网络测量

孙毅

内容

- 网络测量概述

- 网络测量方法
  - 可用带宽测量
  - 瓶颈带宽测量
  - 网络拓扑探测

- 研究实例
  - 利用视频流量预测网络带宽和可用带宽

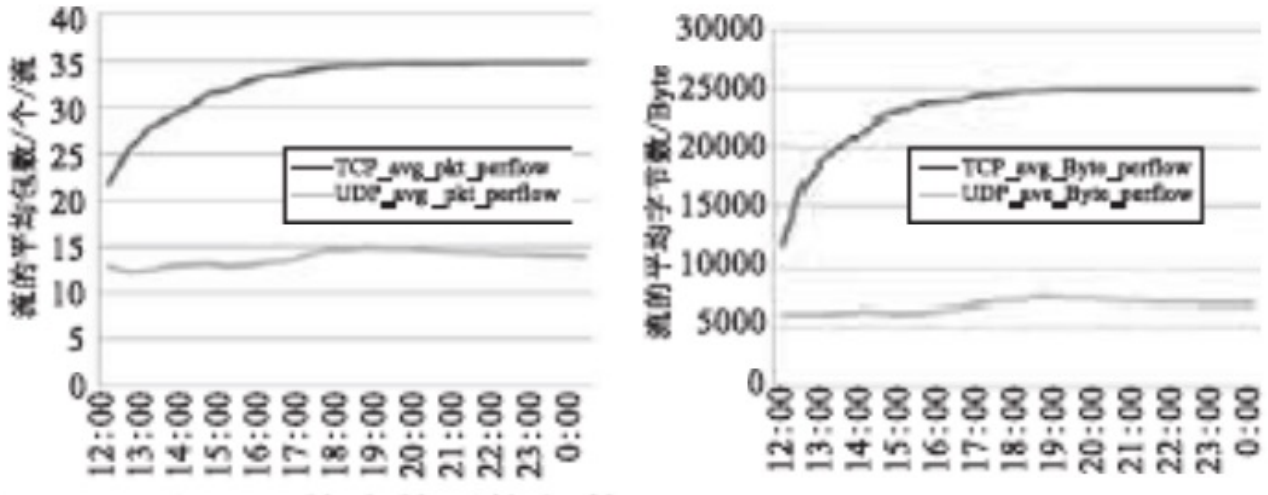
定义

- 网络测量是按照一定的方法和技术，利用软件或硬件工具来量度网络的运行状态、表征网络特性的一系列活动的总和。

- 网络测量的应用：
  - 监测网络故障
  - 测试协议行为
  - 刻画流量特征
  - 评估网络性能

网络流量测量的几个重大发现

- TCP协议占据了大部分网络流量
  - 早期（2006年以前）：TCP占80%以上
  - 中期（2006年~2012年）：TCP、UDP各占半壁江山，音视频、游戏等
  - 现阶段（2013年至今）：TCP重拾优势，http视频



网络流量测量的几个重大发现

- 网络流量是双向的，但通常非对称
  - 大部分应用会产生双向流量，但双向非对称
  - C/S、B/S架构，下载密集型
  - 以太网全双工，上先行带宽一致，2G、3G、4G网络上下行带宽非对称

网络流量测量的几个重大发现

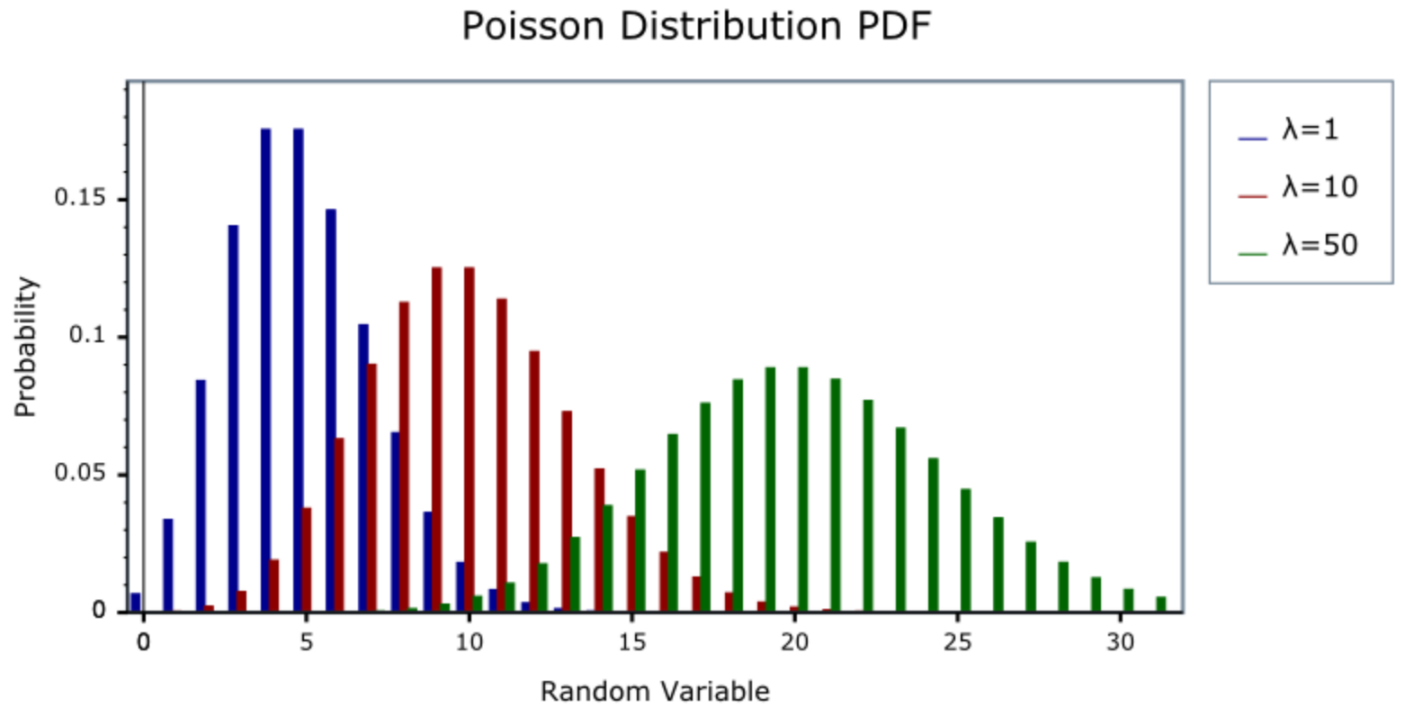
- 大象流与老鼠流
  - 在AS之间占总数9 %的流承载了90 %的流量
  - 数据包大小的分布两极化（50%MTU，40%小于40个字节如TCP ACK）

网络流量测量的几个重大发现

- 蜻蜓流和乌龟流
  - 总量45%的网络流在持续时间上小于2s
  - 不到2%的网络流在持续时间上超过15min并且承载了50%以上的流量

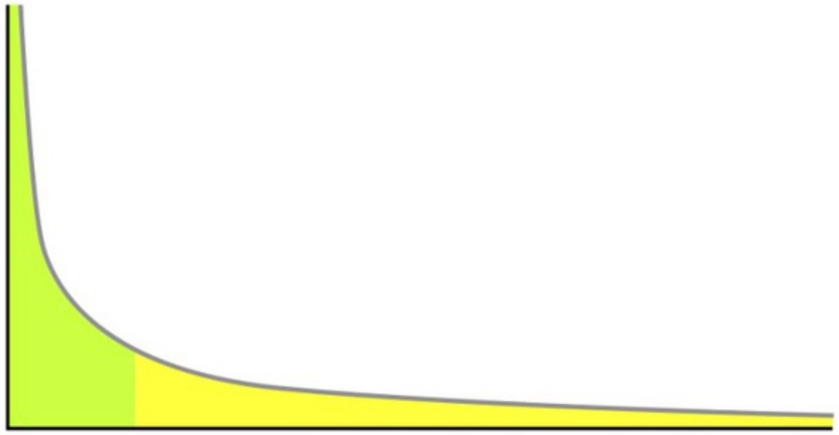
网络流量测量的几个重大发现

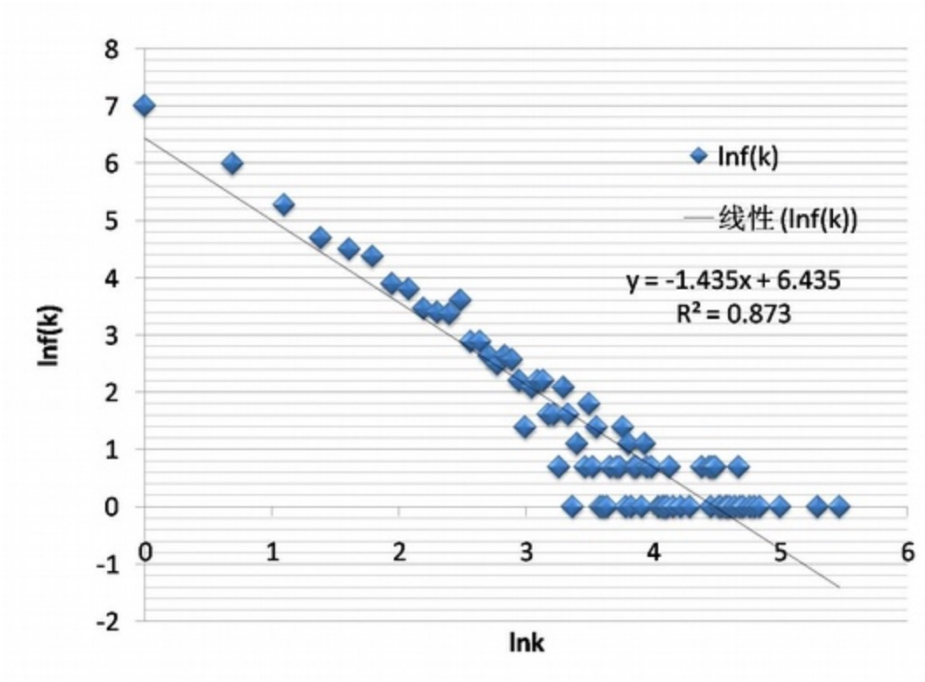
- 数据对话请求(session)的到达则服从泊松分布（用户访问服从泊松分布）
- 数据包(package)的到达不服从泊松分布（数据包到达具有突发性）



网络流量测量的几个重大发现

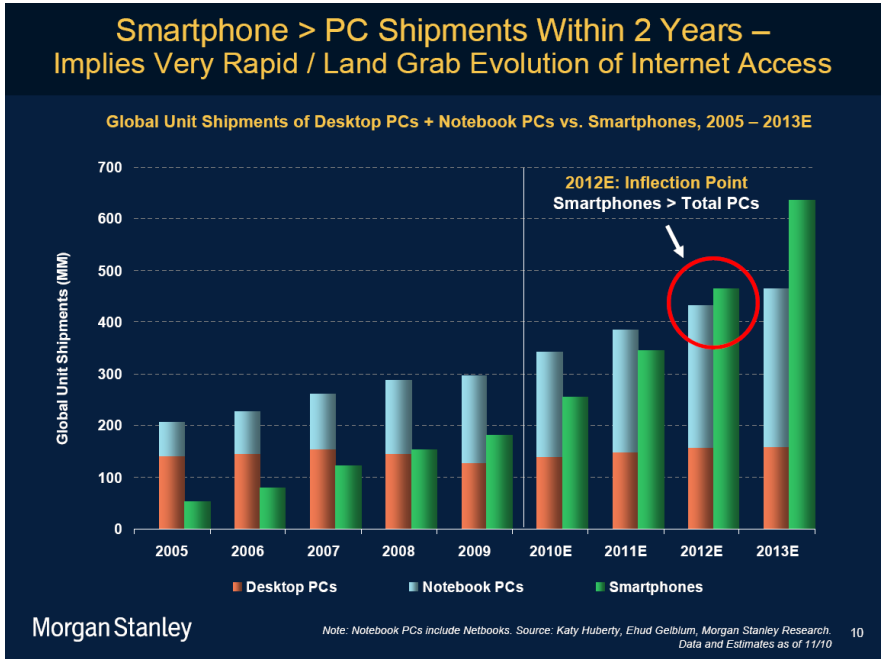
- 流量分布并不均衡
  - 由于C/S模式，10%的主机占据了90%的流量
  - 幂律分布
- 网络流量带有明显“本地”特征
  - 时间局部性、空间局部性

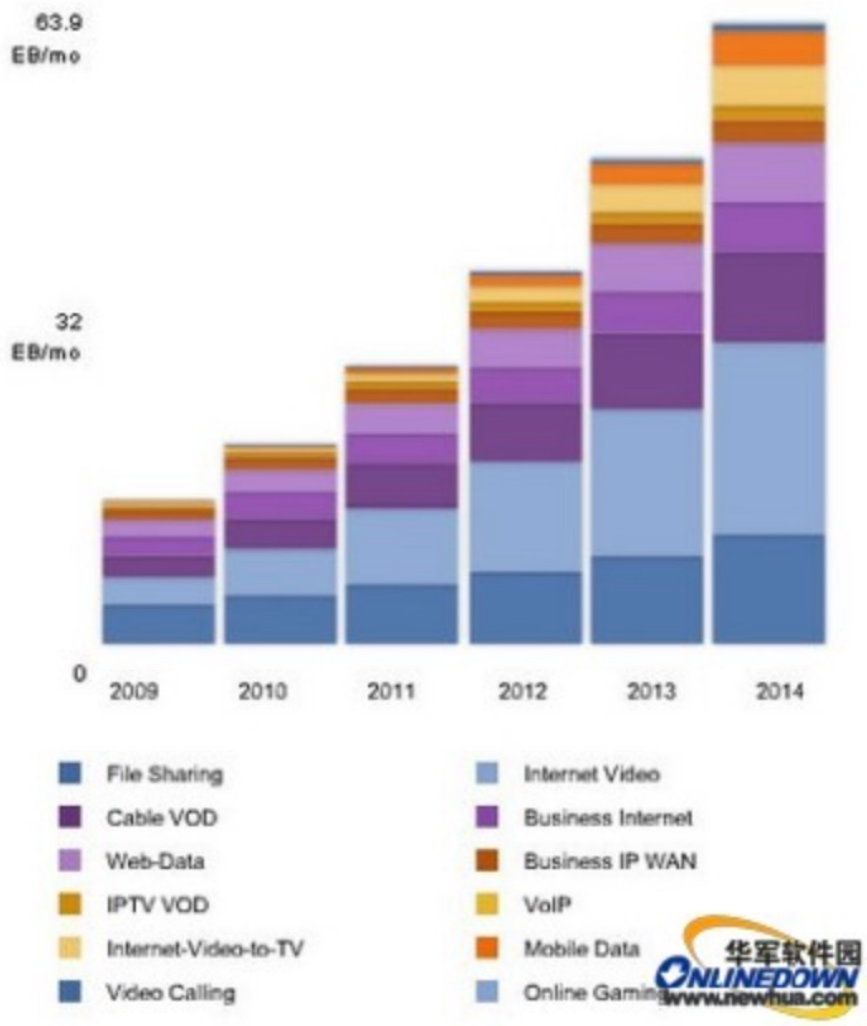




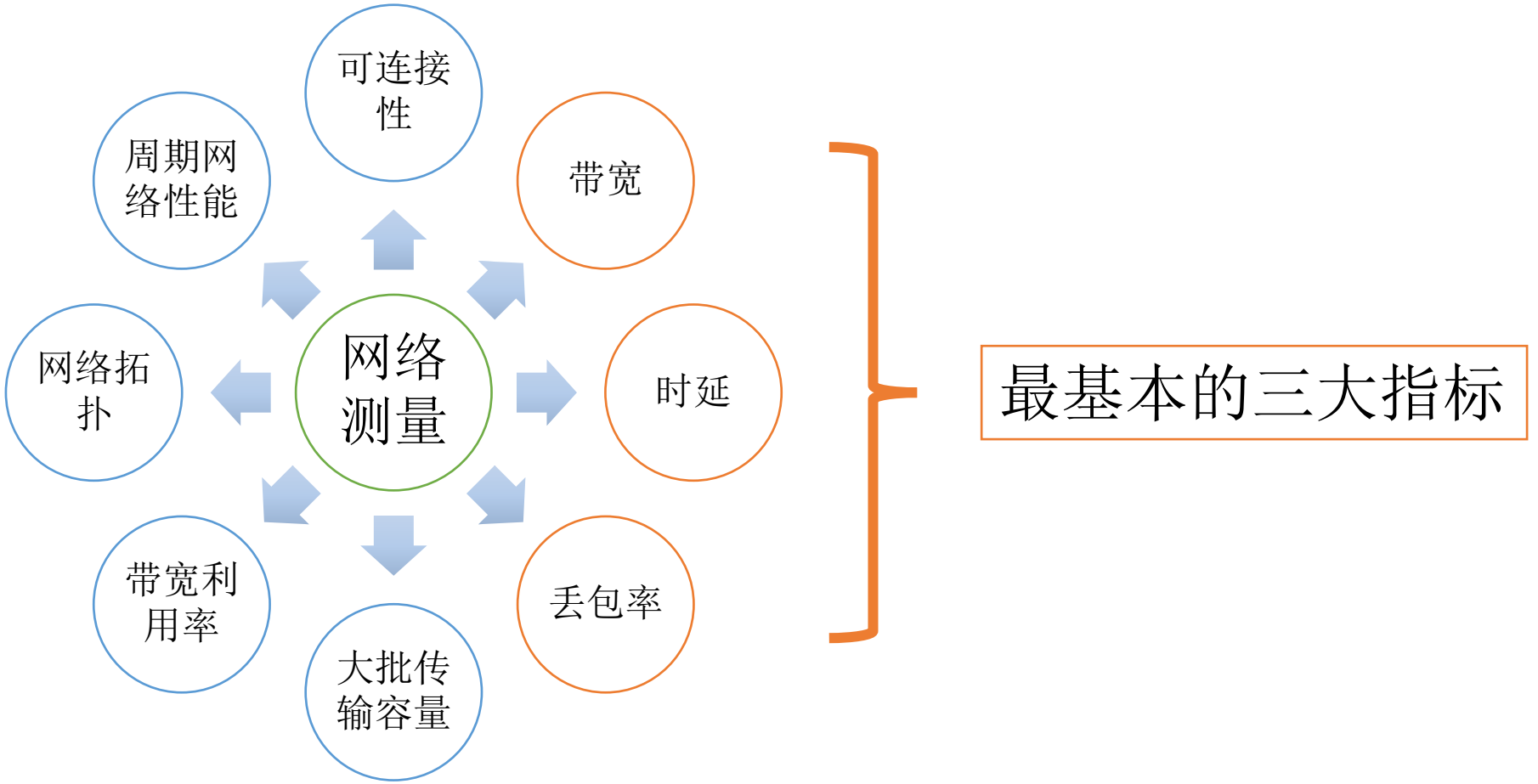
网络流量测量的十大发现

- 网络流量不断在变化
  - 视频流量、移动端流量的增加…

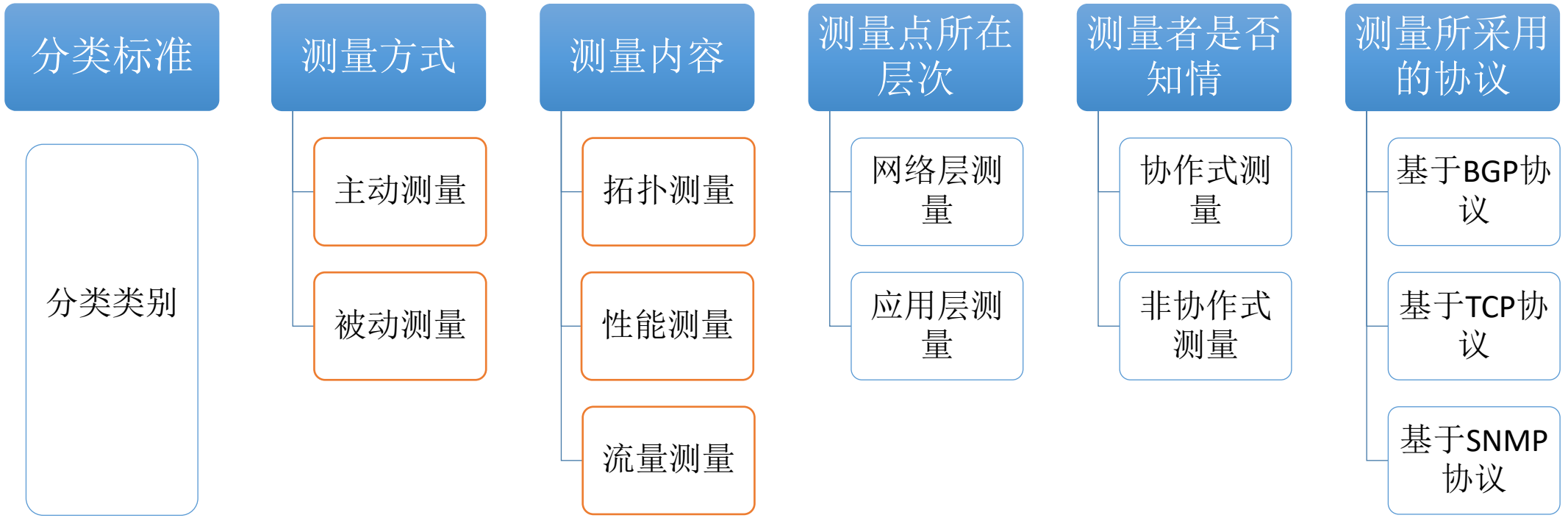




网络测量的指标



网络测量的分类


分类标准
分类类别
测量方式
主动测量
被动测量
测量内容
拓扑测量
性能测量
流量测量
测量点所在层次
网络层测量
应用层测量
测量者是否知情
协作式测量
非协作式测量
测量所采用的协议
基于BGP协议
基于TCP协议
基于SNMP协议

测量方式分类

- 主动测量
  - 指由测量用户主动发起测量，将探测分组注入网络，根据测量数据流的传送情况来分析网络的性能。
  - 优点
    - 使用方便，适合端到端的网络性能测量，对于需要关心的内容只要在本地发送测试包观察网络的响应即可
    - 由于该方法不涉及用户的网络信息，所以对用户而言是很安全的
  - 缺点
    - 增加了网络潜在的负载，尤其是如果该测量未经仔细设计，使产生的流量达不到最小，可能会对网络造成较大的影响
    - 需要消耗较多的计算资源

测量方式分类

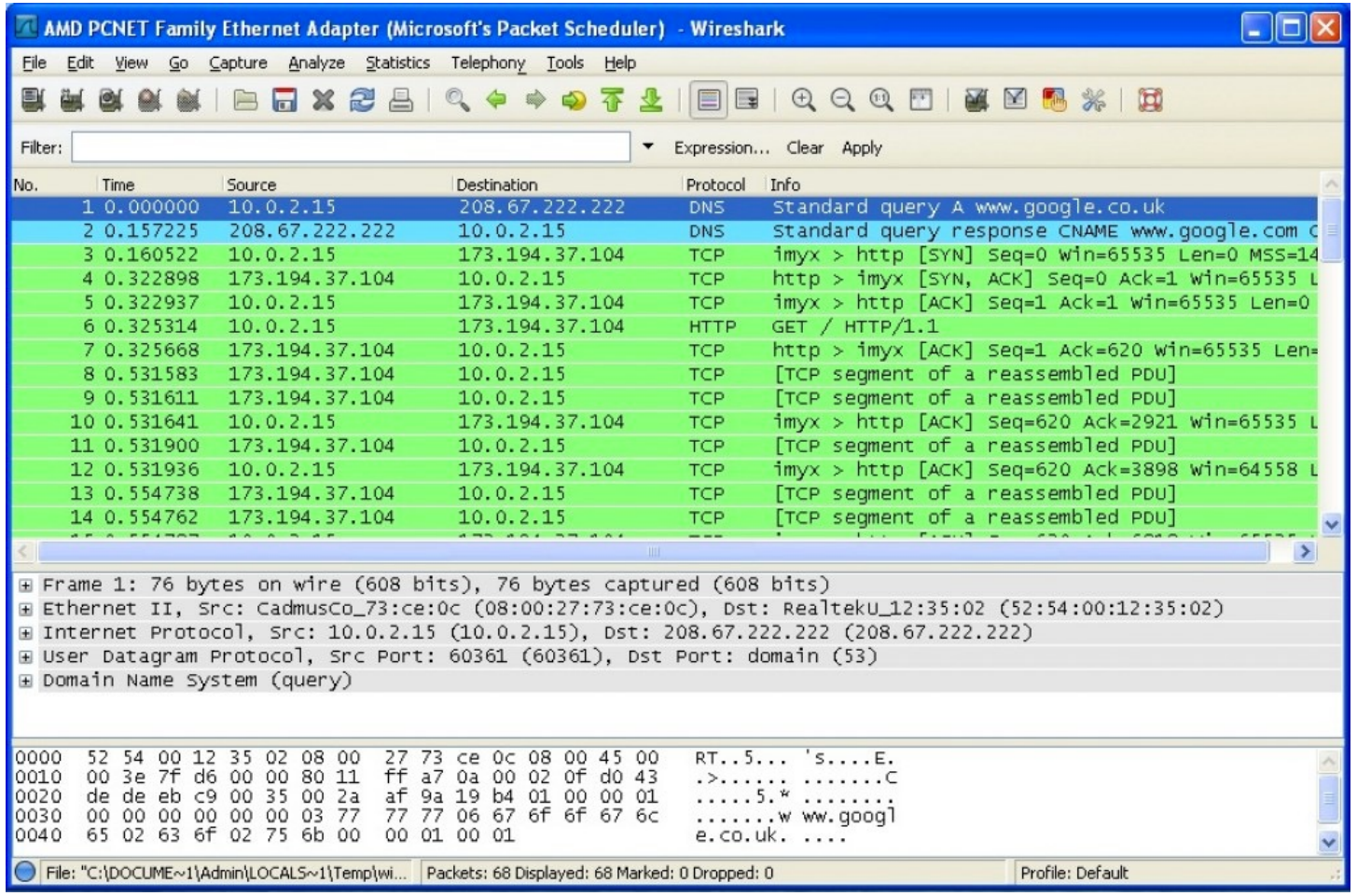
- 被动测量
  - 通过在网络中的链路或设备（如路由器、交换机等）上借助包捕获数据的方式来记录网络流量，分析流量，获知网络的性能状况。
  - 优点
    - 测量的是网络上的真正流量
    - 能够达到对观察点网络行为的详尽理解
  - 缺点
    - 被动测量方式可能要查看网络上的所有数据包，容易捕获网络中的敏感信息，给用户信息的保密和安全带来一定威胁
    - 只能获得网络局部数据，无法了解网络整体状况
    - 测量范围受限

测量内容分类

- 网络拓扑测量
  - 了解网络拓扑结构，用于资源调度和流量分配。
- 性能测量
  - 通过监测网络端到端时延、抖动、丢包率等特性，了解网络的可达性、利用率以及网络负荷等。
- 流量测量
  - 对网络数据流的特性进行监测和分析，以掌握网络的流量特性，如协议的使用情况、应用的分布和用户的行为特征等。

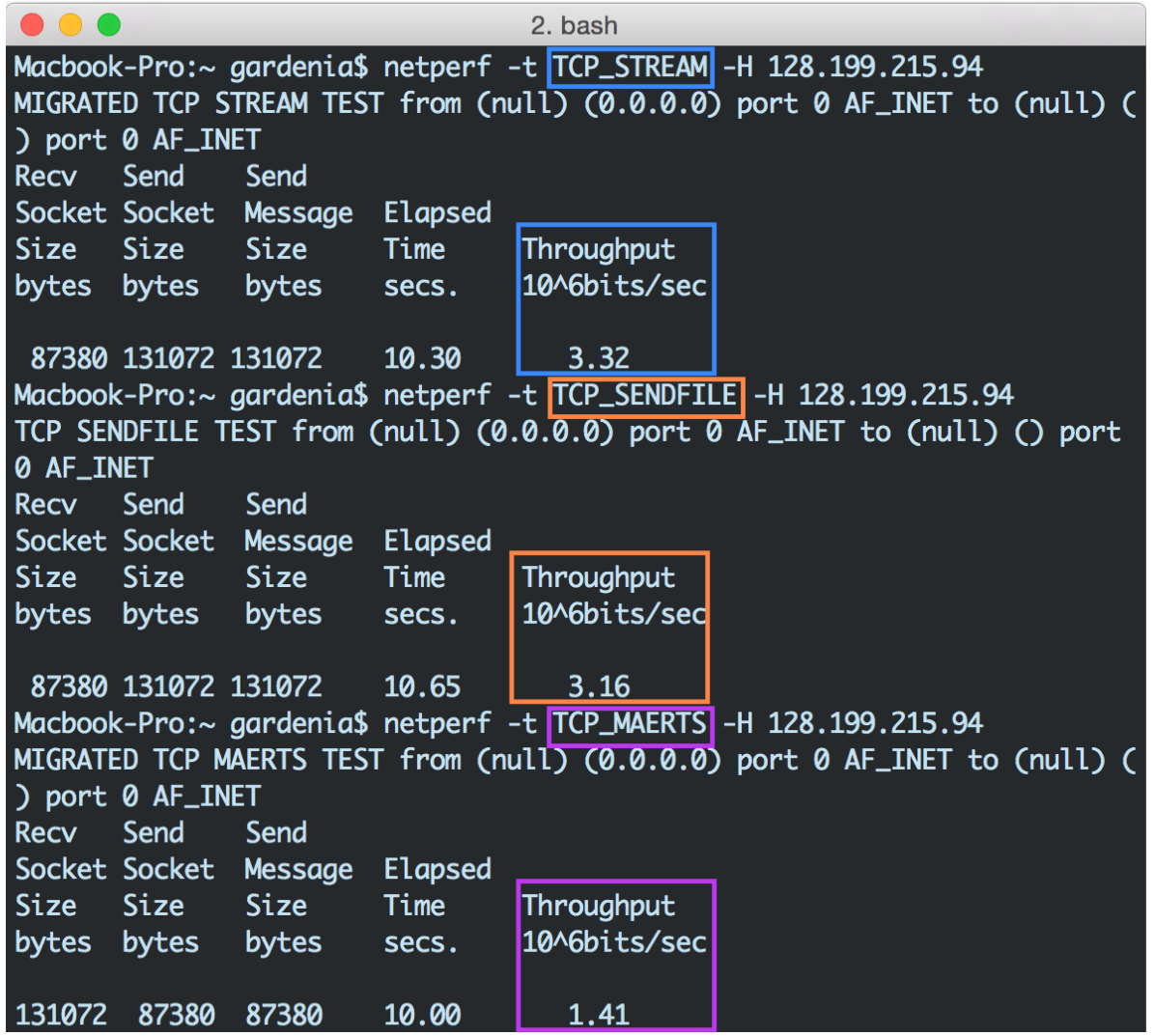
例1：网络嗅探器

- 测量方式：被动测量
- 测量内容：网络流量
- 右图为Wireshark测量的一次访问Google请求，包括DNS解析和TCP三次握手。



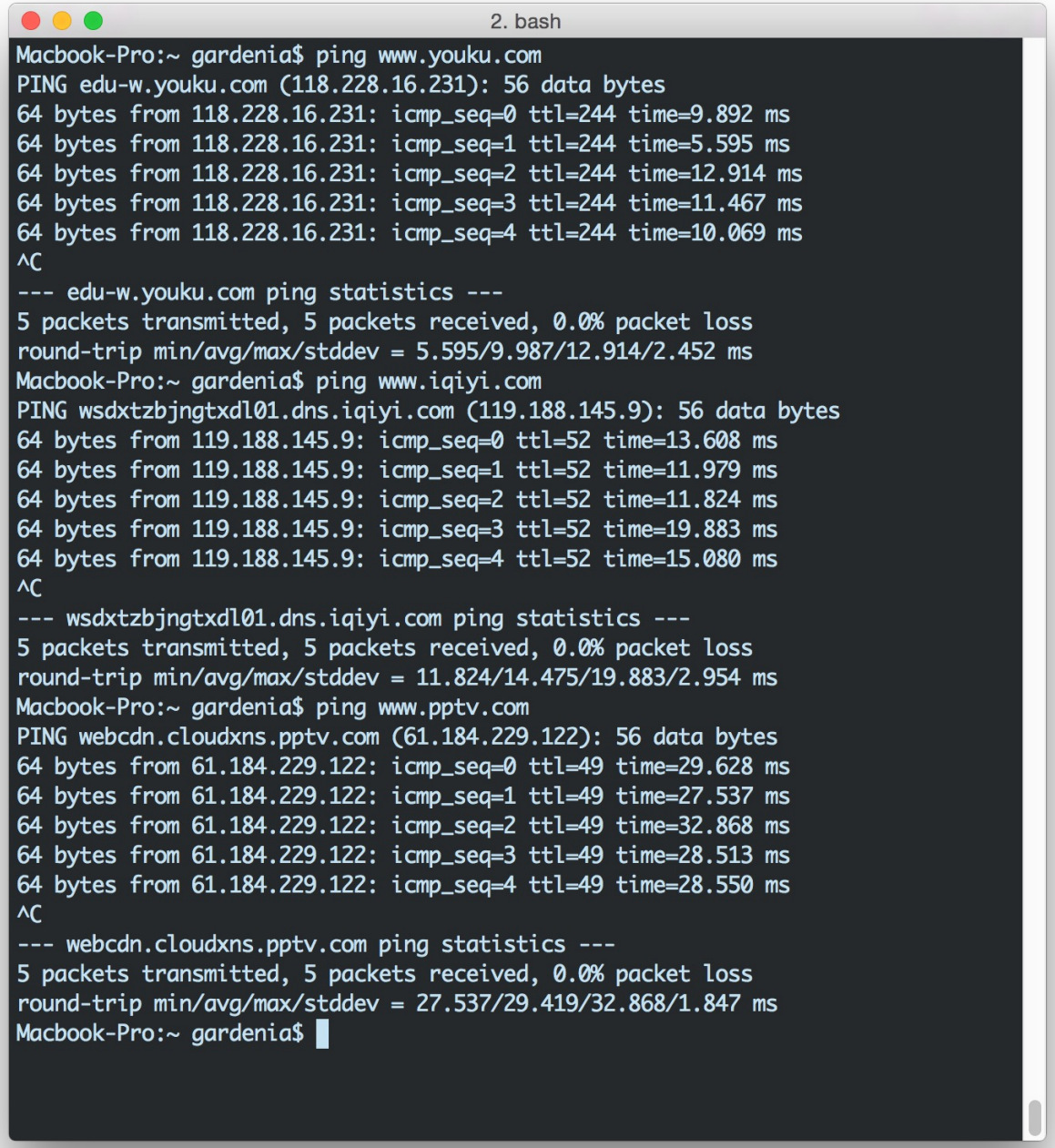
例2：Netperf

- 测量方式：主动测量
- 测量内容：网络性能（带宽）
- 测量点所在层次：网络层
- 测量者是否知情：协作式
- 测量协议：TCP、UDP
- 右图为Netperf对几种TCP协议测量得到的带宽值

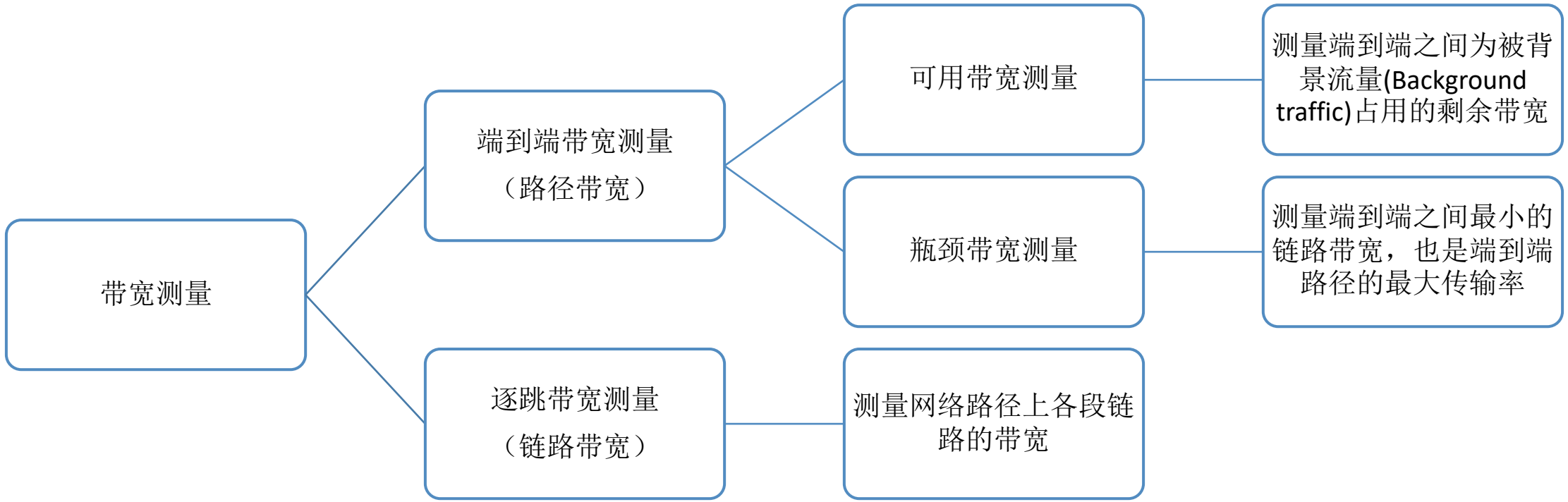


例3：Ping

- 测量方式：主动测量
- 测量内容：网络性能（可连接性、延迟、丢包、抖动）
- 测量点所在层次：网络层
- 测量协议：ICMP
- 右图为Ping测试几个视频站点的结果



网络带宽测量


带宽测量
端到端带宽测量
（路径带宽）
逐跳带宽测量
（链路带宽）
可用带宽测量
瓶颈带宽测量
测量网络路径上各段链路的带宽
测量端到端之间为被背景流量(Background traffic)占用的剩余带宽
测量端到端之间最小的链路带宽，也是端到端路径的最大传输率

带宽/容量的概念

- 链路的带宽(Bandwidth)或者容量(Capacity)是指该链路上数据报文的最大传输速率。

- 网络路径的瓶颈带宽(Bottleneck Bandwidth)或者容量指的是源节点到目的节点之间处理能力最低的链路所能达到的最大的数据传输速率。

- 传输路径上带宽/容量最小的链路称为该路径的瓶颈链路（Narrow Link）。

可用带宽的概念

可用带宽(Available Bandwidth)是指当应用程序和其它背景流(Cross Traffic)共享网络路径时，该应用程序所能得到的带宽。也就是指网络在不降低其它业务流的传输速率的情况下，所能提供给一个业务流的最大传输速率。

- 传输路径上可用带宽最小的链路称为该路径的Tight Link。

可用带宽测量算法

- 用$(C_0,C_1,C_2,\cdots,C_H)$表示一条从源端到目的端的路径，H表示路径的跳数，Ci表示链路i的链路带宽，$u_i(0\leqslant u_i\leqslant 1)$表示链路的利用率
- 路径可用带宽可表示为：

$$A=\min[C_i(1-u_i)] \quad (i=0..H)$$

- 测量前提假设
  1. 网络路径上所有路由器的排队模式都是先进先出(FIFO);
  2. 背景流量处于平稳状态;
  3. 背景流量的平均速率变化比较慢，并且在单个测量周期中保持恒定

可用带宽测量算法

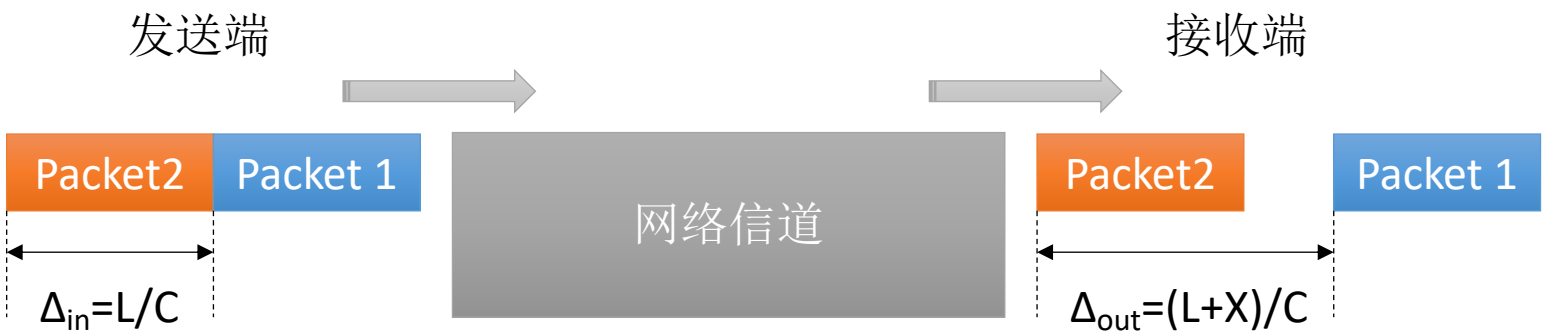
- 探测报文间隔模型(The Probe Gap Model, PGM)
  - 假设1：路径上容量最小的链路(窄链路,narrow link)和可用带宽最小的链路(紧链路, tight link)必须是同一条链路
  - 假设2：路径各链路的带宽C已知
- 从发送端发送两个连续数据包，设数据包发送的时间间隔为 $\Delta_{in}=L/C$ (1)
- 接收端接受这对数据包的时间间隔为 Δ_{out}
- 数据包的大小为L，可用带宽为A
- 那么，在两个数据包发送的时间间隔 Δ_{in} 内引入的额外数据量X为：

$$X=(C-A)\Delta_{in} \quad (2)$$

- 则 $\Delta_{out}=(L+X)/C$ (3)
- 根据(1),(2),(3)，可得：

$$A = C(1 - \frac{\Delta_{out} - \Delta_{in}}{\Delta_{in}})$$

可用带宽测量算法



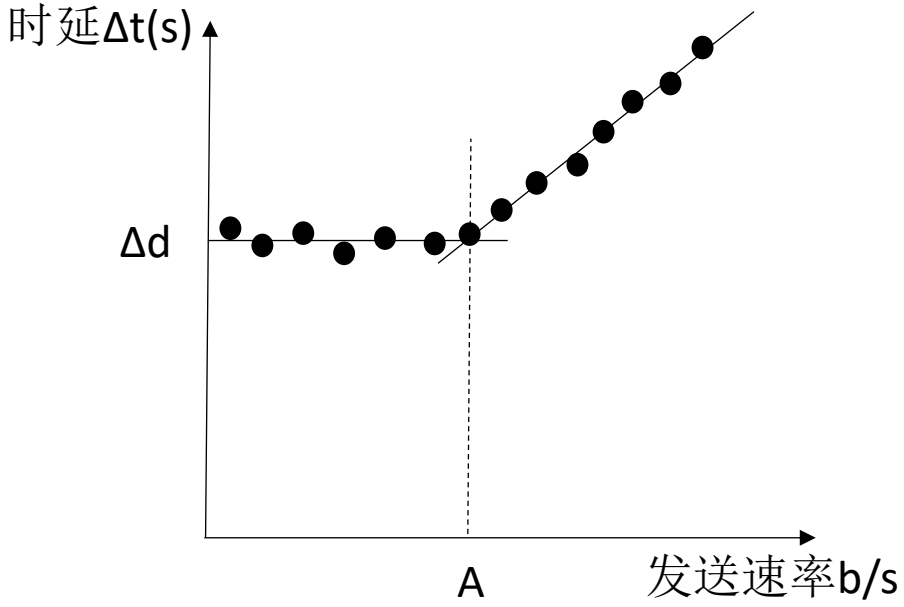
时间间隔Δ_{in}内引入的额外数据量

$X = (C-A)\ \Delta_{in}$

可用带宽计算公式：$A = C(1 - \frac{\Delta_{out} - \Delta_{in}}{\Delta_{in}})$

可用带宽测量算法

- 探测报文速率模型(PRM)
  - 利用自导拥塞思想进行带宽测量
  - 当测试报文发送速率小于链路可用带宽时,传输时延相对固定,由网络物理特性决定
  - 当测试报文发送速率大于链路可用带宽时,网络出现排队现象,传输时延增大,则导致时延增大的发送速率转折点A处对应的速率即代表了该链路最大可用带宽，即路径可用带宽。
  - 缺点：测量过程本身会影响网络状态和已有流量特征，导致网络不稳定和服务质量下降



瓶颈带宽测量

- 用$(C_0,C_1,C_2,\cdots,C_H)$表示一条从源端到目的端的路径，H表示路径的跳数，Ci表示链路i的链路带宽，
- 路径瓶颈带宽可表示为：

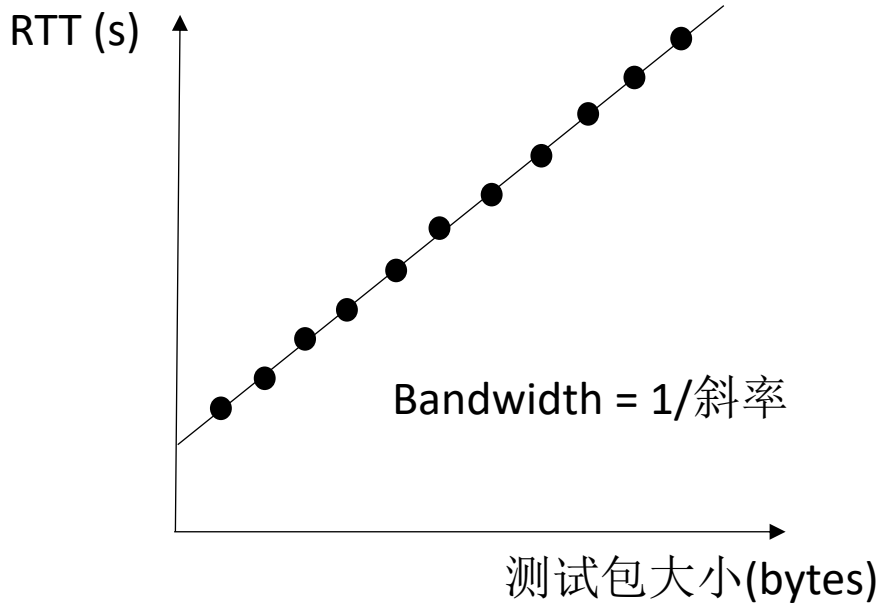
$$B_{bottleneck}=\min[C_i] \quad (i=0..H)$$

瓶颈链路带宽测量

- 变长包序列模型
  - 最初由Steve Bellovin和Van Jacobson提出
  - 代表工具：Pathchar，Clink，pchar，nettimer等
  - 假设条件：
    1. 传输延迟和包大小成线性关系
    2. 背景流量只会增加包延迟
    3. 特定大小的多个测量包中，往返延迟RTT最小的包可近似认为在通路中没有经历排队延迟
    4. 包转发时间可以忽略不计

瓶颈带宽测量

- 变长包序列模型
  - 测量原理：
    - 源端向目的端发送一列相同大小的n个测量包，目的端接收到测量包后发回ack确认包。
    - 源端计算测量包的RTT值，并选取最小值。
    - 这一过程重复k次，每次取不同大小的测量包。
    - 计算k个包大小的差值与相应RTT的差值，取其差获得链路物理带宽。背景流量的影响通过取差值而消除。
    - 实际中使用作图法得到结果，分别以包延迟和包大小为纵横坐标描点，用线性回归法作出一条直线，直线斜率的倒数即为链路物理带宽。通过逐跳链路物理带宽，可以间接获得通路瓶颈带宽。



瓶颈带宽测量

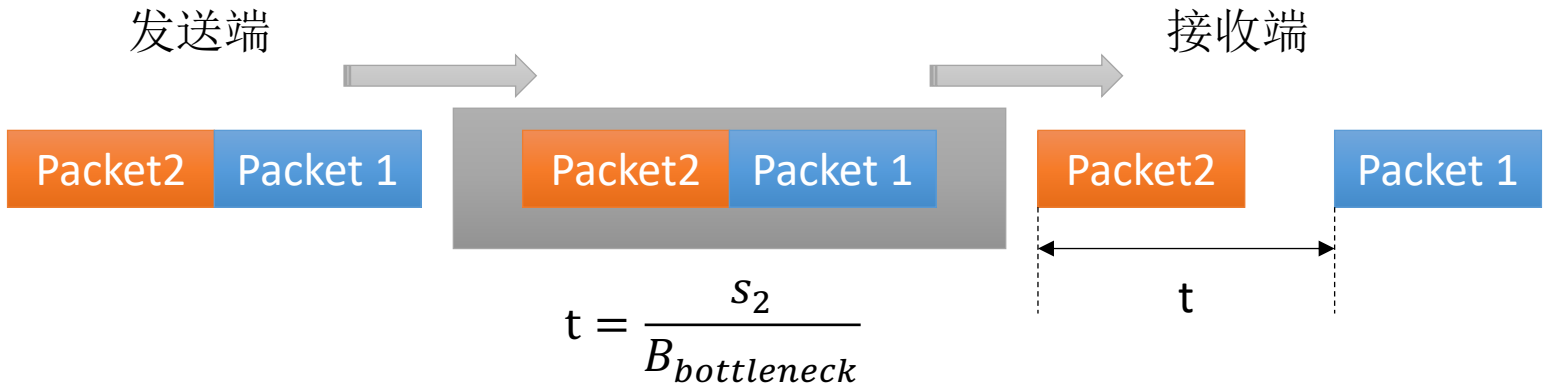
- 变长包序列模型
  - 模型缺陷
    - 测量所需时间较长，消耗的带宽较大
    - 测量工具多用ICMP应答请求包或UDP包执行测量，需要各个目的节点及时反馈，但很多路由器和主机赋予这两种包较低的处理级别甚至无响应，导致测量结果不准确
    - 没有IP地址的二层存储转发设备消耗的处理时间影响测量结果
    - 双向延迟引起不对称的正反向链路对于测量结果的影响
    - 通路最大MTU和端系统的时间粒度决定了测量上限，目前软硬件技术的限制使这类工具难以用于高速网络测量

瓶颈带宽测量

- 基于包对模型的方法
  - 代表工具：bprobe , pathrate ,TOPP
  - 测量原理：
    - 若两个数据包在瓶颈链路处相邻，则两个数据包到达接收端的时间间隔为

$$t = \frac{s_2}{B_{bottleneck}}$$

    - s_2为第二个数据包的大小，$B_{bottleneck}$为瓶颈链路的带宽



$$t = \frac{s_2}{B_{bottleneck}}$$

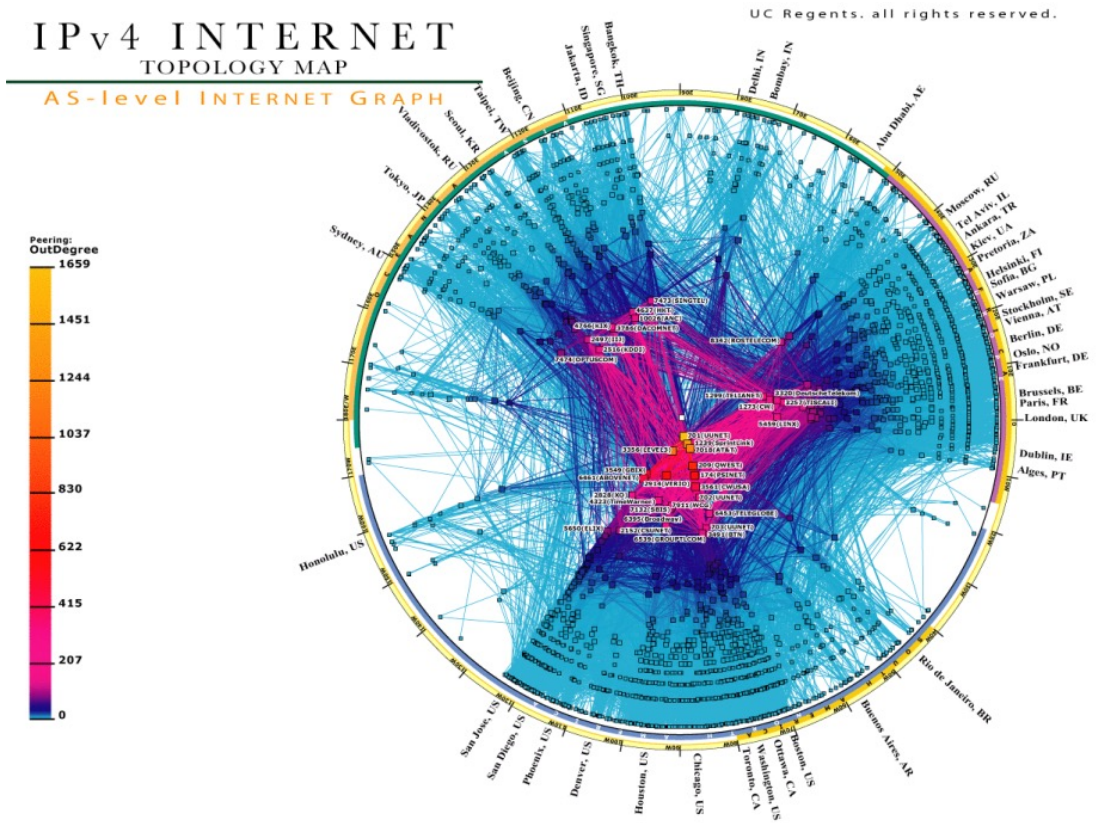
瓶颈链路带宽计算公式：$B_{bottleneck} = \frac{s_2}{t}$

瓶颈带宽测量

- 基于包对模型的方法
  - 假设条件：
    - 两包发送间隔足够小，以使其在瓶颈链路处紧邻
    - 瓶颈处路由器排队机制为先进先出
    - 传输延迟和包大小成线性关系
    - 路由器是存储转发的
  - 影响测量精度的原因：
    - 背景流量影响两包之间的间隔
    - 包丢失:直接丢失或路由器溢出
    - 两包在瓶颈链路后续链路再次排队引起包间隔的变化
    - 系统时钟粒度的限制

拓扑测量

- 拓扑分级
  - IP级
  - 路由器级
  - PoP级
  - AS级



TraceRoute

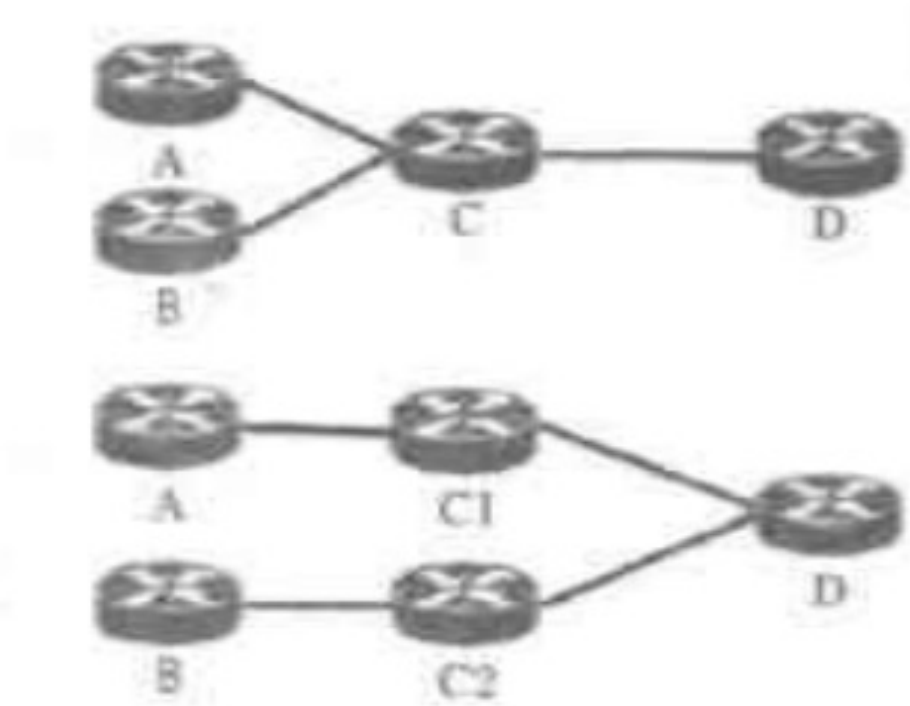
- 原理：利用IP路由器对TTL的处理
- 特点
  - IP级拓扑
  - 探测节点发送UDP消息
  - 中间节点回复ICMP消息
- 优点
  - 不受网络管辖范围限制
- 缺点
  - 效率较低，发送一系列探测包
  - 部分中间路由器不回复ICMP

```
C:\Documents and Settings\IBM>tracert www.baidu.com

Tracing route to www.a.shifen.com [202.108.22.5]
over a maximum of 30 hops:

  1    <1 ms    <1 ms    <1 ms  172.30.50.1
  2     1 ms     1 ms     1 ms  172.25.1.1
  3    <1 ms    <1 ms    <1 ms  172.25.0.2
  4    <1 ms    <1 ms    <1 ms  221.2.164.1
  5     1 ms    <1 ms    <1 ms  221.2.131.25
  6     8 ms     8 ms     8 ms  60.215.131.209
  7    24 ms    23 ms    16 ms  60.215.136.73
  8    19 ms    18 ms    18 ms  219.158.12.97
  9    17 ms    17 ms    17 ms  202.96.12.22
 10    23 ms    23 ms    23 ms  61.148.157.238
 11    17 ms    17 ms    17 ms  61.148.155.230
 12    18 ms    17 ms    17 ms  202.106.43.30
 13    23 ms    23 ms    23 ms  xd-22-5-a8.bta.net.cn [202.108.22.5]

Trace complete.
```

别名解析(Alias Resolution)

- 路由器有多个接口，每个接口对应一个IP地址，别名解析即监测哪些IP地址属于一个路由器，把同一路由器的IP地址聚合起来。

- 不使用别名解析，会导致
  - 虚拟链路
  - 虚拟节点
  - 虚拟拓扑



别名解析

- 使用DNS域名反向查询
  - 对同一路由器的多个可能的IP地址做反向域名查询，假定同一个路由器的多个接口地址都具有相同的域名
- 使用IP数据包Identification字段
  - Identification字段唯一
  - 同一个路由器发出的IP数据包Identification字段是连续的（70%路由器满足这个特征）
- 基于ICMP消息
  - TTL超时，ICMP消息源地址为路由器上探测报文的入口地址
  - 端口不可达，ICMP消息源地址为路由器上探测报文的出口地址

生成AS级网络拓扑

- IP映射到AS需要两个步骤
  - 将一个IP地址映射到一个最佳的IP地址前缀（最长匹配）
  - 将IP地址前缀映射到它的源AS(Origin AS)

表 2 CN0504 和 SKITTER0403 的主要拓扑特征参数

拓扑特征参数	CN0504	SKITTER0403
节点数	84	71
边数	211	160
最大节点度数	38	24
聚集系数	0.187	0.225
Mixing 系数	–0.328	–0.254
平均最短距离	2.54	2.66
10%的高度节点 Rich-club 系数	0.679	0.714
度分布幂律指数	–2.21	–2.20

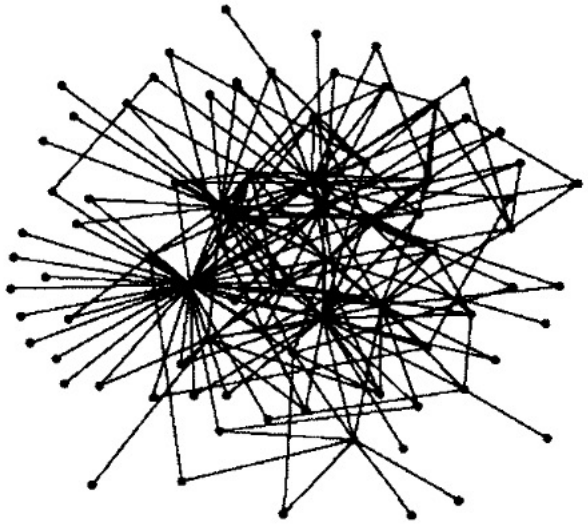


图 2　中国大陆 AS 的拓扑图（CN0504）

中国大陆AS级拓扑现状

- 节点度分布是指一个节点的度数为k的概率 p(k)
- Betweenness是刻画网络中节点重要性的一个重要特征

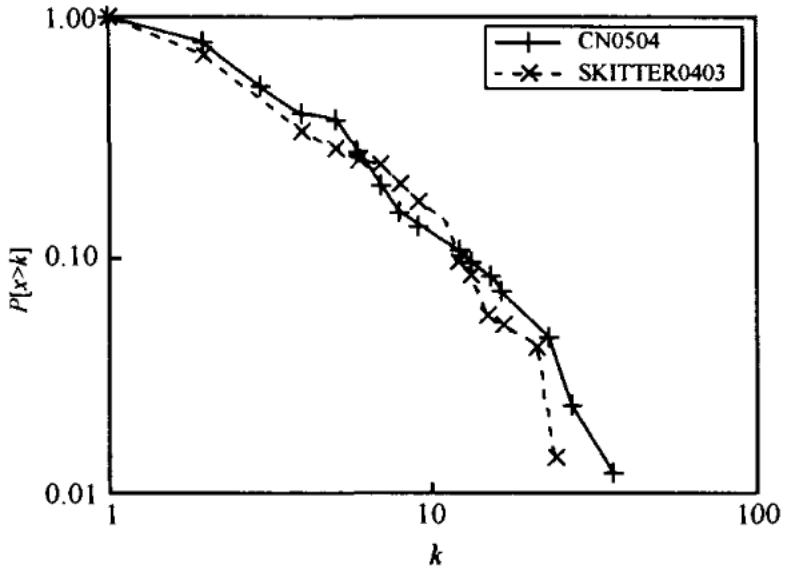


图 3　中国大陆 AS 拓扑节点度的累积分布

服从幂律分布，少数超级节点

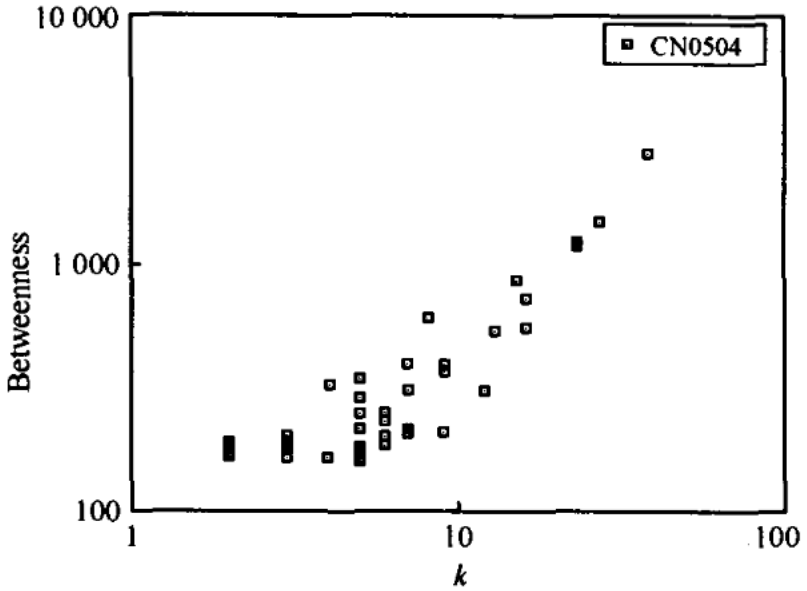
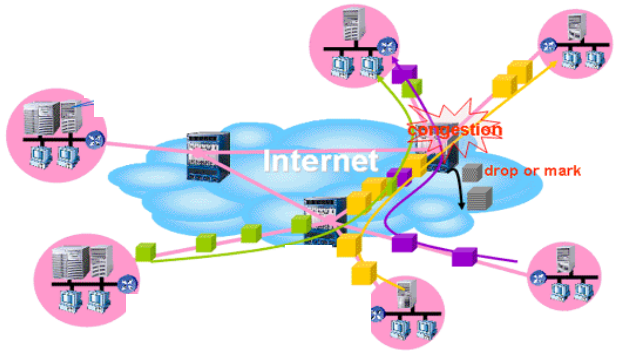


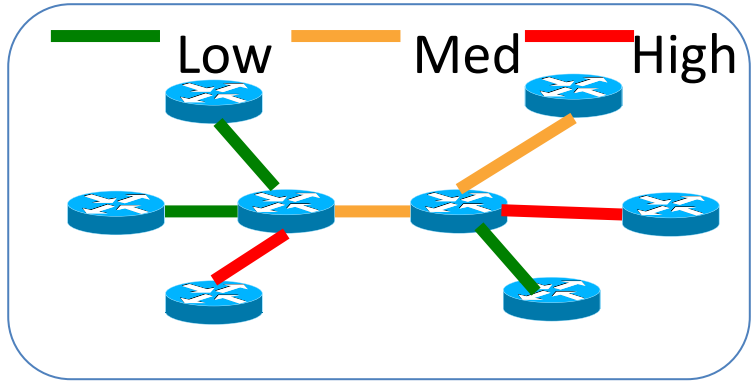
图 5　节点度和 Betweenness 的关联图

节点度高的节点在信息流的控制上也具有更为重要的作用

利用视频流量测量绘制网络流量地图

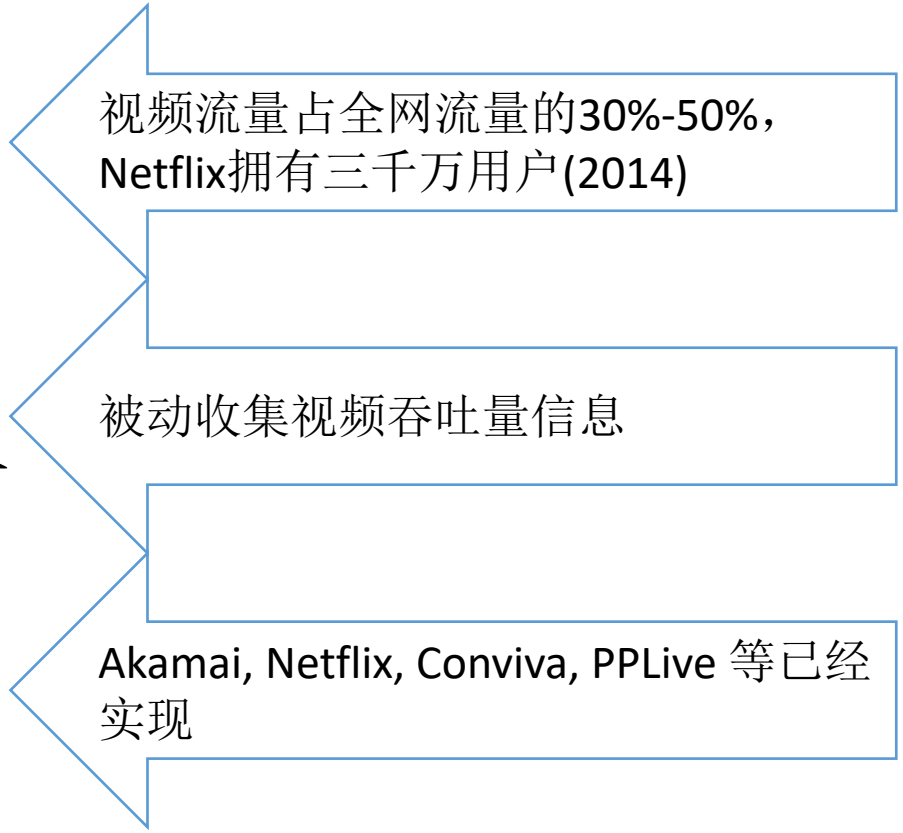
- Using Video-Based Measurements to Generate a Real-Time Network Traffic Map HotNets 2014
- Real world has Google Maps, why don’t we!



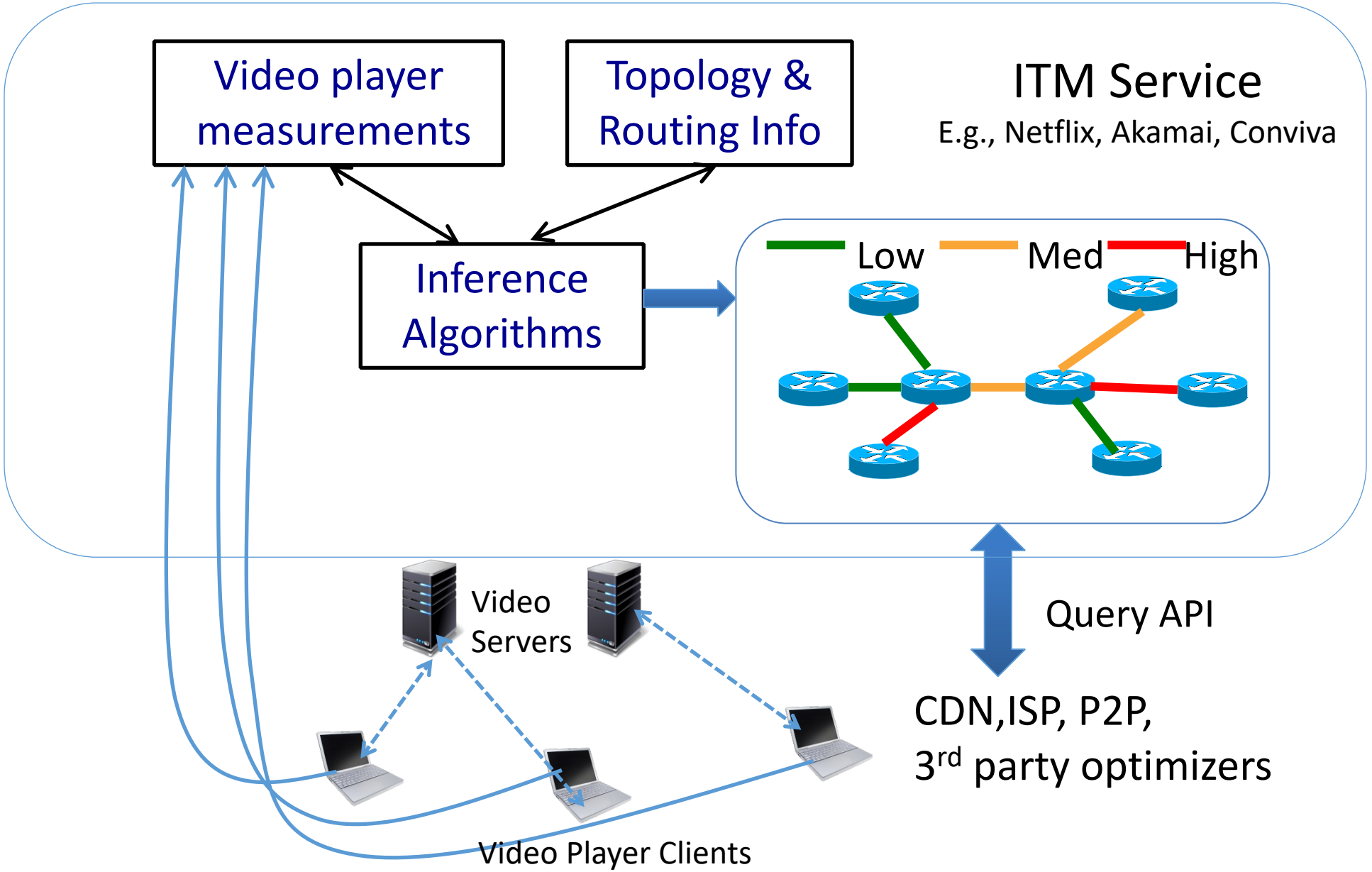


利用视频流量预测链路带宽和可用带宽

- 实现难点：
  - 覆盖率
    - 需要百万级别的监测点
  - 开销
    - 带宽测量引入巨大的计算成本
  - 实时性
    - 实时更新



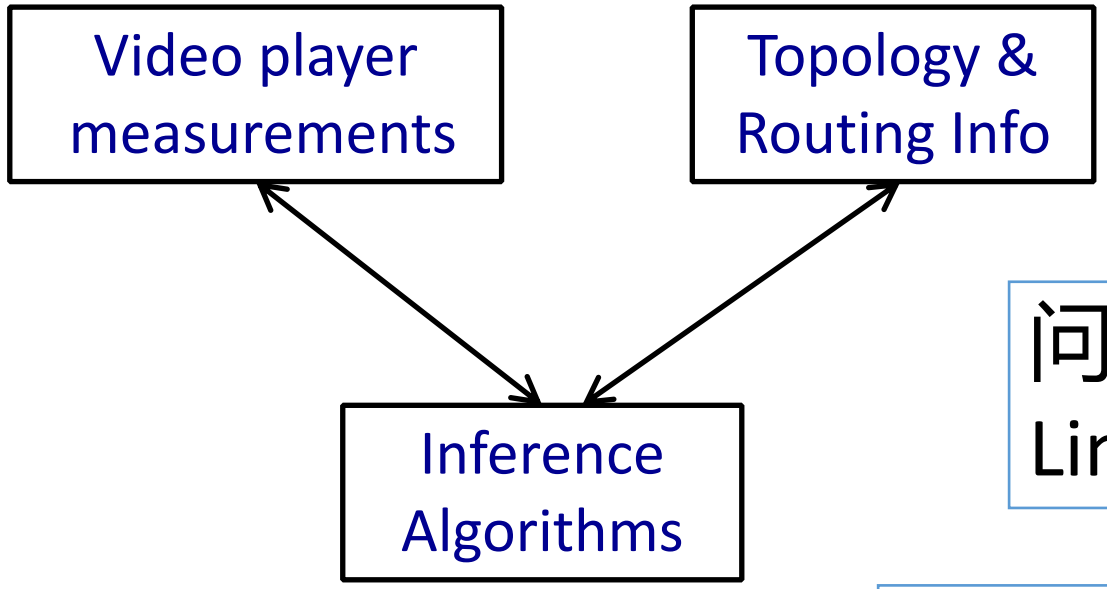
Internet Traffic Map Service



问题定义

For each epoch, src, dst:
Throughput
Bytes

For each epoch, src, dst:
Path (PoP-level)
e.g., iPlane



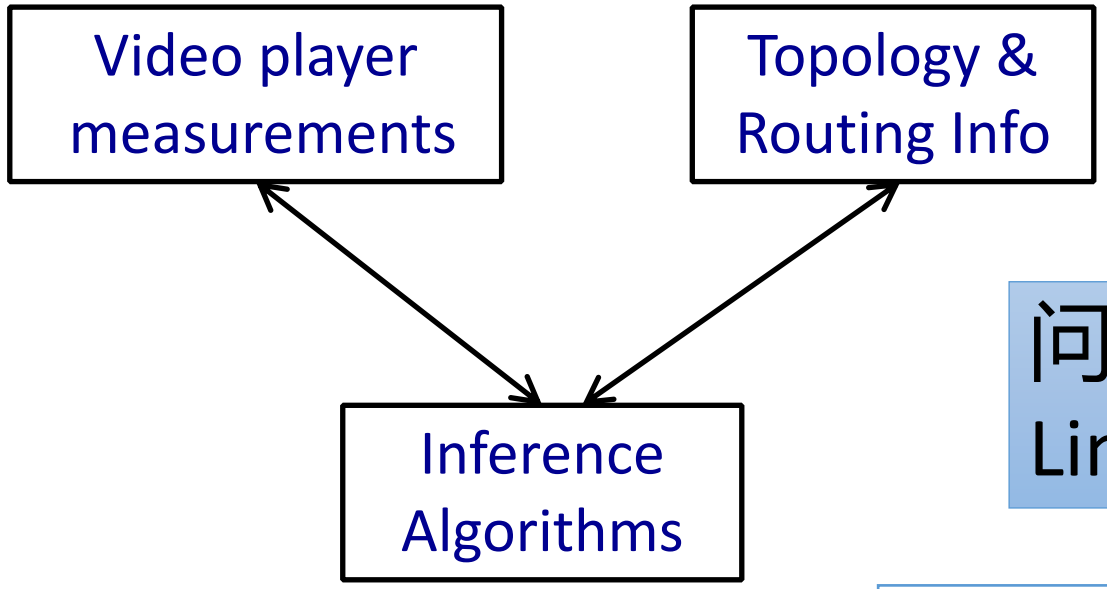
问题1:带宽估计 InferCapacity
Link → Capacity

问题2: 使用率估计InferUtilization
Link,Epoch → Utilization

问题定义

For each epoch, src, dst:
Throughput
Bytes

For each epoch, src, dst:
Path (PoP-level)
e.g., iPlane

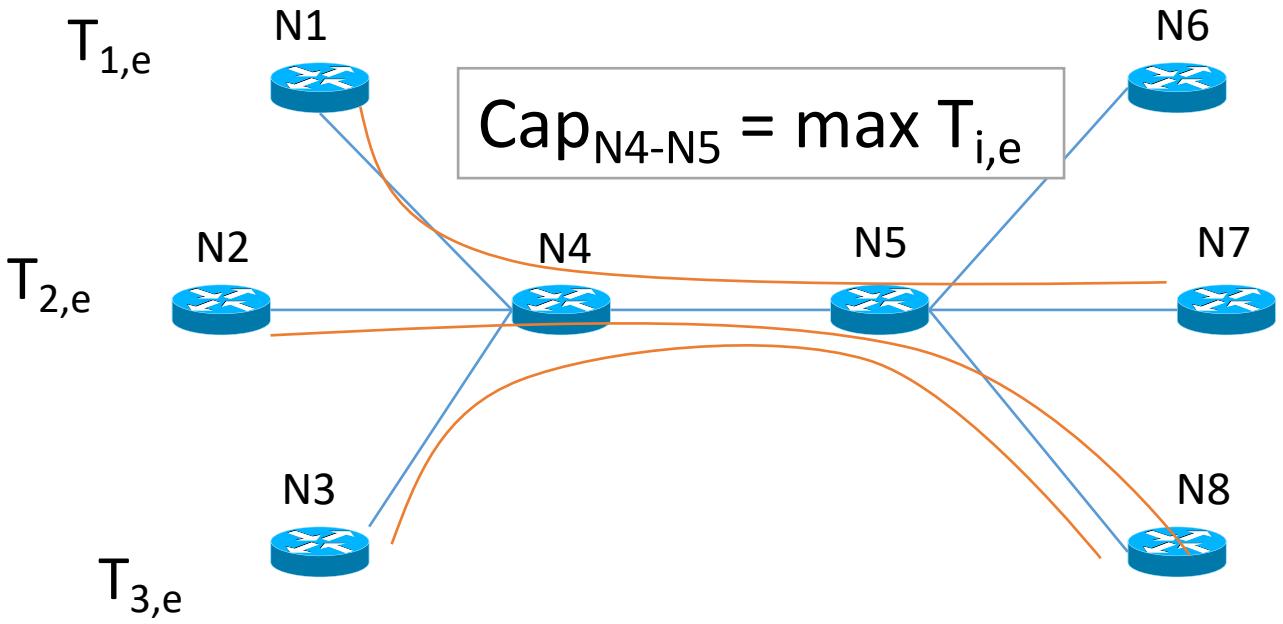


问题1:带宽估计 InferCapacity
Link → Capacity

问题2: 使用率估计InferUtilization
Link,Epoch → Utilization

Strawman 1: Max Estimator

$T_{i,e}$ = throughput measurement in epoch e



低估？

背景流量？

离散值？

使用收集到最大速度？

Strawman 2: Tomography

低估？ 背景流量？ 离散值？

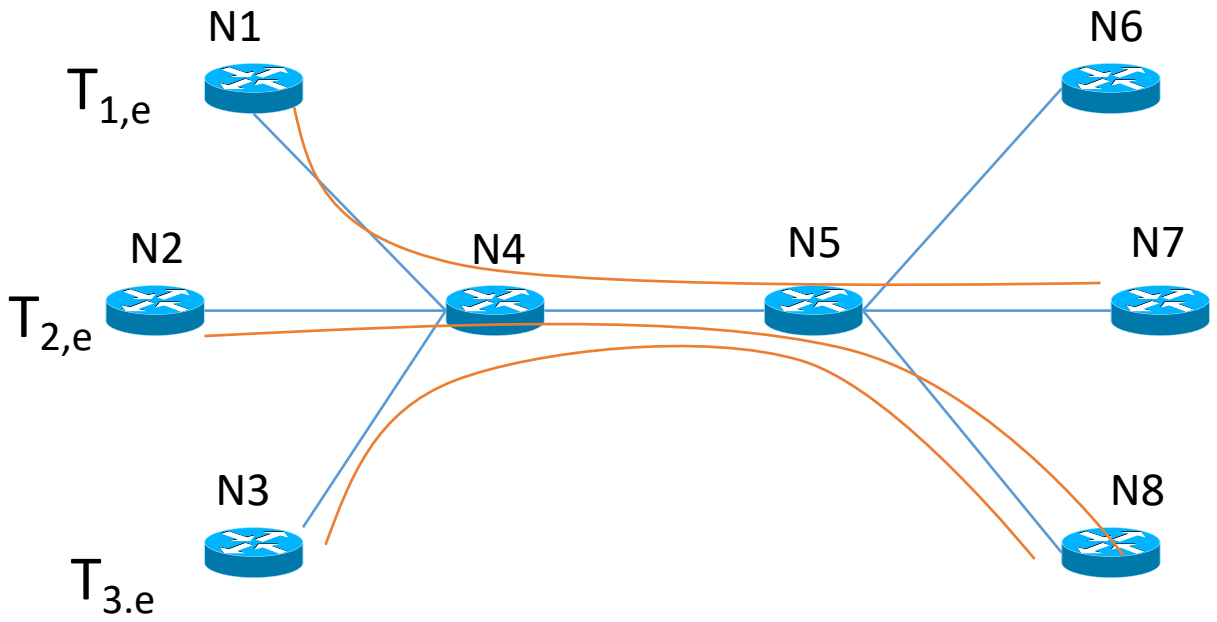
1. Cap_{Ni-Nj}使用离散值
2. 引入背景流量值 $B_{N\{1,2,3\}-N\{6,7,8\}}$
3. 考虑带宽建设成本

约束最优化

Output = Cap_{Ni-Nj}

Idea: Max_e {B+T,Bytes/Time}

约束条件太少!



Idea: Add "Side" information

$$Minimize: \sum_{l} Cost_l \quad (1)$$

$$\forall l, e: \sum_{s,d,l \in \mathcal{P}_{s,d}} \frac{B_{m,s,d,e}}{|e|} + \sum_{s,d,l \in \mathcal{P}_{s,d}} bg_{s,d,e} \leq C_l \quad (2)$$

$$\forall l: C_l \geq T_{m,s,d,e}, if\, l \in \mathcal{P}_{s,d} \quad (3)$$

$$\forall l: C_l = \sum_{c \in CapVals} (d_{l,c} \times Cap_c) \quad (4)$$

$$\forall l: Cost_l = \sum_{c \in CapVals} (d_{l,c} \times Cost_c) \quad (5)$$

$$\forall l: \sum_{c \in CapVals} d_{l,c} = 1 \quad (6)$$

$$\forall l, c: d_{l,c} \in \{0, 1\} \quad (7)$$

1. "Gravity" assumption
2. B/T ratio
3. Expected overprovisioning

1. $Cap_{Ni\text{-}Nj}$使用离散值
2. 引入背景流量值 $B_{N\{1,2,3\}\text{-}N\{6,7,8\}}$
3. 考虑带宽建设成本

<u>约束最优化</u>

Output = $Cap_{Ni\text{-}Nj}$

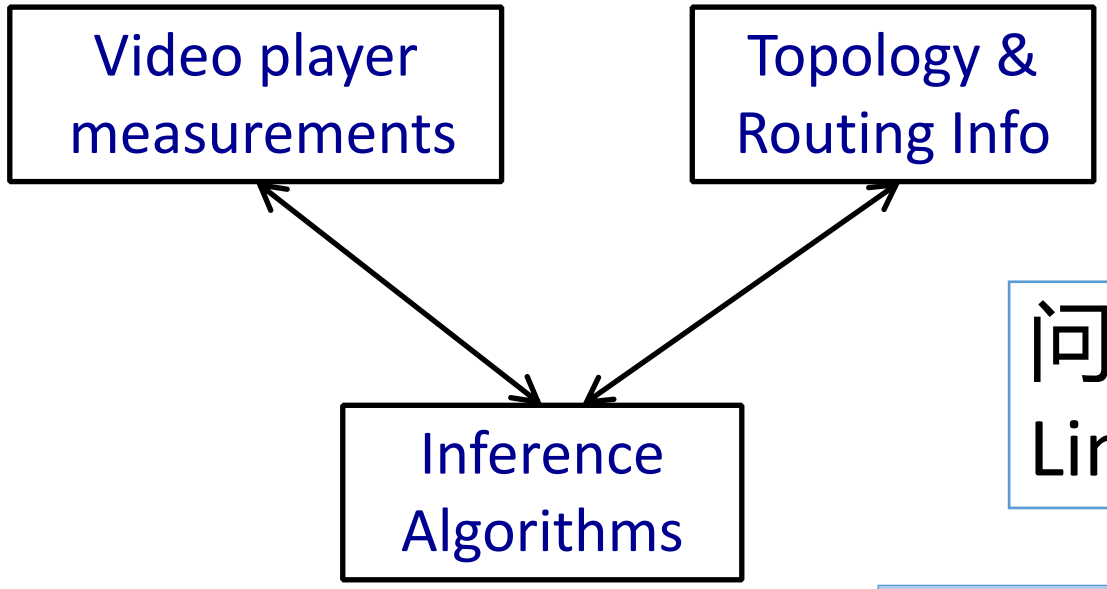
Idea: Max_e {B+T,Bytes/Time}

Underconstrained!

问题定义

For each epoch, src, dst:
Throughput
Bytes

For each epoch, src, dst:
Path (PoP-level)
e.g., iPlane



问题1:带宽估计 InferCapacity
Link → Capacity

问题2: 使用率估计InferUtilization
Link,Epoch → Utilization

Strawman: Tomography+?

1. “Gravity” assumption
2. B/T ratio
3. Expected overprovisioning

Constrained Optimization

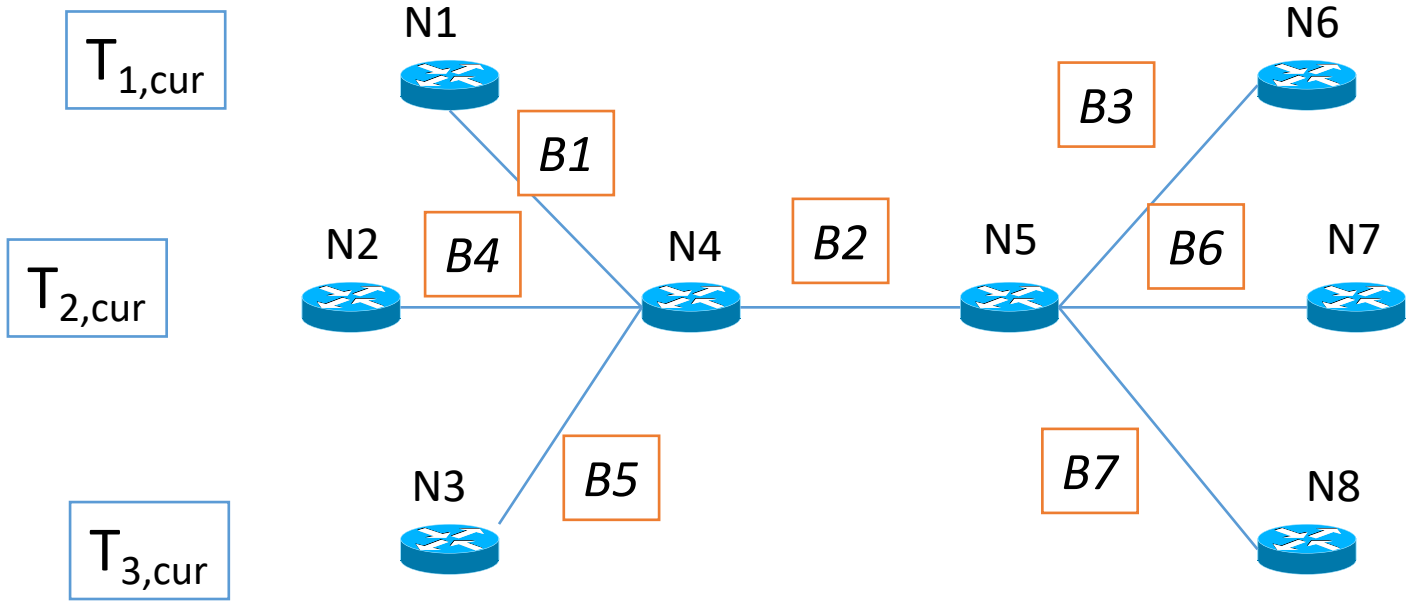
Output = B_{Ni-Nj}

“聚合”效果

Don’t have history

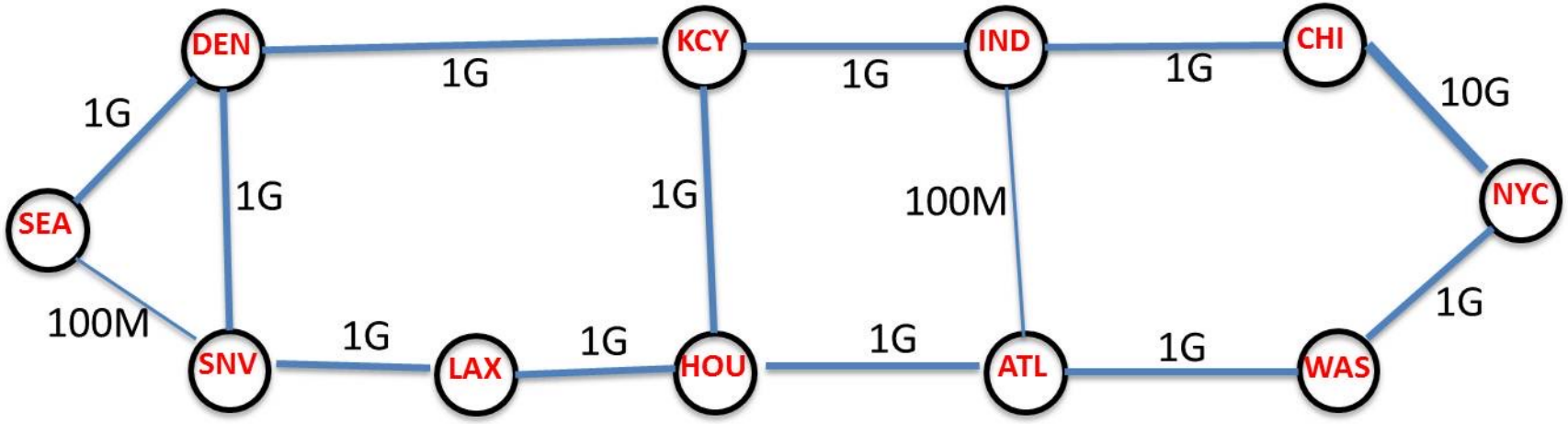
High-level Idea:
Capacity + Max-min fairness

得到问题1估计的带宽值后，
B取什么值最能“解释”收集到的throughput数据？



利用穷举的方法进行极大似然估计
s.t. 预测的throughputs“匹配”收集的数据

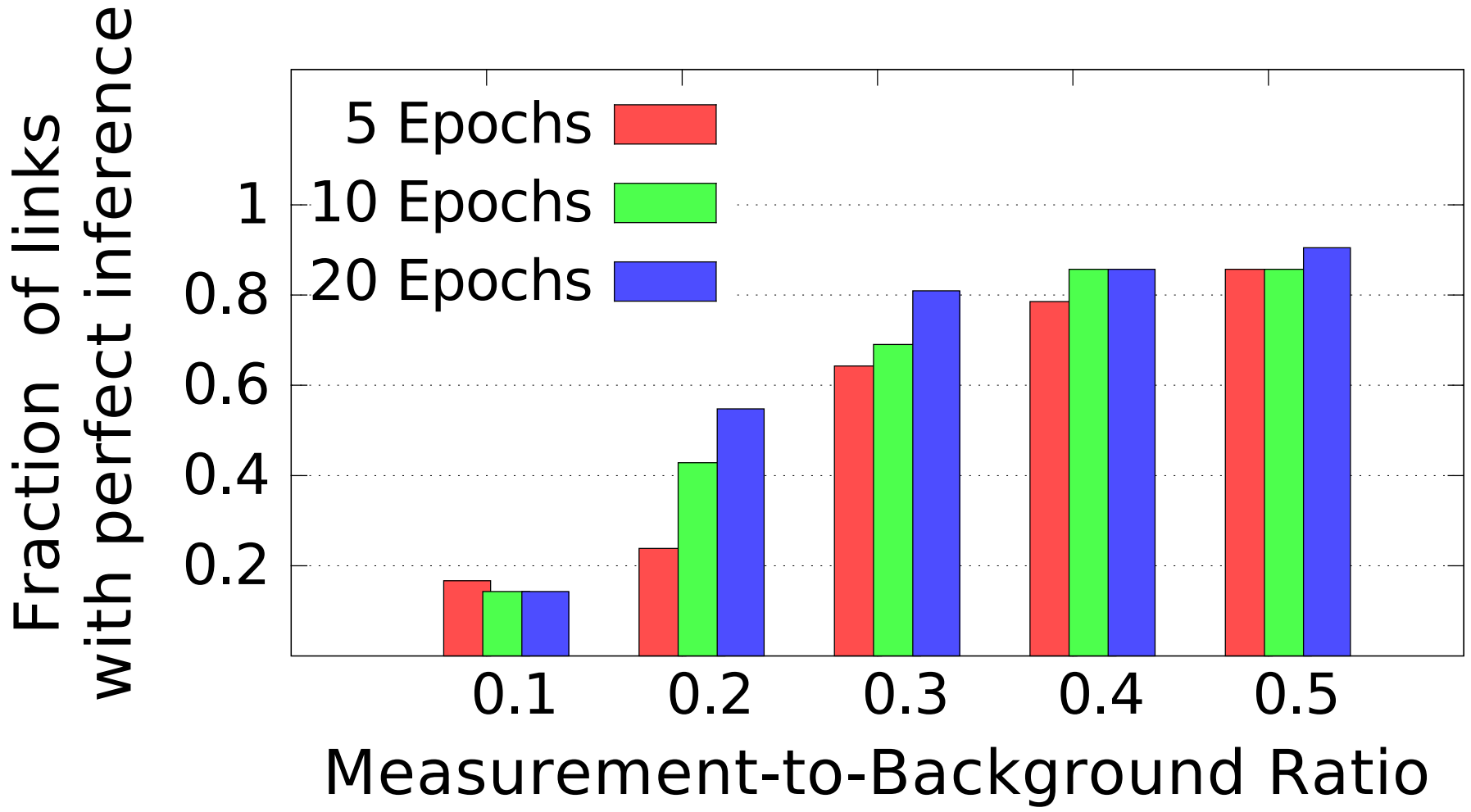
Evaluation Setup



- Custom flow-level simulator
- Sensitivity
  - number of epochs
  - Background vs. measurement traffic ratio
  - Accuracy of capacity inference

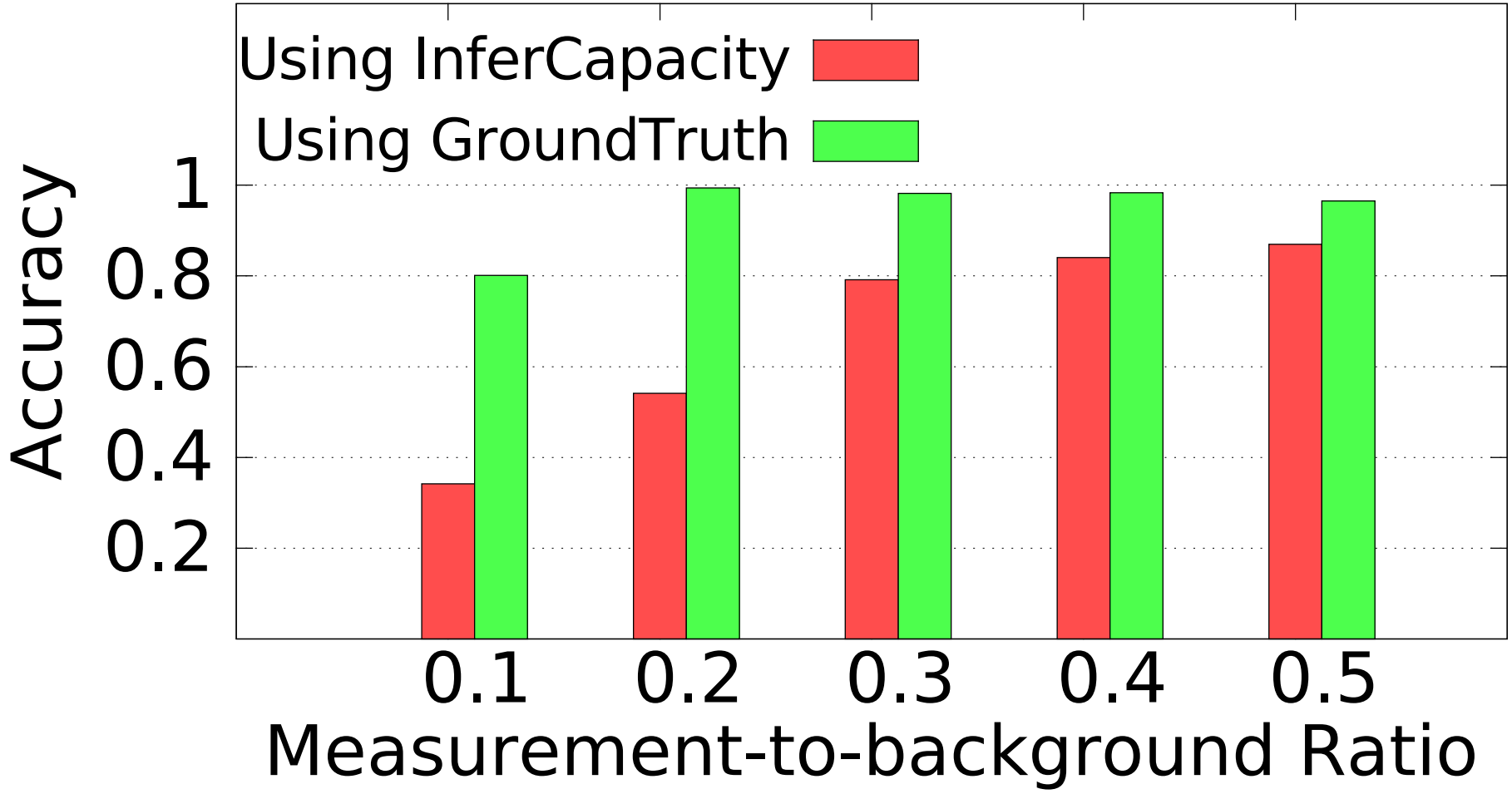
49

Accuracy of Capacity Inference



50

Accuracy of Background Inference



51