Performance Evolution of Mobile Web-Based Services

The mobile Web's widespread diffusion opens many interesting design and management issues about server infrastructures that must satisfy present and future client demand. Future mobile Web-based services will have growing computational costs. Even requests for the same Web resource will require services to dynamically generate content that takes into account specific devices, user profiles, and contexts. The authors consider the evolution of the mobile Web workload and trends in server and client devices with the goal of anticipating future bottlenecks and developing management strategies.

Claudia Canali, Michele Colajanni, and Riccardo Lancellotti
University of Modena and Reggio Emilia

The spread of Web-enabled mobile client devices, such as laptops, handheld PCs, PDAs, and smart phones, has significantly altered the Web landscape. A growing number of mobile users can access vast amounts of data in the form of Web content. This evolution is determining the advent of the so-called *mobile Web*, in which users access Web content anytime, anywhere, and through any class of device. This mobile scenario, coupled with Web resources' complexity and diversity, has opened the possibility of mobile Web-based services that tailor content to user preferences, user locations, and device capabilities.

Future server-infrastructure design must consider mobile Web-based services' performance requirements and how their impact on server infrastructure will evolve. Heterogeneous scenarios require services to dynamically generate and adapt Web content on-the-fly at the moment of the client request.[1,2] However, such operations are computationally expensive. The increasing number of innovative services offered on the Web only exacerbates the costs. Specifically, the success of video sharing (as with YouTube), online games, virtual reality environments (such as Second Life), and mobile IPTV is causing a growing demand for graphics-heavy and multimedia content. Adapting these novel services to a mobile context will represent a major challenge for future server infrastructures.

1089-7801/09/$25.00 © 2009 IEEE

To evaluate the future performance requirements of mobile Web-based services, we consider three main factors: workload characteristics' evolution and technological improvements at both the server-infrastructure and client-device levels. The trade-off should be clear: workload evolution and novel services will increase each user request's computational demand, while server-platform and client-device technology improvements will support more demanding services. Assuming that no disruptive technology — that is, technologies that can upset the computation paradigm — develops in the next decade, we contend that the technology evolution might be insufficient to support a widespread diffusion of mobile Web-based services. Our goal is to anticipate possible bottlenecks in future systems and propose management strategies to cope with them.

Mobile Web-Based Services

Mobile Web-based services fall into three categories: personalization, context awareness, and content adaptation.

Personalization aims to generate customized content to tailor navigation to the user's preferences and needs.[1] User information is typically stored in one or more databases and obtained as a consequence of explicit information coming from the user request or inferred through runtime or offline analysis of user behavior (for example, through data mining a Web site's log files). This information is the basis for deploying personalization services, such as subscriptions to news feeds, customized layouts, request filtering, recommendation systems,[1] and adaptation to user's navigation styles.

Context awareness represents an innovative class of services that generate Web content on the basis of information about the user's context. Services should consider context in the broadest sense, including information about geo-location, ambient location, time of day, or current user activity. This feature is one of the mobile Web's most innovative: the ability to continuously trace context and user position. Typical context-awareness examples are location-based applications that offer tourist information and commercial advertisements, and services that help users discover nearby resources, such as devices, friends, and members of a social community.

Content-adaptation services transform original Web content in the form of images, audio, or video to match client device requirements for network connection, computational power, storage, and display size. Multimedia content represents a growing percentage of Web information. So, server infrastructure should provide efficient transformation within multimedia types (such as changing an image's color depth or converting from a high-fidelity JPEG to a low-fidelity GIF format), across multimedia types (such as from video to still images), or both.

Because the large variety of mobile Web-based services makes it difficult to identify a "typical Web site," we instead established major site categories: news, social, and travel. To do this, we analyzed the 100 most popular sites worldwide, according to the Media Metrix Top 500 (www.mediametrix.com) and Alexa Top 500 (www.alexa.com) lists. News and social sites cover roughly 80 percent of the sites and are likely to remain popular on the mobile Web.[3] Travel sites, the third category, are an emerging site type with expected widespread use among mobile users.[4]

News

The news category includes information portals, such as online newspapers or news broadcasting sites, that offer information from events to stock quotes and sports results. These sites typically deliver news in the form of textual resources and images, which account for 60 percent and 35 percent of requests, respectively. Although growing in use, audio and video content today accounts for only 5 percent of requests (www.stateofthenewsmedia.org/2008). Often, personalization and context-aware services customize textual resources to user preferences. Content adaptation services tailor multimedia resources.

Social

The sites in the social category represent a new form of user communication and interactivity, the qualifying characteristics of Web 2.0. Typical examples are forums, blogs, and content-sharing sites, in which users exchange opinions, stories, and files (examples include MySpace, Flickr, and YouTube), and online gaming and virtual worlds, in which users navigate and socialize in three dimensions. Some 55 percent of the requests for these sites involves personalization and context awareness. Social sites also rely on content-adap-

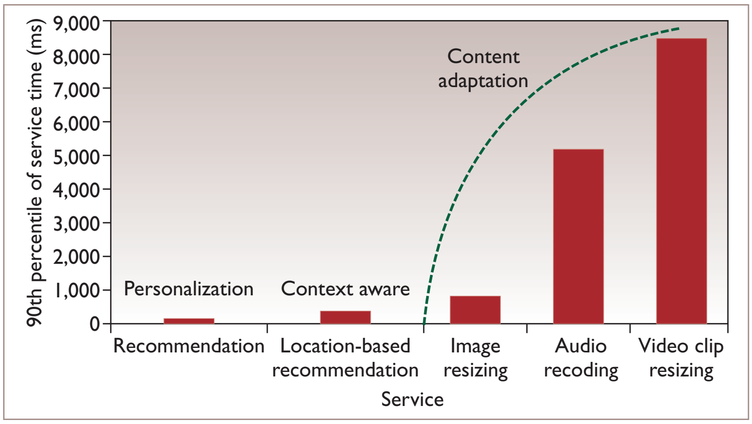

Figure 1. Service times for mobile Web-based services. The 90th percentile of service time is a suitable metric because the Web has heavy-tailed distributions at the system and workload levels.

tation services to tailor multimedia content (roughly 45 percent of the requests) because many user interactions involve exchanging images or audio or video resources.[5]

Travel

The travel category includes Web sites offering location-based information to support user mobility. Examples include sites offering online travel reservations (such as Expedia) or tourist and travel information (such as Transport for London's Journey Planner at www.journeyplanner.org). Roughly 65 percent of requests involve personalization and context awareness. Other requests involve adapting multimedia resources that illustrate tourist locales or events (27 percent for images and 8 percent for audio and video).

Computational Cost

Among personalization, context awareness, and content adaptation, we wanted to determine which service is likely to have the highest computational demand on server infrastructure.

Personalization

To evaluate personalization, we studied the recommendation service that sites such as Facebook and YouTube use to rate and categorize content and users. The service uses offline collaborative filtering and clustering to group users with similar preferences or click history.[1] The resulting associations reside in a recommendation database. When accessing the site, the service identifies each user as belonging to a specific group and retrieves the corresponding recommendation list from the database. This personalization service time takes on the order of 100 milliseconds because the most costly operations occur offline. We implemented online operations through common open source technologies: Apache Web server, MySQL database servers, and PHP applications.

Context Awareness

To evaluate context awareness, we implemented a location-based service resembling that of Food.com. The service exploits user location and time of day to create and display a map of neighborhood restaurants and their daily menus. This location-based service time takes between 100 and 400 ms, depending on the number of accessed databases.

Content Adaptation

To evaluate content adaptation's computational costs, we studied Web portal services, such as AvantGo, that customize content from popular sites to the client device's requirements, including small display, limited rendering capability, and low-speed connection. The typical service time for content adaptation is 1 or more seconds because of the complex operations it must carry out on possibly large files.

Adapting images, audio, and video resources requires a separate analysis because service time depends on resource size. Image adaptation involves various techniques, such as scaling, cropping, or compressing the image.[2] In our experiments, we evaluated reducing the spatial geometry from the original resolution to 480 × 320 pixels, a typical mobile display size. We considered JPEG images with a median size of 80 Kbytes to be a typical resource in the working set of Web workloads.[6]

Most audio resources accessed through the Web are music files in MP3 format with a median size of 3.5 Mbytes.[7] Services usually adapt these resources by reducing the bit rate. We recoded the audio resources to reduce the original bit rate to 32 Kbps.

If we exclude large mobile TV videos for which statistics are unavailable, most video resources on the Web are clips in MPEG format with a median size of 7 Mbytes and duration of a few minutes.[7] Common transformations include frame size or color depth reduction. We reduced the frame size of short video clips from the original to 480 × 320 pixels.

For content adaptation, we relied on open

Adaptation service	Service time per Mbyte (milliseconds)	Service time per Mbyte (normalized)
Image resizing	920	1
Audio recoding	1,471	1.59
Video clip resizing	1,254	1.36

Table 1. Computational cost of content-adaptation services.

source software: ImageMagick library (www.imagemagick.org) to adapt the images and Lame (lame.sourceforge.net) and Transcode software (www.transcoding.org) for audio and video recoding.

Results

We carried out experiments on different server platforms, including a virtual server cluster, and we found no significant difference among the orders of magnitude for service times or the respective service time results. Thus, our performance evaluations' main conclusions don't depend on the hardware or software platform.

Figure 1 shows the 90th percentile of service time — that is, the service fulfills 90 percent of the requests within that time. Because the Web has heavy-tailed distributions at the system and workload levels, this metric is more meaningful than average values.

Figure 1 summarizes two results that emerged from all our experiments: The service times characterizing the mobile Web might span several orders of magnitude, and the lowest service times are related to personalization and context-aware services. We expected the former result but were surprised at the latter because of the complex operations behind these services. The explanation is that the most time-consuming tasks for personalization and context awareness are related to data preparation through data mining, clustering, and indexing[6] that can occur offline. Thus, when a client makes a request, the service accesses one or more databases updated offline.

Another important factor is that the cost of personalization and context-aware services doesn't relate to the requested resource's size because the service time primarily relates to database interactions. However, adaptation's service time directly correlates to the adapted resource's size.[2] The low service time for resizing images (around 10s of Kbytes in size) and the high service time for adapting audio and video files (ranging in size from 3 to 7 Mbytes on average) confirm this observation. Table 1 shows the relationship between resource size and adaptation time. The service times have some value in absolute terms because services obtain them on a specific architecture. However, they demonstrate that the adaptation service times of different resources per unit size are in the same order of magnitude, and the differences in Figure 1 are due to a multimedia resource's "typical" size.

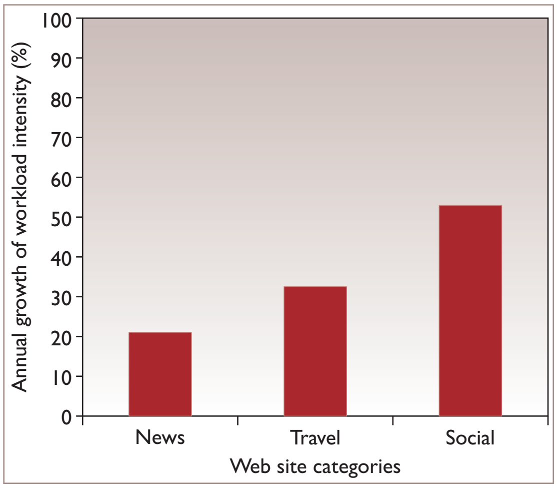

Figure 2. Workload intensity growth. Social sites are becoming more popular overall, but mobile use will spark rapid growth.

Mobile Web-Based Services Evolution

According to workload trend studies, workload intensity, composition, and resource size are the three primary factors that might affect service demand.[4,8,9]

Workload Intensity

From 2000 to 2007, the workload intensity trend showed steady growth for news sites (www.stateofthenewsmedia.org/2007), with mobile use accounting for much of the increase.[8] Workload intensity for travel sites will grow by 29 percent in each of the next six years.[4] Social sites, which also have gained in popularity recently, will experience rapid growth among mobile users in the next six years.[8] Figure 2 shows workload intensity's expected annual growth for each of the site categories.

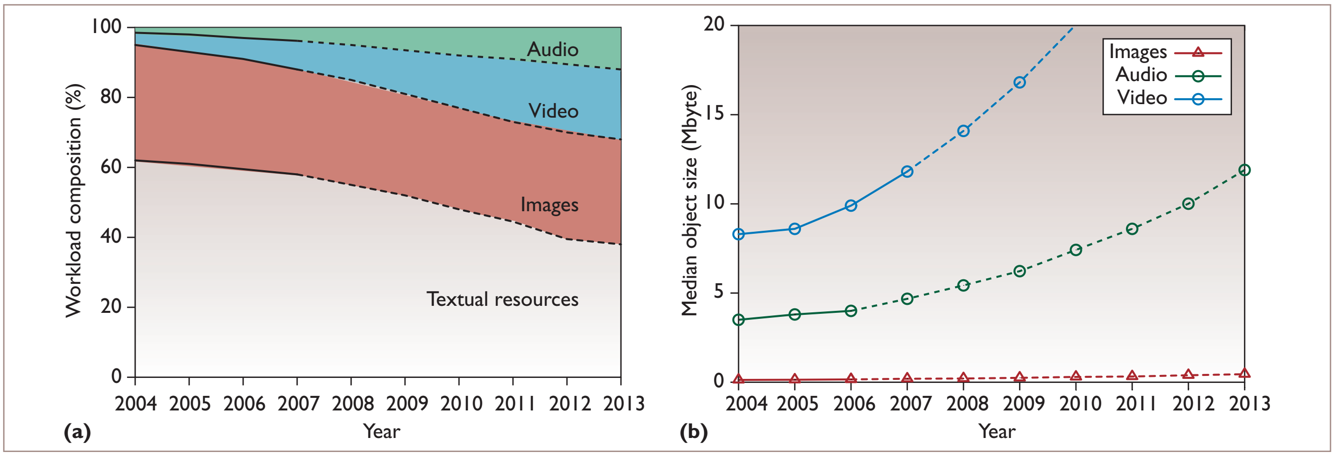

Figure 3. Workload composition growth. (a) The expected workload composition trend for news sites shows the balance shifting to multimedia resources. (b) The trend for all sites shows that the median size of audio and video resources will also rapidly increase.

Workload Composition

The mix of content types is an important workload characteristic for evaluating mobile Web-based services' impact on server infrastructure performance. Images, audio, and video resources might require content-adaptation services. Textual resources might involve personalization or context-aware services that tailor, for example, page layout or recommendations on the basis of user preferences and contexts.

News site workload includes an increasing amount of multimedia content, especially video and audio resources. One news site following this trend is CNN.com, which evolved from text and image information in 1997 to a current rich mix of video and audio content. Figure 3a shows the projected workload composition trend for news sites. We based the projection on data from the *2004 State of the News Media* report (www.stateofthenewsmedia.org/2004) and an analysis of five top news site homepages (CNN, Yahoo! News, MSNBC, BBC, and AOL News) in June 2007. We expect the trend of the past three years to continue for the next six. Even the workload of social sites appears to be moving rapidly toward multimedia. Social sites have evolved from Slashdot, an early example of a blog essentially based on textual content, to multimedia sharing sites[10] such as YouTube, in which interactions occur through video clips.

Because of the large variety of sites and the lack of scientific statistics about the resources that comprise social sites' workload, we assume that the workload composition follows multimedia increments similar to those of news sites. Travel sites' workload composition, however, already contains significant multimedia content and isn't likely to change abruptly.

Future news and social sites will feature a high increase of audio and video resources relative to textual resources and images. Consequently, we can expect that the adaptation of audio and video resources will represent the most expensive service.

Resource Size

For resource size, we focus on multimedia resources because they affect content adaptation services' computational cost. To predict the trend of multimedia resource size, we rely on multiple studies.[2] For image resources, we refer to the Web workload characterization of Adepele Williams and colleagues,[9] who identified size evolution from 1995 to 2004. We assume that the median image size will follow a similar pattern from 2004 to 2013.

We refer to measurements from the period 2004–2006[7] for audio resources and 2004–2007[7,10] for video resources. For both types of resources, we expect that these trends will continue for the next six years. Our measures on resource sizes consider sites that professional providers operate, because they tend to enforce size limits on multimedia resources (especially audio and video resources) to guarantee a high-quality experience for users. Because of this, we corrected the data we obtained from nonprofessionals' sites (such as that of Williams and colleagues[9]) by truncating the resource size distribution tail.

Figure 3b shows the expected trend in each multimedia type's median resource size. The size

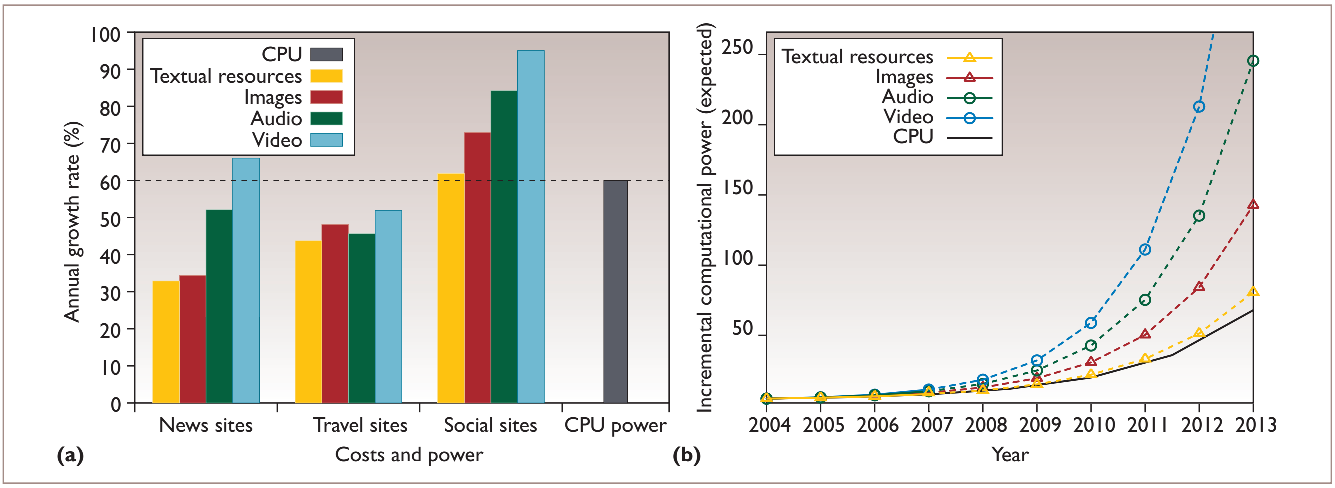

Figure 4. Computational costs versus power. (a) The projected annual growth rate of computational power required to serve each site category shows social sites' demands outpacing the computational power supply. (b) The computational power trends for social sites shows growth in each resource type.

is likely to increase for every multimedia type but at a greater rate for audio and video resources.

Computational Demand vs. CPU Computational Power

Mobile Web-based services' computational demands will grow significantly due to the combination of multiple trends, such as request rate and resource size increments and larger amounts of multimedia content.

Along with workload evolution, we must also consider improvements in the computational power of server platforms. To this end, we refer to Moore's law, which asserts that the number of transistors on an integrated circuit is expected to double every 18 months. We apply the same assumption to computational power. We justify this assumption with the effect of the increased transistor density coupled with improved capabilities for exploiting internal parallelism, which is due to better technologies on system bus and internal logic.

We also assume that over the next decade, this trend will remain valid, and no disruptive technology will emerge. In this case, mobile Web-based services' computational costs are likely to follow a 64-fold decrease in the next decade as a consequence of increasing server platform computational power. So, if we consider the service times Table 1 shows for adapting multimedia resources, they'll decrease in 10 years between 15 and 20 ms per Mbyte for any multimedia resource type.

To evaluate mobile Web-based services' impact on performance, we consider the combined effects of the trends in workload intensity, workload composition, and median resource size. Figure 4a shows the annual growth rate of computational power required to serve each resource type. For example, the annual growth rate for video content on social sites is 96 percent. We obtain this percentage by multiplying the growth in median video size (16 percent) by the growth in workload intensity for social sites (51 percent) by the growth in workload composition for that service, where the fraction of video content grows by 12 percent every year.

In Figure 4a, we also report the improvements of the CPU computational power per year. The dotted threshold denotes the services that are likely to cause problems at the server side. Figure 4b compares the trends of the expected increase in computational power to serve each resource type for social sites and of the CPU power in the decade from 2004 to 2013. The projection shows the computational demand of travel site services increasing at a slower rate than CPU power. However, for the other scenarios, some services are likely to grow at a faster rate. In particular, the CPU power improvements don't seem to counterbalance the explosion in the computational demand of social sites' mobile Web-based services.

Performance Study Consequences

All the performance analyses indicate that mobile Web-based services' computational demand might grow faster than servers' CPU power, requiring large investments to support widespread

Evolution of Mobile Client Devices

The first generation of mobile Web-enabled devices was a spectrum of devices, from laptops to smart phones. They had limited capabilities and heterogeneous characteristics in terms of network connection, computational power, storage, and display size. Mobile devices' technological evolution is substantially improving the scene for the mobile client population.[1]

The bandwidth available to mobile devices has greatly increased with the spread of 3G wireless networks, which have boosted bandwidths from Global System for Mobile technology's few Kbps to rates on the order of Mbps. The advent of 4G and WiMax technologies are expected to bring further enhancements.[2,3]

These significant improvements in computational power have enabled even small devices to consume multimedia content in several formats.[4] Computational power has also significantly affected devices' rendering capabilities. The first generation of devices ranged from monochrome to full-color capabilities, whereas modern devices can display at least 16-bit color images. Previous adaptations from color to black-and-white videos are now useless.

Mobile devices' processing power and storage resources have increased as well, letting them consume larger resources and facilitating content adaptation at the client side. The display size is likely to experience minor improvements owing to portability requirements. The trend, however, is toward devices with at least 3-inch screens and resolutions of 480 × 320 pixels or higher[1] because users are accessing multimedia resources through their mobile devices.

The mobile clients' technological evolution has a positive consequence on the server side because future server infrastructures won't need to tailor resources for every type of client device, as is currently the case. In the future, we can expect that different devices will be able to consume the same version of a multimedia resource with no server adaptation, and, if necessary, some devices will perform final adjustments locally. However, even if mobile clients become powerful devices with medium to large connections, limitations on energy and bandwidth will prevent client-side-only adaptations.

References

1. M. Walker et al., "Future Mobile Devices: An Overview of Emerging Device Trends, and the Impact on Future Converged Services," *BT Technology J.*, vol. 25, no. 2, 2007, pp. 120–125.
2. T.B. Zahariadis, "Migration Toward 4G Wireless Communications," *IEEE Wireless Comm.*, vol. 11, no. 3, 2004, pp. 6–7.
3. A. Ghosh et al., "Broadband Wireless Access with WiMax/802.16: Current Performance Benchmarks and Future Potential," *IEEE Comm. Magazine*, vol. 43, no. 2, 2005, pp. 129–136.
4. Y.-K. Chen et al., "Trends and Challenges with System-on-Chip Technology for Multimedia System Design," *Proc. Emerging Information Technology Conf.*, IEEE Press, 2005; doi: 10.1109/EITC.2005.1544365.

mobile accesses. A critical decision for future mobile Web-based service support concerns tasks the server infrastructure must perform on-the-fly at the moment of the client request and tasks that it can perform offline. Personalization and context-aware services already perform the most expensive tasks offline. More efficient data and text mining techniques will enhance this solution, enabling content pregeneration and server-side caching techniques for specific user profiles.

Adapting multimedia resources can involve large computational costs to the point at which most present server infrastructures become inadequate. Different solutions are practical, depending on the nature of multimedia resources: prerecorded versus live content.

Prerecorded Multimedia Content

Services exploit prerecorded content on the servers and deliver it at clients' request, possibly after adaptation. The clients' technological evolution will let future devices consume directly, or through a client-side adaptation, a wide range of resources (see the sidebar, "Evolution of Mobile Client Devices"). This scenario opens novel content management and architectural solutions. Now, pregenerating or caching all versions of each prerecorded multimedia resource isn't feasible or convenient because there is a combinatorial number of alternatives deriving from different contents, devices, and network bandwidths.

However, in the future, offline pregeneration and server-side caching of adapted multimedia resources can become a convenient way to limit expensive online adaptations. Researchers have demonstrated the effectiveness of caching solutions for multimedia content.[5] For example, caching the most popular 10 percent of resources satisfies 80 percent of the requests owing to the highly skewed popularity distribution of video-on-demand resources. Moreover, server-side caching is simple to deploy and avoids the need to redesign applications to generate offline all the versions of the prerecorded content.

Pregeneration and caching save CPU cycles at runtime but significantly add to storage costs.

So, there's an interesting trade-off between CPU and storage investments. Storage requirements will increase because of multimedia resources' growth in size and number but will decrease thanks to the need to save fewer versions. For example, our experiments show that introducing five versions of each resource in a multimedia workload leads to just a twofold growth of the working set size. The choice of storing multiple versions of multimedia resources will become a convenient alternative for the future by comparing the evolution in storage requirements with the trends in storage capacity. To this end, we consider the social scenario expected to have the highest storage requirements. The storage capacity, in terms of disk density, has increased from 60 percent to 100 percent every 12 months for the past 15 years[11] at the same or lower cost per Gbyte. However, the storage requirement for the social scenario is expected to increase by 54 percent every 12 months for the same period. From this simple analysis, we conclude that if the storage capacity continues to experience at least 60 percent growth per year, the pregeneration and caching of adapted resources will become the most convenient alternative for managing prerecorded multimedia content.

Live Multimedia Content

Live content includes event broadcasting, video conferencing, and online gaming, for which no alternative to expensive on-the-fly adaptations exists. Server side scalability is achievable through software and system solutions at the local or geographical level. At the software level, transcoding research is proposing novel algorithms to reduce content adaptation's computational cost,[12] and parallelism is often a viable solution. At the system level, modern and future clusters will be able to address most computational issues. However, when a system must deliver a large number of streams on a geographic scale, as in the case of IPTV services, resorting to third-party infrastructures might be a valid alternative.

The two main third-party options aren't comparable now, but both might someday result in viable solutions. The consolidated approach relies on a content delivery network, such as Akamai, that takes care of content adaptation and distribution through its geographically spread servers. A more innovative solution refers to peer-to-peer systems, in which the overlay network servers can collaborate for multimedia adaptation and delivery. There's some resistance to thinking of peer-to-peer infrastructures in a commercial context because of the open issues around availability, security, and copyright infringement. However, the efforts of various research groups and companies might change the negative halo surrounding these systems.

Services for the mobile Web are placing and will place an increasing demand on underlying server infrastructures because of the need to tailor contents to user preferences, contexts, and device capabilities. The major performance challenge will likely concern services that adapt multimedia contents. For on-the fly adaptation services, the mobile Web's expected evolution might represent a critical problem because the growth of computational requirements might exceed the increase in CPU computational power.

Key to the evolution of mobile Web-based services is supporting computationally demanding services. A possible first step toward addressing this issue is to limit runtime adaptation to only live multimedia contents and exploit offline solutions, in which pregenerated adapted versions of multimedia resources are stored on the server platform.

References

1. S. Flesca et al., "Mining User Preferences, Page Content and Usage to Personalize Website Navigation," *World Wide Web J.*, vol. 8, no. 3, 2005, pp. 317–345.
2. C. Canali, V. Cardellini, and R. Lancellotti, "Content Adaptation Architectures Based on Squid Proxy Server," *World Wide Web J.*, vol. 9, no. 1, 2006, pp. 63–92.
3. Berg Insight, *Mobile Internet 2.0*, May 2007; www.berginsight.com/ReportPDF/ProductSheet/BI-MI2-PS.pdf.
4. *Portable Navigation &Wireless Tracking: Western European Markets & Forecasts 2007-2012*, Juniper Research, Oct. 2007; http://juniperresearch.com/shop/viewreport.php?id=94.
5. M. Cha et al., "I Tube, You Tube, Everybody Tubes: Analyzing the World's Largest User Generated Content Video System," *Proc. 7th ACM SIGCOMM Conf. Internet Measurement* (IMC 07), ACM Press, 2007, pp. 1–14.
6. S. Chandra et al., "Transcoding Characteristics of Web Images," *Proc. SPIE Vol. 4312: Multimedia Computing and Networking*, SPIE, 2001, pp. 135–149.
7. L. Guo et al., "Delving into Internet Streaming Media

Delivery: A Quality and Resource Utilization Perspective," *Proc. ACM SIGCOMM Internet Measurement Conf.* (IMC 06), ACM Press, 2006, pp. 217–230.

8. N. Jones, *Mobile Web Trends 2007 to 2011*, Gartner Group, June 2007; www.gartner.com/DisplayDocument?id=507337.
9. A. Williams et al., "Web Workload Characterization: Ten Years Later," *Web Content Delivery*, X. Tang, J. Xu, and S.T. Chanson, eds., Springer, 2005.
10. P. Gill et al., "YouTube Traffic Characterization: A View from Edge," *Proc. ACM SIGCOMM Internet Measurement Conf.* (IMC 07), ACM Press, 2007, pp. 15–28.
11. E. Grochowski and R.D. Halem, "Technological Impact of Magnetic Hard Disk Drives on Storage Systems," *IBM Systems J.*, vol. 42, no. 2, 2003, pp. 338–346; www.research.ibm.com/journal/sj/422/grochowski.html.
12. A. Vetro, C. Christopulos, and H. Sun, "Video Transcoding Architectures and Techniques: An Overview," *IEEE Signal Processing Magazine*, vol. 20, no. 2, 2003, pp. 18–29.

Claudia Canali is a researcher at University of Modena and Reggio Emilia. Her research interests include mobile and wireless systems and content adaptation and delivery. Canali has a PhD in computer engineering from the University of Parma. Contact her at claudia.canali@unimore.it.

Michele Colajanni is a full professor at University of Modena and Reggio Emilia. His research interests include distributed systems for Web-based services, performance modeling and evaluation, and security. Colajanni has a PhD in computer engineering from the University of Roma Tor Vergata. Contact him at michele.colajanni@unimore.it.

Riccardo Lancellotti is a researcher at University of Modena and Reggio Emilia. His research interests include distributed systems for ubiquitous Web access, performance evaluation of multitier Web clusters, and peer-to-peer systems. Lancellotti has a PhD in computer engineering from the University of Roma Tor Vergata. Contact him at riccardo.lancellotti@unimore.it.