INTERNET MEASUREMENT: INFRASTRUCTURE

Gaia Maselli
maselli@di.uniroma1.it

SAPIENZA
UNIVERSITÀ DI ROMA

Properties of the Internet's Infrastructure

- *Physical* device properties (physical components that make up the Internet)
- *Topology* properties (how the components are interconnected)

Physical properties of the Internet

- The basic building blocks of the Internet are
  - end systems
  - Links
  - routers
- Links and routers are interesting for Internet measurement

Physical properties of the Internet: links

Links

- A single pint-to-point communication medium
- A sequence of connections that are switched below the IP layer (multiple Ethernet segments)
- A broadcast medium (WiFi)

Link properties

- Propagation delay (time required to traverse the link)
- Capacity (the maximum data rate that can be achieved by the link)

Performance properties associated with a link

- Packet delay
- Packet loss
- Packet jitter: variability of packet inter arrival times

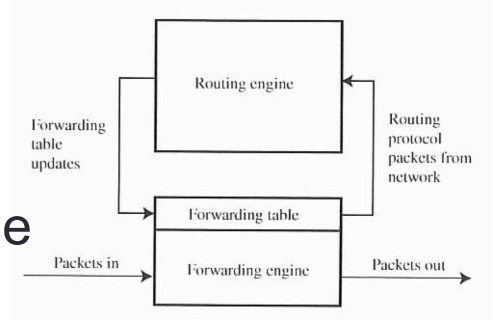
Physical properties of the Internet: routers

Routers

- Move packets from a link to another
- The rate at which an outgoing interface move packets away is the outgoing link's capacity
- Since packets may arrive faster than the outgoing interface can take them away, internal buffers are required to hold packets awaiting transmission on each outgoing interface
- Buffers are FIFO queue
- Packets arriving at a full buffer are discarded (*drop-tail* service)
- Routers can be configured to perform *active queue management*

Routers properties

- Set of IP addresses used on router's interfaces
- Geographic location of the router
- Time a router requires to respond to an ICMP message
- Time a router requires to forward a packet



Physical properties of the Internet: wireless

Wireless connectivity

- Radio frequency media
- Primary goal: to link users to the wired infrastructure

Measurements involving wireless communication

- Signal strength
- Amount of power consumed
- Data bit rates
- Degree of coverage
- Error rates
- Link capacity
- Available and effective bandwidth
- Identifying bottleneck links

Topology properties

Interconnection of physical components can be visualized at four levels

1. Autonomous systems
2. Point of presence (one or more routers in a single location)
3. Router
4. Interface

Topology views

- AS graph
- PoP-level graph
- Router graph
- Interface graph

Interaction of traffic and network

- Physical limits imposed by the infrastructure
  - Minimum possible delay
  - Maximum possible throughput
- Network conditions influence traffic properties
  - Packet delay
  - Packet loss
  - Throughput (total, per connection, goodput)
  - Packet jitter (variability of packet inter-arrival time)

Infrastructure properties that are interesting to measure

Challenges in measurement

Poor observability (observability is not built into the design of Internet protocol and components)

- *Core simplicity*: core elements of the network are deliberately very simple and so do not support detailed measurement
  - Routers are stateless (do not keep track) with respect to the connections or flows passing through them
  - No counters are maintained
- *Hidden layers*: the layered IP model tends to impede visibility of the lower layers
  - Details on packet transmission (layer 2) are hidden at the IP level

Challenges in measurement (cont)

- *Hidden pieces*: measurement of some network components can be hampered by specialized network devices
  - Middleboxes (devices that deviate from the end-to-end architecture principle) impedes visibility of network components
  - Ex. Firewall may block UDP and ICMP packets being used by `traceroute`
  - NAT can prevent discovery of end systems via `ping`
- *Administrative barriers*: operators avoid providing information about their networks for competitive reasons
  - ISP frequently seek to hide internal details (interconnection patterns, amount of traffic carried over network links) of their networks
  - ISP block traffic that may be used to measure infrastructure (`ping`, SNMP)

How to measure infrastructure properties

Tools

- *Active measurement*: adding traffic to the network for the purpose of measurement
- *Passive measurement*: capturing traffic that is generated by other users and applications
- *Fused measurement*: combination of active and passive
- *Bandwidth measurement*
- *Latency measurement and estimation*
- *Geolocation*

Some active measurements

- Ping -> Connectivity, Instantaneous RTT
- Owamp -> one way packet delay
- Traceroute -> network paths, topology
- Multicast-based methods -> packet loss

Active measurement: `ping`

- Active methods involve adding traffic to the network for the purpose of measurement

`Ping`

- Metrics
  - Connectivity
  - Instantaneous RTT between the sender and the target
- Method
  - Sends ICMP ECHO packet to a target and captures the ECHO REPLY
- Characteristics:
  - Only the sender needs to be under experimental control
  - Difficult to determine the direction in which congestion is experienced

Active measurement: owamp

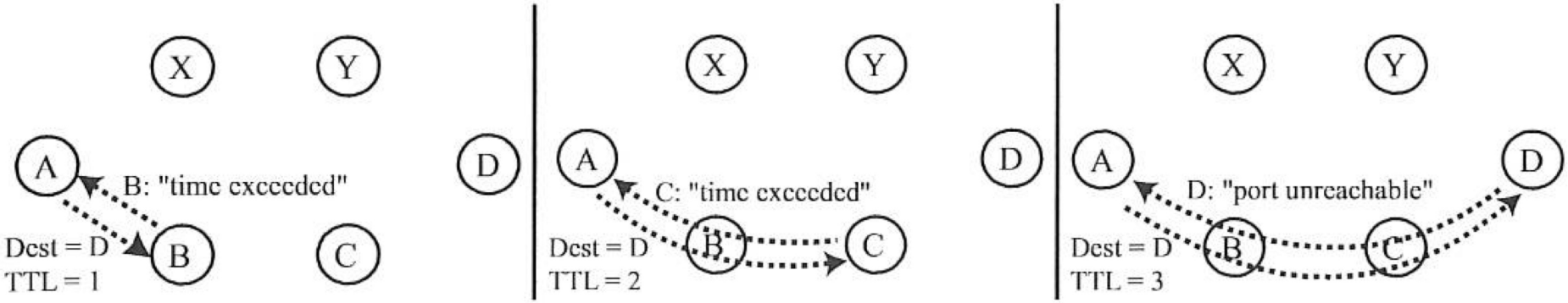
Owamp (one-way ping)

- Metrics
  - One way delay
- Method
  - Sends a probe packet to a demon process running on the target
- Characteristics:
  - Sender and receiver need to be under experimental control
  - Requires a demon process to run on the target, which listens for and records probe packets sent by the sender
  - Requires synchronized time

Active measurement: `traceroute`

`Traceroute`

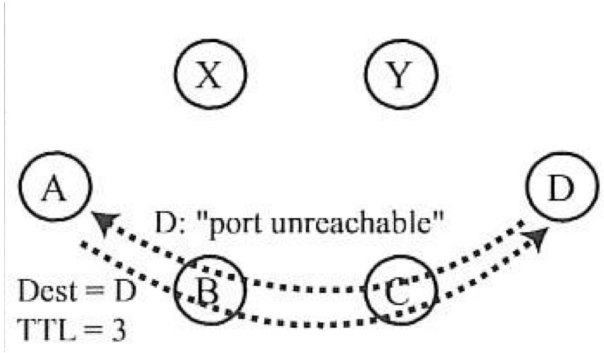
- Metrics
  - Network paths
- Method
  - Sends packets with increasing TTL (starting at 1) to an unlikely port on a destination



`Traceroute`: path asymmentry

- The nodes visited by `traceroute` are those in *forward* path from the source to the destination
- *Reverse* path may be different

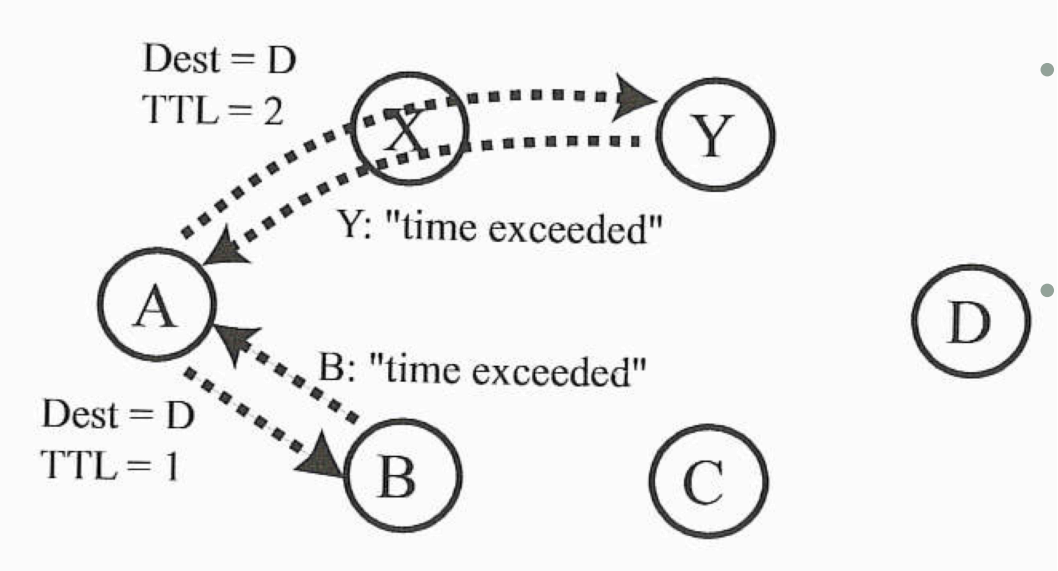
➔The output of `traceroute` must be interpreted only in terms of directed path from source to destination



- If D were sending packets to A, those packets are not guaranteed to follow the path D->C->B->A
- They may pass through X or Y

Traceroute: unstable paths and false edges

- It only reports the nodes visited by successive probe packets with increasing TTL
- This sequence represents a valid path if the path is stable
- If IP paths are not stable over the measurements period, then successive probes may follow different paths

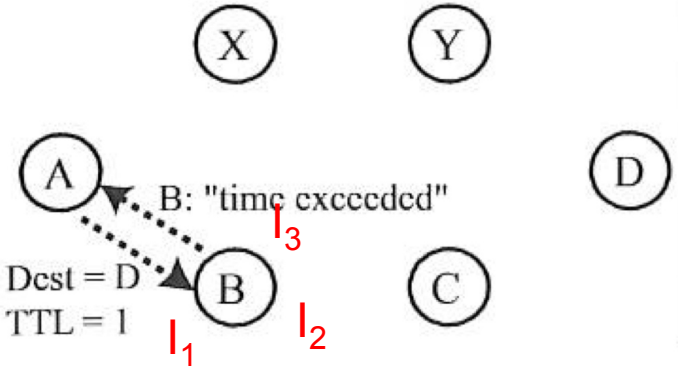


- Node A may alternate between using B and X as the next hop
- Inferred path segment contains *false edge*

A->B->Y

Traceroute: alias resolution

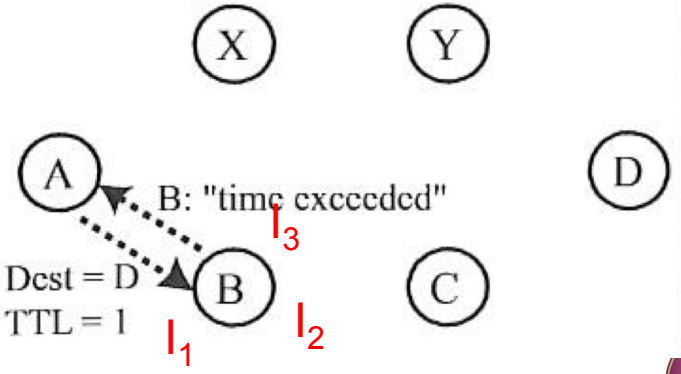
- Traceroute discovers *interfaces* rather than *routers*
- Routers along the path will generally have multiple network interfaces
- Each network interface has a different IP address
- The source IP address of the ICMP TIME EXCEDEED response packet is the address of the interface that the router uses when sending packets to the source



- The IP address in the source field of the TIME EXCEDEED response will be I_1 (address that A is able to discover)

`Traceroute`: alias resolution

- N.B. It not possible to form a router-level topology map from a collection of `traceroute` measurements
- If X were to use `traceroute` to discover the path to D, and if the path passed through B, the interface discovered by X would be I_3.
- Given the two sets of path measurements (A to D and X to D) it would be not clear that both paths passed through the same node B



Methods for alias resolution are needed!!!

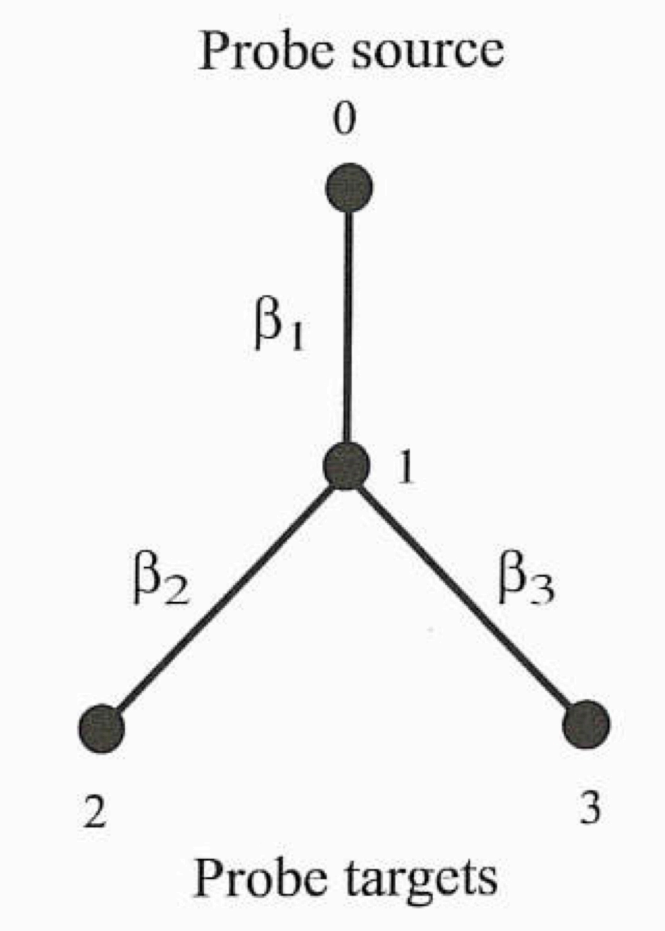
Traceroute: alias resolution

- One of the methods requires to send ICMP ECHO packets to both interfaces from the same source
- If both interfaces belong to the same router, the responses will both be sent from one interface
- By matching ECHO REPLY messages having the same source interface, it is possible to infer that the ECHO packets were sent to a common router

Multicast-based method

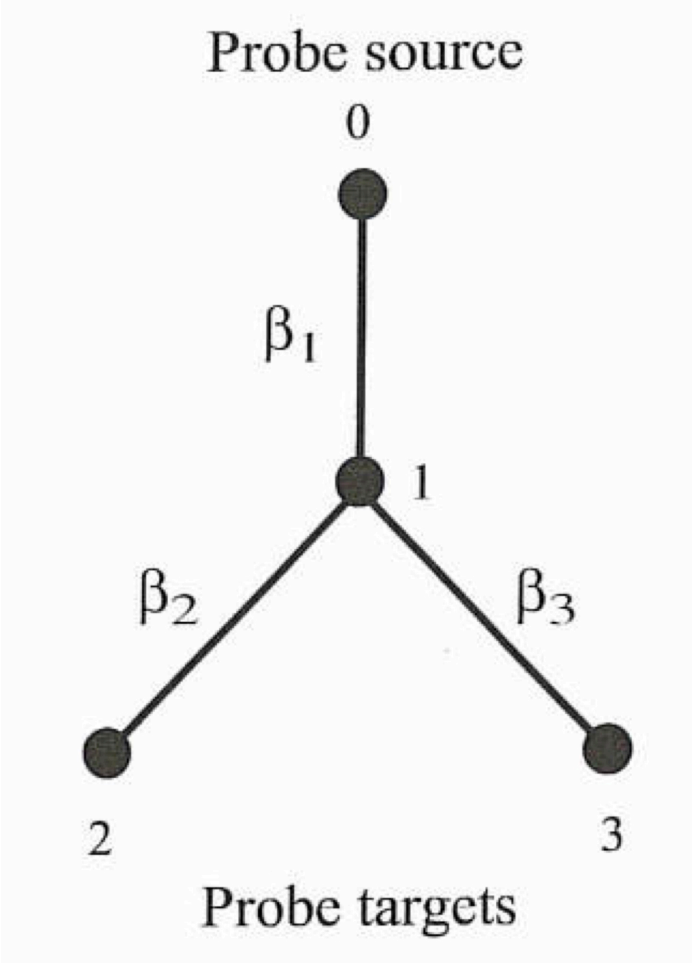
- Multicast: probes sent via multicast have the property that a single probe is replicated by routers along the path, so that network conditions experienced by a single upstream packet are reflected in measurable properties of multiple downstream packets
- Inference technique
- MINC approach allows to estimate *network tomography* (the study of a network's internal characteristics using information derived from end point data)

Multicast-based method (cont)



- Three links, three end systems, and one internal node
- β_i: loss rate associated with link *i*
- Goal: to estimate all the three loss rates from loss measurements made only on the paths 0->2 and 0->3
- MINC works with loss events instead of loss rates

Multicast-based method (cont)



- The probe source (node 0) sends multicast packets toward the end systems (nodes 2 and 3)
- when the multicast packet reaches the branching point 1, a copy of the packet is sent down on each of the links 1->2 and 1->3
- Packets that are not sent at either node 2 or 3 are assumed to be lost on link 0->1
- Packets seen at a node, but not the other, are assumed to be lost on the link leading to the node where the packet is unseen

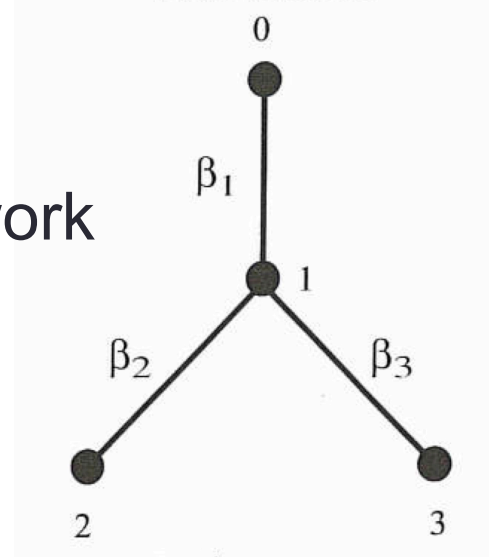
Multicast-based method (cont)

- Repeating the experiment many times it is possible to build up an estimate of the loss rates on each of the three links
- Losses on different links are assumed to be independent

Maximum likelihood estimator of loss rates

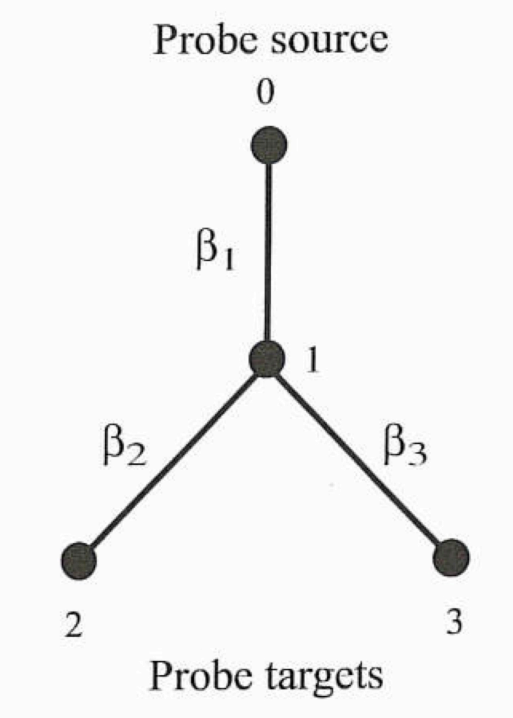
- {00, 01, 10, 11} : four possible events when a probe is sent
- 1: the probe is received at an endpoint
- 0: the probe is lost somewhere in the network
- Ex. Event 01: probe is lost on link 1->2

but successfully transmitted to node 3



Multicast-based method (cont)

- Let p(x) denote the proportion of trials in which the event x is observed
- *Per-link loss rates* can be estimated as



$$\hat{\beta}_1 = 1 - \frac{(p(01)+p(11))\cdot(p(10)+p(11))}{p(11)}$$

$$\hat{\beta}_2 = 1 - \frac{p(11)}{p(01)+p(11)}$$

Proportion of pkts successfully received at 2

$$\hat{\beta}_3 = 1 - \frac{p(11)}{p(10)+p(11)}$$

Passive measurement

- BGP -> Internet AS-level topology
- OSPF -> internal AS topology

Passive measurement: BGP

Goal: Internet AS-level topology

- BGP routing tables provide partial information about the AS-level topology (connections between ASes)
- The fact that two ASes appear in sequence in an AS path is evidence that they are directly connected
- Each AS advertises to is neighbors the routes it knows
- To understand how traffic flows into any particular AS, it is necessary to obtain BGP tables (views) from many other Ases
- `routeviews` repository: collects BGP views from a large set of ASes
- `Routeviews` was mainly intended to aid network operators, but it is used as data source for passive Internet topology monitoring and analysis

Passive measurement: OSPF

- Goal: internal AS topology
- Capturing control plane traffic generated by interior gateway protocol such as OSPF
  - Link state announcements (LSA)
  - Topology changes are indicated in LSA

Fused measurement

- In measuring infrastructure or discovering topology characteristics, it is often useful to fuse different kinds of measurement, including combining both active and passive measurements
- Passive measurement can be used to obtain a first view of the system and then use active measurement for specific and restricted goals

Bandwidth measurement

- Packet pair method
- Size delay method

Bandwidth measurement: motivation

- Measurement of bandwidth is important for applications that intend to adapt their behavior to the properties of the network
  - Streaming media applications (adjust transmission rate to the network bandwidth)
  - Server selection (find a server with an appropriate bandwidth connection to the client)
  - Estimating the bandwidth-delay product for use in TCP flow control
  - Overlay networks (to route data over good-performing path)
  - Verification of service level agreement between network customers and providers

Bandwidth measurement: techniques

- Generally bandwidth measurement is an ***active process*** in which packets are injected into the network and the measurement process is based on resulting observations
- Sometimes both endpoints of the measurement path are assumed to be instrumented
- In other settings only one endpoint is active and the other endpoint is simply expected to respond to an ICMP echo or similar trigger
- ***Passive methods*** have been proposed

Capacity

- *Capacity (single and end-to-end)*: maximum possible throughput (IP layer rate) that a link or path can sustain
- The minimum link capacity in the path determines the end-to-end capacity.
- The hop with the minimum capacity is the *narrow link* on the path

Available bandwidth

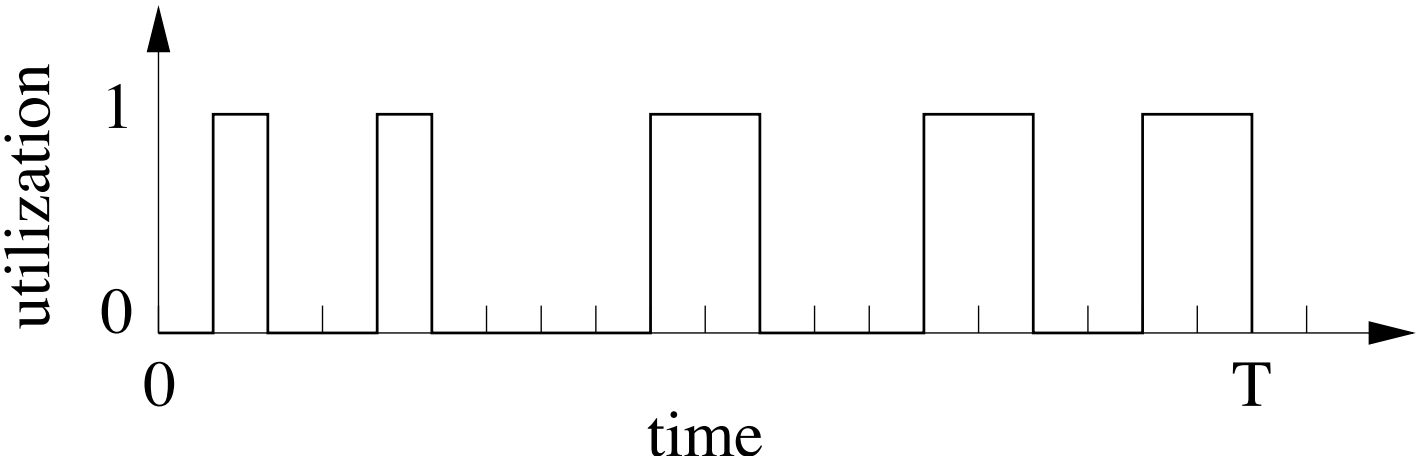
- *Available bandwidth* *(single and end-to-end)*: portion of capacity that is not being used during a given time interval (*residual* capacity)
- Depends on the traffic load and is a time-varying metric
- At any specific instant of time a link is either transmitting a packet at the full capacity (1) or it is idle (0)
- Available bandwidth requires time averaging of the instantaneous utilization over the time interval of interest
- The average utilization $\bar{u}(t-\tau,t)$ for a time period $(t-\tau,t)$

$$\bar{u}(t-\tau,t)=\frac{1}{\tau}\int_{t-\tau}^{t}u(x)d(x)$$

  - Where u(x) is the instantaneous available bandwidth on the link at time x

Available bandwidth

- Example: the link is used during 8 out of 20 time intervals between 0 and T, yielding an average utilization of 40%



Available bandwidth

- ***Single hop***: If C_i is the capacity of hop i and u_i is the average utilization at that hop in the given time interval, the average available bandwidth A_i of hop i is given by the unutilized fraction of capacity

$$A_i = (1 - u_i)C_i$$

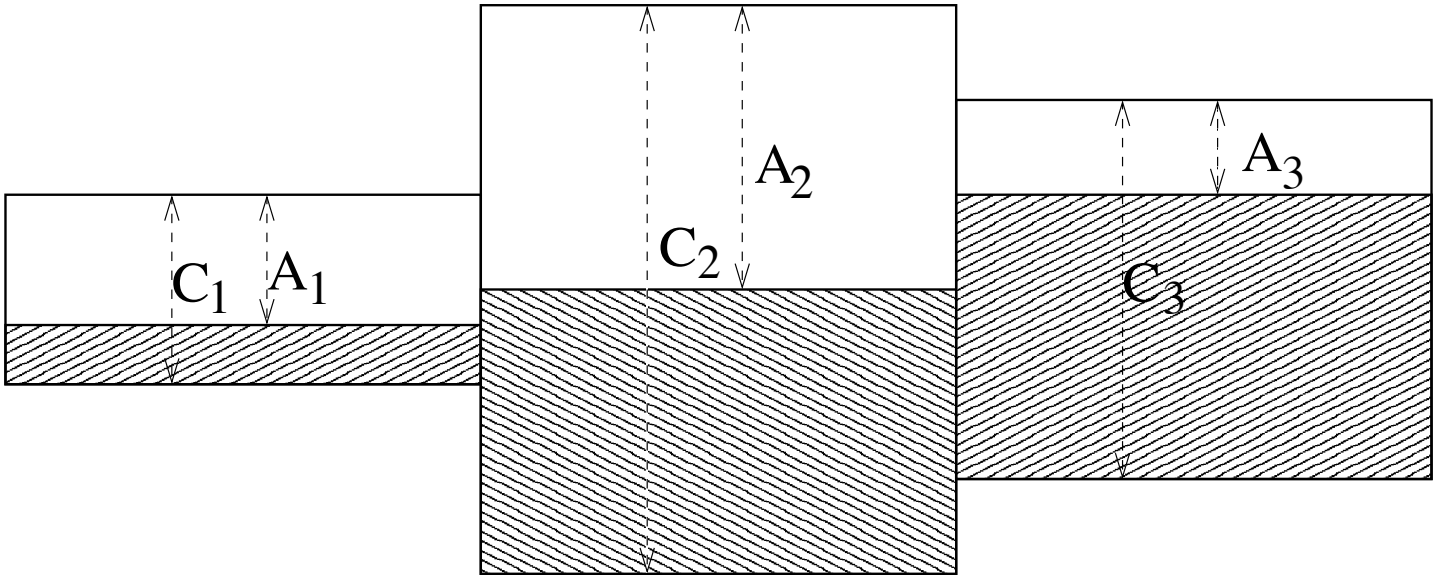
- ***H-hop path***: the available bandwidth of end-to-end path is the minimum available bandwidth of all H hops

$$A = \min_{i=1,\ldots,H} A_i$$

- The hop with the minimum available bandwidth is called the *tight link* of the end-to-end path

Capacity versus available bandwidth

- The minimum link capacity C_1 (*narrow link*) determines the end-to-end capacity
- The minimum available bandwidth A_3 (*tight link*) determines the end-to-end available bandwidth

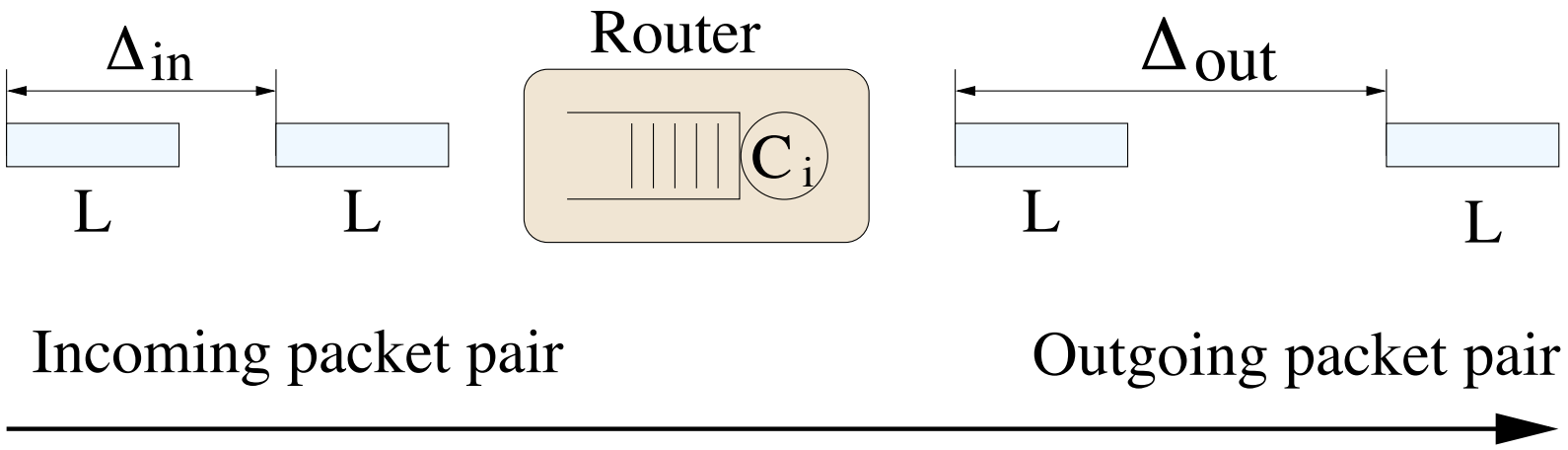


Bulk transfer capacity (BTC)

- Achievable throughput by a TCP connection
- TCP specific metrics
- BTC depends on how TCP share bandwidth with other TCP flows

Packet-pair method to measure end-to-end capacity

- The source sends multiple *packet pairs* to the receiver
- Each packet pair consists of two packets of the same size sent back-to-back.
- The **dispersion** of a packet pair at a specific link of the path is the time distance between the last bit of each packet.



Packet-pair method to measure end-to-end capacity

- If a link of capacity C_0 connects the source to the path and the probing packets are of size L, the dispersion of the packet pair at that first link is

$$\Delta_0 = L/C_0$$

- In general if the dispersion prior to a link of capacity C_i is Δ_{in}, assuming that the link does not carry other traffic, the dispersion after the link will be

$$\Delta_{out} = \max\left(\Delta_{in}, \frac{L}{C_i}\right)$$

Packet-pair method to measure end-to-end capacity

- After a packet pair goes through each link along an otherwise empty path, the dispersion Δ_R that the receiver will measure is

$$\Delta_R = \max_{i=0,\ldots,H} \left(\frac{L}{C_i} \right) = \frac{L}{\min_{i=0,\ldots,H}(C_i)} = \frac{L}{C}$$

- Where C is the end-to-end capacity of the path. Thus the receiver can estimate the path capacity from

$$C = \frac{L}{\Delta_R}$$

Observations on packet-pair method

- The assumption that the path is empty of any other traffic (referred to as *cross traffic*) is far from realistic
- Cross traffic can either increase or decrease the dispersion Δ_R, causing underestimation or overestimation of the path capacity

Capacity underestimation: if cross traffic packets are transmitted between the probing packet pair at a specific link, increasing the dispersion to more than *L/C*

Capacity overestimation: if cross traffic delays the first probe packet of a packet pair more than the second packet