Sustainable Computer System Design

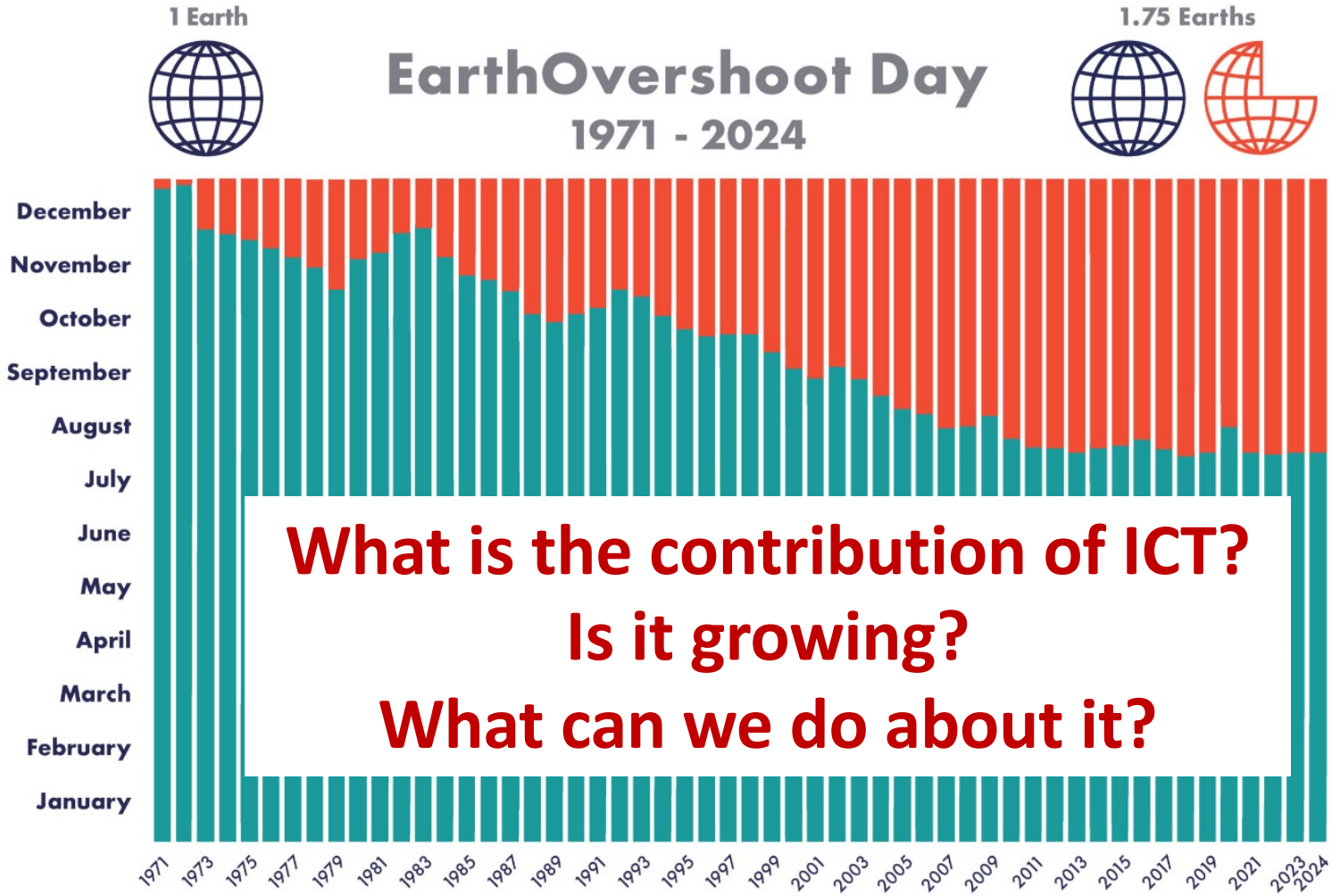
Lieven Eeckhout
Ghent University, Belgium

National University of Singapore (NUS) Computer Science Research Week
Jan 8 – 10, 2025

“*Sustainable development is development that meets the needs of the present without compromising the ability of future generations to meet their own needs.*”

[The Brundtland Report of the World Council on Economic Development, 1987]

How Are We Doing?



[Global Footprint Network]

Agenda: Key Questions

Part I	***What is the environmental impact of computing? What are the challenges?***
Part II	***How does the environmental impact of ICT scale?*** ***What are the contributing factors?***
Part III	***How to reason about sustainable computer system design in light of inherent data uncertainty?***
Part III	***How to design (less un)sustainable microprocessor chips?***

Challenges when Doing Research in Computer System Sustainability

1. ***Sustainability is multi-faceted problem***
   - ***Global warming, raw materials, e-waste, water consumption, biodiversity, etc.***

2. ***Inherent data uncertainty***
   - ***Many unknowns, data limitations, industry secrecy***

3. ***Need to account for entire lifecycle***
   - ***Embodied footprint: raw material extraction, manufacturing, end-of-life recycling***
   - ***Operational footprint: usage of device during lifetime***

GHG Emissions Lead to Global Warming

bush fire

drought

biodiversity loss

flooding

hurricanes

Contribution of ICT to global greenhouse gas (GHG) emissions estimated to be around 2.1–3.9%, and it is rising

…on par with aviation industry…

[Freitag et al., 2020]

U.N. Report Says World Could See 3.1°C Warming by 2100 Without Urgent Action on Climate

HEADLINE OCT 25, 2024

Sustainability is a Multi-Faceted Challenge

Sustainability is much more than combating global warming

It is also about

- **Raw material extraction**

  World Bank projects that demand for metals and minerals will increase rapidly with climate ambition

  - Electric storage batteries: 10x more metals (aluminum, cobalt, iron, lead, lithium, manganese and nickel) needed by 2050 under a 2℃ scenario

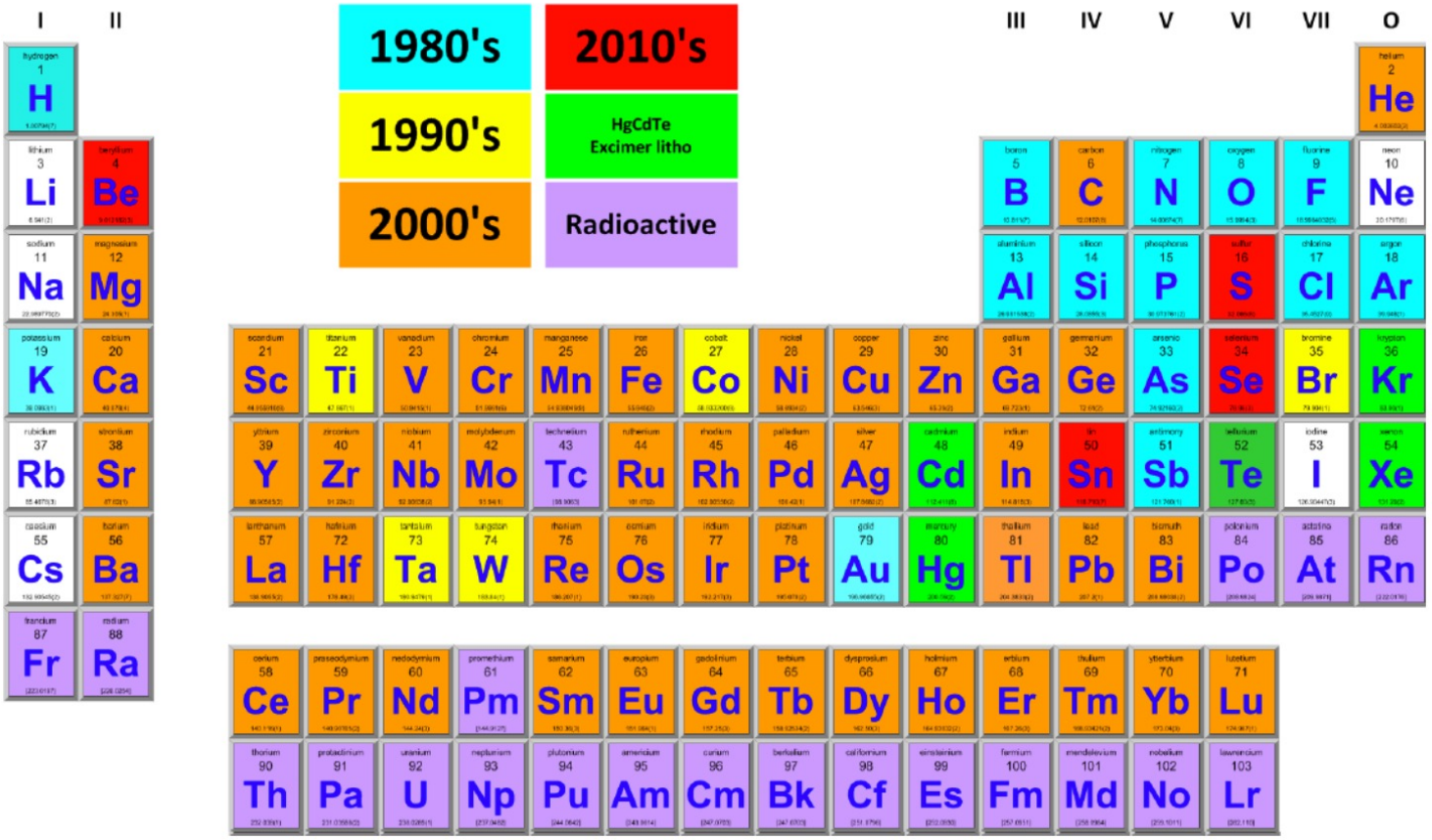
  Under EU's climate-neutrality scenarios for 2050, the EU needs

  - 18x more lithium in 2030, and almost 60x more in 2050
  - 5x more cobalt in 2030, and almost 15x more in 2050
  - 10x more **Rare Earth Elements (REEs)** in 2050
    - REEs for permanent magnets: Dysprosium, Neodymium, Praseodymium, Samarium; The remaining rare earths are Yttrium, Lanthanum, Cerium, Promethium, Europium, Gadolinium, Terbium, Holmium, Erbium, Thulium, Ytterbium, Lutetium

[World Bank (2017): The Growing Role of Minerals and Metals for a Low Carbon Future]

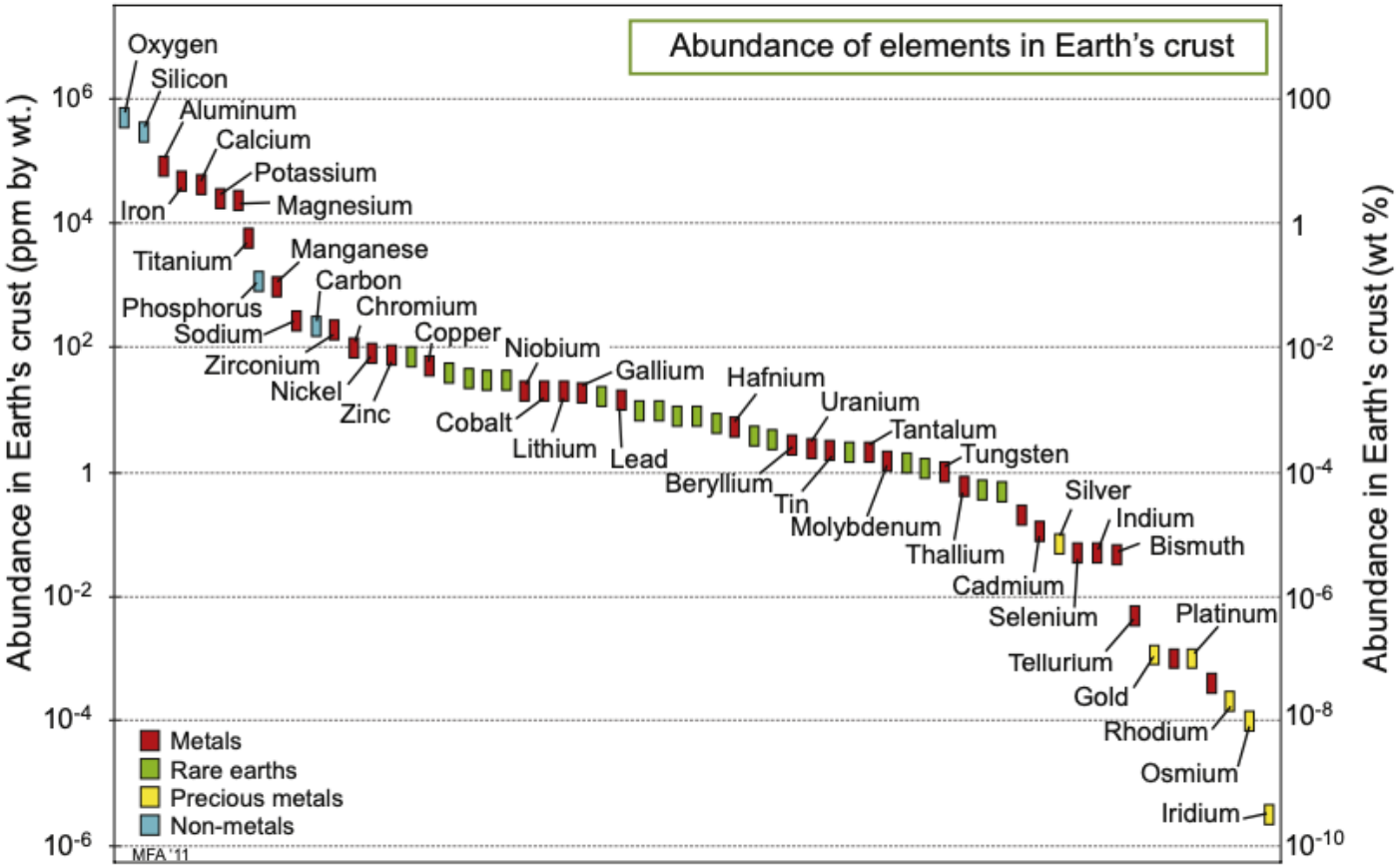
[European Commission 2020: Critical Raw Materials Resilience: Charting a Path towards greater Security and Sustainability]

What Materials Are Needed to Produce Microelectronic Devices?



[Ernst et al., HiPEAC Vision 2024]

Some Materials are Rare



[M. F. Ashby, Materials and Sustainable Development, 2016]

How much of everything have we got?

Enormous range: some elements are abundant, others are rare

Mining rare elements can become extremely expensive and challenging

Raw Material Mining

- Energy/carbon-intensive industry
- Has significant impact on the environment

For example: copper (Cu)

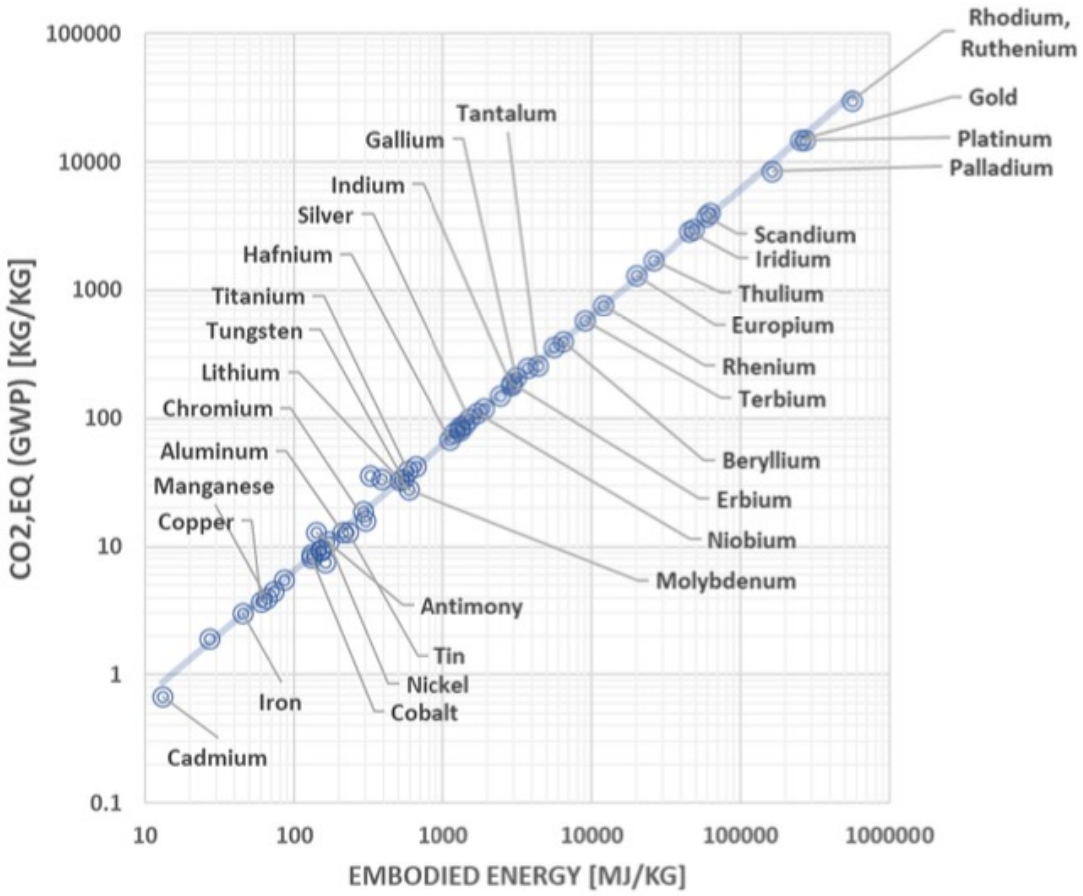
~50 MJ energy for 1 kg of Cu

~4 kg of CO2 for 1 kg of Cu

For example: gold (Au)

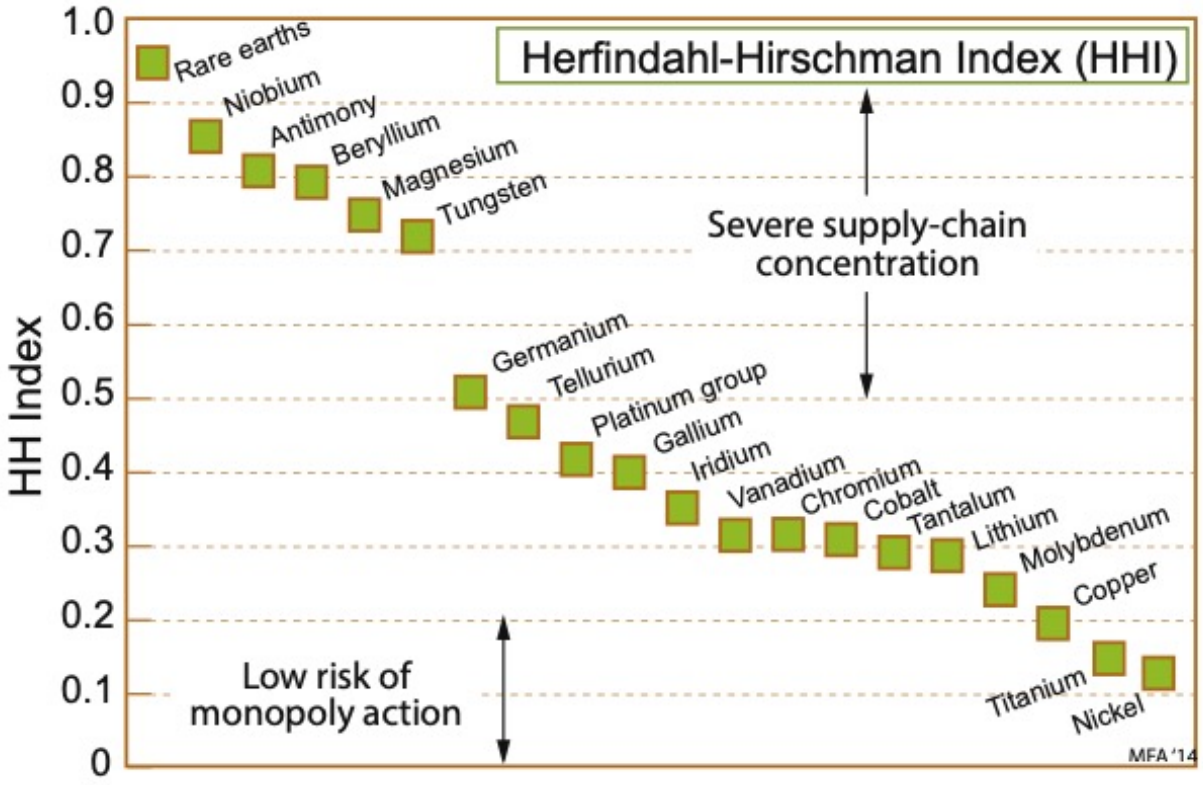
~200 BJ energy for 1kg of Au

~15 tons of CO2 for 1kg of Au



[M. Ashby (2016): Materials and Sustainable Development]

Supply Chain Risk



Herfindahl-Hirschman Index (HHI):

$$HHI = \sum_{i=1}^{n} f_i^2$$

f_i is fraction of market sourced by nation *i*, and *n* is total number of source-nations.

One nation is monopoly: HHI = 1

Two nations with equal share: HHI = $0.5^2 + 0.5^2 = 0.5$

Many source-nations: HHI → 0

Esp. problematic if HHI is high and materials come from politically unstable region(s)

[M. F. Ashby, Materials and Sustainable Development, 2016]

Sustainability is a Multi-Faceted Challenge

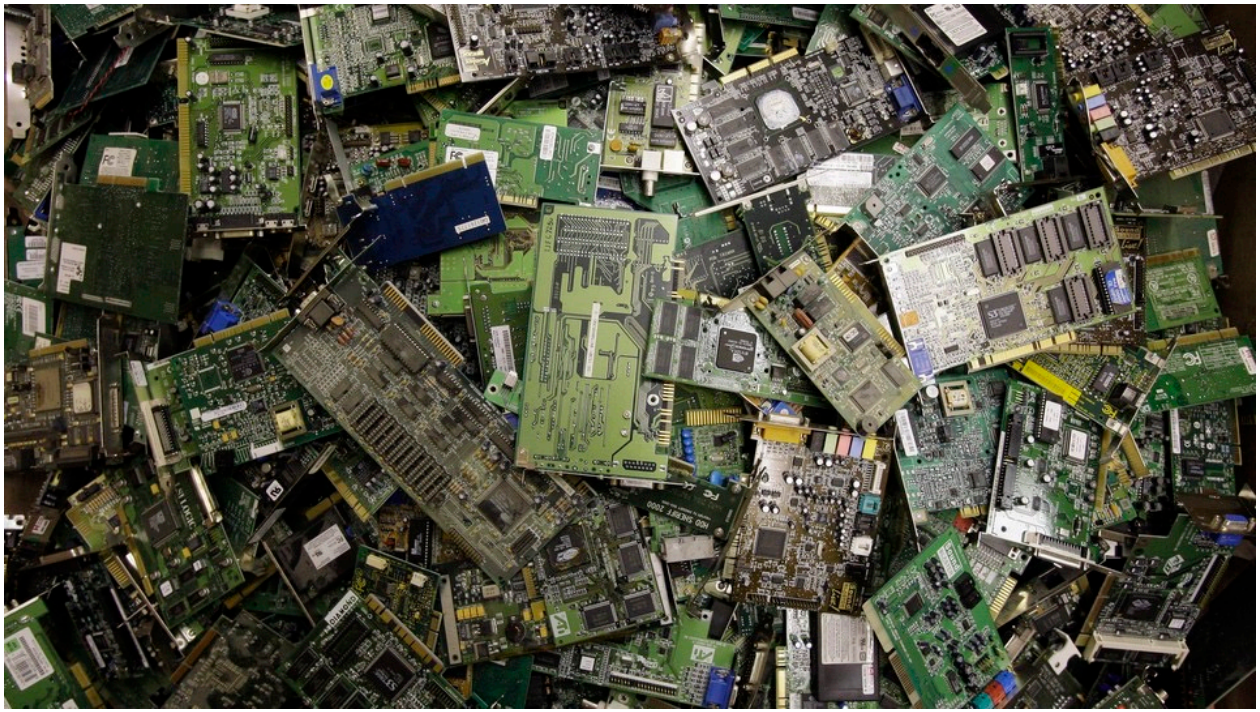
Sustainability is much more than combating global warming

It is also about

- Raw material extraction
- **E-waste**

due to linear economy

[Credit: Michael Conroy, AP]



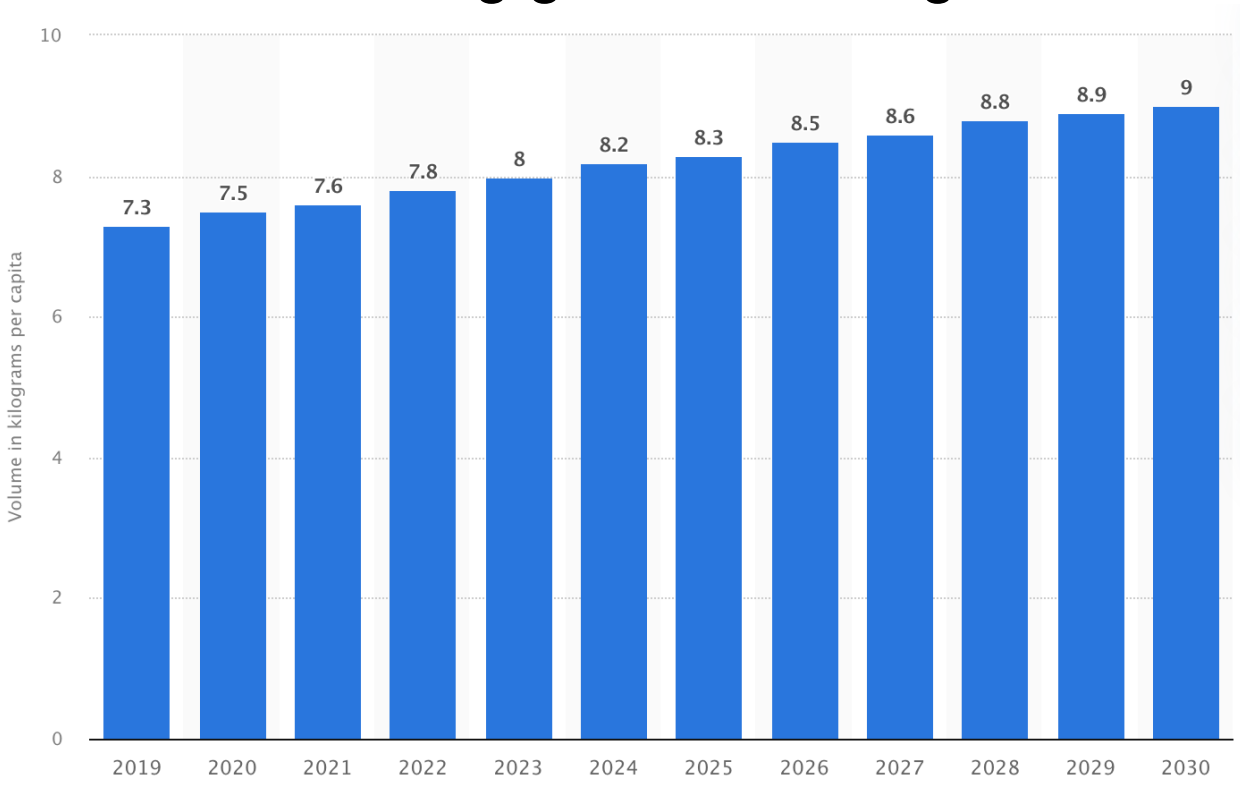
Sustainability is a Multi-Faceted Challenge

Sustainability is much more than combating global warming

It is also about

- Raw material extraction
- **E-waste**

- 8 kg per capita per annum
  - this includes small to large appliances
- only 17% gets recycled



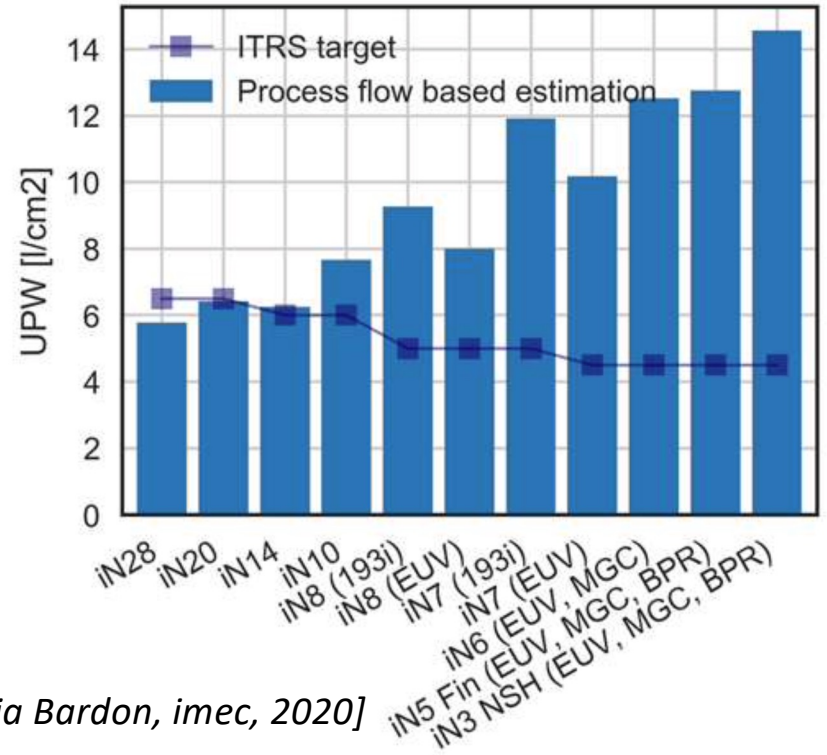
[Statista 2023]

Sustainability is a Multi-Faceted Challenge

Sustainability is much more than combating global warming

It is also about

- Raw material extraction
- E-waste
- **Ultra pure water**



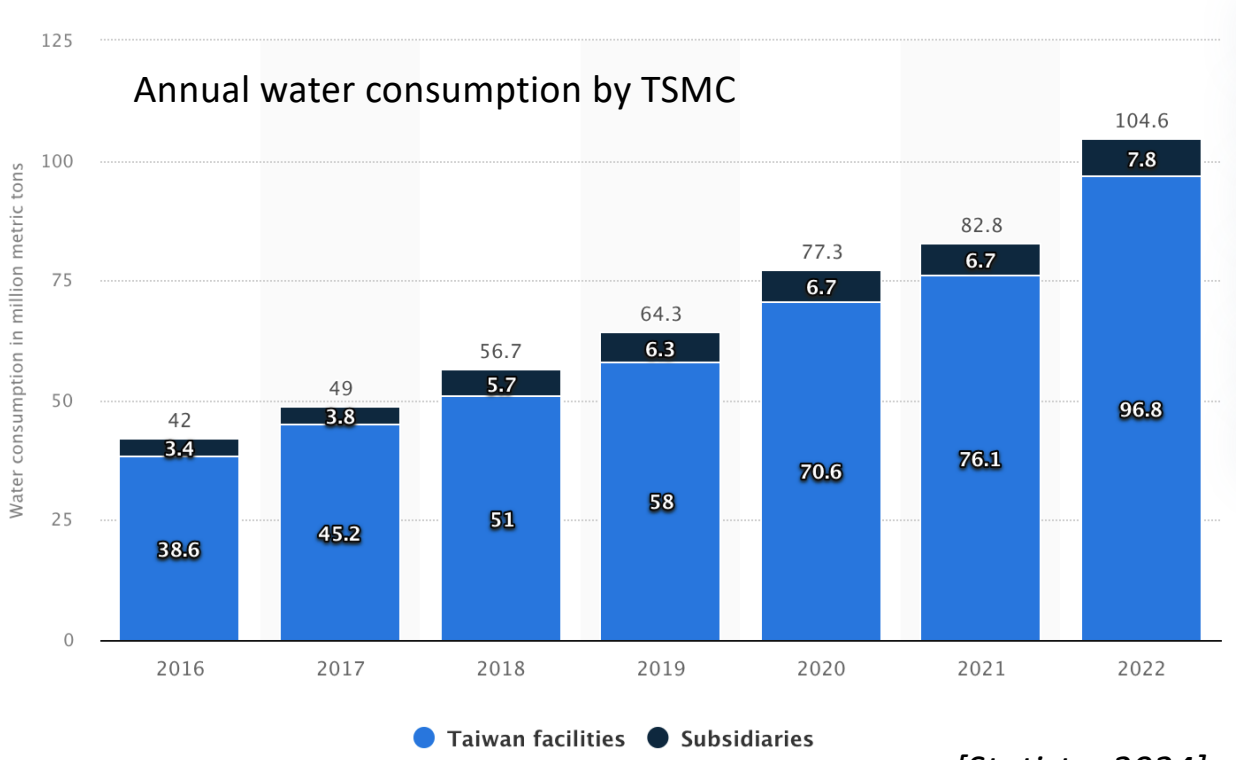
[M. Garcia Bardon, imec, 2020]

Sustainability is a Multi-Faceted Challenge

Sustainability is much more than combating global warming

It is also about

- Raw material extraction
- E-waste
- **Water usage**



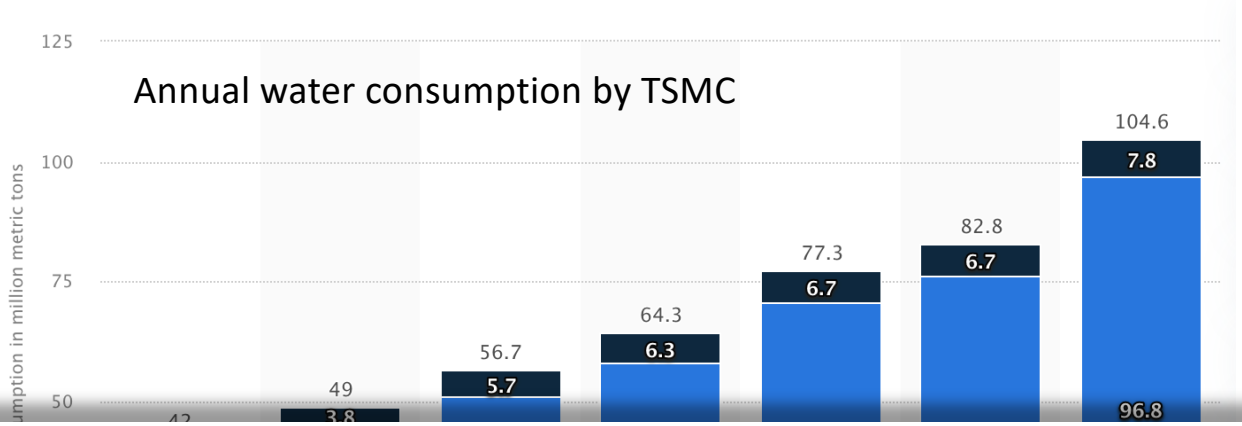
[Statista, 2024]

Sustainability is a Multi-Faceted Challenge

Sustainability is much more than combating global warming

It is also about

- Raw material extraction
- E-waste
- **Water usage**



CLIMATE

Epic drought in Taiwan pits farmers against high-tech factories for water

The island is facing one of its worst dry spells in a century, and both the agricultural and high-tech sectors are competing for scarce water resources.

April 19, 2023 | By: Emily Feng

[NPR, 2023]

Sustainability is a Multi-Faceted Challenge

Sustainability is much more than combating global warming

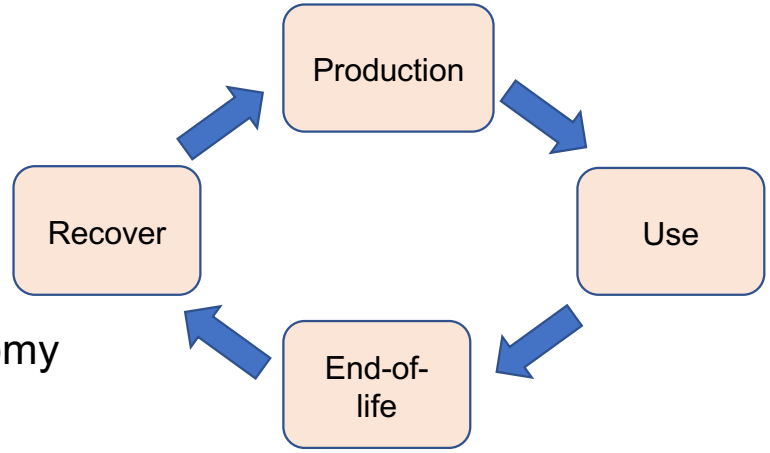
It is also about

- **Raw material extraction**
- **E-waste**
- **Water usage**
- **New business models & legislation**

Key motivation for circular (rather than linear) economy

Keep materials in the economy longer

- Fewer raw materials are needed
- Less impact on climate
- Avoid (e-)waste
- Improved security of material supply
  - Be less depending on third-party countries
- Design for repairability



Selling services instead of goods

Consumer wants (societal needs)

Light, not lamps

Mobility, not cars

Connectivity, not smartphone

Challenges when Doing Research in Computer System Sustainability

1. *Sustainability is multi-faceted problem*
   - *Global warming, raw materials, e-waste, water consumption, biodiversity, etc.*

2. ***Inherent data uncertainty***
   - ***Many unknowns, data limitations, industry secrecy***

3. *Need to account for entire lifecycle*
   - *Embodied footprint: raw material extraction, semiconductor manufacturing*
   - *Operational footprint: usage of device during lifetime*

Life Cycle Assessment (LCA) iPhone 12

Global Warming Potential (GWP 100 years) in CO_2

Carbon footprint: Estimated emissions are calculated in accordance with guidelines and requirements as specified by ISO 14040 and ISO 14044. There is inherent uncertainty in modeling carbon emissions due primarily to data limitations. For the top component contributors to Apple's carbon emissions, Apple addresses this uncertainty by developing detailed process-based environmental models with Apple-specific parameters. For the remaining elements of Apple's carbon footprint, we rely on industry average data and assumptions. Calculation includes emissions for the following life cycle phases contributing to Global Warming Potential (GWP 100 years) in CO_2 equivalency factors (CO_2e):

- **Production:** Includes the extraction, production, and transportation of raw materials, as well as the manufacture, transport, and assembly of all parts and product packaging.
- **Transport:** Includes air and sea transportation of the finished product and its associated packaging from manufacturing site to regional distribution hubs. Transport of products from distribution hubs to end customers is modeled using average distances based on regional geography.

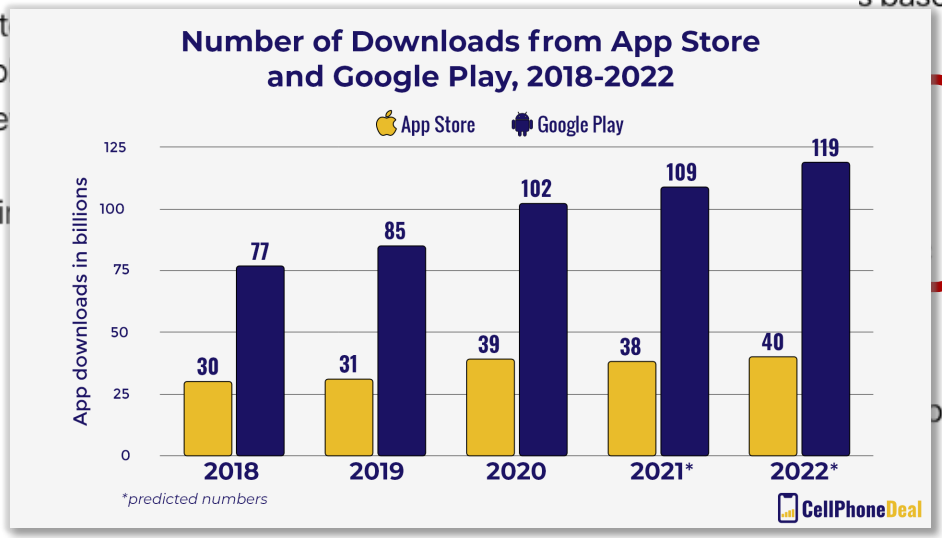
customers is modeled using average distances based on regional geography.

- **Use:** Apple assumes a three- or four-year period for power use by first owners based on the product type. Product use scenarios are based on historical customer use data for similar products. Geographic differences in the power grid mix have been accounted for at a regional level.
- **End-of-life processing:** Includes transportation from collection hubs to recycling centers and the energy used in mechanical separation and shredding of parts. For more information on the carbon footprint, visit apple.com/environment/answers.

Life Cycle Assessment

How predictive is historical data?

Global Warming Potential (GWP 100 years) in CO_2



Carbon footprint: Estimated emissions are calculated in accordance with guidelines and requirements as specified by ISO 14040 and ISO 14044. There is inherent uncertainty in modeling carbon emissions due primarily to data limitations. For the top component contributors to Apple's carbon emissions, Apple addresses this uncertainty by developing detailed process-based environmental models with Apple-specific parameters. For the remaining elements of Apple's carbon footprint, we rely on industry average data and assumptions. Calculation includes emissions for the following life cycle phases contributing to Global Warming Potential (GWP

The Truth About Smartphone Addiction, And How To Beat It

[Forbes, 2019]

as well as the manufacture, tra… …bly of all parts and product packaging.

Do Not Disturb: How I Ditched My Phone and Unbroke My Brain

[New York Times, 2019]

customers is mode… …average distances based on regional geography.

Challenges when Doing Research in Computer System Sustainability

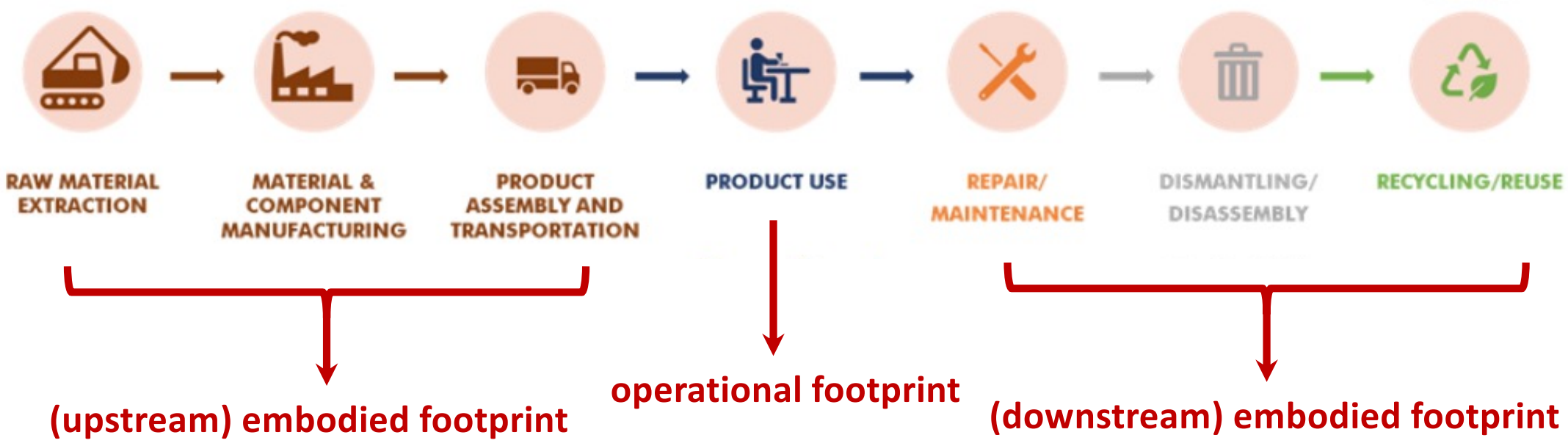
1. *Sustainability is multi-faceted problem*
   - *Global warming, raw materials, e-waste, water consumption, biodiversity, etc.*

2. *Inherent data uncertainty*
   - *Many unknowns, data limitations, industry secrecy*

3. ***Need to account for entire lifecycle***
   - ***Embodied footprint: raw material extraction, semiconductor manufacturing***
   - ***Operational footprint: usage of device during lifetime***

The Life of a Computer Device

Power/energy-efficient computing ignores embodied footprint



[Global Economic Council, 2021: State of Sustainability Research -- Climate Change Mitigation]

Does a more energy/power-efficient computing device lead to an overall reduction in carbon footprint?

Not necessarily!

Making an individual device more carbon-friendly is necessary condition, but not a sufficient condition!

Rebound Effect due to Improved Efficiency

Counter-intuitive finding: making an *individual* system more energy/power-efficient may lead to an *overall* increase in footprint

Rebound effect of making systems more efficient a.k.a. Jevons' paradox

more efficient system → cheaper/easier to use → increased usage and deployment → increased (embodied and operational) footprint

William Stanley Jevons (1865) first describes this rebound effect

- James Watt improved the efficiency of coal-fired steam engine
  - Each steam engine uses less coal, so coal became a more cost-effective fuel
- This led to an increased use of steam engines in a variety of industries
- The result was increased overall coal consumption

Part II

How does the global environmental footprint of computing scale?

What are the contributing factors?

[L. Eeckhout, "Kaya for Architects: Towards Sustainable Computer Systems", IEEE Micro, 2023]

Kaya Identity

Contributing factors to carbon emissions *[by energy economist Yoichi Kaya, 1997]*

$F = P \times G/P \times E/G \times F/E$, with

F = global CO_2 emissions

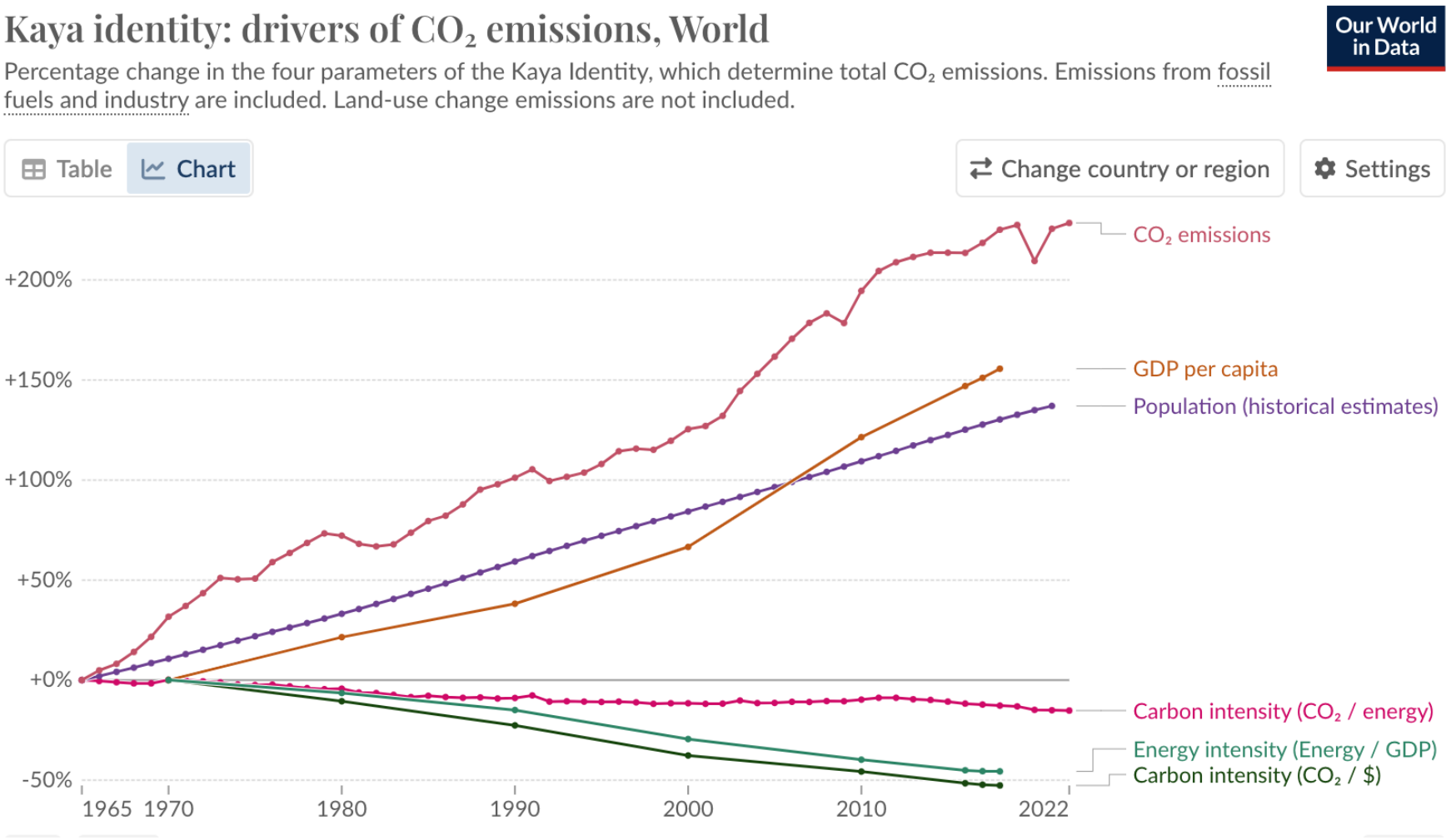
P = global population

G/P = GDP per capita

E/G = energy intensity of the GDP

F/E = carbon footprint of energy

Kaya identity is used by the Intergovernmental Panel on Climate Change (IPCC) to predict world CO_2 emission scenarios and impact on global warming



Despite improvements in energy intensity and carbon intensity, we witness an overall increase in CO_2 emissions

Reformulating Kaya for Architects

Can we reformulate the Kaya identity to something we, computer architects, gain insight from?

… so we can understand how to reduce environmental impact of computing?

We focus on carbon footprint

- But representative for other sustainability issues
- Using recently published numbers, yet to be taken with grain of salt…

Distinction between

- **Embodied emissions**: GHG emissions during manufacturing process
  - **Scope-1:** chemicals and gases emitted
  - **Scope-2:** carbon emissions from energy usage
  - **Scope-3:** due to material extraction [not considered here]
- **Operational emissions**: GHG emissions during product lifetime

Total Carbon Footprint

Embodied Scope-2 (energy usage during production)

$$\mathbf{CO2e_{embodied,\ scope\text{-}2} = \#chips \times \#wafer/chips \times kWh/wafer \times CO2e/kWh}$$

Embodied Scope-1 (chemicals and gases during production)

$$\mathbf{CO2e_{embodied,\ scope\text{-}1} = \#chips \times \#wafer/chips \times CO2e/wafer}$$

Operational (energy usage during lifetime)

$$\mathbf{CO2e_{operational} = \#chips \times kWh/chip \times CO2e/kWh}$$

[L. Eeckhout, "Kaya for Architects: Towards Sustainable Computer Systems", IEEE Micro, 2023]

Demand for Chips is Increasing

Embodied Scope 2 (energy usage during production)

$CO2e_{embodied,\ scope\text{-}2}$ = #chips × wafer/chips × kWh/wafer × CO2e/kWh

Embodied Scope 1 (chemicals and gases during production)

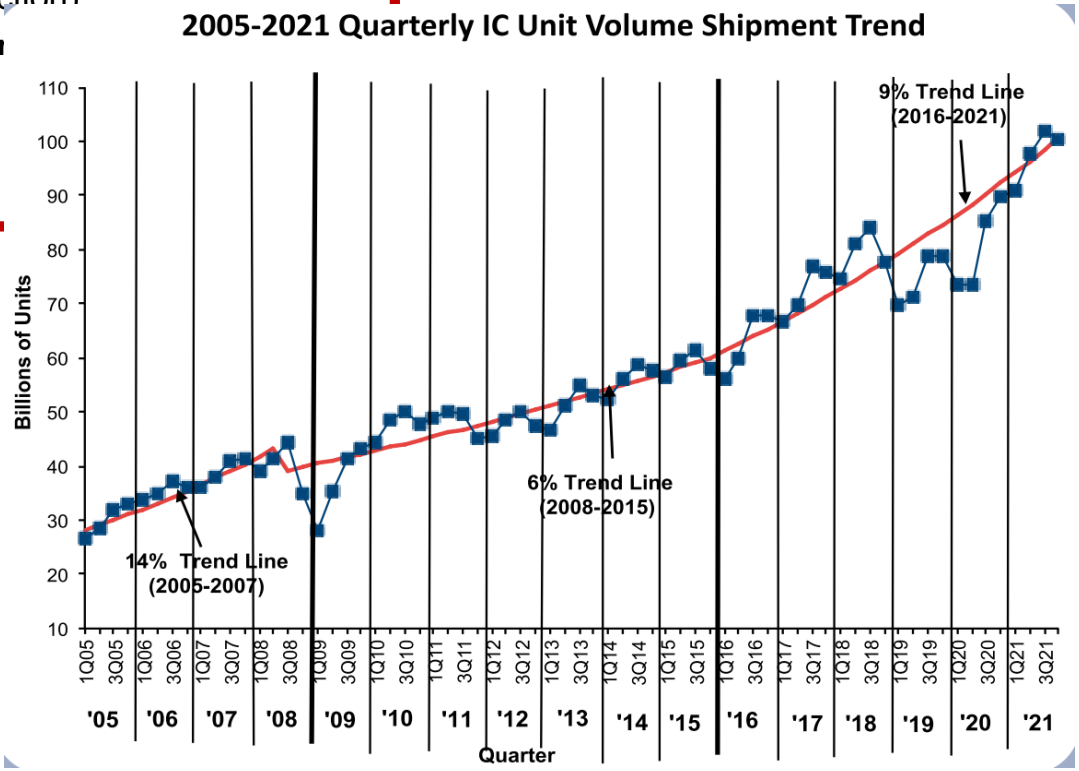
$CO2e_{embodied,\ scope\text{-}1}$ = #chips × wafer/chips × CO2e/wafer

Operational (energy usage during lifetime)

$CO2e_{operational}$ = #chips × kWh/chip × CO2e/kWh

Increasing number of chips:

CAGR = +9%



[IC Insight, 2022]

Die Size Seems to Have Stagnated

Embodied Scope 2 (energy usage during production)

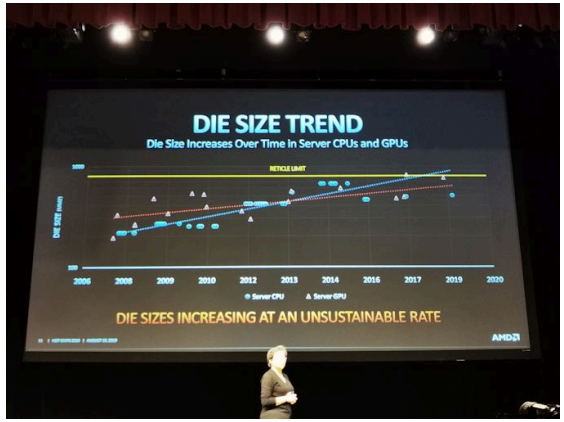
$CO2e_{embodied,\ scope\text{-}2}$ = #chips × #wafer/chips × kWh/wafer × CO2e/kWh

Embodied Scope 1 (chemicals and gases during production)

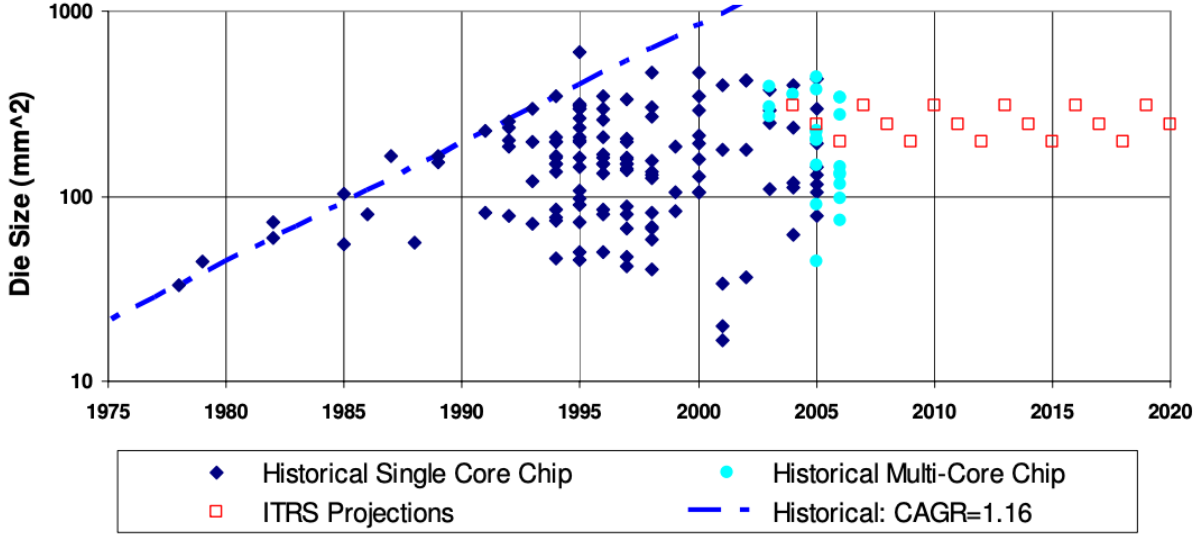
$CO2e_{embodied,\ scope\text{-}1}$ = #chips × #wafer/chips × CO2e/wafer

Operational (energy usage during lifetime)

$CO2e_{operational}$ = #chips × kWh/chip × CO2e/kWh



Number of chips per wafer: CAGR ≃ +0%



[Kogge et al., 2008]

Increasing Energy Demand per Wafer

Embodied Scope 2 (energy usage during production)

$CO2e_{embodied,\ scope\text{-}2}$ = chips × wafer/chips × kWh/wafer × CO2e/kWh

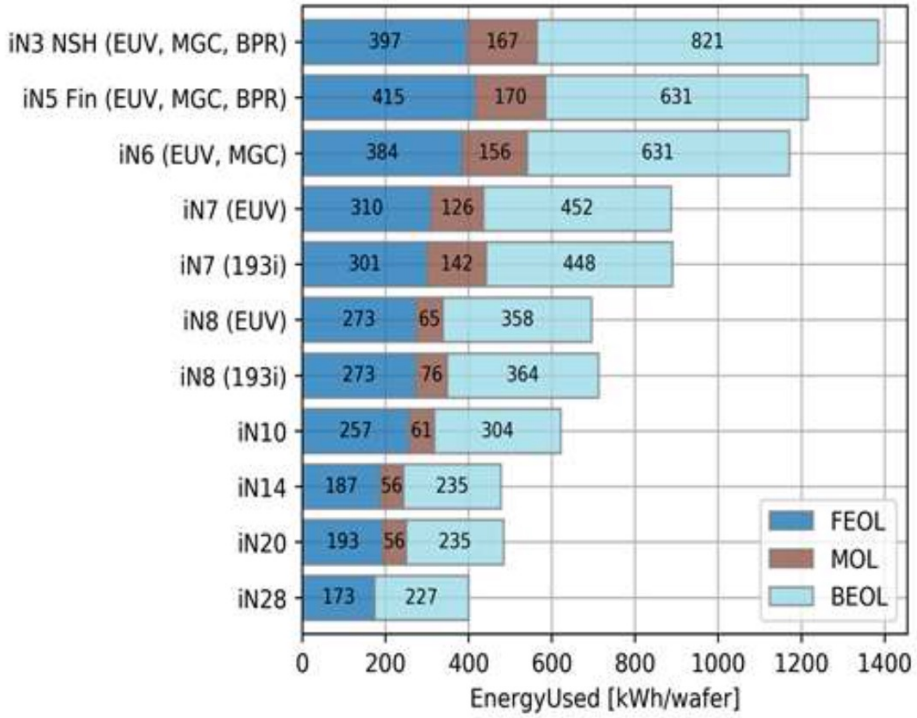
Embodied Scope 1 (chemicals and gases during production)

$CO2e_{embodied,\ scope\text{-}1}$ = chips × wafer/chips × CO2e/wafer

Operational (energy usage during lifetime)

$CO2e_{operational}$ = chips × kWh/chip × CO2e/kWh

[M. Garcia Bardon, imec, 2020]



Increasing energy demand for new tech nodes

- increasing no. processing steps
- CAGR kWh/wafer = +11.9%

Increasing Energy Demand per Wafer

Embodied Scope 2 (energy usage during production)

$CO2e_{embodied,\ scope\text{-}2}$ = chips × wafer/chips × kWh/wafer × CO2e/kWh

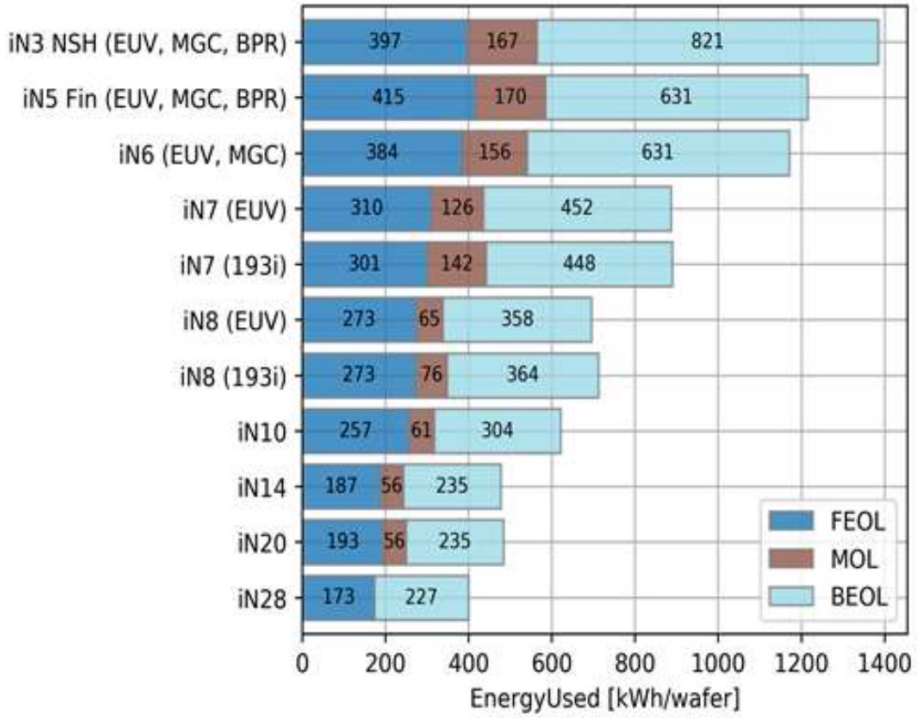
Embodied Scope 1 (chemicals and gases during production)

$CO2e_{embodied,\ scope\text{-}1}$ = chips × wafer/chips × CO2e/wafer

Operational (energy usage during lifetime)

$CO2e_{operational}$ = chips × kWh/chip × CO2e/kWh

[M. Garcia Bardon, imec, 2020]



Press Release > Climate & Energy

Semiconductor industry electricity consumption to more than double by 2030: study

Greenpeace East Asia

April 20, 2023

Increasing Chemicals/Gases per Wafer

Embodied Scope 2 (energy usage during production)

$CO2e_{embodied,\ scope\text{-}2}$ = #chips × #wafer/chips × kWh/wafer × CO2e/kWh

Embodied Scope 1 (chemicals and gases during production)

$CO2e_{embodied,\ scope\text{-}1}$ = #chips × #wafer/chips × CO2e/wafer

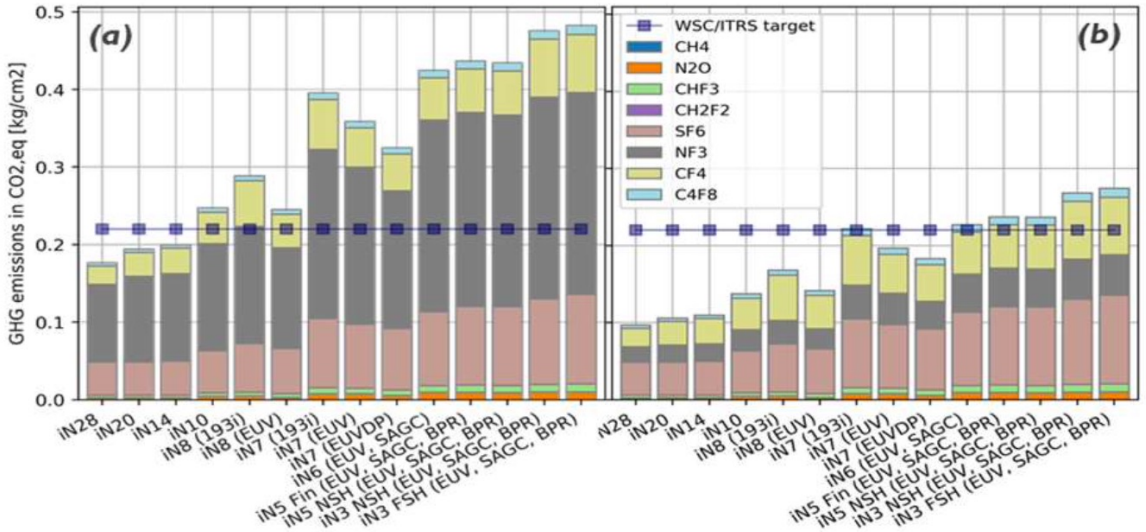
Operational (energy usage during lifetime)

$CO2e_{operational}$ = #chips × kWh/chip × CO2e/kWh

[M. Garcia Bardon, imec, 2020]

Increasing chemical/gas emissions for new tech nodes

CAGR CO2e/wafer = +9.4%



Carbon Intensity Slowly Decreasing

Embodied Scope 2 (energy usage during production)

$CO2e_{embodied,\ scope\text{-}2}$ = #chips × #wafer/chips × kWh/wafer × CO2e/kWh

Embodied Scope 1 (chemicals and gases during production)

$CO2e_{embodied,\ scope\text{-}1}$ = #chips × #wafer/chips × CO2e/wafer

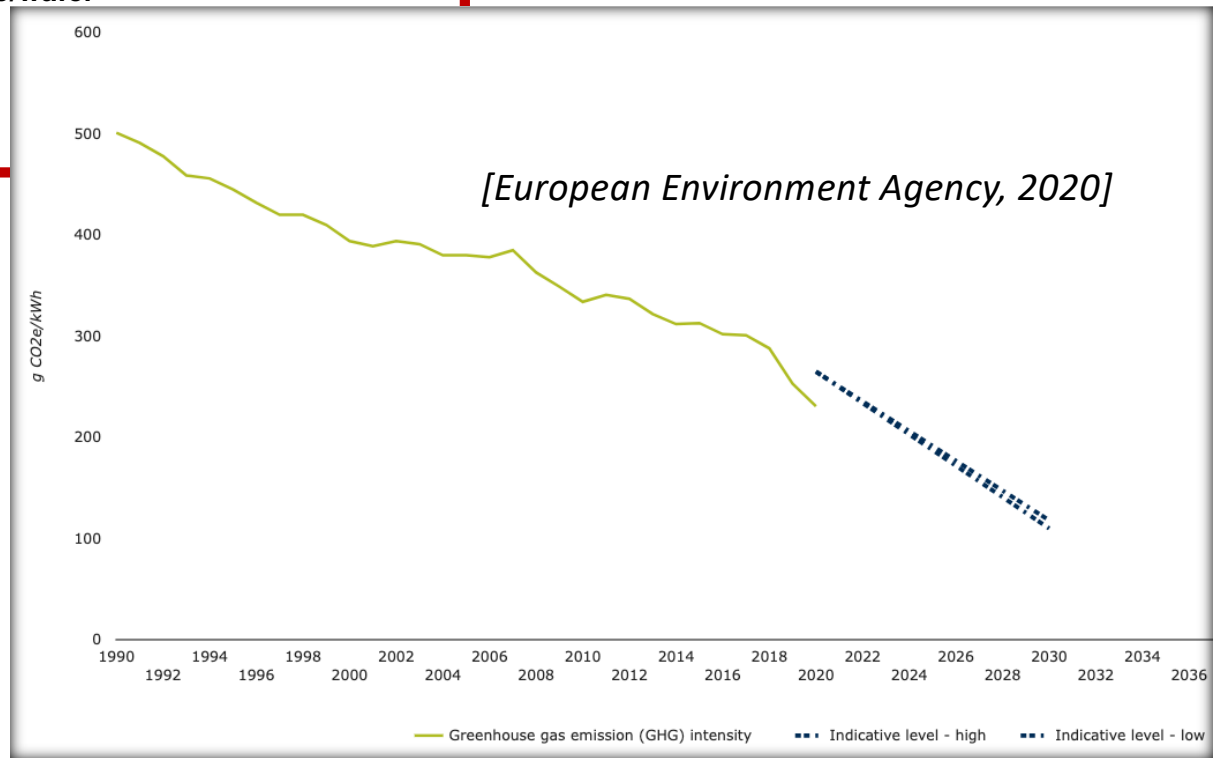
Operational (energy usage during lifetime)

$CO2e_{operational}$ = #chips × kWh/chip × CO2e/kWh

- Transition towards green energy sources

  CO2e/kWh (Europe): CAGR = -2.5%

- Much faster transition to green energy in the datacenter *[Gupta et al., HPCA 2021]*
  - Critical side note: green energy contracts deprive other customers from green energy



Decreasing Operational Energy

Embodied Scope 2 (energy usage during production)

$CO2e_{embodied,\ scope\text{-}2}$ = #chips × #wafer/chips × kWh/wafer × CO2e/kWh

Embodied Scope 1 (chemicals and gases during production)

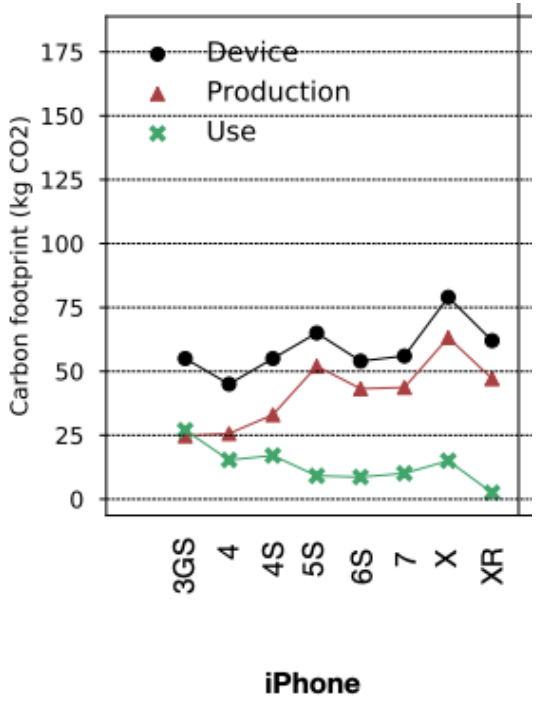
$CO2e_{embodied,\ scope\text{-}1}$ = #chips × #wafer/chips × CO2e/wafer

Operational (energy usage during lifetime)

$CO2e_{operational}$ = #chips × kWh/chip × CO2e/kWh

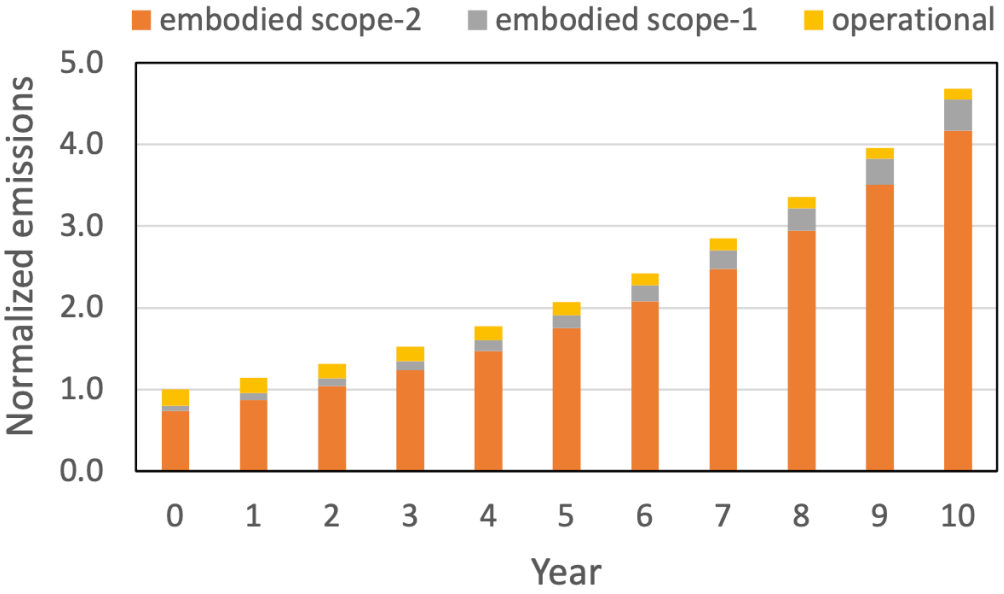
Operational energy consumption is decreasing

kudos to ourselves! ☺

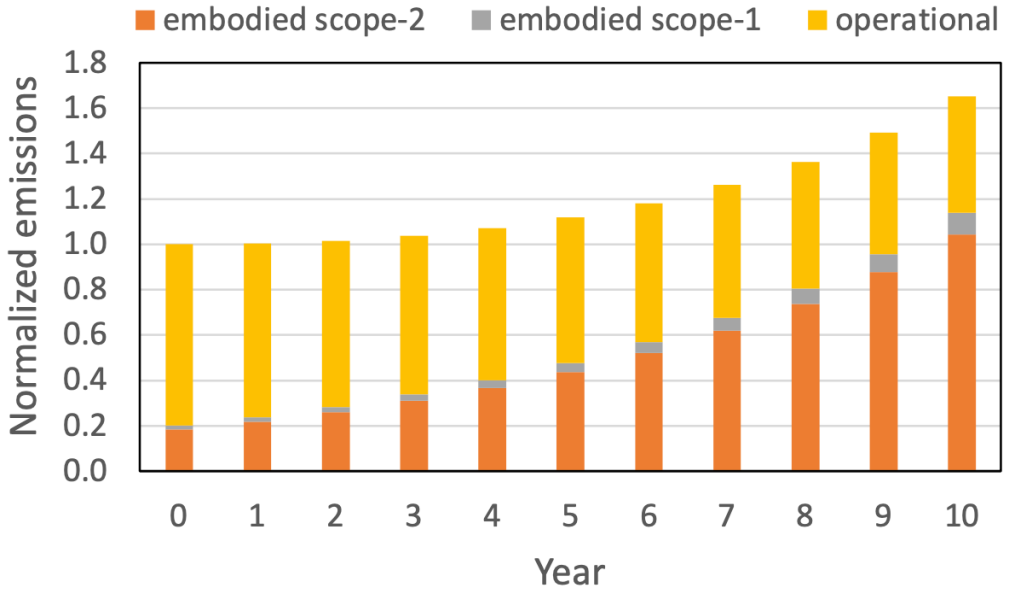


[U. Gupta et al. , HPCA 2021]

Putting it Together: Current Trends for Total Carbon Emissions



(a) initially dominating embodied emissions



(b) initially dominating operational emissions

Total carbon footprint continues to grow

Embodied emissions grow in importance and (will) dominate

Reason: increasing number of chips *and* growing energy intensity of chip manufacturing

Total Carbon Footprint

Embodied Scope-2 (energy usage during production)

$CO2e_{embodied,\ scope\text{-}2}$ = #chips × wafer/chips × kWh/wafer × CO2e/kWh

Embodied Scope-1 (chemicals and gases during production)

$CO2e_{embodied,\ scope\text{-}1}$ = #chips × wafer/chips × CO2e/wafer

Operational (energy usage during lifetime)

$CO2e_{operational}$ = #chips × kWh/chip × CO2e/kWh

Key take-aways:

- Demand for chips keeps increasing (Jevons' paradox?)
- GHG emissions (both scope-1 and 2) increase with new tech nodes
- Transition to green energy not moving fast enough ***and*** it doesn't impact scope-1 nor scope-3 emissions (and other sustainability issues like raw material need, e-waste, water usage, etc.)
- Embodied emissions dominate or will soon dominate

Part III

How to reason about sustainable computer system design in light of inherent data uncertainty?

Embrace it!

FOCAL: First-Order CArbon ModeL

FOCAL is a top-down, parameterized model that

- is deliberately simple,
- is built upon first principles, and
- provides insight

Key idea:

- use proxies for embodied and operational footprint,
- parameterize relative importance of embodied versus operational footprint,
- while considering different use case scenarios, incl. rebound effects

FOCAL enables powerful analyses despite inherent data uncertainty:
- *similar conclusions across a range of scenarios → confident conclusions*
- *otherwise → need to be careful when reaching conclusions*

Proxy for Embodied Footprint?

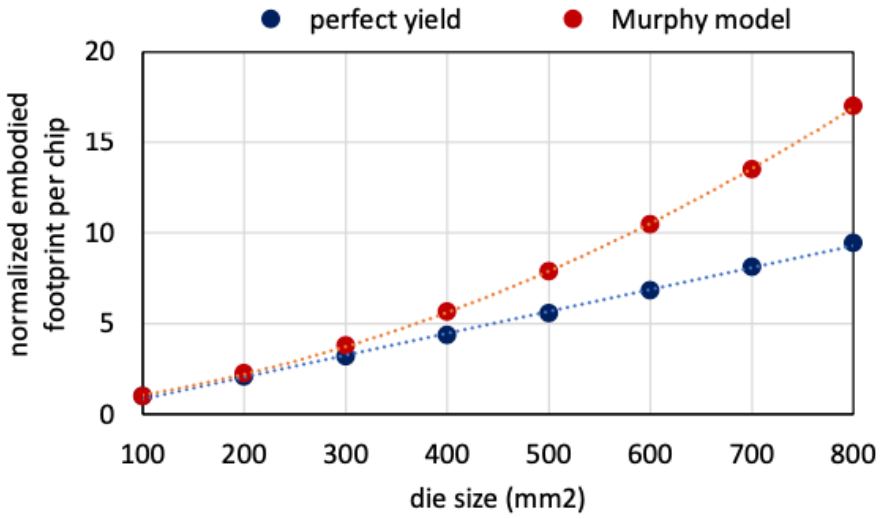
Wafer = production unit in semiconductor fab

- Environmental impact for producing a wafer: energy consumed, chemicals and gases emitted, ultra pure water used, materials used

The bigger the size of a chip, the higher its embodied footprint

- Accounting for lost silicon wafer area
  [de Vries, 2005]
- Accounting for yield issues
  [Murphy model,
  TSMC: 0.09 defect density per cm^2]

Proxy = chip area (A)



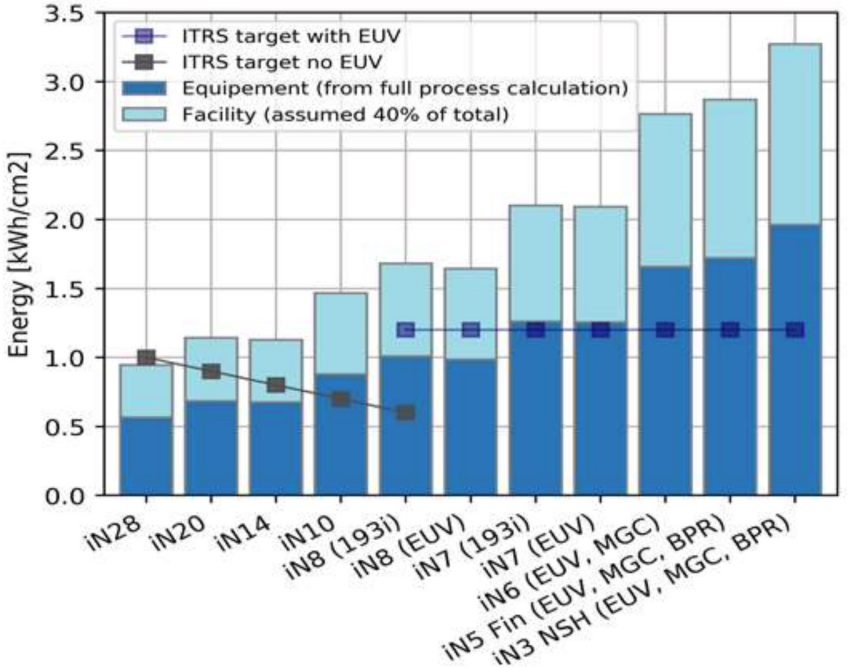
Proxy for Embodied Footprint?

Embodied footprint of an IC is proportional to its area

Amount of energy needed (and chemicals/gases emitted) to produce a wafer increases with newer chip technologies

From imec: iN28 (~2011) to iN3 (~2022)

CAGR = +11.9%



[M. Garcia Bardon, imec, 2020]

Proxy = chip area (A)

Embodied footprint =

$$A\ [cm^2] \times E_l\ [kWh / cm^2] \times C_l\ [CO2e / kWh]$$

Proxy for Operational Footprint? (1/2)

(1) Fixed-work scenario

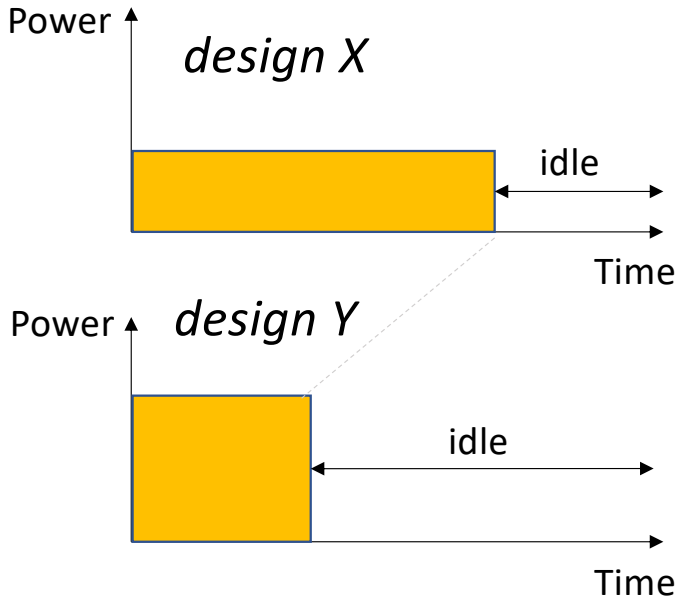
- Assumption: a device performs fixed amount of work over its entire lifetime

The higher energy consumption, the higher its operational footprint

Proxy = energy consumption (E)

Operational footprint =

$$E\ [kWh] \times C_I\ [CO2e / kWh]$$



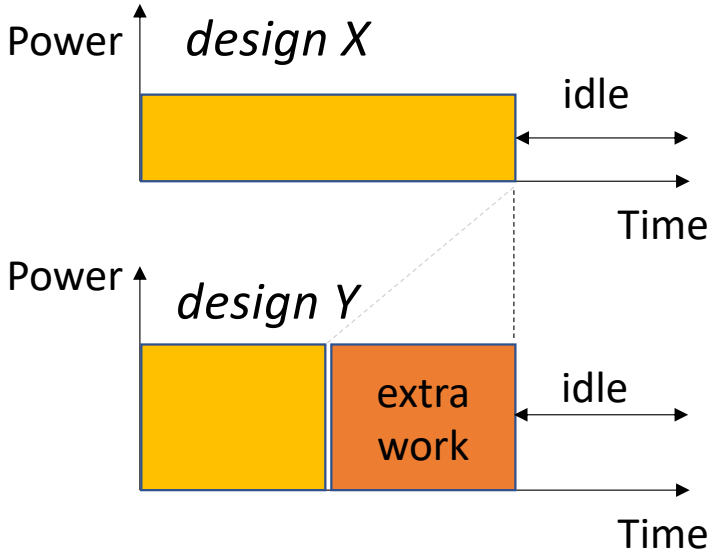
Proxy for Operational Footprint? (2/2)

(2) Fixed-time scenario – *more realistic scenario(?)*

- We perform more work because it is more efficient, cf. Jevons' paradox
- Assumption: we use the device for the same amount of time

The higher power consumption, the higher its operational footprint

Proxy = power consumption (P)



How to Weigh Embodied versus Operational Footprint?

Ratio of embodied vs operational footprint depends on

Device type

Battery-operated vs always-on devices

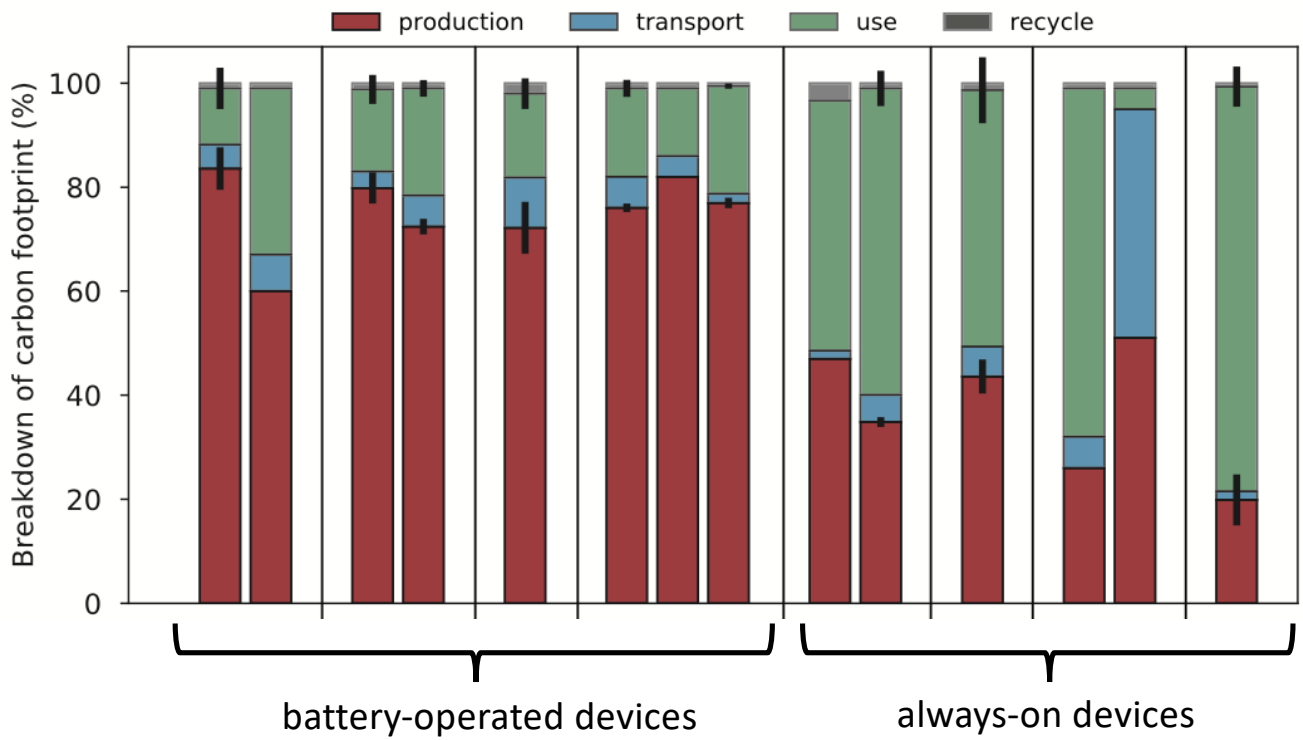
Lifetime

The longer the lifetime, the higher the relative weight of operational footprint

Energy mix

The greener the energy mix during lifetime, the higher the relative weight of embodied footprint

Answer: we parameterize the embodied-vs-operational footprint



[Gupta et al., HPCA 2021]

FOCAL Computes the Normalized Carbon Footprint (NCF)

fixed-work:

$$NCF_{fw,\alpha_{E2O}}(X,Y) = \alpha_{E2O}\frac{A_X}{A_Y} + (1-\alpha_{E2O})\frac{E_X}{E_Y}$$

fixed-time:

$$NCF_{ft,\alpha_{E2O}}(X,Y) = \alpha_{E2O}\frac{A_X}{A_Y} + (1-\alpha_{E2O})\frac{P_X}{P_Y}$$

α_{E2O} parameter is a function of device type/usage, lifetime of device, rebound effect, energy source during manufacturing vs lifetime

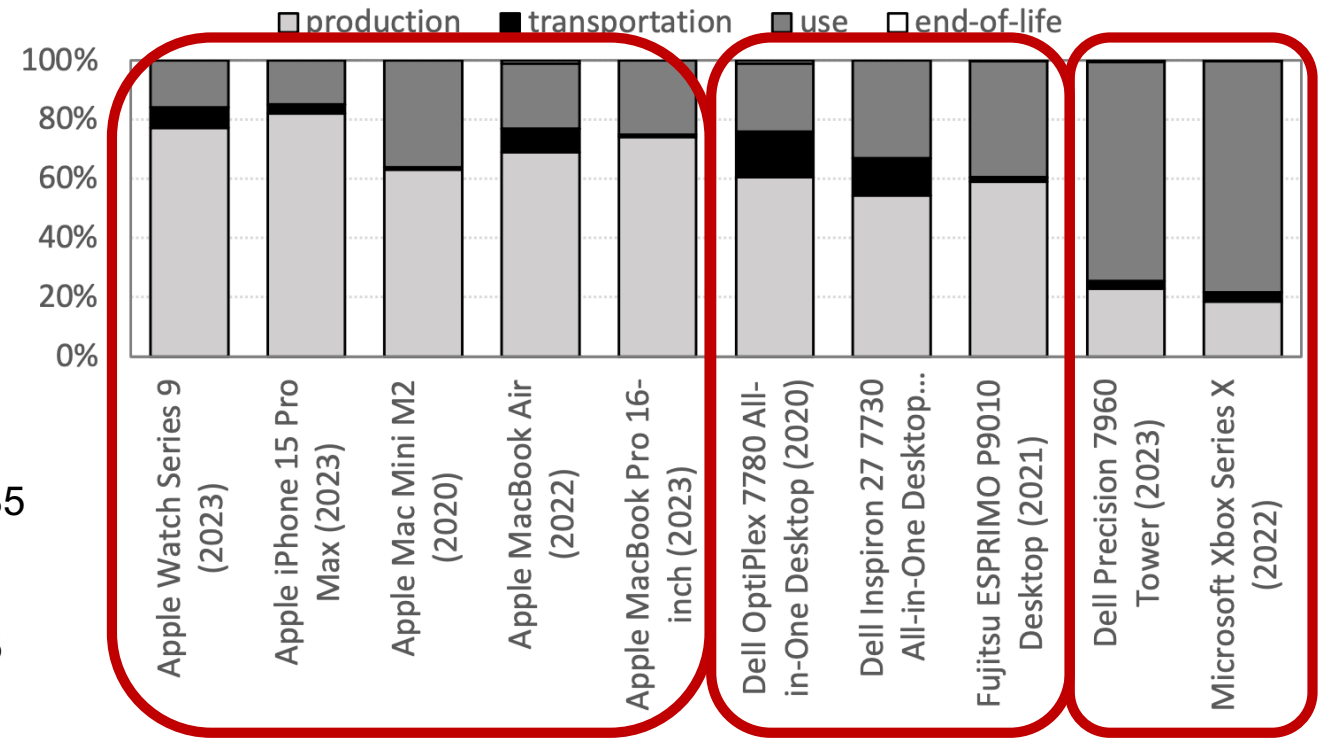
Parameterization allows for considering different scenarios w/ confidence intervals:

- Embodied emissions dominate (assume $\alpha_{E2O} = 0.8 \pm 0.1$) versus
- Operational emissions dominate (assume $\alpha_{E2O} = 0.2 \pm 0.1$)
- Fixed-work versus fixed-time

[L. Eeckhout, "FOCAL: A First-Order Model to Assess Processor Sustainability", ASPLOS, 2023]

Total Carbon Footprint = Embodied + Operational Footprint

$\alpha_{E2O} \in [0,1]$ = relative weight of embodied vs operational footprint



smartphones & smartwatches
$\alpha_{E2O} \simeq 0.80 – 0.85$

laptops
$\alpha_{E2O} \simeq 0.7 – 0.75$

medium-range desktops
$\alpha_{E2O} \simeq 0.55 – 0.60$

high-end desktops & gaming consoles
$\alpha_{E2O} \simeq 0.20 – 0.25$

What insight can we gain from this simple model?

Evaluating Archetypal Processor Design Choices using FOCAL

A design choice is

- **strongly sustainable** if it reduces carbon footprint under both the fixed-work and fixed-time scenarios

  → ***no risk for rebound effect***

  e.g., **die shrink**, **multicore**, pipeline gating, dynamic voltage and frequency scaling

- **weakly sustainable** if it reduces carbon footprint only under a fixed-work scenario

  → ***(substantial) risk for rebound effect***

  e.g., **speculation** (branch prediction, runahead), heterogeneity, **acceleration**, caching

- **less sustainable** if it increases carbon footprint under both the fixed-work and fixed-time scenarios

  e.g., high-complexity microarchitecture (out-of-order vs. in-order), **dark silicon**, turboboosting

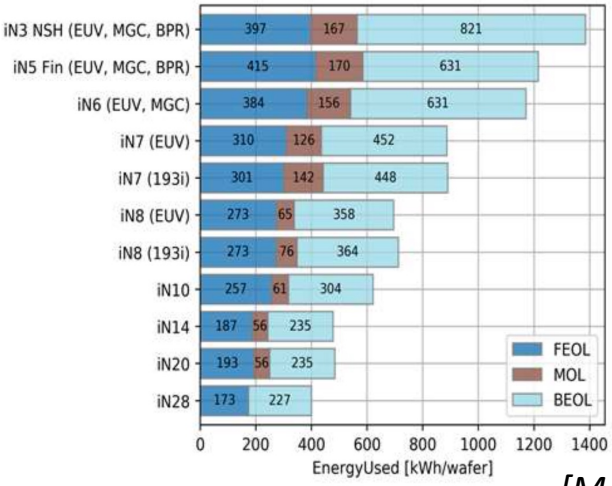
All results obtained using analytical models and published results → low carbon footprint research project ☺
Just a few examples follow – many more in ASPLOS 2024 paper

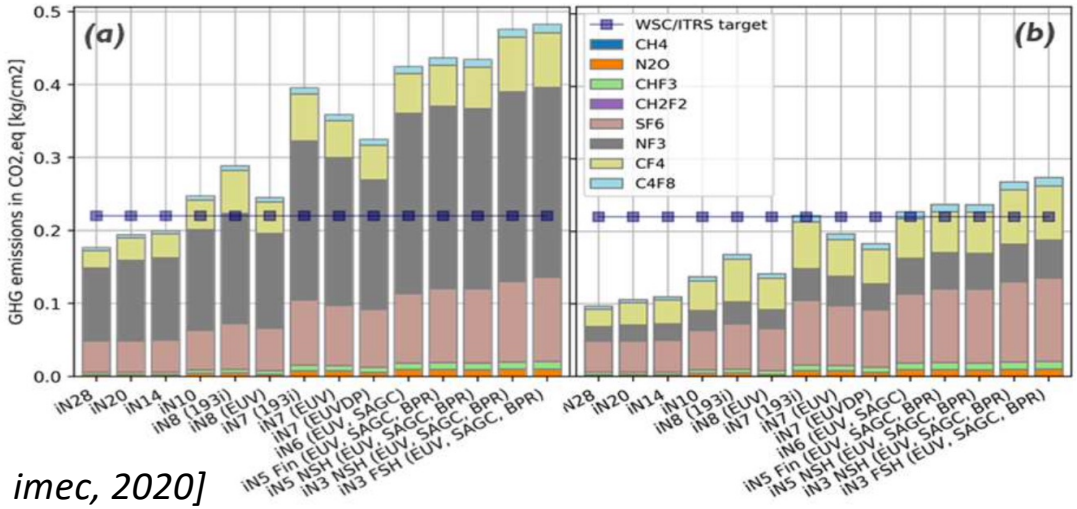
#1: Die Shrink is Strongly Sustainable

Implement an existing microarchitecture in a new tech node

Embodied emissions: net decrease

- Reduction of chip area by 50%
- This offsets the increase in energy consumption during manufacturing (+25%) and increase in chemicals/gases emitted (+19.5%)





[M. Garcia Bardon, imec, 2020]

#1: Die Shrink is Strongly Sustainable

Implement an existing microarchitecture in a new tech node

Embodied emissions: net decrease

- Reduction of chip area by 50%
- This offsets the increase in energy consumption during manufacturing (+25%) and increase in chemicals/gases emitted (+19%) *[Imec, 2020]*

Operational emissions: net decrease or neutral

- Classical scaling: power reduces by 2x, performance increases by 1.41x, and energy reduces by 2.82x
- Post-Dennard scaling: power remains the same, energy reduces by 1.41x

Overall conclusion: net reduction in environmental footprint

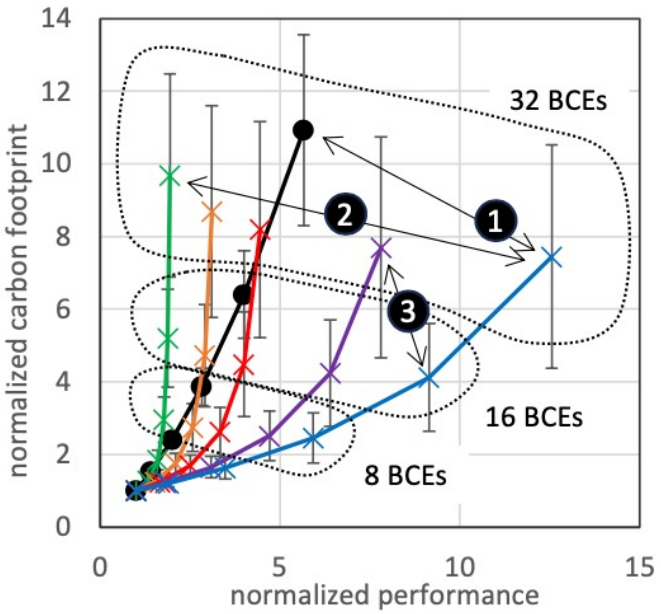
This is not what we've seen, on the contrary – cf. Jevons' paradox

#1: Multi-core is Strongly Sustainable

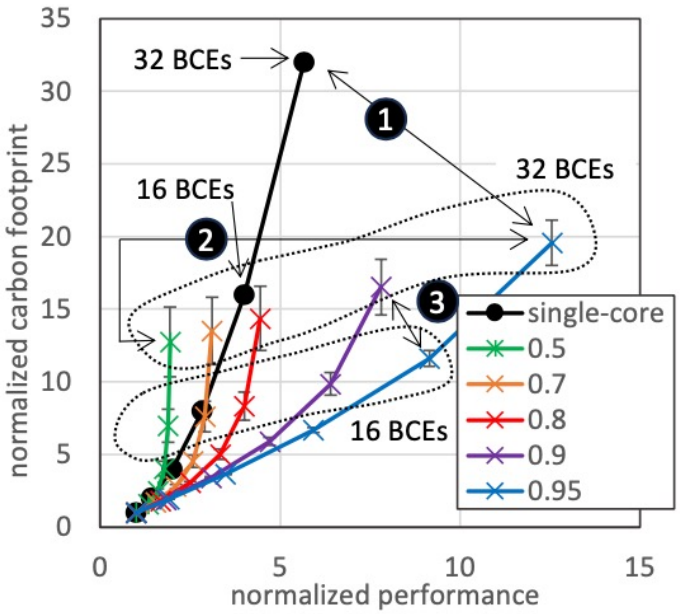
Key insights:

1. Multicore is strongly sustainable compared to single core
2. Parallelizing software is weakly sustainable
3. Parallelizing software is more sustainable than adding cores

fixed-work



fixed-time



Using Amdahl's Law
[Hill & Marty, 2008]
[Woo & Lee, 2008]

f = degree of parallelism
[see legend]

BCE = Base Core Equivalent
= number of cores
= unit of chip area

assuming $\alpha_{E2O} = 0.2 \pm 0.1$

#2: CPU Speculation is Weakly Sustainable

Branch prediction*

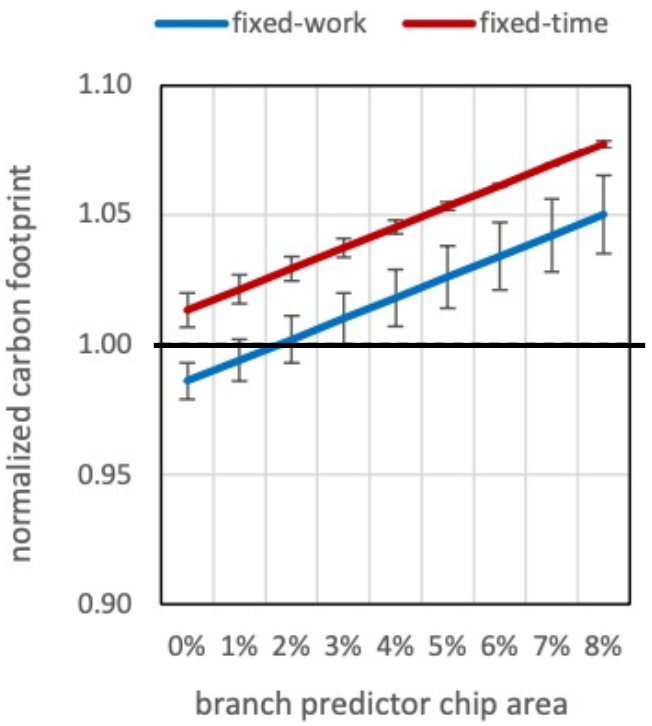
Large hybrid vs small bimodal predictor:

- 14% higher performance
- 7% less energy
- 6.6% higher power consumption

If operational emissions dominate, branch prediction is weakly sustainable

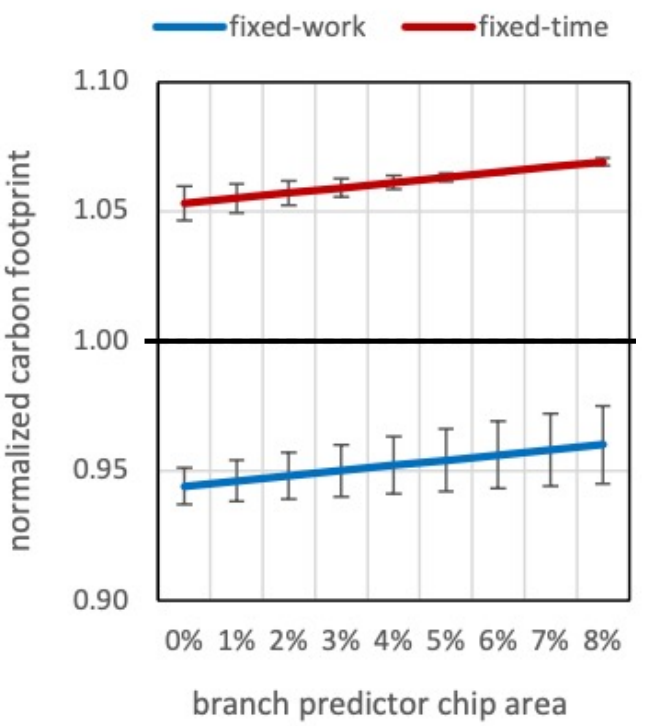
If embodied emissions dominate, branch prediction is less sustainable

**[Parikh et al., ISCA 2002]*



(a) embodied dominated

α_{E2O} = 0.8 ± 0.1



(b) operational dominated

α_{E2O} = 0.2 ± 0.1

#2: Acceleration is Weakly Sustainable

H.264 accelerator* versus general-purpose CPU implementation

- Accelerator consumes 500x less energy
- Accelerator is 15x smaller
- Similar performance

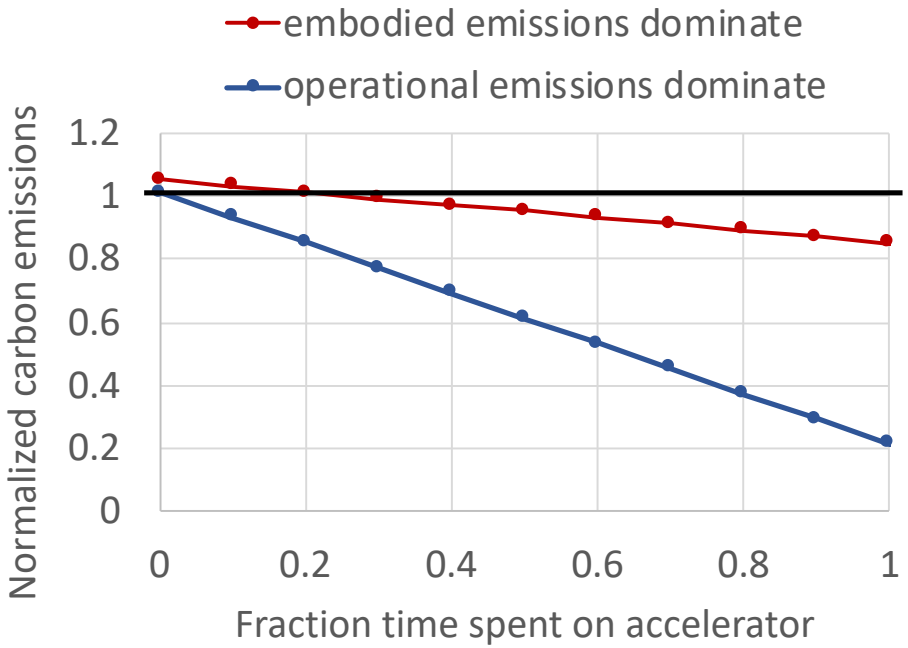
Embodied emissions = CPU + accelerator

- Accelerator = 6.5% extra chip area over CPU

Operational emissions = (1-f) x E_{CPU} + f x $E_{accelerator}$

- Depends on fraction f spent on accelerator

If embodied emissions dominate, it is critical that the accelerator is used for a sufficient fraction of time to be more sustainable than a CPU implementation



*[Hameed et al., ISCA 2010]

#3: Dark Silicon is Not Sustainable

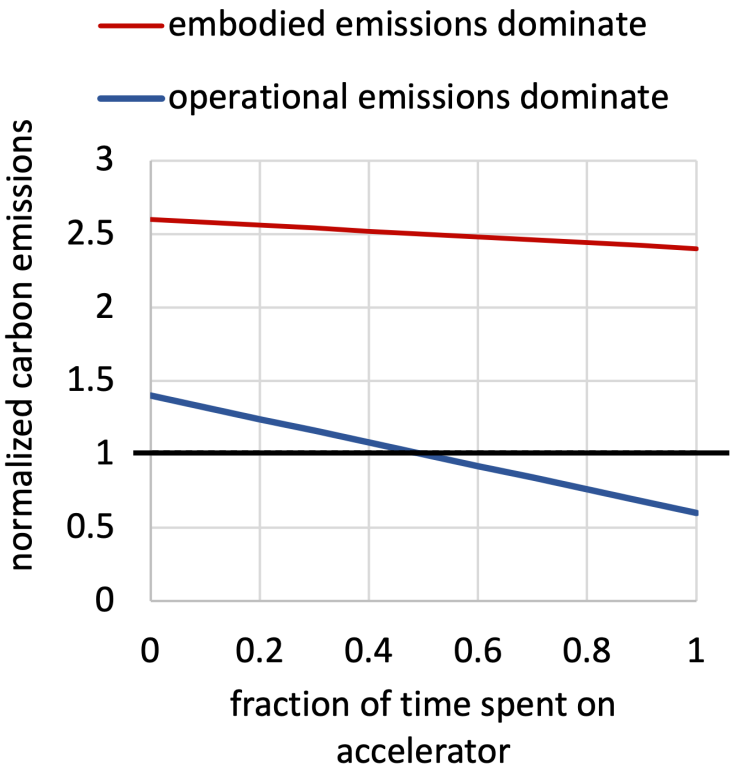
Dark silicon trades off chip area (increased embodied footprint) for power/energy efficiency (reduced operational footprint) – does it increase or decrease overall footprint?

We assume*

- Accelerators consume 500x less energy
- Similar performance
- Accelerators take up 2/3 of total chip area

Dark silicon is harmful if embodied emissions dominate

If operational emissions dominate, we need to use dark silicon very frequently, which is impossible



*[Hameed et al., ISCA 2010]

Part IV

How to design sustainable computer systems?

Three examples: (1) Sustainable multi-core scaling

(2) Hardware design through PPA analysis

(3) Hardware reconfigurability

Case Study: Sustainable Multi-Core Scaling

Baseline: quad-core processor in current tech node

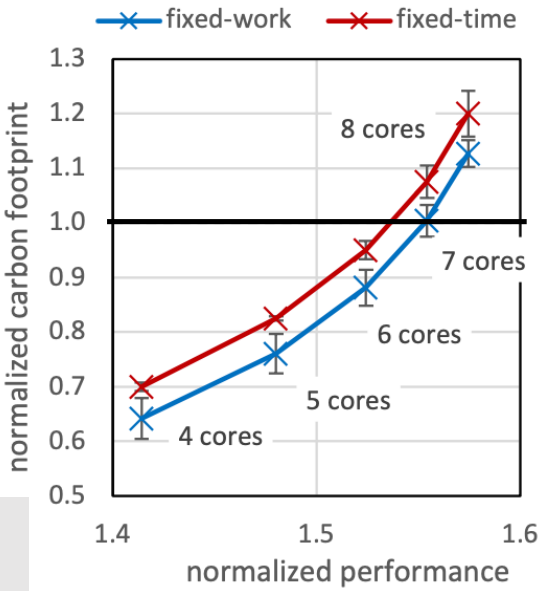
Question: how many cores in next-generation tech node? -- impact of tech node using *[imec, 2020]*

Pathway towards sustainable processor design

4, 5 or 6 cores are strongly sustainable options → significant performance boost *and* lower carbon footprint

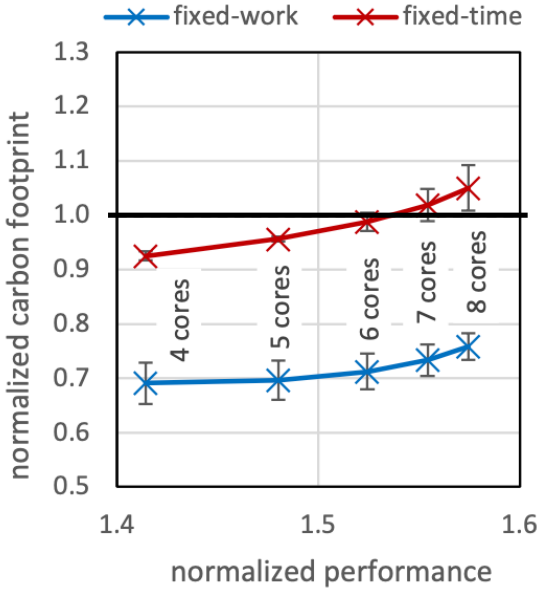
7 or 8 cores are weakly or less sustainable → risk of increased footprint by using all available transistors

Overall insight: *use increase in available transistor count in a sober way and leverage reduced carbon footprint per transistor to design more sustainable processors*



(a) embodied dominated

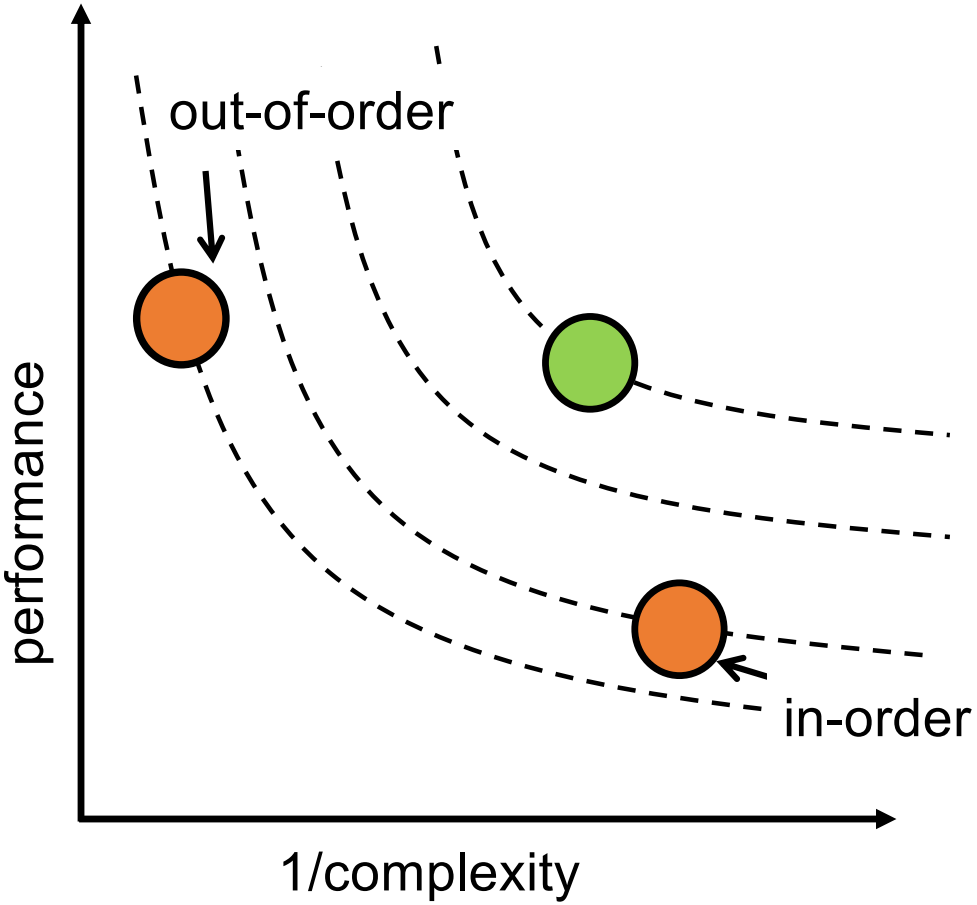
α_{E2O} = 0.8 ± 0.1



(b) operational dominated

α_{E2O} = 0.2 ± 0.1

Case Study: Instruction Selection in Superscalar Processors



Goal: out-of-order performance at in-order complexity

Prior work:

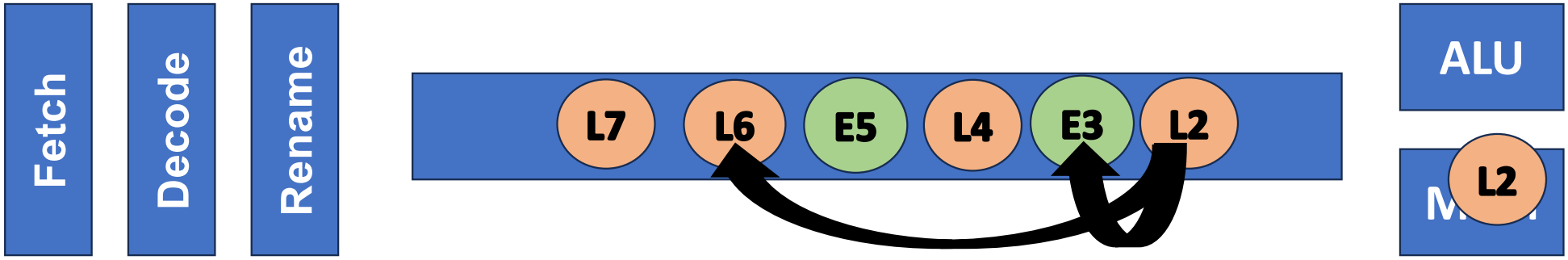
- Load Slice Core (LSC)
- Freeway
- Forward Slice Core (FSC)
- Casino
- Delay-and-Bypass (DnB)

New: FSC+, FSC++, FSC+++

Are these sustainable?

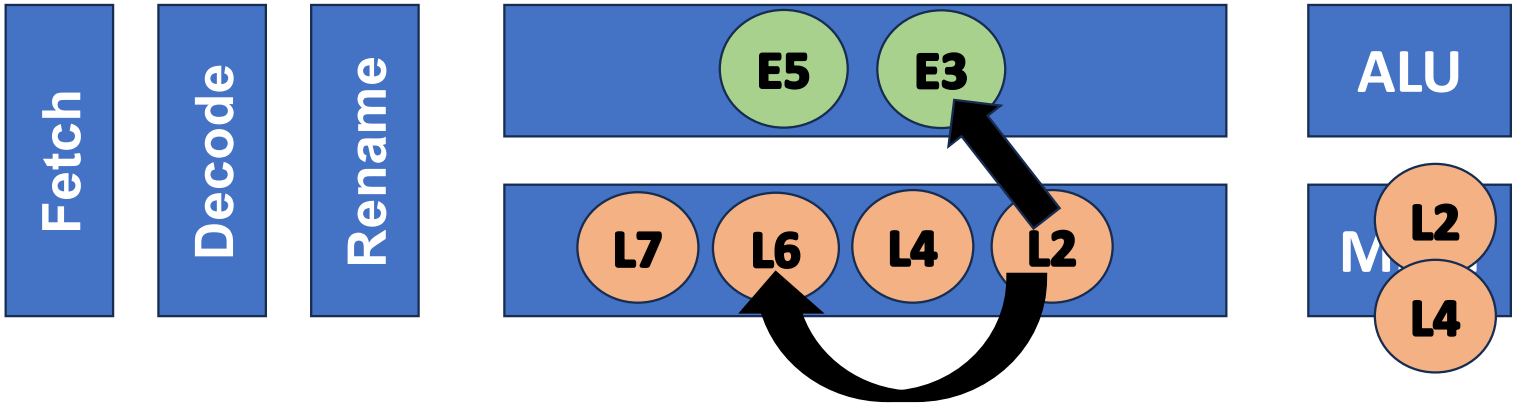
Problem with In-Order Cores

Limited memory-level parallelism (MLP) and instruction-level parallelism (ILP) due to stall-on-use in-order instruction selection



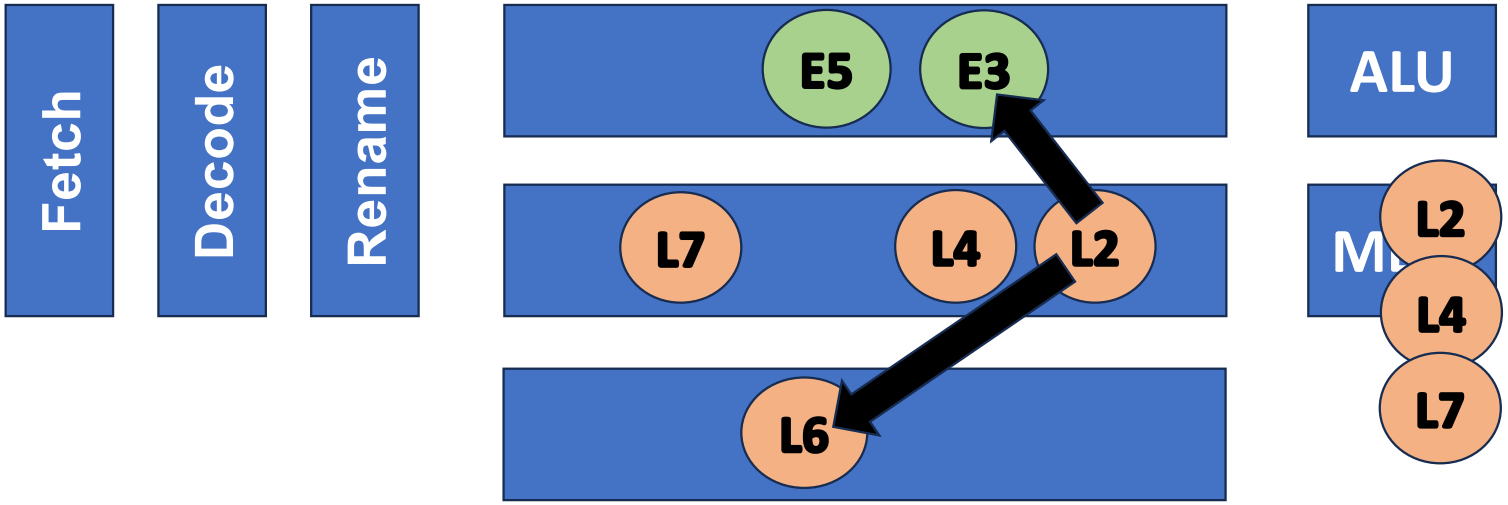
Exploiting MLP using In-Order Queues

Load Slice Core (LSC) *[ISCA'2015]* sends loads (and their address-generating instructions) to a separate in-order queue



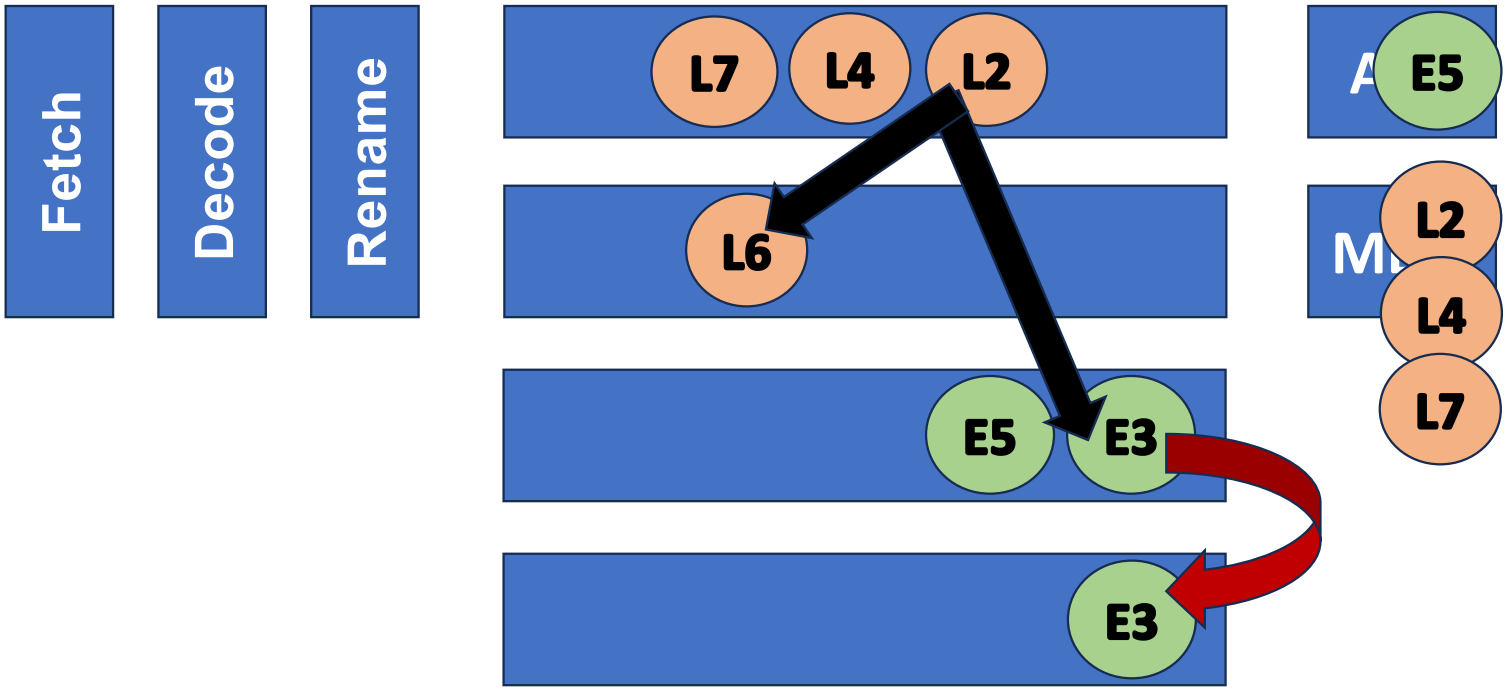
Exploiting MLP using In-Order Queues

Freeway *[HPCA'2019]* sends dependent loads to a third queue to exploit even more MLP



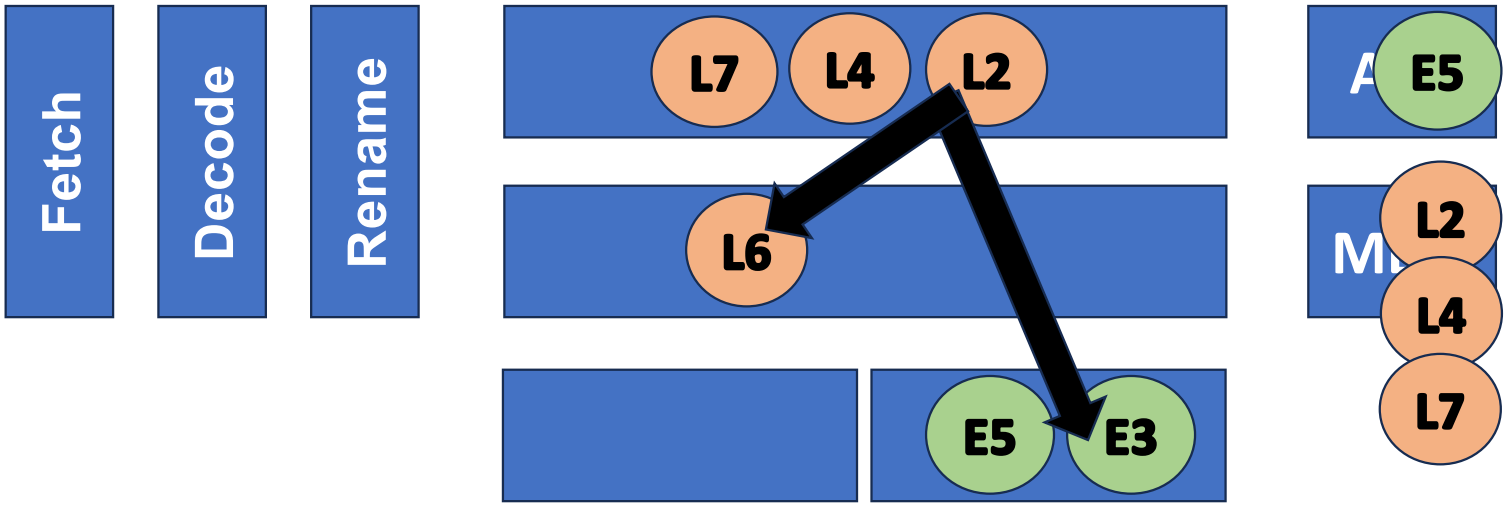
Exploiting MLP using In-Order Queues

Forward Slice Core (FSC) *[PACT'2020]* exposes high MLP ***and*** ILP



FSC++: Further Improving Efficiency

FSC++ exposes (even) higher MLP ***and*** ILP at reduced complexity



Hybrid queue: partly out-of-order, partly in-order

From PPA Analysis to Carbon Footprint Experimental Setup

Timing: Cycle-accurate FPGA simulation using Chipyard FireSim of 9 complete SPEC CPU benchmarks w/ reference inputs, trillions of instructions

Area estimate: synthesize to ASAP7 PDK 7nm FinFET standard-cell, Cadence Innovus 2021

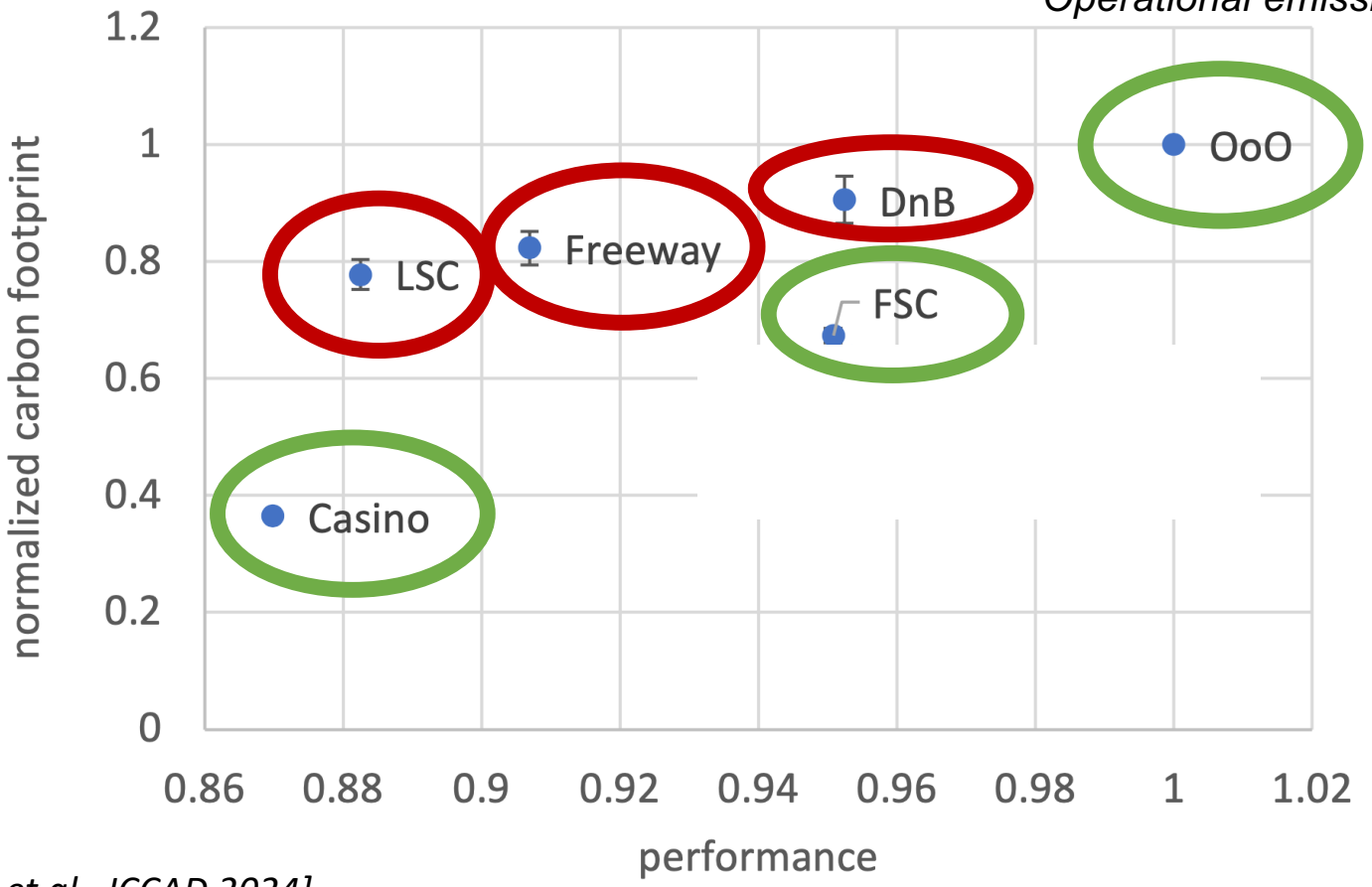
Power estimate: simulate micro-benchmarks using Verilator; activity factors as input to Cadence Voltus @ 0.7V

Simulated baseline architecture: 2-wide superscalar OoO processor @ 3.2GHz, UC Berkeley's SonicBOOM, 64-entry reorder buffer, 32-entry issue queue

Alternative instruction selection policies: total of 32 entries (max) in issue queues for fair comparison

Key Results: Casino, FSC and OoO are Pareto-optimal unlike LSC, Freeway and DnB

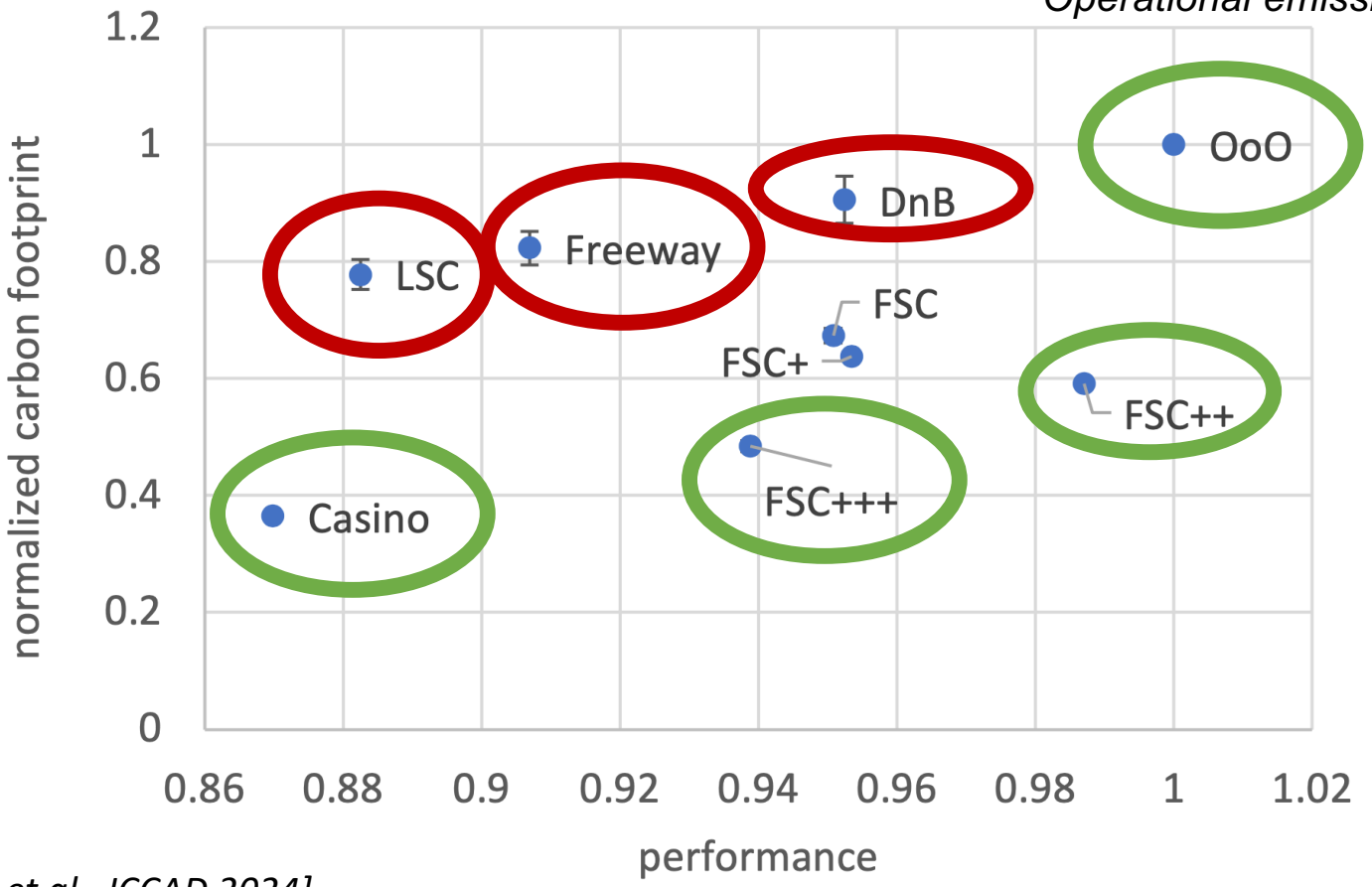
Operational emissions dominate: $\alpha = 0.2 \pm 0.1$



[S. Sheikhpour et al., ICCAD 2024]

Key Results: FSC++ reduces carbon footprint by ~40% at 1.7% performance loss

Operational emissions dominate: $\alpha = 0.2 \pm 0.1$



[S. Sheikhpour et al., ICCAD 2024]

Part IV

How to design sustainable computer systems?

Three examples: *(1) Sustainable multi-core scaling*

(2) Hardware design through PPA analysis

(3) Hardware reconfigurability

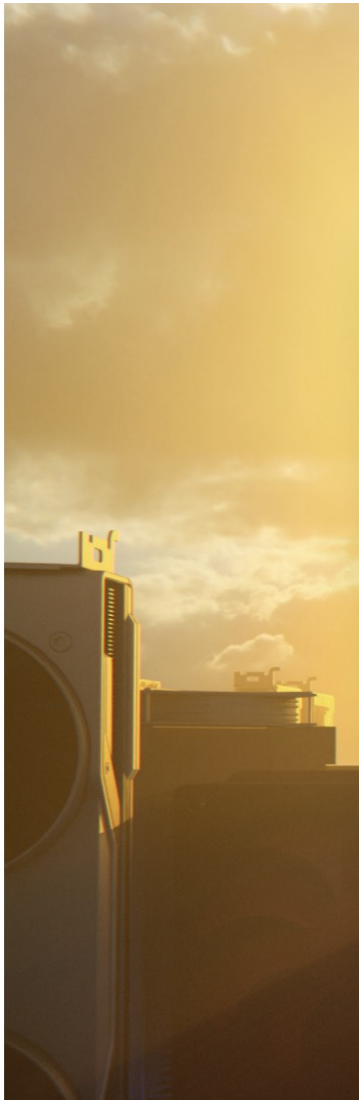
turing lecture

DOI:10.1145/3282307

Innovations like domain-specific hardware, enhanced security, open instruction sets, and agile chip development will lead the way.

BY JOHN L. HENNESSY AND DAVID A. PATTERSON

A New Golden Age for Computer Architecture



[Communications of the ACM, Feb 2019]

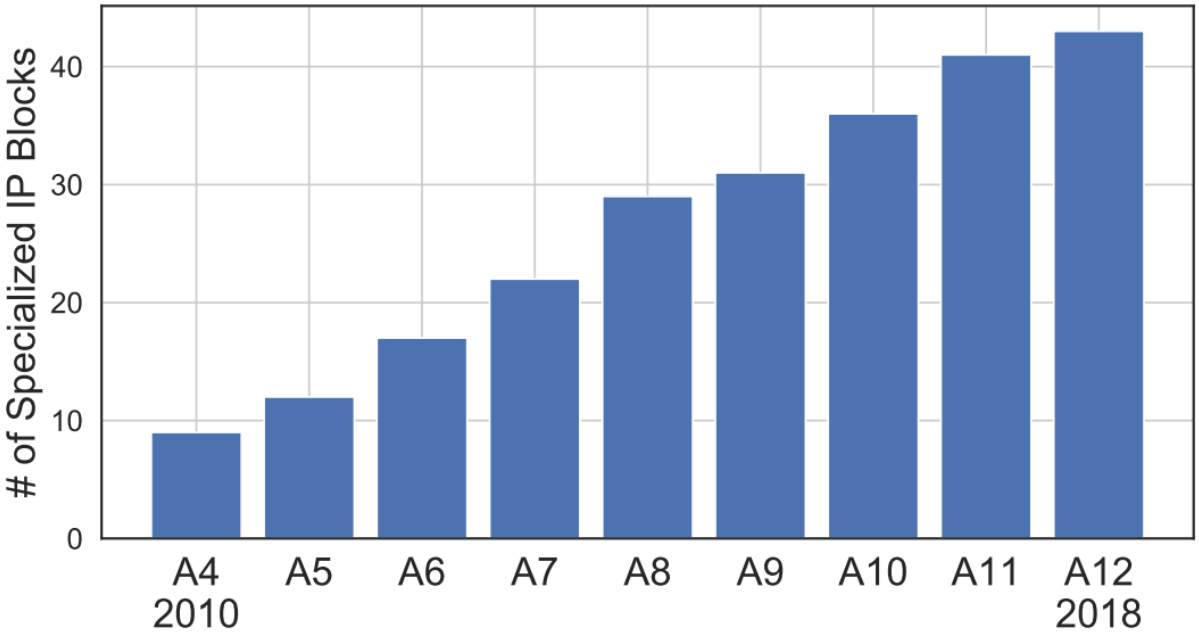
Dark Silicon: Continued Performance Scaling in Post-Dennard Scaling Era

Domain-Specific Accelerators (DSAs) powered on only when needed

Sea of DSAs is widely deployed across modern-day SoCs:

- **Mobile**: e.g., Qualcomm Snapdragon
- **Laptop**: e.g., Apple M2
- **Server**: e.g., IBM Tellum

Dark silicon fundamentally trades off chip area for power/energy efficiency



[Shao et al., ISCA@50, 2023]

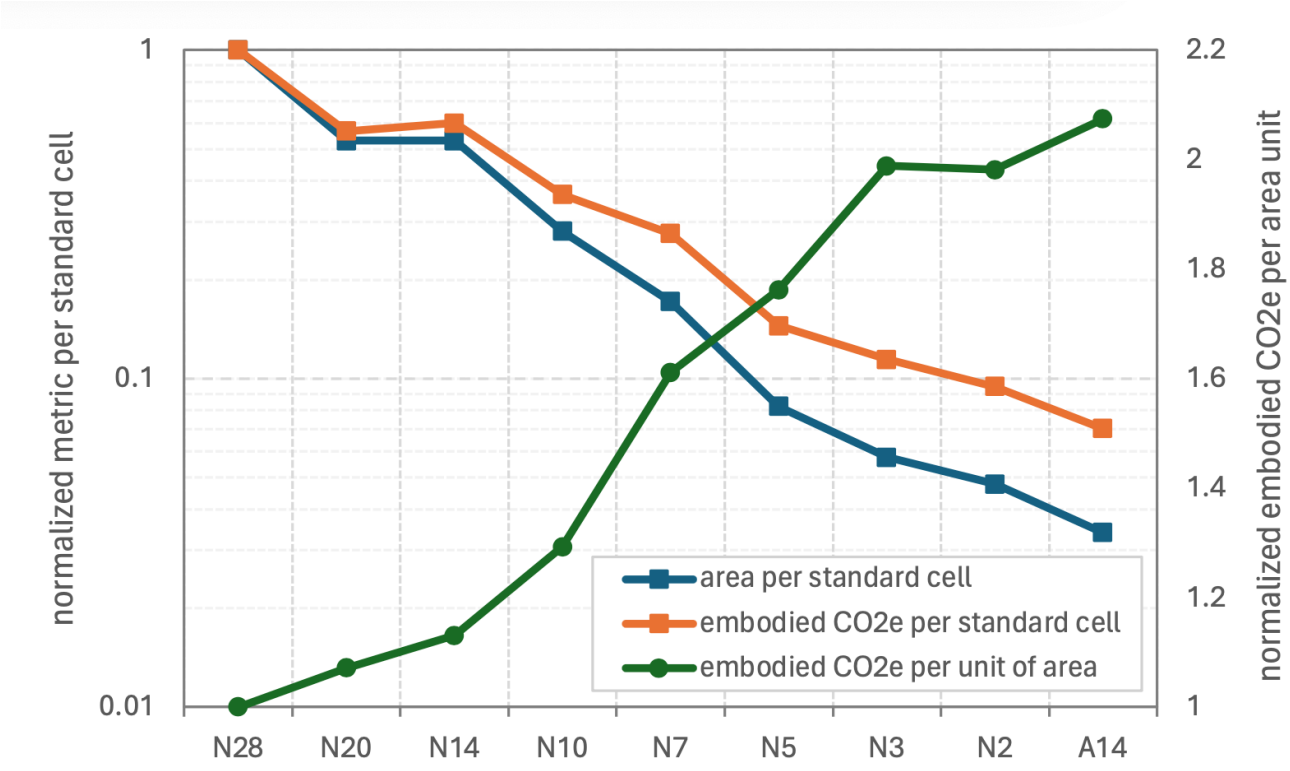
The Dirty Secret of Dark Silicon: Its Embodied Carbon (CO2e) Footprint

Embodied footprint per unit of area increases with new tech nodes

Question: Does the increase in embodied footprint due to dark silicon offset the decrease in operational footprint?

No – *"Dark silicon considered environmentally harmful"*

[Brunvand et al., IGSC, 2019]
[Eeckhout, ASPLOS, 2024]



[Boakes et al., IEDM, 2023]

Is There an Alternative? Reconfigurable Hardware to the Rescue?

Intuition: Reconfigurable hardware incurs

- **smaller embodied footprint** because smaller chip area compared to sea of DSAs, but
- it incurs **higher operational footprint** because it is less efficient

Fundamental question: ***Does the decreased embodied footprint offset the increased operational footprint? – if so, reconfigurable hardware is more sustainable than dark silicon***

[P. Dangi et al., ICCAD 2024]

Modeling Carbon Footprint of Reconfigurable Fabric vs Sea of DSAs

- Assuming serial DSA execution; reconfigurable fabric large enough for a single kernel
- Chip area is proxy for embodied footprint; energy is proxy for operational footprint*

**[Eeckhout, ASPLOS, 2024]*

Reconfigurable fabric incurs smaller environmental footprint than sea of *N* DSAs if

$$\alpha_{E2O} \times N \times A + (1-\alpha_{E2O}) \times E > 1$$

embodied footprint of *N* DSAs

operational footprint of using one DSA at a time

total (embodied + operational) normalized footprint of reconfigurable fabric

A = normalized chip of area of one DSA relative to reconfigurable fabric

E = normalized energy of one DSA relative to reconfigurable fabric

Modeling Carbon Footprint of Reconfigurable Fabric vs Sea of DSAs (bis)

- Assuming *n* DSAs execute concurrently; *n = 1..3* is typical*

**[Hill and Reddi, HPCA 2019] [Bleier et al., ISCA 2022] [Bleier et al., ISCA 2023] [Karageorgos et al., ISCA 2020]*

Reconfigurable fabric incurs smaller environmental footprint than sea of *N* DSAs if

$$\alpha_{E2O} \times N \times A + (1-\alpha_{E2O}) \times n \times E > n$$

normalized embodied footprint of *N* DSAs

normalized operational footprint of using *n* DSAs concurrently

total (embodied + operational) normalized footprint of reconfigurable fabric that replaces *n* concurrent DSAs – *a conservative estimate*

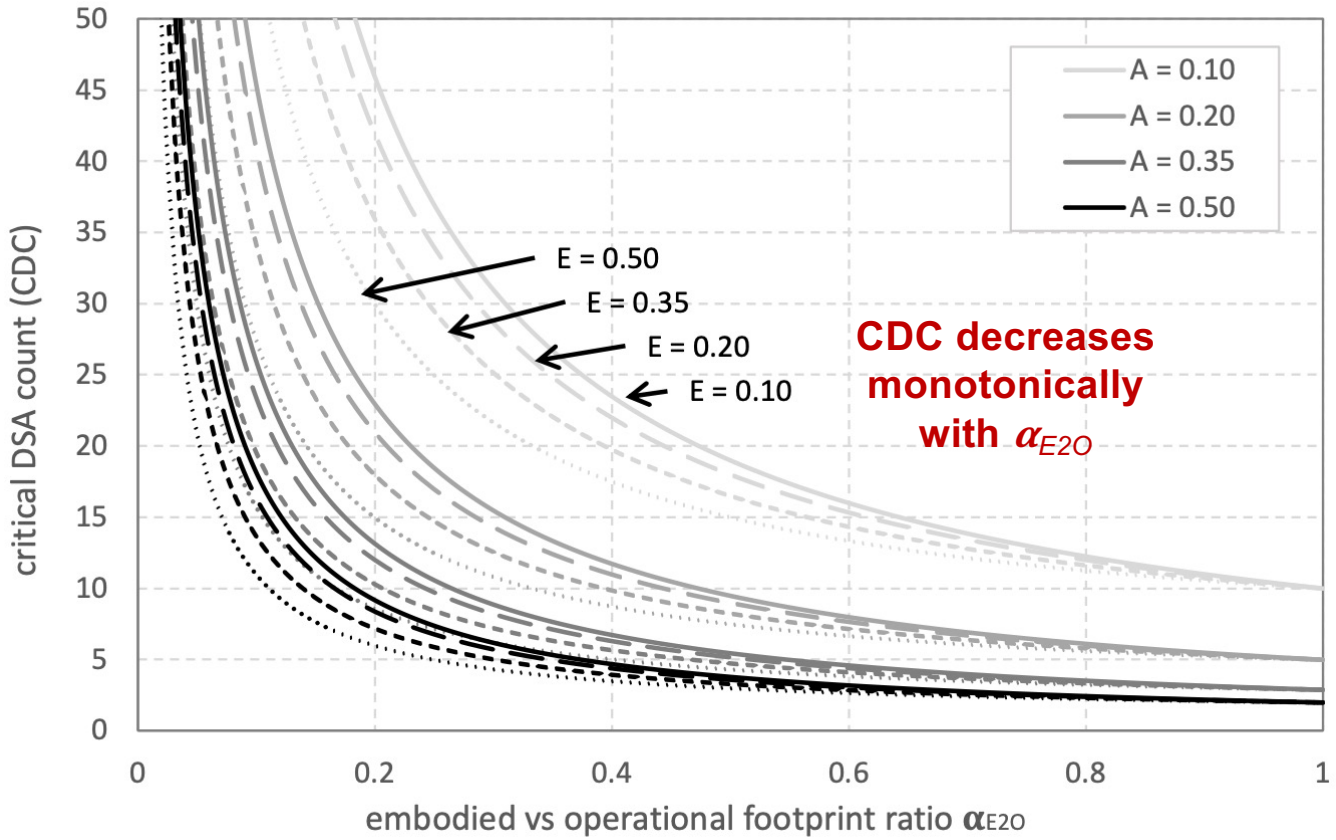
Defining Critical DSA Count (CDC)

Reconfigurable fabric incurs smaller environmental footprint than sea of *N* DSAs if *N* is larger than the Critical DSA Count (CDC)

$$\mathrm{N} > \mathrm{n} \times \left(\frac{\mathrm{E}}{\mathrm{A}} + \frac{1-\mathrm{E}}{\alpha_{\mathrm{E2O}} \times \mathrm{A}}\right) = \mathrm{CDC}$$

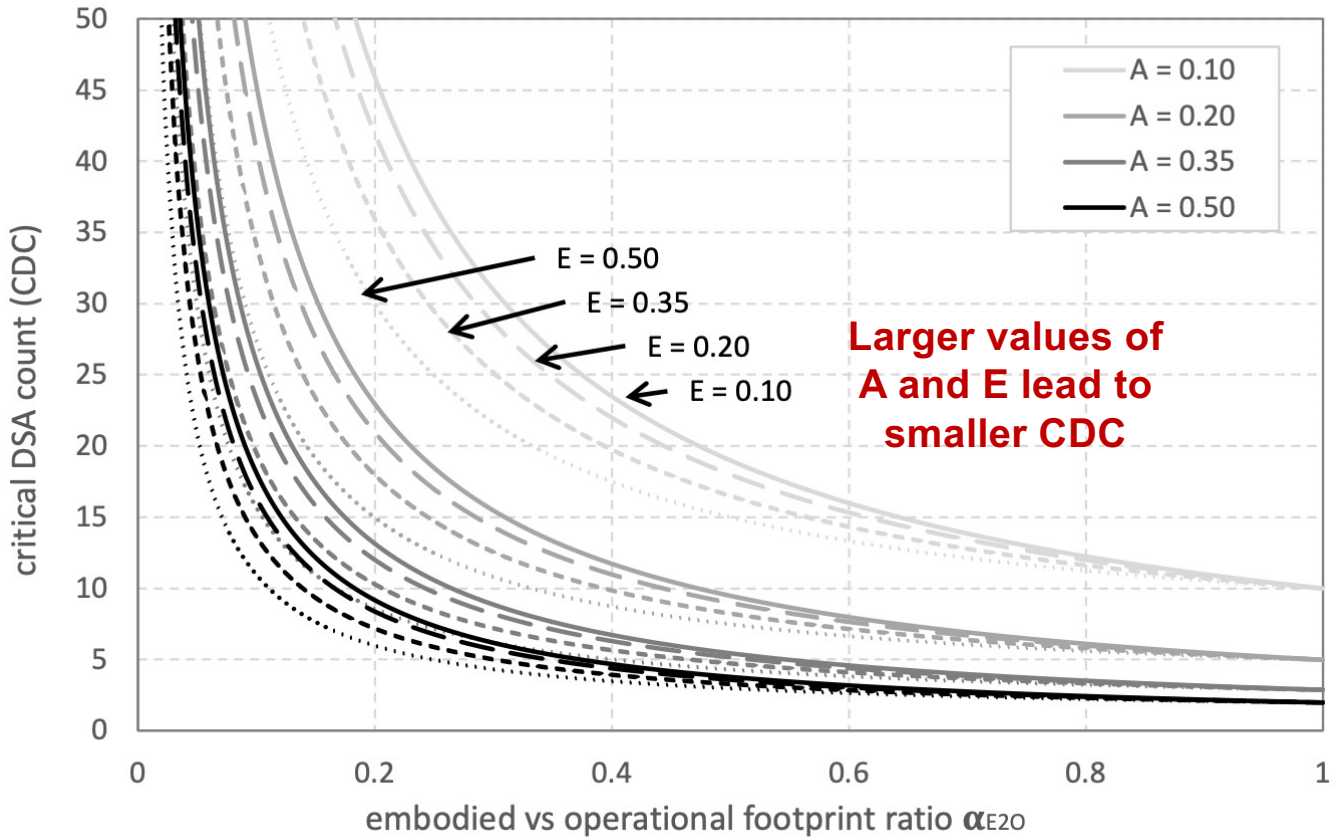
[Note: CDC is a conservative overestimation]

Reconfigurable Fabric is More Environmentally Friendly…



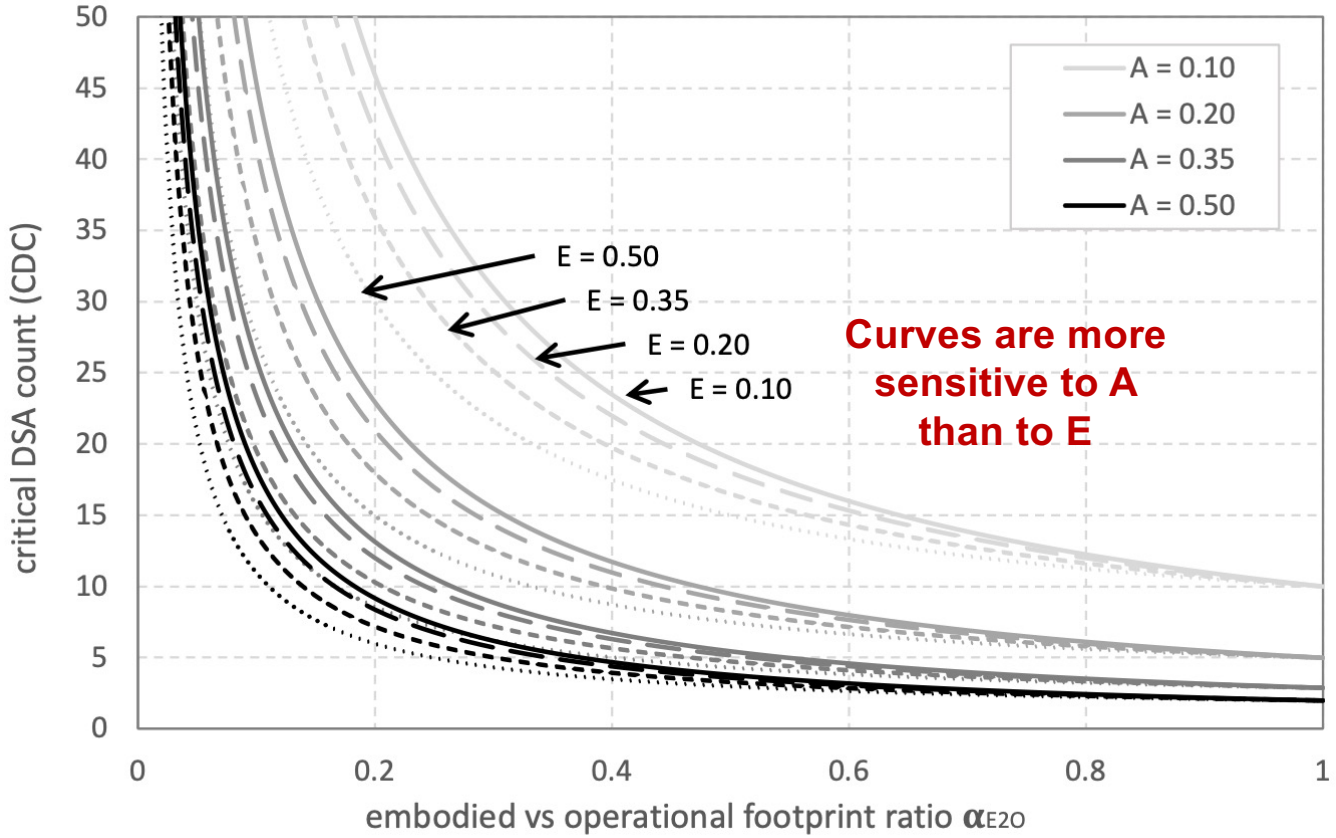
1. ***for embodied-footprint dominated systems***
   - ***Is the case for wearables, mobile devices, laptops***

Reconfigurable Fabric is More Environmentally Friendly…



1. ***for embodied-footprint dominated systems***
   - ***Is the case for wearables, mobile devices, laptops***
2. ***if area/energy reduction of DSA is small***

A Note on Sustainable DSA Design

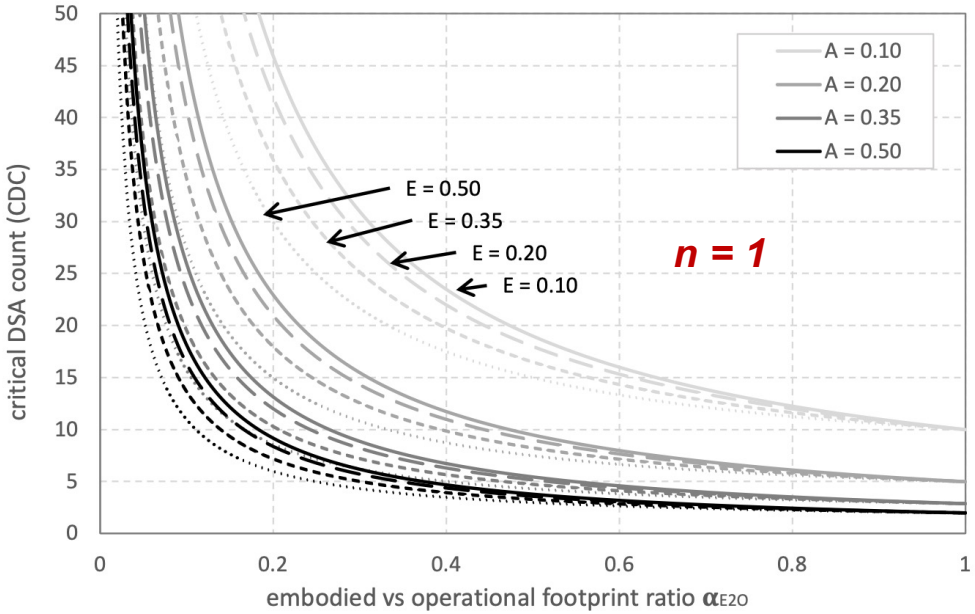


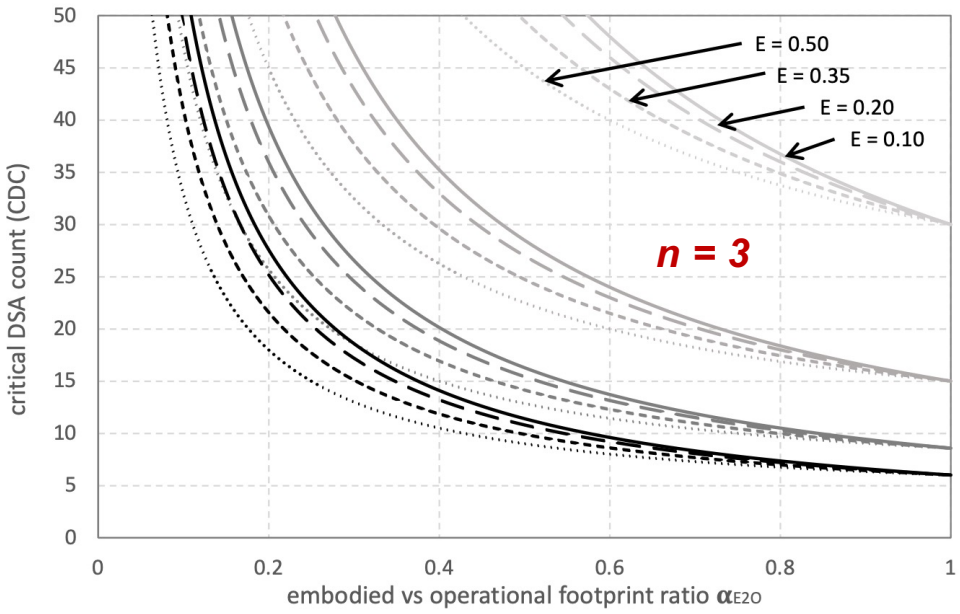
Interesting side note:

Area efficiency is more critical than energy efficiency for DSAs

- ***Contrary to common belief!***

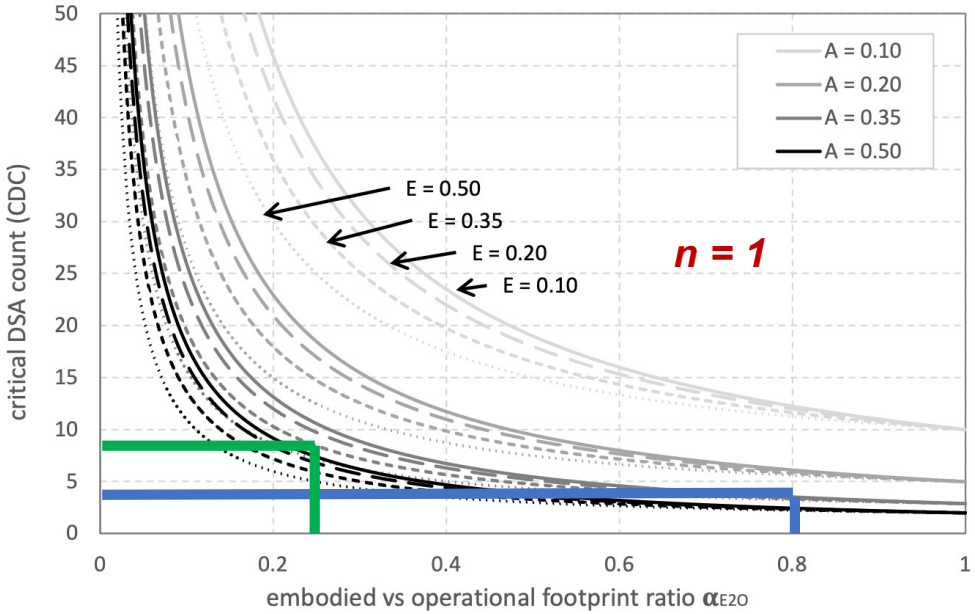
Reconfigurable Fabric is More Environmentally Friendly…

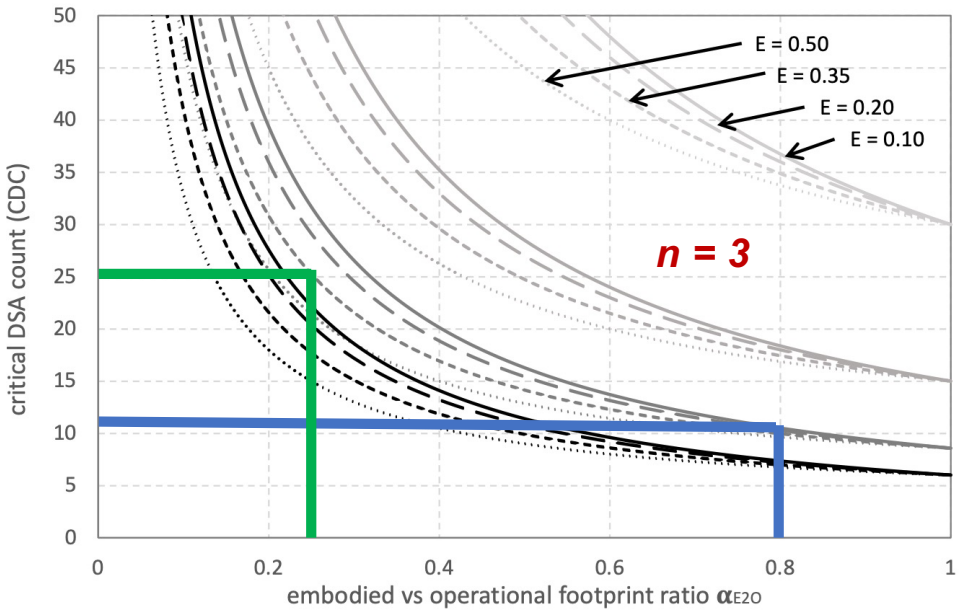




1. ***for embodied-footprint dominated systems***
2. ***if area/energy reduction of DSA is small***
3. ***at limited concurrency***

Is Reconfigurable Fabric More Environmentally Friendly?





for embodied-footprint dominated systems → if it replaces ~4 to ~12 DSAs

for operational-footprint dominated systems → if it replaces ~8 to ~25 DSAs

[assuming A = E = 0.35]

This is (way) fewer than the number of (~40) DSAs in modern-day SoCs

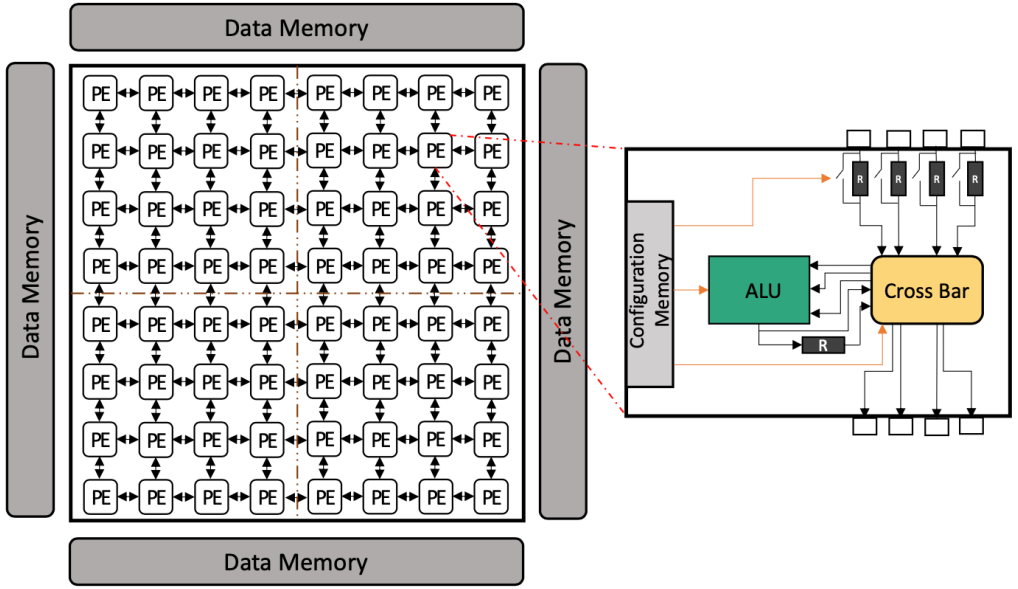
Is Reconfigurable Fabric More Environmentally Friendly? Let's See…

Reconfigurable fabric and mapping: Coarse-Grain Reconfigurable Area (CGRA) & Morpher *[NUS]*

DSA synthesis: Aladdin *[Shao et al., ISCA 2015]*

Workloads:

App. Kernel	Domain	Description	Memory (in KB)
GeMM	Machine Learning	General Matrix Multiplication 32x32 tile size, 96x96 input size	108
KNN		K-nearest neighbour 16 maximum neighbours	22
Conv2D		2D convolution Filter size of 3x3, input size 96x96	72
Stencil3D	Image Processing	3D stencil calculation with data size 16x32x32	256
Viterbi	Speech Recognition	Viterbi algorithm 64 hidden states 32 observations	52
FFT	Signal Processing	128-pt fast Fourier transform	1.5
FIR		32-tap FIR filter	108
AES Encryption	Security	Rijndael ciphers with 16B block size	0.5



Analysis based on iso-performance comparison between CGRA versus DSA

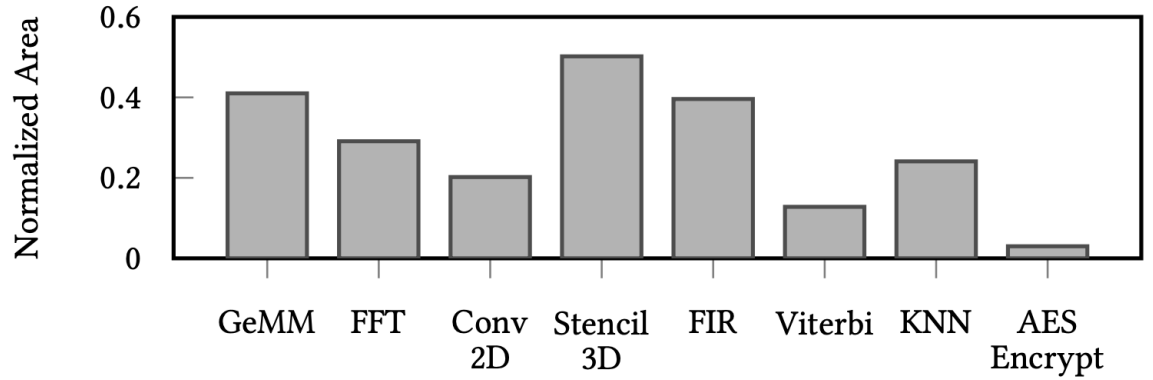
Area/Energy Efficiency of DSA vs CGRA

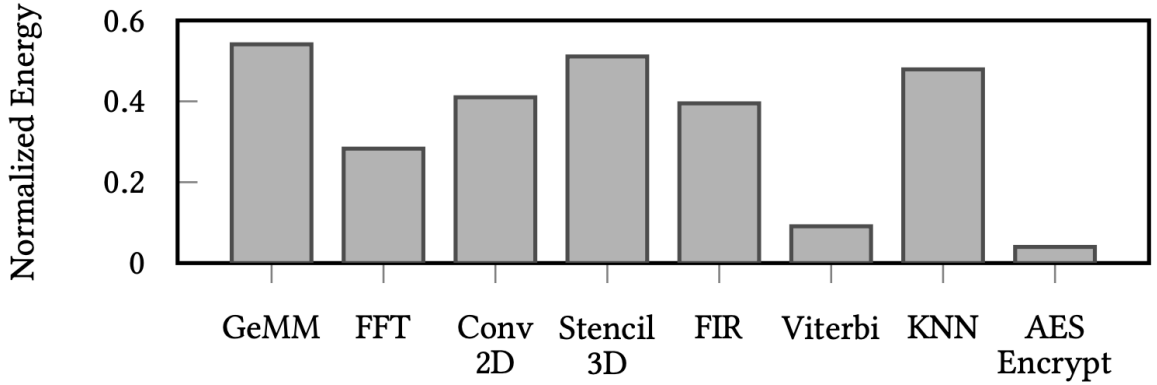
***A* varies b/w 0.03× and 0.55×, 0.27× on average**

***E* varies b/w 0.03× and 0.49×, 0.31× on average**

Viterbi and AES bit-level intensive

Stencil3D, FIR, GeMM are multiply-accumulate intensive

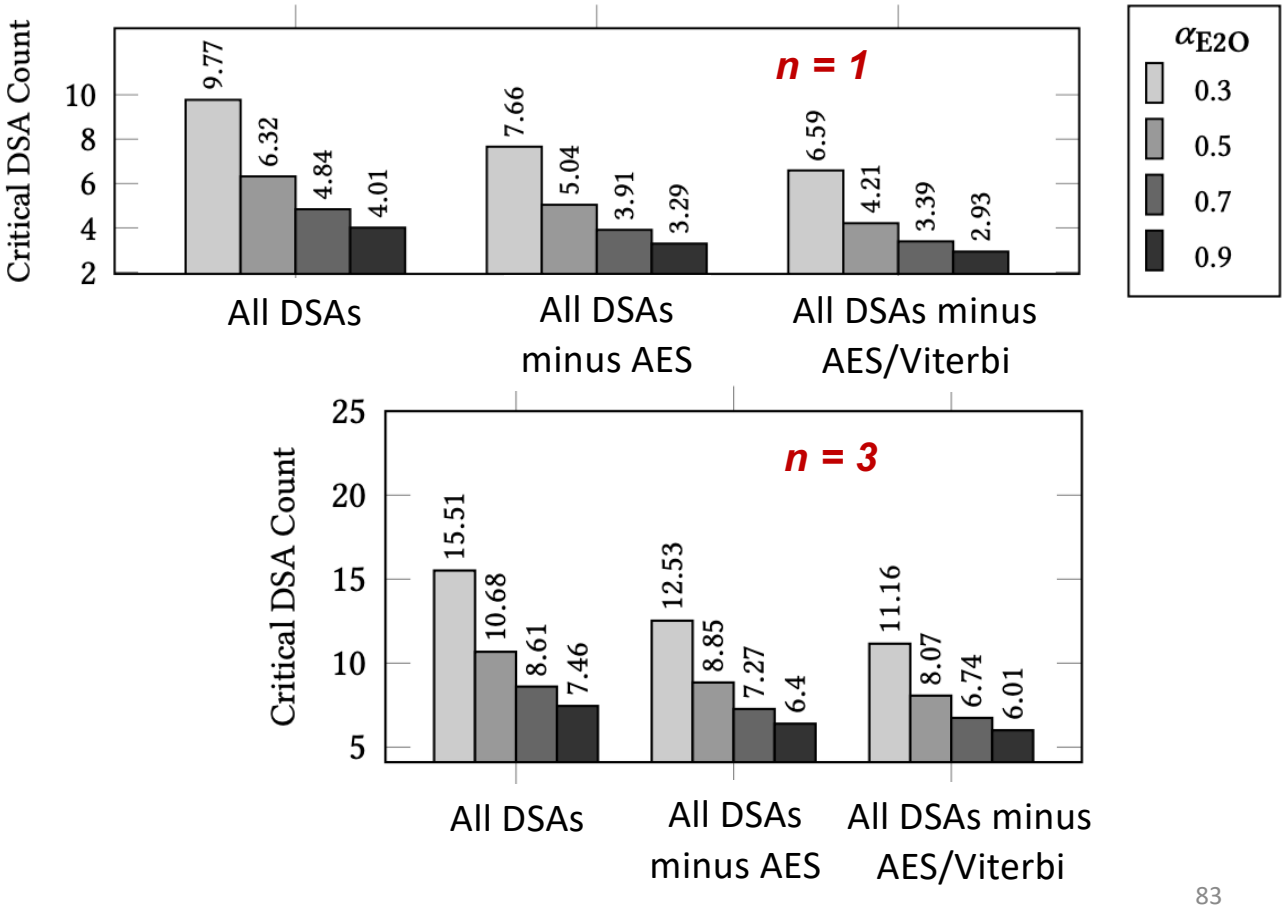




CGRA is More Environmentally Friendly than Sea of DSAs

CDC decreases with increasing embodied-footprint dominance

CDC increases with DSA concurrency



Excluding most area/energy-efficient DSAs from replacement by CGRA further decreases CDC

Replacing a handful to a dozen DSAs is worthwhile

Environmental Footprint Savings Compared to Sea of 40 DSAs…

vary between 2.5x and 7.6x under most probable scenario

vary between 1.6x and 7.6x under worst-case scenario

Concurrency (n)	Carbon Footprint Improvement	
	Avg Util ($n' < n$)	100% Util ($n' = n$)
1	-	7.60×
2	6.10×	3.84×
3	4.12×	2.59×
4	3.12×	1.97×
5	2.53×	1.59×

CGRA is *less than n* times larger if concurrency equals *n*

CGRA needs to be *n* times larger if concurrency equals *n*

Intuition: CGRA is not fully utilized by each kernel, e.g., 100% for GeMM/FIR but (much) less for others

Summary

ICT's contribution to global warming is significant, and rising

Assessing computer architecture sustainability is challenging

- Multi-faceted problem, inherent data uncertainty, need to take whole lifecycle into account

Total carbon emissions continue to grow under current scaling trends

Embodied emissions are, or will soon be, most dominant contributor to the total carbon footprint

Computer architects can (and should) **reduce total carbon footprint by reducing die size** (primarily) ***and*** **operational emissions** (secondarily)

FOCAL: First-order model using proxies for embodied/operational footprint and parameterized embodied/operational ratio to holistically reason about sustainability

- Deliberately simple, yet accounts for Jevons' paradox
- Provides insight and intuition
- Framework to reason about computer architecture sustainability trade-offs for a variety of scenarios
  - Multicore, heterogeneity, caching, speculation, specialization, parallelization, etc.

Exciting and important work ahead of us: call for action for computer scientists and engineers ☺

Sustainable Computer System Design

Lieven Eeckhout

Ghent University, Belgium

[L. Eeckhout, "The Sustainability Gap for Computing: Qua Vadis?", Communications of the ACM, to appear]
[P. Dangi et al., "Sustainable Hardware Specialization", ICCAD 2024]
[S. Sheikhpour et al., "Sustainable High-Performance Instruction Selection for Superscalar Processors", ICCAD 2024]
[L. Eeckhout, "FOCAL: A First-Order Model to Assess Processor Sustainability", ASPLOS 2024]
[L. Eeckhout, "A First-Order Model to Assess Computer Architecture Sustainability", (Best of) IEEE CAL, 2022]
[L. Eeckhout, "Kaya for Architects: Towards Sustainable Computer Systems", IEEE Micro, 2023]

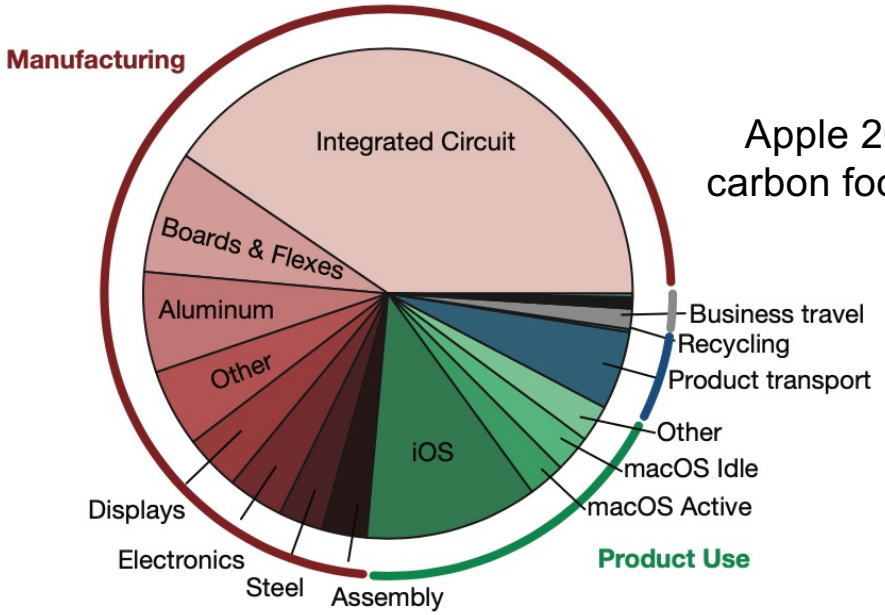
Carbon Footprint of ICT – Impact of IC Manufacturing

Integrated circuits:
manufacturing ICs ~30% of total footprint >>
~15% for product use *[Apple 2019]*

getting worse with technology: +12% CAGR

[L. Boakes et al., IEDM, 2023]

How can computer scientists and engineers reduce the environmental footprint of electronic devices?



Apple 2019
carbon footprint

[Gupta et al., HPCA 2021]