A Hybrid GPU+CPU Speech Recognition Engine

- For intuitive Voice and Interactive Multimodal systems robust and responsive speech recognition is crucial
  - Robust
    - Acoustic robustness: Large Acoustic Models
    - Linguistic robustness: Large Vocabulary (1M+ words)
  - Responsive
    - Low latency: Faster than real-time search
    - Current state-of-the-art speech recognition systems are optimized for either robustness or responsiveness
      - Robustness: 5-10 x real-time >95% accuracy
      - Responsiveness: real-time 85% accuracy
- Models (>20GB)
- Current state
- Language Model
  - 95.40%
  - 330 sentences
  - Nov. 92 ARPA WSJ test set

Linguistic robustness
- Accuracy improvement converges with large N

Acoustic robustness
- Accuracy improves when maintaining more number of N

Hybrid GPU/CPU
- Compute likelihood of phonetic models (Gaussian Mixture Model) for 3000 tied state
- Evaluation Set
- Low latency

Experimental Evaluation
- Acoustic Model
  - 51-284 Data Set
  - 3000 tied state
  - 16 mixture Gaussians
  - 39th MFCCs features
- Language Model
  - Wall Street Journal 5k
    - 1-gram: 5k entries
    - 2-gram: 1.6M entries
    - 3-gram: 2.7M entries
- Evaluation Set
  - Nov. 92 ARPA WSJ test set
  - 330 sentences
- NVIDIA GTX 680
  - Kepler architecture
  - 1536 CUDA cores

- 20x speed-up compared to standard WFST decoding on CPU at word accuracy of 93.80%
  - 95.40% maximum accuracy is achieved.
- Accuracy improves when maintaining more number of N best hypothesis.
- Accuracy improvement converges with large N.

On-The-Fly Partial Hypothesis Rescoring

- How can we decode with large models in real-time?
  - Use hybrid GPU/CPU architectures
  - Perform “On-The-Fly Partial Hypothesis Rescoring”

GPU
- Recognition Network (WFST)
  - Acoustic Model
  - Pronunciation Model
  - Language Model (uni-gram)
- Viterbi Search
  - Observation Probability Calculation
  - WFST Search

CPU
- Feature Extraction
  - Backtrack
- Language Model (n-gram)
- Output

Decoding Process
- Prepare Active Hypotheses Set
  - Gather active speech recognition hypotheses (word and phone sequences) from previous frame.
- Compute Observation Probabilities
  - Compute likelihood of phonetic models (Gaussian Mixture Model) for current input feature.
- On-The-Fly Partial Hypothesis Rescoring
  - On the CPU, rescoring likelihoods of partial hypotheses using a higher order N-gram language model stored in main memory.
  - Partial Hypothesis rescoring and the observation probability computation can be performed concurrently.
- WFST Search
  - Frame synchronous Viterbi search is performed on the GPU using WFST network composed using unigram language model.
  - Maintaining N-best paths during decoding to ensure good hypotheses are not pruned early.

Relationship between word accuracy and Real Time Factor (RTF)

- Accuracy improves when maintaining more number of N best hypothesis.
- Accuracy improvement converges with large N.
Network size does not significantly affect decoding.

Resume the partial hypothesis using likelihood difference.

Language Model

CPU function and GPU kernel can be conducted in parallel.

Choose only best hypothesis when multiple arcs meet in the same destination state.

Why "N-Best"?

Early pruning: Best hypothesis \( g_2 \) is pruned before the rescoring.

Load Balancing Between GPU and CPU using OpenMP

N-Best On-The-Fly Partial Hypothesis Rescoring

- Acoustic Model
  - All WSJ corpus
  - 10,000 tied state
  - 32 mixture Gaussians
  - 39th MFCCs features

- Language Model
  - 1M vocab.
  - 3-gram: 497.6M entries
  - 4-gram: 767.8M entries
  - 5-gram: 977.1M entries

- Evaluation Set
  - WSJ test set
  - 543 sentences

GPU and CPU Parallel Execution

- Language model look up has no data dependency between acoustic likelihood computation.
- CPU function and GPU kernel can be conducted in parallel.
- Language model runtime can be hid by behind GPU run time.

GPU and CPU load Balancing using OpenMP

- Language model look up is longer than acoustic likelihood computation time with small acoustic model
- Language model lookup for each hypothesis is independent.
- Language model lookup phase is parallleized using OpenMP on the CPU to achieve better load balance.

Experimental Evaluation

- 2.74x faster than realtime when the WER is 9.35%.
- 91.23% maximum accuracy is achieved.

1M vocab: network can be decoded on a modern GPU.
- Network size does not significantly affect decoding speed.