中国电信OpenStack研发实践
&
Google数据中心网络技术漫谈

王　峰
中国电信北京研究院

目录

OpenStack研发是中国电信引入开源软件的重要步骤

中国电信 CHINA TELECOM

■ **基于OpenStack开展自主研发，对中国电信拥有重大意义**

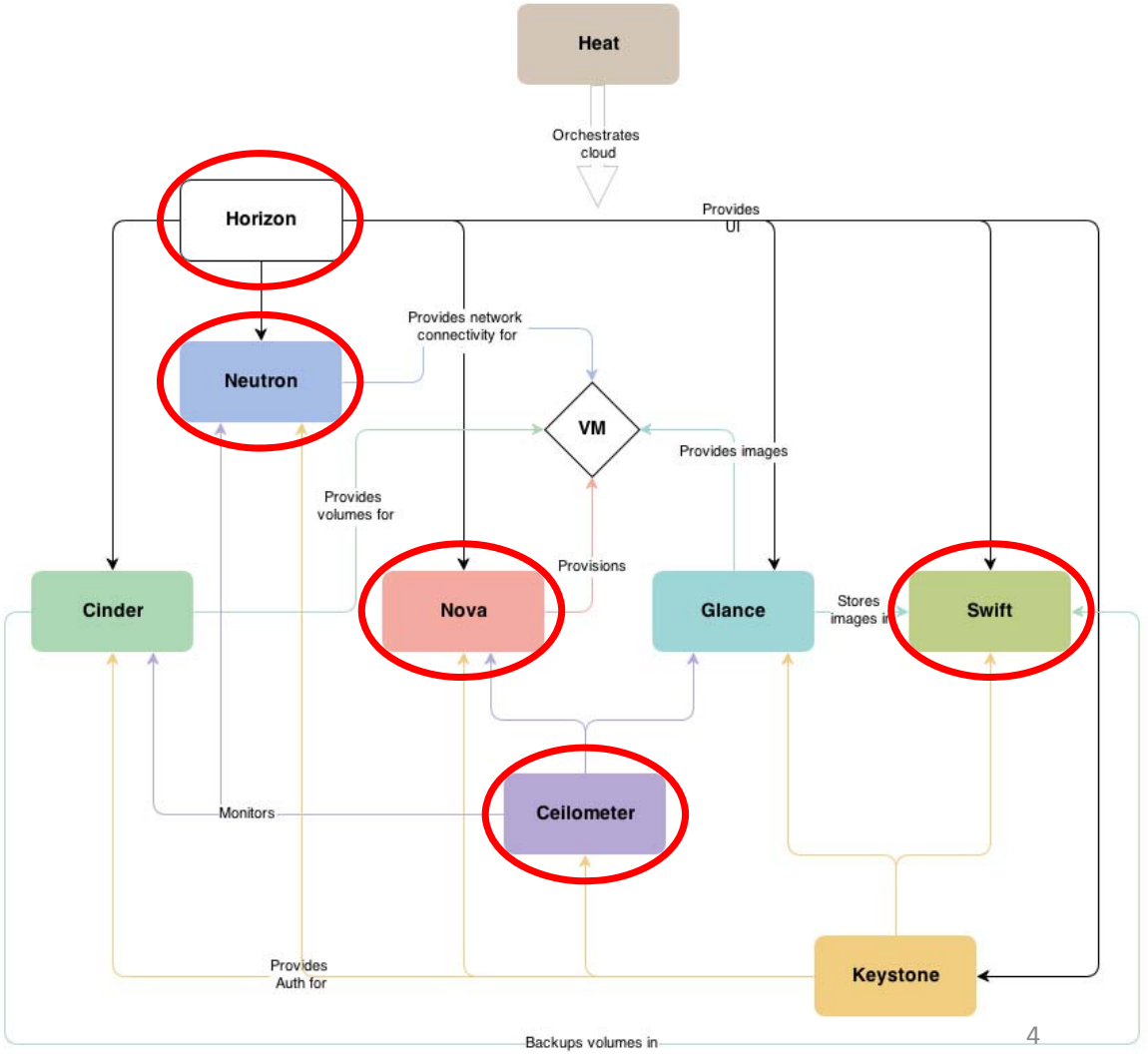
把握业界前沿方向	确立企业技术方向	创新云计算服务
➢ 学习开源软件研发方式 ➢ 掌控云管理平台核心技术	➢ 建立基于开源的研发体系 ➢ 调整优化现有技术架构	➢ 基于开源改进云计算服务 ➢ 针对需求开展云服务创新

■ **中国电信持续推进基于OpenStack开源技术的自主研发**

- ■2013年初，OpenStack被列为集团云计算领域研发重点
- ■以Swift为切入点，研发对象存储技术和产品，并交付客户
- ■总结Swift经验，全面开展基于OpenStack的云管理平台研发
- ■2015年4月，OpenStack云管理平台在企业内外部落地

通过自主研发，积累开源技术研发经验

- **开源之“用”— 有啥吃啥**
  - 基于日渐成熟的OpenStack平台技术，借力社区，跟踪最新版本，更新系统bug
- **开源之“改” — 吃啥点啥**
  - 优化消息队列通信机制，提升组件间通信性能，满足大规模资源池部署的需求
  - 面向存储特性，利用软件程序实现存储的并行读写，实现主机服务秒级开通
- **开源之“创” — 吃啥做啥**
  - 针对技术和业务需求，自主研发异构资源管理、SDN控制器对接、定价计费机制、用户门户定制等创新技术和能力



针对业务实际需求，开展技术服务创新

对接自研控制器 利用SDN实现VPC	实现异构虚拟化软件 统一管理	新增运营平台 实现完善定价和计费	扩展和优化用户平台 Web操作界面
➢ 利用Neutron组件提供虚拟化网络的基本功能，并提供统一框架对接和管理不同厂商的SDN控制器 ➢ 实现自研SDN控制器与Neutron组件对接，基于OpenFlow管控网络数据通路	➢ 利用Nova组件管理虚拟化计算资源，支持虚拟化资源动态配置 ➢ 研发和扩展不同虚拟化技术与Nova组件的对接与整合，实现异构虚拟机的统一管理	➢ 利用Ceilometer组件采集资源使用事件，为计费和资源计量提供基本的基础数据和机制 ➢ 研发完成运营平台，新增和完善产品和服务定义、订单管理、客户管理、价格配置、系统配置等功能	➢ 利用Horizon组件提供平台操控基本GUI ➢ 根据业务逻辑，完善和优化平台操作界面，实现界面能力的扩展和定制化开发，支持用户和管理员进行分权分域的登录和使用

探索基于SDN技术的云计算网络能力提升

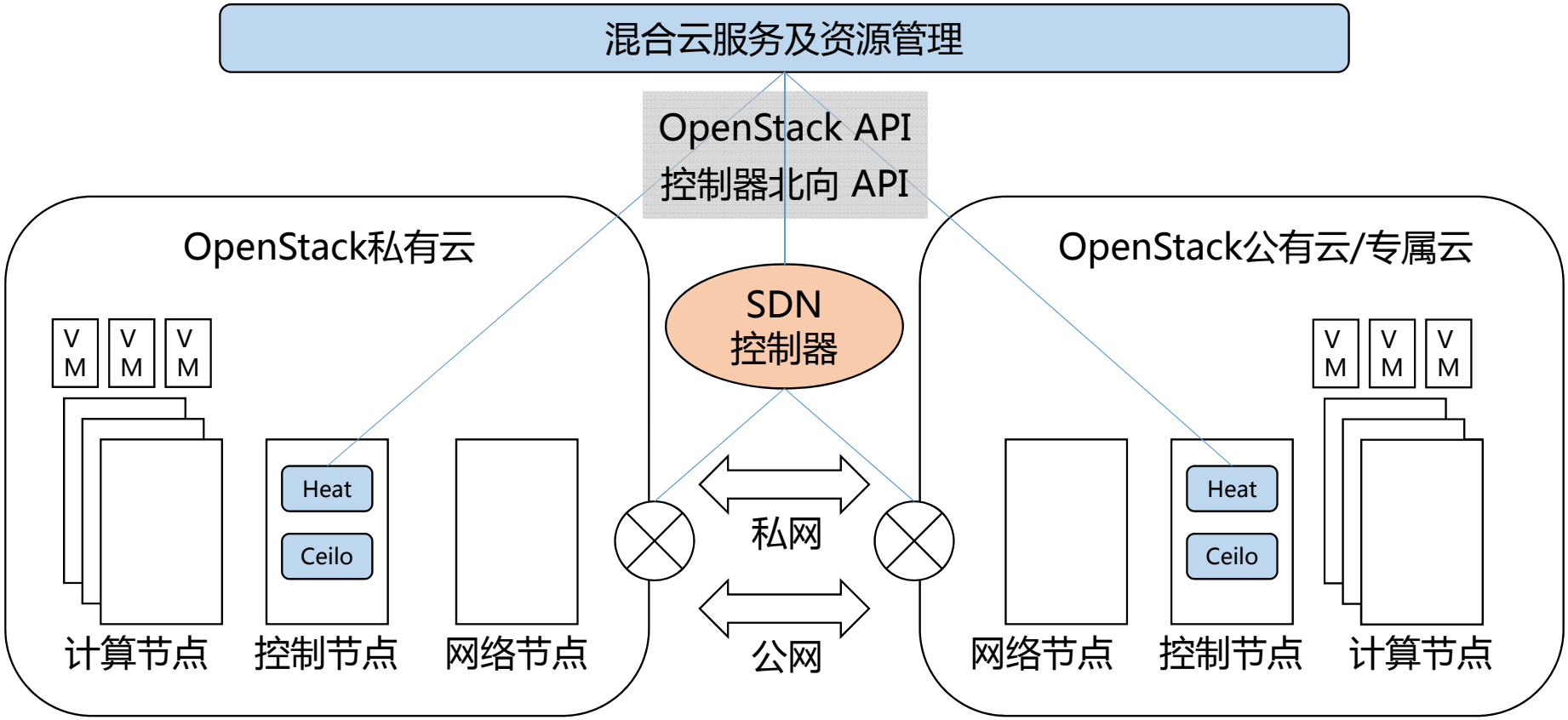
- **自主研发SDN控制器，满足云服务网络需求**
  - 参考Floodlight、Ryu、NOX、MUL等开源技术实现，针对运营商数据中心需求设计控制器
  - 拥有Java和C两个版本，实现控制器核心能力，支持多租户网络、虚拟防火墙等典型网络服务

- **控制器研发成果贡献给国家863 SDN开源社区，并积极参与社区建设**

```
#创建租户虚拟局域网，包括VLAN ID、网络名称、网关等
    if (log.isDebugEnabled()) {
        String gw = null;
        try {
            gw = IPv4.fromIPv4Address(gateway);
        } catch (Exception e) { }
        log.debug("Creating network {} with ID {} and gateway {}",
                new Object[] {network, vlanid, gw});
    }
    if (!nameToVlanid.isEmpty()) {
        for (Entry<String, String> entry : nameToVlanid.entrySet()) {
            if (entry.getValue().equals(vlanid)) {
                nameToVlanid.remove(entry.getKey());
                break;
            }
        }
    }
    nameToVlanid.put(network, vlanid);
    if (vNetsByVlanid.containsKey(vlanid))
        vNetsByVlanid.get(vlanid).setName(network);
    else
        vNetsByVlanid.put(vlanid, new VirtualNetwork(network, vlanid));
     if ((gateway != null) && (gateway != 0)) {
        addGateway(vlanid, gateway);
        if(vNetsByVlanid.get(vlanid)!=null)
            vNetsByVlanid.get(vlanid).setGateway(IPv4.fromIPv4Address(
gateway));
        }
    }
```

发展方向：面向混合云服务的管理平台研发

■突出运营商特色，支持云服务与网络资源的协同调度

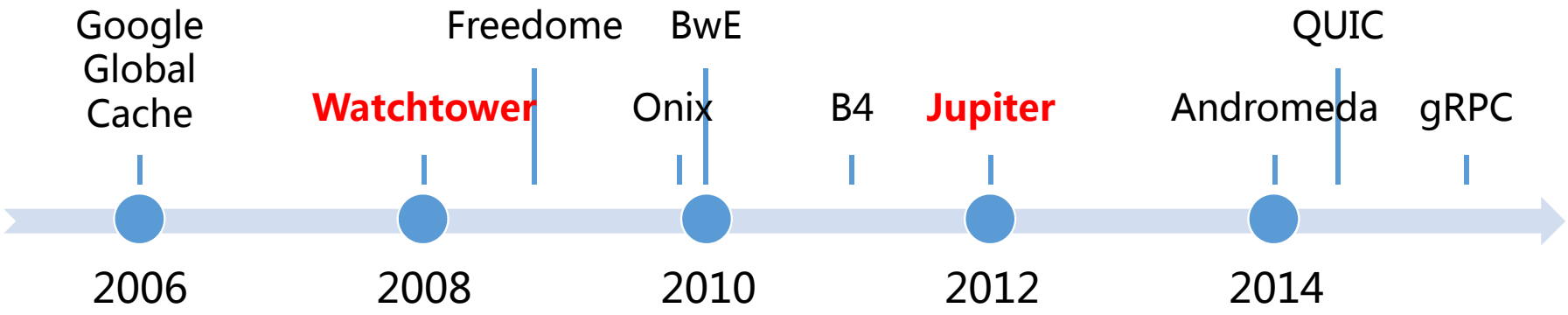


目录

- 中国电信OpenStack研发实践
- **Google数据中心网络技术**

Google网络技术发展路径

■ **Google已经形成完备的网络技术体系，并在诸多领域做出巨大创新**



名称	说明
Google Global Cache	CDN网络技术
Watchtower	第三代数据中心网络技术
Freedome	园区网络技术
Onix	SDN控制器
BwE	？

名称	说明
B4	广域网互连
Jupiter	第五代数据中心网络技术
Andromeda	网络虚拟化技术
QUIC	新的传输层协议 Quick UDP Internet Connections
gRPC	支持多平台的RPC技术

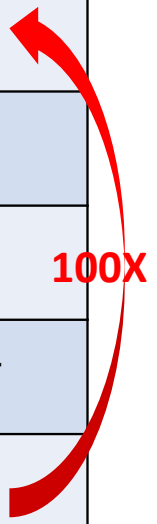
网络成为突破数据中心瓶颈的关键

- **数据中心网络是Google云平台的基础，支持Google高可扩展、高性能、高可用**

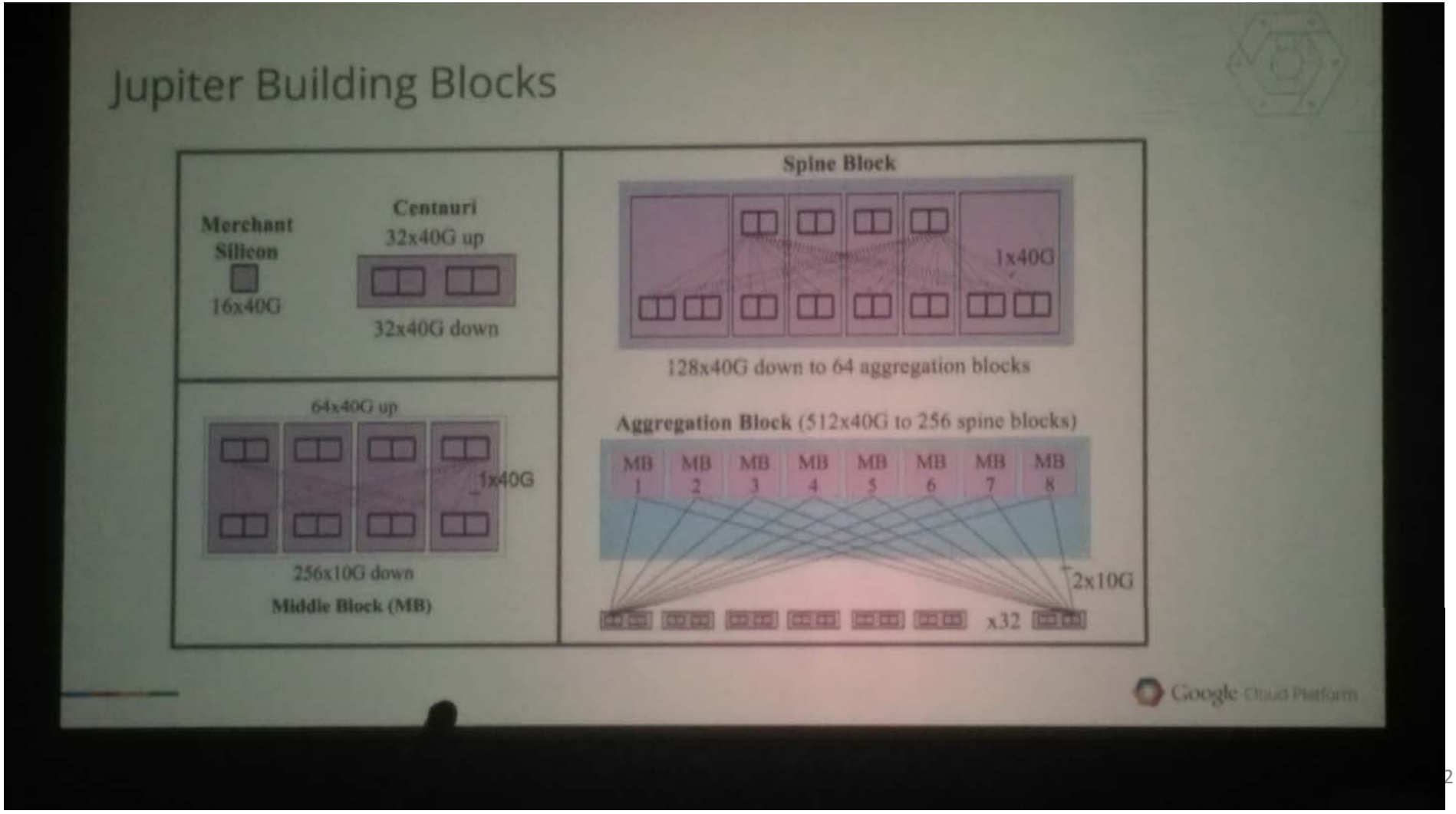
计算	➢ 摩尔定律失效，单点性能受限 ➢ 分布式计算依赖socket编程
存储	➢ 通过分布式实现提升容量 ➢ 存储I/O仍旧为性能瓶颈 ➢ 下一代存储Flash容量有限
网络	➢ **Clos 拓扑：多级交换、无阻塞、每一级每个单元与下一级设备全相连** ➢ **商用芯片（Merchant Silicon）：成本低、开展数据中心新协议定制** ➢ **集中化的软件控制器：数以千计的小交换机 → 一台大型逻辑网络设备**

Google数据中心网络演进

DC Gen.	Year	Merchant Silicon	ToR Config	Aggr. Block	Spine Block	Fabric Speed	Host Speed	Aggr BW
Four-Post CRs	2004	Vender	48 × 1G	--	--	10G	1G	2T
Firehose 1.0	2005	8 × 10G 4 × 10G(ToR)	2 × 10G up 24 × 1G down	2 × (32 × 10G)	32 × 10G	10G	1G	10T
Firehose 1.1	2006	8 × 10G	4 × 10G up 48 × 1G down	64 × 10G	32 × 10G	10G	1G	10T
Watchtower	2008	16 × 10G	4 × 10G up 48 × 1G down	4 × (128 × 10G)	128 × 10G	10G	1G	82T
Saturn	2009	24 × 10G	24 × 10G	4 × (288 × 10G)	288 × 10G	10G	10G	207T
Jupiter	**2012**	**16 × 40G**	**16 × 40G**	**8 × (128 × 40G)**	**128 × 40G**	**10/40G**	**10/40G**	**1.3P**



Jupiter Building Block



Jupiter Superblock

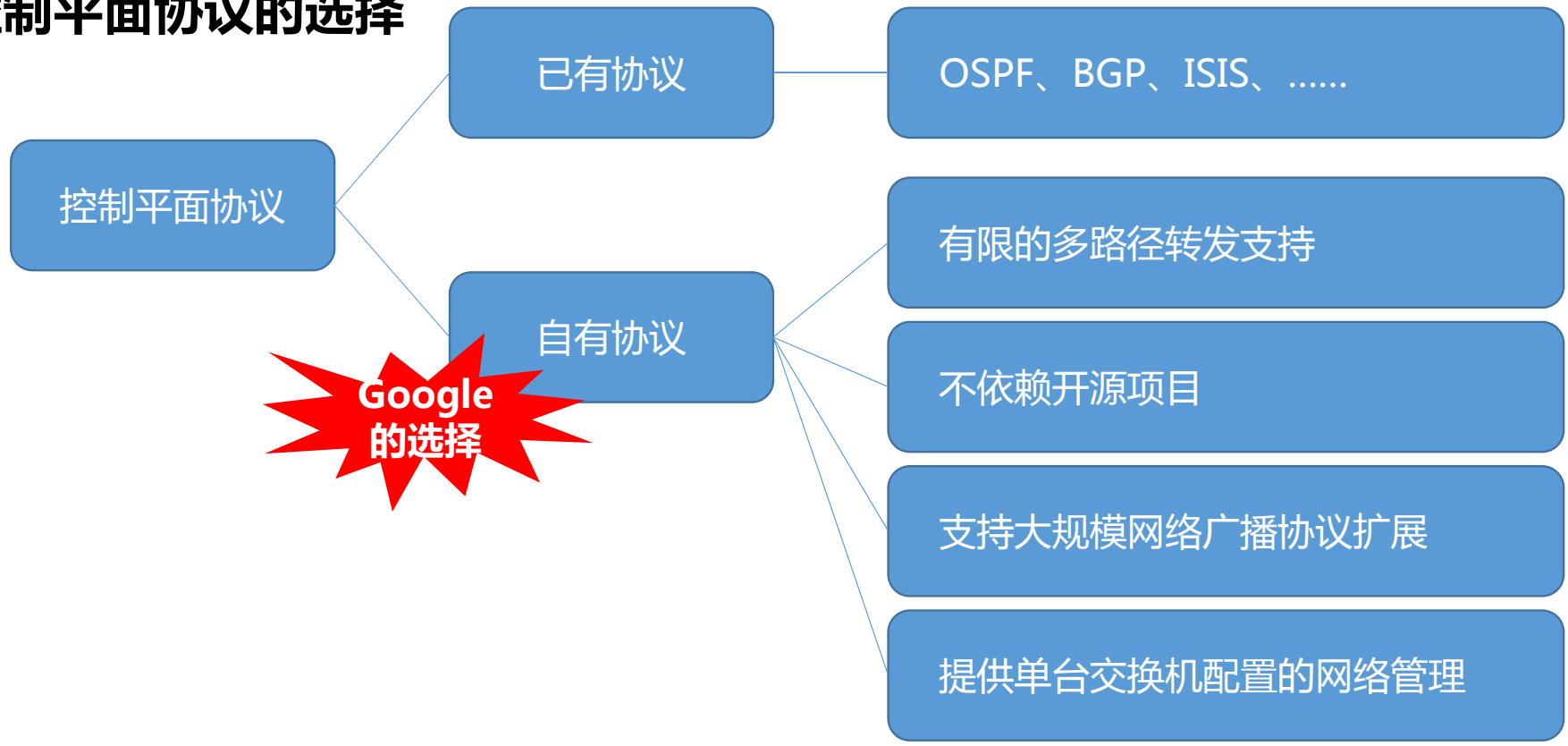
中国电信 CHINA TELECOM

Google Jupiter Superblock

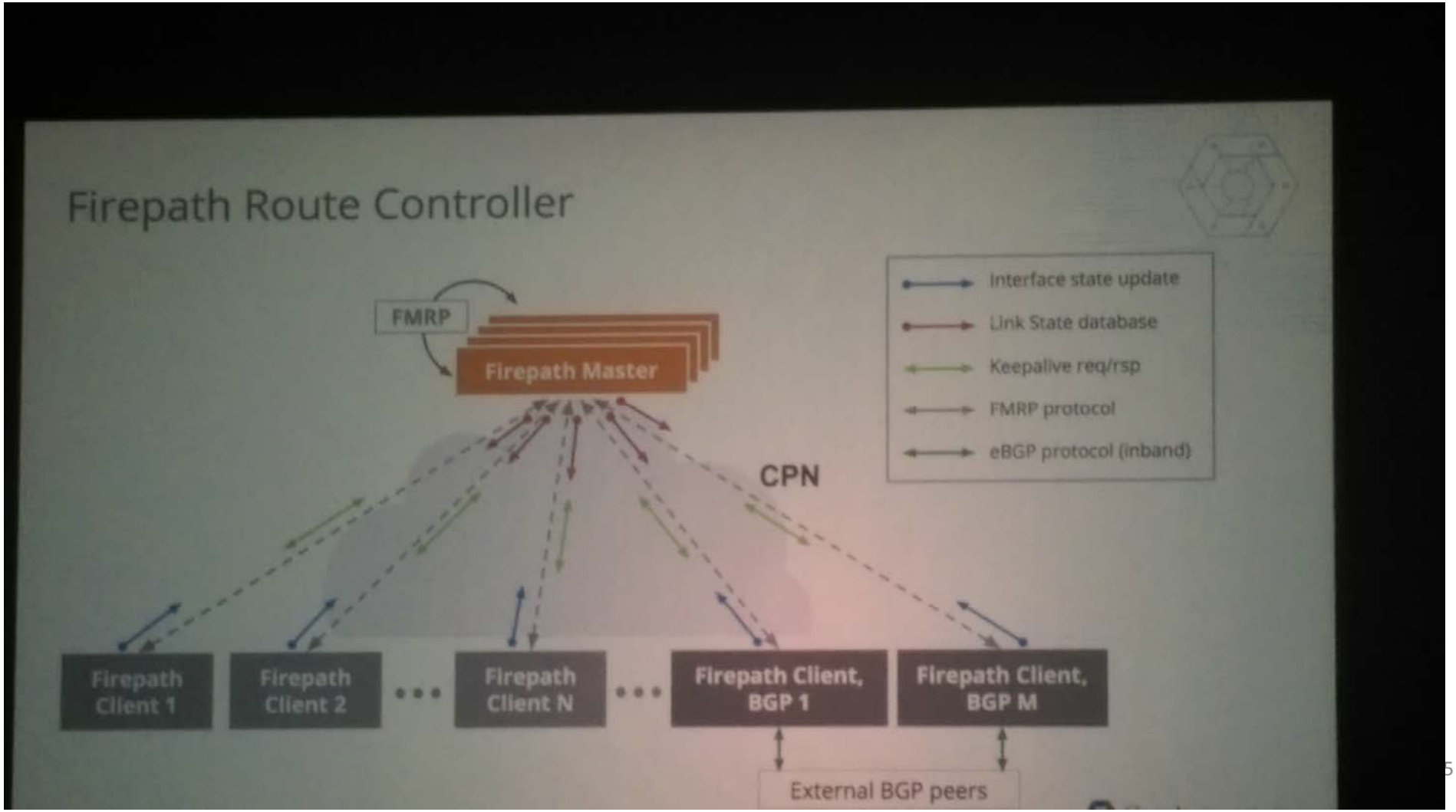
Google Cloud Platform

Google数据中心控制层

- **逻辑上集中部署并与数据层点对点连接的的控制层是大规模分布式系统的必需**
  - GFS、MapReduce、BigTable、Spanner、B4、Andromeda
- **控制平面协议的选择**



Firepath Route Controller



感想 & 小结

- **Google的今天就是大家的明天、后天**
  - 网络是超大规模数据中心的基石
  - 数据中心的东西向流量 >> 南北向流量
- **Google利用分布式理念解决网络问题**
  - 网络设备自研：商用器件 + Linux + 自有协议
  - 网络设备同质，不强调单台设备优势
  - 利用分布式提升性能、扩展性、可用性
  - 逻辑上集中控制是提升网络管控效率的必需
- **以Google为鉴，可以 ……**

中国电信
CHINA TELECOM

感谢聆听