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

D6.1 Report on identification and set-up of key development systems

Work Package	**WP 6, Development and Deployment Platforms and Services**
Lead Author (Org)	**Lubomir RIHA, Ondrej Vysocky (IT4I)**
Contributing Author(s) (Org)	**Jayesh Badwaik, Andrea Herten (FZJ)**
Due Date	**01.07.2021 (M6)**
Date	**15.07.2021**
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	29.04. 2021	Ondrej Vysocky, Lubomir Riha (IT4I), Jayesh Badwaik, Andrea Herten (FZJ)	Initial skeleton of the document
0.2	31.05. 2021	Ondrej Vysocky, Lubomir Riha (IT4I), Jayesh Badwaik, Andrea Herten (FZJ)	First draft of the document
0.3	03.06. 2021	Ondrej Vysocky, Lubomir Riha (IT4I), Jayesh Badwaik, Andrea Herten (FZJ)	Draft prepared for review
0.4	28.06. 2021	Cuidard Romain (CSGROUP), Gabriel Staffelbach (CERFACS), Harald Koestler (UErlangen)	Version after internal review
1.0	7.7.2021	Lubomir Riha (IT4I)	Final version

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 the access to the experimental and cutting-edge computing systems that will be used for code development and optimization, benchmarking, and porting to new architectures by all members of the consortium.
The key development systems at IT4I and FZJ were identified, and through the Director Discretion project (IT4I) and PRACE Preparatory Access project (FZJ) all project partners have now access to the identified development systems for early work on both WaLBerla and ProLB. Access to petascale/pre-exascale systems Karolina and LUMI will be enabled during the upcoming months when the systems become generally available for users.

Both WaLBerla and ProLB were installed and tested using a basic test case at Barbora and JUWELS Booster, so that WP2 can perform benchmarking.

A compute project proposal with several million core hours has been submitted to provide resources for development at IT4I systems, other similar-size projects are planned to propose resources of JUWELS Booster and LUMI systems.

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

1 Introduction

The goal of this deliverable is to report the status of efforts of IT4Innovations (IT4I) and FZ Jülich (FZJ) which provide and maintain access to their experimental and cutting-edge computing systems which will be used to enable single-node and multi-node optimization, GPU acceleration, porting to new architectures, and early code development by all members of the consortium.

This deliverable provides detailed information on:

- Identification of development systems from both supercomputing sites and across PRACE and EuroHPC machines, including key technical parameters,
- Submitted and planned projects for computational resources,
- Account creation for consortium members at both sites,
- Installation and validation of the WaLBerla and ProLB softwares on development systems.

2 Identification of development systems

IT4I and FZJ secure computing resources by providing access to their HPC infrastructures as well as by writing PRACE or EuroHPC proposals to get access to other European systems, which are necessary for successful development of WaLBerla and ProLB to become exascale-ready. Based on work defined in the proposal and after discussion with the partners, the following systems have been identified for development and initial benchmarking.

2.1. IT4I Barbora (IT4I)

partition	#nodes	CPU	accelerator	memory
w/o accelerator	192	2 × Intel Cascade Lake 6240; 2.6 GHz, 16 cores	-	192 GB
accelerated	8	2 × Intel Skylake Gold 6126; 2.6 GHz, 12 cores	4 × NVIDIA Tesla V100-SXM2	192 GB
fat	1	2 x Intel Skylake Platinum 8153; 2.0 GHz, 16 cores	-	6144 GB

Key parameters:

- Total theoretical peak performance (Rpeak) 848.8 TFLOP/s

The key role of the Barbora system is in enabling energy consumption evaluation and tuning. The non-accelerated nodes are equipped with High-Definition Energy Efficiency Monitoring

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

(HDEEM) power monitoring system, which will be used for precise energy efficiency optimization of the code.

In addition, on this system tuning of selected hardware parameters (for instance: CPU core frequency, CPU uncore frequency, powercap level, etc.) can be performed which is essential for the proposed methodology of dynamic tuning.

Also, in the first months of the project, prior to launch of Karolina (see below), this is the main development and benchmarking system.

2.2. IT4I Karolina (IT4I)

partition	#nodes	CPU	accelerator	memory
w/o accelerator	720	2 × AMD EPYC 7H12; 2.6 GHz, 64 cores	-	256 GB
accelerated	70	2 × AMD EPYC 7452; 2.35 GHz, 32 cores	8 × NVIDIA A100	512 GB
data analytics	1	32 × Intel Xeon 8268; 2.9 GHz, 24 cores	-	24 TB
cloud	36	2 × AMD EPYC 7H12; 2.6GHz, 64 core	-	256 GB

Key parameters:

- Theoretical peak performance (Rpeak)
    - CPU only partition: 2.3 PFLOP/s
    - GPU accelerated partition: 6.6 PFLOP/s built from large GPU nodes with 8 GPUs per node
    - data analytics partition: 40 TFLOP/s
    - cloud partition: 131 TFLOP/s

At the time of writing this report, the Karolina system is being installed, validated, and prepared for production. It is expected that production will start in June 2021 and SCALABLE members will have access to the system right after that.

The key role of Karolina as the largest system at IT4I is to optimize the codes for AMD CPUs and perform large scale benchmarking on its CPU partition with over 700 nodes. This node size will correspond to small- and middle-size jobs.

Karolina will also contain state-of-the-art GPU accelerators (Nvidia A100) with full NVlink interconnect between 8 GPUs per node.

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

2.3. LUMI (IT4I)

partition	#nodes	CPU	accelerator	memory
w/o accelerator	> 1500	AMD EPYC	-	unknown
accelerated	unknown	AMD EPYC	AMD MI GPUs	unknown

Key parameters:

- Theoretical peak performance (Rpeak) 552 PFLOP/s

LUMI will be a EuroHPC pre-exascale system. It's the CPU partition of the system is expected to be in production from Q3 of 2021. This partition is expected to be used for large scale CPU runs. In Q4 of 2021 it is expected that the accelerated part of the machine becomes available. This partition will be accelerated by a new generation of AMD server GPUs and expected to have 4 GPUs per node. From a development point of view, this partition will enable development of the GPU-accelerated codes using this new generation of AMD GPU accelerators which are expected to share memory between CPU and GPU.

In addition to code development and performance optimization at scale, the key role of this EuroHPC pre-exascale machine is to run large scale demonstrators identified in WP2.

2.4. ARMv8 server (IT4I)

partition	#nodes	CPU	accelerator	memory
--	1	2 × Hi1616; 2.4 GHz, 64 cores	-	256 GB

The ARMv8 Huawei heterogenous system (without SVE instruction set) can be used for porting the project applications to an ARM architecture, as needed for porting to the EPI infrastructure (Task 4.1).

2.5. DGX-2 (IT4I)

partition	#nodes	CPU	accelerator	memory
--	1	2 × Intel Xeon 8168; 2.7 GHz, 24 cores	16 × NVIDIA Tesla V100-SMX3	512 GB

Key parameters:

- Theoretical peak performance (Rpeak) 130 TFLOP/s

The DGX-2 provides 16 Nvidia V100 GPUs per node, making it the hardware platform with the highest number of GPUs per node from the listed ones. It allows development towards scaling on very fat nodes with large numbers of GPUs used during the computation.

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

2.6. JUWELS Booster (FZJ)

partition	#nodes	CPU	accelerator	memory
--	936	2 × AMD EPYC 7402; 2.8 GHz, 24 cores	4 × NVIDIA A100 Tensor Core GPU, 40 GB, NVLink3	512 GB

Key parameters:

- Theoretical peak performance (Rpeak) 73 PFLOP/s
- Dragonfly-type network topology

JUWELS Booster is currently the fastest and largest GPU installation in Europe. The key role of the system is to optimize the performance of the applications for modern heterogenous architectures in large systems. The Nvidia A100 GPUs in each node provide opportunity for optimization towards high-bandwidth memories. The large number of nodes, which are connected in DragonFly+ topology, allow to optimize communication between different GPUs and try to solve performance challenges arising from the DragonFly+ network topology. The large network injection bandwidth per node (4 × 200 Gbit/s) offers opportunity to study bandwidth-intensive data exchange.

2.7. JUWELS Cluster (FZJ)

partition	#nodes	CPU	accelerator	memory
standard	2271	2 × Intel Xeon 8168; 2.7 GHz, 24 cores	-	96GB
large memory	240	2 × Intel Xeon 8168; 2.7 GHz, 24 cores	-	192 GB
accelerated	56	2 × Intel Xeon 6148; 2.4 GHz, 20 cores	4 × Nvidia V100, 16GB HBM	192 GB

Key parameters:

- Theoretical Peak Performance (Rpeak) : 12 PFLOP/s
- Key role – Enhancing CPU based Scalability of Codes

JUWELS Cluster provides around 2500 nodes which is the largest number of nodes in all the systems in this list. The large number of nodes can be used to test scaling of the target applications in CPU configuration. Specifically, it will be used to enhance CPU-based communication between nodes and optimize execution on CPU using vectorization and other techniques.

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

3 Securing computational resources

In order to provide the SCALABLE project with computational resources needed to perform work described in the proposal, IT4I and FZJ have prepared and submitted the following proposal.

3.1. Submitted proposals

3.1.1 Director Discretion project for IT4I systems

Preliminary, ad-hoc project to enable rapid access to systems.

- Submitted at: 09.02. 2021
- Accepted at: 18.02. 2021
- Duration: 6 months
- Amount of resources: 250 000 core hours
- Systems: Barbora, DGX-2, ARM server
- Purpose: Project used for account creation at IT4I, environment preparation, tools installation, initial benchmarking

3.1.2 Open Project for IT4I systems

This is the main project that secures computational resources for code development and optimization at IT4I infrastructure. It does not contain resources to be used at LUMI, we will submit a separate proposal for LUMI.

- Submitted at: 02.04. 2021 (22nd Open Access Grant Competition)
- Accepted at: 28.5. 2021
- Duration: 3 years
- Amount of resources: 12 200 000 core hours
- Systems: Karolina, Barbora, DGX-2, ARM server, and future complementary systems
- Purpose: used for code development and optimization, continuous integration and benchmarking for small- and middle-size test cases.

3.1.3 EuroHPC/PRACE Preparatory Access for FZJ systems

PRACE Preparatory Access allows PRACE users to optimise, scale, and test codes on PRACE Tier-0 systems before applying to PRACE calls for Project Access. Via Preparatory Access, initial access to computing systems at FZJ is secured. The results from this initial access will be used to apply for PRACE compute calls for both WaLBerla and ProLB. We have submitted two applications for Preparatory Access, one for WaLBerla and another for ProLB.

3.1.3.1 Preparatory Access for WaLBerla

- Submitted at: 17.06. 2021

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

- Accepted at: 23.06. 2021
- Duration: 6 months
- Amount of resources:
  - JUWELS Booster: 25 000 core hours
  - JUWELS Cluster: 50 000 core hours
- Systems: JUWELS Booster, JUWELS Cluster
- Purpose: Preparation for PRACE Large-scale Call

3.1.3.2 Preparatory Access for ProLB

- Submitted at: to be determined
- Accepted at: to be determined
- Duration: 6 months
- Amount of resources:
  - JUWELS Cluster: 50 000 core hours
- Systems: JUWELS Cluster
- Purpose: Preparation for PRACE Large-scale Call

3.2. Planned proposals

3.2.1 Large-Scale Access to PRACE/EuroHPC Systems

As this report is submitted in M6 of the project runtime, in the following months more computational resource proposals will be submitted to be able to provide more compute time and access to larger systems. The main goal of this effort is to be able to run the large-scale demonstrators. We plan to submit following proposals:

- **EuroHPC or PRACE project access for FZJ systems**
  - Purpose: Obtain long-term access and compute time needed to run large-scale demonstrators on FZJ systems.
- **EuroHPC or internal IT4I project for accessing LUMI**
  - Purpose: Obtain long-term access to LUMI with focus on large-scale demonstrators

3.2.2 Access to systems related to EPI

In order to enable partners to perform work in **Task 4.3 (EPI enablement and co-design [Month 12-36])** it is important to secure access to ARM-based systems. ARMv8.2-based systems similar to ones based on the Fujitsu A64FX CPU are especially important in order to perform porting work for the new EPI processor, as the A64FX architecture is expected to be similar to the GPP part of the EPI processor.

IT4I, as of now, have an ARMv8 server, see above, that can be used for initial porting to ARMv8. In addition, IT4I plans to acquire a new complementary system in Q1 2022, which

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

should also contain several A64FX-based nodes. Installation of this system overlaps well with M12 of the Task 4.3 in the project.

Also, CERFACS via a collaboration with the GENCI technology watch group will have access to 80 A64FX nodes where the technologies of SCALABLE could be tested. Access for the other partners will be negotiated if possible.

Another possible ARMv8.2 to which the European research community will have access to is the EuroHPC petascale **Deucalion supercomputer** installed at Minho Advanced Computing Center Portugal. This system consists of 1632 nodes each with one Fujitsu A64FX per node equipped with 32 GB of HBM memory.

Both FZJ and IT4I are members of Horizon 2020 EUPEX project which develops the prototype of the European Exascale supercomputer based on the EPI processor technology. In addition, FZJ is also member of the TEP ("The European Pilot", *Pilot-2*) for the RISC-based European accelerator.

4 Set-up of development systems

4.1. Creation of user accounts at IT4I

Consortium members (FJZ, FAU, CS group, CERFACS) with the help of IT4I support team were able to obtain login credentials for IT4I systems.

4.2. Creation of user accounts at JSC

The process of creation of user accounts at JSC will be started once the preparatory access is finalized. The process will follow the standard procedure established at JSC using JSC's JuDoor portal, https://judoor.fz-juelich.de/. After creation of the preparatory access project, SCALABLE participants will be able to register on JuDoor via the respective project. After successful registration, system access is provided via SSH or Jupyter (https://jupyter-jsc.fz-juelich.de/).

4.3. Installation of WaLBerla

At IT4I, WaLBerla with code generation support was successfully installed on Barbora using the following environment modules:

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

Also, the GPU-accelerated version compiled with CUDA 11 is available. The application is also ready to be analysed for energy efficiency analysis.

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

The validation of the WaLBerla was performed using UniformGridGenerated test case, which is part of the WaLBerla repository, as well as Turbulence modelling test case, which is one of the benchmarks, that will be further used for performance analysis of the libraries.

On Karolina, WaLBerla will be installed as soon as the system becomes generally available to users.
At FZJ, WaLBerla with code generation support was successfully installed on JUWELS using following environment modules:

- Python 3.8.5
- GCC 9.3.0
- ParaStationMPI 5.4.7.-1
- FFTW 3.3.8
- Boost 1.74.0

4.4. Installation of ProLB

At IT4I, the ProLB was successfully installed on Barbora using the following environment modules:

- OpenMPI/3.1.5-GCCcore-8.3.0
- HDF5/1.10.6-GCC-6.3.0-2.27-serial (not need for execution) Xdmf was locally installed for compilation on Barbora. And the version of ProLB installed on Barbora do not support the feature ENSIGHT_READER. This feature is not relevant for performance testing and avoid the dependency with VTK.

Currently, for administrative reason, this version does not use licencing therefor It's only available for CSGROUP member.

At this moment, ProLB on Barbora is successfully tested on unitary tests cases over two processors on the Barbora frontal.

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

5 Conclusions

The key development systems were identified, and through the Director Discretion project (IT4I) and PRACE Preparatory Access project (FZJ) all project partners have now access to the identified development systems for early work on both WaLBerla and the ProLB.

Access to petascale/pre-exascale systems Karolina and LUMI will be possible during the upcoming months when systems become generally available for users. At the current stage of the project, this is not a limiting factor.

Both WaLBerla and ProLB were installed and tested using basic test cases at Barbora and JUWELS Booster, so that WP2 can perform benchmarking.

A compute time project with several million core hours has been submitted to provide resources for development at IT4I systems, other similar-sized projects are planned for resources of JUWELS and LUMI.

Finally, we have also described our approach for access to ARMv8.2-based systems need for work in WP4 related to porting to EPI processor platform.

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