

What is claimed is:

1. A method for managing data in a distributed data processing system including a plurality of processing nodes, the method including:

maintaining a plurality of data stores in the system,

each processing node of the plurality of processing nodes maintaining a respective subset of the data stores,

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~~ having a durability level of a plurality of durability levels, at least some processing nodes each maintaining data stores at multiple durability levels, the plurality of durability levels including a first durability level and a second durability level with a greater degree of durability than the first durability level,

maintaining, by a root node, a plurality of time indicators for the processing system, each time indicator corresponding to a respective different durability level, including maintaining a first time indicator corresponding to the first durability level; and

processing a plurality of sets of data units using two or more of the plurality of processing nodes, including

assigning each data unit a time interval of a plurality of non-overlapping time intervals and assigning a first set of data units to a given time interval; and

storing data units in data stores at multiple levels of durability, including storing at least some data units in data stores having the first durability level and storing at least some data units in data stores having the second durability level;

wherein maintaining the plurality of time indicators includes updating and distributing, by the root node to the plurality of processing nodes, the first time indicator to represent a time value such that the data units assigned a time interval represented by the time value are stored in a data store having at least the first durability level.

~~each data unit of each set of data units being associated with a corresponding time interval of a plurality of time intervals, the plurality of sets of data units including a first set of data units associated with a particular time interval of the plurality of time intervals, the processing including:~~

- ~~—— 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 two or more 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;~~
- ~~—— updating a first indicator associated with the first durability level of the plurality of durability levels, with the first indicator maintaining an updateable first value representing a first time interval value by which all sets of data units associated with the first time interval value or earlier have been stored at the first durability level, after determining that all data units of the first set of data units associated with the particular time interval have been successfully stored in data stores associated with the first durability level, and~~
- ~~—— updating a second indicator associated with the second durability level of the plurality of durability levels, with the second indicator maintaining an updateable second value representing a second time interval value, lagging the first time interval value, by which all sets of data units associated with the second time interval or earlier have been stored at the second durability level, after determining that all data units of the first set of data units associated with the particular time interval have been successfully stored in at least one data store associated with the second durability level.~~

2. The method of claim 1 wherein data stores with the first durability level comprise volatile storage media.

3. The method of any one of claims 1 through 2 wherein data stores with the second durability level comprise persistent storage media.

4. The method of any one of claims 1 through 3 wherein maintaining the plurality of time indicators includes:

updating the first time indicator to represent a first time value corresponding to the given time interval to indicate that the first set of data units assigned to the given time interval are stored in a data store having the first durability level, and

subsequently updating a second time indicator corresponding to the second durability level to represent the first time value corresponding to the given time interval to indicate that the first set of data units assigned to the given time interval are stored in a data store having the second durability level ~~the first durability level is satisfied for the first set of data units prior to the second durability level being satisfied for the first set of data units.~~

5. The method of any one of claims 1 through 4 wherein the plurality of durability levels includes a third durability level having greater durability than the second durability level.

6. The method of claim 5 further comprising updating a third time indicator ~~associated with corresponding to~~ the third durability level of the plurality of durability levels to represent the first time value corresponding to the given time interval to indicate that the first set of data units assigned to the given time interval are stored in a data store having ~~all sets of data units associated with the first time interval are stored at~~ the third durability level, ~~after all data units of the first set of data units are successfully stored in at least one data store associated with the third durability level.~~

7. The method of claim 5 wherein data stores with the third durability level comprise persistent storage media outside of the distributed data processing system.

8. The method of claim 5 wherein maintaining the plurality of time indicators includes:

updating the first time indicator to represent a first time value corresponding to the given time interval to indicate that the first set of data units assigned to the given time interval are stored in a data store having the first durability level,

subsequently updating a second time indicator corresponding to the second durability level to represent the first time value corresponding to the given time interval to indicate that the first set of data units assigned to the given time interval are stored in a data store having the second durability level, and

subsequently updating a third time indicator corresponding to the third durability level to represent the first time value corresponding to the given time interval to indicate that the first set of data units assigned to the given time interval are stored in a data store having the third durability level.

~~the first durability level is satisfied for the first set of data units prior to the second durability level being satisfied for the first set of data units, and the second durability level is satisfied for the first set of data units prior to the first durability level being satisfied for the first set of data units.~~

9. The method of any one of claims 1 through 8 wherein the data units include data processing tasks.

10. The method of claim 9 wherein the data units include results generated by the data processing tasks.

11. The method of any one of claims 1 through 10 wherein the data units include inter-processing node messages.

12. The method of any one of claims 1 through 11 wherein the data units include data records.

13. The method of claim ~~4~~1, wherein updating the first time indicator includes incrementing the first time indicator to a first value representative of a time earlier than a current working time maintained by a working counter, wherein updating the second time indicator includes incrementing the second time indicator to a second value representative of another time earlier than the current working time or the time represented by the first value of the first time indicator, and wherein the method further includes broadcasting a message containing at least the first time indicator, the second time indicator and the working counter to the plurality of processing nodes to cause a set of data units associated with time intervals equal to or earlier than the value of the first time indicator to be stored to storage with the first durability level, and to cause another set of data units associated with time intervals equal to or earlier than the second value of the second time indicator to be stored to storage with the second durability level.

14. The method of claim 1, ~~wherein the first indicator comprises a first time value, and~~ wherein updating the first indicator comprises incrementing the first time value to a value smaller than a current working time value maintained by a working counter to cause a set of data units associated with time intervals equal to or less than the value of the first indicator to be stored to storage with the first durability level.

15. Software stored in a non-transitory form on a computer-readable medium, managing data in a distributed data processing system including a plurality of processing nodes, the software including instructions for causing a computing system to perform all the steps of any one of claims 1-14.

16. An apparatus for managing data in a distributed data processing system including a plurality of processing nodes, the apparatus including:

a plurality of data stores in the distributed data processing system, each processing node of the plurality of processing nodes maintaining a respective subset of the data stores, 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~~ having a durability level of a plurality of durability levels, at least some processing nodes each maintaining data stores at multiple durability levels, the plurality of durability levels including a first durability level and a second durability level with a greater degree of durability than the first durability level,

one or more processors for processing a plurality of sets of data units using two or more of the plurality of processing nodes, each data unit of each set of data units being associated with a corresponding time interval of a plurality of time intervals, the plurality of sets of data units including a first set of data units associated with a particular time interval of the plurality of time intervals, the processing including:

maintaining by a root node from the plurality of processing nodes, a plurality of time indicators for the processing system, each time indicator corresponding to a respective different durability level, including maintaining a first time indicator corresponding to the first durability level;

assigning each data unit a time interval of a plurality of non-overlapping time intervals and assigning a first set of data units to a given time interval; and

storing data units in data stores at multiple levels of durability, including storing at least some data units in data stores having the first durability level and storing at least some data units in data stores having the second durability level;

wherein maintaining the plurality of time indicators includes updating and distributing, by the root node to the plurality of processing nodes, the first time indicator to represent a time value such that the data units assigned a time interval represented by the time value are stored in a data store having at least the first durability level.

~~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 two or more 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;~~

~~—— updating a first indicator associated with the first durability level of the plurality of durability levels, with the first indicator maintaining an updateable first value representing a first time interval value by which all sets of data units associated with the first time interval value or earlier have been stored at the first durability level, after determining that all data units of the first set of data units associated with the particular time interval have been successfully stored in data stores associated with the first durability level, and~~

~~—— updating a second indicator associated with the second durability level of the plurality of durability levels, with the second indicator maintaining an updateable second value representing a second time interval value, lagging the first time interval value, by which all sets of data units associated with the second time interval or earlier have been stored at the second durability level, after determining that all data units of the first set of data units associated with the particular time interval have been successfully stored in at least one data store associated with the second durability level.~~

17. A computing system for managing data in a distributed data processing system including a plurality of processing nodes, the computing system including:

means for maintaining a plurality of data stores in the system, each processing node of the plurality of processing nodes maintaining a respective subset of the data stores, 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 having a durability level of a plurality of durability levels, at least some processing nodes each maintaining data stores at multiple different durability levels, the plurality of durability levels including a first durability level and a second durability level with a greater degree of durability than the first durability level,

means for maintaining, by a root node, a plurality of time indicators for the processing system, each time indicator corresponding to a respective different durability level, including maintaining a first time indicator corresponding to the first durability level;

means for processing a plurality of sets of data units using two or more of the plurality of processing nodes, the processing including:

assigning each data unit a time interval of a plurality of non-overlapping time intervals and assigning a first set of data units to a given time interval; and

storing data units in data stores at multiple levels of durability, including storing at least some data units in data stores having the first durability level and storing at least some data units in data stores having the second durability level;

wherein maintaining the plurality of time indicators includes updating and distributing, by the root node to the plurality of processing nodes, the first time indicator to represent a time value such that the data units assigned a time interval represented by the time value are stored in a data store having at least the first durability level.

~~each data unit of each set of data units being associated with a corresponding time interval of a plurality of time intervals, the plurality of sets of data units including a first set of data units associated with a particular time interval of the plurality of time intervals, the processing including:~~

- ~~—— 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 two or more 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;~~
- ~~—— updating a first indicator associated with the first durability level of the plurality of durability levels, with the first indicator maintaining an updateable first value representing a first time interval value by which all sets of data units associated with the first time interval value or earlier have been stored at the first durability level, after determining that all data units of the first set of data units associated with the particular time interval have been successfully stored in data stores associated with the first durability level, and~~
- ~~—— updating a second indicator associated with the second durability level of the plurality of durability levels, with the second indicator maintaining an updateable second value representing a second time interval value, lagging the first time interval value, by which all sets of data units associated with the second time interval or earlier have been stored at the second durability level, after determining that all data units of the first set of data units associated with the particular time interval have been successfully stored in at least one data store associated with the second durability level.~~

18. The apparatus of claim 16 or the computing system of claim 17, wherein data stores with the first durability level comprise volatile storage media.

19. The apparatus of claim 16 or 18, or the computing system of claim 17 or 18, wherein data stores with the second durability level comprise persistent storage media.

20 The apparatus of any one of claims 16 or 18 through 19, or the computing system of any one of claims 17 through 19, wherein maintaining the plurality of time indicators includes:

updating the first time indicator to represent a first time value corresponding to the given time interval to indicate that the first set of data units assigned to the given time interval are stored in a data store having the first durability level, and

subsequently updating a second time indicator corresponding to the second durability level to represent the first time value corresponding to the given time interval to indicate that the first set of data units assigned to the given time interval are stored in a data store having the second durability level~~the first durability level is satisfied for the first set of data units prior to the second durability level being satisfied for the first set of data units.~~

21. The apparatus of any one ~~of~~ claims 16 or 18 through 20, or the computing system of any one of claims 17 through 20, wherein the plurality of durability levels includes a third durability level having greater durability than the second durability level.

22. The apparatus or the computing system of claim 21, wherein the processing further includes updating a third time indicator ~~associated with corresponding to~~ the third durability level of the plurality of durability levels to represent the first time value corresponding to the given time interval to indicate that the first set of data units assigned to the given time interval are stored in a data store having ~~all sets of data units associated with the first time interval are stored at~~ the third durability level, ~~after all data units of the first set of data units are successfully stored in at least one data store associated with the third durability level.~~

23. The apparatus ~~of claim 16~~ or the computing system of claim ~~2017~~, wherein updating the first time indicator includes incrementing the first time indicator to a first value representative of a time earlier than a current working time maintained by a working counter, wherein updating the second time indicator includes incrementing the second time indicator to a second value representative of another time earlier than the current working time or the time represented by the first value of the first time indicator, and wherein the method further includes broadcasting a message containing at least the first time indicator, the second time indicator and the working counter to the plurality of processing nodes to cause a set of data units associated with time intervals equal to or earlier than the value of the first time indicator to be stored to storage with the first durability level, and to cause another set of data units associated with time intervals equal to or earlier than the second value of the second time indicator to be stored to storage with the second durability level.