EYEORG: Crowdsourcing Web QoE measurements

M. Varvello, J. Backburn, D. Naylor*, D. Papagiannaki
Telefonica, *CMU
CoNEXT 2016
Web QoE Matters

- 1 second slowdown
  - ↓ $1.6 Billion in sales per year
- 0.4 second slowdown
  - ↓ 8 Million searches per day
Much work on improving Page Load Time

<table>
<thead>
<tr>
<th>Research</th>
<th>Standards</th>
<th>CDNs et al.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shandian [NSDI’16]</td>
<td>SPDY [Google]</td>
<td>Level3</td>
</tr>
<tr>
<td>Klotski [NSDI’15]</td>
<td>HTTP/2 [IETF]</td>
<td>CloudFlare</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Conviva</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CacheFly</td>
</tr>
</tbody>
</table>

Measuring PLT is important to evaluating new techniques.
PLT usually measured on OnLoad

OnLoad may overestimate
Include objects users may not care about (ads?)

OnLoad may underestimate
Deferred scripts may load objects after
How do we measure user-perceived PLT?

Challenges

– Consistent experience
  • Participants have different software and network conditions

– Quantitative responses
  • It’s hard to express when a page “seems loaded”

– Trustworthy results
  • Crowd workers are not always reliable
Consistent experience

- Participants’ network condition impact their responses

It was fast!

It was slow!

a.com.ar
Consistent experience

- Videos of pages loading look the same to all

Capture videos in advance

Serve videos, not sites, during tests
Challenges

- **Consistent experience**
  - Participants have different software and network conditions

- **Quantitative responses**
  - It’s hard to express when a page “seems loaded”

- **Trustworthy results**
  - Crowd workers are not always reliable
Two types of tests

- Timeline
  - When does the page looks “ready to use”? 

- A/B testing
  - Which version loaded faster?
Timeline

- ‘Scrub bar’ rater than standard HTML5 control
- Preload video – The issues is not the video
- Frame rewind – When user submits, offer the earliest similar frame to correct for overshooting

Drag the slider to scrub through the video until the page appears “ready to use”
A/B Testing

- Head-to-head comparison – No need to decide precise PLT
- Single video – So A and B never get out of sync
- Random order A is not always left, B not always right
Challenges

- Consistent experience
  - Participants have different software and network conditions

- Quantitative responses
  - It’s hard to express when a page “seems loaded”

- Trustworthy results
  - Crowd workers are not always reliable
Eyeorg filters responses

- Filters based on HCI literature
  1. Control questions
     - Randomly inserted with known answer
     - For timeline – suggest a nearly-blank rewind frame
     - For A/B – same video, 3 secs delayed
  2. Engagement
     - Time on it, seek, switch tabs?
  3. Soft rules
     - Do they watch the video at all?
  4. Wisdom of the crowd
     - Average as ground truth

Evaluation campaign
- 100 crowdsourced workers
- 100 trusted participants as ground truth
- 20 sites from Alexa top 1M
Wisdom of the crowd

- Looking for a consensus rule

UserPerceivedPlt of four representative videos

Removing outliers – stdev as measure of agreement over the whole set

Rule: Keep 25-75th percentiles of responses for timeline
Run 3 measurement campaigns

- PLT metrics
  - How well do existing metrics capture user-perceived PLT?
- HTTP/1.1 vs HTTP/2
  - Can users perceive a diff?
- Ad blockers
  - What is the impact of ad blockers on user perception?
Timeline tests to compare PLT

For each site, measure PLT 5 ways

1. **OnLoad** (from HAR)
   Time it takes for the JavaScript “onLoad” event to fire; once before the page’s embedded resources have been downloaded

2. **First Visual Change**
   First pixel is drawn

3. **Last Visual Change**
   Last pixel stops changing

4. **SpeedIndex** (from video)
   Average time at which visible parts (above-the-fold) of the page are displayed

5. **User-Perceived PLT** from eyeorg

---

**PLT metric campaign**

- 1000 crowdsourced workers
- 100 sites from Alexa top 1M
- $120 total cost
- 1.5 days to collect responses
OnLoad and FVC correlate best with UPLT

- **OnLoad**
  - Correlation: 0.85
  - Graph showing a strong positive correlation between Mean ULTP (sec) and OnLoad.

- **LastVisualChange**
  - Correlation: 0.47
  - Graph showing a weaker correlation between Mean ULTP (sec) and LastVisualChange.

- **FirstVisualChange**
  - Correlation: 0.84
  - Graph showing a strong positive correlation between Mean ULTP (sec) and FirstVisualChange.

- **SpeedIndex**
  - Correlation: 0.68
  - Graph showing a moderate positive correlation between Mean ULTP (sec) and SpeedIndex.

**Higher complexity, but not better results**

Additional observations:
- Despite its higher complexity, OnLoad has a lower correlation (0.85) compared to FirstVisualChange (0.84).
- SpeedIndex provides a moderate correlation (0.68), suggesting it may be more useful in certain contexts.
- LastVisualChange has the lowest correlation (0.47), indicating it may be less reliable as a metric for page load time.

For visibility, we only show results for the first 20 videos used for validation in Section 4. For most videos, the frame selection helper's suggestions for fine-tuning their choices ("submitted") match their final state ("slider") over 7% of responses. On average, the submitted videos' suggestions exactly match the value proposed by the frame selection helper ("slider"). The correlation of 0.68 indicates a strong relationship between the two metrics.
How close are the actual values?

OnLoad within 1 sec of UPLT

Within 200msec for 60% of sites

Within 100msec for 30% of sites
Discussion

- What does “ready” mean?
  - Intentionally left as an open question
  - Looking at UPLT, some easy-to-tell sites and some with wider gap between FVC and LVC or including auxiliary content

- Using a video means no site interaction … for many users, the criteria to decide

- …