# Network Transmission and Storage: Vertical Relationships and Industry Structure

John Chung-I Chuang
University of California, Berkeley
chuang@sims.berkeley.edu

**Abstract**

Distance still matters over the Internet. Network storage services, such as caching and replication, are employed by network operators and publishers to bring data closer to their clients. This paper examines the vertical relationships between the nascent network storage services industry and the more established, but still continually evolving network transmission services industry. We find that significant technological and transactional efficiencies may be realized through the joint provisioning of network storage and transmission services. However, the benefits of vertical integration may be diminished, or even negated, if an integrated provider with market power in the transmission domain exercises it to foreclose competition in the storage domain.

27th Telecommunications Policy Research Conference, September 1999

## 1. Introduction

Telecommunications policy has, historically, been shaped and governed by the unique economics of wires and switches. The construction of any telecommunications infrastructure requires huge, up-front capital investments, but upon its completion, the marginal cost of supporting a communication is virtually zero. The fact that network transmission involves high sunk costs and exhibits significant scale economies creates many opportunities for market failure, thus warranting a watchful eye by the social planner.

Yet, despite the attention paid to transmission links and switches, network storage has always been an important element of any communications network as well. In traditional circuit-switched networks supporting real-time applications, such as the public switched telephone network, control and management information must be stored and made accessible in the network to support phone conversations. In store-and-forward networks such as the Internet, user data (including real-time data) are buffered multiple times in the network as they travel from source to destination. Now, as the Internet becomes the ubiquitous infrastructure for global distributed computing and communication, network storage has taken on a central and explicit role, alongside network transmission. In an attempt to bring data closer to the clients, user data of various shapes and sizes (email messages, web pages, video clips, etc.) are stored and retrieved from network storage elements deployed at distributed locations throughout the network.

This paper focuses on network storage, and its relationship to network transmission, as they pertain to distributed communication over the Internet. Figure 1 provides a sample of network services and applications that are enabled by a combination of transmission and storage resources. In this figure, network storage is shown as downstream of network transmission. A local access transport (LAT) service provides basic connectivity to the point-of-presence (POP) of the retail Internet access provider (IAP); the IAP deploys transmission

resources (aggregators, routers, etc.) to provide connectivity to the Internet proper, as well as storage elements to support value-added applications such as email, web-access, etc. Similarly, a wide-area transport (WAT) service provides transport over the Internet backbone; an Internet Service Provider (ISP) or another independent entity operates data centers at key network nodes with storage and processing elements to support content hosting, distributed databases, application hosting, e-commerce and other capabilities. Clearly, the vertical relationships between network transmission and network storage will have a strong influence on their vertical relationships with the downstream services and applications, and vice-versa.

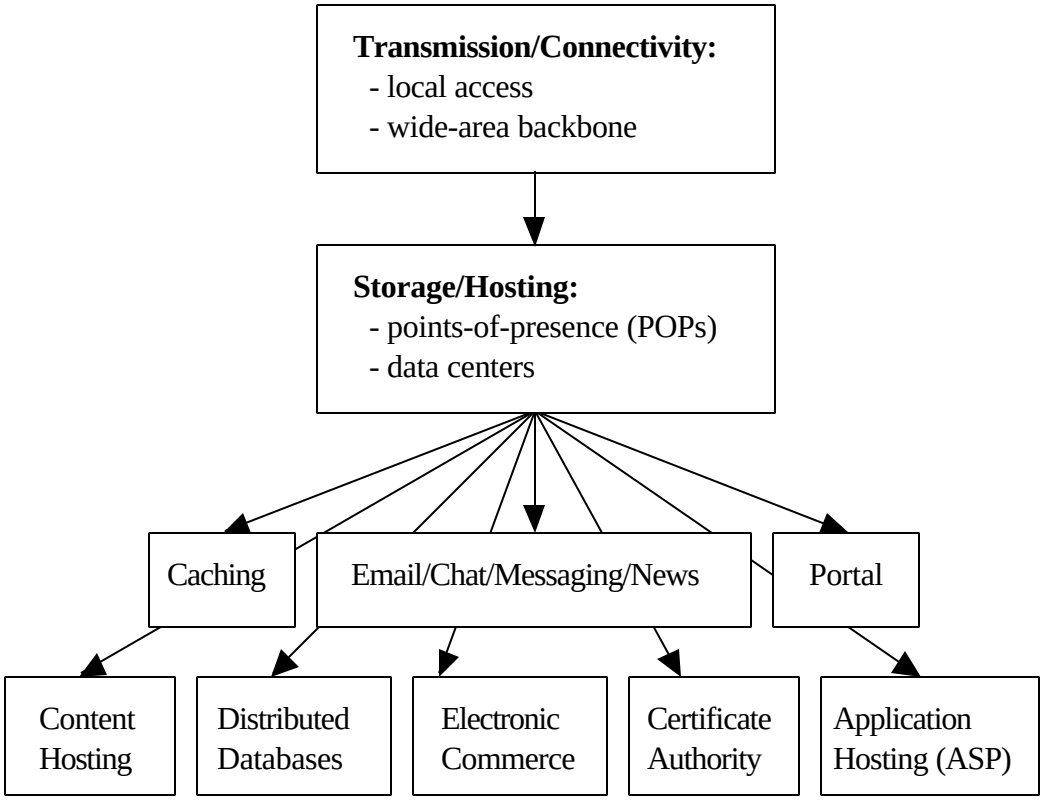



**Figure 1. A sample of downstream services and applications supported by a combination of network transmission and storage resources.**

The network storage service industry is poised for tremendous growth over the next several years. Forrester Research estimates the nascent data center industry to be $2 billion today, and growing by over 700% over the next four years. IAPs and ISPs have also been

investing significantly in the deployment of caching hardware. New and potential entrants to this market include traditional telecommunications powerhouses, startup firms, as well as non-telecommunications companies (e.g., Intel).

Network storage and network transmission have very close relationships at various locations in the network. In fact, the storage and transmission elements are often physically co-located at the strategic network locations, and operated by the same entity. The objective of this paper is to understand, from both technical and economic perspectives, the vertical relationships between various different network storage services and the underlying transmission services, and their implications to the industry structure of the Internet. Since this industry is undergoing rapid growth and change, we caution against any static reading and application of the results presented in this paper.

## 1.1 Related Work

The telecommunications industry is undergoing significant changes in this period of convergence, de-regulation, and competition. In particular, vertical expansion and integration are occurring across traditional market boundaries, even for the Internet industry, which operates on the premise of well-defined architectural layers and open standardized interfaces.

Gong and Srinagesh (1995) offer one explanation. Network transmission is fast becoming a commodity business and the presence of excess capacity is putting pressure on the network operators to price at marginal cost. However, marginal cost pricing will not permit the operators to recover their fixed costs. One way to realize above-marginal-cost revenues is to provide value-added services in downstream markets. This is reflected in the marketplace today, as free Internet access is bundled with various other products and services.[1] Kassavalis et al. (1997) offer an alternate strategy that does not require vertical integration -- product

differentiation via Quality-of-Service (QoS) provisioning. Recognizing that some customers (or applications) are more sensitive to network performance, the Internet engineering community has established frameworks for supporting network transmission services with different QoS levels (Braden et al., 1994, Blake et al., 1998). A network operator can thus offer different grades of transmission services and charge a higher price for those customers with a higher willingness-to-pay. Of course, a combination of vertical integration and product differentiation would work as well. Finally, Katz and Woroch (1997) offer a set of four broad strategies that may be undertaken by a network operator: (i) bundle conduit with content; (ii) bundle intelligence with transmission, (iii) bundle multiple services, and (iv) develop propriety networks. Indeed, bundling (in its various manifestations) and a propriety, closed network are possible outcomes of vertical integration.

Lehr (1998) provides a framework for studying vertical integration in the Internet, though he focused exclusively on the transmission/connectivity segment of the industry. His framework identifies four component stages: local access transport (LAT), retail Internet access provision (IAP), wide-area transport (WAT) and Internet service provision (ISP), which I adopt in this paper for clarity and consistency.

The remainder of the paper is organized as follows. Section 2 provides a technical background of network storage, identifying the different flavors of network storage services and classifying them according to different technical and economic characteristics. Section 3 examines, for each of the major classes of storage services, their vertical relationships with the transmission substrate. Section 4 discusses some of the policy implications of this work.

[1] AltaVista, NetZero, ZapMe! and Bigger.net offer free Internet access supported by advertisement revenue alone.

## 2. Network Storage Services

There are many different flavors of network storage services, each with different characteristics and purposes. They can be distinguished along one or more of the following dimensions:

- type (caching, push caching, replication, hosting, co-location, etc.)
- location (client, server, network: data center, point-of-presence)
- number of copies (from one to millions)
- initiator/beneficiary (consumer, network operator, publisher)
- service "provider" (consumer, network operator, publisher, independent)
- application (different content types & characteristics: size, demand,...)

### 2.1 Type

The most common type of network storage service is caching. A cache makes a local copy of a data object with the hope of satisfying future requests for the same object using this local copy. Caching is primarily motivated by data reuse, with the benefits of bandwidth savings and latency reduction. Data access patterns tend to exhibit strong references of locality (i.e., a recently accessed object is more likely to be accessed again in the near future). Therefore, caching can often eliminate repeated accesses to the more popular objects. However, finite cache sizes mean that old objects will have to be purged to make room for new ones. Inevitably cache misses occur, in which case the object requests are forwarded to another cache higher up in the hierarchy or to the original source itself, incurring a delay penalty. Caches therefore implement different object replacement policies (e.g., least recently used [LRU], least frequently used [LFU], greedy-dual, etc.) to maximize their hit rates.

Push caching and prefetching are variants of simple caching, where objects are pushed or pulled to the cache without an explicit object request from the client. Instead of being demand-driven, these two schemes populate the caches with objects based on projected popularity (or any other criteria). Several companies are using a combination of satellite and multicast technologies to push 'hot' content to network caches.[2] Another variant is differential caching, where 'premium' data objects are subject to a preferential replacement policy when they enter a cache (Chan et al., 1999). With these variants a new business model is emerging -- whereas the duration of cache residency was influenced by object popularity in simple caching, willingness-to-pay may now influence cache residency in these new caching schemes.

Replication, on the other hand, is publisher-driven placement of data objects at one or more distributed locations in the network. Unlike demand-driven caching, replication gives the publishers control over the number, location and lifetime of copies of their objects. Web hosting is the most common replication strategy, where the publisher out-sources the operation of its entire site to a web-hosting service provider. A simple low-traffic site may be hosted at a single network location, while high-traffic sites are increasingly replicated at multiple locations to reduce access latency and distribute server loading. Distributed hosting provides location transparency, in that clients are automatically pointed to the best replica at a given time. This is an improvement over traditional site mirroring where clients have to explicitly select a mirror.

Web hosting requires significant setup and teardown costs, and is usually limited to static placements of entire sites lasting for months or years.[3] Several companies are beginning to offer replication services for individual data objects (especially image files) that are most frequently accessed, though these are still limited to long-term contracts with large publishers.[4] In its most general form, replication should be applicable to dynamic placements of individual data objects, and with service duration of hours or less. The key challenge is to reduce the

[2] SkyCache, iBEAM and Edgix are competing firms in this market.
[3] There are high-profile exceptions to this, such as the official sites for the World Cup, Olympic Games, etc., though these arrangements are set up mainly for marketing and technology demonstration purposes.
[4] Akamai and Sandpiper are two startup firms in this business.

costs of transaction and provisioning, so that replication can become as adaptive as caching (Chuang and Sirbu 1999).

Finally, there are some publishers or merchants who require direct, high speed connectivity to the network, but wish to retain control over the operation of their servers. These publishers would purchase co-location services from ISPs and co-locate their servers within the premises of the ISPs' data centers.

## 2.2 Location and Numbers

Network storage services can be performed at various locations in the network, including the two end-points. Web caching, for example, is performed on the client's desktop by the browser, permitting re-use by the recipient (locations <1> in Figure 2). At the next level, a proxy cache is installed at the edge of the network, either at the point-of-presence (POP) of an IAP, or at the gateway of an organization (locations <2>). Since the POP or the gateway is a point of aggregation (all traffic in/out have to pass through this point) the objects stored at the proxy cache can be re-used by any client on a shared basis. Finally, network caches are installed in the network itself, often at key nodes such as data centers within an ISP backbone network (locations <3>). Network caches themselves may be organized hierarchically to maximize efficiency. Replication, on the other hand, started as server-side solutions (e.g., web hosting and mirroring at location <4>) and progressed into the network (e.g., distributed hosting at the data centers <5> and the POPs <6>).

Figure 2 also illustrates facilities-based competition for residential Internet access, the relevance of which will become apparent in later sections. Today, the majority of clients connect to the Internet via a voiceband dialup IAP subscription (like client #1). Increasingly residential customers are able to connect to the Internet via one of several broadband local access transport (LAT) options, such as digital subscriber line (xDSL, like client #2),

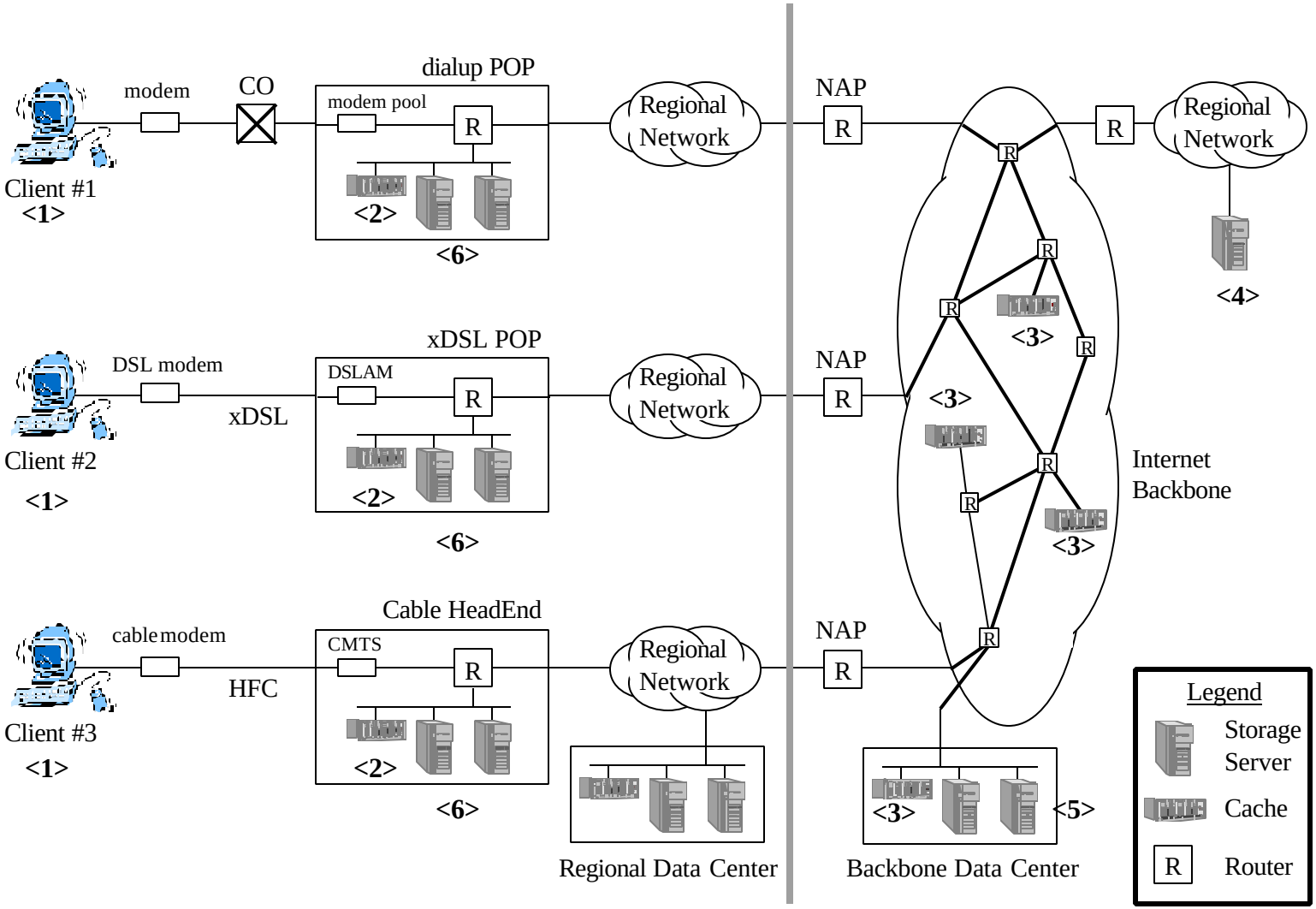



**Figure 2. Locations of storage service elements in a network environment with facilities-based competition.**

hybrid-fiber-coax (HFC) over the cable network (like client #3), satellite and wireless local loop (not shown). Furthermore, it may be possible for multiple IAPs to share the same LAT facility, with each IAP operating its own POP (though I only show one IAP per LAT facility in the figure). From a residential customer's perspective, facilities-based competition will deliver more choices and lower prices for connecting to the Internet. As customers upgrade to a broadband service, the demand for multimedia and other objects with high-bandwidth requirements will rise. This further drives the need to bring data objects closer to the clients.

There is a very strong relationship between the location and the number of the object copies (Table 1). The closer the objects are to the clients, the smaller their coverage area, the more copies are needed. The topology of the Internet plays an important role here. The Internet backbone consists of multiple ISP networks that are well interconnected and highly redundant. Objects placed at a single data center within an ISP network can be reached from practically anywhere on the Internet (barring node or link failures). Local access networks, on the other hand, have a tree-like structure and are highly hierarchical. Therefore, objects placed at a neighborhood POP are intended to be accessible only by clients served by that POP. For example, objects cached or replicated at the cable headend would not be accessible to clients #1 and #2, even if all three clients were geographically located on the same street.

## 2.3 Requesters and Providers

From Table 1, we can see that caching is initiated and implemented by the same entity within their respective 'jurisdictions'. In particular, network operators perform proxy and network caching and realize two tangible benefits: bandwidth savings and latency reduction for their customers. In this sense, caching is both an engineering decision and a business decision. Caching allows the network operator to optimally trade off storage resources for transmission resources, resulting in cost savings over a transmission-only solution. Caching also results in faster data access time for popular objects, and therefore happy customers.

Replication, on the other hand, is always requested by the publisher. Latency reduction, increased data availability and reduced server loading are the primary motivations. Replication services are provided either by independent service providers, or by integrated ISPs/IAPs. In the former case, the independent service providers have to purchase transmission resources (i.e., network connectivity) from the ISPs or IAPs. Then they add storage resources to produce the final product: replication service. In the latter case, the integrated provider jointly provisions the transmission and storage resources to deliver the final replication service. This has important economic ramifications, which I will discuss in the next section.

**Table 1. Characteristics of Selected Network Storage Services**

| Storage Services | Location | # of copies | Requester | Provider |
|---|---|---|---|---|
| Caching | | | | |
| browser caching† | clients | $10^7$ to $10^8$ | client | client |
| proxy caching | POPs | $10^3$ to $10^4$ | IAP‡ | IAP |
| network caching | backbone | $10^1$ to $10^2$ | ISP‡ | ISP |
| Replication | | | | |
| web hosting/mirroring | server(s) | $10^0$ to $10^1$ | publisher | ISP, independent |
| hosting at data centers | backbone | $10^1$ to $10^2$ | publisher | ISP, independent |
| hosting at POPs | POPs | $10^3$ to $10^4$ | publisher | IAP, independent |

† Strictly speaking, browser caching is not a network storage service, and therefore will not be considered in the remainder of this paper.

‡ Following Lehr (1998) I make the distinction between a retail Internet Access Provider (IAP, e.g., AOL, MSN, AT&T WorldNet, MindSpring, @Home, etc.) and a backbone Internet Service Provider (ISP, e.g., AT&T, UUNET, PSINET, GTE/BBN, etc.) However, the boundary between these two markets is rapidly blurring, and both IAPs and ISPs operate POPs and backbone data centers. So the ISP and IAP may in many cases be the same entity (e.g., AT&T).

## 3. Vertical Relationships between Transmission and Storage

This section examines the vertical relationships between the network storage services and the underlying network transmission services. For each of the storage services, I identify possible determinants of vertical integration, empirical evidences of such, and discuss their merits and policy implications.

Economists have identified three main determinants of vertical integration, namely technological efficiency, transactional efficiency, and market imperfections (Perry, 1989). In the first case, vertical integration is motivated by technological efficiencies because there exist economies of scope across the upstream and downstream production stages, such that total production cost is minimized under integrated production. In the second case, vertical integration is motivated by transactional efficiencies, in that market exchanges for the intermediate product is less efficient compared to intra-firm transfer of the intermediate product. This may occur due to the inability to design and/or enforce complete contracts between the parties. Finally, vertical integration may arise due to market imperfections. If either of the upstream or downstream markets is not competitive, those firms that possess market power may employ vertical integration as a leverage instrument to gain unfair advantage in the other, otherwise competitive market. In the first two cases, vertical integration promotes economic efficiency and should be strongly encouraged. In the third case, however, the welfare implications of vertical integration are ambiguous, and careful scrutiny is warranted.

### 3.1 Proxy Caching and Network Caching

Empirically, network caching has always been integrated with the underlying transmission service. Proxy caches are installed and operated by IAPs at their POPs and regional data centers; network caches are installed and operated by ISPs at key locations (e.g., backbone data centers) within their networks.

There are clear technological efficiencies for physically co-locating the caching hardware with the routers and aggregators. The IAPs and ISPs can choose to trade storage resources for transmission resources (since caching realizes bandwidth savings), as well as to trade storage resources for performance improvement (since caching realizes latency reductions).

Furthermore, the IAPs and ISPs are themselves the primary consumers (and beneficiaries) of the caches they operate, so full integration between storage and transmission is the most efficient arrangement from the transactional perspective. Caching is intended to be an internal business/engineering decision, and its deployment and operation should be transparent to both the content providers (publishers) and the content consumers (readers).

### 3.2 Content Hosting

Content providers often choose to outsource the operation of their web servers to web hosting service providers so that they may concentrate on their core business of content creation.[5] Web hosting services range from (i) hosting of personal/community homepages,[6] (ii) hosting of simple sites for small businesses desiring basic web-presence, to (iii) distributed hosting of large commercial sites with replicated sites at various network data centers.

Basic web hosting has little or no barriers of entry -- a provider simply sets up one or more web servers with basic connectivity to the Internet. Therefore, non-integrated independents as well as integrated IAPs can easily enter this commodity business. In the former case, the independents have to purchase network access from an IAP. In this context, network transmission is the intermediate product, and web hosting is downstream from Internet access provisioning. An integrated IAP holds scope (and scale) economies advantages over the independent, especially with respect to facilities co-location and network connectivity. Today,

[5] Effectively, there is vertical dis-integration between content creation and content distribution.
[6] This is exemplified by GeoCities, who was acquired by Yahoo!.

hosting of personal homepages is routinely offered by IAPs as a free, value-added component to the Internet access service bundle.

There is a recent, interesting twist to this story. Residential broadband infrastructures (e.g., xDSL, cable) are now capable of handling upstream bandwidths of 100Kbps to 1Mbps range. This is sufficient bandwidth for an individual to operate a web server from home over a simple subscription service. The IAPs have so far chosen to clamp down on this practice, in an attempt to avoid the cannibalization of its own hosting service.

### 3.3 Distributed Hosting at Backbone Data Centers

Whereas customers of simple hosting services only look to establish a presence on the web, customers of commercial hosting services demand fast, reliable access to their content for their geographically distributed audience. These sites are increasingly replicated across the network at multiple data centers of a hosting service provider.[7] These data centers are located near the major network exchanges and other key network nodes with high bandwidth links.

Like simple hosting and caching, the physical sharing of facilities is a strong technological determinant for integration. The common costs of building and operating the network data centers can be amortized over both the storage and transmission service elements located at the facilities. On the other hand, data center operators are also offering co-location services, where content providers can operate their own hardware in a dedicated and secured floor-space. This suggests that some of the gains from physical co-location can be realized without the bundling of transmission and storage services.

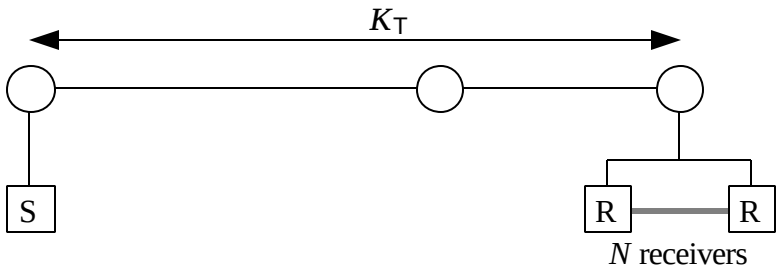
Integration in distributed hosting may also be motivated by transactional economies. Recall that a key goal of distributed hosting is to reduce the average distance traveled by data

packets. Therefore, the bandwidth consumed for transmitting a data packet from any given replicated server is, on average, lower than that for transmitting from a single non-replicated server (Chuang, 1998, Chuang, 1999). This implies that a distributed hosting provider should consume less transmission resource than a simple hosting provider, and should therefore pay less per unit of transmission service purchased from the ISP. Unfortunately this is not true today, where network transmission is generally offered under a distance-insensitive pricing regime (e.g., a packet that travels one network hop is charged the same amount as a packet that travels twenty hops.[8]) This distance-insensitive tariff applies regardless of whether the ISP adopts a flat rate or a usage-based pricing scheme. The cost of metering the distances traveled by data packets is simply too high to justify a distance sensitive pricing scheme.

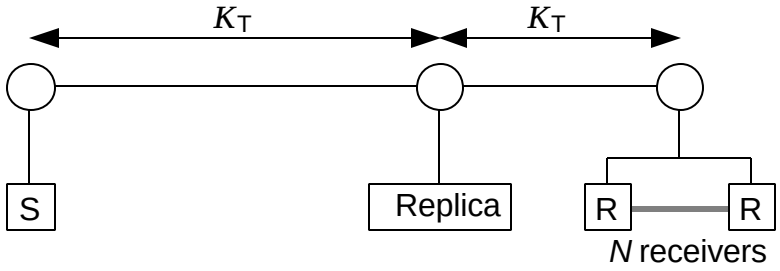
Consider the example in Figure 3, where a content owner wishes to deliver a data packet to $N$ receivers. Under scenario (a), the content owner simply transmits $N$ copies of the packet to the receivers, incurring a transmission charge of $N*K_T$ (where $K_T$ is the cost of transmitting one packet, regardless of distance traveled). In scenario (b), the content owner arranges for a distributed copy to be placed at the data center of an independent storage service provider, bringing the object closer to the receivers. But this independent provider has to purchase transmission service (at a distance-insensitive rate of $K_T$ per packet) from the transmission provider. Therefore, the total transmission charge goes up to $(N+1)*K_T$, not to mention the additional storage costs incurred. In scenario (c), the storage and transmission services are offered by a vertically integrated provider, who is able to internalize the true cost of transmission: $(1-\alpha)*K_T + N*\alpha K_T$. As $\alpha$ approaches zero, the transmission cost approaches $K_T$.

[7] An April 1999 survey by Jupiter Communications reports that 12% of surveyed websites are distributing their content from multiple storage locations, and a further 58% plan to do so in the next year.

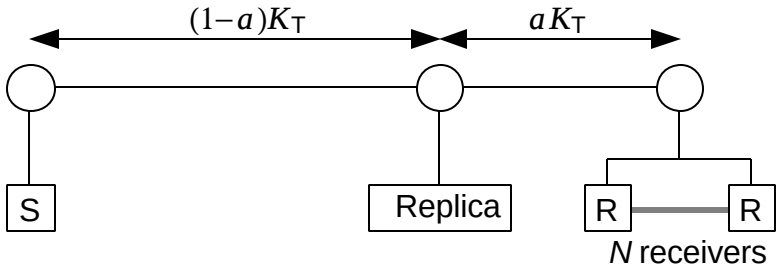
[8] This is analogous to the U.S. postal tariffs, where first-class rate applies to any domestic letters regardless of the distance between sender and receiver.




(a) no replication




(b) replication at independent 3rd party server




(c) replication at vertically integrated server

**Figure 3. Example shows vertically integrated storage provider can internalize transmission cost savings not available to an independent storage provider.**

The net result is that an integrated provider can internalize transmission cost saving that results from distributed hosting, but an independent provider cannot. In effect, the distance insensitive pricing scheme is an incomplete contract, leading to a loss in efficiency that can only be avoided through integration.

There has been significant amount of horizontal consolidation in the national and international ISP market. ISPs derive significant market power from their size and the comprehensiveness of their coverage. Additionally, the bulk of network delay can be attributed to congestion at the inter-ISP exchange or peer points. Therefore, there is strong incentive on the part of publishers to host their content on the ISPs with the largest coverage. Today, the largest web sites are hosted by a small number of large integrated ISPs with national (if not global) coverage.[9] Indeed, some publishers may choose to have their content hosted by multiple ISPs, in order to obtain better latencies for their audiences who are customers of the various ISPs.

There is strong empirical evidence of vertical integration in the distributed hosting/data center market. Exodus, the leading provider in this arena, chose to integrate backwards into the transmission domain. It is operating its own national backbone network with the explicit intent of peering with (rather than being a customer of) other national ISPs. GlobalCenter, the other leading hosting provider, was bought by Frontier who will be bought in turn by GlobalCrossing, a facilities-based carrier. Qwest, the other facilities-based carrier who was interested in Frontier has also entered this market. Digital Island is another integrated ISP. Finally, UUNET recently announced its intention to enter the market with a $100 million capital investment.

### 3.4 Replication at POPs

With the deployment of residential broadband access services, we are beginning to see a rising demand for replication and hosting of broadband content (e.g., video, audio and other streaming content) at the POPs. For example, the @Home network architecture supports replication (as well as caching) of objects at both the headends (POPs) and regional data centers (supernodes).[10] AOL also supports hosting of broadband content at its xDSL POPs.

[9] Top 25 ISPs, Data Communications, June 1999. URL: www.data.com/issue/990607/topisps.html

The same technological efficiencies in Section 3.3 (derived from physical co-location of transmission and storage resources) also apply here at the POPs. However, there is one significant difference between POP replication and data center replication: the lack of interconnection at the local access network level means that objects replicated at POPs are not globally accessible as objects replicated at backbone data centers are. For example, an object replicated at the backbone data center (location <5>) would be accessible to all three clients in Figure 2. Now if the publisher wishes to provide local POP access to its object, and still maintain coverage to all three clients, it will have to place individual copies of the object at each of the three POPs (or headends). The situation is worse if there are multiple IAPs sharing access to the same local access transport (LAT) facility (e.g., xDSL line or HFC plant). Now an object copy has to be placed at each of the POPs attached to the same LAT facility.

So, while facilities-based competition in the LAT market is very desirable from a connectivity perspective (Clark, 1998), the lack of local interconnections between these different LAT facilities may not be so attractive from a hosting/replication perspective.

We can say that the IAPs are operating in two simultaneous markets. In the first, they offer retail Internet access service to subscribers, using storage resources at the POPs to support value-added services such as email, news, chat, caching, portal, and 'fast access to locally hosted content and applications'. In the second market, they sell to publishers and merchants preferential access to their subscription base, via the right to place content, e-commerce capabilities and other applications at the POPs and RDCs.

The first market, faced by the consumers, is a competitive one. Consumers can choose to subscribe to any IAP or IAP-LAT bundle. The second market, faced by the publishers, is not. If a publisher wishes to obtain 100% market coverage, it will have to purchase POP replication service from all of the IAPs. This is similar to the newspaper industry, in which local

[10] @Home publishes information on its network architecture at www.home.net/about/network.html

newspapers face competition in the subscription market, but exercise monopoly power in the advertising market (Slade, 1998).

Many publishers are thus forced to settle for less than full coverage, entering into exclusive agreements to have their content hosted at the POPs and RDCs of a single IAP. AOL, for example, has over the years built up a 'walled garden' of exclusive content available only to its subscribers. Recently, both AOL and @Home have secured media partnerships to host exclusive broadband content at their xDSL POPs and cable headends, respectively. Even though the residential subscribers are not prohibited from visiting sites outside of the 'walled garden', a significant portion of them never leave its confines and venture into the Internet itself.[11]

### 3.5 Co-Location Services and Co-Located Services

Co-location services are offered at data centers to customers who want direct high-bandwidth connectivity to the Internet backbone but wish to supply their own storage and processing hardware. Effectively, co-location is an unbundling of network transmission service from network storage service, and customers are allowed to purchase the former without the latter. While most customers purchase co-location services to run their own web servers, it is possible for a 'broker' to purchase co-location service from one or more data center providers, install its own storage servers, and offer a co-located distributed replication service. By co-locating at data centers of multiple providers, this broker can provide better coverage than any one hosting provider.

Co-location services are also possible at POPs, though none are presently available. Two startup firms have made partnership arrangements with various IAPs and ISPs to place

[11] Almost 75% of AOL members' time are spent within the confines of its walled garden.

storage servers on the premises of POPs and data centers, and sell distributed replication services to major web sites.[12]

## 4. Discussion

We have identified, for the various classes of network storage services, technological and transactional efficiency reasons for joint provisioning with network transmission services. As long as there are no concerns for anti-competitive behavior, vertical integration between network transmission and storage should not be discouraged.

In particular, the various flavors of network caching should be considered a network optimizing effort that is fairly transparent to the end hosts, and implemented by the IAP or ISP as an integrated part of network transmission provisioning. For server-side web hosting, as well as distributed hosting at backbone data centers, vertical integration is fine as long as there are multiple ISPs in the market, i.e., the upstream transmission market is competitive.

The story for POP replication, on the other hand, is a little more ambiguous. While facilities-based competition may become a reality, having multiple IAPs in a given market does not necessarily provide for competition in the POP replication market. This is due to the technical infeasibility of local storage interconnection, which translates into the need to replicate at not one but all of the local facilities to obtain full coverage. It remains to be seen if market forces are sufficient to support co-location, unbundling and resale of POP-based storage services.

[12] Akamai offers a 'FreeFlow ISP' partnership program where ISPs and IAPs can volunteer to house the Akamai servers in their POPs and data centers. Cisco is an investor of Akamai, while AOL (an IAP) and Inktomi (a supplier of network caches) are investors of Sandpiper.

When we accept vertical integration, we are acknowledging that system-based competition is an adequate or even superior alternative to component-based competition. However, as Farrel et al. (1998) have shown, system-based competition may not be as efficient as component-based competition, especially if firm competencies are very different across the different components. One potential danger of vertical integration is that the barrier to entry has been raised, and storage specialists are precluded from the storage market unless they enter the transmission market as well. A critical question is: do we expect firm competencies to be very different for the provisioning of transmission and storage services?

As mentioned at the outset, the network storage service industry is new, expanding, and subject to changes and improvements in technology. The network transmission industry may undergo technological and structural changes as well. Since it is not possible to capture all of these dynamics in this study, I will close this paper by posing some 'what-if' questions.

First, what happens if caching and replication services are unified? There are strong technical and economic motivations (e.g., statistical multiplexing, economies of scope efficiency) for jointly offering caching and replication services from the same physical hardware (Chuang and Sirbu, 1999). For example, storage resources may be reserved for replication services, and any unused capacity may be opportunistically utilized for best-effort caching. What would be the optimal integration strategy for this unified caching and replication service?

Second, economies of scale and market power from network coverage continue to drive horizontal consolidation in the network transmission industry. What might the equilibrium industry structure look like, and what are its implications on the downstream network storage industry structure?

Third, how would the potential unbundling of local access transport (LAT) and retail Internet access provision (IAP) impact the vertical relationships of storage and transmission at the local access network level? For example, if a retail IAP gains access to, and deploys POPs

at, the various facilities-based local access networks, it will obtain a fairly complete coverage of the residential subscription audience. Will this translate into market power for its vertically integrated POP replication service?

In summary, this work is a study of the vertical relationships between the nascent network storage services industry and the more established, but still continually evolving network transmission services industry. These relationships vary significantly depending on the network locations from which the storage services are offered, and the market structure of the underlying transmission service industry. The study finds significant technological and transactional efficiency motivations for integration, and this corroborates well with evidence from the marketplace. However, a watchful eye is needed, especially over the provisioning of POP-based storage services, to ensure that the Internet access providers (who operate the POPs) do not foreclose competition in this increasingly important market.

## References


Blake, S. et al., An architecture for differentiated services, RFC 2475, December 1998.

Braden, R., D. Clark and S. Shenker, Integrated services in the Internet architecture: an overview, RFC 1633, June 1994.

Chan, Y.M., J. Womer, and J. Mackie-Mason, One size doesn't fit all: improving network QoS through preference-driven web caching, Proceedings of Telecommunications Policy Research Conference, 1999.

Chuang, J.C.-I., Economies of scale in information dissemination over the Internet, Ph.D. Dissertation, Carnegie Mellon University, November 1998.

Chuang, J.C.-I., Resource allocation for *stor-serv*: network storage services with QoS guarantees, Proceedings of NetStore'99 Symposium, October 1999.

Chuang, J.C.-I. and M.A. Sirbu, Distributed network storage with quality-of-service guarantees, Proceedings of Internet Society INET Conference, July 1999.

Clark, D., Implications of local loop technology for future industry structure, Proceedings of Telecommunications Policy Research Conference, 1998.

Farrell, J., H.K. Monroe and G. Saloner, The vertical organization of industry: systems competition versus component competition, Journal of Economics and Management Strategy 7(2): 143-182, 1998.

Gong, J. and P. Srinagesh, Network competition and industry structure, Industrial and Corporate Change 5(4):1231-41, 1996.

Kavassalis, P., T.Y. Lee and J.P. Bailey, Sustaining a vertically disintegrated network through a bearer service market. Presented at Internet Telephony Consortium Semiannual Meeting, November 1997.

Katz, M.L. and G.A. Woroch, Introduction: convergence, competition, and regulation, Industrial and Corporate Change 6(4):701-719, December 1997.

Lehr, W., Understanding vertical integration in the Internet, EURO CPR'98.

Perry, M., Vertical integration: determinants and effects. In Handbook of Industrial Organization, Volume 1, edited by R. Schmalensee and R. Willig, Elsevier Science Publishers B.V., 1989.

Slade, M.E., The leverage theory of tying revisited: evidence from newspaper advertising, Southern Economic Journal 65(2), October 1998.