gestão *backups*:
        - EMC NetWorker,
        - 2 x Virtual Edition Client,

        - Agentes online,
        - VTL,
        - Servidor *proxy*,
    - *Proxy backup*:
        - 2 x CPU Quad-Core 2,86 GHz,
        - 32 GB memória,
        - 2 NIC ToE de duas portas 1 Gb,
        - 2 HBAs com uma porta 8 Gb,
        - 2 discos de 146 GB@15k,