从“路由”回归“交换”

--探讨数据中心网络的演变

刘 洋

思科中国互联网运营商事业部
技术总监

“交换”的烦恼

- 物理连接层次
- 透明生成树，二层多路径，网络收敛
- Unicast Flooding，环路，广播风暴

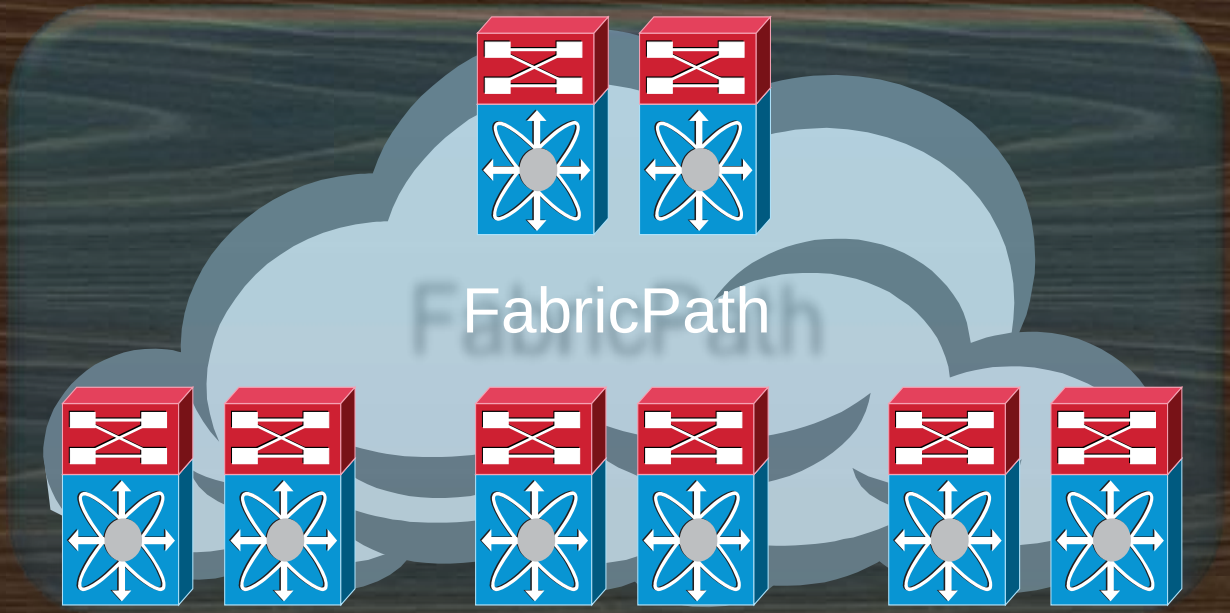
“路由”后的幸福生活

- ECMP（Equal Cost Multi Path）
- 平滑扩展；
- 快速收敛；
- 防止广播风暴；

烦恼

- 集群的规模
- 网段地址规划
- 路由控制平面
- 虚机
- 开放平台，云计算
- 价格
- Dumb Big Flat

从“路由”回归“交换”
--大型数据中心的交换网络



- Turn your network into a Fabric!
- 关键技术：FabricPath / Trill

FabricPath对于二层交换的创新

- 实现交换机间多条路径同时转发流量ECMP（Equal Cost Multi Path）；去除透明生成树
- 类似路由网络的平滑扩展；
- 快速收敛；
- 防止广播风暴（TTL）；
- 保持原有二层网络
- 基于会话的MAC地址学习
- 成本降低

FabricPath的设计目标

Switching

- Minimal Configuration
- Plug & Play
- Auto Discovery
- Auto Learning
- Flat Addressing
- Spanning Tree Protocol (STP)
- Slow Convergence
- Single Path
- Edge-to-Root Rigid Design
- Single Multicast Tree
- Constrained Scaleability

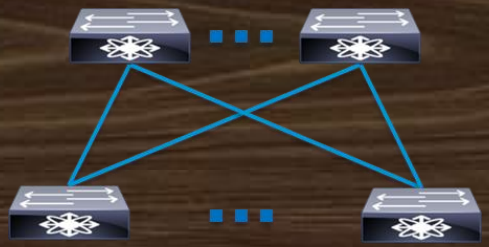
FabricPath

Routing

- Configuration Intense
- Configured Learning
- Configured Discovery
- Plan & Play
- Fast Convergence
- Multiple Paths
- Load Balancing
- Multiple Multicast Trees
- Hierarchical Forwarding
- Any-to-any Flexible Design
- Highly Scalable

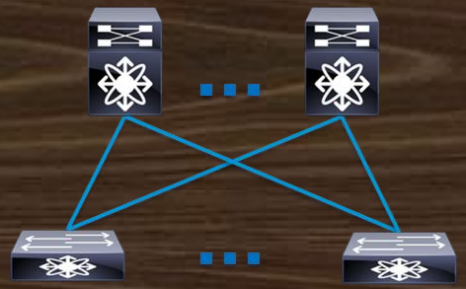
Architectural Approach for MSDC

CLOS



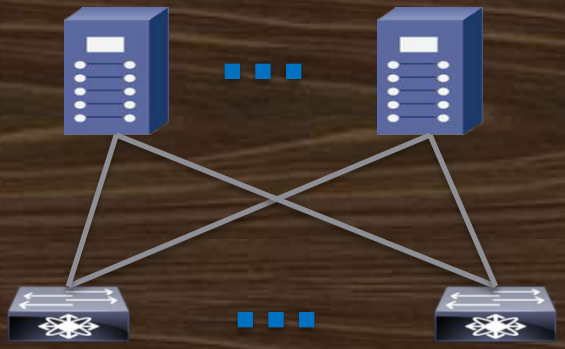
- Same node type used in all roles (Spine and Edge)
- Fine Grain Redundancy
- Additional density provided through density of node or additional layers

Scale-Up Spine Scale-Out Leaf



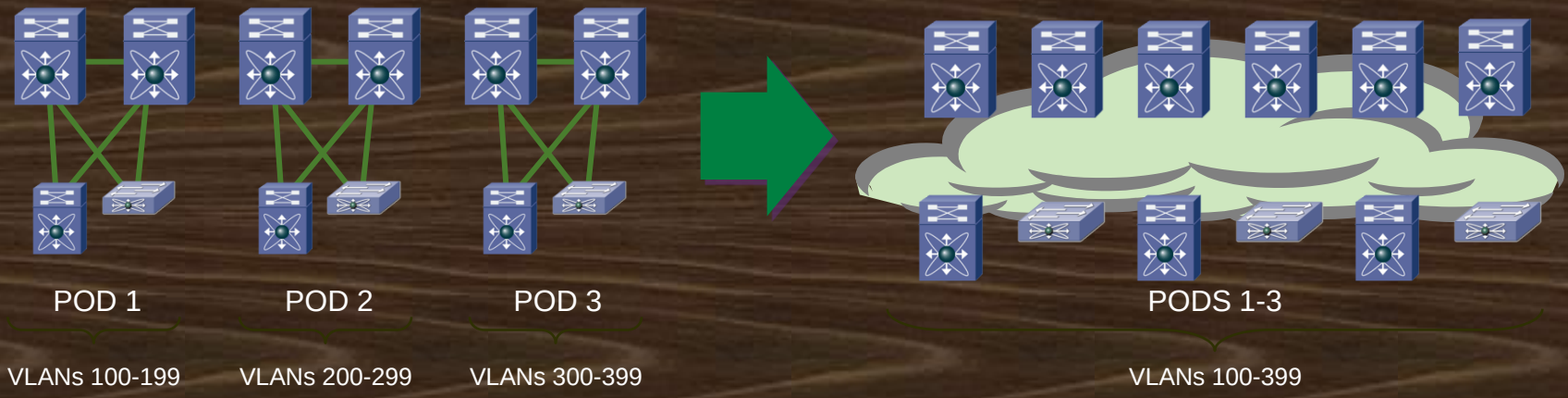
- High density spine node
- Smaller fixed leaf
- Fewer control planes than pure Clos

Lean Core Smart Edge



- Layer-1.5 Spine (Dumb Core)
- Intelligent Edge

Fabricpath 构建通用网络交换平台



大规模数据中心的通用网络交换平台
--网络对业务部署灵活性的支持

- 模块化 → 易扩展
- 网络带宽及延时的一致性 → 与服务器所处位置无关
- 业务的快速部署 → 计算资源的灵活移动和调配

Any service on any server, at any time!!!

- 可扩展性 → 业务/集群的扩展不再受制于网络
- 服务器的使用效率 → 服务器重复利用
- 可管理性 → 即插即用，配置最简化，人工干预少
- 可靠性 → 单点故障对整体业务的影响

从“路由”回归“交换”
--中小型数据中心的交换网络

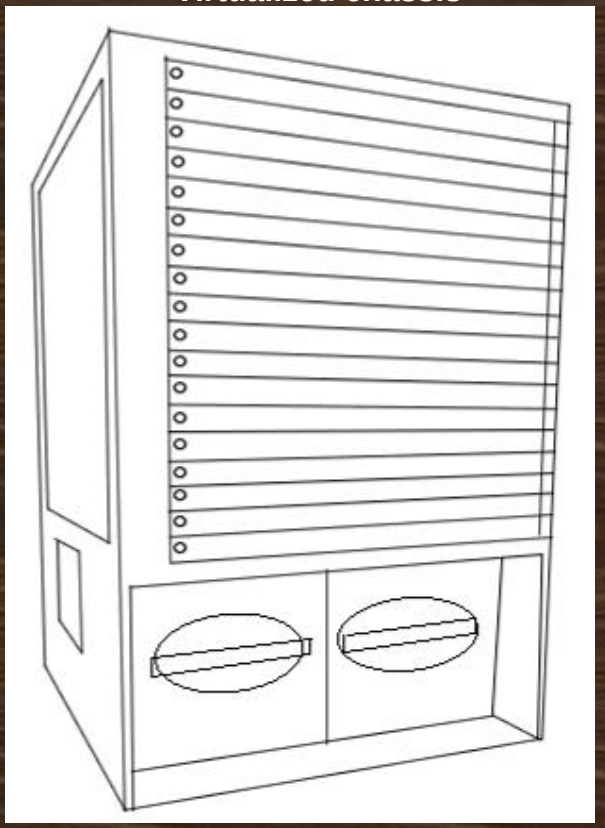
Nexus 5000

+

Nexus 2000 Fabric Extender

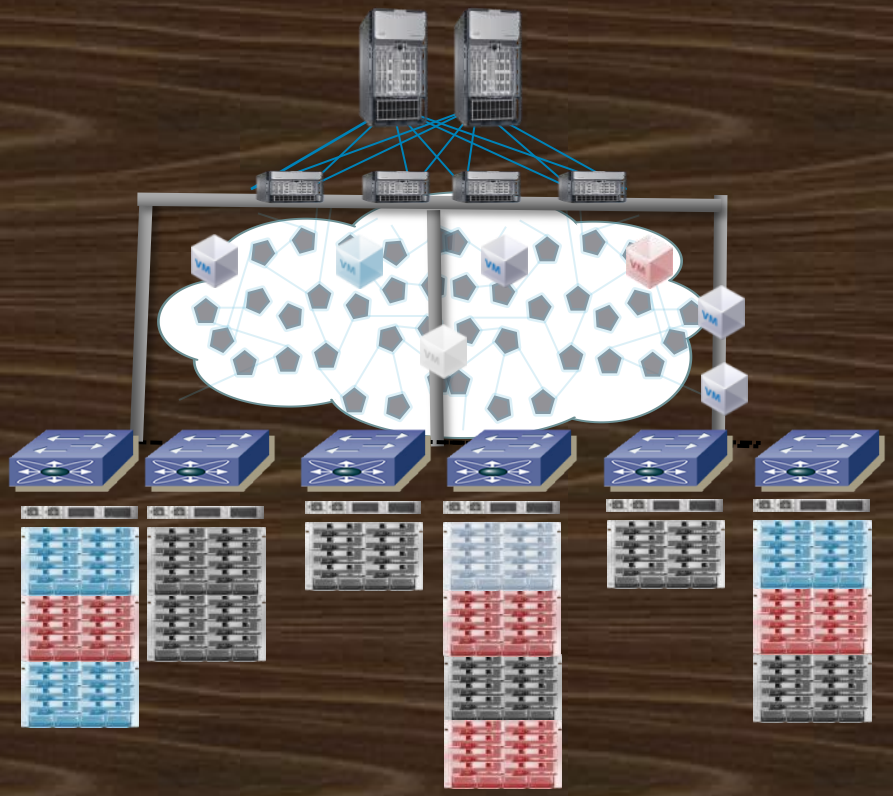
=

Nexus 7000/5000
Virtualized chassis



- Turn your network into a Switch
- 关键技术：远端扩展模块，FEX as TOR

Data Center-Wide Scalability at **Layer 2**



- 扁平化结构
- 应用在更大区域的灵活部署
- 线速的网络

谢谢