Impact of memory technology trends on performance of Web systems

Mauro Andreolini
University of Modena

Michele Colajanni
University of Modena

Riccardo Lancellotti
University of Modena
**Characteristics of today's Web**

- Complex Web-based services (dynamic Web content)
- Technology trends:
  - Increasing capacity of network connections
  - Growing amount of available memory (RAM)

<table>
<thead>
<tr>
<th>Year</th>
<th>Cost [$/Mb]</th>
<th>Typical Amount of RAM</th>
</tr>
</thead>
<tbody>
<tr>
<td>1995</td>
<td>20</td>
<td>128 Mb</td>
</tr>
<tr>
<td>2000</td>
<td>2</td>
<td>1 Gb</td>
</tr>
<tr>
<td>2005</td>
<td>0,2</td>
<td>8 Gb</td>
</tr>
<tr>
<td>2010</td>
<td>0,02</td>
<td>64 Gb</td>
</tr>
</tbody>
</table>

Memory embedded DBs will be common in a near future even for large Web sites.

What is the impact of technology trends on Web system performance?
Multi-tier Web sites

- **Front-end tier**: static Web resources, interaction with clients
- **Middle tier**: generation of dynamic Web resources
- **Back-end tier**: data repository (DBMS)

Focus on available memory for the DBMS
Motivation

- Technology trends in memory lead to changes:
  - In the system performance
  - In the bottlenecks limiting the performance

How do bottlenecks change as a function of technology trends? (no previous studies in literature)

- Bottleneck analysis is essential to plan system upgrade that can improve performance
- Need to understand and anticipate the effect of technology trends
- This approach can be applied to other Web-based applications and Web services
  - Focus on O.S. and server software (Web, DBMS, application servers)
Experimental testbed

- Dynamic Web site
  - Apache + PHP + MySQL
  - TPC-W like workload
- Additional studies with different technologies and workloads (not shown)
- Fine-grained performance analysis (*sar, oprofile*)

- Performance evaluation:
  - *When does a bottleneck appear?*
  - *What is the bottleneck?*
- Three memory scenarios:
  - All in-memory (100% of DB in memory)
  - Partially in-memory (60%)
  - Mostly on-disk (30%)
Impact of available memory on system capacity

(When does a bottleneck appear)
Page response time

- Analysis of the contribution to response time by the three tiers
  - The back-end tier contribution drives the explosion of response time
    → The bottleneck is on the DBMS
  - Confirmation of the impact of DBMS on performance
  - This is true for different technologies and scenarios
    - PHP, J2EE
    - Multiple workload mixes and memory scenarios

Bottleneck analysis focused on DBMS node

Partially in memory scenario
Bottleneck analysis (Mostly on-disk scenario)

- **Bottleneck identification** *(What is the bottleneck)*:
  - Low utilization of sockets
  - Negligible utilization of CPU
  - Full utilization of disk
Bottleneck analysis (Partial in-memory scenario)

- **Bottleneck identification** (*What is the bottleneck)*:
  - Full utilization of sockets
  - High utilization of CPU
  - High utilization of disk
Bottleneck analysis
(All in-memory scenario)

- **Bottleneck identification** (*What is the bottleneck)*:
  - Low utilization of sockets
  - Full utilization of CPU
  - Negligible utilization of disk
Analysis of results

- The amount of available RAM on the DBMS has a significant impact on the causes of poor performance

- Little memory available $\rightarrow$ performance is bounded by disk throughput,
  - Little system level interventions are available (reduced memory $\rightarrow$ caching effectiveness is reduced)
  - hardware upgrade is the most effective approach (e.g., RAID systems, memory)

- More memory available $\rightarrow$ socket descriptors limit system performance
  - high number of parallel requests can be a common situation (e.g., preliminary study on network effects)
  - should reduce request parallelism (e.g., replication of DBMS nodes, exploit of component-based systems)
Analysis of results

- Large amount of memory available → performance is limited by **asynchronous I/O** (interaction with O.S. disk cache),

- Computationally expensive checksumming operations
  - Should reduce asynchronous I/O (e.g., query caching)

- **Message for the future:** Interaction between O.S. disk cache and DBMS buffer cache can be inefficient and this can become a major bottleneck
  - Need for **efficient** DBMS tailored for memory-embedded DB operations
Future work

• Evaluation of the impact of network
  ▪ Increasing capacity of network connections
  ▪ What is the impact of network technology trends on system performance and on system bottlenecks?
• Study with multiple applications and workloads
  ▪ Pub/sub systems (e.g., forums, blogs, ...)
  ▪ Web-based Auctions
  ▪ Web services

WEB Lab group homepage
http://weblab.ing.unimo.it/