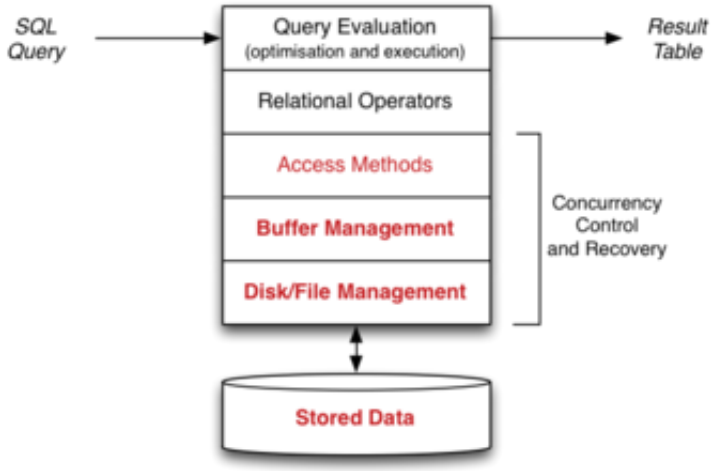
Week 02 Lectures

Storage Manager

Storage Management

2/68

Levels of DBMS related to storage management:



... Storage Management

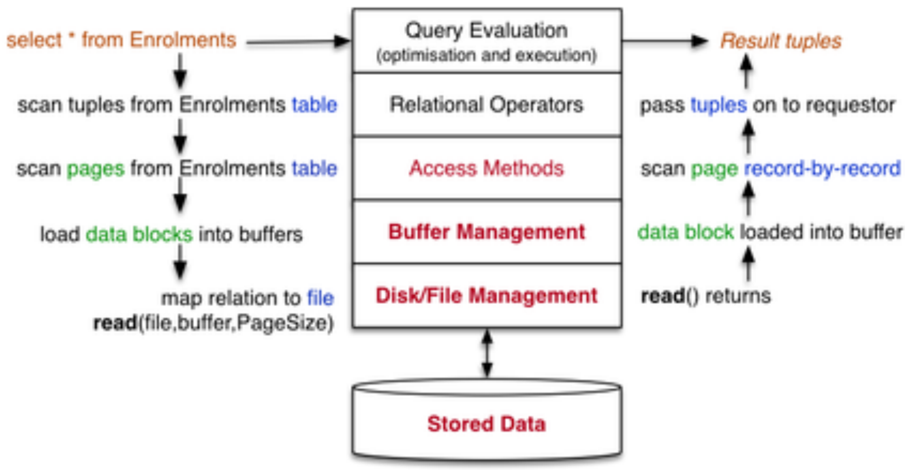
3/68

Aims of storage management in DBMS:

- provide view of data as collection of pages/tuples
- map from database objects (e.g. tables) to disk files
- manage transfer of data to/from disk storage
- use buffers to minimise disk/memory transfers
- interpret loaded data as tuples/records
- basis for file structures used by access methods

Views of Data in Query Evaluation

4/68



... Views of Data in Query Evaluation

5/68

Representing database objects during query execution:

- `DB` (handle on an authorised/opened database)
- `Rel` (handle on an opened relation)
- `Page` (memory buffer to hold contents of disk block)
- `Tuple` (memory holding data values from one tuple)

Addressing in DBMSs:

- `PageID = FileID+Offset` ... identifies a block of data
    - where `Offset` gives location of block within file
- `TupleID = PageID+Index` ... identifies a single tuple
    - where `Index` gives location of tuple within page

Storage Management

6/68

Topics in storage management ...

- Disks and Files
    - performance issues and organisation of disk files
- Buffer Management
    - using caching to improve DBMS system throughput
- Tuple/Page Management
    - how tuples are represented within disk pages
- DB Object Management (Catalog)
    - how tables/views/functions/types, etc. are represented

Storage Technology

Storage Technology

8/68

Persistent storage is

- large, cheap, relatively slow, accessed in blocks
- used for long-term storage of data

Computational storage is

- small, expensive, fast, accessed by byte/word
- used for all analysis of data

Access cost HDD:RAM ≅ 100000:1, e.g.

- 10ms to read block containing two tuples
- 1μs to compare fields in two tuples

... Storage Technology

9/68

Hard disks are well-established, cheap, high-volume, ...

Alternative bulk storage: SSD

- faster than HDDs, no latency
- can read single items
- update requires block erase then write
- over time, writes "wear out" blocks
- require controllers that spread write load

Feasible for long-term, high-update environments?

... Storage Technology

10/68

Comparison of HDD and SSD properties:

	HDD	SDD
Cost/byte	~ 4c / GB	~ 13c / GB
Read latency	~ 10ms	~ 50μs
Write latency	~ 10ms	~ 900μs
Read unit	block (e.g. 1KB)	byte
Writing	write a block	write on empty block

Will SSDs ever replace HDDs?

Cost Models

11/68

Throughout this course, we compare costs of DB operations

Important aspects in determining cost:

- data is always transferred to/from disk as whole blocks (pages)
- cost of manipulating tuples in memory is negligible
- overall cost determined primarily by #data-blocks read/written

Complicating factors in determining costs:

- not all page accesses require disk access (buffer pool)
- tuples typically have variable size (tuples/page ?)

More details later ...

File Management

12/68

Aims of file management subsystem:

- organise layout of data within the filesystem
- handle mapping from database ID to file address
- transfer blocks of data between buffer pool and filesystem
- also attempts to handle file access error problems (retry)

Builds higher-level operations on top of OS file operations.

... File Management

13/68

Typical file operations provided by the operating system:

```
fd = open(fileName,mode)
  // open a named file for reading/writing/appending
close(fd)
  // close an open file, via its descriptor
nread = read(fd, buf, nbytes)
  // attempt to read data from file into buffer
nwritten = write(fd, buf, nbytes)
  // attempt to write data from buffer to file
lseek(fd, offset, seek_type)
  // move file pointer to relative/absolute file offset
fsync(fd)
  // flush contents of file buffers to disk
```

DBMS File Organisation

14/68

How is data for DB objects arranged in the file system?

Different DBMSs make different choices, e.g.

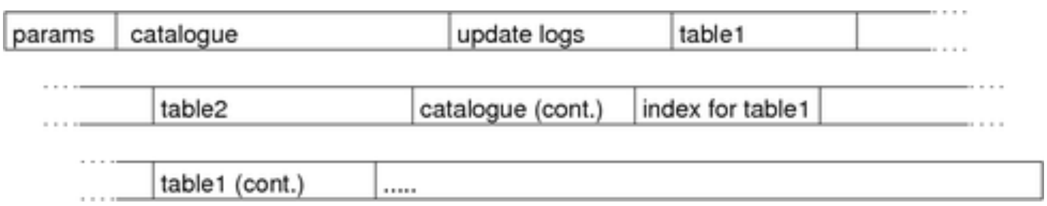
- by-pass the file system and use a raw disk partition
- have a single very large file containing all DB data
- have several large files, with tables spread across them
- have multiple data files, one for each table
- have multiple files for each table
- etc.

Single-file DBMS

15/68

Consider a single file for the entire database (e.g. SQLite)

Objects are allocated to regions (segments) of the file.



If an object grows too large for allocated segment, allocate an extension.

What happens to allocated space when objects are removed?

... Single-file DBMS

16/68

Allocating space in Unix files is easy:

- simply seek to the place you want and write the data
- if nothing there already, data is appended to the file
- if something there already, it gets overwritten

If the seek goes way beyond the end of the file:

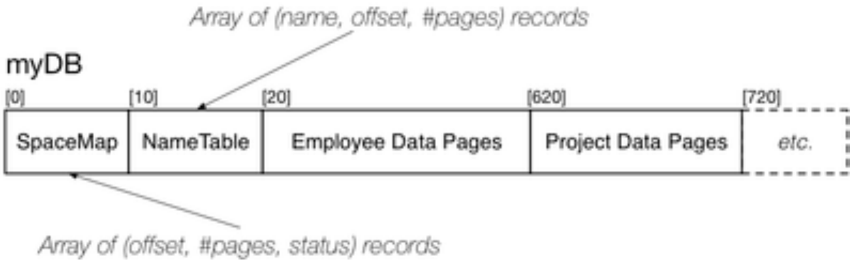
- Unix does not (yet) allocate disk space for the "hole"
- allocates disk storage only when data is written there

With the above, a disk/file manager is easy to implement.

Single-file Storage Manager

17/68

Consider the following simple single-file DBMS layout:



E.g.

SpaceMap = [(0,10,U), (10,10,U), (20,600,U), (620,100,U), (720,20,F)]

TableMap = [("employee",20,500), ("project",620,40)]

... Single-file Storage Manager

18/68

Each file segment consists of a number fixed-size blocks

The following data/constant definitions are useful

```
#define PAGESIZE 2048    // bytes per page

typedef long PageId;     // PageId is block index
                         // pageOffset=PageId*PAGESIZE

typedef char *Page;      // pointer to page/block buffer
```

Typical `PAGESIZE` values: 1024, 2048, 4096, 8192

... Single-file Storage Manager

19/68

Storage Manager data structures for opened DBs & Tables

```
typedef struct DBrec {
   char *dbname;     // copy of database name
   int fd;           // the database file
   SpaceMap map;     // map of free/used areas
   NameTable names; // map names to areas + sizes
} *DB;

typedef struct Relrec {
   char *relname;    // copy of table name
   int   start;      // page index of start of table data
   int   npages;     // number of pages of table data
   ...
} *Rel;
```

Example: Scanning a Relation

20/68

With the above disk manager, our example:

```
select name from Employee
```

might be implemented as something like

```
DB db = openDatabase("myDB");
Rel r = openRelation(db,"Employee");
Page buffer = malloc(PAGESIZE*sizeof(char));
for (int i = 0; i < r->npages; i++) {
   PageId pid = r->start+i;
   get_page(db, pid, buffer);
   for each tuple in buffer {
      get tuple data and extract name
      add (name) to result tuples
   }
}
```

Single-File Storage Manager

21/68

```
// start using DB, buffer meta-data
DB openDatabase(char *name) {
   DB db = new(struct DBrec);
   db->dbname = strdup(name);
   db->fd = open(name,O_RDWR);
   db->map = readSpaceTable(db->fd);
   db->names = readNameTable(db->fd);
   return db;
}
// stop using DB and update all meta-data
void closeDatabase(DB db) {
   writeSpaceTable(db->fd,db->map);
   writeNameTable(db->fd,db->map);
   fsync(db->fd);
   close(db->fd);
   free(db->dbname);
   free(db);
}
```

... Single-File Storage Manager

22/68

```
// set up struct describing relation
Rel openRelation(DB db, char *rname) {
   Rel r = new(struct Relrec);
   r->relname = strdup(rname);
   // get relation data from map tables
   r->start = ...;
   r->npages = ...;
   return r;
}

// stop using a relation
void closeRelation(Rel r) {
   free(r->relname);
   free(r);
}
```

... Single-File Storage Manager

23/68

```
// assume that Page = byte[PageSize]
// assume that PageId = block number in file

// read page from file into memory buffer
void get_page(DB db, PageId p, Page buf) {
```

```
   lseek(db->fd, p*PAGESIZE, SEEK_SET);
   read(db->fd, buf, PAGESIZE);
}

// write page from memory buffer to file
void put_page(Db db, PageId p, Page buf) {
   lseek(db->fd, p*PAGESIZE, SEEK_SET);
   write(db->fd, buf, PAGESIZE);
}
```

... Single-File Storage Manager

24/68

Managing contents of space mapping table can be complex:

```
// assume an array of (offset,length,status) records

// allocate n new pages
PageId allocate_pages(int n) {
   if (no existing free chunks are large enough) {
      int endfile = lseek(db->fd, 0, SEEK_END);
      addNewEntry(db->map, endfile, n);
   } else {
      grab "worst fit" chunk
      split off unused section as new chunk
   }
   // note that file itself is not changed
}
```

... Single-File Storage Manager

25/68

Similar complexity for freeing chunks

```
// drop n pages starting from p
void deallocate_pages(PageId p, int n) {
   if (no adjacent free chunks) {
      markUnused(db->map, p, n);
   } else {
      merge adjacent free chunks
      compress mapping table
   }
   // note that file itself is not changed

}
```

Changes take effect when `closeDatabase()` executed.

Exercise 1: Relation Scan Cost

26/68

Consider a table *R(x,y,z)* with 10^5 tuples, implemented as

- number of tuples $r = 10{,}000$
- average size of tuples $R = 200$ bytes
- size of data pages $B = 4096$ bytes
- time to read one data page $T_r = 10msec$
- time to check one tuple *1 usec*
- time to form one result tuple *1 usec*
- time to write one result page $T_r = 10msec$

Calculate the total time-cost for answering the query:

```
insert into S select * from R where x > 10;
```

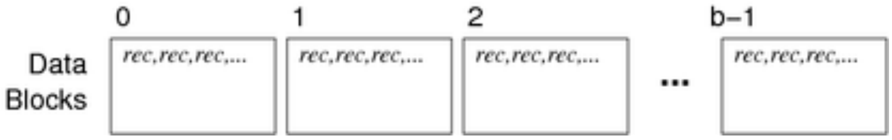
if 50% of the tuples satisfy the condition.

DBMS Parameters

27/68

Our view of relations in DBMSs:

- a relation is a set of r tuples, with average size R bytes
- the tuples are stored in b data pages on disk
- each page has size B bytes and contains up to c tuples
- data is transferred disk↔memory in whole pages
- cost of disk↔memory transfer T_r , T_w dominates other costs



... DBMS Parameters

28/68

Typical DBMS/table parameter values:

Quantity	Symbol	E.g. Value
total # tuples	r	10^6
record size	R	128 bytes
total # pages	b	10^5
page size	B	8192 bytes
# tuples per page	c	60
page read/write time	T_r , T_w	10 msec
cost to process one page in memory	-	$\cong 0$

Multiple-file Disk Manager

29/68

Most DBMSs don't use a single large file for all data.

They typically provide:

- multiple files partitioned physically or logically
- mapping from DB-level objects to files (e.g. via meta-data)

Precise file structure varies between individual DBMSs.

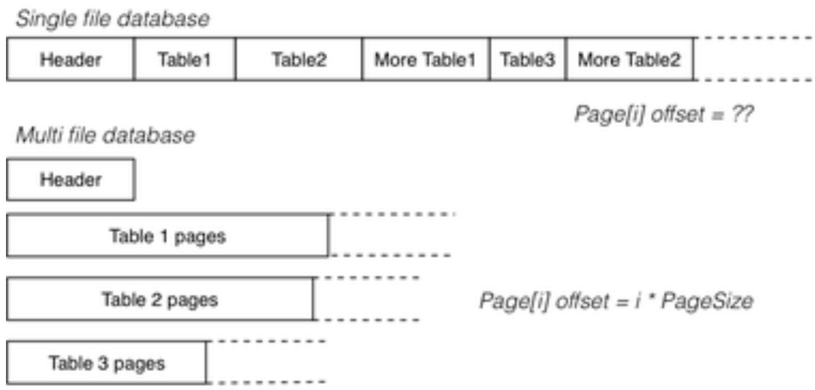
Using multiple files (one file per relation) can be easier, e.g.

- adding a new relation
- extending the size of a relation
- computing page offsets within a relation

... Multiple-file Disk Manager

30/68

Example of single-file vs multiple-file:



Consider how you would compute file offset of page[i] in table[1] ...

... Multiple-file Disk Manager

31/68

Structure of `PageId` for data pages in such systems ...

If system uses one file per table, `PageId` contains:

- relation indentifier (which can be mapped to filename)
- page number (to identify page within the file)

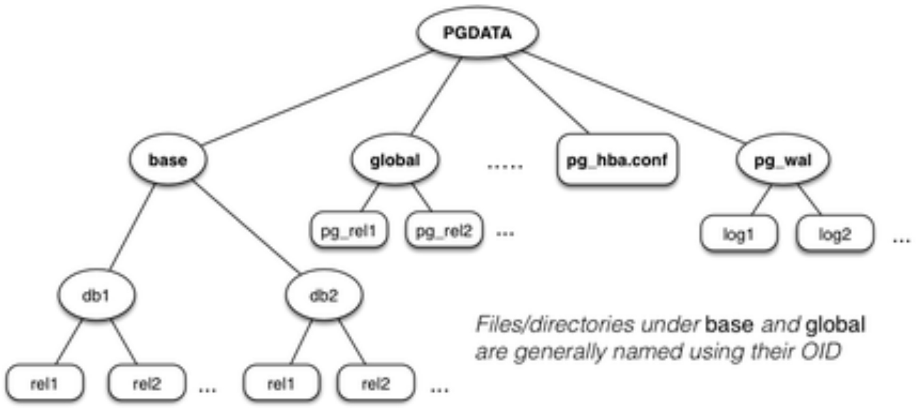
If system uses several files per table, `PageId` contains:

- relation identifier
- file identifier (combined with relid, gives filename)
- page number (to identify page within the file)

PostgreSQL Storage Manager

32/68

PostgreSQL uses the following file organisation ...



... PostgreSQL Storage Manager

33/68

Components of storage subsystem:

- mapping from relations to files (`RelFileNode`)
- abstraction for open relation pool (`storage/smgr`)
- functions for managing files (`storage/smgr/md.c`)
- file-descriptor pool (`storage/file`)

PostgreSQL has two basic kinds of files:

- heap files containing data (tuples)
- index files containing index entries

Note: `smgr` designed for many storage devices; only disk handler provided

Relations as Files

34/68

PostgreSQL identifies relation files via their OIDs.

The core data structure for this is **RelFileNode**:

```
typedef struct RelFileNode {
    Oid  spcNode;  // tablespace
    Oid  dbNode;   // database
    Oid  relNode;  // relation
} RelFileNode;
```

Global (shared) tables (e.g. `pg_database`) have

- `spcNode == GLOBALTABLESPACE_OID`
- `dbNode == 0`

... Relations as Files

35/68

The **relpath** function maps **RelFileNode** to file:

```
char *relpath(RelFileNode r)  // simplified
{
   char *path = malloc(ENOUGH_SPACE);

   if (r.spcNode == GLOBALTABLESPACE_OID) {
      /* Shared system relations live in PGDATA/global */
      Assert(r.dbNode == 0);
      sprintf(path, "%s/global/%u",
              DataDir, r.relNode);
   }
   else if (r.spcNode == DEFAULTTABLESPACE_OID) {
      /* The default tablespace is PGDATA/base */
      sprintf(path, "%s/base/%u/%u",
              DataDir, r.dbNode, r.relNode);
   }
   else {
      /* All other tablespaces accessed via symlinks */
      sprintf(path, "%s/pg_tblspc/%u/%u/%u", DataDir
              r.spcNode, r.dbNode, r.relNode);
   }
   return path;
}
```

Exercise 2: PostgreSQL Files

36/68

In your PostgreSQL server

- examine the content of the `$PGDATA` directory
- find the directory containing the `pizza` database
- find the file in this directory for the `People` table
- examine the contents of the `People` file
- what are the other files in the directory?
- are there *forks* in any of your databases?

File Descriptor Pool

37/68

Unix has limits on the number of concurrently open files.

PostgreSQL maintains a pool of open file descriptors:

办法解决

- to hide this limitation from higher level functions
- to minimise expensive `open()` operations

File names are simply strings: **typedef char *FileName**

Open files are referenced via: **typedef int File**

A **File** is an index into a table of "virtual file descriptors".

... File Descriptor Pool

38/68

Interface to file descriptor (pool):

```
File FileNameOpenFile(FileName fileName,
                      int fileFlags, int fileMode);
     // open a file in the database directory ($PGDATA/base/...)
File OpenTemporaryFile(bool interXact);
     // open temp file; flag: close at end of transaction?
void FileClose(File file);
void FileUnlink(File file);
int  FileRead(File file, char *buffer, int amount);
int  FileWrite(File file, char *buffer, int amount);
int  FileSync(File file);
long FileSeek(File file, long offset, int whence);
int  FileTruncate(File file, long offset);
```

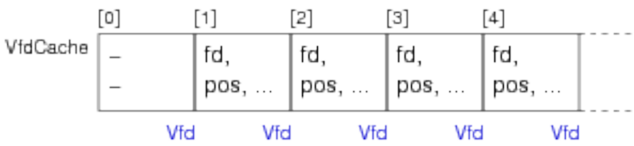
Analogous to Unix syscalls `open()`, `close()`, `read()`, `write()`, `lseek()`, ...

... File Descriptor Pool

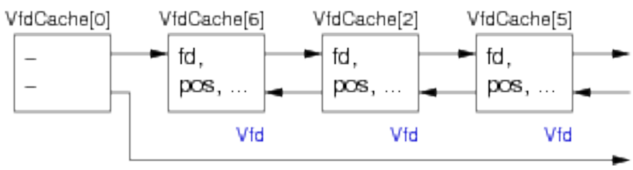
39/68

Virtual file descriptors (**`Vfd`**)

- physically stored in dynamically-allocated array



- also arranged into list by recency-of-use



`VfdCache[0]` holds list head/tail pointers.

... File Descriptor Pool

40/68

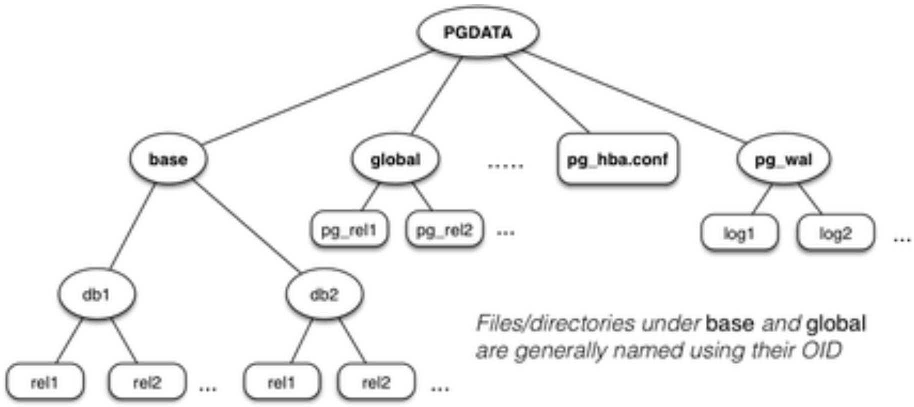
Virtual file descriptor records (simplified):

```
typedef struct vfd
{
    s_short  fd;              // current FD, or VFD_CLOSED if none
    u_short  fdstate;         // bitflags for VFD's state
    File     nextFree;        // link to next free VFD, if in freelist
    File     lruMoreRecently; // doubly linked recency-of-use list
    File     lruLessRecently;
    long     seekPos;         // current logical file position
    char     *fileName;       // name of file, or NULL for unused VFD
    // NB: fileName is malloc'd, and must be free'd when closing the VFD
    int      fileFlags;       // open(2) flags for (re)opening the file
    int      fileMode;        // mode to pass to open(2)
} Vfd;
```

File Manager

41/68

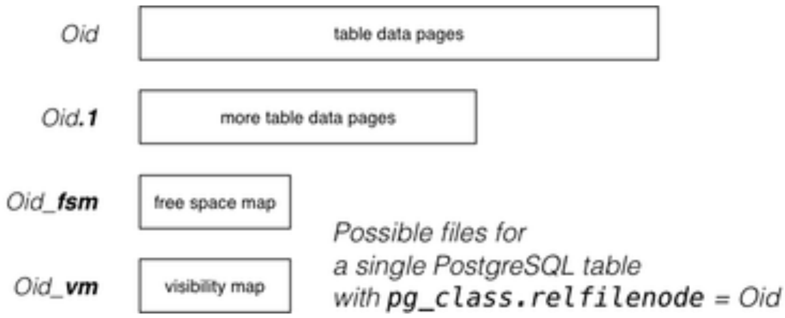
Reminder: PostgreSQL file organisation



... File Manager 42/68

PostgreSQL stores each table

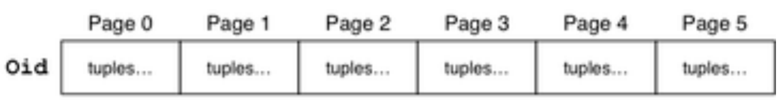
- in the directory `PGDATA/pg_database.oid`
- often in multiple files (aka *forks*)



... File Manager 43/68

Data files (*Oid*, *Oid*.1, ...):

- sequence of fixed-size blocks/pages (typically 8KB)
- each page contains tuple data and admin data (see later)
- max size of data files 1GB (Unix limitation)



PostgreSQL Data File (Heap)

... File Manager 44/68

Free space map (*Oid*_`fsm`):

- indicates where free space is in data pages
- "free" space is only free after `VACUUM`
  (`DELETE` simply marks tuples as no longer in use `xmax`)

Visibility map (*Oid*_`vm`):

- indicates pages where all tuples are "visible"
  (*visible* = accessible to all currently active transactions)
- such pages can be ignored by `VACUUM`

... File Manager 45/68

The "magnetic disk storage manager" (**`storage/smgr/md.c`**)

- manages its own pool of open file descriptors (Vfd's)
- may use several Vfd's to access data, if several forks
- manages mapping from **`PageID`** to file+offset.

PostgreSQL `PageID` values are structured:

```
typedef struct
{
    RelFileNode rnode;    // which relation/file
    ForkNumber  forkNum;  // which fork (of reln)
    BlockNumber blockNum; // which page/block
} BufferTag;
```

... File Manager

46/68

Access to a block of data proceeds (roughly) as follows:

```
// pageID set from pg_catalog tables
// buffer obtained from Buffer pool
getBlock(BufferTag pageID, Buffer buf)
{
   Vfd vf;  off_t offset;
   (vf, offset) = findBlock(pageID)
   lseek(vf.fd, offset, SEEK_SET)
   vf.seekPos = offset;
   nread = read(vf.fd, buf, BLOCKSIZE)
   if (nread < BLOCKSIZE) ... we have a problem
}
```

`BLOCKSIZE` is a global configurable constant (default: 8192)

... File Manager

47/68

```
findBlock(BufferTag pageID) returns (Vfd, off_t)
{
   offset = pageID.blockNum * BLOCKSIZE
   fileName = relpath(pageID.rnode)
   if (pageID.forkNum > 0)
      fileName = fileName+"."+pageID.forkNum
   if (fileName is not in Vfd pool)
      fd = allocate new Vfd for fileName
   else
      fd = use Vfd from pool
   if (offset > fd.fileSize) {
      fd = allocate new Vfd for next fork
      offset = offset - fd.fileSize
   }
   return (fd, offset)
}
```

Buffer Pool

Buffer Pool

49/68

Aim of buffer pool:

- hold pages read from database files, for possible re-use

Used by:

- *access methods* which read/write data pages
- e.g. sequential scan, indexed retrieval, hashing

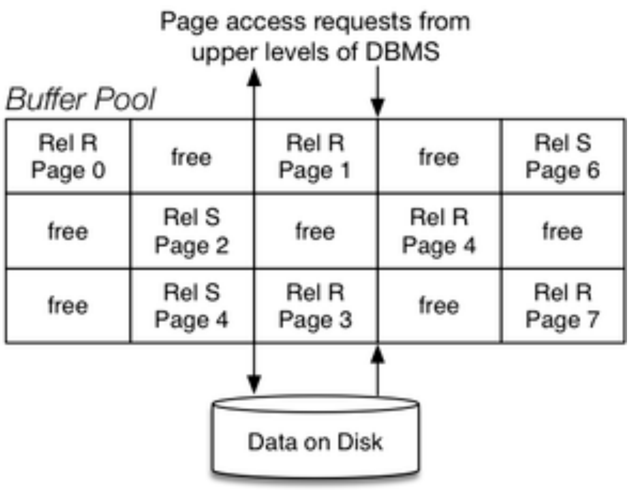
Uses:

- file manager functions to access data files

Note: we use the terms *page* and *block* interchangably

... Buffer Pool

50/68



... Buffer Pool

51/68

Buffer pool operations: (both take single `PageID` argument)

- `request_page(pid)`, `release_page(pid)`, ...

To some extent ...

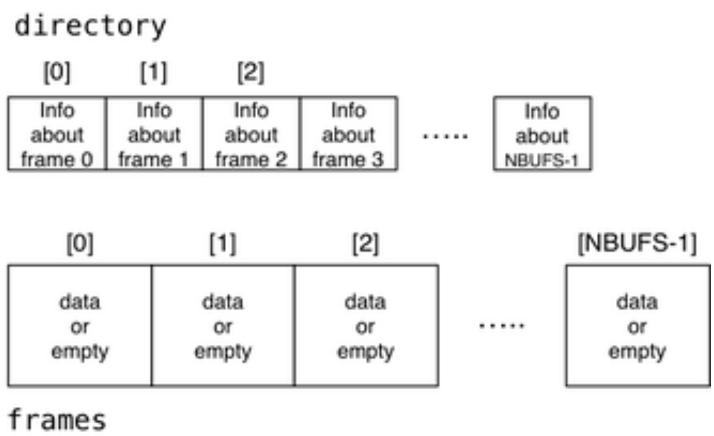
- `request_page()` replaces `getBlock()`
- `release_page()` replaces `putBlock()`

Buffer pool data structures:

- `Page frames[NBUFS]`
- `FrameData directory[NBUFS]`
- `Page` is `byte[BUFSIZE]`

... Buffer Pool

52/68



... Buffer Pool

53/68

For each frame, we need to know: (`FrameData`)

- which Page it contains, or whether empty/free
- whether it has been modified since loading (*dirty bit*)
- how many transactions are currently using it (*pin count*)
- time-stamp for most recent access (assists with replacement)

Pages are referenced by PageID ...

- PageID = BufferTag = (rnode, forkNum, blockNum)

... Buffer Pool

54/68

How scans are performed without Buffer Pool:

```
Buffer buf;
int N = numberOfBlocks(Rel);
for (i = 0; i < N; i++) {
   pageID = makePageID(db,Rel,i);
   getBlock(pageID, buf);
   for (j = 0; j < nTuples(buf); j++)
      process(buf, j)
}
```

Requires `N` page reads.

If we read it again, `N` page reads.

... Buffer Pool

55/68

How scans are performed with Buffer Pool:

```
Buffer buf;
int N = numberOfBlocks(Rel);
for (i = 0; i < N; i++) {
   pageID = makePageID(db,Rel,i);
   bufID = request_page(pageID);
   buf = frames[bufID]
   for (j = 0; j < nTuples(buf); j++)
      process(buf, j)
   release_page(pageID);
}
```

Requires `N` page reads on the first pass.

If we read it again, $0 \leq$ page reads $\leq$ `N`

... Buffer Pool

56/68

Implementation of `request_page()`

```
int request_page(PageID pid)
{
   if (pid in Pool)
      bufID = index for pid in Pool
   else {
      if (no free frames in Pool)
         evict a page (free a frame)
      bufID = allocate free frame
      directory[bufID].page = pid
      directory[bufID].pin_count = 0
      directory[bufID].dirty_bit = 0
   }
   directory[bufID].pin_count++
   return bufID
}
```

... Buffer Pool

57/68

The `release_page(pid)` operation:

- Decrement pin count for specified page

Note: no effect on disk or buffer contents until replacement required

The `mark_page(pid)` operation:

- Set dirty bit on for specified page

Note: doesn't actually write to disk; indicates that page changed

The `flush_page(pid)` operation:

- Write the specified page to disk (using `write_page`)

Note: not generally used by higher levels of DBMS

... Buffer Pool

58/68

Evicting a page ...

- find frame(s) preferably satisfying
  - pin count = 0 (i.e. nobody using it)
  - dirty bit = 0 (not modified)
- if selected frame was modified, flush frame to disk
- flag directory entry as "frame empty"

If multiple frames can potentially be released

- need a policy to decide which is best choice

Page Replacement Policies

59/68

Several schemes are commonly in use:

- Least Recently Used (LRU)
- Most Recently Used (MRU)
- First in First Out (FIFO)
- Random

LRU / MRU require knowledge of when pages were last accessed

- how to keep track of "last access" time?
- base on request/release ops or on *real* page usage?

... Page Replacement Policies

60/68

Cost benefit from buffer pool (with *n* frames) is determined by:

- number of available frames (more ⇒ better)
- replacement strategy vs page access pattern

Example (a): sequential scan, LRU or MRU, $n \geq b$

First scan costs *b* reads; subsequent scans are "free".

Example (b): sequential scan, MRU, $n < b$

First scan costs *b* reads; subsequent scans cost $b - n$ reads.

Example (c): sequential scan, LRU, $n < b$

All scans cost *b* reads; known as *sequential flooding*.

Effect of Buffer Management

61/68

Consider a query to find customers who are also employees:

```
select c.name
from   Customer c, Employee e
where  c.ssn = e.ssn;
```

This might be implemented inside the DBMS via nested loops:

```
for each tuple t1 in Customer {
    for each tuple t2 in Employee {
        if (t1.ssn == t2.ssn)
            append (t1.name) to result set
    }
}
```

... Effect of Buffer Management

62/68

In terms of page-level operations, the algorithm looks like:

```
Rel rC = openRelation("Customer");
Rel rE = openRelation("Employee");
```

```
for (int i = 0; i < nPages(rC); i++) {
    PageID pid1 = makePageID(db,rC,i);
    Page p1 = request_page(pid1);
    for (int j = 0; j < nPages(rE); j++) {
        PageID pid2 = makePageID(db,rE,j);
        Page p2 = request_page(pid2);
        // compare all pairs of tuples from p1,p2
        // construct solution set from matching pairs
        release_page(pid2);
    }
    release_page(pid1);
}
```

Exercise 3: Buffer Cost Benefit (i)

Assume that:

- the `Customer` relation has b_C pages (e.g. 10)
- the `Employee` relation has b_E pages (e.g. 4)

Compute how many page reads occur ...

- if we have only 2 buffers (i.e. effectively no buffer pool)
- if we have 20 buffers
- when a buffer pool with MRU replacement strategy is used
- when a buffer pool with LRU replacement strategy is used

For the last two, buffer pool has $n=3$ slots ($n < b_C$ and $n < b_E$)

Exercise 4: Buffer Cost Benefit (ii)

practice 05

If the tables were larger, the above analysis would be tedious.

Write a C program to simulate buffer pool usage

- assuming a nested loop join as above
- `argv[1]` gives number of pages in "outer" table
- `argv[2]` gives number of pages in "inner" table
- `argv[3]` gives number of slots in buffer pool
- `argv[4]` gives replacement strategy (LRU,MRU,FIFO-Q)

PostgreSQL Buffer Manager

PostgreSQL buffer manager:

- provides a shared pool of memory buffers for all backends
- all access methods get data from disk via buffer manager

Buffers are located in a large region of shared memory.

Definitions: **`src/include/storage/buf*.h`**

Functions: **`src/backend/storage/buffer/*.c`**

Buffer code is also used by backends who want a private buffer pool

---

... PostgreSQL Buffer Manager

Buffer pool consists of:

`BufferDescriptors`

- shared fixed array (size `NBuffers`) of **`BufferDesc`**

`BufferBlocks`

- shared fixed array (size `NBuffers`) of 8KB frames

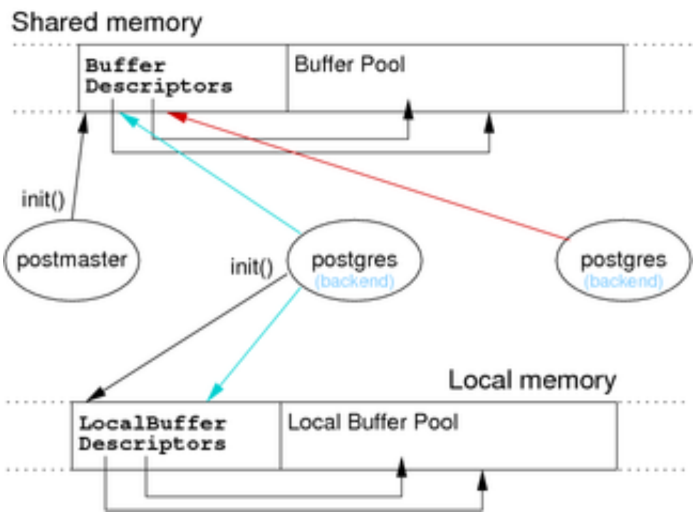
`Buffer` = index values in above arrays

- indexes: global buffers `1..NBuffers`; local buffers negative

Size of buffer pool is set in *postgresql.conf*, e.g.

```
shared_buffers = 16MB    # min 128KB, 16*8KB buffers
```

... PostgreSQL Buffer Manager



... PostgreSQL Buffer Manager

68/68

include/storage/buf.h

- basic buffer manager data types (e.g. **Buffer**)

include/storage/bufmgr.h

- definitions for buffer manager function interface
  (i.e. functions that other parts of the system call to use buffer manager)

include/storage/buf_internals.h

- definitions for buffer manager internals (e.g. **BufferDesc**)

Code: **backend/storage/buffer/*.c**

Commentary: **backend/storage/buffer/README**

Produced: 12 Jun 2019