Self Management of Large-Scale Distributed Systems by Combining Structured Overlay Networks and Components[⋆]

Peter Van Roy[1], Ali Ghodsi[2], Jean-Bernard Stefani[3] Seif Haridi[2], Thierry Coupaye[4], Alexander Reinefeld[5], Ehrhard Winter[6], and Roland Yap[7]

[1] UCL/Universit catholique de Louvain, pvr@info.ucl.ac.be,
[2] KTH/Royal Institute of Technology, {aligh, haridi}@kth.se,
[3] INRIA/Institut National de Recherche en Informatique,
Jean-Bernard.Stefani@inria.fr
[4] France Telecom R&D, thierry.coupaye@francetelecom.com
[5] ZIB/Zuse Institute in Berlin, ar@zib.de
[6] E-Plus Mobilfunk, Ehrhard.Winter@eplus.de
[7] NUS/National University of Singapore, ryap@comp.nus.edu.sg

Abstract. This position paper envisions making large-scale distributed applications self managing by combining *component models* and *structured overlay networks*. A key obstacle to deploying large-scale applications running on Internet is the amount of management they require. Often these applications demand specialized personnel for their maintenance. Making applications self-managing will help removing this obstacle. Basing the system on a structured overlay network will allow extending the abilities of existing component models to large-scale distributed systems. Structured overlay networks provide guarantees for efficient communication, efficient load-balancing, and self-manage in case of joins, leaves, and failures. Component models, on the other hand, support dynamic configuration, the ability of part of the system to reconfigure other parts at run-time. By combining overlay networks with component models we achieve both low-level as well as high-level self-management. We will target multi-tier applications, and specifically we will consider three-tier applications using a self-managing storage service.

1 Introduction

Multi-tier applications are the mainstay of industrial applications. A typical example is a three-tier architecture, consisting of a client talking to a server, which itself interfaces with a database (see Figure 1). The business logic is executed at the server and the application data and meta data are stored on the database. But multi-tier architectures are brittle: they break when exposed to stresses such as failures, heavy loading (the "slash-dot effect"), network congestion, and changes in their computing environment. This becomes especially cumbersome for large-scale systems. Therefore, cluster-based solutions are employed where

[⋆] This collaborative work is supported by the Network of Excellence CoreGRID (contract no. 004265), funded by the EU in the sixth framework programme.

the three-tier architecture is duplicated within a cluster with high speed interconnectivity between tightly coupled servers. In practice, these applications require intensive care by human managers to provide acceptable levels of service, and make assumptions which are *only* valid within a cluster environment, such as perfect failure detection.

Lack of self-management is not only pervasive in multi-tier architectures, but a problem in most distributed systems. For example, deploying a distributed file system across several organizations requires much manual configuration, as does adding another file server to the existing infrastructure. If a file server crashes, most file systems will stop functioning or fail to provide full service. Instead, the system should reconfigure itself to use another file server. This desirable behavior is an example of self management.

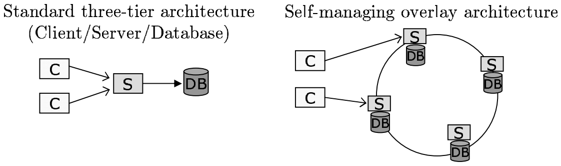
Standard three-tier architecture
(Client/Server/Database)
C
C
S
DB
Self-managing overlay architecture
C
C
S
DB
S
DB
S
DB
S
DB

Fig. 1. Left: Traditional three-tier arch. Right: A self-managing overlay arch.

In our vision, we intend to make large-scale distributed applications such as these self managing. In this position paper we outline our vision of a general architecture which combines research on structured overlay networks together with research on component models. These two areas each provide what the other lacks: structured overlay networks provide a robust communications infrastructure and low-level self-management properties for Internet-scale distributed systems, and component models provide the primitives needed to support dynamic configuration and enable high-level self-management properties.

1.1 Definition of self management

Self management and self organization are overloaded terms widely used in many fields. We define self-management along the same lines as done in [1], which can be summarized in that the system should be able to reconfigure itself to handle changes in its environment or requirements without human intervention but according to high-level management policies. It is important to give a precise definition of self management that makes it clear what parts can be handled automatically and what parts need application programmer or user (system administrator) intervention. The user then defines a self management policy and the system implements this policy. Self management exists on all levels of the system. At the lowest level, self management means that the system should be able to automatically handle frequent addition or removal of nodes, frequent failure of nodes, load balancing between nodes, and threats from adversaries. For large-scale systems, environmental changes that require some recovery by the system become normal and even frequent events. For example, failure becomes a normal situation: the probability that at a given time instant some part of the system is failed approaches 1 as the number of nodes increases. At higher levels, self management embraces many system properties. For our approach, we consider that these properties are classified in four axes of self management: *self configuration*, *self healing*, *self tuning*, and *self protection*.

To be effective, self management must be designed as part of the system from its inception. It is difficult or impossible to add self management a posteriori. This is because self management needs to be done at many levels of the system. Each level of the system needs to provide self management primitives (”hooks”) to the next level.

The key to supporting self management is a service architecture that is a framework for building large-scale self-managing distributed applications. The heart of the service architecture is a component model built in synergy with a structured overlay network providing the following self-management properties:

1. *Self configuration*: Specifically, the infrastructure provide primitives so that the service architecture will continue to work when nodes are added or removed during execution. We will provide primitives so that parts of the application can be upgraded from one version to another without interrupting execution (online upgrade) . We will also provide a component trading infrastructure that can be used for automating distributed configuration processes.
2. *Self healing*: The service architecture will provide the primitives for continued execution when nodes fail or when the network communication between nodes fails, and will provide primitives to support the repair of node configurations. Specifically, the service architecture will continue to provide its basic services, namely communication and replicated storage, and will provide resource trading facilities to support repair mechanisms. Other services are application-dependent; the service architecture will provide the primitives to make it easy to write applications that are fault-tolerant and are capable of repairing themselves to continue respecting service level agreements.
3. *Self tuning*: The service architecture will provide the primitives for implementing load balancing and overload management. We expect that both load balancing and online upgrade will be supported by the component model, in the form of introspective operations (including the ability to freeze and restart a component and to get/set a component’s state).
4. *Self protection*: Security is an essential concern that has to be considered globally. In a first approximation, we will consider a simple threat model, in which the nodes of the service architecture are considered trustworthy. We can extend this threat model with little effort for some parts, such as the structured overlay network, for which we already know how to protect against more aggressive threat models, such as Sybil attacks.

An essential feature of self management is that it adds feedback loops throughout the system. A feedback loop consists of (1) the detection of an anomaly, (2) the calculation of a correction, and (3) the application of the correction. These feedback loops exist within one level but can also cross levels. For example, the low level detects a network problem, a higher level is notified and decides to try another communication path, and the low level then implements that decision. Because of the feedback loops, it is important that the system behavior converges (no oscillatory, chaotic, or divergent behavior). In the future, we intend to model formally the feedback loops, to confirm convergent behavior (possibly changing the design), and to validate the model with the system. The formal model of a computer system is generally highly nonlinear. It may be possible to exploit oscillatory or chaotic behavior to enhance certain characteristics of the system. We will explore this aspect of the feedback loops.

2 Related Work

Our approach to self management can be considered a computer systems approach. That is, we give a precise definition of self management in terms of computer system properties, namely configuration, fault tolerance, performance, and security. To make these properties self managing, we propose to design a system architecture and the protocols it needs. We consider that our approach is an effective one and that our project is a realistic way to achieve self management according to our definition. But in the research community self management is sometimes defined in a broader way, to touch on various parts of artificial intelligence: learning systems, swarm intelligence (a.k.a. collective intelligence), biologically-inspired systems, and learning from the immune system[1]. We consider that these artificial intelligence approaches are worth investigating in their own right. However, we consider that the computer systems approach is a fundamental one that has to be solved, regardless of these other approaches.

Let us characterize the advantages of our proposed architecture with respect to the state of the art in computer systems. There are three areas to which we can compare our approach:

1. *Structured overlay networks and peer-to-peer systems.* Current research on overlay networks focuses on algorithms for basic services such as communication and storage. The reorganizing abilities of structured overlay networks can be considered as low-level self management. We extend this to address high-level self management such as configuration, deployment, online updating, and evolution, which have been largely ignored so far in structured overlay network research.
2. *Component-based programming.* Current research on components focuses on architecture design issues and not on distributed programming. We extend this to study component-based abstractions and architectural frameworks for large-scale distributed systems, by using overlay networks as an enabler.
3. *Autonomic systems.* Most autonomic systems focus on individual autonomic properties, specific self-managed systems, or focus on specific elements of autonomic behavior. Little research has considered the overall architectural implications of building self-managed distributed systems. Our project proposal is unique in this respect, combining as it does component-based system

construction with overlay network technology into a service architecture for large-scale distributed system self management.

We now present these areas in more detail and explain where the contribution of our approach fits.

2.1 Structured overlay networks and peer-to-peer systems

Research on peer-to-peer networks has evolved into research on structured overlay networks, in particular on Distributed Hash Tables (DHTs). The main differences between popular peer-to-peer systems and structured overlay networks are that the latter provide strong guarantees on routing and message delivery, and are implemented with more efficient algorithms. The research on structured overlay networks has matured considerably in the last few years[2–4]. Hardware infrastructures such as PlanetLab have enabled DHTs to be tested in realistically harsh environments. This has led to structured peer-to-peer communication and storage infrastructures in which failures and system changes are handled gracefully.

At their heart, structured overlay networks enable the nodes in a distributed system to organize themselves to provide a shared directory service. Any application built on top of an overlay can add information to this directory locally, which immediately results in the overlay system distributing the data onto the nodes in the system, ensuring that the data is replicated in case some of the nodes become unavailable due to failure.

The overlay guarantees that any node in the distributed system can access data inserted to the directory efficiently. The efficiency, calculated as the number of reroutes, is typically $log_k(N)$, where N is the number of nodes in the system, and k is a configurable parameter. The overlay makes sure that the nodes are interconnected such that data in the directory always can be found. The number of connections needed vary in different system, but are typically in the range $O(1)$ to $O(\log N)$, where N is the number of nodes in the overlay.

Though most overlays provide a simple directory, other abstractions are possible too. More recently, a relational view of the directory can be provided[5], and the application can use SQL to query the relational database for information. Most ordinary operations, such as selection, projection, and equi-joins are supported.

All structured overlays provide self-management in presence of node joins and node departures. This means that a running system will adapt itself if new nodes arrive or if some nodes depart. Self-management is done at two distinct layers: the communication layer and the storage management layer.

When nodes join or leave the system, the communication layer of the structured peer-to-peer system will ensure that the routing information present in the system is updated to adapt to these changes. Hence, routing can efficiently be done in presence of dynamism. Similarly, the storage management layer maintains availability of data by transferring data which is stored on a departing node to an existing node in the system. Conversely, if a new node arrives, the storage management layer moves part of the existing data to the new node to ensure that data is evenly distributed among the nodes in the system. Hence, data is self-configured in presence of node joins and leaves.

In addition to the handling of node joins and leaves, the peer-to-peer system self-heals in presence of link failures. This requires that the communication layer can accurately detect failures and correct routing tables accordingly. Moreover, the communication layer informs the storage management layer such that data is fetched from replicas to restore the replication degree when failures occur.

Much research has also been conducted in making peer-to-peer systems self-tuning. There are many techniques employed to ensure that the heterogeneous nodes that make up the peer-to-peer system are not overloaded[6]. Self-tuning is considered with respect to amount of data stored, amount of routing traffic served, and amount of routing information maintained. Self-tuning is also applied to achieve proximity awareness, which means that routing done on the peer-to-peer network reflects the latencies in the underlying network.

Lately, research has been conducted in modeling trust to achieve security in large-scale systems[7]. In essence, a node's future behavior can be predicted by judging its previous behavior. The latter information can be acquired by regularly asking other nodes about their opinion about other nodes.

2.2 Component-based programming

The main current de-facto standards in distributed software infrastructures, Sun's J2EE, Microsoft .Net, and OMG CORBA, provide a form of component-based distributed programming. Apart from the inclusion of publish-subscribe facilities (e.g. the JMS publish-subscribe services in J2EE), support for the construction of large-scale services is limited. Management functions are made available using the traditional manager agent framework [8] but typically do not support online reconfiguration or autonomous behavior (which are left unspecified). Some implementations (e.g. JBoss) have adopted a component-based approach for the construction of the middleware itself, but they remain limited in their reconfiguration capabilities (coarse-grained, mostly deployment time, no support for unplanned software evolution).

Component models supported by standard platforms such as J2EE (the EJB model) or CORBA (the CCM model) are non-hierarchical (an assemblage of several components is not a component), and provide limited support for component introspection and dynamic adaptation. These limitations have been addressed in work on adaptive middleware (e.g. OpenORB, Dynamic TAO, Hadas, that have demonstrated the benefits of a reflective component-based approach to the construction of adaptive middleware). In parallel, a large body of work on architecture description languages (e.g. ArchJava, C2, Darwin, Wright, Rapide, Piccola, Acme or CommUnity) has shown the benefits of explicit software architecture for software maintenance and evolution. The component models proposed in these experimental prototypes, however, suffer from several limitations:

1. They do not allow the specification of component structures with sharing, a key feature required for the construction of software systems with resource multiplexing.

2. They remain limited in their adaptation capabilities, defining, for those that do provide such capabilities, a fixed meta-object protocol that disallows various optimizations and does not support different design trade-offs (e.g. performance vs. flexibility).
3. Finally, and most importantly, they lack abstractions for building large distributed structures.

Compared to the current industrial and academic state of the art in component-based distributed system construction, our approach intends to extend a reflective component-based model that subsumes the capabilities of the above models (it caters to points (1) and (2)) in order to address point (3).

2.3 Autonomic systems

The main goal of autonomic system research is to automate the traditional functions associated with systems management, namely configuration management, fault management, performance management, security management and cost management [8]. This goal is becoming of utmost importance because of increasing system complexity. It is this very realization that prompted major computer and software vendors to launch major R&D initiatives on this theme, notably, IBM's Autonomic Computing initiative and Microsoft's Dynamic Systems initiative.

The motivation for autonomic systems research is that networked environments today have reached a level of complexity and heterogeneity that make their control and management by human administrators more and more difficult. The complexity of individual elements (a single software element can literally have thousands of configuration parameters), combined with the brittleness inherent of today's distributed applications, makes it more and more difficult to entertain the presence of a human administrator in the "management loop". Consider for instance the following rough figures: One-third to one-half of a company's total IT budget is spent preventing or recovering from crashes, for every dollar used to purchase information storage, 9 dollars are spent to manage it, 40% of computer system outages are caused by human operator errors, not because they are poorly trained or do not have the right capabilities, but because of the complexities of today's computer systems.

IBM's autonomic computing initiative, for instance, was introduced in 2001 and presented as a "grand challenge" calling for a wide collaboration towards the development of computing systems that would have the following characteristics: self configuring, self healing, self tuning and self protecting, targeting the automation of the main management functional areas (self healing dealing with responses to failures, self protecting dealing with responses to attacks, self tuning dealing with continuous optimization of performance and operating costs). Since then, many R&D projects have been initiated to deal with autonomic computing aspects or support techniques. For example, we mention the following projects that are most relevant to our vision: the recovery-oriented computing project at UC Berkeley, the Smartfrog Project at HP Research Labs in Bristol, UK, and the Swan project at INRIA, Alcatel, France Telecom. Compared to these projects, the uniqueness of our approach is that it combines structured overlay networks with component models for the development of an integrated architecture for large-scale self-managing systems. Each complements the other: overlay networks support large-scale distribution, and component models support reconfiguration. None of the aforementioned projects provide such a combination, which gives a uniform architectural model for self-managing systems. Note also that many of the above-mentioned projects are based on cluster architectures, whereas our approach targets distributed systems that may be loosely coupled.

3 Synergy of Overlays and Components

The foundation of our approach is to combine a structured overlay network with a component model. Both areas have much matured in recent years, but they have been studied in isolation. It is a basic premise of our approach that their combination will enable achieving self-management in large-scale distributed systems. This is first of all because structured overlay networks already have many *low-level* self-management properties. Structured overlay network research has achieved efficient routing and communication algorithms, fault tolerance, handling dynamism, proximity awareness, and distributed storage with replication. However, almost no research has been done on deployment, upgrading, continuous operation, and other *high-level* self-management properties.

We explain what we mean with lack of high level self-management in overlay networks by the following concrete problems. An overlay network running on thousands of nodes will occasionally need software upgrade. How can a thousand node peer-to-peer system, dispersed over the Internet, be upgraded on the fly without interrupting existing services, and how do we ensure that it is done securely. How can it be guaranteed that the new version of the overlay software will not break when deployed on a node which does not have all the required software. For example, the new version might be making calls to certain libraries which might not be available on every node.

To continue the example, nodes in the overlay might provide different services or may run different versions of the services. For instance, an overlay might provide a rudimentary routing service on every node. But it might be that high-level services, such as a directory service, do not exist on every node. We need to be able to introspect nodes to find out such information, and, if permitted, install the required services on the remote machine at runtime. Even if the nodes do provide a directory service, it might be of different incompatible versions. For example, a node might be running an old version which stores directory information in memory, while another node has support for secure and persistent data storage.

The above mentioned research issues have been ignored by the peer-to-peer community. By using components, we can add these high-level self-management properties, such as deployment, versioning, and upgrade services. Recent research on component models, such as the Fractal model[9], is adding exactly those abilities that are needed for doing self management (such as reification and reflection abilities).

3.1 A Three-tier e-commerce Application

We now give a motivational example which will show how the overlay and the component model is used to build a scalable fault-tolerant application.

Imagine an e-commerce application, which allows users to use their web browser to buy books. The user can browse through the library of books, and add/remove books to its shopping cart. When the user has finished shopping, it can decide to either make a purchase or cancel it.

Traditionally, the above application is realized by creating a three-tier architecture, where the client makes request to an application server, which uses a database to store session information, such as the contents of the shopping cart.

In our system (see Figure 1), there will be several application servers running on different servers, possibly geographically dispersed running on heterogeneous hardware. Each application server is a node in a structured overlay network and can thus access the storage layer, which is a distributed hash table provided by the overlay. The storage layer has a thin layer which provides a relational view of the directory, allowing SQL queries, and supports transactions ontop of the distributed hash table. A user visiting the shopping site will be forwarded by a load-balancer to an appropriate server which can run the e-commerce application. The component model will enable the load-balancer to find the server which has the right contextual environment, e.g. with J2EE installed and with certain libraries, and which is not overloaded. Thereafter the request is forwarded to the appropriate server, which uses the overlay storage layer to store its session state.

To continue our above example, we would like the e-commerce application to be self-healing and provide *failover*. This can be realized by providing a failover component which periodically checkpoints by invoking an interface in the e-commerce application forcing it to save its entire state and configuration to the overlay storage . Should the application server crash, the failure detectors in the crashed node's neighborhood will detect this. One such neighbor is chosen by the overlay and the application is executed on that node. The component model ensures that the last saved state will be loaded by making calls to a standard interface in e-commerce application which will load the session state.

We might want our application to be self-tuning, such that the e-commerce application running on an overloaded server is migrated to another application server . This could be solved using different approaches. One approach would be to have a component which saves the state of a session, and initiates another server to start the e-commerce application with the saved state. Notice that the level of granularity is high in this case as the component model would only define interfaces for methods which the e-commerce application would implement. These methods would then save the application specific state to the storage layer. Similarly, interfaces would be defined to tell the application to load its state from the storage layer. Another approach, with a low-level of granularity, would be to use a virtual machine such as Xen, or VMWare. With these, the whole e-commerce application, its OS and state, would be moved to another machine. This would nevertheless require that the application is running on a common distributed file system, or is getting its data from a common database.

The overlay could be used to either provide a self-managing distributed file system, or let the application use the overlay storage to fetch and store its data. The virtual machine approach has the additional advantage that it guarantees that applications running on the same machine are shielded securely from each other. At the same time, the virtual machine approach would not be able to run if the servers actual hardware differ.

4 Conclusions

We have outlined an approach for building large-scale distributed applications that are self managing. The approach exploits the synergy between structured overlay networks and component models. Each of these areas has matured considerably in recent years, but in isolation. Each area lacks the abilities provided by the other. Structured overlay networks lack the deployment and configuration abilities of component models. Component models lack the decentralized distributed structure of structured overlay networks. By combining the two areas, we expect to eliminate both of these lacks and achieve a balanced approach to self management.

References

1. Herrmann, K., Mühl, G., Geihs, K.: Self-management: The solution to complexity or just another problem? IEEE Distributed Systems Online (DSOnline) **6**(1) (2005)
2. Stoica, I., Morris, R., Karger, D., Kaashoek, M., Balakrishnan, H.: Chord: A Scalable Peer-to-Peer Lookup Service for Internet Applications. In: ACM SIGCOMM 2001, San Deigo, CA (2001) 149–160
3. Rowstron, A., Druschel, P.: Pastry: Scalable, Decentralized Object Location, and Routing for Large-Scale Peer-to-Peer Systems. Lecture Notes in Computer Science **2218** (2001)
4. Alima, L.O., Ghodsi, A., Haridi, S.: A Framework for Structured Peer-to-Peer Overlay Networks. In: LNCS post-proceedings of Global Computing, Springer Verlag (2004) 223–250
5. Chun, B., Hellerstein, J.M., H., R., Jeffery, S.R., Loo, B.T., Mardanbeigi, S., Roscoe, T., Rhea, S., Shenker, S., Stoica, I.: Querying at internet scale. In: SIGMOD '04: Proceedings of the 2004 ACM SIGMOD international conference on Management of data, New York, NY, USA, ACM Press (2004) 935–936
6. Karger, D.R., Ruhl, M.: Simple efficient load balancing algorithms for peer-to-peer systems. In: SPAA '04: Proceedings of the sixteenth annual ACM symposium on Parallelism in algorithms and architectures, New York, NY, USA, ACM Press (2004) 36–43
7. Aberer, K., Despotovic, Z.: Managing trust in a peer-2-peer information system. In Paques, H., Liu, L., Grossman, D., eds.: Proceedings of the Tenth International Conference on Information and Knowledge Management (CIKM01), ACM Press (2001) 310–317
8. Distributed Management Task Force: http://www.dmtf.org (2005)
9. Bruneton, E., Coupaye, T., Leclercq, M., Stefani, V.Q.J.B.: An Open Component Model and Its Support in Java, Lecture Notes in Computer Science. Lecture Notes in Computer Science **3054** (2004)