GRAPHCHI: LARGE-SCALE GRAPH COMPUTATION ON JUST A PC

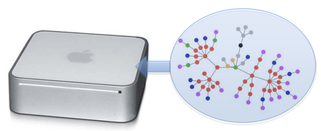
Aapo Kyrölä, Guy Blelloch (Carnegie Mellon University), Carlos Guestrin (University of Washington)

Contact author: akyrola@cs.cmu.edu

OSDI'12

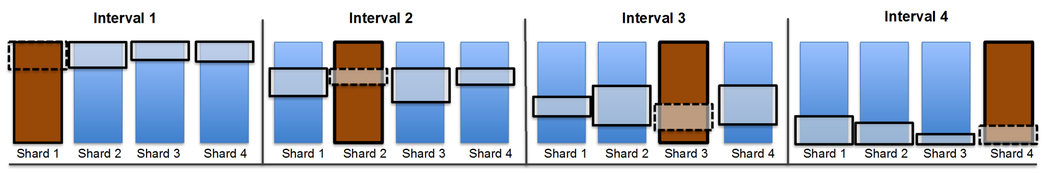
DISK-BASED GRAPH COMPUTATION ON A BASIC PC / LAPTOP (OSDI'12)

- Developing distributed algorithms remains hard.
- Distributing graph computation is particularly challenging.
- **Could we compute Big Graphs on just a single machine?**



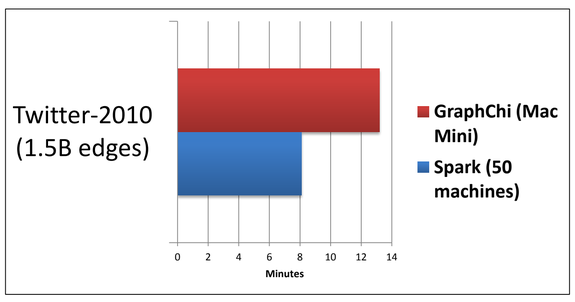
Natural graphs have random structure →
How to eliminate **random disk seeks**?

Our Solution: Parallel Sliding Windows

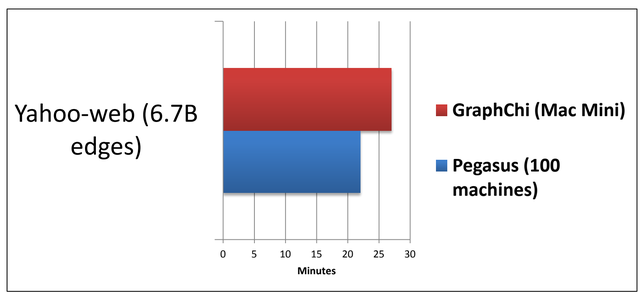


RESULTS: GRAPHCHI CAN SOLVE AS BIG PROBLEMS AS DISTRIBUTED SYSTEMS, ALMOST AS FAST

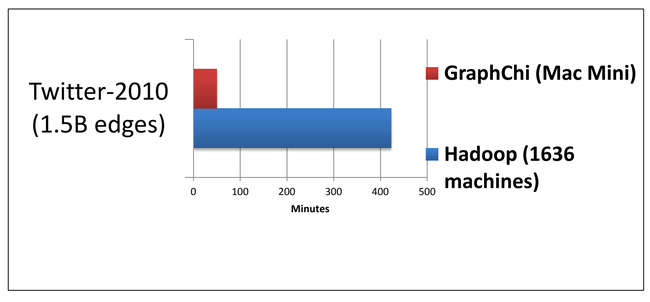
PageRank



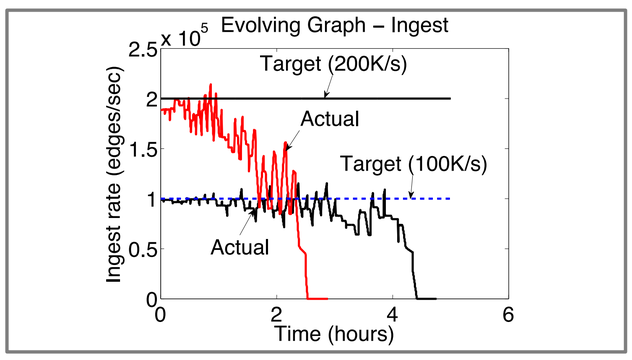
WebGraph Belief Propagation (U Kang et al.)



Triangle Counting



Evolving Graphs: Stream of 100,000 edges / sec while simultaneously computing PageRank [on Mac Mini].



Source code: http://graphchi.org

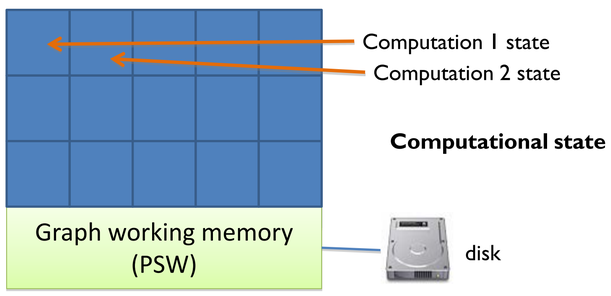
FUTURE WORK: GRAPHCHI ON A BIG MACHINE

Research question: How to compute multiple tasks on the same graph efficiently?

Background

- Computing recommendations for users requires executing graph algorithms with different settings for different groups → hundreds of tasks:
  - Languages
  - Countries
  - Customer segments
  - Interest groups, etc.

Idea: Use RAM for the algorithm state and let Parallel Sliding Windows process the graph from disk.



Early experiments:
On a machine with 144 GB of RAM, compute 20 parallel personalized PageRanks on *the current Twitter graph* → 20x throughput of Hadoop running > 100 workers.

Intel Science & Technology Center for Cloud Computing

Carnegie Mellon University · Georgia Tech · intel · Princeton University · UC Berkeley