Shared Memory Multicore PageRank Estimation

Piyush Garewal
Sindhu Kotapati
Tekenah Opuba

ADVISOR:
DR. APARICIO CARRANZA
Objective

To design, implement and test a multiprocessing parallel computation.

PageRank Algorithm using OpenMP

A random walk methodology using dataset from Stanford network

Management of load balancing, synchronization

Handling deadlocks

Comparison with parallel algorithmic concepts using speedup, runtime and efficiency
Background

The co-founders of Google, Larry Page and Sergey Brin innovated the algorithm the PageRank that took in the vast amount of data that is available on the Internet, organized it and ranked the sites in accordance to the number of links it receives, higher the rank the greater the importance.

This algorithm is unparalleled in its design and accuracy in delivering accurate results to queries, and continues to be the leading innovation that sets Google apart from its competitors.
Consider an imaginary web of 3 web pages. And the inbound and outbound link structure is as shown in the figure. The calculations can be done by following method:

\[
\begin{align*}
PR(A) &= 0.5 + 0.5 \times PR(C) \\
&= 0.5 + (0.5 \times 1) \\
&= 1 \\
PR(B) &= 0.5 + 0.5 \times (PR(A) / 2) \\
&= 0.5 + 0.5 \times (1/2) \\
&= 0.5 + 0.25 \\
&= 0.75 \\
PR(C) &= 0.5 + 0.5 \times ((PR(A) / 2) + PR(B)) \\
&= 0.5 + 0.5 \times ((1/2) + 0.75) \\
&= 0.5 + 0.5 \times 1.25 \\
&= 0.5 + 0.625 \\
&= 1.125
\end{align*}
\]
To design, implement and test a parallel estimator for PAGERANK of all nodes in a graph.

The PageRank of a node \((v)\) in a graph is an indicator of the node's importance. For instance, a web graph, where nodes are web pages and edges are crosslinks (directional; incoming or outgoing) between two webpages.

In this context, the PageRank of a node is the relative importance of that webpage on the internet (higher the better).

The original Pagerank paper was written by Larry Page, Sergey Brin and others (@Stanford, and later Google).
Parallel PageRank

1. Select Graph data
2. Reading and store graph in matrix
3. Select starting vertex to perform random walk
4. Select ending node
5. Provide all vertex equal initial rank starting rank
6. Count incoming and outgoing link (edges) for selected range of vertex.
7. Calculate PageRank
8. Print Top 5 PageRank
Selecting data (Graph)

### Web graphs

<table>
<thead>
<tr>
<th>Name</th>
<th>Type</th>
<th>Nodes</th>
<th>Edges</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>web-BerkStan</td>
<td>Directed</td>
<td>685,230</td>
<td>7,600,595</td>
<td>Web graph of Berkeley and Stanford</td>
</tr>
<tr>
<td>web-Google</td>
<td>Directed</td>
<td>875,713</td>
<td>5,105,039</td>
<td>Web graph from Google</td>
</tr>
<tr>
<td>web-NotreDame</td>
<td>Directed</td>
<td>325,729</td>
<td>1,497,134</td>
<td>Web graph of Notre Dame</td>
</tr>
<tr>
<td>web-Stanford</td>
<td>Directed</td>
<td>281,903</td>
<td>2,312,497</td>
<td>Web graph of Stanford.edu</td>
</tr>
</tbody>
</table>

Source: [https://snap.stanford.edu/data/](https://snap.stanford.edu/data/)
Storing Data
Partitioning the store graph (Load balance)
Key challenge

Avoiding deadlock (Race Case)

A race condition occurs when two or more threads can access shared data and they try to change it at the same time.
Key Challenges

Load Balancing: In parallel programming, the most important part is load balancing i.e.

- All the processes have an approximately equal amount of task.
- In an efficient parallel implementation, all the processes should complete the task approximately the same computation time.
- To achieve the proper load balance graph,
- We have stored the graph in the adjacent matrix but accessing the elements in the graph was not time efficient, then we tried using adjacency list, here all the elements/nodes can be divided equally into all the processes.
Load Balancing on graph
What is Speedup and Efficiency?

The speedup is defined as the ratio of the serial runtime of the best sequential algorithm for solving a problem to the time taken by the parallel algorithm to solve the same problem on \( p \) processors.

The efficiency is defined as the ratio of speed up to the number of processors. Efficiency measures the fraction of time for which a processor is usefully utilized.
Results
Results
Applications of PageRank

Neuroscience - detect brain activity through examining changes in blood flow to the brain. PageRank can be used to generate a network in which the voxels or the subunits of the image of the brain are represented as nodes and the edges between the nodes represent the voxels that are time correlated. With this information, neuroscientists are able to conduct studies to assess the parts of the brain that have changed as the patient ages.

Toxic Waste Management - determine the position of water molecules which can provide an efficient method for waste management teams to remove nuclear waste/pollutants.

Traffic Patterns - Each road is represented as a node on the graph, and each intersection where roads meet are the edges. Understanding traffic patterns can aid civil engineers in figure out how to improve overall traffic flow and prevent congestion.


Question ?
Thank you for your time!