Exploring Recognition Network Representations for Efficient Speech Inference on Highly Parallel Platforms
Jike Chong, Ekaterina Gonina, Kisun You, Kurt Keutzer, Department of Electrical Engineering and Computer Science, University of California, Berkeley

**Maturing Highly Parallel Platforms**
- Architecture trend:
  - Increasing vector unit width
  - Increasing numbers of cores per die
- Maturing HW architecture:
  - Including caches as well as local stores that benefit irregular accesses

Ongoing work investigates performance of alternative approaches to speech recognition on these highly parallel platforms.

**Speech Recognition Inference Engine Characteristics**
- Parallel graph traversal through Recognition network
- Guided by a sequence of input audio vectors
- Computing on continuously changing data working set

**Implementation Architecture**

**Two Recognition Network Representations**

- **LLM Network**
  - Chain of triphone states for each pronunciation
  - Each chain constructed using a separate copy of triphone states – many duplications
  - Evaluate possibility of transition from one word to all other words at the end of each triphone chain

- **WFST Network**
  - FSM of composed pronunciation and language models
  - Across-word transitions explicitly represented
  - Encapsulates large amount of information with little redundancy
  - Fewer tokens required to be maintained for target accuracy

**Wall Street Journal 1 Corpus**
- Based on a 5,000 word vocabulary, 1,050,282 bigrams (291,116 pruned)
- 3,000 16-mixture acoustic models, 39 dim features based on 13 dim MFCC
- WFST network is an HCLG model compiled and optimized offline

**GTX285 Results**
- 22x transitions evaluated
- 1.17x Execution speed measured at 0.09s WER

**Conclusions**
- Simpler LLM network representation performs competitively with highly optimized WFST representation
- WFST representation is a more concise representation requiring traversal of 1/22 number of state transitions to achieve the same accuracy
- Per state transition LLM gathers data 53-65x faster and evaluates transitions 4.7-6.4x faster than WFST
- Uncoalesced memory accesses are still a major bottleneck in implementations using the WFST representation

Emergence of highly parallel platforms brings forth an opportunity to reevaluate computational efficiency of speech recognition approaches.