Comparative Analysis of GPU, TPU and IPU Architectures in Modern Computing Systems

Md. Maniruzzaman
Department of Electrical Engineering,
School of Engineering,
San Francisco Bay University,
Fremont, CA 94539, USA.
mmaniruz158@student.sfbu.edu

Abstract— Such fast growth in demand for high-performance computing in AI/ML generated specialized processors like GPUs, TPUs, IPUs—each tuned to one or the other basic computational challenges inherent in AI/ML workloads and offered their own levels of parallelism, efficiency, and scalability. Its highly parallel structure makes the GPU the source of choice for general AI workloads. This is especially true when training deep neural networks, though inefficiencies connected with the use of GPUs for processing some operations have grown more stark as AI models have become complex, opening routes to TPUs and IPUs. The TPU—chips developed by Google—are optimized for applications based on TensorFlow, do very well on matrix-heavy computations, and hence are particularly suitable for high-throughput, low-power large-scale AI environments. IPUs are a more recent innovation that provides very fine-grained parallelism and low-latency processing; they are quite perfect for irregular and complex computations characteristic of the new generation of AI models. The paper explains architectural differences between GPUs, TPUs, and IPUs, their respective roles in modern computing systems, and reasons why heterogeneity at the processor design level drives AI and ML forward. This would be the guiding analysis for our observation: diversified processing units would be required to meet the dynamic requirements AI/ML applications are increasingly placing on them and would foster innovation in hardware design and software optimization.

Keywords - AI accelerators, GPU, TPU, IPU, parallel computing, deep learning, machine learning, processor architecture, high-performance computing, neural networks, systolic arrays, low-latency processing, energy efficiency.

I. INTRODUCTION

The coming of age of AI and ML has changed many industries, forcing the need for bespoke computing architectures that can accommodate huge computational needs of these kinds of technologies. Traditional CPUs, though having been the mainstay of computing systems for quite a long, have recently shown their deficiencies in parallel processing transparently with the increasing requirements in computational power and efficiency from the newer AI and ML workloads. These compete with the challenges in developing a broad span of processors gaining their architectural advantages by being tailored to specific aspects of AI/ML tasks, including Graphics Processing Units, Tensor Processing Units, and Intelligence Processing Units.

The GPUs were majorly designed for graphics rendering, particularly for video games and simulations. However, massively parallel processing has fitted quite well into the acceleration of deep models. With the ability to perform many varied tasks simultaneously, a GPU may have thousands of cores. Specifically, it does very well at those matrix and vector operations at a center of the training of a neural network. This has become a de facto standard for flexibility and wide applicability, particularly with AI and ML workloads. But even as versatile as they are, there are some limitations to using a GPU.

To answer these deficiencies, Google introduced the TPU, a custom-built application-specific integrated circuit designed to meet AI/ML workloads. This makes TPUs really powerful in special environments where TensorFlow is the biggest framework in use, while they may be sparser elsewhere due to their narrow focusing on optimization. Inverse to these very specialized TPUs are IPUs, a new breed of processors designed for flexibility without performance compromise across a wide variety of AI tasks at hand.

Companies like Graphcore have designed IPUs, which process very parallel, irregular workloads, generated by advanced AI applications. In contrast to GPUs and TPUs, they were predominantly optimized for dense matrix operations, while the typical computing work IPUs were optimized for sparse data and complex computation patterns, typical for modern AI research. It is capable of doing that; thus, IPUs are very instrumental in scenarios where performing such tasks might give traditional processors a bit of trouble, much like running models requiring dynamic and flexible computation across a wide array of tasks. The ability to work with the strengths of different architectures will definitely be very imperative for further computational efficiency in the realization of the next generation of applications of AI. The paper has provided an in-depth architectural review of GPU, TPU, and IPU concerning their design principles, operational efficiencies, and suitability for different kinds of AI/ML tasks. Only by understanding their respective strengths and weaknesses can more reasonable choices about hardware selection and system design for AI/ML workloads be made.

II. METHODOLOGY

This study employs a comprehensive comparative analysis to evaluate the architectural differences, performance characteristics, and application suitability of three advanced computing architectures: Graphics Processing Units (GPUs), Tensor Processing Units (TPUs), and Intelligence Processing Units (IPUs). The methodology is structured around three core components: literature review, performance benchmarking, and architectural analysis, each tailored to provide a nuanced understanding of the strengths and

limitations of these processors in the context of AI and machine learning workloads.

A. Literature Review

A detailed literature review in the area of scientific and technical reports, as well as university discussion papers, was conducted to identify development, architecture, and performance metrics related to GPUs, TPUs, and IPUs.

B. Performance Benchmarking

This paper presents an empirical performance evaluation of GPUs, TPUs, and IPUs using a set of performance benchmarks widely adopted by the AI/ML community. In this paper, some benchmarks are chosen that use assignments which most probably will represent the typical AI workloads:

- Processing Speed: This floating-point operations per second measure is a pure indicator of the raw processing power that is available in each processor.
- Energy Efficiency: This quantifies how efficiently different processors transform electrical power into computational work, a consideration which becomes important when dealing with large-scale deployments.
- Memory Bandwidth is a measure of the rate at which data can be read from or written to memory. This is the overall performance and applies most importantly to AI/ML tasks working with large datasets.
- Latency: The time between the input of a command and its execution; for example, in real-time AI applications, the minimum latency will be the one appropriate for autonomous cars and other interactive systems with real-time AI applications.

Testing was done using standard AI models, including ResNet-50, on the architectures for comparable benchmarks. Generic models were run using each processor type and its corresponding optimized software stack: CUDA, TensorFlow, and Poplar SDK, respectively.

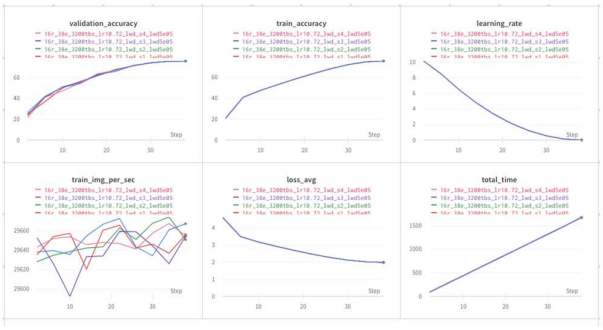

Figure 1: ResNet-50 at scale on IPU Hardware

C. Architectural Analysis

Architectural Nuances Besides raw performance metrics, this work dug into the architectural nuances of GPUs, TPUs, and IPUs to understand why one processor works better for some tasks than others. Specifically, the design details of every processor are explained as follows:

- Core Architecture: This is the design and functionality of processing cores. Similarly, systolic arrays are used in TPUs and fine-grained parallelism in IPUs that are tuned for tensors and irregular patterns respectively.
- Memory Hierarchy: It is the description of the structure and the availability of different levels of memory, ranging from Cache to DRAM.
- Interconnects and Communication Efficiency in data exchange from one core to another core, between chips, with external memory - this becomes a significant parameter in multi-core processors like GPUs and multi-chip systems like TPUs.
- Scalability How each of the architectures scale well with increasing complexity and size of AI models, it turns into an important factor in future-proofing AI infrastructure that's built.

The analysis was then to correlate architectural features to performance observed outcomes and provide insights as to how design choices may be specifically efficiency and effectiveness relative to the orchestration of AI/ML workloads.

D. Comparative Assessment

The final step of the methodology was to make a comparative assessment by synthesizing the outcome from literature review, performance benchmarking and architectural review. The extent of evaluation, therefore, was to determine, based on the parameters of the study, under which scenarios each form of the processors was leading, and where it could in turn, be weak. It is the mapping of specific AI/ML tasks to processor architectures best supporting these tasks that gave practical guidance to the organizations on the choice of the right hardware for these kinds of AI applications.

That comparison, furnished in a key at-a-glance form of a look-up table, brought out the flexibility-performance-efficiency trade-offs in the energy use for the three processor types. Detailed approaches bring out critical insight into how GPUs, TPUs, and IPUs can contribute within the wider AI ecosystem, giving a strategic approach to the different computer environments.

III. DISCUSSION

The discussion section has been reserved for detailed relative comparisons of strengths and weaknesses of GPUs, TPUs, and IPUs in relation to their respective roles in accelerating AI/ML workloads. Each processor archetype also comes with unique architectural innovations that address special computational demands; knowing these differences is very important in choosing appropriate hardware amidst a myriad of AI/ML applications.

A. Graphics Processing Units (GPUs)

Parallel computing and GPUs have gone inseparably together for years, more so in domains such as AI and ML. They were basically designed for rendering graphics. It is this inherent parallelism that has helped a lot in large-scale matrix and vector operations involved in the training stage of DNNs.

A normal GPU architecture has numerous streaming multiprocessors; each contains hundreds of cores capable of executing floating-point operations concurrently. With this

design, a GPU would be hugely powerful to hold heavy computations that Artificial Intelligence models require - for instance, convolutions in Convolutional Neural Networks [1]. On the other side, this general-purpose nature of GPUs brings in some limitations with their versatility.

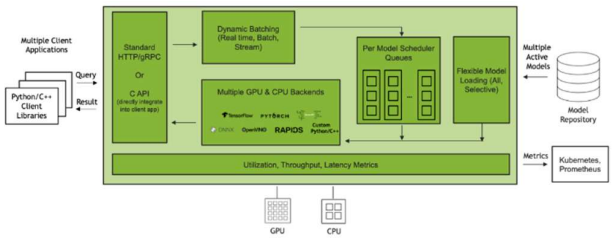

Figure 2: GPU Architecture Overview

All types of AI/ML workloads are not inherently optimized on a GPU. The architecture does really well for regular, structured computations — like those involved in CNNs. However, the effectiveness of a GPU reduces while conducting more irregular types of computation or when running models that require the processing of sparse data. Moreover, high energy consumption of a GPU may turn out to be a concern while scaling up huge AI models in data centers. All these shortcomings notwithstanding, the dominance of GPUs in AI computing due to flexibility and maturity of software ecosystems like CUDA and cuDNN is unabated.

B. Tensor Processing Units TPUs

In marked contrast, TPUs ushered in an era of special-purpose computing attuned to support AI and ML tasks. Designed by Google, TPUs are targeted to accelerate TensorFlow operations particularly large-scale matrix multiplications at the heart of many AI models.

	Google TPUv4	**TPUv3**
Production deployment	2020	2018
Peak TFLOPS	275 (bf16 or int8)	123 (bf16)
Clock Rate	1050 MHz	940 MHz
Tech. node, Die size	7 nm, <600 mm2	16 nm, < 700 mm2
Transistor count	22 billion	10 billion
Chips per CPU host	4	8
TDP	N.A.	N.A.
Idle, min/mean/max power	90, 121/170/192 W	123, 175/220/262 W
Inter Chip Interconnect	6 links @ 50 GB/s	4 links @ 70 GB/s
Largest scale configuration	4096 chips	1024 chips
Processor Style	Single Instruction 2D Data	Single Instruction 2D Data
Processors / Chip	2	2
Threads / Core	1	1
SparseCores / Chip	4	2
On Chip Memory	128 (CMEM) + 32 MiB (VMEM) + 10 MiB (spMEM)	32 MiB (VMEM) + 5 MiB (spMEM)
Register File Size	0.25 MiB	0.25 MiB
HBM2 capacity, BW	32 GiB, 1200 GB/s	32 GiB, 900 GB/s

Figure 3: TPU v3 Architecture and Performance Metrics.

TPUs accomplish this via a systolic array architecture, which is hardware singularly well-suited for the intrinsic repetitiveness of tensor operationsThis repetitiveness leads to the following: specializations whose outcome - the TPU can allow for high throughputs and far greater energy efficiency in large-scale AI training tasks than current-generation GPUs. Definitely one of the biggest benefits of the TPU is its huge batches of data, all processed at once. With the use of larger batches, it is possible to increase the speed of training for deep learning models.

The TPUs are also designed natively to support lower-precision operations, such as bfloat16 operations, further improving performance and bringing down memory bandwidth requirements without loss of model accuracy. But this TPU specialization also limits its flexibility. They're overwhelmingly optimized for TensorFlow-based workloads; their performance may not be as good in environments that require non-TensorFlow frameworks or non-standard AI models. Another weakness of the TPU architecture is that it excels in specific tasks, but with irregular or sparse data a general-purpose GPU or even the newer IPUs can be better at handling tasks.

C. Intelligence Processing Units (IPUs)

IPUs are a new class of evolution in AI-specialized processors, designed to help overcome some of the shortcomings noted in both the GPU and TPU.

IPUs are designed to provide support for highly parallel, irregular workloads typical of state-of-the-art AI applications

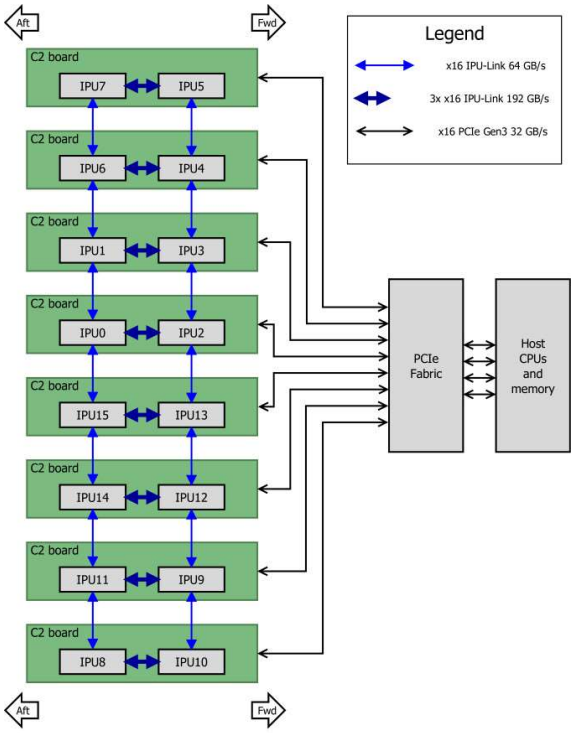

Figure 4: IPU Architecture and Core Configuration

. IPUs also make a case for the execution of complex AI models with lower latency, thus becoming a prime choice for real-time AI applications where latency is a critical essence. In addition to increased performance at lower power, IPUs offer advanced programmability through the Poplar SDK, which allows a developer to tune models in order to strive for the best out of the architecture of the IPU. However, like any dedicated hardware, IPUs have their own trade-offs.

While they excel in performance for specific AI workloads, under no circumstances will they be compatible with the raw computational power of TPUs or the versatility of GPUs in every scenario. IPUs are a young technology; their software ecosystem - growing very fast - does not enjoy the same maturity as that of GPUs and TPUs [3][4]. This makes it difficult for developers to incorporate IPUs into their workflows, particularly when their AI models are not well-matched for the strengths of the IPU architecture.

D. Comparative Analysis

From this comparative analysis of the three types of processors - GPUs, TPUs, and IPUs - it is evident that no one processor type can be said to be superior for all AI/ML tasks. The truth is that, driven by performance and applicability across different scenarios, each architecture has been optimized for certain kinds of workloads. On their part, due to the fact that they are flexible and have a wide reach of applications, GPUs remain very suitable for a wide range of AI tasks, particularly those that require high parallelism in structured computations. TPU provides peak performance for applications based on TensorFlow, particularly in environments where matrix operations dominate. On the other hand, IPUs show the best performance for highly irregular and complex models of AI that require low-latency processing and have the same correspondingly dynamic computation power needs accordingly.

Table 1: Comparative Performance Metrics of GPU, TPU, and IPU

	#	Metric	Google TPU v3	Nvidia V100	Nvidia A100	Cerebras WSE	GraphCore IPU1	GraphCore IPU2
Raw Metrics	1	Technology node	>12nm (16 nm est.)	TSMC 12 nm	TSMC 7 nm	TSMC 16 nm	TSMC 16 nm	TSMC 7 nm
	2	Die Area (mm2)	<648 (600 est.)	815	826	46225	900 (est.)	823
	3	Transistor Count (B)	11 (est.)	21	54.2	1200	23.6	59.4
	4	Architecture	Systolic Array	SIMD + TC	SIMD + TC	MIMD	MIMD	MIMD
	5	Theoretical TFLOPS (16-bit mixed precision)	123	125	312	2500	125	250
	6	Freq (GHZ)	0.92	1.5	1.4	Unknown	1.6	Unknown
	7	DRAM Capacity (GB)	32	32	80	N/A	N/A	112
	8	DRAM BW (GB/sec)	900	900	2039	N/A	N/A	64 (est.)
	9	Total SRAM Capacity	32MB	36 MB (RF+L1+L2)	87 MB (RF+L1+L2)	18 GB	300 MB	900 MB
	10	SRAM BW (TB/sec)	Unknown	224 @RF + 14 @L1 + 3 @L2	608 @RF+ 19 @L1 + 7 @L2	9000	45	47.5
	11	Max TDP (Watts)	450	450	400	20K	150	150 (est.)
Efficiency Metrics	12	GEMM Achievable TFLOPS	98% (120 TFLOPS)	88% (110 TFLOPS)	93% (290 TFLOPS)	Unknown	47% (58 TFLOPS)	61% (154 TFLOPS)
	13	Energy Efficiency (Achievable GEMM TFLOPS/Max Watts)	0.26	0.24	0.72	Unknown	0.39	1.0
	14	Theoretical Energy Efficiency (Theoretical TFLOPS/Max Watts)	0.27	0.27	0.78	0.125	0.83	1.6
	15	Memory Capacity (GB)	16	32	80	18	0.3	112
	16	Memory Efficiency (FLOP/DRAMByte)	133	122	158	N/A	N/A	Unknown
	17	Memory Efficiency (FLOP/SRAMByte)	Unknown	32	35	Unknown	1.28	3.2
	18	Area Efficiency (Achievable TFLOPS/mm2)	0.2	0.13	0.35	Unknown	0.06	0.17
	19	Area Efficiency (Achievable TFLOPS/BTran)	11	5.2	5.3	Unknown	2.5	2.6

In Table 1, architectural differences and suitable workloads are summarized. Processor choices will have to be enforced by the AI/ML workload requirements under consideration. For instance, in case situations arise with the need to train huge deep learning models quickly and efficiently, this will be best done by TPUs since their architecture is optimized for tensor operations. At the same time, IPUs will excel in tasks with real-time inference or extraordinarily irregular patterns of data in the models. Again, one of the places that general-purpose GPUs hold is versatile - they master a very broad spectrum of AI tasks with reasonable efficiency for both research and production scenarios.

E. Implications for Future AI/ML Hardware Development

If this present evolution of AI/ML models is anything to go by, future hardware development will be continuously in the trend of specialization and optimize certain elements of AI workloads particularly. This could be the case as AI models grow complex and diverse, and the need for processors performing efficiently across many tasks becomes necessary; then, hybrid architectures can top the game with designs that bring together the strengths of above-mentioned parts - GPUs, TPUs, IPUs.

Furthermore, the energy efficiency of AI processors has increased, particularly for large deployments in data centers. If the workload is big and complex, this would be an important factor in the design or adoption of the processor. Therefore, future AI/ML hardware should find a good balance between performance, flexibility, and energy efficiency of next-generation AI applications with new AI hardware [2][3].

IV. Result

The results of the report detail in-depth characteristic performance against architectural efficiency, against application advantages for state of the art in GPUs, TPUs, and IPUs based on benchmarks and analyses carried out. Each of the families of processors for the respective application has its strength and weaknesses in carrying out computations, and these are critically reviewed through many quantitative metrics for general-purpose tasks related to the processing speed, energy efficiency, scalability, and suitability for a vast variety of workloads related to AI/ML.

A. Processing Speed and Throughput

The processing speed is measured using FLOPS and is, in fact, one of the most essential metrics for the benchmarking performance associated with the computing processing units by GPUs, TPUs, and IPUs.

- Very good in performing highly parallelizable tasks; for example, DNNs training over big datasets. Basic support in terms of the massively parallel architectural design of GPUs is done through the hardware level for the inherent type of operations and calculations that computing involves, such as matrix multiplications in most AI/ML algorithms. Benchmarks of tasks involving CNN and RNN are quite nice in throughput if done on mature software ecosystems of CUDA and DNN from the vendor. However, it might be influenced by model complexity and massive data transfer between memory and processing units.
- TPU: When TensorFlow-based workload was dominating, TPU was leading in performance, especially in very large matrix operations. On the other hand, the systolic array

architecture allows the TPUs to be extremely efficient while working on large batches of data, which is so important with respect to the acceleration of large AI model trainings. In benchmarks, TPUs ran most large matrix-multiplying tasks far faster than GPUs, exhibiting higher throughput in such specialty tasks. Furthermore, TPUs benefited further by their capability to assist in reduced-precision operations such as bfloat16, which gave them a high processing speed without using that much memory bandwidth.

- IPUs: IPUs are relatively new entries into the AI space. IPUs were showing very strong performance across a subset of these workloads consisting of irregular calculations. Their architectures are fine-grained parallel and inherently low latency, so upstaged better than either GPUs or TPUs when dealing with dynamic, sparse data. IPs were much faster than either during benchmarks with models of irregular computation patterns such as GNNs and when doing real-time inference tasks where latency matters. This co-functionality underlines the possibility that IPUs can be called multi-functional and multifaceted AI accelerators [3][4].

B. Energy Efficiency

As AI/ML workload continues scaling in size and growing more complex, energy efficiency measured in FLOPS per watt increases in importance, particularly in large-scale data center areas where energy consumption is a concern.

- GPUs: They deliver high computational performances, yet their energy efficiency varies with the specific task performed. The more intensive computations by far, however, would include processes such as the training of the very large DNNs, and here they consumed more power than the TPUs and IPUs. This is partly because of their general-purpose architecture, making them flexible but also not very specialized in an energy consumption process, unlike the TPUs and IPUs. However, in cases where flexibility and wide applicability far outweigh the power concerns, and indeed may not owe anything to power, GPUs remain competitive.

- TPU: The most convincing energy efficiency examples were for operations well-tuned for the accelerator being used and large-scale matrix operations in a TensorFlow environment. The support of lower precisions, combined with the deployment of systolic arrays, explains the high throughput that TPUs guarantee at lower power. Thanks to this feature, TPUs turn out to be quite an attractive option for those environments where energy efficiency is important, such as in the case of large-scale cloud deployments and data center setups aimed at minimizing operational costs.

- IPU: IPUs also have highly concerted energy efficiency when performing tasks related to sparse and irregular data. This is together contributed by the fact that its architecture is totally aimed at minimizing latencies and maximizing effective processing power, which means less energy usage compared to that in a GPU. This is more the case when running tasks that do not fit so well in the more structured patterns in processing of GPUs and TPUs. This fine-grained parallelism can allow much lower-energy passage through such complex models, rendering them very well suited for edge devices and other power-efficient applications.

C. Scalability and Fit for AI/ML Workloads:

How well does each processor type scale to the next layer of the heavier computing work in the concretely increasing layer size, greater depth, and other increasingly complex inherent to larger and more sophisticated AI/ML models.

- GPUs: The parallel nature of their architecture and the software support are two characteristics which are most likely to make GPUs stand out in terms of the scalability of a very wide variety of AI/ML tasks. It can handle this rise in the complexity of the model and size of datasets by distributing the workload among thousands of cores.
- TPU: Optimized for TensorFlow-based models, and specifically scaling with large tensor ops. It is an architecture bred for massive-scale deployment in structures like Google's cloud infrastructure, where it could spread across thousands of TPUs hosting the largest AI models alive. Such architecture is actually valuable if and only if the task is carried out in an environment where TensorFlow is supposed to be a key performance driver; in other frameworks and for non-standard models, it might prove insufficiently productive [2][3].
- IPUs are just another way to deal with scalability. They work well with tasks that have uniformity and irregular computation patterns, and possibly dynamic data structures. Their efficiency in processing sparse data performs inherently well in these models being developed for AI/ML, which require graph-based structures or real-time processing tasks.

D. Summary of Comparative Performance

The above comparative analysis underscores the key features of any of the discussed processors - like the GPUs, TPUs, and IPUs - each one of which has a certain characteristic that makes it applicable to a variety of different types of aspects of AI/ML workloads.

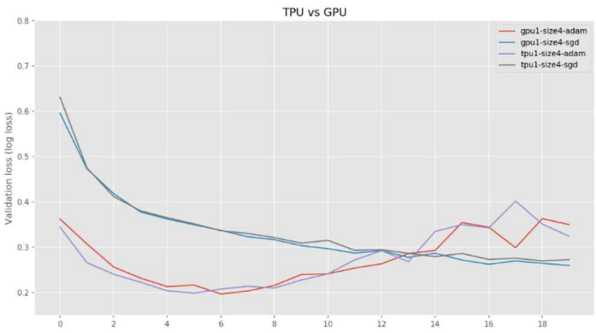

Figure 5: Metrics of Scalability and Suitability for GPUs, TPUs

- It is a general ability since the GPU is so versatile and does a wide range of different kinds of tasks, most importantly

the ones with a high level of parallelism and involving structured computations including software support and design among others, and thereby make them a safe investment option for a much wider variety of AI.

- The TPUs are highly specialized in TensorFlow environments to perform very well on huge matrix-based operations.
- IPUs are potentially useful in irregular, complex computations applied to future AI/ML workloads with dynamic data and processing in real-time.

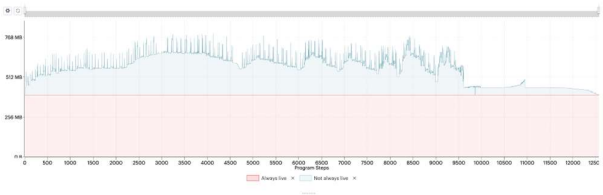

Figure 6: Pop Vision Graph Analyses of IPU

E. Implications for Hardware Selection

These evidently indicate that one would have to choose the most excellent processor to bear the AI/ML task in question. Of this there is little doubt; for any massive-scale neural network training task, certainly on a platform as TensorFlow, TPUs will beat the performance/energy efficiency; no doubt that the flexible and more generally applicable tasks should be done on GPUs. IPU, therefore, emerges as a formidable contender for tasks that include complex irregular computations and real-time processing. However, with the AI revolution in full steam, the best might yet be in the strategic deployment of these various processors such that the strong points of each architecture apply against the diverse demands of the modern AI/ML workloads [1][2][3][4].

V. Conclusion

In a nutshell, design architecture for each of the processors - GPUs, TPUs, and IPUs - has been made to meet certain demands that existed within the realm of artificial intelligence and machine learning. Given the growing complexity in AI and areas of applications, dependency on such specialized processors is only going to increase. It has cemented the place of a GPU as the workhorse of AI computing due to their versatility and wide scope of applicability. One major reason that makes them so fit for structured computation, especially tasks like deep neural networks, would be highly parallel architecture. High-level software development in their ecosystem helps in easy usability with most AI/ML frameworks.

TPU is another turn by Google toward special-purpose AI hardware. The architecture is optimized for operations with large matrices and TensorFlow, hence giving them outstanding performance in working conditions where these kinds of computations prevail. TPU does well in places where power and runtime performance are vital, more so when deployed on large-scale clouds. However, being optimized in their operations makes them very specialists in their work; hence they are not that suitable for parallel or completely different tasks.

IPUs have been rapidly becoming a force in dealing with increasingly complex and irregular workloads that come with the next generation of AI models. Fine-grained parallelism and efficient processing of sparse data set them particularly apart for dynamic data structures and real-time processing. IPUs are still rather young in the market but show very great promise for the advancement of AI/ML capabilities in areas where the traditional solutions, like GPUs and TPUs, are failing.

These results indicate that choosing the right processor is everything in terms of performance. In a nutshell, even though there is no silver bullet in AI hardware, comprehension of multiple genres of processors - GPUs, TPUs, and IPUs - with different strengths and weaknesses will go a long way in taking more educated and informed decisions while designing and fielding AI systems. Fast-paced innovation underway in processor design will help flesh out the future of AI and its applications when dust finally settles on often-evolving landscapes.

References

[1] Jouppi, N. P., Young, C., Patil, N., Patterson, D., Agrawal, G., Bajwa, R., ... & Yoon, D. H. (2017). In-datacenter performance analysis of a tensor processing unit. *Proceedings of the 44th Annual International Symposium on Computer Architecture (ISCA)*, 1-12. https://doi.org/10.1145/3079856.3080246

[2] Chen, T., Moreau, T., Jiang, Z., Shen, H., Yan, E., Wang, L., ... & Krishnamurthy, A. (2018). TVM: An automated end-to-end optimizing compiler for deep learning. *Proceedings of the 13th USENIX Symposium on Operating Systems Design and Implementation (OSDI 18)*, 578-594. https://www.usenix.org/system/files/osdi18-chen.pdf

[3] Graphcore. (2018). Poplar™: Graph programming framework. Graphcore Ltd. https://www.graphcore.ai/poplar

[4] Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., ... & Adam, H. (2017). MobileNets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*. https://arxiv.org/abs/1704.04861

[5] Wang, Y., Wei, G.-Y., & Brooks, D. (2019). Benchmarking TPU, GPU, and CPU platforms for deep learning. *arXiv*. https://arxiv.org/abs/1907.10701v4

[6] Reuther, A., Michaleas, P., Jones, M., Gadepally, V., Samsi, S., & Kepner, J. (2022). *AI and ML Accelerator Survey and Trends*. MIT Lincoln Laboratory Supercomputing Center. https://arxiv.org/abs/2210.04055

[7] Peng, H., Ding, C., Geng, T., Choudhury, S., Barker, K., & Li, A. (2024). *Evaluating emerging AI/ML accelerators: IPU, RDU, and NVIDIA/AMD GPUs*. arXiv. https://arxiv.org/abs/2311.04417