INSTITUTE
OF COMMUNICATION,
INFORMATION
AND PERCEPTION
TECHNOLOGIES

Scuola Superiore
Sant'Anna

ITU ACADEMIA

An Approach for Estimating Energy Consumption of AI Hardware

Luca Valcarenghi
Scuola Superiore Sant'Anna

ITU virtual workshop on AI and environmental efficiency
Session 2
Assessment and Measurement of the Environmental Efficiency of AI and Emerging Technologies

December 9, 2020

Start from the Basics

- Joule (SI)
  - Energy expended (or work done) in applying a force of one newton through a distance of one metre (1 newton metre or N·m), or in passing an electric current of one ampere through a resistance of one ohm for one second

James Prescott Joule
(1818–1889)

POWER

$$J=\boxed{1\ \mathrm{kg}\cdot\mathrm{m}^2/\mathrm{s}^3}\ \mathrm{s} = \boxed{W}\ \mathrm{s}=\boxed{VA}\boxed{\mathrm{s}}$$

TIME

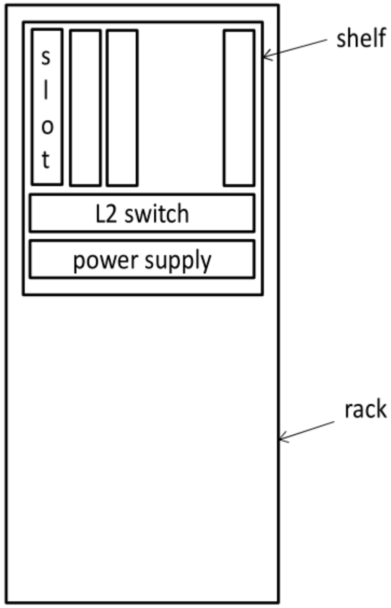
- From Energy to GHG emissions

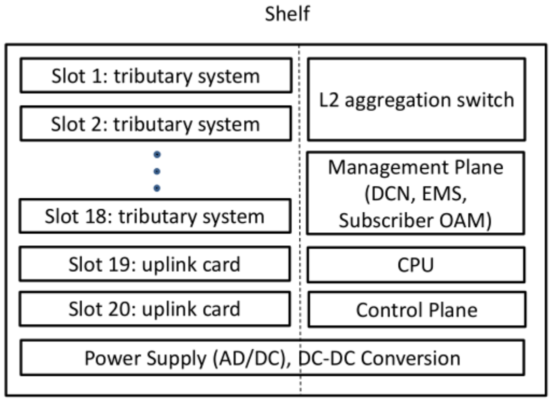
$$7.07 \times 10^{-4}\ \text{metric tons}\ CO_2\ /\ kWh$$

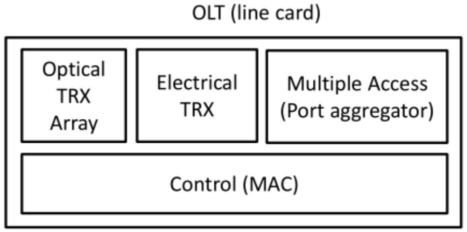
Source: https://www.epa.gov/energy/greenhouse-gases-equivalencies-calculator-calculations-and-references

Take a Holistic View of Energy Consumption

- NG-PON2 OLT real deployment model and power consumption contributions







Always ON

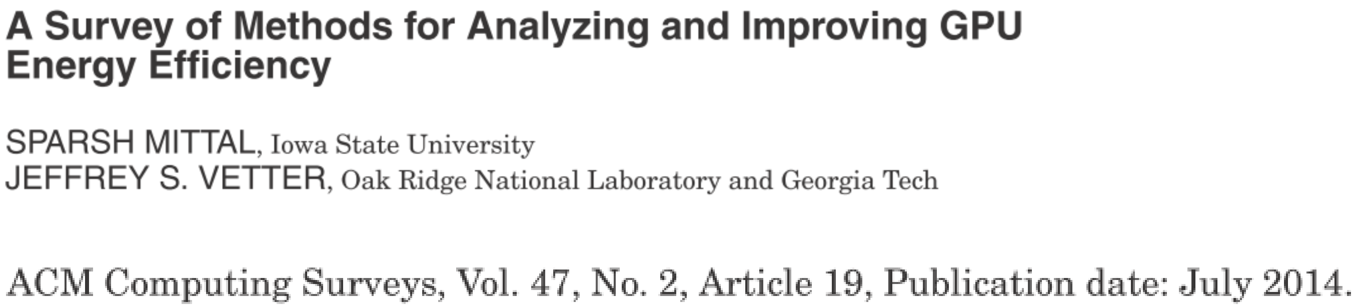
Component or Sub-system	Amount	Power consumption ON [W]	Power consumption OFF [W]
OLT			
4x10G TRX array TDMA coloured line card	2 line cards	16	0
Port aggregator	80 Gb/s	40	20
4xXG-PON Control MAC	8 ports	8	8
Shelf			
Basic shelf power	3 slots (2 TRX array, 1 MUX/DEMUX)	10	10
L2 Aggregation Switch	80 Gb/s	80	80
Total		154	118

Maximum achievable energy savings 23%

Consider Several Factors

- Hardware Type
  - CPU
  - GPU
  - FPGA
- Hardware architecture
- Application type
- Sustainable Development Goals (SDG) 2030 Goal 7
  - affordable, reliable, sustainable, and modern energy for all

Some examples



A Survey of Methods for Analyzing and Improving GPU Energy Efficiency

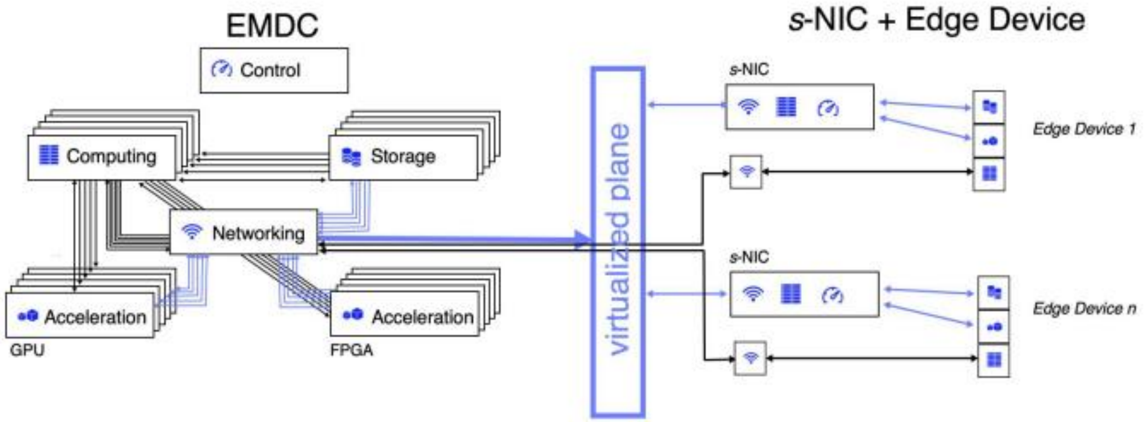
SPARSH MITTAL, Iowa State University
JEFFREY S. VETTER, Oak Ridge National Laboratory and Georgia Tech

ACM Computing Surveys, Vol. 47, No. 2, Article 19, Publication date: July 2014.

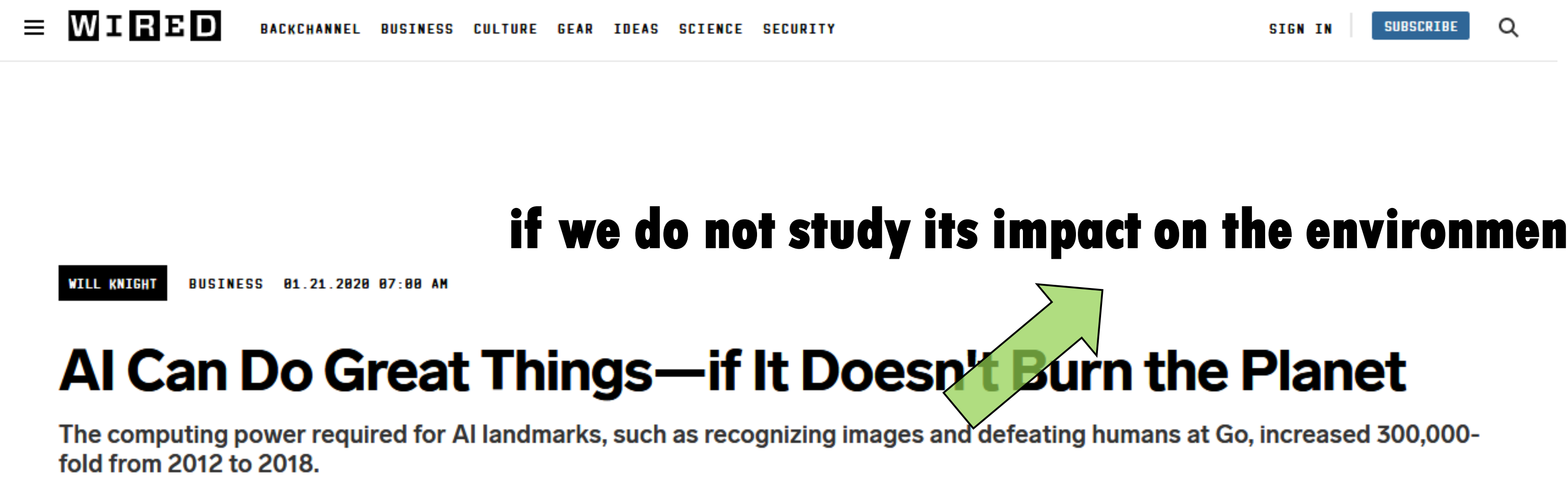
- Power consumption of GPUs can be divided into two parts
  - leakage power
    - Leakage power is consumed when the GPU is powered, even if there are no runtime activities
  - dynamic power
    - Dynamic power arises from switching of transistors and is determined by the runtime activities
- Techniques for improving GPU energy efficiency
  - DVFS (dynamic voltage/frequency scaling)-based techniques
  - CPU-GPU workload division-based techniques
  - Architectural techniques for saving energy in specific GPU components, such as caches
  - Techniques that exploit workload variation to dynamically allocate resources
  - Application-specific and programming-level techniques for power analysis and management

Some examples (2)

- BRAINE project (https://www.braine-project.eu/)
  - Edge Micro Data Center
  - Innovative integration of hardware and software components for efficient operation in embedded edge applications with very limited energy budget
  - Matching of different types of AI with different types of nodes/SoC, workload distribution/placement and switching/ communication costs
  - Exploitation of federated learning and edge cloud approaches
  - Novel cooling solutions



Conclusions



WILL KNIGHT BUSINESS 01.21.2020 07:00 AM

AI Can Do Great Things—if It Doesn't Burn the Planet

The computing power required for AI landmarks, such as recognizing images and defeating humans at Go, increased 300,000-fold from 2012 to 2018.

if we do not study its impact on the environment

Source: https://www.wired.com/story/ai-great-things-burn-planet/

Thank you

BRAINE

Thanks to:

Maria Rita Spada, Fred Buining, Federico Civerchia, Patrick Moder

LUCA.VALCARENGHI@SANTANNAPISA.IT

This work has received funding from the ECSEL Joint Undertaking (JU) under grant agreement No 876967. The JU receives support from the European Union's Horizon 2020 research and innovation programme and the Italian Ministry of Education, University, and Research (MIUR).