

Sustainable AI

Mark Zhao

CSCI 5942

Fall 2026

College of Engineering & Applied Science
UNIVERSITY OF COLORADO **BOULDER**

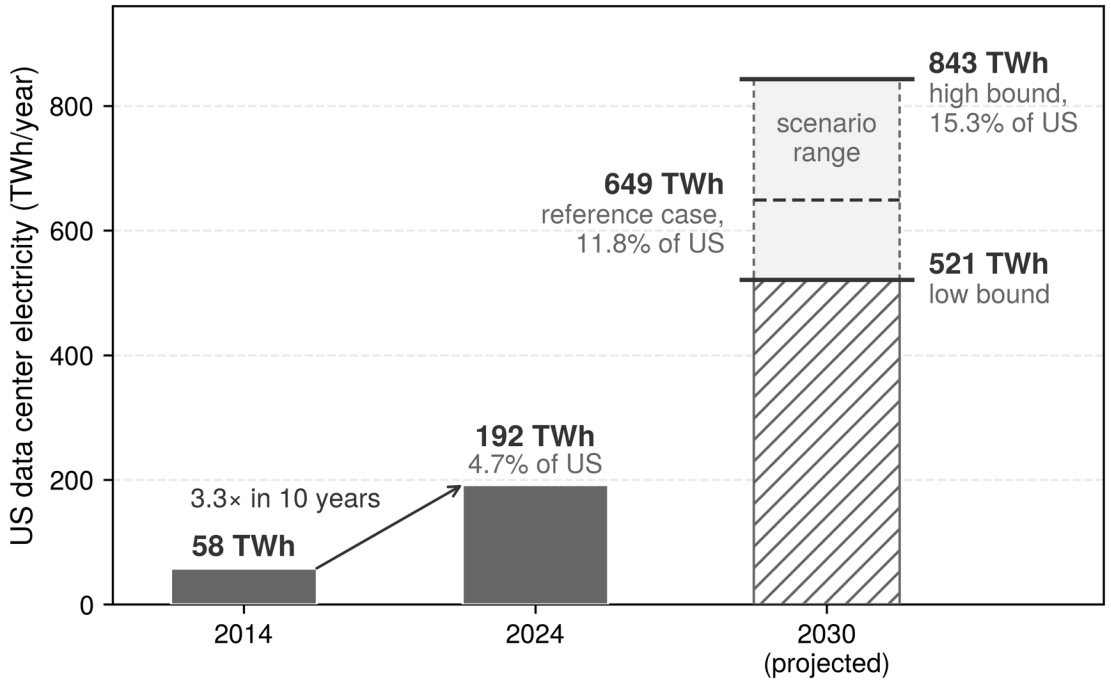
AI growth is a global, geopolitical issue

... Largely restricted by physical electricity grid/generation

- **415 TWh** global datacenter electricity in 2024, about 1.5% of world consumption
- **17%** datacenter electricity growth in 2025 vs 3% globally
- **~945 TWh** base case for 2030, 3x growth in AI-focused datacenters
- **>$400B** capital spending in 2025 by hyperscalers, largely on datacenters

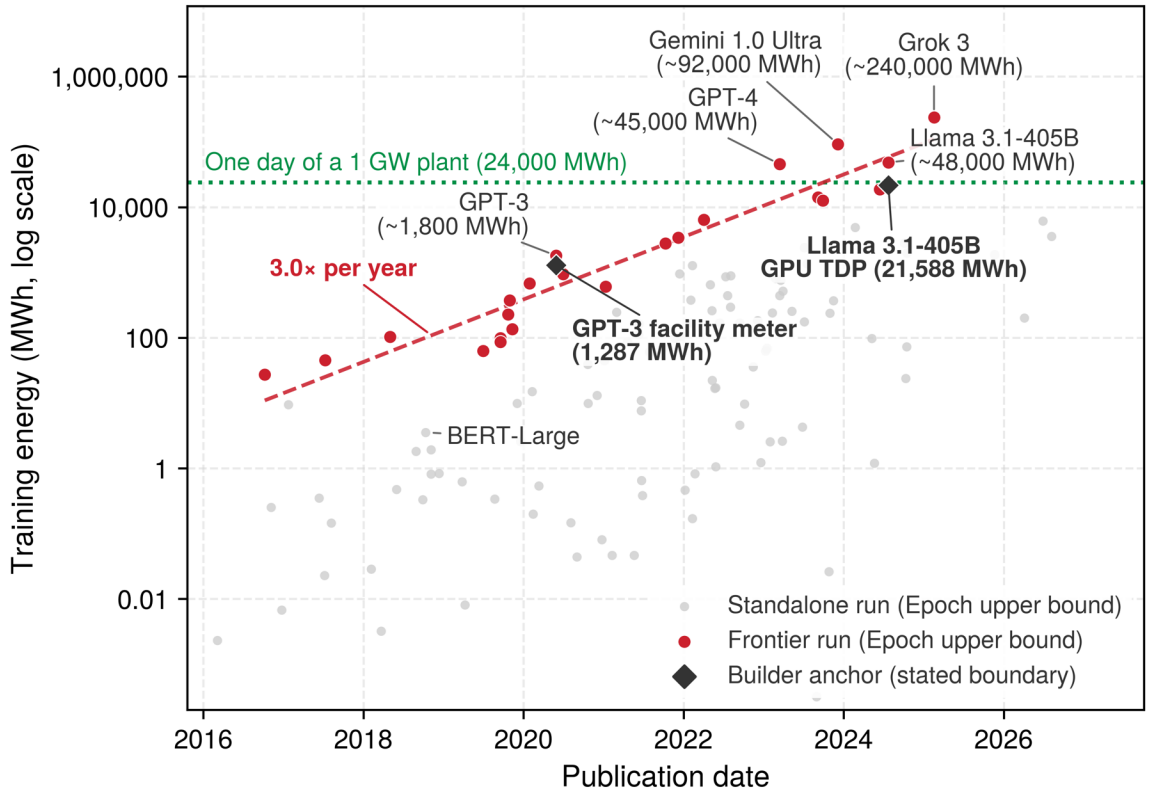
https://www.iea.org/reports/energy-and-ai

US AI Datacenter electricity demand



https://mlsysbook.ai/vol2/15_sustainable_ai/15_sustainable_ai.html#sec-sustainable-ai-data-center-energy-ai-workloads-b1a8

Training Energy of Notable Models

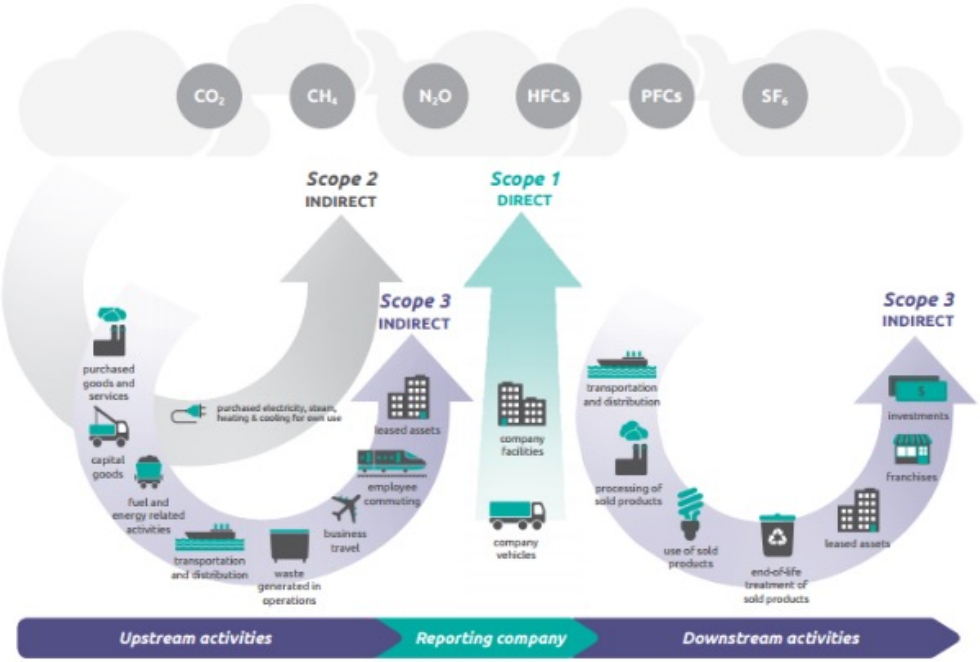


https://epoch.ai/data/ai-models?view=graph&tab=notable

How do we measure AI's environmental footprint?

Greenhouse Gas Protocol (GHG)

- CO_2e (Carbon Dioxide Equivalent)
  - Normalized metric to translate any emission into equivalent amount of CO_2
- Scope 1: Direct GHG emissions
- Scope 2: Indirect emissions
  - GHG of generating energy consumed
- Scope 3: All other activities, including upstream and downstream activities

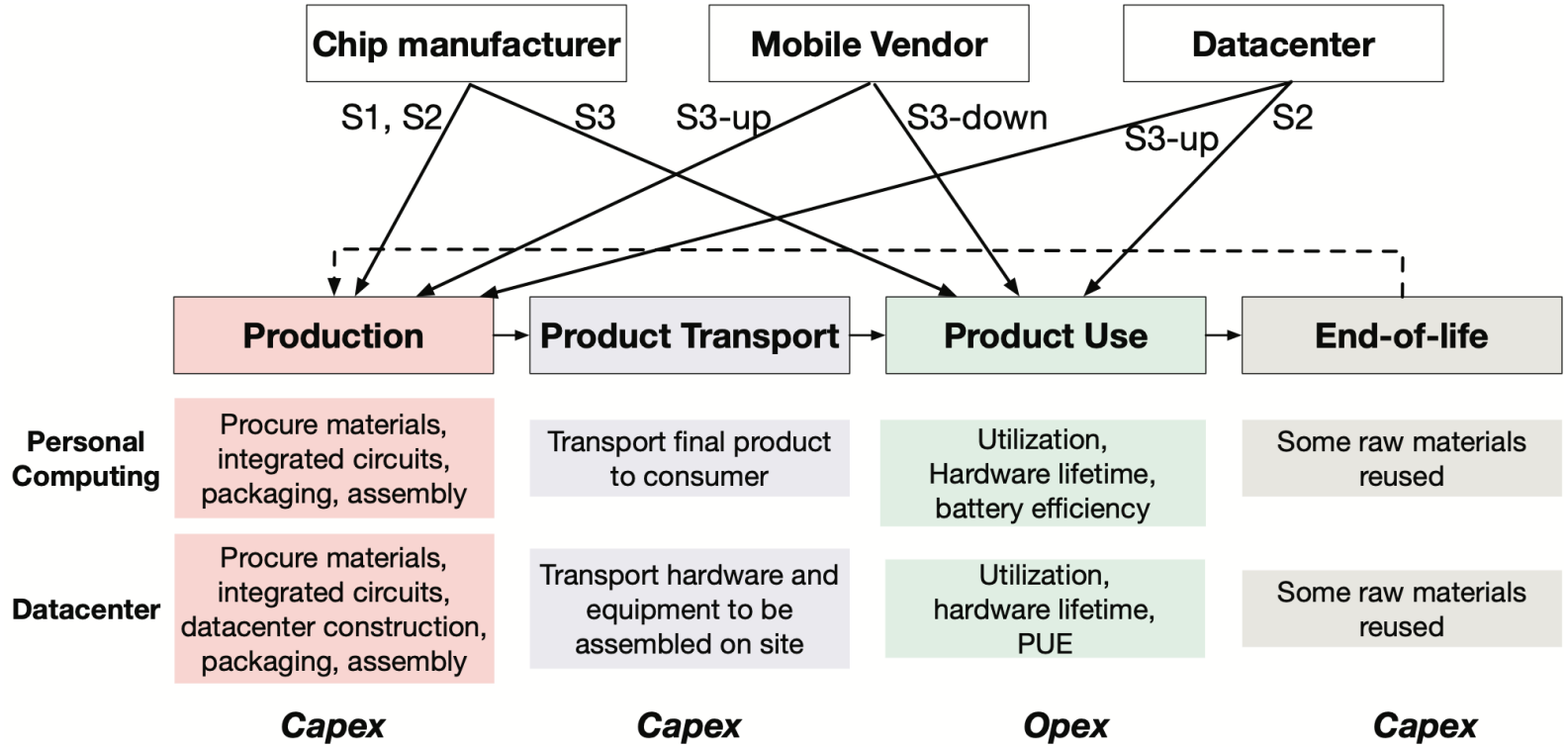


Technology company	Scope 1	Scope 2	Scope 3
Chip manufacturer	Burning PFCs, chemicals, gases	Energy for fabrication	Raw materials, hardware use
Mobile-device vendor	Natural gas, diesel	Energy for offices	Chip manufacturing, hardware use
Data-center operator	Natural gas, diesel	Energy for data centers	Server-hardware manufacturing, construction

https://ghgprotocol.org; https://arxiv.org/abs/2011.02839

How do we measure AI's environmental footprint?

Can also track carbon for individual hardware components through lifecycle analysis:



https://ghgprotocol.org; https://arxiv.org/abs/2011.02839

Carbon footprint

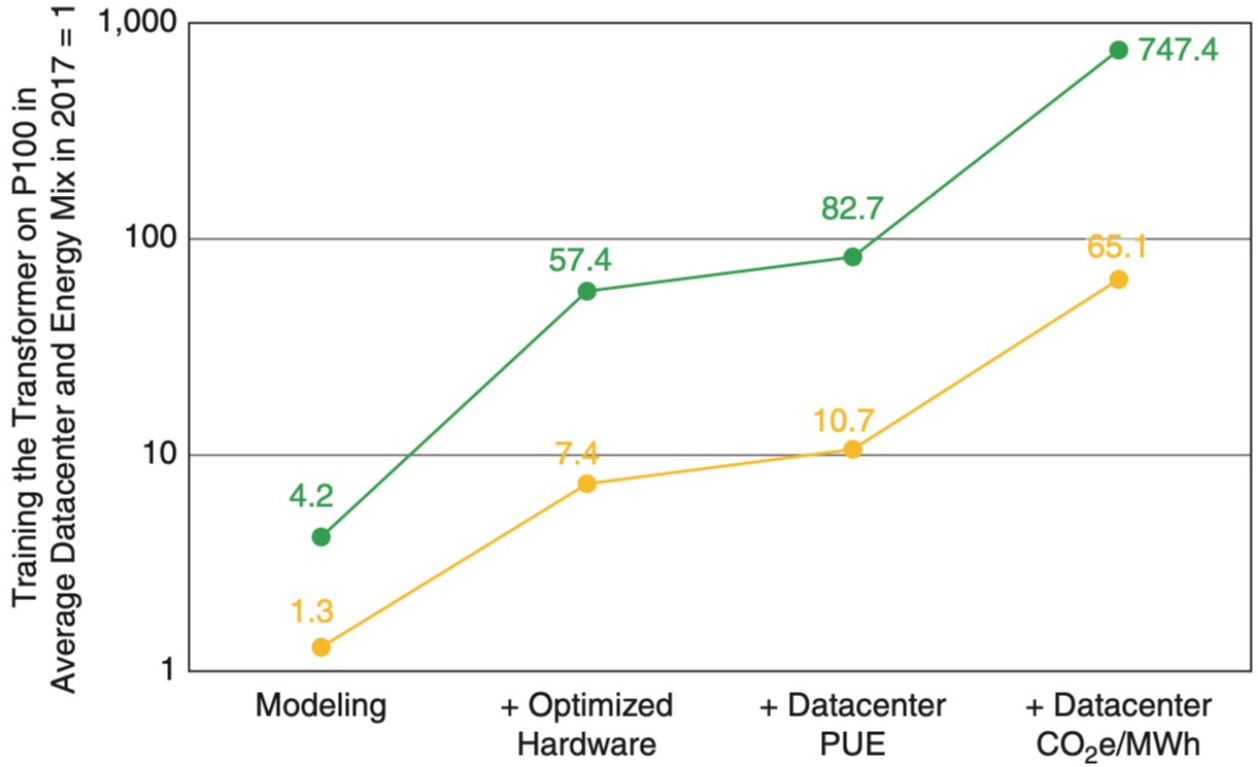
```
hours × N_chips = FLOPs ÷ (3,600 × measured FLOP/s per chip)
kWh = hours × N_chips × P_avg (W) × PUE ÷ 1000
OP_CF (in tCO2e) = kWh × CI (kg CO2e/kWh) ÷ 1000
CF = OP_CF + (T ÷ LT) × E_CF                    T = runtime, LT = hardware lifetime
```

Term	Set by	Typical range / relationships
FLOPs (model, data, search)	ML researcher	Sparse MoE: <1/10 the energy of dense
FLOP/s per chip and P_avg (perf/W)	Architect, compiler, runtime	ML accelerators ~2–5× vs. off-the-shelf
PUE	Data-center operator	1.10 (Google) vs. 1.58 (US average, 2018)
CI, carbon intensity	Location, time, procurement	11 (wind) to 820 (coal) g CO2/kWh
E_CF and LT	Procurement, hardware lifetime	Embodied > operational intensity

How can we bring down AI's operational carbon footprint?

Multiplicative wins from 4 Ms

- **Model**: Improve model quality at fixed compute costs
  - E.g., Neural Architecture Search, sparse MoE, ...
- **Machine**: Improve perf/W
  - Hardware: CPU vs GPU vs TPU
    - E.g., TPU v2 vs P100: 5.6x perf/W
  - Systems: higher MFU
- **Mechanization:**
  - Lower PUE, more efficient datacenters
- **Map:**
  - Lower datacenter CO2e/MWh



David Patterson et al.; The Carbon Footprint of Machine Learning Training Will Plateau, Then Shrink; IEEE Computer

Machine: Improving energy/op

- Specialized accelerators generally require less energy/FLOP

- Most energy due to data movement, not compute units

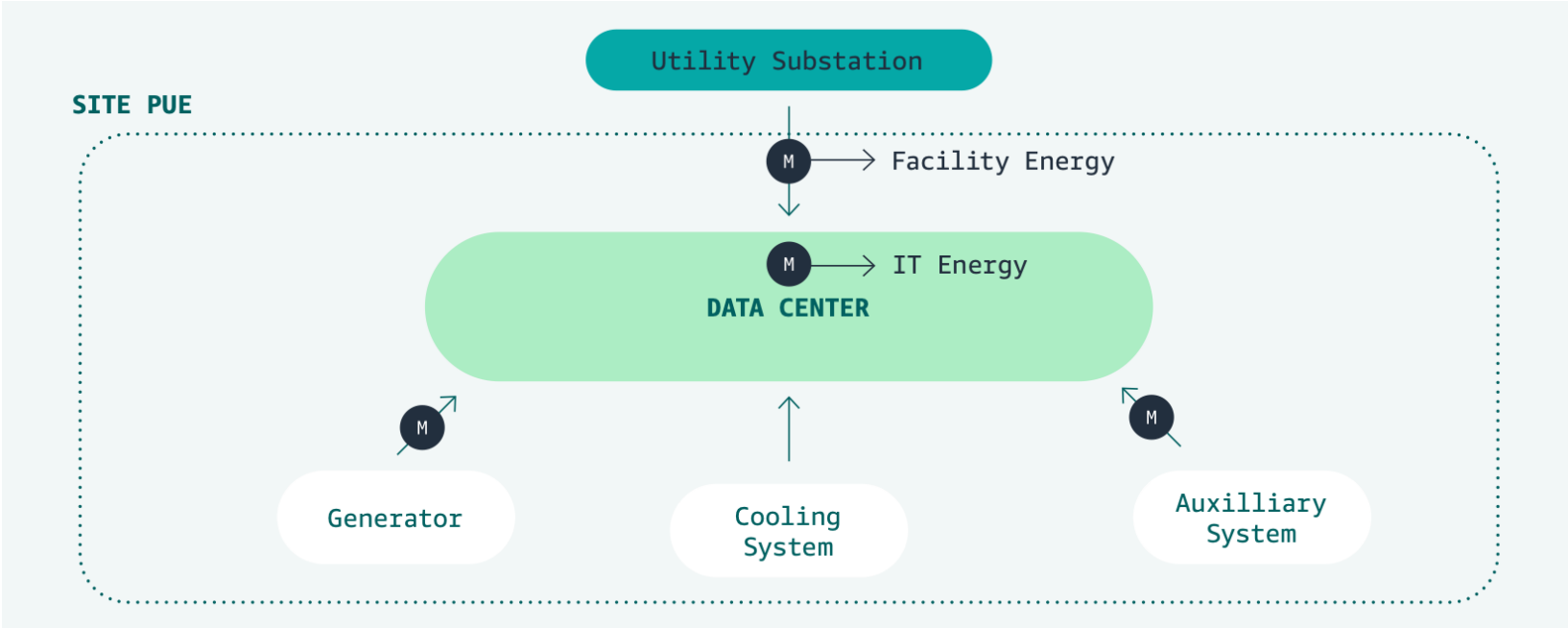
Architecture	Effective energy (model)	Potential energy advantage	Comparison must include
General-purpose CPU	100 pJ/FLOP	Flexible execution for diverse workloads	Instruction overhead and achieved utilization
GPU	10 pJ/FLOP	Parallel arithmetic and data reuse	Precision, occupancy, and memory traffic
Systolic accelerator	1 pJ/FLOP	Local reuse across a regular datapath	Supported shapes, precision, and data movement
Edge accelerator or fixed-function ASIC	0.1 pJ/FLOP	Specialized low-power execution	Host, memory, peripherals, and duty cycle
Analog or in-memory compute	Not modeled	Reduced movement for supported operations	Conversion, peripheral circuits, and achieved quality

Memory level	Energy cost (pJ/byte)
Register	0.1 pJ/byte
L1 Cache	1 pJ/byte
L2 Cache	5 pJ/byte
DRAM	160 pJ/byte
NVMe SSD	1000 pJ/byte
Network	> 10000 pJ/byte

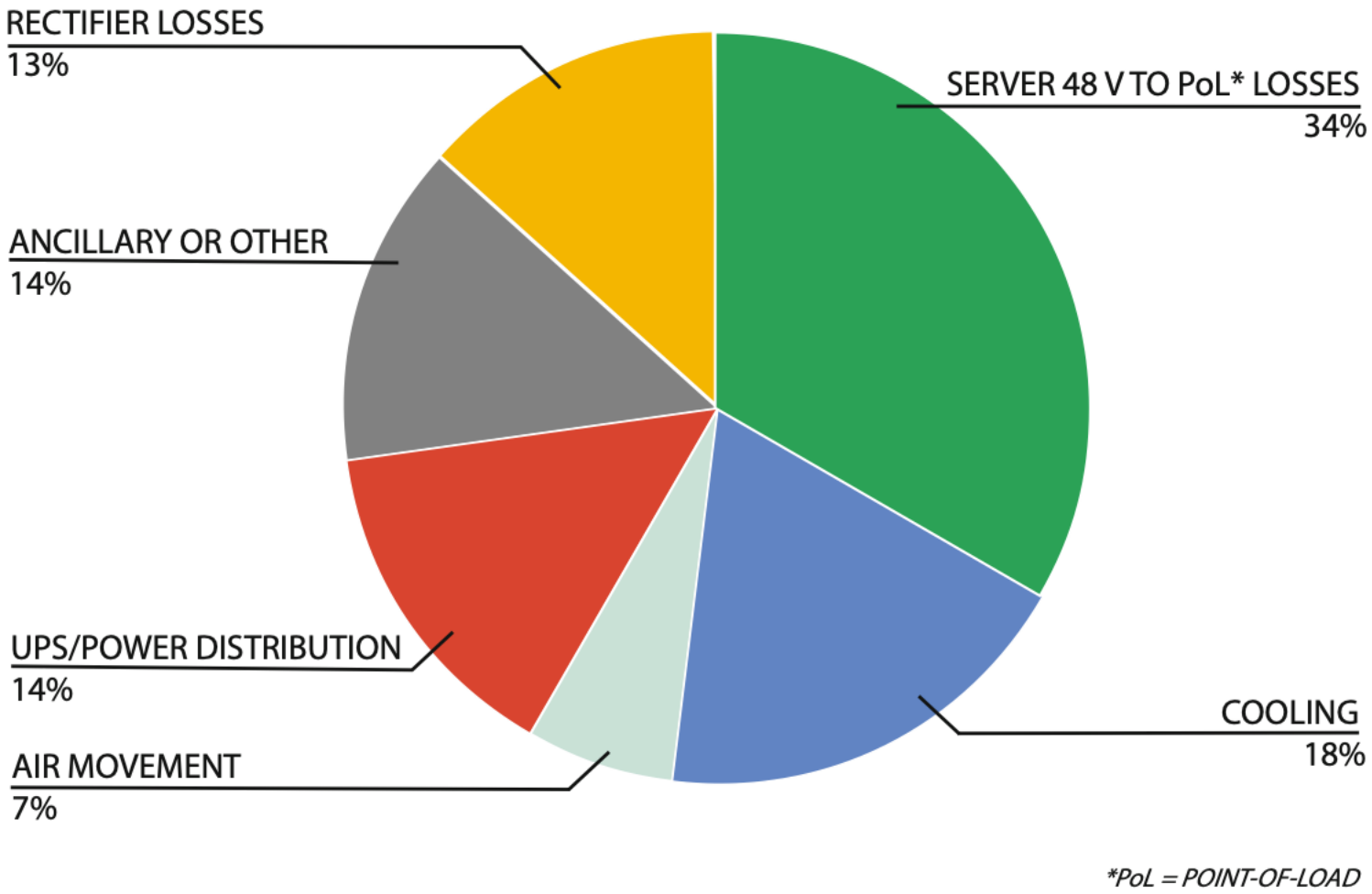
https://mlsysbook.ai/vol2/15_sustainable_ai/15_sustainable_ai.html#sec-sustainable-ai-energy-per-byte-5d87-5d87

Power/Water Usage Effectiveness

- $\text{PUE} = \dfrac{\text{Total Facility Energy}}{\text{Useful (IT) Equipment Energy}}$
- $\text{PUE} = \dfrac{\text{Total Facility Water Usage (L)}}{\text{Useful (IT) Equipment Energy (kWh)}}$



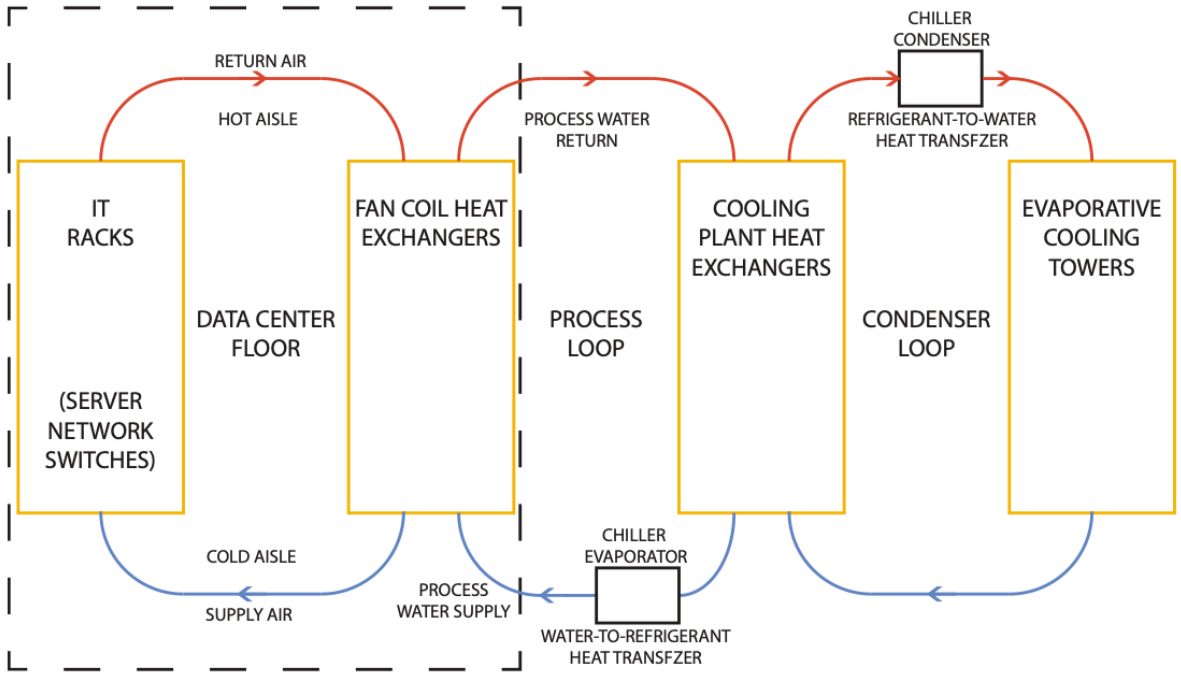
Energy losses in datacenters



L. Barroso et al.; The Datacenter as a Computer: Designing Warehouse-Scale Machines

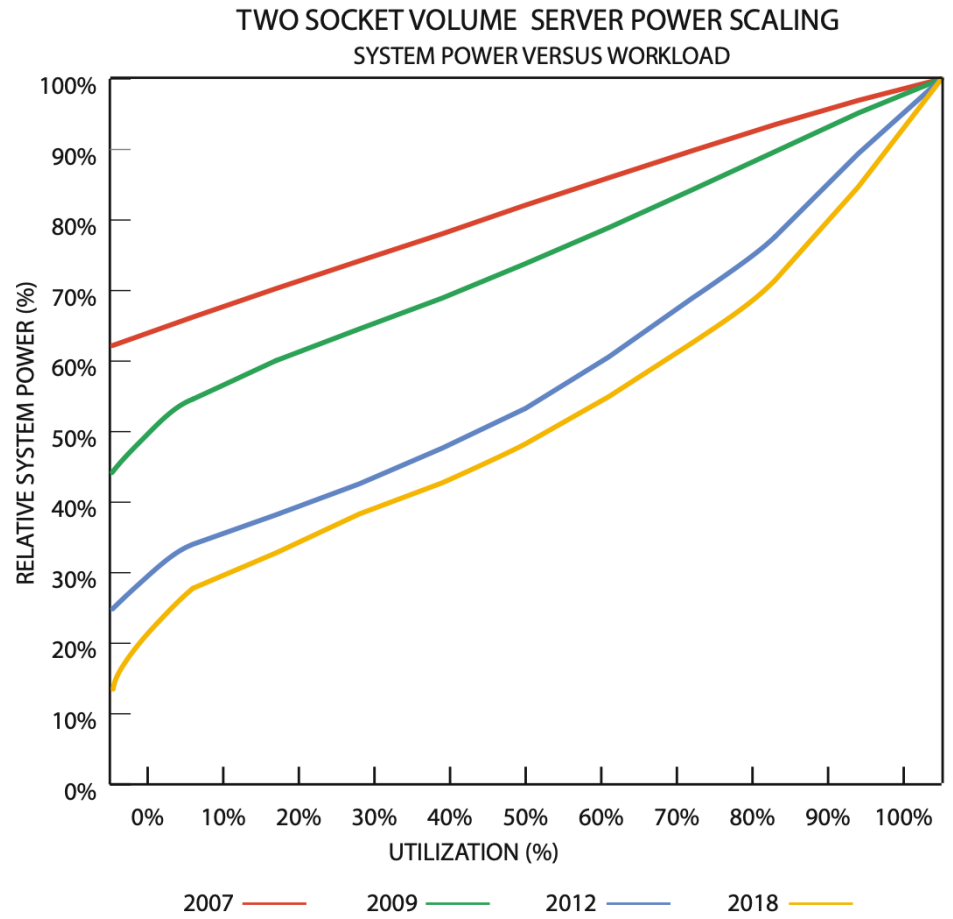
Improving PUE

- **Airflow optimization**: Isolate hot air exhausted from servers from cold air
- **Temperature optimization:** Keep cold aisle ~25-30C
- **Free cooling**: Build in moderate climates
- **Better power delivery** architectures
- **ML**: Apply ML to optimize DC based on sensor readings



Energy Proportionality

- Ideally, power consumption should be directly linear with activity (e.g., 0% power at 0% load, 100% power at 100% load).
  - In reality, not so much…
- How can we better use server power?
  - Higher utilization
  - DVFS, power modes
  - E-cores / P-cores



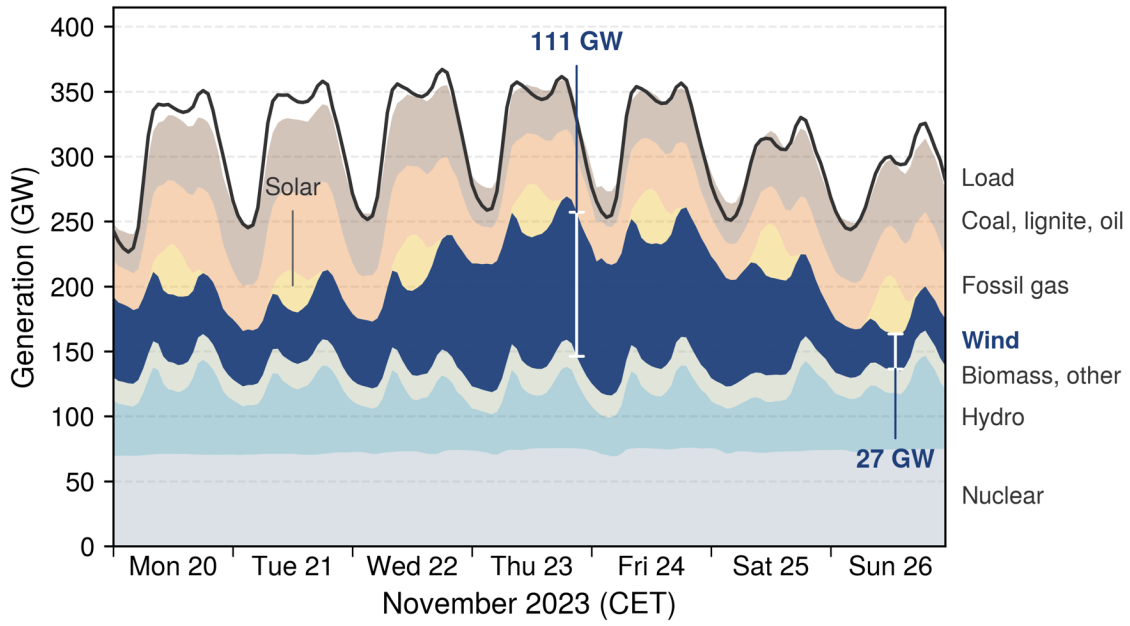
Dynamic electricity carbon intensity

Source (life cycle)	g CO2/kWh
Coal	820
Natural gas	490
Biomass	230
Solar	41
Geothermal	38
Hydropower	24
Nuclear	12
Wind	11

Grid average	g CO2/kWh
India	725
Australia	597
Taiwan	583
Singapore	495
United States	380
Europe	295
Brazil	82
Iceland	28

Dynamic electricity carbon intensity

- https://app.electricitymaps.com/map/fifteen_minutes?signal=carbon-intensity

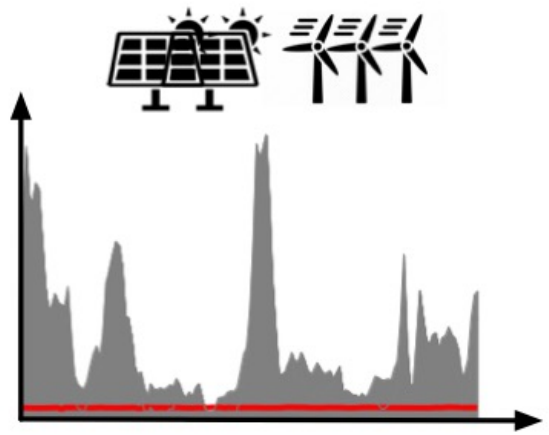


Carbon data across Google Cloud regions

Google Cloud Region	Location	Google CFE%	Grid carbon intensity (gCO_2eq/kWh)	
africa-south1	Johannesburg	15%	641	
asia-east1	Taiwan	15%	428	
asia-east2	Hong Kong	2%	489	
asia-northeast1	Tokyo	23%	449	
asia-northeast2	Osaka	45%	299	
asia-northeast3	Seoul	37%	359	
asia-south1	Mumbai	20%	598	
asia-south2	Delhi	39%	473	
asia-southeast1	Singapore	5%	366	
asia-southeast2	Jakarta	15%	588	
australia-southeast1	Sydney	37%	468	
australia-southeast2	Melbourne	44%	416	
europe-central2	Warsaw	81%	499	Low CO_2
europe-north1	Finland	98%	30	Low CO_2
europe-north2	Stockholm	100%	19	Low CO_2
europe-southwest1	Madrid	87%	96	Low CO_2
europe-west1	Belgium	79%	126	Low CO_2
europe-west2	London	75%	109	Low CO_2
europe-west3	Frankfurt	70%	276	
europe-west4	Eemshaven	82%	194	Low CO_2
europe-west6	Zürich	97%	21	Low CO_2
europe-west8	Milan	79%	202	Low CO_2
europe-west9	Paris	96%	16	Low CO_2
europe-west10	Berlin	70%	276	
europe-west12	Turin	79%	202	Low CO_2

How can we exploit electricity CI?

1. Renewable Deployment

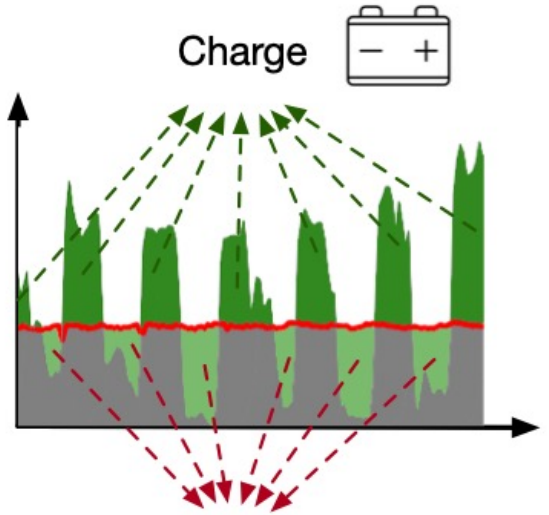


+ Renewable coverage increased

- Additional manufacturing carbon cost for renewables

- Excess renewable supply increased

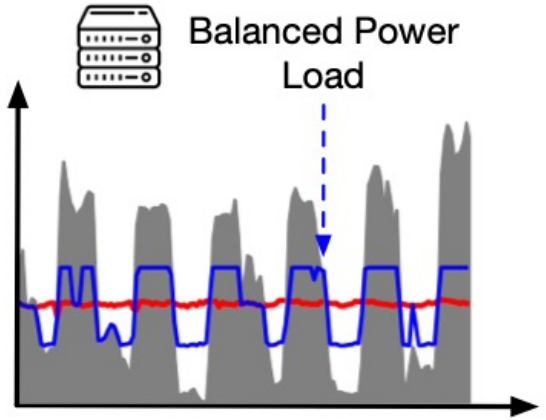
2. Battery Deployment



+ Renewable coverage increased

- Manufacturing carbon cost for batteries

3. Carbon Aware Scheduling



+ Renewable coverage increased

- Peak capacity increased

- Manufacturing carbon cost for additional servers

B. Acun et al.; Carbon Explorer: A Holistic Framework for Designing Carbon Aware Datacenters; ASPLOS'23

Carbon-Aware Placement

- Track carbon intensity of eac datacenter throughout the day, move/route jobs to carbon-efficient datacenters

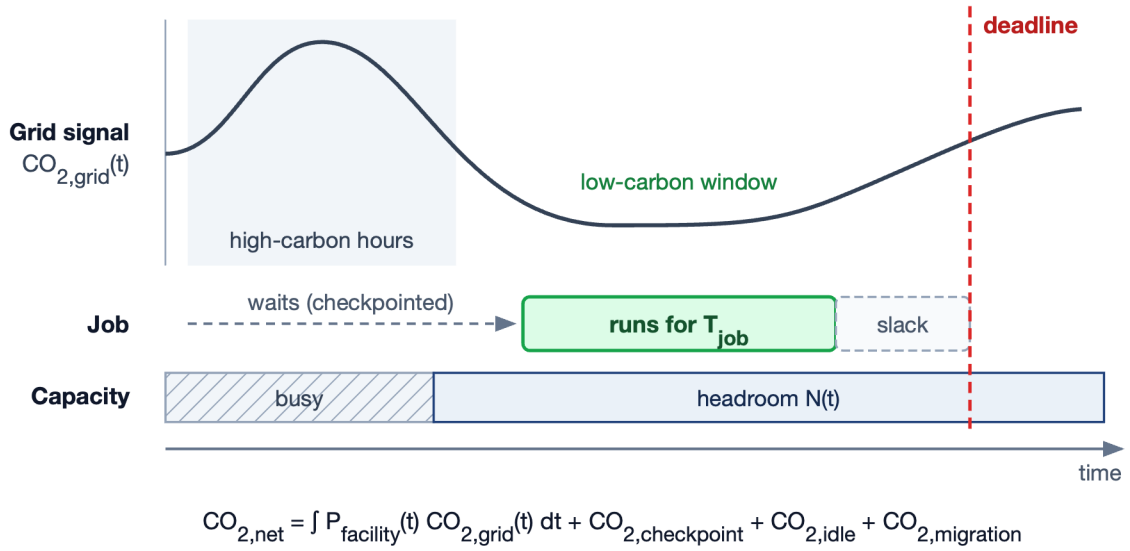
- What are the costs?



https://blog.google/company-news/outreach-and-initiatives/sustainability/carbon-aware-computing-location/

Carbon-Aware Scheduling

- Idea: Shift job in time, not space. Checkpoint job until there is (a) low-carbon window, (b) cluster headroom.
- What can move? What can't?



https://mlsysbook.ai/vol2/15_sustainable_ai/15_sustainable_ai.html#sec-sustainable-ai-time-shifting-deadlines

Embodied carbon emissions

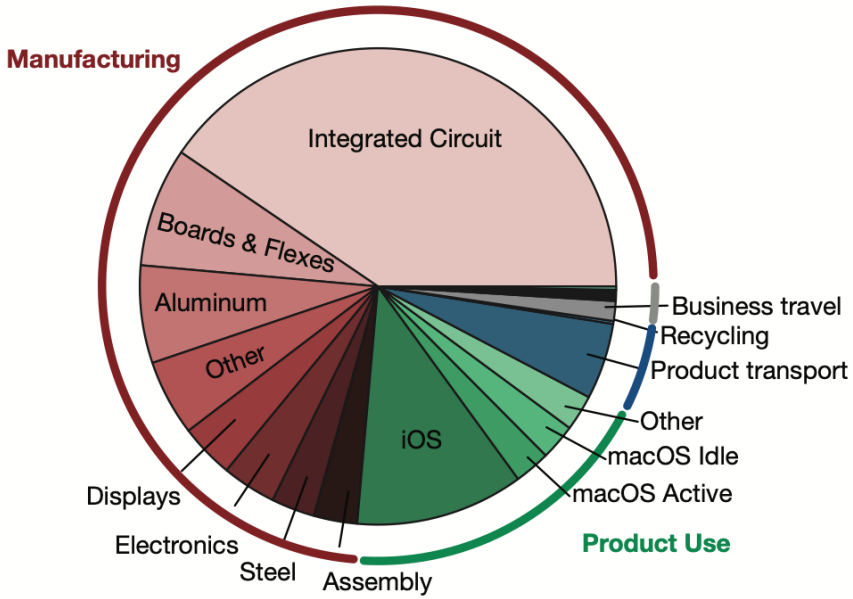
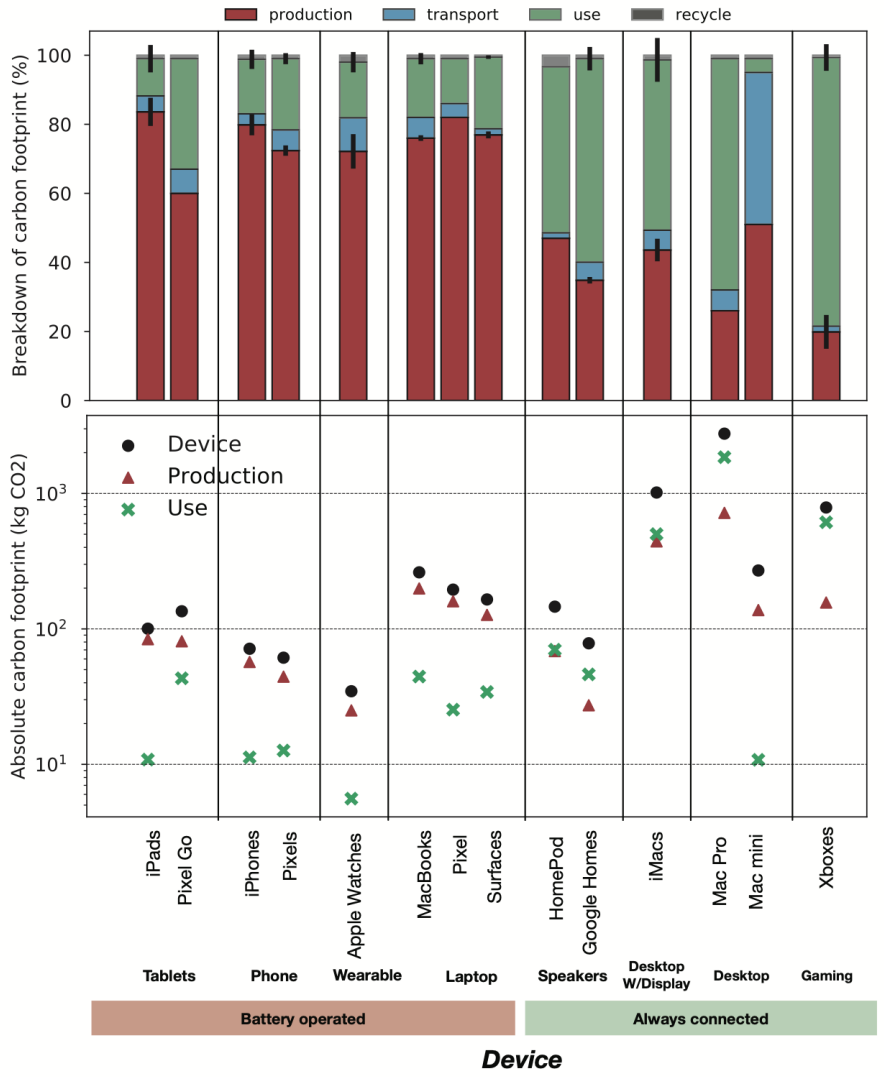


Fig. 5. Apple's carbon-emission breakdown. In aggregate, the hardware life cycle (i.e., manufacturing, transport, use, and recycling) comprises over 98% of Apple's total emissions. Manufacturing accounts for 74% of total emissions, and hardware use accounts for 19%. Carbon output from manufacturing integrated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



Udit Gupta, et al.; Chasing Carbon: The Elusive Environmental Footprint of Computing; HPCA'21

Embodied carbon emissions

- As we leverage lower-carbon intensity energy sources, carbon footprint shift from operational to embodied emissions

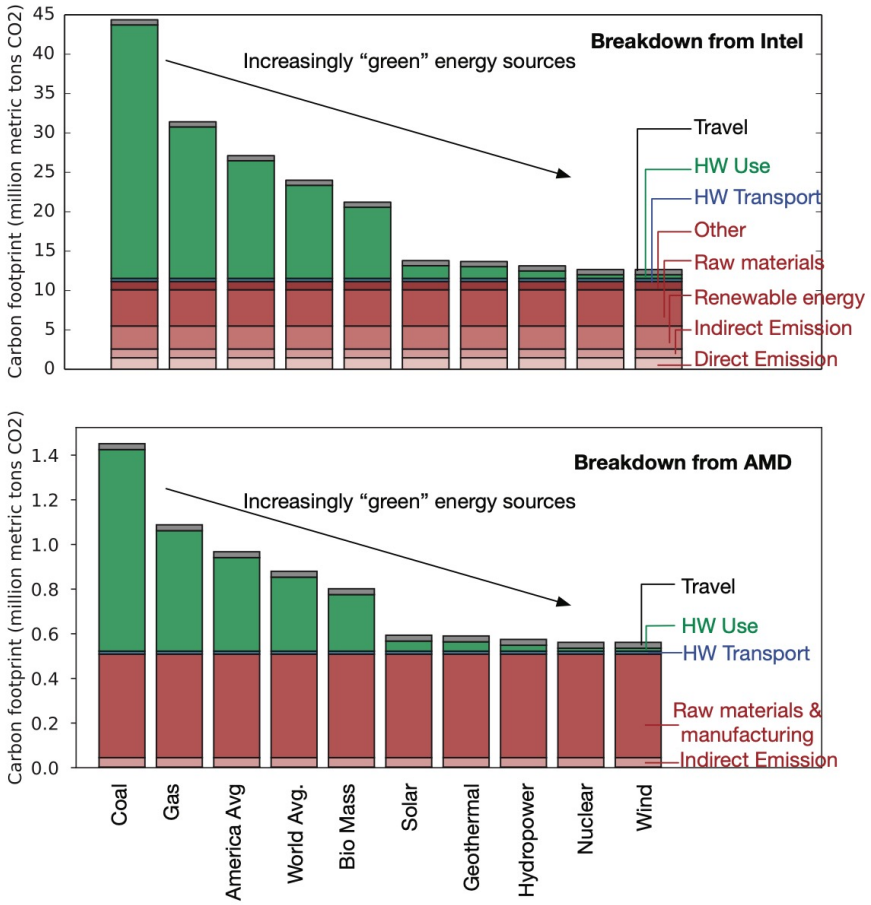
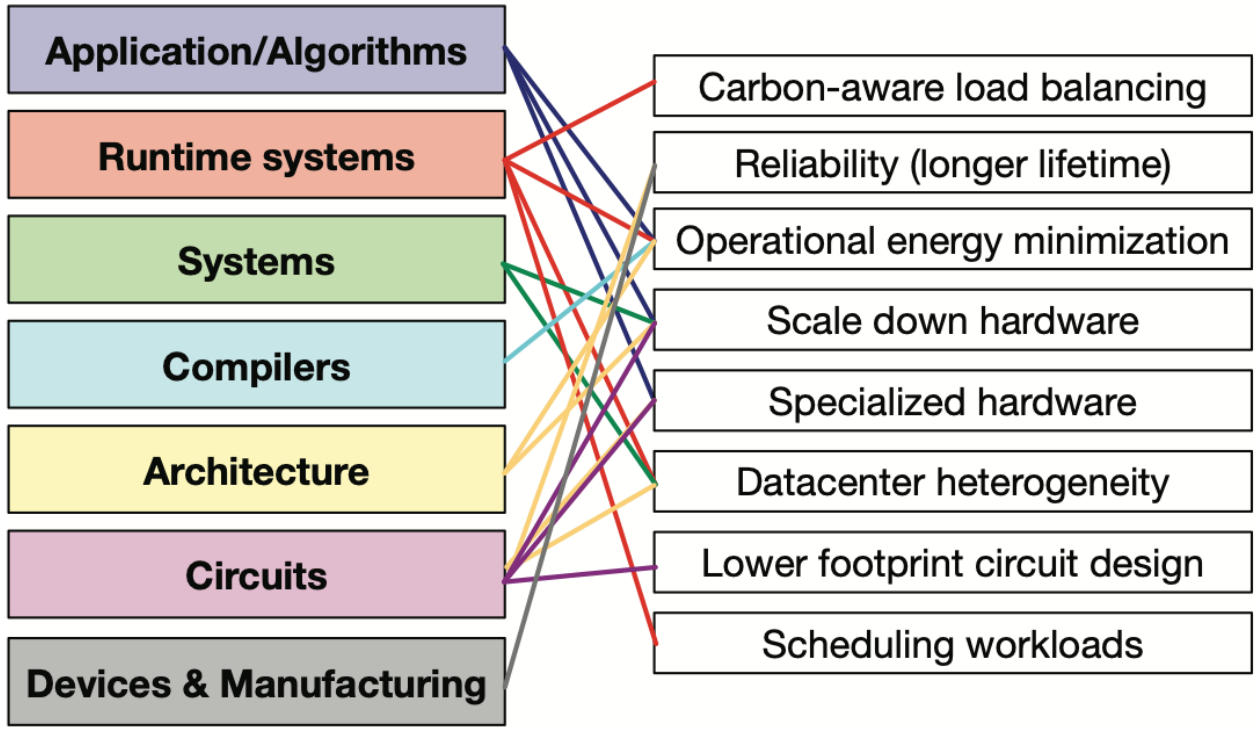


Fig. 13. Reported carbon-footprint breakdown for Intel (top) and AMD (bottom) as renewable energy increasingly (from left to right) powers hardware operation. The use of renewable energy reduces carbon emissions dramatically; most of the remaining emissions are from manufacturing.

Udit Gupta, et al.; Chasing Carbon: The Elusive Environmental Footprint of Computing; HPCA'21

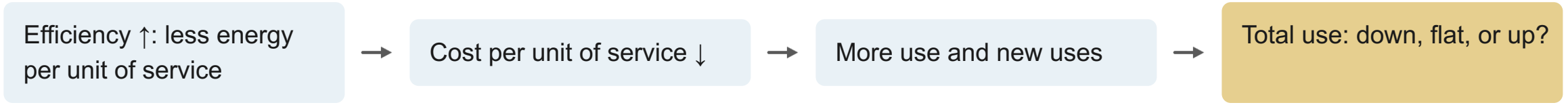
How can we reduce embodied carbon?

- Reduce hardware requirements
  - Train on “scale-down” resources
  - Better scheduling to balance performance for latency/batch workloads
- Use hardware for longer
  - Keep GPUs, enable applications to use heterogeneous clusters
  - Build more reliable, fault-tolerant hardware
- Manufacturing processes
  - Research into sustainable manufacturing processes, better yields, renewable energy for fabs



Jevon’s Paradox

“It is wholly a confusion of ideas to suppose that the economical use of fuel is equivalent to a diminished consumption. The very contrary is the truth.” – W.S. Jevons, The Coal Question, 1865



What can the end result be?

- Rebound: Extra use eats part of the expected savings.
- Backfire: Extra use exceeds the savings, and total consumption rises

Jevon’s Paradox and AI



Where are we now?

Efficiency is up, but total use has skyrocketed

- Google
  - 33x *drop* in energy per prompt, but 50x *growth* in tokens from 2024-2025
- Meta
  - -28.5% operational power saved by optimization from 2019-2021
  - 2.9x growth in AI training capacity growth in 18 months
- IEA
  - Per-task energy falling by ~order of magnitude per year
  - But datacenter energy growth expected to 2x by 2030

Where are we now?

- Backfire?
  - As tokens got cheaper, products spent more per task: reasoning traces, test-time scaling, agents with hundreds of calls
  - The IEA attributes demand growth to more users and more energy-intensive modes such as agents, while plain text generation becomes cheaper.
- Reasons for doubt
  - 2010-2018: datacenter energy rose ~6% while compute capacity rose ~550% -> efficiency outran demand
  - Multiplicative gains from 4 M's: footprint will plateau if best practices spread
- Takeaway: Systems efficiency is necessary, but not sufficient. Need to consider wide breadth of factors to address AI sustainability
  - Operational: Model, Machine, Mechanization, Map
  - Embodied: better accounting and reporting, more/longer effective use of hardware