Characterizing JSON Traffic Patterns on a CDN

Santiago Vargas, Utkarsh Goel, Moritz Steiner, Aruna Balasubramanian
JSON Traffic is Growing on Akamai’s CDN

- JSON is 4x more requested than HTML
- JSON is the leading content type on Akamai’s CDN
What is JSON?

JSON is a text-based data format

1. Key-value Pairs

```
{
  first_name: 'Paul',
  surname: 'Miller',
  cell: 447557505611,
  city: 'London',
  location: [45.123, 47.232],
  profession: ['banking', 'finance', 'trader'],
  cars: [
    { model: 'Bentley',
      year: 1973,
      value: 1000000, ... },
    { model: 'Rolls Royce',
      year: 1965,
      value: 330000, ... }
  ]
}
```

2. Lightweight

Fields can contain arrays

3. Structured

Fields can contain an array of sub-documents

String

Number

Geo-Coordinates

Typed field values

Fields can contain an array of sub-documents
Motivation

• Little is known about JSON usage
• Optimizations exist for types of content, i.e. browsing and media content. Unsure if these apply to JSON content
Akamai’s Network as a Vantage Point

• Collect HTTP request logs from Akamai edge servers
  • Data Fields: Mime Type, Time of Request, User-Agent, Caching Information, HTTP Method, URL, Anonymized IP

• Collect 2 datasets:

<table>
<thead>
<tr>
<th>Dataset</th>
<th># of Logs</th>
<th>Duration</th>
<th># of Domains</th>
</tr>
</thead>
<tbody>
<tr>
<td>Short-term</td>
<td>25 million</td>
<td>10 mins</td>
<td>~5k</td>
</tr>
<tr>
<td>Long-term</td>
<td>10 million</td>
<td>24 hours</td>
<td>~170</td>
</tr>
</tbody>
</table>

• Short-term – wide network coverage for overall characterization
• Long-term – wide temporal coverage for pattern characterization
Extracting Our Data

- Develop a taxonomy to look at JSON traffic on different dimensions:

<table>
<thead>
<tr>
<th>JSON Traffic Taxonomy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source</td>
</tr>
<tr>
<td>Embedded Device, App, Browser, or Unknown</td>
</tr>
<tr>
<td>Human Triggered or Machine Generated</td>
</tr>
<tr>
<td>Request</td>
</tr>
<tr>
<td>Upload or Download</td>
</tr>
<tr>
<td>Response</td>
</tr>
<tr>
<td>Cacheable vs. Uncacheable</td>
</tr>
<tr>
<td>Size</td>
</tr>
</tbody>
</table>
Extracting Our Data

• Develop a taxonomy to look at JSON traffic on different dimensions:

<table>
<thead>
<tr>
<th>JSON Traffic Taxonomy</th>
<th>Source</th>
<th>Request</th>
<th>Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source</td>
<td>Embedded Device, App, Browser, or Unknown</td>
<td>Human Triggered or Machine Generated</td>
<td>Upload or Download</td>
</tr>
</tbody>
</table>

| | Cacheable vs. Uncacheable | Size |
|--------------------------|---------------------------------|

• Device/Application
• Machine-Generated
• Request Type
• Response Type

User Agent
Pattern Based
HTTP Method
Cache Labels
What Devices Request JSON?

- Majority of JSON is from mobile smartphones and embedded devices (>64%)
  - Embedded devices include game consoles, wearables, and smart appliances
- 88% of JSON is non-browser
Identifying Machine Traffic

- Autocorrelation techniques identify 6% of JSON traffic is requested periodically.

- Objects are requested in standard periods:
  - 30s
  - 1m
  - 2 & 3m
  - 10m
  - 15m
  - 30m

- Machine-to-machine traffic:
  - 78% upload traffic & 56% uncacheable
  - One optimization avenue: Deprioritize machine traffic that is not QoE sensitive
Does Caching Help?

- Majority (84%) of requests are downloads making them ideal for caching
  - 55% of requests are uncacheable
  - 50% of domains don’t use caching for JSON
Does Caching Help?

- Majority (84%) of requests are downloads making them ideal for caching
  - 55% of requests are uncacheable
  - 50% of domains don’t use caching for JSON

Since caching doesn’t help JSON content delivery for specific industries, can other optimizations help?
Using Dependencies as avenue for optimization

- Dependencies Example:

  - Methodology: Cluster objects with similar URLs

  - Results: 90% accuracy using ngram model for clustered URLs

  Requests for JSON
  Object A then B then C

*Question: Given request for A, can we predict object B will be requested next?*

- Methodology: Cluster objects with similar URLs

  a.com/abc123/profile
  a.com/def456/profile

  a.com/*/profile

  Clustered URL

- Results: 90% accuracy using ngram model for clustered URLs

Analyzing these patterns can inform Prefetching, Server Push, Anomaly Detection systems

*when considering top 10 predictions*
Conclusions

• JSON is becoming a leading content type on the Web

• Majority of JSON is from mobile & embedded devices and is non-browser

• Avenues for optimization:
  • Periodic machine-to-machine traffic can be deprioritized
  • JSON has ordering patterns that can be predicted