Serving and Optimizing Machine Learning Workflows on Heterogeneous Infrastructures

Yongji Wu
Duke University
yongji.wu769@duke.edu

Matthew Lentz
Duke University
mlentz@cs.duke.edu

Danyang Zhuo
Duke University
danyang@cs.duke.edu

Yao Lu
Microsoft Research
luyao@microsoft.com

ABSTRACT

With the advent of ubiquitous deployment of smart devices and the Internet of Things, data sources for machine learning inference have increasingly moved to the edge of the network. Existing machine learning inference platforms typically assume a homogeneous infrastructure and do not take into account the more complex and tiered computing infrastructure that includes edge devices, local hubs, edge datacenters, and cloud datacenters. On the other hand, recent AutoML efforts have provided viable solutions for model compression, pruning and quantization for heterogeneous environments; for a machine learning model, now we may easily find or even generate a series of model variants with different tradeoffs between accuracy and efficiency.

We design and implement JellyBean, a system for serving and optimizing machine learning inference workflows on heterogeneous infrastructures. Given service-level objectives (e.g., throughput, accuracy), JellyBean picks the most cost-efficient models that meet the accuracy target and decides how to deploy them across different tiers of infrastructures. Evaluations show that JellyBean reduces the total serving cost of visual question answering by up to 58% and vehicle tracking from the NVIDIA AI City Challenge by up to 36%, compared with state-of-the-art model selection and worker assignment solutions. JellyBean also outperforms prior ML serving systems (e.g., Spark on the cloud) up to 5x in serving costs.

PVLDB Reference Format:
Yongji Wu, Matthew Lentz, Danyang Zhuo, and Yao Lu. Serving and Optimizing Machine Learning Workflows on Heterogeneous Infrastructures. PVLDB, 16(3): 406 - 419, 2022.
doi:10.14778/3570690.3570692

PVLDB Artifact Availability:
The source code, data, and/or other artifacts have been made available at https://github.com/libertyeagle/JellyBean.

1 INTRODUCTION

There is a growing complexity in machine learning (ML) inference workloads both in terms of the workloads themselves as well as the computing and networking infrastructures. These workloads often involve multiple ML operators that together form a larger *ML workflow*[1]; each can be a directed acyclic graph (DAG) of ML or relational operators. For each ML operator, there are often choices of models (e.g., YOLO [55], Faster R-CNN [56]) or the same model architectures with different hyperparameters (e.g., number of layers, neural network size, choice of activation functions); inputs to the ML workflows are often collected by sensors deployed at the edge, including video cameras and an ever-expanding array of Internet-of-Things (IoT) devices. These devices may have varying on-board compute [11] and are connected to more powerful edge-local and cloud computing services over the network.

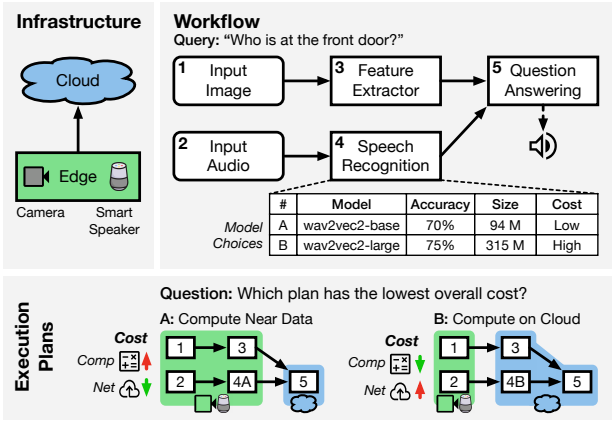


Figure 1: An example ML workflow of Visual Query Answering (VQA) on heterogeneous infrastructures. Execution plans vary by model selection and worker assignment for each operator (in the boxes) and result in different serving costs, i.e., compute and network.

Example. Consider the visual question answering (VQA) workflow in Figure 1 for the query "*Who is at the front door?*". The workflow uses multiple ML models for feature extraction and model inference. The infrastructure includes edge devices (e.g., cameras) as well as cloud datacenters. To deploy ML workflows on heterogeneous infrastructures, the following decisions must be made:

- *Model selection.* With advances of AutoML and model compression techniques (e.g., pruning, quantization [30, 59]), each ML operator in the workflow[1] can use various structures or hyperparameters; e.g., the speech recognition operator in Figure 1

[1]Workflows are generated using a standard parser [33] or a natural language interface [36], which are orthogonal to this paper. See §3 for more details.

Proceedings of the VLDB Endowment, Vol. 16, No. 3 ISSN 2150-8097.
doi:10.14778/3570690.3570692

may use the `base` variant for a faster execution or `large` for a better accuracy. To provide a viable accuracy-efficiency tradeoff, picking individual models in the workflow is non-trivial.
- *Worker assignment.* Each operator must be assigned to a worker for execution. Figure 1 demonstrates two execution plans - placing compute near the data source to reduce communication, or moving them to the cloud to take advantage of more powerful (and likely cheaper) compute resources. Choosing an appropriate plan depends on resource availability and costs.

Goals, challenges, and prior solutions. Given the ML workflow, resource availability, input throughput, and target accuracy, we aim to optimize the total serving costs that consist of both compute and networking. It is easy to see that model selection and worker assignment formulate a complex search space.

Current ML serving platforms such as Ray [47], Clipper [21], PyTorch [52], and Spark [65] focused on homogeneous infrastructures (namely cloud datacenter environments). Unfortunately, ignoring resource heterogeneity (e.g., compute, network) often leads to suboptimal deployments and even feasibility issues given the infrastructure constraints (e.g., on links shared among many high data rate sensors like video cameras). Some prior systems solve this problem in an ad-hoc manner for specific ML workflows, individual models, and fixed infrastructure configurations [12, 21, 32, 38, 52, 57, 60, 66]. Chameleon [32] considers video analytics with one model on a single GPU; Nexus [60] considers workflows on a homogeneous GPU cluster with no model choices. To our best knowledge, there is currently no off-the-shelf system that optimizes the deployments of ML workflows on heterogeneous infrastructures. As a result, users often manually determine how to best deploy ML workflows.

JellyBean **ideas and approaches**. We address some initial problems for optimizing ML workflows on heterogeneous infrastructures, and propose a system JellyBean. Given an ML workflow and specifications of the infrastructures, the JellyBean optimizer quickly finds a cost-efficient execution plan with model choices and worker assignments using the following insights:

First, we formulate the problem within a cost-based optimization [17], minimizing the compute and network costs while meeting the input throughput and accuracy constraints. However, optimizing ML workflows poses novel challenges. In the above example, even though we can profile the accuracy and cost for every single model, understanding how different models interact for estimating the overall query accuracy is non-trivial. We leverage a simple but effective model profiling strategy that relies on sampled measurements of interactions between models to estimate query accuracy.

Next, simultaneously solving for optimal model choices and worker assignment is NP-hard and results in an exponentially large search space. We reduce the search space and provide a fast query optimization by (1) making two simplifying assumptions that hold for many real-world scenarios, and (2) identifying key parts that are amenable to greedy approaches. Our evaluations in §6 show the efficacy in practice.

Lastly, to serve and optimize ML workflows on heterogeneous infrastructures, a flexible runtime is critical such that *the optimizer may explore plans in which models are placed in different workers and locations.* Due to the lack of an existing system to support this, we implemented the JellyBean processor upon Naiad [48] and Timely

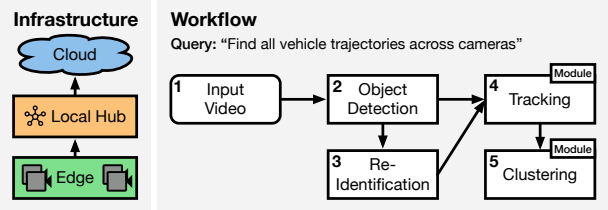


Figure 2: NVIDIA AI City Challenge for Vehicle Tracking. Some pairwise operators are omitted for simplicity.

Dataflow [1], modifying them to enable operator-level parallelism – each worker may handle a subset of the overall workflow. Such a processor and optimizer decide *where to run what*; for *how* to execute each individual operator, we use a containerized runtime with virtualization and ML compiler techniques [10, 18] such that JellyBean can cope with the infrastructure heterogeneity.

We performed experiments on various real-world use cases, including the Nvidia AI City Challenge [3] and Visual Question Answering (VQA) [13]. Compared with running the ML workflows (1) with all data pushing to the cloud, (2) with all computations staying on the edge, and (3) with optimizations carried out by several worker assignment heuristics, better assigning different parts of the workload to different infrastructure is significantly more effective. We also compared with a few recent ML serving platforms and found that JellyBean is significantly better to achieve the user-specified query-level goal. JellyBean achieves close to equivalent performance compared with an exhaustive brute force search on a small-scale experiment and can still generate efficient physical plans when brute force is infeasible on larger-scale experiments. JellyBean can reduce the total serving cost for VQA by up to 58.1%, and for vehicle tracking in AICity by up to 36.3% compared to the best baselines. JellyBean also outperforms prior ML serving systems (e.g., Spark on the cloud) up to 5x in total serving costs. We have open sourced our prototype: https://github.com/libertyeagle/JellyBean.

Contributions of this paper can be summarized as follow:

- The JellyBean optimizer to derive highly effective execution plans for complex ML workflows on heterogeneous infrastructures given the infrastructure constraints and model choices.
- A flexible JellyBean processor based on a graph dataflow to execute the optimized plans and enable operator-level parallelism on heterogeneous infrastructures.
- Evaluations on real datasets show significant performance improvements over state-of-the-art ML serving platforms as well as running the workflows using heuristics.

2 BACKGROUND

We discuss some popular ML workflows, followed by the challenges of running them across heterogeneous infrastructures.

ML Workflows There are many other use cases of ML queries for intelligent Internet of Things (IoT). In addition to the VQA query introduced above, we name a few interesting scenarios for instance:

- *NVIDIA AI City challenge:* Tracking vehicles across neighboring intersections is an important ML query that allows people to understand and improve transportation efficiency [3]. The

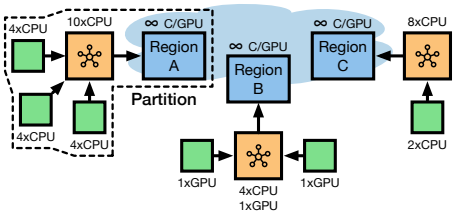


Figure 3: Deploying ML workflows on heterogeneous infrastructure requires designing physical plans for different partitions.

workflow is shown in Figure 2 with video inputs from multiple cameras of neighboring traffic intersections. It first detects objects on each individual video stream, and then performs an object re-identification (ReID) step to extract key features per detected car. A tracking module is used to find car traces in each video stream, followed by a clustering module to trace cars across different video streams.

- *Wearable health:* detecting anomalous heart signals.
- *Personal assistant:* answering complex human voice commands using Internet data.

One common characteristic is that they all rely on a set of loosely-coupled operators (i.e., operators that do not share global states but only depend on prior outputs), each of which uses an ML model or a traditional data processing module; e.g., a model to tokenize the text or relational operators such as reduce and join [16, 44]. The output of a previous operator is the input of the next, therefore formulating a workflow or logic plan in directed acyclic compute graph (DAG). Breaking down an ML query into workflows that consist of independent operators has been highly leveraged in prior research and production [43, 60]. Doing so promotes the reuse of trained models and operators to ease the development of the serving system as well as to boost performance due to shared computations [15, 37, 42, 63]; each module also can be improved independently to accelerate the application development.

Serving ML on Heterogeneous Infrastructures. The above examples also show that many application scenarios have input data injected from edge devices. To deploy ML workflows upon these inputs, one way is to put them in cloud datacenters. Clearly, this can often be suboptimal since raw inputs (e.g., images and videos) can be large and data movement can be costly.

Moving compute to near the data source is a well-known technique in the big-data systems literature and has been proven to be effective in many use cases [29, 54]. However, today developers still have to hard code or manually tune the physical execution plans for each ML workflow depending on the amount of resources on the edge and costs of various types of resources [32, 35, 46]. We believe this manual approach cannot scale with the rapid development of edge data centers and IoT devices.

Figure 3 shows a cloud with three regional datacenters, several local hubs, and edge compute devices. A different execution plan is needed for each partition. For example, different local hubs can have different numbers and types of workers. The cost of running models at different locations can also be different, depending on the cloud region and the resource availability at local hubs. We use the term *partition* to denote the tiered infrastructure where different locations within a tier have similar resources. If a partition contains multiple local hubs, they must have similar worker configurations. JellyBean can be used to generate a physical plan per partition.

Table 1: Comparing current ML systems. MS: model selection. WA: worker assignment.

System	Parallelism	QO		Usage	Heterogeneity	
		MS	WA		Worker	Infra.
PyTorch [52]	Data	×	×	Both	×	×
TF [12]	Data	×	×	Both	×	×
Spark [65]	Data	×	×	Infer	×	×
Clipper [21]	Data	×	×	Infer	×	×
Ray [47]	Data, Model	×	×	Both	×	×
Optasia [42]	Data, Op	×	✓	Infer	×	×
Pathways [14]	Data, Model	×	×	Train	✓	×
Llama [58]	Data, Op	×	✓	Infer	✓	×
Scrooge [27]	Data, Op	×	✓	Infer	✓	×
JellyBean (Ours)	Data, Op	✓	✓	Infer	✓	✓

ML Serving Systems. In order to partially move the ML workflow to the edge devices, besides being able to break it into modules or operators, another necessary condition is a serving system that supports operator-level parallelism on heterogeneous infrastructures. Prior ML systems focused on data, model (i.e., breaking large DNNs into operators) and operator (i.e., breaking workflows into operators) parallelism on homogeneous infrastructures [21, 42, 47, 52], or on heterogeneous workers within a datacenter [14, 27, 58]. We present a qualitative comparison in Table 1. Recently, Google's Pathways [14] has started to investigate operator-level parallelism for training large deep neural networks with hybrid cloud infrastructures of CPUs, GPUs, and TPUs. There still lacks an off-the-shelf system for serving and optimizing ML workflows with model choices on heterogeneous and especially IoT infrastructures. We provide a more detailed comparison with related systems in §8.

3 OVERVIEW

We discuss our JellyBean design and scope in this section.

System scope. JellyBean aims at serving and optimizing ML inference workloads that can be decomposed into multiple operators deployed on heterogeneous infrastructures. We target infrastructures that exhibit resource heterogeneity across tiers and resource homogeneity within a tier. JellyBean operates over an infrastructure configuration that describes a single partition of a potentially larger infrastructure. The optimization takes into account input throughput, resource cost, availability and efficiency, and targets scenarios in which compute and communication are important factors in the total serving cost. The JellyBean processor provides a flexible runtime and decouples resource heterogeneity using a containerized runtime with virtualization and ML compilers, hence targeting a wide spectrum of edge and cloud devices.

System overview. In Figure 4, we present an overview of our JellyBean system architecture and the workflow for processing an ML workflow. There are two main components: the query optimizer (QO) and the query processor (QP). The query optimizer generates an execution plan for the ML workflow, while the query processor runs the execution plan across heterogeneous infrastructure.

JellyBean takes the following inputs:

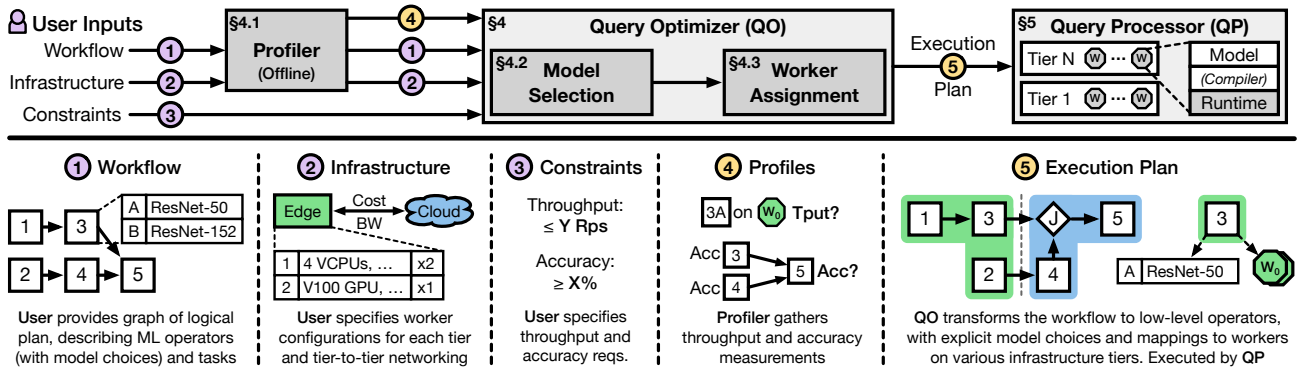


Figure 4: Overview of the JellyBean architecture. There are three main components: Profiler, Query Optimizer (QO), and Query Processor (QP).

- *Workflow.* Each input workflow is a directed acyclic graph (DAG) with compute operators on the nodes and input-output relationships between operators on the edges. The operators can be ML models or relational operations. Declarative queries can be parsed into workflows [33, 36] as is done in [34, 43].
- *Model choices for each ML operator.* Each ML operator may use different models with the same semantics but different structures or hyperparameters. These models have different accuracy and cost profiles. JellyBean may profile these models offline if necessary.
- *Infrastructure specifications.* We consider infrastructures that consist of heterogeneous resources (i.e., compute, storage and networking) in multiple tiers - each tier is a group of efficiently interconnected resources that share common specifications.
- *Input throughput and target accuracy.* Users provide a target accuracy on the query output; meanwhile, JellyBean must keep up with the input throughput. The target accuracy restricts the model selection to generate a low-cost physical plan.

Our query optimizer generates the physical plan in two steps. First, it selects models that satisfy the target accuracy with the least costs (§4.2). Here we do not have worker assignments yet, so the exact costs of deploying the selected models are unknown. We approximate the costs based on the characteristics of the models (e.g., model sizes, the latency of inference on a standard CPU/GPU) and use beam search to select the best K configurations. Each configuration includes the model selection for all models in the workflow.

The second step is to determine the worker assignment (§4.3). We again use a beam search method. We progressively determine the worker assignment by choosing a set of workers for each operator to achieve the lowest compute and networking costs. More than one worker may be assigned to an operator to consolidate the costs. The best worker assignment is derived then for each of the K configurations and choose the best physical execution plan for both model selections and worker assignment.

The JellyBean processor is a distributed query processing engine upon Naiad [48] and Timely dataflow [1] to provide a low-overhead dataflow abstraction. However, Naiad and Timely Dataflow use a homogeneous datacenter setup with data parallelism only JellyBean augmented their codebase to incorporate operator-level parallelism, allowing different workers to run different portions of the workflow. Each worker leverages a containerized runtime with virtualization or ML compilers [2, 18] to offset heterogeneity (§5).

Table 2: Set of common notations used in our description.

	Notation	Definition
User Input	G	Graph of logical plan ($G = \langle V, E, M, m \rangle$)
		Vertices V, Edges E, Models M
		Model Choices $m : V \rightarrow \mathcal{P}(M)$
	I	Set of infrastructure tiers
	W, W_i	Set of workers overall [or for tier $i \in I$]
	C_B	Worker-to-worker communication cost ($C_B : WxW \rightarrow \frac{\$}{\text{byte}}$)
	T, T_v	Input throughput overall [or for node $v \in V$]
	A	Target overall accuracy
Profiler	C_C	Unit compute cost for model on worker ($C_C : MxW \rightarrow \$$)
	t_u^w	Throughput for model u on worker w
	r	Unit input size at v from u ($r : VxV \rightarrow$ byte)
QO	s	Model selection ($s : V \rightarrow M$)
	a	Worker assignment ($a : V \rightarrow \mathcal{P}(W)$)

4 QUERY OPTIMIZER

4.1 Problem formulation

We consider our infrastructure to be composed of a number of workers with diverse computing capability distributed across multiple tiers (e.g., edge, hub, and cloud). Data sources are located on the lowest tier (i.e., W_1), often with some limited compute resources. Workers on higher tiers tend to have more computing capability but are far away from the data sources. We assume a set of workers W, which are partitioned into $|I|$ tiers.

Let the input of our optimizer be a logical plan graph G in which each node $v \in V$ corresponds to an ML or regular relational operator. For each ML operator, the user specifies a list of candidate models $m(v)$, each having a different accuracy and runtime performance. These models can be developed independently or can be variants of other well-known models through quantization [22, 30], distillation [24], and pruning [41, 59]. §7 discusses techniques to generate a diverse set of model choices. A model's accuracy and performance can be either provided by the user or profiled by JellyBean. We use $s(v)$ to denote the model choice for v. Meanwhile, we assign for each logical operator v a list of workers $a(v)$ in the heterogeneous infrastructure. The infrastructure specification contains sets of each type of worker a tier has, the cost of each type of worker, and the communication costs between different tiers. Note here our formulation only considers a single partition. This is because each partition (shown in Figure 3) requires a different physical execution plan. Table 2 illustrates the notations used in this paper as well as inputs to our query optimizer. Note that the

compute and communication costs here as unit monetary costs; the former is the hourly price per worker, and the latter is based on network traffic (i.e., data movement on the DAG edges).

We aim to solve worker assignment[2] $a : V \rightarrow \mathcal{P}(W)$ and model selection $s : V \rightarrow M$ simultaneously, such that the overall query accuracy (acc) is beyond a user-specified target A, and that the system's throughput ($t_{v_{out}}^{a(v_{out})}$ at the output node v_{out}) is no less than a target T. We describe our target cost function and our query optimization as:

$$\arg\min_{a,s} \sum_{v \in V} \sum_{w \in a(v)} C_c(s(v), w) \; + \sum_{(u,v) \in E} \sum_{\substack{(w_u, w_v) \in \\ a(u) \times a(v)}} C_B(w_u, w_v) R(u, v) \tag{1}$$

$$s.t. \;\; acc \geq A, \;\; t_{v_{out}}^{a(v_{out})} \geq T,$$

where $R(u, v)$ denotes the consumed network bandwidth from u to v. The formulation above minimizes the ML workflow's combined compute (first term) and networking (second term) costs and is NP-hard, because the sub-problem of solving only the worker assignment is already a combinatorial optimization that can be reduced to a binary knapsack problem (which is NP-complete [23]).

Assumptions. We make two assumptions in our optimization to reduce the problem complexity without losing generality, as these assumptions hold for many realistic use cases:

- A1: We assume that communication costs $C_B(w_1, w_2)$ to have the following properties: 1) set to 0 if w_1 and w_2 are on the same infrastructure tier and are in the same location, 2) otherwise set to a positive value. This is common in many use cases, as workers in the same tier either do not inter-communicate (e.g., among edge devices at different locations) or use high-speed networking (e.g., among datacenter nodes) with negligible costs.
- A2: We assume that all workers only communicate with peers in the same infrastructure tier or any higher tier, thus making information flow in one direction[3]. This assumption implies that for all edges $(u, v) \in E$, the set of workers $a(v)$ are all on tiers greater than or equal to the highest tier of any worker in $a(u)$. This is reflected in C_B by values of $+\infty$ for pairs of workers that violate this one-way flow assumption.

Model Profiling. JellyBean needs to understand the impact of selecting different models on accuracy and throughput in order to meet the constraints specified by the user for the overall workflow. While users can optionally specify the accuracy and performance of models for different infrastructure workers, JellyBean supports automatic profiling using validation datasets provided by the user. If a worker cannot run a particular model (e.g., model requires a GPU but the worker is CPU-only), we set both the accuracy and the throughput to be zero. Otherwise, JellyBean measures the runtime performance in terms of the throughput for the model on every worker type in the infrastructure. Note that we use the mean throughput of each model (and thus compute cost) relative to the input throughput during cost calculation, since operators in ML workflows may have different output-to-input ratios. For model accuracy, we need to understand the accuracy response of a model with respect to the accuracy of upstream models whose outputs are fed into it. JellyBean varies the input accuracy by selecting different upstream models (with different accuracy profiles) and measures the output accuracy response of the model under test. For example, consider a model with two inputs and exhibits the following accuracy profile: (60%, 50%) → 55%, (50%, 60%) → 60%, (70%, 90%) → 65%. This profile enables us to conservatively estimate the output accuracy by identifying the row that is closest to (but not higher than) the accuracy of all inputs; for example, if the input accuracy is (55%, 83%), then we can conclude the output accuracy is at least 60%. One assumption we make here is that the output accuracy is monotonically increasing with respect to each input accuracy (with the others fixed). In §4.2, we demonstrate how we use the accuracy profile to select models that satisfy the user's target end accuracy.

Next, we describe our solution that finds highly effective execution plans as well as components to derive query-level accuracy and, assign workers across the heterogeneous infrastructure.

4.2 Model Selection

Model selection balances the inference cost and model accuracy:

Satisfying Accuracy Constraints. One challenge in our model selection is to estimate the query-level accuracy given profiles of individual ML models, which can be non-trivial due to the dependencies among them. So far, this has not been discussed in any prior work, and we propose a solution here as follows.

We consider the dependency graph of the ML operators in the logical plan G. For each operator, we can assign (choose) a model variant; the final accuracy for the model selections s should satisfy a user specified accuracy threshold A. We use the model profiles to determine whether a model configuration satisfy the accuracy constraint. In each model's accuracy profile, we need to choose a row such that the output accuracy of a model is larger than a downstream node's required input accuracy. Also, the final output model's accuracy has to be above the target end-to-end accuracy.

Reducing the Total Cost. Another problem during model selection is that we do not know worker assignments yet and thus we cannot use a concrete cost. Thus, we need to choose models based on a different cost definition. We can use the execution latency on a single GPU or the number of parameters in the model. In our current prototype, we use a simple notion of cost: the latency for model inference on the most powerful infrastructure worker (e.g., NVIDIA V100 GPU in our evaluation).

We use the accuracy profiles and perform a *beam-search* to find the model assignments that can attain user's specified end-to-end accuracy threshold. We traverse the graph in reverse topological order, and assign the model for each node. Each candidate is a combination of partial model assignment and the accuracy requirements for upstream nodes. Specifically, we first extract the accuracy requirement for a node that we are currently assigning, and then

[2]We note that assigning for each operator a list of workers is equivalent to picking the model to execute for each worker.

[3]We note that the final result from the workflow may be transferred back to the lowest tier (e.g., user's device), but we do not model this.

iterate through all the candidate models for the node and find models whose output accuracy is greater than the threshold from the downstream models. When there are multiple models satisfying the output accuracy, we pick the ones that have the lowest cost. There can be many model configurations that satisfy the accuracy constraint, and we maintain the best B_{MS} number of model configurations based on their costs. After one model selection is found for this node, we then update the model assignment to propagate the accuracy constraints to upstream nodes until all nodes have a model assignment. We have to maintain more than a single candidate model configuration because our cost estimation can be not accurate. The real cost should be the actual cost of deploying this model on a particular worker type in the edge or cloud; here we simply use the latency or model size as the cost.

4.3 Worker Assignment

The goal here is to take the set of candidate model selections from the previous step and determine the best mapping from models to available infrastructure workers that minimizes the overall cost while meeting the input throughput to our system. We will first present an overview of our worker assignment algorithm, which makes greedy choices along two dimensions to reduce the large search space for worker assignment: 1) the order of assigning nodes $v \in V$ to workers, and 2) the workers $w \in W$ to be assigned. Next, we will describe our approach for determining the per-input cost of assigning the execution of a model to a given worker, which enables our greedy selection of workers. We also discuss key refinements that improve optimality in practice.

Computing Assignments. We present our solution in Algorithm 1. We consider as input a specific candidate model selection (out of the top-K candidates produced by the previous phase). The output consists of a mapping between nodes in the logical graph and sets of available workers.

In Line 2, we start by iterating over each node $v \in V$, using a topological ordering such that parent nodes are assigned before their downstream child nodes. While an optimal solution would need to consider the assignment of all nodes jointly, this is computationally intractable. However, due to the nature of realistic workflows and our assumption A2 that limits communication in one direction between tiers (i.e., from lower to higher), greedily computing worker assignments based on the topological ordering is a reasonable approximation. For any particular V and E, there may be many valid topological orderings; therefore, we extend our approach to also iterate over a constant number of different, randomly-selected topological orderings to improve the optimality.

For a given node v, we need to assign a set of workers to execute ML operator (or task), such that we limit the cost while meeting the input throughput. Each worker can be assigned to a node v[4], and such assignments formulate a combinatorial optimization which is NP-hard [23].

We use a greedy approximation for worker assignment by considering the cost of assigning a worker w to handle the execution of node v (with the assignment cost defined at the end of this section). We assign workers based on availability (i.e., not already assigned)

[4]We use one-to-one mapping due to the low overhead of our processor. See §5 and §7.

Algorithm 1: Worker assignment.

Input : Model selection $s : V \rightarrow M$
Output: Worker assignment $a : V \rightarrow \mathcal{P}(W)$
Function *Avail(W, a, i)* → *Returns unassigned workers in tier ≥ i*
Function *MinCost(W, s, v)* → *Returns worker with min cost (Eq 2)*
Function *TCoeff(w)* → *Returns throughput coefficient based on tier*
Function *Top(a, k)* → *Returns top-k best assignments in set*

1 $a_B = \{\varnothing\}$ // Current set of assignments in beam
2 **for** $v \in Topo(V)$ **do**
3 $a'_B = \{\}$ // Next set of assignments in beam
4 **for** $a_b \in a_B$ // Iterates over current set of assignments in beam **do**
5 **for** $i \in I$ **do**
6 $T_{\text{rem}} = T_v$, $a_{cur} = a_b$
  // Greedily assign workers up to throughput req.
7 **while** $T_{rem} > 0$ **and** $|Avail(W, a_{cur}, i)| > 0$ **do**
8 $a_{cur}[v] \cup= \text{MinCost}(\text{Avail}(W, a_{cur}, i), s, v)$
9 $T_{\text{rem}} -= (t_v^w \times \text{TCoeff}(w))$
10 **if** $T_{rem} \leq 0$ **then** $a'_B = a'_B \cup \{a_{cur}\}$
11 $a_B = \text{Top}(a'_B, B_{WA})$ // Keep only top assignments in beam
12 $a = \text{Top}(a_B, 1)$

and ordering from lowest to highest cost until the input throughput is met, or until we run out of workers to assign (Lines 7-9). Given our assumption of one-way communication between infrastructure tiers (A2), if a node u is greedily assigned to a worker on a higher-tier, then all nodes $v \in V$, where there exists an edge from u to v, are unable to be placed on lower tiers. We modify this by computing the greedy assignment over expanding pools of available workers, where the number of pools is equal to the number of tiers $|I|$ and the i$^{\text{th}}$ pool contains all workers in the i^{th} tier or lower. We use a beam search to reduce the search space by keeping the best B_{WA} candidate assignments (i.e., those with the lowest cost) out of the $B_{WA}|I|$ considered at each step (Line 11).

Since each tier may be distributed among one or more locations, we cannot simply consider the remaining throughput based on that achieved by a candidate worker w for node u (i.e., t_u^w). Instead, we need to multiply this by the TCoeff(w), which computes the factor based on the number of locations from the tier of w up to the root of the partition (e.g., cloud tier). Consider an example infrastructure that consists of the cloud, hub (2 locations), and edge (5 locations); TCoeff(.) is 1, 2, and 10 for workers on the cloud, hub, and edge (respectively).

Assignment Cost. To greedily pick workers with minimal unit (or per-input) cost, we need to take both computation and communication costs into account. Considering the cost for a node $v \in V$, with model selection s, running on a worker w, our overall cost equation is:

$$C_C(s(v), w) + \sum_{(u,v) \in E} \sum_{x \in a(u)} C_B(x, w) \left(\frac{t_u^x}{T_u} \right) r(u, v), \qquad (2)$$

containing the unit cost for computation (first term) and communication (second term). $s(v)$ is the selected model out of all choices for node v, and the unit computation cost is derived from the profiler

using the cost of each worker and the throughput of the worker while executing the selected model.

For the unit communication cost, we leverage all previous assigned nodes $u \in V$ that have edges to the current node v. Hence, the second term involves summing the costs across all workers assigned to u (i.e., $x \in a(u)$) and the worker w that is being considered. Note that we only consider the parents of v and not it's children, since our greedy algorithm operates in the topological ordering of the nodes, such that the assignments $a(u)$ for all child nodes u are already known. If w_u and w are on the same tier, the communication cost between the workers will be zero (A1); otherwise, there is some bandwidth-based cost for the traffic between the infrastructure tiers for x and w. This bandwidth cost is multiplied by the amount of communication for w_u, which is based on the unit input size $r(u, v)$ and the fraction of that input which is handled by x. The fraction of input is equivalent to the ratio of the throughput for u on x compared to the input throughput T_u. For instance, a node v takes inputs from u that is assigned to an edge worker x_1 (40% inputs) and a cloud worker x_2 (60% inputs). If we assign a worker at the cloud, the communication cost has to include the split linkages. The term t_u^x/T_u is the fraction of the $u \rightarrow v$ traffic contributed from x.

5 QUERY PROCESSOR

We prototype JellyBean upon Naiad [48] and Timely Dataflow [1] code base, which offered a low-overhead dataflow abstraction. However, there are additional features that JellyBean requires. We outline the challenges and our implementations in the following.

Operator-level parallelism. Timely Dataflow is designed for data parallelism. Instead, JellyBean aims for operator-level parallelism, spanning the workflow and compute nodes across different workers; hence they can execute different portions of the plan. The challenges here are two-fold: (1) all workers in Timely Dataflow must execute the same set of operators with different data inputs; (2) Timely Dataflow uses all-to-all communications for progress tracking, causing unnecessary overhead.

In the prior sections, we described our optimizer to assign workers to operators, where each worker is responsible for one operator in the graph. Indeed, executions of pipelines or workers that are assigned with multiple nodes are used in production database systems [50]. Our solution is simple but effective; as our experiments will show, we may put multiple workers on a single device, since the compute and network overhead of our processor is low.

Therefore, each worker only acquires its input data from upstream workers and sends its outputs to the downstream workers. We build a relay mechanism to serve as a "broker" between adjacent workers. There can be one or more relays in each worker; each receives input data from the relay nodes in the upstream workers. It also collects the outputs and sends them to the relay nodes in the downstream worker. To implement this, we use a thread for each upstream worker that keeps pulling data from the upstream worker's relays through TCP streams and maintaining proper buffers. There is also a thread for each downstream worker that pulls output data and sends it to the relays of the downstream workers. In such a manner, operator-level parallelism is achieved by properly parallelizing independent workers (which can be on the same device), tracking their progress, and syncing by treating each worker in our compute graph as a Naiad node. Lastly, we modified the progress tracking algorithm to support node-to-node progress updates.

Table 3: Some AICity models/operators used in our experiments.

Model	#Parameters (Millions)				
resnet	**18**	**34**	**50**	**101**	**152**
Object Re-identification	11.7	21.8	25.6	44.5	60.2
YOLO	**v5n**	**v5s**	**v5m**	**v5l**	**v5x**
Object Detection	1.9	7.2	21.2	46.5	86.7
wav2vec2	**base**	**large**			
Speech Recognition	94.4	315.5			

Networking protocols. Timely Dataflow supports communication among the worker nodes only by relaying on the master node; this results in unnecessary data movements. We augment the networking protocols to enable peer-to-peer communications among the workers; a low networking overhead is essential in a dataflow engine that supports operator-level parallelism.

Containerized worker runtime. Timely Dataflow supports homogeneous runtimes only. To offset runtime and hardware heterogeneity in JellyBean, each compute node deploys a containerized runtime with a Linux virtual machine to hold one or more Naiad workers. Table 3 illustrates part of the operators and models used in our experiments; each may contain a feature extraction or classification model. Within each container, JellyBean optionally applies ML compilers [2, 18] to adapt the model assigned by the QO to the worker hardware. By default, the ML models are implemented in PyTorch within the Naiad map functions.

Relational operators support. Timely Dataflow did not support relational operators including filters, join and group-by-aggregation upon columnar inputs. We hence implement these operators in JellyBean. The metadata is packaged with the data being transmitted in-between the workers to facilitate relational operations.

Remark. The runtime backend of our prototype consists of 12K lines of new Rust code beyond the Timely Dataflow v0.12. While our query optimizer is independent to the runtime engine, supporting broader runtime backends can be interesting future work.

6 EVALUATION

We evaluate JellyBean against state-of-the-art techniques for machine learning model serving with the following goals.

- G1 Is it beneficial to use JellyBean for serving ML inference workloads on heterogeneous infrastructures? We showcase end-to-end accuracy and cost measurements comparing with relative systems on two real-world use cases.
- G2 We measure the effectiveness and cost overhead of the JellyBean processor on various cloud and physical runtime.
- G3 To show that our optimizer is near optimal, we tease apart the usefulness of various aspects of the JellyBean optimizer in an ablation study and compare with alternative ML model selection and placement strategies as well as lower bounds.
- G4 We study the robustness and flexibility of JellyBean in a sensitivity analysis by varying the systems and workload settings.

Table 4: Four workload and infrastructure setups. We use $m \times n$ to denote that there exists m servers, each has n vCPUs. We show here the input throughput in frame/request per second (FPS/RPS); we use the mean per-frame/audio size from the input dataset in our cost model.

Dataset	VQA		AICity	
Setups	Objectives	Infras	Objectives	Infras
`small`: (5 nodes)	Accuracy: 0.55, Throughput: 9 rps.	Edge: 1x4, 1x8, 1x16, Cloud: 2xV100.	Accuracy: 0.65, Throughput: 3.5 fps.	Edge: 1x4, 1x8, Hub: 1x16, 1xV100, Cloud: 1xV100.
`medium`: (9 nodes)	Accuracy: 0.56, Throughput: 40 rps.	Edge: 1x2, 1x4, 2x8, 1x16, Cloud: 1x48, 3xV100.	Accuracy: 0.70, Throughput: 8 fps.	Edge: 1x2, 1x4, 1x8, Hub: 1x8, 1x16, 1xV100, Cloud: 1x48, 2xV100.
`large`: (15 nodes)	Accuracy: 0.56 Throughput: 60 rps	Edge: 2x2, 6x4, 1xV100, Cloud: 3x8, 3xV100.	Accuracy: 0.70 Throughput: 11 fps	Edge: 2x2, 2x4, Hub: 4x4, 2xV100, Cloud: 3x8, 2xV100.
`xlarge`: (30 nodes)	Accuracy: 0.57 Throughput: 100 rps	Edge: 6x2, 10x4, 2xV100 Cloud: 2x4, 6x8, 4xV100	Accuracy: 0.75 Throughput: 20 fps	Edge: 6x2, 3x4, 1xV100, Hub: 8x4, 2x8, 2xV100, Cloud: 1x4, 4x8, 3xV100.

6.1 Experiment Setup

Datasets. We consider two realistic machine learning workflows (and associated datasets) for model inference:

NVIDIA AI City Challenge (AICity) [3] is a public dataset and benchmark to evaluate tracking of vehicles across multiple cameras. The dataset is divided into 6 traffic intersection scenarios in a mid-sized US city, which in total contains 3.58 hours of videos collected from 46 cameras. A frame has 1.1MP (megapixels) and 22 objects (cars) on average. The ReID models are trained on the CityFlowV2-ReID dataset [61], while the object detection models are pre-trained on the COCO image dataset [39]. We leverage their testing scenario in our system evaluations. Figure 2 demonstrates a typical workflow upon this dataset with an object detection model, an object Re-identification (ReID) model and the subsequent tracking modules to derive cross-camera vehicle trajectories.

Visual Question Answering (VQA) [13] is another public dataset containing open-ended questions about images from the COCO image dataset [39]. The task is to generate an answer (from a large set of candidate responses) for an image-question pair. This dataset has 614,163 questions on 204,721 images. The mean input image resolution is 0.3MP and the mean input speech length is 1.5sec. The validation set from the original dataset split is used in our evaluation. Figure 1 demonstrates a typical workflow for VQA.

In our offline profiling, we measure the accuracy of 10 model combinations on the VQA validation set with 121,512 samples, taking 10-20 minutes depending on the model combinations. As for AI City, where test labels are not available, we use the official benchmarking API [3] to get the IDF1 scores. We profile 20 model combinations, and the profiling takes 1-2 hours depending on the model combinations. We also use reported accuracy on standard benchmarks whenever available [5, 6, 8]. We note that these are *one-time, per-database* costs and can be amortized among different ML workflows later on. We use P75 efficiency numbers as input to our optimizer to offset runtime variance; our sensitivity analysis in §6.4 discusses using other percentiles.

Workload and infrastructure settings. We conduct our experiments on the IBM cloud where the workload and infrastructure setups are detailed in Table 4. We evaluate four setups ranging from `small` to `xlarge` by varying the number and type of available workers for each infrastructure tier as well as the throughput and accuracy targets. Each compute node represents a virtual machine as described in §5 with the number of vCPUs specified (2-48), while each GPU compute node represents a VM with a 16GB NVIDIA V100 GPU. The memory of each node ranges from 4GB to 192GB and the bandwidth ranges from 3Gbps to 25Gbps. In §6.4, we show experiments when the bandwidth is limited. While the absolute infrastructure tier configurations may not capture all real-world infrastructure setups (e.g., IoT devices with compute <2 vCPUs), we note that the *relative* compute power difference between tiers does capture this. Using these settings strengths our evaluations as our processor offsets hardware heterogeneity by virtualization and ML compilers (§5).

We strive to echo real-world scenarios when setting up the base resource costs in our experiments; nevertheless, there can be orthogonal factors such as dynamic pricing models [26]. Hence, we use the unit compute and networking costs based on the pricing catalog of the IBM Cloud as of April 2022 [7]. The unit costs increases sub-linearly along with the resources used (e.g., 1 and 1.5 unit costs for 2 and 8 vCPUs respectively, and 3 for V100). The communication costs among different tiers (e.g., from edge to cloud) range from 0.1 to 0.3 unit cost per GB; for example, direct communication from edge to cloud bypassing local hubs is more expensive.

We also leverage prior VQA and AICity solutions on top of the benchmarks from [15, 40] and set up the accuracy and throughput targets used in our experiments based on the profiles of these state-of-the-art solutions. The virtual machines are chosen such that `small` and `medium` aim for low serving costs without edge GPUs, while the larger setups aim for low latency with edge GPUs available. The later cases also demonstrate how compute can be moved to the cloud when the edge has not enough compute power.

Evaluation metrics used in our experiment include:

Performance. We report both estimated and actually achieved throughput in one hour, as well as various overheads incurred by our query optimizer and processor. We also aim for a system that provides viable trade-offs between accuracy and throughput; we report the actual accuracy scores on the validation sets described earlier.

Serving costs. We report the compute and networking costs of executing the ML workload on the infrastructures specified in Table 4. We evaluate the costs while varying the target accuracy and input throughput. For JellyBean and all baselines (described next), we report the serving costs and other metrics when the system saturates, excluding model loading, system startup and shutdown time.

Baselines and comparisons. To compare JellyBean (`JB`) over state-of-the-art ML serving solutions on heterogeneous infrastructures, we consider the following baselines in our experiments:

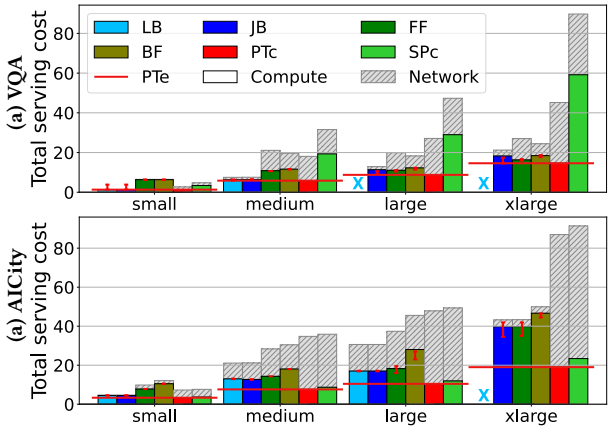
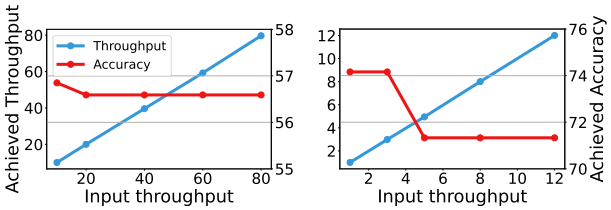


Figure 5: End-to-end evaluations of ML serving. We showcase the actual total serving costs using the P75 profiles; the error bars illustrate the estimated costs using the P50 and P90 profiles (see §6.4). 'X' indicates unsolvable given 1h of QO time.



(a) VQA (b) AICity

Figure 6: Achieved throughput and accuracy given different input throughput on the `medium` setup.

Worker assignment strategies. Inspired by geo-distributed database placement [29, 51, 54] and VM placement strategies [25], we compare with using the following model selection and worker assignment strategies while using the JellyBean processor[5]: (1) Best Fit (`BF`) is inspired by geo-distributed database optimizers [29, 54] to reduce the networking costs; it uses the most accurate model and greedily assigns jobs to the cheapest worker on the same infrastructure tier. (2) First Fit (`FF`) follows a classic VM placement strategy [51] in which each operator uses the most accurate model and assign jobs to the cheapest worker regardless of their location. (3) Lower bound (`LB`): we compute a lower bound of the serving cost by enumerating over all possible model choices and worker assignments when keeping the placement constraints (`A2`). This baseline showcases the optimality of our solution and it is worth noting that `BF` and `FF` may not follow the networking constraints used in `JB` and `LB`.

End-to-end ML serving. To our best knowledge, there lacks an off-the-shelf solution for serving ML on heterogeneous infrastructures while supporting the functionalities that JellyBean can provide. We use the following variants of existing systems to echo the real-world ML deployments. (1) We perform all computation on a single GPU worker using native PyTorch to handle the entire workflow. Doing so has the minimum compute overhead from the software stack beyond PyTorch but has to pay potentially large networking costs if the workers are on the cloud. By default, we use the most accurate models that are available and denote `PTe` as running PyTorch on the edge, *pretending* that there is a V100 GPU and counting the GPU costs; `PTc` runs PyTorch on a cloud V100 GPU, which is equivalent to `PTe` plus networking costs. (2) We assume the data is transferred to the cloud and use the most accurate models in a Spark. This baseline leverages all the cloud GPU workers in each infrastructure setup (Table 4) and performs data parallelism upon native PyTorch wrapped in a map function (`SPc`).

[5]We note that Worst Fit placement [51] that greedily puts models on the most expensive location does not fit in our context.

Table 5: Cost analysis on the AICity dataset. We show the costs for one hour of input data with input throughput specified in Table 4 and the corresponding query optimizing time (QO).

Model		`medium`			`large`		
Select.	Assign.	Comp.	Net	QO	Comp.	Net	QO
JB	JB	12.7	8.4	6.5ms	17.0	13.6	7.8ms
LB	LB	13.0	8.1	2.1**s**	17.0	13.6	27**min**
JB	FF	9.0	14.1	3.9ms	12.0	19.1	5.7ms
JB	BF	16.0	8.3	3.4ms	20.0	13.5	3.4ms

Model selection. The baselines above use the most accurate models available, since none of them solves the model selection problem. We will perform in §6.3 an ablation study to examine the effectiveness of our proposed model selection strategy, showing the optimality gap from using brute force.

6.2 System Evaluations

System efficiency. We showcase `G1` by the the end-to-end evaluations in Figure 5 and Table 5 using various workload and infrastructure settings in Table 4. We note a few observations here:

`JB` demonstrates the best performance with different datasets and setups compared to the baselines. On VQA, `JB` saves the total serving cost up to 58.1% compared to the best-performing baseline (`PTc`) and up to 5x compared to end-to-end ML systems `SPc`. On AICity, `JB` saves the total cost for up to 36.3% compared to the best-performing baseline (`PTc`) and up to 2.1x comparing to `SPc`.

We showcase the *actual* throughput and accuracy in Figure 6. `JB` achieved near 1:1 for actual:expected throughput (diagonal line). The results for the `large` setting is shown in Appendix [9]. With increasing input throughput but fixed available infrastructures, JellyBean successfully trades off throughput with accuracy by picking suitable models.

Comparing `JB` to `LB`, we observe a subtle difference in the overall serving costs – with different input throughput, 94.2% of the chances `JB` provides a total cost that has less than 1% difference to that provided by `LB` on AICity. `LB` requires a large QO time as will be shown next and becomes unusable – in AICity, `medium` has 8K choices while `large` has 7M choices.

Figure 7 illustrates a qualitative example of the execution plans of `JB` and `LB` when they do not match. JB uses 1x16 worker and a larger ResNet model for feature extraction, while LB uses 1x2 and 1x8 which leads to a lower cost. `BF` and `FF` failed to find overall optimal execution plans in our experiments; though in some cases, they find plans with low compute or low network costs solely (e.g., BF with low network cost while FF with low compute cost

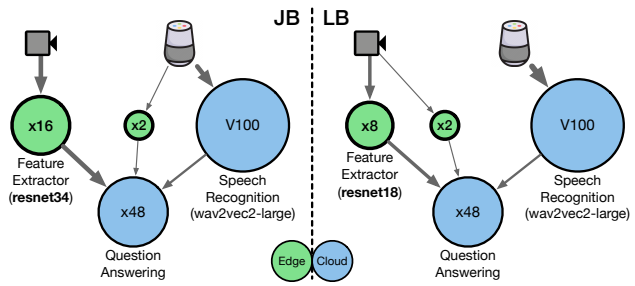


Figure 7: Comparison of the execution plans of JB and LB on VQA using the `medium` setup modified to 20 rps.

Table 6: Costs of operators upon the `medium` setup. E: Edge. H: Hub. C: Cloud. Original: the latency with native PyTorch. QP exec: the overhead of executing the operator in the JellyBean query processor; QP network: the overhead of communication.

VQA Operator	Node	Orignial (ms)		QP exec.		QP network	
		P50	P90	P50	P90	P50	P90
ImgFeat	x2E	152.2	161.9	+2.5%	+4.1%	+1.2%	+1.3%
ImgFeat	x4E	76.6	82.3	+7.4%	+8.1%	+3.9%	+3.6%
ImgFeat	x16E	27.4	29.4	+11.7%	+19.4%	+6.1%	+6.0%
ASR	x8E	251.0	297.6	+2.3%	+1.0%	+0.4%	+0.4%
ASR	V100C	24.1	26.3	+0.8%	+0.8%	+0.6%	+0.5%
VQA	x48C	6.3	8.6	+9.7%	+12.6%	+4.8%	+6.9%

AICity Operator	Node	Orignial (ms)		QP exec.		QP network	
		P50	P90	P50	P90	P50	P90
ObjDet	x2E	1412	1455	+3.4%	+8.2%	+0.3%	+0.4%
ObjDet	x4E	721.0	732.0	+6.4%	+8.7%	+0.4%	+0.7%
ObjDet	x8E	451.7	468.3	+3.5%	+6.1%	+0.8%	+0.8%
ObjDet	V100H	19.7	20.6	+13.7%	+13.3%	+8.2%	+10%
ReID	V100C	8.4	8.7	+6.5%	+7.5%	+6.0%	+6.2%
ReID	V100C	8.5	8.8	+6.1%	+7.6%	+20%	+20%

in Table 5). This can be as expected since their heuristics ignore model accuracy-efficiency trade-offs and the resource availability on heterogeneous infrastructures. In most cases, BF and FF have much higher costs than JB; using heuristics that consider network or compute cost solely is suboptimal. On the other hand, model selection greatly helps to reduce the overall costs, especially when the accuracy target is lower. PTc and SPc use homogeneous GPU computing, which results in lower compute costs than JB yet larger networking costs since the raw data must be transferred from the edge. SPc exhibits more overhead as compared with PTc [45]. PTe is a hypothetical baseline that assumes strong GPUs on the edge, and thus leads to minimum compute costs at zero networking cost. In real-world applications, with no resource constraint, users should adopt this solution; however, this is often not true in practice.

Further evaluations in Appendix [9] show that JellyBean often yields serving costs equal to or close to the lower bound. We also discuss some failure cases in the Appendix. For instance, in the case that Assumption A2 is removed.

We observe that the runtime variance is low across all setups; for example, the standard deviation from five runs on the `large` setup is 0.003% for AICity and 0.020% for VQA. The runtime variance on the `xlarge` setup is reported .

System overhead. Table 5 also illustrates G3 – the JB optimizer has a small overhead with the QO time of JB in a few milliseconds. In comparison, LB uses brute force, which incurs adverse QO time in larger infrastructure settings (e.g., 27 minutes for `large`). Other placement strategies have smaller QO time due to a smaller search space, but the total serving costs are larger.

We further demonstrate in Table 6 the compute overhead of the JB processor. We show the 50th and 90th percentile of various ML operators in native PyTorch and by the JB processor. The overhead caused by JB processor, as partially been discussed in [48], contains that for metadata parsing, data (un)packing, network I/O, and task scheduling. The QO latency is reported on 1x8 virtual CPU node with a Python implementation. Results indicate a small overhead ranging from a few to 19% upon the native PyTorch executions. This is significantly smaller that that of Spark which may take up to 300% (as shown in Figure 5).

Remark. Our evaluations across various workload and infrastructure setups showed that JellyBean efficiently computes and deploys execution plans and significantly reduces the total serving cost of real ML workloads. We believe it is beneficial to leverage JellyBean for serving ML on heterogeneous infrastructures across a wide range of real-world applications.

6.3 Ablation Study

We leverage the `medium` setting and evaluate JellyBean by sweeping different knobs used during query optimization. We also demonstrate similar experiments on other setups in Appendix [9].

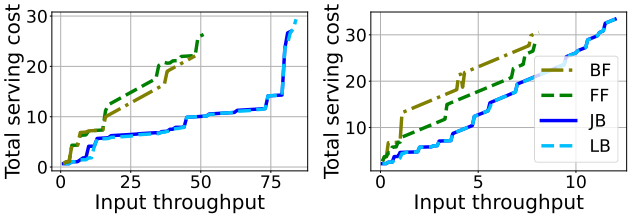
Input throughput. To demonstrate the scalability of JellyBean and to supplement Figure 5, we leverage a fixed target accuracy as in `medium` and demonstrate how the costs change when varying the input throughput. Figure 8 shows the results. We observe that JellyBean can keep up with increasing input throughput and is near optimal – in most situations, JB achieves the same total serving costs as LB. For BF and FF, no valid execution plans can be found beyond 51 rps and 8 fps (VQA and AICity, respectively).

Target accuracy. To show that JellyBean provides viable accuracy-cost trade-offs, we fix the target throughput as in `medium` and demonstrate the total serving costs by varying the target accuracy. Figure 9 shows the results. BF and FF solve only for placement while using the most accurate models, and thus the costs are constant. For the scenarios we examined, JellyBean is near optimal across a range of accuracy targets. JB and LB eventually use the most accurate models, converging with FF for AICity.

Effect of model selection. To examine the model selection strategy used in JellyBean (§4.2), Table 7 illustrates an ablation study in which we substitute our model selection for either the most accurate models or a brute force selection. We also evaluate our model selection strategy for PTc and SPc. Results show that our proposed model selection is effective with our JellyBean processor as well as other ML runtimes.

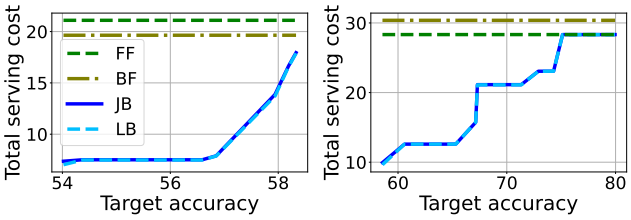
6.4 Sensitivity Analysis

We further study the robustness and flexibility of JellyBean (G4) with the following sensitivity analysis experiments.



(a) VQA (b) AICity

Figure 8: Total serving cost w.r.t. input throughput in JB.



(a) VQA (b) AICity

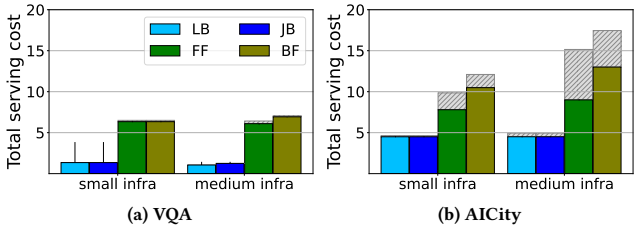
Figure 9: Total serving cost w.r.t. target accuracy in JB.

Table 7: Ablation analysis of model selection on the AICity dataset.

Model		medium			large		
Select.	Assign.	Comp	Net	QO	Comp	Net	QO
JB	JB	12.7	8.4	6.5ms	17.0	13.6	7.8ms
Most acc.	JB	14.3	14.0	1.2ms	17.5	19.1	1.3ms
Brute f.	JB	12.7	8.4	11.6ms	17.0	13.6	15.0ms
JB	PTc	5.3	27.2	N/A	7.3	37.4	N/A
JB	SPc	5.8	27.2	N/A	8.0	37.4	N/A
Most acc.	PTc	7.6	27.2	N/A	10.4	37.4	N/A
Most acc.	SPc	8.7	27.2	N/A	12.0	37.4	N/A

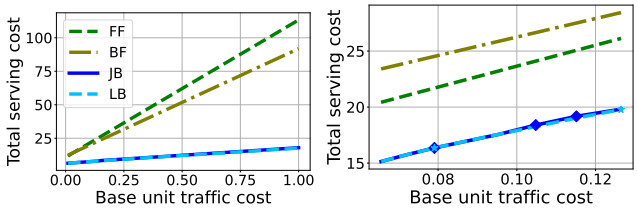
Effect of resource over-subscriptions. When there are more resources than needed, especially on the cloud, can JellyBean handle the workloads without wasting resources? Also, how do the costs change? We answer these questions by deploying the `small` workload on the `medium` infrastructure (Table 4). Figure 10 illustrates the results. We observe that, compared with using the `small` infrastructure, more resource availability will not significantly increase the serving cost for JellyBean with a fixed workload. However, BF and FF cannot guarantee cost efficiency in such a scenario. This is largely due to their sub-optimal worker assignment strategies which disregard resource availability. With JellyBean, users may use large cloud subscriptions without wasting resources.

Base unit network costs. We examine the effect of varying the network costs in a `medium` setup, which play a critical role in the total serving costs. Figure 11 showcases a change in cost from 0 to 1 (per GB). Interestingly, for VQA, we found that the unit network costs actually have minor effects on the execution plans and the plan changes are subtle – this is due to a relatively higher compute cost on the cloud, so the computation is kept at the edge. Meanwhile, on AICity, we use blue dots to show where the plan changes, though the total serving cost is near linear. We present actual query plans in Figure 12 to show an example plan change when the network cost is reduced by 90% and compute is shifted to higher-tier workers.



(a) VQA (b) AICity

Figure 10: Applying the `small` workload setup on the `medium` infrastructures. JellyBean uses the minimum available resource to achieve an optimal performance.



(a) VQA (b) AICity

Figure 11: The overall serving costs w.r.t. base unit traffic cost. Dots indicate when the execution plan changes. On AICity, different lines are near linear after 0.12 and hence we show cropped results.

Effect of infrastructure changes. We examine the flexibility of JellyBean when there are additional resource caps on the `medium` setup. Specifically, we placed a 10Mbps bandwidth constraint over all edge devices, mimicking real-world scenarios with limited networking. The JellyBean optimizer simply applies an additional constraint and limits the search space; there is no change on the processor and execution engine. Figure 13a illustrates the results and our findings. To reduce network costs, the JellyBean optimizer uses more compute resources on the edge. The blue curve ends early since no viable solution can be found.

We change the number of workers allocated to different tiers, and observe how the total serving cost changes (Figure 13b). Since there are 9 total workers in the `medium` setting, we rank them according their cost and place those with higher costs on higher tiers (and vice versa). For instance, in the 5:4 case, the edge has 1x2, 1x4, 2x8 workers, and the cloud has 1x16, 1x48, 3xV100. Results show that JellyBean successfully finds good execution plans in all the settings; more cloud resources does increase the network costs.

Effect of profiling. In our previous experiments in §6.3, we leveraged the profiling of P75 percentiles as inputs to our QO and report actual runtime numbers; doing so gives extra room for the QO to find valid plans. In Figure 5, we also explore the estimated costs using P50 and P90 efficiency profiles on the error bars. We observe a small variance – the the actual runtime using P75 in most cases falls in the middle of estimates using P50 and P90. In some cases, the optimizer chooses different plans which leads to discontinuity of the costs. Overall, we observe that this has minor effects on our end-to-end solution.

7 DISCUSSION

Obtaining diverse model choices. The user optionally provides a list of model choices for each operator in the workflow. Our current

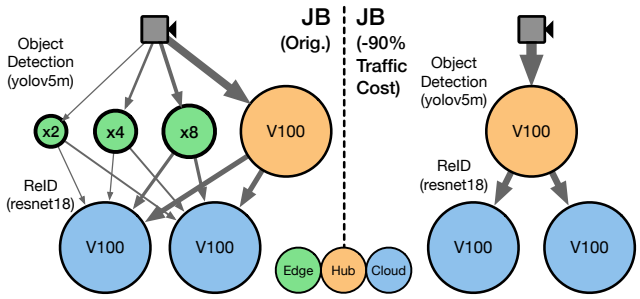
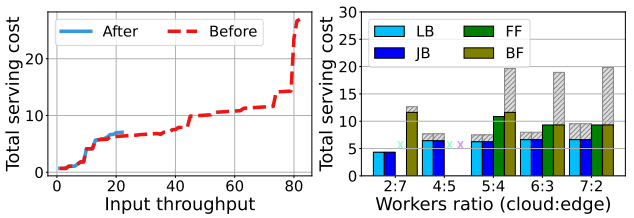


Figure 12: Change of worker assignment when unit traffic cost is 10% of the original traffic cost on `medium` setup for AICity.



(a) (b)

Figure 13: Sensitivity study on VQA for (a) limiting total outbound bandwidth at 10Mbps on edge devices with the `medium` setup, changing input throughput and (b) changing the worker ratios on different tiers; see text for details. 'X' indicates unsolvable inputs.

prototype depends on this provided model choices. However, in the future JellyBean can also enrich the choices using off-the-shelf model quantization, pruning, and distillation tools. Several tools already exist today, and it is an active area of research in ML [19, 22, 30, 41, 53, 59] in order to democratize ML on weak edge devices. To integrate these tools into JellyBean, we can simply invoke them to derive cheaper models offline (similar to how we profile models for their accuracy profiles). We acknowledge that running these tools may require us to access the original training data and labels.

Limitations. As discussed in §4 and §5, we used a one-to-one mapping between the workers and operators. Using a one-to-multiple mapping to consolidate the operators may further improve the performance and can be an interesting further work to explore. Doing so may require automatic grouping of the operators. Nevertheless, we have shown in §6.3 that our processor already has low overhead.

JellyBean also assumes the heterogeneous infrastructures to have near constant input requests on the edge devices; this is true for the use cases discussed in §2 and in our experiments. Exploring use cases that do not fall into this category, such as security sensors or cluster telemetries which send only intermittent signals, can be an interesting future work. Besides, we used one-time profiling and fixed worker costs in our experiments; quickly adapting to changes in these aspects can also improve the usability of our system.

8 RELATED WORK

Edge-cloud systems. Moving compute to the edge can reduce the networking cost and is used in video analytics to eliminate the need to transfer raw video streams. Chameleon [32] leverages temporal and spatial correlations to tune frame resolution, sampling rate, detector model configurations for an optimal resource-accuracy trade-off. In [62], a latency and energy consumption model is considered for choosing the configuration. Jain et al. [31] scale video analytics to large camera deployments using hand-crafted rules that leverage cross-camera correlation to improve cost efficiency and accuracy. Elf [67] applies a content-aware approach to offload smaller inference tasks in parallel to edge servers. These works considered a simple edge-cloud infrastructure and used workload-specific optimization techniques. We support optimizing and running arbitrary ML workflows on a wide range of infrastructures, both of which are inputs to our optimizer.

ML inference systems. Serving machine learning inference has attracted great attention. TensorFlow Serving [49] is one of the first serving systems for production environments. Clipper [21] maximizes throughput under a user-specified latency service-level objective (SLO), model selection policies are also integrated to provide different cost-accuracy trade-offs. Nexus [60] automatically chooses the optimal batch size and the number of GPUs to use according to the request rate and latency SLO. Model DAGs are also considered in other works [4, 20, 27, 28, 57, 58]. JellyBean differs in two ways. First, we choose individual models based on input throughput and target accuracy for the entire ML workflow. Second, we target at deploying ML workflows on heterogeneous infrastructures, where prior works focused on either: a) homogeneous cloud datacenters or edge devices only, or b) heterogeneity within a single tier (i.e., datacenter).

Optimizing ML queries A number of works have been proposed in optimizing ML queries at either logical- or physical-level. Lu et al. [43] filter data that does not satisfy the query predicate by using probabilistic predicates. BlazeIt [34] optimizes aggregation and limit queries for videos. Yang et al. [64] exploit predicate correlations to build proxy models online to avoid exhaustive offline filter construction. Optimization at physical execution-level is addressed in some of the ML serving systems that support model DAGs. For instance, Llama [58] applies a greedy strategy that chooses cost-efficient worker configurations for video analytics pipelines. These works did not consider network cost, because these systems target pure datacenter deployment scenarios. JellyBean optimizes general ML workflows jointly at logical and physical levels for a heterogeneous infrastructure across edges and the cloud.

9 CONCLUSIONS

The rise of smart home devices and the Internet of Things opens up the opportunity for ML serving systems at the level of both the infrastructure and ML workflow to explore new trade-offs between accuracy and performance. We build JellyBean, an ML serving to optimize ML workflows which takes into account the cost, availability, and performance of the increasingly tiered and heterogeneous infrastructures. JellyBean significantly reduces the total serving cost of visual question answering and vehicle tracking from the NVIDIA AI City Challenge compared with state-of-the-art solutions.

ACKNOWLEDGEMENT

We thank VLDB reviewers for their insightful feedback. Our work is partially supported by gifts from Adobe, Amazon, Meta, and IBM.

REFERENCES

[1] 2015. Timely Dataflow. https://github.com/TimelyDataflow/timely-dataflow. [Last accessed:: 11/17/2022].

[2] 2019. NVIDIA TensorRT. https://developer.nvidia.com/tensorrt. [Last accessed:: 11/17/2022].

[3] 2021. AI City Challenge. https://www.aicitychallenge.org/2021-ai-city/. [Last accessed:: 11/17/2022].

[4] 2021. Triton Inference Server. https://github.com/triton-inference-server/server. [Last accessed:: 11/17/2022].

[5] 2021. YOLOv5 releases. https://github.com/ultralytics/yolov5/releases/tag/v6.0. [Last accessed:: 11/17/2022].

[6] 2022. HuggingFace pre-trained models. https://huggingface.co/models. [Last accessed:: 11/17/2022].

[7] 2022. IBM Cloud Pricing. https://www.ibm.com/cloud/vpc/pricing. [Last accessed:: 11/17/2022].

[8] 2022. PyTorch pre-trained models. https://pytorch.org/vision/stable/models.html. [Last accessed:: 11/17/2022].

[9] 2022. Supplementary materials of JellyBean. https://arxiv.org/abs/2205.04713. [Last accessed:: 11/17/2022].

[10] 2022. Virtual GPU (vGPU) NVIDIA. https://www.nvidia.com/en-us/data-center/virtual-solutions/. [Last accessed:: 11/17/2022].

[11] 2022. What's Inside Our New DNNCam? Learn About The Hardware. https://boulderai.com/whats-inside-our-new-dnncam-learn-about-the-hardware/. [Last accessed:: 11/17/2022].

[12] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. 2016. TensorFlow: A System for Large-Scale Machine Learning. In *Symposium on Operating Systems Design and Implementation (OSDI)*.

[13] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *IEEE International Conference on Computer Vision (ICCV)*.

[14] Paul Barham, Aakanksha Chowdhery, Jeff Dean, Sanjay Ghemawat, Steven Hand, Dan Hurt, Michael Isard, Hyeontaek Lim, Ruoming Pang, Sudip Roy, et al. 2022. Pathways: Asynchronous distributed dataflow for ML. *arXiv preprint arXiv:2203.12533* (2022).

[15] Hedi Ben-Younes, Rémi Cadene, Matthieu Cord, and Nicolas Thome. 2017. Mutan: Multimodal tucker fusion for visual question answering. In *IEEE International Conference on Computer Vision (ICCV)*.

[16] Steven Bird. 2006. NLTK: the natural language toolkit. In *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*. 69–72.

[17] Surajit Chaudhuri. 1998. An overview of query optimization in relational systems. In *Proceedings of the seventeenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems*. 34–43.

[18] Tianqi Chen, Thierry Moreau, Ziheng Jiang, Lianmin Zheng, Eddie Yan, Haichen Shen, Meghan Cowan, Leyuan Wang, Yuwei Hu, Luis Ceze, et al. 2018. TVM: An automated end-to-end optimizing compiler for deep learning. In *Symposium on Operating Systems Design and Implementation (OSDI)*. 578–594.

[19] Yu Cheng, Duo Wang, Pan Zhou, and Tao Zhang. 2017. A survey of model compression and acceleration for deep neural networks. *arXiv preprint arXiv:1710.09282* (2017).

[20] Daniel Crankshaw, Gur-Eyal Sela, Xiangxi Mo, Corey Zumar, Ion Stoica, Joseph Gonzalez, and Alexey Tumanov. 2020. InferLine: latency-aware provisioning and scaling for prediction serving pipelines. In *Proceedings of the 11th ACM Symposium on Cloud Computing*.

[21] Daniel Crankshaw, Xin Wang, Guilio Zhou, Michael J Franklin, Joseph E Gonzalez, and Ion Stoica. 2017. Clipper: A low-latency online prediction serving system. In *Symposium on Networked Systems Design and Implementation (NSDI)*.

[22] Steven K Esser, Jeffrey L McKinstry, Deepika Bablani, Rathinakumar Appuswamy, and Dharmendra S Modha. 2019. Learned step size quantization. *arXiv preprint arXiv:1902.08153* (2019).

[23] Michael R Garey and David S Johnson. 1979. *Computers and intractability*. Vol. 174. freeman San Francisco.

[24] Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. 2021. Knowledge distillation: A survey. *International Journal of Computer Vision* 129, 6 (2021), 1789–1819.

[25] Robert Grandl, Ganesh Ananthanarayanan, Srikanth Kandula, Sriram Rao, and Aditya Akella. 2014. Multi-resource packing for cluster schedulers. *ACM SIGCOMM Computer Communication Review (CCR)* 44, 4 (2014), 455–466.

[26] Sangtae Ha, Soumya Sen, Carlee Joe-Wong, Youngbin Im, and Mung Chiang. 2012. TUBE: Time-dependent pricing for mobile data. In *ACM Special Interest Group on Data Communication (SIGCOMM)*.

[27] Yitao Hu, Rajrup Ghosh, and Ramesh Govindan. 2021. Scrooge: A Cost-Effective Deep Learning Inference System. In *Proceedings of the ACM Symposium on Cloud Computing*. 624–638.

[28] Yitao Hu, Weiwu Pang, Xiaochen Liu, Rajrup Ghosh, Bongjun Ko, Wei-Han Lee, and Ramesh Govindan. 2021. Rim: Offloading Inference to the Edge. In *Proceedings of the International Conference on Internet-of-Things Design and Implementation*. 80–92.

[29] Yuzhen Huang, Yingjie Shi, Zheng Zhong, Yihui Feng, James Cheng, Jiwei Li, Haochuan Fan, Chao Li, Tao Guan, and Jingren Zhou. 2019. Yugong: Geo-distributed data and job placement at scale. *Very Large Data Base Endowment (VLDB)* 12, 12 (2019), 2155–2169.

[30] Benoit Jacob, Skirmantas Kligys, Bo Chen, Menglong Zhu, Matthew Tang, Andrew Howard, Hartwig Adam, and Dmitry Kalenichenko. 2018. Quantization and training of neural networks for efficient integer-arithmetic-only inference. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2704–2713.

[31] Samvit Jain, Ganesh Ananthanarayanan, Junchen Jiang, Yuanchao Shu, and Joseph Gonzalez. 2019. Scaling video analytics systems to large camera deployments. In *Proceedings of the 20th International Workshop on Mobile Computing Systems and Applications*.

[32] Junchen Jiang, Ganesh Ananthanarayanan, Peter Bodik, Siddhartha Sen, and Ion Stoica. 2018. Chameleon: scalable adaptation of video analytics. In *ACM Special Interest Group on Data Communication (SIGCOMM)*.

[33] Stephen C Johnson et al. 1975. *Yacc: Yet another compiler-compiler*. Vol. 32. Bell Laboratories Murray Hill, NJ.

[34] Daniel Kang, Peter Bailis, and Matei Zaharia. 2018. Blazeit: Optimizing declarative aggregation and limit queries for neural network-based video analytics. *arXiv preprint arXiv:1805.01046* (2018).

[35] Yiping Kang, Johann Hauswald, Cao Gao, Austin Rovinski, Trevor Mudge, Jason Mars, and Lingjia Tang. 2017. Neurosurgeon: Collaborative intelligence between the cloud and mobile edge. *ACM SIGARCH Computer Architecture News* 45, 1 (2017), 615–629.

[36] Hyeonji Kim, Byeong-Hoon So, Wook-Shin Han, and Hongrae Lee. 2020. Natural language to SQL: where are we today? *Very Large Data Base Endowment (VLDB)* 13, 10 (2020), 1737–1750.

[37] Hung Le, Doyen Sahoo, Nancy F Chen, and Steven CH Hoi. 2019. Multimodal transformer networks for end-to-end video-grounded dialogue systems. *arXiv preprint arXiv:1907.01166* (2019).

[38] Yuanqi Li, Arthi Padmanabhan, Pengzhan Zhao, Yufei Wang, Guoqing Harry Xu, and Ravi Netravali. 2020. Reducto: On-camera filtering for resource-efficient real-time video analytics. In *ACM Special Interest Group on Data Communication (SIGCOMM)*.

[39] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*.

[40] Chong Liu, Yuqi Zhang, Hao Luo, Jiasheng Tang, Weihua Chen, Xianzhe Xu, Fan Wang, Hao Li, and Yi-Dong Shen. 2021. City-scale multi-camera vehicle tracking guided by crossroad zones. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[41] Zhuang Liu, Jianguo Li, Zhiqiang Shen, Gao Huang, Shoumeng Yan, and Changshui Zhang. 2017. Learning efficient convolutional networks through network slimming. In *IEEE International Conference on Computer Vision (ICCV)*.

[42] Yao Lu, Aakanksha Chowdhery, and Srikanth Kandula. 2016. Optasia: A relational platform for efficient large-scale video analytics. In *ACM Symposium on Cloud Computing (SoCC)*.

[43] Yao Lu, Aakanksha Chowdhery, Srikanth Kandula, and Surajit Chaudhuri. 2018. Accelerating machine learning inference with probabilistic predicates. In *ACM SIGMOD International Conference on Management of Data (SIGMOD)*.

[44] Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*. 55–60.

[45] Xiangrui Meng, Joseph Bradley, Burak Yavuz, Evan Sparks, Shivaram Venkataraman, Davies Liu, Jeremy Freeman, DB Tsai, Manish Amde, Sean Owen, et al. 2016. Mllib: Machine learning in apache spark. *The Journal of Machine Learning Research* 17, 1 (2016), 1235–1241.

[46] Massimo Merenda, Carlo Porcaro, and Demetrio Iero. 2020. Edge machine learning for AI-enabled IoT devices: A review. *Sensors* 20, 9 (2020), 2533.

[47] Philipp Moritz, Robert Nishihara, Stephanie Wang, Alexey Tumanov, Richard Liaw, Eric Liang, Melih Elibol, Zongheng Yang, William Paul, Michael I Jordan, et al. 2018. Ray: A distributed framework for emerging AI applications. In *Symposium on Operating Systems Design and Implementation (OSDI)*.

[48] Derek G. Murray, Frank McSherry, Rebecca Isaacs, Michael Isard, Paul Barham, and Martín Abadi. 2013. Naiad: A Timely Dataflow System. In *ACM Symposium on Operating Systems Principles (SOSP)*.

[49] Christopher Olston, Noah Fiedel, Kiril Gorovoy, Jeremiah Harmsen, Li Lao, Fangwei Li, Vinu Rajashekhar, Sukriti Ramesh, and Jordan Soyke. 2017. Tensorflow-serving: Flexible, high-performance ml serving. *arXiv preprint arXiv:1712.06139* (2017).

[50] M Tamer Özsu and Patrick Valduriez. 1999. *Principles of distributed database systems*. Vol. 2. Springer.

[51] Rina Panigrahy, Kunal Talwar, Lincoln Uyeda, and Udi Wieder. 2011. Heuristics for vector bin packing. https://www.microsoft.com/en-us/research/publication/heuristics-for-vector-bin-packing/.

[52] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems (NeurIPS)* 32 (2019).

[53] Antonio Polino, Razvan Pascanu, and Dan Alistarh. 2018. Model compression via distillation and quantization. *arXiv preprint arXiv:1802.05668* (2018).

[54] Qifan Pu, Ganesh Ananthanarayanan, Peter Bodik, Srikanth Kandula, Aditya Akella, Paramvir Bahl, and Ion Stoica. 2015. Low latency geo-distributed data analytics. *ACM SIGCOMM Computer Communication Review (CCR)* 45, 4 (2015), 421–434.

[55] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. 2016. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 779–788.

[56] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems (NeurIPS)* 28 (2015).

[57] Francisco Romero, Qian Li, Neeraja J Yadwadkar, and Christos Kozyrakis. 2019. INFaaS: A model-less inference serving system. *arXiv preprint arXiv:1905.13348* (2019).

[58] Francisco Romero, Mark Zhao, Neeraja J Yadwadkar, and Christos Kozyrakis. 2021. Llama: A Heterogeneous & Serverless Framework for Auto-Tuning Video Analytics Pipelines. In *ACM Symposium on Cloud Computing (SoCC)*.

[59] Victor Sanh, Thomas Wolf, and Alexander Rush. 2020. Movement pruning: Adaptive sparsity by fine-tuning. *Advances in Neural Information Processing Systems (NeurIPS)* 33 (2020), 20378–20389.

[60] Haichen Shen, Lequn Chen, Yuchen Jin, Liangyu Zhao, Bingyu Kong, Matthai Philipose, Arvind Krishnamurthy, and Ravi Sundaram. 2019. Nexus: a GPU cluster engine for accelerating DNN-based video analysis. In *ACM Symposium on Operating Systems Principles (SOSP)*.

[61] Zheng Tang, Milind Naphade, Ming-Yu Liu, Xiaodong Yang, Stan Birchfield, Shuo Wang, Ratnesh Kumar, David Anastasiu, and Jenq-Neng Hwang. 2019. Cityflow: A city-scale benchmark for multi-target multi-camera vehicle tracking and re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[62] Can Wang, Sheng Zhang, Yu Chen, Zhuzhong Qian, Jie Wu, and Mingjun Xiao. 2020. Joint configuration adaptation and bandwidth allocation for edge-based real-time video analytics. In *IEEE Conference on Computer Communications (INFOCOM)*.

[63] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*. 38–45.

[64] Zhihui Yang, Zuozhi Wang, Yicong Huang, Yao Lu, Chen Li, and X Sean Wang. 2022. Optimizing Machine Learning Inference Queries with Correlative Proxy Models. *arXiv preprint arXiv:2201.00309* (2022).

[65] Matei Zaharia, Mosharaf Chowdhury, Tathagata Das, Ankur Dave, Justin Ma, Murphy McCauly, Michael J Franklin, Scott Shenker, and Ion Stoica. 2012. Resilient Distributed Datasets: A Fault-Tolerant Abstraction for In-Memory Cluster Computing. In *Symposium on Networked Systems Design and Implementation (NSDI)*.

[66] Chengliang Zhang, Minchen Yu, Wei Wang, and Feng Yan. 2019. Mark: Exploiting cloud services for cost-effective, slo-aware machine learning inference serving. In *USENIX Annual Technical Conference (ATC)*.

[67] Wuyang Zhang, Zhezhi He, Luyang Liu, Zhenhua Jia, Yunxin Liu, Marco Gruteser, Dipankar Raychaudhuri, and Yanyong Zhang. 2021. Elf: accelerate high-resolution mobile deep vision with content-aware parallel offloading. In *ACM Conference on Mobile Computing and Networking (MobiCom)*.