

CSCF Scaling Management

Call Session Control Function

DESCRIPTION

Copyright

© Ericsson AB 2018. All rights reserved. No part of this document may be reproduced in any form without the written permission of the copyright owner.

Disclaimer

The contents of this document are subject to revision without notice due to continued progress in methodology, design and manufacturing. Ericsson shall have no liability for any error or damage of any kind resulting from the use of this document.

Trademark List

All trademarks mentioned herein are the property of their respective owners. These are shown in the document Trademark Information.



Contents

1	Understanding Scaling Management	1
1.1	Key Scaling Management Concepts	2
2	Basic Scaling Management Operations	5





1 Understanding Scaling Management

Scaling Management provides a management interface to configure the following on the Managed Element (ME):

- Increase capacity, or scale-out, that is, adding one or more Virtual Machines (VMs) when a cluster requires more processing resources.
- Decrease capacity, or graceful or automatic scale-in, that is, removing one or more VMs when a cluster no longer requires as many processing resources. Released resources can be allocated again when more processing resources are required.

The Scaling Management managed area can be found under the CrM Managed Object Class (MOC) in the Managed Object Model (MOM). For general information about the MOM, MOCs, cardinality, and related concepts, see [Managed Object Model User Guide](#).

When the system starts a scaling operation, it enters in the Maintenance Mode, meaning the overload regulation is lowered to the vDicos initial configuration parameter `LOAD_REG_MAINT_LIMIT`.

The scaling operation involves planned reconfiguration of distribution units. This activity is performed in the quickest possible manner with high priority. Hence, load regulation-related alarms can appear during scaling operation. Such alarms are not expected to be present for longer than 2–4 seconds. The effect is minimal on traffic handling capability.

A scale-out operation is performed by adding one or more VMs to the Virtual Network Function (VNF) cluster, see Section 1.1.1 Auto Scale-Out on page 2. For graceful scale-in, the VNF cluster reallocates the resources from VMs to be scaled-in and moves to other VMs to prevent data loss, see Section 1.1.2 Graceful Scale-In on page 2. Performance counters can be used as input to decide which scaling operation is to be performed.

A VM is assigned a role with an attribute that describes the scaling behavior:

- Non-Scalable

When a VM is allocated to a role with the attribute `scalability=NON_SCALABLE`, it cannot be scaled in or scaled out, as it has a system-defined size and a system-defined role. For example, a common configuration is to have two VMs, usually named SC-1 and SC-2, allocated to the Operation and Maintenance (O&M) role within the VNF. These VMs cannot be scaled in or scaled out.

- Scalable

When a VM is allocated to a role with the attribute `scalability=SCALABLE`, it has only one role called Default-Role instead of a specialized role like, for example, SC-1 or SC-2. This means that VMs with the attribute set to

SCALABLE can be scaled in or scaled out, depending on the capacity needs of the VNF.

1.1 Key Scaling Management Concepts

1.1.1 Auto Scale-Out

Auto Scale-Out is an operation where one or more new compute resources are launched, see Figure 1. The system automatically detects, configures, and brings up the nodes as a member of the scaling domain of the cluster. See Figure 2 for an example where one new compute node is added to the cluster.

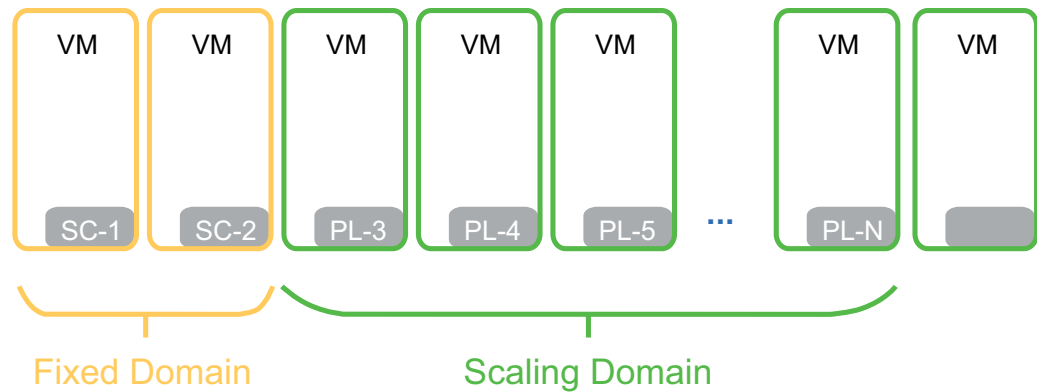


Figure 1 A New Compute Resource Is Spawned and Available

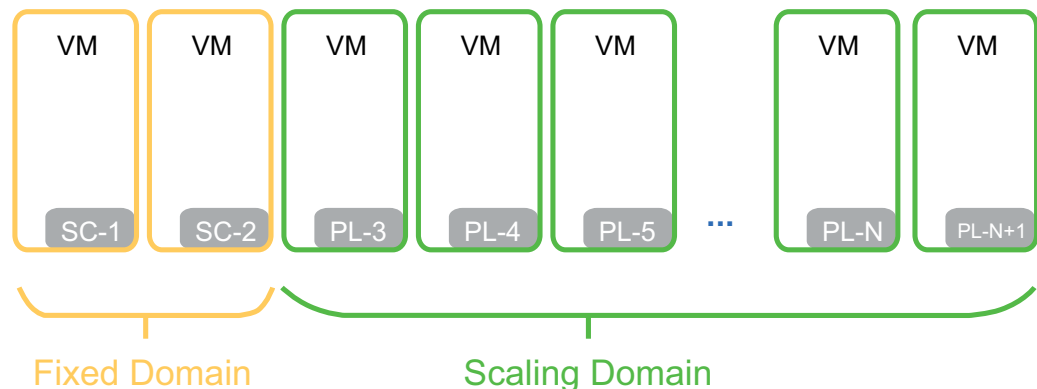


Figure 2 After Auto Scale-Out, a New Resource Is Added to the Cluster

1.1.2 Graceful Scale-In

Graceful Scale-In is an operation where one or more compute resources, part of the scaling domain of the cluster (see Figure 3) are removed from the cluster (see Figure 4) to free up resources.

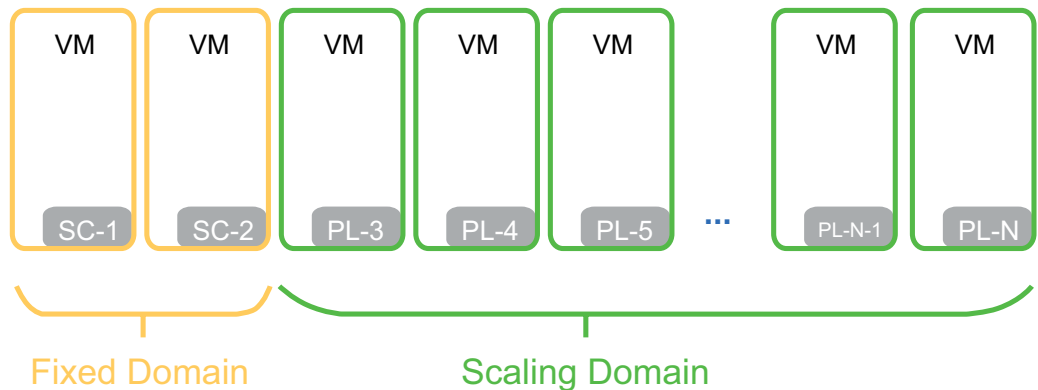


Figure 3 The Node Named PL-(N-1) Is Part of the Cluster

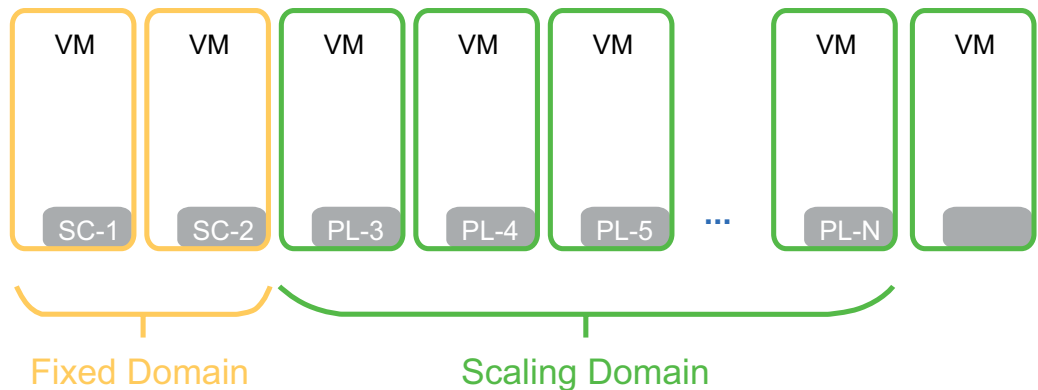


Figure 4 The Node Named PL-(N-1) Is Removed from the Cluster and Its Resources Can Be Released

Note: The Graceful Scale-In operation can be rejected by the cluster if, according to the automatic estimation of the system, the target size of the cluster does not have the memory resources to serve the needed memory capabilities for the ongoing traffic.

1.1.3

Forceful Scale-In

Forceful Scale-In is, similarly to Graceful Scale-In, an operation to remove one or more nodes from the scaling domain of the cluster. The only difference is that in this case, either the node is not available (see Figure 5) or scale-in with potential traffic loss is acceptable. If the node is not available, it can be either because it already freed up its resources or because of a failure. Therefore, the removal is only an administrative operation, see Figure 6.

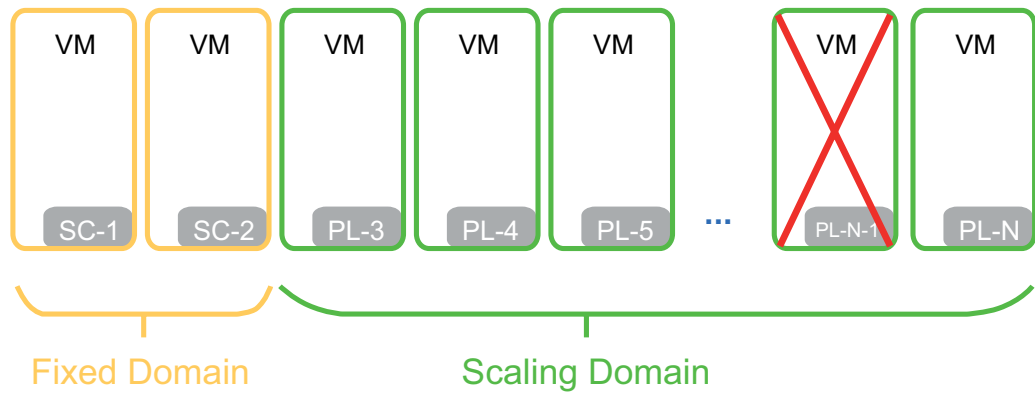


Figure 5 The Node Named PL-(N-1) in the Cluster Scaling Domain Is Unavailable

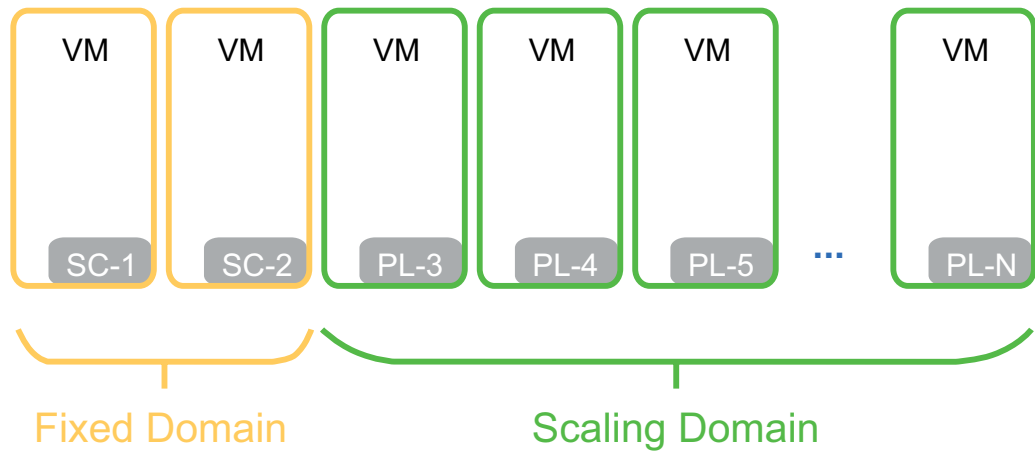


Figure 6 The Node Named PL-(N-1) Is Removed Administratively from the Cluster



2 Basic Scaling Management Operations

Scaling Management is accessed using NETCONF or the Ericsson Command-Line Interface (ECLI) to manipulate the Management Information Base (MIB). The following operations can be performed:

- Manage Scaling Manually, see [Increase Capacity Manually](#) to add nodes to the CSCF cluster, and [Decrease Capacity Manually](#) to remove nodes from the CSCF cluster.
- Manage Scaling with Heat Orchestration, see [Increase Capacity with Heat Orchestration](#) to add nodes to the CSCF cluster, and [Decrease Capacity with Heat Orchestration](#) to remove nodes from the CSCF cluster.
- Manage Scaling with VNF-LCM, see [CSCF VNF Lifecycle Management](#).