An information management system includes a computer and a database comprising a B-Tree data structure comprising a plurality of nodes associated with disk blocks and handles stored in the nodes. At least one left-link handle, \( h_{\text{left}} \), stored in each node points to a left sibling of that node. A mechanism for performing a lookup operation with respect to a key, \( k \), traverses the B-Tree and refers to the left-link handle, \( h_{\text{left}} \), of a node to access a left sibling of the node if the key \( k \) is less than or equal to a value \( k_{\text{min}} \) stored in the node. Mechanisms are also provided for performing insert and delete operations, and the lookup, insert, and delete operations detect if the key range of an index node, \( A \), does not include the key \( k \) that the operation is trying to locate, and follow a handle \( h_{\text{left}} \) to the left sibling when \( k < A.k_{\text{min}} \).

**References Cited**


(Continued)
OTHER PUBLICATIONS


Lomet, D.B., “Multi-Table Search for B-Tree Files,” SIGMOD Conference, 1979, 35-42.


Pollari-Malmi, K. et al., “Concurrency Control in B-Trees with Batch Updates”, IEEE Transactions on Knowledge and Data Engineering, Dec. 1996, 8(6), 975-984.
* cited by examiner
Functions Supported:
1) insert(k, v);
2) lookup(k);
3) delete(k).

FIGURE 2 (PRIOR ART)

FIGURE 3A (PRIOR ART)
LOOKUP PROCEDURE USING HANDLES TO LEFT SIBLINGS

procedure Lookup(h_p : Handle, k : Key)
var h: Handle, A: Node;
P := Read(h_p)
h := last handle in P
repeat
A := read(h)
if k > A.k_max then
  h := A.h_right
else if k <= A.k_min then
  h := A.h_left
else
  choose i such that A.k_{i-1} < k <= A.k_i (let i = 0 if Num(A) = 0)
  h := A.h_i
endif
until A is a leaf and A.k_min < k <= A.k_max
if k is contained in A then
  return corresponding value
else
  return null
endif
endproc

FIGURE 3B

FIGURE 3E
FIGURE 3C
FIGURE 3D
LOOKUP PROCEDURE USING VERSION NUMBERS TO HANDLES AND DISK BLOCKS

procedure Lookup(h_p : Handle, k : key)
  var h: Handle, A: Node;
label start:
  P := Read(h_p)
  h := last handle in P
  repeat
    A := read(h)
    if version(A) = version(h) then goto start
    if k > A.k_max then
      h := A.h_right
    else if k <= A.k_min then
      goto start
    else
      choose i such that A.k_{i-1} < k <= A.k_i (let i=0 if Num(A)=0)
      h := A.h_i
    endif
    until A is a leaf and A.k_min < k <= A.k_max
    if k is contained in A then
      return corresponding value
    else
      return null
    endif
  endproc

FIGURE 3F
GetNextHandle(X, k) entry point

1. If k <= X.k_{min}, return (X.h_{left}, true)
2. If k > X.k_{max}, return (X.h_{right}, true)
3. If Num(X) = 0, return (X.h_0, false)
4. If X.k_{n-1} < k, return (X.h_n, false)
5. Find smallest i such that k <= X.k_i
   a. If IsLeaf(X), return (X.h_{i+1}, false)
   b. Otherwise, return (X.h_i, false)

FIGURE 4
FIGURE 5

InsertPair(X, k, h) entry point

502

504

Yes

No

508

X.k_0 < k ?

Yes

No

512

Find largest i such that X.k_i < k

506

X := (X.h_left, X.k_min, X.h_0, k, h, X.k_max, X.h_right)

508

X := (X.h_left, X.k_min, X.h_0, k, h, X.k_0, ... , X.h_n, X.k_max, X.h_right)

510

516

return

514

X := (X.h_left, X.k_min, X.h_0, ... , X.k_{i-1}, X.h_i, k, h, X.k_i, ... , X.h_n, X.k_max, X.h_right)
CreateTree() entry point

\[ X := (\text{null, -inf, null, +inf, null}) \]

\[ h_x := \text{Alloc(space for } X) \]

\[ \text{Write}(X, h_x) \]

\[ P := (h_x) \]

\[ h_p := \text{Alloc(space for } P) \]

\[ \text{Write}(P, h_p) \]

return \( h_p \)

FIGURE 6
Lookups:

1. Lookup($h_p$, $k$) entry point
2. ($v$, $P$) := Read($h_p$)
3. $h :=$ last handle in $P$
4. ($v$, $X$) := Read($h$)
5. If $h.v = v$ THEN $\text{(h, isLink)} := \text{GetNextHandle}(X, k)$
6. If $h = \text{null}$ THEN return null
7. If $\text{IsLeaf}(X) \text{ and not isLink}$ THEN
   8. ($v$, $D$) := Read($h$)
   9. If $h.v = v$ THEN
      10. If $D.k = k$ THEN return $D.d$
      11. ELSE return null
   12. ELSE return null

FIGURE 7
Insert \((h_p, k, d)\) entry point

\[ D := (k, d) \]

\[ h_d := \text{Alloc}(\text{space for } D) \]

Write\((D, h_d)\)

\[ L_T := 0 \]

\[ S := \text{new empty stack} \]

\[ (v, P) := \text{Read}(h_p) \]

\[ h_x := \text{last handle in } P \]

\[ L_N := \text{number of handles in } P \]

\[ L_C := L_N - 1 \]

\[ (v, X) := \text{Read}(h_x) \]

\[ h_x.v = v? \]

\[ (h', \text{isLink}) := \text{GetNextHandle}(X, k) \]

\[ L_C > L_T? \]

No \(\rightarrow\) B

Yes

\[ L_C := L_C - 1 \]

Push \(h_x\) onto \(S\)

\[ h_x := h' \]

FIGURE 9A
\[ (v, X) := \text{Read}(h_x) \]

\[
\begin{align*}
&\text{Lock}(h_x) \\
& (v, X) := \text{Read}(h_x) \\
& h_x.v = v? \\
& \text{No} \quad \text{A} \\
& \text{Yes} \\
& \text{IsLeaf}(X)? \\
& \text{Yes} \\
& \text{Num}(X) = 0? \\
& \text{No} \\
& \text{X.k_{n-1} < k?} \\
& \text{No} \\
& \text{Find smallest i such that k <= X.k_i} \\
& \text{Yes} \\
& (v, D) := \text{Read}(X.h_{i+1}) \\
& \text{Yes} \\
& D.k = k? \\
& \text{No} \\
& \text{Unlock}(h_x) \\
& \text{return}
\end{align*}
\]
Z := (null, -inf, h_{x}, X, k_{max}, h_{y}, +inf, null)

h_{x} := Alloc(space for Z)

Write(Z, h_{x})

Append h_{x} to end of P

Write (P, h_{p})

Unlock(h_{x})

return

E

h_{x} = last handle in P?

Yes

No

Unlock(h_{x})

h_{p} := h_{y}

k := X, k_{max}

S empty?

Yes

No

h_{x} := pop handle from S

h_{x} := P[L_{n}]

L_{N} := L_{N} + 1

L_{T} := L_{T} + 1

FIGURE 9E
Compress(hₚ) entry point

(v, P) := Read(hₚ)

i := 1

i < number of handles in P?

CompressLevel(P[i])

i := i + 1

h := last handle in P

Lock(h)

(v, X) := Read(h)

Num(X) = 0 and X.hₙᵉᵉᵣ = null and not IsLeaf(X)?

(v, P) := Read(hₚ)

h = last handle in P?

Remove last handle from P

Write(P, hₚ)

Dealloc(h)

Unlock(h)

Pause for some time

FIGURE 10
CompressLevel(h_x) entry point

a := 0

h_A := null

h_x = null? Yes → return

No → Lock(h_x)

(v, X) := Read(h_x)

h_A = null? Yes → a := -1

No → a := -1

i := 0

i := i + 1

i ≤ Num(X)? Yes → X.h_i = h_A?

No → a := i

Yes → a := i

FIGURE 11A
FIGURE 11B
MergeNodes(X, A, B, h_x, h_A, h_B) entry point

IsLeaf(A)?

Yes

A := (A.h_left, A.k_min, null, A.h_0, B.k_0, ..., B.h_n, B.k_max, B.h_right)

No

A := (A.h_left, A.k_min, A.h_0, ..., A.h_n, A.k_max, B.h_0, ..., B.h_n, B.k_max, B.h_right)

B := (B.h_left, -inf, null, -inf, B.h_left)

X := (X.h_left, X.k_min, X.h_0, ..., X.k_{n-2}, X.h_{n-1}, X.k_{n}, X.h_{n+1}, ..., X.h_{n}, X.k_max, X.h_right)

Write(A, h_A)

Unlock(h_A)

Write(X, h_x)

Unlock(h_x)

Write(B, h_B)

Unlock(h_B)

Dealloc(h_B)

return

FIGURE 12
MoveToLeft(X, A, B, hA, hB) entry point

m := 1/2 Num(B) - 1/2 Num(A)

IsLeaf(A)?

A := (A.hleft, A.kmin, null, A.k1, ..., A.hn, B.k1, ..., B.hm, B.km, A.hright)

B := (B.hleft, B.km, null, B.km1, ..., B.hn, B.kmax, B.hright)

X.kn := Akmax

Write(A, hA)

Unlock(hA)

Write(X, hA)

Unlock(hA)

Write(B, hB)

Unlock(hB)

return

FIGURE 13
\[ m := \frac{1}{2} \text{Num}(A) + \frac{1}{2} \text{Num}(B) \]

\[ B := (B_{\text{left}}, A_{k_{\text{min}}}, \text{null}, A_{k_{\text{min}}}, \ldots, A_{k_{n}}, B_{k_{n}}, \ldots, B_{k_{\text{max}}}, B_{k_{right}}) \]

\[ A := (A_{\text{left}}, A_{k_{\text{min}}}, A_{k_{\text{min}}}, \ldots, A_{k_{n}}, A_{k_{n}}, A_{k_{\text{right}}}) \]

\[ X_{k_{a}} := A_{k_{\text{max}}} \]

\[ \text{Write}(B, h_{B}) \]

\[ \text{Unlock}(h_{B}) \]

\[ \text{Write}(X, h_{X}) \]

\[ \text{Unlock}(h_{X}) \]

\[ \text{Write}(A, h_{A}) \]

\[ \text{Unlock}(h_{A}) \]

\[ \text{return} \]
ALGORITHM FOR TREE TRAVERSALS USING LEFT LINKS

BACKGROUND OF THE INVENTION

B-Trees are a core technology in relational and non-relational databases, as well as to file systems and other systems in which a data structure including a set of linked nodes is employed as a way to index and access large amounts of data. A database management system is one example of an information management/retrieval system of the kind for which the present invention is suited. Nevertheless, because the present invention is well suited for use in connection with a database, although by no means limited thereto, the background of the invention and the exemplary embodiments will be discussed with reference to a database.

Concurrent B-Trees with minimal locking are instrumental for building distributed databases, where a single relation may span multiple machines. Distributing relations over multiple machines makes it possible to build scalable databases, where the size of the database can be increased simply by adding more hardware.

The present invention is especially concerned with ways to increase the efficiency of concurrent B-Tree algorithms. As discussed in greater detail below, the subject matter disclosed herein is directed to (a) maintaining extra data that prevents B-Tree operations from getting “lost” in the tree, which may happen in state-of-the-art algorithms due to aggressive minimization of locking operations, and (b) garbage-collecting deleted nodes without the need for any extra synchronization (the state-of-the-art algorithm uses a time-stamp-based approach to garbage collection, and requires extra synchronization to keep track of the start times of all outstanding B-Tree operations).

Further background information about B-Trees may be found in the following documents:


SUMMARY OF THE INVENTION

An information management system in accordance with a first aspect of the present invention includes a computer and a database, wherein the database comprises one or more B-Tree data structure(s) each comprising a plurality of nodes associated with disk blocks (or more generally, “storage blocks,” in the event the invention is employed in connection with a storage medium other than disk) and handles stored in the nodes (and optionally version numbers associated with each of the handles and nodes). At least one left-link handle, h_{left}, stored in each node points to a left sibling of that node. A mechanism for performing a lookup operation with respect to a key, k, traverses the B-Tree and refers to the left-link handle, h_{left}, of a node to access a left sibling of the node if the key k is less than or equal to a value k_{min} stored in the node. Mechanisms are also provided for performing insert and delete operations, and the lookup, insert, and delete operations detect if the key range of an index node, A, does not include the key k that the operation is trying to locate, and follow a handle A_{h_{left}} to the left sibling when k < A_{k_{min}}.

Note that a “node” may span multiple disk or storage blocks, and that links and version numbers may be logically associated with nodes as opposed to individual disk blocks. These and other features of the present invention are described below.
A. Exemplary Computing Environment

FIG. 1 illustrates an example of a suitable computing system environment 100 in which the invention may be implemented. The computing system environment 100 is only one example of a suitable computing environment and is not intended to suggest any limitation as to the scope of use or functionality of the invention. Neither should the computing environment 100 be interpreted as having any dependency or requirement relating to any one or combination of components illustrated in the exemplary operating environment 100.

The invention is operational with numerous other general purpose or special purpose computing system environments or configurations. Examples of well known computing systems, environments, and/or configurations that may be suitable for use with the invention include, but are not limited to, personal computers, server computers, hand-held or laptop devices, multiprocessor systems, microprocessor-based systems, disk controllers, set top boxes, programmable consumer electronics, network PCs, minicomputers, mainframe computers, distributed computing environments that include any of the above systems or devices, and the like.

The invention may be described in the general context of computer-executable instructions, such as program modules, being executed by a computer. Generally, program modules include routines, programs, objects, components, data structures, etc. that perform particular tasks or implement particular abstract data types. The invention may also be practiced in distributed computing environments where tasks are performed by remote processing devices that are linked through a communications network or other data transmission medium. In a distributed computing environment, program modules and other data may be located in both local and remote computer storage media including memory storage devices.

With reference to FIG. 1, an exemplary system for implementing the invention includes a general purpose computing device in the form of a computer 110. Components of computer 110 may include, but are not limited to, a processing unit 120, a system memory 130, and a system bus 121 that couples various system components including the system memory to the processing unit 120. The system bus 121 may be any of several types of bus structures including a memory bus or memory controller, a peripheral bus, and a local bus using any of a variety of bus architectures. By way of example, and not limitation, such architectures include Industry Standard Architecture (ISA) bus, Micro Channel Architecture (MCA) bus, Enhanced ISA (EISA) bus, Video Electronics Standards Association (VESA) local bus, and Peripheral Component Interconnect (PCI) bus (also known as Mezzanine bus).

Computer 110 typically includes a variety of computer readable media. Computer readable media can be any available media that can be accessed by computer 110 and includes both volatile and non-volatile, removable and non-removable media. By way of example, and not limitation, computer readable media may comprise computer storage media and communication media. Computer storage media includes both volatile and non-volatile, removable and non-removable media implemented in any method or technology for storage of information such as computer readable instructions, data structures, program modules or other data. Computer storage media includes, but is not limited to, RAM, ROM, EEPROM, flash memory or other memory technology, CD-ROM, digital versatile disks (DVD) or other optical disk storage, magnetic cassettes, magnetic tape, magnetic disk storage or other magnetic storage devices, or any other medium which can be used to store the desired information and which can be accessed by computer 110. Communication media typically embodies computer readable instructions, data structures, program modules or other data in a modulated data signal such as a carrier wave or other transport mechanism and includes any information delivery media. The term “modulated data sig-
nal” means a signal that has one or more of its characteristics set or changed in such a manner as to encode information in the signal. By way of example, and not limitation, communication media includes wired media such as a wired network or direct-wired connection, and wireless media such as acoustic, RF, infrared and other wireless media. Combinations of any of the above should also be included within the scope of computer readable media.

The system memory 130 includes computer storage media in the form of volatile and/or non-volatile memory such as ROM 131 and RAM 132. A basic input/output system 133 (BIOS), containing the basic routines that help to transfer information between elements within computer 110, such as during start-up, is typically stored in ROM 131. RAM 132 typically contains data and/or program modules that are immediately accessible to and/or presently being operated on by processing unit 120. By way of example, and not limitation, FIG. 1 illustrates operating system 134, application programs 135, other program modules 136, and program data 137.

The computer 110 may also include other removable/non-removable, volatile/non-volatile computer storage media. By way of example only, FIG. 1 illustrates a hard disk drive 140 that reads from or writes to non-removable, non-volatile magnetic media, a magnetic disk drive 151 that reads from or writes to a removable, non-removable magnetic disk 152, and an optical disk drive 155 that reads from or writes to a removable, non-volatile optical disk 156, such as a CD-ROM or other optical media. Other removable/non-removable, volatile/non-volatile computer storage media that can be used in the exemplary operating environment include, but are not limited to, magnetic tape cassettes, flash memory cards, digital versatile disks, digital video tapes, solid state memory, ROM or other optical media. Other removable/non-removable, volatile/non-volatile computer storage media such as a CD-ROM or other optical media.

The computer 110 is connected to the LAN 171 through a network interface or adapter 170. When used in a WAN networking environment, the computer 110 typically includes a modem 172 or other means for establishing communications over the WAN 173, such as the Internet. The modem 172, which may be internal or external, may be connected to the system bus 121 via the user input interface 160, or other appropriate mechanism. In a networked environment, program modules depicted relative to the computer 110, or portions thereof, may be stored in the remote memory storage device. By way of example, and not limitation, FIG. 1 illustrates remote application programs 185 as residing on memory device 181. It will be appreciated that the network connections shown are exemplary and other means of establishing a communications link between the computers may be used.

B. Overview: Versioned Nodes and Left Links

The present invention generally improves on previous methods for traversing concurrent B-Trees, and in particular it improves on an algorithm by Yehoshua Sagiv (we refer to this as the “Sagiv algorithm”). As discussed above, a B-Tree is a data structure that maintains an association of “keys” (such as employee numbers) to “values” (such as employee records). B-Trees are typically stored on disk. B-Trees are at the foundation of most information retrieval systems.

FIG. 2 depicts a typical information retrieval system 200. As shown, such a system can include a server 202 and a persistent store, such as a database, 204. In addition, the data residing in the store 204 may be organized in the form of a tree, e.g., a B-Link-Tree 206. Such a data structure includes nodes, N1, N2, N3 and so on, and, in the case of index nodes, links from each node to at least one other node (data nodes typically have only incoming links). The nodes may be sized to correspond to a disk block, or may be bigger or smaller, and may be formed as data nodes and index nodes (discussed further below in connection with FIG. 3A). Further, there is a root node (node N1 in FIG. 2) and children nodes, with sibling nodes being those nodes that have a common parent (e.g., nodes N2 and N3 are siblings). Index nodes may also be linked to their right siblings, as shown.

Information retrieval systems typically support concurrent access to and updating of the data maintained by them, which means that there may be multiple concurrent lookup and update operations on the underlying B-Tree. In order to prevent these concurrent operations from corrupting the B-Tree, some synchronization scheme is required. Typical concurrent B-Tree algorithms synchronize concurrent operations at the node-level of the tree; that is, an operation that wants to modify a node of the tree has to acquire a lock...
on that node, in order to guarantee that it does not interfere with another concurrent update (or other) operation on the same node.

Lock acquisition is expensive in several respects: It can be computationally expensive (in particular when the B-tree is replicated across multiple computers, meaning that locks have to be acquired from a remote lock server), and it limits concurrency. It is therefore desirable to minimize the number of lock acquisitions (without compromising the correctness of the algorithm). Much research has been devoted to this topic (Paul Wang, An In-Depth Analysis of Concurrent B-Tree Algorithms, cited above, contains a good survey of work on this problem.) To our knowledge, the B-Tree algorithm that performs best with respect to minimizing lock acquisitions is the Sagiv algorithm (see Yehoshua Sagiv, Concurrent Operations on B-Trees with Overwriting, cited above).

The system described herein improves on Sagiv’s algorithm in two respects: First, it avoids an inefficiency of Sagiv’s algorithm, namely that operations may get “lost” while trying to locate a data record and have to be restarted; second, it introduces a scheme for garbage-collecting deleted nodes that does not require any additional lock acquisitions, while Sagiv’s technique requires additional locking. The two optimizations are independent of each other, that is, Sagiv’s algorithm can be improved by adding either one or both.

Before describing the invention, it is helpful to review B-Trees in general and Sagiv’s algorithm in particular.

C. B-Trees and the Sagiv Algorithm

A B-Tree is a data structure that maintains an association of keys with values. A prerequisite is that there exists a total ordering over the keys, i.e., that it is always possible to decide whether one key is larger than the other. As indicated in FIG. 2, reference numeral 208, B-Trees support three basic operations:

1) insert(k, v), which associates the key k with the value v;
2) lookup(k), which returns the value v associated with the key k; and
3) delete(k), which disassociates the key k from its associated value.

B-Trees were first described by Bayer and McCreight (R. Bayer and E. McCreight, Organization and Maintenance of Large Ordered Indexes, cited above). There are many variations of B-Trees, including B*-trees (see H. Wedekind, On the selection of access paths in an information retrieval system, cited above), B+ Trees (see D. Comer, The Ubiquitous B-Tree, ACM Computing Surveys, cited above), and B-Link Trees (see P. L. Lehman and S. B. Yao, Efficient Locking for Concurrent Operations on B-Trees, cited above). Sagiv’s algorithm uses B-Link Trees. The present invention is applicable to all types of B-Trees and variations thereof, and the term B-Tree as used herein in describing the invention is intended to encompass all variants of the basic B-Tree structure. Preferred embodiments of the invention use a new form of B-Link Tree.

A B-Link Tree (and in fact any B-Tree) stores keys and values as well as metadata in nodes. Nodes are kept on disk or some other storage device (B-Trees make sense for any slow and cheap storage device), and are read into main memory on demand, and written back to disk if modified. Nodes on disk are identified by handles. (For the purpose of this discussion, it is sufficient to think of handles as the addresses of contiguous ranges of storage blocks (e.g., disk block addresses) plus optional metadata.) As indicated in FIG. 2, reference numeral 210, an allocator is a software component that maintains nodes on disk and supports four operations:

1) allocate(n), which reserves space on the disk for a node with a maximum size of n bytes and returns a handle to it;
2) deallocate(h), which relinquishes the space at the disk location identified by the handle h;
3) read(h), which reads the node from the disk location identified by the handle h and returns it; and
4) write(h, A), which writes the node A from main memory to the disk location identified by handle h.

In the following discussion, we assume that the allocator operations are atomic, that is, two concurrent operations on the same handle do not interfere with each other.

Nodes in a B-Tree may contain handles referring to other nodes. In most B-Tree variants, the handles connect the nodes to form a tree (hence the name), a directed, connected, and acyclic graph. In the following, we assume the reader to be familiar with the definition of a tree and the terms subtree, link, root, leaf, parent, child, and sibling. B-Link Trees differ from proper trees in that in addition to the links from parents to children, every node has a link to its directly adjacent right sibling (if such a sibling exists). This can be seen in the exemplary B-Link Tree 300 of FIG. 3A, where the “right link” (link to right sibling) is represented by reference numeral 302.

The B-Link Trees used by Sagiv are composed of two different kinds of nodes: data nodes and index nodes, reference numerals 304 and 306, respectively, of FIG. 3A. A data node is simply a key-value pair of the form <kid>. An index node is of the form:

<kmin , k1, k2, k3, k4, k5>, . . . , kn , kmax>

In the following, we refer to field x of node A as Ax. Given an index node A, A.hleft is a handle to the left sibling of A, and A.hright is a handle to its right sibling. A.ki (for 0 ≤ i < n) is the largest key in the subtree rooted at handle A.hleft (or A.hright) if i=n); A.kleft (or A.kright, if i=0) is less than the smallest key in the subtree rooted at handle A.hleft (or A.hright) if A does not have a right (left) sibling, and A.kmax is greater or equal to the largest key in any subtree of A (and per definition = if A does not have a right sibling), and A.kmin is equal to B.kmax, where B is the left sibling of A (or ∞ if A does not have a left sibling). Moreover, A.kleft <A.k1 < . . . <A.kn < A.kmax. Finally, there is a limit on the size of n (which indicates the number of keys and handles in an index node). If n reaches a maximum value (say, 2N), the index node is said to be full. Likewise, if n falls below a certain number (say, t), the node is said to be underfull.

Because of the constraints on the keys in an index node and the keys in the subtrees rooted at that node, B-Link trees are search trees, that is, trees where one can find a particular key by descending into the tree. Intuitively, lookup(k) starts at the root handle, reads in the corresponding node A, and identifies a value i such that A.ki-1 < k ≤ A.ki (or A.kmin if i=0) < k ≤ A.kmax (or A.kmax if i=n). It then recursively proceeds along the handle A.hleft until it reaches a data node B, and returns B’s value if B’s key is indeed k, or null otherwise.

The delete operation is similar to the lookup operation: delete(k) descends into the tree until it reaches a data node D with key k and removes the handle to D from main memory to the disk location identified by handle h.

The insert operations is more complicated: insert(k,v) allocates a new data node D with handle h, writes the pair...
was split (because IsLeaf(l1) (IsLeaf is discussed below-see section F) is shown. The tree contains the same mapping, {(2, Miller), (5, Smith), (7, Jones), (10, Brown), (12, Levin), (15, Dahl), (17, Lewis), (20, Yu)}. The numbers are key values, the names are the associated data values. P is the prime node of the tree; 11, . . . , 17 are index nodes; and D1, . . . , D8 are data nodes. A prime node is a node containing an array of handles, each handle pointing to the leftmost index node of a level of the tree, with the first handle in the prime node pointing to the leftmost leaf index node, and the last one pointing to the root index node. 11 is the root node of the tree (that is, the one index node that does not have a parent). One gets to the root node by following the last handle in the prime node (P[2] in this case). 11, 12, 13 are non-leaf index nodes (that is, index nodes that have index-node children). This means that 11.h12 is non-null, and that IsLeaf(11) (IsLeaf is discussed below—see section F) is false (the same is true for 12, 13). 14, 15, 16, and 17 are leaf index nodes, that is, index nodes that do not have index-node children. This means that 14.h15 is null, and that IsLeaf(14) is true (the same is true for 15, 16, and 17). In this tree, = 2; that is, each index node can hold two keys in addition to kmin and kmax. 11, 12, and 13 are not full (some fields are unused and shaded out). Num(11) (Num is discussed below—see section F) is 1 (the same is true for 12, 13). 11, 12, and 13 can each take one more key-handle pair. 14, 15, 16, and 17 are full, i.e., there are no unused fields. Num(14) is 2 (same for 15, 16, and 17).

FIG. 3D depicts a tree that contains left links, as shown. The tree contains the same mapping, {(2, Miller), (5, Smith), (7, Jones), (10, Brown), (12, Levin), (15, Dahl), (17, Lewis), (20, Yu)}, as the tree in FIG. 3C. As before, the numbers are key values, the names are the associated data values. P is the prime node of the tree; 11, . . . , 17 are index nodes; and D1, . . . , D8 are data nodes. A prime node is a node containing an array of handles, each handle pointing to the leftmost index node of a level of the tree, with the first handle in the prime node pointing to the leftmost leaf index node, and the last one pointing to the root index node. 11 is the root node of the tree (that is, the one index node that does not have a parent). One gets to the root node by following the last handle in the prime node (P[2] in this case). 11, 12, 13 are non-leaf index nodes (that is, index nodes that have index-node children). This means that 11.h12 is non-null, and that IsLeaf(11) (IsLeaf is discussed below—see section F) is false (the same is true for 12, 13). 14, 15, 16, and 17 are leaf index nodes, that is, index nodes that do not have index-node children. This means that 14.h15 is null, and that IsLeaf(14) is true (the same is true for 15, 16, and 17). In this tree, = 2; that is, each index node can hold two keys in addition to kmin and kmax. 11, 12, and 13 are not full (some fields are unused and shaded out). Num(11) (Num is discussed below—see section F) is 1 (the same is true for 12, 13). 11, 12, and 13 can each take one more key-handle pair. 14, 15, 16, and 17 are full, i.e., there are no unused fields. Num(14) is 2 (same for 15, 16, and 17).
nodes, D1, ..., D8 are data nodes. N is the root node of the tree (that is, the one index node that does not have a parent). One gets to the root node by following the last handle in the prime node (P[2] in this case). N, D2, D3 are non-leaf index nodes (that is, index nodes that have index-node children). This means that N.h2 is non-null, and that IsLeaf(N) is false (same for D2, D3). 14, 15, 16, and 17 are leaf index nodes (that is, index nodes that don’t have index-node children). This means that 14.h2 is null, and that IsLeaf(14) is true (same for 15, 16, and 17). In this tree, t = 2 (that is, each index node can hold two keys in addition to k_min and k_max). N, D2, and D3 are not full (some fields are unused and shaded out). Num(N) is 1 (same for D2, D3). N, D2, and D3 can each take one more key-handle pair. 14, 15, 6, and 17 are full (there are no unused fields). Num(14) is 2 (same for 15, 16, and 17).

To further illustrate the idea, we have produced below exemplary pseudocode of Sagiv’s lookup operation (adapted from Yehoshua Sagiv, Concurrent Operations on B-Trees with Overlapping, cited above), including the code for restarting the lookup operation if the lookup goes astray (but omitting the machinery for keeping track of operation start times):

```plaintext
procedure Lookup(h, Handle, k : Key)
var h: Handle; A: Node;
label start:
P := Read(h);
h := last item in P
repeat
A := read(h)
if k > A.k_max then
    h := A.h;
else if k ≤ A.k_min then
    goto start
else
    choose i such that A.k_{i-1} < k ≤ A.k_i (let i be 0
    if Num(A) = 0)
    h := A.h_i
endif
until A is a leaf and A.k_min < k ≤ A.k_max
if k is contained in A then
    return corresponding value
else
    return null
endproc
```

In contrast, here is exemplary pseudocode corresponding to one implementation of a lookup routine in accordance with the present invention. (Note that FIG. 7 depicts a flowchart of another exemplary Lookup procedure.) As can be seen, this routine follows the left-link instead of restarting the operation. This pseudocode is reproduced in FIG. 3E with a box 320 enclosing new lines of code designed to utilize the left-link data. The new lines of code determine whether the key value k is less than or equal to the value stored in the k_min field of node A, and if so it assigns the left-link handle to the handle variable (i.e., h := A.h_left).

```plaintext
procedure Lookup(h, Handle, k : Key)
var h: Handle; A: Node;
P := Read(h);
h := last item in P
repeat
A := read(h)
if k > A.k_max then
    h := A.h;
endproc
```

E. Attach Version Numbers to Handles and Disk Blocks

We will now describe an inventive way to avoid maintaining a table of operation start times. In particular, in a presently preferred implementation, we do this by attaching version numbers to handles and nodes or the “unit of allocation” returned by the allocator 210 (FIG. 2). As mentioned above, when a node is deleted by the delete operation or by the compression thread, it is possible that a concurrent operation has a handle to it. Sagiv’s algorithm deals with this problem by attaching a timestamp noting the time of deletion to every deleted node, and by recording the start time of every operation. Deleted nodes are “garbage collected” only when none of the ongoing operations have been started before the node was marked as deleted. This solution requires the system to maintain a table mapping thread to start times, which requires two synchronizations—one for adding a thread-start time pair to the table and one for removing it—since the table is being mutated by the client threads and accessed by the compression thread.

According to a presently preferred implementation of our invention, version numbers are attached to each handle and to each node. Initially, each node has version number 0. The allocate operation (described above) returns a handle identifying a node and embeds the node’s version number into the handle. The deallocate operation increments the node’s version number (in addition to making it available for subsequent reallocation). The read operation returns a node’s version number in addition to the node. The caller of the read operation (i.e., a lookup, insert, or delete operation) checks if the version number returned by the read operation matches the version number of the handle that was read. If the version numbers do not match, the node has been deleted and subsequently deallocated by a different thread. In this case, the lookup, insert, or delete operation is restarted.

Using this version numbering scheme, it is possible to deallocate nodes right when they are being marked as deleted. As a further optimization, we can defer deallocation deleted nodes for some fixed interval of time, which eliminates the needs for restarts for all but the most long-running operations.

In an alternative embodiment of this aspect of the invention, the read operation does not return the version number of the disk block that was being read, but instead raises an exception whenever the version number of the handle and the version number of the disk block do not match. The caller of the read operation (i.e., a lookup, insert, or delete operation) catches the exception and restarts the lookup, insert, or delete operation.

To further illustrate the idea, here is exemplary pseudocode of Sagiv’s lookup operation (again adapted
from Yehoshua Sagiv, *Concurrent Operations on B-Trees with Overtaking*, this time also including the machinery for keeping track of operation start times. \( T \) is a global variable version does not incorporate the use of left-links to avoid

In contrast, here is an example of pseudocode for a lookup routine in accordance with the present invention, where this routine avoids any operations on the table \( T \) (to emphasize that our invention consists of two independent parts, this version does not incorporate the use of left-links to avoid restarts):

```
procedure Lookup(h: Handle, k : Key)
  var h: Handle, A: Node;
  label start:
  P := Read(h)
  h := last handle in P
  repeat
    lock T
    insert (ID of this thread, current time) into T
    unlock T
    A := read(h)
    if k > A.kmax then
      h := A.h
    else if k < A.kmin then
      goto start
    else
      choose i such that A.k_{i+1} < k <= A.k_{i}
    endif
    if Num(A) > 0
      h := A.h
    endif
    until A is a leaf and A.k_{min} < k <= A.k_{max}
    lock T
    remove (ID of this thread, corresponding time) from T
    unlock T
    if k is contained in A then
      return corresponding value
    else
      return null
   endif
  endproc
```

Note that this pseudocode is also depicted in FIG. 3F, where the box 330 indicates the new line of code ("if version(A) > version(h) then goto start"). Moreover, the new pseudocode does not include the lines of code from the Sagiv algorithm relating to the lock operation. In other words, the following code has been eliminated:

lock T
insert (ID of this thread, current time) into T
unlock T

In addition, the code:

lock T
remove (ID of this thread, corresponding time) from T
unlock T

has been removed.

Of course, the two improvements disclosed herein may be combined to achieve the benefits of both of them. The pseudocode of the insert and delete operations and the compression thread is significantly more complicated but nonetheless well within the capabilities of a person skilled in the art. Moreover, the modifications needed to incorporate our invention should be straightforward after reading this specification.

F. Flowcharts of Exemplary Procedures

FIGS. 4-10 are flowcharts representing several procedures that may be employed in connection with a presently preferred implementation of an algorithm in accordance with the present invention. The following table summarizes the data types, notation and functions (index node and Allocator functions) referred to in these flowcharts.

<table>
<thead>
<tr>
<th>Primitive Types</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>( k )</td>
<td>a key value</td>
</tr>
<tr>
<td>( d )</td>
<td>a data value</td>
</tr>
<tr>
<td>( bs )</td>
<td>a block address</td>
</tr>
<tr>
<td>( v )</td>
<td>a version number</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Composite Types</th>
</tr>
</thead>
<tbody>
<tr>
<td>( h = (v, bs) )</td>
</tr>
<tr>
<td>( D = [k, d] )</td>
</tr>
<tr>
<td>( X, A, B = )</td>
</tr>
<tr>
<td>( (h_{n+1}, h_n, k_n, k_{n-1}, \ldots, h_1, h_0) )</td>
</tr>
<tr>
<td>( P )</td>
</tr>
<tr>
<td>( \nu )</td>
</tr>
<tr>
<td>( X.h_{n} )</td>
</tr>
<tr>
<td>( P[i] )</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Index Node Functions</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>( \text{Num}(X) )</td>
<td>returns ( n ) where ( X ) is an index node ( (h_{n}, h_{n-1}, \ldots, h_1, h_0) )</td>
</tr>
<tr>
<td>( \text{IsLeaf}(X) )</td>
<td>returns true if and only if ( X.h_0 = \text{null} )</td>
</tr>
<tr>
<td>( \text{IsDeleted}(X) )</td>
<td>returns true if and only if ( X.k_{\text{max}} = -\text{inf} )</td>
</tr>
<tr>
<td>( \text{Allocate}() )</td>
<td>allocates a block of the given size, and returns a handle to the block</td>
</tr>
<tr>
<td>( \text{Deallocation}() )</td>
<td>deallocates the block referred to by the handle ( h ), and increments the block's version number</td>
</tr>
<tr>
<td>( \text{Read}(h) )</td>
<td>reads the block referred to by the handle ( h ), and returns ( (v, N) ),</td>
</tr>
</tbody>
</table>
FIG. 4 is a flowchart of a procedure entitled GetNextHandle. This procedure takes an index node X and a key k and returns (h, isLink), where h is a handle referring to a child or sibling of X that is closer in the tree to the data node containing the key k and its associated data value, and isLink is a boolean value indicating whether h is a link to a sibling of X. As shown in FIG. 4, this procedure begins at an entry point for GetNextHandle(X, k) 402. At step 404, the procedure determines whether k is less than or equal to X.k_max, where X.k_max refers to the field k_max of index node X. If so, the procedure returns (X.h_off, true), indicating that the left sibling of X is closer to the desired data node. If not, the procedure determines whether k is greater than X.k_max, and if so, returns (X.h_off, true), indicating that the right sibling of X is closer to the desired data node (Step 410). If k is not greater than X.k_max as determined at step 408, the procedure determines whether Num(X) is equal to 0 (step 412), and if so, at step 414 returns (X.h_off, false). Moving on to step 416, the procedure determines whether X.k_i is less than k. If so, the procedure at step 418 returns (X.h_off, false). If not, the procedure moves on to step 420 and finds the smallest i such that k is less than or equal to X.k_i. Next, at step 422, the procedure tests whether IsLeaf(X) is true. If so, it returns (X.h_off, false) (Step 424), and if not, it returns (X.h_off, false) (Step 426). In sum, the procedure of FIG. 4 returns a handle and a boolean flag indicating whether the handle is a link to a sibling of node X.

FIG. 5 is a flowchart of a procedure entitled InsertPair, which takes an index node X, a key k, and a handle h and inserts the key k and the handle h into X. This procedure returns nothing.

As shown in FIG. 5, the procedure begins at the entry point InsertPair(X, k, h) (Step 502). Next, at step 504, the procedure determines whether Num(X) is equal to 0. If so, at step 506, the key k and handle h are inserted into the node X as shown. If Num(X) is not equal to 0, at step 508 the procedure determines whether X.k_i is less than k; and if not, proceeds to step 510 where the key k and handle h are inserted into the node X as shown. On the other hand, if the procedure determines that X.k_i is less than k, it proceeds to step 512 where it finds the largest i such that X.k_i is less than k, and then at step 514 inserts the key k and handle h into node X as shown. Finally, at step 516 the procedure returns. In sum, the key k and handle h are inserted into the node X at steps 506, 510, or 514, depending on whether Num(X) is equal to 0, and whether X.k_i is less than k.

FIG. 6 is a flowchart of a procedure entitled CreateTree. This procedure receives no argument and returns a prime node handle h_p identifying a new tree. As shown, the CreateTree procedure 600 begins at the entry point CreateTree() (step 602). At step 604, the node X is defined as (null, -inf, null, +inf, null). At step 606, the handle h, is defined as Alloc(space for X) where the function Alloc() allocates a block of a size given by the argument and returns a handle to the block. Next, at step 608 the procedure invokes Write(X, h_p), in order to write the node X to the block referenced by the handle h_p. At step 610 the procedure defines the prime node P to contain the handle h_p. At step 612 the procedure again invokes the function Alloc(space for P), in order to allocate a block sufficiently large to store P, which returns a handle h_p, to the block. At step 614, the procedure writes the node P to the block referenced by the handle h_p, and at step 616 the procedure concludes by returning the handle h_p. In sum, this procedure returns a prime node handle h_p identifying a new tree.

FIG. 7 is a flowchart of a procedure entitled Lookup, which takes a prime node handle h_p and a key k and returns a data item d or null. The Lookup procedure 700 begins at the entry point Lookup(h_p, k) (step 702). At step 704, the procedure invokes the function Read(h_p) in order to read the block referred to by the handle h_p. The Read function returns (v, P), the block’s version number v and the prime node P stored in the block. Next, at step 706, the handle h is defined as the last handle in the prime node P, thereby referring to the root node of the tree. At step 708, the procedure again invokes the function Read() in order to read the block referred to by handle h. This time, the function returns the version number v and the index node X stored in the block referenced by handle h. At step 710, the procedure determines whether h.v (i.e., the version number of handle h) is equal to v. If not, the procedure loops back as shown to step 704. If so, the procedure proceeds to step 712 where the function GetNextHandle (described above in connection with FIG. 4) is invoked. In this instance, the function GetNextHandle(X, k) returns the handle h and boolean isLink indicating whether the handle h is a link to a sibling of node X. Next, at step 714, the procedure tests whether handle h is null. If so, the procedure proceeds to step 716 where null is returned. If not, the procedure proceeds to step 718, where it determines whether both IsLeaf(X) and not isLink are true. If they are not both true, the procedure loops back to step 708 as shown. If, on the other hand, IsLeaf(X) and not isLink are both true, the procedure proceeds to step 720 where the function Read(h) is invoked in order to read the block referred to by handle h and thereby determine the block’s version number v and the data node D stored in the block. Next, at step 722, the procedure determines whether h.v (i.e., the version number of handle h) is equal to v. If not, the procedure loops back to step 704 as shown, and if so, proceeds to step 724. At step 724, the procedure determines whether D.k is equal to k (i.e., whether the key of data node D is equal to k). If not, the procedure proceeds to step 726 where it returns null, and if so, it proceeds to step 728 and returns D.d, that is, the data portion of the data node D. Thus, in sum, the Lookup procedure takes a prime node handle h_p and a key k and returns a data item d or null.

FIG. 8 is a flow chart of a procedure entitled Delete which takes a prime node handle h_p and a key k, and deletes the key and its associated data value from the tree identified by the handle h_p. This procedure returns nothing.

As shown, the Delete procedure 800 begins at the entry point Delete(h_p, k) (step 802). Next, at step 804, the procedure invokes the function Read(h_p) in order to read the block referred to by the handle h_p. This function returns the block’s version number v and the prime node P stored in the block. Next, at step 806 the handle h is defined as the last handle in node P, thereby referring to the root node of the tree. At step 808, the procedure again invokes the Read() function in order to read the block referred to by handle h, returning the block’s version number and the index node X contained therein. At step 810, the procedure tests whether h.v is equal to v, i.e., whether the version number of handle h is equal to the version number returned by the previous invo-
Write() function is invoked to write the node X referenced k and removes the key k and its associated data value from

If whether IsLeaf(X) and not isLink are both true; if not, the procedure loops back to step 808, and if so, proceeds to step 816. At step 816, the procedure tests whether handle h is null, and if so, proceeds to step 818 where it returns. If h is not null, the procedure proceeds to step 820 where the function Lock(h) is invoked in order to lock the block referred to by the handle h. Next, at step 822 Read(h) reads the block referred to by handle h, returning the block’s version number and the index node X contained therein. At step 824 the procedure tests whether h.v is equal to v, and if not, proceeds to step 826 where the block referred to by the handle h is unlocked, after which the procedure loops back to step 804 as shown. On the other hand, if h.v is equal to v, the procedure proceeds to step 828, where the procedure determines whether key k is less than or equal to X.k, i.e., whether the key k is found in a data node attached to one of the left siblings of X. If so, the procedure proceeds to steps 830 and 832, where the block referred to by handle h is unlocked and then the handle h is defined as the left sibling handle of node X. If key k is not less than or equal to X.k, the procedure proceeds to step 834 where it tests whether k is greater than X.k, and if so, proceeds to steps 836 and 838, where it unlocks the block referred to by handle h and defines h to be the right sibling handle of node X. If at step 834 the procedure determines that k is not greater than X.k, it proceeds to step 840 where it tests whether Num(X) is equal to 0, meaning that X (which is a leaf-level index node) has no children, and that if therefore does not contain the key k. If the answer to this test is yes, the procedure moves ahead to step 856, and if it is no, it proceeds to step 842. At step 842, the procedure determines whether X.k is less than k, i.e., whether the field k of node X is less than the key value k. If it is, the procedure jumps ahead to step 856 as shown, and if it is not, it proceeds to step 844, where it finds the smallest i such that k is less than or equal to X.k. Next, at step 846, the procedure invokes Read(X.h[i]), reading a block containing a data node D and a version number v. Next, at step 848 the procedure tests whether D.k is equal to k. If it is not, the procedure jumps to step 856, and if so, it proceeds to step 850 where it deallocates the block referred to by the handle X.h[i+1], which increments the block version number. Next, at step 852 the procedure removes the handles identified by X.k and X.h[i] from node X. At step 854, the Write() function is invoked to write the node X referenced by handle h. Finally, at steps 856 and 858 the block referred to by handle h is unlocked and the procedure returns. In sum, the Delete procedure takes a prime node handle hp, a key k and a data value d and inserts the key and the data value into the tree identified by hp. The Insert function 900 begins at the entry point identified by reference numeral 901, and is invoked by the call Insert(hp, k, d), where hp is a prime node handle, k is a key value, and d is a data value. At step 902, the Data node D is defined by the key value k and data value d. At step 903, the Alloc function is invoked so as to allocate a block large enough to contain D; and the handle identifying the block is assigned to hp. Next, at step 904 the procedure writes the node D to the block referenced by handle hp. At step 905, Lr is set to 0 and then at step 906 S is defined as the new empty stack. (By way of providing further background, we note that the levels of a tree are the sets of siblings. In the terminology used in the instant application, level 0 contains all leaf index nodes, level 1 contains their parents, and so on; the highest level contains the root node. The number of the highest level is one less than the number of handles in a prime node (since we use zero-based indexing). The “target level” variable Lr identifies the level at which the key-handle pair (k, hp) shall be inserted into the tree. Setting Lr to 0 means that the pair shall be inserted into a leaf index node. As the algorithm unfolds, Lr may be incremented, indicating that a (modified) key-handle pair shall be inserted into an index node higher up in the tree (potentially all the way up to the root.) At step 907, the Read( ) function is invoked to read the block referred to by handle h, so as to identify the version number v and the prime node P stored in the block. At step 908, the last handle in node P (the root index node) is assigned to hp. At step 909 Lr is defined to be the number of handles in node P, which is identical to the number of index node levels in the tree. At step 910, Lc is defined as Lr−1. Lc (the “current level”) is the level the insert procedure is currently at. Since Insert starts at the root (which is at level Lr−1, as explained above), Lc is initialized to this value, and is decremented as Insert descends into the tree. At step 912 Read(hp) is invoked to return the version number v and the index node X referred to by handle hp. At step 913, the procedure tests whether h.v is equal to v, and at step 914 the procedure tests whether Lc is greater than Lr. If the answer to both of the tests of steps 913 and 914 is yes, the procedure proceeds to step 915. If, on the other hand, the answer to step 913 is no, the procedure loops back to step 906 as shown. If the answer to the test of step 914 is no, the procedure proceeds to step 920 (see subpart 9B of FIG. 9). At step 915, the procedure invokes the GetNextHandle function with the arguments X and k, whereby the handle h and boolean isLink are returned. As discussed previously, the boolean isLink indicates whether the handle h is a link to a sibling of node X. At step 916, the procedure tests whether isLink is true, and if it is, branches ahead to step 919. If isLink is false, the procedure proceeds to step 917, where it decrements Lc. At step 918, the procedure pushes the handle hp onto stack S. At step 919, the function assigns the value of h’ to hp.

Now referring to subpart 9B of FIG. 9, at step 920 the procedure invokes the Lock( ) function in order to lock the block referred to by the handle hp. At step 921 the procedure invokes Read(hp), which reads the block identified by handle hp and extracts an index node X and a version number v. At step 922, the procedure tests whether h.v is equal to v. If not, it branches to step 906 (subpart 9A) as shown. If, on the other hand, h.v is equal to v, the procedure proceeds to step 923, where it tests whether the function IsLeaf(X) is true. If not, the procedure branches to step 934 (subpart 9C of FIG. 9), and if so, it proceeds to step 925, where it tests whether X.k is less than k. If the answer to the test of step 925 is yes, the procedure branches to step 934, and if it is no, the procedure proceeds to step 926. At step 926, the procedure finds the smallest i such that k is less than X.k. At step 927, the procedure invokes Read(X.h[i]), which reads the block identified by handle X.h[i] and extracts a data node D and a version number v. At step 928, the procedure tests whether D.k is equal to k; if it is, the procedure proceeds to step 929, and otherwise it branches to step 934. At step 929, the
procedure defines node D to be a data node containing the key value k and data value d. Next, at step 930 the function Write(D, X, h x) is invoked in order to write the node D to the block referenced by the handle h x. Next, at step 931, the function Dealloc(h p) is invoked in order to deallocate the block referred to by the handle h p, and increment the block’s version number. At step 932, the block referred to by the handle h y is unlocked, and then at step 933 the procedure returns.

Referring now to subpart 9C of FIG. 9, at step 934 the procedure tests whether k is greater than X.k.max, and less than or equal to X.k.min. If the answer to the test is yes, the procedure proceeds to step 941; if the answer is no, the procedure proceeds to step 935. At step 935, the block referred to by the handle h y is unlocked, and then at step 936 the GetNextHandle function is invoked with the arguments X and k, which thereby returns the handle h' and the boolean isLink. As discussed previously, this function returns the handle h' and the boolean isLink indicating whether this handle is a link to a sibling of node X. At step 937, the procedure tests whether isLink is true; if so, it proceeds to step 938 and if not, it branches to step 920 (see subpart 9B of FIG. 9). At step 938, the handle h x is assigned the value of h; and then at step 939 Read(h x) is invoked to determine the version number v and node X referred to by the handle h x. At step 940, the procedure tests whether h x.v is equal to v; if so, it branches back to step 936 as shown, and if not, it branches to step 906 (subpart 9A of FIG. 9).

As mentioned above, at step 934 the procedure tests whether k is greater than X.k.max, and less than or equal to X.k.min. If the answer to this test is yes, the procedure proceeds to step 941 where it tests whether node X is full. If X is full, the procedure proceeds to step 946 (subpart 9D of FIG. 9), otherwise it proceeds to step 942. At step 942, the procedure invokes InsertPair(X, k, h y), in order to insert the key k and handle h y in node X. At step 943, Write(X, h y) is invoked in order to write the node X to the block referenced by the handle h y. At step 944, the block referred to by the handle h x is unlocked, and then at step 945 the procedure returns.

Referring now to subpart 9D of FIG. 9, if node X is full, then X is split into two nodes X and Y. At step 946 the procedure tests whether the function IsLeaf(X) is true. If it is, the procedure proceeds to step 947, and if it is not, the procedure proceeds to step 951. Steps 947, 948 and 949, on the one hand, and steps 951, 952 and 953 on the other hand are performed in order to define nodes Y and X, depending on whether X is a leaf or not. Steps 947 and 951 are performed in order to define the node Y; steps 948 and 952 are performed to allocate space for Y, and steps 949 and 953 are performed to define the node X. At step 950, the procedure tests whether X.k.max is less than key value k. If not, the procedure proceeds to step 954 where the function InsertPair (discussed above in connection with FIG. 5) is invoked with the arguments X, k and h p to insert k and h p into X. In step 955, Write(Y, h y) is invoked in order to write the node Y. Similarly, at step 957 Write(X, h x) is invoked in order to write the node X. If at step 950 it is determined that X.k.max is less than k, the procedure proceeds to step 956 where the function InsertPair(Y, k, h y) is invoked to insert k and h y into Y; after which steps 955 and 957 are performed as described above. At step 958, Read(h x) is invoked and this function returns the version number v and the prime node P in the block referred to by the handle h x.

The procedure then proceeds to step 959 (see subpart 9E of FIG. 9).

Referring to subpart 9E of FIG. 9, at step 959 the procedure determines whether handle h y is the last handle in prime node P, meaning that the node identified by h y is the root of the tree. If so, the procedure proceeds to steps 960 through 966 and if not, proceeds to steps 967 through 972. At step 960 the procedure defines the index node Z as shown. At step 961 the handle h y is defined with the function Alloc(space for Z), in order to allocate a block large enough to hold Z and return a handle to the block. In other words, the handle h y is returned by the function Alloc(). At step 962 the function Write(Z, h y) is invoked in order to write the index node Z to the block referenced by the handle h y. At step 963, the procedure appends h y to the end of prime node P. At step 964, Write(P, h y) is invoked to write the node P to the block referenced by the handle h y. Next, at step 965 the block referred to by the handle h y is unlocked and then at step 966 the procedure returns.

If at step 959 the answer to the test whether h y is the last handle in P is no, the procedure proceeds to step 967, wherein the block referred to by the handle h y is unlocked. Then, at step 968, the handle h y is set to h x. At step 969 the key value k is set to X.k.min. At step 970, the procedure determines whether the stack S is empty; if not, it proceeds to step 971 wherein the handle h x is defined from the last value on the stack S. On the other hand, if S is empty, the procedure proceeds to step 973 where handle h x is defined with P[L.3], and then at step 974 L.x is defined as L.x+1. Finally, at step 972 L.f is set to L.f+1, and then the procedure loops back to step 920 (see subpart 9B of FIG. 9), as shown.

In sum, the Insert procedure of FIG. 9 (subparts 9A through 9E), takes a prime node handle h y, a key value k, and a data value d and inserts the key-data pair into the tree identified by h y.

FIG. 10 is a flowchart of a procedure entitled Compress. This procedure takes a prime node handle h y and compresses the node; it does not return, as it is executed by a background process. The Compress procedure 1000 begins at an entry point 1002 with the call Compress(h y). At step 1004, Read(h y) is invoked, and this function returns a version number v and the prime node P referred to by the handle h y. At step 1006, the index i is set to 1. At step 1008, the procedure tests whether the number of handles in node P is greater than i. If the number of handles in P is greater than i, the procedure proceeds to step 1010 to cause the function CompressLevel() with the argument P[i]. This function is described below in connection with FIG. 11. Briefly, CompressLevel( ) is a function that takes an index node handle h x and compressed the index node referred to by that handle and all its right siblings. In step 1012, the index i is set to i+1, and then the procedure loops back to decision step 1008. At step 1008 if it is determined that the number of handles in P is not greater i, the procedure proceeds to step 1014, where the handle h is set to the last handle in node P. Next, at step 1016, the block referred to by the handle h is locked, and then at step 1018 Read(h) is invoked, and this function returns the version number v and node X referred to by handle h. Next, at step 1020, the procedure tests whether Num(X) is equal to 0 and X.h.right is equal to null and not IsLeaf(X) are all true. If so, the procedure proceeds to step 1022 and invokes Read(h y), which returns the version number v and prime node P. If the answer to the test at step 1020 is no, the procedure jumps ahead to step 1032 and unlocks the block referred to by handle h. After step 1022 is performed, the procedure moves on to step 1024 and tests whether h is the last handle in P (that is, if h refers to the root node of the tree); if not, the procedure jumps ahead to step 1032, and if so, proceeds to step 1026. At step 1026 the procedure...
removes the last handle from P, and at step 1028 writes the node P referenced by the handle h. Next, at step 1030, the procedure deallocates the block referred to by handle h, and then proceeds to step 1032, where it unlocks the block referred to by handle h. Finally at step 1034, the procedure pauses and then loops back to step 1004, as shown. In sum, this procedure compresses the tree identified by handle h.

The CompressLevel procedure 1100 begins at an entry point 1101 with the call CompressLevel(h). At step 1102, the variable a is set to 0 and at step 1103, the handle h_a is set to null. Next, at step 1104 the procedure tests whether the handle h_a is null, and if so, proceeds to step 1105 where the procedure returns. If h_a is not null, the procedure proceeds to step 1106 and locks the block referred to by handle h_a. Next, at step 1107, Read(h_a) is invoked to read index node X and its version number v. At step 1108, the procedure tests whether handle h_a is null. If h_a is null, the procedure branches to block 115 (subpart 11B of FIG. 11), and if h_a is not null, the procedure proceeds to step 1109 where variable a is set to -1. Next, at step 1110, index i is set to zero, and then at step 1111 the procedure tests whether Num(X) is greater than or equal to i, i.e., the procedure tests whether the number n returned by the function Num(X) is greater than or equal to index i. If it is not, the procedure branches to step 1115 (subpart 11B of FIG. 11), and if so, proceeds to step 1112. At step 1112, the procedure tests whether the handle to the ith child of node X (i.e., X.h_i) is equal to handle h_a. If it is not, the procedure does not proceed to block 1114 and increments index i and then proceeds back to step 1111. If X.h_i is equal to handle h_a, the procedure proceeds to step 1113 where variable a is set to the value of index i.

Referring now to subpart 11B of FIG. 11, at step 1115, the procedure tests whether the handle h_x is null or the variable a is neither equal to -1 nor equal to Num(X). If the answer to the test of step 1115 is no, the procedure proceeds to step 1116 through 1118 and then back to step 1104, as shown. Steps 1116 through 1118 are performed to unlock the block referred to by the handle h_a, set h_a to the field h_left of index node X, and then to set handle h_x to null.

If the answer to the test of step 1115 is yes, the procedure performs a series of steps beginning with step 1119, where it tests whether handle h_d is null. If h_d is null, the procedure at step 120 sets variable a to 0. If handle h_d is not null, the procedure at step 1202 increments a. Then, at step 1212, handle h_a is set to X.h_d. At step 1213, the block referred to by handle h_a is locked, and then at step 1214 the function Read(h_a) is invoked, to read index node A and its associated version number v. At step 1215, the handle h_g is set to A.h_right. Next, at step 1216 the procedure tests whether the handle h_g is null. If it is, steps 1217 through 1233 are performed, i.e., the block referred to by handle h_g is unlocked, the block referred to by handle h_a is unlocked, and then the procedure returns. On the other hand, if handle h_g is not null, the procedure proceeds to step 1227, and locks the block referred to by handle h_g. Next, at step 1228, Read(h_g) is invoked to determine the version number v and node B. At step 1229, B.h_g is set to the value of handle h_x. Next, at step 1300, b is set to 0.

The remaining steps of the CompressLevel procedure are shown in subpart 11C of FIG. 11. Referring to this subpart, at step 1134 the procedure tests whether X.h_d is h_g. If so, it proceeds to step 1144; otherwise, it proceeds to step 1135. At step 1135, variable b is incremented, and at step 1136 the procedure tests whether Num(X) is greater than or equal to b. If so, the procedure loops back to step 1134 as shown. If not, the procedure proceeds to step 1137 and unlocks the block referred to by handle h_g. Next, at step 1138, the procedure unlocks the block referred to by handle h_g, and at step 1139 unlocks the block referred to by handle h. At step 1140 the procedure tests whether B.k_max is less than X.k_max. If not, the procedure performs step 1143, in which handle h is set to X.h_d. If B.k_max is less than X.k_max as determined at step 1140, the procedure proceeds to step 1141 and tests whether Num(A) is less than t or Num(B) is less than t. If either of these is true, the procedure proceeds to step 1142 and sets handle h_d to null, and then branches back to step 1104 (subpart 11A of FIG. 11) as shown.

At step 1144, the procedure again tests whether Num(A) is less than t or Num(B) is less than t. If not, the procedure proceeds to steps 1152 through 1154 as shown, and from there to steps 1150 and 1151, after which the procedure branches back to step 1104 (subpart 11B of FIG. 11). If the answer to the test of step 1144 is yes, the procedure proceeds to steps 1145, 1146, 1147, 1148, and 1149 as shown (these steps should by now be self-explanatory and are obviously not executed in order). After these steps are performed, the procedure tests whether node B is deleted, and if so, branches back to step 1104. If node B is not deleted, the procedure sets at step 1151 h_d to h_g.

FIG. 12 is a flowchart of a procedure entitled MergeNodes. This procedure, as its name implies, merges nodes; it takes index nodes X, A, B, (where A is the left sibling of B and X is their parent) and index node handles h_x, h_A, and h_B referring to these nodes, and merges the contents of A and B into A, discarding B in the process. The MergeNodes procedure 1200 begins at entry point 1202 and is invoked with the call MergeNodes(X, A, B, h_x, h_A, h_B), as shown. At step 1204, the procedure tests whether IsLeaf(A) is true. Recall that IsLeaf(X) is a function that returns true if, and only if, X.h_d is null. If IsLeaf(A) is true, the procedure performs step 1206, and if it is false, the procedure performs steps 1210. Steps 1206 and 1210 assign the values shown in the corresponding blocks to node A. Next, at step 1208 the procedure performs the operation shown with respect to node B. At step 1212, node X is defined as shown. Steps 1214 through 1228 are then performed to write nodes A, X, and B, unlock handles h_A, h_B, and h_x, and to deallocate index node B identified by handle h_x. The procedure returns at step 1228.

FIG. 13 is a flowchart of a procedure entitled MoveToLeft. This procedure is performed in order to move a node to the left in order to move part of the contents of an index node to its left sibling. The MoveToLeft procedure 1300 begins at entry point 1302 with the call MoveToLeft(X, A, B, h_x, h_A, h_B). At step 1304 the variable m is defined as \( \frac{1}{2} \text{Num}(B) - \frac{1}{2} \text{Num}(A) \). Recall that the function Num(X) returns n where X is an index node. At step 1306, the procedure performs steps 1308 and 1310 as shown; and if not, the procedure performs steps 1314 and 1316. These steps assign the values shown in the corresponding blocks to nodes A and B.

Steps 1312 through 1330 write out and unlock the index nodes A, X, and B, and then return.

FIG. 14 is a flowchart of a procedure entitled MoveToRight which takes index nodes X, A, B and index node handles h_x, h_A, h_B. Steps 1402 through 1406 are directly analogous to steps 1302 through 1306 discussed above, except in 1404 the variable m is defined as \( \frac{1}{2} \text{Num}(A) + \frac{1}{2} \text{Num}(B) \) instead of as \( \frac{1}{2} \text{Num}(B) - \frac{1}{2} \text{Num}(A) \). The remaining steps of procedure 1400 are similar to the corre-
sponding steps of procedure 1300 discussed above but are modified in order to accomplish the move the right function. These differences will be apparent to one of ordinary skill in the art and will not be described here.

G. Conclusion

As mentioned above, while exemplary embodiments of the present invention have been described in connection with various computing devices and network architectures, the underlying concepts may be applied to any computing device or system in which it is desirable to traverse and/or perform other functions in connection with a B-Link or other data structure. Thus, the procedures and systems described above may be applied to a variety of applications and devices. While exemplary data structures, programming languages, names and examples are chosen herein as representative of various choices, these are not intended to be limiting.

The various techniques described herein may be implemented in connection with hardware or software, or, where appropriate, with a combination of both. Thus, the methods and apparatus of the present invention, or certain aspects or portions thereof, may take the form of program code (i.e., instructions) included in tangible media, such as floppy diskettes, CD-ROMs, hard drives, or any other machine-readable storage medium, wherein, when the program code is loaded into and executed by a machine, such as a computer, the machine becomes an apparatus for practicing the invention. In the case of program code execution on programmable computers, the computing device will generally include a processor, a storage medium readable by the processor (including volatile and non-volatile memory and/or storage elements), at least one input device, and at least one output device. One or more programs that may utilize the debugging interface aspects of the present invention, e.g., through the use of a data processing API or the like, are preferably implemented in a high level procedural or object-oriented programming language to communicate with a computer system. However, the program(s) can be implemented in assembly or machine language, if desired. In any case, the language may be a compiled or interpreted language, and combined with hardware implementations.

The methods and apparatus of the present invention may also be practiced via communications embodied in the form of program code that is transmitted over some transmission medium, such as over electrical wiring or cabling, through fiber optics, or via any other form of transmission, wherein, when the program code is received and loaded into and executed by a machine, such as a gate array, a programmable logic device (PLD), a client computer, or the like. When implemented on a general-purpose processor, the program code combines with the processor to provide a unique apparatus that operates to invoke the functionality of the present invention. Additionally, any storage techniques used in connection with the present invention may invariably be a combination of hardware and software.

While the present invention has been described in connection with the presently preferred embodiments, it is to be understood that other similar embodiments may be used or modifications and additions may be made to the described embodiment for performing the same function of the present invention without deviating therefrom. For example, one skilled in the art will recognize that the present invention as described in the present application may apply to any computing device or environment, whether wired or wireless, and may be applied to any number of such computing devices connected via a communications network, and interacting across the network. Furthermore, it should be empha-

sized that a variety of computer platforms, including handheld device operating systems and other application specific operating systems are contemplated, especially as the number of wireless networked devices continues to proliferate.

Still further, the present invention may be implemented in or across a plurality of processing chips or devices, and storage may similarly be effected across a plurality of devices. Therefore, the present invention should not be limited to any single embodiment, but rather should be construed in breadth and scope in accordance with the appended claims.

We claim:

1. An information management system, comprising:
   a computer;
   a database operatively connected to said computer, wherein said database comprises a B-Tree data structure comprising a plurality of nodes associated with disk blocks and handles stored in said nodes, wherein at least one left-link handle, hleft, stored in each node points to a left sibling of that node; and
   a mechanism for performing a lookup operation with respect to a key, k, wherein, in traversing the B-Tree at a given node, said lookup operation refers to said left-link handle, hleft, of that node to access a left sibling of the node if the key k is less than or equal to a value K_{min} stored in the node.

2. An information management system as recited in claim 1, further comprising a mechanism for performing an insert operation.

3. An information management system as recited in claim 1, wherein the insert operation detects if the key range of an index node, A, does not include the key k that the operation is trying to locate, and wherein said operation follows a handle A_{hleft} to the left sibling when k \leq A_{k_{min}}.

4. An information management system as recited in claim 1, further comprising a mechanism for performing a delete operation.

5. An information management system as recited in claim 1, wherein the delete operation detects if the key range of an index node, A, does not include the key k that the operation is trying to locate, and wherein said operation follows a handle A_{hleft} to the left sibling when k \leq A_{k_{min}}.

6. An information management system as recited in claim 1, further comprising mechanisms for performing insert and delete operations, wherein said lookup, insert and delete operations detect if the key range of an index node, A, does not include the key k that the operation is trying to locate, and wherein said operations follow a handle A_{hleft} to the left sibling when k \leq A_{k_{min}}.

7. An information management system as recited in claim 1, wherein the B-Tree data structure comprises at least one index node having the following structure:

\[
<h_{left}, h_{min}, h_{0}, k_{0}, h_{1}, k_{1}, \ldots, h_{n-1}, k_{n-1}, h_{n}, k_{n}, h_{right}>
\]

wherein h_{left} is a handle of a left sibling and the remaining data elements (k_{min}, h_{0}, k_{0}, h_{1}, k_{1}, \ldots, h_{n-1}, k_{n-1}, h_{n}) are handles and keys according to a B-Link Tree data structure.

8. A computer-implemented method, comprising maintaining a data structure comprising a plurality of nodes and handles stored in said nodes, wherein at least one left-link handle, h_{left}, stored in each node points to a left sibling of that node; wherein the method further comprises performing a lookup operation with respect to a key, k, wherein the lookup operation involves traversing the data structure and, at a given node, referring to said left-link handle, h_{left}, of that
node to access a left sibling of the node if the key \( k \) is less than or equal to a value \( k_{\text{min}} \) stored in the node.

9. A method as recited in claim 8, further comprising performing an insert operation.

10. A method as recited in claim 9, wherein the insert operation detects if the key range of an index node, \( A \), does not include the key \( k \) that the operation is trying to locate, and wherein said operation follows a handle \( A.h_{\text{left}} \) to the left sibling when \( k \leq A.k_{\text{min}} \).

11. A method as recited in claim 8, further comprising a mechanism for performing a delete operation.

12. A method as recited in claim 11, wherein the delete operation detects if the key range of an index node, \( A \), does not include the key \( k \) that the operation is trying to locate, and wherein said operation follows a handle \( A.h_{\text{left}} \) to the left sibling when \( k \leq A.k_{\text{min}} \).

13. A method as recited in claim 12, further comprising performing insert and delete operations, wherein said lookup, insert and delete operations detect if the key range of an index node, \( A \), does not include the key \( k \) that the operation is trying to locate, and wherein said operations follow a handle \( A.h_{\text{left}} \) to the left sibling when \( k \leq A.k_{\text{min}} \).

14. A method as recited in claim 8, wherein the data structure comprises at least one index node having the following structure:

\[
\langle h_{\text{left}}, h_{\text{main}}, h_0, k_0, h_1, k_1, \ldots, h_{n-1}, k_{n-1}, h_n, k_n, h_{\text{right}} \rangle
\]

wherein \( h_{\text{left}} \) is a handle of a left sibling and the remaining data elements \((h_0, h_1, k_0, k_1, \ldots, h_{n-1}, k_{n-1}, h_n, k_n, h_{\text{right}})\) are handles and keys according to a B-Link Tree data structure.

* * * * *