Distributed Scheduling of Event Analytics across Edge and Cloud

Rajrup Ghosh* and Yogesh Simmhan*
*Department of Computational and Data Sciences,
Indian Institute of Science, Bangalore 560012, India
Email: rajrup@grads.cds.iisc.ac.in, simmhan@cds.iisc.ac.in

Abstract

Internet of Things (IoT) domains generate large volumes of high velocity event streams from sensors, which need to be analyzed with low latency to drive decisions. Complex Event Processing (CEP) is a Big Data technique to enable such analytics, and is traditionally performed on Cloud Virtual Machines (VM). Leveraging captive IoT edge resources in combination with Cloud VMs can offer better performance, flexibility and monetary costs for CEP. Here, we formulate an optimization problem for placing CEP queries, composed as an analytics dataflow, across a collection of edge and Cloud resources, with the goal of minimizing the end-to-end latency for the dataflow. We propose a brute-force optimal algorithm (BF) and a Genetic Algorithm (GA) meta-heuristic to solve this problem. We perform comprehensive real-world benchmarks on the compute, network and energy capacity of edge and Cloud resources for over 17 CEP query configurations. These results are used to define a realistic simulation study that validates the BF and GA solutions for 45 diverse dataflows. Our results show that the GA approaches within 99% of the optimal BF solution that takes hours, maps dataflows with $4 - 50$ queries within $1 - 25$ $secs$, and in fewer than 10% of the experiments is unable to offer a feasible solution.

Index Terms

Internet of Things; Complex Event Processing; Cloud Computing; Big Data platforms; Query Partitioning; Distributed Systems; Low Power Processing; Scheduling

I. Introduction

Internet of Things (IoT) is a new computing paradigm where pervasive sensors and actuators deployed in the physical environment, with ubiquitous networking and communication, allow us to observe, manage and enhance the efficiency of the system or the comfort of humans. The application domains motivated by IoT spans cyber-physical utilities such as smart power and water with their metering infrastructure [1], [2], to health and lifestyle applications like Fitbit and smart watches [3], and onto even mobile platforms such as unmanned drones and vehicles [4].

A key requirement for IoT applications is to apply analytics over the data collected from the distributed sensors in order to make intelligent decisions that can be communicated back to the ecosystem. Often, these decisions need to be performed on data that is *continuously streaming* from the edge devices, sometimes at high input rates [5]. Fitness bands, say like Fitbit constantly track the activity and heart rate of users, and trigger motivational messages based on the current activity level of the user or their social network. Occasionally, these analytics and decision making are time-sensitive, and require a *low latency response*. For e.g., in the power grid, we may need to make decisions on load shifting or curtailment based on power metering data arriving from customers across the city in real time to avoid outages or brownouts [2], [6]. In certain cases, there may be *security and privacy concerns* on the data collected as well [7]. This includes platforms such as drones used for urban surveillance or disaster management where the images and videos may be of a sensitive nature.

Big Data platforms for stream and event processing offer the ability for continuous analytics to IoT applications [8], [9]. These platforms are designed for low latency processing of continuous data or event streams, such as from physical sensors or even from social network feeds [10], [11]. In particular, *Complex*

Event Processing (CEP) engines allow users to define intuitive SQL-like queries over event streams that are executed continuously on tuples as they arrive [12]. These can be used to detect when thresholds are breached to trigger alerts, perform aggregation over temporal windows of events for smoothing or visualization, or detect sequence of events having a specific pattern that indicate a situation of interest. In some cases, they also have domain semantics encoded within the query [13]. These queries can be composed into a dataflow graph for online decision making applications, where vertices are queries and edges are streams carrying events between queries. CEP platforms are common for analytics in IoT domains such as smart power grids, transportation and sports analytics [14], [15], [16].

A common information processing architecture in IoT domains is to move data from thousands or millions of edge devices centrally into *public Clouds*, where CEP or other analytics engines hosted on Virtual Machines (VM) can process the incoming streams, and data can be persisted for subsequent mining and visualization. Cloud vendors even offer customized software stacks for IoT applications that include such event processing and rule-based engines [1] [2]. Here, the edge devices essentially serve as "dumb" sensors that observe and transmit the data to the Cloud, with all intelligence and analytics performed centrally in the data center.

However, there are a few *limitations* to this Cloud-only model.

1) Moving large streams of events from the edge to the Cloud introduces data transfer latency, thereby introducing a fixed network latency time and additional bandwidth transfer time to the decision making. The network may also become the bottleneck when many such edge devices evacuate data from the same region. There are also associated monetary costs for the use of the network and hosting VMs in the Cloud, and energy cost paid by the constrained edge devices in moving large data streams over the network.
2) Edge devices are getting more capable by the day with advances in energy-efficient mobile processors and battery technology that gives them longevity. Commoditization of platforms like Raspberry Pi and Arduino has also made them very affordable. By treating these edge devices purely as sensing platforms, their captive computing capability for analytics is under-used and instead Cloud VMs with a pay-as-you-go pricing model are billed and used.
3) Lastly, having a hub and spoke model where all data is ingested to the Cloud does not give us any fine-grained control over where data from the IoT deployment can go. In applications that are privacy-sensitive or have security implications, there may be requirements for permitting certain analytics only on the private network of the edge, or anonymizing data streams before they go to public Clouds.

In this article, we address some of these limitations by *proposing an approach to distributing event-based analytics across edge and Cloud resources to support IoT applications*. We consider an IoT deployment scenario with multiple event streams generated at the edge at high frequency, a user-defined analytics dataflow composed of CEP queries that needs to be performed on these streams, and several edge devices in a private network and public Cloud VMs available to perform the queries. Our goal is to determine a distributed placement of these queries onto the edge and Cloud resources such that the *end-to-end latency* for performing the event analytics is minimized to support timely decision making. This placement schedule needs to meet *constraints* such as the throughput capacity supported on edge and Cloud machines by the queries, bandwidth and latency limits of the network, and energy capacity of the edge devices. The latter aspect is particularly novel – edge devices like motes and Pi's are often powered by batteries that are themselves recharged through small solar panels, and hence have a *limited energy budget* within a daily cycle that we must meet.

The benefits of computing across edge devices and Cloud is being recognized off-late [17]. There has been significant interest in off-loading computing from smart phones and mobile devices to the Cloud [18]. Many of these only consider moving modular parts of the application from the edge to the

[1] https://aws.amazon.com/iot/how-it-works/

[2] https://www.microsoft.com/en-in/server-cloud/internet-of-things/overview.aspx

Cloud, rather than to other edge devices as well, which is our focus. Some do support an edge-only solution by leveraging intermittently connected mobile devices, but these are not amenable to low latency stream processing which we target [19]. More generally, *Fog Computing* emphasizes the exclusive use of edge devices for computation but lack centralized control and reliability [20]. Such peer-to-peer (P2P) systems have been considered in the past as well for file and compute sharing, but do not address issues of energy constraints that is critical in IoT environments [21]. Literature from a decade ago has examined query operator placement in Wireless Sensor Networks (WSN), and energy and communication costs on constrained motes were addressed [22]. While similar in spirit to our problem, we consider placement at the granularity of a query rather than operator, support streams that are orders of magnitude faster, and also have access to Cloud VMs rather than only on motes.

Our prior short paper has considered a simpler problem of bi-partitioning a CEP pipeline between a single edge device and the cloud, including factors like incoming event rate, throughput of resources and enforcing privacy constraints [23]. This current article addresses a more comprehensive problem, and considers multiple edge devices, energy constraint on those devices and offers detailed experimental results that was absent earlier.

We make the following specific contributions in this article.

1) We formulate the problem of placement of queries in a given directed acyclic graph (DAG) onto distributed edge and Cloud resources, having computing, network and energy constraints, as a *combinatorial optimization problem*, with the objective function being to minimize the end-to-end processing latency (§ IV).
2) We propose an *optimal brute force approach* to solving this problem, but with exponential time complexity, and also a more practical solution based on the *Genetic Algorithm meta-heuristic* (§ V).
3) We perform and present *comprehensive, real-world micro-benchmarks* for a wide class of 17 CEP queries at different input event rates. These benchmarks empirically evaluate the throughput of edge and Cloud resources, using Raspberry Pi and Microsoft Azure VMs as exemplars, the energy capacity of the edge device for different analytics queries, and the network characteristics (§ VI).
4) We use the benchmark results to conduct a *detailed and realistic simulation study* of the effectiveness of the query placement approaches for a wide range of 45 synthetic CEP dataflows, with varying numbers of queries, input event rates, and resource availability. We offer a rigorous analysis of the experimental results, using both qualitative metrics and time complexity analysis (§ VII).

We complement the above contributions with a background motivating the problem scenario and CEP (§ II), a literature review of related work (§ III) and a discussion of open problems and future work in this emerging research area (§ VIII).

II. Background and Motivation

Our problem is motivated by on our prior work at the University of Southern California [2] and ongoing work at the Indian Institute of Science [24] on developing a Smart Campus IoT fabric to support emerging applications such as smart power and water management at a campus or township scale. In such a scenario, there are three functional layers: *the sensing fabric, communication infrastructure,* and *data processing and analytics platforms*. The sensors may include plug load and water level sensors that generate data at rates of hundreds per second to once every few minutes, and number in the thousands. The communication layer uses a mix of wired campus LAN and wireless *ad hoc* networks, with the latter using a network of motes to move data from the sensor location to the nearest wired LAN. A gateway device, such as a Raspberry Pi, typically acts as the interface between tens of motes, receiving and passing along observation streams from hundreds of sensors, and the Big Data platform where the events are processed, analyzed and decisions made. The data platform itself runs on a server on campus, or more frequently, on a set of Virtual Machines (VM) in a public Cloud like Amazon AWS or Microsoft Azure.

Complex Event Processing (CEP) engines are part of the data platform, and offer a query interface to define patterns of interest that need to be detected from one or more event streams. Each event is a

TABLE I: Example Siddhi CEP queries used in experiments

Query Type	Siddhi Query Definition

Filter	`define stream inStream (height int);`
	`from inStream[height < 150]`
	`select height insert into outStream ;`

Sequence
Match 3 events
	`from every e1 = inStream, e2 = inStream[e1.height == e2.height], e3 = inStream[e3.height == e2.height]`
	`select e1.height as hi1, e2.height as hi2, e3.height as hi3`
	`insert into outStream ;`

Pattern
Match 3 events
	`from every e1 = inStream -> e2 = inStream[e1.height == e2.height] -> e3 = inStream[e2.height == e3.height]`
	`select e1.height as hi1, e2.height as hi2, e3.height as hi3`
	`insert into outStream ;`

Aggregate (Batch) *Window Size = 60*	`from inStream #window.lengthBatch(60)` `select avg(height) as AvgHeight insert into outStream ;`
Aggregate (Sliding) *Window Size = 60*	`from inStream #window.length(60)` `select avg(height) as AvgHeight insert into outStream ;`

tuple with a set of named and typed columns and their values, just like a row in a relational database. Frequently, they contain the timestamp and the source stream ID as columns, besides the observed values. Such event-driven decisions can predict energy surges based on patterns of load characteristics, identify demand-supply mismatch to initiate demand response optimization, detect water usage patterns to trigger pumping operations, and identify water quality changes to notify relevant consumers [2], [25]. CEP engines like Siddhi [9] and Esper [26] register these queries and execute them continuously over event streams for long periods of days or weeks.

CEP queries are of 4 four major types – filter, sequence, pattern and aggregate, which are illustrated in Table I using data from a water level sensor that measures the depth of water in an overhead tank. *Filter* queries match a property predicate against fields in the incoming event, and only those events that match the predicate are placed in the output stream. For e.g., '`height < 150`' in row 1 of Table I is the predicate that is matched, detecting a situation where the water level may have dropped below a threshold. A *Sequence* query matches predicates on consecutive events, and if all predicates match the sequence of events, those events are placed in the output stream. For e.g., row 2 of Table I shows a 3-event sequence that detects when *consecutive* events `e1, e2` and `e3` have the same height values, indicating that the water level in the tank has not changed during that period.

A *Pattern* query is similar to a sequence query, except that it relaxes the requirement that the matching events appear contiguously. For e.g., in row 3 of Table I, we match events with the same height values for 3 *successive* events `e1, e2` and `e3`, i.e., there may be other non-matching events between `e1` and `e2`, and/or `e2` and `e3`. Lastly, *Aggregate* queries perform an aggregation function on a window of events. The aggregation may be operations like average ('`avg(···)`' in row 4 and 5 of Table I), summation or maximum. The window is specified by giving a *count* of consecutive events to be included, such as 60 events in the example. The window may be formed by *sliding* over the input stream – by shifting one event out of the window as a new event arrives into the window, or by *batching* events such that every event appears in only one batch.

The CEP queries can be composed into a *directed acyclic graph (DAG)* or a dataflow, where vertices are queries and edges indicate events passed from the output stream of one query to the input stream of the downstream query. Multiple queries may run within the same CEP engine on a machine, or different queries may run on CEP engines in different machines and coordinate their execution using events streamed over the network. This allows complex analytics, such as pre-processing, aggregation, pattern detection and so on to be composed together into IoT applications that distributedly execute on one or more streams. The output stream from the DAG may further be processed by other Big Data platforms, such as a data mining or a notification service.

Typically, the CEP engine runs on Cloud VMs to allow *centralized processing* of all data streams. So

while the source to the DAG is from the edge gateway devices, the vertices and the output streams are all hosted on the Cloud. However, edge devices such as the Raspberry Pi have multiple cores, perform $25 - 50\%$ as fast as desktop or server class CPUs, have $1 - 2$ GB of RAM, SD card storage, and LAN/WLAN connectivity, and run standard Linux distributions. Their relatively small energy footprint ensures that they can run for a whole day on a $\sim 8,600$ mAh battery source (which is about 4 smartphone batteries). In field deployments, they are coupled with a solar panel that re-charges them during the daytime. For e.g., a 4 sq.ft. panel is adequate to fully recharge the Pi's battery in a single day.

Since tens or hundreds of these may be deployed in a campus or township (e.g., one per building), they offer captive and sustainable computing resources to complement Cloud resources, and can be used for distributed execution of CEP queries in a DAG. However, their constrained computing capability and limited energy budget means that we require efficient scheduling strategies to place the queries onto different edge devices. At the same time, since much of the other Big Data platforms and decision making logic run on Cloud VMs, the results from the dataflow need to be pushed to VMs in the Cloud. This means at least one captive Cloud VM can additionally support running these CEP queries besides the edge devices. While more VMs could be provisioned on the Cloud for CEP, the pay-as-you-go model means that for each new VM, there is an additional monetary cost involved. For e.g., the Microsoft Azure D2 VM size we consider in our experiments costs about US$5 per day to rent [3], or US$150 per month – for which cost about three Raspberry Pi 2 edge devices can be purchased.

III. Related Work

There are three primary related research areas relevant to our article: mobile Clouds, Fog and Peer-to-Peer (P2P) computing, and query processing in sensor networks.

Mobile Cloud Computing [17] has grown as a research area that lies at the intersection of mobile devices such as smart phones and the Cloud. The key idea is to use these personal devices as a thin client to access rich services hosted on Clouds, forming a variation of a client-server model. In addition, the concept of Cloudlets has been proposed as an additional layer that sits between the edge and the Cloud to help reduce latencies while offering superior computing power than the edge alone [27]. This is conceptually the computing equivalent of Content Distribution Networks that move data closer to the edge. Both these paradigms conceive of interactions between a single client and a remote Cloud/Cloudlet, which is in contrast to our approach of leveraging the collective capabilities of distributed edge devices and the Cloud.

Specific research papers have extended these broad concepts further. CloneCloud is an application partitioning framework for mobile phones that off-loads a part of the application execution from the edge device to device "clones" in a public Cloud [18]. Partitioning is done by migrating a thread from the mobile device at a chosen point in the application to the clone in the cloud. After execution on the cloud, it is re-integrated back onto the mobile device. It models energy as a function of CPU, display and network state, and this is considered in their partitioning strategy. A further extension tries to improve upon this by reducing state transfer costs for dynamic offloading [28]. Although CloneCloud partitions an interactive application across mobile and Cloud, it does not address streaming analytics applications essential for the IoT domain, where factors like latency and throughput need to be addressed, nor does it use multiple edge devices.

Others do deal with partitioning of data stream applications between mobile devices and Cloud to maximize the throughput of stream processing [29]. This framework considers sharing VM instances among multiple users in the Cloud to improve VM utilization, and solves the problem using a genetic partitioning algorithm like us. The empirical evaluation, however, uses a QR code recognition application, which is unlike the high rate event analytics that we support for IoT domains. Further, it does not consider distributing tasks to multiple edge resources either.

[3] https://azure.microsoft.com/en-in/pricing/details/virtual-machines/

Our own prior work [23] has considered a similar bi-partitioning of a CEP query pipeline between a single edge device and the cloud. Factors like compute time on each query, incoming event rate and latency between resources are used to find an edge-cut in the DAG such that latency is reduced. It also considers enforcing privacy constraints on event streams to determine if a stream is allowed on a resource. It does not, however, consider multiple edge devices and could be solved optimally using a dynamic programming solution. Energy constraints were not considered either, and there was no empirical evaluation. Besides the flexibility of multiple resources and energy constraints considered in our current article, we also present robust benchmarks and empirical evaluation.

Distributed query processing on multiple edges and the Cloud have been considered for feed-following applications [30]. Here, database views of applications that follow social network feeds are distributed to edge devices, with query operators that are applied on the feeds by existing relational databases engines. The problem is modeled as a view placement problem with the goal of optimizing communication between sub-queries running on the edge and the Cloud. However, there is a difference between their view placement and our query placement problem on edge devices. In the former, the edge devices can communicate with each other only through the Cloud, causing a "star" network topology. This reduces the optimization problem to linear time. We instead allow the edges to communicate with each other, which is feasible as they are typically on the same private network.

Serendipity [19] is a more comparable work that uses remote computing collaboratively among closely connected mobile devices. It explores off-loading of computationally intensive tasks onto other intermittently connected mobile devices rather than the Cloud. In their model, jobs are distributed to nearby mobile devices with the aim of reducing the job completion time and conserving the device's energy. Unlike us, their approach does not work for streaming applications, as that would not be viable for transiently connected devices, and they do not consider Cloud resources either.

More broadly, the concept of *Fog Computing* is gaining traction in the IoT domain [31]. Here, the Cloud with a massive centralized data center is supplanted by a fog of wireless edge devices that collaboratively offer computing resources. These have the benefit of low-latency communication and the ability to self-organize locally, but lack full (centralized) control and their availability is unreliable [20]. Fog computing platforms are still in a nascent stage, but our approach to distributed analytics execution across edge devices (Fog) and the Cloud offers a model for coordinating their computation, and leveraging the best features of these two paradigms.

As such, there has historically been work on such *Peer-to-Peer (P2P)* systems, where interconnected nodes can self-organize into a network topology to share files, CPU cycles, storage and bandwidth [21]. Peers can offload execution of tasks to other peers to speed up their job completion, and significant work on lookup services such as distributed hash tables have taken place. This type of content search and retrieval requires guarantees on QoS parameters like timely results, utilization of resources, response time and correctness. For e.g., [32] [33] give algorithms for distributing tasks to a set of peers using hierarchical coalition formation. Our work operates on a more deterministic set of edge devices (peers), but can benefit from the management services developed for *ad hoc* P2P systems. Issues like energy and mobile devices were not relevant for those architectures but gain prominence in IoT ecosystems we target.

Yet another related area from a decade back is on query processing in *Wireless Sensor Networks (WSN)*. Here, distributed sensors (motes) deployed to measure environmental parameters assemble together to solve streaming analytics task [34] [10]. Query plan optimizing and placement techniques have been explored in the context of a large number of sensor nodes. A virtual tree topology is created with an elected leader node which receives query requests from users, and sends smaller tasks to worker nodes having the relevant streams [22]. Intermediate nodes in the tree can partially process the query or forward the results back to the leader to build the final result set. Like us, these strategies try to reduce the energy and communication costs on these embedded devices. Rather than consider individual queries that are partitioned into sub-queries, we consider a DAG of queries with placement at the granularity of query. CEP engines have a richer query model as well, and we also have access to Cloud VMs rather than execute exclusively on motes on the edge.

IV. Problem Formalization

In this section, we formally state the CEP query placement problem for a DAG that we motivated earlier in Sec. II, and formulate it as a constrained optimization problem. The solution approaches to the problem is offered in the next section.

A. Preliminaries

The dataflow application that is a composition of queries executing over event streams is represented as a *Directed Acyclic Graph (DAG)* of vertices and edges: $\mathcal{G} = \langle \mathbb{V}, \mathbb{E} \rangle$. $\mathbb{V} = \{v_i\}$ is the set of *CEP queries* that are the vertices of the DAG, and $\mathbb{E}$ is the set of *event streams* that connect the output of query v_i to the input of the next query v_j, and form the directed edges of the DAG. $\mathbb{E}$ is given by:

$$\mathbb{E} = \{e_i \mid e_i = \langle v_i, v_j \rangle, v_i \in \mathbb{V}, v_j \in \mathbb{V} \cup \phi\}$$

The output event(s) of a query v_i is *duplicated* across all the outgoing edges from that vertex. The inputs for a query v_i from multiple incoming edges follow an *interleave semantics*, meaning events from all incoming edges are appended to a single logical input stream for the vertex. The queries that receive the initial input event streams into the DAG are called *source queries*, and are characterized by having no incoming edges. The set of source queries is given by:

$$\mathbb{V}^{SRC} = \{v_j | \nexists e_i = \langle v_i, v_j \rangle \in \mathbb{E}, \forall v_i \in \mathbb{V}\}$$

Source queries do not perform any computation. These are just no-op tasks that generate and pass events downstream.

The queries that emit the output event streams from the DAG are called *sink queries*, and have unbound outgoing edges that are incident on a dummy vertex ϕ. This set is given by:

$$\mathbb{V}^{SNK} = \{v_j | \exists e_j = \langle v_j, \phi \rangle \in \mathbb{E}, \forall v_j \in \mathbb{V}\}$$

A query executes on a specific *computing resource* r_k and the set of all computing resources available within the IoT infrastructure is given by $\mathbb{R} = \{r_k\}$. We consider two classes of computing resources – *edge devices* such as smart phones and Raspberry Pi's, and *Cloud resources* such as VMs, each having a specific computing capability. These two computing resources form two mutually exclusive sets, $\mathbb{R}_E$ for edge devices and $\mathbb{R}_C$ for Cloud VMs, respectively. Thus, $\mathbb{R}_E \cup \mathbb{R}_C = \mathbb{R}$ and $\mathbb{R}_E \cap \mathbb{R}_C = \varnothing$. A *resource mapping function* $\mathcal{M}$ indicates the resource on which a query executes:

$$\mathcal{M} : \mathbb{V} \rightarrow \mathbb{R}$$

A *path* in a DAG is defined as a unique sequence of edges from the source query to the sink query. A path p_i of length n in the graph $\mathcal{G}$ is an alternating sequences of $n+1$ vertices and n edges, starting at a source and ending at the sink.

$$p_i = \langle v_0, v_1 \rangle, ..., \langle v_k, v_{k+1} \rangle, ..., \langle v_n, v_{n+1} \rangle$$
$$\text{where} \quad v_0 \in \mathbb{V}^{SRC}, v_{n+1} \in \mathbb{V}^{SNK}$$

A *selectivity* function denoted by $\sigma(v_i)$ is associated with each query of the DAG, and is a statistical measure of the average number of output events generated for every input event consumed by the query. Using duplicate semantics, the selectivity of each outgoing edge is same as the selectivity of the vertex writing to that edge. The *stream rate* defines the number of events passing per unit time on a stream. The *incoming stream rate* of the DAG is denoted by Ω^{in}, and is the sum of the stream rates emitted by all source queries in the DAG. Similarly the *outgoing stream rate* denoted by Ω^{out} is the sum of output rate of events emitted by the sink queries. Then *selectivity of the whole DAG*, $\sigma(\mathcal{G})$, for is given by:

$$\sigma(\mathcal{G}) = \frac{\Omega^{out}}{\Omega^{in}}$$

The *incoming stream rate* ω_i^{in} for a vertex v_i is the sum over the stream rates on all the incoming edges, due to interleave semantics. Due to duplicate semantics, the *outgoing stream rate* ω_i^{out} for a vertex v_i can be defined as the product of its incoming stream rate ω_i^{in} and its selectivity $\sigma(v_i)$.

For simplicity, if the output stream rate for all source queries $v_k \in \mathbb{V}^{SRC}$ is uniform, we have ω_k^{out} as:

$$\omega_k^{out} = \frac{\Omega^{in}}{|\mathbb{V}^{SRC}|}$$

Based on this, we can recursively compute the input and output stream rates for downstream vertices v_j in the DAG as follows:

$$\begin{aligned}
\omega_i^{in} &= \omega_k^{out} && \forall \langle v_k, v_i \rangle \in \mathbb{E}, v_k \in \mathbb{V}^{SRC} \\
\omega_j^{in} &= \sum_{\langle v_i, v_j \rangle \in \mathbb{E}} \big(\omega_i^{in} \times \sigma(v_i)\big) && \forall v_i \in \mathbb{V}, v_i \notin \mathbb{V}^{SRC} \\
\omega_j^{out} &= \omega_j^{in} \times \sigma(v_j) && \forall v_j \in \mathbb{V}, v_j \notin \mathbb{V}^{SNK}
\end{aligned}$$

Thus, the outgoing stream rate Ω^{out} for the entire graph $\mathcal{G}$ is given by:

$$\Omega^{out} = \sum_{v_i \in \mathbb{V}^{SNK}} \omega_i^{in} \times \sigma(v_i)$$

Latency (or compute latency) denoted by λ_i^k is defined as the time taken to process one event by a query v_i on an exclusive resource r_k. If n queries are placed on the same resource r_k, the latency for each query becomes $\sum_i \lambda_i^k, \quad \forall (v_i, r_k) \in \mathcal{M}$ due to round-robin scheduling [4]. If λ is the latency time in seconds taken by a query to process a single event on a resource, λ^{-1} is the *throughput* that can be processed by that query on that resource in 1 second.

Let the *size of an event* that is emitted by query v_i on its outgoing edge(s) be denoted by d_i. The *network latency* and *network bandwidth* between two resources r_m and r_n is denoted by $l_{m,n}$ and $\beta_{m,n}$, respectively. Let the *query path* $\mathbb{P}$ be defined as the set of all paths of events flowing from $\mathbb{V}^{SRC}$ to $\mathbb{V}^{SNK}$ in the DAG. Therefore, the *end-to-end latency along a path* $p \in \mathbb{P}$ for a given resource mapping $\mathcal{M}$ can be defined as the sum of the compute latency, and the network transfer time:

$$L_p = \sum_{\substack{\langle v_i, v_j \rangle \in p_i \\ (v_i, r_m) \in \mathcal{M} \\ (v_j, r_n) \in \mathcal{M}}} \Big(\lambda_i^k + \big(l_{m,n} + \frac{d_i}{\beta_{m,n}}\big)\Big)$$

The maximum over the end-to-end latency values along all paths $p \in \mathbb{P}$ gives us the *end-to-end latency for the DAG*,

$$L_{\mathcal{G}} = \max_{\forall p \ \in \mathbb{P}} (L_p)$$

$L_{\mathcal{G}}$ is also called the *makespan* of the DAG and the path along which this maximum time appears is called the *critical path*.

B. Constraints

Based on the motivating scenario introduced in the background, we define several constraints that need to be satisfied when performing a mapping of queries of resources.

[4]For simplicity, we do not consider multi-thread execution of queries by a single CEP engine on a resource. Resources with multiple cores can instead be modeled as multiple resources.

Constraint 1: All source vertices should be mapped to an edge device, while the sink vertices should be on the Cloud.

$$\forall (v_i, r_k) \in \mathcal{M} \quad | \quad v_i \in \mathbb{V}^{SRC} \implies r_k \in \mathbb{R}_E$$
$$| \quad v_i \in \mathbb{V}^{SNK} \implies r_k \in \mathbb{R}_C$$

This constraint ensures that source queries are co-located on the edge device that is generating the input event stream. Likewise, given that analytics performed after the CEP are hosted on the Cloud, the sink queries must be placed on the VM resource.

Constraint 2: Given an input rate ω_i^{in} on vertex v_i, we should select r_k such that,

$$\omega_i^{in} < \frac{1}{\lambda_i^k} \quad \forall v_i \in \mathbb{V}$$

If multiple queries are running on the same resource r_k, then input rate ω_i^{in} on vertex v_i that r_k can handle is given by:

$$\omega_i^{in} < \frac{1}{\sum\limits_{(v_i, r_k) \in \mathcal{M}} \lambda_i^k} \quad \forall v_i \in \mathbb{V}$$

The maximum event rate that a resource r_k can handle when exclusively running a query v_i is given by the inverse of its latency $\frac{1}{\lambda_i^k}$. Hence, we should ensure a mapping of a query to a resource such that it does not receive an input rate greater than the throughput supported by that query on that resource.

Edge resources are generally run on batteries with a fixed power capacity. Let the *power capacity* available for an edge device r_k be C_k, given in mAh, when fully charged. Let ϵ_i^k be the *power drawn* on the edge resource $r_k \in \mathbb{R}_E$ by a query v_i to process a single event. Let the *time interval between charging* this edge device r_k, given in hours, be denoted as τ_k, be it through solar regeneration or by replacing the battery.

Constraint 3: The queries running on a edge device r_k should not fully drain out the battery capacity of that resource within the recharge time period τ_k.

$$(\omega_i^{in} \times \tau_k) \times \epsilon_i^k \leq C_k$$

We assume that DAGs once registered with the engine and placed on resources run for a much longer time (say days or weeks) than the recharge period (say 24 hours). Thus our optimization plan should take this energy constraint into consideration [5].

C. Optimization Problem

Given a DAG $\mathcal{G} = \langle \mathbb{V}, \mathbb{E} \rangle$ and a set of edge and Cloud resources $\mathbb{R}$, find a resource mapping $\mathcal{M}$ for each query $v_i \in \mathbb{V}$ on to a resource $r_k \in \mathbb{R}$ such that the mapping meets the Constraints 1, 2 and 3 while minimizing *the end-to-end latency for the DAG.*

In other words, find the mapping that gives the minimum possible end-to-end DAG latency,

$$\widehat{L_{\mathcal{G}}} = \min_{\forall (\mathbb{V}, \mathbb{R}) \in \mathcal{M}} (L_{\mathcal{G}})$$

V. Approaches to Solve the Optimization Problem

There have been a multitude of techniques that have been proposed to solve optimization problems, much like the ones we have used [35]. Here, we present two approaches for solving the placement problem: one, a *Brute Force (BF) approach* that gives the optimal solution while being computationally intractable for large problem sizes, and the other which translates the problem to a *Genetic Algorithm (GA) meta-heuristic* and gives an approximate but fast solution.

[5]For simplicity, we consider that discharging of a battery by an edge resource is linear with time and its full recharge is instantaneous at every time τ_k. In practice, batteries have non-linear discharge cycles based on their present capacity, and batteries charged by solar panels may have constant charging/discharging occurring concurrently on the time of day.

A. Brute Force Approach (BF)

Given the Constraint 1 that source vertices $v_i \in \mathbb{V}^{SRC}$ are pinned to the edge devices and sink vertices $v_i \in \mathbb{V}^{SNK}$ are pinned to Cloud resources, our goal is to find a mapping for the n intermediate vertices of the DAG to either edge or Cloud resources, where $n = \big(|\mathbb{V}| - (|\mathbb{V}^{SRC}| + |\mathbb{V}^{SNK}|)\big)$. In the process, we wish to minimize the end to end latency of the query DAG and also meet the other two constraints.

The Brute Force (BF) approach is a naïve technique which does a combinatorial sweep of the entire parameter space. Here, each of the n vertices are placed in every possible $|\mathbb{R}|$ resources as a trial mapping. For each trial mapping, the constraints are evaluated and if all are satisfied, the end-to-end latency for the DAG $L_{\mathcal{G}}$ is calculated for this mapping. If this latency is smaller than the previously known minimum latency from an earlier trial, then this smaller latency is set as the current minimum latency and the resource mapping is stored as the current best mapping. Once all possible combinations of placing the queries onto resources have been attempted, the current minimum latency value is reported as the best end-to-end DAG latency, $\widehat{L_{\mathcal{G}}}$, and its corresponding mapping returned.

1) Complexity Analysis: The brute force algorithm is a provably optimal solution since it considers all possible solutions. However, its computational cost is high, as optimal solutions to such optimization problems tend to be NP-Hard.

The number of trials that are performed is an exponential function, given by $|\mathbb{R}|^n$. The primary task performed in each trial is to test if the constraints are met, which requires just an $\mathcal{O}(|\mathbb{V}|)$ pass through all the vertices in the DAG, and to calculate the latency $L_{\mathcal{G}}$ for the trial mapping. Calculating the latency is the dominating cost, and requires us to find the critical path of the DAG for each mapping. Finding the longest paths from a source to a sink in a DAG is a linear time algorithm with an asymptotic time complexity of $\mathcal{O}(|\mathbb{V}| + |\mathbb{E}|)$. So, the overall time complexity for the BF algorithm is given by $\mathcal{O}\big((|\mathbb{V}| + |\mathbb{E}|) \times |\mathbb{R}|^n\big)$.

B. Genetic Algorithm based Optimization Problem Solver

Finding an optimal placement of the query to resources is a non-linear optimization problem, which makes it difficult to use heuristics like integer linear programming [18]. It has been seen that many NP-complete problems have been well-solved heuristically using evolutionary meta-heuristic algorithms like Genetic Algorithm (GA) and Particle Swarm Optimization (PSO) [36]. Our goal of finding the optimal query placement introduces an additional challenge of satisfying the constraints as well – converged solutions from such meta-heuristics may not be feasible due to compute, network, and/or energy violations. These have to be modeled too.

Given that GA has solved hard graph-based problems like Job Scheduling and Traveling Salesman Problem (TSP) with considerable accuracy, we chose this technique to solve this optimization problem. GA also offers the flexibility of allowing itself to be modified to produce solutions which satisfy multiple constraints.

There are four integral components to a GA approach [36]. *Chromosomes* contain solutions to the problem being solved. A *Population* is the set of all chromosomes whose solutions are being considered. A *Generation* is the number (iterations) of evolutions that the chromosomes in the population have undergone. And the *Objective Function* gives the measure of fitness of a chromosome. Defining the GA solution requires us to map our placement problem to each of these stages.

A chromosome $Q = \{q_0, q_1, ...q_{n-1}\}$ gives the placement of a set of $n = \big(|\mathbb{V}| - (|\mathbb{V}^{SRC}| + |\mathbb{V}^{SNK}|)|\big)$ queries onto a set of resources $\mathbb{R}$. n is the number of variables present in the GA. The chromosome's values q_i are encoded with an integer value in the range $(0, |\mathbb{R}| - 1)$ such that it represents the resource number to which the i^{th} query gets mapped to. A set of chromosomes form a population and the size of the population, p, is a fixed parameter that does not change across the generations.

The zeroth generation of the population is initialized randomly with p chromosomes. In every generation, an optimization function F gives the *fitness value* $F(j)$ for the j^{th} chromosome c_j in the population. For our placement problem, the fitness value is calculated as the end-to-end latency of the DAG (in seconds) obtained from the critical path of the DAG for the placement solution provided by a chromosome.

Apart from the population, we maintain a *best-fit chromosome* which is the solution with the best fitness value seen so far across all generations. At the end of each generation, the current population's new chromosomes are compared with the best-fit chromosome to see if an improved solution has been discovered, and if so, the best-fit chromosome is updated to this.

We use a "roulette wheel" algorithm to select the candidate chromosomes from the current population to consider for evolution into the population of the next generation. We first calculate the total fitness value for the current population,

$$\overline{f} = \sum_{j=0}^{p-1} F(j)$$

Then, we calculate the probability mass function (PMF), $\rho_j = \mathcal{P}(J = j)$, which gives the probability of selecting a chromosome c_j from the population $(c_0, c_1, ..., c_{p-1})$, where

$$\rho_j = \frac{F(j)}{\overline{f}}$$

Next, we compute the cumulative distribution function (CDF) of this PMF as

$$\delta_j = \sum_{k=0}^{j} \rho_k$$

A random real number x in the range $[0..1]$ is then generated. If $x \leq \delta_0$, we select c_0 into the population; otherwise if x falls in the range $(\delta_{j-1}, \delta_j]$, we select c_j into the population. This selection step is repeated p times to generate the next population. The selections are independent, and some chromosomes that have a greater PMF may get selected multiple times.

After a new population has been generated, we apply two recombination operators to further its evolution: *crossover* and *mutation*. Crossover picks each chromosome for into the crossover set with a probability χ, thus giving a crossover set size of $p \times \chi$. Chromosomes in this set are randomly paired to form "couples" for crossover; if the crossover set has an odd number of chromosomes, the last added chromosome is dropped. During crossover between a pair of chromosomes $c_i = (q_0, q_1, ..., q_m, q_{m+1}, ..., q_{n-1})$ and $c_j = (q'_0, q'_1, ..., q'_m, q'_{m+1}, ..., q'_{n-1})$, a random crossover point m is selected in the range $[0..n-1]$. Then, the crossover results in the new chromosomes $\widehat{c_i} = (q'_0, q'_1, ..., q'_m, q_{m+1}, ..., q_{n-1})$ and $\widehat{c_j} = (q_0, q_1, ..., q_m, q'_{m+1}, ..., q'_{n-1})$.

The mutation operation helps jump (out of local minimas) to regions of the solution space which may not have been searched before. The probability for mutation μ decides whether a query q_i in a chromosome will change its resource placement value or not, and if it mutates, the new value for the query becomes a random integer in the range $[0..(|\mathbb{R}| - 1)]$. We expect $\mu \times (n \times p)$ number of queries to change their resource mapping values in each population.

For every generation, we repeat the steps of roulette wheel population selection from the previous generation's population; crossover to generate new chromosomes; mutation of these chromosomes; and potentially updating the best-fit chromosome based on the fitness values for chromosomes in this evolved population. The above operations are repeated for g generations, where this value may either be a predefined constant or based on some convergence function.

Our optimization problem requires us to enforce constraints. There are three approaches to doing this – remove solutions which violate constraints in each generation, give a penalty to the fitness value of the violative chromosome, or have an encoder-decoder scheme so that invalid solutions do not occur in the first place [36]. The first approach can constrain the search space of GA and cause the population to die out, while the third approach is difficult, sometimes impossible, to formulate cleanly and also increases time complexity. Instead, we use a high-valued penalty function on invalid solutions to reduce the chance that they will be selected and will cause them to eventually be pushed out. It also holds the possibility that the generation having such invalid solutions can still evolve to a valid one. We add a penalty value of $log(1 + \gamma \times F(j))$ to each chromosome c_j, for each constraint that is violated, where γ is a large constant.

1) Complexity Analysis for GA: In each generation of the GA, we need to find the objective value for all the chromosomes present in the population. This means evaluating the critical path for the DAG based on each mapping solution (chromosome) present in the population. Since we have g generations, a population size of p chromosomes in each generation, and the time to find the longest path in the DAG is $\mathcal{O}(|\mathbb{V}| + |\mathbb{E}|)$ for each candidate solution, the asymptotic time complexity of the GA approach is:

$$\mathcal{O}(g \times p \times (|\mathbb{V}| + |\mathbb{E}|))$$

VI. Micro-benchmarks on Resource Usage for Event Analytics

We perform a series of micro-benchmark experiments to measure the *latency*, λ_k^i, for a query v_i running on resource r_k, and its *energy consumption*, ϵ_i^k. In addition, we also measure *network latency* $l_{m,n}$ and *bandwidth* $\beta_{m,n}$ between pairs of resources r_m, r_n. These offer real-world characteristics of the computing resources consumed by the edge and Cloud resources, their network characteristics and the energy usage of edge devices for event analytics. These measurements are in fact useful in themselves for IoT deployment studies, and also inform the design of our simulation study of the proposed query placement optimization, presented in the next section.

A. Experimental Setup

We run experiments with different configurations of individual CEP queries, on edge and Cloud resources. We use the popular Raspberry Pi 2 Model B v1.1 as our edge device, and a Standard D2 VM in Microsoft' Azure' Southeast Asia data center as our Infrastructure-as-a-Service public Cloud. The Pi has a 900MHz quad-core ARM Cortex-A7 CPU and 1GB RAM, while the Azure D2 VM has a 2.2Ghz dual-core (4 hyper-threads) Intel Xeon E5-2660 CPU and 7GB RAM. As we see, the Azure VM has as many hyper-threads as the Pi's cores, and is rated at about twice the clockspeed. It also has seven times the physical memory. Both run Linux OS distributions, Raspbian Wheezy for Pi and Ubuntu 14.04 for Azure.

For simplicity of the experimental evaluation, all resources on the edge are assumed to be identical, and likewise the Cloud VM instances, though this does not affect the analytical model and optimization solution.

We use *WSO$_2$'s Siddhi* as our Complex Event Processing (CEP) engine [9] on both Pi and Azure. Siddhi is an open source CEP engine written in Java, and popular for IoT applications [6]. Queries are written using Siddhi APIs and compiled into executable JARs that are run on the resources. The Java installation on the Pi is Oracle JDK SE 1.8 for ARM and on Azure is OpenJDK SE 1.7.

We generate input event streams which contain synthetic integer values that represent a sensor's observation stream. Each event is up to 10 bytes of text on the wire, pre-fetched from file into memory for the experiments, and is 4 bytes in size when parsed as an integer and replayed to Siddhi's input event handler. The event values generated are configured to meet the selectivity that we need for a given query that consumes it, as discussed later in Table I.

Output patterns matched for a query are returned through a callback, where a counter is maintained for measuring the output event rate per second while the matched event itself is not considered further. A similar counter for the input rate per second is also maintained at the input stream, and by comparing these two, we arrive at the latency and throughput performance measures for each query.

We design the *query benchmark* using various configurations of the four major query types – filter, sequence, pattern and aggregate (Table I), based on the selectivity and length of patterns matched or aggregated, to give 21 different queries configurations. These are summarized in Table II.

We consider 3 configurations for *filter* queries with different selectivities: $\sigma = 0.0$ which does not match any input events, $\sigma = 0.5$ which matches about half the number of input events, and $\sigma = 1.0$ which matches all input events, as listed in rows 1-3 of Table II. For *sequence* queries, we consider two queries

[6]http://wso2.com/library/articles/2014/12/article-geo-fencing-for-iot-with-wso2-cep/

TABLE II: Summary of query configurations used in micro-benchmarks

Query ID	Selectivity (σ)	Input Event generation for required selectivity	Pattern/ Window length	Peak Rate (Pi) [e/sec]	Peak Rate (Azure) [e/sec]	Energy used by Pi [mA]
Fil 1.0	1.0	Random integer < 150	-	86,971	321,413	337.04
Fil 0.5	0.5	Random integer [0-299]	-	117,150	385,952	336.91
Fil 0.0	0.0	Random integer $<= 150$	-	204,633	529,803	337.41
Seq3 1.0	1.0	Equal integers (10,10,10,...)	3	33,958	248,896	340.91
Seq3 0.5	0.5	5 equal integers followed by a different integer (3,3,3,3,3,10,...)	3	43,470	263,332	342.45
Seq3 0.0	0.0	Non-equal integers (3,7,9,...)	3	62,347	352,617	342.62
Seq5 1.0	1.0	Equal integers (10,10,10,10,10,...)	5	22,693	190,257	341.27
Seq5 0.5	0.5	9 equal integers followed by a different integer (3,3,3,3,3,3,3,3,3,12,...)	5	31,945	224,290	342.42
Seq5 0.0	0.0	Unequal integers (3,4,7,8,9,...)	5	45,304	314,746	344.44
Pat3 1.0	1.0	Equal integers (10,10,10,...)	3	34,198	238,017	340.88
Pat3 0.5	0.5	Sequence of 3 equal and 3 random integers (3,4,3,5,3,100,...)	3	75	311	351.32
Pat3 0.0	0.0	Random integers	3	53	237	343.75
Pat5 1.0	1.0	Equal integers (10,10,10,10,10,...)	5	23,106	190,590	352.33
Pat5 0.5	0.5	Sequence of 5 equal and 5 random integers (3,4,3,5,3,6,3,10,3,11,...)	5	74	311	351.59
Pat5 0.0	0.0	Random integers	5	52	235	352.45
Agg B 60	$1/60$	Random integers	60	107,665	334,008	393.68
Agg B 600	$1/600$	Random integers	600	113,092	331,137	396.55
Agg B 6000	$1/6000$	Random integers	6,000	110,991	332,177	387.84
Agg S 60	1.0	Random integers	60	55,130	236,521	393.92
Agg S 600	1.0	Random integers	600	56,074	225,908	393.72
Agg S 6000	1.0	Random integers	6,000	52,486	230,313	393.41

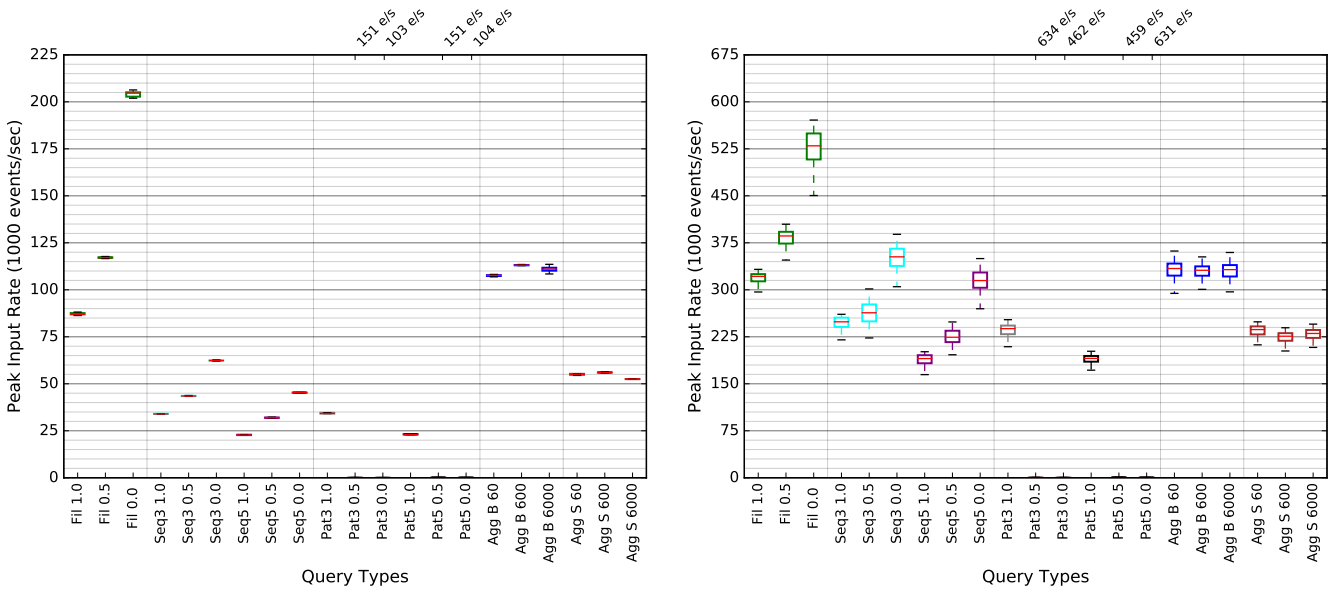
that have sequence lengths of 3 and 5, and within each have selectivities of $0.0, 0.5$ and 1.0. *Pattern* queries of lengths of 3 and 5 are considered as well, with three different selectivities each. *Aggregate* queries are designed with window widths of $60, 600$ and 6000, emulating different temporal sampling frequencies for sensors. We include both *sliding* and *batching* window variants. Details of all these query configurations are given in Table II.

These 21 queries are used to benchmark several performance parameters necessary to solve the optimization problem, and decide the placement of a query on a resource. We measure the **peak throughput rate** of a query on the Pi and Azure VM by replaying input events through Siddhi without any pause at the input. The maximum rate that is sustained is determined by the resource constraints on that device. Since the Siddhi query engine is single threaded, the inverse of the peak throughput rate for a query v_i on a resource r_k gives the expected latency per event, λ_k^i.

We measure the **energy usage** of the Pi in terms of the current drawn (milli-Ampere, mA), measured using a high precision multimeter which samples 4 current values per second. The energy usage is measured under a *base-load* condition, where no queries are running, and under a *load* condition when each query runs on it. We also measure the energy usage under different input rates for the queries.

In real life, the input event rate arriving at a query may be lower than the peak rate. Empirical results for energy used by a CEP query for different input rates helps model the optimization problem more accurately. For this, we introduce a delay between events (using `Thread.sleep()` in Java) when replaying the input stream, and benchmark for *input event rates* of $100\ e/sec$, $1000\ e/sec$, and $10,000\ e/sec$, besides the peak rate. The energy usage for Azure is not relevant here as it is not a constraint.

We measure **network latency** using the `nping` command, which is a part of the `nmap` tool available on Linux distributions. `nping` sends a 40 Bytes TCP packet (the minimum packet size allowed) from a source machine to destination machine, and the destination responds with a 44 byte TCP packet. Using this, the total round trip time (rtt) for the packet is calculated. We average `nping`'s performed for a duration of 1 min from one Pi to another, with both present in the same campus private network, and between a Pi on the campus to an Azure VM on the Cloud. We repeat these two scenarios every minute for 1 hour to get 60 average rtt values, and this allows us to capture any temporal variations in latency.



(a) Raspberry Pi 2

(b) Azure D2 VM

Fig. 1: Peak input rate for all queries on Pi and Azure (in $1000 \times \ e/sec$)

We calculate the network latency in a single direction as $\frac{1}{2} \times rtt$.

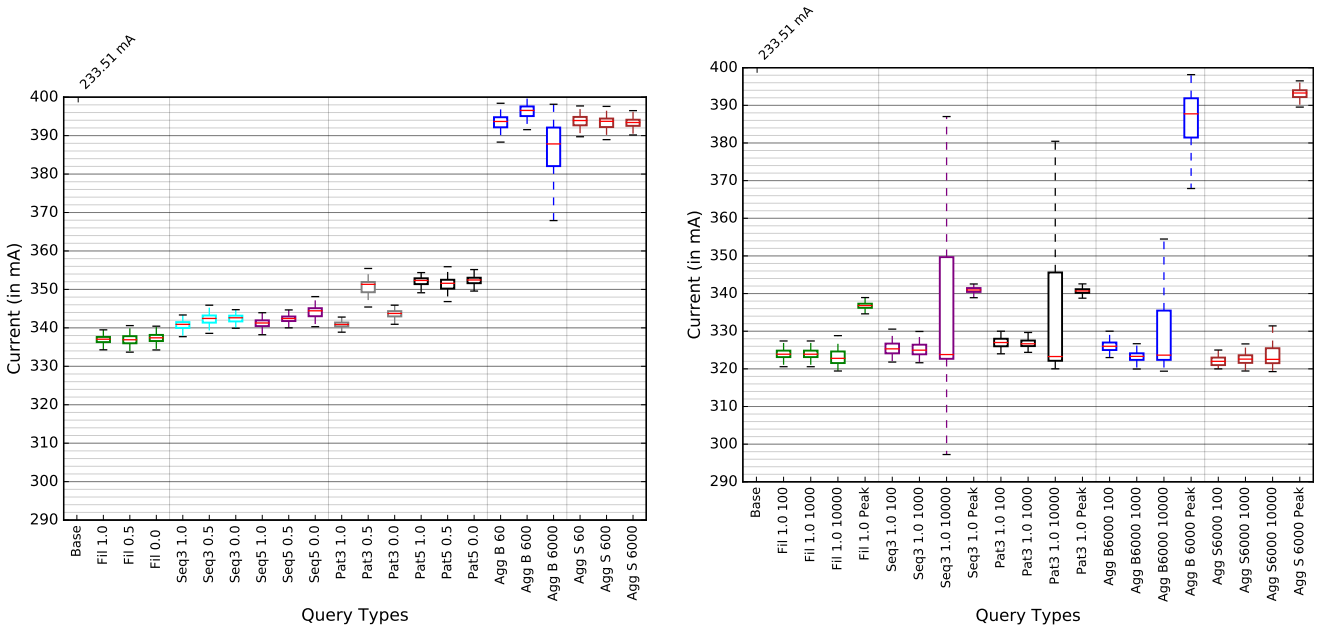
Network bandwidth is measured using the `iperf` tool available for Linux. The destination machine starts an `iperf` server which the source machine connects to as a client and downloads data of a specified size. The time taken for this transfer over multiple trials is used to calculate the average bandwidth. In our experiments, a Pi acts as the client while another Pi in the campus network and an Azure VM serve data, to measure the Pi-Pi and Pi-Azure bandwidths respectively. We repeatedly transfer a data of size 150 MBytes for 1 minute to determine the average bandwidth in that minute, and repeat this for 1 hour to get 60 measurements of network bandwidth.

B. Observations and Analysis

The peak input rate that can be sustained for different CEP queries on Pi and Azure are shown in Fig. 1 as Box and Whiskers plots. In general, we see that filter and batch aggregate queries can support a higher peak rate than the other query types, staying above $90,000\ e/sec$ and $300,000\ e/sec$ for Pi and Azure, respectively. The Pi is about $3\times$ slower than Azure in processing corresponding queries, which is understandable given their different CPU architectures (ARM AArch32 vs. Intel x86-64), and clock speeds (900 MHz vs. 2.2 GHz). The single threaded execution of Siddhi means the number of cores on each resource does not have a direct impact here.

The peak rates of filter and sequence queries are inversely correlated with their selectivity. For e.g., `Seq5 1.0` sequence query of length 5 with $\sigma = 1.0$ supports a median rate of $22,693\ e/sec$ on Pi (Fig. 1a) while `Seq5 0.0`, which is has the same length but lower selectivity at $\sigma = 0.0$, supports a higher rate of $45,304\ e/sec$. A similar behavior is seen for the Azure VM as well, with the corresponding rates for these two queries being $190,257\ e/sec$ and $314,746\ e/sec$, respectively, in Fig. 1b. As the selectivity reduces, fewer output event objects have to be generated and fewer memory states maintained, which allows more input events to be processed.

However, this phenomenon does not hold good for pattern queries and the peak rate supported actually decreases sharply as the selectivity decreases, from $23,106\ e/sec$ for `Pat5 1.0` on the Pi to barely $104\ e/sec$ for `Pat5 0.0`. Since pattern queries allow other events to occur between successive matching events, this means that as fewer events match, Siddhi has to maintain more partially matched states in



(a) All queries, at Peak Input Rate

(b) Subset of queries, at various input rates

Fig. 2: Energy usage (Current, mA) for different queries running on Raspberry Pi

memory, and test each state with every future event to verify if the full pattern matches. This increases the resource usage and hence lowers the throughput. Our experiments show that the longer a pattern query with low selectivity runs for, the lower its throughput becomes, with a long tail.

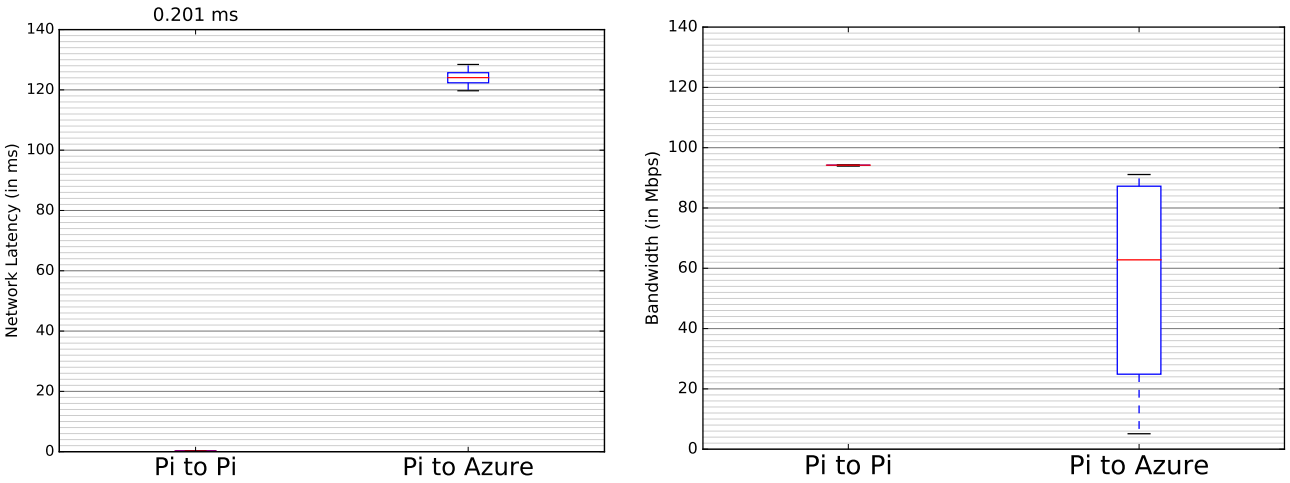
We also observe that for sequence and pattern queries, the peak rate goes down with the increase in the pattern length that is matched, e.g., on the Pi going from a median rate of $33,958$ e/sec for `Seq3 1.0`, which matches 3 consecutive events, to $22,693$ e/sec for `Seq5 1.0`, which matches 5 events. This is understandable – as the match size increases, more in-memory states have to be maintained to match subsequent events against.

In case of aggregate queries, batch windows support a higher peak rate than sliding windows since the number of windows that the latter processes is much higher than the former, and their selectivities are also very different. For e.g., the batch aggregate query rate is double that of the sliding aggregate query on the Pi, through it is less pronounced on Azure. However, the peak rate supported does not change as we increase the window width, for both batch and sliding windows. Even though the window size grows, the number of aggregation operations performed (e.g., sum and division for average) remain almost the same, with the number of output events generated being the only other change. These do not impact the resource usage as much.

All the above relative trends are consistent for both Pi and Azure, shown in Figures 1a and 1b, assuring us that these are characteristics of the query and not the device.

We report the *energy used by the Pi* for each input event for different queries at the *peak rate*, given in terms of current (in mA) drawn in Figure 2a. Multiplying this value by the duration for which the query runs helps map to the energy capacity of the battery used to power the Pi, given in mA·h. The *base load* current drawn by a freshly booted Pi is about 233 mA, shown to the left in the figure. In general, we do not see a significant difference in the energy used by filter, sequence and pattern queries, largely falling between $336 - 353$ mA. Different aggregate queries have comparable energy levels as well, with their boxes falling between $382 - 398$ mA, though they are higher than filter due to the floating-point operations required for aggregation.

The energy consumption for query types with a selectivity of 1.0 but for *different input event rates* of



(a) Latency (b) Bandwidth

Fig. 3: Latency and Bandwidth between two Pi's, and between Pi and Azure VM

$100\ e/sec$, $1000\ e/sec$, $10,000\ e/sec$ and their peak rate is shown in Figure 2b. We see that the median current drawn by the Pi is relatively the same for all rates, other than the peak, which has a higher consumption. For e.g., `Fil 1.0 100`, `Fil 1.0 1000` and `Fil 1.0 10000`, which is the filter query with $\sigma = 1.0$ and at rates of $100\ e/sec$, $1000\ e/sec$, $10,000\ e/sec$, all draw between $322 - 324$ mA while at the peak rate, the same query `Fil 1.0 Peak` uses a higher 338 mA. However, we do see that for input rates of $10,000\ e/sec$, several queries have many outliers which results in taller boxes between Q2 and Q3. This show the transient energy usage behavior for the Pi under certain cases.

The *network latency* between the two Pi's is quite small at $0.20\ ms$ as compared to between Pi and Azure at a median $124\ ms$, as shown in Figure 3a. This is obvious: the two Pi's are located in the same campus private network (in Bangalore, India), while the Azure VM they access is located in the closest Microsoft data center in Singapore. The many network hops required from the campus network to the data center incurs additional latency. The *network bandwidth* between two Pi's is very stable at 96 Mbps, as shown in Figure 3b, which is close to the theoretical limit of 100 Mbps supported by the Pi's network interface. In comparison, the bandwidth between Pi and Azure is both lower, at a median 64 Mbps, and has a wider variation of about $\pm 16\ Mbps$ due to higher congestion in the public network as compared to the private campus network.

VII. Simulation Study on Distributed Query Placement

We use results from these real-world micro-benchmark experiments to perform a simulation study on optimal placement of CEP queries across the Pi edge and Azure Cloud VM resources. This study explores the ability of our brute-force and meta-heuristic optimization solutions to meet the constraints of energy and computing capacity of the resources, while minimizing the end-to-end latency for event processing between edge and Cloud, for a user-specified dataflow graph of CEP queries.

A. Experimental Setup

1) DAG Generation and Static Characteristics: Our evaluation considers a broad collection of synthetically generated DAGs composed out of the CEP queries introduced and benchmarked in the previous section. We use the *Random Task and Resource Graph (RTRG) tool* [37], developed at the University of Southampton in the context of embedded systems research, to generate dataflows with different numbers of CEP queries (vertices) present in them. The tool is configured to generate DAGs with vertices that have a maximum vertex out-degree of between $1 - 5$ edges. We then sample from among the benchmarked

CEP queries to map onto each vertex in the DAG. Specifically, we randomly map a vertex to one of the query types, filter, sequence, pattern, batch aggregate or sliding aggregate, with equal probability. Next, we uniformly select a variant of this query type from Table II [7]. This ensures that we have coverage, both with respect to query types and selectivities, in each DAG. We generate DAGs of smaller sizes with $4-12$ queries, and larger sizes of $20-50$ vertices, to give 9 sizes.

Further, an application performing event analytics is likely to consume streams generated from not just one edge gateway device but several. The number of source vertices ($\mathbb{V}^{SRC}$) has an impact both on the downstream rate and the selectivity of the DAG. As defined earlier, vertices that consume the DAG's input streams are required to be present on the edge. So the number of sources in the DAG also impacts the number of queries placed on the edge. We consider two variants of each DAG size: one with a single input event stream, and for DAGs of size 10 and larger, another with four input streams. We do not impose any such restrictions on the number of sink queries in the DAG ($\mathbb{V}^{SNK}$) that need to be placed on the Cloud as it does not impact the output rate or selectivity.

In order to avoid local effects of the random DAG generator tool, we generate 3 DAGs for each of these sizes and source vertex counts to give us a total of 45 DAGs, as listed in Table III. As can be noted, we see a fair coverage of the different types of filter, sequence, pattern, batch aggregate and sliding aggregate queries in each DAG. The selectivity $\sigma(v_i)$ of each query v_i in the DAG is used to generate the *overall selectivity for the DAG* recursively, and this is listed as well. This, when combined with the input rate to the DAG, determines the *output rate of the DAG*, shown in the last column of Table III for a sample $1000\ e/sec$ input rate. As we see, the selectivities of the DAGs have a wide range from $\sigma = 0.04$ for DAG ID `40_1_1` to $\sigma = 458.28$ for `20_4_1`, depending on the queries mapped to their vertices [8].

We also indicate the expected output rate for a sample input rate of $1000\ e/sec$, with the DAG's output rates ranging from $20-114,000\ e/sec$ [9]. Depend on the random generation, some DAGs may have one query that is particularly bottlenecked due to multiple input edges or a high input rate caused by upstream queries and their selectivities. We identify the *max query* as the one with the highest relative input rate in the DAG, and as an aid to the analysis, we list its *input selectivity* and *peak input rate*. Here, we see that in some cases, the rate processed by the max query, using an input rate of $1000\ e/sec$ for the DAG, goes as high as $250,000\ e/sec$ for the query `50_4_3` even as the DAG's output rate is only $25,000\ e/sec$.

2) Dynamic Characteristics of Generated DAGs: The static characteristics of the DAG are those that we expect users to provide, and remain invariant over time for a given event stream distribution which determines the queries' selectivities. However, there are dynamic runtime characteristics of the edge and Cloud resources, and their network connectivity, that we consider in our study to more accurately model real-world behavior. Specifically, for each DAG, we are interested in the values of the *latency* (λ_i^k) of each query v_i in the DAG running on each resource r_k; the *network latency and bandwidth* (l, β) for each outgoing edge from a vertex, be they from edge-edge, edge-Cloud or Cloud-edge; and the *energy usage* (ϵ_i^k) on the edge device for each query, at the input rate it is processing [10]. These parameters are required as inputs when solving the optimization problem.

We use results from our benchmarks in Sec. VI to provide these values. For simplicity, we assume that all edge devices are of the same type, Raspberry Pi 2 Model B, and the Cloud VM is an Azure Standard D2 instance, which were used in our benchmarks. However, even with the same resource type, there may be temporal or device specific variations in the parameter values. In order to simulate the runtime

[7]For these experiments, we consider all queries in Table II except **Pat3 0.5, Pat3 0.0, Pat5 0.5** and **Pat5 0.0**. These four have a sharply lower peak throughput rate compared to the other queries, and their inclusion makes it difficult to automatically generate synthetic DAGs having a feasible solution. This gives us 17 queries in all as candidates to map onto vertices in the DAG.

[8]In case of multiple sinks in the DAG, the output rate is determined as the sum of the output rates from all of them, and the selectivity of the DAG reflects this as well.

[9]When multiple source input streams are present, the input rate to the DAG is evenly divided between the streams. So an input rate of 1000 events/sec to a DAG with 4 input queries will effectively pass 250 events/sec through each input query.

[10]In the study, for simplicity, we assume that all edge devices have the same computing capacity and network behavior, and similarly with the Cloud VMs. That said, the analytical models does consider edge and Cloud resources of different capabilities, and the study design itself would be identical even with multiple edge resource types.

TABLE III: Configuration of DAGs used in simulation study. Counts of source and sink vertices, and different query types present in each DAG is listed. Selectivity and rates for the DAG and the query in the DAG with the maximum input rate are also shown.

DAG ID[1]	Sources	Sinks	Filter	Seq.	Pattern	Agg B	Agg S	Max Qry I/P σ	DAG σ	Max Qry I/P Rate[2]	DAG O/P Rate[2]
4_1_1	1	1	0	1	2	0	0	1.50	1.50	1,500	1,500
4_1_2	1	2	1	0	1	1	0	1.00	1.00	1,000	1,000
4_1_3	1	1	1	1	0	0	1	2.00	2.00	2,000	2,000
6_1_1	1	3	1	2	1	1	0	6.00	6.00	6,000	6,000
6_1_2	1	3	1	2	1	1	0	2.50	0.06	2,500	60
6_1_3	1	3	1	1	3	0	0	9.50	6.00	9,500	6,000
8_1_1	1	2	1	3	2	0	1	2.00	2.00	2,000	2,000
8_1_2	1	2	0	3	2	1	1	12.0	12.0	12,010	12,010
8_1_3	1	1	0	1	2	3	1	1.00	1.00	1,000	1,000
10_1_1	1	2	4	0	4	0	1	40.5	40.5	40,500	40,500
10_1_2	1	1	0	3	3	1	2	4.13	2.10	4,130	2,100
10_1_3	1	2	0	4	3	0	2	56.0	56.0	56,000	56,000
10_4_1	4	3	1	1	1	2	1	18.9	18.9	4,720	4,720
10_4_2	4	2	1	4	0	0	1	21.7	12.0	5,430	3,000
10_4_3	4	1	0	1	3	1	1	116	116	29,000	29,000
12_1_1	1	2	4	3	3	0	1	41.6	41.6	41,600	41,600
12_1_2	1	2	2	2	3	3	1	1.55	0.01	1,550	10
12_1_3	1	3	2	0	5	2	2	29.0	0.01	29,000	10
12_4_1	4	1	2	1	3	1	1	42.3	21.1	10,570	5,280
12_4_2	4	2	1	5	0	0	2	49.0	14.5	12,250	3,630
12_4_3	4	2	1	3	3	1	0	56.0	56.0	14,000	14,000
20_1_1	1	2	1	6	6	2	4	4.11	2.95	4,110	2,950
20_1_2	1	3	2	5	5	5	2	14.8	14.8	14,780	14,780
20_1_3	1	2	3	3	7	5	1	6.13	6.13	6,130	6,130
20_4_1	4	2	1	2	7	2	4	458	458	114,570	114,570
20_4_2	4	2	0	5	7	2	2	78.0	15.7	19,500	3,930
20_4_3	4	1	3	4	2	3	4	186	62.8	46,510	15,700
30_1_1	1	1	1	6	10	7	5	84.3	1.68	84,260	1,680
30_1_2	1	1	3	11	10	3	2	7.55	7.55	7,550	7,550
30_1_3	1	2	4	10	8	2	5	2.00	0.72	2,000	720
30_4_1	4	1	3	5	8	6	4	40.7	0.20	10,180	50
30_4_2	4	2	2	11	4	5	4	16.0	0.80	4,000	20
30_4_3	4	1	4	8	8	4	2	155	16.0	38,760	4,160
40_1_1	1	2	2	12	13	8	4	9.95	0.04	9,950	40
40_1_2	1	1	5	11	9	6	8	151	76.6	150,690	76,600
40_1_3	1	1	5	8	8	5	13	24.4	0.32	24,410	320
40_4_1	4	2	4	15	9	4	4	104	4.44	26,000	1,110
40_4_2	4	2	1	7	14	9	5	16.0	6.44	4,000	1,610
40_4_3	4	1	3	8	11	8	6	875	2.44	218,840	610
50_1_1	1	1	10	14	10	4	11	122	72.9	121,580	72,950
50_1_2	1	1	6	15	15	4	9	64.7	64.7	64,670	64,670
50_1_3	1	2	7	17	12	4	9	6.00	1.03	6,000	1,030
50_4_1	4	2	3	14	19	5	5	48.5	0.12	12,120	30
50_4_2	4	3	3	11	15	12	5	305	0.44	76,320	110
50_4_3	4	2	9	13	10	5	9	1,003	102	250,780	25,600

[1] *The 1^{st} number in the DAG ID is the number of vertices, 2^{nd} is the number of source vertices, and 3^{rd} is a count for the 3 versions.*
[2] *Based on a DAG input rate of* $1000\ e/sec$

variability of these parameters, we use a simple sampling technique on the box plot distribution that we have measured and shown in Section VI-B.

Our *sampling technique* is based on the three quartiles – Q1, Q2 (median) and Q3 – representing the boxes. We first pick one of the two quartile ranges between Q1–Q2 or Q2–Q3 with equal chance, and then with a uniform probability, select a numerical value that falls in the inter-quartile ranges between $(Q2-Q1)$ or $(Q3-Q2)$, as applicable. This technique is simple and reproducible based on the benchmarking results we have provided. At the same time, it captures the variation in the values of some of these parameters as occasionally, such as in Figure 2a and Figure 2b, the box plots are wider indicating tangible runtime variations that we capture in this study.

For each vertex v_i in each synthetic DAG, we apply the above sampling technique on the box plots

from Sec. VI-B to determine its runtime parameters, given by the tuple $\langle \lambda_i^{pi}, \lambda_i^{azure}, \epsilon_i^{azure} \rangle$, indicating the latencies for running this query on the Pi and Azure VM, and the energy consumed when running it on the Pi, respectively. Similarly, for each edge between vertices v_i, v_j in each DAG, we sample and determine its network characteristics as the tuple $\langle l_{i,j}^{pi-pi}, l_{i,j}^{pi-azure}, \beta_{i,j}^{pi-pi}, \beta_{i,j}^{pi-azure} \rangle$, which represents the latencies between edge to edge and edge to Cloud, and the bandwidth between edge to edge and edge to Cloud.

3) Input Rates to Generated DAGs: Given the diversity in the DAGs that we consider, we need to carefully determine meaningful input rates Ω^{in} passed to the DAGs such that feasible solutions to the optimization problems are possible. As we have seen in Table III, some of the DAGs have high selectivities and the max queries have still higher input selectivities. Giving too large an input rate to the DAG may overwhelm the capacity of even an Azure VM to process such a query, and hence make it impossible to solve the optimization problem without violating the constraints. At the same time, picking too small an input rate may not reflect real-world needs.

We select two different input stream rates for our study, $100\ e/sec$ and $1000\ e/sec$ for this study. These are selected to ensure that a sufficient number of queries can feasibly run on the edge devices, without forcing all queries to run on the Cloud. These two rates were chosen such that 90% and 95% of all queries (respectively, for $100\ e/sec$ and $1000\ e/sec$) present in the set of synthetic DAGs we generate will receive an input rate that is smaller than the first quartile input rate supported for that query type on the Pi. Or, in other words, with an input rate of $100\ e/sec$, no more than 10% of queries in a DAG, on average, will be unable to run on any edge device due to violation of the throughput constraint.

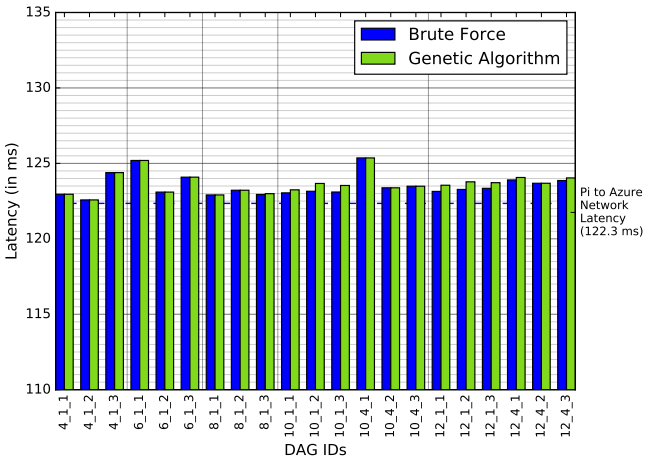
Further, in order to eliminate infeasible solutions or trivial DAGs even with these two rates, we ensure that each synthetic DAG meets the following two safety tests, and otherwise regenerate the DAGs. First, we make sure that a DAG has no queries such that its effective input rate for a $1000\ e/sec$ input to the DAG will be greater than the third quartile input rate supported by the Azure VM for this query type. This ensures that it is not practically impossible to place the query on even the VM and yet overwhelm its throughput capability. The second test eliminates trivial DAGs whose effective selectivity is zero, meaning no output events will generated at the sinks.

4) Edge Resources in Deployment: As the number of queries in a DAG increases, the resources required to support the queries, whether on edge or Cloud, will increase. As stated earlier, the sink queries need to run in the Cloud since the decision making logic for the IoT application lies there, so we are assured of having one VM resource in the Cloud. However, Cloud resources have a pay-as-you-go model, and each additional VM provisioned will have a monetary cost. So, for this study, we limit the number of Cloud resources to a single Azure VM. On the other hand, an IoT deployment on the field may have tens if not hundreds of captive (i.e., "free") gateway edge devices supporting thousands of sensors.

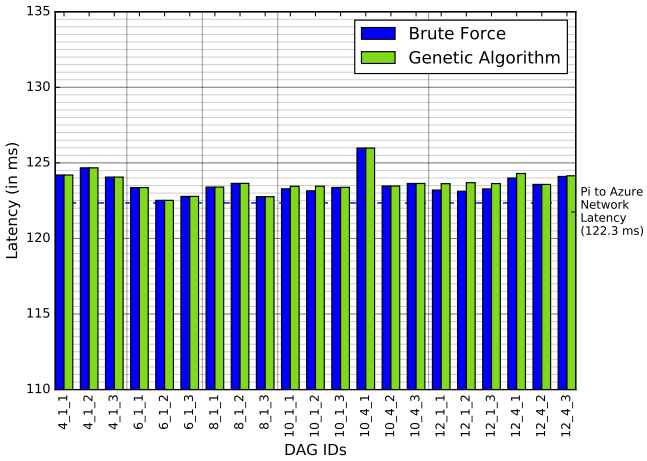
We consider two different scenarios for the number of edge devices that are available to plan the placement of the DAG queries. In a *liberal* setup, the number of edge devices plus the single Azure VM equals the number of queries, i.e., $|\mathbb{R}_E| + |\mathbb{R}_C| = |\mathbb{V}|$, or $|\mathbb{R}_E| = |\mathbb{V}| - 1$. This gives a high degree of confidence in finding a feasible solution to the optimization problem as each edge device can always run an independent query and the Cloud VM runs one as well. In a *conservative* setup, we have one Cloud VM and the number of edge devices is half the number of queries in the DAG, i.e., $|\mathbb{R}_E| = \frac{|\mathbb{V}|}{2}$ and $|\mathbb{R}_C| = 1$. Here, more than one query may have to run on a single edge device, or more of them will need to be co-located in the same Azure VM.

We assume a $24\ hour$ battery recharge cycle for the edge device, i.e., from Sec. IV-B, $\tau_k = 24$ for the edge resource r_k. From Figure 2a it can be seen that mean of the median current draw by the Pi at the peak input rate across all the queries is $358\ mA$. So, in $24\ hours$, the Pi when continuously executing a single query at an average rate will consume approximately $8,600\ mAh$. We use this as the battery capacity parameter $C_k = 8600$ in our study.

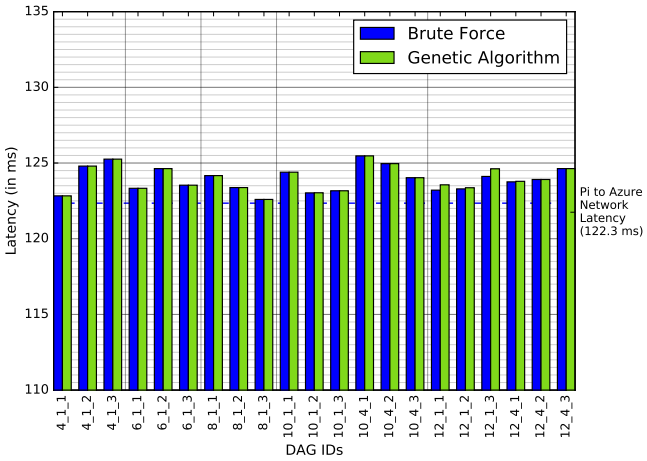
5) Brute Force and Genetic Algorithm Configuration: Both BF and GA algorithms are implemented using `C++`. All experiments to solve the optimization problem are run on a server with AMD Opteron 6376 CPU with 32 cores rated at 2.3GHz, having 128GB of RAM and running CentOS 7. We configure



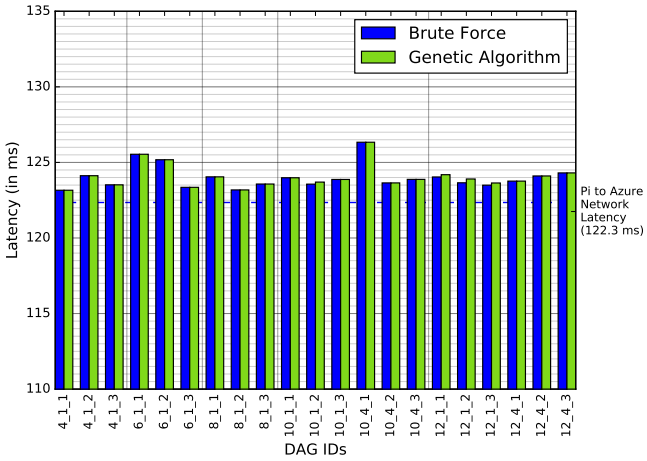
(a) Input Rate=100 e/sec, *Liberal* resources



(b) Input Rate=1000 e/sec, *Liberal* resources



(c) Input Rate=100 e/sec, *Conservative* resources



(d) Input Rate=1000 e/sec, *Conservative* resources

Fig. 4: Comparison of end-to-end latency given by GA and BF solutions for each DAG.

the GA parameters with population size $p = 50$, crossover probability $\chi = 0.50$, and mutation probability $\mu = 0.15$. Rather than fix a static number of generations for g, we instead use a logic to test in the solution has converged as follows: After running the GA for $15,000$ generations, we start checking after each generation if the best fitness value has not changed for the last 50% of the generations. Thus we conservatively run the GA for at least $15,000$ generations to avoid local convergence effects, and have an upper bound of $1,000,000$ generations.

B. Observations and Analysis

Here, we analyze the qualitative performance of the brute-force optimal algorithm (BF) and the Genetic algorithm (GA) meta-heuristic, and also the time complexity of the algorithms that determine their feasibility for use in practical IoT deployments.

1) Latencies of the Solutions: The objective function of our optimization problem is to reduce the end-to-end latency for processing events between the source and the sink queries in a DAG, while meeting the constraints. Here, we evaluate the effectiveness of the GA algorithm in offering a good solution with low latency, and compare its qualitative performance with the BF and a baseline algorithm.

Comparing GA with BF. The BF algorithm gives the theoretical best solution by performing a combinatorial sweep of all possible placements of queries on resources. Figs. 4 show the latency values

of the solutions provided by GA and BF for each DAG in our study. The four plots are for the *liberal* and *conservative* edge resource availability , and for event rates of $100\ e/sec$ and $1000\ e/sec$. Note that these plots only show results for DAG sizes up to 12 queries since the BF algorithm took longer than 10 hours to complete for larger DAG sizes. GA solutions for all DAG sizes are discussed later.

We see that GA performs relatively well in giving a solution that is close to BF's optimal. When the input rate is $100\ e/sec$ for DAGs with liberal edge resources, we see that GA converges to near-optimal values for DAG sizes up to 8 queries, but for 10 queries, the solution is marginally higher than optimal. With half the number of edge resources in the conservative setup, GA reaches the optimal or near-optimal solution in all cases. This is because reducing the number of resources results in a reduction in the search space for GA. In all, of the 44 DAG instances (21 DAGs with 2 different rates) in the conservative setup, GA gives the exact optimal solution for all but 8 DAGs.

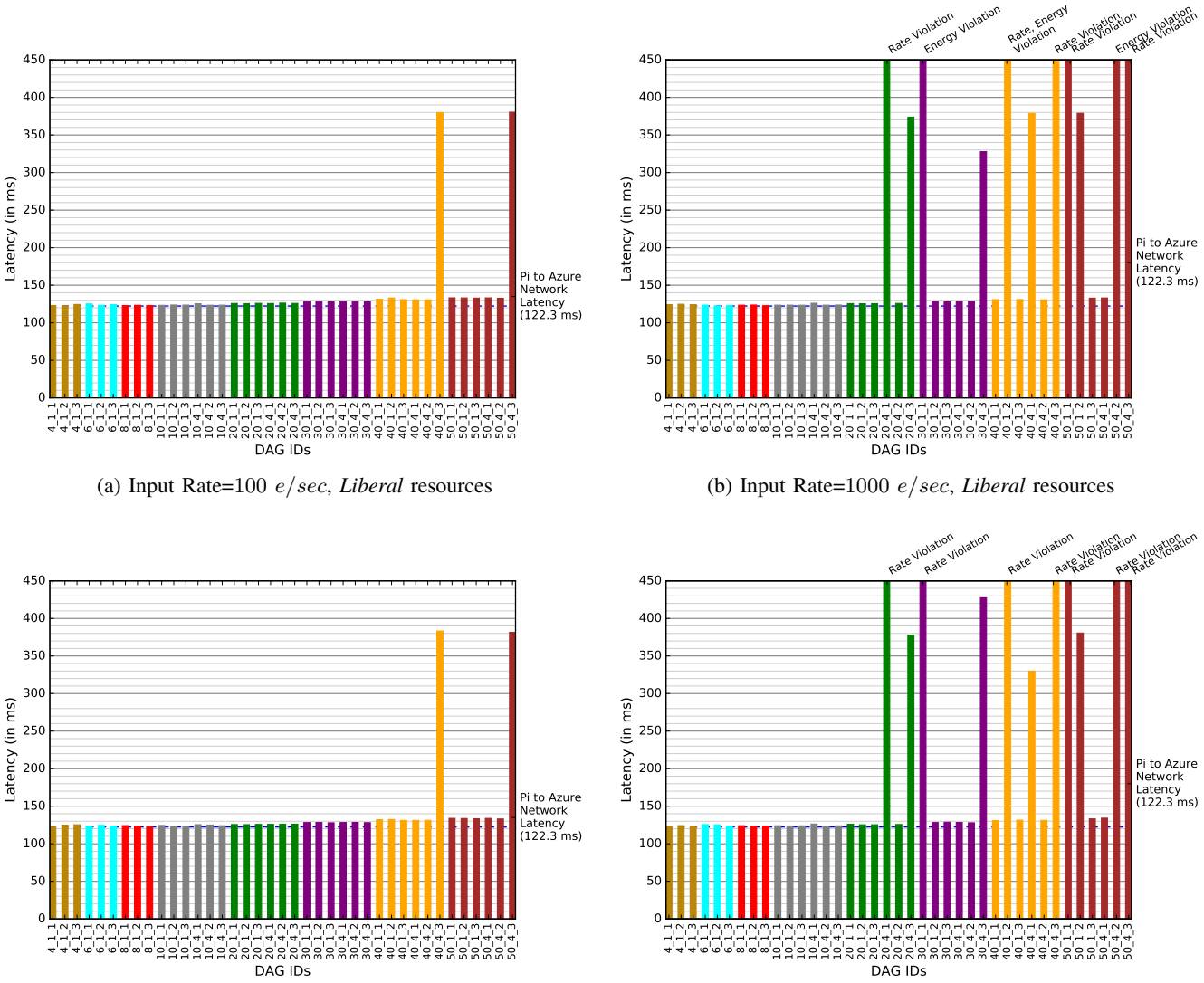
It should be noted that a dominating factor in the end-to-end latency is the network transfer time, specifically the network latency. Our micro-benchmarks show that first quartile network latency between Pi and Azure is $122.3\ ms$, and this is shown in a dotted horizontal line in Figs. 4. Since we constrain the event input stream to be on the edge and the sink to be on the Cloud, this Q1 network latency is a minimum cost paid by every placement solution, with some even having higher network latency costs based on the Pi to Azure box plot in Fig. 3a. That said, depending on the Cloud data center chosen, this network latency can be as low as $25 - 80\ ms$ based on existing literature [38], [39], and in such cases, the benefits of getting a close to optimal solution will be more significant.

GA Solution on Large DAGs. GA solutions were found for 39 DAGs that were proposed in the setup [11], despite BF being intractable for those DAGs with > 12 queries. Figs. 5 show the latency values for these solutions for each DAG, using the two input rates of $100\ e/sec$ and $1000\ e/sec$, and the liberal and conservative edge allocation. We see that the end-to-end latency in most experiments remain below $140\ ms$, and there is a small increase in the latencies as the DAG size grows. This is due to the longer critical path through the DAG as it passes through more numbers of queries between the source and sink. There are also small variations in the latencies for DAGs with similar configurations (shown in the figures with similar shading) due to the randomization during DAG generation. However, there are two outlier scenarios we observe.

One, when the input rate is $100\ e/sec$ for the liberal and conservative resource availability, shown in Figs. 5a and 5c, cases corresponding to `40_4_3` and `50_4_3` have GA solutions that converges to a value about $370\ ms$. This happens when the placement solution is such that that successive queries are placed on the edge followed by Cloud, and back to the edge and back to the Cloud. This causes a ping-pong effect due to which the network latency is paid 4 times. Such cases arise when the solution space considered by the GA results in many input rate constraint violations on the edge due to high input rate selectivity for some queries in those DAGs. For e.g., Table III shows that `40_4_3` and `50_4_3` have queries whose peak input rates are greater than $21,800\ e/sec$, and it turns out there are several such sequence queries with such a high max input rates in those DAGs. As such, several of these may be getting pushed to the VM, which further reduces its capacity and moves other queries with lower throughputs back to the edge, as a result, causing multiple round trips between edge and Cloud.

Such high input rates at some queries exacerbates the problem for DAG input rates of $1000\ e/sec$, shown in Figs. 5b and 5d, where the second set of issues crop up. Here, for DAGs such as `20_4_1`, `30_1_1`, `40_1_2`, `40_4_3`, `50_1_1`, `50_4_2` and `50_4_3`, we see that the solution that GA converges to is invalid due to energy and/or throughput rate constraint violations. The latency value (fitness function) that the GA converges to is in the order of tens of seconds (truncated in the plots), which reflects the penalty function applied by the GA for invalid solutions that violate constraints. In the absence of a BF solution to these, it is not possible to determine if the GA is unable to find a valid solution, or if a valid solution does not even exists.

[11]While 44 DAGs are evaluated for GA, we omit DAGs with 12 vertices in these plots due to limited space. These DAGs were listed in the BF vs. GA plots.



(a) Input Rate=100 e/sec, *Liberal* resources

(b) Input Rate=1000 e/sec, *Liberal* resources

(c) Input Rate=100 e/sec, *Conservative* resources

(d) Input Rate=1000 e/sec, *Conservative* resources

Fig. 5: Comparison of end-to-end latency given by GA solution for all DAG sizes.

2) Comparative Quality of Solutions: In addition to the baseline optimal BF solution, it is worth considering the relative merits of GA with respect to a baseline "sub-optimal" solution. We define a naïve, iterative *random placement* algorithm (RND) that places queries randomly on any available edge or Cloud resources in each trial solution, and calculates the latency for that trial. If a solution is valid (i.e., does not violate constraints) and has a lower latency than a previous best trial, the best solution is updated to the current solution. This is repeated $15,000$ times, which is equal to the median iterations required by GA to converge to a solution.

To offer a cumulative measure of the relative performance of latency, we define *percentage latency deviation* of a "worse" solution over another "better" solution for a set of n DAGs as:

$$\mathcal{E}_{b\to w} = \frac{\sum_{i=1}^{n}(L'_i - L_i)}{n \times \overline{L_i}} \times 100\%$$

where L'_i is the latency for DAG i given by the "worse" solution such as GA, L_i is the latency value for the DAG i by the "better" solution such as BF, and $\overline{L_i}$ is the average of the latencies for the better

TABLE IV: Comparison of the quality of the different placement algorithms

In Rate *(e/sec)*	**Resources**	**Latency Deviation %** $\mathcal{E}_{BF\rightarrow GA}$	$\mathcal{E}_{BF\rightarrow RND}$	$\mathcal{E}_{GA\rightarrow RND}$	**Invalid%** *GA*	*RND*	**Avg. Edge Use%** *RND*	*GA*	*RND*
100	liberal	0.065	0.146	18.414	0.00	0.00	56.39	63.64	65.52
1000	liberal	0.026	0.154	0.367	17.95	23.08	56.61	63.57	66.70
100	conserv.	0.000	0.048	9.288	0.00	0.00	82.00	85.53	86.98
1000	conserv.	0.007	0.045	0.304	17.95	23.08	82.22	86.14	87.89

solution. Here, we consider only DAGs where solutions from both algorithms are valid. Smaller this value, closer the worse solution is to the better one.

A second evaluation measure we use is *percentage invalid*, which reports the fraction of DAGs for which an algorithm was not able to converge to a *valid solution*, i.e., one that meets all the constraints. Here, the lack of convergence could either mean that the heuristic algorithm under-performs, or, in the absence of an optimal solution, it may also indicate that a valid solution is not theoretically possible.

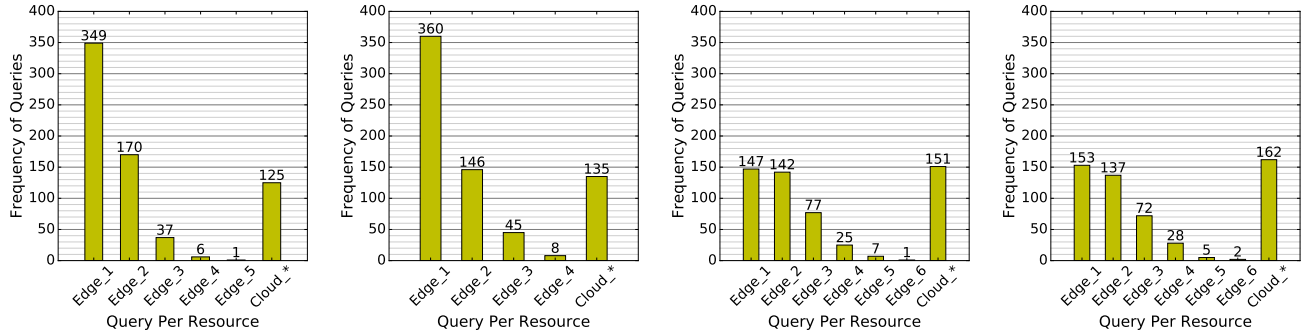
A final metric we use to evaluate the solution provided is the *percentage edge resources used*. This considers the ratio between the number of edge devices on which queries are actually placed by the algorithm compared to the number of edge devices made available. Here, a lower fraction, corresponding to using fewer Pi's, is considered more beneficial. This indicates a better utilization of the *active* devices (i.e. edges having at least 1 query on them), reduces the footprint of the DAG, and will work well for a smaller deployment of edge devices.

Table IV compares these quality metrics for the three solutions, BF optimal baseline, GA, and RND simple baseline. The latency deviation % values compare GA with BF, RND with BF and RND with GA, to see the relative pairwise performance. In each pair, only those DAGs for which valid solutions were available from both algorithms are considered. The invalid % are evaluated for all 39 DAGs (excluding DAGs of size 12) in both GA and RND. The average of the edge use % over all DAGs with valid solutions, evaluated using each algorithm, is also reported.

We observe that the latency deviation of both GA and RND for valid solutions is not far from the optimal solution, with under 1% deviation ($\mathcal{E}_{BF\rightarrow GA}$ and $\mathcal{E}_{BF\rightarrow RND}$). The GA solution does outperform RND consistently, showing a relative improvement of between 1% to 18% ($\mathcal{E}_{GA\rightarrow RND}$). We see that the improvements for input rate of $100\ e/sec$ is much better than for $1000\ e/sec$. That said, the RND solution has a larger fraction of 23% of DAGs where an invalid solution is reached, specifically for $1000\ e/sec$ input rate. On the other hand, GA converges to an invalid result only 18% of the time. So we see that when RND has no invalid solutions, their latencies are worse than GA (input rate of $100\ e/sec$), while in cases where RND gives valid latency solutions comparable to GA, it has a large fraction of invalid solutions ($1000\ e/sec$ input rate). This indicates the robustness of the GA meta-heuristic under different conditions.

The edge resource usage % for the GA placement solutions are quite close to the BF solution, and consistently (though only marginally) smaller than the RND approach. A lower value is better here. In drilling down into the results, we observe that this over usage of resources by GA happens in the liberal case where more edge resources are available. In all the conservative cases, GA takes exactly the same number of edge resources as BF. Intuitively, as the number of resources reduce, GA performs comparatively better due to the reduction in the search space. Thus, in real-world scenarios, where number of resources will be much lesser than the number of queries, we expect GA to use edge resources as prudently as BF.

Figs. 6 further show for the GA solutions, a histogram of the number of queries present in edge devices and the Cloud, aggregated across all the DAGs. These give the frequency of queries (Y axis) present in edge devices hosting 1, 2, 3, ..., etc. queries (X Axis), and the number of queries in the Cloud VM. The frequency under the curve for each plot is 954 queries, which is the total number of vertices for the 39 DAGs listed in Table III, excluding size 12. We see that for the liberal edge availability (Figs.6a and 6b, a large fraction of queries ($349 - 360$, or about 37%) are present in exclusive edge devices, hosting just that query ("Edge_1" in X Axis). About half as many queries are paired up on edge devices ("Edge_2"



(a) Input Rate=100 e/sec, *Liberal* resources (b) Input Rate=1000 e/sec, *Liberal* resources (c) Input Rate=100 e/sec, *Conservative* resources (d) Input Rate=1000 e/sec, *Conservative* resources

Fig. 6: Number of queries placed on each resource obtained from GA solutions [Total Queries = 954]

with $146 - 170$ queries), and in a few cases, 4 or 5 queries are present in the same edge. The Cloud VM hosts about 14% of all queries ("Cloud_*").

But when we halve the number of edge resources in the conservative case, the frequency of queries running on independent edge devices more than halves to $147 - 153$ while there is an increase in the fraction of edge devices hosting multiple queries in them. E.g. there are $72 - 77$ queries that are placed three at a time in a single edge in Figs. 6c and 6d's "Edge_3", compared to $37 - 45$ for the liberal case. Even the fraction of queries running in the Cloud increases to 16%. As the number of edge resources decrease, there is a natural tendency to require more queries to be packed in fewer edge resources. When queries map to the same edge resource, the maximum input rate supported by that edge for each additional query decreases, thus increasing the chance of throughput violations on the edge. This consequently pushes more of the queries to the Cloud.

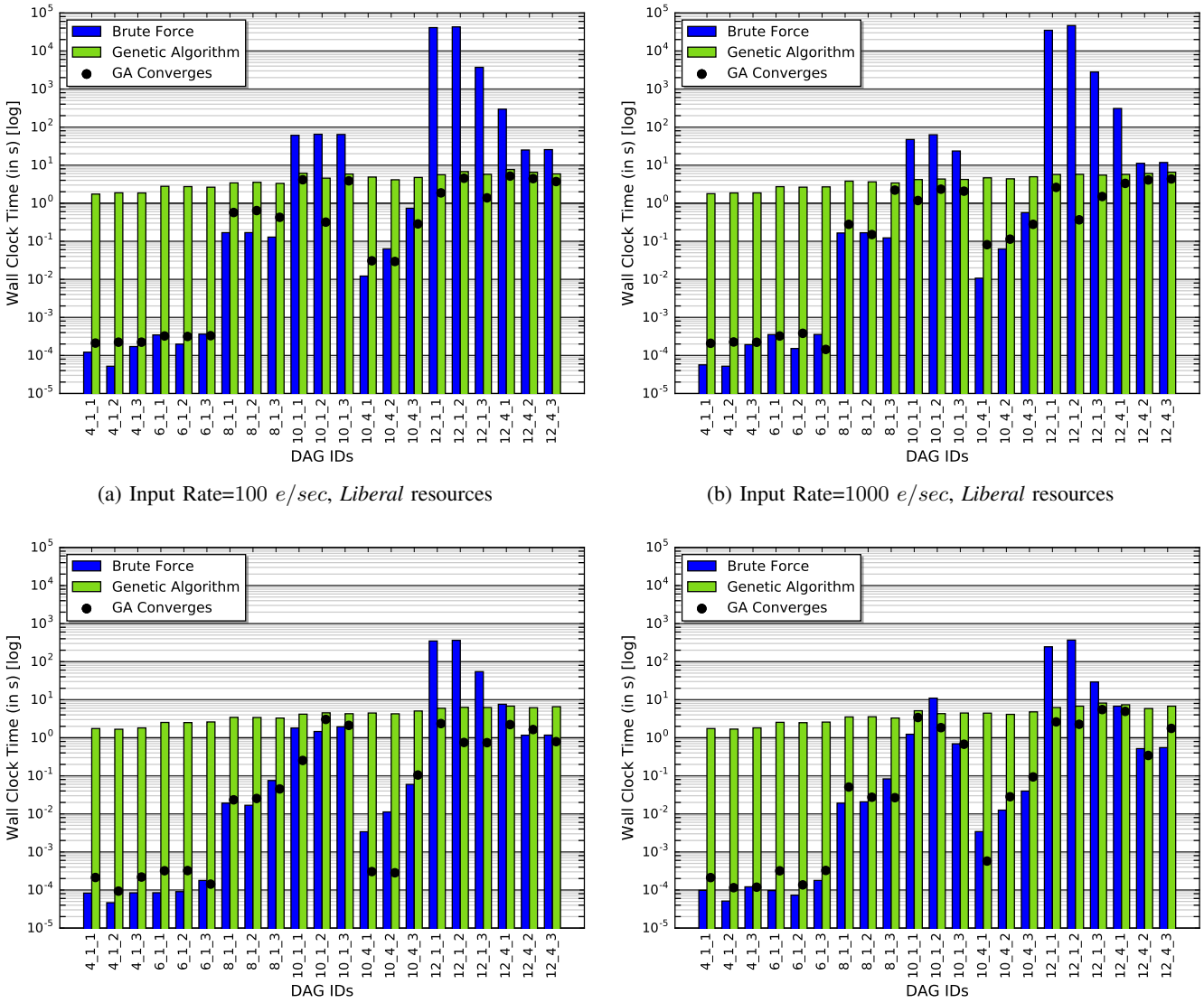
3) Time Complexity of the Solutions: Optimal scheduling is an NP-complete problem, which means the computational cost is intractable for larger problem sizes. GA being an iterative meta-heuristic does not guarantee an optimal solution, but terminates in reasonable time with a (typically) valid and near-optimal solution. Here, we complement the earlier discussions on the qualitative performance of the algorithms with an analysis of their time performance.

Wall Clock Time for BF and GA. Figs. 7 shows the wall clock time in seconds required to run GA and BF programs for the DAG sizes from $4 - 12$ queries, which was the largest we could solve using BF within 13 hours.

BF is expected to take exponentially longer time as the queries and number of resources available to place its queries increase. We do see that BF typically takes longer as the DAG sizes increase, after accounting for the number of source and sink vertices – the search space is reduced by the number of source and sink queries since they are pre-pinned to specific edge and Cloud resources, respectively. So the time taken for, say, `10_1_*` is much higher than `10_4_*` in all cases. This is seen again for DAGs with the same size and number of sources, but different number of sink vertices, caused by the random generation of DAGs. For `10_4_1`, `10_4_2` and `10_4_3`, we see from Table III that they have 3, 2 and 1 sink queries respectively, which are pre-placed in the VM. As a result, the number of queries which need to be placed has reduced between the three, leading to smaller runtime for the latter compared to the former. The time taken for conservative resources is also much lower than the liberal ones, since the search space is much smaller due to half the number of edge resources available.

BF has severe computational complexity limitations. For DAGs of size 12 with a single source, the time taken to find the optimal solution for the liberal case ranges from 46 $mins$ to 12.9 $hours$ (Figs.7a and 7b), and takes several minutes even in the conservative cases (Figs.7c and 7d). For DAGs of size 14 and larger, our BF program did not terminate in even 120 $hours$.

The GA runtime remains narrowly bounded across all the plots in Fig. 7, taking between $1 - 10$ $secs$



(a) Input Rate=100 e/sec, *Liberal* resources

(b) Input Rate=1000 e/sec, *Liberal* resources

(c) Input Rate=100 e/sec, *Conservative* resources

(d) Input Rate=1000 e/sec, *Conservative* resources

Fig. 7: Wall clock time taken by BF and GA to find a solution. The black circle is the earliest time at which the GA started to converge at the final solution but continued to run optimistically. The time is given in seconds, and the Y axis is in log scale.

to run. This is because we run the GA program for at least 15,000 iterations to conservatively ensure that we do not stop at a local minima. A fairer estimate of the wall clock time for GA is the black circles in the plots that show the start of convergence, i.e., the earliest time at which the eventual solution is first seen. As we see, many of these initial convergence times are much smaller, particularly for small DAGs, taking sub-second in most cases. That said, for DAGs of size ≤ 8, it is possible to run BF algorithm quickly, sometimes even faster than GA, and the usefulness of GA is clearly evident for DAGs with sizes ≥ 12.

Predictable Time Complexity for BF and GA. In fact, we can strengthen this analysis further. We plot this empirical data of wall clock times for various DAG experiments against the asymptotic time complexity for BF and GA introduced in Secs. V-A1 and V-B1 respectively. Recollect that the time complexity for BF is $\big((|\mathbb{V}| + |\mathbb{E}|) \times |\mathbb{R}|^n\big)$, where $n = \big(|\mathbb{V}| - (|\mathbb{V}^{SRC}| + |\mathbb{V}^{SNK}|)\big)$ are the number of non-source and non-sink queries. Similarly, the time complexity for GA is $\big(g \times p \times (|\mathbb{V}| + |\mathbb{E}|)\big)$, where

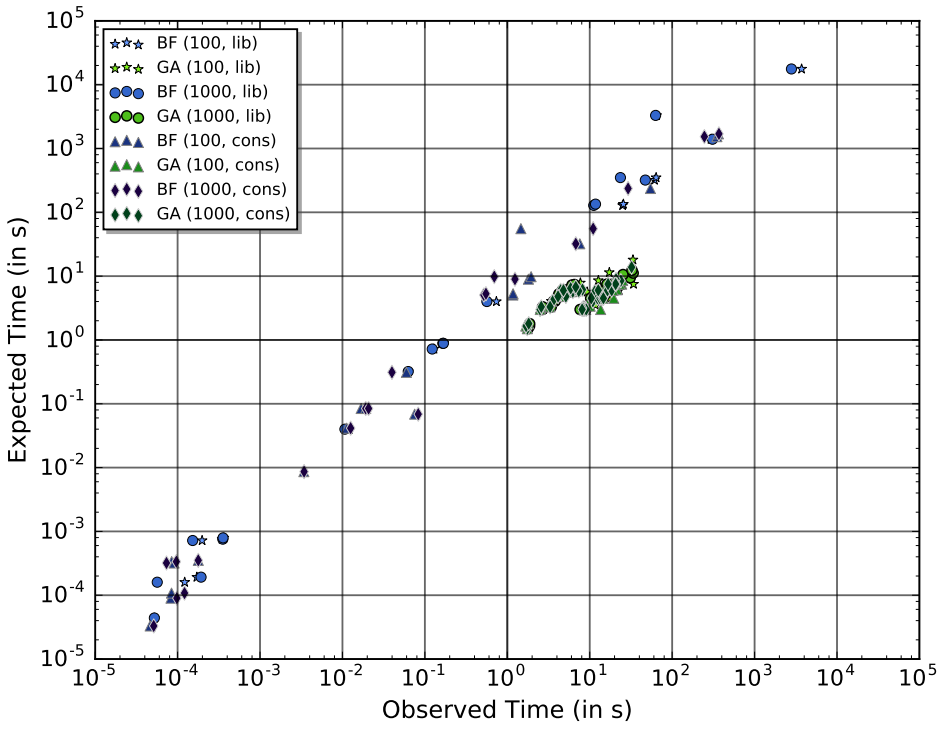


Fig. 8: GA and BF's observed wall clock time (X Axis) compared with their expected time complexity (Y Axis). Log-Log scatter plot with units in seconds.

we have set $p = 50$ and the number of generations g can range from $15,000 - 1,000,000$, and depends on the run. We multiply these complexity measures by a constant scaling factor, and draw a scatter plot in Fig. 8 between the expected algorithm time based on the complexity function, and the observed algorithm wall clock time, both for BF and GA.

In this Log-Log plot, we clearly see that there is a strong linear correlation between the expected and observed wall clock time for both the algorithms. This indicates that our complexity analysis matches with real-world behavior of these algorithms, if only off by a constant scaling factor, and thus can be extrapolated. The BF has a wide range of values, spanning from 10's of μs to 10's of thousands of seconds, for DAG sizes that just range from $4 - 12$ queries. This reflects the exponential time complexity of the algorithm. From this, we can see that a DAG with even 14 queries would expect to take about 5 days to complete for an optimal BF solution. Hence for DAGs of non-trivial sizes, and non-trivial number of resources to place them on, it is practically infeasible to find an optimal schedule using BF.

On the other hand, the GA values are tightly clustered and also show a strong correlation between expected and observed, even as these values are for DAGs sizes that range from $4 - 50$ queries. The complexity for a GA in our optimization problem is proportional to the DAG size and number of generations. For e.g., while DAGs `20_1_2` and `50_1_2` take the same number of generations to converge for 100 e/sec rate and liberal resources, the wall clock time taken for the latter is $3\times$ higher at 15 $secs$ compared to the former. We do however see that there are two clusters in the GA scatter plots, where in one cluster, the observed time has shifted higher, though the general trend holds.

Examining the plot of the number of iterations (g) taken for the GA to converge for different DAG sizes (not shown due to space limits), we see an occasional increase in the number of iterations as the number of queries in a DAG increases. But this is a very nominal growth with a small linear slope, and the g term is unlikely to dominate even for larger DAGs. In fact, the time taken to find the critical path in each iteration, which is a function of $(|\mathbb{V}| + |\mathbb{E}|)$, has a greater impact as the DAG size increases than the number of generations required to converge.

VIII. Discussion

In any emerging area that sees the confluence on multiple technologies such as Internet of Things, Cloud computing, Big Data, and mobile platforms, addressing specific research problems opens the door for more such problems and opportunities that exist. Here, we summarize our key findings, and highlight a host of future experiments and research that arise from this study.

A. Key Outcomes

There are several key takeaways from this article, beyond what was discussed in the analysis sections, that have a broader impact beyond just the immediate problem we address.

Edge and Cloud. We see that edge devices like the Pi perform $\frac{1}{3}^{rd}$ as well as similar Cloud VMs for CEP event analytics, and both have similar performance trends for different query categories. They are also cheaper than the Cloud in the long-term, if they are already deployed and available as part of IoT deployments.

However, Clouds are still useful when we consider aggregation across many streams, or from edge devices that span private networks, and when the throughput limits required are very high. The ability to have many cores in a single VM helps as well, but allowing in-memory communication between many queries present in the same VM. That said, we also see that the punitive cost in the end-to-end latency is the network latency between edge and Cloud, and its variability as well. So for highly time-sensitive applications, a much-closer data center or a private Cloud on the same network will be necessary. The bandwidth appears to be less of a concern, given the small event sizes which even cumulatively or at a high rate are tolerable.

We have used real-world, high-precision measurements of energy usage in our edge resources for a variety of event queries. This goes beyond current literature that limits itself to examining CPU, memory and network usage, which are poorer approximations of energy use. For e.g., in all cases, while the CPU and memory utilization by each query is stable at about 98% (single core) and 3%, respectively, we see two discrete energy levels for the sequence-like and aggregate queries. This shows the importance of practical validation.

Scheduling Approaches. Our benchmarks also show that the Pi consumes discrete levels of power for the different queries, that allows for predictable energy modeling. In fact, given the tight bounds of these consumption levels, we can approximate the power levels to just three categories: base load, filter-like queries, and aggregate queries.

The GA meta-heuristic has shown to be robust and scalable in solving the non-linear optimization problem. It gives optimal or near-optimal solutions, where it is possible to compare against the optimal BF; gives results with low end-to-end latency values for larger problems; has a limited number of cases where it was unable to provide feasible solution (in one was indeed possible); and can be consistently solved within seconds. With the complexity plots showing a high correlation between expected and observed runtime, the GA holds promise for providing good placement solutions for much larger IoT deployments, on the order of thousands of resources and DAGs.

That being said, for small scale IoT deployments with under ten resources and small DAGs, BF offers optimal solutions within a reasonable time and should be chosen. BF can also be complemented with techniques like Dynamic Programming to possible speed up the time, though it would not have a tangible impact on exponential time complexity. GA however offers the flexibility of trading-off runtime and solution quality. It can be limited to finding a solution within a fixed time budget, and the GA evolves for that duration and results the best solution seen that far.

Supplementary benefits. A variation of this simulation study is to estimate the least number of edge and Cloud resources required to support a certain number of streams and query workloads while meeting specific QoS required for the application. Such "What if" studies are crucial for emerging domains in IoT to better plan deployments that may take months and millions of dollars, and conserve the resources required for future workloads.

B. Future Experiments

There are additional experiments and studies that can be considered to validate the proposed solution in a more diverse environment.

1) We do not consider multi-threaded or multi-core execution in our study. Siddhi supports a limited for of multi-threading, which we did not leverage in our study to simplify the experiments, and one can always run multiple copies of Siddhi on independent cores. Given the prevalence multi-core CPUs in even edge devices, this needs to be considered. This will make edge devices like the Pi even more favorable.
2) While our benchmarks considered the most common CEP patterns, these could be complemented with a wider variety of sample queries from among these types. Using queries from real-world deployments will also make the workloads more representative. Also, while our benchmarks considered pattern queries with low selectivity, these were not used in our synthetic DAGs since they produced DAGs with very low output rates and selectivities. A special class of DAGs including such low-rate queries can be considered. In a similar vein, we should also consider a wider variety event types, with different payloads. Our prior work on benchmarking for distributed stream processing systems offers some possibilities [40].
3) Our benchmarks do not consider the energy cost for the network (LAN/WLAN) transmission. In this work current drawn by Pi is measured only for running Siddhi queries on the Pi, with events being generated locally. For a more realistic scenario, the energy cost for both the wireless and the LAN interface need to be measured, and included in the simulation study.
4) We have seen scenarios with just two types of networks, private campus and public Cloud. However, even within these networks, there is bound to be variability. Edge devices (or VMs) could be at different parts of the topology in the private network (or the data center), and the latency costs may be different due to multiple switches coming in the way. While we used a high speed uplink from campus to the public Internet, the network behavior from edge to Cloud may be different when using a home broadband or cellphone carriers. There is also a growth in the number of Cloud data centers with, for e.g., three new Azure data centers coming online in India as we are writing this article. Choosing the best/nearest data center when having a multi-city deployment, and the network costs between the data centers need attention too.
5) Even within edge devices, we have considered the Raspberry Pi 2 Model B, and there exist newer/-faster models like the Pi 3, embedded versions like the Pi Zero, and other DIY platforms like the Arduino, Intel Edison, etc. While our current work targets platforms that run Linux and Java for the analytics platforms, these concepts can also be extended to more constrained devices and recent platforms. Clouds offer different VM favors as well that could be considered.
6) It would be useful to understand the appropriate mix of the different numbers of edge and Cloud resources, e.g. more Cloud VMs, fewer edges, fixed number of edge devices for different DAGs, etc. These would offer better insight on the resource usage by solutions from the optimization solver.
7) While we have considered two input rates for our simulation study, it would be useful to observe the impact on latency, infeasible solutions generated and resource usage as we increase the rate to higher levels as well.
8) We have used real-world benchmarks to drive the simulation study. However, to offer even higher guarantees of the practical viability and relevance of our work, the placement solutions obtained from the optimization solvers should be tested with real life deployments having multiple edge devices and the Cloud. This will be yet another stepping stone toward translating research into practice for analytics across edge and Cloud for IoT.

C. Future Research

There are several promising research avenues to explore further in this emerging area of event analytics across edge and Cloud.

1) Resource usage was not a primary consideration for our problem definition, even though we reported the resource usage on the edge for different solutions. While the constraints ensured that solutions were limited to the full compute and energy capacity of a resource, our optimization goals did not consider resource usage across devices – whether to ensure the utilization of the edge was balanced (e.g. to ensure no single Pi is overloaded, and they all drain their battery at the same rate), or to ensure that the utilization on edge and/or Cloud was high for active devices/VMs (e.g. to ensure we get full value for VMs that are paid for, or to allow some inactive edge devices to be turned-off or duty cycled if others with high utilization can take the workload).
2) This article considers the problem of scheduling a single DAG on to the edge and Cloud that are fully available. However, practical situations have DAGs that may arrive periodically, or exit after a few days or weeks. There may also be analytics from multiple domains that share the same IoT fabric and devices. In such cases it is required to place multiple DAGs on to the same set of resources, or place a DAG on edge and Cloud resources that only have partial capacities available. Knowing the entry and exit schedules of the DAGs will also better inform us as to plan for future submissions or capacity availability.
3) Model VM cost into the equation. In our work VM cost has not been included, but this cost may become significant as we increase the number of VMs. This calls for conservative use of VMs and keep queries on edge as much as possible.
4) Our problem dealt with input rates that arrive at a constant rate, and this is reasonable since many sensors are deployed to generate events at a constant sampling interval. The impact of input rate variability on the optimization solution was not considered. While we observe that there is not a lot of variability in the energy usage for different event rates, it may be that changing the input rate to the DAG will cause different solutions to be generated. Given the long-running nature of the event dataflows, the impact of variable input rates or periodic changes to the input rates on the solutions that are generated should be considered.
   More generally, we assume a fixed set of edge and Cloud resources in our problem, and a single solution that is deployed when the DAG is submitted. Since such event analytics run for days or weeks at a time, many factors may change in this period: event rates may change significantly, edge devices may fail or be taken down for planned maintenance, solar energy generation may be lower due to a cloudy day causing edges with longer recharge cycle, network behavior may vary, and connectivity between edge-edge or edge-Cloud may go down all together, and so on. So we should consider our ability to change the solution on the fly as the environmental conditions change, and also to provide robustness to guarantee the latency QoS. Additional strategies to consider may include dynamically moving tasks between edge and Cloud or vice versa, replicating the queries across multiple devices, etc.
5) Lastly, one aspect that we had introduced briefly in an earlier work and remains relevant still is that of planning placement of queries to preserve the privacy of data [23]. IoT deployments offer an unprecedented ability to observe the environment around us. As a result, some of the sensor streams on which we perform analytics may contain sensitive information that would be embarrassing or illegal if compromised [41], [42]. Incorporating privacy constraints as a first-class entity in the placement problem across edge and Cloud is important. Here, we may wish to limit the event streams that go out of the private network, have variable trust in different edge resources, or introduce "anonymizing" queries at the trust boundaries. This is a vast and important area that requires exploration.

IX. Conclusion

Literature on using both edge and Cloud resources often focus on a few application quality parameters, such as latency and throughput; some system characteristics like CPU, memory, network and power; specific types of architectures such as Cloud-only, Mobile-Cloud and Fog Computing; used to support programming models, such as modular tasks, individual queries or stream processing. In this paper, we

have identified a unique combination of these dimensions that are essential for IoT: reducing latency for event dataflows across edge and Cloud, while conserving energy and bound by compute capabilities of the devices.

Our micro-benchmark results offer a unique glimpse on the compute, network and energy performance of edge and Cloud VMs for individual complex event processing queries. The diverse experiments with different query types and event rates offer a broad set of performance numbers that are valuable to evaluate other resource platforms for event analytics, as well as to include them in further studies, such as the optimization problem of resource placement we have used it in.

We have formally defined the query placement problem for a CEP dataflow on to edge and Cloud resources as an optimization problem with constraints on the compute and energy capabilities of the resource. We have proposed a brute force approach (BF) to solving it optimally. Further, we have mapped this problem to a Genetic Algorithm (GA) formulation with a corresponding solution approach that considers the constraints as well.

We validate and evaluate the problem and solution approaches using a simulation study that include a diverse set of synthetic DAGs that are embedded with static and runtime properties that are sourced from the real-world benchmarks. We have obtained results for up to 45 types of DAGs with two resource variants and two input event rates, using the BF and GA approaches as well as a random placement baseline. Our analysis shows that GA gives optimal or near-optimal solutions comparable to BF, offer a better trade-off between lower latency and more frequent feasible solutions than the random placement baseline. It also offers solutions within seconds for even DAGs as large as 50 queries on 50 edge and Cloud resources, while the BF is unable to complete within 13 hours for 12 or more queries and resources. These are promising results that can inform practical IoT deployments using sound theoretical and experimental results.

Finally, we have also offered a detailed summary of our contributions, additional experiments that are recommended, and a swathe of new research ideas to further pursue in this nascent area of analytics across edge and Cloud.

X. Acknowledgments

We wish to thank several students, interns and staff members of the DREAM:Lab, CDS, IISc, who helped with various aspects of the problem formulation, experiments and discussions, including Nithyashri Govindarajan, Shashank Shekhar, and Pranav Konanur. We also thank Dr. T.V.Prabhakar from the Department of Electronic Systems Engineering (DESE) at IISc, and his staff, Madhusudan Koppal and Abhirami Sampath, for enabling the energy measurements for the Raspberry Pi devices. We also thank our research sponsors, Robert Bosch Center for Cyber Physical Systems (RBCCPS) at IISc, and the Department of Electronics and Information Technology (DeitY). We acknowledge Microsoft for their Azure for Research grant that provided us with the Cloud resources for our experiments.

References

[1] T. Robles, R. Alcarria, D. Martn, A. Morales, M. Navarro, R. Calero, S. Iglesias, and M. Lpez, "An internet of things-based model for smart water management," in *WAINA*, 2014.

[2] Y. Simmhan, S. Aman, A. Kumbhare, R. Liu, S. Stevens, Q. Zhou, and V. Prasanna, "Cloud-based software platform for big data analytics in smart grids," *Computing in Science and Engineering*, 2013.

[3] J. Wei, "How wearables intersect with the cloud and the internet of things : Considerations for the developers of wearables." *IEEE Consumer Electronics Magazine*, 2014.

[4] J. P. Sterbenz, "Drones in the smart city and iot: Protocols, resilience, benefits, and risks," in *DroNet*, 2016.

[5] A. B. Zaslavsky, C. Perera, and D. Georgakopoulos, "Sensing as a service and big data," *CoRR*, vol. abs/1301.0159, 2013.

[6] S. Ciavarella, J. Y. Joo, and S. Silvestri, "Managing contingencies in smart grids via the internet of things," *IEEE Trans. on Smart Grid*, 2016.

[7] R. Roman, J. Zhou, and J. Lopez, "On the features and challenges of security and privacy in distributed internet of things," *Computer Networks*, 2013.

[8] M. Zaharia, T. Das, H. Li, S. Shenker, and I. Stoica, "Discretized streams: An efficient and fault-tolerant model for stream processing on large clusters," in *HotCloud*, 2012.

[9] S. Suhothayan, K. Gajasinghe, I. Loku Narangoda, S. Chaturanga, S. Perera, and V. Nanayakkara, "Siddhi: A second look at complex event processing architectures," in *GCE*, 2011.

[10] S. Chandrasekaran, O. Cooper, A. Deshpande, M. J. Franklin, J. M. Hellerstein, W. Hong, S. Krishnamurthy, S. R. Madden, F. Reiss, and M. A. Shah, "Telegraphcq: Continuous dataflow processing," in *SIGMOD*, 2003.

[11] A. Toshniwal, S. Taneja, A. Shukla, K. Ramasamy, J. M. Patel, S. Kulkarni, J. Jackson, K. Gade, M. Fu, J. Donham, N. Bhagat, S. Mittal, and D. Ryaboy, "Storm@twitter," in *SIGMOD*, 2014.

[12] G. Cugola and A. Margara, "Processing flows of information: From data stream to complex event processing," *ACM Comput. Surv.*, 2012.

[13] Q. Zhou, Y. Simmhan, and V. K. Prasanna, "Incorporating semantic knowledge into stream processing for smart grid applications," in *ISWC*, 2012.

[14] Z. Jerzak and H. Ziekow, "The debs 2014 grand challenge," in *DEBS*, 2014.

[15] A. Arasu, M. Cherniack, E. Galvez, D. Maier, A. S. Maskey, E. Ryvkina, M. Stonebraker, and R. Tibbetts, "Linear road: A stream data management benchmark," in *VLDB*, 2004.

[16] C. Mutschler, H. Ziekow, and Z. Jerzak, "The debs 2013 grand challenge," in *DEBS*, 2013.

[17] N. Fernando, S. W. Loke, and W. Rahayu, "Mobile cloud computing: A survey," *FGCS*, 2013.

[18] B.-G. Chun, S. Ihm, P. Maniatis, M. Naik, and A. Patti, "Clonecloud: Elastic execution between mobile device and cloud," in *EuroSys*, 2011.

[19] C. Shi, V. Lakafosis, M. H. Ammar, and E. W. Zegura, "Serendipity: Enabling remote computing among intermittently connected mobile devices," in *MobiHoc*, 2012.

[20] L. M. Vaquero and L. Rodero-Merino, "Finding your way in the fog: Towards a comprehensive definition of fog computing," *SIGCOMM Comput. Commun. Rev.*, 2014.

[21] K. Aberer, P. Cudré-Mauroux, A. Datta, Z. Despotovic, M. Hauswirth, M. Punceva, and R. Schmidt, "P-grid: A self-organizing structured p2p system," *SIGMOD Rec.*, 2003.

[22] U. Srivastava, K. Munagala, and J. Widom, "Operator placement for in-network stream query processing," in *PODS*, 2005.

[23] N. Govindarajan, Y. Simmhan, N. Jamadagni, and P. Misra, "Event processing across edge and the cloud for internet of things applications," in *COMAD*, 2014.

[24] Smartx, "Iisc smart campus: Closing the loop from network to knowledge," 2016. [Online]. Available: http://smartx.cds.iisc.ac.in

[25] T. Rohrmann, "Introducing complex event processing (cep) with apache flink," 2016. [Online]. Available: https://flink.apache.org/news/2016/04/06/cep-monitoring.html

[26] EsperTech, "Esper: Complex event processing and event series analysis for java," 2006. [Online]. Available: http://www.espertech.com/esper/

[27] M. Satyanarayanan, P. Bahl, R. Caceres, and N. Davies, "The case for vm-based cloudlets in mobile computing," *IEEE Pervasive Computing*, 2009.

[28] S. Yang, D. Kwon, H. Yi, Y. Cho, Y. Kwon, and Y. Paek, "Techniques to minimize state transfer costs for dynamic execution offloading in mobile cloud computing," *IEEE Trans. on Mobile Computing*, 2014.

[29] L. Yang, J. Cao, S. Tang, T. Li, and A. T. S. Chan, "A framework for partitioning and execution of data stream applications in mobile cloud computing," in *CLOUD*, 2012.

[30] B. Chandramouli, S. Nath, and W. Zhou, "Supporting distributed feed-following apps over edge devices," *Proc. VLDB Endow.*, 2013.

[31] F. Bonomi, R. Milito, P. Natarajan, and J. Zhu, *Fog Computing: A Platform for Internet of Things and Analytics*, 2014.

[32] S. Abdallah and V. Lesser, "Organization-based cooperative coalition formation," in *IAT*, 2004.

[33] D. R. Karrels, G. L. Peterson, and B. E. Mullins, "Large-scale cooperative task distribution on peer-to-peer networks," *Web Intelli. and Agent Sys.*, 2013.

[34] Y. Yao and J. Gehrke, "Query processing in sensor networks," in *CIDR*, 2003.

[35] J. Nocedal and S. J. Wright, *Numerical Optimization, second edition*. World Scientific, 2006.

[36] Z. Michalewicz, *Genetic Algorithms + Data Structures = Evolution Programs (3rd Ed.)*. Springer-Verlag, 1996.

[37] R. A. Shafik, B. M. Al-Hashimi, and K. Chakrabarty, "Soft error-aware design optimization of low power and time-constrained embedded systems," in *DATE*, 2010.

[38] S. Choy, B. Wong, G. Simon, and C. Rosenberg, "The brewing storm in cloud gaming: A measurement study on cloud to end-user latency," in *NetGames*, 2012.

[39] R. Shea, J. Liu, E. C. H. Ngai, and Y. Cui, "Cloud gaming: architecture and performance," *IEEE Network*, 2013.

[40] A. Shukla and Y. Simmhan, "Benchmarking distributed stream processing platforms for iot applications," *arXiv preprint*, 2016.

[41] U. I. Voytek, "Rides of glory," 2012.

[42] R. H. Weber, "Internet of things–new security and privacy challenges," *Computer Law & Security Review*, 2010.