Measuring the power draw of computers

What you cannot measure, you cannot improve

Mercredi 19 Mai

Power draw of computers

First message

- Most of IT carbon footprint comes from manufacturing
  building computers, smartphone, internet cables, telecom sattelite,...

Still it's worth to monitor our usage,

- Large waste of the computing power in development data centers (K. Khan et al. 2019)
- Our models are over calibrated(Parcollet and Ravanelli 2021)
- We can do just as good with less

This presentation aims at persuading you to measure the energy used by your algorithm

Power draw of computers

First message

- Most of IT carbon footprint comes from manufacturing
  building computers, smartphone, internet cables, telecom sattelite,...

Still it's worth to monitor our usage,

- Large waste of the computing power in development data centers (K. Khan et al. 2019)
- Our models are over calibrated(Parcollet and Ravanelli 2021)
- We can do just as good with less

This presentation aims at persuading you to measure the energy used by your algorithm

Power draw of computers

First message

- Most of IT carbon footprint comes from manufacturing
  building computers, smartphone, internet cables, telecom sattelite,...

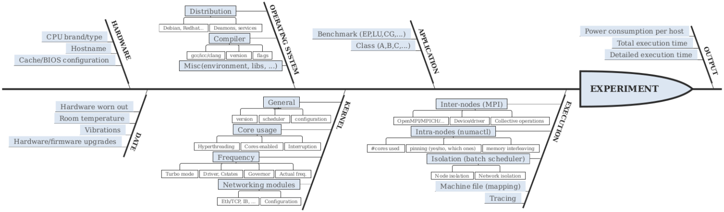
Still it's worth to monitor our usage,

- Large waste of the computing power in development data centers (K. Khan et al. 2019)
- Our models are over calibrated(Parcollet and Ravanelli 2021)
- We can do just as good with less

This presentation aims at persuading you to measure the energy used by your algorithm

A not so trivial topic

- Difficulty to isolate the energy hungry elements
- Dependent on the built in sensor and constructor support.
- Low level (close to hardware) programming
- Energy depends on the lot of parameters



Picture from Orgerie 2020

What we learn in highschool

- **Joule**: energy transferred to an object when a force of one newton acts on that object in the direction of the force's motion through a distance of one metre (1 newton-metre or Nm)
  - The energy required to lift a medium-sized tomato up 1 metre
- **Watt**: 1 joule per seconds
- **kWh**: ????? Joules

What we learn in highschool

- **Joule**: energy transferred to an object when a force of one newton acts on that object in the direction of the force's motion through a distance of one metre (1 newton-metre or Nm)
    - The energy required to lift a medium-sized tomato up 1 metre
- **Watt**: 1 joule per seconds
- **kWh**: **3600000** Joules
    - 3 hours of GPU computation

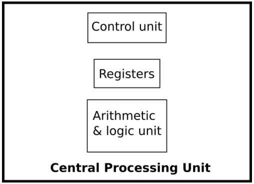
What we learn in highschool

- **Joule**: energy transferred to an object when a force of one newton acts on that object in the direction of the force's motion through a distance of one metre (1 newton-metre or Nm)
  - The energy required to lift a medium-sized tomato up 1 metre
- **Watt**: 1 joule per seconds
- **kWh**: **3600000** Joules
  - 3 hours of GPU computation

How a computer uses energy?

What we learn at the university

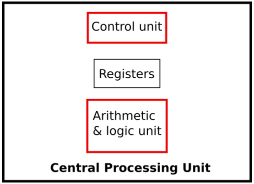
Let's start with the cpu



- From 100Khz in 1971 to some Ghz today
- Composed of millions of transistors (Moore law)
- Cristal of qwartz giving the frequency of the cpu
- Optimization of the frequency to save power (turboboost)

What we learn at the university

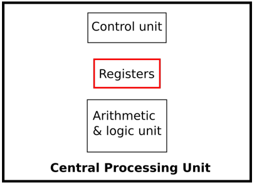
Let's start with the cpu



One cpu Core

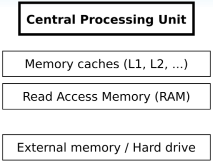
- Instructions set : boolean, floating operations
  - RISC (AMD), CISC (Intel), dedicated FPGA instructions
    `/proc/cpuinfo`
- Conditions the power draw
- Low level programmation with binary networks

Let's start with the cpu



- Registers : fast memory used by the ALU
- 10-100 registers with 8-64 bits

and continue with the memory

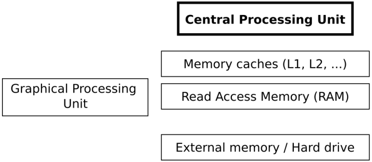


- Memory hierarchy
  - Closer to the cpu → smaller and faster

```
$ lscpu
L1d cache:                        384 KiB
L1i cache:                        256 KiB
L2 cache:                         4 MiB
L3 cache:                         16 MiB
```

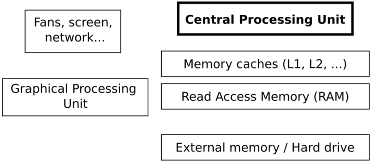
- Moving data up and down the memory hierarchy costs time and power
- Taken into account in optimization code to limit these moves.
  - Eg: Row major or column major storage in matrix multiplication

GPU : major actor in the consumption



- Consumes more than the whole computer (Bridges, Imam, and Mintz 2016)

Other components



- Consumes more than the whole computer (Bridges, Imam, and Mintz 2016)
- Overall a full a diagnostic might be complex
  - lack of available sensors

GPU versus CPU

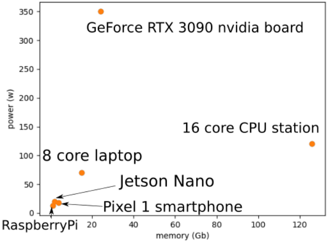
- Invented by nvidia in 1999
- Thousands of cores to enable parallelism
- Lower amount of RAM memory available
- Higher latency : GPU clock speed $<$ CPU clock speed
- Higher memory throughput : GPU operates on larger chunks of data
  - GPU can fetch data from its RAM more quickly
  - CPU bandwidth $<$ GPU bandwidth
- Smaller set of instructions dedicated to graphics and matrix calculus
- More power hungry and requires a CPU

Energy efficient since the computations is faster.

Other hardwares

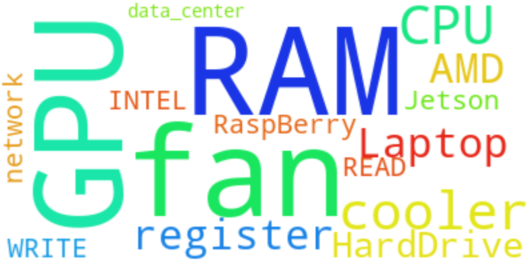
- AMD CPU: RISC instruction set lower energy than Intel processors
- Programmable circuits with custom instruction set
  - Field-programmable gate array
  - Application-specific integrated circuit (ASIC):
    Implements the Tensor Processing Unit.
- Small devices
  - Rasberrypi
  - Jetson Cards
- Neuromorphic sensors (Guillaume Bellec presentation this morning)

Some perspective numbers



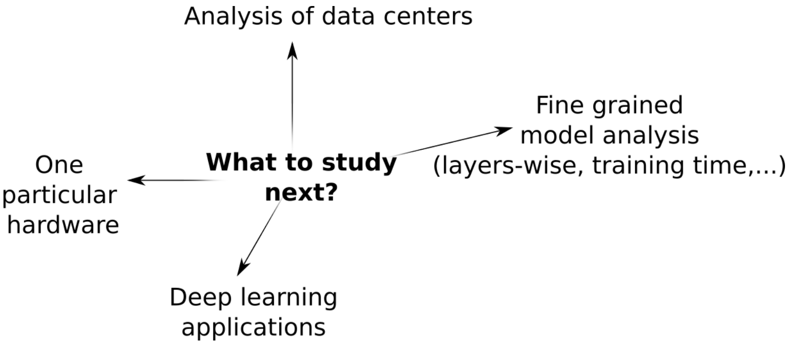
Power usage versus memory capacity

- How to rank machines by efficiency ?
- Compromise between, power, memory, computing capacity



How to measure all of it?

Different angles to tackle



Related work on consumption measurements

- Opensource libraries for machine learning carbon footprint (Henderson et al. 2020; Anthony, Kanding, and Selvan 2020)
  - based on RAPL and nvidia-smi
- Fine grained studies on a specific Jetson hardware (Rodrigues, Riley, and Luján 2018; Holly, Wendt, and Lechner 2020; Arafa et al. 2020)
- Generic libraries from the data center community : Papi, Likwid
- Machine learning based prediction models (Cai et al. 2017, Jia et al. 2015)
- French Startup : https://github.com/hubblo-org

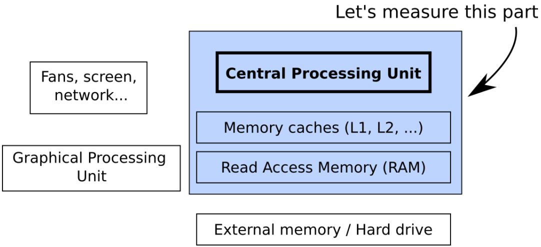
Hard to get recover exactly what you measure on your power meter.
Developping from scratch requires complex low level programming skills

Related work on consumption measurements

- Opensource libraries for machine learning carbon footprint (Henderson et al. 2020; Anthony, Kanding, and Selvan 2020)
  - based on RAPL and nvidia-smi
- Fine grained studies on a specific Jetson hardware (Rodrigues, Riley, and Luján 2018; Holly, Wendt, and Lechner 2020; Arafa et al. 2020)
- Generic libraries from the data center community : Papi, Likwid
- Machine learning based prediction models (Cai et al. 2017, Jia et al. 2015)
- French Startup : https://github.com/hubblo-org

Hard to get recover exactly what you measure on your power meter.
Developping from scratch requires complex low level programming skills

RAPL to measure Intel CPUs



RAPL to measure Intel CPUs

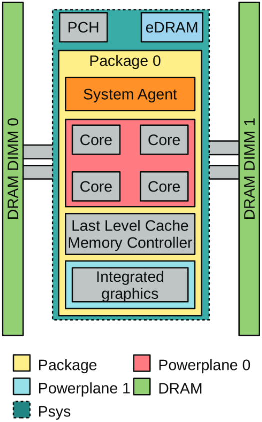
Running Average Power Limit

- Present since the Sandy bridge architecture in 2011
- Now supported by integrated voltage regulators in addition to power models
- Reports the accumulated energy consumption
- Recording at 1000Hz
- Requires administrator privilege

RAPL Organisation

Different counters for physically meaningfull domains:

- Power Plane 0 : CPU
- Power Plane 1 : Processor graphics on the socket.
- DRAM : energy consumption of the RAM
- Psys : System on Chip energy consumption



K. N. Khan et al. 2018

Access to RAPL measurements

- Model specific registers

  `/dev/cpu/core_id/msr`

  - Read MSR register bit by bit (not trivial)
  - See intel documentation (not trivial)
  - And activate the kernel module

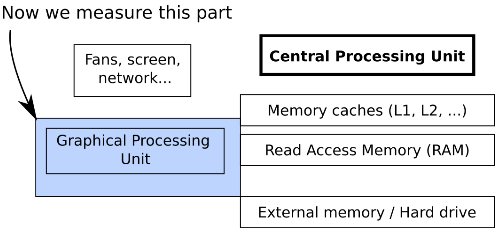
    ```
    sudo modprobe msr
    ```

- **Linux**: Exposition of a sysfs tree with powercap

  Accumulation of energy consumption in Joules

  ```
  sudo chmod -R 755 /sys/class/powercap/intel-rapl/
  ```

nvidia-smi to measure Nvidia GPUs



nvidia-smi

NVIDIA System Management Interface, based on top of the NVIDIA Management Library (NVML, cuda v4.1, 2011)

- Gpu global statisics and memory usage per process

  ```
  $ nvidia-smi -q -x
  ```

  - The power consumption is given for the entire board
  - +/- 5% accuracy of current power draw.
  - Memory usage per gpu and per process
  - Percentage of usage of each gpu

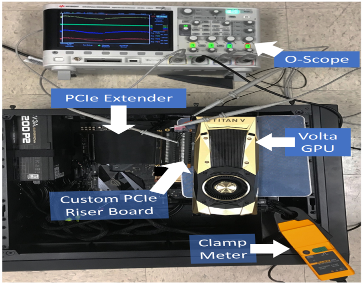
nvidia-smi

NVIDIA System Management Interface, based on top of the NVIDIA Management Library (NVML, cuda v4.1, 2011)

- Per process Average utilization values for streaming multiprocessors (SM)

```
$ nvidia-smi pmon # up  to  4  devices
# gpu        pid  type   sm   mem   enc   dec   command
# Idx          #   C/G    %     %     %     %   name
    0       1114     G    -     -     -     -   Xorg
    0       1289     G    -     -     -     -   gnome-shell
    0    1135553     C   76     0     -     -   python
```

Fine grained measurement



Arafa et al. 2020

- Fine grained measurement at instruction level
- Verification with powermeters.

Let's dive into practice!

Deep Learning Power Measure @UPPA

Clone AIPowerMeter from github!

We are developing (yet another) python module for :

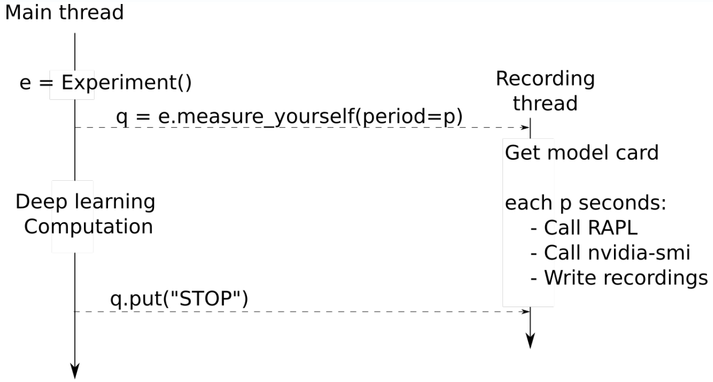
- Recording the power of a specific process
- Focus on accessibility and analysis for data scientist
- Model card, number of parameters and macs

```
process, queue = exp.measure_yourself(period=2)

 ###################
#  place here the code that you want to profile
################

q.put(experiment.STOP_MESSAGE)
```

Multi threading under the hood



- Energy recording only for the main thread
- Queue to communicate between the threads

Get power draw by process

- RAPL and nvidia-smi provides the global power consumption
- Using memory and processor usage from psutil to obtain the consumption by program
- However some of the components are shared from all programs.

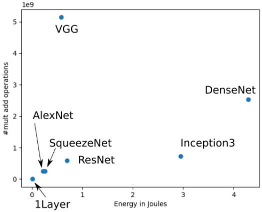
Divide in equal parts? ignore these parts?

Experiment

Let's test classics network on a random synthetic image

- Energy consumed by 200K forward passes
- input image is $(3 \times 128 \times 128)$
- AlexNet, VGG, Resnet, SqueezeNet, DenseNet, Inception
- 1 convolutional layer with a (3×3) kernel

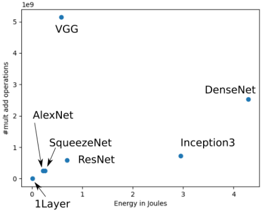
RAPL Organisation

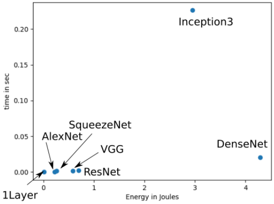


\# mult add *versus* power

- How good or bad are proxy measures ?

RAPL Organisation



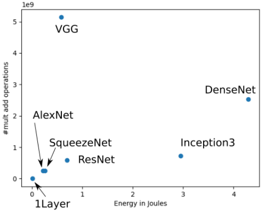


\# mult add *versus* power

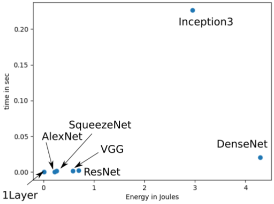
- can you be slow and low power ?

time *versus* power

RAPL Organisation



mult add *versus* power



time *versus* power

- Two factors here : Duration and usage !

A lot to discover for deep learning!

Join the community

- SustaiNLP 2020: Workshop on Simple and Efficient Natural Language Processing
- Low-Power Computer Vision Challenge since 2015

Or just be better at optimizing (understanding) your program:

```
torch.backends.cudnn.benchmark = True
```

A lot to discover for deep learning!

Join the community

- SustaiNLP 2020: Workshop on Simple and Efficient Natural Language Processing
- Low-Power Computer Vision Challenge since 2015

Or just be better at optimizing (understanding) your program:

```
torch.backends.cudnn.benchmark = True
```

References I

Anthony, Lasse, Benjamin Kanding, and Raghavendra Selvan (July 2020). "Carbontracker: Tracking and Predicting the Carbon Footprint of Training Deep Learning Models". In: arXiv preprint https://arxiv.org/abs/2007.03051.

Arafa, Yehia et al. (2020). "Verified instruction-level energy consumption measurement for nvidia gpus". In: **Proceedings of the 17th ACM International Conference on Computing Frontiers**, pp. 60–70.

Bridges, Robert A, Neena Imam, and Tiffany M Mintz (2016). "Understanding GPU power: A survey of profiling, modeling, and simulation methods". In: **ACM Computing Surveys (CSUR)** 49.3, pp. 1–27.

Cai, Ermao et al. (2017). "Neuralpower: Predict and deploy energy-efficient convolutional neural networks". In: **Asian Conference on Machine Learning**. PMLR, pp. 622–637.

References II

Henderson, Peter et al. (2020). "Towards the Systematic Reporting of the Energy and Carbon Footprints of Machine Learning". In: **ArXiv** abs/2002.05651.

Holly, Stephan, Alexander Wendt, and Martin Lechner (2020). "Profiling Energy Consumption of Deep Neural Networks on NVIDIA Jetson Nano". In: **2020 11th International Green and Sustainable Computing Workshops (IGSC)**. IEEE, pp. 1–6.

Jia, Wenhao et al. (2015). "GPU performance and power tuning using regression trees". In: **ACM Transactions on Architecture and Code Optimization (TACO)** 12.2, pp. 1–26.

Khan, K. et al. (2019). "Analyzing the power consumption behavior of a large scale data center". In: **SICS Software-Intensive Cyber-Physical Systems** 34, pp. 61–70.

References III

Khan, Kashif Nizam et al. (2018). "Rapl in action: Experiences in using rapl for power measurements". In: **ACM Transactions on Modeling and Performance Evaluation of Computing Systems (TOMPECS)** 3.2, pp. 1–26.

Orgerie, Anne-Cécile (2020). "From Understanding to Greening the Energy Consumption of Distributed Systems". PhD thesis. Ecole Nationale Supérieure de Rennes.

Parcollet, Titouan and Mirco Ravanelli (2021). "The Energy and Carbon Footprint of Training End-to-End Speech Recognizers". In:

References IV

Rodrigues, Crefeda Faviola, Graham Riley, and Mikel Luján (2018). "SyNERGY: An energy measurement and prediction framework for Convolutional Neural Networks on Jetson TX1". In: **Proceedings of the International Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA)**. The Steering Committee of The World Congress in Computer Science, Computer . . ., pp. 375–382.