2K: A Distributed Operating System for Dynamic Heterogeneous Environments*

Fabio Kon Roy H. Campbell M. Dennis Mickunas Klara Nahrstedt
Department of Computer Science
University of Illinois at Urbana-Champaign
{f-kon,rhc,mickunas,klara}@cs.uiuc.edu

Francisco J. Ballesteros
Rey Juan Carlos University, Madrid
nemo@gsyc.escet.urjc.es

http://choices.cs.uiuc.edu/2K

Abstract

The first decades of the new millennium will witness an explosive growth in the number and diversity of networked devices and portals. We foresee high degrees of mobility, heterogeneity, and interactions among computing devices connected to global networks. While previous research in distributed operating systems solved many problems related to resource management, they seldom addressed the problems of heterogeneity and dynamic adaptability. On the other hand, middleware solutions, like CORBA and Jini, solve part of the heterogeneity problem by permitting seamless communication among different platforms. But, they do not address dynamic resource management and adaptability for applications requiring high-performance distributed computing.

This paper presents *2K*, an integrated operating system architecture that addresses the problems of resource management in heterogeneous networks, dynamic adaptability, and configuration of component-based distributed applications.

1 Introduction

Modern computing environments are characterized by a high level of dynamism. Two major kinds of dynamic changes occur frequently. The first refers to structural changes such as hardware and software upgrades, protocol and API updates, and operating system patches. The second refers to dynamic changes in the availability of memory, CPU, network bandwidth and, in mobile systems, connectivity and location. Drastic changes may occur in a few seconds, impacting the performance of user applications profoundly. Existing operating systems offer very little support for managing, adapting, and reacting to these changes; all the work is left to the applications or to users and system administrators who must take care of them manually. Since large corporate and academic networks tend to be heterogeneous, the configuration work is multiplied by the number of supported platforms.

Thus, we need a very flexible and adaptable architecture that permits the dynamic instantiation of customized user environments at different locations in the distributed system. Existing system architectures do not provide proper management of the dependencies among system and application components, which makes it difficult to support automatic configuration in a reliable way. It is hard to create robust and efficient systems if the dynamic dependencies between components are not well understood. For those reasons, proper dependence management is a major requirement for the next generation middleware and operating systems.

*This research is supported by the National Science Foundation, grants 98-70736, 99-70139, and CCR 96-23867.

A proper management of dynamism is crucial for distributed and parallel applications which suffer from changes in the environment. In many cases, the performance degradation is so high that the distribution overhead becomes larger than the speedup obtained with the additional parallelism. In this paper, we present an overview of *2K*, a novel network-centric operating system designed to solve this problem.

2 Resource Management in Heterogeneous Environments

The basic task of both centralized and distributed operating systems is to manage the resources of a machine (or a collection of machines) and safely export them to their users. Conventional operating systems, however, are not able to manage the resources of collections of heterogeneous machines.

CORBA and Jini emerge as powerful technologies for interoperability in heterogeneous environments. But they both lack the notion of a *user* and do not provide support for dynamic resource management either in a single machine or in a distributed environment.

Our approach combines the benefits of CORBA with those of distributed operating systems. It provides management of distributed resources while being able to handle different hardware platforms and different underlying, single-node operating systems.

The *2K* system supports a completely object-oriented view of the distributed computing environment; distributed hardware and software resources are encapsulated as CORBA objects while distributed operating system services (e.g. file, naming, and execution services) are exported as CORBA services. Applications run within this relatively homogeneous environment built on top of highly heterogeneous distributed environments.

To achieve optimal application performance in a dynamic environment of distributed resources, the middleware must be configurable and able to adapt to dynamic changes in resource availability and in the software and hardware infrastructures. *2K* uses a dynamically configurable reflective ORB [RKC99] to provide a high-level of flexibility to applications that can benefit from it by tuning the CORBA implementation to their specific needs. But it also keeps the complexity away from applications that prefer to use the CORBA distributed object model without worrying about the underlying details.

The reflective ORB solves the problem of *how* to adapt the system to the application needs but it does not address the problem of *when* and *what* to adapt. We do that by maintaining an explicit representation of the dynamic dependencies among system and application components and by allowing the inspection and monitoring of the dynamic state of system components. *2K* enhances resource management with algorithms for QoS provision including admission control, negotiation, reservation, and renegotiation. Application programmers have then complete access to the system's dynamic state and are able to implement application-specific adaptations while the system guarantees that QoS is preserved.

3 *2K* System Model

2K adopts a *network-centric* model in which all entities, users, software components, and devices exist in the network and are represented as CORBA objects. Each entity has a network-wide identity, a network-wide profile, and dependencies upon other network entities. When a particular service is instantiated, the entities that constitute that service are assembled.

In contrast to existing systems where a large number of non-utilized modules are carried along with the basic system installation, our philosophy is based upon a "What You Need Is What You Get" (WYNIWYG) model. The system configures itself automatically and loads the minimum set of components required for executing the user applications in the most efficient way.

As shown in figure 1, this philosophy is realized by leveraging standard CORBA services such as Naming, Trading, Security, and Persistence and extending the CORBA model with the addition of services for QoS-Aware Resource Management, Automatic Configuration, and Code Distribution.

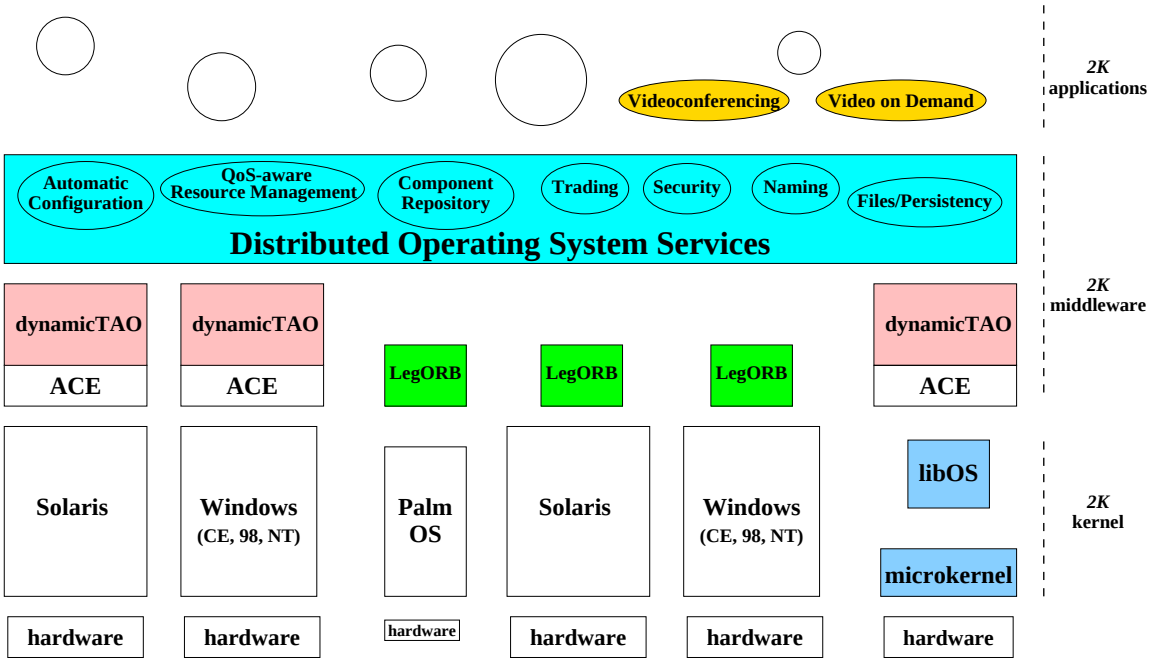

Figure 1: The *2K* overall Architecture

3.1 Automatic Configuration Service

To address the problems described in the previous sections, the *2K Automatic Configuration Service* manages two distinct kinds of dependencies:

1. *prerequisites*, i.e., the requirements for loading an inert component into the runtime system, and
2. *dynamic dependencies* among loaded components in a running system.

As long as the system has access to the requirements for installing and running a software component, the installation and configuration of new components can be automated. As a byproduct of this knowledge, component performance can be improved by analyzing the dynamic state of system resources, by analyzing the characteristics of each component, and by configuring them in the most efficient way.

3.1.1 Prerequisites

The prerequisites for a particular inert component specify any special requirement for properly loading, configuring, and executing that component. Prerequisites specify 1) the type and share of hardware resources that a component needs and 2) the software services (i.e. other components) it requires.

The first kind of prerequisites lets the QoS-aware Resource Management Service determine where, how, and when to execute each component. It uses this data to enable proper admission control, resource negotiation, reservation, and scheduling.

The second kind of prerequisites determines which auxiliary components must be loaded and which other software services must be located. As the Automatic Configuration Service parses the software prerequisite specifications, it verifies whether it is necessary to create new instances of the required components in the *2K* runtime. If necessary, it contacts the Component Repository, fetches the component binary code compiled for that specific platform and dynamically loads it.

As of now, the prerequisite specifications are created manually by component developers. In the future, we expect that this task will also be automated.

3.1.2 Dynamic Dependencies

While the Automatic Configuration Service parses the prerequisite specifications, fetches the required components from the Component Repository, and dynamically loads their code into the runtime, it uses the information in the prerequisite specifications to create a runtime representation of inter-component dependencies. This representation uses CORBA objects called *ComponentConfigurator*s (see figure 2). These objects store the dependencies as lists of CORBA Interoperable Object References (IORs), pointing to other component configurators, forming a dependence graph of distributed components.

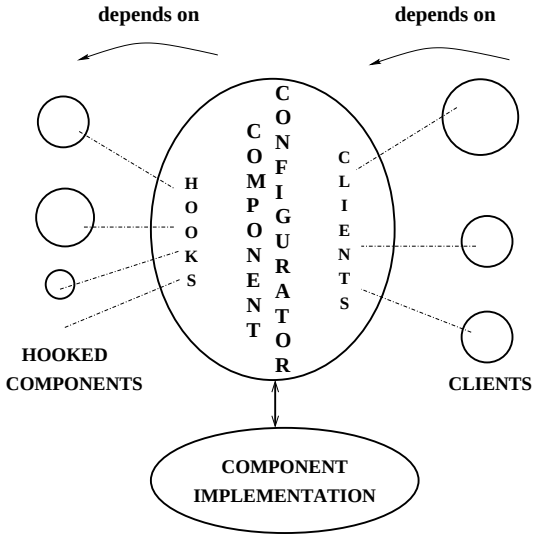

Figure 2: Reification of component dependence

With information about its runtime dependencies, applications can refer to its own requirements, selecting different components to fulfill their needs in different environments and at different times. In addition, the underlying system can manipulate the application dependencies in order to optimize performance or to adapt to dynamic changes in the environment.

When a *2K* component fails, the system inspects its dependencies and informs the proper components about the failure. Applications can customize the system by implementing specialized instances of the *ComponentConfigurator*, for example, to recover from a failures by replacing the faulty component with a new one. The same mechanism can be used for adapting the system and its components to changing parameters such as network bandwidth, CPU load, resource availability, user access patterns, and the like.

3.2 Mobile Configuration Agents

In addition to the *pull*-based approach for code distribution described in 3.1.1 where the system fetches components from the Component Repository, *2K* also supports a *push* mechanism based on mobile agents. In many cases, this alternate mechanism can improve system performance by distributing code updates in a scalable way.

The reflective ORBs are organized as a distribution network through which system administrators or applications can send *configuration* and *inspection* agents [KGCM99]. Agents may contain both configuration commands (to change the configuration of the ORBs and their applications) and new implementation for system and application components (in the form of dynamically loadable libraries or Java bytecode).

The combination of these mechanisms provides a flexible infrastructure for dynamic software updates. By working on an environment that requires less manual administration, users and developers can concentrate on more important tasks and improve their productivity.

3.3 QoS-Aware Distributed Resource Management

The *2K* Resource Management Service [Yam00] relies on Local Resource Managers (LRMs) present in each node of a *2K* cluster and whose task is to export the state of the hardware resources in that node to the whole distributed system. LRMs send periodic updates of the state of their resources to the Global Resource Manager (GRM), a replicated service that maintains an approximate view of the cluster resource utilization state. The GRM then utilizes this information as a hint for performing QoS-aware load distribution within its cluster. Groups of GRMs can be combined hierarchically to provide hardware resource sharing across multiple clusters connected through the Internet.

The LRMs are also responsible for performing QoS-aware admission control, resource negotiation, reservation, and scheduling of tasks on a single node. This is achieved with the help of a Dynamic Soft Real-Time Scheduler (DSRT) [NhCN98] that runs as a user-level process in conventional operating systems.

2K uses a CORBA Trader [OMG98] to supply resource discovery services, which allow applications to request resources based on QoS specifications. In this way, the system helps parallel and distributed applications achieve the best performance with the available resources.

3.4 Dynamic Security

Access to *2K* services is restricted to controlled CORBA interfaces. For that, we utilize the OMG Standard Security Service [OMG98] that comprises authentication, access control, auditing, object communication encryption, non-repudiation, and administration of security information.

Our implementation of the CORBA Security Service utilizes the *Cherubim* security framework [CQ98] which supports dynamic security policies [Qia99]. The reflective ORB allows on-the-fly reconfiguration of the Security Service, facilitating the adoption of situation-specific policies and mechanisms for authentication and encryption. The implementation currently supports various access control models including Discretionary Access Control (DAC), Double Discretionary Access Control (DDAC), and Mandatory Access Control (MAC) [SS94a]. We are now extending it to support Role-Based Access Control (RBAC) [Rav96], which will be the basis for security in large-scale *2K* environments.

The possibilities for dynamically configuring the security subsystem that *2K* provides are very useful for a wide range of applications in several situations. As an example, consider a computationally intensive application that runs initially in a single cluster and later expands itself to use the processors of several clusters connected via the public Internet. It may be acceptable to use lightweight encryption and soft access control in the intranet but it may be necessary to apply strong encryption and very tight access control policies when communicating over the public Internet.

3.5 Reflective ORBs

One of the major constituent elements of *2K* is *dynamicTAO* [RKC99], a CORBA-compliant reflective ORB. *dynamicTAO* is an open source extension of the TAO ORB [SC99]. It enables on-the-fly reconfiguration of the ORB internal engine and of applications running on top of the ORB. In *dynamicTAO*, we used the *ComponentConfigurator* model described in section 3.1.2 to represent the dependence relationships between ORB components and between ORB and application components. The current version supports safe dynamic reconfiguration of the strategies that control aspects such as concurrency, security, and monitoring. *dynamicTAO* exports an interface for loading and unloading modules into the system runtime, and for inspecting and changing the ORB configuration state.

After our experience in developing applications with both open source and commercial ORBs, we came to the conclusion that typical applications utilize just a very small fraction of the services and functionalities offered by common ORBs. Besides, one of the criticism that CORBA often receives is that it is too heavyweight to be used in small devices and embedded systems. Although *dynamicTAO* is configurable dynamically, its memory footprint is never less than a few megabytes, which makes it inappropriate for environments with limited resources and for applications with stringent resource requirements. This motivated our group to develop a new ORB architecture called *LegORB*. It can be dynamically customized to adapt to resource availability and to accommodate the requirements of different applications and devices at different moments.

Unlike TAO, *LegORB* is designed with componentization and dynamic reconfiguration as a fundamental premise. Careful design and implementation has allowed us to achieve surprising results in terms of code size. A minimal configuration of *LegORB* that is able to send simple CORBA requests to standard ORBs occupies only around 6Kbytes on a PalmPilot running PalmOS. The development of *LegORB* is still in its early stages, but the preliminary results indicate that it will be not only a good choice for embedded systems and PDAs, but also for high-performance workstations where *LegORB* can perform even faster than highly-optimized commercial ORBs.

4 Lessons Learned

In the past two years of our work on the design and implementation of *2K*, our research group has learned a number of lessons that we consider significant.

It is unlikely that a large number of users would be willing to adopt a completely new research operating system to use on a daily basis. Thus, we decided that *2K* would have to be able to run on top of other operating systems and, if necessary, co-exist with traditional applications. In that manner, users of traditional systems can extend the functionality of their machines by using the QoS-awareness, network-centrism, code distribution, and dynamic configuration properties of *2K* to manage their conventional system. The users that need the extra control and performance offered by a customizable microkernel can choose to boot their machines with the *2K* microkernel.

By using a platform like CORBA, we have the opportunity of re-using a large number of distributed services and applications that were developed by the CORBA community, saving us a lot of development time. In addition, the use of IDL interfaces among distributed and even co-located system components improves system organization considerably.

The possibility of changing the implementation of the different aspects of the CORBA middleware through the use of a reflective ORB opens new possibilities in terms of code re-use. If a scientific application requires a special underlying protocol or a special optimization, the programmer can implement it as a reflective ORB communication module, making it available to other applications with similar needs. The reflective architecture allows the deployment of these different protocols and optimizations without modifications to the application code.

5 Performance Considerations

Our reflective ORB is an extension of TAO [SC99], a CORBA-compliant ORB that optimizes inter-object communication by using different protocols depending on the location of the objects. Calls to co-located servers can be as fast as virtual method calls on a C++ object[1].

On Linux running on a single 450MHz Pentium II with 256M of RAM, it takes 236 μs for TAO to perform a cross-domain method invocation with a single parameter [LFGS99] which is an acceptable figure for a wide-range of applications. Applications requiring a better performance can customize the ORB (and, if desired, the microkernel) to optimize the system. TAO supports pluggable protocols which allow specific transports to be used to maximize application performance.

We measured the performance of our infrastructure for dynamic configuration based on mobile agents (see section 3.2) by sending configuration and inspection agents into a network of six ORBs running on Sun Ultra-2 machines connected by a 100Mbps Ethernet. The inspection agent carried code to collect information about the state of the six reflective ORBs, bringing it back to the administrator. The total average time for sending, processing, and returning the agent was 101 milliseconds.

The configuration agent carried instructions to load a 30Kbyte component to the runtime of the six ORBs and attach the new component to a running application in each of the ORBs. It took 265 milliseconds, on average, to complete its task and return the results to the administrator.

Although these numbers can be improved significantly with more tuning and optimizations, they show that it is possible to carry out dynamic configuration of a collection of distributed components in few tenths of a second.

[1] The general impression that CORBA is big and slow corresponds to first-generation brokers. Recent performance measurements [POS+00] suggest that contemporary CORBA implementations are efficient and that even faster implementations will appear.

In another experiment, we measured the performance of our infrastructure for dynamic code distribution in a wide-area system composed of nine nodes, three in the USA, three in Brazil, and three in Spain. Figure 3 shows a comparison between the performance of our agent-based approach and a conventional point-to-point approach as the size of the component being uploaded increases.

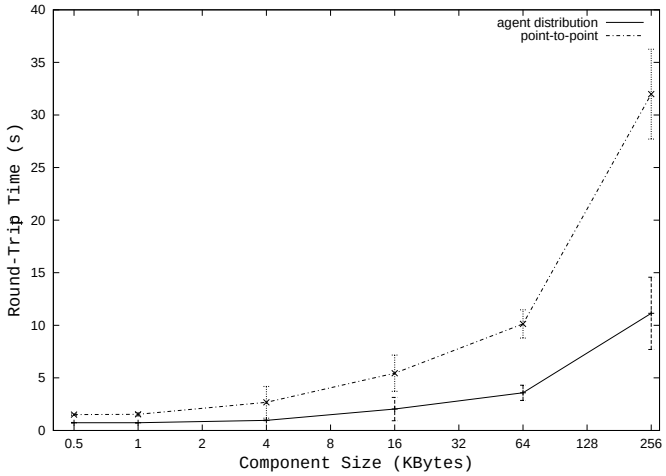

Figure 3: Agent uploading a new component to 9 nodes

6 Related Work

Our work builds on previous and ongoing research in a number of different areas including operating system architecture, cluster computing, configurable middleware, mobile agents, dynamic security, dynamic configuration, and software architecture.

SPIN [B+95] and VINO [SS94b] are adaptable systems which load code into the kernel to allow system extensions. We build on their work, and employ code downloading (through the network) to install new components into *2K* nodes. Choices and its derivatives [CIMR93] implement operating system services by means of a collection of object-oriented frameworks. These systems, however, do not address large-scale, heterogeneous distributed systems.

Spring [M+94] is an object-oriented, distributed operating system which also uses IDL-based interfaces for system services. *2K* takes the ideas introduced by Spring a step further by adopting the CORBA communication model and standard CORBA services as the glue to connect *heterogeneous* hardware and software platforms and by representing and managing inter-component dependence.

Systems like Condor [LLM88] are targeted to high performance computing on clusters of workstations. They rely on a central resource manager that starts processes on workstations with spare cycles. Globus [FK98] provides a "computational grid" integrating distributing resources in the same system. It also supports scalable resource management based on a hierarchy of resource managers, as we do. Unlike Globus, *2K* builds on OMG standards and provides a framework for dealing with component prerequisites and dynamic interdependencies using the Component Configurators.

Recent research targeted at using the Internet as "the computer" has led to systems like Globe [vSHT99], Legion [GW+97], and WebOS [VAD+98]. Although some may be customizable, they do not consider adaptability, dependence management, and automatic configuration as a primary requirement. To the best of our knowledge, none of the systems mentioned above include a model enabling automatic configuration of component-based systems on distributed, heterogeneous environments.

One of the major contributions of our work is to combine these important research results using a completely standard environment based on CORBA objects and standard CORBA services. It also brings automatic configuration – previously limited to isolated tools for application development – to the core of a distributed, object-oriented operating system. With the evolution of mechanisms for automatic configuration, such as the ones available in *2K*, we foresee a bright future for distributed

operating systems for high-performance computing that will be easy to manage, comfortable to use, and extremely powerful.

7 Conclusions

We expect great changes in the environments for high performance distributed computing in the first decade of the new millennium, including higher degrees of dynamism, mobility, heterogeneity, and interactions among heterogeneous computing devices connected to global networks. Traditional middleware and operating system architectures are not prepared to provide efficient resource management for these highly dynamic heterogeneous environments.

In this paper, we presented an integrated operating system architecture for managing distributed heterogeneous resources using recent advances in software systems technology. *2K* provides support for code distribution, configuration of component-based distributed applications, and QoS-aware distributed resource management. *2K* can run both as "middleware" on top of traditional operating systems and as an integrated architecture with our customized microkernel directly on top of the hardware.

Finally, *2K* uses and offers services based on the CORBA standard which opens a wide variety of possibilities for integration with other systems and applications.

Availability

Documentation and source code for the *2K* microkernel, middleware, and distributed services can be found at `http://choices.cs.uiuc.edu/2K`.

References

[B+95] B. N. Bershad et al. Extensibility, Safety and Performance in the SPIN Operating System. In *Proc. of the 15th SOSP*. ACM, December 1995.

[CIMR93] Roy Campbell, Nayeem Islam, Peter Madany, and David Raila. Designing and Implementing Choices: an Object-Oriented System in C++. *Communications of the ACM*, 36(9):117–136, September 1993.

[CQ98] Roy Campbell and Tin Qian. Dynamic Agent-based Security Architecture for Mobile Computers. In *Proceedings of the Second International Conference on Parallel and Distributed Computing and Networks (PDCN'98)*, pages 291–299, Australia, December 1998.

[FK98] I. Foster and C. Kesselman. The Globus Project: A Status Report. In *Proceedings of the IPPS/SPDP '98 Heterogeneous Computing Workshop*, pages 4–18, 1998.

[GW+97] Andrew S. Grimshaw, Wm. A. Wulf, et al. The Legion Vision of a Worldwide Virtual Computer. *Communications of the ACM*, 40(1), January 1997.

[KGCM99] Fabio Kon, Binny Gill, Roy H. Campbell, and M. Dennis Mickunas. Secure Dynamic Reconfiguration of Scalable CORBA Systems with Mobile Agents. Technical Report UIUCDCS-R-99-2131, Department of Computer Science, University of Illinois at Urbana-Champaign, December 1999.

[LFGS99] David L. Levine, Sergio Flores-Gaitan, and Douglas C. Schmidt. An Empirical Evaluation of OS Support for Real-Time CORBA Object Request Brokers. In *Proceedings of the International Symposium on Distributed Objects and Applications DOA99*, Edimburgh, Scotland, September 1999.

[LLM88] M. Litzkow, M. Livny, and M. W. Mutka. Condor - A Hunter of Idle Workstations. In *Proceedings of the 8th International Conference of Distributed Computing Systems*, pages 104–111, 1988.

[M+94] J. Mitchell et al. An Overview of the Spring System. In *Proceedings of Compcon 'Spring 1994*, February 1994.

[NhCN98] Klara Nahrstedt, Hao hua Chu, and Srinivas Narayan. Qos-aware resource management for distributed multimedia applications. *Journal on High-Speed Networking, Special Issue on Multimedia Networking*, 1998. Available at `http://cairo.cs.uiuc.edu/papers.html`.

[OMG98] OMG. *CORBAservices: Common Object Services Specification.* Object Management Group, Framingham, MA, 1998. OMG Document 98-12-09.

[POS+00] Irfan Pyarali, Carlos O'Ryan, Douglas Schmidt, Nanbor Wang, Aniruddha S. Gokhale, and Vishal Kachroo. Using Principle Patterns to Optimize Real-Time ORBs. *IEEE Concurrency*, 8(1):16–25, January-March 2000.

[Qia99] Tin Qian. *Dynamic Authorization Support in Large Distributed Systems.* PhD thesis, Department of Computer Science, University of Illinois at Urbana-Champaign, November 1999.

[Rav96] Ravi S. Sandhu and Edward J. Coyne and Hal L. Feinstein and Charles E. Youman. Role-based Access Control Models. *IEEE Computer*, 29(2):38–47, February 1996.

[RKC99] Manuel Román, Fabio Kon, and Roy H. Campbell. Design and Implementation of Runtime Reflection in Communication Middleware: the *dynamicTAO* Case. In *Proceedings of the ICDCS'99 Workshop on Middleware*, pages 122–127, Austin, TX, June 1999.

[SC99] Douglas C. Schmidt and Chris Cleeland. Applying Patterns to Develop Extensible ORB Middleware. *IEEE Communications Magazine Special Issue on Design Patterns*, 37(4):54–63, May 1999.

[SS94a] Ravi S. Sandu and Pierangela Samarati. Access Control: Principles and Practice. *IEEE Communications Magazine*, 32(9):40–48, September 1994.

[SS94b] Christopher Small and Margo Seltezer. VINO: An Integrated Platform for Operating System and Database Research. Technical report, Computer Science Laboratory, Harvard University, Cambridge, MA 02138, 1994.

[VAD+98] Amin Vahdat, Thomas Anderson, Michael Dahlin, David Culler, Eshwar Belani, Paul Eastham, and Chad Yoshikawa. WebOS: Operating System Services For Wide Area Applications. In *Proceedings of the Seventh Symposium on High Performance Distributed Computing*, July 1998.

[vSHT99] Maarten van Steen, Philip Homburg, and Andrew S. Tanenbaum. Globe: A Wide-Area Distributed System. *IEEE Concurrency*, 7(1):70–78, January 1999.

[Yam00] Tomonori Yamane. The Design and Implementation of the 2K Resource Management Service. Master's thesis, Department of Computer Science, University of Illinois at Urbana-Champaign, February 2000.