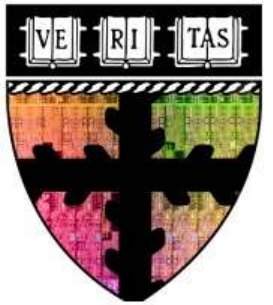


SM6: A 16nm System-on-Chip for Accurate and Noise-Robust Attention-Based NLP Applications

The 33rd Hot Chips Symposium – August 22-24, 2021

Thierry Tambe[1], En-Yu Yang[1], Glenn G. Ko[1], Yuji Chai[1], Coleman Hooper[1],
Marco Donato[2], Paul N. Whatmough[1,3], Alexander M. Rush[4], David Brooks[1], and Gu-Yeon Wei[1]

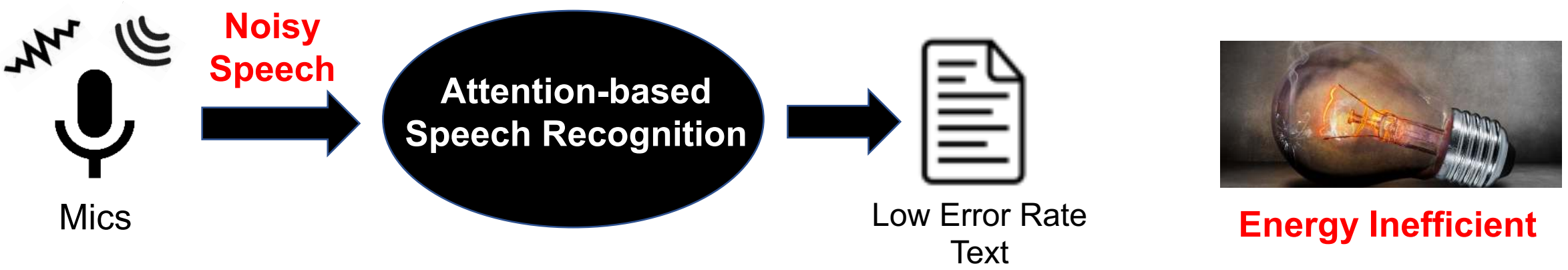
[1]Harvard University, Cambridge, MA, [2]Tufts University, Medford, MA,
[3]ARM Research, Boston, MA, [4]Cornell University, New York, NY

Abstract

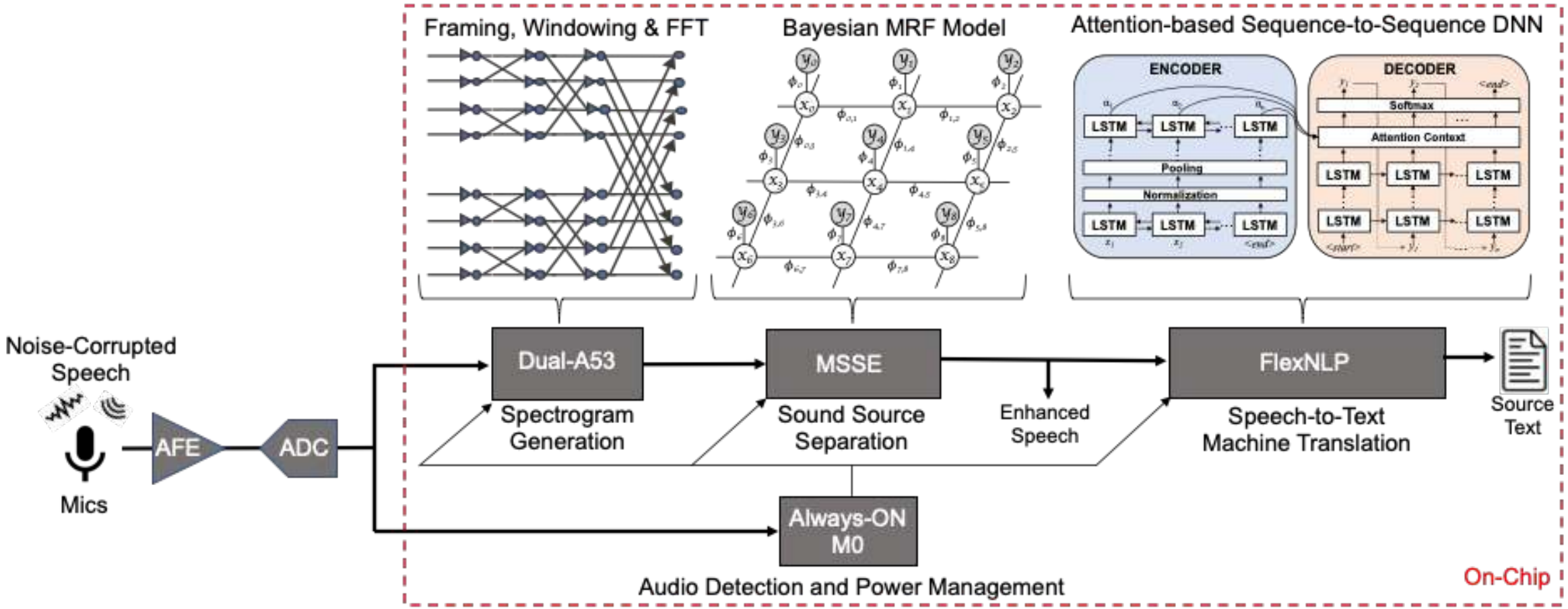
In this work, we present SM6, an SoC architecture for real-time denoised speech and NLP pipelines, featuring (1) MSSE: an unsupervised probabilistic sound source separation accelerator, (2) FlexNLP: a programmable inference accelerator for attention-based seq2seq DNNs using adaptive floating-point datatypes for wide dynamic range computations, (3) a dual-core Arm Cortex A53 CPU cluster, which provides on-demand SIMD FFT processing, and operating system support.

In adverse acoustic conditions, MSSE allows FlexNLP to store up to 6x smaller ASR models obviating the very inefficient strategy of scaling up the DNN model to achieve noise robustness. MSSE and FlexNLP produce efficiency ranges of 4.33-17.6 Gsamples/s/W and 2.6-7.8TFLOPs/W, respectively, with per-frame end-to-end latencies of 15-45ms.

Speech-Enhancing ASR

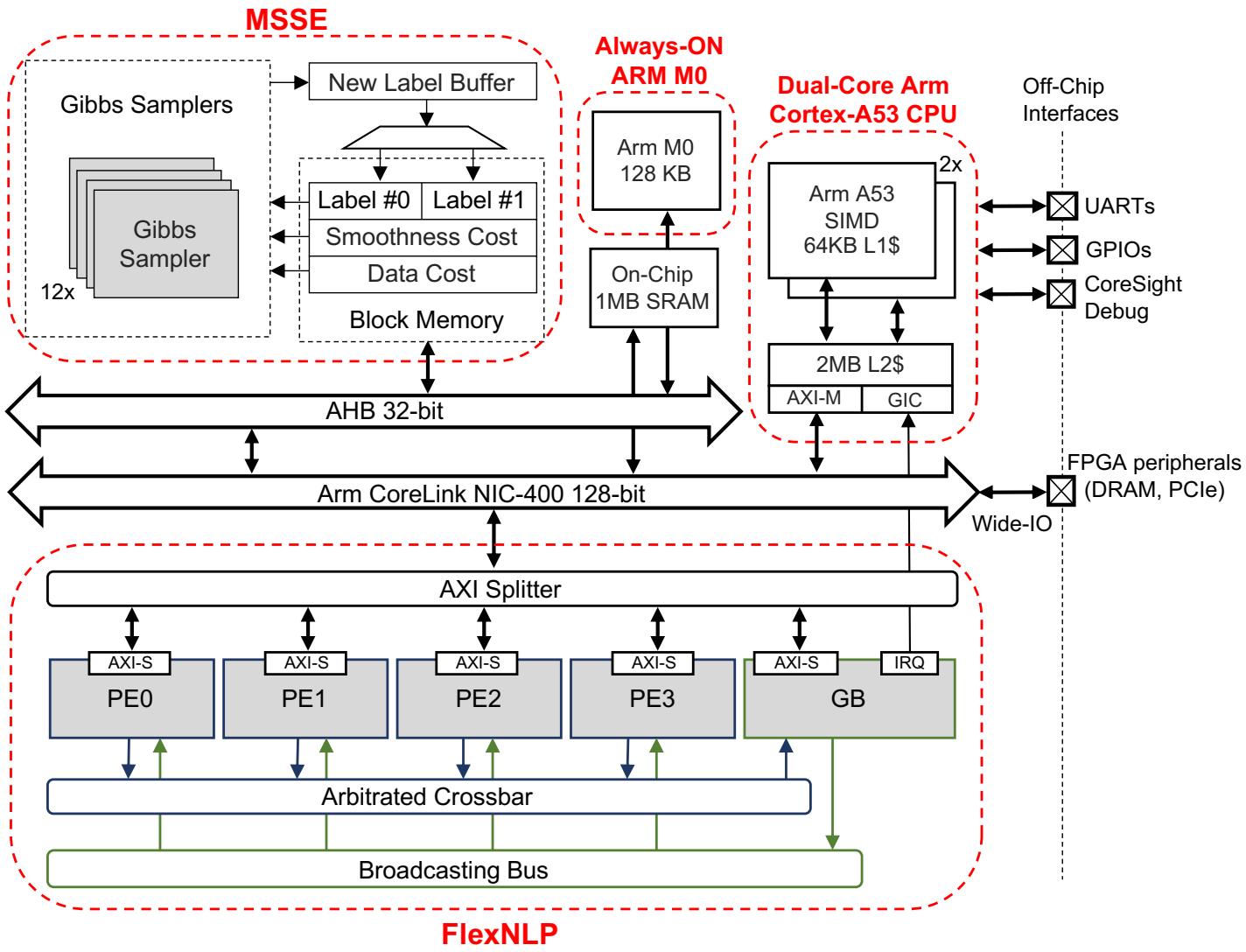


Proposed Universal Translator Pipeline in SM6



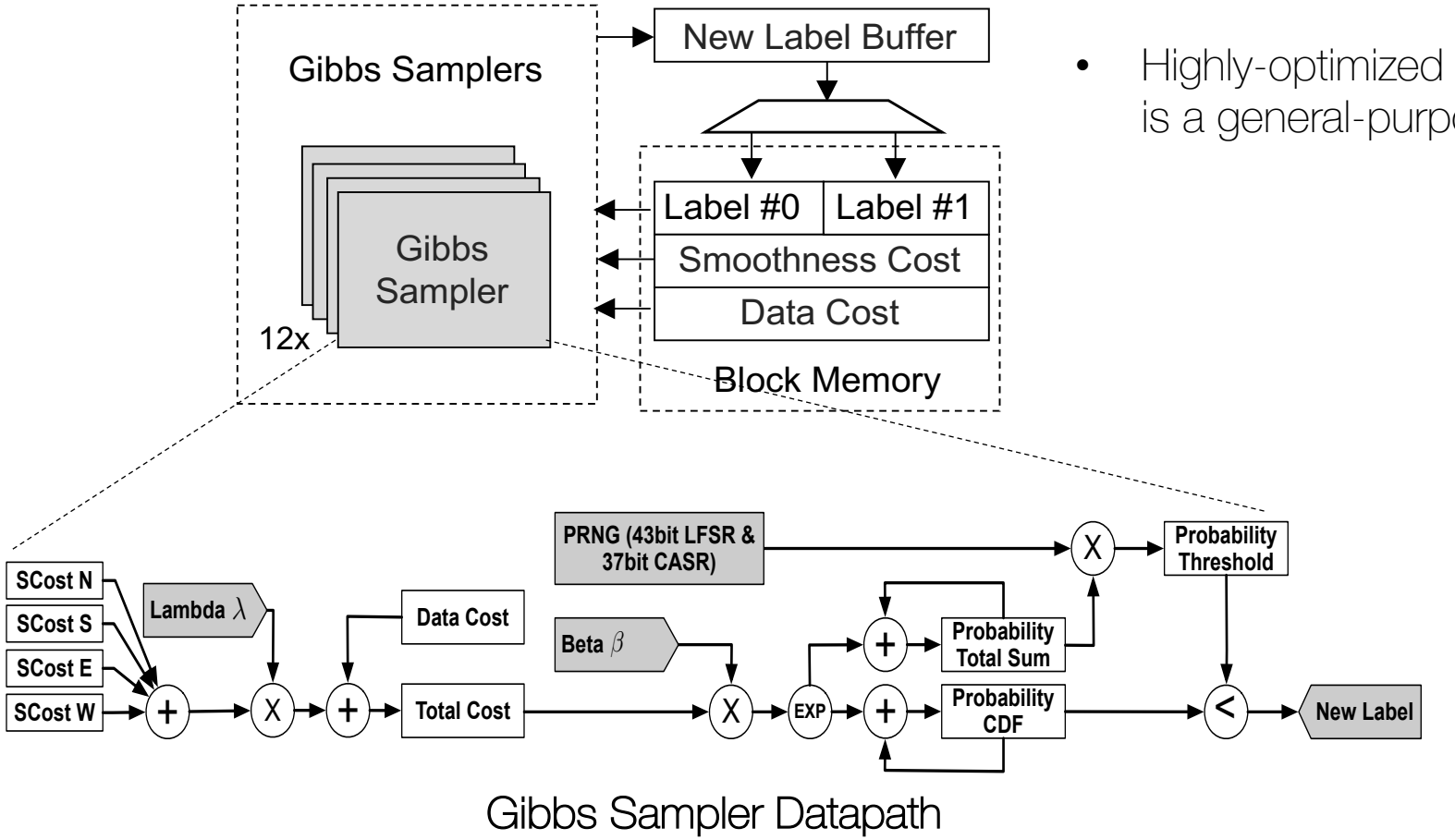
- M0: autonomously monitors incoming audio amplitudes and subsequently boots A53, FlexNLP and MSSE
- Dual A53: performs feature extraction tasks (framing, windowing, 1024-pt FFT) on its dual-issue SIMD datapath
- MSSE: optimized for unsupervised speech enhancement via MCMC Gibbs sampling
- FlexNLP: optimized for whole-model acceleration of large-vocabulary, bidirectional, attention-based DNNs

SM6 SoC Architecture



- FlexNLP accelerator
- MSSE utilizes 12 parallel Gibbs samplers to solve the spectrogram MRF and ultimately produces a binary label corresponding to *noise* or *speech*.
- Dual A53 with 2MB L2 cache.
- M0: always-on for audio detection and power management
- 128-bit AXI and 32-bit AHB NoCs
- Wide-IO to off-chip DRAM

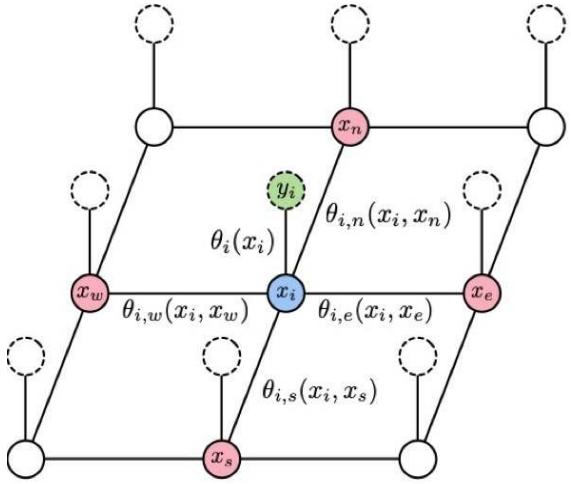
MRF Sound Source Separation Engine (MSSE)



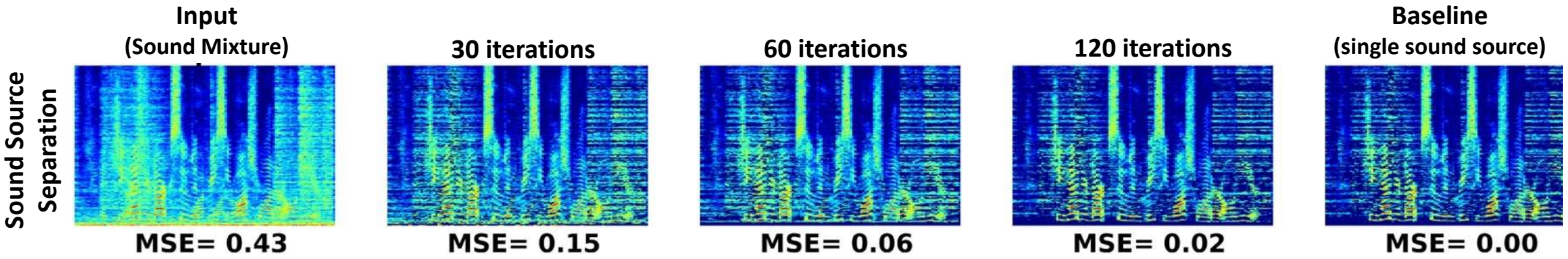
- Highly-optimized for sound source separation while PGMA [1] is a general-purpose Bayesian inference accelerator

[1] G. Ko et al., A 3mm2 Programmable Bayesian Inference Accelerator for Unsupervised Machine Perception using Parallel Gibbs Sampling in 16nm, VLSI Symposium, 2020

Sound Source Separation

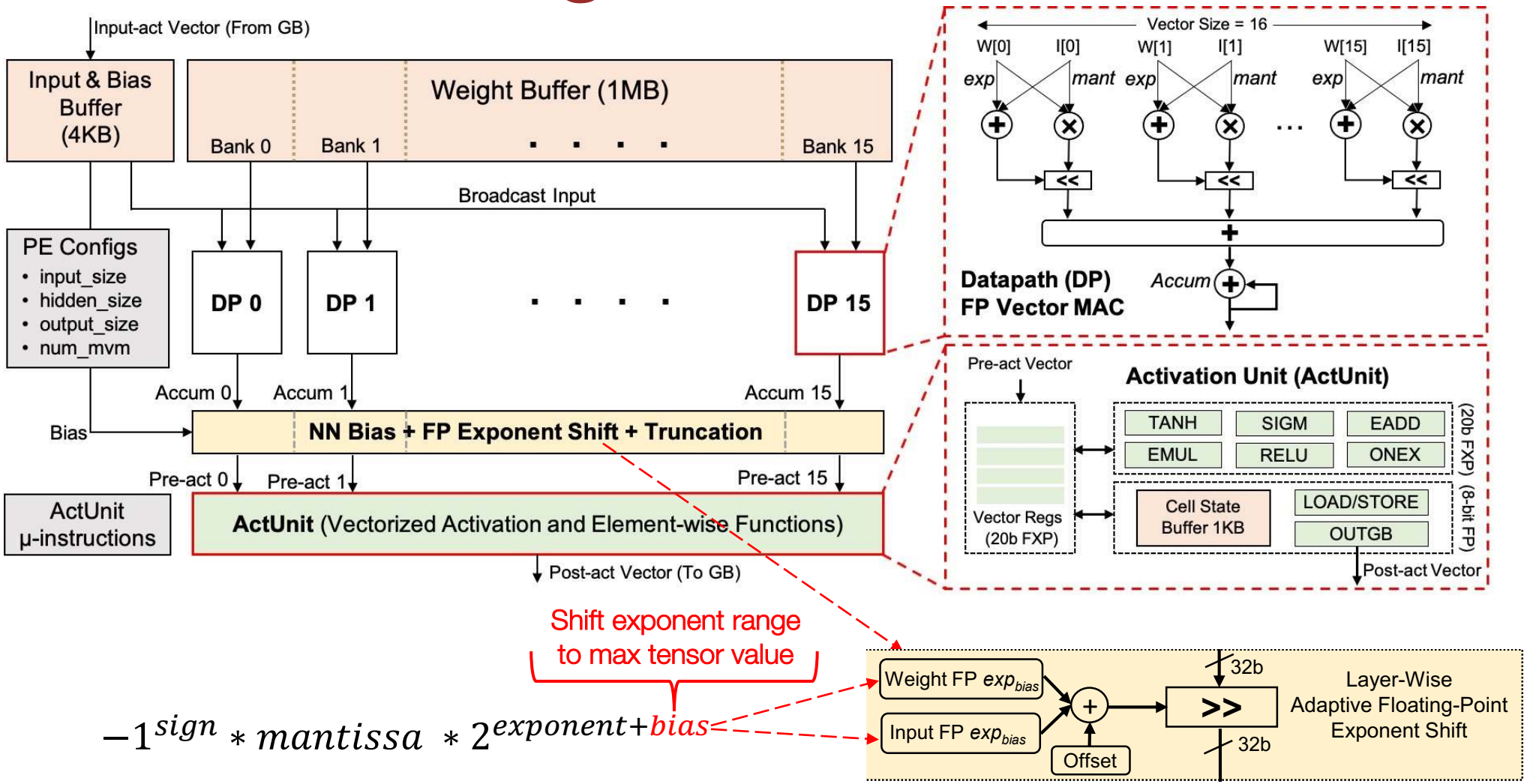


- Nodes representing input speech features
- Nodes representing output labels corresponding to feature locations
- Node being sampled
- Observed node
- Neighbor labels



[2] G. Ko et al., A 3mm2 Programmable Bayesian Inference Accelerator for Unsupervised Machine Perception using Parallel Gibbs Sampling in 16nm, VLSI Symposium, 2020

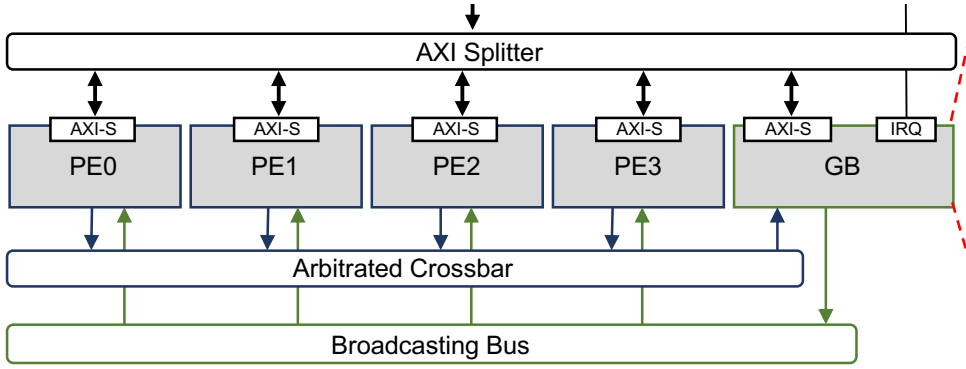
FlexNLP Processing Element

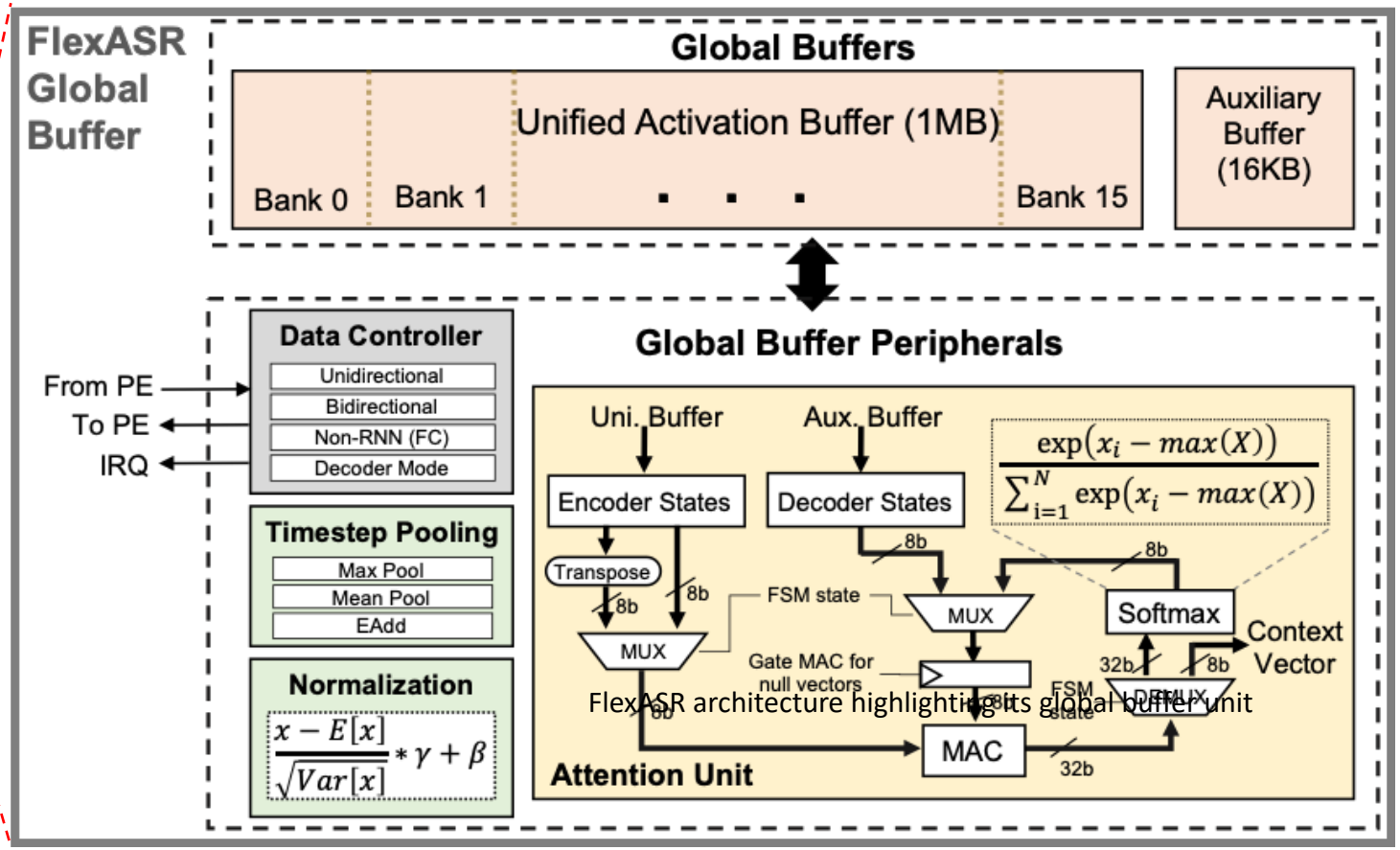


- **Adaptive floating-point computations with per-layer custom exponent bias for high dynamic range computations**

FlexNLP Multi-Function Global Buffer (GB)

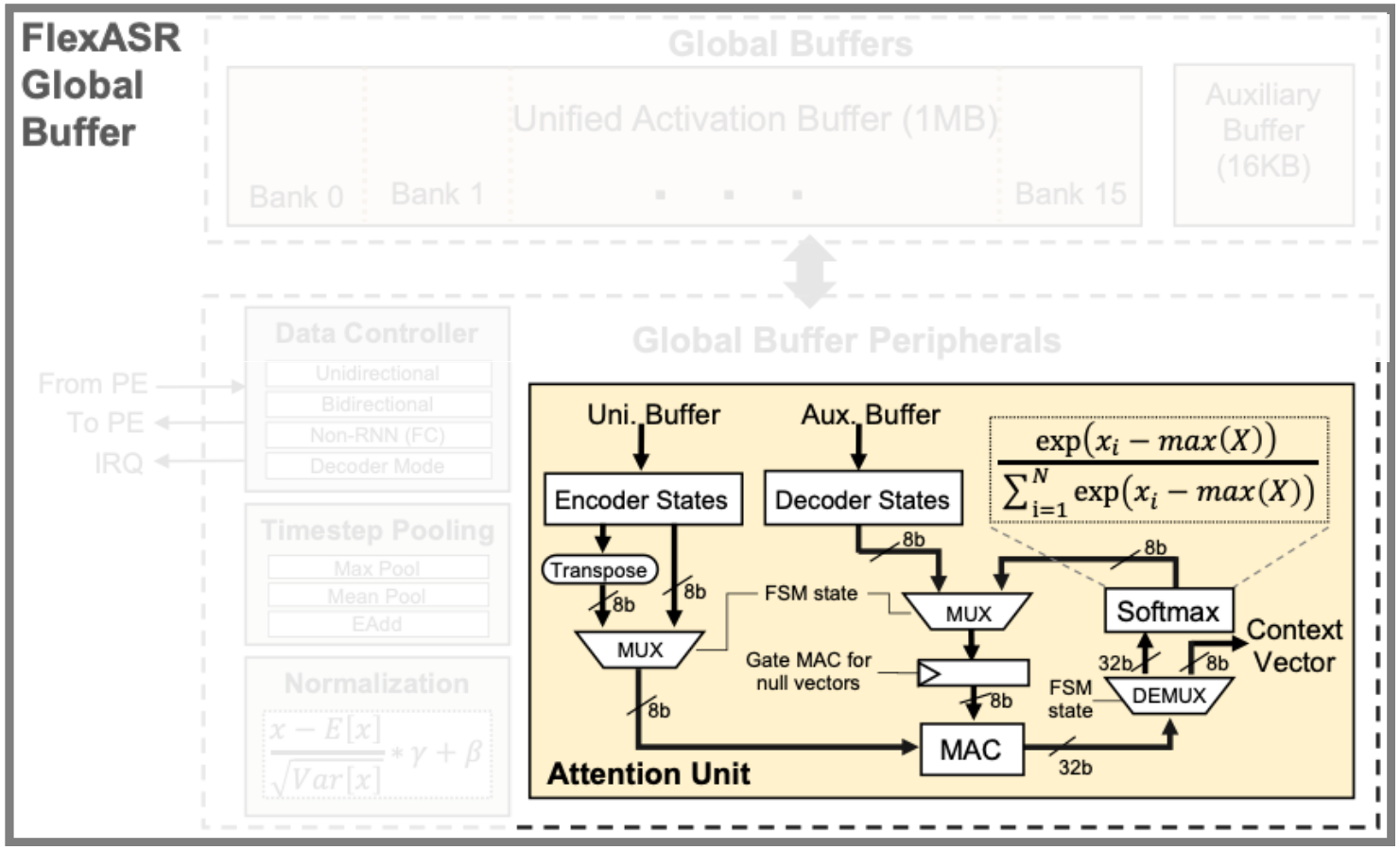
- **Global Buffer (GB) collects and unifies partial RNN outputs from PEs**

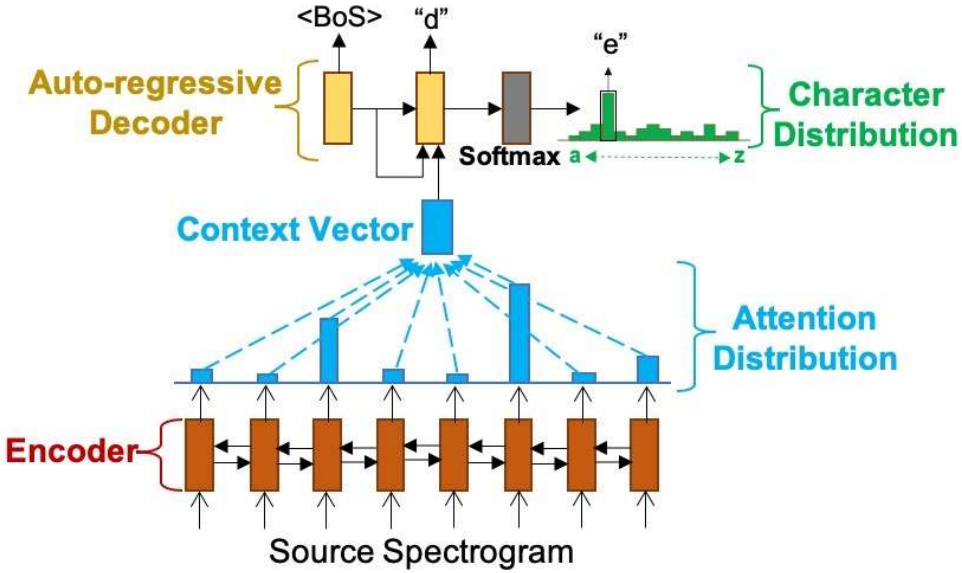




FlexASR architecture highlighting its global buffer unit

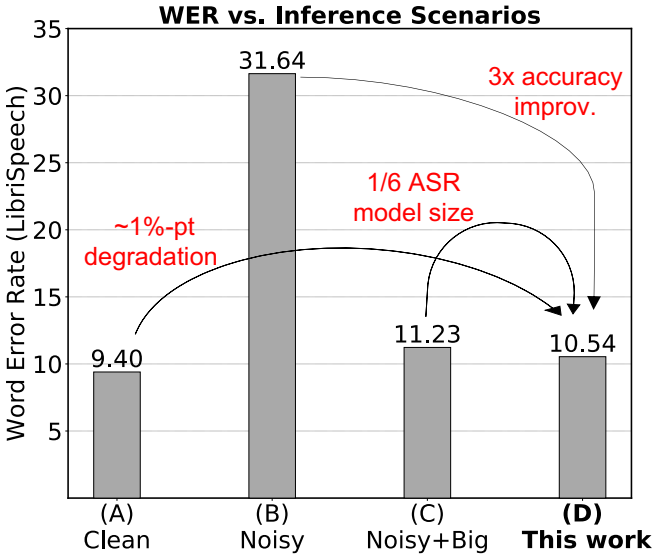
FlexNLP GB: Attention Mechanism

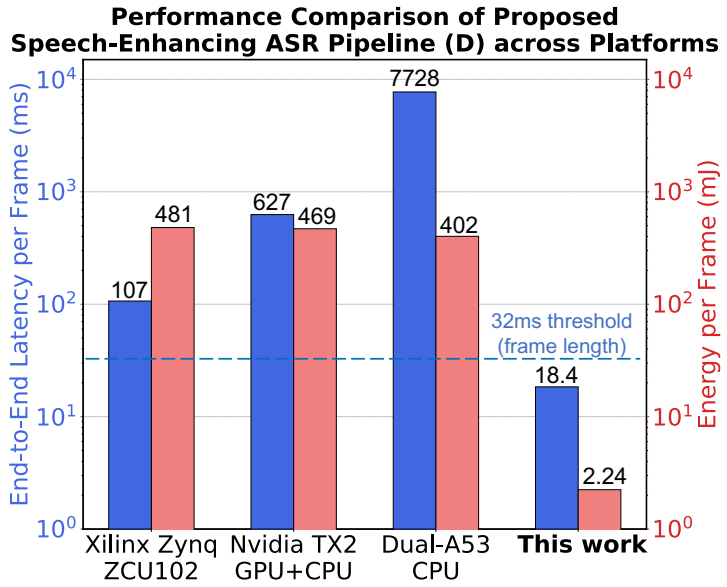


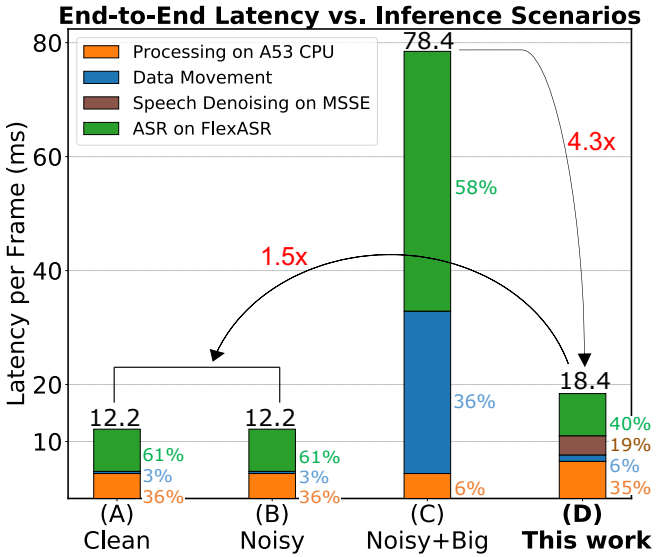


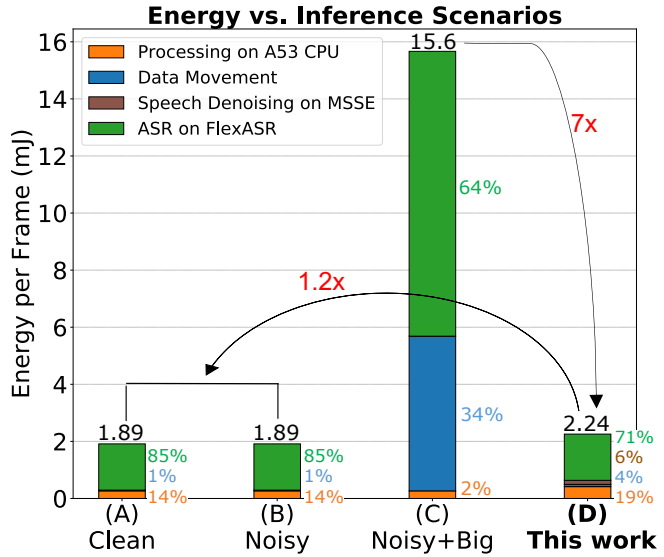
- **Attention mechanism**
  - **Computes numerically stable version of Softmax**
  - **MAC operations are skipped for null decoder states**
    - **Saves energy**

Performance Benefits





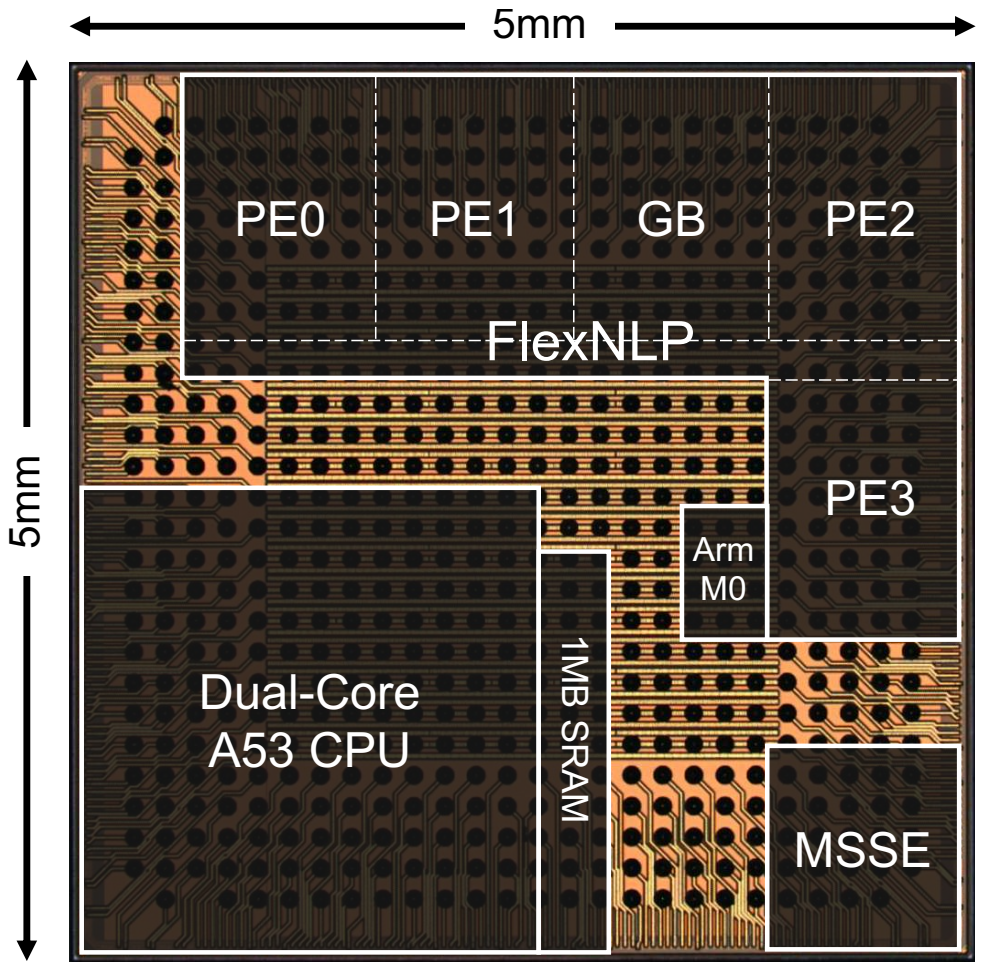




(A) with noiseless audio of the speaker
(B) with noise mixed with the speaker's voice at 0.9dB SNR
(C) with a much larger ASR model (22MB) trained with a noise-corrupted LibriSpeech dataset in order to accommodate noisy inputs
(D) with the proposed pipeline using Bayesian speech enhancement to separate the noise from the speaker's voice

By denoising incoming signals prior to speech recognition, MSSE allows FlexNLP to store a much smaller ASR model (1/6x), which obviates the very inefficient strategy of scaling up the DNN model (C) in order to achieve noise robustness

16nm Chip Summary



	MSSE	FlexNLP	Dual-A53
Workloads	Markov Random Field	Attention-based ASR models	FFT, Application Logic
Data Type	FxP32	FP8	FP64
Area	1.31mm^2	8.84mm^2	6.21mm^2
SRAM	0.103MB	5.03MB	2.41MB
Voltage	0.55 - 1V	0.55 - 1V	0.55 - 1V
Frequency	287 – 651MHz	130 – 573MHz	354 – 775MHz
Power @ Fmax/0.8V	42.2mW	219mW	50.4mW

Technology	TSMC 16nm FFC
Area	25mm^2
Total SRAM	9.8MB
Gate Count	11M
Clock Domains	6
Power Domains	5
Supply Voltage	0.55 – 1V
Packaging	Flip-chip BGA-672

Thank You

- FlexNLP HW architecture and simulator is publicly released at https://github.com/harvard-acc/FlexASR
- FlexNLP leveraged several SystemC/C++ IPs from MatchLib: https://github.com/NVlabs/matchlib
- Development and verification of this test chip leveraged several hardware IPs and tools from the CHIPKIT framework: https://github.com/whatmough/CHIPKIT
- This work is supported in part by JUMP ADA, DARPA CRAFT and DSSoC programs, Intel Corp., and Arm Inc.