ACCELERATING MULTI-USER LARGE VOCABULARY CONTINUOUS CPU-GPU PLATFORMS

Jungsuk Kim, Ian Lane
Carnegie Mellon University

MOTIVATION

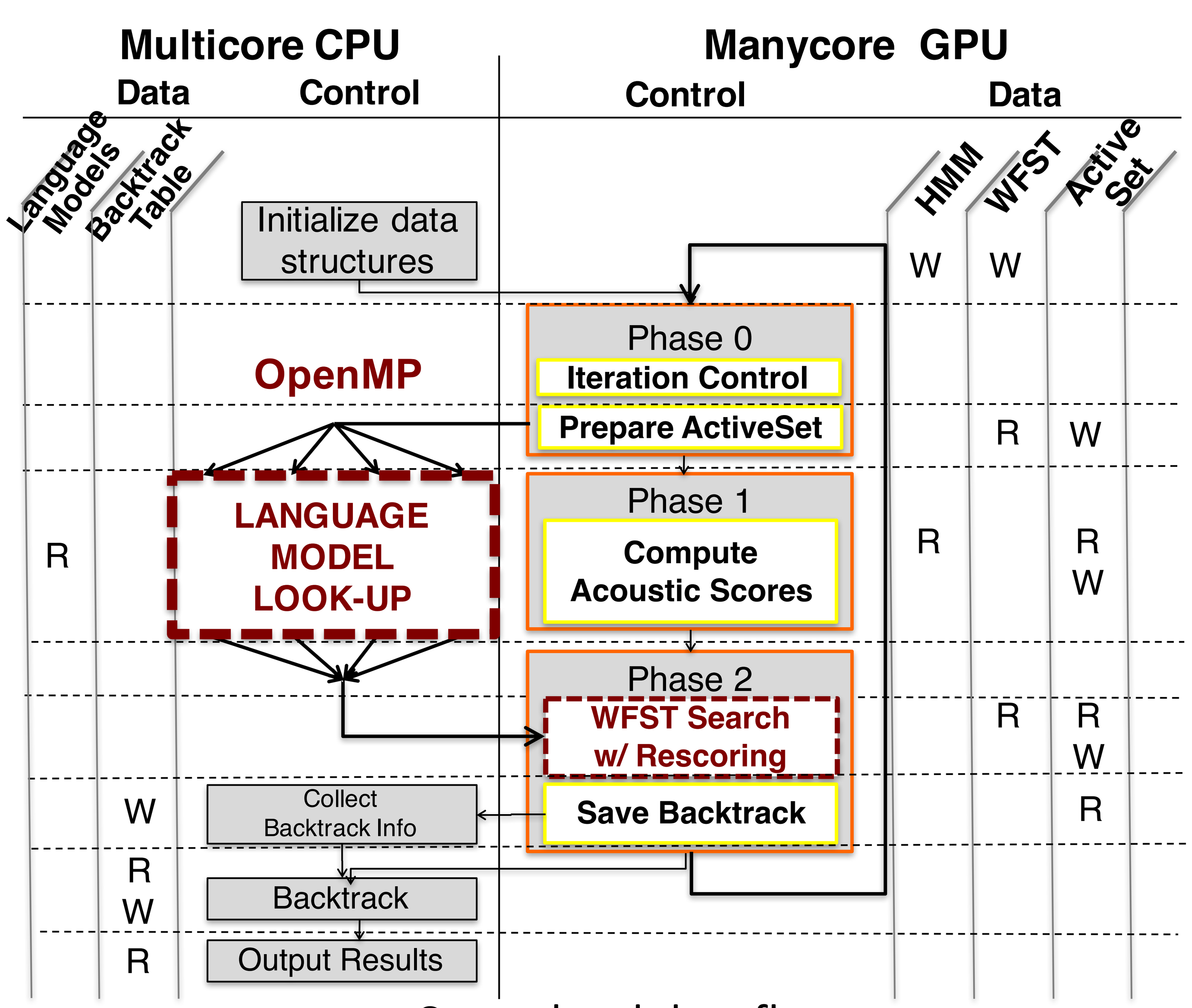
- **Modern *Distributed Speech Recognition* (DSR) system for real-time speech application should be:**
  - **ROBUST**
    - Large Acoustic Models
    - Large Language Models (> 1M words, > 20GB)
  - **RESPONSIVE**
    - GPU-accelerated (> 10X faster than real-time)
  - **EFFICIENT**
    - Support as many concurrent users as possible.

Previous Research:

Heterogeneous CPU-GPU speech recognition is as ***ROBUST*** as "state-of-the-art" lattice rescoring, but more than 22X ***RESPONSIVE***.

How can we make distributed speech recognition more ***Efficient***?

HETEROGENEOUS CPU-GPU LVCSR

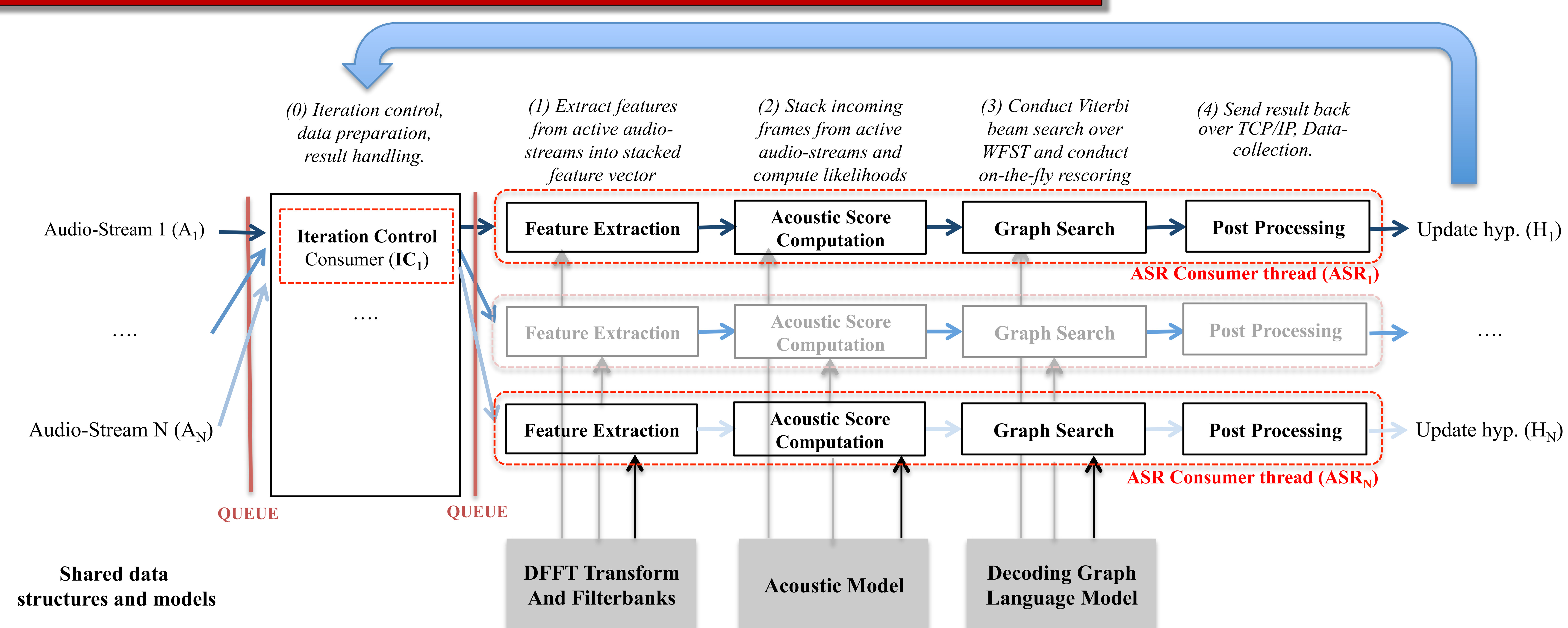


Control and data flow

Decoding Process

- **Prepare Active Hypotheses Set** (Phase 0)
- **Compute Acoustic Scores** (Phase 1)
  - ***On the GPU***, compute acoustic score for current input.
- **Language Model Look-up**
  - ***On the CPU***, compute ***likelihoods difference between large and small language models*** of active hypotheses.
- **WFST Search with Rescoring** (Phase 2)
  - ***On the GPU***, Frame synchronous ***N-best Viterbi search*** is performed on the GPU using WFST network composed with ***small language model.***
  - ***On the GPU***, Rescoring hypotheses ***"on-the-fly"*** using language model likelihood difference from CPU.

BASELINE SYSTEM ARCHITECTURE



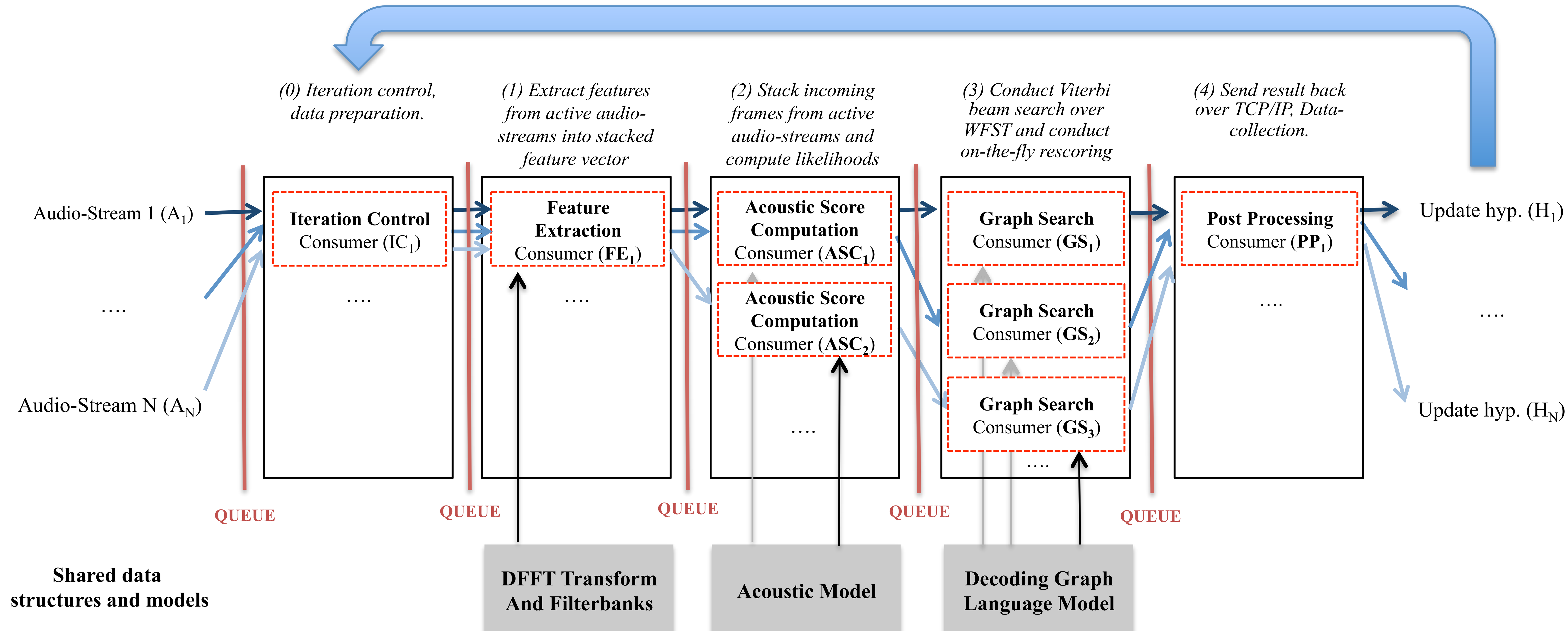
- **Pros.**
  - Simple thread management.

Not suitable for ***multi CPU + single GPU*** configuration

- **Cons.**
  - ***Low throughput and GPU utilization*** if audio stream batch size is small.
  - ***Server capacity limited*** by maximum number of inflight GPU kernels.
  - ***GPU is bottleneck*** due to sequentialization of tasks.

SPEECH RECOGNITON ON HETEROGENEOUS

PROPOSED SYSTEM ARCHITECTURE



- **Pros.**
  - ***Scalable and configurable*** structure.
  - Can assign more threads to bottleneck phase.
  - interleaving frames from different audio streams.
  - Can achieve ***maximum GPU utilization***.
- **Cons.**
  - Complex threads configuration.
  - More queuing overheads

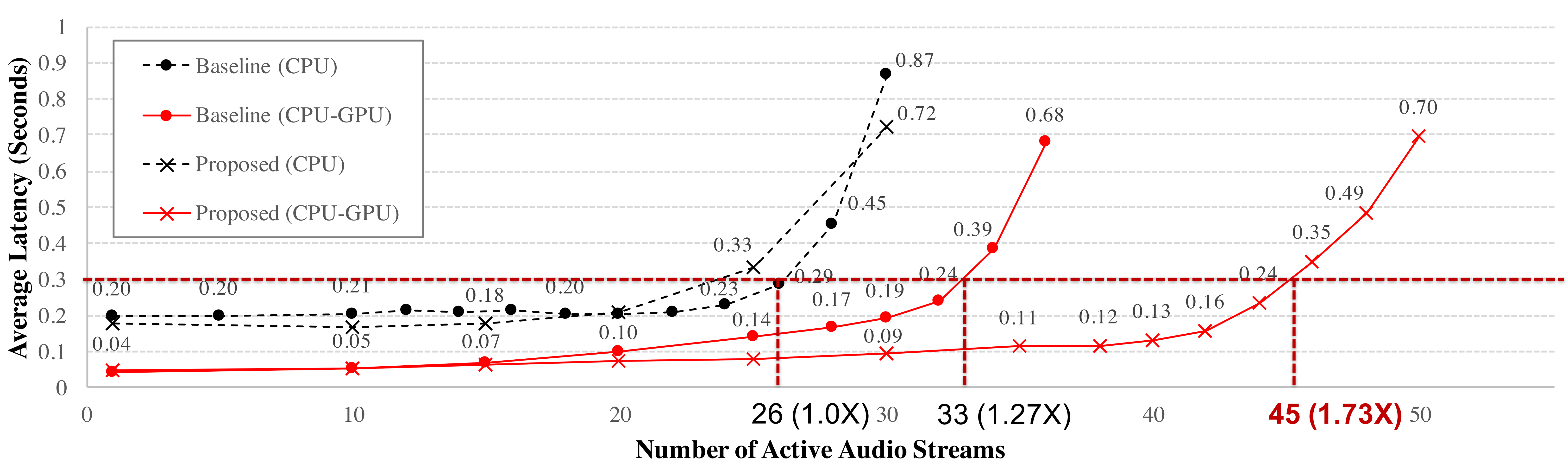
EVALUATION RESULTS

- **Evaluation Platform**
  - 2 *Intel Xeon E5-2697v3 @2.60GHz* = **14 cores** + 128GB DDR4
  - NVIDIA Titan X = **3072 CUDA cores @1.22GHz + 12GB GDDR5**
- **Model Specification**
  - Data set: Wall Street Journal + Web Data
  - Feature: 23th Filterbank coeff.
  - Hybrid DNN/HMM (5 hidden layers, ***22.7M*** parameters)

Vocab.	N-gram	# N-gram	Size (MB)	WFST (MB)
1M	3 (Pruned)	10.1M	407	3,583
	4	***769.9M***	***19,554***	-

- **Thread Configuration**

	CPU		CPU-GPU	
	Baseline	Proposed	Baseline	Proposed
# IC	2	1	2	1
# ASR	14 X 1	-	**2** X 8	-
# FE	-	2	-	2
# ASC	-	10	-	***1***
# GS	-	4 X 1	-	**2** X 8
# PP	-	2	-	1
Total	**16**	**19**	**18**	**22**



- ***"Proposed (CPU-GPU)"*** approach handles ***45 active real-time audio streams*** at an *average latency of 0.3 seconds.* (***73% more*** than CPU baseline, 36% more than GPU baseline)

Proposed CPU-GPU heterogeneous architecture is ***ROBUST, RESPONSIVE*** and ***73% more EFFICIENT*** than *"state of the art"* CPU baseline.

This work was partially supported in part by Samsung GRO program