

- 1 -

Database Management System

DescriptionTechnical Field

5 **[0001]** This invention generally relates to database management systems. More specifically this invention relates a method and apparatus for implementing a multi-user, elastic, on-demand, distributed relational database management system characterized atomicity, performance and scalability.

Background Art

10 **[0002]** Over the past years the use of databases for storing and retrieving messages has emerged as an important tool in a wide variety of commercial applications. Initially, many database systems operated on a single server installation with multiple users. However, various factors have developed over the past years that have required the basic nature of database architecture to change. As a first factor, database storage requirements have become extremely
15 large. Second, the number of users trying to access such databases has also become large. Third, the use of databases to retrieve relatively stable data with minimal updates has been replaced by transactional processing.

[0003] A transaction is a unit of work must be completed in its entirety. A single transaction may include multiple data manipulations. As an example a single transaction may
20 include a reading operation followed by a write operation. In recent years significant effort has been directed to enabling relational databases to support ever increasing rates of transaction processing.

[0004] Databases are now judged by a standard that defines ACID properties, namely: atomicity, consistency, isolation and durability. Atomicity guarantees that all transaction tasks
25 will be completed in their entirety. Consistency assures is that only valid data is written to the database. Isolation assures that other operations cannot access or "see" data in an intermediate state during a transaction. Durability assures that once a transaction has been processed successfully, it cannot be undone.

[0005] Consistency is particularly important in multi-user systems where it is possible
30 for two or more users to seek concurrent access to shared volatile data. Early multi-user systems used locking operations to assure consistency. Locks could be exclusive, or write, locks, or non-exclusive, or read, locks and could be applied to individual records or to pages. However, as

- 2 -

databases have grown in size and as transaction rates have increased, the overhead for managing locks has become significant and, in some cases, prohibitive.

[0006] Multi-version concurrency control (MVCC) is an alternative process for assuring concurrency. MVCC can be more effective than locks with complex databases. MVCC uses timestamps or increasing transaction identifications (IDs) to serialize different versions of a record. Each version permits a transaction to read the most recent version of an object which precedes the timestamp or ID. With this control method, any change to a record, for example, will not be seen by other users until the change is committed. MVCC also eliminates locks with other attendant overhead and establishes a system in which read operations can not block write operations.

[0007] In addition to meeting the ACID tests, there now is a requirement for continuous availability to users. Some database systems dedicate one computer system to transaction processing and another to decision support and other reporting processes. They are interconnected so that other functions can be supported simultaneously. As databases grow in size and complexity, existing data processing systems are replaced by more powerful data processing system. Another approach for accommodating growth involves replicated systems where one machine is designated as a "head" machine that keeps all the replicated machines in synchronism. If a head machine were to fail, a process would assign that function to another replicated machine. Different replicated machines are available to certain users. This approach is not scalable because all the machines have to have the same capability.

[0008] As another approach, multiple autonomous database systems can be integrated into a single "federated" database with a computer network interconnecting the various individual databases. Federated databases require "middleware" to maintain the constituent databases in synchronism. This "middleware" can become very complex. As the database size increases, the resources required for operating the middleware may impose such a sufficiently great overhead that overall system performance degrades.

[0009] "Partitioning" is another approach for implementing databases in which a logical database or its constituent elements are divided into distinct independent parts. In a distributed database management system, each partition may be spread over multiple nodes. Users at given node can perform local transactions on the partition. Partitioning also can be implemented by forming smaller databases or by splitting selected elements of just one table.

- 3 -

[0010] There are two general approaches to partitioning. In horizontal partitioning, also called “sharding”, different rows are placed in different tables and different servers. Generally they have a certain commonality such as a range of zip codes or last names which are divided into different tables by ranges. For example a first database might contain all the records for last names in the range A through M; a second database, in the range N through Z. Sharding, which is a form of horizontal partitioning, involves locating rows of a database on separate servers. Sharding does reduce the number of rows in each table and increases search performance. However, sharding uses a hash code at an application level that makes it much more difficult to implement. It also incorporates a two-phase commit. The complexities of sharding make it suitable for particular applications as the basis for defining the shards is quite well defined.

[0011] Vertical partitioning involves the creation of tables with fewer columns and splitting columns across tables. Like a federated database, vertical partitioning requires middleware to determine how to route any request for a particular field to an appropriate partition. In addition these systems operate using a two-phase commit sequence which is complicated to implement.

[0012] In still another approach, known as a “shared-nothing” architecture, each node is independent and self-sufficient. Shared-nothing architecture is popular for web development because it can be scaled upward simply by adding nodes in the form of inexpensive computers. This approach is popular in data warehousing applications where updates tend to be less frequent than would occur with transaction processing. However, the processing of joins is very complex over large data sets from different partitions or machines.

[0013] Some database systems are referred to as “distributed” systems. One implementation of a distributed system incorporates “clusters” and two communications paths. A high-speed Internet path carries data among the clusters. High-speed dedicated communications paths are required for various control functions, such as lock management. While this approach solves the redundancy and availability issues for databases, lock management, as previously discussed, can limit system performance.

[0014] In a “shared everything” system, super high-speed communications keep the system in synchronism. However lock management can require significant bandwidth resources. To avoid this, such systems incorporate point-to-point communications channels and a very sophisticated disk controller.

- 4 -

[0015] Collectively, those prior art systems satisfy some but not all of the known requirements for a database system. What is needed is a database architecture that is scalable, that meets the ACID properties of atomicity, consistency, isolation and durability. What is also needed is a database system that operates over the Internet without the need for dedicated high-speed communications paths, that provides transaction processing and that is operable over a wide geographic area.

Disclosure of Invention

[0016] Therefore, it is an object of this invention to provide an elastic, scalable, on-demand, distributed data processing system.

[0017] Another object of this invention is to provide an elastic, scalable, on-demand, distributed data processing system that is fault tolerant.

[0018] Still another object of this invention is to provide an elastic, scalable, on-demand, distributed data processing system that has a high degree of availability.

[0019] Yet another object of this invention is to provide an elastic, scalable, on-demand, distributed data processing system that is platform independent.

[0020] Still yet other object of this invention is to provide an elastic, scalable, on-demand, distributed data processing system that is atomic, consistent, isolated and durable.

[0021] Yet still another object of this invention to provide an elastic, scalable, on-demand, distributed data processing system that operates over the Internet without a need for dedicated high-speed communications paths.

[0022] Still yet another object of this invention is to provide an elastic, scalable, on-demand, distributed data processing system that provides transaction processing and that is adapted for implementation over a wide geographic area.

[0023] In accordance with one aspect of this invention a database management system that enables users to interact with a database comprised of data and metadata comprises a plurality of nodes and persistent storage with communications paths therebetween. Each node includes an interface between high level-input and output commands at a user level and input and output commands at the system level that control a sequence of operations for interacting with the database, wherein, in response to certain system level commands, atom objects generate atoms, each atom containing a specified fragment of data or metadata whereby a set of all instances of atoms collectively define all the metadata and data in the database. Each node

- 5 -

additionally includes a communications control for establishing a communications path with each other node in the system, a method responsive to a system command from the interface for requesting from a selected node copy of an atom that is relevant to the query but is not present in that node, a method responsive to a request for an atom from another node for replicating the requested atom for transfer to the requesting node whereby only atoms required to complete a query need be located in any transaction node at any given time, and a method responsive to a change in an atom at that node for replicating that atom for transfer to every other node in the system in which that atom is resident. The persistent storage contains a collection of atoms that collectively contain all the data and metadata in the database.

[0024] In accordance with another aspect of this invention database management system that enables users to interact with a database comprised of data and metadata, said system includes at least one transactional node that provide users with access to the database and at least one archival node that maintains an archive of the entire database. Each transactional node includes a database request engine that provides an interface between high level-input and output query commands at a user level and input and output commands at the system level that control a sequence of operations for interacting with the database. In response to certain system level commands, atom objects generate atoms. Each atom contains a specified fragment of data or metadata whereby a set of all instances of atoms collectively define all the metadata and data in the database. A database system network interconnects all the nodes. A communications control in each of the nodes for establishes a communications path with each other node in the system, A method in each transactional node responds to a system command from the database request engine for requesting a copy of an atom that is relevant to the query command but is not present in that node. Another method in each node responds to a request for an atom from another node for replicating the requested atom for transfer to the requesting node whereby only atoms required to complete a query command need be located in any transactional node at any given time. Another method in each transactional node responds to a change in an atom at that node for replicating that change to every other node in the system that contains a copy of that atom

[0025] In accordance with yet another aspect of this invention a database management system for a logical database comprised of data records organized into tables that is to be accessed from multiple transactional nodes that process transactions related to the logical data base wherein the database is parsed into fragments wherein each fragment stores a portion of the

- 6 -

metadata and/or data pertaining to the logical data base for transfer within the database management system as serialized messages and for storage as a de-serialized message. The system includes at least one archival node that stores all the fragments in a de-serialized form in permanent storage thereby to constitute a single store for the entire data base. Each transactional node responds to queries from users by establishing a sequence of low-level commands for identifying fragments that are relevant to the query and responds to the low-level commands by obtaining only those copies of existing fragments that are relevant to the query being processed thereat whereby a given fragment might exist at some other node or only at an archival node. Each transactional node replicates any changed fragment to the at least one archival node and each transactional node in which a copy of that fragment resides whereby changes are made to fragments in other nodes on a peer-to-peer basis and whereby any transactional node only contains those fragments that pertain to the queries being made users accessing the data base through that transactional node.

Brief Description of the Drawings

[0026] The appended claims particularly point out and distinctly claim the subject matter of this invention. The various objects, advantages and novel features of this invention will be more fully apparent from a reading of the following detailed description in conjunction with the accompanying drawings in which like reference numerals refer to like parts, and in which:

[0027] FIG. 1 is a diagram in schematic form of one embodiment of an elastic, scalable, on-demand, distributed data processing system that incorporates this invention with interconnected transactional and archival nodes;

[0028] FIG. 2 depicts the organization of a transactional node;

[0029] FIG. 3 depicts the organization of an archival node;

[0030] FIGS. 4A and 4B depict the logical organization of "atom" objects generated by atom classes shown in FIGS. 2 and 3 that are useful in implementing this invention and they might appear at any given time in a transactional node;

[0031] FIG. 5 depicts the information in a Master Catalog atom;

[0032] FIG. 6 depicts the information in a Transaction Manager atom;

[0033] FIG. 7 depicts the information in a Database atom;

[0034] FIG. 8 depicts the information in a Schema atom;

[0035] FIG. 9 depicts the information in a Table atom;

- 7 -

[0036] FIG. 10 depicts the information in a Table Catalog atom;

[0037] FIG. 11 depicts the information in an Index atom;

[0038] FIG. 12 depicts the information in a Record States atom;

[0039] FIG. 13 depicts the information in a Data atom;

5 [0040] FIG. 14 depicts the information in a Blob States atom;

[0041] FIG. 15 depicts the information in a Blob atom;

[0042] FIG. 16 depicts the syntax of a exemplary asynchronous message that is transferred among the transactional and archival nodes of the database system of FIG. 1;

10 [0043] FIG. 17 depicts various messages types by which information is transferred among the transactional and archival nodes of the database system of FIG. 1;

[0044] FIG. 18 is a flow diagram useful in understanding the method by which a node joins the database system in FIG. 1;

[0045] FIG. 19 depicts the information in a node object;

15 [0046] FIG. 20 is a flow diagram useful in understanding the method by which a node creates an atom in accordance with this invention;

[0047] FIG. 21 is a flow diagram useful in understanding the method by which unique atom identifications are assigned during the method of FIG. 20;

[0048] FIG. 22 is a flow diagram useful in understanding the method by which one node obtains a copy of an atom from another node; and

20 [0049] FIG. 23 is a flow diagram useful in understanding a method by which this invention commits a transaction.

Best Mode for Carrying Out the Invention

[0050] FIG. 1 depicts one embodiment of an elastic, scalable, on-demand, distributed database system 30 with a plurality of data processing nodes that incorporates this invention.

25 Nodes N1 through N6 are “transactional nodes” that provide user applications access to the database; nodes A1 and A2, “archival nodes” that function to maintain a disk archive of the entire database at each archival node. While an archival node normally stores the entire database, a single transactional node contains only that portion of the database it determines to be necessary to support transactions being performed at that node at that time.

30 [0051] Each node in FIG. 1 can communicate directly with each other node in the system 30 through a database system network 31. For example, node N1 can establish a communications

- 8 -

path with each of nodes N2 through N6, A1 and A2. Communications between any two nodes is by way of serialized messages. In a preferred embodiment, the messaging is performed in an asynchronous manner to maximize the bandwidth used by the system thereby to perform various operations in a timely and prompt manner. Typically the database system network 31 will
5 operate with a combination of high-bandwidth, low-latency paths (e.g., an Ethernet network) and high-bandwidth, high-latency paths (e.g., a WAN network). Each node has the capability to restrict use of a low-latency path to time-critical communications (e.g., fetching an atom). The high-latency path can be used for non-critical communications (e.g. a request to update information for a table). Also and preferably, the data processing network of this invention
10 incorporates a messaging protocol, such as the Transmission Control Protocol (TCP), and assures that each node processes messages in the same sequence in which they were sent to it by other nodes.

[0052] FIG. 2 depicts a representative transactional node 32 that links to the database system network 31 and various end users 33. The transactional node 32 includes a central
15 processing system (CP) 34 that communicates with the database system network 31 through a network interface 35 and with the various users through and user network interface 37. The central processing system 34 also interacts with RAM memory 38 that contains a copy of the database management program that implements a preferred embodiment of this invention. This program functions to provide a remote interface 40, a database request engine 41 and a set 42 of
20 classes or objects.

[0053] The database request engine 41 only exists on transactional nodes and is the interface between the high-level input and output commands at the user level and system level input and output commands at the system level. In general terms, its database request engine parses, compiles and optimizes user queries such as SQL queries into commands that are
25 interpreted by the various classes or objects in the set 42.

[0054] For purposes of the explanation of this invention, the set classes/objects set 42 is divided into a subset 43 of "atom classes," a subset 44 of "message classes" and a subset 45 of "helper classes." Additional details of these classes are described later.

[0055] As will become apparent and in accordance with this invention, at any given time
30 a transactional node only contains those portions of the total database that are then relevant to active user applications. Moreover, the various features of this invention enable all portions of

- 9 -

database in use to be resident in random access memory 38. There is no need for providing supplementary storage, such as disk storage, at a transactional node during the operation of this system.

[0056] Referring to FIG. 3, each archival node 50, such as archival node A1 or A2 in FIG. 1, also connects to the database system network 31. However, in place of end users 33 associated with a transactional node 32 in FIG. 2, an archival node connects only to persistent storage 51, typically a disk-based storage system or a key value store. The archival node 50 includes a central processing system 54 that communicates with the persistent storage 51 through an I/O channel 52 and with the database system network 31 through a network interface 55. The central processing system 54 also interacts with RAM memory 57 that contains a set 62 of classes or objects. Similarly to the transactional node 32 in FIG. 2, the classes/objects set 62 in FIG. 3 includes a set 63 of "atom classes," a set 64 of "message classes" and a set 65 of "helper classes."

[0057] A preferred embodiment of this invention uses object-oriented programming (OOP) wherein, as known in the art, classes and subclasses such as shown in FIGS. 2 and 3 define methods, data structures and processes by which an "instance", or object, of that class or subclass may be generated. An "instance" may be generated using "inheritance" and/or "polymorphism." It will become apparent to those skilled in the art that implementations are possible that do not use object-oriented programming or variations on the specifically disclosed embodiment.

[0058] This invention will now be described in several phases. An "Atoms" section defines the hierarchy and function of objects produced by the atom classes 43 and 63 in FIGS. 2 and 3 respectively. A "Messages" section describes a set of messages which provide communications among the transactional and archival nodes as might be produced by message classes 44 and 64 in FIGS. 2 and 3 respectively. A "Methods" section describes basic operations with respect to database management. An "Example" section describes the interaction of the atom, message and methods by which the objectives of this invention are achieved in response to a specific database query to the database request engine 41.

Atoms

[0059] As previously stated, each of the atom classes 43 in FIG. 2 and 63 in FIG. 3 produce "atoms". More specifically, the atom classes define one or more "atom types" or "atom

- 10 -

objects.” Each "atom type" or “atom object” produces an "instance" of itself, that is, an "atom."

As will become apparent with a more detailed understanding of the purpose of each specific atom object, each "atom" contains a specific fragment of database information. Some atoms contain a portion of the database metadata; others contain data records; still others serve as catalogs that create and track other atom types. Some "atom types" may only instantiate one atom which replicates to all nodes. Other "atom types" may instantiate multiple atoms which are replicated to other nodes on an as-needed basis.

[0060] Atoms have certain characteristics. In a transactional node, an atom exists only in non-persistent memory and in the form of a de-serialized message that has populated a particular atom type to provide an efficient memory resident format for the atom. Each atom has means for encoding its content into a serialized message and means for decoding a serialized message to retrieve the contents of the atom. Such serialized messages are used in connection with a number of operations as will be described later.

[0061] Each serialized message transmitted from a node for replicating an atom includes the content of that atom with an appended node identification and the most recent transaction commit sequence number for that node. When an archival node receives that serialized message, it de-serializes the message, removes the node list and commit sequence number before placing the remaining content of the message (i.e., an atom) in persistent storage.

[0062] A number of rules apply to atoms in accordance with this invention. The reasons and implications of those rules will become more apparent. First, each atom must have a unique identification to provide a reliable identification of that atom anywhere in the database processing network 30 in FIG. 1. Second, any atom must exist in two nodes simultaneously to maintain redundancy, except that a single atom can exist after it is created and before an archive node has requested a copy. Third, a transactional node will load an atom only on demand. Fourth, each time a change is made to an atom in one node, that node must replicate that changed atom on a "peer-to-peer" basis; i.e., to all archival nodes and only those transactional nodes that contain that same atom.

[0063] A “garbage collection” process, described in greater detail later, can occur on archival and transactional nodes. The process removes inactive atoms from transactional and archival nodes. As a result, a transactional node can store those atoms that are then currently relevant to user applications in random access memory at that node. Thus, the database request

- 11 -

engine 41 "sees" the entire database as being local and is not aware that it is operating in a multi-node environment and without a complete copy of the database at its node. Archival nodes have the option of purging the contents of an atom after it has been serialized to disk thereby reducing the size of the memory required for storage. If an archival node receives a replication message
5 for such an atom, the archival node must fetch the content from disk storage before applying the information from the atom being replicated.

[0064] With this general background, each atom type will now be described at a "logical" or functional level. This information, in conjunction with further discussions of the operation of this invention, will enable a person of ordinary skill in the art to make and use this
10 invention in any of a variety of implementations, including implementations based upon object oriented programming.

[0065] FIGS. 4A and 4B depict the database engine 41 and typical set of atoms that might reside in a transactional node 32 at any given time. In this example, the transactional node hosts a Master Catalog atom 70, a Transaction Manager atom 71, a Database atom 72, a Schema
15 atom 73, a Table atom 74 and a Table Catalog atom 75. There is only one Master Catalog atom 70, one Transaction Manager atom 71 and one Database atom 72 per database. The Transaction Manager atom 71, the Database atom 72 and the Schema atom 73 are created when a database request engine 41 creates a new database.

[0066] Referring to FIG. 4A, a Master Catalog atom 70 tracks the status of transactional and archival nodes in database system 30 of FIG. 1. It also can be considered as an active index
20 that creates and monitors the Transaction Manager atom 71, the Database atom 72, each Schema atom 73, each corresponding set of Table atoms 74 and Table Catalog atoms 75, and Sequence ID Manager 82.

[0067] The Table Catalog atom 75 acts as an active index and creates and monitors Index
25 atoms 76, Record States atoms 77, Data atoms 78, Blob States atoms 80 and Blob atoms 81 associated with a single table. That is, there is one Table Catalog atom 75 for each table.

[0068] FIG. 4B is useful in understanding the interaction and management of different atom types. In this context, neither the Master Catalog atom 70 nor the Table Catalog atom 75 performs any management functions. With respect to the remaining atoms, the Database atom
30 70 manages each Schema atom 73. Each Schema atom 73 manages each related Table atom 74 and Sequence ID Manager atoms 82. Each Table atom 74 manages its corresponding Table

- 12 -

Catalog atom 75, Index atoms 76, Record States atoms 77, Data atoms 78, Blob States atom 80 and Blob atoms 87.

[0069] Still referring to FIG. 4B, the database request engine 41 communicates with the Master Catalog atom 70, Transaction Manager atom 71, the Database atom 72, each Schema atom 73, each Table atom 74 and the Sequence ID Managers 82. The database request engine 41 acts as compiler for a high-level language such as SQL. As a compiler, it parses, compiles and optimizes queries and obtains metadata and data from atoms for the formation of the various fragments of data base information.

[0070] Each atom has certain common elements and other elements that are specific to its type. Referring to FIG. 5, a Master Catalog atom 70 includes common elements 70A through 70I. Element 70A is a unique identification for the atom. As there is only one instance of a Master Catalog atom that is replicated to all nodes, the Master Catalog atom ID 70A is given a fixed number, typically "1." As a general rule, pointers 70B and 70C identify the Master Catalog atom and the creating catalog atom, respectively. For the Master Catalog atom, both pointers identify the Master Catalog atom itself.

[0071] Each atom must have a chairman. The chairman performs functions as described later. Element 70D is a pointer to that node where the chairman for that atom resides.

[0072] Each time a copy of an atom is changed at any transactional node, it receives a new change number. Element 70E records that change number.

[0073] Whenever a node requests an atom from another node, there is an interval during which time the requesting node will not be known to other transactional nodes. Element 70F is a list of all the nodes for which the supplying node must relay messages to the requesting node from all other nodes that contain that atom until the request is completed.

[0074] Operations of the database system are also divided into cycles. A cycle reference element 70G provides the cycle number of the last access to the atom. Element 70H is a list of all active nodes that contain the atom. Element 70I includes several status indicators.

[0075] Still referring to FIG. 5, a global node IDS entry 70J contains an ID manger for assigning a unique identifier for each active node in the system. As known, such identifiers are long character strings. A local node IDS entry 70K includes a range of numbers up to the total number of nodes that can attach to the system. Together these entries provide the

- 13 -

correspondence between the two types of identification. Use of local node IDS increases efficiency.

[0076] When a transactional node joins the database system 30 in FIG. 1, a connection manager 70L effects that process. A specific implementation for enabling the transactional node to join the database system is described later. The joining node uses a pending node status entry 70M to indicate that it will not have a global address for a responding node until it receives further communications from that responding node. A database UUI entry 70N contains the universal unique identification for the database.

[0077] The entries at 70P are important because they link all the atoms for which the Master Catalog atom 70 acts as an active index. As previously stated, these include the Database atom 72 and each of Schema atoms 73, Table atoms 74 and Table Catalog atoms 75.

[0078] A password entry 70Q is representative of a means for authenticating a connection into the database. Actual and software version entries 70R and 70S allow the system to operate with backward compatibility when a newer version of the software is installed. The actual software entry 70R identifies the software version then in use; the software version entry 70S, the number corresponding to the most recent installed version. This permits individual nodes to be updated to newer versions without requiring other nodes to be updated and without closing the database for access by all nodes.

[0079] Still referring to FIG. 5, the Master Catalog atom 70 also includes a pointer 70T to the Transaction Manager atom 71, a pointer 70U to a configuration object, a pointer 70V to a garbage collection thread and a pointer 70W to a ping thread. The ping thread operates periodically and independently of other operations at a node. It "pings" each other node to provide information that can be used in the determination of communications efficiency of a corresponding path. For example, if node N1 in FIG. 1 has the option of communicating with either node N2 or N5, node N1 could use the ping information in the selection of the most efficient of the communications paths to nodes N2 and N5 for that communication. Other selection processes might also be substituted or added.

[0080] Referring to FIG. 6, there is one Transaction Manager atom 71 for each database and it is created during the same process that creates the Master Catalog atom 70. The Transaction Manager atom 71 creates, tracks and ends database transactions in response to database commands from the database request engine 41. A Transaction Manager atom 71

includes elements 71A-71I that correspond to like elements in the Master Catalog atom.

However, element 71A is the identification of the Transaction Manager atom 71. Elements 71B and 71C both point to the Master Catalog atom 70.

[0081] An ID manager 71J provides unique transaction sequence identifications and maintains an active transactions list 71K, a committed transactions list 71L and a failed transactions list 71M. Element 71N stores commit sequence information. The ID manager 71J assigns a transaction ID at the start of each transaction. Each transaction ID is unique, but not necessarily sequential. A local Transaction Manager atom assigns a commit sequence number to element 71N when a transaction commits. Sequence numbers are sequential and each is specific to the node that requested the transaction. A transaction-transition event counter 71P identifies discrete events that occur during each transaction, such as the start of the transaction and the successful committing of the transaction. Such counters are useful when multiple transactions involving the same information overlap.

[0082] Referring to FIG. 7, the Database atom 72 is created at the same time the Master Catalog atom 70 and Transaction Manager atom 71 are created. The Database atom 72 identifies each of the Schema atoms 73. The Database atom 72 may be involved with an authentication process when a new user attempts to join the database. It may also include other data concerning the authorization levels.

[0083] Basically, the Database atom 72 includes elements 72A-72I corresponding to like elements in FIG. 5. Element 72A is the Database atom identification. Each of pointers 72B and 72C identify the Master Catalog atom 70. A schema name-schema ID registry 72J relates schema names to Schema atom identifications.

[0084] Referring to FIG. 8, a Schema atom 73 creates and tracks Table atoms for that schema. A Database atom 72 may manage multiple Schema atoms and each Schema atom may interact with multiple Table atoms. The Schema atom 73 includes elements 73A-73I corresponding to elements 70A-70I in FIG. 5. Element 73A is the unique Schema atom identification 73A and elements 73B and 73C are pointers to the Master Catalog atom 70. A table has an unique name within a schema. A table name-table atom ID registry 73J provides the correspondences between each table name and the corresponding table atom. Each schema sequence has a name. A sequence name-sequence ID manager registry 73K provides the

relationship between those names and corresponding sequence ID managers associated with each Schema atom, such as sequence ID managers 82 in FIGS. 4A and 4B.

[0085] FIG. 9 provides a logical view of a Table atom 74 that incorporates metadata relating to fields, formats, indices and types and that manages each of the Index atoms 76,

5 Record States atoms 77 and Blob States atoms 80 for that table. It also creates and tracks data within a table. Table atom 74 includes elements 74A-74I that correspond to elements 70A-70I in FIG. 5. Element 74A includes the unique Table atom identification, elements 74B and 74C both point to the Master Catalog atom. Pointer 74J identifies the corresponding Table Catalog atom. Element 74K contains a list of all fields for the table.

10 **[0086]** Each Table atom has several ID managers. A pointer 74L points to an ID manager that provides each field with a unique identification. Pointers 74M, 74N, 74P and 74Q identify separate ID managers for assigning identifications to Index atoms, Data atoms, Blob atoms and subtypes, respectively. Element 74R is a list of the existing subtypes. Arrays 74S and 74T provide the locations of Record States atoms and Blob States atoms, respectively.

15 **[0087]** Now referring to FIG. 10, there is one Table Catalog atom for each Table atom. Each Table Catalog atom 75 is created when a Table atom is created. In turn, a Table Catalog atom creates and tracks atoms specific to one table including the Index, Records States, Data, Blob States and Blob atoms. Each Table Catalog atom 75 includes elements 75A-75I that correspond to elements 70A-70I in FIG. 5. Element 75A is a unique Table Catalog atom identification as assigned by the Master Catalog atom. Both elements 75B and 75C point to the Master Catalog atom 70. An ID manager 75J provides unique atom identifications for each of the Index, Record States, Data, Blob States and Blob atoms. A list 75K identifies all atoms associated with a corresponding Table atom. Pointers in element 75L identify the location of each atom at the local node associated with the corresponding Table atom. A series of lists 75M
20 identify, for each atom, a list of the nodes with replications of that atom. Bitmaps 75N provide a convenient means for identifying other objects and directories when the atom is in an archival node.

[0088] Referring to FIG. 11, there is one Index atom 76 for each index in a table, and there may be multiple Index atoms per table. Each Index atom includes elements 76A-76I that
30 correspond to elements 70A-70I in FIG. 5, respectively. Element 76A is the unique Index atom identification as assigned by the corresponding Table Catalog atom. Pointers 76B and 76C

identify the Master Catalog atom and the Table Catalog atom, respectively. Element 76J contains a binary tree of index nodes to provide a conventional indexing function. Element 76K contains the index level. Such index structures and operations are known to those skilled in the art.

5 **[0089]** Referring to FIG. 12, a Record States atom 77 manages record versions and state for a fixed range of record numbers within a single table. Thus, a given Table atom may manage multiple Record States atoms. Each Record States atom includes elements 77A-77I that correspond to elements 70A-70I in FIG. 5. Element 77A includes the Record States atom ID that the creating Table Catalog atom assigns. Pointers 77B and 77C identify the Master Catalog and
10 the Table Catalog atoms, respectively. Element 77J is an array for locating all the Data atoms managed by the Record States atom 77. Element 77K contains the record number for a "base record". That is, each Data atom stores multiple records. Element 77 is a pointer to the corresponding Table atom.

[0090] In database applications to which this invention is directed multiple users can
15 produce multiple versions of the same record. A preferred embodiment of this invention uses multi-version concurrency control (MVCC) to ensure that a transaction never has to wait for the database by allowing several versions of a record or other object to exist in the database simultaneously. Consequently, each Record States atom 77 includes metadata about each record version. Entry 77M is a bitmap for identifying the location of each versioned record that is
20 useful in garbage collection.

[0091] The Record States atom 77 includes, for each version 77N of a record, a transaction ID 77P to identify the transaction that generated the version. A format version entry 77Q identifies the version number for the table sub-type that existed when the record was inserted. This format identifies the physical order of the record and identifies the subtype to
25 which the record belongs of the database program that was in use at the time the record version was created. Element 77R includes a record version sequence number; element 77S, the location of the next older, or previous, version of the record. An index 77T to the data atom array 77J and the base record identification 77K together provide the address for the actual slot 77U in the Data atom with the record version.

30 **[0092]** FIG. 13 depicts a Data atom 78 with elements 78A-78I that correspond to elements 70A-70I in FIG. 5. In the Data atom 78 element 78A is the Data atom identification

- 17 -

78A assigned by the Table Catalog atom. Elements 78B and 78C are pointers to the Master Catalog atom and to the corresponding Table Catalog atom, respectively. An ID manager 78J assigns a record slot identification for each record in the Data atom 78. Element 78K identifies, for each record in the Data atom 78, the address and length of that record. Element 78C represents data records and versions thereof.

[0093] Now referring to FIG. 14, databases also store "blob records." A "blob record" typically is a collection of binary data stored as a single entity in the database. Blob records do not exist in versions. A Blob State atom includes elements 80A-80I that correspond to elements 70A-70I in FIG. 5. Element 80A has the Blob States atom unique atom identification. Elements 80B and 80C are pointers to the Master Catalog and the Table Catalog atoms, respectively. List 80J identifies all the Blob atoms managed by a single Blob States atom 80. Entry 80K provides the identification of the base Blob record. Element 80L points to the corresponding Table atom. For each blob record, the Blob States atom includes an index 80M to the Blobs atom. Element 80N identifies the slot in a Blob atom for the blob record.

[0094] FIG. 15 depicts a Blob atom 81 with elements 81A-81I that correspond to elements 70A-70I in FIG. 5, respectively. Element 81A is the atom identification assigned by the Table Catalog atom. Elements 81B and 81C are pointers to the Master Catalog atom and to the corresponding Table Catalog atom, respectively. An ID manager 81J assigns a blob slot identification to each blob in the Blob atom 81. Element 81K identifies, for each blob in the Blob atom 78, its address and length. Element 81L represents all the blob records assigned to the Blob atom.

[0095] In summary, each atom has a relationship with only a fragment of the database. For example, the Database atom 72 contains metadata that identifies the schema for the database. Each Schema atom 73 contains metadata that identifies all the tables associated with that schema. For each table, a Table atom 74 and a corresponding Table Catalog atom 75 provide metadata about the table including such information as the identification of fields and their properties. Record States atoms include metadata about a group of records. Data atoms include information about each data record with pointers to slots that contain these records and various versions. Blob States and Blob atoms contain similar information about blob records.

Messages

[0096] As previously indicated, communications between any two nodes is by way of serialized messages which are transmitted asynchronously using the TCP or another protocol with controls to maintain messaging sequences. FIG. 16 depicts the basic syntax of a typical message 90 that includes a variable-length header 91 and a variable-length body 92. The header 91 includes a message identifier code 93 that specifies the message and its function. As this invention envisions a scenario under which different nodes may operate with different software versions, the header 91 also includes an identification 94 of the software version that created the message. The remaining elements in the header include a local identification 95 of the sender (i.e., from the Master Catalog atom in FIG. 5) and information 96 for the destination of the message, namely a catalog atom (e.g., a Table Catalog ID 75A in FIG. 10) and an atom identification 97 (e.g., a Record States ID 77A in FIG. 12). From this information, the recipient node can de-serialize, decode and process the message.

[0097] FIG. 17 depicts a set of messages having the syntax of FIG. 16 for a specific embodiment of this invention. Each performs a specific function as will now be described.

[0098] As briefly discussed previously, when a message is to be sent, there are different communications paths to different nodes. For example, if one node, as a requesting node, needs to obtain an atom, replications of that atom may be located in multiple other nodes. In this embodiment, "pinging" provides selection information useful for selecting the best path corresponding node. Pinging, as known, involves determining time for a "ping" command to reach its destination and for an acknowledgement message to be received. In this embodiment of the invention, each node periodically uses a helper class to send a Ping message 110 to each of the other nodes to which it connects. Each receiving node uses a helper class to return a Ping Acknowledge message 111 that contains the ping time. Each node accumulates this information about the time to send and receive these messages in a node object described later with respect to FIG. 8. When one node prepares to send a message to one of multiple nodes, the transmitting node analyzes factors including, but not limited to, the accumulated ping data to select one of the nodes as the receiving node for that message.

[0099] A next set of messages are involved with the connection of a new node into or previously inactive node back into the database system 30 in FIG. 1. When such a node, e.g., transactional node N2, wishes to connect to the database system 30, it initiates a connection process described in detail later with respect to FIG. 19. Referring to FIG. 17, once that process

identifies an active node for receiving messages, the joining node sends a Connect message 112 to the selected node. The selected node returns a Welcome message 113 and a New Node message 114 to all other connected nodes in the database system 30. Each of the other connected nodes transmits its own Welcome message 113 to the joining node. When this message sequence completes, the joining node can then take further steps to obtain various atoms.

[0100] Archival nodes can operate in either an active mode or a synchronizing mode when they are synchronizing to another node. A helper class in the archival node transmits a Node State message 115 to alert all other nodes of any status change in that archival node.

[0101] A next set of messages are involved when a node, as a requesting node, retrieves a copy of an atom from another node. For example, after a node joins the database system 30 in FIG. 1, it generally requests a copy of the Master Catalog atom. The process is described in greater detail in connection with the explanation of FIGS. 20 and 21.

[0102] Still referring to FIG. 17, the requesting node issues an Object Request message 116 to a selected node which generally returns an Object message 117 with the requested atom. The selected node also sends an Object Available message 118 to all other nodes with that atom. Each node that receives an Object Available message 118 from the selected node returns an Object Acknowledged message 119 to the selected node. The selected node sends an Object Complete message 120 node to the requesting node after the selected node receives all the Object Acknowledged messages 119.

[0103] In some situations the selected node sends an Object Unavailable message 121 to announce that the selected node has dropped the requested atom. A Bounce Object message from the selected node indicates that the requested atom is not found in the Master Catalog atom or one of the Table Catalog atoms. This may occur when an update transaction is underway and the selected node does not respond to the Object Request message 116 because the garbage collection process has collected that atom before receiving the Object Request message. As a response, the requesting transactional node can select another node in the database system with the atom.

[0104] The database request engine 41 in FIG. 1 and FIGS. 4A and 4B can periodically generate a Register Object message 123 or an Unregister Object message 124. These messages are directed to atoms that include a registry, such as Database and Schema atoms. An Object

1 Delete message 125 is sent from a node when a user at that node issues a command to delete
2 some item, such as a table.

3 **[0105]** Whenever a local node is updated or modified, its change number is incremented.
4 Every replication message contains that atom's local change number. Every atom tracks the
5 most recent change number for each node. When an archival node receives a copy of a modified
6 atom, it copies the change numbers for the atoms, clears the change numbers and then serializes
7 the atom to disk. If a change has been made, the change number will not be zero. The archival
8 node sends an Object Written message 126 with the change number written to each node. Each
9 receiving node compares its own change number with that in the message. Then the node can
10 update the status for the atom to note that the atom has been archived and is a potential candidate
11 for garbage collection.

12 **[0106]** As previously indicated, each atom must have a unique identity. When a first
13 atom of a particular atom type (e.g., a new Table atom) is created, the creating transactional node
14 is designated a Chairman for that Table atom. Rules for governing the "chairmanship" are
15 discussed later. Whenever a transactional node needs to create a new atom type, it sends an ID
16 Request message 127 to the Chairman if it does not have an available identification. The
17 Chairman generally returns an ID Delegation message 128 which includes a block of at least one
18 unique identification from its allocation of free identification values. This process is described
19 in greater detail with respect to the process of FIG. 20.

20 **[0107]** Referring to FIGS. 4A and 4B, a Table atom 74 can send any of a number of
21 messages. If the database request engine 41 initiates a process for adding a new field, the Table
22 atom at that corresponding transactional node generates the new structure and sends a Table
23 Field Added message 129 that replicates the change to all other nodes that include that Table
24 atom. If a transactional node updates a field, a subtype or field characteristic that changes the
25 table format, that node issues a Table Format message 130. A Table Request Records message
26 131 is generated whenever a node needs to create a new Record States or Blob States atom. Only
27 a chairman can create this atom; and the chairman broadcasts a Table Records Object message
28 132 when this occurs.

29 **[0108]** Each time a transactional node inserts a new record in a table, it produces a Table
30 Record message 133. Whenever it becomes necessary to create a new index, as when an indexed
31 field is added to a table, the creating Table atom replicates the new Table atom. At this time, the

1 index is set to be a write-only index. After all the related processes involved have been
2 completed, the node sends a Table Index Added message 134.

3 **[0109]** Any time a Table atom creates a Blob States atom for a blob, the Table atom
4 generates a Table Blobs Object message 135. A Table Blob message 136 indicates that a new
5 blob has been created.

6 **[0110]** In a database that utilizes types and sub-types, the database request engine 41 in
7 FIG. 1 will generate commands to cause a Table atom to assign a new table-type identification.
8 When this occurs, a Table Type message 137 is replicated to all nodes with like Table atoms.

9 **[0111]** A Table Record Expunge message 138 provides the record number on specific
10 Record States atom. A Table Garbage Collect message 139 is generated by a Chairman for a
11 Table atom when it is determined that a record within that table contains an unusually long chain
12 of back versions or other criteria. The result is that "unused" atoms are "emptied" of data.

13 **[0112]** A Record States atom 77 in FIGS. 4A and 4B also produces several specialized
14 messages. If it becomes necessary to update a particular record, the database management
15 system of this invention creates a new version of that record. Referring back to FIG. 17, the
16 corresponding Record States atom generates a Record Update Request message 140 that is
17 directed to the Chairman for that atom to seek permission to update that particular record. The
18 Chairman responds by generating a Records Update Response message 141 that grants or denies
19 permission to update that record. If the Chairman grants permission, the requesting Record
20 States atom performs the update and sends a Records Update message 142 with the new version
21 of the atom to each node with a copy of that atom.

22 **[0113]** A Data atom stores up to some maximum number of records and versions thereof.
23 A corresponding Record States atom monitors the size of the Data atoms it manages. If a Record
24 States atom determines that a managed Data atom has exceeded that size, it generates a new Data
25 atom and replicates that new Data atom by means of a Records Data Object message 143.

26 **[0114]** A Record States atom generates a Records Record message 144 for replicating
27 any new record version. Periodically, a Table atom Chairman initiates a process by which a
28 Record States atom identifies record versions that are older than the oldest active transaction.
29 When this occurs, the Chairman's Record State atom transmits a Records Prune message 145
30 that enables those older record versions to be deleted during a subsequent garbage collection

- 22 -

process. If a situation arises wherein it becomes necessary to rollback a transaction, a Record States atom generates a Backout Record message 146 that updates the backout record.

[0115] There is also a set of index-specific messages. As known an index has an optimum maximum size; and if the index exceeds that size, the index should be split. In accordance with this invention, only a Chairman can split an Index atom. The Chairman can do this unilaterally or in response to an Index Split Request message 147 from another copy of that atom. When the Chairman causes the split, the Chairman generates an Index Split message 148 that contains a split index. After the split, the Chairman sends an Index Prune message 149 to truncate the original Index atom following the split. Each time an index node is added to the index, it generates an Index Node Added message that contains an index key, record identification and other information for the new index. When an index has been fully populated and therefore is ready for use during retrieval operations, a Table Index Ready message 151 is generated. As previously indicated, adding a table index generates an Index atom that is a write-only index and that is not readable. The Table Index Ready message 151 makes such a write-only index readable.

[0116] A Blob States atom identifies a slot in a Blob atom by generating a Blobs Blob message 152.

[0117] A Data Record Message 153 contains the slot number and the record length. It also includes the data record.

[0118] As the state of a transaction changes, the Transaction Manager atom 71 for a given transactional node generates Transaction State messages 154. These indicate whether the current status of the transaction is in an active state, a pre-commit state, a commit state or a rollback state. Under certain circumstances it is possible that a request related to one transaction will be blocked by another transaction on another node. In that event, the Transaction Manager atom receives a Transaction Blockage message 155. If a transaction is deadlocked and must be rolled back, the Transaction Manager atom associated with the node causing the deadlock generates a Transaction Deadlock message 156 causing the transaction to be rolled back.

[0119] The Master Catalog or any Table Catalog atom on an archival node can request the time an atom was last written. This occurs when a requesting node sends a Request Write Times message 158. The recipient of that message then returns those requested times in a Write Times message 159.

1 **[0120]** The foregoing messages constitute a set by which the various procedures needed
2 to maintain a database management system incorporating this invention can be handled properly.

3 As will be apparent, each message has a minimum of overhead in a header. Each message can
4 be reasonably short. When used with TCP or other messaging protocol, the messages must be
5 sent sequentially and, at any given node, on receipt they must be handled in the same sequence
6 in which they were sent even by different nodes.

7 Methods

8 **[0121]** It will now be helpful to a further understanding of this invention to describe
9 some basic methods that relate to various aspects of the operation of this invention. Variations
10 from each will be apparent to those of ordinary skill in the art.

11 **[0122]** FIG. 18 is a flow diagram of the operations that occur in different nodes in the
12 database system 30 of FIG. 1 when a node joins the network. For purposes of this description,
13 assume that Node 5 in FIG. 1 is to join the database system; it is designated to be joining
14 transactional node 170. As a first process, the joining transactional node 170 uses step 171 to
15 establish a TCP connection with a selected node. Basically, the joining transactional node 170
16 sends a message to a fixed location that identifies a database. A connection broker, not shown
17 but as known in the art, responds to this request by denying or granting access to the database
18 system. If the connection broker grants access, it selects a node, such as transactional node N1,
19 as a selected node 172. Then the connection broker sends a message to the joining transactional
20 node 170 with a designation, as by a port number, of the selected node 172. FIG. 18 also
21 represents as a group 173 all other active transactional and archival nodes in the database
22 system.

23 **[0123]** Once the connection is established in step 171, the joining transactional node 170
24 uses step 174 to send a Connect message to the selected node 172. The selected node 172
25 responds to the Connect Message at step 175 by updating its Master Catalog atom with its
26 connection manager, assigning a local node ID to the joining transactional node 170 and adding
27 that node to an array of local nodes in its node object, an example is shown in FIG. 19.

28 **[0124]** FIG. 19 depicts a node object such as described in connection with step 175 of
29 FIG. 18. It contains a pointer 400A to the socket for the node and a pointer 400B to the Master
30 Catalog atom at the node and a pointer 400C to the connection manager in the Master Catalog
31 atom. Pointer 400D identifies the thread used to listen for incoming messages and pointer 400E

- 24 -

1 identifies a socket buffer for receiving messages. The message listener waits for a message,
2 determines the message type and then processes the message to completion.

3 **[0125]** The node object 400, like all the atoms, includes a global node ID 400F and a
4 local node ID 400G of the node to which this node is listening. Element 400H is a pointer to a
5 queue of messages that are waiting to be sent from the node. Elements 400I and 400J contain the
6 identifications for the local port and the remote ports. Element 400K contains the version
7 number for the software operating at the remote node. A node type element 400L indicates
8 whether the remote node is a transactional node, an archive node undergoing a synchronization
9 process or an on-line archival node.

10 **[0126]** Element 400M contains the name for the local node; element 400N, the name for
11 the remote node. Element 400P is a pointer to the current message being processed. Elements
12 400Q and 400R identify the time of the last ping operation and the ping time. As previously
13 indicated, each node generates a sequential commit sequence number in response to each commit
14 operation for a transaction initiated at that node. Element 400S contains this number. Element
15 400T indicates whether this node object is the node object for this node.

16 **[0127]** Referring back to FIG. 18, at step 176 the selected node 172 sends a Welcome
17 message to the joining transactional node 170 that contains the global node ID. Then, the
18 selected node broadcasts a New Node message in step 177 to the group 173 of all other
19 transactional and archival nodes.

20 **[0128]** Each node in the group 173 responds to the New Node message by recording the
21 global ID and assigning a local ID of the joining transactional node and updating a local list of
22 all connected nodes in their respective Master Catalog atoms at step 180. Each then uses step
23 181 to send a Welcome message to the joining transactional node 170. Upon completion of this
24 process, the joining transactional node 170 has a full list of all the active nodes including the
25 selected node 172 and all the nodes in the group 173.

26 **[0129]** When the joining transactional node 170 receives the Welcome message from the
27 selected node at step 182, it sends an Object Request message (step 183) to the selected node 172
28 requesting a copy of the Master Catalog atom. After the selected node 172 has updated the
29 various items of information in its Master Catalog atom, the selected node implements step 184
30 to serialize its Master Catalog atom in an Object message that is sent to the joining transactional

node 170 and broadcasts an Object Available message to all the other nodes in the system. Thus, the Master Catalog atoms in each node are updated and synchronized.

[0130] Without waiting for the receipt of the Object message, the joining transactional node can also begin the process of step 185 for retrieving a copy of the Database atom and the Transaction Manager atom from the selected node 172.

[0131] Each other node that receives the Object Available message responds in step 186 by sending an Object Acknowledged message back to the selected node.

[0132] The database request engine in the joining node initiates this sequence. Thus, when the method of FIG. 18 has been completed, the joining transactional node 170 is connected to the database system and, as shown in FIGS. 4A and 4B, includes a copy of the Master Catalog atom 70, Transaction Manager atom 71 and the Database atom 72. The joining transactional node 170 thereafter will either create or obtain copies of other atoms on an as-needed basis.

[0133] During some operations, a transactional node might create of a new table, necessitating the creation of a new Table atom. FIG. 20 discloses that process 190 where node A is a requesting node, atom X is the atom that requests and will manage a new atom Y. For creating a new Table atom Y, atom X would be a Schema atom and the local catalog Z would be the Master Catalog atom. Step 191 represents preparatory functions for the request to obtain an instance of atom Y. Each time an atom is changed, it is assigned a change number; the initial value typically is "0".

[0134] At step 192, the local catalog Z creates that instance of atom Y with no content and designates the local node as the chairman for atom Y. Then a process 193 allows the local catalog Z to assign an object ID to new atom Y. The specifics of such a process are shown and described with respect to FIG. 21.

[0135] Next, the local catalog Z sets an exclusive lock to allow changes to it to occur without any influences from outside automation. While the lock is in place, the local catalog Z sets the status of new atom Y to a "dirty" status and to a "non-open" status. The "dirty" status indicates that the new atom Y has not been replicated to an archive. The "non-open" status indicates that the new atom Y is not yet available to other nodes. At step 195 the local catalog Z updates itself and then releases the exclusive lock. At step 196 catalog Z broadcasts an Object Available message identifying atom Y to all other instances of catalog Z in the transactional and archival nodes.

1 **[0136]** At step 197 atom X, as the managing atom for new atom Y, populates the instance
2 of atom Y and sets the status for atom Y to “open” indicating that the new atom can be
3 replicated. Some time thereafter an archival node will save a copy of the new atom to persistent
4 storage. That is, the archival node will respond to the receipt of the Object Available message by
5 requesting a copy of atom Y thereby to provide the redundancy. When this is complete, the
6 receipt of a Object Acknowledged message from the archival node will cause the the “dirty”
7 status to be changed and this change will then be reflected to all other nodes that have a copy of
8 atom Y.

9 **[0137]** The assignment process 193 in FIG. 20 and like assignment processes are used in
10 various stages of the operation of this invention for assigning object IDs. FIG. 21 depicts this
11 assignment process in greater detail. Specifically, when a catalog atom wants to assign an object
12 ID it determines at step 200 if it has a local ID available to it. If it does, control passes to step
13 201 that assigns the available ID in response to the request. If a local ID is not available, control
14 passes to step 202 to determine whether the catalog atom at this node has a “chairman” status. If
15 it is, the chairman has the authority to identify available IDs directly and control passes to step
16 201.

17 **[0138]** As will be recalled, when at least one other node exists for the particular atom,
18 each atom contains a list of nodes that include copies of the atom being requested. If this were a
19 first request for the atom after that atom had been created in accordance with the process of FIG.
20 20, the corresponding catalog in the first node on the list is the chairman. Step 203 represents
21 the process of selecting a node and identifying the chairman. For the selection process, it is
22 preferred to establish communications with a transactional node first. Step 204 represents the
23 transmission of an ID Request message to the chairman.

24 **[0139]** When the chairman receives the ID request message at step 205, it obtains a block
25 of available ID numbers (step 206). The chairman identifies the transactional node making the
26 request as the site of this block of available ID numbers (step 207). Then the chairman sends an
27 ID delegation message at step 210 to the requesting node. When the requesting node receives
28 the ID Delegation message from the chairman at step 211, it stores the block of ID numbers at
29 step 212 and then selects the first available ID for assignment in step 201. Referring back to
30 FIG. 20, in that situation the process 193 transfers from step 200 to step 202 directly to step 201
31 because the chairman is designated in step 192 at FIG. 20.

1 **[0140]** Whenever the Database engine 41 in FIGS. 4A and 4B makes a request for an
2 atom, a response 220 in FIG. 22 is processed. For example, assume that the database request
3 engine 41 in a requesting node 221 asks for a copy of atom Y (e.g., a Table atom) from a
4 selected node 222 in a database with other nodes 223. When the database request engine 41
5 makes that demand, step 224 determines whether atom Y is present in the requesting node 221.
6 If it is, step 225 terminates the process because the requested atom is present. If it is not, control
7 transfers to step 226 whereby a local catalog Z creates an empty instance of atom Y and declares
8 atom Y to be "unpopulated." If the Database engine 41 were requesting a Table atom 74, the
9 Master Catalog atom 70 would perform this step. Step 230 then uses the previously described
10 selection process to identify a selected node 222. The preference is to select any of the
11 transactional nodes before selecting the most responsive archival node as the selected node 222.

12 **[0141]** In step 231, the requesting node 221 sends an Object Request message for atom Y
13 to the selected node 222. In response, the selected node 222 uses step 232 to send an Object
14 Message with the requested atom Y in its serialized form with the node and sequence numbers.

15 **[0142]** At the same time, the selected node 222 broadcasts an Object Available message
16 to all other nodes 223. It also creates a relay list for all other nodes with a copy of atom Y. At
17 this point in the process, the other nodes do not communicate directly with atom Y in the
18 requesting node because they are not aware that they should be sending replication messages for
19 that atom to requesting node. Therefore, when any of the other nodes replicates its atom Y, the
20 selected node 222 will relay the message to the requesting node 221.

21 **[0143]** When the Requesting node 221 receives the Object message from the selected
22 node 222, it conducts an accessibility analysis of the message at step 233. If the message
23 contains a current atom, atom Y processes the message at step 236 and sends an Object
24 Acknowledged message to the selected node at step 237.

25 **[0144]** All other nodes 223 use step 240 to respond to the Object Available messages by
26 sending an Object Acknowledged message to the selected node 222. The selected node 222 uses
27 step 241 to monitor the Object Acknowledged messages. Specifically, it removes each of the
28 other nodes from its relay list in response to each Object Acknowledged message and ceases
29 relays to that node. When all the other nodes 223 have been removed from the list, the selected
30 node stops all relaying and broadcasts an Object Complete message.

1 **[0145]** It is possible for another node to send a replication message that reaches the
2 requesting node 221 during the time between steps 226 and 236 in FIG. 22. This can interrupt
3 message processing. Consequently, when a replication message for atom Y in the above
4 example is received at the requesting node while atom Y is unpopulated, it is placed in a pending
5 message list that is part of each atom, although not shown in any of FIGS. 5 through 13.

6 **[0146]** As previously indicated, this invention is particularly adapted to a database that
7 interfaces with transactional processing techniques. As such, an appropriate approach for
8 “committing” messages should be involved. FIG. 23 depicts one approach that assures the
9 consistency of data in such an environment. Specifically, FIG. 23 depicts four nodes, namely:
10 transactional node A 250, transactional node B 251, transactional node C 252 and archival node
11 253. It is assumed that transactional node B 251 sends a Precommit message with a transaction
12 ID supplied by the Transaction Manager atom at that node at step 254. The message is routed to
13 the archival node at step 255. When all the conditions for committing that transaction have been
14 completed, the archival node 255 issues a commit transaction message at step 256 and, at step
15 257, broadcasts that commit message. Each of the transactional nodes updates its corresponding
16 transaction number in response at step 258.

17 **[0147]** As previously described with respect to the process for requesting a copy of an
18 atom in FIG. 21, the requesting node performs the accessibility analysis upon any received
19 message at step 233. This test provides assurance that any transactional node always operates
20 with valid information and involves an analysis of transaction ID numbers and commit sequence
21 numbers. An understanding of this analysis and other features of this invention will be facilitated
22 by a further understanding of transaction IDs and commit sequence numbers, by an analysis of
23 the relative order of transactions and by an understanding of "atom skew."

24 **[0148]** With respect to transaction IDs and commit sequence numbers and as previously
25 stated, each transaction identifier is unique across the entire database system of FIG. 1. The
26 purpose of a transaction ID is to be a unique, permanent, system-wide mark indicating which
27 transaction created a specific record version. Each transaction ID is assigned by a local copy of
28 a Transaction Manager atom 71 and one such Transaction Manager atom will be the chairman
29 that assigns blocks of identifier numbers to the Transaction Manager atoms on each transactional
30 node in the database system. A Transaction Manager atom at a given transactional node assigns
31 an unused number in an assigned block in order to each transaction that node starts. As a result,

1 the transaction IDs of newer transactions started at a given node can be higher than the
2 transaction IDs assigned to older transactions started on that same node. However, on a system-
3 wide basis the transaction IDs do not imply anything about the relative start times of different
4 transactions.

5 **[0149]** Consequently, some method must be provided to assure that a transaction can
6 read a record version only if the transaction that created the record version had committed before
7 the reading transaction started. In this invention such a method for determining which record
8 version to read is the transaction commit time for the transaction that created the record version,
9 not the start time. A transaction ID does not contain its commit time, so the Transaction
10 Manager atom on each node assigns a commit sequence number to each transaction based upon
11 the actual commit operation. Each transactional node generates its commit sequence numbers in
12 an increasing sequence. If a transaction from a given node has a commit sequence number of
13 467, all transactions from that node with lower commit sequence numbers are certain to have
14 been committed. All nodes generate the same commit sequence numbers, so interpreting the
15 numbers requires both the number and the identification of the corresponding node.

16 **[0150]** When a transaction commits, its Transaction Manager atom sends a Commit
17 message to all nodes that include the corresponding transaction ID and commit sequence
18 number. When an atom serializes itself, it includes the highest commit sequence number that it
19 has seen from each node that holds a copy of that atom. For example, assume nodes A, B and C
20 contain copies of an atom Z and that node A generates a Object message for atom Z. The
21 serialized form of that message will include the highest commit sequence number that node A
22 has seen from itself and the highest commit sequence numbers it has seen from nodes B and C.
23 The serialized message from node A describes the state of transactions on all nodes that share a
24 copy of atom Y as node A sees it. It is possible that node B or node C may have actually issued
25 higher commit sequence numbers for transactions that node A has not yet received or processed.

26
27 **[0151]** The Transaction Manager on each node maintains a transaction object for each
28 node in the database system. Each transaction object reflects the state of transactions on all
29 nodes and assigns each object two more numbers that are local and continuously increasing.
30 When the Transaction Manager atom starts a transaction or receives a transaction transition
31 message that indicates that another node has started a transaction, the Transaction Manager atom

1 creates a local transaction object with the new transaction ID and assigns it a start number.
2 When a local transaction commits or when Transaction Manager atom receives a transaction
3 transition message that indicates that a transaction on another node has been committed, a
4 transaction end number is assigned to the transaction object along with the commit sequence
5 number. The start and end numbers come from the same sequence. As an example, if
6 Transaction 123 has a start number that is higher than the end number of Transaction 453,
7 Transaction 123 can read record versions created by Transaction 453 from any node that
8 executed Transaction 453.

9 **[0152]** Transaction start and end numbers are local; that is, they reflect the state of
10 transactions as seen on the local node. Different nodes assign different values and see different
11 ordering of transactions on other nodes. As will be apparent, delays can exist in receiving and in
12 processing Transaction State messages. Each transaction runs only on its local transactional
13 node. This database system prevents one transaction node from "seeing" a transaction as
14 committed before the transaction commits on its local transactional node. For example, if node A
15 starts Transaction 123 after receiving and processing the commit message from transactional
16 node B for Transaction 453, Transaction 123 can read the changes from Transaction 453, even if
17 a transactional node C that received the transaction commit and start messages in another order
18 and considers the two transactions to be contemporary.

19 **[0153]** As neither of the transactions runs on transactional node C, the information is not
20 relevant to transactional node C. However, the difference between the view of the database
21 system by each individual transactional node can cause problems in certain circumstances.

22 **[0154]** With respect to "atom skew", in accordance with one aspect of this invention,
23 each node must process messages from another node in the order that the other node sent the
24 messages. However, all nodes operate independently. Thus, at any given time, some nodes will
25 have received and processed messages that other nodes have not yet processed. Some copies of
26 atoms will include changes that other copies do not. Those differences do not matter to any node
27 that has a copy because each copy of the atom is in a consistent state for that node. As
28 previously described, whenever a transactional node changes an atom, the node replicates the
29 changed atom to all other nodes that contain a copy of that atom. If transactional nodes B and C
30 each contain a a copy of that atom and transactional node B changes that atom, transactional
31 node B sends replications messages to transactional node C. While transactional node B is

- 31 -

1 processing a transaction, it sends a Transaction State message to transactional node C and
2 replicates any atoms that transactional node B changes. When transactional node B commits the
3 transaction by the process shown in FIG. 23, transactional node C should have received the
4 Transaction State message indicating that the transaction has been committed. Transactional
5 node C will process all the messages related to the transaction before it receives the Commit
6 message due to the fact that all messages will be processed in the order in which they were sent.
7 Copies of atoms on different transactional nodes can differ, but all changes made by a
8 transaction on any transactional node will be on every other node before the other node sees the
9 transaction as committed.

10 **[0155]** When viewing transactions from the system level, different transactional nodes
11 are not acting in synchronism. The times for a single message being processed by different
12 nodes can vary due to differences in the efficiency of the communications path and the number
13 of messages that each transactional node needs to process. The differences in processing times
14 at different nodes must be taken into account. Consider three transactional nodes: node A, node
15 B and node C. Assume that node B executes a transaction 768 that changes atom X and that node
16 C has a copy of atom X. Assume that atom X at node B sends change messages to node C and
17 that there is some time lapse before node C processes those changes. Assume also that node B
18 has assigned a commit sequence number of 47 to Transaction 768. The Transaction Manager
19 atom on node B sends a Transaction State message to all nodes in the database system. Also
20 assume that node A requests a copy of atom X from node C after receiving the commit message
21 from node B, but before Node C completes processing the transaction. Node A receives and
22 processes the Transaction State message from node B. From the perspective of node A,
23 transaction 768 is committed and the highest commit sequence number for node B is 47.
24 However, with this timing, the atom returned from node C does not reflect the changes made by
25 the transaction at node B.

26 **[0156]** In accordance with this invention, when atom X on node C serializes itself for
27 transmission to node A, it includes the highest commit sequence number from each node that has
28 a copy of atom X. In this case, assume that the message would include a commit sequence
29 number 97 for node C and 46 for node B.

30 **[0157]** Referring to FIG. 22, node A as a requesting node creates an empty instance of
31 atom X before asking node C for its contents. The accessibility analysis in FIG. 22, obtains the

1 current highest commit sequence number for each node as seen by node A. It also expands the
2 serialized message with Atom X to obtain the current highest commit sequence numbers for each
3 node as seen by node B. In this comparison, the accessibility analysis will determine that the
4 serialized commit sequence number in the message is 46, while its local highest commit
5 sequence number for B is 47. Thus, Node A cannot use the serialized version of atom X in its
6 current state and must wait for atom X on node C to send it an Object Complete message. If the
7 commit sequence number is the same or higher, node A can continue to process the atom.

8 **[0158]** As previously stated, when node C sent the serialized copy of atom X to node A,
9 it sent an Object Available message to other copies of atom X, in this case, to the copy on node
10 B. Atom X on node B sends an Object Acknowledged message to atom X on node C and adds
11 the copy on node A to its list of copies of itself. Thereafter, atom X on node B will send change
12 messages to its copies on both nodes A and C. Meanwhile, atom X on node C processes the
13 messages from node B, relaying them to node A. Atom X on node A processes the messages,
14 bringing itself more and more up to date. When atom X on node C processes the Object
15 Acknowledged message from atom X on node B and all other copies of atom X on other nodes,
16 it sends an Object Complete message to atom X on node A.

17 **[0159]** In accordance with this invention each transactional node should operate without
18 having to transfer any atoms to disk or other persistent storage. This requires removal of
19 inactive atoms from each transactional node in an expeditious manner. To achieve this goal, a
20 helper class in each node provides a garbage collection function called by a cycle manager
21 periodically to initiate a new garbage collection cycle independently of other operations. More
22 specifically, the cycle manager performs an aging cycle whereby any atom referenced in a
23 previous cycle is moved to the front on a least recently used (LRU) list in the Master Catalog
24 atom for the node. The cycle manager obtains an exclusive lock to assure that no other threads
25 are active with respect to a previous cycle.

26 **[0160]** The cycle manager then starts the record garbage collection thread for a table
27 atom that is marked for record garbage collection. If the current amount of memory at the node
28 exceeds a specified amount defined by a configuration object identified by the pointer 70U at
29 FIG. 5, the garbage collection thread process loops through the Master Catalog LRU from the
30 least recently used to the most recently used atoms. It classifies an atom as a candidate for a
31 garbage collection candidate if that atom: (1) is not active, (2) has not been referenced in the

- 33 -

1 current aging cycle, (3) is not “object incomplete” atom, (4) has been changed locally and not
2 yet written to an archival node and (5) is a catalog that has no memory resident atoms. An atom
3 in an archival node additionally must not have been serialized to disk.

4 **[0161]** If the garbage collection process determines that all the foregoing conditions have
5 been met, there are two options. If the node is an archival node and there are other instances of
6 the atom at other nodes, the content of the atom is “purged”. Otherwise the atom is requested to
7 drop itself which it does by first checking with its creating catalog atom for a final test. If that
8 test is passed, the catalog atom clears its pointer to the candidate atom which broadcasts an
9 Object Unavailable message to other nodes that contain instances of the candidate atom and then
10 deletes itself. This loop continues on an atom by atom basis until the working memory is within
11 acceptable limits.

12 **[0162]** Two mechanisms are involved with this process. When a Catalog atom is asked
13 to find an atom for classification, the Catalog atom sets its cycle reference entry to the current
14 cycle number. In addition, Transaction Manager, Database, Schema, Sequence and Table atoms
15 obtain a shared lock on a current cycle so that methods in those atoms can hold atom pointers
16 without having to increment an atom use count to prevent an object or an atom from being
17 garbage collected.

18 **[0163]** FIGS. 19 and 20 discuss “chairmanship.” As indicated in FIG. 20 when an atom
19 is created, the creating node is designated as the atom chairman and establishes an ordered list of
20 nodes for the atom. When a node creates an atom, it is the only entry in the ordered list. The
21 first node in the list is the chairman.

22 **[0164]** When another node later requests a copy of that atom, the selected node places the
23 identification of the requesting node in the ordered list just after itself, whether it is the chairman
24 or not. As a result, if the node designated as a chairman for any atom becomes inactive for any
25 reason, every other copy of that atom has the ordered list of nodes. Going through the ordered
26 list, the next chairman is the first remaining transactional node on the list. If there are no
27 transactional nodes on the list, the first non-synchronizing archival node is designated.

28 **[0165]** It may now be helpful as an aid in understanding the interaction among different
29 nodes and atoms within a node to discuss a simple database query. Assume the database request
30 engine at a transactional node receives a database query to select all records in a table “member”
31 with a postal code of “01944.” It is assumed that the Transaction Manager atom and Table atom

- 34 -

are located at the local node, that all query processing has been completed and that an Index atom exists for a “zip” field in the “member” table.

[0166] Initially the database request engine 41 in FIG. 2 issues a command to the Transaction Manager atom to begin a new transaction whereupon the Transaction Manager atom assigns a transaction ID. Next, the database request engine utilizes the Table atom for the “member” table to determine if an Index atom exists for the “zip” field. If these two atoms are at the local node, they are processed. If they are not, copies of these atoms are obtained from other nodes before processing continues.

[0167] The Table atom for the “member” table utilizes information from the corresponding Table Catalog atom to scan the index. This produces a bitmap that identifies each “record” in the “member” table with the specified zip code. As will be apparent, this bitmap may be limited to a particular field or maybe the result of the logical combination of indexes from multiple fields.

[0168] Next, a loop is established based upon the resulting bitmap. For each iteration of the loop, the database request engine issues a call to the Table atom to process a fetch record method specified with the transaction ID in the record. The Table atom selects an appropriate Record States atom that is related to a Record atom by dividing the record number by a fixed number that corresponds to the maximum number of record IDs a Record States atom manages. Next, the Table atom uses a fetch method in the selected Record States atom using the transaction ID and identified record number. For multi-version records, the Record States atom loops through any record versions to find the correct version and the appropriate pointer to the corresponding Data atom. The Record States atom calls that Data atom with the data atom number that points to the record and permits its retrieval. When this process is completed, the database request engine provides the user with a recordset that lists all the records with the specified postal code.

[0169] In summary, it will be apparent to those of ordinary skill in the art that a database management system constructed in accordance with this invention provides an elastic, scalable demand-distributed data processing system. The system is fault tolerant and has a high degree of availability. It is platform independent and operates to provide an atomic, consistent, isolated and durable database. Moreover, it can operate over the internet without the need for high speed

- 35 -

1 communications paths and is adapted for transactional processing that can be implemented over
2 a wide geographic area.

3 **[0170]** This invention achieves all these objectives by implementing one or more of the
4 following features. This invention fragments the database into distributed objects that replicate
5 on a peer-to-peer basis. Each transactional and archival node determines which atoms are to be
6 memory resident on a local basis. Catalog atoms track locations to local and remote copies of
7 various atoms and identify the catalogs of which they are members. Moreover each node can
8 determine the best of multiple nodes from which to request a copy of the atom enabling
9 geographically dispersed systems.

10 **[0171]** This invention has been disclosed in terms of certain embodiments. It will be
11 apparent that many modifications can be made to the disclosed apparatus without departing from
12 the invention. Therefore, it is the intent of the appended claims to cover all such variations and
13 modifications as come within the true spirit and scope of this invention.

14

Claims

1. A database management system that enables users to interact with a database comprised of data and metadata, said system comprising
- A) a plurality of nodes, each node comprising:
- i) an interface between high level-input and output commands at a user level and input and output commands at the system level that control a sequence of operations for interacting with the data base, wherein, in response to certain system level commands, atom objects generate atoms, each atom containing a specified fragment of data or metadata whereby a set of all instances of atoms collectively define all the metadata and data in the database,
 - ii) a communications control for establishing a communications path with each other node in the system,
 - iii) a method responsive to a system command from the interface for requesting from a selected node copy of an atom that is relevant to the query but is not present in that node,
 - iv) a method responsive to a request for an atom from another node for replicating the requested atom for transfer to the requesting node whereby only atoms required to complete a query need be located in any transaction node at any given time, and
 - v) a method responsive to a change in an atom at that node for replicating that atom for transfer to every other node in the system in which that atom is resident,
- B) means connected to the plurality of nodes for providing communications paths between all the nodes, and
- C) means connected to said communications means for providing persistent storage for information in atom whereby the collection of atoms in persistent storage collectively contain all the data and metadata in the database.
2. A database management system as recited in claim 1 wherein certain of the atom objects produce catalog atoms that keep track of all atoms that are resident at that node and the location of each node that has a copy of each of the resident atoms.

1
2 3. A database management system as recited in claim 2 wherein one of the catalog atoms is
3 a master catalog atom that is replicated in each node connected to the system, wherein another of
4 the atoms is a single transaction manager atom for the database for tracking the progress of
5 transactions initiated by the interface and wherein the master catalog identifies the location of
6 the transaction manager atom.

7 4. A database management system as recited in claim 3 wherein the database at the user
8 level comprises named tables with records and fields and operates in accordance with at least
9 one named schema and others of the atom objects produce:

- 10 i) a table atom for each table that contains a list of fields,
11 ii) a schema atom establishes a correspondence between a schema and the
12 named tables relevant to that schema, and
13 iii) a database atom that establishes a correspondence between the database
14 and the named schema.

15 5. A database management system as recited in claim 4 wherein the database tables at the
16 user level comprises identified records containing fields and data, wherein other atom objects
17 produce data atoms that contain a predetermined number of data records and a record states atom
18 that manages record versions and wherein said record states atom includes an array that
19 identifies the managed data atoms assigned to that record states atom, an identification of each
20 record version and, for each record version, information about that version and the location of
21 the data atom containing that version.

22 6. A database management system as recited in claim 5 wherein each data atom additionally
23 includes the address and length of each record within the data atom and the data records.

24 7. A database management system as recited in claim 4 wherein the database tables at the
25 user level comprises identified blobs, wherein other atom objects produce blob atoms that
26 contain a blob and a blob record states atom that includes a list of all associated blob atoms and
27 the location of the blob in the blob atom.

28 8. A database management system as recited in claim 7 wherein each blob atom includes
29 the address and length of each blob and the blob.

30 9. A database management system as recited in claim 1 wherein each node additionally
31 comprises:

- 38 -

1 i) messaging means for generating and receiving predefined messages in
2 response to system level commands, and

3 ii) means for selecting one of the communications paths for transmitting the
4 message.

5 10. A database management system as recited in claim 9 wherein each node comprises means
6 for asynchronously testing the efficiency of each communications path.

7 11. A database management system as recited in claim 10 wherein said testing means
8 includes means for generating a ping message and receiving a ping acknowledgement message
9 for determining the ping time for each communications path.

10 12. A database management system as recited in claim 9 wherein said messaging means
11 includes means for serializing an atom into a message and de-serializing a message into an atom
12 and wherein each message is sent over the selected communications path asynchronously.

13 13. A database management system as recited in claim 12 wherein messages are sent under a
14 protocol with controls that maintain message sequences.

15 14, A database management system as recited in claim 1 wherein individual nodes may be
16 active and connected to the system and may be disconnected from the system and one of the
17 atoms is a master catalog atom that identifies each active node in the system and that is
18 replicated in each node connected to the system, the system comprising means for enabling a
19 node to join the system including:

20 i) means for establishing a connection between the joining node and a
21 selected one of the active nodes,

22 ii) means in the joining node for asking the selected node for connection,

23 iii) means in the selected node for updating its master catalog atom to indicate
24 the joining node as an active node,

25 iv) means in the selected node for transferring a message to the joining node
26 including the selected node's identification and for broadcasting a
27 message indicating the availability of the joining node to all other nodes in
28 the system and persistent storage,

29 v) means in said joining node for responding to the message by requesting a
30 copy of the master catalog atom from the selected node, and

- 1 vi) means in the selected node for transferring its updated master catalog atom
2 to the joining node.

3 15. A database management system as recited in claim 14 wherein said joining node includes
4 means responsive to the receipt of the updated master catalog atom for requesting copies of
5 additional atoms in response to requests from the interface.

6 16. A database management system as recited in claim 14 wherein the selected node
7 broadcasts the availability of the updated master catalog to all other nodes and to persistent
8 storage and each other node and persistent storage includes means for acknowledging its receipt
9 of the updated master catalog to the selected node.

10 17. A database management system as recited in claim 1 wherein an interface in one of the
11 nodes makes a request that requires the production of a new atom wherein certain of the atom
12 objects produce catalog atoms that keep track of atoms that are resident at that node, said one of
13 said nodes additionally comprising:

- 14 i) means in said catalog atom that will track the new atom responsive to the
15 interface request for causing the catalog atom to create that new atom with
16 no content and to designate the new atom as a chairman for that atom,
17 ii) means in the catalog atom that assigns an object identification to the new
18 atom,
19 iii) means in the catalog atom for establishing a status for the new atom that
20 indicates there is no redundant copy of the new atom and that the new
21 atom is not available to other nodes,
22 iv) means in the catalog atom for broadcasting to all other corresponding
23 catalog atoms in the system the existence of the new atom,
24 v) means for populating the new atom and for making the new atom available
25 on the system.

26 18. A database management system as recited in claim 17 wherein persistent storage includes
27 means responsive to the storage of the new atom for sending a message to node with the new
28 atom updating its status to indicate the existence of a redundant copy.

29 19. A database management system as recited in claim 1 wherein an interface in one of the
30 nodes makes a request for an atom that only exists in other nodes and wherein certain of the

- 40 -

1 atom objects produce catalog atoms that keep track of atoms that are resident at that node and
2 other nodes, said system including:

- 3 i) means in the requesting catalog atom responsive to the interface request
4 that will track the requested atom responsive to the interface request for
5 creating that empty atom,
- 6 ii) means in the requesting catalog atom for establishing selecting the most
7 responsive node and requesting a copy of the atom from the most
8 responsive node,
- 9 iii) means in the selected node for sending a copy of the requested atom in an
10 object message to the requesting node and for broadcasting the availability
11 of the atom to other nodes in which the atom,
- 12 iv) means for assessing the validity of the received object message, for
13 populating the empty atom and for forwarding an object acknowledgment
14 message to the selected node.

15 20. A database management system as recited in claim 19 wherein said selected node
16 generates a relay list of all nodes containing the requested atom to which the selected node will
17 forward messages from other related to that atom to the requesting node.

18 21. A database management system as recited in claim 20 wherein each of the other nodes
19 that receive the broadcast object available message includes means for sending an object
20 acknowledgment message to the selected node, said selected node including means for updating
21 the relay list in response to the receipt of each object acknowledgment message from another
22 node to terminate the relay of any subsequent messages to the requesting node.

23 22. A database management system as recited in claim 1 adapted for transactional processing
24 for assuring consistency of the database wherein the system includes a first node involving an
25 atom to be committed, at least one other node including a copy of that atom and a copy of that
26 atom included in persistent storage, said system including:

- 27 i) means in the first node for generating a precommit message with a
28 transaction identification for transfer to persistent storage,
- 29 ii) means associated with persistent storage responsive to the precommit
30 message for committing the transaction and thereafter broadcasting a
31 commit message with the transaction identification to all the nodes,

1 iii) means in the first and at least one other node responsive to the receipt of
2 the commit message for updating the identified transaction.

3 23. A database management system that enables users to interact with a database comprised
4 of data and metadata, said system comprising

- 5 A) at least one transactional node that provides users with access to the database, and
6 at least one archival node that maintains an archive of the entire database, each
7 transactional node including a database request engine that provides an interface
8 between high level-input and output query commands at a user level and input
9 and output commands at the system level that control a sequence of operations for
10 interacting with the database, wherein, in response to certain system level
11 commands, atom objects generate atoms, each atom containing a specified
12 fragment of data or metadata whereby a set of all instances of atoms collectively
13 define all the metadata and data in the database,
14 B) a database system network interconnecting all the nodes,
15 C) a communications control in each of the nodes for establishing a communications
16 path with each other node in the system,
17 D) a method in each transactional node responsive to a system command from the
18 database request engine for requesting a copy of an atom that is relevant to the
19 query command but is not present in that node,
20 E) a method in each node responsive to a request for an atom from another node for
21 replicating the requested atom for transfer to the requesting node whereby only
22 atoms required to complete a query command need be located in any transactional
23 node at any given time, and
24 F) a method in each transactional node responsive to a change in an atom at that
25 node for replicating that change to every other node in the system that contains a
26 copy of that atom.

27 24. A database management system as recited in claim 23 wherein certain of the atom objects
28 produce catalog atoms that keep track of all atoms that are resident at that node and the location
29 of each node that has a copy of each of the resident atoms.

30
31 25. A database management system as recited in claim 24 wherein one of the catalog atoms

1 is a master catalog atom that is replicated in each node connected to the system, wherein another
2 the atoms is a single transaction manager atom for the database for tracking the progress of
3 transactions initiated by the interface and wherein the master catalog identifies the location of
4 the transaction manager atom.

5
6 26. A database management system as recited in claim 25 wherein the database at the user
7 level comprises named tables with records and fields and operates in accordance with at least
8 one named schema and others of the atom objects produce:

- 9 i) a table atom for each table that contains a list of fields,
- 10 ii) a schema atom establishes a correspondence between a schema and the
11 named tables relevant to that schema, and
- 12 iii) a database atom that establishes a correspondence between the database
13 and the named schema.

14 27. A database management system as recited in claim 26 wherein the database tables at the
15 user level comprises identified records containing fields and data, wherein other atom objects
16 produce data atoms that contain a predetermined number of data records and a record states atom
17 that manages record versions and wherein said record states atom includes an array that
18 identifies the managed data atoms assigned to that record states atom, an identification of each
19 record version and, for each record version, information about that version and the location of
20 the data atom containing that version.

21 28. A database management system as recited in claim 27 wherein each data atom
22 additionally includes the address and length of each record within the data atom and the data
23 records.

24 29. A database management system as recited in claim 26 wherein the database tables at the
25 user level comprises identified blobs, wherein other atom objects produce blob atoms that
26 contain a blob and a blob record states atom that includes a list of all associated blob atoms and
27 the location of the blob in the blob atom.

28 30. A database management system as recited in claim 29 wherein each blob atom includes
29 the address and length of each blob and the blob.

30 31. A database management system as recited in claim 23 wherein each node additionally
31 comprises:

1 i) messaging means for generating and receiving predefined messages in
2 response to system level commands, and

3 ii) means for selecting one of the communications paths for transmitting the
4 message.

5 32. A database management system as recited in claim 31 wherein each node comprises
6 means for asynchronously testing the efficiency each communications path.

7 33. A database management system as recited in claim 32 wherein said testing means
8 includes means for generating a ping message and receiving a ping acknowledgement message
9 for determining the ping time for each communications path.

10 34. A database management system as recited in claim 31 wherein said messaging means
11 includes means for serializing an atom into a message and de-serializing the message into an
12 atom and wherein each message is sent over the selected communications path asynchronously.

13 35. A database management system as recited in claim 34 wherein messages are sent under a
14 protocol with controls that maintain message sequences.

15 36. A database management system as recited in claim 23 wherein individual nodes may be
16 active and connected to the system and may be disconnected from the system and one of the
17 atoms is a master catalog atom that identifies each active node in the system and that is
18 replicated in each node connected to the system, the system comprising means for enabling a
19 node to join the system including:

20 i) means for establishing a connection between the joining transactional
21 node and a selected one of the active nodes,

22 ii) means in the joining transactional node for asking the selected node for
23 connection to the system,

24 iii) means in the selected node for updating its master catalog atom to indicate
25 the joining transactional node as an active node,

26 iv) means in the selected node for transferring a message to the joining
27 transactional node including the selected node's identification and for
28 broadcasting a message indicating the availability of the joining
29 transactional node to all other nodes in the system,

30 v) means in said joining transactional node for responding to the message by
31 requesting a copy of the master catalog atom from the selected node, and

- 44 -

- 1 vi) means in the selected node for transferring its updated master catalog atom
2 to the joining transactional node.

3 37. A database management system as recited in claim 36 wherein said joining transactional
4 node includes means responsive to the receipt of the updated master catalog atom for requesting
5 copies of additional atoms in response to system commands from its database request engine.

6 38. A database management system as recited in claim 36 wherein the selected node
7 broadcasts the availability of the updated master catalog to all other nodes and each other node
8 includes means for acknowledging its receipt of the updated master catalog to the selected node.

9 39. A database management system as recited in claim 23 wherein the database request
10 engine in one of the transactional nodes makes a request that requires the production of a new
11 atom and wherein certain of the atom objects produce catalog atoms that keep track of atoms that
12 are resident at that node, said transactional node additionally comprising:

- 13 i) means in said catalog atom for tracking the new atom responsive to the
14 database request engine request, for causing the catalog atom to create that
15 new atom with no content and for designating the new atom as a chairman
16 for that atom,
17 ii) means in the catalog atom for assigning an object identification to the
18 new atom,
19 iii) means in the catalog atom for establishing a status for the new atom that
20 indicates there is no redundant copy of the new atom and that the new
21 atom is not available to other nodes,
22 iv) means in the catalog atom for broadcasting to all other corresponding
23 catalog atoms in the system the existence of the new atom,
24 v) means for populating the new atom and for making the new atom available
25 to other nodes in the system.

26 40. A database management system as recited in claim 39 additionally including means in
27 each archival node responsive to the storage of the new atom for sending a message to the new
28 atom updating its status to indicate the existence of a redundant copy.

29 41. A database management system as recited in claim 23 wherein a database request engine
30 in one transactional nodes requests for an atom that only exists in other nodes and wherein

1 certain of the atom objects produce catalog atoms that keep track of atoms that are resident at
2 that node and other nodes, said system including:

- 3 i) means responsive to the request in the requesting catalog atom that will
4 track the requested atom for creating that empty atom,
- 5 ii) means in the requesting catalog atom for establishing selecting the most
6 responsive node and requesting a copy of the atom from the most
7 responsive node,
- 8 iii) means in the selected node for sending a copy of the requested atom in an
9 object message to the requesting node and for broadcasting the availability
10 of the atom to other nodes in which the atom resides,
- 11 iv) means for assessing the validity of the received object message, for
12 populating the empty atom and for forwarding an object acknowledgment
13 message to the selected node.

14 42. A database management system as recited in claim 41 wherein said selected node
15 generates a relay list of all nodes containing the requested atom to which the selected node will
16 forward messages from other related to that atom to the requesting node.

17 43. A database management system as recited in claim 42 wherein each of the other nodes
18 that receive the broadcast object available message includes means for sending an object
19 acknowledgment message to the selected node, said selected node including means for updating
20 the relay list in response to the receipt of each object acknowledgment message from another
21 node to terminate the relay of any subsequent messages to the requesting node.

22 44. A database management system as recited in claim 23 adapted for assuring consistency of
23 the database during transactional processing wherein the system includes a first node involving a
24 transaction to be committed and at least one other node including an archival node includes a
25 copy of that atom including , said system including:

- 26 i) means in the first node for generating a precommit message with a
27 transaction identification for transfer to an archival node,
- 28 ii) means associated with archival node responsive to the precommit message
29 for committing the transaction and thereafter broadcasting a commit
30 message with the transaction identification to all the nodes, and

iii) means in the first and at least one other node responsive to the receipt of the commit message for updating the identified transaction.