Green Computing and Sustainability

HW3 due tonight at midnight!

POMONA COLLEGE
COMPUTER SCIENCE
COLLOQUIUM

Sustainable AI: From Datacenter Infrastructure to Algorithm Design

Bilge Acun
META

Machine learning has witnessed exponential growth over the recent years. In this talk, we will first explore the environmental implications of the super-linear growth trend of AI from a holistic perspective, spanning data, algorithms, and system hardware. System efficiency optimizations can significantly help reducing the carbon footprint of AI systems. However, predictions show that the efficiency improvements will not be enough to reduce the overall resource needs of AI as Jevon's Paradox suggests "efficiency increases consumption". Therefore, we need to design our datacenters with sustainability in mind, using renewable energy every hour of every day. Furthermore, we need to design ML model architectures and hardware by taking carbon footprint into account.

This talk will introduce two sustainable design frameworks, Carbon Explorer and CATransformers. Carbon Explorer is a framework to design datacenters operating on renewable energy holistically by including embodied and operational footprints. CATransformers is a carbon-aware neural network and hardware architecture search framework that enables sustainability-driven co-optimization of ML models and accelerator hardware.

Bilge Acun is a Research Scientist at Meta AI (/FAIR). Her research lies in the intersection of energy efficient and sustainable system design and machine learning. Her work at Meta includes making large scale machine learning systems more efficient through algorithmic and system optimizations. She received her Ph.D. degree in 2017 at the Department of Computer Science at University of Illinois at Urbana-Champaign. Her dissertation was awarded 2018 ACM SigHPC Dissertation Award Honorable Mention. Before joining FAIR, she worked at the IBM Thomas J. Watson Research Center as a Research Staff Member.

FRIDAY, NOVEMBER 14
11AM
ZOOM

Outline

- Conventional wisdom concerning carbon and computing
- Re-quantifying carbon costs and proposed solutions
- Other environmental considerations

(From Friday) The GPU Programming Model

```
import pycuda.autoinit
import pycuda.driver as drv
import numpy

from pycuda.compiler import SourceModule
mod = SourceModule("""
__global__ void multiply(float *dest, float *a, float *b) {
    const int i = threadIdx.x;
    dest[i] = a[i] * b[i];
}
""")

a = numpy.random.randn(400).astype(numpy.float32)
b = numpy.random.randn(400).astype(numpy.float32)

dest  = numpy.zeros_like(a)
mod.get_function("multiply")(drv.Out(dest), drv.In(a), drv.In(b), block=(400, 1, 1), grid=(1, 1))

print(dest)
```

GPUs (and other accelerators) have their own instruction set… why?

The holistic functionality of a CPU is unnecessary for a GPU to compute if its task is to efficiently produce graphics!

Power = Energy / Time

There exists a trade-off between *generalizability* of computation and *specialization*

Generality versus Specialization

- Having a general-purpose compute unit (e.g., a CPU) means that there needs to exist components in a data path for *all possible paths* of execution through the instruction set!
- With a smaller instruction set, the data path will not need to have as many heterogeneous components to implement all instructions ➡️ components can be more closely tied to the computation that will happen (less waste)!
- The “heterogeneous accelerator” model of computation means that computation is *driven* by a CPU but all heavy lifting is offloaded to the appropriate accelerator
- In the context of GPUs, there has been a push for GPGPUs (general purpose GPUs) given their widespread adoption ➡️ larger instruction sets to launch programs from the GPU instead of the CPU

Chat with your neighbor(s)!

Suppose you are consulting for a firm that would like to setup a computation cluster for their workload. The client makes the argument that they should have an accelerator-heavy fleet because specialization is more efficient than general purpose compute. Do you agree or disagree with the client? If you agree, why? If you disagree, what would you say instead?

How much of the workload can be offloaded to the accelerators?

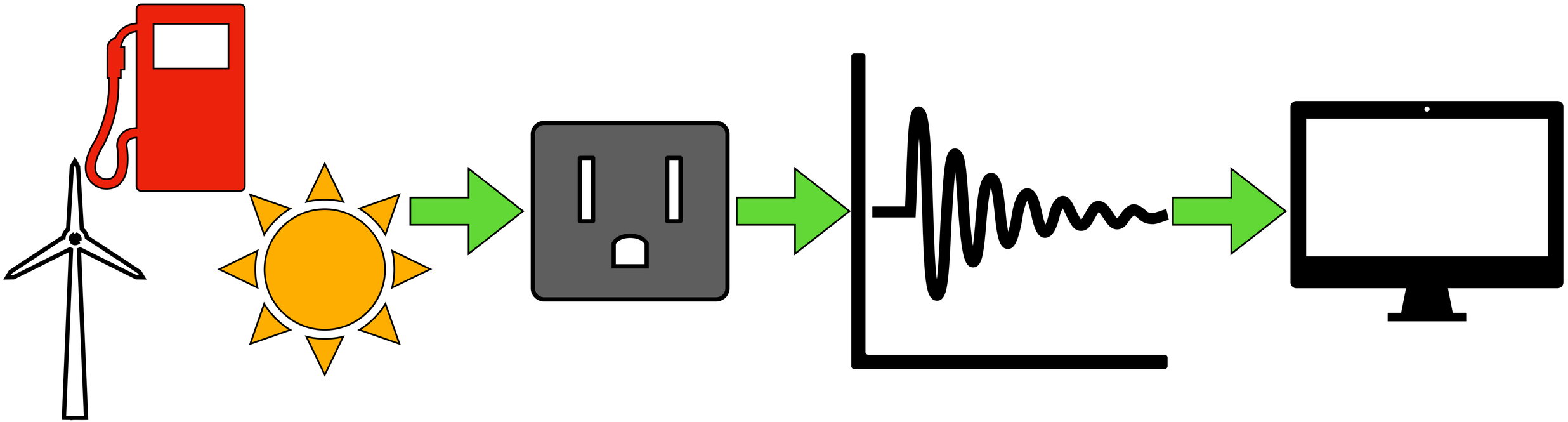
What percent of the accelerators will be idle, and for how long?

How long will it take to replace the accelerators versus the general purpose compute?

Device Lifetime

- Conventionally, specialization had been long-considered the solution to the power/energy problem ➡️ more efficient computation means less energy per unit compute
- Jevons Paradox: “Technological advances make a resource more efficient to use; however, as the cost of using the resource drops, … total consumption will rise”
- Modern reformulations of how we understand carbon costs take both efficiency *and* device lifetime into account ➡️ efficiency accounts for *operational* carbon consumption whereas device manufacturing describes *embodied* carbon costs

Understanding Operational Carbon



Understanding Operational Carbon

How long producing energy until the initial energy to produce plant is regenerated

Source	Carbon intensity (g CO_2/kWh)	Energy-payback time (months)
Coal	820	2 [33]
Gas	490	1 [33]
Biomass	230	∼12 [73]
Solar	41	∼36 [34]
Geothermal	38	72 [74]
Hydropower	24	∼12–36 [33], [75]
Nuclear	12	2 [33]
Wind	11	≤12 [35]

TABLE II
CARBON EFFICIENCY OF VARIOUS RENEWABLE-ENERGY SOURCES.

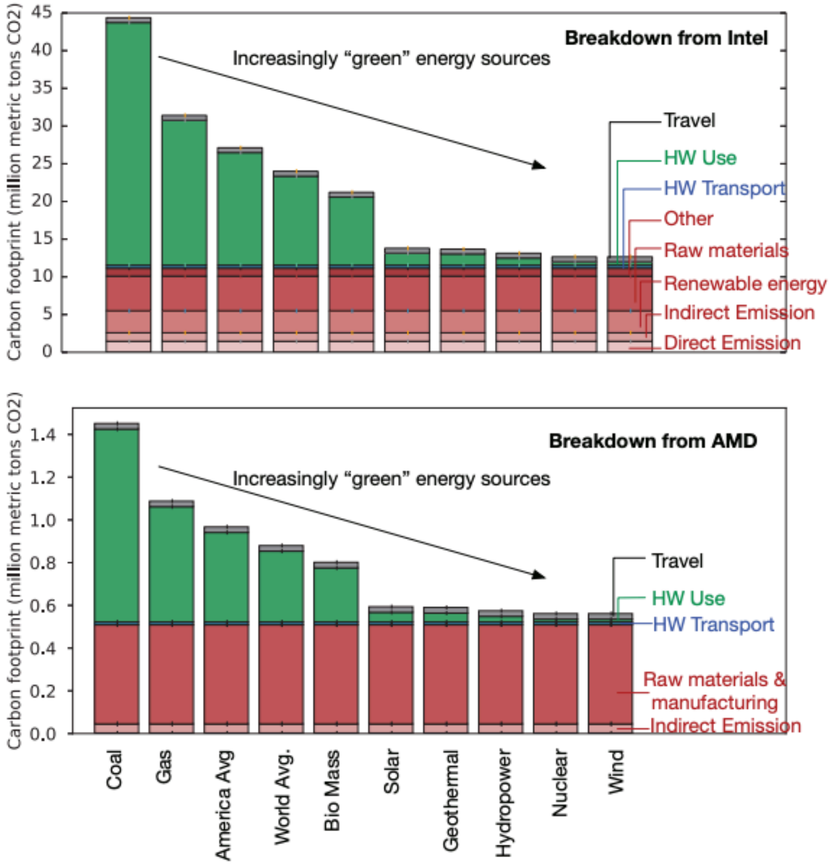


Fig. 13. Reported carbon-footprint breakdown for Intel (top) and AMD (bottom) as renewable energy increasingly (from left to right) powers hardware operation. The use of renewable energy reduces carbon emissions dramatically; most of the remaining emissions are from manufacturing.

Gupta, et al. Chasing Carbon, HPCA 2021

Understanding Embodied Carbon

- The carbon emissions that go into producing a device are a function of the area of that device ➡️ all compute units are derived of integrated circuits and wires to compose our gates, and the number of ICs per device is a function of IC size
- For CPU and GPUs, the amount of carbon consumed is ~0.1-0.4 kg of CO_2 per cm^2; for memory and storage devices this is ~.12-.6 kg of CO_2 per cm^2
- The embodied cost of the device can be *amortized* across the device's lifetime ➡️ if a device lasts a long time, it may make sense to tolerate a higher embodied cost to produce the device

Understanding Embodied Carbon

- Components within a device wear out at different rates ➡️ a device is considered “worn out” when the failures of the components increase beyond the initial tolerable design
- In general, compute units tend to wear out in 3-5 years whereas memory lifetimes are 5-7 years
- Recent literature has proposed reintegrating memory devices from prior devices into newer platforms ➡️ new platforms need to be backwards compatible with the communication protocol

Fig. 2. Moving average (black) of raw (gray) normalized failure rates vs. DDR4 DIMMs’ deployment time in production. Failure rates tend to stay constant over a 7-year period.

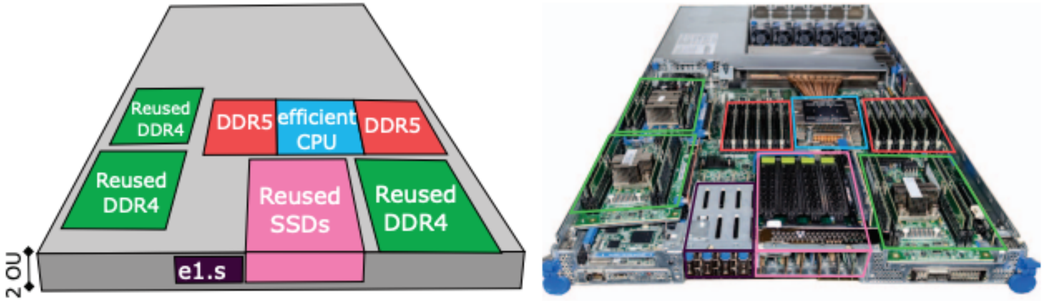
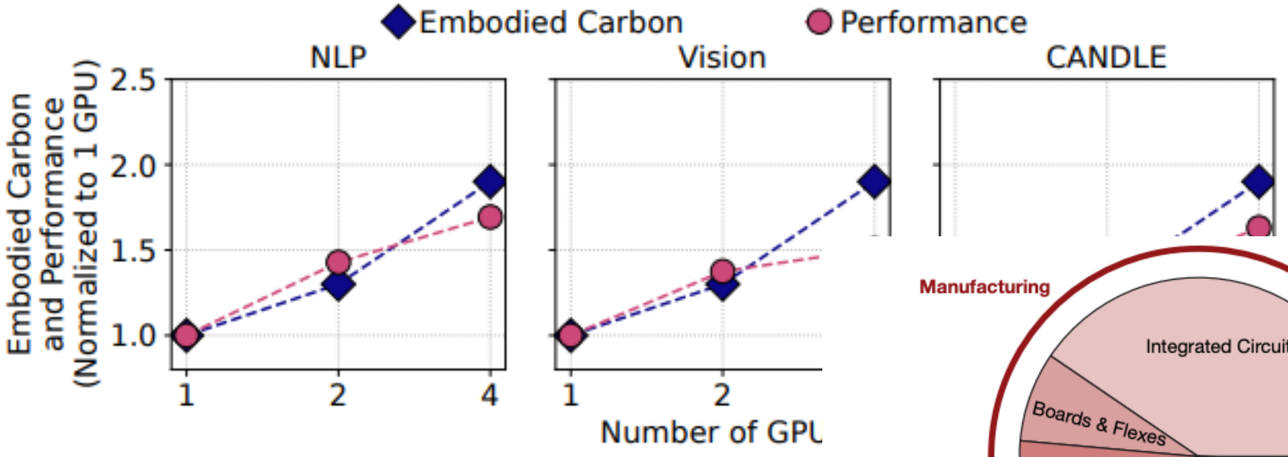


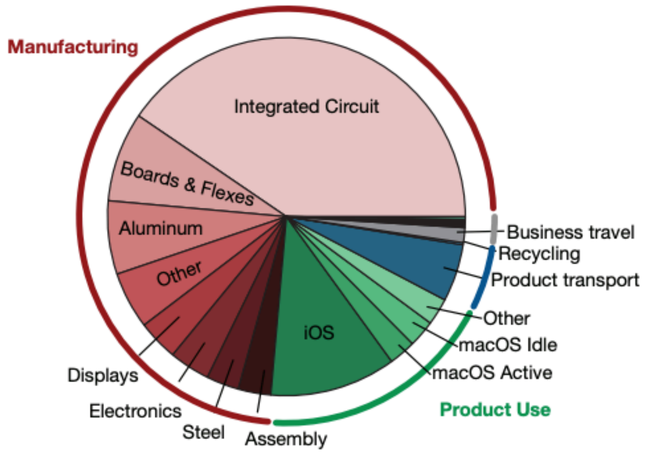
Fig. 5. Our *GreenSKU-Full* design with AMD’s efficient CPU, reused DDR4 DRAM (via CXL), and reused m.2 SSDs (via e1.s and PCIe adapters).

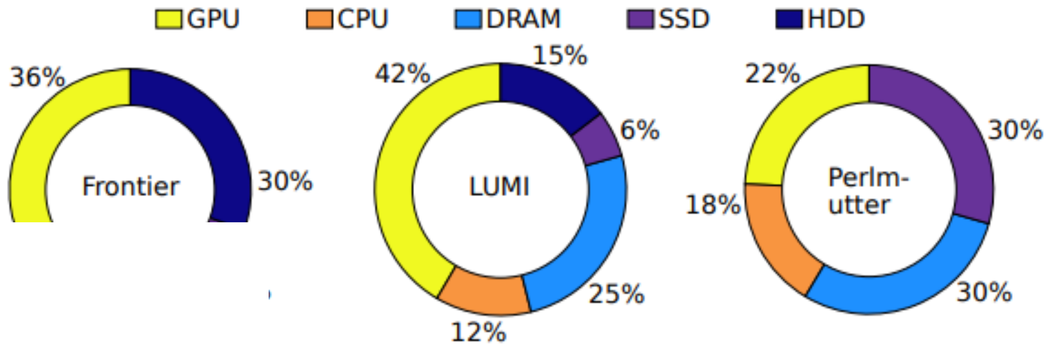
Wang, et al. Designing Cloud Servers for Lower Carbon. ISCA 2024

Holistic Carbon Footprints



Li, et al. Toward Sustainable H





footprint contribution from different com-
eadership supercomputers: Frontier, LUMI,

Fig. 5. Apple's carbon-emission breakdown. In aggregate, the hardware life cycle (i.e., manufacturing, transport, use, and recycling) comprises over 98% of Apple's total emissions. Manufacturing accounts for 74% of total emissions, and hardware use accounts for 19%. Carbon output from manufacturing integrated circuits (i.e., SoCs, DRAM, and NAND flash memory) is higher than that from hardware use.

Further Reading!

- Chasing Carbon: The Elusive Environmental Footprint of Computing
- ACT: Designing Sustainable Computer Systems with an Architectural Carbon Modeling Tools
- Treehouse: A Case for Carbon-Aware Datacenter Software
- Carbon Explorer: A Holistic Framework for Designing Carbon Aware Datacenters
- Toward Sustainable HPC: Carbon Footprint Estimation and Environmental Implications of HPC Systems
- Designing Cloud Servers for Lower Carbon

Chat with your neighbor(s)!

As computer scientists, it's important to remember that our training is not always the best to ask and answer these questions!

What other environmental considerations go into developing an equitable green computing ecosystem?

"Carbon dioxide emissions are a common metric to measure global warming potential" — Climate Change 2007: Synthesis Report, Intergovernmental Panel on Climate Change

Forever chemicals are produced in manufacturing, transport, data center operation, etc…

Water consumption required to cool data centers is extraordinarily high!

When discarding equipment, electronic waste often becomes an issue for the global south to deal with