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.

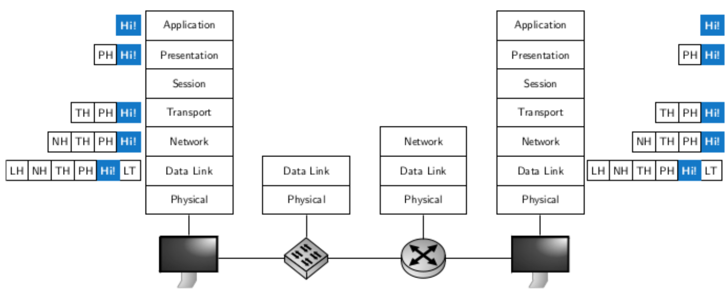


Figure 1: Internet Layers

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 for 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

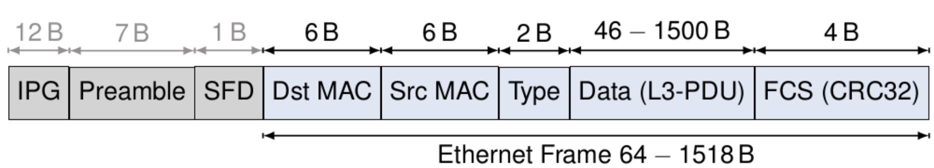


Figure 2: Ethernet Frame

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, $\ldots 2^m - 1$ 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 hubs. 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

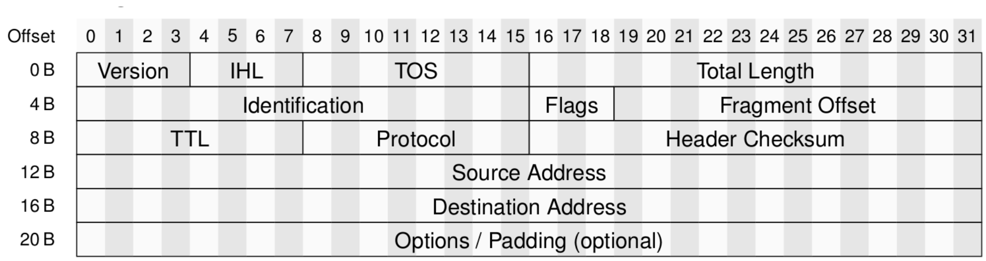
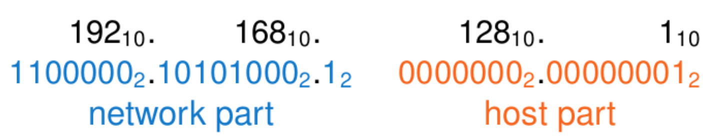


Figure 3: IPv4 Datagram

IPv4 Addressing

IPv4 addresses are 32-bit identifiers for every host and router interface 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):



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

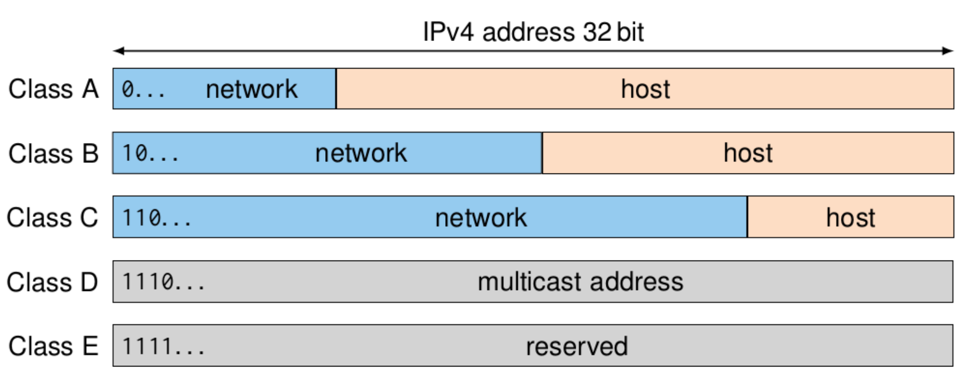


Figure 4: Classful IPs

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.

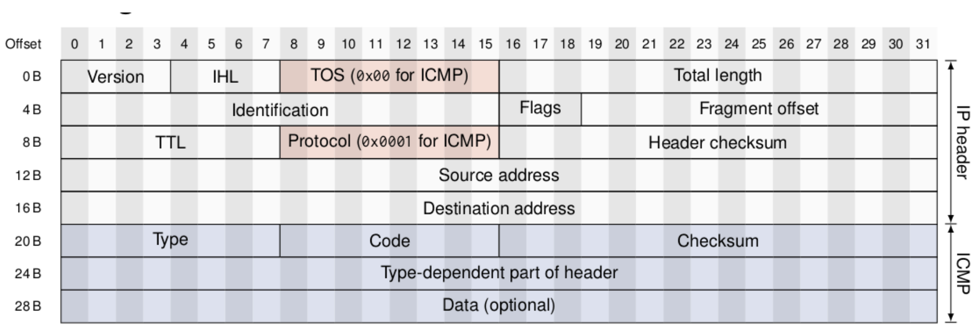


Figure 5: ICMP Message

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 fix this Paris traceroute uses different fields than normal traceroute (e.g. destination port for tcp) to do measurements.

3.4 Address Resolution Protocol (ARP)

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 within a certain time threshold.

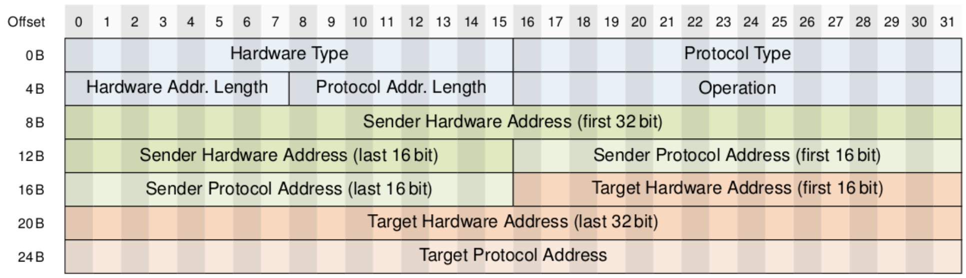


Figure 6: ARP Packet

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), a host joining LAN via VPN (VPN server does proxy ARP) and to include host that are separated via firewalls (firewall handles proxy ARP).

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 now 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 for different metrics like throughput (bandwidth or packet rate), latency (average, median, standard deviation,...), frame loss rate, topology measurements or others with different circumstances (load, traffic type,...). Different RFC 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 (not bandwidth) since routers usually only look at packet headers and not the entire packet, so the actual size has only a minor importance. The worst case scenario regarding costs 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. With this method, the devices might get overloaded though which leads to different behaviors. 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 transferring the network stack 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.

To control the packet rate software implementations push single packets to the NIC where queues cannot be used. Also the NICs work with asynchronous push-pull models which can lead to micro bursts and thus to unreliable, imprecise and bad performance. Hardware generators on the other hand support hardware rate control where queues can be used and have good performance but they are quite inflexible. To combine the advantages of both, one can disable hardware control and use invalid packets in the queues to control the rate since those are simply dropped by the device under test without much overhead.

5.5 Internet Wide Scans

When doing larger scale network scans are performed, several points have to be taken into consideration. For one which targets are selected which might be a specific hitlist provided by e.g. traceroute, web server logs or traffic traces, certain IP addresses per subnet or even a full 0/0 scan. Also there are performance requirements that need to be met and ethical considerations, too, since one causes sometimes large amounts of traffic on the scanned network.

Nmap

Nmap is a common measurement tool which provides host discovery, service detection, OS detection and support for custom scripts. It provides a multitude of scanning techniques:

- TCP scan: Sends TCP packets with different flags set. A SYN scan checks for open ports, ACK scans scan for (un-) filtered ports by a firewall. Some more exist.
- ICMP scan: ping requests
- UDP payload scan: Sends UDP packets with different payloads

While scanning, randomization is used to avoid complaints of system administrators but only groups of 16k hosts are possible. Nmap uses a stateful scanning approach to keep track of every packet in transit and to catch timeouts to try again to send packet.

Zmap

A full Internet scan using nmap takes around 10 days which is quite slow. For that reason, **zmap** was developed by the university of Michigan which is able to do a full scan in around 45 minutes. It uses TCP SYN or UDP payload scans to find open ports and it is possible to distribute the scanning load on different machines and every IP is only scanned once. The performance is reached by using raw sockets and not keeping state of packets which makes it impossible to detect timeouts though. This is handled by cycling through the host IPs that have not yet responded a certain amount of times and then abort. Furthermore without keeping states, incoming packets that belong to the scan are more difficult to identify. Zmap therefore uses IP IDs which are used to generate a validation with AES which is stored in the packet sent e.g. in the sequence number When receiving packets, they are validated using the validation, in the example from before $\textit{sequence number} - 1$.

IPv6 Scanning

IPv6 scanning faces different challenges than IPv4 due to different routing, firewall and host configuration and the huge address space. To to the amount of possible addresses it is not possible to do a 0/0 scan so hitlists are used to define the targets. Sources for such hitlists are for example the Alexa Top 1 Million list (most popular websites), the Rapid7 DNS ANY list or DNS zone files (content of TLD name zones). Also traceroute or passive sources like packet traces or flow data can be taken into account.

IPv6 scans can be evaluated for reachability or stability. Also as long as the targets do not use IPv6 privacy extensions (Interface ID randomly chosen) the target device type can be determined. The criterion is that routers usually have the IID ::1 for the default gateway.

Security Scans

TLS scans are used to scan the state of TLS protocols like HTTPS or IMAPS in the Internet by analyzing certificate chains, expiry and algorithms. Scans are done by identifying hosts that offer TLS services, download the certificate chains and analyse and validate these chains.

SSH scans mainly provide an overview over the security of the SSH configurations of hosts with public IPs. Since SSH is mostly used for administrative or security sensitive contexts it is usually advisable to notify CERTs, watchlist services and blocklist operators about scans. Also other measurements like an own scanning subnet with a abuse WHOIS email contact might be useful to hide one's identity.
In the past several Internet wide SSH scans were performed which analysed key strength (length, debian weak keys, duplicate keys).

IPMI (Intelligent Platform Management Interface) is an out-of-band management system used in servers. It uses a separate OS which has full access to the host OS. IPMI scans try to detect known vulnerabilities in configurations, i.e. hosts should not be reachable from the public internet. When combining IPMI responses with HTTPS reachability to detect IPMI web-interfaces which also might be vulnerable. Compromised web servers again may lead to a compromised OS.
Internet wide scans for IPMI-over-IP devices are showing declining numbers of reachable devices. Those who are reachable are heavily clustered in a view ASes though and detected HTTPS interfaces show that 90% of them have 1024 bit or shorter keys.

BACnet (Building Automation and Control Networks) is a protocol that is used to control heaters, solar panels and other building automation aspects. Access to those systems can lead to real world consequences. It is based on UDP and has no build in security. Devices providing this protocol have properties that can be queried via a SingleProperty or MultipleProperty request which enables attackers to generate larger responses which results in easier DOS attacks. To detect these easily possible attacks, scans are performed to check the BACnet deployment in the internet.

Ethical considerations

It is usually advisable to take some ethical steps before doing Internet wide scans. These include reducing the intrusiveness of scanning by avoiding logins or limit the scanning rate, providing information on the scanning machine's website, respond quickly to every inquiry and abuse email and offer possibilities for blacklisting IPs. A general guideline here is to be the "nice guy".

5.6 Passive Measurements

Passive network measurement does not cause additional traffic like the approaches mentioned above but uses monitoring probes to analyze existing traffic. Important metrics here are traffic volume, traffic composition and packet inter-arrival times with different levels of granularity. When captured, different analyses can be performed, e.g. network utilization, QoS parameters, detection of failures and anomalies or traffic characterization. These informations might be useful for accounting, security or traffic engineering.

The capturing hardware usually has to be able to perform on multi-gigabit/s links with relative cheap costs and simple deployment. Common approaches to these requirements are high-end network adapters with large amount of memory and functionality (still cheap?), sophisticated algorithms by eliminating copying of packets amongst others, sampling (probabilistic filtering of packets), filtering (mask, router state, hash based) and aggregation.

The process passive measurement is depicted in Figure 7.

Figure 7: Passive Network Measurement Process

Packet capturing is assisted by hardware. Server NICs have direct access to the main memory without processor support and do batch processing to reduce copy operations. Also special monitoring interface cards exist which usually only are able to receive data and provide certain processing features like filtering, high-precision time-stamps and others.

Flow-based Traffic Measurements

Flows describe packets which belong together like all packets of a TCP connection (IP-5-Tuple). Flow data is usually measured passively on the network and is exported when on of two timeouts runs out. The inactive timeout starts at the last received packet of a flow and is reset with every packet. The active timeout starts with the first packet of a flow and is only reset if it expires. The inactive timeout was designed to export flow data of short lived flows whereas the active timeout sends data during a flow is active for long lived ones.

IP Flow Information eXport (IPFIX)

The IPFIX protocol was defined by the IETF in RFCs and is an extensible flow exportation protocol. The extensibility is achieved by differentiating between template and data records where the template defines the data format for the data records.
During measurement, statistical counters and values are updated and whenever a flow terminates, the data is exported via SCTP or, if available, TCP or UDP.
The metering process of IPFIX includes packet header capturing, timestamping, classifying and the maintaining of flow records where a flow record contains information about measured properties of the flow like total number of bytes in the flow or IP addresses. Exporting then sends flows to one or more collecting processes. In the end data is collected from all capturing points and further processed.
Metering and exporting can be done on network devices directly, or be on separate hardware. Collecting is usually done separately.
Sampling and filtering can be used for very high-speed networks.

Anomaly Detection with Machine Learning

Machine Learning can be used to detect anomalies based on flow data. For this, feature vectors have to be created from flow data with numerical and categorical semantics. An example for such a transformation of the data is shown in Figure 8.

Flow data	Feature vector
Flow start time & end time	Flow duration (numerical)
Protocol = 6	Protocol TCP (categorical)
TCP Control Bits = 0x16	Connection RST (categorical)
…	…

Figure 8: Feature Vector Creation

Supervised machine learning then uses labeled training data to learn what is benign and malicious traffic. **Unsupervised ML** on the other hand has no training data available. It tries to find clustered data and outliers. The outliers then represent an anomaly.

To assess the quality of these approaches, different metrics can be used:

- Precision $= \frac{\text{true positives}}{\text{true positives + false postives}}$
  How many of the detected anomalies are actual anomalies?
- Recall $= \frac{\text{true positives}}{\text{true positives + false negatives}}$
  How many of all actual anomalies did I detect as anomalies?
- Accuracy $= \frac{\text{true positives + true negatives}}{\text{all}}$
  How many positive and negative classifications are correct?

The goal is to decrease false positives and negatives here, but in reality decreasing on type of error increases the other.

5.7 Amplification Attack Detection

In amplification attacks, the attacker sends small request to an amplifier network with a spoofed IP address which generate large response packets to the victim (owner of the spoofed address).
The amplifier network might be able to detect such an attack though by passively measuring incoming and (potential) outgoing traffic. Different characteristics can be used then to identify an attack:

- Amplification factor: compare incoming and outgoing traffic, if asymmetric an attack is possible
- Packet size similarity: Same sized packets incoming frequently
- Payload similarities: Packets from the attacker have similar payload content. Payload similarities are detected by a low entropy.
- Unsolicited ICMP messages: backscatter ICMP message of the victim are indicator for an attack
- TTL measurements: path form attacker $\neq$ path from amplifier to victim indicates an attack

5.8 Hybrid Measurements

A hybrid approach between active and passive measurement can be taken where packet flows are modified by piggybacking or header modification. This enables adding additional information to packets without applying additional load to the network. This has to be taken with care though, since people might not like this.

5.9 Detecting Traffic Misdirection in Interdomain Routing

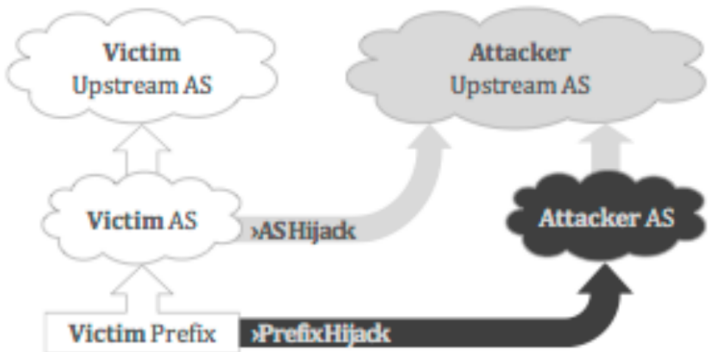


Figure 9: Possible attacks

In prefix hijacking attackers announce a victim's (sub-) prefix to other ASes. We already have real-time detection for that. AS Hijacking is a more sophisticated approach where a letter of authorization has to be accepted by ISPs as a legitimation to advertise resources of a customer's AS. Such an attack is usually carried out over several month. In the example of LinkTel, the attackers re-registered an expiring domain (link-telecom.biz), forged an letter of authorization to the upstream provider and announced false BGP routes.
In 2013 an escalation warning system for AS hijacking was designed at the TUM which uses passive monitoring of DNS expiry an re-registration and analysis of reverse DNS and BGP activity to identify vulnerable targets.
Current hijack detection is done via multiple traceroute scans from multiple vantage points. To identify the poisoned part of the network, last hops to the target prefix and downstream graph of last hops are regarded. Possible detection metrics are a detection of an odd distribution of first-rank countries (5 neighbours Germany, 1 New Zealand), an odd distribution of first-rank ASes or odd RTTs. Also Exclusive AS connectivity can be used to detect how many ASes in the graph are connected exclusively through one neighbour of the target. With that approach a segmentation of the graph is possible which allows the specification of the impact of a possible hack.

6 Software Defined Networking

Traditional networking has the problem that distributed connectivity algorithms do often times not find the best solution (e.g. spanning tree protocol) and scenario-specific requirements are generally hard to implement. Furthermore due to the lack of abstraction, it is hard to manage a network and the innovation is slow. Theses problems are tried to be solved by software defined networking.

The idea behind SDN is to have the control plan logically separated from the forwarding plane so that is handled centrally. This way, the central control point can be used to calculate spanning trees or load balancing and thus render worse distributed algorithms unnecessary. Furthermore the forwarding behavior can be specified more precisely, e.g. forbidding traffic from one VM to another and an higher level of abstraction can be introduced by adding a API level between hardware devices and control plane. This abstraction greatly increases the speed in which the forwarding logic can be modified since instead of having to buy new hardware that all speak the same protocol, only some software changes have to be made. The actual forwarding plan then only executes the specified behavior of the control plane. Figure 10 shows the model of SDN.

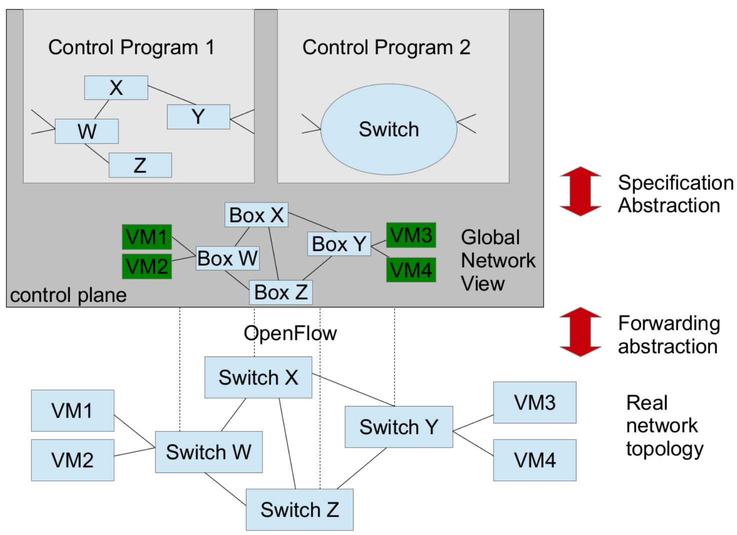


Figure 10: SDN big picture

6.1 Network Operating Systems

A network operating system like for example OpenFlow manages network hardware, provides SDN control plane services and provides a standardized API to hardware resources. Figure 11 shows the abstractions in a NOS. With an NOS we have a central control plane which handles the difficult network and routing computations. This way the actual forwarding switches are only "dumb boxes" which are connected via ssl to the NOS which runs on fairly common CPUs.
Some disadvantages of SDN are that more configuration is required than in traditional forwarding and a single point of failure exists.

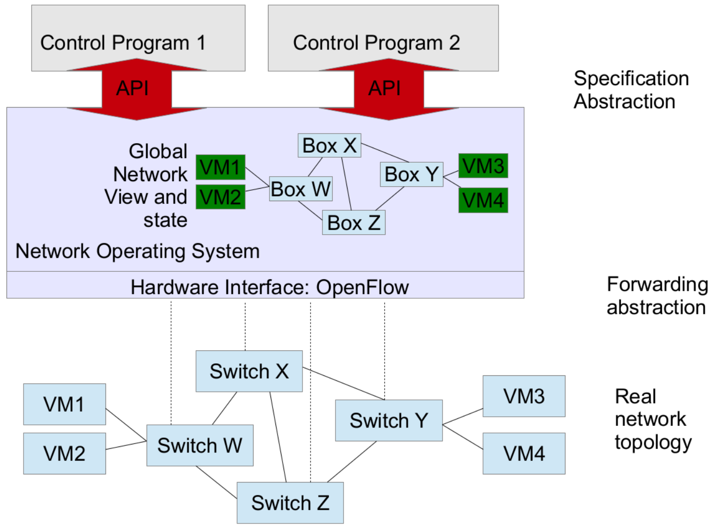


Figure 11: Network Operating System Model

7 Quality of Service

Quality of service are performance guarantees given to customers in a service level agreement (SLA). These performance guarantees might be important for different scenarios like streaming, interactive applications (games,...) or safety-critical application or safety-critical applications. SLAs are applied at different levels: packet, flow, application or user level. To guarantee the specifications in an SLA one has to perform different tasks:

- Modelling: Understand which parts of the network have an impact on the SLA
  These usually are propagation, processing, transmission and queuing delay.

- Classification: Identify which packet need SLA
  Several methods exist for identification like using packet header fields like IP-5-Tuple or IPv4 TOS field or another alternative is to do deep packet inspection.

- Scheduling: Give preferential service to network packets. Two types of schedulers can be differentiated: work-conserving schedulers only are idle if no packet available whereas non-work conserving ones might be idle even if packets are pending.
  The common architecture for scheduling is to have multiple queues with different priorities. The queues are pulled if all queues with higher priority are empty.
  A slight deviation of this approach is round robin, where the queues are polled after one another.
  A third approach namely **weighted fair queuing** tries to solve the problem of round robin queueing, where the bandwidth per queue depends on the packet sizes, and priority queue scheduling (starvation) by splitting the actual available bandwidth according to the weights of different queues. The implementation is much more complex though and thus it is rarely used in real switches.

- Monitoring: Check actively or passively if SLAs are met.
  This can be done with live tests, emulations or simulations or formal verification methods. The more critical real-time requirements an application has, the more precise the measuring algorithms typically are.

7.1 Deterministic Network Calculus

Network calculus is a framework developed for analyzing performance guarantees in networks of queues and schedulers. In the deterministic variant, no randomness is involved whereas the stochastic model includes randomness and thus is characterized in probabilistic terms.

Packets and network protocols are described as **flow**, meaning an unidirectional set of packets going from a sender to a receiver and are modeled by a **cumulative arrival function** A. $A(t)$ represents the amount of data sent by the flow in the time interval $[0, t)$.The The **deterministic arrival curve** then is defined as $A(t) - A(s) \leq \alpha(t-s), \forall 0 \leq s \leq t$. A simple example for this would be the **token bucket** $y_{r,b}(t, s) = r \cdot (t-s) + b$ where r denotes the average rate and b the burstiness parameter.

Queues and schedulers are seen as **servers** in network calculus. They have a **deterministic service curve** β such that the output curve $A^* \geq \inf_{0 \leq s \leq t}\{A(s) + \beta(t-s)\}$ A simple example for this is the **rate-latency** which is defined as $\beta_{R<T}(t) = R[t-T]^+$ where R is the rate, T the processing delay and $[x]^+ = max(0, x)$. Delay is defined as time time it takes for a packet to traverse the queue and queue size as the backlog size at the server. For a visualization see Figure 12.

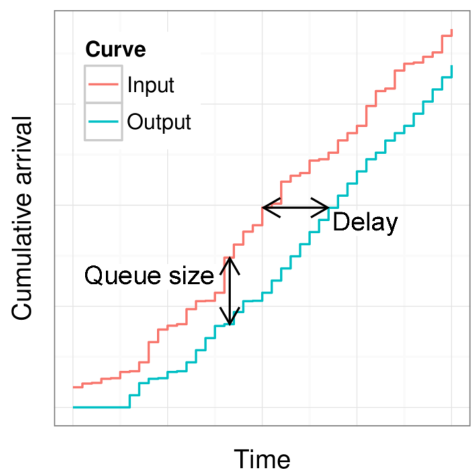


Figure 12: Delay and Backlog

Deterministic network calculus can be used when there are no cyclic dependencies, no feedback loops, no simple analysis exists (state explosion) and when a good model of the traffic is known.

7.2 Stochastic Network Calculus

In stochastic network calculus flows are defined by a sequence of non-negative, real random variables $(a_n)_{n \in \mathbb{N}}$ of random size. Their cumulative arrival up to time n is then defined as $A(n) = \sum_{i=a}^{n} a_i$. (a_n) can follow any random distribution and are considered to be independent and identically distributed. Similar to DNC, service curves are $S(n, m)$ of a server are defined as $A^*(n) \geq \inf_{0 \geq k \geq n}\{A(k) + S(k, n)\}$ where $S(n, m)$ can either be a stochastic or deterministic process.

Stochastic network calculus is mainly used for video streaming, protocols with feedback loops, wireless networks or energy networks.

7.3 QoS in IP Networks

In modern IP networks traffic is usually limited to a fixed set of declared parameters concerning average, peak rate and burst size. To implement these limits, a token bucket can be used which fills with a pre-specified rate until it is full. The limitation is then given by the cost of forwarding of one token. To guarantee an upper bound, this approach can be combined with WFQ.

IETF Integrated Services

The IETF integrated services provide an architecture for providing QoS guarantees in IP networks for individual application sessions.
If a flow/call arrives, resources have to be requested with information about the estimated traffic and reservation characteristics. Routers then decide based on this information and remaining unreserved resources if the request can be accepted and answers accordingly.
Two service models are possible here. A guaranteed service where worst-case traffic is important or a controlled load service where the QoS closely corresponds to the QoS that the same flow would receive from an unloaded network.

IETF Differentiated Services

IEFT Differentiated Services want qualitative service classes (platinum, gold, silver services). Implementation here differs regarding core and edge nodes. The edge routers look at the per-flow traffic, marks packets according to a class as "in-profile" or "out-profile" and forwards to core nodes regarding a token bucket. The core nodes then only look at the packet classes and do buffering and scheduling according to that information where "in-profile" packets have higher priority.

Classification and Conditioning

For classification the TOS field of the IPv4 header or the Traffic class field of the IPv6 header can be used. They consist 2 bits for a explicit congestion notification (ECN) and 6 bits that can be used for Differentiated Service Code Point (DSCP) and per hop behaviors (PHB).
Developed PHBs:

- Expedited Forwarding: logical link with a minimum guaranteed rate.
- Assured Forwarding: 4 classes of bandwidth with defined guaranteed minimum bandwidth and buffering, packets can have one of three possible drop preferences

7.4 Resource Reservation Protocol (RSVP)

RSVP is used to communicate/signal requirements to the network. This is done by sending information about the QoS that is required (r-spec) and the characteristics of the traffic that will be sent into the network (t-spec) to routers.
More precisely senders and receivers first join a multicast group. The sender then sends a path message to make its presence known to the routers which store path states which consist of the IP address of the previous node and a sender template (data format of the sender), sender t-spec and adspec (advertisement data). In the end of the entire communication that required RSVP, a path teardown message is sent to remove the path state from routers.
The receiver send a reservation message to reserve resources from sender(s) to receiver. At every node, the destination IP changes to the next node on the reverse path and the source IP gets the address of the previous node. In the end a reservation teardown message is sent to remove reservations. The network only sends messages when errors occur.

7.5 Maintaining Network State

A state in this context is defined as information stored in network nodes by network protocols.
A **hard state** is installed on receiving a setup message from the sender and only removed when getting a teardown message or sometimes a heartbeat is used. A **soft state** on the other hand is also created when receiving a setup or trigger message from the sender but is removed when the connection times out. For this reason a heartbeat is definitely necessary.
Senders in this context are nodes that (re)generate signaling messages to install, keep and remove states. Receivers on the other hand are nodes that creates, maintains and removes states based on the messages of the sender.

The advantage of soft state over hard state here is that no explicit error messages need to be sent to indicate problems. In case of a failed link the update messages will update the routes and in case of a failed host the connection will just time out. A problem is though that if the first path message gets lost the timeout might be quite long to retry. For that reason ACK messages can be used from receiver to sender to be able to use a quicker timeout than the normal heartbeat.

8 Node Architectures

Nodes/Routers in a network consist of multiple input and output ports and a switching fabric in between. There are three types of switching fabric: memory, a bus, or a crossbar (net of buses).
Usually the goal of a router is to completely process the incoming packets at line speed, but queuing/buffering might be necessary due to a too slow switching fabric or a overloading of one or multiple output ports. Recommended buffer sizes are $RTT \cdot \text{link capacity}$ by RFC 3439 or $\frac{RTT \cdot \text{link capacity}}{\sqrt{\text{number of flows}}}$ more recently.

First generation IP routers had a CPU next to a global memory buffer which was connected to multiple line interfaces. **The second generation** then introduced local memory buffers for all interfaces and the **third** moved the CPU to a separate card besides the network interfaces. In **the forth generation** clustered and multistage network interfaces were added.

9 Internet Transport Layer

9.1 Congestion Control

One of the transports layers jobs is to handle congestion. Congestions happens if e.g. too many sources send too much data to fast for the network to handle which results in packet loss and long packet delays.
Congestion control tries to solve this because without controlling the outgoing traffic, capacity may drop dramatically because of congestion collapse. **End-end congestion control** infers congestion only by observing lost packets and delay whereas **network-assisted congestion control** use informations of routers to detect congestion. Also an explicit rate can be told to the sender by the network with the second approach.

Self clocking congestion control is another approach which sends a new packet for every packet that left the network what the sender knows from ACK messages. It assumes though that packet loss only occurs due to congestion which is not true for wireless networks for example.

9.2 Transport Control Protocol (TCP)

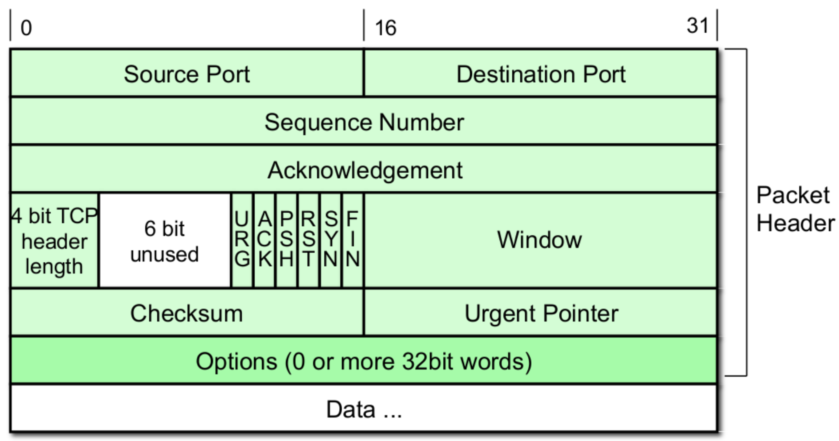


Figure 13: TCP Header

TCP is an connection oriented protocol that sends packets in order and does retransmission for lost ones. For the detection of losses, an acknowledgement flag (ACK) is used. It has an mechanism to avoid losses as good as possible by not overloading the receiver (size of rwin) and network (cwin) that adjusts the sending window to $swin = \min(rwin, cwin)$.

A connection is established with the TCP handshake which consists of a SYN, SYNACK and ACK message. In that, the receiver tells the sender its maximum segment size (MSS) which represents the maximum size of a TCP segment it is able to receive. When this is done, TCP continues with the so called **slow start** where the cwin is 1 MSS. To quickly ramp up the transmission, the cwin is increased exponentially until a sender threshold is met. TCP then enters the congestion avoidance phase where the cwin is increased linearly. If a packet gets lost (e.g. 3 duplicate ACKs), the thrshold is cut in half. **With fast recovery**, cwin also cut in half an congestion avoidance starts again. **Without fast recovery** cwin is set to 1 MSS again and slow start is applied. Besides fast recovery, one is also able to use **fast retransmit**, where not the entire cwin is sent again for missing ACKs but only the missing packet. Figure 14 shows a good overview over this.

State	Event	TCP Sender Action	Comment
Slow Start (SS)	ACK receipt for previously unacked data	CongWin = CongWin + MSS, If (CongWin > Threshold) set state to “Congestion Avoidance”	Resulting in a doubling of CongWin every RTT
Congestion Avoidance (CA)	ACK receipt for previously unacked data	CongWin = CongWin+MSS * (MSS/ CongWin)	Additive increase, resulting in increase of CongWin by 1 MSS every RTT
SS or CA	Loss event detected by triple duplicate ACK	Threshold = CongWin/2, CongWin = Threshold, Set state to “Congestion Avoidance”	Fast recovery, implementing multiplicative decrease. CongWin will not drop below 1 MSS.
SS or CA	Timeout	Threshold = CongWin/2, CongWin = 1 MSS, Set state to “Slow Start”	Enter slow start
SS or CA	Duplicate ACK	Increment duplicate ACK count for segment being acked	CongWin and Threshold not changed

Figure 14: TCP Sender Congestion Control

TCP Fairness

Two competing TCP sessions will over time get the same bandwidth as indicated in Figure 15. If an app opens

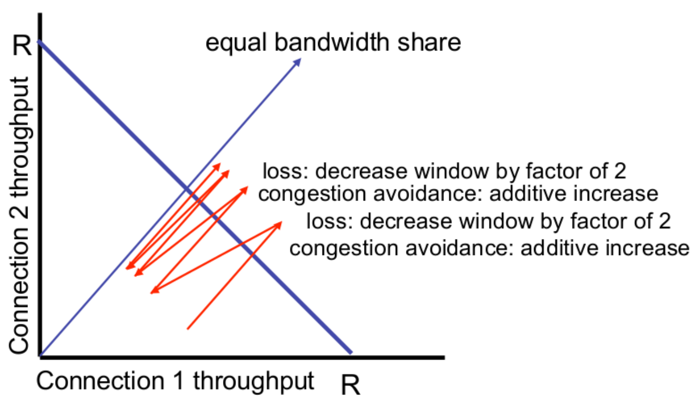


Figure 15: TCP Fairness

multiple connections though, it might get an higher share of bandwidth because every single connections gets an equal amount independent of the app. Also multimedia apps often do not want to use TCP, because they do not want the bandwidth to be throttled. For this reason UDP is oftentimes used with a TCP-friendly rate control.

Buffer Bloat

Large buffers in routes cause problems for TCP connections since once queues are full at the bottleneck, large queueing delays occur. TCP then gets no early warning about congestion cause no duplicate ACKs are sent but instead sudden timeouts are recognized. For this reason a large oscillation happens between sending way too much and send way to little (start of slow start again). Repetition: Rule of thumb for buffers: $\frac{RTT \cdot C}{\sqrt{N}}$ where C is the link bandwidth and N the number of flows.

TCP with Explicit Congestion Notification (ECN)

ECN is a 2 bit field in the IP header. It can be used on the network layer where endpoints set it to ECT(0) or ECT(1) to declare ECN capable transports and routers set ECN to CE (ECT = ECN capable transport, CE = congestion encountered).
On the transport layer, endpoints negotiate ECN at connection setup. TCP then reacts to CE as if packet is lost with a ECN-Echo to echo back the congestion experienced to the sender which reduces the cwin. ECT(0) and ECT(1) can be used to differentiate between packets and flows.

Data Center TCP (DCTCP)

DCTCP is an enhancement of TCP's congestion control algorithm that responds to congestion not only regarding its presence but also the amount of congestion.

TCP CUBIC

TCP CUBIC is a loss based congestion control algorithm that optimizes for high bandwidth high latency. It modifies the window growth algorithm so that the window grows slowly around the maximum size. Standard TCP outperforms CUBIC in short RTTs though.

TCP Vegas

TCP Vegas is a delay based algorithm that uses $i^{th} RTT > min\ RTT + delay\ threshold$ to adjust the window size.

Delay Gradient TCP

Delay Gradient TCP uses the delay gradient as a congestion indicator. It has an average probability of back off independent of RTT and works well in loss-based congestion control flows.

TCP Remy

TCP Remy is an algorithm that tries to find the optimal congestion control protocol for a given network defined by prior network assumptions and goals by using a heuristic search procedure that generates different CC algorithms.

TCP BBR (Bottleneck Bandwidth and RTT)

TCP BBR tries to estimate when the queue will become empty and sends more data just in time. This is realized by estimating the optimal rate on which the maximum bandwidth and RTT is reached, so the maximum bandwidth-delay product (BDP). Every ACK measures the RTT and delivery rate by updating the min_RTT and max_rate and adapts the rate to assure max_rate and min_RTT where the maximum observed bandwidth has priority over $cwin = 2 \cdot min\_RTT \cdot max\_rate$.
TCP BBR delivers very good performance and is already deployed in Google WAN backbones.

Multipath TCP

The IETF develops an multipath-aware congestion control transport protocol that can be used for example in smartphones with WLAN and Mobile networks.

10 Patents and Computer Networks

Companies often have invested a lot of resources into developing a product or service and patents often have significant impact on the economic success of such a firm. Computer networks rely on interoperable standards which often require patented techniques. If these techniques are patented, we speak of **standard-essential patents**. Without the right (Intellectual Property Rights IPRs) to use it, the standard cannot be manufactured.

10.1 Patent vs Copyright Law

The main difference between patents and copyrights are that patents are protections of ideas for which an application is necessary and holds for 20 years from application whereas a copyright includes only a specific form of an invention, does not need an application and lasts 70 years after the death of the author.

10.2 Patent Law

To be granted a patent, one has to fulfill several requirements according to the European Patent Convention (EPC):

- Invention: have to be new, applicable in industry and involve an inventive step, excluding discoveries, scientific theories, mathematical methods, aesthetic creations, schemes, rules and methods for performing mental acts, playing games, doing business, program computers and the presentation of information
- Novelty: absolute novelty with no restriction on time, place or person including all publications available
- Inventive step: Invention has to have something extraordinary in an area not excluded from invention (technical character)
  Methodology:
  1. Determine technical field of invention
  2. Identify closest prior art
  3. Identify differences, effect of differences with an emphasis on technical effects
  4. Deduce technical problem (faster execution speed, higher accuracy,...)
- Industrial applicaiton