智能时代数据中心网络实践与趋势

锐捷网络

权熙哲

关于我……

姓名：权熙哲　　民族：朝鲜族

主要工作经历：

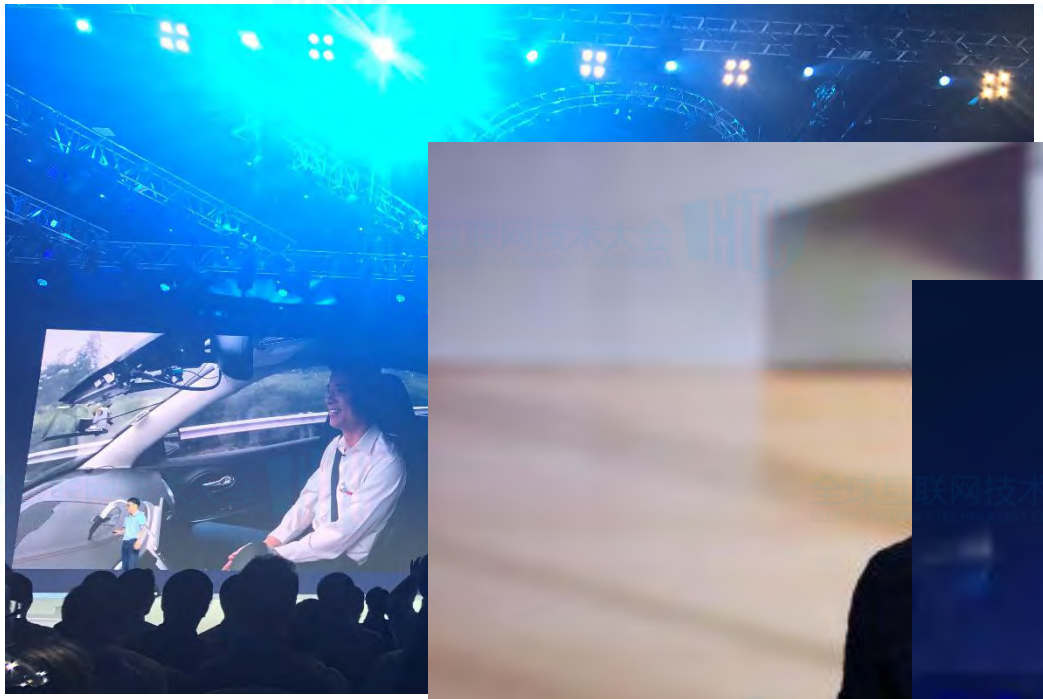
2007年-2010年：中国软件，系统集成

- ✓ 首都机场T3航站楼（安防、办公、商业、无线……）；
- ✓ 国家统计局第二次经济普查（办公、全国专线骨干）；
- ✓ CNGI工程。

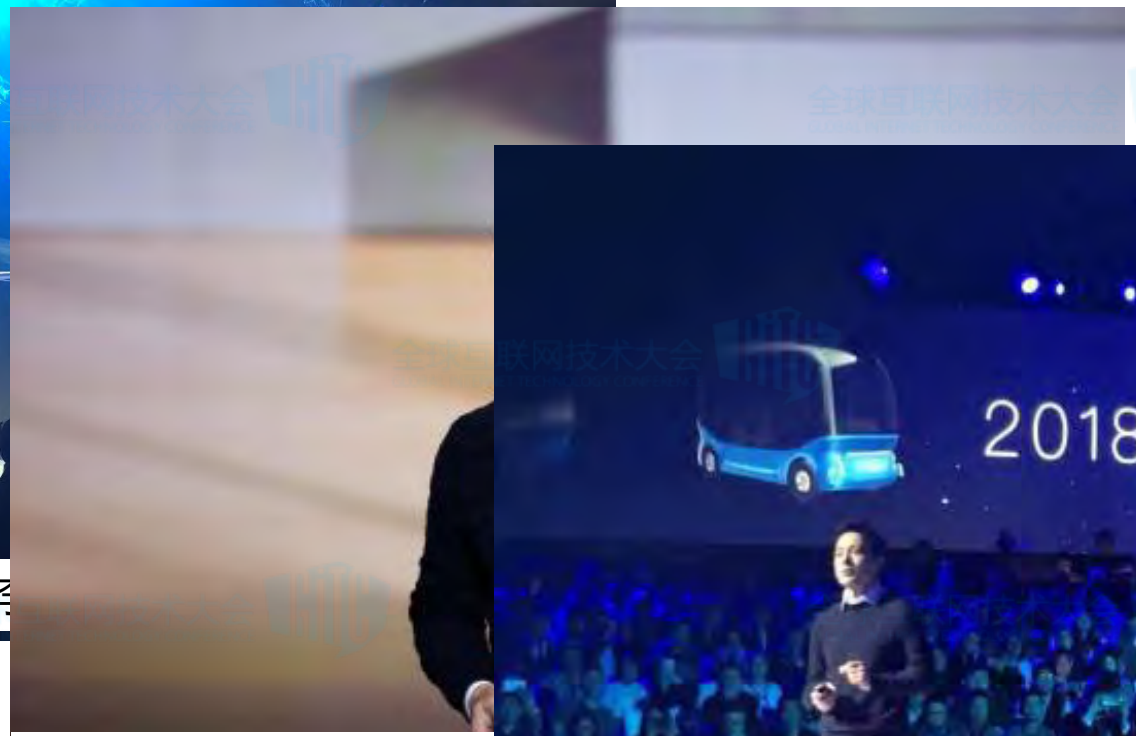
2010年至今：锐捷网络

- ✓ 售后技术服务（2年）：国家某中心、各部委；
- ✓ 售前技术咨询（5年）：政府行业、互联网行业。

无人驾驶汽车小故事



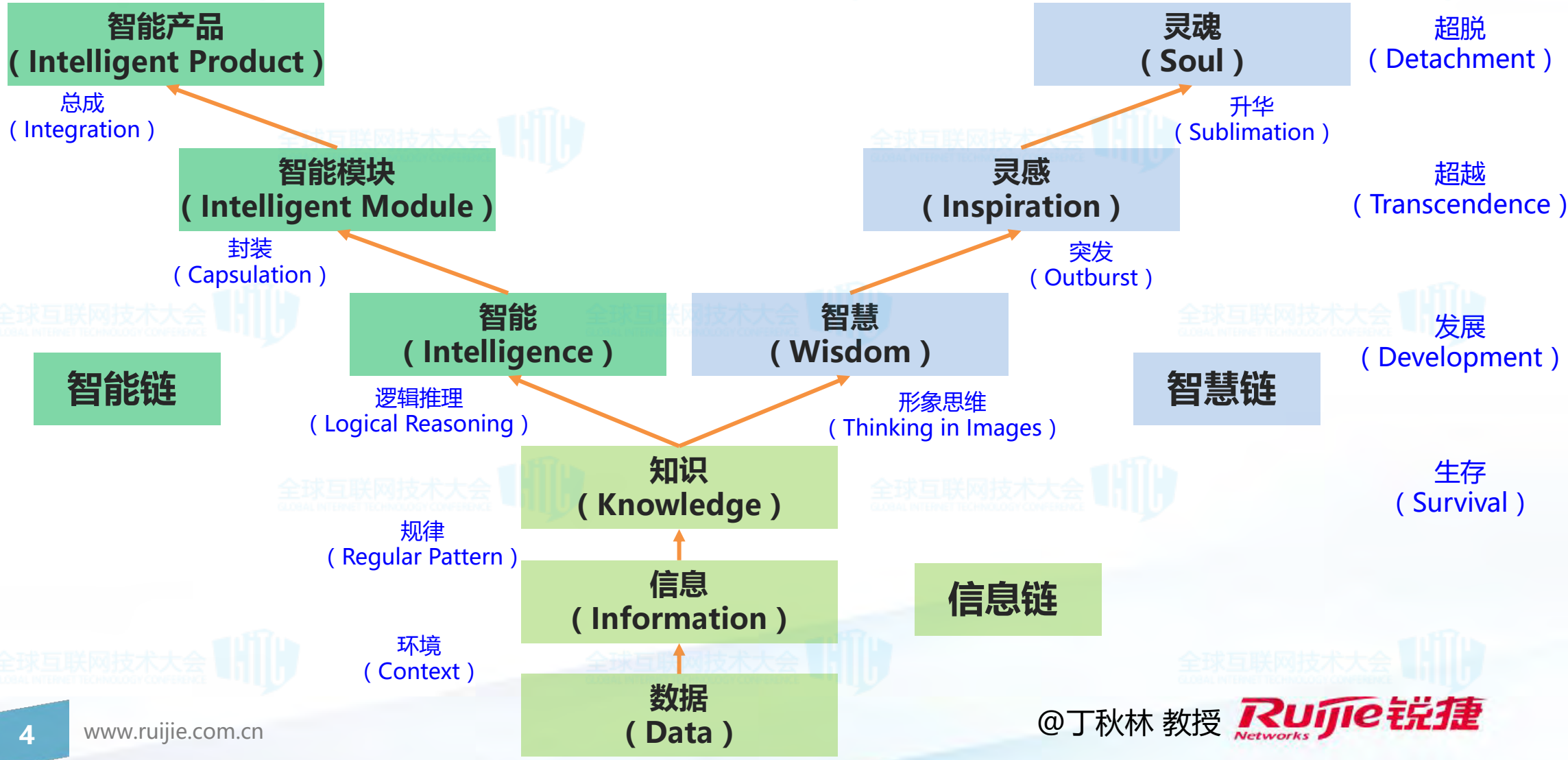
百度AI开发者大会：李



百度世界大会：无人驾驶

百度世界大会：2018无人车实现量产

信息链、智能链、智慧链



大数据、人工智能应用对数据中心网络提出新挑战

人工智能

高性能计算

实时大数据分析

深度学习

分布式计算

分布式存储

计算、存储分离

高性能

低时延

无丢包

业务变化 → 技术引入 → 网络要求

时延的组成

主机处理时延

设备转发时延

重点突破

数据串行时延

重点分析

光电传输时延

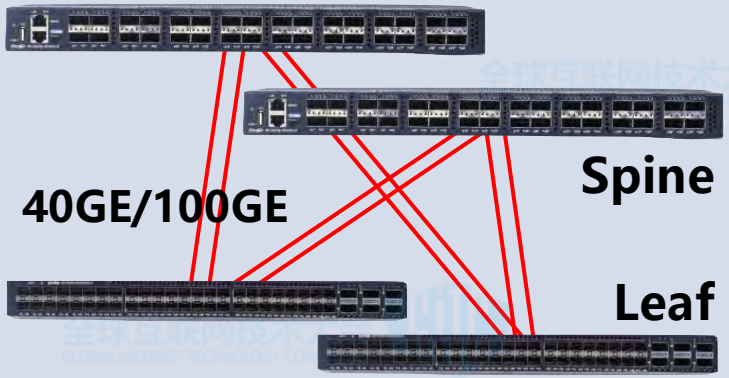
取决于芯片技术，依靠升级芯片来降低时延效果有限

光电传输时延是固定值，没法改变

整合解决方案

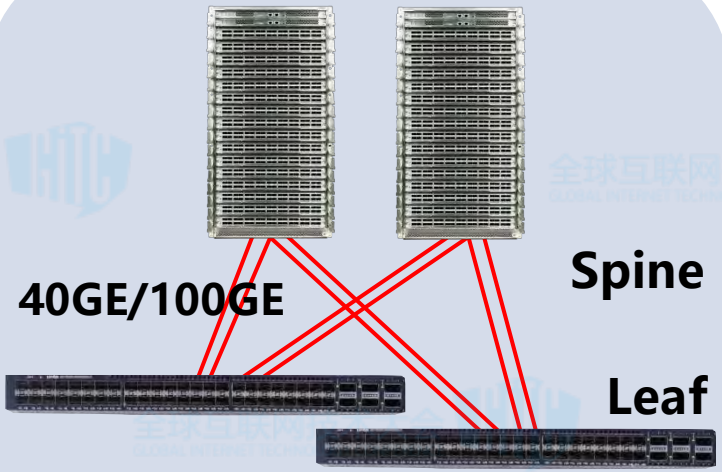
高性能	低收敛比方案（10G/25G网络） 根据集群容量、带宽总需求设计
主机处理时延	RDMA + RoCE 兼顾成本、技术成熟度
网络转发时延 无丢包	PFC + ECN 通过流控技术，避免网络拥 塞造成的业务丢包

10G数据中心网络架构——2级三层架构



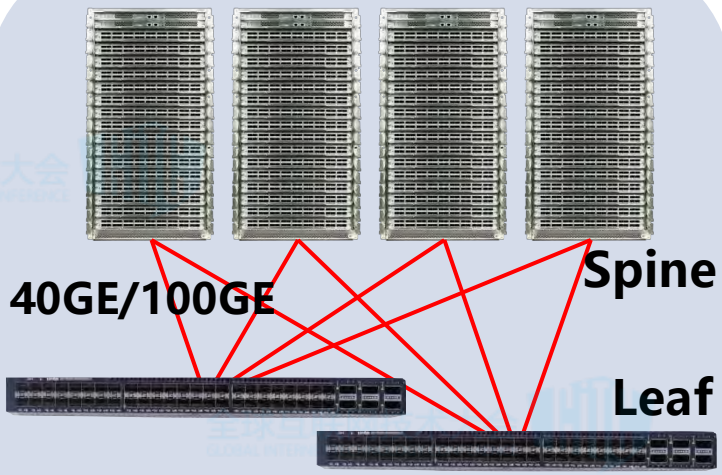
- ✓ 每台TOR **4*40GE/100GE**上联2核心，OSPF组网；
- ✓ 适用集群规模**200~500**台；
- ✓ IDC内交互收敛比**1:1**，集群带宽**2~5Tbps**。

中小型



- ✓ 每台TOR **4*40GE/100GE**上联2核心，OSPF或BGP组网；
- ✓ 适用集群规模**1000~10000**台；
- ✓ IDC内交互收敛比**1:1**，集群带宽**10~100Tbps**。

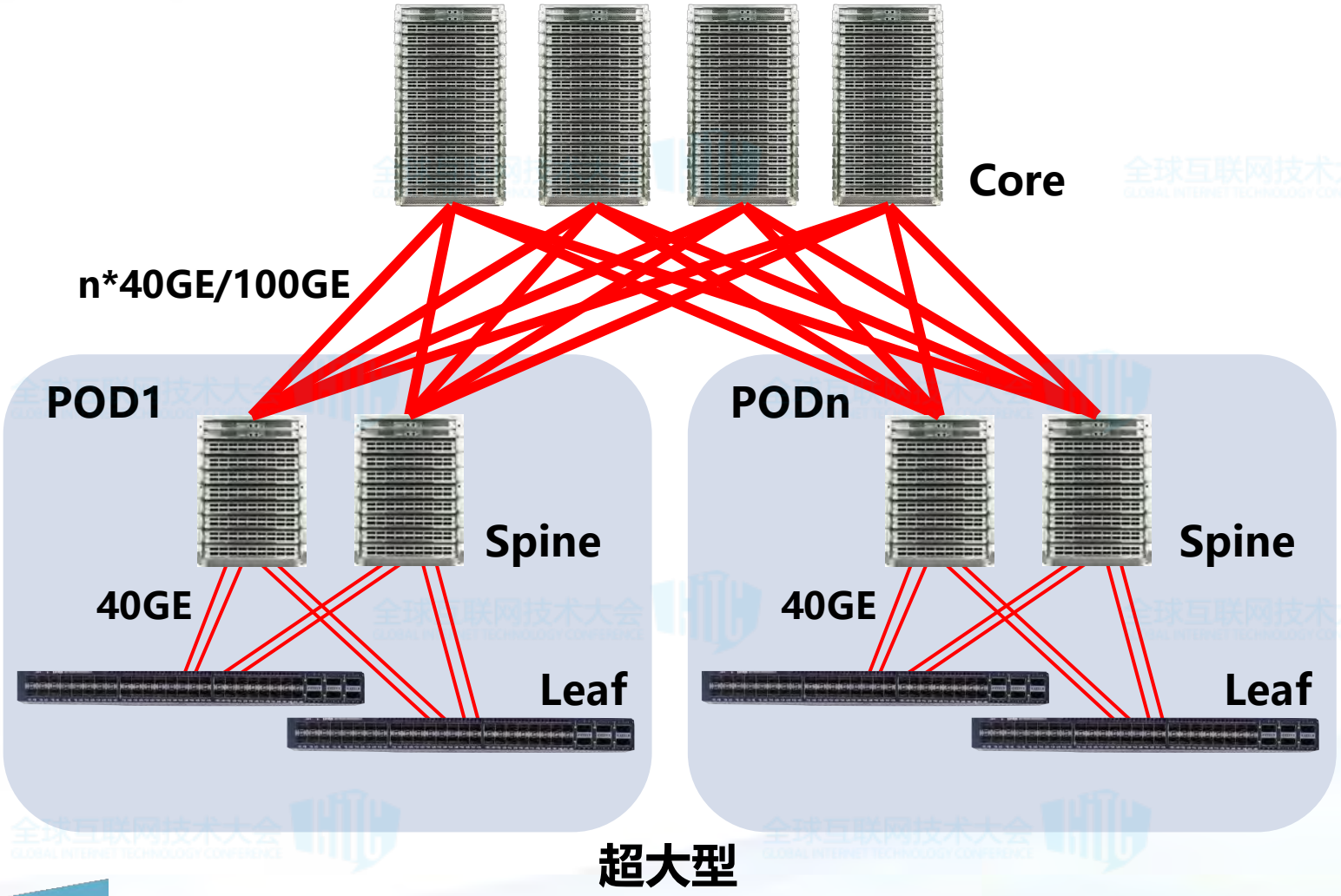
中型



- ✓ 每台TOR **4*40GE/100GE**上联4核心，BGP组网；
- ✓ 适用集群规模**8000~20000**台；
- ✓ IDC内交互收敛比**1:1**，集群带宽**80~200Tbps**。

大型

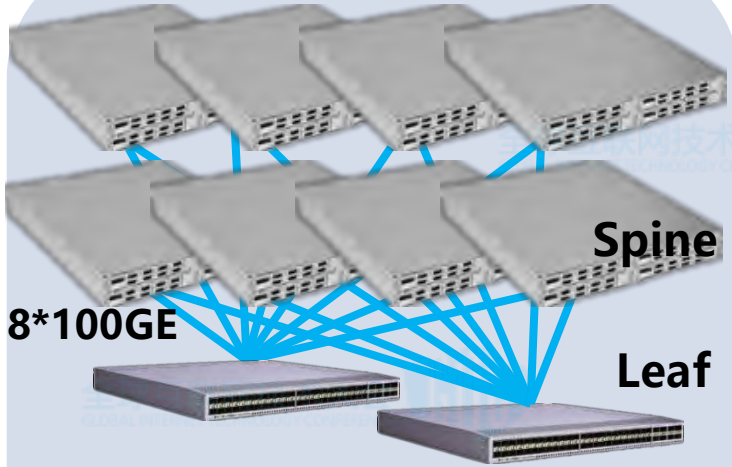
10G数据中心网络架构——3级三层架构



超大型

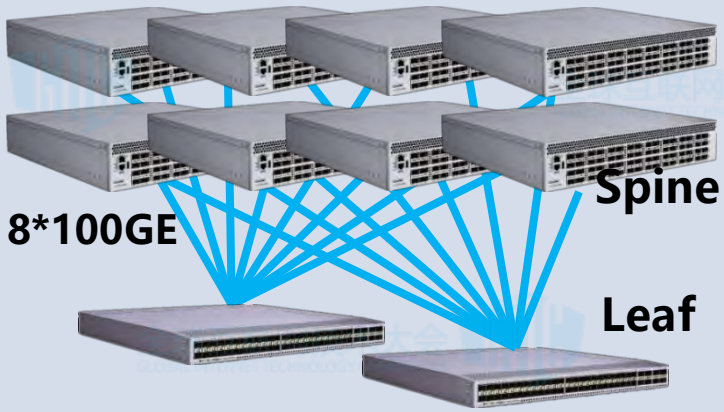
- ✓ 单POD集群规模300~1000台，数据中心集群规模**20000+，BGP组网**；
- ✓ POD内收敛比**1:1**，单POD集群带宽**3~10Tbps**；
- ✓ 上联带宽根据集群规模灵活配置。

25G数据中心网络架构——2级三层架构



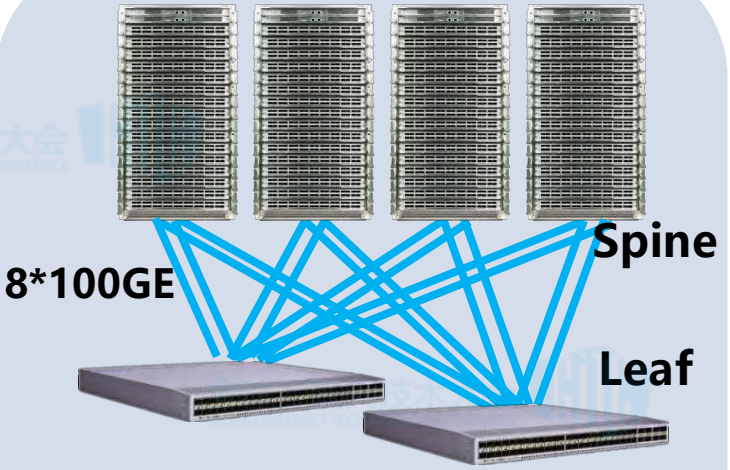
- ✓ 每台TOR **8*100GE**上联8台32口100G BOX，OSPF/BGP组网；
- ✓ 适用集群规模**1000**台；
- ✓ 每台TOR下联32台Servers，IDC内收敛比**1:1** ，集群带宽**25Tbps**。

中小型



- ✓ 每台TOR **8*100GE**上联8台64口100G BOX，OSPF/BGP组网；
- ✓ 适用集群规模**2000**台；
- ✓ 每台TOR下联32台Servers，IDC内收敛比**1:1** ，集群带宽**50Tbps**。

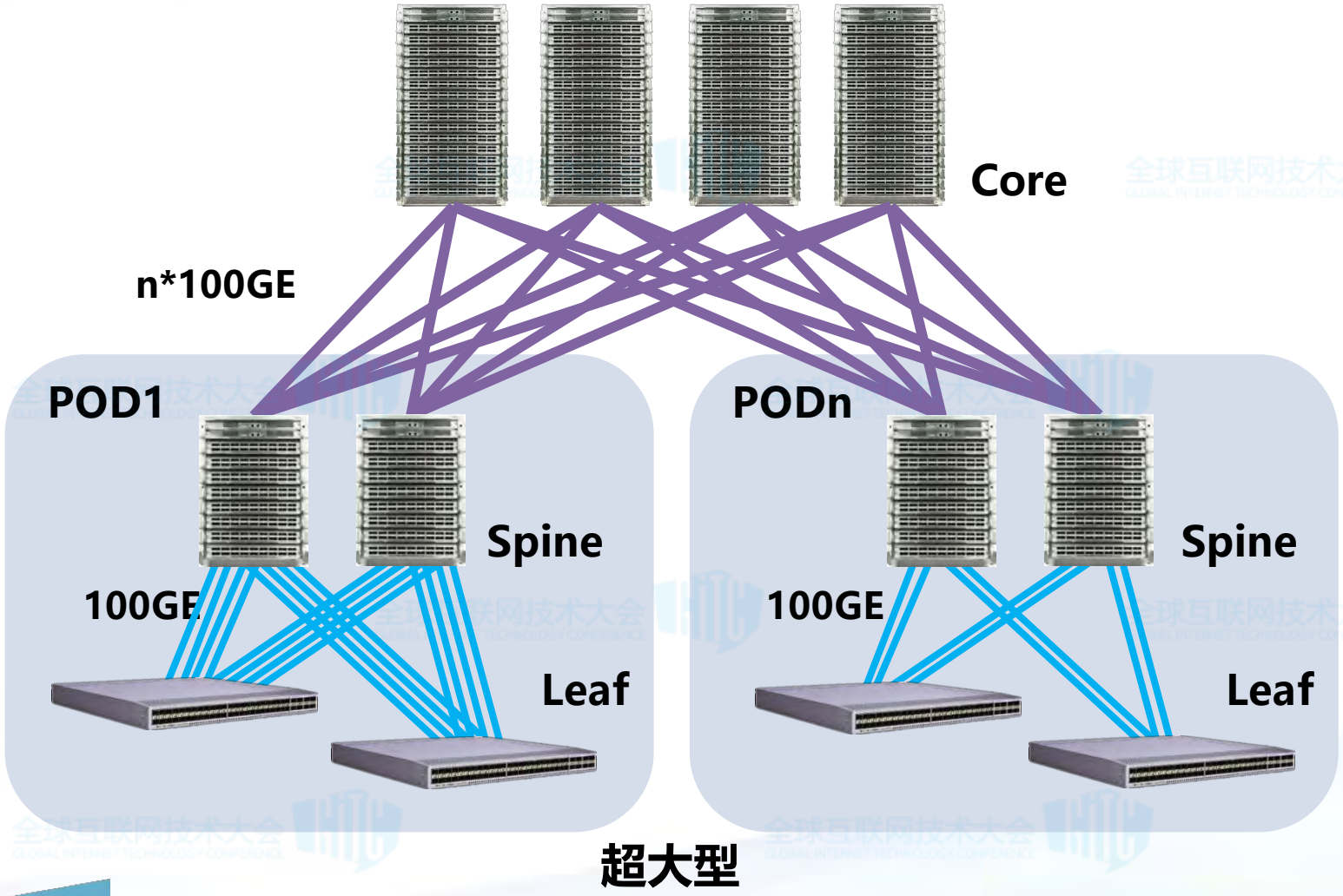
中型



- ✓ 每台TOR **8*100GE**上联4~8核心（机框式），BGP组网；
- ✓ 适用集群规模**2000~18000**台；
- ✓ 每台TOR下联32台Servers，IDC内收敛比**1:1** ，集群带宽**50~450Tbps**。

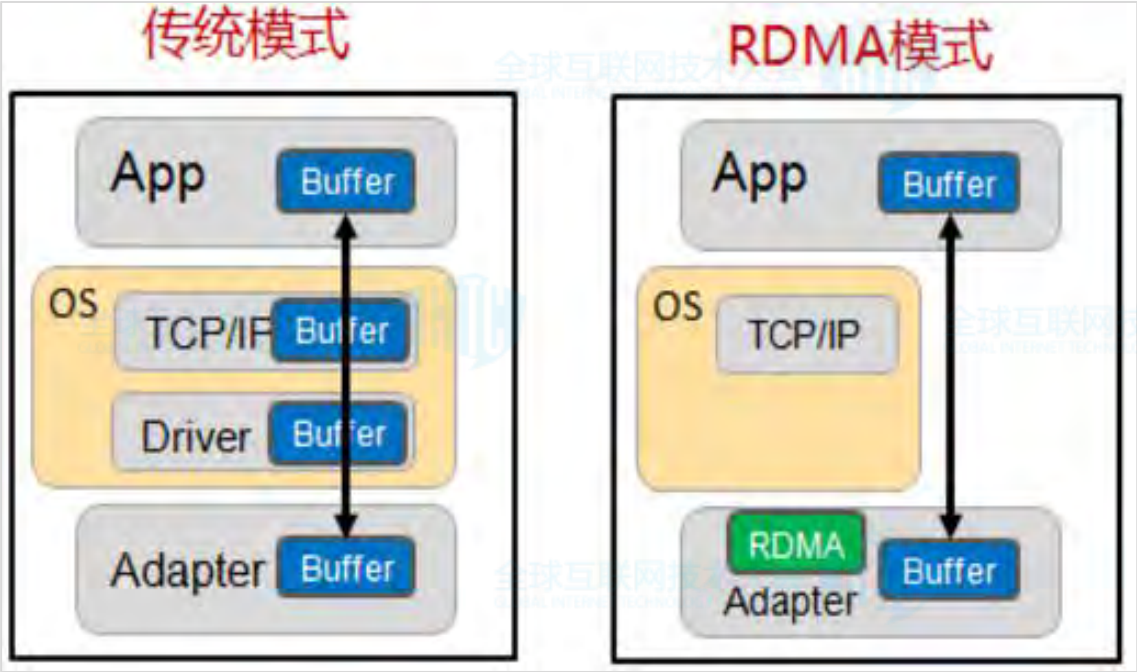
大型

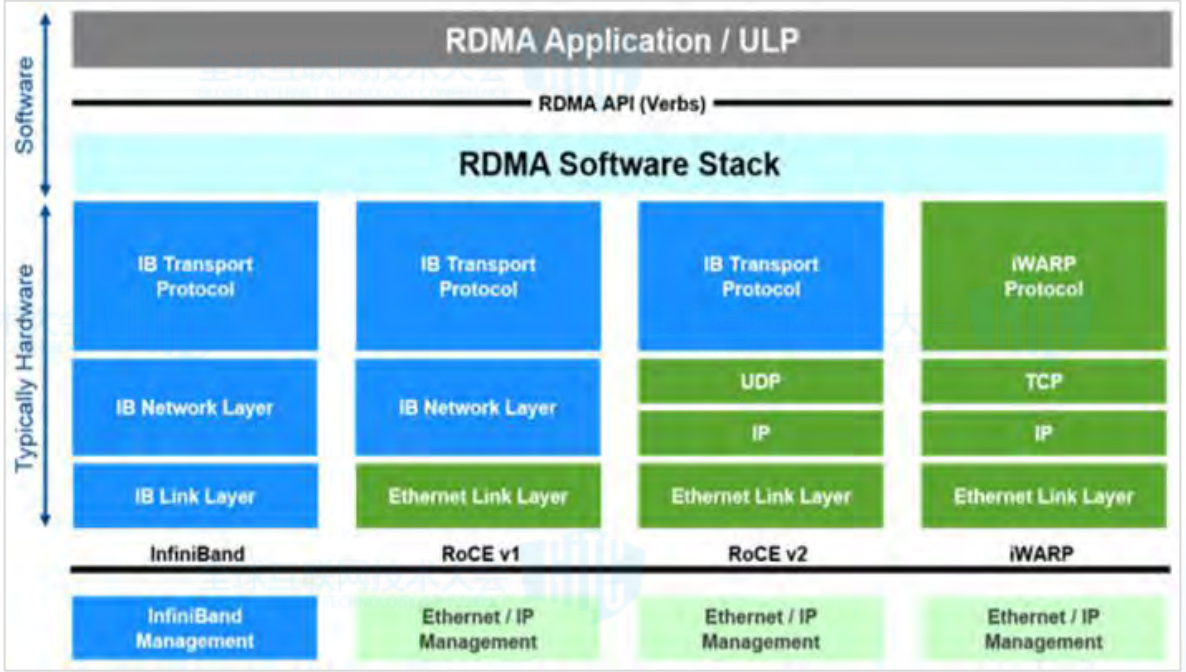
25G数据中心网络架构——3级三层架构



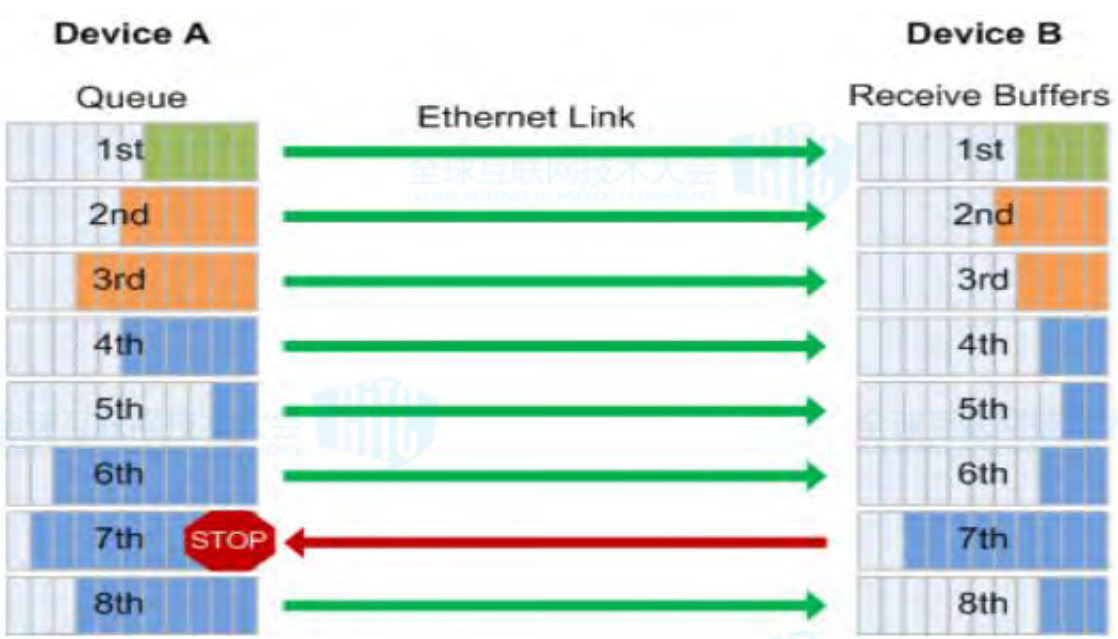
- ✓ 单POD集群规模1000~2000台，数据中心集群规模**20000+，BGP组网**；
- ✓ POD内收敛比**1:1**，单POD集群带宽**25Tbps**，总集群带宽**500Tbps+**；
- ✓ POD内收敛比和上联带宽根据集群带宽需求灵活配置。

主机处理时延——RDMA & RoCE v2





低时延无损网络——PFC功能介绍



PFC 机制将以太链路上的流量区分为不同的等级，基于每条流量单独发送“不许可证”，说明如下：

1）如果本设备所有优先级的流量都没有拥塞，则不发送任何信息给对端发送，对端可以正常发送流量；

2）如果本设备的某一优先级的流量出现了拥塞，则向对端设备发送信息；

3）拥塞结束，停止发送“不许可证”，对端可以正常发送流量，避免了丢包的发生；

4）对于二层报文，其优先级来源为802.1p优先级；对于三层报文，其优先级可以通过将DSCP优先级映射成8个优先级来获取。

优势：

相对于pause帧而言，PFC可以将链路虚拟出几条不同等级的虚拟通道。这样当某条通道出现拥塞后不会影响其他通道。

劣势：

1）只在两台设备的端口之间作用；

2）PFC属于逐级反压，会有较大的延迟。

低时延无损网络——ECN功能介绍

ECN（Explicit Congestion-Notification，显式拥塞通告）主要在TCP报文流的场景中应用，利用IP报文头部中的ECN标志位，在设备中出现拥塞时，对于支持ECN标记的报文，将ECN标志位设置为CE状态。TCP报文的接受方检测到报文中存在CE标志状态时，会在随后的ACK报文的TCP头中设置ECN-Echo标志位来指示拥塞。当发送端接收到该ACK时，就可以根据其ECN-Echo标志位来判断出网络链路上发生了拥塞，从而可以做出相应的调整。

优势：

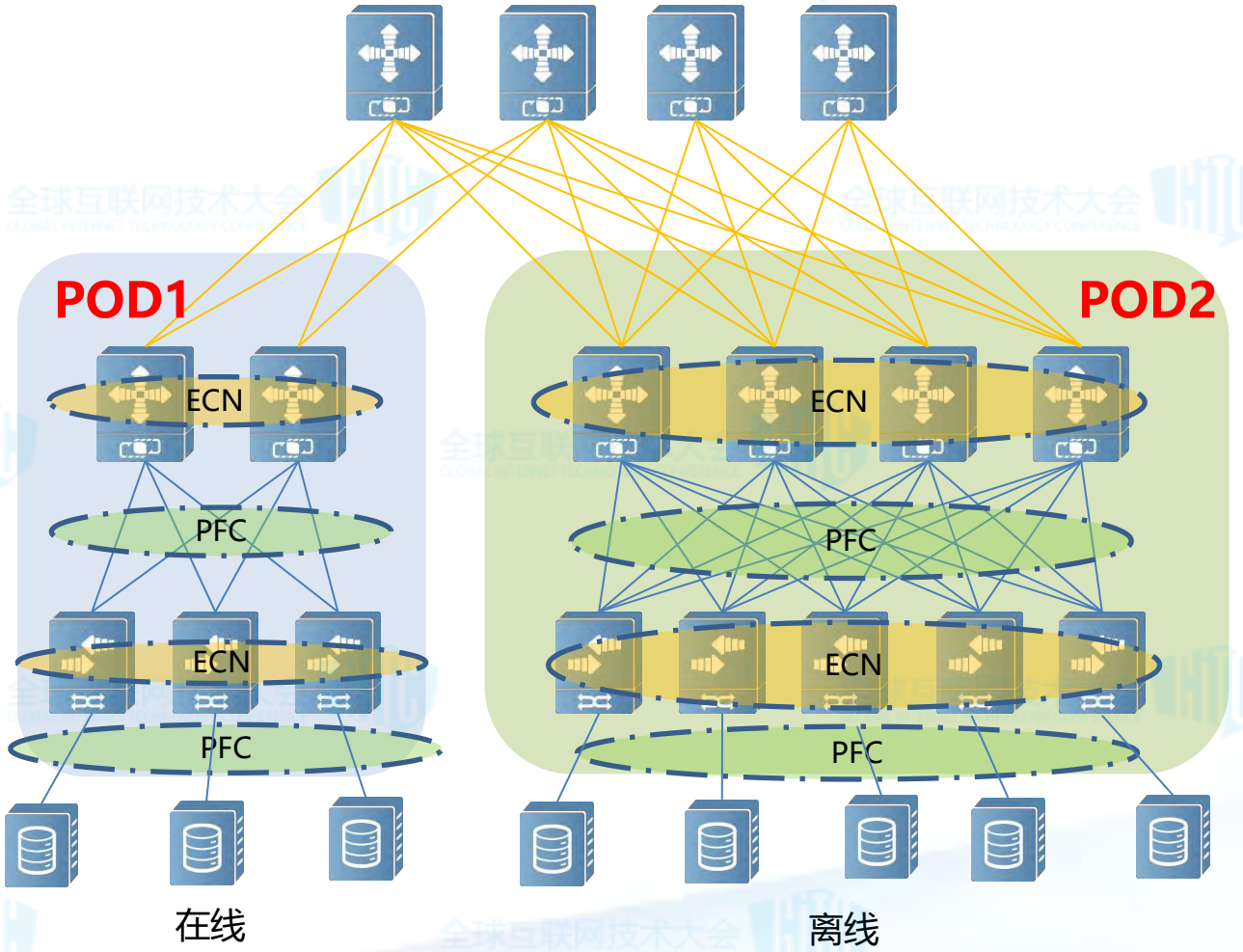
1）该功能基于IP协议及TCP协议，在TCP连接的两端发送方和接受方上发生作用，从而达到流控的目的；

2）实现了端到端的流控，在丢包发生前就主动进行拥塞控制，避免了由于丢包导致的TCP流的慢启动，维持TCP流的稳定，有效避免拥塞。

劣势：

当服务器A发现通路有拥塞的时候，会减少发包，但是实际上这个通路上的拥塞并不是由服务器A的业务造成的。

低时延无损网络应用架构



PFC & ECN 功能改进，提升运维效率

- 统计功能加强（每个port的所有queue）

1. ingress和egress方向的drop count；
2. 发送/接收的PFC个数；
3. PG peak headroom值；
4. egress buffer超过水位和门限的次数；
5. 报文被Mark ECN标记个数；
6. ingress和egress方向的buffer监控。

- 异常情况可以告警：

1. Incast：a) 网卡侧持续发送大量CNP报文（网络incast导致，ECN流控介入，属正常现象）；
2. burst丢包：网卡侧没有持续的发送或收到CNP报文，但交换机ingress或egress方向的RDMA流量有丢包（说明PFC或ECN门限设置不合理）；
3. slow receiver symptom：网卡主动发送PFC pause（正常情况下网卡应该发送CNP来通知发端降速，如果出现网卡主动发PFC属于异常，需要报警）；
4. PFC storm：整个pod下的交换机都有收发PFC pause（PFC storm会引发严重故障，需要严格监控）；
5. PFC deadlock：网卡持续收到PFC pause, 或交换机持续发送PFC pause。

智能运维技术——Buffer水线可视化

可与SDN对接以实时监控Buffer占用情况

Agent生成统计表，并组装多个报文统计信息后上报

Collector (PktStats Client)

统计信息

Network

统计信息

Local CPU (PktsStats Agent/Client)

8 packets required to DMA 8K queues

PCIe

告警水限

TRIDENT 3

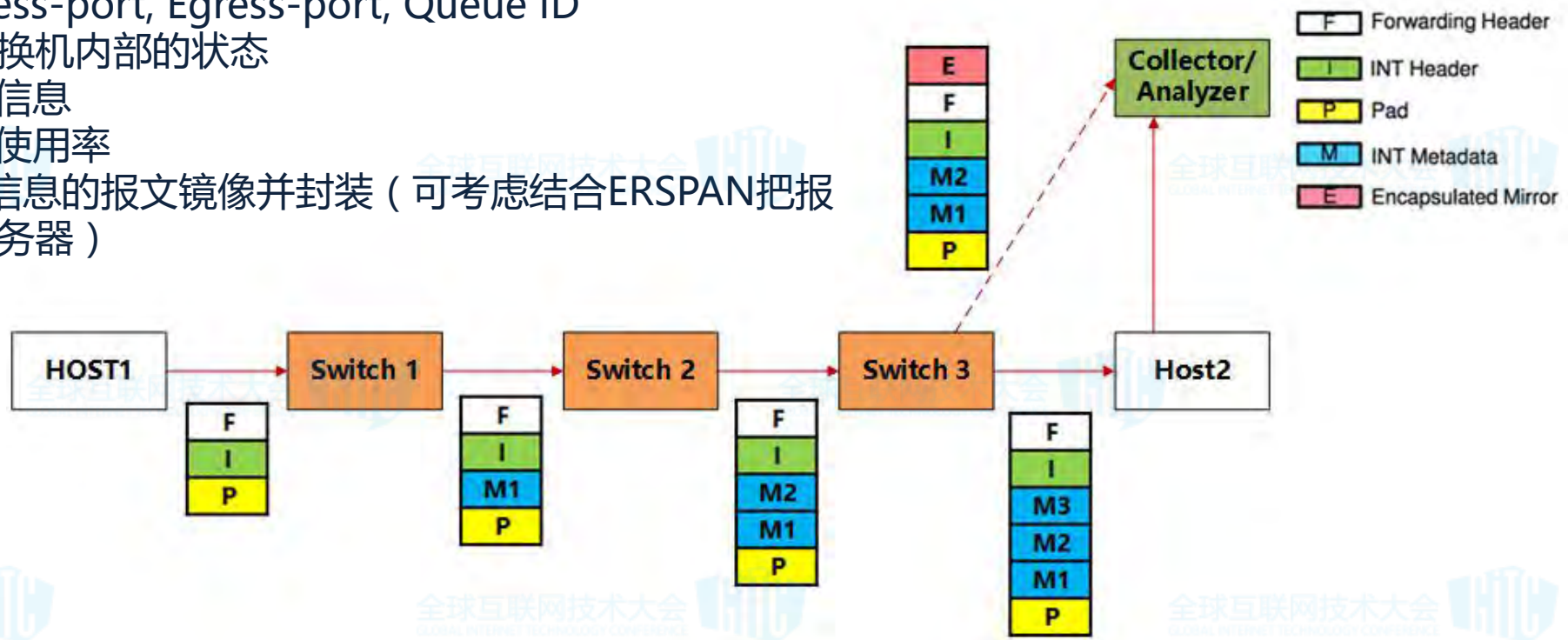
Collector基于*消息类型*或*序列号*对负载(payload)重新排列，并从中提取计数。

Ruijie锐捷 Networks

智能运维技术——报文路径可视化

在报文传输节点添加 INT信息，从而可以：

- 确定报文去向、传输耗时
  - ✓ 添加Switch-ID，时间戳，Residence-time
- 确定报文在每台交换机的选路信息
  - ✓ 添加Ingress-port, Egress-port, Queue ID
- 确定报文在交换机内部的状态
  - ✓ 添加拥塞信息
  - ✓ 添加链路使用率
- 对添加了INT信息的报文镜像并封装（可考虑结合ERSPAN把报文发到分析服务器）



智能运维技术——运用AI技术的网络运维

- 流量特征自分析
- 流量模型自学习
- 流量转发自调度
  - 故障自诊断
  - 故障自恢复

基于AI技术的网络自动化运维已启程

祝大家可以早日实现一边喝着咖啡一边运维

我们锐捷将会为此不断创新产品和方案

最后……

互联网行业网络解决方案主流供应商，服务的互联网企业超过200家

深入业务创新方案，针对互联网IDC、办公网、CDN、商业Wi-Fi等场景推出方案并实现规模应用，获得客户好评

产品及解决方案广泛应用于百度、阿里、腾讯、奇虎360、今日头条、网宿科技、爱奇艺、美团等互联网企业

数据中心核心产品全面应用于阿里巴巴、腾讯、奇虎360、爱奇艺等互联网企业

THANKS 欢迎加v交流

锐捷网络股份有限公司
地址：北京海淀区复兴路29号中意鹏奥大厦东塔A座11层　邮编：100036
Office Tel: 010-51715999　　Fax: 010-51715872

www.ruijie.com.cn

权熙哲

北京 通州

扫一扫上面的二维码图案，加我微信