A 1024 RV-Cores Shared-L1 Cluster with High Bandwidth Memory Link for Low-Latency 6G-SDR

Yichao Zhang*, Marco Bertuletti*, Chi Zhang*, Samuel Riedel*, Alessandro Vanelli-Coralli*† and Luca Benini*†
*ETH Zürich, Zürich, Switzerland †Università di Bologna, Bologna, Italy
Email: yiczhang, mbertuletti, chizhang, sriedel, avanelli, lbenini, @iis.ee.ethz.ch

***Abstract*—We introduce an open-source architecture for next-generation Radio-Access Network baseband processing: 1024 latency-tolerant 32-bit RISC-V cores share 4 MiB of L1 memory via an ultra-low latency interconnect (7-11 cycles), a modular Direct Memory Access engine provides an efficient link to a high bandwidth memory, such as HBM2E (98% peak bandwidth at 910 GBps). The system achieves leading-edge energy efficiency at sub-ms latency in key 6G baseband processing kernels: Fast Fourier Transform (93 GOPS/W), Beamforming (125 GOPS/W), Channel Estimation (96 GOPS/W), and Linear System Inversion (61 GOPS/W), with only 9% data movement overhead.**

***Index Terms*—Many-core, RISC-V, SDR, 6G**

I. Introduction

Beyond 5G, baseband compute demands rapidly increase with the number of antennas and sub-carriers. For example, Physical Uplink Shared Channel (PUSCH) processes massive-Multiple-Input, Multiple-Output (MIMO) transmissions over 100 MHz frequency bandwidth within 1 ms. Energy-efficient many-core Systems-on-Chip (SoCs) can provide the flexibility needed to keep up with evolving standards, following the Software Defined Radio (SDR) paradigm. However, it is essential for these programmable SoCs to meet the tight performance and efficiency requirements.

We introduce the TeraPool-SDR SoC[1], featuring 1024 RISC-V cores, 4 MiB of multi-banked (4096 banks) L1 Scratchpad Memory (SPM), and a minimal overhead link to High Bandwidth Memory (HBM2E) main memory. Our contributions are: ① The physical design of a cluster with the largest core count reported for 6G-SDR workloads (Tab. I). Its ultra-low latency core-to-L1 interconnect, operates at up to 924 MHz (TT/0.80 V/25 °C) in GlobalFoundries' 12 nm FinFET technology. ② A modular Direct Memory Access (DMA) engine enables transfers between L1 and main memory via a hierarchical Advanced eXtensible Interface (AXI) tree. The open-sourced cycle-accurate *DRAMsys5.0* simulator [1] is used for fast runtime main memory co-simulation[2]. We use an HBM2E main memory model to demonstrate our DMA engine capability to manage ultra-high bandwidth data flows, while the large L1 hides 130 cycles average transfer latency, without affecting kernel performance. We achieve 0.18 to 0.84 TOPS at 60 to 125 GOPS/W on key 6G-SDR kernels with only 9% data movement overhead, while meeting 1 ms PUSCH latency requirements at less than 8.8 W cluster power consumption.

[1]https://github.com/pulp-platform/mempool
[2]A QuestaSim linkable dynamic library is created to speed up runtime.

II. Architecture

TeraPool-SDR's Processing Elements (PEs) are single-stage, latency-tolerant *Snitch* cores [2]. The cores-L1 interconnect (snapshot Fig. 2, implementation details Table I) consists of a 3-level hierarchy of Fully-Connected (FC) crossbars (Tile, SubGroup, Group) with pipeline cuts for timing closure (Fig. 1). Therefore, access latency varies across the hierarchy and is parametrizable for the point-to-point connections between Groups (1-3-5-7/9/11 cycles) depending on timing constraints. At the system level (Fig. 3), we design a modular DMA split in three: frontend (configuration), midend (transfer split), and backend (data mover). One Subgroup Tile-shared master AXI port, supports L1 Instruction Cache (I$) refill or DMA-controlled data transfers. The Cluster AXI masters demultiplex to DMA-frontend, L2, and Control Status Register (CSR)/Peripherals. Fig. 2 shows ultra-low (14.1%) Cluster area overhead in Gate Equivalent (GE) for the interconnects (FC-L1-crossbar, AXI & DMA). We connect the L2-AXI master with two 16 GiB HBM2E stacks. An address scrambler aligns the data interleaved across HBM2E channels to burst length, reducing AXI conflicts during transfers.

III. 6G-SDR Key Kernels and Results

Fig. 4 shows key kernels flow for PUSCH Demodulation Reference Symbols (DMRSs) and Data symbols. We consider 64 antennas, 3276 sub-carriers, 32 beams, 4 transmitters [3]. The problem size is large. For example, in small clusters, when Fast Fourier Transform (FFT) data does not fit in L1, demodulated sub-carriers must be stored in L2 before Beam Forming (BF). Multiplication by BF coefficients is tiled. TeraPool-SDR with its workload-sized 4 MiB shared L1, reduces the required data splitting significantly. Data transfer overhead through HBM2E (910 GBps, 98% efficiency) for all benchmarked kernels is <9% (Fig. 4), even in presence of a considerable 130 cycles average latency. All kernels have >0.6 instructions-per-cycle (IPC) (Fig. 5), as we exploit data locality and proximity of memory hierarchies to minimize Load Store Unit (LSU) stalls from contentions to the interconnect shared resources. TeraPool-SDR achieves energy efficiency greater or equal than a 4× smaller, 256 cores shared-1 MiB MemPool cluster [2], thanks to its low-power interconnect (Fig. 6). TeraPool-SDR$_{1\text{-}3\text{-}5\text{-}9}$ is optimal for energy efficiency, and TeraPool-SDR$_{1\text{-}3\text{-}5\text{-}11}$ excels in performance for the selected SDR workload. Achieved <70 µs latency per data symbol at a power consumption below 8.8 W demonstrates that TeraPool-SDR is suitable for baseband operations.

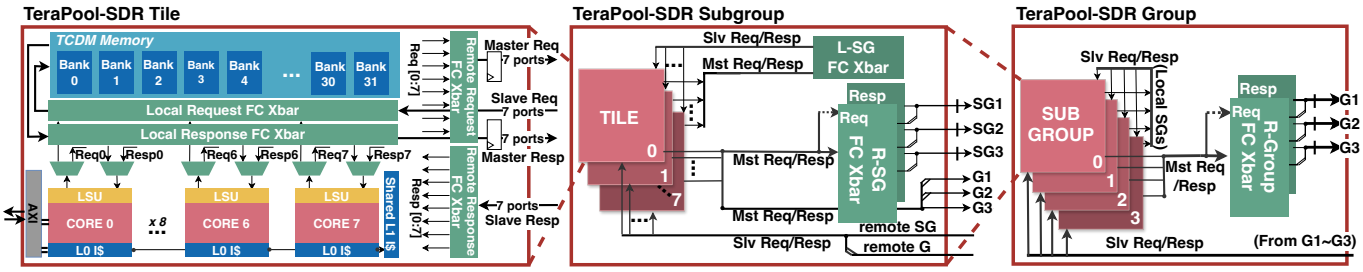

Fig. 1: Architecture of TeraPool-SDR$_{1\text{-}3\text{-}5\text{-}X}$. Subscript stands for cores access latency to banks in each hierarchical level.

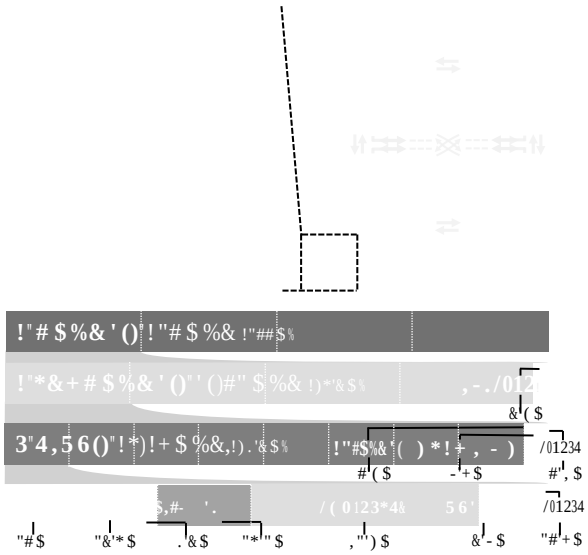

Fig. 2: PnR layout view in GlobalFoundries' 12 nm FinFET and hierarchical area breakdown of TeraPool-SDR Cluster.

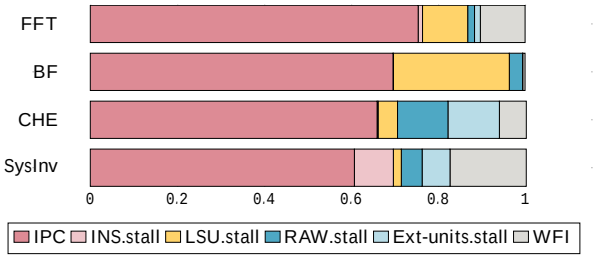

Fig. 5: Fraction of instructions and stalls over the total cycles for the kernel's execution in TeraPool-SDR$_{1\text{-}3\text{-}5\text{-}9}$.

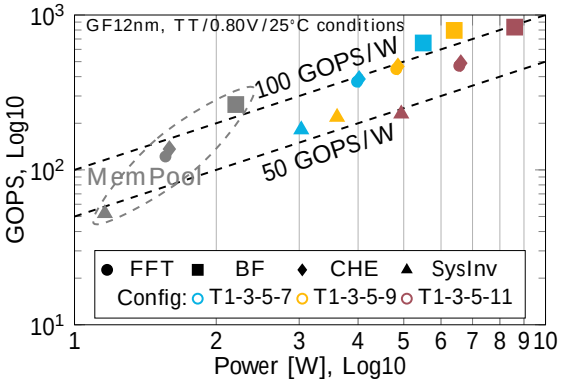

Fig. 6: TeraPool-SDR Energy Efficiency for SDR workloads, compared with a 4× smaller *MemPool* cluster.

ACKNOWLEDGMENT

Funded in part by COREnext supported by Horizon Europe program under grant agreement No. 101 092 598.

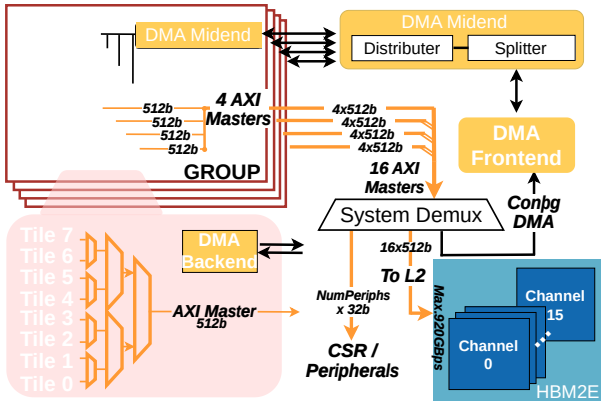

Fig. 3: System-level hierarchical AXI interconnection and modular DMA implementation.

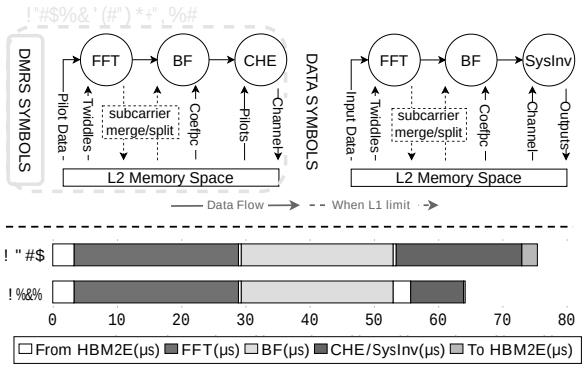

Fig. 4: Up: Key kernels' data flow of baseband receiving; Down: Data move vs. compute time for kernels in TeraPool-SDR$_{1\text{-}3\text{-}5\text{-}9}$.

TABLE I: State of Art Comparison

	This Work GF12	NVIDIA H100 TSMC4N [4]	Kalray MPPA3-80 TSMC16 [5]	Ramon RC64 TSMC65 [6]
Frequency (MHz)	(730-924)@(7-11)cycles	1755	600∼1200	300
Cores / Shared L1	1024 / 4MiB	128 / 256KB	16 / 4MB	64 / 4MB
Shared L1 latency cycles	1∼7/9/11	16.57	N.A.	N.A.
L1 Throughput	4KiB/Cyc	0.125KiB/Cyc	0.25KiB/Cyc	N.A.
Main Mem Latency	130ns	377.9ns	N.A.	N.A.
Main Mem Throughput	920GBPS	2TBPS	600GBPS	0.7GBPS
Power / Shared-L1 Cluster	8.8 W	6.14 W	4.9 W	5 W
Energy-Eff (INT32)	125GOPS/W	73.15GOPS/W	N.A.	15GOPS/W

REFERENCES

[1] L. Steiner *et al.*, "Dramsys4.0: An open-source simulation framework for in-depth dram analyses," *INT J PARALLEL PROG*, 2022.
[2] S. Riedel *et al.*, "Mempool: A scalable manycore architecture with a low-latency shared l1 memory," *IEEE Trans. Comput.*, 2023.
[3] M. Bertuletti *et al.*, "Efficient parallelization of 5G-PUSCH on a scalable RISC-V many-core processor," in *DATE Conference*, 2023.
[4] L. Kundu *et al.*, "Hardware acceleration for open radio access networks: A contemporary overview," *IEEE Commun. Mag.*, 2023.
[5] T. Yabe *et al.*, "Exploring the performance of deep neural networks on embedded many-core processors," in *ICCPS Conference*, 2022.
[6] R. Ginosar *et al.*, "Ramon space rc64-based ai/ml inference engine," in *European Workshop on OBDP*, 2021.