

MTAS Scaling Management

MTAS

DESCRIPTION

Copyright

© Ericsson AB 2018, 2019. 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	1
2	Basic Scaling Management Operations	5





1 Understanding Scaling Management

1.1 Key Scaling Management Concepts

The term scaling refers to the scalability of the system, which is provided by multiple instances to distribute the load in parallel for having the capacity needed.

The following terms are used:

Node	Refers to a compute resource and can be a physical hardware blade or a virtual machine (VM) instantiation.
Fixed Domain	The set of nodes that cannot be subject of a scaling operation. The fixed domain of MTAS consists of SC-1 and SC-2 nodes permanently. The domain cannot be changed.
Scaling Domain	The set of nodes that can be subject of a scaling operation. MTAS scaling domain consists of all traffic nodes (PL-3, PL-4, PL-5 ... PL-N).

The PL-3 and PL-4 nodes are not scalable; even though PL-3 and PL-4 nodes are considered to be part of the scaling domain, they cannot be scaled in.

Traffic Handling: 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 time than 2–4 seconds. The effect is minimal on traffic handling capability.

1.1.1 Auto Scale-Out

Auto Scale-Out is an operation when 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 when one new compute node is added to the cluster.

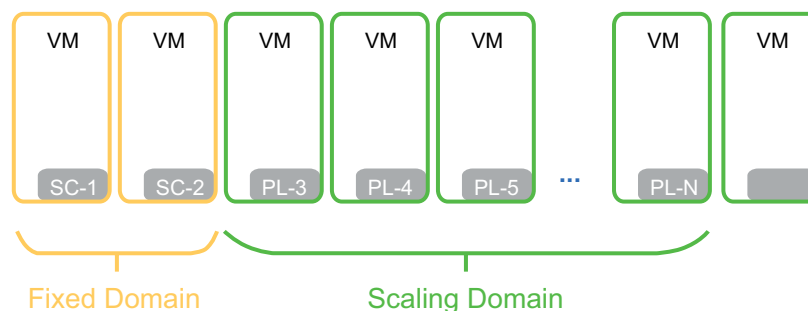


Figure 1 New Compute Resource Spawning and Available

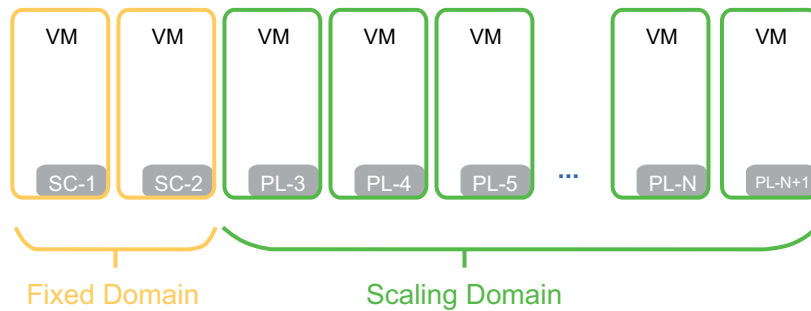


Figure 2 After Auto Scale-Out New Resource Is Added to Cluster

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.

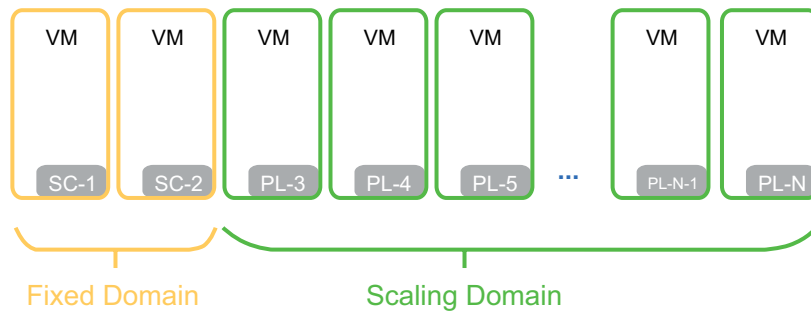


Figure 3 Node Named PL-(N-1) Is Part of Cluster

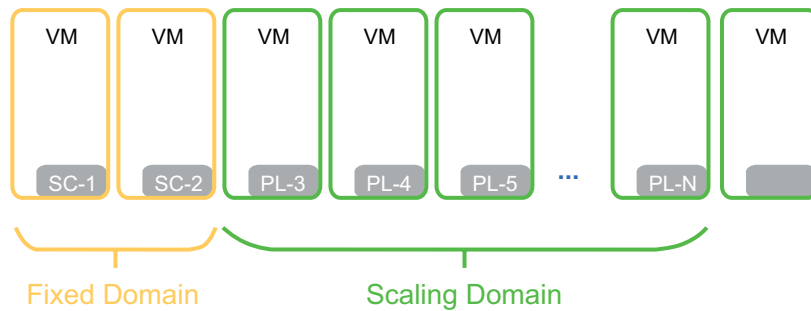


Figure 4 Node Named PL-(N-1) Is Removed from Cluster and Its Resources Can Be Released

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, the node is not available (see Figure 5) either because it already freed up its resources or because of a failure. Therefore the removal is only an administrative operation, see Figure 6.

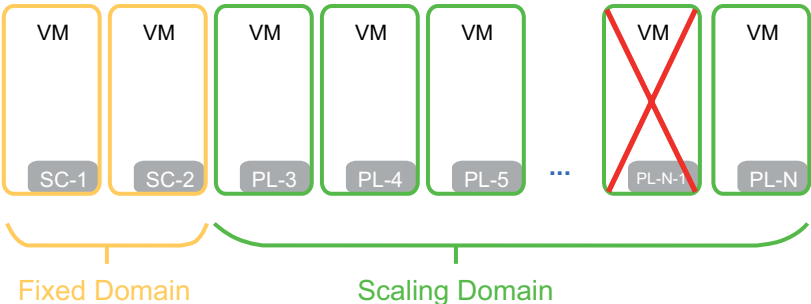


Figure 5 Node Named PL-(N-1) in the Cluster Scaling Domain Is Unavailable

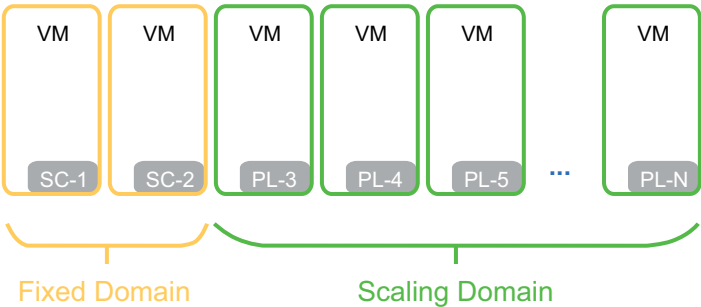


Figure 6 Node Named PL-(N-1) Is Removed Administratively from Cluster





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 by the user and are described in Operating Instructions:

- Manage Scaling Manually; see [Manually Scale Out Cluster](#) to add a node to the cluster (scale out), and [Manually Scale In Cluster](#) to remove nodes from the cluster (scale in).
- Manage Scaling with Heat Orchestration; see [Scale Out Cluster Using Heat Orchestration](#) to add a node to the cluster, and [Scale In Cluster Using Heat Orchestration](#) to remove nodes from the cluster.

Note: There is no direct connection between MTAS and OpenStack. Therefore, the name (number) of the VM present in OpenStack differs from the ComputeResource present in MTAS. To correlate a compute resource with a VM, use the Universally Unique Identifier (UUID).

- Manage Scaling with VNF-LCM, refer to [MTAS VNF Lifecycle Management](#). Do not use manual procedures when scaling is managed by VNF-LCM.