CORBA Replication Support for Fault-Tolerance in a Partitionable Distributed System*

Stefan Beyer, Francesc D. Muñoz-Escoí and Pablo Galdámez
Instituto Tecnológico de Informática, Universidad Politécnica de Valencia
Camino de Vera, s/n , 46022 Valencia, Spain
{stefan, fmunyoz, pgaldamez}@iti.upv.es

Abstract

The Common Request Broker Architecture (CORBA) specification originally did not include any support for fault-tolerance. The Fault-Tolerant CORBA standard was added to address this issue. One drawback of the standard is that it does not include fault-tolerance in the case of network partitioning faults. However, wide area networks, over which distributed systems are often employed, are especially susceptible to network partitioning.

The main contribution of this paper is the design of a fault-tolerance CORBA add-on for partitionable environments. In contrast to other solutions, our modular design separates replication and reconciliation policies from the basic replication mechanisms. This modularity allows the replication and reconciliation strategies to be modified easily.

1 Introduction

The Common Request Broker Architecture (CORBA) [20] is a popular middleware framework to construct distributed object systems. As distributed systems are subject to host and network failures, fault-tolerance is an important aspect in the design of such systems. However, the CORBA specification did originally not include any support for fault-tolerance. Since then, the Fault-Tolerant CORBA specification (FT-CORBA) [22] has been added to introduce a degree of fault-tolerance to CORBA. However, the standard has various drawbacks. One important shortcoming of FT-CORBA is that it does not provide support for fault-tolerance in a partitioned network. Wide area networks, over which distributed systems are often employed, are especially susceptible to network partitioning.

*This work has been funded by the European Community under the FP6 IST project DeDiSys (Dependable Distributed Systems, contract number 004152).

In this paper we present the architecture of a middleware add-on that adds fault-tolerance to CORBA in a partitioned environment by means of replication. The system is part of the DeDiSys project [23]. DeDiSys aims at providing fault-tolerance through add-ons for various middlewares. The CORBA add-on presented here uses CORBA Portable Interceptors [21] to intercept calls to server objects in a transparent manner and divert these calls through a replication manager. An underlying group membership and communication service provides reliable communication.

In contrast to other systems, the modular design of the DeDiSys replication support allows different replication and reconciliation policies to be implemented easily. The design of the replication support is based on a separation of mechanism and policy. Replication mechanisms are basic primitives such as creating a replica and changing its role, or the ability of managing object and replica references. In our system these mechanisms are provided by a distributed replication manager. Many replication protocols will have these mechanisms in common. In contrast, replication and reconciliation policies, such as the update propagation policy or reconciliation strategy, may vary between replication protocols. We extract such policy from the replication manager and place it into a replication protocol component. The replication manager and the replication protocol components provide fixed interfaces. New replication protocols can be implemented by replacing the replication protocol component.

A default replication protocol [3] is included. The protocol allows operations in each partition in a partitioned system to continue. Resulting conflicts can be resolved automatically by the reconciliation support or manually by the application.

A non-CORBA prototype [4] of our architecture has been implemented and we are currently in the process of implementing the full CORBA system, taking into account the lessons learnt from the prototype.

2 Related Work

In order to add fault-tolerance to CORBA, certain mechanisms, such as replication, are required. Existing systems either implement the FT-CORBA [22] standard to provide fault-tolerance or suggest their own fault tolerance extensions. There are two possible reasons for a system not to comply with the FT-CORBA standard. Some systems reviewed here were simply developed before the standard was defined. Other systems try to overcome some of the drawbacks associated with FT-CORBA. As DeDiSys is a research project, aimed at partitionable distributed systems, which are not covered by the FT-CORBA standard, we do not consider FT-CORBA compliance as the main factor for this review.

In literature, approaches to add fault-tolerance mechanisms to CORBA are typically classified into three categories: In the **integration approach**, the ORB itself is modified to include the required fault tolerance mechanisms. It is easy to provide transparency using this approach, but existing commercial ORBs cannot be used. Orbix+Isis [11], Electra [13] and Maestro [28] are examples of systems using the integration approach. More recently, the authors of [14] and [30] have proposed the integration of group communication support by modifying the CORBA Open Communication Interface (OCI) and using the Pluggable Protocols Framework [12]respectively.

In the **service approach**, the mechanisms required to provide fault tolerance are provided as CORBA services. This approach has the advantage that existing ORBs can be used. However, transparency is difficult to achieve with this approach, as applications have to be aware of the fault tolerance services. Object Group Services (OGS) [7] and Newtop Object Group Service [15] provide services for object group support which can be used to provide fault-tolerance. FTS [8], OPEN EDEN [9], IRL [2] and AQuA [29] are examples of reliable CORBA systems using the service approach, although it can be argued that these systems also use elements of the interceptor approach.

In the **interceptor approach**, CORBA invocations are intercepted and redirected to fault tolerance mechanisms. Recent systems make use of CORBA Portable Interceptors [21]. The only systems using a pure interception approach we are aware of are Eternal [18] [19] and DAISY [26].

Three of the systems mentioned above - Maestro, FTS and Eternal - provide some support for network partitioning. Therefore, these systems are reviewed here in more detail. Newtop also provides support for network partitioning, but, as a mere object group toolkit, does not provide any support for reconciling replica state after partitioning. Therefore, we do not discuss Newtop in detail here.

Maestro uses the integration approach. The system was developed before the FT-CORBA specification existed. It is not a pure CORBA implementation, but was designed as a distributed object layer to be used on its own or to be integrated in CORBA or in other distributed object technologies. The system uses Ensemble [27] as an underlying group communication and membership toolkit. Partitioning is supported using a variation of the primary partition model [24]. Only updates in one partition are permanently accepted, but in contrast to the regular primary partition model, the decision on which partition dominates is postponed until recovery time. At recovery time the partition with "the most updated" state is chosen.

FTS is an attempt to remain close to the FT-CORBA specification, whilst also providing support for partitioning. The system uses a mixture of the service and interceptor approaches. A group object adapter (GOA) is provided as a CORBA object adapter. The GOA is implemented on top of the portable object adapter (POA) to allow for object groups; that is, groups of replicas representing the same logical object. The main drawback of FTS is that it only implements active replication, although the authors claim it would be easy to adapt FTS to passive replication. In DeDiSys we use also use the idea of an object adapter providing object group support. FTS uses the primary partition model for consistency in case of network partitioning.

Eternal is probably the most advanced of the systems of which we are aware in terms of support for partitioning, despite being one of the oldest systems. The system allows for active and passive replication. The Eternal replication manager keeps track of replicated objects. CORBA messages are intercepted at the transport level and are redirected using the Totem group communication toolkit [17]. Totem provides Eternal with the extended virtual synchrony model, which allows for network partitioning. As far as we know, Eternal is unique in partition-aware CORBA systems, in that it does not use a variant of the primary partition model, but does allow operations in all partitions to continue. A simple reconciliation algorithm is provided. When the network partitions, a primary subgroup is chosen for each object. However, operations are also allowed to continue in secondary subgroups. When subgroups are re-merged, Eternal gives preference to the state contained in the primary subgroup. However, operations in secondary subgroups are queued and applied after the state of the primary subgroup has been installed in all the merging subgroups during recovery. Conflicts that cannot be resolved are reported to the application.

In DeDiSys we make use of some techniques from Eternal, DAISY and FTS. In particular, we use interception, as in Eternal and DAISY, and the implementation of the replication support as a CORBA object adapter, as in FTS. In contrast to Eternal's interception at the operating system level approach we use DAISY's approach of using portable interceptors, which were not available when Eternal was de-

signed. Furthermore, in DeDisys we aim at making replication and recovery flexible and configurable. To this end we do not embed replication protocol and reconciliation policy in the replication manager, as done in Eternal, but provide an easily interchangeable replication and reconciliation protocol component.

3 Design Principles

In order to design a replication support for partionable environments in CORBA we have followed the following design principles:

Separation of Mechanism and Policy. Replication mechanisms are basic primitives, such as the ability to create a replica or manage the relation between object references and replica references. The provided mechanisms can be used in different ways to implement replication policies, such as the object state transfer policy or the reconciliation strategy. Policies may vary, whereas mechanisms are provided to support different policies. In conventional systems policies and mechanism are often embedded in the same component. This makes it difficult to implement different policies. In DeDiSys, we extract replication and reconciliation policy from the main replication component, which only provides mechanisms that allow to implement a variety of policies.

Interception The DeDiSys concept is to provide a Middleware add-on rather than modifying existing middleware. To achieve this in CORBA, we intercept object invocations. To this end, CORBA portable interceptors are used to pass control to the replication support.

Client-Side Transparency Replication should be transparent to the client application. That is, the client is not aware it is dealing with a replicated object and existing CORBA clients do not have to be modified to use DeDiSys, apart from calling a simple initialisation routine.

Server-Side Transparency It is our goal to make server side integration of the replication support as transparent as possible. However, the server application should have some control over the replication support. Therefore, a simple interface provides mechanisms, such as replica creation, and has to be used by the server application. It is our goal to make CORBA server applications as easy to port to DeDiSys as possible, whilst allowing configurability of key parameters, such as number and location of replicas.

4 The DeDiSys Replication Model

DeDiSys aims to introduce fault-tolerance through replication. In this section we describe the failure model we support and the replication model used to achieve this.

The "crash model" [6] is assumed for node failures, and the "link failure model" [25] for communication services. As we cannot distinguish between a failed node and an isolated node until recovery time, we treat every failure as partitioning. Partitions can occur in any number and order. Recovery of partitioning can be in a different order in which the partitioning originally occurred.

In order to provide support for partitioning, DeDiSys uses the Spread group communication and membership toolkit [1]. Spread provides the extended virtual synchrony model [16]. This model simplifies the reconciliation process of potential replication protocols, as nodes are aware which views have been installed in re-joining partitions.

We employ the *passive replication* model. In passive replication [5] [10] requests are only processed by one *primary copy*. Updates are then propagated to the *secondary copies*. The passive model lends itself to a system where consistency is to be configured as it allows variations in the way updates are propagated. If *synchronous* update propagation is used, a primary copy must propagate any updates immediately; that is, before the result of the operation that has caused the update is returned to the client. In *asynchronous* update propagation the result is returned and the propagation of state changes performed some time later. We leave the choice of which update propagation paradigm to use to the replication protocol.

The default replication protocol [3] allows operations in all partitions to continue. Object state updates are propagated synchronously in a healthy system and asynchronously during partitioning. If a primary copy of an object is not reachable, the protocol promotes a secondary copy to a temporary primary copy. The protocol therefore implements a "primary per partition model". The protocol also includes a reconciliation protocol that restores consistency when partitions are merged. Conflicts that occur when replicas of the same object are written to in different partitions can be resolved automatically by the replication protocol or manually by the application. However, the design of our system is such that many replication and reconciliation protocols based on the passive replication model can be implemented.

5 Replication Support Integration in CORBA

Figure 1 shows how the DeDiSys replication support is integrated in CORBA. Portable interceptors are used to transfer control to the replication support, without client code having to be modified. The client invokes an object in the standard CORBA way, using a logical object reference. The DeDiSys replication support takes care of identifying the real object reference of the primary replica. The **client-side request interceptor** is used to intercept object invocations, before they are sent. This interceptor uses the replication manager (RM) to obtain the reference of the primary

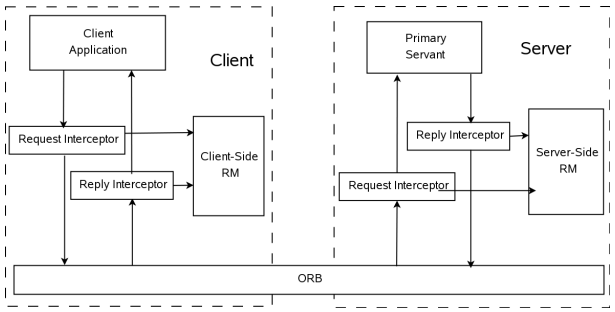

Figure 1. Replication Support Overview

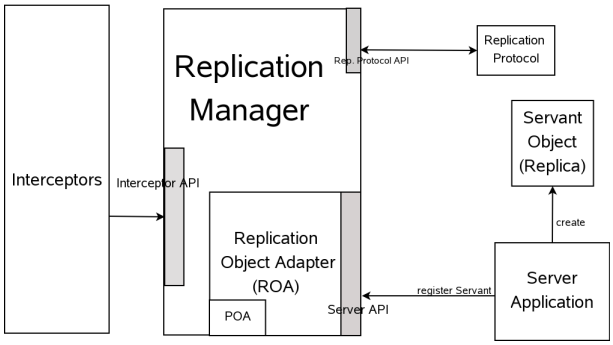

Figure 2. Replication Manager Overview

replica and redirects the invocation to this primary replica. The replication manager is also used to trigger some replication protocol specific tasks that might need to be executed before the invocation can begin.

On the server side, the **server-side request interceptor** also intercepts the incoming request, in order to trigger replication protocol specific tasks. The object invocation is then executed in the standard CORBA way. Before the result is returned to the client, control is again passed to the RM. At this stage the replication protocol might require changes in the accessed object's state to be propagated to the secondary replicas of the object.

Before the request is delivered to the client application the reply is again intercepted on the client side by the **client-side reply interceptor**. At this stage a replication protocol might trigger consistency checks that could cause the invocation to be undone.

6 Object Reference Management

We distinguish between **logical object references** and **replica references**. When using the term logical object reference, we are referring to the reference of a logical object. When using the term replica reference we are referring to the actual reference of an object replica; that is, a reference of a real CORBA implementation of a logical object.

Both types of references are standard CORBA object references. However, internally logical object references only refer to an intermediate "dummy" object which is never invoked. The replication system intercepts calls to these objects and redirects them using the actual replica reference. Only logical object references are visible to client and server applications.

The replication support keeps track of which logical object references are associated with which replica references.

7 Replication Manager

Figure 2 shows the replication manager (RM). The RM consists of various components. Only the **Replication Object Adapter (ROA)** is visible to the server application. The ROA is a CORBA object adapter. It internally uses CORBA's Portable Object Adapter (POA) and provides standard POA functionality, such as associating objects with object references. In addition, it manages object replicas and allows replicas to be created and associated with a logical object. The ROA provides the standard POA API and a small set of additional methods needed to support replication. Client-side-only RMs do not need a ROA.

The RM also interacts with a **replication protocol** component, in which replication protocol details, such as update transfer policies, are implemented. By encapsulating such policy in a separate component with a defined interface, the replication protocol can be changed easily.

The RM also provides an interface to the DeDiSys interceptors, in order to pass control to the replication support.

Furthermore, the RM is an "application" of the Spread group membership and communication service. RMs on different nodes use Spread to exchange information on new replicas or to broadcast replica role changes. Furthermore, the RM keeps track of which replicas are reachable. Therefore Spread callbacks handling the reception of group messages and new membership views need to be implemented in the RM.

8 The Replication Protocol Component

The replication protocol (RP) component encapsulates replication and reconciliation policies. We locate replication policy and reconciliation policy in the same component, as the policies have to match each other. For instance, a replication protocol that allows updates in each of the partitions of a partitioned system requires a reconciliation pol-

icy that allows the system to recover from the inconsistencies this might introduce.

The RM passes control to the RP before and after every object invocation, in order to allow the RP to allow or deny certain object invocations to maintain consistency and to keep track of changes to objects and maintain internal data structures that hold information necessary for reconciliation. The activities of the RP in a healthy system vary from that in a system in which one or more nodes are not reachable, as different data-structures have to be maintained in these different system modes. Furthermore, the RP implements update propagation and reconciliation. Finally, the RP implements Spread callbacks to receive replica update messages and any other messages necessary for synchronising RPs on different nodes.

Different RPs can be implemented by modifying the RP component. To this end, the RP component consists of an abstract `ReplicationProtocol` class, which should be extended, in order to implement a replication protocol. `ReplicationProtocol` also provides default implementations of some methods, that may or may not be overwritten by a particular replication protocol. A default update propagation method is provided. The method can be called by any subclass implementing a specific replication protocol to broadcast the state of a specific primary replica to all secondary copies. Furthermore, a default method handling incoming replica updates is provided. This method just sets the state of all the secondary copies it holds of a particular primary copy to that included in the message. Both the update propagator and the incoming message handler can be overwritten by protocols that require more specialised implementations.

9 Conclusion and Future Work

In this paper we have described the design of our fault-tolerance support for CORBA. In contrast to most approaches to fault-tolerance in CORBA and the Fault-Tolerance CORBA specification [22], the system can cope with network partitioning. The system forms part of the DeDiSys project [23], which aims at providing fault-tolerance add-ons for a variety of middlewares.

We have implemented our design in our own non-CORBA evaluation environment. The DeDiSys Lite platform [4] serves as both a first prototype implementation of DeDiSys and an evaluation platform for replication protocols. However, it does not make use of CORBA.

We are currently implementing the architecture described here in CORBA using Java as an implementation language, taking into account the experiences gained with our non-CORBA prototype. After evaluating our implementation, the results obtained will be compared with implementations of the DeDiSys approach in other middleware architectures which are currently being developed in parallel by our project partners.

Furthermore, we are planning to extend our modular design to allow replication protocols using models other than passive replication to be implemented.

References

[1] Y. Amir, C. Danilov, and J. R. Stanton. A low latency, loss tolerant architecture and protocol for wide area group communication. In *International Conference on Dependable Systems and Networks*, pages 327–336, 2000.

[2] R. Baldoni and C. Marchetti. Three-tier replication for ft-corba infrastructures. *Softw. Pract. Exper.*, 33(8):767–797, 2003.

[3] S. Beyer, M. Bañuls, P. Galdámez, and F. D. Muñoz-Escoí. Increasing availability in a replicated partionable distributed object system. Technical Report ITI-ITE-05/10, Instituto Tecnológico de Informática, 2005.

[4] S. Beyer, A. Sánchez, F. D. Muñoz-Escoí, and P. Galdámez. Dedisys lite: An environment for evaluating replication protocols in partitionable distributed object systems. In *International Conference on Availability, Reliability and Security*, 2006.

[5] N. Budhiraja, K. Marzullo, F. B. Schneider, and S. Toueg. *The primary-backup approach*, pages 199–216. ACM Press, Addison-Wesley, 1993.

[6] F. Cristian. Understanding fault-tolerant distributed systems. *Commun. ACM*, 34(2):56–78, 1991.

[7] P. Felber, B. Garbinato, and R. Guerraoui. The design of a corba group communication service. In *Symposium on Reliable Distributed Systems*, page 150, 1996.

[8] R. Friedman and E. Hadad. Fts: A high-performance corba fault-tolerance service. In *IEEE International Workshop on Object-Oriented Real-Time Dependable Systems*, pages 61–68, 2002.

[9] F. Greve, M. Hurfin, and J.-P. L. Narzul. Open eden: a portable fault tolerant corba architecture. In *International Symposium on Parallel and Distributed Computing*, pages 88–95, 2003.

[10] R. Guerraoui and A. Schiper. Software-based replication for fault tolerance. *Computer*, 30(4):68–74, 1997.

[11] IONA and Isis. An Introduction to Orbix+Isis, IONA Technologies Ltd. and Isis Distributed Sytems Inc., 1994.

[12] F. Kuhns, C. O'Ryan, D. C. Schmidt, O. Othman, and J. Parsons. The design and performance of a pluggable protocols framework for corba middleware. In *IEEE ComSoc TC on Gigabit Networking Sixth International Workshop on Protocols for High Speed Networks VI*, pages 81–98, 1999.

[13] S. Landis and S. Maffeis. Building reliable distributed systems with CORBA. *Theory and Practice of Object Systems*, 3(1):31–43, 1997.

[14] D. Lee, D. Nam, H. Y. Youn, and C. Yu. Oci-based group communication support in corba. *IEEE Transactions on Parallel and Distributed Systems*, 14(11):1126–1139, november 2003.

[15] G. Morgan, S. K. Shrivastava, P. Ezhilchelvan, and M. Little. Design and implementation of a corba fault-tolerant object group service. In *International Working Conference on Distributed Applications and Interoperable Systems*, June 1999.

[16] L. E. Moser, Y. Amir, P. M. Melliar-Smith, and D. A. Agarwal. Extended virtual synchrony. In *The 14th IEEE International Conference on Distributed Computing Systems (ICDCS)*, pages 56–65, 1994.

[17] L. E. Moser, P. M. Melliar-Smith, D. A. Agarwal, R. K. Budhia, and C. A. Lingley-Papadopoulos. Totem: A fault-tolerant multicast group communication system. *Communications of the ACM*, 39(4):54–63, 1996.

[18] L. E. Moser, P. M. Melliar-Smith, and P. Narasimhan. Consistent object replication in the eternal system. *Theor. Pract. Object Syst.*, 4(2):81–92, 1998.

[19] P. Narasimhan, L. E. Moser, and P. M. Melliar.Smith. Replica consistency of corba objects in partitionable distributed systems. *Distributed System Engeneering*, 4:139–150, 1997.

[20] Object Management Group. The common object request broker architecture (corba) v.3.0.3, March 2004.

[21] Object Management Group. The common object request broker architecture (corba) v.3.0.3. chapter 11. portable interceptors, March 2004.

[22] Object Management Group. The common object request broker architecture (corba) v.3.0.3. chapter 23. fault tolerant corba, March 2004.

[23] J. Osrael, L. Froihofer, K. M. Goeschka, S. Beyer, F. D. Muñoz-Escoí, and P. Galdámez. A system architecture for enhanced availability of tightly coupled distributed systems. In *International Conference on Availability, Reliability and Security*, 2006.

[24] A. Ricciardi, A. Schiper, and K. Birman. Understanding partitions and the "non partition" assumption. In *Workshop on Future Trends of Distributed Systems*, 1993.

[25] F. B. Schneider. What good are models and what models are good? In *Distributed Systems*, chapter 2, pages 17–26. ACM Press, Addison-Wesley, 2nd edition, 1993.

[26] Taha Bennani et al. Implementing simple replication protocols using corba portable interceptors and java serialization. In *International Conference on Dependable Systems and Networks*, pages 549–554, 2004.

[27] R. van Renesse, K. Birman, M. Hayden, A. Vaysburd, and D. Karr. Building adaptive systems using ensemble. *Softw. Pract. Exper.*, 28(9):963–979, 1998.

[28] A. Vaysburd and K. Birman. The maestro approach to building reliable interoperable distributed applications with multiple execution styles. *Theor. Pract. Object Syst.*, 4(2):71–80, 1998.

[29] Yansong (Jennifer) Ren et al. Aqua: An adaptive architecture that provides dependable distributed objects. *IEEE Trans. Comput.*, 52(1):31–50, 2003.

[30] W. Zhao, L. E. Moser, and P. M. Melliar-Smith. Design and implementation of a pluggable fault-tolerant corba infrastructure. *Cluster Computing*, 7(4):317–330, 2004.