ETSI GS NFV-REL 002 V1.1.1 (2015-09)

GROUP SPECIFICATION

Network Functions Virtualisation (NFV); Reliability; Report on Scalable Architectures for Reliability Management

ETSI

650 Route des Lucioles
F-06921 Sophia Antipolis Cedex - FRANCE

Tel.: +33 4 92 94 42 00   Fax: +33 4 93 65 47 16

Siret N° 348 623 562 00017 - NAF 742 C
Association à but non lucratif enregistrée à la
Sous-Préfecture de Grasse (06) N° 7803/88

ETSI

Contents

Intellectual Property Rights

IPRs essential or potentially essential to the present document may have been declared to ETSI. The information pertaining to these essential IPRs, if any, is publicly available for **ETSI members and non-members**, and can be found in ETSI SR 000 314: *"Intellectual Property Rights (IPRs); Essential, or potentially Essential, IPRs notified to ETSI in respect of ETSI standards"*, which is available from the ETSI Secretariat. Latest updates are available on the ETSI Web server (http://ipr.etsi.org).

Pursuant to the ETSI IPR Policy, no investigation, including IPR searches, has been carried out by ETSI. No guarantee can be given as to the existence of other IPRs not referenced in ETSI SR 000 314 (or the updates on the ETSI Web server) which are, or may be, or may become, essential to the present document.

Foreword

This Group Specification (GS) has been produced by ETSI Industry Specification Group (ISG) Network Functions Virtualisation (NFV).

Modal verbs terminology

In the present document "**shall**", "**shall not**", "**should**", "**should not**", "**may**", "**need not**", "**will**", "**will not**", "**can**" and "**cannot**" are to be interpreted as described in clause 3.2 of the ETSI Drafting Rules (Verbal forms for the expression of provisions).

"**must**" and "**must not**" are **NOT** allowed in ETSI deliverables except when used in direct citation.

1   Scope

The present document describes a study of how today's Cloud/Data Centre techniques can be adapted to achieve scalability, efficiency, and reliability in NFV environments. These techniques are designed for managing shared processing state with low-latency and high-availability requirements. They are shown to be application-independent that can be applied generally, rather than have each VNF use its own idiosyncratic method for meeting these goals. Although an individual VNF could manage its own scale and replication, the techniques described here require a single coherent manager, such as an orchestrator, to manage the scale and capacity of many disparate VNFs. Today's IT/Cloud Data Centres exhibit very high availability levels by limiting the amount of unique state in a single element and creating a virtual network function from a number of small replicated components whose functional capacity can be scaled in and out by adjusting the running number of components. Reliability and availability for these type of VNFs is provided by a number of small replicated components. When an individual component fails, little state is lost and the overall VNF experiences minimal change in functional capacity. Capacity failures can be recovered by instantiating additional components. The present document considers a variety of use cases, involving differing levels of shared state and different reliability requirements; each case is explored for application-independent ways to manage state, react to failures, and respond to increased load. The intent of the present document is to demonstrate the feasibility of these techniques for achieving high availability for VNFs and provide guidance on Best Practices for scale out system architectures for the management of reliability. As such, the architectures described in the present document are strictly illustrative in nature.

Accordingly, the scope of the present document is stated as follows:

- Provide an overview of how such architectures are currently deployed in Cloud/Data Centres.

- Describe various categories of state and how scaling state can be managed.

- Describe scale-out techniques for instantiating new VNFs in a single location where failures have occurred or unexpected traffic surges have been experienced. Scale-out may be done over multiple servers within a location or in a server in the same rack or cluster within any given location. Scaling out over servers in multiple locations can be investigated in follow-up studies.

- Develop guidelines for monitoring state such that suitable requirements for controlling elements (e.g. orchestrator) can be formalized in follow-up studies.

2   References

2.1   Normative references

References are either specific (identified by date of publication and/or edition number or version number) or non-specific. For specific references, only the cited version applies. For non-specific references, the latest version of the referenced document (including any amendments) applies.

Referenced documents which are not found to be publicly available in the expected location might be found at http://docbox.etsi.org/Reference.

NOTE: While any hyperlinks included in this clause were valid at the time of publication, ETSI cannot guarantee their long term validity.

The following referenced documents are necessary for the application of the present document.

Not applicable.