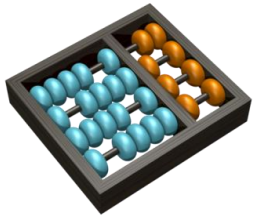


UNICAMP

Topologias de Data Centers

Valdomiro Luis Scannapieco Neto

23/10/2013

Agenda

- Introdução
- O que é Cloud Computing?
- Topologia Convencional de Data Centers
- Objetivos e Requisitos das Novas Topologias
- Novas Topologias
  - Monsoon
  - VL2
  - Portland
  - Bcube
  - Dcell
- Considerações Finais
- Referências
- Questão

Introdução

- Cloud Service: novo modelo de computação que surgiu e alterou a forma como interagimos com a rede e com os serviços e aplicações.

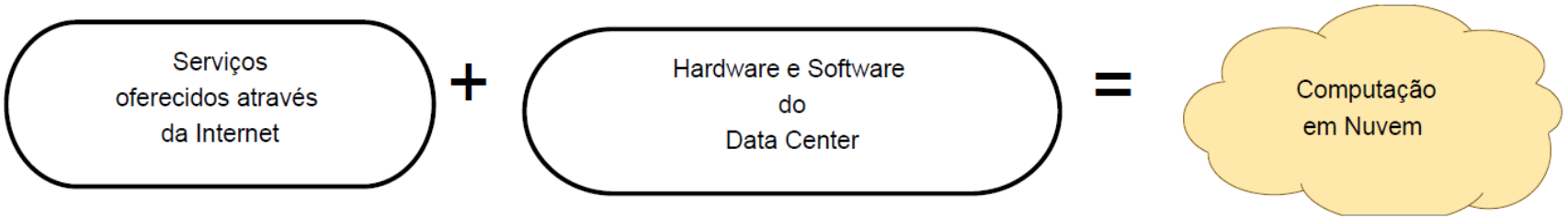
Introdução

- Aplicações são hospedadas e executadas remotamente.
- Comunicação centrada no servidor.
- O desenvolvimento de aplicações é mais simples, pois as mudanças e as melhorias são realizadas em um único local.
- A computação se torna mais barata quando vista como um serviço compartilhado.
- Evolução das técnicas de virtualização.
- Estas diferentes aplicações requerem diferentes arquiteturas de data centers, motivando a pesquisa e o desenvolvimento de soluções que atendam os seguintes requisitos:
  - Escalabilidade
  - Alto desempenho
  - Custo mínimo

O que é Cloud Computing?

O que é Cloud Computing?

- “Cloud computing é um conjunto de recursos virtuais facilmente usáveis e acessíveis tais como hardware, plataformas de desenvolvimento e serviços. Estes recursos podem ser dinamicamente reconfigurados para se ajustarem a uma carga variável, permitindo a otimização do uso dos recursos. Este conjunto de recursos é tipicamente explorado através de um modelo pay-per-use com garantias oferecidas pelo provedor através de acordos de nível de serviço (Service Level Agreements-SLAs). ” [Vaquero et al. 2009]



O que é Cloud Computing?

- Caracteristicas Essenciais:
  - Serviço sob-demanda: funcionalidades computacionais providas automaticamente sem interação humana.
  - Amplo acesso aos serviços: disponíveis através da Internet e acessados via mecanismos padronizados.
  - Resource pooling: recursos alocados e realocados dinamicamente conforme a demanda do usuário.
  - Elasticidade: impressão de recursos ilimitados.
  - Medição dos serviços: sistemas de gerenciamento que monitoram automaticamente o uso dos recursos.

O que é Cloud Computing?

- Modelo de Serviços:
  - SaaS: aplicações passam a ser hospedadas na nuvem, como uma alternativa ao processamento local. Todo o controle e gerenciamento da infraestrutura de rede, sistemas operacionais, servidores e armazenamento é feito pelo provedor do serviço.

O que é Cloud Computing?

- Modelo de Serviços:
  - PaaS: capacidade oferecida pelo provedor para o usuário desenvolver aplicações que serão executadas e disponibilizadas na nuvem.

O que é Cloud Computing?

- Modelo de Serviços:
  - IaaS: capacidade que o provedor tem de oferecer uma infraestrutura de processamento e armazenamento de forma transparente. O usuário não tem o controle da infraestrutura física mas, através de virtualização, possui controle sobre os sistemas operacionais, armazenamento, aplicações e um controle limitado dos recursos da rede.

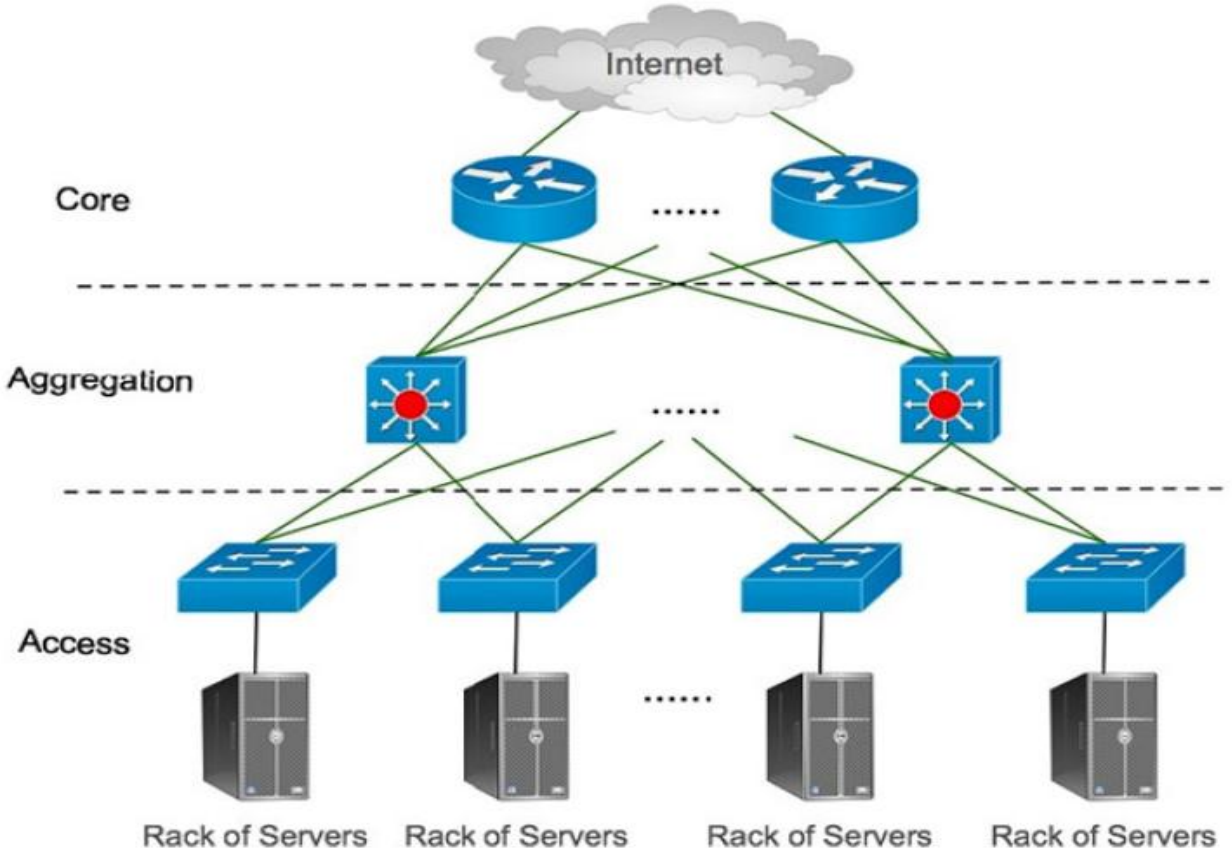
O que é Cloud Computing?

- Novos aspectos em relação ao hardware:
  - Ilusão de recurso computacional infinito disponível sob-demanda.
  - Eliminação de um comprometimento antecipado por parte do usuário.
  - Capacidade de alocar e pagar por recursos usando uma granularidade de horas.

Topologia Convencional de Data Centers

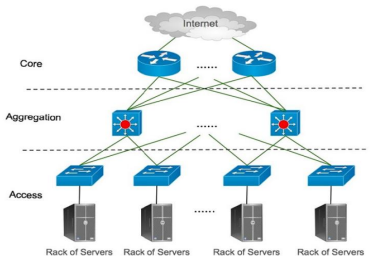
Topologia Convencional de Data Centers

- Dividida em camadas: núcleo, agregação e acesso.



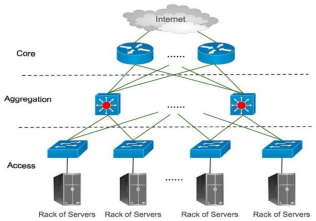
Topologia Convencional de Data Centers

- Camadas:
  - Camada de acesso: onde os servidores, organizados em racks, se conectam fisicamente à rede. Tipicamente existem de 20 a 40 servidores por rack e cada rack é conectado a um switch de acesso por um link de 1 Gbps. Cada switch de acesso geralmente conecta-se a dois switches de agregação por links de 10 Gbps. Links redundantes.
  - Camada de agregação: usualmente provê importantes funções como: balanceamento de carga, serviços de domínio e localização.
  - Núcleo: provê conectividade a múltiplos switches de agregação e resiliente fábrica de rotas sem ponto único de falhas. Conexão com o "mundo" fora do data center.



Topologia Convencional de Data Centers

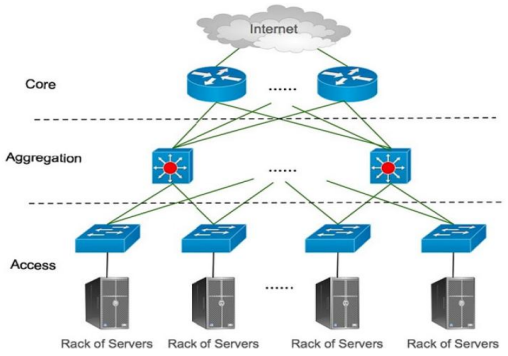
- Funcionamento:
  - Cada aplicação está associada a um (ou mais) IP público através do qual clientes na internet mandam suas requisições e recebem as respostas. Esse IP público é chamado de IP virtual (VIP) e é mapeado, dentro da camada núcleo, para um ou mais IPs direto (DIPs) que referenciam de fato os servidores responsáveis por receber e responder requisições dessa aplicação.
  - As requisições que chegam da internet são IP (camada 3). Ao atingirem a camada de agregação essas requisições trafegam por um domínio de camada 2 (layer 2 domain). Na prática um domínio de camada 2 é limitado a, no máximo, 4.000 servidores, devido à limitação das arquitetura de rede e dos protocolos convencionais.
  - Dentro de cada domínio de camada 2 devemos ficar atentos ao overhead do trafego de broadcast, como do ARP (IP -> MAC), por isso cada domínio de camada 2 precisa ser dividido em sub-redes.



Topologia Convencional de Data Centers

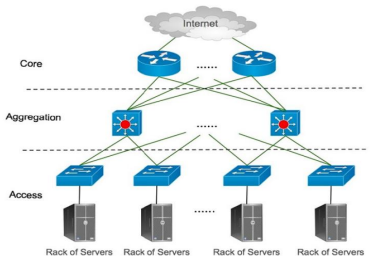
- Problemas com os projetos convencionais:
  - As práticas atuais de engenharia do data center e, em especial, a organização hierárquica L2/L3 da arquitetura de rede, causam uma série de problemas que dificultam a agilidade do sistema e apresentam as seguintes limitações:

  - Fragmentação dos recursos;
  - Alocação estática de serviços;
  - Vazão e latência entre servidores;
  - Escalabilidade;
  - Custo dos equipamentos de rede;
  - Eficiência energética.



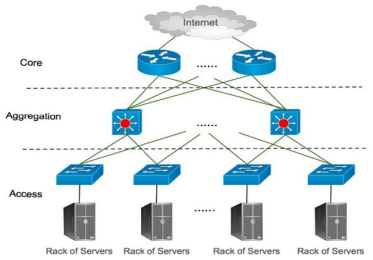
Topologia Convencional de Data Centers

- Fragmentação dos recursos (VLAN, sub-redes):
  - Atribuição de endereços IP que possuem significado topológico, criam um cenário de fragmentação dos servidores. Tal fragmentação introduz rigidez na migração de máquinas virtuais, uma vez que exige a troca do endereço IP para aderir a nova posição topológica.
  - Se uma determinada aplicação cresce e necessita de mais recursos (servidores), tal aplicação não pode utilizar servidores disponíveis e ociosos localizados em outro domínio (sub-rede), causando uma subutilização dos recursos do data center.



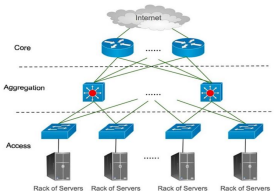
Topologia Convencional de Data Centers

- Alocação estática de serviços:
  - As aplicações ficam mapeadas a determinados switches e roteadores, dificultando a alocação dinâmica de serviços.
  - Quando um determinado serviço enfrenta sobrecarga, a intensa utilização da rede faz com que os serviços que compartilham a mesma sub-rede também sejam afetados.



Topologia Convencional de Data Centers

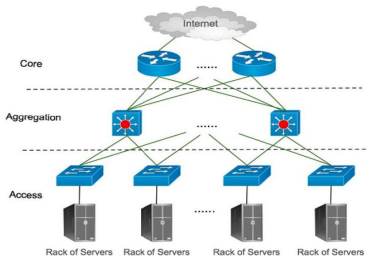
- Vazão e latência entre servidores:
  - As arquiteturas convencionais não oferecem capacidade de transmissão suficiente entre os servidores.
  - A latência e a vazão se tornam não uniformes dependendo do par de servidores.
  - Devido à natureza hierárquica da rede, a comunicação entre servidores localizados em diferente domínios de camada 2 deve ocorrer através de roteamento em camada 3. Para isso, os roteadores do topo da hierarquia devem possuir alta capacidade de processamento e grandes buffers, aumentando ainda mais os custos de TI do data center.
  - Fator 5 de oversubscription na camada de acesso e concentração de tráfego no mais alto nível da árvore variando entre 80 e 240.



Topologia Convencional de Data Centers

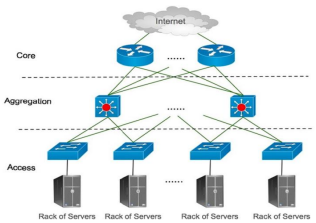
- Fator de oversubscription:
  - Refere-se á prática de realizar a multiplexação dos recursos de banda para fazer um dimensionamento adequado, que economize a quantidade de enlaces e equipamentos.

  - Assumindo um rack com 40 servidores, cada um interligado a uma porta de um 1 Gbps em um switch de 48-portas de 1 Gbps, restarão apenas 8 portas disponíveis para se interligar com a camada de switches de agregação. Se todos os servidores quisessem transmitir no máximo da capacidade da interface de rede (40 Gbps no total), o tráfego agregado nos uplinks seria, no melhor dos casos, de apenas 8 Gbps, o que corresponde a um fator 5 de oversubscription.



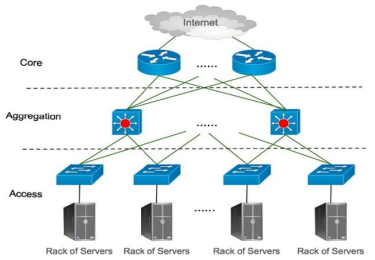
Topologia Convencional de Data Centers

- Escalabilidade:
  - Os protocolos atuais de roteamento, encaminhamento e gerenciamento não oferecem a escalabilidade necessária para atingir a ordem de grandeza necessária nos data centers.
  - As soluções atuais de camada 2 utilizam de broadcast, criando um cenário não escalável devido ao elevado nível de sinalização.
  - As soluções de camada 3 exigem a configuração dos equipamentos (definição de sub-redes). Um data center com 100.000 servidores, cada um executando 32 máquinas virtuais, se traduz em mais de 3 milhões de endereços IP e MAC em um único data center.



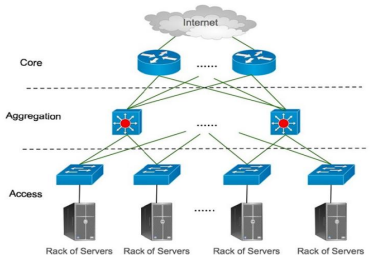
Topologia Convencional de Data Centers

- Custo dos equipamentos de rede:
  - Os balanceadores de carga tradicionais e os switches de auto desempenho são equipamentos caros. Quando estes equipamentos não suportam mais a carga, são substituídos por novos com mais capacidade (scale-up x scale-out).



Topologia Convencional de Data Centers

- Eficiência energética:
  - Os data centers tradicionais usam mecanismos básicos de refrigeração seguindo a premissa de que se o data center cresce, instala-se mais equipamentos de refrigeração, causando um impacto significativo no consumo de energia.
  - Atualmente, equipamentos trabalhando com uma carga mínima, não consome proporcionalmente menos energia do que trabalhando ao máximo de sua capacidade.



Objetivos e Requisitos das Novas Topologias

Objetivos e Requisitos das Novas Topologias

- Objetivos e requisitos:
  - Confiabilidade
  - Desempenho
  - Agilidade: endereçamento e encaminhamento de pacotes que permitam levantar qualquer máquina virtual em qualquer servidor físico – any server, any service.
  - Escalabilidade: poder escalar a uma ordem de centenas de milhares de servidores e milhões de máquinas virtuais. Dois aspectos precisam ser considerados: tamanho das tabelas de encaminhamento e os broadcasts.

Objetivos e Requisitos das Novas Topologias

- Mais objetivos e requisitos:
  - Desempenho: padrão de tráfego entre qualquer par de servidores.
  - Custo: Melhorar os esforços de configuração através de mecanismos automatizados. Usar hardware comoditizado (scale-out).

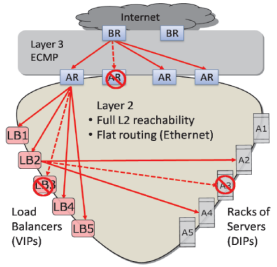
Novas Topologias

Novas Topologias

- Novas topologias foram desenvolvidas afim de atender aos requisitos citados:
  - Monsoon
  - VL2
  - Portland
  - Bcube
  - DCell

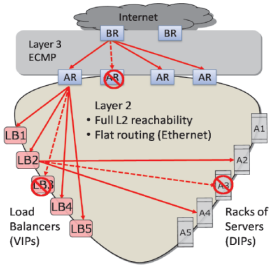
- Algumas arquiteturas privilegiam certos requisitos em detrimento de outros. Entretanto, a maioria das topologias atendem em maior ou menor grau a todos os requisitos.

Novas Topologias



Monsoon

Monsoon

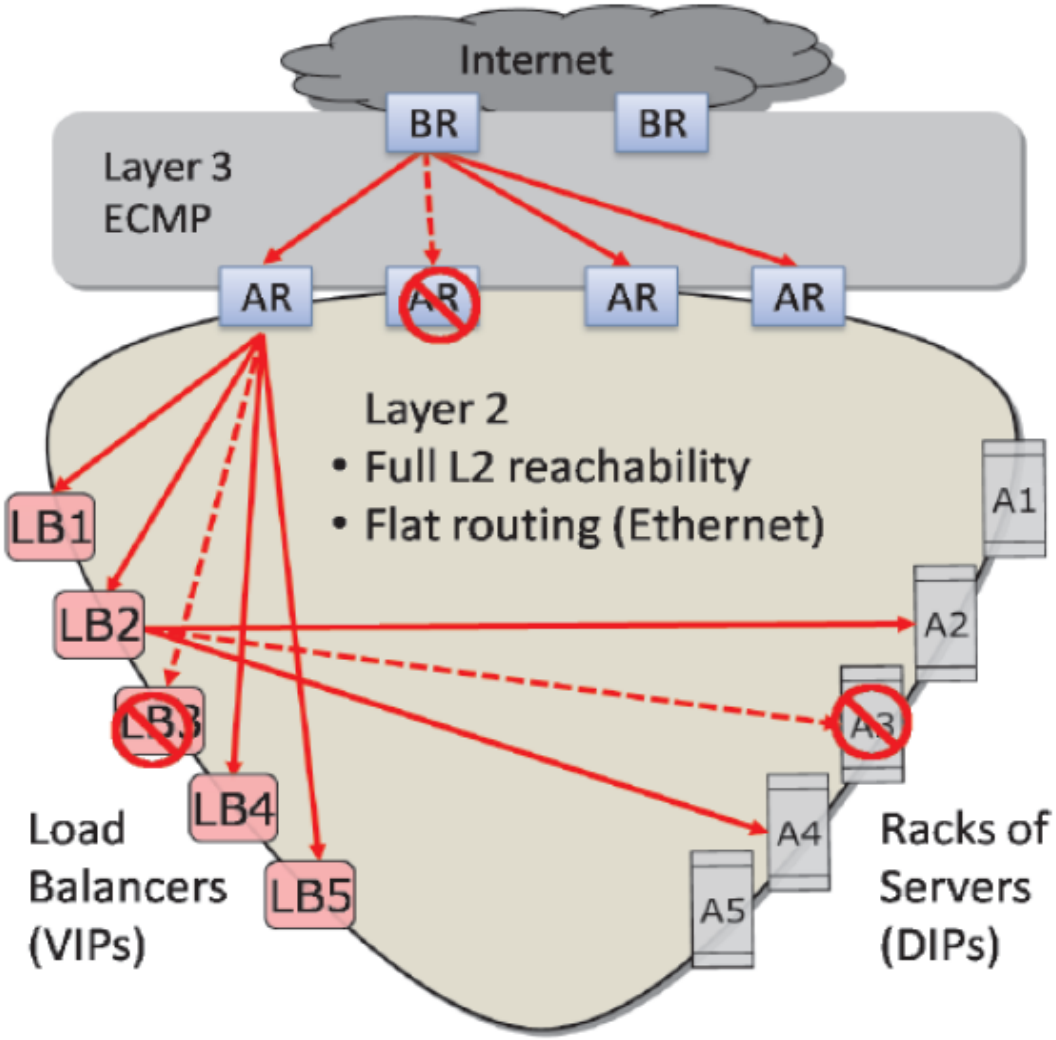


- A filosofia defendida pelo Monsoon é a comoditização como forma de obter escalabilidade e baixo custo, ou seja, o Monsoon adiciona equipamentos novos e baratos (scale-out) para atender a nova demanda.

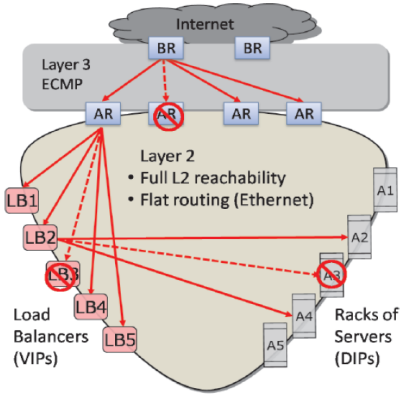
- Estabelece uma arquitetura organizada em formato de malha com o objetivo de comportar 100.000 ou mais servidores. Para criar esta malha são utilizados switches programáveis de camada 2 e servidores.

- Modificações no plano de controle dos equipamentos são necessárias para suportar o encaminhamento de dados por múltiplos caminhos através do *Valiant Load Balancing* (VLB).

Monsoon


Internet
BR
BR
Layer 3
ECMP
AR
AR
AR
AR
Layer 2
• Full L2 reachability
• Flat routing (Ethernet)
LB1
LB2
LB3
LB4
LB5
A1
A2
A3
A4
A5
Load
Balancers
(VIPs)
Racks of
Servers
(DIPs)

Monsoon



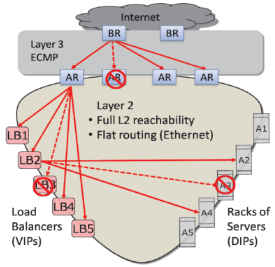
- Características:

- Os dois aspectos principais da arquitetura são:
  - Definição de uma única rede de camada 2 na qual todos os 100.000 servidores são conectados.
  - Flexibilidade pela qual as requisições podem ser distribuídas entre os diversos conjuntos de servidores.

- Fatores que levaram o Monsoon a optar pela camada 2 no único domínio:
  - Cortar custos
  - Eliminar a fragmentação de servidores

- Ethernet foi a tecnologia eleita pois já apresenta custo e desempenho otimizados para o encaminhamento baseado em endereços planos (atualmente uma porta Ethernet custe entre 10% e 50% do valor de uma porta com velocidade equivalente de camada 3).

Monsoon

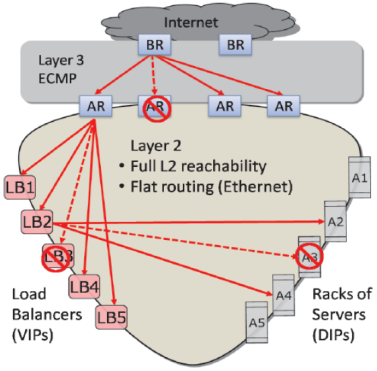


- A comunicação entre os servidores utiliza a taxa máxima das interfaces de rede (1 Gbps).

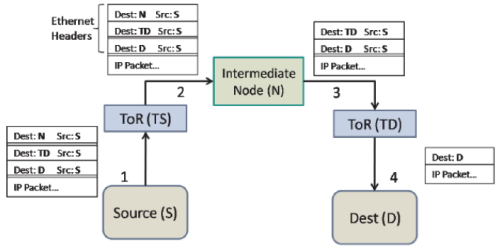
- A camada 3 é necessária para conectar o data center à Internet, utilizando roteadores de borda e o ECMP (Equal Cost MultiPath) para espalhar as requisições igualmente entre os roteadores de acesso.

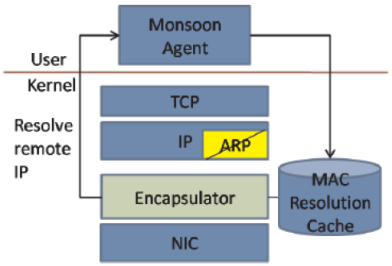
- Os roteadores de acesso utilizam técnicas de hash para distribuir as requisições entre os balanceadores de carga (VIP, público e por aplicação) que, por sua vez, distribuem as requisições para o conjunto de servidores (DIP) de uma determinada aplicação.

Monsoon

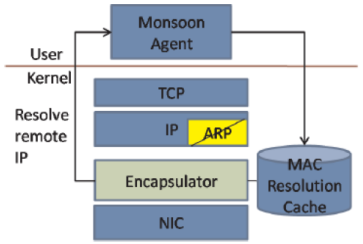


- Encaminhamento Servidor-a-Servidor:
  - Utiliza tunelamento MAC-in-MAC para encaminhar pacotes entre os servidores.
  - É utilizado um serviço de diretório no qual a lista de MAC dos servidores responsáveis pelas requisições, bem como o MAC dos switches nos quais os servidores estão conectados, é mantida.



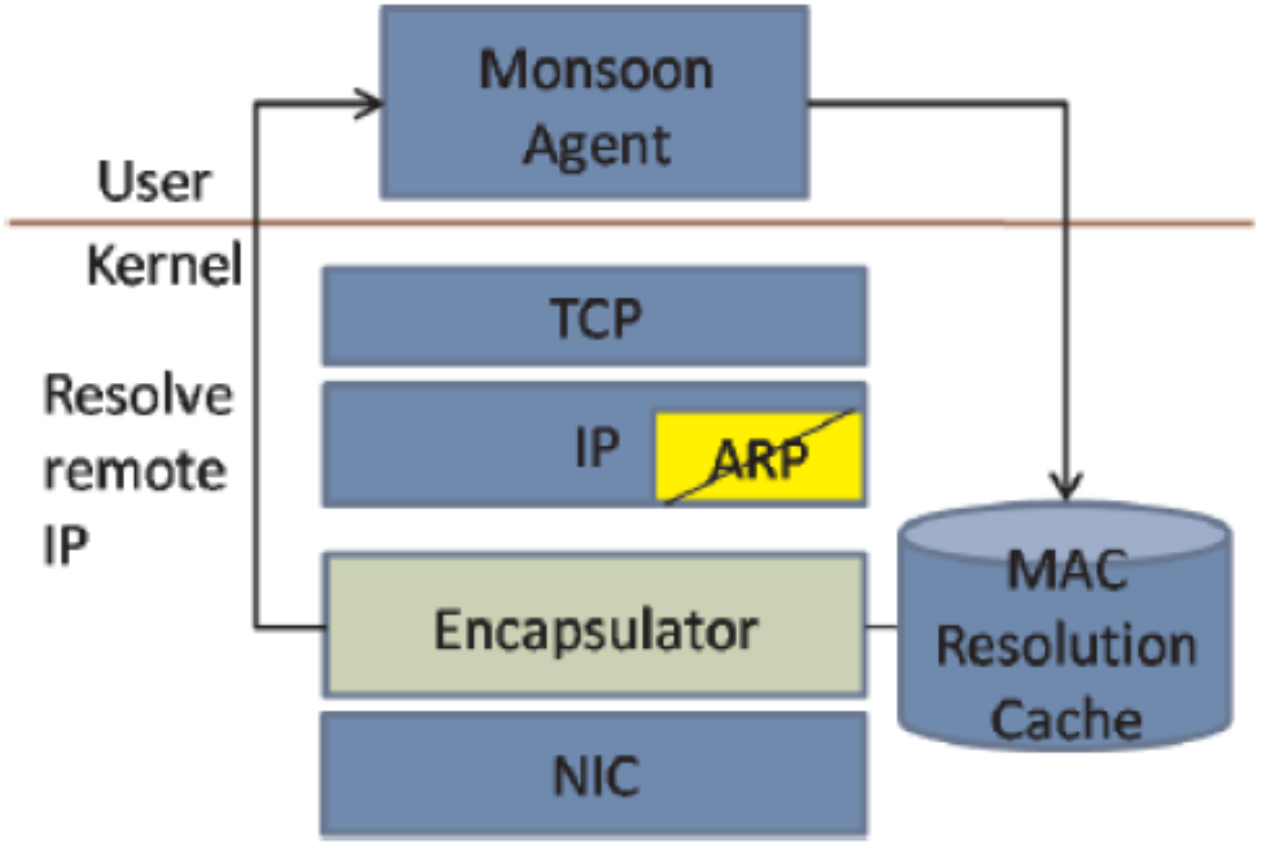


Monsoon

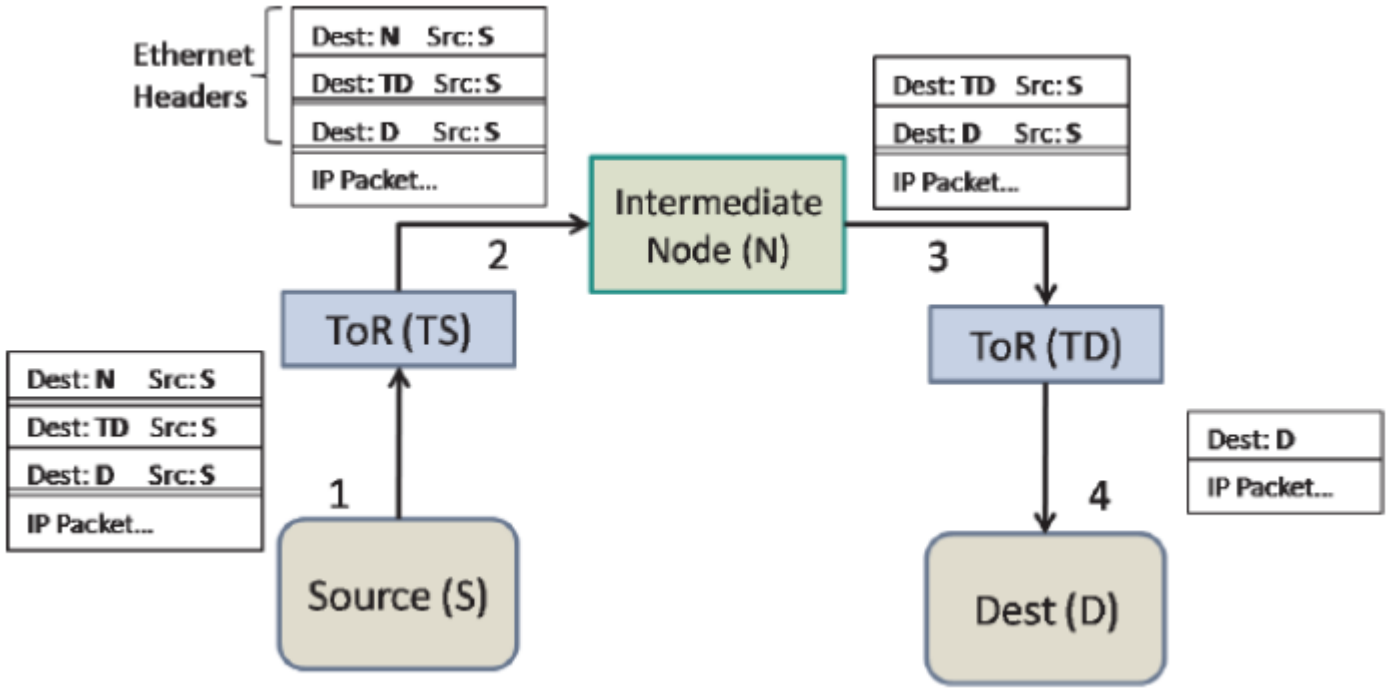


- A tradicional função do ARP é desativada e substituída pelo Monsoon Agent.
- Essas mudanças são imperceptíveis para as aplicações.
- O encapsulador recebe pacotes vindos da camada de rede (IP) e consulta seu cache de resoluções MAC em busca da informação de encaminhamento. Caso não exista, o encapsulador solicita ao agente uma nova resolução através do serviço de diretório.
- O serviço de diretório resolve o endereço IP de destino e retorna uma lista contendo todos os endereços MAC dos servidores associados ao serviço solicitado. Inclusive dos switches e nós intermediários.
- Essas informações são mantidas em cache para que todos os quadros do fluxo recebam o mesmo tratamento.

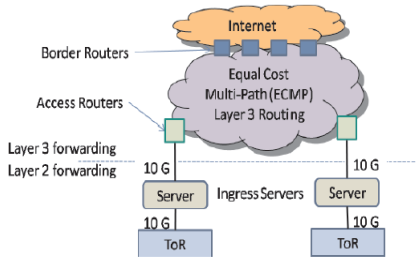
Monsoon


Monsoon Agent
User
Kernel
TCP
Resolve remote IP
IP
ARP
MAC Resolution Cache
Encapsulator
NIC

Monsoon

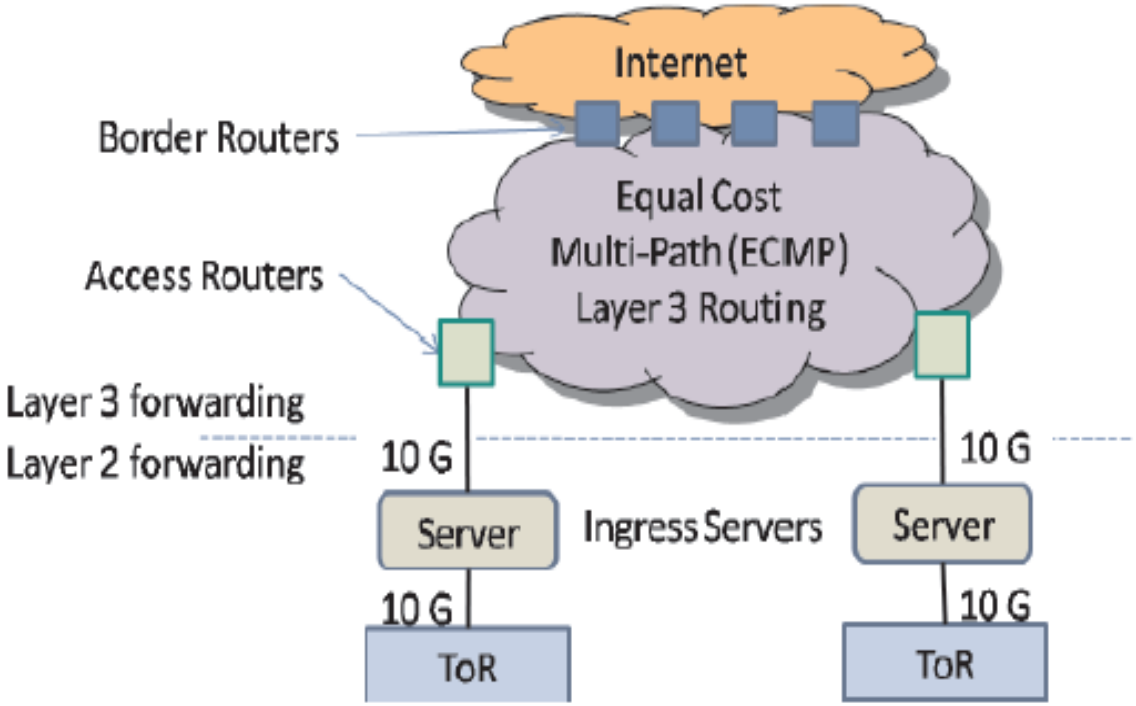

Ethernet Headers
Dest: N Src: S
Dest: TD Src: S
Dest: D Src: S
IP Packet...
Intermediate Node (N)
Dest: TD Src: S
Dest: D Src: S
IP Packet...
2
3
ToR (TS)
ToR (TD)
Dest: N Src: S
Dest: TD Src: S
Dest: D Src: S
IP Packet...
1
4
Dest: D
IP Packet...
Source (S)
Dest (D)

Monsoon

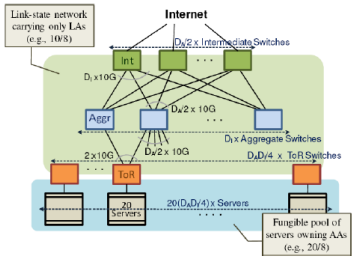


- Conectividade externa:
  - Todo o tráfego entra e sai do data center através dos roteadores de borda que, por sua vez, estão conectados a um conjunto de roteadores de acesso utilizando uma rede de camada 3.
  - No caso dos pacotes vindos da Internet, o servidor de ingresso utiliza o serviço de diretório para resolver o endereço IP e encaminha o tráfego para dentro da rede do data center como qualquer outro servidor.
  - Para os pacotes na direção da Internet, o serviço de diretório mapeia o endereço MAC do gateway default para o endereço MAC dos servidores de ingresso, possibilitando assim a saída dos pacotes para a Internet.

Monsoon

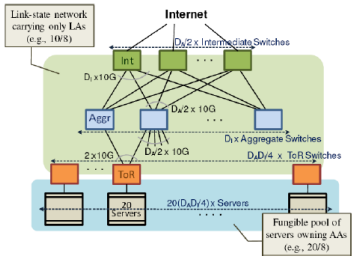

Internet
Border Routers
Equal Cost
Multi-Path (ECMP)
Layer 3 Routing
Access Routers
Layer 3 forwarding
Layer 2 forwarding
10 G
10 G
Server
Ingress Servers
Server
10 G
10 G
ToR
ToR

Novas Topologias



VL2 – Virtual Layer 2

VL2 – Virtual Layer 2

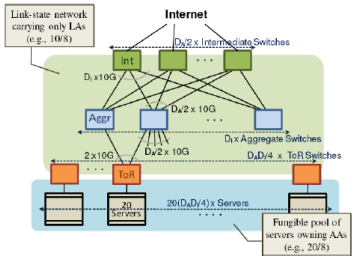


- Pertence ao mesmo grupo de pesquisa da Microsoft que criou o Monsoon. É considerado, portanto, uma evolução do mesmo.

- Criação de uma camada 2 virtual que provê aos serviços do data center a ilusão de que todos os servidores estão conectados através de um único switch de camada 2.

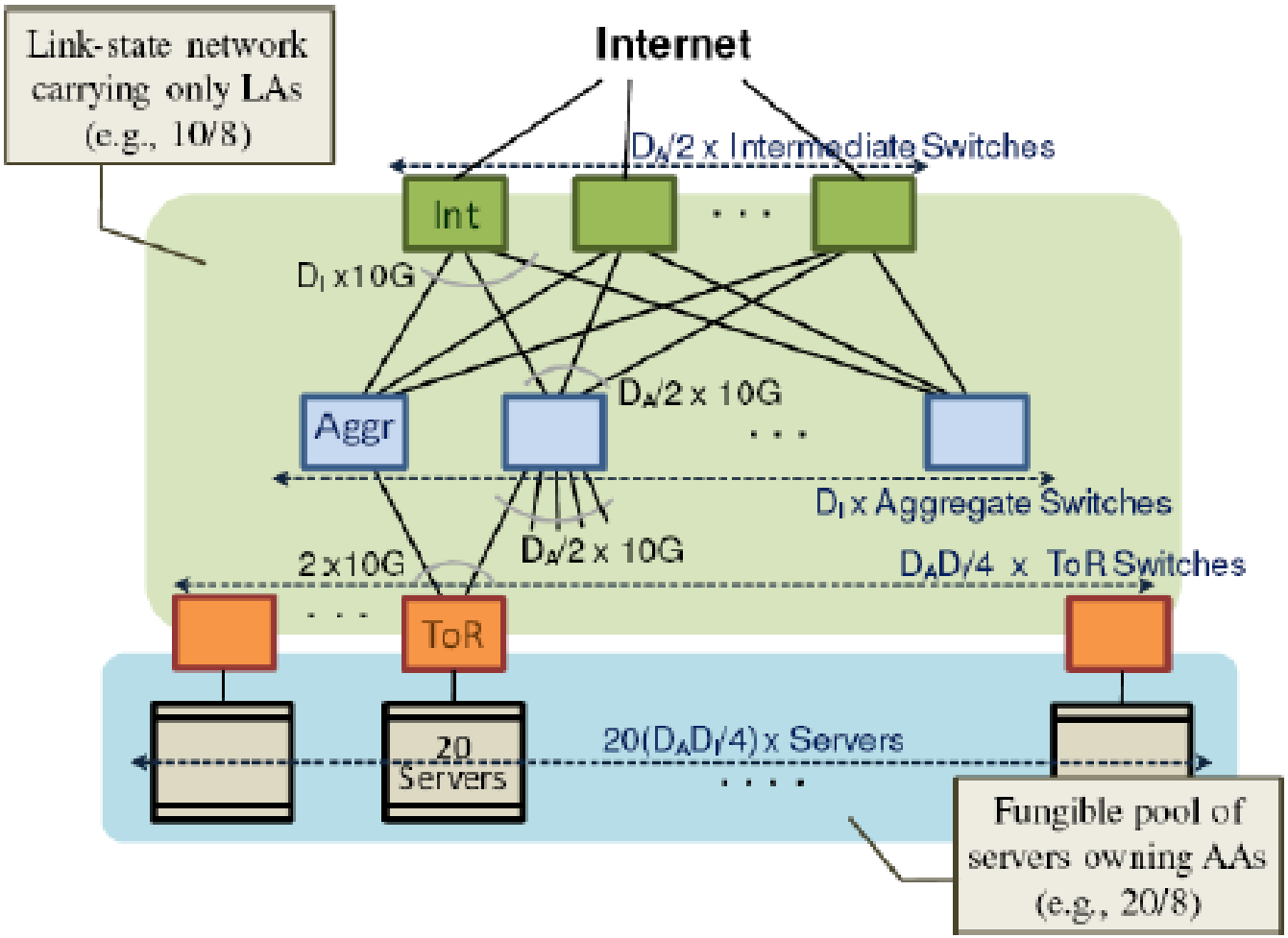
- Propõe uma reorganização nos papéis desempenhados tanto pelos servidores quanto pela rede, através da introdução de uma camada de software na pilha de protocolos.

VL2 – Virtual Layer 2

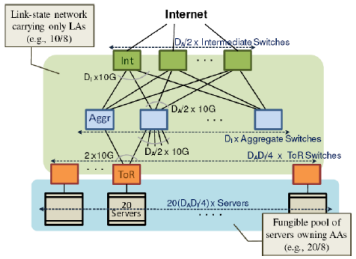


- Para criar essa ilusão é preciso honrar três objetivos:
  - Comunicação entre servidores só pode ser limitada pela taxa de transmissão das interfaces de rede.
  - O tráfego gerado por um serviço deve ser isolado de forma que não afete outros serviços.
  - O data center deve ser capaz de alocar qualquer servidor para atender a qualquer serviço (agilidade).

VL2 – Virtual Layer 2

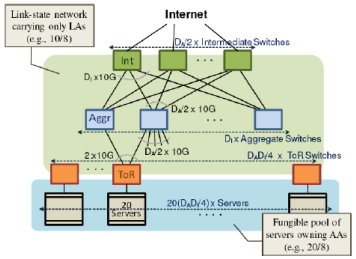


VL2 – Virtual Layer 2



- Elevada conectividade entre os switches de agregação e intermediários. Garantia de lenta degradação do serviço em caso de falhas nos switches intermediários.
- Os switches ToR são conectados a dois switches de agregação.
- A rede é constituída por duas classes de endereços:
  - Endereços com significado topológico (LAs – Locator Addresses)
  - Endereços planos de aplicação (AAs – Application Addresses)

VL2 – Virtual Layer 2



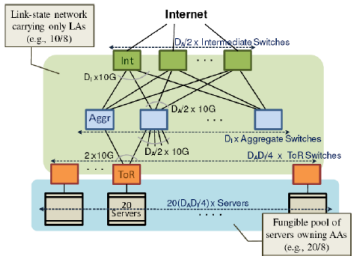
- Os endereços LAs são atribuídos para todos os switches e suas interfaces.

- Todos o switches executam um protocolo de roteamento baseado no estado do enlace para disseminar estes LAs, oferecendo uma visão global da topologia e possibilitando o uso de caminhos mais curtos no encaminhamento de pacotes.

- As aplicações utilizam AAs que permanecem inalterados, independentemente da maneira como os servidores migram no interior do data center.

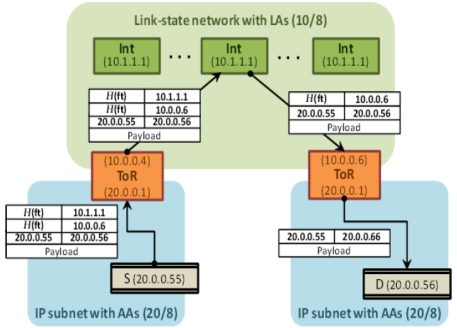
- Para todo AA é associado o LA atribuído ao switch ToR no qual o servidor está conectado. Este mapeamento é mantido por um serviço de diretório do VL2.

VL2 – Virtual Layer 2



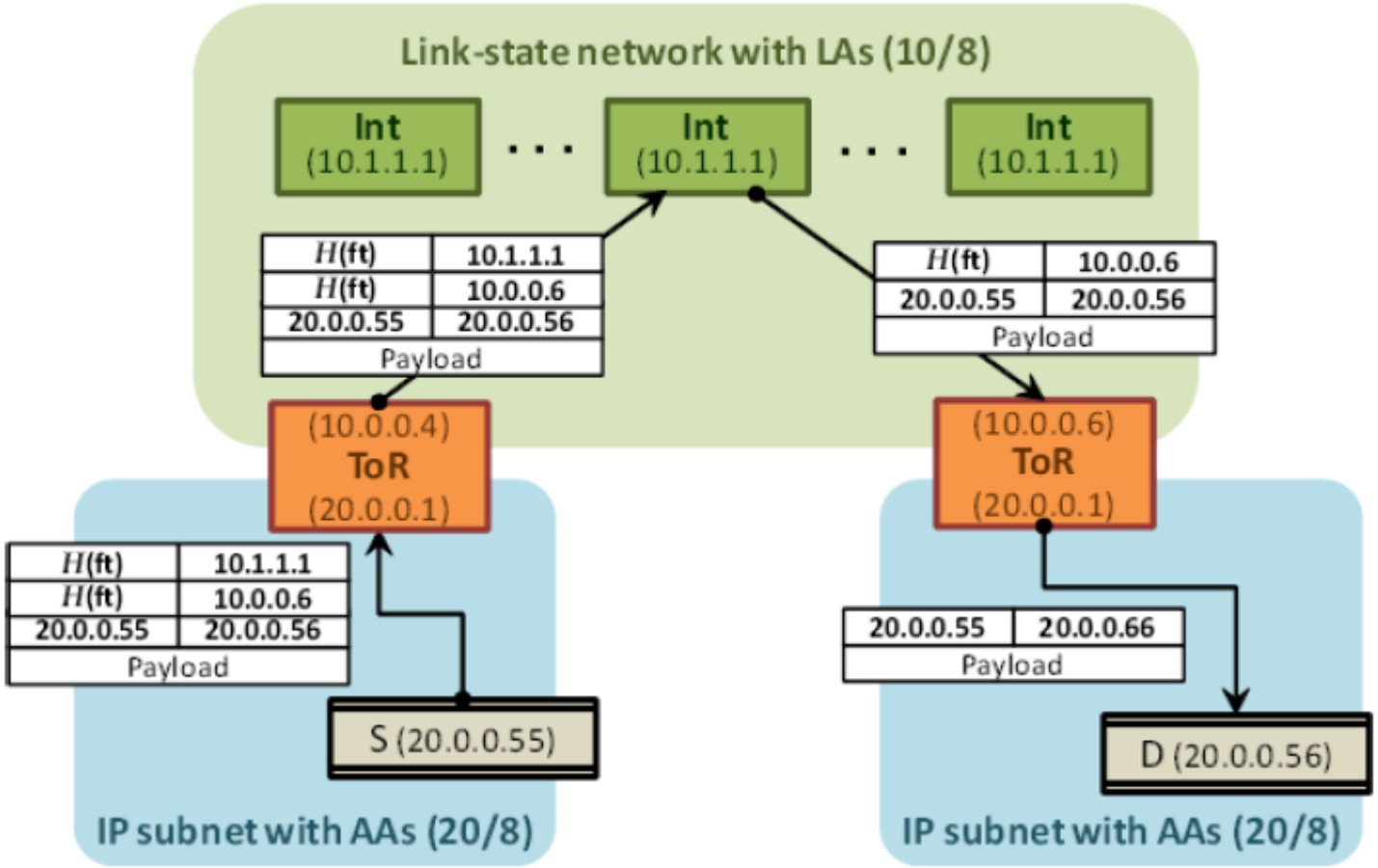
- A malha de camada 3 formada pelo VL2 cria a ilusão de um único domínio de camada 2 para os servidores no interior do data center, uma vez que os servidores imaginam pertencer a uma mesma VLAN.

VL2 – Virtual Layer 2

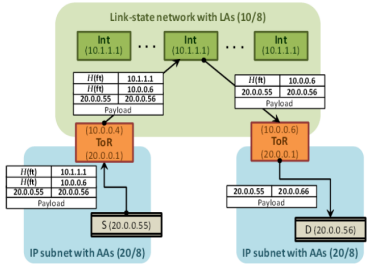


- Endereçamento e roteamento:
  - Para rotear tráfego entre servidores identificados por endereços AA em uma rede que possui rotas formadas a partir de endereços LA, um agente VL2, executando em cada um dos servidores, intercepta os pacotes originados e os encapsula em pacotes endereçados ao LA do switch ToR associado ao servidor de destino.
  - O sucesso do VL2 está associado ao fato dos servidores acreditarem compartilhar uma única sub-rede IP, devido ao encapsulamento efetuado.

VL2 – Virtual Layer 2



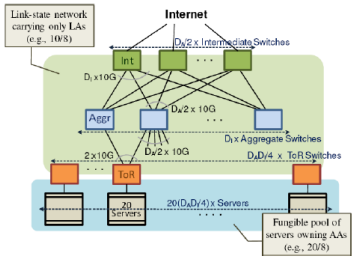
VL2 – Virtual Layer 2



- Para controlar o broadcast gerado pelo protocolo ARP, na primeira vez em que um servidor envia pacotes para um determinado endereço AA, a pilha de rede original do servidor gera uma requisição ARP e a envia para a rede. Nesse instante, o agente VL2 presente no servidor intercepta a requisição e a converte em uma pergunta unicast para o serviço de diretório do VL2.

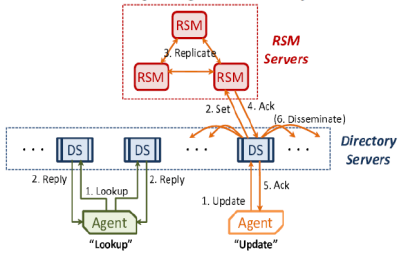
- O serviço de diretório responde, por sua vez, com o endereço LA do ToR de destino e o agente VL2 armazena este mapeamento.

VL2 – Virtual Layer 2



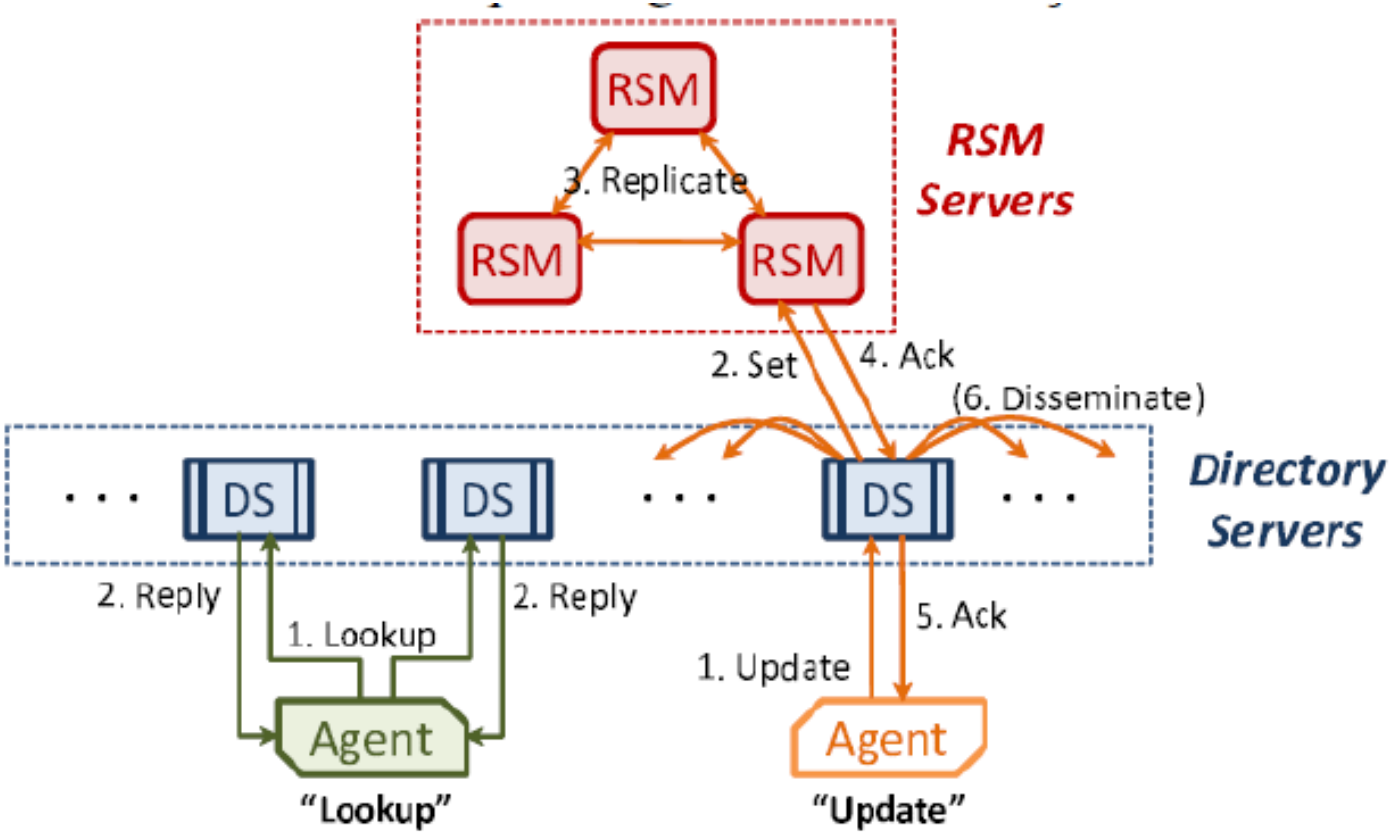
- Espalhamento de tráfego por múltiplos caminhos:
  - O VL2 combina o VLB e o ECMP como forma de evitar áreas de concentração de tráfego. O VLB é utilizado para distribuir o tráfego entre um conjunto de switches intermediários e o ECMP é utilizado para distribuir o tráfego entre caminhos de custo igual.
  - Eventuais falhas no switches intermediários poderiam levar a um cenário no qual um elevado número de agentes VL2 teriam de ser atualizados para obter o novo estado da rede.
  - O VL2 contorna esta situação atribuindo o mesmo endereço LA para todos os switches intermediários (10.1.1.1). Esse switches estão a exatamente três saltos de distância dos servidores de origem criando um cenário adequado para a utilização do ECMP.

VL2 – Virtual Layer 2

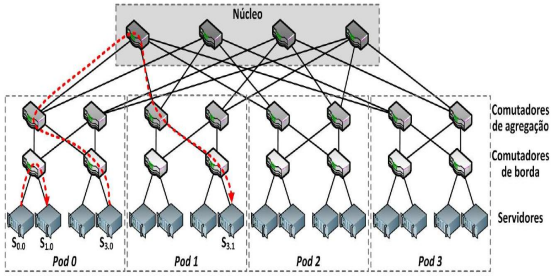


- O serviço de diretório do VL2 provê três serviços principais:
  - Consultas;
  - Atualizações de mapeamento entre AAs e LAs;
  - Um mecanismo de atualização de cache reativo para atualizações sensíveis a atrasos.

VL2 – Virtual Layer 2

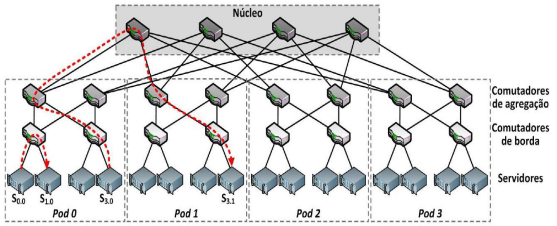


Novas Topologias



Portland

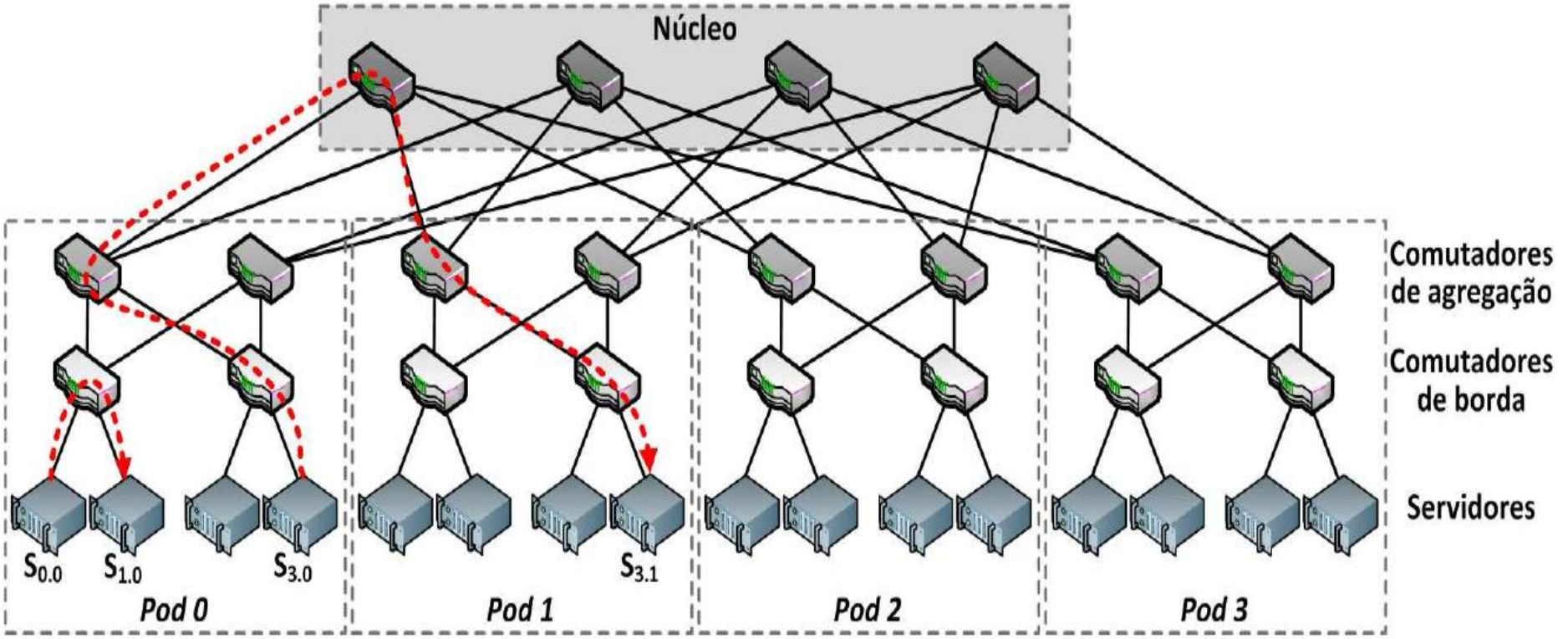
Portland



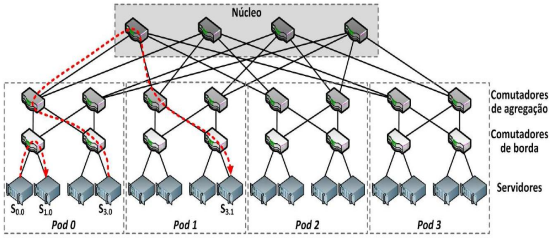
- Conjunto de protocolos compatíveis com Ethernet para efetuar o roteamento, encaminhamento e resolução de endereços.

- Desenvolvido considerando-se a estrutura organizacional comumente encontrada em data centers (Fat-Tree).

Portland – Fat-Tree



Portland - Fat-Tree

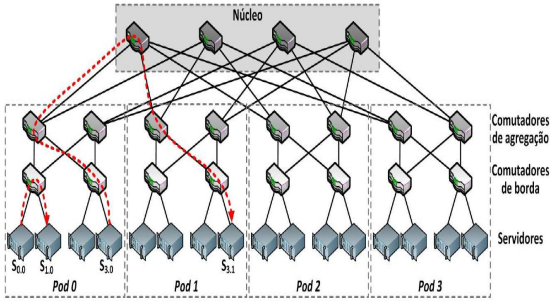


- A árvore possui dois tipos de conjuntos, o núcleo e os pods. O núcleo é formado por comutadores que possuem cada uma de suas portas conectada a um pod diferente. O pod é formado por comutadores de agregação e de borda, além dos servidores em si.

- Todos os comutadores da rede são idênticos e possuem k portas. Assim a rede possui k pods sendo que cada pod possui k/2 comutadores de agregação e outros k/2 de borda. Cada comutador de borda está individualmente ligado a k/2 servidores diferentes.

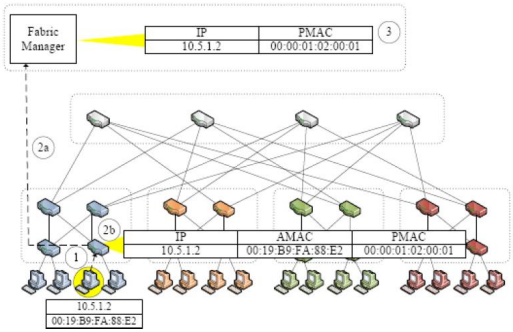
- Portanto, a topologia Fat-Tree possui capacidade para k/2 * k/2 * k servidores (na figura anterior k = 4) e k/2 * k/2 * 5 switches.

Portland



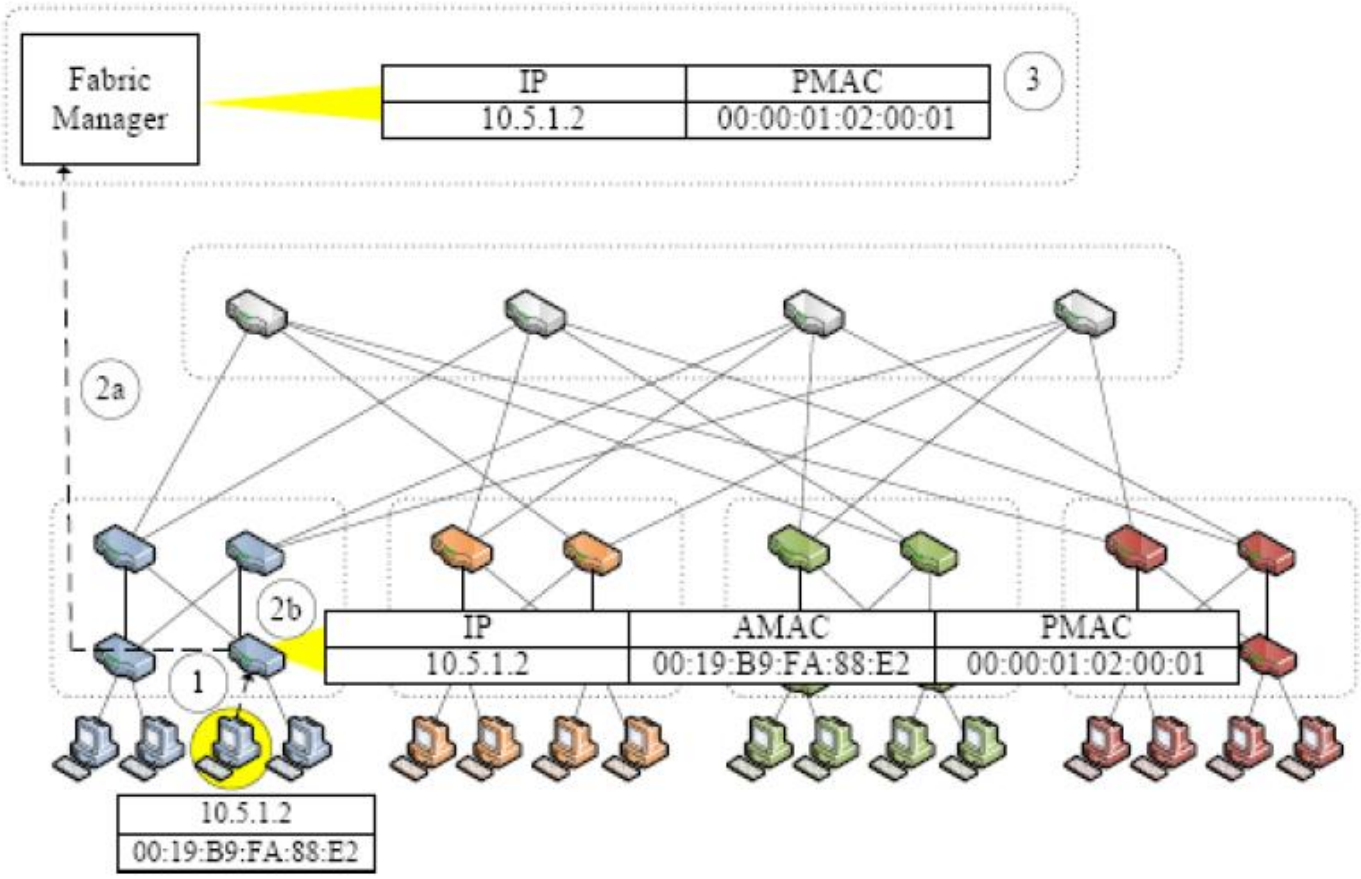
- O Portland propõe:
  - Utilização de um protocolo para possibilitar que os switches descubram sua posição topológica na rede.
  - A atribuição de Pseudo endereços MAC (PMAC) para todos os nós finais, de forma a codificar suas posições na topologia.
  - A existência de um serviço centralizado de gerenciamento da infraestrutura de rede.
  - A implantação de um serviço de proxy para contornar broadcasts inerentes ao ARP.

Portland

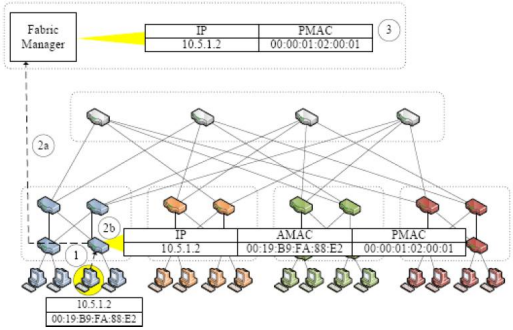


- Fabric Manager:
  - O Portland utiliza um gerenciador de infraestrutura de rede centralizado para manter estados relacionados à configuração da rede, tal como sua topologia.
  - O Fabric Manager é um processo executado em uma máquina dedicada responsável pelo auxílio às requisições do ARP, tolerância a falhas e operações de multicast.

Portland

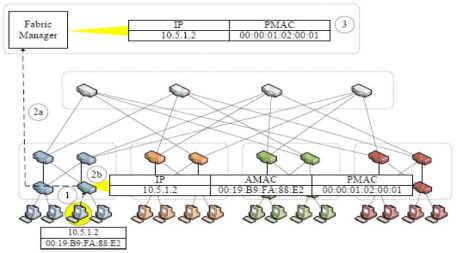

Fabric Manager
IP
PMAC
10.5.1.2
00:00:01:02:00:01
3
2a
2b
1
IP
AMAC
PMAC
10.5.1.2
00:19:B9:FA:88:E2
00:00:01:02:00:01
10.5.1.2
00:19:B9:FA:88:E2

Portland



- Endereços PMAC (Pseudo MAC):
  - O Portland atribui um único PMAC para cada nó final, representando a localização de cada nó na topologia.
  - Cada nó, portanto possui um PMAC e um MAC real (AMAC).
  - As requisições de ARP feitas pelos nós finais são respondidas com o PMAC do nó de destino. Sendo assim, todo o processo de encaminhamento de pacotes ocorre através da utilização dos PMAC.
  - Os switches de egresso são responsáveis pelo mapeamento PMAC para AMAC e reescrita dos pacotes para manter a ilusão de endereços MAC inalterados no nó de destino.

Portland



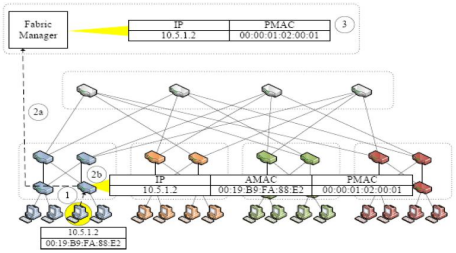
- Os switches de borda aprendem um único número de pod e uma única posição dentro deste pod.

- Para todos os nós diretamente conectados, os switches de borda atribuem um PMAC de 48 bits:

  pod.posição.porta.vmid

  - “pod” possui 16 bits e refere-se ao número do pod onde os nós estão localizados.
  - “posição” possui 8 bits e indica a posição do switch dentro do pod.
  - “porta” possui 8 bits para representar a porta na qual o nó final está ligado ao switch.
  - “vmid” possui 16 bits e é utilizado para multiplexar máquinas virtuais em uma mesma máquina física.

Portland



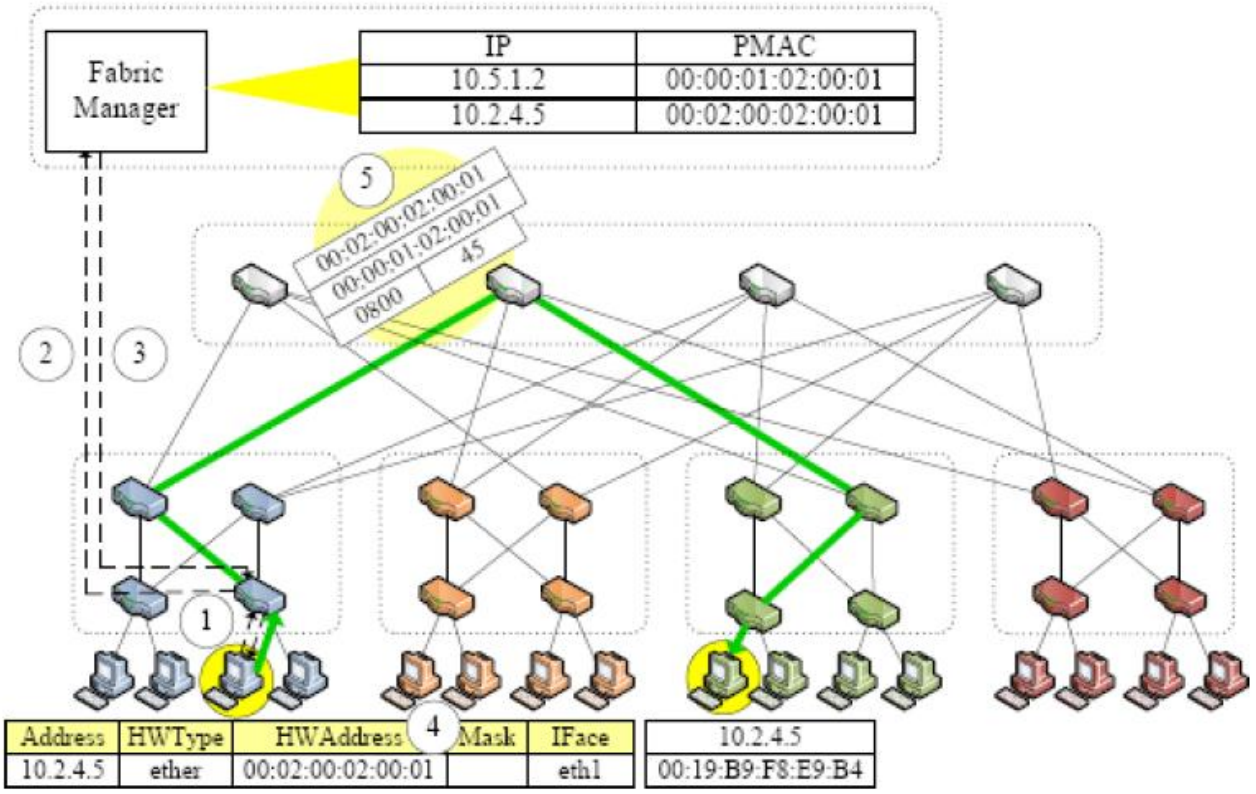
- No instante em que um switch de ingresso observa a existência de um novo endereço MAC, os pacotes com este endereço são desviados para o plano de controle do switch, que cria uma nova entrada na tabela de mapeamento e, na sequência, encaminha este novo mapeamento para o Fabric Manager para futuras resoluções.

- O Portland efetua a separação entre localizador/identificador de forma totalmente transparente aos nós finais e compatível com o hardware dos switches comoditizados disponíveis no mercado.

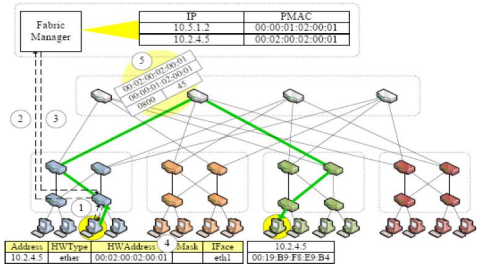
- Outra característica importante do Portland refere-se a não utilização de técnicas de tunelamento para encaminhar os pacotes, sendo apenas necessário a reescrita de endereços PMAC/AMAC nas bordas do data center.

Portland

- Mecanismo de proxy para requisições ARP:
  - O Portland utiliza o Fabric Manager para contornar o overhead de sinalização causado pelo ARP.

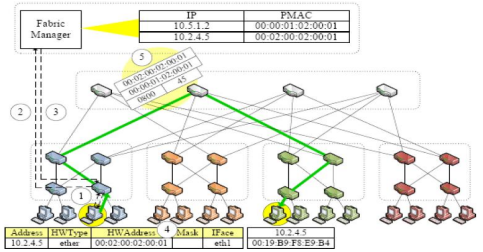


Portland



- Passo 1: o switch de egresso detecta a chegada de uma mensagem ARP requisitando um mapeamento IP para MAC.
- Passo 2: intercepta está mensagem e a encaminha para o Fabric Manager.
- Passo 3: o Fabric manager consulta sua tabela de PMACs em busca do mapeamento e retorna o PMAC para o switch requisitante.
- Passo 4: O switch de borda cria uma mensagem de resposta do ARP e a retorna para o nó que originou a requisição.

Portland

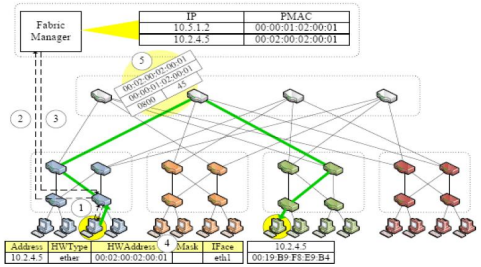


- Para prover suporte à migração de máquinas virtuais, assim que a migração é completada, a máquina virtual envia um ARP gratuito contendo seu novo mapeamento entre endereços IP e MAC. Este ARP é encaminhado até o Fabric Manager.

- Os nós que estavam se comunicando com esta máquina antes da migração manterão o mapeamento antigo e terão de esperar até que o mapeamento expire para prosseguir com a comunicação.

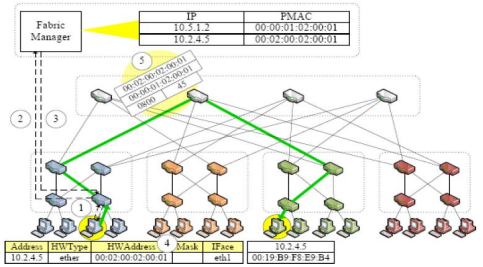
- O Fabric Manager pode encaminhar uma mensagem de invalidação de mapeamento ao switch no qual a máquina virtual estava associada. Desta maneira o switch seria capaz de replicar o ARP gratuito aos nós que continuam a originar pacotes na direção da máquina virtual que migrou.

Portland



- Protocolo de descoberta de posição topológica:
  - Os switches utilizam informações relativas às suas posições na topologia global do data center para efetuar encaminhamento e roteamento mais eficientes através da comunicação em pares, ou seja encaminhamento considerando apenas os vizinhos diretamente associados.
  - O Portland propõe a utilização de um protocolo para descobrir a localização topológica de forma automática.
  - Neste protocolo chamado LDP (Location Discovery Protocol), os switches enviam periodicamente LDMs (Location Discovery Messages) em todas as suas portas para:
    - Definir suas posições
    - Monitorar o estado de suas conexões físicas.

Portland

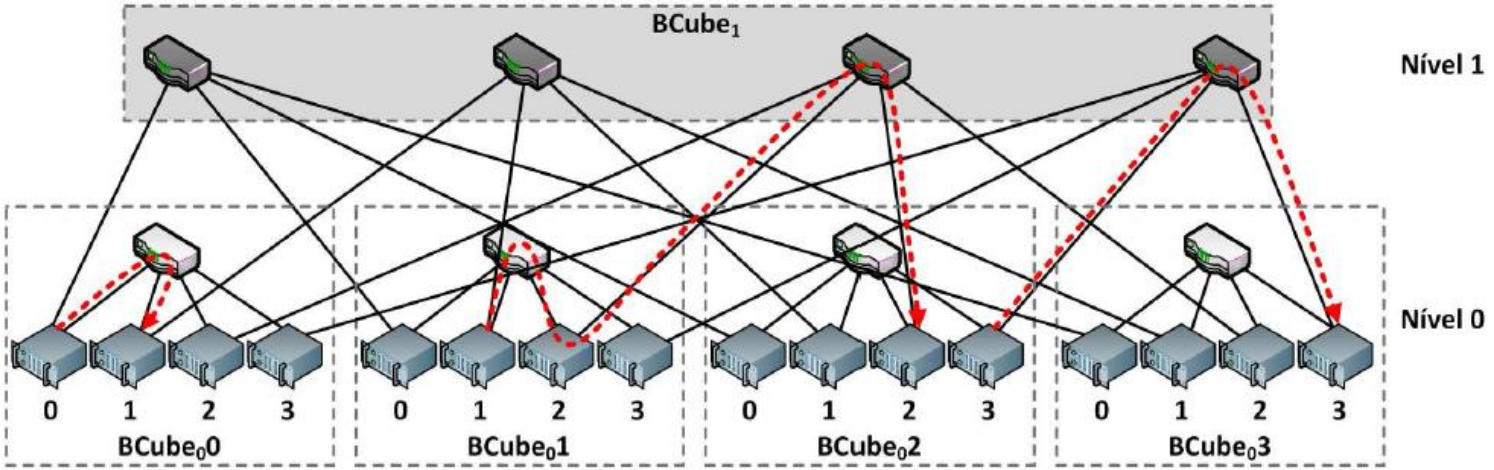


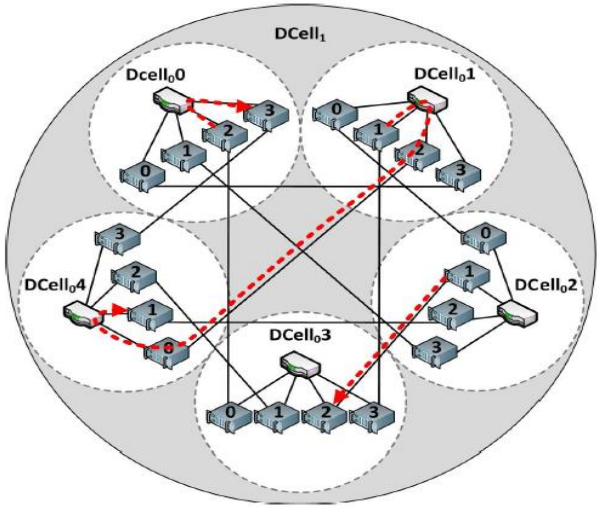
- Em resumo o LDP consegue definir quais são os switches de borda, uma vez que eles recebem LDMs em apenas uma fração de suas portas; as conectadas aos switches de agregação, pois os nós finais não geram LDMs.

- A partir do momento que os switches de borda descobrem sua localização (nível) na topologia, as LDMs disseminadas subsequentemente passam a conter informações referente ao seu nível. Desta forma, o restante dos switches são capazes de aferir suas respectivas posições:
  - switches de agregação recebem LDMs em todas as portas, algumas sem informação de nível (switches de núcleo).
  - Switches de núcleo recebem todos os LDMs contendo informação de níveis.
  - Definido o nível de todos os switches, o Fabric Manager é utilizado para atribuir o mesmo número de pod para os switches de borda pertencentes ao mesmo grupo.

Outras Topologias

Bcube e DCell





	Monsoon	VL2	Portland	BCube e MD-Cube
Realocação dinâmica de servidores (agilidade)	sim	sim	sim	sim
Transmissão à taxa máxima das interfaces/ *oversubscription*	sim (1Gbps) / 1:1	sim (1Gbps) / 1:1	sim (1Gbps) / 1:1	sim (1Gbps) / 1:1
Topologia	*fat tree*	*fat tree*	*fat tree*	hipercubo
Mecanismo de Roteamento/ Encaminhamento	tunelamento MAC-in-MAC	tunelamento IP-in-IP	baseado na posição hierárquica dos nós (PMAC)	rota na origem gerada por mensagens de *probing*
Balanceamento de Carga	VLB + ECMP	VLB + ECMP	Não especificado	trocas periódicas de *probing* modificam a rota
Modificação nos Nós Finais	sim	sim	não	sim
Modificação nos *Switches*	sim	não	sim	não
Serviço de diretório	sim	sim	sim	não

Considerações Finais

- O modelo convencional IP/Ethernet não atende os requisitos de custo, escala e controle dos provedores de computação em nuvem.

- Por este motivo novos projetos e propostas têm emergido para atender os objetivos específicos dos cloud data centers, que são criticamente diferentes dos data centers tradicionais e das redes locais e metropolitanas dos provedores de serviços.

- As novas propostas de arquiteturas de rede prometem atingir uma redução dos custos operacionais e de capital, uma maior confiabilidade, um modelo de escala sob demanda sustentável e uma maior capacidade de inovação.

Referências

1 - Verdi, F. L.; Rothenberg, C. E.; Pasquini, R. e Magalhães, M. F.. Capitulo 3 - “Novas arquiteturas de Data Center para Cloud Computing”. In: Livro Texto dos Minicursos, XXVIII Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos. Gramado: Sociedade Brasileira de Computação (SBC), 2010.

2 - [Greenberg 2009] Greenberg, A. (2009). Networking The Cloud. ICDCS 2009 keynote. Disponível online em http://www.cse.ohio-state.edu/icdcs2009/Keynote_files/greenbergkeynote.pdf.

3 - [Greenberg et al. 2009a] Greenberg, A., Hamilton, J., Maltz, D. A., and Patel, P. (2009a). The Cost of a Cloud: Research Problems in Data Center Networks. SIGCOMM Comput. Commun. Rev., 39(1):68–73.

4 - [Greenberg et al. 2009b] Greenberg, A., Hamilton, J. R., Jain, N., Kandula, S., Kim, C., Lahiri, P., Maltz, D. A., Patel, P., and Sengupta, S. (2009b). VL2: A Scalable and Flexible Data Center Network. In Proceedings of the ACM SIGCOMM 2009 Conference on Data Communication, Barcelona, Spain.

5 - [Greenberg et al. 2008] Greenberg, A., Lahiri, P., Maltz, D. A., Patel, P., and Sengupta, S. (2008). Towards a Next Generation Data Center Architecture: Scalability and Commoditization. In Proceedings of the ACM Workshop on Programmable Routers For Extensible Services of Tomorrow, Seattle, WA, USA.

Referências

6 - [Mysore et al. 2009] Mysore, R. N., Pamboris, A., Farrington, N., Huang, N., Miri, P., Radhakrishnan, S., Subramanya, V., and Vahdat, A. (2009). PortLand: A Scalable Fault-Tolerant Layer 2 Data Center Network Fabric. In Proceedings of the ACM SIGCOMM 2009 Conference on Data Communication, Barcelona, Spain.

7 - [Al-Fares et al. 2008] Al-Fares, M., Loukissas, A., and Vahdat, A. (2008). A Scalable Commodity Data Center Network Architecture. SIGCOMM Comput. Commun. Rev., 38(4):63–74.

8 - Couto, R. S.; Campista, M. E. M. e Costa, L. H. M. K. – “Uma Avaliacão da Robustez Intra Data Centers Baseada na Topologia da Rede”

Questão

- O ARP (Address Resolution Protocol) é um protocolo usado para encontrar um endereço de camada 2 (MAC) a partir de um endereço de camada 3 (IP). O emissor difunde em broadcast um pacote ARP contendo o endereço IP de outro host e espera uma resposta com o endereço MAC respectivo. Ao conectar milhares de servidores usando uma única rede de camada 2 devemos evitar o uso de sinalizações em broadcast. Escolha uma das topologias abaixo e explique como a mesma foi projetada para evitar o broadcast do protocolo ARP:
  - Monsoon (Ref 5 – seção 3.2)
  - VL2 (Ref 4 – seção 4.2)
  - Portland (Ref 6 – seção 3.3)