Karlsruhe Institute of Technology

EXASCALE COMPUTING PROJECT

EuroHPC Joint Undertaking

NATIONALES HOCHLEISTUNGS RECHNEN

HELMHOLTZ RESEARCH FOR GRAND CHALLENGES

Joint Workshop: MathSEE-KCETA

Numerical methods for High Performance Computing

Pratik Nayak

Karlsruhe Institute of Technology

EXASCALE COMPUTING PROJECT

EuroHPC Joint Undertaking

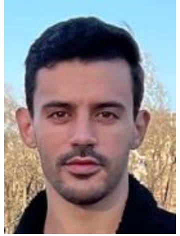
NATIONALES HOCHLEISTUNGS RECHNEN

HELMHOLTZ RESEARCH FOR GRAND CHALLENGES

Fixed Point Numerics for Exascale (FiNE)

Hartwig Anzt

Isha Aggarwal



Rached Chaben

Terry Cojean

Vasileios Georgiou

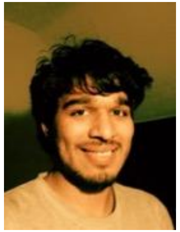
Fritz Göbel

Thomas Grützmacher

Claudius Holeksa

Marcel Koch

Stefano Maurogiovanni



Pratik Nayak

Gregor Olenik

Tobias Ribizel

Yu-Hsiang Tsai

Outline

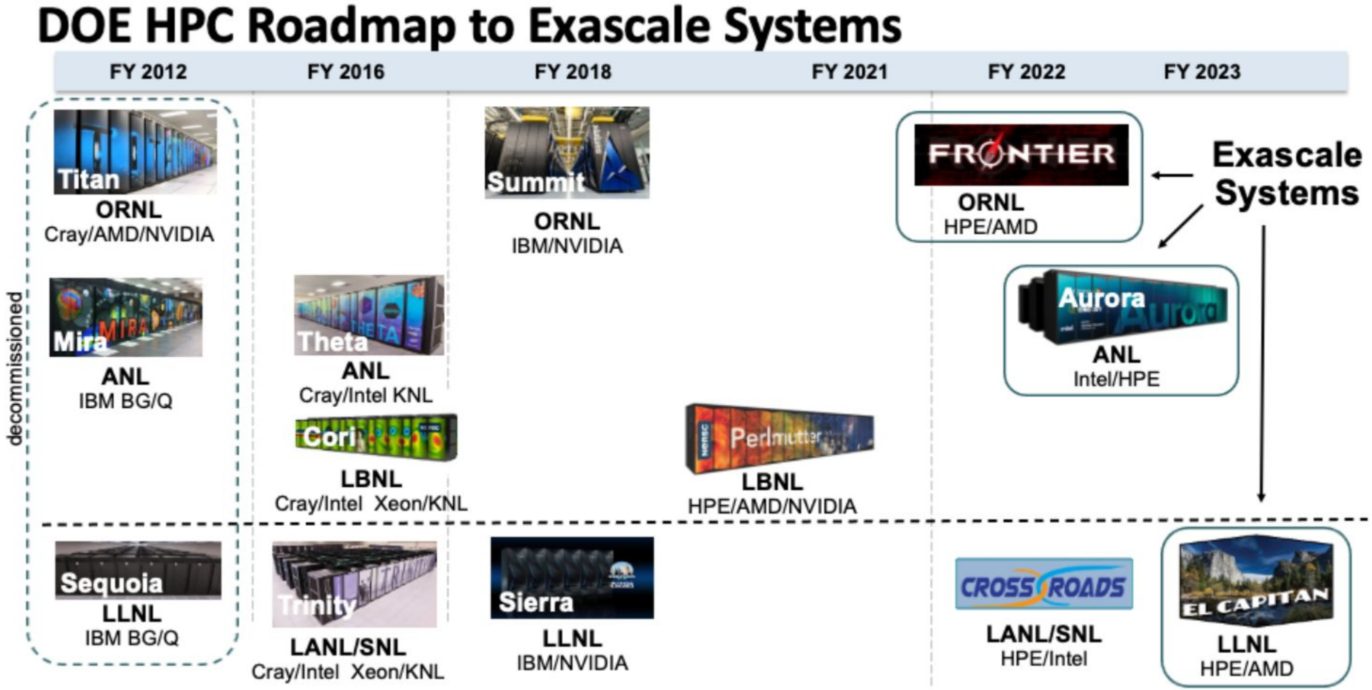
- The HPC landscape
- Ginkgo - A high performance numerical linear algebra library.
- Application 1: MFEM
- Application 2: OpenFOAM
- Application 3: Combustion simulations with PeleLM
- Application 4: Fusion plasma simulations with XGC.
- Conclusions.

The High Performance Computing (HPC) Landscape

- A first exascale system, Frontier (1.1 EFlops).
- European system in top 3, LUMI with 151 PFlops.
- All systems aiming for efficiency use some form of accelerators.

Rank	System	Cores	Rmax (PFlop/s)	Rpeak (PFlop/s)	Power (kW)
1	**Frontier** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE DOE/SC/Oak Ridge National Laboratory United States	8,730,112	1,102.00	1,685.65	21,100
2	**Supercomputer Fugaku** - Supercomputer Fugaku, A64FX 48C 2.2GHz, Tofu interconnect D, Fujitsu RIKEN Center for Computational Science Japan	7,630,848	442.01	537.21	29,899
3	**LUMI** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE EuroHPC/CSC Finland	1,110,144	151.90	214.35	2,942
4	**Summit** - IBM Power System AC922, IBM POWER9 22C 3.07GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband, IBM DOE/SC/Oak Ridge National Laboratory United States	2,414,592	148.60	200.79	10,096
5	**Sierra** - IBM Power System AC922, IBM POWER9 22C 3.1GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband, IBM / NVIDIA / Mellanox DOE/NNSA/LLNL United States	1,572,480	94.64	125.71	7,438

The High Performance Computing (HPC) Landscape



Dongarra et.al May 2022 Tech report ICL-UT-22-05

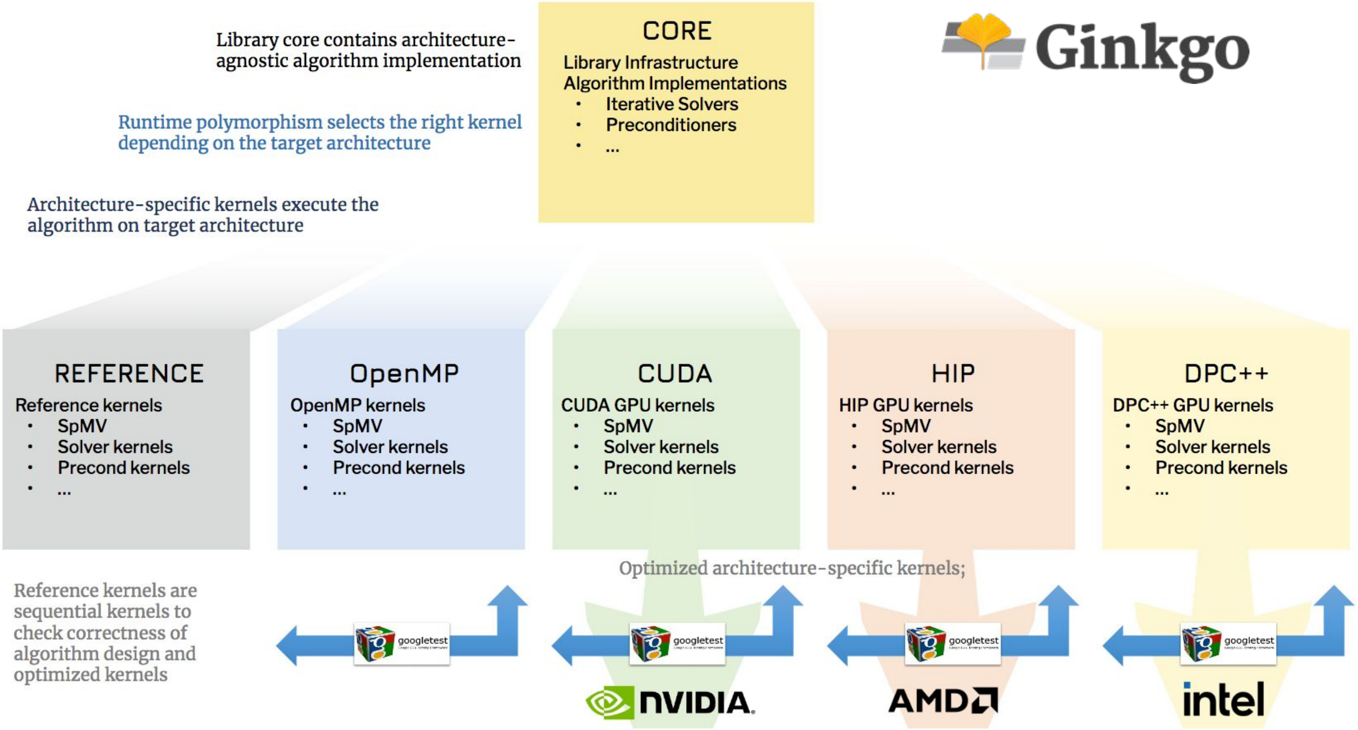
What is the performance distribution ? (GPUs v/s CPUs)

- Consider 1 node of Summit (6 GPUs and 2 sockets of IBM Power 9 with a total of 44 cores)
- Peak Flop of 6 V100 GPU is 40 TFlops. Peak Flop of 2 sockets of IBM Power 9 CPUs is 0.9 TFlops.
- **98%** of the performance on a node is in the GPUs. **Extremely critical** to design implementations and algorithms that work well on GPUs and multi-GPUs.

Dongarra et.al May 2022 Tech report ICL-UT-22-05

The Ginkgo software library.

- Focus on high performance sparse linear algebra.
- Thoroughly tested and benchmarked.
- Linear solvers, matrix formats, preconditioners and more.
- Support for multiple backends: CUDA, HIP, DPC++ and OpenMP.



Ginkgo: Current features

	Functionality	OMP	CUDA	HIP	DPC++
Basic	SpMV	✓	✓	✓	✓
	SpMM	✓	✓	✓	✓
	SpGeMM	✓	✓	✓	✓
Krylov solvers	BiCG	✓	✓	✓	✓
	BiCGSTAB	✓	✓	✓	✓
	CG	✓	✓	✓	✓
	CGS	✓	✓	✓	✓
	GMRES	✓	✓	✓	✓
	IDR	✓	✓	✓	✓
Preconditioners	(Block-)Jacobi	✓	✓	✓	
	ILU/IC		✓	✓	
	Parallel ILU/IC	✓	✓	✓	
	Parallel ILUT/ICT	✓	✓	✓	
	Sparse Approximate Inverse	✓	✓	✓	
	Algebraic Multigrid	✓	✓	✓	
Batched	Batched BiCGSTAB	✓	✓	✓	
	Batched CG	✓	✓	✓	
	Batched GMRES		✓	✓	

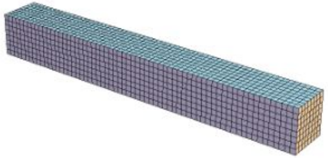
Local parallelization strategy | **Local algorithmic approach**

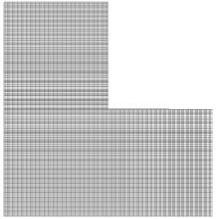
- SpMV, SpMM: Thread-to-nonzero, subwarp-to-row, or thread-to-row mapping — Mixed precision support
- SpGeMM: subwarp-to-row mapping
- Krylov solvers: Thread-to-row mapping — Merged kernels
- (Block-)Jacobi: subwarp to blocks — Advanced Mixed precision support
- ILU/IC: Subwarp-to-row mapping
- Parallel ILU/IC, Parallel ILUT/ICT: Thread-to-nonzero mapping — Asynchronous Fix-point based
- Sparse Approximate Inverse: Subwarp-to-row mapping
- Algebraic Multigrid: Composes of other routines
- Batched: Thread-block to problem mapping — kernel-inlining & single-kernel design to avoid global memory accesses

Single GPU applications

Application 1: MFEM: Poisson equation

$$-\Delta u = 1$$

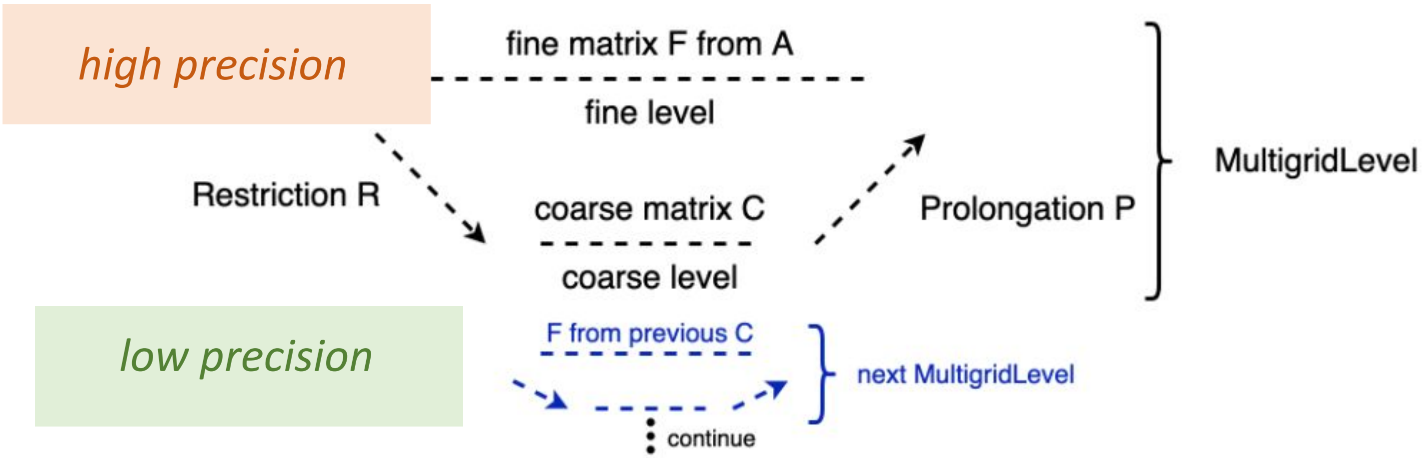




	problem	size	nonzero elements
matrices in MFEM integration test	beam (-o2 -l3)	37,281	21,67,425
	beam (-o3 -l3)*	120,625	14,070,001
	beam (-o4 -l3)	279,873	57,251,713
	beam (-o3 -l4)	924,385	111,573,601
	L-shape (-o3 -l7)*	443,905	11,066,881
	L-shape (-o3 -l8)	1,772,545	44,252,161
	L-shape (-o4 -l7)	788,481	28,323,841
	L-shape (-o4 -l8)	3,149,825	113,270,785

Karlsruhe Institute of Technology
EXASCALE COMPUTING PROJECT
EuroHPC Joint Undertaking
NATIONALES HOCHLEISTUNGS RECHNEN
HELMHOLTZ RESEARCH FOR GRAND CHALLENGES

Application 1: MFEM, Multigrid and mixed precision

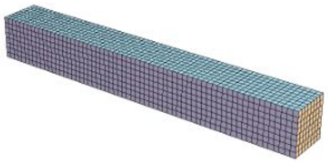


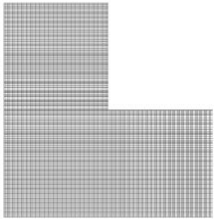
Karlsruhe Institute of Technology · EXASCALE COMPUTING PROJECT · EuroHPC Joint Undertaking · NATIONALES HOCHLEISTUNGSRECHNEN · HELMHOLTZ RESEARCH FOR GRAND CHALLENGES

Application 1: MFEM, Multigrid and mixed precision

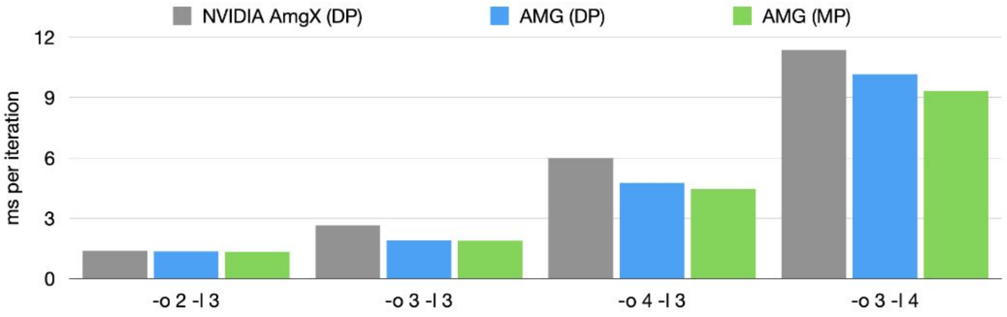
MFEM

Ginkgo

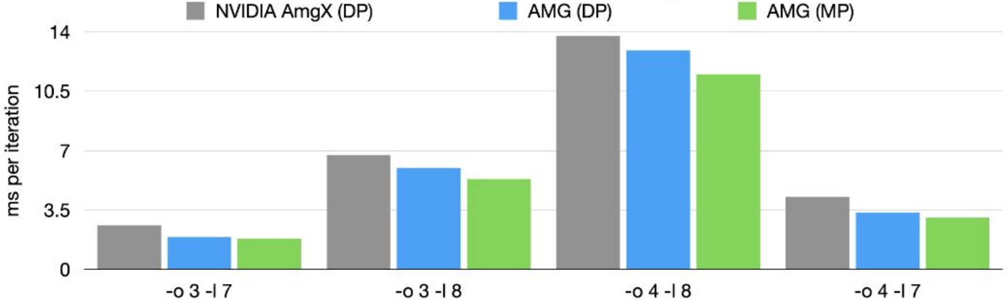




geometry	problem	NVIDIA AmgX (DP) runtime [ms]	NVIDIA AmgX (DP) #iter	Ginkgo AMG (DP) runtime [ms]	Ginkgo AMG (DP) #iter	Ginkgo AMG (MP) runtime [ms]	Ginkgo AMG (MP) #iter
Beam	-o 2 -l 3	20.71	15	20.27	15	19.96	15
	-o 3 -l 3	52.94	20	39.93	21	39.56	21
	-o 4 -l 3	155.47	26	128.69	27	120.41	27
	-o 3 -l 4	329.68	29	294.68	29	270.39	29
L-shape	-o 3 -l 7	242.27	93	178.02	93	170.08	94
	-o 3 -l 8	1211.38	180	1033.96	173	943.27	177
	-o 4 -l 8	3452.91	251	3044.24	236	2722.63	237
	-o 4 -l 7	551.99	129	407.27	122	366.99	120



MFEM beam with different setting



MFEM L-shape with different setting

Karlsruhe Institute of Technology

EXASCALE COMPUTING PROJECT

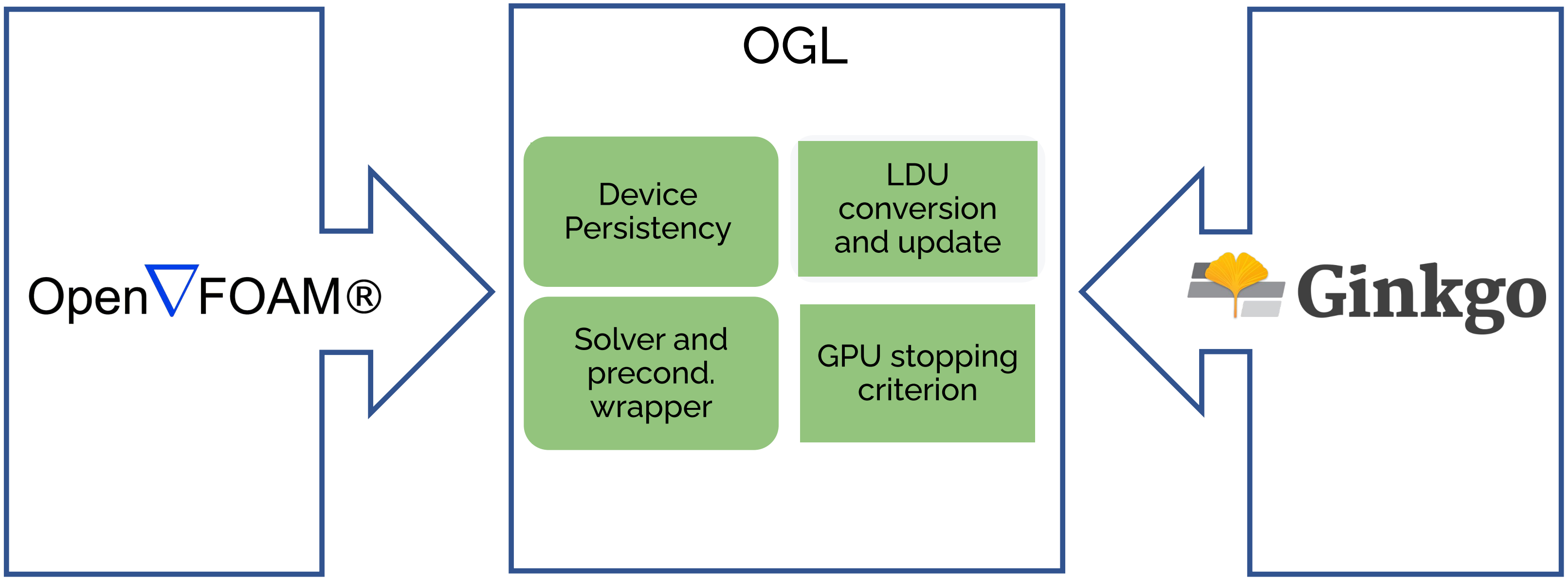
EuroHPC Joint Undertaking

NATIONALES HOCHLEISTUNGS RECHNEN

HELMHOLTZ RESEARCH FOR GRAND CHALLENGES

Application 2: OpenFOAM

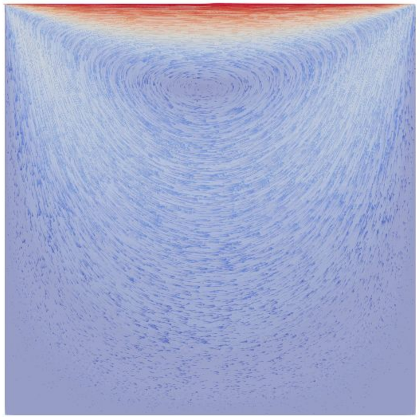
OpenFOAM® + Ginkgo = OGL[1]



[1]Olenik,G: OGL: OpenFOAM Ginkgo layer: https://github.com/hpsim/OGL

Application 2.1 : OpenFOAM (Lid driven cavity)

OpenFOAM® + Ginkgo = OGL[1]

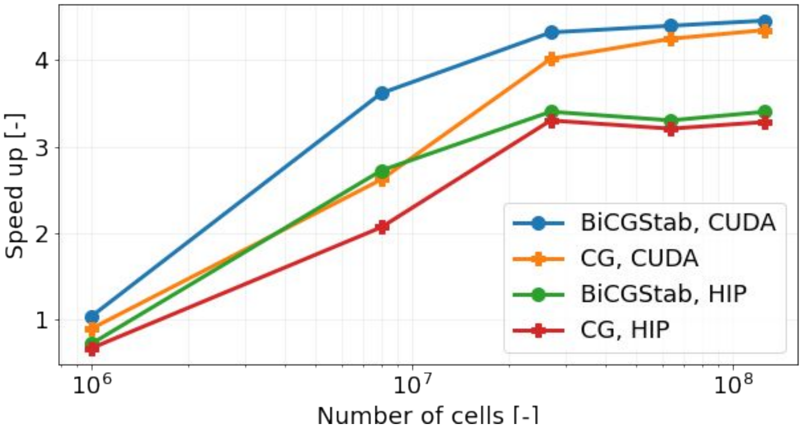


DoFs	1e6 to 25e6
Application	icoFoam
Solver, p	CG, DIC
Solver, U	BiCGStab, DILU

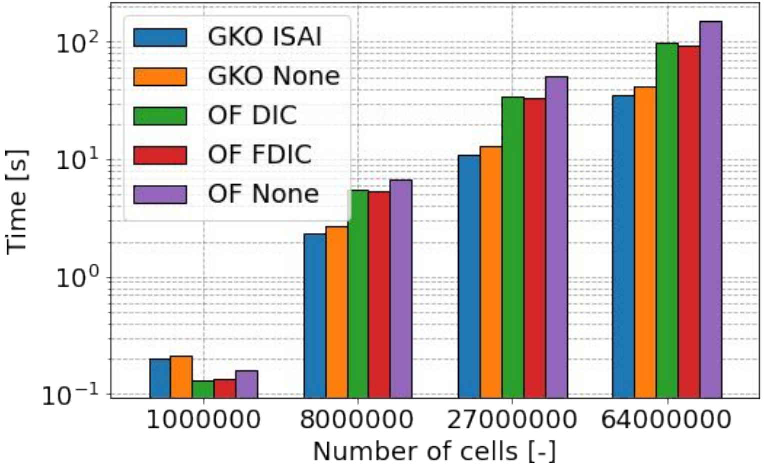
[1]Olenik,G: OGL: OpenFOAM Ginkgo layer: https://github.com/hpsim/OGL

Application 2.1 : OpenFOAM (Lid driven cavity)

OpenFOAM® + Ginkgo = OGL[1]



Single GPU speedup (HIP (MI100) and CUDA (V100)) over CPUs, AMD (32 cores and Intel (76 cores) resp.



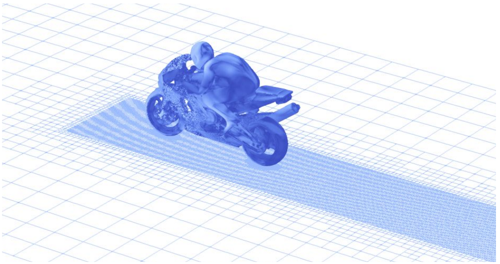
Poisson solver times with different preconditioners (AMD MI100 and AMD CPU, 32 ranks)

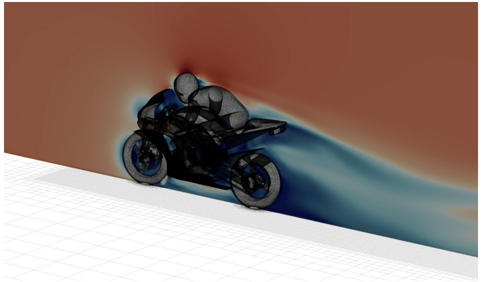
[1]Olenik,G: OGL: OpenFOAM Ginkgo layer: https://github.com/hpsim/OGL

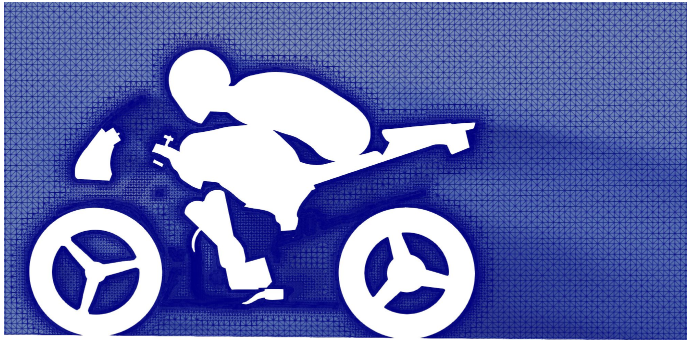
Multi-GPU and distributed

Application 2.2 : OpenFOAM (Motorcycle)

OpenFOAM® + Ginkgo = OGL[1]







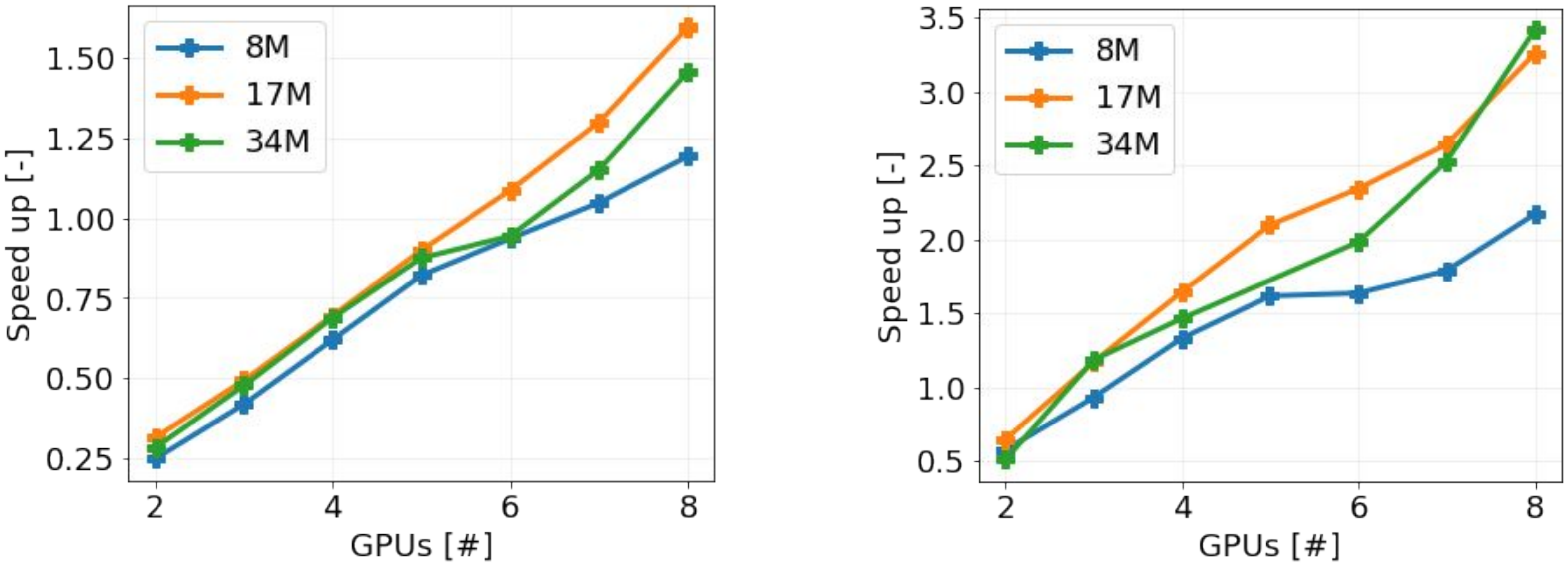
DoF	n*8.678M
Application	simpleFoam
Outer iterations	500
Solver, p	CG
Solver, U	BiCGStab

Case	DoFs
Small	8.678M
Medium	17.356M
Large	26.034M

[1]Olenik,G: OGL: OpenFOAM Ginkgo layer: https://github.com/hpsim/OGL

Application 2.2 : OpenFOAM (Motorcycle)

OpenFOAM® + Ginkgo = OGL[1]



CG speedup with distributed Schwarz preconditioner with ISAI (left), and Multigrid (right) on MI100s versus 32 MPI ranks on an AMD EPYC 7302 with an IC preconditioner.

[1]Olenik,G: OGL: OpenFOAM Ginkgo layer: https://github.com/hpsim/OGL

Karlsruhe Institute of Technology

EXASCALE COMPUTING PROJECT

EuroHPC Joint Undertaking

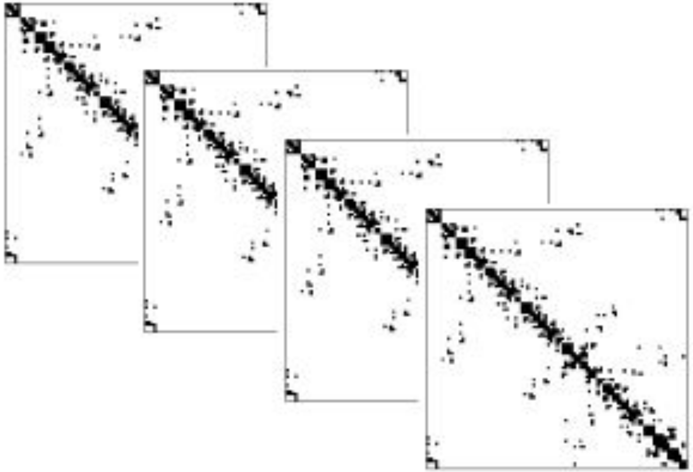
NATIONALES HOCHLEISTUNGS RECHNEN

HELMHOLTZ RESEARCH FOR GRAND CHALLENGES

Batched methods

Batched solvers in Ginkgo

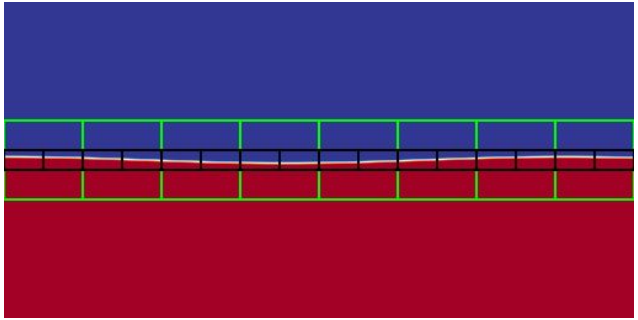
- Batching: Independent computations that can be scheduled in parallel.
- Are highly suitable for GPUs and processors with many parallel computing units.
- Can maximize utilization of the GPU, due to excellent scalability.

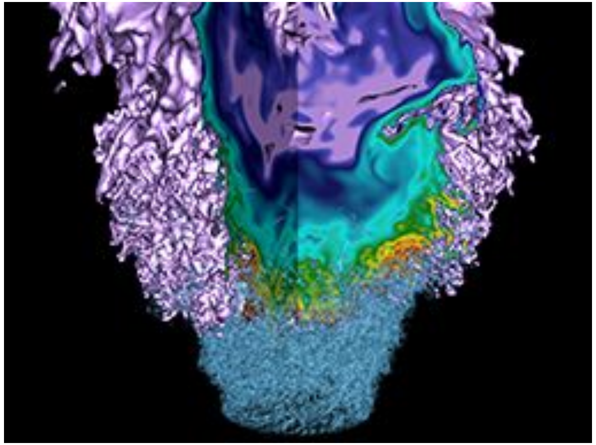


Karlsruhe Institute of Technology | EXASCALE COMPUTING PROJECT | EuroHPC Joint Undertaking | NATIONALES HOCHLEISTUNGSRECHNEN | HELMHOLTZ RESEARCH FOR GRAND CHALLENGES

Application 3: Combustion simulations

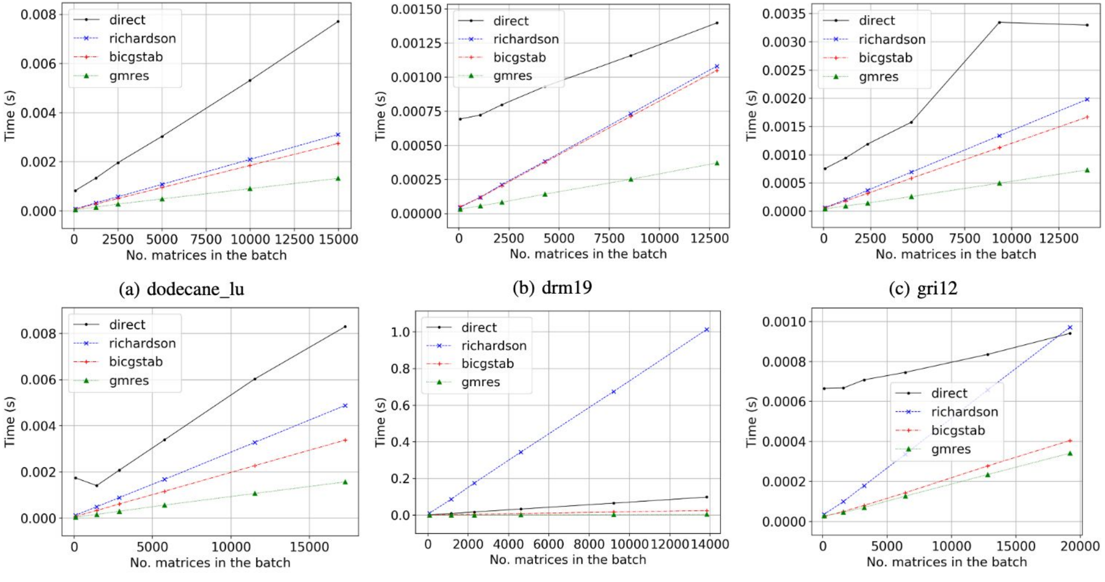
- PeleLM is a parallel, adaptive mesh refinement (AMR) code that solves the Navier-Stokes equations with in the low Mach number regime with the chemical reaction mechanisms.
- https://amrex-combustion.github.io/PeleLM/overview.html



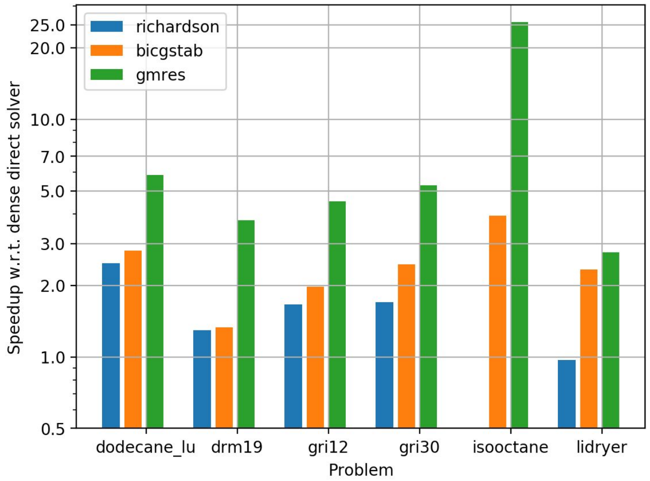


Problem	Size	Non-zeros (A)	Non-zeros (L+U)
dodecane_lu	54	2,332 (80%)	2,754 (94%)
drm19	22	438 (90%)	442 (91%)
gri12	33	978 (90%)	1,018 (93%)
gri30	54	2,560 (88%)	2,860 (98%)
isooctane	144	6,135 (30%)	20,307 (98%)
lidryer	10	91 (91%)	91 (91%)

Application 3: Combustion simulations



(a) dodecane_lu (b) drm19 (c) gri12
(d) gri30 (e) isooctane (f) lidryer



Speedups in production code may be larger:

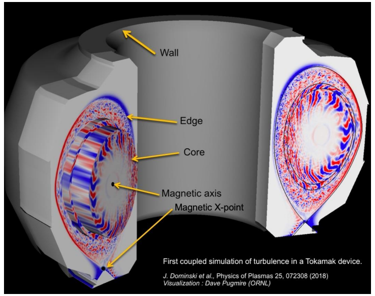
- Use of "suitable" initial guess (last linear solve);
- Less accurate solution often sufficient;

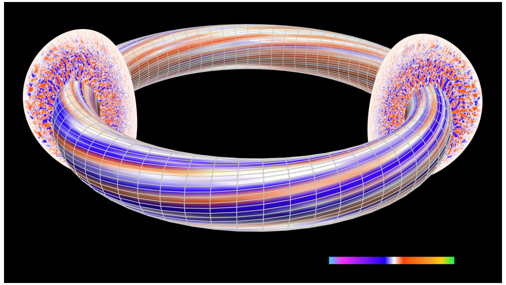
Batched Sparse Iterative Solvers for Computational Chemistry Simulations on GPUs, ScalA 2021, SC21, Aggarwal, Kashi, Nayak, Balos, Woodward and Anzt.

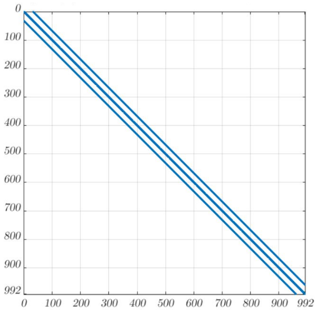
Application 4: Fusion plasma simulations

XGC is a gyrokinetic particle-in-cell code, which specializes in the simulation of the edge region of magnetically confined thermonuclear fusion plasma. The simulation domain can include the magnetic separatrix, magnetic axis and the biased material wall. XGC can run in total-delta-f, and conventional delta-f mode. The ion species are always gyrokinetic except for ETG simulation. Electrons can be adiabatic, massless fluid, drift-kinetic, or gyrokinetic.

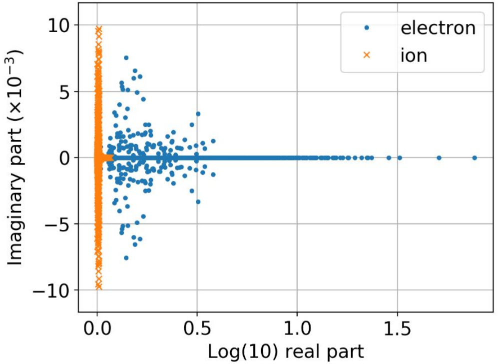
Source: https://xgc.pppl.gov/html/general_info.html



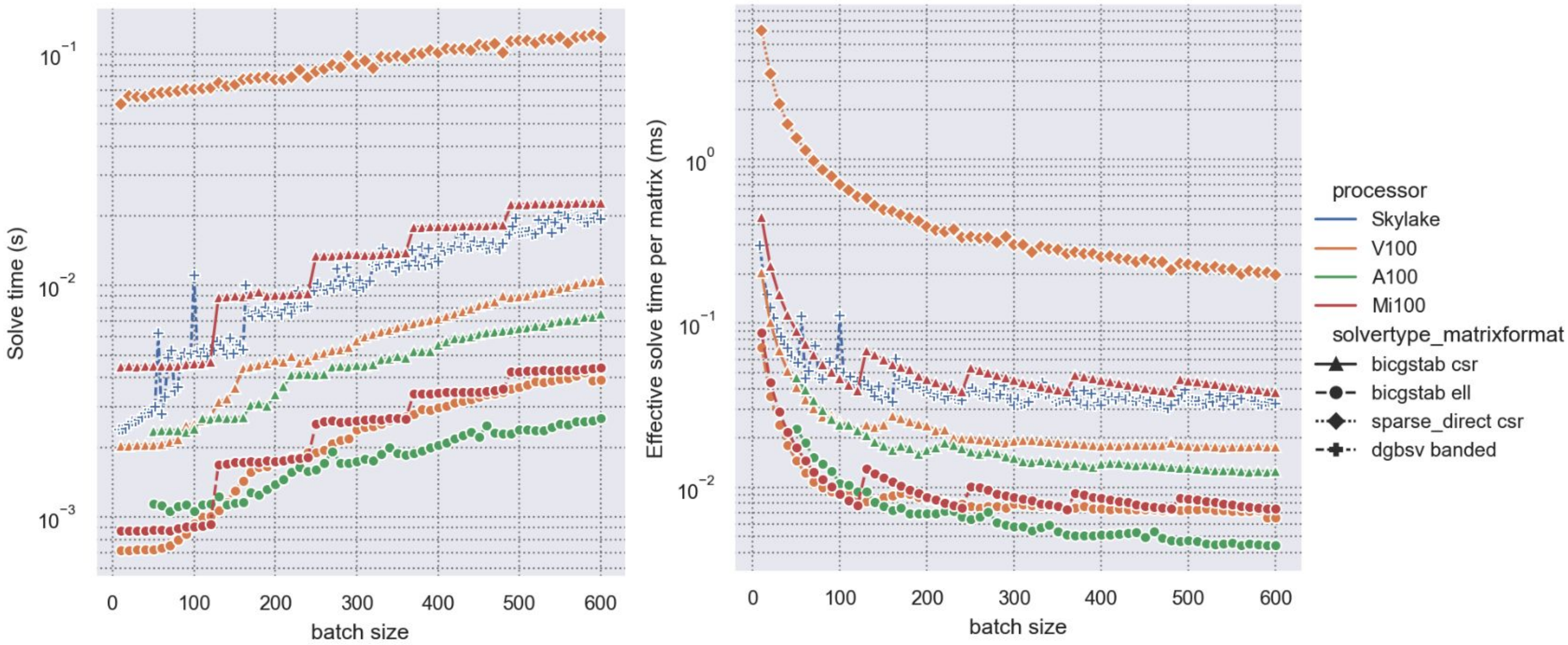




- Two species
- Ions easy to solve
- Electrons hard to solve
- Banded matrix structure
- Non-symmetric, need BiCGSTAB
- n = ~1,000
- nz = ~9,000



Application 4: Fusion plasma simulations



Batched Iterative solvers on GPUs for Fusion Plasma simulations, IPDPS 2022
Kashi, Nayak, Kulkarni, Lin and Anzt.

Outlook

- Efficient usage of GPUs is crucial in the path to Exascale.
- Ginkgo provides high performance implementations of many algorithms and we have shown good performance for a variety of applications.
- Distributed functionality in Ginkgo is bleeding edge, but has shown good scaling and performance.
- More mixed-precision functionality is being investigated (SpMVs, solvers and preconditioners)
- A fully GPU enabled sparse direct solver will also be available soon in Ginkgo.

Thank you!

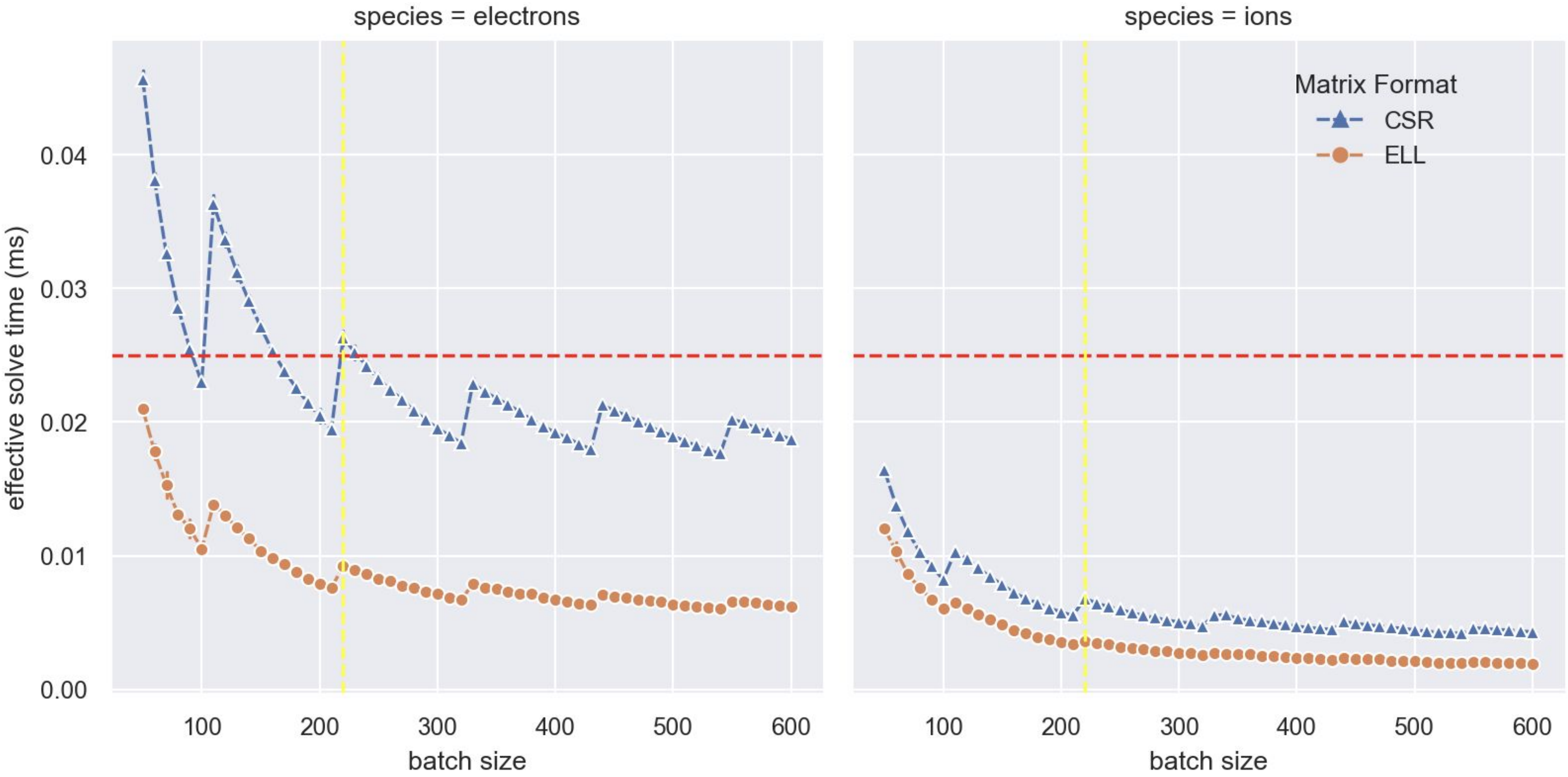
pratik.nayak@kit.edu

Ginkgo

https://github.com/ginkgo-project/ginkgo

Bonus

Application 4: Fusion plasma simulations



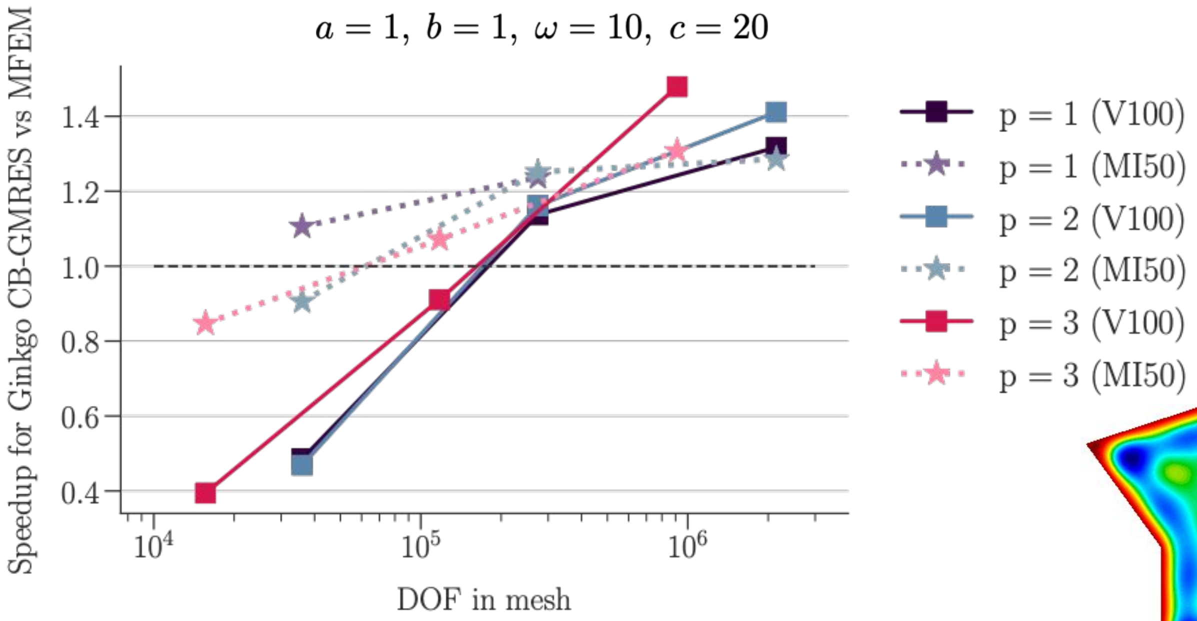
Application 1: Finite element library interfaces.

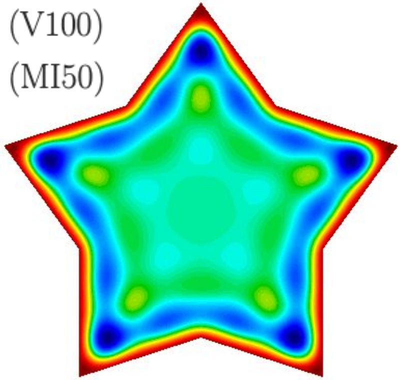
MFEM
Finite
element
library

Example: Speeding up MFEM's "example 22" (damped harmonic oscillator) on NVIDIA and AMD GPUs

$$-\nabla \cdot (a\nabla u) - \omega^2 bu + i\omega cu = 0$$

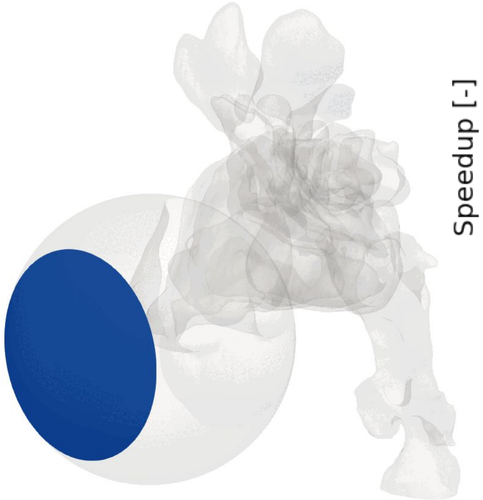
$$a = 1,\ b = 1,\ \omega = 10,\ c = 20$$



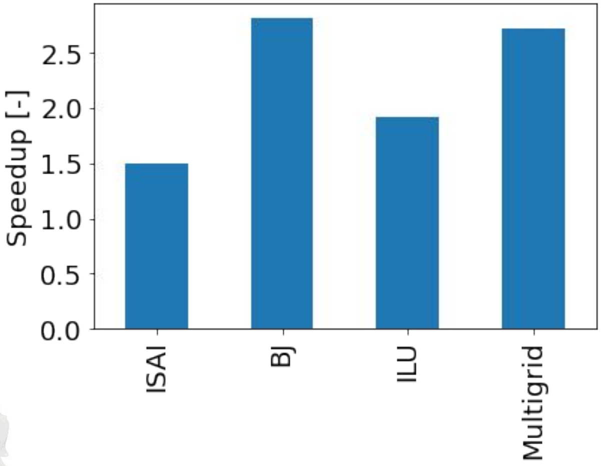


Application 2.3 : OpenFOAM (Nasal cavity)

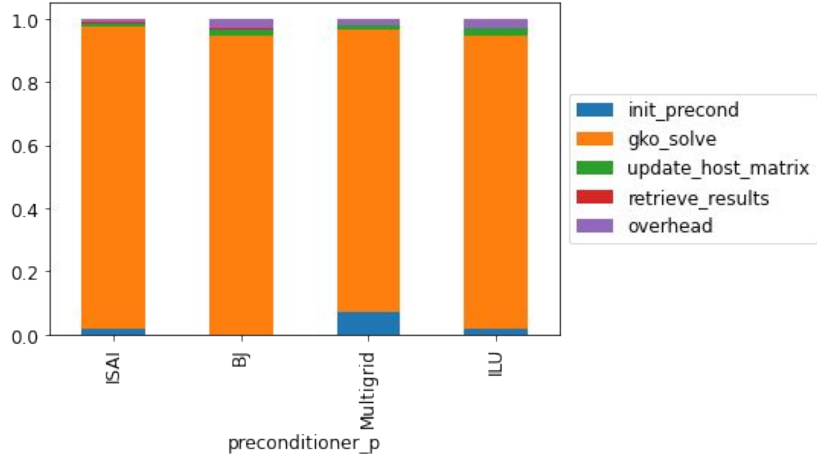
OpenFOAM® + Ginkgo = OGL[1]



Airflow inside a human nose



Total application speedup on one node with 4 GPUs



Solver cost breakdown

[1]Olenik,G: OGL: OpenFOAM Ginkgo layer: https://github.com/hpsim/OGL

Karlsruhe Institute of Technology

EXASCALE COMPUTING PROJECT

EuroHPC Joint Undertaking

NATIONALES HOCHLEISTUNGS RECHNEN

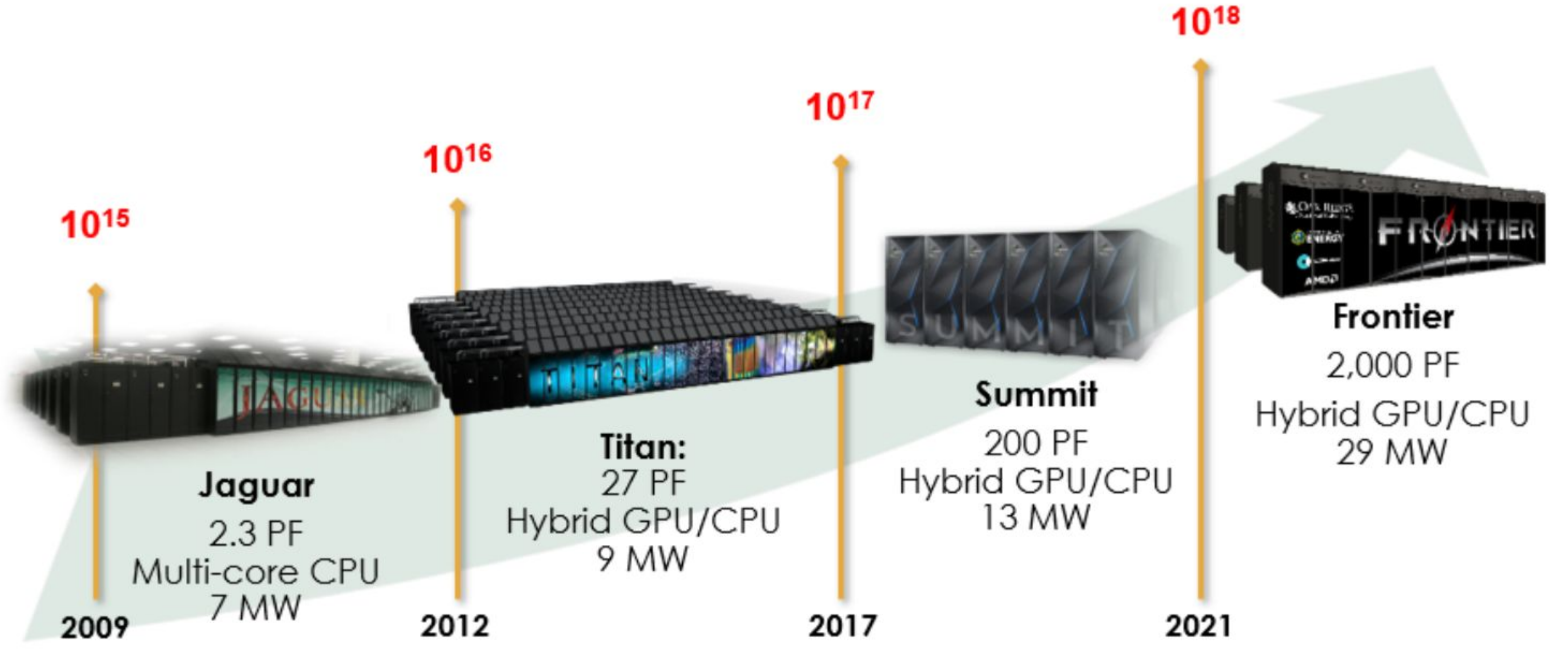
HELMHOLTZ RESEARCH FOR GRAND CHALLENGES

The High Performance Computing (HPC) Landscape

System	Titan (2012)	Summit (2017)	Frontier (2021)
Peak	27 PF	200 PF	2 EF
# nodes	18,688	4,608	9,408
Node	1 AMD Opteron CPU 1 NVIDIA Kepler GPU	2 IBM POWER9™ CPUs 6 NVIDIA Volta GPUs	1 AMD EPYC CPU 4 AMD Radeon Instinct GPUs
Memory	0.6 PB DDR3 + 0.1 PB GDDR	2.4 PB DDR4 + 0.4 HBM + 7.4 PB On-node storage	4.6 PB DDR4 + 4.6 PB HBM2e + 36 PB On-node storage, 66 TB/s Read 62 TB/s Write
On-node interconnect	PCI Gen2 No coherence across the node	NVIDIA NVLINK Coherent memory across the node	AMD Infinity Fabric Coherent memory across the node
System Interconnect	Cray Gemini network 6.4 GB/s	Mellanox Dual-port EDR IB 25 GB/s	Four-port Slingshot network 100 GB/s
Topology	3D Torus	Non-blocking Fat Tree	Dragonfly
Storage	32 PB, 1 TB/s, Lustre Filesystem	250 PB, 2.5 TB/s, IBM Spectrum Scale™ with GPFS™	695 PB HDD+11 PB Flash Performance Tier, 9.4 TB/s and 10 PB Metadata Flash. Lustre
Power	9 MW	13 MW	29 MW

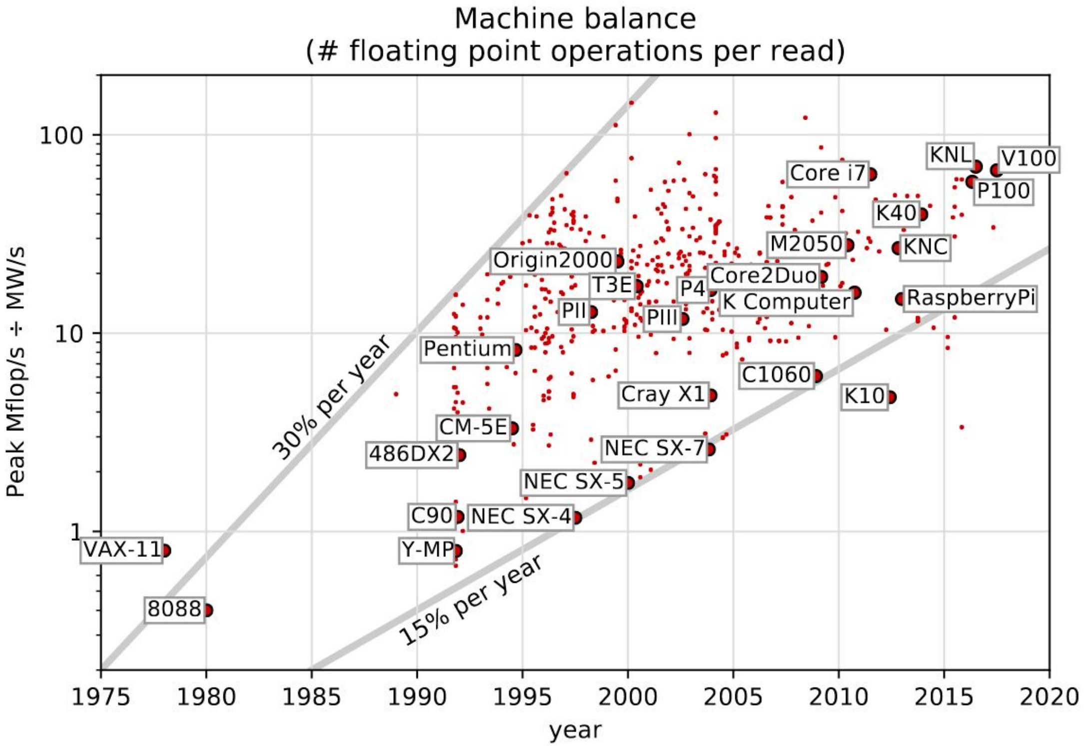
Dongarra et.al May 2022 Tech report ICL-UT-22-05

The High Performance Computing (HPC) Landscape



Dongarra et.al May 2022 Tech report ICL-UT-22-05

Read and writes significantly more expensive !



Abdelfattah et.al Sep 2021 Tech report LLNL-JRNL-826451