



BEIJING 2017

AWS数据中心与VPC揭秘

余骏 Ivan Yu



促进软件开发领域知识与创新的传播



关注InfoQ官方信息
及时获取QCon软件开发者
大会演讲视频信息



扫码，获取限时优惠



全球架构师峰会 2017 [深圳站]

2017年7月7-8日 深圳·华侨城洲际酒店

咨询热线: 010-89880682

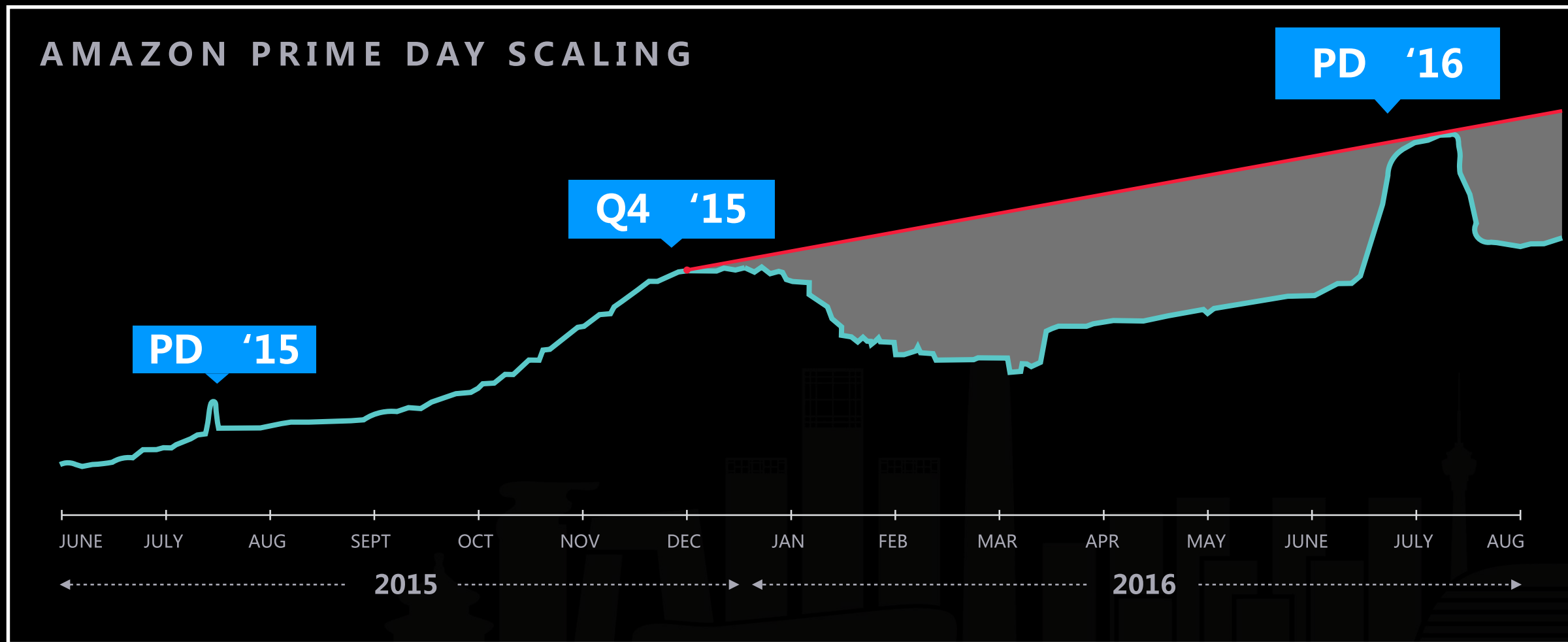


全球软件开发大会 [上海站]

2017年10月19-21日

咨询热线: 010-64738142

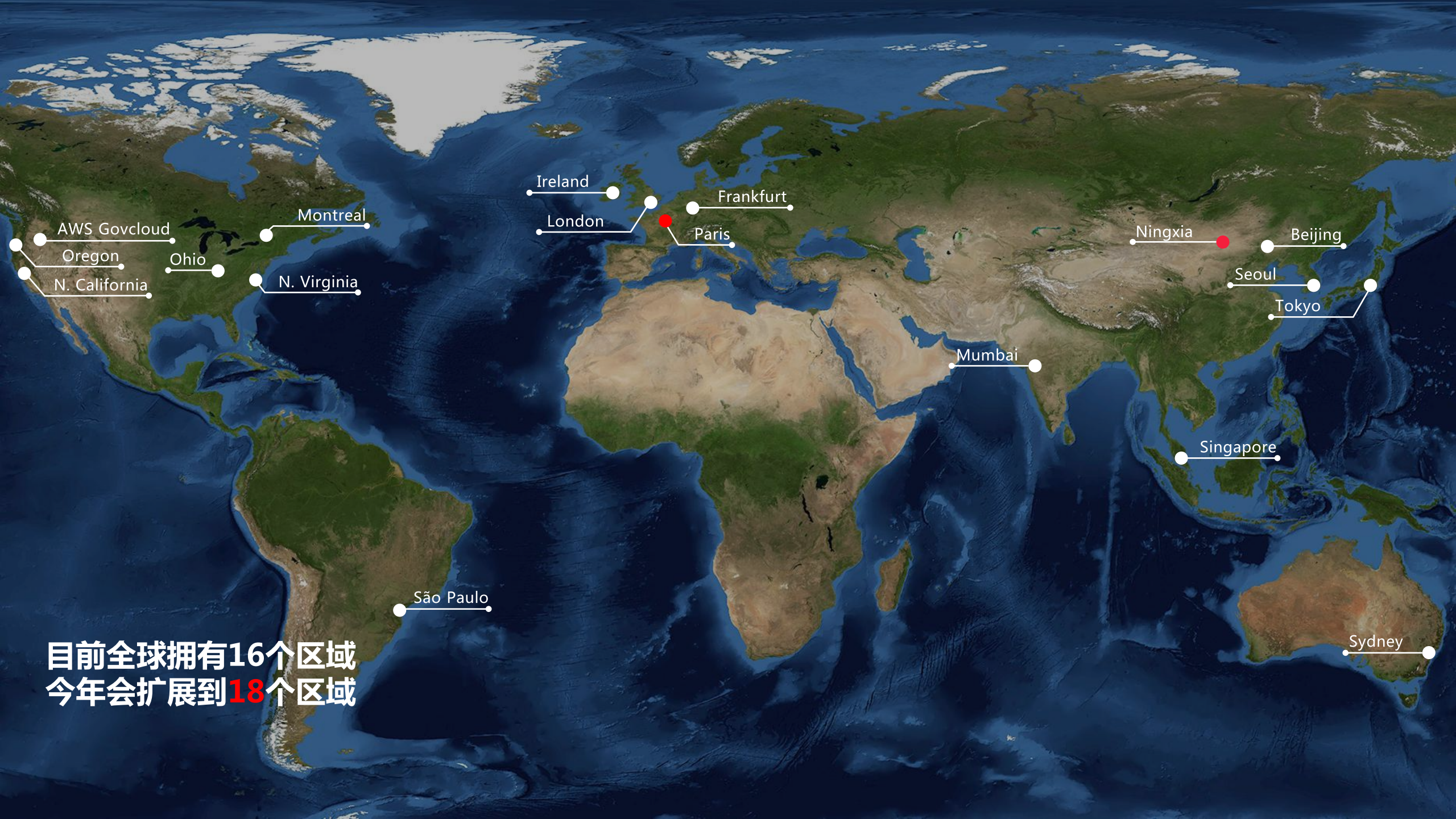
弹性是新常态





In 2015 AWS Deployed Almost
**ENOUGH SERVER CAPACITY
EVERY DAY TO SUPPORT
AMAZON IN 2005**
When it was an \$8.49B Enterprise

AWS adds the capacity equivalent of a FORTUNE 500 Enterprise daily



AWS Govcloud
Oregon
N. California

Montreal
Ohio
N. Virginia

Ireland
London
Frankfurt
Paris

Mumbai

Ningxia

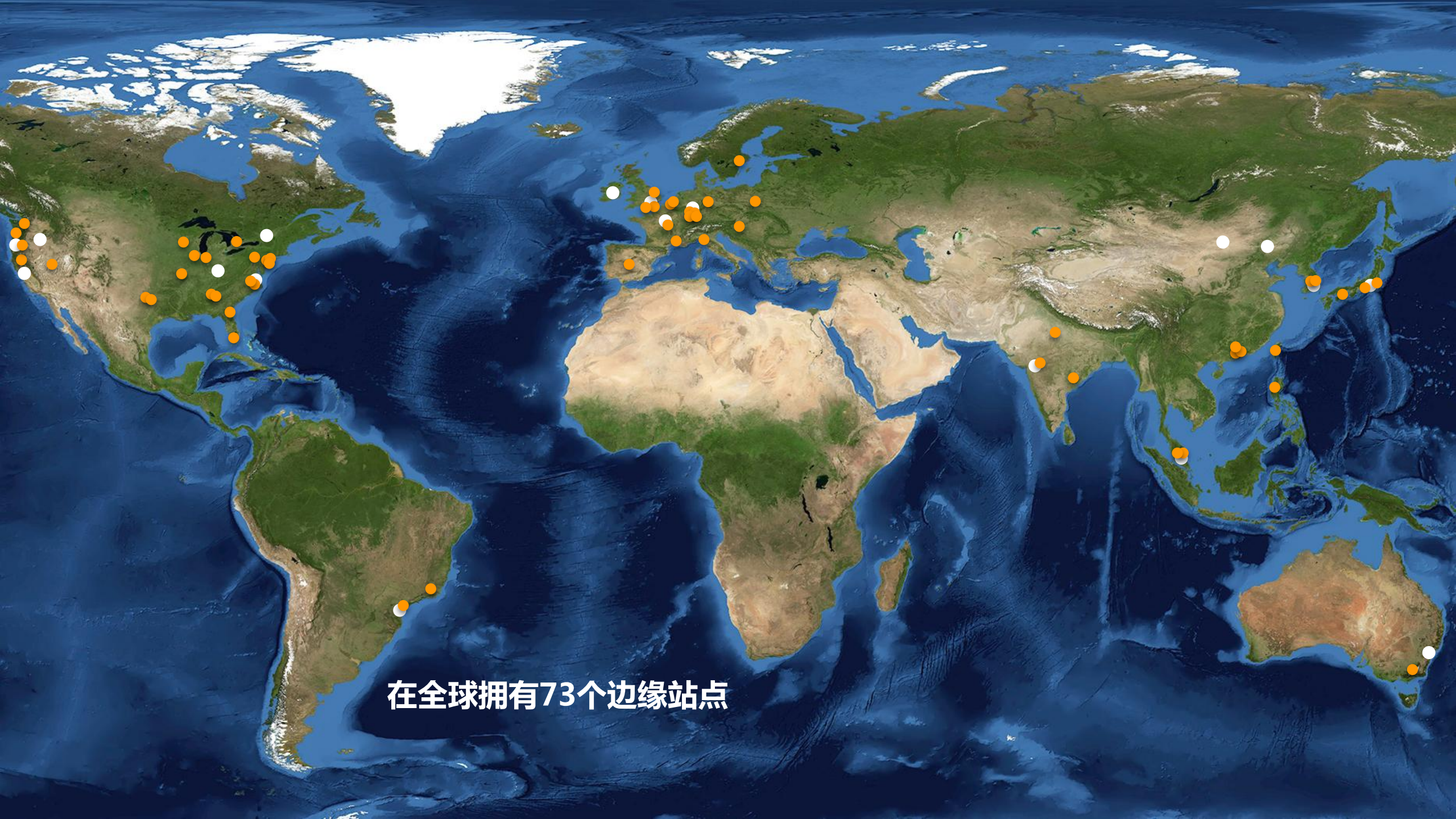
Beijing
Seoul
Tokyo

Singapore

São Paulo

Sydney

目前全球拥有16个区域
今年会扩展到18个区域

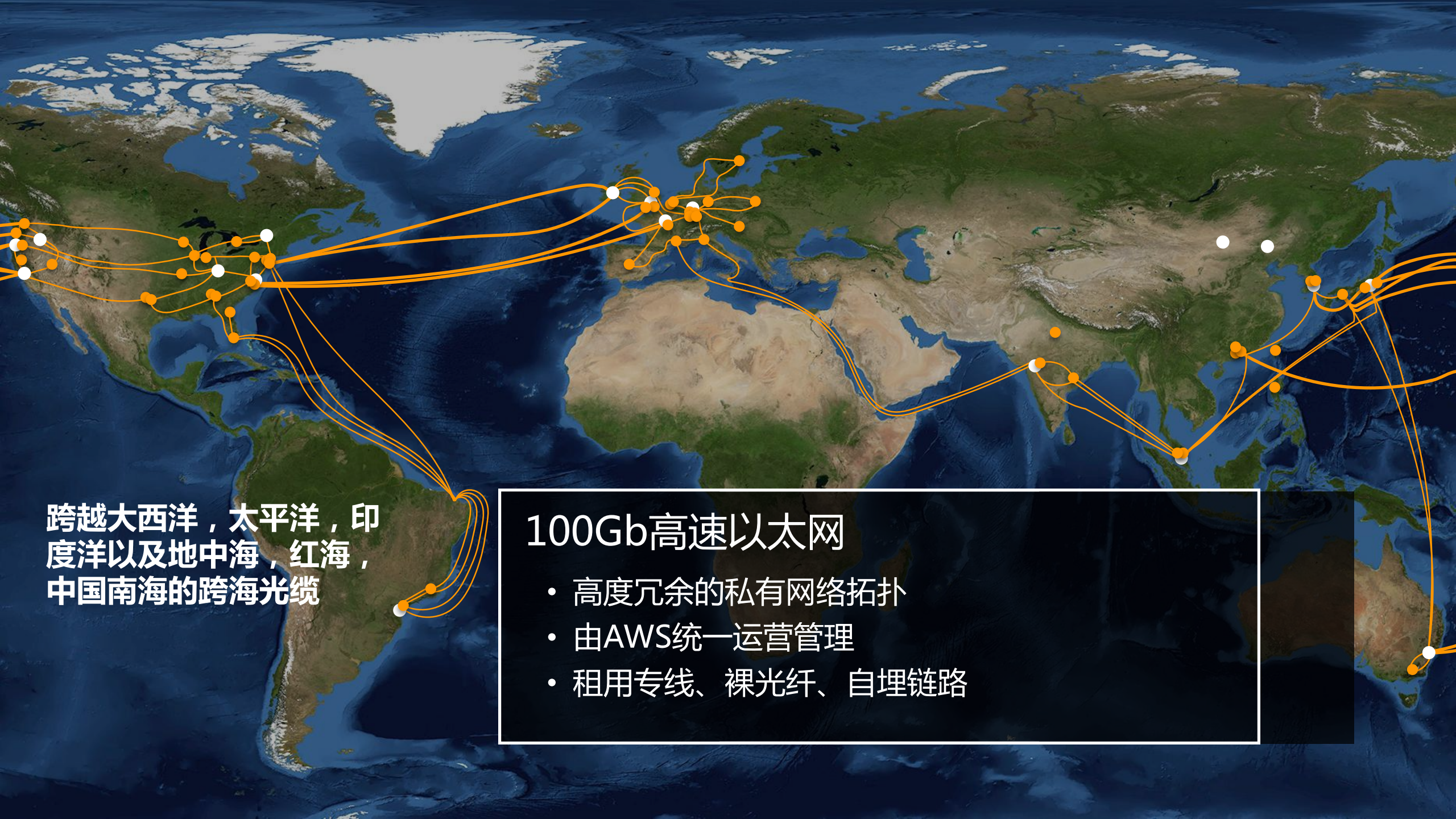


在全球拥有73个边缘站点



AWS全球网络基础设施

- 改善延迟、丢包率及总体网络质量
- 避免网络互联的带宽瓶颈
- 统一管理控制，快速响应



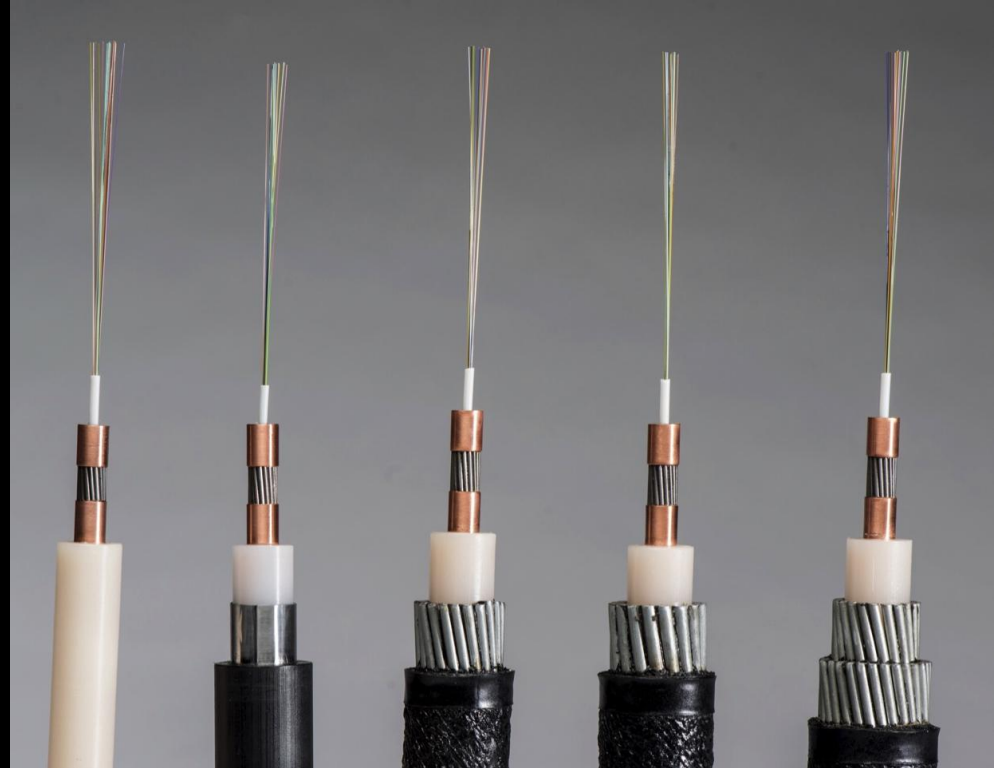
跨越大西洋，太平洋，印度洋以及地中海，红海，中国南海的跨海光缆

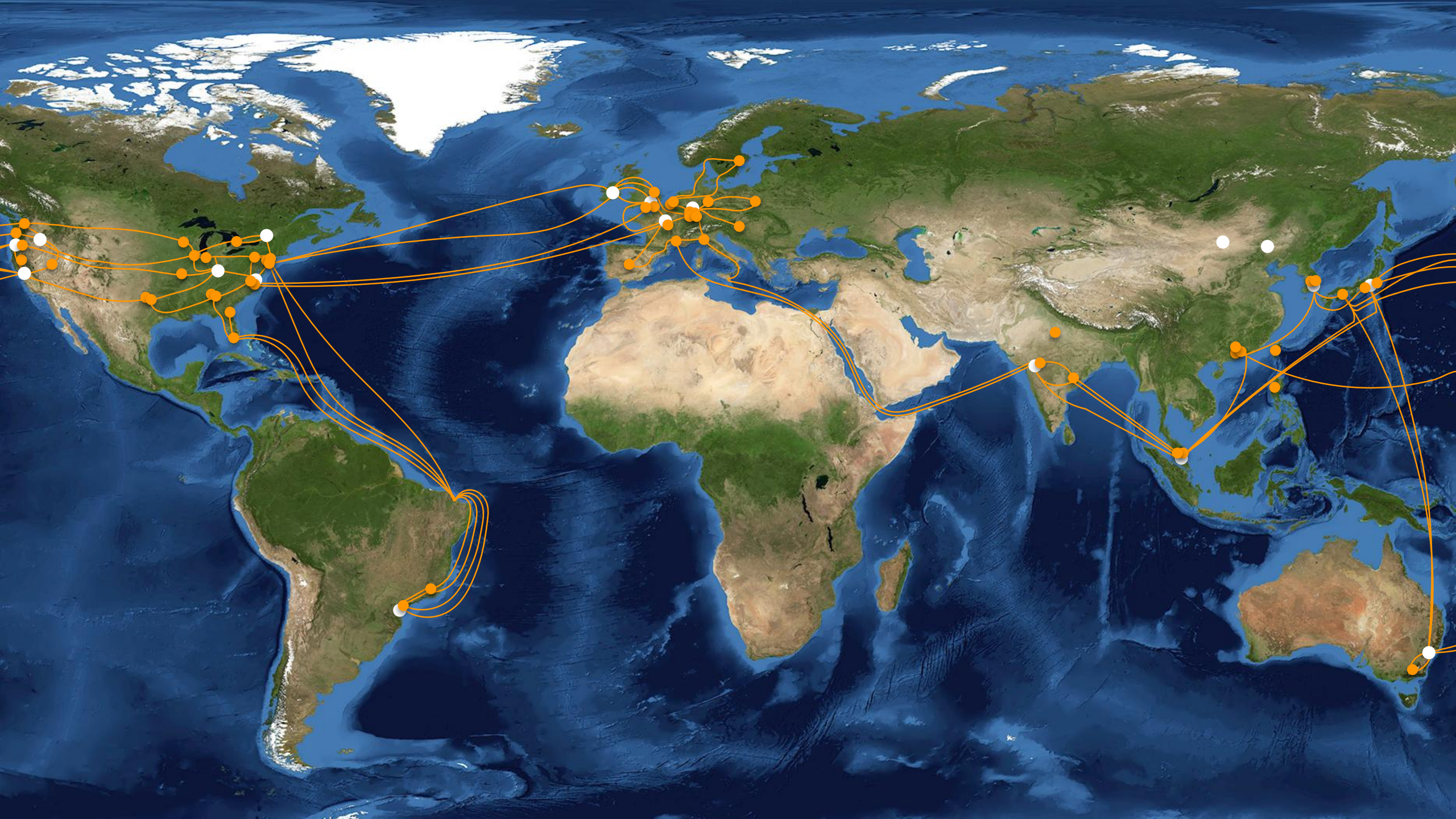
100Gb高速以太网

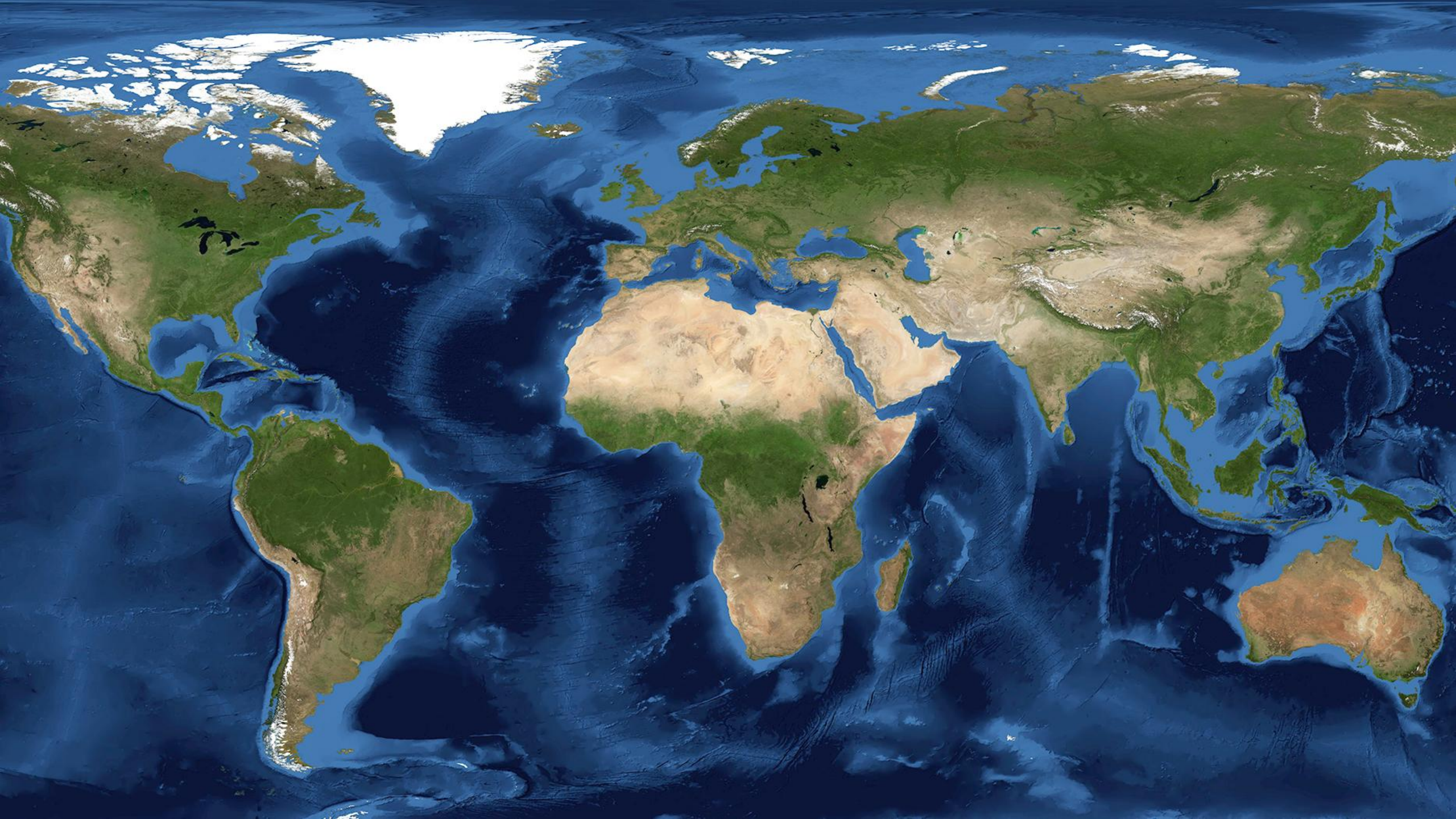
- 高度冗余的私有网络拓扑
- 由AWS统一运营管理
- 租用专线、裸光纤、自埋链路

最新项目

- 跨太平洋海底光缆
- 总长14,000km，连接澳大利亚，新西兰，夏威夷和俄勒冈
- 3对光纤
- 100 waves @ 100G
- 新西兰海岸破土工程已于去年启动

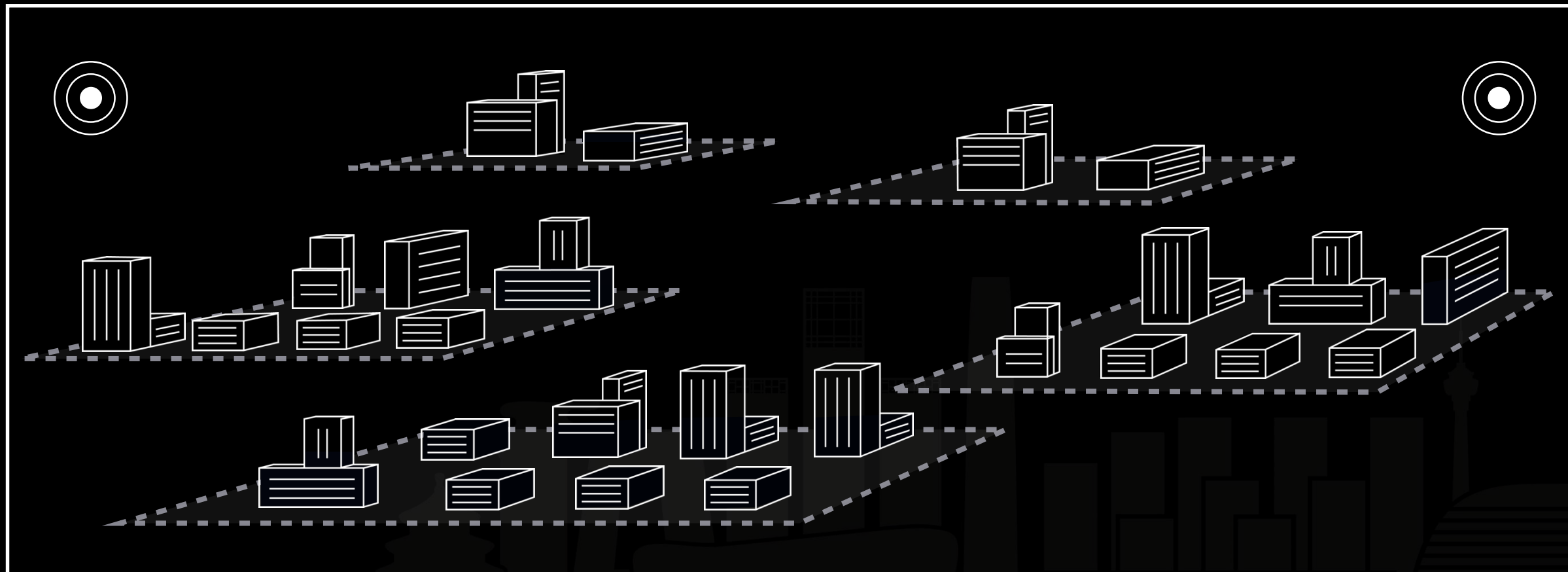




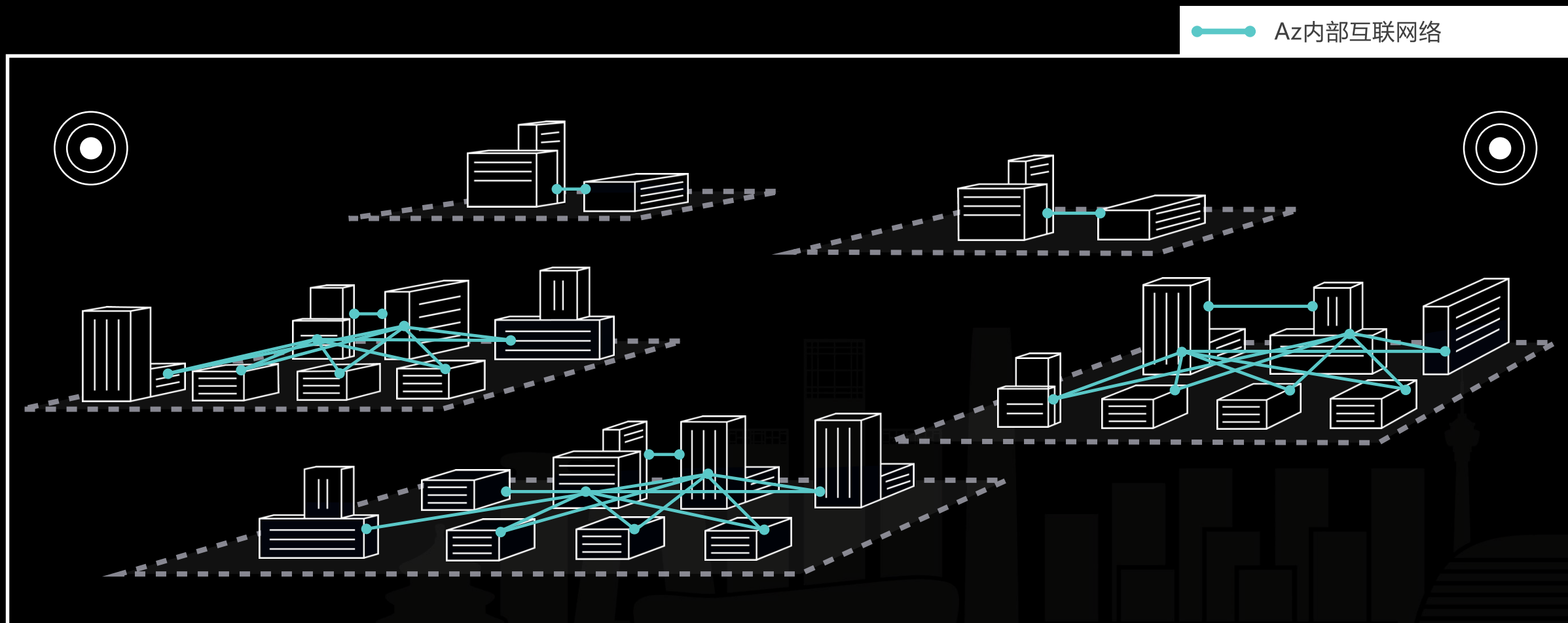


转接中心

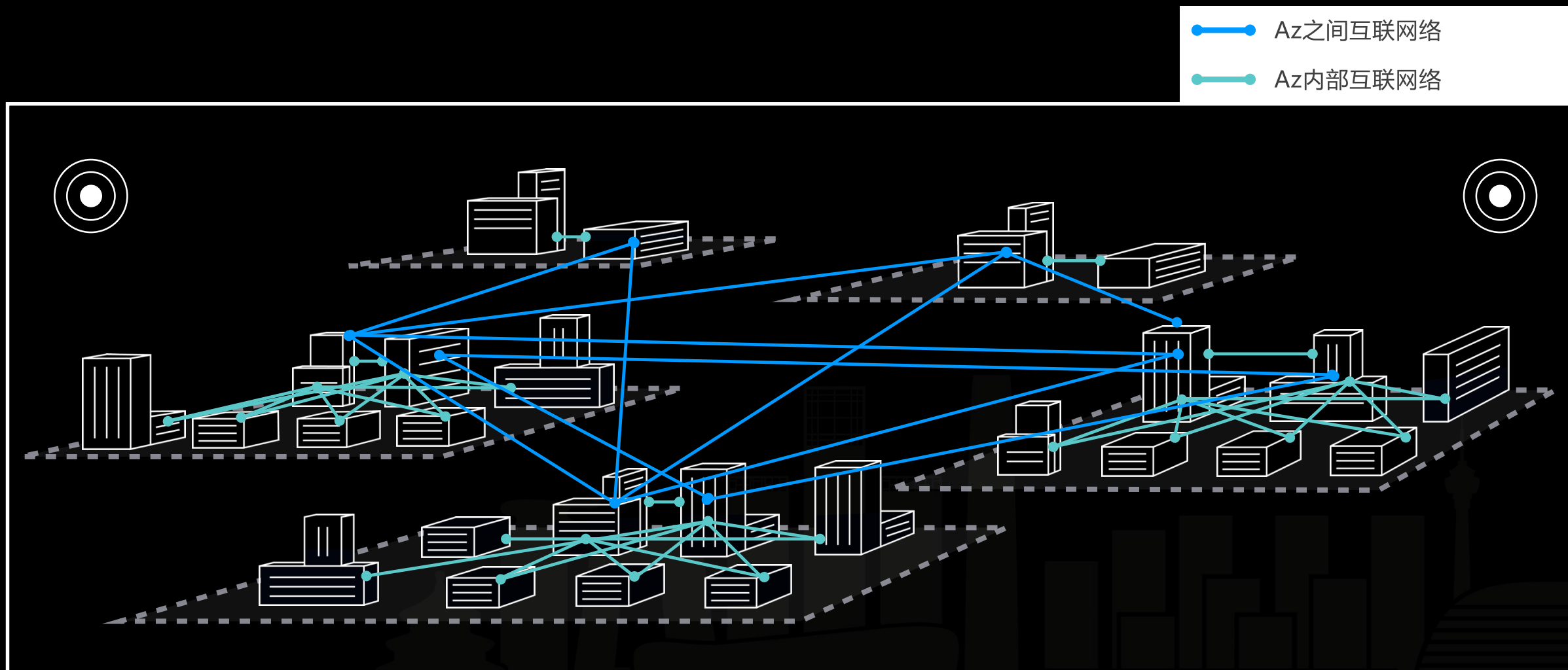
- 每个区域拥有2个冗余的转接中心
- 高速互联的基础设施



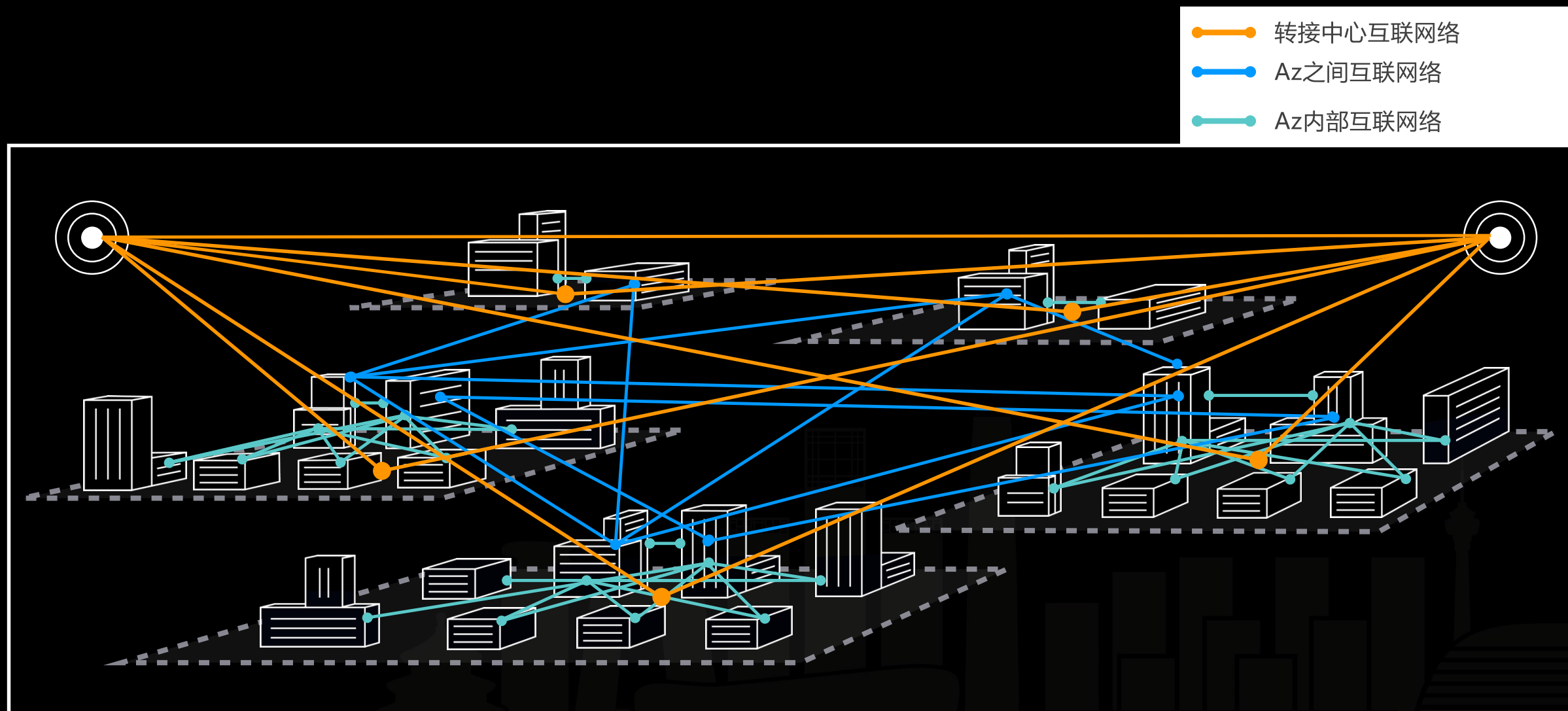
城域光纤网络



城域光纤网络



城域光纤网络



高度可扩展的可用区设计

- 每个可用区包含1个以上的数据中心
- 某些可用区包含多达8个数据中心
- 数据中心之间通过冗余的光纤互联
- 部分可用区拥有超过30W台规模的服务器



数据中心设计

- 很容易实现60–120MW或者更大规模的数据中心
- 更大规模的数据中心会带来过大的故障域



AWS定制化路由器

- 市场上销售的路由器
 - 功能复杂，可靠性低
 - 售价昂贵
 - 故障修复缓慢（6个月）
- AWS定制化路由器
 - 硬件基于AWS标准制造
 - 软件协议由AWS定制开发
- 设计之初便遵循25GbE路线
 - 业界处于10GbE & 40GbE
 - 光模块供应短缺
- 40GbE实际上由4对10GbE光纤构成
- 50GbE (2x 25GbE)拥有比40GbE更低的成本及更高的速度



AWS定制化路由器

- 市场上销售的路由器
 - 功能复杂，可靠性低
 - 售价昂贵
 - 故障修复缓慢（6个月）
- AWS定制化路由器
 - 硬件基于AWS标准制造
 - 软件协议由AWS定制开发
- 设计之初便遵循25GbE路线
 - 业界处于10GbE & 40GbE
 - 光模块供应短缺
 - 40GbE实际上由4对10GbE光纤构成
 - 50GbE (2x 25GbE)拥有比40GbE更低的成本及更高的速度



AWS定制化路由器

- 市场上销售的路由器
 - 功能复杂，可靠性低
 - 售价昂贵
 - 故障修复缓慢（6个月）
- AWS定制化路由器
 - 硬件基于AWS标准制造
 - 软件协议由AWS定制开发
- 设计之初便遵循25GbE路线
 - 业界处于10GbE & 40GbE
 - 光模块供应短缺
- 40GbE实际上由4对10GbE光纤构成
- 50GbE (2x 25GbE)拥有比40GbE更低的成本及更高的速度



AWS定制化路由器

- 基于Broadcom Tomahawk芯片
 - 70亿个晶体管
 - 128个25GbE端口
 - 1RU, 22lbs, <310W
- Amazon Annapurna ASIC
 - 支持第二代增强型网络
 - AWS控制芯片、硬件和软件
 - 更快地迭代创新
- EC2实例的最大带宽高达20GbE
 - 小机型的最大带宽可以达到10GbE



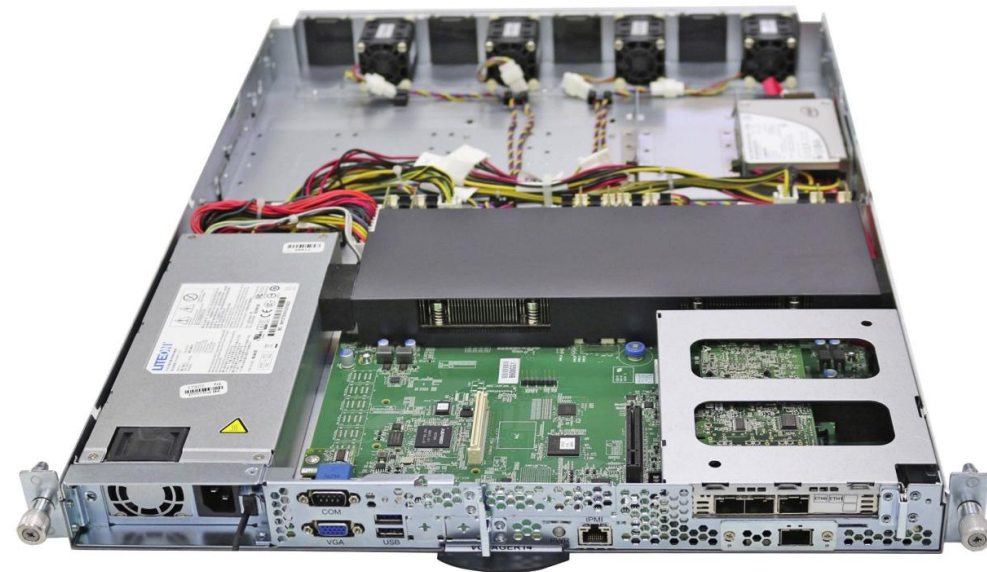
AWS定制化存储

- 2014: 880块磁盘/机架
- 下一代设计:
 - 1,110块磁盘/机架
 - 最初设计容量8.8PB
(目前实际高达11PB)
 - 2,778 lbs of storage



AWS定制化服务器

- 1RU设计
- 散热及电源效率为导向
- PSU & VRD > 90%效率



VPC基本概念

分布式映射服务

Server 192.168.0.3

10.0.0.2

10.0.0.2

Server 192.168.0.4

10.0.0.4

10.0.0.5

⋮

Server 192.168.1.3

10.0.0.3

10.0.0.4

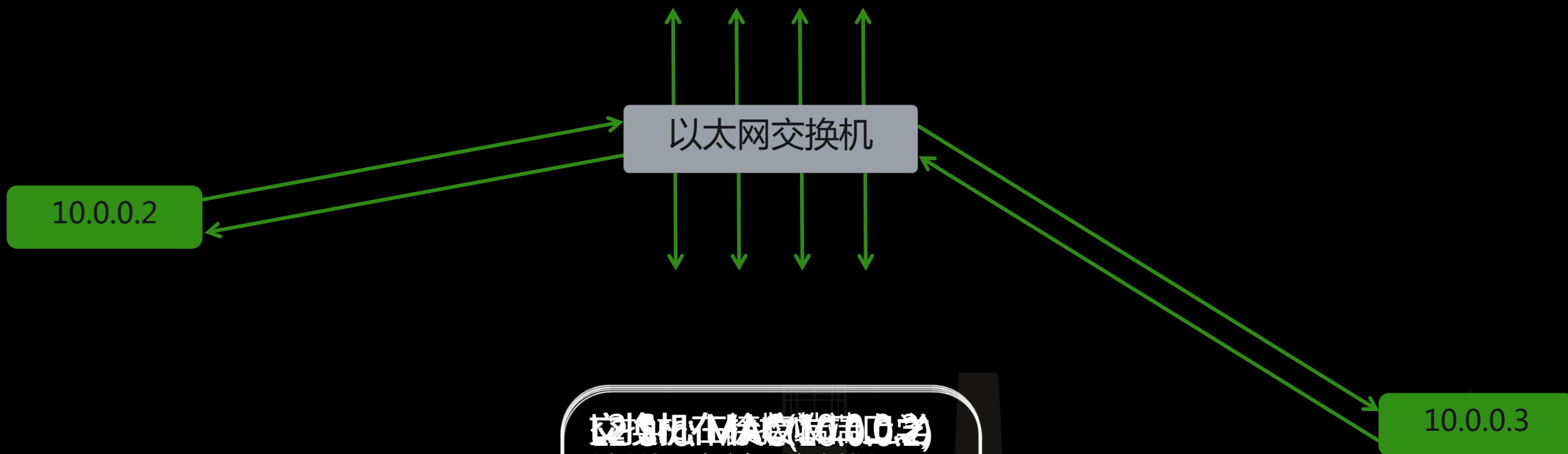
Server 192.168.1.4

10.0.0.3

⋮

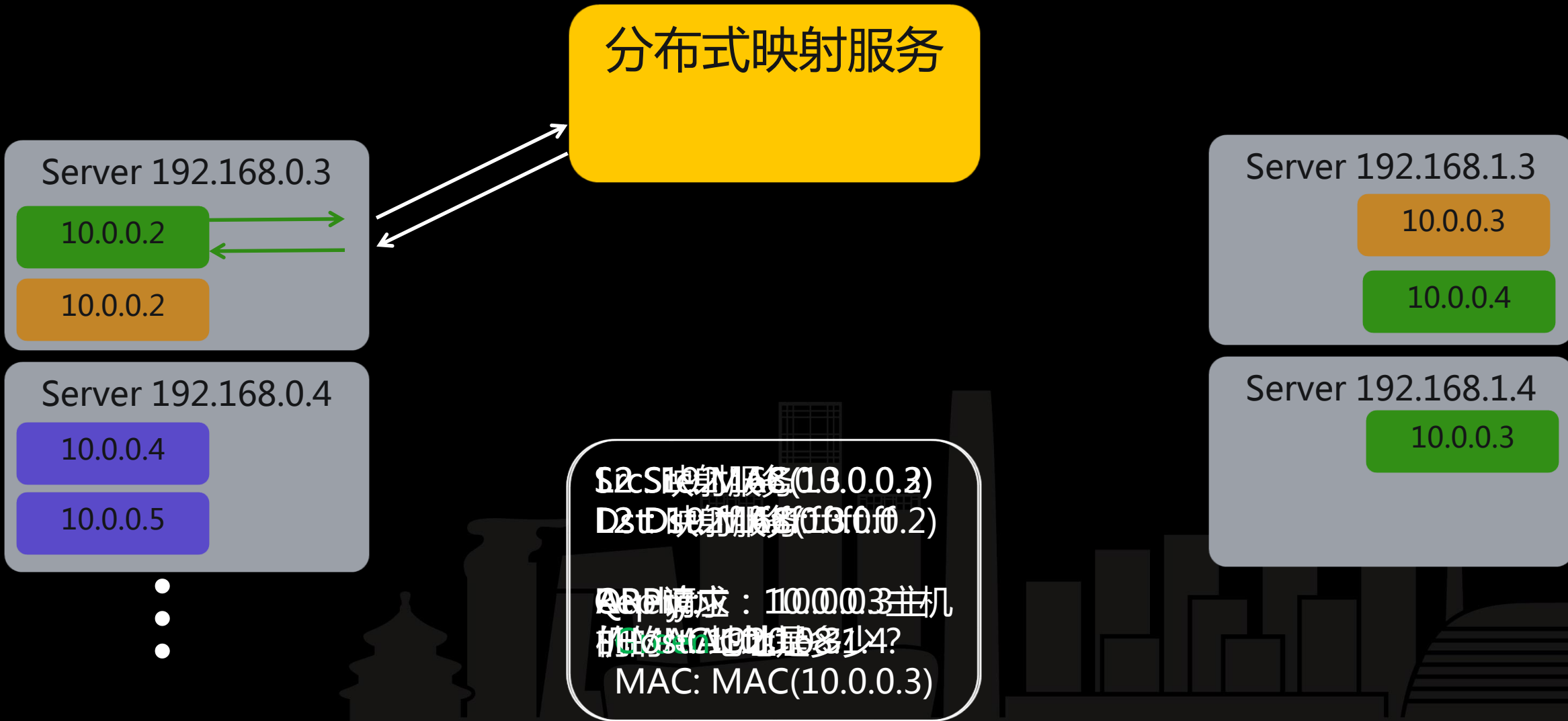
分布式映射服务：
用户请求通过负载均衡
路由策略映射到物理服
务器b3c4d

传统二层以太网的通信机制



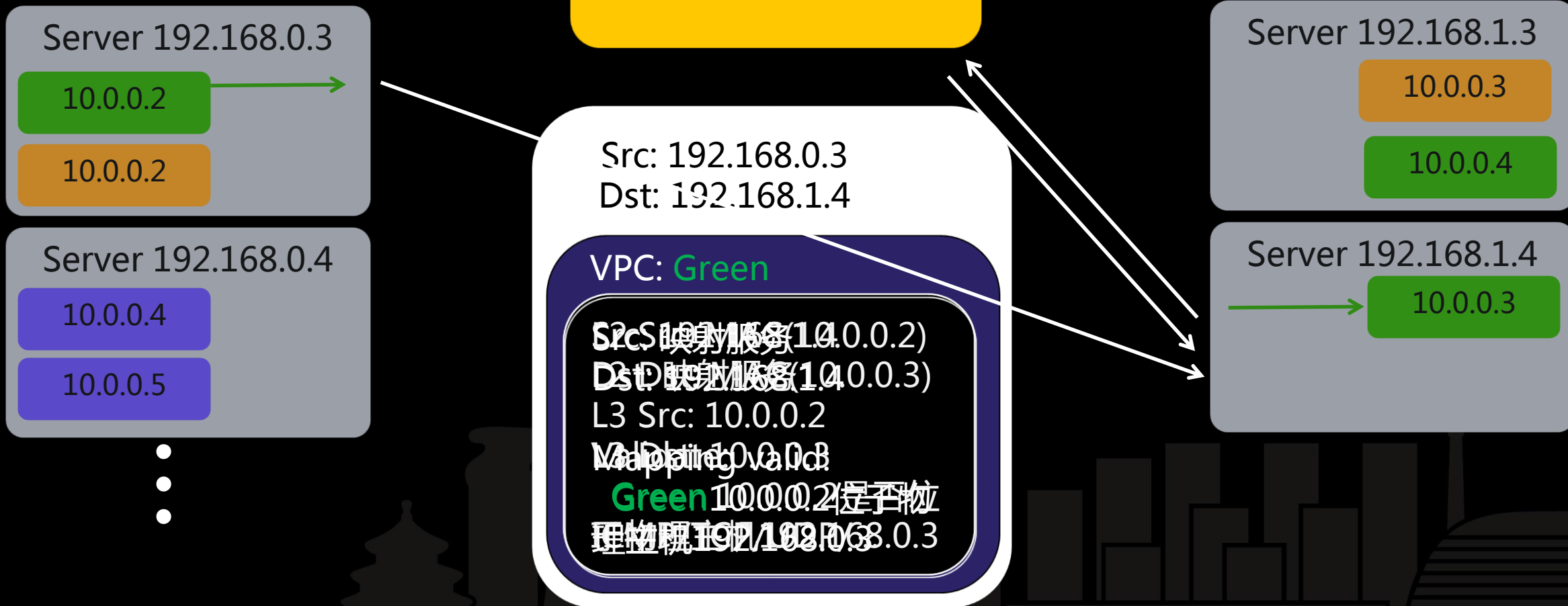
交换机在转发(10.0.0.2)数据包时,发现(10.0.0.3)请求地址0.0更新MAC表项,请问0.0000.3主机的MAC地址是多少?
ICMP/TCP/UDP/...

VPC中的二层通信机制

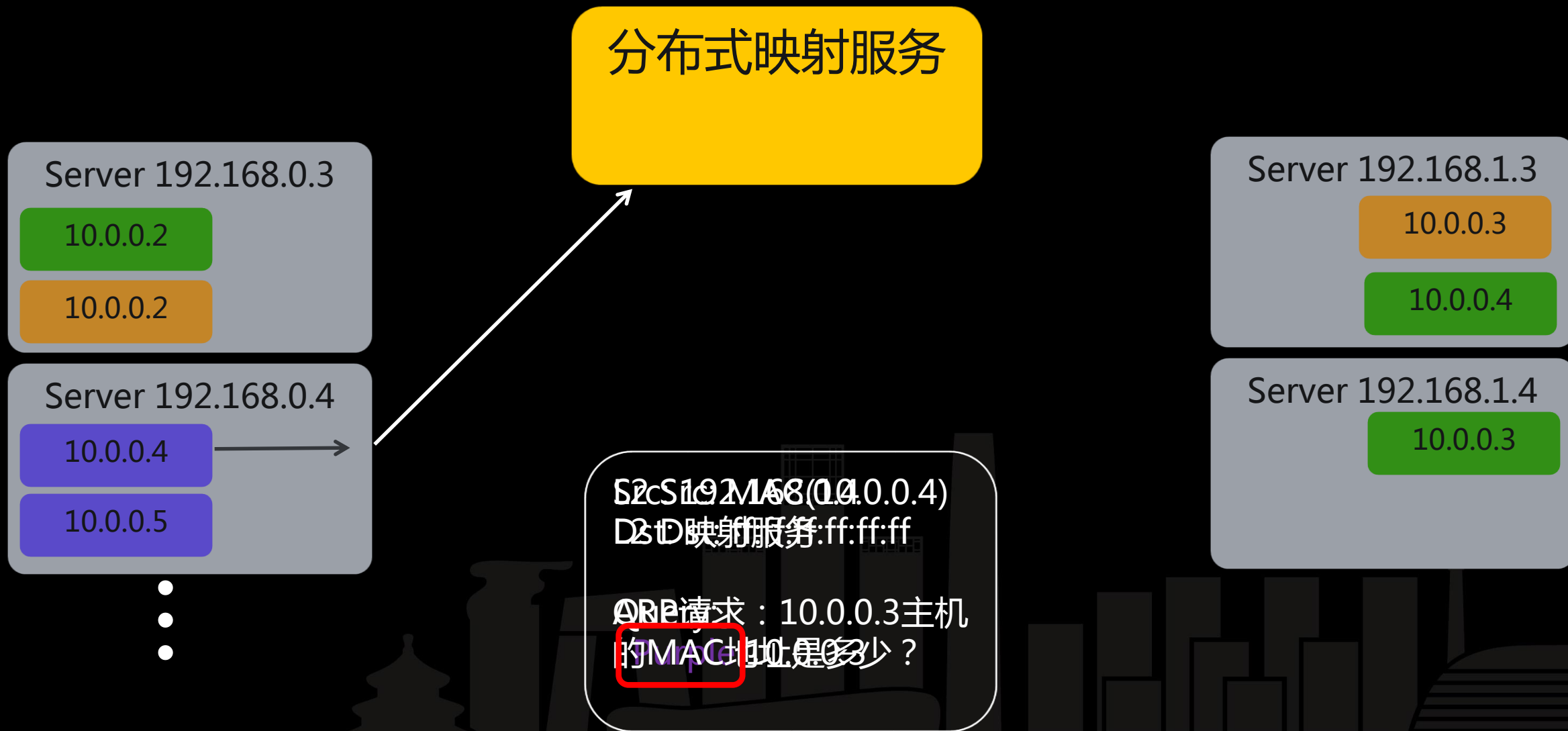


VPC中的二层通信机制

分布式映射服务

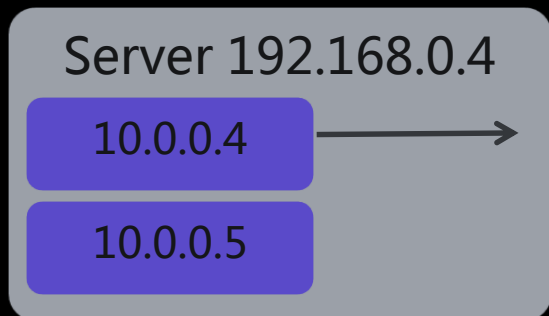
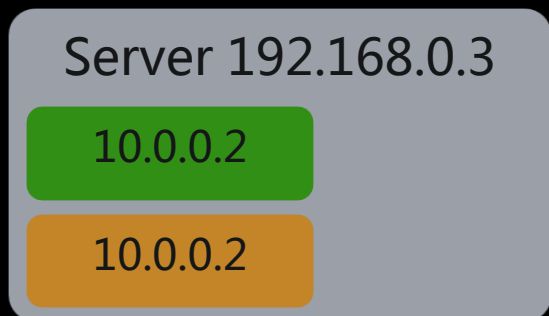


VPC隔离机制

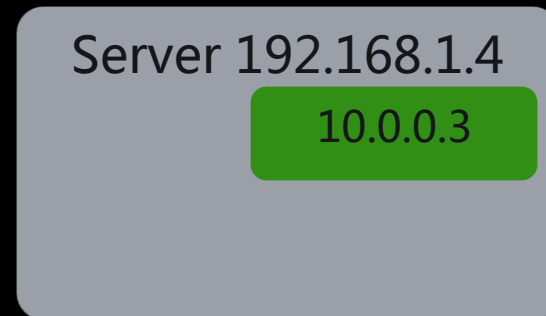
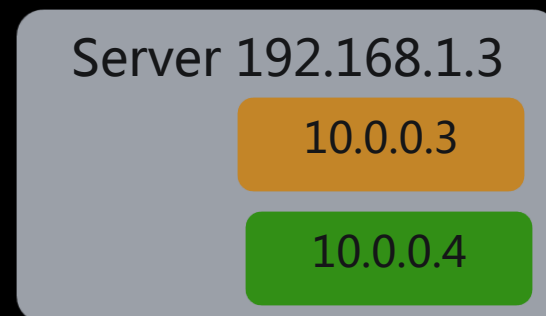


VPC隔离机制

分布式映射服务

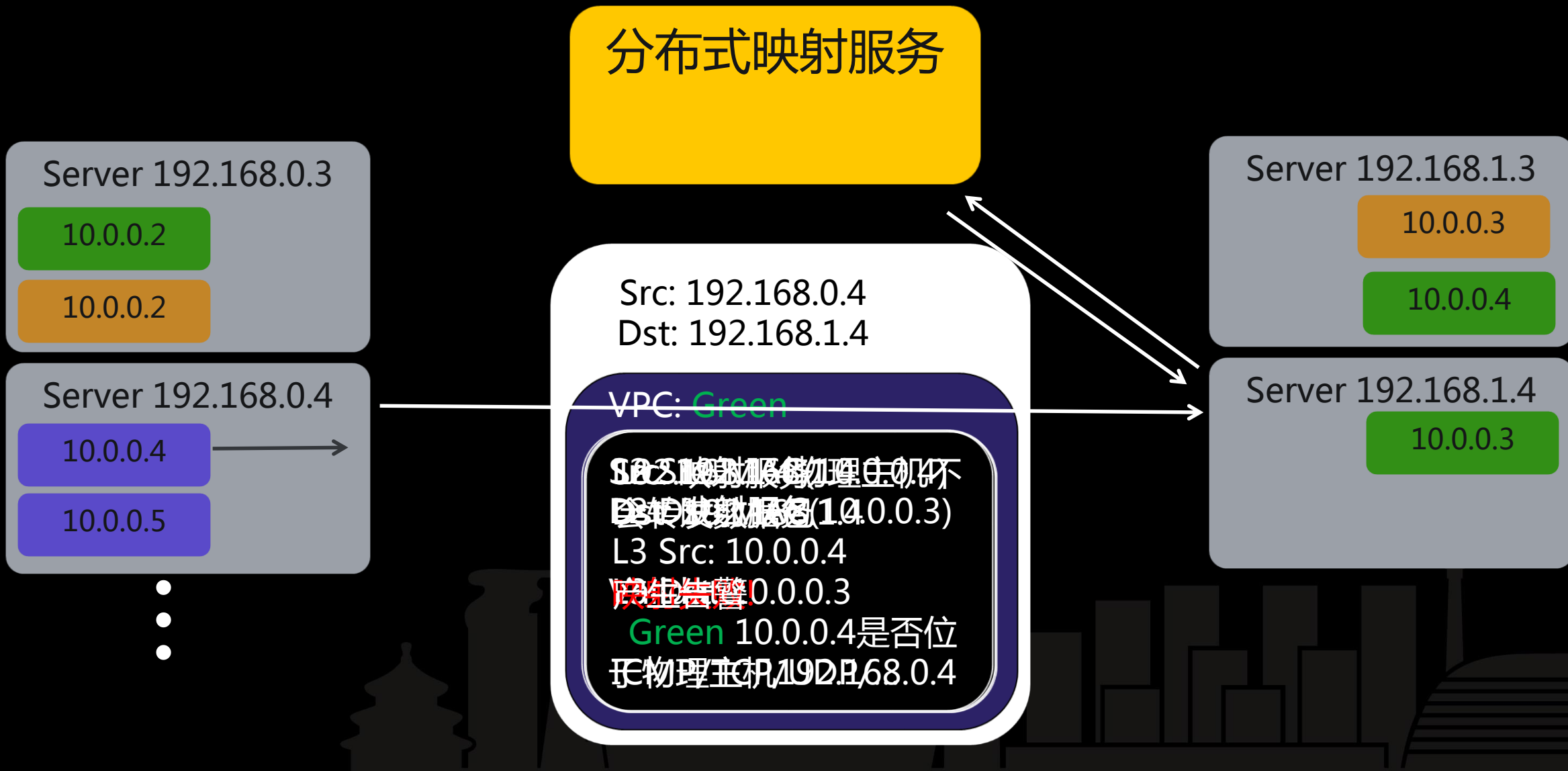


⋮

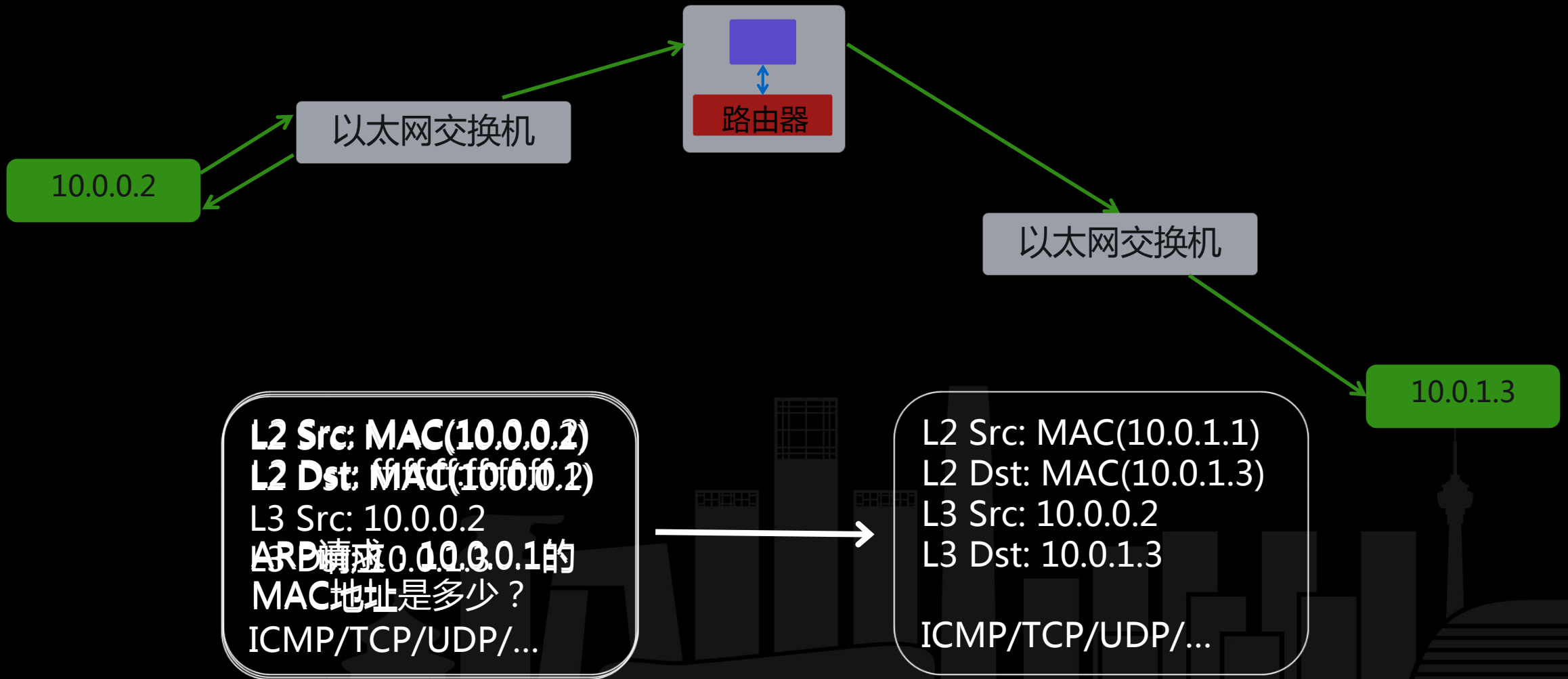


52.10.2.148物理主机上
没有运行服务: Green
请求: 10.0.0.3主机的
映射失败地址是多少?
产生告警

VPC隔离机制

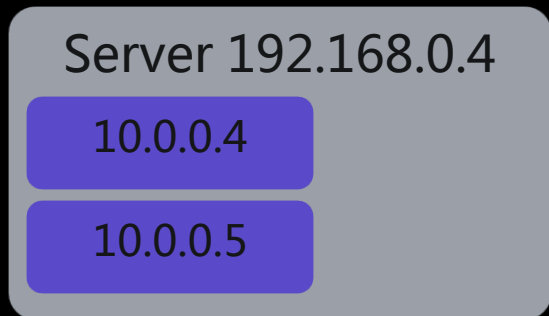
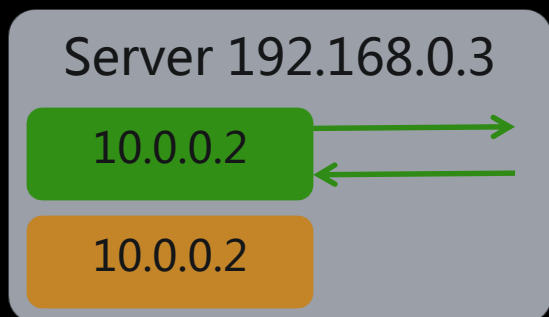


传统三层路由的通信机制

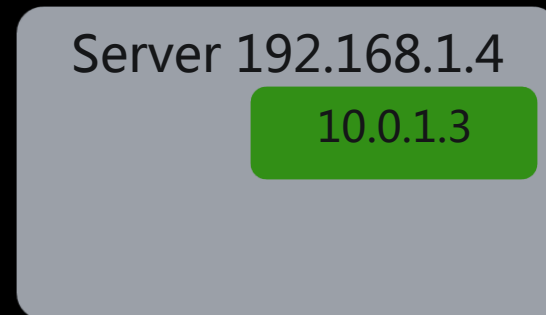
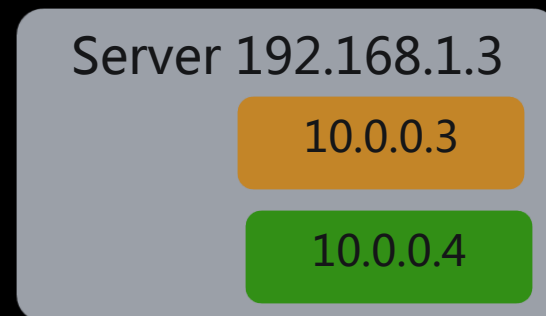


VPC中的三层通信机制

分布式映射服务



...



分布式映射服务(10.0.0.2)

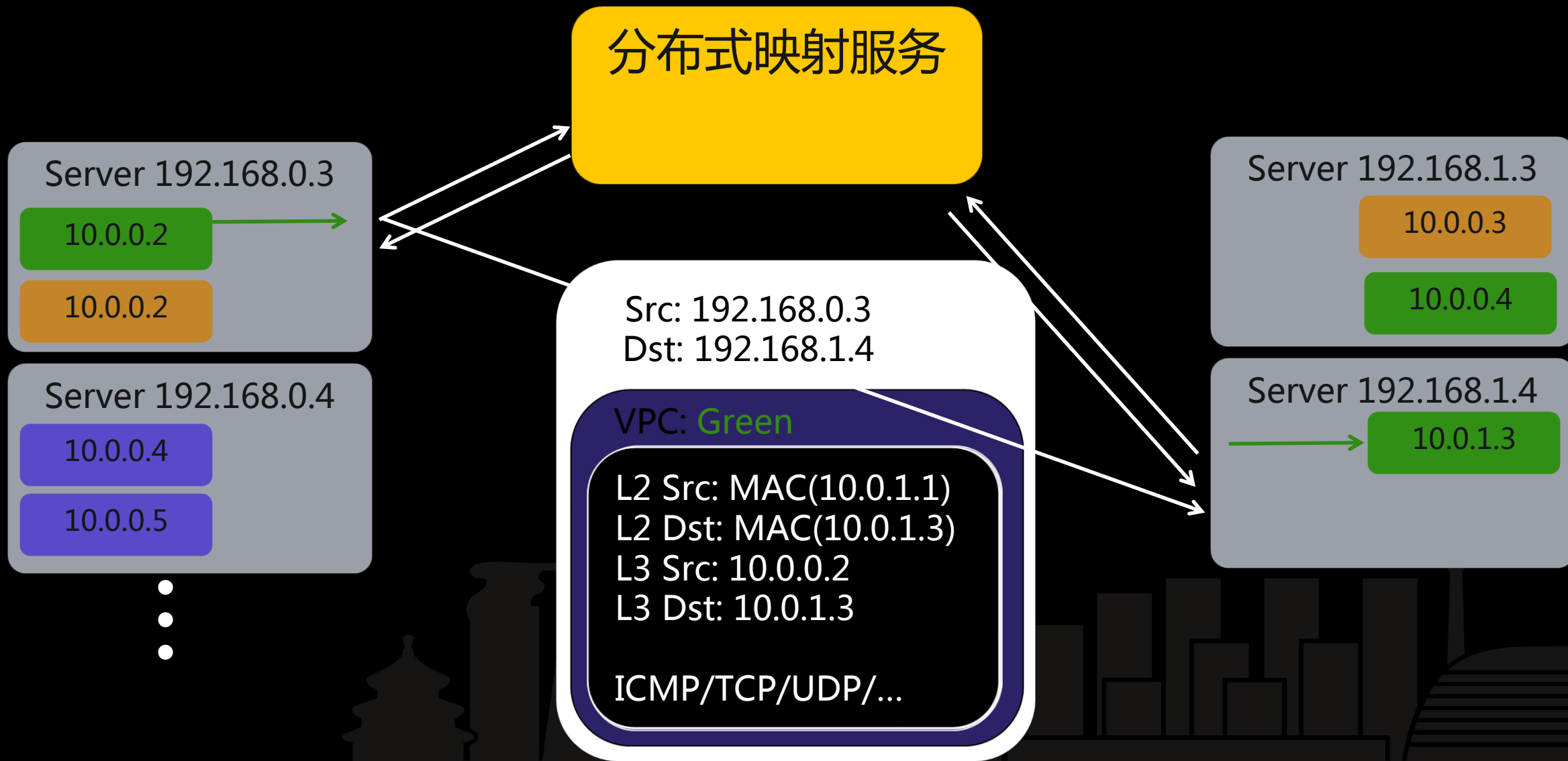
分布式映射服务(10.0.0.2)

ARP请求: 10.0.0.1的

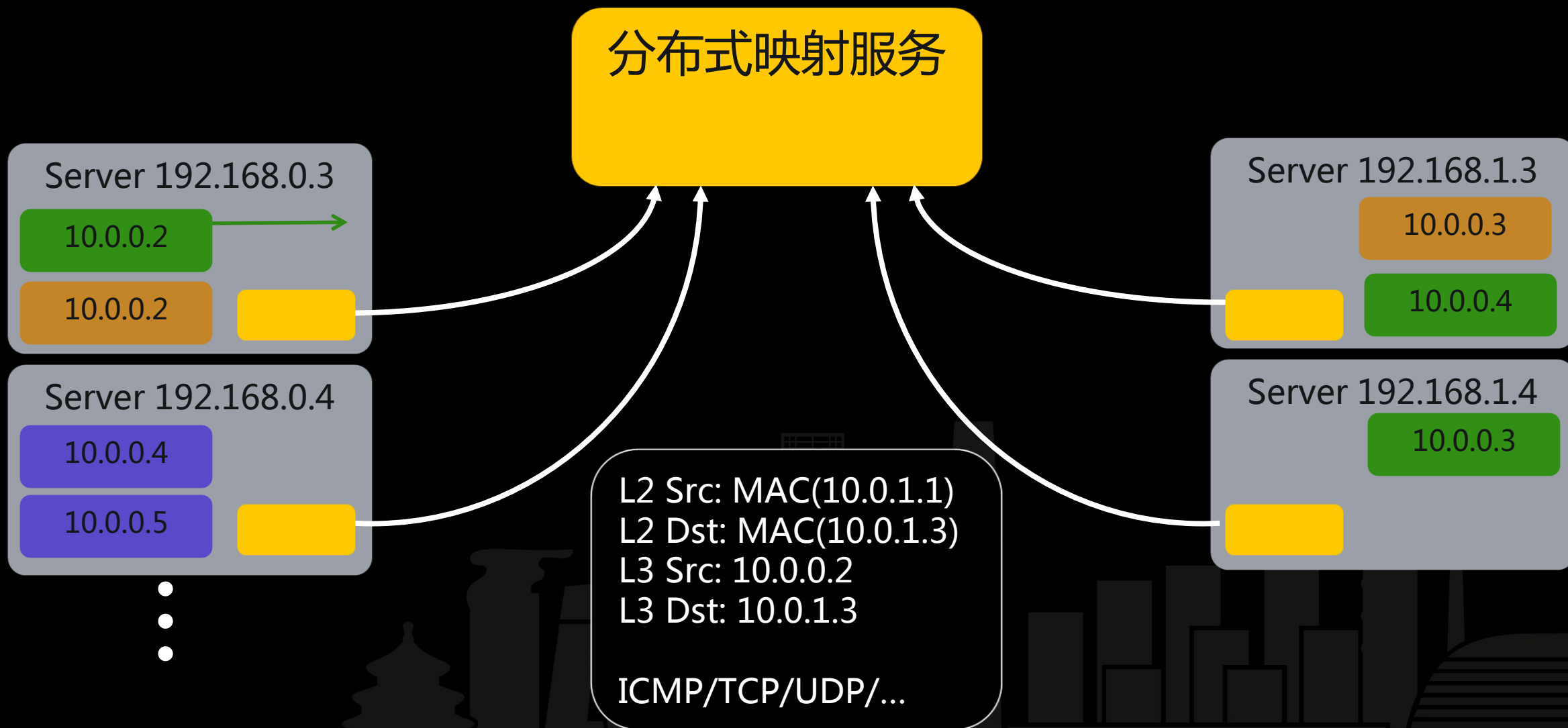
HostID是多少?

MAC: MAC(10.0.0.1)

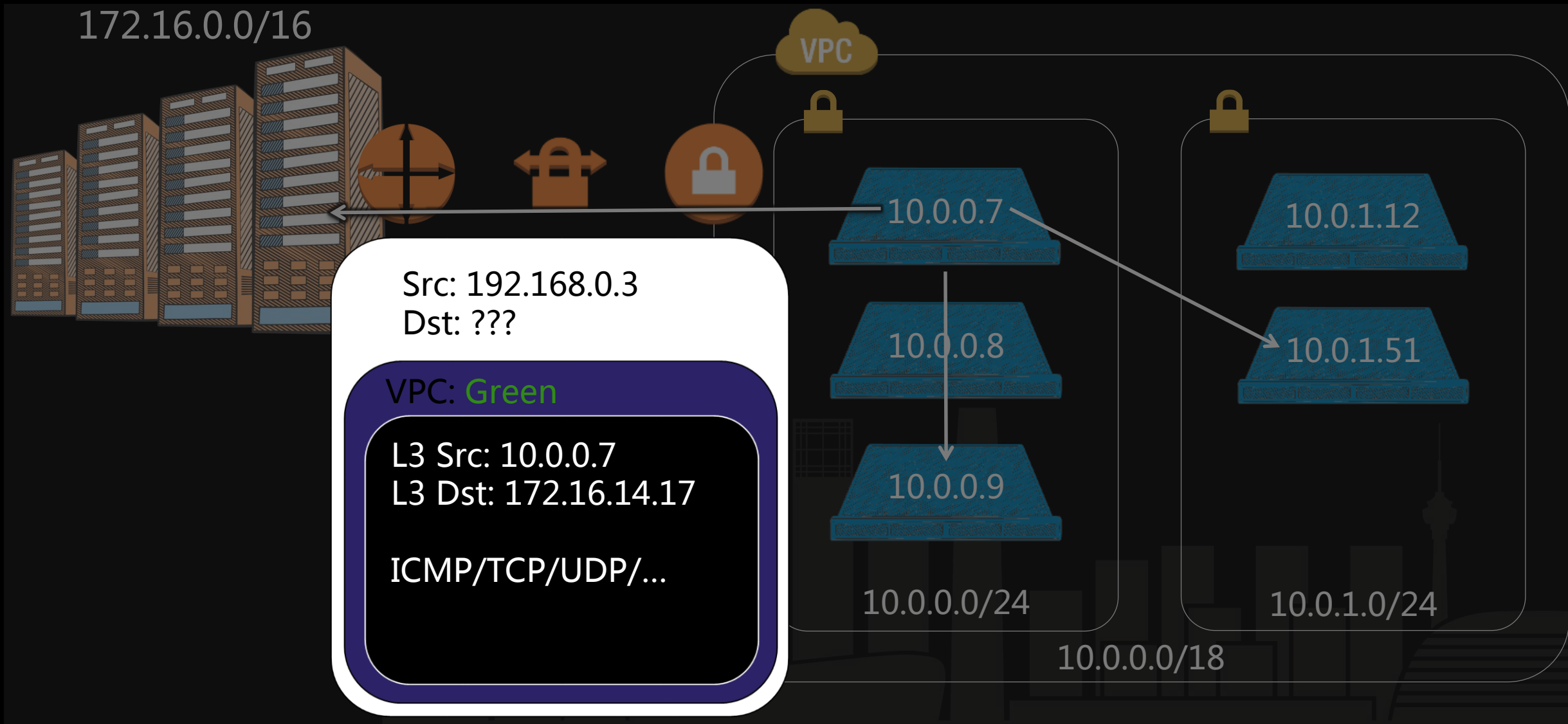
VPC中的三层通信机制



映射缓存

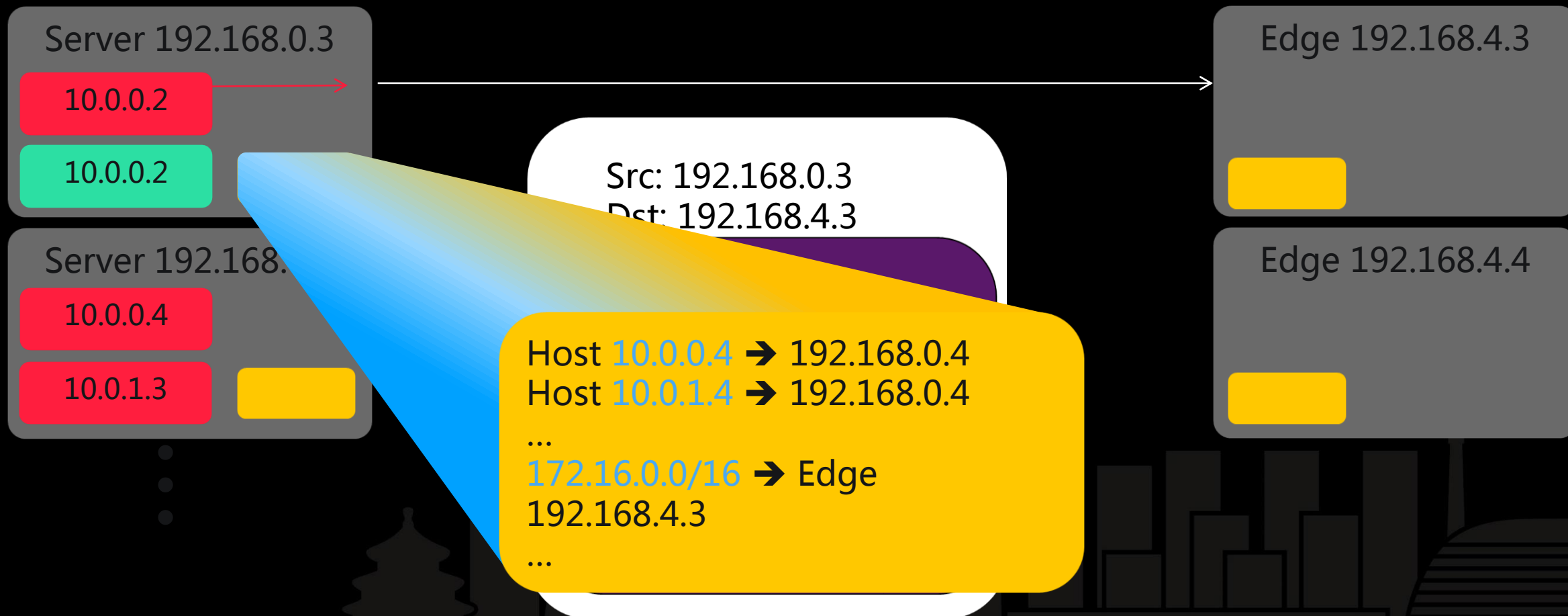


访问VPC外部资源

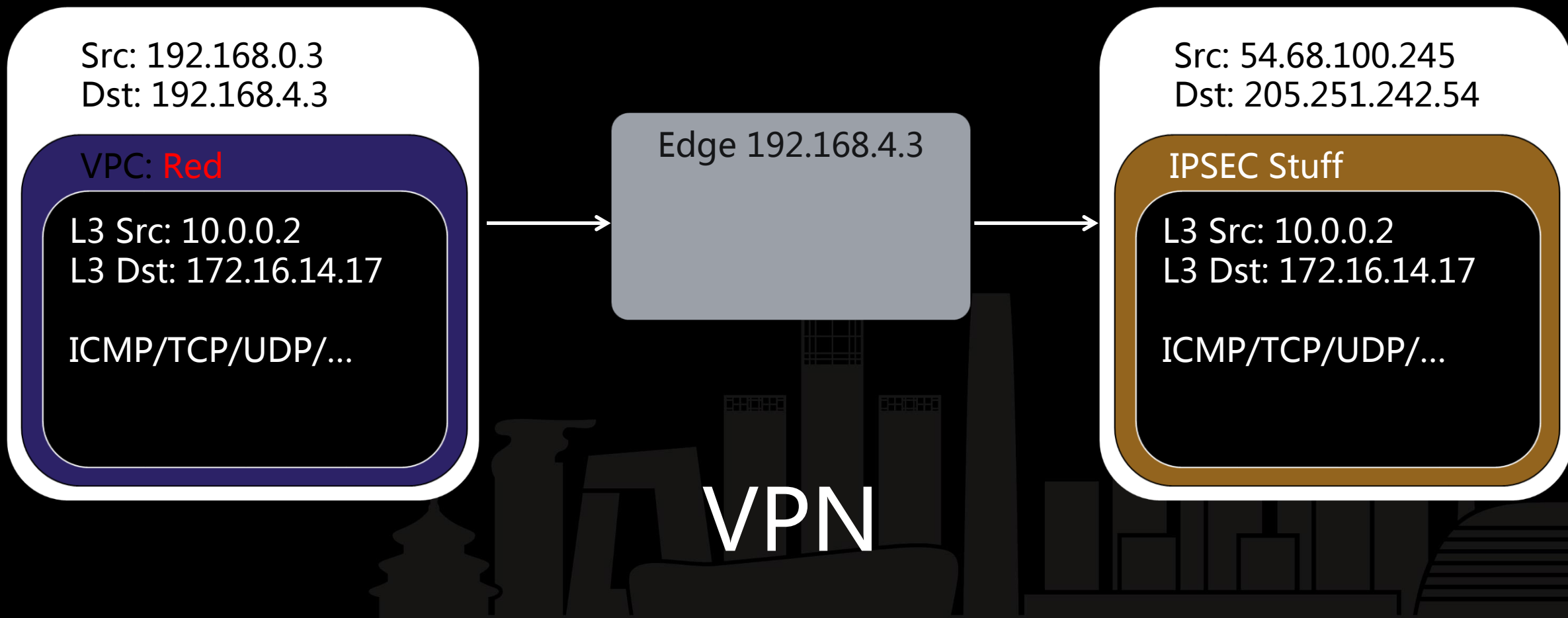


VPC出口服务

分布式映射服务



VPC出口服务—VGW

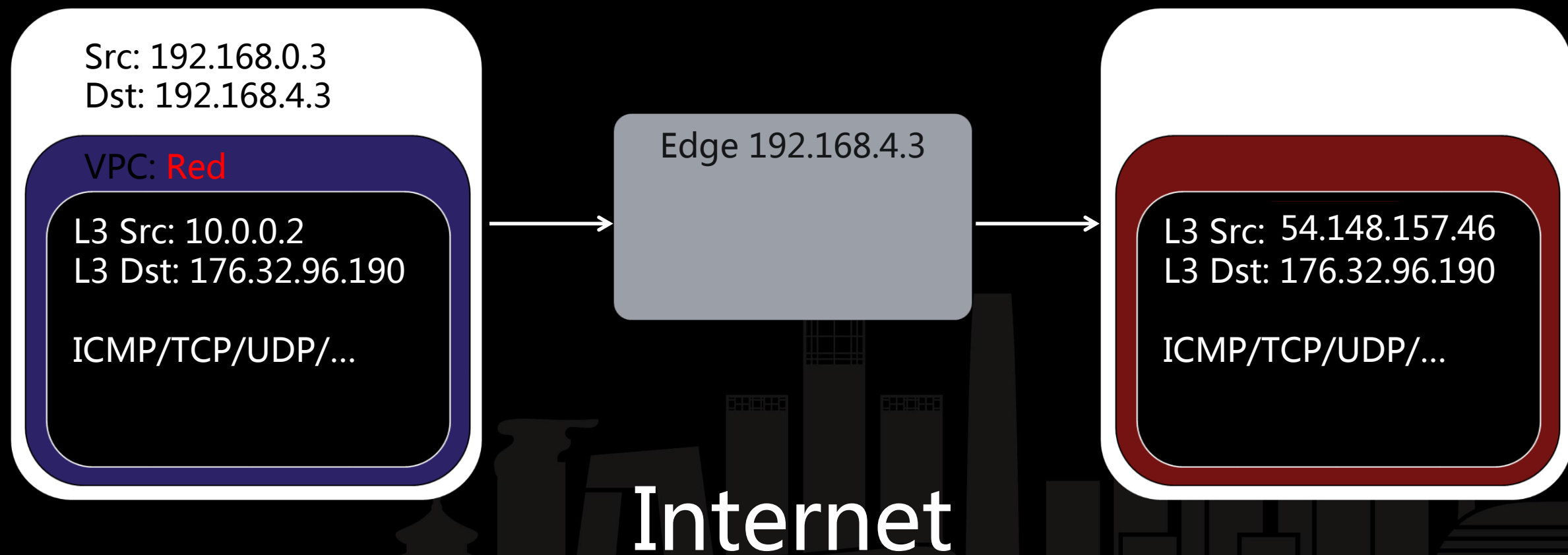


VPC出口服务—VGW



Direct Connect

VPC出口服务—IGW



VPC总结

- 通过分布式映射服务实现VPC id+实例ip与物理服务器ip之间的映射
- 通过全局唯一的VPC id实现租户之间的隔离
- 通过外层封装实现数据包的寻址
- 通过边界服务实现对外访问



关注QCon微信公众号，
获得更多干货！

Thanks!



INTERNATIONAL SOFTWARE DEVELOPMENT CONFERENCE

主办方 **Geekbang** **InfoQ**
极客邦科技