

Oakestra white paper: An Orchestrator for Edge Computing

Giovanni Bartolomeo, Mehdi Yosofie, Simon Bäurle, Oliver Haluszczynski, Nitinder Mohan and Jörg Ott

Abstract—Edge computing seeks to enable applications with strict latency requirements by utilizing compute resources deployed closer to the users. The diverse, dynamic, and constrained nature of edge infrastructures necessitates a flexible orchestration framework that dynamically supports application QoS requirements. However, existing state-of-the-art orchestration platforms were designed for datacenter environments and make strict assumptions about underlying infrastructures that do not hold for edge computing. This work proposes a novel hierarchical orchestration framework specifically designed for supporting service operation over edge infrastructures. Through its novel federated cluster management, delegated task scheduling, and semantic overlay networking, our system can flexibly consolidate multiple infrastructure operators and absorb dynamic variations at the edge. We comprehensively evaluate our proof-of-concept implementation – *Oakestra* – against state-of-the-art solutions in both controlled and realistic testbeds and demonstrate the significant benefits of our approach as we achieve $\approx 10\times$ and 30% reduction in CPU and memory consumption, respectively.

Index Terms—Edge Orchestration, Kubernetes, Resource Allocation, Hierarchical Scheduling



1 INTRODUCTION

WITHIN almost a decade since its inception, edge computing has found a wide range of use cases in both industry and research, especially for supporting sophisticated services like AR/VR, live video analytics etc. [1]–[3]. However, despite significant interest in the field, there have only been a handful of real-world demonstrations of the paradigm so far [4]. We attribute the cause to the following reasons. Firstly, the resource capacity significantly decreases as we move out of the datacenter silos and towards the network’s edge to match the close-proximity requirements [5] and smaller form factor specialized hardware with CPU, GPU, TPU and VPU become more prevalent [6], [7]. However, unlike siloed deployment of cloud datacenters, setting up edge infrastructures requires significant investment and planning as the compute fabric co-exists alongside its users, e.g., in base stations, city-owned properties, public transport, etc. [4], [8]–[11]. Furthermore, the benefits of edge computing are only apparent when there is a dense availability of computing resources [1], [12]–[14].

Secondly, the majority of de-facto resource orchestration and application deployment frameworks, e.g., Kubernetes [15], K3s [16], KubeFed [17], etc., are either off-shoot branches or adoptions of datacenter-oriented solutions and, therefore, struggle to leverage the edge. Specifically, such approaches make strong assumptions regarding the underlying infrastructure’s consistent reliability and reachability, which does not hold for edge computing. Similarly, finding optimal service deployment in a loosely coupled infrastructure spanning vast geographical regions has been proven to be a non-trivial problem [18]–[20], which is almost entirely overlooked by such works. Furthermore, almost none of the existing scheduling platforms can currently support the significant heterogeneity and diversity in processing, networking [13], [21], and availability of resources [22], which are synonymous with edge computing.

This paper presents a flexible hierarchical orchestration

framework that overcomes the many challenges of edge infrastructures and workloads. Our system allows multiple operators to contribute their resources to shared infrastructure and retain administrative control – thereby significantly reducing the effort to achieve a dense computing fabric at the edge. The key innovation lies in the edge-focused design of our system components that allows the framework to cope with infrastructure complexity while providing application developers familiar experience for deploying and managing their services. Specifically, we make the following contributions:

(1) We propose a *hierarchical* resource orchestration scheme that decomposes the control-plane management across segregated tiers of clusters. Each resource is managed by its local cluster orchestrator which coordinates with its parent orchestrator for exchanging aggregated statistics and deployment commands. Our approach allows flexible infrastructure scaling at the edge by providing context separation between resources managed by different cluster operators (§3).

(2) We propose a *delegated service scheduling mechanism* that decentralizes the task placement problem across the hierarchy to effectively support service deployment at scale (§4). Similar to cloud systems, application developers can define high-level service operational requirements using our SLA definition. Our system offloads the SLA to best-candidate cluster orchestrators, which then find a suitable resource for supporting the service within their operational boundaries. We also present a novel *latency and distance placement* (LDP) algorithm that optimizes latency and geographical distance constraints while deploying services at the edge.

(3) We design a robust overlay network that enables service interactions across edge resources in different (private) networks without overheads (§5). Through our novel *semantic addressing* scheme, we can dynamically (and transparently)

adjust communication endpoints in response to infrastructure changes, e.g., service migrations, resource failures, etc., ensuring uninterrupted service interactions. Our networking component also supports edge-oriented load balancing policies, e.g., connecting to the closest instance, effectively utilizing the geographically vast and diverse edge computing infrastructures.

(4) We implement *Oakestra*, which is lightweight and features compatibility with technologies popularly used in modern cloud applications (§6). Our extensive evaluation conducted in both high-performance computing and realistic edge-like infrastructures shows that *Oakestra* consistently outperforms the state-of-the-art by a large margin and efficiently integrates heterogeneous resources (§7). Our experiments using realistic edge workloads (e.g., live video analytics) highlight the effectiveness of *Oakestra* for supporting distributed microservice-based applications on edge infrastructures. Our results show up to $10\times$ lower CPU overhead and 60% reduction in service deployment time. Under heavy loads, our platform reduces resource utilization by $\approx 20\%$ than its closest competitor.

2 BACKGROUND AND RELATED WORK

Most available service scheduling and monitoring frameworks were designed for datacenter environments. Of these, Kubernetes [15] is the most popular orchestration system in production, used by $\approx 59\%$ large organizations [23], and has been touted by many as the primary solution for managing edge infrastructures. However, Kubernetes’ inherent operation makes strong assumptions about the underlying infrastructure, which was found to be its primary limitation when ported to the edge [24]. Specifically, the platform requires all resources to be in the same cluster and directly reachable (similar to datacenters) – requiring a close coupling between the orchestrator and workers. Additionally, the default service scheduling policies in Kubernetes are not suited for heterogeneous and diverse edge infrastructures as they do not consider metrics such as end-to-end latency, geographical locations, etc. Other frameworks, such as KubeEdge [25], K3s [16], Microk8s [26], and KubeFed [17], have re-architected Kubernetes to make it lightweight and suitable for edge computing. However, recent explorations have also found these to be restrictive, partly since they inherit the strong infrastructure assumptions of Kubernetes in their design as well [27]. On the other hand, our proposed hierarchical orchestration framework is designed from the ground-up to support edge computing infrastructures and workloads through constraint-aware delegated scheduling (§4) and semantic overlay networks (§5).

Only a few works have explored the effective orchestration of edge servers from a research standpoint. CloudPath [28] envisions a multi-tier on-path computing paradigm that allows stateless functions to be deployed closer to end-user and IoT devices. Projects like HeteroEdge [29] or SpanEdge [30] cater specifically to support streaming-based applications on edge servers while FogLamp [31] focuses on data management at the edge. VirtualEdge [32] aims at supporting orchestration at the edge but is limited to cellular networks.

Researchers have also proposed hierarchical scheduling solutions for the edge similar to ours [33]–[35]. Most of these approaches exploit different domain knowledge by distributing the scheduling across the cloud-to-edge hierarchy. In [36], the authors present an autonomous hierarchical scheduling approach that distributes service tasks on a cloud-to-edge continuum. However, while the focus of these works is to design a service scheduling solution for the edge, we provide an orchestration solution that offers both service and resource management for edge infrastructures. Additionally, as we show in §7, *Oakestra* significantly outperforms production orchestration frameworks, thereby demonstrating its suitability for edge computing.

3 FRAMEWORK OVERVIEW

3.1 Challenges & Design Motivations

Previous research has shown that both service deployment and resource management in distributed edge infrastructures are non-trivial problems, primarily due to the heterogeneity and dynamicity of the environment, which convolutes with increasing scale [18]–[20]. Simultaneously, since the coverage area and capacity of edge servers is significantly smaller than cloud datacenters, application providers are likely to deploy multiple service instances with specialized operational requirements to maximize their client’s quality of experience (QoE) [37]. As a result, orchestration platforms designed to support edge computing must propose solutions to two significant challenges. *Firstly*, the platforms should not only incorporate the heterogeneous edge hardware but also support a consolidated, shared infrastructure that allows (a) application developers to utilize edge resources regardless of their ownership, and (b) all participating operators to retain complete contextual and management control over their resources [9], [38], [39]. *Secondly*, the framework must adapt to dynamic infrastructure (and environment) changes without significantly affecting already operational applications.

3.2 System Architecture

We propose a hierarchical orchestration framework for enabling edge computing applications over heterogeneous edge resources. Through our system’s unique multi-cluster resource management, multiple edge operators (e.g., ISPs, cloud operators, city administration, private players, etc.) can contribute their local deployments towards a shared infrastructure while retaining administrative control [40]. As a result, our framework provides a mechanism to realize a dense edge computing fabric without requiring significant deployment investments. Simultaneously, we allow application providers to seamlessly migrate their services to the edge by specifying high-level SLAs. To hide the complexity in edge hardware and infrastructure management from application providers, we design a *delegated service scheduling* approach to effectively handle the scale, density, and heterogeneity at the edge. Similarly, our semantic overlay network with in-built traffic tunneling allows application providers to seamlessly utilize edge resources from different operators without additional management overheads.

Figure 1 shows the high-level architecture of our orchestration framework. Instead of the flat master-slave design

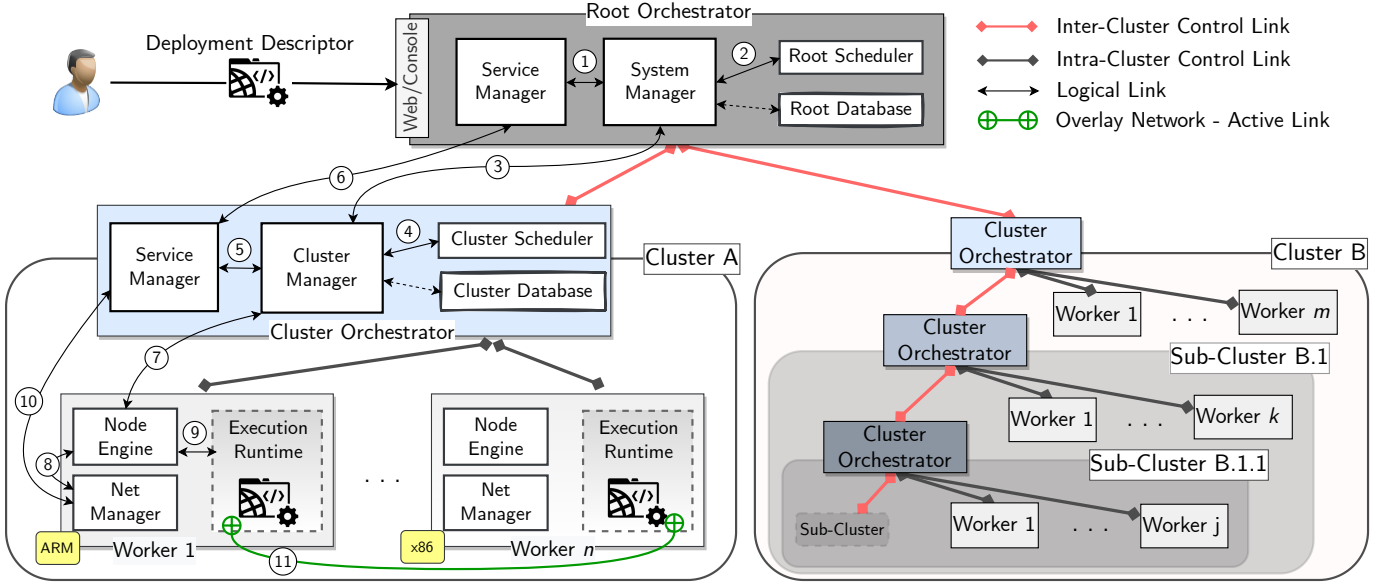


Fig. 1: System Architecture and Workflow.

(inherent to most orchestration solutions [16], [17], [25], [41], [42]), our framework organizes edge infrastructure into hierarchical *clusters* (see clusters A and B). Resources within a cluster (*workers*) are owned and administered by the cluster operator. We leave the definition of “cluster” purposefully abstract as a single operator can deploy several clusters to segregate its resources, e.g., by geographical regions. The hierarchical management extends within each cluster with many sub-clusters attached to their respective parents – forming a tree-like hierarchy (as shown in cluster B). Such a design allows infrastructure providers to logically separate their resources in more specific zones, which would, in turn, ease future scalability. For example, an ISP can operate its infrastructure in different cities as clusters and regions in each city as sub-clusters. We also allow independent providers with under-utilized hardware to participate as a single resource cluster operator in the global infrastructure. We separate the resource and service management responsibilities into different components such that both operations can be performed independently of each other. Specifically, **system manager** is responsible for resource availability and fault-tolerance while **service manager** handles application service deployment and lifecycle management. As shown in the figure, our framework composes of three functional entities – *root*, *cluster*, and *worker*. We now detail the operation of each component.

3.2.1 Root Orchestrator

The *root orchestrator* is the centralized control plane of the framework. The component is analogous to the “control-plane” of Kubernetes [41] and is responsible for managing participating resource clusters. We envision the root orchestrator to be deployed in the cloud or a node reachable from all clusters. Developers interested in deploying their applications at the edge submit the application code and a list of service level agreements (SLA) to the **service manager** in root orchestrator via an API. The SLA includes high-level operational requirements and constraints for service execu-

tion at the edge, e.g., virtualization technology, hardware capacity, geographical location, etc. (detailed in §4.2). The service manager notifies the **system manager** of the new deployment request (step ①), which registers the service in the local database. The system manager contacts the **root scheduler** (step ②) to calculate a priority list of clusters best suited to deploy the application. We design task placement as a multi-step mechanism that distributes the scheduling operation across schedulers in root and clusters (see §4). The system manager is also responsible for registering new clusters to the platform and coordinating information exchange between the cluster and the root orchestrator. Once the service is deployed, the service manager monitors its operational requirements, e.g., service addressing, external discovery requests, inter-cluster service-to-service communication, etc., and takes remedial actions in case of violations (see §5). The database stores the current state of all submitted services and all reported operational information from attached cluster orchestrators [43].

3.2.2 Cluster Orchestrator

Component-wise, the cluster orchestrator is a logical twin of the root, but with management responsibility restricted to resources in the cluster. Any infrastructure provider (e.g., ISP) can register its resources as a consolidated cluster with the root orchestrator via an API. In this case, the operator also assigns the orchestrator role to a node that is ideally reachable by every other resource in the cluster. The cluster orchestrator includes **service manager**, **cluster manager** and **cluster scheduler** components – which perform similar operations to their counterparts at the root. The cluster manager periodically updates the root system manager with *aggregated* statistics of cluster utilization and deployed services via the *inter-cluster control link*. Note that the cluster orchestrator withholds minute information of its member resources to retain administrative control within cluster boundary (see §4.1 for resource management details). Through *delegated service scheduling*, we exploit the logical

information separation between root and the cluster and the increased resource reachability within each cluster to minimize the overhead for task deployment at the root. Specifically, while scheduling services in the infrastructure, the root scheduler only calculates a list of candidate clusters by matching the service SLA constraints to aggregated statistics of attached clusters. It further *delegates* the precise service scheduling task to filtered **cluster schedulers** in highest-priority-first fashion (step ③ and ④). As shown in cluster B, a single cluster can host a multi-tier hierarchy of sub-clusters – each sub-cluster operating as an independent administrative domain of resources. Note that there is no difference between a cluster and a sub-cluster from a system perspective other than the parent resource that acts as the orchestrator. In the case of a multi-tiered cluster hierarchy, the service scheduling task is iteratively delegated down the branch until a suitable edge server for deploying the service is found (see §4 for details). Each cluster (and sub-cluster) **service manager** periodically sends the health and QoS of all operational services within its domain to its respective parent (step ⑤ and ⑥).

3.2.3 Worker Node

We term edge servers responsible for executing services as *workers* which are also the leaves of our orchestration hierarchy. Each worker node has distinct capacity and capability, e.g., CPU cores, local disk size, supported execution runtimes etc., which it reports to the cluster orchestrator at the time of registration. If a worker is found suitable for the requested service’s SLA constraints, the cluster orchestrator instructs the worker’s `NodeEngine` to deploy the service (step ⑦). The worker first reserves the required subnetwork for service communication requirements and instantiates the service inside the execution runtime (step ⑧ and ⑨). Each worker node periodically reports its detailed utilization metrics along with the health of operational services (e.g., SLA default alarms) to its cluster orchestrator via the *intra-cluster control link*. It must be noted that we do *not* require worker machines to have public IP addresses but instead assume that workers within a cluster can only directly access resources within the same (and parent) cluster. In case a deployed service needs to communicate with another service in the system (within same or across clusters), the `NetManager` in the worker fetches the target IP address from the cluster service manager (step ⑩) and establishes the connection via an invisible tunnel (step ⑪). We detail the network management in §5.

4 RESOURCE & SERVICE MANAGEMENT

The majority of existing orchestration frameworks, e.g., Kubernetes [15], K3s [16], etc. were designed for cloud infrastructures and workloads [44] and, therefore, follow a flat centralized (master-slave) management architecture. Such platforms require all resources in the infrastructure to be homogeneous and consistently available – an assumption that does not always hold for edge infrastructures [45]. Previous research has shown service placement at the edge to be an NP-hard problem [18]–[20], [46] which increases in complexity as the variables in the system increase. Consequently, frameworks that inherently rely on a centralized

orchestrator cannot operate at the edge without significant overheads [18]. We overcome these challenges by decentralizing the resource management and service scheduling decisions across cluster tiers. Our design leverages the increased resource reachability within cluster boundaries and, thereby, reduces the dependencies over inconsistent inter-cluster network links for control decisions.

4.1 Resource Management

As discussed in §3.2, edge resources in our system participate in distinct clusters and sub-clusters. We defined such a hierarchy as an oriented tree I such that $I = \langle C, E \rangle$. C is the set of the clusters $\{C_i | C_i = \{R_1^i \dots R_n^i, R_{CO}^i\}; 0 < i, j \leq |C|; i \neq j; R_l^i \neq R_m^j\} \cup C_0$, and $C_0 = \{RO\}$. Here, RO denotes the root orchestrator, R_{CO}^i is the cluster orchestrator of the i -th cluster and R_l^i is the l -th resource of the i -th cluster. We define the edges of this tree $E = E_c \cup \{(C_0, C_i) | \nexists_j (C_j, C_i) \in E_c\}$. Here, E_c is the set of oriented edges, (C_i, C_j) denoting a inter-cluster control link, i.e. a sub-cluster relationship between R_{CO}^i and R_{CO}^j (see fig. 1).

Each resource R_n^i periodically *pushes* its current hardware utilization (U_n^i) and other defining characteristics (e.g. location) to R_{CO}^i with update frequency $\lambda(R_n^i)$ over the intra-cluster link. By correlating U_n^i to the maximum capacity C_n^i of R_n^i reported at registration, R_{CO}^i can monitor the available capacity of each resource in the cluster (denoted by A_n^i). The frequent *push-based* resource status update retains the time-sensitive communication within the cluster where the network is not assumed to be a significant bottleneck. This relieves the cluster orchestrator of unnecessary management and communication overheads as the infrastructure grows and the proximity between resources decreases significantly. $\lambda(R_n^i)$ can be different for each resource and can be adjusted dynamically to balance between “most-updated” information of R_n^i to network overhead caused by frequent updates. For example, a worker may only publish an update if its Δ utilization crosses a threshold; or use age-of-information to dynamically adjust the rate to optimize for the system-at-large [47]–[51]. We leave the exploration of such solutions for tuning update frequency to future work.

Similar to intra-cluster operation, update message exchanges are also push-based over inter-cluster links (C_i, C_j) . Each cluster orchestrator periodically sends the distribution of *available* cluster and sub-clusters capacities, i.e. $\cup(A^i) = \langle \sum(A^i), \mu(A^i), \sigma(A^i) \rangle$ where $A^i = \{A_1^i, A_2^i, \dots, A_n^i\} \cup \{A^j | \exists_j (C_i, C_j) \in E\}$, to the orchestrator of the tier above. The aggregation allows different infrastructure operators to participate in the federated environment while obscuring the minute details of their resources and network. Additionally, each operator can freely scale up/down its cluster density without involving the parent (or root) orchestrator.

4.2 Service Deployment & Scheduling

Application providers can deploy their services on edge servers by specifying QoS requirements as service level agreements (SLA) at the root orchestrator. Schema 1 shows a high-level SLA description supported by our framework. In addition to operational requirements already prevalent in

```

constraints: [{
  microservice_id: {type: number},
  properties: [{
    memory: {type: integer},
    vcpus: {type: integer},
    vgpus: {type: integer},
    vtpus: {type: integer},
    bandwidth_in: {type: integer},
    latency: {type: number},
    area: {type: string},
    location: {type: string},
    threshold: {type: number},
    rigidness: {type: number},
    convergence_time: {type: integer},
    virtualization: {type: string},
    ... }]}
...}]

```

Schema 1: Service Requirement Descriptor.

Algorithm 1: Resource-Only Match

```

Input:  $A_n$  : Information about worker  $n$ .
          $Q_{\tau_{p,i}}$  : Requirements of  $i$ -th task of  $p$ -th service.
          $f(A_n, Q_{\tau_{p,i}})$ : Resource selection strategy.
Output: Best worker  $W$  to run  $\tau_{p,i}$ .

// Resource selection strategy examples:
//  $f(A_n, Q_{\tau_{p,i}}) = \arg \max_n [(A_n^{cpu} - Q_{\tau_{p,i}}^{cpu}) + (A_n^{mem} - Q_{\tau_{p,i}}^{mem})$ 
//  $\quad \quad \quad \wedge Q_{\tau_{p,i}}^{virt} \in A_n^{virt}]$ 
//  $f(A_n, Q_{\tau_{p,i}}) = \text{first}_n [Q_{\tau_{p,i}}^{cpu} \leq A_n^{cpu} \wedge Q_{\tau_{p,i}}^{mem} \leq A_n^{mem}$ 
//  $\quad \quad \quad \wedge Q_{\tau_{p,i}}^{virt} \in A_n^{virt}]$ 
1  $W \leftarrow f(A_n, Q_{\tau_{p,i}})$ 
2 return  $W$ 

```

cloud environments, such as processing performance, networking requirements, virtualization needs, etc., the schema allows developers to specify edge-specific restrictions, e.g., geographical location, specialized hardware, etc. Additionally, developers can fine-tune the precision of scheduling heuristics by enforcing *convergence time* and *decision rigidness* metrics. Convergence time specifies the maximum allowed time within which the scheduler should find the suitable edge server that supports the SLA requirements of the service, and rigidness defines the sensitivity for re-triggering service scheduling in case the selected resource violates the SLA (due to environment/infrastructure changes).

To support application deployment over vast and highly variable edge infrastructures, we propose a *delegated service scheduling* mechanism. As shown in Figure 1, both root and cluster orchestrators have their separate scheduling components that are responsible for solving a subset of the task placement problem within their respective domains. Let $S = \{s_1, s_2, \dots, s_{|S|}\}$ denote the set of services requested to be deployed by the developers at the root. Each service $s_p \in S$ can be composed of n individual microservices or tasks, i.e. $s_p = \{\tau_{p,1}, \tau_{p,2}, \dots, \tau_{p,n}\}$ where $\tau_{p,i}$ denotes i -th task of p -th service. Each task $\tau_{p,i}$ requires a certain capacity (CPU, GPU, memory), denoted by $Q_{\tau_{p,i}}$. Other considerations like geographical location or virtualization technology, specified by the developer in the SLA, are also part of $Q_{\tau_{p,i}}$. The task of the scheduling components (in both root and cluster) is to find a suitable resource in the infrastructure that supports the requirements in $Q_{\tau_{p,i}}$. However, since detailed resource availability and utilization information is restricted within cluster boundaries, service scheduling in a t -tier edge infrastructure hierarchy is conducted in t steps.

In the first step, the root scheduler matches $Q_{\tau_{p,i}}$ to

Algorithm 2: Latency & Distance Aware Placement

```

Input:  $A_n$  : Information about worker  $n$ .
          $Q_{\tau_{p,i}}$ : Requirements of  $i$ -th task of  $p$ -th service.
Output: Best workers  $W$  to run  $\tau_{p,i}$ .

1  $W \leftarrow \{n \in [1, |A|] \mid A_n^{cpu} \geq Q_{\tau_{p,i}}^{cpu} \wedge A_n^{mem} \geq Q_{\tau_{p,i}}^{mem} \wedge$ 
    $\quad \quad \quad Q_{\tau_{p,i}}^{virt} \in A_n^{virt}\}$ 
2 if  $|Q_{\tau_{p,i}}^{s2s}| \geq 1$  then
3   for  $Q_j$  in  $Q_{\tau_{p,i}}^{s2s}$  do
4      $t \leftarrow Q_j^{trg}$ 
5      $W \leftarrow \{n \in W \mid \text{dist}_{gc}(A_n^{geo}, A_t^{geo}) \leq Q_j^{geo\_thr} \wedge$ 
    $\quad \quad \quad \text{dist}_{euc}(A_n^{viv}, A_t^{viv}) \leq Q_j^{viv\_thr}\}$ 
6   end
7 end
8 if  $|Q_{\tau_{p,i}}^{s2u}| \geq 1$  then
9   for  $Q_k$  in  $Q_{\tau_{p,i}}^{s2u}$  do
10     $u \leftarrow Q_k^{lat\_trg}$ 
11     $rtts \leftarrow \{rtt_{i,u} \mid i \in \text{rnd}(W), rtt_{i,u} = \text{ping}(i, u)\}$ 
12     $\text{vivaldiNet} \leftarrow \{A_n^{viv} \mid n \in [1, |A|]\}$ 
13     $\tilde{U} \leftarrow \text{trilateration}(rtts, \text{vivaldiNet})$ 
14     $W \leftarrow \{n \in W \mid \text{dist}_{gc}(A_n^{geo}, Q_k^{geo\_trg}) \leq Q_k^{geo\_thr} \wedge$ 
    $\quad \quad \quad \text{dist}_{euc}(A_n^{viv}, \tilde{U}) \leq Q_k^{lat\_thr}\}$ 
15  end
16 end
17 return  $W$ 

```

$\cup(A^i)$ for each cluster such that $\exists_i(C_0, C_i) \in E$ and calculates a priority list of best-fit clusters. This step filters out all clusters not suitable for the task, e.g., insufficient resource availability, not within target geographical region, no support for the desired virtualization, etc. The root scheduler then offloads the deployment request to the orchestrator of the cluster with the highest priority. The request includes both the task $\tau_{p,i}$ and its requirements $Q_{\tau_{p,i}}$. In the following $t - 1$ steps, the respective cluster schedulers either find a suitable worker for the deployment, resulting in early termination of the t -step scheduling process, or in turn calculate another best-fit sub-cluster and propagate the deployment request down the branch of the tree I . Each cluster scheduler can utilize different task placement algorithms to find suitable workers within its boundary depending on the metrics to be optimized [37]. In this work, we propose and incorporate two different scheduling approaches.

(1) Resource-Only Match (ROM): As the name suggests, in ROM, the cluster scheduler finds a suitable resource within the cluster that satisfies the capacity requirements of the service (see Algorithm 1). The scheduling approach is analogous to greedy-fit and knapsack-based solutions popularly used for placing VMs on cloud servers in datacenters [52].

(2) Latency & Distance Aware Placement (LDP): LDP (shown in Algorithm 2) builds on the ROM scheduler but additionally considers latency and geographical distance constraints for service placement. Since edge applications can be composed of multiple microservices that can either interact amongst each other (in a chain-like fashion) or directly with end users/devices, we allow the application provider to specify constraints for both service-to-service (S2S) and service-to-user (S2U) links. The root scheduler first filters unsuitable clusters by comparing their resource constraints along with approximate geographical operation zones to the SLA requirements. Within each cluster, the algorithm first creates a list of candidate workers that sat-

isfy the resource constraints. Then, for all S2S constraints $Q_{\tau_p, i}^{s2s}$, the algorithm filters out workers that exceed the specified distance $Q_j^{geo_thr}$ and latency thresholds $Q_j^{viv_thr}$ to the target service $t = Q_j^{trg}$. LDP estimates geographic distance as the great circle distance ($dist_{gc}$) between the geographic location of worker n (A_n^{geo}) and the location of the target service A_t^{geo} . The approximated latency is the Euclidean distance ($dist_{euc}$) between the location of worker n (A_n^{viv}) and the location of the target service A_t^{viv} in the Vivaldi network [53]. Vivaldi is a network coordinate system embedding networked nodes into a d -dimensional coordinate system such that the Euclidean distance of two nodes approximates their round-trip time. If the developer has specified any S2U constraints $Q_{\tau_p, i}^{s2u}$, LDP measures the round-trip times ($rtts$) to the target as $Q_k^{lat_trg}$ from a set of random workers in the cluster ($i \in rnd(W)$). The measurements approximate the user’s position within the Vivaldi network via trilateration [54]. Following that, LDP filters out workers that either exceed the distance threshold $Q_k^{geo_thr}$ to $Q_k^{geo_trg}$ or the latency threshold $Q_k^{lat_thr}$ to the approximated user position \tilde{U} .

Suppose the cluster scheduler cannot find appropriate resources for all application microservices within the same cluster. In that case, our federated clustering approach allows the framework to place the services across several multiple clusters. In that case, the root orchestrator iteratively requests the clusters in the priority list to search for a locally optimal worker for service deployment. In case of resource failures, which are highly probable at the edge, the service manager of the associated cluster marks all affected services as failed. The orchestrator then attempts to re-deploy each service on another resource that satisfies the SLA requirements within the same cluster. If unsuccessful, the rescheduling request is recursively propagated to the root orchestrator until a suitable worker is found [54]. Similarly, if the cluster orchestrator observes SLA violations of a running service, it triggers a migration. Service migration follows a similar procedure as service rescheduling in case of failures, except the original service is terminated once the newer instance becomes operational.

Our delegated service scheduling approach significantly reduces the problem search space of multi-objective task placement at the edge by only considering a subset of candidate resources (limited to each cluster). Note that in addition to our proposed ROM and LDP approaches, it is also possible to integrate extensive research on edge service scheduling [37] into our framework, as long as the algorithm can be re-architected to operate on a t -tier hierarchy. However, we admit that our hierarchical scheduling mechanism sacrifices global optimality for reduced complexity. Therefore, in the future, we plan to leverage recent investigations on this topic and incorporate heuristics and deadline-guarantee-based approaches [55] to discover a near-optimal solution with increased probability.

5 SERVICE COMMUNICATION

Supporting reliable intra-service (and service-to-user) networking in edge infrastructures can be quite challenging. Unlike cloud environments, edge infrastructures are susceptible to dynamic changes and hardware failures. Application

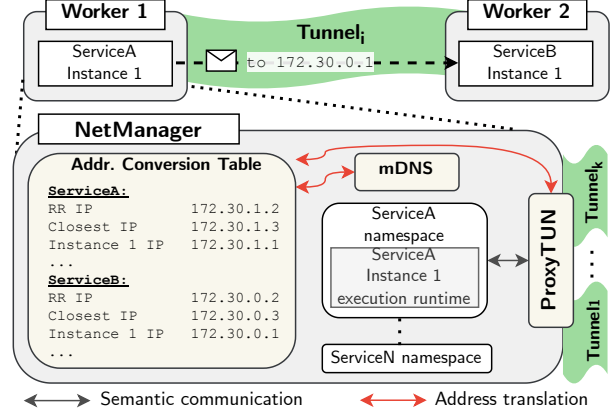


Fig. 2: Service communication across edge servers.

deployment is more fluid at the edge as services migrate to remain close to mobile users and dependent microservices [56]. Furthermore, developers are likely to deploy several instances of their application microservices to cover a larger geographical area. As a result, traditional network load balancing techniques [57] do not function well at the edge as clients are not only interested in connecting to the least loaded instance but also the one deployed “closest” to them. Moreover, with multiple participating infrastructure operators, it is impractical to presume that every edge server will be accessible over a public network – which is an implicit assumption for most existing orchestration frameworks [15], [16], [26].

To overcome the challenges above, we design a robust networking component (named NetManager) that (i) accommodates dynamic infrastructure changes and (ii) transparently integrates resources from multiple operators without imposing overheads on application providers. As shown in Figure 1, NetManager is installed only at the workers, which allows us to separate the bulk of data-plane networking complexity from the control-plane operations. Figure 2 shows the cross-section of the NetManager component enabling communication between two services (service A and B) deployed on different edge servers (worker 1 and 2). We use logical IP addresses to decouple the physical address of the edge server from the address of the deployed service, thereby forming a service overlay that remains oblivious to underlying infrastructure changes. On top of that, the addressing mechanism binds multiple semantic IP addresses, namely `serviceIPs`, to each service operational on the worker. Each `serviceIP` maps to a different instance of that service in the infrastructure according to a balancing policy, which is tracked in the address conversion table (analogous to a routing table). A `serviceIP`, drawing inspiration from semantic routing [58], can be used by an application to find the instance that best suits that policy automatically. For example, in fig. 2, worker 1 maintains ServiceA’s logical address for each instance as well as the *closest* and *round robin* `serviceIPs`.

The address table also tracks `serviceIPs` of target services (in this case, ServiceB) which allows ServiceA to communicate with the closest instance of ServiceB using *closest* `serviceIP`. The NetManager also includes local mDNS which enables services to use load balancing naming

schemes instead of IP addresses (e.g. *serviceB.closest*). At the time $t = 0$, the worker sets all entries in the conversion table, except the local service instance address, to `null`. Suppose an operational service sends a network packet towards an unknown address. In that case, the node requests an IP resolution to the cluster orchestrator’s service manager (see ⑩ in fig. 1) and populates its table entries. Similarly, if the table data is insufficient to instantiate a connection or the `serviceIP` gives a network error, the worker explicitly requests its orchestrator for a conversion table refresh – which is recursively propagated up the hierarchy until resolved. Any future updates to the requested `serviceIPs` are automatically pushed to the worker by the orchestrator.

The `NetManager` natively supports end-to-end encrypted transport-layer tunneling, using `proxyTUN` to (i) transparently maintain the service overlay network across multiple nodes and (ii) allow safe traversal over untrusted networks. Each service requests the `proxyTUN` to establish a connection to a `serviceIP`. The `proxyTUN` first chooses the service instance based on the requested balancing policy, then translates the semantic address to the logical address through a table lookup. The `proxyTUN` actively maintains and dynamically adjusts endpoints of the tunneled connections to adapt to infrastructure changes, meanwhile ensuring that the services continue to communicate uninterrupted. For ingress/egress traffic outside the edge infrastructure (e.g., towards private end-users or third-party endpoints), the `proxyTUN` uses the service manager in cluster orchestrator as a VPN server that either redirects or tunnels the traffic on behalf of the worker. In the future, we plan to utilize multipath transmissions using MPTCP [59]–[61] over multiple available network interfaces simultaneously at the edge to improve network reliability and availability [21].

As the number of simultaneous service connections escalates, maintaining an increasing number of tunnels may become a burden for `proxyTUN`. To overcome this, we distinguish between *configured* and *active* links. We consider a tunnel to be *active* only when services are using it for data transmissions. In contrast, if a tunnel has been inactive for a while, e.g., the service using it has migrated, it is marked as *configured* and becomes a candidate for garbage collection. Therefore, each resource R_j^i has maximum $n - 1$ outbound *configured* links, one for every worker in the infrastructure, where $n = \sum_i (|C_i| - 1)$ (excluding R_{CO}^i). We can define L as the set of all the *configured* links $(R_l^i, R_m^j) \in L$ such that $\forall C_i, C_j \in C; 0 \leq l < |C_i|; 0 \leq m < |C_j|$ and $R_l^i \neq R_m^j$. Consequently, defining k as the maximum number of *active* tunnels that can be maintained in each worker node, the set A of the *active* links is $A \subset L$ iff $k < n$. When the number of required tunnels exceeds k , the tunnel eviction mechanism uses the least recently used (*LRU*) policy.

6 IMPLEMENTATION: OAKESTRA

As discussed in §3, edge infrastructures can be composed of many heterogeneous devices with diverse form factors, hardware characteristics, and energy constraints [6], [7], [62]. From a software perspective, these devices can have different runtime characteristics, including CPU architectures, virtualization support, etc. An ideal orchestration

platform must be capable of absorbing the inherent infrastructure heterogeneity without amplifying operational overheads – all the while providing developers familiar techniques to seamlessly extend their cloud-supported applications to use the edge.

We implement our orchestration framework and its components (shown in fig. 1) as `Oakestra`. Our comprehensive implementation, spanning over 11,000 lines of code, supports popular development tools and virtualization techniques necessary for edge computing. We currently restrict `Oakestra` only to support a two-tier hierarchy (i.e., without sub-clusters) since the topology already embraces most edge computing models [4], [5] and can sufficiently demonstrate the benefits of hierarchical orchestration at the edge. However, `Oakestra` can be extended to support multiple hierarchy tiers with relative ease.

Orchestration. We implement the root and cluster orchestrators in ≈ 4500 and 2800 lines of Python code, respectively. Infrastructure operators can initialize a cluster of edge servers as workers attached to a cluster orchestrator and register it with the root. The key technical difference between the two orchestrators is the communication protocol for the control plane traffic. Within each cluster, control message exchange between workers and the orchestrator (i.e., worker statistics, scheduling directives, etc.) is over MQTT, which is a lightweight message-passing networking protocol that allows `Oakestra` to scale cluster sizes without communication overheads efficiently. On the other hand, the interaction between cluster and root orchestrator utilizes HTTP(S) WebSockets, which implicitly allows us to monitor the liveness of both orchestrator endpoints and trigger remedial actions in case of failures.

Service Scheduling and Management. Application providers can deploy their services at the edge by submitting the SLA and code binaries to the root orchestrator. Using our delegated scheduling scheme, the scheduler component of the root and cluster orchestrator finds a suitable placement for the services in the infrastructure. We implement both ROM and LDP service scheduling algorithms (discussed in §4) as *language-agnostic plugins* to `Oakestra`’s scheduler. Our purposeful design choice allows future extensions as researchers/developers can easily incorporate custom scheduling algorithms without significant implementation overhead. Furthermore, since the scheduling logic of each tier is independent and decoupled, cluster operators can fine-tune the service scheduling and resource utilization behavior of their cluster by customizing the algorithm to their preference.

`Oakestra` also keeps track of deployed applications throughout their lifecycle through a state machine. Each service instance starts as *requested*, indicating that the root scheduler has initiated the scheduling process. Once the cluster orchestrator finds a suitable worker for the service deployment, the service state becomes *scheduled*. The worker deploys the service instance and periodically reports the current QoS and current resource utilization ($\cup(A^i)$) to the cluster orchestrator – changing the service status to *running*. As discussed in §4, if the cluster orchestrator observes lapses in expected service behavior, e.g., the service becomes unavailable or violates its SLA, it triggers an implicit migra-

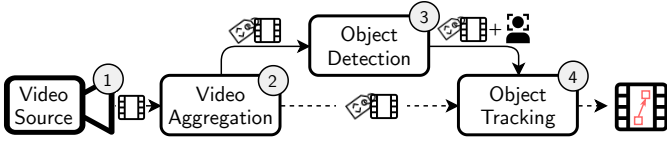


Fig. 3: Cross-section of video analytics application.

tion request that is handled as a (new) scheduling request. Once the migrated service instance becomes operational, the previous instance is undeployed, and its status changes to *terminated*. *Oakestra* also supports service *replication* which follows a similar procedure as service migration, except for the termination of original service. In case of unexpected early termination or failures, the affected service instances are marked as *failed*.

Networking. We implement the *NetManager* component (see fig. 2) of *NodeEngine* in $\approx 3,000$ lines of Go code to ensure a low resource footprint without sacrificing performance. The worker nodes obtain a unique subnetwork upon registering with their cluster orchestrator during the initial handshake. Each deployed service is mapped to a logical address in the local subnetwork. The *NetManager* creates a bridge that connects the virtual interface of the service instance to the *ProxyTUN* component that is responsible for ingress/egress traffic. The *ProxyTUN* handles *ServiceIP* resolution (e.g., round-robin, closest) in a separate thread and finally establishes UDP-based connection tunnels to the resolved node/interface. Additionally, the *NetManager* utilizes the MQTT communication channel between the worker and the cluster orchestrator for receiving service routing updates, e.g., in case of scaling, migration and undeployment.

7 EVALUATION

In this section, we evaluate and compare *Oakestra* to state-of-the-art orchestration frameworks through a series of experiments in realistic infrastructure testbeds (§7.2). We also investigate the effectiveness of our service scheduling solutions (§7.3) and showcase *Oakestra*’s ability to support the operation of realistic edge application workloads (§7.4).

7.1 Experiment Setup

We set up two different infrastructure testbeds for our evaluation. Our *High-Performance Computing (HPC)* testbed is a large VM-based compute cluster which allows us to configure different experiment configurations in a controlled environment. We use VMs of different sizes for our tests, namely *S*, *M*, *L*, *XL* with 1, 2, 4, 8 GB RAM and 1, 2, 4, 8 CPUs, respectively. All VMs used Ubuntu 18.04 LTS over x86 processors. Our other testbed is a *Heterogeneous (HET)* edge-like cluster composed of devices with different hardware configurations, e.g., Raspberry Pis [62], Intel NUCs [6], mini-desktops, and Nvidia Jetson AGX Xavier [63]. While VMs in our HPC testbed run on servers interconnected by 1 Gbps ethernet, our HET devices are interconnected via a mix of WiFi and ethernet links. As a result, our *HET* setup closely emulates a realistic deployment of edge computing with wirelessly-connected constrained resources.

We attempted to compare *Oakestra*’s performance against the most popular orchestration frameworks for our system evaluation. However, despite our careful management, KubeFed [17], KubeEdge [25], ioFog [42] and Fog05 [64] performed quite inconsistently in our setup, often exhibiting random failures. We attribute this behavior to their relatively early development stage and omit them from our analysis. As a result, we compare *Oakestra* against Kubernetes (K8s) [15], MicroK8s [26] and K3s [16], all of which are widely used real-world production-ready systems and have been proposed to operate at the edge [24], [27].

We use two different application workloads for our evaluation. For our stress-test experiments, we utilize an Nginx web server which allows us to control the operational load on workers dynamically. We also developed a (live) video analytics application that detects and tracks objects in a video, which has been touted as the killer application for edge computing [65]. The pipeline (shown in fig. 3) is composed of four microservices [66]. The **video source** sends an RTP encoded video stream ① and can be replaced with a live camera. For repeatable experiments, we use the wildtrack dataset [67] as our source. The **video aggregation service** ② stitches multiple camera feeds together and performs some pre-processing for the rest of the pipeline. The **object detection service** ③ uses YOLOv3 to detect objects in every frame. The processed metadata is sent to the **object tracking service** ④, which tracks the movement of each detected object across frames. All application services were virtualized as Docker containers. We repeat all our experiments *at least* ten times across multiple days. Only one framework was operational at any given time, and we flush the memory and disk of all resources at the end of each run to avoid artifacts due to residual files. Unless otherwise specified, we consolidate all workers in *Oakestra* within a single cluster to architecturally resemble and remain comparable to other master-slave orchestration platforms (with cluster orchestrator analogous to master).

7.2 Orchestration Performance

Service Deployment. Figure 4a compares the time taken by each framework to deploy a low-footprint containerized Python application that tracks its deployment time. To emulate a constrained edge environment, we configure *XL* VM as root, *L* VM as cluster orchestrator in *Oakestra* and master in other platforms. All platforms use *S* VMs as workers. We increase cluster size from 2 to 10 workers and measure overheads due to the service scheduler of each framework by toggling its operation, shown with *s* (scheduler) and *ns* (no scheduler). We observe that MicroK8s and Kubernetes (K8s) perform significantly worse ($\approx 10\times$ slower for MicroK8s) than *Oakestra*. MicroK8s’ performance degrades considerably with increasing infrastructure size. We also find that for frameworks other than MicroK8s, scheduler operation adds almost negligible overhead to service deployment. Our results are in line with other recent measurements on the topic [27]. Note that *Oakestra* exhibits minimal service deployment time, which remains unaffected by infrastructure size.

Since K3s’s performance closely matched *Oakestra*, we tested both platforms in our HET testbed and gradually

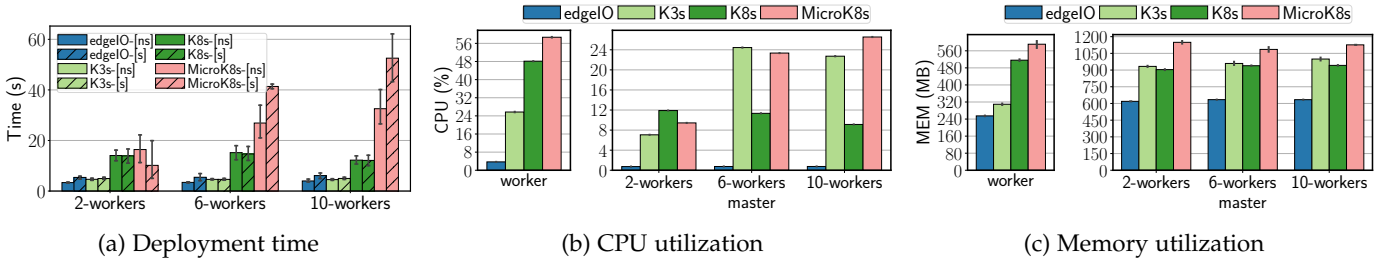


Fig. 4: Orchestration framework performance for different infrastructure sizes.

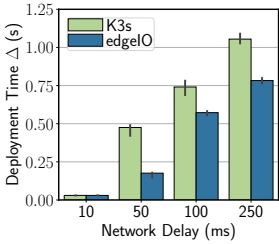


Fig. 5: Deployment time with increasing network delay.

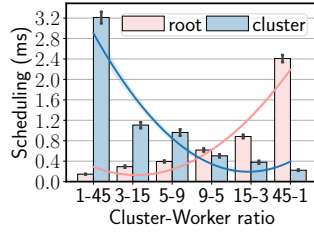
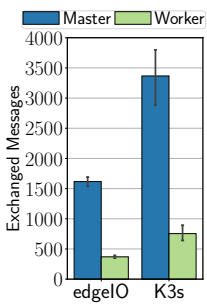
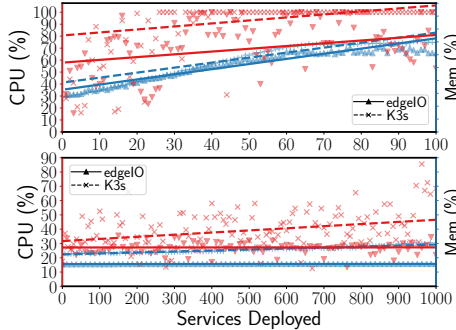


Fig. 6: Schedule processing for cluster-worker ratios.



(a) Total control worker (top) & cluster orch. (bottom) message overhead.



(b) CPU (red)/memory (blue) usage of

Fig. 7: Orchestration overhead in 10 node cluster.

degraded the network conditions by using `tc` utility. Figure 5 shows that `Oakestra`'s deployment time performance surpasses `K3s` by $\approx 20\%$ with increasing network delays. We observe similar behavior with packet losses on the network as well. Specifically, `Oakestra` consistently outperformed `K3s`, achieving $\approx 50\%$ and 60% reduction in deployment time with 20% and 50% losses, respectively (plot not shown for brevity). Note that the above experiments use a single cluster configuration, which is the worst-case scenario for `Oakestra`'s multi-cluster orchestration. To better showcase the capabilities of `Oakestra`, we record the time taken by the root and cluster scheduler for a different number of clusters and workers per cluster configurations (see fig. 6). It can be observed that `Oakestra` achieves better performance when the workers are somewhat balanced across multiple clusters in the hierarchy (see minima around nine clusters with five workers setting).

Scalability. Figures 4b and 4c compares each orchestration platform's idle resource consumption in the HPC testbed with cluster size ranging from two to ten workers. Lower overhead at the worker indicates the platform's capability

to operate on constrained devices. On the other hand, lower overhead at the master highlights the platform's ability to handle scale. We observe that all frameworks consume considerably more CPU for their operation than `Oakestra`. Between the competitors, we find that `K3s` has a low worker node footprint while `K8s` supports scaling better as its performance in the master remains relatively consistent. On the other hand, thanks to its lightweight design, `Oakestra` can support both scale and resource constraints at the edge as it achieves $\approx 6\times$ reduction in CPU and $\approx 18\%$ memory usage on the workers along with $\approx 11\times$ less CPU and $\approx 33\%$ less memory on the master.

Figure 7 shows the CPU, memory, and bandwidth overhead of `Oakestra` against `K3s` for increasing service deployment. From Figure 7a, we find that `K3s` sends $\approx 2\times$ more control messages (from both worker and master) compared to `Oakestra` on the network for orchestration operations. Not only does the increased messages indicate potential overhead of `K3s` in constrained edge network, but also highlights its dependency on network conditions for optimal operation – justifying our results in fig. 5. Figure 7b compares the resource consumption in a cluster of 10 workers as we increasingly schedule up to 100 Nginx containers on each worker (totaling 1000 containers in the cluster). The top half illustrates the worker's utilization, while the bottom shows the performance of the cluster orchestrator (or `K3s` master). Note that the primary overhead in the master is due to the scheduling operation. We find that the `Oakestra` orchestration causes almost negligible overhead and can support numerous services effectively, achieving $\approx 10\text{--}20\%$ better performance than `K3s`. Similarly, the low footprint of `Oakestra` has significant operational advantages for the workers. While `K3s` exhausted the available CPU at the worker at ≈ 60 services, `Oakestra` was able to deploy 100 services with 30% CPU still available.

Networking. We now compare the performance of `Oakestra`'s networking component to the state-of-the-art. For our experiments, we set up multiple Nginx server replicas on different workers, and then deploy a single client performing GET requests to a server instance. Figure 9 (left) shows the average round-trip latency between the client and the closest server. A lower RTT indicates the effectiveness of load balancing by the platform. On average, `K3s` performs $\approx 10\text{--}20\%$ better than `Oakestra` in a single client-server setup, while Kubernetes and `MicroK8s` perform considerably worse – likely due to their substantial operational overhead in constrained resources (as noted in fig. 4). With multiple server replicas, the performance of all

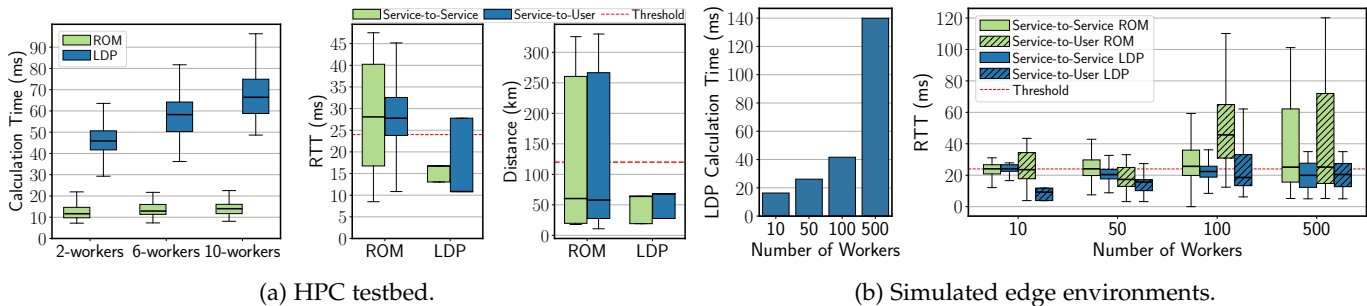


Fig. 8: Resource-Only Match (ROM) and Latency and Distance Aware Placement (LDP) scheduler performance.

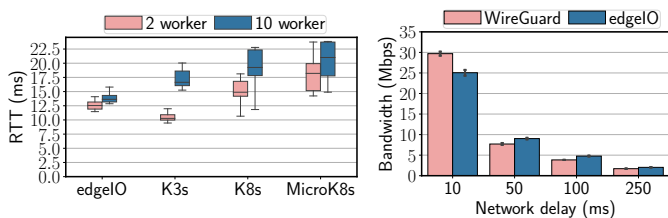


Fig. 9: Network latency and bandwidth overheads.

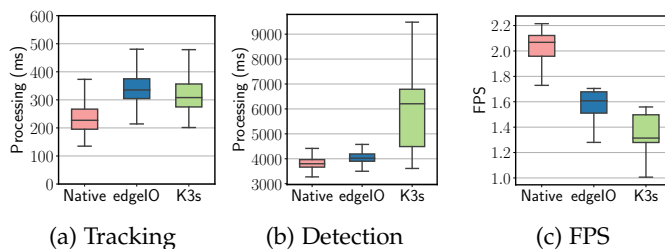


Fig. 10: Live video analytics application performance.

competitors degrades significantly, resulting in RTT inflation of $\approx 20\%$ compared to *Oakestra*. Closer investigation revealed the cause of network overhead in a single service instance setup to *Oakestra*'s traffic tunneling. Recall from §5 that *Oakestra* uses L4-tunneled traffic to allow communication across cluster networks. We evaluate the impact of *Oakestra*'s tunneling on network performance and compare it with *WireGuard* [68] – an open-source tunneling solution used by most frameworks. We emulate the network inconsistencies at the edge [1] by gradually increasing the delay between the client and the servers from 10 to 250 ms. Figure 9 (right) compares the time to download a 100 MB sparse file over HTTP using both approaches. While *WireGuard* achieves $\approx 10\%$ higher bandwidth than *Oakestra* in low latency settings, the performance gap diminishes with network delays. We also measured the performance of both systems for different loss rates (1% to 10%) and always found *Oakestra* to be in competitive range (2–10%) of *WireGuard*. We must stress that *Oakestra* is in its early development stages. We plan to investigate alternate approaches such as L7 connection-level tunneling instead of current L4 per-packet tunnels in the future.

7.3 Service Scheduling

We now evaluate our proposed ROM and LDP schedulers in (a) *HPC* with up to 10 workers and (b) simulated infrastructure with up to 500 edge servers. While our *HPC* experiments provide us an insight into the real-world operation of our schedulers using *Oakestra*, our simulation experiments allow us to investigate the behavior of the schedulers at scale. We configure network latencies between edge servers within 10 - 250 ms, which, as per recent research [12], is a typical latency range between users and cloud datacenters globally. We then instruct the schedulers (using SLA) to find workers that satisfy 1 CPU, 100 MB memory, ≈ 20 ms latency (usual for immersive edge applications [1]) and 120 km operational distance. Figure 8a shows *HPC* results.

Since ROM only performs a best-fit match for overall computational requirements, its calculation time is significantly lower than LDP, whose computational complexity is much larger due to distance calculations and trilateration in the *Vivaldi* network. However, LDP almost always satisfies the latency and geographical SLA constraints (shown as dashed red lines) albeit at a higher calculation cost which increases with infrastructure size.

We investigate the schedulers' behavior further in our simulation experiments (fig. 8b) which shows the LDP calculation time with up to 500 workers and achieved RTT latencies by both ROM and LDP. We find that LDP's calculation time escalates several manifolds with infrastructure size. However, the absolute time is still in the milliseconds' range, which may not be a significant overhead as the service is not yet operational. On the other hand, LDP can effectively support latency-based service constraints at the edge since it usually satisfies the latency thresholds even in large infrastructures (see RTT in 500 worker cluster). We attribute minor lapses in latency thresholds to *Vivaldi*, whose accuracy is significantly affected by triangle inequality violations in large networks [53].

7.4 Realistic Edge Application Support

We now investigate the capability of the orchestration frameworks to support the operation of the video analytics pipeline described in §7.1. We create a cluster of four *S* VMs as workers in our *HPC* testbed and map each microservice of the application to separate workers. Interestingly, we observed that both *Kubernetes* and *MicroK8s* were unable to reliably support the application in our tests since their orchestration components consumed the majority of the resource's capacity (see fig. 4). As a result, we compare *Oakestra*'s performance to *K3s*, and without any orchestration (named native) on the same infrastructure. Since the application can fully utilize the available resource capacity

in the native setting, we consider it our baseline. Figure 10 shows our results. Oakestra and K3s exhibit similar performance for object tracking, taking ≈ 300 -400 ms. However, due to its minimal footprint Oakestra significantly outperforms K3s for supporting the more resource-demanding object detection, achieving results closer to the baseline. Overall, application performance over Oakestra exceeded K3s by almost 10%. We also replicated our experiments in HET but skip its discussion as the application performance was similar to HPC.

CONCLUSION

In this paper we presented a flexible orchestration framework designed to support diverse and heterogeneous edge computing environments. Through our unique hierarchical cluster management, we allow multiple operators to participate in a federated infrastructure while retaining complete administrative control. We also designed a delegated service scheduling mechanism and service overlay networks to effectively deploy and support application services in edge infrastructures spanning vast geographical regions over different networks. We implemented Oakestra and thoroughly evaluated it against the state-of-the-art using various experiments in realistic edge testbeds. Oakestra consistently outperformed its competitors, achieving $\approx 10\times$ reduction in resource usage and 10% improvement in application performance.

REFERENCES

- [1] N. Mohan *et al.*, "Pruning edge research with latency shears," in *Proceedings of the 19th ACM Workshop on Hot Topics in Networks*, ser. HotNets '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 182–189. [Online]. Available: <https://doi.org/10.1145/3422604.3425943>
- [2] A. Y. Ding *et al.*, "Roadmap for edge ai: A dagstuhl perspective," *SIGCOMM Comput. Commun. Rev.*, vol. 52, no. 1, p. 28–33, mar 2022. [Online]. Available: <https://doi.org/10.1145/3523230.3523235>
- [3] V. Cozzolino *et al.*, "Nimbus: Towards latency-energy efficient task offloading for ar services," *IEEE Transactions on Cloud Computing*, pp. 1–1, 2022.
- [4] S. A. Noghabi *et al.*, "The emerging landscape of edge computing," *GetMobile: MCC*, 2020.
- [5] W. Shi and S. Dustdar, "The promise of edge computing," *IEEE Computer*, 2016.
- [6] "Intel nuc," <https://www.intel.com/content/www/us/en/products/details/nuc/boards.html>.
- [7] "Coral edge," <https://coral.ai/products>.
- [8] N. Mohan and J. Kangasharju, "Edge-fog cloud: A distributed cloud for internet of things computations," in *2016 Cloudification of the Internet of Things (CIoT)*, 2016, pp. 1–6.
- [9] A. Zavodovski *et al.*, "Deccloud: Truthful decentralized double auction for edge clouds," in *2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS)*, 2019, pp. 2157–2167.
- [10] A. Silvestro *et al.*, "Mute: Multi-tier edge networks," in *Proceedings of the 5th Workshop on CrossCloud Infrastructures and Platforms*, ser. CrossCloud'18. New York, NY, USA: Association for Computing Machinery, 2018. [Online]. Available: <https://doi.org/10.1145/3195870.3195871>
- [11] N. Mohan and J. Kangasharju, "Placing it right!: optimizing energy, processing, and transport in edge-fog clouds," *Annals of Telecommunications*, vol. 73, no. 7, pp. 463–474, 2018. [Online]. Available: <https://doi.org/10.1007/s12243-018-0649-0>
- [12] L. Corneo *et al.*, "Surrounded by the clouds: A comprehensive cloud reachability study," in *Proceedings of the Web Conference 2021*, ser. WWW '21. New York, NY, USA: Association for Computing Machinery, 2021, p. 295–304. [Online]. Available: <https://doi.org/10.1145/3442381.3449854>
- [13] T. K. Dang *et al.*, "Cloudy with a chance of short rtt: Analyzing cloud connectivity in the internet," in *Proceedings of the 21st ACM Internet Measurement Conference*, ser. IMC '21. New York, NY, USA: Association for Computing Machinery, 2021, p. 62–79. [Online]. Available: <https://doi.org/10.1145/3487552.3487854>
- [14] L. Corneo *et al.*, "(how much) can edge computing change network latency?" in *2021 IFIP Networking Conference (IFIP Networking)*, 2021, pp. 1–9.
- [15] "Kubernetes," <https://kubernetes.io/>.
- [16] "Lightweight kubernetes — k3s," <https://k3s.io>.
- [17] "Kubefed," <https://github.com/kubernetes-sigs/kubefed>.
- [18] J. Liu *et al.*, "Delay-optimal computation task scheduling for mobile-edge computing systems," in *IEEE ISIT*, 2016.
- [19] L. Tianze *et al.*, "An overhead-optimizing task scheduling strategy for ad-hoc based mobile edge computing," *IEEE Access*, 2017.
- [20] A. Brogi *et al.*, "How to place your apps in the fog: State of the art and open challenges," *Software: Practice and Experience*, 2020.
- [21] T. Shreedhar *et al.*, "Qaware: A cross-layer approach to mptcp scheduling," in *2018 IFIP Networking Conference (IFIP Networking) and Workshops*, 2018, pp. 1–9.
- [22] N. Mohan *et al.*, "Anveshak: Placing edge servers in the wild," in *Proceedings of the 2018 Workshop on Mobile Edge Communications*, ser. MECOMM'18. New York, NY, USA: Association for Computing Machinery, 2018, p. 7–12. [Online]. Available: <https://doi.org/10.1145/3229556.3229560>
- [23] "Why large organizations trust kubernetes," <https://tanzu.vmware.com/content/blog/why-large-organizations-trust-kubernetes>.
- [24] A. Jeffery *et al.*, "Rearchitecting kubernetes for the edge," *4th ACM EdgeSys*, 2021.
- [25] "Kubeedge," <https://github.com/kubeedge/kubeedge>.
- [26] "Microk8s," <https://microk8s.io>.
- [27] S. Böhm and G. Wirtz, "Profiling lightweight container platforms: Microk8s and k3s in comparison to kubernetes." in *ZEUS*, 2021.
- [28] S. H. Mortazavi *et al.*, "Cloudpath: A multi-tier cloud computing framework," in *ACM/IEEE SEC*, 2017.
- [29] W. Zhang *et al.*, "Hetero-edge: Orchestration of real-time vision applications on heterogeneous edge clouds," *IEEE INFOCOM*, 2019.
- [30] H. P. Sajjad *et al.*, "Spanedge: Towards unifying stream processing over central and near-the-edge data centers," in *2016 IEEE/ACM SEC*, 2016.
- [31] "Foglamp – simplifying iiot data management from sensors to clouds," <https://dianomic.com/platform/foglamp/>.
- [32] Q. Liu and T. Han, "Virtualedge: Multi-domain resource orchestration and virtualization in cellular edge computing," in *ICDCS*. IEEE, 2019.
- [33] Y. Kim *et al.*, "Collaborative task scheduling for iot-assisted edge computing," *IEEE Access*, 2020.
- [34] C. Ciconetti *et al.*, "A decentralized framework for serverless edge computing in the internet of things," *IEEE Transactions on Network and Service Management*, 2020.
- [35] J. Edinger *et al.*, "Decentralized low-latency task scheduling for ad-hoc computing," in *2021 IEEE IPDPS*, 2021.
- [36] M. Aljarah *et al.*, "Cooperative hierarchical based edge-computing approach for resources allocation of distributed mobile and iot applications," *IJECE*, 2020.
- [37] F. A. Salaht *et al.*, "An overview of service placement problem in fog and edge computing," *ACM Computing Surveys (CSUR)*, 2020.
- [38] A. Zavodovski *et al.*, "Decentralizing computation with edge computing: Potential and challenges," in *Proceedings of the Interdisciplinary Workshop on (de) Centralization in the Internet*, ser. IWCI'21. New York, NY, USA: Association for Computing Machinery, 2021, p. 34–36. [Online]. Available: <https://doi.org/10.1145/3488663.3493689>
- [39] A. Zavodovski *et al.*, "Exec: Elastic extensible edge cloud," in *Proceedings of the 2nd International Workshop on Edge Systems, Analytics and Networking*, ser. EdgeSys '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 24–29. [Online]. Available: <https://doi.org/10.1145/3301418.3313941>
- [40] A. Zavodovski *et al.*, "Open infrastructure for edge: A distributed ledger outlook," in *2nd USENIX Workshop on Hot Topics in Edge Computing (HotEdge 19)*. Renton, WA: USENIX Association, Jul. 2019. [Online]. Available: <https://www.usenix.org/conference/hotedge19/presentation/zavodovski>
- [41] "Kubernetes components," Mar 2021. [Online]. Available: <https://kubernetes.io/docs/concepts/overview/components/>
- [42] "Eclipse iofof," <https://iofof.org/>.

- [43] N. Mohan *et al.*, "Managing data in computational edge clouds," in *Proceedings of the Workshop on Mobile Edge Communications*, ser. MECOMM '17. New York, NY, USA: Association for Computing Machinery, 2017, p. 19–24. [Online]. Available: <https://doi.org/10.1145/3098208.3098212>
- [44] B. P. Rimal *et al.*, "Architectural requirements for cloud computing systems: an enterprise cloud approach," *Grid Computing*, 2011.
- [45] B. Zhang *et al.*, "The cloud is not enough: Saving iot from the cloud," *7th USENIX Workshop HotCloud*, 2015.
- [46] H. Cui *et al.*, "Cloud service reliability modelling and optimal task scheduling," *Iet Communications*, 2017.
- [47] T. Shreedhar *et al.*, "An age control transport protocol for delivering fresh updates in the internet-of-things," in *IEEE WoWMoM*, 2019.
- [48] T. Shreedhar *et al.*, "Acp: Age control protocol for minimizing age of information over the internet," in *Proceedings of the 24th Annual International Conference on Mobile Computing and Networking*, ser. MobiCom '18. New York, NY, USA: Association for Computing Machinery, 2018, p. 699–701. [Online]. Available: <https://doi.org/10.1145/3241539.3267740>
- [49] T. Shreedhar *et al.*, "Coexistence of age sensitive traffic and high throughput flows: Does prioritization help?" 2022. [Online]. Available: <https://arxiv.org/abs/2203.00647>
- [50] T. Shreedhar *et al.*, "An empirical study of ageing in the cloud," in *IEEE INFOCOM 2021 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, 2021, pp. 1–6.
- [51] A. Kosta *et al.*, "Age of information: A new concept, metric, and tool," *Foundations and Trends in Networking*, 2017.
- [52] M. C. Silva Filho *et al.*, "Approaches for optimizing virtual machine placement and migration in cloud environments: A survey," *Journal of Parallel and Distributed Computing*, 2018.
- [53] F. Dabek *et al.*, "Vivaldi: A decentralized network coordinate system," *ACM SIGCOMM Computer Communication Review*, 2004.
- [54] A. Zavodovski *et al.*, "edisco: Discovering edge nodes along the path," *CoRR*, vol. abs/1805.01725, 2018. [Online]. Available: <http://arxiv.org/abs/1805.01725>
- [55] T. Zhu *et al.*, "Task scheduling in deadline-aware mobile edge computing systems," *IEEE Internet of Things Journal*, 2018.
- [56] S. Yi *et al.*, "Lavea: Latency-aware video analytics on edge computing platform," in *ACM/IEEE SEC*, 2017.
- [57] T. Bourke, *Server load balancing*. "O'Reilly Media, Inc.", 2001.
- [58] D. King and A. Farrel, "A Survey of Semantic Internet Routing Techniques," Internet Engineering Task Force, Internet-Draft draft-king-irtf-semantic-routing-survey-03, Nov. 2021.
- [59] N. Mohan *et al.*, "Redesigning mptcp for edge clouds," in *Proceedings of the 24th Annual International Conference on Mobile Computing and Networking*, ser. MobiCom '18. New York, NY, USA: Association for Computing Machinery, 2018, p. 675–677. [Online]. Available: <https://doi.org/10.1145/3241539.3267738>
- [60] T. Shreedhar *et al.*, "A longitudinal view at the adoption of multipath tcp," 2022. [Online]. Available: <https://arxiv.org/abs/2205.12138>
- [61] F. Aschenbrenner *et al.*, "From single lane to highways: Analyzing the adoption of multipath tcp in the internet," in *2021 IFIP Networking Conference (IFIP Networking)*, 2021, pp. 1–9.
- [62] "Raspberry pi 4," <https://www.raspberrypi.org/products/raspberry-pi-4-model-b/specifications/>.
- [63] "Nvidia jetson agx xavier," <https://www.nvidia.com/en-us/autonomous-machines/jetson-agx-xavier/>.
- [64] "Eclipse fog05," <https://fog05.io/>.
- [65] G. Ananthanarayanan *et al.*, "Real-time video analytics: The killer app for edge computing," *computer*, 2017.
- [66] S. Bäurle and N. Mohan, "Comb: A flexible, application-oriented benchmark for edge computing," in *Proceedings of the 5th International Workshop on Edge Systems, Analytics and Networking*, ser. EdgeSys '22. New York, NY, USA: Association for Computing Machinery, 2022, p. 19–24. [Online]. Available: <https://doi.org/10.1145/3517206.3526269>
- [67] T. Chavdarova *et al.*, "WILDTRACK: A Multi-camera HD Dataset for Dense Unscripted Pedestrian Detection," in *IEEE CVPR*, 2018.
- [68] J. A. Donenfeld, "WireGuard," <https://www.wireguard.com/>.