

# Piattaforme Abilitanti Distribuite - PAD -

## Distributed Enabling Platforms

Nicola Tonello  
(ISTI, CNR)  
[nicola.tonello@isti.cnr.it](mailto:nicola.tonello@isti.cnr.it)



# Today



# Who?

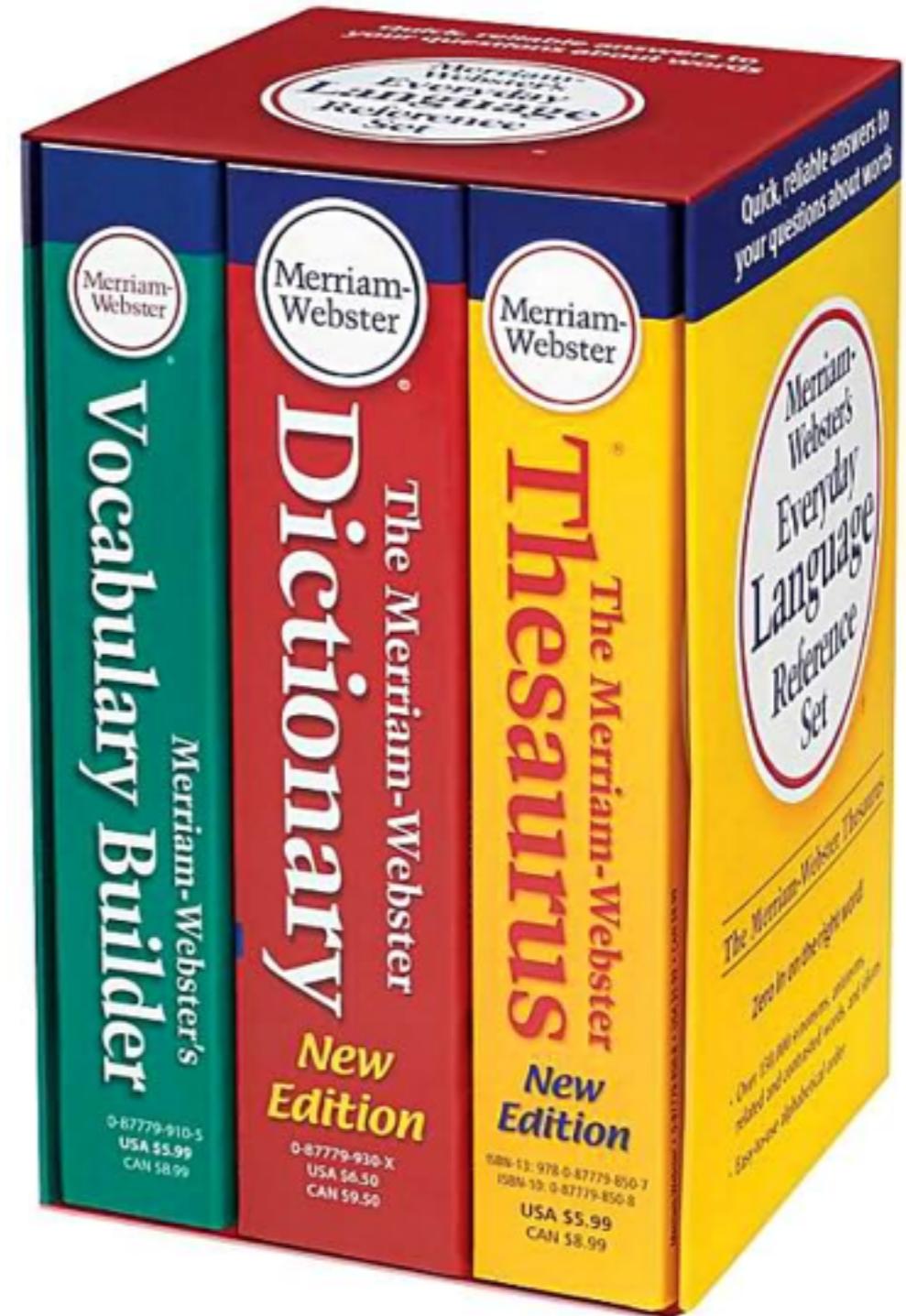


# What?

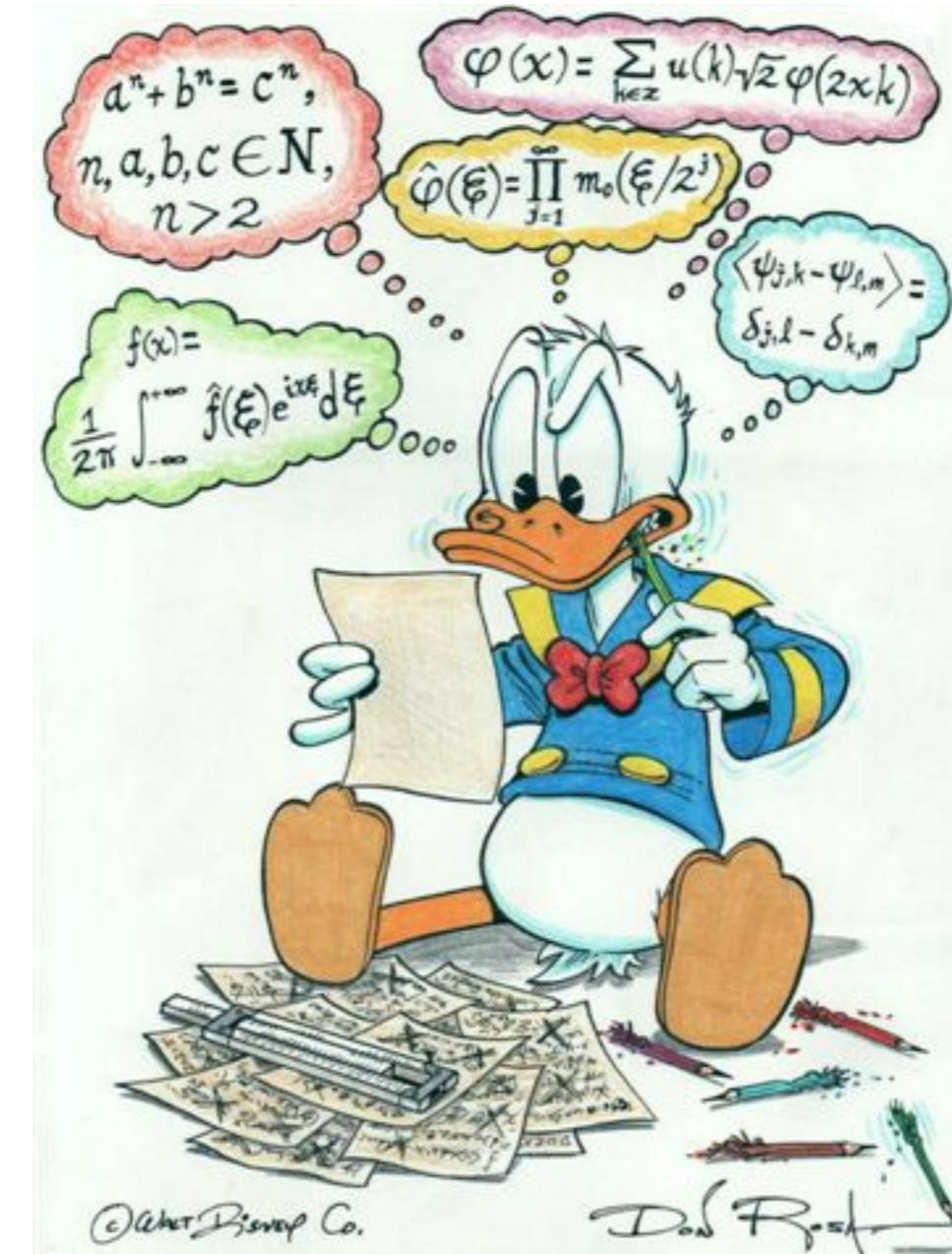


# What is the meaning of words?

- Distributed...
  - relating to a **computer network** in which at least some of the processing is done by the **individual computers** and **information is shared** by and often stored at the computers
- Enabling...
  - to make possible, practical, or easy
- Platforms...
  - the computer architecture and equipment used for a particular purpose



# To do what?

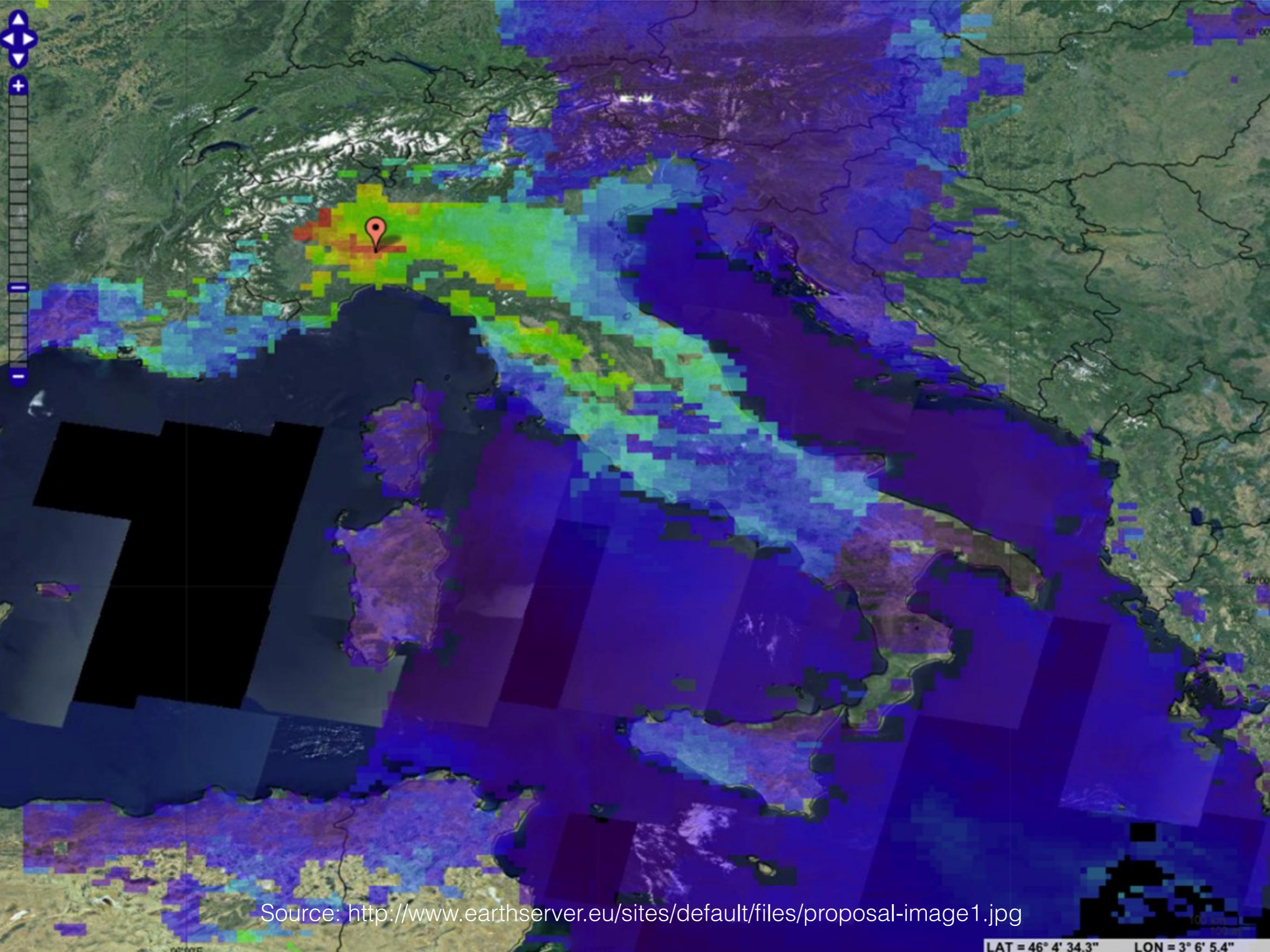


# Solve large scale problems!





Source: [https://upload.wikimedia.org/wikipedia/commons/e/eb/Views\\_of\\_the\\_LHC\\_tunnel\\_sector\\_3-4,\\_tirage\\_1.jpg](https://upload.wikimedia.org/wikipedia/commons/e/eb/Views_of_the_LHC_tunnel_sector_3-4,_tirage_1.jpg)



Source: <http://www.earthserver.eu/sites/default/files/proposal-image1.jpg>

LAT = 46° 4' 34.3" LON = 3° 6' 54.4"



Source: [https://upload.wikimedia.org/wikipedia/commons/1/18/NASDAQ\\_studio.jpg](https://upload.wikimedia.org/wikipedia/commons/1/18/NASDAQ_studio.jpg)

Source: [https://upload.wikimedia.org/wikipedia/commons/1/19/Times\\_Square%2C\\_New\\_York\\_City\\_%28HDR%29.jpg](https://upload.wikimedia.org/wikipedia/commons/1/19/Times_Square%2C_New_York_City_%28HDR%29.jpg)



Source: [https://upload.wikimedia.org/wikipedia/commons/2/24/Huge\\_crowd\\_turns\\_out\\_for\\_MTV\\_EXIT\\_concert\\_against\\_human\\_trafficking\\_and\\_exploitation.jpg](https://upload.wikimedia.org/wikipedia/commons/2/24/Huge_crowd_turns_out_for_MTV_EXIT_concert_against_human_trafficking_and_exploitation.jpg)



# Some Numbers (I)

- 10,033 Tweets sent in 1 second
- 2,623 Instagram photos uploaded in 1 second
- 2,199 Tumblr posts in 1 second
- 1,862 Skype calls in 1 second
- 29,667 GB of Internet traffic in 1 second
- 50,512 Google searches in 1 second
- 107,401 YouTube videos viewed in 1 second
- 2,425,138 Emails sent in 1 second

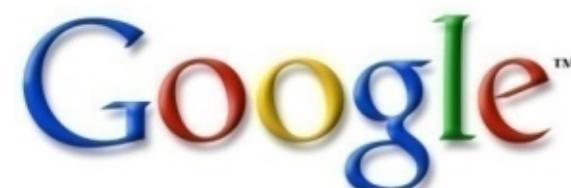
source: <http://www.internetlivestats.com>

# Some Numbers (II)

- 3.2 billions of Internet users
- 928 millions of Web sites
- 168 billions of email sent during the day
- 3.2 billions of google searches during the day
- 2.9 millions blog posts written during the day
- 640 millions of tweets sent during the day
- 6.8 billions of videos viewed on Youtube during the day
- 167 millions of Photos uploaded on Instagram during the day
- 1.5 billions of Facebook active users
- 1.5 billions of Google+ active users
- 328 millions of Twitter active users
- 120 millions of Skype calls during the day
- 40 thousands of Web sites hacked during the day
- 503 thousand of computers sold today
- 4 millions of smartphones sold today
- 772 thousands of tablets sold today
- 1.9 billions of GB (1.9 EB) Internet traffic today

source: <http://www.internetlivestats.com>

# BIG DATA!



Processes 20 PB a day (2008)  
Crawls 20B web pages a day (2012)  
Search index is 100+ PB (5/2014)  
Bigtable serves 2+ EB, 600M QPS (5/2014)



150 PB on 50k+ servers  
running 15k apps (6/2011)



19 Hadoop clusters: 600  
PB, 40k servers (9/2015)



Hadoop: 10K nodes, 150K  
cores, 150 PB (4/2014)



300 PB data in Hive +  
600 TB/day (4/2014)



S3: 2T objects, 1.1M  
request/second (4/2013)

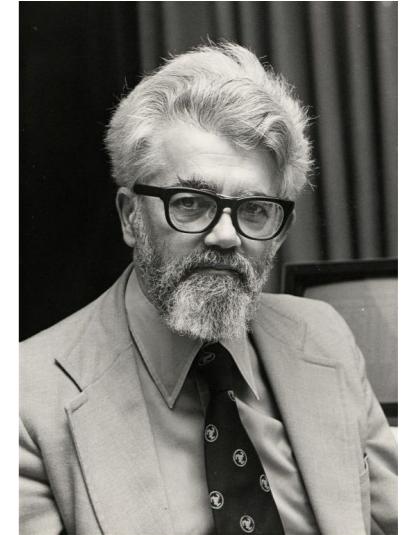
# How?



# Famous predictions

## 1961

[...] computing may someday be organized as a **public utility** just as telephone system is a public utility [...] the computer utility could become the basis of a new and important industry [...]



John McCarthy (1927-2011)  
Turing Award (1971)  
Artificial Intelligence

## 1969

As of now, computer networks are still in their infancy, but as they group up and become sophisticated, we will probably see the spread of **computer utilities** which, like present electric and telephone utilities, will service individual homes and offices across the country.

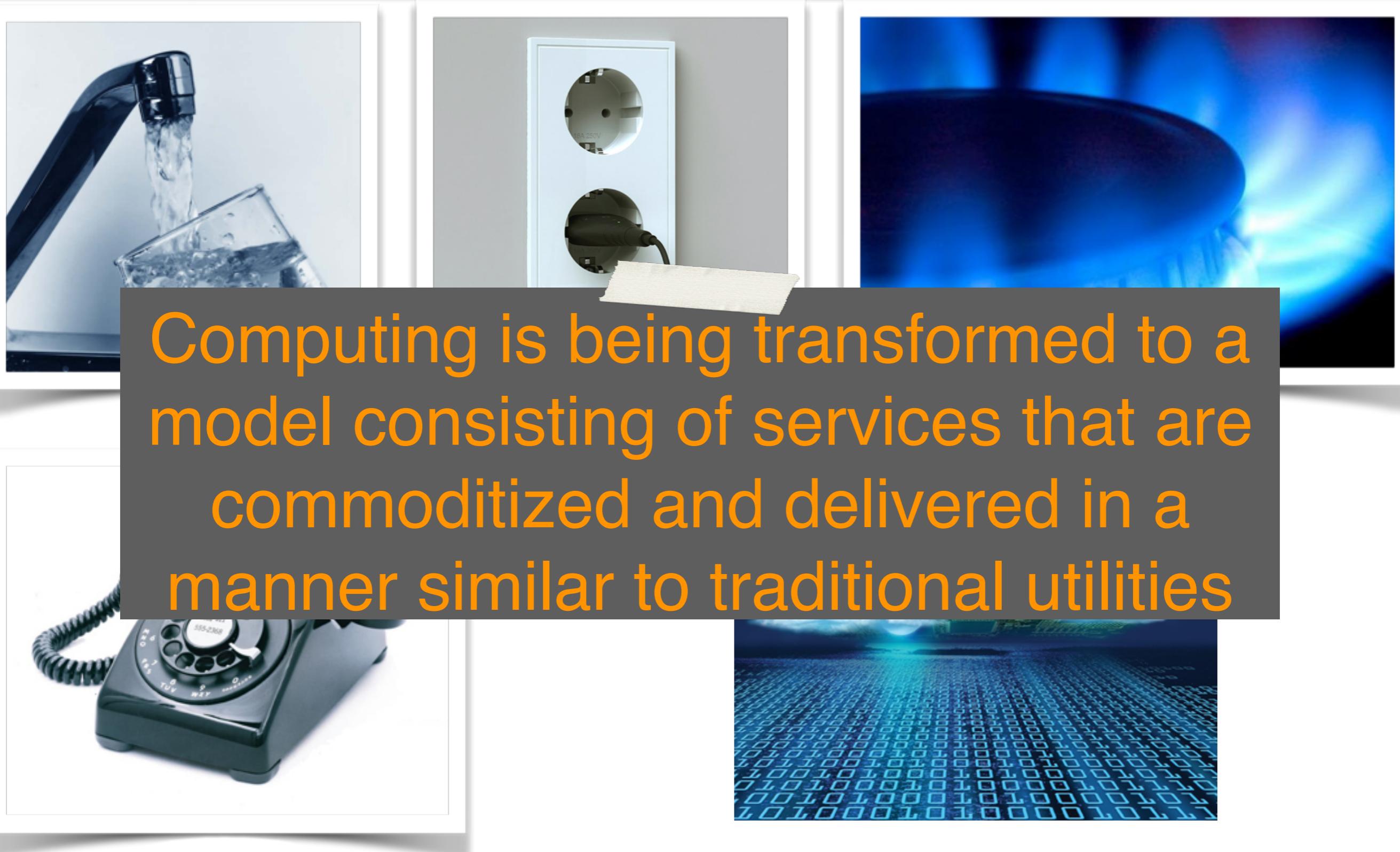


Leonard Kleinrock (1934)  
Queueing Theory

# The 5th Utility



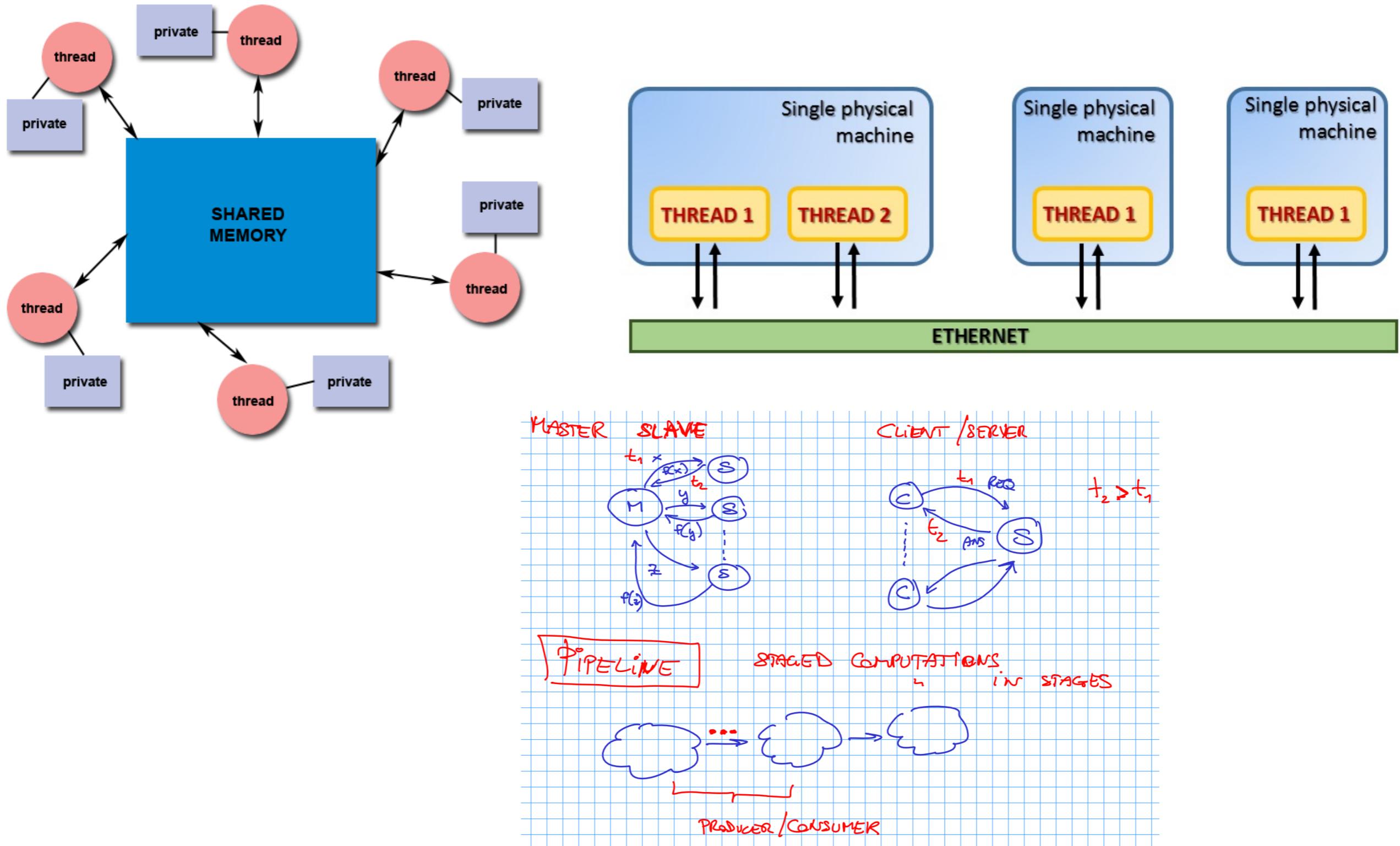
# The 5th Utility



Computing is being transformed to a model consisting of services that are commoditized and delivered in a manner similar to traditional utilities



# Tools you (should) know



# Once upon a time...



Microcomputer



Minicomputer



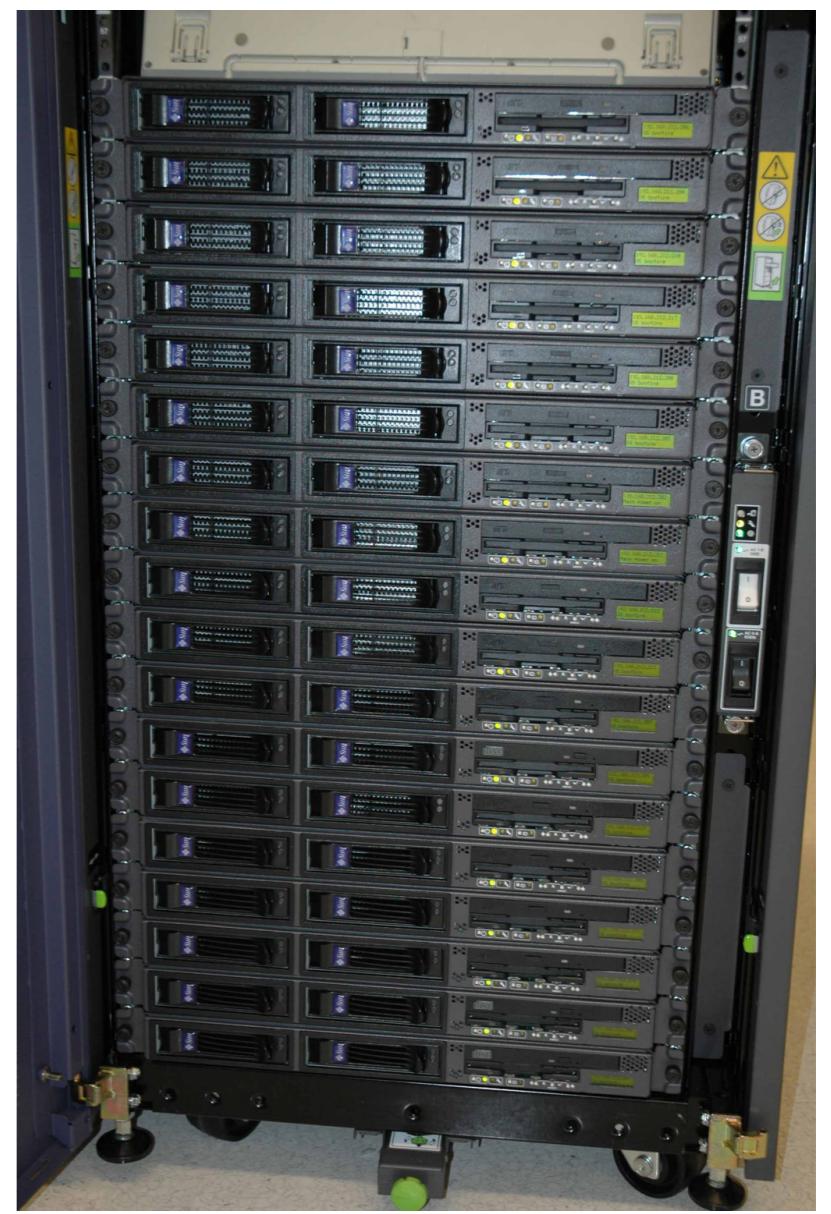
Cluster



Mainframe

# Cluster Computing

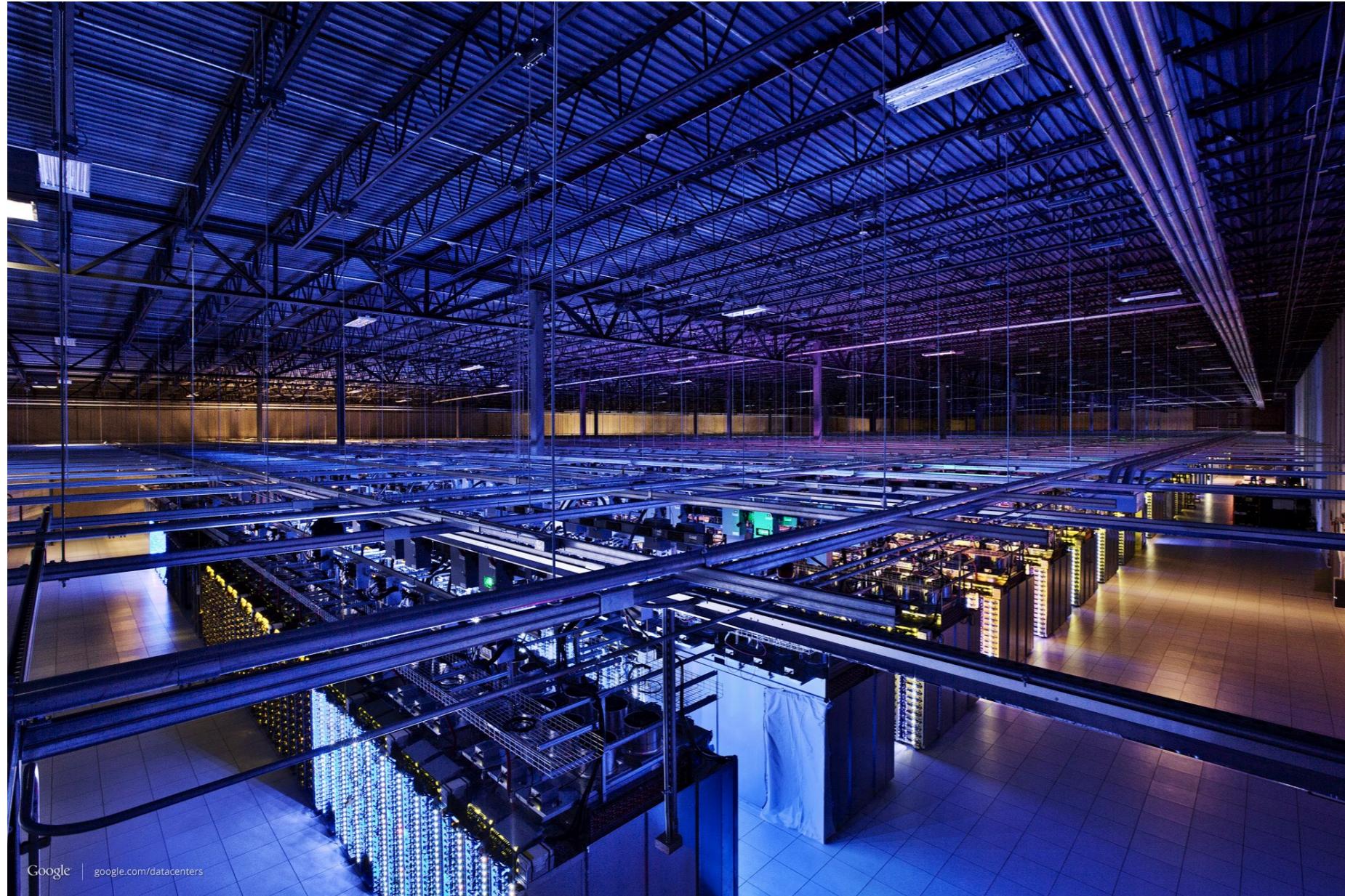
- A cluster is a type of parallel and distributed system, which consists of a collection of **inter-connected stand-alone computers** working together as a **single integrated computing resource**.
- Basic element is the **node**, a single or multiprocessor system with memory, I/O and OS
- Generally two or more nodes connected together
- In a single **rack**, or physically separated and connected via a LAN
- Appears as a single system to users and applications
- Specialized access, management and programming



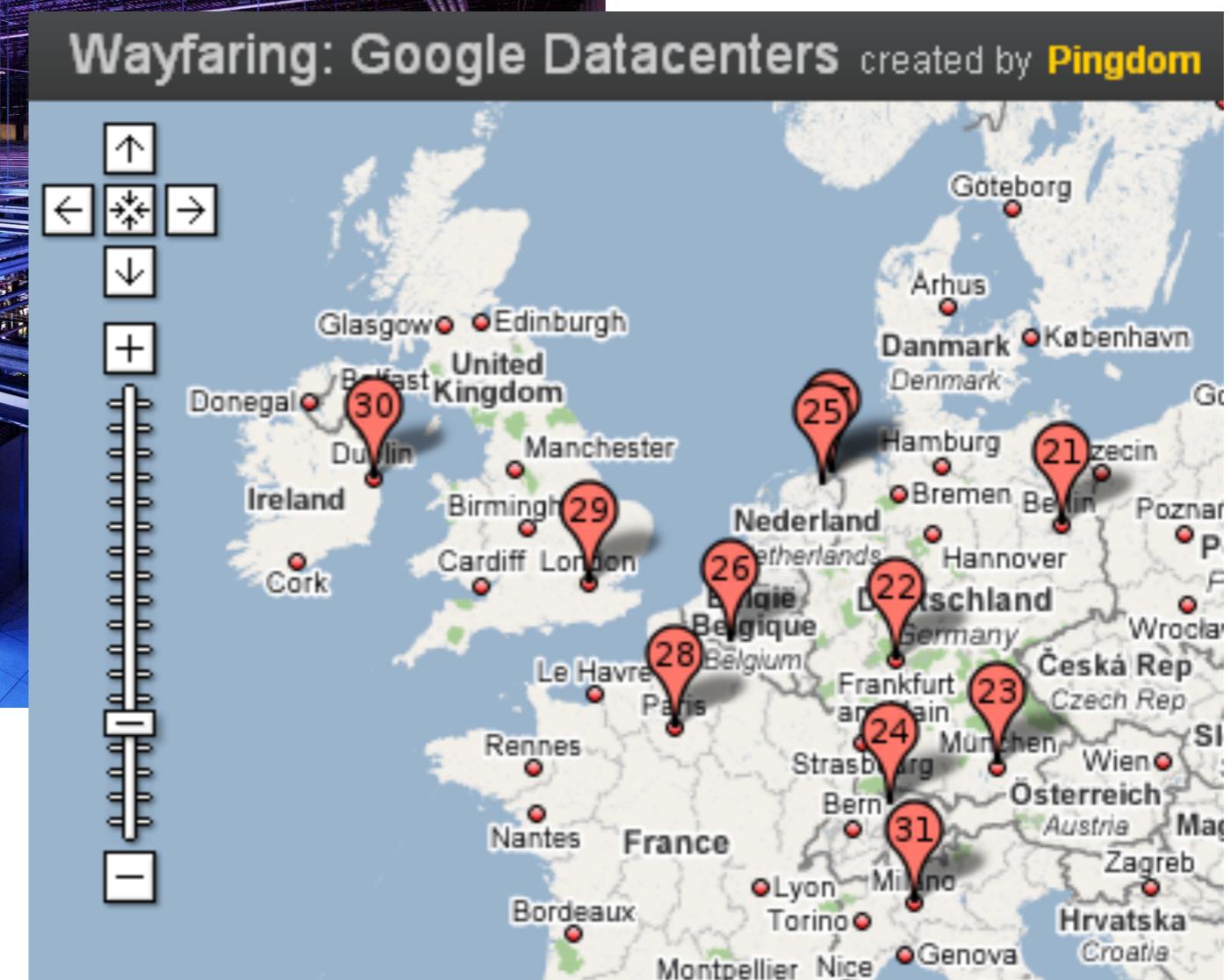
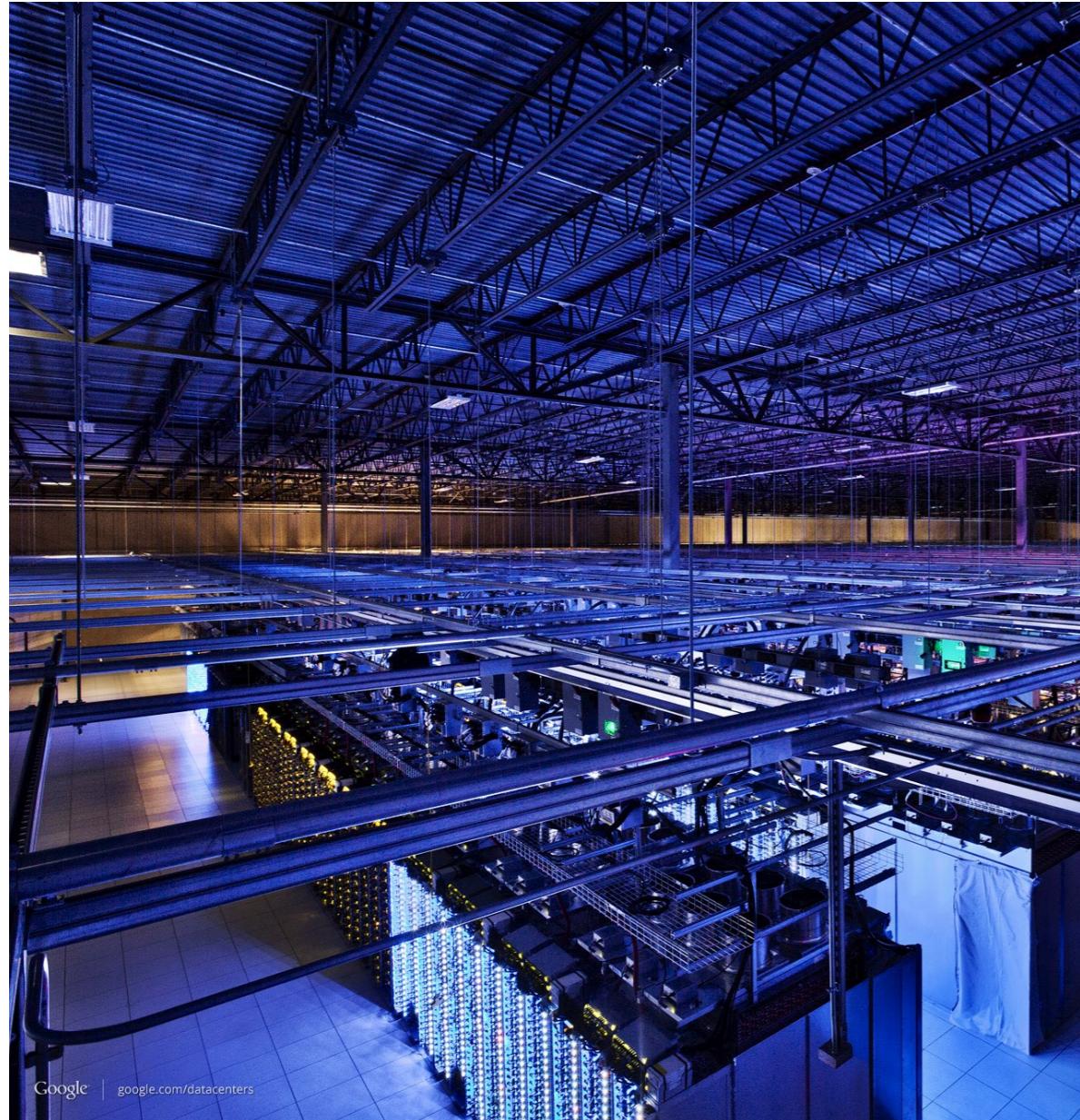
# ...up to the Grid...



# ...up to the Cloud



# ...up to the Cloud





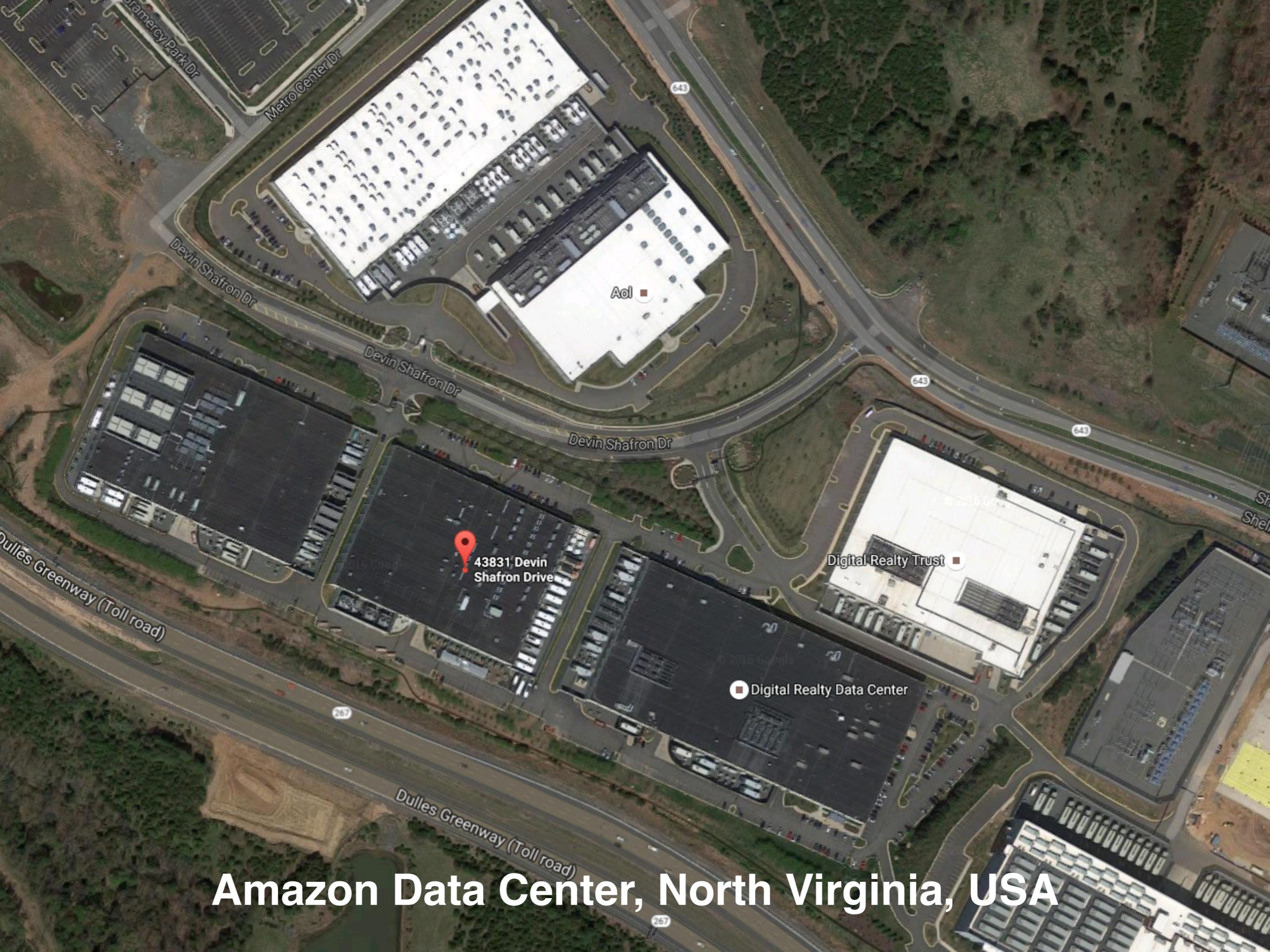
Google Data Center, The Dalles, Oregon

# Facebook Data Center, Luleå, Sweden



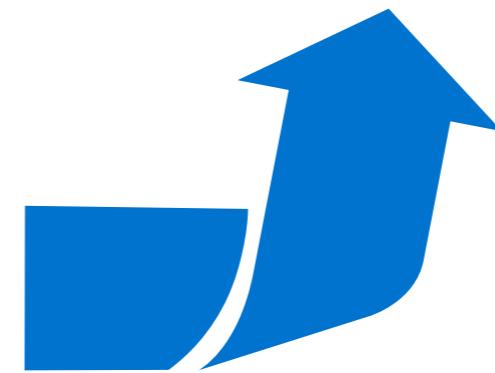


Microsoft Data Center, Dublin, Ireland



**Amazon Data Center, North Virginia, USA**

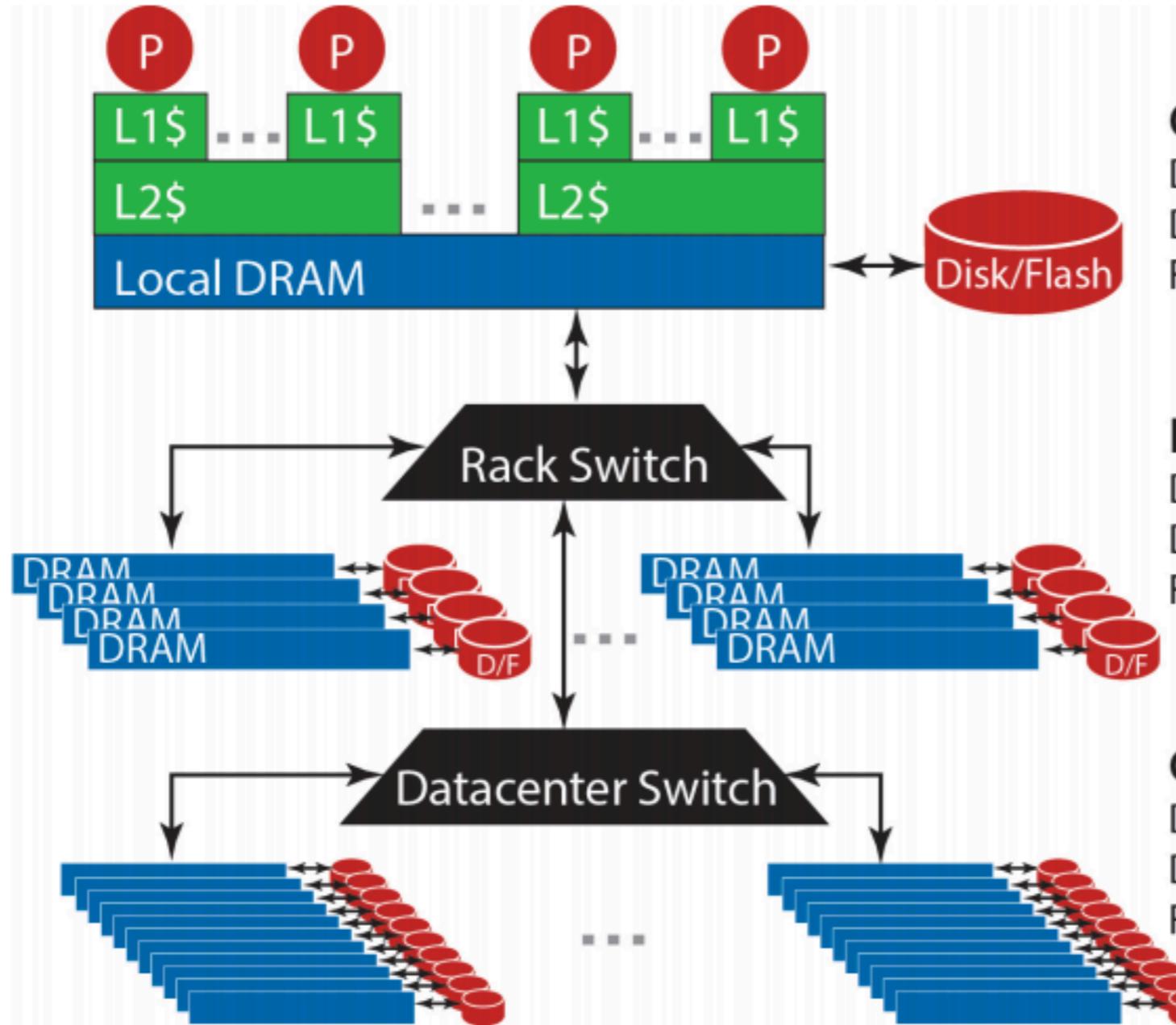
# Datacenter Basic Components





Facebook

# Datacenter Storage



## One Server

DRAM: 16 GB, 100 ns, 20 GB/s  
 Disk: 2TB, 10 ms, 200 MB/s  
 Flash: 128 GB, 100 us, 1 GB/s

## Local Rack (80 servers)

DRAM: 1 TB, 300 us, 100 MB/s  
 Disk: 160 TB, 11 ms, 100 MB/s  
 Flash: 20 TB, 400 us, 100 MB/s

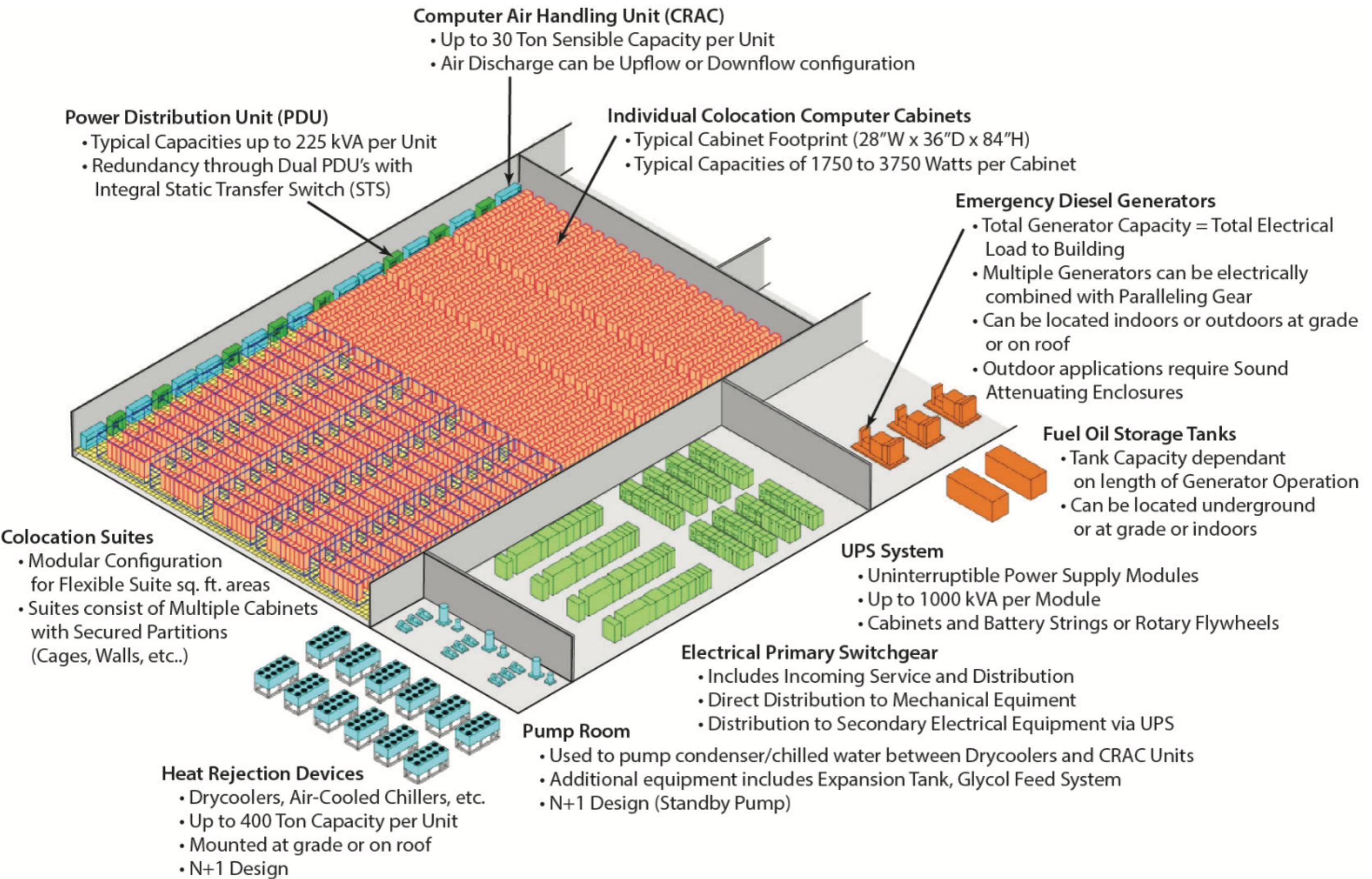
## Cluster (30 racks)

DRAM: 30 TB, 500 us, 10 MB/s  
 Disk: 4.80 PB, 12 ms, 10 MB/s  
 Flash: 600 TB, 600 us, 10 MB/s

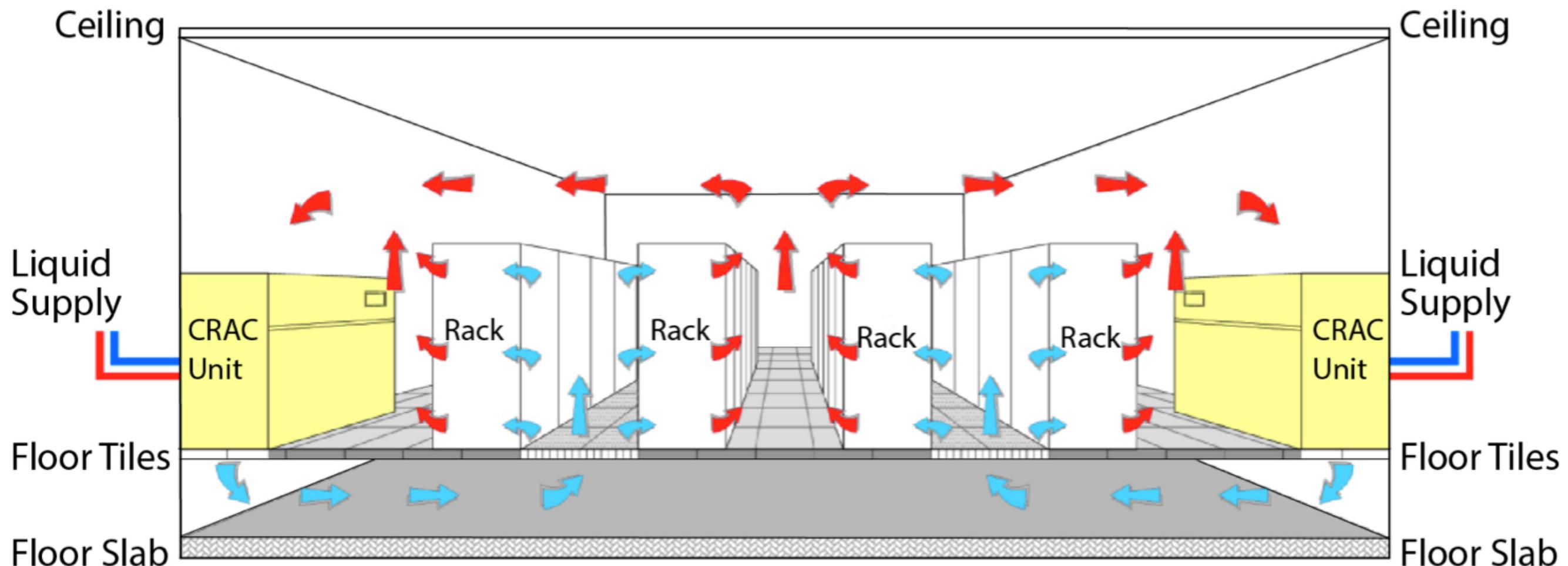
# Datacenter Tiers

- **Datacenters have strict standards for reliability and availability**
  - Tier 1: 99.671% Availability: 28 hours of downtime/year
    - a single path for power distribution, UPS, and cooling distribution, without redundant components.
  - Tier 2: 99.741% Availability: 22 hours of downtime/year
    - redundant components to this design, improving availability.
  - Tier 3: 99.982% Availability: 1.5 hours of downtime/year
    - one active and one alternate distribution path for utilities.
  - Tier 4: 99.995% Availability: 26 minutes of downtime/year
    - two simultaneously active power and cooling distribution paths, redundant components in each path, and are supposed to tolerate any single equipment failure without impacting the load.

# Datacenter Power Systems



# Datacenter Cooling





Google Cooling System

# Latency Numbers Every Programmer Should Know

## Latency Comparison Numbers

L1 cache reference	0.5	ns		
Branch mispredict	5	ns		
L2 cache reference	7	ns	14x L1 cache	
Mutex lock/unlock	25	ns		
Main memory reference	100	ns	20x L2 cache, 200x L1 cache	
Read 4K randomly from memory	1,000	ns	0.001 ms	
Compress 1K bytes with Zippy	3,000	ns		
Send 1K bytes over 1 Gbps network	10,000	ns	0.01 ms	
Read 4K randomly from SSD*	150,000	ns	0.15 ms	
Read 1 MB sequentially from memory	250,000	ns	0.25 ms	
Round trip within same datacenter	500,000	ns	0.5 ms	
Read 1 MB sequentially from SSD*	1,000,000	ns	1 ms	4X memory
Disk seek	10,000,000	ns	10 ms	20x datacenter roundtrip
Read 1 MB sequentially from disk	20,000,000	ns	20 ms	80x memory, 20X SSD
Send packet CA->Netherlands->CA	150,000,000	ns	150 ms	

Source: Jeff Dean and Peter Norvig (Google), with some additions

<https://gist.github.com/hellerbarde/2843375>

## For all scale analogies, consider 1 CPU cycle = 1 second

(In reality, 1 CPU cycle = 0.3 nanoseconds)

**ONE CPU CYCLE**  
=.3NS, WHICH = 1 SEC,  
OR IS EQUAL TO



**Clapping  
your hands**

**L1 CACHE ACCESS**  
.9NS, WHICH = 3 SEC,  
OR IS EQUAL TO



**Blowing  
your nose**

**L2 CACHE ACCESS**  
= 2.8NS, WHICH = 9 SEC,  
OR IS EQUAL TO



**Bill Gates  
earning \$2,250**

**L3 CACHE ACCESS**  
= 12.9NS, WHICH = 43 SEC,  
OR IS EQUAL TO



**COMPLETING AN AVERAGE MARIO BROS.  
Level 1-1 speed run**  
(THE WORLD RECORD IS ABOUT 28 SECONDS)

**MUTEX LOCK/UNLOCK**  
= 17 NS, WHICH = 56 SEC,  
OR IS EQUAL TO



**Washing  
your dishes**

**MAIN MEMORY ACCESS**  
= 100 NS, WHICH = 6 MIN,  
OR IS EQUAL TO



**LISTENING TO QUEEN'S  
“Bohemian Rhapsody”**

**COMPRESS 1KB WITH ZIPPY**  
= 2 $\mu$ S, WHICH = 2 HOURS,  
OR IS EQUAL TO



**Watching  
a movie**

**READ 1M BYTES  
SEQUENTIALLY FROM MEMORY**  
= 9 $\mu$ S, WHICH = 9 HOURS,  
OR IS EQUAL TO



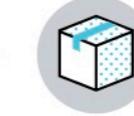
**COMPLETING A STANDARD  
US workday**

**SSD RANDOM READ**  
= 16  $\mu$ S, WHICH = 14 HOURS,  
OR IS EQUAL TO



**TAKING A FLIGHT FROM  
New York to Beijing**

**SOLID-STATE DISK I/O**  
= 50-150  $\mu$ S, WHICH = 2-6 DAYS,  
OR IS EQUAL TO



**WAITING FOR A STANDARD GROUND-SHIPPED  
US domestic package**

**READ 1M BYTES  
SEQUENTIALLY FROM SSD**  
= 200  $\mu$ S, WHICH = 8 DAYS,  
OR IS EQUAL TO



**IF THERE WERE 8 DAYS IN A WEEK,  
IT WOULD NOT BE ENOUGH TIME FOR  
The Beatles  
TO SHOW THEY CARE**

**ROUND TRIP IN THE  
SAME DATACENTER**  
= 500  $\mu$ S, WHICH = 18 DAYS,  
OR IS EQUAL TO



**Free climbing  
EL CAPITAN'S DAWN WALL  
IN YOSEMITE NATIONAL PARK**

**READ 1M BYTES SEQUENTIALLY  
FROM A SPINNING DISK**  
= 2MS, WHICH = 70 DAYS,  
OR IS EQUAL TO



**PLANTING + HARVESTING A  
zucchini**

**DISK SEEK**  
= 4MS, WHICH = 5 MONTHS,  
OR IS EQUAL TO



**TRAINING FOR YOUR  
first marathon**  
IF YOU'VE NEVER DONE ONE • YOU'RE  
AT AN AVERAGE FITNESS LEVEL

**ROTATIONAL DISK I/O**  
= 1-10MS, WHICH = 1-12 MONTHS,  
OR IS EQUAL TO



**WAITING UNTIL THE NEXT SEASON OF  
Game of Thrones**

**INTERNET: SF TO NYC**  
= 71MS, WHICH = 7 YEARS,  
OR IS EQUAL TO



**ATTENDING + GRADUATING  
Hogwarts  
IF YOU'RE A WITCH OR WIZARD**

**OS VIRTUALIZATION REBOOT**  
= 4 S, WHICH = 423 YEARS,  
OR IS EQUAL TO



**423 YEARS AGO,  
Shakespeare  
WROTE RICHARD III**

**SCSI COMMAND TIME-OUT**  
= 30 S, WHICH = 3,000 YEARS,  
OR IS EQUAL TO



**3,000 YEARS AGO, PEOPLE STARTED  
wearing pants**

**HARDWARE VIRTUALIZATION  
REBOOT**  
= 40 S, WHICH = 4,000 YEARS,  
OR IS EQUAL TO



**4,000 YEARS AGO, THE PHARAOHS STILL  
ruled Egypt**

**PHYSICAL SYSTEM REBOOT**  
= 5 MINUTES, WHICH = 32,000 YEARS,  
OR IS EQUAL TO



**32,000 YEARS AGO, THE AREA THAT IS THE  
Sahara desert was  
well-watered**

# Where? & When?



# Course Organization

- Agreement on room and timetable
  - Currently: Mon 16 – 18 (room L1), Fri 9 – 11 (room C1)
  - Depending on availability
- Highly interactive lectures
- Laboratory
  - Java programming skills required
  - Bring your own laptop (don't forget plugs!)
- Slides and references available online
  - Updated in real time on the course wiki
- Final examination: project + oral session
  - To be agreed with teacher