1. Introduction

Many large-scale application-software require Fast Fourier Transforms (FFT), e.g., within the Exascale Computing Project (ECP) of the United States.

Hybrid CPU-GPU systems are widely used and are expected for the upcoming exascale machines. FFT libraries targeting such architectures have been accelerated via tuning and asynchronous kernel evaluation on GPUs [2], obtaining up to 2x speedup compared to fully CPU libraries.

We present techniques to further accelerate FFT computation by overcoming the communication bottleneck, we provide architecture-aware selection of FFT algorithm, a novel All-to-All routine (which can considerably speedup default MPI standard routines), and a mixed-precision implementation.

2. Even faster FFTs

Multidimensional FFT are computed by a sequence of 1D or 2D FFTs, with intermediate data reshapes. The latter is the most expensive task (>80% of runtime) [2], moving data among processors (1), typically in an All-to-All fashion [2, 3].

heFFTe supports any type of reshaping technique (c.f., Fig. 2.1) and provides a tool to create architecture-aware Phase diagrams (Fig. 2.2). In Fig. 2.2, we show the case of Summit, where users can input their resources and FFT size to select the fastest reshape approach.

3. Mixed-Precision FFT

Fig. 3 shows that heFFTe linearly scales. We use a 3D complex-data grid, and compute both: single (FP32) and double (FP64) precision FFTs.

We developed a mixed-precision version for heFFTe, which exploits GPU power to compress data (using casting/ZFP) to save in All-to-All cost (which usually takes over 90% of runtime [2, 3]). We used a ring version of our OSC_A2A routine.

CR2 means a compression ratio of 2 times. CR2 is up to 2.6x faster than FP64 and up to 1.4x faster than FP32. CR4 is up to 5x faster than FP64 and up to 2.6x faster than FP32. CR2 validation error is O(10^-5) while for FP32, it is O(10^-6); i.e., we move the same data volume faster, while getting a better quality FFT output.

4. Communication model

We introduce a novel communication model for hybrid-distributed FFTs which can adapt to any architecture [2], and gives a theoretical estimation of the reshaping cost.

This model assumes that fast communication is available within a node, e.g., Summit at ORNL, which has NVLink connections.

For Summit, we get [2]:

\[ T_{sol} = \frac{7.8125 \times \text{log}(N)}{2^{17}} \]

where \( N \) is the FFT size.

Using this model, we derive a roofline model theoretical peak performance. Fig. 4 shows heFFTe achieving about 90% of peak value up to 48 GPUs; this percentage then decreases, indicating that too many resources are being used for the given FFT size (latency).