Chapter 3

Transport Layer & Flows

How does the internet decide at what quality you can watch a video?

3.1 Flows

The total bandwidth of the internet must be shared by all users; this can be modeled with flows.

Definition 3.1 (Flow, Rate). Let s,t be two nodes in a directed graph. A flow from source s to destination t (also called an s-t-flow) is a function $F:E \rightarrow \mathbb{R}_{\geq 0}$ such that the following hold:

- $F(e) \leq c(e)$ for all $e \in E$ (capacity constraints)
- $\sum_{e \in \text{in}(v)} F(e) = \sum_{e \in \text{out}(v)} F(e)$ for all $v \in V \setminus \{s,t\}$ (flow conservation)

We call $F(e)$ the rate of $F$ on directed edge $e$ and the net flow leaving $s$ ($\sum_{e \in \text{out}(s)} F(e) - \sum_{e \in \text{in}(s)} F(e)$) the rate of $F$, also denoted by $F$.

Remarks:
- Since data flows are inherently directed, we only consider (weighted) directed graphs in this chapter.
- In contrast to Chapter 2, the weights do not indicate the latency of edges, but the bandwidth capacity.
- By $\text{in}(v)$ resp. $\text{out}(v)$ we denote the set of all incoming resp. outgoing edges at node $v$.
- You may wonder what happens if there is not only one flow in the graph, but if there are multiple source-destination pairs. Welcome to the world of multi-commodity flows!

Definition 3.2 (Multi-Commodity Flow). A multi-commodity flow $F = (F_1, \ldots, F_k)$ is a collection of s-t-flows $F_i$ such that for each edge $e \in E$ the sum of the flows’ rates on $e$ does not exceed the capacity of $e$, i.e.,

$$\sum_{i=1}^{k} F_i(e) \leq c(e) \quad \text{for all } e \in E.$$
3.2. LINEAR PROGRAMMING

Remarks:
- Let’s have a look at an example of a linear program. Imagine you want to throw a party. How much beer should you buy? You can buy beer for a liter price of 1, and self-made cocktails where the ingredients for a liter will cost you 3. Your fridge has a capacity of 30 liters, but for each liter of cocktail you only need half a liter of fridge space. You figure that 50 liters in total should be enough for your friends. Here’s the linear program for your problem:

\[
\begin{align*}
\text{Minimize } f(x) &= x_1 + 3x_2 \\
\text{subject to } & \\
1. & x_1 + x_2 \geq 50 \\
2. & x_1 + \frac{1}{2}x_2 \leq 30 \\
3. & x_1 \geq 0 \\
4. & x_2 \geq 0 \\
\end{align*}
\]

Figure 3.4: Linear program for throwing a party

Remarks:
- How is a linear program defined in general?

Definition 3.5 (Linear Program, LP). A linear program (LP) consists of a set of \( m \) inequalities

\[
a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n \leq b_1 \\
a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n \leq b_2 \\
\vdots \\
a_{m1}x_1 + a_{m2}x_2 + \ldots + a_{mn}x_n \leq b_m
\]

and a linear function

\[
f(x) = c_1x_1 + c_2x_2 + \ldots + c_nx_n.
\]

The \( a_{ij}, b_i, \) and \( c_i \) are given real-valued parameters and a vector \( x = (x_1, \ldots, x_n)^T \) is a solution to the linear program if \( x_i \geq 0 \) for all \( 1 \leq i \leq n \) and \( x \) maximizes \( f(x) \).

Remarks:
- If a linear program is specified as in the above definition, then we say that it is given in canonical form. There is also a short hand notation

\[
\max \{ c^T x \mid Ax \leq b, x \geq 0 \}
\]

where \( A \) is the matrix with entries \( a_{ij} \) and \( b \) and \( c \) the vectors given by the \( b_i \) and \( c_i \), respectively.

Algorithm 3.6 Simplex Algorithm
1. choose a vertex \( x \) of the polytope
2. while there is a neighboring vertex \( y \) such that \( f(y) > f(x) \) do
3. \( x \leftarrow y \)
4. end while
5. return \( x \)

Remarks:
- There are other methods for solving LPs, such as interior point methods, where a solution is approached through the interior of the polytope. While the simplex algorithm performs well in practice, there are instances where its runtime is not polynomial in \( n \). For some interior point methods it has been proved that the runtime is polynomial.

- In our party example, the solution of the LP uses fractional amounts of beer and cocktail ingredients. Sometimes fractional solutions are not possible and we need an integer solution. Solving integer linear programs is usually NP-hard.

- LPs can solve flow problems. For simplicity, we only present the LP for maximizing a single-commodity s-t-flow. The multi-commodity case is similar, with the number of inequalities growing roughly linearly with the number of commodities.
Maximize \( f(x) = \sum_{e \in \text{out}(s)} x_e \)
subject to
1. \( x_e \geq 0 \) for all \( e \in E \)
2. \( x_e \leq c(e) \) for all \( e \in E \)
3. \( \sum_{e \in \text{in}(i)} x_e = \sum_{e \in \text{out}(i)} x_e \) for all \( v \in V \setminus \{s,t\} \)
4. \( \sum_{e \in \text{in}(i)} x_e = 0 \)

Figure 3.7: LP for maximizing a single-commodity s-t-flow

Remarks:
- For each edge \( e \), \( x_e \) is a variable indicating the amount of flow on \( e \), i.e. \( x_e \) was previously called \( F(e) \).
- As our goal is to find a maximum s-t-flow, we want to maximize the function \( f(x) \) describing the amount of flow exiting \( s \). The first constraint ensures that the amount of flow is non-negative on each edge, and the second guarantees that no edge capacities are violated. The third enforces flow conservation. The fourth is required because we do not want any part of the flow leaving \( s \) to return to \( s \).
- Some networks indeed adopt centralized approaches for finding good allocations (e.g. using Software Defined Networking, SDN, or Multi-protocol Label Switching, MPLS). However, for large networks with quickly changing data flows, such as the internet, calculating and maintaining a good allocation in a centralized way is difficult. Moreover, who should do it?! There is no central authority for bandwidth allocation. We need a distributed way of avoiding congestion. We will discuss this in Section 3.5.
- So far, a flow was allowed to split up at vertices, resulting in a branched flow. In practice, we often want each flow to follow just a path.

Definition 3.8 (Unsplittable Flow). An s-t-flow \( F \) is called unsplittable if the edges \( e \in E \) with \( F(e) > 0 \) form a path from \( s \) to \( t \). If we do not impose this path restriction on a flow, it is called splittable.

Remarks:
- The notion of an unsplittable flow also extends to multi-commodity flows. If paths are not fixed, we cannot use a simple LP for maximizing an unsplittable multi-commodity flow, as the additional constraint cannot be expressed by linear inequalities.
- Maximizing an unsplittable multi-commodity flow is NP-hard, but various algorithms solve the problem approximately.

3.3 Fairness

Definition 3.9. The demand \( d_i \in \mathbb{R}_{\geq 0} \) of a flow \( F_i \) is the rate at which \( F_i \) wants to transmit. The actual flow rate is always at most as large as the demand, i.e., \( F_i \leq d_i \).

Remarks:
- Due to the capacity restrictions in our network and the presence of other flows, the rate of a flow may be considerably smaller than its demand.
- For convenience we will assume in the following that all considered flows are unsplittable and that, for each flow, we are given a designated path this flow will follow.
- A fundamental problem of managing data flows in a network is how to allocate the bandwidth of a link whose capacity is not sufficient for simultaneously accommodating all flows (at full demand) which are to be routed along this link. On one hand, it may seem reasonable to allocate the available resources in a way that throughput is maximized. On the other hand, if throughput is maximized, some flows may starve. A certain fairness is desirable.

Figure 3.10: We have three flows, all with demand 1.

Remarks:
- What is a fair bandwidth allocation in Figure 3.10? Throughput is maximized if flow \( F_2 \) is ignored and \( F_1 \) and \( F_3 \) are allocated a bandwidth of 1. A fairer allocation that still takes the throughput into account is to allocate a bandwidth of 2/3 to \( F_1 \) and \( F_3 \) each and of 1/3 to \( F_2 \). There is an argument for allocating \( F_3 \) only half of the bandwidth of \( F_1 \) and \( F_3 \) since it uses twice as many edges. If we ignore throughput completely, then allocating a bandwidth of 1/2 to each flow is simple and fair. How can we formalize this intuitive concept of fairness?

Definition 3.11 (Max-Min-Fairness). A bandwidth allocation is called max-min-fair if increasing the allocation of a flow would necessarily decrease the allocation of a smaller or equal-sized flow.
3.4 UDP

Remarks:
- There is only one max-min-fair allocation for a given set of flows in a network. It can be found by Algorithm 3.12.

Algorithm 3.12 Max-Min-Fair Allocation
1. Given a graph $G$, a set $F = \{F_1, \ldots, F_k\}$ of flows with initial rate 0 on all edges $p_1, \ldots, p_k$ along which the respective flows are to be routed and demands $d_1, \ldots, d_k$
2. while $F \neq \emptyset$ do
3. repeat
4. increase rate of all flows in $F$ evenly, but at most up to the respective demands
5. until there is an edge $e \in E$ such that $\sum_{i=1}^{k} x_{i} \leq c(e)$
6. for all such edges $e$ do
7. for all $i$ such that $e \in p_i$ do
8. $F := F \setminus \{F_i\}$
9. end for
10. $E := E \setminus \{e\}$
11. end for
12. end while

3.4 UDP

As multiple applications running on the same computer want to use a network at the same time, it is necessary to distinguish between those applications (and their respective data flows). This distinction is provided by ports.

Definition 3.13 (Port). A port is a numeric identifier used in transport protocols to identify which application sent the packet and which application should receive it on the destination computer.

Definition 3.14 (Client-Server Model). In the client-server model, the client actively initiates the communication, while the server passively waits for a client to connect. The client is regarded as a consumer of the services offered by the server.

Remarks:
- When communicating with a server, a client transmits its port so that the server knows where to reply, if needed.
- There exists a multitude of protocols used when communicating between applications, with various tradeoffs in terms of latency, security and consistency. The most common ones are UDP and TCP.

Protocol 3.15 (UDP). The user datagram protocol (UDP) is a no-frills transport protocol that allows an application to send packets from client to server.

3.5 TCP

Definition 3.16 (Connection). A connection is a bidirectional long-term relationship established between a client and a server in order to transmit data reliably.

Protocol 3.17 (TCP). The transmission control protocol (TCP) is a connection-oriented transport protocol guaranteeing that lost packets are being retransmitted and that packets are delivered in the same order they are sent.

Remarks:
- In Chapter 2 you learned that IP packets consist of header and payload. In the transport layer (when using UDP) the IP payload is divided further into the UDP header and the actual data.
- In the UDP header, the source and destination ports are specified along with a checksum (error detection) and a packet length (packets have individual sizes).
- UDP is connectionless; it simply sends packets.
- UDP does not handle packet loss.
- UDP does not provide any congestion control.
- Furthermore, UDP does not guarantee any order on the delivery of packets.
- Dealing with all these issues is delegated to the client application.
- However, UDP also has very little overhead in terms of packet size and latency, hence it is commonly used in scenarios where the application requires low overhead, e.g., real-time applications.

In the literature, the TCP packets are also called segments.

While UDP simply sends packets, TCP establishes a connection between source and destination before starting to send packets containing the actual data to be transmitted. This procedure is known as a three-way handshake.
• After establishing, a TCP connection remains alive as long as it is not terminated. Although the connection might be idle, the corresponding resources (ports, buffers etc.) are released only after terminating the connection.

Definition 3.18 (Acknowledgement). An acknowledgement (ACK) is the confirmation that a sent packet has actually been received. The ACK is sent from the receiver of the packet to the sender.

Remarks:
• In TCP, each data byte is specified by a sequence number. The sequence number of a packet is the number of the first data byte in the packet. Upon receiving a packet, the receiver sends back a packet where the acknowledgement number is set to the number of the last data byte of the received packet plus 1, i.e., the sequence number of the first byte of the packet it expects to receive next. By sending this acknowledgement packet, the receiver confirms to have received all data up to the specified byte. The acknowledgement packet may be void of any actual data.
• In addition, TCP often also supports non-cumulative acknowledgements known as selective ACKs (SACKs).

Protocol 3.19 (Establishing a Connection),

• The client sends a SYN (synchronize) packet to the server.
• The server acknowledges the packet by sending back a SYN/ACK packet.
• The client acknowledges the reception of the SYN/ACK packet by sending an ACK packet itself.

Remarks:
• The sequence number $x$ of the first SYN packet is not simply set to 0 (for security reasons), but to some arbitrary number. Based on this number the subsequent data is numbered (bytewise). The rules explained above for the used sequence and acknowledgement numbers also apply for establishing the connection: The server’s SYN/ACK packet has acknowledgement number $x + 1$ and the client’s ACK packet, containing also the first actual data, has sequence number $x + 1$.
• Terminating a connection can be done by a similar process where the SYN packets are replaced by FIN packets.
• A packet is specified as a SYN, FIN, or ACK packet by setting the respective binary flag in the header.

Definition 3.20 (Flow Control). Avoiding congestion on the recipient’s side which occurs, e.g., because the recipient is a device processing relatively slowly, is called flow control.

Remarks:
• For flow control, the receiver uses the flow window size field in the header to specify how many bytes it can receive before its buffer is full. The sender accordingly adjusts its sending rate so that the number of bytes in flight is never bigger than the flow window size.
• Finally we come to our application of multi-commodity flows: congestion control.

Definition 3.21 (Congestion Control & AIMD). Avoiding congestion in the network is called congestion control, TCP implements a control mechanism called AIMD, where AIMD stands for additive increase/multiplicative decrease.

Remarks:
• Similar to the flow windows size, the sender has a congestion window to limit the number of bytes in flight. Unlike the flow window size, the congestion window size is not specified by the receiver but by the sender itself.
• The congestion window size is controlled with AIMD. In particular, the rate of any flow continuously changes as follows: If a congestion is reported, each flow repeatedly increases its rate additively. When congestion occurs on some edge, the affected flows decrease their rates by a multiplicative factor. The function describing the rate of a flow thus roughly follows a sawtooth behaviour.
• To be precise, in wired networks, congestion occurs in a node (router), and not on an edge (link). When the router’s buffers are full while data packets come in, those packets are dropped and packet loss occurs. Such a packet loss is used as indicator that the affected flow has to perform a multiplicative decrease.
• In TCP, the sender decreases the congestion window size by a factor of two. Afterwards, the congestion window is additively increased by one packet per RTT.

Definition 3.22 (Round-Trip Time). The time it takes a packet to travel from sender to receiver and back is called round-trip time (RTT).

Remarks:
• In TCP, recognition of dropped packets on the sender side is implemented by timeouts, i.e., if a packet is not acknowledged in some time frame it is considered to be lost. Thus, some time elapses between a congested router dropping a packet and the affected flow decreasing its rate which in turn causes other flows to suffer packet loss in the congested router since the congestion is not remedied immediately.
• How long should a sender wait for an acknowledgment? This waiting time depends on the RTT, more precisely TCP uses a variable called smoothed RTT set initially to the RTT of the first acknowledged packet. The new smoothed RTT is the weighted (‘smoothed’) mean of itself and the RTT of the last acknowledged packet.
Lemma 3.23 (AIMD Fairness). If the bandwidth allocation is performed according to AIMD, then it roughly converges to a max-min-fair allocation (Definition 3.11).

Proof. First note that in AIMD the allocation never reaches a stable state. Now consider what happens if a router used by two flows becomes congested: If both flows drop packets, then their rates are decreased by the same factor (in TCP: halved), and the absolute difference between the two rates decreases. The subsequent additive increase does not change this difference and when the next congestion occurs on this link, the rates converge again. It is possible that only one flow drops a packet during congestion, but this also helps the smaller flow, since the probability of packet loss is larger for the larger flow.

Remarks:

• Over time, various heuristics have been incorporated into TCP to improve performance, e.g., the slow-start algorithm which governs the initial growth of the size of the congestion window. According to slow-start, whenever a packet is acknowledged, the window size is increased by one packet. Thus, the initially small window grows exponentially in size until a certain threshold is reached upon which the additive increase part of AIMD starts. This reduces the time to reach a reasonable bandwidth after a new connection has started.

• TCP relies on the goodwill of the senders as this is where the adjustment of the flow rates takes place. You may tweak your local version of the TCP protocol in order to obtain more bandwidth for yourself, e.g., by simply ignoring the multiplicative decrease.

3.6 NAT

Definition 3.24 (Network Address Translation, NAT). A node systematically exchanges the header of packets in order to be able to route to nodes with private addresses.

Remarks:

• Because of the shortage of IPv4 addresses, ISPs do not want to give many IPv4 addresses to their customers, often each customer gets exactly one IPv4 address. Instead, in a home or a small business, all machines but the entry node (router) only get private addresses.

• The address blocks 10.0.0.0/8, 172.16.0.0/12, and 192.168.0.0/16 are reserved for private networks. In other words, addresses of these blocks are not unique as promised in Definition 2.20, but many nodes may have the same address. Nodes outside the private network cannot route to such a private address.

• We have a client node with a private address in a network with a router, and a server. While the client can easily send a packet with a search query to the server, how does the server send back its answer? When the client packet $p$ arrives at the router, the router will switch the client’s private IPv4 address and port with its own router IPv4 address and an arbitrary unused port. The router forwards that modified packet $p'$ to the server, and memorizes the triple (new unused port, client address, and client port). When the server’s answer comes back to the router, the router will switch back the destination address/port to the client’s address/port, before the router forwards the packet to the client.

• This is a nasty hack because it mixes concepts of the network layer (addresses) and the transport layer (ports).

Chapter Notes

As Leighton and Rao show in [4], for multi-commodity flows, the size of the maximum flow does not equal the size of the minimum cut in general. The NP-hardness of maximizing an unsplitable multi-commodity flow can be inferred from [1].

Two of the first researchers who formulated applied problems from logistics/economics as linear programs were Kantorovich and Koopmans who later received the Nobel Prize in economics for their contributions. The simplex algorithm was developed by Dantzig in 1947. In 1979, Khachiyan showed that linear programs can be solved in polynomial time. In 1984, Karmarkar developed an interior point algorithm that not only had a polynomial-time runtime, but was also practically feasible.

TCP was developed by Cerf and Kahn, based on their work [2]. Analysis of the AIMD algorithm can be found in [3].

This chapter was written in collaboration with Sebastian Brandt.

Bibliography