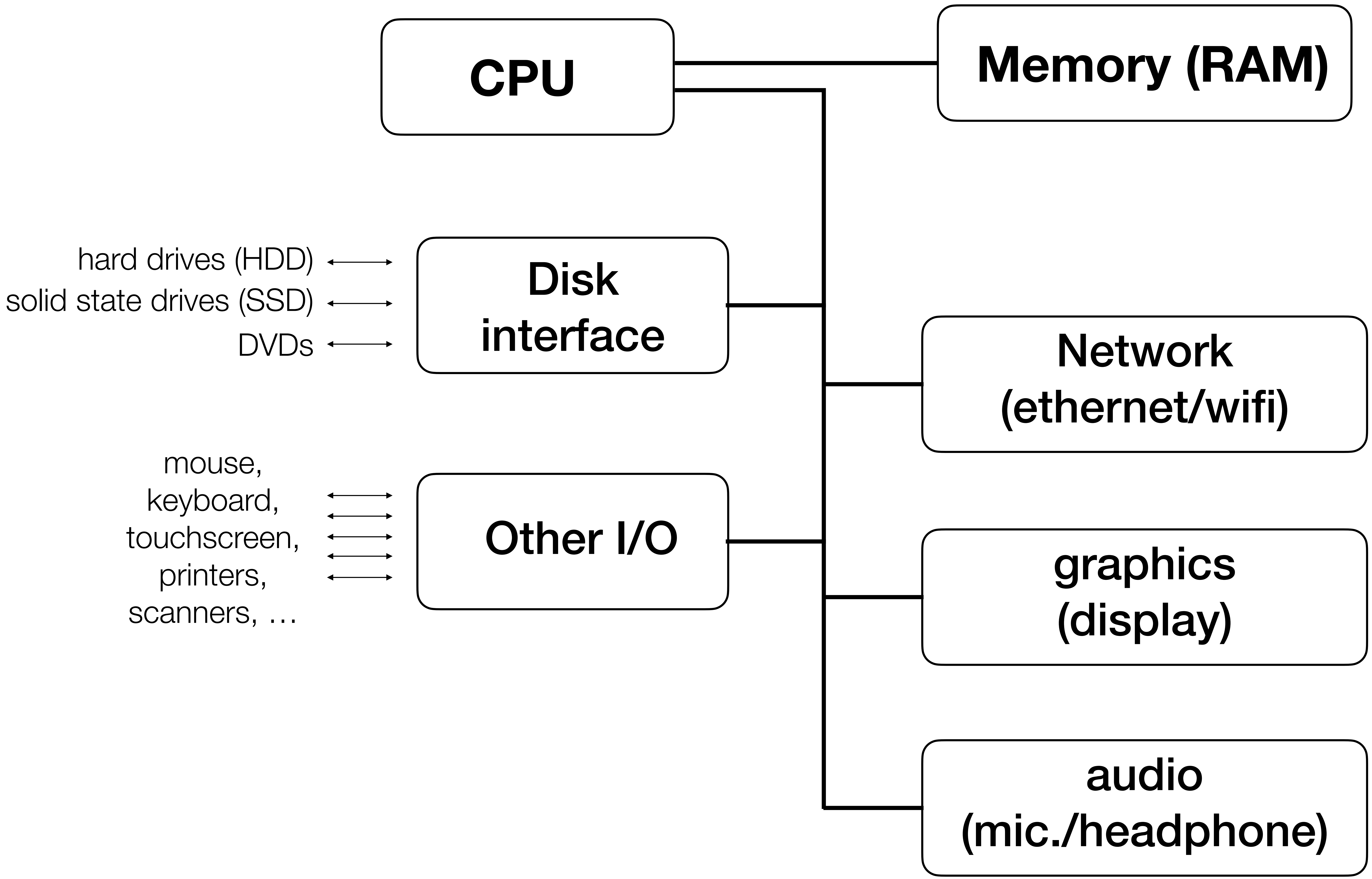
Introduction to computers and computing environments

Mark Stenglein, GDW pre-workshop crash course
StengleinLab.org

Genomics of Disease in Wildlife

Basic computer architecture



Memory (RAM) vs. storage (drives)

	Hard drives (HDD)	Solid state drives (SSD)	Main memory (RAM)
Speed	Slow	Much faster	Fastest
Cost per Gb	$0.025	$0.20	$4
Volatile	No	No	Yes
Uses	Long term storage of large datasets	Laptops, long term storage, OS boot drives	
Other comments	Uses more power	Largest SSD ~4 Tb	Some bioinformatic tasks, like genome assembly, require lots of RAM

Memory (RAM) is like a “mise en place” station in a restaurant kitchen. Storage is like a pantry.

RAM: quick to access

Hard drives: larger capacity, slower to access

How much RAM do you need?

Your MacBooks have **16 Gb** of RAM. That much RAM is plenty for many tasks.

Some bioinformatics tasks are processor (CPU) intensive, but use little RAM. For instance, inferring phylogenetic trees.

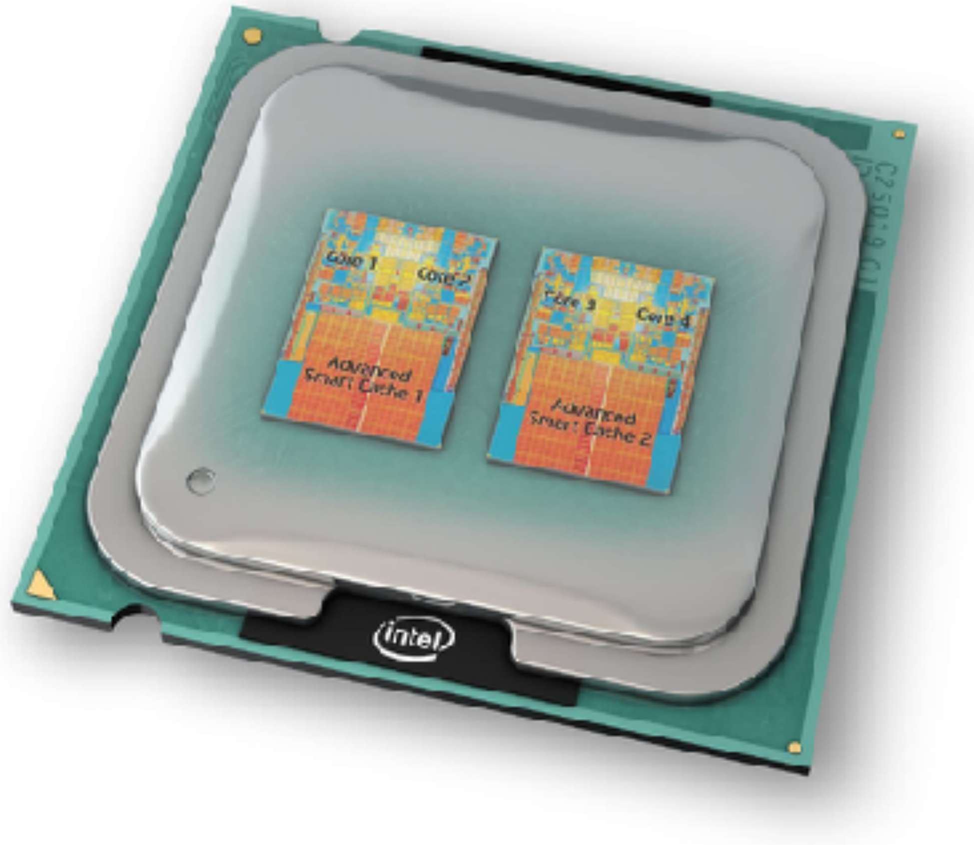
Other tasks, like assembly of large eukaryotic genomes or searching large databases with tools like BLAST, require lots of RAM (like ~512 Gb to 1 Tb). These types of analyses are better done on a dedicated server, a ‘super-computer’, or in the cloud.

Parallelization

Like airport security, computers can run in parallel

Modern computers have more than one processor

Typically, these are referred to as “cores”.



Intel cores are “hyper-threaded”, which means that the operating systems sees these 4 cores as 8 CPUs.

```
[MDSTENGL-M01:Desktop _mdstengl$ sysctl hw.physicalcpu hw.logicalcpu
hw.physicalcpu: 4
hw.logicalcpu: 8
```

You can speed up your analyses by taking advantage of multiple CPUs/cores/threads.

Lots of times, an analysis will run quickly and you don’t need to worry about this.

Some computational tasks, like tree inference or genome assembly are computationally intensive and you should take advantage of multi-threading if available.

1. Some programs have built-in multi-threading.

```
-num_threads <Integer, >=1>
  Number of threads (CPUs) to use in the BLAST search
  Default = `1'
```

1. If your computer has multiple processors/cores, use `-p`

   The `-p` option causes Bowtie 2 to launch a specified number of parallel search threads. Each thread runs on a different processor/core and all threads find alignments in parallel, increasing alignment throughput by approximately a multiple of the number of threads (though in practice, speedup is somewhat worse than linear).

2. You can also run multiple instances of single-threaded programs
   1. You can do this manually
   2. parallel is a linux command line utility for doing this
   3. “Workflow managers” like Nextflow or snakemake allow parallelization.

Speed-up from running an analysis in parallel

map 1M 50 nt reads to the human genome with 1 CPU: 1m 16 seconds

```
mdstengl@cctsi-104:~/analyses/test_human_mapping$ time ./run_bowtie ERR3252925_1_1M.fastq GCA_000001405.15_GRCh38_no_alt_an
alysis_set.fna.bowtie_index
bowtie2 -x GCA_000001405.15_GRCh38_no_alt_analysis_set.fna.bowtie_index -q -U ERR3252925_1_1M.fastq --local --score-min C,1
00,1 --no-unal --threads 1 -S ERR3252925_1_1M.fastq.GCA_000001405.15_GRCh38_no_alt_analysis_set.fna.bowtie_index.sam

real    1m16.427s
user    1m13.836s
sys     0m12.844s
```

map 1M 50 nt reads to the human genome with 24 CPU: 9.6 seconds: ~8x faster

```
mdstengl@cctsi-104:~/analyses/test_human_mapping$ time ./run_bowtie_multiple_threads ERR3252925_1_1M.fastq GCA_000001405.15
_GRCh38_no_alt_analysis_set.fna.bowtie_index
bowtie2 -x GCA_000001405.15_GRCh38_no_alt_analysis_set.fna.bowtie_index -q -U ERR3252925_1_1M.fastq --local --score-min C,1
00,1 --no-unal --threads 24 -S ERR3252925_1_1M.fastq.GCA_000001405.15_GRCh38_no_alt_analysis_set.fna.bowtie_index.sam

real    0m9.641s
user    1m38.696s
sys     0m33.124s
```

How do you check resource usage and availability?

On your workshop laptops:

- How much memory is available? How much is used?
- How much storage is available? What type is it? How much is used?
- How many cores and threads are there?   What fraction of these are being used?

Checking resource usage and availability through the command line

command		better way to run it
`df`	reports disk usage and storage	`df -h`
`top`	reports CPU & memory usage	`htop`

Use these commands to answer the same questions:

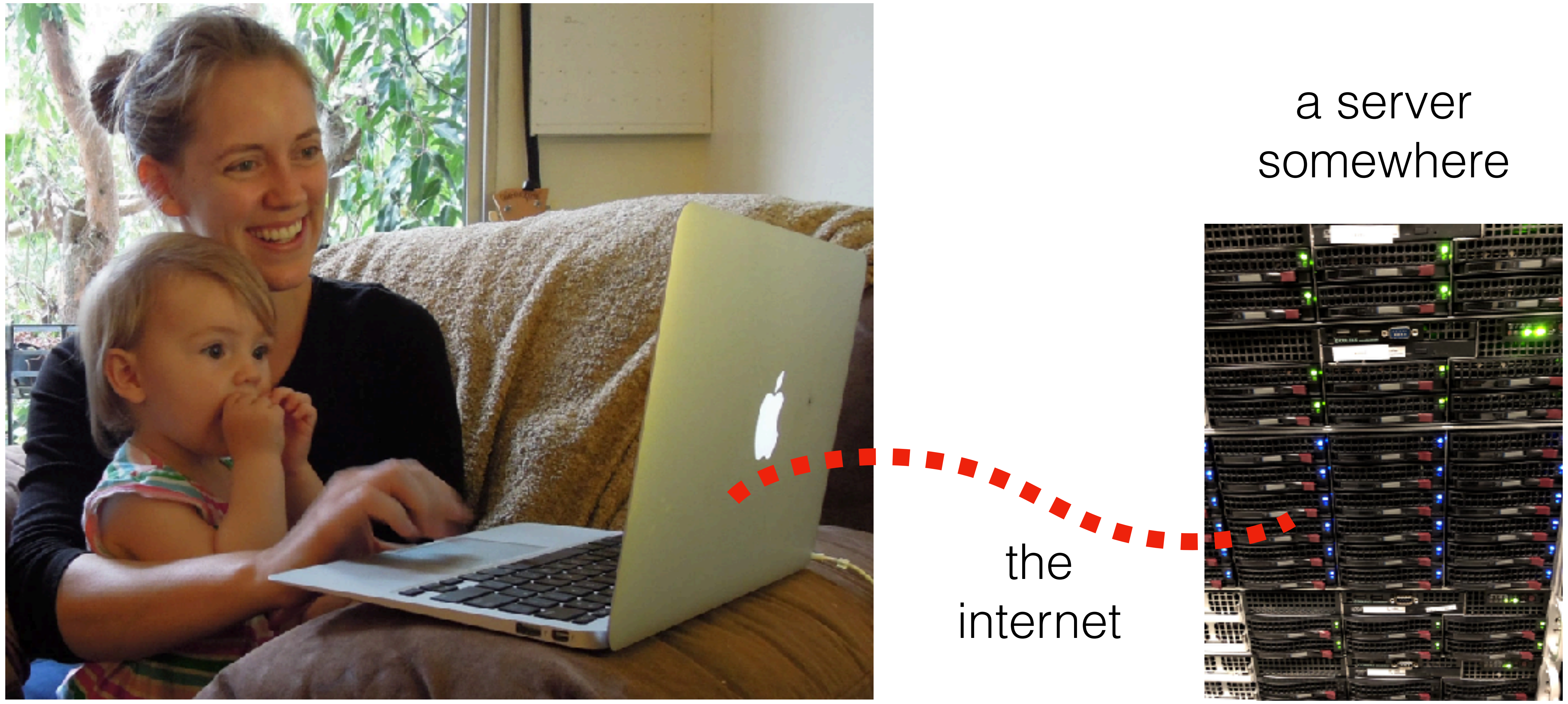
- How much storage is available? How much is used?
- How much memory is available? How much is used?
- How many cores and threads are there? What fraction of these are being used?

Local vs. remote computing

Local: you are using the resources on your computer

Remote: you are using the resources of a distant computer, but probably still connected through your own computer

a server somewhere



Connecting to remote computers on the command line

Typically you connect to a remote server via ssh (secure shell). For instance, to connect to a server at CSU I would run:

```
ssh mdstengl@cctsi-104.cvmbs.colostate.edu
```

Use the df and stop commands to answer the same questions about this server

- How much storage is available? How much is used? (in /home)
- How much memory is available? How much is used?
- How many cores are there? What fraction of these are being used?

Cloud computing is an increasingly popular form of remote computing

Advantages:
- Scalable and flexible
- Don’t have to buy, host, or maintain servers
- Can take advantage of pre-existing images and containers

Disadvantages:
- Can be expensive
- Have to pay to store and transfer data
- Can be slow to transfer data

Real physical computing resources are available as virtual computers through the internet

image: gigabitmagazine.com

Cost of cloud computing depends on how many resources you need

	vCPU	ECU	Memory (GiB)	Instance Storage (GB)	Linux/UNIX Usage
t2.large	2	Variable	8 GiB	EBS Only	$0.1104 per Hour
t2.xlarge	4	Variable	16 GiB	EBS Only	$0.2208 per Hour
r5.12xlarge	48	173	384 GiB	EBS Only	$3.36 per Hour
r5.24xlarge	96	347	768 GiB	EBS Only	$6.72 per Hour

Similar to the workshop laptops

Similar to the servers my lab uses

Plus ~$0.05-$0.15 per Gb-month for (short term) storage: $50-$150 per month per Tb

Plus ~$0.09 per Gb to transfer from Amazon to anywhere else: $90 per Tb

An Amazon Web Services (AWS) linux environment