A MetaObject Framework for QoS-Based Distributed Resource Management

Nalini Venkatasubramanian⋆ Carolyn Talcott⋆⋆ Gul Agha⋆⋆⋆

Abstract: *Systems that provide distributed multimedia services are subject to constant evolution - customizable middleware is required to effectively manage this change. In this paper we present a meta-architectural framework for customizable QoS-based middleware using Actors, a model of concurrent active objects. Middleware services for resource management execute concurrently with each other and with application activities— scheduling, protocols providing security and reliability, load balancing and stream synchronization can therefore potentially interfere with each other. To ensure cost-effective QoS in distributed multimedia systems, safe composability of resource management services is essential. For instance, system protocols and activities must not cause arbitrary delays in the presence of timing based QoS constraints. Using TLAM, a semantic model for specifying and reasoning about components of open distributed systems, we show how a QoS brokerage service can be used to coordinate multimedia resource management services in a safe, flexible and efficient manner.*

In the coming years, distributed multimedia servers will deployed to deliver a variety of interactive, digital multimedia(MM) services over emerging broadband (wide-area) networks [6] to form a wide-area infrastructure. Applications such as telemedicine, distance learning and electronic commerce exhibit varying requirements such as timeliness, security, reliability and availability. Many MM applications can tolerate minor, infrequent violations of their performance requirements specified as a quality-of-service (QoS) parameter, for e.g., tolerable jitter in a video frame. Systems that provide distributed multimedia services are continuously changing and evolving. For instance, the set of servers, clients, user requirements, network and system conditions, in a wide area infrastructure are changing continuously. Future applications will require dynamic invocation and revocation of services distributed in the network without violating QoS constraints of ongoing applications. In order to manage the distributed components and adapt to the above dynamic changes in multimedia applications, customizable middleware services are required. Today, the task of distributed systems management is performed in middleware layers using frameworks such as CORBA and DCOM. Such frameworks are designed for heterogeneous interoperability but are limited in the degree of flexibility and customizability of services. They provide only limited capabilities for the specification and adaptation of end-to-end QoS properties. Customizable middleware allows us to deal with changes in systems and applications in a non-intrusive way.

In this paper, we describe techniques for developing modular components for customizable middleware infrastructures. We develop a model for customizable, cost-effective middleware to enforce QoS requirements in multimedia applications. Specifically, we deal with the modeling of high level behaviors that describe the distributed middleware components and reason about the non-interference requirements of modules implementing these behaviors. For instance, we illustrate a placement policy that determines the degree of replication necessary for popular MM objects using a cost-based optimization procedure based on *a priori* predictions of expected subscriber requests. For scheduling MM requests, we represent an adaptive scheduling policy that compares the relative utilization of resources in a multimedia server to determine an assignment of requests to replicas. To optimize storage utilization, we introduce methods for dereplication of MM objects based on changes in their popularity and in server usage patterns. The TLAM (two level actor model) developed earlier [40, 39] provides a basis for reasoning about composability properties of resource management algorithms. The multimedia meta-architecture developed in this paper has been used in an experimental implementation of resource management mechanisms for distributed MM servers. The semantic model is used to verify safe interaction of the implemented mechanisms. Details of the mechanisms implemented and the performance evaluation of the composite environment is described in [38, 35]; detailed formal specification and correctness reasoning of system activities is described in [41].

Prior Performance Results: Performance studies show that application objects can be managed effectively by composing multiple resource management activities managed at the metalevel [38]. Figure 1 illustrates the performance, measured by request rejection rate, of various policies for load management - (a) purely adaptive (on-the-fly)

⋆ Contact Author: Nalini Venkatasubramanian Dept. of Information and Computer Science, Univ. of California Irvine, Irvine, CA 92697-3425, Email: nalini@ics.uci.edu

⋆⋆ Department of Computer Science, Stanford University, Stanford, CA 94305, Email: clt@cs.stanford.edu

⋆⋆⋆ Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, IL 61801, Email: agha@cs.uiuc.edu

scheduling and placement(P1), (b) purely predictive (decided apriori) scheduling and placement(P2), (c) composite policies that provide adaptive scheduling and predictive placement(P3 and P4, an optimized version of policy P3). The left hand side graph illustrates the request rejection rate under purely adaptive policies for placement and scheduling. *Startup latency* is a QoS factor that indicates how long the user is willing to wait for a replica to be created adaptively. The graph demonstrates that when the startup latency is below a threshold value (2 min), the purely adaptive mechanisms, represented by P1 force a very large fraction of the requests received to be rejected. Assuming that startup latency is sufficiently large, the right hand side depicts the inadequacy of P2, that relies on only predictive policies for scheduling and placement. In comparison, the other 3 policies (P1, P3 and P4), show hardly any rejects (indicated by the overlapping lines in the graph). As can be observed from the performance results, the ability to run multiple policies simultaneously (as in cases P3 and P4) reduced the total number of rejected requests in the overall system. In this paper, we study complex interactions that can arise due to the simultaneous execution of multiple system policies. The

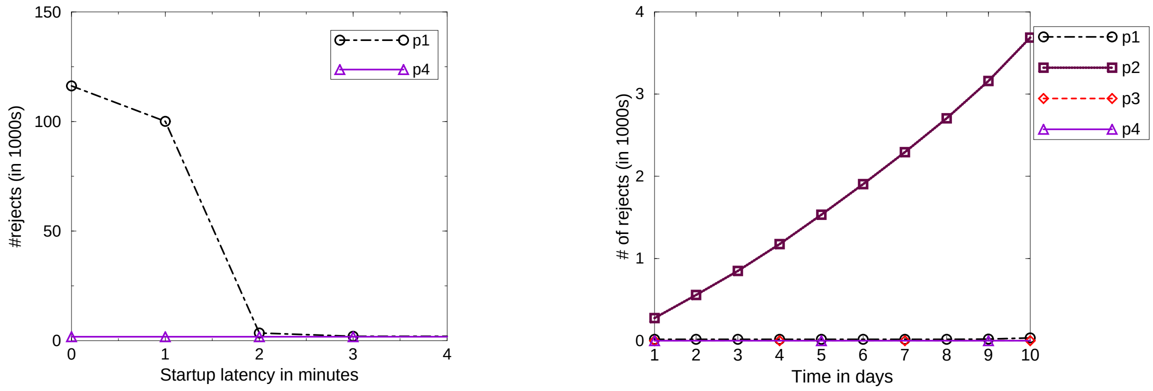

Fig. 1. Comparison of the performance of load management policies for request scheduling and video placement in a distributed video server.

rest of this paper is organized as follows. Section 1 reviews the two level semantic framework for distributed resource management based on Actors, a model of concurrent objects. Section 2 describes a multimedia server: its physical and software architectures, based on the two-level model of Section 1, and a mapping of the software architecture onto the physical architecture. Section 3 describes several resource management policies and optimization strategies that we have used in realizing the software architecture. Section 4 describes how we formally model the architecture and policies in the TLAM framework and the kinds of properties that can be expressed and proved. Section 5 discusses related work and outlines areas for future research.

1 The Two Level MetaArchitecture

The Actor model is a model of concurrent active objects that has a built-in notion of encapsulation and interaction and is thus well-suited to represent evolution and co-ordination among interacting components in distributed multimedia applications. In the actor paradigm, the universe contains computational agents called *actors*, distributed over a network. Traditional passive objects encapsulate state and a set of procedures that manipulate the state; actors extend this by encapsulating a thread of control as well. Each actor potentially executes in parallel with other actors and interacts only by sending and receiving messages. (See [1, 2] for more discussion of the actor model, and for many examples of programming with actors.)

In [40, 35], we presented the TLAM(Two Level Actor Machine) framework for specifying, composing and reasoning about resource management services in open distributed systems. In the TLAM, a system is composed of two kinds of actors, base actors and meta actors, distributed over a network of processing nodes. *Base-actors* carry out application level computation, while *meta-actors* are part of the runtime system which manages system resources and controls the runtime behavior of the base level. A TLAM provides an abstract characterization of actor identity, state, messages, and computation, and of the connection between base and meta level computation. Meta-actors communicate with each other via message passing as do base-actors, and they may also examine and modify the state of the base actors located on the same node. Base level actors and messages have associated runtime annotations that can be set and read by meta actors, but are invisible to base level computation. Actions which result in a change of base-level state are called events. Meta actors may react to events occurring on their node.

A TLAM *configuration*, C, has a set of base and meta level actors and a set of undelivered messages. The actors are distributed over the TLAM nodes. Each actor has a unique name (address) and the configuration associates a current state to each actor name. The undelivered messages are distributed over the network – some are traveling along communication links and others are held in node buffers. The semantics of a TLAM is given by a labeled transition relation on configurations. There are two kinds of transitions: communication and execution. Communication transitions move undelivered messages around the network and are the same in every TLAM. An execution transition is a computation step taken by a base or meta level actor. These transitions are described by *reaction rules* specific to a particular TLAM that determine how individual actors react to messages received, and in the case of meta actors how they react to events. A *computation path* for a configuration is a possibly infinite sequence of labeled transitions. The semantics of a configuration is the set of fair computation paths starting with that configuration. A *system* is a set of configurations closed under the transition relation.

Core Services for Resource Management: Meta-level actors provide services that customize various aspects of distributed systems management. In practice, multiple system and application activities occur concurrently in a distributed system, e.g. scheduling, load balancing, protocol processing, stream synchronization – such compositions can lead to complex interactions. Consider the following example of a system where distributed garbage collection and process migration can proceed concurrently. If processes in transit are not accounted for, the garbage collection process can potentially destroy accessible information. In general, risks that arise due to composition of services include loss of information, possible nonterminations that cause deadlocks and livelocks, dangling resources, inconsistencies, and incorrect execution semantics. Current approaches to deal with such interference include: serializing or delaying activities, which may cause QoS violations; and designing a closed system, which reduces portability of code.

Composability of resource management activities is not just desirable, it is *essential* to ensure cost-effective QoS in distributed multimedia systems. For instance, system protocols and activities must not cause arbitrary delays in the presence of timing based QoS constraints. This adds complexity to the design of such systems. To ensure non-interference and manage the complexity of reasoning about components of open distributed systems in general, our strategy is to identify key system services where non-trivial interactions between the application and system occur, i.e. base-meta interactions. We refer to these key services as *core services*. As a starting point, we have identified three core services:

- Recreation of services/data at a remote site - *remote creation*
- Capturing information at multiple nodes/sites - *distributed snapshot*
- Interactions with a global repository - *directory services*.

We have studied non-interference requirements that would allow these services to safely operate concurrently. A detailed case study of the composition of remote creation and a distributed snapshot formalized in the TLAM framework appears in [35]. Core services may be used in specifying and implementing more complex activities within the actor framework as purely meta-level interactions (Left hand side of Figure 2). For example, remote creation can be used as the basis for designing algorithms for activities such as migration, replication and load balancing. Distributed snapshots (cf. [8]). can also be used as the basis for global activities like quiescence detection, checkpointing and recovery and distributed garbage collection (cf. [37]). Similarly, a directory service can be used to design access control mechanisms, routing policies and group based communication.

2 A Meta-Architecture for QoS-based Services

Building on the two-level model described in the previous section, we develop a meta-architectural model of a multimedia server that provides QoS based services to applications. The physical architecture of the MM server consists of:

- A set of data sources (DS) that provide high bandwidth streaming MM access to multiple clients. Each independent data source includes high capacity storage devices (e.g. hard-disks), a processor, buffer memory, and high-speed network interfaces for real-time multimedia retrieval and transmission
- a specific node designated as the distribution controller (DC) that coordinates the execution of requests on the data sources and
- a tertiary storage server that contains the MM objects; replicas of these MM objects are placed on the data source nodes.

All the above components are interconnected via an external distribution network that also transports multimedia information to computers and set-top devices at the client end. A lower speed back-channel conveys subscriber commands back to the data sources via the DC.

The software architecture of a multimedia server consists of two subsystems - the base level and meta-level subsystems corresponding to the application and system level resource management components respectively. The base level component implements the functionality of the MM application and models both MM data objects and their replicas (e.g. video and audio files), and MM requests to access this data via *sessions*. The corresponding base-level entities are *replica actors* and *request actors*.

The meta-level component deals with the coordination of multiple requests and sharing of existing resources among multiple requests. To provide coordination at the highest level and perform admission control for new incoming sessions, we introduce a meta-level entity, the *QoS Broker* meta-actor, QB. The two main functions of the QoS Broker are data management and request management. The multimedia data and request management functions of the QoS broker in turn require a number of services. We describe a representative sample below. The organization of meta-level services provided by a QoS broker is shown in the right hand side of Figure 2Each of these services in turn can be based on one or more of the core services - remote creation and distributed snapshot.

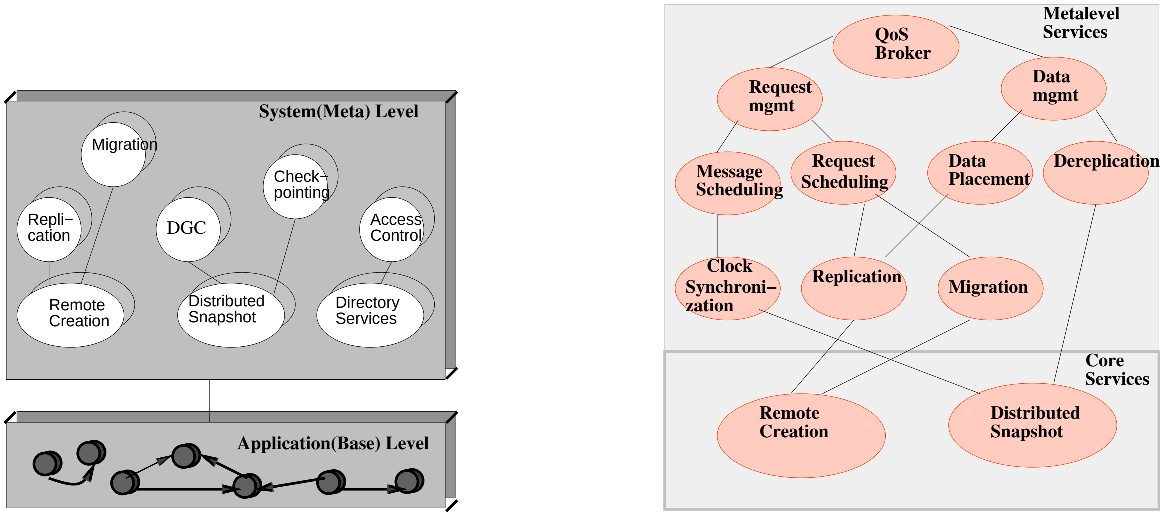

Fig. 2. (a) Classification of Core Services (b) Detailed architecture of the QoS meta-architecture system.

We now describe the main meta-level services and the supporting services in a little more detail: what they accomplish, why they are needed, and some of the issues that must be considered in establishing their correctness when operating in the context of the MM system, possibly concurrently with other services.

Data Management – The data management component decides on the placement of replica actors in the distributed MM system, i.e. it decides when and where to create additional replicas of data. It also determines when additional replicas are no longer needed and can be garbage collected/dereplicated. Data Placement can be performed adaptively or offline using a predefined placement strategy. The adaptive data placement policy dynamically initiates data placement when data or resources that are required to satisfy a request are not available. A predictive placement policy determines the degree of replication necessary for popular objects using a cost-based optimization procedure based on *a priori* predictions of expected subscriber requests.

Request management – Request arrival is unpredictable; often, it is not possible to know when requests will arrive and for how long they will last. This creates a need for adaptive admission control. The feasibility of QoS constraints for a MM session is determined during the *request scheduling* process when a new request arrives. For scheduling requests, we use an adaptive request scheduling policy that compares the relative utilization of resources in a MM server and assigns the request to a replica on a DS node. Once a request has been admitted, a *message scheduling* component assures the satisfaction of QoS constraints for service within the session.

Synchronization service: schedules message service to maintain a uniform notion of time among multiple nodes and thereby among actors in the session that are distributed across multiple nodes. The notion of synchronization can be built by using a uniform time value implemented as a snapshot broadcast phase that informs nodes about a (approximation of) current value of time. In [29], we specify the notion of a session in the context of MM actors and describe a construct, *QoSSynchronizer*, for expressing session specific time and QoS constraints.

Replication service: creates replicas to support availability, load-balancing and fault tolerance. When the existing copies of a MM object are incapable of supporting additional subscriber requests, replication of the MM object becomes necessary. At the time of replication it is necessary to choose a data source on which a new replica should be created. The rate at which replication proceeds also has a direct impact on system performance and application interactivity. Replication at a rate that is faster than real-time minimizes the time for replication, but consumes more resources.

Dereplication service: can dereplicate data. In many time periods, a MM object may drop in popularity. By tracking the demand for different MM objects and determining when copies of a MM object can be dereplicated, the load management procedure can optimize utilization of storage space on a MM server. The copy to be dereplicated must be carefully chosen based on current load on the different data sources, as well as the expected future demands for other MM objects currently being stored or that are expected to be stored on the MM server. Dereplication cannot occur instantly - the service must ensure that a copy that has been chosen for dereplication is removed only after all requests that are currently being serviced by that copy have completed.

Migration service: migrates data or requests to support load balancing, availability and locality. Often, instead of replicating an object to service a new request, we can migrate an existing request to another data source. Such *copy-free* migration is an attractive alternative for Internet Web servers and other services that handle non-real-time data. Often, this requires explicit tear-down and reestablishment of network connections as well as exchange of state information between data sources, both of which can cause significant playback jitter for real-time transmission.

In order to map the QoS meta-architecture to the physical system architecture, we distinguish between local and global components and define interactions between local resource managers on nodes and the global resource management component. The global component which includes the QoS broker and associated meta-actors resides on the distribution controller node. The node local components of the QoS broker metaarchitecture implementation include, for each DS node:

- a DS meta-actor for load-management on that node. The DS meta-actor contains state information regarding the current state of the node in terms of available resources, replicas, ongoing requests and replication processes etc.
- Request base actors corresponding to the requests assigned to that node.
- Replica base actors that correspond to the replicas of data objects currently available on that node.

3 Resource Management Policies for MM Servers

Apart from the QoS broker, the MM system contains a number of meta-actors whose behaviors combine to provide the resource management services discussed above. In this section, we describe the specific load management mechanisms that have been formalized in the meta-architecture. The mechanisms are implemented as meta-level actors and provide a modular and integrated approach to managing the individual resources of a MM server so as to effectively utilize all of the resources such as disks, CPU, memory and network resources. A MM request specifies a client, one or more multi-media objects, and a required QoS. The QoS requirement in turn is translated into resource allocation requirements. The ability of a data source to support additional requests is dependent not only on the resources that it has available, but also on the MM object requested and the characteristics of the request (e.g., playback rate, resolution). We characterize the degree of loading of a data source DS with respect to request R in terms of its load factor, $LF(R, DS)$, as: $LF(R, DS) = max(\frac{DB^R}{DB^{DS}}, \frac{M^R}{M^{DS}}, \frac{CPU^R}{CPU^{DS}}, \frac{NetBW^R}{NetBW^{DS}})$ where $DB^{R(DS)}$, $M^{R(DS)}$, $CPU^{R(DS)}$, and $NetBW^{R(DS)}$ denote the disk bandwidth, memory buffer space, CPU cycles, and network transfer bandwidth, respectively, that are necessary for supporting request R (available on data source DS). The load factor helps identify the critical resource in a data source, i.e., the resource that limits the capability of the data source to service additional requests. By comparing the load factor values for different servers, load management decisions can be taken by the QoS brokerage service. We briefly describe mechanisms for the scheduling of MM requests and placement of MM data. Optimizations to the basic mechanisms are outside the scope of this paper and are presented elsewhere [38, 35].

Scheduling of Multimedia Requests: The Request Scheduling meta-actor (RS) implements an adaptive scheduling policy that compares the relative utilization of resources at different data sources to generate an assignment of requests to replicas, so as to maximize the number of requests serviced. The data source that contains a copy of the MM object requested and which entails the least overhead (as determined by the computed load factor) is chosen as the candidate source for an incoming request. If no candidate data source can be found for servicing request R, then the RS meta-actor can either reject the incoming request or initiate *replication on demand* - implemented via a replication on demand meta-actor (ROD). In the former case, the RS meta-actor analyzes the rate of rejections over longer time frame and triggers appropriate placement policies to reduce the rate of rejection. With *replication on demand*, the RS

meta-actor decides to create a new replica of the MM object on one of the sources on the fly. The source DS on which the new replica is to be made can be determined to be the one that has minimum load-factor with respect to R, i.e., with the minimum value of $LF(R, DS)$. By doing so, the QoS broker attempts to maximize the possibility of servicing additional requests from the same replica. In order for this approach to be feasible and attractive, the replication must proceed at a very high rate, thereby consuming vital server resources for replication.

Placement of MM Objects: The predictive placement meta-actor(PP) implements a placement policy that determines in advance when, where, for how long, and how many replicas of each MM object should be placed in a MM server so as to maximize the number of requests that can be serviced by it. Predictive placement can be initiated periodically. The periodicity can be varied and the frequency with which the predictive placement process is executed governs its effectiveness. Smaller the time period, greater the computational overheads and greater the difficulty in predicting the request rates more accurately. Infrequent execution of predictive placement may result in under-utilization of the MM server's resources. The two main tasks of the predictive placement module are replication of MM objects from one data source to another, or from tertiary storage to secondary storage, and dereplication of MM objects from the data sources. The placement module must determine how many replicas of each MM object are necessary, and which of the data sources these replicas should be allocated to. The number of replicas possible for each MM object is dependent on the data source(s) to which the replicas are allocated, and upon the degree of loading of the data sources. In the process of determining the number and location of replicas that are necessary, the predictive placement module has to derive an allocation of MM objects to data sources, as well as a tentative assignment of expected requests to data sources; we term this the *pseudo-allocation* procedure. Placement mechanisms must be designed to work effectively with request scheduling. The goal of the pseudo-allocation procedure is to facilitate the task of the adaptive scheduler module, by allocating MM objects in such a way as to maximize system-wide revenue, by permitting a maximum number of requests to be admitted and scheduled for service. In order for predictive placement to execute concurrently with the adaptive scheduling process, a current snapshot of the system is taken prior to initiating predictive placement to provide a consistent view of the system state. The predictive placement module does not consider the exact times at which requests may arrive; the adaptive scheduling module makes assignment decisions based on the exact arrival times of requests.

4 Specifying and Reasoning about QoS-Based MM Services

Assuring safe composability of resource management services is essential for efficient management of distributed systems with widely varying and dynamically changing requirements. To analyze designs, clarify assumptions that must be met for correct operation, and establish criteria for non-interference, it is important to have a rigorous semantic model of the system: the resources, the management processes, the application activities, and the sharing and interactions among these.

In this section we briefly and mostly informally describe how to model the multimedia meta-architecture and resource management policies presented above in the TLAM framework. We state the key theorems relating QoS requirements to the QoS broker architecture and discuss reasoning about correctness and non-interference. More detail can be found in [35, 41]. Properties of a system modeled in the TLAM are specified as properties of computation paths. A property can be a simple invariant that must hold for all configurations of a path, a requirement that a configuration satisfying some condition eventually arise, or a requirement involving the transitions themselves. Such properties are checked using the properties of the building blocks for configurations – message contents and actor state descriptions – and of the TLAM reaction rules that determine the behavior of actors in the system.

To set the stage we informally describe the notion of a system providing QoS-based MM Service ($MMService_{QoS}$). We then map QoS requirements to resource requirements and focus on modeling and reasoning about the resource management underlying a QoS-based service. We define the notions of a *Resource-based MM Service* ($MMService$) and of a system having *Resource Based MM Behavior* ($MMBehavior$). $MMService$ reflects the chosen system resource architecture and allows us to reason abstractly about the availability and use of resources. $MMBehavior$ reflects the QoS broker software architecture and models the resource management algorithms as specific meta actor behaviors. Such a behavior specification can serve as a first stage in refining a service specification into an implementation. The main result is that a MM system meeting the Resource Based MM Behavior specification provides Resource-based MM Service, which in turn implies (under the assumptions on the mapping from QoS requirements to Resource requirements) that it provides QoS based MM Service. A consequence of this result is that for any service meta-actor, any algorithm that meets the behavior requirement can be used to implement that meta-actor.

We assume that there is a fixed set, $MMObjects$ of MM objects available in the system and let MM range over $MMObjects$. A MM request message specifies a triple $(\alpha_{cl}, \langle MM_0, \ldots, MM_n\rangle, QoS)$ which is interpreted as a request to initiate a MM streaming service from the server receiving the request, to the client actor α_{cl}, using the MM objects $\langle MM_0, \ldots, MM_n \rangle$, and obeying the QoS requirement QoS. To simplify the presentation, we shall assume that there is only one MM object involved in each request.

Definition 1 **(**$MMService_{QoS}$**)** A system S provides a QoS-based MM Service over the set of MM objects, $MMObjects$, iff for a configuration $C \in S$, if there is an undelivered message $MMreq$, then along any computation path π from C there are configurations such that one and only one of the following properties hold: (1) the request is accepted for service and service is provided with the required QoS until complete, or (2) the request is rejected, for this to happen it must be the case that the requested QoS cannot be provided at the time the request arrives.

4.1 Resource Based MM Service

We assume, given, a function $QoSTranslate$ which maps MM requests to resource requirements which, if met, will ensure the requested QoS (See [25] for examples of such QoS translation functions). The function $QoSTranslate$ maps MM requests to 4-tuples representing resource allocation requirements for the four managed resources: network bandwidth ($NetBW$), CPU cycles (CPU), disk bandwidth ($DiskBW$), and memory buffer ($BufMem$).

MM nodes (data source, distribution controller, and client) are modeled as TLAM nodes. We assume given a function $Capacity$ such that $Capacity(DS, Res) \in Unit_{Res}$ for any data source node DS and resource Res.

To provide an abstraction of the admission and placement processes, we introduce functions characterizing the state of replicas on each data source node and functions characterizing the state of each request that has arrived. For example, $ReplClass(C, DS, MM)$ is the replication class of the multimedia object MM on node DS. The con-

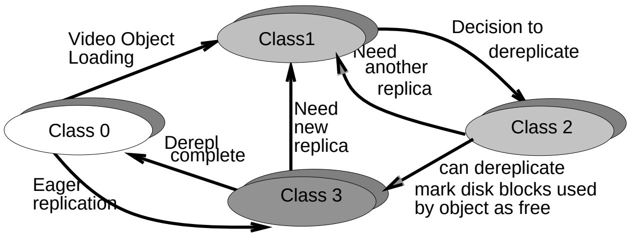

Fig. 3. State Transition Diagram. $Class_0$ indicates that the replica is not present on the node. A $Class_1$ replica is guaranteed to be available. A $Class_2$ replica is marked as dereplicable but remains available until all requests assigned to it have completed. A $Class_3$ replica exists on the node in the sense that it has not been overwritten, but there is no guarantee it will remain that way and can not be considered available until its class is changed. The transitions specify the possible phase changes in the life of a MM object.

straints on this function are given by the class transition diagram in Figure 3. Each request is uniquely associated to a base-level request actor. This relies on the uniqueness of messages and of newly created actors in the TLAM model. $ReqState(C, \alpha^{req})$ is the request status of request actor α^{req}. When initially created a request actor is in state $WaitForAdm$. As the system progresses, the value of this function will eventually change from $WaitForAdm$ to $admGranted$ or $admDenied$. From $admGranted$ it moves to $Servicing$ and then to $Completed$. After the value reaches $Completed$ or $admDenied$ it remains constant. The functions $ReqObjId(C, \alpha^{req})$ and $ReqReplica(C, \alpha^{req})$ identify the replica to which the request has been assigned (the MM object and the DS node containing the assigned replica). $ReqQoS(C, \alpha^{req})$ is the tuple of resources assigned to the request.

The *total resource property* $\phi_{res}^{total}(S)$ states that for every configuration in the system and every data source node the sum of the resources allocated to the requests on the node will not exceed the total capacity of resources on the node. Using the characterizing functions and the corresponding constraints on their values, we define a Resource-based MM service as follows.

Definition 2 **(**$MMService$**)** A system S providesResource-based MM service with respect to functions $QoSTranslate$, $Capacity$, and the functions characterizing replica and request state as specified above iff S obeys the constraints stated above for the replica and resource functions, $\phi_{res}^{total}(S)$ holds, and for configuration $C \in S$, if there is an undelivered message, $MMreq = (\alpha_{cl}, MM, QoS)$, then along any computation path π from C there is one of the following segments:

(D) $C_{\mathrm{start}} \xrightarrow{+} C_{\mathrm{deny}}$

(G) $C_{\text{start}} \xrightarrow{+} C_{\text{grant}} \xrightarrow{+} C_{\text{serve}} \xrightarrow{+} C_{\text{complete}}$

such that

1. C_{start} is the result of delivery of $MMreq$ and there is a new request actor α^{req} such that: $ReqClientId(C_{\text{start}}, \alpha^{req}) = \alpha_{cl}$, $ReqObjId(C_{\text{start}}, \alpha^{req}) = MM$, $ReqQoS(C_{\text{start}}, \alpha^{req})_{Res}$ is greater than $QoSTranslate(QoS)_{Res}$ for each Res, and $ReqState(C_{\text{start}}, \alpha^{req}) = WaitForAdm$.
2. $ReqState(C_{\text{deny}}, \alpha^{req}) = admDenied$, and there is a message to α_{cl} notifying of rejection of $MMreq$. In this case the system had insufficient resources to schedule $MMreq$ in C_{start}.
3. $ReqState(C_{\text{grant}}, \alpha^{req}) = admGranted$, and $ReqReplica(C_{\text{grant}}, \alpha^{req}) = DS$ for some DS node such that $ReplClass(C_{\text{grant}}, DS, MM) = Class_1$.
4. $ReqState(C_{\text{serve}}, \alpha^{req}) = Servicing$ and $ReqState(C_{\text{complete}}, \alpha^{req}) = Completed$ (allocated resources remain allocated during servicing).

Theorem 1 (QoS2Resource) If a system S provides $MMService$, and $QoSTranslate$ satisfies the stated requirements, then the system provides $MMService_{QoS}$.

4.2 QoS Broker Behavior

A system with MM behavior has meta actors: QB (a QoS Broker), RS (a Request Scheduler), ROD (Replication-On-Demand), DR (Dereplication), PP (Predictive Placement), located on the DC node, and a data source manager $DSma(DS)$ on each data source node DS. Information about resource capacity, replication, and request allocation is distributed on the data source nodes. The replica of an MM object MM on data source node DS, if present, is represented by a replica base actor $NodeReplica(C, DS, MM)$ and the request and replica state information is kept in annotations of the base actors state.

The QoS broker maintains, as part of its current state beh^{qb}, representing the behavior of the broker object and a model $MMstate(beh^{\text{qb}})$ of the distributed MM state which is used by the resource managers to make decisions. There are analogs of the configuration based functions characterizing request and replicas in which QoS broker state plays the role of configuration. To support more sophisticated resource management processes, the QoS broker maintains a current prediction/statistics model, $Pmodel(beh^{\text{qb}})$, and knowledge of what management processes are ongoing at any time.

The dynamic behavior of the QoS meta actors is given by TLAM rules specifying the effects of reaction to events and messages: updating the actors state, creating new actors, sending messages, modifying base-level annotations. There are QoS broker rules for receiving requests, invoking resource management processes (request scheduling, predictive placement, dereplication, ...), and updating the broker model in response to reports from these processes. In addition there are rules specifying the behavior of DS node meta actors that manage the replicas and local request actors.

All such broker processes are required to terminate and to accurately report effects to QB. DR can only change the class of replicas and can only change class 1 to class 2. ROD and PP can change class annotations and initiate replication (changing replication status and bandwidth). They can only change a replicas class from 0, 2, or 3 to 1. Processes that can change resource allocation are required to maintain the total resource invariant of the model they are provided with. A replication in progress must complete and we require that request servicing is a finite process.

We assume that only the MM meta actors set or modify the MM annotations of replica and request base actors. The QoS broker must coordinate the management activities sufficiently to assure that the model used is accurate. Thus we require that in any configuration, the QOS brokers model agree with the distributed state, modulo the effects of delivery of pending update notifications from the managers.

Theorem 2 If a system S has QoSBroker MM behavior as specified above, then S provides $MMService$.

The shared information used and modified by the resource managers— resources allocated to requests, replication status and replica class— is distributed on the DS nodes as base actor annotations. Interleaving of incremental modifications keeps the QoS brokers model sufficiently accurate to prevent interference among these processes. In fact $\phi_{res}^{total}(S)$ and the class transition constraints provide simple criteria for admitting other management processes in the interleaved setting.

Ensuring Safe Composability: A rigorous framework such as we have outlined is crucial in analyzing and preventing interference between concurrent activities. As an example of problems that arise, we might imagine that DR (dereplication) and RS (request scheduling) could be allowed to proceed concurrently. An attempt to establish that the system invariants are maintained reveals the following problem. Suppose MM on datasource DS is $Class_1$, RS

is invoked with request α^{req} requiring a copy of MM and concurrently DR is invoked. Then we have the following possible scenario: DR decides to mark MM on DS as $Class_2$ and RS decides to allocate MM on DS to α^{req}; the notification from DR reaches $DSma(DS)$ first and all requests previously allocated to MM complete so MM becomes $Class_3$; then the notification from RS arrives and now we have a request allocated to a $Class_3$ replica which can be overwritten. Another example of interference in this domain is the interaction between ROD (replication-on-demand) and PP (predictive placement) resulting in the allocation of multiple replicas of the same object on a single data source (causing potential conflicts in scheduling, access and deallocation). Currently, the QoS broker coordinates the safe interaction of these conflicting policies. Protocols and annotations that support increased concurrency of QoS-based resource management are the topic of current investigations.

5 Related Work and Future Research Directions

A preliminary implementation of the QoS broker and its associated components is discussed in [35]. Using the two level multimedia metaarchitecture described in this paper, a QoS-enabled customizable middleware framework, called $\mathbf{CompOSE|Q}$ is currently being developed at the University of California, Irvine [36]. $\mathbf{CompOSE|Q}$ includes modules that implement the 3 basic composable core services with specific interface definitions and interaction constraints built in, common services built out of the core services - object migration, distributed snapshot, directory services, scheduling etc. The $\mathbf{CompOSE|Q}$ environment consists of the following components: (1) a programming environment based on concurrent objects (2) a middleware library that manages execution of applications on nodes, coordinates distribution and provides services. (3) a compact runtime component that resides on the nodes of the distributed system which interacts with the distributed component of the middleware.

Commercially available object-based middleware infrastructures such as CORBA and DCOM represent a step toward compositional software architectures but do not support the development and maintenance of large applications. Specifically, they do not deal with interactions of multiple object services executing at the same time, or the implication of composing object services. For instance, the Electra framework [24] extends CORBA to provide support for fault tolerance using group-communication facilities and protocols like reliable multicast. Architectures that provide real-time extensions to CORBA [31, 44] necessary to support timing-based QoS requirements [47] have been proposed. TAO is a framework that supports real-time CORBA extensions to provide end-to-end QoS; it has been used to study performance optimizations [13], and patterns for extensible middleware [30]. Similarly, real-time method invocations have been explored by transmitting timing information in CORBA data structures [44].

The Java Development Environment is a distributed object implementation framework that provides mobility; however the semantics of interaction with other customizations is dependent on the implementation. The ability to deal with the management of thread priorities for real-time thread management is dependent on the underlying threads implementation, making QoS support complicated to achieve. Various systems such as the Infospheres Infrastructure [7] and the Globe System [34] explore the construction of large scale distributed systems using the paradigm of distributed objects. Various systems use resources on the Internet for wide-area parallel computing, e.g. Legion [14], Atlas [3]. Globus, a metacomputing framework for networked virtual machines defines a QoS component called Qualis [19] where low level QoS mechanisms can be integrated and tested. The composable framework presented in this paper is suitable for modeling and reasoning about activities in a metacomputing environment where computational nodes are in different administrative domains. *Reflection* allows application objects to customize the system behavior [32] as in Apertos [15] and 2K [18]. The Aspect Oriented Programming paradigm [17] makes it possible to express programs where design decisions can be appropriately isolated permitting composition and re-use. Some of the more recent research on actors has focused on coordination structures and meta-architectures [2] and runtime systems such as Broadway [33] and the Actor Foundry.

Multimedia QoS enforcement has been a topic of extensive research. Related work in this area includes projects such as Omega [27], *QualMan* [26] and several algorithms for MM server management [45, 42, 22, 16, 11, 43, 10]. *QualMan* [26] is a QoS aware resource management platform which contains a set of resource brokers that provides negotiation, admission and reservation capabilities for sharing end-system resources such as CPU, memory and network bandwidth.

Much of the work on formal models for QoS has been in the context of QoS specification mechanisms and constructs. In some implementation driven methods of QoS specification, the specification of QoS requirements is intermixed with the service specification [20, 21]. Other approaches address the representation of QoS via multiparadigm specification techniques that specify functional behavior and performance constraints distinctly using multiple languages [4, 46, 5, **?**]. Synchronizers and *rt* [12, 28] allow us to express QoS constraints via coordination constraints in the actor model either as local synchronization constraints or multi-actor coordination constraints. A number of language independent formalisms have been developed for specifying and reasoning about concurrent systems such as Unity [9] and I/O automata [23]. The TLAM approach models runtime services and the application using a single framework and uses the framework to reason about interactions between application actors, between metalevel services as well as ensure the correct behavior of base-meta interactions.

We are actively working on extending the existing meta-architecture to support more services. Modeling client interaction requires a notion of session and resources within a session. Further work is required to provide a generalized model that captures the architectural resources required in the server and network to support the session connection. *Dynamic negotiation protocols* involves negotiation of resources on the fly to degrade QoS of ongoing requests in order to admit more requests. In addition to providing a clear model of negotiation, this requires developing a mechanism to allow requests to decide when and how to renegotiate. For end-to-end QoS, it is necessary to determine how *real-time scheduling* strategies for time constrained task management interact with strategies for other tasks such as CPU intensive calculations, or network communication with clients. Also, further work is required in order to model the request migration service in the meta-architecture and develop strategies for its effective use.

In general, the dynamic nature of applications such as those of multimedia under varying network conditions, request traffic, etc. imply that resource management policies must be dynamic and customizable. Current mechanisms, which allow arbitrary objects to be plugged together, are not sufficient to capture the richness of interactions between resource managers and application components. For example, they do not allow customization of execution protocols for scheduling, replication, etc. This implies that the components must be redefined to incorporate the different protocols representing such interaction. We believe that a cleanly defined meta-architecture which supports customization and composition of protocols and services is needed to support the flexible use of component based software.

References

1. G. Agha. *Actors: A Model of Concurrent Computation in Distributed Systems*. MIT Press, Cambridge, Mass., 1986.
2. G. Agha, S. Frølund, W. Kim, R. Panwar, A. Patterson, and D. Sturman. Abstraction and Modularity Mechanisms for Concurrent Computing. *IEEE Parallel and Distributed Technology: Systems and Applications*, 1(2):3–14, May 1993.
3. J. E. Baldeschwieler, R. D. Blumofe, and E. A. Brewer. ATLAS: An Infrastructure for Global Computing. In *Proceedings of the The Seventh ACM SIGOPS European Workshop on Systems Support for Worldwide Applications*, 1996.
4. G. Blair, L. Blair, H. Bowman, and A. Chetwynd. *Formal Specifications of Distributed Multimedia Systems*. UCL Press, 1988.
5. L. Blair and G. Blair. Composition in multiparadigm specification techniques. In *ECOOP Workshop on Aspect Oriented Programming*, July 1999.
6. M. Buddhikot and G. Parulkar. Efficient data layout, scheduling and playout control in mars. In *Proceedings of NOSSDAV'95*, pages 339–351, April 1995.
7. K. Chandy, A. Rifkin, P. A. Sivilotti, J. Mandelson, M. Richardson, W. Tanaka, and L. Weisman. A world-wide distributed system using java and the internet. In *Proceedings of IEEE International Symposium on High Performance Distributed Computing (HPDC-5), Syracuse, New York*, Aug. 1996.
8. K. M. Chandy and L. Lamport. Distributed snapshots: determining global states of a distributed system. *ACM Transactions on Computing Systems*, 3(1):63–75, 1985.
9. K. M. Chandy and J. Misra. *Parallel Program Design*. Addison-Wesley, 1988.
10. A. Dan and D. Sitaram. An online video placement policy based on bandwidth to space ratio (bsr). In *SIGMOD '95*, pages 376–385, 1995.
11. A. Dan, D. Sitaram, and P. Shahabuddin. Dynamic batching policies for an on-demand video server. *ACM Multimedia Systems*, 4:112–121, 1996.
12. S. Frølund. *Coordinating Distributed Objects: An Actor-Based Approach to Synchronization*. MIT Press, 1996.
13. A. Gokhale and D. C. Schmidt. Evaluating the Performance of Demultiplexing Strategies for Real-time CORBA. In *Proceedings of GLOBECOM '97*, Phoenix, AZ, 1997.
14. A. Grimshaw, W. Wulf, and et al. The Legion Vision of a Worldwide Virtual Computer. *Communications of the ACM*, 40(1), January 1997.
15. J. ichiro Itoh, R. Lea, and Y. Yokote. Using meta-objects to support optimization in the Apertos operating system. In *USENIX COOTS (Conference on Object-Oriented Technologies*, June 1995.
16. K. Keeton and R. Katz. The evaluation of video layout strategies on a high-bandwidth file server. In *Proceedings of the Fourth International Workshop on Network and Operating System Support for Digital Audio and Video, Lancaster, UK*, pages 237–250, Nov. 1993.
17. G. Kiczales, J. Lamping, A. Mendhekar, C. Maeda, C. V. Lopes, J.-M. Loingtier, and J. Irwin. Aspect-Oriented Programming. In *Proceedings of ECOOP'97 European Conference on Object-Oriented Programming*, June 1997.

18. F. Kon, A. Singhai, R. H. Campbell, D. Carvalho, R. Moore, and F. J. Ballesteros. 2K: A Reflective, Component-Based Operating System for Rapidly Changing Environments . In *Proceedings of ECOOP'98 Workshop on Reflective Object-Oriented Programming and Systems*, Brussels, Belgium, July 1998.
19. C. Lee, C. Kesselman, J. Stepanek, R. Lindell, S. Hwang, B. S. Michel, J. Bannister, I. Foster, and A. Roy. The quality of service component for the globus metacomputing system. *Proceedings of IWQoS '98*, pages 140–142, 1998.
20. P. Leydekkers and V. Gay. Odp view on qos for open distributed mm environments. In J. d. Meer and A. Vogel, editors, *4th International IFIP Workshop on Quality of Service, IwQos96 Paris, France*, pages 45–55, Mar. 1996.
21. F. H. d. S. Lima and E. R. M. Madeira. Odp based qos specification for the multiware platform. In J. d. Meer and A. Vogel, editors, *4th International IFIP Workshop on Quality of Service, IwQos96 Paris, France*, pages 45–55, Mar. 1996.
22. P. Lougher and D. Shepherd. The design of a storage server for continuous media. *The Computer Journal - Special Issue on Distributed Multimedia Systems*, 36(1):32–42, February 1993.
23. N. A. Lynch and M. R. Tuttle. An introduction to input/output automata. Technical Report MIT/LCS/TM-373, MIT Laboratory for Computer Science, 1988.
24. S. Maffeis and D. Schmidt. Constructing reliable distributed communication systems with corba. *IEEE Communications*, 14(2), February 1997.
25. K. Nahrstedt. *Network Service Customization: End-Point Perspective*. PhD thesis, University of Pennsylvannia, 1995.
26. K. Nahrstedt, H.-H. Chu, and S. Narayan. Qos-aware resource management for distributed multimedia applications.
27. K. Nahrstedt and R. Steinmetz. Resource management in networked multimedia systems. *IEEE Computer*, 28(5):52–65, May 1995.
28. S. Ren, G. Agha, and M. Saito. A modular approach for programming distributed real-time systems. *Journal of Parallel and Distributed Computing*, 36(1), July 1996.
29. S. Ren, N. Venkatasubramanian, and G. Agha. Formalizing qos constraints using actors. In *Proceedings of Second IFIP International Conference on Formal Methods for Open Object Based Distributed Systems,FMOODS'97*, July 1997. Also post-conference book, Editors Bowman and Derrick, Published by Chapman and Hall, pp. 139-157.
30. D. C. Schmidt and C. Cleeland. Applying patterns to develop extensible and maintainable orb middleware. *Communications of the ACM*, 1998. (to appear).
31. D. C. Schmidt, D. Levine, and S. Mungee. The design of the tao real-time object request broker. *Computer Communications Special Issue on Building Quality of Service into Distributed System*, 1997.
32. B. C. Smith. *Reflection and Semantics in a Procedural Langauge*. PhD thesis, Massahusetts Institute of Technology, Jan. 1982.
33. D. Sturman. *Modular Specification of Interaction Policies in Distributed Computing*. PhD thesis, University of Illinois at Urbana-Champaign, May 1996. TR UIUCDCS-R-96-1950.
34. M. van Steen, A. Tanenbaum, I. Kuz, and H. Sip. A scalable middleware solution for advanced wide-area web services. In *Proc.Middleware '98, The Lake District, UK*, 1998.
35. N. Venkatasubramanian. *An Adaptive Resource Management Architecture for Global Distributed Computing*. PhD thesis, University of Illinois, Urbana-Champaign, 1998.
36. N. Venkatasubramanian. Compose—q - a qos-enabled customizable middleware framework for distributed computing. In *Proc. of the ICDCS Workshop on Distributed Middleware Austin, TX*, June 1999.
37. N. Venkatasubramanian, G. Agha, and C. L. Talcott. Scalable distributed garbage collection for systems of active objects. In *International Workshop on Memory Management, IWMM92, Saint-Malo*, LNCS, 1992.
38. N. Venkatasubramanian and S. Ramanathan. Effective load management for scalable video servers. In *Proceedings of the International Conference on Distributed Computing Systems (ICDCS97)*, May 1997.
39. N. Venkatasubramanian and C. Talcott. Composition of resource management activities in distributed systems. *Formal Aspects of Computing (to be submitted)*, 1998.
40. N. Venkatasubramanian and C. L. Talcott. Reasoning about Meta Level Activities in Open Distributed Systems. In *14th ACM Symposium on Principles of Distributed Computing*, pages 144–152, 1995.
41. N. Venkatasubramanian, C. L. Talcott, and G. Agha. Specifying Composable Services for QoS-Based Distributed Resource Management . In *Submitted for publication*, 1999.
42. H. M. Vin and P. V. Rangan. Designing a multi-user hdtv storage server. *IEEE Journal on Selected Areas in Communications*, 11(1):153–164, Jan. 1993.
43. J. L. Wolf, P. S. Yu, and H. Shachnai. Dasd dancing: A disk load balancing optimization scheme for video-on-demand computer systems. In *Proceedings of ACM SIGMETRICS '95, Performance Evaluation Review*, pages 157–166, May 1995.
44. V. F. Wolfe, J. K. Black, B. Thuraisingham, and P. Krupp. Real-time method invocations in distributed environments. In *Proceedings of the HiPC'95 Intl. Conference on High Performance COmputing*, 1995.
45. P. Yu, M. Chen, and D. Kandlur. Design and analysis of a grouped sweeping scheme for multimedia storage management. *Proceedings of Third International Workshop on Network and Operating System Support for Digital Audio and Video, San Diego*, pages 38–49, November 1992.
46. P. Zave and M. Jackson. Requirements for telecommunications services: An attack on complexity. In *IEEE International Symposium on Requirements Engineering*, 1997.
47. J. Zinky, D. Bakken, and R. Schantz. Architectural support of quality of service. *Theory and Practice of Object Systems*, 3(1), 1997.