Sustainable AI Processing at the Edge

Sébastien Ollivier, Sheng Li, Yue Tang, Stephen Cahoon, Ryan Caginalp, Chayanika Chaudhuri, Peipei Zhou, Xulong Tang, Jingtong Hu, and Alex K. Jones, *University of Pittsburgh, Pittsburgh, PA, 15260, USA*

Edge computing is a popular paradigm for accelerating light- to medium-weight machine learning algorithms initiated from mobile devices without requiring the long communication latencies to send them to remote datacenters in the cloud. Edge servers primarily consider traditional concerns, such as size, weight, and power constraints for their installations. However, such metrics are not entirely sufficient to consider environmental impacts from computing given the significant contributions from embodied energy and carbon. In this article we explore the tradeoffs of hardware strategies for convolutional neural network acceleration engines considering inference and online training. In particular, we explore the use of mobile graphics processing unit (GPU) accelerators, recently released edge-class field-programmable gate arrays, and novel processing in memory (PIM) using dynamic random-access memory (DRAM) and emerging Racetrack memory. Given edge servers already employ DRAM and sometimes GPU accelerators, we consider the sustainability implications using breakeven analysis of replacing or augmenting DDR3 with Racetrack memory. We also consider the implications for provisioning edge servers with different accelerators using indifference analysis. While mobile GPUs are typically much more energy efficient, their significant embodied energy can make them less sustainable than PIM solutions in certain scenarios that consider activity time and compute effort.

Deep neural networks have become a popular algorithm used by a variety of applications on mobile devices including smart phones, autonomous vehicles, robotics, unmanned aerial vehicles, and other smart and connected devices. Convolutional neural networks (CNNs) have been demonstrated as an effective deep learning implementation methodology that trades computational complexity for accuracy.

There have been many proposals to improve the performance and energy efficiency of CNN inference. However, these algorithms may still be too compute and data intensive to execute directly on mobile nodes that typically have limited energy and computational capabilities. In addition, due to changes or drift in input datasets over time, it is sometimes necessary to adjust the parameters of CNN inference algorithms through online training. Online training is typically intractable for mobile connected devices.

Thus, edge servers, now often being deployed in conjunction with advanced (e.g., 5G) wireless networks, have become a popular target to accelerate CNN inference and training. Moreover, due to their deployment in the field, edge servers must operate under size, weight, and power (SWaP) constraints, while serving many concurrent requests from mobile clients. Thus, to accelerate CNNs, these edge servers often use energy-efficient accelerators, sometimes employing reduced precision approximate models. Their goal is to achieve fast response time while balancing requests from multiple clients and maintaining a low operational energy cost.

Moreover, keeping online training local to edge server nodes avoids communicating large datasets from edge to cloud servers.[1] However, online training typically requires much higher precision and floating-point computation. This can be a heavier burden to edge servers compared to inference.

0272-1732 © 2022 IEEE
This article has supplementary downloadable material available at https://doi.org/10.1109/MM.2022.3220399, provided by the authors.
Digital Object Identifier 10.1109/MM.2022.3220399
Date of publication 8 November 2022; date of current version 13 January 2023.

While edge servers can dramatically improve capabilities to deploy deep learning more broadly, this proliferation of lightweight computing from mobile devices and medium-weight computing from edge servers can create negative environmental impacts. Manufacturing new mobile and edge computing infrastructure requires problematic emissions of everything from carcinogens to volatile organic compounds, not to mention greenhouse warming gases (GWGs). These include most notably carbon dioxide (CO_2) but also methane (CH_4) and nitrous oxide (N_2O), among others.

As such, there is a significant and growing aspect of environmental impacts that come from embodied impacts of computing.[2] Embodied impacts include the energy, GWGs, waste water generation, etc., from manufacturing computing infrastructure, particularly the semiconductor elements that form the heart of all computing systems.

Recent evidence shows that for cloud servers, embodied impacts are equally as high as operational (runtime) effects.[3] For mobile devices and compact computers, embodied impacts can reach 80%–90% of total lifecycle impacts and that these impacts are dominated by their integrated circuits (ICs).[3,4] Thus, for systems already optimized for SWaP constraints, embodied energy will be a higher proportion of the total energy footprint, making its amortization an important sustainability goal. Accelerated deployment of mobile and edge systems to support deep learning exacerbate these concerns. Specialty processing units, including field-programmable gate arrays (FPGAs) and graphics processing units (GPUs), can accelerate CNN applications while meeting low operational energy constraints. However, this operational efficiency comes at the cost of increasing the silicon area of these edge systems. This creates a significant tradeoff between embodied energy from including accelerators and the operational energy impacts from executing deep learning algorithms.

THE PRIMARY SOURCE OF ENVIRONMENTAL IMPACTS FOR COMPUTING SYSTEMS COMES FROM THE CHIPS THAT IMPLEMENT THE CORE FUNCTIONALITY OF PROCESSING, MEMORY, AND DATA STORAGE.

In this article, we explore the several state-of-the-art proposals to accelerate CNN inference and training using GPUs, FPGAs, and processing in memory (PIM) with commodity DRAM and recently proposed Racetrack memory (RM) PIM.[5,6] Our comparison considers the main two phases of energy consumption of embodied and operational energy.[2] Thus, we explore total lifetime energy efficiency of the state-of-the-art computing targets allowing a evaluation of the sustainability of these different system choices.

We select energy as our metric as it bridges the manufacturing and operational phase of the system into a single metric that can be directly compared. However, we will also discuss how these energy values inform other environmental metrics including GWG when including electrical grid mix profiles.

In particular, this article makes the following contributions.

- We provide estimates of the embodied energy to fabricate edge class GPU, FPGA, and in-memory computation comparison points.
- We characterize the operational power and performance of representative CNN applications for edge-scale execution including both inference and training.
- We conduct indifference and breakeven analyses of different target systems and usage scenarios to determine holistic sustainability calculations.
- We explore the carbon impacts of these systems for different grid-mix choices.

In the next section we discuss the background and related work to conduct these analyses.

BACKGROUND

In this section, we provide a background on sustainability analysis through lifecycle assessment (LCA), indifference, and breakeven analyses. We also provide background on RM, including how it is used for PIM and its required extension for LCA. We also mention the features about CNN inference and training that lead to different assumptions about datatypes.

Lifecycle Assessment

The primary source of environmental impacts for computing systems comes from the chips that implement the core functionality of processing, memory, and data storage.[2] To determine the holistic environmental impacts in terms of energy, GWG, and other concerns of a product or process, such as semiconductor fabrication, typically involves a technique called LCA.[7] LCA is most accurate when a detailed analysis of the process is used to determine the assessment, but sometimes relative costs to similar processes can be used as a coarse-grain assessment called economic input/ouput LCA.

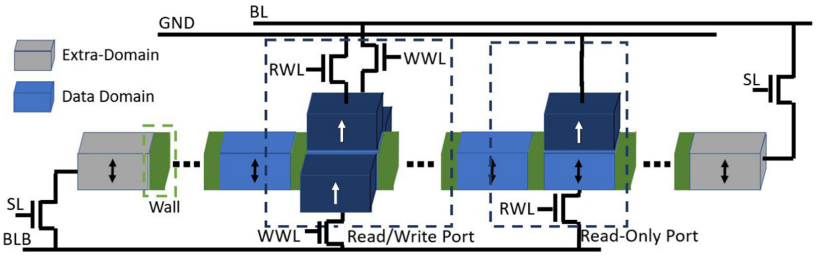

FIGURE 1. Anatomy of a domain-wall memory nanowire.

Semiconductor process LCA explores the impact of the different steps of the approximately 20 masks required to build complementary metal–oxide–semiconductor (CMOS) circuits. These masks can be broken down into their individual steps, such as deposition, lithography, etching, and metrology, per wafer. As the technology scales to smaller feature sizes, these steps become increasingly costly due to several factors. These include slower throughput and higher energy cost of the machines, more costly high fidelity clean rooms, and more process steps required for things, such as multipatterning lithography, high-κ dielectrics, more exotic transistor shapes and materials (e.g., III–V gate channels). A particular culprit is multipatterning[4] and extreme ultraviolet lithography steps.[8]

Relatively few process LCAs have been undertaken of semiconductor fabrication. One assessment considered CMOS, flash, and DRAM fabrication covering technologies from 350 nm down to 32 nm.[9] A hybrid (mixing process and economic) LCA combined process technology estimations with reported cost trends to create a semiconductor fabrication model estimating embodied energy scaling to 7 nm.[4] Recently, a process LCA was conducted for IC fabrication from 28–3-nm feature sizes.[8] Additional background on LCA for semiconductors can be found in the supplementary material.

Indifference and Breakeven Analyses

One motivation to use a single metric of energy for both manufacturing and operational sustainability evaluation of the system is to allow quantitative comparison metrics, such as indifference and breakeven analyses. To compare two design choices of the system for deployment we can use the indifference formula t_I, as shown in[10]

$$t_I = \frac{M_1 - M_0}{P_0 - P_1} \qquad t_B = \frac{M_1}{P_0 - P_1}. \tag{1}$$

For a system with higher embodied energy (M) and lower operational energy (P), t_I is the time at which the increase in embodied energy will be completely amortized by the savings in operational energy.

Thus, if the proposed service time $t < t_I$ the architecture with the lower embodied energy minimizes environmental impact. In contrast, for a proposed service time $t > t_I$ the architecture with the lower operational energy minimizes impact. If one choice is lower in both embodied and operational energy, then indifference analysis is not needed and the lower energy system can be selected independent of service time.

A similar calculation can be considered for the breakeven time t_B, also defined in (1).[10] Consider the case that an existing system is already deployed. Replacing the existing system is like assuming embodied energy of the deployed system is 0. Thus, t_B is the time it takes for the replacement system to overcome the embodied energy of the replacement through operational energy savings, i.e., $t_B = t_I$ when $M_0 = 0$.

While we characterize several accelerators in this work for CNN acceleration, we also consider an exotic technology that uses spintronics to store data and has been explored for PIM called RM.[11] We provide some background on RM in the next section.

Racetrack Memory

Spintronic RM[11] is made of ferromagnetic nanowires consisting of many magnetic domains separated by domain walls (DWs), as shown in Figure 1. Each domain has its own magnetization direction such that binary values are represented by the magnetization direction of each domain, either parallel/antiparallel to a fixed reference. For a planar nanowire, several domains share an access point for read and write operations.[12]

RM is similar to and has many of the same advantages as spin-transfer torque magnetic memory (STT-MRAM), including high endurance, fast access time, low energy. Energy is particularly low as static energy is nearly eliminated due to the device's nonvolatility. RM can have a density $\leq 2F^2$ because it can store multiple bits in a nanowire accessed using one transistor. In contrast, STT-MRAM requires 6–50F^2.[13]

Hence RM, which was originally conceived for secondary storage, has been proposed at several memory levels, from L1 cache to main memory. RM achieves

this density by requiring shifting if data are not aligned with an access point. Shifting occurs through DW motion in the nanowire.

Racetrack Memory Architecture

DW motion is controlled by applying a short current pulse laterally along the nanowire. Random access requires shifting the target domain to align it with an access point (dark blue) and apply a current to read or write the target bit. To avoid data loss when shifting, the blue domains store actual data while the grey domains are overhead domains to prevent data loss. Shift-based writing (read/write port)[14] allows slower current writes to be replaced with orthogonal shifts from fixed magnetic alignment domains to reduce latency and energy.

RM structures are typically built by bundling multiple tracks that are shifted together. Each track represents a different bit that can be accessed in parallel, while different memory addresses can be accessed by shifting the bundled tracks as a group to other positions.[15] Larger memory structures can be build from these groups of tracks to form tiles, subarrays, banks, etc.[5] Thus, the biggest challenge for RM is to accelerate and minimize shifting for fastest and more energy efficient operation.[11]

Processing in RM

Processing using memory has recently received considerable attention. DRAM-based techniques use multiple row simultaneously[16] and/or in sequence[17] to allow sensing amplifiers to achieve two-operand bulk bitwise logical operations. Higher level arithmetic logic is constructed out of a sequence of these logical operations.

RM has also received significant attention for PIM, particularly for deep learning.[5,6,18] The state-of-the-art approach uses a multidomain read to sense the number of 1's in a segment of the nanowire, such as between the two access points in Figure 1. From this access and 1's counting, it is possible to construct multioperand bulk bitwise logical operations. The number of operands is dictated by the size of the multidomain read.

Arithmetic structures, such as addition, can be constructed by converting a multidomain read into a local sum and carry logic. Multiplications are possible by summation of partial products.[5] Floating-point versions of these operations, particularly multiply accumulate, can be achieved by using these logical and arithmetic primitives on the sign, mantissa, and exponent components individually.[6] We provide more background on these ideas in the supplementary document.

Racetrack Memory LCA

RM, like many other novel memories, requires additional process steps during fabrication to realize the magentic nanowires and access ports. The process LCA for ICs including RM must be adjusted to account for the embodied cost of wafers including these additional steps.

In particular, additional layers of ferromagentic materials and insulators are placed on top of the completed CMOS layers. Typically these are added in between the lower levels of the metal stack. The spintronic devices are composed of three conceptual layers, a fixed magnetic layer, an MgO barrier that separates the fixed layer from the free layer in the form of a nanowire, often made out of a ferromagentic material, such as CoFeB. CoFeB with different doping properties can also be used for fixed magnetic layers.

In terms of the process steps, they are essentially the same between STT-MRAM and RM, which have been studied for the former.[19] Thus, during manufacture, in addition to the CMOS and metal layers, while circa 10 material layers are required for the magnetic devices, a total of three additional mask layers on top of the circa 20 CMOS layers are required to add these devices into the evaluation. According to process LCA study of these devices, they are composed of three lithography, three dry etching, three deposition steps, and a polishing step.[19]

We provide more background on RM including the process LCA methodology in the supplementary material.

Convolutional Neural Networks

CNNs are a popular method to compute deep learning algorithms. CNNs are dominated by the convolution operation, which is a windowed pointwise multiplication accumulation of multiple channels of input features with a set of weights to generate output features. As an example, for the input features $\mathbf{I}$ and weights $\mathbf{K}$ of size $N \times R_{\text{in}} \times C_{\text{in}}$ and $M \times N \times 3 \times 3$, respectively, the convolution operation for the window at m (output channel index), r (row), c (column) is

$$\text{Conv}(\mathbf{I}, \mathbf{K})(m, r, c) = \sum_{n=0}^{N-1} \sum_{j=0}^{2} \sum_{t=0}^{2} \mathbf{K}_{m,n,j,t} \times \mathbf{I}_{n,r+j,c+t}$$

where M is the number of output channels, N is the number of input channels, $R_{\text{in}} \times C_{\text{in}}$ is the size of an input feature map.

While deep learning with CNNs presumes calculations with floating-point values, CNN inference calculations can often be reduced to integer computation with as few as 8 bits achieving reasonable accuracy.

TABLE 1. Accelerator statistics, embodied energy, and embodied carbon emissions for grid mixes from Table 2.

	RM	DDR3	RM	RM	FPGA	GPU
Tech node	32[a,d]	55[a]	32[b,d]	32[c,d]	7[c]	14[c]
Die size (mm^2)	38	73	38	38	324	350
Die per wafer	1,847	967	1,847	1,847	217	201
PE (kWh/Wafer)	1,600	1,200	1,206	753	1,482	882
Energy (MJ/die)	3.12	4.47[e]	2.35	1.47	24.59	15.80
AZ (gCO_2eq/die)	343	490[e]	259	162	2,698	1,734
CA (gCO_2eq/die)	203	291[e]	153	95	1,598	1,027
TX (gCO_2eq/die)	380	544[e]	286	179	2,992	1,922
NY (gCO_2eq/die)	163	233[e]	123	77	1,284	825

[a]Calculated using process LCA from Boyd.[9]
[b]Calculated using process LCA from Higgs et al.[21]
[c]Calculated using process LCA from Garcia Bardon et al.[8]
[d]Requires extra steps for spintronics.[19]
[e]Requires 16 dies to build a the tested 1-GB DIMM.

Recent DRAM PIM work has shown that in many cases this can be further reduced to ternary $w \in \{-1, 0, 1\}$ or even binary $w \in \{0, 1\}$ computations operations to replace the multiplications. However, online training for all but the simplest CNNs still requires full 32-bit floating-point computations to work properly. Without this accuracy, the weight updates can be ineffective and possibly even detrimental.

In the next section, we explore embodied energy calculations of a variety of accelerators suitable for CNN acceleration.

EVALUATION OF EDGE ACCELERATION SUSTAINABILITY

To consider holistic energy across embodied and operational phases of potential edge accelerators requires use of the LCA of the semiconductor fabrication process discussed previously. In the next section, we discuss how to obtain embodied energy and carbon footprint for different accelerators.

Determining Embodied Energy and Carbon

As process LCA studies, including our modified process to include spintronics, report embodied energy per wafer, to determine the embodied energy of the DRAM, RM, FPGA, and GPU we require the IC die area and technology node. The die area determines what portion of the wafer is required for each die, from which the portion of the embodied energy of the wafer is a result of that die.

TABLE 2. Energy to (gCo_2eq/kwh)[22] and grid mixes.[23]

Source gCO_2eq/kWh		AZ	CA	TX	NY
Coal	980	20%	3%	19%	–
Natural gas	465	40%	39%	53%	37%
Geothermal	27	–	5%	–	–
Hydroelectric	24	5%	18%	–	22%
Solar PV	65	7%	20%	2%	2%
Wind	11	–	7%	17%	4%
Nuclear	27	28%	7%	9%	33%
Biopower	54	–	3%	–	–
Mix (gCO_2eq/kWh)		395	234	438	188

We use reported die areas for DDR3 DRAM, FPGAs, and GPUs for the selected devices reported in Table 1. For RM we used a modified version of NVSIM[20] to calculate the die area. We also are studying a version of RM that is extended with PIM capabilities to serve as an accelerator using the processing capabilities of CORUSCANT[5] and POD-RACING.[6] Thus, we calculated the additional die area of the PIM peripheral circuitry.[5] Thus, the RM-based accelerator has both an increased embodied energy per die area due to the exotic memory process as well as a larger die area than traditional RM due to the additional logic required for PIM.

There are CMOS process LCAs reported in the literature for 350–32-nm[9] processes and for 28–3 nm.[8] There are also DRAM process LCAs down to 55 nm[9] that were in service to produce DDR3 parts. There is a significant gap between the two studies as noted by the gap between the reported 32[9] and 28 nm,[8] such that a third study that reports 32 nm[21] sits between the two. Thus, in our work we do not compare nodes that cross the studies.

Because we report RM at 32 nm, for which there are three process LCA studies, we estimated the total cost based on the CMOS estimates from each of the three studies and make comparisons to devices that can be estimated using the same process LCA study. We discuss this in more detail in the supplementary material.

System Embodied Energy and Carbon Study

Several grid mix scenarios for CO_2 eq based on CO_2 eq per generation method[22] and reported grid mix per state[23] for states that have significant semiconductor manufacturing activities are presented in Table 2. These states, Arizona (AZ), California (CA), Texas (TX), and New York (NY), all have very different grid mixes.

AZ and TX have significant electrical generation from coal and the highest generation from natural gas. While AZ has significant generation from nuclear plants and TX has significant wind energy, their 395 and 438 gCO_2eq/kWh (CO_2 equivalent generated per kilowatt hour) are much higher than CA and NY, which still get more than a third of their electricity from natural gas. CA is very balanced on renewable energy and NY has significant hydroelectric and nuclear power generation, thus their grid mix generates about half the GWG emissions at 234 and 188 gCO_2eq/kWh, respectively.

In Table 1, we report the embodied energy and embodied carbon using the grid mixes from Table 2 for different accelerators. We targeted DDR3-1600 for DRAM as this is the device that has been used to implement DRAM PIM using ELP^{2}IM[17] and subsequently used to implement a ternary model reduction of CNN inference.

For dedicated accelerators we selected edge server appropriate low-energy devices including the Versal Prime FPGA (VM1802) from AMD/Xilinx and the NVIDIA Jetson NX mobile GPU. Note that, we were somewhat limited in our choice of, particularly FPGA, devices as die area is necessary to estimate embodied energy/carbon and not typically reported.

The RM is extremely dense, even with the additional PIM logic,[5] it has a low embodied energy even compared to the DRAM. The GPU and FPGA require an order of magnitude more embodied energy due to their much larger die sizes.

Holistic Sustainability Evaluation

To determine the overall energy (and carbon footprint) of these acceleration choices we compared a CNN conducting inference using hand-designed ternary approximations targeting DRAM PIM[17] and RM PIM[5] against the GPU using 8-bit integer precision from a PyTorch-based flow. Between the PIM solutions, RM provides both an embodied and operational energy improvement, ultimately providing order-of-magnitude benefits in mega frames per gCO_2eq.

RM is also competitive with the GPU, with the GPU having an approximately 30% latency and throughput advantage. However RM is clearly more sustainable having an order-of-magnitude improvement in both embodied and operational energy.

Breakeven Inference Analysis

We conducted two studies, presuming the edge system already contains DDR3 with PIM capabilities or a GPU. We illustrate this using the GreenChip tool[10] in Figure 2(a). The chart shows the comparisons between the two systems in terms of activity ratio on the y-axis versus sleep ratio on the x-axis. The sleep ratio is the ratio of active to sleep time. The activity ratio is, of the active time, the ratio of compute to idle time.[10] More details on how the GreenChip tool represents breakeven and indifference scenarios is included in the supplementary document.

In the comparison of adding RM to a server using DDR3 as a PIM accelerator [see Figure 2(a)], if the system is heavily loaded (bottom left) it can take on the order of a month before the RM upgrade saves overall energy. As the system becomes more idle (toward top left) or sleeping (toward bottom right) or both (toward top right), it can take months to recover the embodied cost. However, unless the machine is sleeping more than 75%–80% of the time, the upgrade will be recovered in less than one year. The time for RM to overtake the GPU is faster, with a busy server requiring days and lightly loaded server requiring months. This is because the embodied cost is lower in the design technology co-optimization (DTCO) estimation and the RM has a substantial advantage over the GPU in both dynamic and static power.

Indifference Online Training Analysis

To explore CNN training we compare the GPU and FPGA implementations using a PyTorch-based flow with hand optimization of AlexNet and VGG-16[1] as well as hand mapped designs for the RM accelerator.[6] From Tables 1 and 3, both embodied and operational energies for the FPGA are higher than both the RM and the GPU, so the indifference calculation will never pick the FPGA. The FPGA does have a lower power than the GPU, so its best use case is if the system has a hard power upper limit.

A notable sustainability comparison is that for training the RM has a lower embodied energy and a higher operational energy than the GPU. The indifference results are shown in Figure 2(c) and (d) for AlexNet and VGG-16, respectively. We note that GreenChip normalizes the usage scenario to the slower system. Thus, in these online training comparisons there is considerable headroom in terms of capacity of training jobs that can be handled by the GPU versus the RM. However, recent discussions with edge system usage scenarios reports that online training jobs are currently less than 30% of overall system loads and are more likely to be around 5%, placing these comparisons in the right ballpark for RM. Both training applications can benefit from the GPU in high usage scenarios (bottom left), but if the system does training infrequently, as is currently the likely scenario, the GPU savings during training cannot overcome the higher embodied

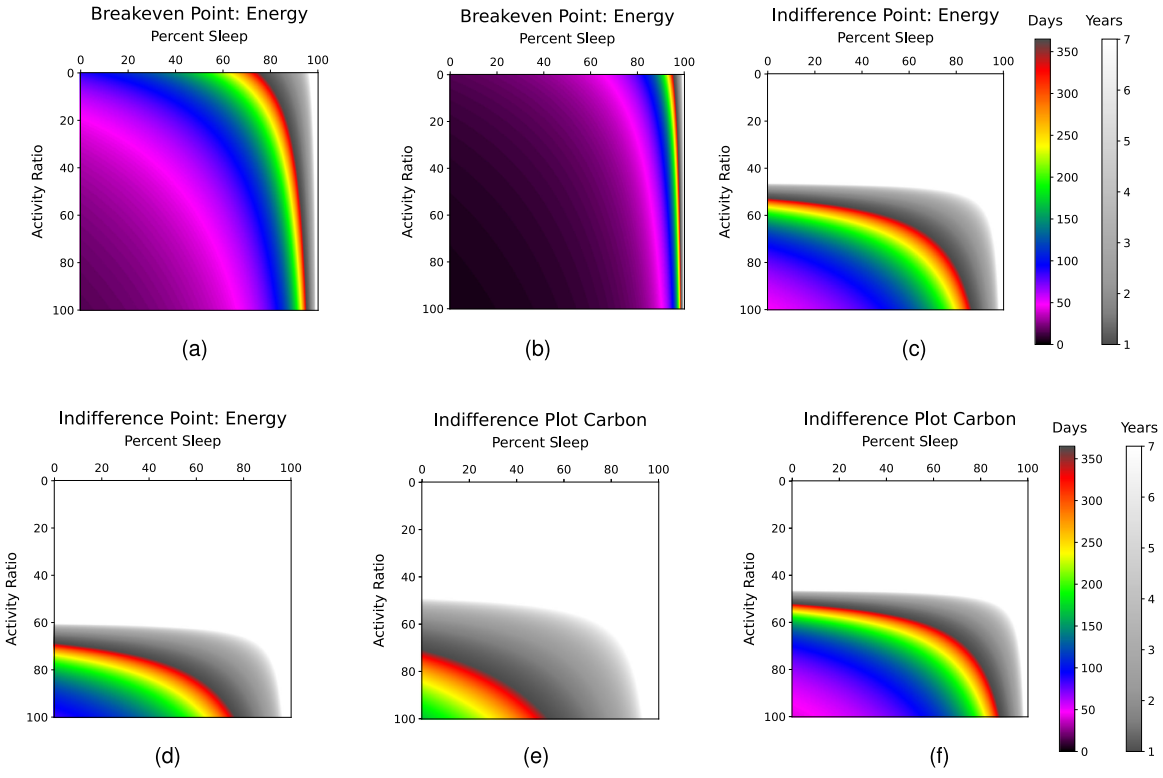

FIGURE 2. Sustainability analyses of different accelerator choices for edge systems. (a) AlexNet inference DDR3→RM. (b) AlexNet inference GPU → RM. (c) AlexNet training GPU versus RM. (d) VGG-16 training GPU versus RM. (e) AlexNet training GPU fabricated in AZ deployed in NY. (f) AlexNet training GPU fabricated in CA deployed in TX.

TABLE 3. Performance, operational power, and efficiency per power and carbon of different edge accelerators.

Inference acceleration using ternary model reduction and PIM						
Benchmark	**Target**	**Performance**		**Power**	**Efficiency**	
		Lat.(S)	FPS	W	FPS/W	MF/gCO_2eq
Alexnet	GPU	0.0014	705.9	9.54	74	0.61–1.42
	DDR3[17]	0.0118	84.8	2	42.4	0.35–0.81
Ternary[17]	RM	0.0020	490	0.93	526	4.6–10.8
Training acceleration using floating-point 32 data						
Benchmark	**Target**	**Performance**		**Power**	**Efficiency**	
		Lat.(S)	GFLOPS	W	GFLOPS/W	TFLOPS/gCO_2eq
Alexnet	GPU	0.005	1335	21.05	63.4	521–1,214
	RM	0.128	50.72	5.65	8.97	74–172
	FPGA	0.13	49.97	16.78	2.98	25–57
VGG-16	GPU	0.11	848	20.37	41.6	342–797
	RM	1.12	81.95	5.7	14.37	118–275
	FPGA	1.03	89.48	18.02	4.97	41–95

energy. For an deeply under loaded server, it becomes impossible for the GPU to benefit due to its higher static power. The activity ratio cutoff for Alexnet is around 50% and VGG-16 cuts off in the 40% range.

When considering the energy grid mix in the calculation this can deflect the indifference calculation substantially. In Figure 2(e) for online training of AlexNet, we explore the case where fabrication takes place in AZ, which has a comparatively high CO_2eq/kWh and the system is deployed in NY with a relatively low CO_2eq/kWh. Even in the highest utilization case the indifference point becomes six months, and in lower utilization (circa 70%) it becomes one year, and quickly grows to multiple years as the utilization drops toward 50%, favoring the RM for relatively more usage scenarios.

In Figure 2(f), a lower embodied carbon grid mix and higher operational carbon grid mix is explored for the same application. As expected the indifference times are much shorter favoring the GPU in more scenarios. Considering a deployment lifetime of circa two years, the AZ, NY scenario requires more than 60% training computation for the GPU to be worthwhile while in the CA, TX comparison this drops to 50% if the server remains active, but could drop to less than 20% if the server can sleep while not in use.

As online training becomes more popular in edge systems, the GPU becomes a more attractive alternative and we provide analyses normalized to the GPU in the supplementary document.

CONCLUSION

In this work, we compared several SWaP-optimized CNN accelerators popular for edge servers for both inference and online training metrics. The breakeven point analysis suggests replacing DRAM PIM with RM PIM results in a benefit in total energy within the $0 \leq t \leq 1$ years for most usage scenarios. The replacement time is likely on the low end of that time frame if the server is heavily used for this task, which is reasonably popular given the rising popularity of CNN acceleration on edge servers. The breakeven time is even more striking for a system using a Jetson Xavier NX mobile GPU, suggesting replacement with RM always yields a savings within just a few months.

In our indifference comparison between RM and the GPU, the edge server activity ratio needs to be at least 50% for lightweight CNN training algorithms, such as Alexnet, and higher for VGG-16 to make a GPU lower overall energy than RM. Because of the higher static power, lower utilization will always favor RM due to its lower embodied and static energy costs. To understand the carbon relationship we can see that the grid mix from manufacturing and use have a significant impact.

A SYSTEM CAN ACHIEVE BETTER SUSTAINABILITY EVEN IF IT IS NOT THE MOST OPERATIONALLY ENERGY EFFICIENT.

It is clear that embodied effects can remain high compared to operational effects. Even an energy-efficient GPU can be inefficient compared to reduced precision models for inference if the accuracy is sufficient. While one takeaway is that RM is an interesting compromise between efficient inference calculation and infrequent online training compared to the GPU, the more salient point is that a system can achieve better sustainability even if it is not the most operationally energy efficient.

The somewhat nonintuitive takeaway is that systems that dramatically reduce embodied energy in general, and static power particularly for underloaded servers, have a place for more sustainable edge computing. This is possible even if the accelerator has higher latency and operational energy than other accelerators. Designing accelerators for holistic sustainability remains an important challenge. Emerging architectures, such as tensor processors, should be studied. Emerging technologies, such as analog crossbars, should also be evaluated, in spite of their increases in embodied energy per area. We plan to explore these approaches in more detail in our future work.

ACKNOWLEDGMENTS

This work was supported in part by the National Science Foundation under Grants CNS-1822085 and CNS-2133267, in part by the National Security Agency, and in part by the Laboratory of Physical Sciences.

REFERENCES

1. Y. Tang, X. Zhang, P. Zhou, and J. Hu, "EF-train: Enable efficient on-device CNN training on FPGA through data reshaping for online adaptation or personalization," *ACM Trans. Des. Autom. Electron. Syst.*, vol. 27, no. 5, Jun. 2022, Art. no. 49, doi: 10.1145/3505633.

2. A. K. Jones et al., "Considering fabrication in sustainable computing," in *Proc. IEEE/ACM Int. Conf. Comput.-Aided Des.*, 2013, pp. 206–210.
3. R. Bennis, "Life cycle assessment of dell poweredge r740," Dell, Jun. 2019. [Online]. Available: https://corporate.delltechnologies.com/content/dam/digitalassets/active/en/unauth/data-sheets/products/servers/lca_poweredge_r740.pdf
4. E. Brunvand, D. Kline, and A. K. Jones, "Dark silicon considered harmful: A case for truly green computing," in *Proc. 9th Int. Green Sustain. Comput. Conf.*, 2018, pp. 1–8.
5. S. Ollivier, S. Longofono, P. Dutta, J. Hu, S. Bhanja, and A. K. Jones, "CORUSCANT: Fast efficient processing-in-racetrack memories," in *Proc. IEEE/ACM Symp. Microarchit.*, 2022, pp. 784–798.
6. S. Ollivier, X. Zhang, Y. Tang, C. Choudhuri, J. Hu, and A. K. Jones, "POD-RACING: Bulk-bitwise to floating-point compute in racetrack memory for machine learning at the edge," *IEEE Micro*, vol. 42, no. 6, pp. 9–16, Nov./Dec. 2022.
7. "Environmental management–life cycle assessment – requirements and guidelines," ISO, Geneva, Switzerland, Tech. Rep. 14044, 2006.
8. M. Garcia Bardon et al., "DTCO including sustainability: Power-performance-area-cost-environmental score (PPACE) analysis for logic technologies," in *Proc. IEEE Int. Electron Devices Meeting*, 2020, pp. 41.4.1–41.4.4.
9. S. B. Boyd, *Life-Cycle Assessment of Semiconductors*. Berlin, Germany: Springer, 2011.
10. D. Kline et al., "GreenChip: A tool for evaluating holistic sustainability of modern computing systems," *Sustain. Comput., Inform. Syst.*, vol. 22, pp. 322–332, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2210537917300823
11. R. Bläsing et al., "Magnetic racetrack memory: From physics to the cusp of applications within a decade," *Proc. IEEE*, vol. 108, no. 8, pp. 1303–1321, Aug. 2020.
12. Y. Zhang, W. Zhao, D. Ravelosona, J.-O. Klein, J.-V. Kim, and C. Chappert, "Perpendicular-magnetic-anisotropy CoFeB racetrack memory," *J. Appl. Phys.*, vol. 111, no. 9, 2012, Art. no. 093925.
13. J. S. Vetter and S. Mittal, "Opportunities for nonvolatile memory systems in extreme-scale high-performance computing," *Comput. Sci. Eng.*, vol. 17, no. 2, pp. 73–82, 2015.
14. R. Venkatesan, M. Sharad, K. Roy, and A. Raghunathan, "DWM-TAPESTRI-an energy efficient all-spin cache using domain wall shift based writes," in *Proc. Des., Autom. Test Eur. Conf. Exhib.*, 2013, pp. 1825–1830.
15. R. Venkatesan, V. Kozhikkottu, C. Augustine, A. Raychowdhury, K. Roy, and A. Raghunathan, "TapeCache: A high density, energy efficient cache based on domain wall memory," in *Proc. ACM/IEEE Int. Symp. Low Power Electron. Des.*, 2012, pp. 185–190.
16. V. Seshadri et al., "Ambit: In-memory accelerator for bulk bitwise operations using commodity dram technology," in *Proc. 50th Annu. IEEE/ACM Int. Symp. Microarchit.*, 2017, pp. 273–287.
17. X. Xin, Y. Zhang, and J. Yang, "ELP2IM: Efficient and low power bitwise operation processing in dram," in *Proc. IEEE Int. Symp. High Perform. Comput. Archit.*, 2020, pp. 303–314.
18. H. Yu et al., "Energy efficient in-memory machine learning for data intensive image-processing by non-volatile domain-wall memory," in *Proc. 19th Asia South Pacific Des. Autom. Conf.*, 2014, pp. 191–196.
19. I. Bayram, E. Eken, D. Kline, N. Parshook, Y. Chen, and A. K. Jones, "Modeling STT-ram fabrication cost and impacts in NVSIM," in *Proc. 7th Int. Green Sustain. Comput. Conf.*, 2016, pp. 1–8.
20. X. Dong, C. Xu, Y. Xie, and N. P. Jouppi, "NVSim: A circuit-level performance, energy, and area model for emerging nonvolatile memory," *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.*, vol. 31, no. 7, pp. 994–1007, Jul. 2012.
21. T. Higgs, M. Cullen, M. Yao, and S. Stewart, "Developing an overall CO_2 footprint for semiconductor products," in *Proc. IEEE Int. Symp. Sustain. Syst. Technol.*, 2009, pp. 1–6.
22. T. Mai et al., "Exploration of high-penetration renewable electricity futures," National Renewable Energy, Golden, CO, USA, Tech. Rep. NREL/TP-6A20-52409-1. [Online]. Available: http://www1.eere.energy.gov/library/viewdetails.aspx?productid=5846
23. N. Popovich and B. Plumer, "How does your state make electricity?," *The New York Times*, Oct. 2020. [Online]. Available : https://www.nytimes.com/interactive/2020/10/28/climate/how-electricity-generation-changed-in-your-state-election.html

SÉBASTIEN OLLIVIER is a postdoctoral associate at the University of Pittsburgh, PA, 15260, USA. His research interests include domain-wall memory, memory reliability, and processing in memory applications. Ollivier received a Ph.D. degree in electrical and computer engineering from the University of Pittsburgh, Pittsburgh. Contact him at sbo15@pitt.edu.

SHENG LI is a doctoral student in the Department of Computer Science, University of Pittsburgh, Pittsburgh, PA, 15260, USA. His research interests include computer architecture, edge computing, and systems for machine learning. Li received a B.E. degree in software engineering from Sichuan University, Chengdu, China. Contact him at shl188@pitt.edu.

YUE TANG is a Ph.D. candidate in electrical and computer engineering at the University of Pittsburgh, Pittsburgh, PA, 15260, USA. Her research interests include FPGA-based CNN training and on-device artificial intelligence. Tang received an M.S. degree from the School of Automation Science and Electrical Engineering, Beihang University, Beijing, China. Contact her at yut51@pitt.edu.

STEPHEN CAHOON is a graduate student studying computer engineering at the University of Pittsburgh, Pittsburgh, PA, 15260, USA. His research interests include processing in memory and domain-wall memory. Cahoon received a B.E. degree in computer engineering from the University of South Alabama, Mobile, AL, USA. Contact him at stc127@pitt.edu.

RYAN CAGINALP is an undergraduate student in electrical and computer engineering at the University of Pittsburgh, Pittsburgh, PA, 15260, USA. Through his curriculum, he worked on sustainability and Flash and DRAM radiation testing under the supervision of professor Alex K. Jones. Contact him at rlc113@pitt.edu.

CHAYANIKA CHAUDHURI is a research volunteer in electrical and computer engineering at the University of Pittsburgh, Pittsburgh, PA, 15260, USA, under the supervision of professor Alex K. Jones. Her research interests focus on novel memories as domain-wall memory. Contact her at roc74@pitt.edu.

PEIPEI ZHOU is an assistant professor in the Department of Electrical and Computer Engineering at the University of Pittsburgh, Pittsburgh, PA, 15260, USA. Her research interests include customized computer architecture and programming abstraction for applications including healthcare, e.g., precision medicine and artificial intelligence. Zhou received a Ph.D. degree in computer science from the University of California, Los Angeles, CA, USA. Contact her at peipei.zhou@pitt.edu.

XULONG TANG is an assistant professor in the Department of Computer Science, University of Pittsburgh, Pittsburgh, PA, 15260, USA. His research interests include high-performance computing, advanced computer architecture designs, and compilers. Tang received a Ph.D. degree in computer science and engineering from The Pennsylvania State University, State College, PA. Contact him at xulongtang@pitt.edu.

JINGTONG HU is an associate professor and William Kepler Whiteford faculty fellow in the Department of Electrical and Computer Engineering, University of Pittsburgh, Pittsburgh, PA, 15260, USA. His research interests include hardware/software co-design for machine learning algorithms, on-device AI, and embedded systems. Hu received a Ph.D. degree in computer science from the University of Texas at Dallas, Richardson, TX, USA. He is a Senior Member of IEEE. Contact him at jthu@pitt.edu.

ALEX K. JONES is a professor of electrical and computer engineering and computer science at the University of Pittsburgh, Pittsburgh, PA, 15260, USA. He is currently a program director at the U.S. National Science Foundation in the CNS Division of the CISE Directorate. His research interests include compilation for configurable systems and architectures, scaled and emerging memory, reliability, fault tolerance, and sustainable computing. Jones received a Ph.D. degree in electrical and computer engineering from Northwestern University, Evanston, IL, USA. He is a Senior Member of IEEE and ACM. Contact him at akjones@pitt.edu.