

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

Abstract—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 SWaP constraints (Size, Weight, and Power) 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 paper we explore the tradeoffs of hardware strategies for convolutional neural network acceleration engines considering inference and on-line training. In particular, we explore the use of mobile GPU accelerators, recently released edge-class FPGAs, and novel PIM using 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.

■ INTRODUCTION

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. Additionally, 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.

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 green-house warming gases. 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, green-house warming gases (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 (run-time) effects [3]. For mobile devices and compact computers, embodied impacts can reach 80-90% of total life-cycle impacts and that these impacts are dominated by their integrated circuits [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.

In this paper we explore 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 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 paper 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 life-cycle assessment, indifference, and breakeven analyses. We also provide background on Racetrack memory including how it is used for PIM and its required extension for life-cycle assessment. We also mention the features about CNN inference and training that lead to different assumptions about datatypes.

Life-cycle Assessment

The primary source of environmental impacts for computing systems comes from the integrated circuits (ICs) 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 Life Cycle Assessment (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/output LCA.

Semiconductor process LCA explores the impact of the different steps of the approximately 20 masks required to build CMOS circuits. These masks can be broken down into their individual steps, such as deposition, lithography, etching, metrology, etc., 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 like 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 (EUV) 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 nm to 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

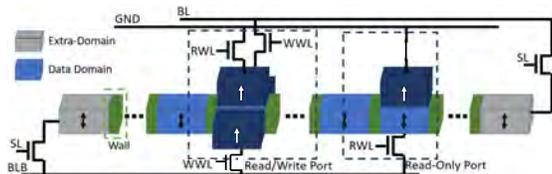


Figure 1. Anatomy of a domain-wall memory nanowire.

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 Eq. 1 [10]. 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 Eq. 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 $M_0 = 0$.

$$t_I = \frac{M_1 - M_0}{P_0 - P_1} \quad t_B = \frac{M_1}{P_0 - P_1} \quad (1)$$

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 Race-track memory [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 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 non-volatility. 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 is 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 Racetrack Memory

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 *multi-domain 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 multi-operand bulk bitwise logical operations. The number of operands is dictated by the size of the multi-domain read.

Arithmetic structures such as addition can be constructed by converting a multi-domain 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 magnetic 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 ferromagnetic 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 ferromagnetic material like 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 point-wise 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 $N \times R_{in} \times C_{in}$ 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:

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

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

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 nm–32 nm [9] processes and for 28 nm–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 nm [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

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 ^{1,4}	55 ¹	32 ^{2,4}	32 ^{3,4}	7 ³	14 ³
Die Size (mm ²)	38	73	38	38	324	350
Die per wafer	1847	967	1847	1847	217	201
PE (kWh/Wafer)	1600	1200	1206	753	1482	882
Energy (MJ/die)	3.12	4.47 ⁵	2.35	1.47	24.59	15.80
AZ (gCO ₂ eq/die)	343	490 ⁵	259	162	2698	1734
CA (gCO ₂ eq/die)	203	291 ⁵	153	95	1598	1027
TX (gCO ₂ eq/die)	380	544 ⁵	286	179	2992	1922
NY (gCO ₂ eq/die)	163	233 ⁵	123	77	1284	825

¹ Calculated using process LCA from [9].

² Calculated using process LCA from [21].

³ Calculated using process LCA from [8].

⁴ Requires extra steps for spintronics [19].

⁵ Requires 16 dies to build a the tested 1GB DIMM.

Table 2. Energy to gCO₂eq/kWh [22] and Grid Mixes [23]

Source	gCO ₂ eq/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 ₂ eq/kWh)		395	234	438	188

process LCA study. We discuss this in more detail in the supplementary material.

System Embodied Energy and Carbon Study

Several grid mix scenarios for CO₂eq based on CO₂eq 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, California, Texas, and New York, all have very different grid mixes.

Arizona and Texas have significant electrical generation from coal and the highest generation from natural gas. While Arizona has significant generation from nuclear plants and Texas has significant wind energy, their 395 and 438 gCO₂eq/kWh (CO₂ equivalent generated per kWh) are much higher than California and New York, which still get more than a third of their electricity from natural gas. California is very balanced on renewable energy and New York has significant hydroelectric and nuclear power generation, thus their grid mix generates about half the GWG emissions at 234 and 188gCO₂eq/kWh, respectively.

In Table 1 we report the embodied energy

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 W	Efficiency	
		Lat.(S)	FPS		FPS/W	MF/gCO ₂ eq
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
	RM	0.0020	490	0.93	526	4.6–10.8
Training Acceleration using Floating-Point 32 Data						
Benchmark	Target	Performance		Power W	Efficiency	
		Lat.(S)	GFLOPS		GFLOPS/W	TFLOPS/gCO ₂ eq
Alexnet	GPU	0.005	1335	21.05	63.4	521–1214
	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
Idle Power (W)		FPGA = 9.6	GPU = 3.0	DDR3 = 1.2	RM = 0.8	

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 ELP²IM [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, 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 gCO₂eq.

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.

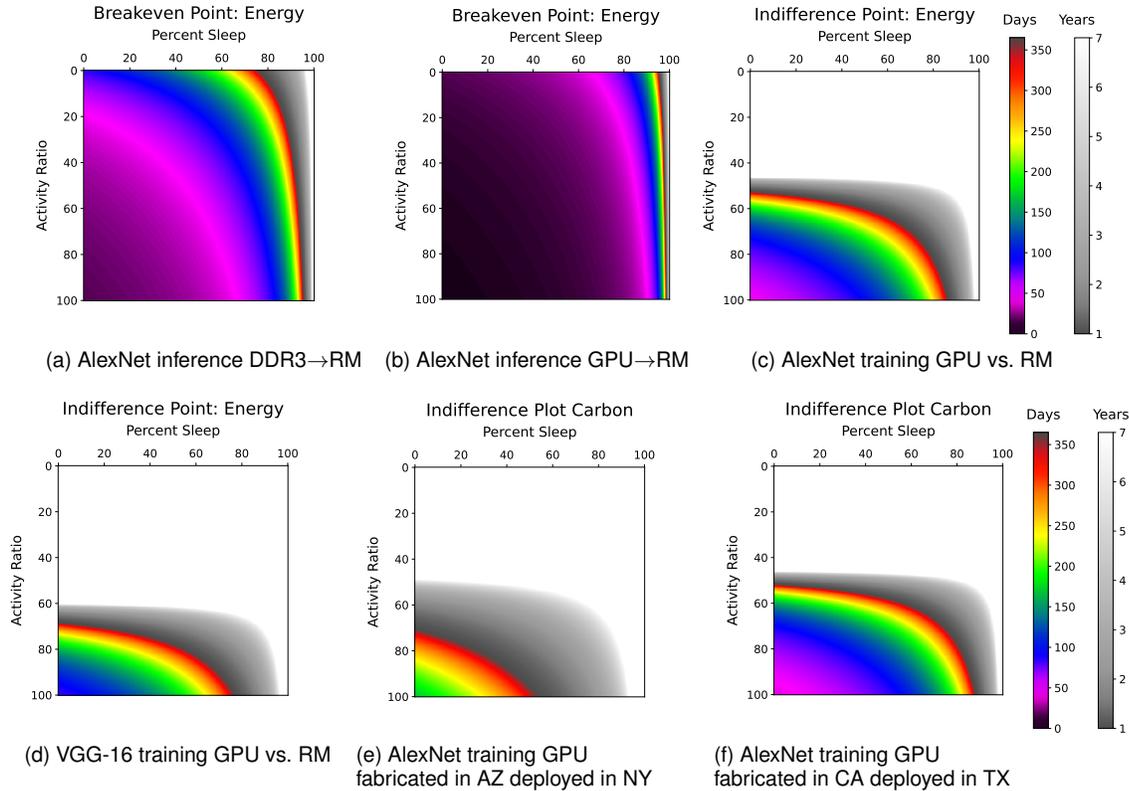


Figure 2. Sustainability analyses of different accelerator choices for edge systems.

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 2a. 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 (Figure 2a), 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 (towards top left) or sleeping (towards bottom right) or both (towards 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 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 energy 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 Figures 2c and 2d for AlexNet and VGG-16, respectively. Both applications can benefit from the GPU in high usage scenarios (bottom left), but if the system does training infrequently, the GPU savings during training cannot overcome the higher embodied energy. For an 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 2e for on-line training of AlexNet, we explore the case where fabrication takes place in AZ which has a comparatively high CO₂eq/kWh and the system is deployed in NY with a relatively low CO₂eq/kWh. Even in the highest utilization case the indifference point becomes six months, and in lower utilization (circa 70%) it becomes 1 year, and quickly grows to multiple years as the utilization drops towards 50%, favoring the RM for relatively more usage scenarios.

In Figure 2f, 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 2 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.

CONCLUSION

In this work, we compared several SWaP optimized CNN accelerators popular for edge servers for both inference and on-line 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 like 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.

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 non-intuitive 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.

ACKNOWLEDGMENT

This work was supported in part by the NSF under grants CNS-1822085, CNS-2133267, the National Security Agency, and 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. [Online]. Available: <https://doi.org/10.1145/3505633>
2. A. K. Jones, Y. Chen, W. O. Collinge, H. Xu, L. A. Schaefer, A. E. Landis, and M. M. Bilec, "Considering fabrication in sustainable computing," in *2013 IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, 2013, pp. 206–210.
 3. R. Bennis, "Life cycle assessment of dell poweredge r740," Dell, June 2019, 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 *2018 Ninth International Green and Sustainable Computing Conference (IGSC)*, 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 *Proceedings of the IEEE/ACM Symposium on Microarchitecture*, 2022, [preprint available online:] <https://arxiv.org/abs/2108.01202>.
 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*, 2022, [Preprint Appears Online] <https://arxiv.org/abs/2108.01202>.
 7. ISO, "Environmental management – life cycle assessment – requirements and guidelines," Tech. Rep. 14044, 2006.
 8. M. Garcia Bardon, P. Wuytens, L.-Å. Ragnarsson, G. Mirabelli, D. Jang, G. Willems, A. Mallik, A. Spessot, J. Ryckaert, and B. Parvais, "Dtco including sustainability: Power-performance-area-cost-environmental score (ppace) analysis for logic technologies," in *2020 IEEE International Electron Devices Meeting (IEDM)*. IEEE, 2020, pp. 41–4.
 9. S. B. Boyd, *Life-cycle assessment of semiconductors*. Springer Science & Business Media, 2011.
 10. D. Kline, N. Parshook, X. Ge, E. Brunvand, R. Melhem, P. K. Chrysanthis, and A. K. Jones, "Greenchip: A tool for evaluating holistic sustainability of modern computing systems," *Sustainable Computing: Informatics and Systems*, vol. 22, pp. 322–332, 2019. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S2210537917300823>
 11. R. Bläsing, A. A. Khan, P. C. Filippou, C. Garg, F. Hameed, J. Castrillon, and S. S. P. Parkin, "Magnetic racetrack memory: From physics to the cusp of applications within a decade," *Proceedings of the IEEE*, vol. 108, no. 8, pp. 1303–1321, 2020.
 12. Y. Zhang, W. Zhao, D. Ravelosona, J.-O. Klein, J.-V. Kim, and C. Chappert, "Perpendicular-magnetic-anisotropy cofeb racetrack memory," *Journal of Applied Physics*, vol. 111, no. 9, p. 093925, 2012.
 13. J. S. Vetter and S. Mittal, "Opportunities for nonvolatile memory systems in extreme-scale high-performance computing," *Computing in Science Engineering*, 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. of DATE*, 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. of ISLPED*, 2012, pp. 185–190.
 16. V. Seshadri, D. Lee, T. Mullins, H. Hassan, A. Boroumand, J. Kim, M. A. Kozuch, O. Mutlu, P. B. Gibbons, and T. C. Mowry, "Ambit: In-memory accelerator for bulk bitwise operations using commodity dram technology," in *2017 50th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO)*. IEEE, 2017, pp. 273–287.
 17. X. Xin, Y. Zhang, and J. Yang, "Elp2im: Efficient and low power bitwise operation processing in dram," in *2020 IEEE International Symposium on High Performance Computer Architecture (HPCA)*. IEEE, 2020, pp. 303–314.
 18. H. Yu, Y. Wang, S. Chen, W. Fei, C. Weng, J. Zhao, and Z. Wei, "Energy efficient in-memory machine learning for data intensive image-processing by non-volatile domain-wall memory," in *2014 19th Asia and South Pacific Design Automation Conference (ASP-DAC)*. IEEE, 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 *2016 Seventh International Green and Sustainable Computing Conference (IGSC)*, 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 Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 31, no. 7, pp. 994–1007, 2012.
 21. T. Higgs, M. Cullen, M. Yao, and S. Stewart, "Developing an overall co₂ footprint for semiconductor products," in *2009 IEEE International Symposium on Sustainable Systems and Technology*, 2009, pp. 1–6.

22. T. Mai, R. Wiser, D. Sandor, G. Brinkman, G. Heath, P. Denholm, D. Hostick, N. Darghouth, A. Schlosser, and K. Strzepek, "Exploration of high-penetration renewable electricity futures," Tech. Rep. NREL/TP-6A20-52409-1, <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.

Sebastien Ollivier graduated from the French engineering school ENSEA (Ecole Nationale Supérieure de l'Electronique et ses Applications) and the University of Pittsburgh with a master's degree in Electrical Engineering in 2018. He completed his PhD in 2022 at the University of Pittsburgh in electrical and computer engineering. He published seven articles focusing on domain-wall memory using novel memory reliability and processing in memory applications.

Sheng Li is currently a doctoral student in the Department of Computer Science, University of Pittsburgh. He received his B.E. degree in Software Engineering from Sichuan University in 2020. His research interests include computer architecture, edge computing, and systems for machine learning.

Yue Tang received the BS and MS degrees from the School of Automation Science and Electrical Engineering at the Beihang University, China. She is currently a PhD candidate at the University of Pittsburgh in Electrical and Computer Engineering. Her research interests include FPGA-based CNN training and on-device artificial intelligence.

Chayanika Chaudhuri is a research volunteer at the University of Pittsburgh in Electrical and Computer Engineering under the supervision of Professor Alex K. Jones. Her research interests are focused on novel memories as Domain-Wall Memory.

Peipei Zhou received her BE degree from Chiung Wu Honor College at Southeast University, China in 2012 and MS in Electrical Engineering, PhD in Computer Science from the University of California, Los Angeles in 2014, 2019 respectively. She is currently an assistant professor in the Department of Electrical and Computer Engineering at the University of Pittsburgh, Pittsburgh, PA, USA. Dr. Zhou's research interests include customized computer architecture and programming abstraction for applications including healthcare, e.g., precision medicine and artificial intelligence.

Xulong Tang is currently an assistant professor in the Department of Computer Science, University of Pittsburgh. He received his Ph.D. degree in Computer Science and Engineering from The Pennsylvania State University in 2019. His research interests include high-performance computing, advanced computer architecture designs, and compilers.

Jingtong Hu received his BE degree from School of Computer Science and Technology at Shandong University, China in 2007 and his MS and PhD degrees in Computer Science from University of Texas at Dallas, USA in 2010 and 2013, respectively. He is currently an Associate Professor and William Kepler Whiteford Faculty Fellow in the Department of Electrical and Computer Engineering at University of Pittsburgh, Pittsburgh, PA, USA. His current research interests include hardware/software co-design for machine learning algorithms, on-device AI, embedded systems.

Alex K. Jones received the BS degree in 1998 in physics from the College of William and Mary in Williamsburg, VA, USA, and the MS and PhD degrees in 2000 and 2002, respectively, in ECE from Northwestern University, Evanston, IL, USA. He is a Professor of ECE and CS at the University of Pittsburgh, Pittsburgh, PA, USA. He is currently serving as a Program Director at the US NSF in the CNS Division of the CISE Directorate. Dr. Jones' research interests include compilation for configurable systems and architectures, scaled and emerging memory, reliability, fault tolerance, and sustainable computing. He is the author of more than 200 publications in these areas. His research is funded by the NSF, DARPA, NSA, and industry. He is currently an associate editor of IEEE Transactions on Computers and Journal of Sustainable Computing: Informatics and Systems. He is a senior member of the IEEE and the ACM.