

What is claimed is:

1. A method for processing tasks in a distributed data processing system including a plurality of processing nodes, the method including:
 - processing a plurality of sets of tasks using two or more of the plurality of processing nodes, each set of tasks configured to generate a corresponding set of results and being associated with a corresponding time interval of a plurality of time intervals,
 - maintaining, at a first processing node of the plurality of processing nodes, a plurality of counters, the plurality of counters including:
 - a working counter indicating a current time interval of the plurality of time intervals in the distributed data processing system, and
 - a replication counter indicating a replication time interval of the plurality of time intervals, earlier than the current time interval, for which at least one of (1) all tasks associated with the replication time interval, or (2) all corresponding results associated with the replication time interval, are replicated at multiple processing nodes of the plurality of processing nodes, and
 - providing messages from the first processing node to other processing nodes of the plurality of processing nodes, the messages including the working counter and the replication counter.
2. The method of claim 1 wherein the processing includes performing computations associated with at least some tasks at a primary processing node of the plurality of processing nodes and performing computations associated with replicas of the at least some tasks at one or more backup processing nodes of the plurality of processing nodes.

3. The method of claim 2 wherein performing computations associated with a first task at the primary processing node includes generating a first result at the primary processing node, and performing computations associated with a replica of the first task at the backup processing node includes generating the first result at the backup processing node.
4. The method of any one of claims 2 through 3 wherein the computations associated with the first task performed at the primary processing node and the computations associated with the replica of the first task performed at the backup processing node both begin after a commit operation indicating that both the first task and the replica of the first task have been persistently stored.
5. The method of any one of claims 2 through 3 wherein the first result includes a modified version of original data that has been replicated at the primary processing node and the backup processing node.
6. The method of any one of claims 2 through 3 wherein the computations associated with the first task performed at the primary processing node are identical to the computations associated with the replica of the first task performed at the backup processing node.
7. The method of claim 6 wherein the computations associated with the first task performed at the primary processing node and the computations associated with the replica of the first task performed at the backup processing node are deterministic and do not depend on which of the multiple processing nodes perform the computations.
8. The method of claim 2 wherein the replication counter indicates a time interval of the plurality of time intervals for which all tasks associated with that time interval and all corresponding results associated with that time interval are replicated at multiple processing nodes of the plurality of processing nodes.

9. The method of any one of claims 1 through 8 wherein the processing includes performing computations associated with at least some tasks at a primary processing node of the plurality of processing nodes while a replica of the task remains dormant at a backup processing node, and sending a result corresponding to a task for which computations were performed from the primary processing node to the backup processing node.

10. The method of claim 9 wherein performing computations associated with a first task at the primary processing node includes generating a first result, and the first result is sent to from the primary processing node to the backup processing node after the first task has been completed at the primary processing node.

11. The method of any one of claims 9 through 10 wherein the first result includes a modified version of original data that has been replicated at the primary processing node and the backup processing node.

12. The method of claim 9 wherein the replication counter indicates a time interval of the plurality of time intervals for which all tasks associated with that time interval and all corresponding results associated with that time interval are replicated at multiple processing nodes of the plurality of processing nodes.

13. The method of claim 1, further including:
maintaining a plurality of data stores in the system, each data store of the plurality of data stores being associated with a corresponding processing node of the plurality of processing nodes and being associated with a durability level of a plurality of durability levels, the plurality of durability levels including a first durability level and a second durability level with a relatively greater degree of durability than the first durability level;

processing a plurality of sets of data units using two or more processing nodes of the plurality of processing nodes, each data unit of each set of data units being associated with a corresponding time interval of the plurality of time intervals, the plurality of sets of data units including a first set of data units associated with a first time interval of the plurality of time intervals, the processing including, for each particular durability level, updating an associated indicator to indicate that all sets of data units associated with the first time interval are stored at that particular durability level;

wherein processing the plurality of sets of tasks includes processing a plurality of sets of requests using the two or more of the plurality of processing nodes, each request of each set of requests being configured to cause a state update at a processing node of the plurality of processing nodes and being associated with a corresponding time interval of the plurality of time intervals, the plurality of sets of requests including a first set of requests associated with a second time interval of the plurality of time intervals.

14. The method of claim 13 wherein the plurality of counters further include a persistence counter indicating a time interval of the plurality of time intervals for which all requests associated with that time interval are stored in persistent storage associated with at least one processing node of the plurality of processing nodes.

15. The method of claim 13 wherein for each data unit of the first set of data units, storing the data unit in data stores of the plurality of data stores associated with respective processing nodes of the plurality of processing nodes, including storing the data unit in data stores of the plurality of data stores associated with the first level of durability and storing the data unit in one or more data stores of the plurality of data stores associated with the second level of durability.

16. The method of claim 1, further including:

receiving input data at a distributed data processing system interface component in communication with the distributed data processing system;

providing the received input data to the distributed data processing system, wherein the distributed data processing system assigns an indicator associated with a first time interval of the plurality of time intervals to the input data;

receiving, at the distributed data processing system interface component, result data associated with the input data from the distributed data processing system, wherein the result data includes the indicator associated with the first time interval;

determining, at a first processing node of the plurality of processing nodes, an indicator associated with a second time interval of the plurality of time intervals;

comparing, at the distributed data processing system interface component, the indicator associated with the second time interval to the indicator associated with the first time interval included in the result data and, if the indicator associated with the second time interval corresponds to a time interval identical to or later than the first time interval, releasing the result data from the distributed data processing system interface component;

maintaining a plurality of data stores in the system, each data store of the plurality of data stores being associated with a corresponding processing node of the plurality of processing nodes and being associated with a durability level of a plurality of durability levels, the plurality of durability levels including a first durability level and a second durability level with a relatively greater degree of durability than the first durability level; and

processing a plurality of sets of data units using the two or more processing nodes of the plurality of processing nodes, each data unit of each set of data units being associated with a corresponding time interval of the plurality of time intervals, the plurality of sets of data units including a first set of data units associated with a third time interval of the plurality of time intervals, the processing including, for each particular durability level, updating an associated indicator to indicate that all sets of data units associated with the third time interval are stored at that particular durability level.

17. The method of claim 16 wherein for each data unit of the first set of data units, storing the data unit in data stores of the plurality of data stores associated with respective processing nodes of the plurality of processing nodes, including storing the data unit in data stores of the plurality of data stores associated with the first level of durability and storing the data unit in one or more data stores of the plurality of data stores associated with the second level of durability

18. The method of claim 16 wherein the indicator associated with the second time interval is provided to the distributed data processing system interface component.

19. The method of claim 1, further including:

receiving input data at a distributed data processing system interface component in communication with the distributed data processing system;

providing the received input data to the distributed data processing system, wherein the distributed data processing system assigns an indicator associated with the first time interval of a plurality of time intervals to the input data;

receiving, at the distributed data processing system interface component, result data associated with the input data from the distributed data processing system, wherein the result data includes the indicator associated with the first time interval;

determining, at the first processing node of the plurality of processing nodes, an indicator associated with a second interval of the plurality of time intervals;

comparing, at the distributed data processing system interface component, the indicator associated with the second time interval to the indicator associated with the first time interval included in the result data and, if the indicator associated with the second time interval corresponds to a time interval identical to or later than the first time interval, releasing the result data from the distributed data processing system interface component;

wherein processing the plurality of sets of tasks includes processing a plurality of sets of requests using the two or more of the plurality of processing nodes, each request of each set of requests being configured to cause a state update at a processing node of the plurality of processing nodes and being associated with a corresponding time interval of the plurality of time intervals, the plurality of sets of requests including a first set of requests associated with a third time interval of the plurality of time intervals.

20. The method of claim 19 wherein the indicator associated with the second time interval is provided to the distributed data processing system interface component.

21. The method of claim 19 wherein the plurality of counters further include a persistence counter indicating a time interval of the plurality of time intervals for which all requests associated with that time interval are stored in persistent storage associated with at least one processing node of the plurality of processing nodes.

22. Software stored in a non-transitory form on a computer-readable medium, for processing tasks in a distributed data processing system including a plurality of nodes, the software including instructions for causing a computing system to perform all of the steps of the method of any one of claims 1 through 21.

23. An apparatus for processing data, the apparatus including:
a distributed data processing system including a plurality of processing nodes,
each processing node including at least one processor; and
a communication medium connecting the plurality of processing nodes for
sending and receiving information between processing nodes of the
plurality of processing nodes;
wherein the distributed data processing system, is configured to perform all of
the steps of the method of any one of claims 1 through 21.
24. A computing system for processing tasks in a distributed data processing
system including a plurality of processing nodes, the computing system configured to
perform all of the steps of the method of any one of claims 1 through 21.