Comparing Hybrid Peer-to-Peer Systems

Beverly Yang Hector Garcia-Molina

Stanford University
Stanford, CA, USA
{byang, hector}@cs.stanford.edu

Abstract

"Peer-to-peer" systems like Napster and Gnutella have recently become popular for sharing information. In this paper, we study the relevant issues and tradeoffs in designing a scalable P2P system. We focus on a subset of P2P systems, known as "hybrid" P2P, where some functionality is still centralized. (In Napster, for example, indexing is centralized, and file exchange is distributed.) We model a file-sharing application, developing a probabilistic model to describe query behavior and expected query result sizes. We also develop an analytic model to describe system performance. Using experimental data collected from a running, publicly available hybrid P2P system, we validate both models. We then present several hybrid P2P system architectures and evaluate them using our model. We discuss the tradeoffs between the architectures and highlight the effects of key parameter values on system performance.

1 Introduction

In a *peer-to-peer* system (P2P), distributed computing nodes of equal roles or capabilities exchange information and services directly with each other. Various new systems in different application domains have been labeled as P2P: In Napster [6], Gnutella [2] and Freenet [1], users directly exchange music files. In instant messaging systems like ICQ [3], users exchange personal messages. In systems like Seti-at-home [9], computers exchange available computing cycles. In preservation systems like LOCKSS [5], sites exchange storage resources to archive document collections. Every week seems to bring new P2P startups and new application areas.

**Proceedings of the 27th VLDB Conference,
Roma, Italy, 2001**

All these companies and startups tout the big advantage of P2P: the resources of many users and computers can be brought together to yield large pools of information and significant computing power. Furthermore, because computers communicate directly with their peers, network bandwidth is better utilized. However, there are often inherent drawbacks to P2P solutions precisely because of their decentralized nature. For example, in Gnutella, users search for files by flooding the network with queries, and having each computer look for matches in its local disk. Clearly, this type of solution may have difficulty scaling to large numbers of sites or complex queries. In Napster, on the other hand, users cannot search for files globally; they are restricted to searching on a single server that has only indexed a fraction of the available files.

Our goal is to study the scalability and functionality of P2P architectures, in order to understand the tradeoffs. Since we cannot possibly study *all* P2P systems at once, in this our initial paper we focus on *data-sharing, hybrid P2P systems*. The goal of a data-sharing system is to support search and exchange files (e.g., MP3s) found on user disks. In a data-sharing *pure* P2P system, all nodes are equal and no functionality is centralized. Examples of file-sharing pure P2P systems are Gnutella and Freenet, where every node is a "servent" (both a client and a server), and can equally communicate with any other connected node.

However, the most widely used file-sharing systems, such as Napster and Pointera [8], do not fit this definition because some nodes have special functionality. For example, in Napster, a server node indexes files held by a set of users. (There can be multiple server nodes.) Users search for files at a server, and when they locate a file of interest, they download it directly from the peer computer that holds the file. We call these types of systems *hybrid* because elements of both pure P2P and client/server systems coexist. Currently, hybrid file-sharing systems have better performance than pure systems because some tasks (like searching) can be done much more efficiently in a centralized manner.

Even though file-sharing hybrid P2P systems are hugely popular, there has been little scientific research done on them (see Section 2), and many questions remain unanswered. For instance, what is the best way to organize indexing servers? Should indexes be replicated at multiple servers? What types of queries do users typically submit in

such systems? How should the system deal with users that are disconnected often (dial-in phone lines)? How do different query patterns affect performance of systems from different application domains?

In the paper we attempt to answer some of these questions. In particular, the main contributions we make in this paper are:

- We present (Section 3) several architectures for hybrid P2P servers, some of which are in use in existing P2P systems, and others which are new (though based on well-known distributed computing techniques).
- We present a probabilistic model for user queries and result sizes. We validate the model with data collected from an actual hybrid P2P system. (Section 4.)
- We develop a model for evaluating the performance of P2P architectures. This model is validated via experiments using an open-source version of Napster [7]. (Section 5.)
- We provide (Section 6.1) a quantitative comparison of file-sharing hybrid P2P architectures, based on our query and performance models. Because both models are validated on a real music-sharing system, we begin experiments by focusing on systems in the music domain.
- We provide (Section 6.2) a comparison of strategies in domains other than music-sharing, showing how our models can be extended to a wide range of systems.

We note that P2P systems are complex, so the main challenge is in finding query and performance models that are simple enough to be tractable, yet faithful enough to capture the essential tradeoffs. While our models (described in Sections 4 and 5) contain many approximations, we believe they provide a good and reliable understanding of both the characteristics of P2P systems, and the tradeoffs between the architectures. Our extended report [24] discusses in further detail the steps we took to validate our models.

We also note that the only current, available experimental data on hybrid P2P systems are in the music-sharing domain. Hence, because it is important to have a validated model as a starting point for analysis, our experimental results in Section 6 begin with scenarios from Napster and other music-sharing systems. However, the models presented in this paper are designed to be general, and we show in Section 6.2 how they are applied to domains beyond music-sharing as well.

Finally, note that by studying hybrid P2P systems, we do not take any position regarding their *legality*. Some P2P systems like Napster are currently under legal attack by music providers (i.e., RIAA), because of copyright violations. Despite their questioned legality, current P2P systems have demonstrated their value to the community, and we believe that the legal issues can be resolved, e.g., through the adoption of secure business models involving royalty charging mechanisms. Thus, P2P systems will continue to be used, and should be carefully studied.

2 Related Work

Several papers discuss the design of a P2P system for a specific application without a detailed performance analysis. For example, reference [1] describes the Freenet project which was designed to provide anonymity to users and to adapt to user behavior. Reference [12] describes a system that uses the P2P model to address the problems of survivability and availability of digital information. Reference [15] describes a replicated file-system based on P2P file exchanges.

Adar and Huberman conducted a study in [10] on user query behavior in Gnutella, a "pure" P2P system. However, their focus was on the implications of "freeloading" on the robustness of the Gnutella community. There was no quantitative discussion on performance, and the study did not cover hybrid P2P systems.

Performance issues in hybrid file-sharing P2P systems have been compared to issues studied in information retrieval (IR) systems, since both systems provide a lookup service, and use inverted lists. Much work has been done on optimizing inverted list and overall IR system performance (e.g., [17, 25]). However, while the IR domain has many ideas applicable to P2P systems, there are differences between the two types of systems such that many optimization techniques cannot be directly applied. For example, IR and P2P systems have large differences in update frequency. Large IR systems with static collections may choose to rebuild their index at infrequent intervals, but P2P systems experience updates every second, and must keep the data fresh. Common practices such as compression of indexes (e.g., [18]) makes less sense in the dynamic P2P environment. While some work has been done in incremental updates for IR systems (e.g., [11, 23]), the techniques do not scale to the rate of change experienced by P2P systems.

Research in cooperative web caching (e.g., [13, 19]) has also led to architectures similar to the ones studied in this paper. However, performance issues are very different due to several key differences in functionality and context. First, most web caches search by key (i.e., URL), making hash-based cooperative web caches more effective than hash-based P2P systems, which allow multi-term keyword queries. Second, URL "queries" in web caches map to exactly one result, whereas queries in file-sharing systems return multiple results. Third, there is no concept of logins or downloads in web caching, which is a large part of P2P system performance. Fourth, bandwidth consumed by transferring a page from cache to client is an important consideration in cooperative web caching, but not in P2P systems. Finally, user behavior in the Web context is fairly well understood, whereas we will see that user behavior in P2P systems vary widely depending on system purpose.

Several of the server architectures we present in the next section either exist in actual P2P systems, or draw inspiration from other domains. In particular, the "chained" architecture is based on the OpenNap [7] server implementation, and resembles the "local inverted list" strategy [20] used in IR systems, where documents are partitioned across modes and indexed in subsets. The "full replication" architecture uses the NNTP [16] approach of fully replicating information across hosts, and a variation of the architecture is implemented by the Konspire P2P system [4].

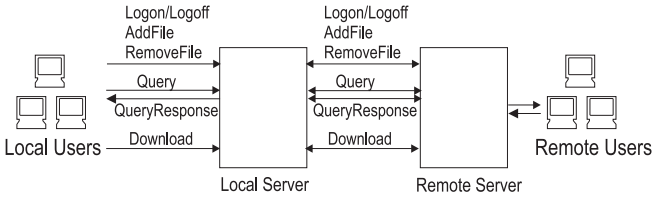

Figure 1: System Components and Messages

The "hash" architecture resembles the "global inverted list" strategy [20] used in IR systems, where entire inverted lists are partitioned lexicographically across nodes. Finally, the "unchained" architecture is derived from the current architecture in use by Napster [6].

3 Server Architecture

We begin by describing the basic concepts in a file-sharing, hybrid P2P system, based on the OpenNap [7] implementation of a file-sharing service.[1] We then describe each of the architectures in terms of these concepts. Figure 1 shows the components in the basic hybrid P2P system and how they interact.

General Concepts. There are three basic actions supported by the system: login, query and download.

On *login*, a client process running on a user's computer connects to a particular server, and uploads metadata describing the user's library. A *library* is the collection of files that a user is willing to share. The *metadata* might include file names, creation dates, and copyright information. The server maintains an index on the metadata of its client's files. For now, we assume the index takes the form of inverted lists [21]. Every file's metadata is considered a document, with the text of the file name, author name, and so on, being its content. The server also maintains a table of user *connection information*, describing active connections (e.g., client IP address, line speed).[2] By logging on, the user is now able to query its server, and is allowing other users to download her files.

A system may contain multiple servers, but a user is logged in and connected to only one server, its *local server*. To that user, all other servers are considered *remote servers*. From the perspective of one server, users logged on to it directly are its *local users*. Depending on the architecture, servers may index the library information of both local and remote users.

A *query* consists of a list of desired words. When a server receives a query, it searches for matches in its index. The server sets a maximum number of results to return for any query, and a query is said to be *satisfied* if the maximum number of results are returned. A query is processed by retrieving the inverted lists for all its words, and intersecting the lists to obtain the identifiers of the matching documents (user files). Clearly, other query and indexing schemes are possible (e.g., relational queries), and we discuss how to evaluate these schemes with our models in Section 6.2.

[1] We are not following the exact protocol used by OpenNap, but rather use a simplified set of actions that represent the bulk of the activity in the system.

[2] Often, clients use dial-in connections, so their IP address can vary from connection to connection.

Parameter Name	Default	Description
FilesPerUser	168	Average files per user library
FracChange	.1	Average percent of a user's library that is changed offline
WordsPerFile	10	Average words per file title
WordsPerQuery	2.4	Average keywords per query
CharPerWord	5	Average characters per word
QueryPerUserSec	.000833	Average number of queries per second per user
QueryLoginRatio	.45	Ratio of queries to logins per second per user
QueryDownloadRatio	.5	Ratio of queries to downloads per second per user
ActiveFrac	.05	Percent of the total user population that is active at any given time
λ_f	100	Inverse frequency skew (Section 4)
r	4	Query skew to occurrence skew ratio (Section 4)

Table 1: User-Defined Parameters

The user examines query results, and when she finds a file of interest, her client directly contacts the client holding the file, and *downloads* the file. After a successful download, or after a file is added through some other means, the client notifies the local server of the new addition to the library. The local server will add this information to its index. The local server is also notified when local files are deleted. Depending on architecture, remote servers may also be notified of the addition/deletion of local files.

Upon logoff, the local server updates the index to indicate that the user's files are no longer available. Again, remote servers may have to update their indices as well, depending on architecture. The options for handling logoffs are discussed in the next subsection.

Most hybrid file-sharing P2P systems offer other types of services to users other than just file sharing, such as chat rooms, hot lists, etc. These services are important in building community and keeping users attached to the main service. However, for our study we do not consider the effects of these activities on the system, as our experiments show that they do not significantly contribute to the workload.

As we discuss and study our hybrid P2P architectures, we will introduce a number of descriptive parameters. We show all parameters and their base values in Table 1, Table 2, and Table 3, even though many of the parameters and their values will be described in later sections. Base values for the parameters were determined by statistics collected from a live OpenNap system [24]. Parameters are divided into user-dependent parameters (Table 1) – those parameters that describe characteristics of user behavior, system parameters (Table 2) – those parameters that determine available system resources, and derived parameters (Table 3) – those parameters that are derived from other user and system parameters. The last derived parameter, *UsersPerServer*, is the value we want to maximize for each server. Our performance model calculates the maximum users per server supportable by the system, given all other

Parameter Name	Default	Description
LANBandwidth	80 Mb/s	Bandwidth of LAN connection in Mb/s
WANBandwidth	8 Mb/s	Bandwidth of WAN connection in Mb/s
CPUSpeed	800 MHz	Speed of processor in MHz
NumServers	5	Number of servers in the system.
MaxResults	100	Maximum number of results returned for a query
User-Server Network	WAN	The type of network connection between users and servers
Server-Server Network	LAN	The type of network connection between servers

Table 2: System-Defined Parameters

Parameter Name	Description
ExServ	Expected number of servers needed to satisfy a query
ExTotalResults	Expected number of results returned by all servers
ExLocalResults	Expected number of results returned by the local server
ExRemoteResults	Expected number of results returned by remote servers
UsersPerServer	Number of users logged on to each server

Table 3: Derived Parameters

parameters.

Batch and Incremental Logins. In current hybrid P2P systems, when a user logs on, metadata on her entire library is uploaded to the server and added to the index. Similarly, when she logs off, all of her library information is removed from the index. At any given time, only the libraries of connected, or active, users are in the index. We call this approach the *batch* policy for logging in. While this policy allows the index to remain small and thereby increases query efficiency, it also generates expensive update activity during login and logoff.

An alternative is an *incremental* policy where user files are kept in the index at all times. When a user logs on, only files that were added or removed since the last logoff are reported. If few user files change when a user is offline, then incremental logins save substantial effort during login and logoff. (The parameter *FracChange* tells us what fraction of files change when a user if offline.) However, keeping file metadata of *all* users requires filtering query results so that files belonging to inactive users are not returned. This requirement creates a performance penalty on queries. Also notice that in some architectures, the incremental policy requires that a user must always reconnect to the same server. This restriction may be a disadvantage in some applications.

Chained Architecture. In this architecture, the servers form a linear chain that is used in answering queries. When a user first logs on, only the local server adds library metadata to its index; remote servers are unaffected. When a user submits a query, the local server attempts to satisfy the query alone. However, if the local server cannot find the maximum number of results, it will forward the query to a remote server along the chain. The remote server will return any results it finds back to the first server, which will then forward the results to the user. The local server continues to send the query out to the remaining remote servers in the chain until the maximum number of results have been found, or until all servers have received and serviced the query. In this architecture, logins and downloads are fast and scalable because they affect only the local server of a user. However, queries are potentially expensive if many servers in the chain are needed to satisfy the query.

Full Replication Architecture. Forwarding queries to other servers can be expensive: each new server must process the query, results must be sent to the originating server, and the results must be merged. The full replication architecture avoids these costs by maintaining on each server a complete index of all user files, so all queries can be answered at a single server. Even with incremental logins, users can now connect to any server. The drawback, however, is that all logins must now be sent to every server, so that every server can maintain their copy of the index (and of the user connection information). Depending on the frequency ratio of logins to queries, and on the size of the index, this may or may not be a good tradeoff. If servers are connected by a broadcast network, then login propagation can be more efficient.

Hash Architecture. In this scheme, metadata words are hashed to different servers, so that a given server holds the complete inverted list for a subset of all words. We assume words are hashed in such a way that the workload at each server is roughly equal. When a user submits a query, we assume it is directed to a server that contains the list of at least one of the keywords. That server then asks the remaining servers involved for the rest of the inverted lists. When lists arrive, they are merged in the usual fashion to produce results. When a client logs on, metadata on its files (and connection information) must be sent to the servers containing lists for the words in the metadata. Each server then extracts and indexes the relevant words.

This scheme has some of the same benefits of full replication, because a limited number of servers are involved in each query, and remote results need not be sent between servers. Furthermore, only a limited number of servers must add file metadata on each login, so it is less expensive than full replication for logins. The main drawback of the hash scheme is the bandwidth necessary to send lists between servers. However, there are several ways to make this list exchange more efficient [22].

Unchained Architecture. The "unchained" architecture simply consists of a set of independent servers that do not communicate with each other. A user who logs on to one server can only see the files of other users at the same local server. This architecture, currently used by Napster, has a clear disadvantage of not allowing users to see all other users in the system. However, it also has a clear advantage of scaling linearly with the number of servers in the system. Though we cannot fairly compare this architecture with the rest (it provides partial search functionality), we still study its characteristics as a "best case scenario" for performance.

4 Query Model

To compare P2P architectures, we need a way to estimate the number of query results, and the expected number of servers that will have to process a query. In this section we

describe a simple query model that can be used to estimate the desired values.

We assume a universe of queries $q_1, q_2, q_3, \ldots$. We define two probability density functions over this universe:

- g – the probability function that describes query popularity. That is, $g(i)$ is the probability that a submitted query happens to be query q_i.
- f – the probability density function that describes query "selection power". In particular, if we take a given file in a user's library, with probability $f(i)$ it will match query q_i.

For example, if $f(1) = 0.5$, and a library has 100 files, then we expect 50 files to match query q_1. Note that distribution g tells us what queries users like to submit, while f tells us what files users like to store (ones that are likely to match which queries).

Calculating ExServ for the Chained Architecture. Using these two definitions we now compute *ExServ*, the expected number of servers needed in a chained architecture (Section 3) to obtain the desired R = *MaxResults* results. Let $P(s)$ represent the probability that exactly s servers, out of an infinite pool, are needed to return R or more results.

The expected number of servers is then:

$$ExServ = \sum_{s=1}^{k} s \cdot P(s) + \sum_{s=k+1}^{\infty} k \cdot P(s) \qquad (1)$$

where k is the number of servers in the actual system. (The second summation represents the case where more than k servers from the infinite pool were needed to obtain R results. In that case, the maximum number of servers, k, will actually be used.)

In [24] we show that this expression can be rewritten as:

$$ExServ = k - \sum_{s=1}^{k-1} Q(s \cdot \textit{UsersPerServer} \cdot \textit{FilesPerUser}).$$

Here $Q(n)$ is the probability that R or more query matches can be found in a collection of n or fewer files. Note that $n = s \cdot$ *UsersPerServer* $\cdot$ *FilesPerUser* is the number of files found on s servers.

To compute $Q(n)$, we first compute $T(n, m)$, the probability of exactly m answers in a collection of exactly n files. For a given query q_i, the probability of m results can be restated as the probability of m successes in n Bernoulli trials, where the probability of a success is $f(i)$. Thus, $T(n, m)$ can be computed as:

$$T(n,m) = \sum_{i=0}^{\infty} g(i) \left(\binom{n}{m} (f(i))^m (1 - f(i))^{n-m} \right).$$

$Q(n)$ can now be computed as the probability that we do not get $0, 1, \ldots$ or $R-1$ results in exactly n files, or

$$Q(n) = 1 - \sum_{m=0}^{R-1} T(n, m).$$

Calculating Expected Results for Chained Architecture. Next we compute two other values required for our evaluations, *ExLocalResults* and *ExRemoteResults*, again for a chained architecture. *ExLocalResults* is the expected number of query results from a single server, while *ExRemoteResults* is the expected number of results that will be returned by a remote server. The former value is needed, for instance, to compute how much work a server will perform as it answers a query. The latter value is used, for example, to compute how much data a server receives from remote servers that assist in answering a query.

Both values can be obtained if we compute $M(n)$ to be the expected number of results returned from a collection of n files. Then, *ExLocalResults* is simply $M(y)$, where y = *UsersPerServer* $\cdot$ *FilesPerUser*. Similarly, the expected total number of results, *ExTotalResults*, is $M(k \cdot y)$. Then *ExRemoteResults* is *ExTotalResults* - *ExLocalResults* $= M(k \cdot y) - M(y)$.

In [24] we show that $M(n) =$

$$\sum_{i=0}^{\infty} g(i) \left(R - \sum_{m=0}^{R-1} \binom{n}{m} (f(i))^m (1-f(i))^{N-m} (R-m) \right).$$

Calculating Expected Values for Other Architectures. So far we have assumed a chained architecture. For full replication, *ExServ* is trivially 1, since every server contains a full replica of the index at all other servers. Because *ExServ* is 1, all results are local, and none are remote. Hence *ExRemoteResults* $= 0$, and *ExLocalResults* = *ExTotalResults* $= M(N)$, where M is defined above, and $N = k \cdot$ *UsersPerServer* $\cdot$ *FilesPerUser* is the number of files in the entire system.

For the unchained architecture, *ExServ* is trivially 1, since queries are not sent to remote servers. *ExRemoteResults* $= 0$ and *ExLocalResults* $= M(n)$, where $n =$*UsersPerServer* $\cdot$ *FilesPerUser* is the number of files at a *single* server.

Finally, for the hash architecture, again *ExRemoteResults* $= 0$ and *ExLocalResults* $= M(N)$, since all the results are found at the server that the client is connected to. *ExServ* is more difficult to calculate, however. The problem again takes the form of equation (1), but now the probability $P(s)$ that exactly s servers are needed to to return R or more results is equal to the probability that exactly s servers contain w inverted lists, where w is the number of words per query. Finding this probability is the same as finding the probability that w balls being assigned to exactly s bins, a well-known probability problem. We refer the reader to [24] for a complete solution.

Distributions for f and g. While we can use any f and g distributions in our model, we have found that exponential distributions are computationally easier to deal with and provide accurate enough results in the music domain. Thus, we assume for now that $g(i) = \frac{1}{\lambda_g} e^{-\frac{i}{\lambda_g}}$. Since this function monotonically decreases, this means that q_1 is the most popular query, while q_2 is the second most popular one, and so on. The parameter λ_g is the mean. If λ_g is small, popularity drops quickly as i increases; if λ_g is large, pop-

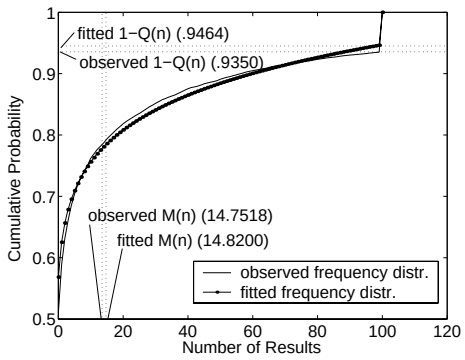

Figure 2: Observed and Fitted Frequency Distributions of Results in OpenNap

ularity is more evenly distributed. Similarly, we assume that $f(i) = \frac{1}{\lambda_f} e^{-\frac{i}{\lambda_f}}$. Note that this assumes that popularity and selection power are correlated. In other words, the most popular query q_1 has the largest selection power, the second most popular query q_2 has the second largest power and so on.

From this point on, we will express λ_g as $r \cdot \lambda_f$, where r is the ratio λ_g / λ_f. For a given query popularity distribution g, as r decreases toward 0, selection power f becomes more evenly distributed, and queries tend to retrieve the same number of results regardless of their popularity.

Section 6.2 describes how the query model behaves with different distributions for f and g.

Validation of Query Model. Using data obtained from a live OpenNap P2P system, we were able to experimentally validate our query model with the given assumptions. We were also able to give qualitative and quantitative evidence to support the assumption of correlated exponential distributions for f and g. Due to lack of space, we must refer the reader to [24] for a full description of our experiments and validation process. Here, we only describe briefly the results of the model validation.

Figure 2 shows the observed cumulative frequency distribution of the number of results returned for a sample of 16958 queries submitted to an OpenNap system, as well as the best-fit cumulative distribution derived from our query model. As seen from the two curves, the observed and derived distributions match very closely. In addition, we can also see that the observed values for $Q(1)$ and $M(n)$ (where n is the number of files at the server) match closely with the derived values of $Q(1)$ and $M(n)$. Hence we conclude that the query model can predict well the actual query behavior of hybrid P2P systems, given the assumptions and the correct parameter values. The fitted parameter values derived from this experiment are $\lambda_f = 400$ and $r = 10$.

5 Performance Model

In this section, we describe the performance model used to evaluate the architectures. We begin by defining our system environment and the resources our model will consider. We then present the basic formulas used to calculate the cost of actions in the system. Finally, we illustrate how we put the parameters and costs together to model the performance of the chained architecture. Because of space limitations, we are only able to give detailed examples of just a portion of what our model covers. For a complete description, please refer to [24].

Action	Formula for CPU instructions
Batch Login/off	$152817.6 \cdot$ *FilesPerUser* $+ 200000$
Batch Query	$(4000 + 3778) \cdot$ *ExTotalResults* $+ 100000 \cdot$ *ExServ* $+2000 \cdot$ *ExRemoteResults*
Batch Download	222348
Incr. Login/off	$77408.8 \cdot$ *FilesPerUser* $\cdot$ *FracChange* $+ 200000$
Incr. Query	$(4000 + 3778/$*ActiveFrac*$) \cdot$ *ExTotalResults* $+100000 \cdot$ *ExServ* $+ 2000 \cdot$ *ExRemoteResults*
Incr. Download	222348

Table 4: Formulas for cost of actions in CPU instructions

System Environment and Resources. In our model, we assume the system consists of *NumServers* identical server machines. The machines are connected to each other via a broadcast local area network (LAN), for example, if they all belong to the same organization, or through a point-to-point wide area network (WAN). Similarly users may be connected to servers via LAN or WAN. Our system can model any of the four combinations; let us assume for the examples in the remainder of the section that servers are connected to each other via LAN, and to users via WAN.

We calculate the cost of actions in terms of three system resources: CPU cycles, inter-server communication bandwidth, and server-user communication bandwidth. If users and servers are all connected via the same network (e.g., the same LAN), then the last two resources are combined into one. For the time being, we do not take I/O costs into consideration, but assume that memory is cheap and all indexes may be kept in memory. Memory is a relatively flexible resource, and depends largely on the decisions of the system administrator; hence, we did not want to impose a limit on it. Later in Section 6.1, we will discuss the memory requirements of each architecture, and point out the tradeoffs one must make if a limit on memory is imposed.

Table 2 lists the default system parameter values used in our performance analysis. The bandwidth values are based on commercial enterprise level standards, while the CPU represents current high-end commodity hardware. *NumServers* and *MaxResults* were taken from the OpenNap chain settings.

CPU Consumption. Table 4 lists the basic formulas for the cost of actions in CPU instructions, for the simple architecture. In this table, we see that the cost of a query is a function of the number of total and remote results returned, the cost of logging on is a linear function of the size of the library, and the cost of a download is constant. The coefficients for the formulas were first estimated by studying the actions of a typical implementation, and by roughly counting how many instructions each action would take. For example, in the formula for batch query cost, we estimated that the startup overhead at each server is due mostly to the cost of an in-memory transaction, which is listed in [14] as

Action	Formula
Batch Login	$42 + (75 +$ *WordsPerFile* $\cdot$ *CharPerWord*$)$ $\cdot$*FilesPerUser*
Incr. Login	$42 + (46 +$ *WordsPerFile* $\cdot$ *CharPerWord*$)$ $\cdot$*FilesPerUser* $\cdot$ *FracChange*
Query	*WordsPerQuery* $\cdot$ *CharPerWord* $+ 100$
QueryResponse	$90 +$ *WordsPerFile* $\cdot$ *CharPerWord*
Download	$81 +$ *WordsPerFile* $\cdot$ *CharPerWord*

Table 5: Formulas for cost of actions in bytes

roughly 100000 instructions. When it was hard to estimate the costs of actions, we ran measurement tests using simple emulation code. Finally, we experimentally validated the overall formulas against our running OpenNap server. The performance predicted by our formulas matched relatively well the actual values, at least for the chained architecture (batch policy) that OpenNap implements. For additional details on the validation process, see [24].

To calculate the cost of actions in incremental mode, we use the same formulas derived for batch mode, but incorporate the changes in the underlying action. Servicing a query in incremental mode has the same formula and coefficients as in batch mode, except that only *ActiveFrac* of the elements in the inverted lists pertain to files that are owned by currently active users, and all other elements cannot be used to answer the query. As a result, the cost of reading list elements per returned result needs to be divided by *ActiveFrac*.

The CPU cost of actions may also vary depending on architecture. In both the unchained and full replication architectures, the formula for batch and incremental query costs are exactly the same; however, *ExServ* = 1 and *ExRemoteResults* = 0 always. In the hash architecture, there is an additional cost of transferring every inverted list at a remote server to the local server. Hence, the coefficient describing the cost of list access per returned result is increased.

Network consumption. Table 5 lists formulas for the cost, in bytes, of the messages required for every action in the chained architecture. The coefficients come directly from the Napster network protocol, user-defined parameters, and a number of small estimates.

For example, when logging on, a client sends a Login message with the user's personal information, an AddFile message for some or all files in the user's library, and possibly a few RemoveFile messages if the system is operating in incremental mode. The Napster protocol defines these three messages as:

- Login message format: *(MsgSize MsgType <user> <passwd> <port> "<version>" <link-speed>)*.
- AddFile message format: *(MsgSize MsgType "<filename>" <md5> <size> <bitrate> <frequency> <time>)*.
- RemoveFile message format: *(MsgSize MsgType <filename>)*.

Using the user-defined parameter values in Table 1, and estimating 10 characters in a user name and in a password, 7 characters to describe the file size, and 4 characters to describe the bitrate, frequency, and time of an MP3 file, the sizes of the Login, AddFile and RemoteFile messages come to 42 bytes, 125 bytes, and 67 bytes, respectively. When a client logs on in *batch* mode, a single Login message is sent from client to server, as well as *FilesPerUser* AddFile messages.[3] Average bandwidth consumption for payload data is $42+168 \cdot 125 = 21042$ bytes. When a client logs on in incremental mode, a single Login message is sent from client to server, as well as approximately *FilesPerUser* $\cdot$ *FracChange* $\cdot 0.5$ AddFile messages, and the same number of RemoveFile messages. Average bandwidth consumption is therefore $42+168 \cdot 0.1 \cdot 0.5(125+67) = 1654.8$ bytes.

Network usage between servers varies depending on the architecture. For example, for login, the unchained architecture has no inter-server communication, so the required bandwidth is 0. Servers in the chained architecture send queries between servers, but not logins, so again, the bandwidth is 0. With the full replication architecture, all servers must see every Login, AddFile and RemoveFile message. If servers are connected on a LAN, then the data messages may be broadcast once. If the servers are connected on a WAN, however, then the local server must send the messages *NumServers*-1 times. Similarly, in the hash architecture, if servers are connected via LAN, then the messages may be broadcast once. If the servers are connected via WAN, however, then the AddFile messages should only be sent to servers storing lists for words contained in the name of the file. Calculating the expected number of servers to receive an AddFile message is analogous to calculating *ExServ* described in Section 4.

Modeling Overall Performance of an Architecture. After deriving formulas to describe the cost of each action, we can put everything together to determine how many users are supportable by the system. Our end goal is to determine a maximum value for *UsersPerServer*. To do this, we first calculate the maximum number of users supportable by each separate resource, assuming infinite resources of the other two types. Then, to find the overall *UsersPerServer*, we take the minimum across the three resources. We cannot directly directly calculate the maximum *UsersPerServer* for a particular resource, however, because the amount of resource consumed is a complex function of *UsersPerServer*. Instead, we use the iterative secant method for zero-finding, where we guess two values for the parameter, calculate consumed resources for each of these guesses, and interpolate a new value for *UsersPerServer* where zero resource is unused, which is maximum capacity. Please see [24] for an example.

6 Experiments

In this section, we present the results of our performance studies. We first study music sharing systems today (Section 6.1), and then focus on the performance of systems in other domains, which may have different query characteristics from music-sharing applications (Section 6.2). For

[3] Yes, a separate message is sent to the server for each file in the user's library. This is inefficient, but it is the way the Napster protocol operates.

each experiment we will highlight the important conclusions and give a condensed, high-level explanation for the results. For a more detailed discussion of results, and to view the results of the full range of experiments we conducted, please refer to [24].

Throughout the performance comparisons, the metric of performance is the maximum number of users each server can support. Hence, we concern ourselves with throughput, and not response time. Unless otherwise specified, default values for parameters (see Tables 1, 2, and 3) were used during evaluation. Note that several of these parameters could be better represented as a distribution, which we replace with the mean. Since we wish to see the relative performance of the systems in terms of average throughput, rather than precise performance numbers, such a simplification is tolerable. For brevity, we will refer to the chained architecture as CHN, the full replication architecture as FR, the hash architecture has HASH, and the unchained architecture as UNCH. Because each architecture can be implemented using one of two login policies, we refer to the combination of a particular architecture and policy as a "strategy". For example, "batch CHN" is the strategy where the chained architecture is implemented using the batch login policy. There are a total of 8 strategies.

6.1 Music-Sharing Today

We begin by evaluating performance of systems with behavior similar to that of music-sharing systems today (e.g., Napster, OpenNap), described by default user and system parameter values listed in Tables 1 and 2. Figure 3 shows the overall performance of the strategies over various values of *QueryLoginRatio*. For example, at *QueryLoginRatio* $= 1$, incremental FR can support 54203 users per server, whereas batch FR can only support 7281 users per server. The dashed line represents the experimentally derived value of *QueryLoginRatio* for OpenNap. Figure 4 shows the expected number of results per query for each of the strategies shown in Figure 3, assuming that *MaxResults* = 100. As *QueryLoginRatio* increases, the number of logins per second decreases, meaning more users can be supported by the system, thereby making more files available to be searched. As a result, when *QueryLoginRatio* increases, the expected number of results increases as well.

From this experiment, we can make several important conclusions:

- Incremental strategies outperform their batch counterparts. In particular, incremental CHN and UNCH have the best performance and are recommended in this scenario. Currently, the incremental policy is not used by music-sharing systems, but it should clearly be considered (see end of section for memory considerations).
- Batch UNCH is the strategy that most closely describes Napster's architecture. As seen in the figures, surprisingly, adding chaining to the servers (switching from UNCH to CHN) does not affect performance by much, but returns significantly more results. Assuming the cost of maintaining a LAN between the servers is acceptable, batch CHN is clearly recommended over batch UNCH.
- Most policies are very sensitive to *QueryLoginRatio* near our measured value of 0.45. Small changes in *QueryLoginRatio* can significantly increase or reduce the maximum number of users supported by the system, thereby making capacity planning difficult. This sensitivity is especially important to consider if large increases in *QueryLoginRatio* are expected in the future when user network connections become more stable (see [24]).

Memory Requirements Thus far, we have evaluated the strategies assuming that there was enough memory to hold whatever indexes a server needed. We will now take a closer look at the memory requirements of each strategy.

Figure 5 shows the memory requirement in bytes of the various strategies, as a function of the number of users. Please refer to [24] for a description on how memory usage is calculated. Here, we assume *ActiveFrac* = .1, to keep all architectures within roughly the same scale. Clearly, the batch strategies are far more scalable than the incremental strategies. For example, when there are 10000 users in the system, batch CHN requires .35 GB of memory, while incremental CHN requires 10 times that amount. Also, CHN requires the least amount of memory, while FR requires the most. However, it is important to note that memory requirement is a function of several parameters, most importantly *NumServers* and *ActiveFrac*. As *NumServers* decreases, FR requires proportionally less memory. On the flip side, as *NumServers* increases, FR also requires proportionally more memory. Likewise, incremental strategies require 1/*ActiveFrac* as much memory as batch. As connections become more stable and *ActiveFrac* increases, memory required by incremental strategies will decrease inverse proportionally, until it is much more comparable to batch memory requirements than it currently is. Furthermore, as memory becomes cheaper and 64-bit architectures becomes widespread, memory limitations will become much less of an issue than it is now.

Today, it is likely that system administrators will limit the memory available on each server. By imposing a limit, several new tradeoffs come into play. For example, suppose a 4GB memory limit is imposed on each server, shown by the dashed line in Figure 5. Now, consider a Napster scenario where r = 4, $\lambda_f = 100$, and *ActiveFrac* = .1. Say we determine that *QueryLoginRatio* = .75. Our model predicts that the maximum number of users supported by batch CHN is 26828, and by incremental CHN is 69708. The memory required by these two strategies is shown in Figure 5 by the large 'x' marks. While incremental CHN can support over twice as many users as batch CHN, it also requires very large amounts of memory – far beyond the 4GB limit. If we use the incremental CHN strategy with the 4GB limit, then our system can only support 12268 users per server, shown as a large triangle in Figure 5, which is fewer than the users supported by batch CHN. Hence, batch CHN is the preferred architecture for this scenario.

However, let us now suppose *QueryLoginRatio* is .25. Then, the maximum number of users supported by batch CHN is 9190, and by incremental CHN is 52088. The

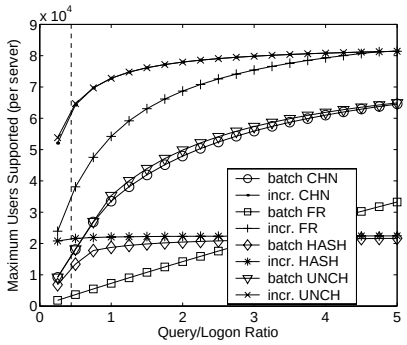

Figure 3: Overall Performance of Strategies vs. *QueryLoginRatio*

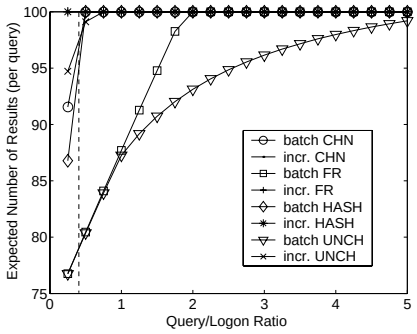

Figure 4: Expected Number of Results

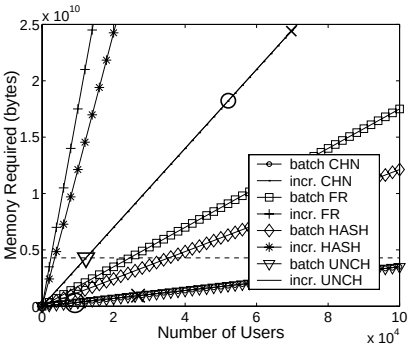

Figure 5: Memory Requirements

memory required by these two strategies is shown in Figure 5 by the large 'o' marks. Again, the amount of memory required by incremental CHN is far too large for our limit. However, looking at incremental CHN performance at the 4 GB limit, we find that the 12268 users supported by incremental CHN is greater than the 9190 users supported by batch CHN. Hence, incremental CHN is still the better choice, because within the limit, it still has better performance than batch CHN.

6.2 Beyond Music: Systems in Other Domains

In Section 4, we describe a query model where, given distributions for query frequency and query selection power, we can calculate the expected number of results for a query, and the expected number of servers needed to satisfy the query. The model itself is completely general in that it makes no assumptions about the type of distributions used for f and g; however, for the purposes of modeling music-sharing systems and validating the model, we made the temporary assumption that f and g are exponential distributions. Implied by this assumption is the additional assumption that f and g are *positively correlated* – that is, the more popular or frequent a query is, the greater the selection power it has (i.e. the larger the result set).

While we believe that our "music" model can be used in many other domains, one can construct examples where the f and g distributions have different properties. Differences in f and g distributions can be caused by different characteristics of application domains, and/or varying expressiveness of queries. For example, if the system supported complex, expressive relational queries, an obscure query such as *select * from Product where price* > 0 would return as many results as a common query such as *select * from Product*. In this case, there is very little, if any, correlation between the popularity of a query and the size of its result set. We say that such systems have *no correlation* distributions. Another example might be an "archive-driven" system, where users provide old, archived files online but keep their current, relevant information offline. This scenario might occur in the business domain where companies do not want their current data available to the public or their competitors, but might be willing to provide outdated information for research purposes. Because archives hold historical data, popular queries which refer to recent times return less data, whereas rare queries which refer to the past will return more. In this case, there is a *negative correlation* between query popularity and selection power.

In [24], we show through detailed analysis how the behavior of systems with negative and no correlation distributions can in fact be approximated very closely by systems with positive correlation distributions. Although the input to the query model, the f and g distributions, are very different, the output of the model, values for *ExServ* and *ExResults*, are closely matched. In particular, *no correlation* distributions are closely approximated by positive correlation distributions where $r = 1$, and *negative correlation* distributions are closely approximated when r is very large. Thus, we can understand the behavior of systems with negative and no correlation distributions by studying the performance of systems with a wide range of positive correlation distributions.

To illustrate, Figure 6 shows CPU performance of each strategy as r is varied. For comparison, the r value derived for the OpenNap system is shown by a dotted line in the figure. In addition, in [24] we presented several "representative" curves for negative and no correlation distributions, and determined the best-fit positive correlation approximations for each. For comparison, the r values derived for the approximations of the negative and no correlation curves are also shown in Figure 6 by dotted lines. From this figure, we can make several observations about the performance of strategies in different query models:

- With negative correlation distributions, or positive correlation with high r, incremental strategies are recommended. CPU performance of incremental strategies is usually bound by expensive queries (see [24]), but with negative correlation or high r, the expected number of results is very low, thereby making queries inexpensive.
- With no correlation distributions, or positive correlation with low r, batch strategies outperform incremental (except for FR). This is because when r is low, the expected number of results is very high, thereby making queries expensive and incremental performance poor. Batch FR performs so poorly because of expensive logins.
- CHN and HASH show greater improvement than FR as r increases. FR, which has poor login performance but

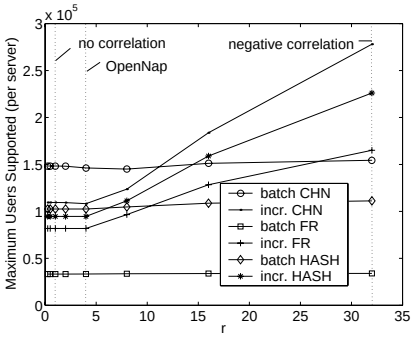

Figure 6: CPU Performance of Strategies vs. Query Model Parameter r

excellent query performance, is least affected by a decrease in *ExResults* caused by an increase in r. Hence, its performance improves the least as r increases.

7 Conclusion

In this paper, we studied the behavior and performance of hybrid P2P systems. We developed a probabilistic model to capture the query characteristics of these systems, and an analytical model to evaluate the performance of various architectures and policies. We validated both models using experimental data from actual hybrid P2P systems. Finally, we evaluated and compared the performance of each strategy. A summary of our findings (including results in our extended report [24]) for each architecture and policy are as follows:

- In our opinion, the **chained** architecture is the best strategy for today's music-sharing systems, and will continue to be unless network bandwidth increases significantly, or user interests become more diverse. This architecture has fast, scalable logins and requires the least amount of memory. However, query performance can be poor if many servers are involved in answering a single query.
- The **full replication** architecture has good potential in the future when network connections will be more stable and memory is cheaper. The architecture performs comparatively well in applications where user interests are diverse, and when result sets are relatively large.
- The **hash** architecture has very high bandwidth requirements. Hence, it is the best choice only if network bandwidth increases significantly in the future, or in systems where it is not necessary for servers to exchange large amounts of metadata (e.g., systems supporting only single-word queries such as search by key or ID).
- The **unchained** architecture is generally not recommended because it returns relatively few results per query and has only slightly better performance than other architectures. It is only appropriate when the number of results returned is not very important, or when no inter-server communication is available.
- The **incremental** policy is recommended in systems with negative correlation (e.g., historical or "archive-driven" systems), and performs best when sessions are short and network bandwidth is limited.

Acknowledgments. We would like to thank Daniel Paepcke for administrating relations with the OpenNap network used for our studies.

References

[1] Freenet Home Page. http://freenet.sourceforge.com/.
[2] Gnutella Development Home Page. http://gnutella.wego.com/.
[3] ICQ Home Page. http://www.icq.com/.
[4] Konspire Home Page. http://konspire.sourceforge.com/.
[5] LOCKSS Home Page. http://lockss.stanford.edu/.
[6] Napster Home Page. http://www.napster.com/.
[7] OpenNap Home Page. http://opennap.sourceforge.net/.
[8] Pointera Home Page. http://www.pointera.com/.
[9] SETI@home Home Page. http://setiathome.ssl.berkely.edu/.
[10] Eytan Adar and Bernardo A. Huberman. Free Riding on Gnutella. http://www.firstmonday.dk/-issues/issue5_10/adar/index.html, September 2000.
[11] Eric W. Brown, James P. Callan, and W. Bruce Croft. Fast incremental indexing for full-text information retrieval. In *Proc. of 20th Intl. Conf. on Very Large Databases*, pages 192–202, September 1994.
[12] Brian Cooper, Arturo Crespo, Hector Garcia-Molina. Implementing a reliable digital object archive. *4th European Conf. on Digital Libraries*, September 2000.
[13] Sandra Dykes. *Cooperative Web Caching: A Viability Study and Design Analysis*. PhD thesis, University of Texas at San Antonio, 2000.
[14] Jim Gray and Andreas Reuter. *Transaction Processing: Concepts and Techniques*. Morgan Kauffman Publishers, Inc., San Mateo, 1993.
[15] R. Guy, P. Reicher, D. Ratner, M. Gunter, W. Ma, and G. Popek. Rumor: Mobile data access through optimistic peer-to-peer replication. In *ER'98 Workshop on Mobile Data Access*, 1998.
[16] Brian Kantor, Phil Lapsley. Network News Transfer Protocol. RFC 977, February 1986.
[17] A. Moffat and J. Zobel. Self-indexing inverted files for fast text retrieval. *ACM Transactions on Information Systems*, 14(4):349–379, October 1996.
[18] Anh NgocVo and Alistair Moffat. Compressed inverted files with reduced decoding overheads. In *Proc. of the 21st Intl. Conf. on Research and Development in Information Retrieval*, pages 290–297, August 1998.
[19] Michael Rabinovich, Jeff Chase, and Syan Gadde. Not all hits are created equal: Cooperative proxy caching over a wide area network. In *Proc. of the 3rd Intl. WWW Caching Workshop*, June 1998.
[20] B. Ribeiro-Neto and R. Barbosa. Query performance for tightly coupled distributed digital libraries. *3rd ACM Conf. on Digital Libraries*, pages 182–190, June 1998.
[21] G. Salton. *Information Retrieval: Data Structures and Algorithms*. Addison-Wesley, Massachussetts, 1989.
[22] Anthony Tomasic. *Distributed Queries and Incremental Updates in Information Retrieval Systems*. PhD thesis, Princeton University, 1994.
[23] Anthony Tomasic, Hector Garcia-Molina, and Kurt Shoens. Incremental update of inverted list for text document retrieval. In *Proc. of the 1994 ACM SIGMOD Intl. Conf. on Management of Data*, pages 289–300, May 1994.
[24] Beverly Yang and Hector Garcia-Molina. Comparing Hybrid Peer-to-Peer Systems. Technical report, Stanford University, February 2001. Available at http://dbpubs.stanford.edu/pub/2000-35.
[25] J. Zobel, A. Moffat, and R. Sacks-Davis. An efficient indexing technique for full-text database systems. In *18th Intl. VLDB Conf.*, pages 352–362, August 1992.