中国科学技术大学
University of Science and Technology of China

VL2: a scalable and flexible data center network

SIGCOMM 2009

授课教师：赵功名

中国科大计算机学院

2025年秋·高级计算机网络

Outline

I. Introduction

II. Background

III. Measurements & Implications

IV. Virtual Layer Two Networking

V. Evaluation

VI. Review

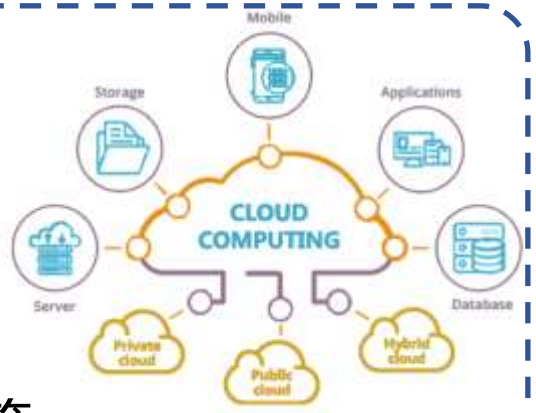
Outline

I. Introduction
   1. Cloud services require AGILITY
   2. Current DCNs lack AGILITY
   3. VL2 goals & contributions
II. Background
III. Measurements & Implications
IV. Virtual Layer Two Networking
V. Evaluation
VI. Review

Introduction：Cloud services require AGILITY

云服务

- 搜索
- 邮箱
- Mapreduce计算
- …



需要

大规模数据中心

- 规模：几十~上百万台服务器
- 业务：多服务混合部署

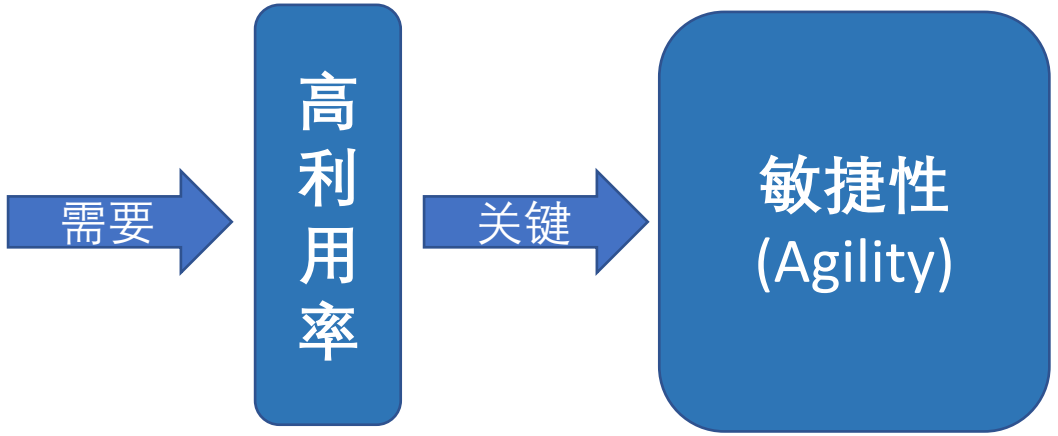
为云服务建立大规模数据中心的动因

- **经济**考量：批量采购/部署，摊薄成本 → **规模经济**
- **技术**考量：按负载动态在不同服务间调度服务器 → **资源弹性**

Introduction：Cloud services require AGILITY

尽管如此，数据中心的成本依然高昂...

- 十万台服务器的数据中心每月开销高达1200万美元
- 成本大部分来自服务器本身



敏捷性(Agility)：为任意服务分配任意服务器 → **风险管理 成本控制**

缺乏敏捷性	具有敏捷性
➢ 每个服务得提前囤足服务器防突发 ➢ 需求难预测 → 服务器过多→ 利用率低 ➢ 服务器过少→ 无法应对流量高峰 ➢ 结果：整体利用率低，成本高	➢ 所有服务共享服务器池 ➢ 为突发需求的服务弹性分配服务器，分钟级弹性伸缩 ➢ 不用各自分配过量资源，利用率提升，硬件预算降低

Outline

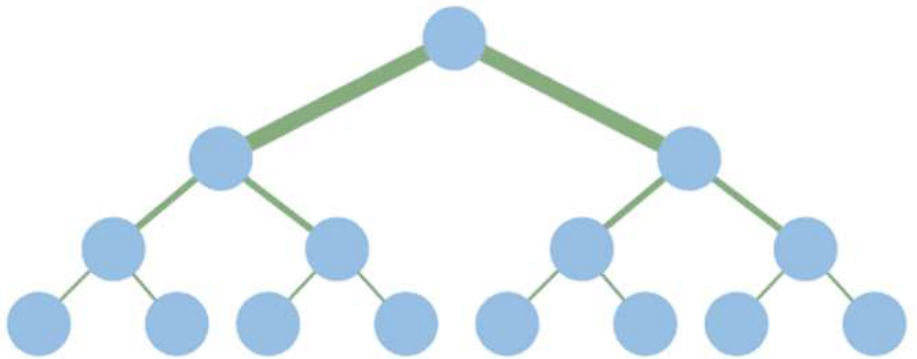
I. Introduction
   1. Cloud services require AGILITY
   2. Current DCNs lack AGILITY
   3. VL2 goals & contributions
II. Background
III. Measurements & Implications
IV. Virtual Layer Two Networking
V. Evaluation
VI. Review

Introduction：Current DCNs lack AGILITY

当前数据中心网络缺乏敏捷性

- **服务器缺乏足够互连带宽**
- 服务缺乏性能隔离
- 虚拟机缺乏灵活迁移机制

- 传统数据中心网络依赖昂贵硬件搭建树形网络拓扑
- 越向上链路越贵，所以采用over subscription
  - 一般在1:5以上
  - 最高层可达1:240
- 后果：服务器池被分隔，空闲服务器存在，但因为跨段带宽不足，没法把热点任务迁过去。



Introduction：Current DCNs lack AGILITY

当前数据中心网络缺乏敏捷性

- 服务器缺乏足够互连带宽
- **服务缺乏性能隔离**
- 虚拟机缺乏灵活迁移机制

- 数据中心虽然部署了许多服务，但通常没有机制来保证一个服务的突发流量不会影响到其他服务
- 当一个服务经历突发流量时，与之位于同一子树上的其他服务也会受到附带的影响

Introduction：Current DCNs lack AGILITY

当前数据中心网络缺乏敏捷性

- 服务器缺乏足够互连带宽
- 服务缺乏性能隔离
- **虚拟机缺乏灵活迁移机制**

传统数据中心的路由设计：

- 服务器的IP地址和拓扑相关
- 同一服务的的服务器被划分至同一VLAN

带来的问题：

- 虚拟机无法保持原有IP地址热迁移
- 配置繁琐、负担巨大，往往需要人工介入，进一步限制了部署速度

Outline

Introduction：VL2 Goals & Contributions

目标

为了解决当下存在的问题，一种理想的网络模型：每一个服务都认为分配给它的所有服务器通过一个独立、无干扰的以太交换机（即虚拟二层 Virtual Layer 2）相连，并且服务器的数量可以从1扩展到100,000。为了实现这样的愿景，需要构建满足以下三个特征的网络：

统一高带宽	性能隔离	二层语义
➢ 服务器到服务器的最大传输速率应该只受发送端和接收端服务器的网络接口速率限制 ➢ 为一个服务分配服务器应该和网络拓扑无关	一个服务的流量不应该受到其他服务的流量影响，达到每个服务都是通过一个独立的物理交换机相连的效果	➢ 任意 IP 可绑任意端口 ➢ VM 热迁不改地址 ➢ ARP/DHCP/广播 legacy 应用零修改

Introduction：VL2 Goals & Contributions

VL2要点

- **可部署**：不改变交换机和服务器的硬件，不修改遗留应用代码
- **VLB的应用**：使用Valiant Load Balancing应对数据中心流量的多变。VLB不需要中心化的协调或流量工程，服务器自动把每条流分摊到随机路径来达到负载均衡和减轻拥塞效果
- **IP地址分离**：为了灵活性，将拓扑意义从应用的IP地址中分离，采用目录系统建立应用IP和物理IP的映射

Introduction：VL2 Goals & Contributions

本文贡献

- 对生产环境的数据中心进行流量模型研究，发现流量具有巨大的波动性
- 在80台服务器的集群中设计、实现和部署了VL2的所有组件，验证VL2满足了前面提出的所有目标，同时网络的速度优良
- 第一次把Valiant Load Balancing(VLB)算法用于数据中心交换机之间，并验证按流粒度对流量进行划分能达到几乎相同的划分效果
- 通过和同等能力的传统网络比较价格，证明了VL2所做trade-off的合理性

Outline

I. Introduction

II. Background

1. Why conventional DCN arch falls short

2. Valiant Load Balancing

III. Measurements & Implications

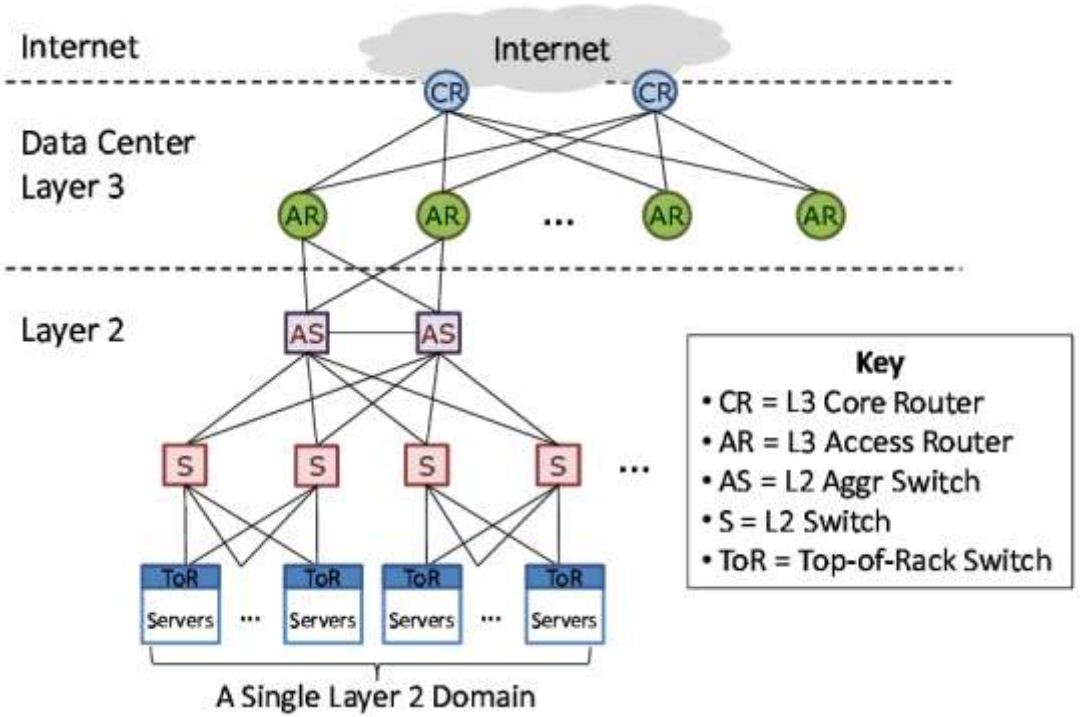
IV. Virtual Layer Two Networking

V. Evaluation

VI. Review

Background：Why conventional DCN arch falls short

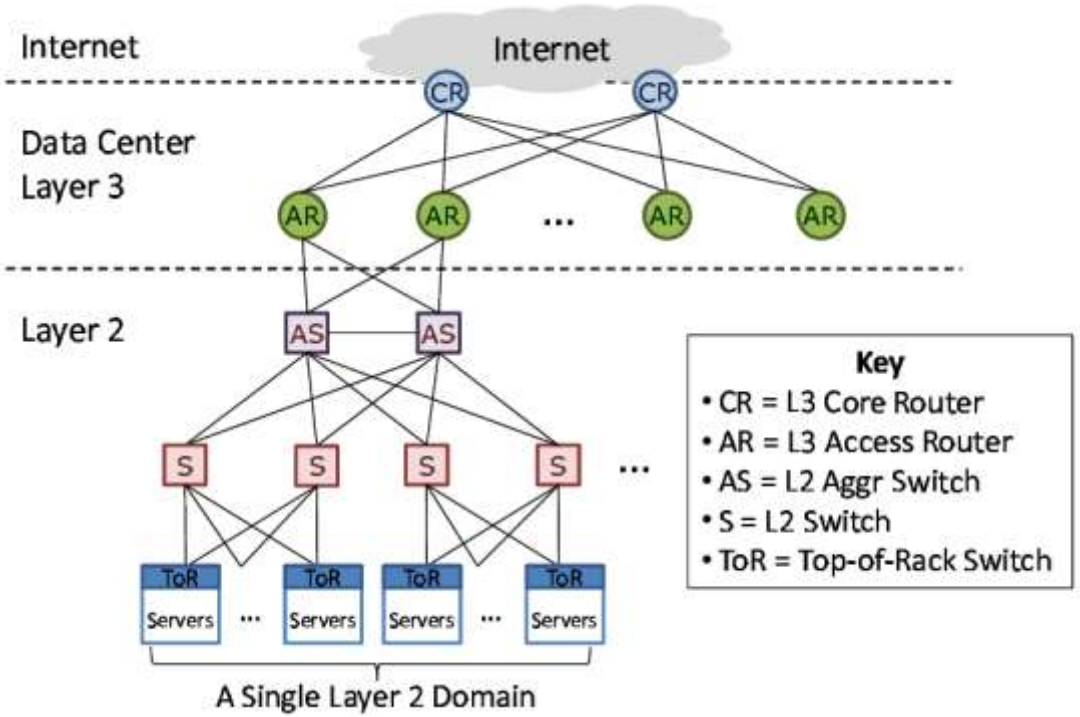
传统数据中心网络架构



- 从ToR交换机到核心路由器形成**层级结构**
  - 每个机柜通常有20到40台服务器，每台服务器连接到ToR交换机
  - 每个ToR交换机连接到两个汇聚交换机AS来达到冗余
  - 汇聚交换机AS继续连接接入路由器AR达到进一步的汇聚
  - 在层级结构的顶端，核心路由器完成接入路由器间的数据交换，同时管理进出数据中心的流量
- 每一对接入路由器下的所有交换机形成一个L2域。
- 为了限制开销(例如包泛洪、ARP广播)以及隔离不同服务，服务器划分为VLAN。

Background：Why conventional DCN arch falls short

传统数据中心网络架构的不足

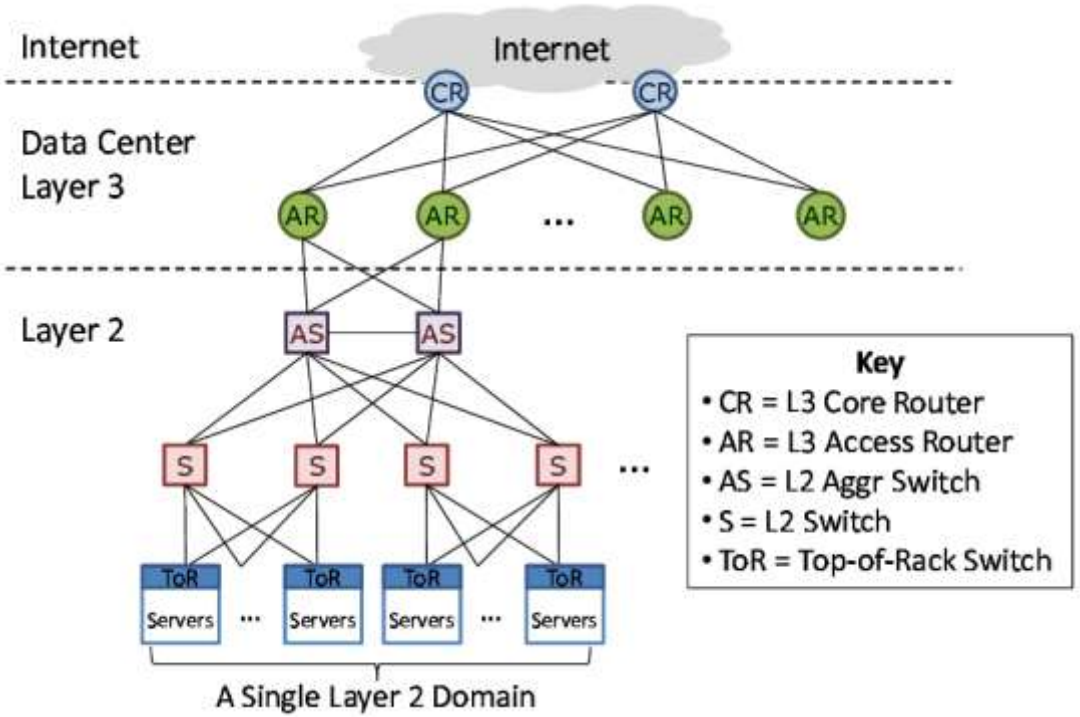


有限的服务器-服务器容量

- 层级结构向上时，面临严重的技术和经济障碍，所以over-subscription的比率急速上升。
  - 机柜内1：1
  - ToR交换机向上1:5到1:20
  - 最上层的达到1:240
- 严重的over-subscription导致服务器池被分隔，阻碍了空闲的服务器分配给超载的服务

Background：Why conventional DCN arch falls short

传统数据中心网络架构的不足

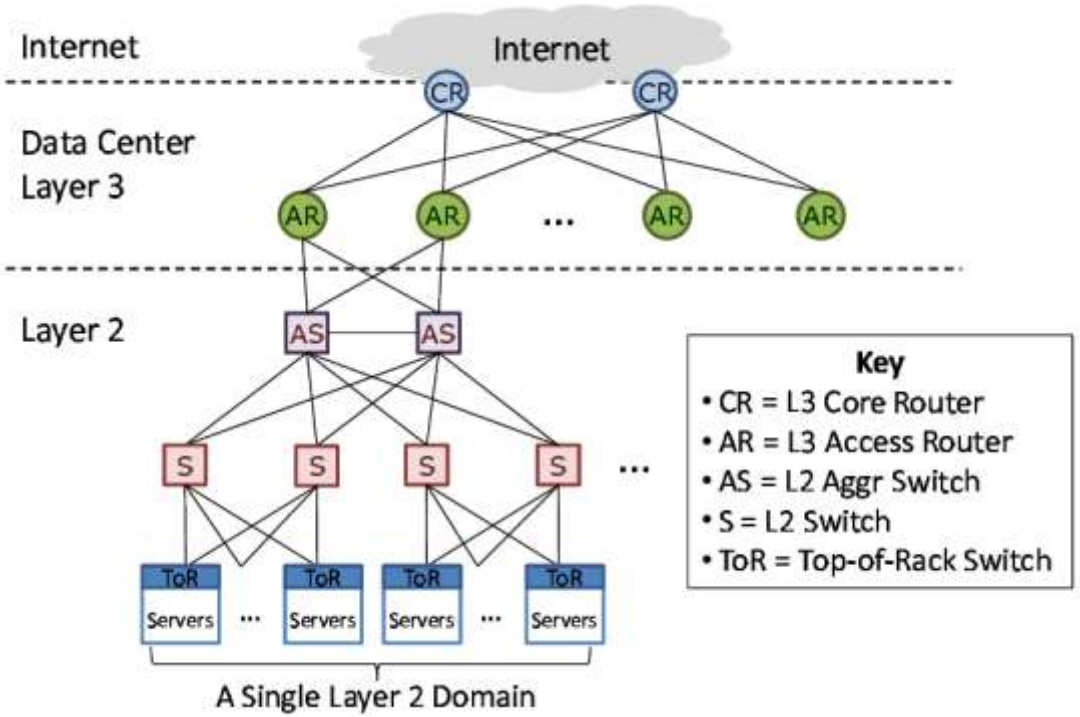


资源碎片化

- 服务器在层级结构中的距离影响着通信的开销和性能 -> 传统设计鼓励将距离近的服务器归为一簇
- 把服务扩展出layer-2域需要经常配置IP地址和VLAN，因为服务器的IP地址由其上的接入交换机(AR)决定。如今的方案用资源换时间，即为每个服务分配冗余的资源，使其在遇到流量高峰时能能扩展到临近的服务器。即便如此，有时仍需要管理员驱逐临近服务器上的服务

Background：Why conventional DCN arch falls short

传统数据中心网络架构的不足



糟糕的可靠性和利用率

- ToR以上部分，冗余度达50%
  - 汇聚交换机和接入路由器成对使用，并且每一个都要能够提供足够带宽，即使一个发生故障，也不影响网络功能
  - 导致链路只能跑到最大容量的50%
- 多路径方案不存在或低效
  - L2：Spanning Tree协议只能利用一条链路
  - L3：ECMP可以利用多路，但传统拓扑一般只有两路

Outline

I. Introduction

II. Background

1. Why conventional DCN arch falls short
2. Valiant Load Balancing

III. Measurements & Implications

IV. Virtual Layer Two Networking

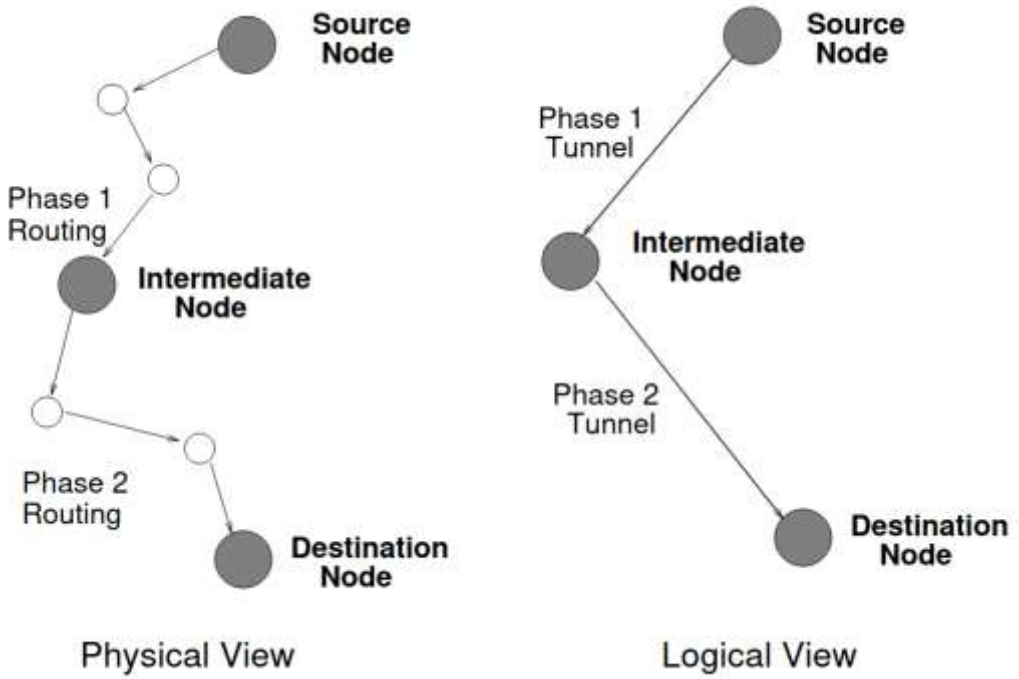
V. Evaluation

VI. Review

Background：Valiant Load Balancing

在一个多路径网络中，直接将源节点的流量发送到目标节点可能会导致链路或交换机上的不均衡。**Valiant Load Balancing**(VLB)通过**两阶段转发**来解决。

- **随机中转**：源节点先将流量随机选择一个“中间节点”（intermediate node），并把数据包转发过去
- **最终转发**：该中间节点再将数据转发到真正的目的节点。



Background：Valiant Load Balancing

VLB的特点：

- **随机化路由**：目标位置无关，均匀分散流量
- **理论保障**：VLB证明了网络在满足
  a) 随机化在小的**数据包**的粒度上进行
  b) 发送到网络中的流量满足**hose model**

这两个条件时，网络可以视为无阻塞模型。

Hose model中，对于每个节点，只规定其最大输入/输出带宽，而不关心具体的流量对端是谁。例如，节点A的出口<x Gbps，入口<y Gbps（例如网卡速度）。这是一种灵活的模型。

VLB**优点**：

- 可以适应多种流量模式，尤其适用于流量模式未知或复杂的网络
- 简单可扩展，不依赖复杂的集中式流量工程

缺点：

- 路径绕行，可能引入额外延迟
- 大流若集中可能导致短期不均衡

Outline

I. Introduction

II. Background

III. Measurements & Implications

IV. Virtual Layer Two Networking

V. Evaluation

VI. Review

Measurements & Implications

为了设计适合云服务的网络机制，作者对用于实际生产的数据中心网络进行观测，并得到了两条关键的信息，以及对于网络设计的指导：

1. 流量模式高度发散（种类多，50个流量模式矩阵只能勉强代表DCN的流量模式），变化迅速且不可预测
2. 层次拓扑不可靠：即使投入大量工作和开销，顶层网络设备的故障仍然造成严重downtime

Outline

I. Introduction

II. Background

III. Measurements & Implications

1. Data-Center Traffic Analysis

2. Flow Distribution Analysis

3. Traffic Matrix Analysis

4. Failure Characteristics

IV. Virtual Layer Two Networking

V. Evaluation

VI. Review

数据中心流量分析

对数据中心NetFlow和SNMP数据的观测展现了4个宏观特征

- 东西向 > 南北向：数据中心服务器间的流量与进出数据中心的流量的比例大致是4:1
- 数据中心计算集中在那些能够快速、低成本地访问内存或磁盘数据的地方
- 数据中心内部服务器间的带宽需求增长速度快于对外部主机的带宽需求增长
- 网络成为计算瓶颈：ToR交换机带宽往往可达80%以上

Outline

I. Introduction

II. Background

III. Measurements & Implications

1. Data-Center Traffic Analysis
2. Flow Distribution Analysis
3. Traffic Matrix Analysis
4. Failure Characteristics

IV. Virtual Layer Two Networking

V. Evaluation

VI. Review

3.2 Flow Distribution Analysis

流分布分析

- 流大小的分布
  - 90 % 流 < 100 KB（老鼠）
  - 90 % 字节落在 1 MB–1 GB（大象）
- 并发流数量
  - 常模 10 条/机，峰值 80 条/机

流大小和并发数量的分布都表明VLB在云服务数据中心中会表现良好：

- 最大流也不过100MB，可以按流粒度随机*
- 自适应路由方案很难实现：因为任何基于反应的流量调控都至少需要每秒运行一次

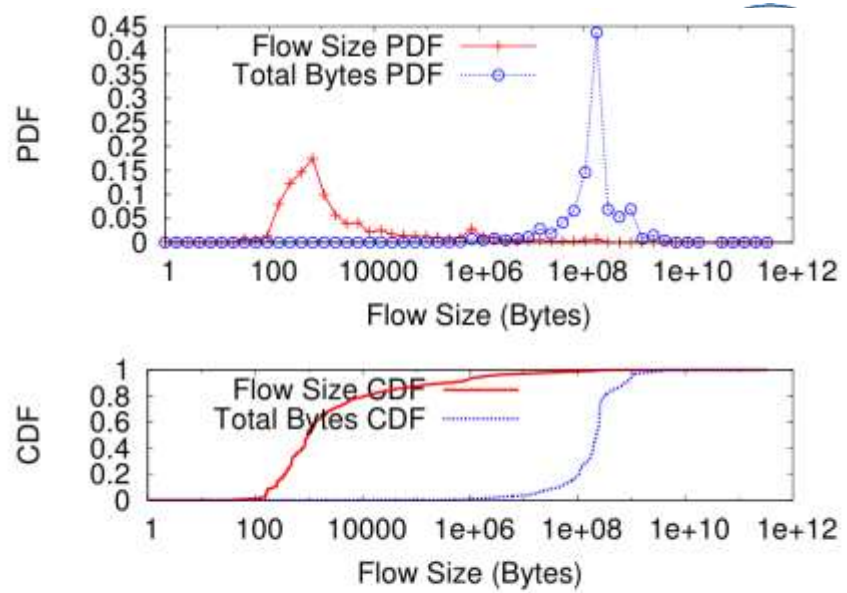


Figure 2: Mice are numerous; 99% of flows are smaller than 100 MB. However, more than 90% of bytes are in flows between 100 MB and 1 GB.

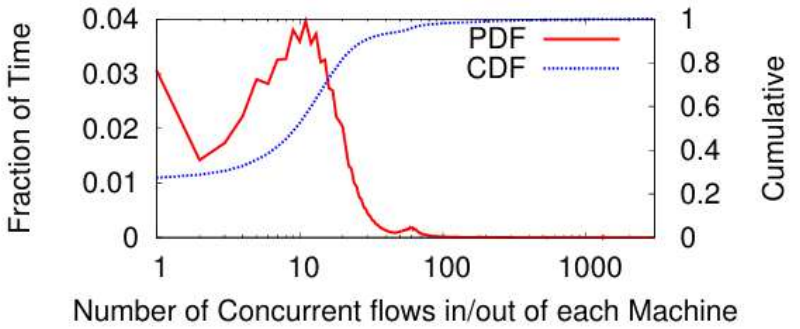


Figure 3: Number of concurrent connections has two modes: (1) 10 flows per node more than 50% of the time and (2) 80 flows per node for at least 5% of the time.

*解释：VLB以往在包粒度进行随机化，并被证明能有效缓解阻塞；2009年时云服务DCN大象流也不大，故作者按流粒度处理

Outline

I. Introduction

II. Background

III. Measurements & Implications

1. Data-Center Traffic Analysis
2. Flow Distribution Analysis
3. **Traffic Matrix Analysis**
4. Failure Characteristics

IV. Virtual Layer Two Networking

V. Evaluation

VI. Review

流量矩阵分析 – 不可总结性

流量中是否存在某种规律？可以通过仔细观测和流量工程利用？

- 作者以100s为时间区间测量1500服务器集群的流量矩阵(TM)
- $TM(t)_{i,j}$代表ToRi和ToRj在时间t开始的100s间的流量
- 根据TM距离进行聚类
  - 递归地把两个相似的TM归为一类
  - 两个TM的距离用交换(shuffle)多少流量可以让两个矩阵变得相似来衡量
  - 每个cluster使用一个TM作为代表，其他TM和这个TM的距离作为拟合误差
  - 当拟合误差快速增大时停止
- 结果：即便用多达50-60个cluster进行拟合，拟合误差仍高达60%
- **说明：数据中心流量的多变性不遵循某种简洁的表达方式，仅对几种TM做流量工程很难适用于实际的流量**

流量矩阵分析 – 不稳定性

给定当前流量模式，下一时刻的流量模式可预测吗？

- 使用前面的聚类技术找到40个最优的cluster，并对流量进行分类
- 结果如下图
- (a) 流量模式变化快速，无周期规律
- (b) 流量持续时间短
- (c) 相同模式重复出现无规律

流量模式的随机性来源于数据中心应用采用随机数来提升性能。

流量模式观测表明其他路由策略的性能很难优于VLB。

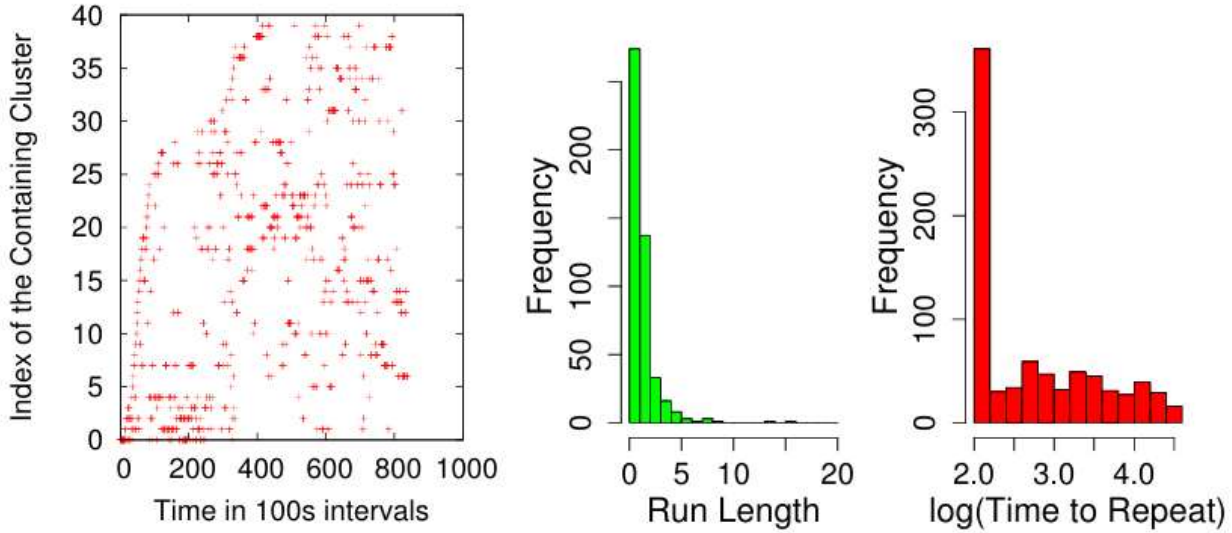


Figure 4: Lack of short-term predictability: The cluster to which a traffic matrix belongs, i.e., the type of traffic mix in the TM, changes quickly and randomly.

Outline

I. Introduction

II. Background

III. Measurements & Implications

1. Data-Center Traffic Analysis
2. Flow Distribution Analysis
3. Traffic Matrix Analysis
4. **Failure Characteristics**

IV. Virtual Layer Two Networking

V. Evaluation

VI. Review

故障特征

作者收集了有十几万服务器，部署了上百云服务，服务百万用户的8个数据中心长达一年的故障记录

网络设备故障的模式	网络设备故障的影响
➢ 大部分故障规模很小：50% < 4设备 ➢ 大型连锁故障很少见 ➢ 然而，故障时间(downtime)可能很长	➢ 如前所述，传统网络采用1:1冗余来提升容灾能力，但在0.3%的故障中，一个网络设备组的所有设备不可用。 ➢ Core switch的故障可能影响几百万用户几小时。

Outline

I. Introduction

II. Background

III. Measurements & Implications

IV. Virtual Layer Two Networking

V. Evaluation

VI. Review

设计原则

- **用随机应对波动**
  VL2通过Valiant Load Balancing将流量分发到多个中间节点，这一过程目的地无关，从而应对流量矩阵的发散和不可预测。
- **基于已被验证的网络技术**
  VL2基于商用交换机可用的IP路由和转发技术。使用链路状态路由协议维护交换机级别的拓扑，用 ECMP + anycast 地址实现VLB，同时减少控制面消息和震荡。
- **分离名字和位置**
  数据中心网络必须支持敏捷性，这尤其意味着：能够在任意服务器上承载任意服务，以及支持虚拟机的快速迁移。为此，就需要将名称与位置分离。VL2 使用一个可扩展、可靠的目录系统来维护名称与位置之间的映射关系。
- **拥抱终端系统**
  将部分控制能力从网络侧转移到编程灵活的终端侧，例如调整VLB的随机化方式

Outline

I. Introduction

II. Background

III. Measurements & Implications

IV. Virtual Layer Two Networking

1. Scale-out Topologies

2. VL2 Addressing and Routing

3. VL2 Directory System

V. Evaluation

VI. Review

4.1 Scale-out Topologies

- Scale-out, not scale-up
  使用大量简单便宜的设备搭建“更宽”的网络，提供强大的汇聚能力
- 核心交换机(Int)和汇聚交换机(Aggr)形成完全二分图
- 即使一个核心交换机发生故障，二分链路带宽只会减少1/n
- 搭建没有over-subscription的Clos网络更加简单、便宜

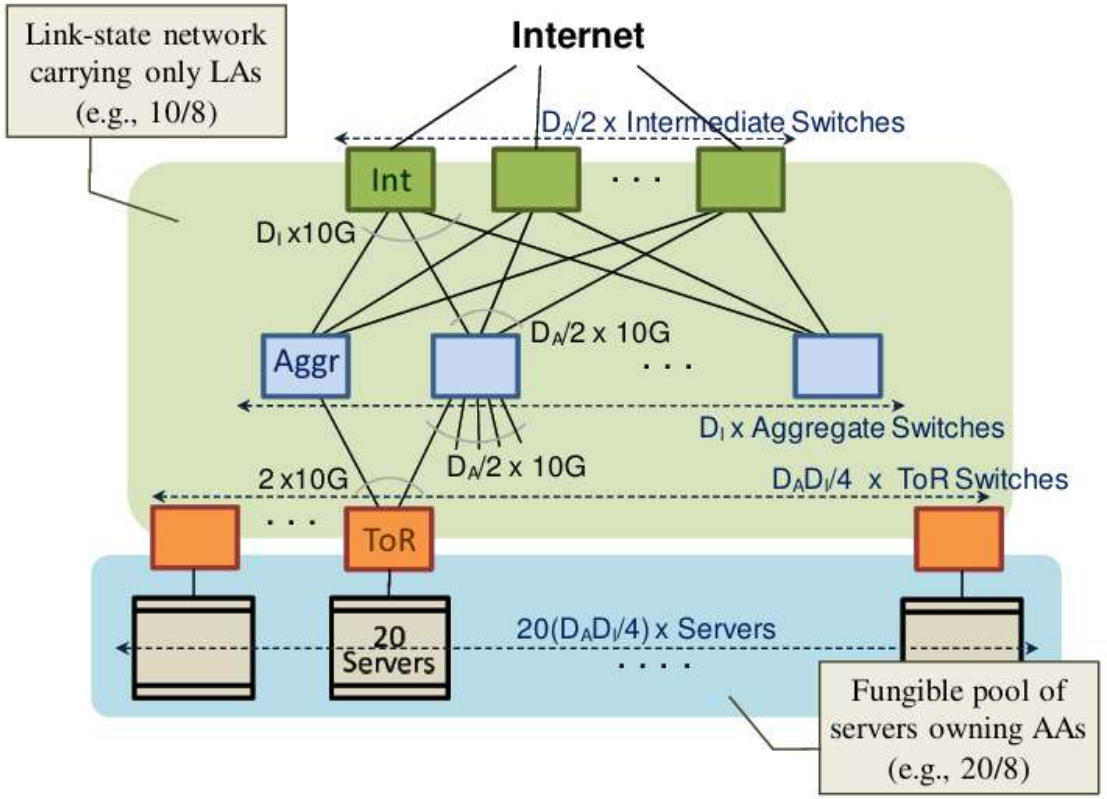


Figure 5: An example Clos network between Aggregation and Intermediate switches provides a richly-connected backbone well-suited for VLB. The network is built with two separate address families — topologically significant Locator Addresses (LAs) and flat Application Addresses (AAs).

4.1 Scale-out Topologies

- 天然适合VLB
  - 通过让流量先绕到网络顶层再下转，网络就能对符合 hose 模型*的任意流量矩阵提供带宽保证
  - 路由简单可靠
- 利用服务器-交换机带宽通常小于交换机-交换机带宽的特性，相比fat-tree，使用更少的链路实现Clos拓扑

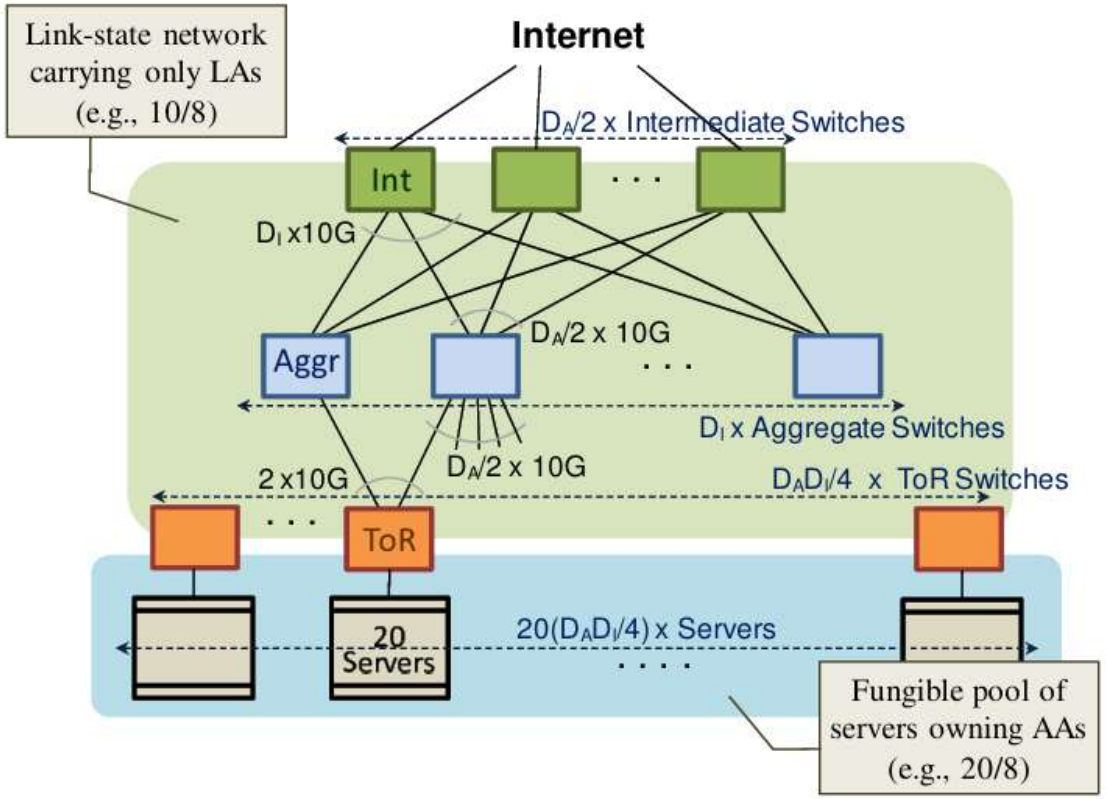


Figure 5: An example Clos network between Aggregation and Intermediate switches provides a richly-connected backbone well-suited for VLB. The network is built with two separate address families — topologically significant Locator Addresses (LAs) and flat Application Addresses (AAs).

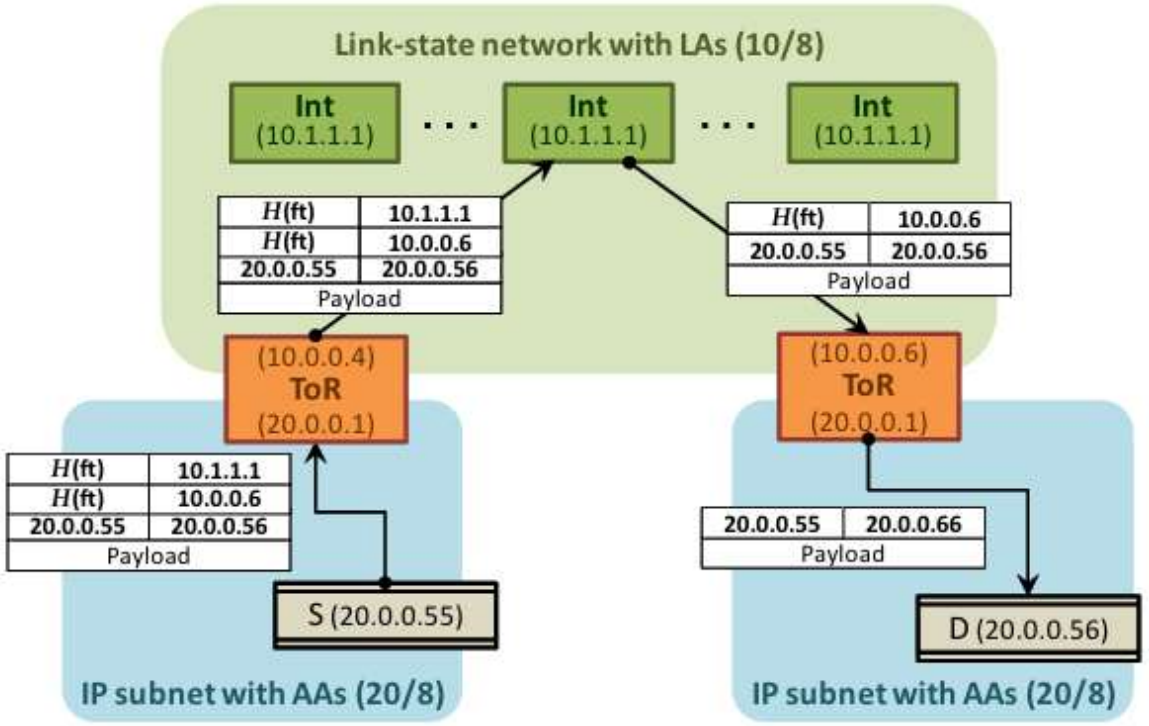
*hose model：只规定每台服务器进/出总速率 ≤ 某一限速，不单独限制谁跟谁通信的速率

Outline

I. Introduction
II. Background
III. Measurements & Implications
IV. Virtual Layer Two Networking
1. Scale-out Topologies
2. VL2 Addressing and Routing
3. VL2 Directory System
V. Evaluation
VI. Review

4.2.1 地址解析与包转发

- VL2使用两个不同的IP地址族
- **网络设备**使用**位置**相关地址(Location-specific IP address, **LA**)
  - 交换机及其接口分配了LA地址
  - 交换机运行L3链路状态协议获取完整的交换机级别的拓扑
- **应用程序**使用**应用**相关地址(Application-specific IP address, **AA**)
  - 无论虚拟机怎么迁移都保持不变
- **VL2目录系统**存储AA -> LA的映射



- AA: 20/8
- LA: 10/8

4.2.1 地址解析与包转发

包转发

1. Server上的VL2 agent**截获**来自host的包，使用目的ToR的LA作为目的地址**封装**包
2. 当包到底目的LA(目的ToR)，交换机对包**解封装**并交给目的AA

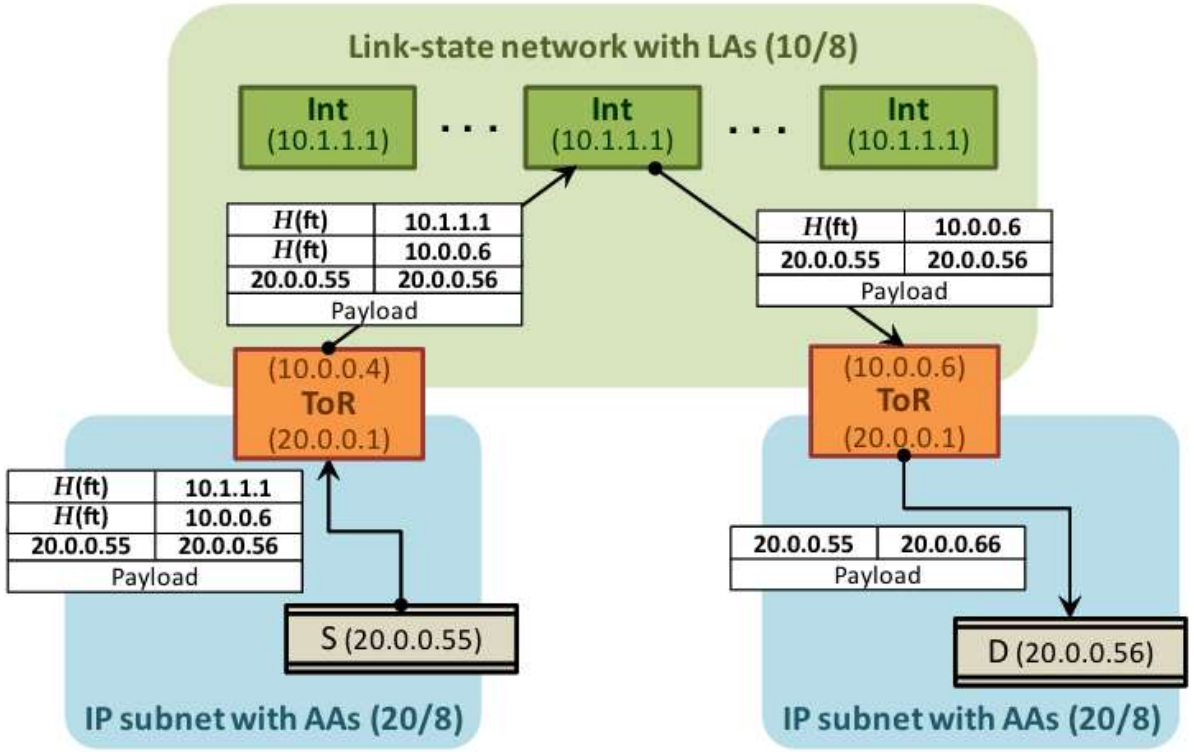


Figure 6: VLB in an example VL2 network. Sender S sends packets to destination D via a randomly-chosen intermediate switch using IP-in-IP encapsulation. AAs are from 20/8, and LAs are from 10/8. H(ft) denotes a hash of the five tuple.

4.2.1 地址解析与包转发

地址解析

1. 当一个应用程序第一次向某AA发送数据包时，host上的网络栈发送ARP广播请求
2. Host上运行的VL2 agent会截获ARP请求，并将其转换成一个单播请求发送给VL2目录系统
3. 目录系统使用目的ToR的LA做应答
4. VL2 agent缓存AA -> LA的映射

通过目录服务实现访问控制

没有AA->LA映射就无法完成通信，所以目录系统可以实现访问控制，例如只允许同一服务的host进行地址映射等。

VL2编址和转发规则的优势

1. 可以使用路由表较小的廉价交换机(只用存储LA，无需管理数量众多的AA)
2. 降低控制面的负担，将负载转移到可扩展性更强的目录系统

4.2.2 多路随机流量分摊

- 为了给任意流量矩阵提供无热点的性能，VL2采用了两种相互关联的机制：**VLB**和**ECMP**。VLB采用流作为流量分摊的粒度，避免乱序。
- 如图所示，VL2 agent把流量发送给随机选择的核心交换机，数据包在核心交换机和目的ToR经过两次解包之后送达目的AA

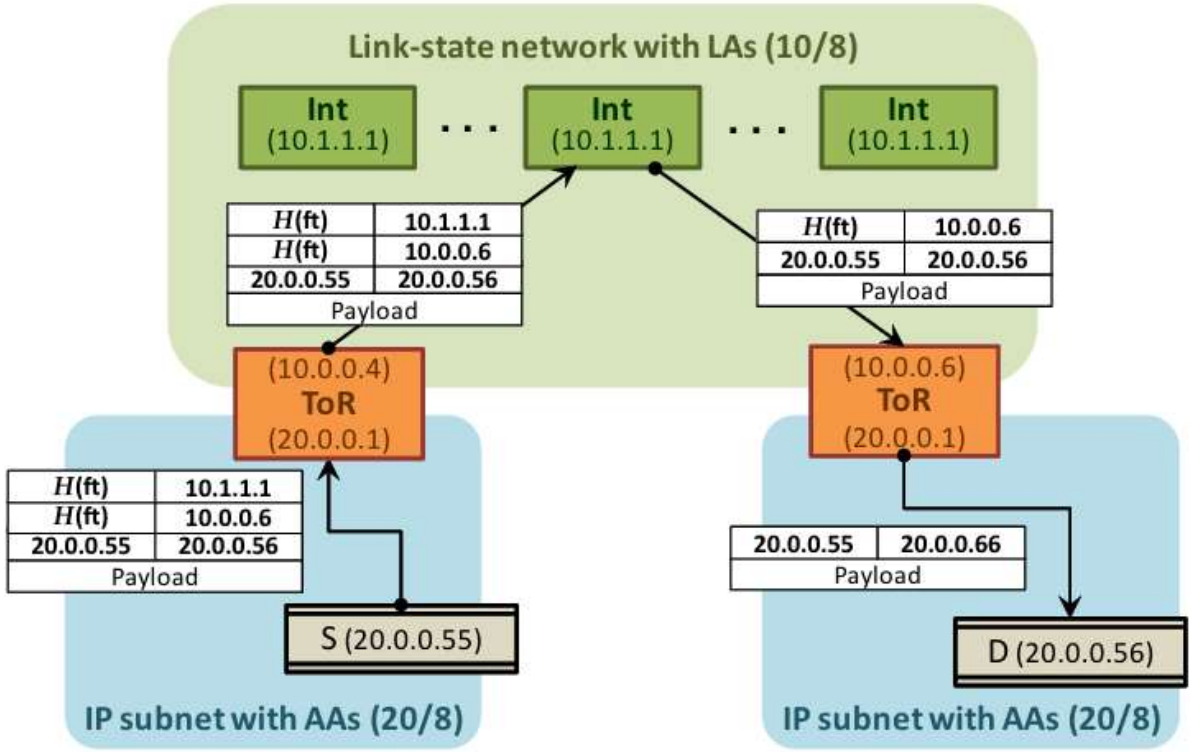


Figure 6: VLB in an example VL2 network. Sender S sends packets to destination D via a randomly-chosen intermediate switch using IP-in-IP encapsulation. AAs are from 20/8, and LAs are from 10/8. H(ft) denotes a hash of the five tuple.

4.2.2 多路随机流量分摊

- **任播的使用**：把数据包封装到一个随机选择的特定核心交换机也能实现VLB，但是如果核心交换机的可用性发生变化(由于交换机/链路故障)，需要更新大量的VL2 agent。所以作者选择给所有的核心交换机相同的LA
- **ECMP**：由于所有的核心交换机距离host都是3跳远，ECMP可以把封装的数据包发送到任意活动的核心交换机。当链路发生故障时，ECMP可以自行做出处理
- **实际问题**
  - 许多交换机只支持16路ECMP：VL2定义多个任播地址，核心交换机发生故障时，对该交换机的任播地址进行重新分配
  - 有些交换机在多层IP封装时无法正确获得五元组：VL2 agent预先把五元组哈希算好，结果写进外层 IP 的源地址字段；交换机即使只看外层 srcIP，也能得到均匀分布。

4.2.3 向后兼容性

和互联网中的主机通信

- 由于VL2部署了L3路由交换结构来实现虚拟的L2网络，外部流量可以直接进入VL2交换机高速硅片，无需现在网关改header
- 需要被互联网访问的服务器(例如前端web服务器)在用于DC内部通信的AA以外还分配一个LA。该LA由BGP管理，可以由外部访问。

处理广播

VL2为应用提供了L2语义保证向后兼容性，其中包括支持广播和组播。VL2完全消除了大部分广播的源头：ARP和DHCP。

- ARP由目录系统取代
- DHCP消息被ToR使用DHCP relay agent拦截，单播到DHCP服务器
- 为了处理其他的L2广播流量，每个服务分配一个IP组播地址，广播被转换为该服务的组播

Outline

I. Introduction
II. Background
III. Measurements & Implications
IV. Virtual Layer Two Networking
    1. Scale-out Topologies
    2. VL2 Addressing and Routing
    3. VL2 Directory System
V. Evaluation
VI. Review

4.3 VL2 Directory System

VL2的目录系统提供了3项**关键功能**:

(1) 查询AA->LA映射

(2) 更新AA->LA映射

(3) 响应式缓存更新机制快速响应延迟敏感的更新

工作负载特征

- lookup频繁，突发：服务器可能短时间和大量其他服务器通信，造成大量AA->LA查询
- update通常由故障或服务器启动事件驱动

系统需求

- 性能需求
  - lookup：突发性->需要高带宽，低响应时间；
  - update：可靠性更重要
- 一致性需求
  L2网络可以通过ARP缓存超时的机制保证IP->MAC的最终一致，目录系统也需要AA->LA的一致

4.3 VL2 Directory System

目录系统设计

根据查找和更新不同的性能需求和工作负载特征，设计了两层的目录系统结构

- 中等数量的读优化的**目录服务器(DS)**，缓存映射，处理请求
- 少量写优化的**复制状态机(RSM)服务器**提供强一致，可靠的AA->LA映射

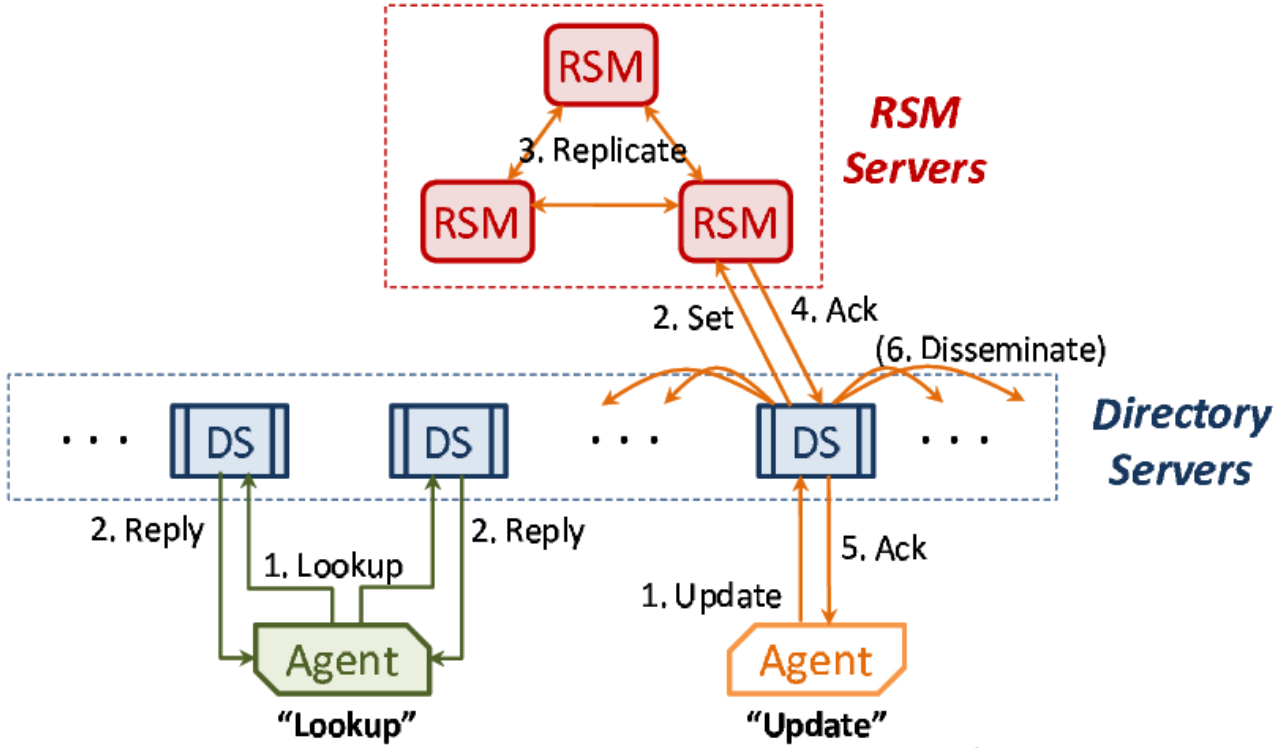


Figure 7: VL2 Directory System Architecture

4.3 VL2 Directory System

查询路径

- 每个agent向k个随机选择的目录服务器发送查询请求
- 每个目录服务器根据缓存独立地回应agent的请求
- agent选择最快的回复保存在本地缓存
- 目录服务器定期和RSM服务器同步映射信息

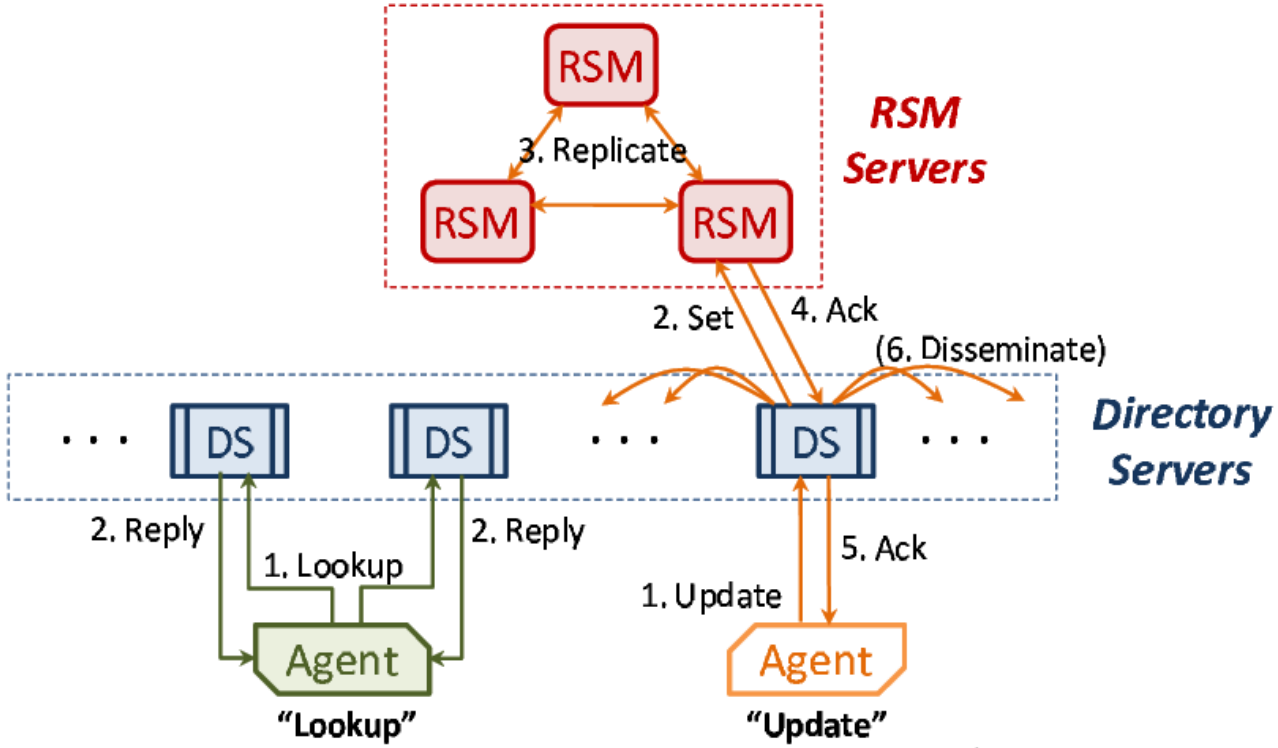


Figure 7: VL2 Directory System Architecture

4.3 VL2 Directory System

更新路径

- 网络资源分配系统向一个随机选择的目录服务器发送目录更新
- 目录服务器把目录更新转发给RSM服务器
- RSM服务器根据Paxos算法可靠地和其他RSM服务器复制更新，并向目录服务器发送ACK
- 目录服务器向原客户端发送ACK
- 如果timeout，客户端会向另一个目录服务器发送更新请求

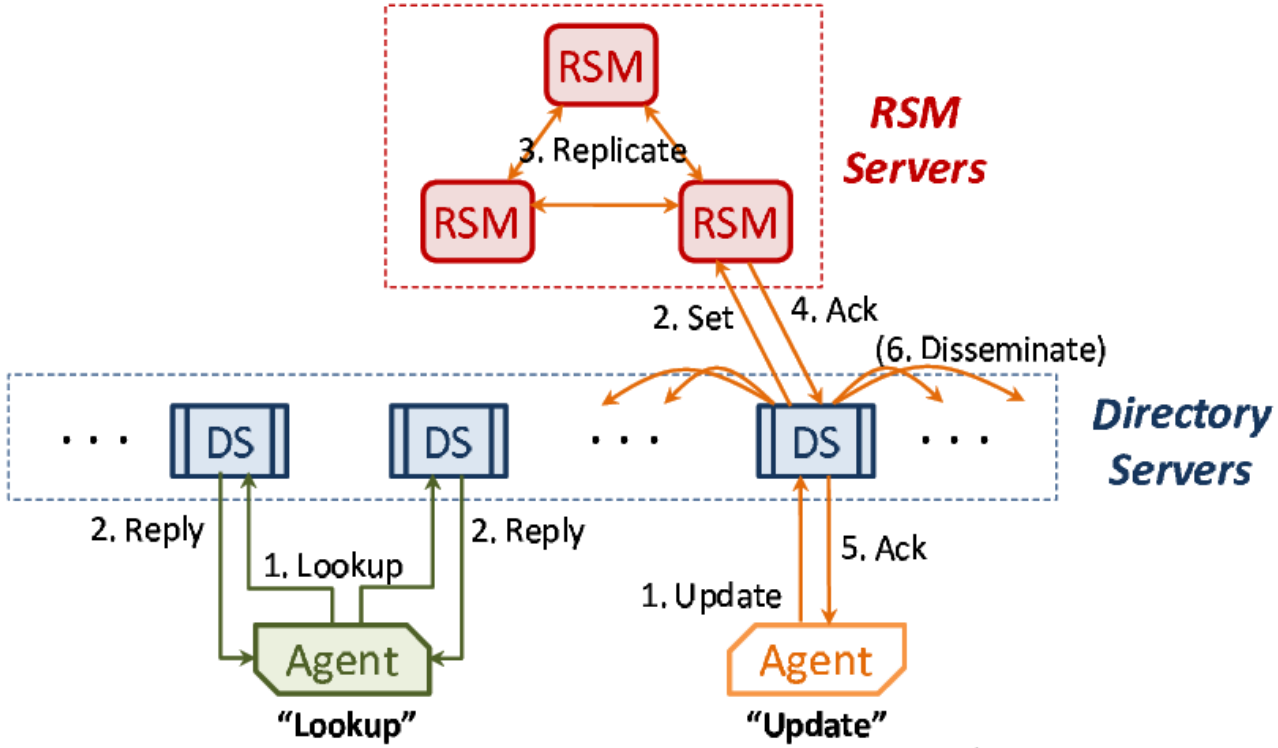


Figure 7: VL2 Directory System Architecture

4.3 VL2 Directory System

响应式更新缓存

- 由于AA->LA映射在目录服务器和VL2 agent中缓存，可能出现不一致的情况
- 过时的映射只需要在下一次使用时更新
- 当服务器使用过时映射时，目的ToR把不可送达的包发给目录服务器，后者会更新agent的缓存

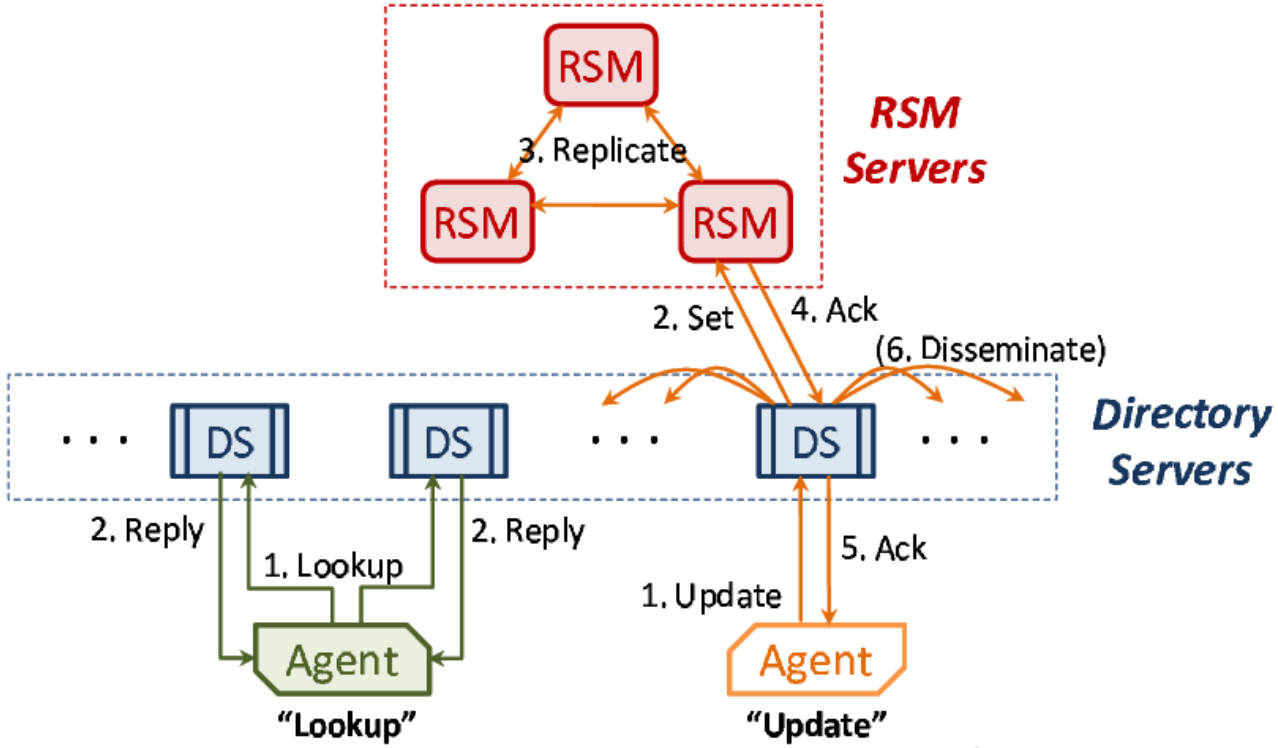


Figure 7: VL2 Directory System Architecture

Outline

I. Introduction

II. Background

III. Measurements & Implications

IV. Virtual Layer Two Networking

V. Evaluation

VI. Review

5 Evaluation

- 在80台服务器，10太商用交换机的testbed进行实验
- Clos架构：3个核心交换机(启用3端口)，3个汇聚交换机(启用6端口)，4个ToR交换机(2路上行，20路下行)
- 交换机-交换机带宽：10Gbps
- 服务器-交换机带宽：1Gbps

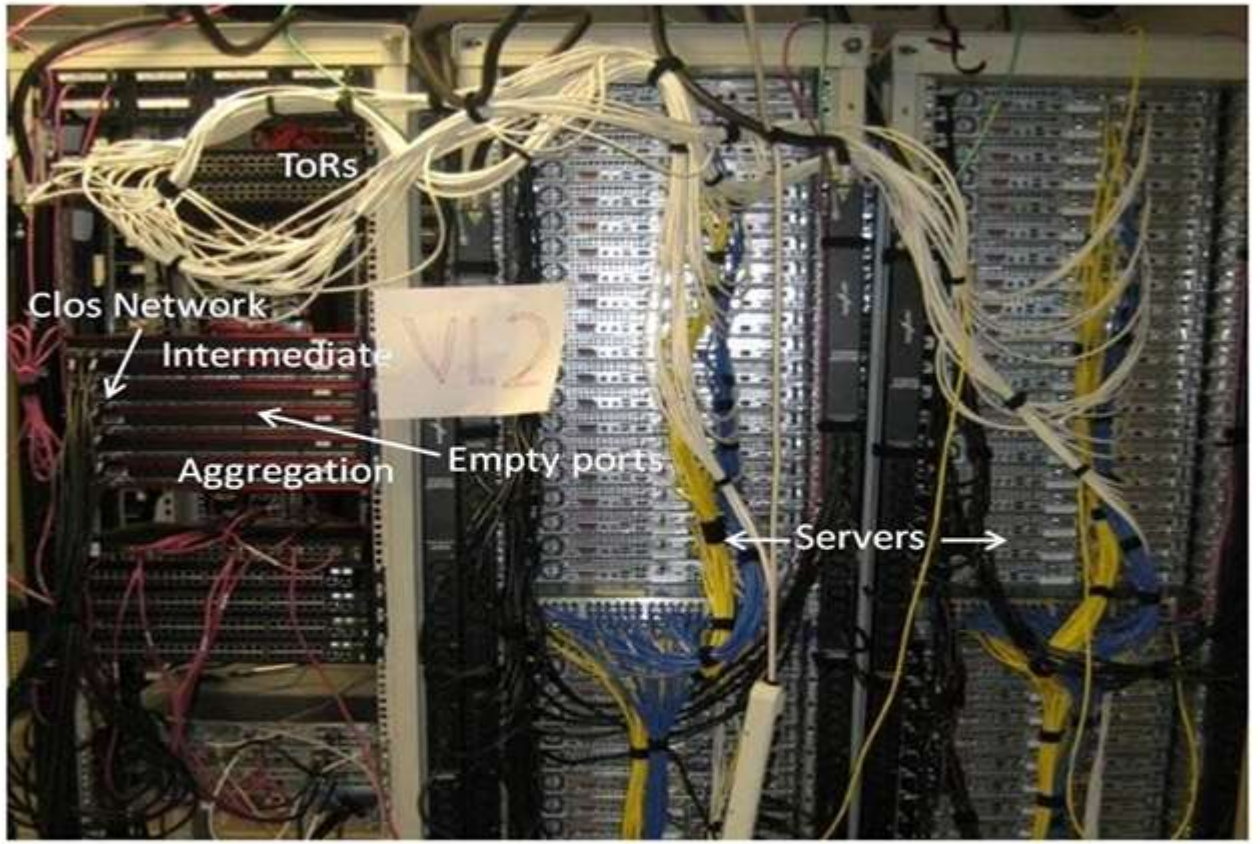


Figure 8: VL2 testbed comprising 80 servers and 10 switches.

5.1 VL2 Provides Uniform High Capacity

- 为了测量server-server通信的带宽，作者利用75台服务器进行all-to-all data shuffle。每台服务器向其他74台服务器各发送500MB数据
- VL2完成shuffle的时间：395s
- 链路利用率：86%
- 实验的大部分时候，VL2能达到58.8Gbps的汇聚goodput
- 流的公平性指数：0.995
- VL2 goodput效率*：94%

*goodput效率的分母不是网络接口的带宽之和，而是考虑了驱动问题和以太/IP包头后实际可达的最大goodput传输速率

94%的goodput效率和0.995的公平性指数表明VL2可以为DCN服务器提供统一高带宽。

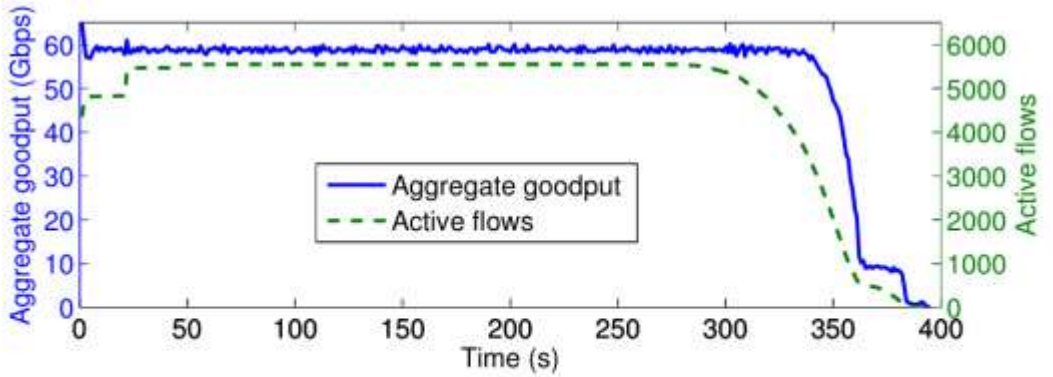


Figure 9: Aggregate goodput during a 2.7TB shuffle among 75 servers.

5.2 VL2 Provides VLB Fairness

- 测量原因：ECMP以流为粒度实现VLB，短时间内可能存在大象流和老鼠流在核心交换机划分不均匀的情况。
- 在75节点testbed实验，模仿实际的流量特征，流的并发数和大小根据第三章观测实验得到的分布随机选取。
- 每隔10s通过SNMP数据获取每个汇聚交换机到核心交换机的链路情况。
- 计算到核心交换机的Jain公平性指数的时间序列，如图所示
- 每个汇聚交换机在实验的时间跨度上的平均公平性指数均在0.98以上
- 这种结果的出现是因为汇聚交换机上的流足够多，静态多路有效

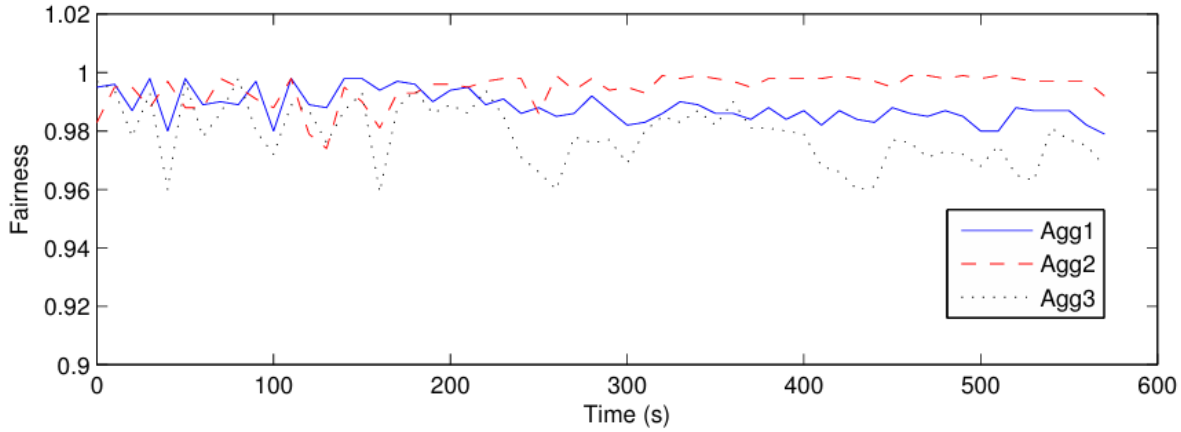


Figure 10: Fairness measures how evenly flows are split to intermediate switches from aggregation switches.

5.3 VL2 Provides Performance Isolation

- VL2的敏捷性依赖于VLB的数学证明：只要流量模型符合hose model，即每台服务器的进/出速率低于某个限速，流量就能被均匀分摊避免拥塞。
- VL2没用采用复杂的准入机制或流量调控，而是采用TCP来限制每个流的速率。

两个问题：

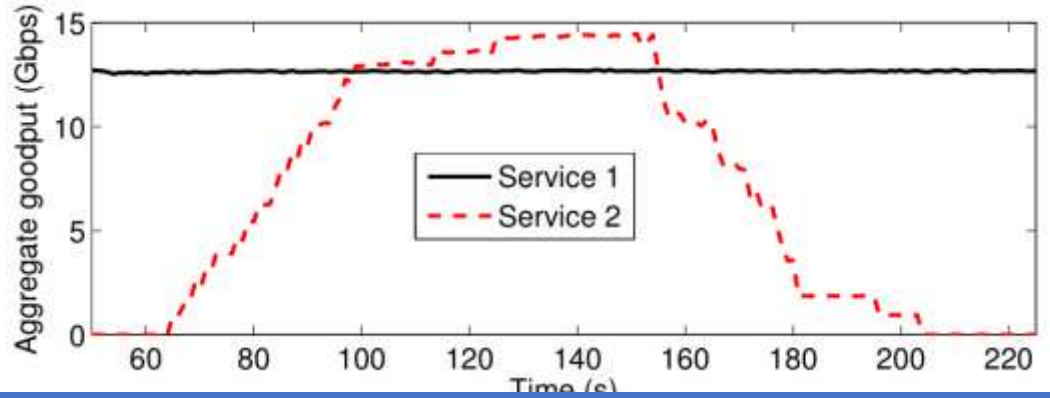
- TCP调整流量以RTT为时间跨度，可能难以满足速率的及时性调控要求
- 大量突发的mice flow，由于slow start，可能短时间内造成突发流量，违反hose model

作者设计实验验证了TCP足以胜任hose model的速率调控。

5.3 VL2 Provides Performance Isolation

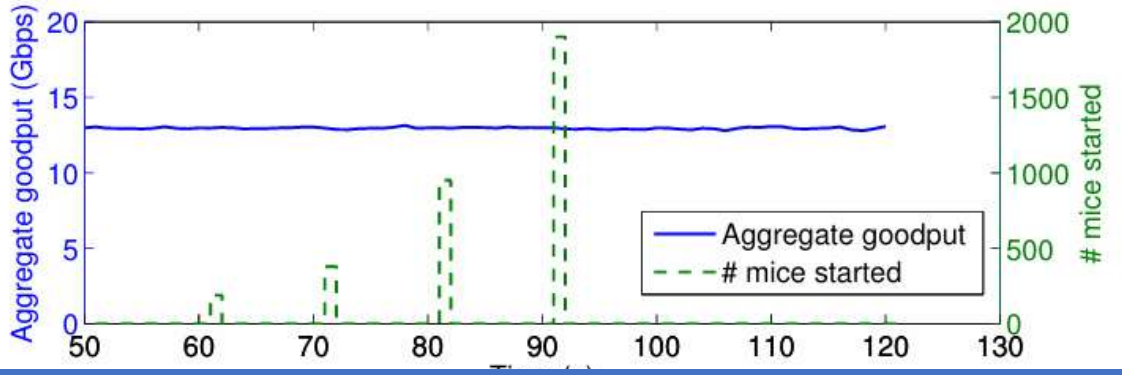
实验一：大象流隔离

- 服务一有一组持续通信的服务器
- 服务二每隔2s加入新的服务器发送8GB大象流
- 观测服务一的goodput随时间变化
- 结果：无明显变化



实验二：老鼠流隔离

- 服务一有一组持续通信的服务器
- 服务二产生老鼠流burst，且每次burst规模逐渐增大
- 观测服务一的goodput随时间变化
- 结果：无明显变化



实验证明TCP天然适合与VLB配合执行hose model；没有over-subscription的网络也可以做到服务间性能隔离！

vice two ramps traffic up and down.

5.4 VL2 Convergence After Link Failures

错误处理能力

- 服务为数据混洗(data shuffle)
- 把聚合交换机和核心交换机之间的链路逐渐断掉，直到只剩下一条；再逐渐恢复
- 图为服务的聚合带宽；图中竖线为链路断开和恢复的时间

- 结果：
  - 链路故障后可以快速重新收敛
  - 聚合速率可以优雅地降低
  - 链路恢复后速率恢复较慢(由于OSPF计时器的保守设置)
  - 结果也展示了VL2在over-subscription情形下也可以充分利用带宽

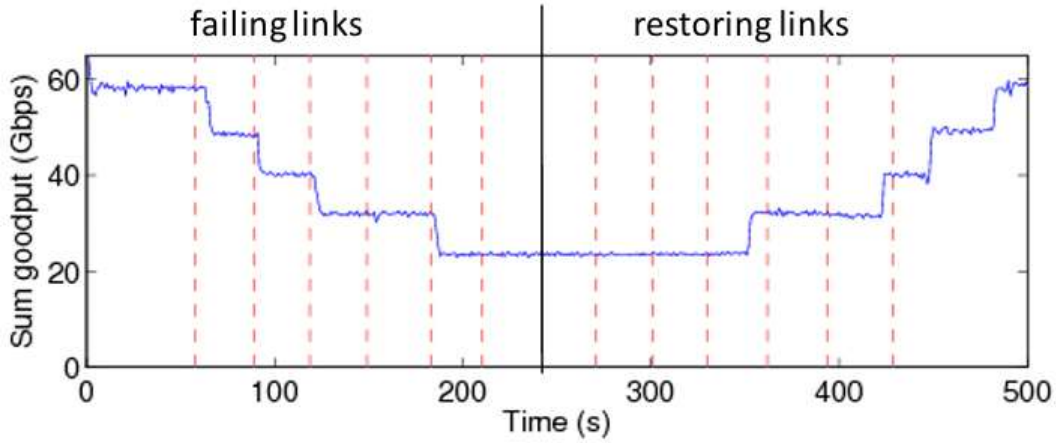


Figure 13: Aggregate goodput as all links to switches Intermediate1 and Intermediate2 are unplugged in succession and then reconnected in succession. Approximate times of link manipulation marked with vertical lines. Network re-converges in < 1s after each failure and demonstrates graceful degradation.

5.5 Directory-system performance

- 目录服务器提供高带宽，低延迟的查询服务
- 目录系统处理更新的速度远快于现实网络发生变化的速度
- 扩展性良好：响应查询的速度随目录服务器数量线性增长
- 目录系统对组件的故障鲁棒性良好

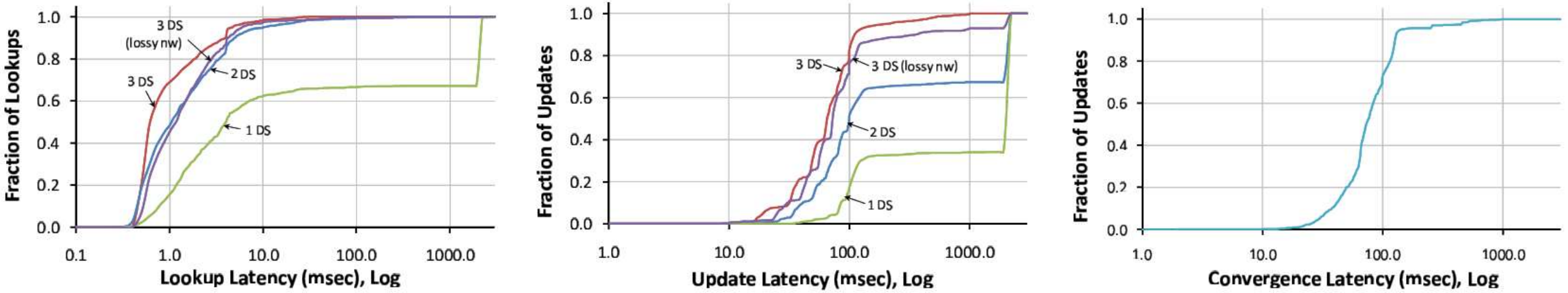


Figure 14: The directory system provides high throughput and fast response time for lookups and updates

Outline

I. Introduction
II. Background
III. Measurements & Implications
IV. Virtual Layer Two Networking
V. Evaluation
VI. Review

Review

核心目标

- 实现数据中心的敏捷性（Agility）：支持任意服务器分配给任意服务
- 提供统一高带宽：服务器间通信速率仅受限于网卡，而非网络瓶颈
- 性能隔离：服务之间互不干扰，像连接在独立交换机上
- 保持L2语义：支持VM热迁移、ARP、广播等，兼容传统应用

Review

关键设计

- Scale-out 拓扑（Clos/Fat-tree）
  - 使用大量廉价交换机构建高带宽、无阻塞网络。
  - 多路径容错，链路故障时带宽优雅降级。
- Valiant Load Balancing (VLB)
  - 每个流随机选择中间交换机转发，避免热点。
  - 基于流级别（flow-level）而非包级别，避免乱序。

Review

关键设计

- 地址分离机制
  - AA（Application Address）：服务使用的地址，与拓扑无关。
  - LA（Locator Address）：服务器实际位置，网络路由使用。
  - 通过目录系统（Directory System）维护AA→LA映射。
- 向后兼容
  - 无需修改交换机硬件或应用。
  - 支持IP广播、ARP、DHCP等传统网络功能。

Review

- VL2通过简单的设计、可部署的技术，解决了传统DCN的带宽瓶颈、资源碎片化和性能隔离缺失等问题。
- 实现了理论上的高带宽与实际部署的可行性之间的良好平衡。
- 是首个将VLB应用于数据中心网络、并验证其有效性的研究。

Review

我们上节课学习的Fat-tree和这节课介绍的VL2有相似之处：

- 目标：都采用廉价的商用交换机构建可扩展的数据中心网络
- 拓扑：都基于叶-聚合-核心的Clos结构，利用路径冗余来提高可用带宽与容错性

Review

但二者存在关键不同。从定位上讲，Fat-tree主要关注**拓扑/物理结构**，以及基于该拓扑的**简单路由算法**；而VL2是完整的**云服务**数据中心网络设计，更加全面。VL2可以运行在Fat-tree或其他Clos架构上，并在此之上做了较多的设计来满足**云和虚拟化场景**的需求，例如：

	Fat-tree	VL2
编制方式	直接使用位置相关的寻址	强调位置无关，便于VM热迁移
流量分散	通常使用静态路由	使用随机化方案，实现负载均衡、消除热点、减少服务间干扰程度

Review

由此可见，**VL2**会更加适用于以下场景：

- 云环境，支持大量VM，支持VM的灵活迁移
- 未知或突发的流量模型，避免复杂流量工程
- 东西向流量主导的应用：避免热点