CERNET 2019

IPv4与IPv6网流量行为对比分析

东南大学　冯丹

目录

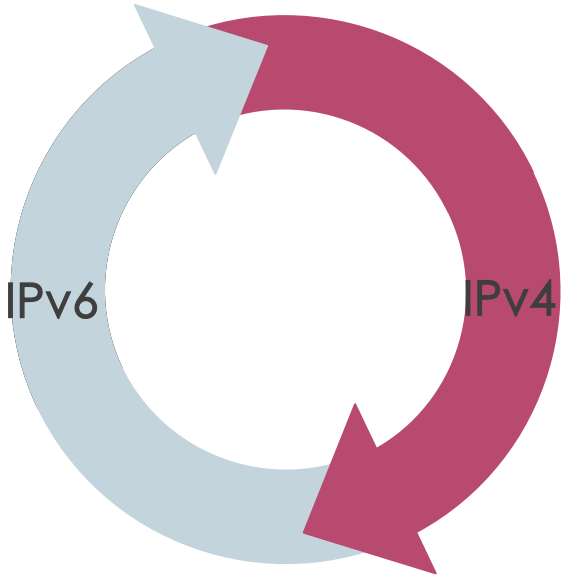
CONTENTS

01

Part one

研究背景及意义

01 / 研究背景及意义



随着IPv4地址分配日渐枯竭和各种弊端的显露，下一代互联网时代即将到来。

虽然新的IPv6协议拥有诸多优势，但目前只有大约20%的互联网流量是使用IPv6协议的，而且过渡阶段发展缓慢，尤其是各种网络服务的部署和建设。

为何造成这种现象？部分文献从技术和管理的角度给出了答案。

1. 设备及兼容问题
2. 安全问题
3. 惰性问题

01 / 研究背景及意义

从体系结构的角度看，IPv4网和IPv6网没有本质的区别，大部分应用都是采用C/S结构；
从网络实际服务的角度出发，IPv6较IPv4有更复杂的应用场景，流量行为表现也愈加错综复杂。
除此之外，IPv6网络还存在以下流量威胁：

- 大量占用带宽，严重影响网络性能和网络正常业务运行的P2P应用流量
- 与业务无关，影响工作和学习效率的网络游戏、网上购物等网络行为产生的流量
- 各种DOS/DDOS攻击；病毒、蠕虫、木马等恶意代码
- 垃圾邮件；反动、迷信、色情、赌博、邪教等危害人们身心健康的网络流量

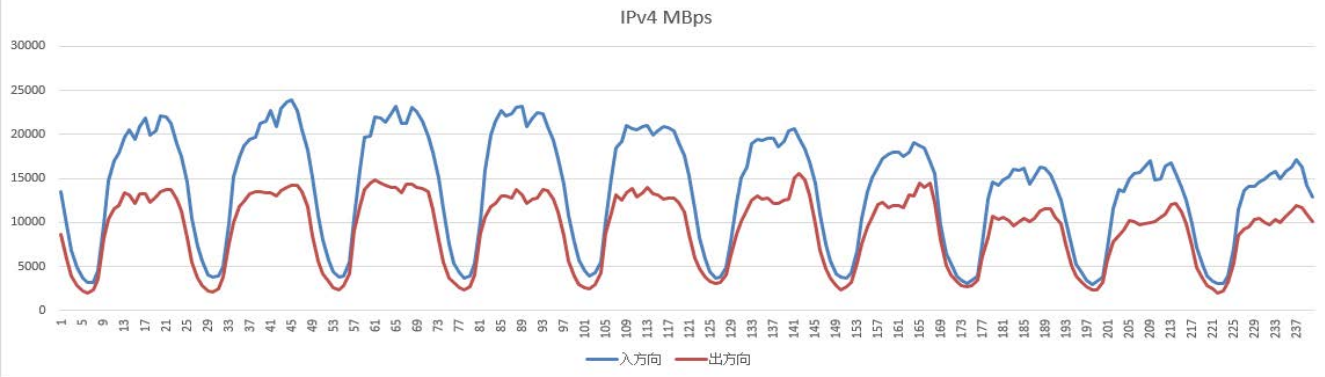
01 / 研究背景及意义

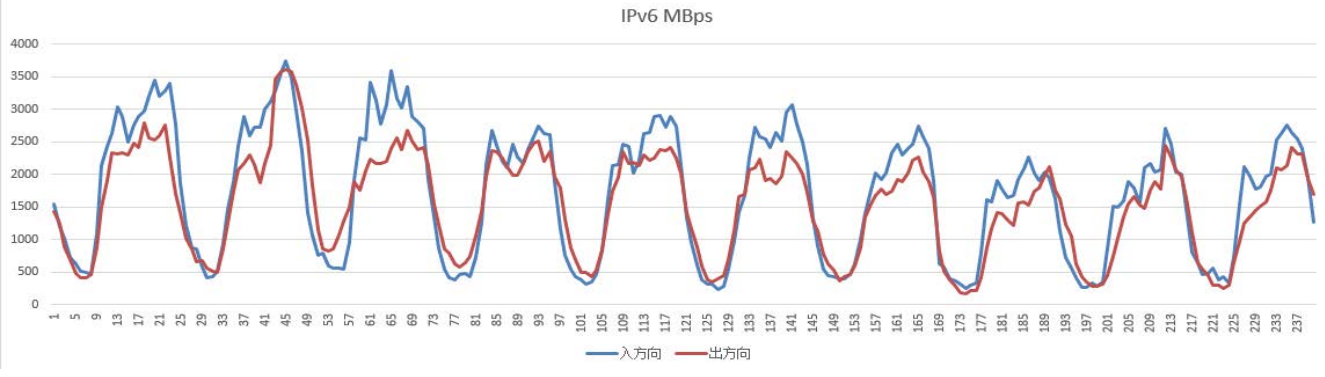
通过对IPv4和IPv6网络流量行为进行对比，为今后发现IPv6网络流量行为特点，对提高IPv6网络的安全性和可靠性，提高服务质量，提供重要的参考依据。

为此，本文尝试基于CERNET南京主节点IPv4和IPv6网络边界路由器提供的流记录数据，分别进行面向基础流量行为和端口流量行为的比对分析。

01 / 研究背景及意义

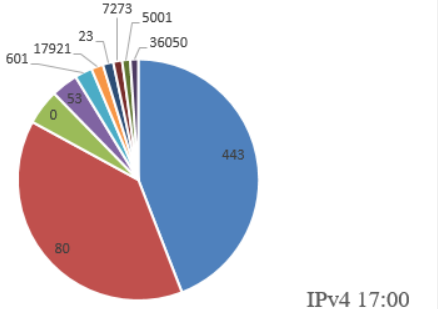
基础流量行为是指以带宽占用为基本测度的时间序列。



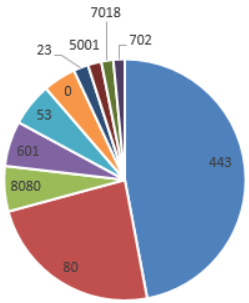


01 / 研究背景及意义

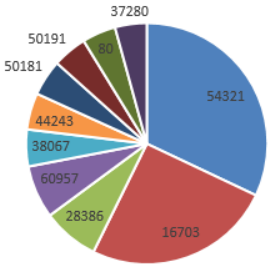
在C/S结构中一般是通过“IP地址+端口号”来定位应用的。端口流量的分布规律，可用于从整体上了解网络服务的状态、发现可能的异常。



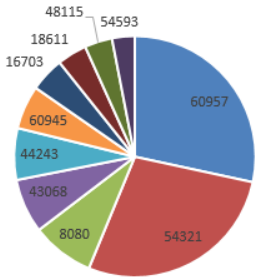
IPv4 17:00



IPv4 02:00



IPv6 17:00



IPv6 02:00

02

Part two

数据源与研究方法

02 / 数据源与研究方法

观测位置

CERNET南京主节点所管辖的IPv4和IPv6网络边界，流记录抽样比均为256，目前IPv4网有153个接入单位，接入带宽为40Gbps；IPv6网有148个接入单位，接入带宽为10Gbps。

共有120个单位既处于IPv4网又在IPv6网。二者接入单位数目相似，覆盖面和用户群体相同，因此具有良好的对比条件。

分析数据源

根据观测角度，需要两方面的数据源，分别是IPv4和IPv6网络总流量和端口流量。在同等的硬件设施、数据源基本情况、统计算法和用户群等条件下，选择从2019年7月10日起连续7天的数据作为分析源，以1小时为一个时间粒度，共有7*24=168个时间粒度的数据。

02 / 数据源与研究方法

网络总流量数据源

入方向报文数	出方向报文数	入方向字节数	出方向字节数

表1 网络总体流量

端口流量数据源

入宿端口报文数	出源端口报文数	入宿端口字节数	出源端口字节数

表2-1 网内端口流量

入源端口报文数	出宿端口报文数	入源端口字节数	出宿端口字节数

表2-2 网外端口流量

研究方法：端口流量分布熵

熵代表一个系统“内在的混乱程度”，体现了系统的不确定性和无序性。本文用香农熵代表端口流量的整体分布情况。

研究方法：Spearman相关系数

利用两变量的秩次大小作线性相关分析，它对数据条件的要求比较宽松，不论两个变量的总体分布形态、样本容量的大小如何，都可以用该方法来进行研究.

±0.8-±1.0	极强相关
±0.6-±0.8	强相关
±0.4-±0.6	中等程度相关
±0.2-±0.4	弱相关
±0.0-±0.2	极弱相关或无相关

03

Part three

研究方案及结果

基础流量行为对比测度：

数据源中的4种原始测度。

端口流量行为对比测度：

对数据源中的8种原始测度进行处理，得到每个时间粒度排在前三名的端口流量占比、HTTP类服务流量占比和端口流量分布熵。

最后对IPv4和IPv6网的上述测度的时间序列进行Spearman相关性分析。

以入方向报文数为例，描述具体对比过程

1）**获取分析测度**

设统计集$in_pkts_i=\{m_j, j=1,2,...,24\}$，（i=1,2...,7），表示第i天的入方向报文数统计集，

m_j表示第i天第j个时间粒度的入方向报文数；

设统计集$in_pkts7days=\{m_j, j=1,2,...,168\}$，in_pkts7days表示7天的测度汇总后的统计集。

流量数据在观测位置被截取并保存，程序从IPv4网和IPv6网数据库中分别获取统计集。

2）**相关性分析**

在数学软件MATLAB中将IPv4和IPv6网每天的统计集分别存储为24*1的向量Ai、Bi（i=1,2,...,7），将7天总的统计集存储为168*1的向量C、D，分别调用corr()函数进行相关性分析，即

corr(Ai,Bi,'type','spearman')和corr(C,D,'type','spearman')。

基础流量行为对比结果

日期	按入方向报文数计算的相关系数	按出方向报文数计算的相关系数	按入方向字节数计算的相关系数	按出方向字节数计算的相关系数
7月10日	0.957391	0.963478	0.909565	0.921739
7月11日	0.962609	0.938261	0.975652	0.92087
7月12日	0.945217	0.903478	0.956522	0.856522
7月13日	0.890435	0.869565	0.822609	0.843478
7月14日	0.923478	0.895652	0.854783	0.786087
7月15日	0.947826	0.936522	0.951304	0.910435
7月16日	0.993043	0.98087	0.973043	0.934783
7天	0.933069	0.929703	0.910438	0.854798

前三名端口流量占比对比结果

名次	按入宿报文数占比计算的相关系数	按出源报文数占比计算的相关系数	按入宿字节数占比计算的相关系数	按出源字节数占比计算的相关系数
1	-0.108576	-0.128916	0.298577	0.095559
2	-0.354984	-0.156101	0.585086	-0.190443
3	0.596921	0.395012	0.673752	0.490696

名次	按入源报文数占比计算的相关系数	按出宿报文数占比计算的相关系数	按入源字节数占比计算的相关系数	按出宿字节数占比计算的相关系数
1	-0.020077	-0.296977	-0.073663	0.086512
2	-0.200317	-0.432381	-0.196118	0.109729
3	0.448376	0.556374	0.268869	0.114828

443和80端口流量占比对比结果

HTTP类端口	按入宿报文数占比计算的相关系数	按出源报文数占比计算的相关系数	按入宿字节数占比计算的相关系数	按出源字节数占比计算的相关系数
443	-0.435176	-0.366873	-0.206806	-0.336634
80	0.247399	0.138621	0.299937	0.296139

HTTP类端口	按入源报文数占比计算的相关系数	按出宿报文数占比计算的相关系数	按入源字节数占比计算的相关系数	按出宿字节数占比计算的相关系数
443	0.101813	0.325571	0.125168	0.322383
80	0.249733	0.415244	0.437586	0.348288

端口流量分布熵对比结果

日期	按入宿报文数熵计算的相关系数	按出源报文数熵计算的相关系数	按入宿字节数熵计算的相关系数	按出源字节数熵计算的相关系数
7天	0.636834	0.605239	0.746694	-0.62735

日期	按入源报文数熵计算的相关系数	按出宿报文数熵计算的相关系数	按入源字节数熵计算的相关系数	按出宿字节数熵计算的相关系数
7天	-0.28301	0.642412	-0.361256	0.829047

04

Part four

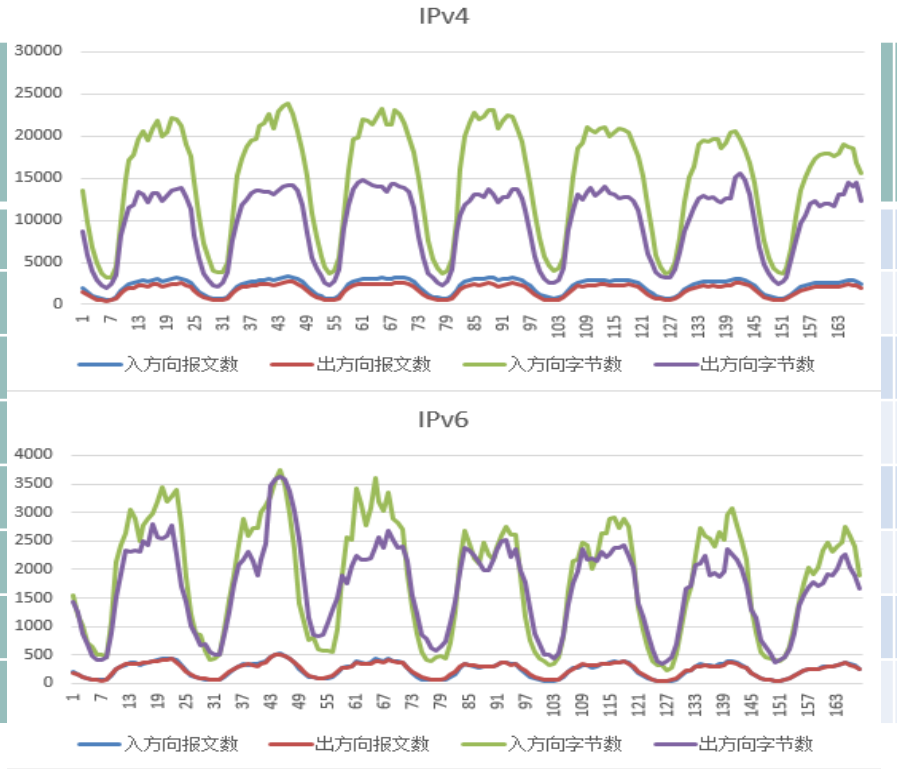
结果分析及总结

04 / 结果分析与总结

网络基本流量

基础流量行为对比结果

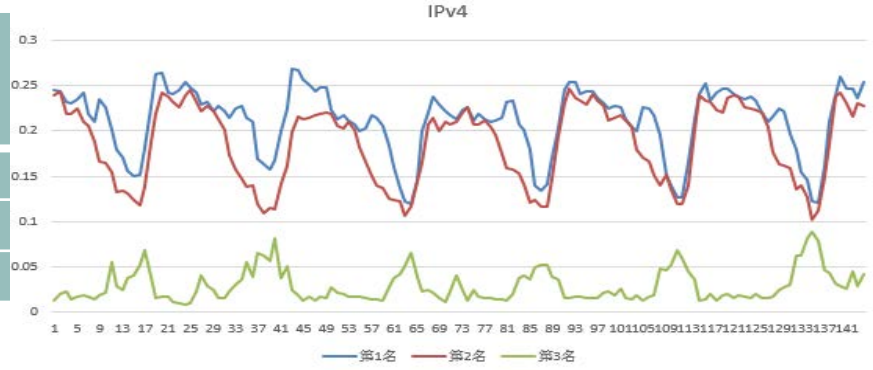
日期	按出方向字节数计算的相关系数
7月10日	0.921739
7月11日	0.92087
7月12日	0.856522
7月13日	0.843478
7月14日	0.786087
7月15日	0.910435
7月16日	0.934783
7天	0.854798



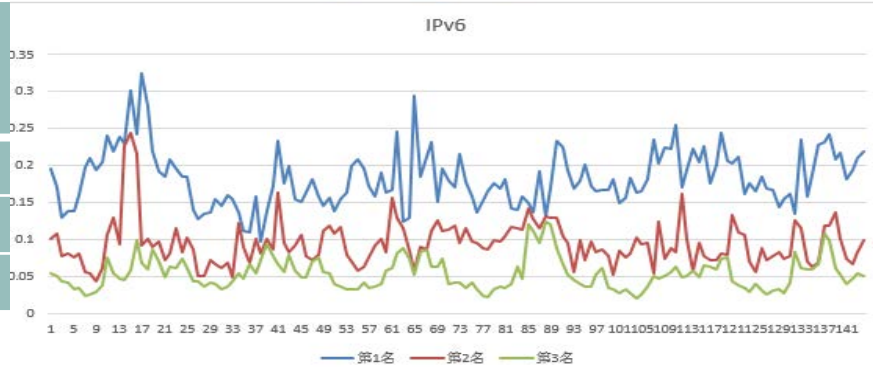
04 / 结果分析与总结

前三名端口流量占比对比结果

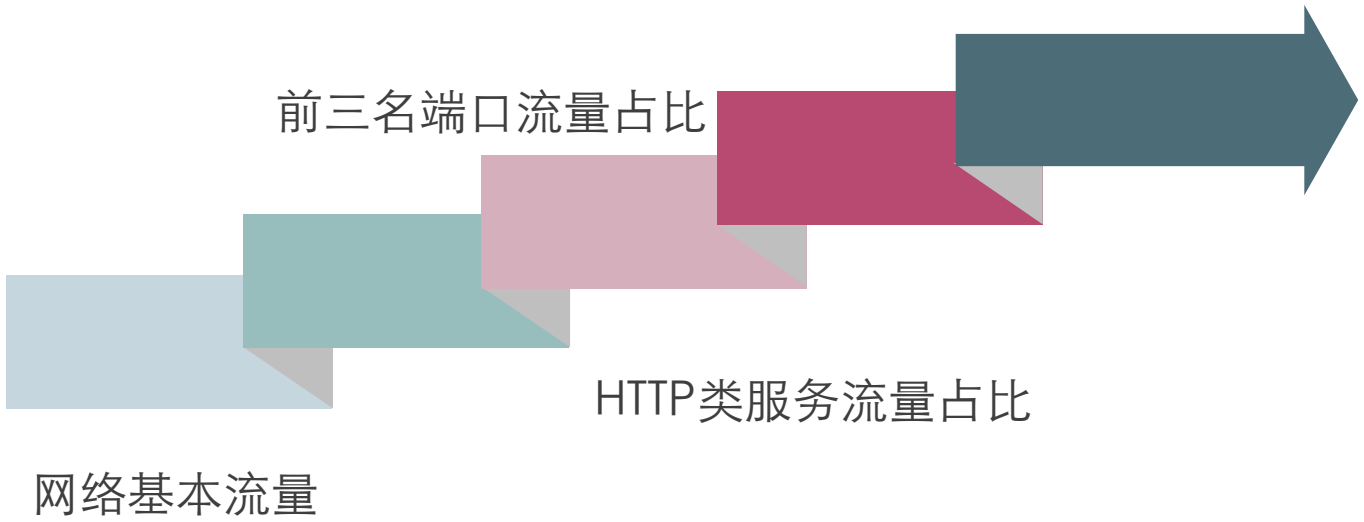
名次	按出源字节数占比计算的相关系数
1	0.095559
2	-0.190443
3	0.490696



名次	按出宿字节数占比计算的相关系数
1	0.086512
2	0.109729
3	0.114828



04 / 结果分析与总结



443和80端口流量占比对比结果

HTTP类端口	按入宿报文数占比计算的相关系数	按出源报文数占比计算的相关系数	按入宿字节数占比计算的相关系数	按出源字节数占比计算的相关系数
443	-0.435176	-0.366873	-0.206806	-0.336634
80	0.247399	0.138621	0.299937	0.296139

HTTP类端口	按入源报文数占比计算的相关系数	按出宿报文数占比计算的相关系数	按入源字节数占比计算的相关系数	按出宿字节数占比计算的相关系数
443	0.101813	0.325571	0.125168	0.322383
80	0.249733	0.415244	0.437586	0.348288

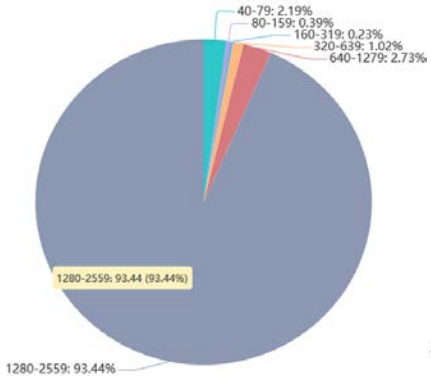
IPv4和IPv6网流量前三名端口

名次	按入报文数排名	按出报文数排名	按入字节数排名	按出字节数排名
1	443	80	443	80
2	80	443	80	443
3	53	123	5001	123

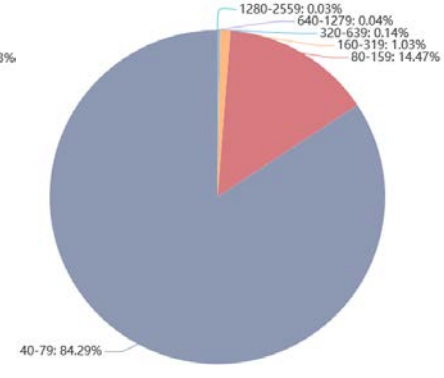
名次	按入报文数排名	按出报文数排名	按入字节数排名	按出字节数排名
1	16703	54321	16703	54321
2	54321	16703	10137	16703
3	10137	10137	54319	51413

04 / 结果分析与总结

基于统计数据和报文头



SrcPort=54321包长分布



DstPort=54321包长分布

可视信息：

（1）数据均来源于TCP报文，校园网内的一组服务器使用54321端口向外发送大量的数据包而反方向相对较少。

（2）校园网内与54321端口所绑定服务相关的IP地址前缀范围7天内没有太大变化，地址个数些许变化，即有特定前缀的IP地址与54321服务相关。

	srcPort =54321平均报文长	dstPort=54321平均报文长	使用54321的校园网地址个数	常用地址范围
10.24	1378.89	72.37	65	
10.25	1387.08	71.62	25	
10.26	1374.66	70.17	60	2001:da8:1006:4000::
10.27	1381.55	70.77	48	2001:da8:1006:9000::
10.28	1370.91	70.4	57	2001:da8:1006:9001::
10.29	1364.95	71.85	54	
10.30	1385.67	70.75	50	

04 / 结果分析与总结

基于报文体

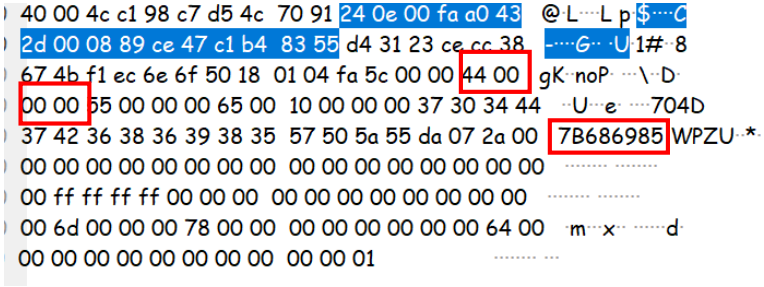


图1

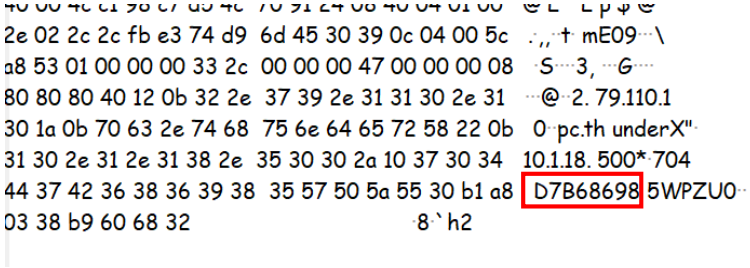


图2

可视信息：

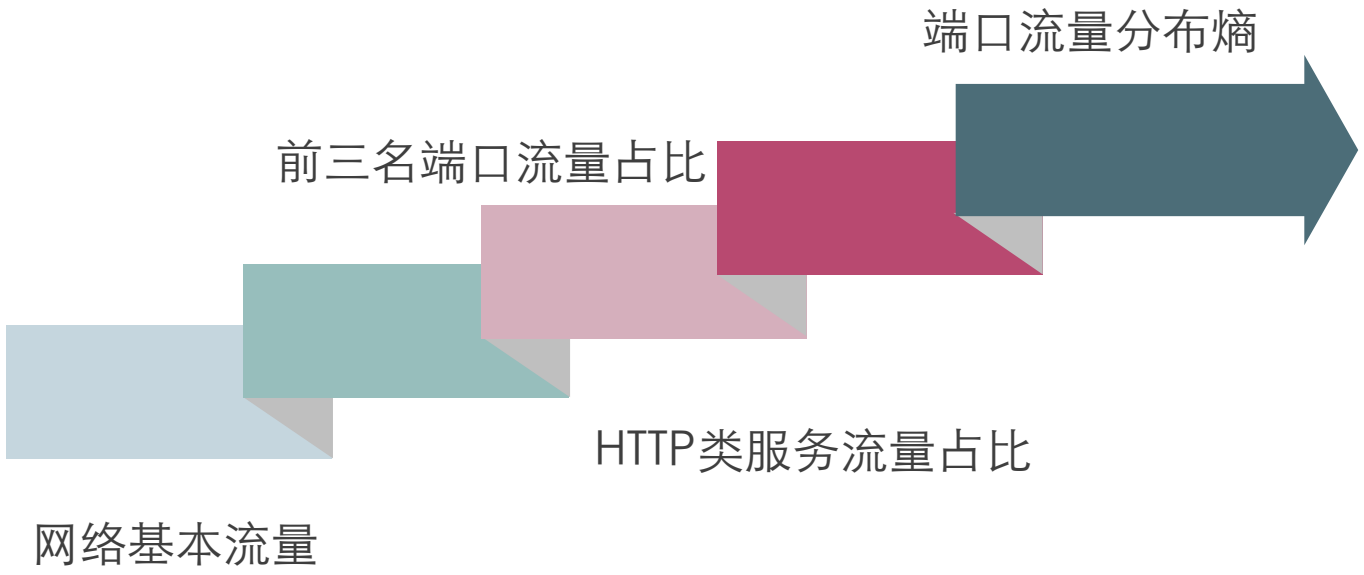
（1）图1：观察报文内容，传送54321端口服务的报文中，第一个分片数据字段都带有440000开头，因此可判定为一个协议头。

（2）图2：含有440000报文内容中出现疑似序列号字段，在相应IP但非54321端口报文中发现一模一样的序列号字段，且含有thunderX字段。

结论：

54321端口流量属于正常的P2P流，从它的流量占比情况来看，P2P流占据相当大一部分的网络资源，该现象在IPv6网中极为常见。

04 / 结果分析与总结



端口流量分布熵对比结果

日期	按入宿报文数熵计算的相关系数	按出源报文数熵计算的相关系数	按入宿字节数熵计算的相关系数	按出源字节数熵计算的相关系数
7天	0.636834	0.605239	0.746694	-0.62735

日期	按入源报文数熵计算的相关系数	按出宿报文数熵计算的相关系数	按入源字节数熵计算的相关系数	按出宿字节数熵计算的相关系数
7天	-0.28301	0.642412	-0.361256	0.829047

04 / 结果分析与总结

总结:

对于网络基本流量行为，IPv4和IPv6网表现基本一致，符合以天为单位的季节模型，存在一些细微的差别，针对可能存在的差别，对端口流量进行了对比研究。

通过比较和查询排名靠前的端口，得知IPv4网络应用中HTTP类服务占绝对地位，IPv6网络中未知端口却异常活跃；
通过对比端口熵值，可见IPv4网络流量在绝大部分时间处于稳定状态，而IPv6网入方向源端口流量行为分布更加不稳定。

经过改进的IPv6协议并不完全可靠，端口流量变化存在很多不确定性，且针对端口的漏洞仍然不可忽视，更应加强对IPv6网的管理和掌控。

CERNET 2019

感谢您的聆听

Thanks for listening!