1. What are the basic steps (show all steps) in building a parallel program? Show

at least one example.

* Parallel programming had been developed to improve performance and efficiency. First, we must identify sets of tasks that can run concurrently or partitions of data which can be processed concurrently on a parallel program. A common implementation technique is master/worker. Master/Worker uses static load balancing. Which divides tasks among the processors in a parallel system to avoid processors being idle to increase efficiency and performance.

1. What is MapReduce?

* MapReduce is a programming model. The map and reduce combinators from Lisp, a functional language originates this model.

1. What is map and what is reduce?

* In Lips, map takes a function as an input and a sequence of values. After this, it applies the function to each value in the sequence. Map which is written by a user of the MapReduce library takes input pairs and produces a set of intermediate key values. Then, all the intermediate values associated with intermediate keys are grouped together by the MapReduce library passes them to reduce function.

Reduce combines all the elements od the sequence taken by map using a binary operation. This is also written by the user which accepts an intermediate key and a set of values for that key. It forms a conceivably smaller set of values by merging those values together.

1. Why MapReduce?

* It is an abstraction which allows to write scalable applications which uses parallel processing to process a huge amount of data. It was introduced by Google, so that their engineers could perform simple computation while hiding details of parallelization, load balancing, data distribution and fault tolerance. It processes a large amount of raw data and then distributes it across thousands of machines to be processed in a reasonable time. After Google, Yahoo, Netflix, Facebook and many more companies starting using MapReduce.

1. Show an example for MapReduce.

* Source Code:

map(String key, String value):

for each word w in value:

EmitIntermediate(w, "1");

reduce(String key, Iterator values):

int result = 0;

for each v in values:

result += ParseInt(v);

Emit(AsString(result));

Here, key is document name and value is document contents, under map function. It emits each word plus an associated count of occurrence using the function map. After that, the reduce function sums together all that for a specific word.

1. Explain in your own words how MapReduce model is executed?

* MapReduce is a programming model which is used to process huge amount of data. The programs are parallel in nature. The model has two phase first one is Map and the second one is Reduce. First, phase of this model splits the input data in sets. The input data after the split can be processed on different machines on parallel. On the second phase, the partitioning of the intermediate key space happens. Here, the number of partitions and the functions of partitions are specified by the user.

1. List and describe three examples that are expressed as MapReduce computations.

* Three examples which are expressed as MapReduce computations are Distributed Grep, Reverse Web-Link Graph and Inverted Index.

Distributed Grep: This map function emits a line if it matches a given pattern. The reduce function then copies the supplied intermediate data to output.

Reverse Web-Link Graph: Here map function outputs pairs of target and source links to a specific URL found in a page named source. After that the reduce function takes all source URL’s associated with the given target URL and emits the pairs.

Inverted Index: Word and document ID pairs are emitted by the map function. The reduce function then accepts all the pairs of given words and sorts the document ID’s and emits the pairs. All output pair forms inverted index.

1. When do we use OpenMP, MPI and, MapReduce (Hadoop), and why?

* In a code, OpenMP is used to introduce shared memory parallelism. It is an effective directive-based library which is powerful yet simple. It it used mostly to get the best performance. But it has its limitations, sometimes it causes performance issues.

Message Passing Interface or MPI is used to develop scientific applications. It uses a distributed memory parallel model implementation. The codes are usually load balanced and highly synchronous. It can be used in codes which runs over multiple machines. User can also combine MPI and OpenMP to do hybrid programming.

Hadoop MapReduce gives user two constructs to apply over large amount of data. This is used when data is too large to organize. User can do some reduction over the results also by using this. If users have terabytes of data to extract, transform and load they should use Hadoop MapReduce. Users can also use this for scientific applications, but it will not be as good as MPI.

1. In your own words, explain what a Drug Design and DNA problem is.

* Our DNA contains the instructions for making proteins in our bodies. Different proteins perform different tasks in a body. To identify which protein does which task we need to see the protein shape. Small molecules called ligands found in protein is used to design drugs. Some ligands will fit, and some will not. The one which fit will be used to produce a desired shape change to design a drug. A collection of ligands is used and tested against the protein to check if they are binding in a useful way. They are scored on binding properties and the best are marked and identified which will make good drug candidates.

Here, MapReduce strategy is used to solve the problem by implementing a master worker pattern. First, we use a function which many ligands will be tested against the given protein. Then Map function will be used to compute binding score for pairs. Then Reduce function will identify the highest scoring ligands.