Resource management strategies for Mobile Web-based services

Claudia Canali
Michele Colajanni
Riccardo Lancellotti

University of Modena and Reggio Emilia
The mobile Web

- **Web access form mobile devices**
  - Access to services tailored to device
    - On-the-fly adaptation
    - Small display
    - No keyboards
  - Services based on user preferences
  - Mobile Web increases the complexity of Web-based services

- **Growth of mobile Web**
  - Mobile users expected to grow by 900% within 2013

- **Will current architectures support future demands of mobile Web?**
Mobile Web-based services

- **Focus on two significant categories of site**
  - 80% of top 100 most popular sites

- **Online news sites**
  - Information portals (sports, economy)
  - Newspaper and news broadcasting sites (e.g., cnn.com)

- **Social-multimedia sites**
  - Web 2.0 sites
  - Social networking (e.g., Facebook, blogsphere)
  - Resource sharing networks (e.g., YouTube, Flickr)
Workload evolution trends

- Workload composition
- Size of workload resources
- Workload intensity

→ Growth of computational demands
Workload composition

Online news

Social multimedia

Growing amount of multimedia resources
Size of workload resources

• **Resources are getting larger**
  – Picture size
  – Video resolution and length

• **Growth of median resource size**
  – 12% per year for images
  – 16% per year for audio and video
Workload intensity

- **Growth of workload intensity**
  - Low growth scenario
    - 20%-40% per year
  - High growth scenario
    - 35%-55% per year

- **Moore's law:**
  - Computational power doubles every 18 months
  - **Is it enough?**
Experimental testbed

- Simulation based on Omnet++ with Inet package
- Server model:
  - Working set description (type and size of resources)
  - Dynamic services (depends on resource size and CPU)
  - Internal server resources (time shared CPU)
  - HTTP 1.1 interactions (chucked downloads and uploads)
- Mobile Web clients (workload intensity based on clients)
  - Use of HTTP streaming for multimedia resources
Experimental scenarios

• **Current scenario**
  - Nowadays workload models
  - Current CPUs

• **Low-growth scenario**
  - Conservative assumptions on workload evolution
  - Future CPUs

• **High-growth scenario**
  - Worst-case for supporting architectures
Performance impact

Response time

Online news

CPU power growing more than workload

Social multimedia

CPU power growing less than workload
CPU Utilization

CPU overload occurring in 3 out of 4 scenarios
Resource management strategies

• Need to reduce computational demand

• Avoid adaptation of multimedia resources on-the-fly

  → Pre-generation of multimedia content

• Pre-generating every content
  – Not every resource can be pre-generated
  – Highly volatile workload
  – High computational and storage demands
  – → Unfeasible
Resource management strategies

- **Pre-generating a fraction of the contents**
  - Focus only on the most popular resources
  - Exploit Zipf-like popularity distribution
  - How much pre-generation is required?

- **Workload characteristics:**
  - No clear model for popularity distribution
  - Zipf $\alpha$ parameter
    - From 0.8 (typical Web workload)
    - To 1.0 (highly skewed workload)
Performance impact

Online news: High growth

![Graph showing the impact of online news growth on 90th percentile response time. The graph compares two scenarios: Zipf parameter $\alpha = 1$ (solid line) and $\alpha = 0.84$ (dashed line). The x-axis represents the pre-generation fraction in percentages, while the y-axis shows the 90th percentile response time in seconds. The graph demonstrates that as the pre-generation fraction increases, the response time decreases.]
Performance impact

Social multimedia

Low growth  High growth

Pre-generating up to 15% is good for most scenarios
CPU Utilization

High growth scenario

- Online News
- Social Multimedia

October 12-14
Wimob 2008 - Avignon, France
Conclusion and open problems

- **Focus on Mobile Web**
- **Workload evolution 2008 → 2013**
  - Social networking + Multimedia will be the killer application of future mobile Web
  - Computational demand will grow faster than CPU power in most considered scenarios
- **Possible solution: pre-generating the most popular resources**
  - 5%-15% of the working set may be sufficient
- **Open problem: identifying the popular resources**
  - Highly volatile workload (the read-write Internet)
  - Short resource life span (~ 24-48 hours)
  - Need for early detection of popular resources
Resource management strategies for Mobile Web-based services

Claudia Canali
Michele Colajanni
Riccardo Lancellotti

University of Modena and Reggio Emilia