Review Form: 1st International Workshop on

Services and Infrastructure for the Ubiquitous and Mobile Internet (SIUMI'05)

SIUMI 2005

Columbus, Ohio, USA, June 6th, 2005

In conjunction with the 25th Int. Conference on Distributed Computing Systems (**ICDCS'05**)

Paper Number: 01
Paper Title: "Performance comparison of distributed architectures for content adaptation and delivery of Web resources"
Authors: Claudia Canali, Valeria Cardellini, Riccardo Lancellotti

---

Reviewer1:

Familiarity Rate your familiarity with the topic	1	2	**3X**	4
	Novice	Some knowledge	**Familiar**	Expert

Significance Technical relevance and practicality of ideas in the paper	1	**2X**	3
	Not significant	**Somewhat significant**	Highly significant

Novelty How original the problem and/or solution method is	1	**2X**	3
	Not novel	**Somewhat novel**	Highly novel

Quality of Presentation Writing and presentation style/accuracy	1	2	**3X**
	Poorly written	Could be improved	**Well written**

Overall Recommendation	1	2	3	**4X**	5
	Strong reject	Weak reject	Weak accept	**Accept**	Strong accept

Contributions

The paper presents an evaluation of three content-delivery architectures in a "controlled" environment, consisting of 16 nodes, a synthetic client workload generator and a network packet "reshaper" that allows the modeling of different WAN organizations.
The major result of the paper is the demonstration that edge-servers are effective and much better than centralized servers (even more effective when network resources are scarce) and that some cooperative caching among the edge-servers may further improve the overall response time (and degrades more gracefully).

Strengths and weaknesses

The paper is well written, with just a couple of minor English flaws ("apart the network location", "the number of alternative", ...). The work appears to be robust and technically sound. Experiments are realistic and the results well presented and discussed.
The only weakness of the paper is that it proves results that are somewhat well known or intuitive and it misses in what one could consider some more interesting measurements (see below).

Detailed public comments

In the introduction, it is not entirely clear what "transcodings" are. For example, are the authors targeting streaming media delivery (eg, live shows), or only static content? Streaming media would probably change the nature of the bandwidth allocation. Either way, it would be good to clarify, and perhaps provide a few (brief textual) examples. Also in the introduction, it would be interesting to understand what leads to the difference between the controlled environment and the real-world experiments. The benefits of a controlled environment are easy to understand, but if the results don't match what then gets observed in the real world, perhaps someone ought to take a look to understand where are the differences.

In section 2, the authors claim that the drawback of on-the-fly adaptation is that it requires more computing power. Agreed --- but, why is this a big problem? Computing is cheap today, and one can easily throw a few more blades into a rack and not have to deal with the pain of a distributed edge-server network. Would be interesting to read what the authors think about this.

In section 2.2, an example is badly needed. As a reader, I have a hard time to understand what adaptation services I can use for a photo album (perhaps image resizing for a PDA client?). At any rate, some example would be rather useful.

In section 2.3, there's something unclear about the cooperative caching. When you get a "remote hit" (ie, a hit on the cache of a peer server), why do you need to fetch the data and then send it to the client? As you'd do in most peer-to-peer systems, you'd simply send the handle to the remote hit to the client, which can go and fetch it by itself from the peer server (this way, you save one network transfer).

In section 4, an important experiment could have been added. The fact that edge servers are more effective is not news. However, a quantification would be important. In other words, there are additional interesting questions: (1) is there an origin-server configuration that enables you to match an edge-server configuration? (2) how many "nodes" does it buy to move the computation to the edge? (3) how does that relationship change with network/workload/etc? Basically -- and the authors are saying that in the intro -- edge-server architectures are not free, because one has to deal with the management of a distributed CDN (far from trivial, Akamai docet).

If one knew that an edge-server configuration is equivalent to adding k more origin-server nodes, one can use that "rule" to make a good decision. An experimental setting would allow you to discover that, and it would have been useful if such an evaluation had been presented in the paper.

When the authors discuss cache effects, they should mention the cache sizes vs. the dataset sizes, and the fact that the results are highly dependent on the locality of the traces. It would be interesting to know how much locality the traces have and whether the authors had to do some form of cache management (replacement) in the experiment.

Finally, and surprisingly, there were no references to Akamai papers, considering their pioneering role in CDNs.

<u>Reviewer2:</u>

Familiarity Rate your familiarity with the topic 3	1	2	3	4
	Novice	Some knowledge	Familiar	Expert

Significance Technical relevance and practicality of ideas in the paper 3	1	2	3
	Not significant	Somewhat significant	Highly significant

Novelty How original the problem and/or solution method is 2	1	2	3
	Not novel	Somewhat novel	Highly novel

Quality of Presentation Writing and presentation style/accuracy 3	1	2	3
	Poorly written	Could be improved	Well written

Overall Recommendation	1	2	3	4	5
4	Strong reject	Weak reject	Weak accept	Accept	Strong accept

Contributions

The paper provides performance comparison of three architecture for dynamic content adaptation and delivery. All considered architectures consider multiple intermediate servers for content adaptation placed between the servers providing the original content (origin servers) and clients. They differ for the positions of the intermediate servers with respect to origin serves and clients, and for the degree of cooperativeness between these intermediate servers. The comparison demonstrates that the solution with the intermediate servers close to the clients and capable to cooperate has the best performance.

Strengths and weaknesses

Strenghts: The paper is technically sound and very well written. It presents original experimental data assessing the relative performance of the three considered architectures in a number of significant scenarios. The use of experimental data, obtained in a controlled real testbed and not through simulations is particularly interesting.

Weaknesses: The three considered architectures have already been proposed in the literature (the cooperative one by the authors themselves in a previous paper); this somewhat limits the novelty of the paper. Nevertheless there is sufficient original content to deserve publication.

Detailed public comments

The paper provides performance comparison of three architecture for dynamic content adaptation and delivery. All considered architectures consider multiple intermediate servers for content adaptation placed between the servers providing the original content (origin servers) and clients. The comparison demonstrates that the solution with the intermediate servers close to the clients and capable to cooperate has the best performance.

The paper is technically sound and very well written. It presents original experimental data assessing the relative performance of the three considered architectures in a number of significant scenarios. The use of experimental data, obtained in a controlled real testbed and not through simulations is particularly interesting. However, the three considered architectures have already been proposed in the literature; this somewhat limits the novelty of the paper.

Minor comments are:

- page 4, col 2, lines 11-12: I guess that the mentioned rate is measured in data units per time unit and not requests per time unit, but it would be better to specify
- page 5, col 1, lines 24-28: It is possible for the reader to understand that values in bold are used to produce data in tables 3 and 4, but it would be easier if this information were put closer to those tables in the text.

<u>Reviewer3:</u>

Familiarity Rate your familiarity with the topic			3	
			Familiar	
Significance Technical relevance and practicality of ideas in the paper			3	
			Highly significant	
Novelty How original the problem and/or solution method is		2		
		Somewhat novel		

<table>
<tr><td rowspan="2">Quality of Presentation
Writing and presentation style/accuracy</td><td colspan="2"></td><td colspan="2">2</td><td></td></tr>
<tr><td colspan="2"></td><td colspan="2">Could be improved</td><td></td></tr>
<tr><td rowspan="2">Overall Recommendation</td><td></td><td></td><td></td><td>4</td><td></td></tr>
<tr><td></td><td></td><td></td><td>Accept</td><td></td></tr>
</table>

Contributions

The paper studies what is called architectural alternatives (I'd call this just functionality placement because the architecture is a middle box one) in a somewhat new (at least to me) settings that differ from a large pack of CDN and cache optimization paper.

Strengths and weaknesses

Problem statement is somewhat new (to my understanding).
Formal (or semi-formal) statement of research problem is missing.

Detailed public comments

The novelty is in considering placement of media (content) adaptation functionality and the overall impact on system performance. I might be not aware of majority of related work in this area, however I find a bit strange that authors are not specifying their assumptions, e.g. under what conditions the problem statement is valid? (One case I may suggest is that user signals to media what and adaptation to make and **where**).
WRT the paper's conclusions section: I'd recommend authors to make these more attractive to a reader with some insights that are already clear (cooperation → adaptation as the next step).
Can you generalize results for other mixes of traffic?