An information management system includes a database comprising a B-Tree data structure. The B-Tree includes a plurality of nodes associated with disk blocks, handles stored in the nodes, and version numbers attached to each of the handles and disk blocks. In addition, the system includes a mechanism for initially assigning each disk block a first prescribed version number (e.g., version number 0), and a mechanism for performing an allocate operation whereby a handle identifying a disk block is returned along with the disk block’s version number. The system may also include a mechanism for performing a deallocate operation that increments the disk block’s version number, as well as a mechanism for performing a read operation that returns a disk block’s version number in addition to the node stored at that disk block. Moreover, a caller of the read operation checks whether the version number returned by the read operation matches the version number of the handle that was read. The caller is preferably restarted if the version numbers do not match, thus indicating that the node has been deleted and subsequently deallocated by a different thread. The caller may be a lookup operation, an insert operation, or a delete operation.

19 Claims, 23 Drawing Sheets
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)
**FIGURE 3B**

B-LINK TREE DATA STRUCTURE

(INDEX NODES: \( V \), \( h_{\text{left}} \), \( k_{\text{min}} \), \( h_p \), \( k_p \), \( h_{\text{right}} \), \( k_{\text{left}} \), \( h_{n+1} \), \( k_{n+1} \), \( h_{\text{max}} \))

**FIGURE 3E**

LOOKUP PROCEDURE USING HANDLES TO LEFT SIBLINGS

procedure Lookup(h : Handle, k : Key)
var h: Handle, A: Node;
P := Read(h);
h := last handle in P
repeat
A := read(h)
if \( k > A.k_{\text{max}} \) then
  \( h := A.h_{\text{right}} \)
else if \( k \leq A.k_{\text{min}} \) then
  \( h := A.h_{\text{left}} \)
else
  choose \( i \) such that \( A.k_{i-1} < k \leq A.k_i \) (let \( i = 0 \) if Num(A) = 0)
  \( h := A.h_i \)
endif
until A is a leaf and \( A.k_{\text{min}} < k \leq A.k_{\text{max}} \)
if k is contained in A then
  return corresponding value
else
  return null
endif
FIGURE 3C
FIGURE 3D
procedure Lookup(h : Handle, k : key)
    var h: Handle, A: Node;
label start:
    P := Read(h)
    h := last handle in P
repeat
    if 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
    else
        A := read(h)
        if version(A) != version(h).then goto start
        if k > A.k.max then
            h := A.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
endproc
FIGURE 4

GetNextHandle(X, k) entry point

\[ k \leq X.k_{min} \rightarrow \text{return } (X.h_{left}, \text{true}) \]

\[ k > X.k_{\max} \rightarrow \text{return } (X.h_{right}, \text{true}) \]

\[ \text{Num}(X) = 0? \rightarrow \text{return } (X.h_{0}, \text{false}) \]

\[ X.k_{i} < k \rightarrow \text{return } (X.h_{i}, \text{false}) \]

Find smallest i such that \( k \leq X.k_{i} \)

\[ \text{IsLeaf}(X) \rightarrow \begin{cases} \text{return } (X.h_{i+1}, \text{false}) & \text{Yes} \\ \text{return } (X.h_{i}, \text{false}) & \text{No} \end{cases} \]
InsertPair(X, k, h) entry point

504

Num(X) = 0 ?

Yes → X := (X_{left}, X_{min}, X_{h0}, k, h, X_{max}, X_{right})

No

X_{k0} < k ?

No → X := (X_{left}, X_{min}, X_{h0}, k, h, X_{k0}, ..., X_{hn}, X_{max}, X_{right})

Yes → Find largest i such that X_{ki-1} < k

512

X := (X_{left}, X_{min}, X_{h0}, ..., X_{ki-1}, X_{hi}, k, h, X_{ki}, ..., X_{hn}, X_{max}, X_{right})

514

516

return

FIGURE 5
CREATE TREE ENTRY POINT

X := (null, -inf, null, +inf, null)

h_x := Alloc(space for X)

Write(X, h_x)

P := (h_x)

h_p := Alloc(space for P)

Write(P, h_p)

return h_p

FIGURE 6
Lookup(h_p, k) entry point

(v, P) := Read(h_p)

h := last handle in P

(v, X) := Read(h)

h.v = v?

(h, isLink) := GetNextHandle(X, k)

h = null?

IsLeaf(X) and not isLink?

(v, D) := Read(h)

h.v = v?

D.k = k?

return D.d
Delete(hv, k) entry point

Yes

No

(v, P) := Read(hv)

806

h := last handle in P

808

(v, X) := Read(h)

810

h.v = v?

Yes

812

(h, isLink) := GetNextHandle(X, k)

814

IsLeaf(X) and not isLink?

Yes

return

816

h = null?

No

Lock(h)

820

(v, X) := Read(h)

822

Unlock(h)

824

h.v = v?

Yes

828

k <= X.kmax?

Yes

Unlock(h)

h := X.hmax

830

No

832

k > X.kmax?

Yes

Unlock(h)

h := X.hmin

834

No

836

838

FIGURE 8
Insert \((h_p, k, d)\) entry point

1. \(D := (k, d)\)

2. \(h_o := \text{Alloc}(\text{space for } D)\)

3. \(\text{Write}(D, h_o)\)

4. \(L_1 := 0\)

5. \(S := \text{new empty stack}\)

6. \((v, P) := \text{Read}(h_p)\)

7. \(h_x := \text{last handle in } P\)

8. \(L_n := \text{number of handles in } P\)

9. \(L_C := L_n - 1\)

10. \((v, X) := \text{Read}(h_x)\)

11. \(h_x.v = v?\)

12. \((h', \text{isLink}) := \text{GetNextHandle}(X, k)\)

13. \(\text{isLink?}\)

14. \(L_C > L_1?\)

15. \(h_x := h'\)

16. Push \(h_x\) onto \(S\)

17. \(L_C := L_C - 1\)

18. No

19. Yes

20. No

FIGURE 9A
(v, X) ::= Read(h_x)

A

No

D := (k, d) Yes

Write(D, X.h_{i+1}) No

Dealloc(h_o)

Unlock(h_x)

return

FIGURE 9B
Unlock(h_x) 935

(h', isLink) := GetNextHandle(X, k) 936

isLink ? 937

Yes

h_x := h'

(ν, X) := Read(h_x)

h_x.ν = ν ? 940

Yes

No

A

B

C

D

X_k_min < k <= X_k_max ? 934

No

Unlock(h_x)

Yes

X is full ? 941

No

InsertPair(X, k, h_o)

Write(X, h_x)

Unlock(h_x)

return

Yes

FIGURE 9C
\[ Y := (h_Y, X, k_B, \text{null}, X, k_A, \ldots, X, k_{n-1}, X, h_{iv}, X, k_{max}, X, h_{reg}) \]

\[ h_{iv} := \text{Alloc} \text{(space for } Y) \]

\[ X := (X, h_{left}, X, k_{min}, X, h_{lo}, X, k_{v}, \ldots, X, k_{t-1}, X, h_{lo}, X, k_{max}, \ldots, X, k_{t}, X, h_{hy}) \]

\[ X, k_{max} < k? \]

\[ \text{InsertPair}(X, k, h_{lo}) \]

\[ \text{Write}(Y, h_{iv}) \]

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

\[ (v, P) = \text{Read}(h_{hy}) \]

**FIGURE 9D**
Z := (null, -inf, h_x, X.k_max, h_y, +inf, null)

h_z := Alloc(space for Z)

Write(Z, h_z)

Append h_z to end of P

Write(P, h_p)

Unlock(h_x)

return

Unlock(h_x)

h_D := h_y

k := X.k_max

S empty?

No

h_x := pop handle from S

Unlock(h_x)

L_T := L_T + 1

Yes

h_x := P[L_T]

L_N := L_N + 1

FIGURE 9E
Compress\(h_p\) entry point

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

\(i := 1\)

\(i < \text{number of handles in } P?\)

Yes: CompressLevel\(P[i]\)

No: \(h := \text{last handle in } P\)

\(\text{Lock}(h)\)

\((v, X) := \text{Read}(h)\)

\(\text{Num}(X) = 0 \text{ and } X.h_{\text{left}} = \text{null and not IsLeaf}(X)?\)

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

No: \(h = \text{last handle in } P?\)

Yes: Remove last handle from \(P\)

Write\((P, h_p, )\)

Dealloc\((h)\)

Unlock\((h)\)

Pause for some time

FIGURE 10
CompressLevel(h_A) entry point

1101

a := 0

1102

h_A := null

1103

h_x = null ?

1104

Yes

return

No

1105

Lock(h_x)

1106

(v, X) := Read(h_x)

1107

h_A = null ?

1108

Yes

No

a := -1

1109

i := 0

1110

i <= Num(X) ?

1111

No

Yes

X.h_i = h_A ?

1112

Yes

a := i

1113

No

i := i + 1

1114

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.k_0, ..., A.h_n, 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_{b-2}, X.h_{b-1}, X.k_b, X.h_{b+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, h_x, h_y, h_y) entry point

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

IsLeaf(A)?

Yes

\[ A := (A._{\text{left}}, A.k_{\text{max}}, \text{null}, A.k_{\text{min}}, \ldots, A.h, B.k_{\text{left}}, B.h, B.k_{\text{right}}) \]

\[ B := (B._{\text{left}}, B.k_{\text{left}}, \text{null}, B.k_{\text{right}}, B.h, B.k_{\text{max}}) \]

No

\[ A := (A._{\text{left}}, A.k_{\text{left}}, A.h, A.k_{\text{min}}, \ldots, A.h, A.k_{\text{max}}, B.h, \ldots, B.h, B.k_{\text{left}}, B.k_{\text{right}}) \]

\[ B := (B._{\text{left}}, B.k_{\text{left}}, B.h, B.k_{\text{right}}, B.h, B.k_{\text{max}}) \]

1312

\[ X.k_{\text{left}} := A.k_{\text{max}} \]

1318

Write(A, h_x)

1320

Unlock(h_x)

1322

Write(X, h_y)

1324

Unlock(h_y)

1326

Write(B, h_y)

1328

Unlock(h_y)

1330

return

FIGURE 13
MoveToRight(X, A, B, h_x, h_y, h_y) entry point

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

IsLeaf(A)?

Yes

No

B := (B.h_left, A.k_m, null, A.k_m, ..., A.h_m, B.k_0, ..., B.h_m, B.k_max, B.h_right)

A := (A.h_left, A.k_m, A.h_m, ..., A.h_m, A.k_m, A.h_right)

X.k_a := A.k_max

Write(B, h_y)

Unlock(h_y)

Write(X, h_y)

Unlock(h_y)

Write(A, h_y)

Unlock(h_y)

return

FIGURE 14
DELETION AND COMPACTION USING VERSIONED NODES

CROSS REFERENCE TO RELATED APPLICATION

The present application is related by subject matter to application Ser. No. 10/308,293, filed on Dec. 2, 2002, entitled “Algorithm for Tree Traversals Using Left Links,” now pending.

COPYRIGHT NOTICE/PERMISSION

A portion of the disclosure of this patent document contains material that is subject to copyright protection. The copyright owner has no objection to the facsimile reproduction by anyone of the patent document or the patent disclosure, as it appears in the Patent and Trademark Office patent files or records, but otherwise reserves all copyright rights whatsoever.

FIELD OF THE INVENTION

The present invention relates generally to the field of information management systems (sometimes also called information retrieval systems) that store, organize and retrieve data, including but not limited to database management and file systems, and more particularly to novel data structures and algorithms for use in such systems. The invention is particularly suited for, but by no means limited to, use in connection with B-Tree data structures and variants thereof, and SQL-type databases.

BACKGROUND OF THE INVENTION

B-Trees are a core technology to 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 algorithms 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), handles stored in the nodes, and version numbers associated with each of the handles and nodes. In addition, the system includes a mechanism for initially assigning each node a first prescribed version number (e.g., version number 0), and a mechanism for performing an allocate operation whereby a handle identifying a node (e.g., a previously unused disk block) is returned along with the node’s version number.

The system may also include a mechanism for performing a deallocate operation that increments the node’s version number and returns the node to a pool of unused nodes, as well as a mechanism for performing a read operation that returns a node’s version number in addition to the node stored at that disk block. (In an alternative embodiment, the read operation does not return the version number of the node being read but instead raises an exception if the version number of the handle and the version number of the node do not match.) Moreover, in a preferred implementation of the first aspect of the present invention, a caller of the read operation checks whether the version number returned by the read operation matches the version number of the handle that was read. The caller is preferably restarted if the version numbers do not match, thus indicating that the node has been deleted and subsequently deallocated by a different thread. The caller may be, for example, a lookup operation, an insert operation, a delete operation, or other operation that requires a read operation to be executed.

The first aspect of the present invention may also be embodied as a B-Tree data structure stored on a computer readable medium, comprising a plurality of nodes associated with disk blocks, handles stored in the nodes, and version numbers attached to each of the handles and nodes.

Note that a “node” may span multiple disk or storage blocks, and that preferably version numbers are logically associated with nodes as opposed to individual disk blocks. The read operation returns a node and its version number. Moreover, it should be noted that an allocator may allocate, deallocate, read, and write variable length storage units.

40 number and returns the node to a pool of unused nodes, as
45 node being read but instead raises an exception if the version number of the handle and the version number of the node do not match, thus indicating that the node has been deleted and subsequently deallocated by a different thread.

The caller may be, for example, a lookup operation, an insert operation, a delete operation, or other operation that requires a read operation to be executed.

The first aspect of the present invention may also be embodied as a B-Tree data structure stored on a computer readable medium, comprising a plurality of nodes associated with disk blocks, handles stored in the nodes, and version numbers attached to each of the handles and nodes.

Note that a “node” may span multiple disk or storage blocks, and that preferably version numbers are logically associated with nodes as opposed to individual disk blocks. The read operation returns a node and its version number. Moreover, it should be noted that an allocator may allocate, deallocate, read, and write variable length storage units.
FIGS. 1 and 2 are schematic illustrations of exemplary computing environments suitable for the present invention, with FIG. 2 depicting an example of a B-Link-Tree data structure stored in a persistent store such as (but not limited to) a database.

FIG. 3A is a simplified schematic illustration of a B-Link Tree in accordance with the prior art.

FIG. 3B is a schematic illustration of a B-Link Tree in accordance with one aspect of the present invention. FIGS. 3C and 3D show examples of a Sagiv-style B-Link Tree (see discussion below) in accordance with the prior art and a B-Link Tree containing left links in accordance with the present invention, respectively.

FIGS. 3E and 3F show examples of pseudocode for lookup procedures employing links to left siblings and version numbers, respectively.

FIG. 4 is a flowchart of a procedure, GetNextHandle, which procedure takes an index node X, a key k, and a data value d and inserts the key value and the data value into the tree identified by h.

FIG. 5 is a flowchart of a procedure, InsertPair, which takes an index node X, a key k, and a handle h and inserts the key-handle pair into X.

FIG. 6 is a flowchart of a procedure, CreateTree, which returns a prime node handle h_p, identifying a new tree.

FIG. 7 is a flowchart of a procedure, Lookup, which takes a prime node handle h_p, a key k and returns a data item d or null.

FIG. 8 is a flowchart of a procedure, Delete, which takes a prime node handle h_p, a key k and deletes the key and its associated data value from the tree identified by h_p.

FIG. 9, made up of subparts 9-9E, is a flowchart of a procedure, Insert, which takes a prime node handle h_p, a key value k, and a data value d and inserts the key value and the data value into the tree identified by h_p.

FIG. 10 is a flowchart of a background procedure, Compress, that takes a prime node handle h_p and compresses the tree identified by h_p.

FIG. 11, made up of subparts 11-11C, is a flowchart of a procedure, CompressLevel, which compresses a level.

FIG. 12 is a flowchart of a procedure, MergeNodes, which merges two nodes.

FIG. 13 is a flowchart of a procedure MoveToLeft that moves parts of a node to its left sibling.

FIG. 14 is a flowchart of a procedure MoveToRight that moves parts of a node to its right sibling.

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 media, removable and non-removable media. By way of example, and not
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, magnet cassette, 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 signal" 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-removable 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 tape, solid state RAM, solid state ROM, and the like. The hard disk drive 141 is typically connected to the system bus 121 through a removable memory interface such as interface 140, and magnetic disk drive 151 and optical disk drive 155 are typically connected to the system bus 121 by a removable memory interface, such as interface 150.

The drives and their associated computer storage media, discussed above and illustrated in FIG. 1, provide storage of computer readable instructions, data structures, program modules and other data for the computer 110. In FIG. 1, for example, hard disk drive 141 is illustrated as storing operating system 144, application programs 145, other program modules 146, and program data 147. Note that these components can either be the same as or different from operating system 134, application programs 135, other program modules 136, and program data 137. Operating system 144, application programs 145, other program modules 146, and program data 147 are given different numbers here to illustrate that, at a minimum, they are different copies. A user may enter commands and information into the computer 110 through input devices such as a keyboard 162 and pointing device 161, commonly referred to as a mouse, trackball or touch pad. Other input devices (not shown) may include a microphone, joystick, game pad, satellite dish (where the dish is used to receive signals to input to a computer), scanner, or the like. These and other input devices are often connected to the processing unit 120 through a user input interface 160 that is coupled to the system bus, but may be connected by other interface and bus structures, such as a parallel port, game port or a universal serial bus (USB). A monitor 191 or other type of display device is also connected to the system bus 121 via an interface, such as a video interface 190. In addition to the monitor, computers may also include other peripheral output devices such as speakers 197 and printer 196, which may be connected through an output peripheral interface 190.

The computer 110 may also operate in a networked environment using logical connections to one or more remote computers, such as a remote computer 180. The remote computer 180 may be a personal computer, a server, a router, a network PC, a peer device or other common network node, and typically includes many or all of the elements described above relative to the computer 110, although only a memory storage device 181 has been illustrated in FIG. 1. The logical connections depicted include a local area network (LAN) 171 and a wide area network (WAN) 173, but may also include other networks. Such networking environments are commonplace in offices, enterprise-wide computer networks, storage area networks (SANs), intranets and the Internet.

When used in a LAN networking environment, 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. 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's algorithm (see Yehoshua Sagiv, Concurrent Operations on B-Trees with Overtaking, 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.

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) 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;
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 <k,d>. An index node is of the form:

<k_min, h_right, k_0, h_1, k_1, ..., h_{n-1}, k_{n-1}, h_n, k_{max}, h_right>

In the following, we refer to field x of node A as A.x. Given an index node A, A.h_0, ..., A.h_n are handles to the n+1 children of A, and A.h_right is a handle to its right sibling. A.k_i (for 0 ≤ i ≤ n) is the largest key in the subtree rooted at handle A.h_i, and A.k_{max} (or A.k_{max} if i=0) is less than the smallest key in the subtree rooted at handle A.h_i. A.k_{min} is greater or equal to the largest key in any subtree of A (and per...
definition \( k \) if \( A \) does not have a right sibling, and \( A.k_{\text{min}} \)

is equal to \( B.k_{\text{max}} \), where \( B \) is the left sibling of \( A \) (or \( \sim \) if \( A \) does not have a left sibling). Moreover, \( A.k_{\text{max}} < A.k_{\text{low}} \ldots \cdot A.k_{\text{min}} \). Finally, there is a limit on the size of \( n \) (which indicates the number of keys and handles in an index node). If \( n \) receives a maximum value (say, 2), then the index node is said to be full. Likewise, if \( n \) falls below a certain number (say, 1), 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, \( \text{lookup}(k) \) starts at the root handle, reads in the corresponding node \( A \), and identifies a value \( i \) such that \( A.k_{i-1} < k \leq A.k_{i} \) (or \( A.k_{\text{max}} \) if \( k \) is equal to \( A.k_{\text{max}} \)). Then it recursively proceeds along the handle \( A.h \), until it reaches a data node \( D \), and returns the value \( D \)’s value if \( D \)’s key is indeed \( k \), or null otherwise.

The delete operation is similar to the lookup operation: \( \text{delete}(k) \) descends into the tree until a data node \( D \) with key \( k \) is discovered (if such a node exists). The operation then marks \( D \) as deleted (\( D \) is not immediately deallocated, because other ongoing operations may have a handle to \( D \) but not yet have read \( D \)), and removes the handle to \( D \) from \( D \)’s parent node \( A \). This may cause \( A \) to become underfull.

The insert operations are more complicated: \( \text{insert}(k,v) \) allocates a new data node \( D \) with handle \( h \), writes the key-handle pair \( (k,v) \) to it, and then recursively descends into the tree the same way as lookup does, until it finds the leaf index node \( A \) (the index node whose children are data nodes) that should receive \( h \). If \( A \) is not full, \( \text{insert}(k,v) \) simply inserts \( h \) and \( k \) at the appropriate places into \( A \); otherwise, it allocates a new index node \( \Delta \), moves half of \( A \)’s key-handle pairs over to \( \Delta \), inserts \( k \) and \( h \) into \( \Delta \) or \( A \), and finally adds the handle to \( \Delta \) and \( A \)’s new \( k_{\text{max}} \) to \( A \)’s parent (this may in turn cause \( A \)’s parent to become overfull, causing the node splitting process to move up the tree).

As mentioned above, the delete operation may cause nodes to become underfull. To prevent too many nodes from becoming underfull (which would cause the tree to become deeper than it needs to be, which would increase the number of disk accesses required by each operation), a compression thread is run in the background. The thread repeatedly traverses the tree, searching for underfull nodes. When it locates an underfull node \( A \), it either rebalances it with its left or right sibling (which entails moving key-handle pairs from the sibling to \( A \), and adjusting a key in the parent node), or it outright merges \( A \) with its left or right sibling (which entails moving all of \( A \)’s content to the sibling, marking \( A \) as deleted, and removing \( A \)’s handle and corresponding key from \( A \)’s parent, which in turn may cause the parent to become underfull). In Sagiv’s algorithm, nodes that are marked as deleted are deallocated only when it is certain that no operations have handles to those nodes. This is achieved by adding a timestamp to each deleted node, indicating the time it was deleted, and maintaining a table that records the start time of every lookup, insert, and delete operation in flight. The system maintains a list of deleted nodes. The compression thread periodically goes through the list, and deletes exactly those nodes whose timestamp predates the start times of all ongoing operations. The drawback of this scheme is that the table of start times is accessed by many threads, which means it has to be protected by a lock. Each operation requires two lock acquisitions and releases (one for adding a thread-start time pair to the table, and one for removing it).

The situation is further complicated by the fact that concurrent lookup, insert, and delete operations may interfere with one another. This could be avoided by locking every node visited during an operation, and unlocking it only once the child or sibling referenced by any extracted handle has been read (this is called “lock coupling” in the literature). Sagiv’s algorithm avoids such excessive locking by using the optimistic assumption that operations usually don’t interfere, detecting the cases in which this assumption was unwarranted, and recovering from the misstep.

To make things concrete, consider the case where the operation \( \text{lookup}(k) \) decides to follow a handle \( h \), but before it can read the corresponding node \( A \), another thread inserts a key-handle pair into \( A \), and in the process splits \( A \) into two nodes, \( A \) and \( \Delta \) (the newly allocated right sibling of \( A \)). After reading \( A \), the lookup operation can detect the fact that \( A \) was split (because \( k > A.k_{\text{max}} \)), and can recover from this situation by following the \( A.h_{\text{right}} \) handle, leading it to \( \Delta \) (which indeed contains the appropriate range of keys). However, this inexpensive recovery is not always possible. Consider the case where the operation \( \text{lookup}(k) \) decides to follow a handle \( h \), but before it can read the corresponding node \( A \), the compression thread rebalances node \( A \) (which happens to be underfull), moving some of \( A \)’s content, including the handle that \( \text{lookup} \) should follow, to \( A \)’s left sibling \( \Delta \). After reading \( A \), the lookup operation can detect that \( A \) has been rebalanced (because \( k < A.k_{\text{min}} \)), but since lookup does not know the handle of \( \Delta \), the only way to recover is to restart the lookup operation at the root of the tree.

D. Maintaining Handles to Left Siblings

According to first aspect of the present invention, to avoid restarts, we propose to modify the structure of index nodes as follows:

\[
\text{<left, known: } h_0, k_0, h_1, k_1, \ldots, h_{n-1}, k_{n-1}, h_n, k_{\text{max}}, h_{\text{right}}>\]

where \( h_{\text{left}} \) is the handle of the left sibling (if there is one, and null otherwise), and all other fields are defined as in Sagiv’s algorithm. This data structure is depicted in FIG. 3B, where reference 300 refers to the modified B-Link data structure and reference numeral 310 refers specifically to the left link, i.e., \( h_{\text{left}} \). As in Sagiv’s algorithm, 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. As in Sagiv’s algorithm, the operation reacts to \( k > A.k_{\text{max}} \) by following the handle \( A.h_{\text{right}} \) to the right sibling. However, while Sagiv’s algorithm reacted to \( k < A.k_{\text{min}} \) by restarting the operation at the root, our algorithm reacts by following the handle \( A.h_{\text{left}} \) to the left sibling. Following the handle to the left sibling requires one extra read operation, while restarting the operation at the root would require up to (and typically close to) \( d \) read operations, where \( d \) is the depth of the tree.

FIGS. 3C and 3D provide more detailed examples of a Sagiv-style B-Link Tree in accordance with the prior art, and a B-Link tree in accordance with the present invention,
respectively. These are intended to further illustrate the left-link aspect of the present invention, and to distinguish it from the prior art.

FIG. 3C depicts a Sagiv-style B-Link tree that contains the mapping \{(2, Miller), (5, Smith), (7, Jones), (10, Brown), (12, Levin), (15, Dahl), (17, Lewis), (20, Yu)\}. The numbers are key values and the names are the associated data values. P is the prime node of the tree, 11, \ldots , 17 are index nodes; and D1, \ldots , 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. P 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 don’t have index-node children). This means that \(11.h_0\) is null, and that IsLeaf(14) is true (the same is true for 15, 16, and 17). In this tree, \(t=2\); that is, each index node can hold two keys in addition to \(k_{\text{min}}\) and \(k_{\text{max}}\). 11, 12, and 13 are not full (some fields are unused and shaded out). Num(11) is 1 (same 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).

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 Overtaking, 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):

```
procedure Lookup(hp: Handle, k: Key)
  var h: Handle, A: Node;
  label start:
  p := Read(hp)
  h := last item in p
  repeat
    A := read(h)
    if k > Ak" then
      h := Ahright
    else if k < Ak" then
      goto start
    else
      choose i such that \(Ak'_i < k < Ak'_{i+1}\) (let i be 0 if Num(A) = 0)
      h := Ah_i
    endif
  until A is a leaf and Ak" < k < Ak"", x
  if k is contained in A then
    return corresponding value
  else
    return null
  endif
end proc
```

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 \(k_{\text{min}}\) field of node \(A\), and if so it assigns the left-link handle to the handle variable (i.e., \(h := A.h_{left}\)).
procedure Lookup(hp: Handle, k : Key)
var h: Handle, A: Node;
P := Read(hp)
h := last item in P
repeat
A := read(h)
if k > A.kmax then
    h := A.hright
else if k ≤ A.kmin then
    h := A.hlstart
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
endif
end proc

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 (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 threads 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 deallocating 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 referring to a table mapping thread identifiers to operation start times:

procedure Lookup(hp: Handle, k : Key)
var h: Handle, A: Node;

E. Attach Version Numbers to Handles and Disk Blocks

procedure Lookup(hp: Handle, k : Key)
var h: Handle, A: Node;

procedure Lookup(hp: Handle, k : Key)
var h: Handle, A: Node;

procedure Lookup(hp: Handle, k : Key)
var h: Handle, A: Node;
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(hp: Handle, k : Key)
var h: Handle, A: Node;
label start:
P := Read(hp)
h := last handle in P
repeat
    A := read(h)
    if version(A) > version(h) then goto start
    if k > Akmin then
        h := Ahright
    else if k <= Akmin then
goto start
    else
        choose i such that Ak-1 < k <= Ak, (let i be 0 if Num(A) = 0)
        h := Ai
    endif
    until A is a leaf and Ak = in
if k is contained in A then
    return corresponding value
else
    return null
endif
end proc
```

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 go to 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>Composite Types</th>
</tr>
</thead>
<tbody>
<tr>
<td>k</td>
<td>a data node</td>
</tr>
<tr>
<td>d</td>
<td>a key value</td>
</tr>
<tr>
<td>ba</td>
<td>a block address</td>
</tr>
<tr>
<td>v</td>
<td>a version number</td>
</tr>
<tr>
<td>h</td>
<td>a handle (may be null)</td>
</tr>
<tr>
<td>D</td>
<td>(h, ba)</td>
</tr>
<tr>
<td>X, A, B</td>
<td>X, A, B = (hmin, hmax, h0)</td>
</tr>
<tr>
<td></td>
<td>h0, h1, k1, ...</td>
</tr>
<tr>
<td></td>
<td>hmin, hmax, h0</td>
</tr>
<tr>
<td></td>
<td>P = (h0, ... , hn) a prime node</td>
</tr>
</tbody>
</table>
The procedure at step 418 returns \((X, h)\) (the leftmost leaf-level index node in the tree) if \(k\) is less than or equal to \(X.k_{min}\), the key \(k\) and its associated data value, and 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 \(X.h\) refers to handle \(h\) of prime node \(P\) (where \(h\) refers to the leftmost leaf-level index node in the tree).

The procedure determines whether \(\text{Num}(X)\) is equal to 0. If so, the procedure tests whether IsLeaf \((X)\) is true. If so, it returns \((X, h)\) where \(h\) is null. If not, the procedure proceeds to step 424, where it finds the largest \(i\) such that \(X.k_i = k\) and \(X.h_i = h\). If \(i\) is less than \(X.k_{min}\), the key \(k\) and handle \(h\) are inserted into the node \(X\) as shown. If \(\text{Num}(X)\) is not equal to 0, at step 408 the procedure determines whether \(X.k_{min}\) 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 \(X.k_{min}\) is less than \(k\), it proceeds to step 512 where it finds the largest \(i\) such that \(X.k_{i-1}\) 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 \(\text{Num}(X)\) is equal to 0, and whether \(X.k_{min}\) is less than \(k\).

FIG. 4 is a flowchart of a procedure entitled GetNextHandle. This procedure begins at the entry point GetNextHandle \((X, k)\) 402. At step 404, the procedure determines whether \(X.k_{min}\), the key \(k\) and handle \(h\) are inserted into the node \(X\) as shown. On the other hand, if the procedure determines that \(X.k_{min}\) is less than \(k\), it proceeds to step 512 where it finds the largest \(i\) such that \(X.k_{i-1}\) 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 \(\text{Num}(X)\) is equal to 0, and whether \(X.k_{min}\) is less than \(k\).

FIG. 5 is a flowchart of a procedure entitled InsertPair, which takes a prime node handle \(h\) and a key \(k\) and returns a data item \(d\) or null. The InsertPair procedure 500 begins at the entry point InsertPair \((X, k, h)\) (step 502). At step 504, the procedure determines whether \(\text{Num}(X)\) is equal to 0, and if not, the procedure loops back as shown to step 704. If, on the other hand, IsLeaf \((X)\) is not true, the procedure proceeds to step 724. At step 724 the procedure determines whether the key \(k\) is equal to \(X.k_{min}\) and returns a handle \(h\) to the block. If not, the procedure proceeds to step 720 where the function Read \((h)\) is invoked in order to read the block referenced by the handle \(h\). 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 GetNextHandle \((X, k)\) (step 716) where it determines whether both IsLeaf \((X)\) and not isLink are both true, the procedure proceeds to step 714. Proceeds to 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 IsLeaf indicating whether the handle \(h\) is a link to a sibling of node \(X\). Next, at step 714, the procedure sends the block \(P\) to the block referenced by the handle \(h\). This time, the function returns the version number \(v\) and the index node \(X\) stored in the block referenced by handle \(h\). As shown in FIG. 6, the procedure loops back to step 704 as shown, and so forth, until the entire tree is built.
FIG. 8 is a flowchart 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 invocation of Read()). If not, the procedure loops back to step $804$ as shown; and if so, it proceeds to step $812$. At step $812$ the function GetNextHandle (discussed above in connection with FIG. 4) is invoked with the arguments $X$, $k$. In this manner, the procedure determines the handle $h$ and the boolean isLink indicating whether the handle $h$ is a link to a sibling of node $X$. Next, at step $814$, the procedure tests whether IsLeaf(X) and not isLink are both true; if so, 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 $828$ (see subpart $8B$ of FIG. 4) is invoked with the arguments $X$, $k$. As the algorithm unfolds, $L_c$ is initialized to this value, and is decremented as insert descends into the tree. At step $832$ the Read($h$) is invoked to return the version number $v$ and the index node $X$ referred to by handle $h$. At step $834$, the procedure determines that $k$ is not greater than $X.k_{max}$, it proceeds to step $840$ where it tests whether Num($X$) is equal to 0, meaning that $X$ (which is a leaf-level node) has no children, and that it therefore does not contain the key $k$. If the answer to this test is yes, the procedure moves ahead to step $856$, and if so, it proceeds to step $842$. At step $842$, the procedure determines whether $X.k_{i+1}$ is less than $k$, i.e., whether the field $k_{i+1}$ 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 $X.k_i$. Next, at step $846$, the procedure invokes Read($X.k_{i+1}$), 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 $k$ 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 $X.h_{a+1}$, which increments the block version number. Next, at step $852$ the procedure removes the handles identified by $X.k_i$ and $X.h_{a+1}$ 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 $h_p$ and a key $k$ and removes the key $k$ and its associated data value from the tree identified by $h_p$.
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 h. At step 921 the procedure invokes Read(h), which reads the block identified by handle h, 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.kmax 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+i) is invoked in order to write the node D to the block referenced by the handle h+i. Next, at step 931, the function Dealloc(h+i) is invoked in order to deallocate the block referred to by the handle h+i and increment the block’s version number. At step 932, the block referred to by the handle h+i 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.kmax and less than or equal to X.kmax. If the answer to this 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 is unlocked, and then at step 936 the GetNextHandle function is invoked with the arguments X and k, which thereby returns the handle h of 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 is assigned the value of h’ and then at step 939 Read(h) is invoked to determine the version number v and node X referred to by the handle h. At step 940, the procedure tests whether h.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.kmax and less than or equal to X.kmax. 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, h, h), in order to insert the key k and handle h in node X. At step 943, Write(X, h) is invoked in order to write the node X to the block referenced by the handle h. At step 944, the block referred to by the handle h 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.kmax 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 to insert k and h into X. In step 955, Write(Y, h) is invoked in order to write the node Y. Similarly, at step 957 Write(X, h) is invoked in order to write the node X. If at step 950 it is determined that X.kmax is less than k, the procedure proceeds to step 956 where the function InsertPair(Y, k, h) is invoked to insert k and h into Y, after which steps 955 and 957 are performed as described above. At step 958, Read(h) is invoked and this function returns the version number v and the prime node P in the block referred to by the handle h.

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 is the last handle in node P, meaning that the node identified by h, 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 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 is returned by the function alloc ( ). At step 962 the function Write(Z, h) is invoked in order to write the index node Z to the block referenced by the handle h. At step 963, the procedure appends h to the end of prime node P. At step 964, Write(P, h) is invoked to write the node P to the block referenced by the handle h. Next, at step 965 the block referred to by the handle h is unlocked and then at step 966 the procedure returns. If at step 959 the answer to the test whether h is the last handle in P is no, the procedure proceeds to step 967, wherein the block referred to by the handle h is unlocked. Then, at step 968, the handle h, is set to h. At step 969 the key value k is set to X.kmax. At step 970, the procedure determines whether the stack S is empty; if not, it proceeds to step 971 wherein the handle h 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 is defined with P[L-1], and then at step 974 L is defined as L+1. Finally, at step 972 L is set to L+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, a key value k, and a data value d and inserts the key-data pair into the tree identified by h.
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 than 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_{rght} is equal to null and not IsLeaf(X) are all true. If so, the procedure proceeds to step 1022 and invokes Read(h), 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_{rght}.

The CompressLevel procedure 1100 begins at an entry point 1101 with the call CompressLevel(h_{sght}). At step 1102, the variable a is set to 0 and at step 1103, the handle h_{sght} is set to null. Next, at step 1104 the procedure tests whether the handle h_{sght} is null, and if so, proceeds to step 1105 where the procedure returns. If h_{sght} is not null, the procedure proceeds to step 1106 and locks the block referred to by handle h_{sght}. Next, at step 1107, Read(h_{sght}) is invoked to read index node X and its version number v. At step 1108, the procedure tests whether handle h_{rght} is null. If h_{rght} is null, the procedure branches to block 115 (subpart 11B of FIG. 11), and if h_{rght} 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 h_{sght} is equal to handle h_{sght} to the ith child of node X (i.e., X.h_{k max} is equal to handle h_{sght} to the ith child of node X). If it is not, the procedure proceeds to block 1114 and increments index i and then proceeds back to step 1111. If X.h_{k max} is equal to handle h_{sght} to the block 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_{rght} 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_{sght} set to the field h_{rght} of index node X, and then to set handle h_{rght} 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_{rght} is null. If h_{rght} is null, the procedure at step 1120 sets variable a to 0. If handle h_{rght} is not null, the procedure at step 1122 increments a. Then, at step 1121 handle h_{rght} is set to X.h_{rght}. At step 1123, the block referred to by handle h_{rght} is locked, and then at step 1124 the function Read(h_{rght}) is invoked, to read index node A and its associated version number v. At step 1125, the handle h_{rght} is set to A.h_{rght}. Next, at step 1126 the procedure tests whether the handle h_{rght} is null. If it is, steps 1131 through 1133 are performed, i.e., the block referred to by handle h_{rght} is unlocked, the block referred to by handle h_{rght} is unlocked, and then the procedure returns. On the other hand, if handle h_{rght} is not null, the procedure proceeds to step 1127, and locks the block referred to by handle h_{rght}. Next, at step 1128, Read(h_{rght}) is invoked to determine the version number v and node B. At step 1129, B.h_{rght} is set to the value of handle h_{rght}. Next, at step 1130, 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_{sght} is h_{sght}. 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 step 1134 as shown. If not, the procedure proceeds to step 1137 and unlocks the block referred to by handle h_{rght}. Next, at step 1138, the procedure unlocks the block referred to by handle h_{sght}, and at step 1139 unlocks the block referred to by handle h_{sght}. Next, at step 1140, the procedure tests whether B.h_{rght} is null or the variable a is not null, the procedure sets at step 1141 and sets handle h_{sght} 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 b to 0.

FIG. 12 is a flowchart of a procedure entitled MergeNodes. This procedure, as its name implies, merges nodes; it takes index nodes A, B, (where A is the left sibling of B and X is their parent) and index node handles h_{sght}, h_{rght}, and h_{rght} 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_{sght}, h_{rght}, h_{rght}). 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_{sght} is null. If IsLeaf(A) is true, the procedure performs steps 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_{sght}, h_{rght}, and h_{rght}, and to deallocate index node B identified by handle h_{rght}. 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 is performed 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_{sght}, h_{rght}, h_{rght}). At step 1304 the variable m is
the underlying concepts may be applied to any computing 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 emphasized 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 data store operatively connected to said computer, wherein said data store comprises a B-Tree data structure comprising a plurality of nodes associated with disk blocks, handles stored in said nodes, and version numbers attached to each of said handles and disk blocks; and
   a software-based mechanism for initially assigning each disk block a first prescribed version number, and a software-based mechanism for performing an allocate operation whereby a handle identifying a disk block is returned along with the disk block’s version number.

2. An information management system as recited in claim 1, further comprising a software-based mechanism for performing a deallocate operation that increments the disk block’s version number.

3. An information management system as recited in claim 1, further comprising a software-based mechanism for performing a read operation that returns a disk block’s version number in addition to the node stored at that disk block.

4. An information management system as recited in claim 1, further comprising a software-based mechanism for performing a read operation that does not return the version number of the disk block being read but instead raises an exception if the version number of the handle and the version number of the disk block do not match.

5. An information management system as recited in claim 1, wherein a caller of the read operation checks whether the version number returned by the read operation matches the version number of the handle that was read.

6. An information management system as recited in claim 1, wherein the caller is restarted if the version numbers do not match, thus indicating that the node has been deleted and subsequently deallocated.
7. An information management system as recited in claim 5, wherein the caller comprises a lookup operation.
8. An information management system as recited in claim 5, wherein the caller comprises an insert operation.
9. An information management system as recited in claim 5, wherein the caller comprises a delete operation.
10. An information management system, comprising:
a computer; and
a data store operatively connected to said computer, wherein said data store comprises a B-Tree data structure comprising a plurality of nodes associated with disk blocks, handles stored in said nodes, version numbers attached to each of said handles and disk blocks, wherein at least one left-link handle, $h_{left}$, stored in each node points to a left sibling of that node.
11. An information management system as recited in claim 10, further comprising a software-based mechanism for initially assigning each disk block a first prescribed version number, and a software-based mechanism for performing an allocate operation whereby a handle identifying a disk block is returned along with the disk block’s version number.
12. An information management system as recited in claim 10, further comprising a software-based mechanism for performing a deallocate operation that increments the disk block’s version number.
13. An information management system as recited in claim 18, wherein a caller of the read operation checks whether the version number returned by the read operation matches the version number of the handle that was read.
14. An information management system as recited in claim 13, wherein the caller is restarted if the version numbers do not match, thus indicating that the node has been deleted and subsequently deallocated by a different thread.
15. An information management system as recited in claim 13, wherein the caller comprises a lookup operation.
16. An information management system as recited in claim 13, wherein the caller comprises an insert operation.
17. An information management system as recited in claim 13, wherein the caller comprises a delete operation.
18. An information management system as recited in claim 10, further comprising a software-based mechanism for performing a read operation that returns a disk block’s version number in addition to the node stored at that disk block.
19. An information management system as recited in claim 10, further comprising a software-based mechanism for performing a read operation that does not return the version number of the disk block being read but instead raises an exception if the version number of the handle and the version number of the disk block do not match.

* * * * *