思科下一代数据中心交换网络架构分析

马旻

行业解决方案部

思科系统（中国）网络技术有限公司

2013年9月

前言

在最近短短的3个月内，思科连续宣布两个重大的数据中心网络架构Dynamic Fabric Automation (DFA) 和Application Centric Infrastructure (ACI)说明数据中心产品线和解决方案在整个公司研发战略中的重要位置。

对于金融行业来讲，数据中心不仅是金融信息化的生命线，也是思科的生命线所在。所以学习和运用好DFA和ACI架构，以及如何在此基础上进行新的数据中心架构设计对于金融行业的SE具有重要的意义。

© 2013 Cisco and/or its affiliates. All rights reserved.

前言

Dynamic Fabric Automation (DFA) 架构的重点在于对交换矩阵网络结构的重新优化。
在技术底层上加强了控制层面的处理手段，使之支持更大规模的DC网络和更可靠稳定的DC网络。
相对之前的VPC技术和FabricPATH技术更前进一步，解决了VPC和FabricPATH存在的扩展性和稳定性的问题。
在对于多租户的虚拟化的支持能力上也进行了加强。
在管理层面，加强DCNM的能力，提出集中管理、自动化配置的方法。
DFA主要解决现有DC架构的问题，进一步优化提升。其产品的组成（N7K F3，N6K+N2K，DCNM）和运用部署模式比较传统，应用场合的普适性较强。

Application Centric Infrastructure (ACI) 架构的重点在于提供网络、应用、安全和IT运维的公共的管理框架，极大的方便应用的迁移和部署，旨在减少应用部署时间。
ACI提供集中化的应用驱动的策略自动化操作，对于物理和虚拟网络整体提供可视化和管理。
ACI将应用连接需求自动化和架构部署结合在一起，简化架构的运维复杂度，提供高效、可扩展、应用感知的基础架构。
ACI主要面向大规模云计算环境，强调架构对应用部署的感知能力，这也是ACI运用部署的适用面问题。
个人认为在国内金融领域推广ACI，需要针对金融客户的运维模式自上而下进行顶层设计，所解决的首要问题可能不是技术本身，而是像FCOE、UC推广所遇到的同样的问题。

© 2013 Cisco and/or its affiliates. All rights reserved.

前言

对于DFA架构的几点误解：

误解1：DFA只适用于大型DC网络
DFA重新设计优化了DC网络的控制平面、转发平面、管理平面，提供了更加稳定可靠、易于扩展、部署灵活的架构设计方法，与客户网络的大小规模没有关系。DFA架构的精髓在于其控制平面和转发平面，并由此带来的网络结构设计的灵活性，对于小型DC网络，管理平面和自动化配置都是可选项。

误解2：采用MP-BGP承载主机路由将降低整体DC网络的收敛速度。
DFA架构针对MP-BGP做了多项的Enhancements。
在收敛速度上“The convergence is expected to be similar to that of an IGP protocol”
在功能上，添加了诸多的增强，如 Host Route Distribution，Route-version Cost Community，iBGP “maximum-paths” ，
Blocking Host Routes to the External Network，Subnet/Host Route Advertisement，“additional-paths” 等。

误解3：DFA设计过于复杂，担心客户难以维护。
DFA完全不需要客户处理复杂的BGP部署策略，对外配置将非常简化，并且可以通过DCNM实现自动化配置。

误解4：DFA将被ACI取代。
正如前面所述，两者应用场景和适用范围有所区别。

其实我们的竞争对手已经有DFA like的解决方案和产品，所以了解和掌握DFA的设计思想是必要的。

© 2013 Cisco and/or its affiliates. All rights reserved.

议程

- 动态交换矩阵自动化架构（DFA）的 需求和功能
- 交换矩阵管理
- 工作负载自动化
- 最优化的网络
  - 设备角色和拓扑
  - 控制平面
  - 转发平面
- 虚拟交换矩阵
- DCI, 服务集成, 迁移和互操作
- 路标和产品

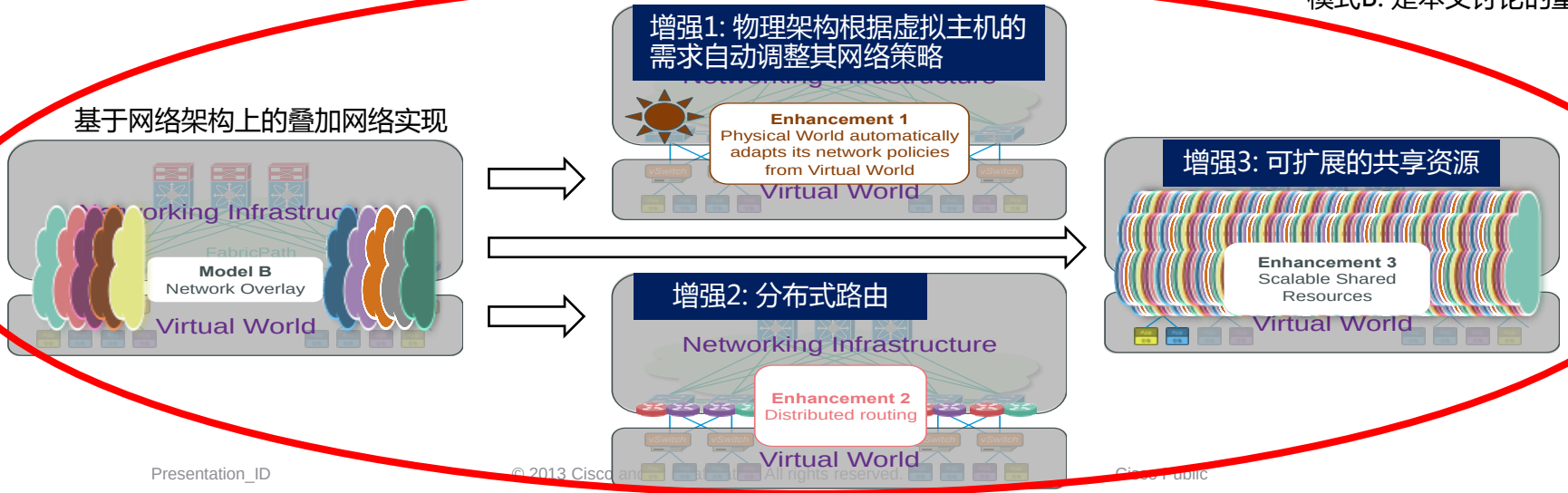
© 2013 Cisco and/or its affiliates. All rights reserved.

思科集成的叠加网络传输模式

模式A：
思科在主机叠加网络实现基础上(如VxLAN) ,
结合网络叠加功能实现混合叠加传输,
方便虚拟服务器和物理服务器之间的通信

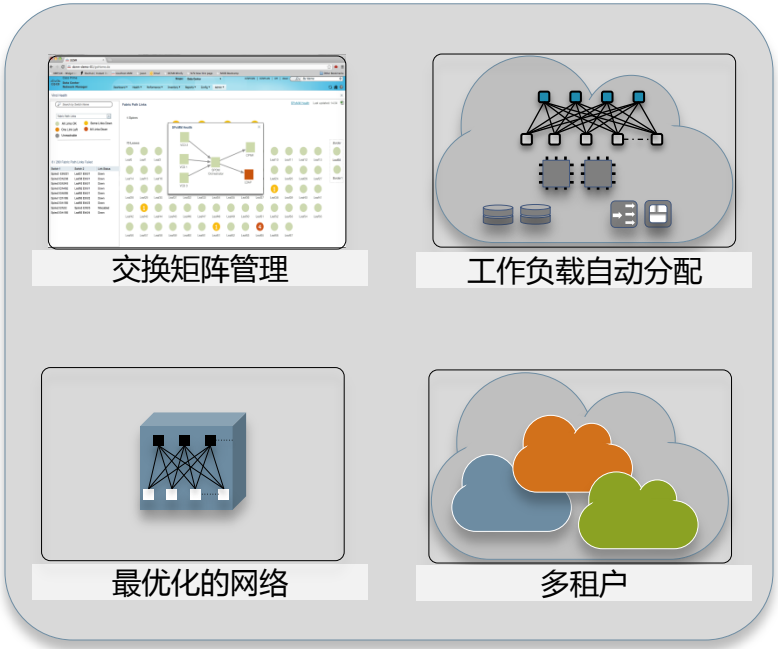
模式B: 是本文讨论的重点

基于网络架构上的叠加网络实现



© 2013 Cisco and/or its affiliates. All rights reserved.
Cisco Public

思科统一交换矩阵演进
—— 功能性



- 功能组件完全集成
- 为了操作的简化可以解耦和
- 交换矩阵的功能可以提供不断演变的优势

- 功能组件:
  - 互操作并且是松耦合
  - 可针对高级的需求开发功能
  - 可针对当前的需要增加功能
  - 和工作负载的变化高度集成

© 2013 Cisco and/or its affiliates. All rights reserved.

议程

- DFA 需求和功能
- 交换矩阵管理
- 工作负载自动化
- 最优化的网络
  - 设备角色和拓扑
  - 控制平面
  - 转发平面
- 虚拟交换矩阵
- DCI, 服务集成, 迁移和互操作
- 路标和产品

© 2013 Cisco and/or its affiliates. All rights reserved.

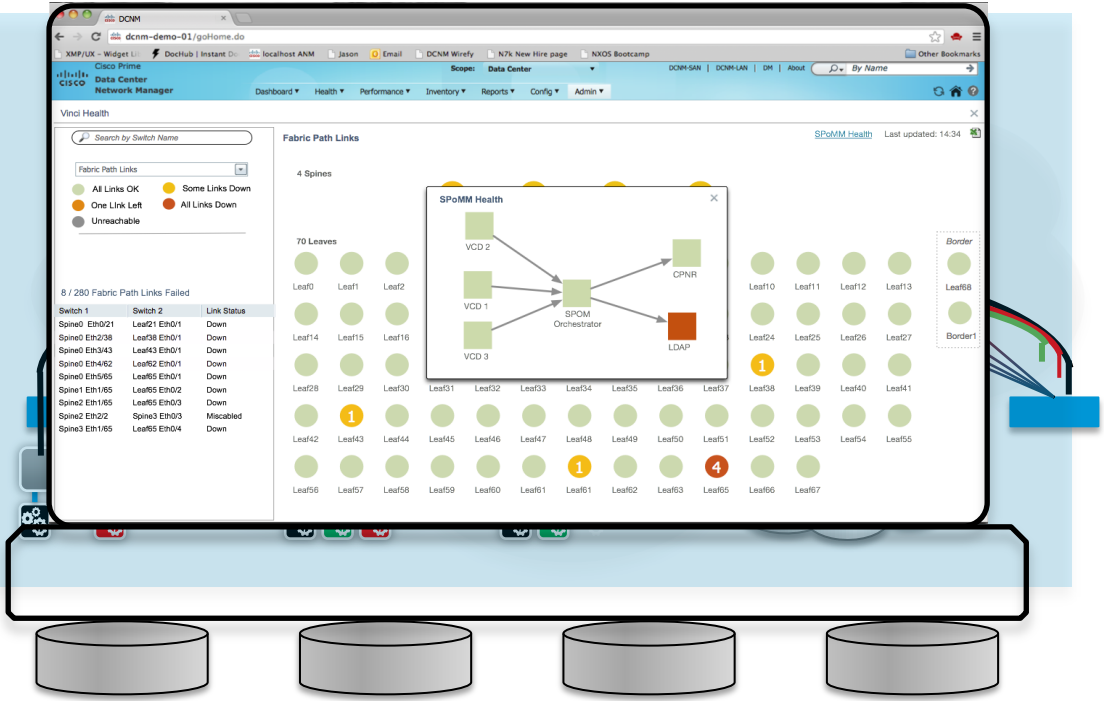
交换矩阵管理的优势

交换矩阵管理的优势

- 减少管理接触点
- 缩短排障时间
- 提供公共和单一的交换矩阵接入
- 支持网络自举
- 确保物理结构到逻辑上的一致性
- 提供可预测的设备配置
- 改善图形用户界面
- 可扩展的管理域

© 2013 Cisco and/or its affiliates. All rights reserved.

交换矩阵管理
—— 统一、自动化、感知



集成化的管理

集成管理组件

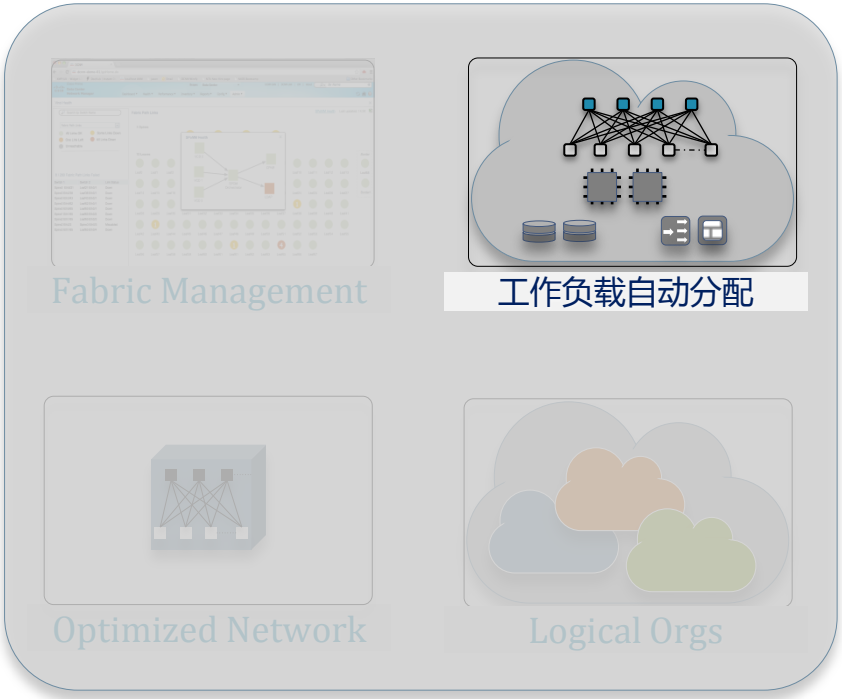
© 2013 Cisco and/or its affiliates. All rights reserved.

议程

- DFA 需求和功能
- 交换矩阵管理
- 工作负载自动化
- 最优化的网络
  - 设备角色和拓扑
  - 控制平面
  - 转发平面
- 虚拟交换矩阵
- DCI, 服务集成, 迁移和互操作
- 路标和产品

© 2013 Cisco and/or its affiliates. All rights reserved.

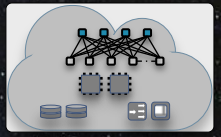
工作负载自动分配的优势

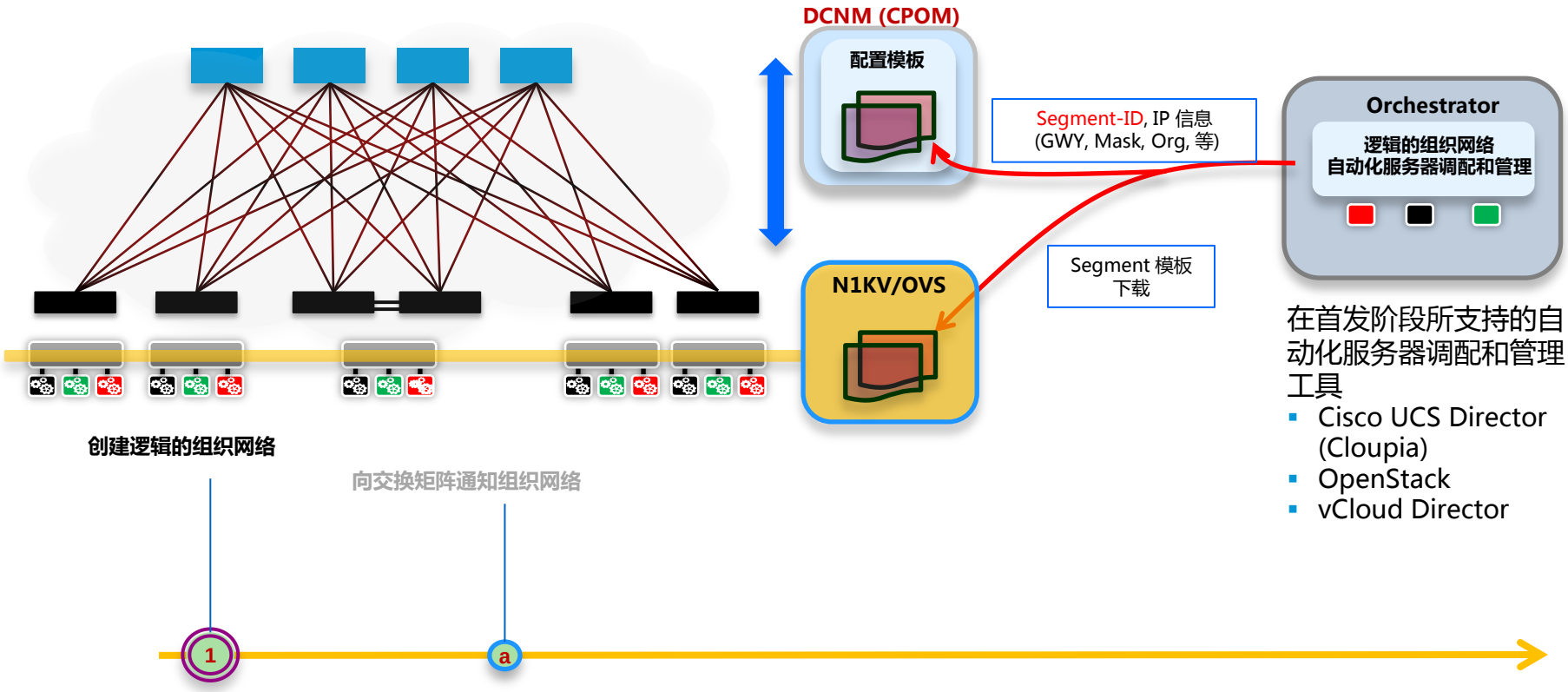


- 工作负载自动分配的优势
  - 随时、随地提供所需工作负载
  - 模板驱动的网络配置
  - 自动化的网络开通
  - 工作负载的自动化服务器调配和管理
  - 物理和虚拟服务器工作负载的归一化
  - 端到端的工作负载工作流: 计算、存储、网络和服务
  - Hypervisor, vSwitch 以及 orchestrator 自动化服务器调配和管理

© 2013 Cisco and/or its affiliates. All rights reserved.

工作负载的自动化
—— 借助VDP实现叶子节点自动配置



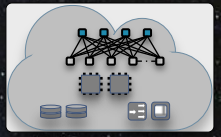


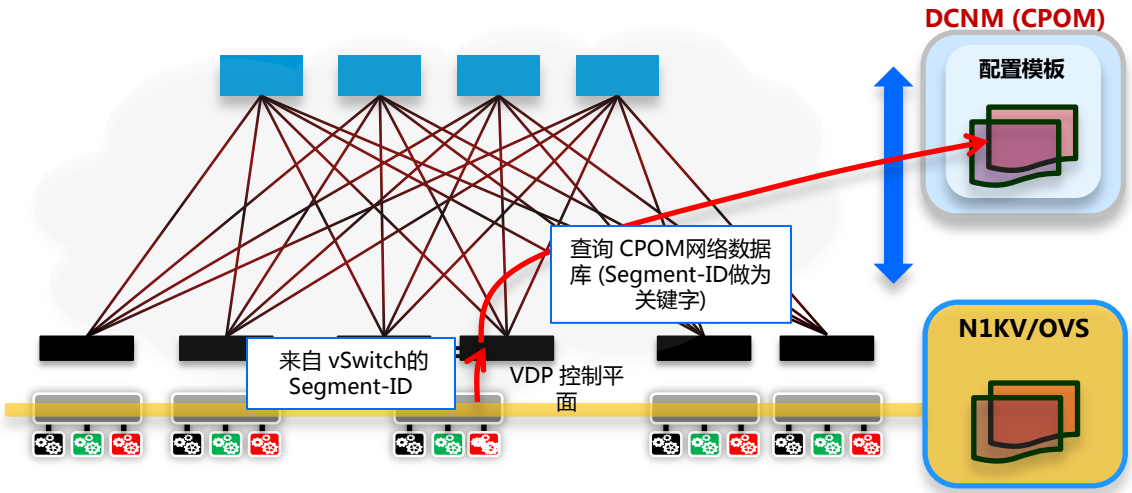
在首发阶段所支持的自动化服务器调配和管理工具

- Cisco UCS Director (Cloupia)
- OpenStack
- vCloud Director

© 2013 Cisco and/or its affiliates. All rights reserved.

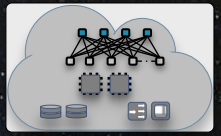
工作负载的自动化
—— 借助VDP实现叶子节点自动配置（2）

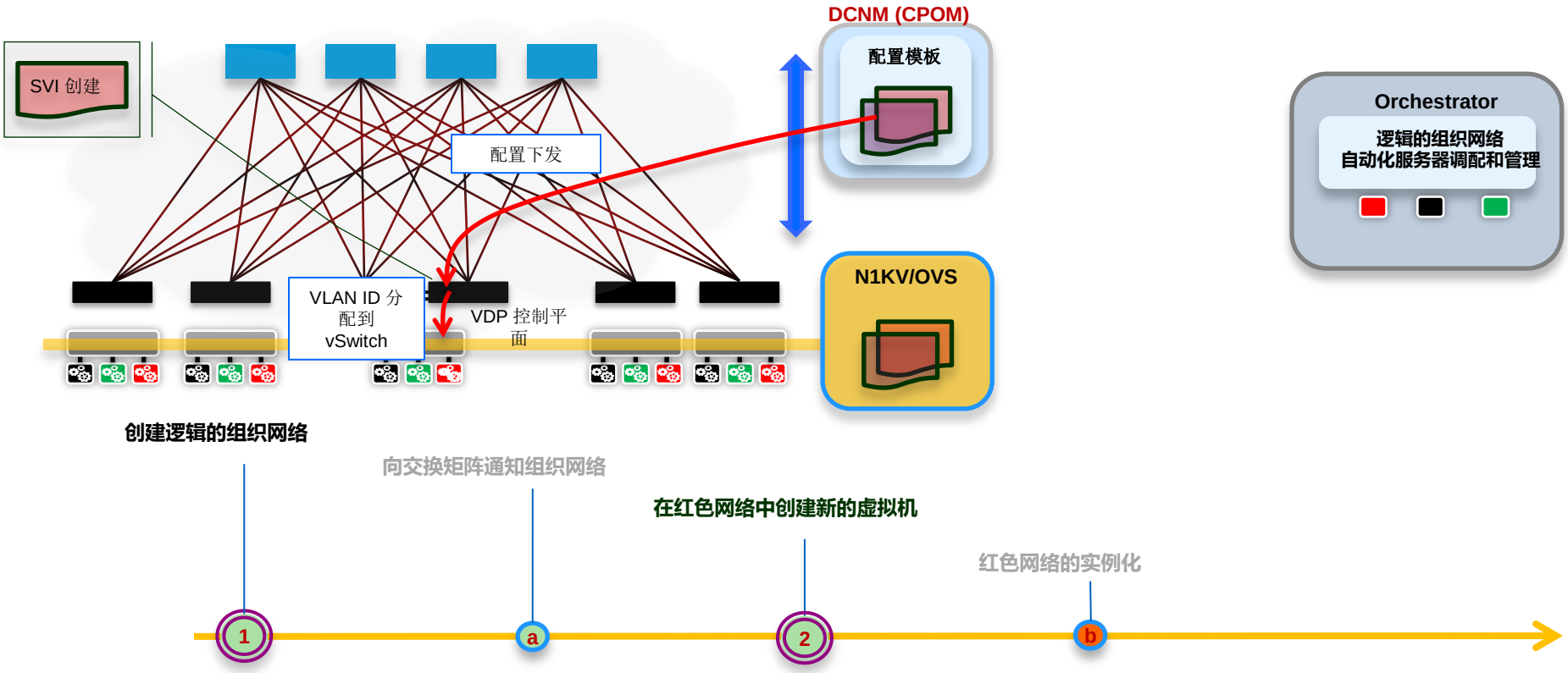




*VDP (VSI 发现和配置协议，是 802.1Qbg Draft 的一部分）

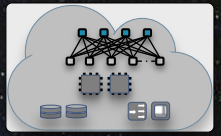
工作负载的自动化
—— 借助VDP实现叶子节点自动配置（3）

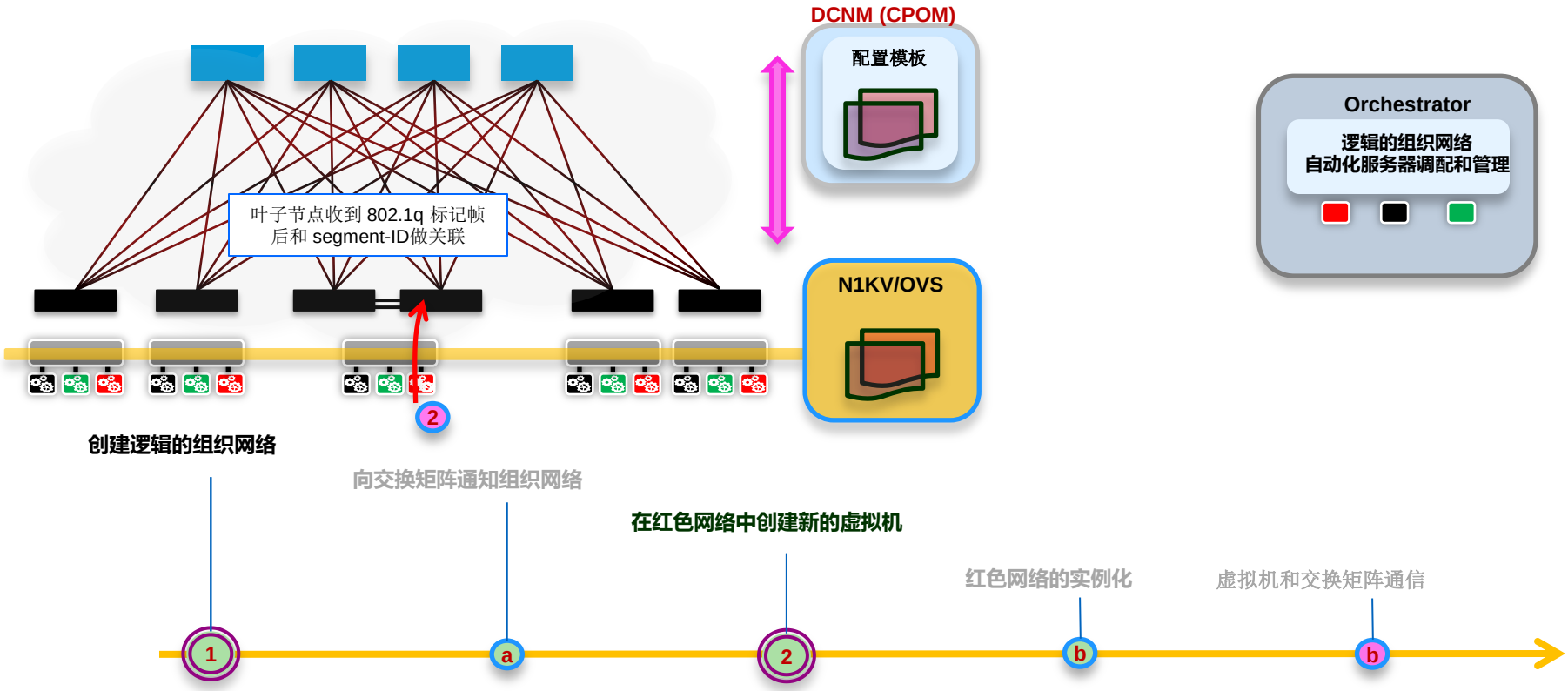




*VDP (VSI 发现和配置协议，是 802.1Qbg Draft 的一部分）

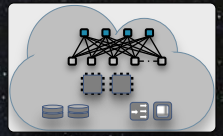
工作负载的自动化
—— 借助VDP实现叶子节点自动配置（4）

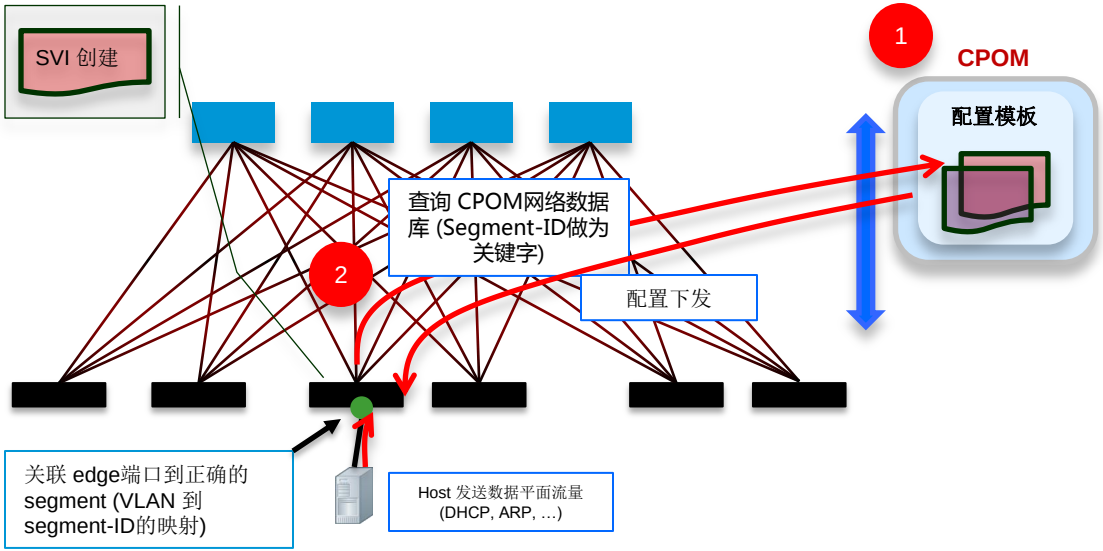




© 2013 Cisco and/or its affiliates. All rights reserved.

工作负载的自动化
—— 如何实现物理服务器的自动配置





- 当物理主机连接到叶子节点所需要自动配置的两点需求
  - 添加配置模板到CPOM 网络数据库
    - 手工配置
    - 利用自动化服务器调配和管理工具配置物理服务器
    - 使用相同的已经创建的针对虚拟工作负载的配置模板
  - 检测主机连接到查询数据库，在叶子节点实例化配置

注意: 同样的模型能被应用到无法被自动化服务器调配和管理工具所管理的VM

© 2013 Cisco and/or its affiliates. All rights reserved.

议程

- DFA 需求和功能
- 交换矩阵管理
- 工作负载自动化
- 最优化的网络
  - 设备角色和拓扑
  - 控制平面
  - 转发平面
- 虚拟交换矩阵
- DCI, 服务集成, 迁移和互操作
- 路标和产品

© 2013 Cisco and/or its affiliates. All rights reserved.

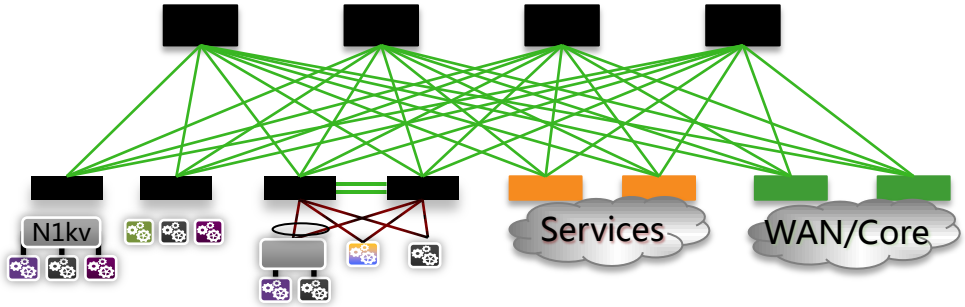
最优化的网络架构

最优化的网络架构

- 支持任何工作负载规模，任何地点的快速部署
- 减少故障域
- 是目前处理L3和L2协同工作的最佳方式
- 简化添加和删除网络设备的处理
- 可扩展的控制平面
- 支持叠加网络
- 利用硬件处理能力
- 优化ECMP转发
- 动态扩展的资源利用率

© 2013 Cisco and/or its affiliates. All rights reserved.

DFA 动态交换矩阵自动化架构 —— 设备角色和命名



虚拟机
物理机
FEXs
第三方交换机
UCS FI
刀片交换机
存储设备

防火墙
负载均衡设备
第三方设备

路由器
L3 交换机
第三方设备

- 针节点
- 叶子节点
- 边界叶子节点
- 服务叶子节点
- 虚拟叶子节点* N1kv

注意：不同的叶子节点的角色是逻辑定义的，不是物理定义的，例如同一个叶子交换机可以完成所有三大类角色功能 (标准、服务、边界)

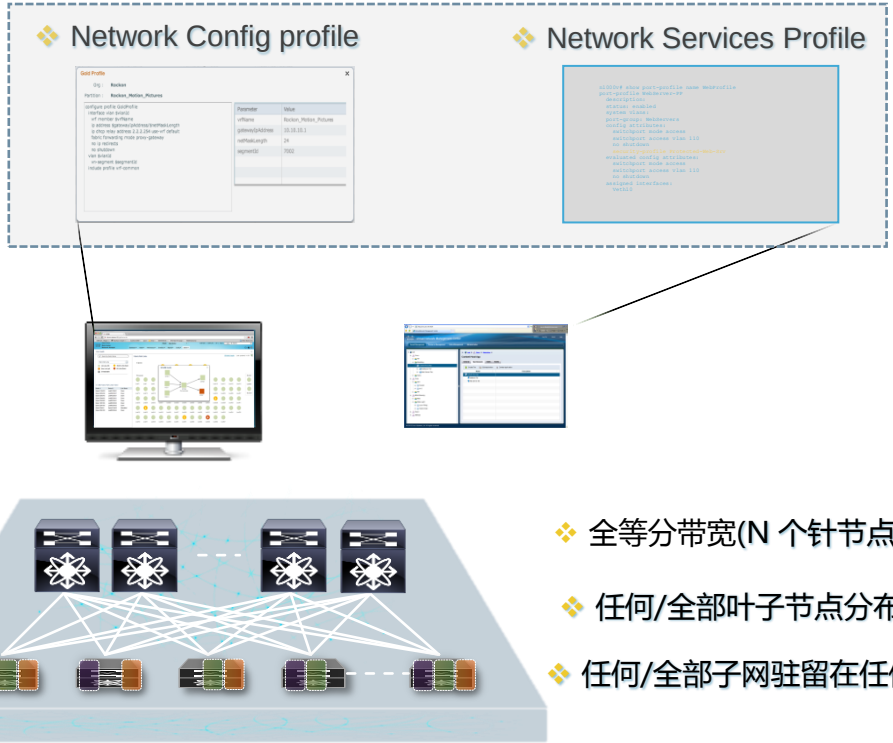
*虚拟叶子节点: N1kv 可以看成在控制平面协议方面一个 “轻型” 参与者– 虚拟叶子节点是支持VDP的通用DVS

© 2013 Cisco and/or its affiliates. All rights reserved.

动态交换矩阵自动化
—— 高度扩展性、弹性和高效率

优势

- 任何网络结构、任何位置、快速部署
- 减小故障域
- 高度的扩展性和弹性
- 模板控制方式的自动配置

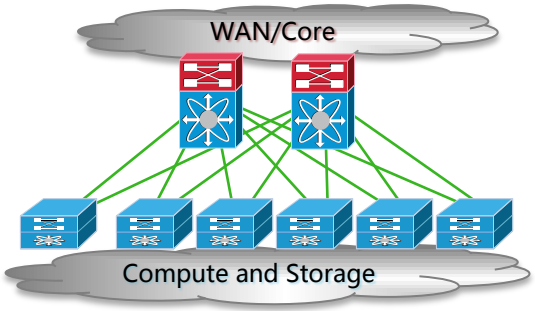


- 全等分带宽(N 个针节点)
- 任何/全部叶子节点分布缺省网关
- 任何/全部子网驻留在任何叶子节点

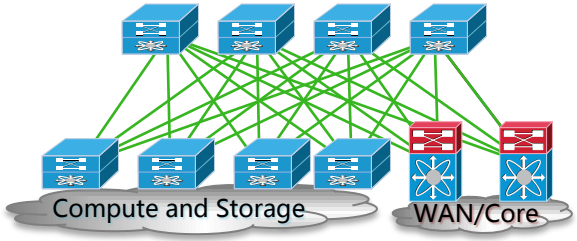
© 2013 Cisco and/or its affiliates. All rights reserved.

动态交换矩阵自动化
—— 支持灵活的拓扑结构

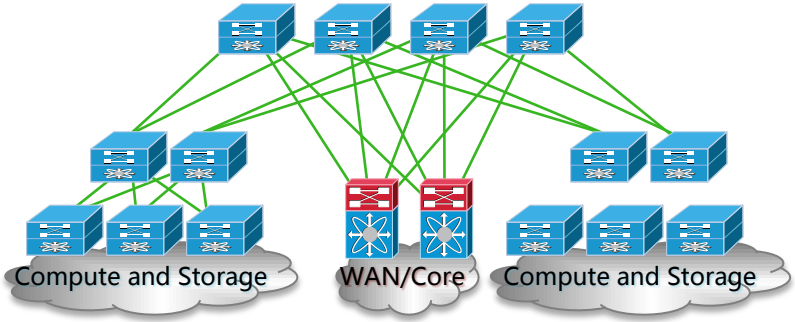
传统的接入和汇聚结构



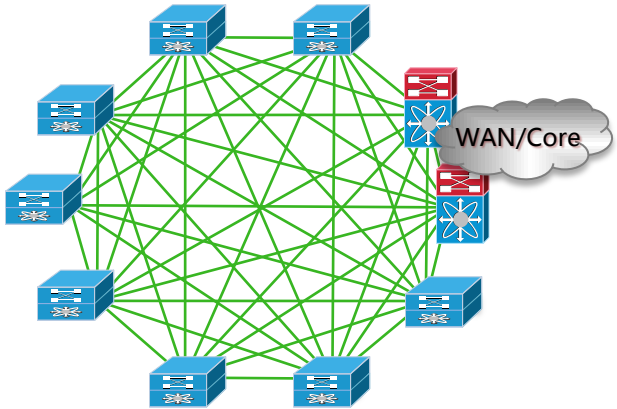
折叠式 CLOS结构



三层结构(胖树)



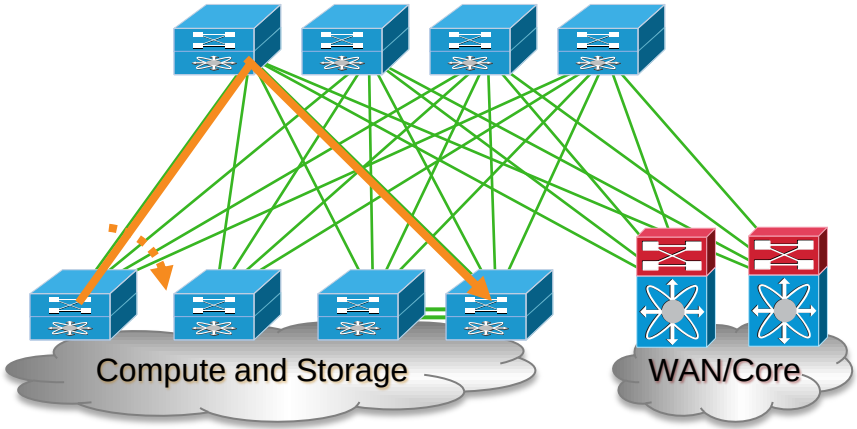
完全对等连接结构



© 2013 Cisco and/or its affiliates. All rights reserved.

动态交换矩阵自动化 —— DFA架构特性

- 更高的全等分带宽
- 更多数量的等价多路径: 单播或者组播
- 一致的网络可达性, 可预计的固定时延
- 高度冗余: 节点/线路 失效
- 针对所有流量的线速、低时延转发

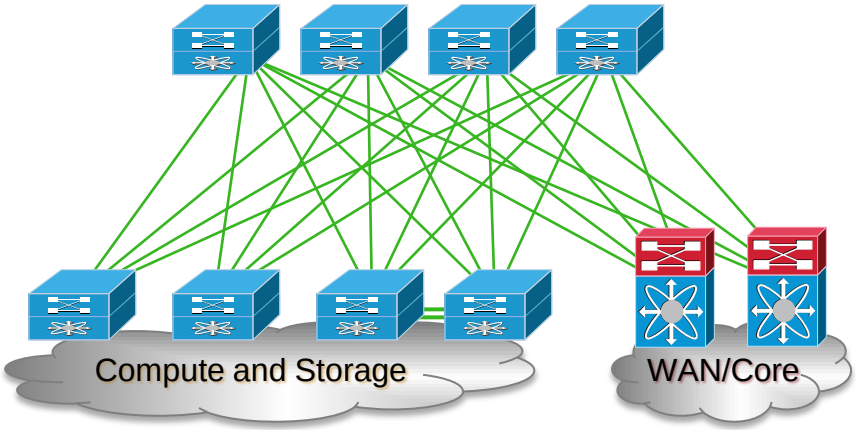


§§ 动态交换矩阵架构的特性可以应用到所有类型的网络拓扑中 §§

© 2013 Cisco and/or its affiliates. All rights reserved.

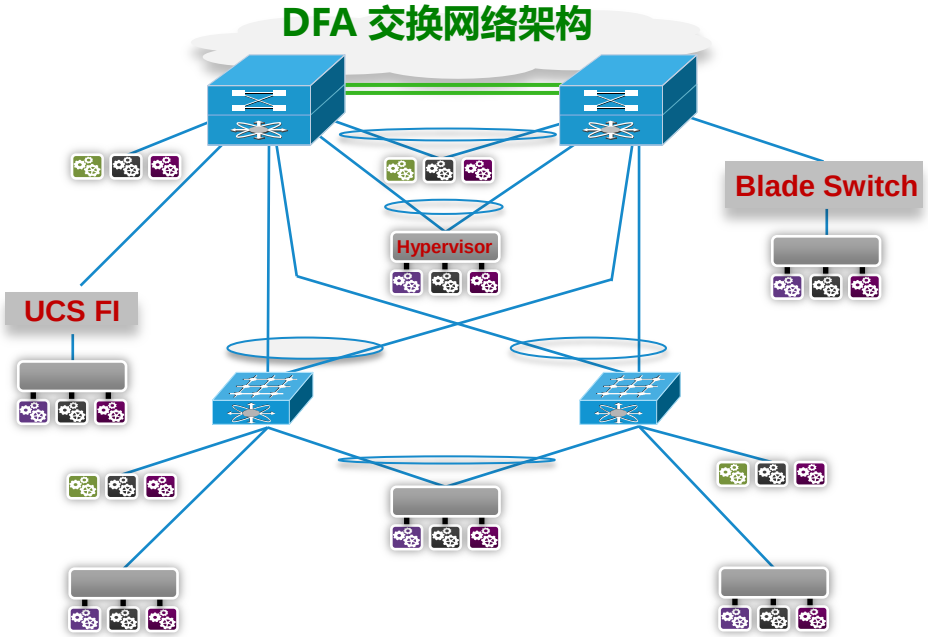
动态交换矩阵自动化
—— 支持各种规模的交换矩阵架构

- 交换矩阵架构规模: 几百到数万 10G 端口
- 可变的组件:
  - ✓ 可变的规模
  - ✓ 可变的容量
  - ✓ 预期的收敛比
  - ✓ 模块化或固定配置
- 横向扩展架构
  - ✓ 按照增长需要平滑增加计算、服务、外部连接节点。



© 2013 Cisco and/or its affiliates. All rights reserved.

动态交换矩阵自动化
——支持各种南向拓扑连接



- 叶子节点有多种灵活的接入选项
  - ✓ 工作在直通或者双活上联模式(eVPC)的FEX
  - ✓ UCS Fabric Interconnects
  - ✓ vPC模式连接的裸机或者Hypervisor
- FEX 仅仅做为 “远程线卡” 并不参与到 DFA 的控制平面和数据平面封装

© 2013 Cisco and/or its affiliates. All rights reserved.

议程

- DFA 需求和功能
- 交换矩阵管理
- 工作负载自动化
- 最优化的网络
  - 设备角色和拓扑
  - 控制平面
  - 转发平面
- 虚拟交换矩阵
- DCI, 服务集成, 迁移和互操作
- 路标和产品

© 2013 Cisco and/or its affiliates. All rights reserved.

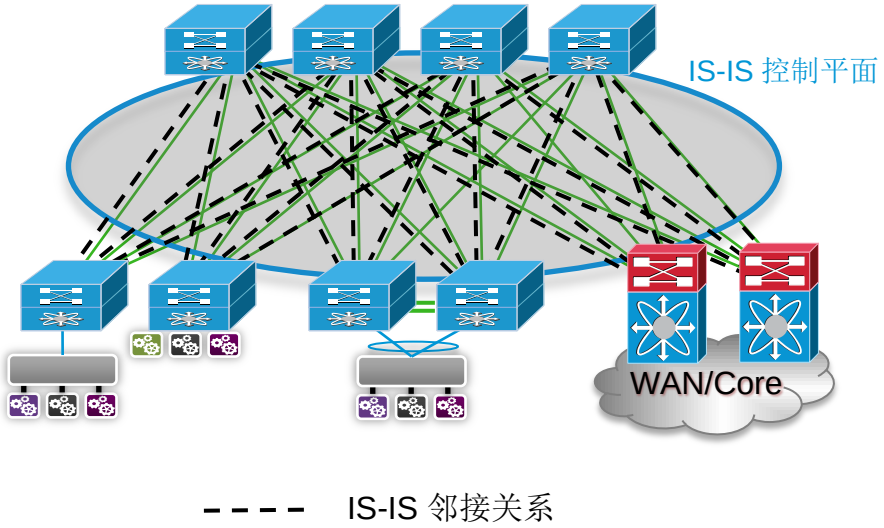
控制平面
1 - IS-IS 做为动态交换矩阵架构控制平面

ISIS 用做动态交换矩阵架构链路状态分布

- 针对叠加封装的交换网络节点的可达性
- 为组播和广播流量构建多目的树
- 对于节点和链路失效可以快速反应
- 网状拓扑的增强

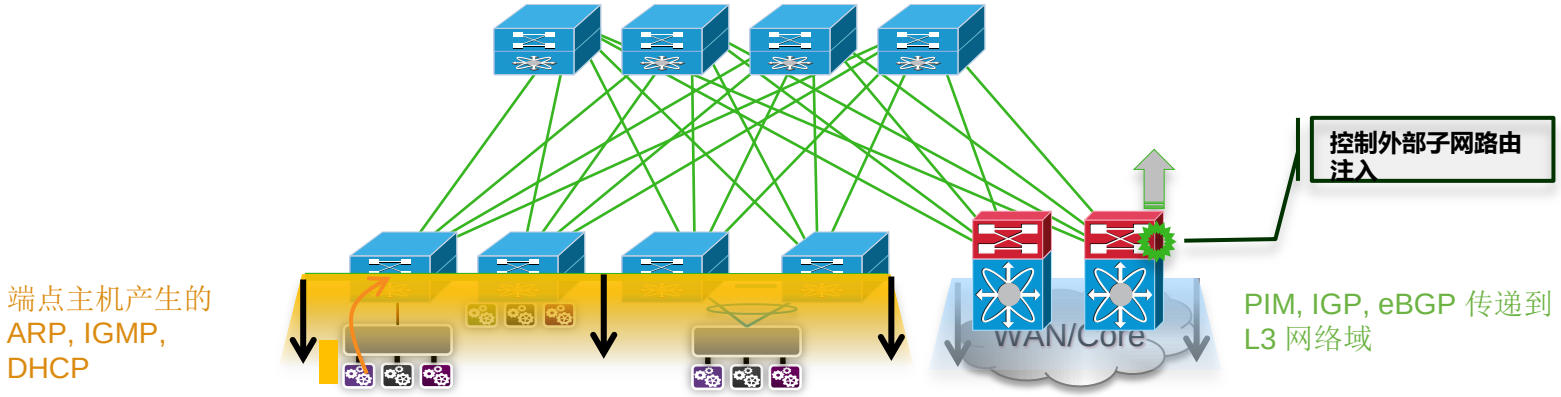
动态交换矩阵架构的控制协议并不分布

- 主机路由
- 主机产生的控制流量
- 服务器子网信息



© 2013 Cisco and/or its affiliates. All rights reserved.

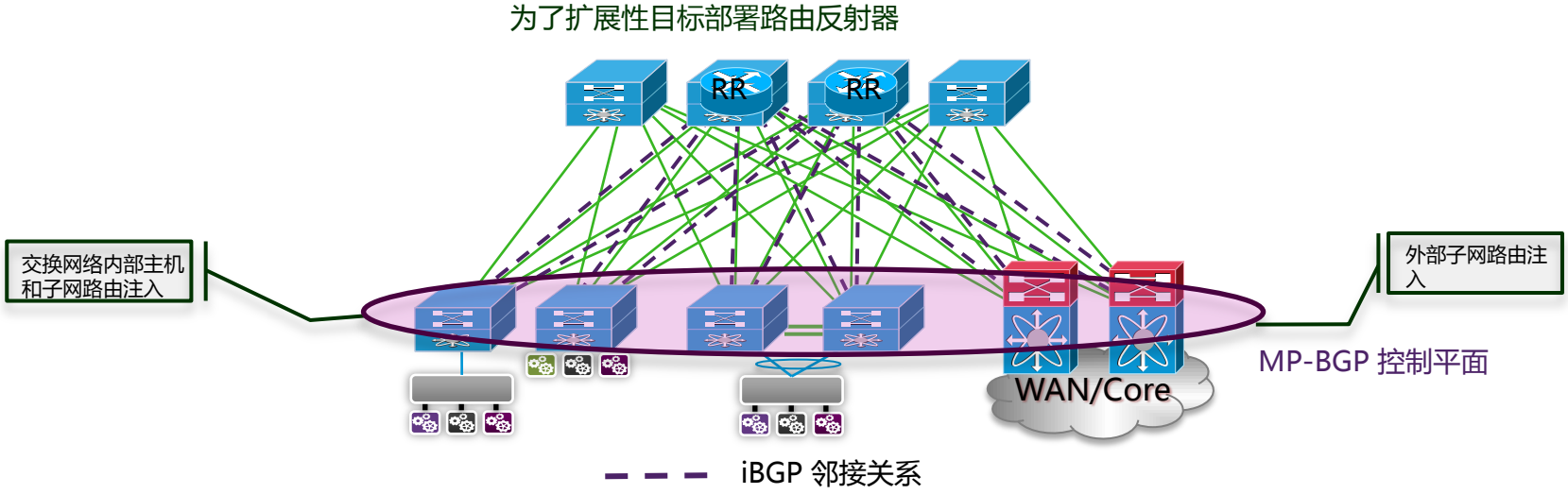
控制平面
2 – 主机产生的控制协议



- 服务器产生的ARP, IGMP, DHCP消息中止在叶子节点。
- 控制泛洪和故障域，分布控制包处理。
- 来自外部网络的 PIM, OSPF, eBGP 等路由消息中止在边界叶子节点

© 2013 Cisco and/or its affiliates. All rights reserved.

控制平面
3 – 主机和子网路由分发



层次化路由设计：

- 主机路由分发工作从动态交换矩阵链路状态协议中分离
- 叶子节点采用MP-BGP分布内部主机/子网路由和外部可达性信息。
- MP-BGP 增强处理 以提供承载数十万路由同时减少收敛时间。

© 2013 Cisco and/or its affiliates. All rights reserved.

控制平面
—— 主机检测和删除

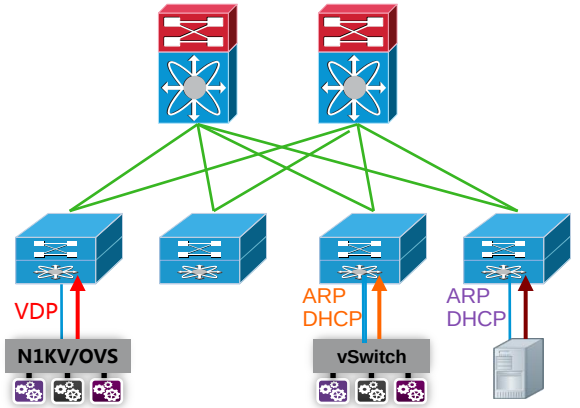
- 为了宣告主机可达性信息，叶子节点必须首先发现本地连接的设备
- 本地主机的检测

  基于 VDP或者ARP/DHCP
- 远程主机的检测

  收到MP-BGP通知消息，安装路由到Unicast RIB表

  （当配置L3 会话式学习时，只是选择性安装路由)
- 本地和远程主机信息删除

  依靠 inactivity 计时器或者显式的 VDP 消息通知



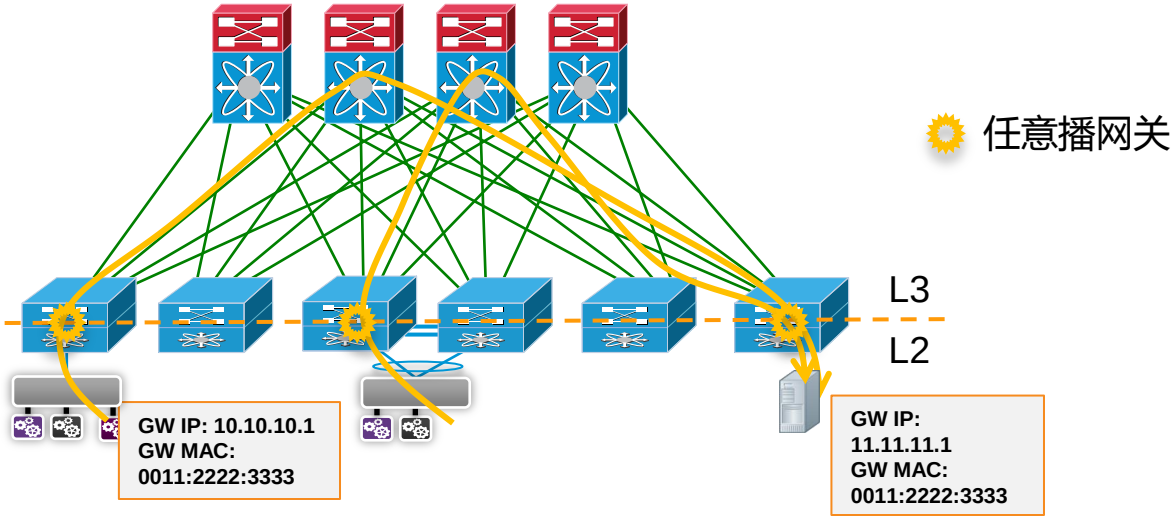
© 2013 Cisco and/or its affiliates. All rights reserved.

议程

- DFA 需求和功能
- 交换矩阵管理
- 工作负载自动化
- 最优化的网络
  - 设备角色和拓扑
  - 控制平面
  - 转发平面
- 虚拟交换矩阵
- DCI, 服务集成, 迁移和互操作
- 路标和产品

© 2013 Cisco and/or its affiliates. All rights reserved.

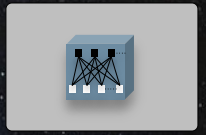
动态交换矩阵自动化架构
—— 叶子节点的分布式网关设计



- 任何子网任何位置 => 任何叶子节点能够提供任何子网的服务
  所有叶子节点上共享所有子网的网关IP和MAC(非HSRP操作)
  ARP报文中止在叶子节点, 在叶子节点外没有任何泛洪。
- 帮助实现虚拟机迁移, 工作负荷分布, 组建任意的工作集群系统
- 在物理主机和虚拟机之间提供无封的L2 或者L3 通信

© 2013 Cisco and/or its affiliates. All rights reserved.

数据平面
——转发模式



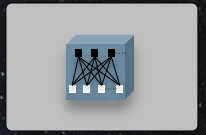
- 针对多个叶子节点下面的每个VLAN，存在两种转发模式：
  - Proxy-Gateway 代理网关模式 （缺省配置模式）

    利用 proxy-ARP 确保两个主机之间的通信永远是被路由的 (不论是同一子网内部还是不同子网之间的通信) → 叶子节点总是执行L3 查询操作
  - Anycast-Gateway 任意播网关模式

    确保分布式缺省网关部署在每个叶子节点但是针对同一子网内通信转发采用传统的FabricPath行为(例如： ARP是被泛洪到整个交换矩阵, 叶子节点执行L2查询, 数据平面针对终端的MAC地址基于会话式学习)
- 转发模式基于每个子网SVI配置！
- 两种模式下，主机路由都是在叶子节点之间宣告。
- 注意: L2非IP报文永远通过交换矩阵进行桥接, 而不论是配置哪种模式。

© 2013 Cisco and/or its affiliates. All rights reserved.

数据平面
—— Proxy-Gateway 代理网关模式



- 在 SVI 控制粒度层级配置

```
interface vlan 10
  vrf member Coke
  fabric forwarding mode proxy-gateway
  ip address 10.1.1.1/24
  ip dhcp relay address 200.200.200.100
  no shutdown
```

- 叶子交换机总是代理连接服务器的接口的入站ARP请求报文

  Proxy-ARP功能在SVI接口是隐含激活的。

  通过交换矩阵的ARP泛洪是被禁止的。

- 仅仅当路由表中存在到目的地址的有效路由时，产生Proxy-ARP 应答

  当目的地址不知道时， 针对这个地址的ARP 请求被丢弃。

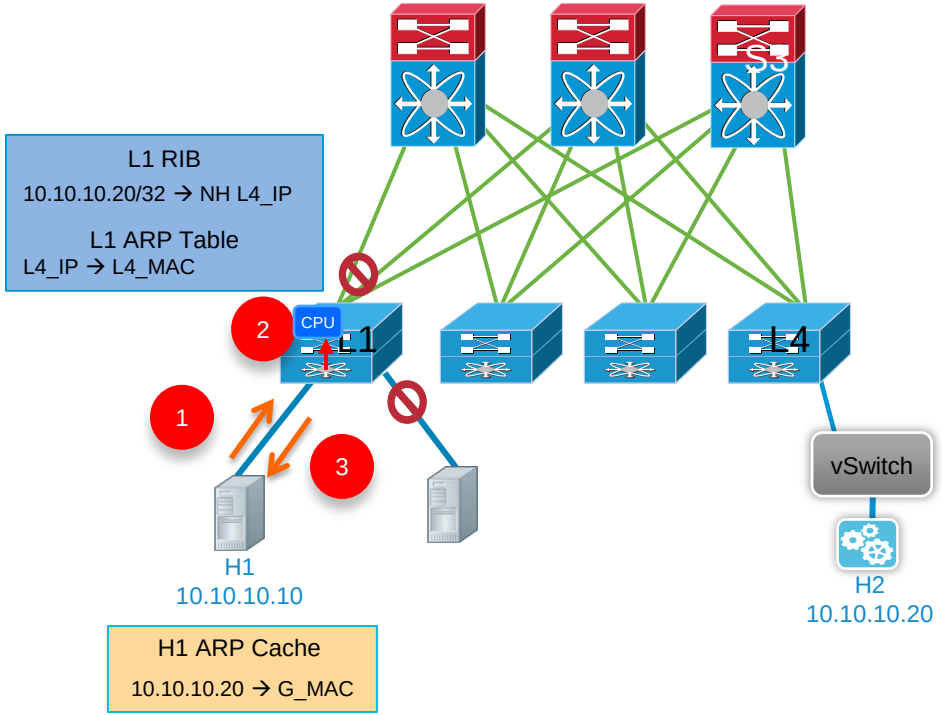
  缺省路由是不会被考虑做为触发 proxy-ARP 应答的。

© 2013 Cisco and/or its affiliates. All rights reserved.

Proxy-Gateway 代理网关模式 —— 同一子网内的IP转发操作

1. H1 向H2（10.10.10.20）发送一个 ARP 请求。
2. 该ARP 请求被拦截在 L1 叶子节点，并发送到控制引擎。
3. 假设在单播RIB表中有一条到达H2的有效的路由, L1发送带有 G_MAC 的ARP应答以至于H1 能够构建自己的 ARP 缓存表

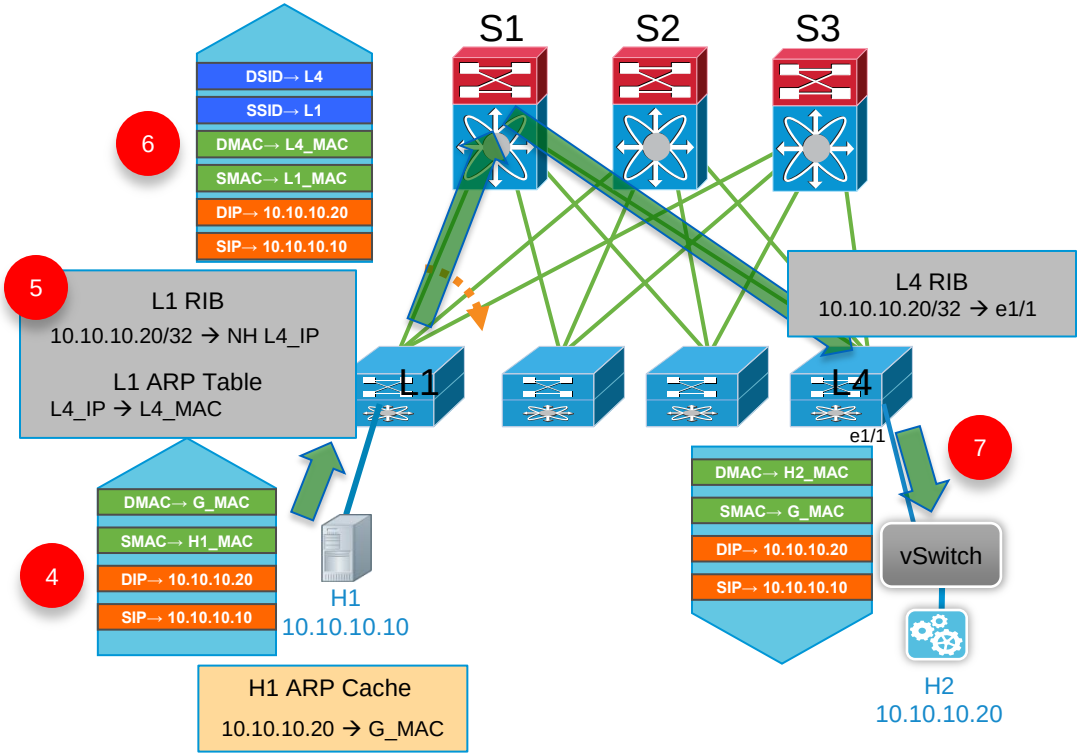
注意: ARP请求没有在整个动态交换矩阵架构中泛洪 也没有在同一个L2 域中其它逻辑接口中泛洪。



© 2013 Cisco and/or its affiliates. All rights reserved.

Proxy-Gateway 代理网关模式
—— 同一子网内的IP转发操作（2）

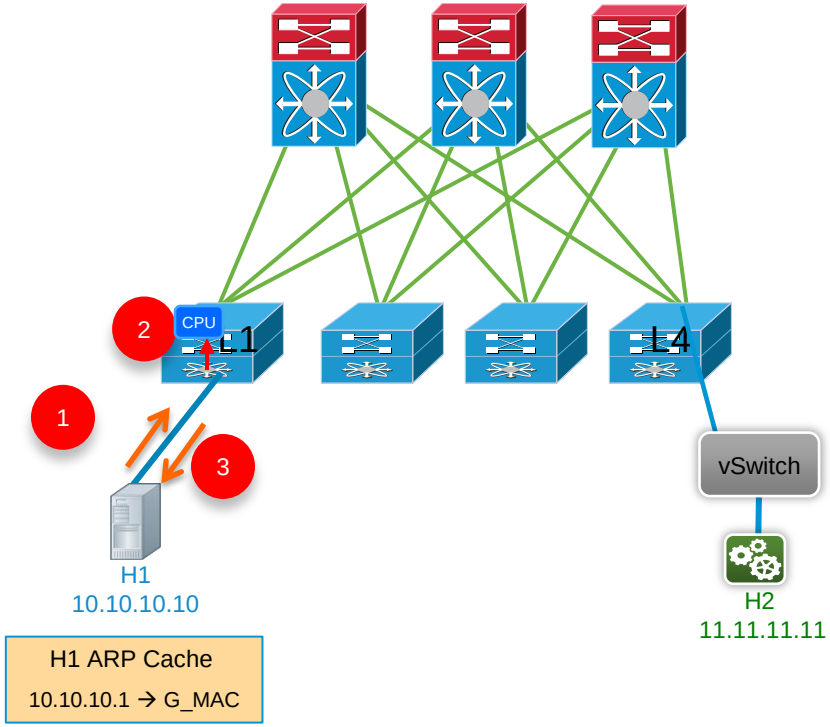
4. H1 产生一个数据报文，该报文以G_MAC地址做为目的MAC
5. L1 收到该报文，删除L2 报文头后执行L3 目的地址查询
6. L1 添加 L2 和 FP 报头，之后在整个交换网络中转发FP帧, 实际转发是通过S1，S2，S3 三个针节点的三条等价路径中的一条。
7. L4接收到报文后，去掉FP 和 L2报头 后执行L3 查询， 然后转发到H2 主机。



© 2013 Cisco and/or its affiliates. All rights reserved.

Proxy-Gateway 代理网关模式 —— 不同子网内的IP转发操作

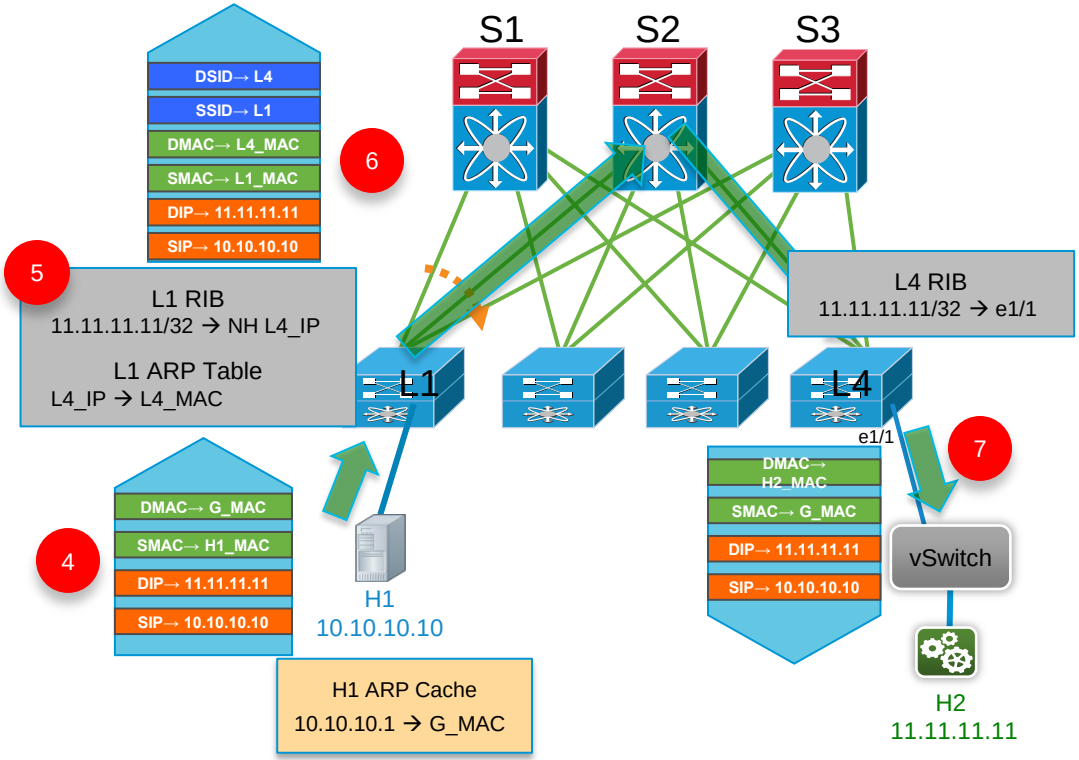
1. H1 向缺省网关（10.10.10.1）发送一个ARP 请求。
2. 该ARP 请求被拦截在 L1 叶子节点，并发送到控制引擎。
3. L1 如同标准的缺省网关的行为一样，发送含有G_MAC地址的ARP响应



© 2013 Cisco and/or its affiliates. All rights reserved.

Proxy-Gateway 代理网关模式
—— 不同子网内的IP转发操作（2）

4. H1 产生一个数据报文，该报文以G_MAC地址做为目的MAC，H2的IP地址做为目的IP
5. L1 收到该报文，删除L2 报文头后执行L3 目的地址查询
6. 如果在单播路由表中存在到达H2的有效路由，L1 添加 L2 和 FP 报头，之后在整个交换网络中转发FP帧, 实际转发是通过S1，S2，S3 三个针节点的三条等价路径中的一条。
7. L4接收到报文后，去掉FP 和 L2报头 后执行L3 查询， 然后转发到H2 主机。



© 2013 Cisco and/or its affiliates. All rights reserved.

动态交换矩阵自动化架构
—— L3 会话式学习(可选操作)

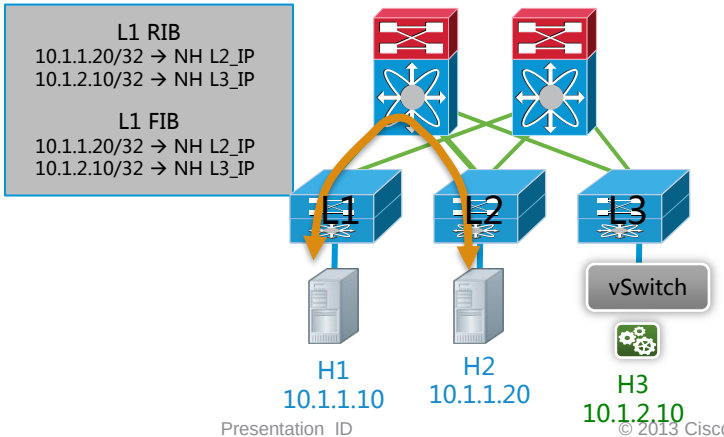
- 使用 /32 主机路由 在大型网络中会导致扩展性问题，例如在所有叶子节点的硬件转发表中安装所有的路由

  L3 会话式学习功能被用来减轻这种情况的发生。

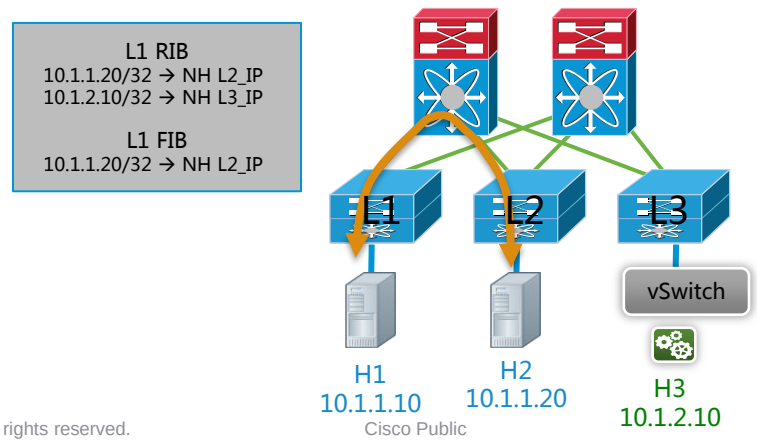
  该功能缺省是关闭的 → 即所有主机路由将被下载到硬件的转发表中。

- 当使用 L3 会话式学习功能后，来自远程节点的主机路由是否从控制层面的RIB表下载到转发层面的硬件的FIB表依赖于该主机是否与本地主机存在活动的会话。

缺省行为 (关闭 L3 会话学习功能)



在激活L3 会话学习功能后



© 2013 Cisco and/or its affiliates. All rights reserved.

L3 会话式学习
—— 活动会话检测

- 活动会话检测需要流量暂时被跳转到CPU

  通过在叶子节点安装“glean-aggregate-prefixes” 以覆盖被给定租户的整个子网的范围所实现

  在租户被第一次检测之后“glean-aggregate-prefixes”被安装在叶子节点的硬件FIB表（带有一个PUNT指针）

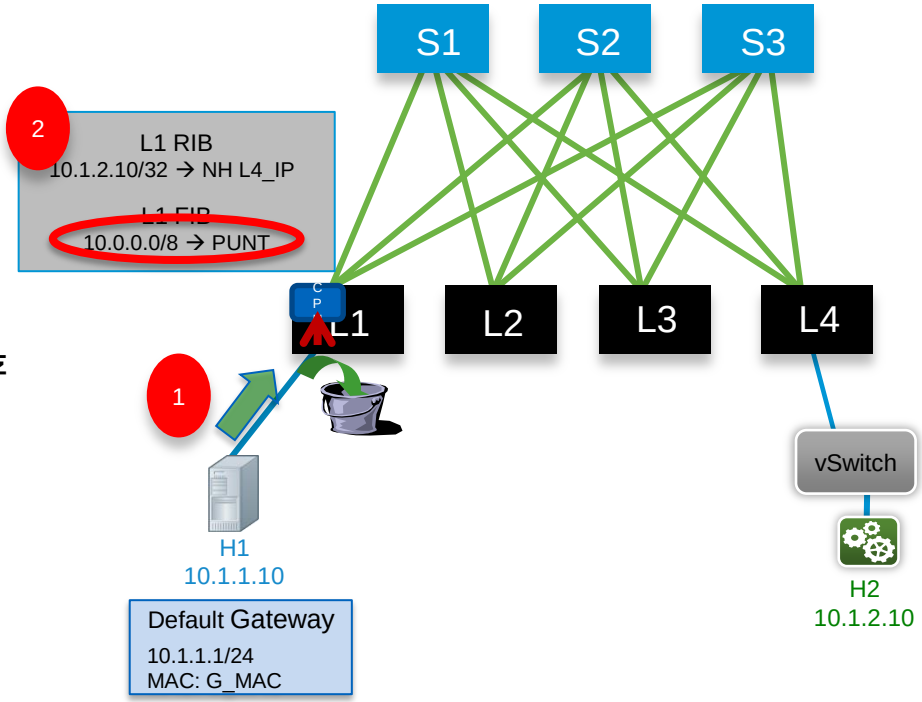
  任何数据报文其目的地址如果匹配到“glean-aggregate-prefix” 则将被跳转到CPU，同时触发硬件FIB表操作（假定该主机路由已经存在于RIB表中）

- 由于报文临时跳转到CPU直到主机路由从RIB表下发到FIB表的操作完成，因此初始流量存在丢包的问题。DFA转发者技术将被用作解决该问题。

© 2013 Cisco and/or its affiliates. All rights reserved.

动态交换矩阵自动化架构
—— L3 会话式学习

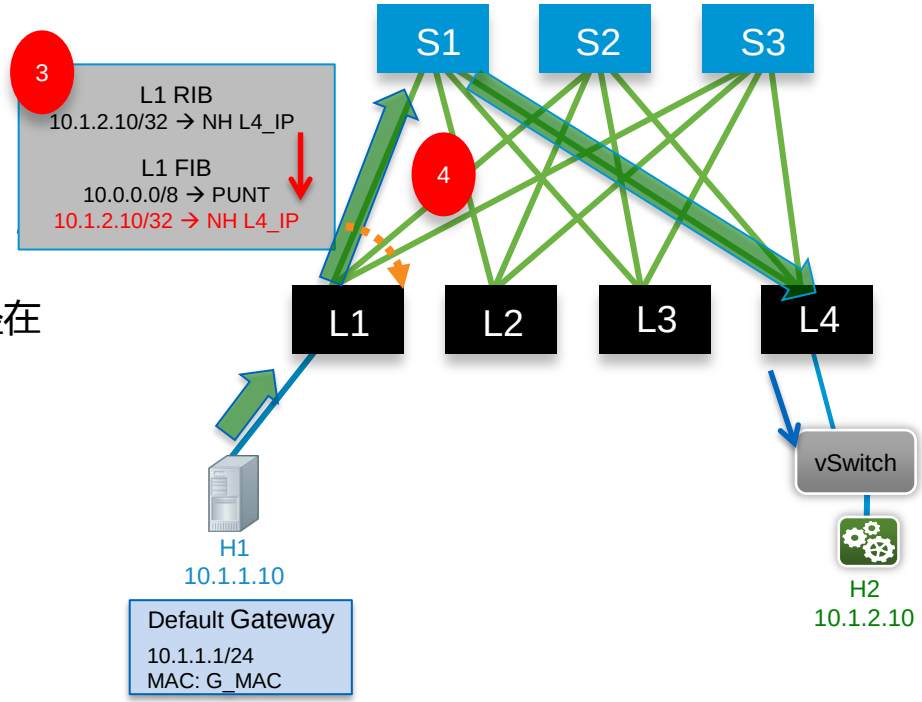
1. H1发送到H2 IP 地址为 10.1.2.10的数据报文
2. L3 会话学习功能在L1上被激活, L1 的FIB 表存在 10.0.0.0/8 “glean-aggregate-prefix”，流量被跳转到CPU (在转发平面被丢弃)



© 2013 Cisco and/or its affiliates. All rights reserved.

动态交换矩阵自动化架构
—— L3 会话式学习

3. 由于在RIB表中存在10.1.2.10/32的主机路由，所以该表项被下载到FIB表
4. 去往H2 的流量开始通过ECMP中的可用路径在交换矩阵中转发



© 2013 Cisco and/or its affiliates. All rights reserved.

动态交换矩阵自动化架构 —— 介绍“DFA转发者”

- 转发者叶子节点功能被设计用来避免当使能L3 会话学习时引起的初始流量丢失问题

  需要用户静态配置（或者通过PoAP模板配置）

- 转发者总是安装所有与租户子网相关的所有主机路由在硬件FIB表中

  设计目标是当指定的叶子节点下载主机路由到硬件FIB表时，用转发者做为跳板去转发流量到指定的目的地址

  因此转发者设备需要具备足够大的FIB容量去保存所有主机路由信息。

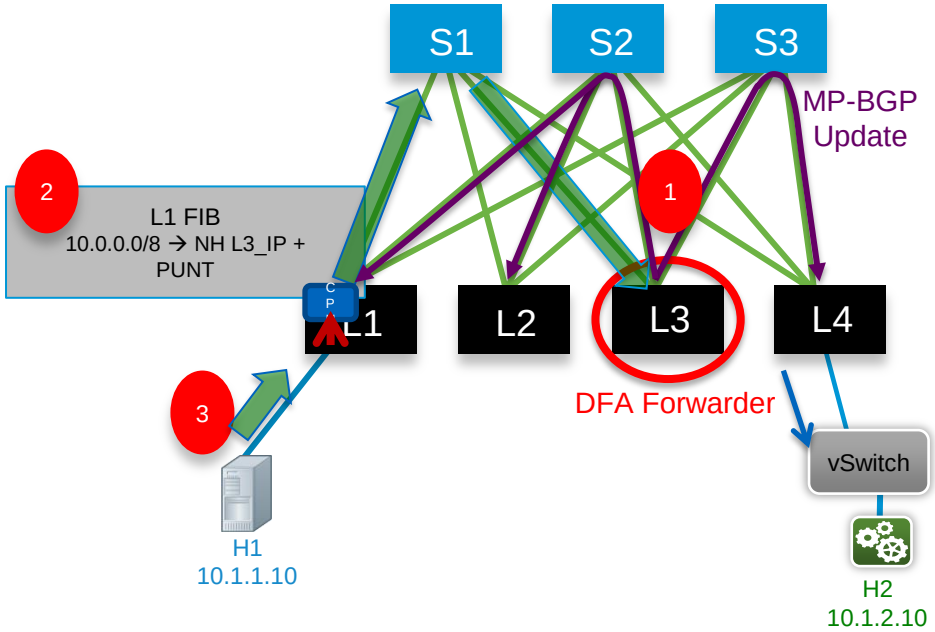
- 转发者宣告聚合路由前缀（汇总了所有租户的子网）给其它叶子节点

  这些聚合路由前缀总是安装在硬件FIB表中，标识为“punt + forward”指针。

© 2013 Cisco and/or its affiliates. All rights reserved.

动态交换矩阵自动化架构
—— 利用“DFA转发者”

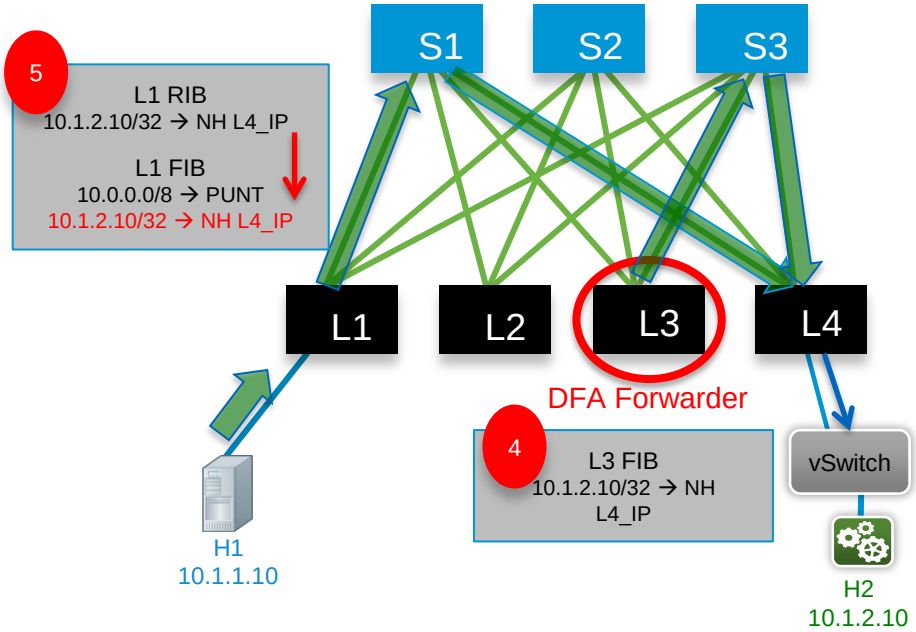
1. L3 被配置为转发者，通过MP-BGP宣告聚合路由前缀10.0.0.0/8 给所有其它的叶子节点
2. 其它叶子节点识别来自L3 的该路由前缀作为“forwarder route” (基于特殊的 BGP community 配置)，因此安装该路由前缀到FIB表，标识为“punt + forward”指针
3. H1发送到H2 IP 地址为 10.1.2.10的数据报文 ，根据FIB表信息，该报文被跳转到CPU ，同时转发给L3 设备



© 2013 Cisco and/or its affiliates. All rights reserved.

动态交换矩阵自动化架构
—— 利用“DFA转发者”

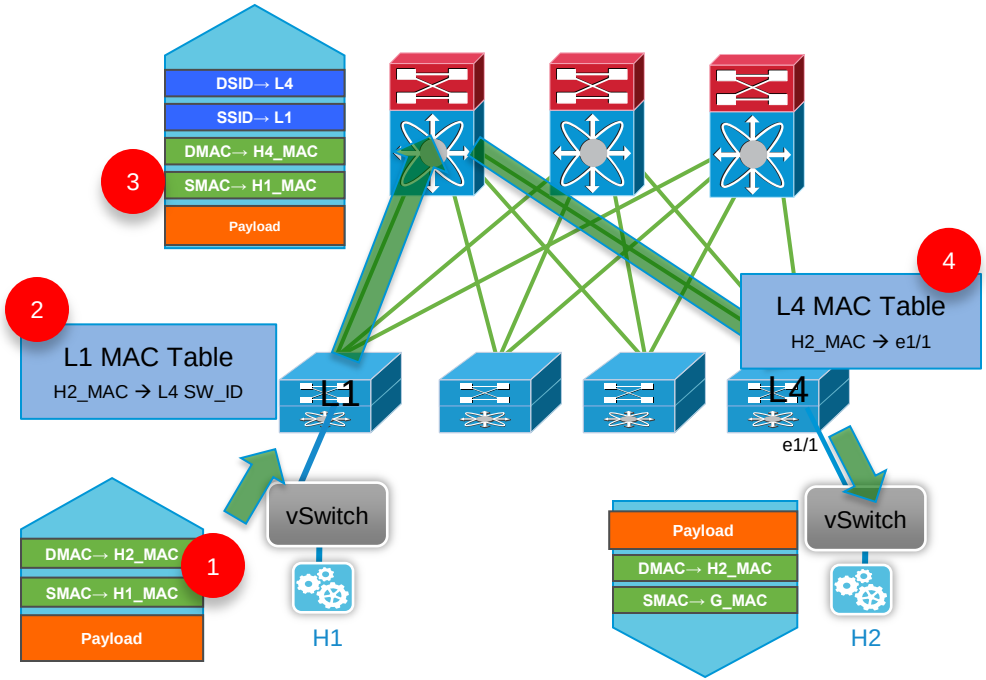
4. L3转发者接收到该报文后，基于自己FIB表中的主机路由信息发送它到L4（否则丢弃）
5. 同时H2 的该主机路由信息被安装到L1 FIB表中，流量开始按照最优路径发送到L4。



© 2013 Cisco and/or its affiliates. All rights reserved.

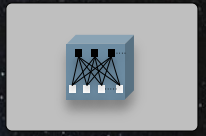
转发二层非IP流量

1. H1 产生一个目的为H2 MAC地址的报文
2. L1 节点针对该帧在所属VLAN的MAC表中执行L2查询
3. L1 在发送到交换网络之前 添加 L2 和FP 报头
4. L4 接收到该帧, 去掉 FP 报头, 执行 L2查找然后发送到H2 主机



© 2013 Cisco and/or its affiliates. All rights reserved.

数据平面
Anycast-Gateway 任意播网关模式



- 在 SVI 控制粒度层级配置

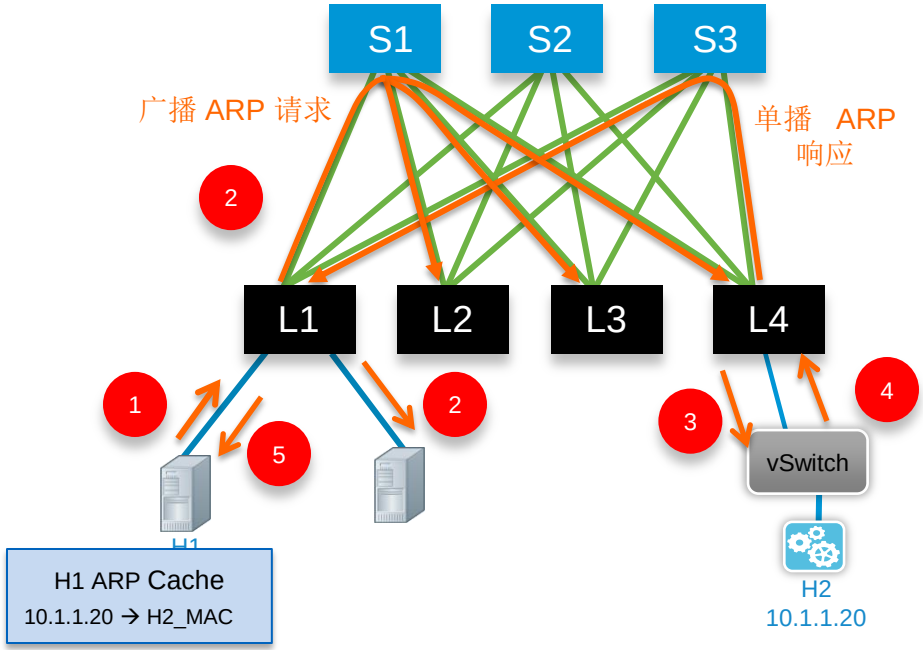
```
interface vlan 10
  vrf member Coke
  fabric forwarding mode anycast-gateway
  ip address 10.1.1.1/24
  ip dhcp relay address 200.200.200.100
  no shutdown
```

- 每个叶子节点针对所有定义的IP子网提供本地网关服务
  - 同一子网内的转发总是被桥接的（传统的FP行为）—— 利用会话式学习MAC地址
  - 不同子网的转发是按照三层转发操作 (和Proxy-Gateway模式的操作行为一致)
- 允许动态发现连接到叶子节点的静默主机

© 2013 Cisco and/or its affiliates. All rights reserved.

Anycast-Gateway 任意播网关模式
—— 同一子网内的IP转发操作

1. H1 发送到 H2 –10.1.1.20主机的ARP请求。
2. 该ARP 请求被泛洪到整个L2 域 ，通过多目的树到达所有远程的叶子节点。
3. L4 转发该ARP请求到属于该L2域的本地接口。
4. H2发送一个单播 ARP应答到H1主机, L4挑选一个可用的路径发送数据帧到L1*. 这也允许L4发现H2 主机(万一L4 还没有发现)
5. L1 转发 ARP 应答给 H1，同时更新它的ARP Cache。

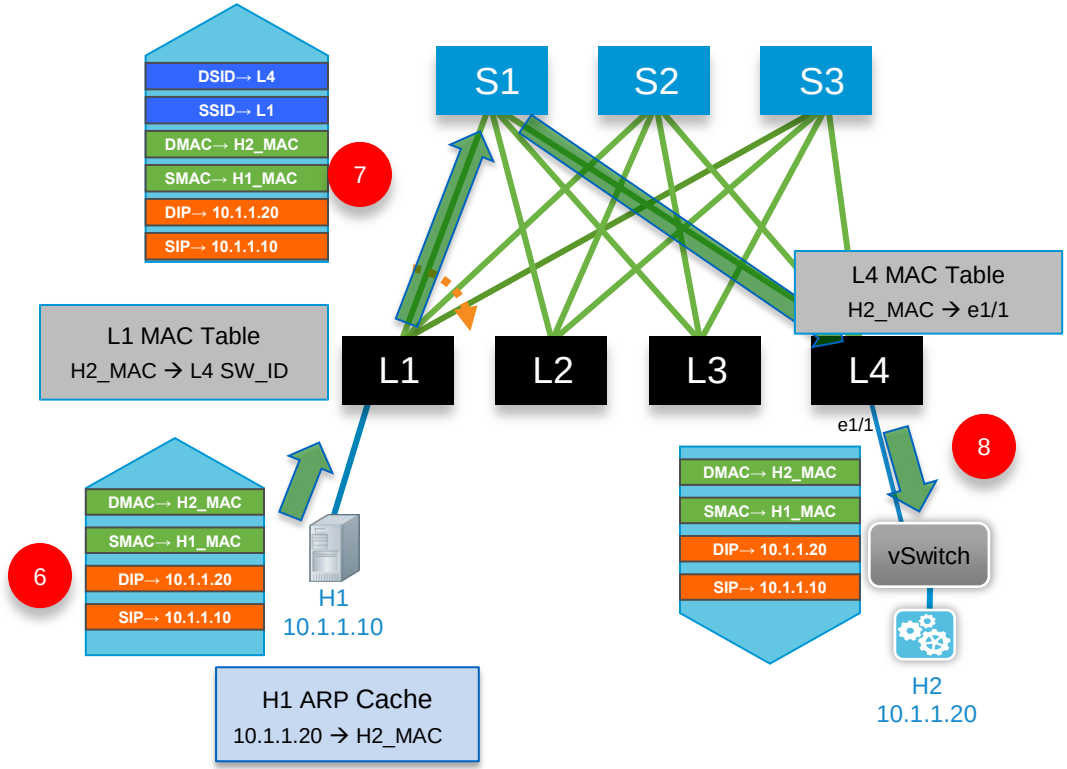


* 如果L4在它的MAC地址表中没有H1的地址，该ARP应答将被作为未知单播地址进行泛洪操作。

© 2013 Cisco and/or its affiliates. All rights reserved.

Anycast-Gateway 任意播网关模式
—— 同一子网内的IP转发操作（2）

6. H1产生一个目的为H2 的数据报文 。
7. L1 收到该报文, 执行 L2 查询 ， 利用FP 规则发送该报文到 L4 。
8. L4 收到该报文，删去 FP 报头, 执行 L2 查询然后转发数据帧到H2 主机。

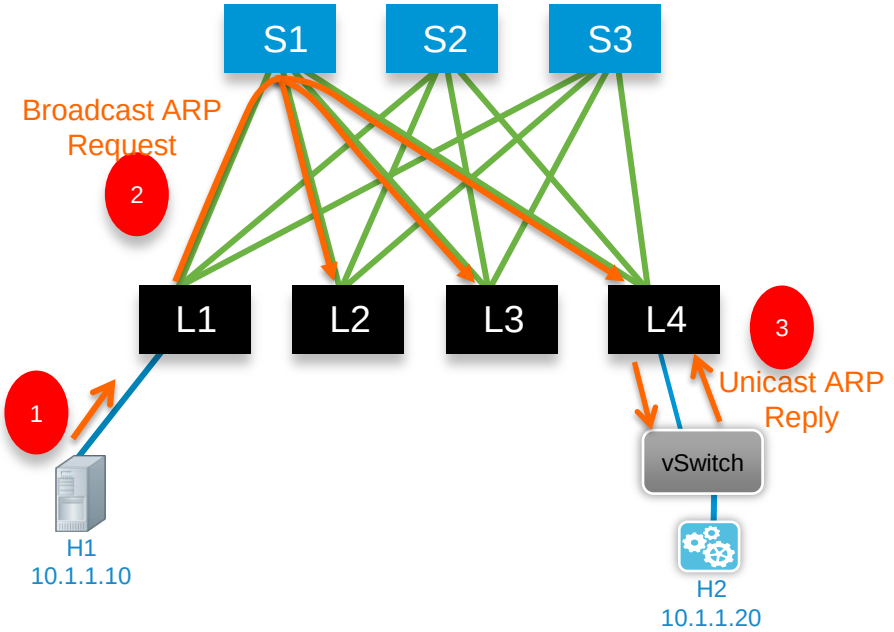


© 2013 Cisco and/or its affiliates. All rights reserved.

Anycast-Gateway 任意播网关模式 —— 发现同一子网内的静默主机

假设: H2是静默主机，因此没有被交换矩阵所发现

1. H1 针对H2 主机的IP地址10.1.1.20 产生一个广播的ARP请求报文
2. 该 ARP请求报文被泛洪到整个 L2 域通过交换矩阵到达 H2 主机
3. 由于ARP请求中包含了H1的MAC地址作为 “sender hardware address” ，所以H2 产生一个单播到H1的ARP应答

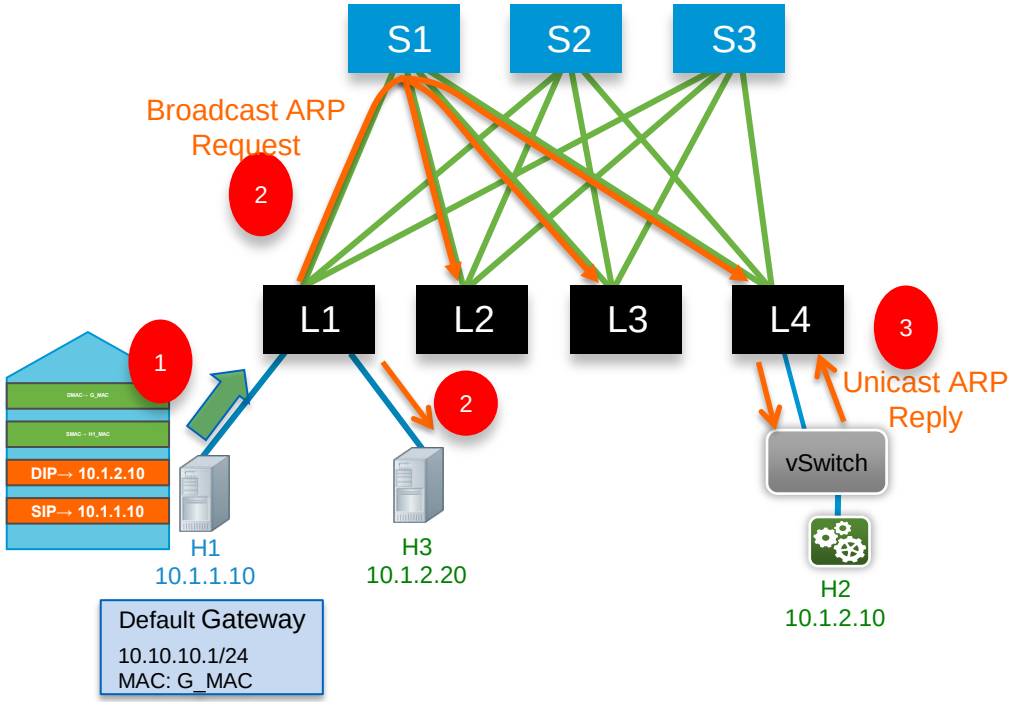


© 2013 Cisco and/or its affiliates. All rights reserved.

Anycast-Gateway 任意播网关模式 —— 发现不同子网内的静默主机

假设: H2是静默主机，因此没有被交换矩阵所发现

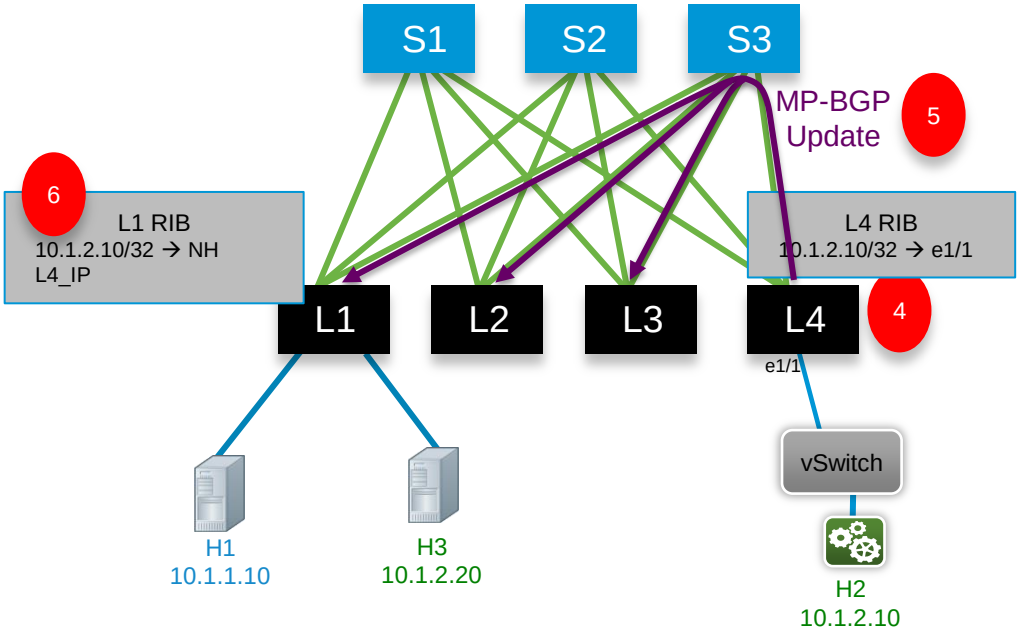
1. L1收到源自H1目的到H2 的IP报文
2. H2的子网在L1上有本地定义（本例中有H3连接到L1 ），因此L1执行针对H2的ARP请求（由于H2未知）。该ARP请求在整个L2 域中泛洪后达到H2
3. 由于ARP请求中包含了G_MAC地址作为“sender hardware address” ，所以H2 产生一个单播到G_MAC的ARP应答



© 2013 Cisco and/or its affiliates. All rights reserved.

Anycast-Gateway 任意播网关模式 —— 发现不同子网内的静默主机

4. 该ARP应答被L4 截获，由于L4 本地定义了G_MAC，因此触发了H2 的发现
5. L4 利用iBGP控制平面宣告H2 主机路由到所有的叶子节点
6. L1 添加L2 主机路由到它的RIB表（同样的操作发生在其它叶子节点）



如果在L1 节点下没有H2的子网的主机连接? 此时需要部署“子网所有者”

© 2013 Cisco and/or its affiliates. All rights reserved.

Anycast-Gateway 任意播网关模式
—— DFA 子网所有者Subnet Owner

- Anycast-Gateway 模式允许发现静默主机
- 需要叶子节点为静默主机产生广播ARP请求

  即使目前没有主机连接，也必须在叶子节点创建针对目的子网的SVI接口
- 通过创建DFA子网所有者以降低上述影响扩展性的担心

  子网所有者是按需动态配置（不需要用户配置）

  每个叶子节点被指派作为本地实例化的所有子网的子网所有者（因为本地被发现的主机属于那些子网）

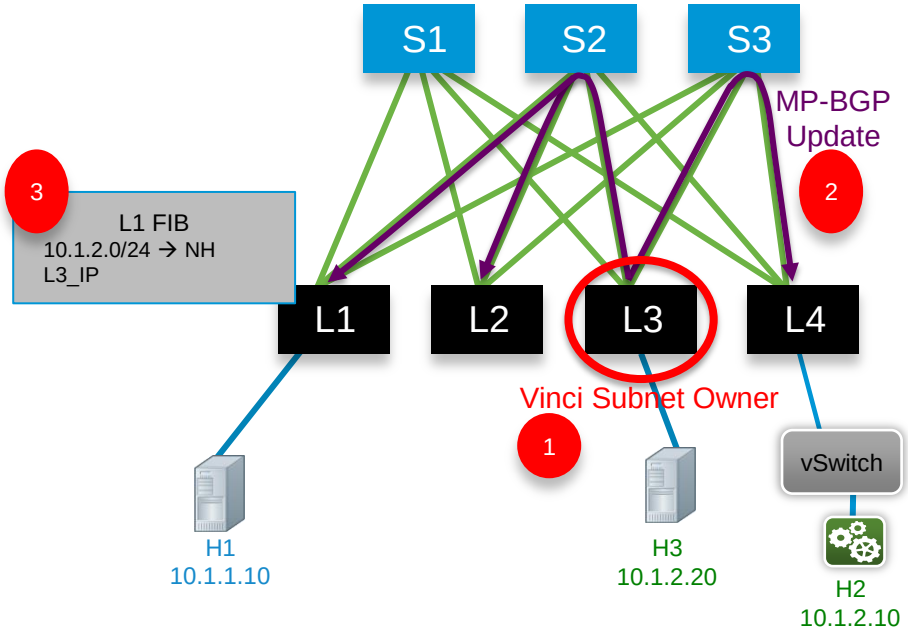
  每个子网所有者通过MP-BGP宣告子网的路由前缀到其它叶子节点，该路由携带特殊的团体属性值以便能被直接安装到硬件FIB表中

© 2013 Cisco and/or its affiliates. All rights reserved.

Anycast-Gateway 任意播网关模式 ——利用 DFA 子网所有者发现主机

假设: H2是静默主机，因此没有被交换矩阵所发现

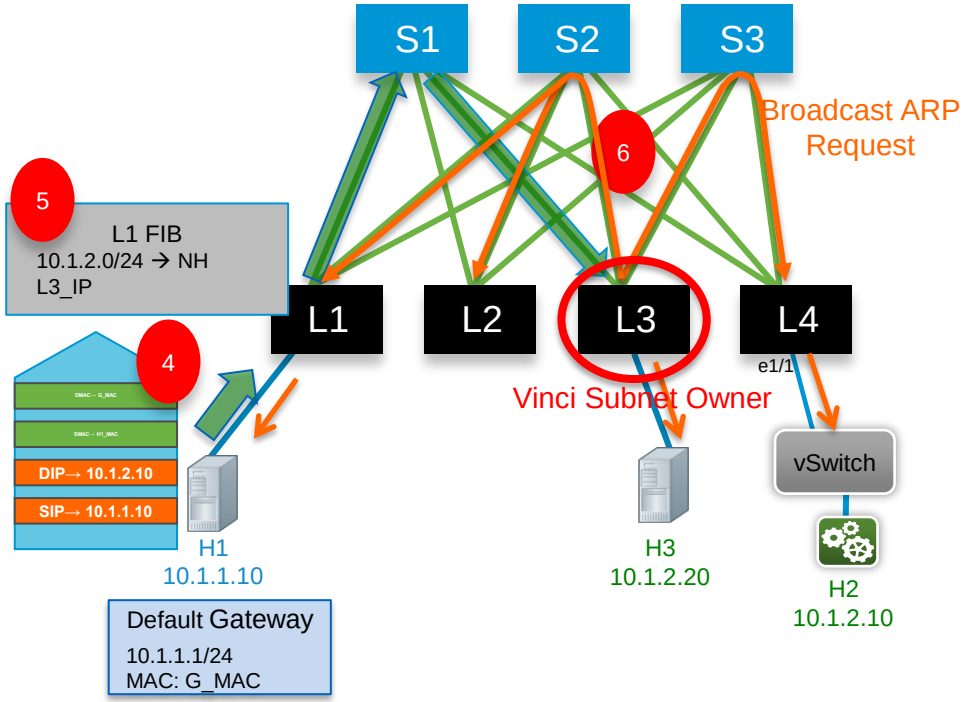
1. L3 被指派为10.1.1.10/24的子网所有者，因为本地连接了H3主机
2. L3 通过 MP-BGP宣告10.1.1.10/24 到其它叶子节点
3. L1 收到该MP-BGP 更新，添加前缀到硬件FIB表（其它叶子节点采用相同的操作）



© 2013 Cisco and/or its affiliates. All rights reserved.

Anycast-Gateway 任意播网关模式 ——利用 DFA 子网所有者发现主机

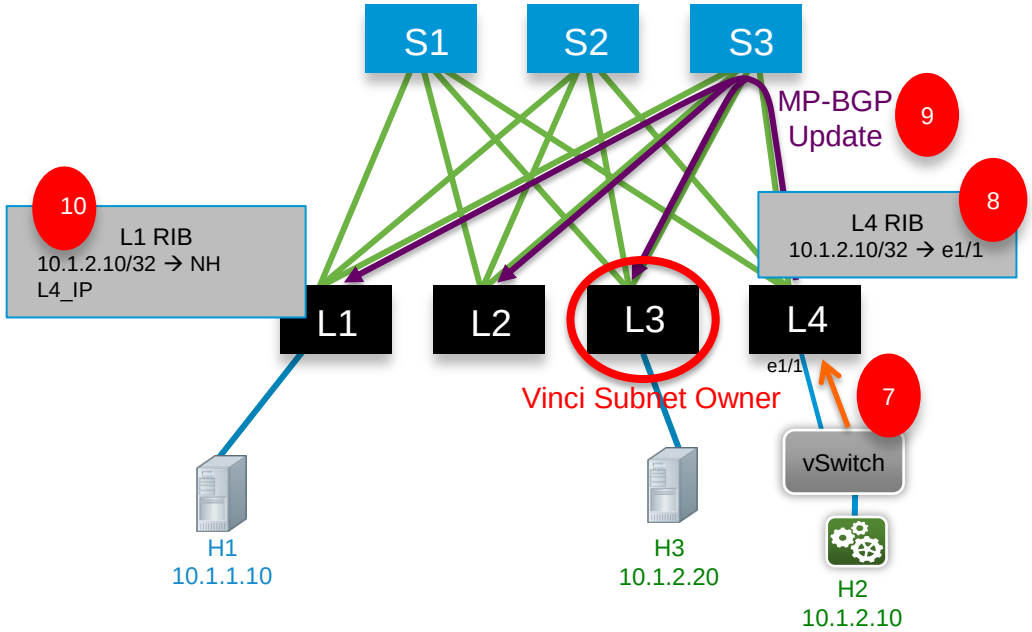
4. L1收到源自H1目的到H2 的IP报文
5. L1执行FIB查询，然后转发该报文到子网所有者
6. L3子网所有者执行针对H2主机的ARP查询，由于H2未知，该ARP请求在整个L2 域中泛洪后达到H2



© 2013 Cisco and/or its affiliates. All rights reserved.

Anycast-Gateway 任意播网关模式 ——利用 DFA 子网所有者发现主机

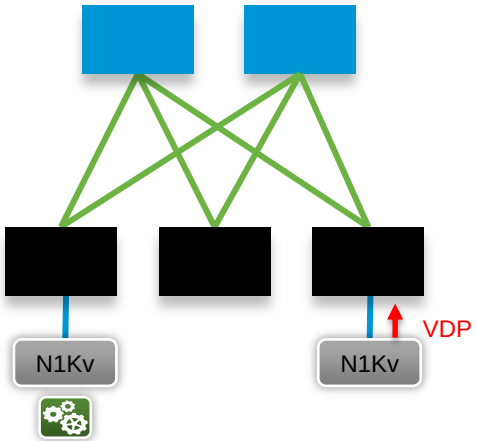
7. 由于ARP请求中包含了G_MAC地址作为“sender hardware address”，所以H2 产生一个单播到G_MAC的ARP应答
8. 该ARP应答被L4 截获，由于L4 本地定义了G_MAC，因此触发了H2 的发现
9. L4 利用iBGP控制平面宣告H2 主机路由到所有的叶子节点
10. L1 添加L2 主机路由到它的RIB表（同样的操作发生在其它叶子节点）

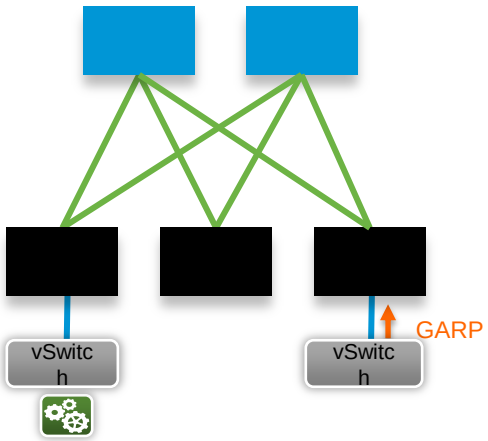


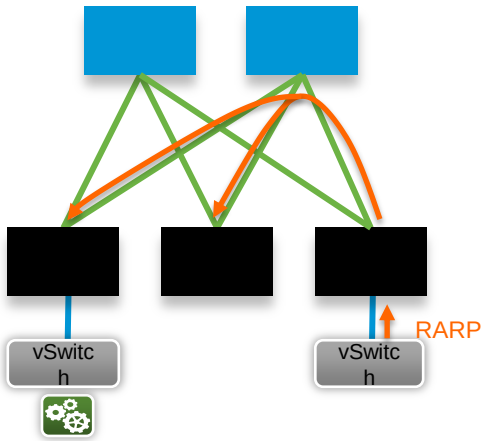
© 2013 Cisco and/or its affiliates. All rights reserved.

主机迁移
—— VM 迁移发现

- 当一个VM迁移到新的叶子节点，该新的VM位置是被如下几种方法之一来检测：
- VDP: 一旦VM迁移，N1Kv发送显式控制平面消息给叶子节点
- GARP: 一旦VM迁移，该hypervisor虚拟层发送免费ARP（GARP）帧〉本地叶子节点能利用包含在报文负荷中的信息发现VM的IP地址
- RARP: 一旦VM迁移，该hypervisor虚拟层发送反向ARP（RARP）帧〉由于在报文负荷中没有IP地址信息，交换矩阵需要处理RARP消息去触发新位置上VM的发现



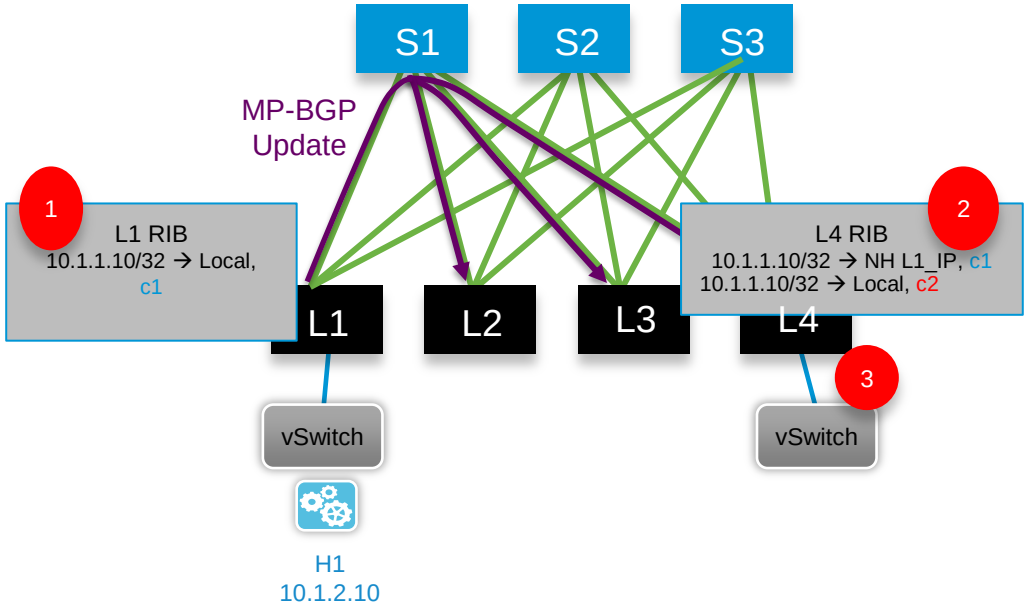




© 2013 Cisco and/or its affiliates. All rights reserved.

通过 VDP/GARP 发现主机迁移 —— 步骤(1)

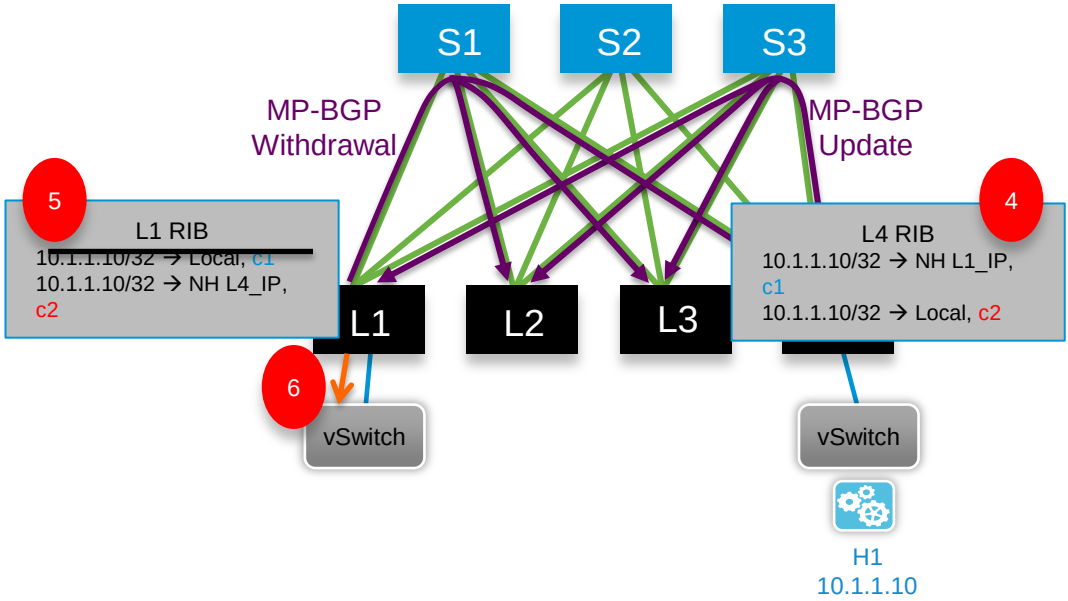
1. L1 发现H1，H1主机路由被L1 通过MP-BGP宣告，附带的Cost 团体属性为c1
2. L4收到该MP-BGP更新后安装该路由到它的RIB表和FIB表（假设L3 会话学习功能关闭）
3. H1主机迁移到L4 所连接的vSwitch，并且被本地检测到（通过VDP/GARP方法））》L4 添加一条主机路由表项，其附带的Cost 团体属性为c2，并且c2<c1



© 2013 Cisco and/or its affiliates. All rights reserved.

通过 VDP/GARP 发现主机迁移 —— 步骤(2)

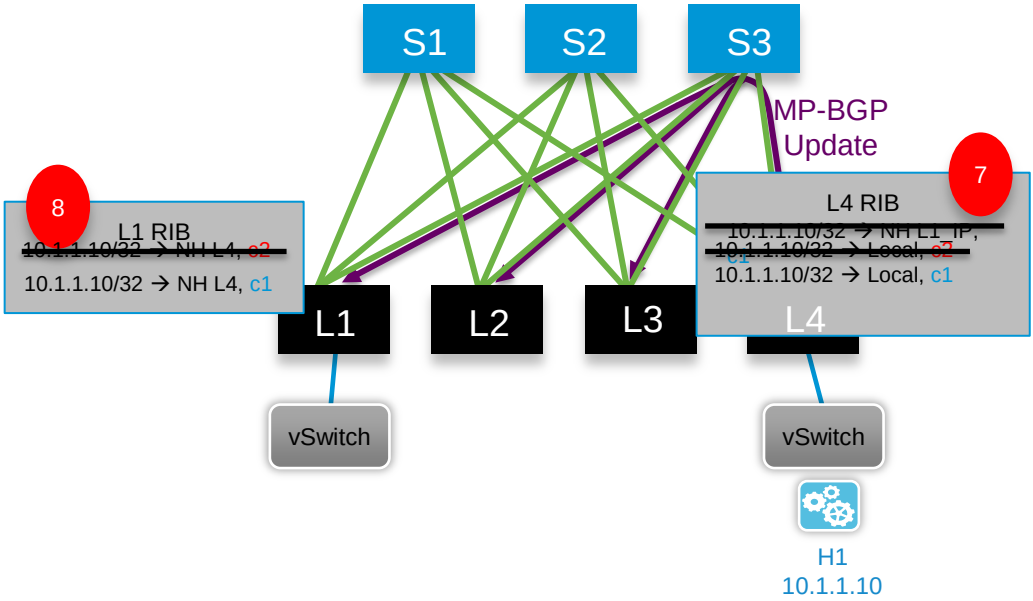
4. 在L4的RIB表中存在关于H1的表项（下一跳指向L1），现在L4 对外宣告新的带有更低cost community 为c2 的相同路由表项
5. L1 收到该路由，由于c2<c1,新的路由表项将立刻被安装到RIB表〉该操作将重定向访问H1 本地流量到L4
6. 针对H1 发出本地ARP请求，该请求带有很短的超时以确保VM真正迁移走了〉一旦ARP更新超时计数器，L1 删除旧的本地路由并发送MP-BGP撤消消息更新



© 2013 Cisco and/or its affiliates. All rights reserved.

通过 VDP/GARP 发现主机迁移 —— 步骤(3)

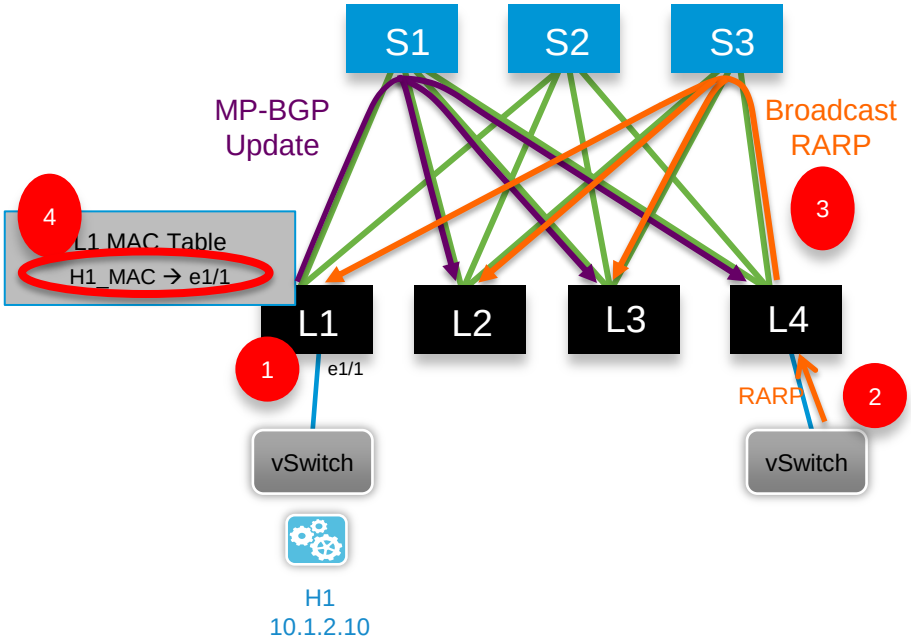
7. L4 收到路由撤销宣告，删除带有cost community c1的旧的路由表项，并且修改本地路由的cost community 为c1, 对外宣告该路由表项
8. L1 收到路由更新，修改RIB表中该表项的cost community 〉上述操作保证H1迁移回到L1的处理步骤和之前的步骤一致



© 2013 Cisco and/or its affiliates. All rights reserved.

通过 RARP 发现主机迁移
—— 步骤(1)

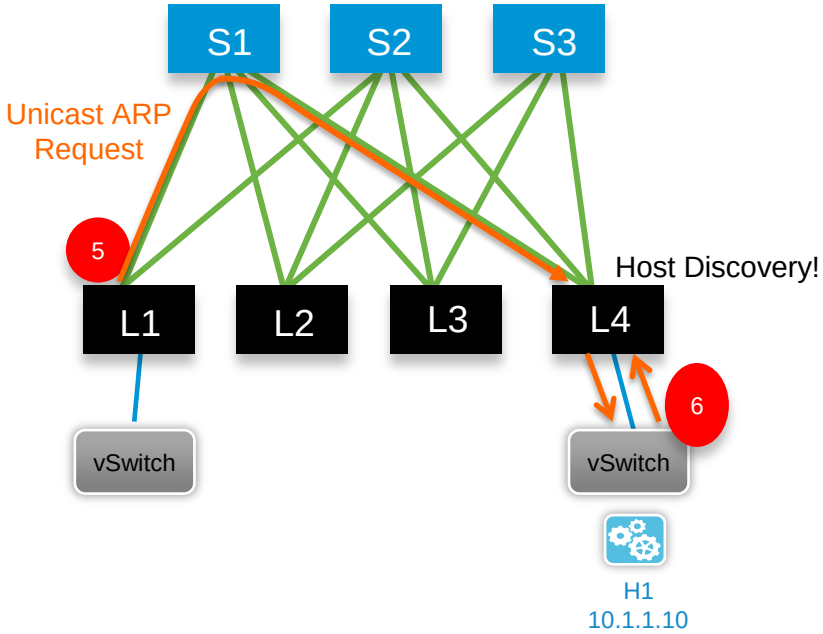
1. H1主机路由被L1 通过MP-BGP宣告
2. H1主机迁移到L4 所连接的vSwitch，虚拟层hypervisor 产生RARP 帧
3. 该RARP帧广播到整个交换矩阵，所有叶子节点接收到它。
4. L1 接收到该RARP帧，检查MAC地址表去校验是否那个主机是本地知道的（暗示是否VM迁移）



© 2013 Cisco and/or its affiliates. All rights reserved.

通过 VDP/GARP 发现主机迁移 —— 步骤(2)

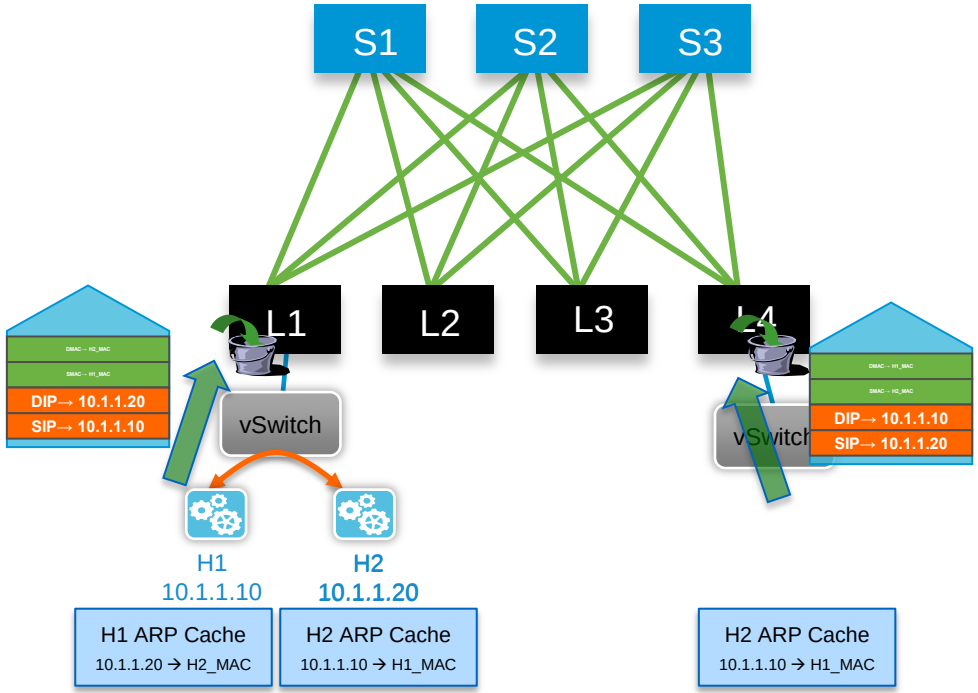
5. 在检测到可能的H1迁移后， L1 发送一个单播的 ARP请求到H1 (报文负荷中包括src-MAC=G_MAC, src-IP=anycast-VIP and dst-IP=10.1.1.10)
6. H1 收到该ARP请求后响应，该响应被L4 处理，因此可以在本地发现H1 ，此后所有的处理操作和前面讨论的VDP/GARP场景完全一致。



© 2013 Cisco and/or its affiliates. All rights reserved.

主机移动
—— VMs 被连接到同一个vSwitch

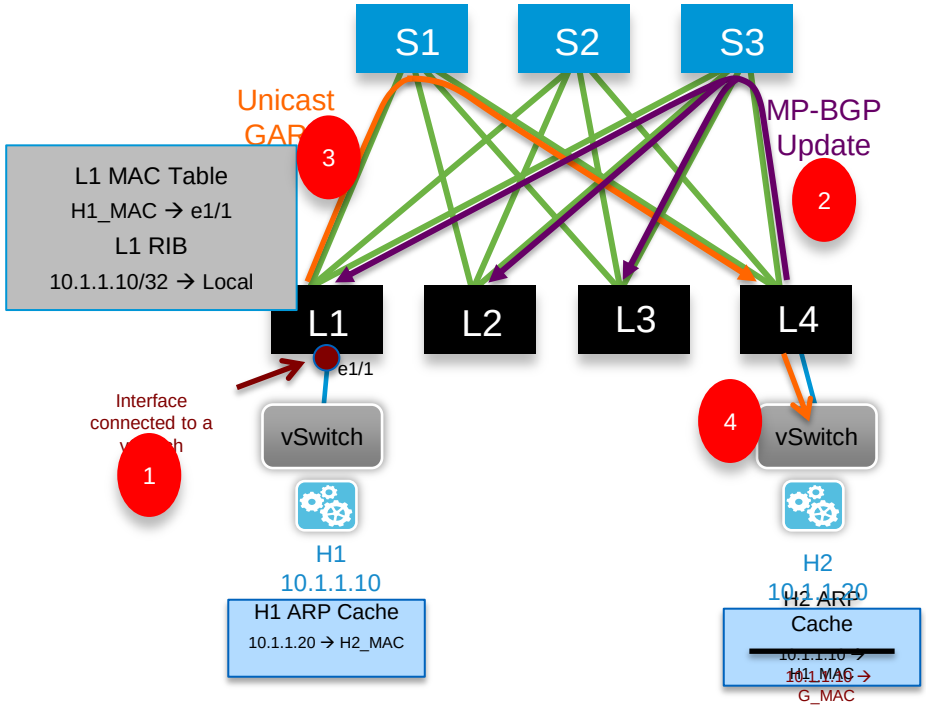
- 主机H1和H2配置在Proxy-Gateway模式下的同一个子网中
- H1和 H2 能够 通过vSwitch ARP和学习彼此的MAC地址 → 它们之间直接通信
- H2主机迁移到L4 所连接的vSwitch，并且被本地发现，现在H1和H2的通信中断直到它们的ARP缓存中修改到G_MAC地址
- 两个办法去减少该情况下发生的丢包：
  1. 在交换矩阵层面处理VM ARP缓存的刷新
  2. 部署N1Kv作为vSwitch，以至于VEM能够重写L2 报头信息



© 2013 Cisco and/or its affiliates. All rights reserved.

VMs 被连接到同一个 vSwitch
——在交换矩阵层面处理VM ARP缓存的刷新 (1)

1. 用一个CLI命令来指定L1的本地接口连接到一个vSwitch (eth1/1)〉需要跟踪所有连接到这个vSwitch上的主机
2. 一旦vMotion，H2被L4发现（利用GARP/RARP），L4发送一个MP-BGP 更新给其它叶子节点
3. 当L1 收到路由更新，它将发送一个GARP帧给H2，报文负荷中将包括G-MAC作为src-MAC 和H1 IP (10.1.1.10) 作为 src-IP*
4. H2 收到该GARP帧，升级它的ARP缓存中H1 和MAC的绑定信息

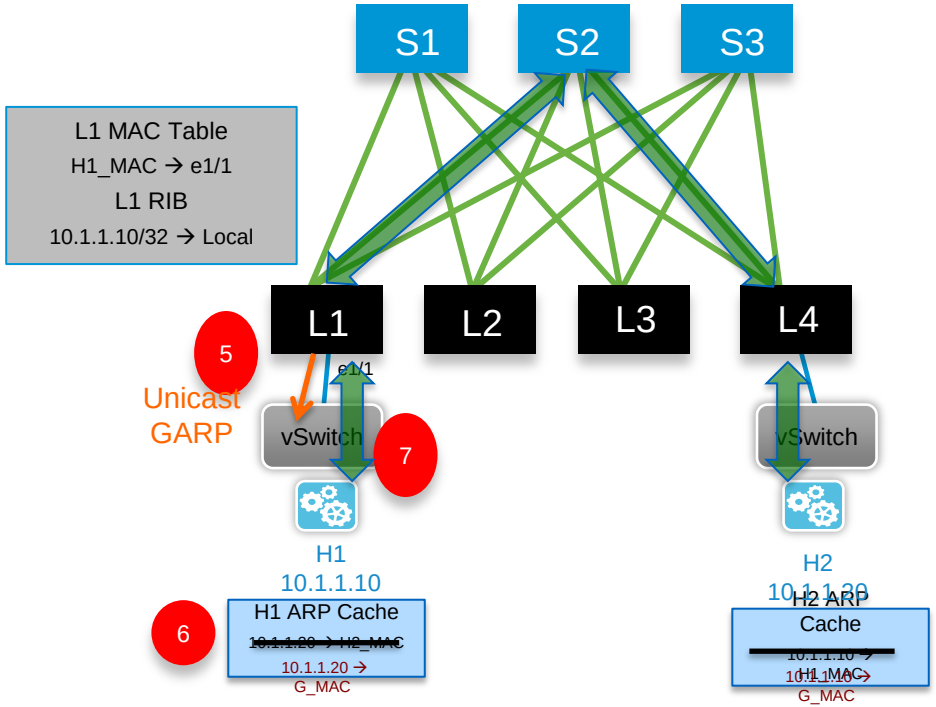


*针对该vswitch port (eth1/1)所发现的每个本地主机，一个独立GARP帧将被L1发送到H2

© 2013 Cisco and/or its affiliates. All rights reserved.

VMs 被连接到同一个 vSwitch
——在交换矩阵层面处理VM ARP缓存的刷新（2）

5. L1也产生一个GARP帧发送到从vSwitch 端口(eth1/1) 本地学习到的每个主机〉报文负荷将包括G_MAC作为src-MAC, H2 IP (10.1.1.20) 作为src-IP
6. H1 收到GARP帧，更新ARP表中H2和MAC的绑定
7. 两个主机之间的流量被成功重建

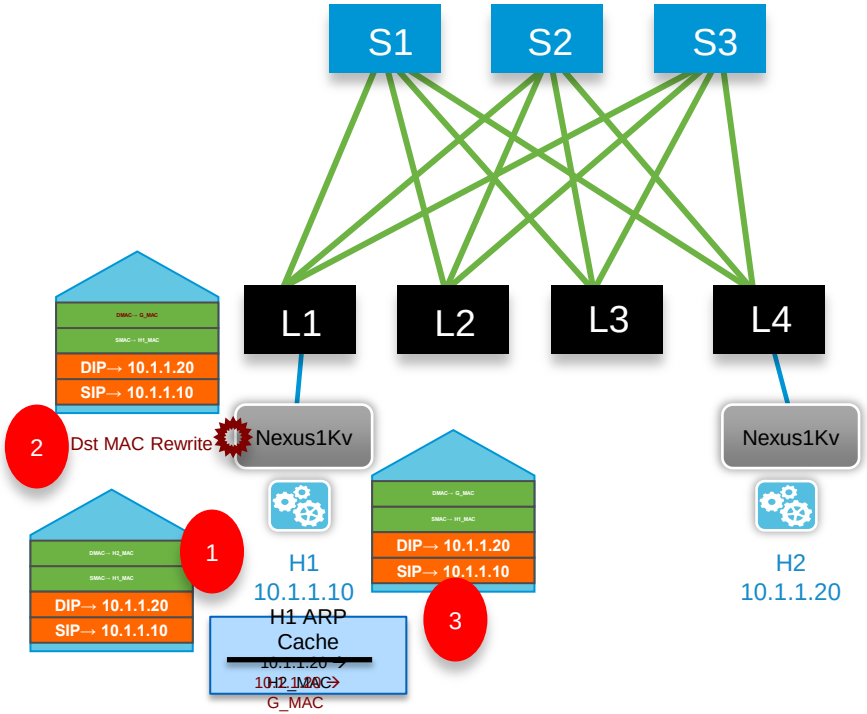


© 2013 Cisco and/or its affiliates. All rights reserved.

VMs 被连接到同一个 vSwitch
—— 利用N1K处理VM ARP缓存的刷新——部署N1K的优势

部署Nexus1000v作为vSwitch 可以减轻由于VM迁移带来的VM ARP缓存更新的问题

1. 一旦H2迁移到L4所连接的N1Kv，H1 和H2试图通过发送流量到它们各自的MAC地址来彼此通信
2. 流量被N1Kv收到，通过连接DFA叶子节点的上联链路泛洪出去〉在发出该帧之前，N1Kv VEM 重写了目的MAC地址为anycast gateway MAC (G_MAC)
3. 一旦ARP表项自然刷新（在ARP刷新计数器过期后），VM将产生针对G_MAC的数据报文



© 2013 Cisco and/or its affiliates. All rights reserved.

主机迁移
—— 如何处理 IP 哄骗情况

- 需要区别合法的主机迁移的场景和在两个独立的DFA叶子节点上出现重复的IP地址的情况
- 重复IP地址检测依赖于给定子网的DFA转发模式：
  - 在Anycast-Gateway 模式下, VM发出免费 ARP 去检查重复的IP地址，ARP报文在整个交换矩阵中广播, 所连接的VM都将接收到
  - 在Proxy-Gateway 模式下, 该GARP被限制在本地, 所以带有相同IP地址的VM主机可以被发现在不同的位置 → 目前的做法是通过显示日志来警告用户

© 2013 Cisco and/or its affiliates. All rights reserved.

DFA 转发模式比较

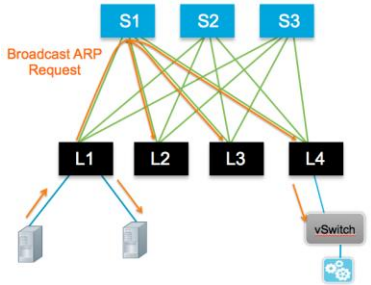
	Proxy-Gateway	Anycast-Gateway	Non DFA Mode*
VLANs/Subnets 在叶子节点间延展	支持	支持	不支持
公共的Anycast GWY IP 遍历叶子节点	支持	支持	不支持
公共的Anycast GWY MAC 遍历叶子节点	支持	支持	不支持
使用 Proxy ARP/ND	是 (如果目的前缀存在在RIB标会响应ARP/ND)	不使用	不使用 (如果显式配置，响应不依赖于目的前缀是否是已知还是未知)
在L2 域内 进行 ARP 泛洪	不会	是 (整个 DFA Fabric)	是
同一子网内转发	路由(TTL减少)	桥接	桥接
未知单播帧处理	丢弃	泛洪(整个 DFA Fabric)	泛洪 (本地)
静默主机发现	不支持	支持	支持

* 在DFA叶子节点后面的VLANs/IP 子网仅仅是本地定义的

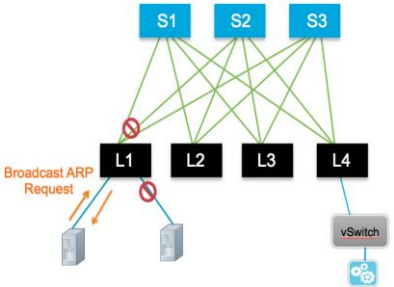
© 2013 Cisco and/or its affiliates. All rights reserved.

DFA 转发模式总结
—— 如何选择？

- Anycast-Gateway模式 在以下的场景下推荐部署：
  - ✓ 传统的交换基础设施 需要连接到DFA交换矩阵的边缘，工作负载在传统交换机和DFA 叶子节点之间流动
  - ✓ 需要支持和发现静默主机



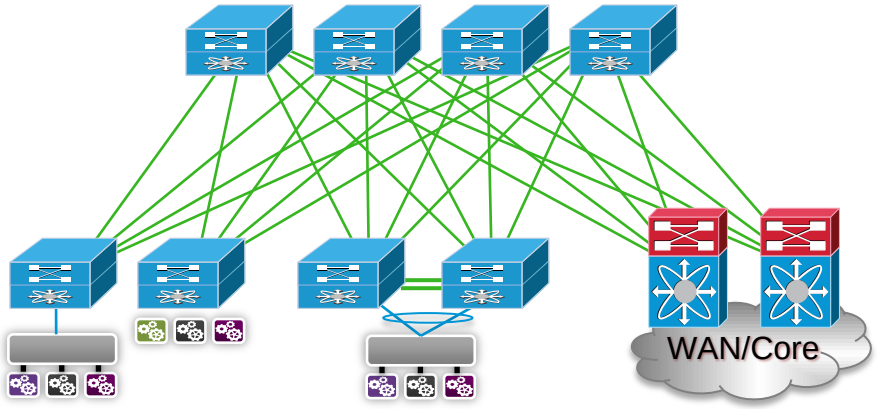
- Proxy-Gateway 模式适用于其它所有场景（这也是缺省配置）



© 2013 Cisco and/or its affiliates. All rights reserved.

动态交换矩阵自动化架构 —— 组播转发

- 动态交换矩阵架构借助于IS-IS计算多分布树，被用来转发组播和广播流量
  - 在动态交换矩阵架构内部不需要其它组播协议(如PIM, etc.)
- 多目的树（MDTs）的根在针节点
- 入口的叶子节点通过在多路径上分布流量实现负载均衡
  - 有效地利用动态交换矩阵架构的内部链路



© 2013 Cisco and/or its affiliates. All rights reserved.

动态交换矩阵自动化架构 —— 组播转发（2）

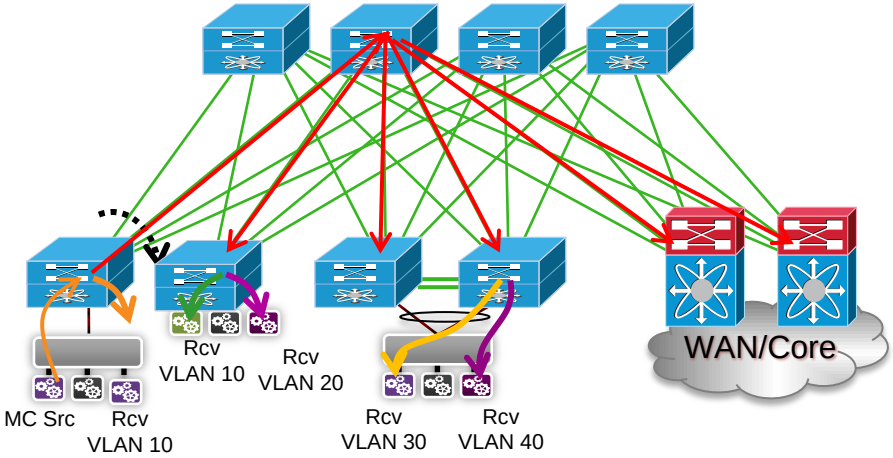
- 两级的组播复制技术
  - 入口叶子节点总是执行组播路由功能，发送单一复制（用Segment-ID 标识VRF）到动态交换矩阵

  (仅仅当本地连接的接收者和源是同一个Segment-ID中，需要L2组播复制)
  - 针节点复制流量到叶子节点
  - 目的叶子节点本地复制到服务器所在端口
- 合理的优化以允许在针节点执行修剪操作 (每租户/VRF或者每组）

  通过IGMP学习的组播组信息被BGP宣告

  →针节点必须参与到BGP控制平面

  需要 N6K 或者 N7K F3 做为针节点设备



© 2013 Cisco and/or its affiliates. All rights reserved.

议程

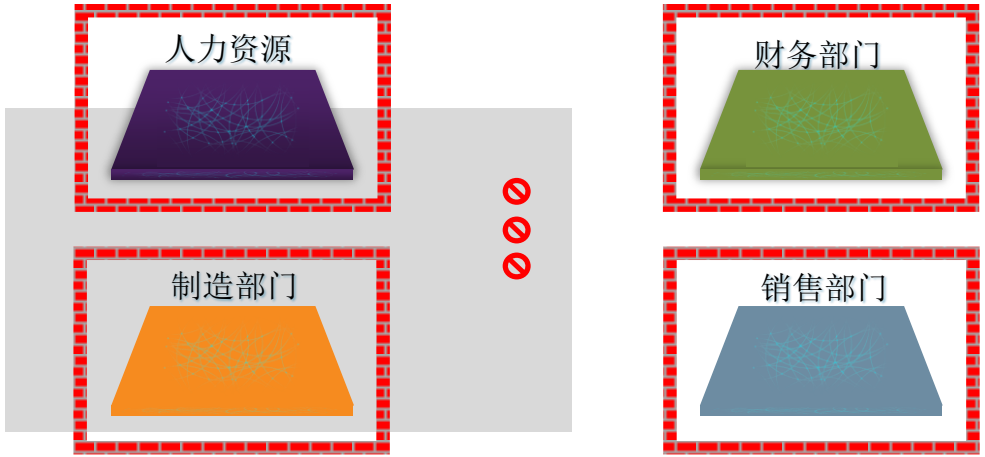
- DFA 需求和功能
- 交换矩阵管理
- 工作负载自动化
- 最优化的网络
  - 设备角色和拓扑
  - 控制平面
  - 转发平面
- 虚拟交换矩阵
- DCI, 服务集成, 迁移和互操作
- 路标和产品

© 2013 Cisco and/or its affiliates. All rights reserved.

针对公共或者私有云环境的虚拟交换矩阵架构

优点

- 任何工作负载, 任何虚拟交换矩阵, 快速部署
- 可扩展的安全虚拟交换矩阵
- 虚拟交换矩阵的租户可视化
- 路由/交换的分段



© 2013 Cisco and/or its affiliates. All rights reserved.

虚拟交换矩阵
—— 引入分段ID Segment-ID概念

- 传统的VLAN空间用12位表示 (802.1Q tag)

  限制了在数据中心内部最大分段数量是4096个

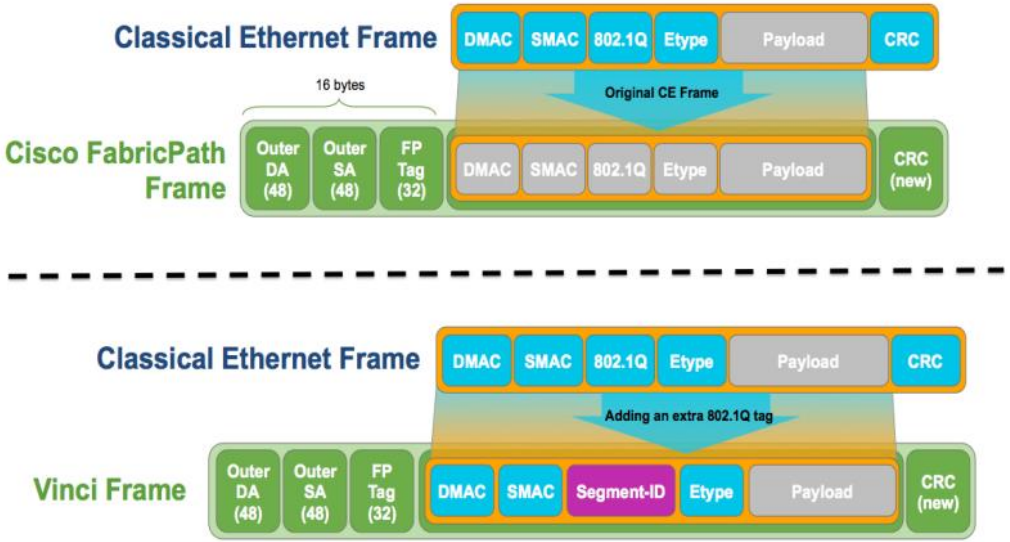
- 分段ID Segment-ID 解决方案通过使用两个 802.1Q tag ，从而使地址空间扩展到24位，理论上可以支持到大约 ~16M 个L2分段 (首发销售时可以支持到16K个)

  分段ID Segment-ID 做为L2报头的一部分 在叶子节点添加和删除。

- 分段ID Segment-ID做为整个集成的动态交换矩阵的一部分，是在叶子和针节点通过硬件方式提供

- 针节点通常转发流量基于FP的 Switch-ID 值, 但是也能通过解析Segment-ID值，修剪多目的流量

FabricPath 帧格式

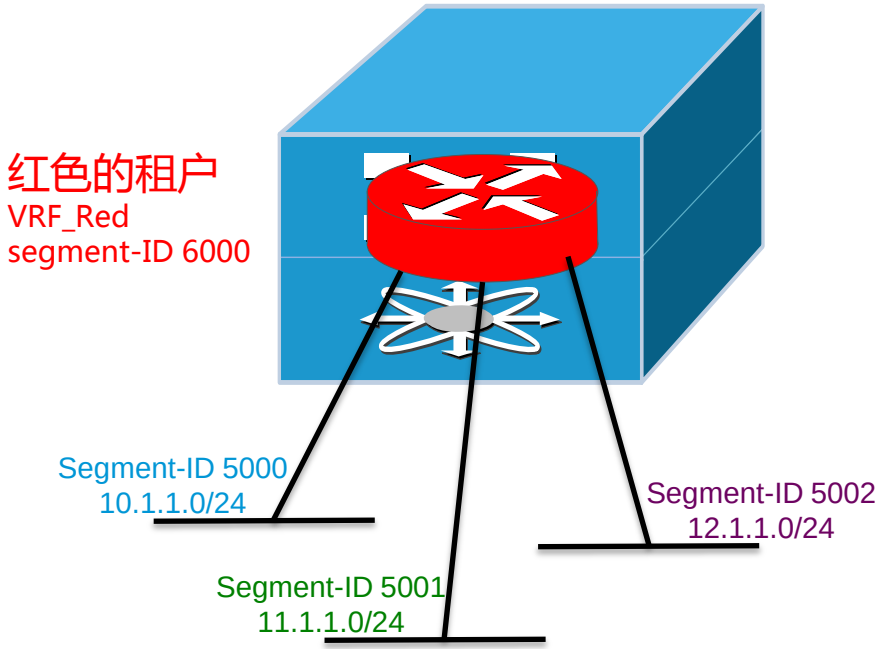


集成的交换网络帧格式

© 2013 Cisco and/or its affiliates. All rights reserved.

虚拟交换矩阵
—— Segment-IDs 的运用方式

- 在集成的动态交换矩阵边缘所定义的每个IP子网被关联到一个L2 域，也就是被一个分段ID所标识
- 多个分段能被定义到一个租户，并且通常被映射到租户所唯一标识的一个L3 VRF
- 在集成的动态交换矩阵中每个VRF用一个专用的分段ID 来唯一标识



© 2013 Cisco and/or its affiliates. All rights reserved.

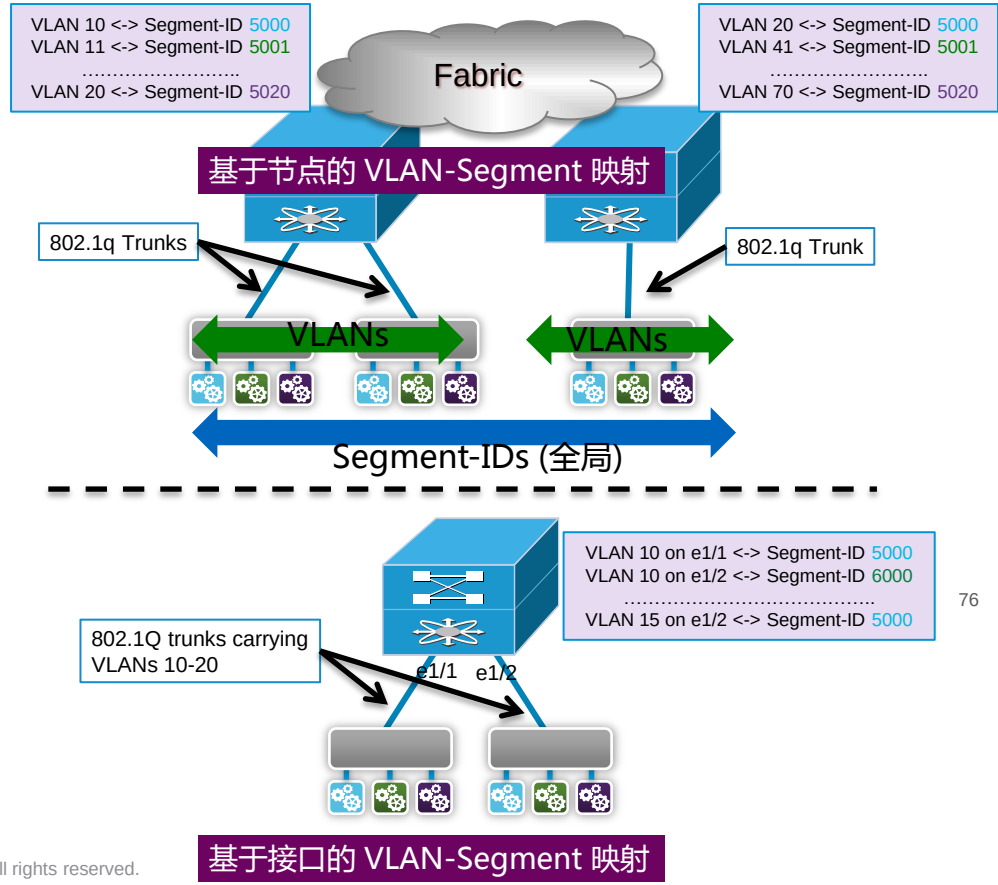
虚拟交换矩阵
——802.1Q 标记的流量与分段ID映射

- 当L2和L3 网络穿越集成的动态交换矩阵，Segment-IDs 被用来提供隔离功能
- 叶子节点从边缘设备接收到的802.1Q 标记的帧必须映射到指定的分段ID
- 在叶子节点设备层执行 VLAN到分段ID的映射操作

  VLANs 成为在叶子节点的本地标识，它被1:1映射到分段ID
- 分段ID 是全局唯一标识的，VLAN ID 成为仅在叶子节点的本地标识
- 可选的另外一种配置, 在叶子节点的物理接口层级执行映射*

  VLANs仅在每个链路接口上定义

  多个 VLAN_ID 值能被关联到每个节点设备的同一个Segment 上

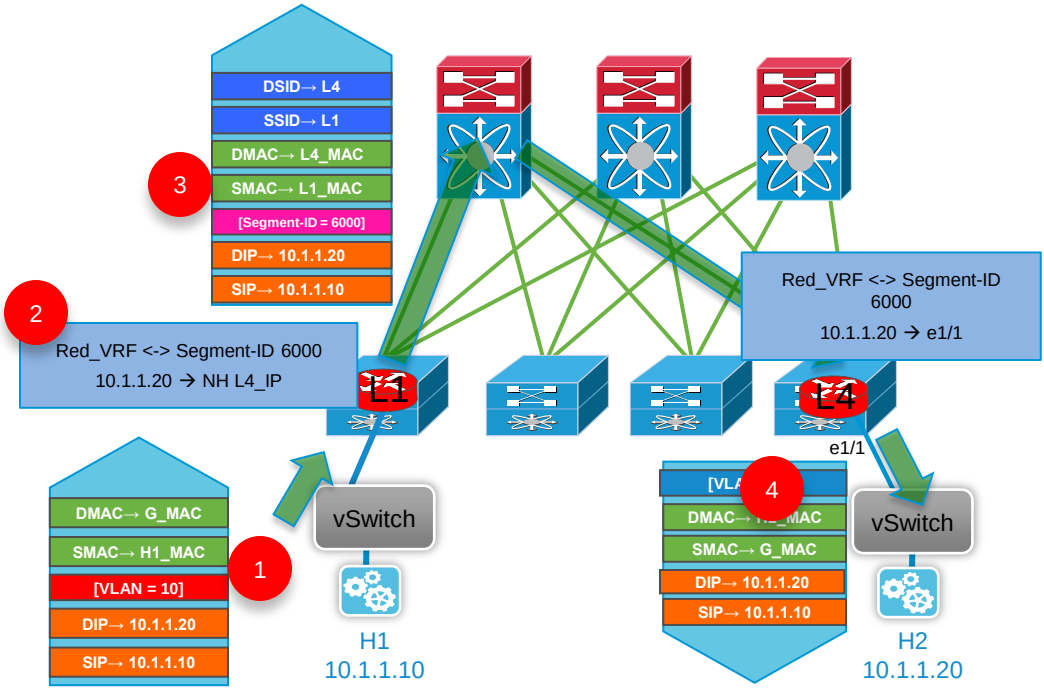


Presentation_ID
© 2013 Cisco and/or its affiliates. All rights reserved.

虚拟交换矩阵
—— 转发可路由的数据流

1. H1 发送报文到 H2 → 流量在vSwitch和叶子节点之间被标记为本地的 VLAN-ID 10
2. L1 节点在 Red VRF 中执行L3 查询
3. L1 在将报文发送到动态交换矩阵之前 添加L2 和 FP 帧头。标识红色VRF的分段ID 被添加到 L2 帧头
4. L4 接收该帧后通过分段ID关联到红色VRF.然后用本地的VLAN-ID 20 发送到主机H2

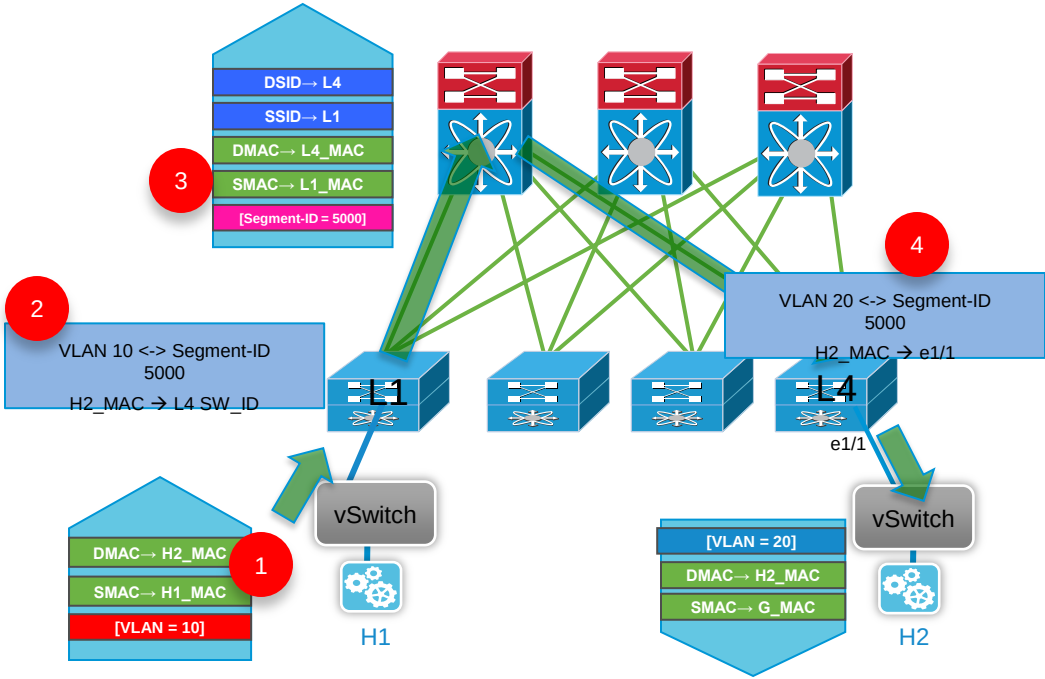
注意: 上述处理行为应用到所有的动态交换矩阵的可路由的数据流处理，包括 Proxy-Gateway Mode下同一子网或者不同子网， Anycast-Gateway Modes下的不同子网。



© 2013 Cisco and/or its affiliates. All rights reserved.

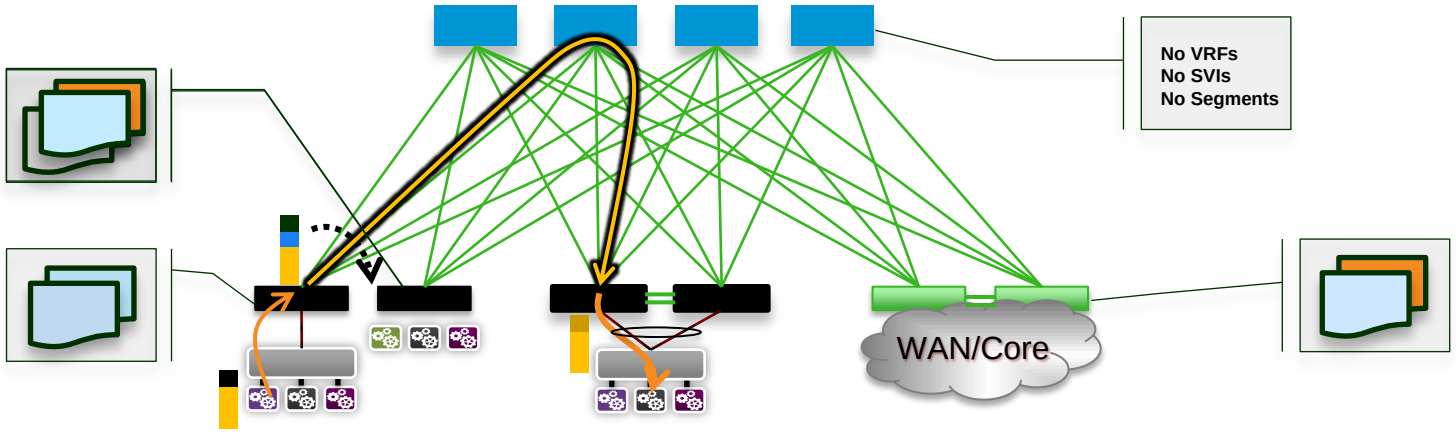
虚拟交换矩阵
——转发L2非IP的数据流或者任意播模式下的同一子网内的数据流

1. H1 发送报文到 H2 →流量在vSwitch和叶子节点之间被标记为本地的 VLAN-ID 10
2. L1在VLAN 10 所关联的分段ID (5000)的MAC地址表中执行L2 查询
3. L1 在将报文发送到动态交换矩阵之前 添加L2和 FP 帧头. 与VLAN 10 所关联的分段ID被插入到 L2 帧头
4. L4 接收该帧后通过分段ID 执行L2 查询.然后用本地的VLAN-ID 20 发送到主机H2



© 2013 Cisco and/or its affiliates. All rights reserved.

虚拟交换矩阵
—— 扩展性



- 叶子节点虚拟化配置: Segments, SVIs, VRFs
- 针节点/叶子节点传输: 转发基于外层封装
- Border Leafs边界叶子节点: 出站流量仅仅需要 VRFs 信息
- 当租户内最后一个主机下线时，在叶子节点删除 相关Segments, SVIs, VRFs 配置

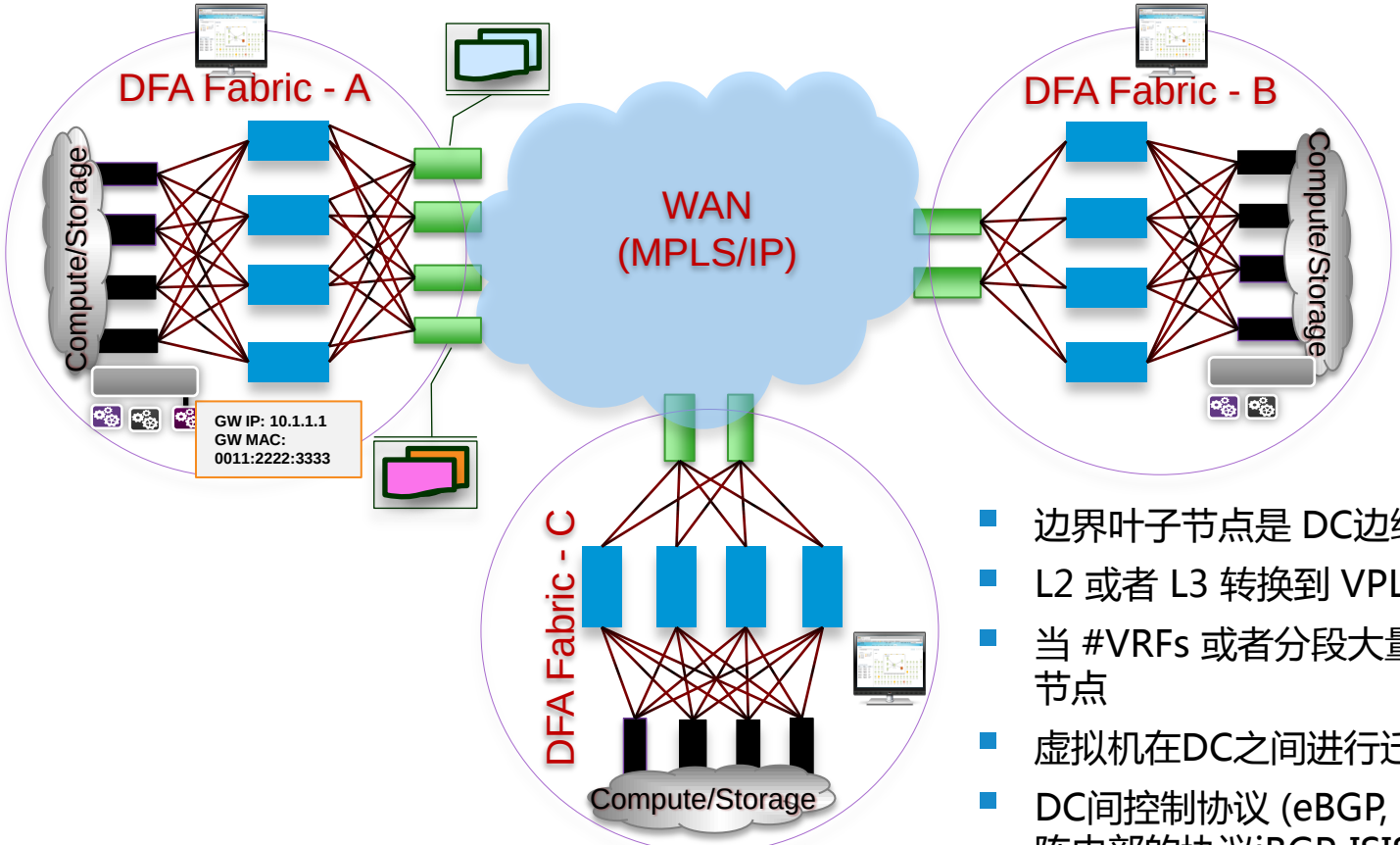
© 2013 Cisco and/or its affiliates. All rights reserved.

议程

- DFA 需求和功能
- 交换矩阵管理
- 工作负载自动化
- 最优化的网络
  - 设备角色和拓扑
  - 控制平面
  - 转发平面
- 虚拟交换矩阵
- DCI, 服务集成, 迁移和互操作
- 路标和产品

© 2013 Cisco and/or its affiliates. All rights reserved.

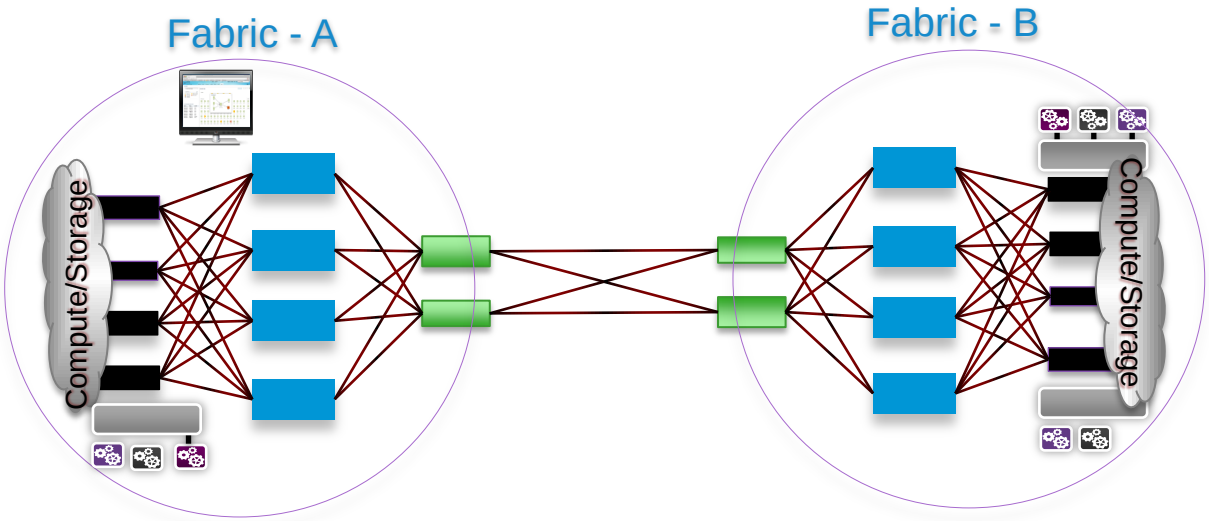
动态交换矩阵自动化架构
DC 互联通过通用传输方式（“WAN DCI”）



- 边界叶子节点是 DC边缘或者连接到DC边缘
- L2 或者 L3 转换到 VPLS, L3-VPNs, OTV, LISP
- 当 #VRFs 或者分段大量增加时，可添加边界叶子节点
- 虚拟机在DC之间进行迁移 (L2 或者L3 相邻)
- DC间控制协议 (eBGP, MPLS) 不同于动态交换矩阵内部的协议iBGP, ISIS

© 2013 Cisco and/or its affiliates. All rights r

动态交换矩阵自动化架构
DC 互联通过裸光纤（“Metro DCI”）



- 在DC之间扩展动态交换矩阵
- 整个网络很像一个巨大的三级或者四级交换网络
  - ⟹ 可为DC内部或者穿越DC的任何IP流量提供最优转发路径
- 整个动态交换矩阵可部署单一的DCNM实例

© 2013 Cisco and/or its affiliates. All rights reserved.

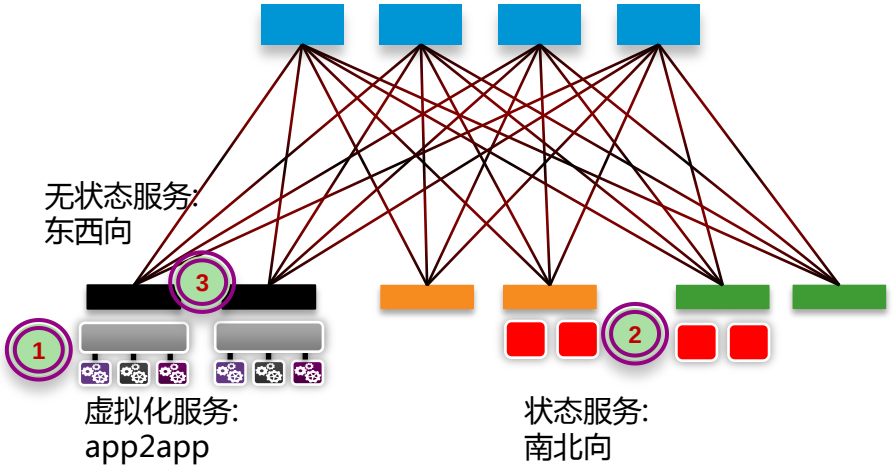
动态交换矩阵自动化架构
—— 网络服务集成

网络服务部署模式

- ✓ 虚拟化服务: 控制特定的应用到应用的流量。
- ✓ 状态服务设备: 采用专门的硬件控制南北向流量，性能受限于出站带宽容量。
- ✓ 无状态服务设备: 采用专门的硬件控制东西向流量（基于ACL处理），可处理大容量的流量转发。

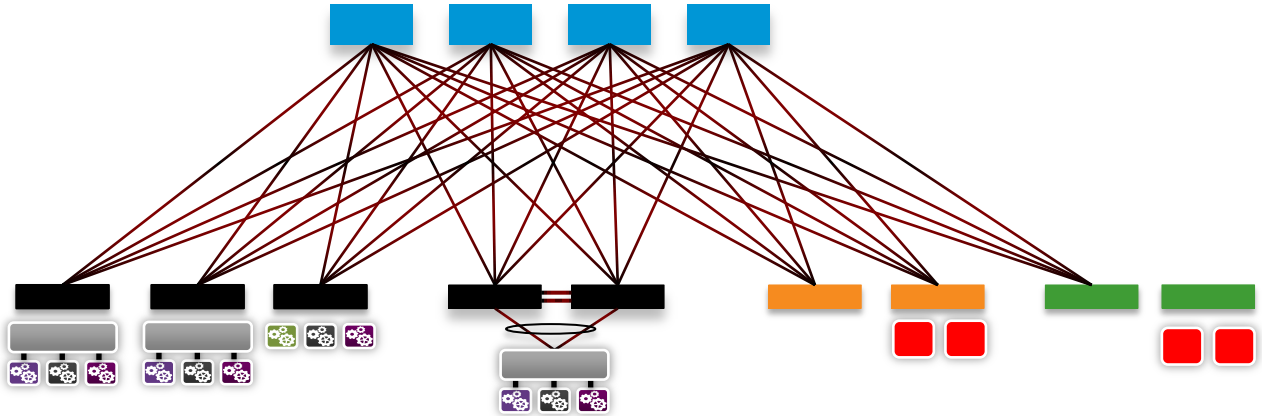
网络服务部署方式

1. 虚拟化服务: 主机范围的，和DFA无缝集成
2. 状态服务设备放置于: 服务叶子节点或者边界叶子节点
   - ✓ 通过重定向机制集成
   - ✓ 部署时需要考虑处理能力
   - ✓ 产生发卡弯和可能出现瓶颈
3. 无状态服务设备– 基于ACL的策略部署
   - ✓ 部署到适当的叶子节点
   - ✓ 线速处理



© 2013 Cisco and/or its affiliates. All rights reserved.

动态交换矩阵自动化架构
—— 服务节点的连接



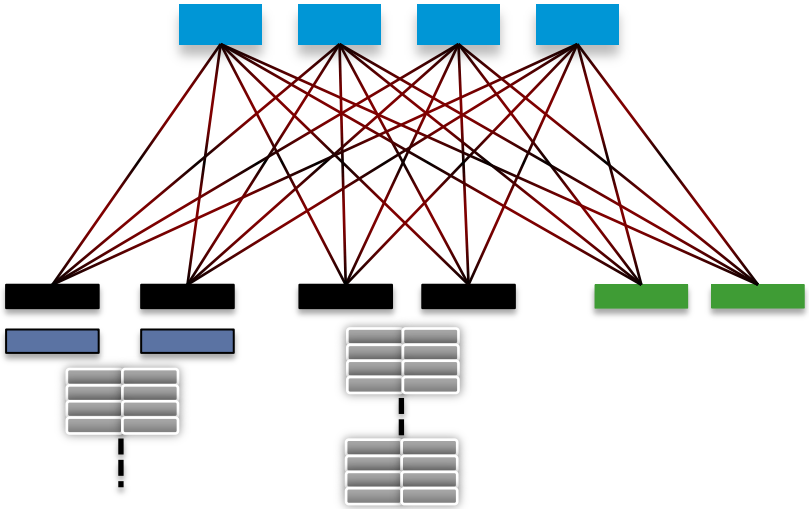
- 和物理服务器与叶子节点连接方式相似，服务节点设备被连接到服务叶子节点
- 交换矩阵处理服务流量 (IP 或者 L2) 的方式等同于服务器到服务器的流量处理
  => 交换矩阵提供泛洪避免和多租户
- 服务节点能被整个交换矩阵所共享。
- 服务节点随时可以添加
  => 当租户增加或者为了某些租户的增长而分段时

© 2013 Cisco and/or its affiliates. All rights reserved.

动态交换矩阵自动化架构
—— UCS 集成选项

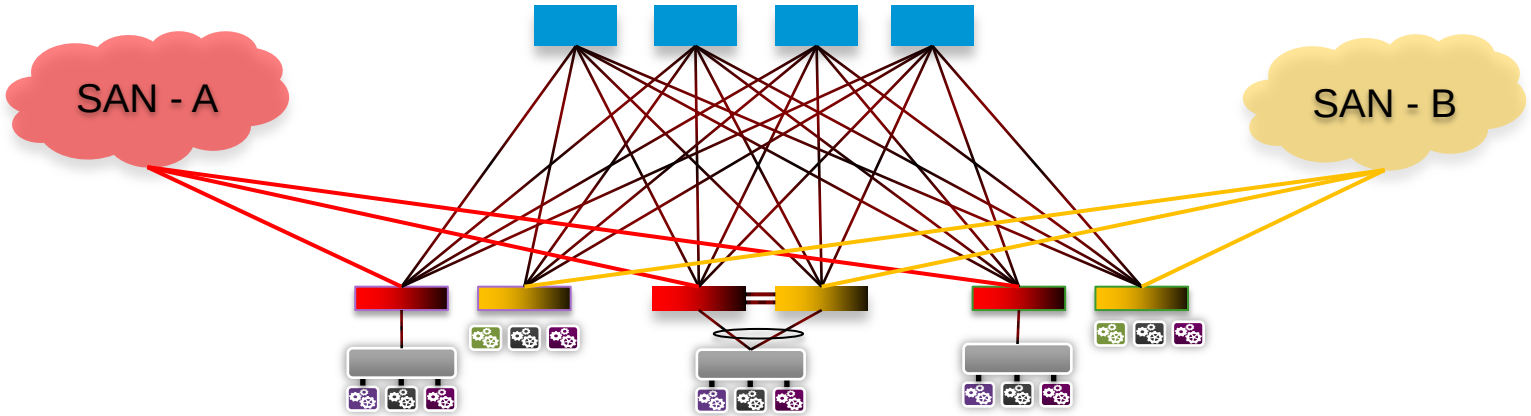
集成模式：

- ✓ 基本集成: FIs 被连接到DFA的叶子节点
- ✓ 高级集成: 下一代的 FI 将集成DFA 叶子节点芯片处理能力: 即工作在交换模式下， FIs 可以做为DFA 叶子节点
- ✓ 完全集成: Orchestration自动化服务器调配和管理工具的APIs被集成: UCSM, CPoM 和 第三方的Oschestrators自动化服务器调配和管理工具



© 2013 Cisco and/or its affiliates. All rights reserved.

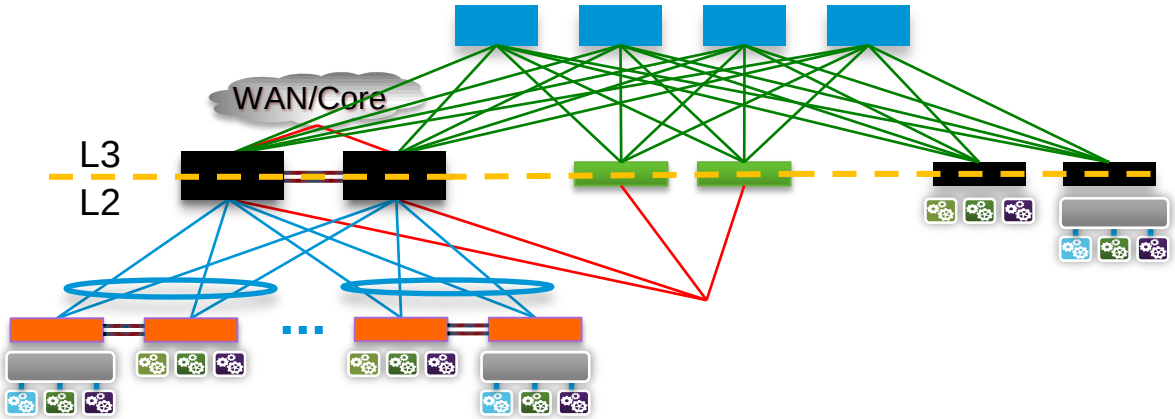
动态交换矩阵自动化架构
—— DFA和FCOE



- 主机节点采用统一的IO (聚合网络适配器Converged Network Adapters – CNAs)
- 选项 1: 在叶子节点， SAN A/B 网络的连接相互独立 (和目前支持的模式一致)
- 选项 2: FCoE 被承载在公共的交换矩阵结构中 (实现真正的统一交换矩阵部署)

© 2013 Cisco and/or its affiliates. All rights reserved.

迁移/共存场景
—— 从基于 vPC的接入/汇聚拓扑中迁移



- vPC 使能的设备
- DFA 针节点
- DFA 叶子节点
- DFA 边界叶子节点

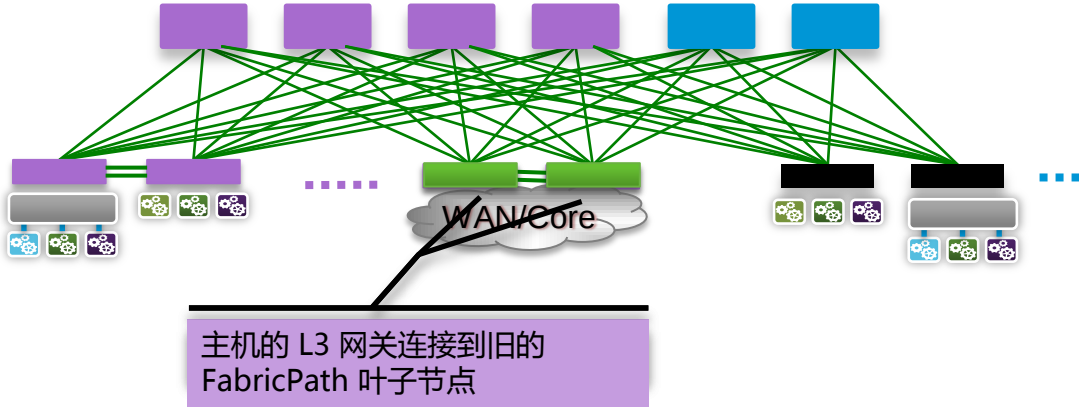
- 引入新的针节点和边界叶子节点
- 转换 vPC 汇聚设备到新的DFA 叶子节点
- 让外部连接通过新的边界叶子节点
- 按照需要添加新的叶子节点
- 按照需要添加新的针节点

部署考虑

- 可能延展旧的分区到新的交换矩阵的叶子节点
- 允许网关分布在新的交换矩阵的叶子节点和旧的汇聚设备上

© 2013 Cisco and/or its affiliates. All rights reserved.

迁移/共存场景
—— 从Spine/Leaf 的FabricPath拓扑中迁移(场景1)



- 转换 FP 边界叶子节点到DFA 边界叶子节点
- 按照需要添加新的叶子节点
- 按照需要添加新的针节点
- 重用旧的针节点

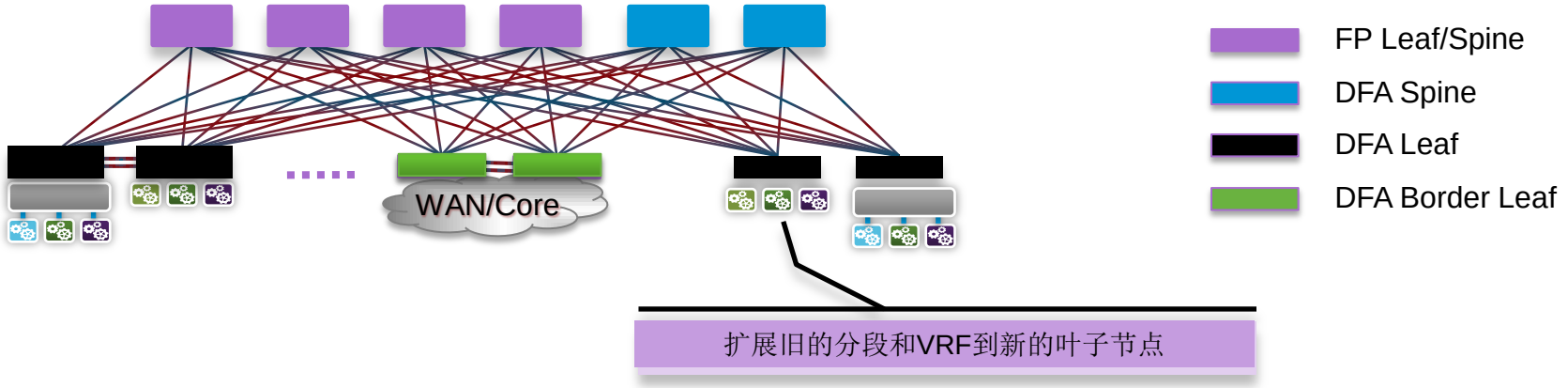
  => 替换: 针节点不修剪基于租户/组的多目的流量

部署考虑

- 可能延展旧的分区到新的交换矩阵的叶子节点
- 如果主机连接到旧的FP 叶子节点部署集中式L3网关
- 如果主机连接到新的交换矩阵叶子节点后可实现分布式网关

© 2013 Cisco and/or its affiliates. All rights reserved.

迁移/共存场景
—— 从Spine/Leaf 的FabricPath拓扑中迁移（场景2）



- 该场景假设所有FP 叶子节点具备DFA 硬件处理能力
- 转换汇聚层 (FP 边界叶子节点) 到DFA边界叶子节点以及其它 FP叶子节点到新的交换矩阵叶子节点
  不需要改变布线, 只需要升级软件和更新配置
- 按需增加新的叶子和针节点
- 重用旧的针节点
  =>替换: 针节点不修剪基于分段ID信息的多目的流量

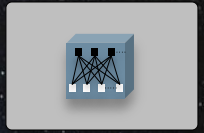
© 2013 Cisco and/or its affiliates. All rights reserved.

议程

- DFA 需求和功能
- 交换矩阵管理
- 工作负载自动化
- 最优化的网络
  - 设备角色和拓扑
  - 控制平面
  - 转发平面
- 虚拟交换矩阵
- DCI, 服务集成, 迁移和互操作
- 路标和产品

© 2013 Cisco and/or its affiliates. All rights reserved.

DFA 产品支持情况 @ FCS (Q4CY13)



Platforms	N7K + F2	N7K + F3	N6K	N5500	N3548
Time	1HCY13	Q4CY13	Q4CY13	Q4CY13	Q1CY14
Software	6.2	7.0	6.2	6.2	7.0
Leaf	✖	✔	✔	✖	✖
Spine	✔	✔	✔	✖	✖
Border Leaf	✖	✔	✔	✖	✖

© 2013 Cisco and/or its affiliates. All rights reserved.

动态交换矩阵自动化部署方案（首发销售时）

Cloud Stacks & Orchestration Tools

Compute & Storage

cloudstack

Microsoft System Center

vmware

openstack

cloupia

Network

DCNM/CPoM

Network Services

Services Controller

Nexus 7X00 (F2/F2e/F3)

High Density Spine

Nexus 6004

Medium Density Spine

ToR Leafs

Nexus 6001

High Density Leafs

Nexus 6004

Fabric Extenders

Nexus 2x00

Fabric Extenders

Nexus 2x00

Nexus 7X00 (F3)

Border Leafs

CISCO
TM