开放交换机与可编程交换芯片在Telemetry中的应用

中国银联　　电子商务与电子支付国家工程实验室

周雍恺

目录 CONTENT

01 背景概要

金融网络转型
技术发展趋势
建设思路

金融网络转型的驱动力

银联网络

持卡人

2017年：

- 交易笔数：293.5亿笔
- 交易金额：93.9万亿元
- 累计发行银联卡：65亿张

商户

- 活动商户：2238 万户
- 非接受理终端：1359.8万台
- 云闪付无障碍街区：600个

业务驱动

- 一行三会监管要求
- 银联2020技术规划
- 金融行业的科技转型

金融网络转型

技术驱动

- 微服务架构变革
- 智能化浪潮
- 开源开放的技术趋势

战略

- 打造开放式、平台型综合支付服务商，向数据公司、科技公司转型。
- 加快推进市场化转型、加大对新技术和新模式的探索和投入，加速银联网络全球布局。

银行系开始设立科技子公司，当前已有6家

中国民生银行 CHINA MINSHENG BANKING CORP., LTD.

中国建设银行 China Construction Bank

招商银行 CHINA MERCHANTS BANK

兴业银行 INDUSTRIAL BANK CO.,LTD.

中国光大银行 CHINA EVERBRIGHT BANK

中国平安 保险·银行·投资 平安银行 PING AN BANK

FinTech

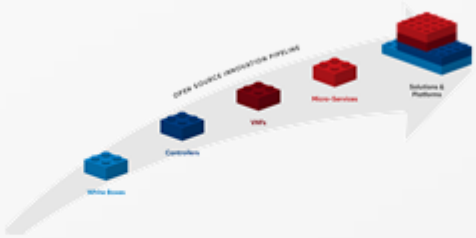
1 **集聚效应凸显，金融科技的行业化服务兴起**

金融行业云将发展成为金融IT一种新的业务盈利模式；
有实力的金融机构将在维护私有系统的同时，转型成为大型金融IT运营商。

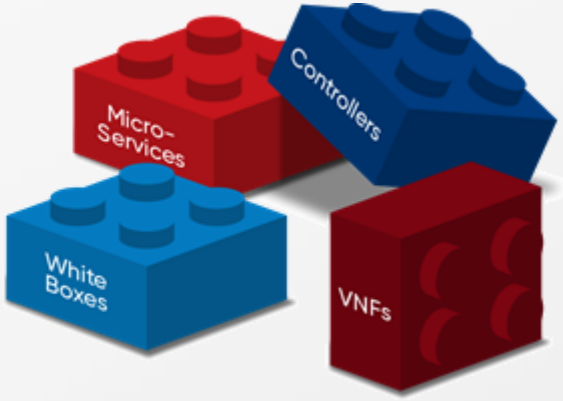
2 **两地三中心向多地多中心的物理基础设施演变，互联网流量大规模接入**

多中心的物理基础可以提供更多的地域选择与资源调度可能。
随着业务向互联网转型，互联网流量将大规模接入。

分布式模块化、微服务的设计



“简洁明了”的设计理念
Simple & Clear



开放开源的技术趋向

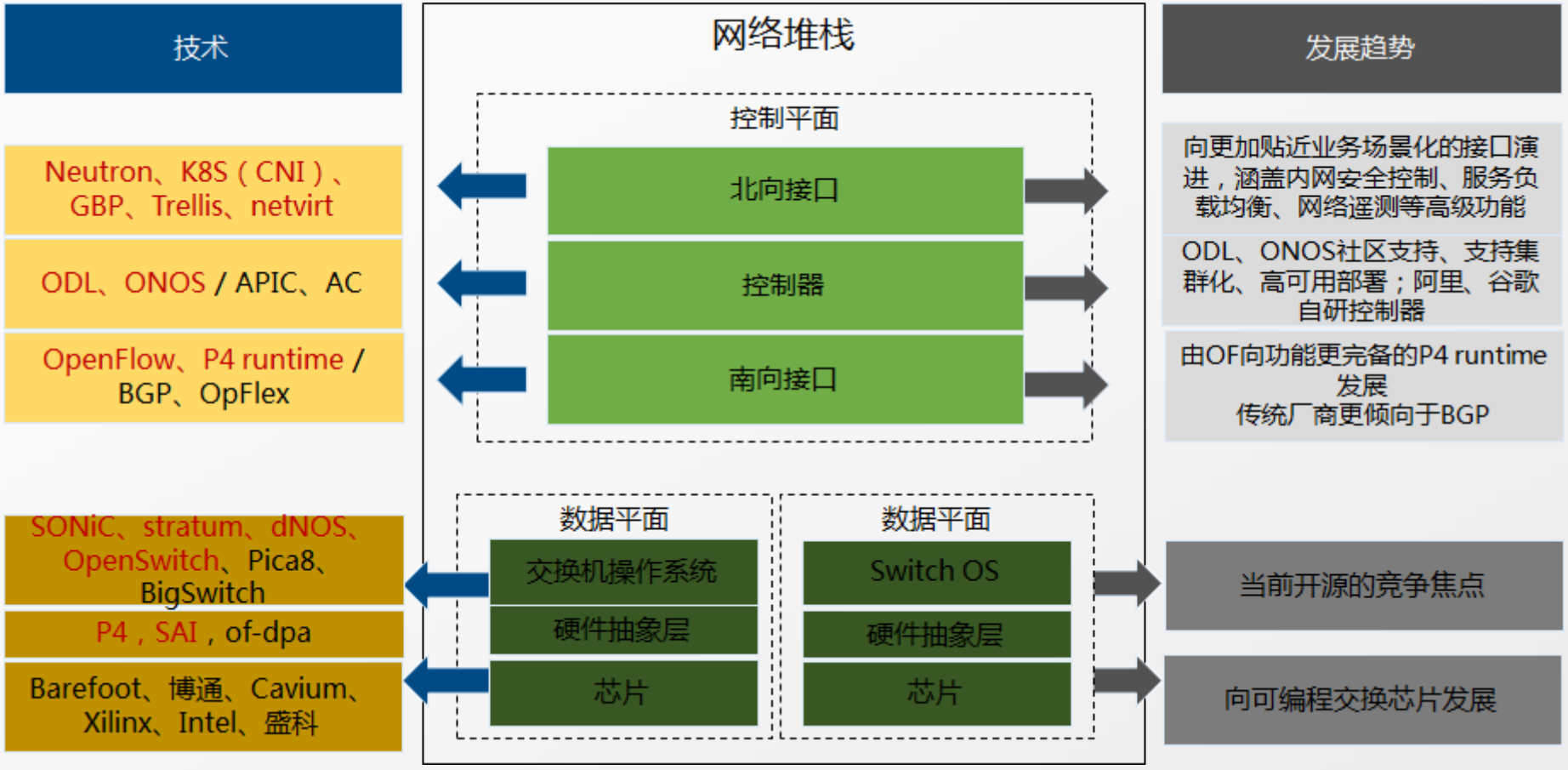
- 复杂技术向**极简架构**转化，降低准入门槛
- **开放控制**增强管控粒度与可控能力
- **开源生态**聚合资源，形成事实标准
- **平台弹性化**：前端微服务化，持久型服务分布化

02

网络技术前研动态

网络技术堆栈
网络控制演进
交换机操作系统的路线之争
可编程芯片革命

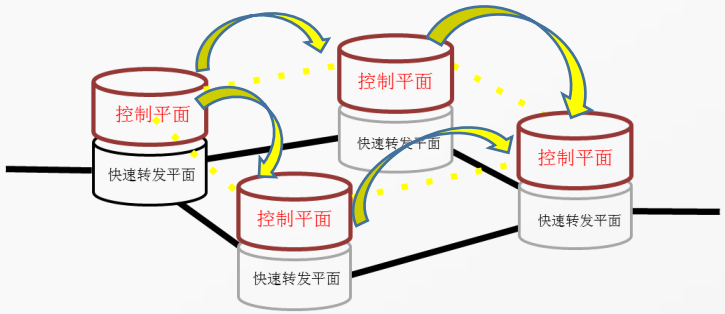
02 网络前研 | 网络技术堆栈



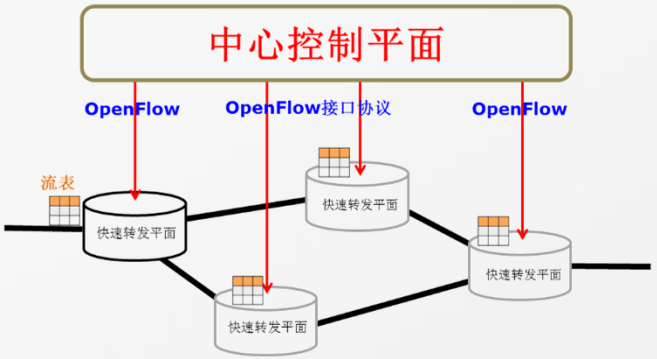
- 网络开放大势所趋，开源软件已经能够实现对网络堆栈的全覆盖
- 既去IOE之后，标准硬件+开源软件的开放交换机生态已经形成

02 网络前研 | 网络控制的演进

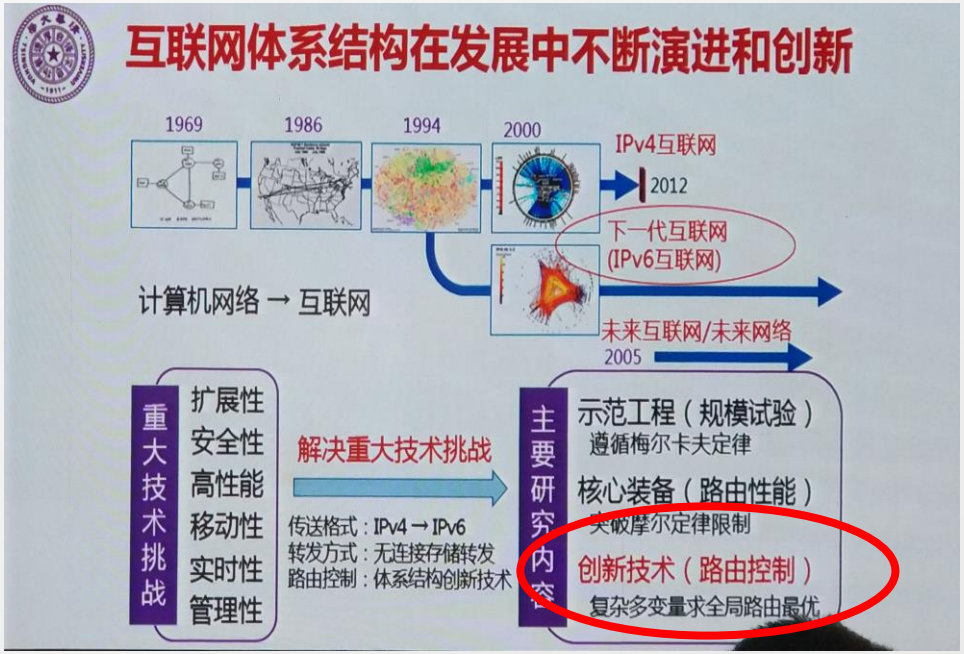
传统网络



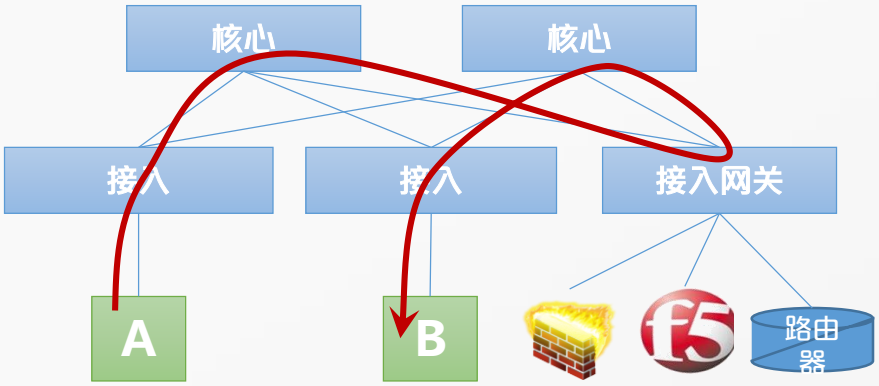
SDN



吴建平院士：GNTC大会

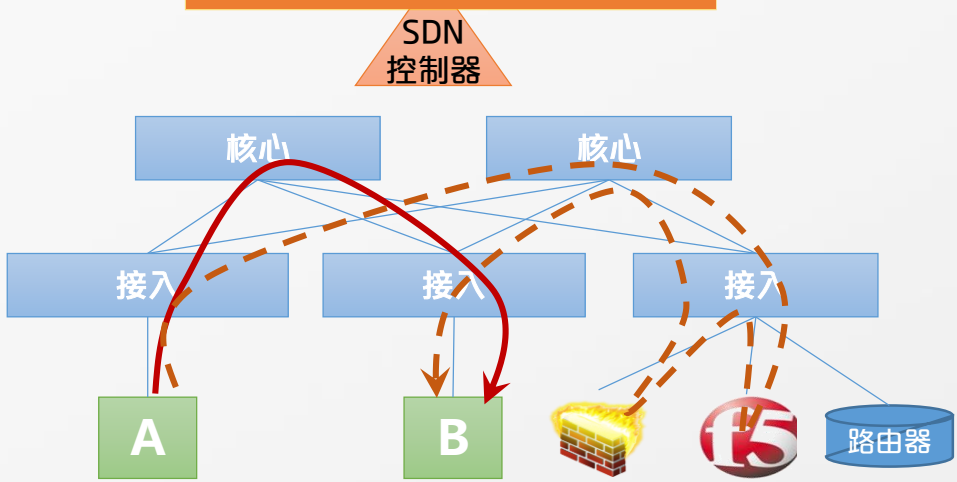


经典路由控制



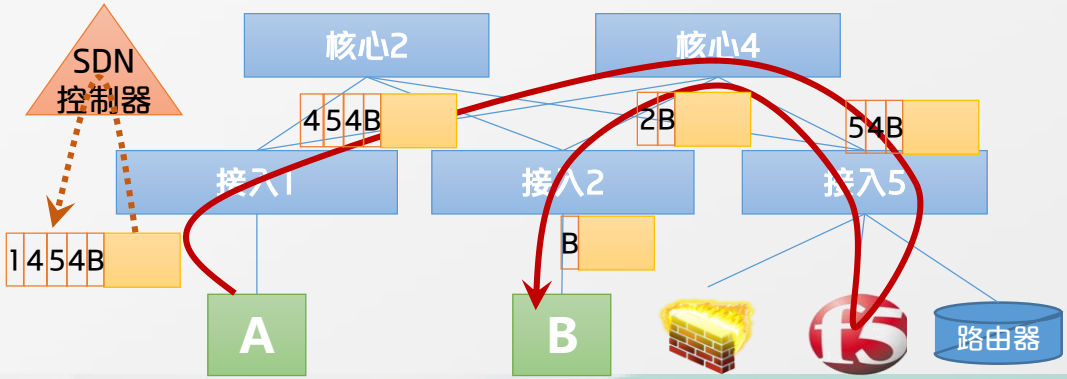
1. 必须先制定标准路由协议
2. 网关、路由等经典概念
3. 典型代表：BGP

SDN强控制



1. 中心控制平面具有全局拓扑，直接下发控制信息
2. 典型：OpenFlow流控制

源路由控制



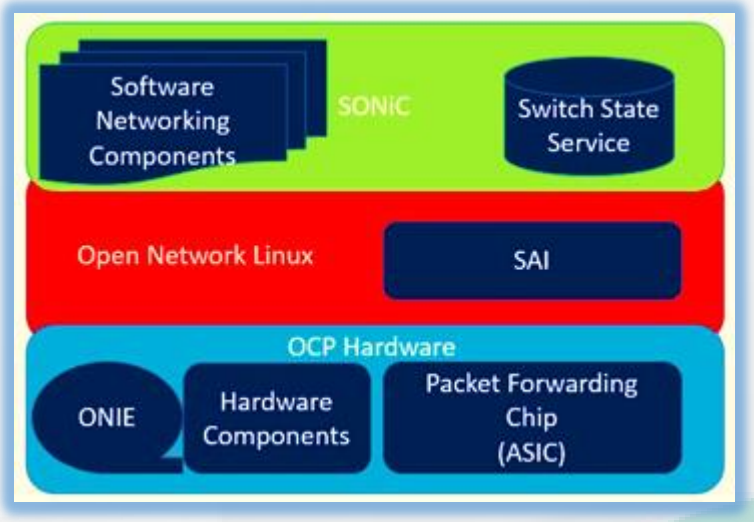
1. 源节点有选路权
2. 中间结点无需查表
3. TE（流量工程）
4. 典型：Segment Routing（段路由）

Microsoft Alibaba.com Google Tencent 腾讯

BGP

SONiC：经典SwitchOS

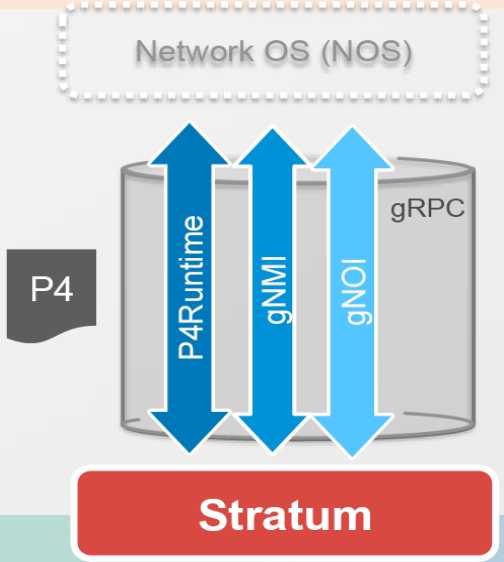
1. 传统交换机的开源版，保留精简的路由协议（BGP）
2. Underlay尽量简单，BGP解决一切问题
3. Overlay逻辑通常在主机侧通过ovs或智能网卡（SmartNIC）实现



SDN

Stratum：轻量级SwitchOS

1. 控制策略由SDN控制器通过P4 runtime（革新版的OpenFlow）下发
2. 侧重于灵活的转发行为
3. 本身不带网络协议，需要SDN控制器/控制平面配合



02 网络前研 | 可编程芯片的革命

网络芯片需要解决的核心问题

一个端口到另外一个端口的处理转发

快
- 单芯片总交换容量Tbps
- 全线速转发

90%
SW ASIC

BROADCOM
Connecting everything

SDN时代网络芯片所存在的问题

数据平面功能固化，无法灵活定制

固定流水线芯片性能高，灵活性不够
新功能需要重新设计制造芯片
用户不可编程

Nick McKeown
SDN共同发明人与推动者，斯坦福大学教授，美国两院院士，英国皇家科学院院士

vmware
nicira
OvS Open vSwitch

2013年，Nicira被Vmware以12.6亿美金收购，轰动业界

软件交换机带动了Overlay的灵活组网模式

硬件交换芯片也需要可编程性的灵活赋能

2013年创立Barefoot，2016年底全流水线可编程的Tofino芯片诞生

全流水线可编程的ASIC设计

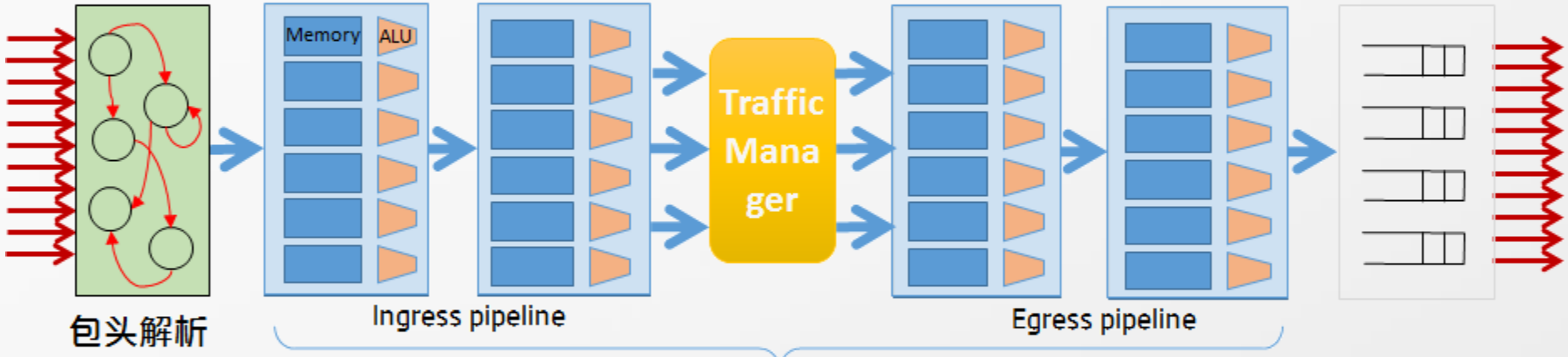
RMT架构 / PISA

简明的架构设计
基本单元：Match-Action Table
- Match：任意的报文偏移与匹配长度
- Action：丰富的报文编辑功能

Ingress/Egress多级流水线
片上资源（SRAM、TCAM）灵活配置

实体芯片

2016年底Tofino芯片诞生
- 第一款RMT架构的ASIC交换芯片
- 12级用户可编程流水线
- 512字节自定义偏移
- 全线速转发




```
parser parse_ethernet {
   extract(ethernet);
   return
switch(ethernet.ethertype) {
      0x8100 : parse_vlan_tag;
      0x0800 : parse_ipv4;
      default: ingress;
}
```

自定义报文头

```
control ingress {
    apply(port_table);
    if (l2_meta.vlan_tags == 0) {
        process_assign_vlan();
    }
}
```

控制流

```
header_type  ethernet_t { … }
header    ethernet_t
vlan_tag[2];
metadata  l2_metadata_t l2_meta;
table port_table { …

}
```

表项定义

02 网络前研 | P4的业界成果

学术界：解决转发面性能和灵活性问题

- SIGCOMM 2013年提出了基于RMT架构的灵活交换芯片
- SIGCOMM 2014年CCR提出用户快速可编程语言P4
- SIGCOMM 2017年可编程的Paper和NFV的Paper一样多，成为新的学术热点

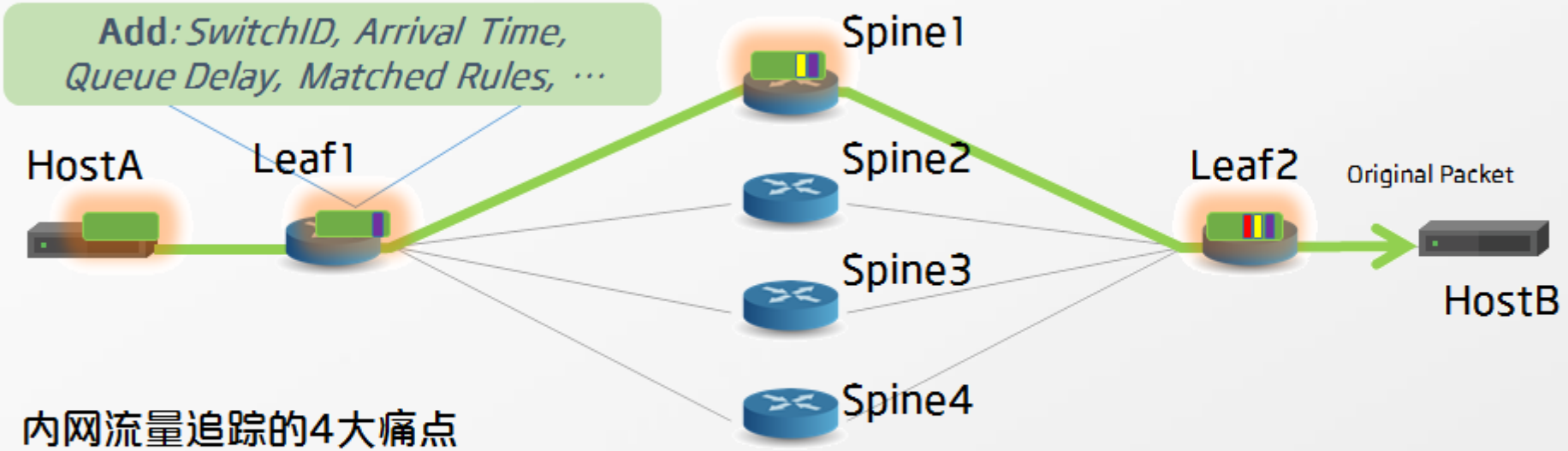
工业界：基于学术的成果

- 2017年，发布P4 Spec 16， 以及对应编译器和模拟器
- 2017年，业界第一款6.5T RMT架构的芯片由Barefoot公司发布
- 2017年，Broadcom公司也积极跟进，会推出有限可编程芯片TD3和TM2

2016	PISCES: A Programmable, Protocol-Independent Software Switch	Princeton University, Stanford University, VMware, Barefoot	基于 OVS 的软件交换机
2016	Packet Transactions High-Level Programming for Line-Rate Switches	VMWare Research, Barefoot Networks, Microsoft Research, Stanford University	表达数据面的调度、拥塞控制、网络度量、队列管理等算法
2016	Programmable Packet Scheduling at Line Rate	MIT CSAIL, Barefoot Networks, Cisco Systems, Stanford University	可编程调度算法
2017	dRMT: Disaggregated Programmable Switching	Cisco, MIT, VMware	RMT 架构的改进
2017	SilkRoad: Making stateful Layer-4 Load Balancing Fast and Cheap Using Switching ASICs	USC, Facebook, Barefoot, Yale	Layer-4 LB
2017	Language-directed hardware design for network performance monitoring	MIT, IIT Guwahati, Cisco, Barefoot	一种硬件原语，用于描述网络性能由交换机支持

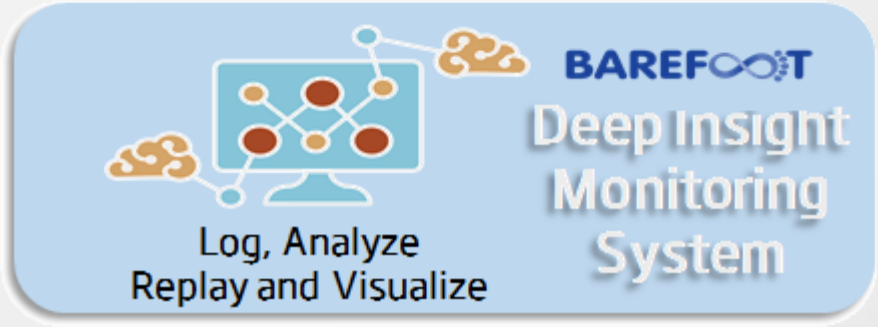
INT（In-band Network Telemetry）报文级可视化

- 每跳进出前后利用强大的可编程能力打上元数据tag



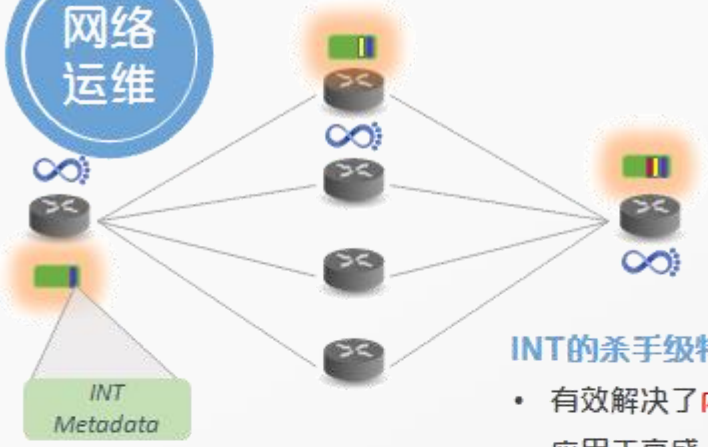
内网流量追踪的4大痛点

1. 走了哪条路
2. 依据哪条规则
3. 每跳逗留了多长时间
4. 一条物理链路有哪些流在共享



02 网络前研 | P4的明星功能

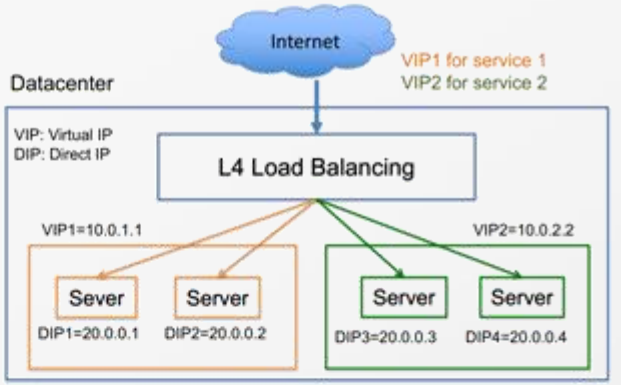
网络运维



INT的杀手级特性

- 有效解决了内网流量的监控问题
- 应用于高盛、纳斯达克等北美重要金融系统中

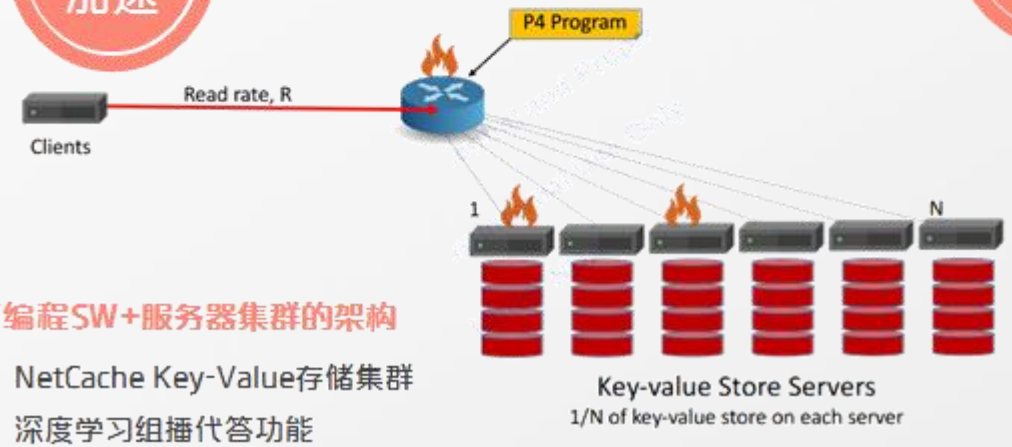
Middle Box



FaceBook的四层负载均衡实现

- 10M的有状态流表连接
- 单片Tofino芯片相当于100台服务器的LB性能
- 类似可以实现DDoS防火墙的功能

集群加速



可编程SW+服务器集群的架构

- NetCache Key-Value存储集群
- 深度学习组播代答功能

功能精简

IPv4 and IPv6 routing
- Unicast Routing
  - Routed Ports & SVI
  - VRF
- Unicast RPF
  - Strict and Loose
- ~~Multicast~~
  - ~~PIM-SM/DM & PIM-Bidir~~

Ethernet switching
- ~~VLAN Flooding~~
- MAC Learning & Aging
- STP state
- ~~VLAN Translation~~

Load balancing
- ~~LAG~~
- ECMP & WCMP
- Resilient Hashing
- ~~Flowlet Switching~~

Fast Failover
- LAG & ECMP

Tunneling
- IPv4 and IPv6 Routing & Switching
  - ~~IP-in-IP (6in4, 4in4)~~
  - VXLAN, NVGRE, GENEVE & GRE
  - ~~Segment Routing, ILA~~

MPLS
- ~~LER and LSR~~
- ~~IPv4/v6 routing (L3VPN)~~
- ~~L2 switching (EoMPLS, VPLS)~~
- ~~MPLS over UDP/GRE~~

ACL
- MAC ACL, IPv4/v6 ACL, RACL
- ~~QoS ACL, System ACL, PBR~~
- Port Range lookups in ACLs

最小功能实现

- 降低复杂性，提升稳定性
- 最大化利用片上资源

03 验证与后续规划

研究计划
验证拓扑与内容
小结与展望

03 研究验证 | 验证计划

开放交换机

- 前期准备工作
- 趋势调研
- 专家讲座

- SoNIC组网
- K8S系统
- INT Telemetry

- stratum项目跟踪
- 基于ONOS Trellis的Underlay+Overlay+Edge的统一组网
- 高级L4-7功能
- INT功能

《基于开放交换机的金融数据中心组网方案》

1-3 | 4-6 | 6-12 | 年底

可编程交换芯片

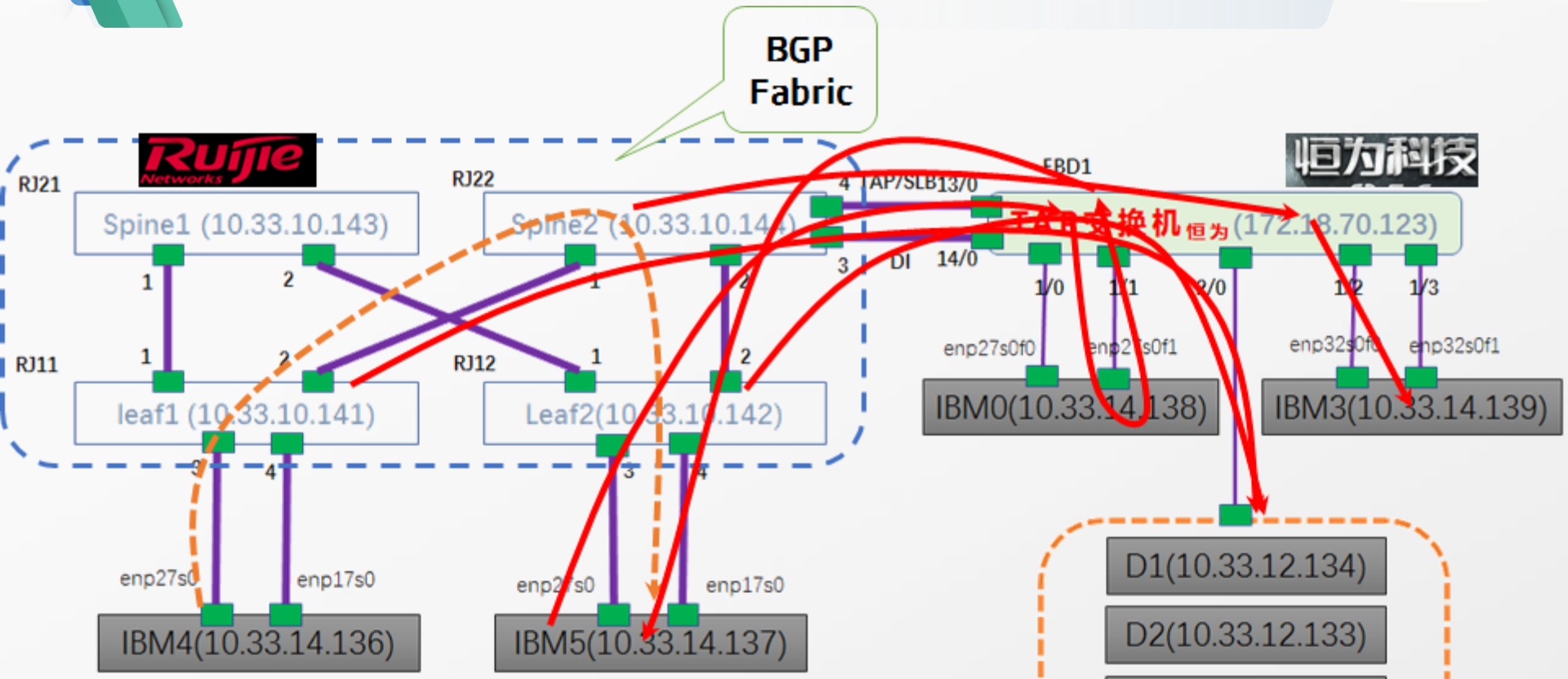
- 搭建开发与测试环境
- 熟悉开发流程与工具
- 编译测试P4示例程序

- 基础FW功能
- 基础LB功能
- 云网监控TAP功能实现

- 高级LB、FW、INT等的实现方式
- 与K8S的组网模型进行对接

《可编程交换芯片在金融数据中心组网中的应用场景与前景》

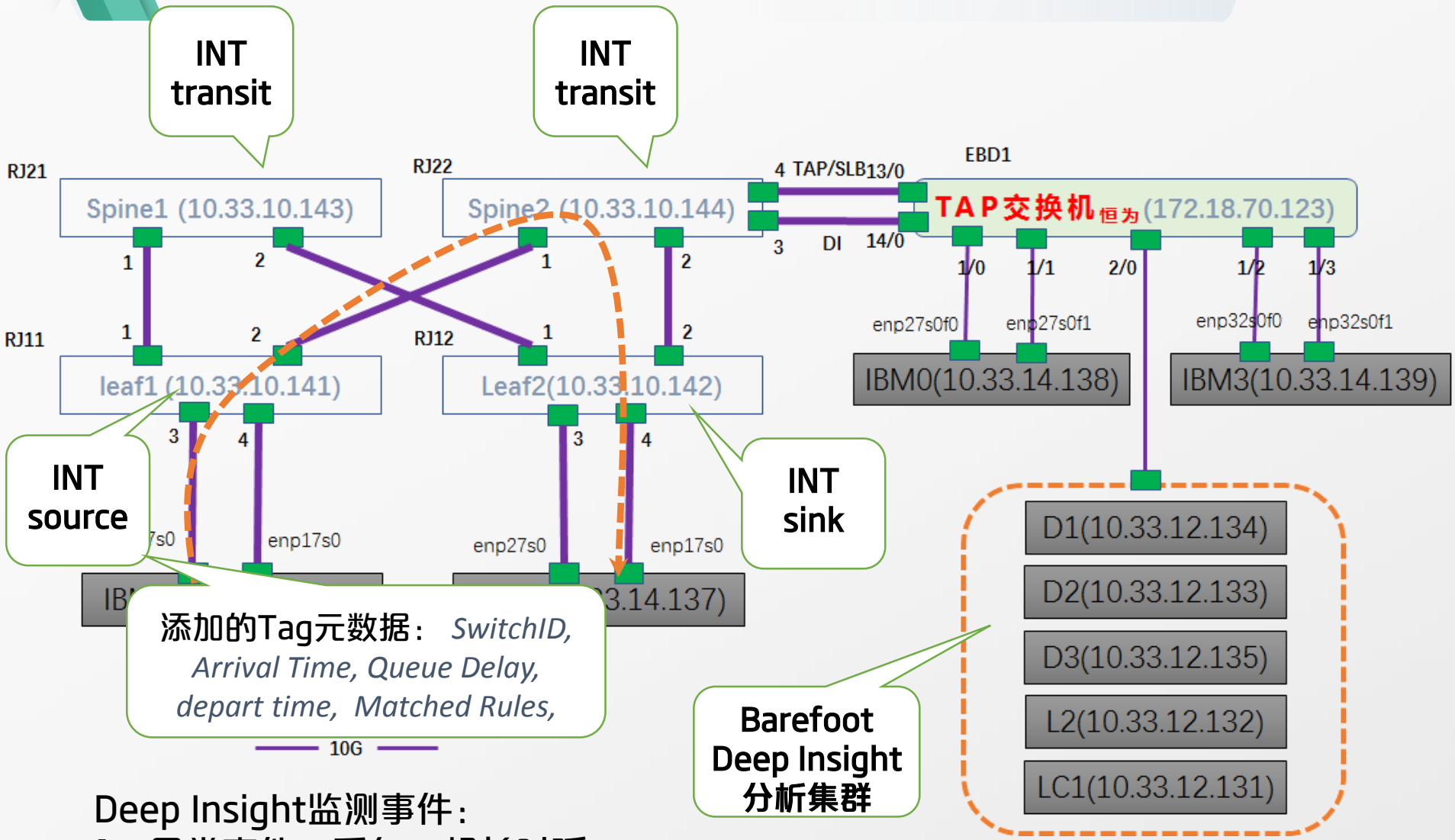
03 前期验证 | 实验拓扑



数据流示意：

1. 业务流+INT
2. 负载均衡流量
3. ERSPAN镜像流量以及TAP分流

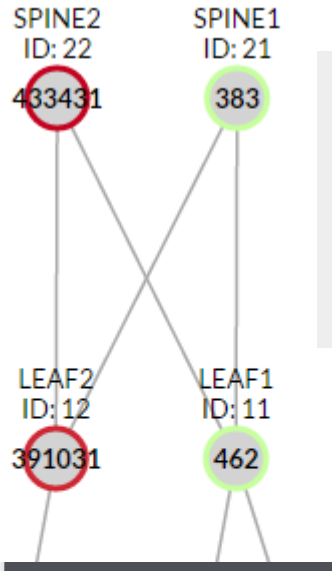
03 研究验证 | SONiC组网+INT遥测

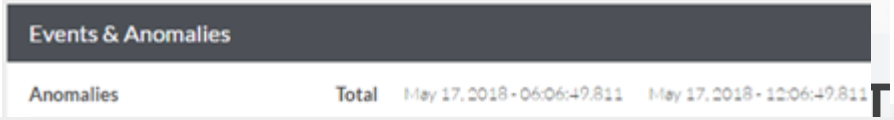


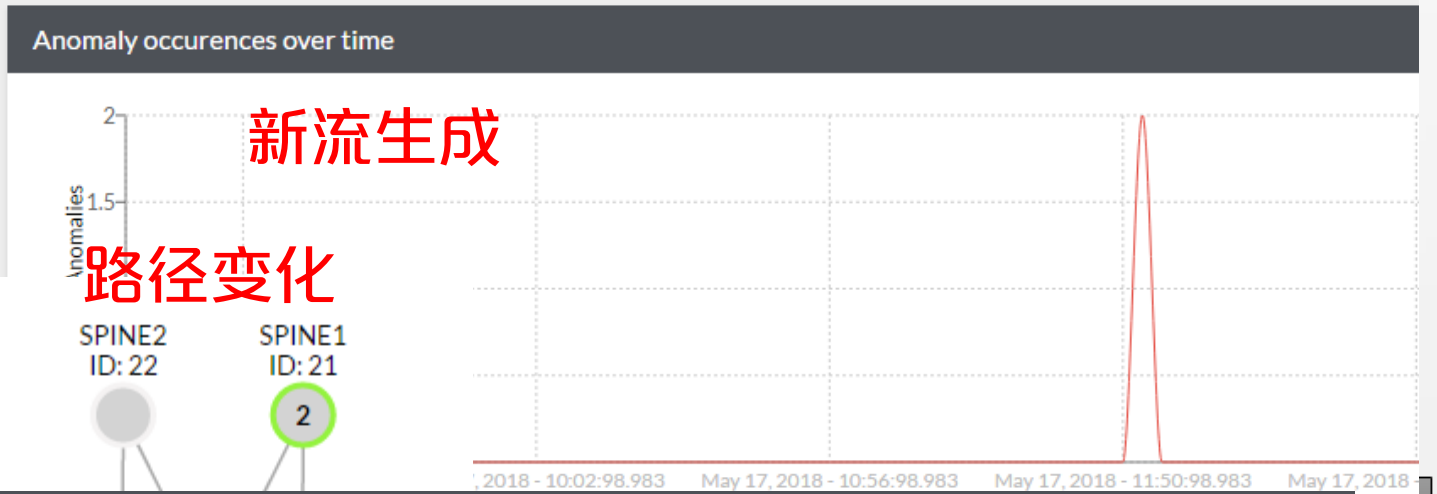
Deep Insight监测事件：

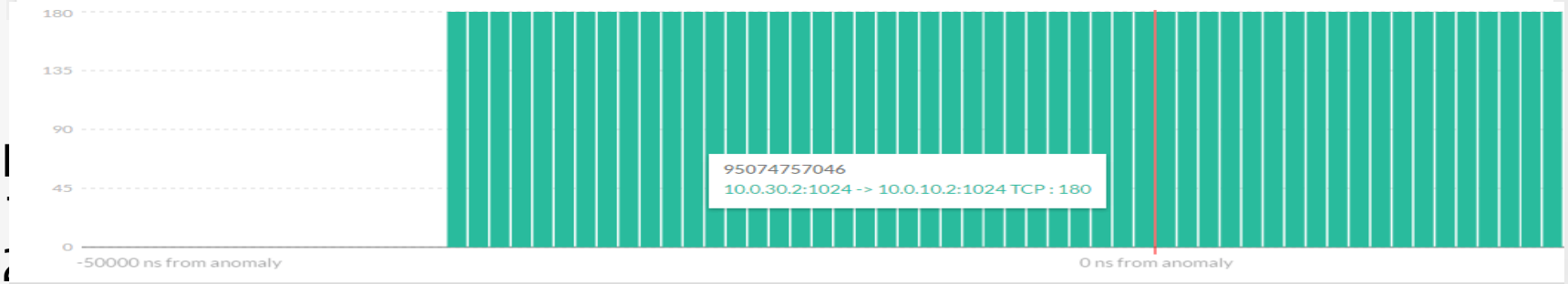
1. 异常事件：丢包，超长时延
2. 其他事件：新流产生，ECMP路径变化

事件汇总



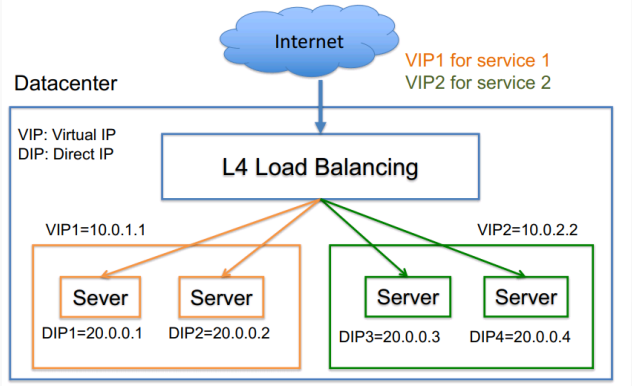




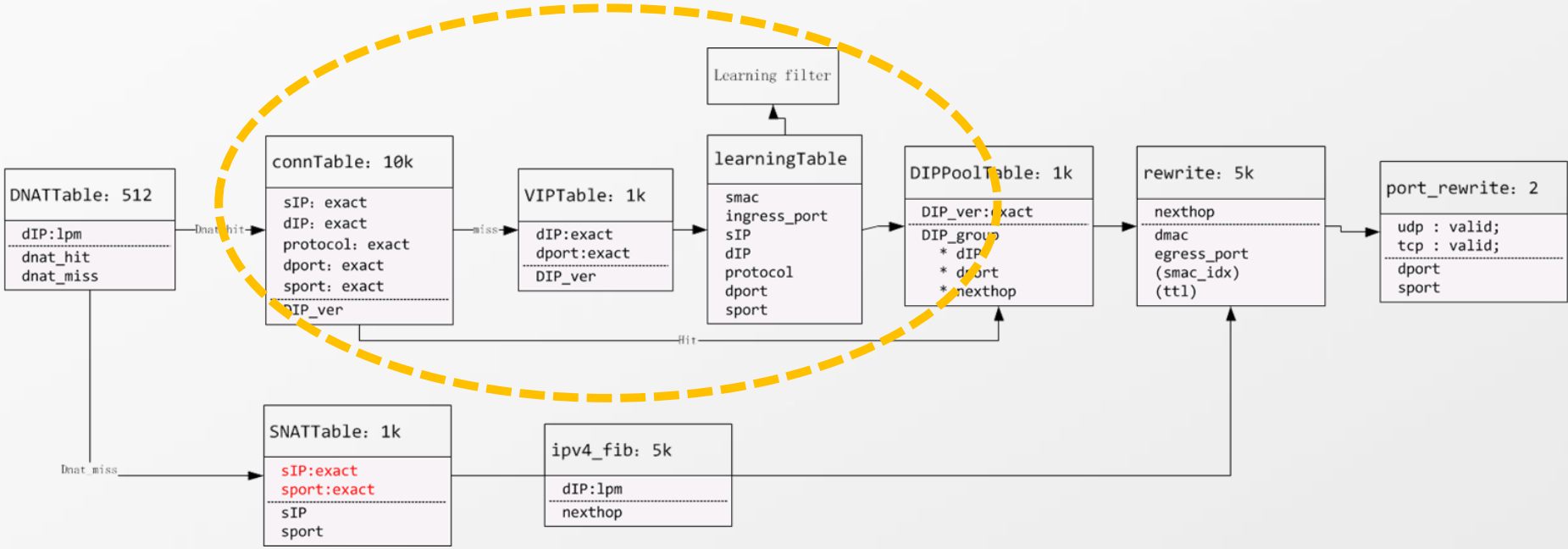


y 17, 2018	10.0.30.2	1024	10.0.10.2	1024	6	1526529325101	144115188075855914	0	12	0	8

03 研究验证 | 四层负载均衡的实现



实现难点：
- PCC（per connection consistency）有状态连接保持



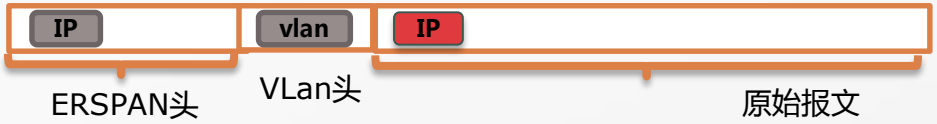
去程：DNAT+hash
回程：SNAT

03 研究验证 | TAP的实现

SDN云网监控的项目中，有两类封装镜像报文

类型1

类型2

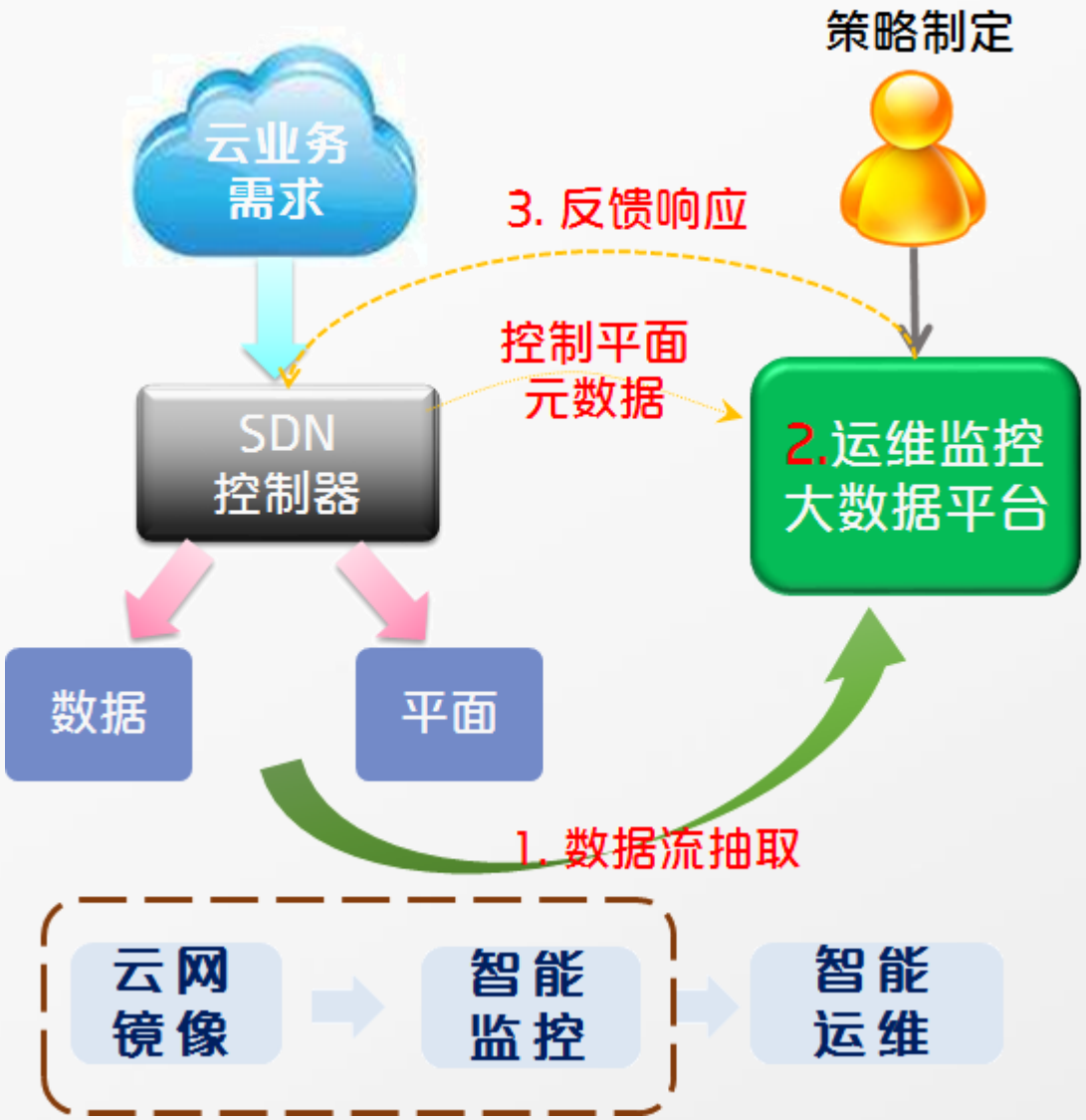


需求

- 镜像网络根据最内层IP进行hash
- （optional）去除ERSPAN头，并提取其中的erspan id，打成vlan tag

核心主干代码

```
control tap_group {
    if (valid(gre)) {
        apply(add_vlan);
        apply(remove_erspan);
        if(valid(vxlan)) {
            apply(inner_vxlan_meta_mapping);
        } else {
            apply(inner_meta_mapping);
        }
        apply(tap_table);
    }
}
```



架构特征

- 数据流提取：按需流量镜像，或者采用最新的带内监控技术
- 技术综合：SDN、大数据、机器学习等开放技术来分析网流数据，未来将有很大的应用前景
- 数据开放：运维开发一体化，可提供网络、应用、安全等各种形态的分析

It is all about how to better control the distributed network

YunShan Networks

IT大咖说 知识共享平台

银联

感谢！