ERICSSON

DESIGN PRINCIPLES FOR CLOUD-OPTIMIZED NETWORK APPLICATIONS

Introduction

Cloud-optimized network applications are developed specifically for cloud environments. Requirements for network applications include massive scalability, fault tolerance, infrastructure agnosticism, and automated lifecycle management. These characteristics are achieved through mechanisms like application decomposition and model-driven management. The promise of the cloud is to reduce time to market (TTM) and total cost of ownership (TCO).

Finding the right balance

The telecom environment puts special demands on network applications and cloud infrastructures. Services are expected to have predictable performance and latency. They must also be available at any time without disruption. Managing today's telecom networks is often complex and resource intensive. Many applications from different vendors need to interoperate to create a reliable system. New applications and cloud infrastructures must reduce this complexity and the associated costs.

Many web-scale applications demonstrate scalability, fast TTM, extensibility and automation. There is no doubt that these benefits are desirable in the telecom domain as well. The architectural patterns, often referred to as Cloud Native, are well documented for web applications, and some of them are directly applicable to network applications. Others need to be adapted to the characteristics and requirements of the telecom environment.

Jonas Falkenå
Senior Expert, Cloud Application Architecture
jonas.falkena@ericsson.com

Tamas Zsiros
Expert End-to-end systems architecture
Chief Architect Telecom Core Implementation Domain
tamas.zsiros@ericsson.com

Optimizing for cloud deployment is about finding the right balance and trade-off for each application and service. Ericsson has six main design principles for cloud-optimized network applications. These principles are used in the development of all current and future applications.

1. Automation of application, service and network slice management
2. Distributed cloud deployment of applications, services and network slices
3. Portability across execution environments
4. Separation of business logic and application data/state
5. Microservice modularity
6. Fault tolerance in applications and services

Automation of application, service and network slice management

5G systems with network slicing enable networks to be tailored to different market segments, industries, services, and enterprises. Cloud based systems provide the flexibility needed for diverse requirements with respect to characteristics, services and deployment. However, this flexibility requires a high level of automation to provide a cost-efficient solution.

The scope of the automation covers the entire lifecycle, including onboarding, instantiation, scaling, upgrade, backup and restore, acceptance tests and termination. Closed-loop automation is achieved through a combination of control, orchestration, management, policy and analytics (COMPA), where each domain (e.g. network applications) has its own functionality and where cross-domain tools are used to automate on a higher level. Machine learning is important to create and fine-tune policies for automated decision-making.

Ericsson's network applications are designed for automated lifecycle management on the application, service and network slice levels. Model-based management, based on standards like YANG and TOSCA, is used to specify applications and services as well as to facilitate integration with external management systems.

Distributed cloud deployment of network services and applications

Distributed cloud environments provide specific advantages in the context of 5G. The physical separation between datacenters enables geographical network redundancy for critical services. Distribution of user plane functions allows for local breakout of traffic, which lowers costs and improves service characteristics. Latency-critical services, e.g. critical machine-type communication, benefit from a distributed deployment of network applications close to the customer.

Many operators are building or planning distributed cloud infrastructures that often utilize assets in existing network topologies. Networks are likely to evolve as more diverse services and use cases emerge. Harmonized execution environments and orchestration systems allow for flexible deployment of network applications that are scaled independently and distributed across the network.

> "Distributed cloud environments provide specific advantages in the context of 5G."

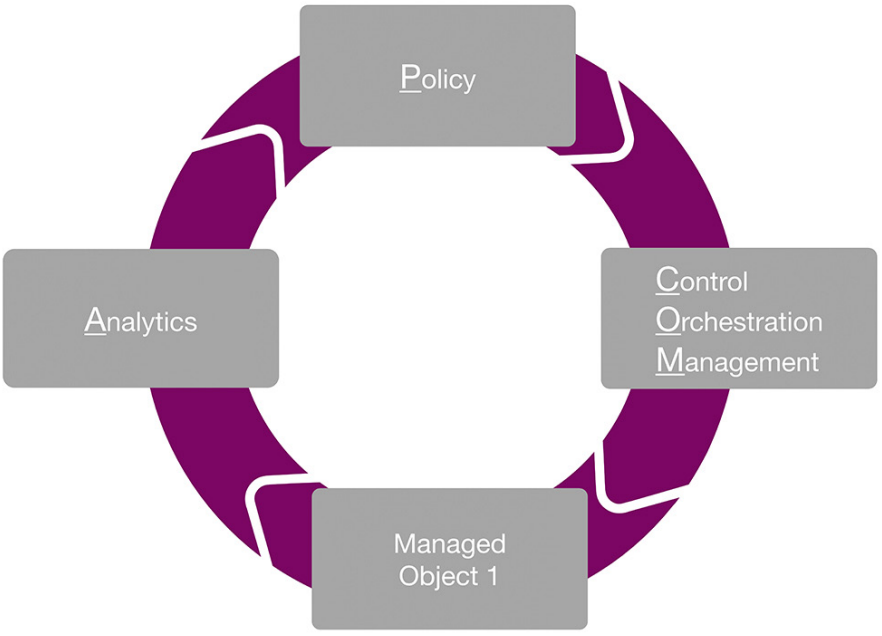

Figure 1: Closed-loop automation via COMPA

A typical example is network slice elasticity. Closed-loop automation can scale out an application, but can also spin up one or more application instances to adjust the network slice capacity. In the latter scenario, automation policies consider both network and data-center topologies in order to create the application in the right place. If application instances benefit from being close to the end user, they are instantiated in a local (regional) datacenter. If not, it is better to optimize for scale and resource utilization and deploy the application(s) in a national (global) datacenter.

Network applications are decomposed into microservices to maximize re-use and development speed. This modularity creates opportunities for flexible deployment, e.g. in the separation of control and user planes or between business logic and application data. In all cases, the service performance needs to be considered to find the best possible deployment. Real-time sensitive applications that need fast access to other applications or data should be co-located while others can tolerate geographic separation.

Ericsson's network applications are designed for distributed cloud infrastructures. Management, performance and modularity are done so that software can be flexibly deployed across network topologies depending on the use case and service requirements.

Portability across execution environments

Cloud systems and cloud execution environments (virtual machines (VM), containers, bare metal, etc.) are not created equal. Each have their own feature set, virtues and disadvantages. VMs can host a dedicated operating system at the expense of resource consumption and start-up times. Containers provide a fast and efficient method for packaging user-space applications but offer fewer capabilities for dedicating resources to workloads. Bare metal has the least flexibility from an automation perspective but offers rewards in terms of performance. There are new emerging technologies that introduce new execution environments (e.g. micro-kernels).

> "... the choice of execution environment also depends on the desired level of automation"

> "Network applications are decomposed into microservices to maximize re-use and development speed."

At some point in the future, the industry may have a cloud environment where automation capabilities of different execution environments are standardized. Until then, the choice of execution environment also depends on the desired level of automation.

Ericsson's network applications are portable between execution environments. This means that the trade-off between characteristics and automation capabilities is a choice that can be made by the service provider.

Separation of business logic and application data/state

Cloud-optimized network applications use separation of business logic and data/state storage. This creates the opportunity to distribute data/state in a more flexible way. The business logic does not have to deal with the complexity of state replication since that is managed by the state repository. User plane performance is optimized by making state available on-demand where the traffic hits the system. Co-locating data with processing can be a conscious choice and not one dictated by necessity.

Data separation from business logic is done considering performance and characteristics of the end-user service. These characteristics become more diverse in IoT and 5G industry applications. The application data belongs to different parts of a 5G cloud data layer, depending on the requirements and characteristics of each application and the use cases. Some applications require high-resilience and geographically redundant data management, e.g. for semi-static subscriber and user data. Other applications use real-time-sensitive and short-lived data (service or activity related) that is stored or cached close to the business logic.

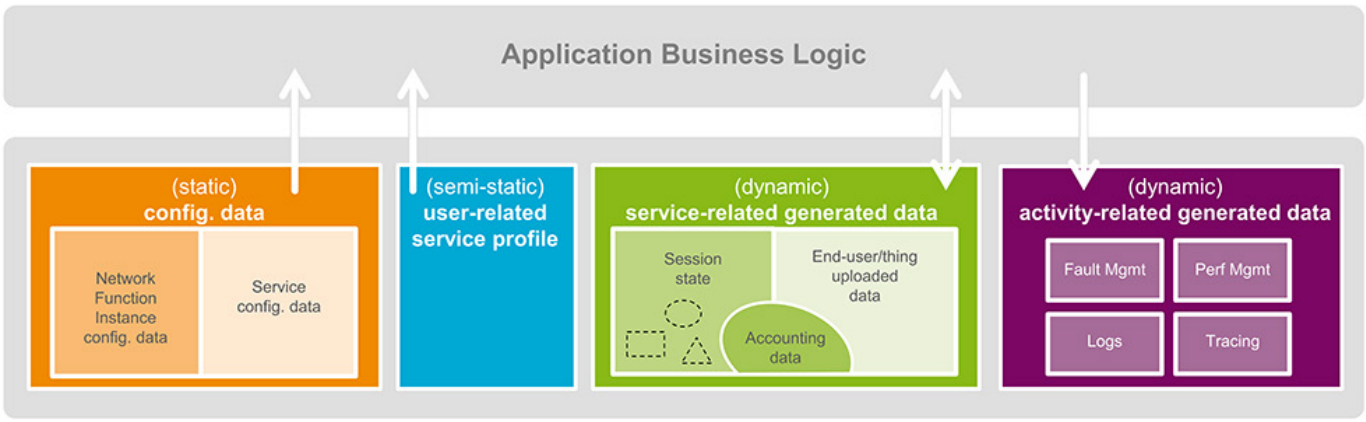

Figure 2: State and business logic separation principles

Ericsson's network applications have separated business logic from application data and are managed in one or several cloud data layers. The software architecture and deployment modules are optimized for performance while meeting telecom-grade requirements for resilience and fault tolerance.

Microservice modularity

Applications that are decomposed into microservices are faster to develop and easier to maintain and deploy. A microservice architecture allows for re-use of components and services between applications.

Microservices are a key enabler for agile software development. Comprehensive development efforts are distributed among many, small, independent teams that each have responsibility for the design lifecycle, from concept to testing. A team that has the responsibility for new functionality will carry out the necessary changes in existing microservices or create new ones if needed. A continuous integration activity with automated testing ensures the integrity, functionality and characteristics of the system at all times.

> "The separation of business logic from data allows for a smooth transition to Ericsson's multi-tiered 5G cloud data layer, with a fully preserved, proven feature set."

The challenge is to prevent an application that is comprised of a large number of microservices from creating complexity during deployment and operation. A large number of deployable modules can quickly become cumbersome if the models required to describe the application and its automation workflows are complex.

Ericsson separates the development architecture, where there are no limits to disaggregation and modularity, from the deployment architecture, which exposes the right level of APIs for easy customization and efficient lifecycle management.

Fault tolerance in applications and services

The mean time between failure (MTBF) of hardware has been one of the most important metrics for the resiliency architecture of physical network functions. Because of the long MTBF of hardware, only support for single HW failures has been required. Virtual machines and containers may have a considerably shorter MTBF. A different resiliency architecture is needed for cloud-optimized network applications.

Fault tolerance is achieved on both the application and service levels. The application is available as long as there are single instances of a network application running. In the unlikely event of an entire network application being lost, another instance is stand-by or can be made dynamically available to take over traffic.

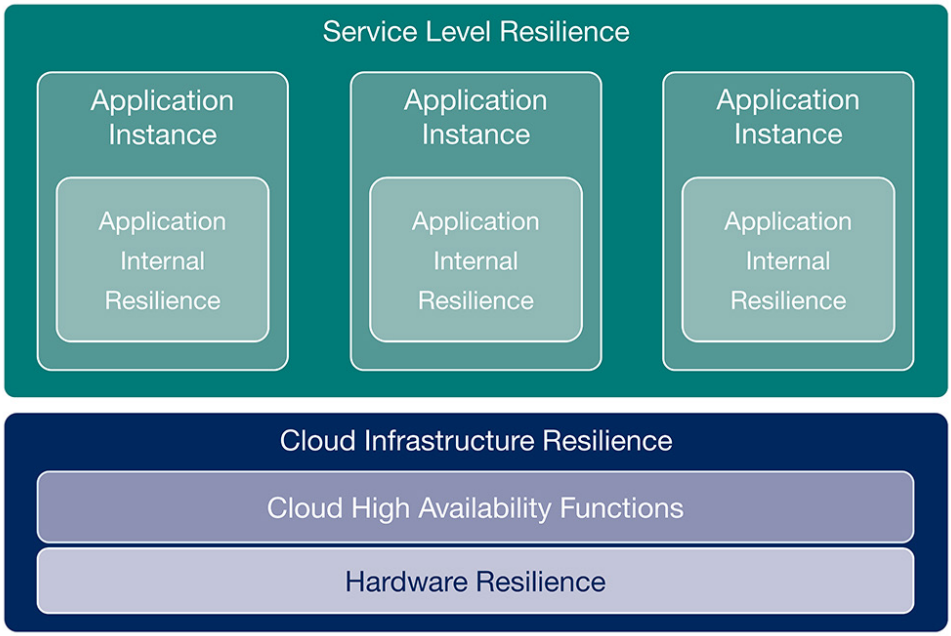

Figure 3: Resilience on different levels

Ericsson's network applications are designed for resiliency and fault tolerance on multiple levels, based on cloud infrastructure characteristics. This service level resilience results in an architecture that can withstand failures on any level without being dependant on long MTBFs for any single component.

SUMMARY

Network requirements are changing as new services develop toward 5G use cases. The combination of distributed-cloud infrastructure and dynamically created network slices generates opportunities for more flexibility for addressing new business opportunities and increased operational efficiency. The cloud environment also creates challenges for applications, especially in a demanding telecom environment where requirements for performance, predictability and resilience are high.

Ericsson's cloud-optimized network applications are designed according to the principles described in this article. These principles are used to optimize development and deployment efficiency for Ericsson and our customers. They also ensure consistent high performance and manageability across the Ericsson portfolio of network applications.

Ericsson AB
SE-126 25 Stockholm, Sweden
Telephone +46 10 719 00 00
Fax +46 8 18 40 85
www.ericsson.com

3/287 01-FGB 101 0325