Benchmarking In The Dark: On The Absence of Comprehensive Edge Datasets

Oleg Kolosov, Gala Yadgar
Technion

Sumit Maheshwari, Emina Soljanin
Rutgers University
MOTIVATION

Need a workload

Important for system research, design, and optimization
  - Define system design objectives
  - Identify optimization goals
  - Make appropriate tradeoffs
  - Evaluate and compare

Use case: Design and evaluation of an edge-based storage service

- Multiple providers
- Considerable heterogeneity
  *Optimization not trivial*

Susceptive to fluctuations
EXISTING WORKLOADS

- Existing data center workloads rarely reflect
  - Edge infrastructure
  - Edge application requirements
- In existing edge papers:
  - Some aspects are irrelevant
  - Some aspects can be modeled by general datasets
  - Some examples:

  **Applications**
  - App data is easy to obtain (HotEdge ’18, HotEdge ’19)

  **Security & Privacy**
  - System (SEC ’16, GLOBECOM ’17) and data (IEEE IRI ’14, GLOBECOM ’16) are trivial

  **Mobility**
  - Geolocation data is easy to obtain (TON Vol.25, SEC ’17)

  **Infrastructure**
  - System dataset is trivial, synthetic workloads are used (ICDCS ‘17, MECOMM ‘17)

  **Our use case is focused on storage**
  - Key aspects aren’t trivial

  **There are no operational edge systems that can provide the desired workload**

  **Small number of deployed real edge systems**
DATASETS AND ATTRIBUTES

The datasets we need:

<table>
<thead>
<tr>
<th>Storage</th>
<th>Compute</th>
<th>User/App.</th>
<th>Location</th>
<th>Architecture</th>
<th>Availability</th>
</tr>
</thead>
<tbody>
<tr>
<td>FIU, Umass, MSR</td>
<td></td>
<td></td>
<td>ECMWF, UBC, FSL</td>
<td>FB, SNAP, Alexa</td>
<td></td>
</tr>
<tr>
<td>Austin, NYC, SFO</td>
<td></td>
<td></td>
<td>BORG, Azure, LANL</td>
<td>RIPE, CAIDA</td>
<td></td>
</tr>
</tbody>
</table>

< Data Object, Time, Location, Node >
The datasets we have:

<table>
<thead>
<tr>
<th>Datasets/Attributes</th>
<th>Storage</th>
<th>Compute</th>
<th>User/App.</th>
<th>Location</th>
<th>Architecture</th>
<th>Availability</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Storage workloads</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FIU, Umass, MSR...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>FS snapshots</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ECMWF, UBC, FSL</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Object Popularity</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FB, SNAP, Alexa...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Mobility</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Austin, NYC, SFO</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Cluster</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BORG, Azure, LANL...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Network Arch.</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RIPE, CAIDA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Device failures</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Backblaze</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
WORKLOAD COMPOSITION

How to bridge the gap?
- Join attributes from several available datasets

NYC Hotspots
NYC Taxi Zones
NYC Yellow Taxis Trip Data

< Data Object, Node, Location, Time >

- Taxi drop-offs represent demand in a zone
- A ‘browsing session’ starts at a drop-off time and zone
- Starts at drop-off node_{h} - Random hotspot from the drop-off zone

Use case: Design and evaluation of an edge-based storage service
The ‘browsing session’

\[
\text{Trace of GET requests: } \langle \text{Data Object, Node, Location, Time} \rangle \text{ for } 0 \leq i < n. \quad \varepsilon \text{ – request rate within a session.}
\]
CHARACTERIZING THE SYSTEM AND ITS USERS

- The workloads are lightly correlated
- The workload composition is not random
GENERALIZATION

User Requests Across NYC

Alternatives

NYC Hotspots

NYC Taxi Zones

Refinements

Wikipedia Pages

NYC Yellow Taxis Trip Data

System Arch.

#Sessions / Arrival Times

Any Trace of Object Requests

Requests with Location

Finer Location Granularity

Subway Station Exists
SUMMARY

- Conclusions
  - The problem is not unique for this specific case (general problem)
  - Described important categories of attributes
  - Showed how partial datasets can be used to compose a workload

- Discussion
  - Is the absence of datasets really temporary?
  - Which basic workloads to use?
  - How can we leverage synthetic distributions?
  - How to generate realistic and useful compositions?

Thank you