基于SDN的路由控制及流量调度

Tencent SDN实践分享

腾讯网络架构－业务需求是第一出发点

业务众多且发展迅速，需要稳定运营&高效交付的网络！

网络媒体

- 流量最高的中国门户网站
- 腾讯微博日活跃帐户数0.87亿

无线业务

- 国内领先的无线门户网站

即时通信

- 国内最大的在线社区
- 最高同时在线数1.76亿
- 微信注册用户突破3亿

网络社区

- 国内最大的互动社区网站
- QZONE月活跃帐户数6.03亿

网络游戏

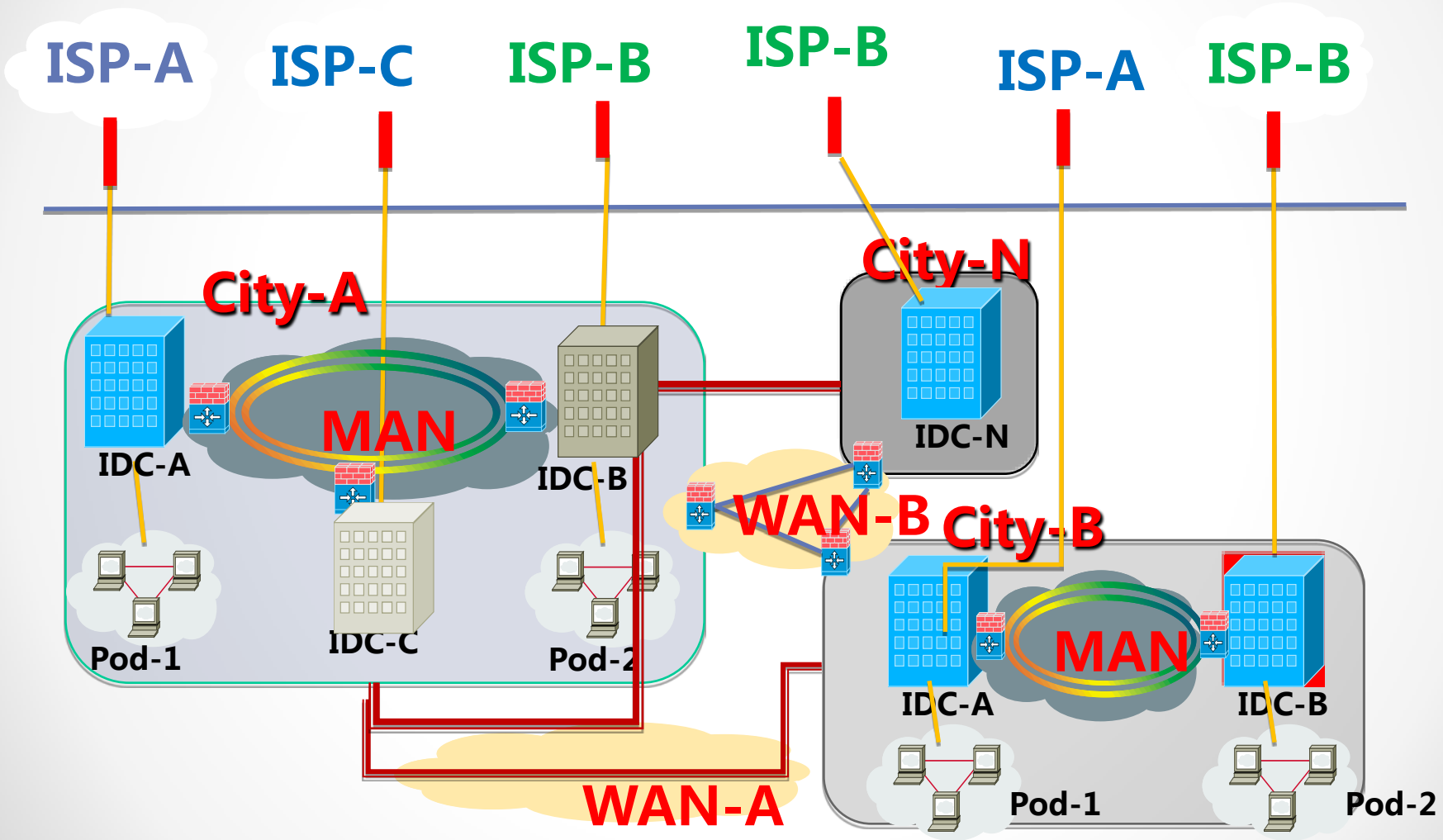
- 国内第一互动娱乐游戏平台
- QQ游戏最高同时在线账户数880万

开放平台

注：所有数字的统计口径为2012年年底

腾讯网络互联

50+DC，30w+服务器，1k+专线，3T+出口带宽



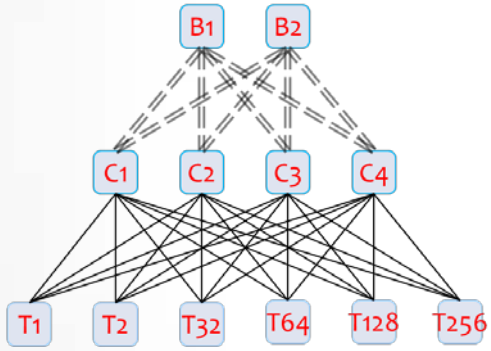
面临的挑战及SDN实践

- 高性能、安全、多特性的大规模数据中心网络
- 大容量、差异化、高利用率的广域网通讯
- 高效、自动化、智能的网络管理

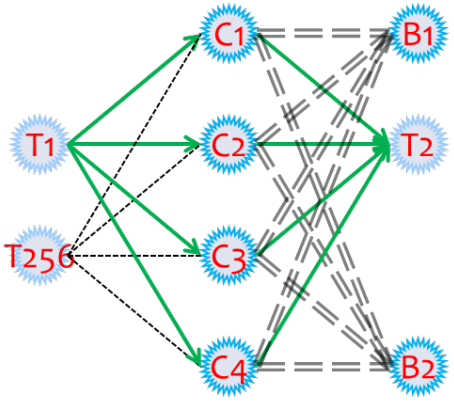
大规模网络路由协议
Sequoia Routing Procotol

目前DC内网络路由协议大多数使用OSPF

1. 生成并全局同步LSA



2. 把LSA拼接成拓扑



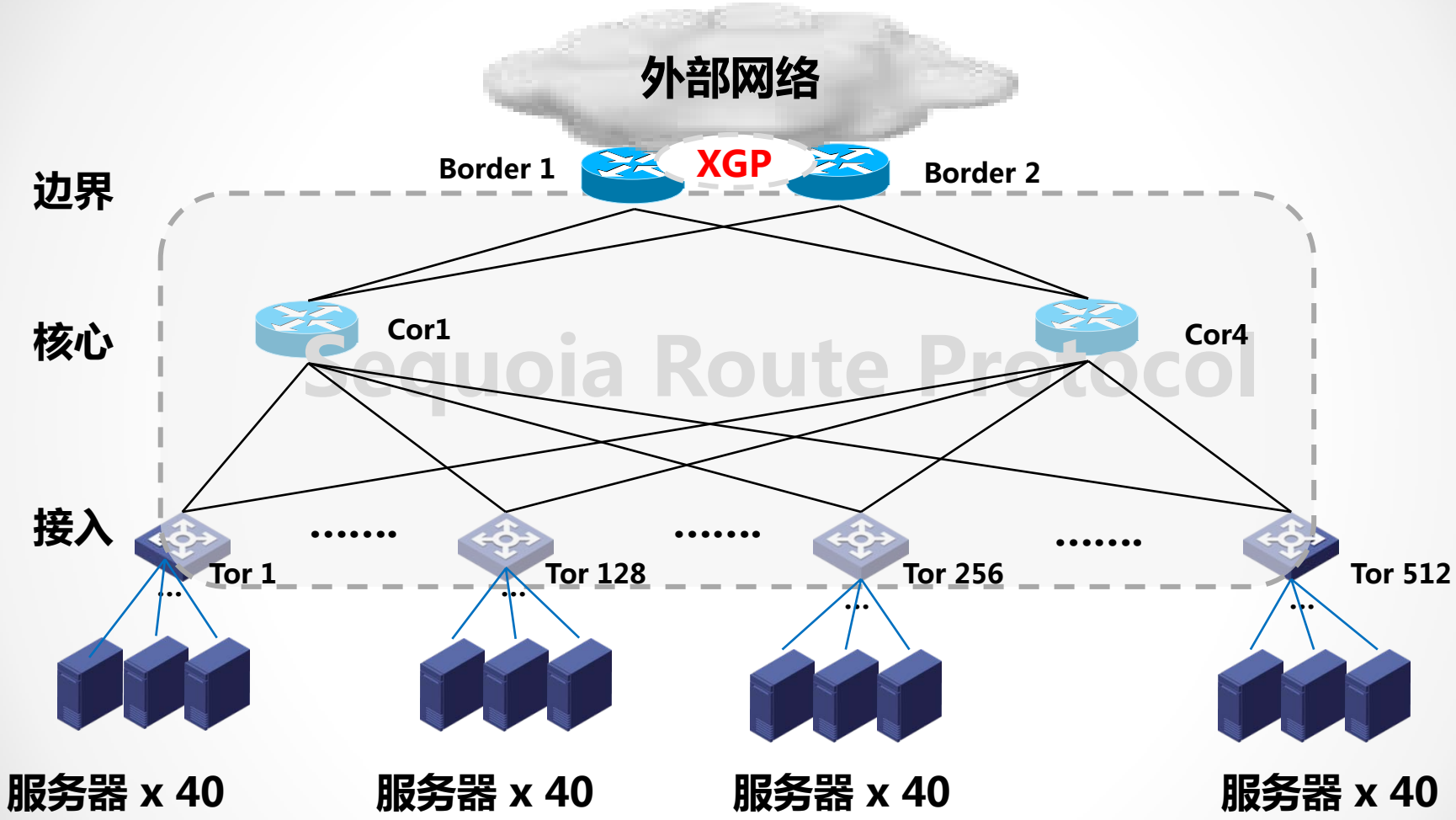
3. 计算最短路径

随网络规模增大，LSA同步和拼接性能恶化，路径难以被控制

Sequoia项目简介

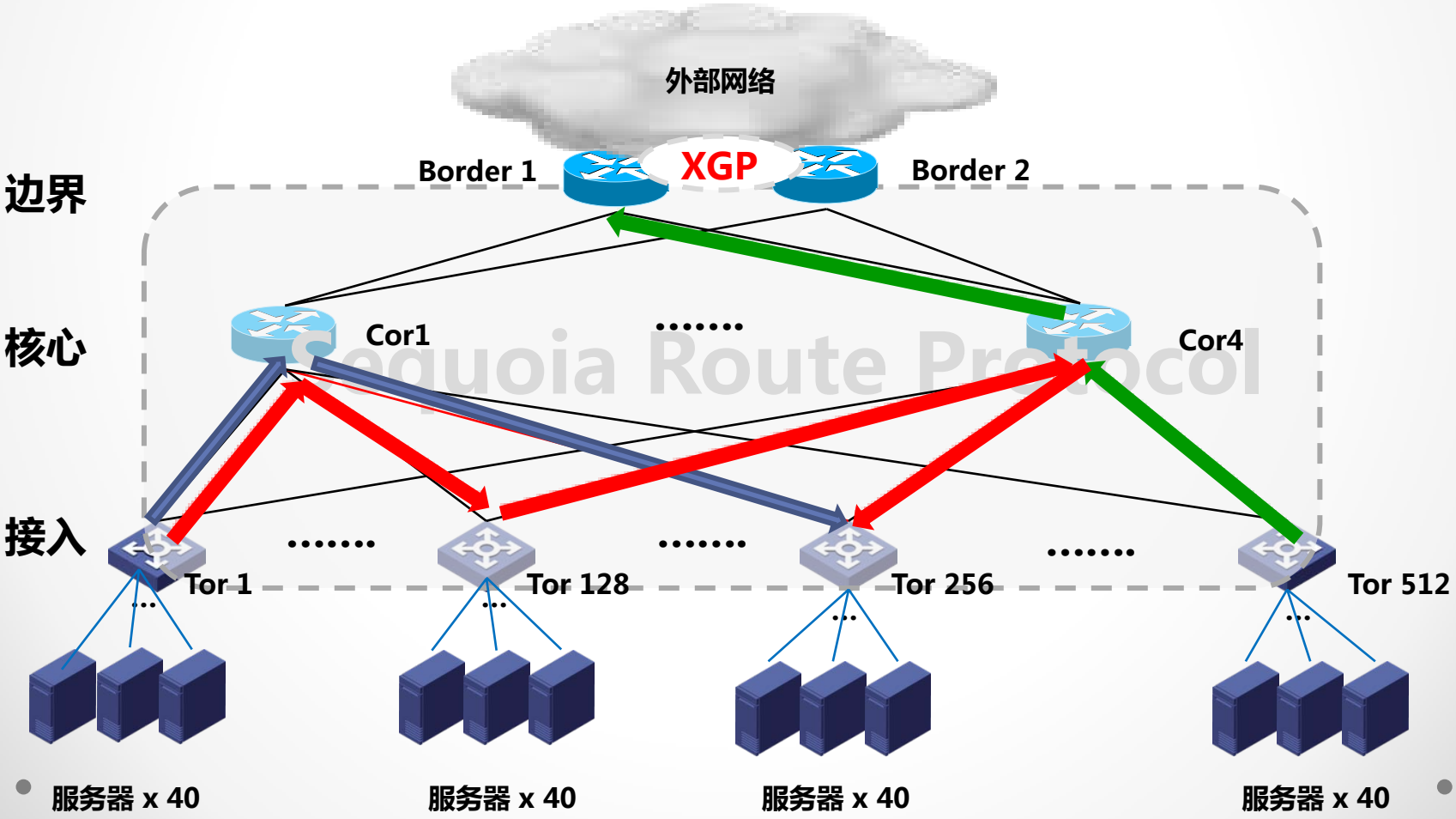
- 目标：大规模IDC架构(2W+服务器）下路由简单可控
- SRP（Sequoia Routing Procotol)
  - 价值
    - 控制器运算，减少路由协议（如OSPF）对控制平面的负担
    - 适应大规模多级CLOS架构，收敛时间不随规模增加而增加
    - 可集中调度，复杂度可控、可预测（代码极度精简）
  - 手段
    - 静态路由 + 动态收敛，简单、轻量的距离矢量协议
    - 依赖固定的网络拓扑，水平分割原则避免环路
    - 中央控制器实时监管、动态调整

SRP典型拓扑

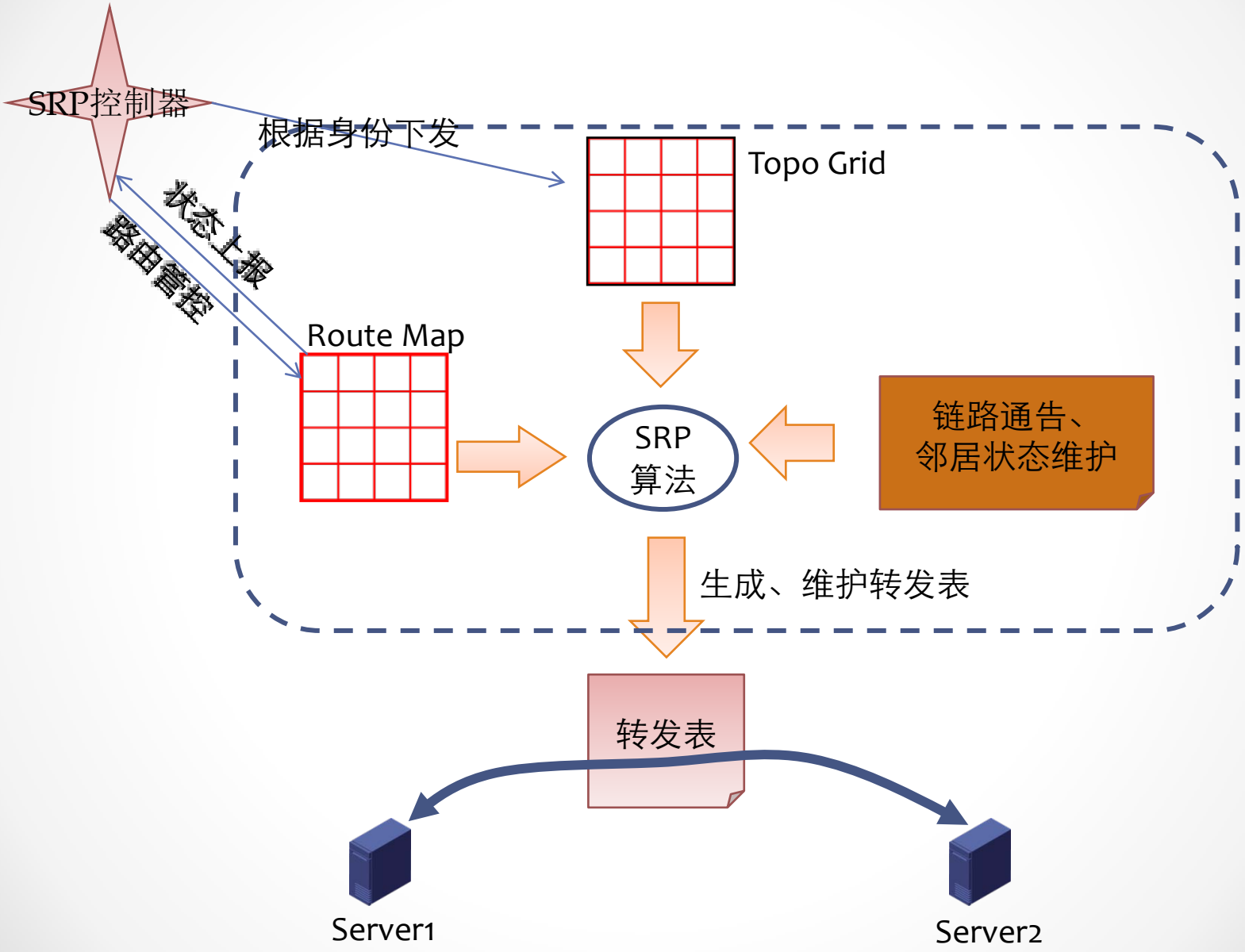

外部网络
XGP
Border 1
Border 2
边界
核心
接入
Cor1
Cor4
Sequoia Route Protocol
Tor 1
Tor 128
Tor 256
Tor 512
服务器 x 40
服务器 x 40
服务器 x 40
服务器 x 40

SRP水平分割原则

- 所有流量必须遵循最短路径（忽略所有其他较长路径）：
  - DC内：接入 --> 核心 --> 接入（2跳）-- 蓝色路径
  - DC间：接入 --> 核心 --> border（3跳）-- 绿色路径
  - 被禁止的路径：接入 --> 核心 --> 接入 --> 核心 --> 接入 (4跳) – 红色路径



SRP 实现框架


SRP控制器
根据身份下发
状态上报
路由管控
Topo Grid
Route Map
SRP
算法
链路通告、
邻居状态维护
生成、维护转发表
转发表
Server1
Server2

SRP Topo Grid示例

描述互联关系

行表示目的设备

列表示邻居

Devices	Core1	Core2	…	Core4
ToR1	*	*	*	*
ToR2	1	1	1	1
……				
Core1	1	*	*	*
Core2	*	1	*	*
……				
Border1	1	1	1	1
Border2	1	1	1	1

ToR1's Grid

1表示邻居为目的有效next-hop

Devices	Core1	Core2	…	Core4
ToR1	1	1	1	1
ToR2	1	1	1	1
……				
Core1	1	*	*	*
Core2	*	1	*	*
……				
Border1	*	*	*	*
Border2	*	*	*	*

Border1's Grid

*表示邻居永远为无效next-hop

Devices	ToR1	ToR2	…	ToR128	Border1	Border2
ToR1	1	*	*	*	*	*
ToR2	*	1	*	*	*	*
……						
Core1	*	*	*	*	*	*
Core2	*	*	*	*	*	*
……						
Border1	*	*	*	*	1	*
Border2	*	*	*	*	*	1

Core1's Grid

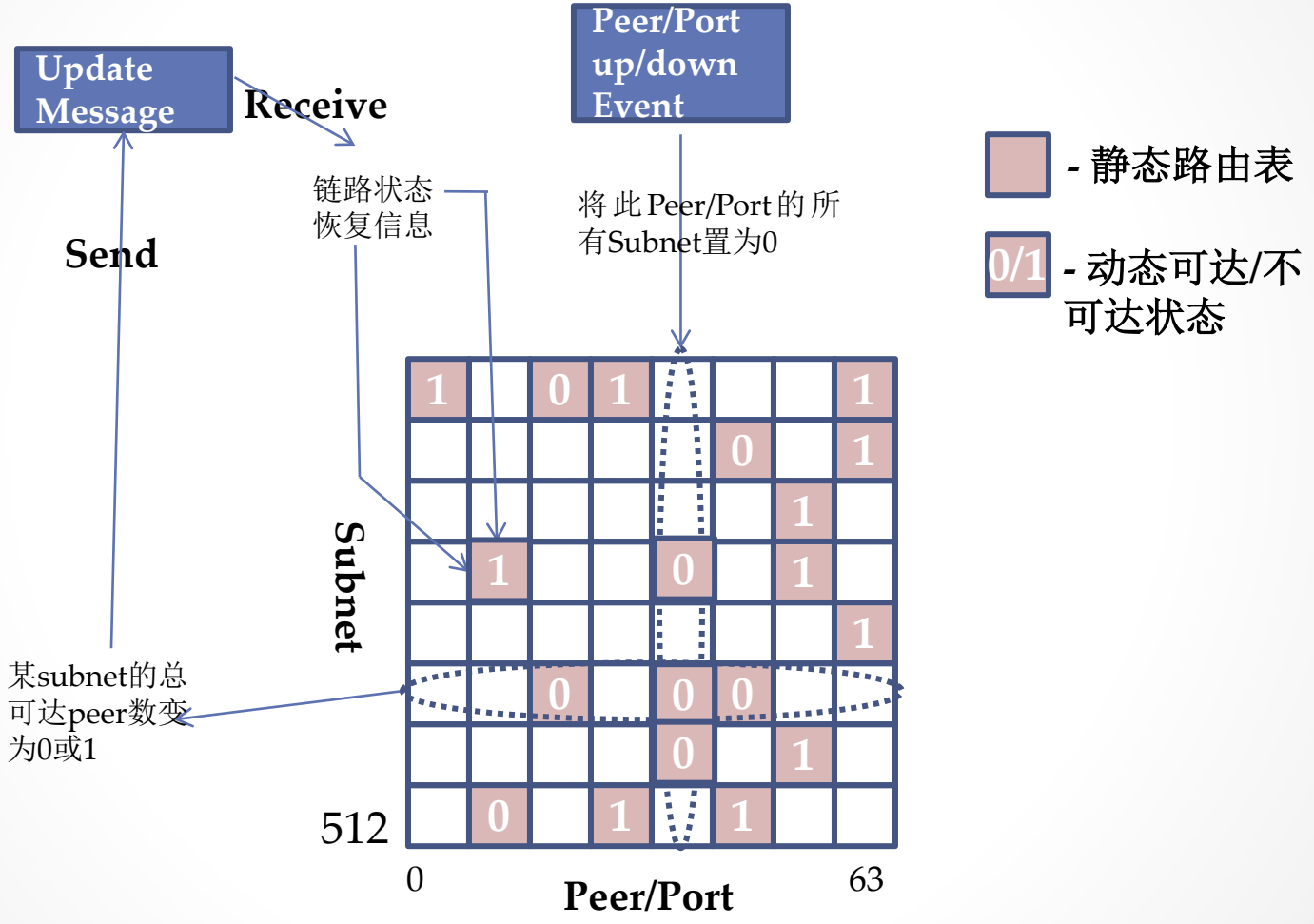
SRP Route Map示例

实现路由管控

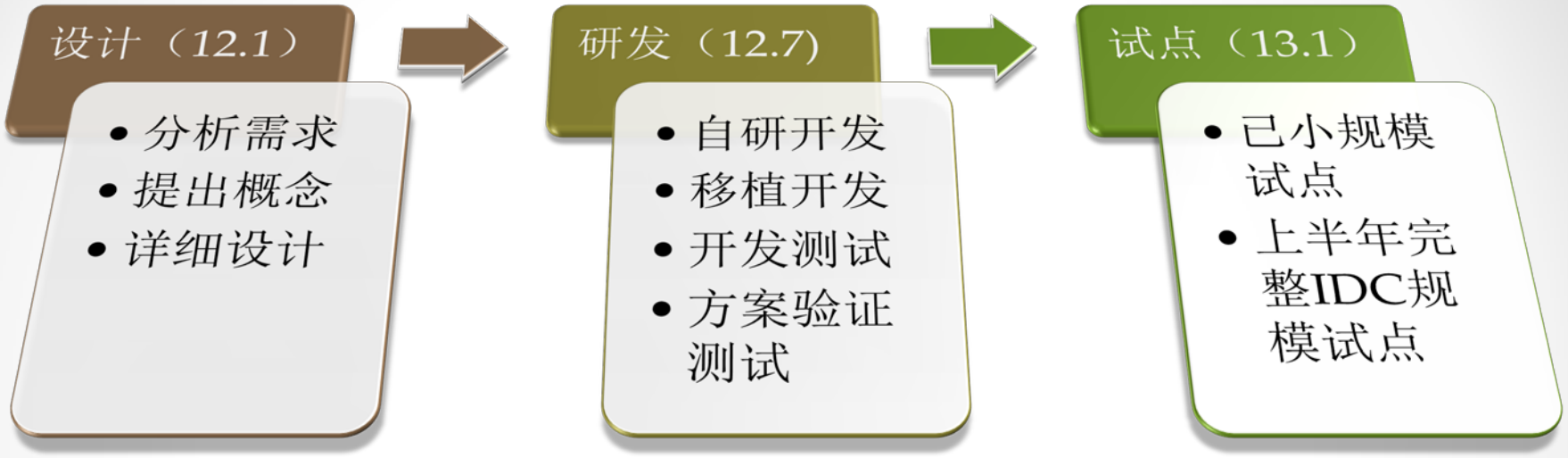
根据设备身份配置，确定采用哪一列设置
每一行确定对每条路由的学习策略

Subnet	From	Description	Status	ToR1'P	ToR2'P	Core1'P	Core4'P	Border1'P
10.1.1.0/26	ToR1	Production	有效	Direct	Learn	Learn	Learn	Learn
10.1.33.0/26	ToR1	Server ILO	失效	Direct	Black hole	Learn	Learn	Learn
10.1.0.41/32	ToR1	Loopback	有效	Direct	Learn	Learn	Learn	Learn
10.255.254.0/24	ToR1	Deployment	有效	Direct	Black hole	Learn	Learn	Black hole
10.1.1.0/26	ToR2	Production	失效	Learn	Direct	Learn	Learn	Learn
10.1.33.0/26	ToR2	Server ILO	有效	Black hole	Direct	Learn	Learn	Learn
10.1.0.41/32	ToR2	Loopback	有效	Learn	Direct	Learn	Learn	Learn
10.255.254.0/24	ToR2	Deployment	有效	Black hole	Direct	Learn	Learn	Black hole
10.1.0.1/32	Core1	Loopback	有效	Learn	Learn	Direct	Black hole	Learn
10.0.0.0/8	Border 1	Away from IDC	有效	Learn	Learn	Learn	Learn	Redistribute
10.0.0.0/8	Border 2	Away from IDC	失效	Learn	Learn	Learn	Learn	Redistribute

SRP 算法纵览


Update Message
Receive
Peer/Port up/down Event
链路状态恢复信息
将此Peer/Port的所有Subnet置为0
Send
某subnet的总可达peer数变为0或1
Subnet
512
0
Peer/Port
63
- 静态路由表
0/1
- 动态可达/不可达状态

SRP 已经上线运行



	OSPF	BGP	SRP
IDC内收敛时间	> 2s	< 1s	< 1s
规模敏感性	规模增长性能恶化	不敏感	不敏感
实现代码	50万级别	30万行级别	7000行
功能扩展	管控程度低	极限：分布式调度	集中式调度

SRP 愿意和各位一起完善

六项相关专利介绍设计思路！

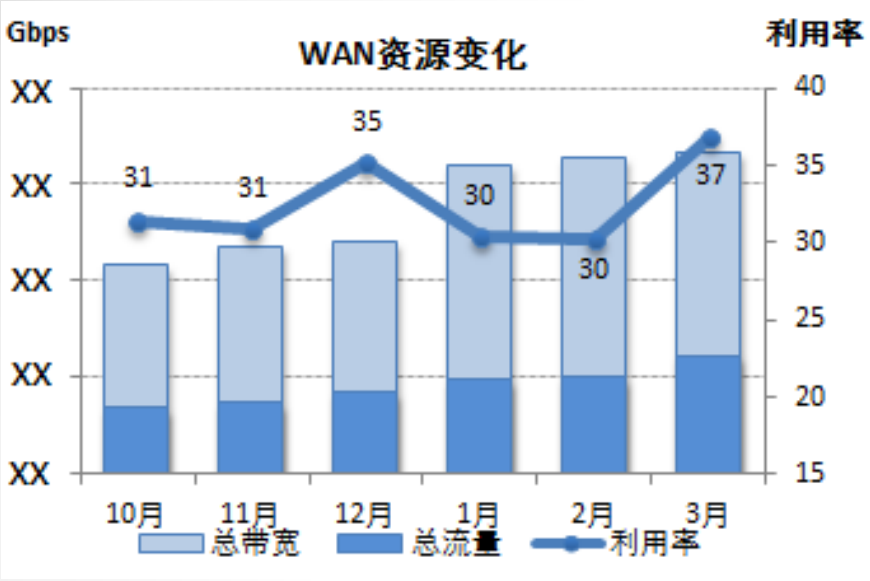
多种交换机平台上代码实现！

完整的SRP白皮书欢迎了解！

广域网流量调度

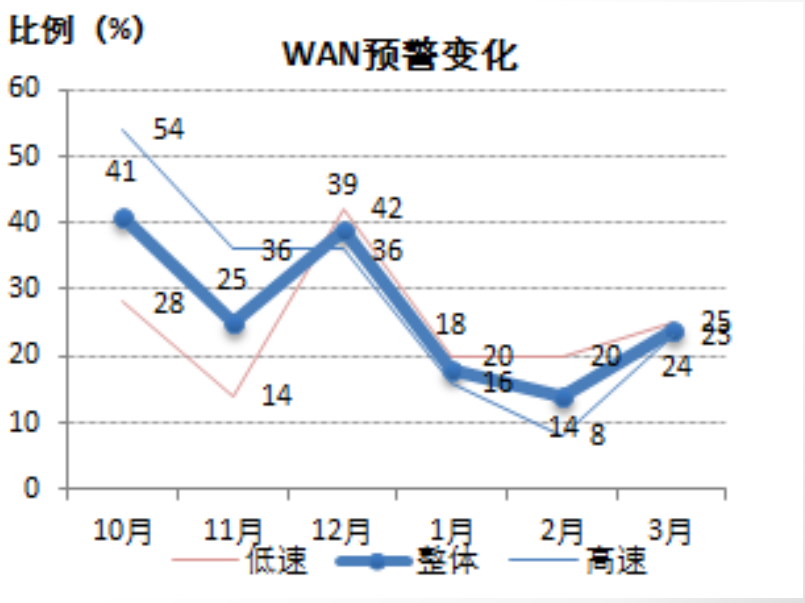
CarrierBackboneNetwork Controller

专线容量管理窘境



专线带宽利用率低

成本压力很大

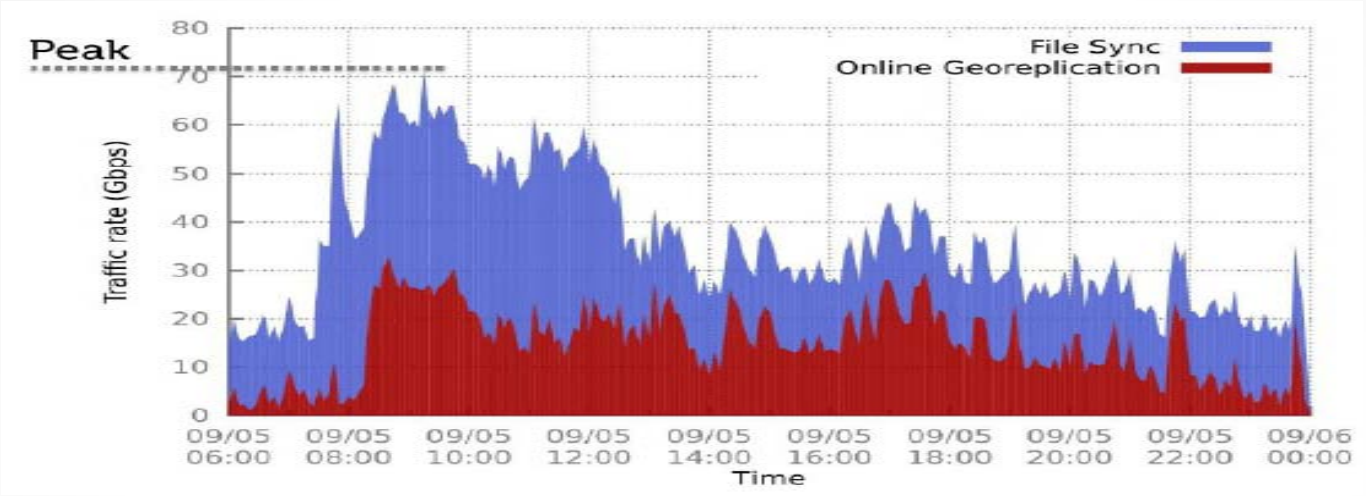


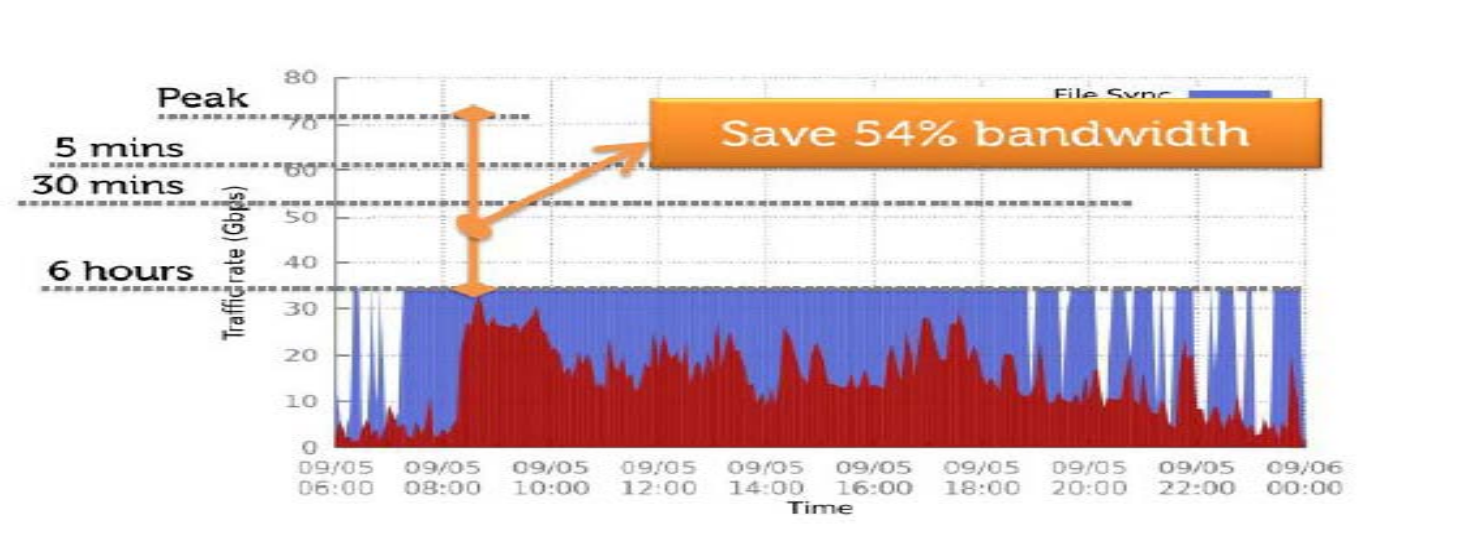
专线容量预警多

BG用户很不爽

更窘迫的是突发流量

Microsoft在其骨干网同样碰到了链路利用不足的问题



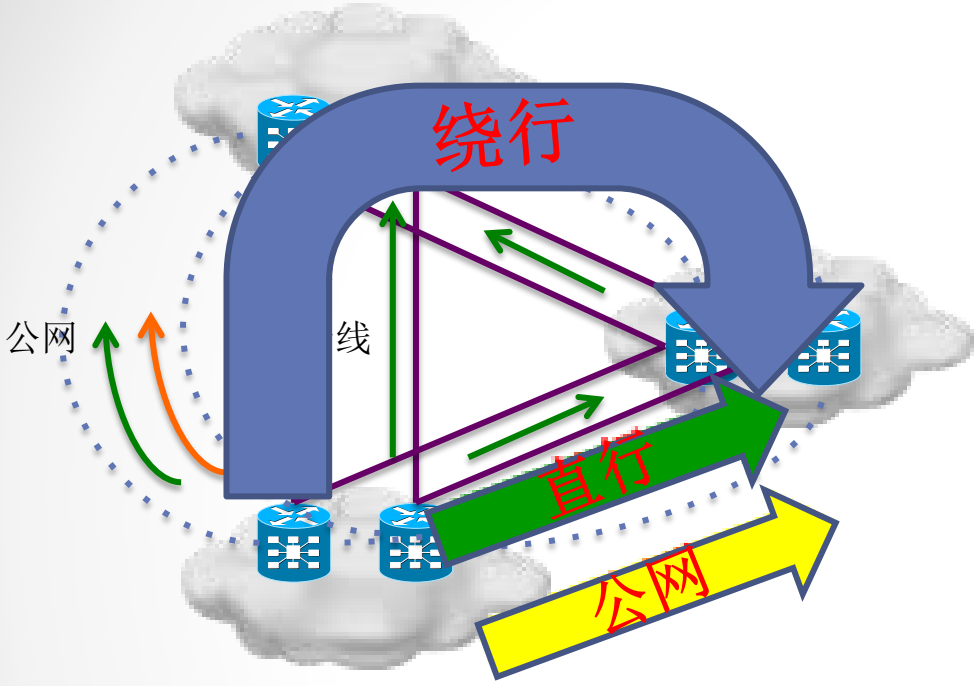


原因在哪里？

传统网络架构类似于每个路口的交警们。根据路口情况独立做出判断指挥交通。分布式，健壮性很好。但缺乏全局视图，无法解决所有人可能都走一条路，或突发的问题；

提前规划、区分流量、智能分流很关键

建立多条通路



流量原则：

①先保障承诺流量优先、最优路径传输

②非保障流量在拥塞时，先切换到次优路径

选路原则：

①有带宽资源的最优专线

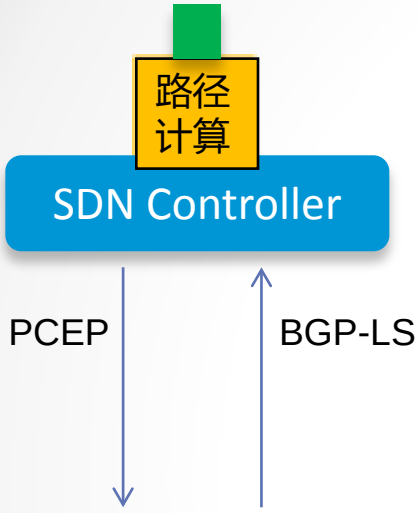
②有带宽资源的次优专线

③有带宽资源的最优公网传输

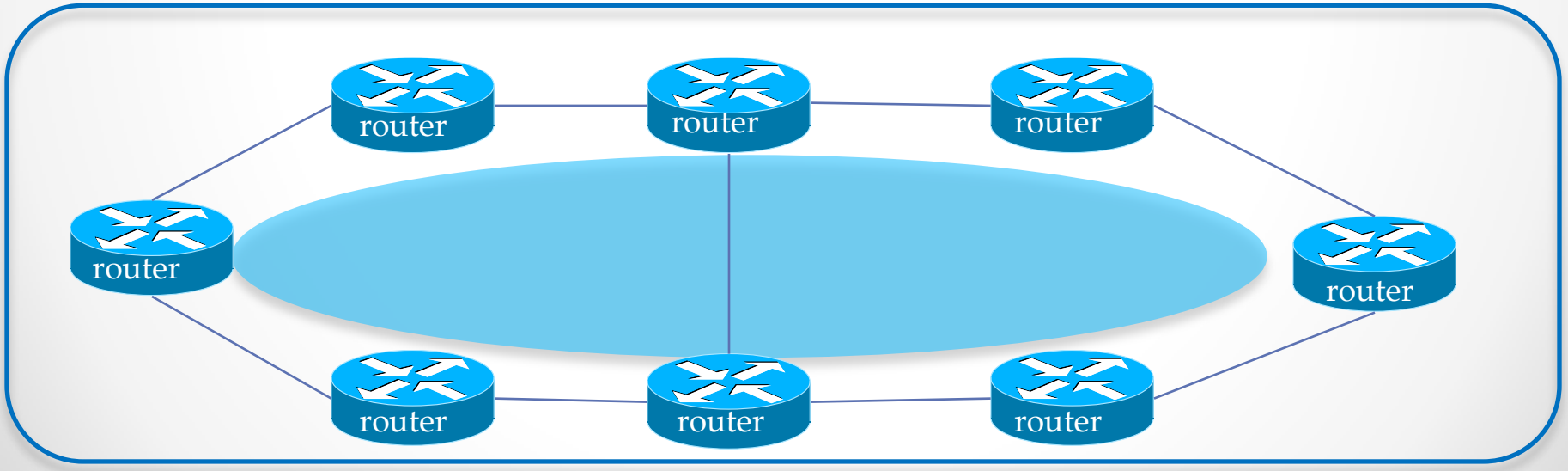
④有带宽资源的次优公网传输

- 利用auto-te技术解决分布式计算带宽
- 实现流量根据资源池资源按需调度
- 更强的容灾架构
- 提高专线利用率，降低运营成本

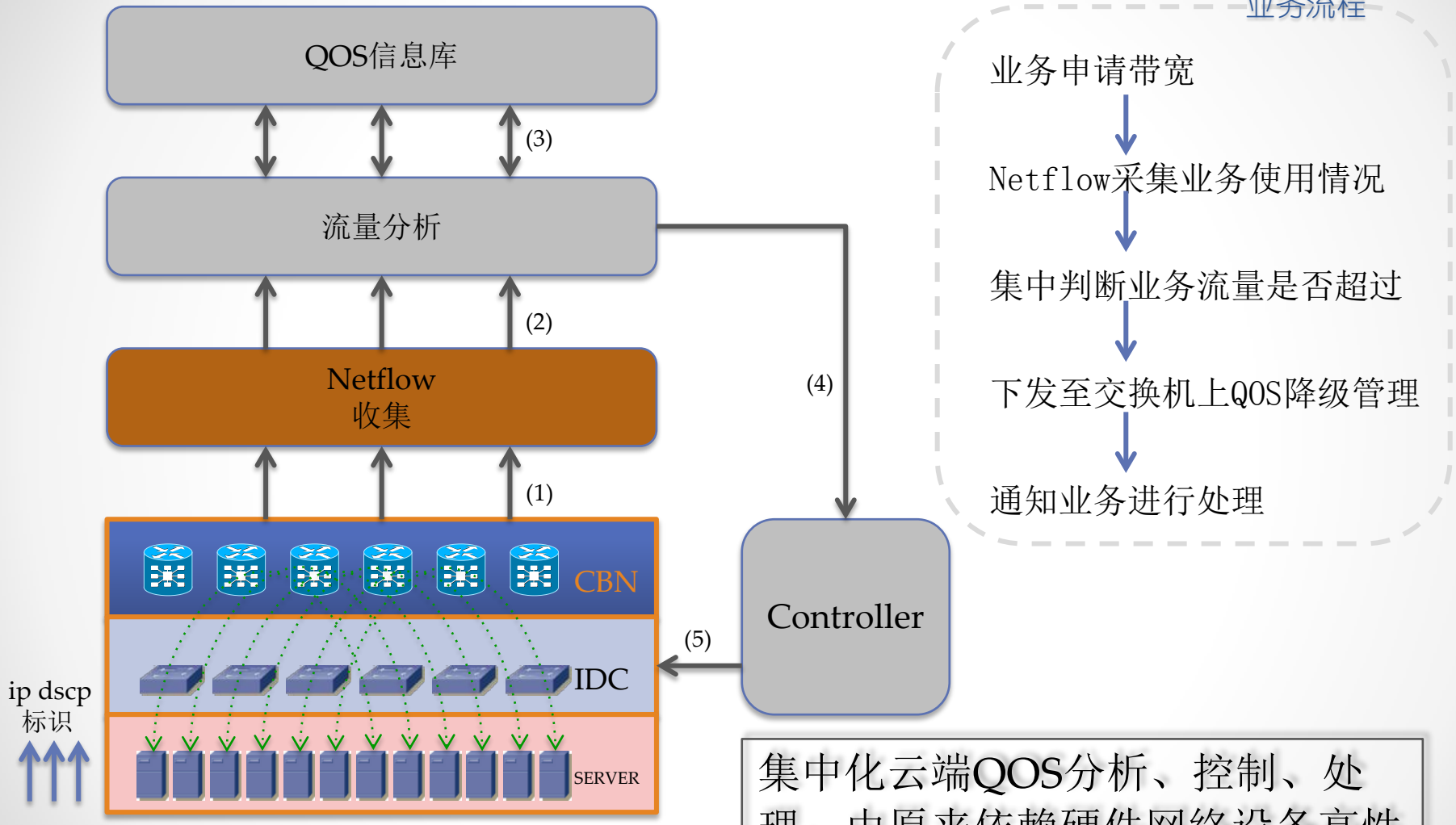
集中式路径计算与调度



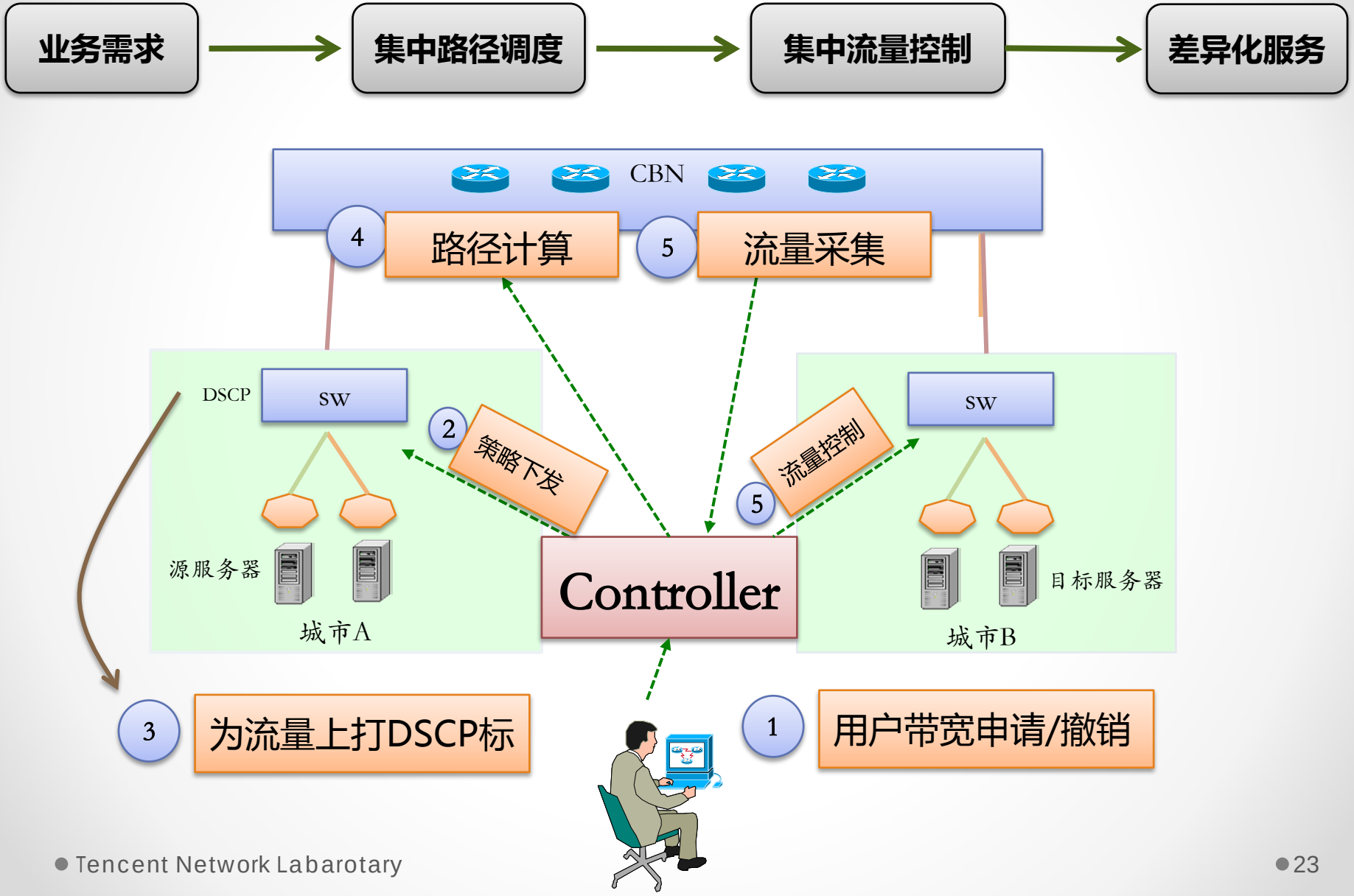
- SDN Controller与路由器之间运行BGP-LS/SNMP协议，路由器将通过IGP搜集的网络拓扑和带宽情况通过BGP-LS传递给SDN Controller，每个域内只需一台或两台路由器（通常为ABR）与Controller建立邻居关系即可，解决了系统扩张的压力；
- SDN Controller基于BGP-LS/SNMP传递的信息完成路径选择（Path Computation），并通过PCEP协议将结果下发给路由器
- 集中式路径计算避免分布式模式下的带宽利用不合理



集中式业务流量规划及QOS管理



差异化广域网流量服务（路径计算+流量控制）



Tencent Network Laboratory愿与您探讨&分享：

大规模IDC网络架构设计
广域网流量调度
虚拟化网络
智能网络管理
更多SDN内容

Wechat ID: sagezou
QQ: 335957600
Mail: sagezou@tencent.com
Tencent Weibo：腾讯云数据中心