Fraunhofer
SCAI

FRAUNHOFER INSTITUTE FOR ALGORITHMS AND SCIENTIFIC COMPUTING SCAI

LAMA
www.libama.org

HETEROGENEOUS SOFTWARE DEVELOPMENT ACCELERATED

Fraunhofer Institute for Algorithms and Scientific Computing SCAI
Schloss Birlinghoven
53754 Sankt Augustin
Germany

Contact:
Dr. Thomas Soddemann
Phone +49 2241 14-3414
thomas.soddemann@
scai.fraunhofer.de

www.scai.fraunhofer.de/hpc

Distributed by:
scapos AG
Schloss Birlinghoven
53574 Sankt Augustin
Germany

Phone: +49 2241 14-2820
info@scapos.com

www.scapos.com

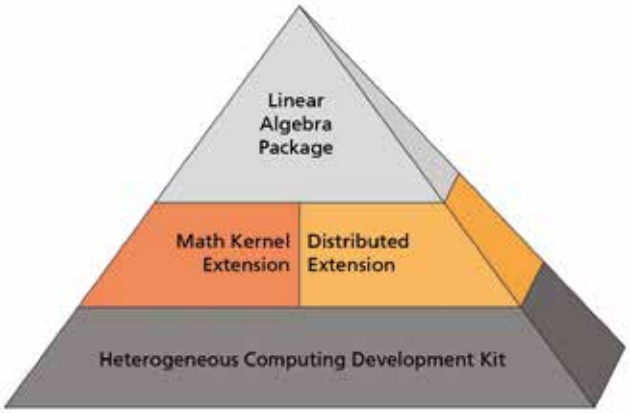
LAMA is a framework for developing hardware-independent, high performance code for heterogeneous computing systems. It facilitates the development of fast and scalable software that can be deployed on nearly every type of system, from embedded devices to highly parallel supercomputers, with a single code base. By using LAMA for their application, software developers benefit from higher productivity in the implementation phase and stay up to date with the latest hardware innovations, both leading to shorter time-to-market.

The framework supports multiple target platforms within a distributed heterogeneous environment. It offers optimized device code on the back-end side and high scalability through latency hiding and asynchronous execution across multiple nodes. LAMA's modular and extensible software design supports the developer on several levels, regardless of whether writing his own portable code with the *Heterogeneous Computing Development Kit* or using prepared functionality from the *Linear Algebra Package*, the user always gains high productivity and maximum performance.

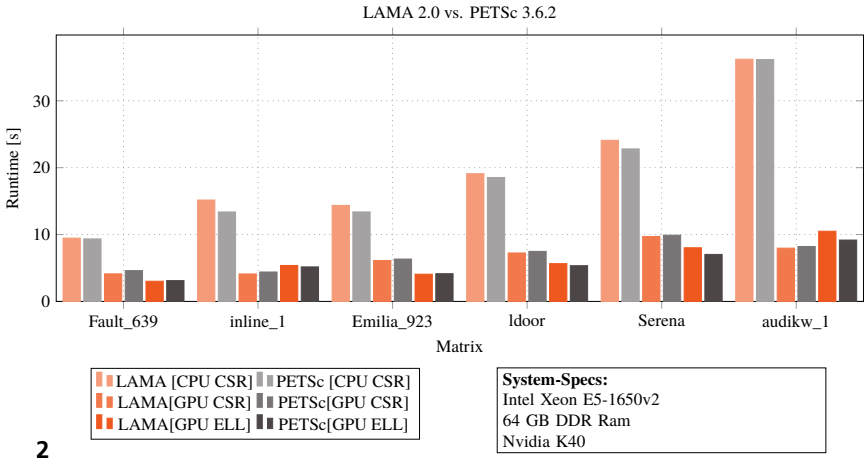
The integration of LAMA into other software products is easy and industry-friendly due to the dual license model: both the open-source AGPL and a commercial licence are offered.

Overview

LAMA is a multi-layer framework (fig. 1) offering four modules for fast heterogeneous software development. It targets multi-core CPUs, NVIDIA® GPUs and Intel® Xeon® Phi™'s – for single-node or multi-node usage. LAMA's flexible plug-in architecture allows a seamless integration of tomorrow´s CPU´s and accelerator hardware architectures, thus reducing maintenance costs. The *Heterogeneous Computing Development Kit* provides the management of heterogeneous memory and compute kernels. Asynchronous



1



2

1 *LAMA's multi-layer software stack: Encapsulated modules enable clean software development.*

2 *Performance Comparison with PETSc: Single Node Implementation of a CG solver on CPU and GPU.*

executing is a key capability. The *Math Kernel Extension* gives uniform access to dense and sparse compute kernel´s on all platforms while the *Distributed Extension* supplies full cluster support for scalability on data parallel applications. The *Linear Algebra Package* enables programming using mathematical notation and prepared iterative solvers.

Heterogeneous Computing Development Kit

LAMA's base module facilitates three issues:
(i) A consistent data usage on heterogeneous devices via dedicated read and write accesses within the memory management.
(ii) Decisions about the execution context in the application are separated from implementing the kernels by the kernel management.
(iii) Asynchronous execution of these kernels, memory transfer and communication is handled by the tasking layer.
This combination leads to a clean software development, accomplishing a good maintainability on the user's side with minimal runtime overhead.

Math Kernel Extension

The *Math Kernel Extension* provides access to hardware independent kernel routines for multiple purposes: basic utility functionality on arrays, dense and sparse BLAS operations as well as sparse conversion routines encapsulating MKL (BLAS), cuBLAS and cuSPARSE or own optimized kernels. Different sparse matrix formats are available to facilitate the application in various use cases and target architectures.

Distributed Extension

The *Distributed Extension* realises strategies for the distribution of data among processes and handles the communication between those. LAMA is compliant to MPI or GPI, giving the developer both options.

Linear Algebra Package

The *Linear Algebra Package* facilitates the development of (sparse) numerical algorithms for various application domains. Code can be written in text-book-syntax as

y = A * x

(where x and y are vectors and A is a matrix). Due to the underlying layers, the problem formulation is handled independently of the implementation details regardless of the target architecture and distribution strategy as memory management and communication is processed internally. Furthermore, with load balancing between different components and asynchronous execution, full system performance can be obtained.
In addition, LAMA offers various iterative solvers like Jacobi or CG methods, that can be used directly or preconditioned, with a combination of several user-definable stopping criteria. Furthermore, the integration of a custom-built solver is straightforward.

Performance

Productivity is combined with performance in execution – which is not mutually exclusive. LAMA's flexible software design introduces only a minimal overhead, conserving the full performance of the underlying BLAS implementations from the hardware vendors and from the highly optimized kernel back-ends. Performance comparison to concurring software libraries in the field of linear algebra show comparable results for single node implementations (fig. 2). On distributed systems the asynchronous execution model guarantees efficient overlapping of calculation, memory transfer and communication reaching linear scaling on GPUs. Convince yourself by having a look at our benchmarks on the LAMA website...

... or simply try it out yourself.

Website

Find more information about LAMA on www.libama.org