Next Meeting

1. Demo Twitter on Accumulo, Dylan
2. Discuss FPGA speeding up Accumulo idea, Dr. Ganesan
    a. Paper
    b. http://newsoffice.mit.edu/2014/storage-system-for-big-data-dramatically-speeds-access-to-information-0131
    c. http://people.csail.mit.edu/wjun/ssd.htm

Prep notes for next meeting 2 October

1. Dr. Ganesan speaks on BLAST & related scoring functions - biology application
2. Yao presents insights from papers on CUDA GPU graph algorithms
3. Dylan presents an overview of DB computation-- MapReduce vs. NoSQL vs. Percolator-like hybrids vs. SQL
4. Real demo of Accumulo.

Meeting 2 October

Xin, nice job getting a website up! Everyone can look at https://stevens-graphgroup.github.io/ You can make changes via the git on Github.

Yao: BFS on GPUs --

- Uses Compressed Sparse Row storage CSR
- parallelism on all the nodes reached by the previous step
- works well for Erdos-Renyi random graphs, not as well for power law real world graphs because of the degree distribution
- What is graph partitioning?

Ganesan: HMMER - Hidden Markov model -

- model how proteins are synthesized. What are the domains that a protein contains? Compare a generating function for a protein.
- Proteins formed as chains of 15-300 amino acids. The states of a HMM generate amino acids in the sequence with a per-state probability, then transition to a new state with a transition probability.
- Viterbi algorithm - determine the protein sequence that has the highest probability of generating an observed sequence. Dynamic programming.
- The big compute part is doing this for every model. Also, there are tens of millions of sequences, which we evaluate all the models for.
- Want the most likely model within a particular family of models for a sequence, say 10-100.
- The problem statement: for each model (10-100 models), determine the most likely sequences for that model

Scantime Iterator design: run scoring on each protein, retain the top k proteins for each model

- Prefiltering --
- Global coarse parallelism - running multiple models at once, running multiple sequences at once
- Local fine parallelism - in the dynamic programming for a single model
- or change algorithm for more parallelism

Dylan: Discussed MapReduce framework, compared NoSQL vs. SQL vs. NewSQL databases. Discussed Google Percolator design (scratch notes here) as an intermediary between NoSQL and SQL.
Demoed Accumulo monitor, ingest and scans via the shell, simple summing iterator.

For HMMER specific questions, reach out to Hanyu Jiang <usukdream@gmail.com>.

> Super meeting today. It's awesome to see that we have so many things to discuss. Meeting notes here, copied below for the lazy.
>
> I recommend everyone finds a particular part to specialize, whether it is GPUs or the HMMER algorithm or other graph algorithms or Accumulo... etc. We'll discuss the big picture during our weekly meetings, but I imagine looking at the technical details of every component may be overwhelming.

Please contact me if you are not sure what to do or don't understand something we discuss. I can help or point you to the right resources, but I can't do that if you don't let me know.