GPU-Accelerated Large Vocabulary Continuous Speech Recognition for Scalable Distributed Speech Recognition

Jungsuk Kim Ian Lane

Electrical and Computer Engineering

Carnegie Mellon University

March 20, 2015 @GTC2015

Overview

- **Introduction**
- **Background**
  - *Weighted Finite State Transducers in Speech Recognition*
- **Proposed Approach**
  - *GPU-Accelerated scalable DSR*
- **Evaluation**
- **Conclusion**

Introduction

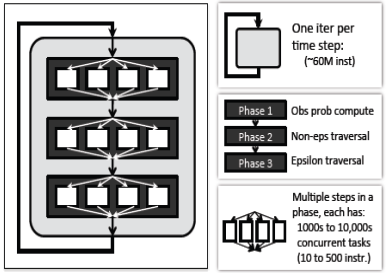
- Voice interfaces a core technology for User Interaction
  - Mobile devices, Smart TVs, In-Vehicle Systems, …
- For a captivating User Experience, Voice UI must be:
  - **Robust**
    - Acoustic robustness → **Large Acoustic Models**
    - Linguistics robustness → **Large Vocabulary Recognition**
  - **Responsive**
    - Low latency → **Faster than real-time search**
  - **Adaptive**
    - User and Task adaptation

Introduction

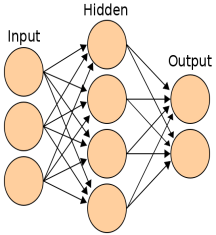
- Large models critical for accurate speech recognition:
  - Large acoustic models ➔ **Tens of Millions of parameters**
  - Large vocabulary ➔ **Millions of words**
  - Large language model ➔ **Billions of n-gram entries (>= 20GB)**

  - Examples include:
    - Acoustic modeling for telephony [Mass 2014] or Youtube [Bacchiani 2014]
      - **~200M** parameter Deep Neural Networks
    - Language model rescoring for Voice Search [Schalkwyk 2010]
      - 1.2M vocabulary, 5-gram LM, **12.7B** n-gram entries

Introduction



Speech recognition
contains many highly parallel tasks

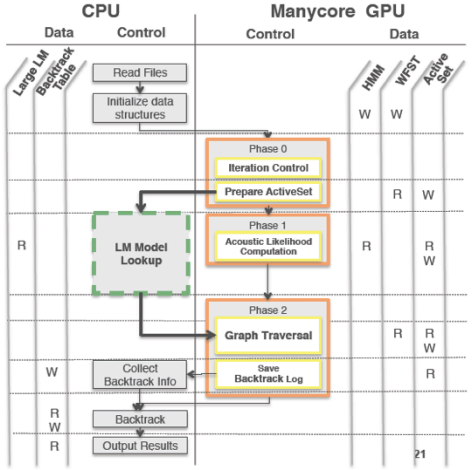


Large Models
More Accurate

\+

Graphic Processing Units
(SIMT, ~3000 cores, <24GB)
optimized for parallel computing

=



ASR engine designed specifically for GPUs

Introduction

- **1 Million** Vocabulary (3-gram)
- **30 Million** parameter Deep Neural Network

Tesla K40
Kepler, 2880 cores

Titan X
Maxwell, 3072 cores

Tegra K1
Kepler, 192 cores

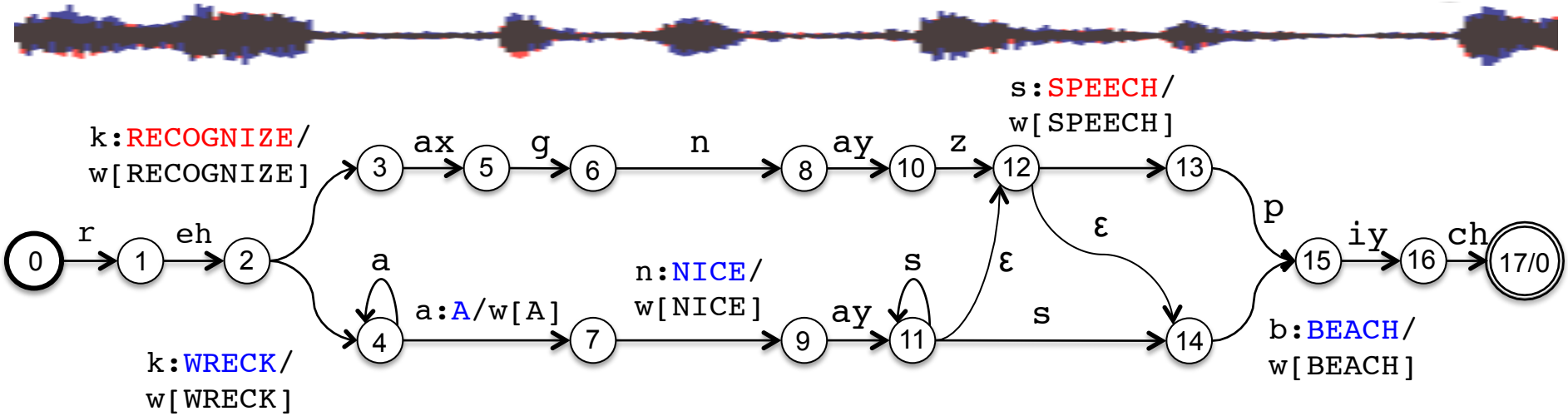
Tegra X1
Maxwell, 256 cores

RTF	0.02	0.01	0.17	0.14
xRT	50X	100X	6X	7X
1hour	72s	***36s***	612s	***504s***

Background

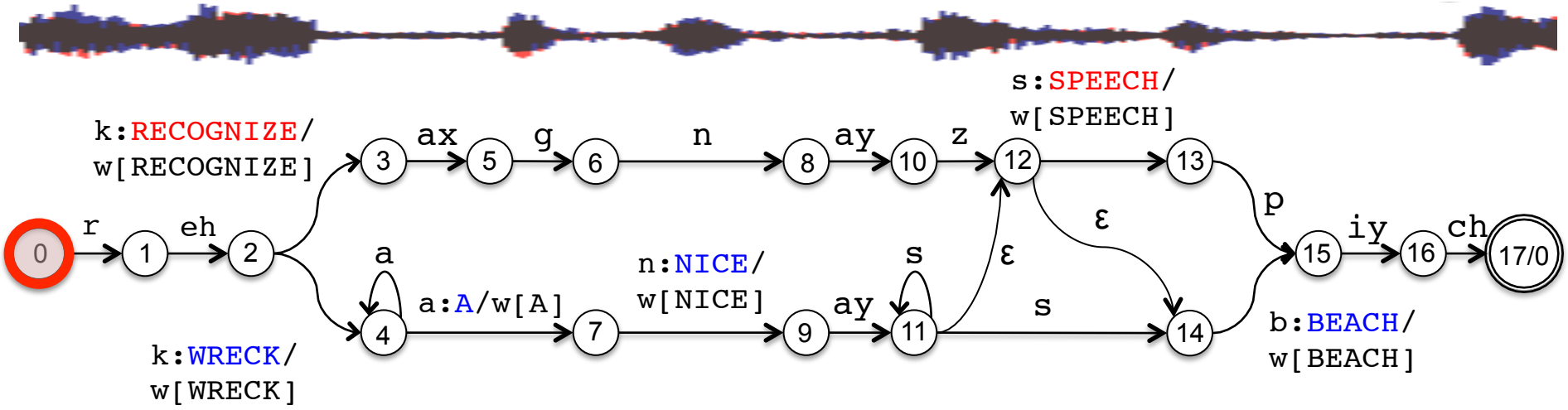
Weighted Finite State Transducers (WFSTs) in Speech Recognition

WFST in Speech Recognition



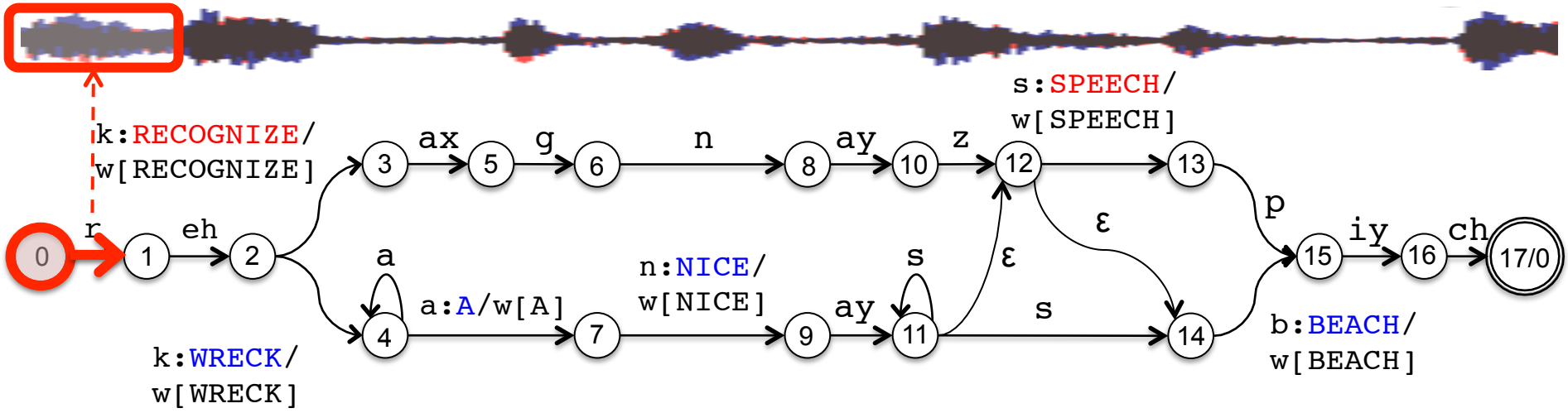
- "***Recognize speech***" v.s. "***Wreck a nice beach***"...
- Search is performed in 3 phases.
  - **Phase 0**: Active Set Preparation.
  - **Phase 1**: Acoustic Score Computation.
  - **Phase 2**: WFST Search.

WFST in Speech Recognition



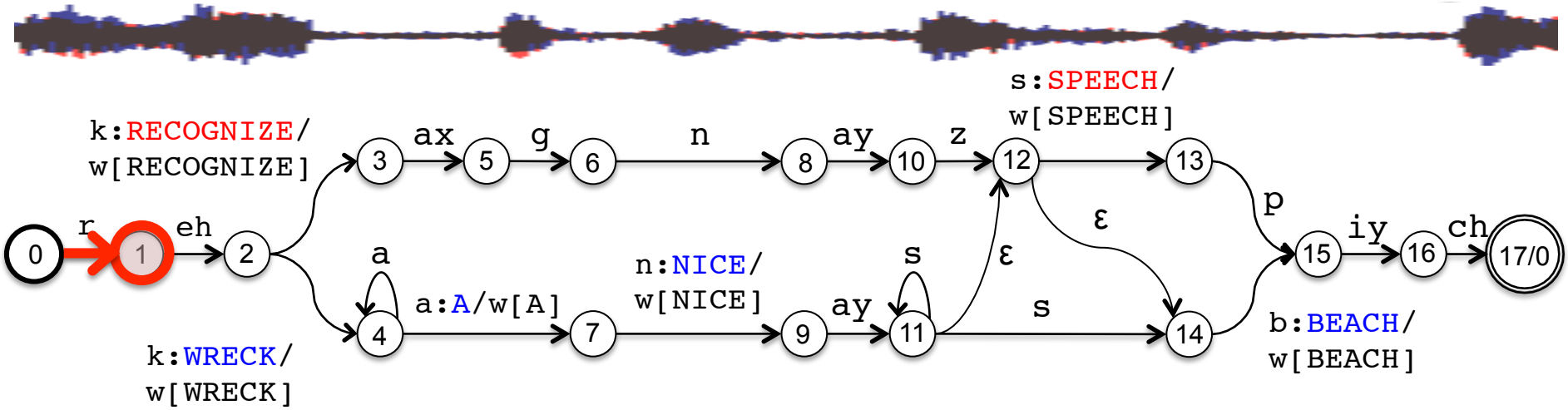
- **Phase 0:** ***Active Set Preparation***
  - Collect active hypotheses from previous frame.

WFST in Speech Recognition



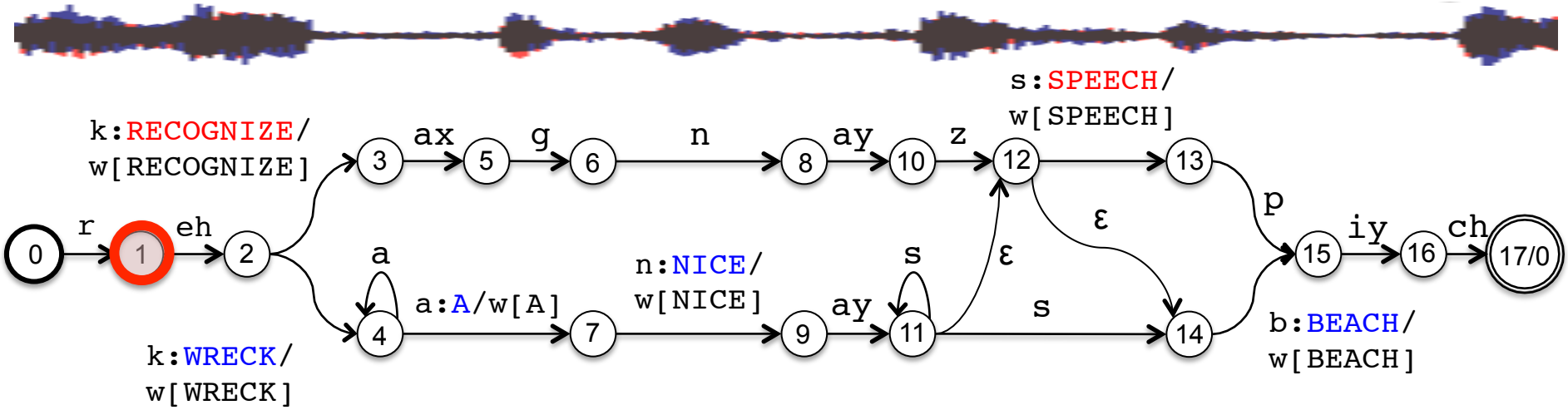
- **Phase 1:** ***Acoustic Score Computation***
  - Compute acoustic similarity between given speech and phonetic models using Deep Neural Network

WFST in Speech Recognition



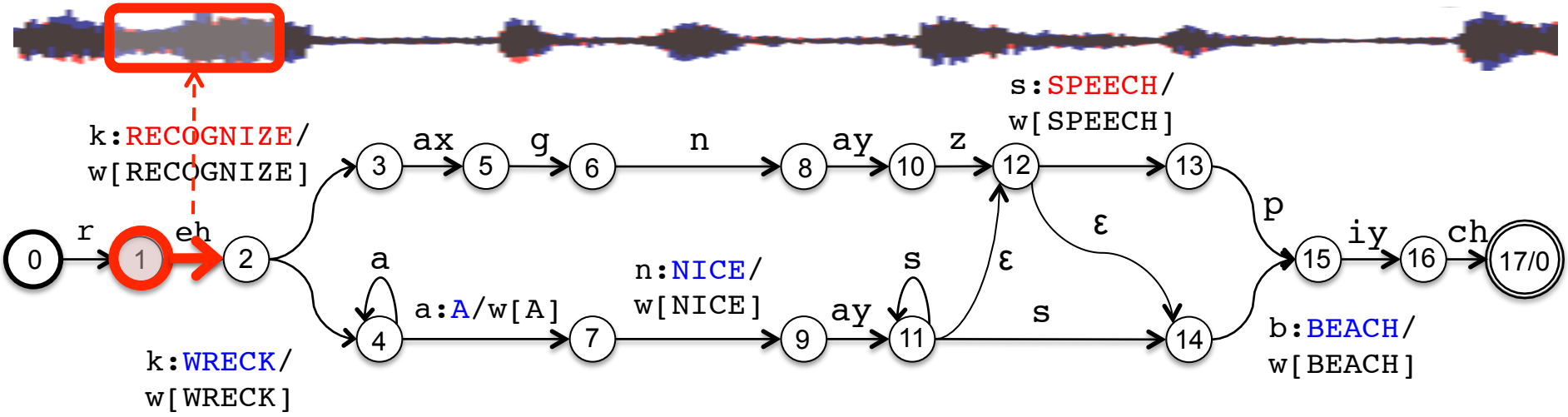
- **Phase 2: *WFST Search***
  - Perform frame synchronous Viterbi beam search on WFST network.
  - If multiple transitions have same next state *s*, then *the most likely (minimum score) hypothesis is retained* (i.e. state 12, 14, 15…)

WFST in Speech Recognition



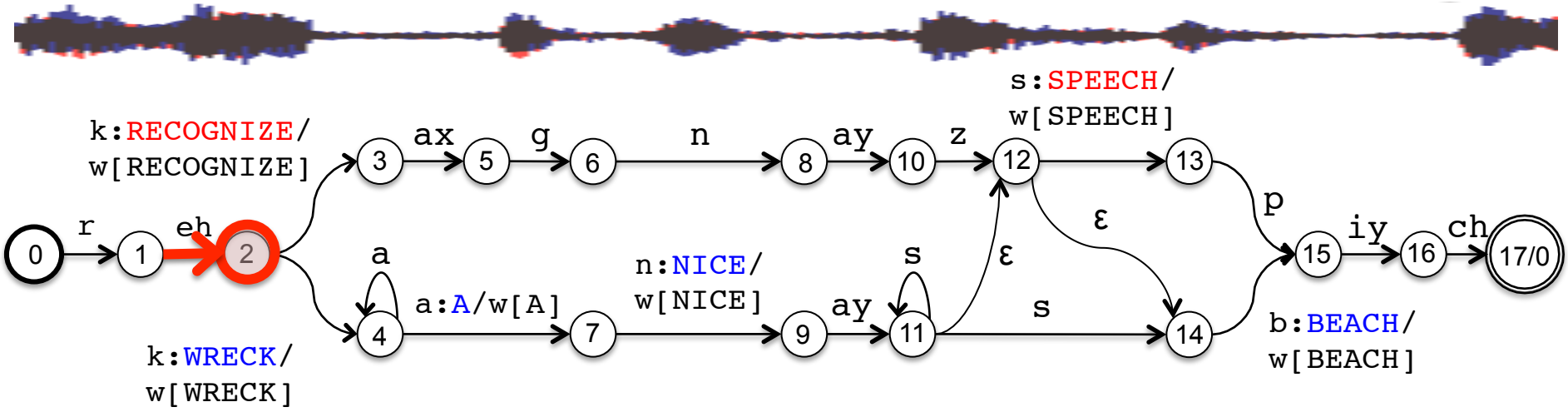
- Iterate these 3 phases until input audio ends.
- **Phase 0:** ***Active Set Preparation***

WFST in Speech Recognition



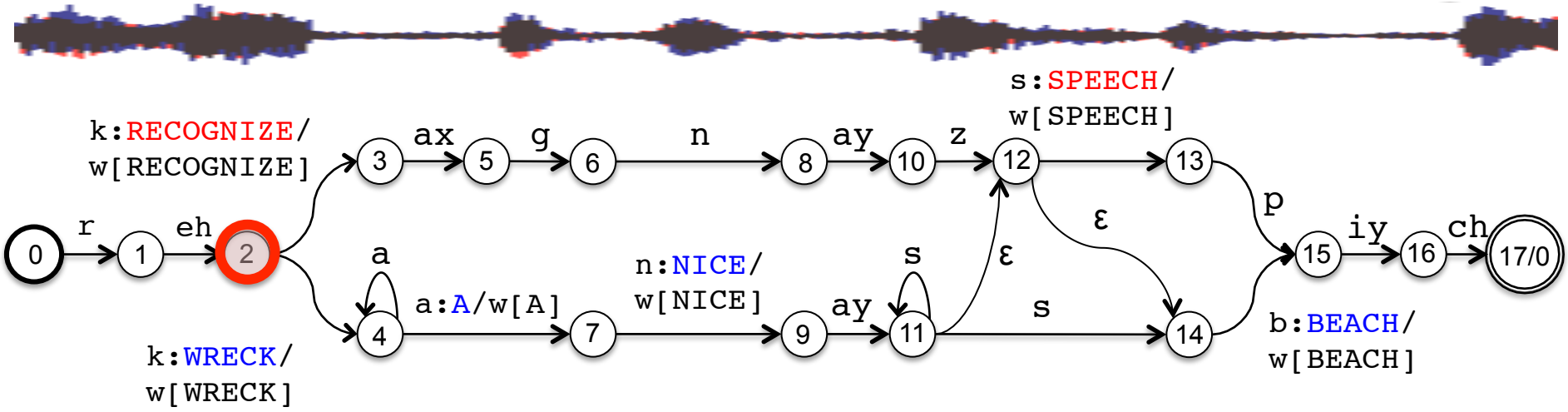
- **Phase 1:** ***Acoustic Score Computation***

WFST in Speech Recognition



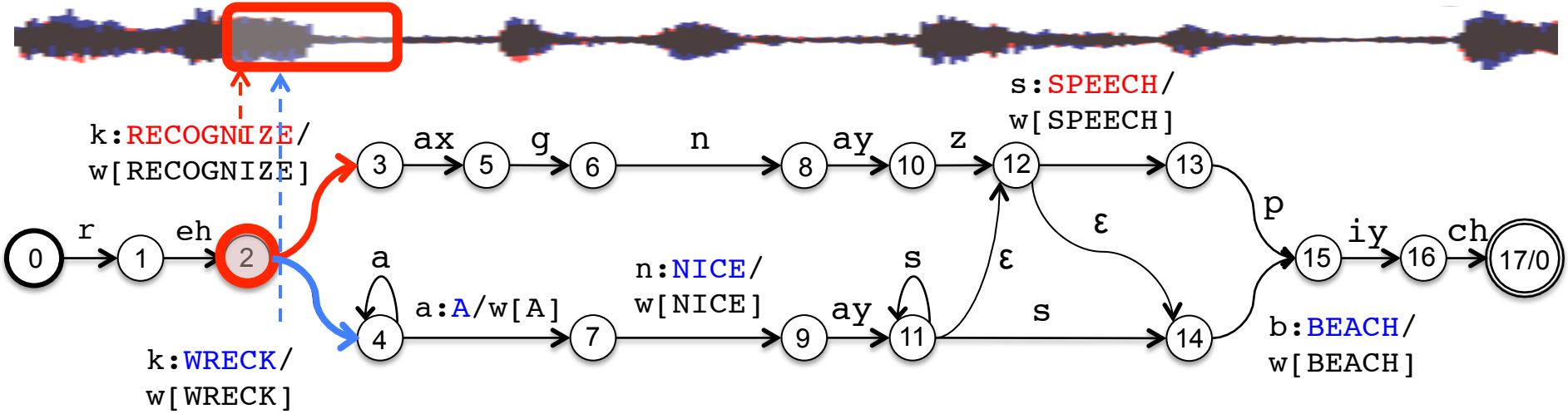
- **Phase 2:** ***WFST Search***

WFST in Speech Recognition



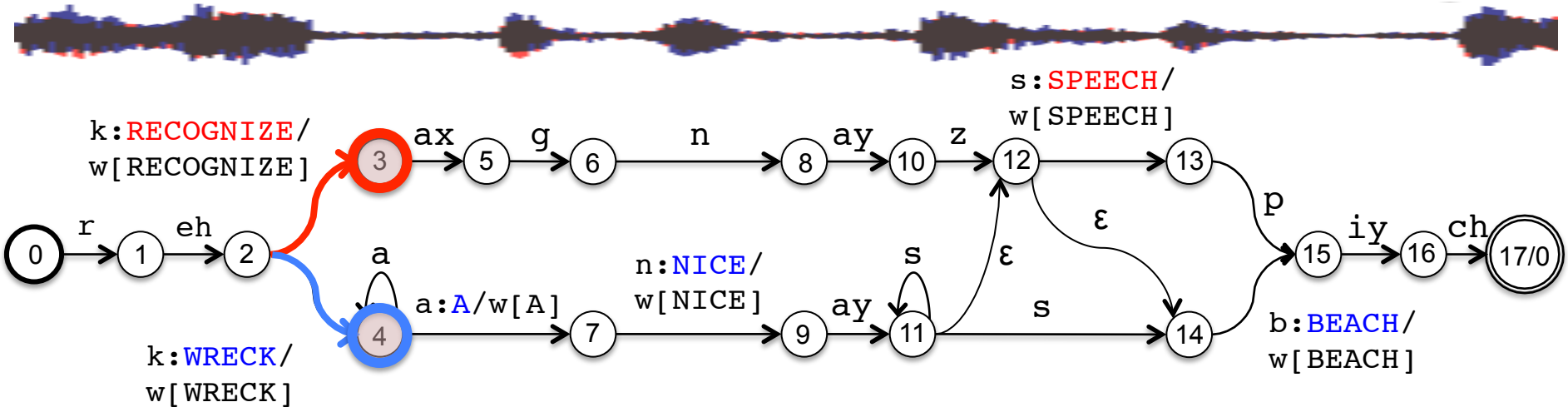
- **Phase 0:** ***Active Set Preparation***

WFST in Speech Recognition



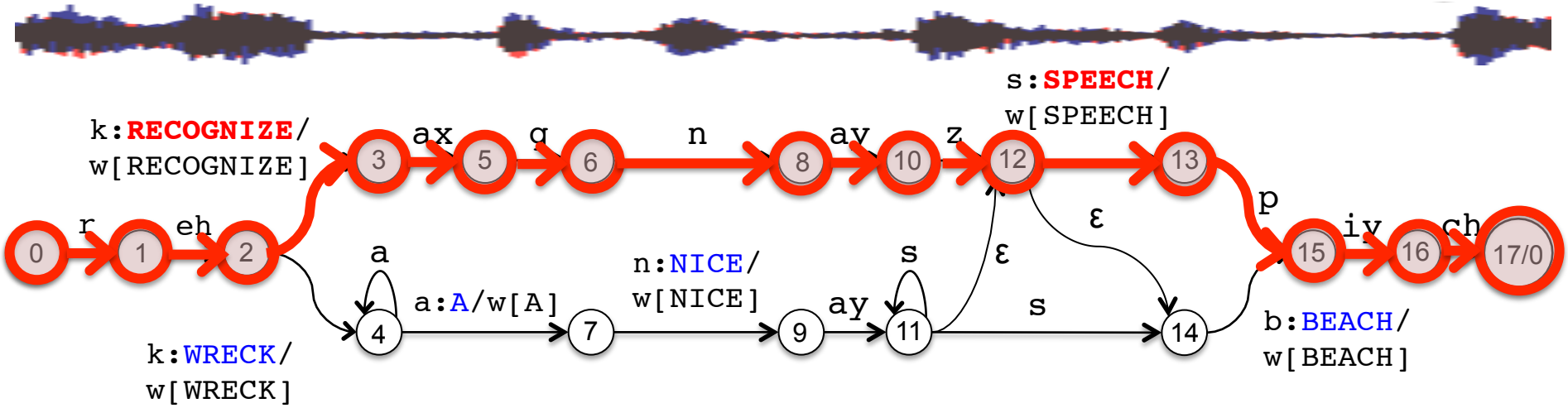
- **Phase 1:** ***Acoustic Score Computation***

WFST in Speech Recognition



- **Phase 2:** ***WFST Search***

WFST in Speech Recognition

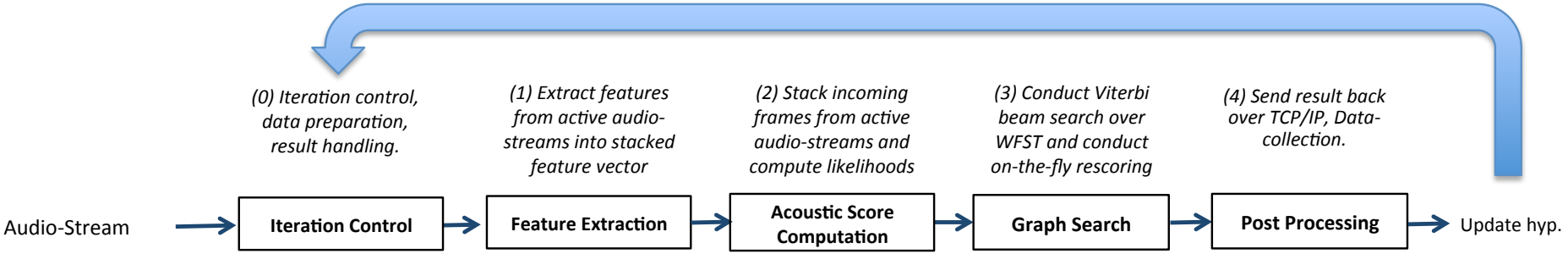


- Recognized result is an output symbol sequence over the best path.
  - **Result: "RECOGNIZE SPEECH"**

Proposed Approach

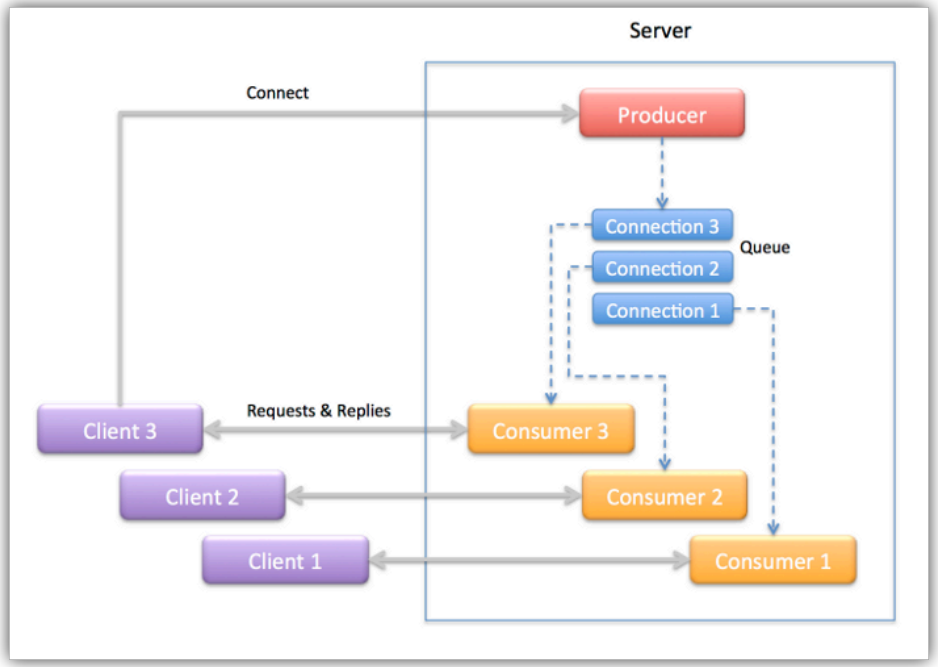
GPU-Accelerated Scalable DSR

Distributed Speech Recognition (DSR)



- ***Iteration control***
  - Allocate or deallocate data structures.
  - Terminate decoding task.
- ***Feature extraction***
  - Receive audio and extract feature for current iteration (batch).
  - Speaker dependent adaptation.
- ***Acoustic score computation***
  - Deep Neural Network (Forward Propagation).
- ***Graph search***
  - Conduct frame synchronous WFST search.
  - End-of-utterance detection.
- ***Post processing***
  - Output (Lattice) processing.
  - Sending result back to client.

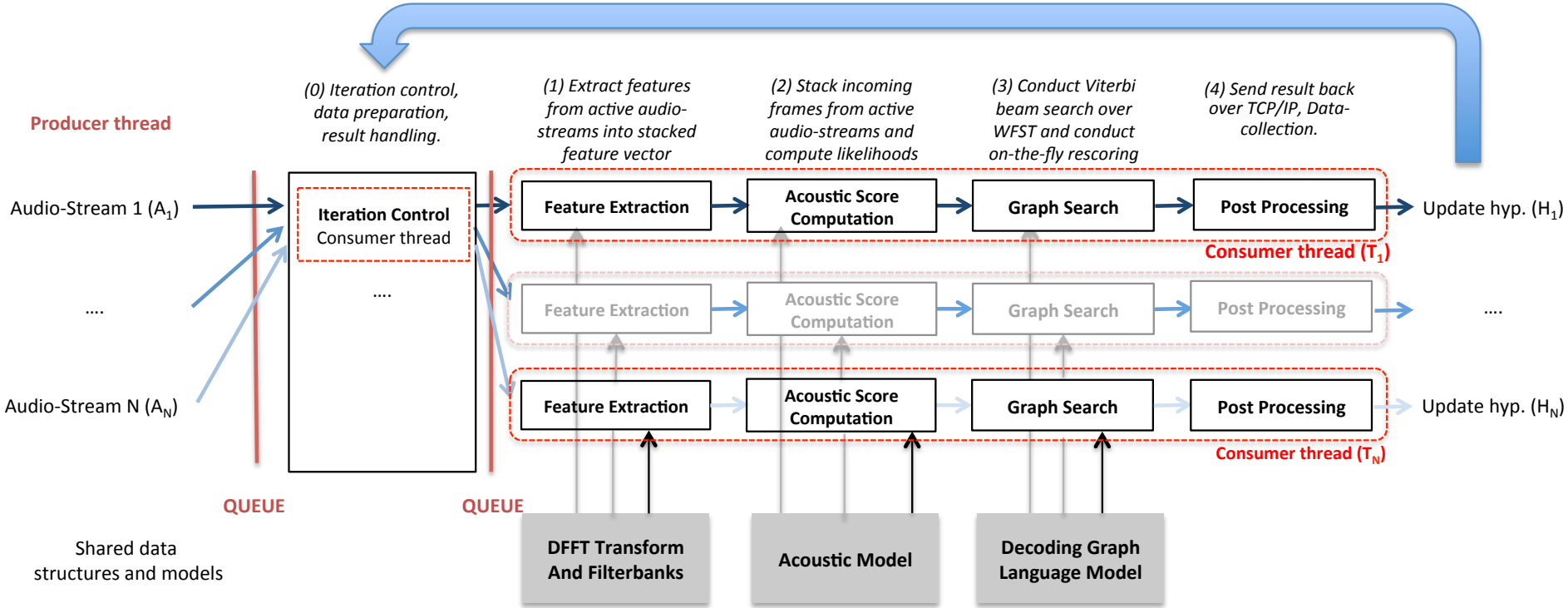
Producer/Consumer design pattern



Producer-Consumer multi-threaded model

- Master/Slave pattern.
- Decuple processes that produce and consume data at different rates.
- **Advantages**:
  - Enhanced data sharing
  - Processes can run in different speeds.
  - Buffered communication between processes.

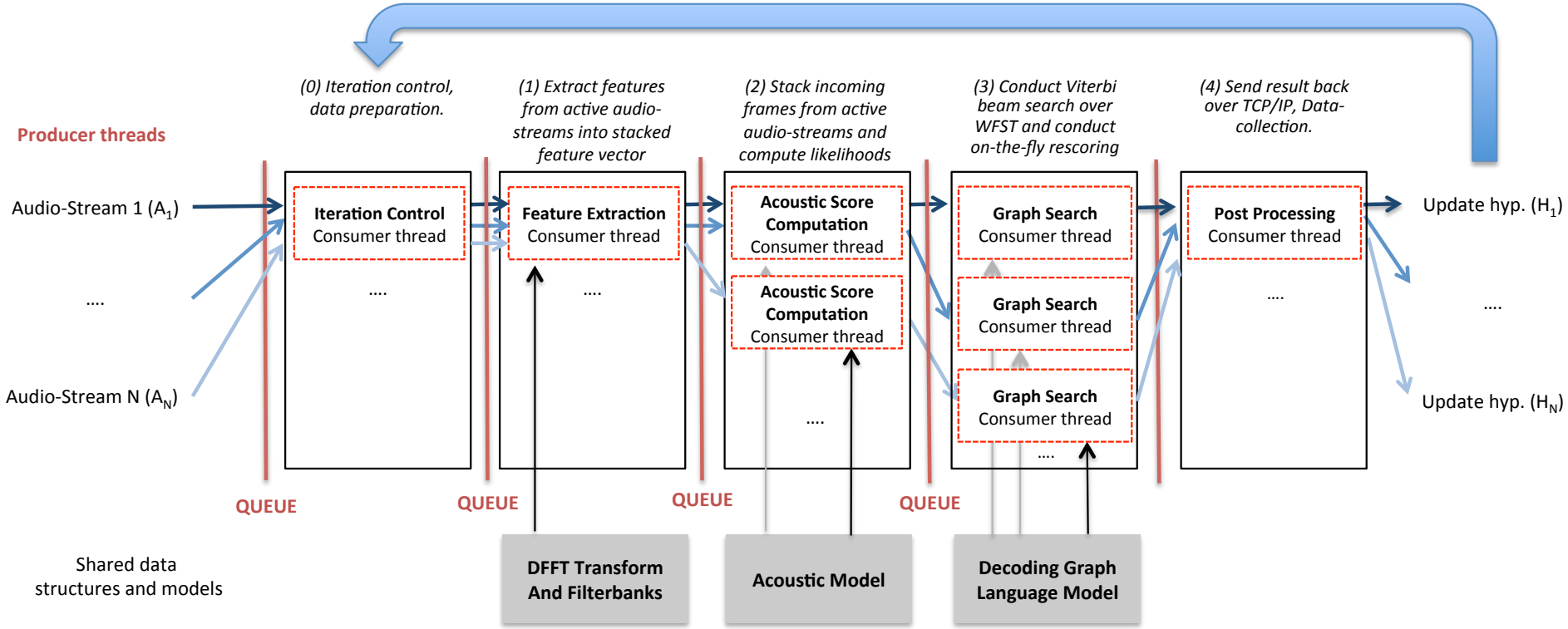
Architecture 1 (Naïve)



Architecture 1 (Naïve)

- **Pros.**
  - *Maximum decoding performance*.
  - Simple thread management.
- **Cons.**
  - Low throughput and GPU utilization if batch size is small.
  - Number of consumer threads can be limited by GPU (by maximum inflight kernels)
  - Not suitable for many CPU + single GPU configuration.

Architecture 2



Architecture 2

- **Pros.**
  - *More scalable and configurable* structure.
  - Can assign more threads to bottleneck phase.
  - interleaving frames from multiple tasks.
  - Can achieve maximum utilization of GPU.
- **Cons.**
  - Complex threads configuration.
  - More queuing overheads
  - Expected relatively *higher latency* compared to “structure 1”

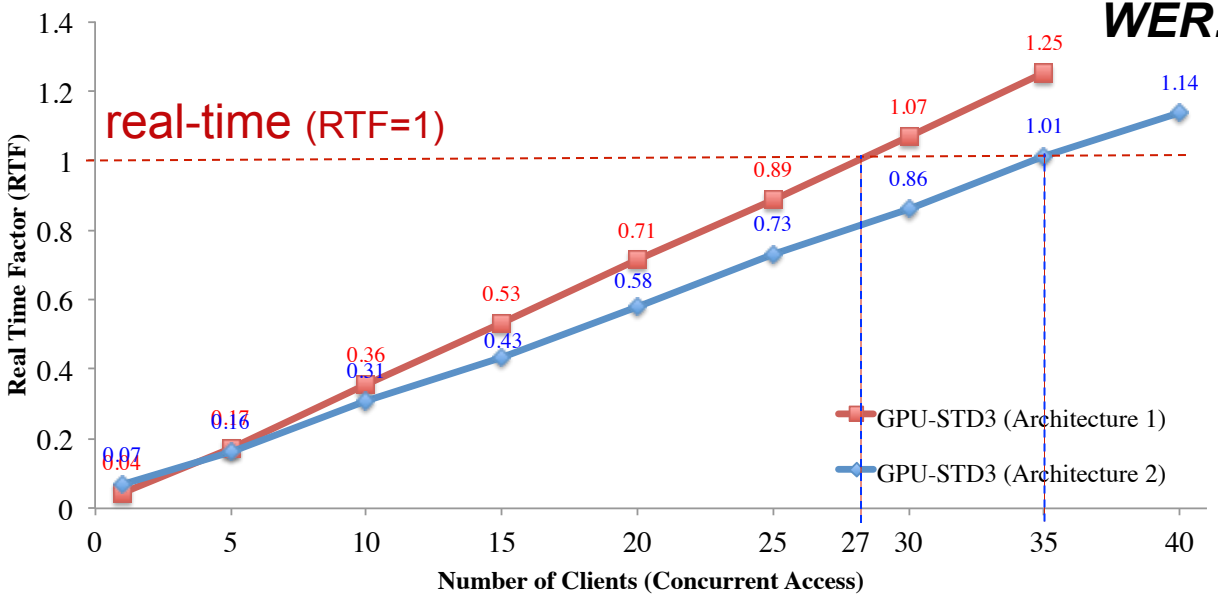
Evaluation Results

GPU-Accelerated Scalable DSR

Evaluation Setup

- **Language Model:**
  - 1 Million Vocab. 3-gram (10.1M n-gram)
- **Acoustic Model:**
  - **DNN**: (in) 253 X 2048 X 2048 X 2048 X 2048 X 2048 X 3432 (out)
- **Feature type:**
  - 23th Filterbank coefficient with CMVN
- **Evaluation Set:**
  - WSJ eval92 (20K, 333 utts.)
- **Platform**:
  - Core i7-2600K + NVIDIA Tesla K40

Evaluation Results

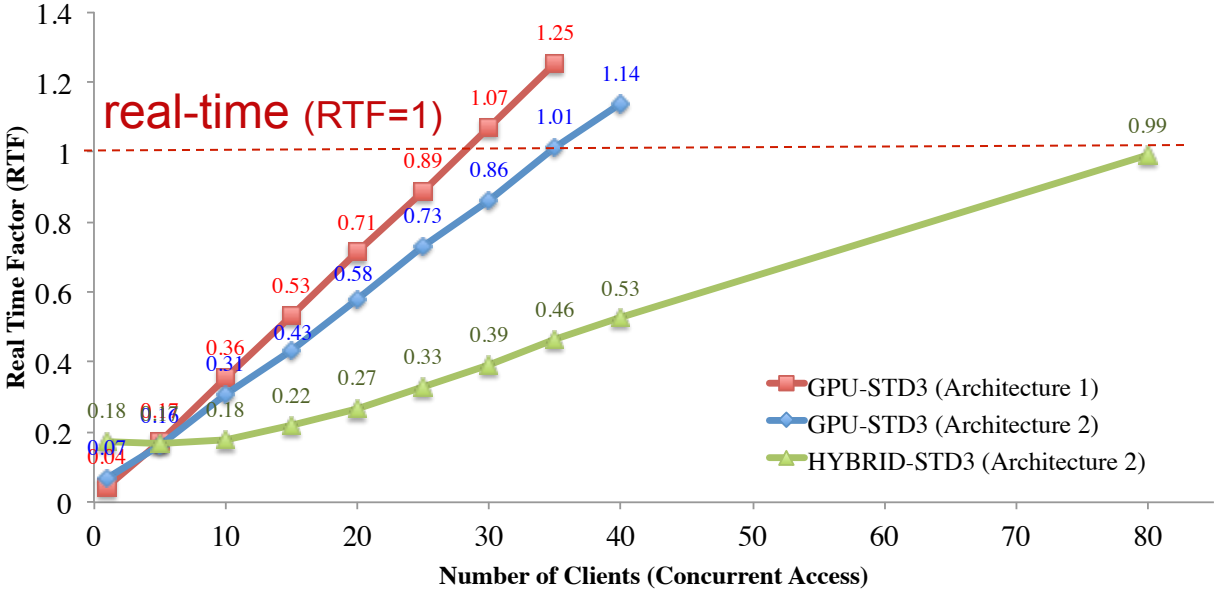


	G1	G2
Iteration control	1	1
ASR decoder	2*	0
Feature extraction	0	1
Acoustic score comp.	0	1*
Graph search	0	2*
Post processing	0	2

* use GPU

- ***GPU only configuration*** (G1, G2): 1 Tesla K40.
- Architecture 2 improves speed by ***0.24 RTF*** (N=35)
- "***Architecture 2***" processes ***35*** concurrent audio streams in real-time.

Evaluation Results



	G1	G2	H2
Iteration control	1	1	1
ASR decoder	2	0	0
Feature extraction	0	1	2
Acoustic score comp.	0	1*	1*
Graph search	0	2*	12
Post processing	0	1	2

* use GPU

- ***Hybrid configuration*** (H2): 1 GPU + 2 CPU (16 cores).
- "***Architecture 2***" processes ***80*** concurrent audio streams in real-time.

Conclusion

GPU-Accelerated Scalable DSR

Conclusions

- Proposed ***scalable*** and ***configurable*** DSR server architecture**.**
- "***Architecture 2***" was able to process …
  - ***40*** concurrent audio streams in real-time with 1 GPU (K40c)
  - ***80*** concurrent audio streams in real-time with 1 GPU + 16 CPU cores.
- Performance can be improved further
  - Lock-free task queue.
  - Optimal / Adaptive Thread configuration.
  - Smart task scheduling.

References

References

[Mass, 2010] Andrew L. Maas, Awni Y. Hannun, Christopher T. Lengerich, Peng Qi, Daniel Jurafsky, and Andrew Y. Ng. " Increasing Deep Neural Network Acoustic Model Size for Large Vocabulary Continuous Speech Recognition", ArXiv: 1406.7806 [cs.CL], 2010

[Schalkwyk, 2010] J. Schalkwyk, D. Beeferman, F.Beaufays, B. Byrne, C.chelba, M. Cohen, M. Kamvar, and B. Stropek, *Google Search by Voice*: A case study, Springer, 2010

[Bacchiani, 2014] M. Bacchiani, David Rybach, "*Context Dependent State Tying For Speech Recognition Using Deep Neural Network Acoustic Models*," in *Proc.* ICASSP, 2014, pp.230-234.

[Mohri, 2002] M. Mohri, F. Pereira, and M. Riley, "Weighted Finite-State Transducers in Speech Recognition," *Computer Speech and Language*, vol. 16, no. 1, pp. 69-88, 2002

[Kanthak, 2002] S. Kanthak, H. Ney, M. Riley, and M. Mohri. A comparison of two LVR search optimization techniques. In Proc. ICSLP, pp. 1309-1312, 2002.

[Chong, 2009] J. Chong, E. Gonina, Y. Yi, and K. Keutzer, "A Fully Data Parallel WFST-based Large Vocabulary Continuous Speech Recognition on a Graphix Processing Unit," in *Proc. Interspeech*, Sep. 2009, pp. 1183-1186.

[Ljolje, 1999] A. Ljolje, F. Pereira, and M. Riley, "Efficient general lattice generation and rescoring," in *Proc. Eurospeech*, 1999

[Hori, 2007] T.Hori,C.Hori,Y.Minami,andA.Nakamura,"EfficientWFST-BasedOne- Pass Decoding With On-The-Fly Hypothesis Rescoring in Extremely Large Vocabulary Continuous Speech Recognition," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 4, pp. 1352 –1365, may 2007.

Q&A

Thank you for your attention.