A Platform for Crowdsourcing Web Quality of Experience Measurements
Web quality of experience matters a lot

**Amazon**
- 1 second slowdown
- ▼$1.6 Billion in sales per year

**Google**
- 0.4 second slowdown
- ▼8 Million searches per day
A lot of people are working to improve page load time (PLT)

**RESEARCH**
- Polaris [NSDI '16]
- Shandian [NSDI '16]
- Klotski [NSDI '15]

**STANDARDS**
- QUIC [Google]
- SPDY [Google]
- HTTP/2 [IETF]

**CDNs**
- Akamai
- Level 3
- CloudFlare
- Limeligh
- CacheFly
- MaxCDN
- Instart Logic
- Speedera
- EdgeCast
- Aryaka
- Incapsul
- Aryaka
- ...
Measuring PLT is important for evaluating new technologies.

CONTENT OPTIMIZATION
- Domain sharding
- Image Spriting

DELIVERY TECHNIQUES
- QUIC, HTTP/2
- Polaris [NSDI ‘16]
- Shandian [NSDI ‘16]
- Klotski [NSDI ‘15]
PLT is usually measured with *OnLoad*
OnLoad might not reflect *user-perceived* PLT

- **OnLoad might overestimate**
  Includes objects user might not care about (e.g., ads)

- **OnLoad might underestimate**
  Deferred scripts might load objects after OnLoad
How do we measure *User-Perceived* Page Load Time?
A platform for crowdsourcing Web quality of experience measurements.
Challenges

1. **Consistent experience**
   Participants have different software and network conditions

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

3. **Trustworthy results**
   Crowd workers are not always reliable
Challenges

1. Consistent experience
   Participants have different software and network conditions

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

3. Trustworthy results
   Crowd workers are not always reliable
Participants’ network connections impact their responses

- Fast connection: “It was fast!"
- Slow connection: “It was slow!”

foo.com
Videos of pages loading look the same to everyone

Capture videos in advance

foo.com

foo.webm

Serve videos, not sites, during tests

FAST

SLOW
Challenges

1. **Consistent experience**
   Participants have different software and network conditions

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

3. **Trustworthy results**
   Crowd workers are not always reliable
Challenges

1. Consistent experience
   Participants have different software and network conditions

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

3. Trustworthy results
   Crowd workers are not always reliable
We designed two types of test

**Timeline**
When does the page look “ready to use”?

**A/B**
Which version loaded faster?
Timeline
When does the page look “ready to use”?

Drag the slider to scrub through the video until the page appears “ready to use.”
Timeline
When does the page look “ready to use”?

“Scrub bar”
Rather than standard HTML5 video controls

Preload the video
To avoid “is the page in the video still loading, or is the video itself still loading?”

Frame rewind
When user submits, offer the earliest similar frame to correct for overshooting
We designed two types of test

**Timeline**
When does the page look “ready to use”?

**A/B**
Which version loaded faster?
A/B
Which version loaded faster?

1 Play the videos.

2 Select which video (“Left” or “Right”) loaded faster or choose “No Difference.”
A/B
Which version loaded faster?

Head-to-head comparison
No need to decide precise PLT; simpler to just choose winner

Single video
So A and B never get out of sync

Random order
A is not always left, B is not always right

1. Play the videos.
2. Select which video (“Left” or “Right”) loaded faster or choose “No Difference.”
We designed two types of test

**Timeline**
When does the page look “ready to use”? 

**A/B**
Which version loaded faster?
Challenges

1. **Consistent experience**
   Participants have different software and network conditions

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

3. **Trustworthy results**
   Crowd workers are not always reliable
Challenges

1. **Consistent experience**
   Participants have different software and network conditions

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

3. **Trustworthy results**
   Crowd workers are not always reliable
Eyeorg filters responses using techniques from HCI literature

Evaluation Campaign

- 100 crowdsourced workers
- 100 trusted participants as ground truth
- 20 sites from Alexa top 1M

Filtering techniques:

1. Control questions
2. Engagement
3. Soft rules
4. Wisdom of the Crowd
Challenges

1. **Consistent experience**
   Participants have different software and network conditions

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

3. **Trustworthy results**
   Crowd workers are not always reliable
Challenges

1. Consistent experience
   Participants have different software and network conditions

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

3. Trustworthy results
   Crowd workers are not always reliable
We ran three measurement campaigns on eyeorg

1. **PLT metrics**
   How well do existing metrics capture user-perceived PLT?

2. **HTTP/1.1 vs. HTTP/2**
   Do users perceive a PLT difference between the two?

3. **Ad Blockers**
   Do users perceive a PLT difference between popular ad blockers?
We ran three measurement campaigns on eyeorg

1. **PLT metrics**
   How well do existing metrics capture user-perceived PLT?

2. **HTTP/1.1 vs. HTTP/2**
   Do users perceive a PLT difference between the two?

3. **Ad Blockers**
   Do users perceive a PLT difference between popular ad blockers?

See Paper
We use timeline tests to compare PLT metrics

PLT Metric Campaign

1000 crowdsourced workers

100 sites from Alexa top 1M

$120 total cost 1.5 days to collect responses

For each site, measure PLT 5 ways:

1. OnLoad (from HAR)
2. First Visual Change (FVC)
3. Last Visual Change (LVC)
4. SpeedIndex (from video)
5. User-Perceived PLT (from eyeorg)
OnLoad and First Visual Change correlate best with UPLT
**OnLoad** is usually within 1 second of **UPLT**

For 30% of sites, onload within 100 ms of **UPLT**

For 60% of sites, onload within 200 ms of **UPLT**
We ran three measurement campaigns on eyeorg

1. **PLT metrics**
   How well do existing metrics capture user-perceived PLT?

2. **HTTP/1.1 vs. HTTP/2**
   Do users perceive a PLT difference between the two?

3. **Ad Blockers**
   Do users perceive a PLT difference between popular ad blockers?

---

See Paper
Want to use eyeorg?
Get in touch!
https://eyeorg.net