Savannah: A Real-time Programmable mmWave Baseband Processing Framework

Zhenzhou Qi[*]
Duke University

Chung-Hsuan Tung[*]
Duke University

Anuj Kalia
Microsoft

Tingjun Chen
Duke University

ABSTRACT

5G new radio (NR) frequency range 2 (FR2) in the millimeter-wave (mmWave) band has a much shorter baseband processing deadline compared to that in the sub-7 GHz FR1 band. This tight deadline requires an efficient real-time system for baseband processing using minimal computational resources. We demonstrate Savannah, a software framework for efficient mmWave baseband processing using minimal and heterogeneous computing resources, including CPU and eASIC. Savannah vectorizes matrix operations and memory access patterns in multi-input multi-output (MIMO) arithmetic, offloads low-density parity-check (LDPC) coding to an eASIC, and enables single-core operation. We demonstrate that Savannah, using a single CPU core and an eASIC, can support a 2×2 MIMO link with 100 MHz bandwidth under full uplink traffic load, yielding a data rate of up to 487 Mbps.

CCS CONCEPTS

• **Networks** → **Wireless access points, base stations and infrastructure**; **Network performance analysis**.

KEYWORDS

Millimeter-wave, Baseband processing, ACC100 accelerator

ACM Reference Format:
Zhenzhou Qi, Chung-Hsuan Tung, Anuj Kalia, and Tingjun Chen. 2024. Savannah: A Real-time Programmable mmWave Baseband Processing Framework. In *The 30th Annual International Conference on Mobile Computing and Networking (ACM MobiCom '24), November 18–22, 2024, Washington D.C., DC, USA.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3636534.3698843

1 INTRODUCTION

Millimeter-wave (mmWave) transmission band in 5G frequency range 2 (FR2) [4] has been actively studied [9, 12, 14, 20, 21] and shown the potential of supporting higher data rates leveraging the larger available bandwidth compared to the sub-7 GHz band (5G FR1) [24]. Meanwhile, the virtualized radio access network (vRAN) emerged for fast deployment, efficient resource allocation, and interoperability [27]. However, the research community lacks an open-sourced baseband processing framework for vRAN that can handle 5G FR2 traffic in real time. We present Savannah [19] – a solution that supports up to 2×2 MIMO in an FR2 link with 100 MHz bandwidth using a single CPU core and an ACC100 accelerator [25] for low-density parity check (LDPC) decoding. Savannah vectorizes matrix operations with SIMD instructions (e.g., AVX-512 [13]) to accelerate MIMO DSP stages (e.g., precoder calculation and equalization [19, Fig. 2]) and adopts the ACC100 accelerator to reduce the required CPU core counts for energy efficiency and cost-effectiveness.

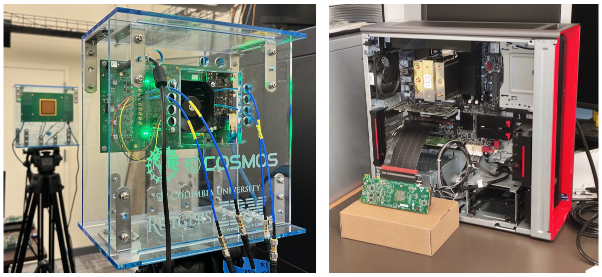

(a) IBM 28GHz PAAM [23]
(b) Lenovo P5 with ACC100

Figure 1: (a) PAWR COSMOS testbed with FR2 front ends, (b) a local setup with a workstation.

We will demonstrate Savannah on both commodity servers and edge workstations. We present the real-time quality of transmission (QoT) using the PAWR COSMOS testbed [22] with 28 GHz front ends [23] connected to servers. We also show Savannah's adaptability to be deployed on a workstation with an Intel ACC100 accelerator card. The codebase for Savannah is open-sourced [1], and the experimentation using COSMOS testbed is publicly available as a tutorial [2].

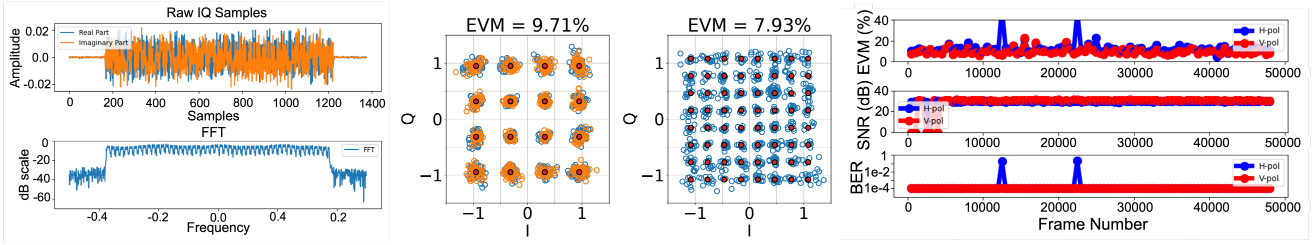
2 DEMO SETUP

Savannah supports two configurations for flexibility and broad applicability. Savannah-mc, designed to operate without the ACC100 accelerator but at the cost of requiring additional CPU cores, applies Intel's FlexRAN SDK [6] for LDPC

This work is licensed under a Creative Commons Attribution International 4.0 License.

ACM MobiCom '24, November 18–22, 2024, Washington D.C., DC, USA
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0489-5/24/11
https://doi.org/10.1145/3636534.3698843

[*] Both authors contributed equally to this work.



(a) IQ samples and FFT results snapshot. (b) Constellation diagrams with EVM. (c) EVM, SNR and BER over time.

Figure 2: Examples of the real-time visualization of FR2 links with 100 MHz bandwidth. (a) Received baseband signal in the time and frequency domains. (b) Constellation diagrams of 16QAM in 2×2 MIMO and 64QAM in SISO. (c) Per-frame EVM, SNR and BER in the 2×2 MIMO, 16 QAM link.

decoding and operates in a multi-core setting. Savannah-sc incorporates Intel's ACC100 accelerator and runs with a single CPU core. The minimum hardware requirement of the base station in Savannah-mc is a workstation whose CPUs support AVX-512 [13], and Savannah-sc requires one additional ACC100 accelerator via the PCIe port. The radio unit (RU) can be any UHD-based [17] software-defined radio (SDR) if not emulated. We select two scenarios to show Savannah's capability and portability: (*i*) Savannah-mc on the COSMOS testbed with FR2 front ends and (*ii*) Savannah-sc on a local workstation with minimum hardware resources.

COSMOS testbed with FR2 front ends. We leverage the open-access PAWR COSMOS testbed [5] to demonstrate Savannah with FR2 front ends. In particular, we use the COSMOS Sandbox 2 (sb2) [18], which includes two IBM 28 GHz phased array antenna module (PAAM) boards [23] (as shown in Fig. 1(a)) and two USRP N310 SDRs. Both the 28 GHz front ends and SDRs are controlled by a Dell PowerEdge R740 server with a 48-core Intel Xeon Gold 6226 CPU @2.7 GHz.

Local setup with a workstation. As shown in Fig. 1(b), we deploy Savannah-sc on a Lenovo TS P5 workstation with an 8-core Intel Xeon W3-2435 CPU and a DPDK [26] controlled Silicom's Lisbon P2 ACC100 card [19, Sec. 6]. Due to the lack of portable mmWave front ends, we set up a sub-6 GHz wireless link between two USRP X310 SDRs and employ Physical layer (PHY) parameters that reflects FR2 settings.

5G NR FR2 PHY configurations. We focus on FR2 numerology 3 ($\mu = 3$) with 120 KHz subcarrier spacing (SCS) and 0.125 ms slot duration. Every five slots are arranged in a TDD format of `DDDSU` [8, 10, 15, 16]. Allowing two slot durations for MAC scheduling [11], we set the PHY processing deadline to be three slots, i.e., 0.375 ms. We consider 100 MHz channel bandwidth and set a fast Fourier transform (FFT) size of 1,024 (equals to the number of subcarriers), out of which 792 are allocated as data subcarriers. We select modulation and coding schemes (MCS) 17 (i.e., 64 QAM and a code rate of 438/1,024) [3], as a higher modulation order and smaller code rate imply heavier baseband processing workload [7].

3 REAL-TIME PROCESSING DEMO

Experiment 1: Peak real-time baseband processing. We will showcase the peak performance of our system in a 2×2, 100 MHz link, using local workstations with an ACC100 accelerator with emulated RUs. We apply MCS17 which translates to a data rate of 487 Mbps. In this case, the system is dedicated to real-time processing, and all non-critical functionalities are disabled. Log recording for the link quality metrics and timestamps is assigned to separate threads. The log includes the processing time of each DSP stage in a frame, bit error rate (BER), and block error rate (BLER). The log, in the standard output, is kept minimal for the system's performance.

Experiment 2: Over-the-air (OTA) transmission with visualization. We develop Python scripts to visualize the DSP results for each stage simultaneously with real-time processing of OTA transmissions through a customized graphical user interface (GUI). Savannah saves the frame buffers for each DSP stage to a `.bin` file, including the raw I/Q samples and FFT results (see Fig. 2(a)), and constellation diagrams after equalization (see Fig. 2(b)). Similar to Agora [7], Savannah saves error vector magnitude (EVM), signal-to-noise ratio (SNR), and BER for each frame into CSV files (see Fig. 2(c)). Dedicated Python scripts repeatedly read from the files and update the corresponding plots in the GUI. This demonstration will be showcased using both demo setups.

ACKNOWLEDGMENTS

This work was supported in part by NSF grants CNS-1827923, CNS-2211944, AST-2232458, and an IBM Academic Award. This work was also supported in part by the Center for Ubiquitous Connectivity (CUbiC), sponsored by Semiconductor Research Corporation (SRC) and Defense Advanced Research Projects Agency (DARPA) under the JUMP 2.0 program. We thank Arun Paidimarri, Bodhisatwa Sadhu, and Alberto Valdes-Garcia (IBM Research), Rahman Doost-Mohammady and Andrew Sedlmayr (Rice University), and Ivan Seskar (Rutgers University) for their contributions to this work.

REFERENCES

[1] 2024. Savannah: Efficient mmWave Baseband Processing with Minimal and Heterogeneous Resources. https://github.com/functions-lab/Savannah.

[2] 2024. Savannah: Efficient mmWave Baseband Processing with Minimal and Heterogeneous Resources. https://wiki.cosmos-lab.org/wiki/Tutorials/Wireless/mmwavePaamRealTimePHY.

[3] 3GPP. 2018. *5G; NR; Physical Layer Procedures for Data.* Technical Specification (TS) 38.214. 3rd Generation Partnership Project (3GPP). https://www.etsi.org/deliver/etsi_ts/138200_138299/138214/15.02.00_60/ts_138214v150200p.pdf Version 15.2.0.

[4] 3GPP. 2022. *5G; NR; User Equipment (UE) radio transmission and reception; Part 2: Range 2 Standalone.* Technical Specification (TS) 38.101-2. 3rd Generation Partnership Project (3GPP). https://www.etsi.org/deliver/etsi_ts/138100_138199/13810102/17.06.00_60/ts_13810102v170600p.pdf Version 17.6.0.

[5] Tingjun Chen, Prasanthi Maddala, Panagiotis Skrimponis, Jakub Kolodziejski, Abhishek Adhikari, Hang Hu, Zhihui Gao, Arun Paidimarri, Alberto Valdes-Garcia, Myung Lee, et al. 2023. Open-access millimeter-wave software-defined radios in the PAWR COSMOS testbed: Design, deployment, and experimentation. *Computer Networks* 234 (2023), 109922.

[6] Intel Corporation. 2019. FlexRAN LTE and 5G NR FEC software development kit modules. https://www.intel.com/content/www/us/en/developer/articles/technical/flexran-lte-and-5g-nr-fec-software-development-kit-modules.html.

[7] Jian Ding, Rahman Doost-Mohammady, Anuj Kalia, and Lin Zhong. 2020. Agora: Real-time massive MIMO baseband processing in software. In *Proc. ACM CoNEXT '20*.

[8] Rostand Fezeu, Jason Carpenter, Claudio Fiandrino, Eman Ramadan, Wei Ye, Joerg Widmer, Feng Qian, and Zhi-Li Zhang. 2023. Mid-band 5G: A measurement study in Europe and US. *arXiv preprint arXiv:2310.11000* (2023).

[9] Zhihui Gao, Zhenzhou Qi, and Tingjun Chen. 2024. Mambas: Maneuvering analog multi-user beamforming using an array of subarrays in mmWave networks. In *Proc. ACM MobiCom '24*.

[10] Global System for Mobile Communications. 2020. 5G TDD synchronisation guidelines and recommendations for the coexistence of TDD networks in the 3.5 GHz range. https://www.gsma.com/spectrum/wp-content/uploads/2020/04/3.5-GHz-5G-TDD-Synchronisation.pdf.

[11] Junzhi Gong, Anuj Kalia, and Minlan Yu. 2023. Scalable distributed massive MIMO baseband processing. In *Proc. USENIX NSDI '23*.

[12] David Hunt, Kristen Angell, Zhenzhou Qi, Tingjun Chen, and Miroslav Pajic. 2024. MadRadar: A Black-Box physical layer attack framework on mmWave automotive FMCW radars. In *Proc. ISOC NDSS '24*.

[13] Intel. 2023. Intel ® intrinsics guide. https://www.intel.com/content/www/us/en/docs/intrinsics-guide/index.html.

[14] Ish Kumar Jain, Raghav Subbaraman, Tejas Harekrishna Sadarahalli, Xiangwei Shao, Hou-Wei Lin, and Dinesh Bharadia. 2020. mMobile: Building a mmWave testbed to evaluate and address mobility effects. In *Proc. ACM mmNets '20*.

[15] Nikita Lazarev, Tao Ji, Anuj Kalia, Daehyeok Kim, Ilias Marinos, Francis Y. Yan, Christina Delimitrou, Zhiru Zhang, and Aditya Akella. 2023. Resilient baseband processing in virtualized RANs with slingshot. In *Proc. ACM SIGCOMM '23*.

[16] Biraja Mahapatra. 2022. Intel + Radisys FlexRAN success stories. https://networkbuilders-cdn.s3.us-east-2.amazonaws.com/Intel_Radisys_FlexRAN_Success_Stories_Biraja_Prasad_Mahapatra.pdf.

[17] National Instruments. 2024. UHD (USRP Hardware Driver). https://www.ettus.com/sdr-software/uhd-usrp-hardware-driver/.

[18] Zhenzhou Qi, Zhihui Gao, Chung-Hsuan Tung, and Tingjun Chen. 2023. Programmable Millimeter-Wave MIMO Radios with Real-Time Baseband Processing. In *Proc ACM WiNTECH '23*.

[19] Zhenzhou Qi, Chung-Hsuan Tung, Anuj Kalia, and Tingjun Chen. 2024. Savannah: Efficient mmWave Baseband Processing with Minimal and Heterogeneous Resources. In *Proc. ACM MobiCom '24*.

[20] Sundeep Rangan, Theodore S. Rappaport, and Elza Erkip. 2014. Millimeter-wave cellular wireless networks: Potentials and challenges. *Proc. of the IEEE* 102, 3 (2014), 366–385.

[21] Theodore S Rappaport, Robert W Heath Jr, Robert C Daniels, and James N Murdock. 2015. *Millimeter wave wireless communications*. Pearson Education.

[22] Dipankar Raychaudhuri, Ivan Seskar, Gil Zussman, Thanasis Korakis, Dan Kilper, Tingjun Chen, Jakub Kolodziejski, Michael Sherman, Zoran Kostic, Xiaoxiong Gu, et al. 2020. Challenge: COSMOS: A city-scale programmable testbed for experimentation with advanced wireless. In *Proc. ACM MobiCom '20*.

[23] Bodhisatwa Sadhu, Yahya Tousi, Joakim Hallin, Stefan Sahl, Scott K Reynolds, Örjan Renström, Kristoffer Sjögren, Olov Haapalahti, Nadav Mazor, Bo Bokinge, et al. 2017. A 28-GHz 32-element TRX phased-array IC with concurrent dual-polarized operation and orthogonal phase and gain control for 5G communications. *IEEE J. Solid-State Circuits* 52, 12 (2017), 3373–3391.

[24] Clayton Shepard, Hang Yu, Narendra Anand, Erran Li, Thomas Marzetta, Richard Yang, and Lin Zhong. 2012. Argos: Practical many-antenna base stations. In *Proc. ACM MobiCom '12*.

[25] Silicom Ltd. 2023. Lisbon P2 ACC100 FEC accelerator extended temp. https://www.silicom-usa.com/wp-content/uploads/2023/09/Lisbon-P2-ACC100-FEC-Accelerator-Extended-temp-1v1.pdf.

[26] The Linux Foundation. 2023. Data plane development kit (DPDK). https://doc.dpdk.org/guides-21.11/tools/testbbdev.html.

[27] Qing Yang, Xiaoxiao Li, Hongyi Yao, Ji Fang, Kun Tan, Wenjun Hu, Jiansong Zhang, and Yongguang Zhang. 2013. BigStation: Enabling scalable real-time signal processing large MU-MIMO systems. *ACM SIGCOMM Computer Communication Review* 43, 4 (2013), 399–410.