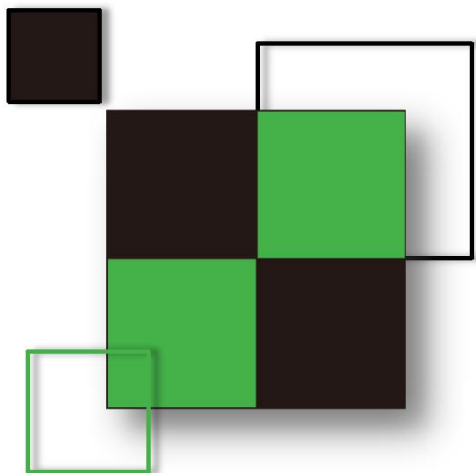


QUIC 协议在分布式系统架构中的实践

OPPO / 李龙彦

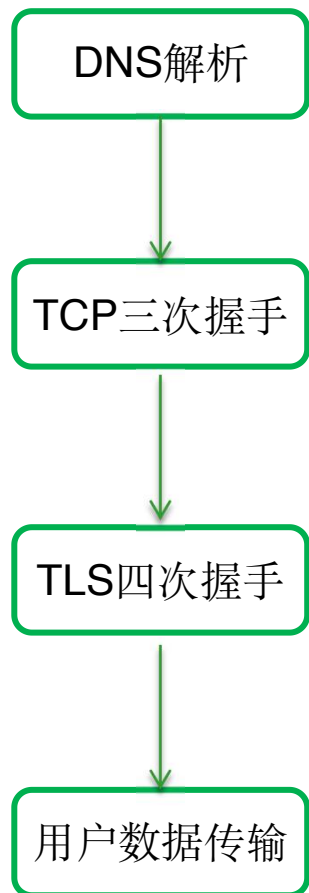
OPPO网络优化领域技术专家，多年网络协议栈开发经验。目前主要负责OPPO的移动端网络库、接入层安全的架构设计和研发工作。从0-1建设了OPPO特色的QUIC（HTTP/3）协议，并成功在OPPO各个业务的进行上线，对网络传输效率以及安全有较大的提升。

目录

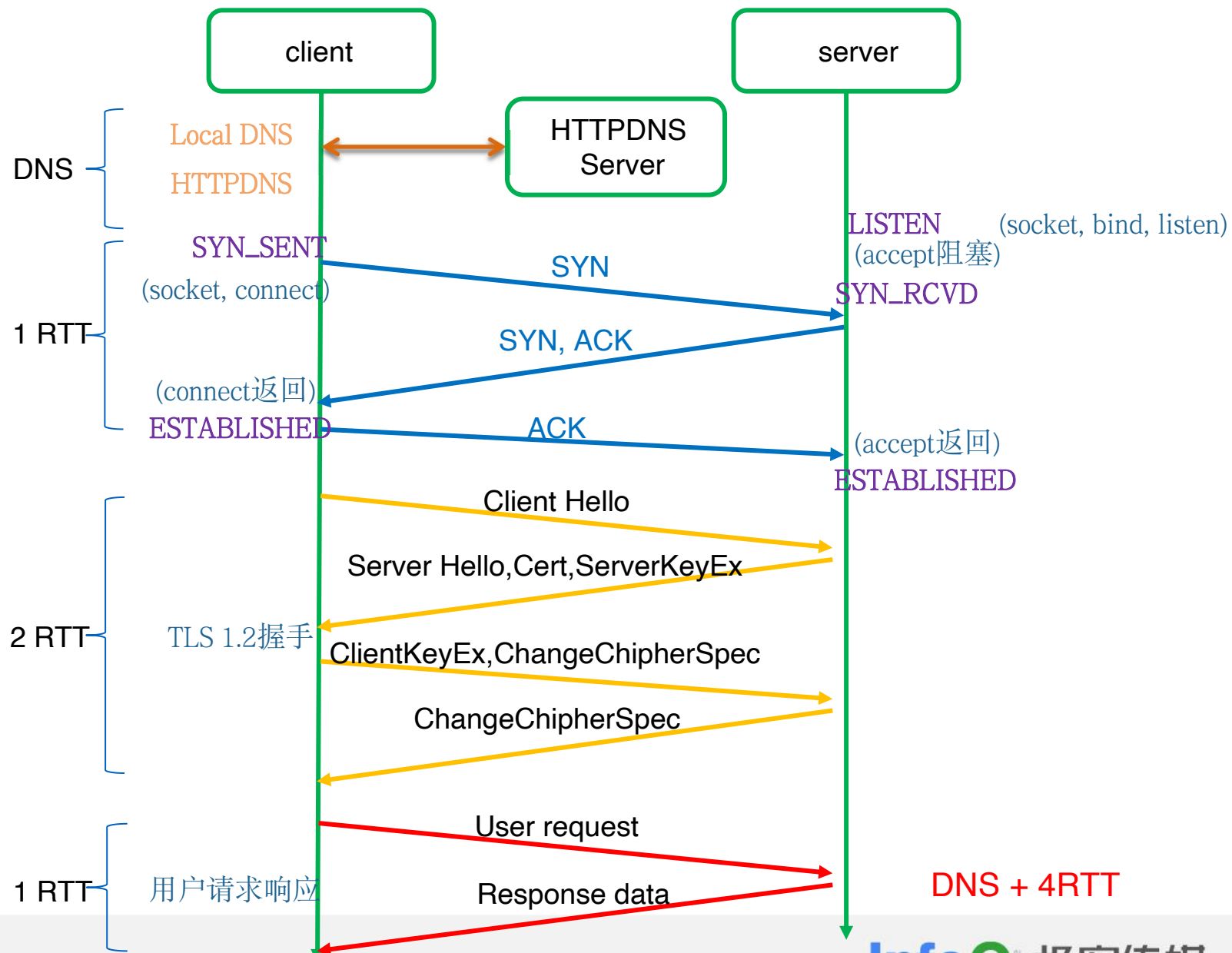


- 01 . Quic协议介绍
- 02 . OPPO分布式系统架构
- 03 . Quic协议在分布式架构中的优化
- 04 . Quic协议实现中的问题案例分享

从一个小请求说起



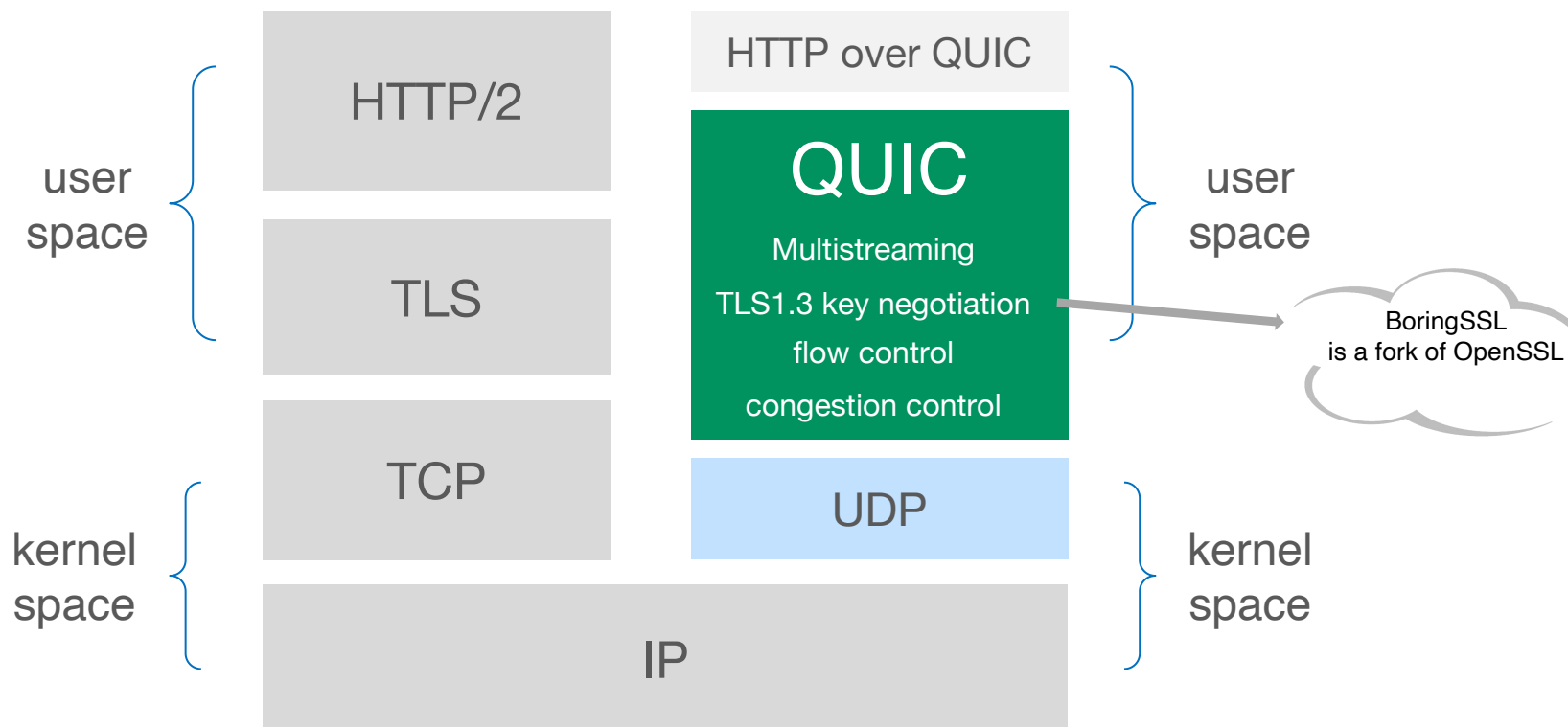
一个请求的过程



01

QUIC协议介绍

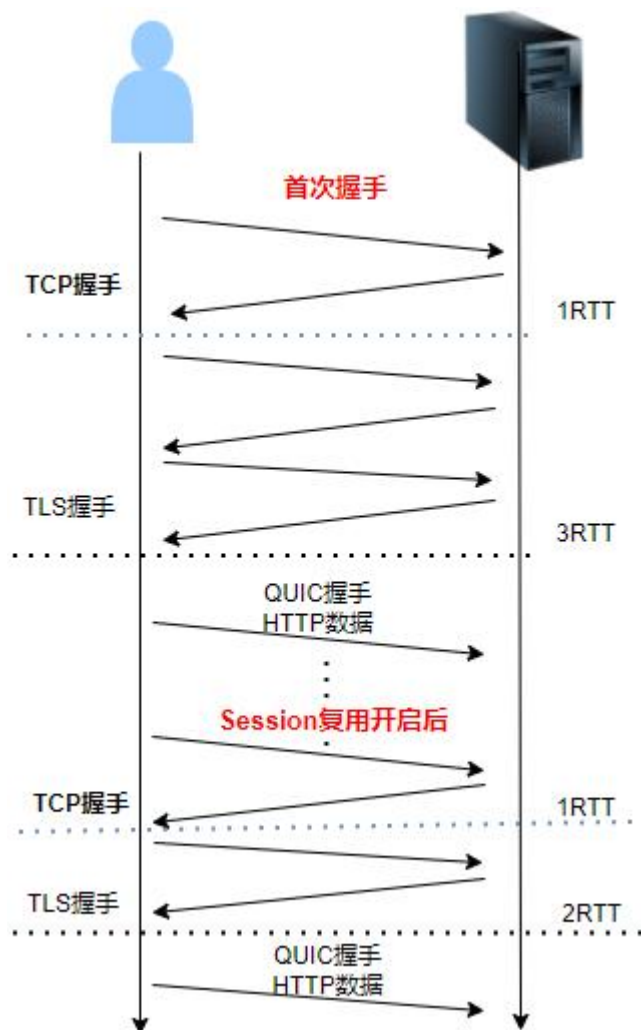
QUIC介绍



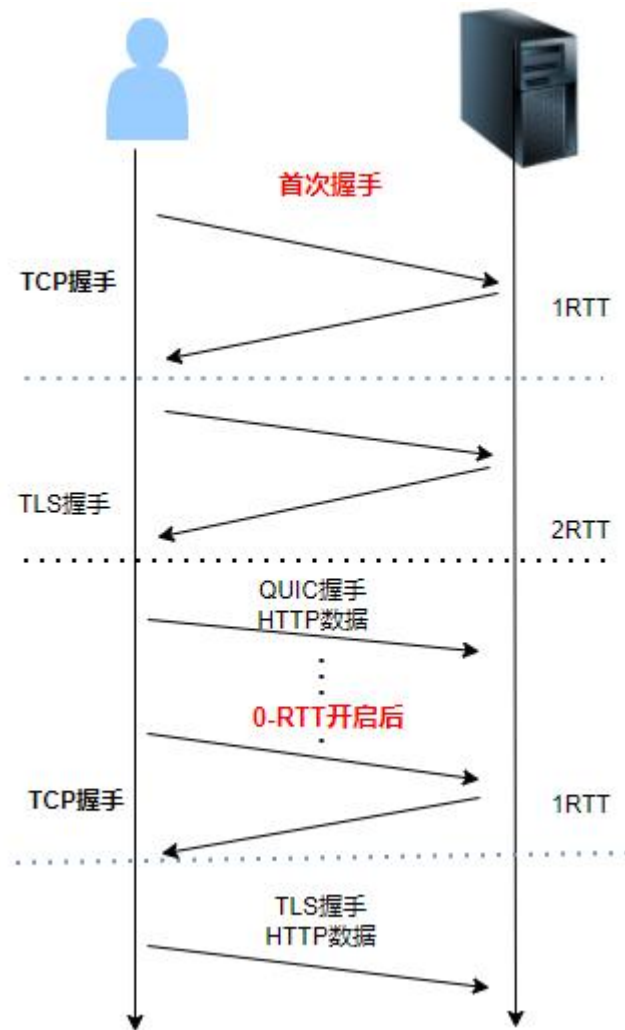
Quic与TCP的对比



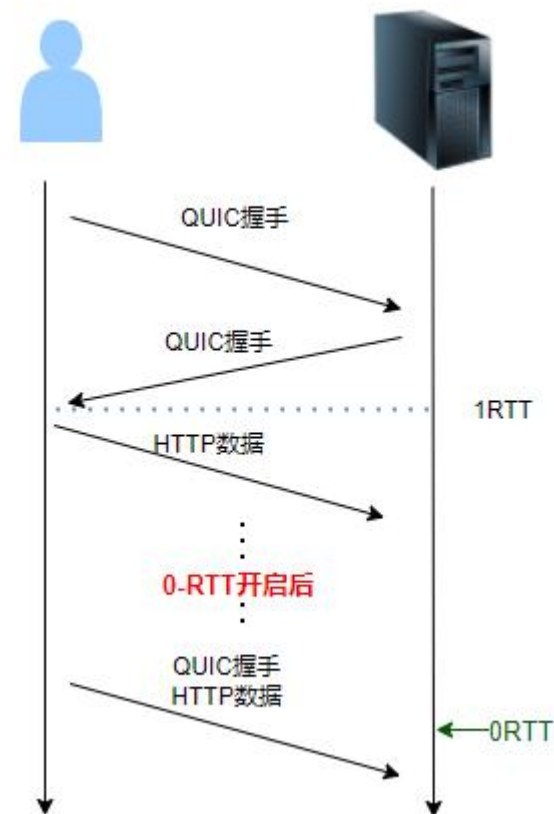
QUIC-0RTT建连



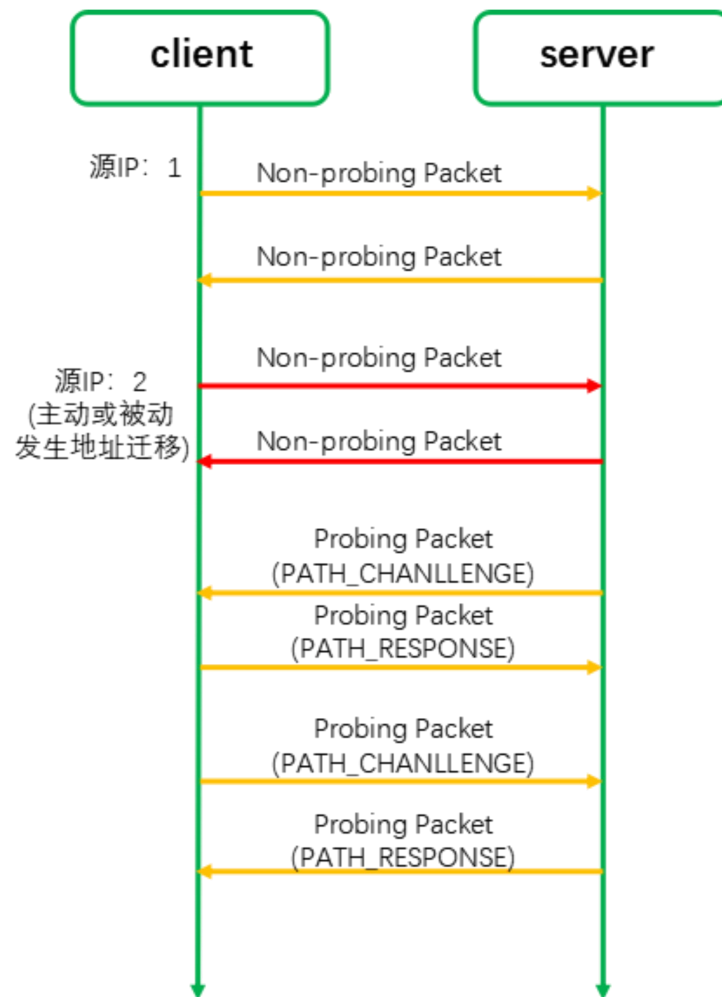
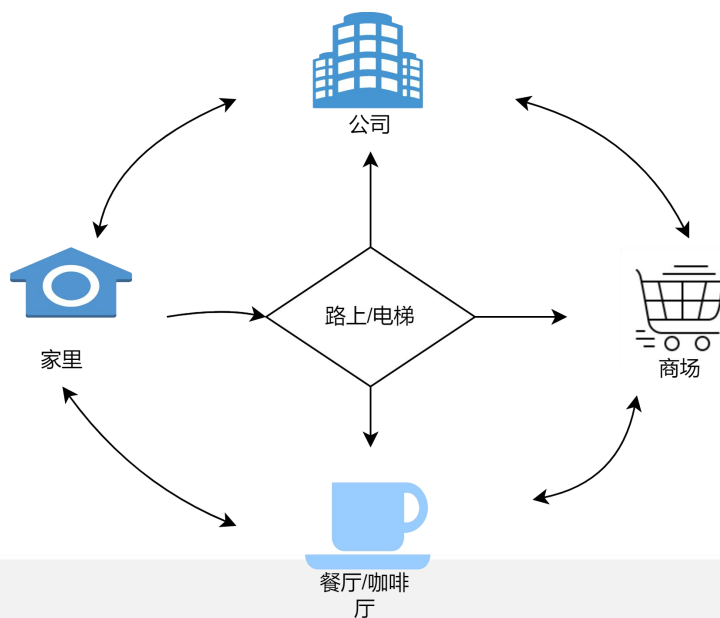
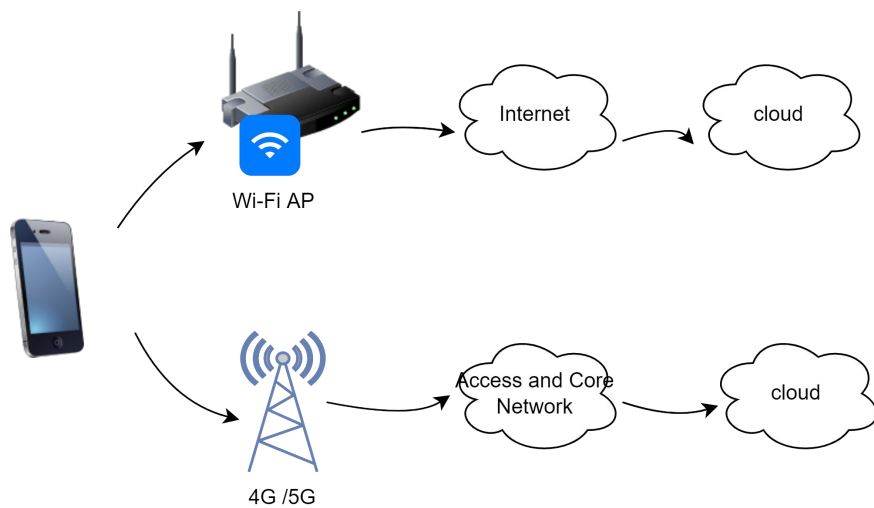
TCP+TLS1.2



TCP+TLS1.3

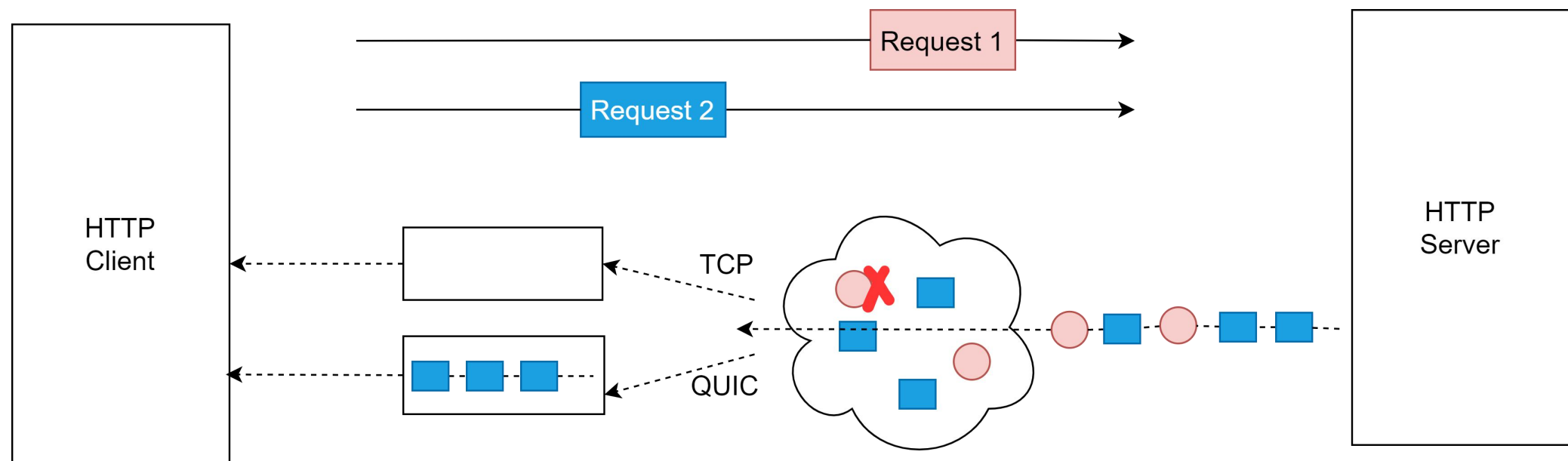


QUIC-连接迁移



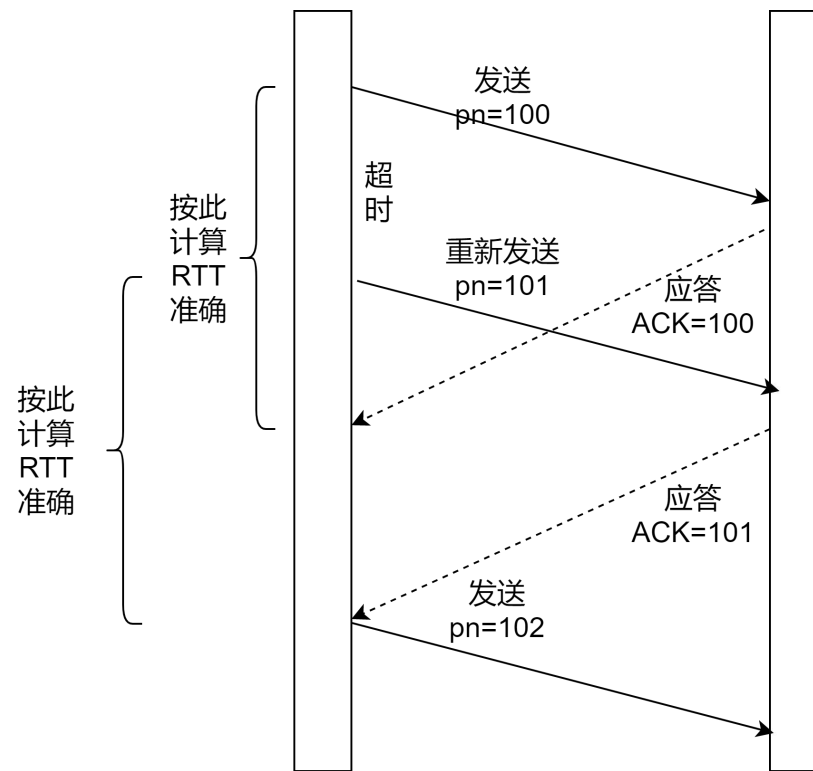
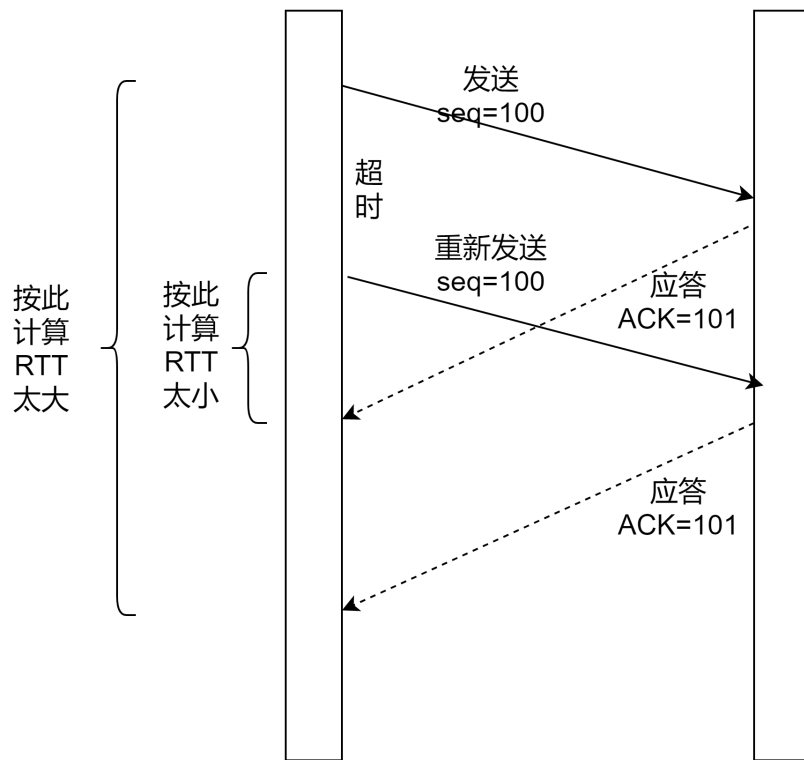
Quic-多流复用

QUIC协议解决了TCP的队头阻塞问题



- HTTP/2提出了“流”的概念，实现了在一条连接上的并发请求。
- 但是TCP协议是内核态协议，无法识别“流”，QUIC连接实现在用户态，可以识别“流”。

Quic-更精准的RTT测量



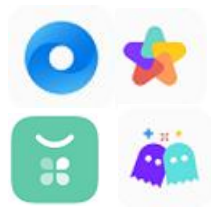
单调递增的包序号，解决了TCP重传包的二义性。

02

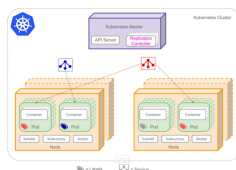
OPPO分布式系统架构

OPPO的在线业务量

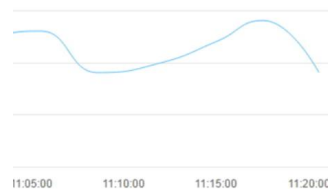
OPPO统一接入层业务规模



数千级业务集群



数十万级容器实例

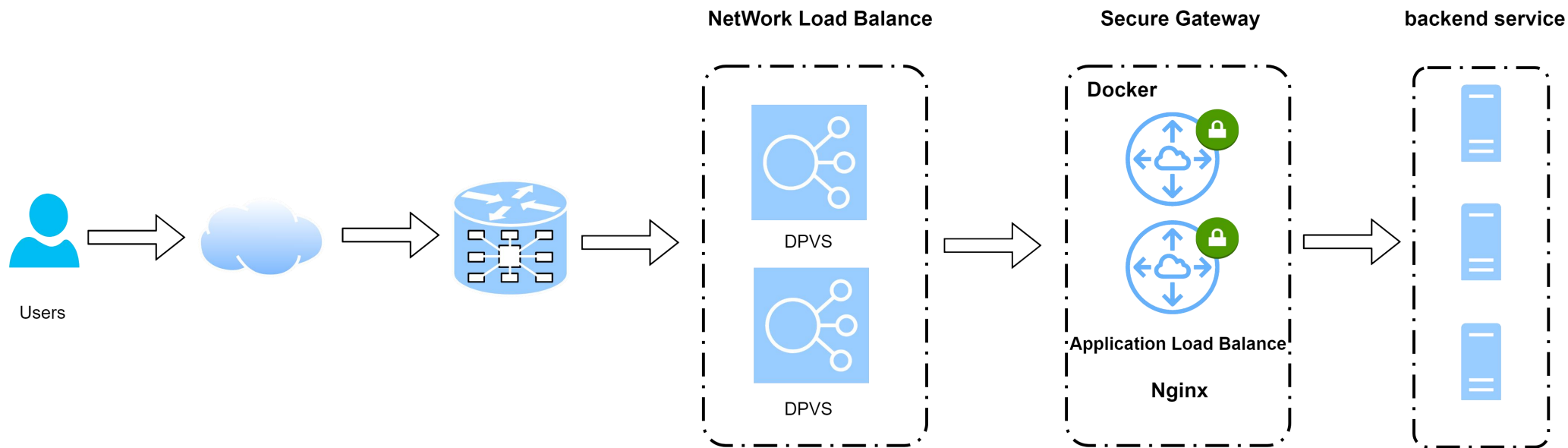


千万级QPS

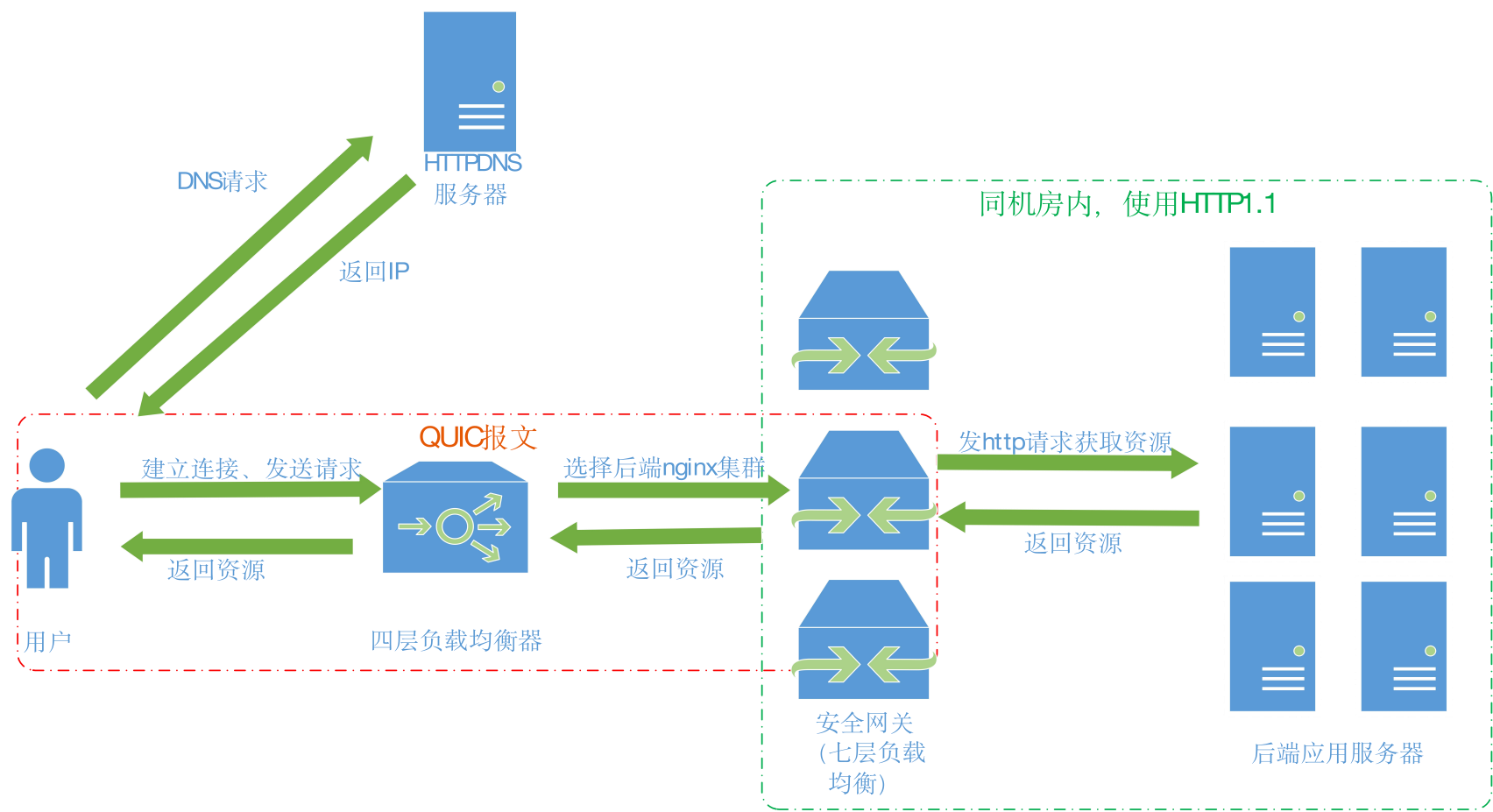


数亿级用户

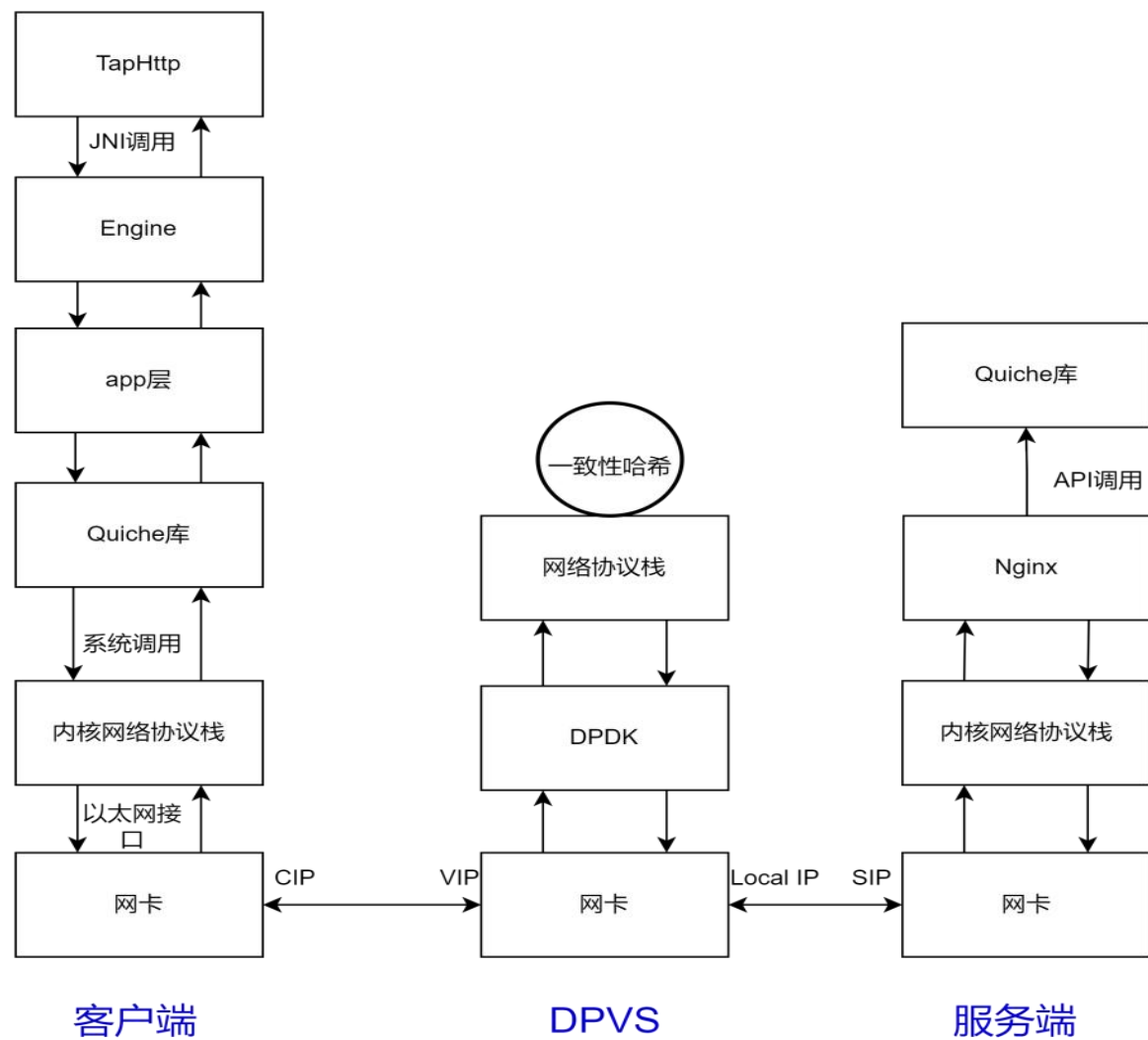
OPPO接入层架构



接入层QUIC架构



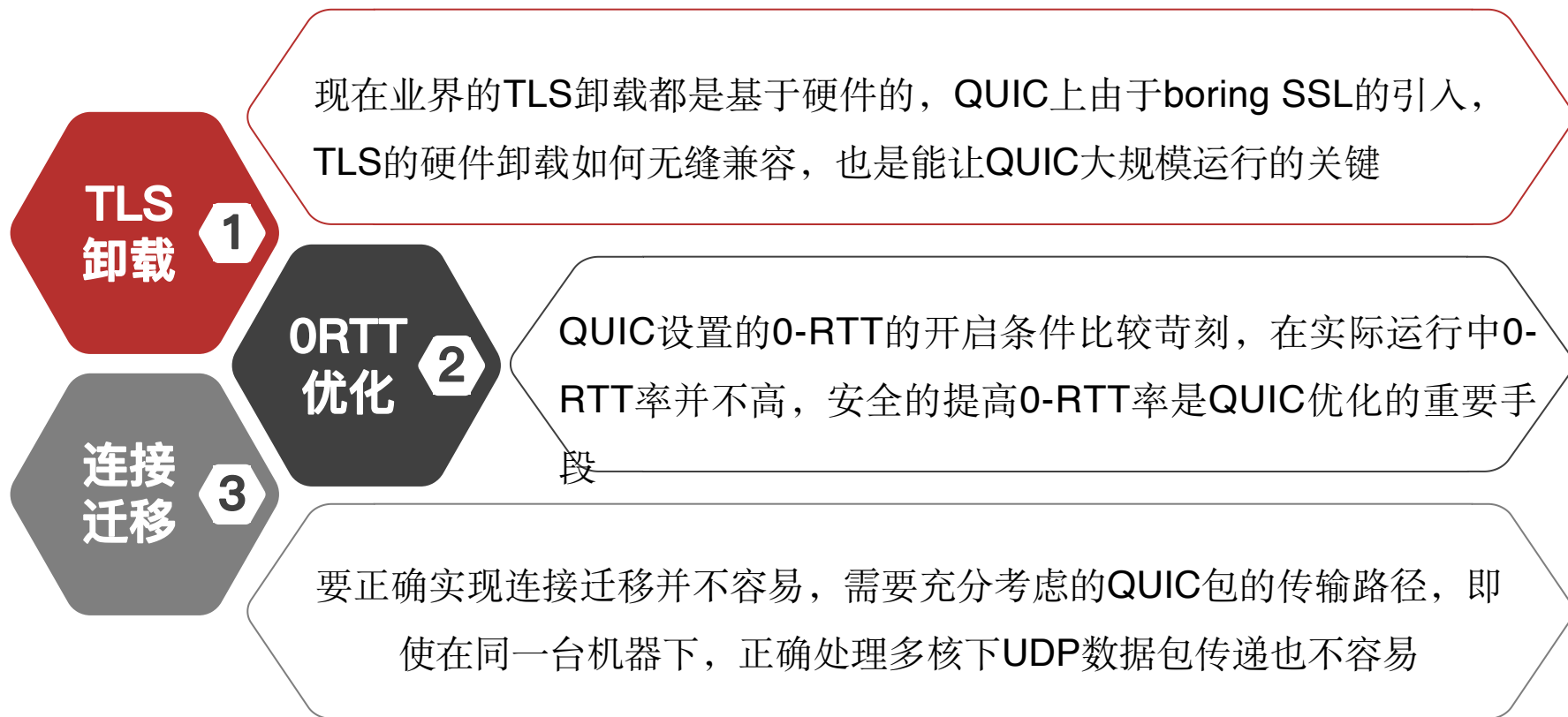
接入层QUIC架构



03

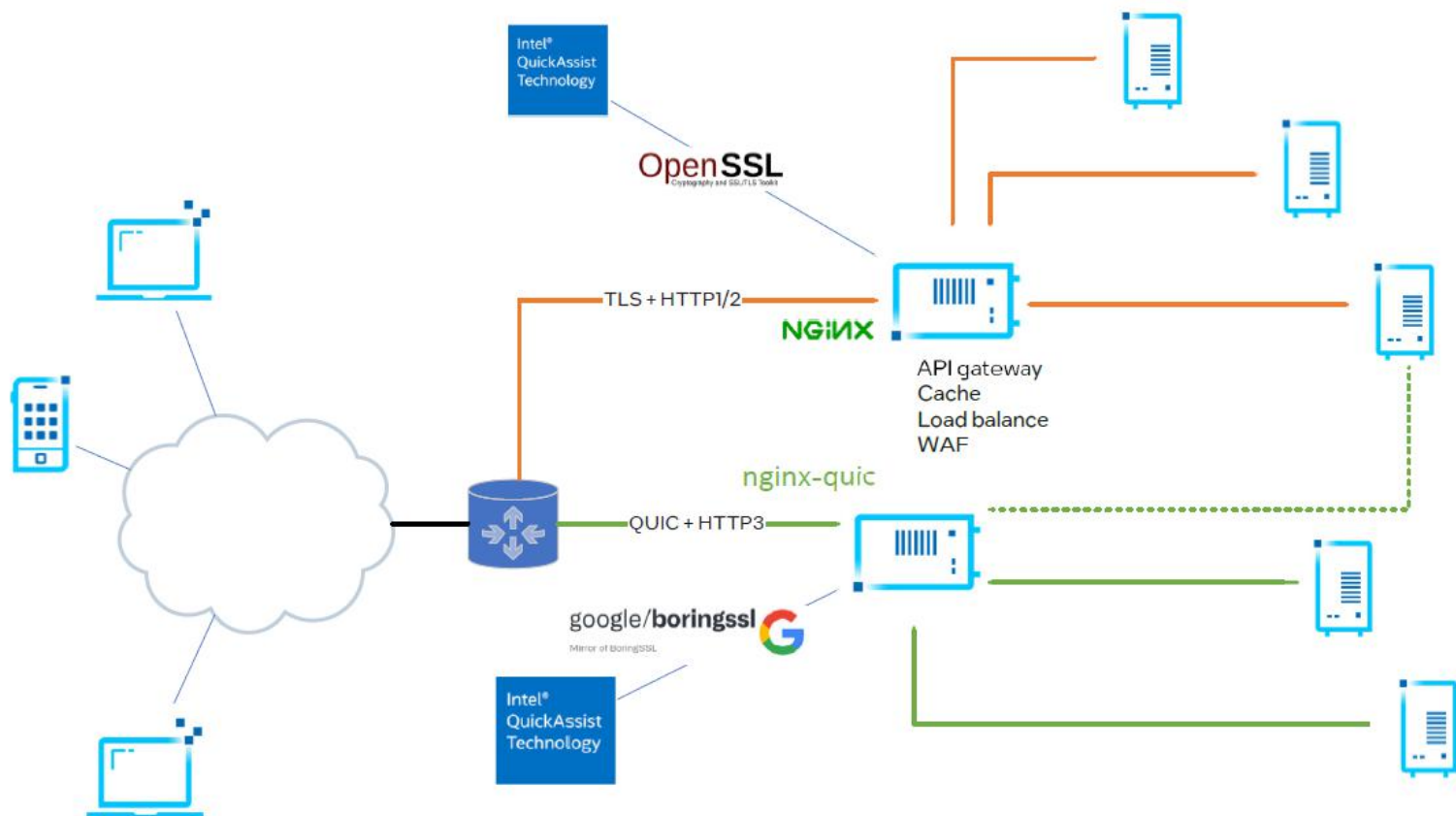
Quic协议在分布式架构中的优化

QUIC的实现会有哪些问题



QUIC的SSL硬件卸载架构

- 将非对称加密过程卸载到QAT加速卡中
- 针对Nginx进行适配
- 异步处理方式
- 支持强大的压缩加速能力
- 实现了对BoringSSL库的支持，支持Quic协议

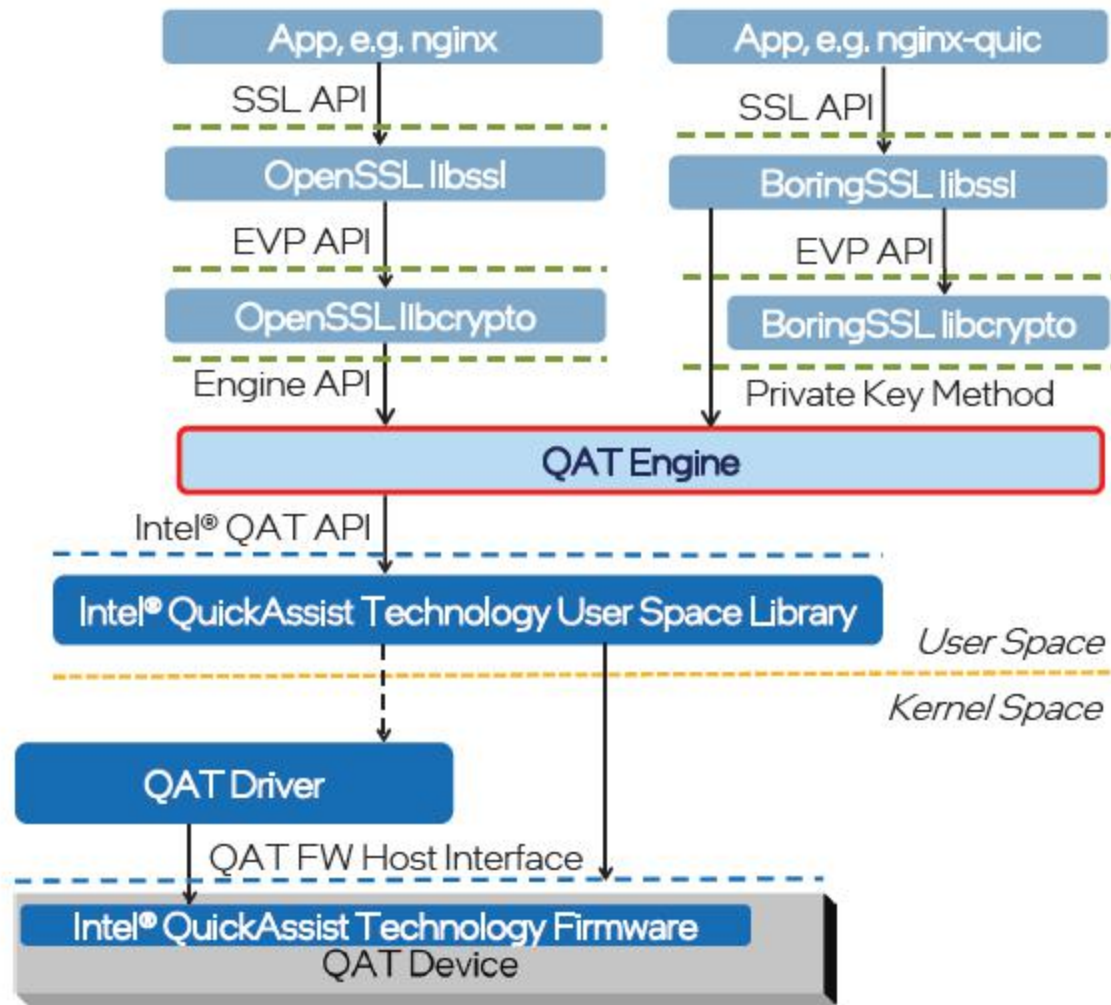
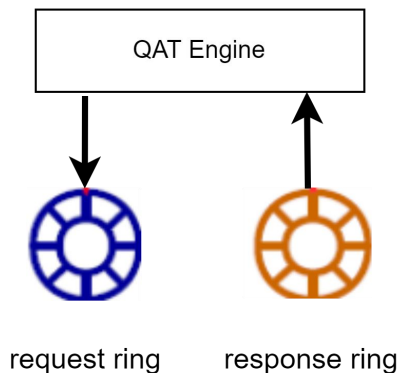


QUIC-SSL卸载的软件架构

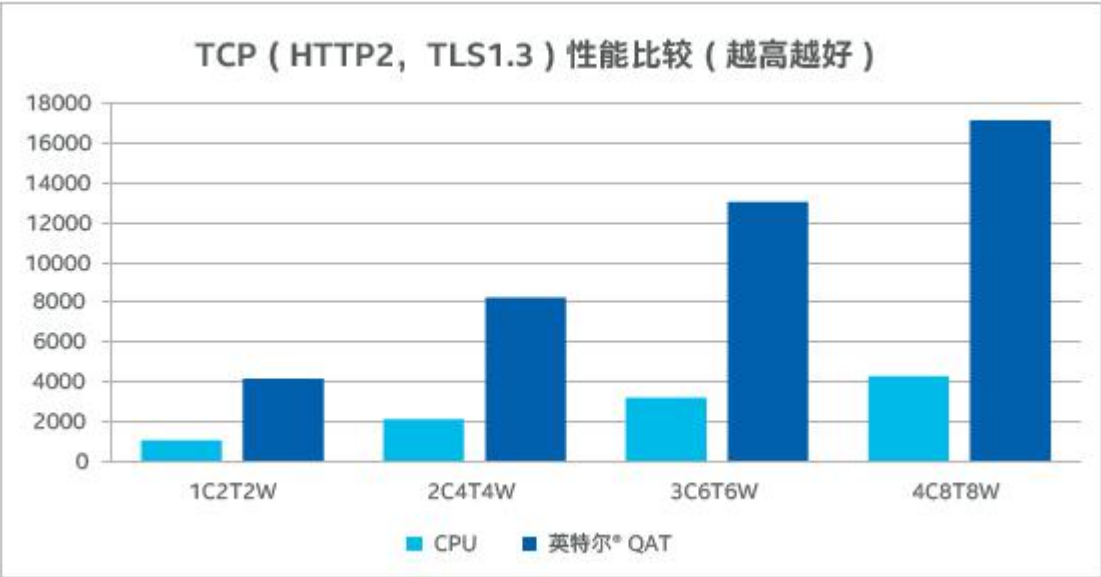
OPPO的安全网关基于Nginx进行自研定制化，继承了Nginx的异步设计特性。

OPPO与英特尔联合开发基于Quic协议的SSL卸载方案，包括加解密库的QAT加速适配、Quic协议栈和QAT引擎的异步化等措施。

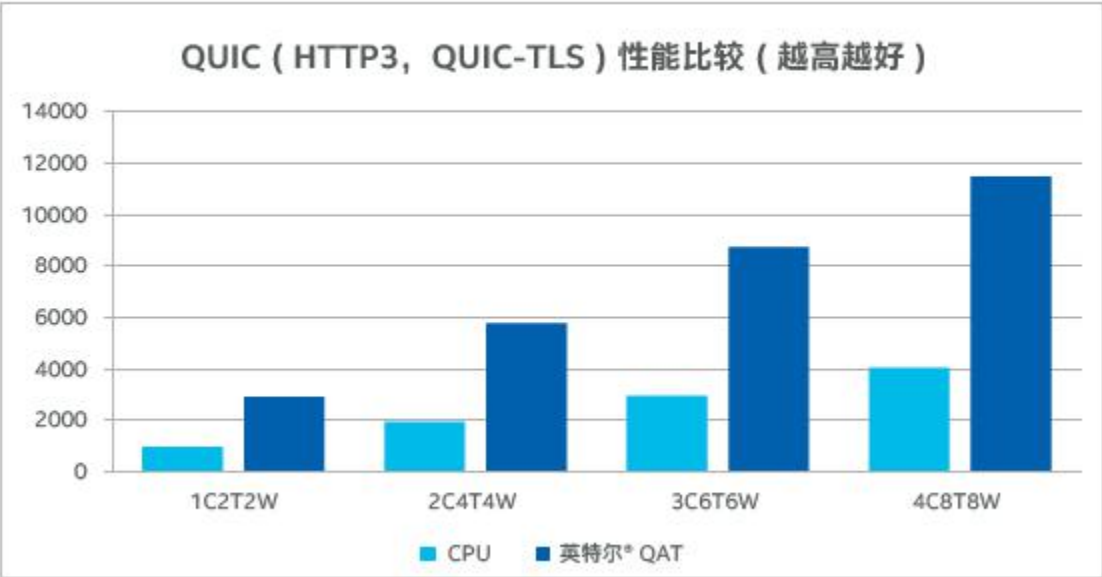
通过这些优化，在一台web服务器中能够并发进行TLS1.2/TLS1.3/Quic的加速，在QAT引擎库中实现了对Open SSL和Boring SSL的同时支持。



QUIC的SSL硬件卸载性能对比



TCP+TLS1.3 性能提升达4.05倍



QUIC 性能提升达3倍

0-RTT率低的问题

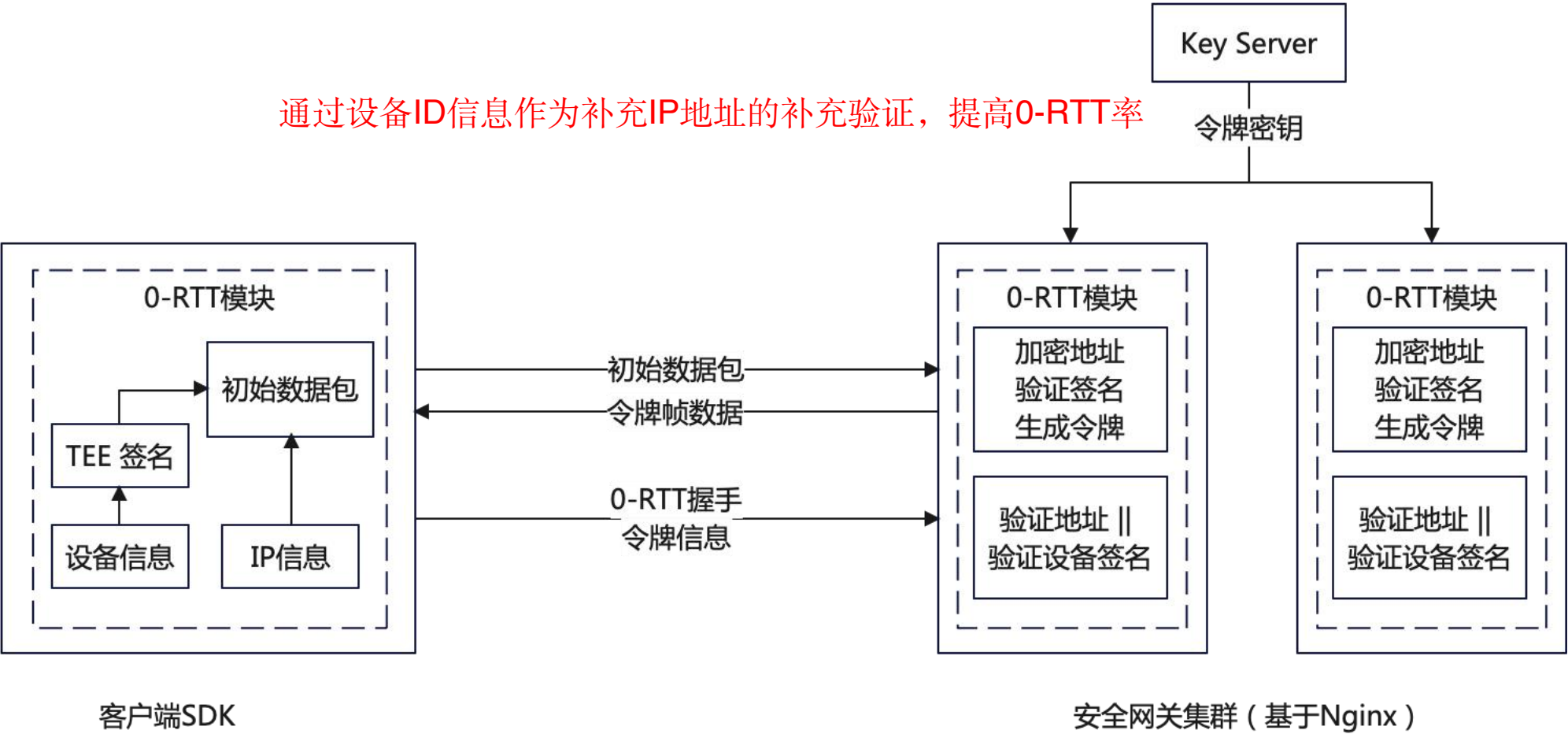
0-RTT的2个问题

客户端ip并不是固定的，在4G和WIFI切换IP地址都会发生变化，服务端下发的token融合了客户端的IP，这个IP发生变化时，携带的token服务端校验不通过，无法0-RTT

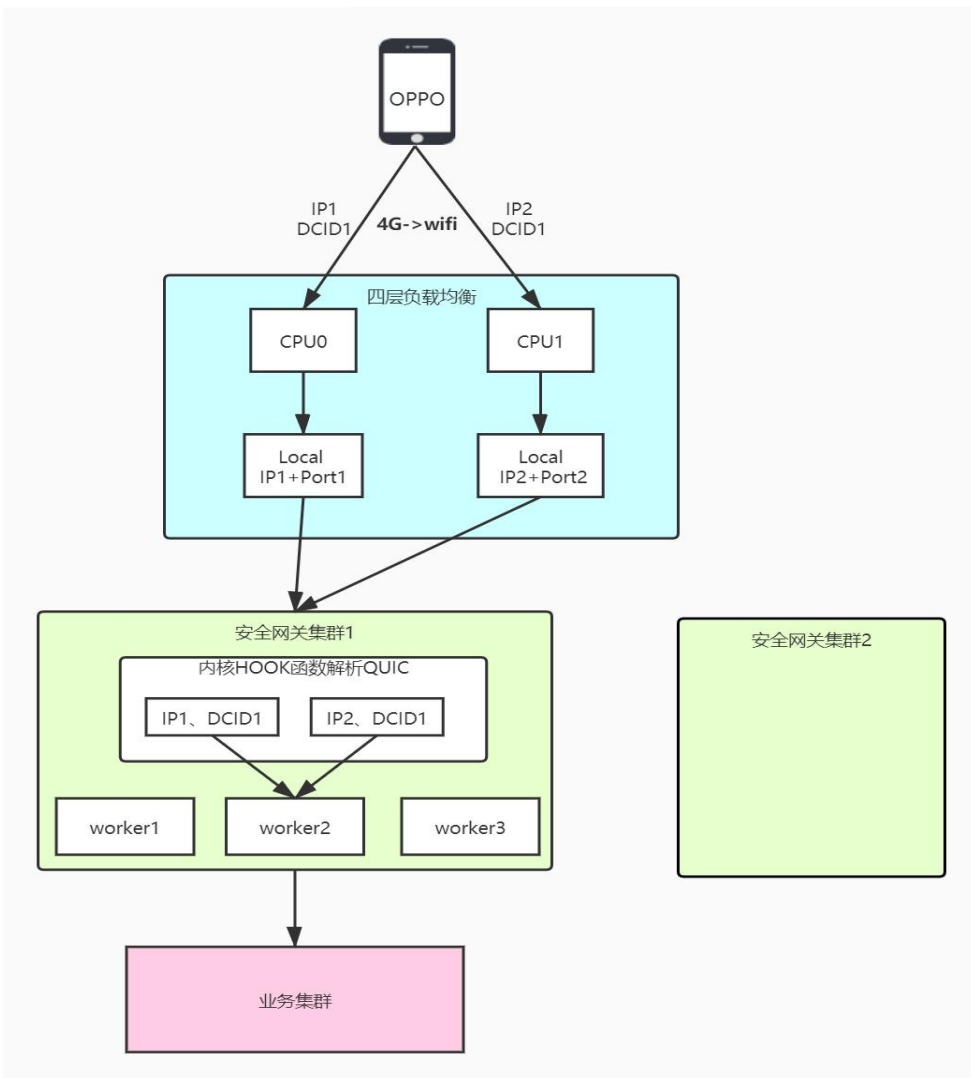
服务端都有集群，后端实际处理请求的服务器对于客户端来说不是固定的，新的请求到来时，如果当前client没有请求过该服务器，则服务器上没有相关会话信息，会把该请求当做一个新的连接来处理，重新走1-RTT

0-RTT的优化

通过设备ID信息作为补充IP地址的补充验证，提高0-RTT率



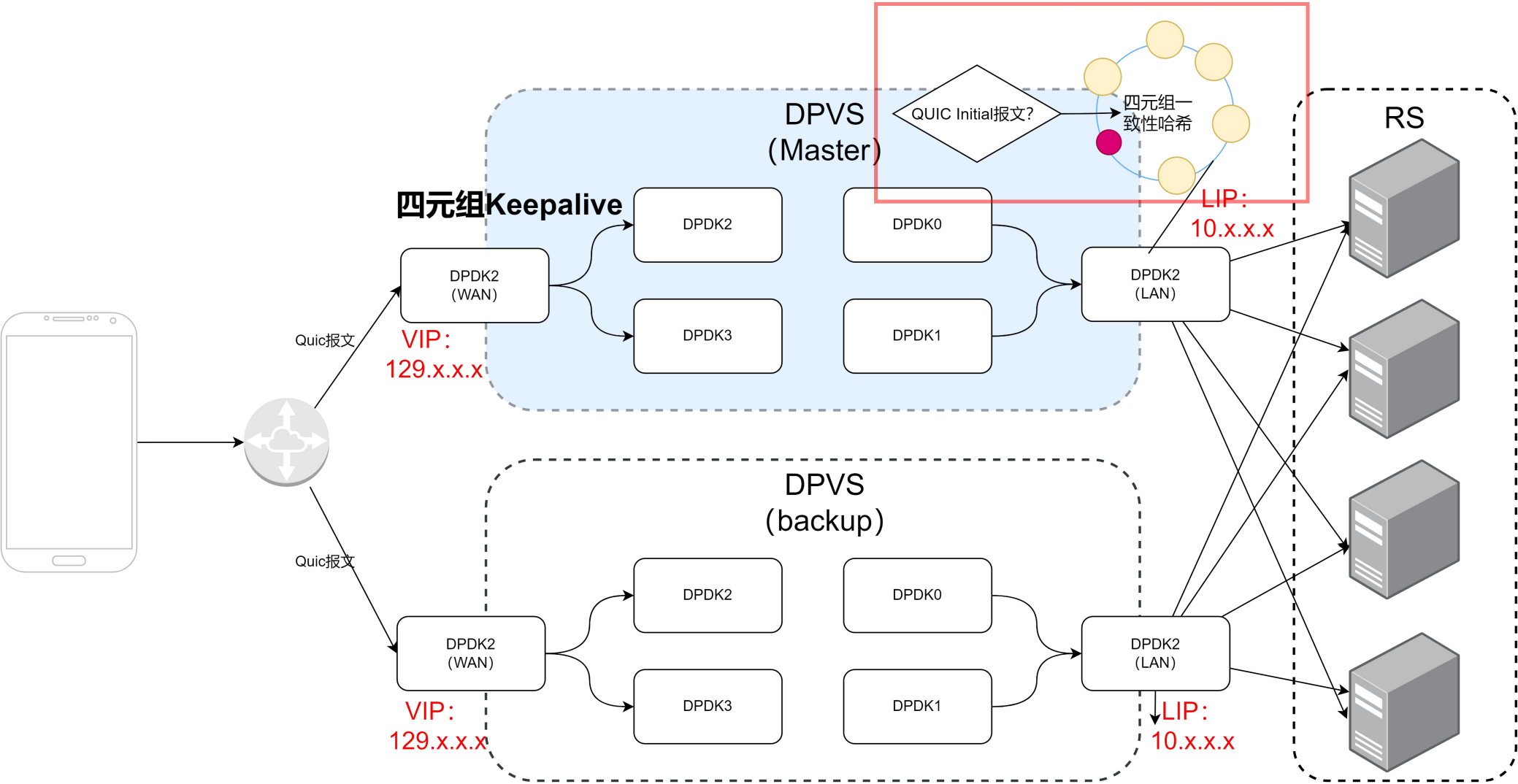
QUIC连接迁移



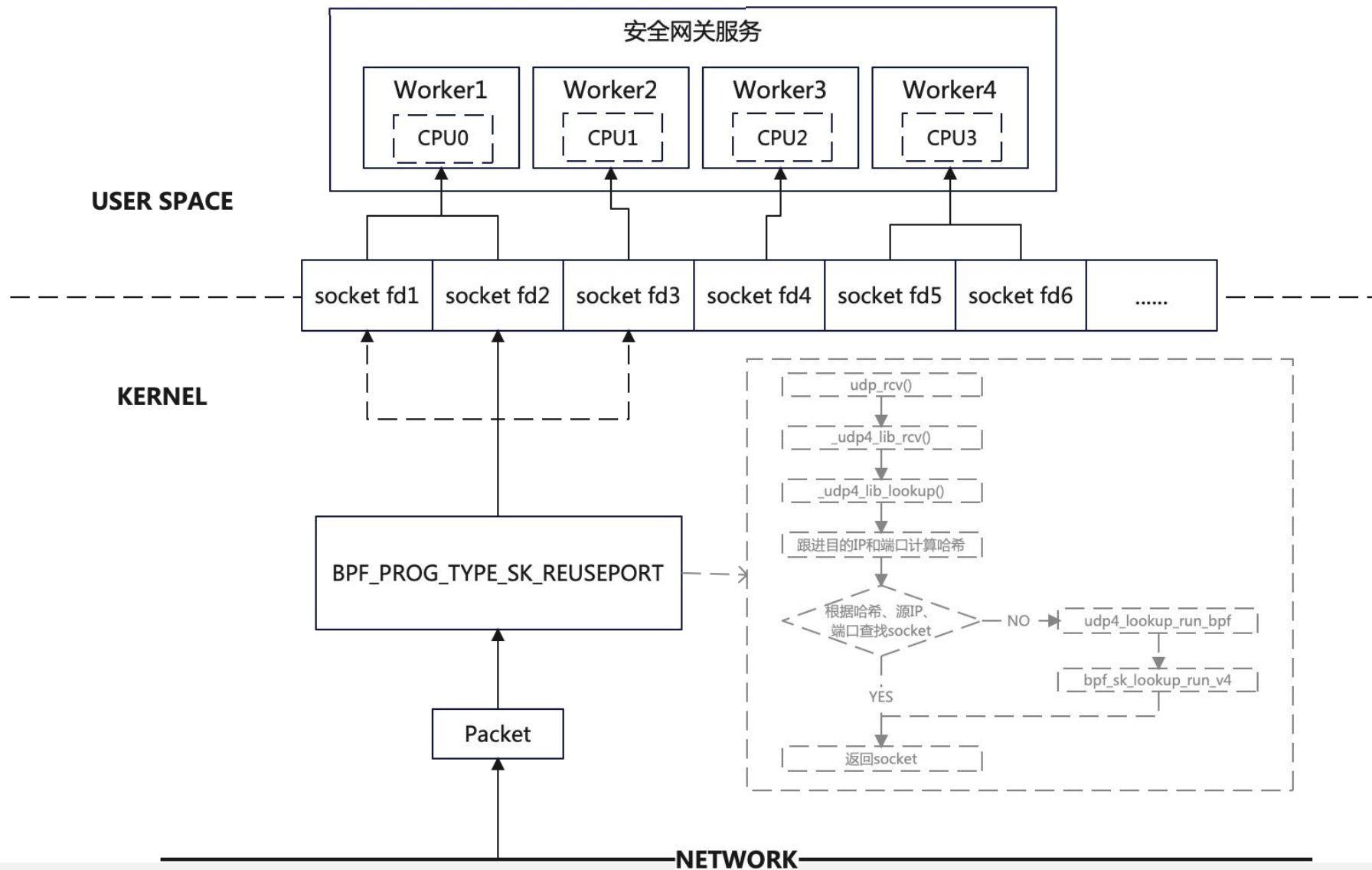
实际过程中会产生问题:

1. 四层架构的影响: LVS、DPVS等四层负载均衡工具基于四元组进行转发, 会转发到不同的后端服务上, 导致无法连接迁移;
2. 多核的影响: 由于多核的原因, 连接迁移中源地址的改变可能会让接下来的数据包去到不同的进程, 影响socket数据的接收

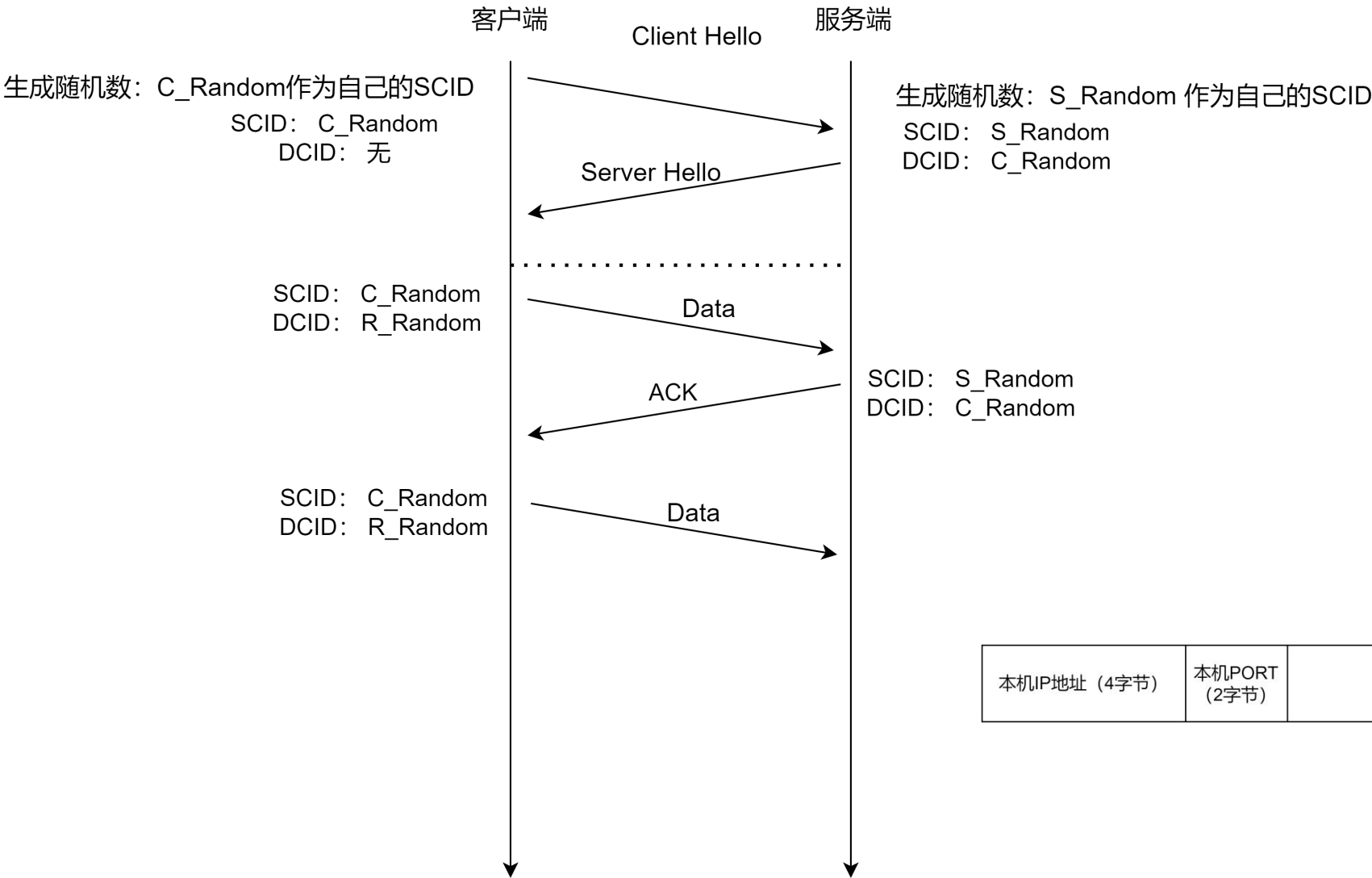
QUIC连接迁移-四层负载均衡器架构



QUIC连接迁移-七层负载均衡器架构

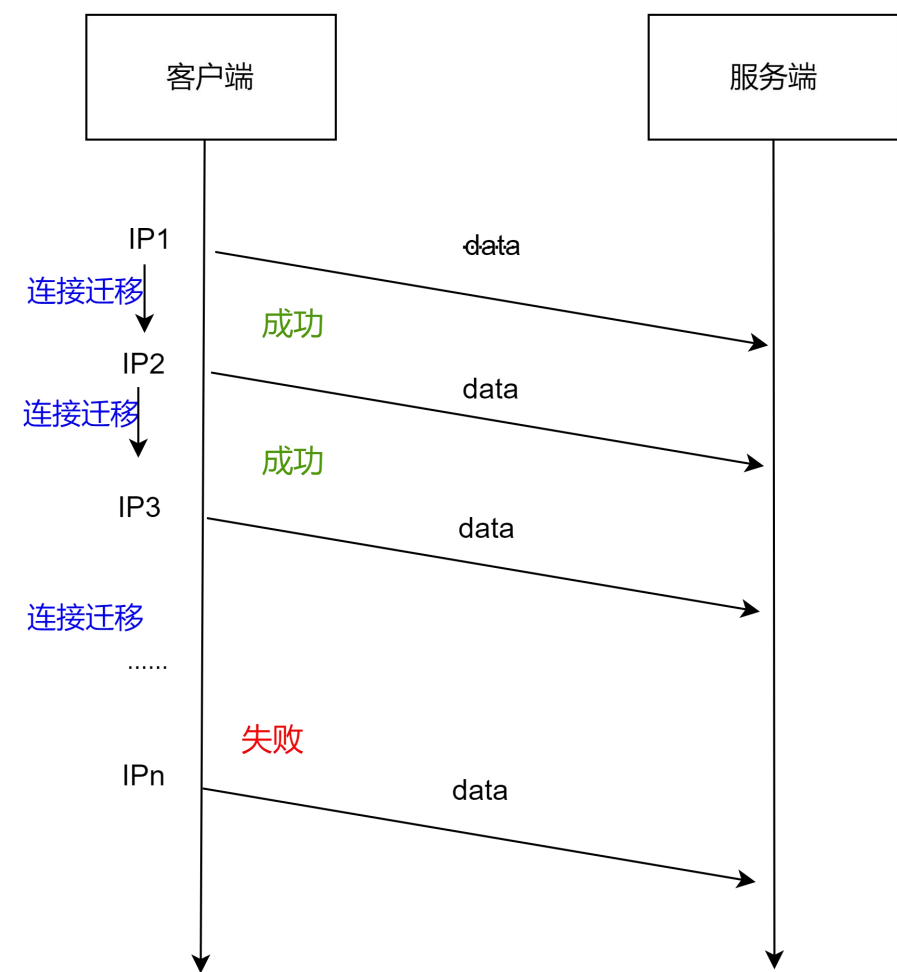
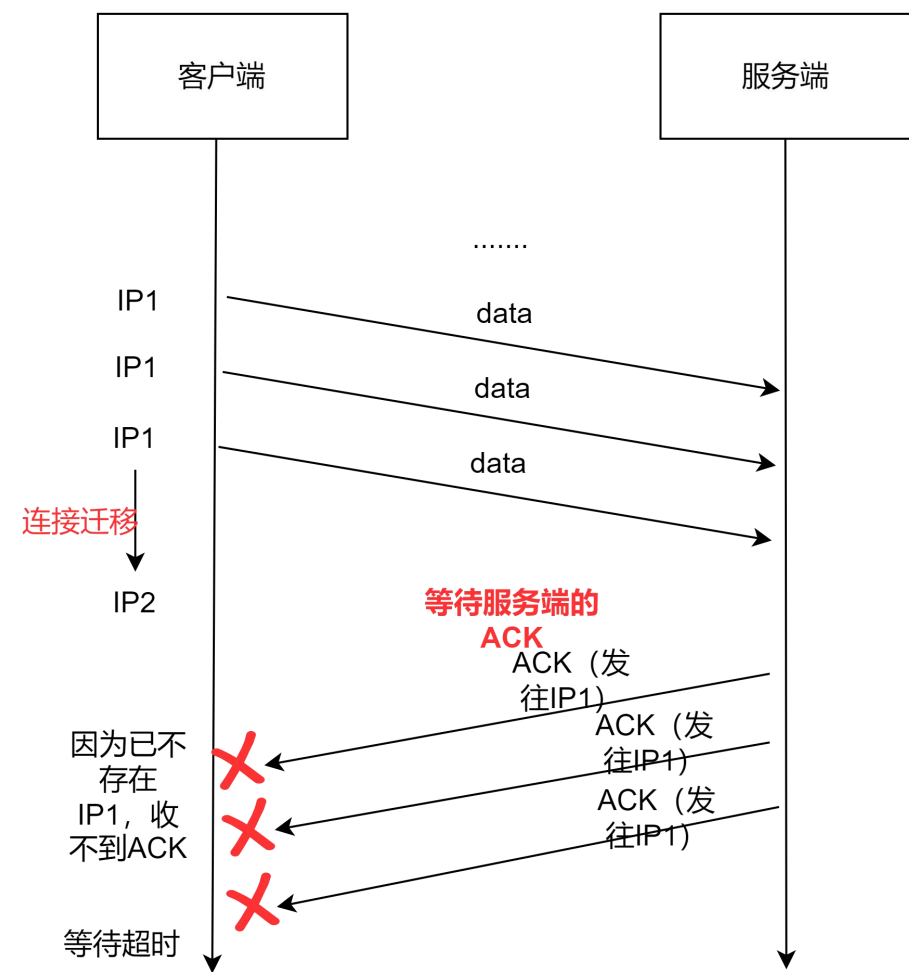


QUIC连接迁移优化



本机IP地址 (4字节)	本机PORT (2字节)	随机数 (16字节)
--------------	--------------	------------

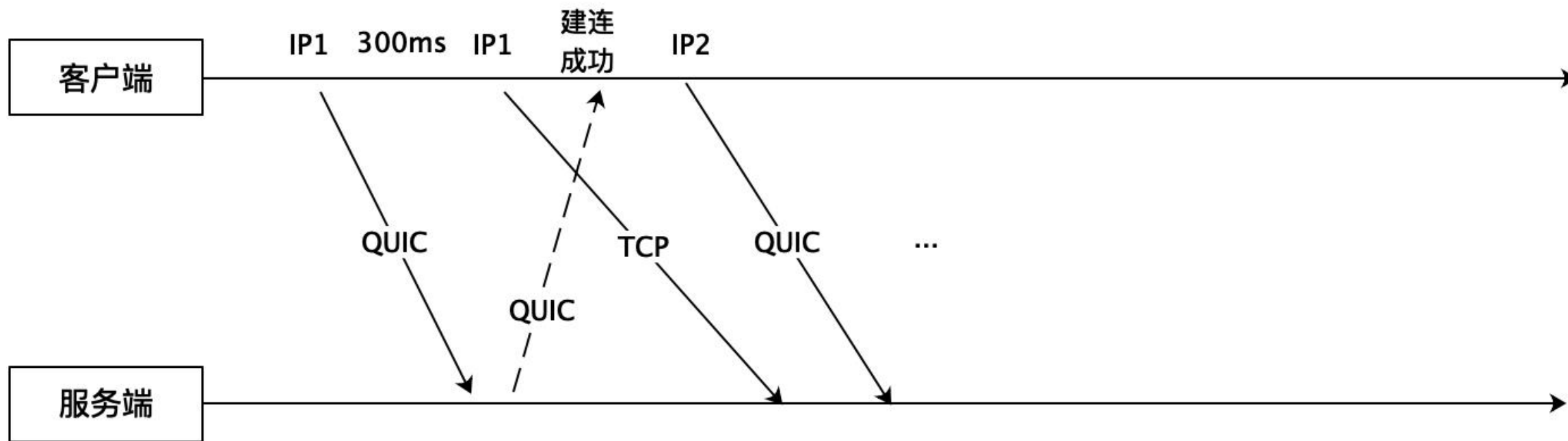
QUIC连接迁移-为什么会迁移失败？



QUIC-竞速机制

业内统计数据全球有7%地区的运营商对UDP有限速或者禁闭，除了运营商还有很多企业、公共场合也会限制UDP流量甚至禁用UDP。这对使用UDP来承载QUIC协议的场景会带来致命的伤害。

对此，OPPO安全网关采用多路竞速的方式使用TCP和QUIC同时建连。除了在建连进行竞速以外，还可以对网络QUIC和TCP的传输延时进行实时监控和对比，如果有链路对UDP进行了限速，可以动态从QUIC切换到TCP。



QUIC-PING 帧机制



作用

1. 用于测试网络连接的可达性和延迟情况
2. 解决连接迁移失败问题
3. 触发网络拥塞控制

交互过程

1. 服务端不需要回PONG帧，ACK即可。
2. PING帧中包含Token

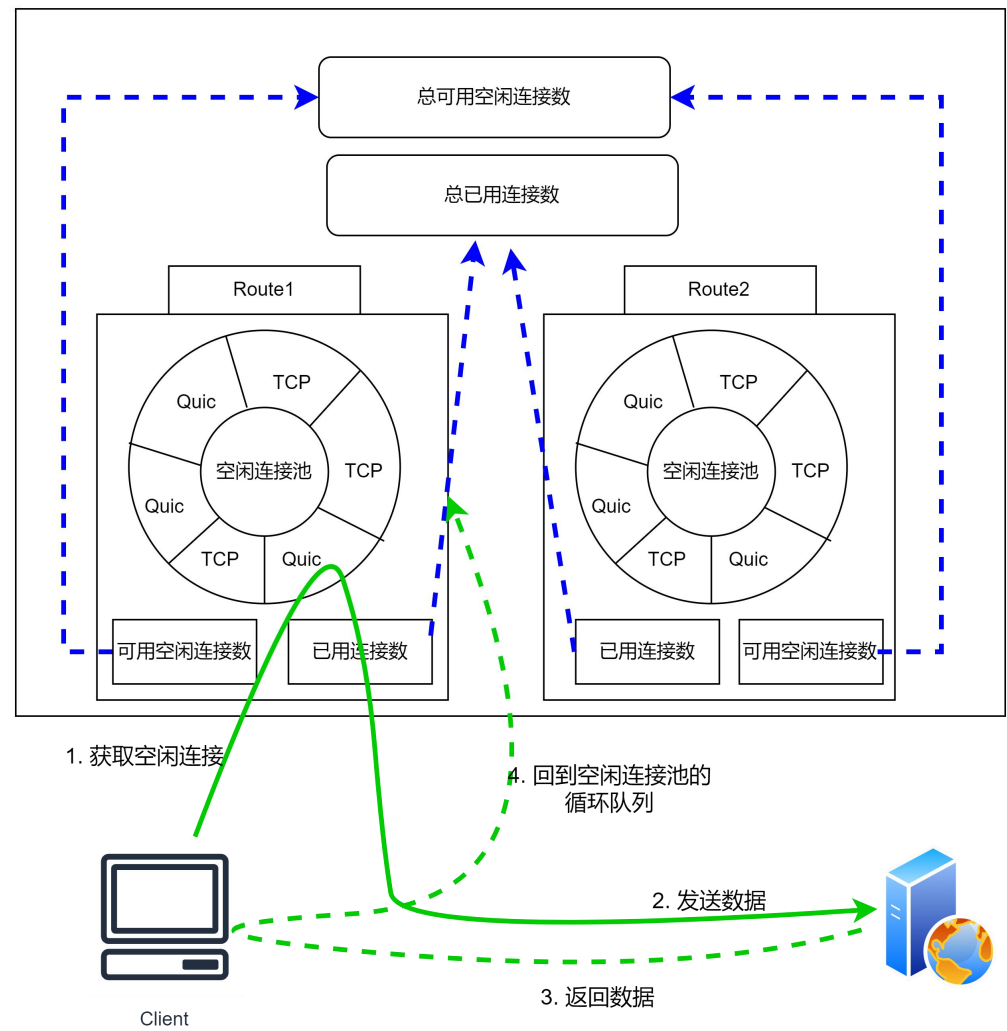
如何实现

1. 1-RTT的时候，需要在连接建立完成后开始发送PING帧。
2. 0-RTT的时候，无法确定连接是否建立完成，PING帧该何时发送？

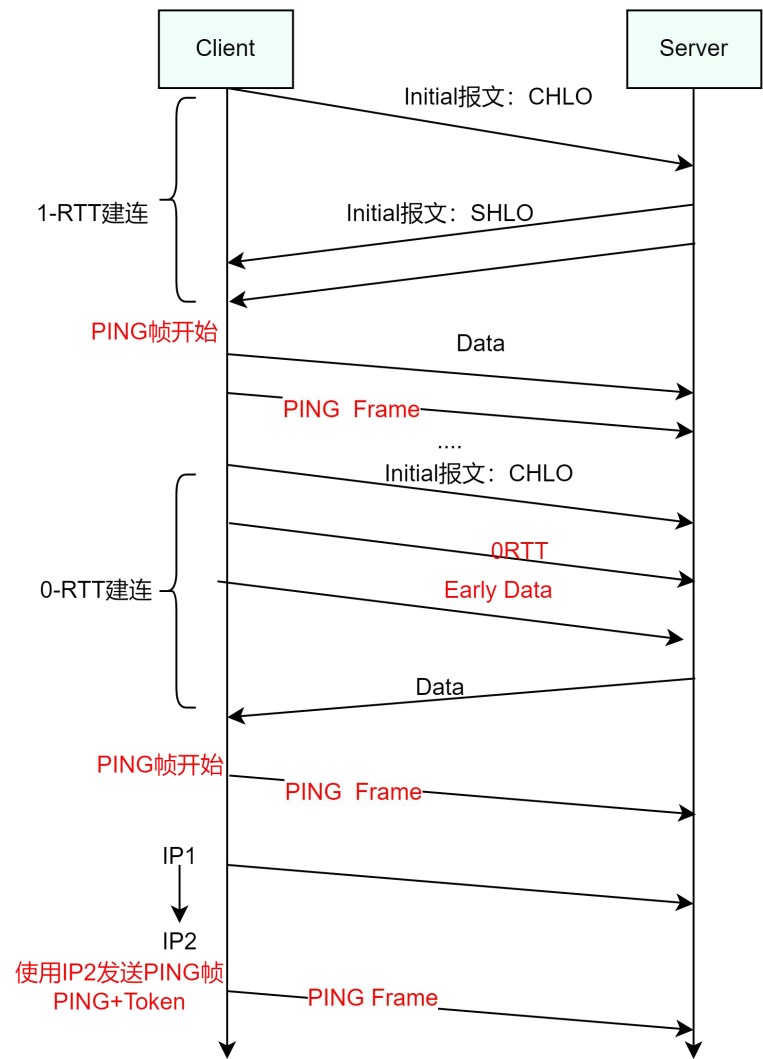
接入层架构中如何设计间隔时间

1. 发送时间间隔如何与四层负载均衡的超时时间进行权衡？又该如何与七层网关进行权衡？

QUIC-PING 帧存在的作用



保持和探测连接活性、及时关闭“坏死”连接，提升连接复用率



解决连接迁移偶然失败的问题

QUIC-IP地址透传

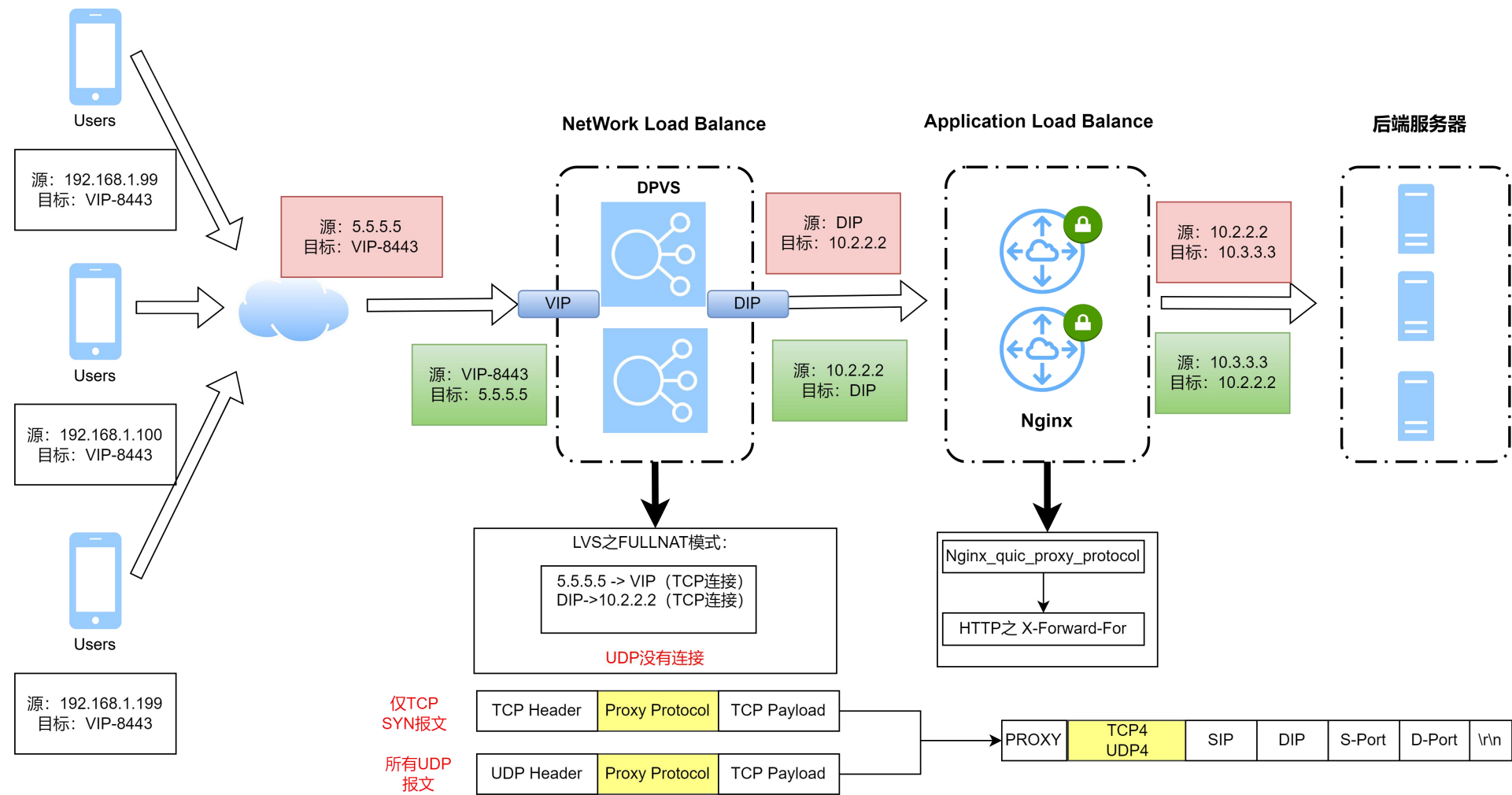
TOA / UOA: 通过将源IP/端口信息放置到IP头的options中, 服务端通过ip头中的option获取源IP信息, 支持的信息有限, V6存在局限性;

ProxyProtocol: 基于TCP的协议都是使用这种方式, 通过将源IP/端口信息放置到每个UDP的payload的起始位置, 服务器端在获取报文后, 将前面payload的源IP信息取出来, 支持V4/V6;

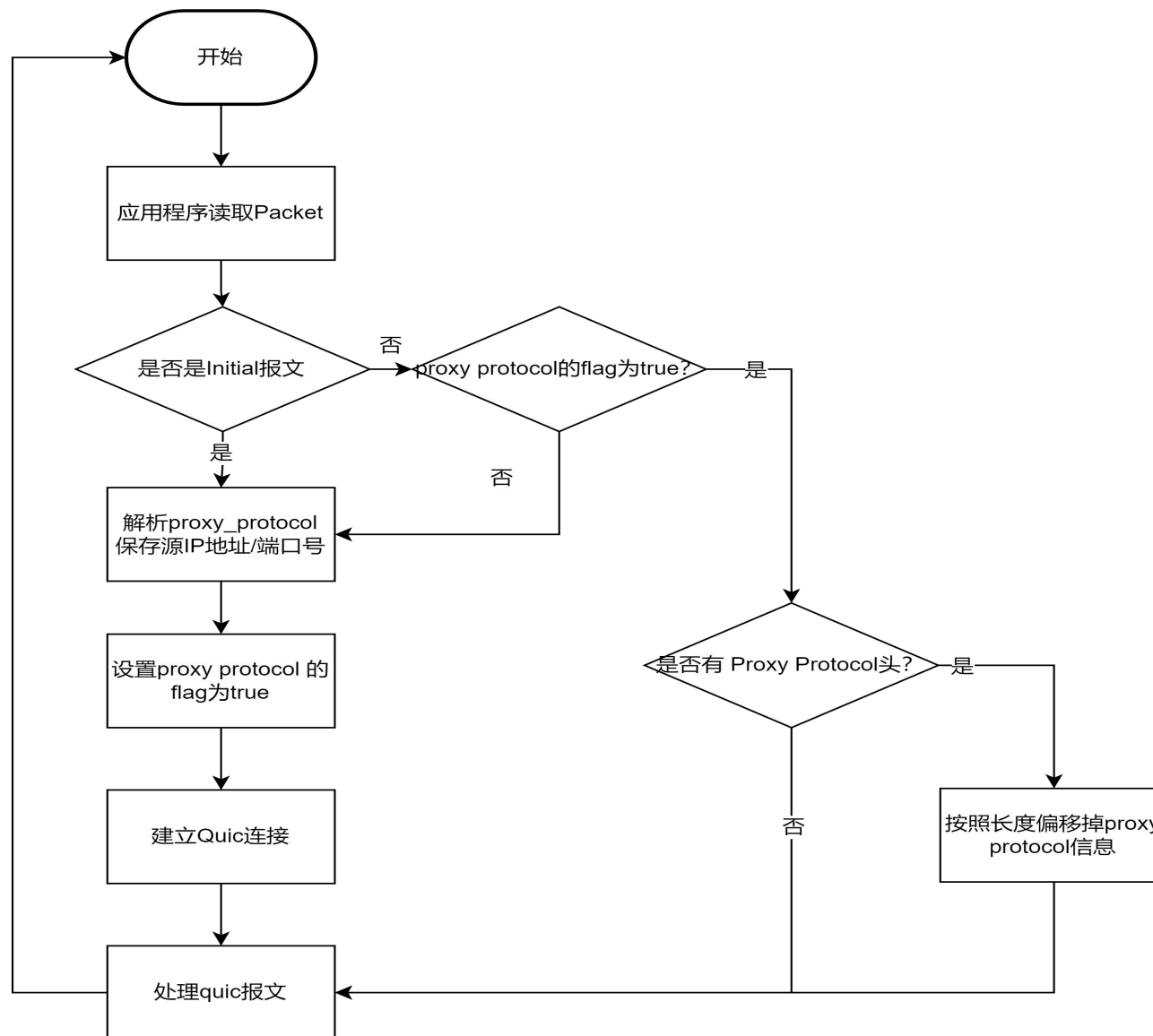
```
PROXY TCP4 192.168.0.1 192.168.0.11 56324 443\r\n
GET / HTTP/1.1\r\n
Host: 192.168.0.11\r\n
\r\n
```

PROXY	TCP4	172.19.0.1	172.19.0.3	42272	80	\r\n
Signature	proxied INET protocol and family	source address	destination address	source port	destination port	CRLF

QUIC-分布式架构中如何实现IP地址透传



QUIC-IP地址透传



QUIC-性能优化效果

通过弱网实验测试，QUIC在开启0-RTT时，其延迟要比HTTP降低20%，比HTTPS要降低50%以上。
现在主要在海外商店、小布助手等多个业务上线使用QUIC。

		quic（竞速模式）		http		https	
		成功率	延时	成功率	延时	成功率	延时
25ms, 0%	1k	100%	87.583ms	100%	128.105ms	100%	292.903ms
	100k	100%	263.083ms	100%	470.3ms	100%	618.12ms
100ms, 0%	1k	100%	256.611ms	100%	467.594ms	100%	902.486ms
	100k	100%	852.666ms	100%	1678.594ms	100%	2200.135ms
100ms, 10%	1k	100%	875.388ms	99.41%	953.425ms	100%	1746.848ms
	100k	99.85%	2552.287ms	99.85%	4364.791ms	99.85%	5386.407ms

QUIC-性能优化效果

在海外软件商店、小布助手上线后，性能有显著提升：接口成功率提升3%~13%，秒开率提升2%~19%，平均延时提升27%~50%。

接口成功率提升百分比	首页	应用文件夹
印度	3%	9%
东南亚 (印尼)	4%	13%

请求秒开率提升百分比	首页	应用文件夹
印度	4%	19%
东南亚 (印尼)	2%	7%

04

Quic协议实现中的问题案例分享

QUIC常见问题（一）

握手包乱序导致建连时间特别长

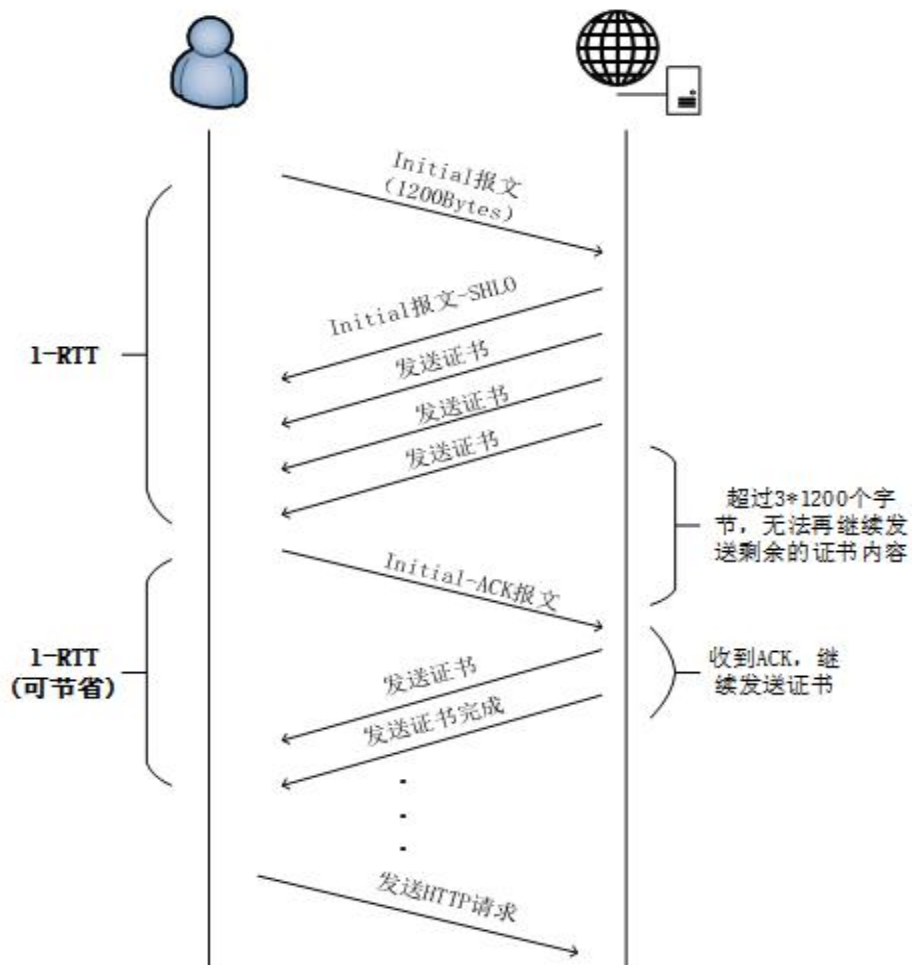
46	*REF*	10.	223.53	10.	3.34	QUIC	49236	9999	1242 Initial, DCID=702e2d7036b8f4920d220f5db34c8c02, SCID=2b425582b9c08536359307b2108c2a9b27443d69, PKN: 0, CRYPTO
47	0.030221	10.	34	10.	3.53	QUIC	9999	49236	1242 Handshake, DCID=2b425582b9c08536359307b2108c2a9b27443d69, SCID=0ab13522270f8393358b760c350fa80ecd365e7b
48	0.030680	10.	34	10.	53	QUIC	9999	49236	1242 Handshake, DCID=2b425582b9c08536359307b2108c2a9b27443d69, SCID=0ab13522270f8393358b760c350fa80ecd365e7b
49	0.030860	10.	34	10.	53	QUIC	9999	49236	481 Handshake, DCID=2b425582b9c08536359307b2108c2a9b27443d69, SCID=0ab13522270f8393358b760c350fa80ecd365e7b
50	0.031074	10.	34	10.	53	QUIC	9999	49236	1242 Handshake, DCID=2b425582b9c08536359307b2108c2a9b27443d69, SCID=0ab13522270f8393358b760c350fa80ecd365e7b, PKN: 0, CRYPTO
51	0.044754	10.	53	10.	34	QUIC	49236	9999	1242 Handshake, DCID=0ab13522270f8393358b760c350fa80ecd365e7b, SCID=2b425582b9c08536359307b2108c2a9b27443d69, PKN: 0, ACK
52	0.162713	10.	3.53	10.	34	QUIC	49236	9999	115 Handshake, DCID=0ab13522270f8393358b760c350fa80ecd365e7b, SCID=2b425582b9c08536359307b2108c2a9b27443d69, PKN: 1, ACK, PING
53	0.170087	10.1	3.34	10.	53	QUIC	9999	49236	114 Handshake, DCID=2b425582b9c08536359307b2108c2a9b27443d69, SCID=0ab13522270f8393358b760c350fa80ecd365e7b, PKN: 4, ACK
54	0.171475	10.1	3.34	10.	53	QUIC	9999	49236	1242 Handshake, DCID=2b425582b9c08536359307b2108c2a9b27443d69, SCID=0ab13522270f8393358b760c350fa80ecd365e7b, PKN: 5, ACK, CRYPTO
55	0.171623	10.1	34	10.	53	QUIC	9999	49236	119 Handshake, DCID=2b425582b9c08536359307b2108c2a9b27443d69, SCID=0ab13522270f8393358b760c350fa80ecd365e7b, PKN: 6, CRYPTO
56	0.188789	10.1	3.53	10.	34	QUIC	49236	9999	114 Handshake, DCID=0ab13522270f8393358b760c350fa80ecd365e7b, SCID=2b425582b9c08536359307b2108c2a9b27443d69, PKN: 2, ACK
57	0.199387	10.1	3.34	10.	53	QUIC	9999	49236	1242 Handshake, DCID=2b425582b9c08536359307b2108c2a9b27443d69, SCID=0ab13522270f8393358b760c350fa80ecd365e7b, PKN: 7, CRYPTO
58	0.199801	10.1	34	10.	53	QUIC	9999	49236	481 Handshake, DCID=2b425582b9c08536359307b2108c2a9b27443d69, SCID=0ab13522270f8393358b760c350fa80ecd365e7b, PKN: 8, CRYPTO
59	0.207117	10.	53	10.	34	QUIC	49236	9999	154 Handshake, DCID=0ab13522270f8393358b760c350fa80ecd365e7b, SCID=2b425582b9c08536359307b2108c2a9b27443d69, PKN: 3, ACK, CRYPTO
60	0.210954	10.	53	10.	34	QUIC	49236	9999	94 Protected Payload (KP0), DCID=0ab13522270f8393358b760c350fa80ecd365e7b, PKN: 0, STREAM(0)
61	0.215807	10.	34	10.	53	QUIC	9999	49236	601 Protected Payload (KP0), DCID=2b425582b9c08536359307b2108c2a9b27443d69, PKN: 0, DONE, CRYPTO
62	0.217699	10.	53	10.	34	QUIC	49236	9999	85 Protected Payload (KP0), DCID=0ab13522270f8393358b760c350fa80ecd365e7b, PKN: 1, STREAM(0)
63	0.220090	10.1	34	10.	3.53	QUIC	9999	49236	1115 Protected Payload (KP0), DCID=2b425582b9c08536359307b2108c2a9b27443d69, PKN: 1, ACK, STREAM(0)

Retry报文引起建连需要2-RTT

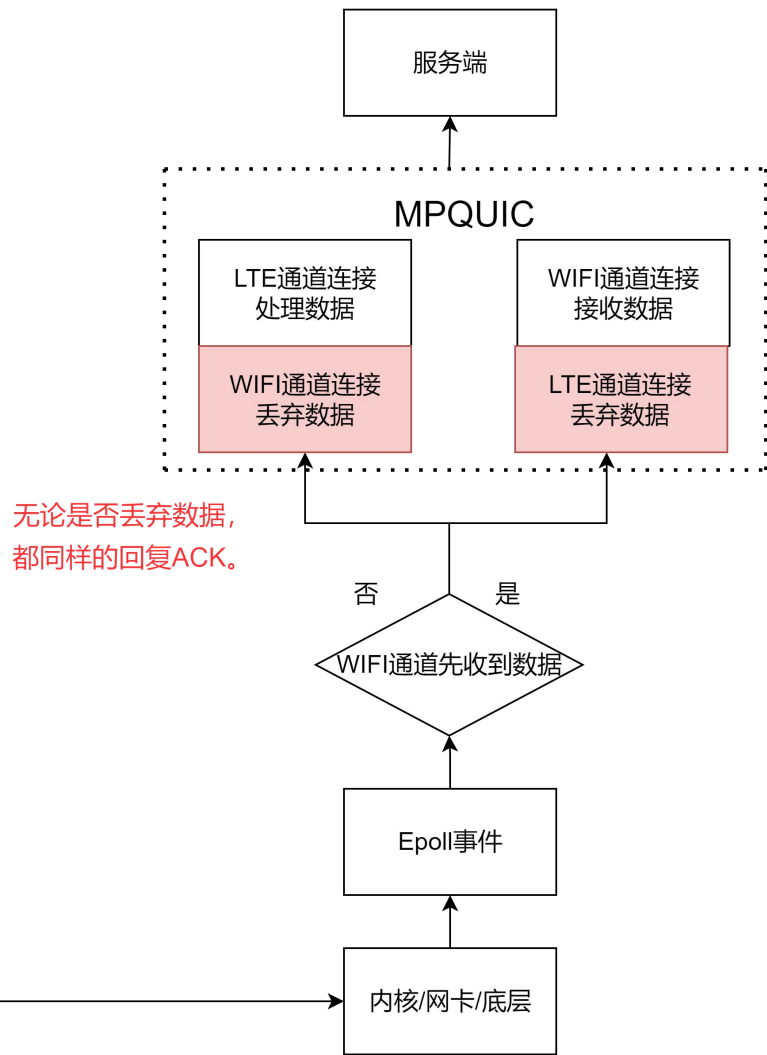
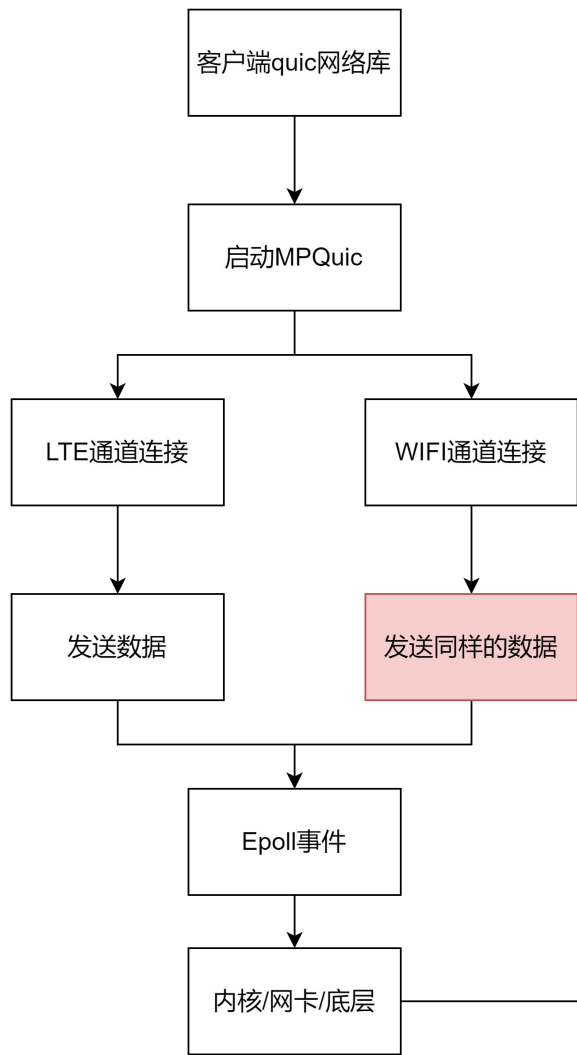
10.	126	10.1	1.143	QUIC	1242 Initial, DCID=1fdea9b0b56c475b825cf8937b20e81c, SCID=46855778bf87eb4b2af5e4e0a07d16ce7d5d7004, PKN: 0, CRY
10.	143	10.	126	QUIC	130 Retry, DCID=46855778bf87eb4b2af5e4e0a07d16ce7d5d7004, SCID=e8f2f806056af590638fdaaf5650e89d9b60bfef
10.	126	10.	143	QUIC	1242 Initial, DCID=e8f2f806056af590638fdaaf5650e89d9b60bfef, SCID=46855778bf87eb4b2af5e4e0a07d16ce7d5d7004, PKN
10.	143	10.	126	QUIC	1242 Handshake, DCID=46855778bf87eb4b2af5e4e0a07d16ce7d5d7004, SCID=e8f2f806056af590638fdaaf5650e89d9b60bfef
10.	143	10.	126	QUIC	1242 Handshake, DCID=46855778bf87eb4b2af5e4e0a07d16ce7d5d7004, SCID=e8f2f806056af590638fdaaf5650e89d9b60bfef
10.	143	10.	126	QUIC	1242 Handshake, DCID=46855778bf87eb4b2af5e4e0a07d16ce7d5d7004, SCID=e8f2f806056af590638fdaaf5650e89d9b60bfef
10.	143	10.	126	QUIC	498 Handshake, DCID=46855778bf87eb4b2af5e4e0a07d16ce7d5d7004, SCID=e8f2f806056af590638fdaaf5650e89d9b60bfef
10.	126	10.	143	QUIC	1242 Handshake, DCID=e8f2f806056af590638fdaaf5650e89d9b60bfef, SCID=46855778bf87eb4b2af5e4e0a07d16ce7d5d7004
10.	126	10.	143	QUIC	114 Handshake, DCID=e8f2f806056af590638fdaaf5650e89d9b60bfef, SCID=46855778bf87eb4b2af5e4e0a07d16ce7d5d7004
10.	126	10.	143	QUIC	114 Handshake, DCID=e8f2f806056af590638fdaaf5650e89d9b60bfef, SCID=46855778bf87eb4b2af5e4e0a07d16ce7d5d7004
10.	126	10.	143	QUIC	154 Handshake, DCID=e8f2f806056af590638fdaaf5650e89d9b60bfef, SCID=46855778bf87eb4b2af5e4e0a07d16ce7d5d7004
10.	126	10.	143	QUIC	92 Protected Payload (KP0), DCID=e8f2f806056af590638fdaaf5650e89d9b60bfef

QUIC常见问题（二）

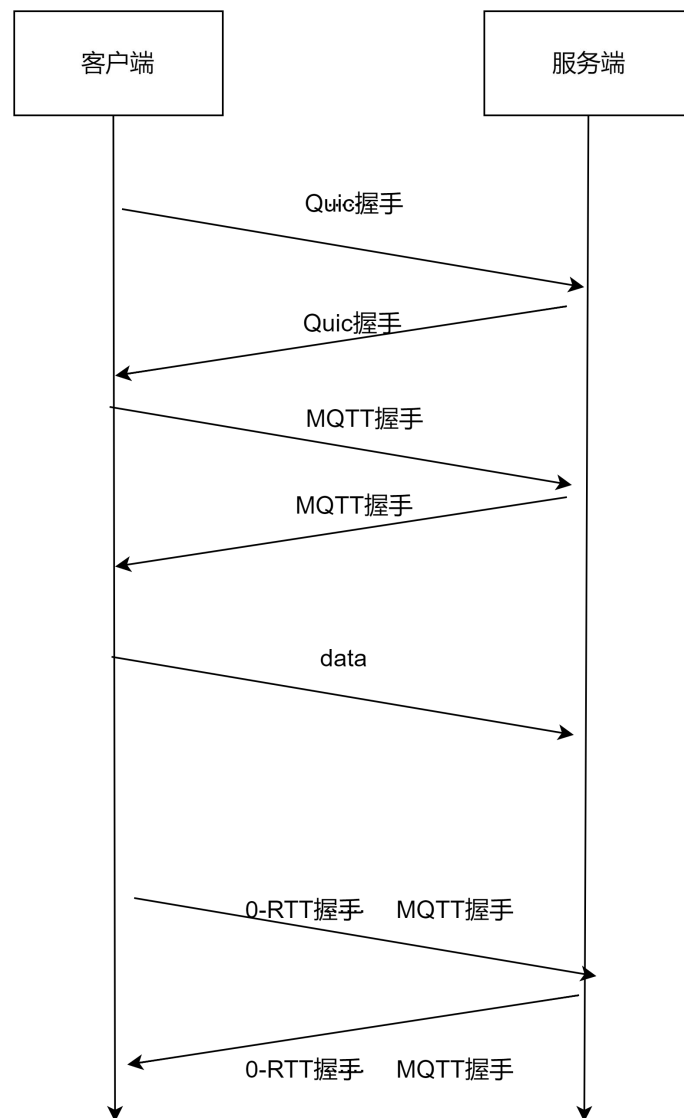
为了防止放大攻击，握手需要多个RTT



一种简化版的MPQUIC



MQTT Over Quic



更多网络相关知识，请关注我的
微信公众号：网络小菜鸟

