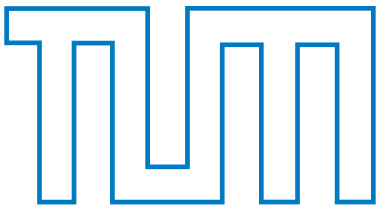

Advanced Computer Networking

Summary

Author: Thomas Pettinger

2017–03–03

Advanced Computer Networking
Technische Universität München

Table of Contents

1 Introduction

Terminology:

Protocols control sending and receiving of messages

Internet loosely hierarchical global network

Internet Standards
- RFC: Request for comment
- IETF: Internet Engineering Task Force
- IANA: Internet Assigned Numbers Authority

1.1 Protocols

Protocols take care of addressing, fragmentation & re-sequencing, error control, congestion control, compression, privacy and more.
The internet has an layered architecture of protocols. On the sender side, protocols take the PDU (Protocol Data Unit) from layer N+1, add their header and trailer and pass the SDU (Service Data Unit) to layer N-1. On the receiver side, the corresponding protocol takes the PDU from layer N-1, strips header and trailer again and passes the SDU to layer N+1.

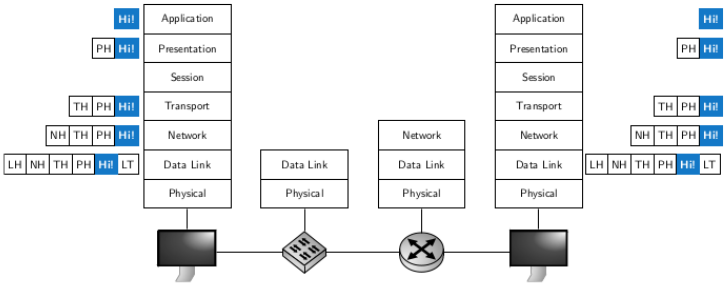

Figure 1: Internet Layers

Protocol layering is necessary because one does not want to implement everything to the physical layer when writing a networking application. On the other hand, layering also introduces some problems like protocol layers are sometimes reusing techniques of other layers like ARQ (Automatic Repeat Query) and layers might need informations of other layers.

1.2 Node Forwarding Performance

During transmission, packets might get delayed or even lost for several reasons. First, the packets need some time to get written to router buffers, secondly the packet arrival rate might exceed the output link capacity and lastly the packets need to wait again for being sent from the packet queue in routers.
The sources for these delays are listed below.

1. Processing delay: interrupt handling when receiving new packets and processing for further transmission
2. Queuing delay: waiting time in output queue
3. Transmission delay: time to send bits into link: $= \frac{\text{packet length L (bit)}}{\text{link bandwidth (bps)}}$
4. Propagation delay $= \frac{\text{length of physical link d}}{\text{propagation speed} \approx 2 \cdot 10^8 m/s}$

The total amount of delay is then $d_{nodal} = d_{proc} + d_{queue} + d_{trans} + d_{prop}$
To reduce total packet delays for a connection consisting of several links one can use circuit switching, where packets do not have to be received entirely to be sent to the next link. Another alternative is to split packets into (very) small sub-parts (= segmenting) and using pipelining (parallel computing of packets).

2 Link Layer

Terminology:

- Hosts and routers are nodes
- Communication channels between adjacent nodes are links
- A layer 2 packet is a frame and encapsulates a layer 3 packet called datagram

The data-link layer has the responsibility of transferring a datagram from one node to an adjacent node over a link.

Services

- Framing, link access, MAC addressing
- Reliable delivery between adjacent nodes (mostly in wireless transmission)
- Flow control: Pacing between sending and receiving nodes
- Error detection
- Error correction
- Half- and full-duplex (half = both ends can transmit, but not simultaneously)

Multiple Access Protocols

When sharing a single channel, a distributed algorithm manages how nodes share it. This management is done via the same channel as the actual communication and does not require a separate one coordination.

Medium Access Control (MAC) Protocols Taxonomy

Channel Partitioning divides channel into smaller pieces (time, frequency, ...)

Random Access does not divide channels, but try to recover from collisions

Taking turns Nodes take turns, requesting turns by polling or token passing

2.1 Ethernet

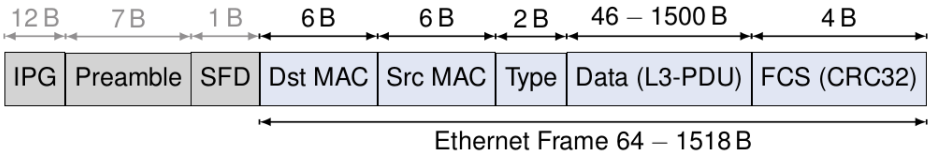

Figure 2: Ethernet Frame

IPG = Inter packet gap, minimum idle period Preamble = 7 byte (10101010...) SFD = Start-of-frame delimiter (10101011) Type = Ethernet II: Protocol type of payload, Ethernet I: length of payload in bytes PAD = Padding if data length smaller than 46 byte FCS = Frame check sequence (CRC-32)
There are several Ethernet standards, but they all share a common MAC protocol and frame format. They provide different bandwidth (from 10M to 200/400G (planned for 2017)) and have different physical layer media like twisted pairs (xBase-T), optical fibres or even chip to chip interfaces on NIC.

2.1.1 Carrier Sense Multiple Access - Collision Detection (CSMA/CD)

CSMA/CD is used for detecting and reacting to collisions. Its steps are

1. NIC receives datagram and creates frame
2. If NIC sees channel idle, it starts transmission, if channel busy, wait until idle
3. If NIC does not detect another transmission during its own transmission, it is done
4. If NIC does detect another transmission, jam signal is sent and transmission is aborted
5. NIC enters exponential backoff: after m-th collision, NIC chooses k at random from 0, 1, $\dots 2^m - 1$ and waits $k \cdot 512bit$ times and returns to step 2. Bit time is $0.1\mu s$ for 10MbE

2.2 Limitations of Layer 2

- Flat addresses
- No hop count (dangerous when having loops)
- Missing protocols like ICMP
- Missing features: fragmentation, error messages, congestion feedback

2.3 MAC addresses

MAC addresses are 6 Byte long unique identifiers for NICs. Manufacturers can buy portions of the total MAC address space from the IEEE Registration Authority, which assures uniqueness. The first 3 bytes of the address in transmission order represent the Organization Unique Identifier (OUI). If the 2nd least significant byte is 0, the MAC is OUI enforced, otherwise its locally administered. MACs are transmitted in canonical form which stands for sending the least significant bit of each byte first (in memory, token ring and FDDI it is the other way around).

2.4 Layer 2 Switching

Hubs

Hubs are repeaters which means they send every bit arriving out to all other links. Because of this, frames from all connected nodes can collide with each other. Furthermore there is no frame buffering or CSMA/CD.

Switch

Switches are a lot smarter when compared to routers. They store and forward Ethernet frames only to the node that the destination MAC address belongs to. Furthermore they use CSMA/CD to access links. Hosts do not need to be aware of the presence of switches and they do not need to be configures and learn themselves. Learning is done when receiving packets: The switch then knows the location of the sender MAC address and stores it in a switch table. An entry expires after a specified amount of time. If a packet arrives, the switch table is checked if the destination is known. If yes, the packet is only sent to that node, otherwise it is sent to all.

If more switches are involved, the **spanning tree protocol** is used. It calculates a loop-free subnet of the given physical network and determines routing. The calculation steps are as followed:

1. Select root bridge, i.e. bridge with lowest bridge_ID (concatenation of 16bit bridge_priority and MAC address)

2. determine least cost paths to root
    - Every bridge determines cost of each path to root
    - Every bridge picks least cost path
    - port connecting to that path is root port
    - Bridges on network segment determine bridge port with least-cost-path to root, i.e. designated port
3. disable all other ports

Bridge Protocol Data Units (BPDUs) are used to transmit configuration information about bridge_IDs and root path costs, to notify about topology changes (TCN = Topology Change Notification) and for TCN acknowledgements.

3 Network Layer

The network layer serves the following functions:

- IP protocol for addressing, datagram format and packet handling conventions
- Routing protocols for path selection
- ICMP protocol for error reporting and router signaling

3.1 Internet Protocol

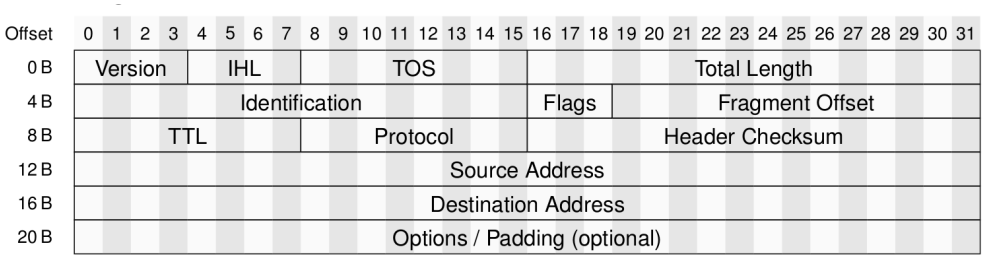

Figure 3: IPv4 Datagram

IPv4 Addressing

IPv4 addresses are 32-bit identifiers for every host and router interfaces where interfaces represent the connection between host/router and physical link.
Subnets are device interfaces with the same subnet part of the IP address which can physically reach each other without intervening router.
Splitting the IP address into network and host part is done in the following way (for the address 192.168.128.1/17):

$192_{10}.$ $168_{10}.$ $128_{10}.$ 1_{10}
$1100000_2.10101000_2.1_2$ $0000000_2.00000001_2$
network part host part

From 1982 to 1993, IP addresses were classfully divided as shown in Figure 4. In 1993, Classless Inter-Domain

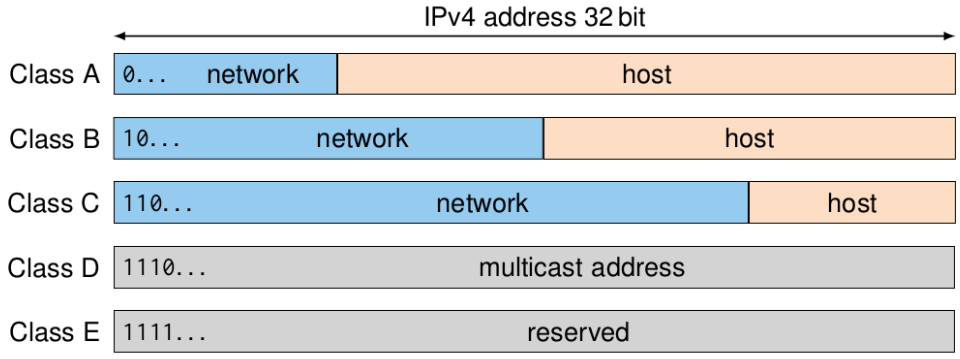

Figure 4: Classful IPs

Routing (CIDR) was introduced which allowed arbitrary subnet length. To route packets, prefix matching is used which checks which entry in the routing table fits best for the incoming packet's network prefix.

3.2 ICMP

The Internet Control Message Protocol (ICMP) are located above IP but can be considered as part of the IP layer. It is used for communicating error messages and other attention requiring conditions for IP and TCP or UDP. Two classes of ICMP messages are possible:

1. Query messages: only kind that generates other ICMP messages
2. Error messages: contain IP header and first 8 bytes (today as much as possible up to 572 bytes) of datagram that caused the ICMP message which allows the receiver to put it into context

The structure of an ICMP message is shown in Figure 5.

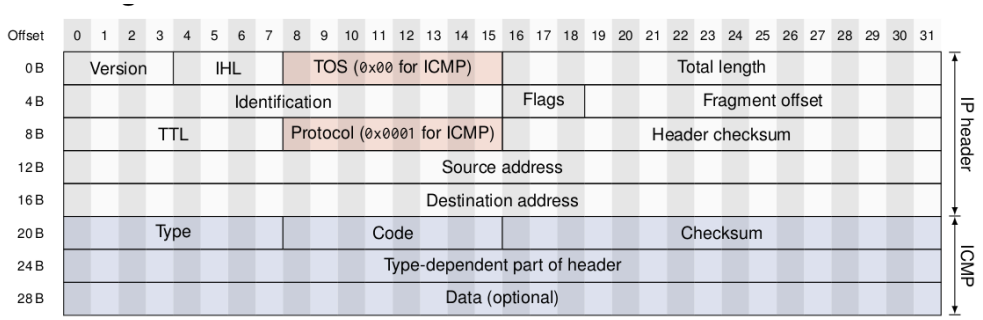

Figure 5: ICMP Message

3.3 Active Network Measurements

Network is actively measured by several parties like network providers (to manage traffic or reduce cost), service providers (to adjust service, get information about clients, ...), clients (to check services, get best one) or researchers (for performance evaluation of algorithms). Furthermore malicious traffic can be detected.

Measurements are done with probe packets and looking at the packet loss, one-way delay, RTTs or packet inter-arrival times.

(Paris-) Traceroute

Traceroute uses different TTLs in the IP header to get the route from the source to the destination. In case of load balancing though, traceroute might fail due to the appearance of ghost paths when successive packets are routed on different routes.
Load balancing routers usually use the IP-5-Tuple to determine routes, so to fixthis Paris traceroute uses different fields than normal traceroute (e.g. destination port for tcp) to do measurements.

3.4 Address Resolution Protocol (ARP)

The ARP is used to map IP addresses to MAC addresses. For that, an ARP broadcast is sent by the sender of an IP packet to get the MAC address of the next hop. The node with the specified IP address responds and the sender caches the mapping and is able to send the resulting Ethernet frame. In case we have a network with routers, the router then again does an ARP request for the IP address specified in the received

IP header and so the procedure begins again at that point. Cached information times out when not refreshed in a certain threshold.

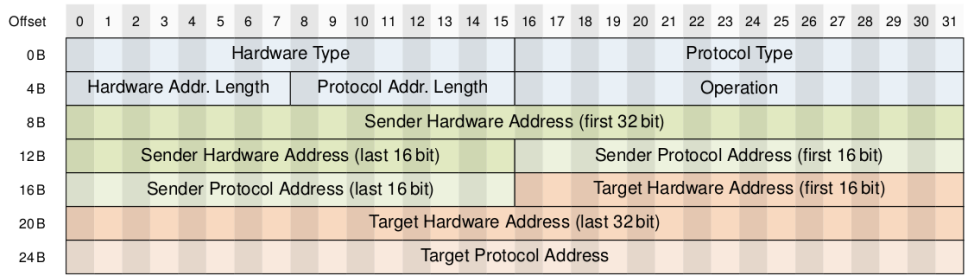

Figure 6: ARP Packet

Reverse ARP also exists, but is rarely used.
Proxy ARP also responds for ARP request of one of its networks with ARP responses for hosts of another network. This enables transparent subnet gatewaying (two LANs with in same subnet), Host joining LAN via VPN and host separated via firewalls.

Since ARP is stateless and not authenticated, ARP responses can easily be forged to poison the cache of hots which can be used to redirect traffic.

3.5 Routing

Routers are layer 3 devices that maintain forwarding tables, implement routing protocols and forward IP packets based on the forwarding table and the destination IP address.
Routing is the process on the control plane, where the forwarding table an hence the path incoming packets will follow is calculated. **Forwarding** then is the actual directing of packets to an outgoing link according to the previously calculated forwarding table.

4 Structure of the Internet

The Internet is separated into regions called **autonomous systems (AS)**. Routers in the same AS use **intra-AS routing** protocols whereas routers connecting different ASes, called **gateway/border routers** use **inter-AS routing** protocols. **Transit domains** are ASes, that forward traffic from one AS to another where in contrast a **stub domain** is an AS without transit traffic. Internet service providers are divided hierarchically: Tier-1 providers are on the top level and connected to each other. They can send traffic to one another without paying (peering). Tier-2 providers are connected to one or multiple Tier-1 providers and possibly to other Tier-2 providers. Tier-3 providers and local ISPs then are the last hop to the end systems. Every ISP has its own IP range purchased at the regional Internet Registrars which they are able to divide amongst their customers.

4.1 Associations of Internet Names and Numbers

ICANN Internet Corporation for Assigned names and numbers: Administration of DNS TLDs

IANA Internet Assigned Numbers Authority: Assignment of Internet Numbers, administration of DNS root name servers and reverse DNS infrastructure, Assignment of protocol names and numbers

NRO Number Resource Organization: Association of the 5 Regional Internet Registrars (RIR)

Regional Registras Assigns IP addresses and AS numbers, administration of local Internet Registers (LIR)

RIPE Registration and administration of Internet resources: AS, prefix and routing information

4.2 Routing Algorithms

Routing algorithms are usually an applied approach of least-cost path search in weighted graphs. The costs are represented for example by the inverse link bandwidth.
They can be classified by several criteria:

- Global or decentralized
  - Global/Link State algorithms (L-S): All routers know the graph topology and link costs (usually through broadcasts) and are able to calculate the routing table by themselves (usually via Dijkstra)
  - Decentralized/Distance Vector algorithms (D-V): Routers only know neighbours and link costs to neighbours, routing tables are computed in collaboration
- Static or dynamic
  - Static: Routes change slowly over time
  - Dynamic: Routes change more quickly due to periodic update and in response to link cost changes
- Scope: Intra- vs Inter- vs special purpose
- Type of traffic: Unicast vs multicast
- Trigger type: permanent routing vs on-demand routing (create routing table only if necessary)

D-V Algorithm

A typical example for a distance vector algorithm is the Bellman-Ford algorithm:

1. Define $D_x(y)$ as the estimate of the least cost from x to y
2. Node x knows all costs to each neighbour v: $c(x, v)$
3. Every node x maintains a distance vector $D_x = [D_x(y) : y \in N]$ where N is the set of nodes
4. Node x also maintains the distance vectors for each neighbour $D_v = [D_v(y) : y \in N]$
5. Update messages for the estimated distances are sent from time to time to neighbours and might lead those to update its own distance vectors according to the B-F equation: $D_x(y) \leftarrow \min_v c(x, v) + D_v(y)$ for each node $y \in N$
6. Under minor, natural conditions these estimates of $D_x(y)$ to the actual least costs $d_x(y)$

A problem which occurs with this approach is that if a link becomes unavailable and thus its cost infinity, the algorithm will encounter the count to infinity problem. The paths to the disconnected node are increased per update by one, infinitely. Solutions for this are

- Finite infinity: set infinite costs to a specific number, e.g. 16 in RIP
- Split Horizon: Tell neighbours that they are part of the best path to a destination that the destination cannot be reached from the original node
- Poisoned Reverse: Actively adverse a route as unreachable to neighbours from which the route was learned

Path Vector Protocols

Path vector protocols try to improve the fact of D-V protocols that they do not include topology information. For each destination, the entire path for each destination is told to neighbours and then the cost calculation is done by looking at the paths. Furthermore loop detection can easily be done by searching if the own node ID appear in the paths. PV protocols are quite rarely used though, mainly in BGP but that is much more complex than just paths.

Intra-AS Routing/Interior Gateway Protocols (IGP)

1. RIP: Routing Information Protocol
2. OSPF: Open Shortest Path First (hierarchical LSA), usually in medium to large systems
3. IS-IS: Intermediate System to Intermediate System, medium-sized ASes
4. (E)IGRP: (Enhanced) Interior Gateway Routing Protocol, CISCO proprietary, hybrid of LS and DV

The open shortest path first protocol (OSPF) uses an link state algorithm to generate routing tables. Advertisement of topology and costs of the directed graph is done via advertisement flooding. All messages are authenticated to prevent malicious intrusion (e.g. with IPsec). Furthermore multiple same-cost paths are supported and different metrics are considered to define the costs for links. The protocol has integrated unicast and multicast support (Multicast OSPF) that uses the same topology database as OSPF which lowers traffic. To even further reduce the traffic, hierarchical OSPF can be used in large domains where a two-level hierarchy is created. On the one side the backbone which are running OSPF among themselves and on the other hand local areas. Area border routes summarize distances to networks in the own area and advertises them to other area border routers.

Inter-domain routing

Inter domain routing is almost exclusively handled with the Boarder Gateway Protocol (BGP). It provides means to obtain subnet reachability from neighbouring ASes (external BGP, eBGP), propagate that information in the AS internally (internal BGP, iBGP) and determine good routes according to that information and router policies via semi-permanent TCP connections. ASes advertise reachable network prefixes to others and give a promise to forward traffic to that IP address space. These advertisements include a multitude of BGP attributes like AS-Paths (Path of AS-Numbers the advertisement has passed through) or the Next-Hop (gateway router to the next-hop AS).
BGP messages can have the following types:

- OPEN: open a BGP session
- NOTIFICATION: error occurred, close BGP session
- KEEPALIVE: null data to prevent closing of TCP session
- UPDATE: about changed routes, also removed routes
  These messages consist of the destination IP prefix, the AS path and the next hop and other attributes related to local preferences, route origins or others. Routers than can make routing decisions based on this information and their policies.

Routers may learn about multiple routes for a prefix. If that is the case, one of those routes has to be selected due to criteria like an policy decision, shortest AS-Path or closest next hop (hot-potato-routing) amongst others.

In the context of inter-domain routing, we define the following **terminology**:

Transit AS Relays traffic between other ASes

Stub AS Buys transit from one other AS but does not offer transit

Multi-homed AS Buys transit from ≥ 2 other ASes, does not offer transit

Peering having a BGP relationship

- Private peering: peering between ASes in private locations like ASes or neutral server rooms
- Public peering: "official" peering locations ("Room full of switches") like in Frankfurt or London

Provider Offers transit traffic for receiving money

Customer Gets transit for paying money

Siblings Mutual transit agreement to provide connectivity of the rest of the Internet for each other, so kind of an very extensive peering

Business and Policy Routing

Routing is done by the policy

$$\text{Routes via customer} > \text{Routes via peer} > \text{routes via provider}$$

In route announcement on the other hand first announce routes that incur financial gain if others use them, then routes that reduce costs if others use them and especially do not advertise routes that incur financial loss as long as an alternative exists. ASes might add the same AS number subsequently to an AS-Path to increase path costs if they prefer another connection over the one this announcement was sent, might be due to lower costs.

Tiers and Default-Free-Zone

Like mentioned in the introduction to this chapter, different tiers of providers exist. With our definitions in inter-domain routing of costumers, providers and peering, we can now better define them:

Tier-1/Default-Free-Zone (DFZ) Only have customers and peers, no providers

Tier-2 only peerings and only tier-1 providers

Tier-n at least noe tier-(n-1) provider

Internet Fixed Points

Internet fixed points are ASes that are stable over a long period of time from different perspectives. Together these form the so called backbone of the Internet. To find those fixed points, the **k-core algorithm** can be applied:

1. Remove all nodes with $degree = 1$ so long until no degree 1 nodes are left
2. Remove all nodes with $degree = 2$ so long until no degree 2 nodes are left
3. Do this until no nodes left $\Rightarrow (Steps - 1) - core$ found.

5 Network Measurement

Network performance can be measured with different metrics like throughput (bandwith or packet rate), latency (average, median, standard deviation,...), frame loss rate and others with different circumstances (load, traffic type,...). Different RFCs standards exist as guideline.

5.1 Throughput

Throughput is usually limited by the line rate and the speed and size of the lookup tables. It is measured in packets per second over the bandwidth since routers usually only look at packet headers and not the entire packet, so the actual size has only a minor importance. With this measurement unit the worst case scenario is network traffic at line rate and minimum packet size which is the minimum sized Ethernet packet plus the 7 byte preamble, 1 byte start-of-frame delimiter and the minimum inter-packet gap of 12 bytes, thus 84 bytes.

When testing, different measuring methodologies can be applied. The simplest one is to apply the highest possible packet rate on A and measure the packet rate at B. Though with this method, the devices might get overloaded which leads to different behavior. So a better version is to apply varying rates on A and find the highest rate were no loss occurs (RFC 2544). Problems of this approach again are that some devices loose packets when suddenly facing high packet rates due to energy saving mechanisms. As a summary the best approach depends on the device under test.

Improving Throughput

Potential bottlenecks for packet forwarding are CPU processing power, NIC processing power, Bus bandwidth, memory bandwidth or CPU caches. As researches found out, the biggest limitation origins in the CPU. The most time is spent to process, receive and transmit packets there. When switching from kernel to user space, performance can be significantly increased due to fewer expensive system calls, simplified memory management and batch processing through the whole application. The disadvantages on the other hand are that only raw packets are handled, so protocols have to be reimplemented for every application, NICs can only be used by one application and there is no API compatibility to traditional user space applications.

5.2 Parallel Packet Processing

Modern NIC cards hare configurable to use multiple rx and tx queues to support multi-core parallelization to improve performance. Several metrics to distribute incoming traffic on the queues exist:

- Per-packet basis: Slow when protocol state has to be synchronized and might cause packet reordering
- Per-flow basis: Fast, protocols handled in the same core and cache and prevents packet reordering
- Explicitly: Useful for e.g. virtual machines, slower than flow-based though

Usually packet forwarding is done in kernel space due to better performance than the socket API.

5.3 Latency

Sources of latency are serialization, propagation and calculations where buffers usually are the biggest bottleneck. Also the technique to receive packets plays a role:

- one interrupt per packet: low latency but also low throughput because interrupts are expensive
- one interrupt for multiple packets: high throughput but also high latency

- no interrupts but polling based: low latency and high throughput but inefficient at low packet rates (busy waiting)

5.4 Packet Generators

Packet generators exist in hardware and software varieties. Hardware generators are fast, precise and accurate. Software ones run on cheap hardware and are very flexible but face challenges with rate control and time stamping.