HYDRA A Hybrid CPU/GPU-based Speech Recognition Engine for Real-Time LVCSR

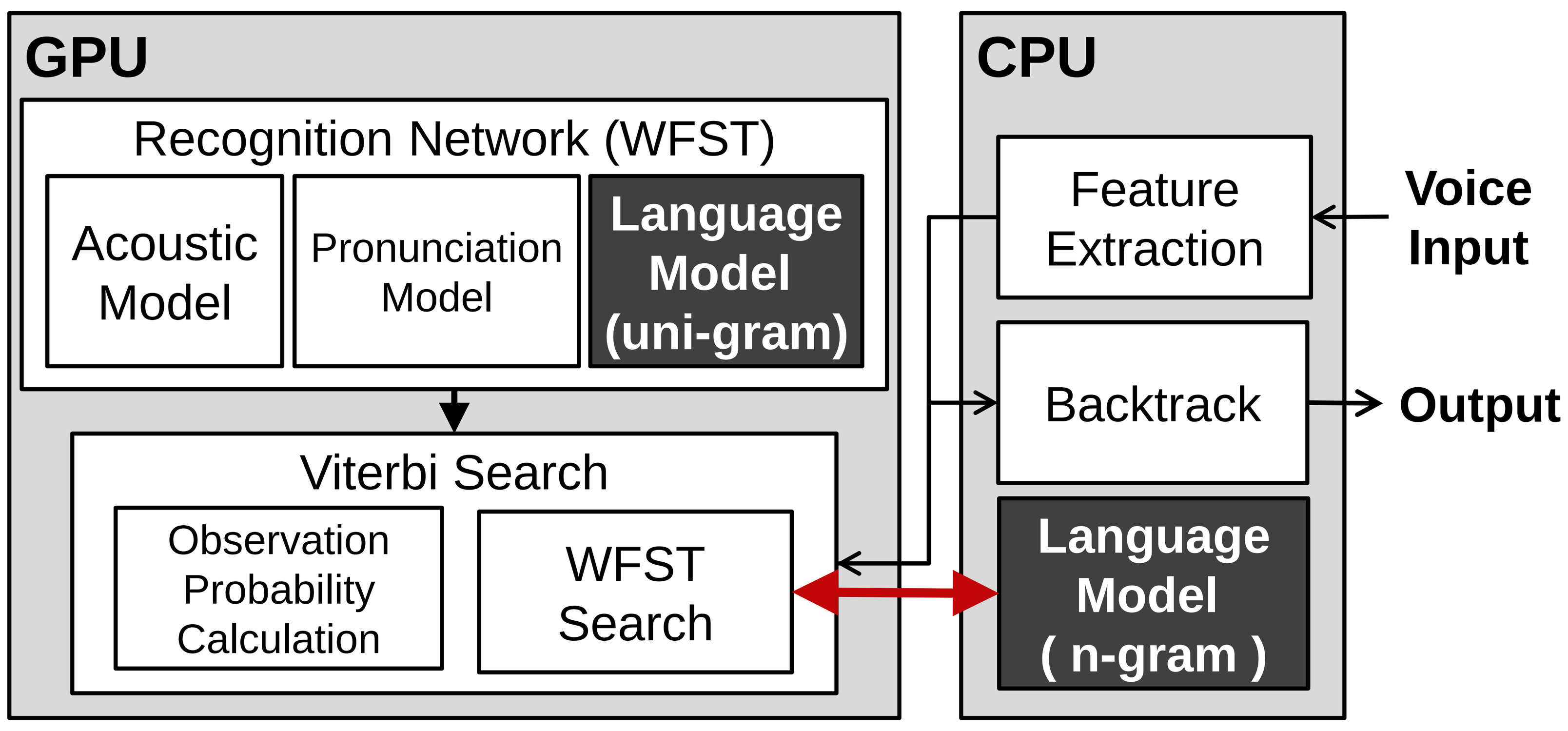
Jungsuk Kim, Jike Chong, Ian Lane

Carnegie Mellon University

A Hybrid GPU+CPU Speech Recognition Engine

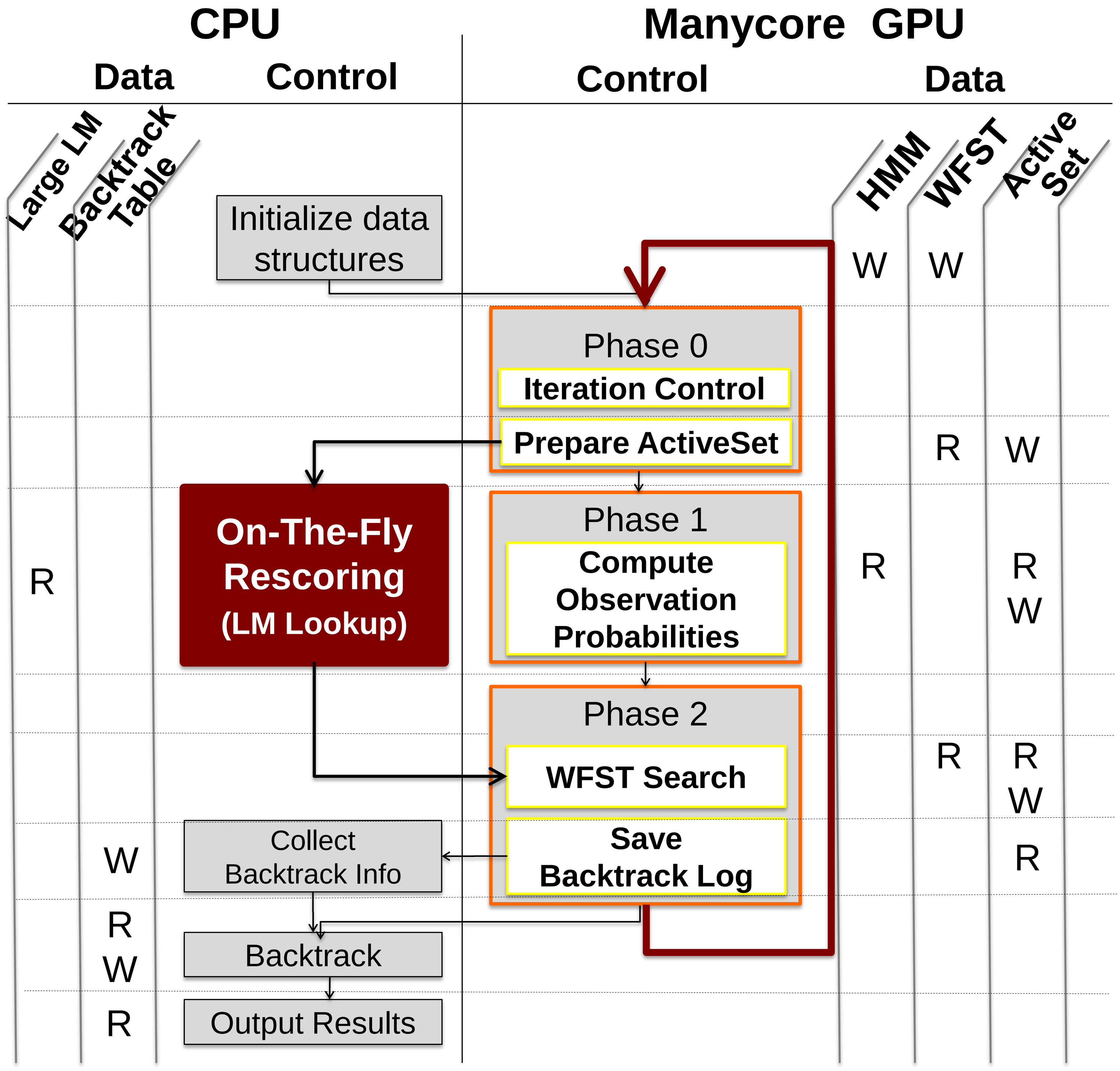
- **For intuitive Voice and Interactive Multimodal systems robust and responsive speech recognition is crucial**
  - **Robust**
    - Acoustic robustness ➔ **Large Acoustic Models**
    - Linguistic robustness ➔ **Large Vocabulary (1M+ words)**
    ➔ **Large Language Models (>20GB)**
  - **Responsive**
    - 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

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



GPU/CPU Hybrid LVCSR Engine architecture

On-The-Fly Partial Hypothesis Rescoring



Control and data flow for the proposed approach

Decoding Process

Prepare Active Hypotheses Set
- Gather active speech recognition hypotheses (word and phone sequences) from previous frame.

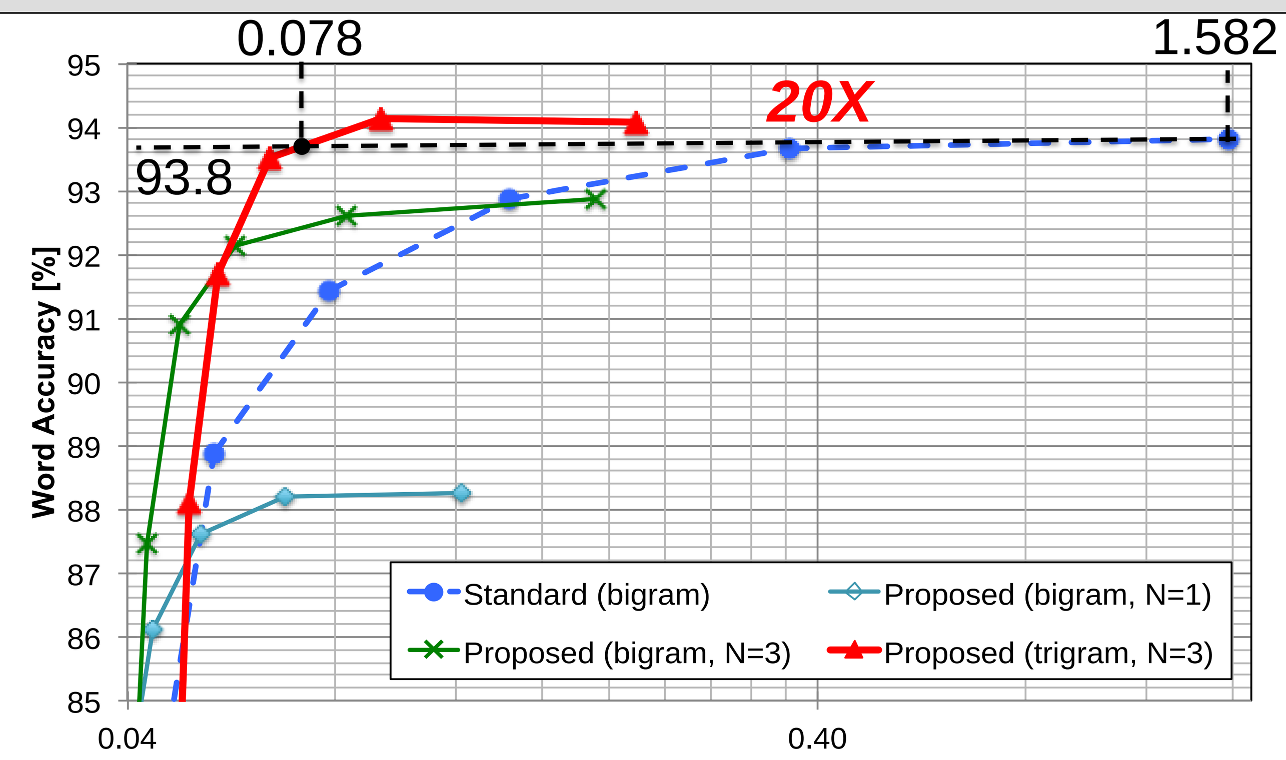
Compute Observation Probabilities
- Compute likelihood of phonetic models (Gaussian Mixture Model) for current input feature.

On-The-Fly Partial Hypothesis Rescoring
- **On the CPU,** rescore 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.

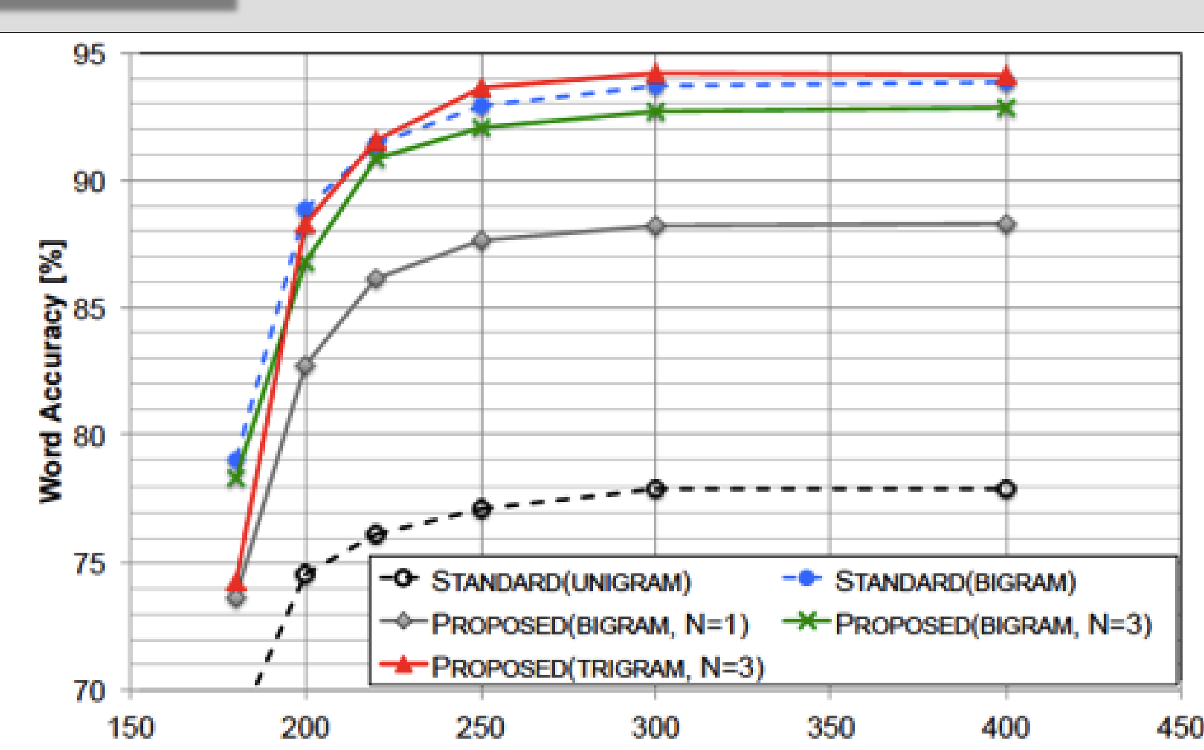
Experimental Evaluation

- Acoustic Model
  - SI-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
  - Keplar architecture
  - 1536 CUDA cores



Relationship between word accuracy and Real Time Factor (RTF)

- ***20x speed-up*** compared to standard WFST decoding on CPU at word accuracy of ***93.80%***
- ***95.40%*** maximum accuracy is achieved.

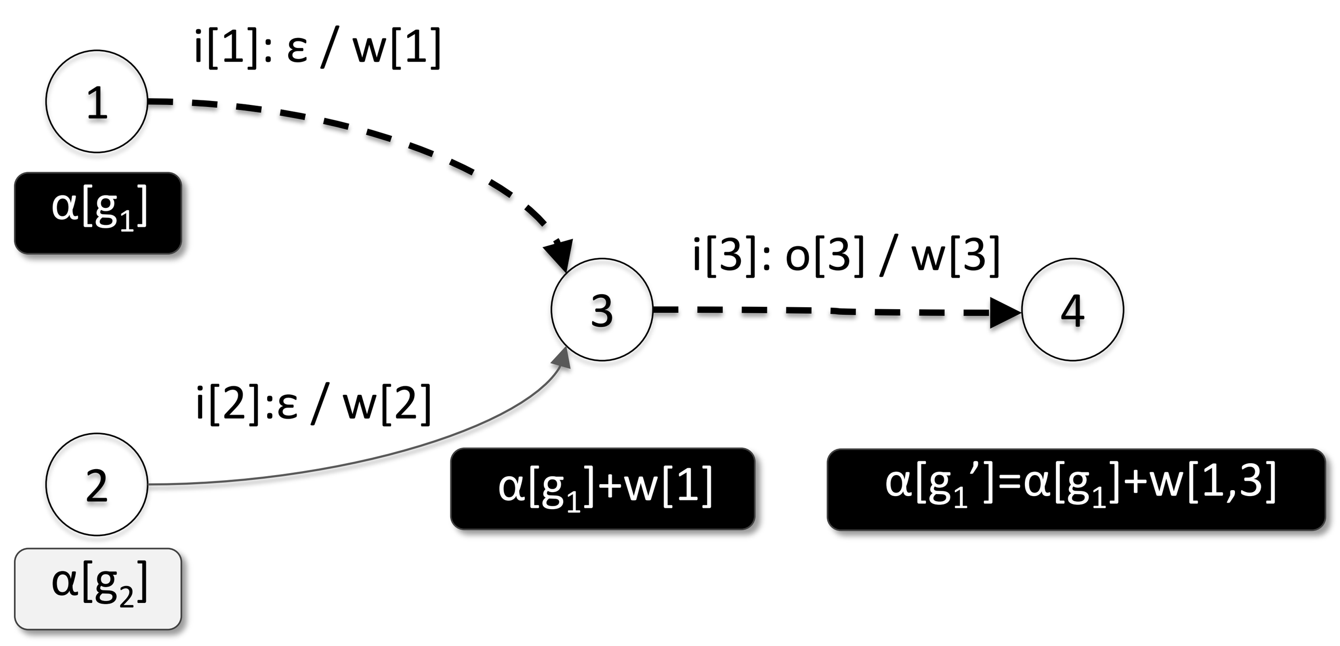


Relationship between word accuracy and beamwidth for maintaining different number of N-best hypothesis

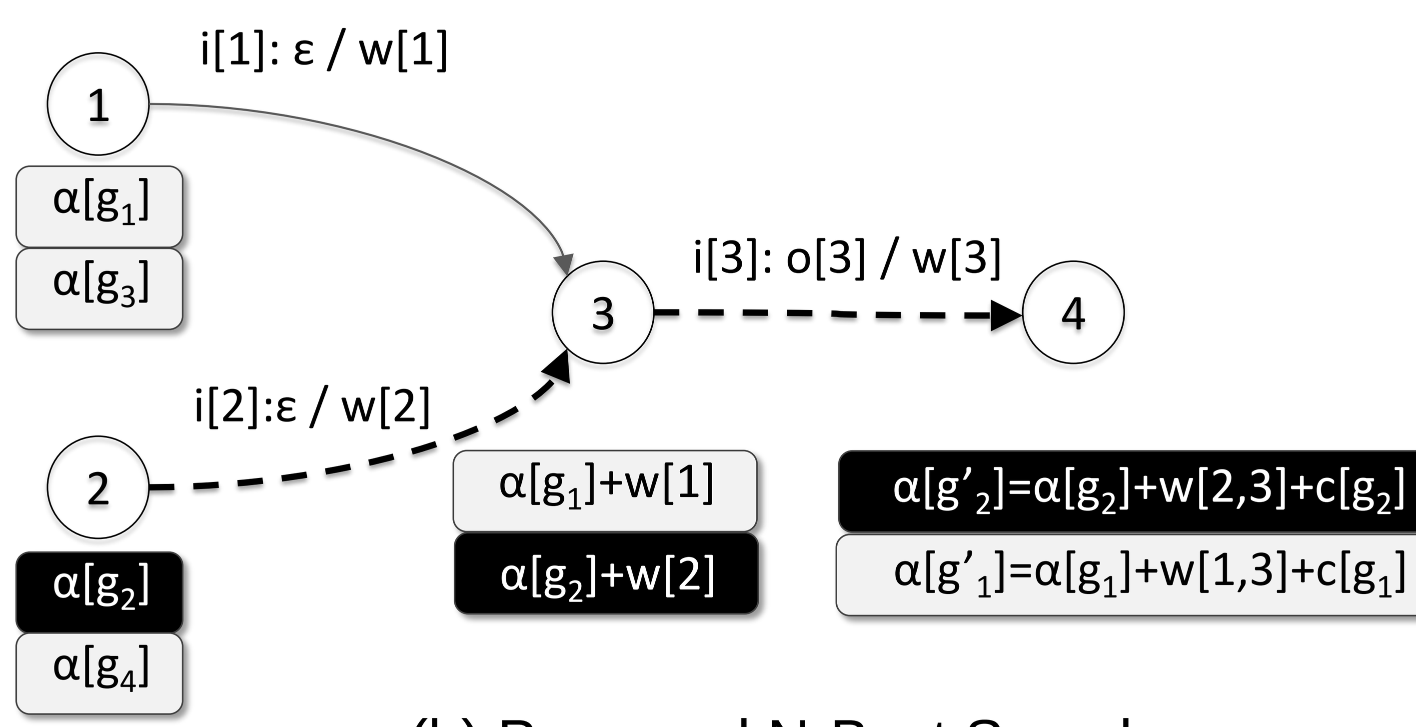
- Accuracy improves when maintaining more number of N-best hypothesis.
- Accuracy improvement converges with large N.

Carnegie Mellon University

N-Best On-The-Fly Partial Hypothesis Rescoring



(a) Standard One-Best Search



(b) Proposed N-Best Search

------→ Best Path	g: Partial hypothesis	i[.]: Input symbol	w[3] > w[2] > w[1]
α[.]: State likelihood	c[.]: Language model likelihood difference	o[.]: Output symbol	$\alpha[g_1] \approx \alpha[g_2] > \alpha[g_3] \approx \alpha[g_4]$
		w[.]: Arc weight	$c[g_1] >> c[g_2]$

Standard One-Best Search

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

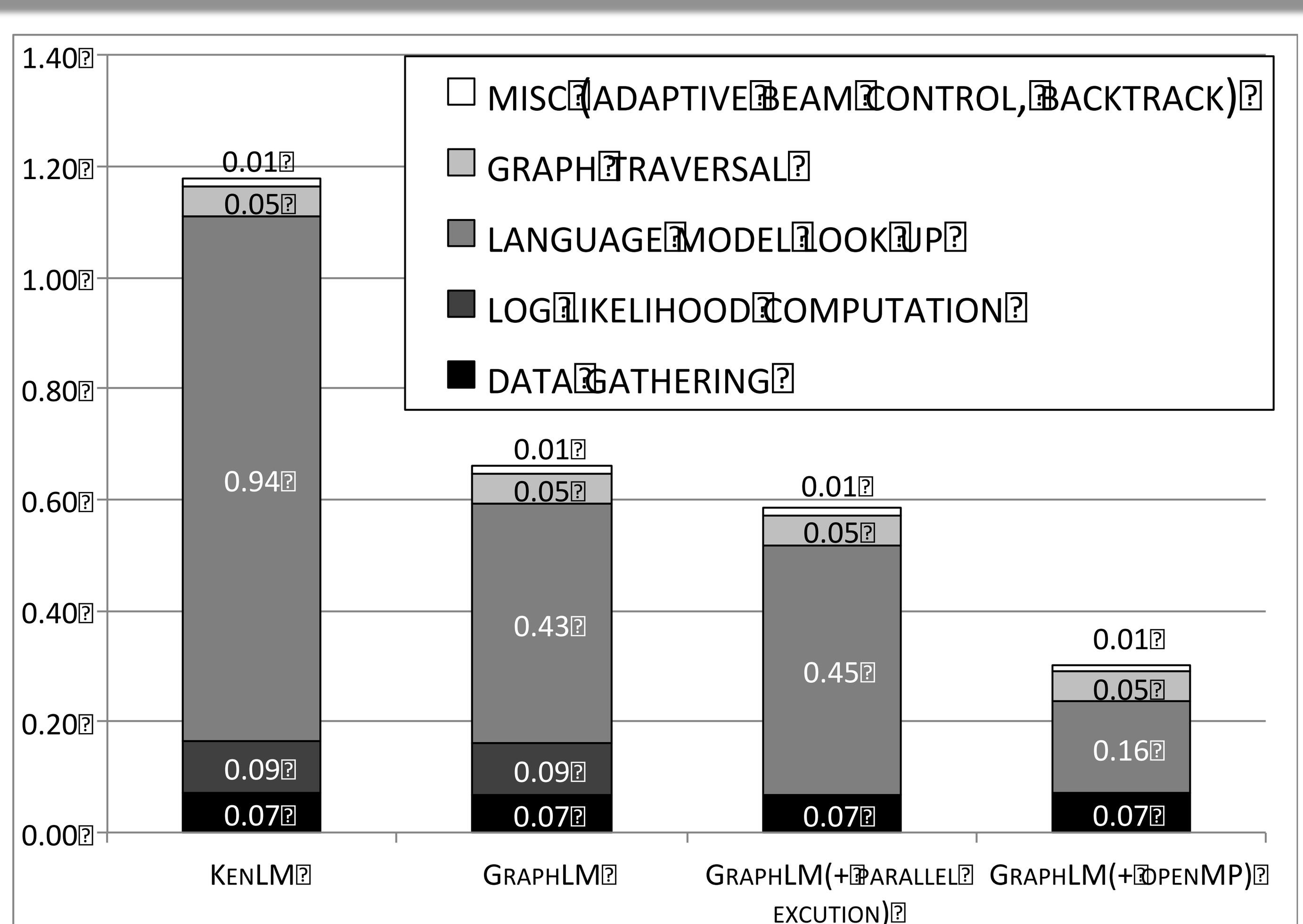
Why "N-Best"?

➔ ***Early pruning:*** Best hypothesis g_2 is pruned before the rescoring.

Proposed N-Best Search with Rescoring

- Rescore the partial hypothesis using likelihood difference between larger N-gram and unigram (c[.]) when hypothesis outputs word symbol.
- Maintaining N-best paths effectively allows multiple word hypotheses to be kept until rescoring can be applied

Load Balancing Between GPU and CPU using OpenMP



Ratio of the processing time per phase

GPU and CPU Parallel Execution

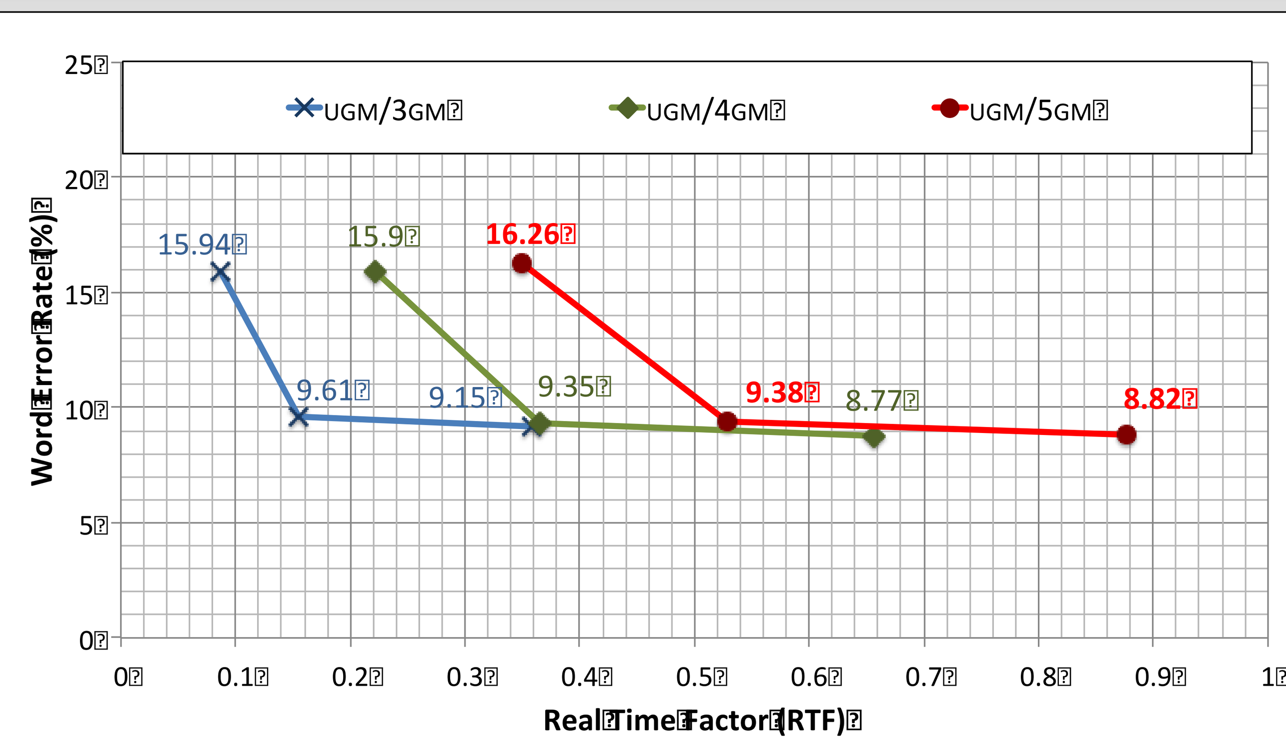
- Language model look up has no data dependency between Acoustic likelihood computation.
- CPU function and GPU kernel can be conducted in parallel
- Language model runtime can be hided behind GPU run time.

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 parallelized using OpenMP on the CPU to achieve better load balance.

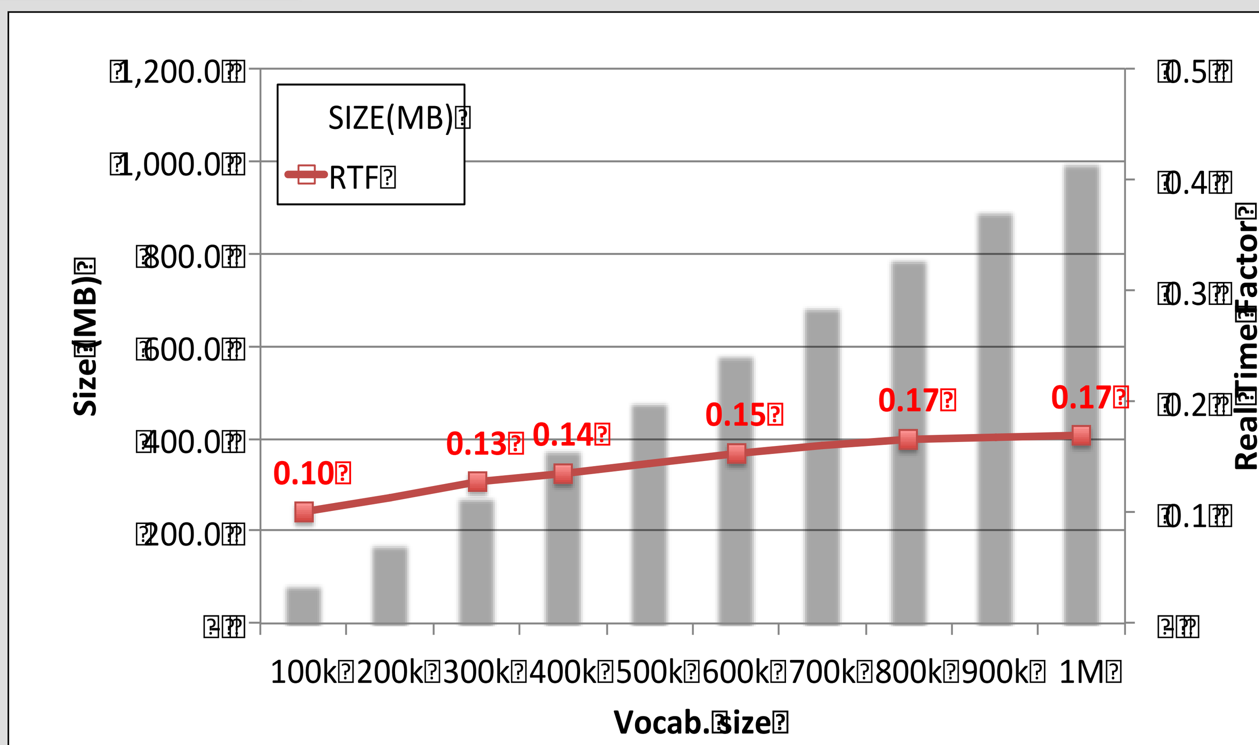
Experimental Evaluation

- 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



Relationship between vocab. Size and Real Time Factor (RTF)

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



Relationship between vocab. Size and Real Time Factor (RTF)

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