Pregel: A System for Large-Scale Graph Processing

Grzegorz Malewicz, Matthew H. Austern, Aart J. C. Bik, James C. Dehnert, Ilan Horn,
Naty Leiser, and Grzegorz Czajkowski
Google, Inc.

Brad huntington
11/24/2013

Main idea

- The main idea of this paper is to address distributed processing of large scale graphs.

- Pregel: a fault-tolerant platform with an API that is sufficiently flexible to express arbitrary graph algorithms.

- The Pregel process is to divide the work load amongst multiple machines in order to increase processing power.

- The paper describes different aspects of Pregel: the model, its expression as a C++ API, implementation issues, present several applications of the model, and performance results.

Implementation

- Pregel was designed for the Google cluster architecture.
  - applications typically execute on a cluster management system to optimize resource allocation.
- Basic architecture:
  - Many copies of the program are executed an a large group of machines, one machine is made the “master”
  - The master determines how many partitions the graph will have, and then assigns each machine one or more partitions.
  - User input is sent to each machine and used as a set of records
  - The worker then executes a function on each active vertex, and then lets the master know how many active vertexes will be available for the next super step.
- When a worker fails their current progress is lost, the master then reassigns the tasks and reloads from the latest checkpoint.

My Analysis

- I think the idea is interesting, it reminds me of Peer-To-Peer processing.
  - For example when using something like a BitTorrent many seeders feed data to the person using the torrent.
  - In the case with Pregel the seeders would be the workers who feed information to the master.
- Pregel is impressive because no matter how large a project is put into it, the run time can stay relatively low with a large amount of workers.

Advantages and Disadvantages

- Advantages:
  - Allows large scale graph data processing.
  - Increase in processing power due to multiple machines working towards 1 common goal.
  - Assists in many real world applications.
- Disadvantages
  - If a worker malfunctions its progress is lost and the master has to find more workers to make up for the lost work.
    - If a lot of work needs to be made up multiple machines would need to be utilized to make up for lost time, this results in inefficient work because the machines could be working on the next project

Real world uses

- PageRank
  - rank websites in Google search engine results.
- Shortest Paths
  - Used in navigation
- Bipartite Matching
- Semi Clustering