Paper Review: The Emerging Landscape of Edge-Computing

Tanjila Ahmed

NetLab Group Meeting

Friday, February 21, 2020

https://www.microsoft.com/en-us/research/uploads/prod/2020/02/GetMobile__Edge_BW.pdf

Authors

S. A. Noghabi, L. Cox, S. Agarwal, and G. Ananthanarayanan

Mobility and Networking Group

Microsoft Research

March 2020

https://www.microsoft.com/en-us/research/uploads/prod/2020/02/GetMobile__Edge_BW.pdf

Motivation

Edge Computing

Enterprise-driven implementation:

on-premise, single-tenant edges with shared, redundant outbound links

limited bandwidth & un-reliability of the link to the cloud

Initial Vision

Multi-tenant resource that will be used opportunistically for **low-latency** mobile applications

Edge Deployment Today: Enterprise Driven

PRIMARY place of storage & computation. **IF** network conditions allow, the cloud is opportunistically used along- side

This Study

Enterprise deployment driving Innovation in Edge Computing

Cyber Foraging

Interactive applications for lightweight devices can benefit from by opportunistically launching short-lived, low-latency jobs *to more powerful machines over a low-latency, high-bandwidth network*

Example

a cognitive-assistance application on a wearable computer could overlay real-time guidance on a display by streaming video to an inference model running on a nearby tensor-processing unit (TPU)

Edge-Sites: Why do these applications run on the edge rather than offloading to the cloud?

- Long-running applications: industrial sensing, video analytics, ML inferences
- Single tenant, & they rarely host transient jobs for mobile devices

Example

1. **Restaurant: customer arrivals & schedule meal preparation (10s of sec)**
2. **Oil rig: process live video to identify safety hazards (100s of msec)**

Not as strict latency requirements as cyber-foraging applications: cognitive assistance (must process video frames in 10ms or less)!

With such relatively high latency tolerance, & high availability, scalability, & low-cost computation offered by the cloud

Why do these applications run on the edge rather than offloading to the cloud?

- Mission- and safety-critical applications, any downtime is unacceptable
- *Need to tolerate cloud outages and the scarcity of network bandwidth*

Example

1. **A restaurant must serve food & an oil rig must prevent injuries, even when they are disconnected from the cloud**
2. **An offshore oil rig analyzing dozens of 8K video feeds may have hundreds of ms to prevent an accident, but it would be infeasible (and ex-pensive) for this application to transfer video frames to the cloud within that time (much less analyze the video)**

Objective

Highlight how edge-sites are used by critical applications to survive transient network disruptions & how their structure differs from the cyber-foraging model

Real World Deployments of Edge Computing

Industry		Company	Use case	Edge location
Business	Restaurants	Chick-fil-A	– forecast food preparation (e.g., more food needs to be fried)	In store
	Retail	Walmart, Coca Cola (vending machines)	– monitoring (e.g., fridge temperature ensuring produce quality) – tracking customers & improving sales (e.g., customized coupons)	In store
	Gas station	Shell	– detect safety hazards (e.g., a person smoking a cigarette) across their 44,000 gas stations	In gas stations
Smart Cities	Cities	City of Bellevue	– traffic administration (e.g., intelligent control of traffic light) – safety at intersections (e.g., alerting drivers to prevent accidents)	Intersections & City clusters
	Construction	PCL, ATF Services	– increase safety, efficiency, and productivity (e.g., detecting a temperature spike or gas leak in a unit) – increase security of construction cites (e.g., protecting equipment over night)	Construction site

Business Intelligence

Chick-fil-A restaurant chain has an IoT application to forecast when more food needs to be fried

- In-restaurant prediction platform that relies on edge computing
- Video analytics and machine learning to predict the number of customers and cars entering the store

Smart Cities

City of Bellevue uses cameras at traffic intersections for both controlling the traffic flow across the city and alerting drivers to avoid fatal accidents (e.g., a bicyclist approaching on the right)

With the deployment of HD cameras, the data volume can be multiple Mbps for a single video stream. In city deployments, there is often sufficient bandwidth for all these streams to reach the edge (e.g., at the local traffic control center), but not beyond that to reach the cloud.

Real World Deployments of Edge Computing

Industry		Company	Use case	Edge location
Transportation	Aviation	Airbus, Bombardier	– analyze in-flight experience of customers – monitor aircraft operations and maintenance	On Plane
	Railway	CAF	– monitor train tracks, freight cars, and wheels for problems that lead to derailment	On the train
	Road Control	Alaska DOT	– monitor quality of roads and detect roads with need of maintenance (e.g., finding spots that need snow plowing to prevent icing)	On Trucks
Industrial Plants	Oil Refinary	Schneider Electric, ExxonMobil	– predictive maintenance (of the pumps and equipment) – workplace safety	Oil rig or pump
	Manufacturing	GE, CPG, DAI-HEN, Airbus	– improve manufacturing yields (e.g., automation or detecting defected products) – monitor equipment & predict need for maintenance	In factory
	Manufacturing	BMW	– manage fleet of robots aiding in production pipeline	In factory
	Agriculture	Buhler	– control quality of produce at harvest, storage, and processing using imagery (e.g., for grains, processing 20,000 kernels/s).	In field
	Agriculture	DroneWorks, FarmBeats	– observe and monitor agricultural fields using sensors and drone imagery (e.g., detect areas that need water or pesticides)	In field

Intelligent Transportation

The railway industry uses high-definition cameras in bungalows along the track, to detect cracks in train wheels. Cracks can cause the wheel to break and derail the entire train

Bandwidth demand for this case is dynamic. When a train passes a bungalow, it will generate GBs of data over a small period. At the same time, cracks should be detected and reported reliably within minutes to avoid severe casualties (financial and human lives)

Industrial Plants

ExxonMobile continuously monitors its multi-million dollar oil rigs. They use edge computing at each rig to detect equipment maintenance needs. They use this solution across their 11 million acres of land with ≈50,000 producing wells.

They cut the need for manual inspection across large fields and reduce the reaction time to unexpected issues. The motivation for using the edge is the unreliable and low-bandwidth satellite connectivity available in such areas

Edge-site Deployment Architecture

A dedicated set of edges are used as a primary place of storage and computation. If the network permits, the cloud is also used alongside the edge-site

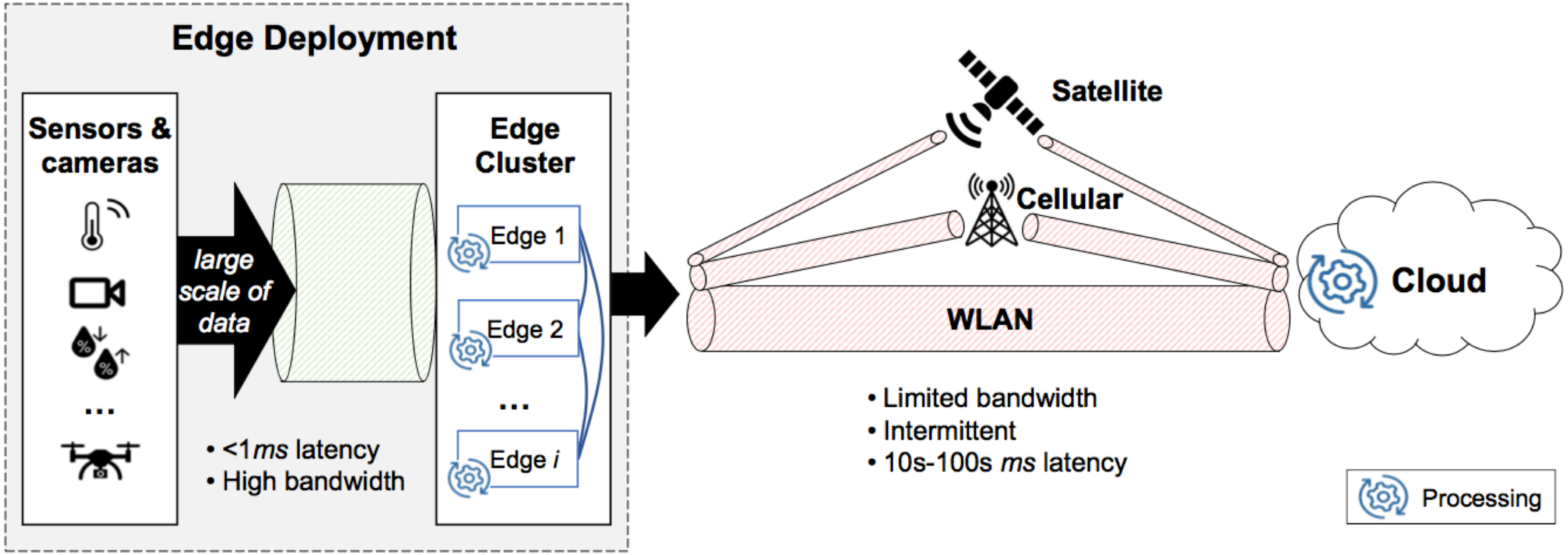

Figure 1: **Typical edge-site deployment architecture in enterprise environments.**

Input Devices

A retail store will consist of tens to hundreds of cameras, tracking items picked by customers along with several other sensors for humidity, temperature, etc

- Always have good connectivity to a nearby edge cluster
- Do not face networking issues in migrating between edges

Edge compute

Run both fridge monitoring and customer tracking applications in a retail store. City running traffic control and accident prevention on a shared edge cluster at the intersection

- Edges in an edge-site deployment are owned by & dedicated to one enterprise, i.e., edges are not shared among various enterprises or tenants.
- Multiple applications of a single enterprise share the edge cluster

Connectivity to cloud:

- **Outlinks:** Edge-site deployment is connected to the cloud via a number of network links
- Outlinks are shared among all input devices, edges, and applications within a deployment
- Multiple outlink connectivity options: cellular & satellite (back-up)

Edge Deployments for Enterprise Environments (edge-site) vs. Mobile-Computing (Cyber Foraging)

Characteristic	Cyber foraging deployment	Edge-site deployment
Input devices	mobile phones, wearables, etc.	cameras, sensors, etc.
Edges	resource rich computer or cluster of computers	diverse hardware (Raspberry pi to rack of machines)
Outlink to cloud	i. single-link, ii.limited bandwidth	i. single- or multi-link, ii.limited bandwidth
Applications	i.high-volume, ii.short-running, iii.best-effort, iv.interactive (*ms* latency)	i.high-volume, ii.long-running, iii.critical, iv. latency-tolerant (hundreds of *ms* – *mins*)
Multi-tenancy	multi-tenant, multi-application	single-tenant, multi-application
Mobility	*input devices*: mobile *edges*: stationary	*input devices*: mobile in a boundary or stationary *edges*: stationary
Ownership & Management	*input devices*: consumers *edges*: 3rd party providers	*input devices*: enterprise *edges*: (typically) enterprise
Energy	*input devices*: battery-powered *edges*: plugged-in	*input devices*: battery-powered & plugged-in *edges*: plugged-in

Future Directions for Research on Edge-Sites

- Edge-sites helps mission-critical applications remain operational in transient changes to cloud-edge connectivity and workload variations
- However, current system support for developing, managing, and scheduling these applications is poor, in a way that they use both the edge and cloud effectively

Graceful Adaptation of Applications

- Adapt gracefully in the presence of disconnections, drops in bandwidth, or workload spikes
- In the cyber foraging model, applications adapt to changes in connectivity to an edge by simply using another nearby edge, and if not possible, the cloud with presumed infinite-resource

Graceful Adaptation of Applications

Cyber foraging model: knowledge of application logic not required

Edge-site: appropriate adaptation strategies are application specific and can vary significantly.

After a change to network conditions, an application could adjust the

a) ***input-data* quality**, e.g., reducing the resolution of videos or increasing sensor reading periods
b) ***compute* quality**, e.g., using larger batch sizes or larger aggregation windows, or
c) ***output-data* quality**, e.g., compressing more aggressively or filtering/discarding parts of the data

Collaborative & App-aware Network Orchestration

Cyber Foraging:

- Multiple untrusted user devices
- Building performance & security isolation between tenants on edges

Edge-sites:

- a collaborative, single-tenant environment with no adversarial application trying to greedily use the entire link
- Instead of a traditional network orchestration mechanism dividing link capacity, there is an opportunity for a new collaborative approach that leverages applications' knowledge of how to adapt

Collaborative & App-aware Network Orchestration

Cyber Foraging: mobile devices have intermittent connectivity to the edge and the edge would have stable connectivity to the cloud

Edge-sites:

- Input devices have stable connectivity to the edge and the edge has intermittent connectivity to the cloud

Better managing & prioritizing the connectivity of edge applications to cloud

Test and Verification Frameworks

- Adding adaptation logic to applications in the edge-site model increases the application complexity. This differs from the cyber foraging model where the complexity of adaptation is mostly placed on the edge infrastructure,

1. *Simplify choosing the right adaptation strategies for developers and to verify their choices*
2. *Build frameworks that can test and verify the correctness of adaptation decisions under various conditions:* length of disconnection, bandwidth values, and the set of other competing applications

Conclusion

- This paper challenges long-held predictions on why and how edge computing will be deployed
- Not end-user interactive mobile applications opportunistically using the edge as originally envisioned. Rather, they are geographically constrained, mission-critical, industrial or enterprise applications that primarily rely on the edge and opportunistically use the cloud
- Not motivated by the need for low latency access to the cloud, but rather by the lack of sufficient and reliable bandwidth to the cloud
- Outlined a number of interesting research challenges that need to be solved in this context