Optimising `MPI_Alltoallv` for 3D FFTs

Supervisors: **Steven Wright** (Computer Science), **Phil Hasnip** (PET)
RSE: **Matt Smith** (PET)

Distributed 3D Fast Fourier Transforms (FFTs) are a core computational component of materials modelling codes such as VASP and CASTEP (castep.org), often representing a significant scalability bottleneck due to their reliance on global communication. Recent research conducted on the Frontier supercomputer has shown that near Exascale, latency dominates the cost of `MPI_Alltoall` communication, even for very large message sizes [1].

White introduces several GPU-aware implementations designed to reduce latency at extreme scales. These include one-sided `MPI_Get` routines and hierarchical 2D/3D node-wise strategies that significantly outperform vendor-provided MPI libraries. However, previous work does not consider the `MPI_Alltoallv`, a variant where each task may send and receive different amounts of data. Adapting these advanced, uniform-size algorithms to support the variable message sizes of `MPI_Alltoallv` is crucial for realising these performance gains in some scientific computing applications.

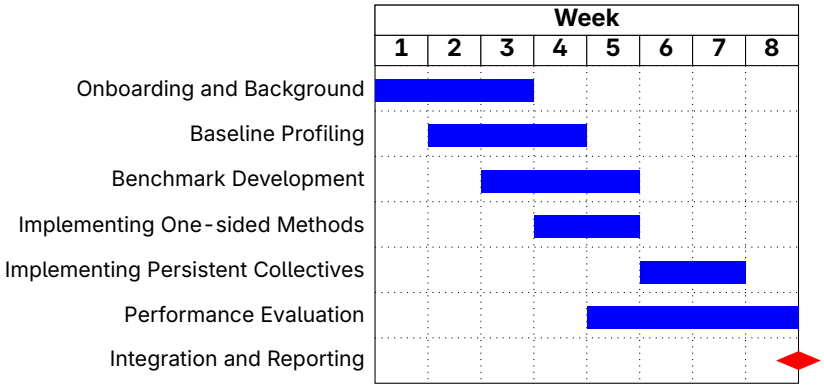
Project Goals

- **Analyse Performance Bottlenecks**: Profile existing `MPI_Alltoallv` implementations in 3D FFTs at scale to establish baseline performance metrics.
- **Algorithm Adaptation**: Adapt the high-performance `MPI_Alltoall` algorithms from recent literature to support the variable message sizes and offsets required by `MPI_Alltoallv`.
- **Implementation and Benchmarking**: Implement these GPU-aware communication strategies and benchmark them against vendor MPI libraries on a modern HPC system.
- **Integration**: Develop a roadmap for integrating the best solution into a relevant codebase.

Computational Aspects

The project will involve low-level programming in either C/C++ or Fortran, and interaction with GPU-aware MPI communication libraries, at scale on modern heterogeneous HPC systems. The intern will leverage benchmarking methodologies similar to the “Always” benchmark discussed in the literature and extend them to variable data sizes [1]. The overall aim of the project is to develop an optimised `MPI_Alltoallv` implementation, and demonstrate its performance on a realistic problem with a scientific application such as CASTEP.

Timeline



References

[1] James Buford White III, *Large-Message All-to-All Communication at Frontier Scale*. In Proceedings of the SC'25 Workshops. 2025. ACM, USA, 461–467.