

MO655 - Comunicação em Datacenters

Leandro Souza da Silva - RA: 191082

Luís Felipe Mattos - RA: 107822

15 de Dezembro de 2016

Conteúdo

1	Introdução	2
1.1	Requisitos de Rede	3
1.1.1	Escalabilidade	3
1.1.2	Tolerância a Falhas	4
1.1.3	Latência	4
1.1.4	Capacidade da Rede	4
1.1.5	Virtualização	4
2	Motivação	5
3	Topologias	6
4	Protocolos	7
4.1	Roteamento	7
4.1.1	Equal Cost Multipath (ECMP)	7
4.1.2	VL2	8
4.1.3	Distributed Adaptive Routing for Datacenter Networks (DRAD)	8
5	Tendências	9
5.1	Facebook Fabric	9
6	Conclusão	12
7	Referências	13

1

Introdução

Com o crescimento da demanda dos usuários por poder computacional e armazenamento, cada vez mais as grandes empresas estão investindo em estruturas próprias de datacenters. Esta estrutura inclui tanto os computadores em si, os discos de armazenamento e os racks como também inclui a própria sala que ficarão estes racks. Estas salas devem ter uma arquitetura própria, como por exemplo, o piso elevado, sistema de refrigeração e circulação de ar. Um exemplo pode ser visto na figura 1.1.

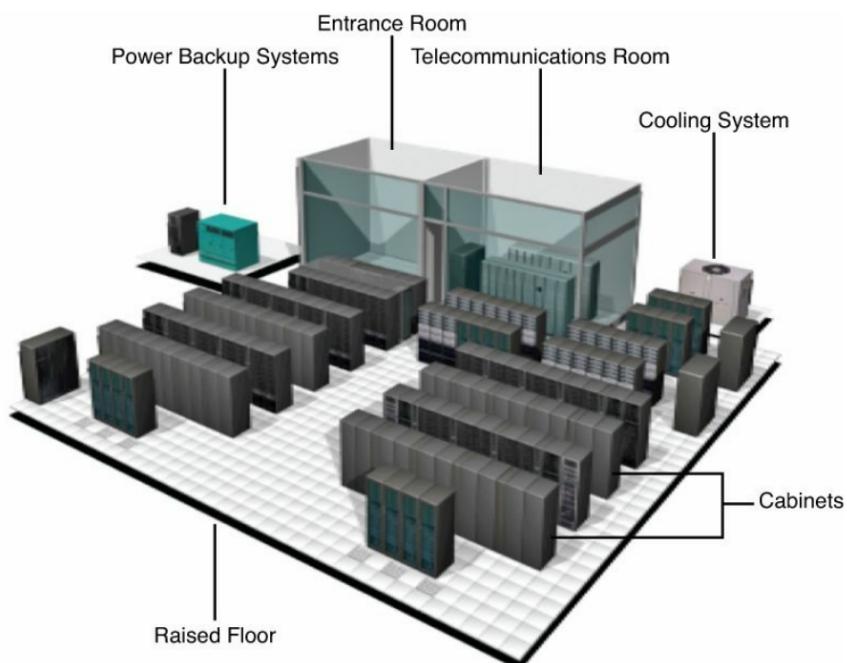


Figura 1.1: Visão geral de um Datacenter

Além da estrutura física, a parte computacional, principalmente relacionada à comunicação interna e externa do datacenter possui alguns requisitos básicos para que possa oferecer um serviço de qualidade para os usuários. Estes requisitos são citados a seguir:

- Escalabilidade
- Tolerância a Falhas

- Latência
- Capacidade da Rede
- Virtualização

A seguir, os requisitos citados serão mais detalhados.

1.1 Requisitos de Rede

1.1.1 Escalabilidade

O sistema deve ser construído de tal forma que seja possível haver uma expansão, caso a demanda aumente. Este requisito diz respeito tanto ao hardware como ao software. Para o hardware, a estrutura das máquinas deve permitir que o sistema seja melhorado e também deve haver espaço físico para a inclusão de novas máquinas. Atualmente, existem alguns sistemas modulares que possuem uma fácil integração de novos módulos.

Um exemplo é a utilização de datacenters em containers, cada container possui um sistema completo com refrigeração própria e é facilmente transportado. Com isso, pode-se expandir facilmente uma estrutura de um datacenter. Um exemplo de container pode ser visto na figura 5.4.

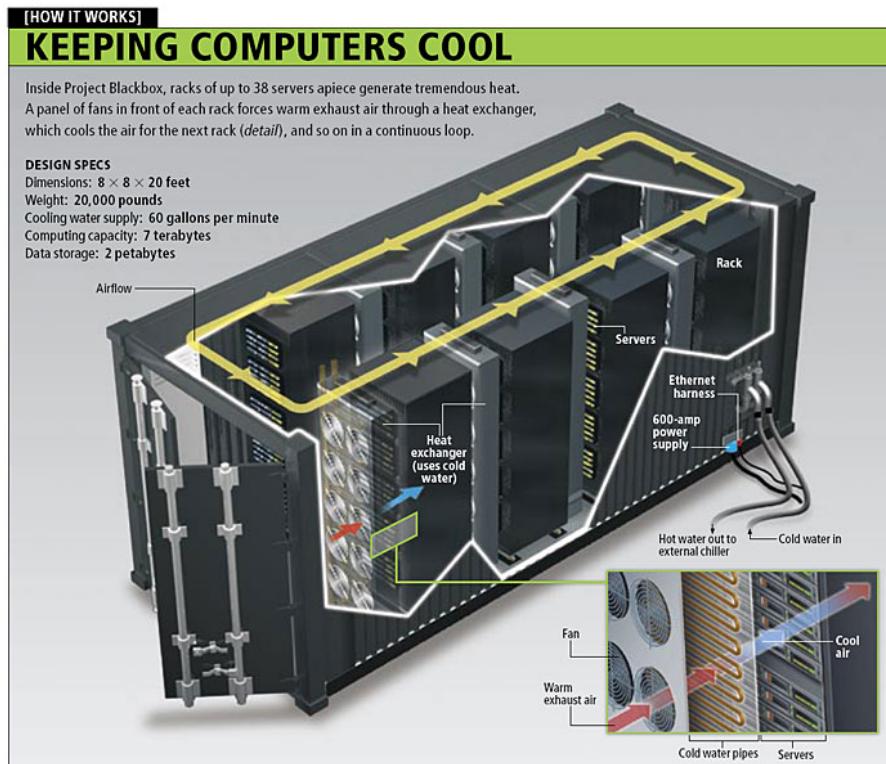


Figura 1.2: Datacenter em container

1.1.2 Tolerância a Falhas

O sistema deve ser capaz de prevenir e corrigir falhas. Por causa disso, a maioria dos sistemas de datacenters possuem redundâncias em quase todos aspectos do datacenter. Existem backups dos dados dos usuários, a comunicação interna é feita de modo que existam vários caminhos possíveis da fonte para o destino e além disso, alguns datacenters possuem backups em outros datacenters. Apesar disso, existe o custo de manter estas cópias atualizadas.

Mais a frente falaremos um pouco mais sobre as redundâncias dos caminhos de comunicação interna dos datacenters, tanto relacionados à topologia como relacionado aos protocolos de comunicação e roteamento.

1.1.3 Latêcia

Um dos principais desafios dos datacenters é possuir baixa latência, assim, a performance do sistema como um todo se mantém em um nível aceitável pelos usuários. Para isso, a topologia é muito importante, uma vez que quanto menor o caminho entre a fonte e o destino, menor a latência. Porém, outro fator que influencia muito a latência é o nível de congestionamento da rede, mas a frente iremos tratar sobre os protocolos e como estes controlam o nível de congestionamento da rede.

1.1.4 Capacidade da Rede

1.1.5 Virtualização

2

Motivação

3

Topologias

4

Protocolos

4.1 Roteamento

Embora a maioria dos esquemas básicos de roteamento busquem rotas entre dois servidores com latência curta, um encaminhamento mais sofisticado exige maior consideração e otimização da latência, confiabilidade, throughput, energia e etc. Esse tipo de otimização é conhecido como problema de Traffic Engineering (TE).

Existem poucos mecanismos para a otimização de roteamento Data Center Network (DCN) hoje em dia. Essa seção apresenta três diferentes esquemas de roteamento, ECMP, VL2 e DRAD.

4.1.1 Equal Cost Multipath (ECMP)

Em um ambiente como datacenters, onde topologias como Fat-Tree, Clos proporcionam múltiplos caminhos, o ECMP é utilizado com intuito de balancear a carga, dividindo a quantidade de pacotes ou fluxos ao longo da topologia pelos seus múltiplos caminhos existentes. O ECMP possui dois modos de operação:

1. Balanceamento baseado em fluxo

Utiliza o hash na tupla de 5 campos de cada pacote e o encaminha para uma interface de saída. Todos os pacotes do mesmo fluxo TCP/IP serão encaminhados pela mesma interface. Este é o modo padrão de funcionamento do ECMP.

2. Balanceamento baseado em pacote

Utiliza o modelo round robin para cada pacote, fazendo com que pacotes do mesmo fluxo sigam por rotas diferentes.

A diferença básica destes dois modos de operação do ECMP impacta diretamente nos protocolos, tais como TCP ou MPTCP. Quando utiliza-se o modo de balanceamento baseado em pacote, pacotes que pertencem ao mesmo fluxo TCP/IP são roteados por rotas diferentes o que pode causar reordenação nos hosts finais e consequentemente diminuir a taxa de transferência.

Quando o balanceamento baseado em fluxos é utilizado, o problema da entrega desordenada não ocorre pois todos os pacotes de um mesmo fluxo TCP/IP seguem a mesma rota. O ECMP, mesmo usando o modelo baseado em fluxo que normalmente se apresenta como melhor solução, ainda sofre com as colisões que ocorrem um vez que soluções baseadas em hashing estatisticamente podem gerar as mesmas saídas quando aplicadas em diferentes entradas.

No caso da utilização do ECMP para balancear fluxos MPTCP, isso pode significar alocar os subfluxos de uma mesma conexão MPTCP em uma mesma rota.

4.1.2 VL2

É outro mecanismo de seleção de caminhos distribuídos. Diferente do ECMP, ele coloca a lógica de seleção nos switches da borda. Em VL2, um switch da borda encaminha primeiro um fluxo a um switch do núcleo, selecionado aleatoriamente, que então encaminha o fluxo para o destino. Como resultado, múltiplos fluxos ainda podem colidir na mesma porta de saída como ECMP.

4.1.3 Distributed Adaptive Routing for Datacenter Networks (DARD)

Difere do ECMP e VL2 em dois aspectos. Primeiro, seu algoritmo de seleção de caminho é sensível à carga. Se vários fluxos colidem no mesmo caminho, o algoritmo deslocará os fluxos do caminho colidido para os caminhos mais ligeiramente carregados. Em segundo lugar, coloca a lógica de seleção de caminhos em sistemas finais, em vez de em switches, para facilitar a implantação.

Um módulo de seleção de caminho que executa em um sistema de extremidade monitora o estado do caminho e comuta-o de acordo com a carga. Uma DCN pode implantar o DARD atualizando a pilha de software do seu sistema final em vez de atualizar switches.

5

Tendências

5.1 Facebook Fabric

Para a próxima geração de projeto de rede de data center, O Facebook se desafiou a fazer todo o data center construindo uma rede de alto desempenho, em vez de um sistema hierarquicamente super-assinado (oversubscription) de clusters. O Facebook Também queria um caminho claro e fácil para a rápida implantação de rede e escalabilidade de desempenho sem perder as infraestruturas maciças já construídas, sempre que precisasse de mais capacidade.



Figura 5.1: Data Center do Facebook em Altoona Pennsylvania - EUA

Para conseguir isso, foi utilizada uma abordagem desagregada: em vez dos grandes dispositivos e clusters, a rede foi quebrada em pequenas unidades idênticas, pods de servidor, e foi conectividade uniforme de alto desempenho entre todos os pods do data center.

Cada pod é servido por um conjunto de quatro dispositivos que são chamados de fabric switches, utilizando a arquitetura de 4-postos 3 + 1 para uplinks TOR de rack de servidor e escalável além disso, se necessário. Cada TOR tem 4 x 40G uplinks, fornecendo 160G de capacidade total de largura de banda para um rack de servidores conectados a 10G.

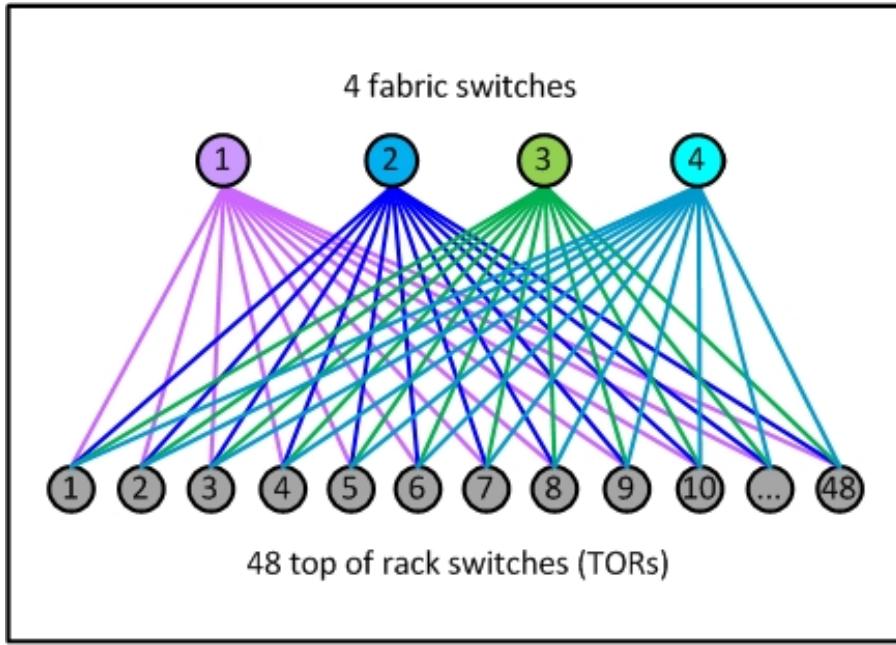


Figura 5.2: Um pod simples - nova unidade de rede do Facebook Fabric

O que difere das arquiteturas anteriores é que cada pod tem apenas 48 racks de servidor, e este fator é sempre o mesmo para todos os pods. É um bloco de construção eficiente que se encaixa bem em várias plantas de centro de dados, e requer apenas switches básicos de tamanho médio para agregar os TORs. A menor densidade de portas dos switches de fabrica torna essa arquitetura interna muito simples, modular e robusta, e há várias opções de fácil acesso disponíveis em várias fontes.

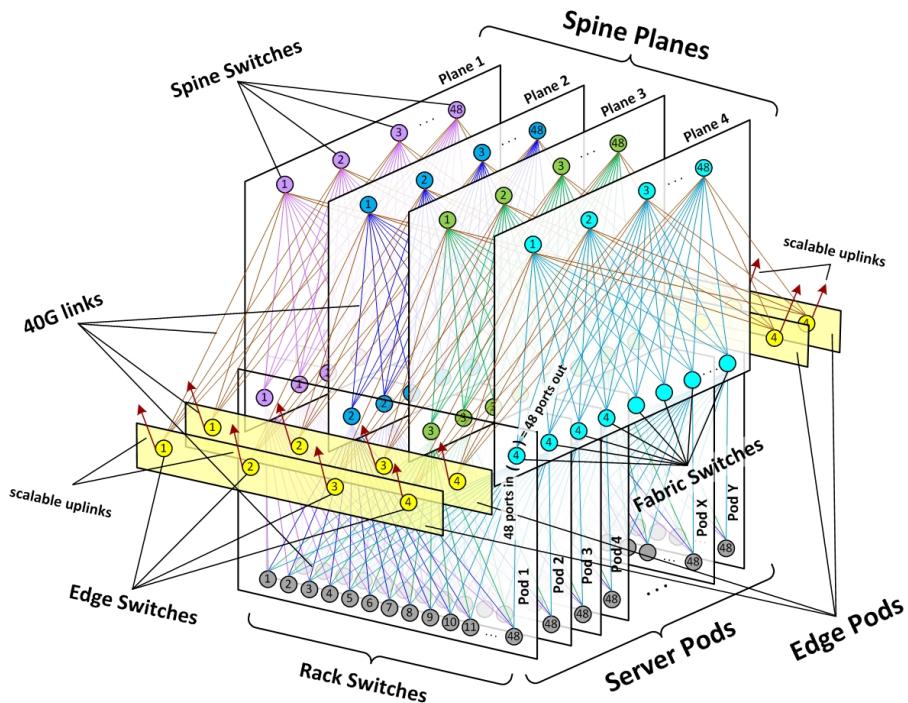


Figura 5.3: Esquemático da topologia do Fabric Data Center

Este design altamente modular permite dimensionar rapidamente a capacidade em qualquer dimensão, dentro de uma estrutura simples e uniforme.

O Fabric foi construído usando padrão BGP4 como o único protocolo de roteamento. Para manter as coisas simples, foi usado apenas os recursos mínimos de protocolo necessários. Isso nos permitiu aproveitar o desempenho e a escalabilidade de um plano de controle distribuído para convergência, oferecendo gerenciamento de propagação de roteamento rígido e granular e garantindo compatibilidade com uma ampla gama de sistemas e software existentes. Ao mesmo tempo, foi desenvolvido um controlador BGP centralizado que é capaz de substituir quaisquer caminhos de roteamento no fabric por decisões de software puro.

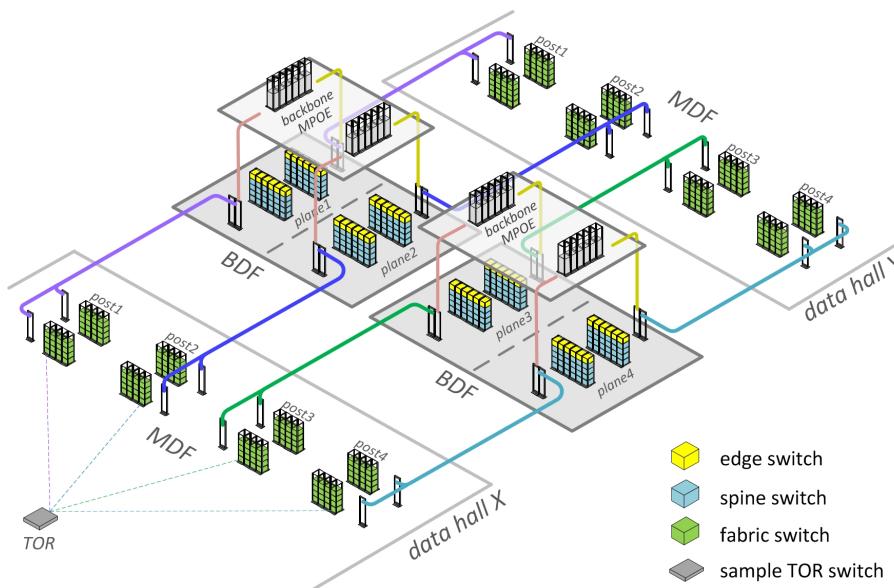


Figura 5.4: Esquemático otimizado da topologia física do Fabric Data Center

Apesar da grande escala de centenas de milhares de fios de fibra, a infraestrutura física e de cabeamento é muito menos complexo do que pode parecer nos desenhos de topologia de rede lógica. Os projetos de construção do Fabric foram otimizados, para encurtar comprimentos de cabeamento e permitir uma rápida implantação. Altoona foi a primeira implementação deste novo tipo de layout.

6

Conclusão

Referências

- Manual de referência do NS-3

<https://www.nsnam.org/docs/release/3.8/manual.pdf>

- Documentação

<https://www.nsnam.org/doxygen/index.html>