Mobile Site Speed and the Impact on E-Commerce
We Are Baqend

We bring performance research to practice.

40+ man-years of web performance research

Novel technology for caching dynamic data

Speed Kit – SaaS for e-commerce speed
3 Things Make Your Website Slow

1. Backend Processing
2. Network Delays
3. Client
Why Do Businesses Care About Web Performance?
<table>
<thead>
<tr>
<th>Delay</th>
<th>User Perception</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 – 100 ms</td>
<td>Instant</td>
</tr>
<tr>
<td>100 – 300 ms</td>
<td>Small perceptible delay</td>
</tr>
<tr>
<td>300 – 1000 ms</td>
<td>Machine is working</td>
</tr>
<tr>
<td>1+ s</td>
<td>Mental context switch</td>
</tr>
<tr>
<td>10+ s</td>
<td>Task is abandoned</td>
</tr>
</tbody>
</table>

Stay under 1000 ms to keep users’ attention

I. Grigorik, High performance browser networking, O’Reilly Media, 2013
Jakob Nielsen, Usability Engineering, Morgan Kaufmann, 1994
You Heard The Stories

Amazon: 100 ms slower → -1% Conversion Rate

Zalando: 100 ms faster → +0.7% Revenue Per Session

Walmart: 100 ms faster → +1% Revenue

References:
- Greg Linden, Make Data Useful, Stanford Data Mining Class CS345A, 2006
- Shuhel Kagawa, Jeff Cybulski, David Martin Jones, et al., Loading Time Matters, Zalando Tech Blog, 2018
Load Time & SEO

GQ
From 7 s to 2 s Loads → +80% Traffic

Google
500 ms Slower Loads → -20% Traffic

Pinterest
40% Faster Loads → +15% SEO Traffic

Load Time & User Engagement

Forrester®
-80% load time → +60% Session Length (Mobile)

Otto
-42% time to FCP → +25% Session Length

Akamai
+2s load time → +103% Bounce Rate

Forrester. The Total Economic Impact™ Of Accelerated Mobile Pages, 2017
Lars Bognar. Mobile Speed Race der Otto Group Verbessert Mobile Ladezeiten, TWG, 2019
Load Time & User Satisfaction

Radware: +500 ms network delay → +26% peak frustration

Aberdeen Group: +1 s delay in response times → -16% customer satisfaction

Imperial College London: +50% response time → -50% productivity

References:
- The Performance of Web Applications: Customers Are Won or Lost in One Second. Aberdeen Group, 2008
Summary: The Business Impact of Site Speed
Summary: The Business Impact of Site Speed

Page Speed = Money
**How Speed Kit Works**

Website + Speed Kit JS (Service Worker)

- Fast Requests
- Speed Kit Cloud
- Real-Time Sync

3rd Parties

Origin Backend
Background: Service Workers

- Programmable **Network Proxy**, running as a **Background** Process, without **DOM** Access.

- **Capabilities:**
  - Intercept & rewrite **HTTP requests**
  - **Cache** data (CacheStorage)
  - **Store** data (IndexDB)
  - Respond **offline** or in slow network
  - Sync data & handle push
How We Solved Cache Coherence

Automatic Browser Cache Coherence

Validate Freshness

Expiration Cache

Invalidation Cache

invalidate

Add to Server Cache Sketch

Compact Cache Sketch

01011

03041
How We Solved Cache Coherence

Automatic Browser Cache Coherence

Validate Freshness

Compact Cache Sketch

Add to Server Cache Sketch

False-Positive Rate: 
\[ f \approx \left(1 - e^{\frac{kn}{m}}\right)^k \]

Hash-Functions: 
\[ k = \left\lceil \ln(2) \cdot \left(\frac{n}{m}\right) \right\rceil \]

With 20,000 entries and a 5% false positive rate: **11 Kbyte**

**Consistency**: Delta-Atomicity, Read-Your-Writes, Monotonic Reads, Monotonic Writes, Causal Consistency
USP: HTML Caching With Dynamic Blocks

Browser
(blocks marked by selector)

Replace Dynamic Blocks

Fast & Anonymous

Speed Kit Cloud

Origin Server

Slow & Personalized
In Action: Speed Kit

- Built-in cache coherence mechanism
- Lower server time (TTFB)
- On-the-fly image optimization
- Automatic browser + CDN caching

Accelerated personalization
Accelerated 3rd parties
HTTP/2 multiplexing
Network stack tuning: TCP, TLS, IPv6
All resources offline-available
This Talk = Teaser of Ongoing Study

Speedstudy.info

Mobile Site Speed and the Impact on E-Commerce

Google

Baqend
Study Setup

Implementation
A/B-Tested Speedup

Evaluation
Quantified On-Site Uplift

Long-Term Effects
Not Evaluated

- First Paint
- Time to Interaction
- ...

- Conversion Rate
- Time on Site
- ...

- SEO Rank
- Returning Visitors
- ...
How Does Speed Correlate To Business Success Exactly?
## Overview: Measuring Performance

<table>
<thead>
<tr>
<th>Synthetic Tests</th>
<th>Real-User Monitoring</th>
<th>Log Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>WebPageTest</td>
<td>Google Analytics</td>
<td>CrUX</td>
</tr>
<tr>
<td>Lighthouse</td>
<td>mPulse</td>
<td>CDN Logs</td>
</tr>
<tr>
<td>GTMetrix</td>
<td>Adobe Analytics</td>
<td>Server Logs</td>
</tr>
</tbody>
</table>

- **User-centric metrics**
  - ✔️ User-centric metrics
  - ✗ Only simulated

- **Data from actual users**
  - ✔️ Data from actual users
  - ✗ Complex to operate

- **Readily available**
  - ✔️ Readily available
  - ✗ Hard to interpret
Measuring the Uplift – With Science

CDNs, Manual Optimizations
- Only before-after comparison

Speed Kit

Application Features
- Measurable business impact through A/B tests
Measuring the Uplift – With SCIENCE

**CDNs, Manual Optimizations**
- Only before-after comparison

**Speed Kit**
- Statistically sound split testing
- Clean measurement of performance & business uplifts

**Application Features**
- Measurable business impact through A/B tests
Before Speed Kit

1.5x faster

After Speed Kit
Overall Performance

First Contentful Paint Histogram

Median 1361 ms
Median 2087 ms

*Histogram of first contentful paint on PDV pages compared between the two A/B test split groups
Without Speed Kit

2.1x faster

With Speed Kit
Stylefile: Business Uplift

User-Based Conversion Rate: +1.9%

Average Order Value: +3.8%
Directions for Research

- Prefetching & Click Prediction
- Anomaly detection on Real-User Data
- Content-based Staleness Minimization
- Dynamic Block Inference
- Dual & Adaptive Bloom Filter

- Vision-based
  - Image Optimization
  - Regression Testing
  - Speed Metrics
- Workload-based TTL Estimation
- Connection-Aware Compression & Push
How Do We Measure Web Performance?
Real-User Monitoring (RUM)

**Collection**
- Raw PL tracking & meta data
- Custom tracking

**Ingestion**
- Tracking (RUM)

**Analytics**
- Materialized views & aggregations
- Historical data

**Reporting**
- Performance Dashboard
- QA Dashboard
- Real-Time Alerting
- Ad-hoc SQL Interface
- Custom Reporting

SQL Interface
Goal: Performance & Business Insights

Browser ➔ Cloud Backend

- Tracking Beacon

Timing API
- Service Worker
- Unhandled Errors

- Time-to-First-Byte
- First (Contentful) Paint
- DOM Timer
- First Input Delay

- Session Length
- Time on Site
- First User Interaction
- Bounce Rate

- Cart Size
- Transactions
- Conversion Rate
- Revenue

- Page Views & Sessions
- Browser Distribution
- JavaScript Errors
- Caching Insights

Performance
User Engagement
Business KPIs
QA Metadata
How to Collect the Performance Data?

- Logging requests is not enough:
  - User? Rendering? ...
  - Browser cache (invisible)
  - Origin requests (no logs)
  - CDN requests

- Solution: Tracking every PI (page impression)
1. 1 for **static info**
   (URL, user agent, session ID, ...)

2. 1 for **timings**
   (TTFB, load time, FCP, ...)

3. 0–n for **events**
   (first input, add-to-cart, ...)

**Types of Data Beacons**
- **Beacon Join → PI**: How do we handle events that come late?
  - Simply wait 5 minutes?
  - Wait for next PI or session timeout?
  - ...
- How to resolve **user agents**?
- **Aggregate events**: collect all events per PI
- **Join 3 Collections**: put together PI from navigation/load/event beacons
- **Resolve User Agents**: derive browser, device, etc. from UA string
- **Unique conversions**: remove phantoms
- **Session timeout** after 30 minutes of *inactivity*
Aggregation by **session ID**: min, max, count, sum, avg, median, ...

**Deduplicate Conversions**: only 1st occurrence per conversion is valid
No Way, MongoDB!

**Indexing**
Queries over non-indexed attributes were infeasible

**Runtime**
Even with indexes in place, queries could take 30+ min.

**Scalability**
Queries got slower with increasing amounts of data

**Complexity (Joins!)**
MongoDB aggregation pipelines become sophisticated quickly
Fixing My Life With Flex Tape Athena
The „A“ Stands for „AWSome“

- Desperate attempt:
  1. Dump MongoDB collection
  2. Upload to S3
  3. Query with Athena

- Typical analysis:
  - 1 equi-join
  - 3 mio. Pls
  - ~15+ min.
The „A“ Stands for „AWSome“

- Desperate attempt:
  1. Dump MongoDB collection
  2. Upload to S3
  3. Query with Athena

- Typical analysis:
  - 1 equi-join
  - 3 mio. Pls
  - ~15+ min.
The „A“ Stands for „AWSome“

Desperate attempt: New best practice:

1. Dump MongoDB collection
2. Upload to S3
3. Query with Athena

- Typical analysis:
  - 1 equi-join
  - 3 mio. Pls
  - ~10 seconds
What's an Athena?

Athena = presto

- **Managed Presto:**
  - Interactive analytics with SQL
  - Heterogeneous datastores
  - Petabyte-scale (Facebook)

- **Pricing** by scanned data volume:
  - Efficient storage formats!
  - Partitioning or clustering!
  - Careful query design!

Raghav Sethi, Martin Traverso, Dain Sundstrom, David Phillips et al. *Presto: SQL on Everything*, ICDE 2019
Upgrading Our ETL Pipeline

- **Simplicity:**
  - Everything in one place
  - Easy to access (SQL)

- **Scalability & efficiency:**
  - Hundreds of gigabytes scanned in a query
  - Response time on the order of seconds
Our Batch Analytics Tech Stack

Issues:
- Many joins → slow queries
- 90 minutes discovery time
- No continuous dashboard (daily materialization)
2020 Real-Time Analytics Tech Stack

Prototyping Engagement with AWS

Benefits:
- No legacy tech → stability & efficiency
- Faster ingestion → Live performance charts
- Fewer joins → faster analytics
Zero-Latency Analytics

Data Beacons (collection) -> Kinesis -> Normalization (Legacy Compatibility & Validation) -> Unique Conversions (Remove Duplicate Order Events) -> UA Resolution (Derive Browser, Device, etc. From User Agent) -> PI Window (Beacons to PI) -> Session Window (PIs to Session) -> Bucketing (Histograms/Counts) -> S3 (Invalid Beacons) -> S3 (All PIs) -> S3 (All Sessions) -> elastic (1-Min aggregates)
Split Testing for Web Performance

Speed Kit Users vs. Normal Users

- Speed Kit enabled
- Measurable uplift:
  - Performance
  - User engagement
  - Business success
- Speed Kit disabled (no acceleration)
THE LARGEST SYSTEMATIC STUDY OF

Mobile Site Speed and the Impact on E-Commerce

Your Email

SUBSCRIBE FOR UPDATES  SUBSCRIBE & PARTICIPATE

Google  Baqend
Ready to Load **Instantly?**

Join the study!

Details & newsletter on  

speedstudy.info

Felix Gessert  
Wolfram Wingerath

felix.gessert@baqend.com  
wolle@baqend.com