Hunter Damron

Graham McDonald

Jeffrey Segars

Home Assignments

1. Performance Evaluation

Searching for ‘Vronsky’ in Anna Karenina (timed using sys time of `time` command)

* Part 1 (pipes): 0.23 s
* Part 2 (sockets): 0.25 s
* Part 3 (MapReduce): 0.02 s

Pipes and sockets perform roughly the same because they are very similar structures (sockets are essentially bidirectional pipes). The imbalance between reading and writing causes inefficiencies which slows down performance. The MapReduce implementation, on the other hand, is much faster at the cost of a larger memory footprint. The speed improvement is achieved by loading more of the input file to operate on at once and by allowing four threads to concurrently process part of the file with only the reduce portion of the task requiring a single dedicated process.

2. Linux supports a variety of methods for interprocess communication. These include pipes, FIFO’s, messsage queues, unix-domain sockets, and shared memory. A pipe is a technique used for passing information from one process to another using a buffer in the kernel. It also utilizes two file descriptors one representing the read end (fd[0]) and the other the write end (fd[1]) of the pipe. Pipes can only be used among a parent and a child process. A FIFO, or named pipe, is similar to a pipe but is not limited to parent-child processes. Additionally, a file is created on the hard disk and there can be multiple writers with one reader. Message queues conceptually is simply a queue in the kernel. These queues utilize POSIX API functions such as send () and receive () to pass messages allowing for the processes to communicate. Unix domain sockets are the recommended way for data passing. They combine the power of a pipe and a message queue to form a socket. It is also bi-directional meaning both ends of the socket can send and receive data. Finally, shared memory is a form of interprocess communication in which a region of physical memory is mapped between two or more processes. Shared memory is useful because it is the only form that doesn’t require mode (user/kernel) switching. Hence why it is the most efficient one.

3. Map-reduce is a model and implementation for processing and generating data with a parallel computing algorithm. A typical program consists of a map procedure which performs sorting and a reduce procedure which performs a summary operation. In our project, the child process created 4 threads each acting as a mapper. The child’s main thread acted as the single reducer. Map-reduce libraries have been written in many languages with different levels of optimization such as Apache’s Hadoop described below.

4. Apache’s Hadoop is a collection of open-source software utilities that solve problems involving massive amounts of data and computation. It was inspired by papers on big data processing such as the Google paper by Jeffrey Dean and Sanjay Ghemawat. This paper led Doug Cutting to develop an open- source implementation of the map-reduce framework. It was named Hadoop after his son’s toy elephant. Hadoop is popular and important in industry because it simplifies the processing of large data by splitting files into large blocks and distributes them across nodes in a cluster. Then packaged code is transferred into nodes to process the data in parallel. Hadoop is so popular that over half of the Fortune 500 companies use it today.