Scalable and Real-time Baseband Processing using Heterogeneous Compute Resources

Tingjun Chen

Nortel Networks Assistant Professor

Departments of Electrical & Computer Engineering and Computer Science

Duke University

First OAI Foundation U.S. Hands-on Workshop

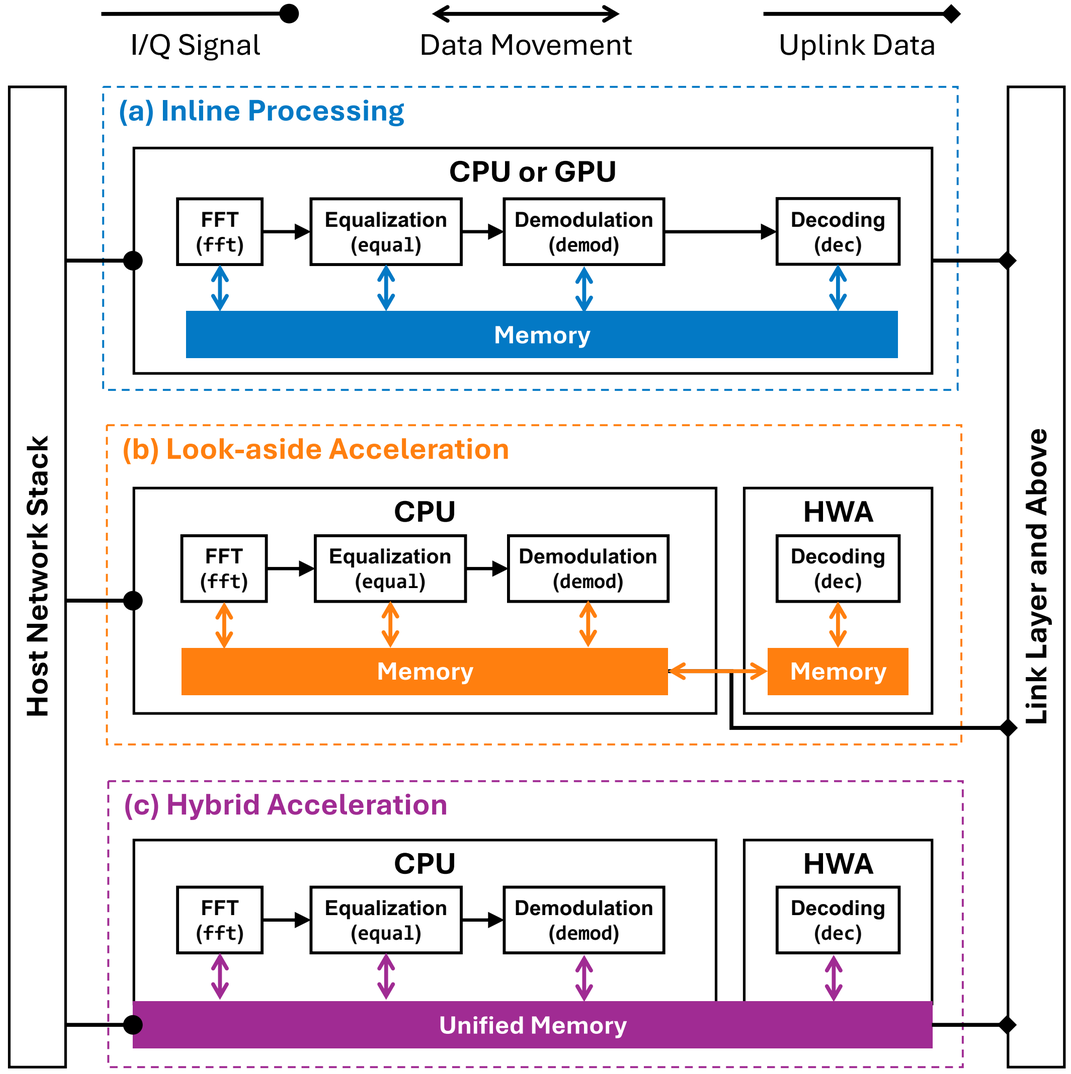
November 17, 2025

Duke ENGINEERING

FuNCtions Lab

L1 (Baseband) Processing Models for vRANs

- Leveraging the powerful **hardware accelerators (HWAs)** for L1 processing in virtualized radio access networks (vRANs)
- **Cloud-native scaling** favors flexibility, software-centric upgrades, and multi-tenancy supports
- **Power-latency tradeoffs** determine the overall system efficiency, but also remains difficult to model
- Requires both significant R&D and engineering efforts!



FlexRAN Aerial

Intel ACC100 eASIC

NVIDIA Jetson Orin AGX

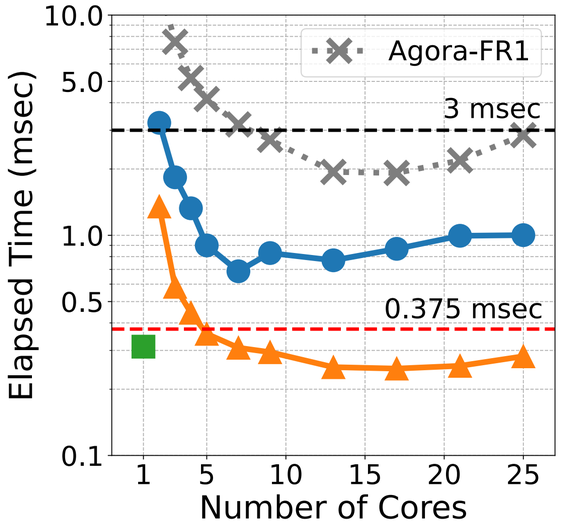
- L. Kundu, X. Lin, E. Agostini, V. Ditya, and T. Martin, "Hardware acceleration for open radio access networks: A contemporary overview," *IEEE Communications Magazine*, vol. 62, no. 9, 2023.
- Z. Qi, Y. Yao, Y. Li, C.-H. Tung, J. Zheng, D. Zhuo, and T. Chen, "DecodeX: Exploring and benchmarking of LDPC decoding across CPU, GPU, and ASIC platforms," *arXiv:2511.02952 [cs.NI]*, Nov. 2025.

Duke ENGINEERING FuNCtions Lab

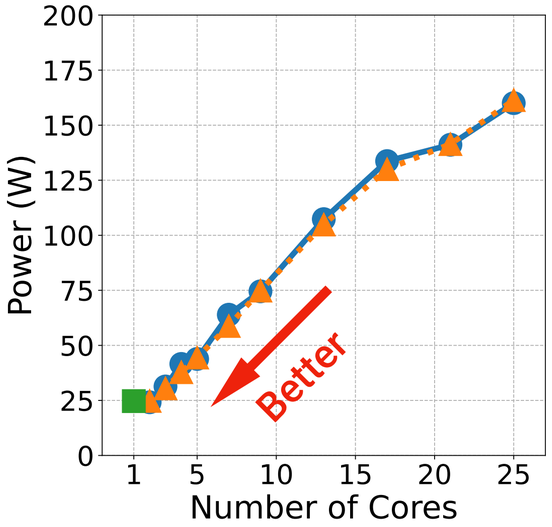
Savannah: Real-time Efficient mmWave Baseband Processing

- Existing software-based PHY processing frameworks are designed for FR1, while FR2 has much tighter PHY processing latency requirements due to larger subcarrier spacing

99.9th-percentile Elapsed Time



Power Consumption



*Same total fronthaul traffic for FR1/FR2

Agora [*ACM CoNEXT'20*]

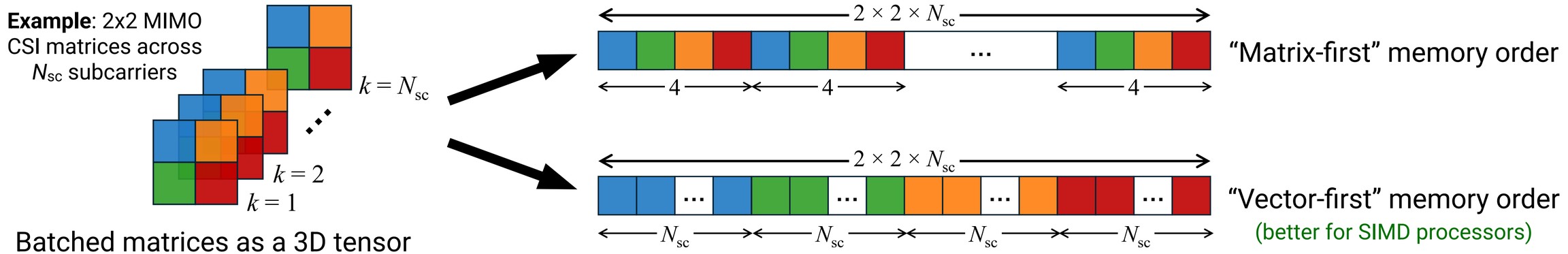
Agora-FR1	CPU only (multi-core)
Agora	CPU only (multi-core)
Savannah-mc	CPU only (multi-core)
Savannah-sc	CPU (single-core) + ASIC (ACC100)

- 2×2 MIMO link with 100 MHz bandwidth, 100% uplink traffic load
- FR2 (μ = 3) with 0.125 ms slot, FR1 (μ = 0) with 1.0 ms
- PHY processing deadline set to 3-slot (frame schedule: `DDDSU`)

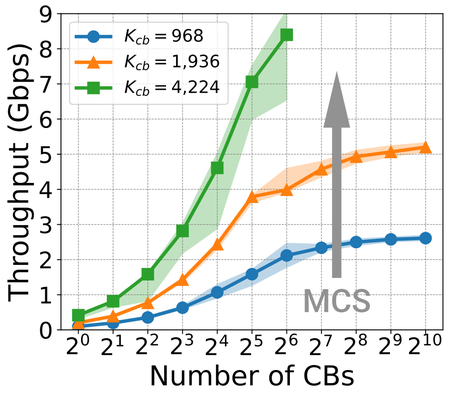
- Z. Qi*, C.-H. Tung*, A. Kalia, and T. Chen, "Savannah: Efficient mmWave baseband processing with minimal and heterogeneous resources," in *Proc. ACM International Conference on Mobile Computing and Networking (MobiCom'24)*, 2024.
- Code and data: https://github.com/functions-lab/Savannah

Savannah: Real-time Efficient mmWave Baseband Processing

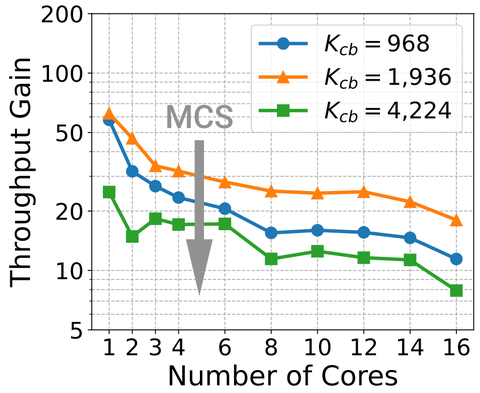
- **Key insight #1**: Vectorizing small matrix operations with optimized memory layout



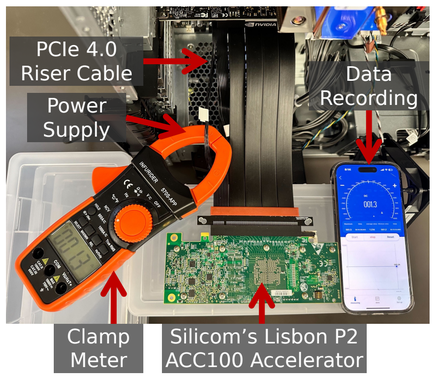
- **Key insight #2**: Leveraging heterogeneous compute resources (e.g., CPU + ASIC)

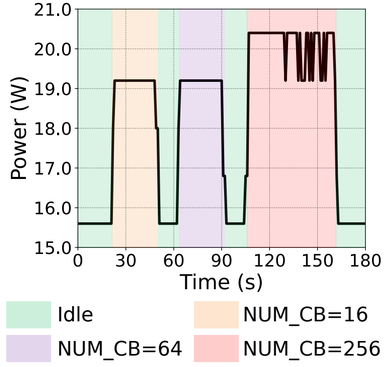


ASIC "parallelism" →

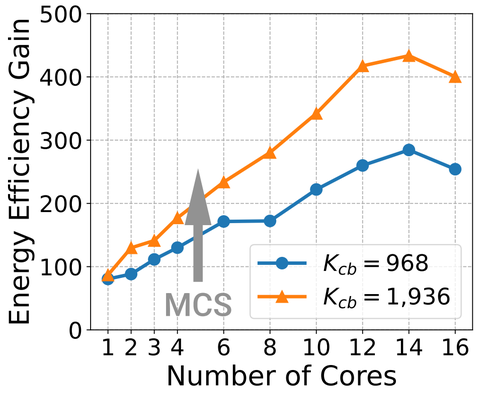


CPU power →





Silicom's Lisbon P2 accelerator with Intel's ACC100 ASIC



CPU power →

- Z. Qi*, C.-H. Tung*, A. Kalia, and T. Chen, "Savannah: Efficient mmWave baseband processing with minimal and heterogeneous resources," in *Proc. ACM International Conference on Mobile Computing and Networking (MobiCom'24)*, 2024.
- Code and data: https://github.com/functions-lab/Savannah

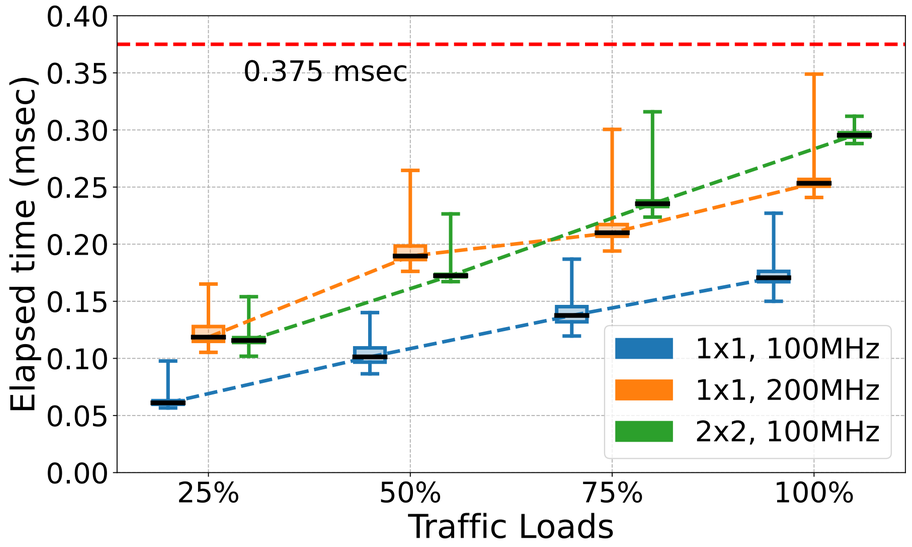
Duke ENGINEERING FuNCtions Lab

CUbiC Center for Ubiquitous Connectivity

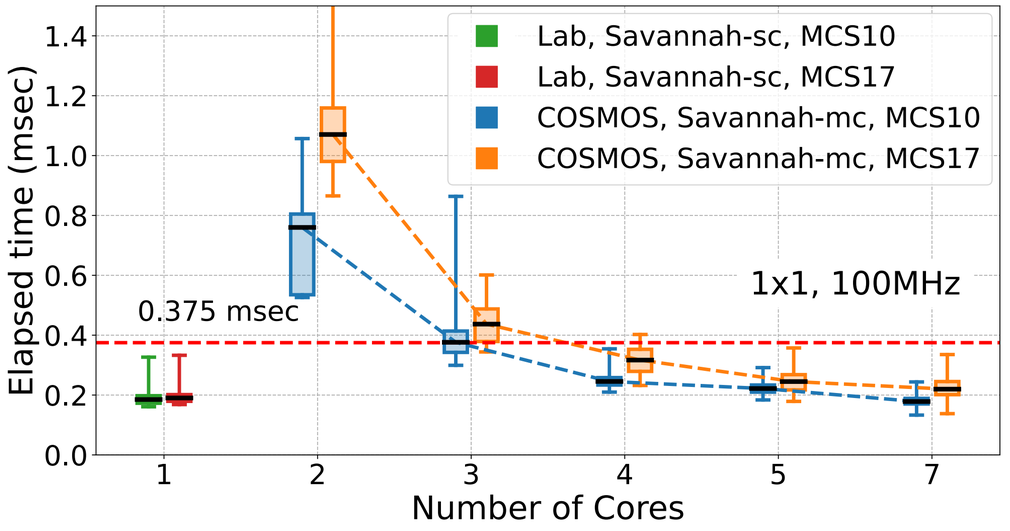
Savannah: Real-time Efficient mmWave Baseband Processing

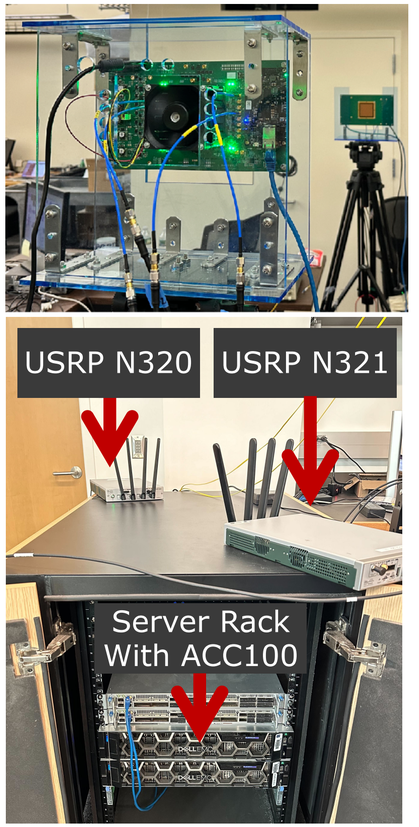
- **Lab**: Dell R750 server w/ Intel Xeon Gold 6348 CPU (56-core, @2.6 GHz), Silicom's Lisbon P2 ACC100 card
- **COSMOS**: Dell R740 server w/ Intel Xeon Gold 6226 CPU (48-core, @2.7 GHz)

Achievable link data rate:
243.9 Mbps, 487.8 Mbps, 487.8 Mbps



MCS10: 16QAM, LDPC code rate = 340/1,024
MCS17: 64QAM, LDPC code rate = 438/1,024





ReACT Testbed Phase 1

Savannah supports 2x2 MIMO, 100 MHz bandwidth in FR2 with a single CPU core + ASIC with 2x reduced power consumption with optimized signal processing and compute resource sharing

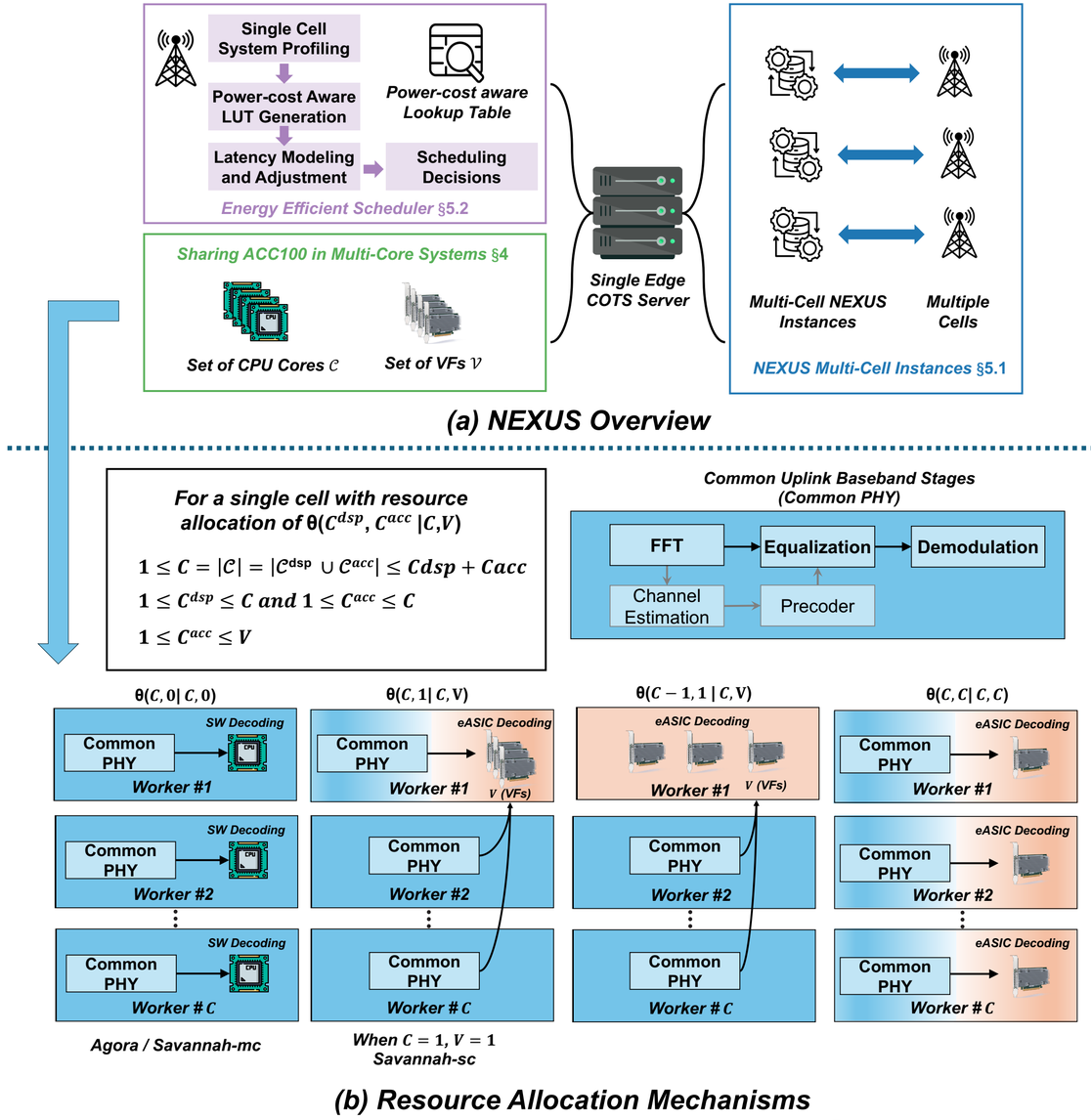
- Z. Qi*, C.-H. Tung*, A. Kalia, and T. Chen, "Savannah: Efficient mmWave baseband processing with minimal and heterogeneous resources," in *Proc. ACM International Conference on Mobile Computing and Networking (MobiCom'24)*, 2024.
- Code and data: https://github.com/functions-lab/Savannah

Duke ENGINEERING FuNCtions Lab

CUbiC Center for Ubiquitous Connectivity

NEXUS: Efficient Multi-cell mmWave Baseband Processing

- **Problem**: Use a single server with a pool of (`cpu`, `acc100_vf`) resources supporting PHY processing of heterogeneous cells with varying (`bw`, `mimo`) configurations, traffic load, and MCSs
- **Objective**: Minimize the total power consumption while ensuring that individual cells meet the latency requirements
- **Highlights**:
  - **Co-optimized software-based DSP (CPU) and LDPC decoding offloaded to ACC100 virtual functions (VFs)**
  - **Power-aware compute resource allocation**
  - **A single server can support 16 cells with a total uplink data rate of ~5.4 Gbps using 16 CPU cores and one ACC100 HWA (16 VFs) while being scalable**



- Z. Qi, C.-H. Tung, Z. Gao, and T. Chen, "NEXUS: Efficient and scalable multi-cell mmWave baseband processing with heterogeneous compute," *arXiv:2509.04625 [cs.NI]*, Sept. 2025.

NEXUS: Single-cell Optimization

- **Cell configuration**: $\sigma\,(B, M, \rho_t, \rho_f, \chi)$

- **Resource allocation strategy**: $\theta\,(C^{\text{dsp}}, C^{\text{acc}} | C, V)$
  - With $C \leq C_{\text{max}}$ cores and $V \leq V_{\text{max}}$ ACC100 VFs
  - C^{dsp} cores are allocated for software DSP
  - C^{acc} cores are allocated for ACC100 offloading tasks

Dimension	Notation	Range	Granularity
Channel BW	B	{100, 200, 400} MHz	Discrete
MIMO Size	M	{1, 2, 4}	Discrete
Traffic Load	ρ_t	[0,100] %	6.25%
Transmission BW %	ρ_f	[0,100] %	Multiple PRBs
MCS Index	χ	$\{0, 1, \ldots, 28\}$	Integer

- Feasible compute resource allocation strategies:

$$\Theta(C, V) = \left\{\theta\left(C^{\text{dsp}}, C^{\text{acc}} | C, V\right)\right\}$$

$$\text{s.t. } 1 \leq C = \left|\mathcal{C}^{\text{dsp}} \cup \mathcal{C}^{\text{acc}}\right| \leq C^{\text{max}},$$

$$1 \leq C^{\text{dsp}} \leq C,\ 0 \leq C^{\text{acc}} \leq C,$$

$$1 \leq C^{\text{acc}} \leq V.$$

Each core is dedicated to either software DSP, ACC100 offloading (LDPC decoding), or both

Each core responsible for ACC100 offloading is mapped to a unique set of VFs, and each VF cannot be managed by more than one core

▸ Z. Qi, C.-H. Tung, Z. Gao, and T. Chen, "NEXUS: Efficient and scalable multi-cell mmWave baseband processing with heterogeneous compute," *arXiv:2509.04625 [cs.NI]*, Sept. 2025.

NEXUS: Single-cell Optimization

- **Power-aware compute resource optimization** for a single cell:

$$\theta^\star := \arg\min_{\theta\in\Theta} : P(\theta) = \alpha_1 P_1(C^{\text{dsp}}) + \alpha_2 P_2(C^{\text{acc}}) + \alpha_3 P_3(V), \text{ s.t. } \mathcal{L}(\theta) < 3 \text{ slots.}$$

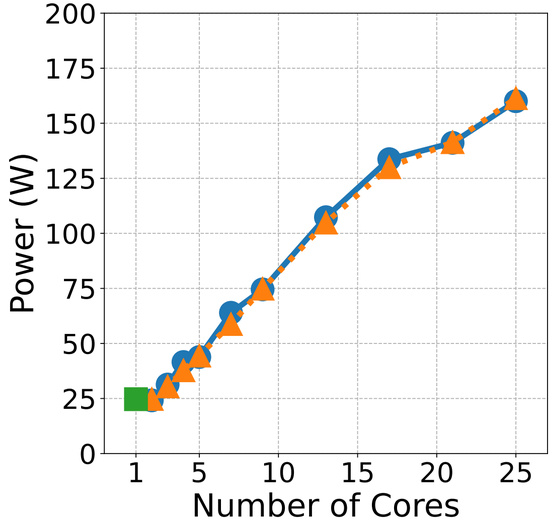
$\approx 7.0 \cdot C$ (Watt)

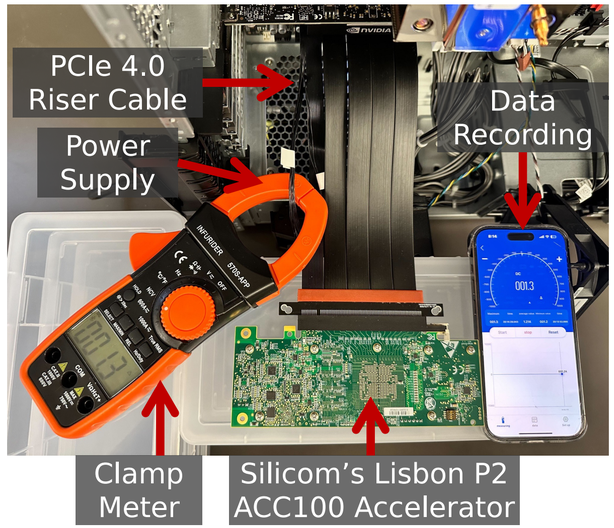
Using Intel's Performance Counter Monitor (PCM)

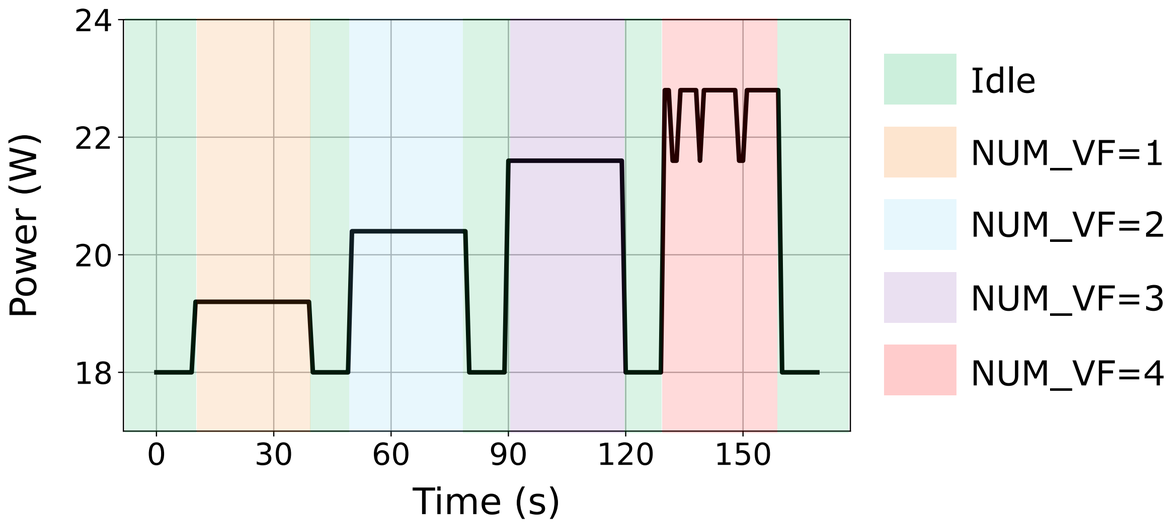
$\approx 1.2 \cdot V$ (Watt)

Verified method using a clamp meter

Power Consumption







- Z. Qi, C.-H. Tung, Z. Gao, and T. Chen, "NEXUS: Efficient and scalable multi-cell mmWave baseband processing with heterogeneous compute," *arXiv:2509.04625 [cs.NI]*, Sept. 2025.

NEXUS: Single-cell Optimization

- **Power-aware compute resource optimization** for a single cell:

$$\theta^\star := \arg\min_{\theta\in\Theta} : P(\theta) = \alpha_1 P_1(C^{\text{dsp}}) + \alpha_2 P_2(C^{\text{acc}}) + \alpha_3 P_3(V), \text{ s.t. } \mathcal{L}(\theta) < 3 \text{ slots.}$$

Large configuration space

$\approx 7.0 \cdot C$ (Watt)

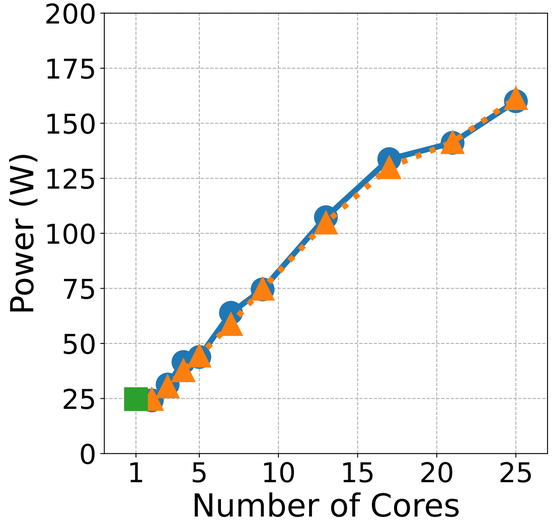
Using Intel's Performance Counter Monitor (PCM)

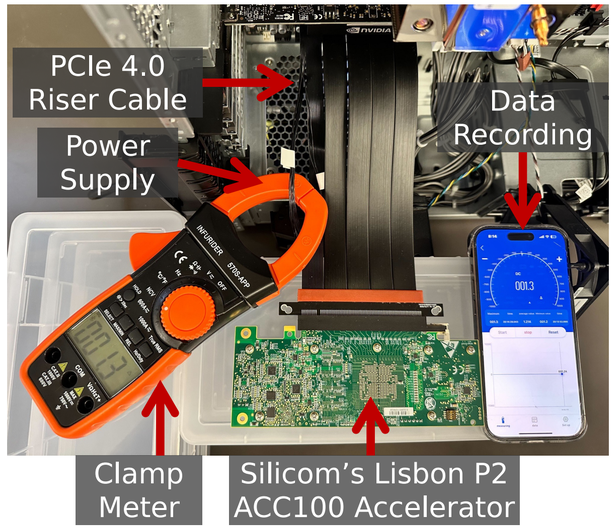
$\approx 1.2 \cdot V$ (Watt)

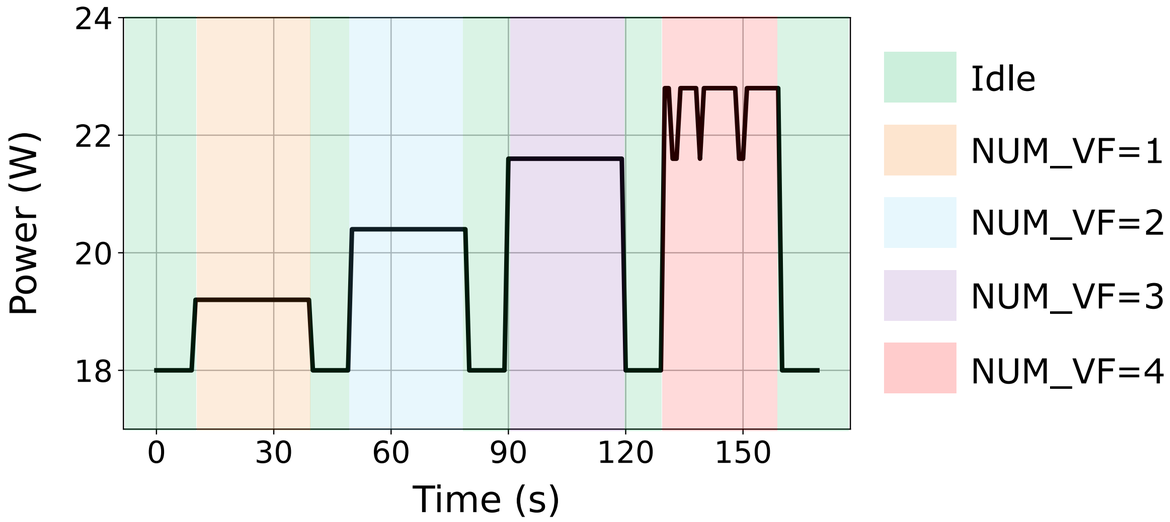
Verified method using a clamp meter

Difficult to model

Power Consumption



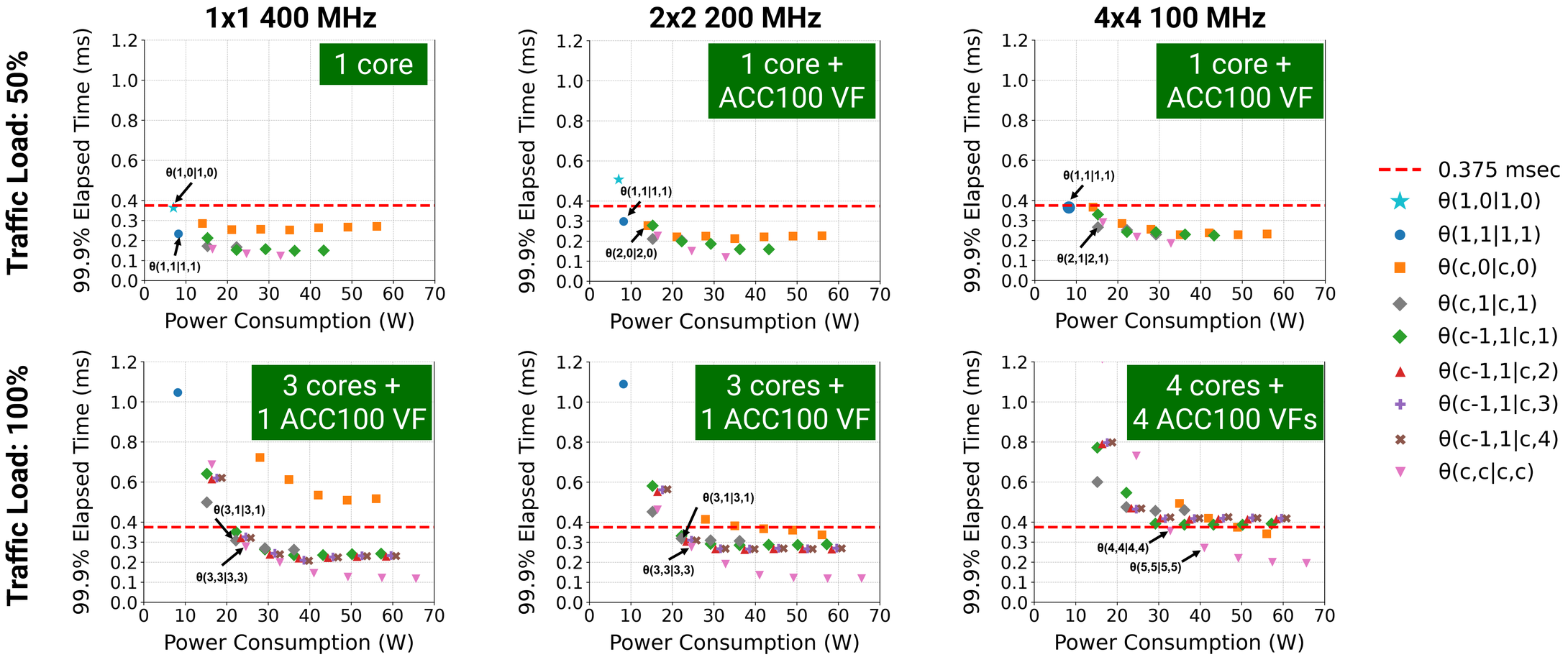




- Z. Qi, C.-H. Tung, Z. Gao, and T. Chen, "NEXUS: Efficient and scalable multi-cell mmWave baseband processing with heterogeneous compute," *arXiv:2509.04625 [cs.NI]*, Sept. 2025.

NEXUS: Single-cell Optimization

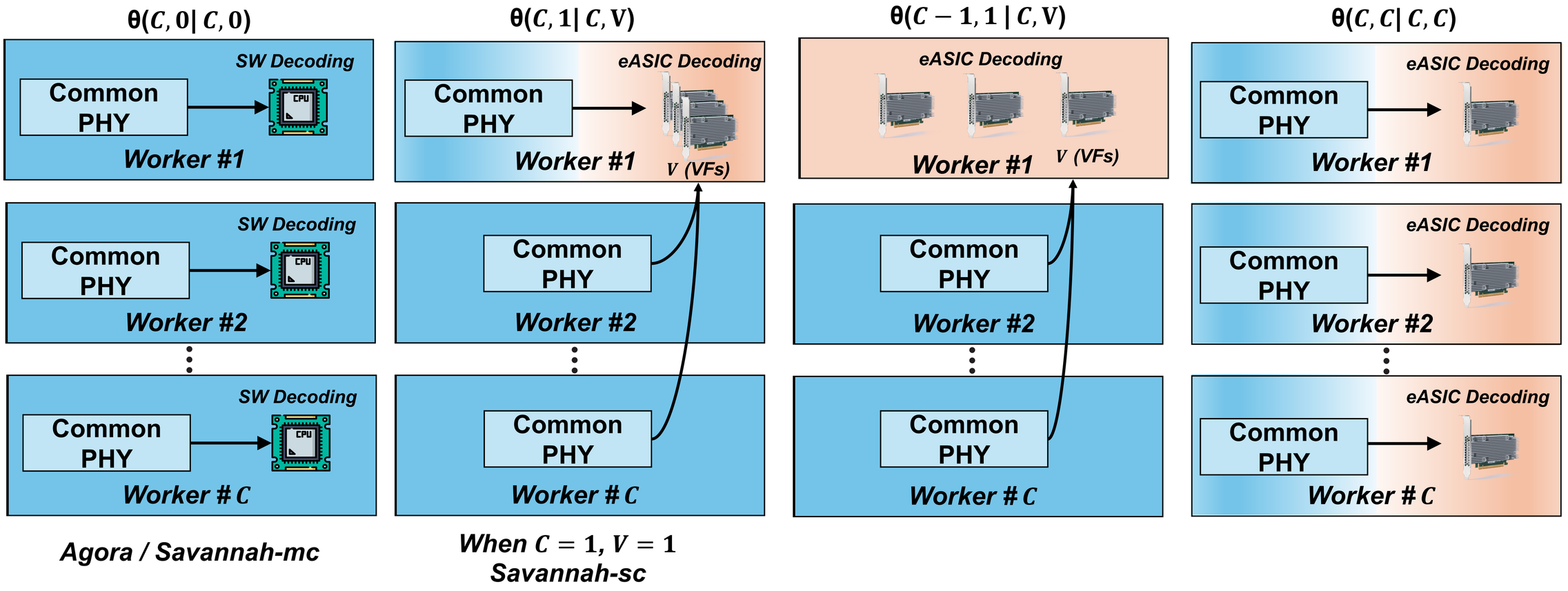
- **Configuration space reduction** based on key observations from comprehensive system profiling
  - E.g., from 152 to 15 with $(C, V) = (26, 16)$



- Z. Qi, C.-H. Tung, Z. Gao, and T. Chen, "NEXUS: Efficient and scalable multi-cell mmWave baseband processing with heterogeneous compute," *arXiv:2509.04625 [cs.NI]*, Sept. 2025.

NEXUS: Single-cell Optimization

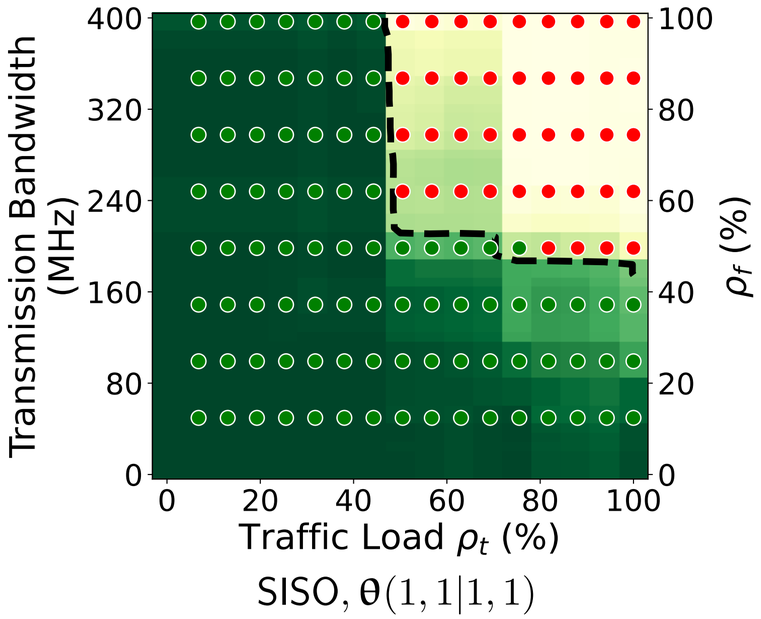
- **Configuration space reduction** based on key observations from comprehensive system profiling
  - E.g., from 152 to 15 with $(C, V) = (26, 16)$
- We consider four resource allocation strategies:

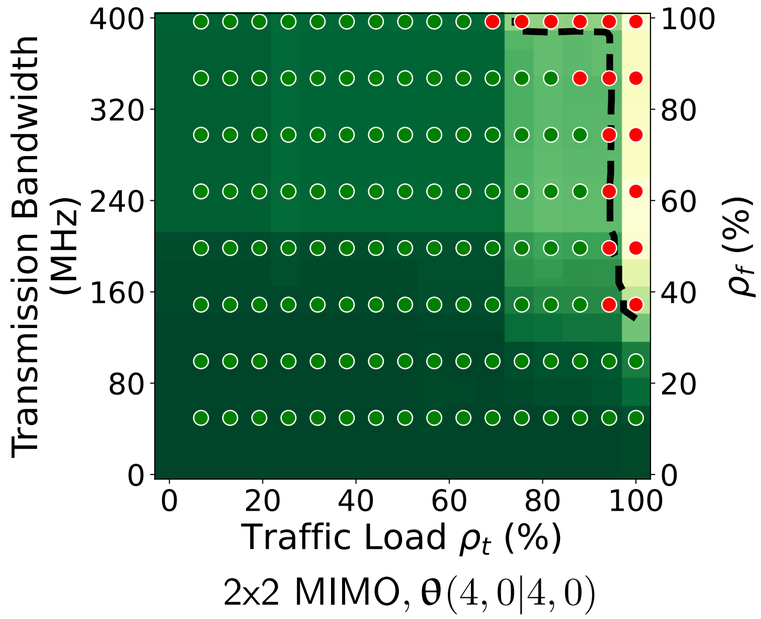


▸ Z. Qi, C.-H. Tung, Z. Gao, and T. Chen, "NEXUS: Efficient and scalable multi-cell mmWave baseband processing with heterogeneous compute," *arXiv:2509.04625 [cs.NI]*, Sept. 2025.

NEXUS: Single-cell Optimization

- **Configuration space reduction** based on key observations from comprehensive system profiling
  - E.g., from 152 to 15 with $(C, V) = (26, 16)$
- **Random forest (RAF) based feasibility prediction**: $f : (\sigma, \theta) \mapsto [0, 1]$
  - Trained on >9K profiling experiments, each with 20K frames with detailed PHY processing timing





RAF-based feasibility prediction (dashed lines) vs. ground truth (*green*: feasible; *red*: infeasible), for an unseen MCS-18

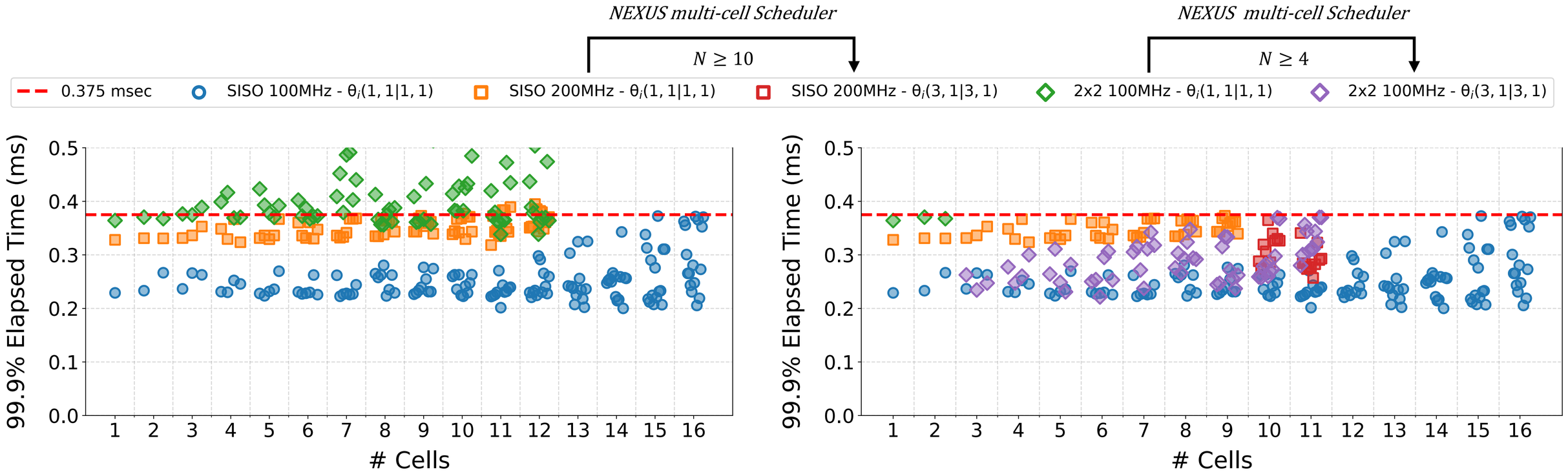
	Accuracy	Inference Time
MLP	99.1%	582 usec
1D CNN	99.2%	615 usec
TinyNet	99.3%	570 usec
RAF (10 trees)	99.0%	1.4 usec
RAF (100 trees)	99.1%	4.3 usec

Accuracy and inference time comparison between different methods

▸ Z. Qi, C.-H. Tung, Z. Gao, and T. Chen, "NEXUS: Efficient and scalable multi-cell mmWave baseband processing with heterogeneous compute," *arXiv:2509.04625 [cs.NI]*, Sept. 2025.

NEXUS: Multi-cell Optimization

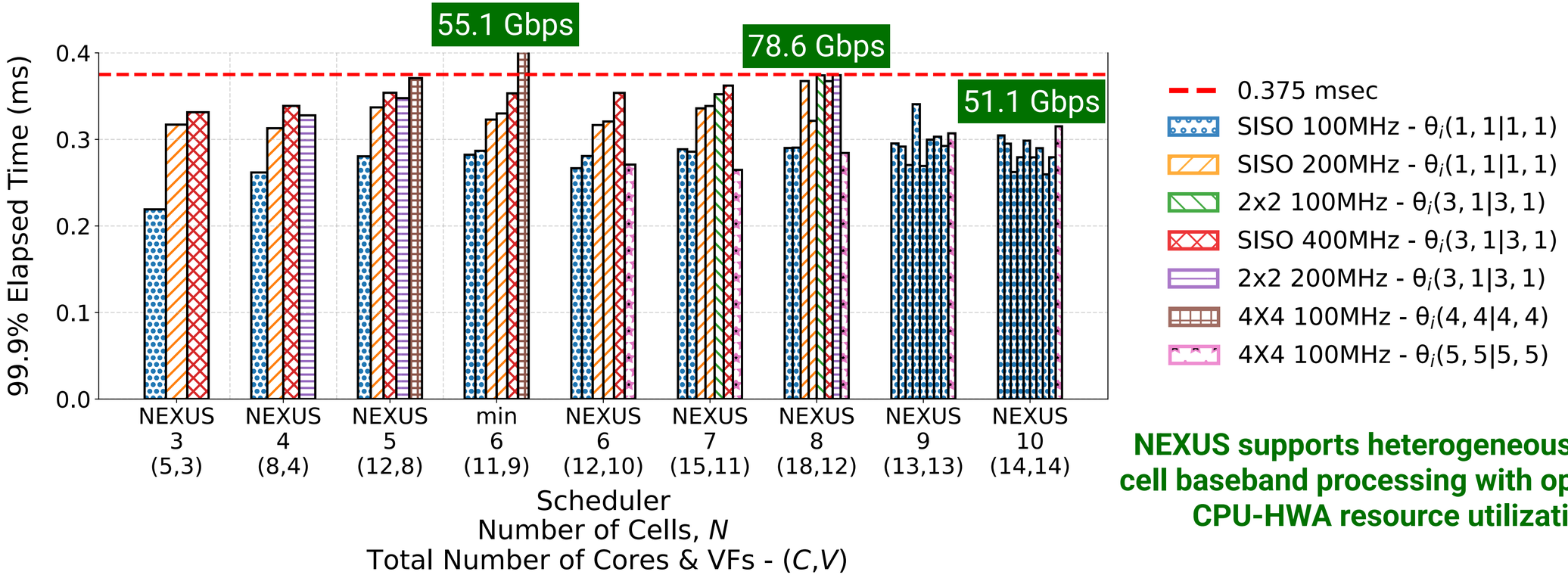
- Assume **non-overlapping allocation** of compute resources (CPU cores, ACC100 VFs) across *N* cells
- Leverage single-cell optimization to reduce configuration space across *N* cells
- **Contention-aware latency adjustment** of the RAF-based feasibility prediction:
  - If the predicted cell configuration cannot meet the latency requirements, go to the next best configuration



- Z. Qi, C.-H. Tung, Z. Gao, and T. Chen, "NEXUS: Efficient and scalable multi-cell mmWave baseband processing with heterogeneous compute," *arXiv:2509.04625 [cs.NI]*, Sept. 2025.

NEXUS: Implementation and Results

- ~5K LOC (C++), ~9K configuration files, ACC100 VF management based on Intel's **`bbdev`** tool (**`pf_bb_config`**) and DPDK's threading model (**`rte_thread`**)
- Two servers emulating the set of RUs and DUs, connected via 100 GbE links (Mellanox CX6)
- Example: NEXUS with heterogeneous cell configurations and heterogeneous resource allocations



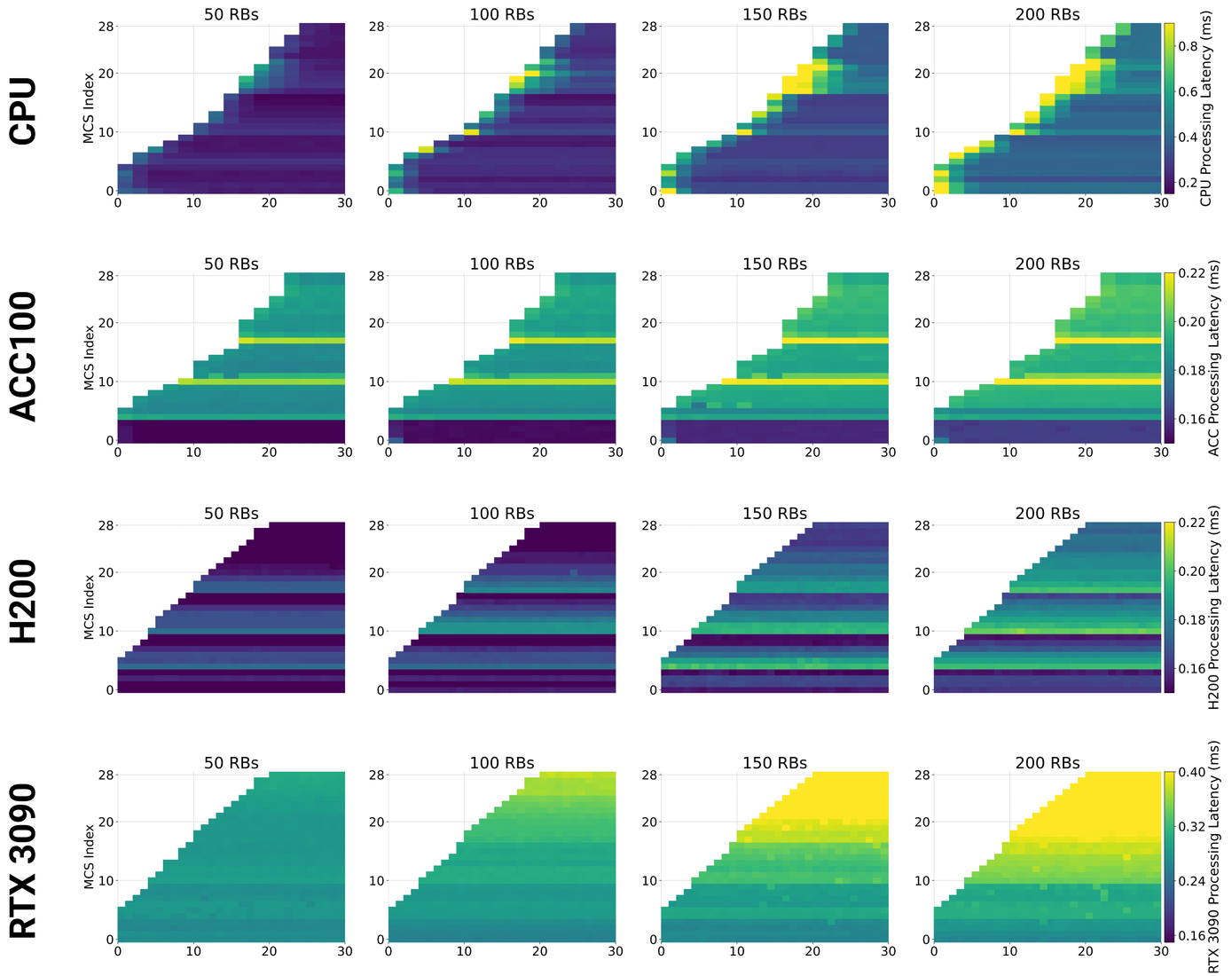
NEXUS supports heterogeneous multi-cell baseband processing with optimized CPU-HWA resource utilization

▸ Z. Qi, C.-H. Tung, Z. Gao, and T. Chen, "NEXUS: Efficient and scalable multi-cell mmWave baseband processing with heterogeneous compute," *arXiv:2509.04625 [cs.NI]*, Sept. 2025.

DecodeX: LDPC Decoding Across CPU, GPU, and ASIC Platforms

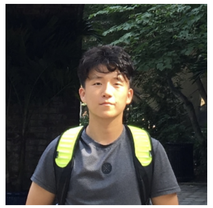
- **A comprehensive suite of LDPC decoder implementations** (kernels, APIs, and test vectors) for benchmarking and R&D:
  - CPU (FlexRAN)
  - GPU (NVIDIA Aerial and Sionna-RK)
  - ASIC (ACC100)

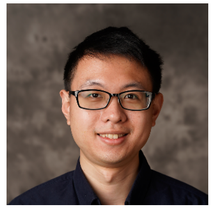
- Allow for evaluation of L1 processing models with different HWAs:
  - **Inline**: CPU or GPU only (e.g., FlexRAN, Aerial)
  - **Lookaside**: CPU + ACC100 (e.g., Savannah, NEXUS), CPU + GPU
  - **Hybrid**: CPU-GPU with unified memory (e.g., Sionna-RK)



▸ Z. Qi, Y. Yao, Y. Li, C.-H. Tung, J. Zheng, D. Zhuo, and T. Chen, “DecodeX: Exploring and benchmarking of LDPC decoding across CPU, GPU, and ASIC platforms,” *arXiv:2511.02952 [cs.NI]*, Nov. 2025.

Thank you! Questions?


Zhenzhou Qi


Chung-Hsuan Tung


Yuncheng Yao


Yiming Li


Yunyao Zheng


Zhihui Gao


Anuj Kalia


Danyang Zhuo

tingjun.chen@duke.edu; https://tingjunchen.com

Platforms for Advanced Wireless Research

COSMOS

ATHENA

NSF AI Institute for Edge Computing

CUbiC

Center for Ubiquitous Connectivity

Semiconductor Research Corporation

sigmobile

Google

IBM

NEC Laboratories America

Relentless passion for innovation

NTT

NVIDIA

Duke ENGINEERING

FuNCtions Lab