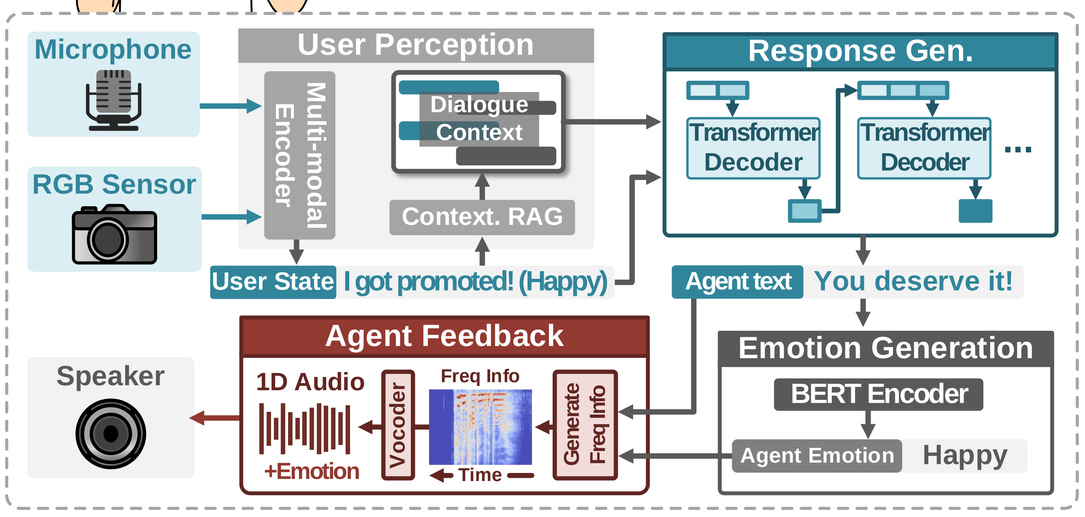
BROCA: A Low-power and Low-latency Conversational Agent RISC-V System-on-Chip for Voice-interactive Mobile Devices

Wooyoung Jo, Seongyon Hong, Jiwon Choi, Beomseok Kwon, Haoyang Sang, Dongseok Im, Sangyeob Kim, Sangjin Kim, Chaeyun Jeong, Yujin Moon and Hoi-Jun Yoo

Motivation

▶ Mobile Voice-interactive Conversational Agent

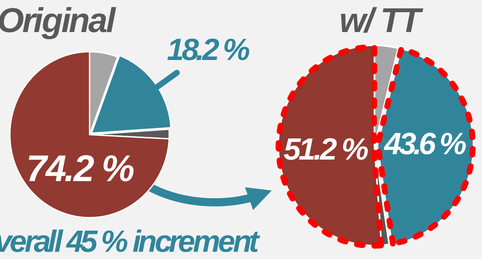
- Voice Conversation w/ Contexts
- Multimodal w/ Emotions
- Constrained On-chip Memory



▶ Design Challenges for On-device Conversational Agent

System Energy Problem

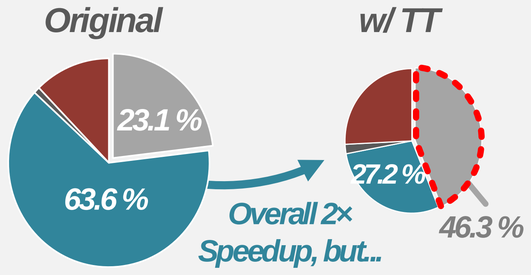
< System Energy Breakdown w/ TT >



Increased energy consumption due to TT

System Latency Problem

< System Latency Breakdown w/ TT >

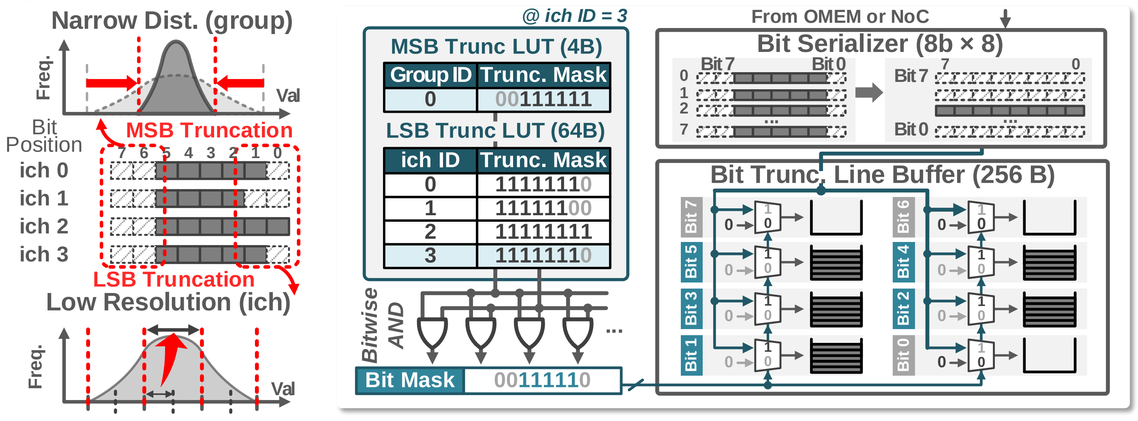


Latency bottleneck by RAG

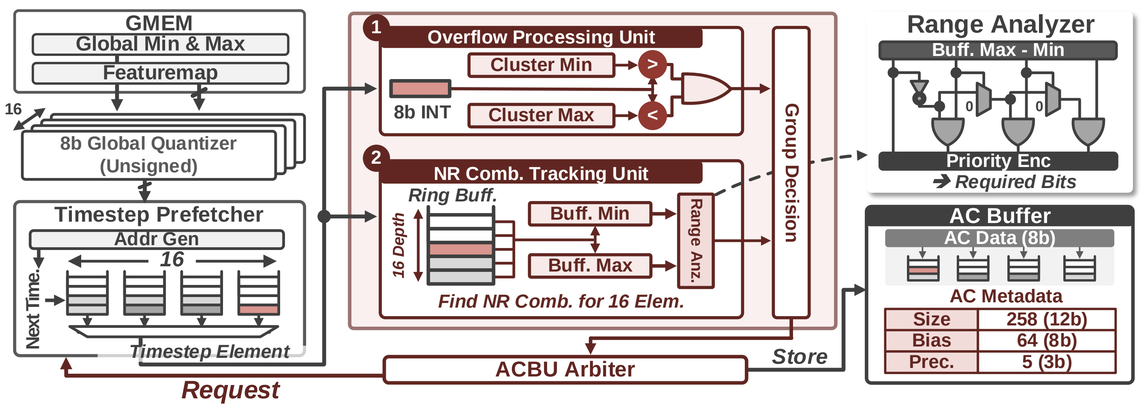
➔ **Bit redundancy** in modalities & **context loading** latency

BROCA Architecture

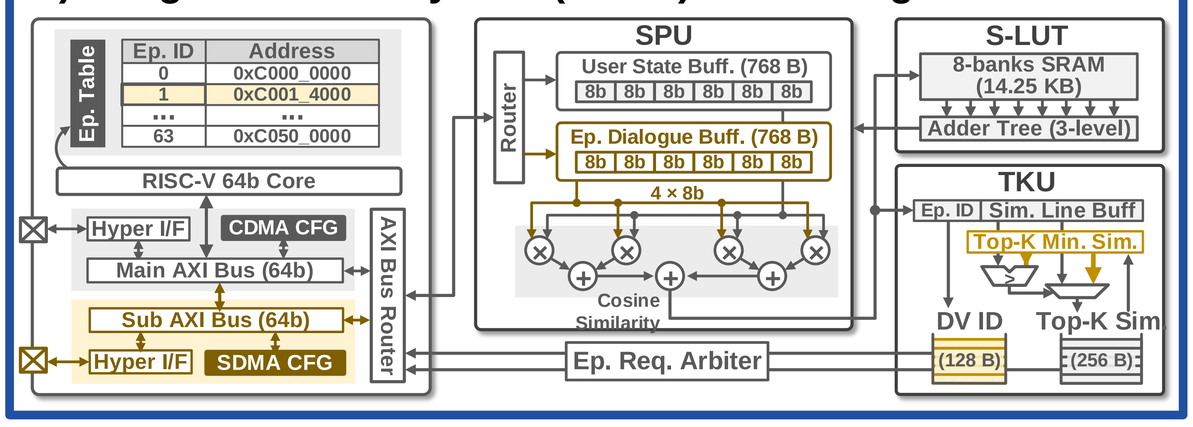
1) Adaptive Bit-truncate Unit (ABTU) ➔ RG Stage



2) Acoustic Cluster Bit-grouping Unit (ACBU) ➔ AF Stage

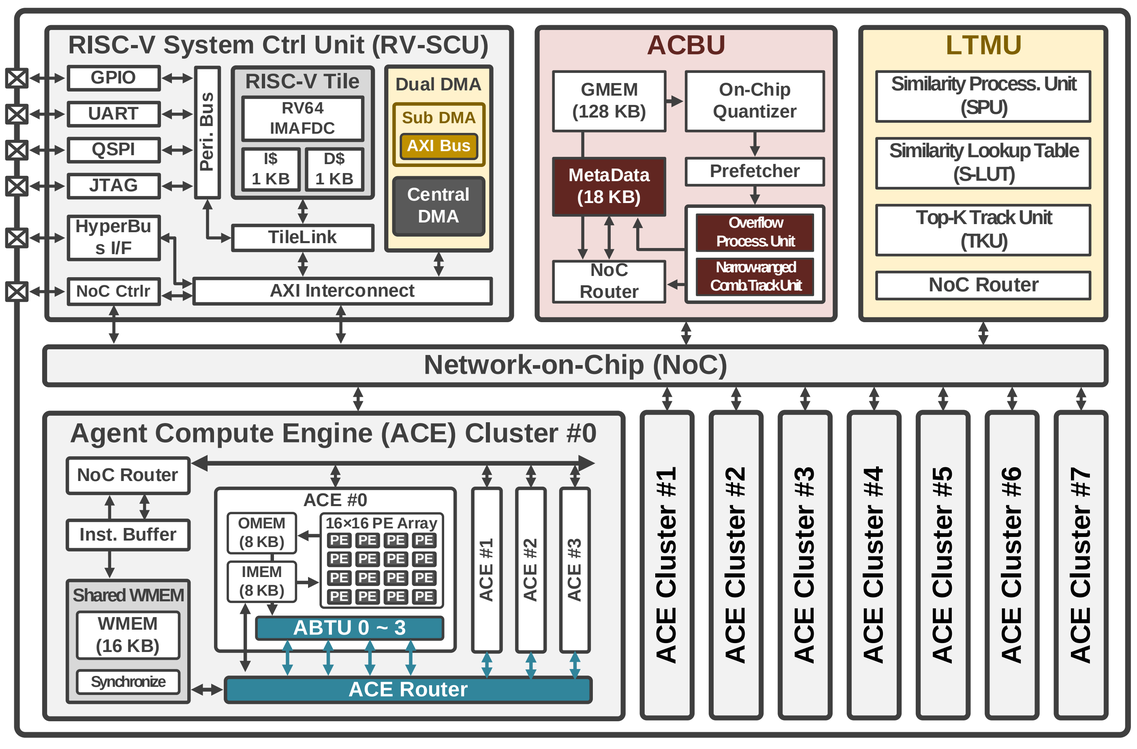


3) Long-term Memory Unit (LTMU) ➔ UP Stage



System Implementation

▶ Overall Architecture of BROCA

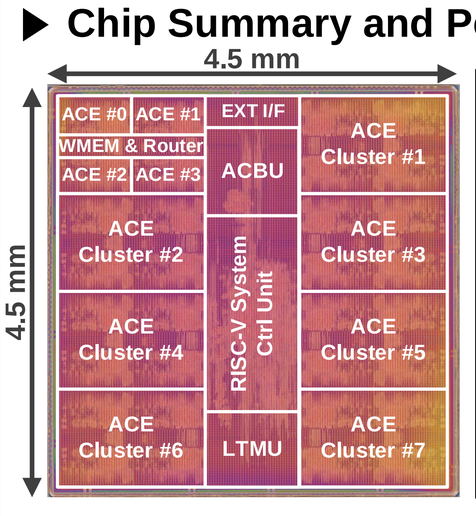


▶ Benchmark Results of a Conversational Agent

		User Perception	Response Gen.	Agent Feedback	Dialogue Context.
Task		Speech-to-Text	Text Generation	Voice Synthesis	Vector Retrieval
Dataset		LibriSpeech	WikiText-103	RAVDESS	DailyDialogues
Model		ConformerASR	GPT-2 (Small)	HiFi-GAN	DPR Enc. for RAG
Precision	Input	INT8	INT2 - 8	INT2 - 8	INT8
	Weight	INT8	INT8	INT8	INT8
Parameters		10.1 M	124.4 M → 17 M	14.4 M	85 M
Accuracy		WER: 2.8 (+0.1)	ppl: 27.6 (+3.3)	MCD: 1.8 (Lossless)	R@1: 0.73

System Verification

▶ Chip Summary and Performance



		Specifications
Technology		Samsung 28nm LPP
Die Area		4.5 mm × 4.5 mm
SRAM		834 KB
System ISA		RV64-IMAFDC
Supply Voltage		0.7 – 1.1 V
Max. Frequency		200 MHz
Data Type	Input	INT2 – INT8
	Weight	INT8

System Performance		
Peak Performance	1.64 TOPS	
Operating Condition	50 MHz 0.7 V	200 MHz 1.1 V
Power Consump.	52.4 mW	559.2 mW
Overall Latency	1480.1 ms	370 ms

▶ Demonstration System of a Conversational Agent