Optimizing SPH-EXA for AMD GPUs

LUMI Optimizing for AMD GPUs Hackathon
October 14-18, 2024, Brussels

Osman Seckin Simsek

SPH EXA SKACH

pasc Platform for Advanced Scientific Computing

https://github.com/unibas-dmi-hpc/SPH-EXA

Download SPH-EXA

Universität Basel

Florina Ciorba (PI)
Ruben Cabezon (Co-PI)
Osman Seckin Simsek
Yiqing Zhu
Lukas Schmidt
José Escartin

Universität Zürich UZH

Lucio Mayer (Co-PI)
Noah Kubli
Darren Reed

cscs

Sebastian Keller
Jean-Guillaume Piccinali
Jean Favre
Jonathan Coles

Axel Sanz (UPC)
Joseph Touzet (Paris-Saclay)

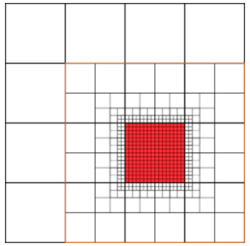
SPH-EXA: Smoothed Particle Hydrodynamics at Exascale

SPH-EXA is a *scalable* **s**moothed **p**article **h**ydrodynamics simulation framework **interdisciplinarily co-designed** by computational physicists and computer scientists to exploit **Exa**scale supercomputers.

SPH-EXA: Framework Components

SPH-EXA application front-end

Parallel I/O and test case setup (ICs)

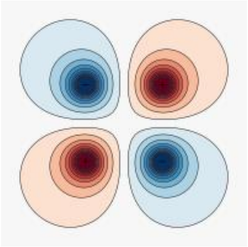


Cornerstone

Octree and domain decomposition framework

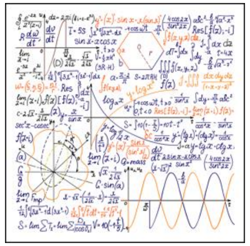
SPH

Hydrodynamics solver



Ryoanji

N-body gravity solver



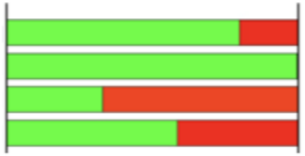
Physics modules

Radiative cooling
Nuclear reactions
Star formation
Stellar feedback

SPH-EXA: Optimization Strategy

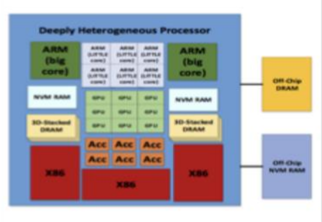
Scalability

weak, strong



Scheduling & Load-balancing

dynamic, adaptive asynchronous execution



Heterogeneity

portability on various CPU and GPU architectures

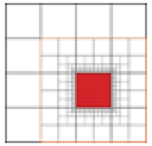
Fault-tolerance

silent error detection and recovery

Energy

measurement reporting efficiency

SPH-EXA: Framework Components



Cornerstone octree

Domain Decomposition

- Space-filling curves and octrees
- Global and locally essential octrees
- Octree-based domain decomposition
- 21’600 lines of code

SPHYNX

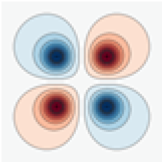
ChaNGa

SPH Solver
Physics modules

Modern SPH and physics implementation with key features

(astro.physik.unibas.ch/sphynx, https://github.com/N-BodyShop/changa):

- Generalized volume elements
- Integral approach to derivatives
- Artificial viscosity with switches
- Sub-grid physics
- 3’800 lines of code



Ryoanji N-body solver

Gravity-solver on GPUs with:

- Cornerstone octrees
- Breadth-first traversal inspired by Bonsai (https://github.com/treecode/Bonsai)
- EXA-FMM multipole kernels (https://github.com/exafmm)
- 4’100 lines of code

SPH-EXA

SPH-EXA application front-end

- Handling of initial conditions, checkpointing and I/O, including compression
- Flexible combination and addition of additional physics for domain scientists
- Performance data and energy consumption measurements
- In-situ visualization
- 7’200 lines of code

SPH-EXA: A Production Code, Easy to Use like a Mini-app

https://github.com/unibas-dmi-hpc/SPH-EXA, v0.8

MPI OpenMP

C++ 20 (GCC 11+)

Cmake 3.22+

CUDA 11.2+

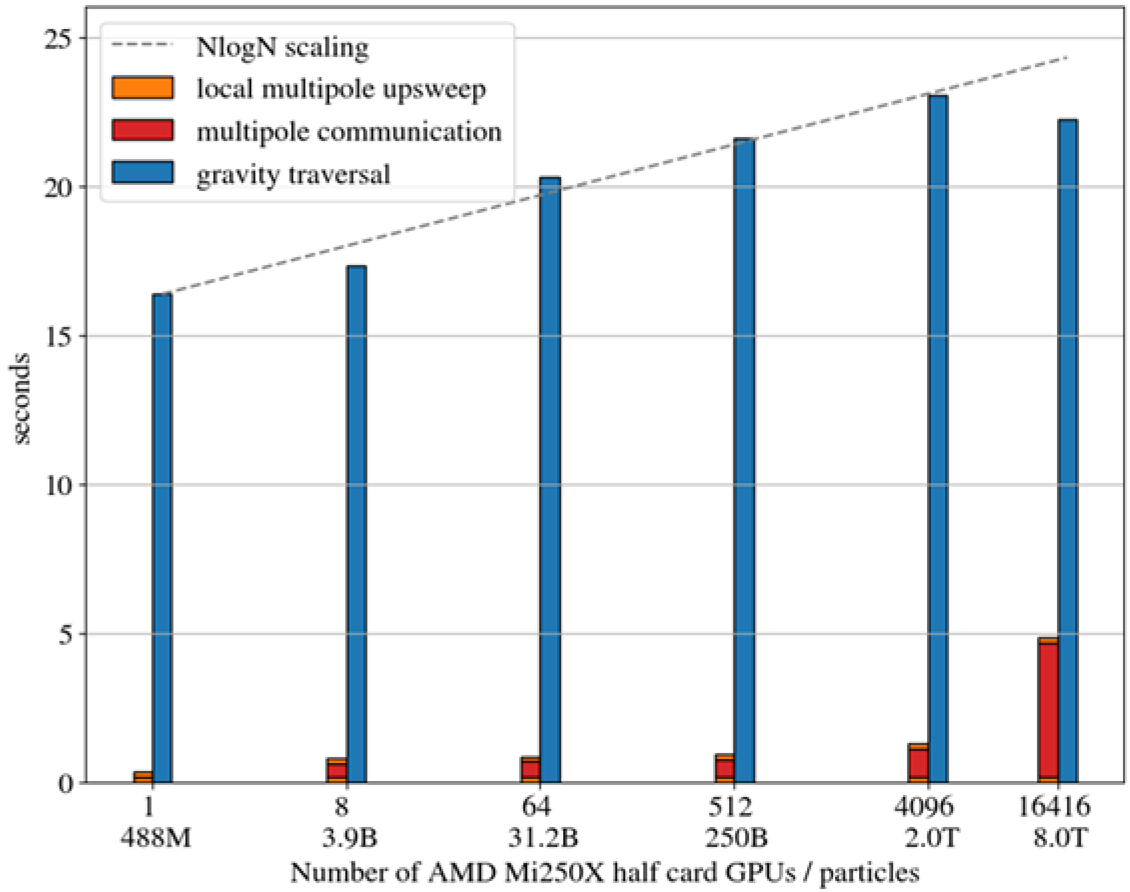
HIP 5.2+

HDF5 1.10+

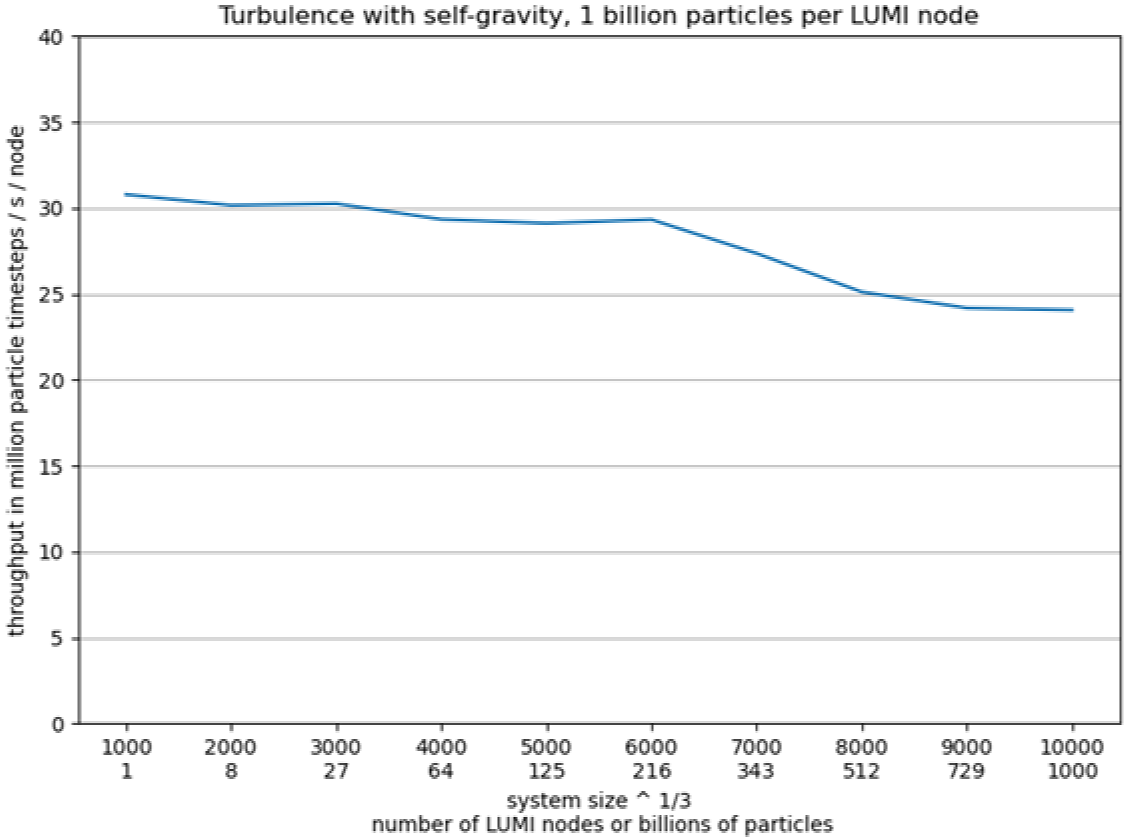
```
$> git clone https://github.com/unibas-dmi-hpc/SPH-EXA.git
$> cd SPH-EXA
$SPH-EXA> mkdir build
$SPH-EXA> cd build
$SPH-EXA/build> cmake ..
.
. Output
.
$SPH-EXA/build> make -j
.
. Output
.
$SPH-EXA/build> cd main/src/sphexa
$SPH-EXA/build/main/src/sphexa> ls
sphexa
sphexa-cuda
$SPH-EXA/build/main/src/sphexa>
```

OPTIONAL [Ascent ParaView Catalyst ADIOS2]

SPH-EXA: Scalability

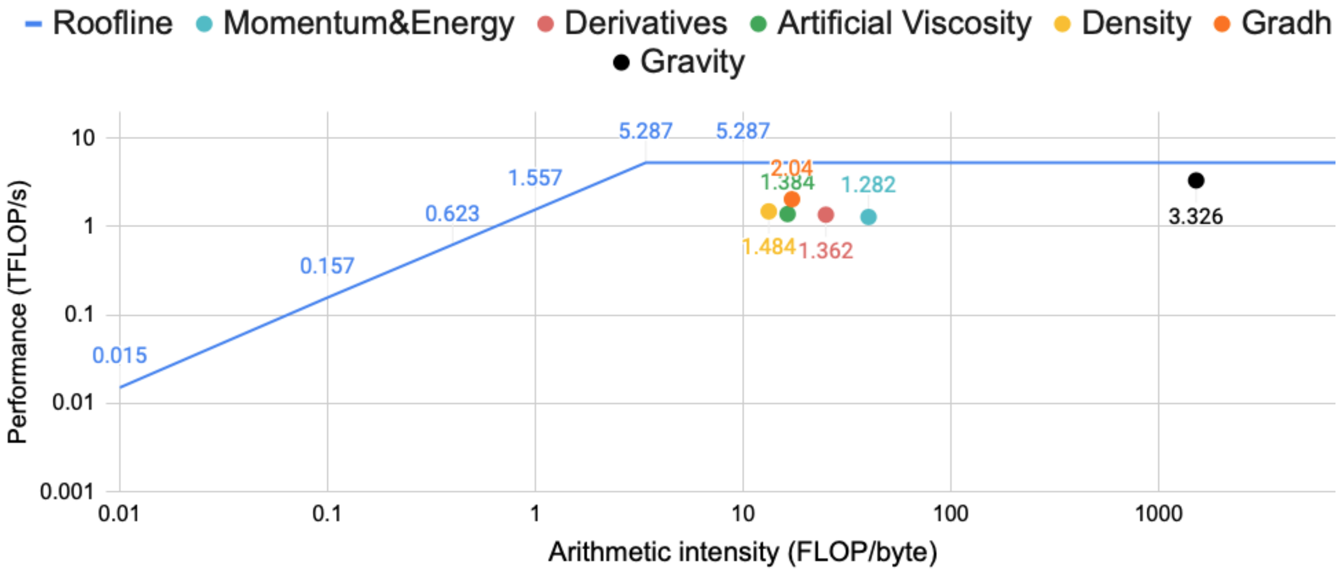


Close to logarithmic (Nlog(N)) weak scalability of the gravity solver in SPH-EXA on LUMI-G ***up to 8 trillion particles.*** *(Keller et al. PASC'23 Proceedings, 18, 2023)*

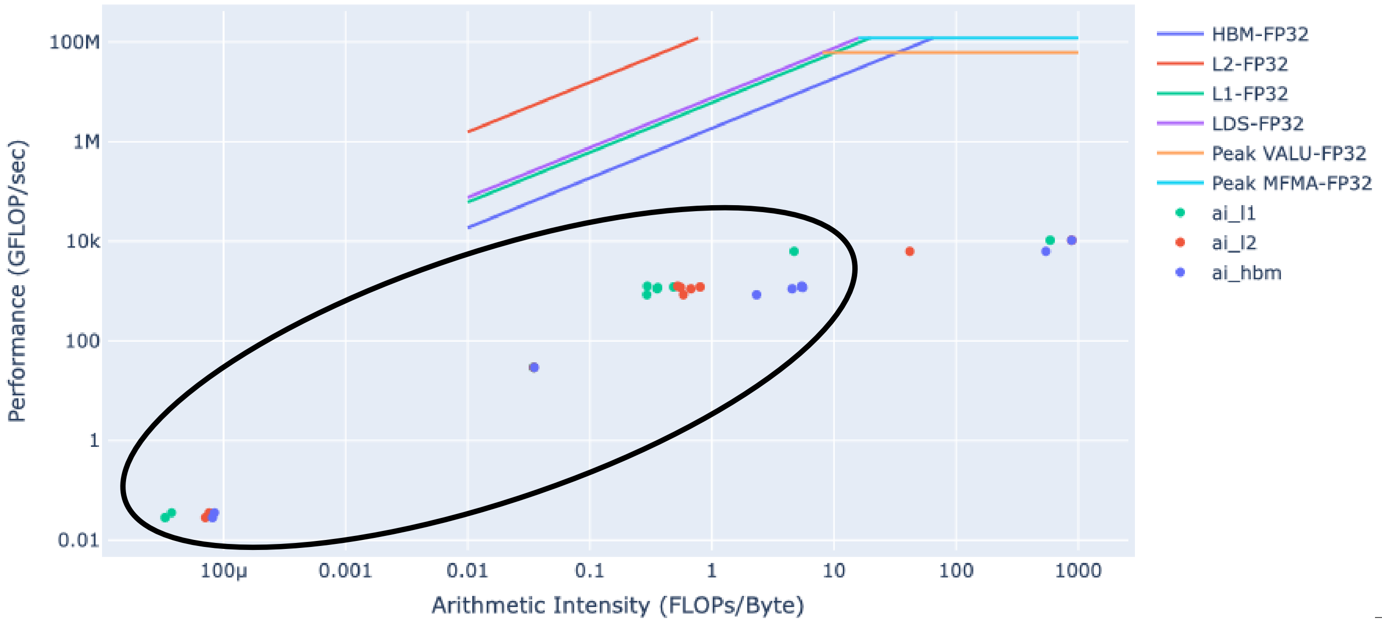


Weak scaling of SPH-EXA on LUMI-G, running 8 MPI ranks/node (1 MPI rank/ GPU half-card) and 1 billion particles/node ***up to 1 trillion particles*** *(SPH-EXA team, 2023)*

Motivation for Joining the LUMI Hackathon



Performance of the most time consuming kernels on a single Nvidia A100 GPU.

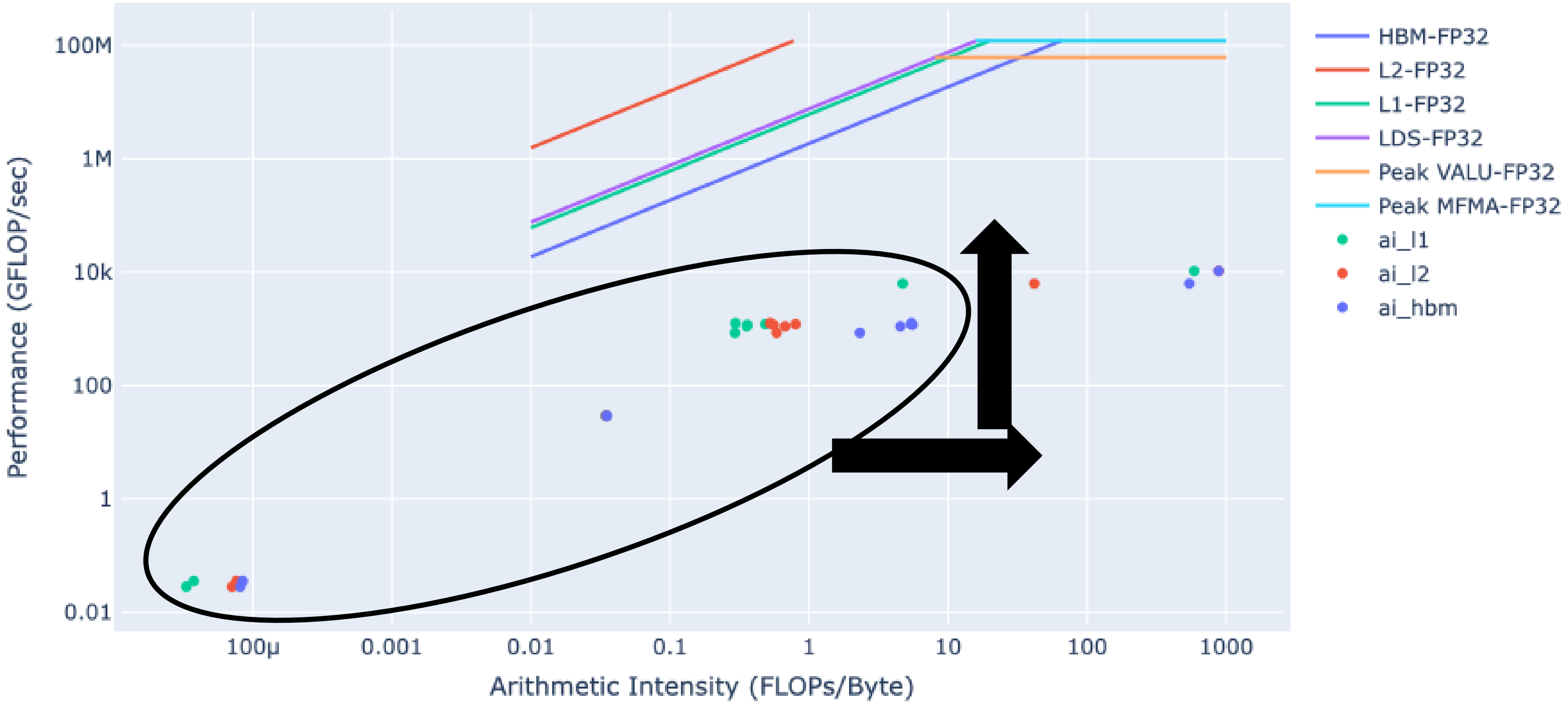


Performance of the most time consuming kernels on a single MI250X GCD.

SPH-EXA: Details of Functions per time-step

Function Name	Percentage of time per time-step
momentumEnergy	25.46%
gravity	21.70%
iadDivvCurlv	13.21%
AVSwitches	12.78%
veDefGradh	10.77%
xMass	10.03%
Total	**93.95%**

SPH-EXA: Plan for LUMI Hackathon



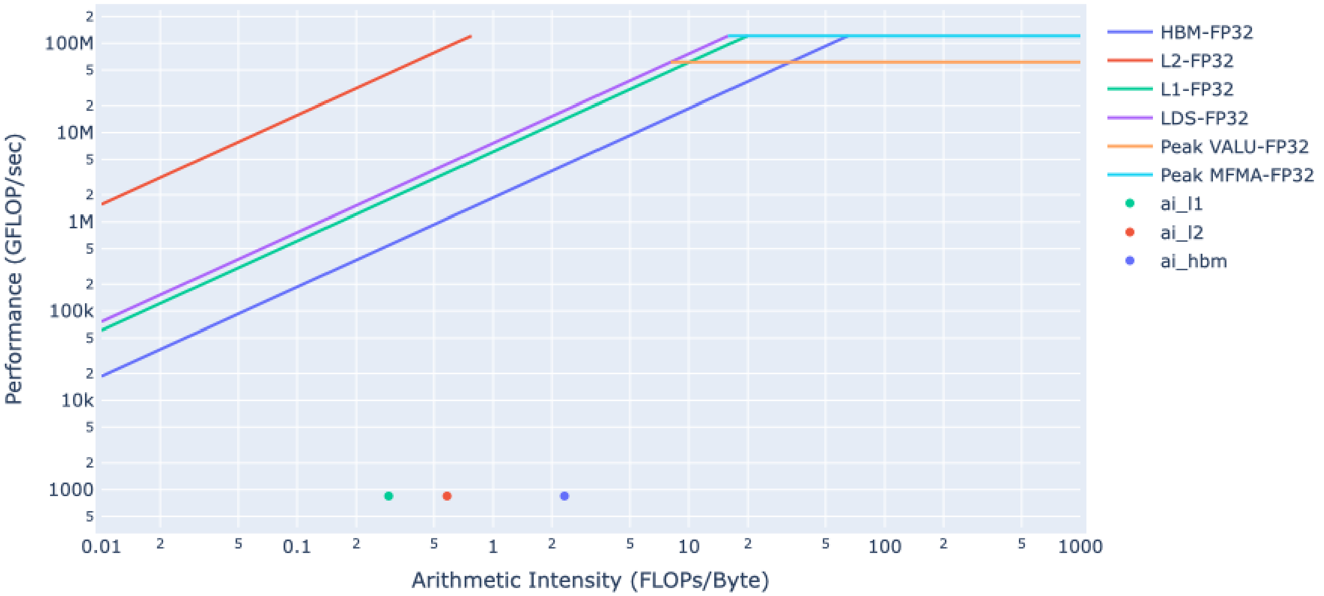
- Optimize per function for:
  - Increasing the arithmetic intensity
  - Increasing the performance
- Porting Nuclear-networks computations to GPU

Backup Slides

SPH-EXA: Details of momentumEnergy Function

0.1 Top Kernels

	Count	Sum(ns)	Mean(ns)	Median(ns)	Pct
::momentumEnergyGpu<false, double, double, unsigned long>(double, double, float, unsigned int, uble>, unsigned int const*, unsigned ctreeNsView<double, unsigned long>, double const*, double const*, float const*, float const*, float const*, float const*, float const*, float const*, float const*, float const*, float const*, float const*, float const*, float const*, float const*, float const*, float	1.00	5625098041.00	5625098041.00	5625098041.00	25.46



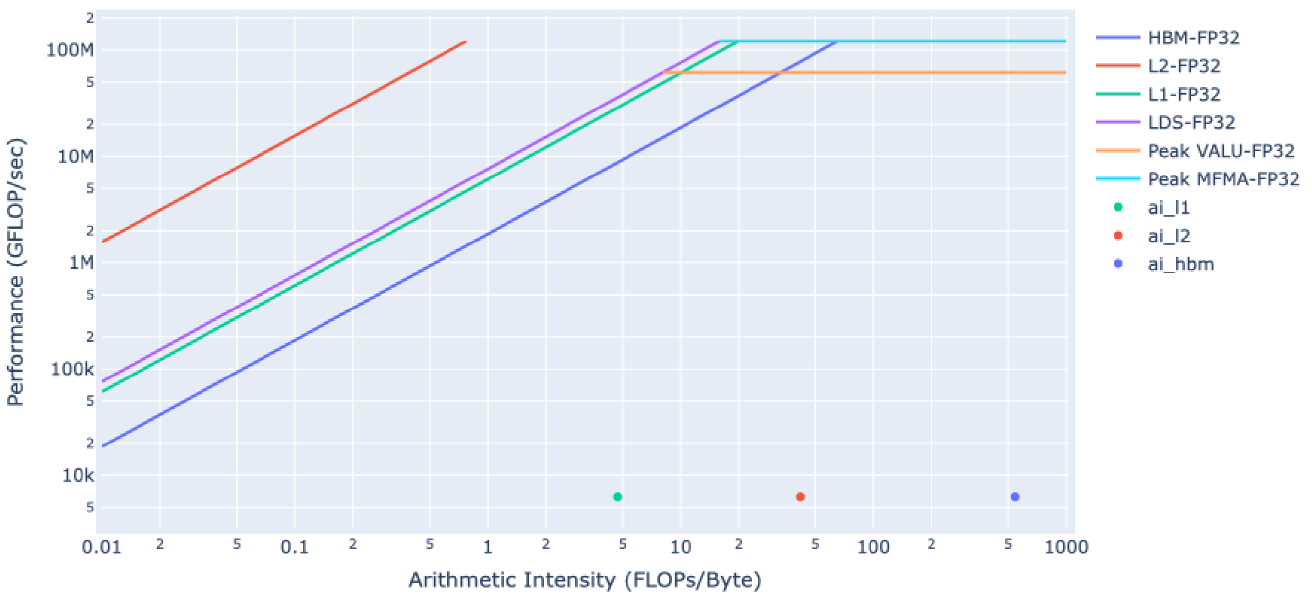
2.1 Speed-of-Light

Metric	Avg	Unit	Peak	Pct of Peak
VALU FLOPs	847.95	Gflop	23936.00	3.54
VALU IOPs		Giop	23936.00	
MFMA FLOPs (BF16)	0.00	Gflop	191488.00	0.00
MFMA FLOPs (F16)	0.00	Gflop	191488.00	0.00
MFMA FLOPs (F32)	0.00	Gflop	47872.00	0.00
MFMA FLOPs (F64)	0.00	Gflop	47872.00	0.00
MFMA IOPs (Int8)		Giop	191488.00	
Active CUs		Cus	110.00	
SALU Utilization		Pct	100.00	
VALU Utilization		Pct	100.00	
MFMA Utilization		Pct	100.00	
VMEM Utilization		Pct	100.00	
Branch Utilization		Pct	100.00	
VALU Active Threads		Threads	64.00	
IPC		Instr/cycle	5.00	
Wavefront Occupancy		Wavefronts	3520.00	
Theoretical LDS Bandwidth	2745.04	Gb/s	23936.00	11.47
LDS Bank Conflicts/Access	0.00	Conflicts/access	32.00	0.00
vL1D Cache Hit Rate	49.79	Pct	100.00	49.79
vL1D Cache BW	2893.61	Gb/s	11968.00	24.18
L2 Cache Hit Rate		Pct	100.00	
L2 Cache BW		Gb/s	3481.60	
L2-Fabric Read BW	291.57	Gb/s	1638.40	17.80
L2-Fabric Write BW	73.79	Gb/s	1638.40	4.50
L2-Fabric Read Latency		Cycles		
L2-Fabric Write Latency		Cycles		
sL1D Cache Hit Rate		Pct	100.00	
sL1D Cache BW		Gb/s	6092.80	
L1I Hit Rate		Pct	100.00	
L1I BW		Gb/s	6092.80	
L1I Fetch Latency		Cycles		

SPH-EXA: Details of gravity Function

0.1 Top Kernels

	Count	Sum(ns)	Mean(ns)	Median(ns)	Pct
traverse<double, float, float, float, array<float, 8ul> >(unsigned int const*, le const*, double const*, double const*, float const*, int const*, int const*, onst*, util::array<double, 4ul> const*, oat, 8ul> const*, double, float*, , float*, int*) [clone .kd]	1.00	4795169057.00	4795169057.00	4795169057.00	21.70



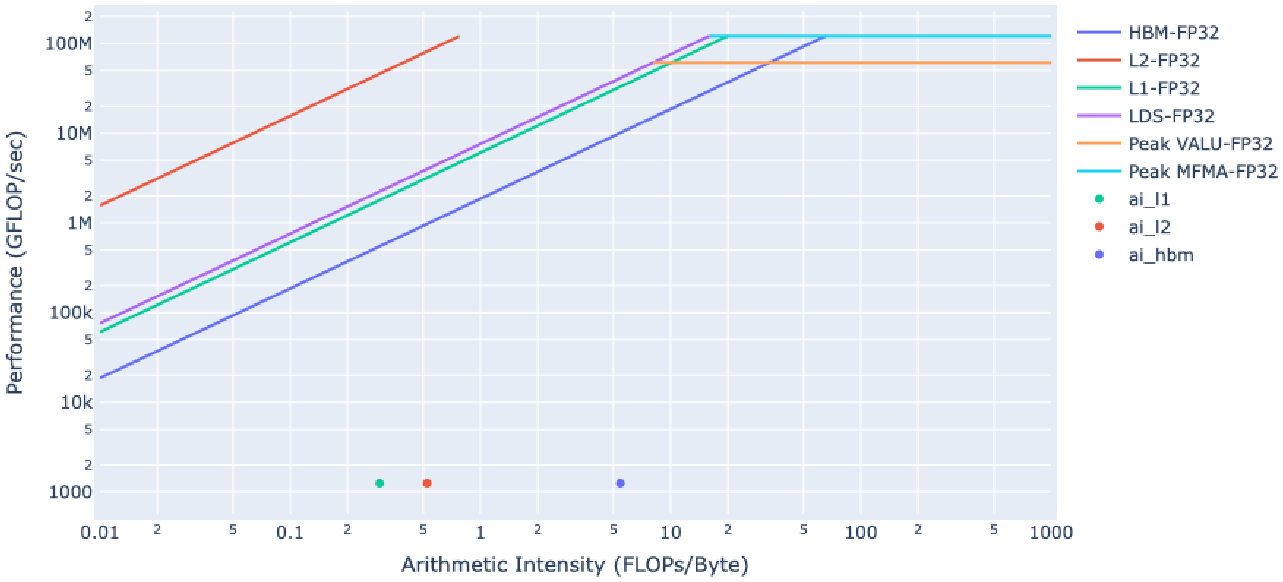
2.1 Speed-of-Light

Metric	Avg	Unit	Peak	Pct of Peak
VALU FLOPs	6285.58	Gflop	23936.00	26.26
VALU IOPs		Giop	23936.00	
MFMA FLOPs (BF16)	0.00	Gflop	191488.00	0.00
MFMA FLOPs (F16)	0.00	Gflop	191488.00	0.00
MFMA FLOPs (F32)	0.00	Gflop	47872.00	0.00
MFMA FLOPs (F64)	0.00	Gflop	47872.00	0.00
MFMA IOPs (Int8)		Giop	191488.00	
Active CUs		Cus	110.00	
SALU Utilization		Pct	100.00	
VALU Utilization		Pct	100.00	
MFMA Utilization		Pct	100.00	
VMEM Utilization		Pct	100.00	
Branch Utilization		Pct	100.00	
VALU Active Threads		Threads	64.00	
IPC		Instr/cycle	5.00	
Wavefront Occupancy		Wavefronts	3520.00	
Theoretical LDS Bandwidth	12116.87	Gb/s	23936.00	50.62
LDS Bank Conflicts/Access	0.00	Conflicts/access	32.00	0.00
vL1D Cache Hit Rate	88.71	Pct	100.00	88.71
vL1D Cache BW	1334.77	Gb/s	11968.00	11.15
L2 Cache Hit Rate		Pct	100.00	
L2 Cache BW		Gb/s	3481.60	
L2-Fabric Read BW	10.75	Gb/s	1638.40	0.66
L2-Fabric Write BW	0.83	Gb/s	1638.40	0.05
L2-Fabric Read Latency		Cycles		
L2-Fabric Write Latency		Cycles		
sL1D Cache Hit Rate		Pct	100.00	
sL1D Cache BW		Gb/s	6092.80	
L1I Hit Rate		Pct	100.00	
L1I BW		Gb/s	6092.80	
L1I Fetch Latency		Cycles		

SPH-EXA: Details of iadDivvCurlv Function

0.1 Top Kernels

	Count	Sum(ns)	Mean(ns)	Median(ns)	Pct
::iadDivvCurlvGpu<double, float, (double, unsigned int, uble>, unsigned int const*, unsigned ctreeNsView<double, unsigned long>, double const*, double const*, float const*, float const*, float const*, float const*, float const*, float , float*, float*, float*, float*, , float*, float*, float*, float*, , float*, unsigned int*, int*, bool)	1.00	2918764091.00	2918764091.00	2918764091.00	13.21



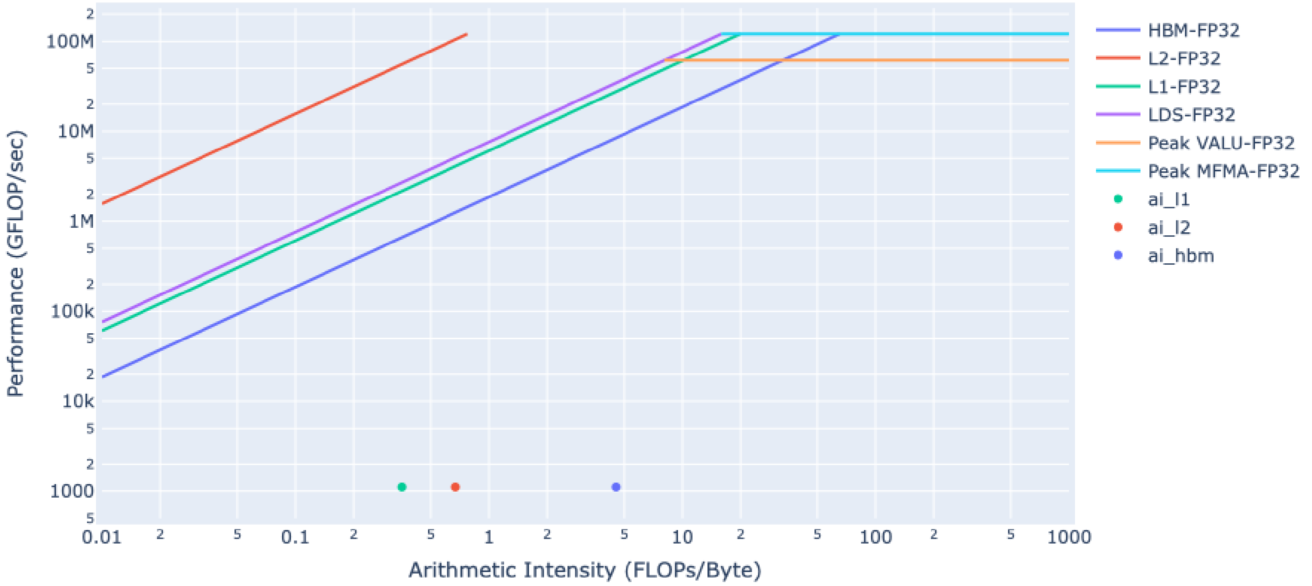
2.1 Speed-of-Light

Metric	Avg	Unit	Peak	Pct of Peak
VALU FLOPs	1257.45	Gflop	23936.00	5.25
VALU IOPs		Giop	23936.00	
MFMA FLOPs (BF16)	0.00	Gflop	191488.00	0.00
MFMA FLOPs (F16)	0.00	Gflop	191488.00	0.00
MFMA FLOPs (F32)	0.00	Gflop	47872.00	0.00
MFMA FLOPs (F64)	0.00	Gflop	47872.00	0.00
MFMA IOPs (Int8)		Giop	191488.00	
Active CUs		Cus	110.00	
SALU Utilization		Pct	100.00	
VALU Utilization		Pct	100.00	
MFMA Utilization		Pct	100.00	
VMEM Utilization		Pct	100.00	
Branch Utilization		Pct	100.00	
VALU Active Threads		Threads	64.00	
IPC		Instr/cycle	5.00	
Wavefront Occupancy		Wavefronts	3520.00	
Theoretical LDS Bandwidth	5290.29	Gb/s	23936.00	22.10
LDS Bank Conflicts/Access	0.00	Conflicts/access	32.00	0.00
vL1D Cache Hit Rate	43.70	Pct	100.00	43.70
vL1D Cache BW	4257.68	Gb/s	11968.00	35.58
L2 Cache Hit Rate		Pct	100.00	
L2 Cache BW		Gb/s	3481.60	
L2-Fabric Read BW	91.75	Gb/s	1638.40	5.60
L2-Fabric Write BW	139.46	Gb/s	1638.40	8.51
L2-Fabric Read Latency		Cycles		
L2-Fabric Write Latency		Cycles		
sL1D Cache Hit Rate		Pct	100.00	
sL1D Cache BW		Gb/s	6092.80	
L1I Hit Rate		Pct	100.00	
L1I BW		Gb/s	6092.80	
L1I Fetch Latency		Cycles		

SPH-EXA: Details of AVSwitches Function

0.1 Top Kernels

	Count	Sum(ns)	Mean(ns)	Median(ns)	Pct
::AVswitchesGpu<double, float, unsigned unsigned int, cstone::Box<double>, unsigned long, NsView<double, unsigned long>, double const*, double const*, float const*, float const*, float const*, float const*, float const*, float const*, float const*, float const*, float const*, float const*, float const*, double, float, float, float, float*, int*) [clone .kd]	1.00	2822758316.00	2822758316.00	2822758316.00	12.78



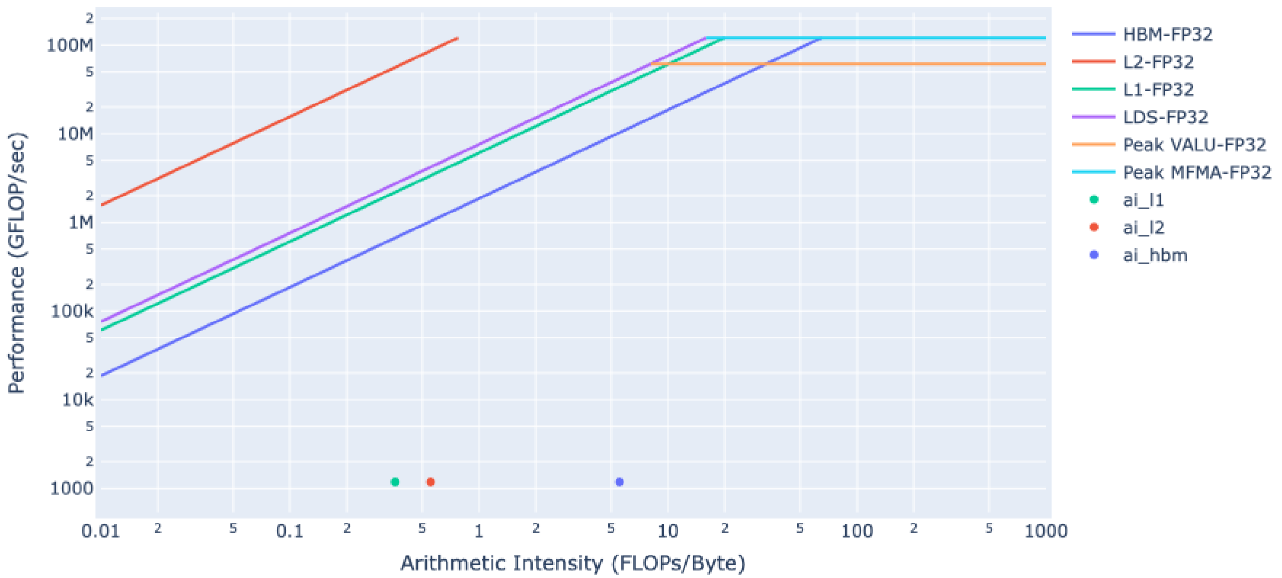
2.1 Speed-of-Light

Metric	Avg	Unit	Peak	Pct of Peak
VALU FLOPs	1107.69	Gflop	23936.00	4.63
VALU IOPs		Giop	23936.00	
MFMA FLOPs (BF16)	0.00	Gflop	191488.00	0.00
MFMA FLOPs (F16)	0.00	Gflop	191488.00	0.00
MFMA FLOPs (F32)	0.00	Gflop	47872.00	0.00
MFMA FLOPs (F64)	0.00	Gflop	47872.00	0.00
MFMA IOPs (Int8)		Giop	191488.00	
Active CUs		Cus	110.00	
SALU Utilization		Pct	100.00	
VALU Utilization		Pct	100.00	
MFMA Utilization		Pct	100.00	
VMEM Utilization		Pct	100.00	
Branch Utilization		Pct	100.00	
VALU Active Threads		Threads	64.00	
IPC		Instr/cycle	5.00	
Wavefront Occupancy		Wavefronts	3520.00	
Theoretical LDS Bandwidth	5470.21	Gb/s	23936.00	22.85
LDS Bank Conflicts/Access	0.00	Conflicts/access	32.00	0.00
vL1D Cache Hit Rate	47.09	Pct	100.00	47.09
vL1D Cache BW	3114.24	Gb/s	11968.00	26.02
L2 Cache Hit Rate		Pct	100.00	
L2 Cache BW		Gb/s	3481.60	
L2-Fabric Read BW	105.84	Gb/s	1638.40	6.46
L2-Fabric Write BW	138.21	Gb/s	1638.40	8.44
L2-Fabric Read Latency		Cycles		
L2-Fabric Write Latency		Cycles		
sL1D Cache Hit Rate		Pct	100.00	
sL1D Cache BW		Gb/s	6092.80	
L1I Hit Rate		Pct	100.00	
L1I BW		Gb/s	6092.80	
L1I Fetch Latency		Cycles		

SPH-EXA: Details of veDefGradh Function

0.1 Top Kernels

	Count	Sum(ns)	Mean(ns)	Median(ns)	Pct
::veDefGradhGpu<double, float, float, (double, unsigned int, uble>, unsigned int const*, unsigned ctreeNsView<double, unsigned long>, double const*, double const*, float const*, float const*, float const*, float*, float*, unsigned int*, int*)	1.00	2379395042.00	2379395042.00	2379395042.00	10.77



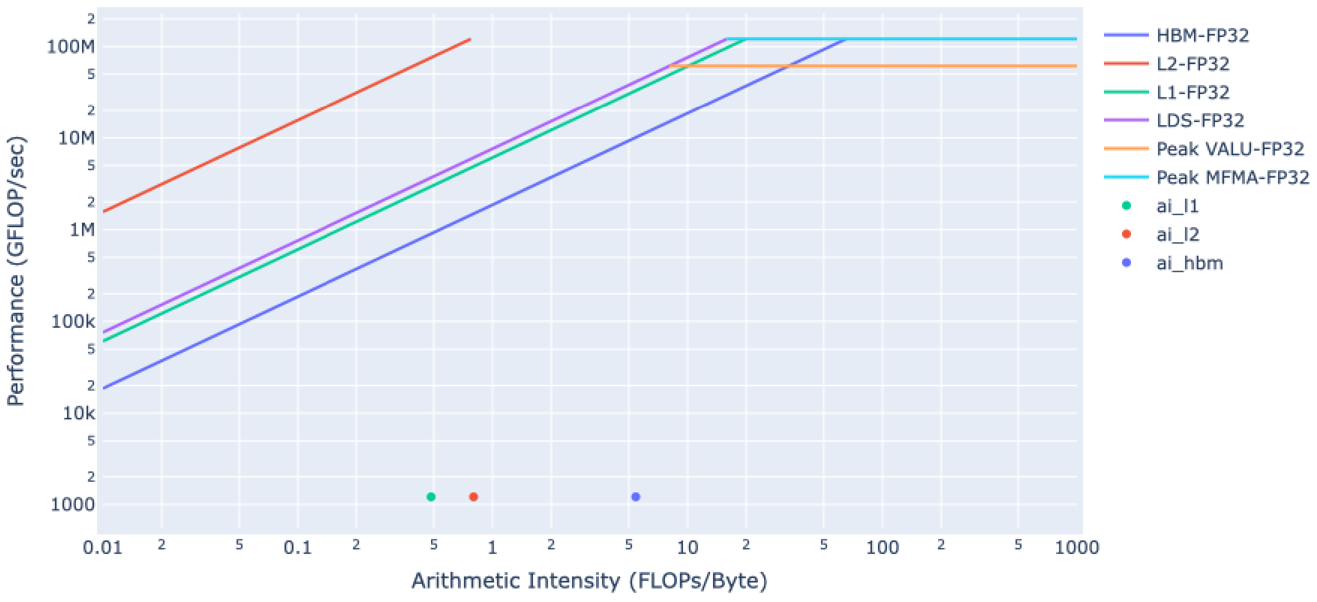
2.1 Speed-of-Light

Metric	Avg	Unit	Peak	Pct of Peak
VALU FLOPs	1178.38	Gflop	23936.00	4.92
VALU IOPs		Giop	23936.00	
MFMA FLOPs (BF16)	0.00	Gflop	191488.00	0.00
MFMA FLOPs (F16)	0.00	Gflop	191488.00	0.00
MFMA FLOPs (F32)	0.00	Gflop	47872.00	0.00
MFMA FLOPs (F64)	0.00	Gflop	47872.00	0.00
MFMA IOPs (Int8)		Giop	191488.00	
Active CUs		Cus	110.00	
SALU Utilization		Pct	100.00	
VALU Utilization		Pct	100.00	
MFMA Utilization		Pct	100.00	
VMEM Utilization		Pct	100.00	
Branch Utilization		Pct	100.00	
VALU Active Threads		Threads	64.00	
IPC		Instr/cycle	5.00	
Wavefront Occupancy		Wavefronts	3520.00	
Theoretical LDS Bandwidth	6489.51	Gb/s	23936.00	27.11
LDS Bank Conflicts/Access	0.00	Conflicts/access	32.00	0.00
vL1D Cache Hit Rate	35.10	Pct	100.00	35.10
vL1D Cache BW	3276.66	Gb/s	11968.00	27.38
L2 Cache Hit Rate		Pct	100.00	
L2 Cache BW		Gb/s	3481.60	
L2-Fabric Read BW	67.34	Gb/s	1638.40	4.11
L2-Fabric Write BW	145.28	Gb/s	1638.40	8.87
L2-Fabric Read Latency		Cycles		
L2-Fabric Write Latency		Cycles		
sL1D Cache Hit Rate		Pct	100.00	
sL1D Cache BW		Gb/s	6092.80	
L1I Hit Rate		Pct	100.00	
L1I BW		Gb/s	6092.80	
L1I Fetch Latency		Cycles		

SPH-EXA: Details of xMass Function

0.1 Top Kernels

	Count	Sum(ns)	Mean(ns)	Median(ns)	Pct
::xmassGpu<double, float, float, (double, unsigned int, unsigned int, uble>, unsigned int const*, unsigned ctreeNsView<double, unsigned long>, double const*, double const*, double , float const*, float const*, float , unsigned int*, int*) [clone .kd]	1.00	2215289298.00	2215289298.00	2215289298.00	10.03



2.1 Speed-of-Light

Metric	Avg	Unit	Peak	Pct of Peak
VALU FLOPs	1210.36	Gflop	23936.00	5.06
VALU IOPs		Giop	23936.00	
MFMA FLOPs (BF16)	0.00	Gflop	191488.00	0.00
MFMA FLOPs (F16)	0.00	Gflop	191488.00	0.00
MFMA FLOPs (F32)	0.00	Gflop	47872.00	0.00
MFMA FLOPs (F64)	0.00	Gflop	47872.00	0.00
MFMA IOPs (Int8)		Giop	191488.00	
Active CUs		Cus	110.00	
SALU Utilization		Pct	100.00	
VALU Utilization		Pct	100.00	
MFMA Utilization		Pct	100.00	
VMEM Utilization		Pct	100.00	
Branch Utilization		Pct	100.00	
VALU Active Threads		Threads	64.00	
IPC		Instr/cycle	5.00	
Wavefront Occupancy		Wavefronts	3520.00	
Theoretical LDS Bandwidth	6970.24	Gb/s	23936.00	29.12
LDS Bank Conflicts/Access	0.00	Conflicts/access	32.00	0.00
vL1D Cache Hit Rate	39.46	Pct	100.00	39.46
vL1D Cache BW	2497.59	Gb/s	11968.00	20.87
L2 Cache Hit Rate		Pct	100.00	
L2 Cache BW		Gb/s	3481.60	
L2-Fabric Read BW	66.49	Gb/s	1638.40	4.06
L2-Fabric Write BW	156.36	Gb/s	1638.40	9.54
L2-Fabric Read Latency		Cycles		
L2-Fabric Write Latency		Cycles		
sL1D Cache Hit Rate		Pct	100.00	
sL1D Cache BW		Gb/s	6092.80	
L1I Hit Rate		Pct	100.00	
L1I BW		Gb/s	6092.80	
L1I Fetch Latency		Cycles		

TGSF: The role of Turbulence and Gravity in Star Formation

Extreme Scale Access

Allocation: **22,000,000 GPUh*** on LUMI-G
Duration: 12 months, Nov.'23 – Oct.'24
*Largest allocation in Europe to date.

Objectives

Cosmology & Astrophysics

- Study the formation of pre-stellar cores and their initial mass function at unprecedented resolution — Scalability limitation for previous codes
- Study turbulent transport and mixing
- Contribute to the general theory of turbulence (Lyapunov exponents)

(Study turbulent transport and mixing; Contribute to the general theory of turbulence: More natural with Lagrangian codes)

Computer Science

- Study the load imbalance, performance, and energy consumption at unprecedented scales
- Study large scale compression techniques for checkpointing, compression, and visualization

(HPC research at extreme scale)

Publications | News
PASC project principle investigators discussing their new astrophysical simulation code, which helped them win a large allocation on LUMI-G
https://bit.ly/cscs-sph-exa2