

A 410 GFLOP/s, 64 RISC-V Cores, 204.8 GBps Shared-Memory Cluster in 12 nm FinFET with Systolic Execution Support for Efficient B5G/6G AI-Enhanced O-RAN

Yichao Zhang* Marco Bertuletti* Sergio Mazzola* Samuel Riedel* Luca Benini*[†]

*IIS, ETH Zürich [†]DEI, University of Bologna

*{yiczhang, mbertuletti, smazzola, sriedel, lbenini}@iis.ee.ethz.ch

Abstract—We present HeartStream, a 64-RV-core shared-L1-memory cluster (410 GFLOP/s peak performance and 204.8 GBps L1 bandwidth) for energy-efficient AI-enhanced O-RAN. The cores and cluster architecture are customized for baseband processing, supporting complex (16-bit real&imaginary) instructions: multiply&accumulate, division&square-root, SIMD instructions, and hardware-managed systolic queues, improving up to 1.89× the energy efficiency of key baseband kernels. At 800 MHz@0.8 V, HeartStream delivers up to 243 GFLOP/s on complex-valued wireless workloads. Furthermore, the cores also support efficient AI processing on received data at up to 72 GOP/s. HeartStream is fully compatible with base station power and processing latency limits: it achieves leading-edge software-defined PUSCH efficiency (49.6 GFLOP/s/W) and consumes just 0.68 W (645 MHz@0.65 V), within the 4 ms end-to-end constraint for B5G/6G uplink.

Index Terms—6G, many-core, O-RAN, shared-memory, systolic

I. INTRODUCTION

The evolution of 5G Cloud Radio Access Networks (RAN) toward 6G Open-RAN (O-RAN) (Fig. 1) relies on open programmable hardware&software and distributed intelligence for the densification of network functions at the edge and the inter-operability of multi-vendor components [1], [2].

Beyond-5G (B5G) and 6G require >20 Gbps uplink throughput, and <4 ms end-to-end latency for high-end processing use-cases [3], [4]: Ultra-Reliable Low-Latency (URLL), massive Machine Type Communications (mMTC), and enhanced Mobile Broadband (eMBB). This corresponds to an increase in the compute density of base stations, the most latency-critical processing components of the RAN. Furthermore, the integration of Artificial Intelligence (AI) and communication will play a strategic role in the transition to 6G [5], improving performance in complex deployment scenarios, but significantly increasing the computational demands on base stations. Base station power is approaching thermal dissipation limits (10 kW), while performance must meet a 50× increase in uplink throughput by 2029 [6], [7]. Thus, baseband processing requires ever-increasing energy efficiency at high performance at the RAN edge, coupled with programmability and flexibility to track fast-evolving heterogeneous workloads.

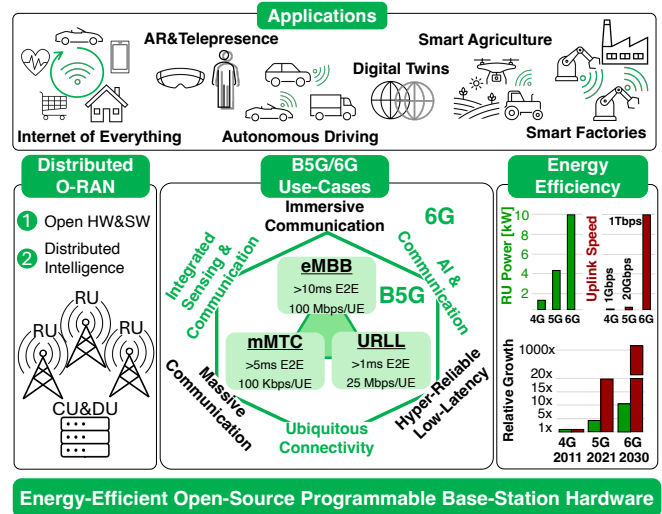


Fig. 1. Energy-efficient open-HW&SW designs support B5G/6G O-RAN demanding use-case scenarios and a wide range of applications.

We present HeartStream, an open-source*, RISC-V (RV) cluster for efficient, software-defined O-RAN. Our design demonstrates in silicon three innovations: (A) a fully programmable 64-core cluster, with 256×1-KiB banks of shared-L1-memory (204.8 GBps bandwidth) through a scalable, hierarchical, low-latency (1-5 cycles) interconnect; (B) efficient integer, floating-point (FP) (32/16/8-bit) and complex (16-bit real&imaginary) enhanced cores, delivering 410 GFLOP/s (800 MHz@0.8 V) peak-performance; and (C) interconnect and core extensions for systolic execution, boosting energy efficiency up to 213 GFLOP/s/W in baseband and deep learning data-parallel workloads (1.89× over non-systolic kernel baseline). HeartStream achieves up to 8.99 Gbps@0.8 V for Physical Uplink Shared Channel (PUSCH) computing. Even at the low-voltage, high-energy-efficiency corner (645 MHz@0.65 V), it meets the 4 ms end-to-end constraint for the baseband uplink of an 8×8 Multiple-Input, Multiple-Output (MIMO) transmission, with 32 antennas, 8 beams, 8 users, and 15 kHz sub-carrier (SC)-spacing on a 15 MHz-FR1 band.

*<https://github.com/pulp-platform/mempool>

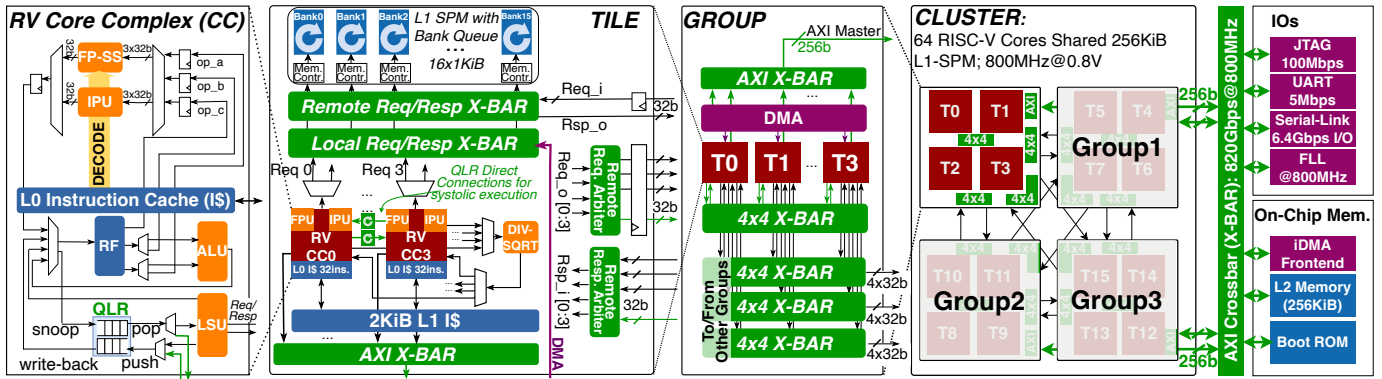


Fig. 2. HeartStream’s 64 RISC-V cores shared-L1-memory hierarchical design architecture. L1 memory addresses are 32-bit interleaved across banks of 16 Tiles in 4 Groups. Each Tile’s cores share an FP division/square-root unit. Core-Complex includes a 32b RISC-V core, IPU, FP-SS, and Systolic QLR.

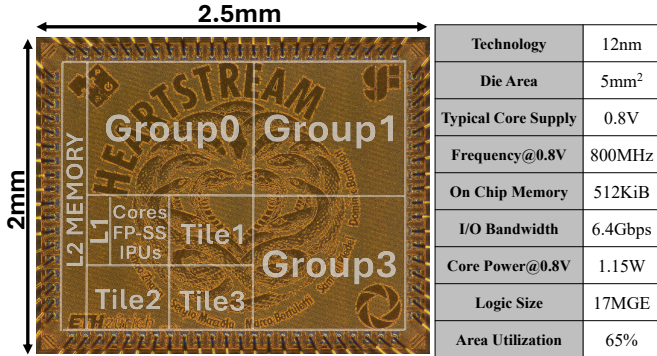


Fig. 3. Die micrograph and design summary. HeartStream was implemented in GlobalFoundries’ 12nm FinFET technology on a 5 mm² die. It achieves a 65% high utilization logic cell placement in the core area.

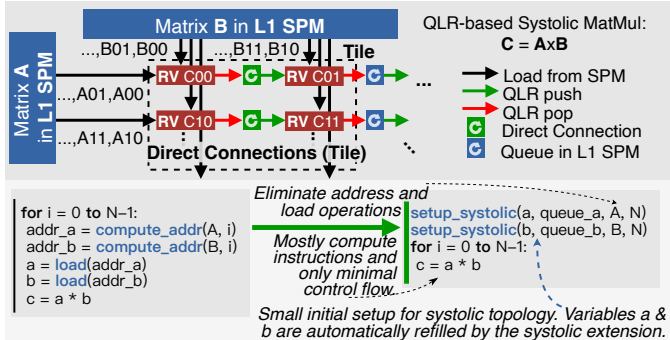


Fig. 4. In a systolic MatMul cores at the edge of the topology fetch from L1 and then forward data through QLR/memory queues, eliminating memory and control instructions; pseudocode shows that implicit inter-core communication eliminates many memory access and control instructions, boosting performance.

II. ARCHITECTURE

HeartStream’s shared-L1-memory architecture (Fig. 2) is inspired by MemPool [8], aiming to reduce buffering, data movement, and synchronization overhead in transferring large baseband data chunks across memory hierarchies. Fig. 3 presents the die shot and floorplan. It features 64 cores connected to 256 1-KiB L1-memory banks hierarchically (4 Groups of 4 Tiles each), resulting in a 1-5 cycles low-latency access. Each Tile contains 4 Core Complexes (CCs) combining a single-stage latency-tolerant 32-bit RV general-purpose (GP) programmable core with an Integer Processing Unit (IPU) and

a Floating Point Sub-System (FP-SS). For efficient software-defined O-RAN, both units support domain-specific instructions: Multiply&Accumulate (MAC), load-post-increment operations, SIMD operations, efficient complex arithmetic, widening sum-of-dot-product, and three-term addition instructions [9]. One Tile-shared FP division&square-root unit helps to accelerate matrix-inversion for MIMO detection. A key efficiency booster is hardware-supported flexible systolic execution with programmable topology [10] within the shared-memory cluster. Each core’s Queue-Linked Registers (QLRs) enable implicit inter-core register-file reads&writes. The QLRs have direct connections within a Tile. Across different Tiles, QLR access requests are memory-mapped and routed by the cluster crossbar interconnects. From the programming viewpoint, the QLRs can be configured at the beginning of the program execution (pseudocode shown in Fig. 4). After configuration, QLR access is fully hardware-managed: data is automatically pushed/popped to/from the corresponding register-linked systolic streams.

The execution of a Matrix Multiplication (MatMul) in systolic configuration is represented in Fig. 4. In this topology, cores exchange the data of input matrices via QLR connections and accumulate the outputs locally, reducing control-flow and memory-access overheads. Complex Fast Fourier Transform (CFFT) is also highly suitable for systolic execution. We adopt a Cooley-Turkey decimation-in-time algorithm and map the butterfly stages to different groups of cores. Cores pass each other the butterfly inputs/outputs without the need for a global inter-stage synchronization. Additionally, twiddle-coefficients and bit-reversal addresses are assigned statically to a core, drastically reducing memory access.

Each Tile has a 256-bit AXI port into a hierarchical interconnect to 256 KiB of L2 memory, peripherals, and off-chip access. A custom-designed DMA handles transfers between the physically-distributed L1 and L2 or off-chip memory. The 16-channel DDR Serial Link delivers 6.4 Gbps throughput.

III. RESULTS

HeartStream is designed for energy-efficient software-defined baseband processing. Moreover, its fully programmable cores can also execute data-parallel AI workloads on the received data streams. This architectural flexibility plays a strategic role in

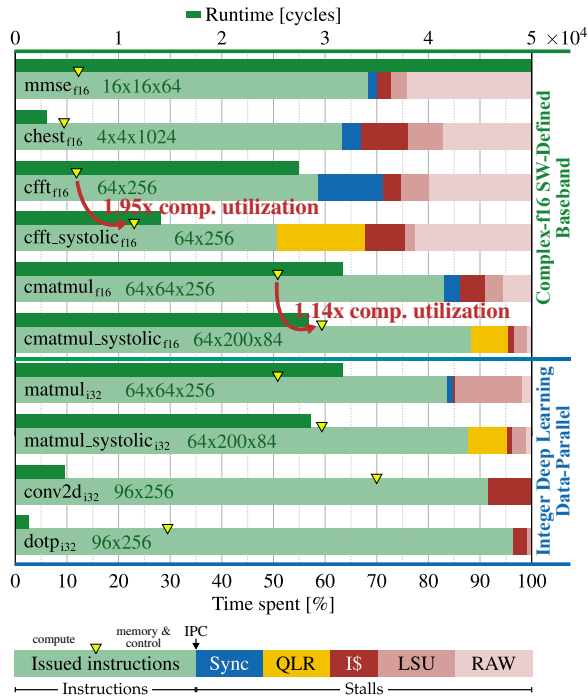


Fig. 5. Absolute runtime and instruction/stall cycle fractions for 16-bit complex (real&imaginary) baseband and 32-bit integer deep learning kernels. Systolic kernels achieve higher compute utilization and performance by reducing overhead instructions.

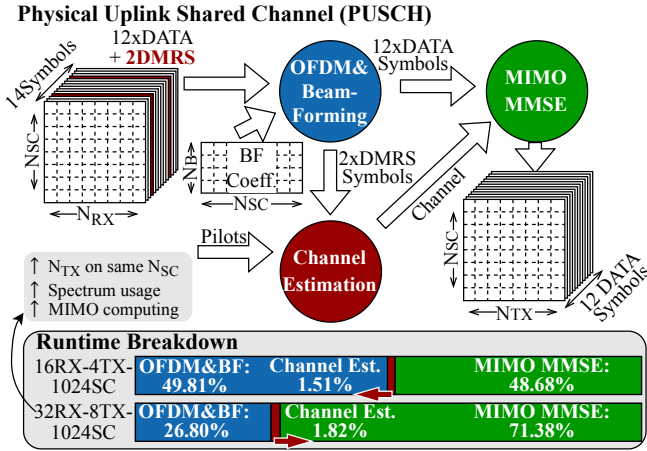


Fig. 6. The baseband PUSCH processing steps for a transition time interval with 14 symbols, 1024 Sub-Carriers (SC) in 15kHz spacing, and compute runtimes breakdown for different $N_{RX} \times N_{TX}$ scenarios.

enabling the convergence between wireless and AI workloads planned by next-generation RAN systems. Complex-valued baseband workloads achieve up to 243 GFLOP/s@0.8 V with instructions-per-cycles (IPCs) of 0.52-0.88 (Fig. 5). On typical deep learning integer benchmarks (MatMul, 2D-Convolution (Conv2D), and Dot Product (DOTP), with the largest input size fitting in the cluster L1 memory), HeartStream achieves 0.84-0.96 IPC and up to 72 GOP/s@0.8 V.

PUSCH is one of the most compute- and time-critical channels of baseband processing. The uplink steps are represented in Fig. 6: in <4 ms end-to-end latency, 14 symbols, each consisting of a N_{RX} antennas \times N_{SC} subcarriers matrix

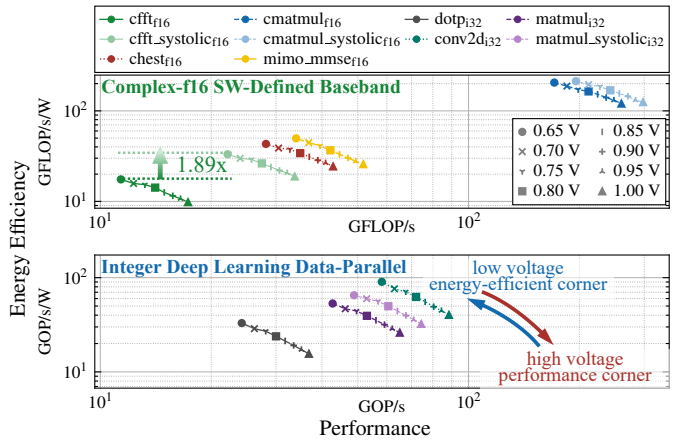


Fig. 7. HeartStream's efficiency and performance on key kernels for baseband and deep learning processing. The different core supply voltages target high energy efficiency or high performance. The systolic extension improves energy efficiency up to 1.89x.

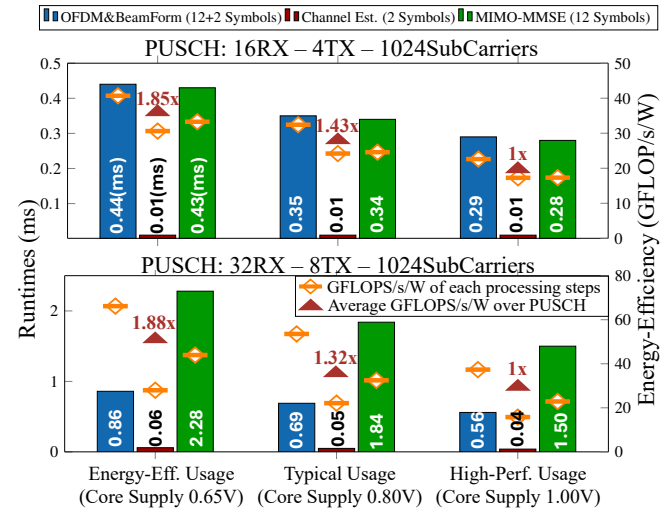


Fig. 8. Runtime and energy efficiency breakdown of PUSCH processing steps: low-voltage for low $N_{RX} \times N_{TX}$ achieves energy efficiency, and high-voltage for high $N_{RX} \times N_{TX}$ achieves high performance.

of complex numbers, undergo CFFT and CMatMul with known beamforming coefficients. Two Demodulation Reference Symbol (DMRS) symbols are used to estimate the transmission channel, and 12 data symbols are fed to the Minimum Mean Squared Error (MMSE) equalization, resulting in a detected complex number for each one of the N_{TX} transmitters sending data over a subcarrier. In OFDM&beamforming (CFFT&CMatMul), systolic extensions allow 50% and 12% runtime reduction, respectively. Pre-configuring a kernel-specific systolic computation reduces control and inter-core synchronization overhead in baseband and integer deep learning kernels. The low-voltage operation (645 MHz@0.65 V) allows energy-efficient processing (Fig. 7): 33.2 GFLOP/s/W OFDM and 213 GFLOP/s/W beamforming, improving up to 1.89x thanks to systolic extensions. Even for 16×16 MIMO, energy-efficiency gains +13 GFLOP/s/W, with only +0.25 ms/symbol in runtime compared to 800 MHz@0.8 V.

Two scenarios (Fig. 8) show different runtime distribution

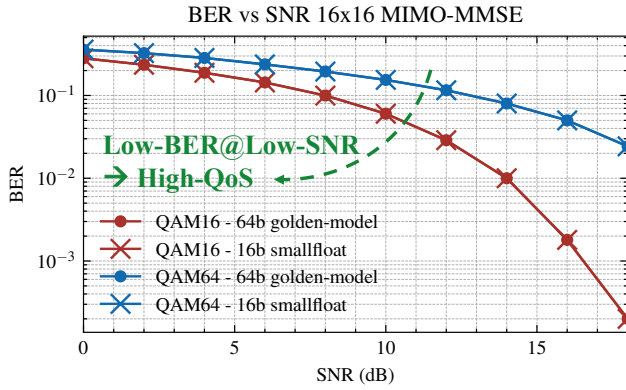


Fig. 9. BER vs. SNR of a 16x16 MIMO MMSE (AWGN channel), implemented with mixed-precision 16/32-bit floating-point extensions, yields the same results as the 64b golden model. Lower SNR at a given BER indicates higher Quality of Service (QoS).

over PUSCH steps: a 4×4 or 8×8 MIMO transmission, with $N_{RX}=16\text{-}32$ antennas, $N_B=4\text{-}8$ active beams, $N_{TX}=4\text{-}8$ transmitters, and 15kHz SC-spacing on a 15MHz-FR1 band. HeartStream achieves an in-phase and quadrature antenna PUSCH computing of up to 8.99 Gbps@0.8 V. In both scenarios, low-voltage operation increases energy efficiency up to 49.6 GFLOP/s/W (1.88 \times) and incurs only 1.1 ms slowdown in runtime, keeping the latency <4 ms (3.2 ms). The compute budget available in the 4×4 MIMO can be invested in additional post-processing: all the presented deep learning benchmarks, for the largest problem size that fits in the L1, achieve high 37-89 GOP/s@1 V and latencies far below the runtime constraints (1-32 μ s). Widening sum-of-dot-product of *xsmallfloat* extensions is essential to keep complex arithmetic precision high in the matrix inversion operations implemented by MIMO-MMSE, yielding 16.5dB SNR at BER@ 10^{-3} dB (Fig. 9) for a 16×16 QAM16 MMSE problem.

TABLE I
COMPARISON WITH STATE-OF-THE-ART BASEBAND PROCESSING DESIGN

	This Work ^a	ESSERC ²⁴ [11] ^c	CoolChip ²² [12] ^d	ISSCC ¹⁴ [13] ^e	ESSCIRC ²² [14] ^e	ISSCC ²⁴ [15] ^b	VLSI ²⁰ [16] ^b
Technology	12nm FinFET	22nm FDX	28nm	65nm	22nm FDX	40nm	40nm
Freq.(MHz)	800@0.8V	200-400@0.8V	800@0.9V	445@1.2V	293@0.8V	200@1.1V	290@1.1V
I/O Gbps	6.4	-	-	10	-	9.6	1.96
Processing Element	64 RISC-V cores	1 RISC-V & 1 Vector, 1 Systolic, 1 Accel.	1 RISC & 2 ARC, 4 ASIPs, 2 Accel.	8 RISC &, 8 Vector, 4 ASIPs	16 PEs Program. ASIP.	ASIC Accel.	ASIC Accel.
Data Precision	Int32/16 FP 32/16/8	-	Int32/-, -	Int16/-, FP32/-	FP32/16/8 bfloat16	-	-
GP-Program.	Yes	No	No	No	No	No	No
Execution	SPMD, SIMD, Systolic	SIMD, Systolic	VLIW, SIMD	SIMD, ASIP-MIMO	ASIP-MIMO	-	-
Baseband Processing	Full B5G/6G SW-Defined O-RAN ^c	Partial RAN: MIMO Detector	Partial RAN: MIMO Detector	Partial RAN: LTE MIMO, WiMAX	MU-MIMO Det./Dec.	MIMO Det./Dec, Channel Est.	MIMO Detector
Peak Perf.	410 GFLOP/s	-	-	3.6 GFLOP/s	37.5 GFLOP/s	-	-
MIMO Workload	Flex. Size ^d	16x16 QAM64	4x4 QAM16	4x4 QAM64	Flex. Size ^d	8x8 QAM16	256x32 QAM256
PUSCH Comp. Gbps	8.99 ^e	5.58 ^f	7.30 ^f	1.73 ^f	1.76 ^f	32.77 ^f	26.13 ^f
PUSCH Comp.Gbps/W	8.26 ^e	159.93 ^f	7.18 ^f	40.42 ^f	33.26 ^f	453 ^f	1893 ^f
Deep-Learn. Workloads	45.2 ^g GOP/s/W	-	-	-	-	-	-

^{a/b} Program/Non-Programmable solution. ^c Open HW&SW lower-PHY processing chain. ^d Software-Defined MIMO supports flexible No.TX/RX. ^e In-phase and quadrature antenna computing divided by runtime, 800MHz@0.8V. ^f MIMO processed by accelerator or specialized datapaths ASIP; technology normalized to 12nm and 0.8V core supply. ^g The energy efficiency of deep learning workload (Conv2D example).

In Table I, we compare HeartStream to Application Specific Integrated Circuits (ASICs) & Application Specific Instruction Processors (ASIPs) for baseband processing. HeartStream is the first open-source, RV-based, and fully programmable O-RAN processor. It delivers the highest peak performance (GFLOP/s), and it achieves competitive throughput and energy efficiency compared to partially programmable designs with inflexible datapaths tailored for MIMO decoding [11]–[14], while offering much greater flexibility to support multiple network scenarios with respect to fixed-function accelerators [15], [16]. Further, it supports diverse MIMO-sizes, a full B5G/6G uplink, and deep learning operators for the convergence between wireless and AI processing in 6G AI-native RANs.

ACKNOWLEDGMENT

This work has received funding from the Swiss State Secretariat for Education, Research, and Innovation (SERI) under the SwissChips initiative.

REFERENCES

- [1] L. Gavrilovska, V. Rakovic, and D. Denkovski, "From Cloud RAN to Open RAN," *Wirel. Pers. Commun.*, vol. 113, no. 3, pp. 1523–1539, Aug. 2020.
- [2] S. Marinova and A. Leon-Garcia, "Intelligent O-RAN Beyond 5G: Architecture, Use Cases, Challenges, and Opportunities," *IEEE Access*, vol. 12, pp. 27 088–27 114, Feb. 2024.
- [3] R. Kumar, D. Sinwar, and V. Singh, "QoS aware resource allocation for coexistence mechanisms between eMBB and URLLC: Issues, challenges, and future directions in 5G," *Comp. Commu.*, vol. 213, pp. 208–235, Jan. 2024.
- [4] ITU, "Recommendation ITU-R M.2160-0: Framework and overall objectives of the future development of IMT for 2030 and beyond," International Telecommunication Union, Tech. Rep. M.2160-0, Nov. 2023.
- [5] F. A. Aoudia and J. Hoydis, "End-to-end learning for OFDM: From neural receivers to pilotless communication," *IEEE TWC*, vol. 21, no. 2, pp. 1049–1063, 2021.
- [6] C.-L. I, S. Han, and S. Bian, "Energy-efficient 5G for a Greener Future," *Nat. Electron.*, vol. 3, pp. 182–184, April 2020.
- [7] Ericsson, "Ericsson Mobility Report," Tech. Rep., 2021 & 2024.
- [8] S. Riedel, M. Cavalcante, R. Andri, and L. Benini, "MemPool: A Scalable Manycore Architecture With a Low-Latency Shared L1 Memory," *IEEE TCOMP*, vol. 72, no. 12, pp. 3561–3575, Aug. 2023.
- [9] L. Bertaccini *et al.*, "MiniFloats on RISC-V Cores: ISA Extensions With Mixed-Precision Short Dot Products," *IEEE TETC*, vol. 12, no. 4, pp. 1040–1055, Feb. 2024.
- [10] S. Mazzola, S. Riedel, and L. Benini, "Enabling Efficient Hybrid Systolic Computation in Shared-L1-Memory Manycore Clusters," *IEEE TVLSI*, vol. 32, no. 9, pp. 1602–1615, Sept. 2024.
- [11] M. Attari, J. R. Sánchez, O. Edfors, and L. Liu, "A 1095 pJ/b 219 Mb/s Application-specific Instruction-set Processor for Distributed Massive MIMO in 22FDX," in *2024 IEEE ESSERC*, Sept. 2024, pp. 257–260.
- [12] Y. Chen, L. Liu, X. Feng, and J. Shi, "DXT501: An SDR-Based Baseband MP-SoC for Multi-Protocol Industrial Wireless Communication," in *2022 IEEE COOL CHIPS*. Los Alamitos, CA, USA: IEEE Computer Society, April 2022, pp. 1–6.
- [13] B. Noethen *et al.*, "A 105GOPS 36mm2 heterogeneous SDR MPSoC with energy-aware dynamic scheduling and iterative detection-decoding for 4G in 65nm CMOS," in *2014 IEEE ISSCC*, Feb. 2014, pp. 188–189.
- [14] O. Castañeda, L. Benini, and C. Studer, "A 283 pJ/b 240 Mb/s Floating-Point Baseband Accelerator for Massive MU-MIMO in 22FDX," in *2022 IEEE ESSCIRC*, Sept. 2022, pp. 357–360.
- [15] Y. Zhang *et al.*, "BayesBB: A 9.6Gbps 1.61ms Configurable All-Message Passing Baseband-Accelerator for B5G/6G Cell-Free Massive-MIMO in 40nm CMOS," in *2024 IEEE ISSCC*, vol. 67, Feb. 2024, pp. 48–50.
- [16] C.-C. Wen *et al.*, "A 1.96 Gb/s Massive MU-MIMO Detector for Next-Generation Cellular Systems," in *2020 IEEE Symp. VLSI*, June 2020, pp. 1–2.