A Case for Embedded FPGA-based SoCs in Energy-Efficient Acceleration of Graph Problems

Pradeep Moorthy, Siddhartha, Nachiket Kapre
Nanyang Technological University, Singapore
nachiket@ieee.org

***Abstract*—**
Sparse graph problems are notoriously hard to accelerate on conventional platforms due to irregular memory access patterns resulting in underutilization of memory bandwidth. These bottlenecks on traditional x86-based systems mean that sparse graph problems scale very poorly, both in terms of performance and power efficiency. A cluster of embedded SoCs (systems-on-chip) with closely-coupled FPGA accelerators can support distributed memory accesses with better matched low-power processing. We first conduct preliminary experiments across a range of COTS (commercial off-the-shelf) embedded SoCs to establish promise for energy-efficiency acceleration of sparse problems. We select the Xilinx Zynq SoC with FPGA accelerators to construct a prototype 32-node Beowulf cluster. We develop specialized MPI routines and memory DMA offload engines to support irregular communication efficiently. In this setup, we use the ARM processor as a data marshaller for local DMA traffic as well as remote MPI traffic while offloading compute-intensive portions on the FPGA. Across a representative set of benchmark graphs, we show that embedded SoCs with FPGA accelerators can exceed the energy efficiency of an Intel E5-2407 by as much as 1.7× at a total graph processing capacity of 91–95 MTEPS.

I. Introduction

During the pioneering years of HPC, computer architects exclusively built systems from specialized vector hardware; such as the Cray-I [7] and other bespoke machines like as the NEC SX-3 and Fujitsu Numerical Wind Tunnel. The early 90s saw x86-based systems rise in popularity due to their low cost, simplicity and standardization of the ISA/floating-point system (Intel 8087 was an early example of IEEE-754 compliant processor hardware). Beowulf clusters of these x86 platforms began as low cost hobbyist alternative to state-of-art HPC systems. Based on the idea of connecting relatively inexpensive COTS computers to collectively solve a particular problem, the first such cluster was developed in 1994 by connecting 16 Intel DX4 processors with 10Mbps Ethernet. This eventually paved way for the creation of the first cluster based supercomputer in 1997, the ASCI Red, which employed 7,246 Intel x86 Pentium Pro processors linked using a custom-interconnect architecture. Peaking the TOP500 list for nearly three years, it set out the foundation for the dominance of x86 cluster systems we see today.

The same era saw the introduction of Reduced Instruction Set Architecture (RISC) based systems in place of Complex Instruction Set Architecture (CISC) machines in the form of PowerPC processors used in the IBM BlueGene. This supercomputer series was launched in 2004 to exploit the low power capabilities of RISC instead of CISC chips by combining multiple PowerPC processors onto each chip. Thus, the usage of multiple low power processors, typically RISC based, in place of a single power hungry "fat" processor was recognized as a way to improve energy efficiency. In lieu of PowerPC hardware, ARM chips are gaining more interest in the research community since they are fabricated extensively in mobile devices yielding economies of scale to deliver low power at low cost. The largest ARM-based cluster studied was the Tibidabo cluster [6], which consisted of 192 NVIDIA Tegra-2 SoCs, interconnected using 1GbE network. The study concluded the lack of high-bandwidth I/O interfaces such as 10GbE/InfiniBand and the absence of hardware support for interconnection protocols on the Tegra-2's ARM Cortex-A9 processor as the sole limiting factors in adopting the SoC for HPC usage. While the present day performance gap between HPC-grade x86 processors and commercial ARM processors can be as high as an order of magnitude, large graph problems with low spatio-temporal locality can eliminate the performance gap between the two architectures while retaining the energy efficiency advantages. To investigate this claim, we prototype a Beowulf cluster composed of 32 Xilinx FPGA-based Zynq SoC boards, interconnected using a Gigabit Ethernet Switch. We map sparse-graph oriented irregular computations of varying dimensions to stress the memory and network throughputs of the cluster nodes. Figure 1 shows a photograph of our "Zedwulf" (ZedBoard+Beowulf) cluster.

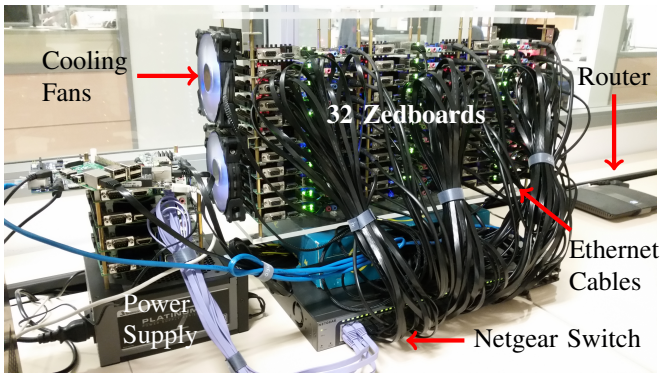

Fig. 1: Zedwulf Cluster: 32 Zynq Z7020 SoC boards

In this paper, we make the following key contributions:

- **Microbenchmarking of COTS SoCs**: We analyze the memory potential and network characteristics of various embedded SoCs using micro-benchmarking tools.

TABLE I: Comparing datasheet specifications and microbenchmark performance of various COTS embedded SoC platforms

	Zedboard	Microzed	Parallella Epiphany	Intel Galileo	Raspberry Pi	Beaglebone Black
Technology	28nm	28nm	28nm	32nm	40nm	45nm
SoC	Xilinx	Xilinx	Xilinx	Intel	Broadcom	TI
	Zynq 7020	Zynq 7010	Zynq 7010	Quark X1000	BCM2835	AM3359
Processor	ARMv7, FPGA	ARMv7, FPGA	ARMv7, FPGA, Epiphany III	i586	ARMv6	ARMv7
Clock Freq.	667 MHz CPU	667 MHz CPU	667 MHz CPU	400 MHz	700 MHz	1 GHz
	250 MHz FPGA	250 MHz FPGA	250 MHz FPGA			
On-chip	32 KB L1	32 KB L1	32 KB L1	16 KB L1	16 KB L1	32 KB L1
Memory	512 KB L2	512 KB L2	512 KB L2		128 KB L2	256 KB L2
	560 KB FPGA	560 KB FPGA	240 KB FPGA		-	
Off-chip	512 MB	1024 MB	1024 MB	256 MB	512 MB	512 MB
Memory	32b DDR3-1066	32b DDR3-1066	32b DDR3-1066	32b DDR3-800	32b DDR2-400	16b DDR3-606
DMIPS	1138	1138	1138	237	862	1778
Coremark	1591	1591	1782	526	1314	2457
Network (Bi-dir)[3]	57 MB/s	59 MB/s	32 MB/s	18 MB/s	10 MB/s	21 MB/s
L1 B/W	7.7 GB/s	7.7 GB/s	7.5 GB/s	2.8 GB/s	2.7 GB/s	7.6 GB/s
L2 B/W	1.4 GB/s	1.4 GB/s	1.4 GB/s	-	1.4 GB/s	3.4 GB/s
DRAM Seq.	654 MB/s	641 MB/s	537 MB/s	270 MB/s	187 MB/s	278 MB/s
DRAM Random	32 MB/s	32 MB/s	28 MB/s	12 MB/s	10 MB/s	11 MB/s
Power	5 Watts	3.6 Watts	7.5 Watts	4 Watts	3.75 Watts	3.25 Watts

[3]Intel MPI Benchmark Suite result for `MPI_Sendrecv` for all systems in 2-node configurations

- **Prototype a 32-node Zynq SoC cluster**: We physically prototype a 32-node Zynq SoC cluster using the Xilinx Zedboard and Microzed platforms.
- **Communication optimization for sparse-graph access on the Zynq cluster**: We develop customized Message Passing Interface (MPI) routines and DMA engines optimized for irregular access exhibited by graph problems
- **Performance and power evaluation of the Zynq cluster vs an x86 server node**: We benchmark our cluster for a few representative sparse graphs and compare against the Intel E5-2407 CPU.

II. Microbenchmarking COTS SoC Platforms

We first evaluate a range of COTS embedded SoC-based platforms listed in Table I to assess their feasibility for scaling to larger-scale systems. Our characterization experiments focus on a single chip and measure raw compute throughput, memory performance as well as MPI support for these systems.

Recent academic studies have examined the feasibility of HPC systems based on mobile SoCs [5] for HPC-oriented workloads and investigated the status of networking support in these SoCs. Additionally, there are many contemporary hobbyist clusters built from **Apple TV** [2], **Raspberry Pi** [1], and **Beagleboard xM** [4] that use off-the-shelf devices for delivering proof-of-concept systems with high power efficiency. These studies are insightful but it remains to be seen if pure ARM-based SoCs have future prospects in the cluster computing space.

Our preliminary experiments on the **Intel Galileo 2** platform indicate the Quark SoC would not be competitive at this stage with its under-powered 400 MHz 32b CPU when compared to ARM-based embedded SoC platforms. It reported the lowest DMIPS score of 237 and had poor Ethernet throughput of 10 MB/s (100M Ethernet NIC). Occupying the lower-end of the ARM spectrum, the **Raspberry Pi** reported a 3x higher DMIPS/Coremark score than the Galileo. Nevertheless, its relatively slower DDR2 memory limits the overall performance gains. The **Beaglebone Black** further doubles the compute performance to 1778 DMIPS. However, the 16b 400 MHz DDR3 memory barely keeps up with its superior compute capabilities constraining overall performance. Besides, these devices are also limited by 100 Mb/s network links. The Zynq SoC-based platforms (Zedboard, Microzed and Parallella) overcome some of these shortcomings by coupling the Zynq SoC to a 1 Gb/s network link and a respectable 32b DDR3 1066 MHz memory. The **Zedboard** and **Microzed** delivered the highest sequential and random access memory bandwidths. Complemented by the Gigabit Ethernet connectivity, these platforms averaged bi-directional network throughput at a high 60 MB/s. Nonetheless, that corresponds only to a network efficiency of 24%. This behavior is attributed to the slower clock rate of the ARM cores (35% slower ARM CPU relative to the Beaglebone running at 1 GHz). In addition to the Zynq SoC, the **Adapteva Parallella** [3] platform also attaches an Epiphany floating-point co-processor as a separate chip thereby improving its compute capability substantially. We recorded comparable DMIPS and memory bandwidth scores on the Zedboard/Microzed, but the network throughput saturated at a disappointing 32 MB/s. The high local DRAM and remote MPI throughputs suggest that the Zedboard can

become a viable candidate for energy-efficient operation for sparse irregular workloads. It is worth noting that these Zynq platforms are development systems with extraneous supporting logic for audio, video and configurable IOs that can be eliminated in a pure datacenter/HPC-focussed design.

III. Zedwulf Organization

The Zedwulf cluster is composed of 32 Zedboards (Rev. D) or 32 Microzed (eval. kit), interconnected using a Netgear GS748T 48-port Gigabit Smart Switch. With a rated switching capacity of 96 Gb/s, the switch can sustain 2 Gb/s duplex bandwidth per 1GbE ethernet link connecting each Zedboard. We powered the system using a Seasonic Platinum 1KW PSU, from the PCIe EPS12 power rail with fuse protection. We stacked the Zedboards on top each other in three columns with 10/11 boards on each column. We provided air cooling from 2 fans placed on either sides of the stack (4 fans total) as shown in Figure 1. While every Zedboard has a SanDisk Ultra 32 GB SD card attached to host the OS, the master node has an additional Samsung 840 Pro SSD attached to the USB2 port using a SATA-USB adapter. We setup the SSD as the primary secondary storage device for our cluster to hold our large graphs and it offers a convenient lower latency solution for quickly loading and distributing sub-graphs. A single Zynq node with various interface bandwidths is shown in Figure 2. We also built a 32-node Microzed cluster by simply replacing the Zedboard with Microzeds.

The Zynq is a heterogeneous multicore system architecture consisting of a dual-core ARM Cortex-A9 on the Processing System (PS) and a FPGA fabric on the Programmable Logic (PL). Residing on the same chip, the PS and PL are interconnected using AXI on-chip buses. This contrasts to traditional FPGA implementations, whereby the latter is connected to an x86 host using PCIe buses. This approach allows ARM processors to benefit from low-latency links to the FPGA which allow tightly-coupled CPU-based control of FPGA operation.

We configured each Zedboard to run Xillinux-1.3a, an Ubuntu-12.04 based Linux distribution with Xillybus drivers to communicate with the FPGA using an AXI 2.4 GB/s channel. We compiled software libraries such as MPI and other utilities with `gcc-4.6.3` with appropriate optimization flags enabled. We use NFS (Network File System) to synchronize files (graphs) across all 32 nodes. We setup MPI to use Myrinet Open-MX patch to deliver a marginal improvement in network latency. We also choose MPICH over OpenMPI as it provided a 20–30% lower latency and higher bandwidth in our initial stress benchmarks.

IV. Communication Optimization

Graph processing is a communication-dominated algorithm that can often be organized as lightweight computations on vertices and message-passing along edges. We map bulk-synchronous parallel (BSP) graph computations to our cluster by careful optimization of local communication (irregular memory access) and remote communication (MPI access).

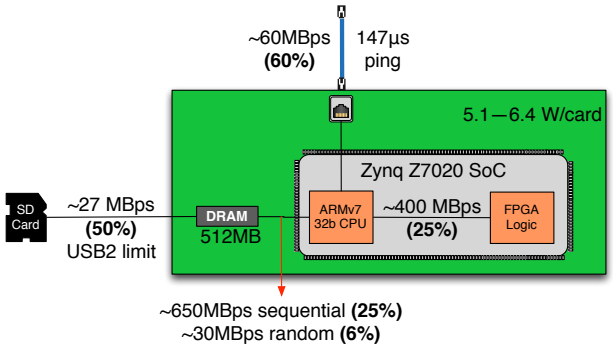

Fig. 2: A Zynq node (Zedboard) with peak and achieved bandwidths

A. MPI Optimization

Partitioning the graph structure to fit across multiple *Processing Elements (PEs)* creates network traffic which connect local vertices to vertices present in other PEs. Unlike local edges, which connect vertices present within the same PE, updating remote edges is typically an order of magnitude slower as the data needs to be transferred from the origin PE to the target PE using the ARM CPUs to handle network packet transfers. Hence, there is an inherent need to reduce the time spent in fulfilling the network operations for maximizing performance gains while using distributed systems. We designed an optimized graph-oriented global scatter technique using the Message Passing Interface (MPI) library.

Our approach leveraged coalesced data transfers between PEs to take advantage of the network bandwidth, rather than being limited by the high network latencies. We used **MPI_type_indexed** API to encode the *send* and *receive* buffer displacements in an MPI friendly manner. We then employed **MPI_Sendrecv** as the building block of our scatter routine. The send-recv operations were scheduled in a periodic fashion to avoid network contention across MPI nodes. This coalesced approach of edge updates offered a speedup of **60**× when compared to performing fine-grained message transfers.

B. Memory Access Optimization

For each vertex in the graph, the graph processor needs to fetch adjacent vertex data from local memory wherever possible. The graph data is conventionally stored in a compressed sparse format (row based or column based), which is a memory storage optimization for sparse graph structures. However, memory access patterns can still result in frequent cache misses under this memory organization scheme.

While the FPGA on the Zedboard has 560KB of on-chip memory, they can barely accommodate 100-1000s of graph vertex and edges. Using the off-chip DRAM memory carelessly would result in poor DRAM bandwidth utilization. Hence, we designed a Memory Management Unit (MMU) for Zedwulf to optimize irregular data transfers. We configure the AXI DMA IP block to use low-level AXI descriptor chains to encode the sparse graph access sequence. With our approach we are able to improve random DRAM access throughput for graph operations by as much as **3–4**×.

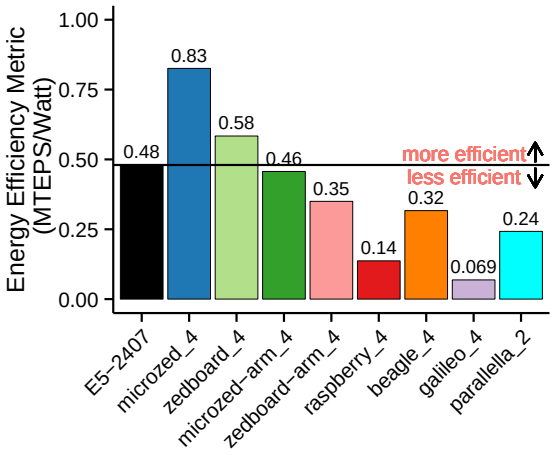

Fig. 3: Performance-Power Tradeoffs across embedded SoCs platforms (4-node and 2-node SoCs) and a single x86 node ("-arm_4" versions exclude FPGA and only use ARM)

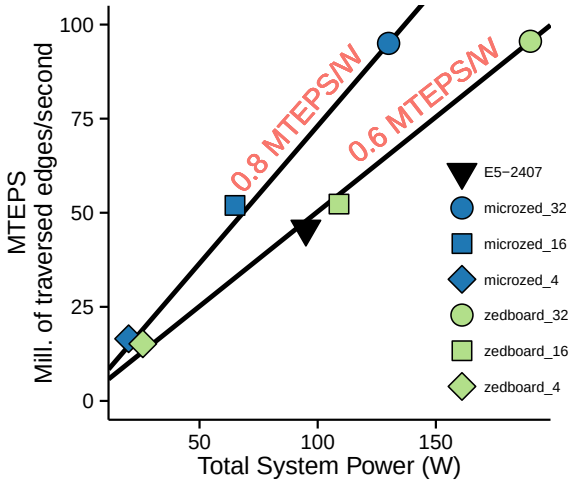

Fig. 4: Energy Efficiency of Zynq FPGA cluster against an x86 node (4-node, 16-node and 32-node Zynq setups)

V. Results

We analyze the performance and energy-efficiency of various embedded clusters for sparse graph processing. We perform bulk-synchronous evaluation on randomly-generated graphs with upto 32M node and edges. For the first experiment, we setup 4-node clusters of each unique embedded platform (except Parallella with 2 nodes) and compare it against one x86 node. We scale our setup for the second experiment where we compare the 32-node Zedboard and 32-node Microzed clusters against a single x86 node. For compleness, our power measurements include the Ethernet switch and PSU along with the Zynq boards.

In Figure 3, we plot processing efficiency (MTEPS/W) across various embedded and x86 platforms. The Galileo and Raspberry Pi clusters have the lowest performance while demanding high power usage. The Beagle cluster doubles the performance achieved while consuming 10% less power, therby improving the power efficiency. The 2-node Parallella cluster nearly matches the performance of the 4-node Beaglebone but it needs more power for the extra Epiphany co-processor. The Zedboard and Microzed boards offer the highest energy efficiency when using the FPGA accelerators instead of simply relying on their ARM CPUs. The Microzed stands out with its 30% less power use over the Zedboard as it eschews unnecessary development suport (audio, video, IO chips) in favor of a low-cost implementation.

In Figure 4, we show the performance (in MTEPS, millions of traversed edges per second) of the x86 node and the Zynq clusters plotted against their measured power consumption. We are able to marginally exceed the energy efficiency of the x86 node (0.48 MTEPS/W vs. 0.58 MTEPS/W) when using the Zedboard cluster. However, the lower-power and cheaper Microzed-based cluster is able to deliver a 1.7× improvement in energy efficiency (0.83 MTEPS/W) due to its lean design.

This measured 0.83 MTEPS/W energy efficiency figure compares favorably to select entries in the Green Graph500 list. We look forward to implementing the Graph500 benchmarks in the near future that builds upon this work.

VI. Conclusions

We show how to use the Zynq SoC with ARMv7 32b CPUs supported by FPGA-accelerators to prototype energy-efficient HPC systems for sparse graph acceleration. For a range of graphs, we are able to deliver a performance of 91–95 MTEPS at an energy-efficiency of 0.58 MTEPS/Watt (32-node Zedboard), and 0.83 MTEPS/Watt (32-node Microzed) which exceeds the x86 efficiency of 0.48 MTEPS/Watt by as much as 1.7×. While the Zynq SoC we evaluated in this study is promising, performance gains were limited by (1) slow 1G Ethernet speeds of 50% peak, (2) limited DRAM capacity per node 512 MB, (3) poor CPU-FPGA link bandwidth of 400 MB/s, and (4) extraneous devices and interfaces for audio/video processing. Upgraded Zynq SoCs optimized for data-center processing that address these concerns can further improve performance and energy efficiency of these systems.

References

[1] S. J. Cox, J. T. Cox, R. P. Boardman, S. J. Johnston, M. Scott, and N. S. O'Brien. Iridis-pi: a low-cost, compact demonstration cluster. *Cluster Computing*, 17(2):349–358, June 2013.

[2] K. Fürlinger, C. Klausecker, and D. Kranzlmüller. The AppleTV-cluster: Towards energy efficient parallel computing on consumer electronic devices. *Whitepaper, Ludwig-Maximilians-Universitat*, 2011.

[3] L. Gwennap. Adapteva: More Flops, Less Watts. *Microprocessor Report*, pages 1–5, June 2011.

[4] E. Principi, V. Colagiacomo, S. Squartini, and F. Piazza. Low power high-performance computingon the Beagleboard platform. In *Education and Research Conference (EDERC), 2012 5th European DSP*, pages 35–39, 2012.

[5] N. Rajovic, P. M. Carpenter, I. Gelado, N. Puzovic, A. Ramirez, and M. Valero. Supercomputing with commodity CPUs. In *the International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–12, New York, New York, USA, 2013. ACM Press.

[6] N. Rajovic, A. Rico, N. Puzovic, C. Adeniyi-Jones, and A. Ramirez. Tibidabo: Making the case for an ARM-based HPC system. *Future Generation Computer Systems*, 2013.

[7] R. Russell. The CRAY-1 Computer System. *Commun. ACM*, 21(1):63–72, Jan. 1978.