Computer Systems

Bits, Bytes and Words

- Elemental information on a computer is a *binary digit* (bit)
  - All information is broken down into a series of zeros and ones
  - Abbreviated with lower case “b”
- Convenient to group bits into sets of eight, called byte
  - Byte can have 255 different values ($2^{8)}$
  - Abbreviated with upper case “B”
- Bytes (or bits) may be grouped into words
  - Typically 32 bits (4 bytes) or 64 bits (8 bytes)

Powers of two

- Because of extensive use of binary, units or values are often organized into groups based on the powers of two

Power	Value	Binary			
2^0	1	1	2^8	256	10000000
2^1	2	10	2^9	512	100000000
2^2	4	100	2^{10}	1024	1000000000
2^3	8	1000	2^{11}	2048	1000000000
2^4	16	10000	2^{12}	4096	10000000000
2^5	32	100000	2^{13}	8192	100000000000
2^6	64	1000000	2^{14}	16384	1000000000000
2^7	128	10000000	2^{15}	32768	10000000000000

- To convert from decimal, just keep subtracting out…
  - 97 = 64 + 32 + 1 = 26 + 2+5 + 20 = 1100001

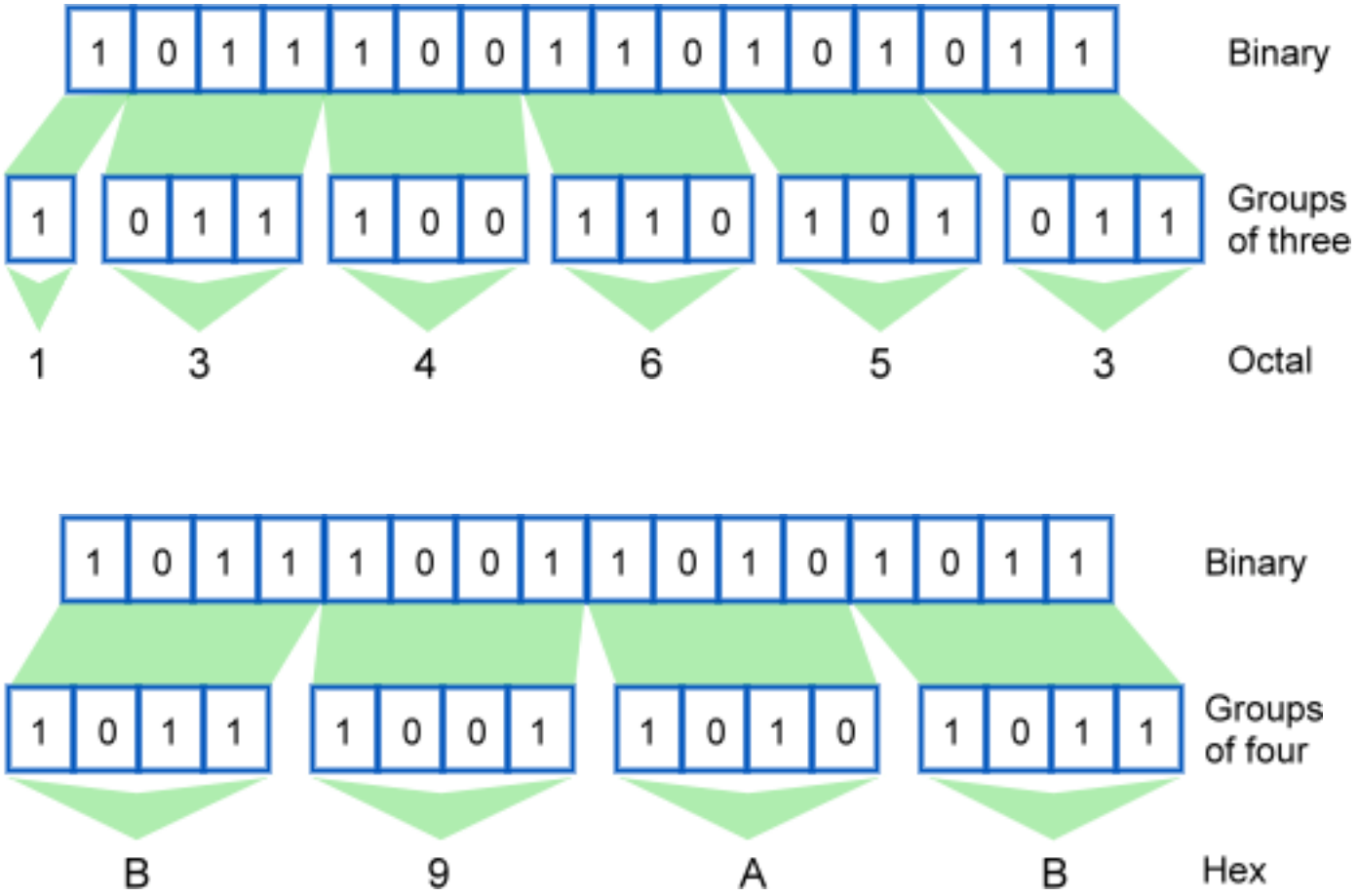
Octal and Hex

- Binary is a pain to write as its long
- For convenience, we frequently group into sets of three or four bits.
  - Groups of three gives us eight values (0-7) and its called octal
  - Groups of four give us 16 values (0-15) and its called hexidecimal or hex.
- Octal numbers are often indicated by leading zero, e.g. 056
- Hex numbers indexed by following with 0x, e.g. 0x56

The Yuki language in California and the Pamean languages in Mexico have octal systems because the speakers count using the spaces between their fingers rather than the fingers themselves.

Decimal	Binary	Octal	Hex
0	00000	0	0
1	00001	1	1
2	00010	2	2
3	00011	3	3
4	00100	4	4
5	00101	5	5
6	00110	6	6
7	00111	7	7
8	01000	10	8
9	01001	11	9
10	01010	12	A
11	01011	13	B
12	01100	14	C
13	01101	15	D
14	01110	16	E
15	01111	17	F
16	10000	20	10
17	10001	21	11
18	10010	22	12
19	10011	23	13

Converting from binary



What are the main pieces?

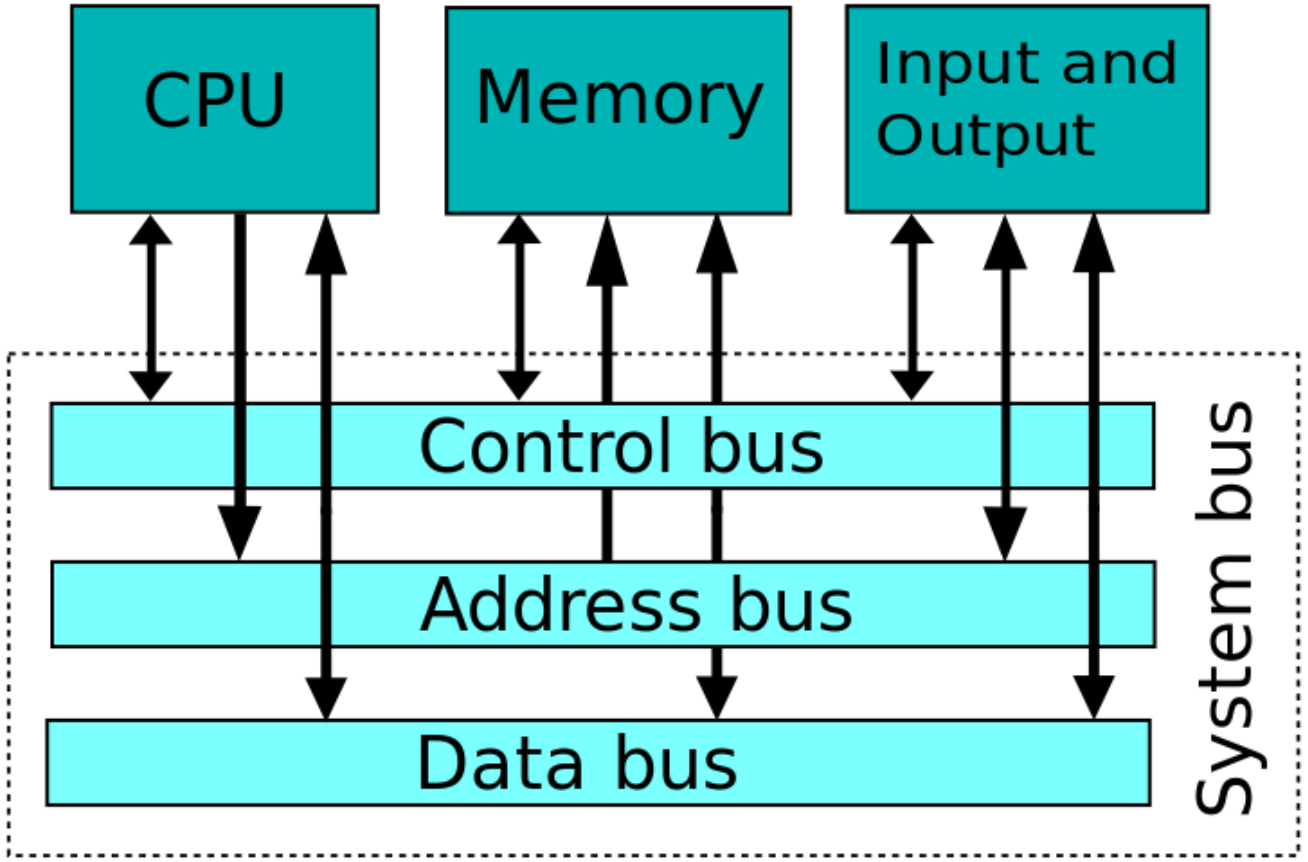
- Computing elements
  - Servers
  - Laptops
  - Desktops
- Networking
- Data centers

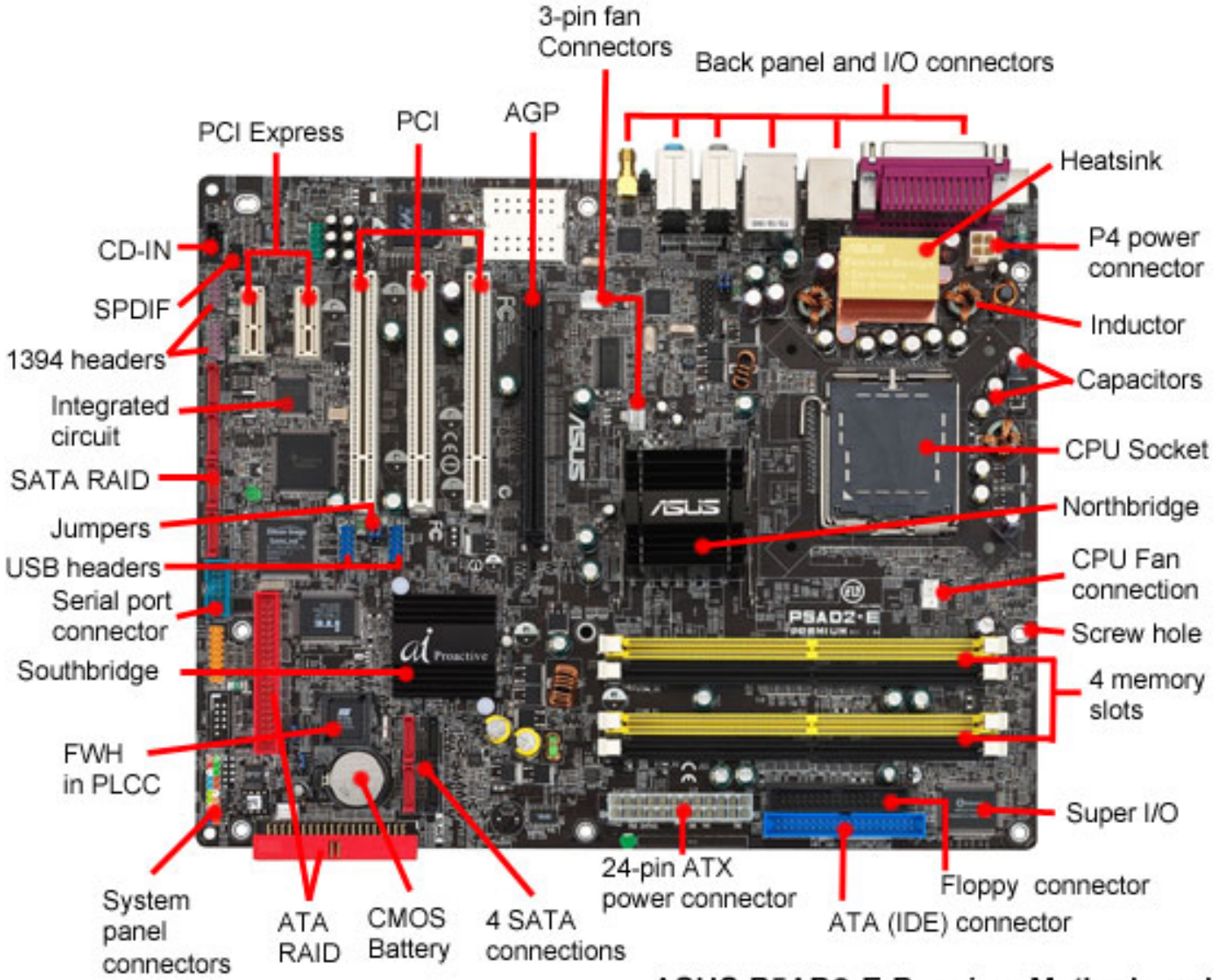
What do we care about

- Performance
  - Time to complete a desired task
  - Time to complete a set of tasks
  - Number of users that can work at the same time…
- Non-functional behavior
  - Reliability, maintainability, …
- Costs
  - Capital expenses (CapEx)
    - cost of acquisition
  - Operational costs (OpEx)
    - how much power does it take, what are the support costs
  - Total cost of ownership (TCO)
    - CapEx + OpEx

Components of a Computer

- processor to interpret and execute programs
- memory to store both data and programs
- mechanism for transferring data to and from the outside world.



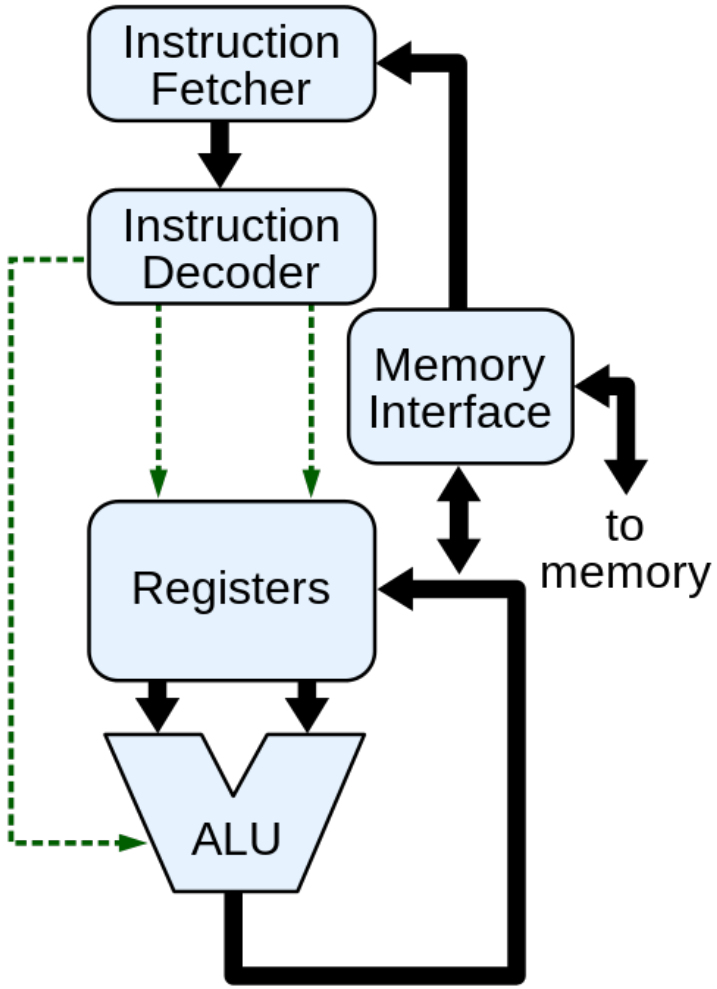


ASUS P5AD2-E Premium Motherboard

http://www.computerhope.com

Central Processing Unit (CPU)

- Executes instructions
  - Typical form operation with locations in memory for inputs and outputs
- Performance is driven by the number of instructions you can execute in a given time
- Factors are:
  - Cost of executing the operation
  - How long it takes to get the inputs and store the outputs
  - How soon before you can start the next instruction (overlap)

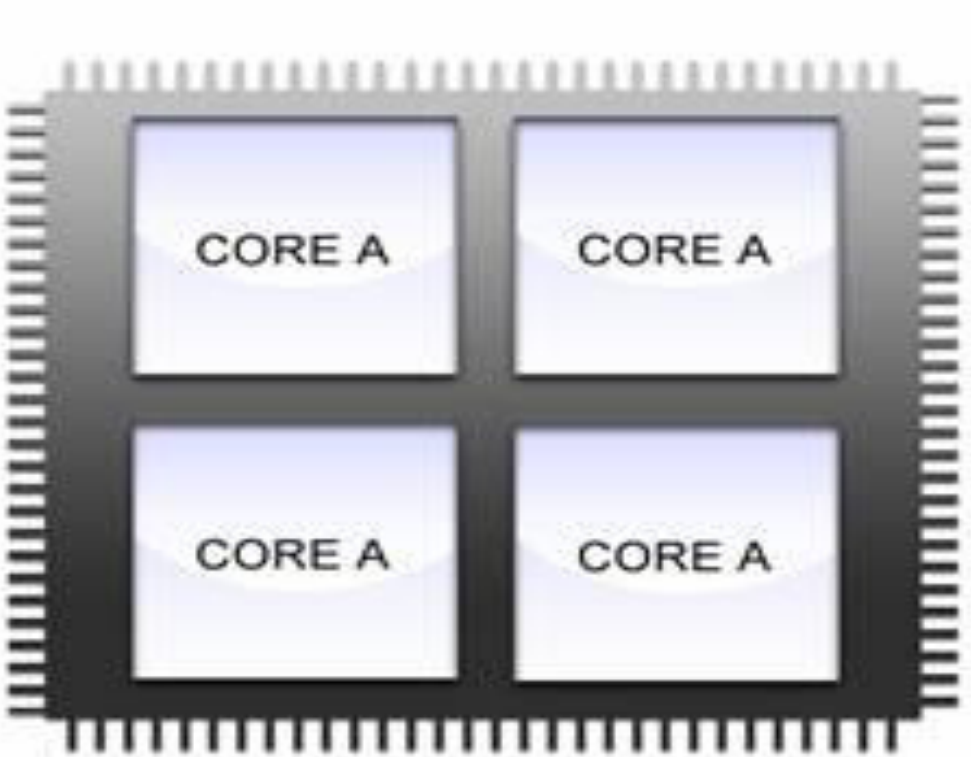


Problems with Single Core

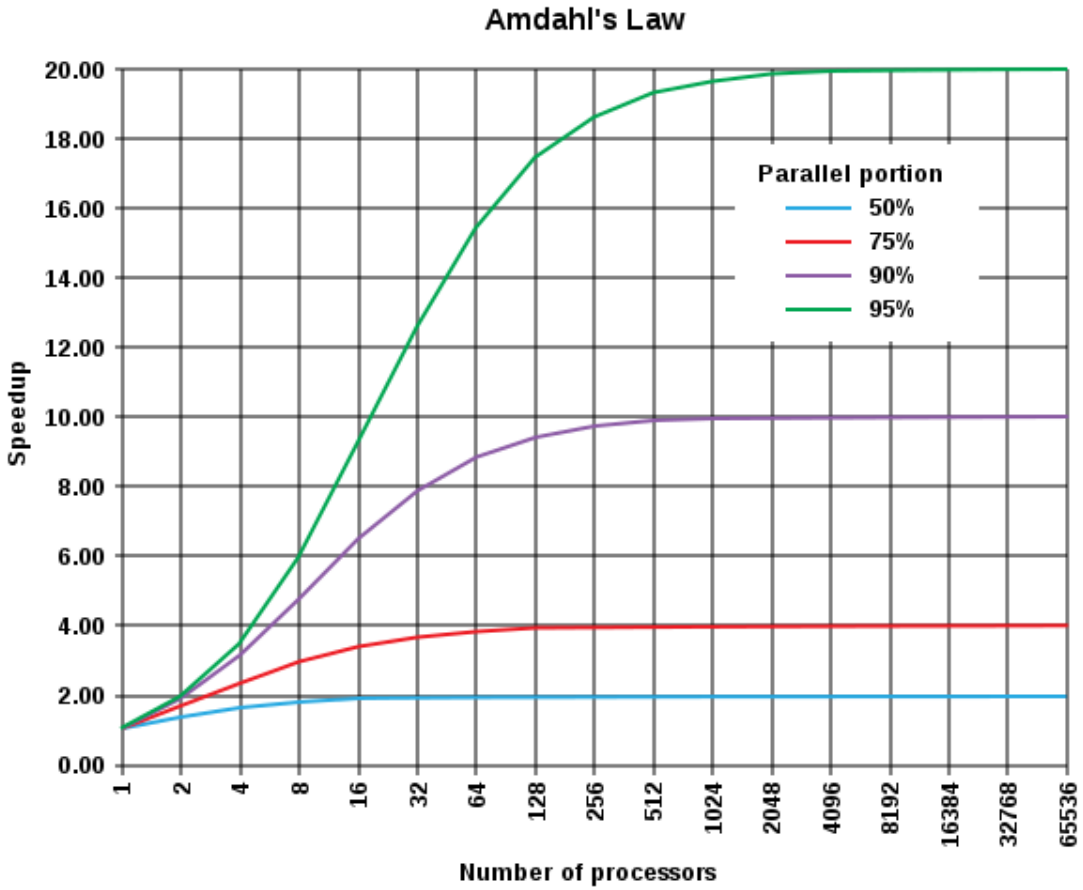
- To execute the tasks faster you must increase the clock time.
- Increasing clock times increases power consumption and heat dissipation to extremely high levels, making the processor inefficient.

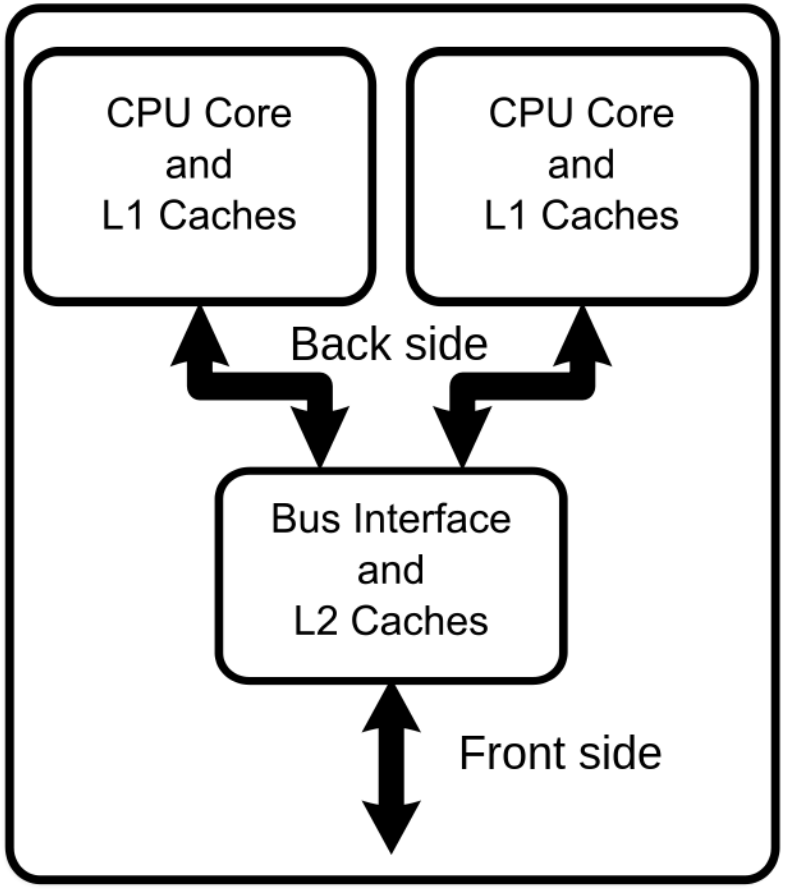
Multi Core Processors

- Creating two cores or more on the same Die increases processing power while limiting clock speeds at an efficient level.
- 2 cores can execute instructions equivalent to a single core processor running at twice the clock speed with less energy.

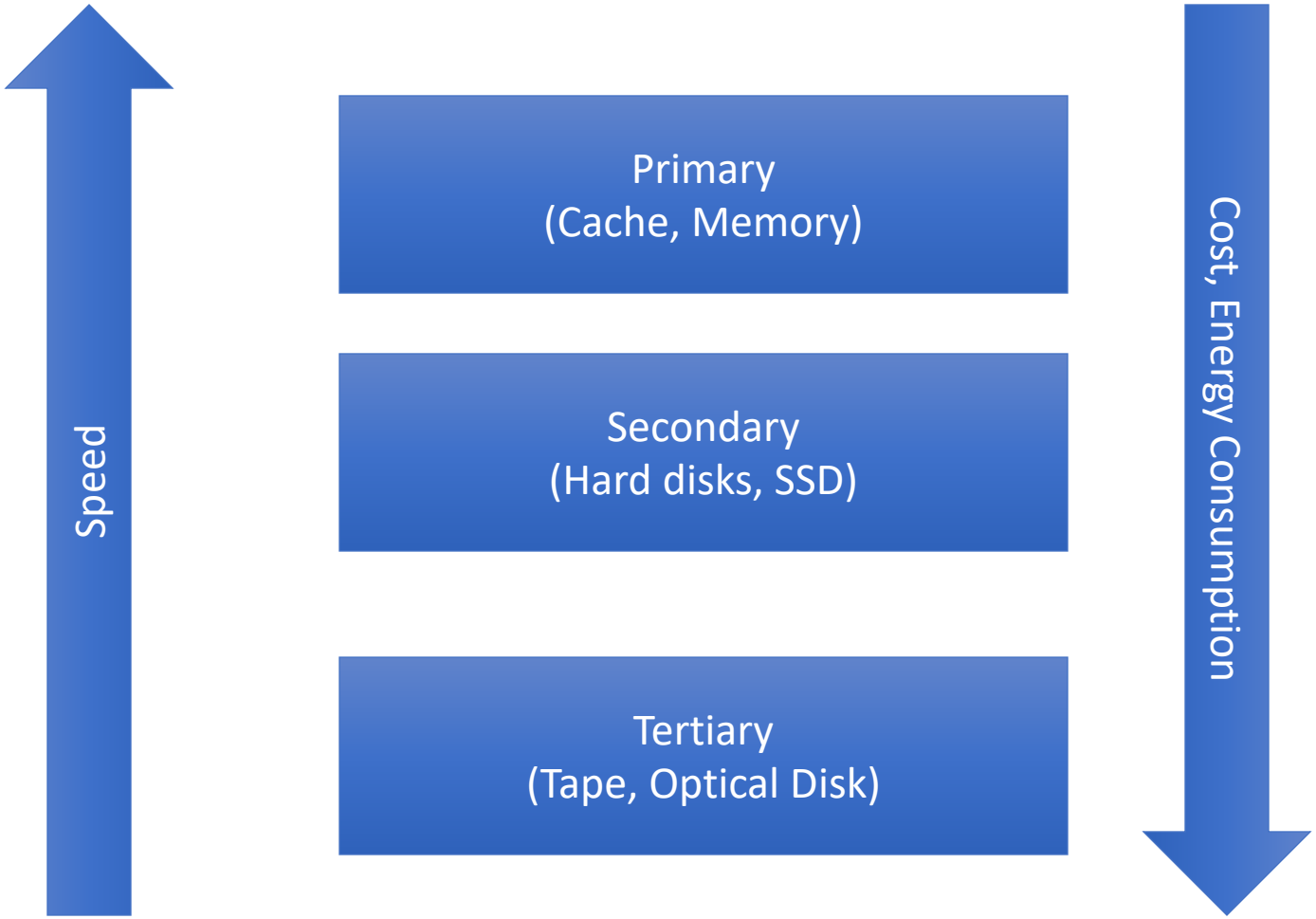


Performance Limits





Computer Storage Hierarchy – feeding the processor

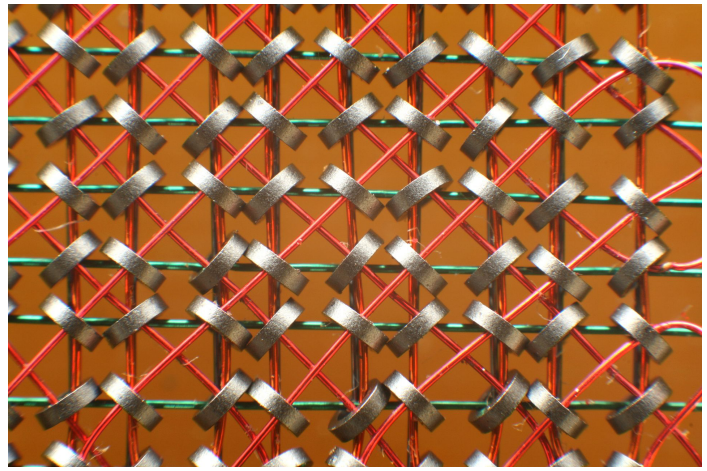


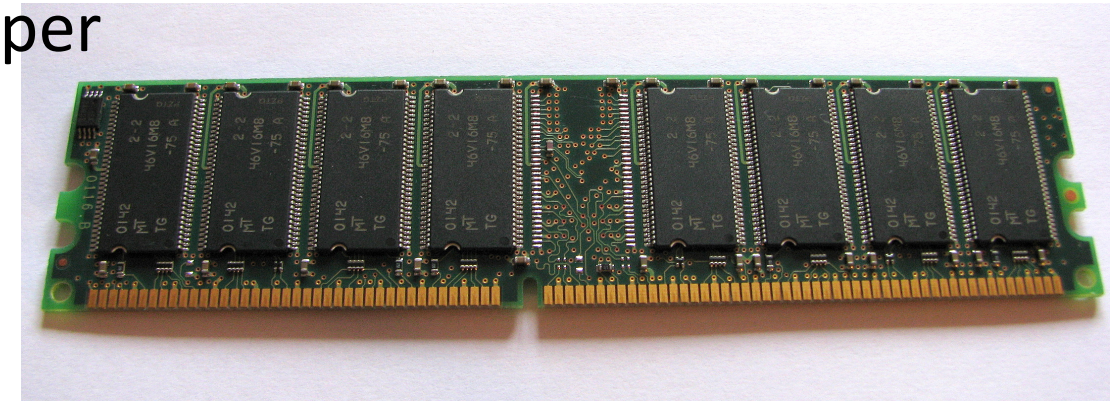
Storage Performance

- Basic operation is to provide access to a “chunk” of data
  - Access may be read, write, update
  - Chunk is named by an “address”, typically a number
  - Chunk size will vary depending on the technology
- Bandwidth
  - How quickly can you move data from one point to another
  - Measured in speed (like miles per hour, but size per unit time)
- Latency
  - Time between information is requested and when you have access to it
  - Measured in time

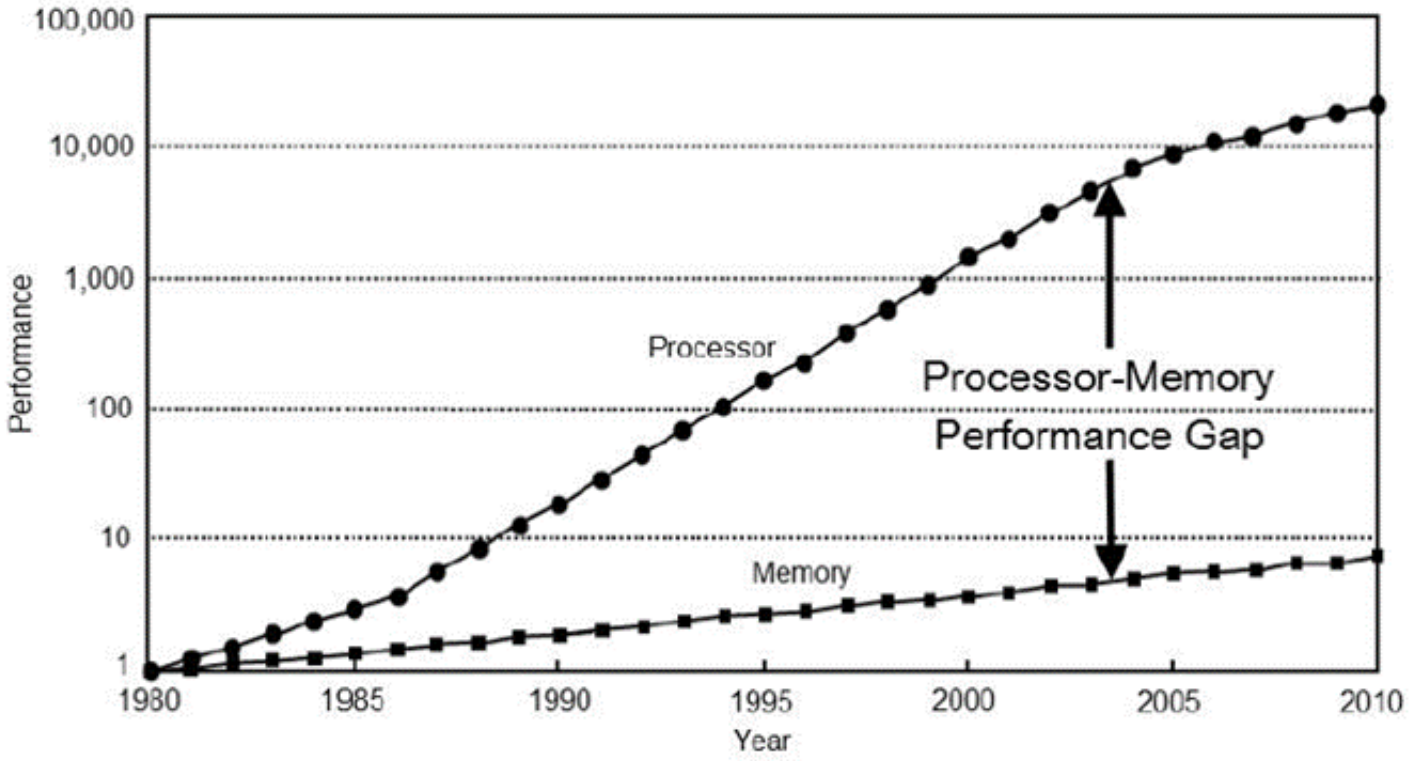
Random Access Memory (RAM)

- Current technology is called
  - Double data rate synchronous dynamic random-access memory (DDR)
- In DRAM, data is lost unless it is continuously refreshed
  - Hence the dynamic part
- Current memory size is up to 64GB per module
- Peak transfer rate of 19200 MB/sec with 12.5 ns latecy



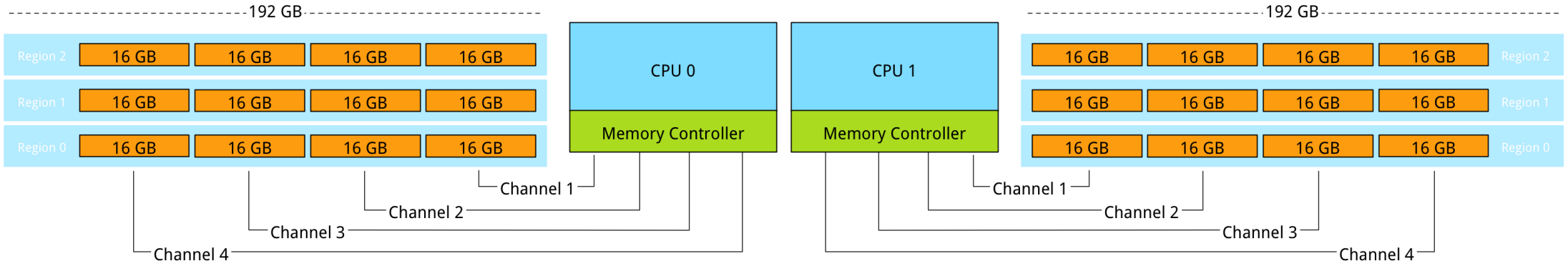


Processors get faster, faster then memory…



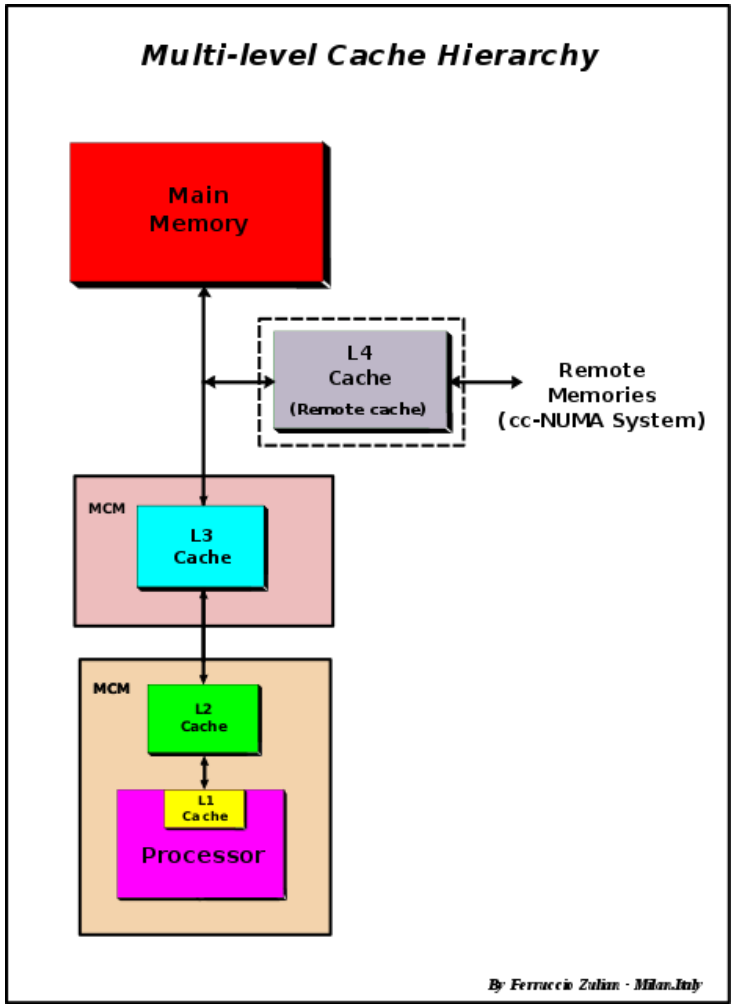
Solution: Parallelism

- Have multiple memories (banks) and process multiple requests at once
- Only impacts bandwidth, not latency
- Need to figure out how to keep all modules busy



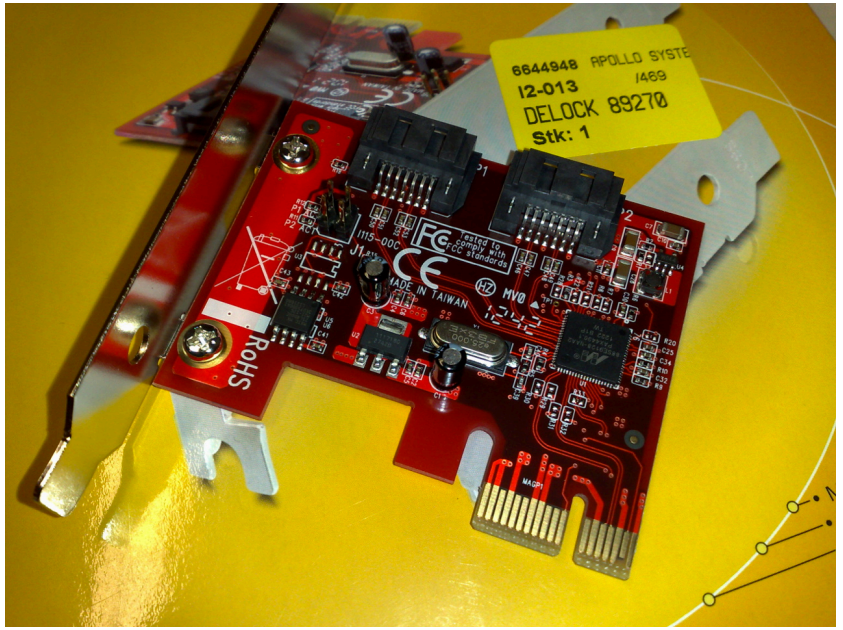
Solution: Caching

- Add faster memory to system
  - Implemented using different design approach (static RAM)
  - More expensive, consumes more power, physical limitations
- Cache size will be smaller then main memory
  - Must be organized differently as we will need to keep track of relationship between copies in cache and where in main memory it came from
- Processor operates out of cache, rather then main memory
- Need to figure out when to move stuff in and out of cache: cache policy
  - When to add new items, what items to add, what items to remove
  - Goal it to avoid “cache misses”



Disk Interfaces

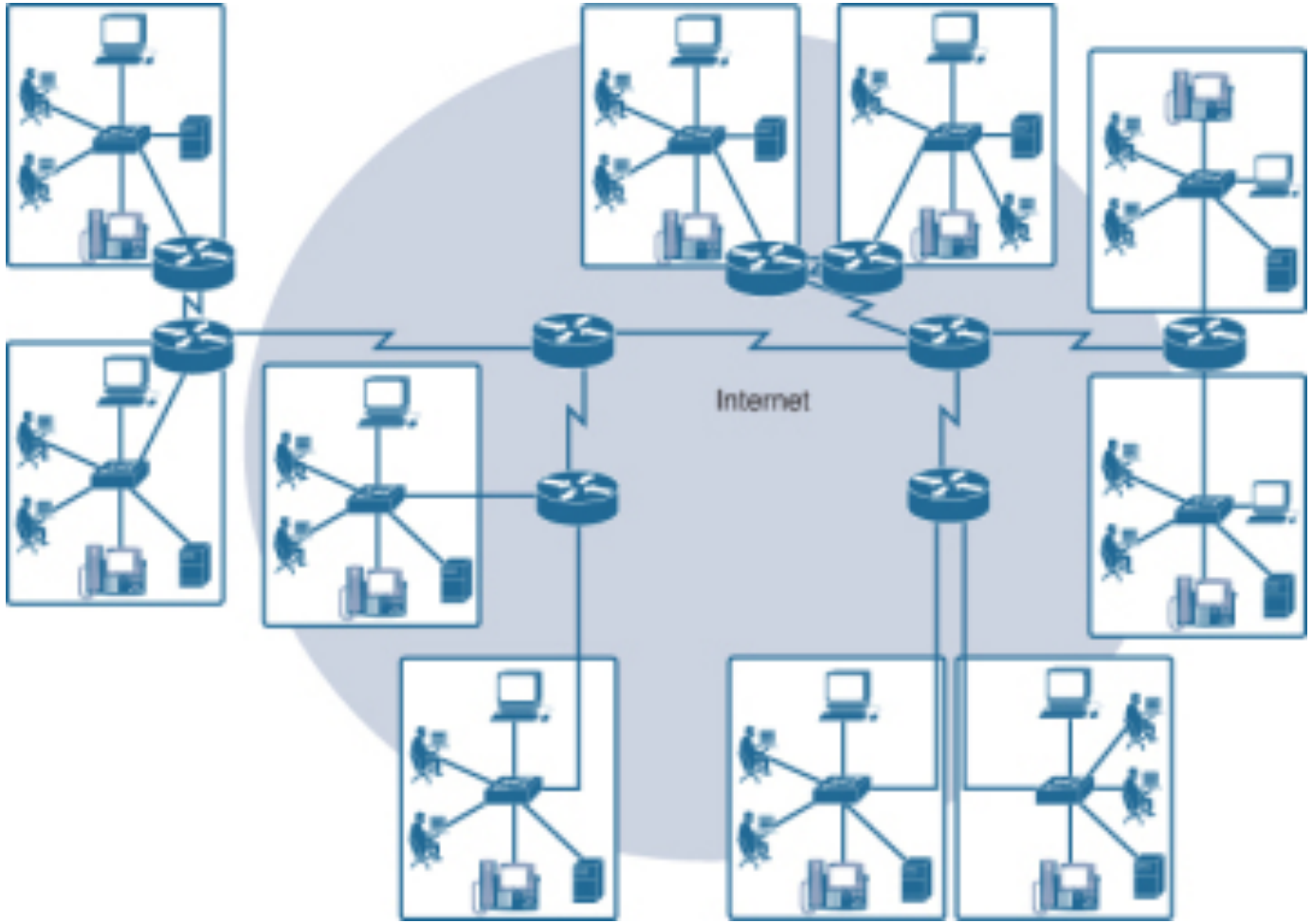
- We will talk about storage technology next week
- Primary interface to storage is Serial ATA, or SATA
  - Up to 6gb/s transfer rates
  - Solid state disks can use other interfaces (M5, PCI)
- Can interface to storage over other interfaces
  - USB, SCSI,



Network Basics

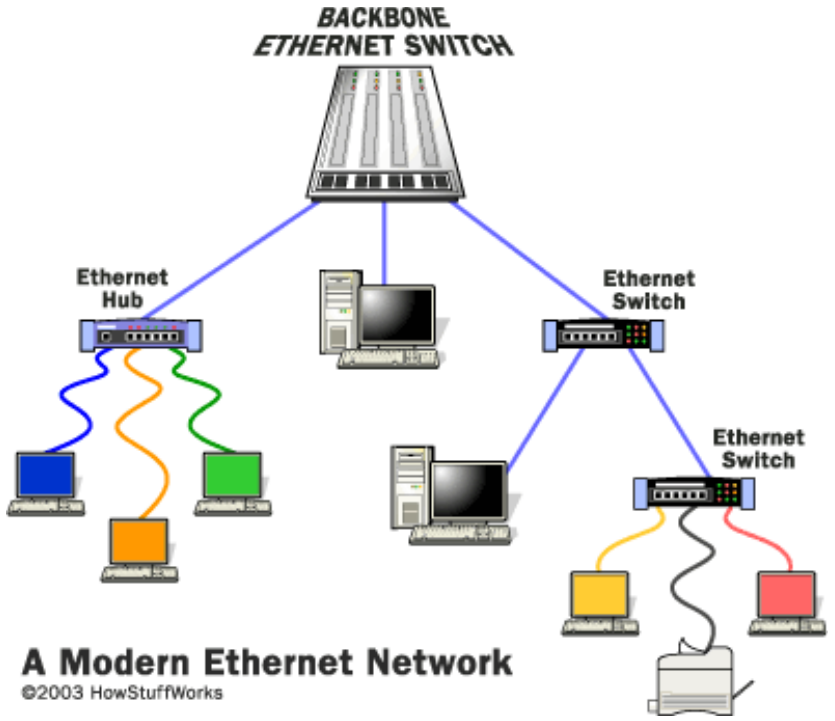
- Main components of a network are:
  - Network interface (plugs into the computer)
  - Physical medium (twisted pair wire, coax, radio, optical fiber…)
  - Networking hardware (routers, switches, firewalls, …)
  - Lots of software
- General types of networks
  - Local Area Network -- within a building, data center, most often “Ethernet”
  - Wide Area Network – up to global scale,

LANs and WANs



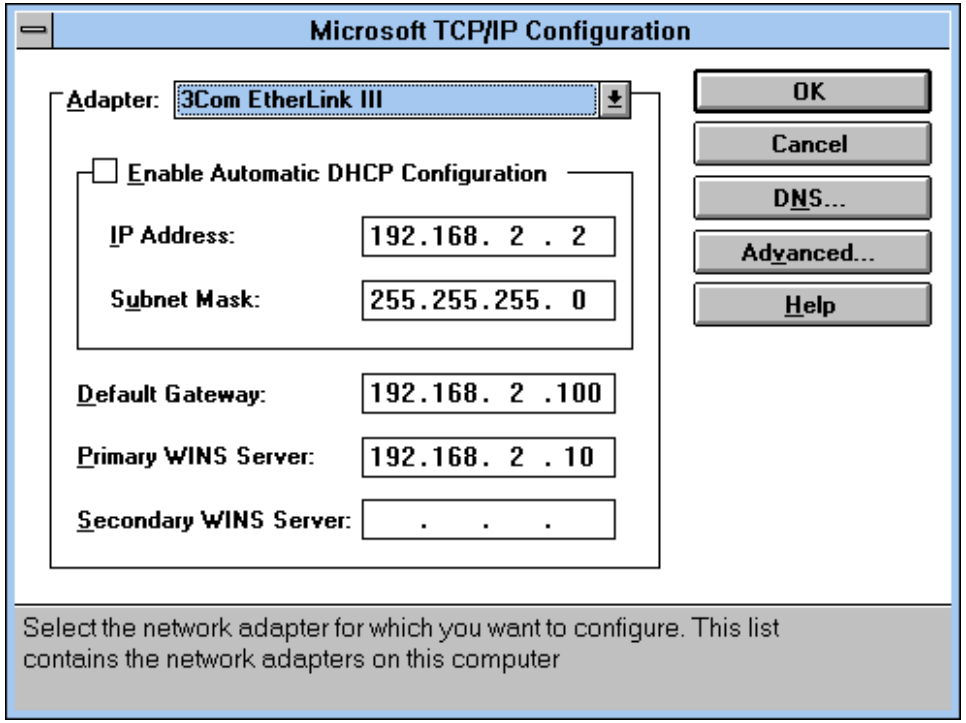
Local Area Networks

- Ethernet is most prevalent
  - Logically a “bus” architecture
    - Carrier sense multiple access with collision detection (CSMA/CD)
  - Each endpoint has a unique Ethernet address (MAC address)
    - 0c:4d:e9:a9:73:f3
- Most modern deployments are switched
- Speeds of 10/100/1G/10G/40G/100G bit/sec
- Physical Media:
  - Unshielded Twisted Pair (UTP) for 1G at 100M using CAT6
  - Fiber, required for longer reach and higher bandwidths
  - Radio, for WiFi



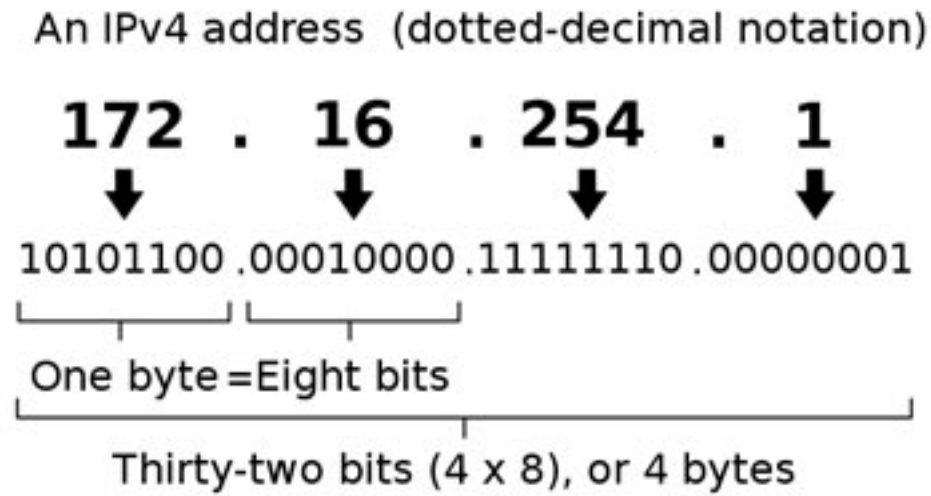
Moving Data over the Network

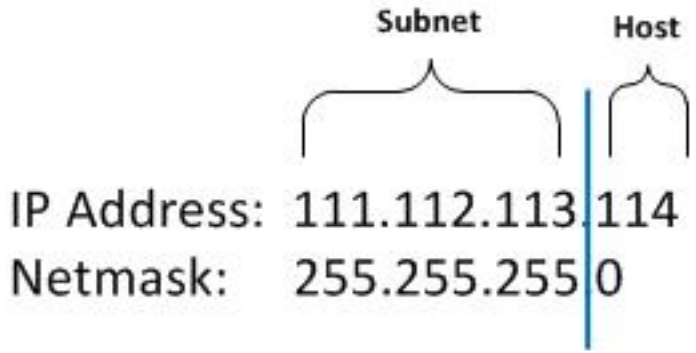
- Transmission Control Protocol / Internet Protocol – TCP/IP
  - Defines how to package up data into chunks called packets and move the packets from the source to a destination.
  - Every TCP/IP endpoint has a unique name called its TCP/IP address

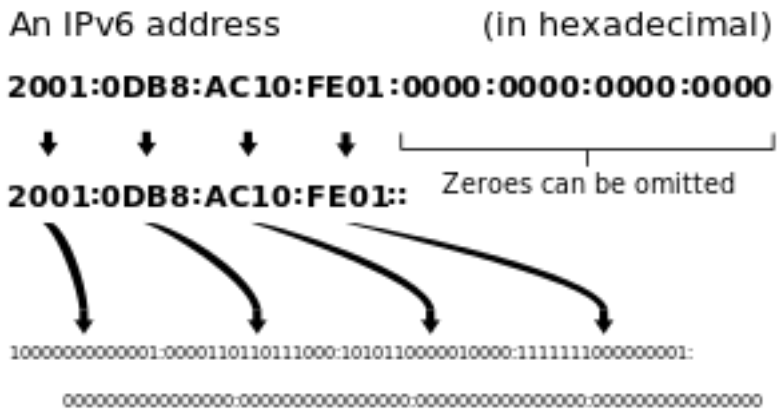


IP Address

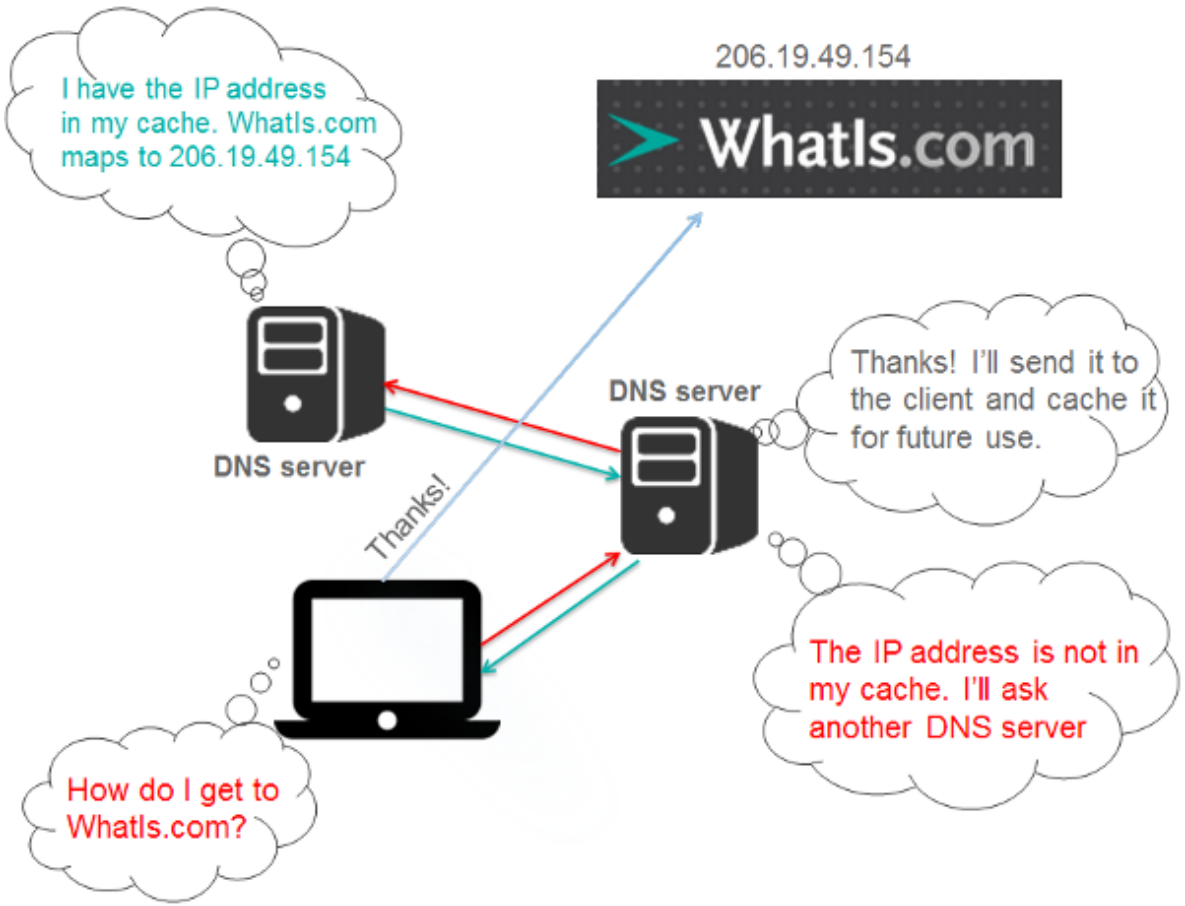
- Uniquely names each network endpoint
- Logically split into a network part and a host part.







Domain Name Service (DNS)



Network Security

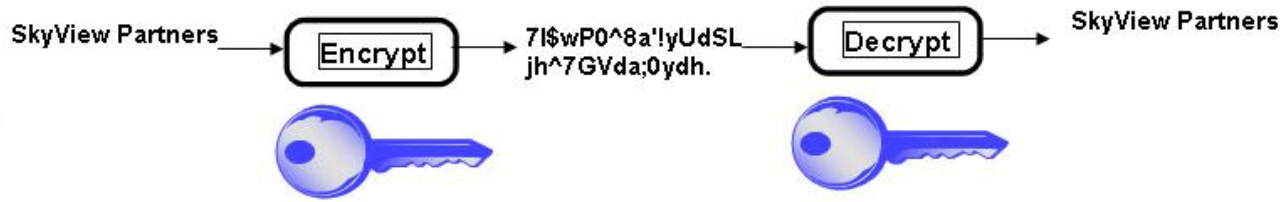
- Authentication – who are you
- Authorization – what can you do
- Privacy – no one else can see what you are doing
- Encryption is a core element

Types of encryption

DES
TripleDES
AES
RC5

Symmetric Keys

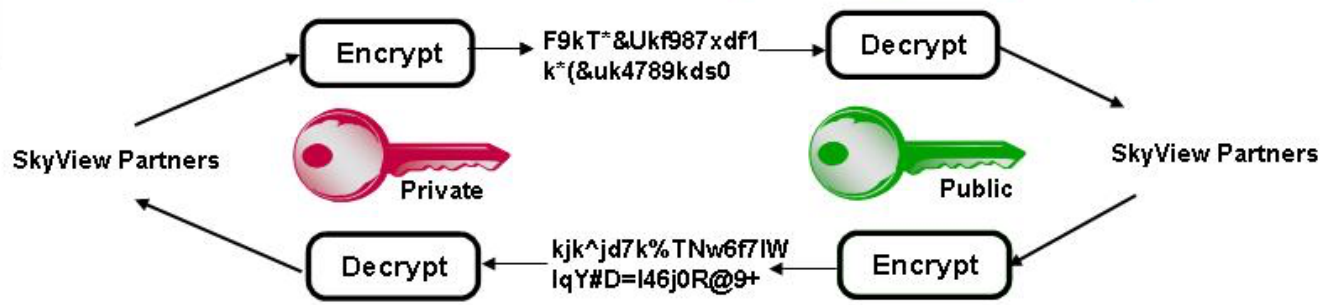
- Encryption and decryption use the **same key**.



RSA
Elliptic Curve

Asymmetric keys

- Encryption and decryption use different keys, a **public key** and a **private key**.



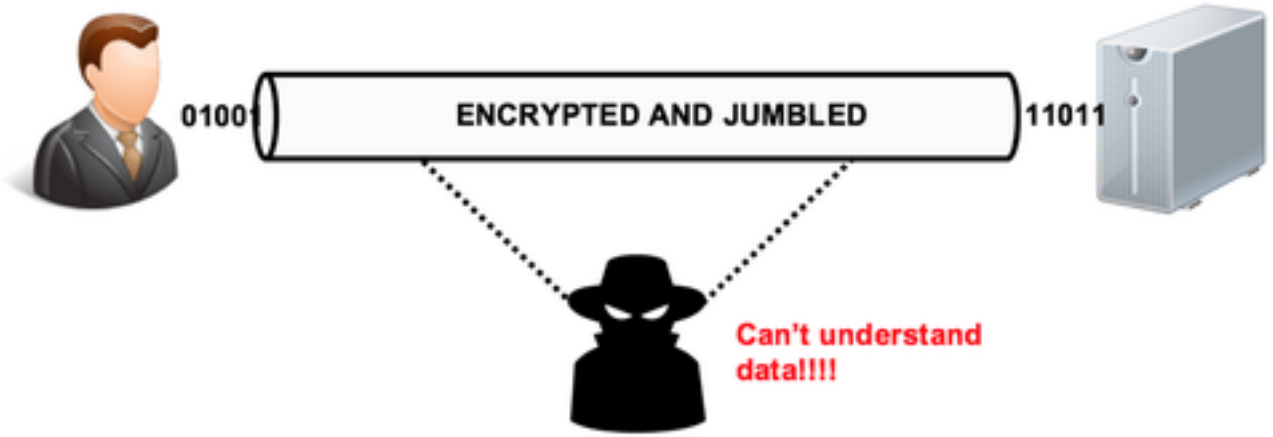
Typically, a password is used to encrypt the private key

MD5
SHA-1

One-way hash

Privacy on the network

- Firewalls
  - Limit access to network services based on characteristics such as IP address
- Authenticating to services
  - Examples include SSL, TLS, SSH
  - Certificate may be used to associate a public key with an identify
  - Network



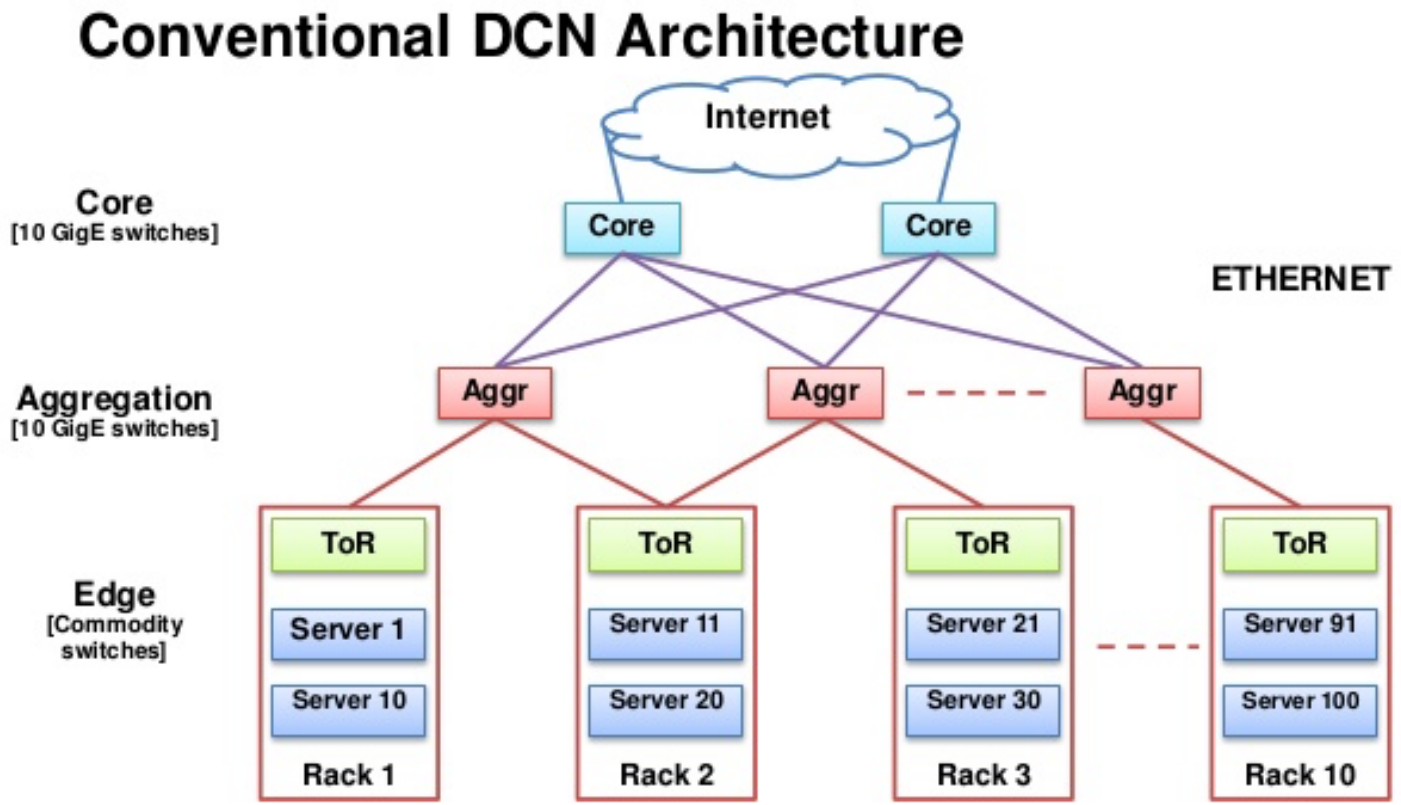
Data Centers

Centralize computing and storage resources in a single location to get economies of scale

The building as system

- Management of 10s of thousands of computers
  - Reliability issues
- Heat management/cooling
- Power management
- Networking

Data centers build on racks of servers



Cloud

The U.S. National Institute of Standards and Technology (NIST) defines cloud computing as:

- Cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction.

What is Cloud Computing?

- Cloud computing is a model for enabling *convenient*, *on-demand network access* to a *shared pool* of *configurable computing resources* (e.g., networks, servers, storage, applications, and services)
- It can be *rapidly provisioned* and *released* with minimal management effort.
- It provides *high level abstraction* of computation and storage model.
- It has some essential **characteristics, service models**, and **deployment models**.

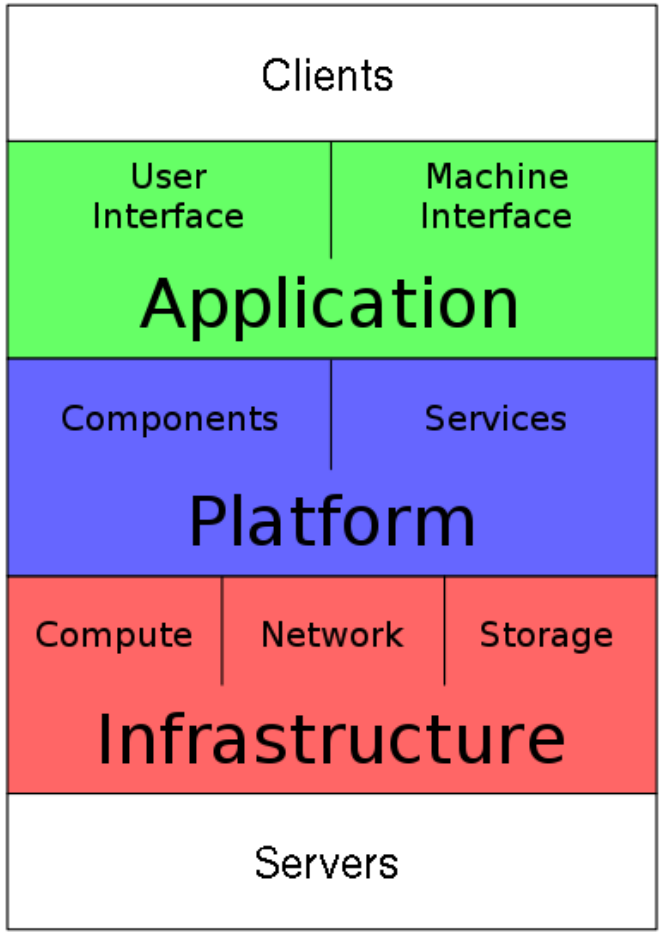
Essential Characteristics

- ***On-Demand Self Service:***
  - A consumer can unilaterally provision computing capabilities, automatically without requiring human interaction with each service's provider.
- ***Heterogeneous Access:***
  - Capabilities are available over the network and accessed through standard mechanisms that promote use by heterogeneous ***thin*** or ***thick*** client platforms.

Essential Characteristics (cont.)

- ***Resource Pooling:***
  - The provider's computing resources are pooled to serve multiple consumers using a *multi-tenant model.*
  - Different physical and virtual resources dynamically assigned and reassigned according to consumer demand.
- ***Measured Service:***
  - Cloud systems *automatically control* and *optimize* resources used by leveraging a metering capability at some level of abstraction appropriate to the type of service.
  - ***It will provide analyzable and predictable computing platform.***

Cloud Computing Service Models



Cloud Computing Stack

Service Model at a glance: Picture From http://en.wikipedia.org/wiki/File:Cloud_Computing_Stack.svg

Software as a Service (SaaS)

- The capability provided to the consumer is to use the provider's applications running on a cloud infrastructure.
- The applications are accessible from various client devices such as a web browser (e.g., web-based email).
- The consumer does not manage or control the underlying cloud infrastructure including network, servers, operating systems, storage,…
- ***Examples: Exchange, Google Apps, Salesforce, DropBox***

Platform as a Service (PaaS)

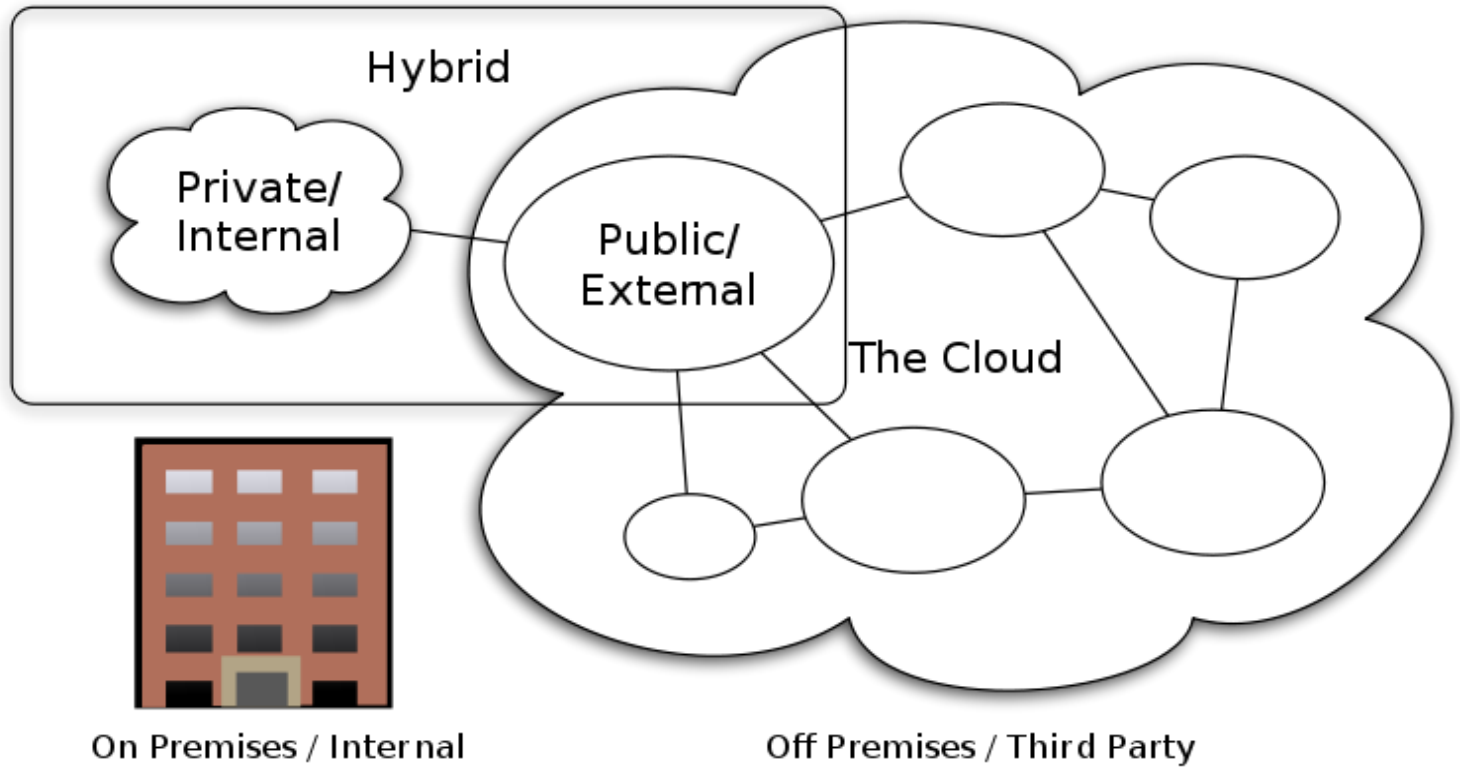
- The capability provided to the consumer is to deploy onto the cloud infrastructure *consumer-created* or *acquired applications* created using *programming languages and tools* supported by the provider.
- The consumer does not manage or control the underlying cloud infrastructure.
- Consumer has control over the deployed applications and possibly application hosting environment configurations.
- ***Examples: Windows Azure, Google App Engine, Amazon Web Services.***

Infrastructure as a Service (IaaS)

- The capability provided to the consumer is to provision processing, storage, networks, and other fundamental computing resources.
- The consumer is able to deploy and run arbitrary software, which can include operating systems and applications.
- The consumer does not manage or control the underlying cloud infrastructure but has control over operating systems, storage, deployed applications, and possibly limited control of select networking components (e.g., host firewalls).
- ***Examples: Amazon EC2, Rackspace Cloud Servers, ReliaCloud.***

Deployment Models

- Private Cloud
  - The cloud is operated **solely** for an organization. It may be managed by the organization or a third party and may exist on premise or off premise.
- Community Cloud
  - The cloud infrastructure is shared by several organizations and supports a specific community that has **shared concerns**
  - May be managed by the organizations or a third party and may exist on premise or off premise
- Public Cloud
  - The cloud infrastructure is made available to the general public or a large industry group and it is owned by an organization selling cloud services.
- Hybrid Cloud
  - The cloud infrastructure is a composition of two or more clouds (private, community, or public).



Cloud Computing Types