Bending the ICT curve: Evaluating options to achieve 2030 sector-wide climate goals & projecting new technology impacts

By

Allison Bell

B.S. Environmental Science
University of California, Los Angeles, 2018

Submitted to the Institute for Data, Systems, and Society in partial fulfillment of the requirements for the Degree of

MASTER OF SCIENCE IN TECHNOLOGY AND POLICY

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2023

Authored by: Allison Bell
Institute for Data, Systems, and Society
May 12, 2023

Certified by: Elsa A. Olivetti
Professor, Materials Science and Engineering
Thesis supervisor

Accepted by: Noelle Eckley Selin
Professor, Institute for Data, Systems, and Society and
Department of Earth, Atmospheric, and Planetary Sciences
Director, Technology and Policy Program

Bending the ICT curve: Evaluating options to achieve 2030 sector-wide climate goals & projecting new technology impacts

By

Allison Bell

Submitted to the Institute for Data, Systems, and Society in partial fulfillment of the requirements for the Degree of Master of Science in Technology and Policy

ABSTRACT

The global impact of the information and communications technology (ICT) sector is both growing and changing as computing technologies continue to develop and industry leaders make more efforts towards emissions reductions. Recent work highlights the increasing importance of manufacturing emissions in regards to the total impact of computing systems, but the tradeoff space in which decisions made to reduce emissions or energy in one part of a device lifecycle might increase emissions or energy demand in another remains largely unexplored. We evaluate several options for global impact reduction within the ICT sector, namely within data center (server) and smartphone footprints, focusing both on the maximal potential impact of each intervention and highlighting associated tradeoffs and limitations. We find that the ICT sector’s 2030 target of a 45% emissions reduction from 2020 levels is potentially achievable through the mechanisms proposed, including: renewable energy for operation, low-carbon electricity for manufacturing, extended device lifetimes, and the harnessing of energy efficiency improvements for impact reduction. In addition, we propose a method for evaluating the total carbon footprint benefits of a new computing technology through a detailed case study of a prototypical analog accelerator device. We provide an example of underspecified estimation of scaled device manufacturing impacts obtained through a reorganization of existing process emissions data. We then demonstrate the use of that estimate to evaluate the benefits of adoption of the new technology from the perspective of total footprint reduction under varying device usage conditions. Both our framework for estimating global ICT sector impact reduction strategies and our framework for assessing tradeoffs associated with new computing technology adoption are intended to serve as starting points for continued discussion and to align different, often siloed, stakeholders within the computing industry towards effectively “bending the curve” of ICT sector emissions growth.

Thesis supervisor: Elsa A. Olivetti

Title: Professor, Materials Science and Engineering

Acknowledgements

I am incredibly grateful for all of the guidance and feedback that I received around this thesis work. First and foremost, this analysis would not have come together in this way without the support and insights of my advisor Elsa Olivetti. Her constant excitement around this topic provided such a positive motivation to dig deeper and contribute as much as I could to furthering this work in the short time that a master's degree research timeline allows and I am beyond thankful for her mentorship. I also want to thank MIT professors Jesus del Alamo and Vivienne Sze, as well as postdoctoral researcher Murat Onen and graduate student Tanner Andrulis, for their insightful feedback and expert inputs as we sought to undergo research that would prove relevant across different engineering disciplines. The analysis in this thesis was further informed by a number of conversations with individuals outside of MIT. Carole-Jean Wu in particular shared in our excitement over the potential of this work and was incredibly giving in terms of her time and ideas. I sincerely thank her and everyone else who took the time to discuss this work with me, including Joe Palazzo, Sarah Boyd, Patrick Wall, and Monique McIntosh. The conversations we had made the process of writing this thesis much more interesting and engaging and the ideas we took away from those conversations greatly improved this work.

Pursuing a graduate degree is not easy, and I owe so much to the support of my loved ones throughout this process. My mom and dad have an incredible (although at times frustrating) confidence in my ability to do absolutely anything that I decide to and I wouldn't be where I am without them cheering me on. My grandparents, brother, aunts, uncles, and cousins are right there behind them, sporting their MIT memorabilia with pride. My grandmother Mary, who passed away last June, always wanted to know what I was working on and for me to share breakdowns of my research alongside pictures of East Coast trees and flowers. Beyond my family, it is difficult to put words to my gratitude for the love and support of my partner Aaron. He makes me smile every day and has given so much of himself these past two years in support of my graduate pursuits. I am so thankful to have him in my life. Finally, to my friends, both old and new, thank you for the video calls, group chats, and afternoon runs that have kept me sane and happy.

Finally, I want to thank the TPP program team, including Barbara DeLaBarre, Frank Field, Elena Byrne, and Noelle Selin. Together, they have done an incredible job at cultivating an inclusive and supporting environment for all of us TPPers that has made this experience all the better.

Table of Contents

List of Figures

List of Tables

Introduction

Projections of information and communication technology (ICT) growth indicate that the sector could account for as much as 20% of global electricity demand by 2030, with networks and data centers highlighted as significant contributors to this growth.[1] In addition, recent research suggests embodied emissions (emissions arising from the manufacturing of a computing device) are increasingly important compared to use phase emissions across both consumer device and data center lifecycles, especially as computing technologies are optimized for energy efficiency and run on cleaner energy sources.[2] While the importance of embodied carbon in computing devices is gaining attention, the tradeoff space in which decisions made to reduce emissions or energy in one part of a device lifecycle might increase emissions or energy demand in another remains largely unexplored.

Existing literature provides “business as usual” estimates of current and future energy and emissions impacts of the ICT sector. Leading examples of assessments of global ICT impact in recent work include Belkhir and Elmeligi (2018)[3], Malmodin and Lundén (2018)[4], and Andrae and Edler (2015)[5] which was subsequently updated by Andrae (2020)[6]. Belkhir and Elmeligi (2018) estimate the global ICT footprint, including operational and embodied emissions, stemming from data centers, consumer devices, and networks and derive ICT impact predictions to 2040 from models fitted to total aggregated global ICT footprint across previous years. Because of this, they do not explicitly include assumptions of growth in computing efficiency or greening of global electrical grids into their future estimates. While the authors define several recommendations to reduce ICT impacts, including running data centers on 100% renewable energy, manufacturing smartphones with renewable energy, and extending device lifetimes, the potential impacts of these recommendations are not quantitatively defined within this work. Malmodin and Lundén (2018) similarly include the embodied and use phase impacts of data centers, consumer devices, and networks in their assessment, but only provide predictions to 2020. Unlike Belkhir and Elmeligi (2018) their study predicts a leveling off or decline in sector emissions in future years. This study does not provide recommendations on ways to reduce ICT sector impacts, perhaps as a result of that predicted decline. Andrae and Edler (2015) focus on operational and production-based electricity usage (and the emissions resulting from that electricity usage) of data centers, consumer devices, and networks and explicitly incorporate assumptions of increasing data center efficiency each year and global electricity emissions intensity into predictions of sector impacts to 2030. The authors mention the potential to power data centers with renewable energy in order to reduce future emissions, a perspective that is reiterated in Andrae (2020) with the caveat that there are geospatial limitations on the availability of renewable energy that may limit adoption. However, neither Andrae and Edler (2015) nor Andrae (2020) provide further quantitative exploration of the potential for future impact reduction. While these three bodies of work provide estimates of future ICT sector impacts under “business as usual” scenarios, they do not attempt to quantitatively demonstrate the impact of interventions to reduce future impacts, even when interventions like renewable energy development and device lifetime extension are discussed.

Separately from the literature focused on estimating global ICT impacts, there are examples of literature that does attempt to quantify the impacts of strategies to reduce emissions from computing devices.

Acun et al. (2023)[7] provide a method for minimizing both the operational and embodied footprint of data center systems through renewable energy deployment, in combination with battery storage and additional server capacity development to enable flexibility in data center energy use. They apply this method to a case study of Meta's datacenters and highlight differences in total footprint reduction across different US locations. Gupta et al. (2020)[2] offers a quantification of the potential emissions benefits that could come from powering wafer fabrication facilities with renewable energy in terms of a percentage decrease in emissions from a wafer fabrication facility. While both Acun et al. (2023) and Gupta et al. (2020) offer methods for quantifying the impacts of renewable energy interventions on ICT sector emissions, they do not attempt to estimate the potential global impacts of these strategies. Bashroush (2018) provides a method to evaluate the total footprint benefits of refreshing server hardware including embodied and operational emissions trade-offs and finds results generally in favor of frequent hardware updates. However, the author does not evaluate how future renewable energy incorporation or grid greening could alter trade-offs in upgrading systems. Jattke et al. (2020)[8] quantifies the environmental implications of service life extension of mobile devices by modeling the total footprint impact of a repair scenario on a single smartphone device, but does not extend that analysis beyond the single device level. Finally, Freitag et al. (2021)[9] proposes a global carbon constraint as a way to control the rebound effect of efficiency increases, and models a scenario in which ICT emissions are held at 2020 levels through 2050, but does not directly compare the benefits of that alternative scenario as compared to a baseline 2050 prediction of ICT emissions or evaluate the benefits of that intervention in comparison to others. While methods exist to quantitatively evaluate the impacts of various interventions to reduce future emissions of the ICT sector, including renewable energy for operation and fabrication of computing devices, extending computing device lifetimes, and limiting sector growth to efficiency gains, we do not find evidence of these interventions being modeled simultaneously at a global scale within previous work.

Novel computing technologies may also prove a tool for future ICT sector impact reduction, but we need a way to evaluate any embodied versus operational emissions tradeoffs associated with their adoption. Existing literature to estimate the embodied footprint of computing devices, specifically integrated circuits, focuses on current and future impacts of iterations upon existing computing technologies. Boyd (2012)[10], for example, provides an in depth life cycle assessment of computational logic and memory technologies manufactured between 1995 and 2010. Bardon et al. (2020)[11] provides updated estimates of embodied energy and emissions for more recent logic technology nodes, and predicts the impacts of future nodes based on Design Technology CoOptimization methodologies. However, Bardon et al. (2020) only projects future computing technology impacts according to Moore's law evolution of existing technologies, leaving a gap in terms of predictions of the embodied impacts of future computing technologies that do not fall into that existing technology evolution paradigm. For example, the impacts of computing device specialization, coined as "More than Moore" technology evolution by the ITRS [12], on the embodied carbon of computing devices is notably unexplored to date.[13] While there is an emerging body of literature around prospective life cycle analysis techniques which focus on estimating the embodied and operational impacts of emerging technologies[14], we do not find evidence of prospective LCA methods applied towards future integrated circuit technologies. A method for estimating the future, scaled impacts of prototypical computing technologies under development today is needed in order to ensure that new technology adoption is aligned with larger ICT sector emissions reduction efforts.

Given the rich body of existing literature surrounding these topics, we keep this initial review brief, but contextualize our approach and results within relevant literature throughout this thesis document. We seek to address these gaps within existing literature through a two-pronged approach. First, we perform a high-level sectoral accounting of the primary levers for carbon emissions reduction within the lifecycle of computing devices that highlights both where embodied versus operational emissions trade-offs might matter and where the greatest opportunities for overall impact reductions lie. Secondly, we provide an in-depth exploration of one of the levers for impact reduction noted in the first section - adoption of emerging energy efficient computing technologies. We conduct a detailed case study exploration where we estimate the embodied impact of a prototypical analog accelerator device through a manufacturing process model, and use that estimate to demonstrate the circumstances under which operational impact benefits can outweigh the embodied impacts of manufacturing that specialized hardware.

Using this approach, we identify opportunities and limits to achieving 2030 ICT sector global climate goals. In addition, we evaluate the operational versus embodied carbon tradeoffs associated with interventions aimed at reducing overall emissions and new technology adoption. We hope that this work can begin to provide a common framework for decision makers at all levels of computing device development and operation to evaluate their decisions in terms of total lifecycle impact benefits and encourage further data collection, sharing, and model refinement across different components of the computing sector.

Reducing ICT sector emissions: Opportunities to reduce 2030 operational and embodied impacts, along with their challenges and limitations

We begin with a high-level sectoral impact accounting of a few of the most promising strategies for reducing the projected global impact of the ICT sector in 2030. Our analysis includes both options to reduce the operational footprint of computing systems (that can also impact embodied carbon) such as investing in renewable energy systems to provide low carbon operational energy or increasing the energy efficiency of a system through new technology adoption, and options to reduce the embodied carbon footprint of computing systems such as through manufacturing devices using low-carbon energy, increasing device lifetimes, or other “green fabrication” strategies. In addition, we consider “new learning” opportunities in which artificial intelligence could reduce overall growth in data center energy demand. However, neither “green fabrication” interventions nor “new learning” impacts are modeled explicitly within this work, unlike the other interventions, due to a lack of data availability for modeling.

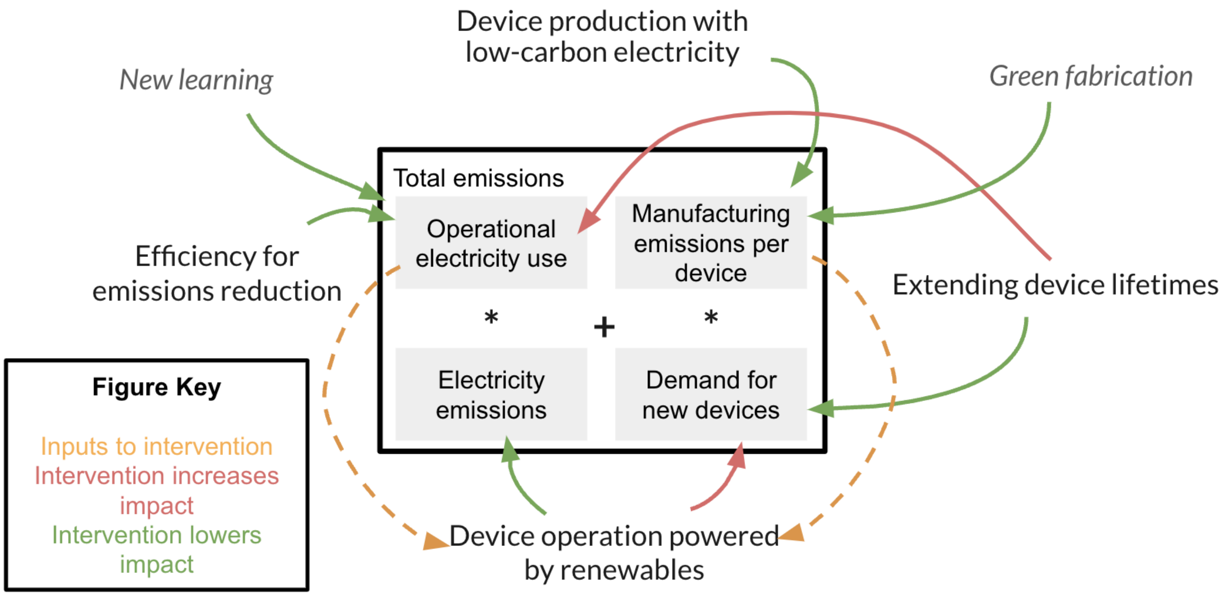

Figure 1. Mapping the relationships between different interventions to reduce the impact of the ICT sector.

In this analysis, we focus on the impacts of smartphones (as a subcategory of user devices) and servers (as the main ICT component of data centers) as two significant contributors to ICT sector impact with differing challenges and opportunities for impact reduction. Note that this excludes other important contributors to ICT sector impact, including networks and other connected devices, for the purpose of limiting the scope of this study. Future work could replicate the trend exploration in this study for other ICT segments that we now omit. The impact of networks in particular may be of interest, given that the scale of data transmission network energy usage in 2021 rivaled that of data centers.[15] In order to more effectively estimate the potential for different interventions to decrease the overall footprint of these two ICT device segments, we decided to build our own estimates of embodied and operational impact for these segments in order to 1) have estimates of current and future impacts using the most up-to-date data available, 2) be exceedingly transparent around the assumptions going into each estimate, and 3) enable a zeroing in on those elements relevant to the lever for impact reduction that we wish to explore. While we offer an analysis of the impact of these different levels on a global level, the methodology we propose for estimating the limitations and opportunities behind different sectoral impact reduction levers can be repurposed and built on for prioritizing and evaluating interventions within the supply chain of a given individual firm or product of interest.

Global footprint estimates, 2021

We begin by estimating the baseline footprint of global data center (server) and smartphone production based on the most recent year that existing data was available for across the majority of our data sources, 2021.

Estimating operational footprints, 2021

As an approximation of global data center operational energy footprints, we utilize a recent estimate from the IEA, which reports non-crypto data center energy usage at 220-320 TWh in 2021.[15] For the purposes of this assessment, we consider all data center operational energy as relevant to server operation since this value includes auxiliary services like cooling and other facility operations that are required for server operation. However, for smartphones, only the energy use of the actual smartphone is considered.

In order to build an estimate of smartphone operational energy usage in 2021, we combine a range of estimates of the total number of smartphones subscriptions worldwide in 2021 (2.9[16]-6.2[17] billion smartphones) with average yearly energy consumption estimates reported by Google for Pixel 3-6 generation devices released between 2018 and 2021[18] (~7.5 kWh per year expected per device). With these two estimates, we estimate global smartphone yearly operational energy consumption at ~36 TWh.

In order to estimate the carbon emissions footprint of that operational energy usage, we use an estimate of average global carbon intensity of electricity, similar to the method employed by previous sector wide carbon footprint estimates.[4] Using a 2021 estimate of 459 gCO^2e/kWh average carbon intensity reported by the IEA,[19] we estimate the operational carbon footprints of data center and smartphone ICT components to be ~124 Mt CO^2e and ~16 Mt CO^2e, respectively.

We acknowledge that the ICT sector has made significant efforts towards the purchase and use of renewable energy,[15] in particular to power data centers, so our estimate provides an upper bound on emissions for data centers rather than a more realistic estimate that takes those investments into account. For example, in 2021, Google reported an average carbon intensity of 0.1006 tCO^2e/MWh (100.6 gCO^2e/kWh) across their operations,[20] nearly a quarter of the emissions intensity that we consider using the global average.

Estimating embodied footprints, 2021

In order to generate a baseline estimate of the embodied footprint for smartphone and server devices, we combine estimates of server and smartphone shipments in recent years with published product carbon footprint data. IDC statistics indicate that 1660 million new smartphones shipped in 2021[21] and 9.53 million new servers shipped in 2020[22]. In order to estimate a "typical" production footprint per device, we average across reported manufacturing emissions of 48 server products included within a dataset of product carbon footprints compiled by Boavizta, including servers from HP and Dell, to obtain a general server manufacturing footprint estimate of ~1230 kg CO^2e per device.[18] For smartphones, we similarly average across reported manufacturing footprints of multiple device generations and manufacturers in the Boavizta dataset[18] including recent generations of iPhones, Google Pixel, Samsung Galaxy, and Fairphones to obtain a general smartphone manufacturing footprint estimate of ~52 kg CO^2e per device.

Combining those numbers, we estimate the global yearly server manufacturing footprint to be ~12 Mt CO^2e and global yearly smartphone manufacturing footprint to be ~87 Mt CO^2e.

Combined baseline footprint estimates, 2021

To incorporate parameter uncertainty into our analysis, we assign uncertainty distributions to each parameter in our analysis (see Appendix) and sample from those distributions to generate a range of potential outcome estimates.

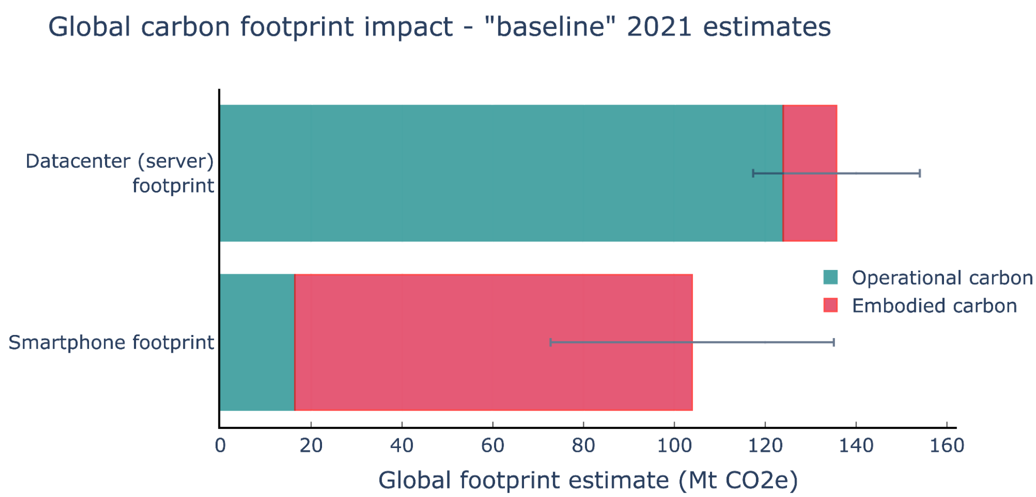

Figure 2. Estimate of global datacenter (server) and smartphone global footprints in 2021, inclusive of operational and embodied carbon emissions

Our baseline footprint estimate is consistent with previous estimates of data center and consumer devices footprints in that the footprint of data centers tends to be dominated by use phase impacts, while consumer device impacts tend to be dominated by embodied impacts prior to significant impact-limiting interventions.[4,2]

Global footprint estimates, 2030

Next, we estimate the future global footprints of data center (servers) and smartphones by building off of existing 2030 parameter projections.

Estimating operational footprints, 2030

To estimate the operational energy usage of global data centers in 2030, we replicate the method used to estimate growth in data center energy usage to 2030 implemented in Andrae (2020)[6] and Andrae and Edler (2015).[5] We use estimates of 2021 total datacenter traffic (EB) and the estimated compound annual growth rate (CAGR) of total datacenter traffic between 2016 and 2021 from Cisco's Global Cloud index report[23] to predict 2030 total datacenter traffic, assuming that datacenter traffic continues to grow at the same constant rate between 2021 and 2030. In order to estimate total electricity demand from data centers in 2030, we combine that estimate of 2030 total datacenter traffic (EB) with expectations of yearly increases in the energy efficiency of global server stocks, which we allow to vary between 5 and 15% per year. This method resulted in a range of estimated global data center energy consumption in 2030 of between ~450 and 1250 TWh, consistent with the lower range of predictions proposed in previous work (500 TWh to 3,000 TWh).[24]

To estimate the operational energy usage of smartphones in 2030, we assume that the energy usage per device per year does not change from our 2021 estimate (remains ~7.5 kWh per year expected per device) based on the stability of total energy consumption by Google Pixel phones over recent generations (see energy efficiency section). Estimates of total smartphones in operation in 2030 range from 5 billion [25] - 7.861 billion (estimate is for 2028 rather than 2030)[17].

We combine those estimates of 2030 energy demand with IEA predictions of global average emissions intensity of electricity in 2030, ranging from 165-330g CO^2/kWh in 2030 across the modeled scenarios.[19] Taken together, we estimate the 2030 operational carbon footprints of data center and smartphone ICT components to be ~195 Mt CO^2 and ~12 Mt CO^2, respectively.

Estimating embodied footprints, 2030

For both servers and smartphones, we utilize the same base estimates of manufacturing emissions per device as in the 2021 scenario to predict 2030 impacts, but allow the manufacturing emissions from energy usage to decrease in accordance with expected lowering of average global emissions per unit of electricity generated to 2030.

To isolate the portion of device embodied carbon stemming from fabrication electricity, we make several assumptions using previously reported data. First, we isolate the proportion of manufacturing emissions of servers and smartphones that are attributable to integrated circuit fabrication (including logic and memory production). For servers, we utilize a Dell 2019 life cycle analysis[26], which attributes ~78% of total manufacturing impact of a server to SSD and large IC production. For smartphones, reported life cycle analysis data for Fairphones 3[27] and 4[28] indicate that 66-69% (for the purposes of having one number, we use 67.5% on average) of their manufacturing impact stems from integrated circuit production. Other embodied impact contributors that are not included in the integrated circuit manufacturing impact include the impact of device transportation, displays, batteries, etc.[27]

Next, we estimate the portion of the integrated circuit manufacturing impact that can be attributed to electricity as opposed to process emissions. In 2021, TSMC[29], SK Hynix[30], and Micron[31] reported between 45% and 54% of their total emissions as Scope 2 impacts, defined within the Greenhouse Gas Protocol as indirect emissions stemming from the purchase of electricity.

Finally, we convert the resulting emissions value from electricity into a quantity of electricity demanded in order to be able to estimate the value of reducing the emissions factor of that electricity. To obtain an estimate of global electricity use for production of each device type, we use an estimate of Taiwan's grid emissions factor for 2021 reported by the Bureau of Energy of 0.509 kg CO2e/kWh[32] to back-calculate total yearly electricity needs. We utilize Taiwan's grid emissions factor for this estimate because over 60% of the world's semiconductors (and over 90% of the world's advanced semiconductors) are produced there.[33]

Once we have the quantity of electricity required for production of each device, we calculate the estimated reduction in device embodied emissions by subtracting the original per device emissions from fabrication electricity from a 2030 per device emissions from fabrication electricity calculated by multiplying the expected global average carbon intensity of electricity in 2030 (165-330 g CO^2/kWh[19]) by the quantity of electricity required for device production.

We combine those lowered per device manufacturing emissions estimates with projections of the number of server and smartphone devices expected to be produced in 2030: ~60 million servers[6] and ~1.7 billion smartphones.[6,34] Taken together, we estimate the global embodied emissions impact of servers and smartphones in 2030 to be ~59 Mt CO^2e and ~74 Mt CO^2e, respectively.

Combined baseline footprint estimates, 2030

As for the 2021 baseline estimates, we assign uncertainty distributions to each parameter in our analysis (see Appendix) and sample from those distributions to generate a range of potential outcome estimates for 2030 impacts.

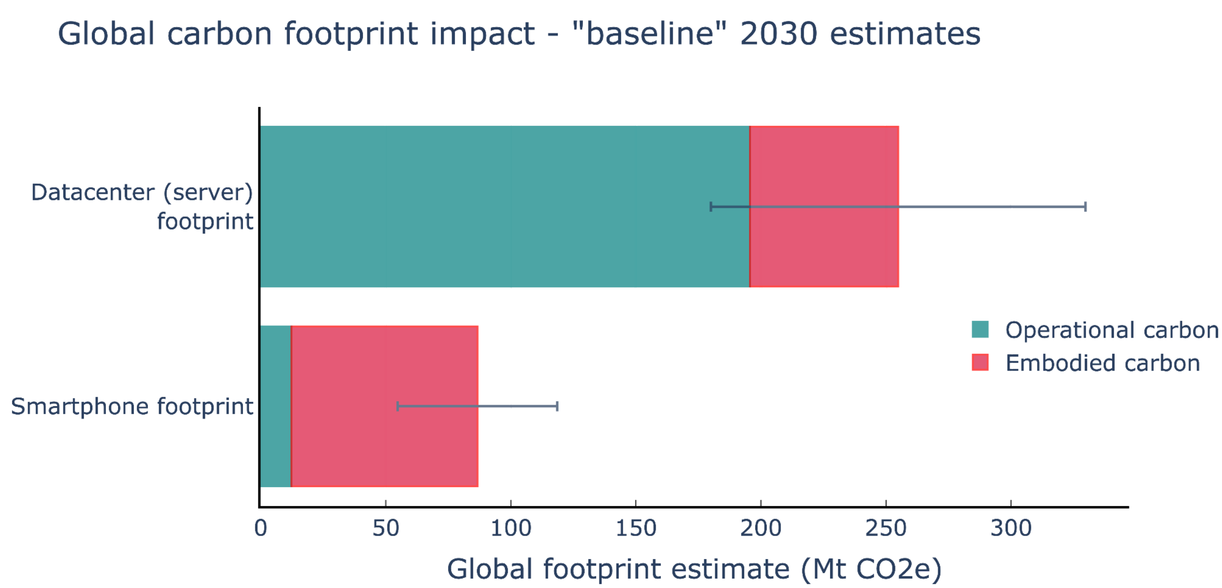

Figure 3. Prediction of global datacenter (server) and smartphone global footprints in 2030, inclusive of operational and embodied carbon emissions

We can now visualize the expected changes in impacts between 2021 and 2030 for both the smartphone and datacenter (server) impact segments. Notably, our estimates indicate no significant change in global smartphone embodied and operational impacts between 2021 and 2030 (note that we do not include the energy associated with network and data center activity here, just direct energy consumption by smartphones and emissions from their production) and an increase in the global datacenter (server) footprint between 2021 and 2030. In a subsequent section of this thesis, we evaluate the sources of uncertainty in these baseline estimates. We identify the operational footprint parameters as the largest contributors to uncertainty for our estimates of both smartphones and data centers (servers). Meanwhile within embodied emissions estimates for both smartphones and data centers (servers) the uncertainty in estimates derives primarily from uncertainty over the manufacturing emissions associated with an individual device.

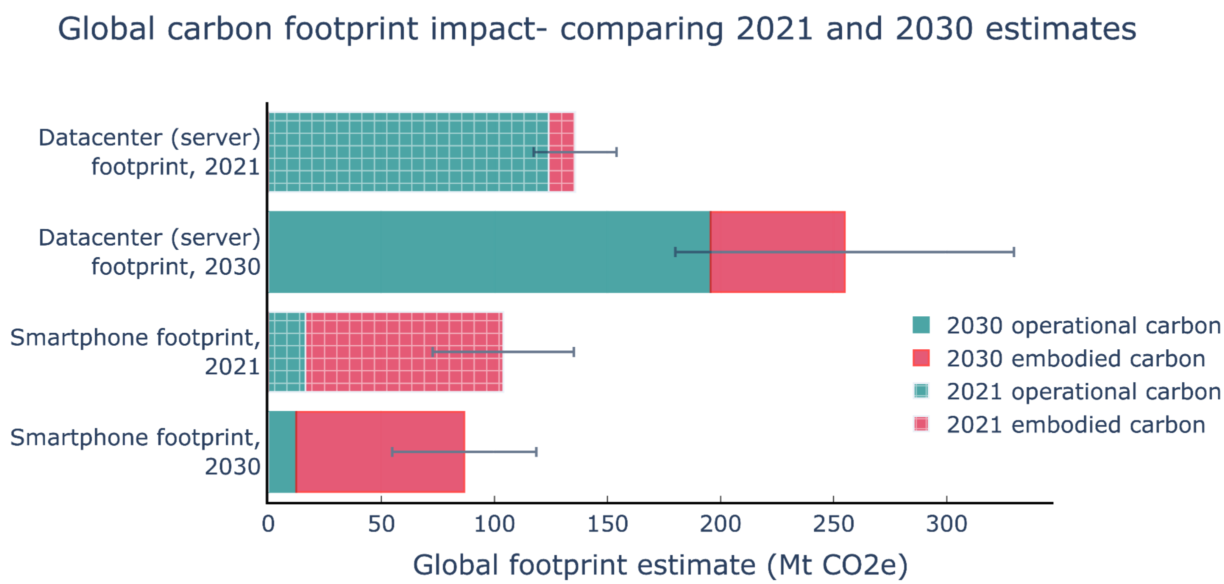

Figure 4. Comparison of 2021 and 2030 estimates of global data center (server) and smartphone carbon footprints

In the following sections, we evaluate several strategies for reduction of these global footprints in 2030. This exercise is especially relevant in the context of ICT industry science-based targets which set a sector wide goal of achieving 45% emissions reductions between 2020 and 2030.[15] Our analysis provides context on the feasibility of achieving those reductions and the combination of interventions that will be required to do so. We construct 2030 “targets” for the portions of datacenter (server) and smartphone global footprints that we have outlined as a 45% reduction from our 2021 baseline values for each component, or ~74 Mt for data centers (servers) and ~57 Mt for smartphones.

Renewable energy for operational footprint reduction

First, we evaluate the potential impact renewable energy integration could have on reducing operational footprints in 2030, including the additional embodied emissions that would result from the intervention. The operational footprint impacts of both smartphones and data centers (servers) derive from use-phase electricity consumption. Given the significant global energy demand of the ICT sector, a shift towards cleaner energy has the potential to reduce sector operational emissions. However, since consumer devices like smartphones are used decentrally and subject to the emissions intensity of a user's electrical grid connection,[9] powering consumer device operation with renewable energy requires continued grid-level investments that are the focus of governmental organizations and the utility sector. Data centers, on the other hand, are centralized and amenable to renewable energy integration. Given that data center computing footprints are historically dominated by use-phase impacts,[4] decarbonizing their electrical supply is an important lever for impact reduction within the sector. While one study notes that the business case for investing in renewable energy to power data centers is less clear than investments in energy efficiency (which can directly cut costs),[35] today large datacenter operators are making significant investments in renewable energy.[3] The ICT sector is currently responsible for over half of global renewable energy power purchase agreements in terms of capacity, with firms Amazon, Microsoft, Meta and Google listed as the top four corporate purchasers of power purchase agreements worldwide.[15] In an effort to further reduce their impacts beyond matching renewable energy purchases with their energy consumption, Google and Microsoft have both recently set 2030 goals of running their data centers entirely on 24/7 carbon free energy.[36,37]

While renewable energy deployment can reduce data center impacts, renewable energy incorporation for data center operations is only a partial solution to reducing the sector's carbon footprint. Firstly, not every data center is located in a geography that features suitable renewable energy resources and this resource availability will limit the extent to which renewable energy can replace existing power sources.[7] In addition, taking full advantage of clean energy resources may require changes to how compute resources in a data center are utilized. For example, Google's Carbon-Intelligent Compute Management capitalizes on flexible workloads within their enterprise to deploy those non-time sensitive workloads when more clean energy is available.[38] While this strategy is effective at taking advantage of clean energy availability to reduce data center use phase impacts, it also has an impact on hardware utilization and requires additional server capacities for the same amount of compute in order to allow for flexibility in workload deployment.[7] Even if a datacenter is operated on 100% renewable energy, there are embodied carbon costs associated with the renewable energy infrastructure, excess server capacities, and batteries that enable use of that clean energy[7] - meaning that the use phase electricity impact of data centers will never be fully eliminated, but may shift towards embodied impacts. This embodied versus use phase energy tradeoff (more hardware is needed to allow for lower use-phase impacts) might be negligible at lower levels of renewable energy incorporation, where use phase impacts dominate over the embodied impacts of data center systems. However, as more clean energy is incorporated into a data center, the embodied impact of the system becomes an increasingly large portion of the remaining impact and continued renewable energy integration may no longer optimally reduce the overall footprint of the data center system.

"Best case" estimation of global impact reduction

To develop a "best case" estimate of the potential for data centers to reduce their operational impacts through dedicated renewable resources, we utilize Facebook Research's Carbon Explorer tool which aims to minimize the total footprint of a data center system inclusive of the embodied carbon tradeoffs of additional renewable energy, battery, or server capacities.[39] We input our expected 2030 data center energy demand, estimates of server embodied carbon in 2030, and 2030 global average carbon emissions from electricity to explore an optimistic 24/7 renewable global data centers scenario. We use renewable energy profile data from Utah, which Acun et al.[7] found was a good location for minimizing total carbon footprint out of the locations surveyed, and obtain a carbon optimal scenario for a five day sample (using the default date range of May 10-15, 2021 within the Carbon Explorer tool[39]) of global data center operation in order to demonstrate the potential opportunities and tradeoffs associated with high levels of renewable energy incorporation.

The lowest carbon scenario found utilizes both renewables and carbon aware scheduling to reduce the operational footprint of the datacenters to ~2 Mt CO^2e. Those operational gains are offset to some extent by an increase in system embodied footprint by ~25 Mt CO^2e to account for ~500 GW of wind capacity, and ~26% additional server capacity. In addition, we note that the optimal percentage of renewable energy coverage the model finds is consistently below 100% (~99% on average across sampled iterations), with the remaining electricity sourced from the grid to meet our hypothetical global data center demand. While some amount of dedicated renewable resources can reduce data center operational impacts, achieving 24/7 renewable energy coverage for data centers may not be carbon-optimal once embodied emissions impacts and the limits of workload flexibility are taken into account.

Figure 5 below demonstrates the change in data center impacts following "best case" renewable energy integration for operational footprint reduction:

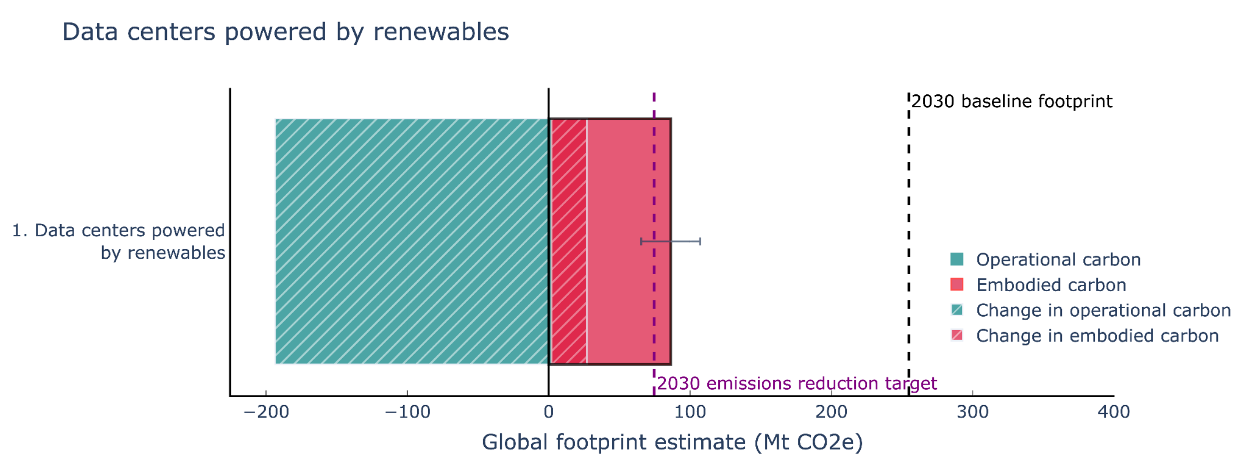

Figure 5. Best case impact estimate of providing renewable energy for data center operation on global data center (server) footprint in 2030

This "best case" scenario is likely unattainable for global data centers for multiple reasons: many data center locations may not have access to reliable renewable energy resources to take advantage of, the required renewable capacity is significant (the 500 GW of wind capacity selected in this analysis would require significant wind energy expansion given that 830 GW of wind energy capacity was installed worldwide in 2021[40]), and we make assumptions about the degree of workload flexibility available across global data centers (assuming 50% of workloads can by flexibly deployed). However, while unrealistic to some extent, this exercise demonstrates the limitations of operational renewable energy integration to reduce overall impacts. In this best case scenario, on average, a ~98% reduction in data center operational footprint from the 2030 baseline is accompanied by a 42% increase in data center embodied footprint, shifting the impact towards embodied footprint dominance. Taking this embodied footprint increase into account, in this scenario total emissions of the data center subsector are on average maximally reduced by ~66%. Additional levers for impact reduction are needed to address the remaining embodied footprint impacts of data centers and to potentially further reduce the operational footprint of data centers beyond what is achievable by renewable energy integration alone. As noted in Gupta et al. 2020[2], once data center operators like Facebook and Google successfully reduce their Scope 2 emissions through strategies like purchasing renewable energy, the emissions of their supply chains emerge as the dominant contributor to remaining footprints. While we may be able to achieve the 2030 sector emissions reduction targets by this lever alone (the 2030 emissions reduction target lies within range of the error bars for our post-intervention estimate), on average the estimated impacts after this intervention exceed the 2030 targeted emissions even in this "best case" scenario. In order to ensure that the 2030 emissions reduction targets are met, other interventions to reduce total emissions are needed.

Renewable energy for embodied footprint reduction

Next, we examine the extent to which the impact of both server and smartphone life cycles could be reduced through renewable or cleaner electricity development for fabrication facilities. This intervention would act upon the portion of the embodied footprints of the devices attributable to fabrication electricity.

To isolate the portion of device embodied carbon stemming from fabrication electricity, we utilize the same impact attribution strategy reported above to obtain the quantity of electricity required for smartphone and server device production (see Estimating embodied footprints, 2030).

"Best case" estimation of global impact reduction

Unlike the previous example of renewable energy for datacenter operation, load flexibility is less salient for fabs, as they are typically operated at near max capacity[41] around the clock[42] and thus redundancy in equipment is a less viable option as compared to the server operation scenario in Lever 1. Because of this, it does not make as much sense to invest in dedicated renewable resources for fabs, as significant additional battery capacity would be required to compensate for the lack of load flexibility and provide the constant power supply needed, implying significant additional embodied carbon costs. This may help explain why, in general, fabs tend to rely on grid resources for their electricity supply, while sometimes supplementing those supplies with fab-owned fossil fuel power plants[43] that are oriented more towards concerns about electrical supply reliability than emissions reductions. As in the operational energy context, some manufacturing firms, like Intel, are increasing purchasing of renewable energy certificates and credits in order to offset their Scope 2 emissions, providing investments for local grid greening.[44]

However, it is important to note that 1:1 purchasing of renewable energy for energy consumed does not mean that green electricity is being used at all times by the specific manufacturing plant.[44]

For a best case impact reduction scenario, we consider the potential reduction in global device impacts observed if all fabrication facilities were located in a grid system with a low carbon emissions factor (we use Sweden as an example of a "best case" low carbon grid, with a reported grid emissions factor of 0.028 kg CO2e/kWh in 2020[45]). This could be realized either through strategic siting of fabrication plants in areas with low grid emissions factors, or through large-scale greening the grid systems of areas where fabs are currently located.

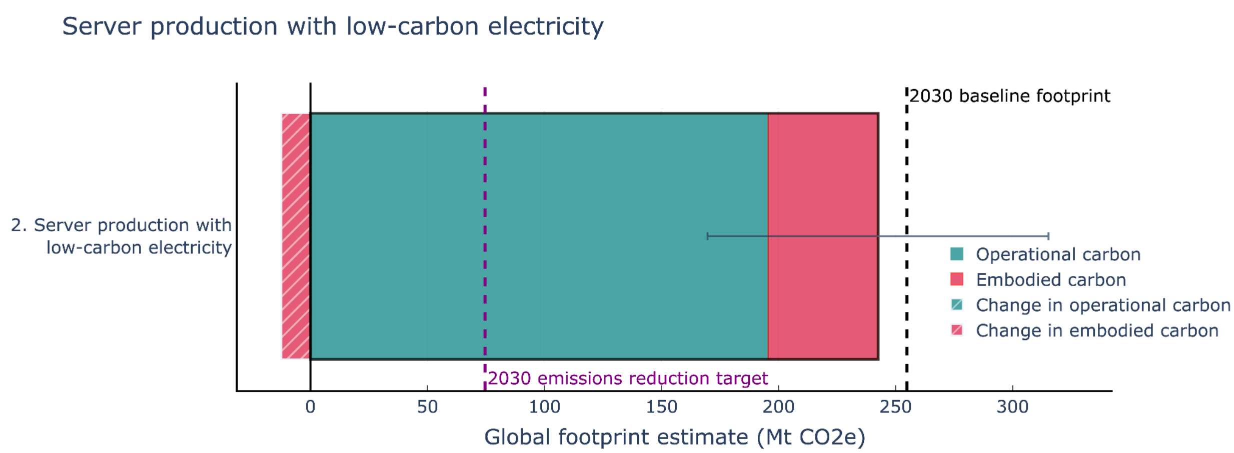

Figure 6. Best case impact estimate of manufacturing server integrated circuits with low-carbon electricity on global data center (server) footprint in 2030

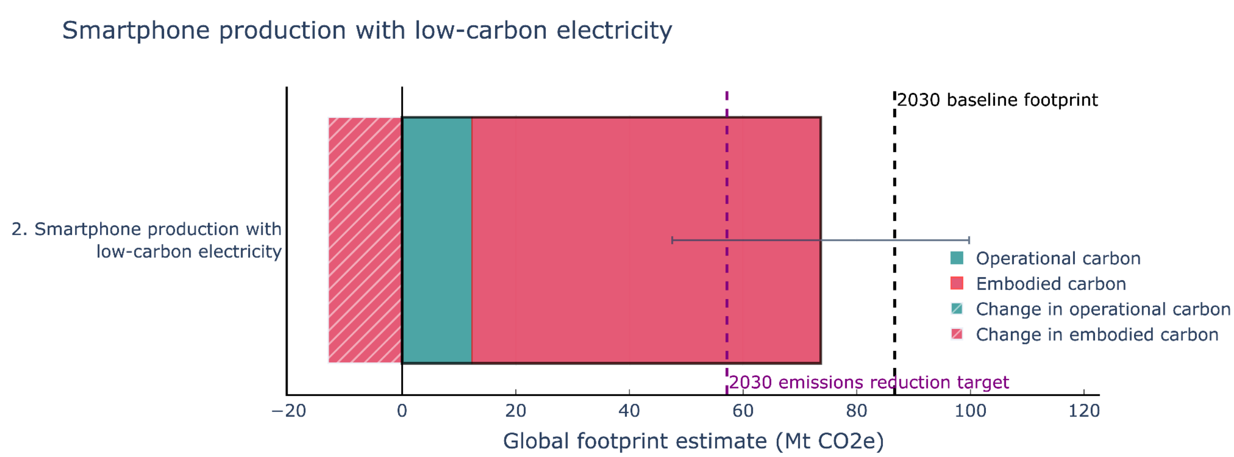

Figure 7. Best case impact estimate of manufacturing smartphone integrated circuits with low-carbon electricity on global data center (server) footprint in 2030

In simulating global fabrication of server and smartphone devices in Sweden rather than Taiwan, we see a ~12% decrease in global server embodied footprints and a ~17% decrease in global smartphone embodied footprints. For smartphones, this results in ~15% decrease in the total global footprint on average. The impact on total global datacenter (server) footprints is much lower, at ~4.8% on average across our simulations due to the dominance of the operational footprint contributions to the total footprint for the datacenter segment.

It is important to note that this scenario is a simplification that likely overestimates the carbon footprint benefits of transferring production to a new grid system or greening an existing grid system by failing to account for additional grid infrastructure investments (and therefore additional embodied carbon costs) that could be considered attributable to the fabrication facility's energy demand. While it may be difficult to estimate the portion of the embodied impact of a grid energy system that is attributable to a specific fabrication facility, it is important to consider how the external embodied costs of grid infrastructure should factor into device footprints. Accounting for those external embodied costs would turn our "best case" impact reduction estimate for this lever into strictly an upper bound. For example, Intel's plans to construct a new manufacturing facility in Ohio are related to ongoing large-scale solar farm developments in the area - and some portion of the embodied costs associated with that investment in additional renewable resources should likely be attributed to the carbon footprint of the firms, including Intel, that plan to utilize energy from those renewable resources.[46]

Unlike the operation of servers and smartphones, the fabrication of integrated circuits for electronic devices is highly concentrated within certain regions of the world. The geographic and geopolitical context of semiconductor manufacturing complicates the realization of the proposed "best case" scenario for impact reduction. For example, much of the logic manufacturing market is concentrated in Taiwan, characterized by a small electricity grid that the semiconductor sector has a large impact on. In 2022, TSMC accounted for around 6% of total electricity demand in Taiwan - and that proportion is expected to continue to increase in the near future.[47] By 2030, TSMC is aiming for 25% renewable energy to power their fabrication facilities.[48] Meanwhile, Taiwan has set a nationwide goal of 27-30% renewable energy by 2030.[49] These 2030 targets would fall short of achieving the best case electricity emissions factor presented in this analysis, given the much higher penetration of renewables in Sweden (~54%), whose grid emissions factor we utilize in the "best case" calculations.[50]

Further, the size of Taiwan's grid and the renewable energy resources available locally may complicate the realization of lower-carbon electricity for fabrication facilities in the future. While TSMC has announced an ambitious goal of sourcing its electricity from 100% renewable energy by 2050, Taiwan recently reduced their 2025 renewable energy target from a goal of sourcing 20% of total electricity generation from renewables[51] to 15% in light of the challenges in quickly scaling up renewable capacity beyond the existing 6.3% on the grid system,[47,52] underlining the difficulty of achieving that goal. These challenges are motivating development of new fab facilities in areas with more renewable energy resource availability, such as Arizona and Japan,[47] but geopolitical interests may complicate ideal fab siting locations from the perspective of renewable energy availability.

Increased lifetime for embodied footprint reduction

Thirdly, we estimate the potential for increasing the lifetime of computing devices as a way to reduce embodied impacts, while highlighting the operational emissions tradeoffs associated with this intervention. In general, the lifetimes of servers and computing hardware within a data center or HPC are determined by business decisions rather than the technical lifetime limits of a device. This is evident in recent trends, as slowdowns in generational efficiency changes, supply chain issues, and algorithmic efficiency gains have altered the business optimal lifetime of data center hardware and resulted in firms using hardware longer.[53] Since 2020, Amazon, Microsoft, and Alphabet have all announced increases in their expected useful lifetimes of servers by as much as 1-2 years, with expected monetary savings in the billions.[54,55] The extendable nature of hardware refresh cycles demonstrates that data center hardware is generally replaced before it reaches the end of its useful life. Unlike for smartphone devices, where multiple studies have emphasized the benefits of lifetime extension[9] and second-life refurbishment[8] for device lifecycle impact reductions (with one study even suggesting that a smartphones lifetime is much more important to overall impact than use patterns[56]), the impacts of lifetime extension of datacenter hardware on a systems overall footprint are less explored and more controllable from the standpoint of ICT operators (smartphones lifecycles are in the end determined by individual consumer decisions). Given the level of previous work around lifetime extension of consumer devices and smartphones and that the decision of lifetime extension for smartphones lies in the hands of individual consumers rather than ICT companies, we do not include an estimate for this lever for smartphones in this work, but rather focus on the data center (server) impact segment.

"Best case" estimation of global impact reduction

Because the operational efficiency of servers generally increases over time, extending the lifetime of datacenter hardware will result in some loss in operational efficiency as compared to more frequent device refresh cycles. To begin to estimate what the potential embodied emissions reduction and operational energy trade off might be from server lifetime extension within a data center, we assume that server lifetime extension is practically limited to a few years of extension based on current consumption patterns and that the intervention applies only to reduce server production for servers intended to replace existing stock, rather than servers produced to satisfy new demand.

Using year on year server shipment growth to estimate the portion of manufacturing impact attributable to new device production, we simulate an extension to 6 year average refresh cycles (from a 4 year baseline, mimicking the lifetime assumption in the Dell server LCA[26]) for data center hardware to see the potential overall footprint reduction that lifetime extension can have in a given year.

In order to estimate interventions on the lifetime of servers, we first need a way to estimate server shipments in the years leading up to 2030. We fit a second order polynomial equation to bridge the historic data on IDC server shipments[22] with the 2030 projection of 60 million servers shipped.[6] Then, we use the fitted model to predict server shipments in the years leading up to 2030.

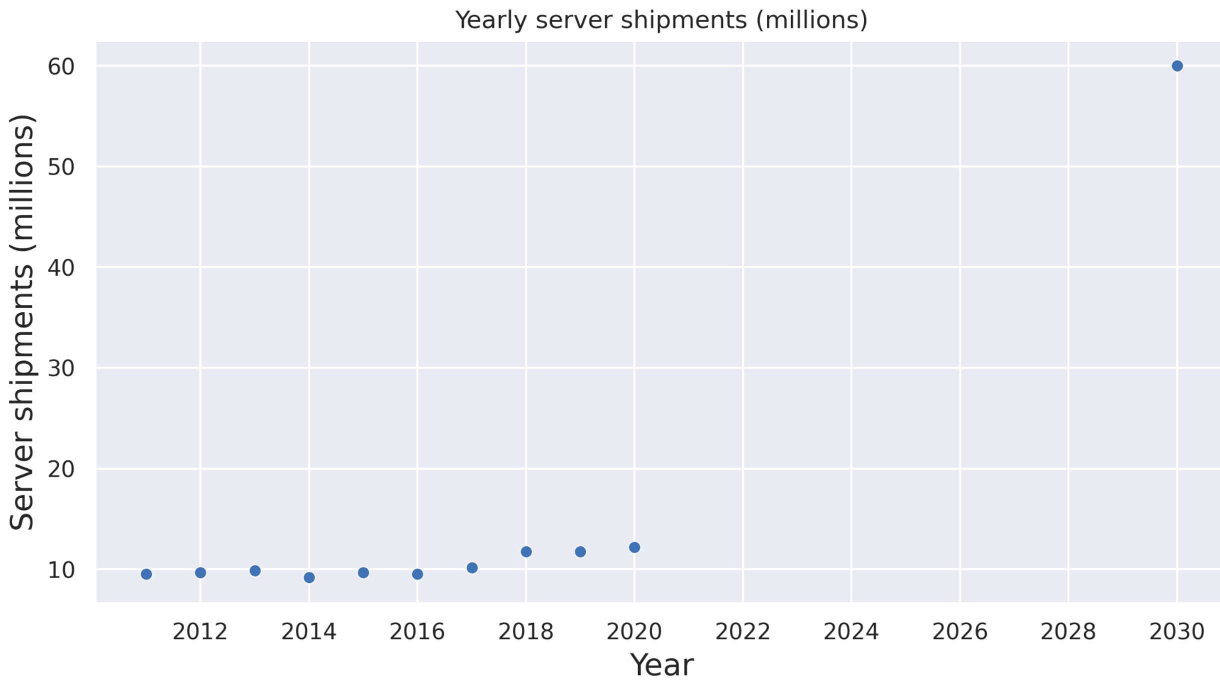

Figure 8. Trend in yearly server shipments, historical and predicted

Future years	Model predicted server production
2026	36.42
2027	41.56
2028	47.14
2029	53.15
2030	59.59

Table 1. Model predicted server production in the years leading up to 2030, baseline scenario

Assuming a 4 year server lifetime, the established server stock going into 2030 would be:

server_stock_baseline = servers_2029 + servers_2028 + servers_2027 + servers_2026 = 178 million servers

To maintain the same amount of server stock, a 6 year replacement cycle would allow for fewer servers to be replaced in any given year than a 4 year replacements cycle, at the cost of less efficient, older servers remaining in operation.

To estimate the server shipments required to maintain the same server stock with a 6-year refresh cycle, we calculate the excess server stock from adding up servers shipped between 2024 and 2029 as compared to the 4 -year baseline:

server_stock_6year = servers_2029 + servers_2028 + servers_2027 + servers_2026 + servers_2025 + servers_2024 = 237 million servers

excess_stock = server_stock_6year - server_stock_baseline = 59 million

Then, we adjust the intercept on the equation to estimate yearly server shipments to eliminate the excess stock in a 6-year refresh scenario:
intercept_adjustment = excess_stock/6= 9.8

6-year refresh adjusted server shipment equation: 0.2166094765(year**2) - 872.775611*year + 879168.0953-intercept_adjustment*

Using this new equation, we re-estimate yearly server shipments under the extended, 6 year, lifetime scenario:

Future years	6-year refresh scenario shipment predictions
2024	17.58
2025	21.86
2026	26.57
2027	31.71
2028	37.29
2029	43.29
2030	49.74

Table 2. Model predicted server production in the years leading up to 2030, extended server lifetime scenario

This results in an estimate of 49.74 million servers to be produced in 2030. We now use data on the increases in Dell server efficiency since 2015[57] to estimate an year on year average efficiency improvement for servers of ~18% per year).

Next, in order to estimate the operational footprint cost of the delayed server refresh, we use the below equation to solve for a baseline energy use under the baseline 4-year replacement scenario, assuming yearly server efficiency improvements of 18%:

*datacenter_op_eng = servers_2029*base_energy*0.82**3 + servers_2028*base_energy*0.82**2 + servers_2027*base_energy*0.82 + servers_2026*base_energy*

We then use that base_energy value to calculate the total energy demand for the same amount of work in the 6-year refresh scenario (where some of the servers in the stock are operating less efficiently).
*datacenter_op_eng_6year = estimate_server_shiptments_6year(2029)*base_energy*0.82**3 + estimate_server_shiptments_6year(2028)*base_energy*0.82**2 + estimate_server_shiptments_6year(2027)*base_energy*0.82 + estimate_server_shiptments_6year(2026)*base_energy+ estimate_server_shiptments_6year(2025)*base_energy*0.82**-1 + estimate_server_shiptments_6year(2024)*base_energy*0.82**-2*

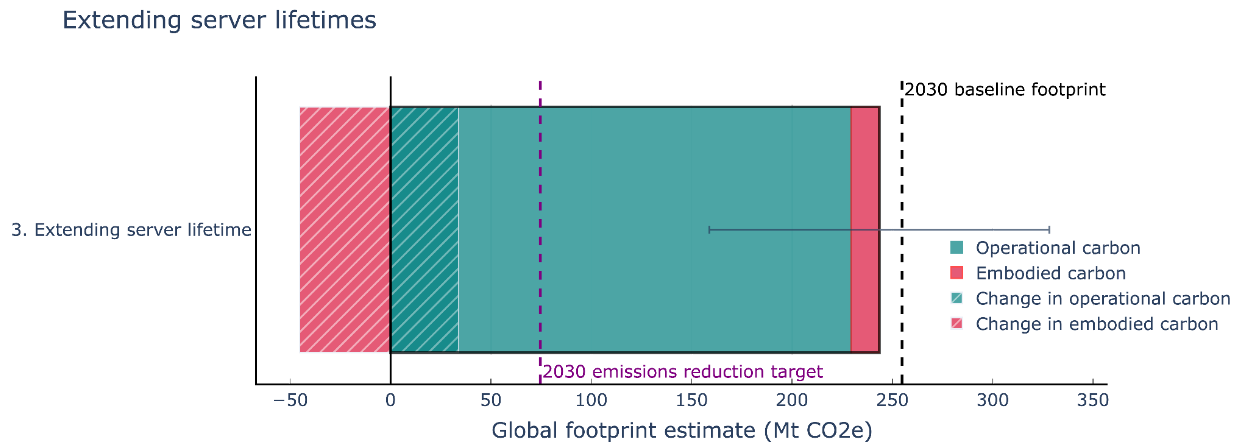

Figure 9. Best case impact estimate of extending server lifetime on global data center (server) footprint in 2030

By these estimations, the server lifetime intervention (from 4-year to 6-year retirement age) results in an 4.4% decrease in the overall global datacenter (server) footprint due to a 76% decrease in embodied carbon and 17% increase in operational carbon emissions. While lifetime extension may not be as effective of a strategy for reducing the global impact of data centers (servers) in isolation, as use-phase interventions like renewable energy powered data centers come into play, we expect that extending the lifetime of devices will prove more valuable as a way to achieve lower embodied emissions and entail more limited rebounds in operational carbon.

Energy efficiency for operational impact reduction

Lastly, we explore the potential benefits of harnessing energy efficiency gains in computing devices for operational footprint reduction. Increasing the energy efficiency of computing devices has the potential to limit the operational footprint of the ICT sector by requiring less energy for the same amount of computing power. We highlight three integrated circuit hardware-related modalities for energy efficiency improvement within computing systems: continued energy efficiency improvements through Moore's law transistor scaling, energy efficiency improvement through hardware specialization, and energy efficiency improvements through new technologies (new memories, new transistors, etc.).

Moore's law

Moore's law has traditionally operated to reduce operational emissions through additional computation per unit energy afforded by continued transistor scaling. In general, previous work shows that Moore's law scaling acts to reduce the embodied impacts of devices alongside increases in device operational efficiency- while the per wafer impacts of more scaled technology generations increase owing to additional manufacturing steps and complexity, area scaling trends allow for an overall decrease in manufacturing impact when normalized by cell area, even for highly scaled nodes.[10,11] Despite the embodied footprint reduction benefits of area scaling, over time the embodied impact of a given computing system tends to rise alongside progressive node scaling as computational demand on a given system increases along with increased device efficiency, requiring additional cell area and thus additional manufacturing impacts.[10]

Specialization

In recent years, efficiency gains in the computing sector have been driven not only by continued CMOS device scaling according to Moore's law, but through device specialization with functionalities not necessarily in line with larger, general device trends (coined as "More than Moore" development by ITRS in 2005[12]). Specialized devices, like GPUs and domain-specific accelerators provide more efficient operations on specific workflows that the device is optimized for (such as neural network training and inference) than general purpose hardware like CPUs.[58] Specialized hardware designs have the potential to reduce operational emissions across both data centers and consumer devices. It is important to note that specialization, like the other levers associated with efficiency gains, can fail to reduce overall impacts in response to increased computing demands with increased specialization. An OpenAI article from 2018[59] notes that, since 2012, the doubling time for compute requirements of AI algorithms has been around 3.4 months, far outpacing Moore's law technology scaling (which has a 2 year doubling time). This trend is supported to some extent by efficiency gains from specialized hardware utilization which have allowed an acceleration in model growth since 2012, when GPUs began to be utilized more commonly for acceleration.[59] Google's analysis of their datacenter systems reports energy efficiency benefits from specialized hardware use alone of 2-5x performance per watt.[60] Meanwhile, Facebook reported a 10.1x energy efficiency improvement on their ML workloads from specialized hardware utilization.[61] Differences in the magnitudes of energy efficiency improvement from specialized hardware utilization may depend on the workload - the 2-5x improvement reported by Google refers to improvements on machine learning workloads generally,[60] while the 10.1x improvement reported by Facebook was referring to improvements seen on specific LM (transformer based language model) tasks.[61]

Unlike with Moore's law scaling, the efficiency gains from hardware specialization are only applicable to a portion of a computing system's total workloads (those specific functions that an accelerator is designed to perform) - this lever does not increase the efficiency of computing performed by general purpose hardware (eg. CPUs). While it is difficult to estimate the proportion of global computing workloads eligible for hardware acceleration, we can use recent data on hardware acceleration and AI workloads published by Google to get an idea of the upper bounds for what hardware acceleration could accomplish on a global scale. Based off of surveys of its datacenter workloads in 2019, 2020, and 2021, Google reported between 70 and 80% of total FLOPs as attributable to ML workloads, but that a variety of efficiency optimizations (algorithmic and hardware-based) have enabled ML-related operations to consume only 10-15% of total data center energy during those surveys.[60] The embodied emissions implication of shifts towards specialized hardware is underexplored,[13] with more research needed to evaluate how specialization impacts both operational and embodied emissions within data center environments.

In terms of the operational footprint of accelerator workloads for smartphone devices, another estimate by Google, this time describing the energy usage of the TPU in the Google Pixel 6 gives us some notion of the potential for operational impact reductions. Google reported that the Edge TPU within the Pixel 6 made up less than 1% of device energy consumption, with ~8% of device energy consumption attributed to device CPU and GPU.[60] From this they estimate an upper bound for consumer device ML operational energy consumption of 5%.[60] While the inclusion of specialized hardware in consumer devices can lead to more efficient operation, Gupta et al (2022)[2] notes the additional embodied emissions cost of including the additional hardware within a mobile device, using the Google Pixel 3 as an example. Depending on the utilization of a specific device, the operational efficiency benefits of additional specialized hardware inclusion within a consumer device may not exceed the embodied carbon costs of producing the additional integrated circuit area.

New technologies

New technologies could allow additional efficiency gains for integrated circuit devices beyond those provided by specialization and Moore's law efficiency improvement trends. For example, new memories, such as RRAM, STT-MRAM, and PCM have the potential to replace existing memory technologies or allow new computing paradigms (such as in memory computation) that could increase the efficiency of operational workloads. Similarly to trends around specialization, more research is needed to evaluate the potential tradeoffs that could be associated with the fabrication of those new devices as compared to the use phase benefits they could offer across different computing contexts.

Limitations of energy efficiency as a means for global impact reduction

While computing technologies are continuously increasing in efficiency, in practice, this lever for ICT sector footprint reduction is practically limited by Jevons Paradox, a paradigm in which increases in computing efficiency are in turn met with increasing computing demand. A 2015 study explored the causal relationship between energy efficiency gains and energy consumption in the ICT sector and estimated rebound effects between 115-161%, indicating that increases in consumption effectively outweighed efficiency gains.[62]

Using IEA estimates of data center energy growth and total workloads capacity installed in 2015 versus 2021 (displayed in the Table 3 below),[15] we see that increases in total data center energy consumption reported are more or less consistent with the rebound effects estimated in Galvin (2015).[62] Between 2015 and 2021, the number of installed workloads per unit of energy increased from 0.9 to between 2.03 and 2.95 installed workloads per kWh based on the IEA data. A 100% rebound effect on the per workload efficiency gain would have resulted in the same level of energy use as in 2015 with comparably more installed workloads (between 406 and 590 million). However, the number of installed workloads reported for 2021 exceeds this range, pointing towards a potential rebound effect somewhere above 100%. Thus, while these data indicate substantial increases in datacenter efficiency, the overall energy demand from the sector continues to increase in line with what we might expect from Jevons paradox.

IEA Data Center Workload & Energy Estimates	2015	2021
Data centre workloads**	180 million	650 million

Data centre energy use (excluding crypto)	200 TWh	220-320 TWh

Table 3. IEA Data Center Workload & Energy Estimates.[15]
***Here, workloads refers to the unit defined within the Cisco Global Cloud Index: "A server workload and compute instance is defined as a virtual or physical set of computer resources, including storage, that are assigned to run a specific application or provide computing services for one to many users. A workload and compute instance is a general measurement used to describe many different applications, from a small lightweight SaaS application to a large computational private cloud database application. For the purposes of quantification, we consider each workload and compute instance being equal to a virtual machine or a container."*[23]

Looking at technology development trends for servers alone, operational efficiency in terms of operations per watt of energy increased steadily with server generations over the last decade (Figure 10). This trend goes hand in hand with increasing computational power per individual chip within a server - over time servers are able to do more, more efficiently (Figure 11). However, increased computational demand beyond what server technology development can offer in terms of increased computational capacity per server results in increasing yearly server shipments over the same time period.[22]

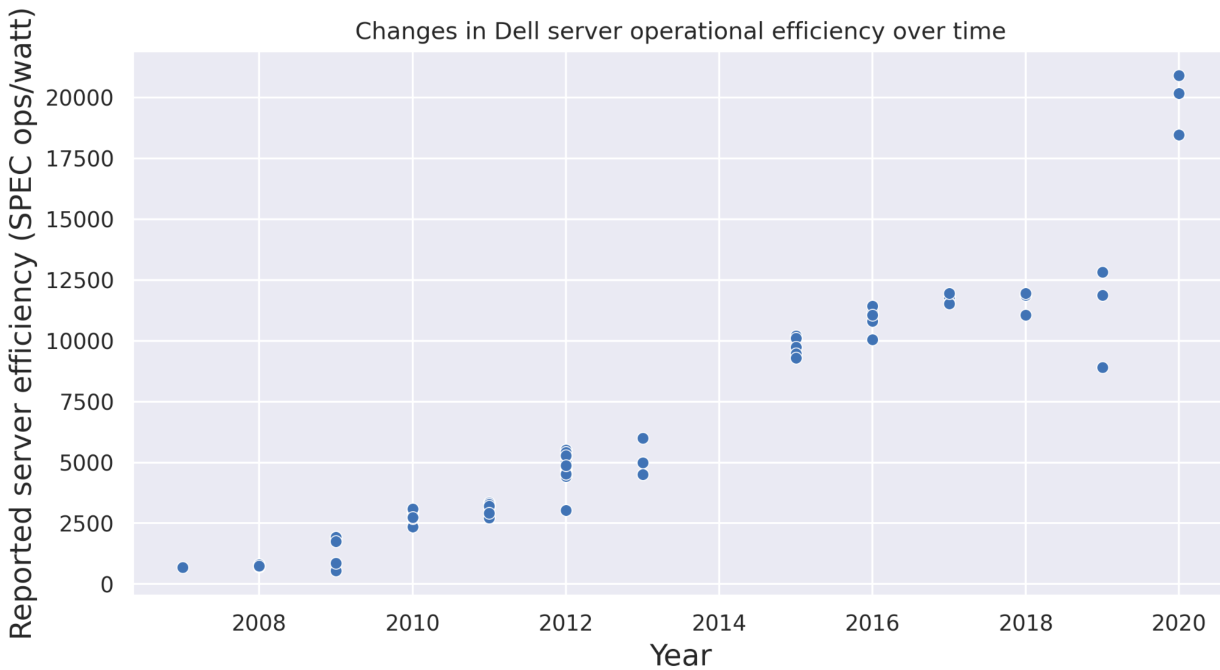

Figure 10. Increasing Dell server operational efficiency over time

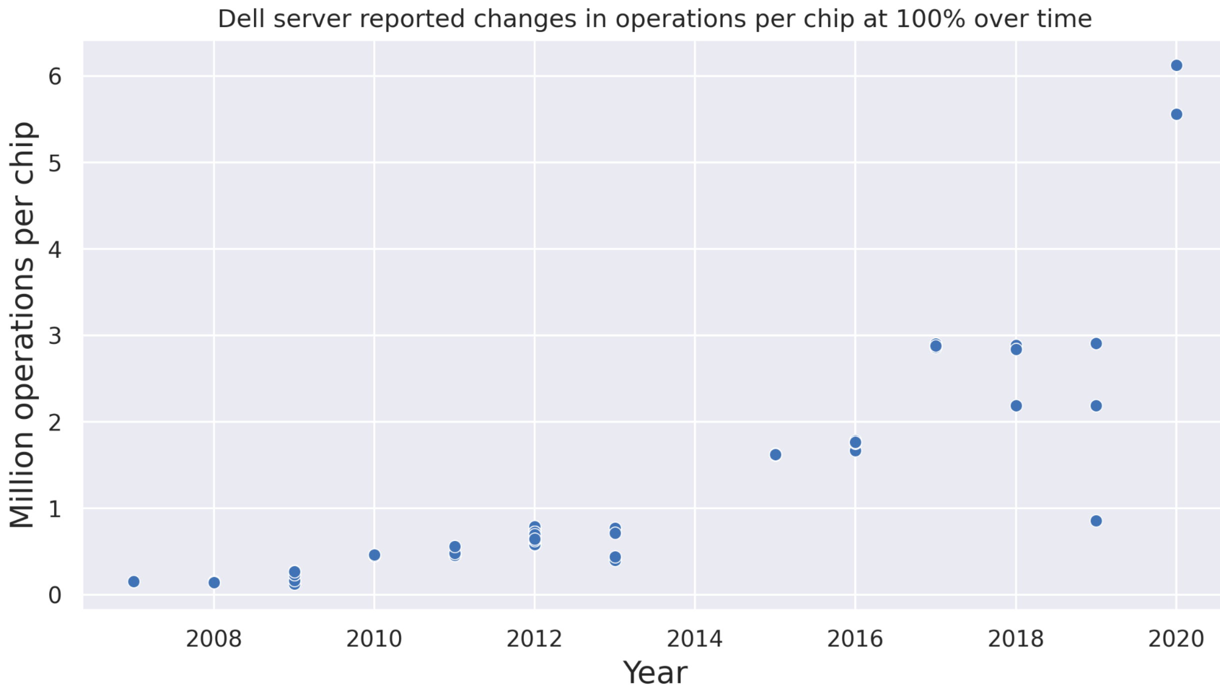

Figure 11. Increasing Dell chip computational capacity over time (in terms of operations per chip @ 100%)

Looking at smartphones, the operational footprint benefits of energy efficiency increases across technology generations are less clear. Recent Apple iPhone and Google Pixel generations demonstrate inconsistent trends in operational impact across subsequent device generations (see Figures 12 and 13 below). Also notable is the trend of increasing embodied to operational emissions impact of Apple devices, likely driven in part by increasing hardware capacities across generations that are perhaps enabled by increasing integrated circuit efficiency (see Table 4).

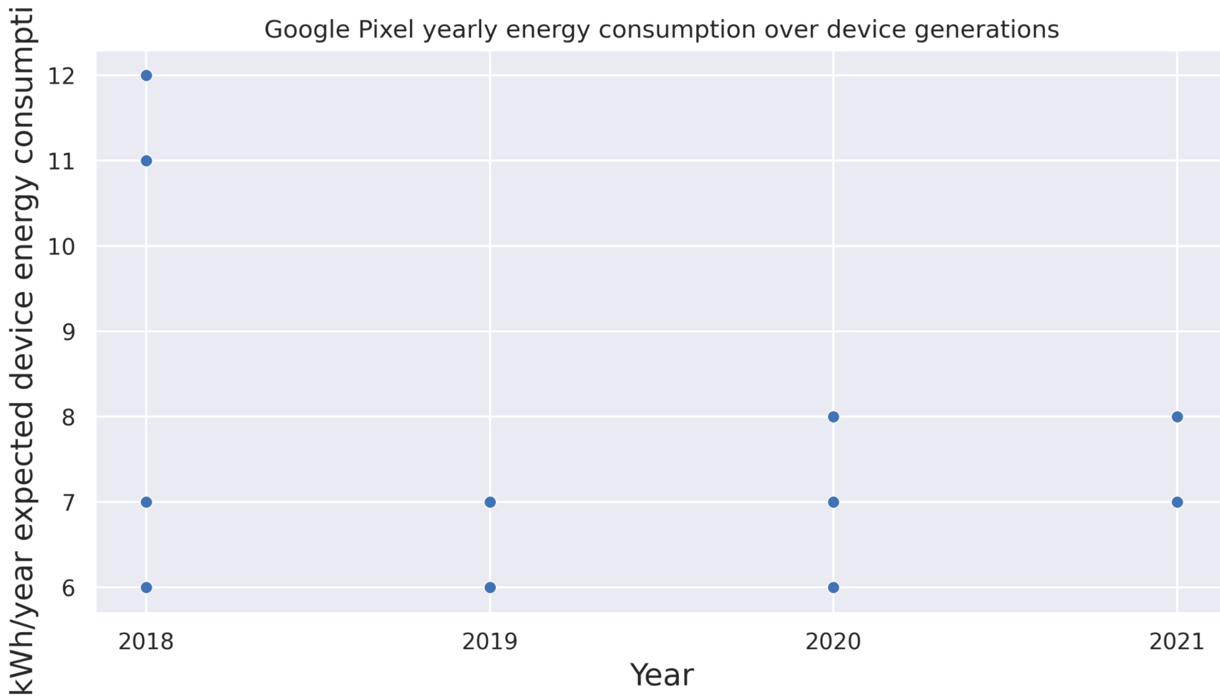

Figure 12. Google Pixel yearly energy consumption over device generations

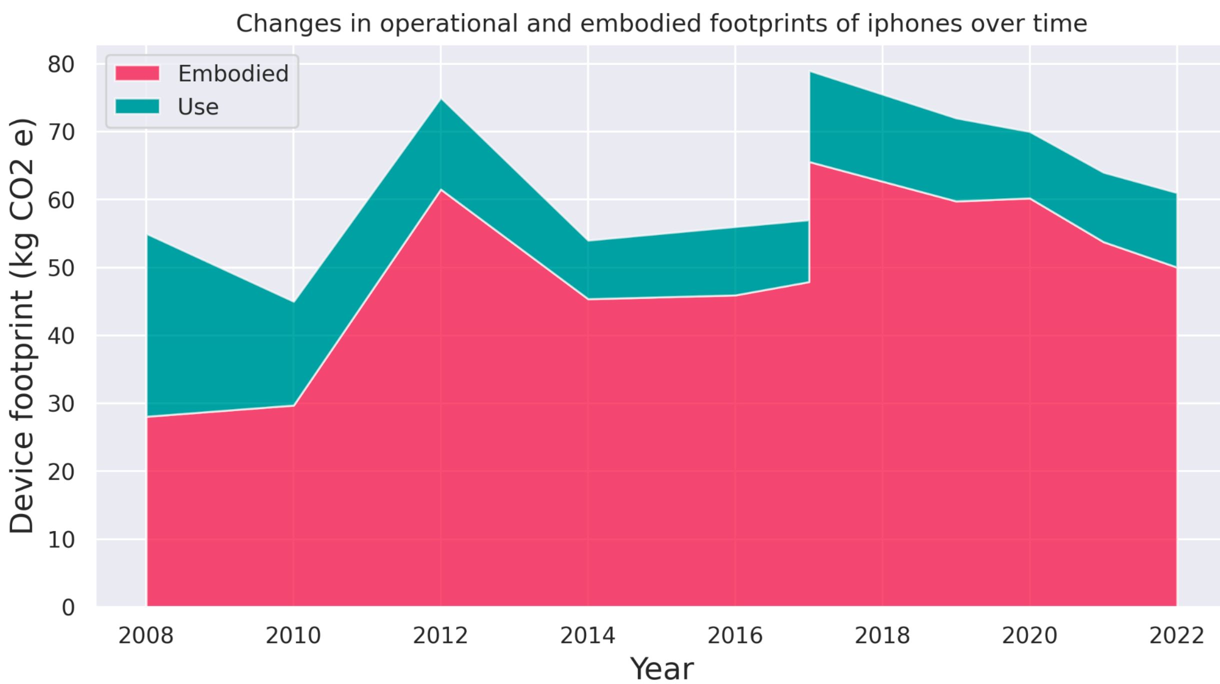

Figure 13. iPhone operational and embodied footprint changes over device generations

Device	Chip	CPU	GPU	Neural engine
8	A11 Bionic Chip	6-core	3-core	2-core

X	A11 Bionic Chip	6-core	3-core	2-core
11	A13 Bionic Chip	6-core	4-core	8-core
12	A14 Bionic Chip	6-core	4-core	16-core
13	A15 Bionic Chip	6-core	4-core	16-core
14	A15 Bionic Chip	6-core	5-core	16-core

Table 4. Increases in CPU, GPU and neural engine cores across iPhone device generations

Overall, with new technology generations comes consolidation and declining overall emissions on a per compute basis, but increasing demand can negate that emissions benefit and lead to higher operational and embodied carbon impacts if the increase in computing demand (whether across data centers or within a single smartphone device) erases the benefits that efficiency increases offer.

"Best case" estimation of global impact reduction

To demonstrate the potential benefits of escaping Jevons Paradox and capping computational growth to a rate that matches efficiency gains, we present an alternative 2030 data center scenario, in which the energy-use rebound effect on efficiency gains is limited to 100%, keeping 2030 data center energy usage at the same level as in 2021. This comes at the cost of limiting total data center computational work to a value below those that might otherwise be executed in 2030. Based on our initial assumptions of data center operational efficiency increases ranging between 5-15% annually (see Estimating operational footprints, 2030), we can estimate the cost of limited data center growth in terms of total datacenter traffic (EB). Under this restricted growth scenario, data center traffic (EB) is limited to between 21 and 59% of our baseline estimated data center traffic in 2030. We do not attempt to estimate the potential impact of this lever on smartphone emissions because of the non-linearity in smartphone energy consumption per device trends noted above.

By limiting the rebound effect of data center energy efficiency gains between 2021 and 2030 to 100%, we see a 28% decrease in total global data center footprints. This is likely an underestimate, representing a lower bound for total footprint benefits, as we are not considering the additional embodied carbon benefits that would come from decreased server demand with more limited computational demand growth.

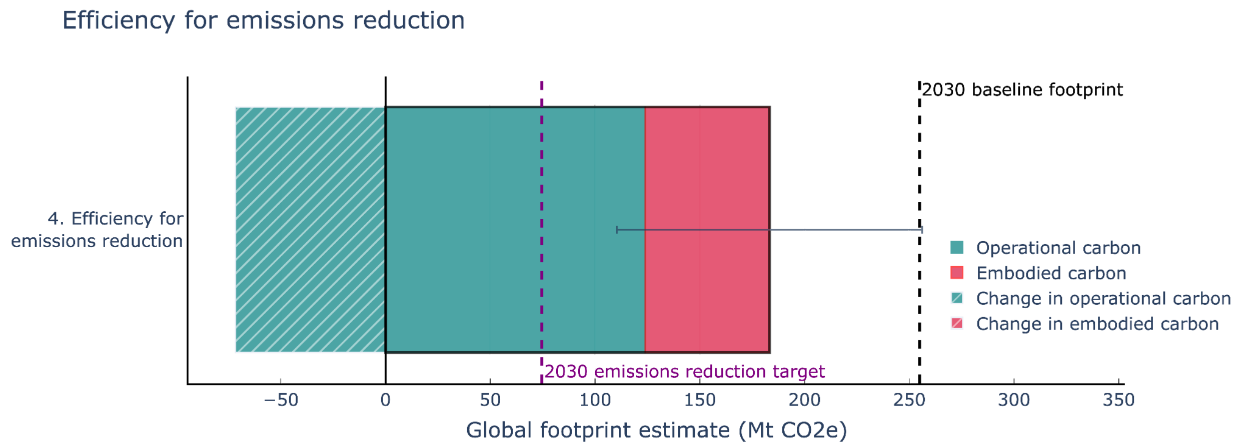

Figure 14. Best case impact estimate of limiting data center growth to efficiency gains on global data center (server) footprint in 2030

While this strategy appears to provide valuable reduction potential for data center operational emissions, in practice this kind of limitation on growth is difficult to achieve. Other work on this topic notes that rebound effects within the ICT sector are unlikely to end without significant intervention or changes in efficiency growth trends,[9] and that sustainability strategies oriented towards efficiency improvement are then liable to 'backfire'.[63] Rebound effects and Jevons Paradox could potentially be mitigated through public policy interventions within the ICT sector, such as carbon-taxes, cap and trade systems, or other regulations that aim to limit emissions growth.[64] However the effectiveness of these mechanisms at limiting rebound effects at a global scale remains to be seen.[64]

Additional opportunities for footprint reduction

The four strategies for ICT sector footprint reduction to 2030 presented in the previous sections are not meant to represent a comprehensive evaluation of all of the opportunities that exist to reduce sector impacts. Rather they represent interventions that are both widely discussed and that were practically modelable within this work given data availability. In order to represent some additional opportunities for reducing ICT sector impacts, here we discuss the potential for limiting direct emissions from fabrication and reducing overall computing demand via new learning technologies. We encourage further exploration of additional impact reduction strategies in future work.

Limiting direct emissions from fabrication

Direct emissions from the fabrication of integrated circuits for electronic devices are another important contributor to smartphone and server embodied footprints that both fabrication companies and the retail companies that they supply are interested in reducing. In 2021, SK Hynix reported that 23% of their total emissions were attributable to Scope 1 (encompassing direct emissions from their facilities),[30] while TSMC reported that direct emissions composed 13% of their corporate emissions.[29] While direct emissions of fabs may be lower than the emissions stemming from purchased electricity, they still comprise a sizeable portion of the sector's emissions and thus merit consideration of interventions to reduce those emissions as well. Among fab direct emissions, a recent McKinsey report cites that ~80% are attributable to process emissions.[43] Previously, the World Semiconductor Council (WSC) has set voluntary goals for reducing the process emissions intensity of semiconductor fabrication.[65] The latest goal was to achieve a Normalized Emissions Rate of 0.22 kg CO^2e/cm2 over the decade leading up to 2020 - representing a 30% reduction in emissions per cm2 from a 2010 baseline.[66] While a new PFC reduction goal was set to be announced by the end of 2021, we were unable to find any public communication of the new target.[67] In a recent communication with the US government, the Semiconductor Industry Association noted that there are technical challenges limiting progress in further reducing process emissions.[65] Reductions fab process emissions are generally achieved through process improvement, the use of alternative chemistries, gas abatement, and gas recycling.[43] However, the use of these strategies to reduce emissions can be constrained by the availability of floorspace in fabs for abatement equipment, difficulties in separating and purifying process-gas outflows, and the lack of known substitutes for some of the process gases.[65] Given that further reduction of direct emissions from fabs is an issue of technical feasibility, we do not attempt to estimate a best case for this lever in this study, but highlight this as an area of ongoing effort to reduce emissions by the industry. More data on the current status of global process emissions rates and reasonable medium-term reduction targets would be needed to generate an estimate of this lever at a similar level of detail to the others in this study.

The impacts of "new learning" on data center demand

As previously noted, the computational requirements for training state of the art artificial intelligence models have increased dramatically in recent years.[59] These large computational requirements in turn result in large carbon footprints associated with both ML training and inference.[61] While the technology to enable autonomous learning is not fully developed to date, this new machine learning paradigm could enable a significant reduction in the energy and computational costs of training artificial intelligence models.[68] Given that growth in AI computational needs is driving growth in data center infrastructure,[69] a fundamental change to the computational requirements for AI models through "new learning" could potentially reduce 2030 datacenter traffic demands and therefore total emissions. This topic should continue to be explored in the future as autonomous technologies develop to evaluate potential changes to the global footprint of AI technologies and data centers in general, and the extent to which the increase in efficiency provided though more energy efficient model training would be mitigated by increases in demand for AI according to Jevons paradox.

Pathways to achieve 2030 emission reduction targets

Finally, we evaluate the results of the four interventions proposed in the context of the 2030 proposed emissions reduction targets for both data centers (servers) and smartphones. Taken separately, none of the interventions that we present within this work, on average, achieve the desired ICT sector wide reduction targets for 2030 based on the assumptions of this model. However, the 2030 emissions reduction target for data centers (servers) does become achievable when we combine strategies.

When we combine renewable energy for datacenter (server) operation with low-carbon electricity for device fabrication and increased device lifetimes, we achieve an estimated total impact well below the 2030 emissions target. Notably, device lifetime extension is effective at reducing the total footprint of data centers (servers) in this scenario, as the negative operational energy backlash of using older servers longer is largely mitigated by the reduction in emissions per unit of operational electricity offered by developing renewable energy to power data center operations. Further, the embodied carbon savings from lifetime extension effectively mitigate the embodied carbon costs associated with developing that renewable energy and additional server capacity to enable low operational emissions. While these results demonstrate that the desired 2030 emissions reduction target could conceivably be met by exercising just these three interventions for impact reduction, actually achieving these results may not be feasible. For one, as mentioned previously in the discussion of using renewable energy to power datacenter operations, this would require a sizable scale-up in existing wind energy capacity to provide operational energy for the data centers,~500 MW. Further, as noted in the discussion of low carbon fabrication, the geopolitical feasibility of relocating fabrication facilities to areas with lower grid emission factors is of concern for fully realizing that reduction potential.

The requirements to achieve the 2030 emissions reduction target become more flexible when we are able to also harness energy efficiency increases for emissions reductions. We demonstrate an extreme example, where growth is limited to efficiency gains from 2021 to 2030 and therefore global data center energy demand remains constant over that time period. By exercising each intervention in combination, we achieve 2030 impacts that are again well below the 2030 emissions reduction target. Importantly, Including this restriction on data center growth reduces the infrastructure cost, and likely monetary costs, of achieving the 2030 targets. For example, the excess server capacity required to support the interventions is reduced from 32.1% to 7.8% when the energy efficiency intervention is present on top of the other strategies. However, as previously discussed, there is no established pathway for how we might exercise this lever within the sector.

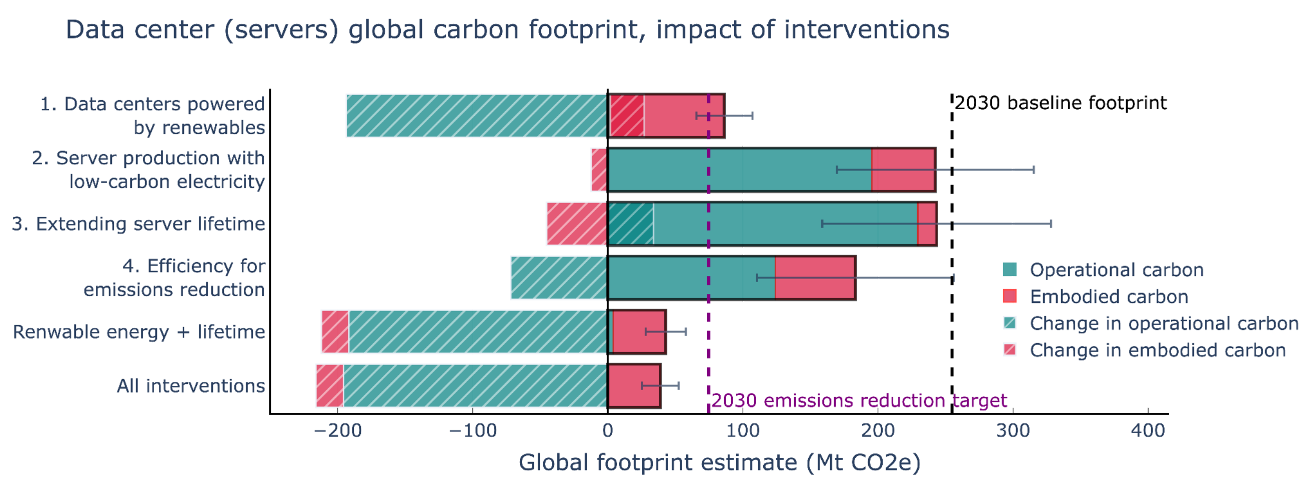

Figure 15. Comparison of "best case" estimates of interventions individually and in combination for data center (sever) global impact reduction

For smartphones, our results emphasize the need for more clear pathways for impact reduction of consumer devices. While utilizing low carbon electricity for smartphone production does improve overall emissions from smartphones, it is not enough to move the average impact across our simulations below the 2030 emissions reduction target for total smartphone impacts.

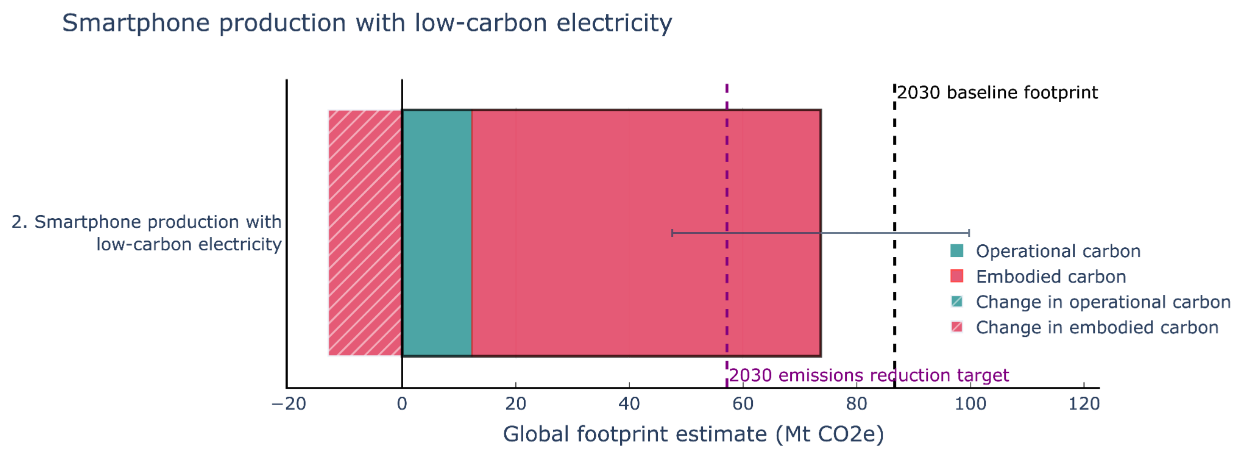

Figure 16. Result of modeled "best case" intervention for smartphones, low-carbon electricity for integrated circuit fabrication

Policy options for achieving ICT sector emissions reductions

In order to tie our proposed interventions back to a real world context, we include a discussion of existing policy levers that could help achieve the ICT sector's 2030 emission reduction targets. The review article Freitag et al. (2021)[9] provides a robust overview of policies oriented towards reducing ICT sector emissions. The authors highlight both governmental policies in Europe, including a commitment by the European Commission to carbon neutral data centers by 2030 and the New Circular Economy Action Plan "Circular Electronics Initiative" oriented towards improving device lifetimes, and self-regulation policies within the ICT industry, including corporate pledges to be carbon neutral, net zero, or carbon negative. The authors note the limitations of these existing policies, including the limited enforcement and incentives for industry compliance with targets for carbon neutral data centers as part of the European Green Deal, the varying degrees of supply chain coverage as part of corporate carbon neutrality goals, and differences in the additionality of different corporate renewable energy procurement strategies. In response to these limitations, the authors call for additional policy mechanisms that would allow achievement of ICT sector emissions targets, suggesting that this could take the form of a global carbon constraint implemented via a carbon tax or a cap on emissions.

To build upon the policy discussion in Freitag et al. (2021), we highlight additional corporate and European policies focused on the embodied emissions of computing devices. ICT sector companies are increasingly focused on quantifying and addressing the embodied emissions footprints associated with their products. Microsoft, for example, acknowledged the significance of its scope 3 emissions within its 2030 net-negative emissions pledge and set intermediate targets that included scope 3 emissions reductions goals.[70] One way that manufacturing emissions are accounted for on an ICT product level is through product carbon footprint evaluation and reporting. For example, Apple publishes product environmental reports of their products on their website that include carbon footprint numbers and comparisons to the previous device model[71] and Logitech recently began printing carbon footprint labels directly onto their device packaging.[72] Firms may be motivated to track the embodied carbon impacts of their products in order to comply with eco-labels, such as those from EPEAT[73] and the Carbon Trust,[74] or in order to self-report product carbon footprint data to customers. Discussions with ICT sector sustainability leaders indicate that customer requests for product carbon footprint information are increasing, and that customers are using this information to guide their purchase decisions.[75] Other motivations for product carbon footprint reduction may include progress towards announced climate goals, supply chain risk reduction, and marketing.[76] Carbon footprinting can also help a firm to identify solutions that increase product manufacturing efficiency which both save the firm money and reduce environmental impact.[76] Lastly, the expectation of future regulatory requirements is likely a large part of the motivating force behind product carbon footprinting efforts.[76]

To that end, two emerging European policies have particular relevance to the issue of device embodied carbon emissions: the EU digital product passport and the EU Carbon Border Adjustment Mechanism. The European Commission's digital product passport program will focus on tracking products across their supply chains with the idea that increased knowledge of the components and materials within a product can promote end of life recycling and reuse in accordance with circular economy principles.[77] The passport will “serve as an inventory of all materials, components and raw materials used in a product”[78] with information on each individual product and component, including their carbon footprints,[79] stored within a blockchain structure. In this way, the digital product passport system could enable consistent and comparable reporting of carbon footprints across products beyond what exists today. The digital product passport will be applied to many sectors, including ICT.[78] The European Commission announced that the passport would be introduced in early 2022,[77] however, implementation will likely take several years.[79] The EU Carbon Border Adjustment Mechanism (CBAM) is another emerging policy aimed at minimizing carbon leakage through imports of embodied product carbon emissions to the EU by applying EU carbon pricing to equalize the price of imported goods.[80] Importers will be required to purchase carbon certificates that correspond to the level of carbon tax that they would have paid if a product had been produced in the EU under carbon-pricing rules.[80] Determination of embodied carbon levels will require product carbon footprinting by all importers according to guidelines set by the European Commission.[81] While initial implementation (through 2026) will focus on select basic materials,[80] the mechanism will have impacts for industries using the products that are covered: including the high-tech goods and consumer appliances sectors.[82] Therefore the mechanism is relevant to ICT even in its current form, and could be expanded to directly apply fees to imported ICT product embodied carbon in the future. Although it is not yet clear when or to what extent these policies will impact ICT product footprinting and supply chain emissions reductions, taken together, the proposed programs offer potentially beneficial policy frameworks for standardization of product supply chain emissions tracking and embodied carbon disclosures, as well as an imported carbon pricing mechanism that could encourage ICT product footprint reductions.

Lastly, we touch upon opportunities for the development of US policies around sustainable ICT product manufacturing in the context of renewed investment in US semiconductor manufacturing as a result of the CHIPS and Science Act.[83] While the Act itself does not explicitly comment on or set goals for environmental sustainability within new investments in semiconductor manufacturing, this renewed cycle of investment and manufacturing capacity building is seen as an opportunity to reduce the negative impacts of manufacturing semiconductor devices.[84,85] The state of New York, in particular, is capitalizing on this opportunity with the passage of Green CHIPS Legislation that complements the goals of the federal CHIPS and Science Bill while making clean energy and carbon emissions mitigation programs a central part of the strategy for industry development.[86] Given the extent of interventions needed to achieve 2030 ICT sector emissions reduction goals, as outlined in this work, it is important that the rest of the US pursues similar sustainability oriented strategies when it comes to new investments in semiconductor fabrication, and that additional policies are developed to regulate and encourage reduction of emissions in other portions of the ICT sector supply chain, including data center operation.

In summation, while there are existing examples of policies aimed at reducing both embodied and operational impacts of the ICT sector, these policies are not consistent across geographies, and many policies are still undergoing implementation or represent voluntary commitments, which may complicate achievement of future emissions reduction goals.

Analysis limitations and takeaways

While the above analyses rely on numerous assumptions in order to estimate the relative impacts and limitations of different strategies for reducing the global carbon footprint of a few exemplary ICT devices, it is our hope that the overall framework we offer for the consideration of tradeoffs between operational and embodied carbon in regards to different interventions and approach for ordering interventions by potential impact can be reapplied to more specific use cases in order to orient ICT sector company strategies along different portions of their supply chains.

While each individual assumption in our analysis may be liable to impact over or underestimation, we believe there is some value in the overall ordering of magnitude of interventions into ICT device supply chains as presented in these results. For data centers (servers), we reported the largest potential decrease in overall footprint as stemming from interventions on the operational emissions component of the total footprint. Notably, we demonstrate that if we were able to harness renewable energy for data center operations and/or energy efficiency for impact reduction, global data center impacts could be greatly reduced. Lowered grid emissions factors for embodied footprint reduction offered comparably lower benefits in terms of carbon footprint reduction as compared to either of the operational energy interventions. Finally, we show that the effectiveness of server lifetime extension as a means for data center (server) segment impact reduction depends on the operational interventions that have already taken place - this intervention was much more effective at reducing the overall footprint of the data center (server) segment once operational energy emissions were already successfully reduced by cleaner energy sources (as in the combined intervention scenarios).

For smartphones, we only model one intervention - renewable energy for embodied emissions reduction, because the smartphone lifecycle is fundamentally different from the lifecycle of a server in a datacenter, limiting utilization of the same strategies for impact reduction. For example, we do not include renewable energy for operation as a viable lever for smartphone impact reduction, as where consumer devices are used and charged is outside of ICT sector control. Interventions on the lifetime of smartphone use are also determined by individual consumers and ICT firms may be motivated, from a business perspective, to reduce smartphone lifetimes rather than extend them in order to sell more products. Finally, energy efficiency could be used as a lever for smartphone impact reduction, but the impact of smartphone operation is already much smaller than the emissions associated with making the device, and the trend towards less energy usage per smartphone is not present as it is with servers, indicating that again business decisions by designers to increase usefulness of a device may fall contrary to the use of increased integrated circuit efficiency for overall reductions in energy use. It would be interesting to explore the potential to harness device energy efficiency trends for impact reduction further in the future with input from smartphone manufacturers. The intervention we do model - the use of lower carbon grid emissions energy to power smartphone fabrication effectively decreases smartphone emissions, reducing smartphone overall emissions further than data center (server) emissions are reduced with the same lever.

Evaluating sources of baseline model uncertainty

There is a large amount of uncertainty associated with our baseline estimates of global smartphone and datacenter (server) impacts for 2021 and 2030. We conduct an analysis of variance in order to determine the parameters contributing most to this uncertainty and identify where additional data collection and refinement could reduce uncertainty in the future.

2021 estimates

By calculating the first-order sobol indices for each parameter relating to our baseline model outputs, we can see that the variation in our estimate for the 2021 operational footprint of data centers (servers) results from the variance of both input variables. When we look at the variation in the server embodied carbon footprint estimate for 2021, the variance is dominated by uncertainty around the manufacturing emissions per server parameter. Finally, when we look at the variability of the total footprint for datacenter (servers) in 2021, the two operational impact variables, 2021 data center energy usage and 2021 global electricity emissions intensity, are the primary sources of variability in that final estimate.

Parameters contributing to 2021 data center (server) baseline estimates	First order Sobol index values		
	Operational carbon variance	Embodied carbon variance	Total footprint

	contributors	contributors	
2021 global data center energy use	0.53		0.51
2021 global electricity emissions intensity	0.45		0.43
2021 global new servers shipped		0.12	0.0
2021 estimate of server manufacturing emissions		0.82	0.0

Table 5. Decomposition of variance in 2021 data center (server) global footprint estimate using first order Sobol indices

For smartphones, the uncertainty in our 2021 operational footprint estimate results firstly from uncertainty around the number of global active smartphones, with sizable uncertainty contributions stemming from variability in the smartphone energy usage and 2021 global electricity emissions intensity parameters as well. When we look at the variation in the smartphone embodied carbon footprint estimate for 2021, the variance is dominated by uncertainty around the manufacturing emissions per smartphone parameter. Finally, when we look at the variability of the total footprint for smartphones in 2021, the three operational impact variables, 2021 global active smartphones, 2021 smartphone energy usage, and 2021 global electricity emissions intensity, are the primary sources of variability in that final estimate.

Parameters contributing to 2021 smartphone baseline estimates	First order Sobol index values		
	Operational carbon variance contributors	Embodied carbon variance contributors	Total footprint
2021 global active smartphones	0.53		0.53
2021 smartphone energy usage	0.33		0.33
2021 global electricity emissions intensity	0.15		0.15
2021 smartphone manufacturing emissions estimate		0.98	0.0
2021 new smartphones shipped		0.02	0.0

Table 6. Decomposition of variance in 2021 smartphone global footprint estimate using first order Sobol indices

2030 estimates

We perform the same Sobol first-order index calculations for our input parameters and baseline estimates for 2030. We find that our 2030 estimate of the global datacenter (server) operational footprint is dominated by uncertainty surrounding the expected yearly data center energy efficiency improvements parameter (followed by uncertainty around the 2030 grid emissions intensity), while the global datacenter (server) embodied footprint estimate for 2030 is dominated by uncertainty surrounding the 2021 server manufacturing emissions parameter. Our estimate of the total 2030 datacenter (server) global footprint can be primarily attributed to the expected yearly data center energy efficiency improvements parameter, with additional contributions to variability stemming from the 2030 average grid emissions factor parameter and, to a lesser extent, 2021 data center energy demand.

Parameters contributing to 2030 data center (server) baseline estimates	First order Sobol index values		
	Operational carbon variance contributors	Embodied carbon variance contributors	Total footprint
2030 global electricity emissions intensity	0.28		0.25
2021 datacenter traffic (EB)	0.0		0.0
Expected yearly data center efficiency improvements	0.88		0.79
Data center energy demand 2021	0.08		0.07
New servers shipped 2030		0.11	0.0
2021 estimate of server manufacturing emissions		0.88	0.0
% of server emissions from manufacturing		0.02	0.0
% of manufacturing emissions from electricity		0.0	0.0
Grid emissions factor of production location		0.0	0.0
2030 average grid emissions factor		0.01	0.0

Table 7. Decomposition of variance in 2030 data center (server) global footprint estimate using first order Sobol indices

Our 2030 estimate of the global smartphone operational footprint is composed of uncertainty deriving from all three input parameters, while our 2030 estimate of the global smartphone embodied footprint is mainly attributed to two input parameters, 2021 smartphone manufacturing emissions, primarily, and the number of new smartphones shipped in 2030, secondarily. Our total 2030 footprint estimate for smartphones again derives uncertainty primarily from the three operational footprint input parameters: 2030 global active smartphones, 2021 estimate of smartphone yearly energy usage, and 2030 global electricity emissions intensity.

Parameters contributing to 2030 smartphone baseline estimates	First order Sobol index values		
	Operational carbon variance contributors	Embodied carbon variance contributors	Total footprint
2030 global active smartphones	0.28		0.27
2021 smartphone energy usage	0.17		0.17
2030 global electricity emissions intensity	0.29		0.29
New smartphones shipped 2030		0.29	0.0
Smartphone manufacturing emissions, 2021 est		0.56	0.0
Proportion of smartphone emissions from IC manufacturing		0.0	0.0
% of manufacturing emissions from electricity		0.0	0.0
Grid emissions factor of production location		0.0	0.0
2030 average grid emissions factor		0.0	0.0

Table 8. Decomposition of variance in 2030 smartphone global footprint estimate using first order Sobol indices

In summation, we see larger contributions to the uncertainty to our total global footprint estimates for both smartphones and data centers (servers) stemming from uncertainty associated with the operational footprint input parameters than the embodied footprint input parameters. Within our embodied footprint estimates for both datacenters (servers) and smartphones in 2021 and 2030, the uncertainty in the per device manufacturing emissions parameters stands out as the largest contributor to estimate variability. Efforts to reduce the uncertainty surrounding these parameters would aid in future global impact modeling efforts as well as further refinement of the model presented in the following sections.

A case study exploration of device specialization as a means for emissions reductions

In this next section, we utilize a detailed case study of a prototypical analog accelerator device to explore the emissions reduction potential and tradeoffs associated with adoption of emerging computing technologies. Within the prior discussion around energy efficiency, we highlight the trend of device specialization as a mode for achieving higher levels of energy efficiency with new technology adoption. While the operational efficiency benefits of specialized devices are increasingly well documented,[58] the impact of device specialization on the embodied carbon impacts of computing systems requires further exploration.

Life cycle analyses of computing technologies are often performed after a technology achieves widespread adoption - which may limit the design options available to reduce the impact of a device. Prospective life cycle analyses, or life cycle analyses performed on emerging technologies, are useful both as an input to inform the evolution of a design,[14] and as a way to assess potential future system benefits of new technology adoption.[87] ICT sector interest in the Interuniversity Microelectronics Centre's (imec) Sustainable semiconductor technologies and systems (SSTS) project demonstrates a growing recognition of the value of prospective sustainability assessments for new computing technologies.[88] While that effort is making great strides towards assessing the embodied impacts of computing technologies at more scaled technology nodes,[11] its focus is primarily on CMOS technology scaling with Moore's law rather than new technologies that diverge from that paradigm, as with the trend of device specialization. Therefore, the goal of this analysis is twofold: 1) to develop a framework for assessing the embodied versus operational emissions tradeoffs that are associated with specialized device adoption, and 2) to provide an example of how the embodied impact of a prototypical computing device can be estimated prior to technology maturation & industrial-scale fabrication.

Developing a process model of a prototypical analog accelerator device

We select an experimental analog accelerator device recently fabricated by researchers at MIT as the basis for our analysis. We obtain information on fabrication processes for the device prototype from their published work[89] and patent filing[90], with additional fabrication details supplied through communications with the researchers. Table 9 below provides an overview of the main process steps required.

	Reported fabrication steps + *additional notes from researchers*	**Process type**
1	(*diesaw + spincoat + wafer clean step*) + Atomic Layer Deposition (ALD) of 10/40 nm HfO_2/Al_2O_3 on 1x1 cm2 SiO_2/Si pieces.	Deposition
2	(*wafer clean step*) + Patterning of poly(methyl methacrylate) (PMMA, e-beam resist) with Elionix FLS-125 for channel layer lift-off. + *(development of resist + SEM inspection)*	Lithography
3	Reactive sputtering of WO_3 layer from metallic target at room temperature in O_2/Ar RF plasma using AJA sputtering system. + (*liftoff)*	Deposition
4	Annealing of the WO_3 layer in 8:2 $N_2:O_2$ environment at 400 °C for 1 hour following a liftoff step.	Annealing

5	(*wafer clean step + spin coating)* Patterning of poly(methyl methacrylate) (PMMA, e-beam resist) with Elionix FLS-125 for source/drain contact layer lift-off.	Lithography
6	Electron-beam evaporation of 35/5 nm of Au/Cr layer using AJA evaporation system, followed by lift-off step. + *(liftoff + SEM inspection)*	Deposition
7	Plasma-Enhanced Chemical Vapor Deposition (PECVD) of PSG layer using 1420 sccm N2O, 12 sccm SiH4, and 12 sccm PH3 (2% in H2) at 100°C, with a RF plasma power of 60 W at 380 kHz.	Deposition
8	(*wafer clean step + spin coating) +* Patterning of poly(methyl methacrylate) (PMMA, e-beam resist) with Elionix FLS-125 for gate contact layer lift-off.	Lithography
9	Electron-beam evaporation of 10 nm of Pd layer using AJA evaporation system, followed by lift-off step.+ *(liftoff + SEM inspection)*	Deposition
10	Reactive Ion Etching (RIE) of the PSG layer using Pd layer as the hard mask, under CF4 plasma at 100W.	Etch
11	(*SEM inspection + cleaning + spincoat)* Patterning the bilayer of poly (methylglutarimide) and Microposit S1813 positive tone photoresist, using Heidelberg-MLA 150 for pad layer lift-off.	Lithography
12	Electron-beam evaporation of 150/15 nm of Au/Cr layer using AJA evaporation system, followed by lift-off step. + (*liftoff + SEM inspection)*	Deposition

Table 9. Process steps required for lab-scale analog accelerator device fabrication on top of a finished CMOS wafer

One challenge with modeling a prototypical device is that the processes used to manufacture the device at a laboratory scale are not necessarily the same as those that would be used in large-scale production. In order to estimate an at-scale embodiment of production of the device, we make several assumptions on modifications to the published process flow based on conversations with the researchers, including that: 1) liftoff steps would be replaced by etch steps and followed by a stripping of the photoresist and clean step, 2) e-beam lithography would be replaced by photolithography.

Table 10 provides an overview of the estimated at-scale process steps, along with the general process category that each step falls into.

Step	Process Step Category	Description
1	Wafer Cleaning	Cleaning pieces: acetone sonication (5m), IPA sonication (5m)
2	Deposition	Atomic Layer Deposition (ALD) of 10/40 nm HfO2/Al2O3 on 1x1 cm2 SiO2/Si pieces.
3	Wafer Cleaning	Cleaning with Acetone+Methanol+IPA and Spincoating PMMA
4	Photolithography	Patterning of poly(methyl methacrylate) (PMMA, e-beam resist) with Elionix FLS-125 for channel layer lift-off.
5	Deposition	Reactive sputtering of WO3 layer from metallic target at room temperature in O2/Ar RF plasma using AJA sputtering system.
6	Etch	Liftoff
7	Strip	Stripping of photoresist following etch step
8	Wafer Cleaning	Cleaning step
9	Oxidation and Annealing	Annealing of the WO3 layer in 8:2 N2:O2 environment at 400 °C for 1 hour following a liftoff step.
10	Wafer Cleaning	Cleaning with Acetone+Methanol+IPA and Spincoating PMMA
11	Photolithography	Patterning of poly(methyl methacrylate) (PMMA, e-beam resist) with Elionix FLS-125 for source/drain contact layer lift-off.
12	Deposition	Electron-beam evaporation of 35/5 nm of Au/Cr layer using AJA evaporation system, followed by lift-off step.
13	Etch	Liftoff

14	Strip	PR Strip
15	Wafer Cleaning	Cleaning step
16	Deposition	Plasma-Enhanced Chemical Vapor Deposition (PECVD) of PSG layer using 1420 sccm N2O, 12 sccm SiH4, and 12 sccm PH3 (2% in H2) at 100°C, with a RF plasma power of 60 W at 380 kHz.
17	Wafer Cleaning	Cleaning with Acetone+Methanol+IPA and Spincoating PMMA
18	Photolithography	Patterning of poly(methyl methacrylate) (PMMA, e-beam resist) with Elionix FLS-125 for gate contact layer lift-off.
19	Deposition	Electron-beam evaporation of 10 nm of Pd layer using AJA evaporation system, followed by lift-off step.
20	Etch	Liftoff
21	Strip	PR Strip
22	Wafer Cleaning	Cleaning step
23	Etch	Reactive Ion Etching (RIE) of the PSG layer using Pd layer as the hard mask, under CF4 plasma at 100W.
24	Wafer Cleaning	Cleaning with Acetone+Methanol+IPA and Spincoating PMGI/S1813
25	Photolithography	Patterning the bilayer of poly (methylglutarimide) and Microposit S1813 positive tone photoresist, using Heidelberg-MLA 150 for pad layer lift-off.
26	Deposition	Electron-beam evaporation of 150/15 nm of Au/Cr layer using AJA evaporation system, followed by lift-off step.
27	Etch	Liftoff
28	Strip	Strip photoresist
29	Wafer Cleaning	Cleaning step

Table 10. Approximate process steps required for analog accelerator device fabrication on top of a finished CMOS wafer at scale

To estimate a fully functional version of the analog accelerator with electrical connections to the underlying CMOS transistors, we include additional process steps beyond initial device fabrication for via formation and filling. The accelerator device patent described an additional proton barrier layer that can be deposited over the finished device, into which vias may be etched to allow for electrical connection to the device contacts.[90] Using this information, a via etch and fill process described in the patent of a similar RRAM device[91] and the process flow for interconnect formation reported in Krishnan et al. (2008)[92] we estimate the additional required steps for electrical interconnection of the device (Table 11).

30	Deposition	Deposit proton barrier layer
31	Photolithography	Pattern a contact via in proton barrier layer
32	Etch	Etch contact via
33	Strip	PR Strip
34	Wafer Cleaning	Wafer clean
35	Deposition	Ti Liner
36	Deposition	TiN Barrier
37	Deposition	Deposit Cu seed layer
38	Electro-chemical plating	Cu ECP
39	Oxidation and Annealing	Cu Anneal
40	Chemical Mechanical	Cu CMP and Clean

	Planarization (CMP)	

Table 11. Additional process steps required for electrical interconnection of the analog accelerator device

Because of the lack of high quality LCA data for many specific fabrication processes and the remaining uncertainty over the exact future embodiment of this analog accelerator device at-scale (specific processes used and materials are subject to change to some degree), we use an underspecified approach to estimating the manufacturing impact of this device based on the general process category for each step.

Using the process inventory tables for various fabrication steps reported in Krishnan et al. (2008), we assign each individual process recorded within that work to a more general process step category: Wafer Cleaning, Deposition, Photolithography, Etch, Strip, Chemical Mechanical Planarization (CMP), Electro-chemical plating, and Oxidation and Annealing. Then, we estimate the global warming impact (g CO2 eq) of each individual inventory step using emissions factors reported in Boyd (2012).[10] By aggregating the individual steps by their general process category, we obtain a distribution that we can sample from to obtain an underspecified impact estimation based on the process flow for the analog accelerator device.

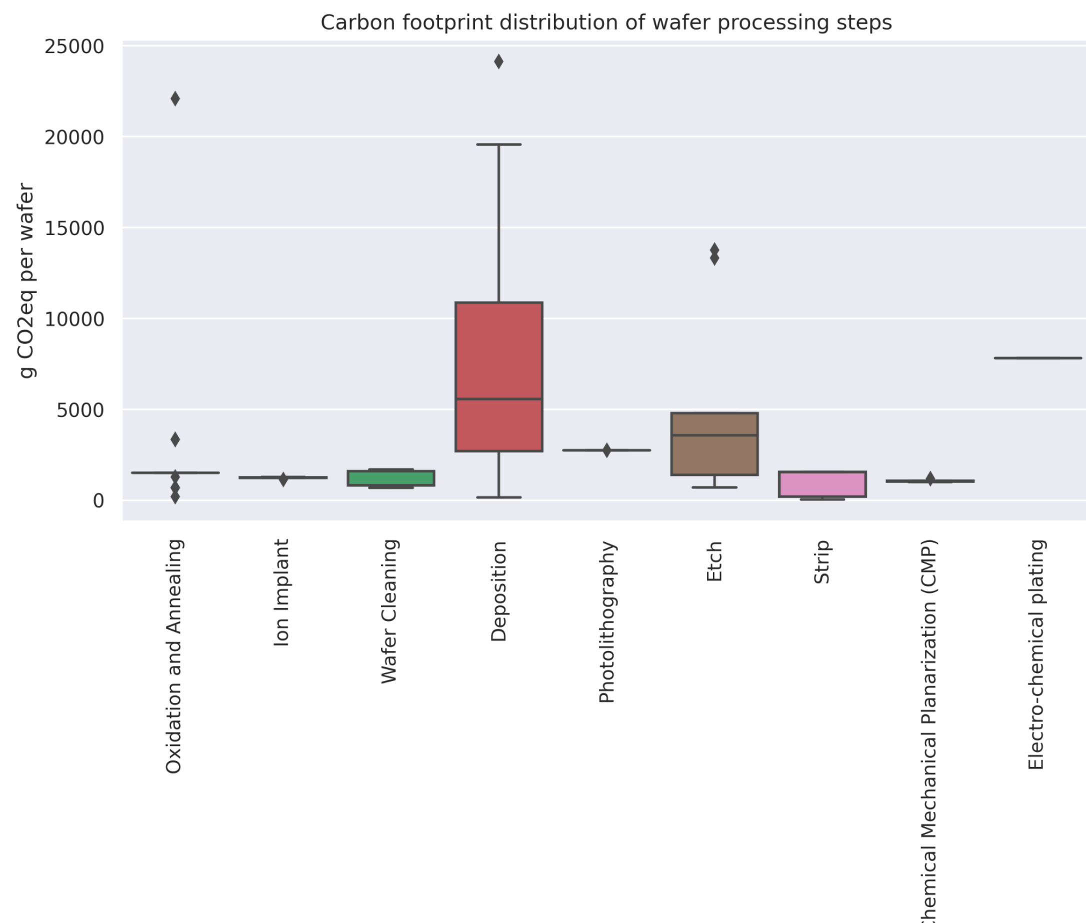

Figure 17. Carbon footprint distribution of aggregated wafer fabrication process categories

To simulate a range of potential impacts for the analog accelerator device, we run 100 trials, sampling from log normal distributions based on the mean and standard deviations of each aggregated process step category. The result is a range of estimated total impacts for the device at the wafer level.

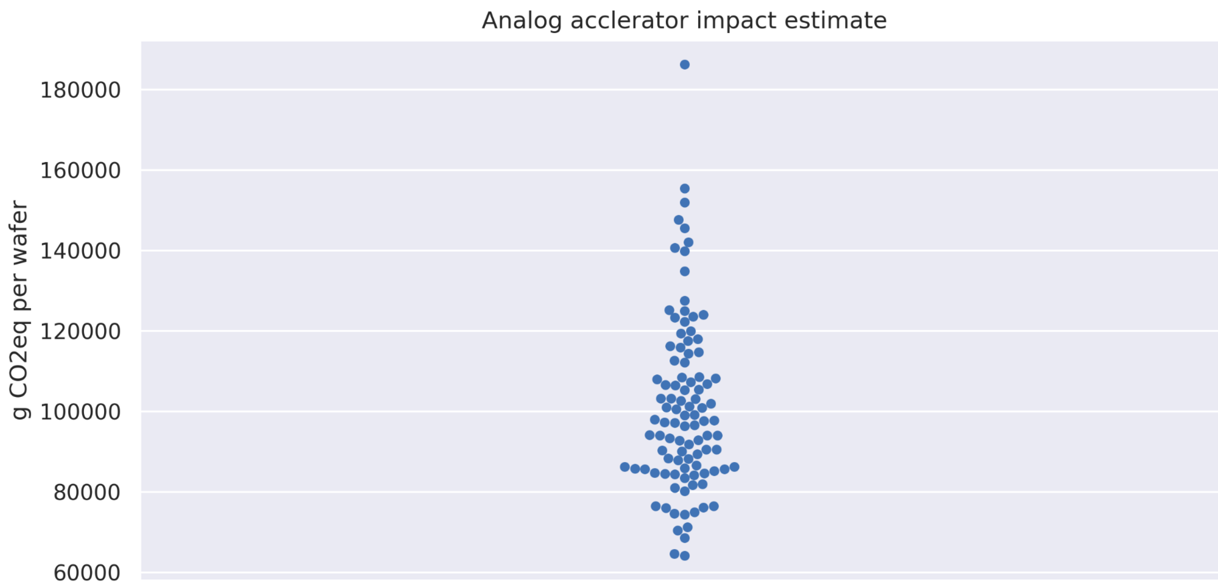

Figure 18. Simulated impact of additional impact per wafer on top of CMOS production required for analog accelerator fabrication

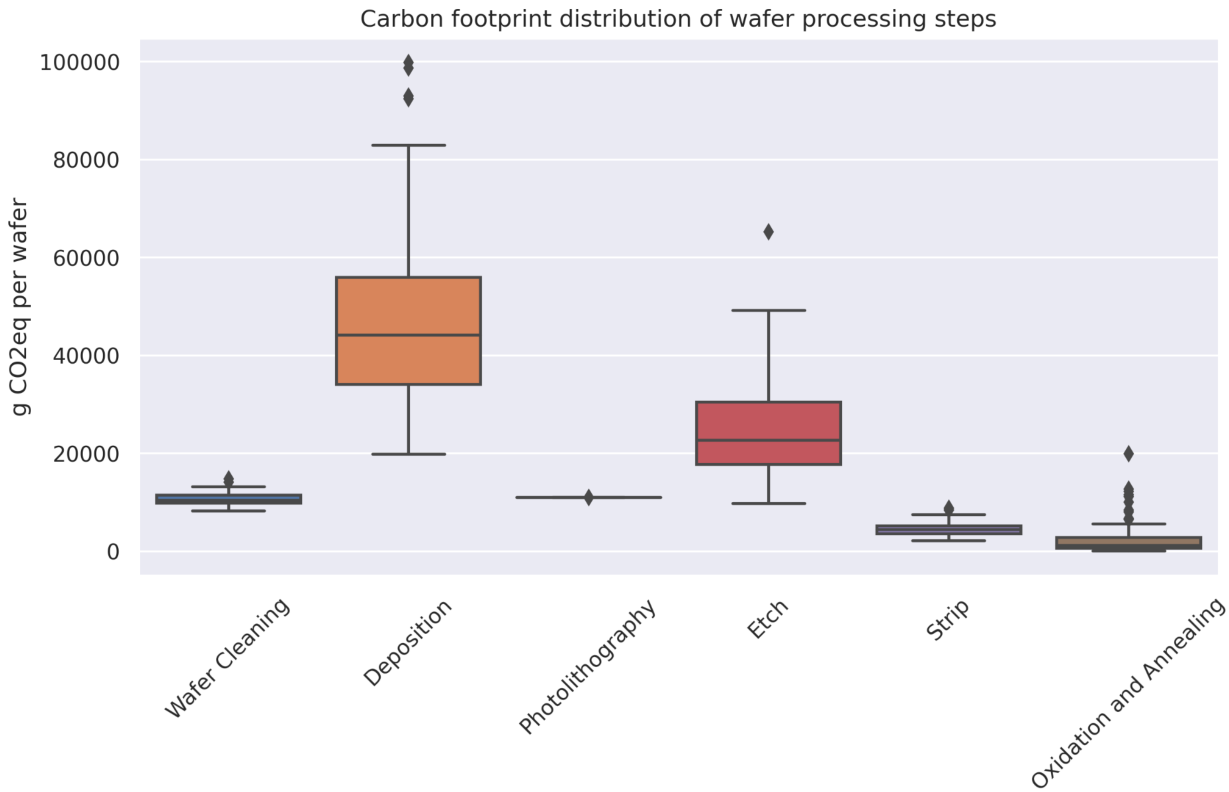

Figure 19. Simulated impact of additional impact per wafer on top of CMOS production required for analog accelerator fabrication broken down by process category

As shown in Figure 19, the deposition steps are responsible both for the highest greenhouse gas impacts from device fabrication and are a significant contributor to the overall variance of the impact estimate. The simulations result in an average impact of 151.051 kg CO^2e/wafer with a standard deviation of 29.574 kg CO^2e/wafer.

Evaluating tradeoffs in device adoption

Now that we have an estimate of the additional manufacturing impact required for analog accelerator device fabrication on top of CMOS processes, we can use that estimate, along with various other parameters, to specify the operational efficiency benefits that the analog accelerator device would need to have over a comparable digital accelerator device in order to have a lower life cycle impact. For this analysis we assume an embodiment of the analog accelerator device with identical fabrication impacts to a digital accelerator apart from the additional process steps as outlined in the above section. The lifecycle benefits of analog accelerator adoption will depend on contextual information like the degree of utilization of the device, the expected device lifetime, the operational electricity emissions factor, any area scaling benefits of the new device as compared to what it is replacing (for this example, a digital accelerator device), and any change in die yield resulting from the additional fabrication steps. We can use these variables to construct a tradeoff space that will address the question of when the additional manufacturing impact required to produce an analog accelerator device is outweighed by use-phase efficiency benefits.

We propose the following formula to determine the minimum workload operational efficiency gain required for lifecycle emissions to be reduced by adoption of this new accelerator device in place of digital accelerator capacity:

digital_operational = digital_active_power*exp_lifetime*activity_ratio*elec_impact
digital_embodied = wafer_impact/dies_per_wafer_digital
analog_embodied = (wafer_impact+additional_acc_embodied)/
(dies_per_wafer_analog*analog_specific_die_yield)
analog_op_required for breakeven = digital_operational + digital_embodied - analog_embodied

Where:

- additional_acc_embodied refers to additional manufacturing impact required to manufacture the analog accelerator device as compared to a digital accelerator, which we have estimated above
- wafer_impact refers to the manufacturing impact of a CMOS wafer, derived from previous estimate of impact at 32nm: GWP of 84 kg CO^2e/die * 347 chips/wafer = 29148 kg CO^2e/wafer[10]
- dies_per_wafer_analog is intended to capture any area scaling cost or benefit that the analog accelerator device might have over a digital accelerator
- analog_specific_die_yield refers to the change in the proportion of viable finished dies out of all dies fabricated attributable to the additional process steps required for analog accelerator fabrication as opposed to digital accelerator fabrication
- digital_active_power refers to the average power usage by the digital accelerator device when in use
- activity_ratio refers to the proportion of time that the accelerator is actively in use
- elec_impact refers to the emissions factor of the electricity used to power the device
- exp_lifetime refers to the expected operating lifetime of the device, assumed to be the same for both the analog and digital accelerator

Using our estimate of the additional analog accelerator embodied impact from the above section (~151.051 kg CO2eq/wafer), we utilize this formula to explore how an operator's decision to adopt an analog accelerator technology as a way to lower the carbon footprint of a system might change depending on the circumstances of device use.

We utilize characteristics of the Google TPU v1 digital accelerator[93] as a baseline for comparison with reported characteristics of two leading hypothetical analog accelerator chip designs (Newton[94] and ISAAC[95]).

Chip	Accelerator Types	Density (TOPS/mm2)	Efficiency (TOPS/W)	Process Technology	Precision	Die area	Power
Newton	Analog (NVM based)	0.68	0.92	32 nm	16-bit	Not specified	
ISAAC	Analog (NVM based)	0.47895	0.6275	32 nm	16-bit	85.4 mm2	
Google TPU v1	Digital	0.07	0.51	28 nm	16-bit (scaled from 8-bit)	330 mm2	45 W

Table 12. Characteristics of proposed analog accelerator chip designs in comparison to a digital accelerator.

First, we look at how a decision to adopt a new analog accelerator device might change based on the emissions factor of the electrical grid where the device is used.

When the electricity emissions factor is lower (as in the case of Arizona below), we see that the the decision boundary is more strict in terms of the area penalty that an analog accelerator device could have over a digital accelerator device, and is more forgiving towards operational efficiency penalties when there is an area benefit of the analog accelerator over a digital one.

Overall, the Newton and ISAAC analog accelerator designs, based on the parameters reported, are well outside of the decision boundary for preferring an analog accelerator to a digital accelerator for all electricity grid emissions factors and utilization rates we simulate.

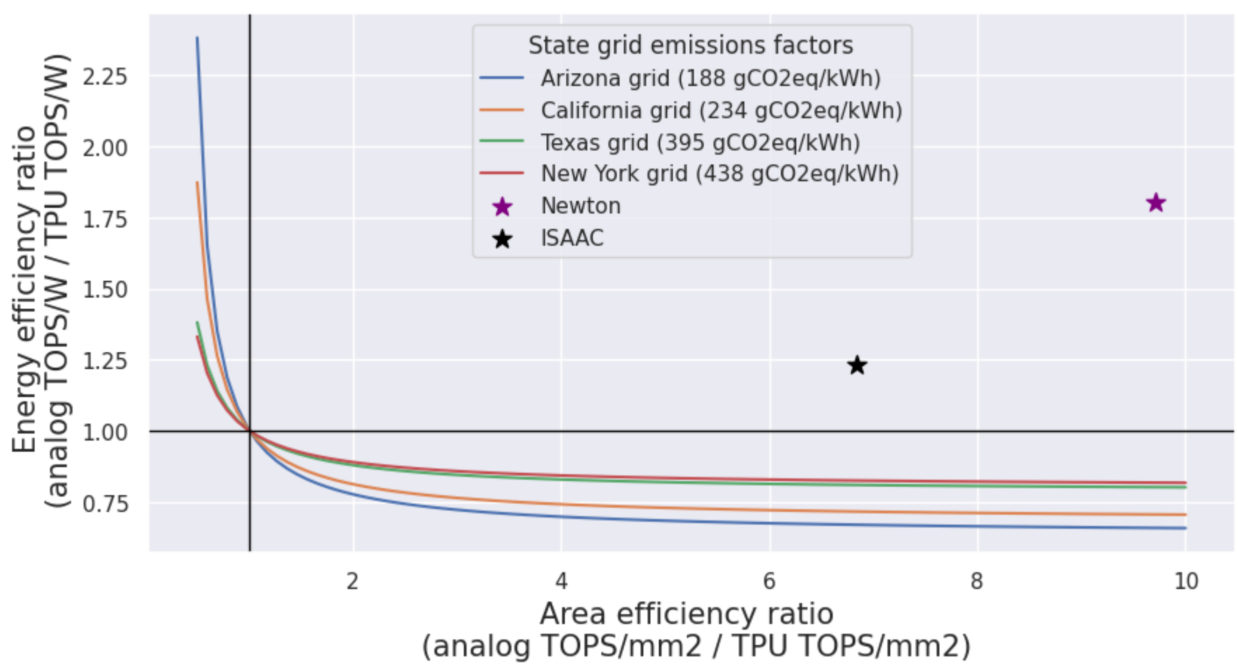

Figure 20. Plot of breakeven boundaries to determine total carbon footprint preference for an analog accelerator device over a digital accelerator, taking embodied and operational impacts into account under varying grid emissions factors

When we look at how the decision boundary changes with different device utilization rates, we see similar trends, with lower utilization rates having more strict area requirements and more lenient requirements for energy efficiency improvements, while the opposite holds true for higher utilization rates.

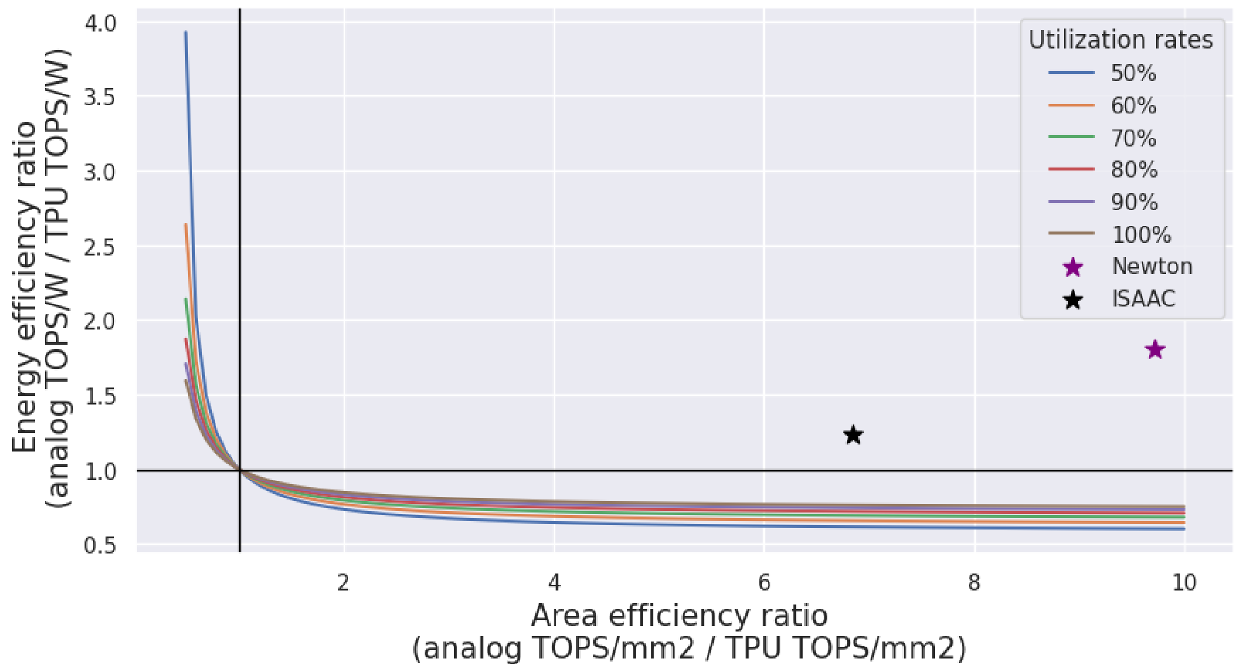

Figure 21. Plot of breakeven boundaries to determine total carbon footprint preference for an analog accelerator device over a digital accelerator, taking embodied and operational impacts into account under utilization rates

While the Newton and ISAAC analog accelerator designs make the case for analog accelerator devices being clearly better in terms of total carbon footprint impacts over digital accelerator devices, it is important to note the performance gap between these modeled designs and the characteristics of non-volatile memory analog accelerator devices that have been successfully fabricated. A recent review of analog non-volatile memory accelerator designs notes that both the energy efficiency and compute densities of fabricated eNVM-based IMCs (similar to the analog device we model) tend to be lower than comparable digital accelerators.[96]

Given this gap between realized and hypothetical area and energy efficiency benefits of analog accelerator devices as compared to digital accelerators, the tradeoff space that we construct here will be useful in framing ongoing evaluations of the benefits of emerging analog accelerator devices as their designs continue to evolve.

In addition, our analysis allows us to highlight when different decisions might be made in terms of deciding whether an analog accelerator device is more or less carbon intensive over its lifetime as compared to a digital accelerator device. In the Figure 22 below, we demonstrate this with a more zoomed in look at the difference in the tradeoff boundaries when the additional manufacturing processes are taken into account versus when they aren't (with 80% utilization and CA grid emissions intensity).

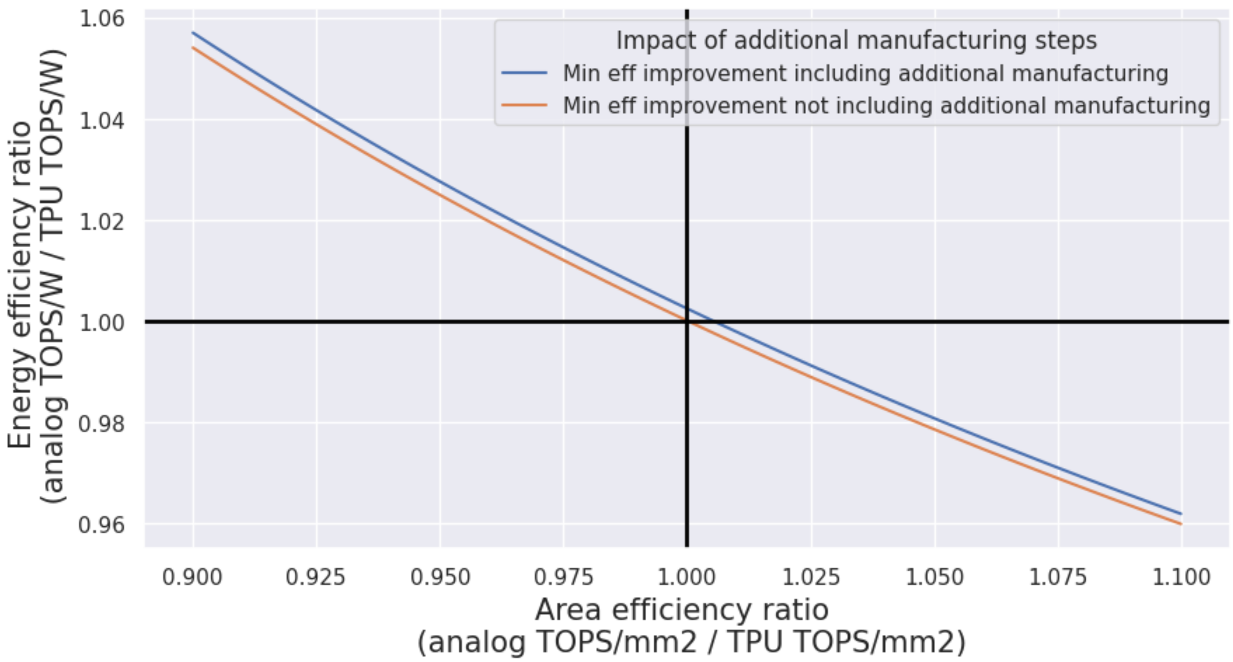

Figure 22. Change in breakeven boundary for total carbon footprint preference for analog accelerator over digital accelerator resulting from inclusion of additional embodied carbon emissions estimate in decision making

Having an estimate of the additional manufacturing impact of a new analog accelerator device matters most when you are close to the decision boundary, or more generally, when the computational density of the analog accelerator is lower in comparison to the digital accelerator it is replacing (more area = more additional cost), the energy efficiency of the analog accelerator is very close to the digital accelerator (eg <1% apart) or when one of the variables improves upon digital accelerators (computational density or energy efficiency) but the other does not. Likewise, the additional embodied energy required is less important for decision making when the energy efficiency benefits of the analog accelerator exceed 1% and the analog accelerator device features a similar computational density (area scaling =1), or neither computational density nor energy efficiency of the analog accelerator is an improvement over the digital accelerator you are comparing to. Further, while not visually apparent in this chart, the additional manufacturing impacts of the analog accelerator device matter more (i.e. the lines below are further apart) when the area efficiency of the device is lower (i.e. more area is required to match the throughput of a digital accelerator) as compared to when the analog accelerator device is more area efficient.

Future work should more closely examine some of the variables we have included here but not explored, including the impact of additional processing steps on yield. In addition, this framework should continue to be updated as analog accelerator technologies continue to advance in order to help technology adopters to understand when total footprint “carbon parity” between fabricated analog and digital accelerators is reached and re-implemented to aid in decision making around other emerging computing technologies.

Summary of contributions

In the first portion of this work, we highlight and attempt to quantify some of the existing opportunities to reduce the impact of the ICT sector with a primary focus on strategies to reduce the impacts of large scale data centers. We find that 2030 ICT sector emissions reduction goals for data centers (servers) are achievable through a combination of interventions, including renewable energy for operation, low-carbon electricity for fabrication facilities, and server lifetime extension. For smartphones, we find that low-carbon electricity for fabrication facilities alone will not achieve 2030 emissions reduction targets, and therefore call for additional solutions to reduce the impacts of that consumer device segment. For data centers (servers) we find that interventions focused on operational emissions (renewable energy for operation and limiting growth in operational energy usage) were more effective in terms of overall emissions reductions than the interventions focused on embodied emissions reductions (low-carbon electricity for fabrication and lifetime extension). Importantly, we note the “missed opportunity” associated with efficiency improvements in computing devices and data center operations as a lever for emissions impact reduction as well as the limitations of interventions focused solely on reducing the emissions associated with energy use in different portions of server device lifecycles. While we find that the 2030 emissions reduction targets for data centers (servers) were achieved without harnessing efficiency improvements for impact reduction, we also find that the infrastructure costs of achieving those targets were reduced when a limit on sector growth to match efficiency improvements was also employed. Emissions reduction solutions such as 24/7 renewable energy for datacenters are necessarily limited in their impact reduction potential once the additional embodied impacts, including from dedicated renewable generation infrastructure and any additional server capacity, are taken into account. In addition, this strategy is limited by spatial and temporal availability of renewable energy resources and the degree of temporal flexibility of a data center’s workloads. Efficiency gains of computing devices, on the other hand, offer a clear reduction in use-phase energy by reducing the energy required for the same amount of work, and, as long as area scaling benefits continue to reduce the overall manufacturing impacts per chip in conjunction with that increased operational efficiency,[11] those gains come without a similar backlash in embodied carbon - making this the only lever that we propose that might be considered something of a “free lunch”. Future work should continue to explore how we might motivate the use of achieved computational efficiency gains for sectoral impact reduction and effectively reign in excessive growth in computational demand in accordance with Jevons Paradox.

In addition, as the ICT sector continues to pursue efficiency gains, it will be important to continue to assess whether embodied cost tradeoffs for those operational efficiency gains do become relevant with future technologies (i.e. the lunch is no longer "free"). In this vein, the second half of this work proposes a way to estimate the potential scaled impacts of future computing devices through an underspecified sampling approach based on fabrication process steps. We then demonstrate how one might construct a tradeoff space to allow comparison to an existing technology, accounting for both embodied emissions costs and operational emissions benefits to determine when new technology adoption would be beneficial from a total carbon footprint perspective. Further, we highlight how decisions around a new technology might change depending on use conditions, such as the grid emissions factor of a use location or a device's utilization rate by allowing for shifting decision boundaries within the constructed tradeoff space. In our case study of a prototypical analog accelerator device, we find that incorporating the additional embodied carbon costs into the tradeoff space did not change decisions in terms of total carbon footprint benefits of device adoption under a variety of use scenarios if that analog accelerator was able to obtain the operational characteristics of a two published hypothetical analog accelerator chip designs. However, given the disparity between characteristics of hypothetical and fabricated analog accelerator devices, the tradeoff space we present can help to track the progression of the technology and determine when analog accelerators reach total footprint parity with comparative existing devices. We note that incorporating the additional embodied carbon estimate into decision making is more important when device characteristics are near to the decision boundary, or when the analog accelerator device may exceed performance of a digital accelerator device on one metric (eg. energy efficiency) but perform worse on another (eg. area efficiency), or vice versa.

Lastly, we make note of the continued uncertainty and difficulty in estimating impact parameters within the ICT sector owing to a lack of up to date and transparent reporting of data within the sector and no one central authority/standardized reporting methodology. If we want to truly reduce the global impact of this sector going forward, we will need better data to feed into models such as the one we have proposed in order to assess the best solutions for achieving that impact reduction moving forward.

Areas of future research

There are many opportunities for future research work within this space. Here we highlight additional topics that arose in discussions of this work that remain, to our knowledge, unaddressed.

Firstly, the methods and approach formed within this work could be extended to new areas of the ICT sector and other emerging technologies. The 2030 global footprint reduction analysis could be extended to also apply to other consumer devices (such as desktop computers, tablets, etc.) and communication networks. Communication networks in particular might prove important for future work as they have a similar operational energy footprint to data centers.[15] Previous work notes the relative difficulty of addressing communication network emissions as compared to data centers, given the non-centralized and diverse technological nature of communication network systems as well as a lack of data on renewable energy incorporation.[3] In terms of applying this methodology to other consumer devices, this would become more impactful if other interventions that could reduce the impact of consumer devices were considered (for smartphones, we only model an intervention on the emissions intensity of fabrication electricity). Expanding this analysis in either of these directions would likely benefit from engagement with industry partners in order to generate additional intervention ideas and obtain additional data on the current state of systems and emissions reduction strategies already in place.

There are additional large-scale trends within the ICT industry that we do not touch upon within this work, but that would make interesting areas for future inquiry, including the rise of the internet of things (IoT) and a shifting of computing burdens from data centers to edge devices. Previous work suggests that the potential future implications of IoT in terms of embodied and operational emissions of the ICT sector are underexplored.[9,4,3] In addition, as artificial intelligence continues to become pervasive within computing technologies, there is a trend towards moving away from centralized AI inference in data centers towards inference that happens on individual edge devices to benefit user privacy, the speed of an application, and allow access to AI models without internet connectivity.[97,98] While some initial work on this topic indicates that edge computing may be preferable from a carbon emissions perspective,[99] this topic merits further research, especially in light of the increasingly low operational emissions footprints of data centers with increases in renewable energy procurement.

Our method for new computing technology impact estimation could be extended to model operational versus embodied tradeoffs associated with other emerging technologies. In particular, new memory technologies, including non-volatile memory technologies like resistive random access memory or phase change memory,[100] are of interest given that memories are a significant user of energy within a computing system,[101] and these new memory technologies offer potentially significant operational energy reduction opportunities.[102] It would also be interesting to evaluate the embodied and operational emissions tradeoffs of newer high-bandgap semiconductor technologies, such as those based on SiC rather than Si wafers. This would likely require engagement with an industry partner involved in SiC semiconductor manufacturing in order to fill in the relevant data gaps that have been identified by previous work[103] as necessary to perform a life cycle analysis assessment on this technology. In addition, there is an opportunity to improve upon our estimates in this analysis. More updated data on the emissions associated with different fabrication processes, as well as the emissions associated with newer process technologies would help to increase the accuracy of our estimates. In addition, it would be interesting to obtain additional data on variables that we were unable to fully explore within this analysis, such as the effect of additional processing steps on device yields.

Eventually, we hope that this and future work can manifest itself as a tool for optimizing sustainability within computing system designs, from the level of an integrated circuit to the operation of a completed device. The concepts we explore within this work, such as device lifetimes or specialization, could be thought of as design parameters to be optimized for minimizing the total emissions impact of a system and our estimates of the scaled carbon footprint of an emerging computing device could be used to reduce a device’s footprint before a device is scaled. However, additional research, discussion with ICT industry experts, and data collection will be required before this kind of tool can be achieved.

Bibliography

1. Jones, N. Data centres are chewing up vast amounts of energy. *Nature* **561**, (2018).
2. Gupta, U. *et al.* Chasing Carbon: The Elusive Environmental Footprint of Computing. Preprint at https://doi.org/10.48550/arXiv.2011.02839 (2020).
3. Belkhir, L. & Elmeligi, A. Assessing ICT global emissions footprint: Trends to 2040 & recommendations. *J. Clean. Prod.* **177**, 448–463 (2018).
4. Malmodin, J. & Lundén, D. The Energy and Carbon Footprint of the Global ICT and E&M Sectors 2010–2015. *Sustainability* **10**, 3027 (2018).
5. Andrae, A. S. G. & Edler, T. On Global Electricity Usage of Communication Technology: Trends to 2030. *Challenges* **6**, 117–157 (2015).
6. Andrae, Anders S.G. New perspectives on internet electricity use in 2030. *Eng. Appl. Sci. Lett.* **3**, 19–31 (2020).
7. Acun, B. *et al.* Carbon Explorer: A Holistic Approach for Designing Carbon Aware Datacenters. (2023) doi:10.1145/3575693.3575754.
8. Jattke, M., Bieser, J., Blumer, Y., Itten, R. & Stucki, M. Environmental implications of service life extension of mobile devices. in 163–170 (Fraunhofer IZM, 2020). doi:10.21256/zhaw-20808.
9. Freitag, C. *et al.* The real climate and transformative impact of ICT: A critique of estimates, trends, and regulations. *Patterns* **2**, 100340 (2021).
10. Boyd, S. B. *Life-Cycle Assessment of Semiconductors*. (Springer, 2012). doi:10.1007/978-1-4419-9988-7.
11. Garcia Bardon, M. *et al.* DTCO including Sustainability: Power-Performance-Area-Cost-Environmental score (PPACE) Analysis for Logic Technologies. in *2020 IEEE International Electron Devices Meeting (IEDM)* 41.4.1-41.4.4 (2020). doi:10.1109/IEDM13553.2020.9372004.

12. International Roadmap for Devices and Systems 2022 Edition. (2022).

13. Nagapurkar, P. & Das, S. Economic and embodied energy analysis of integrated circuit manufacturing processes. *Sustain. Comput. Inform. Syst.* **35**, 100771 (2022).

14. Hetherington, A. C., Borrion, A. L., Griffiths, O. G. & McManus, M. C. Use of LCA as a development tool within early research: challenges and issues across different sectors. *Int. J. Life Cycle Assess.* **19**, 130–143 (2014).

15. Data Centres and Data Transmission Networks – Analysis. *IEA* https://www.iea.org/reports/data-centres-and-data-transmission-networks.

16. IEA. Total Energy Model. *4E Energy Efficient End-use Equipment* https://www.iea-4e.org/edna/tem/.

17. Ericsson Mobility Visualizer - Mobility Report. https://www.ericsson.com/en/reports-and-papers/mobility-report/mobility-visualizer.

18. Boavizta Project - Environmental Footprint Data. (2023).

19. IEA. World Energy Outlook 2022. *IEA* (2022).

20. Google. Google Environmental Report 2022. (2022).

21. Mobile phone shipments worldwide 2009-2026 | Statista. https://www-statista-com.libproxy.mit.edu/statistics/272696/mobile-phone-shipments-worldwide-by-quarter/.

22. Server unit shipments by quarter 2021. *Statista* https://www.statista.com/statistics/287005/global-server-shipments/.

23. Cisco Global Cloud Index: Forecast and Methodology, 2016–2021. (2018).

24. Borderstep Institute *et al. Energy-efficient cloud computing technologies and policies for an eco-friendly cloud market: final study report*. (Publications Office of the European Union, 2020).

25. Strategy Analytics: Half the World Owns a Smartphone. https://www.businesswire.com/news/home/20210624005926/en/Strategy-Analytics-Half-the-World-

Owns-a-Smartphone (2021).

26. thinkstep. Life Cycle Assessment of Dell R740. (2019).

27. Proske, M., Sánchez, D., Clemm, C. & Baur, S.-J. LIFE CYCLE ASSESSMENT OF THE FAIRPHONE 3. (2020).

28. Sánchez, D., Proske, M. & Baur, S.-J. LIFE CYCLE ASSESSMENT OF THE FAIRPHONE 4. (2022).

29. TSMC 2021 Sustainability Report. (2021).

30. SK hynix Sustainability Report 2022. (2022).

31. 2022 Micron Sustainability Report. (2022).

32. Bureau of Energy, M. of E. A. 2021 Electricity Carbon Emission Factor. *Bureau of Energy, Ministry of Economic Affairs, R.O.C.* https://www.moeaboe.gov.tw/ECW/english/content/Content.aspx?menu_id=20721 (2022).

33. Taiwan's dominance of the chip industry makes it more important. *The Economist* (2023).

34. Global Industry Analysts. Global Smartphones Industry. (2023).

35. Flucker, S., Tozer, R. & Whitehead, B. Data centre sustainability – Beyond energy efficiency. *Build. Serv. Eng. Res. Technol.* **39**, 014362441775302 (2018).

36. Sverdlik, Y. Microsoft Pledges to Run Data Centers, Offices Carbon-Free 24/7. *Data Center Knowledge* (2021).

37. Tracking Our Carbon-Free Energy Progress. *Google Sustainability* https://sustainability.google/progress/energy/.

38. Radovanovic, A. *et al.* Carbon-Aware Computing for Datacenters. Preprint at http://arxiv.org/abs/2106.11750 (2021).

39. Carbon Explorer. *Github* https://github.com/facebookresearch/CarbonExplorer (2023).

40. Wind Electricity – Analysis. *IEA* https://www.iea.org/reports/wind-electricity (2022).

41. Ravi, S. Chipmakers Are Ramping Up Production to Address Semiconductor Shortage. Here's Why

that Takes Time. *Semiconductor Industry Association* https://www.semiconductors.org/chipmakers-are-ramping-up-production-to-address-semiconductor-shortage-heres-why-that-takes-time/ (2021).

42. Y, J. The Economics of TSMC's Giga-Fabs. *The Asianometry Newsletter* https://asianometry.substack.com/p/the-economics-of-tsmcs-giga-fabs (2022).
43. Burkacky, O., Nikolka, M., Göke, S., Patel, M. & Spiller, P. Sustainability at semiconductor fabs. *McKinsey & Company* https://www.mckinsey.com/industries/semiconductors/our-insights/sustainability-in-semiconductor-operations-toward-net-zero-production (2022).
44. Ticoras, M. With Ohio about to become a chips-making hub, the time is now to embrace renewable energy. *cleveland.com* https://www.cleveland.com/opinion/2022/07/with-ohio-about-to-become-a-chips-making-hub-the-time-is-now-to-embrace-renewable-energy-mitchell-ticoras.html (2022).
45. carbon footprint. County Specific Electricity Grid Greenhouse Gas Emission Factors. (2022).
46. Ohio Solar Industry Welcomes Intel. *Business Wire* https://www.businesswire.com/news/home/20220121005479/en/Ohio-Solar-Industry-Welcomes-Intel (2022).
47. Net zero targets could force Taiwan's chipmakers abroad. *Clean Energy Wire* https://www.cleanenergywire.org/news/net-zero-targets-could-force-taiwans-chipmakers-abroad (2022).
48. Lin, P. & Sun, L. TSMC Becomes the World's First Semiconductor Company to Join RE100, Committed to 100% Renewable Energy Usage. *TSMC ESG* https://esg.tsmc.com/en/update/greenManufacturing/caseStudy/37/index.html (2020).
49. Chuang, B. Renewable energy to be 2nd largest source of power in Taiwan in 2030, says MOEA.

DIGITIMES https://www.digitimes.com/news/a20221230PD201/renewable-energy.html (2023).

50. Transforming homes into power stations - how Sweden is disrupting energy production. *World Economic Forum* https://www.weforum.org/agenda/2020/09/sweden-energy-production-renewable-power-district-heating/ (2020).
51. International Trade Administration. Taiwan Renewable Energy Market. https://www.trade.gov/market-intelligence/taiwan-renewable-energy-market (2021).
52. Staff Writer. Renewables to fall short of 2025 target, report says. *Taipei Times* (2022).
53. Davis, J. Uptime Institute Global Data Center Survey 2022. (2022).
54. Horizon Editorial. Analyzing Hardware Refresh Cycles in the Data Center. *Horizon Technology* https://horizontechnology.com/news/data-center-hardware-refresh-cycles/ (2022).
55. Caballar, R. D. Data Center Hardware Refresh Cutback by Microsoft — What's Next? *Data Center Knowledge* (2022).
56. Ercan, M., Malmodin, J., Bergmark, P., Kimfalk, E. & Nilsson, E. Life Cycle Assessment of a Smartphone. in 124–133 (Atlantis Press, 2016). doi:10.2991/ict4s-16.2016.15.
57. SPEC. All Published SPEC SPECpower_ssj2008 Results. https://www.spec.org/power_ssj2008/results/power_ssj2008.html.
58. Jouppi, N. P. *et al.* In-Datacenter Performance Analysis of a Tensor Processing Unit. Preprint at http://arxiv.org/abs/1704.04760 (2017).
59. Amodei, D. & Hernandez, D. AI and compute. *OpenAI* (2018).
60. Patterson, D. *et al.* The Carbon Footprint of Machine Learning Training Will Plateau, Then Shrink. *Computer* **55**, 18–28 (2022).
61. Wu, C.-J. *et al.* Sustainable AI: Environmental Implications, Challenges and Opportunities. Preprint at http://arxiv.org/abs/2111.00364 (2022).

62. Galvin, R. The ICT/electronics question: Structural change and the rebound effect. *Ecol. Econ.* **120**, 23–31 (2015).

63. Xu, Z., Wang, Z. & Liao, H.-T. People-centered Computing Within Limits: System Thinking on Interventions of Internet Platforms. in *Proceedings of the 2019 3rd International Conference on Cloud and Big Data Computing* 16–20 (Association for Computing Machinery, 2019). doi:10.1145/3358505.3358523.

64. The Jevons Paradox and Rebound Effect: Are we implementing the right energy and climate change policies? *The OECD Forum Network* http://www.oecd-forum.org/posts/the-jevons-paradox-and-rebound-effect-are-we-implementing-the-right-energy-and-climate-change-policies (2022).

65. Comments of the Semiconductor Industry Association on the Proposal of the Securities and Exchange Commission: The Enhancement and Standardization of Climate-Related Disclosures for Investors. (2022).

66. Best Practice Guidance for Semiconductor PFC Emission Reduction. (2017).

67. JOINT STATEMENT OF THE 25 th MEETING OF THE WORLD SEMICONDUCTOR COUNCIL (WSC). (2021).

68. LeCun, Y. A Path Towards Autonomous Machine Intelligence Version 0.9.2, 2022-06-27. (2022).

69. Artificial Intelligence: What is its future and how will it impact data center markets? *CBRE* (2023).

70. Microsoft will be carbon negative by 2030. *The Official Microsoft Blog* https://blogs.microsoft.com/blog/2020/01/16/microsoft-will-be-carbon-negative-by-2030/ (2020).

71. We’re carbon neutral. And by 2030, every product you love will be too. *Apple Environment* https://www.apple.com/environment/.

72. Product Lifecycle Approach | Logitech Sustainable Designs. https://www.logitech.com/en-us/sustainability/environment.html.

73. EPEAT Registry. https://www.epeat.net/.

74. Product carbon footprint label.
https://www.carbontrust.com/what-we-do/assurance-and-labelling/product-carbon-footprint-label
https://www.carbontrust.com/what-we-do/assurance-and-labelling/product-carbon-footprint-label
(2020).

75. MIT Climate and Sustainability Consortium Workshop on Carbon Footprinting of Electronic and Communication Devices. (2022).

76. Vasan, A., Sood, B. & Pecht, M. Carbon footprinting of electronic products. *Appl. Energy* **136**, 636–648 (2014).

77. Taylor, K. EU plans 'digital product passport' to boost circular economy. *www.euractiv.com*
https://www.euractiv.com/section/circular-economy/news/eu-plans-digital-product-passport-to-boost-circular-economy/ (2021).

78. Totaro, A. Europe: Digital Product Passport is Coming Soon. *Renewable Matter*
https://www.renewablematter.eu/articles/article/europe-digital-product-passport-is-coming-soon
(2022).

79. EU digital passport scheme: using blockchain to decentralise Europe's circular economy. *European Scientist*
https://www.europeanscientist.com/en/big-data/eu-digital-passport-scheme-using-blockchain-to-decentralise-europes-circular-economy/ (2022).

80. Carbon Border Adjustment Mechanism. *European Commission - European Commission*
https://ec.europa.eu/commission/presscorner/detail/en/qanda_21_3661.

81. Mertenskötter, C. G. M., Péter Balás, Paul. Twelve Things to Know About the Upcoming EU Carbon Border Adjustment Mechanism. *Global Policy Watch*
https://www.globalpolicywatch.com/2021/06/twelve-things-to-know-about-the-upcoming-eu-carbon-border-adjustment-mechanism/ (2021).

82. Figures, T., Gilbert, M., McAdoo, M. & Voigt, N. The EU's Carbon Border Tax Will Redefine Global Value Chains. *BCG Global* https://www.bcg.com/publications/2021/eu-carbon-border-tax (2021).

83. The White House. FACT SHEET: CHIPS and Science Act Will Lower Costs, Create Jobs, Strengthen Supply Chains, and Counter China. *The White House* https://www.whitehouse.gov/briefing-room/statements-releases/2022/08/09/fact-sheet-chips-and-science-act-will-lower-costs-create-jobs-strengthen-supply-chains-and-counter-china/ (2022).

84. James, S. B., Modrich, S. & Price, S. Path to net-zero: US chipmakers balance growth vs. going green. *S&P Global* https://www.spglobal.com/marketintelligence/en/news-insights/latest-news-headlines/path-to-net-zero-us-chipmakers-balance-growth-vs-going-green-70486324.

85. Lloyd, J. Sustainability for Semiconductors. *Issues in Science and Technology* https://issues.org/sustainability-semiconductors-harrington-dhople-wang-choi-koester/ (2022).

86. Governor Hochul Signs Transformative Green CHIPS Legislation to Create Jobs and Lower Emissions by Boosting Semiconductor Manufacturing in New York. *New York State* https://www.governor.ny.gov/news/governor-hochul-signs-transformative-green-chips-legislation-create-jobs-and-lower-emissions.

87. Sacchi, R. *et al.* PRospective EnvironMental Impact asSEment (premise): A streamlined approach to producing databases for prospective life cycle assessment using integrated assessment models. *Renew. Sustain. Energy Rev.* **160**, 112311 (2022).

88. imec. SSTS: achieving sustainability in semiconductor industry. https://www.imec-int.com/en/expertise/cmos-advanced/sustainable-semiconductor-technologies-and-systems-ssts.

89. Onen, M. *et al.* Nanosecond protonic programmable resistors for analog deep learning. *Science* **377**, 539–543 (2022).

90. Onen, O. M., Del Alamo, J., Li, J. & Yildiz, B. CMOS-Compatable Protonic Resistive Devices. (2022).

91. Liao, Y.-W. *et al.* Resistive random access memory (rram) and method of making. (2014).

92. Krishnan, N. *et al.* A Hybrid Life Cycle Inventory of Nano-Scale Semiconductor Manufacturing. *Environ. Sci. Technol.* **42**, 3069–3075 (2008).

93. Ankit, A. *et al.* PUMA: A Programmable Ultra-efficient Memristor-based Accelerator for Machine Learning Inference. Preprint at http://arxiv.org/abs/1901.10351 (2019).

94. Xiao, P., Bennett, C. H., Feinberg, B., Agarwal, S. & Marinella, M. J. Analog architectures for neural network acceleration based on non-volatile memory. *Appl. Phys. Rev.* **7**, (2020).

95. Shafiee, A. *et al.* ISAAC: A Convolutional Neural Network Accelerator with In-Situ Analog Arithmetic in Crossbars. in *2016 ACM/IEEE 43rd Annual International Symposium on Computer Architecture (ISCA)* 14–26 (2016). doi:10.1109/ISCA.2016.12.

96. Shanbhag, N. R. & Roy, S. K. Benchmarking In-Memory Computing Architectures. *IEEE Open J. Solid-State Circuits Soc.* **2**, 288–300 (2022).

97. Coral. *Coral* https://coral.ai/.

98. Edge TPU - Run Inference at the Edge. *Google Cloud* https://cloud.google.com/edge-tpu.

99. Kimovski, D. *et al.* Cloud, Fog, or Edge: Where to Compute? *IEEE Internet Comput.* **25**, 30–36 (2021).

100. Vetter, J. S. & Mittal, S. Opportunities for Nonvolatile Memory Systems in Extreme-Scale High-Performance Computing. *Comput. Sci. Eng.* **17**, 73–82 (2015).

101. Kultursay, E., Kandemir, M., Sivasubramaniam, A. & Mutlu, O. Evaluating STT-RAM as an energy-efficient main memory alternative. in *2013 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS)* 256–267 (IEEE, 2013). doi:10.1109/ISPASS.2013.6557176.

102. Gamatié, A. *et al.* Emerging NVM Technologies in Main Memory for Energy-Efficient HPC: an Empirical Study. *HAL Open Sci.* (2019).

103. Schmidt, S., Díaz, A., Glaser, S. & Makoschitz, M. A “life cycle thinking” approach to assess

differences in the energy use of SiC- vs. Si power semiconductors. *Power Electron. Convers. Technol. Annex PECTA* doi:https://www.iea-4e.org/wp-content/uploads/publications/2022/01/DiazSchmidtGlaserMakoschitz_LCThinkingWBG_For_PECTA.pdf.

Appendix

Parameter estimates and uncertainty assignments for global data center and smartphone footprints

Parameter	**Estimate**	**Source**	**Link**	**Uncertainty distribution assigned**
Global data center energy usage, 2021	220-320 TWh	IEA	Link	Uniform
Global number of active smartphones, 2021	2.967 billion, 6.262 billion	IEA Electronic Devices & Networks Annex (EDNA) Total Energy Model, Ericsson	Link Link	Uniform
Yearly energy consumption of a smartphone	7.75 kWh average & 1.24 std dev	Google Pixel published TEC data	Link	Normal distribution
Global average carbon intensity of electricity, 2021	459 gCO^2/kWh	IEA World Energy Outlook 2022	Link	Data quality index
Yearly new server shipments, 2020	9.53 million	IDC	Link	Data quality index
Yearly new smartphone shipments, 2021	1660 million	IDC	Link	Data quality index
Server manufacturing carbon emissions	mean 1230 std dev 374 kg CO^2e	Boavizta Environmental Footprint Dataset (filtered by Subcategory = "Server" & to remove entries without gwp_manufacturing_ratio - 7 removed, all Lenovo)	Link	Normal distribution
Smartphone manufacturing carbon emissions	mean 52.7, standard deviation 18.3 kg CO^2e	Boavizta Environmental Footprint Dataset (filtered by Subcategory = "Smartphone" & to remove entries without GWP total - 3	Link	Normal distribution

		removed)		
Proportion of manufacturing emissions of servers that are attributable to integrated circuit fabrication (including logic and memory production)	~78%	2019 Dell Server LCA	Link	Data quality index
Proportion of manufacturing emissions of smartphones that are attributable to integrated circuit fabrication (including logic and memory production)	66 & 69%	Fairphone 3 & 4 LCAs	Link Link	Uniform
Portion of the integrated circuit manufacturing impact can be attributed to electricity as opposed to process emissions	In 2021, 50% of TSMC's emissions came from Scope 2 impacts, SK Hynix attributing 45% of their total emissions to Scope 2 in 2021, and Micron reporting 54% of their emissions as stemming from purchased energy in 2021 (but note that Micron did not estimate Scope 3, so Scope 2 may be overrepresented)	TSMC reporting, SK Hyinx, Micron	Link Link Link	Uniform
Grid emissions factor of Taiwan, 2021	0.509 kg CO^2e/kWh	Taiwan Bureau of Energy	Link	Data quality index
"Best case" low carbon grid emissions estimate - Sweden	0.02881 kg CO^2e/kWh in 2020	Carbon Footprint Ltd.	Link	Data quality index
2030 average grid intensity	165-330g CO^2/kWh	IEA World Energy Outlook 2022	Link	Uniform

2030 new servers shipped	60 million	New perspectives on internet electricity use in 2030 (2020)	Link	Data quality index
2030 active smartphones	5 billion - 7.861 billion (in 2028)	Half the World Owns a Smartphone by Strategy Analytics, Ericsson Mobility Visualizer	Link Link	Uniform
2030 smartphone shipments	1.7 billion units, 1774 million	Global Smartphones Industry report, New perspectives on internet electricity use in 2030	Link Link	Data quality index
Data center traffic (EB), 2021	20,555 EB	Cisco Annual Internet Report (2018–2023) White Paper	Link	Data quality index
CAGR of total datacenter traffic (EB) 2021-2030	24.70% CAGR between 2016 and 2021	Cisco Global Cloud Index	Link	No uncertainty assigned, taken as an assumption
Yearly data center energy efficiency improvements	5-15%	N/A	N/A	Uniform