GLASs: a GPU-accelerated Linear Algebra framework for exascale problems in physics and engineering

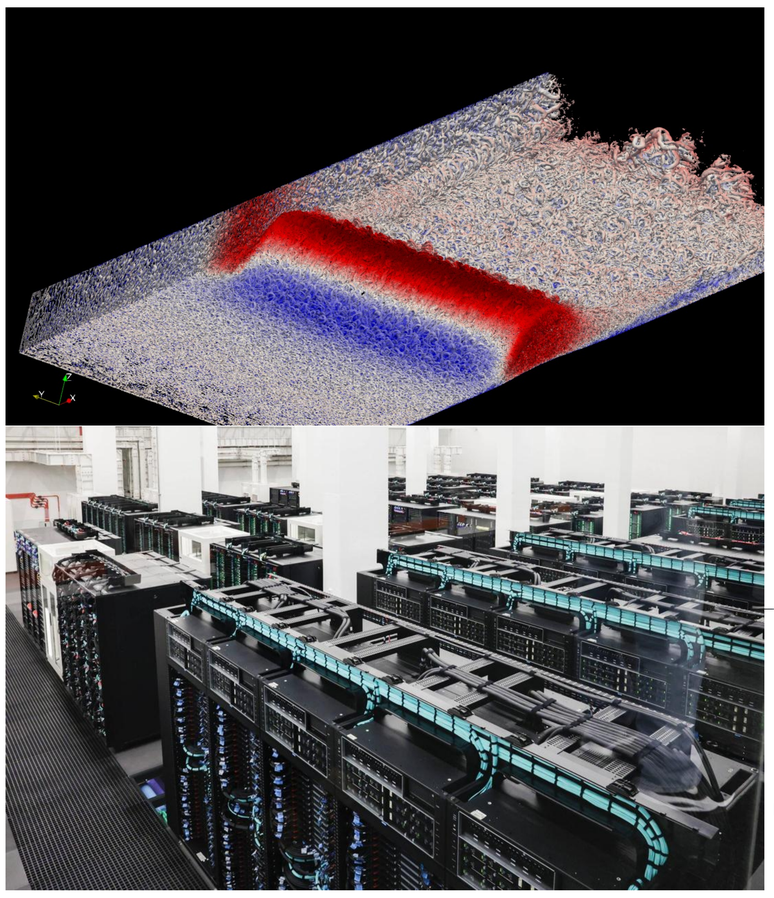
L. Gasparino[a], A. Quintanas[a], C. McDonnell[b], F. Spiga[c], O. Lehmkuhl[a]

[a]BSC-CNS (CASE), [b]UPM, [c]NVIDIA Corp.

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

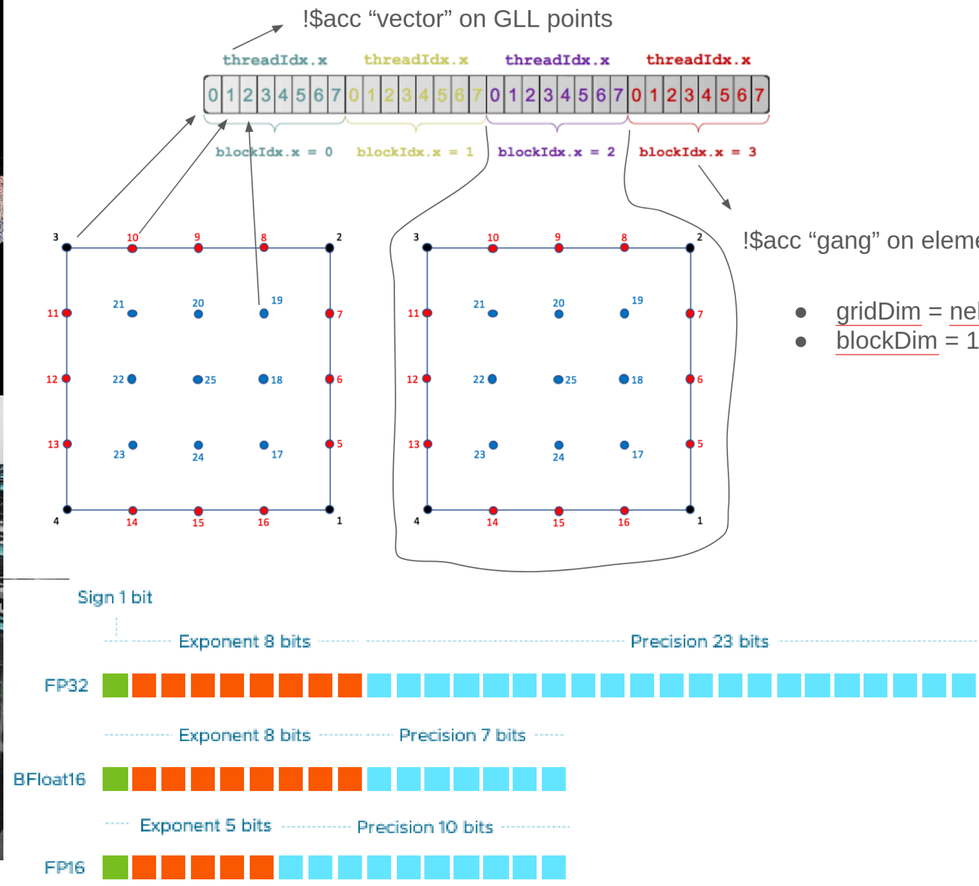
Motivation

- Enabling high-order methods for industrial simulations and research
- Reducing time-to-solution for large-scale numerical solutions
- Improve on existing solution frameworks (PETSc, Trilinos...)



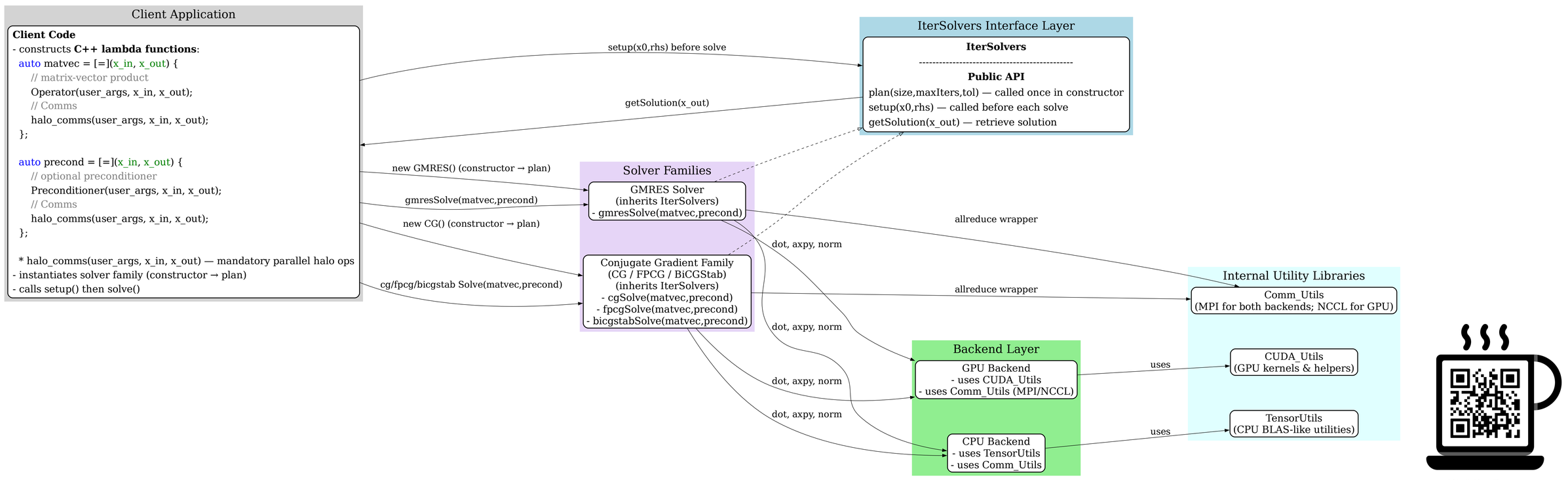
Challenges

- Ensure a flexible, portable design
- Easy client-framework integration
- Achieve high performance on modern GPU-based systems
- Support for accurate mixed-precision operations, including bfloats



GLASs Framework

- Modular C++ framework for building GPU-accelerated iterative linear algebra solutions (linear system solvers and eigensolvers)
- GPU kernels implemented in detached module, allowing for easy switch between CUDA, HIP or future backends at compile time
- Routine-driven design: client codes provide operations to methods, not data. Flexible design with minimal assumptions on data layout
- Templated design for easy mixed-precision support, including for experimental NVIDIA bfloat16 formats
- Full parallel execution over multiple GPUs: internal collectives rely on either GPU-aware MPI or vendor-specific CCL libraries (NCCL, RCCL)



GLASs Repo

FPCG test

- Simple tridiagonal, symmetric system
- Basic capability showcase
- Uses FP32 version of solver (mixed-precision)
- Tested on MN5 ACC nodes
- Preconditioner: Laplacian diagonal inverse
- Similar to a 1D FEM test, including halo communications with MPI

Scalability results