Project Title	SCAlable LAttice Boltzmann Leaps to Exascale
Project Acronym	SCALABLE
Grant Agreement No.	956000
Start Date of Project	01.01.2021
Duration of Project	36 Months
Project Website	www.scalable-hpc.eu

D6.2
Report on deployment of SCALABLE codes on peta-scale and pre-exascale systems [M18]

Work Package	**WP 6, Development and Deployment Platforms and Services**
Lead Author (Org)	**IT4I**
Contributing Author(s) (Org)	**FZJ, CERFACS**
Due Date	**01.07.2022 (M18)**
Date	**01.07.2022**
Version	**V1.0**

Dissemination Level

X	PU: Public
	PP: Restricted to other programme participants (including the Commission)
	RE: Restricted to a group specified by the consortium (including the Commission)
	CO: Confidential, only for members of the consortium (including the Commission)

The SCALABLE project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 956000.

Versioning and contribution history

Version	Date	Author	Notes
0.1	27.5.2022	Lubomir Riha (IT4I)	Initial skeleton of the document
0.2	6.6.2022	Lubomir Riha, Ondrej Vysocky, Radim Vavrik (IT4I)	First draft of the document
0.3	10.6.2022	Lubomir Riha, Ondrej Vysocky, Radim Vavrik (IT4I), Jayesh Badwaik (FZJ), Marcus Holzer (CERFACS)	Draft prepared for review
0.4	26.6.2022	Gabriel Staffelbach (CERFACS), Cuidard Romain (CSGROUP), Harald Koestler (FAU)	Internal review
1.0	1.7.2021	Lubomir Riha, Ondrej Vysocky (IT4I)	Final version

Disclaimer

This document contains information which is proprietary to the SCALABLE Consortium. Neither this document nor the information contained herein shall be used, duplicated or communicated by any means to a third party, in whole or parts, except with the prior consent of the SCALABLE Consortium.

The SCALABLE project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 956000.

Table of Contents

The SCALABLE project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 956000.

TERMINOLOGY

Terminology/Acronym	Description
SCALABLE	SCAlable LAttice Boltzmann Leaps to Exascale
HPC	High-Performance Computing
EPI	European Processor Initiative
PRACE	Partnership for Advanced Computing in Europe
HDEEM	High-Definition Energy Efficiency Monitoring
TEP	The European Pilot
EUPEX	EUropean Pilot for Exascale

The SCALABLE project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 956000.

Executive Summary

The goal of this deliverable is to report the status of work related to the efficient deployment of the SCALABLE codes on the peta-scale and the pre-exascale systems.

The supercomputing centers IT4I and FZJ are continuously deploying the newly developed versions of WaLBerla and ProLB codes on their peta-scale systems throughout the lifetime of the project. They also provide support for optimal code and tool compilation including environment setup for (i) particular CPU architectures, (ii) GPU accelerators (as of now only Nvidia GPUs), (iii) network topology and communication libraries (MPI, UCX, GPU-Direct, etc.). This work is mainly needed by both WP2 and WP4.

In addition, in collaboration with local infrastructure support teams, the IT4I team enables the hardware parameter tuning and energy consumption monitoring on their both major production systems (Barbora and newly also Karolina) which is essential for work in WP2. On the Karolina system, we are now able to tune both CPU and GPU parameters as well as measure their energy consumption. These features will be used in D2.3.

As of now, WP6 secured access to 4 peta-scale systems, that are used by all WPs dealing with code development and optimization. These systems are: Barbora (IT4I), Karolina (IT4I), JUWELS Cluster (FZJ) and JUWELS Booster (FZJ). We have used several mechanisms including IT4I Open-access, PRACE Preparatory access, EuroHPC code development access and GCS Compute Access.

Finally, there is only one EuroHPC pre-exascale machine that is partially available to the general public at the time of submission of this report. This is the LUMI machine and, the non-accelerated LUMI-C partition to which access has been secured as well.

1 Introduction

The goal of this deliverable is to report the status of efficient deployment of the WaLBerla and ProLB codes on the selected peta-scale and the pre-exascale systems of the IT4I and FZJ supercomputing centers as well as the LUMI pre-exascale system. The deployed codes are then used by WP2 to evaluate the performance and scalability of the codes in the D2.3 deliverable which this deliverable contributes to.

This deliverable provides detailed information on:

- updated list of IT4I, FZJ and LUMI-C peta-scale and pre-exascale systems
- report on deployment of SCALABLE codes and related tools on these systems
- finished, submitted, and planned projects for computational resources needed to access the systems
- updated report on the EPI related development systems access

The SCALABLE project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 956000.

2 Peta-scale and Pre-Exascale Systems for SCALABLE

The supercomputing centers IT4I and FZJ are continuously deploying the newly developed versions of both WaLBerla and ProLB codes on their peta-scale systems throughout the project's runtime. In this section, we provide a detailed list of machines and their partitions that are used for code development, optimization, and benchmarking in work packages 2, 3, 4, and 5. These systems are Barbora (IT4I), Karolina (IT4I), JUWELS Cluster (FZJ) and JUWELS Booster (FZJ).

In addition, WP6 also secured access to LUMI pre-exascale system. As of June 2022, the general access to LUMI supercomputer is only available to its CPU partition, also called LUMI-C. This partition contains 1536 compute nodes and in total 196 608 CPU cores and it is the largest CPU partition in terms of the number of CPU cores available for the project.

Based on LUMI website, the installation of the LUMI-G partition started in March 2022 and the general availability is expected in September 2022 [1] [2] .

2.1 Barbora system at IT4I

partition	#nodes	CPU	accelerator	memory	network
w/o accelerator	192	2× Intel Cascade Lake 6240; 2.6 GHz, 18 cores with AVX-512	-	192 GB	1x 100Gbit/s (Infiniband HDR100)
accelerated	8	2× Intel Skylake Gold 6126; 2.6 GHz, 12 cores with AVX-512	4 × NVIDIA V100, 16GB of HBM2 memory, NVLink2 w/o NVSwitch	192 GB	2x 100Gbit/s (Infiniband HDR100)
fat	1	2× Intel Skylake Platinum 8153; 2.0 GHz, 16 cores with AVX-512	-	6144 GB	2x 100Gbit/s (Infiniband HDR100)

Table 1. Key parameters of the Barbora system

Other key parameters of the system:

- 192 * 36 = 6 912 CPU cores in total in CPU partition,
- total theoretical peak performance (Rpeak) 848.8 TFLOP/s,
- fully non-blocking fat-tree InfiniBand network.

Barbora, even though it is a smaller system, contains a unique energy consumption monitoring system called High-Definition Energy Efficiency Monitoring (HDEEM) developed by Atos. This system enables energy consumption monitoring of an entire compute node with

[1] https://www.lumi-supercomputer.eu/lumis-full-system-architecture-revealed/

[2] https://docs.lumi-supercomputer.eu/computing/systems/lumig/

The SCALABLE project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 956000.

1kHz sampling frequency. The HDEEM is installed only in the non-accelerated partition of the Barbora.

The key difference with other HPC systems lies in the fact, that HDEEM is able to monitor the real power consumption of all compute nodes with all their components at high frequency and not only consumption of selected components, i.e. CPU, DRAM or GPU. In contrast, all other systems can only use RAPL[3] (Running Average Power Limit) to measure energy consumption, which only measures the consumption of the CPU itself.

In addition, this system also enables tuning of the following hardware parameters: (i) CPU core frequency, (ii) CPU uncore frequency and (iii) CPU powercap level.

Together with the possibility of tuning the runtime system parameters, such as number of active OpenMP threads, Barbora provides a complete environment for evaluation of the energy efficiency of the SCALABLE codes using a dynamic tuning methodology.

This system was and it is used to provide energy efficiency evaluation of the codes for two deliverables in WP2:

- D2.2: Report on up-to-date application performance, accuracy, and energy efficiency #1
- D2.3: Report on up-to-date application performance, accuracy, and energy efficiency #2

2.2 Karolina system at IT4I

Karolina has been partially described in D6.1, however as the system was not installed at the time of its submission, we provide a complete description of the system here. Karolina has two major partitions non-accelerated, and GPU accelerated. Both partitions are used for SCALABLE. The uniqueness of the GPU partition is given by fact, that it contains 8 GPU per node with NVSwitch based network, while Juwels Booster contains more common 4 GPUs per node.

NOTE: As of now Karolina does not offer GPU-direct support. This forces MPI to use the CPU network and affects efficient scalability. This issue is being investigated together with the IT4I support team and HPE (machine vendor). For this reason, scalability tests on GPU accelerated version of WaLBerla presented in the deliverable D2.3 are executed on JUWELS Booster, which also enables significantly larger runs.

[3] Running Average Power Limit Energy Reporting: https://www.intel.com/content/www/us/en/developer/articles/technical/software-security-guidance/advisory-guidance/running-average-power-limit-energy-reporting.html

The SCALABLE project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 956000.

partition	#nodes	CPU	accelerator	memory	network
w/o accelerator	720	2 × AMD Zen 2 EPYC™ 7H12, 2.6 GHz, 64 cores with AVX2	-	256 GB 2 GB per core	1x 100Gbit/s (Infiniband HDR100)
accelerated	72	2 × AMD Zen 3 EPYC™ 7763, 2.45 GHz, 64 cores with AVX2	8 × NVIDIA A100, 40GB of HBM2 memory, NVLink3 with NVSwitch	1024 GB 8GB per core	2x 200Gbit/s (Infiniband HDR200)
data analytics	1	32 × Intel Xeon-SC 8268; 2.9 GHz, 24 cores with AVX2	-	24 TB	2x 200Gbit/s (Infiniband HDR200)
cloud	36	2 × AMD Zen 2 EPYC™ 7H12, 2.6 GHz, 64 cores with AVX2	-	256 GB	1x 100Gbit/s (Infiniband HDR100) 2x 10 Gbit/s (Ethernet)

Table 2. Key parameters of the Karolina system

Other key parameters of the Karolina system:

- non-accelerated partition:
  - 720 * 128 = 92 160 CPU cores in total in the non-accelerated partition
  - 720 * 256 = 184 320 GB RAM
  - total theoretical peak performance (Rpeak) 3.83 PFLOP/s
  - HPL performance 2.84 PFLOP/s which means 74% efficiency wrt. to Rpeak.
  - number 183 in Top500 in November 2021
- GPU accelerated partition:
  - 72 * 8 = 576 A100 GPUs in total
  - 72 * 8 * 40 = 23 040 GB of GPU HBM2 memory
  - 72 * 128 = 9 126 CPU cores in total
  - 72 * 1024 = 73 728 GB RAM
  - total theoretical peak performance (Rpeak) 11.6 PFLOP/s
  - number 71 in Top500 in November 2021
- Network: non-blocking Fat Tree which consists of 60 x 40-ports HDR switches (40 Leaf HDR switches and 20 Spine HDR switches)

The SCALABLE project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 956000.

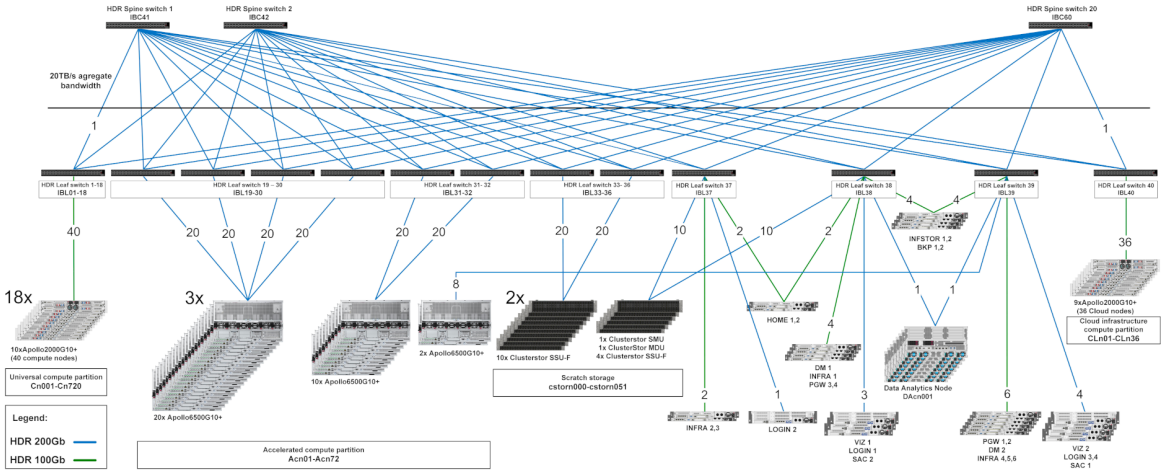


Figure 1. Diagram of the Karolina machine

2.3 JUWELS Cluster system at FZJ

The JUWELS Cluster system is described in more detail in D6.1. In this section, we only summarize its key parameters, which are compared to other systems in the sections below.

partition	#nodes	CPU	accelerator	memory	network
standard	2271	2 × Intel Xeon 8168; 2.7 GHz, 24 cores with AVX-512	-	96GB 2 GB per core	1x 100Gbit/s (Infiniband EDR)
large memory	240	2 × Intel Xeon 8168; 2.7 GHz, 24 cores with AVX-512	-	192 GB 4GB per core	1x 100Gbit/s (Infiniband EDR)
accelerated	56	2 × Intel Xeon 6148; 2.4 GHz, 20 cores with AVX-512	4 × NVIDIA V100, 16GB of HBM2 memory, NVLink2 w/o NVSwitch	192 GB	2x 100Gbit/s (Infiniband EDR)

Table 3. Key parameters of the JUWELS Cluster system

Other key parameters of the system:

- (2271 + 240) * 48 = 120 528 CPU cores in total in both non-accelerated partitions
- (2271 * 96 + 240 * 192) = 264 096 GB of RAM
- number 77 in Top500 in November 2021
- total theoretical peak performance (Rpeak) 10.4 of the two non-accelerated partitions (HPL efficiency 62.4%)
- Mellanox InfiniBand EDR fat-tree network with 2:1 pruning at leaf level and top-level HDR switches

2.4 JUWELS Booster system at FZJ

The SCALABLE project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 956000.

The JUWELS Booster system is described in more detail in D6.1. In this section, we only summarize its key parameters, which are compared to other systems in the sections below.

partition	#nodes	CPU	accelerator	memory	network
	936	2 × AMD Zen 2 EPYC™ 7402; 2.8 GHz, 24 cores with AVX2	4 × NVIDIA A100, 40GB of HBM2 memory, NVLink3 w/o NVSwitch	512 GB 10.6 GB per core	4x 200Gbit/s (Infiniband HDR200)

Table 4. Key parameters of the JUWELS Booster system

Other key parameters of the system:

- 936 * 4 = 3744 A100 GPUs in total
- 936 * 4 * 40 = 149 760 GB of GPU HBM2 memory
- 936 * 48 = 44 928 CPU cores in total
- 936 * 512 = 479 232 GB of RAM
- total theoretical peak performance (Rpeak) 73 PFLOP/s (HPL efficiency 62.4%)
- number 8 in Top500 in November 2021
- Mellanox InfiniBand HDR DragonFly+ topology with 20 cells

2.5 LUMI-C System

LUMI is one of the three EuroHPC Pre-exascale systems. As of now, it is the only one that is installed and running with general availability through EuroHPC access projects for at least its CPU partition called LUMI-C. The GPU partition, LUMI-G, will be available later in 2022.

LUMI-C partition provides the largest number of CPU cores from all the systems available to SCALABLE and therefore it will be mostly used for scalability evaluations and large-scale calculations.

The SCALABLE project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 956000.

partition	#nodes	CPU	accelerator	memory	network
LUMI-C	1376	2 × AMD Zen 3 EPYC™ 7763, 2.45 GHz, 64 cores with AVX2	-	256 GB 2 GB per core	1x 100 Gbit/s (HPE-Cray Slingshot-10)
LUMI-C	128	2 × AMD Zen 3 EPYC™ 7763, 2.45 GHz, 64 cores with AVX2	-	512 GB 4GB per core	1x 100 Gbit/s (HPE-Cray Slingshot-10)
LUMI-C	32	2 × AMD Zen 3 EPYC™ 7763, 2.45 GHz, 64 cores with AVX2	-	1024 GB 8 GB per core	1x 100 Gbit/s (HPE-Cray Slingshot-10)

Note: *The LUMI-C nodes will later be upgraded to 200 Gb/s Slingshot-11 interconnect when LUMI-G becomes operational in 2022.*

Table 5. Key parameters of the LUMI system

Other key parameters of the LUMI-C system:

- 1536 non-accelerated nodes in total
- (1376 + 128 + 32) * 128 = 196 608 CPU cores in total
- 1376 * 256 + 128 * 512 + 32 * 1024 = 450 560 GB RAM in total
- total theoretical peak performance (Rpeak) 7.6 PFLOP/s (HPL efficiency 82.5%)

2.6 *Summary of the systems and range of key parameters*

This section provides a summary of all the system parameters which have major impact on the performance of the codes.

The SCALABLE project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 956000.

Cluster name	Barbora	Karolina	Karolina	JUWELS Cluster	JUWELS Booster	LUMI-C
Partition	**CPU**	**CPU**	**GPU**	**CPU**	**GPU**	**CPU**
Top500 Nov 2021	**--**	**183**	**71**	**77**	**8**	**76**
Peak performance [PFLOP/s]	**0.849**	**3.83**	**11.6**	**10.5**	**73**	**7.6**
# of nodes	192	720	72	2271+240	936	1376 + 128 + 32
# of cores	6,912	92,160	9,216	120,528	44,928	196,608
# of accelerators	32*	--	576	224*	3 744	--
Instruction set	AVX-512	AVX2	AVX2	AVX-512	AVX2	AVX2
CPU architecture	AMD Zen 2	Intel Cascade Lake	AMD Zen 3	Intel Skylake	AMD Zen 2	AMD Zen 3
Number of cores per node / socket	36 / 18	128 / 64	128 / 64	48 / 24	48 / 24	128 / 64
Memory size total [GB]	36,864	184,320	73,728	264,096	479,232	450,560
Memory size						
per node [GB]	192	256	1024	96 / 192	512	256 / 512 / 1024
socket [GB]	96	128	512	48 / 96	256	128 / 256 / 512
NUMA node [GB]	96	32	128	48 / 96	64	32 / 64 / 128
core [GB]	5.3	2	8	2 / 4	10.6	2 / 4 / 8
Mem. bandwidth						
total [GB/s]	53,760	295,200	29,520	642,816	383,760	629,760
per node [GB/s]	280	410	410	256	410	410
per socket [GB/s]	140	205	205	128	205	205
per core [GB/s]	7.8	3.2	3.2	5.3	17.1	3.2
Accelerators	V100*	--	A100	V100*	A100	--
per node	4x		8x	4x	4x	
total number	32		576	224	3,744	
Accel. memory						
type	HBM2	--	HBM2	HBM2	HBM2	--
size per card [GB]	16		40	16	40	
size total [GB]	512		23,040	3,584	149,760	
Network topology	non-blocking Fat-Tree	non-blocking Fat-Tree	non-blocking Fat-Tree	Fat-Tree with 2:1 pruning at leaf level	DragonFly+	Dragonfly
Net. bandwidth						
per node [Gbit/s]	1x100	1x100	4x200	1x100	4x200	1x100
socket [Gbit/s]	50	50	2x200	50	2x200	50
avg. core [Gbit/s]	2.7	0.8	6.25	2.1	16.6	0.8
Network type	InfiniBand HDR 100	InfiniBand HDR 100	InfiniBand HDR 200	InfiniBand EDR	InfiniBand HDR 200	HPE-Cray Slingshot-10

Note: * GPUs are not part of the main (CPU) partition but are available in other partitions of the system.

Table 6. Summary table of peta-scale and pre-exascale systems parameters

To evaluate the scalability of the codes and their capability to solve very large problems in WP2, we have the following options:

The SCALABLE project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 956000.

Codes can be scaled up to the following sizes:

- **92,160** **MPI ranks** on AMD Zen 2 cores **Karolina**
- **120,528** **MPI ranks** on Intel Skylake cores **JUWELS Cluster**
- **196,608** **MPI ranks** on AMD Zen 3 cores **LUMI-C**

GPU accelerated codes can be scaled up to

- **576** **Nvidia A100 accelerators** **Karolina**
- **3,744** **Nvidia A100 accelerators** **JUWELS Booster**

Total memory size is the key parameter for solving the large-scale benchmarks:

- **184TB of CPU memory** with bandwidth **295 TB/s** **Karolina**
- **264 TB of CPU memory** with bandwidth **642 TB/s** **JUWELS Cluster**
- **450 TB of CPU memory** with bandwidth **629 TB/s** **LUMI-C**
- **479 TB of CPU memory** with bandwidth **383 TB/s** **JUWEL Booster**

2.6.1 Potential for co-design and identification of optimal system parameters

In the deliverable D6.3 "*Best practice guide for next-gen industrial systems for large scale CFD simulations based on LB method*" WP6 must identify the optimal system configuration for CFD simulations based on LB method. Based on the system, that we have currently available, we can already start this work and identify the optimal system configuration for the new version of the SCALABLE codes:

Memory size and average memory bandwidth per core

- 3.2, 5.3, 7.8 and 17.1 GB/s average memory bandwidth per CPU core
- 2, 4, 5.3, 8 and 10.6 GB of memory per CPU core

Single-core performance and vector instruction set:

- AVX2: AMD Zen 2 and 3 generations
- AVX512: Intel Skylake and Cascade Lake generations

GPU accelerated codes architecture

- Nvidia V100 and A100 GPU architectures
- intra-node architectures and effect of different type of NVLink network
  - 4 GPUs per node – connected directly via NVLink
  - 8 GPUs per node – connected via NVLink and NVSwitch
  - one node with 16 GPUs per node – DGX-2 machine with V100 with 32 GB per card and 512 GB total

The SCALABLE project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 956000.

Network

- 1, 2 or 4 links per node at a rate 100 or 200 Gbit\s
- different ratios of GPU per NIC
    - 4:4 on JUWELS Booster
    - 8:2 on Karolina
    - 4:2 on Barbora and JUWELS Cluster
- can be evaluated for network topologies (systems contain the two most common topologies)
    - Fat tree on Infiniband network
    - Dragonfly+ and DragonFly on Infiniband and Slingshot networks

3 Providing computational resources for the project

This section describes the status of computational resource projects that have been submitted or are planned for the SCALABLE consortium. First, we summarize the finished projects and the statistics of used resources. Finally, we describe the planned activities for the second half of the project. Please note that the description of the user account creation has been done in D6.1, and it is not presented here as the only new set of accounts is for LUMI system.

3.1 Finished projects

In this section we report the projects that have been submitted at the beginning of the projects to get access to the development systems. Most of the were short-running projects and their results were used to submit long-term projects with more computational resources.

3.1.1 Director Discretion project for IT4I systems

Preliminary, ad-hoc project to enable rapid access to systems. The project was used for account creation at IT4I, GIT access, environment preparation, tools installation and initial benchmarking.

- Project ID: DD-21-3
- Submitted at: 09.02. 2021
- Accepted at: 18.02. 2021
- Duration: 6 months
- Amount of resources: 250 000 core hours
- Systems: GIT, Barbora, DGX-2

The SCALABLE project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 956000.

Status of the project:

Status	Resource	Validity Period	Requested	Allocation	Usage	
Finished	Core Hours	2021-02-18 to 2021-08-20	250,000	250,000	253,580	101.4%

This project was already reported in D6.1. Here we report the usage of the project. All the allocated computational resources have been successfully used by the consortium.

3.1.2 PRACE Preparatory Access Projects for FZJ systems

Two PRACE Preparatory Access projects have been submitted by WP6, one for WaLBerla and one for ProLB to get access to JUWELS Cluster and JUWELS Booster systems. In PRACE access these projects are necessary to be able to submit the Project Access proposal. Via these two projects, the initial access to computing systems at FZJ was secured and the results obtained were used to apply for PRACE Project Access.

Preparatory Access for WaLBerla

- Project ID: 2010PA5894
- Submitted at: 01.05.2021
- Accepted at: 01.06.2021
- Duration: 6 months
- Amount of resources:
  - JUWELS Booster: 25 000 core hours
  - JUWELS Cluster: 50 000 core hours
- Systems: JUWELS Booster, JUWELS Cluster

Status of the project:

Status	Resource	Validity Period	Requested	Allocation	Usage	
Finished	Core Hours	2021-06-01 to 2022-05-31	25,000 50,000	25,000 50,000	25,000 49,000	100%

This project was successfully finished and a report "***2010PA5894 - Performance Analysis of waLBerla in SCALABLE EU Project***" has been submitted to PRACE.

Preparatory Access for ProLB

- Project ID: 2010PA5930
- Submitted at: 01.06.2021
- Accepted at: 01.07.2021
- Duration: 01.07.2021 to 30.11.2022

The SCALABLE project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 956000.

- Amount of resources:
  - JUWELS Cluster: 50 000 core hours
- Systems: JUWELS Cluster
- Purpose: Preparation for PRACE Large-scale Call

Status of the project:

Status	Resource	Validity Period	Requested	Allocation	Usage	
Active	Core Hours	2021-07-01 to 2022-11-30	50,000	50,000	0	0%

Due to delays in the legal process, the transfer of source code to FZJ systems has been delayed. Therefore, the compute time has not yet been used. The compute project has therefore been extended up to 30.11.2022. The legal issues have now been sorted and we expect the compute time to be fully used during this session of the allocation.

3.2 *Active projects*

3.2.1 Open Access Project for IT4I systems

This is the main project that secures computational resources for code development, optimization and performance analysis at the IT4I infrastructure. It is used for code development and optimization, continuous integration, and benchmarking. It provides access to both CPU and GPU accelerated nodes.

- Project ID: OPEN-22-7
- Submitted at: 02.04. 2021 (22nd Open Access Grant Competition)
- Accepted at: 28.05. 2021
- Duration: 3 years
- Amount of resources: 12 200 000 core hours
- Systems: Karolina, Barbora, DGX-2, ARM server, and future complementary systems

The SCALABLE project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 956000.

Status of the project:

Status	Resource	Validity Period	Requested	Allocation	Usage Used | Used [%] | Remaining		
Expired	Core Hours	2021-05-24 to 2022-02-23	1,700,000	1,700,000	1,701,565	100.1	0
Active	Core Hours	2022-02-23 to 2022-11-25	3,300,000	3,201,000	611,371	19.1%	2,589,628
Planned	Core Hours	2022-11-25 to 2023-08-27	4,300,000	4,171,000	0	0%	4,171,000
Planned	Core Hours	2023-08-27 to 2024-05-28	2,900,000	2,813,000	0	0%	2,813,000

This project was already reported in D6.1. Here we report the usage of the core-hours within the project. The project is broken into 4 periods and there is a given allocation for each period.

By the time of submission of this deliverable, the 1st period had expired, and all allocated resources had been used. Currently, the second period is in place.

3.2.2 EuroHPC code development projects for IT4I Karolina CPU and GPU partitions

In addition to the open-access project described above, WP6 has also submitted two more proposals for computational resources of the IT4I Karolina machine using EuroHPC JU development and benchmarking call. The first is for CPU partition and the second one for GPU partition.
Submission of these proposals was necessary to mitigate the risk of temporarily losing access to IT4I systems due to the bankruptcy of the electricity provider for IT4Innovations and sudden increase in electricity price. The kind of work done on these projects is the same as in case of the Open Access Project above.

EuroHPC code development project for IT4I Karolina GPU partition:

- Project ID: EU2021D02-048, DD-22-12
- Submitted at: 01.02. 2022 (EuroHPC EuroHPC Development Access)
- Accepted at: 24.02. 2022
- Duration: 1 year
- Amount of resources: 384 000 core hours
- Systems: Karolina GPU partition

The SCALABLE project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 956000.

Status of the project:

Status	Resource	Validity Period	Requested	Allocation	Usage Used | Used [%] | Remaining		
Active	Core Hours	2022-02-24 to 2023-02-24	N/A	384,000	25,468	6.6%	358,531

EuroHPC code development project for IT4I Karolina CPU partition:

- Project ID: EU2021D02-048, DD-22-11
- Submitted at: 01.02. 2022 (EuroHPC EuroHPC Development Access)
- Accepted at: 24.02. 2022
- Duration: 1 year
- Amount of resources: 1 920 000 core hours
- Systems: Karolina CPU partition

Status of the project:

Status	Resource	Validity Period	Requested	Allocation	Usage Used | Used [%] | Remaining		
Active	Core Hours	2022-02-24 to 2023-02-24	N/A	1,920,000	214,858	11.2%	1,705,141

3.2.3 EuroHPC code development project for LUMI-C partition

This is the first project that secures the SCALABLE project to access the EuroHPC pre-exascale system LUMI. At the time of writing the report, only the LUMI-C partition is available to the general public. Currently for SCALABLE this partition will be used for development and optimization to improve scalability of the CPU based version of the codes when scaling over 100 000 CPU cores

- Project ID: EHPC-DEV-2021D04-132
- Submitted at: 30.04. 2022
- Accepted at: 23.05. 2022
- Duration: 1 year
- Amount of resources: 1 920 000 core hours
- Systems: LUMI-C

The SCALABLE project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 956000.

Status of the project:

Status	Resource	Validity Period	Requested	Allocation	Usage Used | Used [%] | Remaining		
Active	Core Hours	2022-07-01 to 2023-07-01	N/A	1,920,000	0	0%	1,920,000

The GPU accelerated partition will be available in a few months and we will prepare and submit the proposal for this partition as our partners are already working on porting the WaLBerla code to AMD GPUs.

3.2.4 GCS Compute Access for FZJ Cluster and Booster

Gauss Center for Supercomputing offers computing allocations to scientists and researchers performing ground-breaking research projects dealing with complex, demanding simulations. In context of the SCALABLE project, the primary goal of the GCS compute call is to provide compute resources for large scale testing and optimization of the compute codes in the project.

- Project ID: scalable2022
- Submitted at: 23.02. 2022
- Accepted at: 01.05. 2022
- Duration: 1 year
- Systems: JUWELS Cluster and JUWELS Booster

Status of the project:

Status	Resource	Validity Period	Requested	Allocation	Usage Used | Used [%] | Remaining		
Active	Core Hours	2022-05-01 to 2023-04-30	7,700,000	5,000,000	0	0%	5,000,000
Active	Core Hours	2022-05-01 to 2023-04-30	6,800,000	4,500,000	0	0%	4,500,000

3.3 Planned proposals

In the near future, we plan to submit the following **EuroHPC code development projects** for selected EuroHPC peta-scale and pre-exascale systems.

3.3.1 LUMI-G partition of LUMI pre-exascale system to evaluate

The main performance of the LUMI-G partition comes from AMD MI250X GPUs and therefore in order to take the advantage of the system the codes have to support one of the

The SCALABLE project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 956000.

programming languages. WaLBerla has very recenty been extended to support AMD GPUs and we plan to evaluate its performance and scaling.

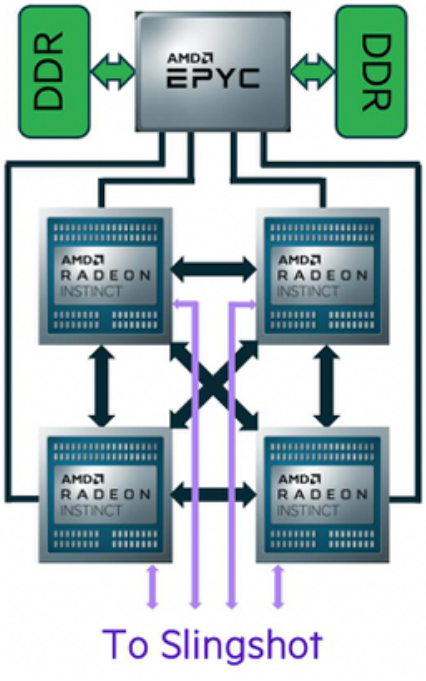


Figure 2. LUMI-G compute node diagram

LUMI-G key parameters:

- 2,560 GPU accelerated nodes
- potential peak performance 550 PFlop/s
- committed Linpack performance is 375 PFlop/s (to be confirmed)

Node architecture:

- 1x 64 core AMD Trento CPU, 256 GB RAM
- 4x AMD MI250X GPUs
- 4x 200 Gbit/s network interconnect cards

The MI250X GPU key features:

- 42.2 TFLOP/s of performance in the HPL benchmarks
- two compute dies, each with 110 compute units, and each computes unit has 64 stream processors for a total of 14,080 stream processors.
- 128 GB of HBM2e memory with over 3.2 TB/s of memory bandwidth.

3.3.2 Leonardo pre-exascale system

Similar to JUWELS Cluster and Booster systems, Leonardo is built as two modules, one CPU based (data-centric) and one GPU accelerated (Booster). As such this system can be used for both CPU (WaLBerla and ProLB) and GPU accelerated (GPU version of WaLBerla) codes. More precisely we plan to evaluate following technologies:

- the next generations of Intel Xeon Ice Lake and Saphire Rapids CPU generations (both without HBM memory),
- performance and scalability of the GPU accelerated WaLBerla as it will be equipped with approximately 14,000 NVIDIA Ampere GPUs (3,456 compute nodes, each with 4 GPUs),
- this system will also help us to evaluate the impact of network performance as its GPU nodes will have 4x 100Gbit links while JUWELS Booster has 4x 200Gbit links.

The SCALABLE project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 956000.

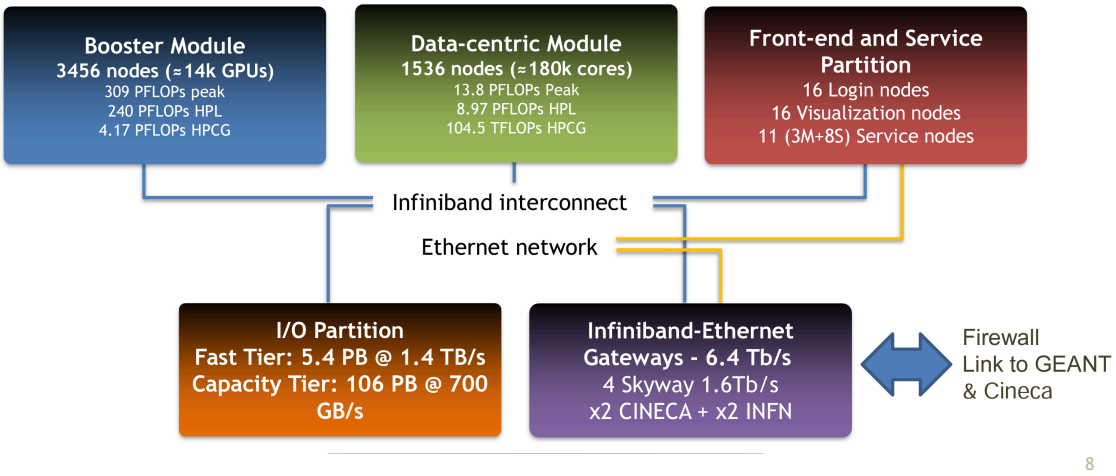


Figure 3. EuroHPC pre-exascale Leonardo system specification

BullSequana X2135 "Da Vinci" single-node GPU Blade

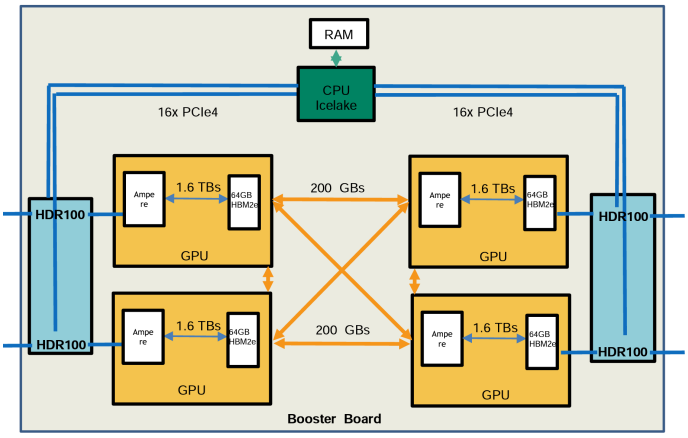
- 1 x CPU Intel Xeon 8358 32 cores, 2,6 GHz
- 8 x 64 GB (512GB) RAM DDR4 3200 MHz
- 4 x NVidia custom Ampere GPU 64GB HBM2
- 2 x NVidia HDR100 dual port card

BullSequana X2140 three-node CPU Blade

For each node:

2 x CPU Intel Sapphire Rapids 56 core 350W
16 x 32 GB RAM (512 GB) DDR5 4800 MHz
1 x NVidia HDR100 single port card
1 x M.2 SSD 3,84 TB

Figure 4. Leonardo CPU node (top) and GPU accelerated node (bottom) specification



- Ad hoc board from ATOS
- 4 NVidia custom Ampere GPUs (SXM)
- CPU-GPU connection via PCIe4 16x connection through HDR Connect6 HCA
  - PCI passthrough
  - 16 PCI links towards CPU, 16 links towards GPU
  - Bandwidth: 32 GB/s
- Full NVLink GPU-GPU connection
  - 200 GB/s bi-directional
- No PCI switch between host and external network
  - Low latency
- Out-of-band telemetry information
- GPUdirect

Figure 5. Leonardo GPU accelerated node architecture

The SCALABLE project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 956000.

3.3.3 Deucalion ARM-based EuroHPC peta-scale system

Deucalion system consists of 1,632 nodes each with one Fujitsu A64FX per node equipped with 32 GB of HBM memory and it is installed at Minho Advanced Computing Center Portugal. This system will enable us:

- performance and scaling on ARMv8.2 based system which will be used for work in Task 4.3 *"EPI enablement and co-design"* [Month 12-36]

The system can be also used for code porting and single node optimization. However, this work is already being carried out on smaller systems described in the next section.

3.4 Access to small system related to EPI processor porting

In order to enable the consortium to perform work in Task 4.3 (EPI enablement and co-design) it is important to have access to ARMv8.2-based systems such as ones based on the Fujitsu A64FX CPU. Rhea processor, which is developed under the EPI project, has following key features:

- SVE vectorization support in CPU cores
- fast HBM memory

These two key features are also available in the A64FX, therefore it is a suitable platform to perform porting work for the new EPI processor in the SCALABLE.

In the previous section we have already mentioned that we plan to submit the EuroHPC code development project for **Deucalion** system. At the time of submission of D6.1 we had expected to have access to this system by now. However, its general availability has been delayed and as of writing this deliverable the system is still not available.

The current status of access of the consortium members to the A64FX based system is as follows:

- CERFACS has access to the GENCI's systems with 80 A64FX nodes where the technologies of the SCALABLE are tested and ported.
- FZJ has access to an internal system with 28 Cortex A72 nodes where WaLBerla is in process of being ported.
- IT4I has a single server with two ARMv8 CPUs without SVE instruction set and HBM memory (2 × Hi1616; 2.4 GHz and 64 cores, 256 GB RAM)

IT4I has successfully finished the procurement of its Complementary systems I., which contains 8 nodes of HPE Apollo 80 server[4], each with A64FX CPU and Infiniband 100Gbit/s interconnect. In addition, it contains the Cray Programming Environment and Cray Scientific Libraries, which is a powerful alternative software stack to the Fujitsu one, which is available on **Deucalion** and other Fujitsu systems. This system is expected to be operational in August

[4] HPE Apollo 80 System: https://buy.hpe.com/cz/en/servers/apollo-systems/apollo-80-system/apollo-80-system/hpe-apollo-80-system/p/1012970957

The SCALABLE project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 956000.

2022 (M20 of the SCALABLE project) and immediately it will be available to all projects members through the existing Open-Access computational resource project for the IT4I systems.

Both FZJ and IT4I are also members of EuroHPC **EUPEX project** which develops the prototype of the European Exascale supercomputer based on the EPI processor technology. Thanks to this close link to the project which develops the actual target platform of the T4.3 we can also provide feedback to EUPEX as well exploit lesson learned in it.

In addition, FZJ is also member of the TEP ("The European Pilot", Pilot-2) for the RISC-based European accelerator.

4 Set-up of the peta-scale and pre-exascale systems

The SCALABLE codes have been installed on the selected system. Not all codes and tools are installed on all systems presented above. Here is the complete list.

4.1 Barbora CPU partition

The installation on Barbora has been described in D6.1, for sake of completeness, these are the modules that has been used:

WaLBerla

- GCC 9.3.0,
- OpenMPI 4.0.3,
- Boost 1.72.0,
- FFTW 3.3.8,
- Python 3.8.6.

ProLB

- OpenMPI/3.1.5-GCCcore-8.3.0
- HDF5/1.10.6-GCC-6.3.0-2.27-serial

Installation of Energy Efficiency tools

- MERIC runtime system
  - HDEEM 2.2.7
  - Libmsr 0.3.1

4.2 Karolina CPU partition

Installation of WaLBerla

WaLBerla has not been installed on this partition as most of its development is related to GPU acceleration.

Installation of ProLB

At IT4I, the ProLB was successfully installed on Karolina CPU partition using the Intel-based toolchain including compilers and MPI library contained in the following modules:

The SCALABLE project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 956000.

- HDF5/1.10.6-intel-2020b-parallel
- impi/2019.9.304-iccifort-2020.4.304
- EXTRAE/4.0.0-impi-2019.9

The Extrae library is an additional dependency due to manual instrumentation for POP2 performance analysis.

The current version of the ProLB uses a licensing, thus a FlexLM license server manager (LMGRD v11.13.1.2) was installed on the system.

Installation of POP2 analysis tools

The POP2-based performance analysis is conducted primarily using the BSC (Barcelona Supercomputing Center) performance toolchain that is composed of the following modules and custom-installed tools:

- impi/2019.9.304-iccifort-2020.4.304
- libxml2/2.9.10-GCCcore-10.3.0
- libunwind/1.4.0-GCCcore-10.3.0
- EXTRAE/4.0.0-impi-2019.9
- Paraver 4.10.0
- Dimemas 5.4.1
- PAPI 6.0.0.1
- Likwid 5.2.0

Installation of Energy Efficiency tools

- MERIC runtime system
    - AMD E-SMI In-band Library 1.5.0

4.3 Karolina GPU partition

Installation of WaLBerla

- OpenMPI/4.1.2-NVHPC-22.2-CUDA-11.6.0
- CMake/3.20.1-GCCcore-10.2.0
- Python/3.8.6-GCCcore-10.2.0

Installation of POP2 analysis tools

On the GPU partition of the Karolina cluster, the same set of tools as in the case of the CPU partition is used. The only extra dependency for the Extrae library is the following CUDA module:

- CUDAcore/11.6.0

Installation of Energy Efficiency tools

The SCALABLE project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 956000.

- MERIC runtime system
    - AMD E-SMI In-band Library 1.5.0

4.4 JUWELS Cluster

Installation of WaLBerla

At FZJ, WaLBerla with code generation support was successfully installed on JUWELS Cluster using following environment modules:

- Python 3.9.6
- GCC 11.2.0
- ParaStationMPI 5.5.0.-1
- FFTW 3.3.10
- Boost 1.78.0
- Extrae 3.8.3
- Score-P 7.1

4.5 JUWELS Booster

Installation of WaLBerla

At FZJ, WaLBerla with code generation support was successfully installed on JUWELS Booster using following environment modules:

- Python 3.9.6
- GCC 11.2.0
- ParaStationMPI 5.5.0.-1
- FFTW 3.3.10
- Boost 1.78.0
- CUDA 11.5
- Extrae 3.8.3
- Score-P 7.1

4.6 LUMI-C

Lumi is a Cray system, so the preferred software stack is Cray Programing Environment. Due to 4 to 6 week long maintenance break that started on June 6th related to installation of LUMI-G partition the installation will be finished few weeks after submission of this deliverable.

The SCALABLE project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 956000.

5 Conclusions

This deliverable summarizes the status of the systems available to the SCALABLE consortium. It analyzes the key parameters of the system and provides information on how the systems differ. Based on system variability we start asking questions about what the optimal system parameters for the SCALABLE codes are and how the optimal system for CFD codes based on the Lattice Boltzmann method should be built. The conclusion of this analysis will be presented in the next deliverable D6.3.

Then, it presents the status of the deployment of the codes on the systems and the status of the projects securing computational resources on these systems. We present the status of all projects (past, running and planned). The planned projects are for the upcoming pre-exascale systems LUMI, in particular LUMI-G, and the Leonardo system to be installed in CINECA.

The key conclusion of this deliverable is that WP6 keeps providing the access to important development and production systems at IT4I, FZJ, and LUMI to the consortium members and that we continuously watch for new technologies being deployed.

The SCALABLE project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 956000.