Chair of Distributed Systems and Security
School of Computation, Information and Technology
Technical University of Munich

Computer Networking and IT Security (INHN0012)

Tutorial 1

Problem 1 Binary prefixes

The difference between binary and SI prefixes always causes confusion. The problem consists in contradictory information, especially on the part of the operating systems: The memory allocation on mass storage devices such as hard disks is commonly counted with binary prefixes while the stated units contain SI prefixes.
An example: you are buying a hard disk with a manufacturer's stated capacity of 3 TB. In the small print on the packaging you will find the note „1 TB = 10^{12} B“. Obviuosly, SI prefixes are used. Now assume that your operating system counts the storage using binary prefixes.

SI prefix		Value	Binary prefix		Value
k	(kilo)	10^3	Ki	(Kibi)	2^{10}
M	(Mega)	10^6	Mi	(Mebi)	2^{20}
G	(Giga)	10^9	Gi	(Gibi)	2^{30}
T	(Tera)	10^{12}	Ti	(Tebi)	2^{40}
P	(Peta)	10^{15}	Pi	(Pebi)	2^{50}

Table 1.1: Comparison between SI and binary prefixes

a)* State the capacity of your hard disk in TiB.

$$3\,\text{TB} = 3 \cdot 10^{12}\,\text{B} = \frac{3 \cdot 10^{12}}{2^{40}}\,\text{TiB} \approx 2.73\,\text{TiB}$$

b)* Determine the percentage difference between SI and binary prefixes for all values given in Table 1.1.

$$\frac{\text{k}}{\text{Ki}} = \frac{10^3}{2^{10}} \approx 97.66\,\% \quad \Rightarrow \quad e = 2.34\,\%$$
$$\frac{\text{M}}{\text{Mi}} = \frac{10^6}{2^{20}} \approx 95.37\,\% \quad \Rightarrow \quad e = 4.63\,\%$$
$$\frac{\text{G}}{\text{Gi}} = \frac{10^9}{2^{30}} \approx 93.13\,\% \quad \Rightarrow \quad e = 6.87\,\%$$
$$\frac{\text{T}}{\text{Ti}} = \frac{10^{12}}{2^{40}} \approx 90.95\,\% \quad \Rightarrow \quad e = 9.05\,\%$$
$$\frac{\text{P}}{\text{Pi}} = \frac{10^{15}}{2^{50}} \approx 88.82\,\% \quad \Rightarrow \quad e = 11.18\,\%$$

Note: The specification of binary prefixes is only usual for byte values. Bit values, e. g. kbit or Mbit, are specified with SI prefixes only.

Problem 2 Sneakernet

BigCorp™ is moving its data to the cloud as it has outgrown its data centre in Akron, Ohio. In order to do so, it has to move all of its data to Amazing Web Services (AWS), the cloud provider it chose to go with. Moving the total of 91 PB is no easy task. Therefore, BigCorp™ has chosen to shut down all of its services for as long as necessary, move the data and finally restart all services in the cloud.

a)* How long does it take to transfer the data using BigCorp's high-speed 8 Gbit/s internet uplink? As for now, ignore the time it takes for the serialized data to propagate to the cloud location. Compare the given uplink with your uplink at home.

$$t = \frac{91\,\text{PB}}{8\,\text{Gbit/s}} = \frac{728\,\text{Pbit}}{8\,\text{Gbit/s}} = \frac{728 \cdot 10^6\,\text{Gbit}}{8\,\text{Gbit/s}} = 91 \cdot 10^6\,\text{s} \approx 25\,277.78\,\text{h} \approx 1053.24\,\text{d} \approx 2.89\,\text{yr}$$

Seeing these numbers, BigCorp™ realizes that it clearly cannot shut down its services for so long without going out of business. AWS offers a product called Snowmobile: a truck, able to carry 100 PB, is sent to the customer's on-premise location. Once arrived, the customer transfers the data onto the storage medium inside the truck. Finally, the Snowmobile is sent to the nearest AWS location and the data is transferred onto AWS' systems.

b) Assume the infrastructure and Snowmobile allow loading and unloading data via eight parallel 400 Gbit/s fibre connections. What is the total data rate? How long does it take to load or unload all of BigCorp's data to and from the Snowmobile?

$$r_{load} = r_{unload} = 400\,\text{Gbit/s} \cdot 8 = 3200\,\text{Gbit/s} = 3.2\,\text{Tbit/s}$$

$$t_{load} = t_{unload} = \frac{91\,\text{PB}}{3.2\,\text{Tbit/s}} = \frac{728 \cdot 10^3\,\text{Tbit}}{3.2\,\text{Tbit/s}} = 227.5 \cdot 10^3\,\text{s} \approx 2.63\,\text{d}$$

Assume that the distance between BigCorp's datacentre and AWS' data centre is 186 km. The truck travels at an average speed of 93 km/h.

c) For this subtask, assume the entire 100 PB capacity of the truck as utilized. Given the capacity and travel time, what data rate in Tbit/s does the truck achieve? For this subtask, ignore the time it takes to load and unload the truck.

$$t_{travel} = \frac{186\,\text{km}}{93\,\text{km/h}} = 2\,\text{h} = 7200\,\text{s}$$

$$r_{fulltruck} = \frac{100\,\text{PB}}{7200\,\text{s}} = \frac{800 \cdot 10^3\,\text{Tbit}}{7200\,\text{s}} \approx 111.11\,\text{Tbit/s}$$

d) Now consider the amount of data BigCorp™ actually has, as well as the time it takes to load and unload the truck. How long does the entire data transfer take? Which data rate in Gbit/s is achieved overall?

$$t_{all} = t_{load} + t_{travel} + t_{unload} = 227.5 \cdot 10^3\,\text{s} + 7200\,\text{s} + 227.5 \cdot 10^3\,\text{s} = 462.2 \cdot 10^3\,\text{s} \approx 5.35\,\text{d}$$

$$r_{overall} = \frac{91\,\text{PB}}{462.2 \cdot 10^3\,\text{s}} = \frac{728 \cdot 10^3\,\text{Tbit}}{462.2 \cdot 10^3\,\text{s}} = 1575.03\,\text{Gbit/s}$$

e) Assume that BigCorp™ still cannot shut down its business for that long. Which alternative does it have to moving these vast amounts of data at once. Consider how real networks function and try applying it to this scenario. Sketch how BigCorp™ can reduce the downtime of the entire system.

BigCorp™ can split its data into smaller parts, given that it can migrate parts of its system at a time (e. g. one service at a time). Splitting the data into smaller chunk allows cutting down on the time needed to load and unload the truck, which is the majority of time needed. Therefore, smaller chunks can be transferred quicker than all data at once, allowing to migrate one service at a time, causing multiple shorter downtimes. This allows to e.g. move the entire data piece by piece over multiple weekends.

"Never underestimate the bandwidth of a station wagon full of tapes hurtling down the highway."
— Andrew S. Tanenbaum

Problem 3 Source entropy

As we have seen in the lecture, a message source can be abstractly modeled as follows:

- A message source Q emits statistically independent characters from an alphabet $\mathcal{X}$.
- The source Q is assumed to be memoryless, i. e., the output in step n does not depend on the output of any previous step.
- The emitted symbols carry a different amount of information, depending on their probability of occurrence – in general, the less likely a symbol is being emitted, the higher the information content.
- Information is measured in bit

Assume that a source Q emits characters of the alphabet $\mathcal{X} = \{a, b\}$. We model this message source as a discrete random variable X. The probability that the source emits the character $X = a$ is $p_a = \Pr[X = a] = 0.25$.

a)* Determine the probability p_b that the character $X = b$ is emitted.

Since $p_a + p_b = 1$, it follows that $p_b = 0.75$.

b) Determine the information content $I(a)$ and $I(b)$ of both symbols.

$$I(a) = -\log_2 p_a = 2.00\,\text{bit}$$
$$I(b) = -\log_2 p_b \approx 0.42\,\text{bit}$$

c) Determine the entropy H of the source.

$$H(X) = \sum_{x \in \mathcal{X}} p_x I(x) \approx 0.81\,\text{bit/symbol}$$

d) Determine the occurrence probabilities p_0 and p_1 of another binary message source Q' such that its entropy H is maximal.

First we express p_1 in dependence of p_0 and write $p_1 = 1 - p_0$. For simplicity we write $p_0 = p$. Finally, the entropy H can be expressed as a function of p and the probability we are looking for can be determined by means of the derivative.

$$H = -p \log_2(p) - (1-p) \log_2(1-p)$$

$$\frac{dH}{dp} = -\log_2(p) - \frac{p}{p \ln(2)} + \log_2(1-p) + \frac{1-p}{(1-p)\ln(2)}$$

$$\Rightarrow \log_2(p) + \frac{p}{p \ln(2)} \stackrel{!}{=} \log_2(1-p) + \frac{1-p}{(1-p)\ln(2)}$$

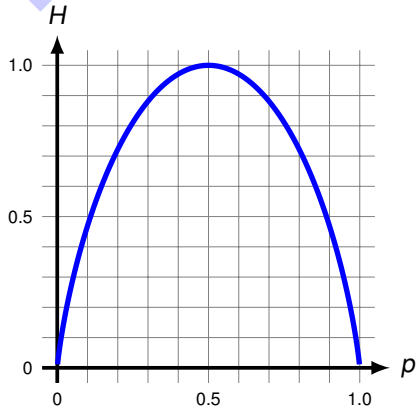
Comparing both sides yields $p = 1 - p = 1/2$.

e) What is the maximum entropy of a binary source?

The entropy is maximised when $\Pr[X = a] = \Pr[X = b] = 0.5$ holds. The maximum entropy is therefore

$$H_{\max} = -2 \cdot 0.5 \cdot \log_2(0.5) = 1 \text{ bit/symbol}.$$

f) Sketch the source entropy H of a binary source in general as a function of the probability of occurrence p.



g) Obviously, the entropy $H(X) < 1$ is not maximal. What conclusion can be drawn from this fact for the data stream emitted by source Q with respect to redundancy?

The string emitted by Q, which is nothing but different instantiations of the random variable X, contains redundancy. The data stream generated by Q is therefore representable on average with less than 1 bit/ symbol.

h) Generalise the results of the subtasks d) and e) to a source emitting N different symbols.

In general, the following applies to entropy

$$H(X) = \sum_{x \in \mathcal{X}} I(x) p_i.$$

With the requirement $p_i = p$, i.e. all characters occur with the same probability, it immediately follows $p = 1/N$ and thus

$$H = \sum_{x \in \mathcal{X}} I(x) p = -\sum_{i=1}^{N} \log_2 \left(\frac{1}{N} \right) \frac{1}{N} = \log_2(N).$$