Batching and Dynamic Allocation Techniques for Increasing the Stream Capacity of an On-Demand Media Server *

Divyesh Jadav
ECE Department
Syracuse University, Syracuse, NY 13244
{divyesh, csrinlt}@cat.syr.edu

Alok Choudhary
ECE Department & Technological Institute
Northwestern University, Evanston, Illinois 60208.
choudhar@ece.nwu.edu

March 1, 1997

Abstract

A server for an interactive distributed multimedia system may require thousands of gigabytes of storage space and high I/O bandwidth. In order to maximize system utilization and thus minimize cost, the load must be balanced among the server’s disks, interconnection network and scheduler. Many algorithms for maximizing retrieval capacity from the storage system have been proposed. This paper presents techniques for improving server capacity by assigning media requests to the nodes of a server so as to balance the load on the interconnection network and the scheduling nodes. Five policies for dynamic request assignment are developed. An important factor that affects data retrieval in a high performance continuous media server is the degree of parallelism of data retrieval. The performance of the dynamic policies on an implementation of a server model developed earlier is presented for two values of the degree of parallelism.

1 Introduction

Digitalization of traditionally analog data such as video and audio, and the feasibility of obtaining networking bandwidths above the gigabit-per-second range are two key advances that have made possible the realization, in the near future, of interactive distributed multimedia systems. Multimedia data differs from unimedia data in the diversity of data sizes and the need to provide real-time guarantees for playback (video and audio data). One of the most pervasive interactive multimedia

---

*This work was supported in part by Intel Research Council, NSF Young Investigator Award CCR-9357840, NSF grant CCR-9509143, and the New York State CASE Center at Syracuse University. The authors thank the Caltech CCSF facilities for providing access to the Intel Paragon.
applications is media-on-demand which refers to the possibility of a consumer interactively retrieving multimedia data over high-speed networks from geographically distributed media servers. It is anticipated that a distributed media-on-demand (MOD) system will be built in a hierarchical manner, with clients connected to neighbourhood servers, which are connected to metropolitan servers, which in turn are connected to massive archive servers. This hierarchy of servers is similar to the memory hierarchy in a computer system. When the processor issues a request for data, the time penalty for retrieving the data is directly proportional to the hierarchical distance of the data from the processor, which constitutes the apex of the hierarchy. In a hierarchical MOD system, a similar relationship exists in this case with regard to the time delay for data retrieval by a client. On account of the real-time and storage capacity requirements of multimedia data, the servers are high-end machines with gigabytes of storage space and high I/O bandwidth. Moreover, higher up the hierarchy is the server, the higher its storage and I/O capacity. In order for the entire system to be cost effective, each server must be cost effective. Hence, it is essential to maximize the utilization of each resource type of the server. This paper deals with techniques to maximize the utilization of one of the resource types of a MOD server.

1.1 Related Work

Researchers have proposed various approaches for the storage and retrieval of multimedia data. [1, 2] have proposed a disk arm scheduling approach for multimedia data, and characterized the disk-level tradeoffs in a multimedia server. [3] proposed a model based on constrained block allocation, which is basically non-contiguous disk allocation in which the time taken to retrieve successive stream blocks does not exceed the the playback duration of a stream block. Storing a stream on a single disk restricts the retrieval bandwidth to the data transfer bandwidth of the disk. [4] get around this problem by striping media data across several disks in a round robin fashion. The effective retrieval bandwidth is then proportional to the number of disks used. Our server model (section 2) is similar to this model. Issues in designing MOD servers are discussed in [6]. [5, 14] studied efficient memory allocation and utilization techniques to maximize the number of supported users. [5] also studied cost tradeoffs and scalability issues in high performance MOD servers. Techniques for improving reliability and availability of the storage subsystem are studied in [13]. A component-wise instrumentation of the delays in a MOD [7], showed that variable delays become performance bottlenecks at high loads. Techniques for balancing the load on the storage devices of a MOD server were developed in [8]. Given the fact that a high performance MOD server consists of multiple processors (nodes) connected by an interconnection network, not much work has been reported on efficient use of the interconnection network so as to maximize server capacity. In this paper we address this issue by developing five policies - round robin (RR), minimum link allocation (MLA), minimum contention allocation (MCA), weighted minimum link allocation (WMLA) and weighted contention allocation (WMCA). An important factor that affects data retrieval time in a high performance continuous media server is the degree of parallelism of data retrieval. We investigate the effect of varying the degree of parallelism of data retrieval on the performance of the dynamic allocation policies. We have developed and implemented a logical model for a MOD server [7]. Performance results of the five policies under this model implemented on the Intel Paragon parallel computer are presented.

The rest of this paper is organized as follows: Section 2 explains the server and data scheduling model. Section 3 presents the five allocation strategies. Section 4 discusses various striping tradeoffs for distributing file data across storage nodes. We present and analyze performance results in
2 Server and Scheduling Model

At the heart of the system is a high-performance server optimized for fast I/O. A parallel machine is a good candidate for such a server because of its ability to serve multiple clients simultaneously, its high disk and node memory, and the parallelism of data retrieval that can be obtained by data striping. We assume that (a) the server is connected to clients by a high-speed wide-area network, which delivers data to clients reliably and at the required bandwidth. (b) Clients have hard deadlines i.e. they cannot tolerate jitter. (c) The data are stored at the server in compressed digital form, with the decompression being done at the client end. The server consists of multiple nodes interconnected by a network. Each node is a computer in its own right, with a CPU, RAM and secondary storage. There are three types of nodes, interface (I) nodes, storage (S) nodes and the object manager (O) node. The object manager receives all incoming requests for media objects, and delegates the responsibility of serving a request to one of the interface nodes. I nodes are responsible for scheduling and serving stream requests that have been accepted. Storage nodes store multimedia data on their secondary storage devices in a striped fashion, and retrieve and transmit the data to an interface node on request. The assumption about the architecture of the interconnection network is that any node can transfer data to and from any other node with approximately the same latency, under conditions of light load. For the purposes of this paper, we assume that the interconnection network is a direct network[9].

The data is compressed and striped across the storage nodes in a round-robin fashion. The number of nodes across which a data object is striped is called the stripe factor (SF). The collection of SF storage nodes that store an object is called a striping group. The data stored at a storage node consists of chunks of the object. The collection of chunks is called a subobject. Note that the collection of chunks of a subobject do not constitute a contiguous portion of the object: however the data within a chunk is a contiguous part of the entire object. This contiguous data is called a stripe fragment. Note that each storage node may have a single high-performance disk, or an array of slower, but cheaper disks. The point to note is that a storage node represents a virtual disk to an interface node. Since the stripe fragments on any given storage node’s disk are not consecutive fragments, it is not necessary to store them contiguously. Disk scheduling algorithms to optimize retrieval from the disk surface have been proposed [1, 11, 5], and can be used in our model. We are concerned with harnessing the parallelism provided by striped storage and balancing the load across the server subsystems.

Table 1 shows the parameters used by our model. $\delta_I$ is the time for which a packet sent by an I node to a client will last at the client. Hence this is also the deadline by which the next packet from the I node must be received at the client. Its value is given by:

$$\delta_I = \frac{P_I}{R_{pl}} \quad (1)$$

Once the requested SF stripe fragments from the S nodes have arrived at the destination I node, the latter arranges them in the proper sequence and continues sending packets of size $P_I$ to the client no less than every $\delta_I$ seconds. The buffer at the I node will last for $\delta_S$ time, before which the next set of stripe fragments must have arrived from the S nodes. The average time to
<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
<th>Units</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R_{pf}$</td>
<td>Required playback rate</td>
<td>bytes/sec</td>
</tr>
<tr>
<td>$P_I$</td>
<td>Size of packets sent by an $I$ node</td>
<td>bytes</td>
</tr>
<tr>
<td>$\delta_I$</td>
<td>Duration of a packet sent by an $I$ node</td>
<td>sec</td>
</tr>
<tr>
<td>$B_I$</td>
<td>Buffer size at an $I$ node</td>
<td>bytes</td>
</tr>
<tr>
<td>$P_S$</td>
<td>Size of packets sent by a $S$ node</td>
<td>bytes</td>
</tr>
<tr>
<td>$\delta_S$</td>
<td>Duration of data in $B_I$</td>
<td>sec</td>
</tr>
<tr>
<td>$T_f$</td>
<td>Period of issuing fetches to $S$ nodes from $I$ node</td>
<td>sec</td>
</tr>
<tr>
<td>$SF$</td>
<td>Stripe factor</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 1: The parameters used in this paper

retrieve $P_S$ bytes from a $S$ node is given by

$$\delta_{io} = \delta_{rq} + \delta_{avgseek} + \delta_{avgrot} + \delta_{tr_{ps}} + \delta_{nwp_{ps}}$$

(2)

where $\delta_{rq}$ is the time delay for a request from an $I$ node to reach a $S$ node, $\delta_{avgseek}$ and $\delta_{avgrot}$ are the average seek and rotational latencies for the disks being used, $\delta_{tr_{ps}}$ is the disk data transfer time for $P_S$ bytes, and $\delta_{nwp_{ps}}$ is the network latency to transport $P_S$ bytes from a $S$ node to an $I$ node. Note that equation 2 uses average seek and rotational latencies for disk accesses, for reasons explained in[7].

3 Stream Assignment Policies

In a given server configuration, only a finite number of nodes (interface nodes) would be connected to the high speed wide area network. Moreover, their position in the server architecture would be fixed a-priori. Secondly, since the secondary storage capacity for a given server configuration is finite, only a finite number of media objects can be stored in the secondary storage system at a time. In order to maximize the pool of potential clients, there would exist a tertiary storage system from which the most frequently requested objects are materialized on the secondary storage subsystem. However, the number of objects stored on secondary storage would be a slowly changing set. Given the fact that the position of the $I$ nodes and the storage pattern of data on $S$ nodes is fixed, the problem is one of assigning accepted media requests to $I$ nodes so as to minimize the incremental workload due to the new requests on the server's resource types. This allows the server to maximize the number of streams that it can source. The network communication time is the sum of two factors - the network latency in the absence of blocking, and the blocking time due to link contention in the interconnection network i.e.,

$$\delta_{nwp_{ps}} = \delta_{nwcomm} + \delta_{nwbl}$$

(3)

For a given message size and interconnection network, the former is fixed; while the latter depends on the network traffic. There is another variable delay in the retrieval time: when multiple requests arrive at a $S$ node, only a finite number of them can be served at a given time; this causes a queueing delay at the $S$ nodes. If $\delta_{sq}$ denotes this queueing delay, equation 2 requires to be modified to:

$$\delta_{io} = \delta_{rq} + \delta_{seek} + \delta_{rot} + \delta_{tr_{ps}} + (\delta_{nwcomm} + \delta_{nwbl}) + \delta_{sq}$$

(4)
The effect of the variable delays $\delta_{nwx}$ and $\delta_{SQ}$ on total retrieval time at various workloads was studied in [7]. At heavy workloads, these delays become the limiting factors on stream capacity. Any mechanism that reduces either or both of these quantities improves performance. In this paper we show that techniques that reduce $\delta_{nwx}$ by minimizing link contention translate into the ability to support more streams.

### 3.1 RR Assignment Policy

This is the simplest policy. If $n$ is the total number of interface nodes, and the $i$th request was assigned to interface node $k$, then the $(i + 1)$th request will be assigned to interface node $(k + 1)$ mod $n$. This policy is simple, requiring $O(1)$ time to execute. The maximum workload imbalance in terms of number of streams served per I node is at most 1. However, this policy does not balance or minimize the load imposed on the interconnection network.

### 3.2 MLA Policy

This policy aims to minimize the total number of links that the data for an object has to travel from the $SF$ storage nodes on which it is stored to the I node which sends it to the outside world. If $l_{i_s}$ denotes the number of links between interface node $I_i$ and storage node $S_j$, then the cost of assigning a stream request to interface node $I_i$ under this policy is:

$$CA_{MLA}(I_i) = \sum_{j=1}^{SF} l_{i_s}$$

This policy will assign a request to interface node $p$.

$$p = i : (\min(CA_{MLA}(I_i)) \quad i = 1, 2, ..., n)$$

i.e., the request is assigned to that interface node which is closest in terms of the total number of links that need to be traversed from the S nodes to the I node. This policy tries to minimize the number of streams using a given link. But, it does not take into account the pre-existing link load. Nor does it balance the total stream load among all the interface nodes.

### 3.3 MCA Policy

In this policy, state information is maintained about the usage of each link. Specifically, whenever a new request is assigned to an interface node, the load imposed on the interconnection network by data traffic due to that stream is calculated and the total load on the network due to all streams is updated. When a stream terminates, the load due to it is decremented from the total network load. If $c_{i_s, S_j}(k)$ is the cost of using the $k$th link on the path from storage node $S_j$ to interface node $I_i$, then the cost of assigning a stream request to interface node $I_i$ under this policy is:

$$CA_{MCA}(I_i) = \sum_{j=1}^{SF} \sum_{k=1}^{l_{i_s}} c_{i_s, S_j}(k)$$

This policy will assign a request to interface node $p$.

$$p = i : (\min(CA_{MCA}(I_i)) \quad i = 1, 2, ..., n)$$
The cost of using a link is directly proportional to the traffic that the link carries. The link traffic due to accepting a new request is updated as follows:

\[
\begin{align*}
\text{for } (j = 1 \text{ to } SF) \\
\text{for } (k = 1 \text{ to } l_{ip_j}) \\
c_{ip_j[k]} &= c_{ip_j[k]} + ld
\end{align*}
\]

(9)

where \(ld\) is a scalar that reflects the load imposed by the new stream on link \(k\). Its value is implementation-dependent: it depends on the network, the size of packets being transferred and the playback rate of the stream. The premise behind this policy is that the load should be distributed evenly over the interconnection network. If some links are more heavily used than others, contention in these links increases network blocking effects, which in turn degrades server performance. Note that since the traffic pattern in the interconnection network for data packets consists of storage nodes sending data to interface nodes, there is a possibility of hot spots developing at the links around the interface nodes. This policy tries to prevent the formation of such hot spots by allocating requests to interface nodes so that aggregate link traffic is distributed as evenly as possible over the entire interconnection network.

### 3.4 WMLA and WMCA Policies

The MLA policy tries to minimize the total number of links that data for a stream will have to travel, while the MCA policy tries to minimize link contention by distributing traffic over more lightly used links. However, neither of them tries to balance the load across the interface nodes. An interface node can source only a finite number of streams; beyond this limit client deadlines may be missed due to excessive scheduling overhead. The weighted MLA and MCA policies try to balance the load across both the network and the I nodes. This is done by factoring in the number of streams that a candidate I node is serving in the cost equation. Specifically, if \(M_I\) is the number of streams being served by interface node \(I_i\), then the cost of assigning to it the responsibility of serving a request under WMLA and WMCA (respectively) is:

\[
CA_{WMLA}(I_i) = \alpha \times M_I + \beta \times CA_{MLA}(I_i)
\]

(10)

\[
CA_{WMCA}(I_i) = \alpha \times M_I + \beta \times CA_{MCA}(I_i)
\]

(11)

where \(\alpha\) and \(\beta\) are fractions that sum to 1, and \(CA_{MLA}(I_i)\) and \(CA_{MCA}(I_i)\) are given by equations 5 and 7 respectively. The criterion for selecting a candidate I node is similar to that for the respective unweighted cases (equations 6 and 8 respectively); so are the running times. The value to assign to the weight is a design choice that depends on the network size and topology, routing strategy and the maximum number of streams that an I node can source. Note that WML(C)A with \(\alpha = 1, \beta = 0\) is equivalent to RR, while WML(C)A with \(\alpha = 0, \beta = 1\) is equivalent to ML(C)A.

### 4 Concurrency of Retrieval

An important factor that affects data retrieval time is the stripe factor \((SF)\). We differentiate between *wide striping*, in which a media object is striped across all the storage nodes of a server, and *narrow striping*, in which an object is striped across a fraction of the total storage nodes in a
server. Each approach has its advantages and disadvantages. In the latter case, a striping group containing a frequently accessed object can become a bottleneck for the server in the absence of object replication. For example, in a video-on-demand case, it is but natural that some videos will be more frequently accessed than others: newly released videos are likely to be more frequently accessed than older videos. Wide striping avoids the formation of such bottlenecks by striping an object across all the storage nodes, which has the effect of balancing the load across all storage nodes even for skewed access patterns. However, wide striping also has some drawbacks. In wide striping, there is only one striping group: this complicates system reconfiguration. Since each storage node has some data for all objects stored in the server, most of the chunks of all the objects may have to be moved. This may lead to an undesirable reconfiguration load on all storage nodes. In contrast, in narrow striping, only the storage nodes in the striping group where the object is stored incur the penalty. Secondly, the larger the size of a striping group (i.e. the wider the striping), the lower is the reliability, and also the availability, of the entire system [12, 13] in the event of a single disk failure. In this paper we consider only narrow striping.

We mentioned earlier that a storage node represents a virtual disk to an interface node. Whether wide striping or narrow striping is used to store data files, the important parameter for data retrieval and delivery at run-time is the degree of parallelism seen by an interface node. We define now a new parameter, the concurrency factor (CF). This is an integer, $1 \leq CF \leq SF$, and denotes the degree of parallelism employed by an interface node in requesting data from the storage nodes. For example, figure 1 shows a media object, $M_1$, with a stripe factor of 6 and concurrency factor of 2. In the figure, the interface node (I) concurrently retrieves data from $S_1$ and $S_2$ in service rounds 0, 3, 6,... from $S_3$ and $S_4$ in service rounds 1, 4, 7,..., and from $S_5$ and $S_6$ in service rounds 2, 5, 8,.... In other words, in each service round of a given media stream, an interface node sends out a batch of CF requests to CF of the SF storage nodes on which the stream's data is stored. Earlier work [17] (where the concurrency factor was called the batch size)
<table>
<thead>
<tr>
<th>Description</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Required playback rate (R_pl)</td>
<td>1.5 Mbits/sec</td>
</tr>
<tr>
<td>Size of packets sent by an I node (P_I)</td>
<td>80 Kbytes</td>
</tr>
<tr>
<td>Size of packets sent by a S node (P_S)</td>
<td>160/320 Kbytes</td>
</tr>
<tr>
<td>Minimum disk seek time</td>
<td>4 msec</td>
</tr>
<tr>
<td>Maximum seek time</td>
<td>30 ms</td>
</tr>
<tr>
<td>Time for one rotation</td>
<td>10 ms</td>
</tr>
<tr>
<td>Disks per storage node</td>
<td>2</td>
</tr>
<tr>
<td>Disk data transfer rate</td>
<td>7.5 MBytes/sec</td>
</tr>
<tr>
<td>Stripe Factor (SF)</td>
<td>4</td>
</tr>
<tr>
<td>Concurrency Factor (CF)</td>
<td>4/2</td>
</tr>
<tr>
<td>Num. of Interface nodes</td>
<td>11</td>
</tr>
<tr>
<td>Num. of Storage nodes</td>
<td>52</td>
</tr>
<tr>
<td>Num. of media objects</td>
<td>100</td>
</tr>
<tr>
<td>Evaluation machine</td>
<td>Intel Paragon</td>
</tr>
</tbody>
</table>

Table 2: The parameter values used

showed that when the buffer space at an I node and the stripe factor are kept fixed, and the concurrency factor is varied, the concurrency factor has an inverse relationship with the maximum number of concurrent streams that can be supported. We investigate below the effect of varying the concurrency factor on the performance of the dynamic allocation policies developed above.

5 Performance Evaluation

We have implemented our logical server model on the Intel Paragon parallel computer [16]. The Intel Paragon is a mesh-based architecture with Intel i860XP microprocessors. Interprocessor communication is done using wormhole routing [9]. The data access pattern is assumed to follow a Zipfian distribution with parameter 0.271[10]. Due to storage space and availability of real-world data limitations. the disk access part was simulated by elapsing the system timer on each storage node. Disk retrieval was simulated by assuming that the stripe fragments are stored on the disk using a random placement model [15]. Table 2 shows the values of the parameters defined in table 1 that we used for our simulation. We used two concurrency factors, 4 and 2 respectively. The stripe factor was kept fixed at 4. and the buffer size at each I node for each stream was kept fixed at 640 Kbytes. Accordingly. P_S is 160 Kbytes for CF = 4 and 320 Kbytes for CF = 2. The traffic was generated as follows. Initially. requests for videos are sent to the object manager at random times. with average inter-arrival time of 2 seconds. Each video lasts for about 10 minutes. As soon as a video terminates, a new request for a video is sent to the object manager. Figure 2 shows the distribution of I nodes and S nodes used. The value of the parameter ld for the MCA policy (subsection 3.3) we used was 1. since. for a given value of the concurrency factor. the value of P_S and R_pl is the same for all streams. The load on the server was increased by incrementally increasing the number of object requests. The same data distribution and request pattern were used for each experiment.

1Numerous tradeoffs are possible with respect to the data partitioning strategy, which are well reported in [4]. However. these are not the subject of this paper.
Figure 2: Distribution of I, S and O nodes

<table>
<thead>
<tr>
<th>Policy</th>
<th>Average streams per I node</th>
<th>Standard Deviation ($\sigma_i$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>RR</td>
<td>54.54</td>
<td>0.49</td>
</tr>
<tr>
<td>MLA</td>
<td>47.64</td>
<td>29.79</td>
</tr>
<tr>
<td>MCA</td>
<td>54.54</td>
<td>24.58</td>
</tr>
<tr>
<td>WMLA ($\alpha = 3 = 0.5$)</td>
<td>54.54</td>
<td>12.94</td>
</tr>
<tr>
<td>WMCA ($\alpha = 3 = 0.5$)</td>
<td>54.54</td>
<td>4.46</td>
</tr>
</tbody>
</table>

Table 3: Standard Deviation of stream load per I node

5.1 Comparison of load balancing ability

In order to compare the distribution of stream requests to the I nodes, the number of streams served by each interface node was measured for the same total number of streams served for each policy. We start with identical values of the concurrency factor and the stripe factor ($CF = SF$). Figure 3 illustrates these values for each I node in figure 2 for a total server load of 526 streams for the MLA policy and for a load of 600 streams for the other four policies. (the maximum number of streams supported by MLA was only 526. Each of the other 4 policies supported at least 600 streams). We first compare the RR, MLA and MCA policies. We note from the table that the RR policy performs best in terms of balancing stream load across the interface nodes. A measure of the degree to which a request assignment policy balances stream load across the interface nodes is the standard deviation of the number of streams per interface node, $\sigma$ (Table 3). The standard deviation of the number of requests per I node for MLA, $\sigma_{MLA}$, is the worst among the standard deviations of RR, MCA and MLA. The reason for this is the local nature of the algorithm used by the MLA policy to determine an I node to which an incoming request is to be assigned. When the data access pattern is skewed (i.e., some files are more popular than others), as it is in the case of the Zipfian distribution, use of the MLA policy results in severe load imbalance on the I nodes. Note that the Zipfian distribution approximates data access patterns observed in practice[10]. Consider now the WMLA and WMCA policies. The graphs in figure 3 for the WMLA and WMCA policies are for values of $\alpha$ and $\beta$ of 0.5 each. In this case too, the load balancing
Figure 3: Comparison of request assignment of RR, MLA, MCA, WMLA and WMCA policies
of WMCA is better than that of WMLA. Although \( \sigma_{WMLA} \) is lesser than \( \sigma_{MLA} \), it is still greater than \( \sigma_{WMCA} \). In summary, the weighted assignment policies improve the load balancing ability at the I nodes as compared to the pure schemes; however, MCA gives better performance than MLA, and WMCA gives better performance than WMLA.

5.2 Comparison of Network Blocking Time

With reference to equation 4, the networking blocking time for each packet requested by an I node from a S node, \( \delta_{\text{network}} \), was measured as follows: \( \delta_{\text{seek}} \) and \( \delta_{\text{rot}} \) were measured at run time. Given a disk and a value of \( P_S \), \( \delta_{\text{tr}_{PS}} \) can be computed. \( \delta S_Q \) is given by \( \delta S_Q = \Delta_S - (\delta_{\text{seek}} + \delta_{\text{rot}} + \delta_{\text{tr}_{PS}}) \), where \( \Delta_S \) is the time interval between arrival of the packet request at the S node, and the time when the packet is sent to the requesting I node. \( \delta_{\text{network}} \) is a known when \( P_S \) and network bandwidth in the absence of blocking is fixed. The round trip time for the sequence of events represented by equation 4, \( \delta_{\text{rt}} \), is measurable at run time. Hence, the only unknowns in equation 4 are \( \delta_{\text{rt}} \) and \( \delta_{\text{network}} \), from which the latter can be approximated (by neglecting \( \delta_{\text{rt}} \)). Figure 4 shows the distribution of packet network blocking time, \( \delta_{\text{network}} \), for the 5 policies, for a concurrency factor of 2, stripe factor of 4, and a total server load of 700 streams (except for MLA, which supported a maximum of only 541 streams for this configuration). Bins of size 5 ms each (horizontal axis) were used to count the distribution of network blocking time for each packet. Overflow values were placed in the last bin. The vertical axis shows the percentage of packets that fall in each bin. For real-time retrieval of data with a high quality of service (QOS), it is desirable that the variable components in equation 4 be bounded and of minimal value. The higher the cumulative percentage of packet blocking times falling in the leftmost bins, the better is the performance of the policy. Accordingly, the performance of the policies with respect to this metric (in ascending order) is RR, MLA, MCA, WMLA and WMCA. The frequency distribution of blocking times for the last four are not very different from each other, suggesting that the number of supportable streams for (W)MLA and (W)MCA policies should be nearly the same. However, this is not the case, as shown below.

5.3 Stream Sourcing Capacity

We now compare the policies with respect to the more important metric of stream sourcing capacity. Table 4 shows the maximum number of streams that were supported by each policy, together with the percentage improvement over the RR policy. The table also shows the effect of varying the concurrency factor, keeping the buffer size at an I node fixed. The second column is for a \( CF \) of 4, and the third column is for a \( CF \) of 2. Consider first column 2 of the table (\( CF = 4 \)). The MLA policy performed the worst for this configuration. This is because of the local nature of the algorithm used by the MLA policy to determine an I node to which an incoming request is to be assigned. For highly skewed data access patterns, like the Zipfian distribution of requests used here, this local nature of the MLA policy results in severe load imbalance among the I nodes. For any server configuration, each I node can source only a finite number of streams. When the number of requests assigned to an I node exceeds this limit, deadlines for requests are likely to be missed. In the case of the MLA policy, even though most I nodes are assigned few requests, and can handle more, some I nodes are assigned more requests than a single I node can handle (Figure 3). Such I nodes become the bottleneck for the entire server. For this reason, MLA performs worse than even the RR policy.
Figure 4: Frequency distribution of packet network blocking time for RR, MLA, MCA, WMLA and WMCA policies
<table>
<thead>
<tr>
<th>Policy</th>
<th>Concurrency Factor = 4</th>
<th>Concurrency Factor = 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Weight α(1 – β)</td>
<td>Max. # of streams</td>
</tr>
<tr>
<td>RR</td>
<td>1.00</td>
<td>623</td>
</tr>
<tr>
<td>MLA</td>
<td>0.00</td>
<td>526</td>
</tr>
<tr>
<td>MCA</td>
<td>0.00</td>
<td>683</td>
</tr>
<tr>
<td>WMLA</td>
<td>0.25</td>
<td>722</td>
</tr>
<tr>
<td>WMCA</td>
<td>0.25</td>
<td>847</td>
</tr>
<tr>
<td>WMLA</td>
<td>0.50</td>
<td>790</td>
</tr>
<tr>
<td>WMCA</td>
<td>0.50</td>
<td>886</td>
</tr>
<tr>
<td>WMLA</td>
<td>0.75</td>
<td>828</td>
</tr>
<tr>
<td>WMCA</td>
<td>0.75</td>
<td>902</td>
</tr>
</tbody>
</table>

Table 4: The maximum number of streams supported for the request assignment policies for concurrency factors of 4 and 2 respectively.

The RR policy best balances the stream among the I nodes (minimum σ), but, it makes no effort to balance the load on the interconnection network. Hence, with the exception of the MLA policy, all other policies perform better than RR. At the other end of the spectrum are the MLA and MCA policies: they try to reduce load on the interconnection network by minimizing link contention; however, they do not try to balance the load across the I nodes. MCA outperforms RR by 9.6%, but MLA underperforms RR by 15.6%. In between RR, on the one hand, and MLA and MCA, on the other, are the WMLA and WMCA policies that try to balance the load on both the network as well as the I nodes. This translates into superior performance over RR, MLA and MCA.

These observations about the relationship between the performance of RR, MLA, MCA, WMLA and WMCA policies are true even when the concurrency factor is varied, as can be seen from column 3 of table 4. In all cases, the lower concurrency factor resulted in more streams being supported for the same allocation policy.

In summary, although the performance of (W)MLA is similar to that of (W)MCA as far as network blocking time is concerned, the load imbalance on the I nodes is much higher for the former than for the latter (Table 3). This explains why (W)MCA consistently outperforms (W)MLA.

6 Conclusions

Issues similar to those addressed in this paper were considered in [18]. The authors in that paper compared and evaluated two physical architectures for a logical media server model which was similar to ours. However, their work differed from ours in that the results reported were for simulations based on an analytical model, whereas the results reported in this paper are for an implementation. In this paper, we considered narrow striping, whereas [18] considered wide striping of the file data. Moreover, we used a Zipfian distribution of accesses to files, whereas the request pattern used in [18] were not reported.

In this paper we developed five policies for assigning requests to the interface nodes in a high-performance multimedia server. MCA, WMLA and WMCA each outperformed RR in terms of
Figure 5: Effect of changing weight values on performance of WMLA and WMCA policies, for concurrency factor of 4. Graph on the left is for maximum streams supported, while that on the right is for standard deviation of stream load per node.

Figure 6: Effect of changing weight values on performance of WMLA and WMCA policies, for concurrency factor of 2. Graph on the left is for maximum streams supported, while that on the right is for standard deviation of stream load per node.
number of streams. RR best balances inter I node load, followed by WMCA and WMLA. WMCA with proper choice of weights gave highest throughput. The (W)MCA policy is a global one, as it takes into account the load on a link due to the existing traffic. (W)MLA, on the other hand is a local optimization that is oblivious of the load imposed by other nodes. This explains why WMCA gave the best throughput. Figure 5 shows the effect of varying the weight values on the maximum number of supported streams for the five policies for a concurrency factor of 4, and figure 6 shows the effect of varying the weights for a concurrency factor of 2.

In this paper, the experiments that we conducted also looked at the effect of varying the concurrency factor for a fixed stripe factor and I-node buffer size on the performance of the dynamic policies. It was reported in [17] that for a given stripe factor and buffer size at an interface node, the maximum number of concurrent streams that could be supported had an inverse relation to the concurrency factor. Our experiments confirm this relationship. However, for a given concurrency factor, the relative performance of the dynamic allocation policies holds: the weighted policies perform better than RR and the non-weighted policies. Varying the concurrency factor does not change this relationship.

For the parameters and data access pattern considered in this paper, α = 0.75, β = 0.25 gave the best performance for both WMLA and WMCA policies. The optimum values to assign to the weights is an implementation-dependent problem that depends on the network topology, routing strategy and the maximum streams that an I node can source. However, changing the ratio values of α and β in one direction makes the assignment criterion tend to RR (α = 1, β = 0). While changing the ratio in the opposite direction will make it tend to ML(C)A (α = 0, β = 1). We have shown that values in between give better performance than these extremes. This is so because such values try to balance both the load on the I nodes and the load on the interconnection network, unlike the extreme cases, which balance one or the other.

References


