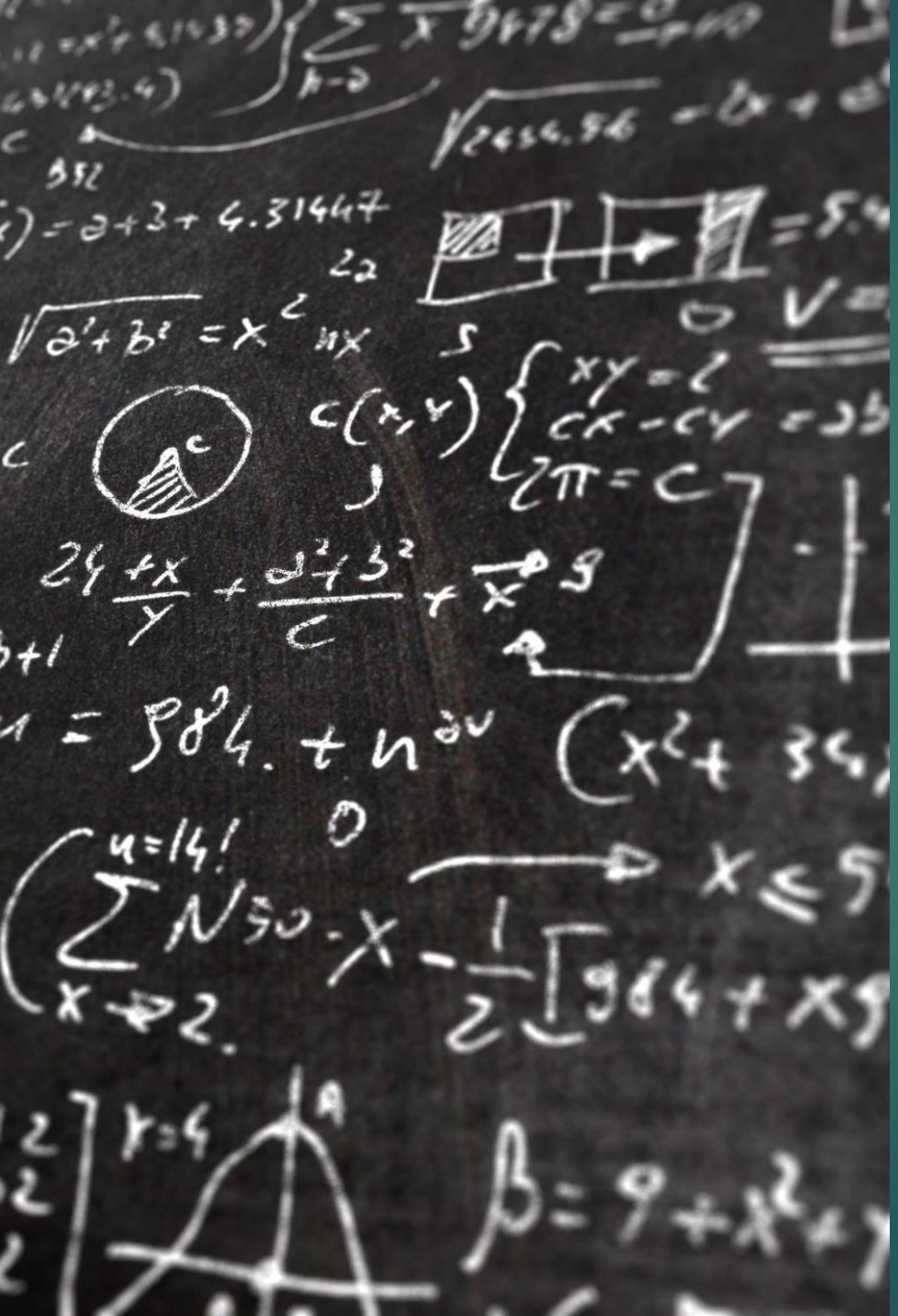
Cloud Computing Theory and Practice

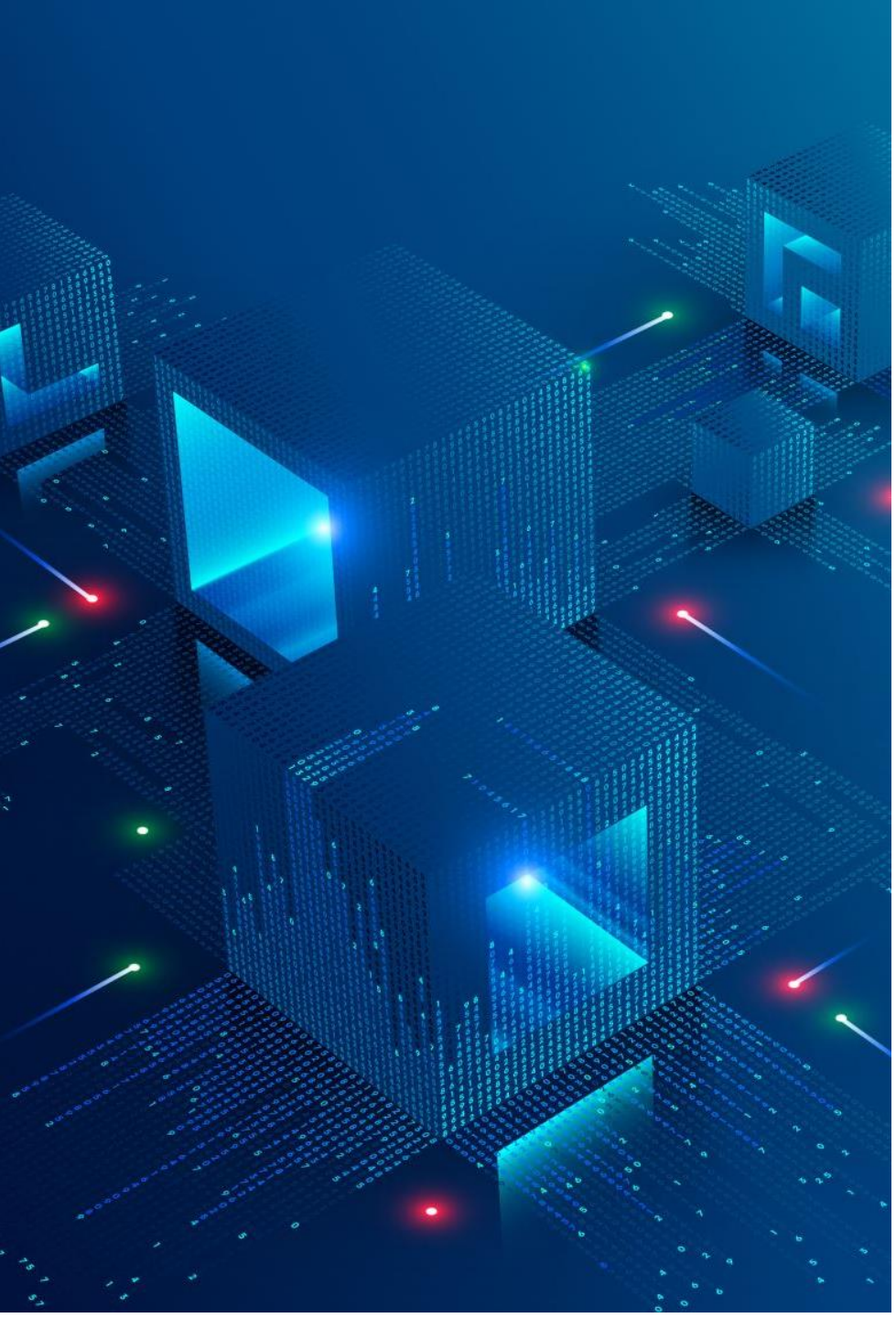
INSY 5345 & INSY 4307

DR. SANTOSO BUDIMAN



Computer Basics

SECTION 2



Section 2

- Computer
  - Computer History
  - CPU
  - Disk
  - Operating Systems
  - File System
- Network
  - Network Infrastructure
  - TCP/IP
- Virtualization
- Hypervisor
- AWS account
- AWS Cost monitoring
- Assignment 1

Computer History

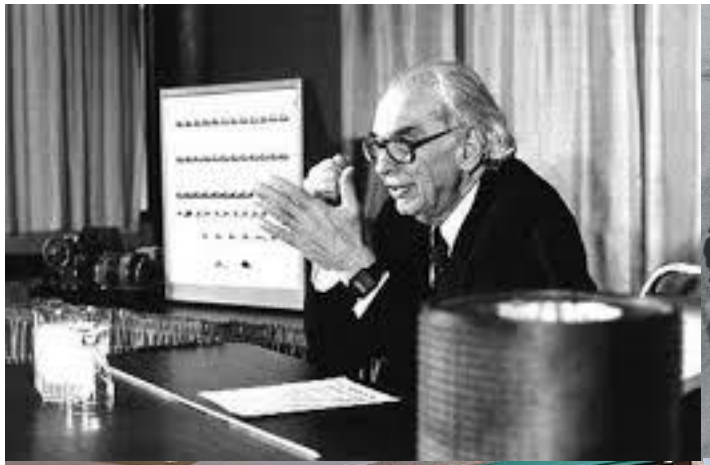
The First Electronic Digital Computer

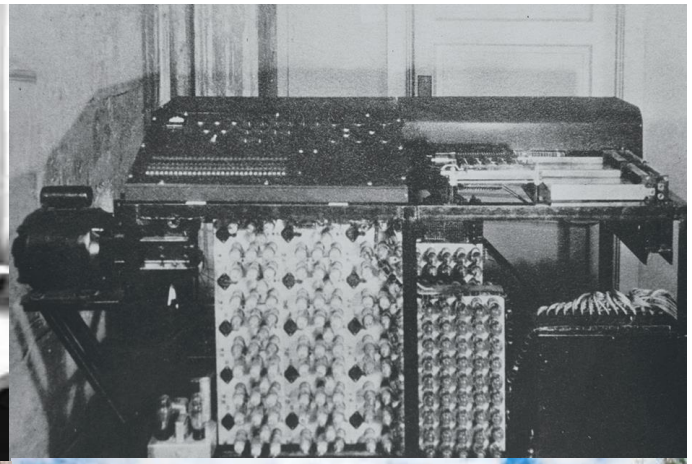
1937

Dr. John Vincent Atanasoff

(October 4, 1903 – June 15, 1995)

Iowa State University

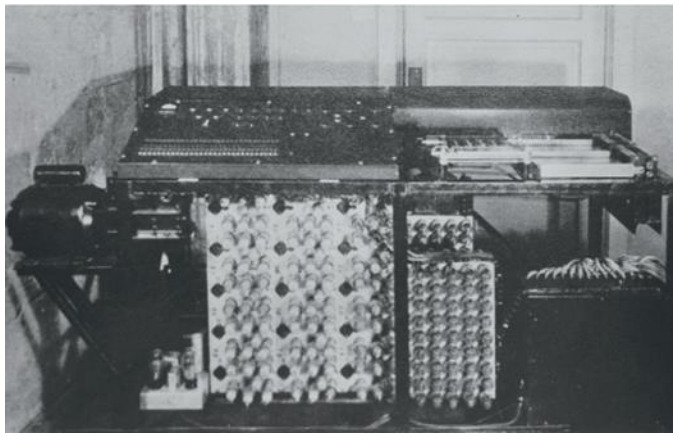




http://jva.cs.iastate.edu/operation.php

https://archive.inside.iastate.edu/2010/0325/abc.php

Computer Evolution



	ABC (1937)	Dell XPS 15
Speed	60 Hz	8th Generation Intel® Core™ i7-8750H Processor (4.1GHz)
Weight	1 ton = 2,000 lbs	4 lbs

Microchip Technology evolution

Moore's law is the observation that the number of transistors in a dense integrated circuit doubles about every two years.

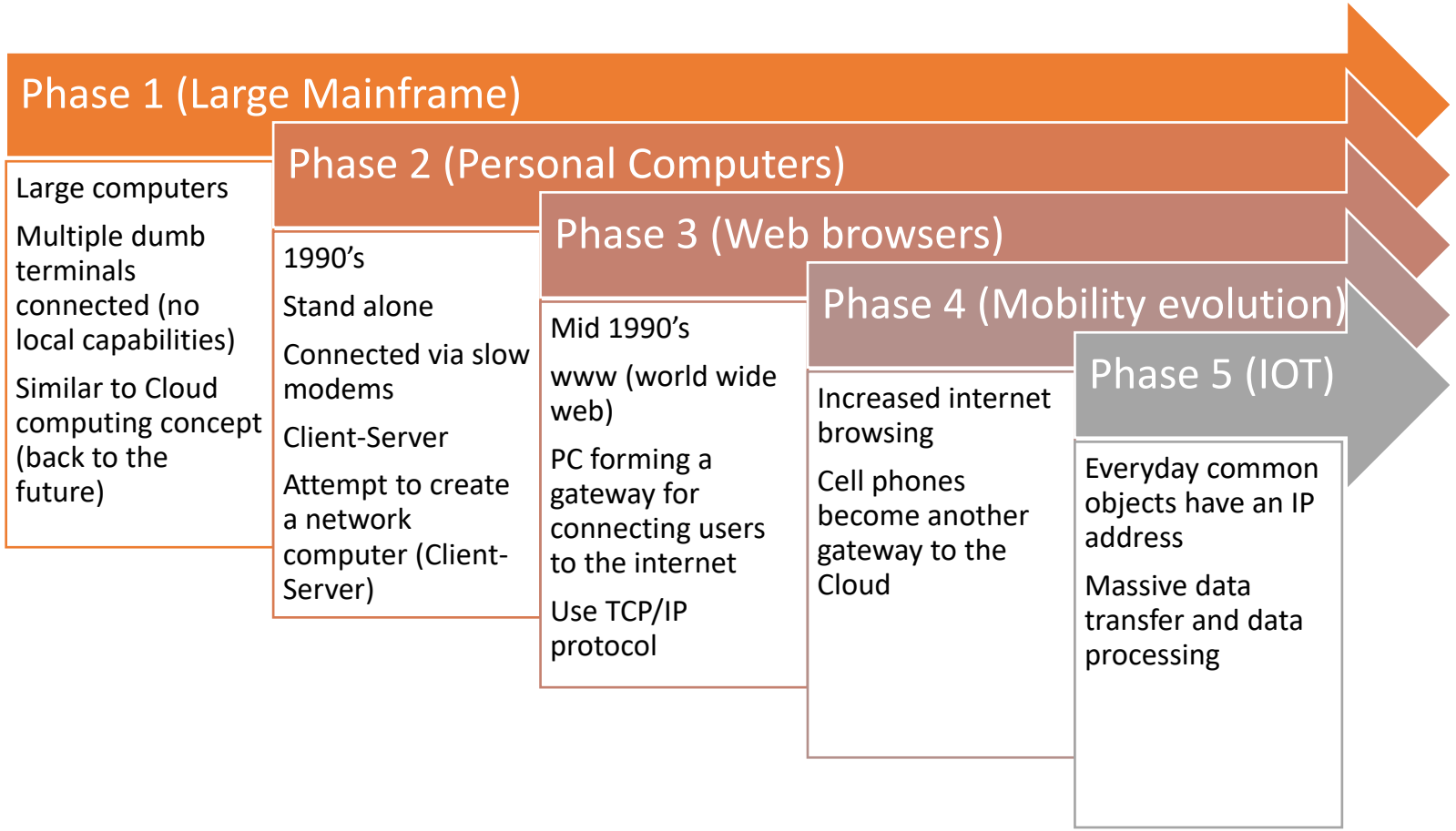
Processor	Transistor count	Date of introduction	Designer	Area
Intel 4004	2,300	1971	Intel	12 mm²
Core i7 (Quad)	731,000,000	2008	Intel	263 mm²
10-core Core i7 Broadwell-E	3,200,000,000	2016	Intel	246 mm²

While price becomes cheaper:

	Price per Million Transistors	Price of Storage Capacity per GB
1995	$110.73	$782
2005	$0.98	$0.58
2015	$0.58	$0.03

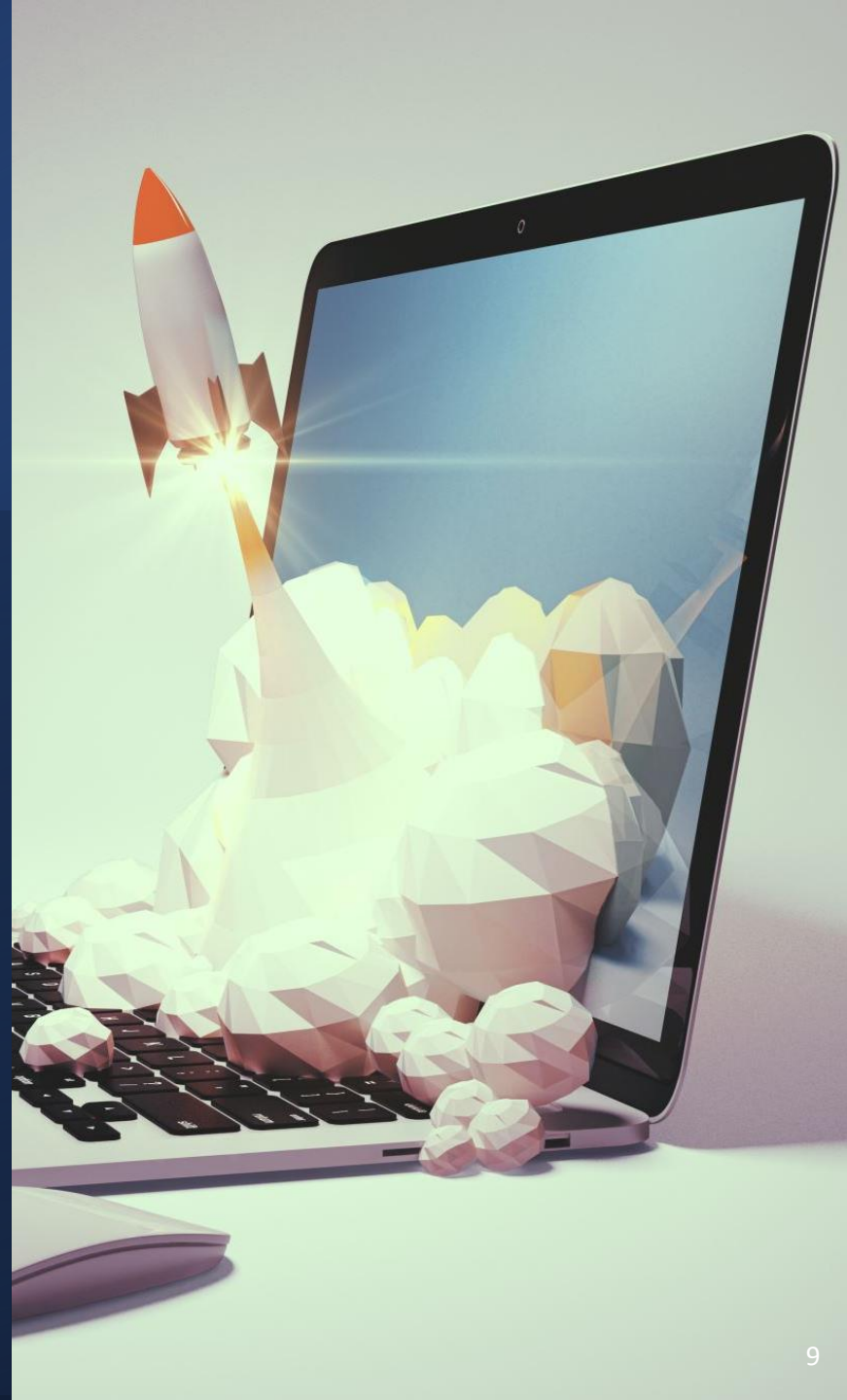
Computer processors become more powerful, smaller, and cheaper which enable fast processing of huge data. Enables development of new digital technologies

Computer Technology Evolution

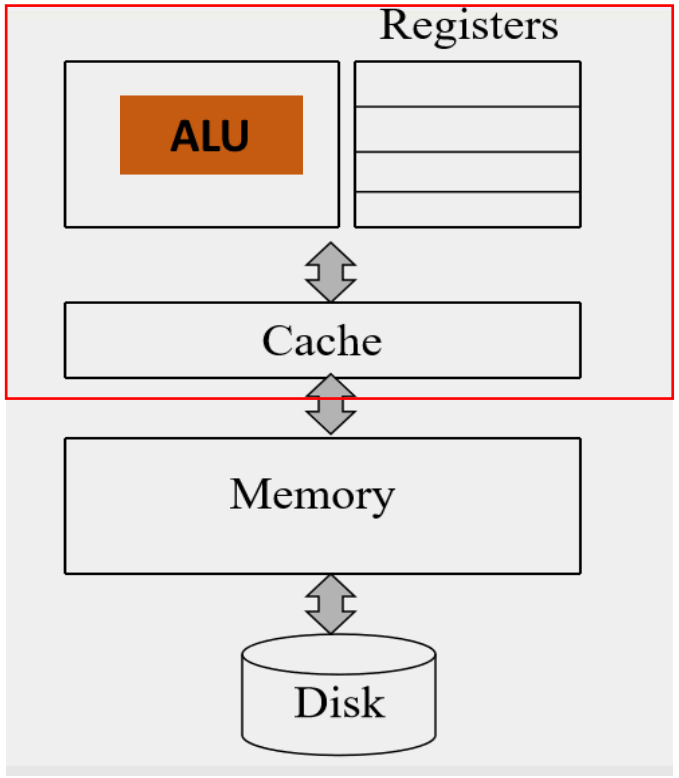


Central Processing Unit

Computer



Basic Computer Architecture



3 Basic components:

- ❑ **CPU**
- ❑ **Memory**
- ❑ **Disk**

- A program written in a high-level language (C++, Java, etc.) gets compiled to low-level machine instructions – Stored in file system on disk
- CPU loads instructions in batches into memory (and cache, and registers)
- It executes each instruction
- CPU loads data for instruction into memory (and cache, and registers) – And does any necessary stores into memory

Computer Configuration Examples

	Inspiron 15 5000	Inspiron 15 7000 2-in-1	Inspiron 14 7000	New Inspiron 15 5000 2-in-
CPU	Up to 8th Generation Intel® Core™ i7-8550U Processor (8MB Cache, up to 4.0 GHz)	Up to 8th Generation Intel® Core™ i5-8250U Processor (6MB Cache, up to 3.4 GHz)	Up to 8th Generation Intel® Core™ i7-8550U processor (8MB Cache, up to 4.0 GHz)	Up to 8th Generation Intel® Core™ i7-8565U Processor (8MB Cache, up to 4.6 GHz)
	Up to Windows 10 Home 64-bit English	Up to Windows 10 Home 64-bit English	Up to Windows 10 Home 64-bit English	Up to Windows 10 Home 64-bit English
Disk	Up to Dual drives with 128GB Solid State Drive+ 1TB 5400 rpm Hard Drive	Up to 256GB Solid State Drive	Up to Dual drives with 128GB PCIe NVMe SSD + 1TB 5400 rpm Hard Drive	Up to 512GB M.2 PCIe NVMe Solid State Drive
	Up to AMD Radeon 530 Graphics with 4G GDDR5 graphics memory	Up to Intel® UHD Graphics 620 with shared graphic memory	Up to NVIDIA® GeForce® MX150 with 2GB GDDR5 graphics memory	Up to Intel® UHD Graphics 620 with shared graphics memory
Memory	Up to 8GB 8GBx1 DDR4 2400MHz Single Channel	Up to 8GB, DDR4, 2400MHz; up to 16GB (additional memory sold separately)	Up to 16GB, DDR4, 2400MHz; up to 16GB (additional memory sold separately)	Up to 16GB, 16GBx1, DDR4 2666MHz

https://www.dell.com/en-us/shop/dell-laptops/inspiron-15-5000/spd/inspiron-15-5570-laptop?view=configurations

CPU

A core is a single processing unit, multi-core processor has multiple processing units.

Multi-core processors are created because technologically it became increasingly difficult to increase clock speed on single core processors.

A six-core 3.0GHz processor has six processing units each with a clock speed of 3.0GHz.

The six-core processor above has a total clock speed of 18.0GHz.

CPU for Virtualization

General virtualization rule: more core is better than higher-speed processors (more VMs)

All processors manufactured after 2003 have hardware assisted virtualization built in

For example: Choose between

- Two 6-core processors @ 2.2 GHz
- Two 4-core processors @ 3.3 GHz

Two 6-core processors is a better choice

- Virtualization can be spread over more cores (faster and more consistent performance)
- Note both have the same total speed 2x6x2.2 GHz (26.4) = 2x4x3.3 GHz (26.4 GHz)

Intel i7

Intel® Core™ i7-9850H Processor (12M Cache, Up to 4.60 GHz)

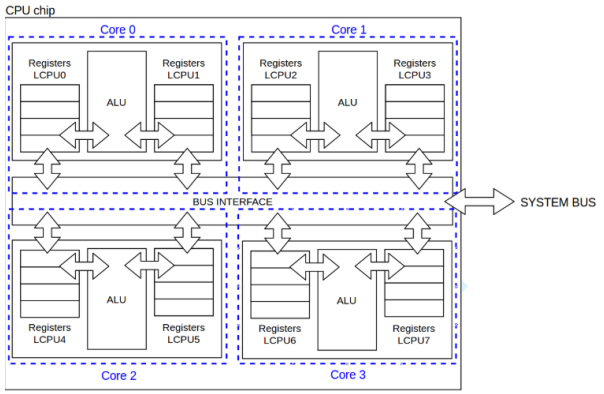
- 12 MB Cache
- 6 Cores
- 12 Threads
- 4.60 GHz Max Turbo Frequency
- H - High performance graphics
- 9th Generation

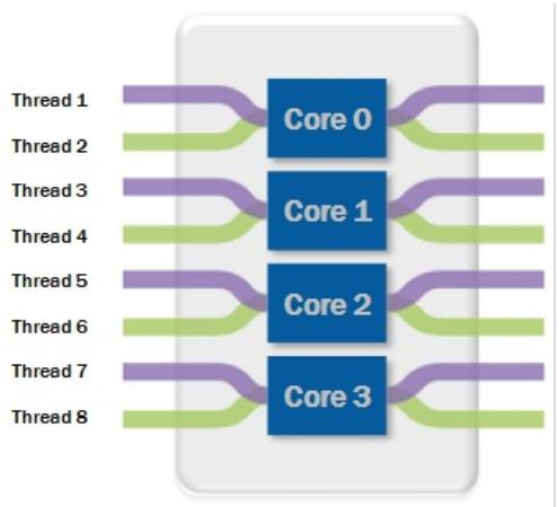
Hyperthreading

- **Hyperthreading** – technology that creates **2 logical CPUs** for **each CPU Core** that support Hyperthreading.
  - Intel Proprietary technology introduced in 2002
  - Increase parallel processes.
  - Hyperthreading is not the same as having multiple cores.
    - Hyperthreading shares the CPU ‘s pipeline, cache, and system bus interface instead of dedicated cache ad interfaces to distinct cores.
- Hyperthreading **allows higher over commitment ratios** of **vCPUs to physical CPUs**
- Hypervisor with few processor cores will see greater advantage from hyperthreading (it sees the logical CPUs (threads) not physical ones.

Central Processing Unit

- Quad-Core CPU (4 cores).
- a ‘thread’ is a logical core.
- In this example, the **OS sees 8** “logical” **processors.**
- Frequency – clock speed – the higher the faster





Intel® Core™ i7-8550U Processor 8M Cache, up to 4.00 GHz

https://ark.intel.com/content/www/us/en/ark/products/122589/intel-core-i7-8550u-processor-8m-cache-up-to-4-00-ghz.html

Performance	
# of Cores	4
# of Threads	8
Processor Base Frequency	1.80 GHz
Max Turbo Frequency	4.00 GHz
Cache	8 MB SmartCache
Bus Speed	4 GT/s OPI
TDP	15 W
Configurable TDP-up Frequency	2.00 GHz
Configurable TDP-up	25 W
Configurable TDP-down Frequency	800 MHz
Configurable TDP-down	10 W

- **Core** is a **hardware** term that describes the number of **independent central processing units (CPUs)** in a single computing component.
- **Thread (**thread of execution) is a **software term** for the basic ordered sequence of instructions that can be passed through or processed by a single CPU core. **Logical CPU**.
- CPU Cache is an area of fast memory located in the processor.

Semiconductor Manufacturers

Rank	2020[25]	2018[26]	2017[26]	2011[27]	2006[28]	2000[28]	1995[28]
1	Intel	Samsung	Samsung	Intel	Intel	Intel	Intel
2	Samsung	Intel	Intel	Samsung	Samsung	Toshiba	NEC
3	TSMC	SK Hynix	TSMC	TSMC	TI	NEC	Toshiba
4	SK Hynix	TSMC	SK Hynix	TI	Toshiba	Samsung	Hitachi
5	Micron	Micron	Micron	Toshiba	ST	TI	Motorola
6	Qualcomm	Broadcom	Broadcom	Renesas	Renesas	Motorola	Samsung
7	Broadcom	Qualcomm	Qualcomm	Qualcomm	Hynix	ST	TI
8	Nvidia	Toshiba	TI	ST	Freescale	Hitachi	IBM
9	TI	TI	Toshiba	Hynix	NXP	Infineon	Mitsubishi
10	Infineon	Nvidia	Nvidia	Micron	NEC	Philips	Hyundai

https://en.wikipedia.org/wiki/Semiconductor_industry

Name	Country
Samsung Electronics	South Korea
Intel	United States
TSMC	Taiwan
SK Hynix[a]	South Korea
Micron[b]	United States
Qualcomm	United States
Broadcom	United States
Toshiba	Japan
Texas Instruments (TI)	United States
Analog Devices	United States
Microchip	United States
NXP	Netherlands/United States
MediaTek	Taiwan
Infineon	Germany
STMicroelectronics	France/Italy
Sony	Japan
ARM	United Kingdom/United States
AMD	United States
Nvidia	United States
Renesas[c]	Japan
GlobalFoundries[d]	United States

Computer - Disk

Disk Storage Systems

- **Disk storage** is a **generic term** to describe **storage mechanisms** where data is digitally recorded by various electronic, magnetic, optical, or mechanical methods on a rotating disk, or media.
  - A disk drive is a device that uses this storage mechanism with fixed or removable media.
  - Removable media refers to an optical disc, memory card, flash media, USB drive

	Solid State Drive (SSD)	Hard Disk Drive (HDD)
Startup time	Almost instantaneous	Disk spin-up takes seconds
Noise	None	Varies between models
Susceptibility to Failure	Resistant to shock and vibrations	Susceptible to shock and vibrations
Power consumption	Average half of HDD	0.35 – 20 watts depending on size
Cost	More expensive	Less expensive
Data transfer rate	Consistent	Slower response time

Solid State Drive (SSD)

- SSD – high performance storage device that contains no moving parts.
- It includes:
  - Dynamic Random-Access Memory (DRAM) or Flash memory boards
  - CPU
  - Memory bus
- SSDs produce the highest possible I/O rates because they contain CPUs to manage data storage.
- Faster access time and lower latency than HDDs.
- SSDs and HDDs have the same I/O interface
  - allowing SSDs to easily replace HDDs without changing computer hardware

Disk Interface

Advanced Technology Attachment (ATA/PATA) – an interface standard for connecting storage devices within computers

Integrated Drive Electronics (IDE) – the integration of the controller and the hard drive itself.

Serial ATA (SATA) – used to connect host bus adapters to mass storage devices. Designed to replace ATA.

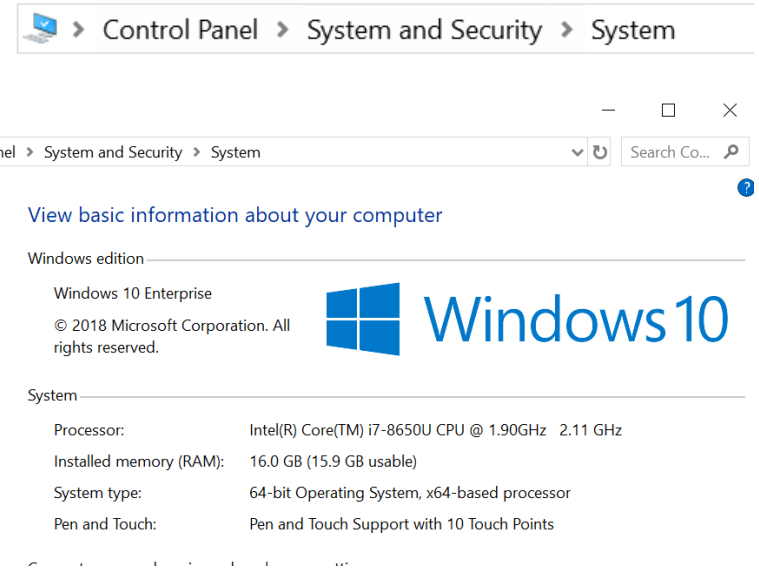
Small Computer System Interface (SCSI) – a set of standard electronic interfaces accredited by ANSI for connecting and transferring data between computers and storage devices.

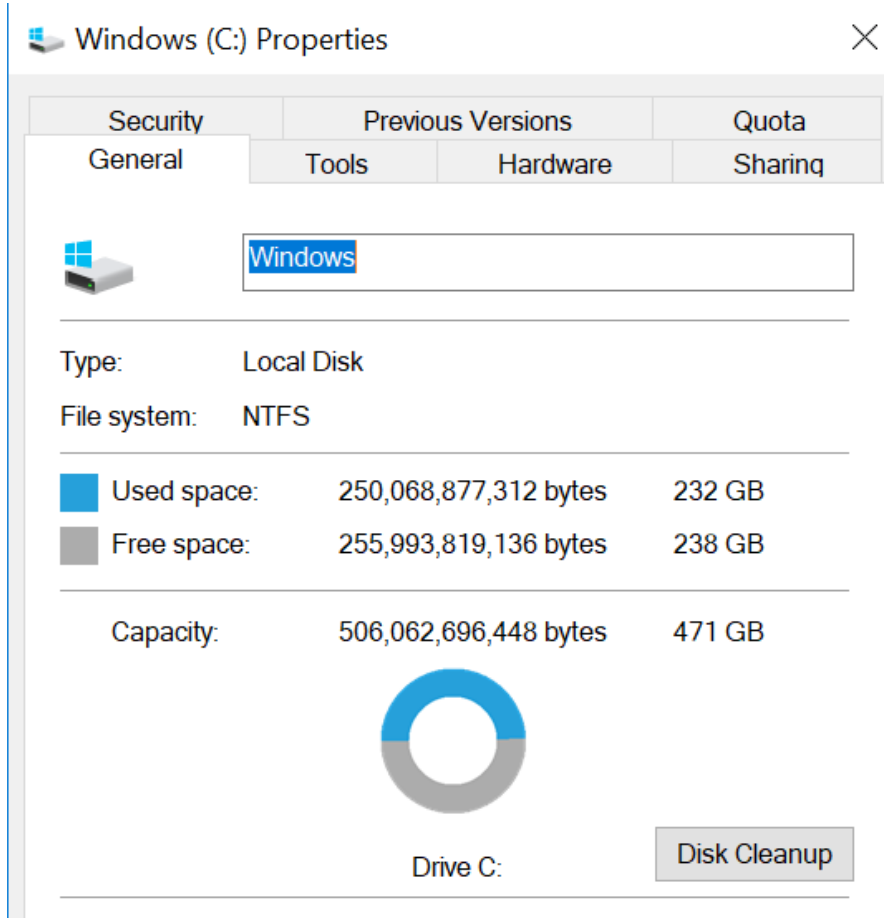
Serial Attached SCSI (SAS) – a data transfer technology that was designed to replace SCSI and to transfer data to and from storage devices.

Fiber Channel (FC) – a high-speed network technology used in storage networking. High speed transfer rate up to 16 gigabits per second.

Find Your Laptop's OS, CPU, and Memory

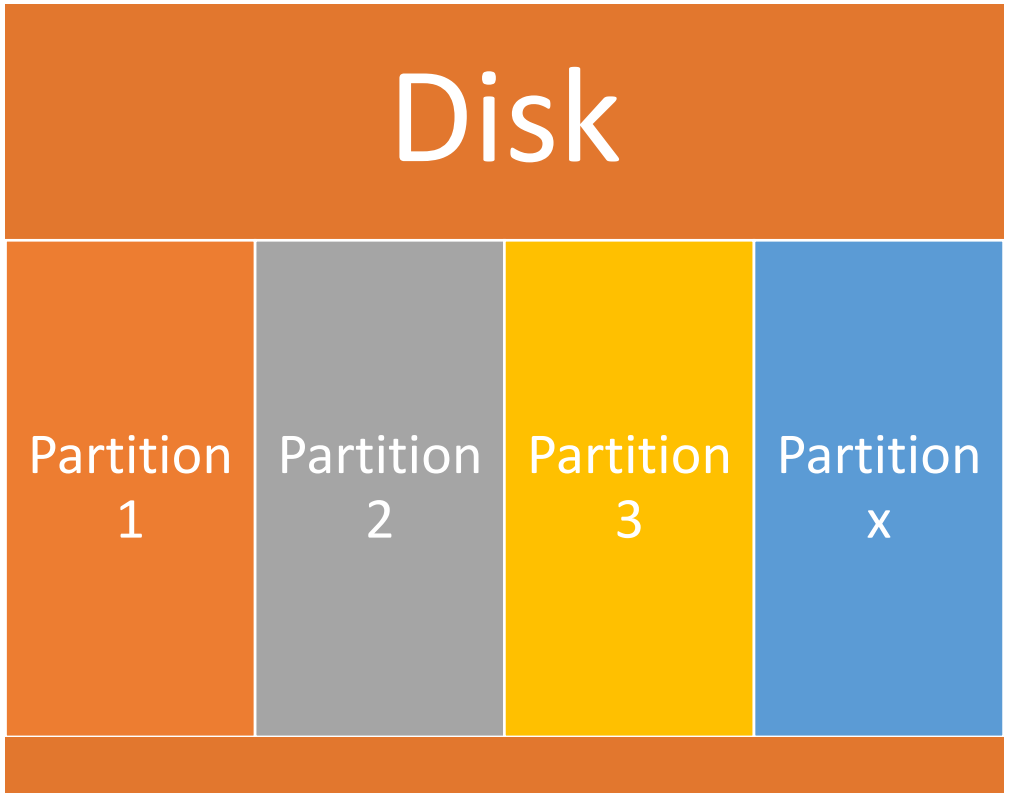
- List the OS version, CPU type and speed, Memory, and Hard drive sizes
- To check the hard drive, right click on C drive and go to property





Disk Partitioning

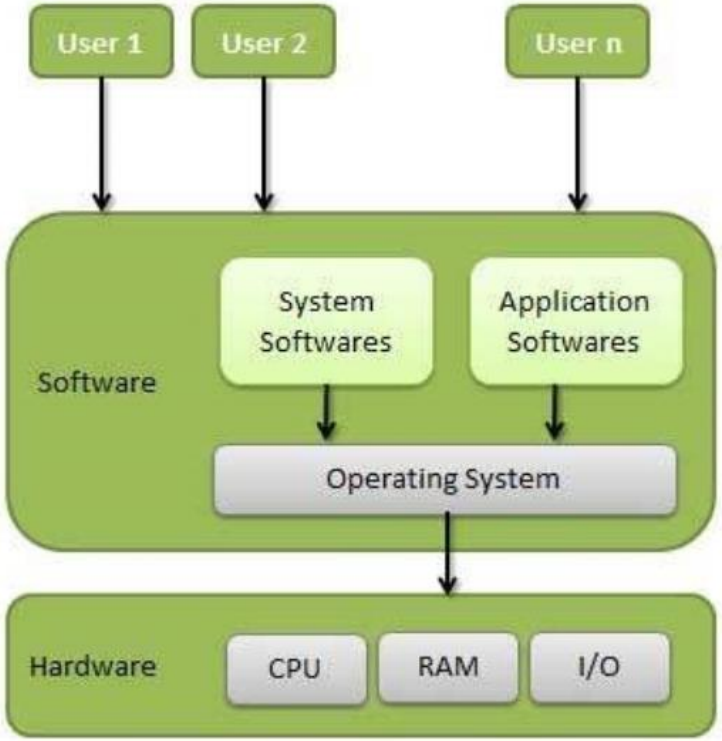
- Partitioning : dividing a physical disk (SSD/HDD) into logical drives (volumes).
- Each partition can have different size.
- Use a disk management tool to format.
- Purpose:
  - To organize files
  - Different OS
  - Different file systems
  - Can be hidden
  - Can be used for recovery



Computer
Operating Systems

Operating System

An **operating system** is system **software** that **manages** computer **hardware and software resources** and provides common services for computer programs.



A kernel is the lowest level of an OS, just above the hardware.

- Kernel manages CPU resources, memory, other resources, and drivers on a computer

Popular OS:

- Linux
- Microsoft Windows
- Apple MacOS
- Android
- Apple's iOS

Shell

Shell

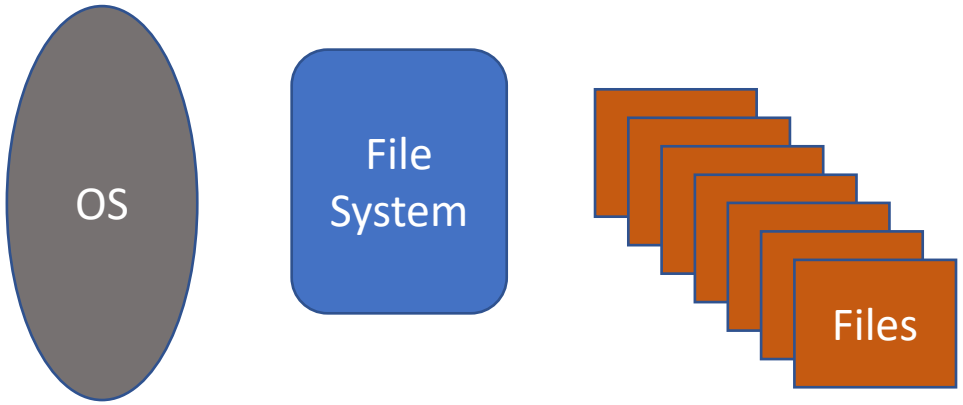
Screen to interface with OS

- Graphical User Interface
  (ex. Windows)
- Line User Interface
  (ex. Command Prompt)

Computer

File System

File System



File System – **software** that accepts the **commands from the OS to read and write data to the disk**.
It is responsible for:

- how the files are named and stored on the disk
- Managing access to the file's metadata (data about the data) and data itself
- Overseeing the relationships to other files and file attributes
- How much available space the disk has
- The reliability of the data on the disk
- Organizing that data in an efficient manner

File System

File Systems	Operating System	Notes
Unix File System (UFS)	Unix and Unix Based OS	Hierarchical file system structure where the highest level of the directory is called the root (/). All files are related in parent-child relationship.
Extended File System (EXT)	**LINUX**	The metadata and file structure is based on the UFS
File Allocation Table File System (FAT)		Mostly been replaced by NTFS (Microsoft)
New Technology File System (NTFS)	Microsoft Windows	A proprietary file system for Windows.
Encrypting File System (EFS)	Microsoft Windows	Provides encryption method for any file or folder on an NTFS partition and is transparent to the user. EFS encrypts a file using a file encryption key (FEK)
Virtual Machine File System (VMFS)		Vmware's cluster file system
Z File system (ZFS)		A combined file syste and logical volume manager designed by Sun Microsystems

Network Infrastructure

Network Types

Network: interconnected computers and peripherals that can share resources, including software, hardware, and files.

Intranet

- **Private Network**
- Use **TCP/IP**
- Controlled by one organization
- Web pages can only be accessed by authorized users
- A **VPN** is a "virtual private network", a piece of software that creates an encrypted communication between two (potentially) far-away computers such that nobody in between can see the contents of the communication.

Extranet

- An **extension of an Intranet**
- Use **TCP/IP**
- Allows controlled access from outside the organization.
- Typically a web site that a company publishes for the benefit of its **vendors, partners, or customers**. Ex: UTA application website, you upload/download documents
- **Not the same as VPN, Extranet** only gives access to the Intranet **via Web browser**.

Internet

- **Global system** of **interconnected computer networks**
- Use **TCP/IP**
- Not controlled by one organization and serves users around the world
- The Internet Corporation for Assigned Names and Numbers (ICANN) is a nonprofit organization coordinating the Internet's system of unique identifiers, including domain names and IP addresses.

Network Scope – defines its boundaries

Local Area Network (LAN)

- Network topology spans small area like office building.
- Ethernet networks 3 data rates:
  - Fast Ethernet (100 Mbps)
  - Gigabit Ethernet (1,000 Mbps)
  - 10 Gigabit Ethernet (10,000 Mbps)

Metropolitan Area Network (MAN)

- Spans a city or a large campus
- Connects multiple LANs
- Use high speed data carriers such as Fiber Optics

Wide Area Network (WAN)

- Large geographic area
- Multiple MANs and LANs
- Internet is the largest WAN
- Some corporations leased lines to create corporate WAN

Internet

- The largest network on earth

Fiber Optic Map in the US

https://www.youtube.com/watch?v=jZOg39v73c4

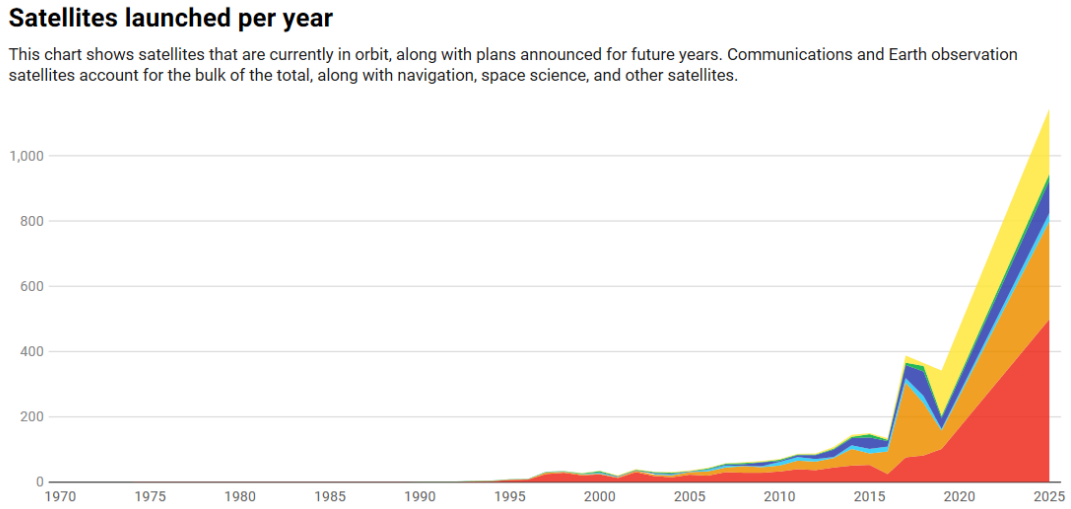
Long-haul Fiber Optics

https://www.youtube.com/watch?v=dGRiVVIUATY

Optical Fiber Cable

- Speed of light
  - Vacuum: 300,000,000 m/s
  - Fiber: ~ 210,000,000 m/s
- Refraction due to different density of mediums
- Dallas – Los Angeles :
  - Distance 2,300,000m
  - Fiber: 2,3/210 = 0.01 sec
- Under ground
- Below the ocean

Amazon Satellites



https://www.technologyreview.com/s/613746/satellite-constellations-orbiting-earth-quintuple/?utm_campaign=the_download.unpaid.engagement&utm_source=hs_email&utm_medium=email&utm_content=74455360&_hsenc=p2ANqtz-_ihmP58oU-aba2demgimKqCNFzm-9IOsz7jh01K9WbfQnzMU0QZo-7hYpKWyBDjsg6pJUgIMpm8AYLetj-Qg6NgLlYSQ&_hsmi=74455360

- Currently: 2000 satellites orbiting
- Amazon asked FCC permission to join SpaceX and launch more than 3,236 satellites
- SpaceX has permission to fly 12,000 small satellites by 2027
- Wants to connect tens of millions around the globe without broadband internet access

Network Topologies – how different nodes are connected

Bus	Star	Ring	Mesh	Tree
Every node is connected to a central cable	Each node is connected to a central switch/hub	Nodes are connected like a ring/circle.	Every node is connected to others.	Multiple star networks are connected through a linear bus backbone
Easy setup and cost effective	Nodes communicate by sending data through the central hub	Each packet is sent around the ring until it reaches its destination.	Difficult to configure and expensive to implement	
Not recommended for large networks (limitations to the number of nodes)	Can easily add new nodes	Hardly used today because one link is broken, the entire connection is broken	Not commonly used	
	Failure of one node does not affect others			
	If switch fails, all are disconnected			

The most popular are **Star** and then **Bus**.

Network Performance - Bandwidth and Latency

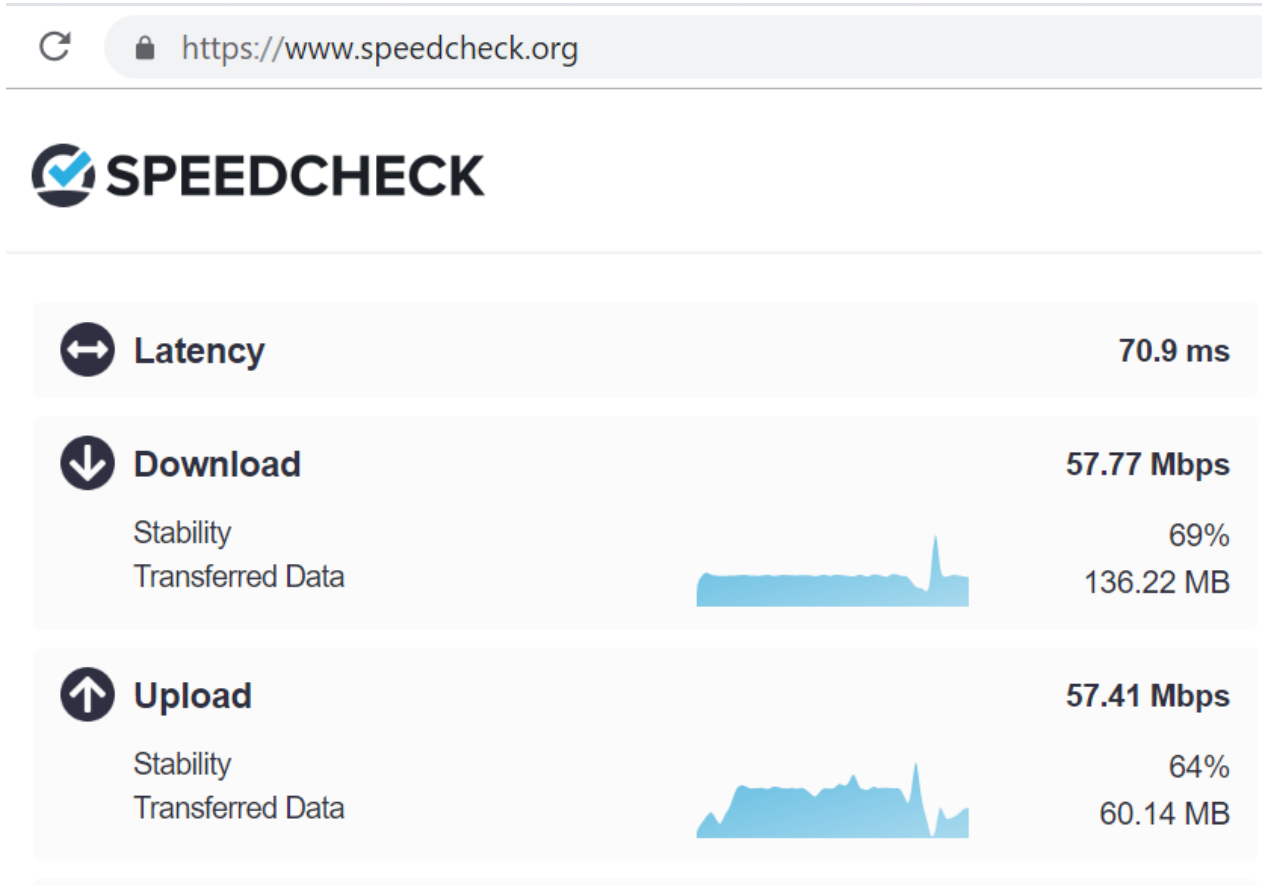
Bandwidth

- Speed of the network
- Bits/second
- If bandwidth is congested, latency increases

Latency

- Time delay (in msec) while data is being sent
- Generally, less than 100ms is acceptable (depending on the application)

Bandwidth check



https://www.speedcheck.org/

Latency comparisons

cmd

```
C:\Users\budimans>ping google.com

Pinging google.com [172.217.12.78] with 32 bytes of data:
Reply from 172.217.12.78: bytes=32 time=8ms TTL=55
Reply from 172.217.12.78: bytes=32 time=10ms TTL=55
Reply from 172.217.12.78: bytes=32 time=10ms TTL=55
Reply from 172.217.12.78: bytes=32 time=11ms TTL=55

Ping statistics for 172.217.12.78:
    Packets: Sent = 4, Received = 4, Lost = 0 (0% loss),
Approximate round trip times in milli-seconds:
    Minimum = 8ms, Maximum = 11ms, Average = 9ms

C:\Users\budimans>ping baidu.com

Pinging baidu.com [220.181.38.148] with 32 bytes of data:
Reply from 220.181.38.148: bytes=32 time=217ms TTL=46
Reply from 220.181.38.148: bytes=32 time=248ms TTL=46
Reply from 220.181.38.148: bytes=32 time=227ms TTL=46
Reply from 220.181.38.148: bytes=32 time=244ms TTL=46

Ping statistics for 220.181.38.148:
    Packets: Sent = 4, Received = 4, Lost = 0 (0% loss),
Approximate round trip times in milli-seconds:
    Minimum = 217ms, Maximum = 248ms, Average = 234ms
```

- Latency is the amount of time network traffic is delayed while the system is processing it.
- The left shows the latency difference between google (server in the US somewhere) and Baidu (server in China).
  - Time = round trip time of the ping message
  - TTL (Time-to-Live): the maximum number of IP routers that the packet can go through before being thrown away.
  - Bytes: Each ping message request is 32 bytes in size. This is a default setting.
- Distance makes a difference.
- This is the reasons why global Cloud providers (AWS, GCP, Azure) have regions worldwide to provide minimum latency.

Ping - localhost

```
C:\Windows\System32>ping localhost

Pinging INSY-125218.uta.edu [::1] with 32 bytes of data:
Reply from ::1: time<1ms
Reply from ::1: time<1ms
Reply from ::1: time<1ms
Reply from ::1: time<1ms

Ping statistics for ::1:
    Packets: Sent = 4, Received = 4, Lost = 0 (0% loss),
Approximate round trip times in milli-seconds:
    Minimum = 0ms, Maximum = 0ms, Average = 0ms

C:\Windows\System32>ping 127.0.0.1

Pinging 127.0.0.1 with 32 bytes of data:
Reply from 127.0.0.1: bytes=32 time<1ms TTL=128
Reply from 127.0.0.1: bytes=32 time<1ms TTL=128
Reply from 127.0.0.1: bytes=32 time<1ms TTL=128
Reply from 127.0.0.1: bytes=32 time<1ms TTL=128

Ping statistics for 127.0.0.1:
    Packets: Sent = 4, Received = 4, Lost = 0 (0% loss),
Approximate round trip times in milli-seconds:
    Minimum = 0ms, Maximum = 0ms, Average = 0ms
```

- localhost – the host-name of the machine you are on.
- localhost IP = 127.0.0.1
- ping localhost = ping 127.0.0.1 – loop back

TCP/IP

TCP/IP – the Internet Protocol

Application Layer	Provides protocols that define how application processes on different hosts exchange messages with each other.
Transport Layer	Control **communications** between the **end-to-end** hosts: flow control, segmentation and de-segmentation, and error control of the application messages.
Internet Layer	responsible for logical transmission of data packets over the internet (connection). It controls the routing of packets using IP addresses. Routers are in this layer.
Network Access Layer	The combination of OSI Data Link Layer and Physical Layer. It looks out for hardware addressing and the protocols present in this layer allows for the physical transmission of data. Controls the hw devices and media that make up the network. Switches are in this layer.

TCP/IP - internet protocol suite provides end-to-end data communication specifying how data should be packetized, addressed, transmitted, routed, and received.

- Implemented in the Operating System.

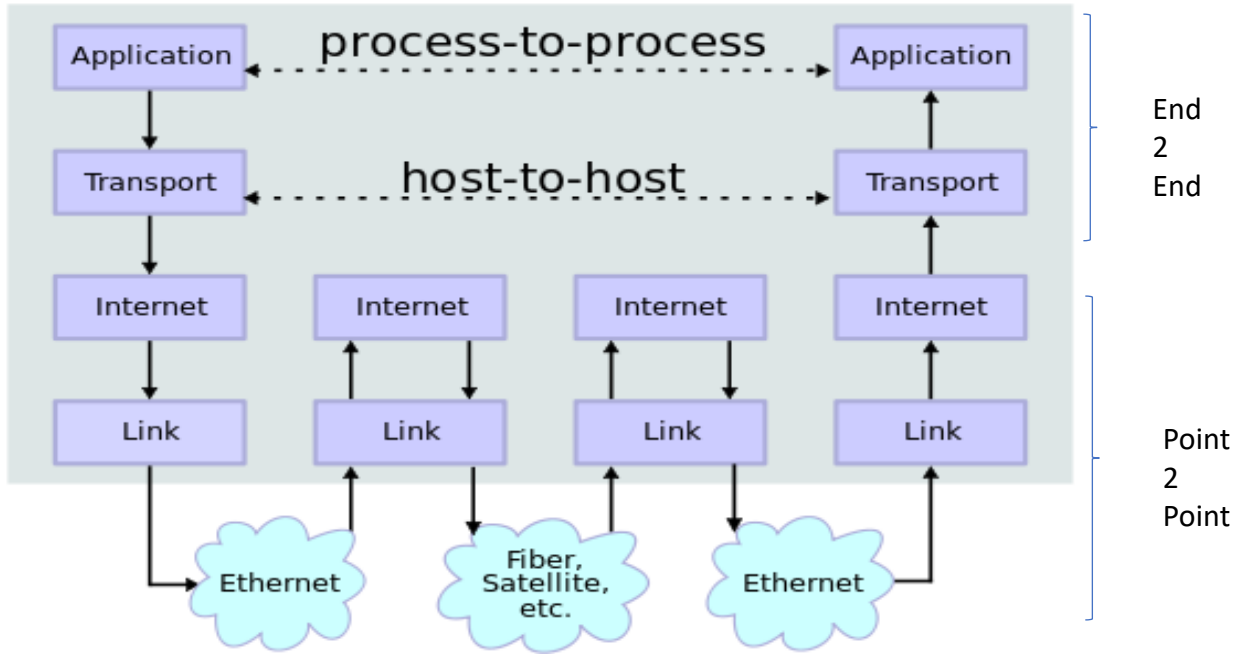
OSI (7 layers) Vs TCP/IP(4 layers):

- The **OSI Model** is a **logical** and conceptual model and better structured.
- The **TCP/IP** model is more **practical** and is popularly used. It is **older than OSI** model.

- Protocol – the format and sequence of messages exchanged between communicating entities.

TCP/IP Layers

TCP/IP Layers	Protocols	Address
Application	http, https, ftp, telnet, smtp, dns, ssh	
Transport	TCP, UDP	Port #s
Internet/Network	IP, ICMP, IGMP	IP Address (IPv4, IPv6)
Network Access Layer	ARP, ATM, Frame Relay, Bluetooth, Ethernet, Wifi,	MAC Address

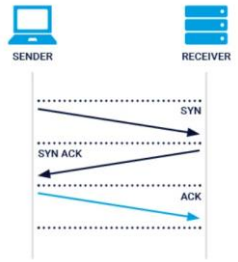


- The main protocols are TCP and IP.
- TCP is a fault tolerant protocol, if a packet is lost or corrupted, the packet is sent again. UDP is not.
- **ARP (Address Resolution Protocol)** – associate IP address and MAC address. It is in between TCP/IP Network Access Layer and Internet layer.

TCP Vs. UDP (Transport Layer)

TCP (Transmission Control Protocol)

- connection-oriented protocol (connection is to be established first)
- Reliable transport
- Speed: slower
- uses handshake protocol



- does error checking and makes error recovery.
- has acknowledgment segments, heavy-weight.
- An application binds a socket to its endpoint of data transmission, which is a combination of destination IP address and port, and source IP address and port.

UDP (User Datagram Protocol)

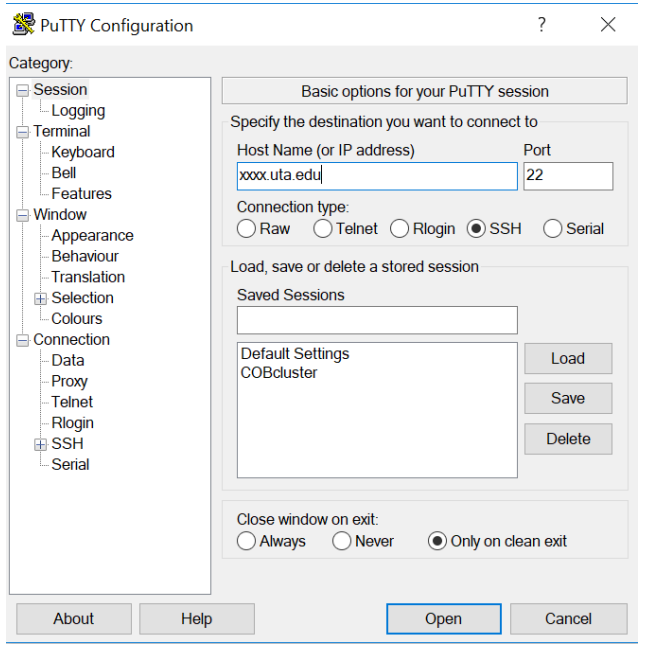
- Connectionless protocol (no session established)
- Not Reliable transport (no guaranteed packets are received)
- Speed: faster
- No handshake protocol
- performs error checking, and discards erroneous packets
- does not have acknowledgment segment
- Lightweight
- An **application** binds a socket to its endpoint of data transmission, which is a combination of an **IP address** and a **port**.
- applications where speed is more critical than reliability

TCP/IP Port Numbers

Port #	Portocol	Description	Status
0	TCP, UDP	Reserved; do not use (but is a permissible source port value if the sending process does not expect messages in response)	Offical
1	TCP, UDP	TCPMUX	Offical
5	TCP, UDP	RJE (Remote Job Entry)	Offical
7	TCP, UDP	ECHO protocol	Offical
9	TCP, UDP	DISCARD protocol	Offical
11	TCP, UDP	SYSTAT protocol	Offical
13	TCP, UDP	DAYTIME protocol	Offical
17	TCP, UDP	QOTD (Quote of the Day) protocol	Offical
18	TCP, UDP	Message Send Protocol	Offical
19	TCP, UDP	CHARGEN (Character Generator) protocol	Offical
20	TCP	FTP - data port (FTP-d)	Offical
21	TCP	FTP - control (command) port (FTP-c)	Offical
22	TCP, UDP	SSH (Secure Shell) - used for secure logins, file transfers (scp, sftp) and port forwarding	Offical
23	TCP, UDP	Telnet protocol - unencrypted text communications	Offical
25	TCP, UDP	SMTP (Simple Mail Transport Protocol) - used for e-mail routing between mailservers	Offical
26	TCP, UDP	RSFTP - A simple FTP-like protocol	Unoffical
35	TCP, UDP	QMS Magicolor 2 printer	Unoffical
443	TCP	HTTPS - HTTP Protocol over TLS/SSL (used for transferring web pages securely using encryption)	Offical

http://www.meridianoutpost.com/resources/articles/well-known-tcpip-ports.php

- A network protocol is an understood set of rules agreed upon by two or more parties.
- A network port is an application specific endpoint to a logical connection.
- Some are reserved, example port 22 is for SSH
- Note: SSH does not use SSL



TCP Port Numbers

A TCP port is a 16-bit number - 65,535 available TCP ports

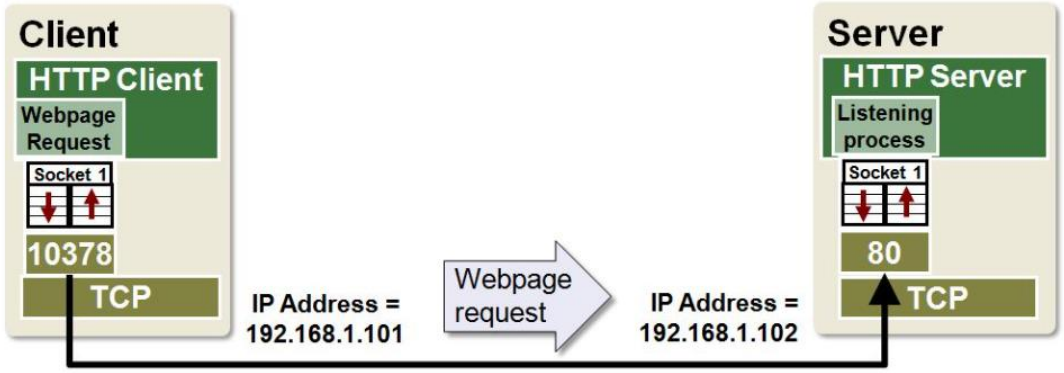
- 0-1023: Well-known ports (agreed upon among technology vendors to support commonly used services). Ex: **22 is for ssh**, **80 for http**. 0 is not a valid TCP port number
- 1024-49151: Registered ports. Assigned to specific services, based on service applications submitted to IANA
  - Ex. MySQL is assigned port 3306
- 49152-65535: Dynamic/private/ephemeral ports. Available for use by any application to use in communicating with any other application.
  - assigned to a process or service at the time the port is needed by the OS, usually when the process or service is started.

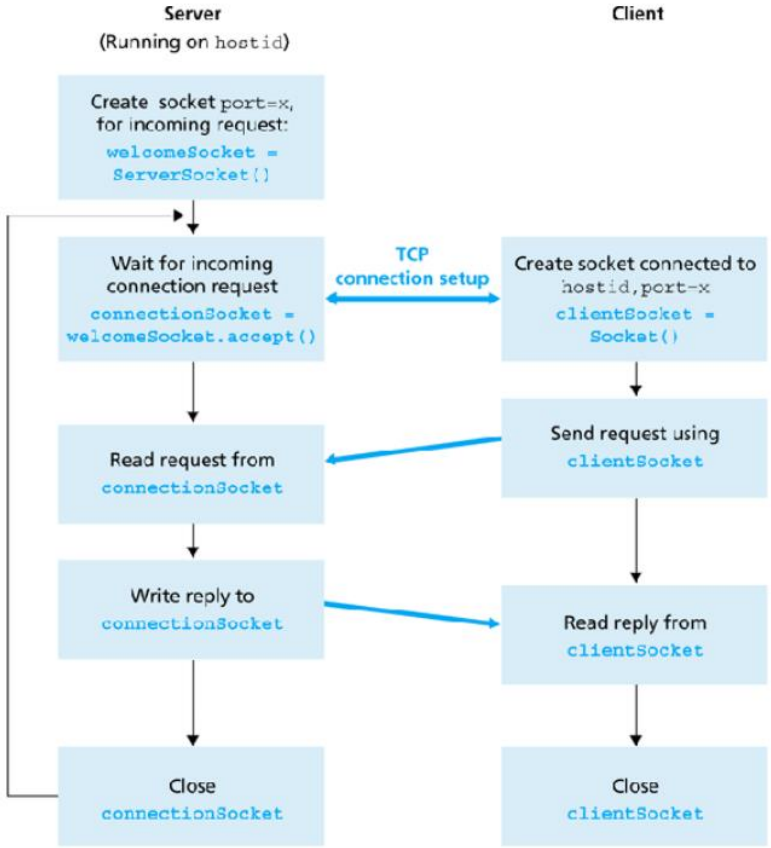
A static port is one whose association with a process or service does not change.

TCP Socket

A socket is the **software interface** between an **application process** and the **transport layer**.

- A port number is assigned to it.
- TCP socket per (source IP, source port #, destination IP, destination port #).
- Source port # and destination port # do not need to be the same.
- A socket per client.



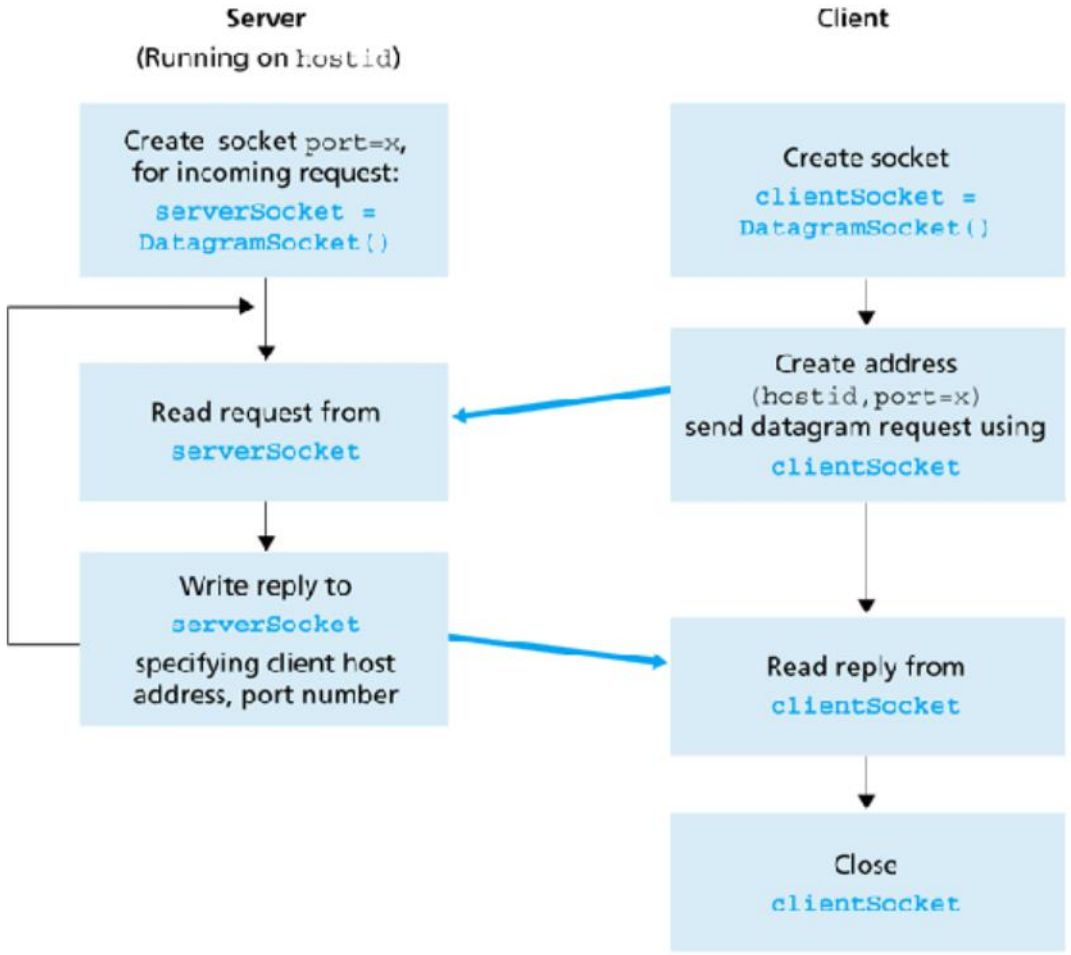


https://microchipdeveloper.com/tcpip:use-sockets-to-establish-a-tcp-connection

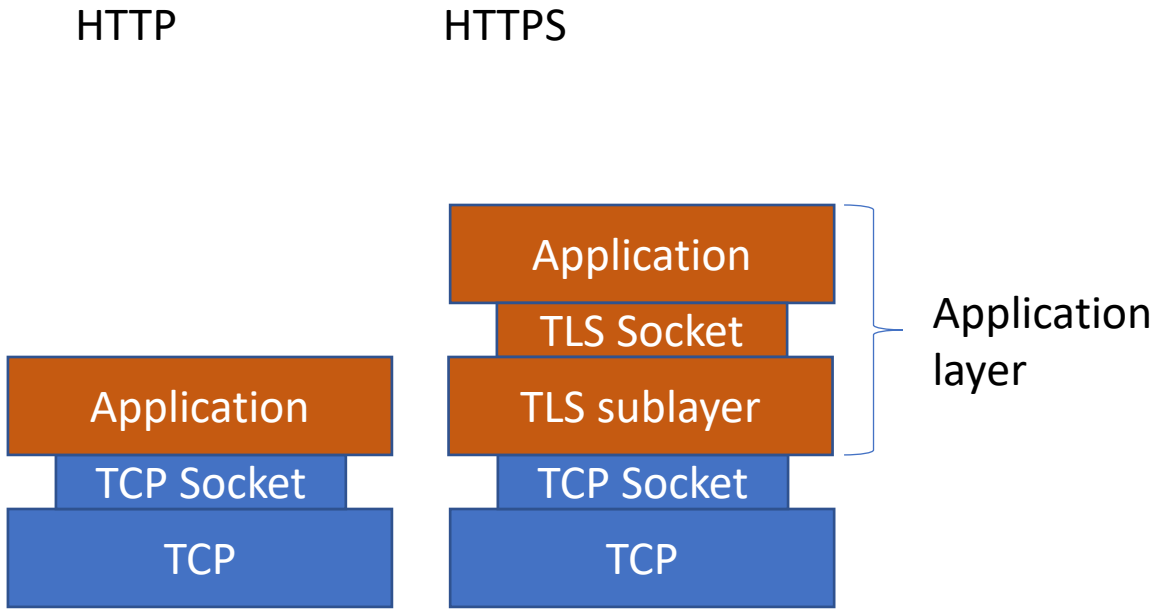
UDP Socket

A socket is the software interface between an application process and the transport layer.

- one socket server to receive from many endpoints, and to send to many endpoints.
- Source IP address and port number are attached by the OS.



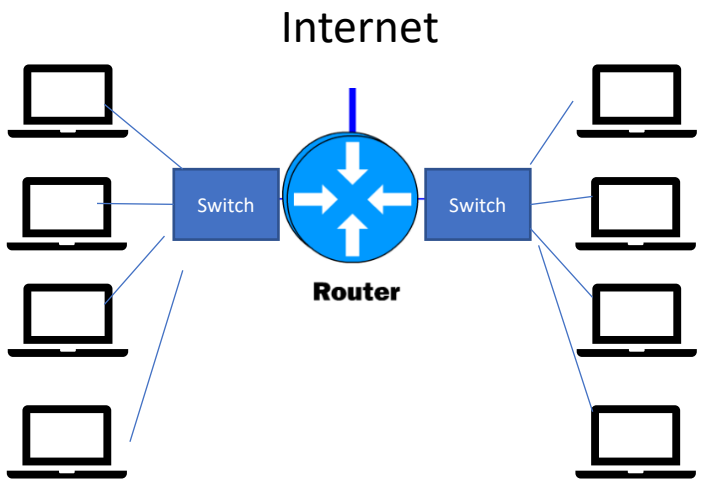
HTTPS – HTTP over TLS



- Transport Layer Security (TLS) – advanced version of SSL
- TLS provides confidentiality, data integrity, server authentication, and client authentication

Routing and Switching

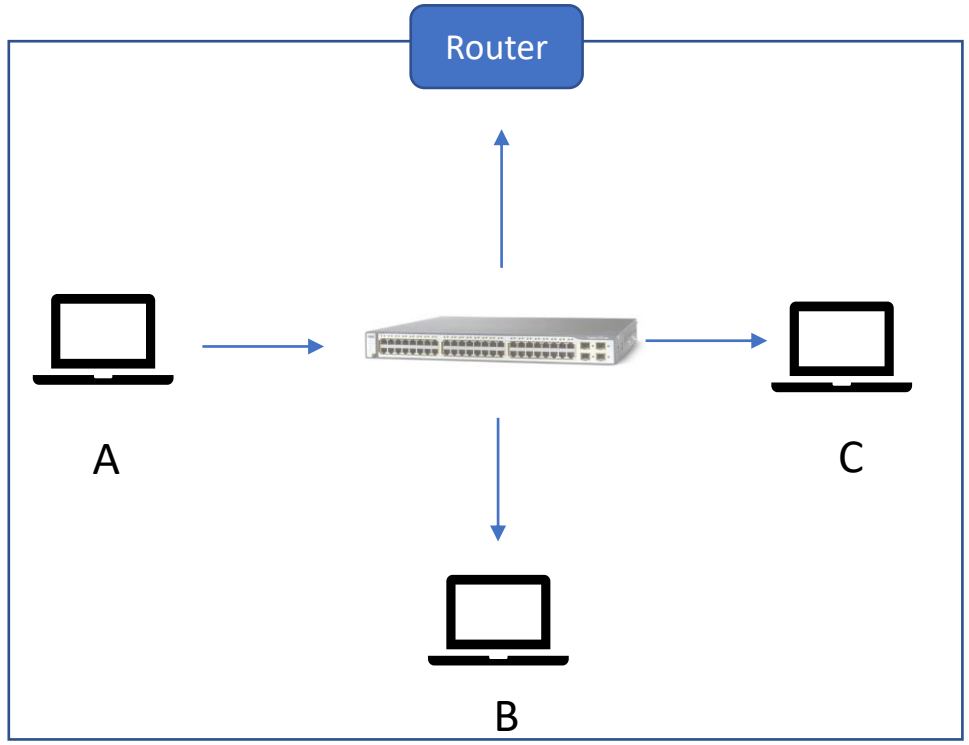
Routers and switches are the networking devices that enable other devices on the network to connect and communicate with each other and with other networks



Switch	Router
Network Access Layer	Network (Internet) layer
To connect multiple devices in the same network or LAN: computers and printers	To connect multiple networks together and allows a network to communicate with the outside world
Typically included in a home router	Makes routing decisions based on the routing protocol configured on it.
Use **Mac address** (not IP address). It is an Ethernet switch. **ARP (Address Resolution Protocol) map the IP addresses to MAC addresses**.	Use **IP address**
Packets send to all devices, some switches will remember the MAC addresses of devices and not send to all any more.	Some routing protocols: Border Gateway Protocol (BGP), Interior Gateway Routing Protocol (IGRP), etc

Media Access Control address (**MAC address**): unique identifier assigned to a network interface controller(NIC).

Switch

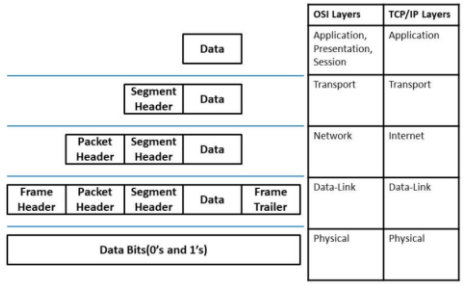


A network switch connects devices in a computer network.

Switch works in **Data Link Layer (A part of TCP/IP Network Access Layer)**

- To connect each device needs a NIC (Network Interface Card).
- **Each NIC** has **a MAC address** (networking hardware address) .
  - Note: VM will have vNIC
- Switches use MAC Address not IP address
- MAC Addresses are unique 48-bits number
  - Ex: 00-14-22-01-23-45
- When first connected to a switch, the switch does not know the devices.
  - Message A to B will be sent to all devices connected.
  - The intended device will acknowledge, others ignore.
  - Switches remember the MAC and can map the logical IP address from acknowledgement and will send to that device only the next time.
    - Older HUB does not have the ability to map ports to MAC addresses and will continue send future messages to all devices
  - Note: you still need the logical private IP address since it is designed that way.

Encapsulation



TCP/IP Layers		
Application	Data	Application produces data to be sent across the network
Transport	Segments	The transport layer breaks the data into segments. Each segment has Transport Header + Part of the app data
Internet/Network	Packets	Network header will be added to each segment and become a packet
Network Access Layer: Data Link	Frames	Frame header and trailer will be added to each packet
Network Access Layer: Physical	Bits	Frames will be converted to bits

Some Notes

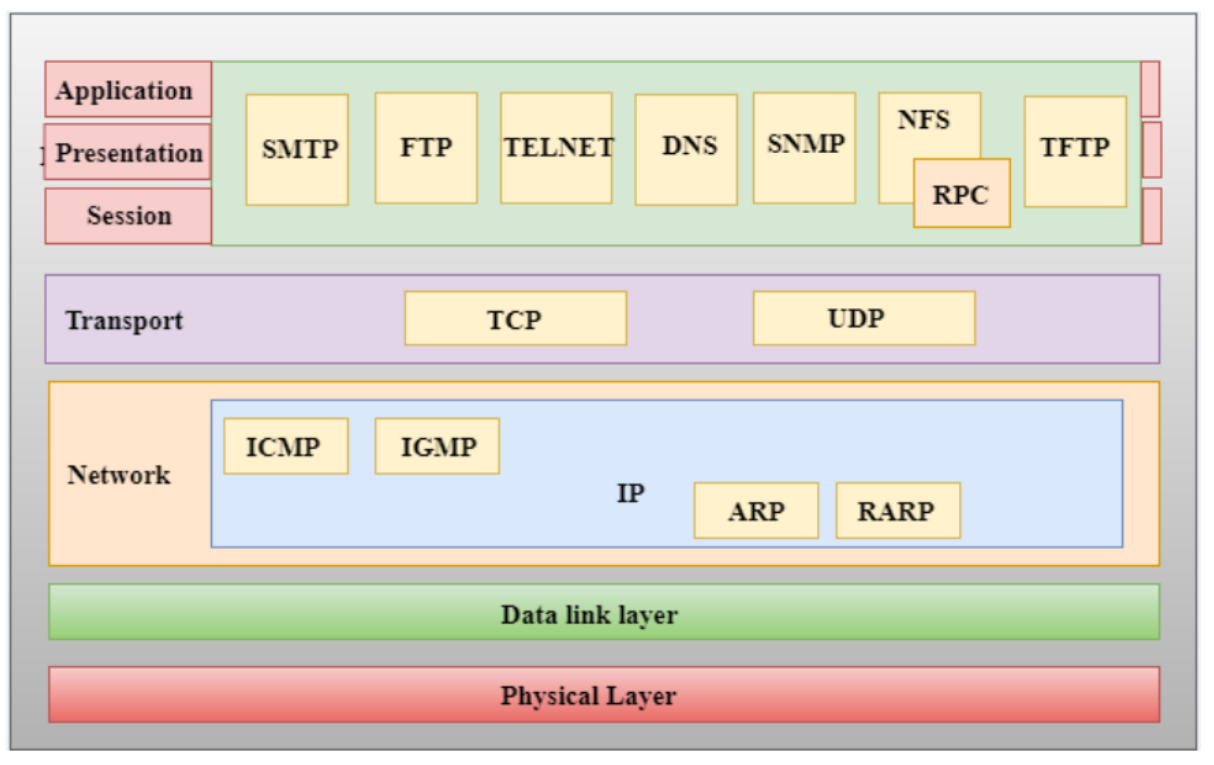
- Remember
  - Transport layer deals with end-to-end communication
  - Network layer deals with point-to-point communication
- Transport : TCP (connection oriented), UDP (connection less)
- Network: Route and Forward
  - IP is connection less
    - Segments of a TCP session may have different paths.

Commonly Used Protocols

			Transport Protocol	Port#
SNMP	Simple Network Management Protocol	Monitor & Modify configurations network devices. Use OID and MIB to identify devices.	UDP	161, 162
NTP	Network Time Protocol	To synch the time between servers	UDP	123
SIP	Session Initiation Protocol	For voice and video	UDP	5060, 5061 (with TLS)
RTSP	Real Time Streaming Protocol	For real time streaming media	UDP	554
FTP	File Transfer Protocol	Transfer data/file from 1 computer to another with username & password	TCP	20 (transfer), 21 (establish)
TFTP	Trivial File Transfer Protocol	Transfer data/file from 1 computer to another without username & password	TCP	69
SFTP	Secure Fie Transfer Protocol	Transfer encrypted data/file from 1 computer to another with username & password. Using SSH	TCP	22 (using SSH)
SSH	Secure Shell	Encrypted communication between the two computers. Often **used** to "login" and perform operations on remote computers, and data transfer.	TCP	22

			Transport Protocol	Port#
LDAP	Lightweight Directory Access Protocol	Directory services authentication. LDAP is a way of speaking to Active Directory.	TCP/UDP	389
RDP	Remote Desktop Protocol	Connect and manage a computer from another using GUI. Mostly TCP but can use UDP as well	TCP/UDP	3389
Telnet	Telnet	Connect and manage a computer from another using Command line interface.	TCP	23
SMTP	Simple Mail Transfer Protocol	Mail servers and other message transfer agents use SMTP to send and receive mail messages.	TCP	25, 465 with TLS/SSL
IMAP4	Internet Message Access Protocol	Download just email from mail server and synchronize.	TCP	143, 993 with TLS/SSL
POP3	Post Office Protocol	Download email from mail server and delete email from server.	TCP	110,995 with TLS/SSL
DNS	Domain Name System		TCP/UDP	53
DHCP	Dynamic Host Configuration Protocol		UDP	67, 68
HTTP	Hypertext Transfer Protocol		TCP	80
HTTPS	Hypertext Transfer Protocol Secure		TCP	443

Ping command to check connectivity

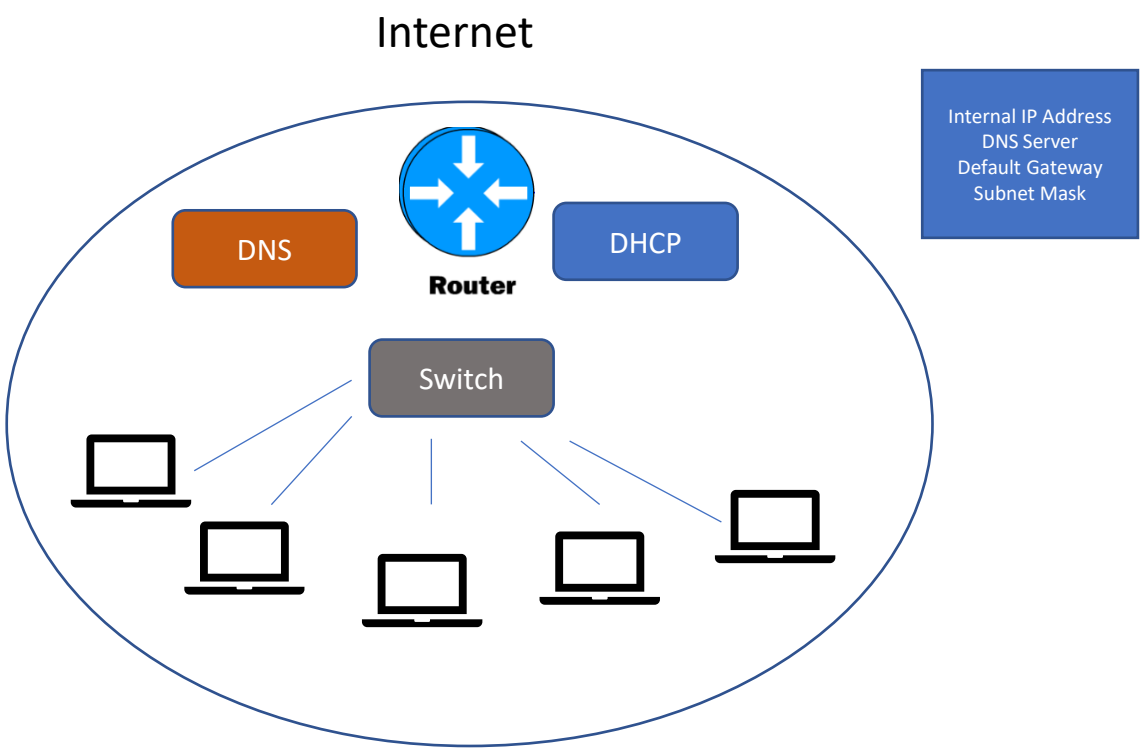


- The Internet Control Message Protocol (**ICMP**):
  - A network level protocol (above IP but not in Transport)
  - not associated with transport layer protocol (TCP or UDP).
  - For error reporting and network diagnostics
  - Sends management messages between systems (ping request).
- **Ping** is a command used to test the reachability (connectivity) of a host on an Internet Protocol (IP) network. It is available for virtually all OS that have networking capability.
  - In linux you can use ping ipaddress or ping –c3 ipaddress (-c3 if you want to limit to 3 times).
  - Note: Command Prompt does not recognize –c3 (ping ipaddress)
- **Ping** uses **ICMP**. The tool sends **ICMP** Echo Request packets to the destination host and waits for **ICMP** Echo Replies.

DDoS

- **denial-of-service** (**DDoS**) **attack** is a malicious attempt to disrupt the normal traffic of a targeted server, service or network by overwhelming the target or its surrounding infrastructure with a flood of Internet traffic.
- **ICMP flood attack** - A ping flood or ICMP flood is when the attacker attempts to overwhelm a targeted device with ICMP echo-request packets.
- **A ping of death** is a type of attack on a computer system that involves sending a malformed or otherwise malicious ping to a computer.

TCP/IP Network



A host in a subnet can only talk to a host in another subnet through a Router.

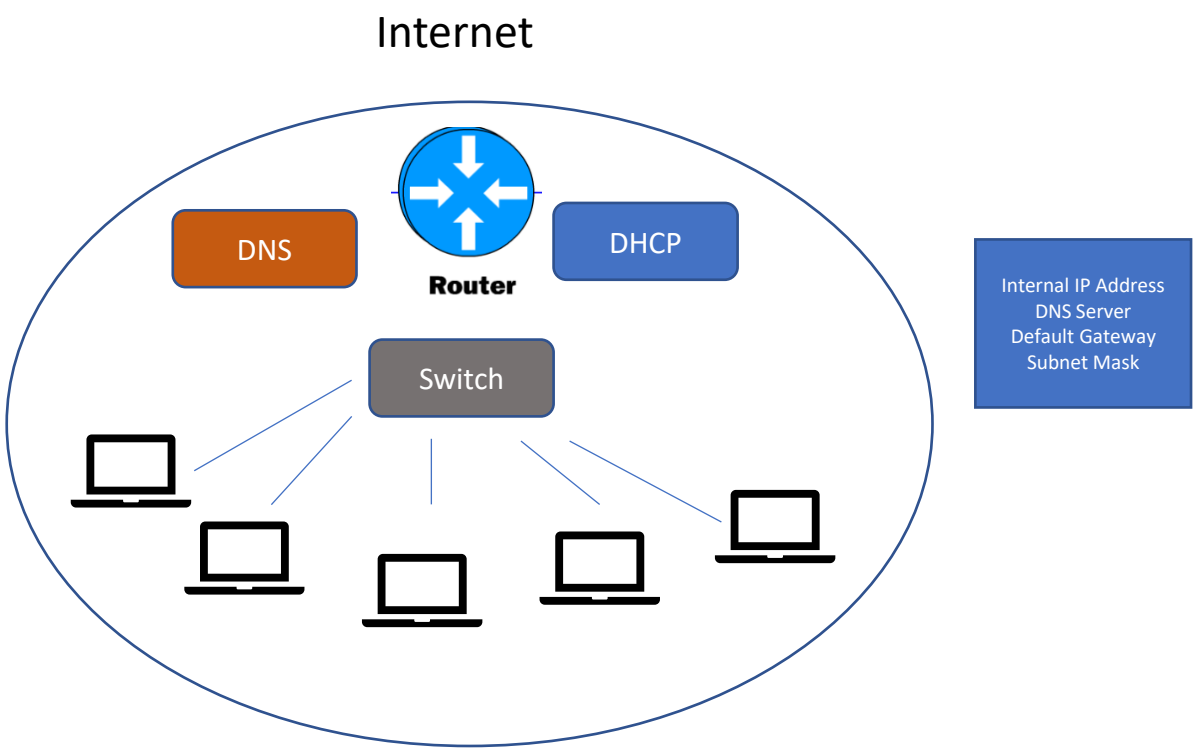
- This router is sometimes called a **Default Gateway**.

If a computer is looking for an IP address, it looks at its own subnet.

- If not found then goes to the Default Gateway.

DNS maps a domain name, such as uta.edu to an IP address. Computers only cares about IP address.

TCP/IP Network



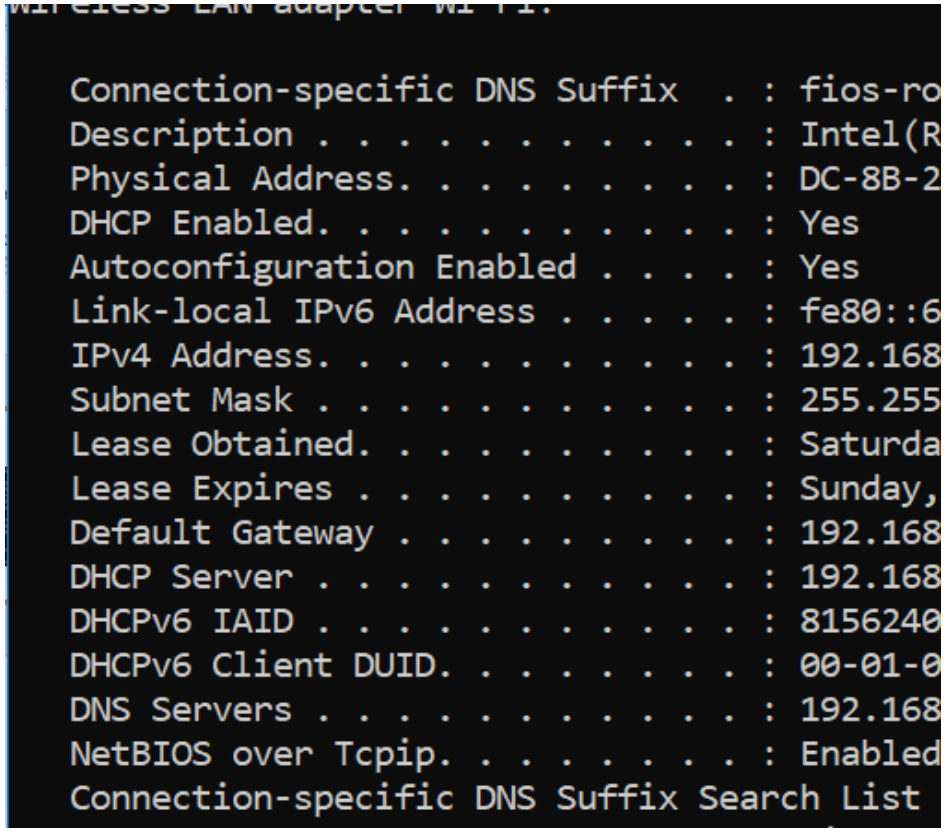
Internal IP Address
DNS Server
Default Gateway
Subnet Mask

Every host needs a **unique** internal IP address.

IP address assignment:

- Static IP address (user manually assigns IP address, subnet mask, Default Gateway, DNS server).
- Dynamic IP address using **DHCP (Dynamic Host Control Protocol)**
  - When a computer connects to a network, it calls out to DHCP server.
  - The DHCP is programmed/configured by an administrator on the **IP address it can provide (the range), the subnet mask, Default Gateway IP address, lease time**.
  - DHCP server provides: assigned IP address (with a lease time), subnet mask, DNS server, Default Gateway.
  - Lease time: how long the host can keep the IP address.
    - Halfway, the host will contact DHCP to renew the least time.
    - The shorter the lease time, the lesser IP address (recirculated)
  - DHCP server keeps track of IP addresses and ensure they are unique (imagine if there are hundreds of hosts).

- In your laptop, you can use: ipconfig/all
- In my case the DHCP Server, DNS Server, and Default Gateway have the same private IP address, which indicate that all functions are collocated in my Wifi router.



AWS Domain Name System

Domain Name System (DNS), translates human readable domain names to IP addresses.

DNS service:

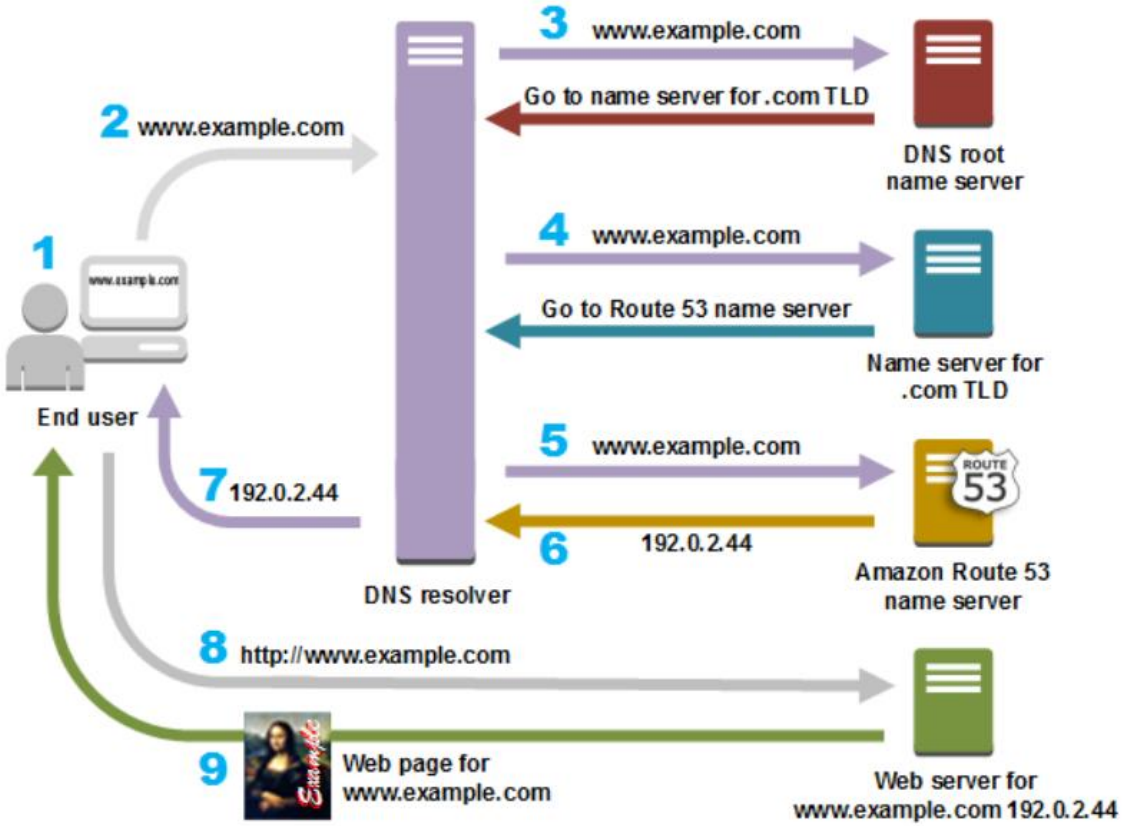
- **Authoritative DNS** has the final authority over a domain and is responsible for providing answers to recursive DNS servers with the IP address information. Ex. Amazon Route 53
- **DNS Resolver or Recursive Resolver** is a server designed to receive DNS queries from web browsers and other applications and track down the IP address. Typically resides in an ISP or corporate office.

Amazon Route 53 is a DNS provider and an Authoritative DNS.

TLD- Top Level Domain

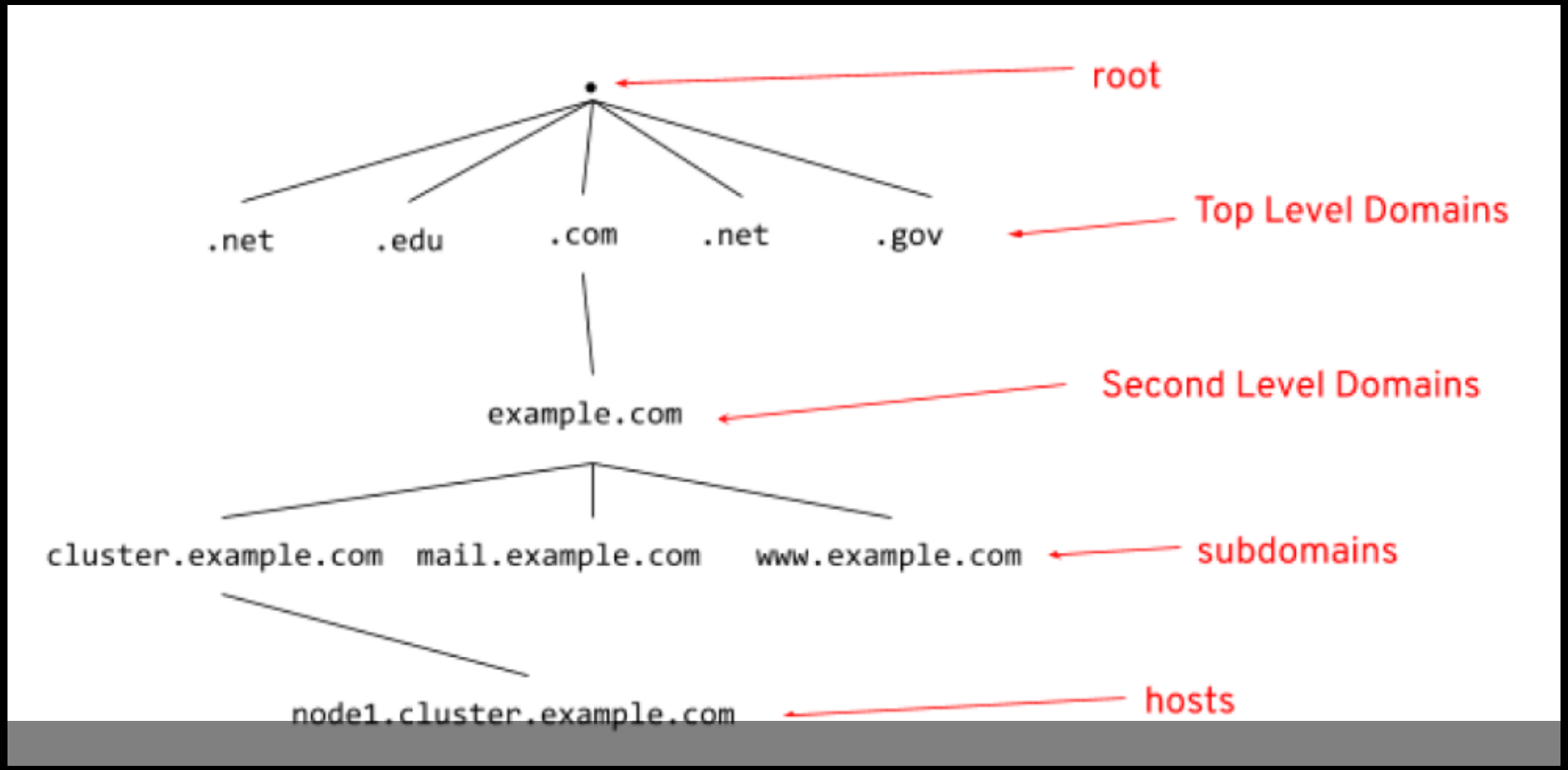
Root Name server - a name server for the root zone of the Domain Name System of the Internet.

https://aws.amazon.com/route53/what-is-dns/

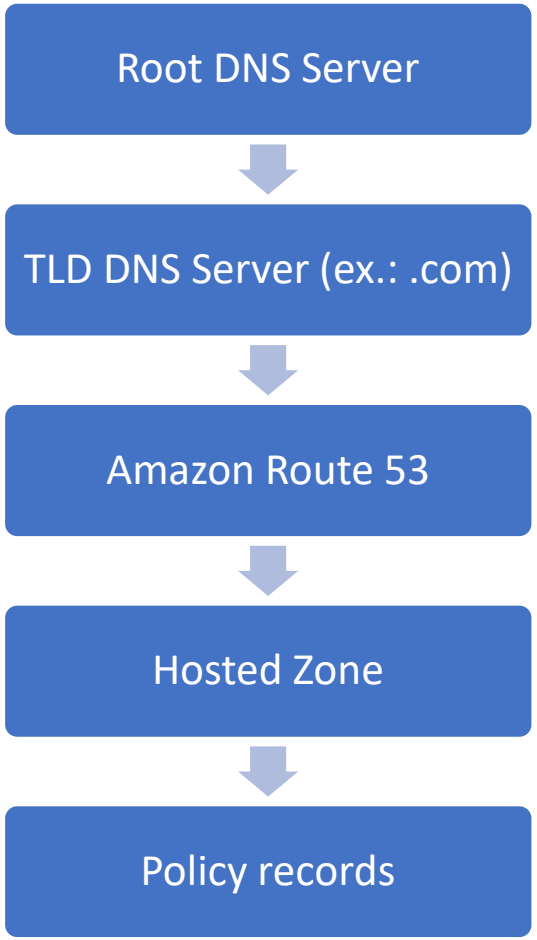


DNS Hierarchy

The Domain Name System (DNS) is hierarchical and decentralized



The Internet Corporation for Assigned Names and Numbers (ICANN) manages the root and TLD.


Root DNS Server
TLD DNS Server (ex.: .com)
Amazon Route 53
Hosted Zone
Policy records

DNS steps explanation (Amazon Route 53)

1. A user opens a web browser, enters www.example.com in the address bar, and presses Enter.
2. The request for www.example.com is routed to a DNS resolver, typically managed by the user's ISP, or a corporate network.
3. The DNS resolver forwards the request for www.example.com to a DNS root name server.
4. The DNS resolver forwards the request for www.example.com again, this time to one of the TLD name servers for .com domains. The name server for .com domains responds to the request with the names of the four Amazon Route 53 name servers that are associated with the example.com domain.
5. The DNS resolver for the ISP chooses an Amazon Route 53 name server and forwards the request for www.example.com to that name server.
6. The Amazon Route 53 name server looks in the example.com hosted zone for the www.example.com record, gets the associated value, such as the IP address for a web server, 192.0.2.44, and returns the IP address to the DNS resolver.
7. The DNS resolver for the ISP finally has the IP address that the user needs. The resolver returns that value to the web browser. The DNS resolver also caches (stores) the IP address for example.com for an amount of time that you specify so that it can respond more quickly the next time someone browses to example.com. For more information, see time to live (TTL).
8. The web browser sends a request for www.example.com to the IP address that it got from the DNS resolver. This is where your content is, for example, a web server running on an Amazon EC2 instance or an Amazon S3 bucket that's configured as a website endpoint.
9. The web server or other resource at 192.0.2.44 returns the web page for www.example.com to the web browser, and the web browser displays the page.

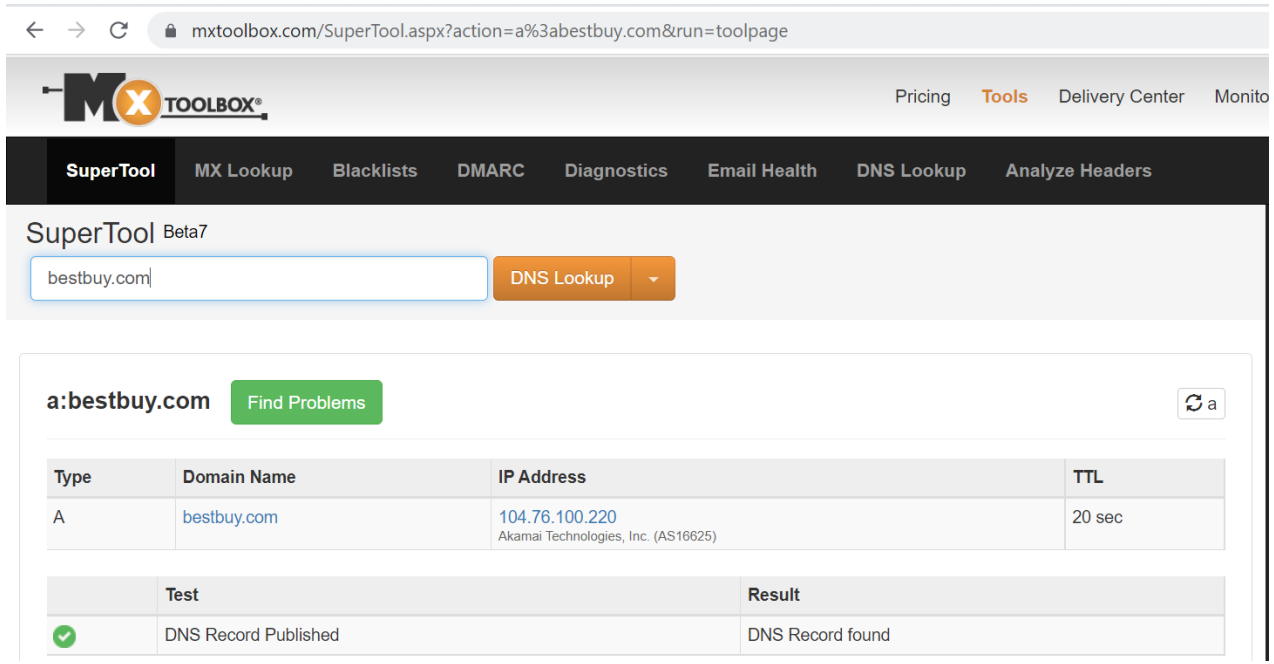
Amazon Route 53

- 53 is the port number for DNS
- Each record has:
  - Domain/Subdomain name. Ex. www.bestbuy.com
  - Record type. The important once:
    - A – translates a hostname to IPv4
    - AAAA – translates a hostname to IPv6
    - CNAME – maps a hostname to another hostname
    - NS - Name Servers for the Hosted Zone
    - Note: Alias (aws specific) – maps a hostname to an aws resource (similar to CNAME), used in A/AAAA

Route 53 Routing Policy

- Simple – route traffic to single resource
- Weighted - % of traffic to specific resource
- Failover – active to passive (based on health check)
- Latency based – to the resource with the lowest latency
- Geolocation – based on user location
- Multi-Value Answer – multiple resources for the client to choose (combined with health check)
- Geoproximity – based on geographic location of the resources.

Ex




```
ubuntu@ip-172-31-11-176:~$ dig bestbuy.com

; <<>> DiG 9.18.1-1ubuntu1.2-Ubuntu <<>> bestbuy.com
;; global options: +cmd
;; Got answer:
;; ->>HEADER<<- opcode: QUERY, status: NOERROR, id: 61010
;; flags: qr rd ra; QUERY: 1, ANSWER: 1, AUTHORITY: 0, ADDITIONAL: 1

;; OPT PSEUDOSECTION:
; EDNS: version: 0, flags:; udp: 65494
;; QUESTION SECTION:
;bestbuy.com.                   IN      A

;; ANSWER SECTION:
bestbuy.com.            20      IN      A       23.39.52.177

;; Query time: 31 msec
;; SERVER: 127.0.0.53#53(127.0.0.53) (UDP)
;; WHEN: Wed Dec 28 14:11:12 UTC 2022
;; MSG SIZE  rcvd: 56

ubuntu@ip-172-31-11-176:~$ nslookup bestbuy.com
Server:         127.0.0.53
Address:        127.0.0.53#53

Non-authoritative answer:
Name:   bestbuy.com
Address: 23.34.76.219

ubuntu@ip-172-31-11-176:~$
```

Local Area Network Physical Layer Protocols (Ethernet & Wifi)

IEEE 802 is a **collection of networking standards** that cover the **physical and data-link layer** specifications for technologies such as Ethernet and wireless.

Ethernet (IEEE **802.3**x)

- **Ethernet** is a standard **wired** system for connecting computers to a local area network (LAN).
- Speed ranging from 10 Mbps to 10Gbps
- Operates at the physical and data link layers of the OSI model (layer 1 and 2)
- Limited by both the **length** and the **type of cables**.
- The Ethernet standard divides its data traffic into groupings called frames.

Wireless LAN Wifi (IEEE **802.11**x)

- **Wi-Fi - Wireless** Fidelity
- Speed ranging from 11 Mbps to 3.4 Gbps

IEEE 802 standards (not exhaustive)

IEEE 802.1	Bridging (networking) and network management
IEEE 802.2	Logical link layer
IEEE 802.3	Ethernet (CSMA/CD)
IEEE 802.4	Token bus (disbanded)
IEEE 802.5	Defines a MAC layer for a token ring (inactive)
IEEE 802.6	Metropolitan Area Networks (disbanded)
IEEE 802.7	Broadband LAN using coaxial cable (disbanded)
IEEE 802.8	Fiber optic TAG (disbanded)
IEEE 802.9	Integrated Services LAN (disbanded)
IEEE 802.10	Interoperable LAN Security (disbanded)
IEEE 802.11	Wireless LAN and mesh (Wi-Fi certification)
IEEE 802.12	Demand Priority (disbanded)
IEEE 802.13	Not used
IEEE 802.14	Cable modems (disbanded)
IEEE 802.15	Wireless PAN
IEEE 802.15.1	Blue-tooth certification
IEEE 802.15.4	ZigBee Certification
IEEE 802.16	Broadband Wireless Access (WiMax Certification)

Ethernet IEEE Standards

IEEE Standard	Speed
802.3	10 Mbps
802.3u	100 Mbps
802.3z	1 Gbps
802.3ab	1 Gbps
802.3an	10Gbps

https://study-ccna.com/ieee-ethernet-standards/

Ethernet Speed: Cable type and Distance

ETHERNET CABLE PERFORMANCE SUMMARY

CATEGORY	SHIELDING	MAX TRANSMISSION SPEED (AT 100 METERS)	MAX BANDWIDTH
Cat 3	Unshielded	10 Mbps	16 MHz
Cat 5	Unshielded	10/100 Mbps	100 MHz
Cat 5e	Unshielded	1000 Mbps / 1 Gbps	100 MHz
Cat 6	Shielded or Unshielded	1000 Mbps / 1 Gbps	>250 MHz
Cat 6a	Shielded	10000 Mbps / 10 Gbps	500 MHz
Cat 7	Shielded	10000 Mbps / 10 Gbps	600 MHz

https://www.electronics-notes.com/articles/connectivity/ethernet-ieee-802-3/how-to-buy-best-ethernet-cables-cat-5-6-7.php

Suitable Cable

Informal name	Common name	Formal IEEE name	Speed	Cable and Max Length	Suitable Cable
10 BASE-T	Ethernet	802.3	10Mbps	Copper,100 m	category3, 5
100 BASE-T	Fast Ethernet	802.3u	100Mbps	Copper,100m	Category5
1000BASE-T	Gig Ethernet	802.3z	1000Mbps	Fiber,5000m	Category5e

The appropriate cable type and distance must be considered when designing Ethernet LAN to enable speed

WIFI IEEE Standards

IEEE standard	Speed	Frequency	Different Naming convention
802.11a	Up to 54 Mbps	5 GHz	WiFi2
802.11b	Up to 11 Mbps	2.4 GHz	WiFi1
802.11g	Up to 54 Mbps	2.4 GHz	WiFi3
802.11n	Up to 600 Mbps	2.4 and 5 GHz	WiFi4
802.11ac	Up to 3.46Gbps	5 GHz (some vendors include 2.4 GHz)	Wifi5
802.11ax	**Up to 10Gbps**	**2.4 and 5GHz**	**Wifi6 (2019)**

Wifi6 and 5G – complementary technology, launched about the same time, enhanced IOT – (seamless roaming)

https://www.networkworld.com/article/3238664/80211-wi-fi-standards-and-speeds-explained.html

IP Address

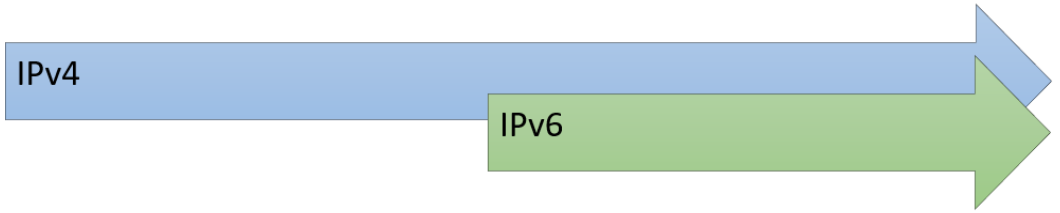
Internet Protocol address (**IP address**) is a unique string of numbers separated by periods that identifies each computer using the Internet Protocol to communicate over a network.

IPv4:

- Internet Protocol version 4 (**IPv4**) defines an IP address with **32-bit** number (**4 octets**). xxxxxxxx.xxxxxxxx.xxxxxxxx.xxxxxxxx
- IPv4 is still widely used and some expect to continue decades to come.
- This course will use IPv4.

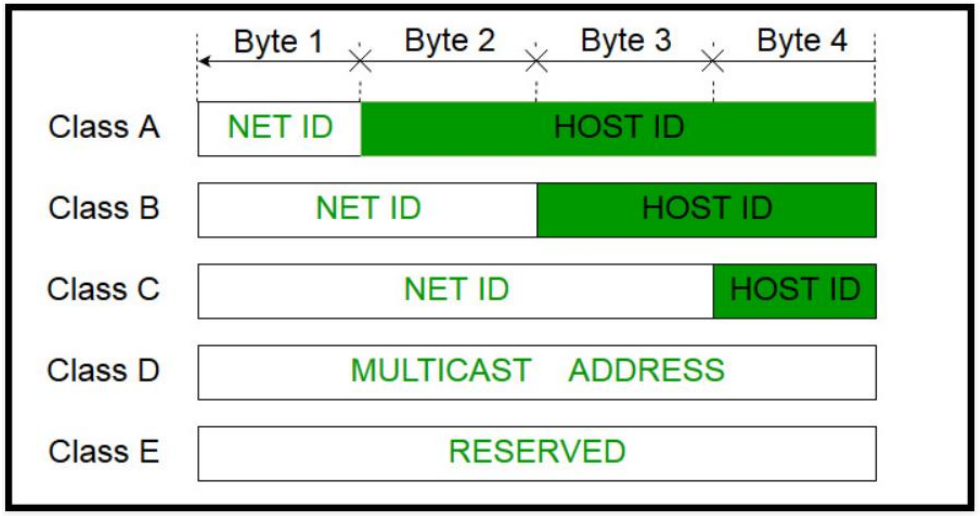
IPv6:

- Due to proliferation of internet devices, **IPv6** with **128-bit** number is developed and standardized in 1998.



IPv4

- IPv4 address is divided into two parts: **Network ID** + **Host ID**
- It uniquely identifies a machine (host) on a network.
- All hosts in a network have the same Network ID but a unique Host ID.
- IP address can be private (internal) or public (external):
  - a device on a network has a private IP address only seen by other devices on that local network.
  - an ISP assigns a public IP address to the network or the device so that other devices on the Internet can locate.
- Previously, the **IP** hierarchy is broadly divided into 5 classes of the **IP addresses**.
- To extend the life of IPv4, **Classless Inter-Domain Routing (CIDR)** --an IP addressing scheme that improves the allocation of IP addresses-- replaces the class system above.



Class	Address Range	Supports
Class A	1.0.0.1 to 126.255.255.254	Large networks with many devices
Class B	128.1.0.1 to 191.255.255.254	Medium-sized networks.
Class C	192.0.1.1 to 223.255.254.254	small networks (fewer than 256 devices)
Class D	224.0.0.0 to 239.255.255.255	Reserved for multicast groups.

Classless Inter-Domain Routing (CIDR)

Classless Inter-Domain Routing is a method for allocating IP addresses and IP routing.

The Internet Engineering Task Force (IETF) introduced CIDR in 1993 to replace the previous addressing architecture of classful network design.

With CIDR, a network of IP addresses is allocated in 1-bit increments as opposed to 8-bits in classful network.

CIDR notation: 192.17.15.6/18 (an IP address with slash and a number).

- The **/x** is called the **IP or network prefix (bit length of the prefix)**.
  - x is the number of bits assigned to the network address (fixed and the remaining is flexible)
- The network prefix must match the subnet mask (network prefix is for human, subnet mask is for machine).
- a /18 block is a CIDR block with a 18-bit prefix.
- Short prefixes (ex./8) allow for more addresses (more hosts) while large prefixes (ex. /24) identify small blocks (less hosts).

The use of a CIDR notated address can easily represent classful addresses (**Class A = /8, Class B = /16, and Class C = /24**).

Its goal:

- to slow the growth of routing tables on routers across the Internet, and
- to help slow the rapid exhaustion of IPv4 addresses.

CIDR and Netmask

- A **subnet mask** separates the IP address into the network and host addresses (<network><host>).
- Subnetting further divides the host part of an IP address into a subnet and host address (<network><subnet><host>) if additional subnetwork is needed. We will see this when we create subnets in a VPC.
- A **Subnet mask** is a **32-bit number** that masks an IP address (AND operation).
- Subnet Mask is made by setting **network bits to all "1"s** and setting **host bits to all "0"s**.
- Within a given network, at least two host addresses are reserved for special purpose, and cannot be assigned to hosts:
  - "0" address is assigned a network address
  - "255" is assigned to a broadcast address.

CIDR and Netmask

n bits m bits

n +m = 32 bits

The maximum number of network in theory is **2^n** . n= the number of bits from the left (used for network ID). Note, there might be some reserved. n = x in /x (the network prefix).

The maximum number of hosts within a network is **2^m-2**. m= the number of bits from the right (used for host ID). 2 is subtracted per reservation above. m = 32-n. Note: **Different providers may reserve more IP addresses.**

- https://www.iplocation.net/subnet-mask

CIDR problem

	IP Address	129.107.58.60																															
	Subnet Mask	255.255.192.0																															
		Octet 4		129							Octet 3		107							Octet 2		58							Octet 1		60				
	Dec value	128	64	32	16	8	4	2	1		128	64	32	16	8	4	2	1		128	64	32	16	8	4	2	1		128	64	32	16	8	4	2	1
	0									0									0									0							
Subnet Mask	0									0									0									0							

	IP Address	129.107.58.60																																	
	Subnet Mask	255.255.192.0	Only 1 subnet mask for a network, this is to separate the network ID and the host ID																																
		Octet 4		129							Octet 3		107							Octet 2		58							Octet 1		60				
	Dec value	128	64	32	16	8	4	2	1		128	64	32	16	8	4	2	1		128	64	32	16	8	4	2	1		128	64	32	16	8	4	2	1
129	1	0	0	0	0	0	0	1	107	0	1	1	0	1	0	1	1	58	0	0	1	1	1	0	1	0	60	0	0	1	1	1	1	0	0
255	1	1	1	1	1	1	1	1	255	1	1	1	1	1	1	1	1	192	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	n	18	the subnet mask has 18 "1" bits (from the left - must be consecutive, can not have any 0 in between).																																
	Max # network	262,144	=2^18																																
	m	14	the number of 0s of the subnet mask (from the right).																																
	Max # of hosts	16,382	=2^14-2		At least 2 is reserved but it depends on the provider, sometimes more																																
IP Address with CIDR notation	129.107.58.60/18		The number next to the slash (18) must be consistent with the number of 1s in the subnet mask																																

TCP/IP problem

A computer in your network can not access external websites and you are asked to fix it. The following is the configuration:

- Router IP address: 173.32.2.62/27
- The computer setup:
  - internal IP address: 173.32.2.65
  - Subnet mask: 255.255.255.224
  - Default Gateway: 173.32.2.62

Find the problem.

TCP/IP Problem - Answer

Default Gateway	173.32.2.62/27																															
Subnet Mask	255.255.255.224																															
	Octet 4		173							Octet 3		32							Octet 2		2							Octet 1		62			
Dec value	128	64	32	16	8	4	2	1		128	64	32	16	8	4	2	1		128	64	32	16	8	4	2	1		128	64	32	16	8	4	2	1
173	1	0	1	0	1	1	0	1	32	0	0	1	0	0	0	0	0	2	0	0	0	0	0	0	1	0	62	0	0	1	1	1	1	1	0
255	1	1	1	1	1	1	1	1	255	1	1	1	1	1	1	1	1	255	1	1	1	1	1	1	1	1	224	1	1	1	0	0	0	0	0

Computer	173.32.2.65																															
Subnet Mask	255.255.255.224																															
	Octet 4		173							Octet 3		32							Octet 2		2							Octet 1		65			
Dec value	128	64	32	16	8	4	2	1		128	64	32	16	8	4	2	1		128	64	32	16	8	4	2	1		128	64	32	16	8	4	2	1
173	1	0	1	0	1	1	0	1	32	0	0	1	0	0	0	0	0	2	0	0	0	0		0	1	0	65	0	1	0	0	0	0	0	1
255	1	1	1	1	1	1	1	1	255	1	1	1	1	1	1	1	1	255	1	1	1	1	1	1	1	1	224	1	1	1	0	0	0	0	0

- The Default Gateway IP address in the computer matches the Router's IP address (173.32.2.62)
- The CIDR prefix /27 matches the subnet mask
- But the Computer IP address indicates that it is in a different network

Uniform Resource Locator

- **Uniform Resource Identifier (URI) -** any character string that identifies a resource.
  - Uniform Resource Locator – using the location to find the resource
    - http://example.com/resource?foo=bar#fragment
  - Uniform Resource Name – using the name to find the resource
    - urn:uuid:6e8bc430-9c3a-11d9-9669-0800200c9a66
- **URL (Uniform Resource Locator) -** a URI type that identify a resource by its location or by the means used to access it.
- **URL contains:**
  - Required: Protocol and domain name
  - Optional: Port #, path, others
  - Ex: https://policies.google.com/privacy?hl=en

How to Find IP Addresses

- To find your device's Public IP Address: The easiest way to find it from the internet.
- http://ip4.me/

- To find your device's Private IP Address (windows), DHCP, DNS, Default Gateway, and subnet mask
- C:\Users\budimans>ipconfig/all
- Windows IP Configuration
- :
- :
- IPv4 Address. . . . . :

- Try to find the public IP address of hosts connected to a WLAN.

```
Wireless LAN adapter Wi-Fi:

   Connection-specific DNS Suffix  . : fios-ro
   Description . . . . . . . . . . . : Intel(R
   Physical Address. . . . . . . . . : DC-8B-2
   DHCP Enabled. . . . . . . . . . . : Yes
   Autoconfiguration Enabled . . . . : Yes
   Link-local IPv6 Address . . . . . : fe80::6
   IPv4 Address. . . . . . . . . . . : 192.168
   Subnet Mask . . . . . . . . . . . : 255.255
   Lease Obtained. . . . . . . . . . : Saturda
   Lease Expires . . . . . . . . . . : Sunday,
   Default Gateway . . . . . . . . . : 192.168
   DHCP Server . . . . . . . . . . . : 192.168
   DHCPv6 IAID . . . . . . . . . . . : 8156240
   DHCPv6 Client DUID. . . . . . . . : 00-01-0
   DNS Servers . . . . . . . . . . . : 192.168
   NetBIOS over Tcpip. . . . . . . . : Enabled
   Connection-specific DNS Suffix Search List
```

- In your laptop, you can use: ipconfig/all
- In this example:
  - DHCP Server,
  - DNS Server, and
  - Default Gateway
- have the same private IP address, which indicate that all functions are collocated in the Wifi router.

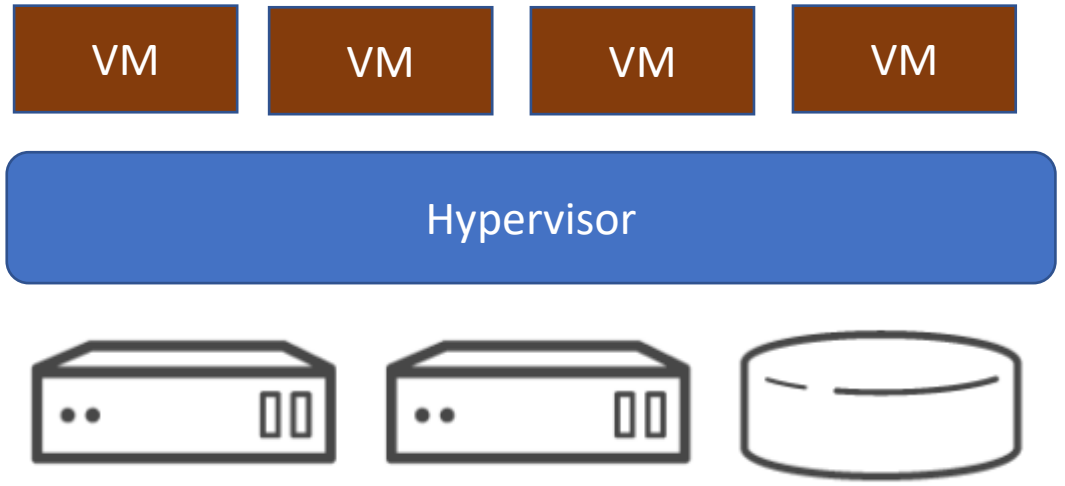
Virtual Machines

Virtualization

What is Virtualization?

not physically exist but made by software to appear to be so

The creation of a virtual resource (server, storage, etc)



Virtualization

Virtualization refers to the **creation of a virtual resource (an abstraction)** such as a server, desktop, operating systems, file, storage or network.

- software that manipulates hardware

The **main goal** of virtualization is to **manage workloads** by transforming traditional computing to make it more **scalable**.

Cloud computing uses virtualization extensively.

VMware is an example of platform virtualization software.

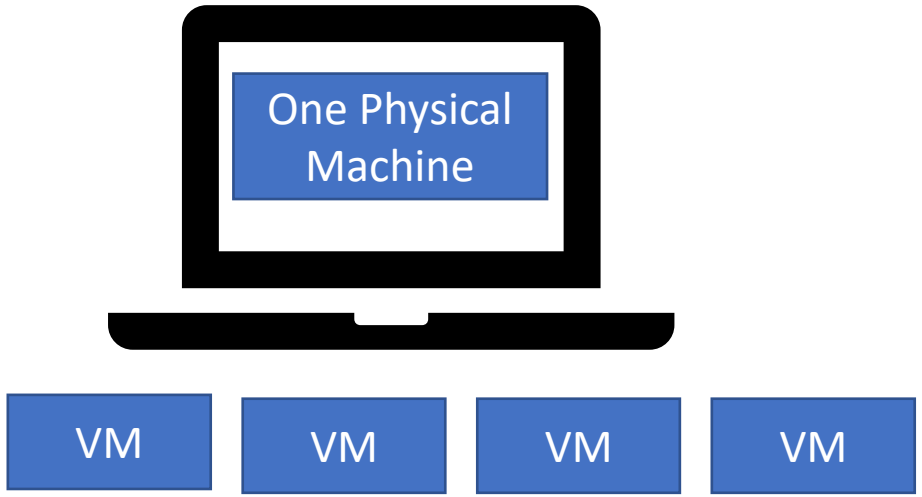
Virtualization technologies have grown significantly over the years

- Previously many SW vendors did not support virtualized environment
- Now it becomes a standard

Resource flexibility and scalability are key elements for rapid adoption.

- IT world shifts from one-to-one to many to one model
- running multiple virtual servers on one physical server.

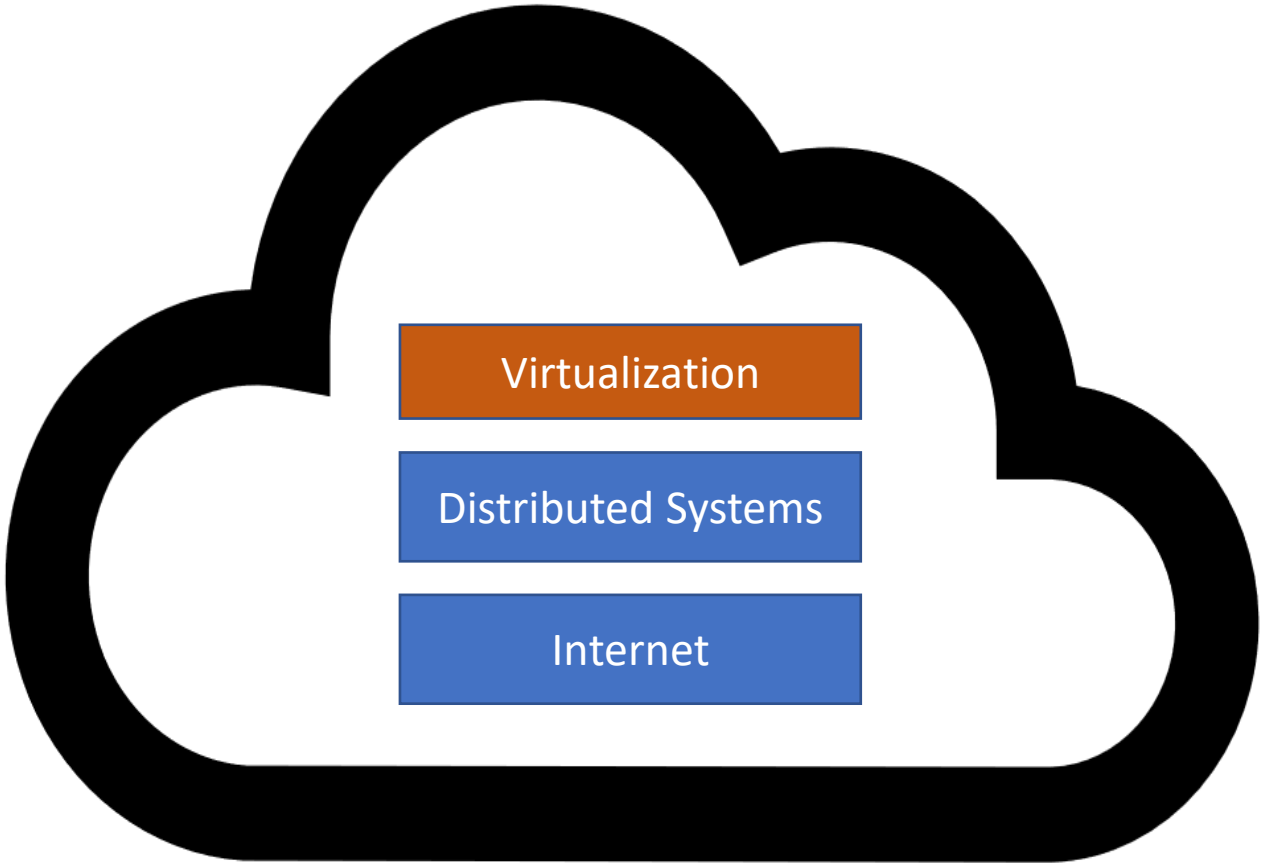
Example



Resources are shared

You can create VMs from your laptop by using Virtualbox

Virtualization In Cloud Computing



Virtualization is a key building block to cloud computing

- Used by Cloud providers to offer services
- It allows **scalability, elasticity, and on-demand**
- For example, when a cloud consumer requests a new server, the cloud provider provisions a new VM, no new physical hardware needs to be put in place to service the request

Virtualization makes Cloud computing more efficient and easier to manage

Benefits of Virtualization in Cloud Environment

Shared Resource

Increase Hardware Utilization (efficiency)

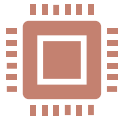
Cloud management enable quick creation of VMs

Elasticity

Resources can be adjusted dynamically based on workload

Resource pooling

Quick adjustment



Network and Application Isolation

Enhance network security, application agility, scalability and availability

Infrastructure Consolidation

Consolidate servers and infrastructure

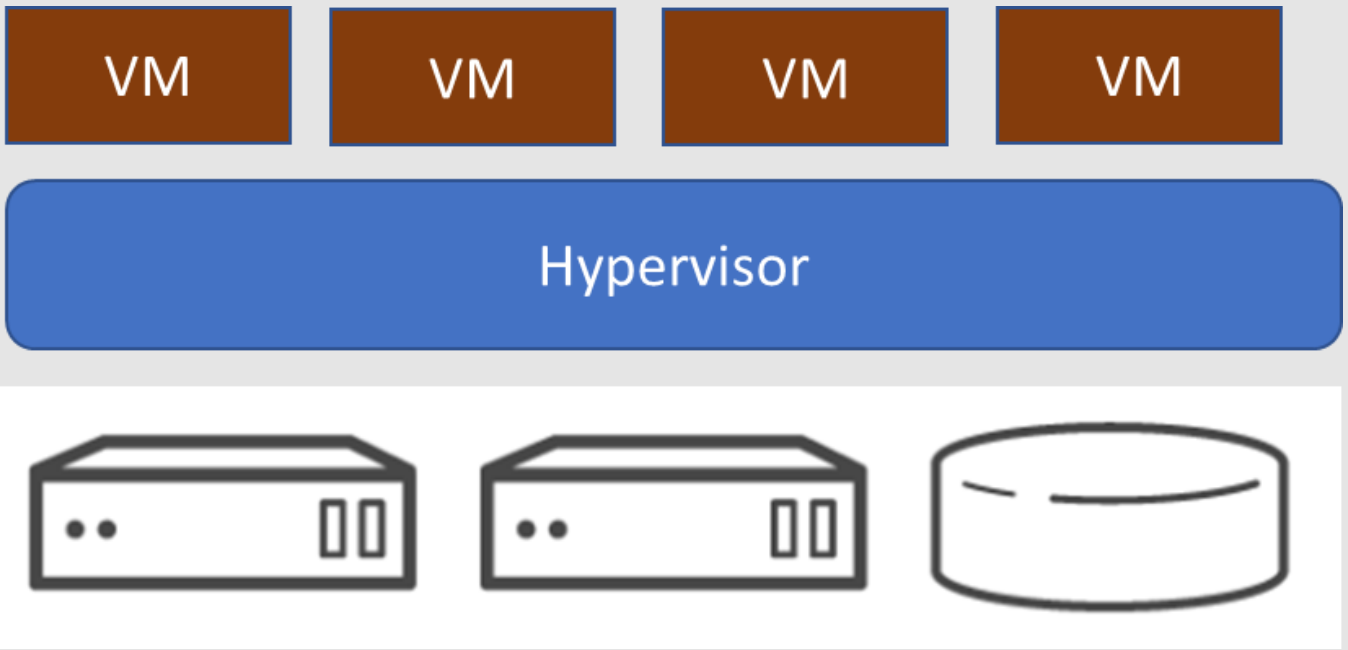
- Cost
- Energy savings
- Dedicated vs. Shared compute environment

Virtual Data Center Creation

Data Center infrastructure as a service

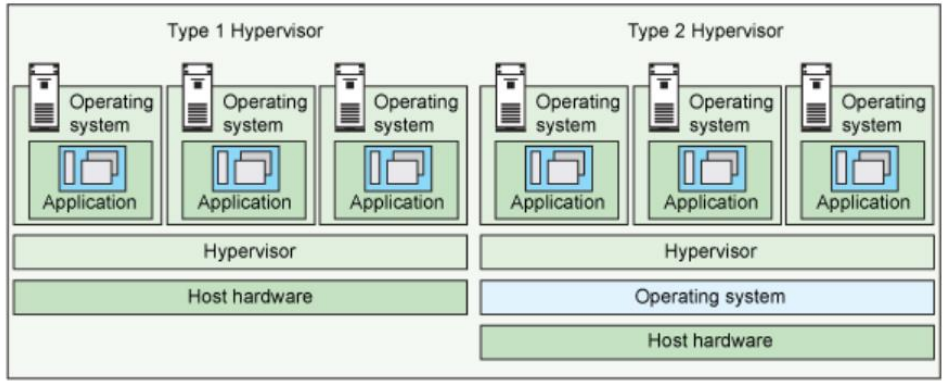
Hypervisor

Hypervisor



- **Hypervisor is a software that creates and manages the virtual infrastructure**, including virtual switch (vSwitch), virtual CPU (vCPU), virtual memory, virtual disks, and virtual machine.
- Allows **multiple OS** to run on one physical machine
- The computer **running the hypervisor** is defined as the "**host**" computer.
- **Virtual machines** running on the host are called "**guest**" machines
- Hypervisor manages the guest OS resources (memory, CPU, etc.)

2 types of Hypervisor



Type 1	Type 2
Hypervisor interfaces directly with hardware resources/**Bare metal**	Hypervisor is loaded on top of an already existing OS installation.
Boots before the OS	Hypervisor can not boot until the OS is loaded and operational
Less complex and less overhead	Less scalable and more complex to manage
Best choice for high performance, scalability, and reliability	
Most major virtualization distributors use this type	
Ex: VMware, Microsoft, Citrix, RedHat, Google	Ex: VirtualBox

Hypervisor Proprietary Vs. Open Source

Hypervisor	Organization	Proprietary/Open Source
Hyper-V	Microsoft	Proprietary
KVM	KVM Project	Open Source
vSphere/ESXi	Vmware	Proprietary
VirtualBox	Oracle	Open Source

- Proprietary Hypervisor: developed and licensed under an exclusive legal right of the copyright holder. It is created and distributed under a license agreement to the customer.

- Open-Source Hypervisor:
  - Provided at “no cost” and delivers similar basic functionality as a proprietary hypervisor
  - Open-source market is growing and advancing faster
  - Probably more secure since more tested (more users since free)
- Selection criteria:
  - Security and reliability
  - OS supported by the hypervisor
  - Staff familiarity

Random Access Memory (RAM for VM)

94

- The **more RAM** and the faster the RAM speed, the **better for virtualization host**
- Hypervisors have a virtual allocation table (VAT) that uses methods such as nested page tables or shadow page to map virtual memory to that of the host
- VMs often require more memory when starting up or when loading processes for the first time
- **Memory ballooning** comes into play when there are **not enough resources** available to handle new memory requests from VMs.
  - Ballooning **requests memory resources from other VMs**
  - Free space are loaned to the hypervisor

Virtualization – Memory technologies

Memory Ballooning

- Request memory from other VMs

Memory Bursting

- VMs can be configured with a min and max memory size (dynamic memory). Burst memory is the max the VM can utilize

Transparent Page Sharing

- A technology that deduplicates hypervisor memory allocated to VMs
- Several VMs may load the same data
- In Virtual Desktop Infrastructure (VDI) this is prevalent

Memory Compression

- When memory is entirely consumed, OS' are configured to dump data from memory to a page file located on a disk (much slower access)

Over-commitment

- Over-commitment ratio is lower than CPU over-commitment
- Generally 1.25:1
- Mostly unsafe to go beyond 1.5:1

Virtual Machine

- VMs can be moved to other hw platform
- The underlying hw can be upgraded while VM stay the same
- Less burden for IT maintenance
- Organization virtualize desktop known as Virtual Desktop Infrastructure (VDI)
  - End user login remotely
  - Effectively managed and secured

Virtual Disks

- A file that represents a physical disk drive to the virtual machine
- Vmware virtual machine disks (VMDKs) have extension .vmdk.
- Hyper-V virtual hard disks (VHD) have extention .vhdx

vNIC

- Network Interface Card allows physical computer to interact with other devices and VMs on the network
- A vNIC is associated with a physical NIC and allows a VM to communicate on the network

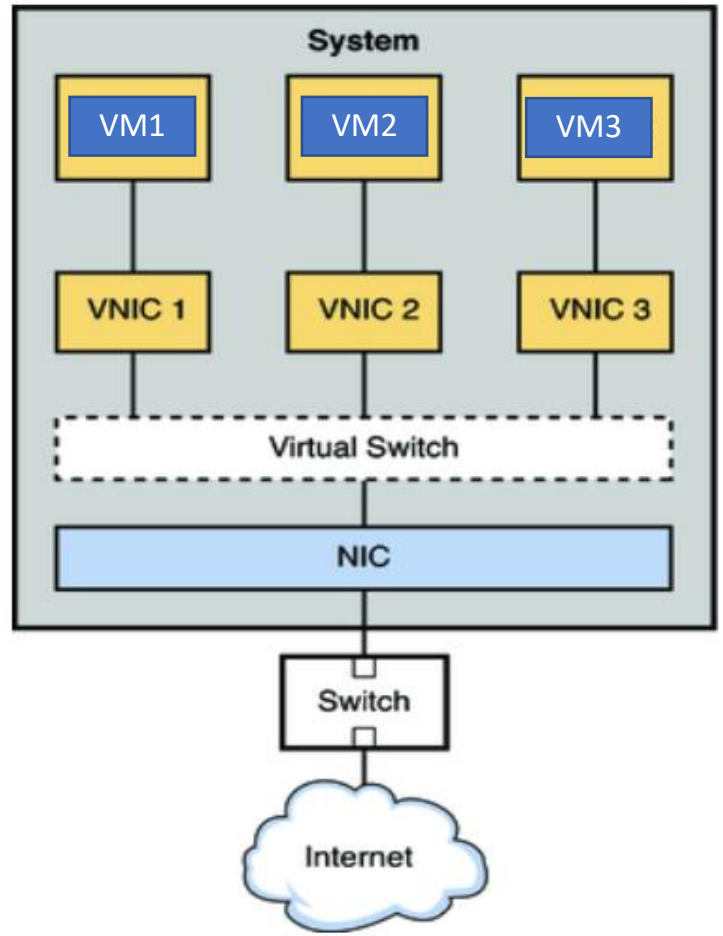
Virtual Switches

- vSwitch controls network traffic between VMs and the host computer and other devices

Over-commitment Ratio (CPU)

- It is possible to assign **more vCPUs** to VMs than **available physical CPU cores** in the hypervisor
- This is called **oversubscription or over-commitment**
- Over-commitment can result in contention for CPU resources when multiple machines attempt to utilize all their vCPUs at the same time
- It is generally safe to maintain an **over-commitment ratio 3:1** (3 vCPU per each physical CPU).
- Most of the time it is unsafe to assign more than 6 vCPU per physical CPU.
- The most important metric to monitor is the
  - CPU ready metric – measures the amount of time a VM has to wait for a physical CPU to become available
  - CPU utilization is also important to measure (each VM and the host)
    - If CPU utilization is high in one host but not others, maybe move some VMs

vNIC-vSwitch-VMs



Network Interface

- A **network interface** is the point of interconnection between a **computer** and a private or public **network**.
- Network Interface Card (NIC)
- A unique MAC address is tied to a NIC or vNIC

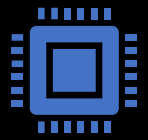
Managing Memory on a VM

3 things to be considered:

OS requirements	Application Requirements	VM neighbors
• Ensure that the amount of memory meets the minimum recommendation	• Consider the apps the VM will be running	• Consider other VMs which will be competing with the VM for memory

Type 1 Hypervisors – best practice- no additional SW runs in the host computer
Type 2 Hypervisors – **other apps may** be running on the host computer, so memory must be considered

How Memory is assigned



Static

Fixed amount is allocated to VMs

Host computer must have at least enough physical memory to support the VMs

Dynamic

Min, max is assigned to each VM

Allows over commitment

Can be enabled per VM basis

AWS Account

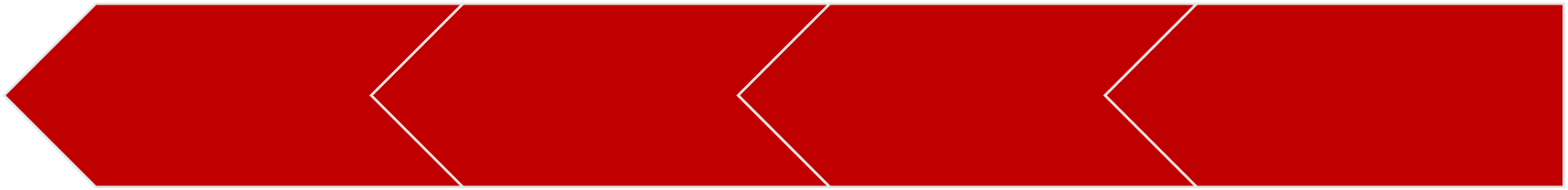
AWS Account

- Open AWS Account
  - We are using real aws account so there might be charges.
  - **You are responsible for payment of all charges.**
- There is limit for free-tiers and not all services used in this class is free.
- Spin up resources you need to do your assignments
  - We will try to use free-tiers as much as possible. If you have the account for a while, the free-tiers might have expired already.
- Always delete resources you no longer need
- General rule of thumb of deleting resources:
  - Start deletion from the last resource you created and move backward
  - Follow instructions they provide.
- Monitor cost and usages frequently.
- You may want to set “Budget”.
- **It is each student’s responsibility for any cost and payment of it. This is the requirement of this course.**

Don’t forget to delete AWS resources

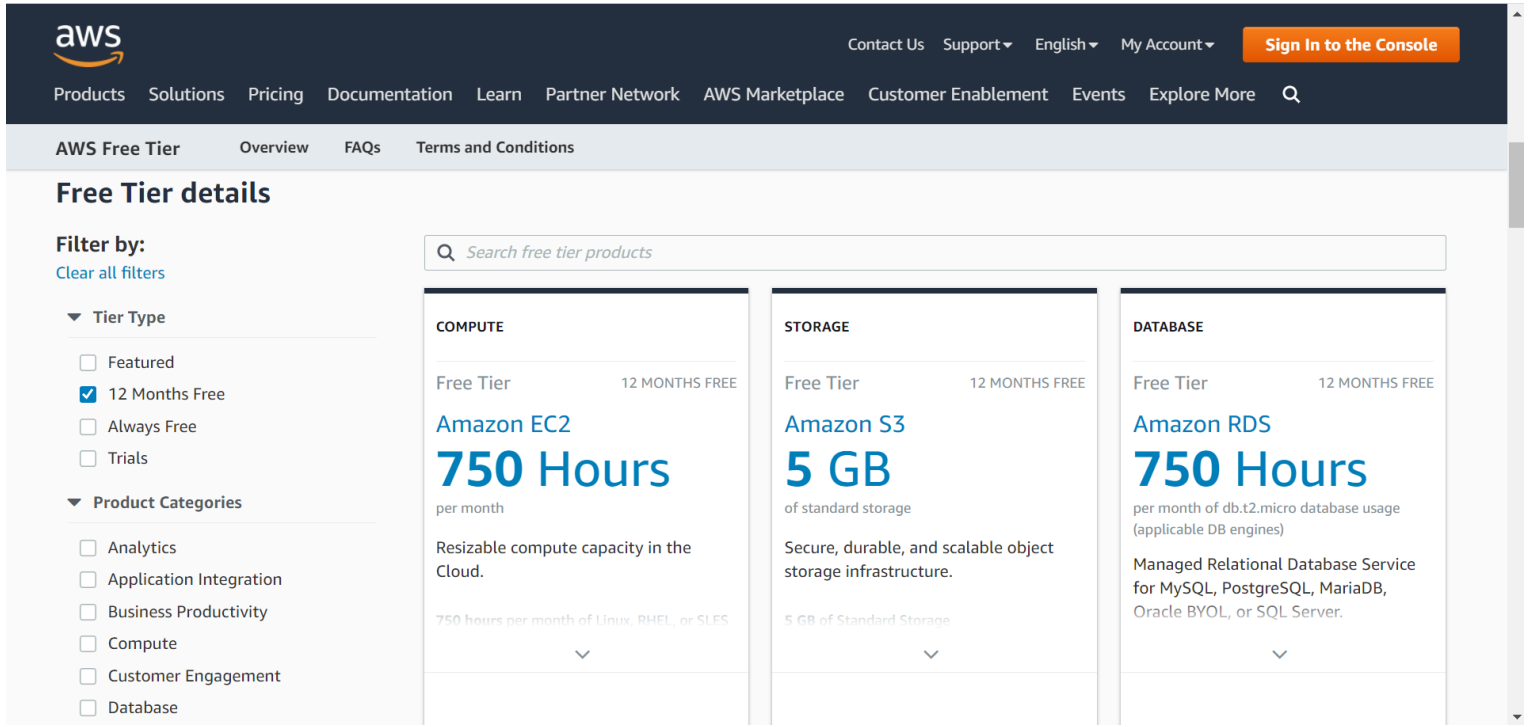
Create aws resources

Rule of thumb for deletion: delete the last aws resources created and move backward

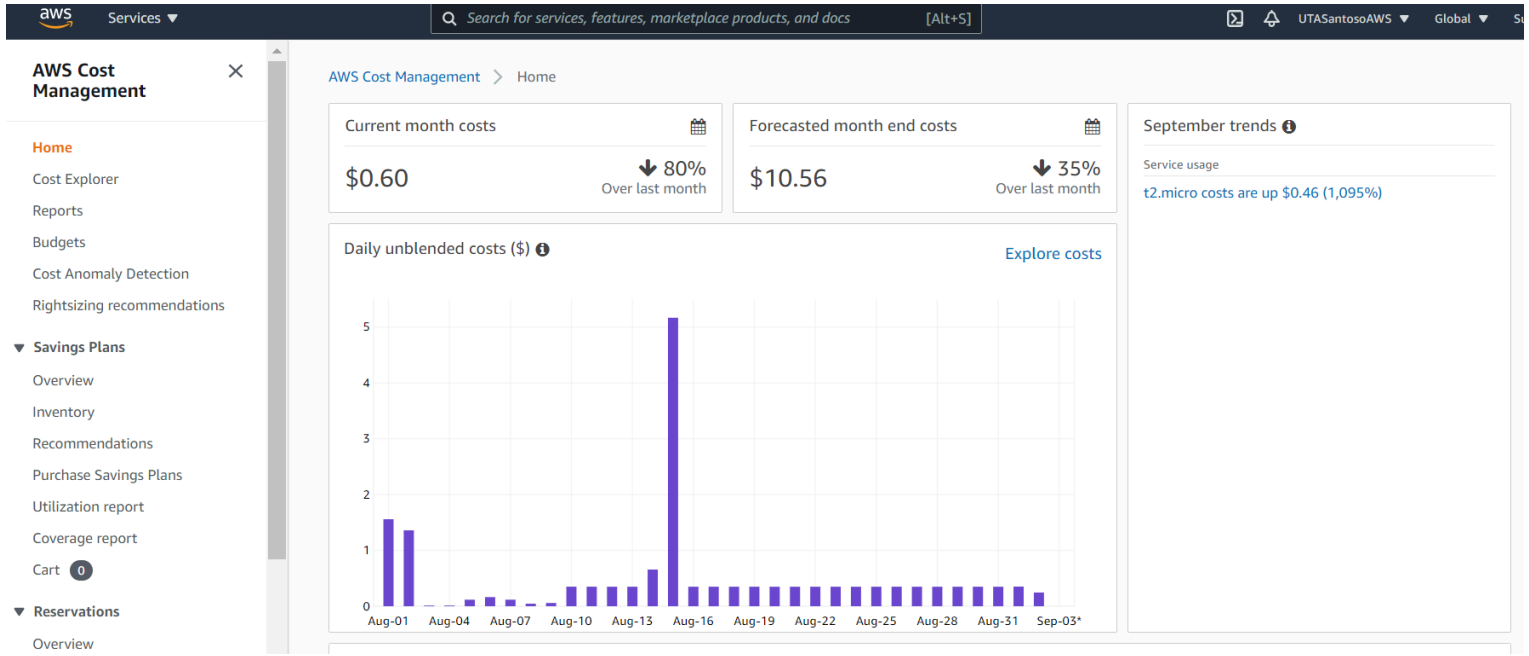


Always monitor costs

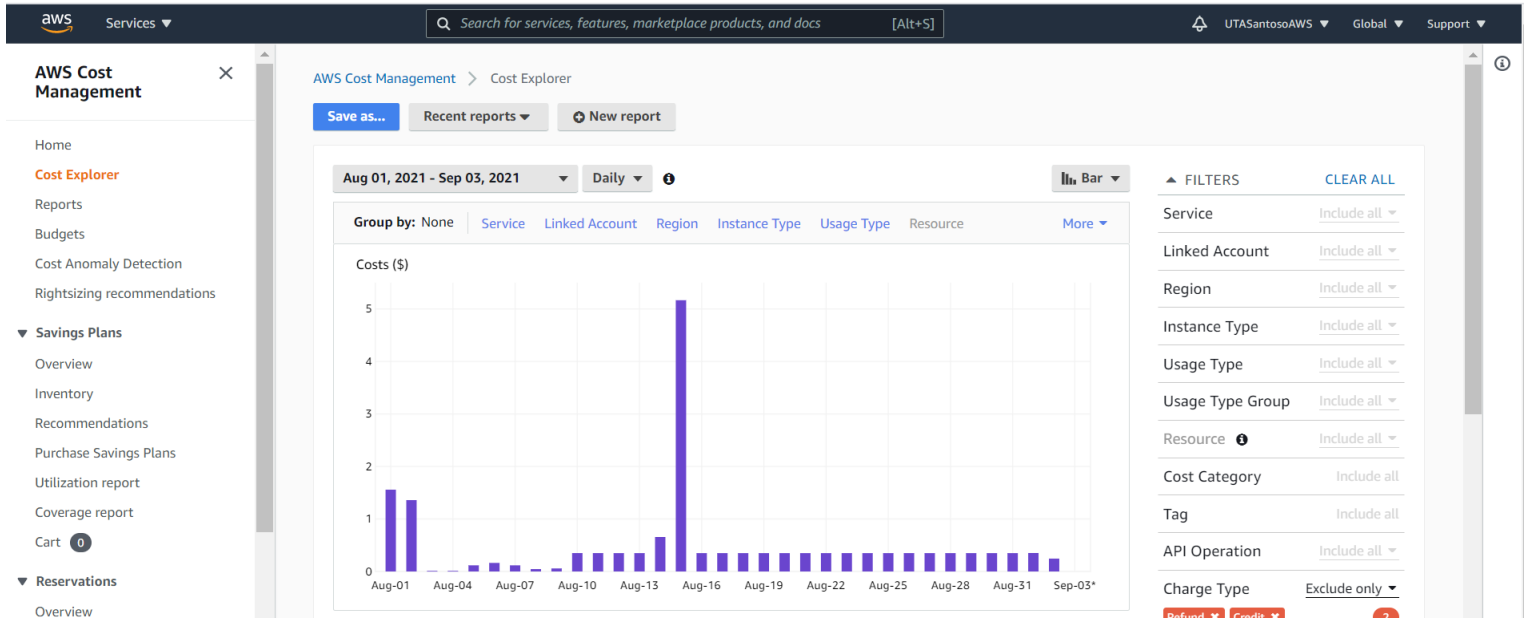
Free Tier


aws
Contact Us
Support
English
My Account
Sign In to the Console
Products
Solutions
Pricing
Documentation
Learn
Partner Network
AWS Marketplace
Customer Enablement
Events
Explore More
AWS Free Tier
Overview
FAQs
Terms and Conditions
Free Tier details
Filter by:
Clear all filters
Tier Type
Featured
12 Months Free
Always Free
Trials
Product Categories
Analytics
Application Integration
Business Productivity
Compute
Customer Engagement
Database
Search free tier products
COMPUTE
Free Tier
12 MONTHS FREE
Amazon EC2
750 Hours
per month
Resizable compute capacity in the Cloud.
STORAGE
Free Tier
12 MONTHS FREE
Amazon S3
5 GB
of standard storage
Secure, durable, and scalable object storage infrastructure.
DATABASE
Free Tier
12 MONTHS FREE
Amazon RDS
750 Hours
per month of db.t2.micro database usage (applicable DB engines)
Managed Relational Database Service for MySQL, PostgreSQL, MariaDB, Oracle BYOL, or SQL Server.

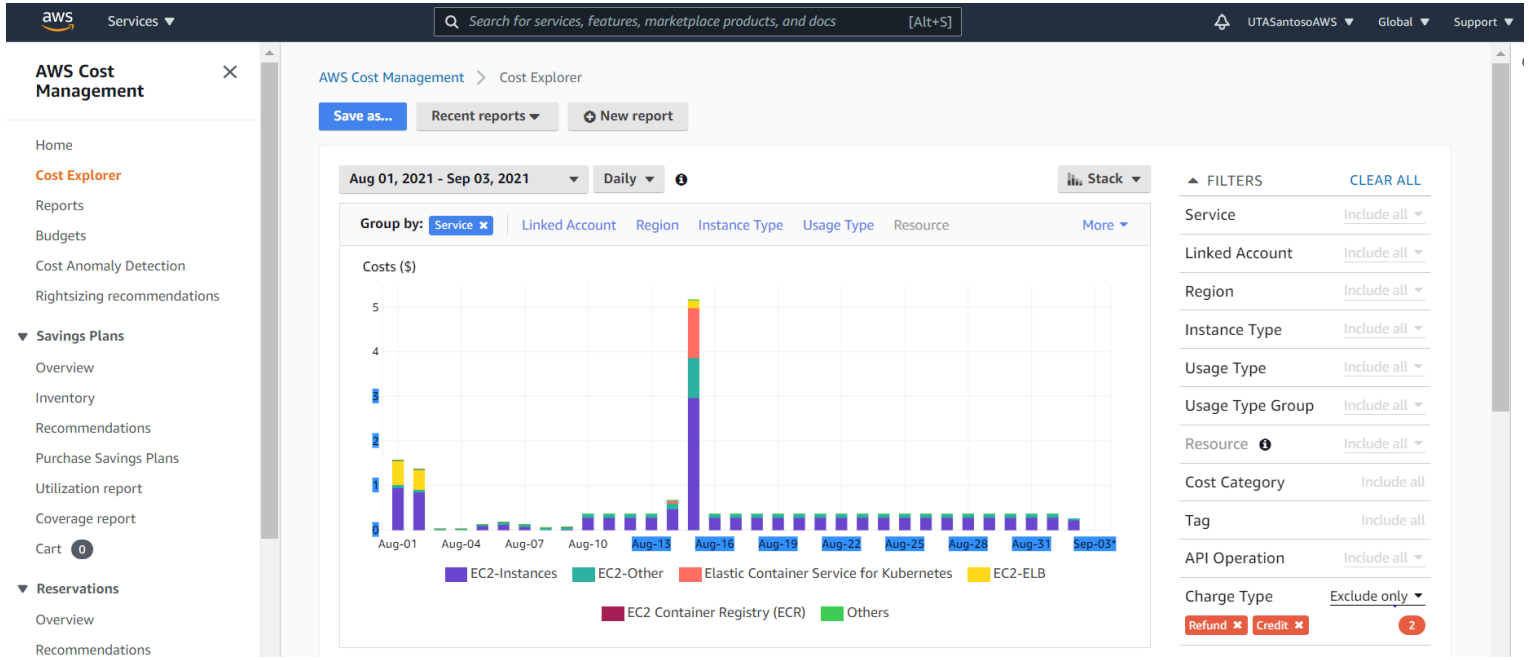
AWS Cost Management - Home



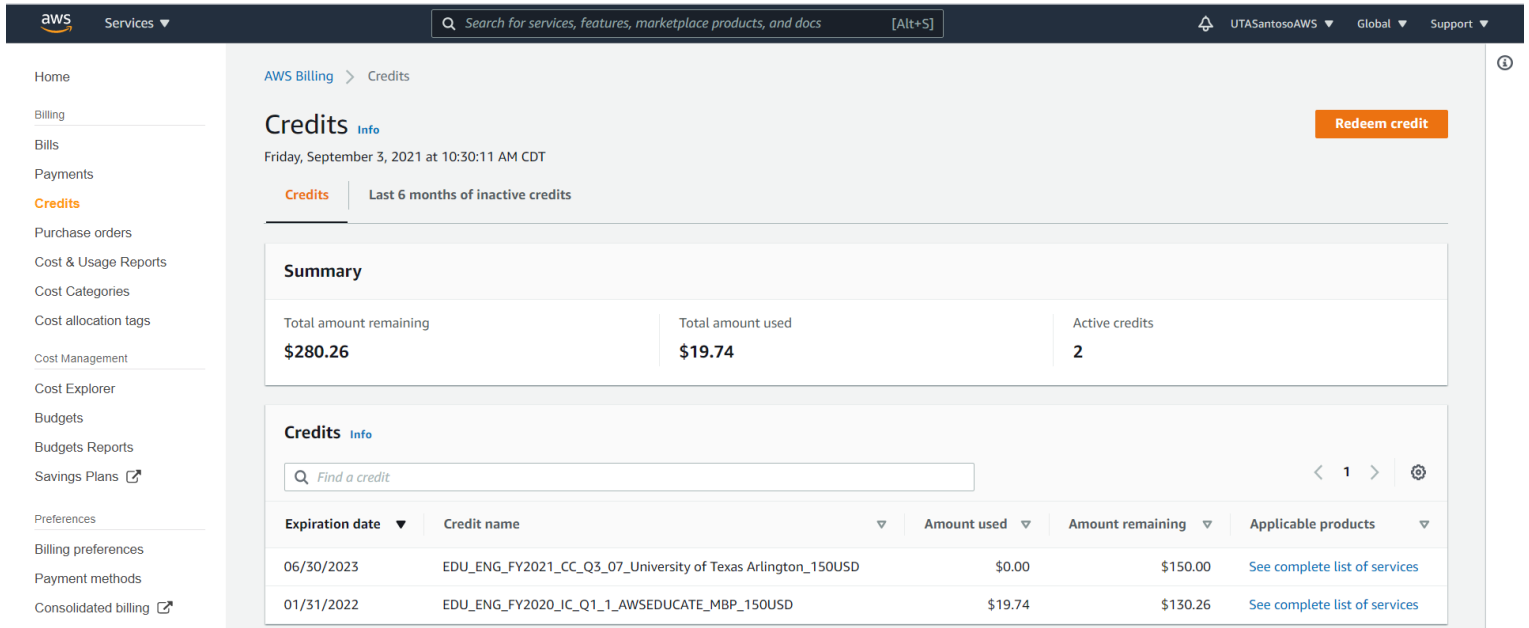
Check Cost Explorer



Cost Explorer - Services



Check frequently your Credits and amount used



https://aws.amazon.com/getting-started/hands-on/control-your-costs-free-tier-budgets/

AWS Cost Monitoring

Control your AWS costs

With the AWS Free Tier and AWS Budgets

In this tutorial you will learn how to control your costs while exploring AWS service offerings using the AWS Free Tier then using AWS Budgets to set up a cost budget to monitor any costs associated with your usage.

Whether you're looking for compute power, database storage, content delivery, or other functionality, AWS has the services to help you build sophisticated applications with increased flexibility, scalability and reliability. But how do you get started experimenting and building with AWS services while keeping your costs low or free?

The AWS Free Tier is a discount program that lets you gain free, hands-on experience with AWS products and services. All new AWS accounts include the Free Tier so you don't have to sign up for it, allowing you to try out the services you need to build your workloads from day 1. With over 80 services in the Free Tier, you can do lots of exploring at a reasonable cost, or even for free.

Monitoring your service usage and associated costs while you are exploring and scaling your usage of AWS is often cited as a top concern. To make sure you don't exceed the Free Tier usage thresholds and your overall budget, we recommend using AWS Budgets. AWS Budgets is cost control tool that allows you to create custom cost budgets that alert you when you exceed your budgeted threshold.

In the next few minutes, you will learn about the AWS Free Tier offering, discover how AWS Budgets monitors your Free Tier usage by default, and create a total monthly cost budget that alerts you when you exceed (or are forecasted to exceed) using AWS Budgets.

It is a best practice to create a total monthly cost budget for each AWS account you use. AWS Budgets has a Free Tier limit of 62 budget days per month, so creating a single budget falls within the AWS Free Tier limit. As the name implies, AWS Free Tier-eligible service usage is free.

About this Tutorial	
Time	10 minutes
Cost	Free Tier Eligible
Use Case	All
Products	AWS Budgets, AWS Free Tier
Audience	All
Level	Beginner
Last Updated	December 18, 2018

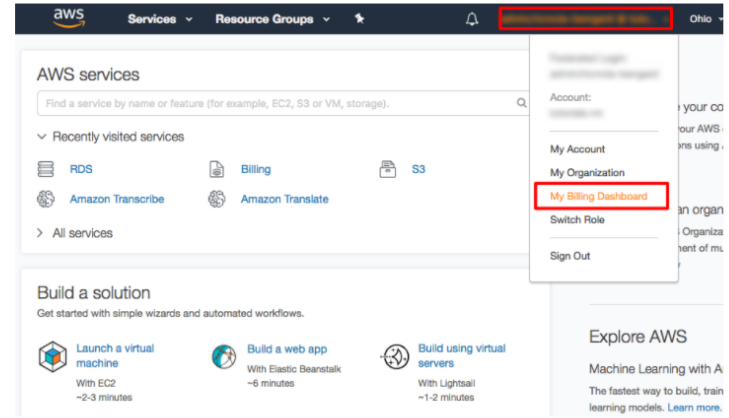
3. Review your spend and Free Tier usage

In this step you will use the AWS Billing Console to review your overall AWS spend and Free Tier usage.

a. Access the billing dashboard

After you have logged in to your account, from the account menu choose **My Billing Dashboard**.

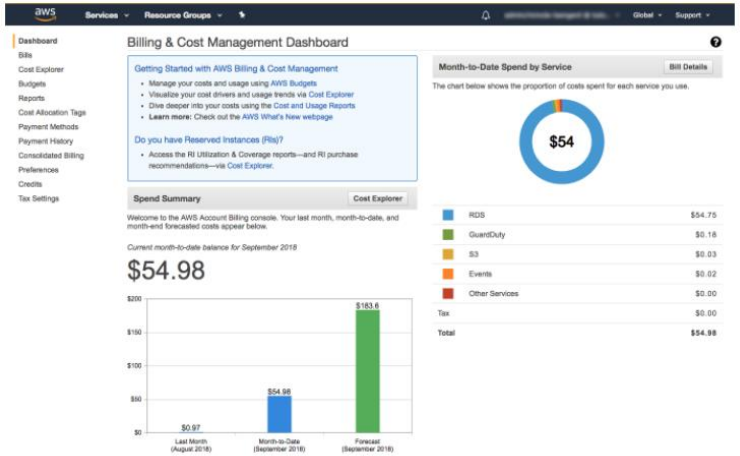
Go to the console page this step describes >>



b. Review your billing dashboard

Once you reach the **Billing & Cost Management Dashboard** page, you can view a summary of your month-to-date costs in the Spend Summary section, as well as a service-based breakdown in the Month-to-Date Spend by Service section.

Go to the console page this step describes >>



c. Analyze top free tier service usage

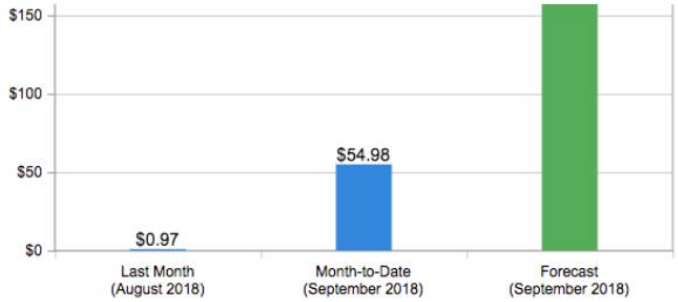
To see an overview of what part of your usage falls into the Free Tier, examine the **Top Free Tier Services by Usage** pane. This pane highlights the highest percentage of month-to-date usage against Free Tier limits categorized by service.

The screenshot to the right highlights Amazon S3 usage. This can be used as an example of how to use this information to analyze your usage.

The **Free Tier usage limit** column outlines the Free Tier discount. In this example, the Free Tier gives you the first 2,000 S3 Put Requests for free.

The **Month-to-date usage** column shows the percentage of the Free Tier benefit you have used this month. In this example, exactly 2,000 Put Requests to S3 have been made (100% of the Free Tier limit), which means any additional usage in excess of this limit will be billed at normal AWS prices.

Go to the console page this step describes >>



▸ Important Information about these Costs

Top Free Tier Services by Usage View all

Service	Free Tier usage limit	Month-to-date usage
AmazonS3	2,000 Put Requests of Amazon S3	100.00% (2,000.00/2,000 Requests)
AmazonEC2	750 hours of Amazon EC2 Linux t2.micro instance usage	29.60% (222.00/750 Hrs)
AmazonS3	20,000 Get Requests of Amazon S3	25.67% (5,133.00/20,000 Requests)
AmazonEC2	30 GB of Amazon Elastic Block Storage in any combination of General Purpose (SSD) or Magnetic	8.22% (2.47/30 GB-Mo)
AmazonS3	5 GB of Amazon S3 standard storage	0.71% (0.04/5 GB-Mo)

d. Access all your Free Tier usage

To dive deeper into your Free Tier-eligible usage, choose the **View all** button in the top right corner of the **Top Free Tier Service by Usage** widget.

Go to the console page this step describes >>

▸ Important Information about these Costs

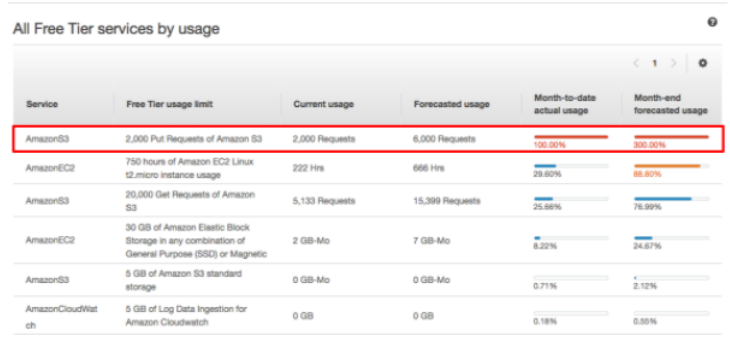
Top Free Tier Services by Usage — View all

Service	Free Tier usage limit	Month-to-date usage
AmazonS3	2,000 Put Requests of Amazon S3	100.00% (2,000.00/2,000 Requests)
AmazonEC2	750 hours of Amazon EC2 Linux t2.micro instance usage	29.60% (222.00/750 Hrs)
AmazonS3	20,000 Get Requests of Amazon S3	25.67% (5,133.00/20,000 Requests)
AmazonEC2	30 GB of Amazon Elastic Block Storage in any combination of General Purpose (SSD) or Magnetic	8.22% (2.47/30 GB-Mo)
AmazonS3	5 GB of Amazon S3 standard storage	0.71% (0.04/5 GB-Mo)

e. Analyze all your Free Tier usage

On the **All Free Tier services by usage** page, all of your usage for all services in the Free Tier are listed. In addition to your month-to-date actual usage, how much service usage you are forecasted to have by the end of the month is detailed in the **Month-end forecasted usage** column.

In the example in the screenshot to the right, note that your forecasted usage of S3 Put requests is 6,000. Exceeding the limit of the Free Tier generally results in a billable charge.



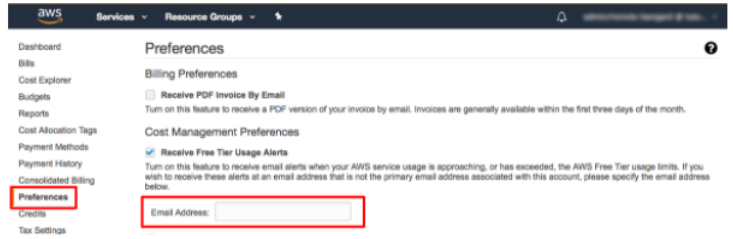
Go to the console page this step describes >>

f. Modify your AWS Free Tier Usage Limit email alerts

By default, most accounts are automatically opted in to receiving AWS Free Tier Usage Limit email alerts when their service usage exceeds 85% of a given Free Tier usage limit.

To change who gets these email alerts, choose **Preferences** from the left navigation bar.

To opt other people in to receiving Free Tier Usage Alerts, in the **Email Address** field add their email address and choose **Save preferences**.



Set up a cost budget

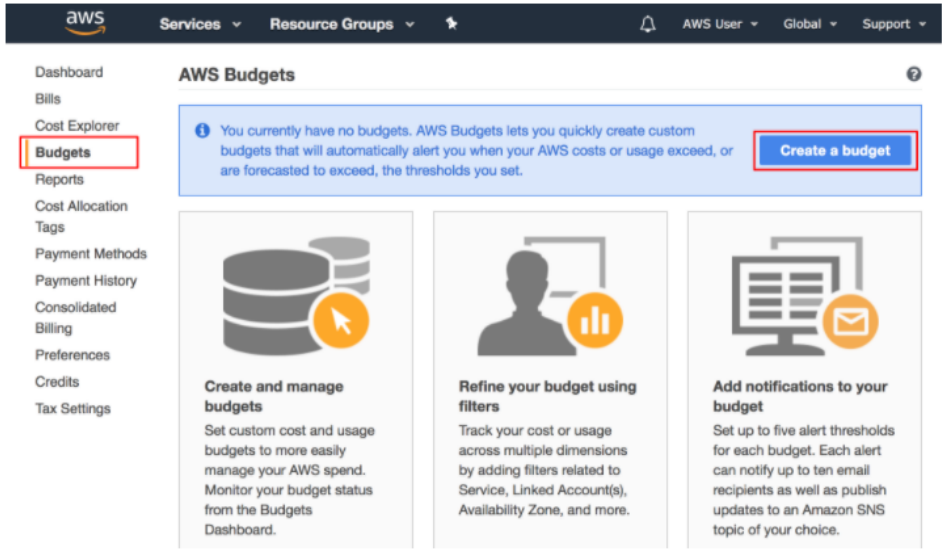
4. Set up a cost budget

In this step you will set up a cost budget in the AWS Billing Console using AWS Budgets. As part of your cost budget, you will set up three notifications: one for if your costs reach 50% of your budget, one for if your costs are forecasted to exceed your budget, and one if your costs do exceed your budget.

a. Create budget

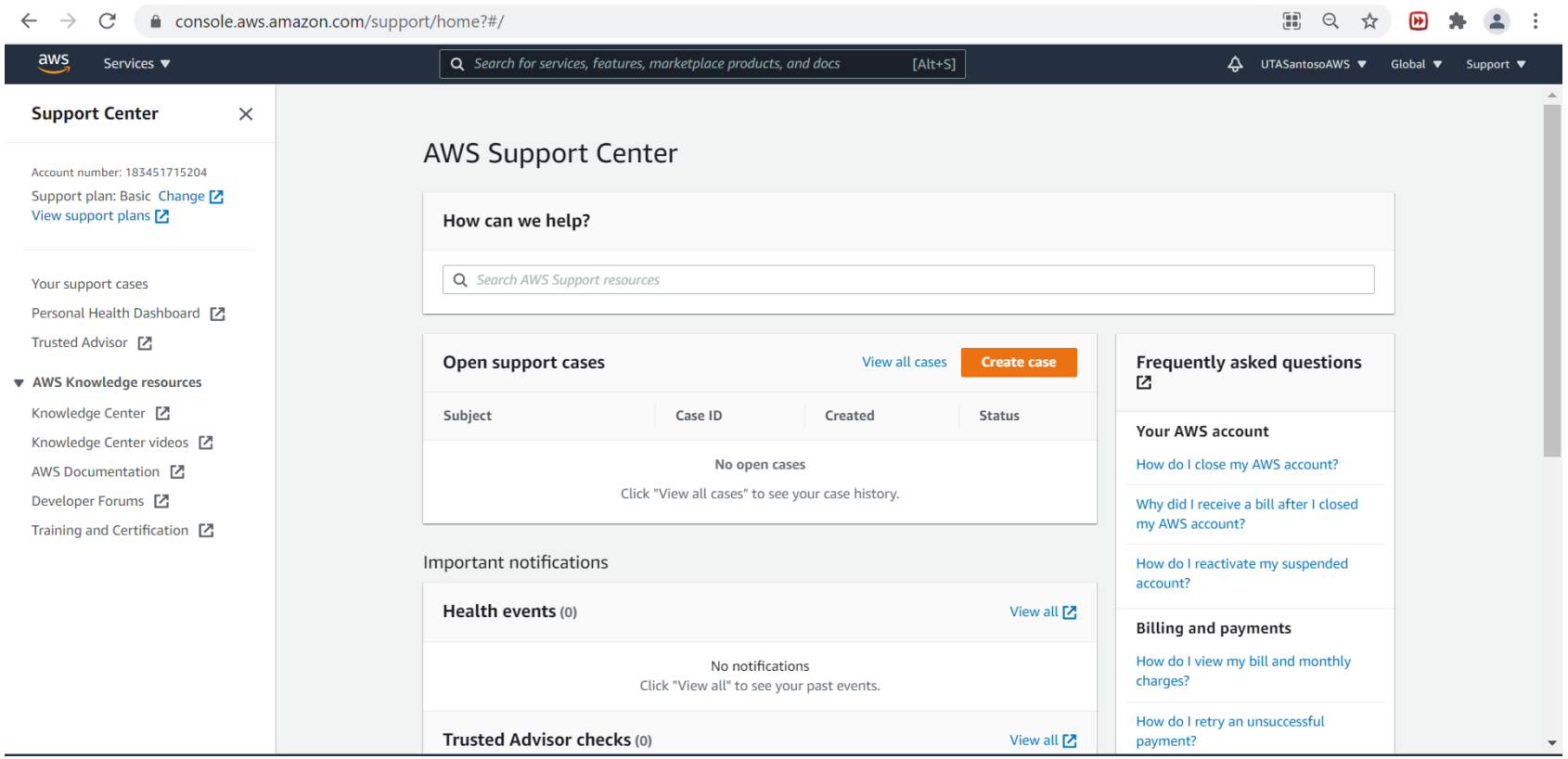
From the navigation menu on the left, select **Budgets** then choose **Create budget**. On the **Create budget** page, choose **Cost** as the **Budget Type**.

Go to the console page this step describes >>



https://aws.amazon.com/getting-started/hands-on/control-your-costs-free-tier-budgets/

Aws Support



https://console.aws.amazon.com/support/home?#/

End Lecture 2