Rhea1: From Compilation to Execution

By **Etienne Renault**, Performance Engineering Manager

Introduction

European semiconductor industry has reached a new milestone!

SiPearl has officially achieved the bring up phase of the Rhea1 processor, coming out directly from the European Processor Initiative (EPI).

From the Mont-Blanc project to collaborative initiatives and Center of Excellence projects (AERO, Emopass, Excellerat, Helios, HIGHER, MAX, MicroLab, ODISSEE, OpenCUBE, PLASMA-PEPSC, RISER), Rhea1 is the most complex server CPU ever developed in Europe. With it, SiPearl is paving the way for demanding HPC and AI applications.

Let's dive into the architecture behind SiPearl's first-generation CPU.

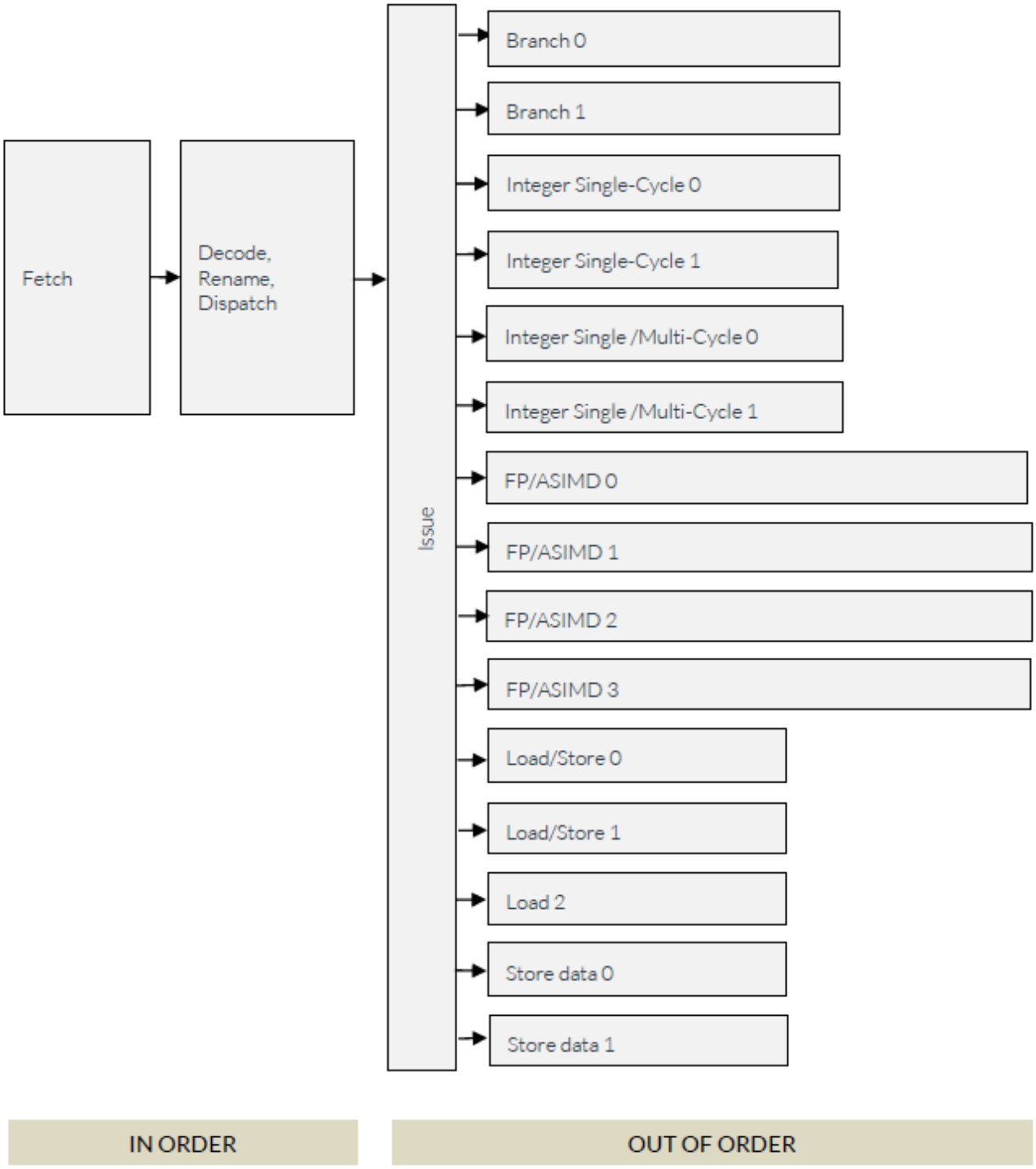
Key Features & Technical Specifications

Rhea1 is designed to handle efficiently memory bound workloads while offering an impressive compute capacity. This makes it the largest CPU built in Europe. Using 61 billion transistors and TSMC N6 process node, Rhea1 features:

- Compute Engine & ISA: Rhea1 is the biggest chip based on Neoverse V1 with 80 Arm Neoverse V1 cores, each core featuring two 256-bit Scalable Vector Extension (SVE) units. It enables a vector length-agnostic programming model for various datatypes: FP64 and FP32 for HPC simulation code and BF16 and INT8 for AI. This allows Rhea1 to be versatile among the workload it can handle.
- Cache Hierarchy & Coherent Mesh (NoC): To handle cache bounded applications and ensure a fast processing of data, each core is associated with a dedicated 64KB L1 instruction/data caches and a 1MB private L2 cache. The System Level Cache (SLC) is a pool of 80MB distributed across an advanced Arm CoreLink CMN-700 Coherent Mesh Network.
- High-Speed I/O Fabric: 104 lanes of PCIe Gen5 are also available, ensuring connectivity with a wide range of devices, including accelerators, NIC, and high-speed storage arrays.
- Tiered Memory Subsystem: Rhea1 proposes an innovative memory layout designed to handle efficiently all kind of workloads
  - In-package High Bandwidth Memory (4 HBM stacks) delivers the massive throughput to tackle memory-bound HPC applications like Computational Fluid Dynamics or Weather Modeling Grids.
  - High-Capacity Memory (4 DDR5 channels, supporting 2 DIMMs per channel) provides expansive system memory pools for massive datasets that exceed HBM capacity. It is also the best memory to handle latency bound applications.

ISA & Architecture

Rhea1 implements the ARMv8.4 instruction set, which brings specific instructions to execute advanced vector operations (like dot product, range extend single precision FMA), security (MPAM, nested virtualization) and cryptography (SHA2-512, SHA3, SM3, SM4). At the microarchitectural level, Rhea1 cores are based on Neoverse V1, which enriches the ISA enabling features that are crucial for HPC and AI, like SVE (with 256bit vectors) and bfloat16 support. For a more detailed description of the Neoverse V1 architecture, see the Arm Neoverse V1 documentation.



Source: Arm Neoverse V1 Software Optimization Guide. https://support.arm.com/documentation/110659/latest/

The backend includes functional units capable of executing scalar and vectorized code, supporting both NEON (fixed 128bit vectors) and SVE (Scalable Vector Extension). The scalable part of SVE is in the programming model: the developer doesn't need to know the vector length for the specific architecture in advance, but rather the ISA operates on the assumption that vectors are of arbitrary power of 2 lengths, and active elements are specified via predicates. Due to this, scalable vector code can run on machines with different hardware vector lengths.

On Rhea1, SVE vectors are 256bit wide, allowing four FP64 or eight FP32 elements per vector. Rhea1 provides two 256-bit SVE vector execution units and supports FP64, FP32, BF16, and INT8 operations.

Compilers, Flags & Optimization Impact

Neoverse V1 is well supported across the Arm compiler ecosystem, offering the developers several options from both open-source toolchains (GNU Compiler Collection, LLVM, Arm Toolchain for Linux) and vendor specific compilers (NVIDIA HPC Compilers).

To squeeze the best performance out of Rhea1, the developer needs to target both the correct instruction set and target microarchitecture. Using Clang, a typical compile command is:

```shell
clang -mcpu=neoverse-v1 source.c -o output
```

In this command, the -mcpu option describes the target CPU, specifying both its instruction set (ARMv8.4-a with, among other things, the SVE and BF16 extensions enabled) and the way the instructions should be generated in order to optimise pipeline scheduling.

Compiling using the -mcpu option does not necessarily force the compiler to emit only SVE instructions for loop vectorization, the strategy can be still controlled explicitly. For instance, this command would allow the compiler to only emit NEON instructions for vectorization.

```shell
clang -march=armv8.4-a+nosve \
      -mllvm -scalable-vectorization=off \
      source.c -o output
```

Whereas this command allows the compiler to assume a specific length of SVE registers (in this case, 256bit), which could enable more optimizations.

```shell
clang -O3 -march=armv8.4-a+sve \
      -mllvm -scalable-vectorization=preferred \
      -msve-vector-bits=256 \
      source.c -o output
```

These two configurations provide a useful basis for comparing conventional NEON vectorization with SVE-based scalable vectorization on Rhea1.

For further reading on benchmarking and vectorization impact on Arm architecture, see our article published at the ISC High Performance 2025 conference, also available as a video.

Current Software Ecosystem Tested on Rhea1

Fully exploiting Rhea1 requires a mature software stack. Thanks to its Arm architecture, the ecosystem is fully prepared out-of-the-box.

To demonstrate this maturity, we have tested a wide variety of software. As of today, it includes open-source **Compilers and Toolchains** (GCC, LLVM/Clang); **Numerical and System Libraries** (OpenBLAS, FFTW, SLEEF, Boost, HDF5, netCDF); **Runtimes** (OpenMP, Kokkos); **Frameworks and AI Stacks** (PyTorch, TensorFlow, ONNX Runtime); and **Inference Engines** (vLLM, llama.cpp). In the background, various hardware **Verification and Stress Suites** have also been successfully executed. The table below depicts the software that has been tested, whilst this catalog provides an up-to-date list.

All these results demonstrate that Rhea1 is ready to drive European exascale computing and AI from day one.

SOFTWARE	VERSION	LICENSE	DESCRIPTION	CATEGORY	URL
ARM Compute Library	52.4.0	MIT	The Compute Library is a collection of low-level machine learning functions optimized for Arm® Cortex®-A, Arm® Neoverse™ and Arm® Mali™ GPUs architectures	Library	link
BigDFT	1.9.6	https://l_sim.gitlab.io/bigdft-suite/overview/license.html	BigDFT is a program for performing Density Functional Theory calculations on a wide class of systems	Application	link
Blogbench	1.2	ISC	Filesystem benchmark that tries to reproduce the load of a real-world busy file server	Benchmark	link
Cloverleaf	1.3.0	LGPL-3.0	A hydrodynamics mini-app to solve the compressible Euler equations in 2D, using an explicit, second-order method.	Benchmark	link
CodeSaturne	v9.2.0	GPL-2.0	CFD tool, EDF's general purpose Computational Fluid Dynamics (CFD) software.	Application	link
Dwarf-p-cloudsc	1.5.3	Apache License 2.0	Dwarf-p-cloudsc is intended to test the CLOUDSC cloud microphysics scheme of the IFS model	Application	link
EcTrans	1.8.0	Apache License 2.0	ecTrans is a library for performing efficient and scalable spectral transformations	Library	link
Ecrad	1.7.2	Apache License 2.0	The ECMWF atmospheric radiation scheme	Benchmark	link
FLEUR	MaX-8.0	MIT	Simulation tool for materials properties using density functional theory and related methods	Application	link
HPCG	3.1	BSD-3	Synthetic benchmark using conjugate gradient computation	Benchmark	link
HPCG - Nvidia	26.02	Apache License 2.0	Synthetic benchmark using conjugate gradient computation, optimized for Nvidia GPUs and ARM CPUs	Benchmark	link
HPCW	3.0	Apache License 2.0	Benchmark suite for (pre)-exascale application of climate and weather codes	Benchmark	link
HPL	2.3	Knoxville university - https://github.com/icl-utk-edu/hpl/?tab=License-1-ov-file	Synthetic benchmark for dense linear systems	Benchmark	link
HydroCplus	614628c	CeCILL	A 2D Hydro code for benchmarking purpose.	Benchmark	link
Icon	2025.10-2	Apache License 2.0	Numerical weather forecasting and climate simulation	Benchmark	link
Incompact3d	2.1.1	BSD-3	Navier-Stokes equations CPU Solver	Benchmark	link

Jemalloc	5.3.3	https://github.com/jemalloc/jemalloc?tab=License-1-ov-file#License-1-ov-file	jemalloc is a general purpose malloc(3) implementation that emphasizes fragmentation avoidance and scalable concurrency support.	Library	link
KleidiAI	b8de5c1aa	BSD-3-Clause	KleidiAI is an open-source library that provides optimized performance-critical routines, also known as micro-kernels, for artificial intelligence (AI) workloads tailored for Arm® CPUs	Library	link
LMBench	Commit 5a386c1	GPL-2.0	Memory characterization benchmark suite	Benchmark	link
LULESH	2.0	BSD-3	LLNL Proxy-application	Benchmark	link
Linux kernel	6.18	GPL-2.0	The Linux kernel is a free, open-source, monolithic operating system core created by Linus Torvalds in 1991 that acts as the primary bridge between computer hardware and software applications	Kernel	link
MLPerf - ResNet	5.1.1	Apache-2.0	MLPerf Resnet classification benchmark. It is used to test the performance of platforms while running this model	Benchmark	link
MLPerf - RetinaNet	5.1.1	Apache-2.0	MLPerf RetinaNet object detection benchmark. It is used to test the performance of platforms while running this model	Benchmark	link
MLPerf - llama3.2-1B	5.1.1	Apache-2.0	MLPerf llama3.2 large language model benchmark. It is used to test the performance of platforms running this LLM	Benchmark	link
MiniFE	2.2.0	LGPL-3.0	Finite Element Mini-Application of the Mantevo Project	Benchmark	link
Mocassin	2.02.73.2	GPL	MOCASSIN is 3D or 2D photoionisation and dust radiative transfer code used to study transfer of radiation through media of arbitrary geometry and density distribution	Application	link
Nastja	b72c25ae	Mozilla Public License Version 2.0	The NAStJA framework provides an easy way to enable massively parallel simulations for a wide range of multi-physics applications based on stencil algorithms. It supports the development of parallel strategies for high-performance computing.	Application	link
Neko	v1.1.2	BSD-3-Clause	Portable framework for high-order spectral element flow simulations.	Application	link
Nemo	5.0	Apache License 2.0	Ocean General Circulation Model	Benchmark	link
Nicam-DC	2024.06-04	Apache License 2.0	Non-hydrostatic ICosahedral Atmospheric Model (NICAM), a global climate and weather simulation model	Benchmark	link
ONNXRuntime	1.22.1	MIT	ONNX Runtime is a cross-platform inference and training machine-learning accelerator	Framework	link
OSU micro-benchmark	7.0.1	BSD license	The OSU Micro-Benchmarks (OMB) are a suite of MPI performance tests designed to measure latency, bandwidth, and other communication	Application	link

			metrics for parallel computing applications.		
OneAPI Construction Kit	v5.0.0	Apache 2.0 with LLVM Exceptions	Framework to provide implementations of open standards, such as OpenCL, for a wide range of devices.	Tool	link
PENNANT	1.1.0	Triad National Security	LANL Unstructured mesh hydrodynamics for advanced architectures	Benchmark	link
PyTorch	2.9.1	BSD-3-Clause	PyTorch is a Python package used for tensor computation with strong GPU acceleration and with DNN built on a tape-based autograd system.	Framework	link
Quantum Espresso	7.6	GPL v2.0	Quantum ESPRESSO suite of codes (ESPRESSO:opEn-Source Package for Research in Electronic Structure, Simulation, and Optimization	Application	link
SPECCPU2017	V1.1	SPEC CPU®2017 License Agreement	The SPEC CPU® 2017 benchmark benchmark package contains SPEC's next-generation, industry-standardized, CPU intensive suites for measuring and comparing compute intensive performance, stressing a system's processor, memory subsystem and compiler.	Application	link
STREAmS-2	v2.1.0	GPL-3.0	Direct Numerical Simulations of compressible turbulent flows solving the unsteady, fully compressible Navier-Stokes equations for a perfect gas.	Application	link
Siesta	5.4.2	GPLv3	Performs efficient electronic structure calculations and ab initio molecular dynamics simulations of molecules and solids.	Application	link
Stream	5.10	GPL-like	The STREAM benchmark is a simple synthetic benchmark program that measures sustainable memory bandwidth	Application	link
TensorFlow	2.14.0	Apache-2.0	TensorFlow is an end-to-end open source platform for machine learning	Framework	link
TornadoVM	v6.0.1-jdk21	Apache-2.0, GPL-2.0	GPU/CPU programming framework for Java that works with JDK 21+	Tool	link
WarpX	24.10	BSD-3-Clause	WarpX is an advanced electromagnetic & electrostatic Particle-In-Cell code	Benchmark	link
XGBoost	3.4.1	Apache License 2.0	XGBoost is an optimized distributed gradient boosting library.	Library	link
YCSB	0.17.0	Apache-2.0	Yahoo! Cloud Serving Benchmark	Benchmark	link
Yambo	5.3	GPL-2.0	Code implementing first-principles methods based on Green's function theory to describe excited-state properties of realistic materials	Application	link
acfl	24.10.1	Arm End User License Agreement	Arm Compiler for Linux	Tool	link
adativecpp	25.02	BSD-2-Clause	AdaptiveCpp is the independent, community-driven modern platform for C++-based heterogeneous programming models targeting CPUs and GPUs from all major vendors	Library	link
arm_neon_tests	Trunk	MIT	Tests for the ARM/Neon instructions	Testsuite	link

atfl	25.07.1	Apache 2.0	Arm Toolchain for Linux	Tool	link
binutils-gdb	16.3	GPL-2	This directory contains various GNU compilers, assemblers, linkers, ...	Tools	link
blis	2.0	BSD-3-Clause	BLAS-like Library Instantiation Software Framework	Library	link
boost	1.84.0	BSL-1.0	The Boost C++ Libraries are a set of libraries for the C++ programming language that provides support for tasks and structures such as linear algebra, pseudorandom number generation, multithreading, image processing, regular expressions, and unit testing.	Library	link
boost	1.88.0	Boost Software License v1.0	The Boost project provides free peer-reviewed portable C++ source libraries.	Library	link
circustent	Trunk	Apache 2.0	Atomic Memory Operation System Benchmarks	Library	link
cmake	3.30.1	BSD 3-Clause	CMake is a cross-platform, open-source build system generator	Tool	link
farm_sve	Trunk	MIT	A scalar C++ implementation of the ARM® Scalable Vector Extension (SVE) testsuite	Testsuite	link
fftw3	0.3.10	GNU General Public v2.0	FFTW is a free collection of fast C routines for computing the Discrete Fourier Transform in one or more dimensions.	Library	link
fleetbench	2.1	Apache-2.0	Benchmarking suite for google workloads	Benchmark	link
gcc	15.2.0	GNU LGPLv2.1	The GNU Compiler Collection is free software	Compiler	link
gdb	16.3	GPL-2	This directory contains various GNU compilers, assemblers, linkers, ...	Tool	link
gmp	6.3.0	GNU LGPLv3	The GNU MP Library is free software	Library	link
hdf5	1.14.5	Modified BSD 3-Clause	HDF5 is a file format for efficiently storing and organizing large datasets, metadata, and hierarchical data	Library	link
hecbench	2e066aca	BSD-3-Clause	collection of heterogeneous computing benchmarks written with CUDA, HIP, SYCL/DPC++, and OpenMP-4.5 target offloading for studying performance, portability, and productivity.	Benchmarks	link
kokkos	5.1.1	Apache 2.0	Kokkos Core implements a programming model in C++ for writing performance portable applications targeting all major HPC platforms	Library	link
libc-test	Trunk	MIT	libc-test is developed as part of the musl project	Testsuite	link
libflame	5.2.0	BSD-3-Clause	High-performance object-based library for DLA computations (LAPACK)	Library	link
libpfm4	4.13.0	MIT	A helper library to program the performance monitoring events	Library	link
llama.cpp	f7873fc69	MIT	llama.cpp is a framework to enable LLM inference with minimal setup and state-	Framework	link

			of-the-art performance on a wide range of hardware.		
llvm-project	22.1.4	Apache 2.0	LLVM, a toolkit for the construction of highly optimized compilers, optimizers, and run-time environments.	Compiler	link
llvm-test-suite	llvmorg-22.1.1	Apache 2.0	LLVM "test-suite"	Testsuite	link
maqao	2025.1.0	LGPL-3.0	MAQAO is a performance analysis and optimization framework operating at binary level with a focus on core performance	Tool	link
miniBUDE	40254ff	Apache-2.0	A BUDE virtual-screening benchmark, in many programming models	Benchmark	link
mpfr	4.2.2	GNU LGPLv3	The GNU MPFR Library is free software	Library	link
netcdf-c	4.10.1	BSD-3-Clause	netCDF-C is a C library for reading and writing scientific data in netCDF files	Library	link
netcdf-fortran	4.6.4	Apache 2.0	netCDF-Fortran is a Fortran interface for reading and writing scientific data in netCDF files	Library	link
ninja	1.13.1	Apache-2.0	Ninja is a small build system with a focus on speed.	Tool	link
nvpl	25.5	NVIDIA Software License Agreement	NVIDIA Performance Libraries	Tool	link
oneDNN	3.9.1	Apache-2.0	oneAPI Deep Neural Network Library (oneDNN) is an open-source cross-platform performance library of basic building blocks for deep learning applications	Library	link
openblas	0.3.31	BSD-3-Clause	OpenBLAS is an optimized BLAS (Basic Linear Algebra Subprograms) library based on GotoBLAS2 1.13 BSD version.	Library	link
openmp_vv	Trunk	BSD-3-Clause	OpenMP Validation and Verification project	Testsuite	link
openmpi	5.0.2	BSD-3-Clause	The Open MPI Project is an open source Message Passing Interface implementation that is developed and maintained by a consortium of academic, research, and industry partners.	Library	link
redis	8.10.2	AGPLv3	The Redis in memory database	Application	link
scalapack	2.2.3	BSD-3-Clause	ScaLAPACK is a library of high performance linear algebra routines for distributed memory computers supporting MPI	Library	link
sleef	3.7	Boost Software-1.0	SLEEF is a library that implements vectorized versions of C standard math functions	Library	link
stress-ng	V0.22.00	GPL-2.0	stress-ng will stress test a computer system in various selectable ways. It was designed to exercise various physical subsystems of a computer as well as the various operating system kernel interfaces.	Kernel	link
sycl-cts	6a7dd02b	Apache-2.0	This is the SYCL Conformance Test Suite for the Khronos Group SYCL standard.	Testsuite	link

vLLM	0.15.0	Apache-2.0	vLLM is a fast and easy-to-use library for LLM inference and serving	Library	link
valgrind	3.25.1	GNU Free Documentation v1.2	Valgrind is a framework for building dynamic analysis tools	Tool	link
vlasiator	271ab544 (dev)	GPL-2	Vlasiator is the state-of-the-art hybrid-Vlasov simulation for ion-scale physics in a global magnetospheric setting	Application	link
zoltan	3.901	BSD-3-Clause	Parallel Partitioning, Load Balancing and Data-Management Services	Library	link