Architectures for Personalized Multimedia

Srinivas Ramanathan and P. Venkat Rangan
University of California at San Diego

Personalized multimedia on-demand services are fast evolving from a symbiosis of storage, network, and content providers. A major obstacle to their practical realization is the unprecedented cost of storage and transmission. Architectures and caching techniques can minimize the costs of delivering personalized multimedia programs across metropolitan networks.

Until recently, information services have been available to us in more ways than one. On the educational front, information acquisition, authoring and composition, and subsequent publishing and delivery occur predominantly in printed form, as journals, newspapers, and books. Entertainment, on the other hand, is produced and delivered via television and radio broadcasts, primarily in analog form. Both these spheres are in for radical changes: Advances in hardware and software are computerizing the processes of information acquisition, authoring, composition, and publishing. Large-capacity storage disks and high-speed fiber-optic networks are enabling the integration of diverse media, such as text, audio, and video, during storage and delivery. Together, these advances are stimulating the development of multimedia technologies in which information (such as journals, newspapers, books, movies—nay, entire encyclopedias even) are all digitally composed and published on multimedia platforms, stored on high-capacity servers, and accessed by clients via broadband networks.

The promise of multimedia technologies to have a wide-ranging impact on both education and entertainment has sparked a number of nationwide ventures (see sidebar on next page). These ventures, targeted at providing multimedia services on demand, represent a radical shift from the conventional broadcast mode of service. In the traditional *broadcast* mode, typified by cable television, clients can neither control the programs they view (for example, by skipping portions they find uninteresting) nor schedule the viewing time of programs to suit their preferences. Moreover, to view programs of interest, clients must subscribe to all the channels that broadcast at least one such program. Thus, clients not only must bear the additional costs for subscribing to a large number of channels, but also withstand an enormous overload of irrelevant information. In contrast, the *on-demand* mode permits clients to procure only what they desire, schedule the viewing times of programs, and control programs by pausing, resuming, fast-forwarding, or rewinding.

In this article, we explore how on-demand services are likely to evolve into *personalized* multimedia services customized to suit the individual needs and preferences of clients. By effectively combining the selectivity and flexibility of on-demand services with the versatility and programmability of computers, personalized multimedia services promise a new epoch in which clients no longer have to search, locate, and schedule media presentations. Rather, intelligent Personal Service Agents (PSAs), acting on behalf of clients, search for and locate information that matches clients' needs and schedule presentations at clients' preferred viewing times.

However, the network bandwidths necessary for supporting simultaneous, independent program selections in such personalized services are prohibitively large. To amortize network usage among clients with similar preferences, PSAs employ intelligent caching strategies that judiciously store media information at strategic locations in the network. We now discuss in detail the architectural considerations underlying the practical realization of personalized multimedia services.

Personalized multimedia services

Personalized multimedia services are expected to evolve from a symbiosis of several enterprises: storage providers, who manage information storage at multimedia servers (akin to analog video stores and libraries); network providers, who are responsible for media transport over integrated networks (a role similar to that of telephone and cable companies); and content providers (such as publishing houses, news distributors, entertainment houses, and radio and television stations), who offer a multitude of services using multimedia servers for storage and retrieval, and high-speed networks for media transmission between the servers and clients' sites. Different types and sizes of clientele (for example, residential homes, educational institutions, and commercial organizations) subscribe to services offered by the content providers.

1070-986X/94/$4.00 © 1994 IEEE

Initiatives for Multimedia Technologies

Federal agencies such as the National Science Foundation (NSF) are promoting concerted efforts to develop digital multimedia libraries that "support research, education, and commerce by providing ubiquitous access to relevant, high-quality, usable information."[1] Such multimedia libraries promise to be instrumental in enabling self-paced education that "encourages students to take the most efficient path to mastery," since "the audio-visual presentation is readily absorbed; immediate interaction and feedback reinforces concepts; and one-on-one instruction accommodates different learning styles."[2] Efforts are already underway to build such "libraries without walls," including an on-line version of the entire Library of Congress' collection of 100 million items (which includes books, lectures, historical documents, photos, maps, laws, cartoons, and software).

In the entertainment sector, several industrial joint ventures aimed at developing interactive multimedia entertainment formed in recent months. The $33-billion merger of Bell Atlantic and Tele-Communications plans on "blending technology and assets, so that phone lines are enhanced with video and cable networks to provide two-way communication."[3] Cable giant Time Warner and telephone carrier US-West announced plans to equip clients' sites with fiber-optic networks and to support hundreds of entertainment channels. Plans even include a teleshopping channel, called "The Catalog Channel," through which clients can view merchandise and make purchases from their homes.[4] Another telephone carrier, Nynex, joined hands with Dow Jones, a financial company, to develop a video news distribution service that offers investment professionals fast access to the latest financial news and trends.[5]

Computer companies have not lagged behind. IBM teamed up with Blockbuster to develop new distribution technology for video rental services. Kaleida, a joint venture of IBM and Apple, announced an alliance with Motorola and Scientific Atlanta to develop the technology for "interactive television" that will permit clients to interactively choose and view programs of their interest from different entertainment channels. A competing alliance involving Intel, Microsoft, and General Instruments also outlined plans to develop hardware and software for interactive television. Their end product, a smart TV-top box resembling the black box that now sits on top of cable TVs, hopes "to keep channel-surfers afloat in an era of overflowing entertainment."[6] This is only a small sampling of the various industrial initiatives underway to develop multimedia technology that can offer greater variety and higher selectivity in home entertainment.

References

1. E.A. Fox, "Advances in Interactive Digital Multimedia Systems," *Computer* (special issue on multimedia information systems), Vol. 24, No. 11, Oct. 1991, pp. 9-19.
2. J.A. Adam, "Interactive Multimedia: Applications, Implications," *IEEE Spectrum*, Vol. 30, No. 3, Mar. 1993, pp. 24-31.
3. "Big Brother's Holding Company," *Newsweek*, Oct. 25, 1993, pp. 38-43.
4. D. Dishneau, "New Player Joins Home Shopping Fray," *San Diego Union-Tribune*, Sept. 28, 1993, p. C-2.
5. G. Miller, G. Baber, and M. Gilliland, "News On Demand for Multimedia Networks," *Proc. ACM Multimedia 93*, ACM, New York, 1993, pp. 383-392.
6. S. Bielski, "Coming Home: Long Focused on Office Technology, Computer-Makers Are Eyeing the Emerging World of Interactive TV," *Boston Globe*, June 13, 1993, p. 87.

To support various clientele, multimedia servers of suitable storage capacities and network switches of suitable bandwidths are necessary. For instance, for a residential neighborhood with 100 clients, each requiring independent selectivity, a neighborhood server must simultaneously store 100 videos. If the videos are one hour long and are MPEG-2 encoded (requiring data rates of 4 Mbits/s), the total storage requirement is 4 × 100 × 60 × 60 Mbtes, divided by 8, which equals 180 Gbytes.

Assuming storage on magnetic disks, each with a capacity of 10 Gbytes and a transfer rate of 4 Mbytes/s, the neighborhood server requires an array of 18 disks. As for the transmission speeds, an asynchronous transfer mode (ATM) switch of 400 Mbits/s bandwidth suffices for the residential neighborhood (since upstream bandwidth for communicating clients' requests to the content providers is negligible compared to the bandwidths necessary for downstream transmission of digital video and audio programs).

A university campus digital library, in addition to a variety of audio-visual information, must store a vast amount of textual information as well. For instance, the entire Encyclopaedia Britannica spans 29 volumes and more than 30,000 pages, constituting close to 300 Mbytes of text. The Scripps collection (which includes historical oceanographic data, manuscripts, notes, slides, charts, and the like) contains up to roughly 100 Gbytes of text. Assuming that the audio-visual information requires 20 times as much storage, and that up to 10 percent of all information is stored on-line, the university library needs a multimedia server with an array of 21 disks.

The different multimedia servers in a metropolis are colocated with the network switches, to

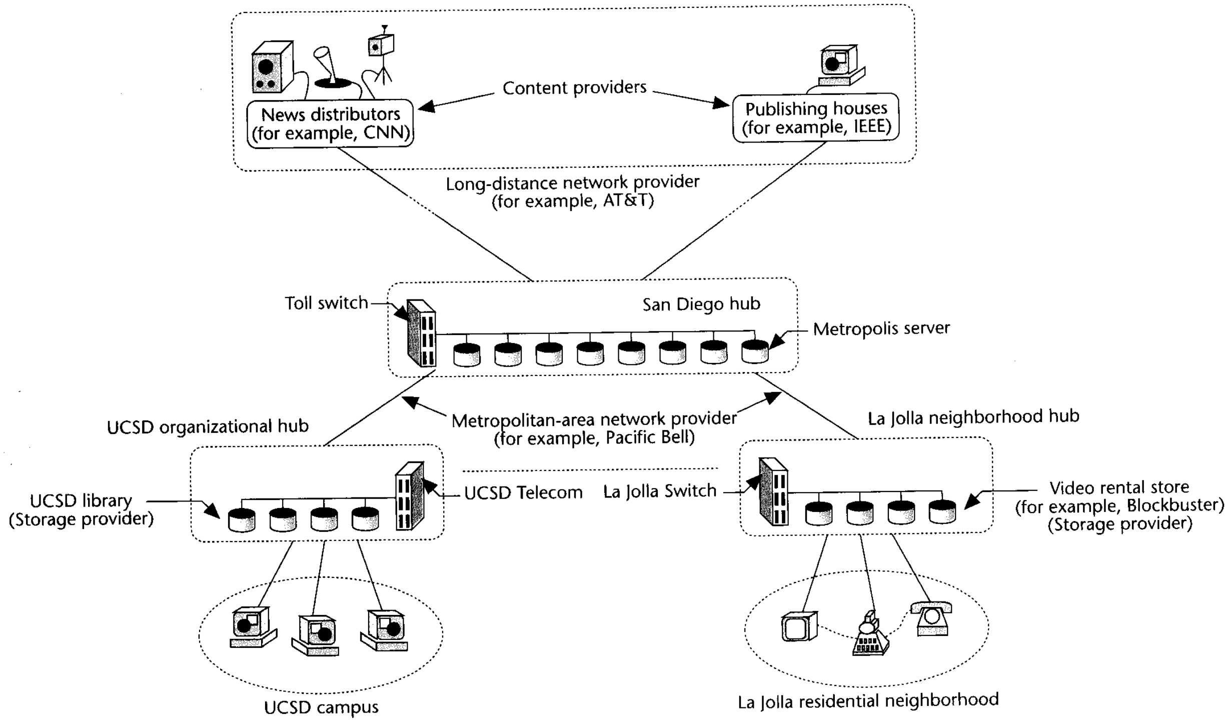

Content providers
News distributors (for example, CNN)
Publishing houses (for example, IEEE)
Long-distance network provider (for example, AT&T)
Toll switch
San Diego hub
Metropolis server
UCSD organizational hub
Metropolitan-area network provider (for example, Pacific Bell)
La Jolla neighborhood hub
UCSD library (Storage provider)
UCSD Telecom
La Jolla Switch
Video rental store (for example, Blockbuster) (Storage provider)
UCSD campus
La Jolla residential neighborhood

Figure 1. A hierarchical configuration of multimedia servers and network switches for the San Diego metropolis: Organizational and residential clients subscribe to the content providers (for example, CNN and IEEE), via the library and video rental caches.

constitute multimedia hubs at strategic places in the network. The arrangement of these hubs mirrors the hierarchical nature of the interconnecting network, with a large-capacity, metropolitan-area hub at the highest level functioning as a gateway to a host of services. The lower levels contain smaller capacity neighborhood and organizational hubs that serve as caches for information originating from the metropolitan-area hub. Figure 1 illustrates a two-level service configuration for the metropolis of San Diego.

The distinguishing characteristic of personalized services is amenability to client requirements. Clients not only have the freedom to choose and procure from the various service offerings, but also to access an entirely new spectrum of services automatically customized (by PSAs) to suit their individual needs. The quintessential example of such a service is a personal entertainment channel: Programs on a personal channel embody a client's viewing preferences, relating both to the choice of programs and viewing times. For instance, for baseball buffs, many of the programs on their personal channels pertain to baseball; even the news program presents baseball results and highlights in detail.

While the content providers function mainly as program creators, PSAs are mediators that negotiate with the various content providers to ensure that clients receive personalized channels. To choose programs for personal channels, PSAs continuously monitor and analyze clients' preferences and construct profiles characterizing both the momentary and long-term preferences of each client. The PSAs use the profiles thus constructed, in conjunction with content descriptions provided by the content providers, to ensure that clients receive only what they desire and no more. By doing so, PSAs also benefit the content providers, since PSAs help direct potential clients to content providers. However, some clients may prefer to explicitly specify (instead of letting the PSAs automatically select) programs on their personal channels. For such clients, PSAs function merely as program schedulers: third parties responsible for media distribution and retrieval at appropriate times, but who have no say in choosing programs.

Content providers

Multimedia presentations, such as news, lectures, and movies, are structured as multimedia documents. Therefore, the primary function of content providers is electronic publishing. For instance, a news provider, such as Cable News

Network (CNN), composes different news segments pertaining to political, domestic, international, financial, and sports news into a complete news document. Content providers make their published documents available at metropolitan area storage providers, relying on directory servers to advertise those documents. The directory servers maintain not only locations but also descriptions of documents, so as to handle content queries from clients and PSAs. The Global Network Navigator and the World-Wide Web are examples of emerging directory servers.

PSAs continuously monitor and analyze clients' preferences and construct profiles characterizing both momentary and long-term preferences of each client.

Most existing directory servers only manipulate alphanumeric information, and hence rely on simple search techniques that match keywords, author names, or document titles to locate information. However, such rudimentary keyword, name, and title entries do not sufficiently describe the content of multimedia documents. Consequently, directory servers must employ sophisticated content-based indexing techniques. For content extraction and analysis, they employ image segmentation, feature extraction, and speech processing techniques. Swanberg et al.[1] observed that in many cases, video information is structured: There exists both a strong spatial order within individual frames and a strong temporal order among different frames pertaining to the same scene. An illustrative example is CNN News. The scenes featuring the news anchor have a specific structure with the CNN logo at the top, the anchor's identity at the bottom left, and the Headline News banner at the bottom right. It is this inherent structure of video scenes that can be exploited to extract interesting features of documents. The three major steps in the content extraction phase are[1]

- Identification of key features characterizing the data contained in a document: These features may be histograms, connected components within a section of a scene, or features within a portion of the digital audio signal, extracted via image analysis or speech processing techniques, respectively.
- Comparison of the extracted features with chosen a priori models to identify objects in which clients are interested: In most cases, these objects do not correspond to physical units (such as video frames) that storage servers manage; rather, they refer to scene changes, episodes, and the like. Consequently, storage servers are not suited for performing content extraction. Moreover, the models used for object identification are domain-specific. Therefore, content providers are responsible for performing content extraction.
- Measurement of unique and descriptive parameters of the identified objects: The catalogues maintained by directory servers store these parameters, which the PSAs can later access for use in program selection.

Furthermore, content extraction of programs explicitly selected by clients enables PSAs to deduce clients' predilections and thereby construct clients' profiles.

Storage providers

Storage providers handle real-time storage and retrieval of multimedia documents. Unlike content providers, whose services are, in general, targeted at particular clientele (for example, an entertainment house targets residential clients, and a financial news provider targets business professionals), storage providers support service-independent, back-end access to multimedia documents. Consequently, the design considerations for storage providers differ from those of content providers. Since multimedia services are in general real-time and interactive, the storage provider architecture is tailored specifically for enforcing real-time performance guarantees that preserve continuity and synchronization of multimedia playback at all times. While continuity is an intramedia requirement (that is, between successive video frames or audio samples, both henceforth referred to as media units), synchronization is mainly an intermedia requirement (for example, between the audio and video tracks of a movie). In addition to the "simultaneity" required of intermedia synchronization, playback of higher level components of a document may also be temporally related, as in sequential or partially overlapped playback of two video programs on a personal channel.

To address these requirements of media playback, storage providers employ a hierarchy of three abstractions:

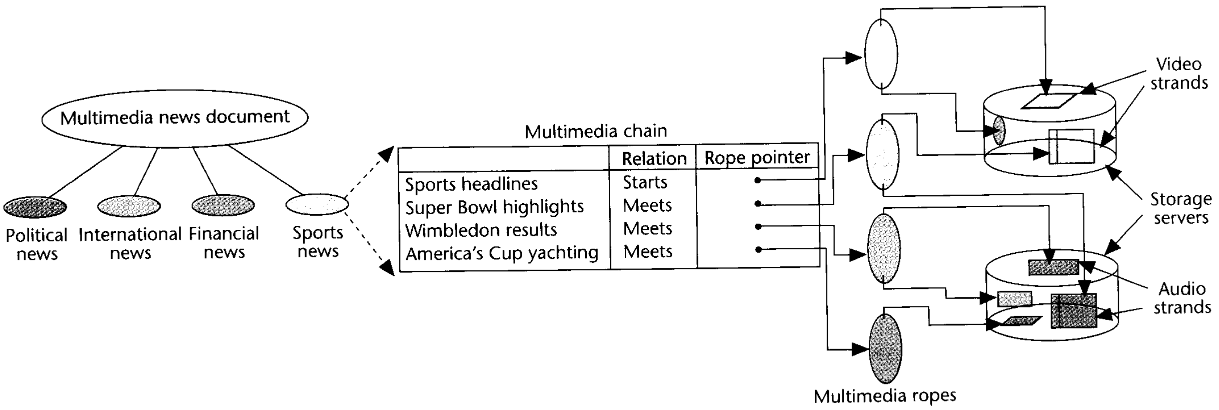


Figure 2. The structure of a multimedia news document. It consists of multiple news segments, each represented as a chain. The Sports News segment is composed of multiple ropes to be played back sequentially (indicated by the meets relation), starting with Sports Headlines.

- A *media strand* embodies the continuity requirement of media playback. A strand is a continuously recorded sequence of media units (of one medium), all stored at the same multimedia server. In Figure 2, while Super Bowl Highlights comprises an audio strand and a video strand, America's Cup Yachting comprises only an audio strand.

- A *multimedia rope* encapsulates synchronization among media strands. A rope is a collection of media strands tied together with simultaneity relations. In the example of Figure 2, each of Super Bowl Highlights, Wimbledon Results, and America's Cup Yachting is a rope.

- A *multimedia chain* captures higher level temporal relationships among ropes. In Figure 2, the Sports News Segment is a chain that consists of ropes: Super Bowl Highlights, Wimbledon Results, and America's Cup Yachting, played in sequence. Chains can be recursive: a higher level chain can be made up of temporally related lower level chains. Hence, personal channels can be represented as chains, with constituent programs being ropes or chains.

Let's look at these abstractions in more detail.

Media strands

To ensure continuous playback of media strands, storage providers employ constrained placement of media blocks. The *granularity* (the size of each media block) and *scattering* (the separation between successive media blocks) of a media strand are determined such that the time to access each media block is bounded by the playback duration of the block.[2]

While constrained placement suffices for ensuring the continuity of one client request, satisfying multiple clients without violating any of their continuity requirements forces the storage providers to employ sophisticated admission control mechanisms. For continuity-guaranteed servicing of multiple clients, storage servers employ the optimal Quality Proportional Multi-Client Servicing (QPMS) algorithm, in which the server proceeds in periodic rounds, retrieving blocks proportional to playback rates of clients' requests in each round.[2]

Multimedia ropes

Multimedia ropes maintain simultaneity relations between media strands. The media strands constituting a rope may be recorded at different media capture sites on the network (for example, audio at a microphone, video at a camera), possibly at different content providers (entertainment houses, news distributors, and so forth), sometimes even at different times (for instance, audio dubbing in movies), but they might need to be played back synchronously. Hence, the simultaneity relations among the media strands of a rope are represented as *relative time stamps* (RTSs), with each of the media units of the rope assigned an RTS indicating its time of recording relative to that of all other units.[3]

A storage provider can implicitly determine the RTS values of media units based on the recording times of media units. Synchronization of this type, called *implicit* or *natural* synchronization, is typified by the simultaneity relations between audio and video strands of a live program. Alternatively, simultaneity relations may be explicitly specified, possibly by the content provider, at the time of creating the rope. A typical example of such *explicit*, or *synthetic*, synchronization is dubbing a video strand with an audio strand, as in movie editing.

The RTS representation is sufficient to enforce other synchronization requirements of multimedia services as well: Whereas the definition of an RTS implies *relative* synchronization, by basing the time scale of the relative time system on a real-time clock, a storage provider can enforce *absolute* synchronization. Furthermore, RTSs assigned to media units can be enforced throughout the duration of playback (for example, "lip-synching"), or only at discrete instants (for example, displaying textual subtitles for video images). The former is a case of *continuous* synchronization, while the latter is a case of *discrete* synchronization.

Consider a chain composed of a tennis match followed by a movie: Any unanticipated extension in the duration of the match delays the start of the movie.

For interstrand synchronization, a storage provider must ensure that playbacks of all the media strands of that rope commence simultaneously and progress in lockstep. To do so, the storage provider utilizes lightweight feedbacks transmitted by clients' display sites at the time of media units' playback initiation. Using these feedbacks, the storage provider estimates the RTS values of media units played back together. Mismatches in these values indicate asynchrony among the display sites. The storage provider then steers the display sites back to synchrony by selectively skipping or pausing an appropriate number of media units of the corresponding media strands.[3]

Multimedia chains

Multimedia chains capture higher-level temporal relations among ropes. Content providers specify these temporal relations at the time of chain creation using mechanisms such as timed petri nets[5] and temporal dependency graphs. Thirteen temporal relations are possible among ropes: the equals relation, together with the before, meets, overlaps, during, starts, and finishes relations and their respective inverses.[4] These temporal relations are coarse-grained (in contrast to the fine-grained synchronization requirements amongst strands of a rope) and generally elastic. For instance, consider a chain composed of a tennis match followed by a movie; any unanticipated extension in the duration of the tennis match delays the start of the movie.

During retrieval of a multimedia chain, a storage provider composes a retrieval schedule consistent with the temporal relations among the ropes of that chain. For coordinating retrieval of multiple ropes, the storage provider employs protocols, such as those proposed by Little and Ghafoor,[4] for enforcing petri-net specifications of temporal relations. In the event of unanticipated changes in the the playback durations of any of the ropes, the storage provider dynamically recomputes the retrieval schedule for the chain.

Network providers

Network providers are responsible for media transmission between storage servers and clients' display sites. Both local-area and long-distance networks are evolving towards ATM technology. In such networks, media units are packetized and transmitted by the storage servers and routed via switches to clients' display sites. Video and audio (being continuous media) require guarantees of minimum bandwidth and maximum end-to-end delay, delay jitter, and loss. For instance, for live viewing of JPEG video, the minimum bandwidth required is 4 Mbits/s, and the acceptable limits on end-to-end delay, jitter, and fractional loss are 200 ms, 5 ms, and 0.1, respectively. These requirements are in general specific to the media and the encoding scheme employed. For example, pulse code modulation (PCM)-encoded audio is much more susceptible to jitter and loss than JPEG-encoded video. To support these requirements, network providers reserve bandwidth on all the links and buffer space at all the switches along routes between storage servers and clients, employing techniques such as those proposed by Ferrari and Verma.[5]

Network providers might also be required to support non-real-time transmissions. As we will see later, a neighborhood server can cache media information from a metropolitan area storage server. At the time of caching, the intervening transmission between the two storage providers must preserve the integrity of media information, for which the network providers can trade delay and jitter.

Personal service agents

Personal service agents, as their name implies, play a central role in tailoring the fabric of multi-

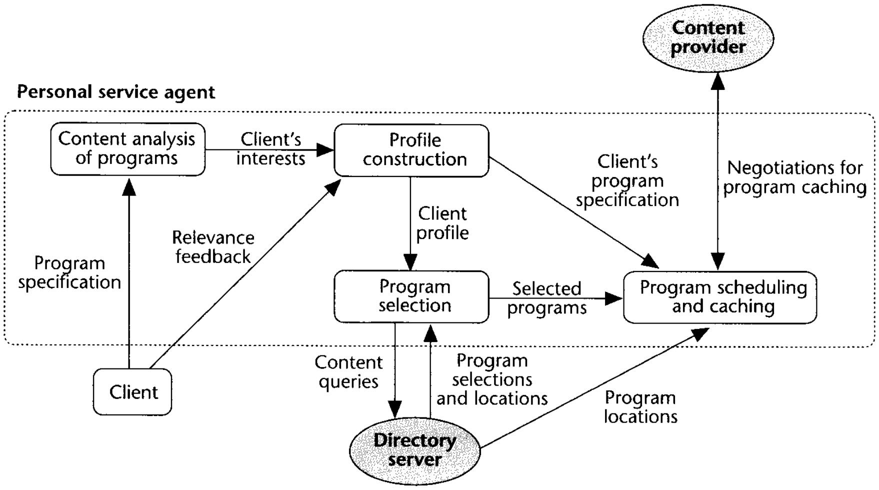


Figure 3. Architecture of a PSA. Program analysis, profile construction, program selection, and program scheduling and caching constitute the main functions of the PSA.

media services to fit the individual needs and preferences of clients. These PSAs, located at clients' sites, employ content analysis techniques to select programs that closely match clients' interests, then schedule the selected programs for clients' preferred viewing times by judiciously caching the programs at neighborhood servers. In this process, PSAs trade storage for network transmissions to minimize service costs borne by clients. These functions are elaborated in the following sections.

Program selection

Program selection by the PSAs is based on clients' preferences. However, no such preferences are available to a PSA at the time when it first receives a request from a client. Hence, initially, the client has to explicitly choose (or state preferences for) programs to be viewed. By analyzing the content of the programs chosen by the client (content descriptions are available in directory servers' catalogues), the PSA starts to infer the client's likes and dislikes (see Figure 3). Since the client's preferences could vary depending upon the context (like time of day), the PSA must not only maintain a history of programs chosen by the client, but also information about the context in which those choices were made. The context may be explicitly specified by the client, or alternatively, inferred by the PSA itself, depending on the client's choice of programs. For instance, the choice of a movie classified as a comedy or tragedy leads the PSA to infer that the client's mood is boisterous or melancholy, respectively. By monitoring the client's choices, the PSA constructs a behavioral profile for the client, characterizing contextual preferences.

Based on the client profile thus constructed, the PSA queries directory servers to determine programs relevant to the client. Owing to the multitude of programs offered by content providers, program selection is typically a multistage process. The preliminary stages filter out programs that are totally uninteresting to the client. The final stages involve more detailed comparisons of program content descriptions with client preferences.

The PSA solicits explicit "relevance feedback" from the client on programs it selects. In the rudimentary case, the client's feedback indicates the programs to be very relevant, relevant, irrelevant, or very irrelevant. A much more comprehensive feedback can identify program features (such as scenes, episodes, and players) that the client found appealing or distasteful. The PSA uses such feedback to continuously refine the client's profile, using powerful learning algorithms such as those proposed by Belew.[6]

The PSA arranges into a multimedia chain all the programs selected on a personal channel, together with the timing relationships between their preferred display times. It then negotiates with the content providers to retrieve the chain.

Program retrieval and caching

The different programs constituting a chain that represents a personal channel could be offered by different content providers. The PSA judiciously schedules retrieval (from storage providers) and subsequent transmission (via network providers) of the programs to minimize client costs. For instance, a PSA can retrieve programs a priori and cache them either at the client's site or at a neighborhood server during a

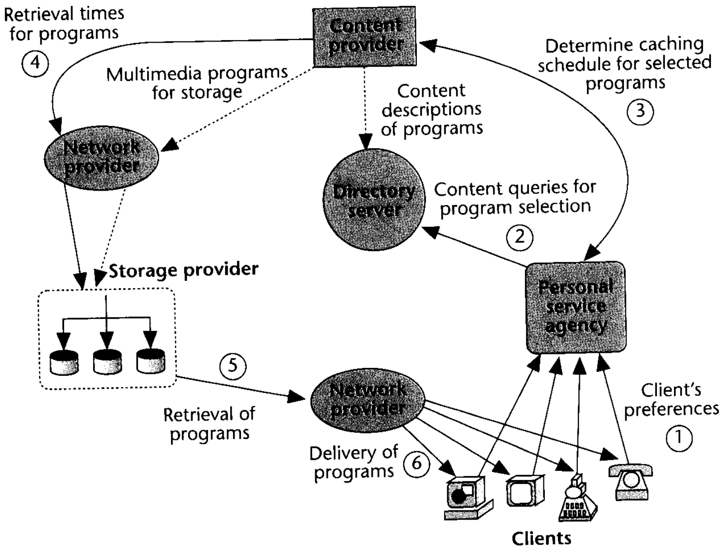


Figure 4. System architecture of a personalized multimedia service. The dotted lines represent interactions at the time of creation and storage of new programs; the solid lines denote program selection and retrieval. The numbered arrows indicate the chronology of interactions. The PSA first creates the client's profile characterizing preferences. Based on this profile, the PSA selects a program for the client's personal channel with the help of the directory server. Then, in conjunction with the content provider, the PSA determines a caching schedule for the selected program. This schedule is then conveyed to the storage provider, which is thereafter responsible for retrieving and delivering the program to the client.

period when the network and server are relatively underutilized. In some cases, these schedules might need to be dynamically altered. Such alterations can be client-initiated. For instance, a client who is delayed at the office might ask the PSA to delay his or her favorite programs that evening. Alternatively, the PSA itself may reschedule programs when it detects the availability of newer programs that are more relevant to the client. To detect such programs, the PSA can either periodically browse through directory servers' catalogues or employ intelligent "knowbots" that constantly look out for new program offerings and inform the PSA when they locate a new program.

Content providers also implement resource optimizations. They attempt to amortize retrieval and transmission costs among several clients choosing the same programs. In the best case, multiple clients may have similar preferences, both in their choice of programs and in their viewing times. For such clients, a content provider simulcasts programs of their common choice. When clients' preferred viewing times do not match, PSAs can arrange with the content providers to cache their program at a neighborhood server, rather than incur the cost of repeated transmission all the way from a metropolitan server. Figure 4 depicts the various interactions between PSAs and the other system components.

In practice, interesting trade-offs exist between the cost of renting storage space at a neighborhood server and the cost of repeated transmissions from a metropolitan server. PSAs and content providers must evaluate these costs before deciding whether to cache programs. To see why, consider the simple, tandem configuration of storage servers shown in Figure 5a, where S_1 is a metropolitan-area server and S_2 and S_3 are neighborhood servers. Suppose that a program P, created by a content provider at 2 p.m. and stored at S_1, is requested by clients C_1, C_2, and C_3 for viewing at 2 p.m., 3 p.m., and 12 a.m., respectively. Further, suppose that (1) the storage cost for program P at a server increases with duration of storage and (2) the cost of each transmission of P over the network links is fixed.

Clearly, no optimizations are possible for servicing the first client C_1 at 2 p.m.: The program must be transmitted all the way from server S_1. However, during this transmission, any (or even both) of the servers S_2 and S_3 may cache the program for future usage. Caching at S_3 entails only a storage cost. Caching at S_2 entails an additional network cost, but a lower storage cost. Considering such trade-offs, for client C_2 it is optimal to cache the program at S_3 during the period [2 p.m., 3 p.m.], incurring a storage cost of $0.50. On the other hand, from client C_3's perspective, it is better to retransmit the program from S_1 rather than cache it at either S_2 or S_3. The caching schedule thus derived (see Figure 5b) entails a total cost of $3.92. Although it optimizes individual service costs for each of the clients, it does not constitute

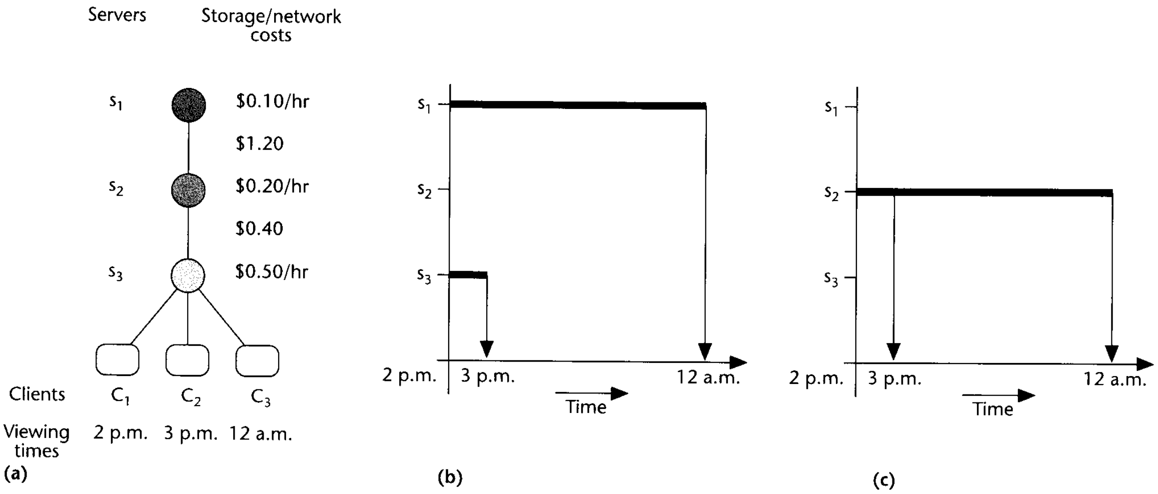


Figure 5. Network-storage trade-offs. (a) The network configuration: S_1 is the metropolitan-area server; S_2 and S_3 are the neighborhood caches. Clients C_1, C_2, and C_3 request program P, created and stored at S_1 at 2 p.m., for viewing at 2 p.m., 3 p.m., and 12 a.m., respectively. The storage and transmission costs for P are indicated in the figure. (b) A caching schedule that minimizes individual costs for each client. The total cost of this schedule is $3.92. (c) The overall optimal schedule, the total cost of which is $3.83.

the overall optimum. For instance, a schedule that caches program P at server S_1 throughout the period [2 p.m., 12 a.m.] (see Figure 5c) amortizes storage costs between C_2 and C_3 and hence entails a lower overall service cost, amounting to $3.83.

In general, storage-network optimization is performed at the time of scheduling program retrieval for clients' personal channels. The content provider must determine when, where, and for how long the program must be cached to minimize the cumulative storage and transmission cost, amortized over all the clients choosing that program. Such an optimal caching schedule must not only account for differences in storage and transmission costs, but also must adapt to any changes that occur in these costs. (For instance, servers and networks are in greater demand during "peak hours," hence more expensive to use.) Cost changes may occur even on the fly; for example, a storage provider, anticipating an overcommitment of server resources, might increase the storage cost to discourage further demand.

At the University of California, San Diego (UCSD), we have developed architectures and caching techniques to optimize delivery schedules for programs both within a single metropolitan area network and across multiple networks.[7] The techniques use an intricate dynamic programming method to minimize the combined costs of storage and transmission to clients. The techniques also adapt to changes in storage and network parameters.

Conclusions

Until now, attention has focused mainly on designing storage servers and networks to support multimedia storage and delivery, respectively. Many prototype implementations of multimedia servers exist,[3,8] and several telephone companies are carrying out field trials to evaluate the effectiveness of multimedia services in real-life scenarios.[9] Owing to the lucrative residential consumer market, video-on-demand entertainment services are likely to be the first of numerous multimedia services offered in the near future.

We envision that personalized multimedia services described in this article will trigger a radical change in the role of many enterprises. For instance, video stores will become storage providers; cable companies and telephone carriers will become network providers. Libraries too are expected to take on a different role: Rather than being a central storehouse for information, libraries will function more as rapid-access mechanisms to geographically distributed multimedia databases.

The practical realization of personalized multimedia services poses a number of interesting and challenging problems in the areas of multimedia content-extraction, database organization, information retrieval, and, most importantly, the cost feasibility of unprecedented storage and transmission demands.

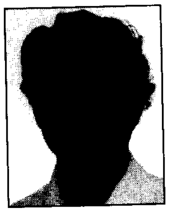
At the University of California, San Diego, we have developed prototype architectures for the storage providers[3] and optimal caching techniques to minimize the combined costs of storage and transmission of multimedia programs to clients across metropolitan-area networks. We are working on information analysis and feature-extraction methods for content-based multimedia retrieval by the PSAs and are integrating them with adaptive, connectionist approaches developed for retrieving text information from a corpus in which various documents have little structure in common.[6] Researchers in library science, visual arts, music, medicine, and psychophysics are working with us to expand applications of personalized multimedia services. **MM**

References

1. D. Swanberg, C.-F. Shu, and R. Jain, "Architecture of a Multimedia Information System for Content-Based Retrieval," In *Proc. Third Int'l Workshop on Network and Operating Systems Support for Digital Audio and Video*, Springer-Verlag, Berlin, 1992, pp. 387-392.
2. H.M. Vin and P.V. Rangan, "Designing a Multi-User HDTV Storage Server," *IEEE J. on Selected Areas in Comm.*, Vol. 11, No. 1, Jan. 1993, pp. 153-164.
3. P.V. Rangan, H.M. Vin, and S. Ramanathan, "Designing an On-Demand Multimedia Service," *IEEE Comm.*, Vol. 30, No. 7, July 1992, pp. 56-65.
4. T.D.C. Little and A. Ghafoor, "Multimedia Synchronization Protocols for Broadband Integrated Services," *IEEE J. on Selected Areas in Comm.*, Vol. 9, No. 9, Dec. 1991, pp. 1,368-1,482.
5. D. Ferrari and D.C. Verma, "A Scheme for Real-Time Channel Establishment in Wide-Area Networks," *IEEE J. on Selected Areas in Comm.*, Vol. 8, No. 3, Apr. 1990, pp. 368-379.
6. R.K. Belew, "Adaptive Information Retrieval," *Proc. 12th Int'l Conf. on Research and Development in Information Retrieval*, June 1989, pp. 11-20.
7. P. Venkat Rangan, "System for Efficient Delivery of Multimedia Information," US Patent pending, San Diego, Calif., 1994.
8. F.A. Tobagi et al., "Streaming RAID: A Disk Storage System for Video and Audio Files," *Proc. ACM Multimedia 93*, ACM Press, New York, pp. 393-400.
9. J. Sutherland and L. Litteral, "Residential Video Services," *IEEE Comm.*, Vol. 30, No. 7, July 1992, pp. 36-41.

Srinivas Ramanathan is a doctoral candidate in the Department of Computer Science and Engineering at the University of California, San Diego. His research focuses on architectures and protocols for multimedia services.

Ramanathan received the B.Tech. degree in chemical engineering from Anna University, Madras, India, in 1988, and the M.Tech. degree in computer science from the Indian Institute of Technology, Madras, India, in 1990. He is a recipient of several awards, including an IBM Doctoral Fellowship.



P. Venkat Rangan directs the Multimedia Laboratory at the University of California, San Diego, where he is an assistant professor of computer science. He serves as editor in chief of the ACM/Springer-Verlag *Journal of Multimedia Systems* and was the program chair of ACM Multimedia 93 (First International Conference on Multimedia).

Rangan received his B.Tech degree in electrical engineering at the Indian Institute of Technology, Madras, India, where he was awarded the "President of India Gold Medal" in 1984. He earned his PhD in computer science at the University of California, Berkeley, in 1988. Recently, he received an NSF National Young Investigator Award.

Readers can contact Rangan at the Multimedia Laboratory, Department of Computer Science and Engineering, University of California at San Diego, La Jolla, CA 92093, e-mail venkat@chinmaya.ucsd.edu.