

Article development led by **acmqueue**
queue.acm.org

Revisiting Gray and Putzolu's famous rule in the age of Flash.

BY GOETZ GRAEFE

The Five-Minute Rule 20 Years Later (and How Flash Memory Changes the Rules)

IN 1987, JIM Gray and Gianfranco Putzolu published their now-famous five-minute rule¹⁵ for trading off memory and I/O capacity. Their calculation compares the cost of holding a record (or page) permanently in memory with the cost of performing disk I/O each time the record (or page) is accessed, using appropriate fractional prices of RAM chips and disk drives. The name of their rule refers to the break-even interval between accesses. If a record (or page) is accessed more often, it should be kept in memory; otherwise, it should remain on disk and be read when needed.

Based on then-current prices and performance characteristics of Tandem equipment, Gray and Putzolu found the price of RAM to hold a 1KB record

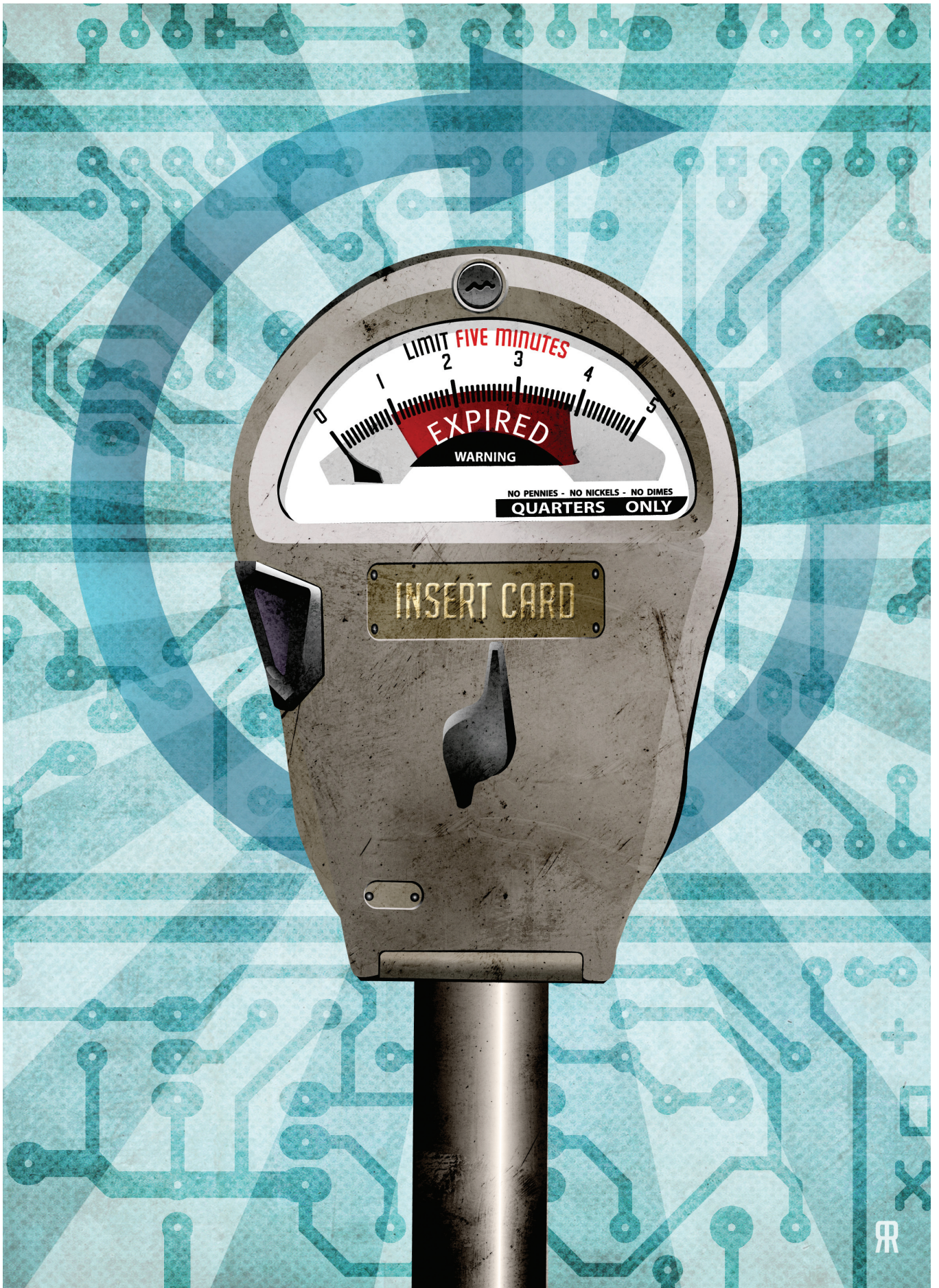
was about equal to the (fractional) price of a disk drive required to access such a record every 400 seconds, which they rounded to five minutes. The break-even interval is about inversely proportional to the record size. Gray and Putzolu reported one hour for 100-byte records and two minutes for 4KB pages.

The five-minute rule was reviewed and renewed 10 years later.¹⁴ Lots of prices and performance parameters had changed (for example, the price of RAM had tumbled from \$5,000 to \$15 per megabyte). Nonetheless, the break-even interval for 4KB pages was still around five minutes. The first goal of this article is to review the five-minute rule after another 10 years.

Of course, both previous articles acknowledged that prices and performance vary among technologies and devices at any point in time (RAM for mainframes versus mini-computers, SCSI versus IDE disks, and so on). Interested readers are invited to reevaluate the appropriate formulas for their environments and equipment. The values used here (in Table 1) are meant to be typical for 2007 technologies rather than universally accurate.

In addition to quantitative changes in prices and performance, qualitative changes already under way will affect the software and hardware architectures of servers and, in particular, database systems. Database software will change radically with the advent of new technologies: virtualization with hardware and software support, as well as higher utilization goals for physical machines; many-core processors and transactional memory supported both in programming environments and hardware;²⁰ deployment in containers housing thousands of processors and many terabytes of data;¹⁷ and flash memory that fills the gap between traditional RAM and traditional rotating disks.

Flash memory falls between traditional RAM and persistent mass storage based on rotating disks in terms of acquisition cost, access



latency, transfer bandwidth, spatial density, power consumption, and cooling costs.¹³ Table 1 and some derived metrics in Table 2 illustrate this point (all metrics derived on 4/11/2007 from dramexchange.com, dvnation.com, buy.com, seagate.com, and samsung.com).

Given the number of CPU instructions possible during the time required for one disk I/O has steadily increased, an intermediate memory in the storage hierarchy is desirable. Flash memory seems to be a highly probable candidate, as has been observed many times by now.

Many architecture details remain to be worked out. For example, in the hardware architecture, will flash memory be accessible via a DIMM slot, a SATA (serial ATA) disk interface, or yet another hardware interface? Given the effort and delay in defining a new hardware interface, adaptations of existing interfaces are likely.

A major question is whether flash memory is considered a special part of either main memory or persistent storage. Asked differently: if a system includes 1GB traditional RAM, 8GB flash memory, and 250GB traditional disk, does the software treat it as

250GB of persistent storage and a 9GB buffer pool, or as 258GB of persistent storage and a 1GB buffer pool? The second goal of this article is to answer this question and, in fact, to argue for different answers in file systems and database systems.

Many design decisions depend on the answer to this question. For example, if flash memory is part of the buffer pool, pages must be considered “dirty” if their contents differ from the equivalent page in persistent storage. Synchronizing the file system or checkpointing a database must force disk writes in those cases. If flash memory is part of persistent storage, these write operations are not required.

Designers of operating systems and file systems will want to use flash memory as an extended buffer pool (extended RAM), whereas database systems will benefit from flash memory as an extended disk (extended persistent storage). Multiple aspects of file systems and database systems consistently favor these two designs. Presenting the case for these designs is the third goal of this article.

Finally, the characteristics of flash memory suggest some substantial

differences in the management of B-tree pages and their allocation. Beyond optimization of page sizes, B-trees can use different units of I/O for flash memory and disks. These page sizes lead to two new five-minute rules. Introducing these two new rules is the fourth goal of this article.

Assumptions

Forward-looking research relies on many assumptions. This section lists the assumptions that led to the conclusions put forth in this article. Some of these assumptions are fairly basic, whereas others are more speculative.

One assumption is that file systems and database systems assign the same data to the flash memory between RAM and the disk drive. Both software systems favor pages with some probability that they will be touched in the future but not with sufficient probability to warrant keeping them in RAM. The estimation and administration of such probabilities follows the usual lines, such as LRU (least recently used).

We assume that the administration of such information uses data structures in RAM, even for pages whose contents have been removed from RAM to flash memory. For example, the LRU chain in a file system’s buffer pool might cover both RAM and flash memory, or there might be two separate LRU chains. A page is loaded into RAM and inserted at the head of the first chain when it is needed by an application. When it reaches the tail of the first chain, the page is moved to flash memory and its descriptor to the head of the second LRU chain. When it reaches the tail of the second chain, the page is moved to disk and removed from the LRU chain. Other replacement algorithms would work *mutatis mutandis*.

Such fine-grained LRU replacement of individual pages is in contrast to assigning entire files, directories, tables, or databases to different storage units. It seems that page replacement is the appropriate granularity in buffer pools. Moreover, proven methods exist for loading and replacing buffer-pool contents entirely automatically, with no assistance from tuning tools or directives by users or administrators needed. An extended buffer pool in

Table 1: Prices and performance of flash and disks.

	RAM	Flash disk	SATA disk
Price and capacity	\$3 for 8×64Mbit	\$999 for 32GB	\$80 for 250GB
Access latency		0.1ms	12ms average
Transfer bandwidth		66MB/s API	300MB/s API
Active power		1W	10W
Idle power		0.1W	8W
Sleep power		0.1W	1W

Table 2: Relative costs for flash memory and disks.

	NAND Flash	SATA disk
Price and capacity	\$999 for 32GB	\$80 for 250GB
Price per GB	\$31.20	\$0.32
Time to read a 4KB page	0.16ms	12.01ms
4KB reads per second	6,200	83
Price per 4KB read per second	\$0.16	\$0.96
Time to read a 256KB page	3.98ms	12.85ms
256KB reads per second	250	78
Price per 256KB read per second	\$3.99	\$1.03

flash memory should exploit the same methods as a traditional buffer pool. For truly comparable and competitive performance and administration costs, a similar approach seems advisable when flash memory is used as an extended disk.

File systems. Our research assumed a fairly traditional file system. Although many file systems differ from this model, most still generally follow it.

In our traditional system, each file is a large byte stream. Files are often read in their entirety, their contents manipulated in memory, and the entire file replaced if it is updated. Archiving, version retention, hierarchical storage management, data movement using removable media, among others, all seem to follow this model as well.

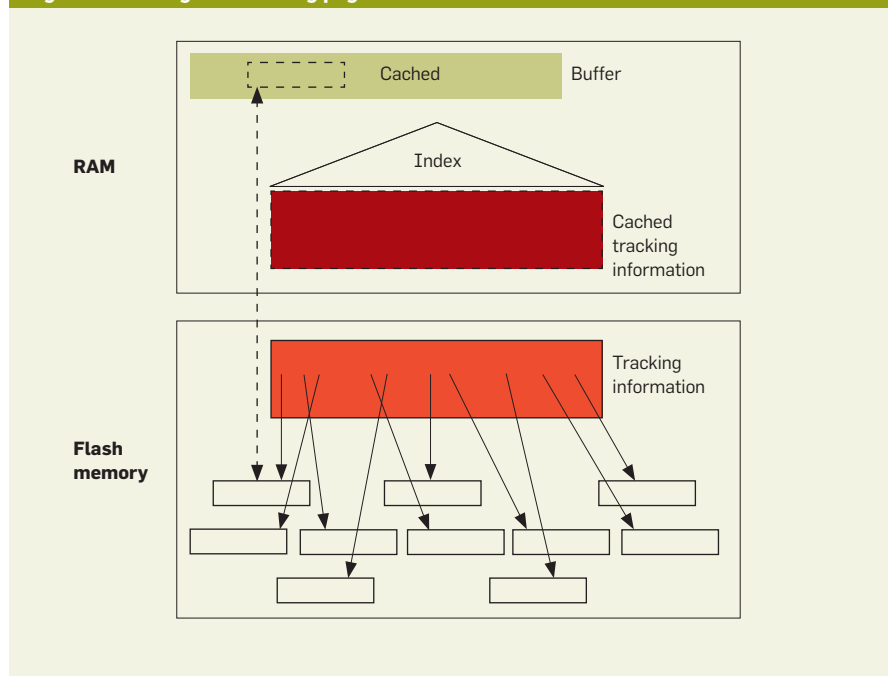
Based on this model, space allocation on disk attempts to use contiguous disk blocks for each file. Metadata is limited to directories, a few standard tags such as a creation time, and data structures for space management.

Consistency of these on-disk data structures is achieved by careful write ordering, fairly quick write-back of updated data blocks, and expensive file-system checks after any less-than-perfect shutdown or media removal. In other words, we assume the absence of transactional guarantees and transactional logging, at least for file contents. If log-based recovery is supported for file contents such as individual pages or records within pages, then a number of the arguments presented here need to be revisited.

Database systems. We assume fairly traditional database systems with B-tree indexes as the workhorse storage structure. Similar tree structures capture not only traditional clustered and nonclustered indexes, but also bitmap indexes, columnar storage, contents indexes, XML indexes, catalogs (metadata), and allocation data structures.

With respect to transactional guarantees, we assume traditional write-ahead logging of both contents changes (such as inserting or deleting a record) and structural changes (such as splitting B-tree nodes). Efficient log-based recovery after failures is enabled by checkpoints that force dirty data from the buffer pool to persistent storage.

Figure 1: Caching and indexing page locations.



Variations such as “second-chance” or fuzzy checkpoints fit within our assumptions. In addition, nonlogged (allocation-only logged) execution is permitted for some operations such as index creation. These operations require appropriate write ordering and a “force” buffer pool policy.¹⁸

Flash memory. Hardware and device drivers are assumed to hide many implementation details such as the specific hardware interface to flash memory. For example, flash memory might be mounted on the computer’s motherboard, a DIMM slot, a PCI board, or within a standard disk enclosure. In all cases, DMA transfers (or something better) are assumed between RAM and flash memory. Moreover, we assume there is either efficient DMA data transfer between flash and disk or a transfer buffer in RAM. The size of such a transfer buffer should be, in a first approximation, about equal to the product of transfer bandwidth and disk latency. If it is desirable that disk writes should never delay disk reads, the increased write-behind latency must be included in the calculation.

Another assumption is that transfer bandwidths of flash memory and disk are comparable. While flash write bandwidth has lagged behind read bandwidth, some products claim a difference of less than a factor of two

(for example, Samsung’s Flash-based solid-state disk used in Table 1). If necessary, the transfer bandwidth can be increased by using array arrangements, as is well known for disk drives; even redundant arrangement of flash memory may prove advantageous in some cases.⁶

Since the reliability of current NAND flash suffers after 100,000–1,000,000 erase-and-write cycles, we assume that some mechanisms for *wear leveling* are provided. These mechanisms ensure that all pages or blocks of pages are written similarly often. It is important to recognize the similarity between wear-leveling algorithms and log-structured file systems,^{22, 27} although the former also move stable, unchanged data such that their locations can absorb some of the erase-and-write cycles.

Note that traditional disk drives do not support more write operations, albeit for different reasons. For example, six years of continuous and sustained writing at 100Mbps overwrites an entire 250GB disk fewer than 80,000 times. In other words, assuming that a log-structured file system is appropriate for RAID-5 or RAID-6 arrays, the reliability of current flash seems comparable. Similarly, overwriting a 32GB flash disk 100,000 times with a sustained average bandwidth of 30Mbps takes about 3.5 years.

In addition to wear leveling, we assume that an asynchronous agent moves stale data from flash memory to disk and immediately erases the freed-up space in flash memory to prepare it for write operations without further delay. This activity also has an immediate equivalence in log-structured file systems—namely, the cleanup activity that prepares space for future log writing. The difference is that disk contents must merely be moved, whereas flash contents must also be erased before the next write operation at that location.

In either file systems or database systems, we assume separate mechanisms for page tracking and page replacement. A traditional buffer pool, for example, provides both, but it uses two different data structures for these two purposes. The standard design relies on an LRU list for page replacement and on a hash table for tracking pages (that is, which pages are present in the buffer pool and in which buffer frames). Alternative algorithms and data structures also separate page tracking and replacement management.

The data structures for the replacement algorithm are assumed to be small and have high traffic and are therefore kept in RAM. We also assume that page tracking must be as persistent as the data, including free-space information. Thus, a buffer pool's hash table is reinitialized during a system reboot, but tracking information for pages on a persistent store such as a disk must be stored with the data. The tracking information may well be cached in RAM while a volume is active, but any changes must be logged and written back to permanent storage. The index required to find the current location of a page may exist only in RAM, being reconstructed every time a volume is opened and the tracking information loaded into the cache in RAM.

As previously mentioned, we assume page replacement on demand. In addition, automatic policies and mechanisms may exist for prefetch, read-ahead, and write-behind.

Based on these considerations, we assume the contents of flash memory are pretty much the same, whether the flash memory extends the buffer pool

or the disk. The central question is therefore not what to keep in cache but how to manage flash-memory contents and its lifetime.

In database systems, flash memory can also be used for recovery logs, because its short access times permit very fast transaction commit. However, limitations in write bandwidth discourage such use. Perhaps systems with dual logs can combine low latency and high bandwidth, one log on a traditional disk and one log on an array of flash chips, with a slightly optimistic policy to consider a transaction committed as soon as the write operation on flash is complete.

Other hardware. In all cases, RAM is assumed to be a substantial size, although probably less than flash memory or disk. The relative sizes should be governed by the five-minute rule.¹⁵ Note that, despite similar transfer bandwidth, the short access latency of flash memory compared with disk results in surprising retention times for data in RAM.

Finally, we assume sufficient processing bandwidth as provided by modern many-core processors. Moreover, forthcoming transactional memory (in hardware and in the software runtime system) is expected to permit highly concurrent maintenance of complex data structures. For example, page replacement heuristics might use priority queues rather than bitmaps or linked lists. Similarly, advanced lock management might benefit from more complex data structures. Nonetheless, we neither assume nor require data structures more complex than those already in common use for page replacement and location tracking.

The Five-Minute Rule

If flash memory is introduced as an

intermediate level in the memory hierarchy, relative sizing of memory levels demands renewed consideration.

Tuning can be based on purchasing cost, total cost of ownership, power, mean time to failure, mean time to data loss, or a combination of metrics. Following Gray and Putzolu,¹⁵ this article focuses on purchasing cost. Other metrics and appropriate formulas to determine relative sizes can be derived similarly (for example, by replacing dollar costs with energy use for caching and moving data).

Gray and Putzolu introduced the following formula:^{14,15}

$$\text{BreakEvenIntervalInSeconds} = (\text{PagesPerMBofRAM} / \text{AccessesPerSecondPerDisk}) \times (\text{Price-PerDiskDrive} / \text{PricePerMBofRAM}).$$

It is derived using formulas for the cost of RAM to hold a page in the buffer pool and the cost of a (fractional) disk to perform I/O every time a page is needed, equating these two costs, and solving the equation for the interval between accesses.

Assuming modern RAM, a disk drive using 4KB pages, and the values from Table 1 and Table 2, this produces

$$(256/83) \times (\$80/\$0.047) = 5,248 \text{ seconds} \approx 90 \text{ minutes} = 1\frac{1}{2} \text{ hours}$$

(The “=” sign often indicates rounding in this article.)

This compares with two minutes (for 4KB pages) 20 years ago. If there is a surprise in this change, it is that the break-even interval has grown by less than two orders of magnitude. Recall that RAM was estimated in 1987 at about \$5,000 per megabyte, whereas the 2007 cost is about \$0.05 per megabyte, a difference of five orders of magnitude. On the other

Table 3: Break-even intervals [seconds].

Page size	1KB	4KB	16KB	65KB	256KB
RAM-SATA	20,978	5,248	1,316	334	88
RAM-flash	2,513	876	467	365	339
Flash-SATA	32,253	8,070	2,024	513	135
RAM-\$400	1,006	351	187	146	136
\$400-SATA	80,553	20,155	5,056	1,281	337

hand, disk prices have also tumbled (\$15,000 per disk in 1987), and disk latency and bandwidth have improved considerably (from 15 accesses per second to about 100 on consumer disks and 200 on high-performance enterprise disks).

For RAM and flash disks of 32GB, the break-even interval is

$$(256 / 6,200) \times (\$999 / \$0.047) = 876 \text{ seconds} \approx 15 \text{ minutes}$$

If the 2007 price for flash disks includes a “novelty premium” and comes down closer to the price of raw flash memory—say, to \$400 (a price also anticipated by Gray and Fitzgerald¹³)—then the break-even interval is 351 seconds \approx 6 minutes.

An important consequence is that in systems tuned using economic considerations, turnover in RAM is about 15 times faster (90 minutes / 6 minutes) if flash memory rather than a traditional disk is the next level in the storage hierarchy. Much less RAM is required, resulting in lower costs for purchase, power, and cooling.

Perhaps most interesting, applying the same formula to flash and disk results in the following:


$$(256 / 83) \times (\$80 / \$0.03) = 8,070 \text{ seconds} \approx 2\frac{1}{4} \text{ hours}$$

Thus, all active data will remain in RAM and flash memory.


Without a doubt, two hours is longer than any common checkpoint interval, which implies that dirty pages in flash are forced to disk not by page replacement but by checkpoints. Pages that are updated frequently must be written much more frequently (because of checkpoints) than is optimal based on Gray and Putzolu’s formula.

In 1987, Gray and Putzolu speculated 20 years into the future and anticipated a “five-hour rule” for RAM and disks. For 1KB records, prices and specifications typical in 2007 suggest 20,978 seconds, or just under six hours. Their prediction was amazingly accurate.

All break-even intervals are different for larger page sizes (64KB or even 256KB). Table 3 shows the break-even intervals, including those just cited, for a variety of page sizes and



Flash memory falls between traditional RAM and persistent mass storage based on rotating disks in terms of acquisition cost, access latency, transfer bandwidth, spatial density, power consumption, and cooling costs.



combinations of storage technologies. (“\$400” stands for a 32GB NAND flash drive available in the future rather than for \$999 in 2007; in fact, 32GB SLC SATA drives are available at retail for \$400 in 2009.)

The old five-minute rule for RAM and disk now applies to 64KB page sizes (334 seconds). Five minutes had been the approximate break-even interval for 1KB in 1987¹⁵ and for 8KB in 1997.¹⁴ This trend reflects the different rates of improvement in disk-access latency and transfer bandwidth.

The five-minute break-even interval also applies to RAM and the expensive flash memory of 2007 for page sizes of 64KB and above (365 seconds and 339 seconds). As the price premium for flash memory decreases, so does the break-even interval (146 seconds and 136 seconds).

Two new five-minute rules are indicated with values in **bold italics** in Table 3. We will come back to this table and these rules in the discussion on optimal node sizes for B-tree indexes.

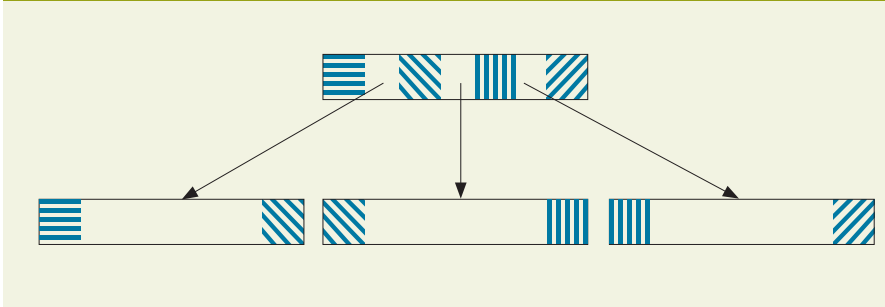
Page Movement

In addition to I/O to and from RAM, a three-level memory hierarchy also requires data movement between flash memory and disk storage.

The pure mechanism for moving pages can be realized in hardware (for example, by DMA transfer), or it might require an indirect transfer via RAM. The former case promises better performance, whereas the latter design can be realized entirely in software without novel hardware. On the other hand, hybrid disk manufacturers might have cost-effective hardware implementations already available.

The policy for page movement is governed or derived from demand-paging and LRU replacement. As mentioned earlier, replacement policies in both file systems and database systems may rely on LRU and can be implemented with appropriate data structures in RAM. As with buffer management in RAM, there may be differences resulting from prefetch, read-ahead, and write-behind. In database systems these may be directed by hints from the query execution layer, whereas file systems must detect page-access patterns

Figure 2: A write-optimized B-tree with fence keys instead of neighbor pointers.



and worthwhile read-ahead actions without the benefit of such hints.

If flash memory is part of the persistent storage, page movement between flash memory and disk is similar to page movement during defragmentation, both in file systems and database systems. The most significant difference is how page movement and current page locations are tracked in these two kinds of systems.

Tracking Page Locations

The mechanisms for tracking page locations are quite different in file systems and database systems. In file systems, pointer pages keep track of data pages or runs of contiguous data pages. Moving an individual page may require breaking up a run. It always requires updating and then writing a pointer page.

In database systems, most data is stored in B-tree indexes, including clustered (primary, nonredundant) and nonclustered (secondary, redundant) indexes on tables, materialized views, and database catalogs. Bitmap indexes, columnar storage, and master-detail clustering can be readily and efficiently represented in B-trees.¹² Tree structures derived from B-trees are also used for *blobs* (binary large objects) and are similar to the storage structures of some file systems.^{5, 25}

For B-trees, moving an individual page can be very expensive or very cheap. The most efficient mechanisms are usually found in utilities for defragmentation or reorganization. Cost or efficiency results from two aspects of B-tree implementation—namely, maintenance of neighbor pointers, and logging for recovery.

First, if physical neighbor pointers are maintained in each B-tree page, moving a single page requires updating two neighbors in addition to the

parent node. If the neighbor pointers are logical using *fence keys*, only the parent page requires an update during a page movement.¹⁰ Figure 2 shows such a B-tree, with neighbor pointers replaced by copies of the separator keys propagated to the parent node during leaf splits. If the parent page is in memory, perhaps even pinned in the buffer pool, recording the new location is rather like updating an in-memory indirection array. The pointer change in the parent page is logged in the recovery log, but there is no need to force the log immediately to stable storage because this change is merely a structural change, not a database contents change.

Second, database systems log changes in the physical database, and in the extreme case both the deleted page image and the newly created page image are logged. Thus, an inefficient implementation fills two log pages whenever a single data page moves from one location to another. A more efficient implementation logs only allocation actions and delays deallocation of the old page image until the new image is safely written in its intended location.¹⁰ In other words, moving a page from one location (for example, on persistent flash memory) to another (for example, on disk) requires only a few bytes in the database recovery log.

The difference between traditional file systems and database systems is the efficiency of updates enabled by the recovery log. In a file system, the new page location must be saved as soon as possible by writing a new image of the pointer page. In a database system, only a single log record or a few short log records must be added to the log buffer. Thus, the overhead for a page movement in a file system is writing an entire pointer

page using a random access, whereas a database system adds a log record of a few dozen bytes to the log buffer that will eventually be written using large sequential write operations.

If a file system uses flash memory as persistent storage, moving a page between a flash memory location and an on-disk location adds substantial overhead. Thus, most file-system designers will probably prefer flash memory as an extension to the buffer pool rather than as an extension of the disk, thus avoiding this overhead.

A database system, however, has built-in mechanisms that can easily track page movements. These mechanisms are inherent in the “workhorse” data structure, B-tree indexes. Compared with file systems, these mechanisms permit efficient page movement, each one requiring only a fraction of a sequential write (in the recovery log) rather than a full random write.

Moreover, the database mechanisms are reliable. Should a failure occur during a page movement, database recovery is driven by the recovery log, whereas a traditional file system requires checking the entire volume during reboot.

Checkpoint Processing

To ensure fast recovery after a system failure, database systems use checkpoints. Their effect is that recovery needs to consider database activity only from the most recent checkpoint, plus some limited activity explicitly indicated in the checkpoint information. The main effort is writing dirty pages from the buffer pool to persistent storage.

If pages in flash memory are part of the buffer pool, dirty pages must be written to disk during database checkpoints. Common checkpoint intervals are measured in seconds or minutes. Alternatively, if checkpoints are not truly points but intervals, it is reasonable to flush pages and perform checkpoint activities continuously, starting the next checkpoint as soon as one finishes. With flash memory as part of the buffer pool, many writes to flash memory require a write to disk soon thereafter as part of checkpoint processing, and flash memory as the intermediate level in the memory hierarchy fails to absorb write activity. Recall, this effect may be exacerbated

if RAM is kept small because of the presence of flash memory.

If, on the other hand, flash memory is considered persistent storage, writing to flash memory is sufficient. Write-through to disk is required only as part of page replacement (such as, when a page's usage suggests placement on disk rather than in flash memory). Thus, checkpoints do not incur the cost of moving data from flash memory to disk.

Checkpoints might even be faster in systems with flash memory because dirty pages in RAM need to be written merely to flash memory, not to disk. Given the very fast random access in flash memory relative to disk drives, this difference might speed up checkpoints significantly.

To summarize, database systems benefit if flash memory is treated as part of the system's persistent storage. In contrast, traditional file systems do not have systemwide checkpoints that flush the recovery log and any dirty data from the buffer pool. Instead, they rely on carefully writing modified file-system pages because of the lack of a recovery log protecting file contents.

Page Sizes

In addition to tuning based on the five-minute rule, another optimization based on access performance is sizing of B-tree nodes. The optimal page size minimizes the time spent on I/O during a root-to-leaf search. It balances a short I/O (that is, the desire for small pages) with a high reduction in remaining search space (that is, the desire for large pages).

Assuming binary search within each B-tree node, the reduction in remaining search space is measured by the logarithm of the number of records within each node. This measure was called a node's *utility* in our earlier work.¹⁴ This optimization is essentially equivalent to one described in the original research on B-trees.³

Table 4 illustrates this optimization for 20-byte records (typical with prefix and suffix truncation⁴) and for nodes filled at about 70%.

Not surprisingly, the optimal node size for B-tree indexes on modern high-bandwidth disks is much larger than the page sizes in traditional database systems. With those disks,

the access time dominates for all small page sizes, such that additional byte transfer and thus additional utility are almost free.

B-tree nodes of 256KB are near optimal. For those, Table 3 indicates a break-even time for RAM and disk of 88 seconds. For a \$400 flash disk and a traditional rotating hard disk, Table 4 indicates 337 seconds or just over five minutes. This is the first of the two new five-minute rules.

Table 5 illustrates the same calculations for B-trees on flash memory. Because there is no mechanical seeking or rotation, transfer time dominates access time even for small pages. The optimal page size for B-trees on flash memory is 2KB, much smaller than for traditional disk drives. In Table 3, the break-even interval for 4KB pages

is 351 seconds. This is the second new five-minute rule.

The implication of two different optimal page sizes is that any uniform node size for B-trees on flash memory and traditional rotating hard disks is suboptimal. Optimizing page sizes for both media requires a change in buffer management, space allocation, and some of the B-tree logic.

Fortunately, Patrick O'Neil of the University of Massachusetts already designed a space allocation scheme for B-trees in which neighboring leaf nodes usually reside within the same contiguous extent of pages.²³ When a new page is needed for a node split, another page within the same extent is allocated. When an extent overflows, half its pages are moved to a newly allocated extent. In other words, the

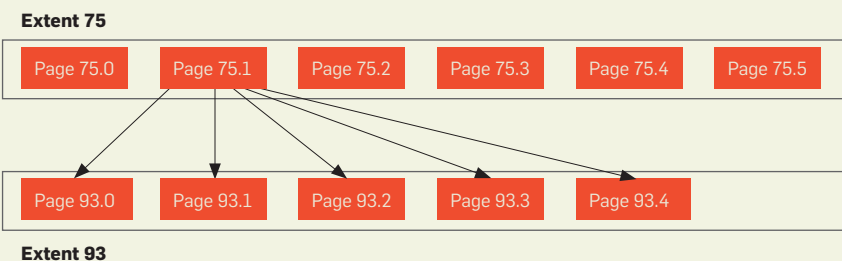
Table 4: Page utility for B-tree nodes on disk.

Page size	Records per page	Node utility	Access time	Utility/time
4KB	140	7	12.0ms	0.58
16KB	560	0	12.1ms	0.75
64KB	2,240	11	12.2ms	0.90
128KB	4,480	12	12.4ms	0.97
256KB	8,960	13	12.9ms	1.01
512KB	17,920	14	13.7ms	1.02
1MB	35,840	15	15.4ms	0.97

Table 5: Page utility for B-tree nodes on flash memory.

Page size	Records per page	Node utility	Access time	Utility/time
1KB	35	5	0.11ms	43.4
2KB	70	6	0.13ms	46.1
4KB	140	7	0.16ms	43.6
8KB	280	8	0.22ms	36.2
16KB	560	9	0.34ms	26.3
64KB	2,240	11	1.07ms	10.3

Figure 3: Pages and extents in an SB-tree.



“split in half when full” logic of B-trees is applied not only to pages containing records, but also to contiguous disk extents containing pages.


Using O’Neil’s SB-trees (*S* meaning *sequential*), 256KB extents can be the units of transfer between flash memory and disk, whereas 4KB pages can be the unit of transfer between RAM and flash memory. Figure 3 shows pages within two extents. Child pointers in a B-tree (also shown) refer to individual pages. If multiple neighboring child pointers refer to neighboring pages (as shown), the pointer values can be represented compactly with run-length encoding applied not to a set of duplicate key values but to a series of values with constant increments. For example, the five child pointers in extent 75.1 in Figure 3 can be represented by the page identifier 93.0 and the counter 5.

Similar notions of self-similar B-trees have also been proposed for higher levels in the memory hierarchy, for example, in the form of B-trees of cache lines for the indirection vector within a large page.¹⁹ Given that there are at least three levels of B-trees and three node sizes now (cache lines, flash memory pages, and disk pages), research into cache-oblivious B-trees² might be promising.


Database-Query Processing

Self-similar designs apply both to data structures such as B-trees and to algorithms. For example, sort algorithms already use algorithms similar to traditional external merge sorts in multiple ways—to merge runs not only on disk but also in memory, where the initial runs are sized to limit run creation to the CPU cache.^{11,21}

The same technique might be applied three times instead of twice: first, cache-size runs in memory are merged into memory-size runs in memory; second, in larger sort operations, memory-size runs in flash memory are merged into runs on disk; and third, runs on disk are merged to form the final sorted result. Read-ahead, forecasting, write-behind, and page sizes all deserve a new look in a multilevel memory hierarchy consisting of cache, RAM, flash memory, and traditional disk drives. These page sizes can then inform the break-even calculation for page retention versus I/O and thus



The 20-year-old five-minute rule for RAM and disks still holds, but for ever-larger disk pages. Moreover, it should be augmented by two new five-minute rules: one for small pages moving between RAM and flash memory and one for large pages moving between flash memory and traditional disks.



guide the optimal capacities at each level of the memory hierarchy.

We can surmise that a variation of this sort algorithm will be not only fast but also energy efficient. While energy efficiency has always been crucial for battery-powered devices, research into energy-efficient query processing on server machines is only now beginning.²⁴ For example, for both flash memory and disks, energy-optimal page sizes might well differ from performance-optimal page sizes.

The I/O pattern of an external merge sort is similar (albeit in the opposite direction) to the I/O pattern of an external distribution sort. Figure 4 illustrates how merging combines many small files into a large file, with many seek operations in the small files as demanded by the merge logic, and how partitioning divides a single large file into many small files, with many seek operations in the small files as demanded by the partitioning function. The I/O pattern of a distribution sort is equal to that of partitioning during hash join and hash aggregation.⁸ All of these algorithms require reevaluation and redesign in a three-level memory hierarchy, or even a four-level hierarchy if CPU caches are also considered.²⁶

Flash memory with its very fast access times may well revive interest in index-based query execution.^{7,9} Instead of large scans and memory-intensive operations such as sorting and hash join, fast accesses to index pages shift the break-even point toward index-to-index navigation. For example, assume a table with 100 million rows of 100 bytes and table scans at 100MB per second. A table scan takes 100 seconds. Searching a secondary index requires fetching individual rows from the table. If the table is stored on a traditional disk, then a period of 100 seconds permits fetching about 10,000 records. If more than 10,000 rows satisfy the query predicate, then the table scan is faster. If, however, the table is stored on a flash device, 100 seconds will permit fetching about 500,000 records. Thus, flash storage shifts the break-even point between table scan and index search from 10,000 to 500,000 rows satisfying the query predicate, and many more query execution plans will rely on index-to-index navigation rather than large scans and hash joins.

Multiple secondary indexes for a single table can be exploited into index intersection, index joins, among others. Fast access to individual pages and records also benefits those query execution plans. Like secondary indexes, column stores or more generally vertical partitioning also require fetching records from multiple places to assemble complete rows. Thus, as seen in the example of database query processing, using flash memory in addition to or even as replacement of traditional disks not only forces reevaluation of optimal use of the hardware but also means some substantial software changes.

Record and Object Caches

Page sizes in database systems have grown over the years, although not as fast as disk-transfer bandwidth. On the other hand, small pages require less buffer-pool space for each root-to-leaf search. For example, consider an index with 20 million entries. With index pages of 128KB and 4,500 records, a root-to-leaf search requires two nodes and thus 256KB in the buffer pool, although half of that (the root node) can probably be shared with other transactions. With 8KB index pages and 280 records per page, a root-to-leaf search requires three nodes or 24KB in the buffer pool, or one order of magnitude less.

In traditional database architecture, the default page size is a compromise between efficient index search (using large B-tree nodes as previously discussed here and in the original B-tree papers³) and moderate buffer-pool requirements for each index search. Nonetheless, the previous example requires 24KB in the buffer pool for finding a record of perhaps only 20 bytes,

Figure 4: Merging and partitioning files.

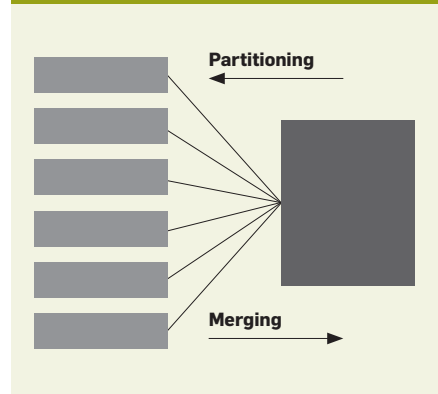
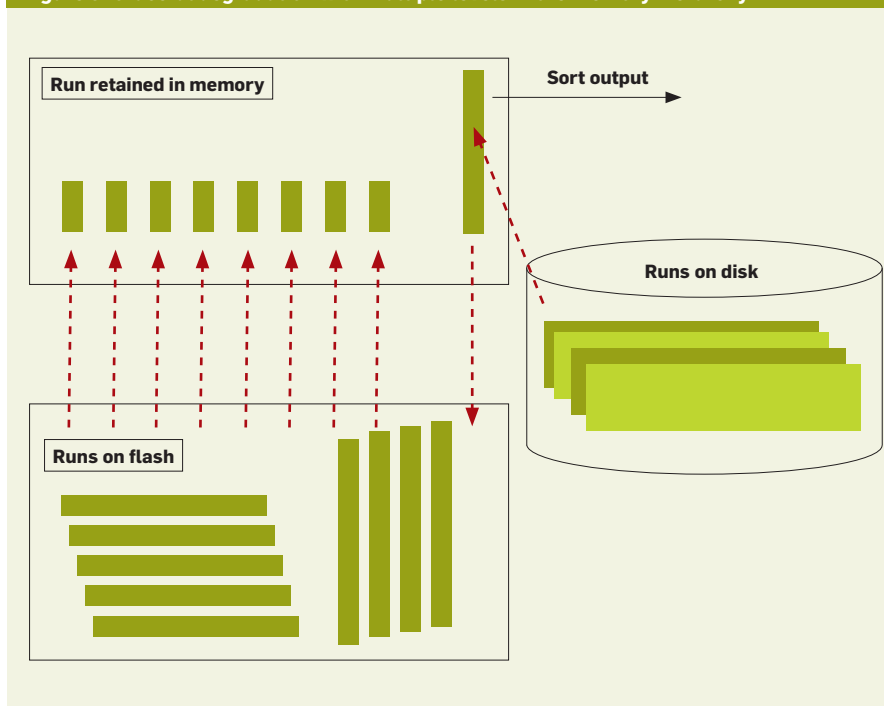


Figure 5: Graceful degradation with multiple levels in the memory hierarchy.



and it requires 8KB of the buffer pool for retaining these 20 bytes in memory. An alternative design uses large on-disk pages and a record cache that serves applications, because record caches minimize memory needs yet provide the desired data retention. In-memory databases represent a specific form of record caches when used as front ends for traditional disk-based databases.

The introduction of flash memory with its fast access latency and its small optimal page size may render record caches obsolete. With the large on-disk pages in flash memory and only small pages in the in-memory buffer pool, the desired compromise can be achieved without the need for two separate data structures (such as, a transacted B-tree and a separate record cache).

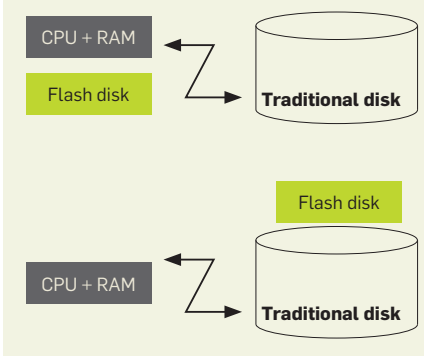
Future Work

Several directions for future research suggest themselves. First, while the analyses in this article focused on purchasing costs, a consideration of other costs could capture the total cost of ownership. A focus on energy consumption, for example, could lead to different break-even points or even entirely different conclusions. Along with CPU scheduling, algorithms for staging data in the memory hierarchy—including buffer-pool replacement and asynchronous I/O—may be the soft-

ware techniques with the highest impact on energy consumption. Note that traditional database-query processing relies on asynchronous I/O to reduce response times; if the primary cost metric for query processing is energy consumption, asynchronous I/O has no advantage over synchronous I/O.

Second, the five-minute rule applies to permanent data and its management in a buffer pool. The optimal retention time for temporary data such as run files in sorting and overflow files in hash join and hash aggregation may be different. For sorting, as for B-tree searches, the goal should be to maximize the number of comparisons per unit of I/O time or per unit of energy spent on I/O. Our initial research and algorithm design has focused on algorithms with graceful degradation in sorting and for hybrid hash join (that is, spilling memory contents to flash only when and as much as truly required, and similarly spilling flash contents to disk only when and as much as truly required). The different optimal page sizes can be exploited to achieve very high effective merge fan-in and partitioning fan-out with relatively little main memory. Figure 5 shows the final merge step—very large runs on disk use large pages that are buffered in flash memory (shown as vertical boxes), a few small runs have remained in flash

Figure 6: Local flash drives versus hybrid drives in network-attached storage.



and never were merged to form very large runs on disk (shown as horizontal boxes), and the available RAM is used to merge a very large number of runs exploiting the small page size optimal for flash devices.

Third, Gray and Putzolu offered further rules of thumb, such as the 10-byte rule for trading memory and CPU power. These rules also warrant revisiting for both costs and energy. Compared with 1987, the most fundamental change may be that CPU power should be measured not in instructions but in cache line replacements. Trading off space and time seems like a new problem in an environment with multiple levels in the memory hierarchy. A modern memory hierarchy might be very deep: multiple levels of CPU caches, main memory (possibly in a NUMA design), flash devices, and finally performance-optimized “enterprise” disks and capacity-optimized “consumer” disks. The lower levels may rely on various software techniques with different trade-offs between performance and reliability, such as striping, mirroring, single-redundancy RAID-5, dual-redundancy RAID-6, log-structured file systems, and write-optimized B-trees.

Fourth, what are the best data movement policies? One extreme is a database administrator explicitly moving entire files, tables, or indexes between flash memory and traditional disk. Another extreme is automatic movement of individual pages, controlled by a replacement policy such as LRU. Intermediate policies may focus on the roles of individual pages within a database or on the current query-processing activity. For example, all catalog pages may be moved as a

unit after schema changes to facilitate fast recompilation of all cached query execution plans, and all relevant upper B-tree levels may be prefetched and cached in RAM or in flash memory during execution of query plans relying on index-to-index navigation. The variety of possibilities may overwhelm automatic policies and may require hints or directives from applications or database software.

Fifth, what are the secondary and tertiary effects of introducing flash memory into the memory hierarchy of a database server? For example, short access times permit a lower multi-programming level, because only short I/O operations must be hidden by asynchronous I/O and context switching. A lower multi-programming level in turn may reduce contention for memory in sort and hash operations, locks (concurrency control for database contents), and latches (concurrency control for in-memory data structures). Should this effect prove significant, the effort and complexity of using a fine granularity of locking may be reduced. Page-level concurrency control may also be sufficient simply as a result of small page sizes. Similarly, in-page data structures may require less optimization, although some techniques may apply to small pages (optimized for flash) within large pages (optimized for disks)—for example, clustering records versus clustering fields.¹

Sixth, will hardware architecture considerations invalidate some of the findings and conclusions of this article? For example, disks are currently separated from the main processors (for example, in network-attached storage or storage-area networks). Will flash devices be placed with the main processors? If so, is it still a good idea to use flash devices as extended disk rather than extended buffer pool? Figure 6 shows two of these alternatives. In the top arrangement, questions arise about the scope and effectiveness of centralized storage management, the granularity of failures and replacement, and so on, whereas many of these questions have much more obvious answers in the bottom arrangement.

Seventh, how will flash memory affect in-memory database systems? Will they become more scalable,

affordable, and popular based on memory inexpensively extended with flash memory rather than RAM? Will they become less popular as a result of very fast traditional database systems using flash memory instead of (or in addition to) disks? Can a traditional code base using flash memory instead of traditional disks compete with a specialized in-memory database system in terms of performance, total cost of ownership, development and maintenance costs, or time to market of features and releases? What techniques in the buffer pool are required to achieve performance competitive with in-memory databases? For example, the upper levels of B-tree indexes can be pinned in the buffer pool and augmented with memory addresses of all child pages (or their buffer descriptors) also pinned in the buffer pool, and auxiliary structures may enable efficient interpolation search instead of binary search.

Finally, techniques similar to generational garbage collection may benefit storage hierarchies.²² Selective reclamation applies not only to unreachable in-memory objects but also to buffer-pool pages and favored locations on permanent storage. Such research also may provide guidance for log-structured file systems, wear leveling for flash memory, and write-optimized B-trees on RAID storage.

Conclusion

The 20-year-old five-minute rule for RAM and disks still holds, but for ever-larger disk pages. Moreover, it should be augmented by two new five-minute rules: one for small pages moving between RAM and flash memory and one for large pages moving between flash memory and traditional disks. For small pages moving between RAM and disk, Gray and Putzolu were amazingly accurate in predicting a five-hour break-even point two decades into the future.

Research into flash memory and its place in system architectures is urgent and important. Within a few years, flash memory will be used to fill the gap between traditional RAM and traditional disk drives in many operating systems, file systems, and database systems.

Flash memory can be used to extend

RAM or persistent storage. These models are called *extended buffer pool* and *extended disk* here. Both models may seem viable in operating systems, file systems, and in database systems. The different characteristics of each of these systems, however, will require different usage models.

In both models, contents of RAM and flash will be governed by LRU-like replacement algorithms that attempt to keep the most valuable pages in RAM and the least valuable pages on traditional disks. The linked list or other data structure implementing the replacement policy for flash memory will be maintained in RAM.

Operating systems and traditional file systems will use flash memory mostly as transient memory (for example, as a fast backup store for virtual memory and as a secondary file-system cache). Both of these applications fall into the extended buffer-pool model. During an orderly system shutdown, the flash memory contents must be written to persistent storage. During a system crash, however, the RAM-based description of flash-memory contents will be lost and must be reconstructed by a contents analysis similar to a traditional file-system check. Alternatively, flash-memory contents can be voided and reloaded on demand.

Database systems, on the other hand, will employ flash memory as persistent storage, using the extended disk model. The current contents will be described in persistent data structures, such as parent pages in B-tree indexes. Traditional durability mechanisms—in particular, logging and checkpoints—ensure consistency and efficient recovery after system crashes. Checkpoints and orderly system shutdowns have no need to write flash memory contents to disk, and the pre-shutdown of flash contents is required for a complete restart.


There are two reasons for these different usage models for flash memory. First, database systems rely on regular checkpoints during which dirty pages are flushed from the buffer pool to persistent storage. If a dirty page is moved from RAM to the extended buffer pool in flash memory, it creates substantial overhead during the next checkpoint. A free buffer must be found

in RAM, the page contents must be read from flash memory into RAM, and then the page must be written to disk. Adding such overhead to checkpoints is not attractive in database systems with frequent checkpoints. Operating systems and traditional file systems, on the other hand, do not rely on checkpoints and thus can exploit flash memory as an extended buffer pool.

Second, the principal persistent data structures of databases, B-tree indexes, provide precisely the mapping and location-tracking mechanisms needed to complement frequent page movement and replacement. Thus, tracking a data page when it moves between disk and flash relies on the same data structure maintained for efficient database search. In addition, avoiding indirection in locating a page also makes database searches as efficient as possible.

Finally, as the ratio of access latencies and transfer bandwidth is very different for flash memory and disks, different B-tree node sizes are optimal. O’Neil’s SB-tree exploits two different node sizes as needed in a multilevel storage hierarchy. The required inexpensive mechanisms for moving individual pages are the same as those required when moving pages between flash memory and disk.

Acknowledgments

This article is dedicated to Jim Gray, who suggested this research and helped the author and many others many times in many ways. Barb Peters, Lily Jow, Harumi Kuno, José Blakeley, Mehul Shah, the DaMoN 2007 reviewers, and particularly Harumi Kuno suggested multiple improvements after reading earlier versions of this work. 

References

1. Ailamaki, A., DeWitt, D.J. and Hill, M.D. Data page layouts for relational databases on deep memory hierarchies. *VLDB Journal* 11, 3 (2002), 198–215.
2. Bender, M.A. Demaine, E.D. and Farach-Colton, M. Cache-oblivious B-trees. *SIAM Journal on Computing* 35, 2 (2005), 341–358.
3. Bayer, R. and McCreight, E.M. Organization and maintenance of large ordered indexes. SIGFI-DET Workshop (1970), 107–141.
4. Bayer, R. and Unterauer, K. Prefix B-trees. *ACM Transactions on Database Systems* 2, 1 (1977), 11–26.
5. Carey, M.J., DeWitt, D.J., Richardson, J.E. and Shekita, E.J. Storage management in EXODUS. In *Object-Oriented Concepts, Databases, and Applications*. W. Kim and F. Lochovsky, Eds. ACM, N.Y., 1989, 341–369.
6. Chen, P.M., Lee, E.L. Gibson, G.A., Katz, R.H. and Patterson, D.A. 1994. RAID: high-performance, reliable secondary storage. *ACM Computing Surveys* 26(2): 145–185.
7. DeWitt, D.J., Naughton, J.F. and Burger, J. Nested loops revisited. *Parallel and Distributed Information Systems* (1993), 230–242.
8. Graefe, G. Query evaluation techniques for large databases. *ACM Computing Surveys* 25, 2 (1993), 73–170.
9. Graefe, G. Executing nested queries. *Database Systems for Business, Technology and Web* (2003), 58–77.
10. Graefe, G. Write-optimized B-trees. *VLDB Journal* (2004), 672–683.
11. Graefe, G. Implementing sorting in database systems. *ACM Computing Surveys* 38, 3 (2006), 69–106.
12. Graefe, G. Master-detail clustering using merged indexes. *Informatik-Forschung und Entwicklung*, 2007.
13. Gray, J. and Fitzgerald, B. 2007. Flash disk opportunity for server-applications; <http://research.microsoft.com/~gray/papers/FlashDiskPublic.doc>.
14. Gray, J., Graefe, G. 1997. The five-minute rule ten years later, and other computer storage rules of thumb. *SIGMOD Record* 26, 4 (1997), 63–68.
15. Gray, J. and Putzolu, G.R. The 5-minute rule for trading memory for disk accesses and the 10-byte rule for trading memory for CPU time. *SIGMOD Journal* (1987), 395–398.
16. Härder, T. Implementing a generalized access path structure for a relational database system. *ACM Transactions on Database Systems* 3, 3 (1978), 285–298.
17. Hamilton, J. An architecture for modular data centers. In *Proceedings of the Conference on Innovative Data Systems Research*, 2007.
18. Härder, T. and Reuter, A. Principles of transaction-oriented database recovery. *ACM Computing Surveys* 15, 4 (1983), 287–317.
19. Lomet, D.B. The evolution of effective B-tree page organization and techniques: a personal account. *SIGMOD Record* 30, 3, 64–69.
20. Larus, J.R. and Rajwar, R. *Transactional Memory. Synthesis Lectures on Computer Architecture*. Morgan & Claypool, 2007.
21. Nyberg, C., Barclay, T., Cvetanovic, Z., Gray, J. and Lomet, D.B. AlphaSort: A cache-sensitive parallel external sort. *VLDB Journal* (1995), 603–627.
22. Ousterhout, J.K. and Douglass, F. Beating the I/O bottleneck: A case for log-structured file systems. *Operating Systems Review* 23, 1 (1989), 11–28.
23. O’Neil, P.W. The SB-tree: An index-sequential structure for high-performance sequential access. *Acta Informatica* 29, 3 (1992), 241–265.
24. Rivoire, S., Shah, M., Ranganathan, P. and Kozyrakos, C. JouleSort: A balanced energy-efficiency benchmark. *SIGMOD Record*, 2007.
25. Stonebraker, M. Operating system support for database management. *Commun. ACM* 24, 7 (July 1981), 412–418.
26. Shatdal, A., Kant, C. and Naughton, J.F. Cache-conscious algorithms for relational query processing. *VLDB Journal* (1994), 510–521.
27. Woodhouse, D. JFFS: The Journaling Flash File System. Ottawa Linux Symposium, Red Hat Inc., 2001.

Related articles on queue.acm.org

Flash Storage Today

Adam Leventhal

<http://queue.acm.org/detail.cfm?id=1413262>

Flash Disk Opportunity for Server Applications

Jim Gray, Bob Fitzgerald

<http://queue.acm.org/detail.cfm?id=1413261>

Enterprise SSDs

Mark Moshayedi, Patrick Wilkison

<http://queue.acm.org/detail.cfm?id=1413263>

Goetz Graefe (Goetz.Graefe@HP.com) joined Hewlett-Packard Laboratories after seven years as an academic researcher and teacher followed by 12 years as a product architect and developer at Microsoft. He was recently named an HP Fellow. His Volcano research project was awarded the 10-year Test-of-Time Award at ACM SIGMOD 2000 for work on query execution.

An earlier version of this article was originally published in *Proceedings of the Third International Workshop on Data Management on New Hardware* (June 15, 2007), Beijing, China.

© 2009 ACM 0001-0782/09/0700 \$10.00