

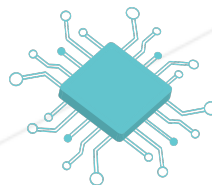


APCSP - When Data Gets Big

Navigating Big Data in Space and Time

Kai kai@42.us.org

Summary: This PDF will cover topics around time complexity, unreasonable running time, unsolvable problems, heuristics, distributed computing, and data compression.



HACK
HIGH
SCHOOL



This work is licensed under a [Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License](https://creativecommons.org/licenses/by-nc-sa/4.0/).

Contents

I	Introduction	2
II	Measuring Storage Space	4
II.1	Binary	4
II.2	Bits and Bytes	5
II.3	Moore's Law	6
III	Time Complexity	7
III.1	Clock Speed is Measured in Hertz	7
III.2	Counting Steps	7
III.3	Reasonable vs Unreasonable	7
IV	Unsolvable & Undecidable Problems	8
IV.1	Mathematical Proofs	8
IV.2	Heuristics	8
V	Distributed Computing	9
VI	Compression	10

Chapter I

Introduction

What makes Big Data such an important topic?



As we interact with each other daily on the internet for social, educational, economic, governmental, and health-related purposes, we leave behind a data footprint. Before the internet became mainstream in the 1980s, this possibility didn't exist. A private eye could follow one person around to find out what they are up to, but if you wanted to research what one million customers think about your brand of clothing, you would have to mail them all paper surveys to fill out, or call them on the phone! And then hire someone to type up and analyze the results by hand!

In today's economy, data collection is relatively easy. Now, researchers around the world are focusing their effort to develop new algorithms that teach computers how to use this digital information to make predictions about the future.

Part of the reason Big Data is such a hot topic is because it is still difficult to work with. Large amounts of data can be messy or inaccurate, they take up a lot of space, and the calculations take a long time to finish. But if the answers are anywhere close to correct then they can be very useful!

For questions that follow the pattern "find the best / shortest / fastest solution...", known as **optimization problems**, we often use an approximation of the right answer instead of waiting around to get the perfect answer. The topics in the PDF below will help you understand the boundary between problems that do have an exact answer, and problems where a 'good-enough' answer is actually the best we can do within current technology.

Chapter II

Measuring Storage Space

II.1 Binary


All data is stored in binary format. In the physical world, binary can be represented as anything that has two distinct states, basically "on" and "off".

- On CDs and DVDs, if you look under a microscope, you will see tiny bumps which represent the binary "1". The flat parts represent "0".
- On a CD-ROM, instead of bumps and flat parts there are shiny dots and non-shiny dots. The computer reads "1" if a laser reflects off the disk and "0" if the laser does not reflect.
- Data can be sent over copper wire. In that context, a "1" is the presence of an electrical charge (electrons travelling through the wire) and a "0" is the charge being turned off.
- Data can be sent through optical cables. Inside the cables, a "1" is represented by a flash of light that travels down a tiny tube, and a "0" is a flash of darkness.
- Data can be sent over radio waves. In this context, there are multiple ways of encoding "1" and "0". AM radio uses differences of **amplitude** to encode the binary. FM radio uses differences of **frequency** to encode the binary. Don't worry if you don't understand the physics of it, but remember that binary data can be carried by radio waves.
- Computer memory is made out of **transistors**, which are a type of electrical circuit which can hold an electrical charge. The transistor either holds a high charge ("1") or a lower charge ("0").

II.2 Bits and Bytes

Humans have a lot of trouble reading 1's and 0's directly. If binary numbers are printed on a page, our eyes tend to glaze over and skip from one line to another. To save space when printing out raw data, computer scientists convert the binary representations to other number bases such as **octal** and **hexadecimal**.

In addition, instead of talking about data in units of "bits", we talk about it in units of "bytes". One byte is a group of 8 bits.




APCSP/images/bitbyte2.png

Most of the time in our daily lives we deal with information amounts on the level of 1,000 (one thousand) to 1,000,000,000,000 (one trillion) bytes.

- **Kilobyte** - 1,000 bytes - 1 thousand bytes - Size of a small image or text file
- **Megabyte** - 1,000,000 bytes - 1 million bytes - Size of a large image or text file
- **Gigabyte** - 1,000,000,000 bytes - 1 billion bytes - Size of a movie
- **Terabyte** - 1,000,000,000,000 bytes - 1 trillion bytes - Size of a really useful hard drive that can back up all your data.

However, the measures go up much higher. A large data center could hold perhaps a trillion *terabytes* - calling for the rarely used word "**yottabyte**".



APCSP/images/bytesizes.png

II.3 Moore's Law

The field of computer science has long been blessed with constant improvements in the engineering of computer hardware. The size of the machines needed to store and process data have gone down over time, from a basic computer the size of a large classroom, to a tiny microchip that fits on your smartwatch.

Do a little bit of research on Moore's law and cite some articles about it in your homework sheet. To what extent do you think that computer components will continue to shrink in size, relative to their power, over time?

Chapter III

Time Complexity

III.1 Clock Speed is Measured in Hertz

Every computer is governed by an electronic clock that regulates the speed of the rest of the system. The clock sends pulses of electricity at a constant frequency through the system.

As a unit of frequency, clock speed is measured in Hertz. At the beginning stages of computer technology the clock speeds were measured in megahertz (MHz, millions of cycles per second). In recent years, computer systems have a top speed of 4 to 5 gigahertz (GHz, billions of cycles per second) - although in many systems it's more efficient to use a lower clock speed due to physical limitations of the hardware.

III.2 Counting Steps

III.3 Reasonable vs Unreasonable

Chapter IV

Unsolvable & Undecidable Problems

IV.1 Mathematical Proofs

IV.2 Heuristics

Chapter V

Distributed Computing

Chapter VI

Compression