Characterization of Accelerated 2D FFT with Off-Chip Optical Channels and Kernel Adaptation for Efficient Utilization

Ke Wen[1], Sébastien Rumley[1], Paolo Mantovani[2], Luca Carloni[2] and Keren Bergman[1]
[1] Department of Electrical Engineering, Columbia University, New York, NY
[2] Department of Computer Science, Columbia University, New York, NY

***Abstract*—Off-chip data movement capability is critical to the performance of current accelerator-oriented embedded systems. Integrated optical chip-memory links provide a promising alternative to electrical wires with unprecedented bandwidth and scalable energy efficiency. This work characterizes 2D-FFT, an off-chip-bandwidth consuming workload, with such optical communication capability. While efficient utilization of optical channels cannot solely rely on wire-to-wire replacement, a tiling-based kernel adaptation strategy is proposed and shown to achieve 22% reduction in execution time compared to the non-adapted implementation under the same bandwidth provision.**

I. Introduction

High-performance embedded systems are often limited by off-chip data movement capabilities. Such limitations are typically attributed to the number of available off-chip pins and their constrained bandwidth (typical a few Gb/s) [1]. This leaves an overall maximal I/O envelope for current systems of less than a few hundreds of Gb/s based on electrical technologies, which is insufficient for workloads with high byte/FLOP ratios, such as FFT (256-point FFT with 0.4 B/FLOP can consume 2.5 TB/s off-chip bandwidth). Beyond the bandwidth, the I/O distance constraints of electrical wires can be a second contributor to performance limitations [2]. With only a few cm's reach, components such as memory have to be located in close proximity to the processing in order to suppress energy dissipation or signal degradation.

In comparison, integrated optical communication promises to provide high off-chip bandwidths with extremely scalable energy efficiency [3]. With wavelength division multiplexing, an optical waveguide or fiber can support hundreads of wavelengths each with tens of Gb/s bandwidth. Furthermore, with low loss, optical links can span tens of meters without significantly impairing the signal quality. The properties of optical data movement enable designs of flattened-distributed computing architectures [4] and potentially distance-independent memory hierarchies.

However, successful utilization of optical bandwidth cannot rely solely on wire-to-wire replacements. This work explores the communication characteristics and performance of accelerated 2D-FFT workloads running with optical off-chip communication capabilities. A tiling-based kernel adaptation strategy is implemented to enable the 2D-FFT workload for more efficient bandwidth utilization and performance scaling.

II. Characterization of Accelerated 2D-FFT Kernel

The architecture of the accelerated 2D-FFT embedded system is shown in Fig. 1. An optical link provides high-bandwidth read/write transaction capability between the on-chip scratchpad (input/output buffer) and the off-chip memory. The accelerator on chip represents a hardware-accelerated 2D-FFT algorithm from the PERFECT benchmark suite [5]. The

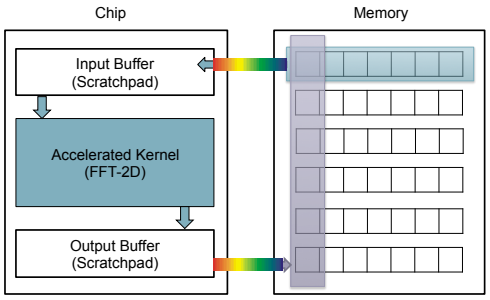

Fig. 1. Architecture of accelerated 2D-FFT kernel with optical chip-memory communication capability.

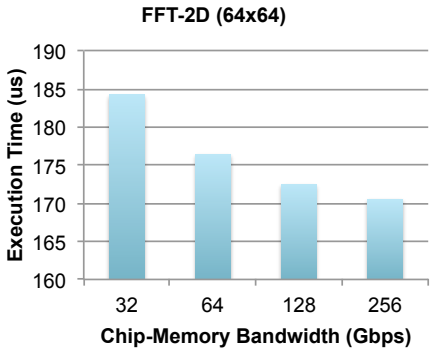

Fig. 2. Perforamce of 2D-FFT (basic algorithm implementation) versus provisioned optical chip-memory bandwidth.

algorithm follows a basic workflow of 2D-FFT: first, DFT is performed on the row dimension, the matrix is then transposed, and another DFT can be performed on the column dimension. Large data set is considered such that the on-chip scratchpad can hold at most one row of the original matrix at one time. The transpose operation is performed during intermediate data write-back, where data elements are written into the memory in a column-wise manner.

Table 1 presents the communication characterization of the above 2D-FFT kernel. In this basic algorithm, only the memory READ operation comprises continuous data access (row access), while the WRITE operation, due to the need for transpose, has to access memory in a column-wise manner. Due to the strided memory addresses, the WRITE operation can only write one data element at one time. Thus, under traditional algorithmic flow, the WRITE operation of the 2D-FFT does not fully benefit in sustainable performance speedup with increased bandwidth, leading to highly-unbalanced I/O requests (8192 WRITEs versus 128 READs, Table 1). The execution time result in Fig. 2 confirms this effect with an execution time reduction that is not proportional to the bandwidth scaling. For example, with bandwidth increasing from 32 to 256 Gb/s, the execution time decreases by only 8%.

Table 1. Memory communication characteristics of *Basic* and *Tiled* implementations for a 64×64 2D-FFT. A tile is of size 8×8.

	Basic Alg	*Tiled* Alg
Read burst length	64	8
Write burst length	1	8
Read requests	128	1088
Write requests	8192	1088

III. Kernel Adaptation Strategy for Efficient Bandwidth Utilization

To improve the kernel's capability in utilizing the optical bandwidth, a 2D data-decomposition based 2D-FFT algorithm [6] is implemented. The implementation is intended to 1) eliminate the need for column-wise memory operation and 2) balance the number of read/write requests.

The tiling-based algorithm partitions an input matrix ($M \times N$) into a $p \times q$ mesh, where each sub-block in the mesh is of size $M/p \times N/q$. The algorithmic workflow comprises row-wise and column-wise strided FFT's and local tile-size 2D-FFT's. In the phase of row-wise strided FFT, each element in a sub-block performs data exchange with the corresponding elements in the other (q-1) sub-blocks of the same row. Similarly, in column-wise strided FFT, each element performs data exchange with corresponding elements in the other (p-1) sub-blocks of the same column. Under these two operations, one method to create burst (instead of single-element) communication is to fetch a burst of continuous data from each sub-block, where each bursts have the same in-block position and together fill up the scratchpad. Row-wise and col-wise strided FFT's are then performed on the data in the scratchpad. Since the data needed is fully on-chip, no off-chip transpose or col-wise memory WRITE is required. For the local sub-block 2D-FFT's, the scratchpad fetch or write back a full row of current sub-block per memory transaction, hence avoiding single-element WRITEs as well.

Characterizations of both the basic (w/o tiling) and the tiled 2D-FFT workloads are compared here. As Table 1 shows, the request number and burst length of read/write operations are fully balanced in the tiled implementation. As a result, the amount of time spent by the kernel in waiting for WRITE transactions to finish significantly reduces (by 81%, Fig. 3). Fig. 4 compares the execution time of the basic versus the tiled implementation. When both implementations are given the same amount of bandwidth increase (32 Gb/s to 256 Gb/s), the basic algorithm only reduces 8% of the execution time due to the incapability in utilizing the increased bandwidth. Whereas, the tiled implementation reduces the execution time by 28% compared to the baseline, cutting down 22% of the time of the basic algorithm. Such speedup is not based on increase of arithmetic on chip, but based on the much more efficient utilization of the output bandwidth and the significant decrease in total WRITE time as shown be Fig. 3.

IV. Conclusion

In this work, we show that optical off-chip bandwidth integration can effectively boost the performance of embedded 2D FFT kernels. Such integration, however, might not solely rely on wire-to-wire replacement, but can also require kernel-specific adaptation. A 2D data-decomposition based adaptation technique is applied to 2D-FFT to avoid single-element column WRITEs. Such adaptation is shown to effectively unleash the advantage of high-bandwidth optics and balance the requests for READs and WRITEs. For a 64 × 64 2D-FFT, the tiled algorithm achieves 22% reduction in execution time compared to the non-adapted version under the same provision of bandwidth increase.

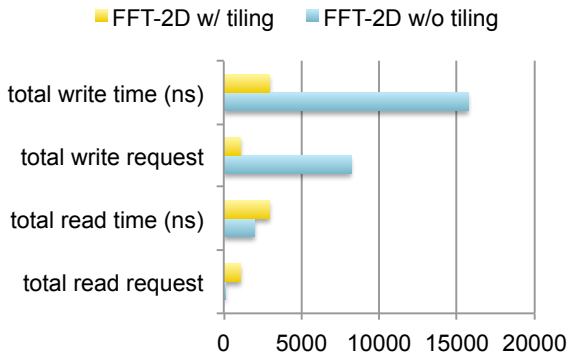

Fig. 3. Comparison of total memory transaction time and requests between basic and tiled 2D-FFT implementations.

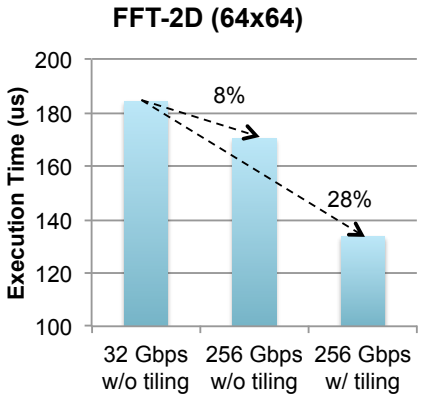

Fig. 4. Performance comparison between basic and tiled 2D-FFT implementations under same amount of bandwidth increase (8x).

References

[1] Young, Ian A., et al. "Optical I/O technology for tera-scale computing." Solid-State Circuits, IEEE Journal of 45, no. 1 (2010): 235-248.

[2] Miller, David AB. "Rationale and challenges for optical interconnects to electronic chips." Proceedings of the IEEE 88.6 (2000): 728-749.

[3] Ophir, Noam, Christopher Mineo, David Mountain, and Keren Bergman. "Silicon photonic microring links for high-bandwidth-density, low-power chip I/O." Micro, IEEE 33, no. 1 (2013): 54-67.

[4] Wen, Ke, et al. "Reuse Distance Based Circuit Replacement in Silicon Photonic Interconnection Networks for HPC." High-Performance Interconnects (HOTI), 2014 IEEE 22nd Annual Symposium on.

[5] Kevin Barker, Thomas Benson, Dan Campbell, David Ediger, Roberto Gioiosa, Adolfy Hoisie, Darren Kerbyson, et al. "PERFECT (Power Efficiency Revolution For Embedded Computing Technologies) Benchmark Suite Manual." Pacific Northwest National Laboratory and Georgia Tech Research Institute, December 2013.

[6] Yu, Chi-Li, Jung-Sub Kim, Lanping Deng, Srinidhi Kestur, Vijaykrishnan Narayanan, and Chaitali Chakrabarti. "FPGA architecture for 2D Discrete Fourier Transform based on 2D decomposition for large-sized data." Journal of Signal Processing Systems 64, no. 1 (2011): 109-122.