Hardware and Infrastructure

What a Data Scientist should know

—

Michael Gat

General Assembly Santa Monica, DSI Summer 2016

@michaelgat

http://github.com/michaelgat/Presentations/DS_Hardware.pdf

Agenda

- Run through material (30 min)
- Q & A as long as you like
- Optional
  - Play with hardware/demo
  - PC Part Picker exercise

Hardware? Why should I care?

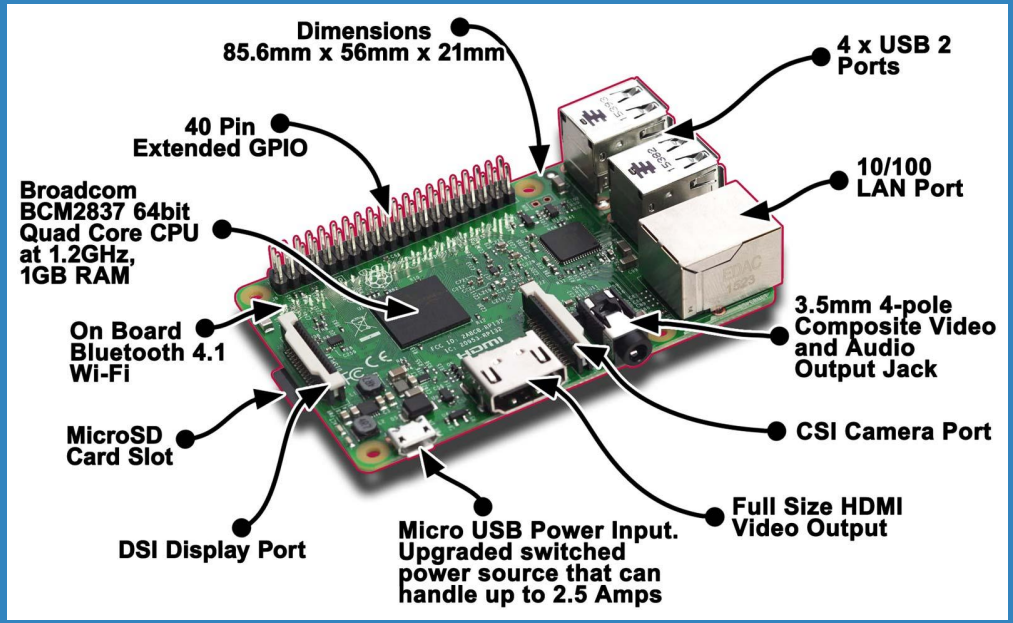
- Hardware/network/infrastructure knowledge was once necessary
  - You needed to have detailed knowledge just to get your programs to run!
  - Later it became merely advantageous
- Today knowledge of underlying systems seems unimportant
  - But is it really?
- It is easy to write code that works, but...
  - Writing *optimal* code still benefits from a bit of knowledge about how systems work
  - Defining what systems you need requires some knowledge
    - This is true whether it's on your desktop, on a server, or on AWS
    - As a Data Scientist you can provide value just by having a bit of informed input
- Google cares, so you should care!

Besides...

You don’t want to be this guy:

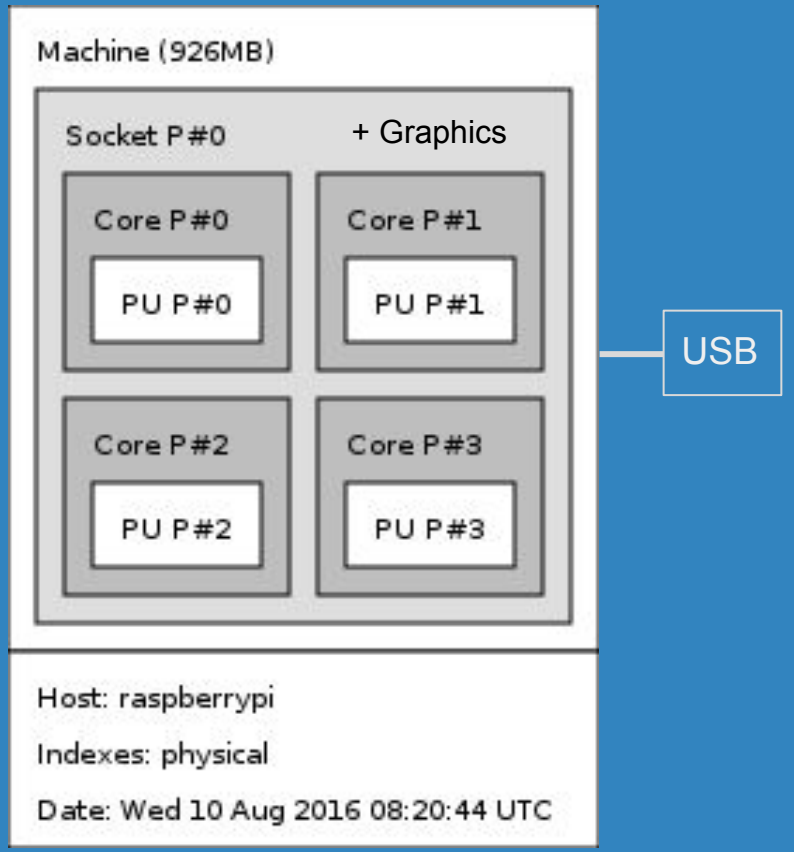
What's a computer? Here's a simple one

- Does anything any other computer does
  - Has a CPU
  - Has memory
  - Has storage
  - Has input/output
  - Has networking
  - Has connections for peripherals/accessories



Really, really simple

- Not much to see here
  - Four cores
  - Memory
  - CPU manages graphics
- Also
  - USB devices
    - Storage
    - Network
  - Graphics (on CPU)

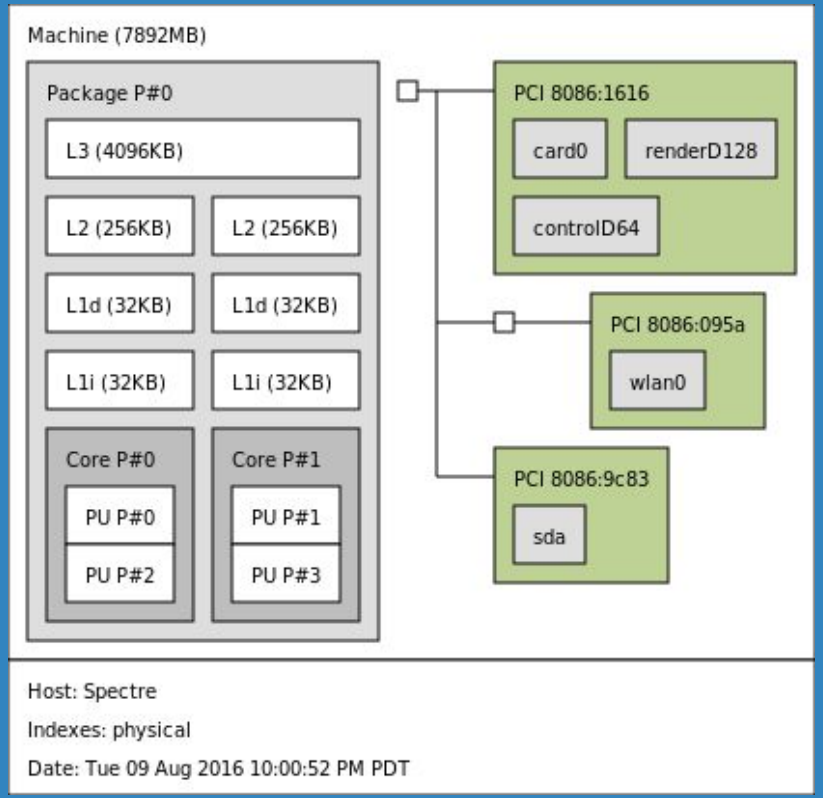


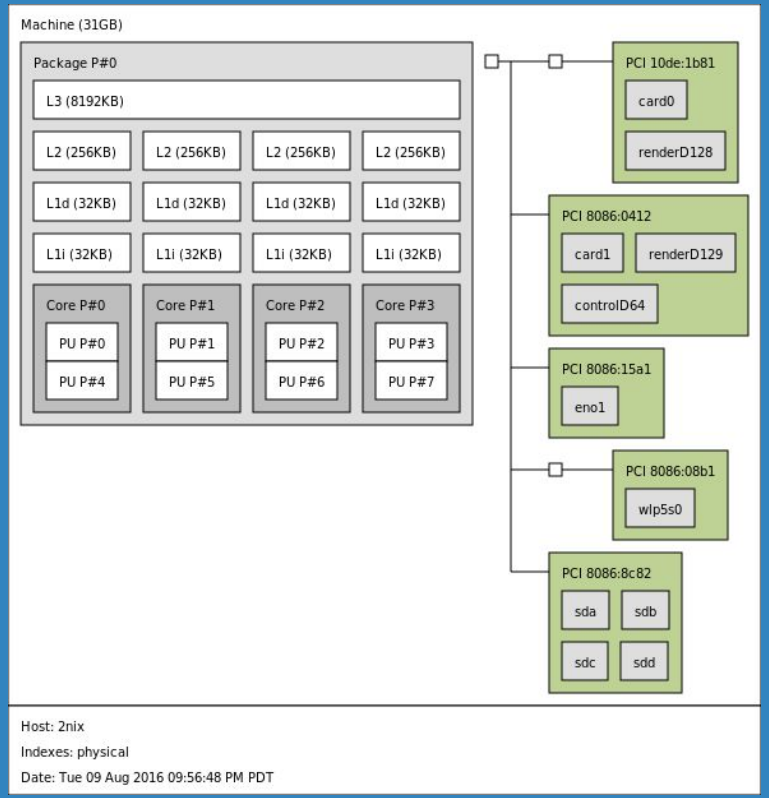
Don’t believe me?
(quick demo)

Every computer does the same things

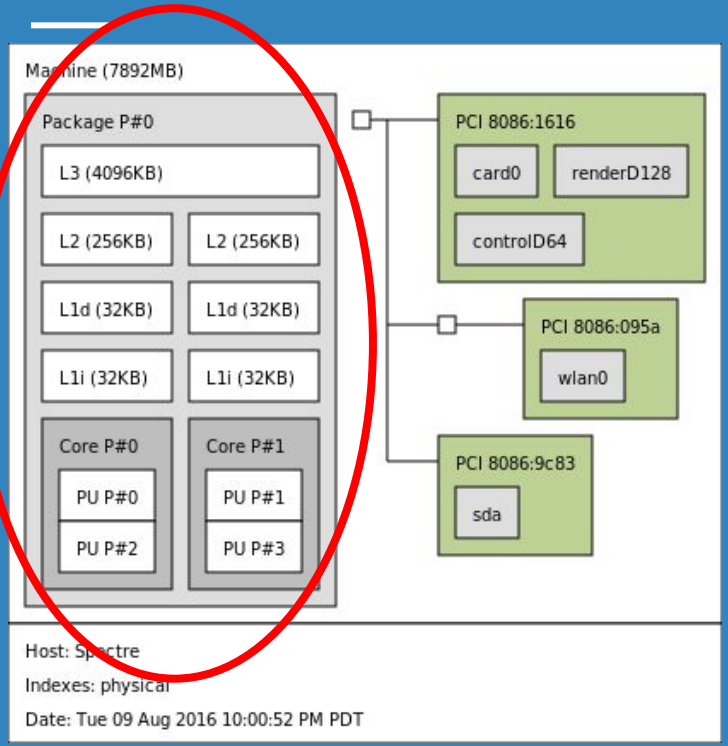
- A CPU is at the center
  - It may have more than one “core”
  - It may include graphics or other “helper” pieces on board
- The CPU reads from and writes to dynamic (short term) memory
- The CPU reads from and writes to persistent storage
- The CPU receives user input (keyboard, mouse, microphone, camera, etc.)
  - Much of a CPU’s time in general purpose computing is spent waiting for input
- The CPU creates user output (screen/graphics, speaker, VR)
  - The CPU will often use a special sub-processor or accessory processor to handle video. This is called a Graphics Processing Unit or GPU. (Remember this, it’s important for Data Science!)
- It uses an Operating System to manage how the pieces work together

Some more examples





CPU: The heart of it all

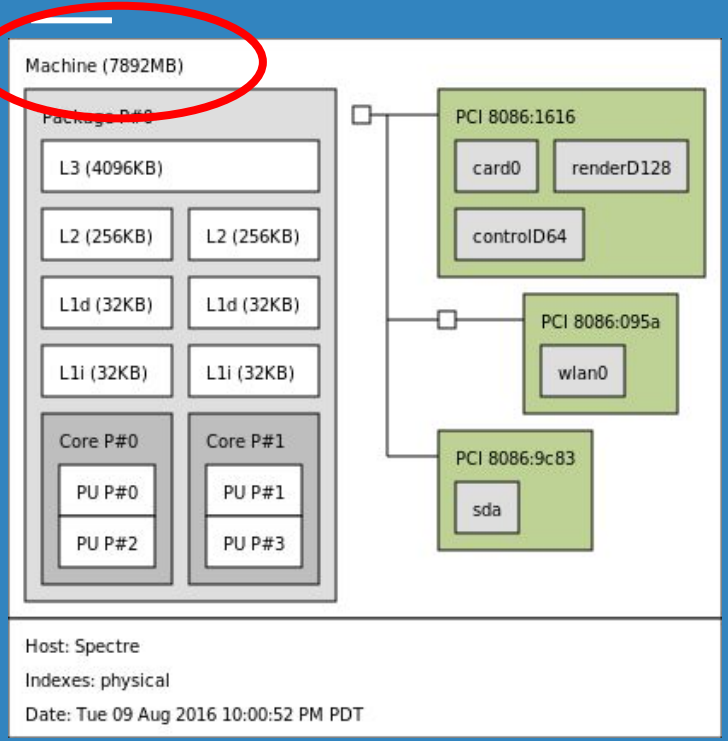


- Performs general purpose processing
  - Jack of all trades, master of none
- Modern CPUs include multiple *cores* that share some memory and resources
- Each core can have multiple *processing units* (aka “threads”) that allow two programs to run simultaneously
  - Threads share resources with other threads in the same core
  - This can be beneficial... or not

CPU considerations

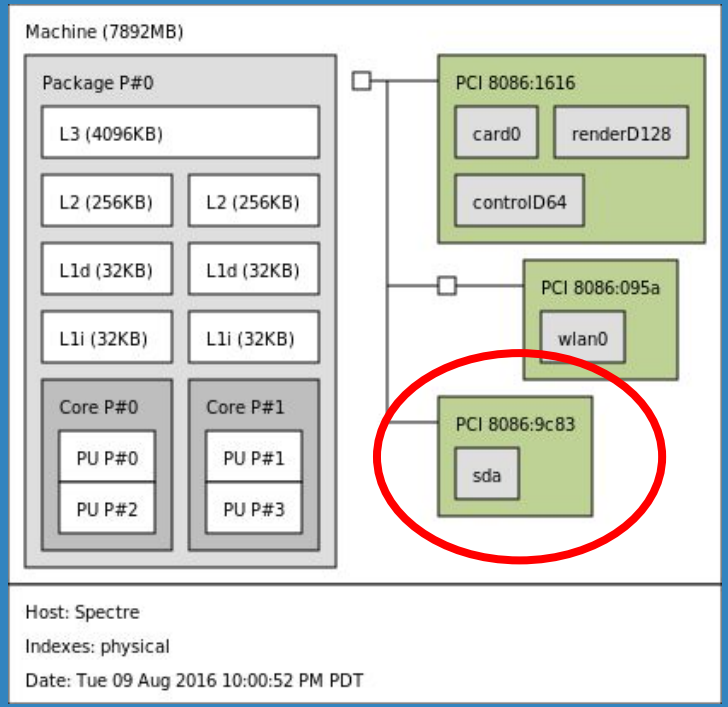
- Lots of cores is nice, especially if you have a parallelizable model
- Multithreading is not helpful with mathematically dense computation
  - http://eli.thegreenplace.net/2016/c11-threads-affinity-and-hyperthreading/
  - It rarely hurts, but doesn't benefit us much in many cases.
- Processor "speed" is not always the best metric to consider
- Mobile versions are always pared down and/or different
  - Product names may mean different things in the "mobile" version!
- The limits on general purpose computation are often outside the CPU
- The limits on mathematically intensive computation are in the CPU

Memory: Your short-term workspace



- It's where programs run
- Lots of it is good
- If you don't have at least 8MB WAYSA?
  - For Data Science, more is better
- Essential for large "in memory" datasets
  - PANDAS, Spark, etc.
- Essential for virtualization
- Makes things run faster even when data is primarily stored on disk
  - SQL DBs can use a lot of memory

Storage: Where your data lives



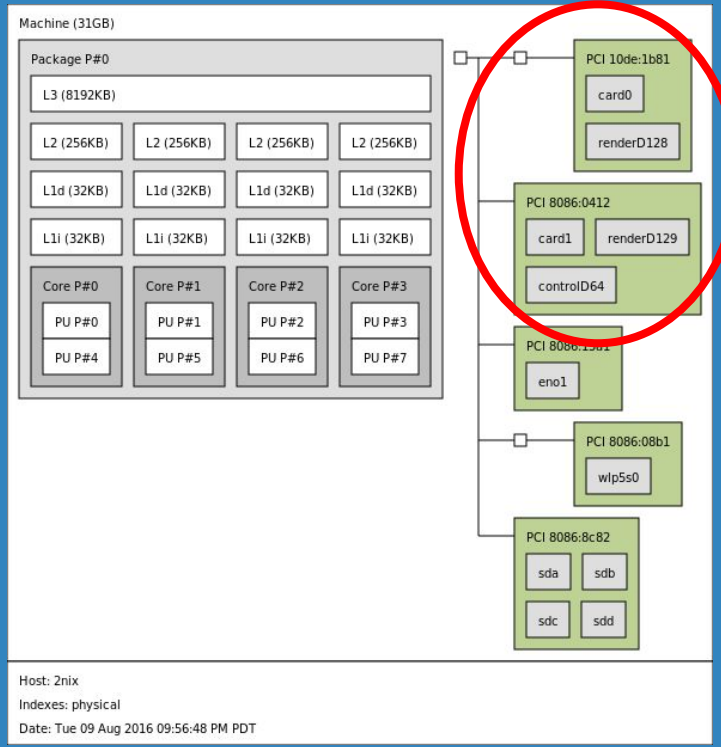
- Solid state (Flash memory)
  - Fast, expensive
  - Great for running programs
  - Limited read/write cycles
- Disk (electro-mechanical)
  - Slower, less expensive
  - Good if there is processing lag
  - Scale and combine well
  - Limited lifetime
- Some devices combine them

More considerations for storage/memory

- Most programs will shift data to memory when they can
- But in other situations, storage substitutes for memory
  - SSD as primary storage helps with this
- On larger servers or workstations, there may be multiple disk or storage devices
  - They may be organized into large logical devices/arrays
- Backups or redundancy are essential.
- Network storage is usually disk, and usually slower
  - This includes some of the storage on AWS

What's with GPUs?

- GPU - Graphics Processing Unit
  - A specialized processor that creates video output
  - High end GPUs designed for 3D gaming
  - Most CPUs have limited built-in GPU functionality -- good enough for watching video
- GPUs are optimized for floating-point and vector math
  - Very parallelized
  - Very fast on limited types of instructions
  - First discovered by Bitstream miners, now used for many math-intensive operations.
    http://www.nvidia.com/object/tesla-supercomputing-solutions.html
  - Ideal for many data science applications
- Available as a service on AWS, but beware the costs!



CUDA - the dominant GPU standard

- Developed by nVidia
- Many levels of parallel structures
- Each green square is a math processing (CUDA) unit
- 20 SM units, each of which is a full-blown processor
- Can provide 10-20x improvement over CPU for many Data Science applications.

http://www.nvidia.com/object/data-science-analytics-database.html

Some thoughts about GPUs

- There is little availability of top GPUs on consumer notebooks
  - Including Macbooks
- Mostly on “mobile workstations” running Linux
- Mobile versions (as with CPUs) are pared down/throttled
- Likely to become a de-facto requirement for mathematically dense applications in the next few years
- CUDA is emerging as the most supported architecture
- Look for “Pascal” GPUs (GTX10xx)

Key points for Data Science

- There is a GPU in your future whether you like it or not
  - Make it an *nVidia GTX 10xx*
- Memory is a big deal for us.
- CPU physical cores matter, logical ones (threads) less
- CPU and basic storage are cheap, other items less so
- You can do anything on any computer
  - But your time is worth something

You shouldn't have asked...

- Intel i7-4790K (4.0GHz)
  - Liquid cooling
- 32 GB RAM
- 240 GB SSD
- 3 x 2TB disk in RAID 5 array
  - 4TB useful storage
- nVidia GTX 1070 GPU
- 650w Power Supply
- 2 x 140mm fans in
- 1 x 140mm fan out
- 1x120mm fan out (on radiator)
- Wifi card (unused)
- 2 x 24" 1920 x 1200 monitors
- https://pcpartpicker.com/user/michaelgat/saved/#view=PT6cCJ

Questions?

http://github.com/michaelgat/Presentations/DS_Hardware.pdf

@michaelgat