An Analytical Performance Model of Robotic Storage Libraries

Theodore Johnson
Dept. of CISE, University of Florida
ted@cis.ufl.edu

June 25, 1996

Abstract

Large scale scientific projects generate and use huge amounts of data. For example, the NASA EOSDIS project is expected to archive one petabyte per year of raw satellite data. This data is made automatically available for processing into higher level data products and for dissemination to the scientific community. Such large volumes of data can only be stored in *robotic storage libraries* (RSLs) for near-line access. A characteristic of RSLs is the use of a robot arm that transfers media between a storage rack and the read/write drives, thus multiplying the capacity of the system.

The performance of the RSLs can be a critical limiting factor of the performance of the archive system. However, the many interacting components of a RSL make a performance analysis difficult. In addition, different RSL components can have widely varying performance characteristics. This paper describes our work to develop performance models of a RSL. We first develop a performance model of a RSL in isolation. Next, we show how the RSL model can be incorporated into a queuing network model. We use the models to make some example performance studies of archive systems.

The models described in this paper, developed for the NASA EOSIDS project, are implemented in C with a well-defined interface. The source code and accompanying documentation are available through WWW at:

`http://www.cis.ufl.edu/~ted/`

1 Introduction

Large scale scientific projects generate and use huge amounts of data. For example, the NASA EOSDIS project is expected to archive one petabyte per year of raw satellite data [28]. This data is made automatically available for processing into higher level data products and for dissemination to the scientific community (see, for example, the reports in [17]). Automatic management of such large data sets requires the use of tertiary storage, typically implemented using *robotic storage libraries* (RSLs). In addition to EOSDIS and related projects, many organizations and scientific disciplines make use of mass storage archives (for example high energy physics [30] and digital libraries [11]).

The database community has also become interested in the use of RSLs. [13, 6, 36, 35]. This interest is motivated in part by scientific database problems such as EOSDIS. Another motivation for integrating RSLs with on-line database systems is to facilitate data warehousing.

Tertiary storage is required when the managed data set becomes too large to store economically with conventional magnetic disk devices. The point at which tertiary storage becomes necessary is an economic

tradeoff. Currently, it seems that tertiary storage is needed to manage more than a terabyte of data. A RSL is much slower than magnetic disk storage, and data access latencies can run into minutes even on unloaded systems. However, RSL-resident data can accessed automatically. Hierarchical storage management systems, such as Unitree, Filestore, and Amass, provide the illusion that the RSL is an extension of the file system. Access to archived data incurs a short delay. The storage capacity of a data system can also be increased by using *off-line storage* – i.e. tape racks with human operators. Access latencies with off-line storage can be very large, ranging into hours or days, but the data storage capacity is limited only by the size of the warehouse that one can afford to rent. Since RSL provides data volumes and access latencies between those provided by on-line and off-line storage, it is often referred to as *near-line storage*. A cost analysis of on-line, near-line, and off-line archives can be found in [27].

A characteristic of RSLs is the use of removable media and a robot arm. The removable media (e.g. magnetic tape, optical disk, etc.) are normally located in a *storage rack*. To service a request for a file, the robot arm fetches the proper media from the storage rack and delivers it to a read/write drive. The media is accessed in the normal way to fetch the file. Finally, the media is returned to the storage rack. The capacity of RSL is the product of the capacity of the media and the size of the storage rack. Recent magnetic tapes have a data capacity on the order of 10 Gbytes, and storage racks sizes range from 10 to 1000 media (approximately). The time to fetch and mount the media which holds the requested file can be a large component of the access latency.

The performance of the RSLs can be a critical limiting factor of the performance of the archive system. Given the high data request rates expected for EOSDIS, attention to handling these requests efficiently is critical [28, 17]. However, the many interacting components of a RSL make a performance analysis difficult. In addition, different RSL components can have widely varying performance characteristics.

This paper describes our work to develop performance models of tertiary storage. We first develop a performance model of a RSL in isolation. Next, we show how the RSL model can be incorporated into a queuing network model. Finally, we model fork-join jobs to study the tradeoffs of using multiple devices. We use the models to make some example performance studies of archive systems.

The models described in this paper, developed for the NASA EOSIDS project, are implemented in C with a well-defined interface. The source code and accompanying documentation are available through WWW at:

`http://www.cis.ufl.edu/~ted/`

1.1 Previous Work

Considerable work has been done to develop performance models of mass storage. Rahm [32] presents a simulation study of a database system with a hierarchy of storage devices. Ramakrishnan and Emer [33] present a queuing model of a client/server file system. Drakopoules and Merges [14] present a closed queuing model of a client/server storage system with hierarchical storage. Kelly, Haynes, and Ernest [26] discuss a benchmark for network storage systems. Hauser, Rivera, and Thoma [20] discuss the performance of their networked WORM server.

Some work has been done to characterize the performance of mass storage devices. Waters [41] presents a validated model of seek times in hard disk drives. More recently, Ruemmler and Wilkes [34] present a detailed model of a modern disk drive, and discuss the difficulties inherent in I/O modeling. Christodoulakis and Ford [10] and Christodoulakis [9] present analytical performance models of optical drives. Chinnaswamy [8] presents performance models of a streaming tape drive to investigate the benefit of a cache.

Models of disk arrays resemble the models presented in this paper in several aspects. Burkhard, Claffy, and Schwarz [3] present a simulation study of a disk array scheme. Lee and Katz [29] and Yang, Hu and Yang [42] present analytical models of disk arrays. Chen et al. [7] and Thomasian [37] present surveys of research in RAID modeling.

Several authors have modeled a RSL. Butturini [4] presents the results of a simulation study of an optical disk jukebox system. Hevner [21] presents a model of an optical jukebox that is used for a database application. Howard [22] gives a performance model for data duplication from an archive. Finestead and Yeager [18] give performance measurements of a Unitree file server at the National Center for Hull and Ranade [23] present measurements of tape loading and unloading, and of data throughput, in a tape silo. Supercomputer Applications. Bedet et al. [2] discuss the results of a detailed simulation model of the Goddard DAAC. Pentakalos, Menasce, Halem, and Yesha [31] develop a queuing network model that incorporates a RSL. Daigle, Kuehl, and Langford [12] present a queuing model of an optical disk jukebox. Golubchik, Muntz and Watson [19] analyze tape striping on a RSL.

The analyses most closely related to the one in this paper are [31, 12, 19]. The analysis in [12] gives a detailed model of access times to data on an optical platter. However, only one drive is permitted and contention for the robotic arm is not modeled. In [31], the authors present a detailed model of a data center, incorporating RAID disk caches and user computation. However, the authors assume that contention for the drives in the RSL is negligible, and model the RSL as a delay server. Contention for the robotic arm due to batch arrivals is modeled in [19], but contention between jobs is not modeled.

The contribution of this work is to present a validated model of a RSL that accounts for batch arrivals, multiple drives, contention for the robotic arm, and realistic operation. We show how the model can be used to make a variety of data layout and device comparison studies. Finally, we show how to incorporate the RSL model into a queuing network model.

2 Model of a Robotic Storage Library

Our model of a RSL is illustrated in Figure 1. Previous studies of mass storage archive log files (see, for example, [24, 12]) indicate that requests to a mass storage device come in batches. This study has been corroborated by our studies of access to preliminary versions of the EOSDIS archives (the V0 archives) [1, 16]. As a result, our RSL model model uses batch arrivals.

A user requests that f files be loaded into on-line storage, and these files are distributed over m media in the RSL. The request is satisfied when every file has been loaded into on-line storage. So, a user *request* consists of m *jobs*, each of which must be completed before the request in finished. A RSL consists of n_d drives, each of which can read or write any of the media in the RSL,[1] a storage rack containing the removable media, and a robot arm for transferring the media between the drives and the storage rack. The model of a RSL is illustrated in Figure 1.

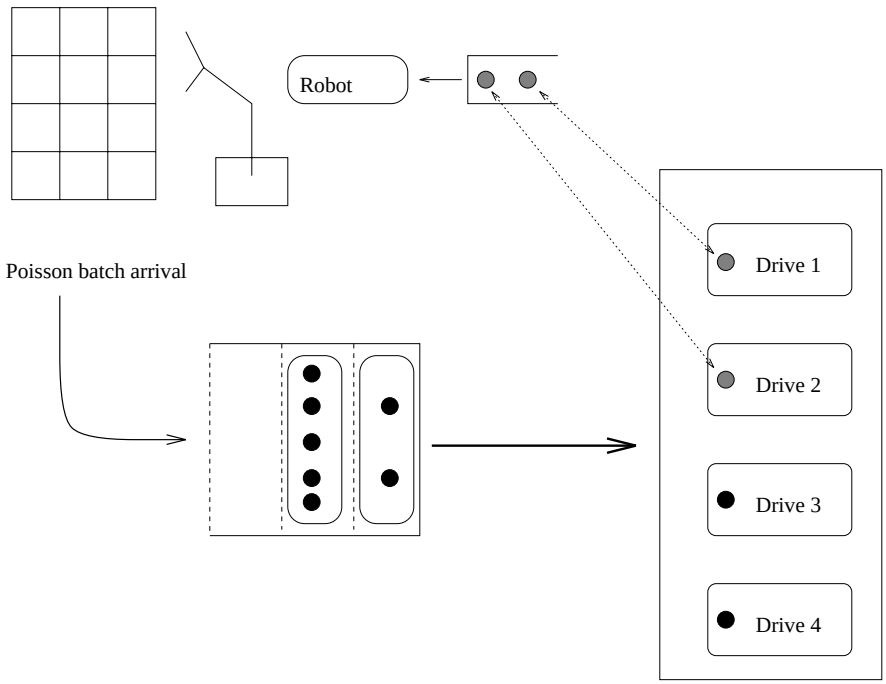

Figure 1: Architecture of a robotic storage library.

[1] In some installations, a subset of the drives are designated as read-only or write-only. We will address this complication in a later version of the model.

The steps taken by a drive in retrieving files from a media is illustrated in Figure 2. When a request arrives, its jobs are placed in the job queue. If there are jobs in the RSL queue and a drive is idle, the drive allocates one of the jobs for execution. First, the robot arm fetches the appropriate media from the storage area and loads it into the drive. If the robot arm is busy serving other drives, the drive must wait for service. After the media is brought to the drive, it must be mounted. For every file of interest on the media, the drive must seek to the start of the file, spend a settling time for precise positioning and opening communications channels, and then transfer the file to on-line storage. After all files have been transferred, the media is rewound and returned to the storage rack by the robot arm. However, the job is finished once all of the files have been transferred.

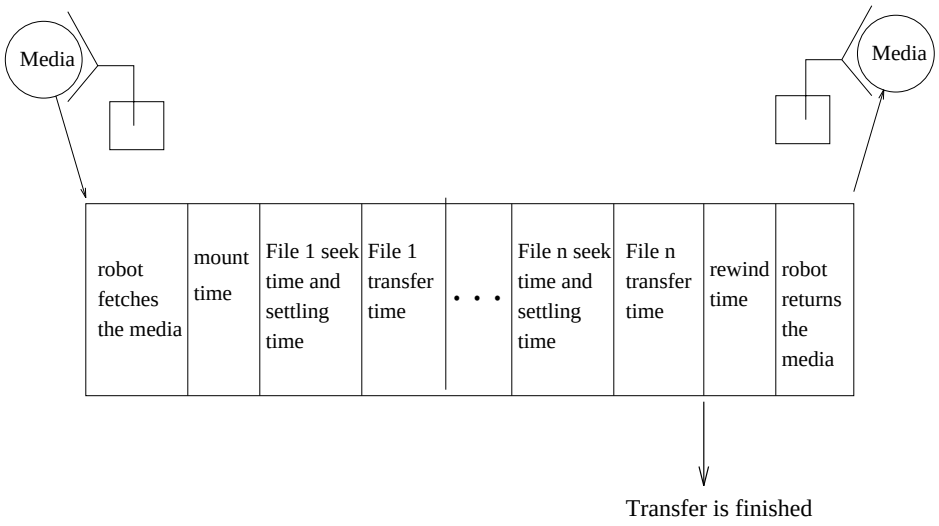

Figure 2: Steps in loading files from a media.

In the next section, we present an analytical performance model of a RSL system. In this preliminary model, we make the following assumptions:

1. Requests arrive in a Poisson process.
2. The distribution of the number of media per request and the number of files per request must be specified. In the model discussion, we assume that the number of media per request and the number of files per request have geometric distributions. These can be replaced by user-specified distributions (e.g. empirically determined), but at the cost of requiring the user to specify more parameters.
3. The RSL can contains one robot arm. The robot arm can access every media, and every drive.

4. Every drive can read and write every media.

5. Requests (i.e., jobs) are serviced first-come-first-serve[2].

6. The service for a request is completed when the last file of the batch has been read (written).

7. Network or communication channel contention is not significant[3].

8. Service times at the drives are independent.

3 Analytical model

A RSL presents many difficulties for performance modeling, including batch arrivals, multiple servers, derived parameters, and interacting components. The primary component of the RSL model, the $M^x/G/c$ queue, has been studied and solved in the literature [40]. Solving the actual $M^x/G/c$ queue is intractable, so the solution technique is to interpolate between the results for the $M^x/M/c$ queue and the $M^x/D/c$ queue using the coefficient of variation of the service time as the interpolation parameter. Because of the potential complexity of the batch arrival distributions, we do not use explicit (i.e., generating function) formulas. Instead, we numerically solve the recurrence equation that defines the state occupancy probabilities. If the occupancy probabilities of the first N states must be computed for an error bound of ϵ, then solving the $M^x/M/c$ queue requires $O(N^2)$ time and solving the $M^x/D/c$ queue requires $O(N^3)$ time.

Fortunately, we can take advantage of the nature of the problem to speed up the solution times. The state occupancy distributions eventually converge to a geometric distribution (i.e., $p_N = t_0 p_{N-1}$). Therefore the recurrence equations only need to be solved up the first N_0 states, and the remainder can be computed using the t_0 ratio (or perhaps the performance metrics can be computed directly). N_0 depends primarily on the distribution of the size of the batch arrival. Fortunately, the batch arrival distribution will have a short tail – one cannot request that more media than exist in the storage rack be mounted, and usually only a few media are required to satisfy a request. By using these tricks, we implemented batch queue solvers that are fast enough to be incorporated into a higher level model which calls them many times.

We now define some variables. The following are the inputs to the model.

- λ : request arrival rate.
- f_r : Average number of files per request.

[2] A simple optimization is to load files for all requests once a media has been mounted. We assume this situation has a negligible impact on performance in this model

[3] Potential model users indicated that communication channel contention is not a problem for their systems. Communication contention can be incorporated into the seek times or mount times using standard techniques [25]

- m_r : Average number of media per request.
- $b(.)$: Distribution of the number of media per request.
- n_d : number of drives.
- E_{tr} : Average robot fetch time.
- V_{tr} : Variance of the robot fetch time.
- T_{mt} : Media mount time.
- T_{fs} : Time for a full seek across the media.
- T_{stl} : Settling time after a seek.
- X_b : Transfer rate.
- E_{sz} : Average file size.
- V_{sz} : Variance in the file size.
- E_{rwd} : Expected time to re-wind the media.

The performance measures are:

- E_{rob} : Average delay for the robot fetch.
- ρ_{rob} : Robot arm utilization.
- ρ_d : Effective drive utilization utilization.
- $p_d(.)$: Queue length distribution for the drives.
- $R_{request}$: ATL request response time.

We model the RSL as an $M^x/G/c$ queue – that is, a queue with Poisson batch arrivals, general service time distribution, and c servers. The parameters of a $M^x/G/c$ queue are:

- λ : Arrival rate
- W_d : Mean service time
- cv_d : Co-efficient of variation of service time
- $b(.)$: Batch size distribution

- n_d : Number of servers

All but the service time distribution are input parameters, so our analysis is focused on how to compute E_d and $cv_d = \sqrt{V_d}/E_d$. To compute queue length distributions and expected waiting times properly, we need to compute the time that a drive is unable to serve other jobs per media that it serves. This period includes the time to fetch the media, mount it, seek to each file, transfer each file, rewind and eject the media, and return it to the storage rack. We will incorporate the time to return the media as part of the media fetch time, so we have:

$$\text{drive service} = (\text{robot fetch}) + (\text{mount time}) + (\text{seek time}) + (\text{transfer time}) + (\text{rewind time})$$

Analyzing the RSL reduces to analyzing each component of the drive service time. Some of these components are easy to estimate, others difficult. We will work from easiest to most difficult.

Mount time The mount time T_{mt} is the time period after the robot arm has placed the media in the drive and the drive can accept commands to seek to the first file. Typical actions include precise positioning of the media, drive initialization, and reading a directory sector (to ensure that the proper tape has been loaded, etc.). These actions are nearly deterministic [23] and are characteristics of the drive used in the RSL. Therefore, the mount time is a parameter, and we assume that mounting makes a negligible contribution to the drive service time variance.

Rewind time The rewind time is time period after the last file has been transferred from the media and the media can be ejected to the robot arm (it is typical to rewind tapes after use to ensure longevity, hence the name). The rewind time is device dependent [15], so we require that the mean and variance of the rewind time be supplied as parameters.

Transfer time Given that a file of Sz bytes is to be transferred, a commonly accepted model of the transfer time is a simple linear model:

$$T_{xf} = a + Sz/X_b$$

where T_{xf} is the transfer time, X_b is the transfer rate and a is transfer initialization time [8]. In this model, we will absorb the transfer initialization time into the seek time, so $a = 0$. Ideally, one could supply the mean and the variance of the volume of data that is to be transferred from the media. However, this might be more information than is available. More typically, one can estimate the mean and variance of the

individual files that are transferred. Then, one can use the distribution of the number of files transferred per media to estimate the mean and variance of the transfer time.

Let $b_f(.)$ be the probability density function of the number of files transferred per media. If k files are transferred, then the mean volume of data transferred is kE_{sz} and the variance is kV_{sz}. Therefore, the second moment of the volume of data transferred is

$$M^2_{sz}(k) = kV_{sz} + k^2E^2_{sz}$$

The first and second moments of the volume of data transferred per media is

$$E_{vol} = \sum_{k=1}^{\infty} kE_{sz}b_f(k)$$
$$M^2_{vol} = \sum_{k=1}^{\infty} M^2_{sz}(k)b_f(k)$$

To obtain transfer times, we divide by the transfer rate.

$$E_{xf} = E_{vol}/X_b$$
$$V_{xf} = (M^2_{vol} - E^2_{vol})/X^2_b$$

If the number of files per request and the number of media per request have a geometric distribution, and the number of files per media is distributed iid, then the number of files per media has a geometric distribution also[4]. If an average of f_m files are transferred from each media, then

$$E_{xf} = f_mE_{sz}/X_b$$
$$V_{xf} = f_m(V_{sz} + E_{sz}(f_m - 1))/X^2_b$$

Seek time Unlike the transfer times, there is no commonly accepted model for seek times (magnetic disk seek times are difficult enough [34]). The problem is compounded by the random seek distances, disk vs. single-pass vs. serpentine tape, and fast-seek modes. One simple model is to assume that files are randomly distributed across a single-pass tape. Suppose that f files are to be transferred. Then [15]:

$$E_{seek}(f) = \frac{f}{f+1}T_{fs} + fT_{stl}$$
$$M^2_{seek}(f) = fT_{fs}/(f+2) + 2f^2T_{stl}T_{fs}/(f+1) + f^2T_{fs}$$

Our solver assumes that user supplies a function which takes the number of files f as parameter and returns the mean and the second moment of the seek time. The unconditional seek time and variance is

[4] **Note to referee**: Here and at other places I suppress some calculations and details in order to save space

computed as:

$$\begin{aligned} E_{seek} = &\quad \sum_{f=1}^{\infty} E_{seek}(f) b_f(f) \\ M^2_{seek} = &\quad \sum_{f=1}^{\infty} M^2_{seek}(f) b_f(f) \\ V_{seek} = &\quad M^2_{seek} - E^2_{seek} \end{aligned}$$

Robot arm service time Computing E_{rob} is more complex, because the fetch time includes contention for a robot arm. If the ATL is empty and a batch request arrives, then a batch request is made to the robot arm. The queuing model that is appropriate for modeling the robot arms is the $M^X/G/1$ queue. Then, E_{rob} is the response time of the average job.

There are some complications. Suppose that n_{idle} of the drives are idle when a batch of size B arrives. If $B > n_{idle}$, the n_{idle} idle drives will request will request service immediately. The remaining $B - n_{idle}$ jobs will request service from the robot arms as the drives finish serving the preceding jobs. Each of these $B - n_{idle}$ requests will be preceded by a matching request to remove the old media from the drive. Therefore, the $B - n_{idle}$ jobs have a doubled service time requirement. The n_{idle} requests also generate matching requests to unload the drives, but these requests are not correlated with any requests to load a drive.

Let $s(n_{idle})$ be the probability that the batch is of size n_{idle} or smaller, and let $\overline{b}(n_{idle})$ be the average size of a batch request, given that the batch is for n_{idle} or fewer jobs. That is,

$$\begin{aligned} s(n_{idle}) &= \sum_{i=1}^{n_{idle}} b(i) \\ \overline{b}(n_{idle}) &= \sum_{i=1}^{n_{idle}} i * b(i) \end{aligned}$$

The arrivals to the robot arm queue consists of the truncated job arrival stream (with rate λ), an arrival stream that corresponds to jobs that must wait for service from a drive (rate λ_{doub}), and an arrival stream that corresponds to drives returning media without an immediate fetch of new media (rate λ_{sing}). Then,

$$\begin{aligned} \lambda_{doub}(n_{idle}) = &\quad \lambda(m_r - \overline{b}(n_{idle})) \quad n_{idle} > 0 \\ \lambda_{sing}(n_{idle}) = &\quad \lambda \overline{b}(n_{idle}) \quad n_{idle} > 0 \\ \lambda_{rob}(n_{idle}) = &\quad \lambda + \lambda_{doub} + \lambda_{sing} \quad n_{idle} > 0 \end{aligned}$$

Let $b_{rob,n_{idle}}$ be the distribution of batch arrivals to the robot, given that n_{idle} drives are idle. Then,

$$b_{rob,n_{idle}}(1) = \quad (\lambda * b(1) + \lambda_{sing})/\lambda_{rob} \quad n_{idle} > 2$$

$$
\begin{aligned}
b_{rob,n_{idle}}(2) = &\quad (\lambda * b(2) + \lambda_{doub})/\lambda_{rob} && n_{idle} > 2 \\
b_{rob,n_{idle}}(i) = &\quad \lambda * b(i)/\lambda_{rob} && n_{idle} > 2 \text{ and } 2 < i < n_{idle} \\
b_{rob,n_{idle}}(n_{idle}) = &\quad \lambda * (b(n_{idle}) + (1 - s(n_{idle}))/\lambda_{rob}) && n_{idle} > 2
\end{aligned}
$$

$$
\begin{aligned}
b_{rob,2}(1) = &\quad (\lambda * b(1) + \lambda_{sing})/\lambda_{rob} \\
b_{rob,2}(2) = &\quad (\lambda * b(2) + (1 - s(2)) + \lambda_{doub})/\lambda_{rob}
\end{aligned}
$$

$$
\begin{aligned}
b_{rob,1}(1) = &\quad (\lambda + \lambda_{sing})/\lambda_{rob} \\
b_{rob,1}(2) = &\quad \lambda_{doub}/\lambda_{rob}
\end{aligned}
$$

If all drives are busy, then all jobs are serviced when a busy drive finishes services. Therefore:

$$
\begin{aligned}
\lambda_{rob}(0) = &\quad \lambda m_r \\
b_{rob,0}(2) = &\quad 1
\end{aligned}
$$

Given the drive queue length distribution p_d, let p_{busy} be the probability that all drives are busy (i.e., $p_{busy} = 1 - (p_d(0) + \cdots + p_d(n_d - 1))$. We can compute the parameters to the robot queue by derandomizing $b_{rob}(.)$ and and $\lambda_{rob}(.)$. For example,

$$
\lambda_{rob} = p_{busy}\lambda_{rob}(0) + \sum_{i=0}^{n_d - 1} p_d(i)\lambda_{rob}(n_d - i)
$$

Then, the parameters to the $M^x/G/1$ queue [40] representing the robot arm are:

Arrival rate	λ_{rob}
Service time	$E_{rob} = E_{tr}$
Coefficient of variation	$cv_{rob} = \sqrt{V_{tr}}/E_{rob}$
Batch size distribution	$b_{rob}(.)$

The queuing model provides for us the robot arm utilization ρ_{rob}, and the waiting time for robot arm service W_{rob}. The $M^x/G/1$ model returns "average" waiting times. That is, the increase in waiting time due to correlated arrivals are not fully accounted for. Recall that if a job is serviced by a drive that was busy immediately before starting service, then the job must wait for both the return of the old media and

the fetch of the new media. This correlation is captured in the waiting times by the batch arrival stream $\lambda_d oub(n_{idle})$. However, the reported service time accounts for only half of the increased increased robot service time for these jobs. Therefore, we compute the additional waiting time due to correlated arrivals at the robot arm to be:

$$E_{xtra}(n_{idle}) = E_{tr}\frac{\lambda_{doub}(n_{idle})}{\lambda+\lambda_{doub}(n_{idle})}$$

$$E_{xtra} = p_{busy}E_{xtra}(0) + \sum_{i=0}^{n_d-1} p_d(i)E_{xtra}(n_d - i)$$

We compute the robot arm response time to be:

$$E_{rob} = W_{rob} + E_{tr} + E_{xtra}$$

The $M^x/G/1$ queue solvers compute the state occupancy distributions, so we compute the variance of the response time V_{rob} by using this distribution.

The Drives The parameters to the queue representing the RSL are:

Arrival rate	λ
Service time	$E_d = E_{xf} + E_{seek} + T_{mt} + E_{rob} + E_{rwd}$
Coefficient of variation	$cvs_d = \frac{\sqrt{V_{xf}+V_{seek}+V_{rob}+V_{rwd}}}{E_d}$
Batch size distribution	$b(.)$
Number of servers	n_d

Since we are interested in the response time of the last job in the batch to finish (i.e., instead of the average job), we need to modify the response time computation. An efficient algorithm for computing the response time of the last job in the batch is given in [25].

The modified $M^x/G/c$ queue provides the batch response time R_{batch}, the drive utilization ρ_d, and $p_d(0), \ldots, p_d(n_d - 1)$, the probability $0, \ldots, n_d - 1$ servers are busy on a request arrival. The effective drive service time (and $p_d(.)$) depends on the robot response time, which in turn depends on $p_d(.)$. We use iteration to converge to the solution.

Finally, the job is finished when the last file has been transferred, there is no need to wait for the tape rewind. Therefore:

$$R_{request} = R_{batch} - E_{rwd}$$

3.1 Validation Study

We wrote a simple RSL simulator. The simulation accepts batch arrivals, requires that a robot unload and fetch a media before a drive can service a job, handles multiple drives, and accounts for media rewind times. The components of E_d and V_d, except for E_{rob} and V_{rob}, are pre-computed and sampled from an Erlang distribution.

We used the following values of the parameters in the validation study.

- $f_r = 20$.
- $b(.)$: Geometric distribution.
- $n_d = 4$.
- $E_{tr} = 10.0$ seconds.
- $V_{tr} = 10.0$.
- $T_{mt} = 10.0$ seconds.
- $X_b = 1.0$ Mbyte/sec.

We ran four sets of experiments to test the model. In the "large files" experiments, $E_{sz} = 50$ and $V_{sz} = 100$, and $T_{fs} = 50$, $T_{stl} = 1$. In the "small files" experiments, $E_{sz} = 5$ and $V_{sz} = 10$, and $T_{fs} = 20$, $T_{stl} = 2$. We tested the model with $m_r = 2$ and $m_r = 6$.

The results of the validation study are shown in Figures 3 through 6. In each case there is close agreement between the analytical and the simulation models. The most difficult case is when the files are small and distributed over an average of six media, because the robot fetch times constitute a large portion of the drive service times (about 22% of the total drive service time when the robotic arm waiting time is added). However, the analytical model is accurate enough to predict response times and drive utilizations. Charts comparing analytical and simulation drive utilizations are shown in Figures 7 and 8.

3.2 Performance Study

A performance model is useful for studying implementation alternatives. In this section, we present three sample performance studies based on the RSL model.

3.2.1 Clustering

Conventional wisdom holds that striping or declustering is necessary for obtaining high transfer rates from tertiary storage (by making use of parallel I/O). So, one should spread the files of a typical request around as many media as possible. Conventional wisdom also holds that swapping media is a source of great inefficiency in RSL access, so that one should try to ensure that the files of a typical request are placed on as few media as possible.

Neither argument is convincing, unless one has a predictive performance model. We ran the "small files" experiment with m_r ranging between 1.2 and 10. In Figure 9, we plot the response time of a request against the number of media per request for varying arrival rates (A similar chart can be found in an analysis of tape striping [19]). For low arrival rates, setting $m_r \approx n_d$ produces the best results. When $\lambda = .0001$, setting $m_r = 3$ results in a 22% lower response time than setting $m_r = 1.2$. For high arrival rates, setting $m_r = 2$ gives lower response times than other choices.

In Figure 10, we plot the drive utilization against m_r for varying arrival rates. Increasing m_r causes a linear increase in the drive utilization. As the arrival rate increases, it becomes less likely that all n_d drives are available to service the request. So, distributing the files over a smaller number of media reduces queuing delays. If the demand on the RSL is expected to be close to the device's capacity, then m_r should be small to increase the maximum throughput of the device.

The question of whether to cluster or decluster the files on the media can be summarized as:

- If the expected drive utilization is low [31] and fast response is important, then declustering can be a good strategy. However, the decrease in transfer times must be larger than the increase in queuing delays.
- If high throughput is important, clustering is a good strategy.

3.2.2 Device Selection

A performance model can be used to determine the most appropriate equipment for an application. Consider the following hypothetical problem. You need to choose between two devices of approximately equal cost. The workload is the "small files" workload, except that an average of 40 files are loaded per request.

The first device uses optical disks as the storage media. Seek times are negligible, but the storage capacity per disk is limited, you can afford to purchase only two drives, and the transfer rate is 1 Mbyte/second. The second device uses tapes. The capacity of the tapes is larger than the capacity of the optical disks, and you can afford to purchase four drives. However, the seek times are large.

Because of the difference in the capacities of the media, you need to load an average of 6 media per request if you use the optical disks, and 2 media per request if you use tape. Using these parameters, we can plot the average response time of the two RSLs under an increasing workload. This chart is shown in Figure 11.

For the parameters of this study, the optical disk system is better than the tape system if the workload is light, but worse if the workload is heavy. The fast seek times of the optical disks allow you to transfer the large number of files quickly. However, the tape based system has a higher data capacity. While the parameters of this study are engineered to give an interesting comparison, the point remains that device performance depends on the workload offered to the device.

3.2.3 Number of Drives

Many RSLs allow the user to install a ranging number of drives. Adding drives to a RSL can improve the performance of the device. But after a threshold, adding drives does not significantly improve performance.

We ran a sample study using the "small files" parameters and four media per request. In Figure 12, we plot the response time versus the arrival rate for a number of drives varying between 2 and 8. Adding a drive significantly improves performance up to four drives, but gives less benefit after four drives. In Figure 13, we plot the drive utilization against the arrival rate. Adding a drive to the RSL increases the capacity of the device. However, the robot arm will start to become a bottleneck. This can be seen in the non-linear increase in utilization of some of the curves, for example for $n_d = 8$.

3.3 Multiple Robotic Storage Libraries

A large scale data center is likely to have multiple RSLs. The devices might be acquired to handle data center growth, or multiple small devices might be less expensive than a single large device. In this section, we discuss an approximation to the request response time when the request is served by multiple RSLs.

If a request is services by two different RSLs, the request is finished when both devices have completed their part of the request. Since we assume that requests are independent, we need to analyze a fork-join queue with interfering traffic. Thomasian and Tantawi [39, 38] have found that a good approximation to the response time of the fork-join job is to take the maximum of the response times of each device.

If we have the response time distribution of the different storage devices, we can compute the expected value of the maximum response time by using order statistics. However, the exact response time distribution of an $M^x/G/c$ queue is difficult to compute. However, we can approximate the variance of the response time by using the state occupancy distributions computed by the $M^x/M/c$ and $M^x/D/c$ queue solvers, and

approximating the variance of the batch service time using a technique similar to the one used to compute the mean [25]. We model the response time distribution as an $\text{Erlang}_{1,k}$ distribution (the tail of the waiting time distribution of the $M^x/G/c$ queue approaches an exponential distribution [40]). The expected value of the maximum of two Erlang distributions is computed by integrating each component of the distribution separately, then summing. The coding is easy, and the routine runs in $O(k_1 k_2)$ time to compute the expected value of an Erlang_{1,k_1} and an Erlang_{1,k_2} distribution.

For an experiment, we applied the "large files" workload to two RSLs, with both receiving the same arrival rate. We considered a request that required files from six media. In Figure 14, we plot the response time of this request against the arrival rate, and varied the number of media that must be loaded from each device. The results show that when the load on the tertiary storage devices is low, it is better to divide the request evenly between the two devices. But, when the load is high it is better to use one device only. The reason for this result is that splitting the request between the two devices provides parallel I/O, but if the request load is high, then the variance in the response times becomes large. Thus, the decision to allocate files so that most requests use only one device or that most requests use both devices depends on the expected load placed on the devices.

3.4 Queuing Network Model

A mass storage data system consists of many components in addition to the RSLs. Typical hierarchical storage management systems use a database to track file to media location mappings, and maintain a sizeable staging and caching area. The computing centers that use tertiary storage often have large scale computing tasks. For example, EOSDIS archives must perform *product generation* to filter, correct, remap, and fuse satellite images (see the reports in [17], and also the discussion in [31]).

To capture the effects of RSLs in computing systems, we need to integrate the RSL model into a queuing network model. The typical approach for incorporating devices with unusual response time characteristics into a queuing network model is to use mean value analysis (MVA), and develop a MVA recurrence for the device in question [25]. However, it is difficult to develop such a recurrence even for multiple server devices. Therefore, we take the approach of integrating the open RSL queue into a MVA model.

Although the RSL model solver is fast (about 2 seconds of execution time), an exact MVA solver requires an iteration over every possible population vector. If the population is large and there are many job classes, solution times become intolerably large. We instead used an approximate MVA solver, making use of Schweitzer's approximation on queue lengths and Bard's approximation for the load dependent servers [25]. The approximate MVA solver built using these approximations compute the throughput at each iteration,

which we use as the arrival rate at the RSL (after scaling by the visit ratio).

Incorporating an open queuing model into a closed queuing network requires some care to ensure accurate solutions. Generally, accuracy is good for large populations but worse for small populations [5]. To improve the accuracy of the solution, we modified the $M^x/G/c$ solvers to incorporate a heuristic that accounts for the finite customer population. Suppose that there are i jobs in the RSL, n customers, and an average batch size of $\overline{b}$. Then, the arrival rate in state i is:

$$\lambda_i = \lambda \frac{n\overline{b} - i}{n\overline{b}}$$

To test the accuracy of the approximate MVA model, we simulated a computer system with a RSL and three other queuing devices. The requests to the RSL used the "large file" workload, and every customer submits a single request to the RSL per task execution. There are three queuing devices, with per-task service demands of 250, 400, and 350 units of work, respectively.

We plot the response time of the RSL against the number of customers in the system Figure 15 for a sleep time of 5000 and 9000. The model is accurate even for a small number of customers. However, the accuracy declines when the number of customers is large and the sleep time is small. This problem is occurring because one of the queuing devices is saturated, and the approximate MVA solver becomes inaccurate in these situations.

4 Conclusions

We have developed an analytical model of a robotic storage library and validated the model by comparison to simulations. The RSL consists of a storage rack for removable media, a set of drives that read and write the media, and a robotic arm that transfers the media between the storage rack and the drives.

The RSL model can be used for many useful studies. We provide examples of data layout and device selection studies. A RSL is used as a part of a larger computing system. We incorporated the RSL solver into an approximate MVA queuing network model, and validated the model by comparison to a simulation.

We have developed this model to support NASA's EOSDIS on-line archiving efforts. Future work will be directed towards refining the model and providing studies useful to archive sites. This work include:

- Refined models of seek times, request sizes, and files per media.
- Incorporating scheduling policies into the waiting queue (combining jobs that access the same media, for example).

- Heterogeneous drives. Some drives might be faster than others, some might accept a subset of the media, and some might be read-only.
- Incorporating multiple classes of RSL access into the queuing network model. Currently, multiple customer classes are supported, but all classes must have the same RSL access characteristics.
- Incorporating fork-join jobs into the queuing network model.
- Developing models of archive data centers, including models of caching, data ingest, and data maintenance activities.

5 Acknowledgements

We'd like to thank Ben Kobler and Chris Daly of NASA GSFC, and Bob Howard of Hughes for their comments, and Alex Thomasian for his advice regarding fork-join jobs.

References

[1] Jean-Jacques Bedet. Goddard daac v0 log files., 1996. Private communication.

[2] J.J. Bedet and et al. Simulation of a data archival and distribution system at GSFC. In *Goddard Conference on Mass Storage Systems and Technology*, pages 139–160, 1993.

[3] W.A. Burkhard, K.C. Claffy, and T.J.E. Schwarz. Performance of balanced array schemes. In *Mass Storage Systems Symposium*, pages 45–50, 1991.

[4] R.S. Butturini. Performance simulation of a high capacity optical disk system. In *Mass Storage Systems Symposium*, pages 147–153, 1988.

[5] J.P. Buzen and P.S. Goldberg. Guidelines for the use of infinite source queuing models in the analysis of computer systems performance. In *Proc. AFIPS 1974 National Computer Conf. Vol. 43*, pages 471–474, 1974.

[6] M.J. Carey, L.M. Haas, and M. Linvy. Tapes hold data too: Challences of tuples on tertiary store. In *Proc. ACM SIGMOD*, pages 413–418, 1993.

[7] M.P. Chen, E.K. Lee, G.A. Gibson, R.H. Katz, and D.A. Patterson. RAID: High-performance reliable secondary memory. *Computing Surveys*, 26(2):145–185, 1994.

[8] V. Chinnaswamy. Analysis of cache for streaming tape drive. In *Goddard Conference on Mass Storage Systems and Technology*, pages 299–310, 1992.

[9] S. Christodoulakis. Analysis of retrieval performance for records and objects using optical disk technology. *ACM Transactions on Database Systems*, 12(2):137–169, 1987.

[10] S. Christodoulakis and D.A. Ford. Performance analysis and fundamental performance tradeoffs for clv optical disks. In *ACM SIGMOD*, pages 286–294, 1988.

[11] R.A. Coyne and H. Hulen. Toward a digital library strategy for a national information infrastructure. In *Proc. 3rd NASA Godddard Conf. on Mass Storage Systems and Technologies*, pages 15–18, 1993.

[12] J.N. Daigle, R.B. Kuehl, and J.D. Langford. Queuing analysis of an optical disk jukebox based office system. *IEEE Trans. on Computers*, 39(6):819–828, 1990.

[13] J. Dozier, M. Stonebraker, and J. Frew. Sequoia 2000: A next-generation information system for the study of global change. In *Proc. 13th IEEE Mass Storage Systems Symposium*, pages 47–53, 1994.

[14] E. Drakopoulos and M.J. Merges. Perofrmance analysis of client-server storage systems. *IEEE Transactions on Computers*, 41(11):1442–1452, 1992.

[15] A. Drapeau and R.H. Katz. Striped tape arrays. In *Proc. 12th IEEE Mass Storage Systems Symposium*, pages 257–265, 1993.

[16] J. Dunham and B. North. Eosdis statistics collection and reporting system, 1996. Available by anonymous FTP at eos.nasa.gov:/EosDis/Daacs/Statistics.

[17] Esdis document catalog. http://spsosun.gsfc.nasa.gov/ESDIS_Docs.html.

[18] A. Finestead and N. Yeager. Performance of a distributed superscaler storage server. In *Goddard Conference on Mass Storage Systems and Technology*, pages 573–580, 1992.

[19] L. Golubchik, R.R. Muntz, and R.W. Watson. Analysis of striping techniques in robotic storage libraries. In *Proc. 14th IEEE Mass Storage Systems Symposium*, pages 225–238, 1995.

[20] S.E. Hauser, C. Rivera, and G.R. Thoma. Factors affecting the performance of a dos-based WORM file server. In *Mass Storage Systems Symposium*, pages 33–37, 1991.

[21] A.R. Hevner. Evaluation of optical disk systems for very large database applications. In *ACM SIGMETRICS conference*, pages 166–172, 1985.

[22] K. Howard. High speed data duplication/data distribution – an adjunct to the mass storage equation. In *Goddard Conference on Mass Storage Systems and Technology*, pages 123–133, 1992.

[23] G. Hull and S. Ranade. Performance measurements and operational characteristics of the Storage Tek ACS 4400 tape library with the Cray Y-MP EL. In *Goddard Conference on Mass Storage Systems and Technology*, pages 111–122, 1993.

[24] T. Johnson. Analysis of the request patterns to the nssdc on-line archive. In *Proc. 4th NASA Goddard Conf. on Mass Storage Systems and Technologies*, 1995.

[25] K. Kant. *Introduction to Computer System Performance Evaluation.* McGraw Hill, 1992.

[26] S.M. Kelly, R.A. Haynes, and M.J. Ernest. Benchmarking a network storage service. In *Mass Storage Systems Symposium*, pages 38–44, 1991.

[27] K.F. Kenk, J.L. Green, and L.A. Treinish. A cost model for NASA data archiving. Technical Report 90-08, National Space Science Data Center, NASA Goddard Space Flight Center, 1990.

[28] B. Kobler, J. Berbert, P. Caulk, and P.C. Hanrahan. Architecture and design of storage and data management for the NASA Earth Observing System Data and Information System (EOSDIS). In *Proc. 14th IEEE Mass Storage Systems Symposium*, pages 65–78, 1995.

[29] E.K. Lee and R.H. Katz. An analytic performance model of disk arrays. In *ACM SIGMETRICS*, pages 98–109, 1993.

[30] L. Lueking. Managing and serving a multi-terabyte data set at the Fermilab D0 experiment. In *Proc. 14th IEEE Mass Storage Systems Symposium*, pages 200–208, 1995.

[31] O.I. Pentakalos, D.A. Menasce, M. Halem, and Y. Yesha. Analytical performance modeling of hierarchical mass storage systems. Technical Report TR-CS-96-01, Dept. of Computer Science, University of Maryland, 1995. A short version appears in the 14th IEEE Mass Storage Symposium proceedings.

[32] E. Rahm. Performance evaluation of extended storage architectures for transaction processing. In *ACM SIGMOD*, pages 308–317, 1992.

[33] K.K. Ramakrishnan and J.S. Emer. Performance analysis of mass storage service alternatives for distributed systems. *IEEE Trans. on Software Engineering*, 15(2):120–133, 1989.

[34] C. Ruemmler and J. Wilkes. An introduction to disk drive modeling. *IEEE Computer*, 27:17–28, 1994.

[35] S. Sarawagi. Database systems for efficient access to tertiary memory. In *Proc. 14th IEEE Mass Storage Systems Symposium*, pages 120–126, 1995.

[36] Sequoia 2000 home page. http://s2k-ftp.cs.berkeley.edu:8000/.

[37] A. Thomasian. Surveyor's forum: High performance secondary memory. *Computing Surveys*, 27(2):292–295, 1995.

[38] A. Thomasian. Approximate analyses for fork/join synchronization in RAID 5. *Computer Systems: Science and Engineering*, 1996. To appear.

[39] A. Thomasian and A.N. Tantawi. Approximate solutions for m/g/! fork/join synchronization. In *Proc. 1994 Winter Simulation Conference*, 1994.

[40] H.C. Tijms. *Stochastic Models: An Algorithmic Approach*. Wiley, 1994.

[41] S.J. Waters. Estimating disk seeks. *The Computer Journal*, 18(1):12–17, 1974.

[42] T. Yang, S. Hu, and Q. Yang. A closed-form formula for queuing delays in disk arrays. In *Proc. Intl. Conf. on Parallel Processing*, pages II:189–192, 1994.

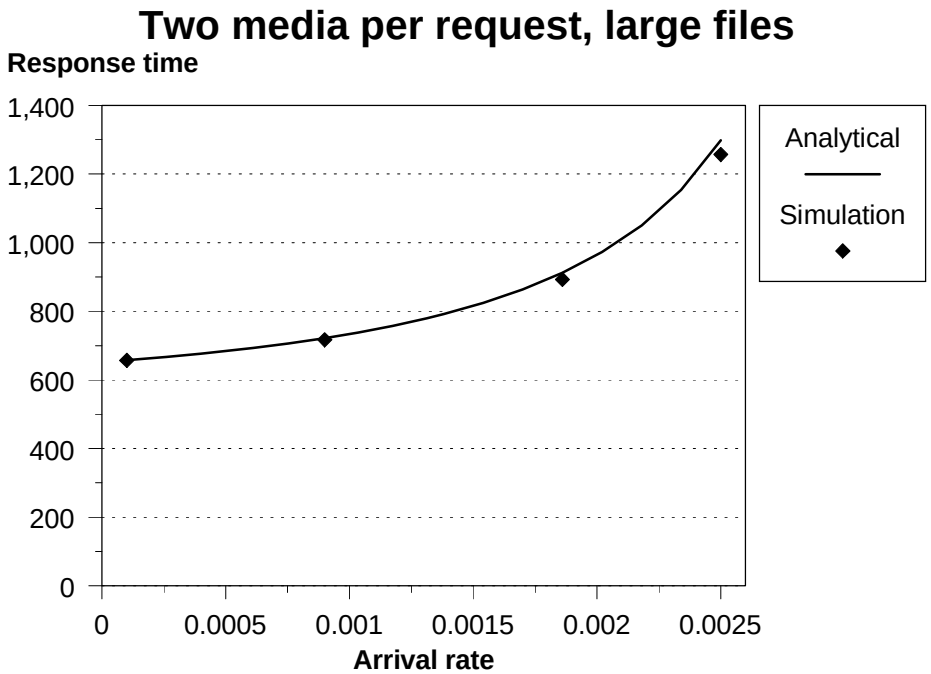

Figure 3: Validation study, response time vs. arrival rate.

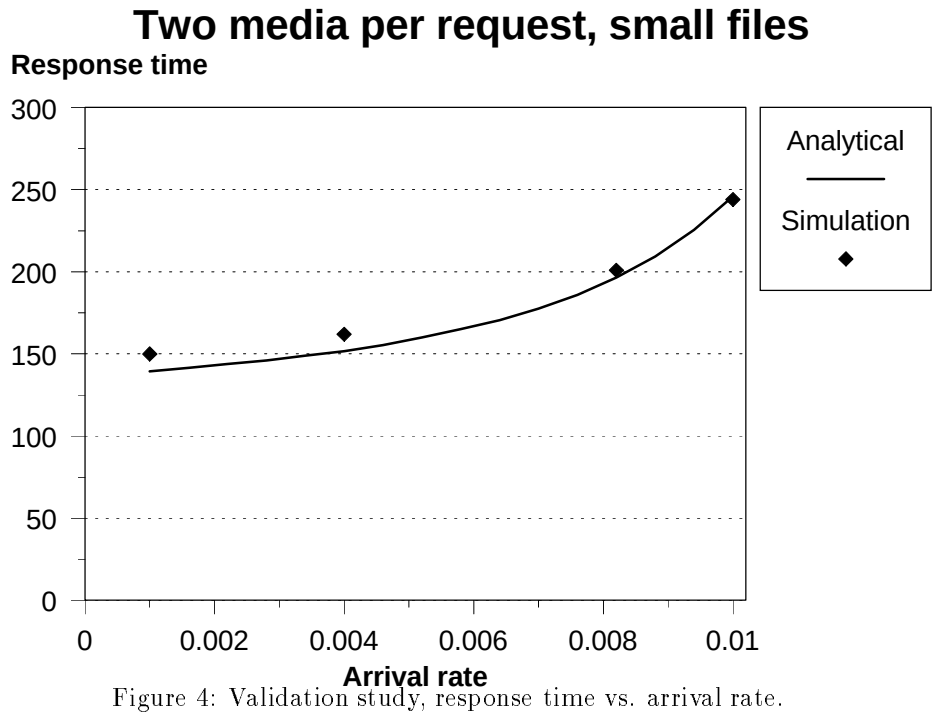

Figure 4: Validation study, response time vs. arrival rate.

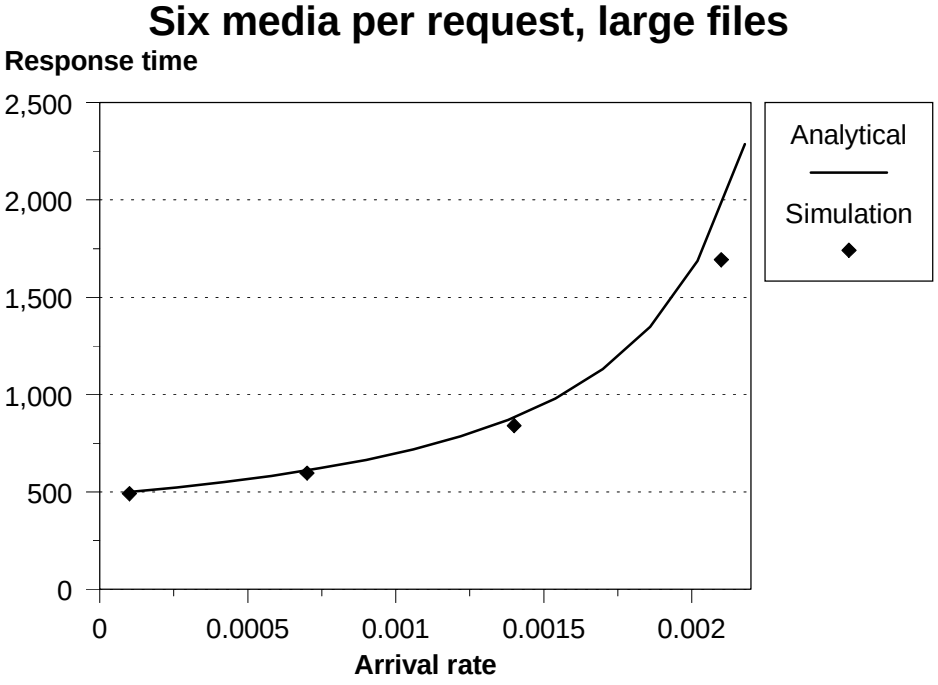

Figure 5: Validation study, response time vs. arrival rate.

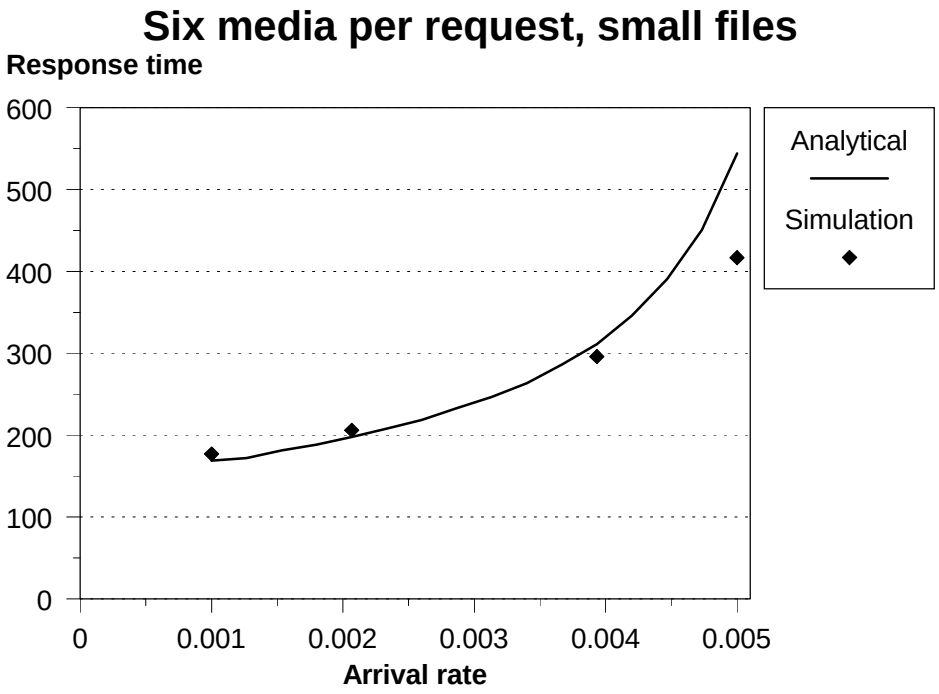

Figure 6: Validation study, response time vs. arrival rate.

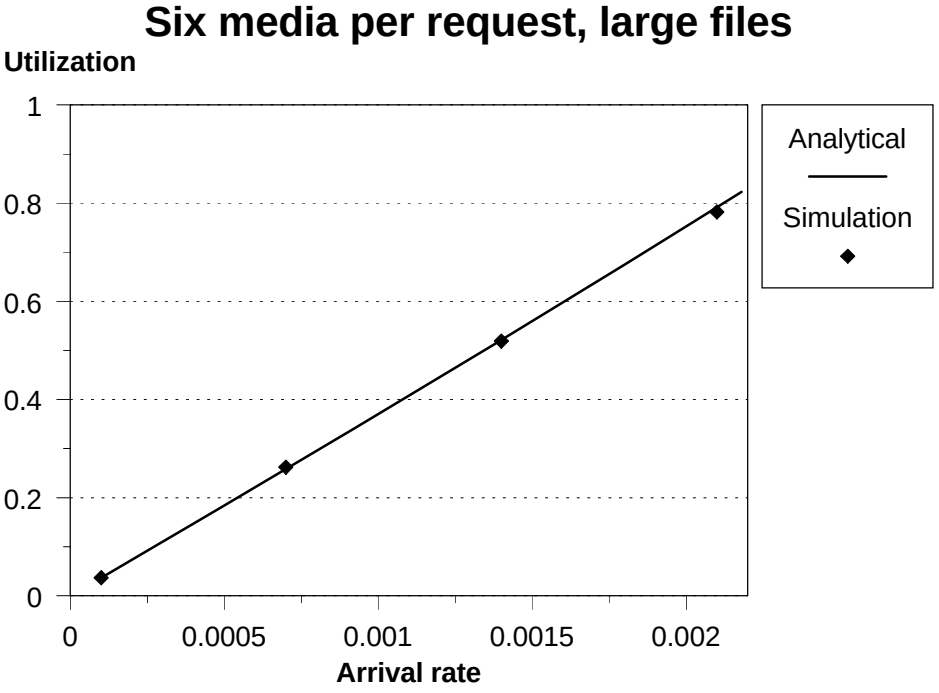

Figure 7: Validation study, utilization vs. arrival rate.

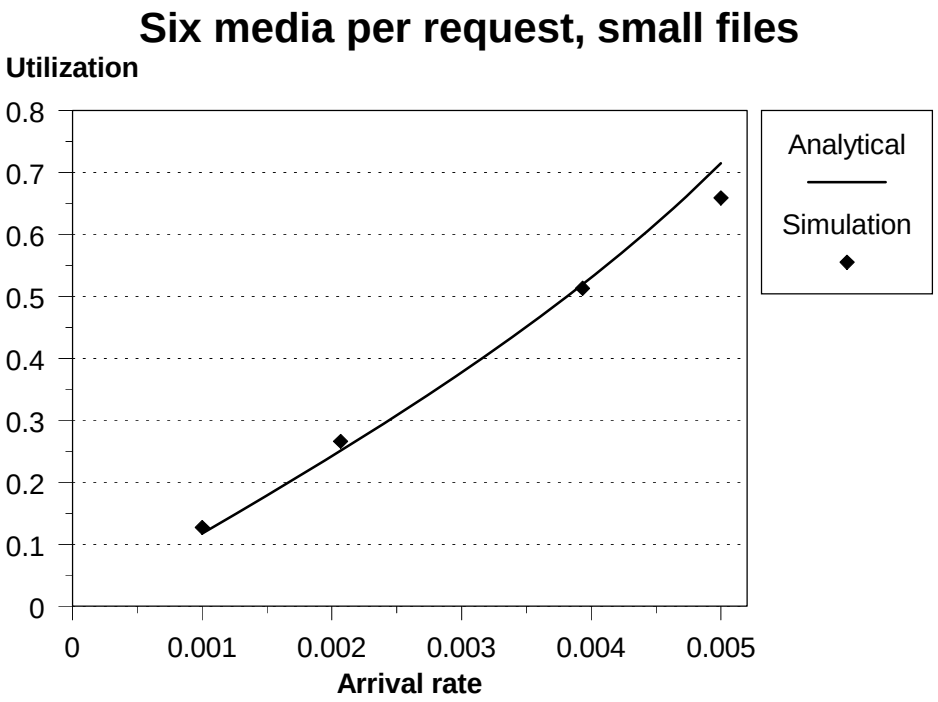

Figure 8: Validation study, utilization vs. arrival rate.

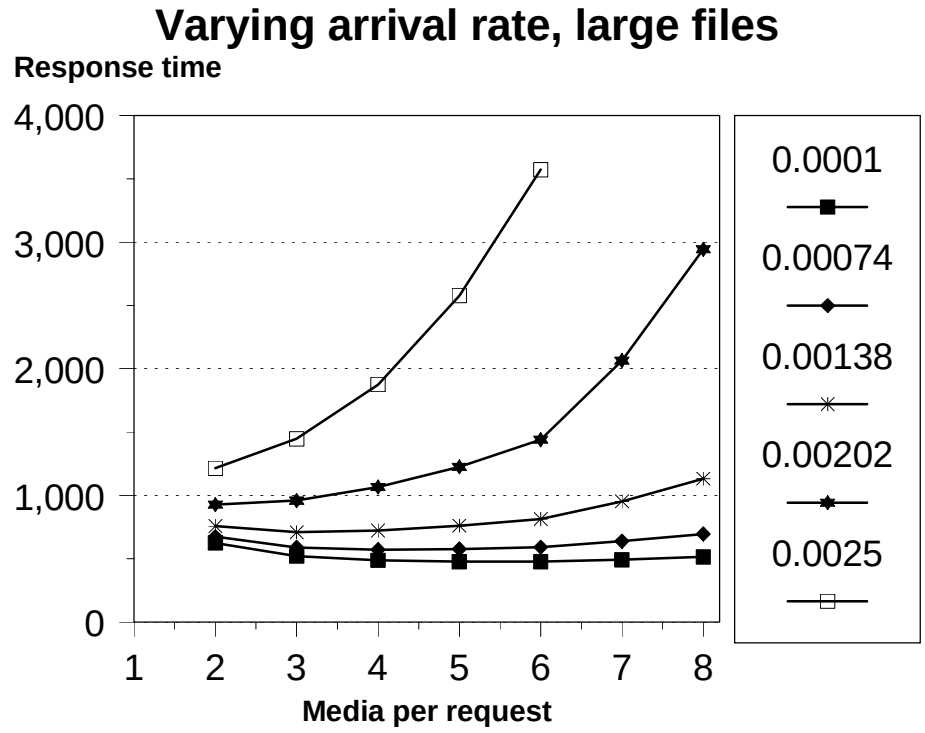

Figure 9: Response time vs. media per request.

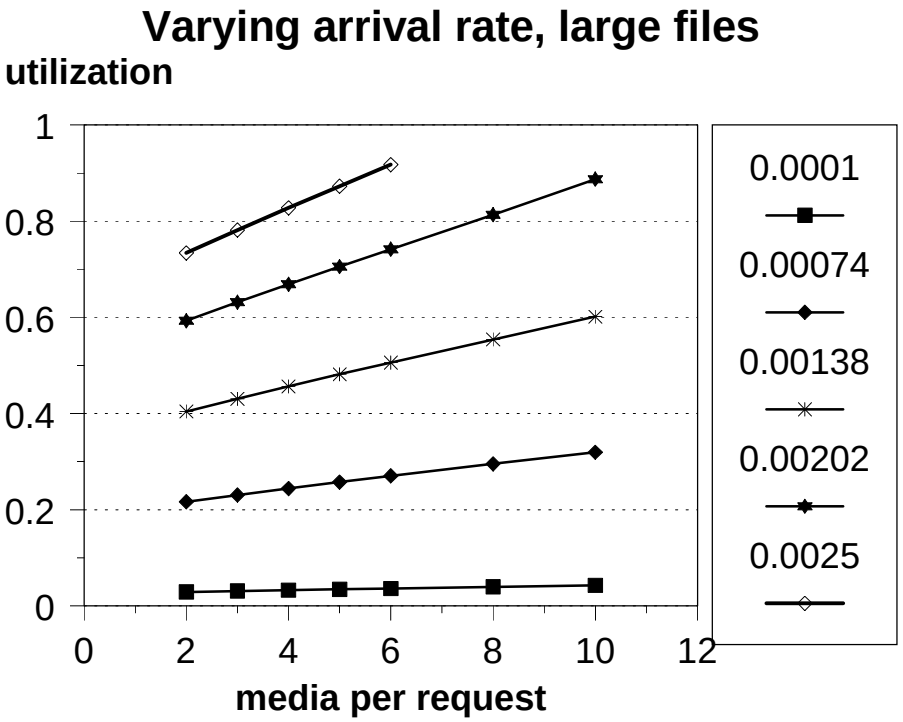

Figure 10: Drive utilization vs. media per request.

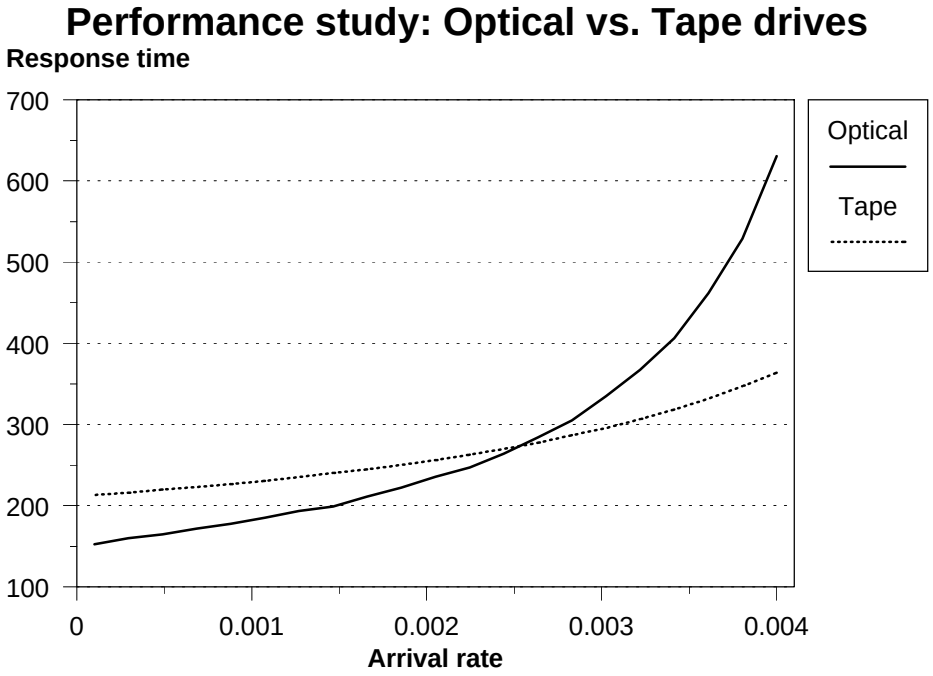

Figure 11: Hypothetical robotic storage library comparison.

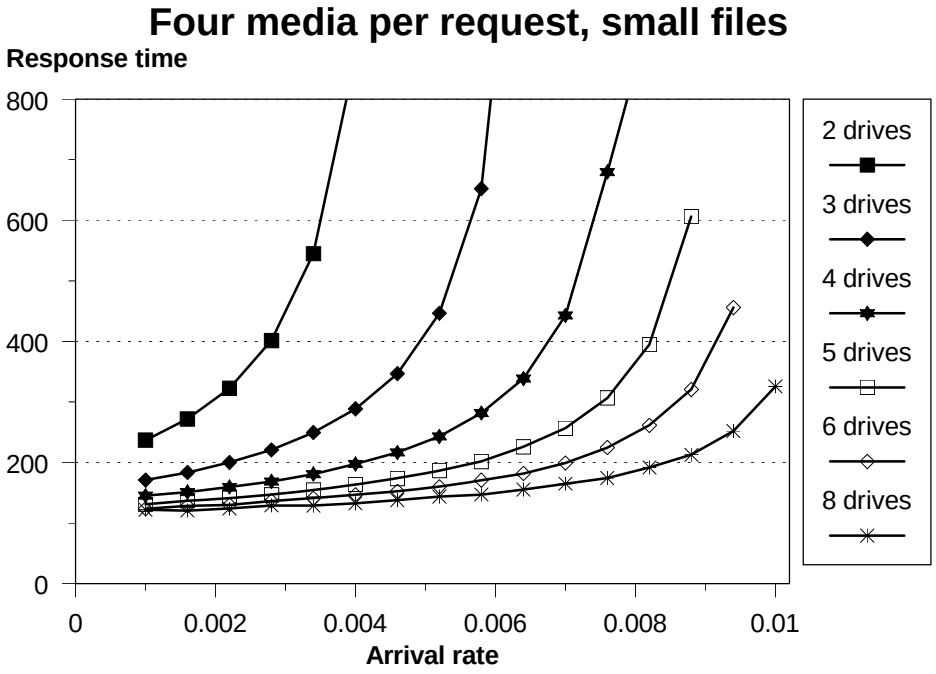

Figure 12: Response time vs. number of drives.

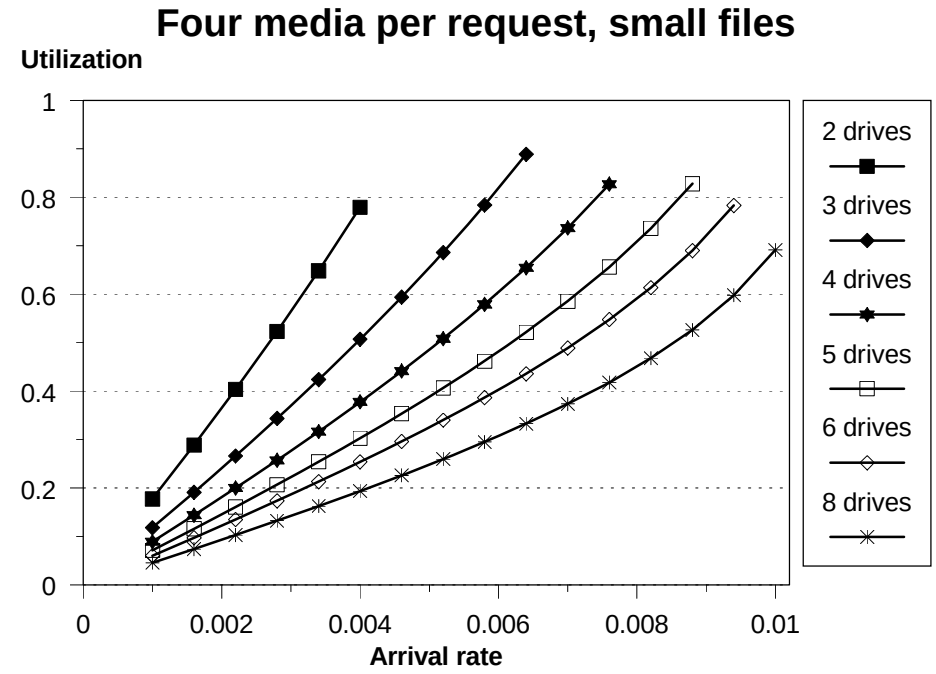

Figure 13: Drive utilization vs. number of drives.

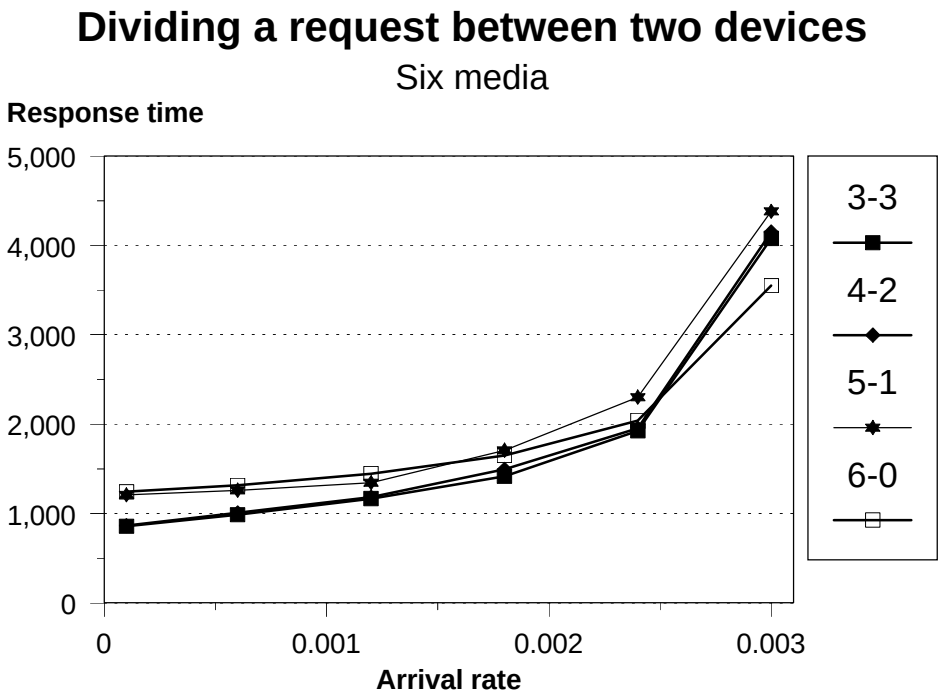

Figure 14: Response time of a fork-join request.

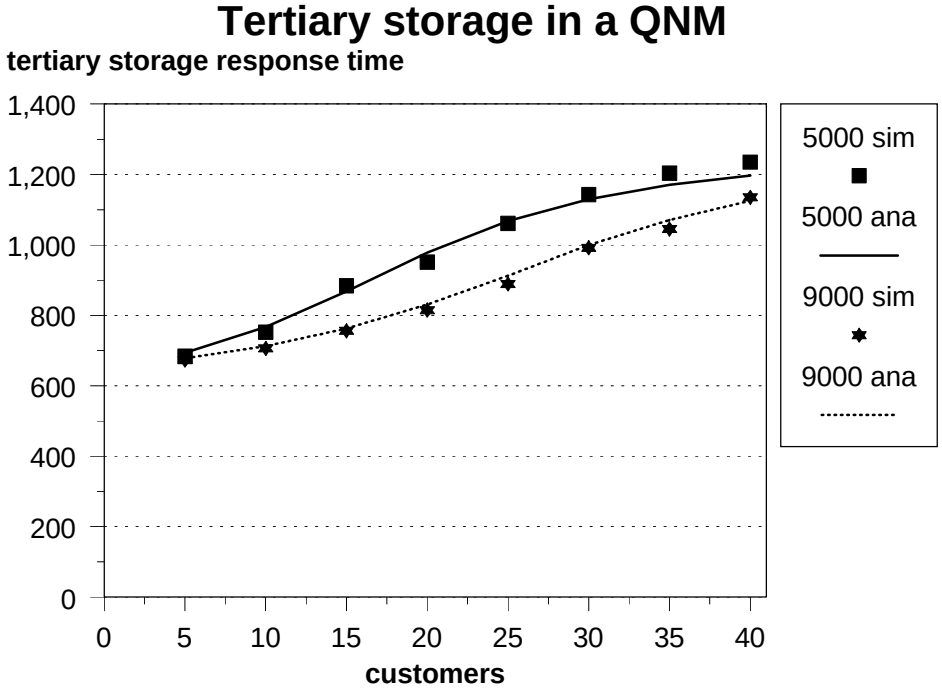

Figure 15: Incorporating the RSL model into a queuing network.