The Autonomous Cloud

Karl-Erik Årzén, Lund University

Vision for the Project

An increasing amount of computing and information services are moving to the cloud, where they execute on virtualized hardware in private or public data centers. Hence, the cloud can be viewed as an underlying computing infrastructure for all systems of systems. The architectural complexity of the cloud is rapidly increasing. Modern data centers consist of tens of thousands of components, e.g., compute servers, storage servers, cache servers, routers, PDUs, UPSs, and air-conditioning units, with configuration and tuning parameters numbering in the hundreds of thousands. Currently, there is also an ongoing development of more modular computing nodes, often called disaggregated or rack-scale systems, where aggregated resources (like compute, memory, network, etc) of a large set of servers are treated as large pools of compute, memory, network, etc.

The same increasing trend holds for the operational complexity. The individual components are themselves increasingly difficult to maintain and operate. The strong connection between the components furthermore makes it necessary to tune the entire system, which is complicated by the fact that in many cases the behaviors, execution contexts, and interactions are not known a priori. The term autonomous computing or autonomic computing was coined by IBM in the beginning of the 2000s for self-managing computing systems with the focus on private enterprise IT systems. However, this approach is even more relevant for the cloud. The motivation is the current levels of scale, complexity, and dynamicity which make efficient human management infeasible. In autonomous cloud control, AI, and machine learning/analytics techniques will be used to dynamically determine how applications should be best mapped onto the server network, how capacity should be automatically scaled when the load or the available resources vary, and how load should be balanced.

Currently there is also a growing interest in applying cloud techniques, such as virtualization and collocation, in the access telecommunication network itself. The unification of the telecom access network and the traditional cloud data centers, sometimes referred to as the distributed cloud or edge cloud, provide a single distributed computing platform. Here the boundary between the network and the data centers disappears, allowing application software to be dynamically deployed in all types of nodes, e.g., in edge nodes, e.g., base stations, near end-users, in remote large-scale datacenters, or anywhere in between. In these systems the need for autonomous operation and resource management becomes even more urgent as heterogeneity increases, when some of the nodes may be mobile with varying availability, and when new 5G-based mission-critical applications with harder requirements on latency, uptime, and availability are migrated to the cloud.

Research Challenges

In the project distributed control and real-time analytics will be used to dynamically solve resource management problems in the distributed cloud. The management problem consists of deciding the types and quantities of resources that should be allocated to each application, and when and where to deploy them. This also includes dynamic decisions such as automatic scaling of the resource amount when the load or the available resources vary, and on-line migration of application components between nodes. Major scientific challenges include

- How to create models for the distributed cloud infrastructure, useful for both system design and optimization of resource and application management?
- How to model and predict workloads, including variations in both time and locality?
- How to perform on-line distributed analytics for creating the dynamically maintained knowledge base needed for optimization of resource and application management?
- How to best use model-based feedback control and distributed control for controlling both system throughput and end-user response times?

- How to design an autonomous management system that dynamically maintains a sufficient degree of decentralization to handle the scale while still being able to make highly optimized management decisions?
- How to, in such a system, perform the most fundamental management optimizations, like vertical and horizontal capacity scaling, geo-placement of application components, service differentiation, and management?
- How to design and manage a distributed cloud to meet the requirements of mission critical applications?
- How to integrate and tailor such an autonomous system to intrinsically distributed and dynamic applications with massive data producers, such as, e.g., video cameras for surveillance and supervision?
- How to perform resource management and orchestration in disaggregated hardware architectures?

The interdependencies between these questions and the expected interactions in order to solve the problems can be illustrated as follows. In order to develop efficient methods for resource management, it is crucial to understand the performance aspects of the infrastructure, what the workloads look like, and how they vary over time. Due to user mobility and variations in usage and resource availability, applications using many instances are constantly subject to changes in the number of instances; the individual instances relocated or resized; the network capacity adjusted; etc. Hence, infrastructure modeling and workload modeling for the distributed cloud are fundamental. On-line analytic and learning based on extreme amounts of monitoring data can create knowledge to be used in autonomous management. Capacity autoscaling is needed to determine how much capacity should be allocated for a complete application or any specific part of it; and Dynamic geo-placement complements by determining when, where, and how instances should be relocated, e.g., from a data center to a specific base station; Since not all applications are equally important, e.g., due to differently priced service levels or due to some being critical to society (emergency, health care, etc.), all management must take into account Quality of Service differentiation, and for mission critical applications, there is also a need to intelligently provide redundancy and fall-back solutions. The management systems themselves needs to be capable at handling systems of extreme scale, and to handle both the highly distributed infrastructure and the individual nodes, including so called rack-scale systems.

Industrial Challenges

Sweden has a long industrial tradition in telecommunication and recently there has been a rapid increase in the cloud industry, from large-scale datacenter establishments to start-ups providing novel cloud services. Most industries, however, have difficulties in addressing the ambitious long-term basic research challenges required to carry out truly game changing technology shifts. We believe our proposed research will provide great opportunities for such industries. The project has particularly strong connections to Ericsson Research who claims that the project "addresses important scientific problems that have to be solved in order to establish the distributed cloud with dependable performance, low latency, and minimal environmental impact. This is an important enabler for the networked society." However, once the autonomous cloud is in place it will open up new applications for a lot of other industry sectors, e.g., process automation and cloud robotics, where part of the optimization-based computations can be moved into the cloud, and automated transport systems where fleet management operations as well as individual vehicle optimizations can be performed in the cloud. The project connects to several of the other WASP start projects in different ways. An immediate connection is the fact that cloud technology is of interest in almost all of the WASP projects as a way of implementing the compute-intensive parts of their systems. Examples of this are the two projects "Automated Transport Systems" and "Interaction and Communication with Autonomous Agents in Sensor-Rich Environments". The second and most important connection, though, is that other WASP projects may provide traffic with the requirements that the distributed edge cloud is designed to support, i.e., low-latency, high bandwidth, mobility, and dynamically changing QoS requirements. This could be the case, e.g., for "Automated Transport Systems" and for "Integrating Perception, Learning and Verification in Interactive Autonomous Systems".

Sub-projects

The personnel consist of Karl-Erik Årzén (project coordinator) and Maria Kihl, LU, who participate with one PhD student (Tommi Nylander) and Erik Elmroth, UmU, who participates with one researcher (Cristian Klein) and one PhD student (Chanh Nguyen). In addition to this, there are three industrial PhD students in the project, at Ericsson Kista (Amir Roozbeh), Ericsson Lund (Per Skarin), and at Axis Communications (Alexandre Martins). The Lund and Umeå groups have collaborated on cloud resource management since 2013 and have so far produced 15 publications with joint authorship. Both groups also have additional PhD students and post-docs working on related areas which guarantees the critical mass for the project. An initial project meeting was held at the 8th Cloud Control Workshop organized in Lövånger, Feb 1-3, 2016 and the next meeting will be held in connection with the 9th Cloud Control Workshop organized, June 27-29, 2016 at Friiberghs Herrgård, by the Mälaren Sea.

The three industrial PhD students each have their own subproject:

- **Per Skarin** from Ericsson (Lund) will work on "**Mission-critical cloud**" with the objectives to design a cloud for mission-critical use cases such as industrial automation, transport, e-health, etc. that requires deterministic and highly available services and to investigate how distributed control and real-time analytics will be used to dynamically solve resource management problems in future converged cloud solutions.

- **Alexandre Martins** from Axis (Lund) will work on "**Autonomous learning camera systems in resource constrained environments**" with the objective to develop dynamic resource management techniques for distributed camera-based vision systems for surveillance and supervision. The major resource considered is network bandwidth, but also compute resources will be considered.

- **Amir Roozbeh** from Ericsson (Kista) will work on "**Autonomous network resource management in disaggregated datacenters**", i.e., develop new dynamic network management techniques that in addition to managing the traditional datacenter traffic also must be able to support new traffic requirements caused by the resource disaggregation.

The two university PhD students **Tommi Nylander** (Lund) and **Chanh Nguyen** (Umeå) together with the researcher **Cristian Klein** (Umeå) will work together on the different challenges of the main project. Here, Nylander will mostly focus on modeling and control-theoretical approaches to resource management, while Nguyen will mostly focus on distributed analytics-based approaches to identifying performance anomalies, detecting bottlenecks and gathering actionable insight. Klein will be working on mitigating resource wastage due to tail latency, through techniques such as adaptive capacity usage and collaboration among different levels of resource management.