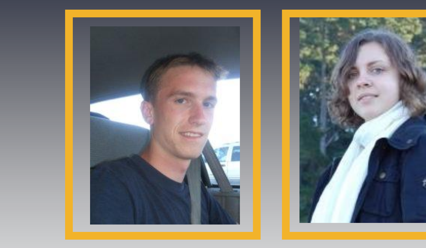
Creating a Scalable HMM based Inference Engine for Large Vocabulary Continuous Speech Recognition

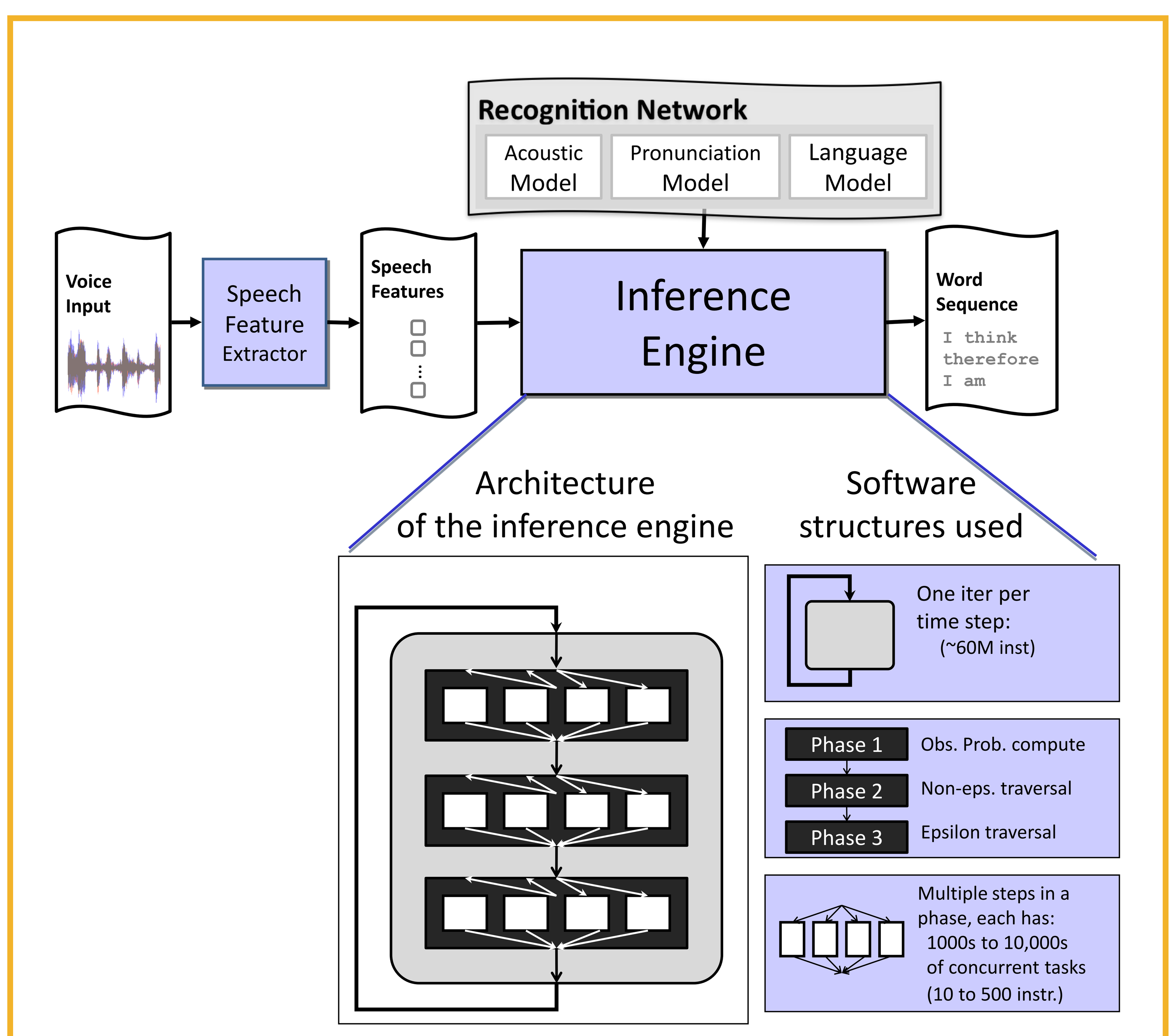
Ekaterina Gonina and Henry Cook



PROBLEM STATEMENT

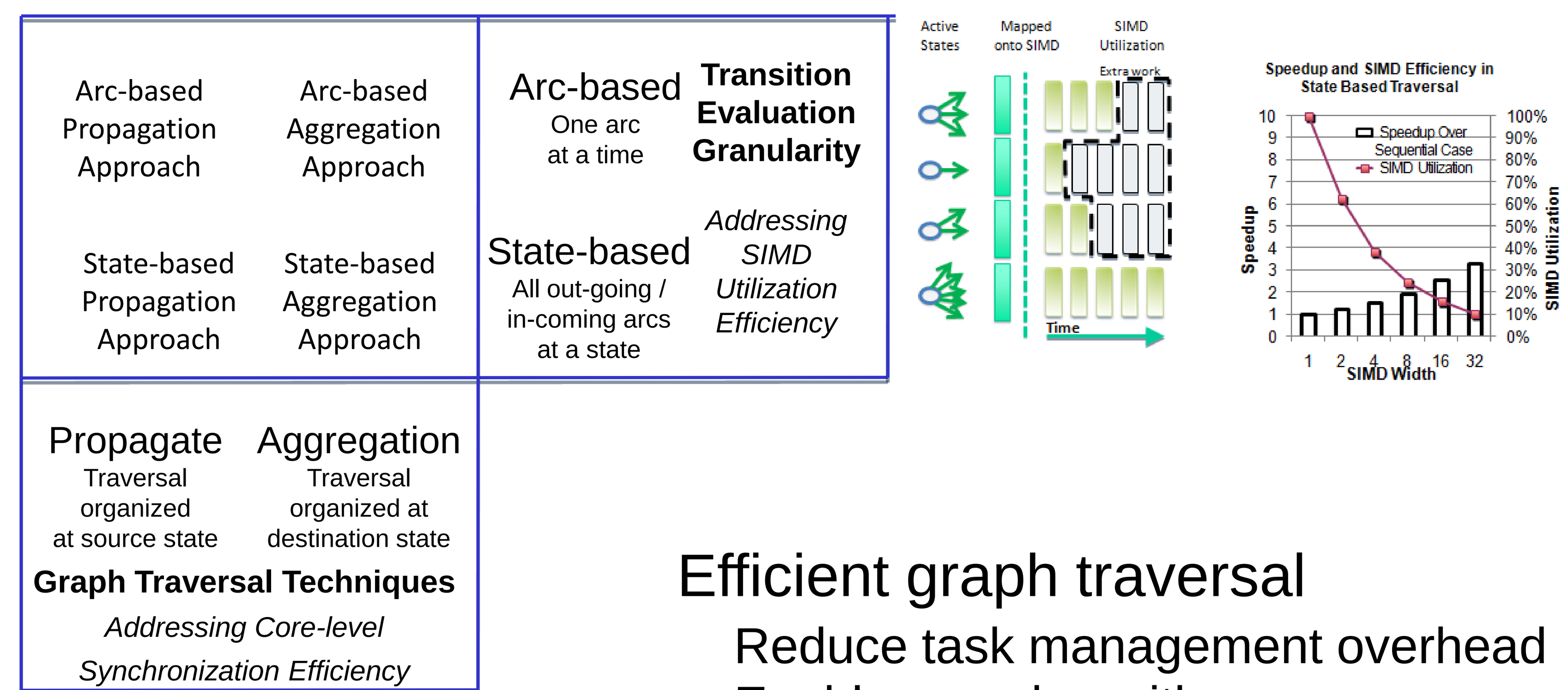
Our goal is to create algorithms for speech recognition inference that scale with increasing vector unit widths, increasing number of cores per die, and increasing complexity of the memory hierarchy. To do this we optimize SIMD efficiency, synchronization costs, and data locality and placement. We evaluate which algorithmic techniques are applicable across a diverse set of architectures including multicore processors, manycore GPUs, and virtual local stores.

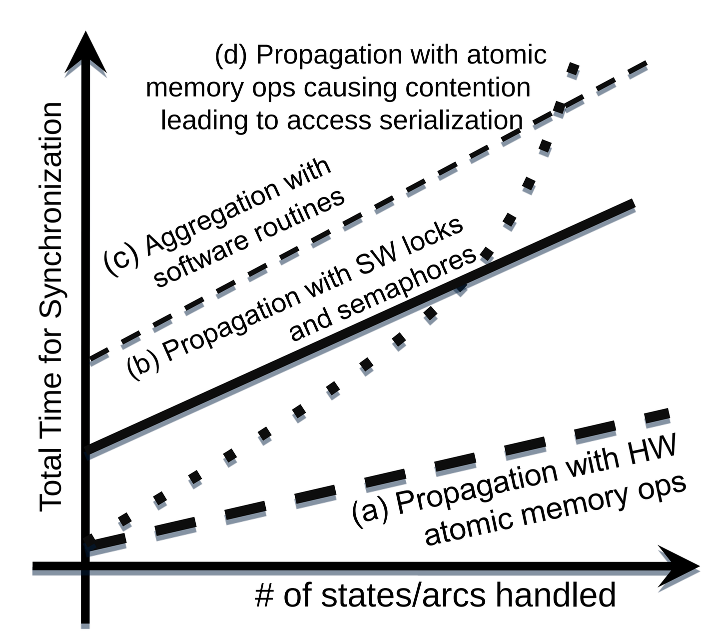
INFERENCE ENGINE CHARACTERISTICS



- The inference process takes the form of a parallel graph traversal through an irregular network of states and arcs.
- Traversal is guided by a sequence of input audio vectors.
- The algorithm has multiple phases, each of which express fine-grained parallelism. As the algorithm runs, it computes on a continuously changing data set corresponding to the currently active states in the underlying Hidden Markov Model.
- Fine-grained parallelism and changing work set size are the primary features that our algorithms must address to achieve scalability.

ALGORITHM DESIGN SPACE EXPLORATION



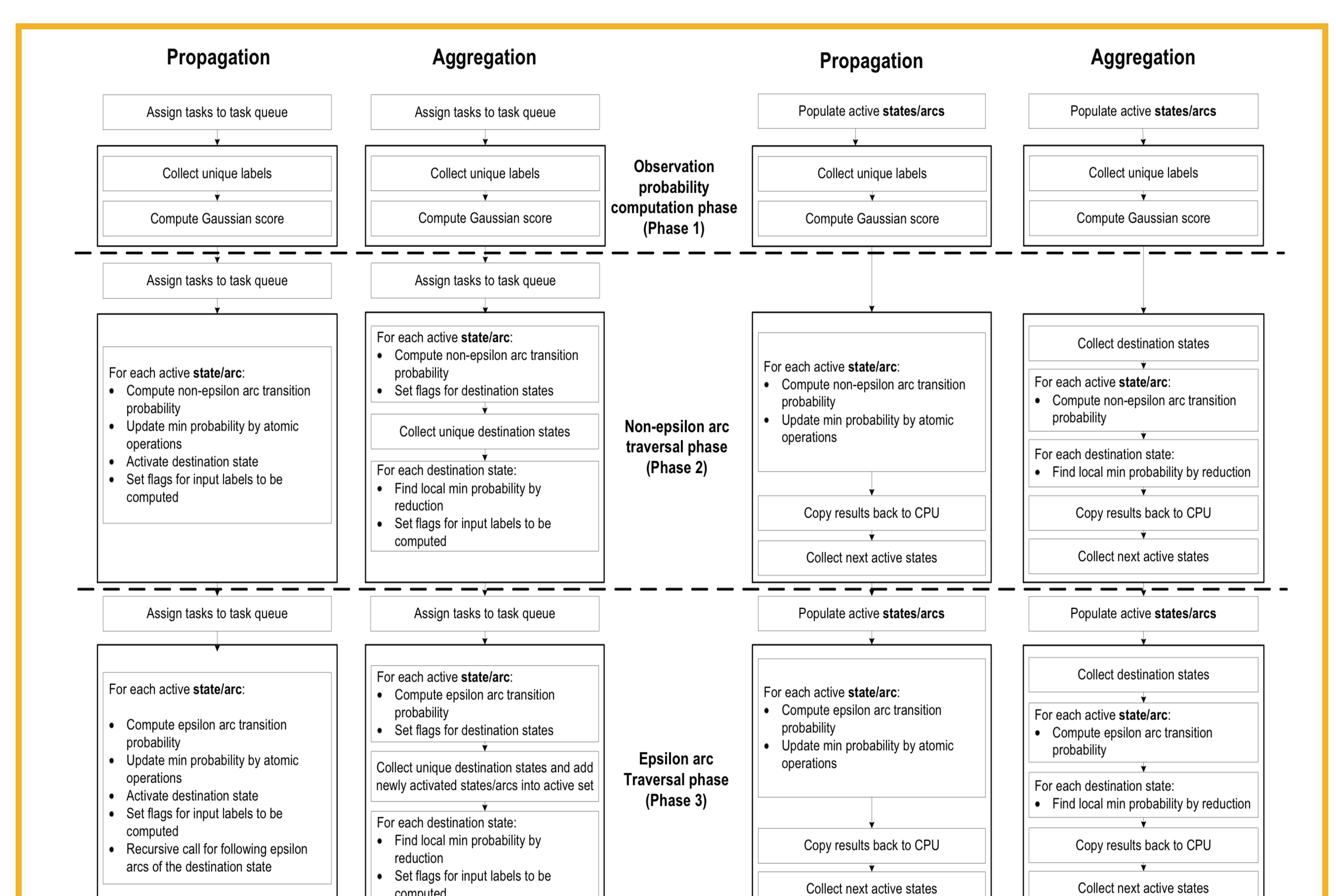


Efficient graph traversal
Reduce task management overhead
Enable speedup with more cores

Efficient transition evaluation
Enable speedup with SIMD width

Efficient data placement
Locality, alignment, coalescence
Software vs hardware management

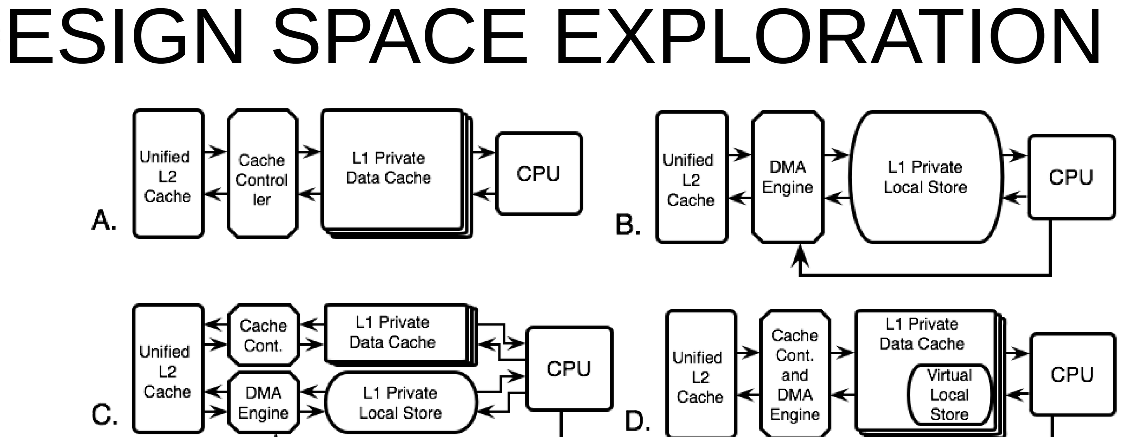
IMPLEMENTATION DETAILS



HARDWARE DESIGN SPACE EXPLORATION

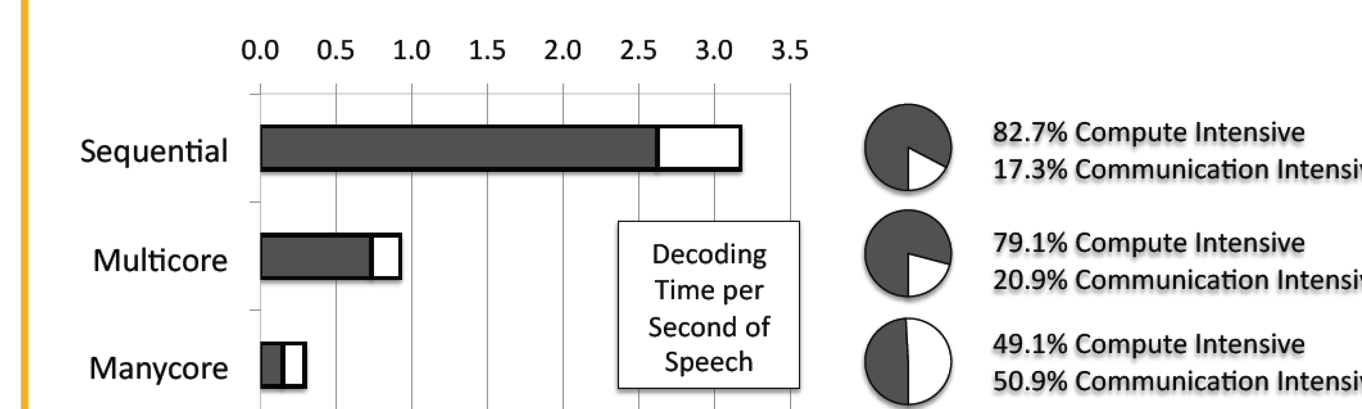
Multicore:
Intel Nehalem 920

Manycore:
NVIDIA G280

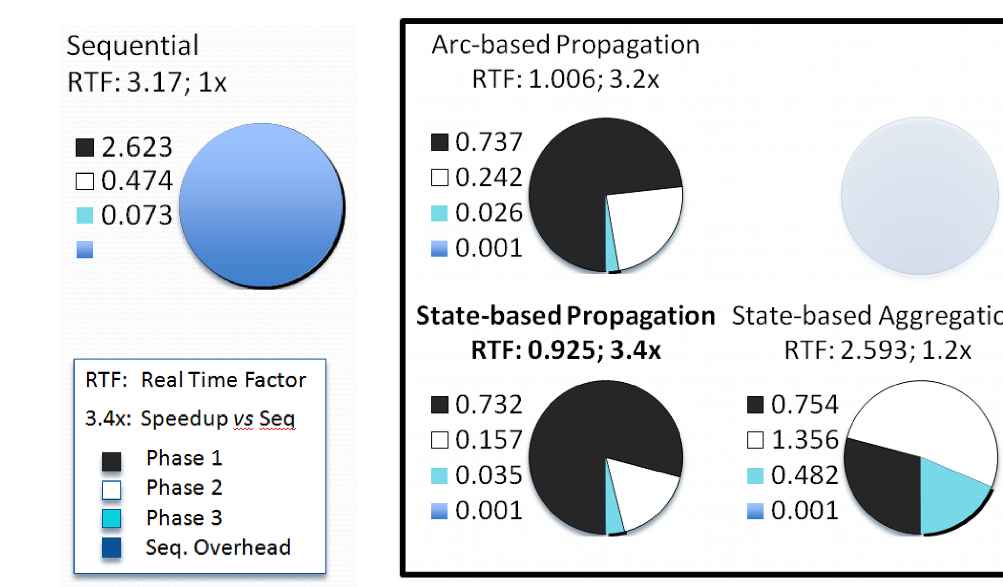


RESULTS

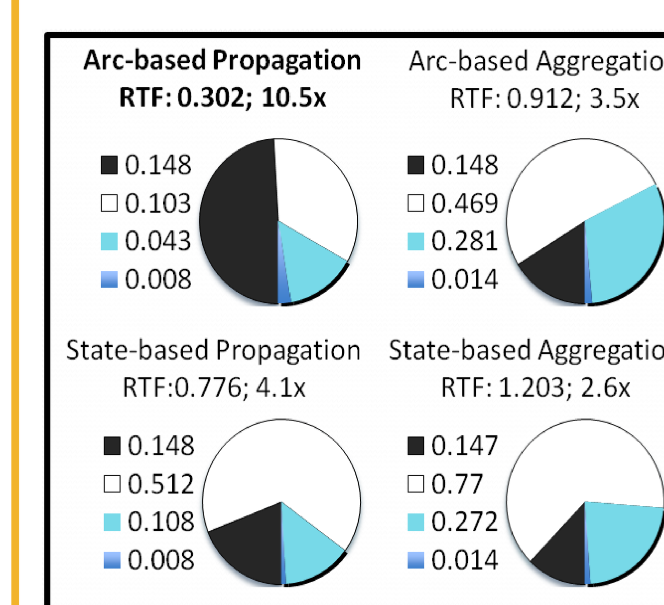
Communication vs. Computation

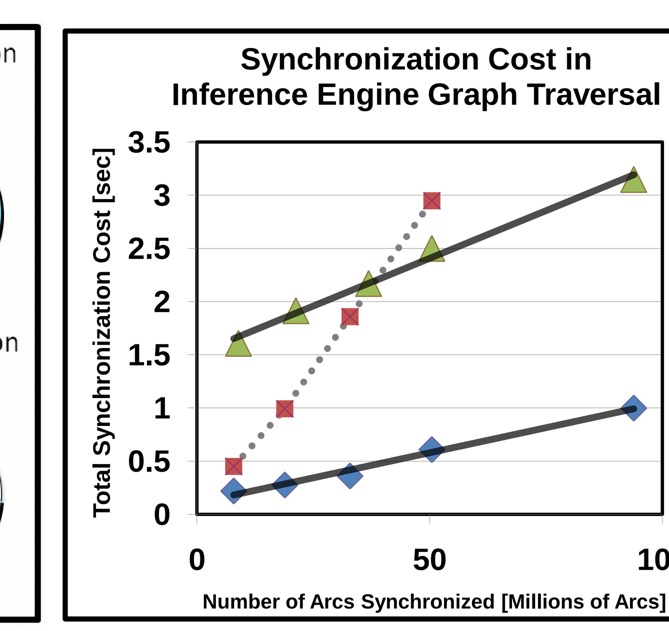


Multicore

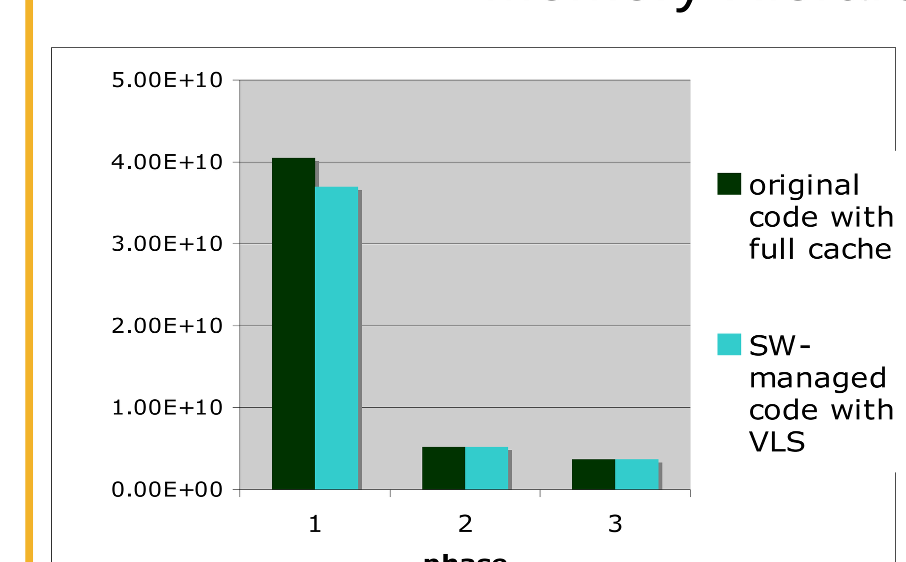


Manycore





Memory Hierarchy



Software management is most effective for phases with streaming memory accesses, but is less effective for phases with fine-grained random access patterns

FUTURE WORK

- Evaluate scaling on future platforms with wider SIMD or more cores
- Optimizations of private storage management for task queue model

COLLABORATORS: JIKE CHONG, KISUN YOU, YOUNGMIN YI, CHRISTOPHER HUGHES, WONYONG SUNG, KURT KEUTZER