
# BalticLSC Platform Component Selection Report

The selection of components used to build BalticLSC Platform

Version 1.00


Priority 1: Innovation

Warsaw University of Technology, Poland
RISE Research Institutes of Sweden AB, Sweden
Institute of Mathematics and Computer Science, University of Latvia, Latvia
EurA AG, Germany
Municipality of Vejle, Denmark
Lithuanian Innovation Center, Lithuania
Machine Technology Center Turku Ltd., Finland
Tartu Science Park Foundation, Estonia

# BalticLSC Platform Component Selection Report

## The selection of components used to build BalticLSC Platform

| | |
|---|---|
| Work package | WP4 |
| Task id | A4.2 |
| Document number | O4.2 |
| Document type | Component selection document |
| Title | BalticLSC Platform Component Selection Report |
| Subtitle | The selection of components used to build BalticLSC Platform |
| Author(s) | Daniel Olsson (RISE), Filip Blylod (RISE), Magnus Nilsson-Mäki (RISE) |
| Reviewer(s) | Agris Šostaks (IMCS), Radosław Roszczyk (WUT) |
| Accepting | Michał Śmiałek (WUT) |
| Version | 1.0 |
| Status | **Final version** |

# History of changes

| Date | Ver. | Author(s) | Change description |
|---|---|---|---|
| 20.06.2019 | 0.01 | Daniel Olsson (RISE) | Document creation and initial contents |
| 26.06.2019 | 0.02 | Magnus Nilsson-Mäki (RISE) | Added production cluster topology |
| 27.06.2019 | 0.03 | Daniel Olsson (RISE) | Added more content |
| 12.12.2019 | 0.1 | Daniel Olsson (RISE) | Finalized document with content from discussions with parties doing procurement as well as review inputs. |
| 16.12.2019 | 0.11 | Daniel Olsson (RISE) | Resolved comments by reviewer (Agris). |
| 04.02.2020 | 1.00 | Daniel Olsson (RISE) | Final version |

# Executive summary

This document contains information on which technologies (computing and networking hardware, operating system software) should be used to develop the BalticLSC platform. Also, it will propose optimal solutions for combining selected hardware units into coherent computation grids, and networking solutions that would allow combining such small grids into large computation networks on the transnational level.

# Table of Contents

# 1. Introduction

## 1.1 Objectives and scope

The BalticLSC Platform is where computation tasks compiled by the BalticLSC Software are to be executed. The scope of this document is to specify requirements on hardware and operating systems to build a cluster running the BalticLSC Platform software.

## 1.2 Relations to other documents

This document gives guidelines and suggestions when choosing hardware to build the platform described in the BalticLSC Platform Vision: O4.1 document.

## 1.3 Intended audience and usage guidelines

This document is intended for internal use within BalticLSC consortium.

# 2. Requirements and recommendations

This section suggests two different configurations. One for small-scale and development purposes with low hardware requirements, and one for large-scale production environments with higher requirements. A small development configuration can be a single machine. The minimal recommended production cluster should consist of a 9+ servers to be able to provide high-availability.

## 2.1 Production cluster configuration

The production cluster consists of two Kubernetes clusters. One running Rancher and BalticLSC Platform components, called the Management Cluster. The other cluster is where all workloads are executed, called the Compute Cluster. The reason for running Rancher in it's own cluster is to provide high-availability.

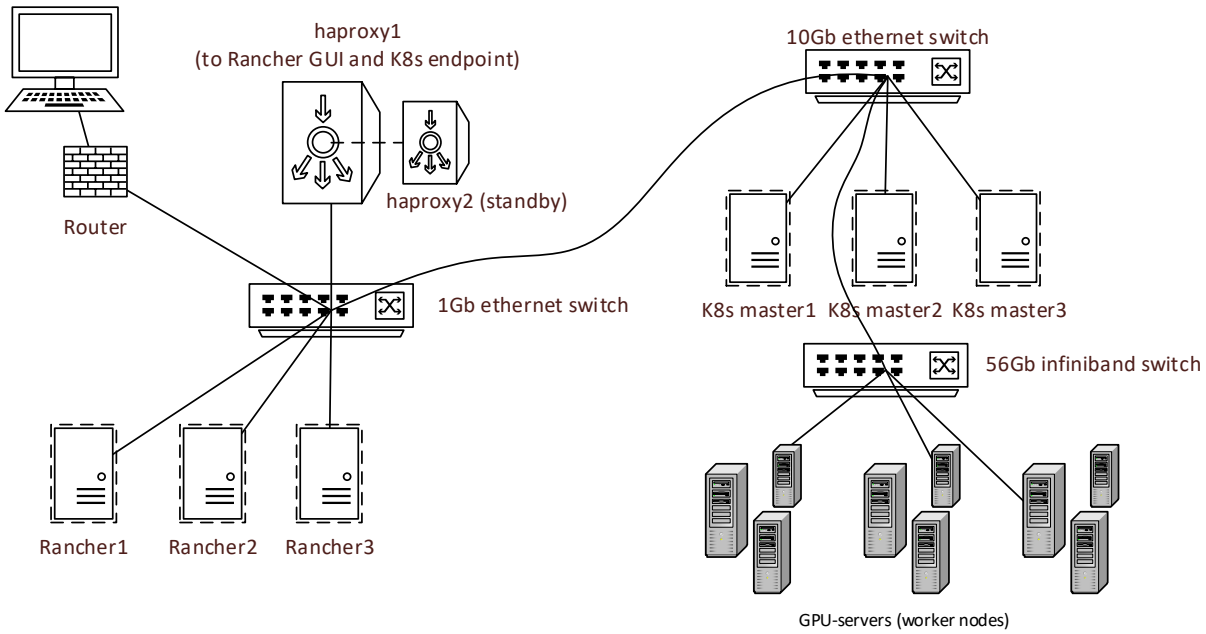


*Figure 1: Production cluster topology*

Imagine that the figure above is splitted vertically in the middle. Then the left side represents the Management Cluster with Rancher and BalticLSC Platform components. Everything to the right is then the Compute Cluster.

### 2.1.1 Management cluster

Rancher is running in its own Kubernetes cluster of three nodes. To provide load-balancing to the rancher nodes, a HA proxy can be used. This should also be configured in a highly available configuration. Details of this is omitted in this document. BalticLSC Platform specific components is also running in this cluster.

Hardware requirements on management nodes scale based on the total size of the clusters managed by Rancher. Provision each individual node according to the requirements in following table (taken from Rancher website[1]). If nodes are VMs, they need to run on separate physical hosts.

| Deployment size | Clusters | Nodes | vCPUs | RAM |
|---|---|---|---|---|
| Small | Up to 5 | Up to 50 | 2 | 8 GB |

[1] https://rancher.com/docs/rancher/v2.x/en/installation/requirements/


EUROPEAN REGIONAL DEVELOPMENT FUND
EUROPEAN UNION

| | | | | |
|---|---|---|---|---|
| Medium | Up to 15 | Up to 200 | 4 | 16 GB |
| Large | Up to 50 | Up to 500 | 8 | 32 GB |
| X-Large | Up to 100 | Up to 1000 | 32 | 128 GB |

Table 1: Rancher node requirements in high-availability configuration

### 2.1.2 Compute cluster

It is recommended to use Rancher to install the compute cluster. The compute cluster consists of at least three master nodes and one or more worker nodes. Three nodes is needed for redundancy.

Following server specifications can be used as inspiration when writing procurements.

**Master node / Low performance compute node**

- 1U server
- 2x CPU
- 32GB RAM
- SSD
- 10Gb Ethernet

**Medium performance compute node**

- 2U server
- 2 x CPU
- 128 GB RAM
- 1 x Nvidia GPU
- SSD
- 10Gbps Ethernet

**High-performance GPU compute node**

- Supermicro 4029GP-TRT3 server chassis
- Single root PCIe architecture
- 2 x Intel Xeon Gold 6130
- 256 GB RAM
- 4TB SSD
- 8 x Nvidia GTX 2080ti
- Infiniband 56 Gbps

### 2.1.3 Operating system requirements

Rancher is tested on the following operating systems and their subsequent non-major releases with a supported version of Docker[2].

- Ubuntu 16.04 (64-bit x86)
- Docker 17.03.x, 18.06.x, 18.09.x
- Ubuntu 18.04 (64-bit x86)
- Docker 18.06.x, 18.09.x
- Red Hat Enterprise Linux (RHEL)/CentOS 7.6 (64-bit x86)
- RHEL Docker 1.13
- Docker 17.03.x, 18.06.x, 18.09.x

[2] https://docker.com

- RancherOS 1.5.1 (64-bit x86)
- Docker 17.03.x, 18.06.x, 18.09.x

The recommendation is to run Ubuntu or CentOS because that is what we are experienced with. This concerns both Rancher and Compute nodes.

## 2.2 Development/minimal cluster configuration

For development and test there are almost no requirements on the hardware. It is possible to run everything on a standard consumer laptop. When running on laptop, and a cluster of several nodes is wanted, then each node should run in its own virtual machine. Choose your favorite hypervisor for running VMs, like KVM, VirtualBox or VMWare.

There are other more lightweight Kubernetes alternatives like Lightweight Kubernetes[3] which has very low system requirements. It has following minimum system requirements which means that it can even run on a Raspberry Pi.

- 512 MB of RAM per server
- 75 MB of RAM per node
- 200 MB of disk space

[3] https://k3s.io/

# 3. Selection guidelines

Clearly, the amount of hardware to choose from are mind boggling. To choose the right combination of hardware components that will work well together, have good performance and not being too expensive is not easy. In this section we list some guidelines to help navigating the hardware jungle.

## 3.1 Network

In the Rancher cluster, 1Gb Ethernet is sufficient. But for the compute cluster, network connectivity performance between compute nodes is critical. At least 40/56 Gbps Infiniband is recommended which is the most cost-effective network technology. Network redundancy is not required.

## 3.2 CPU vs Memory

Total server hardware cost when deciding between higher performing CPU versus the presence of more memory modules in the server was considered. As earlier usage of similar hardware in the facility has proven. The decision for a higher performance CPU was chosen. This of course depends on what type of calculations will be performed by the cluster. More RAM is also a good choice if the majority of calculations are RAM demanding.

## 3.3 Nvidia vs AMD

Nvidia has been the number one leader in the realm of HPC with its CUDA driver. However, AMD is a catching up with their GPUs and the corresponding ROCm driver. Early tests by Jim Dowling [4]have shown that you will get better performance for the money by going the AMD route. The tests have also shown that the driver is still a bit shaky, but improving for every release. The ROCm driver is open-source which gives the community the ability to fix and improve it. CUDA is proprietary and comes with a license[5] prohibiting usage of consumer (GeForce) GPUs in datacenters. For some reason it is okay to use for mining which is strange. Probably Nvidia will modify this license when AMD starts picking market shares in this realm.

Our recommendation is to go with Nvidia because of current stability issues. If going with AMD it is suggested to choose one of the supported GPUs found in this list https://github.com/RadeonOpenCompute/ROCm#supported-gpus. Also do some research of the current state of the ROCm driver. Buy one or two cards and do some testing.

## 3.4 Storage types

To ensure optimal speed, we recommend using NVMe SSD disks. An option is to use fast storage for metadata and caching and normal HDDs as the end storage.

## 3.5 Server manufacturer

The server manufacturer Supermicro has been chosen as the hardware-platform of choice due to cost / performance ratio and equipment familiarity. The hardware has proven easy to manage and use in the past and has no proprietary limitations for open source software and compatibility with the GPU and network requirements. The TRT2 generation or newer is preferred because it has a single root PCIe architecture which is preferred for optimal GPU to GPU communication performance. The alternative is dual root.

[4] https://youtu.be/neb1C6JlEXc

[5] https://www.nvidia.com/content/DriverDownload-March2009/licence.php?lang=us&type=GeForce

## 3.6 UPS – Uninterruptible Power Supply

Because BalticLSC network is distributed, UPS is not mandatory. If the cluster running compute tasks shuts down, the compute tasks should be restarted on another cluster.

## 3.7 PCIe 3.0 or PCIe 4.0

Generally newer is better, but it is also a question of price. If the selected GPUs does not require PCIe 4.0, then it may be sufficient to use a more inexpensive server offering only PCIe 3.0. However, if the server supports PCIe 4.0, then the latest and future GPUs will be supported.