

# The OoO VLIW JIT Compiler for GPU Inference

Paras Jain, Xiangxi Mo, Ajay Jain<sup>‡</sup>, Alexey Tumanov, Joseph E. Gonzalez, Ion Stoica  
UC Berkeley, MIT<sup>‡</sup>

## ABSTRACT

Current trends in Machine Learning (ML) inference on hardware accelerated devices (e.g., GPUs, TPUs) point to alarmingly low utilization. As ML inference is increasingly time-bounded by tight latency SLOs, increasing data parallelism is not an option. The need for better efficiency motivates GPU multiplexing. Furthermore, existing GPU programming abstractions force programmers to micro-manage GPU resources in an early-binding, context-free fashion. We propose a VLIW-inspired Out-of-Order (OoO) Just-in-Time (JIT) compiler that *coalesces* and *reorders* execution kernels at runtime for throughput-optimal device utilization while satisfying latency SLOs. We quantify the inefficiencies of space-only and time-only multiplexing alternatives and demonstrate an achievable 7.7x opportunity gap through spatial coalescing.

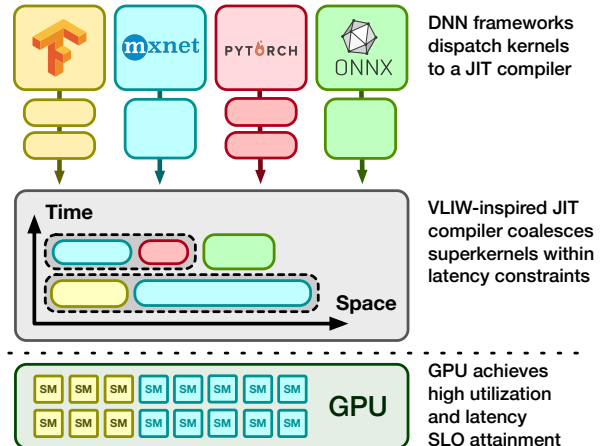
## 1 INTRODUCTION

As deep learning is deployed in applications ranging from video monitoring to language translation, there is an emerging need for parallel hardware accelerators to support inference. Deep learning inference needs to scale to billions of queries per day and is rapidly outpacing training in datacenters [21]. Amazon estimates that 90% of production ML infrastructure costs are for inference, not training [27].

While there are numerous specialized inference processors, the widespread availability of GPUs and their support for general deep learning models renders them indispensable for inference. By leveraging substantial parallelism, high memory bandwidth, and tensor acceleration, GPUs quickly evaluate deep neural networks.

Interactive inference workloads typically have tight latency objectives, especially when revenue-critical services issue inference queries. We note a concerning trend that inference latency on a CPU has been on a rise (Figure 2); the state-of-the-art SENet-184 [24] model takes 4.1s for a single CPU inference. As models grow larger over time, CPUs will continue to struggle to serve interactive model serving workloads. With real revenue costs associated with user-facing latency [19, 37], GPUs will remain a favorite for inference.

While some training workloads continue to scale and can often easily saturate modern GPUs, ML *inference* has distinctly different performance requirements that often



**Figure 1: A proposed OoO VLIW JIT compiler for on-GPU inference *coalesces* and *reorders* heterogeneous kernels from multiple streams of execution, effectively creating an efficient space-time schedule for on-GPU execution. Coalesced kernel will increase GPU compute and memory utilization. Interleaving multiple execution streams extracts OoO parallelism and reorders execution to fit in latency SLO budgets.**

result in poor GPU utilization, given the current GPU programming abstractions. In practice, online inference queries often cannot realize the high levels of parallelism that offline iterative minibatch training achieves, leading to poor GPU utilization. AWS reports p3 GPU instances are only 10-30% utilized [27]. Low resource-efficiency is not isolated to GPUs as Google’s Tensor Processing Unit reports a 28% mean utilization [28].

We propose an Out-of-Order (OoO) Just-in-Time (JIT) GPU kernel VLIW-inspired compiler for DNN inference (Figure 1). VLIW refers to computer architecture designed to extract instruction level parallelism (ILP) by packing multiple mutually-independent instructions that utilize different arithmetic processing units into a single large instruction word. VLIW design pushed the complexity of superscalar execution to the compiler while trying to keep the hardware simple. Analogously, we can improve accelerator utilization by reordering and packing kernels on-the-fly. We borrow inspiration from VLIW by coalescing execution kernels into superkernels that can more fully leverage a large pool of massively

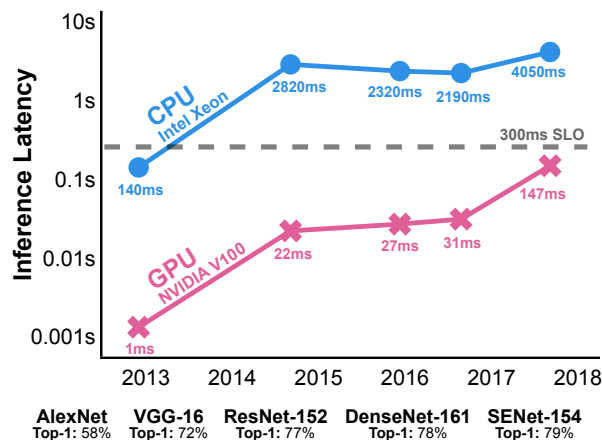


Figure 2: DNN model complexity and inference latency is increasing over time on CPUs and GPUs. Most models fail to meet the 300ms latency SLO on a CPU. To meet low latencies, GPUs are seeing widespread use for DNN inference.

parallel GPU compute units. Out-of-order execution between independent execution streams then increases efficiency while meeting latency deadlines.

Two reasons often attributed to the VLIW’s failure are: (a) overwhelming multitude of operations to coalesce, making it hard for the compiler to solve the packing problem; (b) difficulty with extracting enough instruction level parallelism such that instructions don’t have data dependencies between them. Despite its history, we believe the VLIW approach holds promise for GPU inference because (a) the set of operations to coalesce is restricted largely to algebraic tensor operations (e.g., matrix multiplies), (b) ability to leverage multiple inference streams, which, by construction, consist of mutually independent operations, and (c) the shape of execution kernels is adjustable, while VLIW compilers operated on immutable instructions.

We emphasize that the JIT compiler is *dynamic*. It operates on multiple streams of execution (analogous to instruction streams) in real-time, coalescing and reordering the operations in GPU space-time (Figure 1). To the best of our knowledge, this is the first such proposition for on-GPU DNN inference.

## 2 DNN INFERENCE REQUIREMENTS

Deep learning workloads like object detection, speech recognition, and natural language processing are being rapidly deployed in datacenters. These models are computationally intensive demanding 10s of GFLOPS of computation per query and a single model can easily see

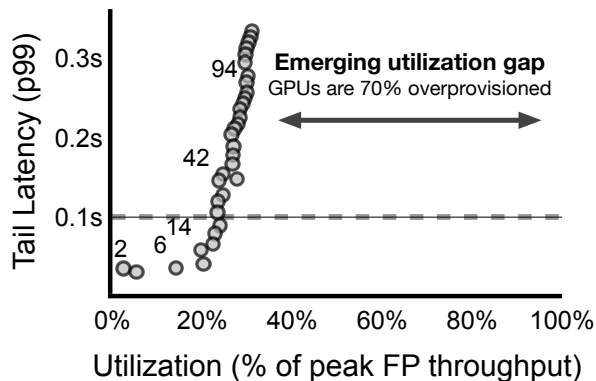


Figure 3: An emerging utilization gap: In order to meet latency SLOs, small batch sizes must be used, resulting in low GPU utilization. We profile ResNet50 across many batch-sizes (annotated above points) and see that single model inference will not fully utilize the GPU.

millions of inference queries per day. Facebook reports that its compute requirements for DNN inference have increased 3.5x in just 2 years [32].

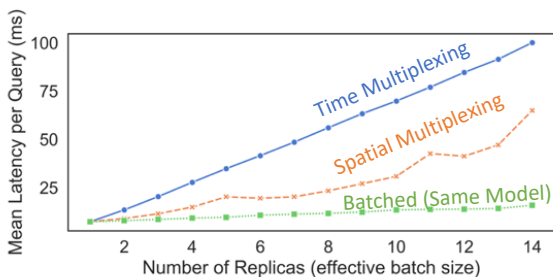
Typical applications see both batch inference queries and interactive latency-sensitive queries. Batch inference queries have no strict latency deadline and are primarily concerned with maximizing throughput. System designers aim to optimize the *cost-per-query* for batch inference. However, interactive inference queries have strict latency deadlines. With a real revenue cost for increased tail latencies in user-facing applications [37], system designers aim to control the *p99 latency* for online model serving. Typical latency SLOs can vary from 10ms for search ranking to several seconds for explicit content recognition [21, 28, 32]. Furthermore, to meet interactive latency requirements, these models must run on specialized hardware accelerators (typically GPUs).

Interactive inference queries present a challenging dilemma for system designers. In order to meet their strict deadlines, small batch-sizes must be used Figure 3. However, those small batch-sizes lead to poor resource-utilization and therefore high-costs under heavy load.

## 3 AN EMERGING UTILIZATION GAP

Deep learning has motivated the design of specialized hardware optimized for its predictable yet computationally expensive workloads. Recent accelerators are being deployed in the datacenter [6, 10, 11, 15, 16, 28, 39] and the edge [4, 5, 7, 8, 30].

However, throughput-optimized inference accelerators see low utilization when serving interactive DNN workloads. Throughput-oriented accelerators like GPUs



**Figure 4: Mean latency for 1 to 15 replicas of ResNet-50 on a V100 GPU. Time multiplexing is not resource-efficient as it is dramatically slower than batched inference. Spatial multiplexing has unpredictable performance still-degraded from batched inference.**

and the TPU require substantial available parallelism to achieve peak throughput.

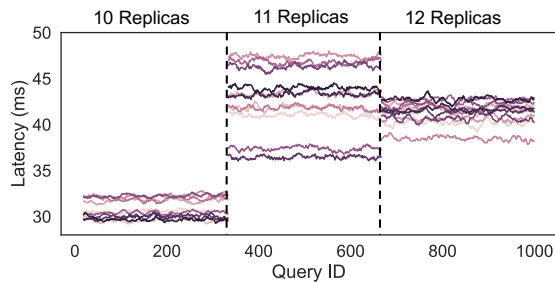
We observe an emerging *utilization gap* between hardware peak throughputs and actual observed performance. In Figure 3, we benchmark ResNet-50 on NVIDIA’s V100 GPU. At interactive latencies, throughput is less than 25% of peak. Larger batch sizes struggle to achieve 40% of NVIDIA’s advertised 15.7 TFLOPS throughput. DNN hardware accelerators suffer similar issues – the TPU v1 sees just 8.2% utilization for text processing workloads [28].

Why are throughput-oriented accelerators underutilized during DNN inference?

First, *tight latency objectives limit available parallelism* during inference. Figure 3 demonstrates that ResNet-50 inference on the NVIDIA V100 GPU achieves under 30% utilization. Kernels with small batch sizes have a low arithmetic intensity and thus poor peak floating-point throughput by the roofline model [41]. Individually, these kernels do not expose sufficient parallelism to saturate all the cores on modern GPUs.

Second, *resources must be provisioned for peak demand* rather than the average. As requests arrive stochastically, occasional bursts in request volume require over-provisioning accelerator resources. For an organization, peak load provisioning common in deep learning inference pipelines like those in [14] leads to excess purchases several times what average load would require.

Third, growth in *compute throughput outpaces memory bandwidth*. Memory bandwidth no longer scales with increasing parallelism and compute throughputs [12, 35, 42], and the ratio of compute throughput to memory bandwidth is rising rapidly. For GPUs, op to byte ratios have risen from 18 with the K80 to 139 for the V100. Specialized hardware fares even worse; Google’s



**Figure 5: Spatial multiplexing on GPUs has unpredictable latency when different number of processes are running concurrently. As we add replicas to a GPU running 10 multi-tenant models, some tenants encounter unpredictable SLO misses. A software-level JIT compiler will be able to recognize SLO misses and reprioritize kernels and pick the right execution streams.**

TPUv2 [16] has an op to byte ratio of 300. We estimate AWS Inferentia [20] has op to byte ratio of almost 500.

These factors will lead to continued underutilization of DNN accelerators. Ultimately, we believe intelligent software-level scheduling will enable more efficient hardware accelerators. The VLIW-inspired approach we propose is one of many potential optimizations that operate at a higher level than hardware.

## 4 INEFFECTIVE GPU MULTIPLEXING

From operating systems [18, 34] to cluster management [17, 23, 43], multiplexing is a well-established approach to improve resource utilization. Systems are usually either time-multiplexed (e.g. CPU core) or space-multiplexed (virtual memory). Current approaches to multiplex GPUs also fall under these two categories. In this section, we discuss why space-only and time-only multiplexing approaches fail to deliver both reliable performance with improved utilization.

### 4.1 Inefficient GPU Time Multiplexing

Each process that interacts with NVIDIA GPU owns a CUDA context. GPU can multiplex multiple CUDA contexts dynamically using an on-device scheduler. This approach enables interleaved (but not parallel) execution kernels. Kernels are serialized and processes are preempted periodically. Therefore temporal multiplexing doesn’t increase parallelism nor GPU utilization.

As shown in Figure 4, the inference latency increased linearly as we increase the number of concurrent processes. In addition, we noticed the context switching overhead is high because GPUs need to flush the execution pipeline.

## 4.2 Unpredictable Spatial Multiplexing

Modern GPUs from NVIDIA and AMD can be spatially multiplexed which enables concurrent overlapping kernel execution if resource permits. NVIDIA spatial multiplexing support is provided by Hyper-Q [1] while AMD’s MxGPU [3] utilizes an SR-IOV approach. CUDA Streams and NVIDIA Multi Process Service (MPS) [9] provided application support for spatial multiplexing. Model inference platforms like ModelBatch [31] and NVIDIA TensorRT [2] utilize CUDA Streams to achieve spatial multiplexing.

However, these approaches to spatial multiplexing result in poor performance isolation and unpredictable execution times. The spatial multiplexing approach is extremely sensitive to the choice of the number of tenants. When there are odd number of tenants on the same GPU, there is greater variability in latency among different processes (Figure 5).

Finally, because many kernels are tuned assuming they are single-tenant and own the entire GPU, the performance of concurrent execution of such kernels leads to lower throughput (Table 1).

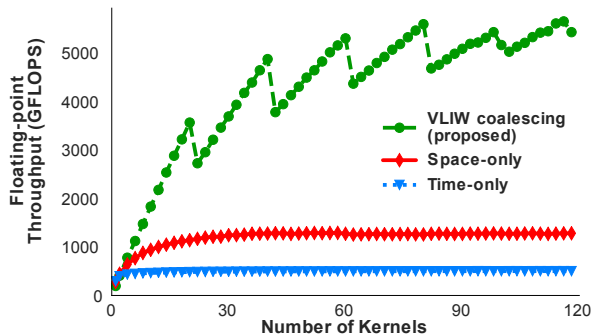
## 5 PROPOSED SOLUTION

We propose dynamic, just-in-time coalescing of execution kernels for GPU inference across multiple streams of execution and over time. This proposal draws inspiration from VLIW compilers by dynamically packing heterogeneous execution kernels to better utilize available hardware resources (spatial dimension). We also reorder and queue execution kernels in a latency SLO-aware manner to maximize the efficiency of packing. Thus, we advocate for a late-binding, context-aware approach to kernel scheduling on the GPU, in contrast to the early-binding, context-free abstractions currently exposed to the GPU programmer.

We show that the throughput-optimal convolutional block size depends on the size and shape of concurrent blocks of execution; ahead-of-time autotuning informs JIT decisions. Further, we demonstrate that inter-kernel optimization via coalescing yields substantial throughput gains. In the time dimension, reordering queued kernels of execution (a) prioritizes streams with tighter latency budgets, (b) purposefully delays/staggers ill-fitting kernels for better coalescing at a (slightly) later time.

### 5.1 Declarative Kernel Dispatch

The CUDA programming API is a form of *early-binding*. A programmer is required to specify the dimensionality and shape of the program without any contextual



**Figure 6: Coalesced kernels achieve ideal FP throughput. Coalescing the SGEMM that backs conv2\_2 from ResNet-18 with similar problems yields 3.23× throughput speedup over space-only multiplexing and 7.71× throughput over time-only multiplexing.**

information on the available GPU resources or other kernels of execution that may be in flight. This is inherent to the current low-level set of programming abstractions aimed at proactively controlling how much GPU compute and memory will be utilized. We refer to this approach as early-binding and context-free.

In contrast, we believe in *late-binding* and *context-aware* dynamic resource allocation, leveraging runtime information about the number of concurrent kernels of execution, and device context, such as the typical problem sizes served by a device. Given this information, individual kernels can be (a) retuned for better spatial multiplexing, (b) coalesced for better utilization, and (c) reordered for better spatiotemporal packing, while meeting the latency SLOs of individual streams of execution.

Therefore, instead of specifying how GPU should allocate threads across blocks, the programmer should interact with the GPU at a higher level, via a declarative API by specifying the operators, the inputs, and latency constraints. The JIT compiler will then execute them with contextual knowledge of the current GPU state and other streams of execution, thereby improving utilization while satisfying latency SLOs.

### 5.2 OoO Execution for SLO Attainment

We preserve predictability and isolation during virtualization by monitoring inference latencies per-kernel. This allows reallocating resources between tenants on-the-fly. Our approach dynamically adjusts to running workloads on the GPU unlike current static compilers like TensorRT [2] and others [13, 33, 36, 40].

Moreover, we notice that CUDA Stream scheduling anomalies typically only create a few stragglers, so we can simply evict degraded workers without significantly

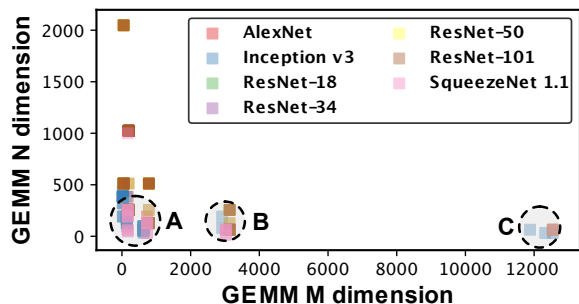


Figure 7: Matrix-multiply kernels from a wide class of models are concentrated into several clusters. Problems within certain clusters coalesce into efficient super-kernels (A, B and C).

impacting total system throughput. We are further investigating this approach in ongoing work.

A kernel that cannot yet be coalesced with those of other applications can be delayed via reordering. Thus, a dynamic JIT compiler overlaps waiting time from coalescing with computation from other execution streams.

### 5.3 VLIW Compilation for Efficiency

Conventional VLIW compilers only modify programs ahead-of-time. However, our dynamic approach uses both ahead-of-time tuning and runtime packing.

By retuning superkernels ahead-of-time in order to improve co-tenancy, we relieve pressure on the runtime JIT compiler. As GPU programs have many tunable parameters, we rely on auto-tuning. Preliminary results for our VLIW auto-tuner show *1.25x throughput gains* in a co-tenancy optimized kernel (Table 1).

At runtime, our VLIW JIT compiler repacks multiple small kernels into a large execution block. The compiler can apply pre-computed parameters from the auto-tuning phase to further optimize these larger kernel configurations. Scheduling a single superkernel on the GPU better utilizes compute and memory as compared to time-slicing. Unlike CISC execution streams, DNN kernels have extremely predictable performance and perform a small class of operations; comparably coarse-grained runtime packing decisions are efficient.

In Figure 7, we show that the matrix multiply kernels from multiple frequently used DNNs can be clustered by their dimensions. Within each cluster, problems can be coalesced with minimal padding overhead resulting in efficient co-execution. Coalescing a ResNet-18 intermediate convolution using the `cudaSgemvBatched` kernel achieves a geometric mean *7.71x throughput increase*

Auto-tuning configuration	Uniplexed throughput	Multiplexed throughput
Greedy kernel	2.2 TFLOPS	4.5 TFLOPS
Collaborative kernel	1.5 TFLOPS	<b>6.1 TFLOPS</b>

Table 1: Auto-tuning the blocking configuration on GPU leads to a different type of kernel. *Multi-tenant kernels* achieve 1.25x maximum throughput speedups when dispatched concurrently, despite small (20%) degradation when run in isolation.

over time-slicing, and *3.23x* over Hyper-Q spatial multiplexing (Fig. 6). Further, coalescing matrix-vector multiplications common in RNN/LSTM inference yields a *2.48x* throughput speedup over time-slicing [26]. VLIW compilation captures this large opportunity gap.

## 6 FUTURE RESEARCH DIRECTIONS

With the end of Moore’s Law and Dennard Scaling, performance gains in hardware will come from workload specialization. This implies increased heterogeneity in the types of specialized devices. We envision JIT compilation across multiple streams of execution to extend over multiple devices, such as ASICs, GPUs, TPUs, specialized inference processing units, and FPGAs. This will enable a more dynamic tradeoff of latency and throughput and improve hardware utilization in heterogenous settings. Furthermore, dynamic JIT compilation will also be able to explore different latency-accuracy tradeoffs.

## 7 CONCLUSIONS

The need to execute computationally intensive models within tight interactive latency deadlines has moved DNN inference workloads onto GPUs. However, the variability in kernel sizes, bursty arrival processes, and tight latency requirements often lead to poor GPU utilization. We propose a VLIW-inspired JIT compiler for GPU inference capable of coalescing, reordering, and retuning execution kernels across multiple streams to maximize throughput while meeting latency SLO constraints.

## 8 ACKNOWLEDGMENTS

We thank Hari Subbaraj and Rehan Sohail Durrani who helped profile kernels as well as Steven Hand, Koushik Sen, Eyal Sela, Zongheng Yang, Anjali Shankar and Daniel Crankshaw for their insightful feedback. In addition to NSF CISE Expeditions Award CCF-1730628, this research is supported by gifts from Alibaba, Amazon Web Services, Ant Financial, Arm, CapitalOne, Ericsson, Facebook, Google, Huawei, Intel, Microsoft, Scotiabank, Splunk and VMware.

## REFERENCES

- [1] 2012. NVIDIA Kepler GK110 whitepaper. <https://www.nvidia.com/content/PDF/kepler/NVIDIA-kepler-GK110-architecture-whitepaper.pdf>.
- [2] 2017. NVIDIA TensorRT. <https://developer.nvidia.com/tensorrt>.
- [3] 2018. AMD MxGPU: Hardware-based virtualization. <https://www.amd.com/en/graphics/workstation-virtualization-solutions>.
- [4] 2018. Edge TPU - Run Inference at the Edge | Edge TPU | Google Cloud. <https://cloud.google.com/edge-tpu/>
- [5] 2018. The future is here: iPhone X. <https://www.apple.com/newsroom/2017/09/the-future-is-here-iphone-x/>
- [6] 2018. groq. <https://groq.com/>
- [7] 2018. Jetson AGX Xavier Developer Kit. <https://developer.nvidia.com/embedded/buy/jetson-agx-xavier-devkit>
- [8] 2018. Kirin 980. <https://consumer.huawei.com/en/campaign/kirin980/>
- [9] 2018. NVIDIA Multi-process service. [https://docs.nvidia.com/deploy/pdf/CUDA\\_Multi\\_Process\\_Service\\_Overview.pdf](https://docs.nvidia.com/deploy/pdf/CUDA_Multi_Process_Service_Overview.pdf).
- [10] 2018. NVIDIA DGX-2: Enterprise AI Research System. <https://www.nvidia.com/en-us/data-center/dgx-2/>
- [11] 2018. NVIDIA T4 Tensor Core GPUs for Accelerating Inference. <https://www.nvidia.com/en-us/data-center/tesla-t4/>
- [12] Krste Asanovic, Rastislav Bodik, James Demmel, Tony Keaveny, Kurt Keutzer, John Kubiatowicz, Nelson Morgan, David Patterson, Koushik Sen, John Wawrzynek, et al. 2009. A view of the parallel computing landscape. *Commun. ACM* 52, 10 (2009), 56–67.
- [13] Tianqi Chen, Thierry Moreau, Ziheng Jiang, Haichen Shen, Eddie Q. Yan, Leyuan Wang, Yuwei Hu, Luis Ceze, Carlos Guestrin, and Arvind Krishnamurthy. 2018. TVM: End-to-End Optimization Stack for Deep Learning. *CoRR* abs/1802.04799 (2018). [arXiv:1802.04799](https://arxiv.org/abs/1802.04799) <http://arxiv.org/abs/1802.04799>
- [14] Daniel Crankshaw, Gur-Eyal Sela, Corey Zumar, Xiangxi Mo, Joseph E. Gonzalez, Ion Stoica, and Alexey Tumanov. 2018. InferLine: ML Inference Pipeline Composition Framework. *CoRR* abs/1812.01776 (2018). [arXiv:1812.01776](https://arxiv.org/abs/1812.01776) <http://arxiv.org/abs/1812.01776>
- [15] Bill Dally. 2017. Efficient Methods and Hardware for Deep Learning. In *Proceedings of the Workshop on Trends in Machine Learning (and Impact on Computer Architecture) (TiML '17)*. ACM, New York, NY, USA, Article 2. <https://doi.org/10.1145/3149166.3149168>
- [16] Jeff Dean. 2017. Recent Advances in Artificial Intelligence via Machine Learning and the Implications for Computer System Design. In *2017 IEEE Hot Chips 29 Symposium*.
- [17] Jeffrey Dean and Sanjay Ghemawat. 2008. MapReduce: simplified data processing on large clusters. *Commun. ACM* 51, 1 (2008), 107–113.
- [18] Kinshuk Govil, Dan Teodosiu, Yongqiang Huang, and Mendel Rosenblum. 1999. Cellular Disco: Resource management using virtual clusters on shared-memory multiprocessors. In *ACM SIGOPS Operating Systems Review*, Vol. 33. ACM, 154–169.
- [19] James Hamilton. 2009. The Cost of Latency. <https://perspectives.mvdirona.com/2009/10/the-cost-of-latency/>
- [20] James Hamilton. 2018. AWS Inferentia Machine Learning Processor. <https://perspectives.mvdirona.com/2018/11/aws-inferentia-machine-learning-processor/>
- [21] Kim Hazelwood, Sarah Bird, David Brooks, Soumith Chintala, Utku Diril, Dmytro Dzhulgakov, Mohamed Fawzy, Bill Jia, Yangqing Jia, Aditya Kalro, et al. 2018. Applied Machine Learning at Facebook: A Datacenter Infrastructure Perspective. In *High Performance Computer Architecture (HPCA), 2018 IEEE International Symposium on*. IEEE, 620–629.
- [22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.
- [23] Benjamin Hindman, Andy Konwinski, Matei Zaharia, Ali Ghodsi, Anthony D Joseph, Randy H Katz, Scott Shenker, and Ion Stoica. 2011. Mesos: A Platform for Fine-Grained Resource Sharing in the Data Center. In *NSDI*, Vol. 11. 22–22.
- [24] Jie Hu, Li Shen, and Gang Sun. 2018. Squeeze-and-Excitation Networks. *IEEE Conference on Computer Vision and Pattern Recognition*.
- [25] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q Weinberger. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.
- [26] Paras Jain, Xiangxi Mo, Ajay Jain, Harikaran Subbaraj, Rehan Sohail Durrani, Alexey Tumanov, Joseph Gonzalez, and Ion Stoica. 2018. Dynamic Space-Time Scheduling for GPU Inference. *arXiv preprint arXiv:1901.00041* (2018).
- [27] Andy Jassy. 2018. Amazon AWS ReInvent Keynote. <https://www.youtube.com/watch?v=ZOIkOnW640A>.
- [28] Norman P Jouppi, Cliff Young, Nishant Patil, David Patterson, Gaurav Agrawal, Raminder Bajwa, Sarah Bates, Suresh Bhatia, Nan Boden, Al Borchers, et al. 2017. In-datacenter performance analysis of a tensor processing unit. In *Computer Architecture (ISCA), 2017 ACM/IEEE 44th Annual International Symposium on*. IEEE, 1–12.
- [29] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*. 1097–1105.
- [30] Greg Kumparak. 2018. Tesla is building its own AI chips for self-driving cars. <https://techcrunch.com/2018/08/01/tesla-is-building-its-own-ai-chips-for-self-driving-cars/>
- [31] Deepak Narayanan, Keshav Santhanam, and Matei Zaharia. 2018. Accelerating Model Search with Model Batching. *SysML 2018* (2018).
- [32] Jongsoo Park, Maxim Naumov, Protonu Basu, Summer Deng, Arvind Kalaiah, Daya Shanker Khudia, James Law, Parth Malani, Andrey Malevich, Nadathur Satish, Juan Pino, Martin Schatz, Alexander Sidorov, Viswanath Sivakumar, Andrew Tulloch, Xiaodong Wang, Yiming Wu, Hector Yuen, Utku Diril, Dmytro Dzhulgakov, Kim M. Hazelwood, Bill Jia, Yangqing Jia, Lin Qiao, Vijay Rao, Nadav Rotem, Sungjoo Yoo, and Mikhail Smelyanskiy. 2018. Deep Learning Inference in Facebook Data Centers: Characterization, Performance Optimizations and Hardware Implications. *CoRR* abs/1811.09886 (2018). [arXiv:1811.09886](https://arxiv.org/abs/1811.09886) <http://arxiv.org/abs/1811.09886>
- [33] Jonathan Ragan-Kelley, Connelly Barnes, Andrew Adams, Sylvain Paris, Frédo Durand, and Saman Amarasinghe. 2013. Halide: A Language and Compiler for Optimizing Parallelism, Locality, and Recomputation in Image Processing Pipelines. *SIGPLAN Not.* 48, 6 (June 2013), 519–530. <https://doi.org/10.1145/2499370.2462176>
- [34] Dennis M Ritchie and Ken Thompson. 1978. The UNIX time-sharing system. *Bell System Technical Journal* 57, 6 (1978), 1905–1929.
- [35] Brian M Rogers, Anil Krishna, Gordon B Bell, Ken Vu, Xiaowei Jiang, and Yan Solihin. 2009. Scaling the bandwidth wall: challenges in and avenues for CMP scaling. In *ACM SIGARCH Computer Architecture News*, Vol. 37. ACM, 371–382.

- [36] Nadav Rotem, Jordan Fix, Saleem Abdulrasool, Summer Deng, Roman Dzhubarov, James Hegeman, Roman Levenstein, Bert Maher, Nadathur Sathish, Jakob Olesen, Jongsoo Park, Artem Rakhov, and Misha Smelyanskiy. 2018. Glow: Graph Lowering Compiler Techniques for Neural Networks. *CoRR* abs/1805.00907 (2018).
- [37] Eric Schurman and Jake Brutlag. [n. d.]. The User and Business Impact of Server Delays, Additional Bytes, and HTTP Chunking in Web Search.
- [38] K. Simonyan and A. Zisserman. 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. *CoRR* abs/1409.1556 (2014).
- [39] Ola Tørudbakken. 2018. Introducing the Graphcore Rackscale IPU-POD. <https://www.graphcore.ai/posts/introducing-the-graphcore-rackscale-ipu-pod>
- [40] Nicolas Vasilache, Oleksandr Zinenko, Theodoros Theodoridis, Priya Goyal, Zachary DeVito, William S. Moses, Sven Verdoolaege, Andrew Adams, and Albert Cohen. 2018. Tensor Comprehensions: Framework-Agnostic High-Performance Machine Learning Abstractions. *CoRR* abs/1802.04730 (2018). arXiv:1802.04730 <http://arxiv.org/abs/1802.04730>
- [41] Samuel Williams, Andrew Waterman, and David Patterson. 2009. Roofline: an insightful visual performance model for multicore architectures. *Commun. ACM* 52, 4 (2009), 65–76.
- [42] Wm A Wulf and Sally A McKee. 1995. Hitting the memory wall: implications of the obvious. *ACM SIGARCH computer architecture news* 23, 1 (1995), 20–24.
- [43] Matei Zaharia, Mosharaf Chowdhury, Tathagata Das, Ankur Dave, Justin Ma, Murphy McCauley, Michael J Franklin, Scott Shenker, and Ion Stoica. 2012. Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing. In *Proceedings of the 9th USENIX conference on Networked Systems Design and Implementation*. USENIX Association, 2–2.