In-Memory Big Data Management and Processing: A Survey

Hao Zhang, Gang Chen, *Member, IEEE*, Beng Chin Ooi, *Fellow, IEEE*,
Kian-Lee Tan, *Member, IEEE*, and Meihui Zhang, *Member, IEEE*

Abstract—Growing main memory capacity has fueled the development of in-memory big data management and processing. By eliminating disk I/O bottleneck, it is now possible to support interactive data analytics. However, in-memory systems are much more sensitive to other sources of overhead that do not matter in traditional I/O-bounded disk-based systems. Some issues such as fault-tolerance and consistency are also more challenging to handle in in-memory environment. We are witnessing a revolution in the design of database systems that exploits main memory as its data storage layer. Many of these researches have focused along several dimensions: modern CPU and memory hierarchy utilization, time/space efficiency, parallelism, and concurrency control. In this survey, we aim to provide a thorough review of a wide range of in-memory data management and processing proposals and systems, including both data storage systems and data processing frameworks. We also give a comprehensive presentation of important technology in memory management, and some key factors that need to be considered in order to achieve efficient in-memory data management and processing.

Index Terms—Primary memory, DRAM, relational databases, distributed databases, query processing

1 Introduction

The explosion of Big Data has prompted much research to develop systems to support ultra-low latency service and real-time data analytics. Existing disk-based systems can no longer offer timely response due to the high access latency to hard disks. The unacceptable performance was initially encountered by Internet companies such as Amazon, Google, Facebook and Twitter, but is now also becoming an obstacle for other companies/organizations which desire to provide a meaningful real-time service (e.g., real-time bidding, advertising, social gaming). For instance, trading companies need to detect a sudden change in the trading prices and react instantly (in several milliseconds), which is impossible to achieve using traditional disk-based processing/storage systems. To meet the strict real-time requirements for analyzing mass amounts of data and servicing requests within milliseconds, an in-memory system/database that keeps the data in the random access memory (RAM) all the time is necessary.

Jim Gray's insight that "Memory is the new disk, disk is the new tape" is becoming true today [1]—we are witnessing a trend where memory will eventually replace disk and the role of disks must inevitably become more archival. In the last decade, multi-core processors and the availability of large amounts of main memory at plummeting cost are creating new breakthroughs, making it viable to build in-memory systems where a significant part, if not the entirety, of the database fits in memory. For example, memory storage capacity and bandwidth have been doubling roughly every three years, while its price has been dropping by a factor of 10 every five years. Similarly, there have been significant advances in non-volatile memory (NVM) such as SSD and the impending launch of various NVMs such as phase change memory (PCM). The number of I/O operations per second in such devices is far greater than hard disks. Modern high-end servers usually have multiple sockets, each of which can have tens or hundreds of gigabytes of DRAM, and tens of cores, and in total, a server may have several terabytes of DRAM and hundreds of cores. Moreover, in a distributed environment, it is possible to aggregate the memories from a large number of server nodes to the extent that the aggregated memory is able to keep all the data for a variety of large-scale applications (e.g., Facebook [2]).

Database systems have been evolving over the last few decades, mainly driven by advances in hardware, availability of a large amount of data, collection of data at an unprecedented rate, emerging applications and so on. The landscape of data management systems is increasingly fragmented based on application domains (i.e., applications relying on relational data, graph-based data, stream data). Fig. 1 shows state-of-the-art commercial and academic systems for disk-based and in-memory operations. In this survey, we focus on in-memory systems; readers are referred to [3] for a survey on disk-based systems.

In business operations, speed is not an option, but a must. Hence every avenue is exploited to further improve performance, including reducing dependency on the hard disk, adding more memory to make more data resident in

- *H. Zhang, B.C. Ooi, and K.-L. Tan are with the School of Computing, National University of Singapore, Singapore 117417. E-mail: {zhangh, ooibc, tankl}@comp.nus.edu.sg.*
- *G. Chen is with the College of Computer Science, Zhejiang University, Hangzhou 310027, China. E-mail: cg@cs.zju.edu.cn.*
- *M. Zhang is with the Information Systems Technology and Design Pillar, Singapore University of Technology and Design, Singapore 487372. E-mail: meihui_zhang@sutd.edu.sg.*

Manuscript received 16 Jan. 2015; revised 22 Apr. 2015; accepted 25 Apr. 2015. Date of publication 28 Apr. 2015; date of current version 1 June 2015.
Recommended for acceptance by J. Pei.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TKDE.2015.2427795

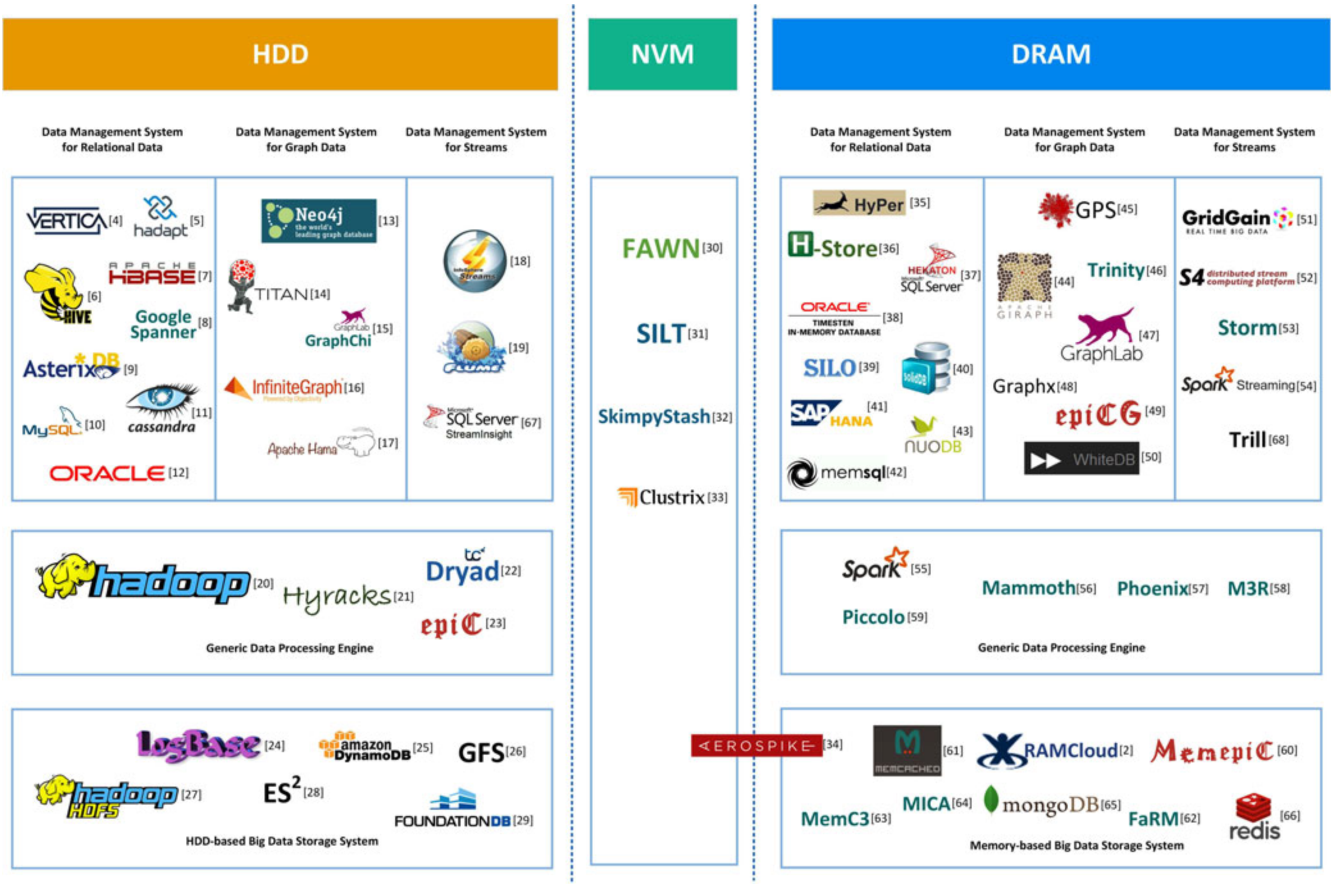

Fig. 1. The (Partial) landscape of disk-based and in-memory data management systems.

the memory, and even deploying an in-memory system where all data can be kept in memory. In-memory database systems have been studied in the past, as early as the 1980s [69], [70], [71], [72], [73]. However, recent advances in hardware technology have invalidated many of the earlier works and re-generated interests in hosting the whole database in memory in order to provide faster accesses and real-time analytics [35], [36], [55], [74], [75], [76]. Most commercial database vendors have recently introduced in-memory database processing to support large-scale applications completely in memory [37], [38], [40], [77]. Efficient in-memory data management is a necessity for various applications [78], [79]. Nevertheless, in-memory data management is still at its infancy, and is likely to evolve over the next few years.

In general, as summarized in Table 1, research in in-memory data management and processing focus on the following aspects for efficiency or enforcing ACID properties:

- Indexes. Although in-memory data access is extremely fast compared to disk access, an efficient index is still required for supporting point queries in order to avoid memory-intensive scan. Indexes designed for in-memory databases are quite different from traditional indexes designed for disk-based databases such as the B^+-tree, because traditional indexes mainly care about the I/O efficiency instead of memory and cache utilization. Hash-based indexes are commonly used in key-value stores, e.g., Memcached [61], Redis [66], RAMCloud [75], and can be further optimized for better cache utilization by reducing pointer chasing [63]. However hash-based indexes do not support range queries, which are crucial for data analytics and thus, tree-based indexes have also been proposed, such as T-Tree [80], Cache Sensitive Search Trees (CSS-Trees) [81], Cache Sensitive B^+-Trees (CSB$^+$-Trees) [82], Δ-Tree [83], BD-Tree [84], Fast Architecture Sensitive Tree (FAST) [85], Bw-tree [86] and Adaptive Radix Tree (ART) [87], some of which also consider pointer reduction.
- Data layouts. In-memory data layouts have a significant impact on the memory usage and cache utilization. Columnar layout of relational table facilitates scan-like queries/analytics as it can achieve good cache locality [41], [88], and can achieve better data compression [89], but is not optimal for OLTP queries that need to operate on the row level [74], [90]. It is also possible to have a hybrid of row and column layouts, such as PAX which organizes data by columns only within a page [91], and SAP HANA with multilayer stores consisting of several delta row/column stores and a main column store, which are merged periodically [74]. In addition, there are also proposals on handling the memory fragmentation problem, such as the slab-based allocator in Memcached [61], and log-structured data organization with periodical cleaning in RAMCloud [75], and better utilization of some hardware features (e.g., bit-level parallelism, SIMD), such as BitWeaving [92] and ByteSlice [93].
- Parallelism. In general, there are three levels of parallelism, i.e., data-level parallelism (e.g., bit-level

TABLE 1
Optimization Aspects on In-Memory Data Management and Processing

Aspects	Concerns	Related Work
Index	cache consciousness, time/space efficiency	T-Tree [80], CSS-Trees [81], CSB$^+$-Trees [82], Δ-Tree [83], BD-Tree [84], FAST [85], ART [87]
Data Layout	cache consciousness, space efficiency	PAX [91], columnar layout [41], [88], HANA Hybrid Store [74], slab allocator [61], log-structure [75]
Parallelism	linear scaling, partitioning	BitWeaving [92], bit-parallel aggregation [94], SIMD sorting [95], SIMD scanning [96], [97], multi-core join [98], distributed computing [2], [55], [99], [100]
Concurrency Control/ Transaction Management	overhead, correctness	virtual snapshot [35], lock-eliding [101],transactional memory [102], [103], PALM [104], LIL [105], VLL [106], OCC [39], [107], MVCC [108], [109], DGCC [110]
Query Processing	code locality, register temporal locality, time efficiency	stored procedure [111], JIT compilation [112], [113], join [98], [114], [115], [116], [117], [118], [119], [120], [121], [122], sort [95], [123], [124]
Fault Tolerance	durability, correlated failures, availability	Copyset [125], fast recovery [126], group commit and log coalescing [37], [127], NVM [128], [129], [130], command logging [131], adaptive logging [132], remote logging [2], [40]
Data Overflow	locality, paging strategy, hot/cold classification	Anti-caching [133], Hekaton Siberia [134], data compression [74], [89], [135], virtual memory management [136], pointer swizzling [137], UVMM [138]

parallelism, SIMD),[1] shared-memory scale-up parallelism (thread/process),[2] and shared-nothing scale-out parallelism (distributed computation). All three levels of parallelism can be exploited at the same time, as shown in Fig. 2. The bit-parallel algorithms fully unleash the intra-cycle parallelism of modern CPUs, by packing multiple data values into one CPU word, which can be processed in one single cycle [92], [94]. Intra-cycle parallelism performance can be proportional to the packing ratio, since it does not require any concurrency control (CC) protocol. SIMD instructions can improve vector-style computations greatly, which are extensively used in high-performance computing, and also in the database systems [95], [96], [97]. Scale-up parallelism can take advantage of the multi-core architecture of supercomputers or even commodity computers [36], [98], while scale-out parallelism is highly utilized in cloud/distributed computing [2], [55], [99]. Both scale-up and scale-out parallelisms require a good data partitioning strategy in order to achieve load balancing and minimize cross-partition coordination [100], [139], [140], [141].

- Concurrency control/transaction management. Concurrency control/transaction management becomes an extremely important performance issue in in-memory data management with the many-core systems. Heavy-weight mechanisms based on lock/semaphore greatly degrade the performance, due to its blocking-style scheme and the overhead caused by centralized lock manager and deadlock detection [142], [143]. Lightweight Intent Lock (LIL) [105] was proposed to maintain a set of lightweight counters in a global lock table instead of lock queues for intent locks. Very Lightweight Locking (VLL) [106] further simplifies the data structure by compressing all the lock states of one record into a pair of integers for partitioned databases. Another class of concurrency control is based on timestamp, where a predefined order is used to guarantee transactions' serializability [144], such as optimistic concurrency control (OCC) [39], [107] and multi-version concurrency control (MVCC) [108], [109]. Furthermore, H-Store [36], [101], seeks to eliminate concurrency control in single-partition transactions by partitioning the database beforehand based on a priori workload and providing one thread for each partition. HyPer [35] isolates OLTP and OLAP by *fork*-ing a child process (via *fork()* system call) for OLAP jobs based on the hardware-assisted virtual snapshot, which will never be modified. DGCC [110] is proposed to reduce the overhead of concurrency control by separating concurrency control from execution based on a dependency graph. Hekaton [104], [107] utilizes optimistic MVCC and lock-free data structures to achieve high concurrency efficiently. Besides, hardware transactional memory (HTM) [102], [103] is being increasingly exploited in concurrency control for OLTP.
- Query processing. Query processing is going through an evolution in in-memory databases. While the traditional Iterator-/Volcano-style model [145] facilitates easy combination of arbitrary operators, it generates a huge number of function calls (e.g., *next()*) which results in evicting the register contents. The poor code locality and frequent instruction miss-predictions further add to the overhead [112], [113]. Coarse-grained stored procedures (e.g., transaction-level) can be used to alleviate the problem [111], and dynamic compiling (Just-in-Time) is another approach to achieve better code and data locality [112], [113]. Performance gain can also be achieved by optimizing specific query operation

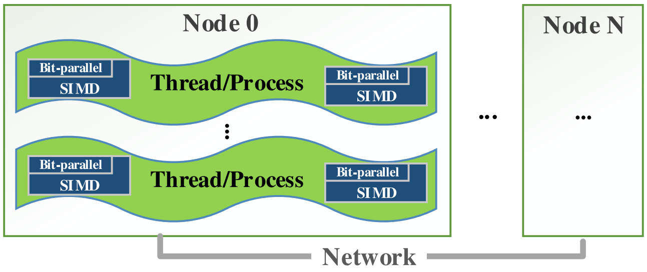

Fig. 2. Three levels of parallelism.

[1] Here data-level parallelism includes both bit-level parallelism achieved by data packing, and word-level parallelism achieved by SIMD.

[2] Accelerators such as GPGPU and Xeon Phi are also considered as shared-memory scale-up parallelism.

such as join [98], [114], [115], [116], [117], [118], [119], [120], [121], [122], and sort [95], [123], [124].

- Fault tolerance. DRAM is volatile, and fault-tolerance mechanisms are thus crucial to guarantee the data durability to avoid data loss and to ensure transactional consistency when there is a failure (e.g., power, software or hardware failure). Traditional write-ahead logging (WAL) is also the de facto approach used in in-memory database systems [35], [36], [37]. But the data volatility of the in-memory storage makes it unnecessary to apply any persistent undo logging [37], [131] or completely disables it in some scenarios [111]. To eliminate the potential I/O bottleneck caused by logging, group commit and log coalescing [37], [127], and remote logging [2], [40] are adopted to optimize the logging efficiency. New hardware technologies such as SSD and PCM are utilized to increase the I/O performance [128], [129], [130]. Recent studies proposed to use *command logging* [131], which logs only operations instead of the updated data, which is used in traditional ARIES logging [146]. [132] studies how to alternate between these two strategies adaptively. To speed up the recovery process, a consistent snapshot has to be checkpointed periodically [37], [147], and replicas should be dispersed in anticipation of correlated failures [125]. High availability is usually achieved by maintaining multiple replicas and stand-by servers [37], [148], [149], [150], or relying on fast recovery upon failure [49], [126]. Data can be further backuped onto a more stable storage such as GPFS [151], HDFS [27] and NAS [152] to further secure the data.
- Data overflow. In spite of significant increase in memory size and sharp drop in its price, it still cannot keep pace with the rapid growth of data in the Big Data era, which makes it essential to deal with data overflow where the size of the data exceeds the size of main memory. With the advancement of hardware, hybrid systems which incorporate non-volatile memories (NVMs) (e.g., SCM, PCM, SSD, Flash memory) [30], [31], [32], [118], [127], [153], [154], [155], [156] become a natural solution for achieving the speed. Alternatively, as in the traditional database systems, effective eviction mechanisms could be adopted to replace the in-memory data when the main memory is not sufficient. The authors of [133], [134], [157] propose to move cold data to disks, and [136] re-organizes the data in memory and relies on OS to do the paging, while [137] introduces pointer swizzling in database buffer pool management to alleviate the overhead caused by traditional databases in order to compete with the completely re-designed in-memory databases. UVMM [138] taps onto a hybrid of hardware-assisted and semantics-aware access tracking, and non-blocking kernel I/O scheduler, to facilitate efficient memory management. Data compression has also been used to alleviate the memory usage pressure [74], [89], [135].

The focus of the survey is on large-scale in-memory data management and processing strategies, which can be broadly grouped into two categories, i.e., in-memory data storage systems and in-memory data processing systems. Accordingly, the remaining sections are organized as follows. Section 2 presents some background on in-memory data management. We elaborate in-memory data storage systems, including relational databases and NoSQL databases in Section 3, and in-memory data processing systems, including in-memory batch processing and real-time stream processing in Section 4. As a summary, we present a qualitative comparison of the in-memory data management systems covered in this survey in Section 5. Finally, we discuss some research opportunities in Section 6, and conclude in Section 7.

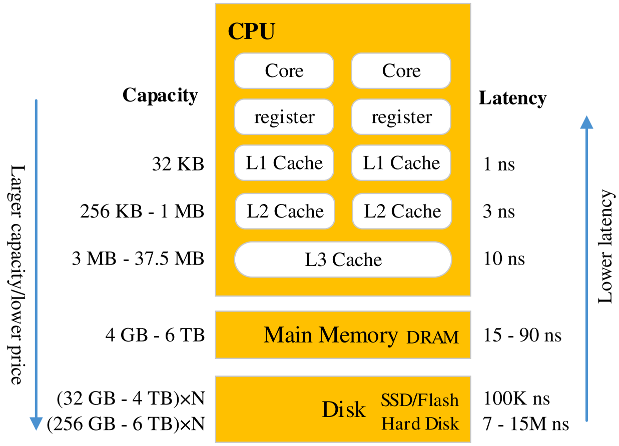

Fig. 3. Memory hierarchy.

2 Core Technologies for In-Memory Systems

In this section, we shall introduce some concepts and techniques that are important for efficient in-memory data management, including memory hierarchy, non-uniform memory access (NUMA), transactional memory, and non-volatile random access memory (NVRAM). These are the basics on which the performance of in-memory data management systems heavily rely.

2.1 Memory Hierarchy

The memory hierarchy is defined in terms of access latency and the logical distance to the CPU. In general, it consists of registers, caches (typically containing L1 cache, L2 cache and L3 cache), main memory (i.e., RAM) and disks (e.g., hard disk, flash memory, SSD) from the highest performance to the lowest. Fig. 3 depicts the memory hierarchy, and respective component's capacity and access latency [158], [159], [160], [161], [162], [163], [164], [165], [166]. It shows that data access to the higher layers is much faster than to the lower layers, and each of these layers will be introduced in this section.

In modern architectures, data cannot be processed by CPU unless it is put in the registers. Thus, data that is about to be processed has to be transmitted through each of the memory layers until it reaches the registers. Consequently, each upper layer serves as a cache for the underlying lower layer to reduce the latency for repetitive data accesses. The performance of a data-intensive program highly depends on the utilization of the memory hierarchy. How to achieve both good spatial and temporal locality is usually what matters the most in the efficiency optimization. In particular, spatial locality assumes that the adjacent data is more likely to be accessed together, whereas temporal locality refers to the observation that it is likely that an item will be accessed again in the near future. We will introduce some important efficiency-related properties of different memory layers respectively.

2.1.1 Register

A processor register is a small amount of storage within a CPU, on which machine instructions can manipulate directly. In a normal instruction, data is first loaded from the lower memory layers into registers where it is used for arithmetic or test operation, and the result is put back into another register, which is then often stored back into main memory, either by the same instruction or a subsequent one. The length of a register is usually equal to the word length of a CPU, but there also exist double-word, and even wider registers (e.g., 256 bits wide YMMX registers in Intel Sandy Bridge CPU micro architecture), which can be used for single instruction multiple data (SIMD) operations. While the number of registers depends on the architecture, the total capacity of registers is much smaller than that of the lower layers such as cache or memory. However, accessing data from registers is very much faster.

2.1.2 Cache

Registers play the role as the storage containers that CPU uses to carry out instructions, while caches act as the bridge between the registers and main memory due to the high transmission delay between the registers and main memory. Cache is made of high-speed static RAM (SRAM) instead of slower and cheaper dynamic RAM (DRAM) that usually forms the main memory. In general, there are three levels of caches, i.e., L1 cache, L2 cache and L3 cache (also called last level cache—LLC), with increasing latency and capacity. L1 cache is further divided into data cache (i.e., L1-dcache) and instruction cache (i.e., L1-icache) to avoid any interference between data access and instruction access. We call it a cache hit if the requested data is in the cache; otherwise it is called a cache miss.

Cache is typically subdivided into fixed-size logical cache lines, which are the atomic units for transmitting data between different levels of caches and between the last level cache and main memory. In modern architectures, a cache line is usually 64 bytes long. By filling the caches per cache line, spatial locality can be exploited to improve performance. The mapping between the main memory and the cache is determined by several strategies, i.e., direct mapping, N-way set associative, and fully associative. With direct mapping, each entry (a cache line) in the memory can only be put in one place in the cache, which makes addressing faster. Under fully associative strategy, each entry can be put in any place, which offers fewer cache misses. The N-way associative strategy is a compromise between direct mapping and fully associative—it allows each entry in the memory to be in any of N places in the cache, which is called a cache set. N-way associative is often used in practice, and the mapping is deterministic in terms of cache sets.

In addition, most architectures usually adopt a least-recently-used (LRU) replacement strategy to evict a cache line when there is not enough room. Such a scheme essentially utilizes temporal locality for enhancing performance. As shown in Fig. 3, the latency to access cache is much shorter than the latency to access main memory. In order to gain good CPU performance, we have to guarantee high cache hit rate so that high-latency memory accesses are reduced. In designing an in-memory management system, it is important to exploit the properties of spatial and temporal locality of caches. For examples, it would be faster to access memory sequentially than randomly, and it would also be better to keep a frequently-accessed object resident in the cache. The advantage of sequential memory access is reinforced by the prefetching strategies of modern CPUs.

2.1.3 Main Memory and Disks

Main memory is also called internal memory, which can be directly addressed and possibly accessed by the CPU, in contrast to external devices such as disks. Main memory is usually made of volatile DRAM, which incurs equivalent latency for random accesses without the effect of caches, but will lose data when power is turned off. Recently, DRAM becomes inexpensive and large enough to make an in-memory database viable.

Even though memory becomes the new disk [1], the volatility of DRAM makes it a common case that disks[3] are still needed to backup data. Data transmission between main memory and disks is conducted in units of pages, which makes use of data spatial locality on the one hand and minimizes the performance degradation caused by the high-latency of disk seek on the other hand. A page is usually a multiple of disk sectors[4] which is the minimum transmission unit for hard disk. In modern architectures, OS usually keeps a buffer which is part of the main memory to make the communication between the memory and disk faster.[5] The buffer is mainly used to bridge the performance gap between the CPU and the disk. It increases the disk I/O performance by buffering the writes to eliminate the costly disk seek time for every write operation, and buffering the reads for fast answer to subsequent reads to the same data. In a sense, the buffer is to the disk as the cache is to the memory. And it also exposes both spatial and temporal locality, which is an important factor in handling the disk I/O efficiently.

2.2 Memory Hierarchy Utilization

This section reviews related works from three perspective—register-conscious optimization, cache-conscious optimization and disk I/O optimization.

2.2.1 Register-Conscious Optimization

Register-related optimization usually matters in compiler and assembly language programming, which requires

[3] Here we refer to hard disks.

[4] A page in the modern file system is usually 4 KB. Each disk sector of hard disks is traditionally 512 bytes.

[5] The kernel buffer is also used to buffer data from other block I/O devices that transmit data in fixed-size blocks.

utilizing the limited number of registers efficiently. There have been some criticisms on the traditional iterator-style query processing mechanisms for in-memory databases recently as it usually results in poor code and data locality [36], [112], [167]. HyPer uses low level virtual machine (LLVM) compiler framework [167] to translate a query into machine code dynamically, which achieves good code and data locality by avoiding recursive function calls as much as possible and trying to keep the data in the registers as long as possible [112].

SIMD is available in superscalar processors, which exploits data-level parallelism with the help of wide registers (e.g., 256 bits). SIMD can improve the performance significantly especially for vector-style computation, which is very common in Big Data analytics jobs [95], [96], [97], [112], [168].

2.2.2 Cache-Conscious Optimization

Cache utilization is becoming increasingly important in modern architectures. Several workload characterization studies provide detailed analysis of the time breakdown in the execution of DBMSs on a modern processor, and report that DBMSs suffer from high memory-related processor stalls when running on modern architectures. This is caused by a huge amount of data cache misses [169], which account for 50-70 percent for OLTP workloads [91], [170] to 90 percent for DSS workloads [91], [171], of the total memory-related stall. In a distributed database, instruction cache misses are another main source of performance degradation due to a large number of TCP network I/Os [172].

To utilize the cache more efficiently, some works focus on re-organizing the data layout by grouping together all values of each attribute in an N-ary Storage Model (NSM) page [91] or using a Decomposition Storage Model (DSM) [173] or completely organizing the records in a column store [41], [90], [174], [175]. This kind of optimization favors OLAP workload which typically only needs a few columns, but has a negative impact on intra-tuple cache locality [176]. There are also other techniques to optimize cache utilization for the primary data structure, such as compression [177] and coloring [178].

In addition, for memory-resident data structures, various cache-conscious indexes have been proposed such as Cache Sensitive Search Trees [81], Cache Sensitive B^+-Trees [82], Fast Architecture Sensitive Trees [85], and Adaptive Radix Trees [87]. Cache-conscious algorithms have also been proposed for basic operations such as sorting (e.g., burst sort) [179] and joining [98], [114].

In summary, to optimize cache utilization, the following important factors should be taken into consideration:

- Cache line length. This characteristic exposes spatial locality, meaning that it would be more efficient to access adjacent data.
- Cache size. It would also be more efficient to keep frequently-used data within at least L3 cache size.
- Cache replacement policy. One of the most popular replacement policies is LRU, which replaces the least recently used cache line upon a cache miss. The temporal locality should be exploited in order to get high performance.

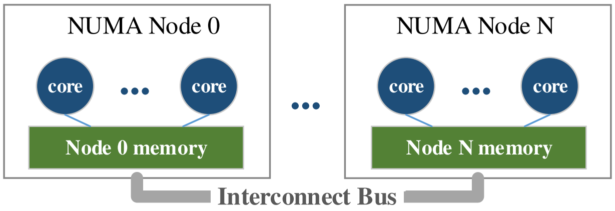

Fig. 4. NUMA topology.

2.3 Non-Uniform Memory Access

Non-uniform memory access is an architecture of the main memory subsystem where the latency of a memory operation depends on the relative location of the processor that is performing memory operations. Broadly, each processor in a NUMA system has a *local memory* that can be accessed with minimal latency, but can also access at least one *remote memory* with longer latency, which is illustrated in Fig. 4.

The main reason for employing NUMA architecture is to improve the main memory bandwidth and total memory size that can be deployed in a server node. NUMA allows the clustering of several memory controllers into a single server node, creating several *memory domains*. Although NUMA systems were deployed as early as 1980s in specialized systems [185], since 2008 all Intel and AMD processors incorporate one memory controller. Thus, most contemporary multi-processor systems are NUMA; therefore, NUMA-awareness is becoming a mainstream challenge.

In the context of data management systems, current research directions on NUMA-awareness can be broadly classified into three categories:

- partitioning the data such that memory accesses to remote NUMA domains are minimized [115], [186], [187], [188], [189];
- managing NUMA effects on latency-sensitive workloads such as OLTP transactions [190], [191];
- efficient data shuffling across NUMA domains [192].

2.3.1 Data Partitioning

Partitioning the working set of a database has long been used to minimize data transfers across different data domains, both within a compute node and across compute nodes. Bubba [186] is an example of an earlier parallel database system that uses a shared-nothing architecture to scale to hundreds of compute nodes. It partitions the data using a hash- or range-based approach and always performs the analytics operations only in the nodes that contain the relevant partitions. Gamma [187] is another example that was designed to operate on a complex architecture with an Intel iPSC/2 hypercube with 32 processors and 32 disk drives. Like Bubba, Gamma partitions the data across multiple disk drives and uses a hash-based approach to implement join and aggregate operations on top of the partitioned data. The partitioned data and execution provide the partitioned parallelism [193]. With NUMA systems becoming mainstream, many research efforts have started to address NUMA issues explicitly, rather than just relying on data partitioning. Furthermore, modern systems

TABLE 2
Comparison between STM and HTM

	Performance Penalty	Hardware Support	Transaction Size	Implementations
STM	Much	Atomic Operation	Large	TinySTM [180], Clojure [181], Haskell [182]
HTM	No or Little	Cache, Bus Protocol	Small	Intel TSX [183], AMD ASF [184]

have increasingly larger number of cores. The recent change in memory topology and processing power have indeed attracted interest in re-examining traditional processing methods in the context of NUMA.

A new sort-merge technique for partitioning the join operation was proposed in [115] to take advantage of NUMA systems with high memory bandwidth and many cores. In contrast to hash join and classical sort-merge join, the parallel sort-merge strategy parallelizes also the final merge step, and naturally operates on local memory partitions. This is to take advantage of both the multi-core architecture and the large local memory bandwidth that most NUMA systems have.

Partitioning the database index was proposed for the Buzzard system [188]. Index operations typically incur frequent pointer chasing during the traversal of a tree-based index. In a NUMA system, these pointer operations might end up swinging from one memory domain to another. To address this problem, Buzzard proposes a NUMA-aware index that partitions different parts of a prefix tree-based index across different NUMA domains. Furthermore, Buzzard uses a set of dedicated worker threads to access each memory domain. This guarantees that threads only access their local memory during index traversal and further improves the performance by using only local comparison and swapping operations instead of expensive locking.

Partitioning both the input data and query execution was proposed in [189]. In contrast to plan-driven query execution, a fine-grained runtime task scheduling, termed "morsel query execution" was proposed. The morsel-driven query processing strategy dynamically assembles small pieces of input data, and executes them in a pipelined fashion by assigning them to a pool of worker threads. Due to this fine-grained control over the parallelism and the input data, morsel-driven execution is aware of the data locality of each operator, and can schedule their execution on local memory domains.

2.3.2 OLTP Latency

Since NUMA systems have heterogeneous access latency, they pose a challenging problem to OLTP transactions which are typically very sensitive to latency. The performance of NUMA-unaware OLTP deployments on NUMA systems is profiled in [190], where many of these systems are deemed to have achieved suboptimal and unpredictable performance. To address the needs for a NUMA-aware OLTP system, the paper proposes "hardware islands", in which it treats each memory domain as a logical node, and uses UNIX domain sockets to communicate among the NUMA memory domains of a physical node. The recently proposed ATraPos [191] is an adaptive transaction processing system that has been built based on this principle.

2.3.3 Data Shuffling

Data shuffling in NUMA systems aims to transfer the data across the NUMA domains as efficiently as possible, by saturating the transfer bandwidth of the NUMA interconnect network. A NUMA-aware coordinated ring-shuffling method was proposed in [192]. To shuffle the data across NUMA domains as efficiently as possible, the proposed approach forces the threads across NUMA domains to communicate in a coordinated manner. It divides the threads into an inner ring and an outer ring and performs communication in a series of rounds, during which the inner ring remains fixed while the outer ring rotates. This rotation guarantees that all threads across the memory domains will be paired based on a predictable pattern, and thus all the memory channels of the NUMA interconnect are always busy. Compared to the naive method of shuffling the data around the domains, this method improves the transfer bandwidth by a factor of four, when using a NUMA system with four domains.

2.4 Transactional Memory

Transactional memory is a concurrency control mechanism for shared memory access, which is analogous to atomic database transactions. The two types of transactional memory, i.e., software transactional memory (STM) and hardware transactional memory (HTM), are compared in Table 2. STM causes a significant slowdown during execution and thus has limited practical application [194], while HTM has attracted new attention for its efficient hardware-assisted atomic operations/transactions, since Intel introduced it in its mainstream Haswell microarchitecture CPU [102], [103]. Haswell HTM is implemented based on cache coherency protocol. In particular, L1 cache is used as a local buffer to mark all transactional read/write on the granularity of cache lines. The propagation of changes to other caches or main memory is postponed until the transaction commits, and write/write and read/write conflicts are detected using the cache coherency protocol [195]. This HTM design incurs almost no overhead for transaction execution, but has the following drawbacks, which make HTM only suitable for small and short transactions.

- The transaction size is limited to the size of L1 data cache, which is usually 32 KB. Thus it is not possible to simply execute a database transaction as one monolithic HTM transaction.
- Cache associativity makes it more prone to false conflicts, because some cache lines are likely to go to the same cache set, and an eviction of a cache line leads to abort of the transaction, which cannot be resolved by restarting the transaction due to the determinism of the cache mapping strategy (refer to Section 2.1.2).
- HTM transactions may be aborted due to interrupt events, which limits the maximum duration of HTM transactions.

There are two instruction sets for Haswell HTM in Transactional Synchronization Extensions (TSX),[6] i.e., Hardware Lock Ellison (HLE) and Restricted Transactional Memory (RTM). HLE allows optimistic execution of a transaction by eliding the lock so that the lock is free to other threads, and restarting it if the transaction failed due to data race, which mostly incurs no locking overhead, and also provides backward compatibility with processors without TSX. RTM is a new instruction set that provides the flexibility to specify a fallback code path after a transaction aborts. The author [102] exploits HTM based on HLE, by dividing a database transaction into a set of relatively small HTM transactions with timestamp ordering (TSO) concurrency control and minimizing the false abort probability via data/index segmentation. RTM is utilized in [103], which uses a three-phase optimistic concurrency control to coordinate a whole database transaction, and protects single data read (to guarantee consistency of sequence numbers) and validate/write phases using RTM transactions.

2.5 NVRAM

Newly-emerging non-volatile memory raises the prospect of persistent high-speed memory with large capacity. Examples of NVM include both NAND/NOR flash memory with block-granularity addressability, and non-volatile random access memory with byte-granularity addressability.[7] Flash memory/SSD has been widely used in practice, and attracted a significant amount of attention in both academia and industry [32], [33], but its block-granularity interface, and expensive "erase" operation make it only suitable to act as the lower-level storage, such as replacement of hard disk [30], [32], or disk cache [198]. Thus, in this survey, we only focus on NVRAMs that have byte addressability and comparable performance with DRAM, and can be brought to the main memory layer or even the CPU cache layer.

Advanced NVRAM technologies, such as phase change memory [199], Spin-Transfer Torque Magnetic RAM (STT-MRAM) [200], and Memristors [201], can provide orders of magnitude better performance than either conventional hard disk or flash memory, and deliver excellent performance on the same order of magnitude as DRAM, but with persistent writes [202]. The read latency of PCM is only two-five times slower than DRAM, and STT-MRAM and Memristor could even achieve lower access latency than DRAM [118], [128], [129], [203]. With proper caching, carefully architected PCM could also match DRAM performance [159]. Besides, NVRAM is speculated to have much higher storage density than DRAM, and consume much less power [204]. Although NVRAM is currently only available in small sizes, and the cost per bit is much higher than that of hard disk or flash memory or even DRAM, it is estimated that by the next decade, we may have a single PCM with 1 TB and Memristor with 100 TB, at price close to the enterprise hard disk [128], [129]. The advent of NVRAM offers an intriguing opportunity to revolutionize the data management and re-think the system design.

It has been shown that simply replacing disk with NVRAM is not optimal, due to the high overhead from the cumbersome file system interface (e.g., file system cache and costly system calls), block-granularity access and high economic cost, etc. [127], [130], [205]. Instead, NVRAM has been proposed to be placed side-by-side with DRAM on the memory bus, available to ordinary CPU `loads` and `stores`, such that the physical address space can be divided between volatile and non-volatile memory [205], or be constituted completely by non-volatile memory [155], [206], [207], equipped with fine-tuned OS support [208], [209]. Compared to DRAM, NVRAM exhibits its distinct characteristics, such as limited endurance, write/read asymmetry, uncertainty of ordering and atomicity [128], [205]. For example, the write latency of PCM is more than one order of magnitude slower than its read latency [118]. Besides, there is no standard mechanisms/protocols to guarantee the ordering and atomicity of NVRAM writes [128], [205]. The endurance problem can be solved by wear-leveling techniques in the hardware or middleware levels [206], [207], [210], which can be easily hidden from the software design, while the read/write asymmetry, and ordering and atomicity of writes, must be taken into consideration in the system/algorithm design [204].

Promisingly, NVRAM can be architected as the main memory in general-purpose systems with well-designed architecture [155], [206], [207]. In particular, longer write latency of PCM can be solved by data comparison writes [211], partial writes [207], or specialized algorithms/structures that trade writes for reads [118], [204], [212], which can also alleviate the endurance problem. And current solutions to the write ordering and atomicity problems are either relying on some newly-proposed hardware primitives, such as atomic 8-byte writes and epoch barriers [129], [205], [212], or leveraging existing hardware primitives, such as cache modes (e.g., write-back, write-combining), memory barriers (e.g., `mfence`), cache line flush (e.g., `clflush`) [128], [130], [213], which, however, may incur non-trivial overhead. General libraries and programming interfaces are proposed to expose NVRAM as a persistent heap, enabling NVRAM adoption in an easy-to-use manner, such as NV-heaps [214], Mnemosyne [215], NVMalloc [216], and recovery and durable structures [213], [217]. In addition, file system support enables a transparent utilization of NVRAM as a persistent storage, such as Intel's PMFS [218], BPFS [205], FRASH [219], ConquestFS [220], SCMFS [221], which also take advantage of NVRAM's byte addressability.

Besides, specific data structures widely used in databases, such as B-Tree [212], [217], and some common query processing operators, such as sort and join [118], are starting to adapt to and take advantage of NVRAM properties. Actually, the favorite goodies brought to databases by NVRAM is its non-volatility property, which facilitates a more efficient logging and fault tolerance mechanisms [127], [128], [129], [130]. But write atomicity and deterministic orderings should be guaranteed and achieved efficiently via carefully designed algorithms, such as group

[6] Intel disabled its TSX feature on Haswell, Haswell-E, Haswell-EP and early Broadwell CPUs in August 2014 due to a bug. Currently Intel only provides TSX on Intel Core M CPU with Broadwell architecture, and the newly-released Xeon E7 v3 CPU with Haswell-EX architecture [196], [197].

[7] NVM and NVRAM usually can be used exchangeably without much distinction. NVRAM is also referred to as Storage-Class Memory (SCM), Persistent Memory (PM) or Non-Volatile Byte-addressable Memory (NVBM).

commit [127], passive group commit [128], two-step logging (i.e., populating the log entry in DRAM first and then flushing it to NVRAM) [130]. Also the centralized logging bottleneck should be eliminated, e.g., via distributed logging [128], decentralized logging [130]. Otherwise the high performance brought by NVRAM would be degraded by the legacy software overhead (e.g., centention for the centralized log).

3 In-Memory Data Storage Systems

In this section, we introduce some in-memory databases, including both relational and NoSQL databases. We also cover a special category of in-memory storage system, i.e., cache system, which is used as a cache between the application server and the underlying database. In most relational databases, both OLTP and OLAP workloads are supported inherently. The lack of data analytics operations in NoSQL databases results in an inevitable data transmission cost for data analytics jobs [172].

3.1 In-Memory Relational Databases

Relational databases have been developed and enhanced since 1970s, and the relational model has been dominating in almost all large-scale data processing applications since early 1990s. Some widely used relational databases include Oracle, IBM DB2, MySQL and PostgreSQL. In relational databases, data is organized into tables/relations, and ACID properties are guaranteed. More recently, a new type of relational databases, called NewSQL (e.g., Google Spanner [8], H-Store [36]) has emerged. These systems seek to provide the same scalability as NoSQL databases for OLTP while still maintaining the ACID guarantees of traditional relational database systems.

In this section, we focus on in-memory relational databases, which have been studied since 1980s [73]. However, there has been a surge in interests in recent years [222]. Examples of commercial in-memory relational databases include SAP HANA [77], VoltDB [150], Oracle TimesTen [38], SolidDB [40], IBM DB2 with BLU Acceleration [223], [224], Microsoft Hekaton [37], NuoDB [43], eXtremeDB [225], Pivotal SQLFire [226], and MemSQL [42]. There are also well known research/open-source projects such as H-Store [36], HyPer [35], Silo [39], Crescando [227], HYRISE [176], and MySQL Cluster NDB [228].

3.1.1 H-Store / VoltDB

H-Store [36], [229] or its commercial version VoltDB [150] is a distributed row-based in-memory relational database targeted for high-performance OLTP processing. It is motivated by two observations: first, certain operations in traditional disk-based databases, such as logging, latching, locking, B-tree and buffer management operations, incur substantial amount of the processing time (more than 90 percent) [222] when ported to in-memory databases; second, it is possible to re-design in-memory database processing so that these components become unnecessary. In H-Store, most of these “heavy” components are removed or optimized, in order to achieve high-performance transaction processing.

Transaction execution in H-Store is based on the assumption that all (at least most of) the templates of transactions are known in advance, which are represented as a set of compiled stored procedures inside the database. This reduces the overhead of transaction parsing at runtime, and also enables pre-optimizations on the database design and light-weight logging strategy [131]. In particular, the database can be more easily partitioned to avoid multi-partition transactions [140], and each partition is maintained by a *site*, which is single-threaded daemon that processes transactions serially and independently without the need for heavy-weight concurrency control (e.g., lock) in most cases [101]. Next, we will elaborate on its transaction processing, data overflow and fault-tolerance strategies.

Transaction processing. Transaction processing in H-Store is conducted on the partition/site basis. A site is an independent transaction processing unit that executes transactions sequentially, which makes it feasible only if a majority of the transactions are single-sited. This is because if a transaction involves multiple partitions, all these sites are sequentialized to process this distributed transaction in collaboration (usually 2PC), and thus cannot process transactions independently in parallel. H-Store designs a skew-aware partitioning model—*Horticulture* [140]—to automatically partition the database based on the database schema, stored procedures and a sample transaction workload, in order to minimize the number of multi-partition transactions and meanwhile mitigate the effects of temporal skew in the workload. *Horticulture* employs the large-neighborhood search (LNS) approach to explore potential partitions in a guided manner, in which it also considers read-only table replication to reduce the transmission cost of frequent remote access, secondary index replication to avoid broadcasting, and stored procedure routing attributes to allow an efficient routing mechanism for requests.

The *Horticulture* partitioning model can reduce the number of multi-partition transactions substantially, but not entirely. The concurrency control scheme must therefore be able to differentiate single partition transactions from multi-partition transactions, such that it does not incur high overhead where it is not needed (i.e., when there are only single-partition transactions). H-Store designs two low overhead concurrency control schemes, i.e., light-weight locking and speculative concurrency control [101]. Light-weight locking scheme reduces the overhead of acquiring locks and detecting deadlock by allowing single-partition transactions to execute without locks when there are no active multi-partition transactions. And speculative concurrency control scheme can proceed to execute queued transactions speculatively while waiting for 2PC to finish (precisely after the last fragment of a multi-partition transaction has been executed), which outperforms the locking scheme as long as there are few aborts or few multi-partition transactions that involve multiple rounds of communication.

In addition, based on the partitioning and concurrency control strategies, H-Store utilizes a set of optimizations on transaction processing, especially for workload with interleaving of single- and multi-transactions. In particular, to process an incoming transaction (a stored procedure with concrete parameter values), H-Store uses a Markov model-based approach [111] to determine the necessary optimizations by predicting the most possible execution path and the set of partitions that it may access. Based on these predictions, it applies four major optimizations accordingly, namely (1) execute the transaction at the node with the partition that it will access the most; (2) lock only the partitions that the transaction accesses; (3) disable undo logging for non-aborting transactions; (4) speculatively commit the transaction at partitions that it no longer needs to access.

Data overflow. While H-Store is an in-memory database, it also utilizes a technique, called *anti-caching* [133], to allow data bigger than the memory size to be stored in the database, without much sacrifice of performance, by moving cold data to disk in a transactionally-safe manner, on the tuple-level, in contrast to the page-level for OS virtual memory management. In particular, to evict cold data to disk, it pops the least recently used tuples from the database to a set of block buffers that will be written out to disks, updates the evicted table that keeps track of the evicted tuples and all the indexes, via a special eviction transaction. Besides, non-blocking fetching is achieved by simply aborting the transaction that accesses evicted data and then restarting it at a later point once the data is retrieved from disks, which is further optimized by executing a pre-pass phase before aborting to determine all the evicted data that the transaction needs so that it can be retrieved in one go without multiple aborts.

Fault tolerance. H-Store uses a hybrid of fault-tolerance strategies, i.e., it utilizes a replica set to achieve high availability [36], [150], and both checkpointing and logging for recovery in case that all the replicas are lost [131]. In particular, every partition is replicated to k sites, to guarantee k-safety, i.e., it still provides availability in case of simultaneous failure of k sites. In addition, H-Store periodically checkpoints all the committed database states to disks via a distributed transaction that puts all the sites into a copy-on-write mode, where updates/deletes cause the rows to be copied to a shadow table. Between the interval of two checkpointings, command logging scheme [131] is used to guarantee the durability by logging the commands (i.e., transaction/stored procedure identifier and parameter values), in contrast to logging each operation (insert/delete/update) performed by the transaction as the traditional ARIES physiological logging does [146]. Besides, memory-resident undo log can be used to support rollback for some abort-able transactions. It is obvious that command logging has a much lower runtime overhead than physiological logging as it does less work at runtime and writes less data to disk, however, at the cost of an increased recovery time. Therefore, command logging scheme is more suitable for short transactions where node failures are not frequent.

3.1.2 Hekaton

Hekaton [37] is a memory-optimized OLTP engine fully integrated into Microsoft SQL server, where Hekaton tables[8] and regular SQL server tables can be accessed at the same time, thereby providing much flexibility to users. It is designed for high-concurrency OLTP, with utilization of lock-free or latch-free data structures (e.g., latch-free hash and range indexes) [86], [230], [231], and an optimistic MVCC technique [107]. It also incorporates a framework, called Siberia [134], [232], [233], to manage hot and cold data differently, equipping it with the capacity to handle Big Data both economically and efficiently. Furthermore, to relieve the overhead caused by interpreter-based query processing mechanism in traditional databases, Hekaton adopts the compile-once-and-execute-many-times strategy, by compiling SQL statements and stored procedures into C code first, which will then be converted into native machine code [37]. Specifically, an entire query plan is collapsed into a single function using *label*s and *goto*s for code sharing, thus avoiding the costly argument passing between functions and expensive function calls, with the fewest number of instructions in the final compiled binary. In addition, durability is ensured in Hekaton by using incremental checkpoints, and transaction logs with log merging and group commit optimizations, and availability is achieved by maintaining highly available replicas [37]. We shall next elaborate on its concurrency control, indexing and hot/cold data management.

Multi-version concurrency control. Hekaton adopts optimistic MVCC to provide transaction isolation without locking and blocking [107]. Basically, a transaction is divided into two phases, i.e., normal processing phase where the transaction never blocks to avoid expensive context switching, and validation phase where the visibility of the read set and phantoms are checked,[9] and then outstanding commit dependencies are resolved and logging is enforced. Specifically, updates will create a new version of record rather than updating the existing one in place, and only records whose valid time (i.e., a time range denoted by start and end timestamps) overlaps the logical read time of the transaction are visible. The uncommitted records are allowed to be speculatively read/ignored/updated if those records have reached the validation phase, in order to advance the processing, and not to block during the normal processing phase. But speculative processing enforces commit dependencies, which may cause cascaded abort and must be resolved before committing. It utilizes atomic operations for updating on the valid time of records, visibility checking and conflict detection, rather than locking. Finally, a version of a record is garbage-collected (GC) if it is no longer visible to any active transaction, in a cooperative and parallel manner. That is, the worker threads running the transaction workload can remove the garbage when encountering it, which also naturally provides a parallel GC mechanism. Garbage in the never-accessed area will be collected by a dedicated GC process.

Latch-free Bw-Tree. Hekaton proposes a latch-free B-tree index, called Bw-tree [86], [230], which uses delta updates to make state changes, based on atomic compare-and-swap (CAS) instructions and an elastic virtual page[10] management subsystem—LLAMA [231]. LLAMA provides a

[8] Hekaton tables are declared as "memory optimized" in SQL server, to distinguish with normal tables.

[9] Some of validation checks are not necessary, depending on the isolation levels. For example, no validation is required for *read committed* and *snapshot isolation*, and only read set visibility check is needed for *repeatable read*. Both checks are required only for serializable isolation.

[10] The virtual page here does not mean that used by OS. There is no hard limit on the page size, and pages grow by prepending "delta pages" to the base page.

virtual page interface, on top of which logical page IDs (PIDs) are used by Bw-tree instead of pointers, which can be translated into physical address based on a mapping table. This allows the physical address of a Bw-tree node to change on every update, without requiring the address change to be propagated to the root of the tree.

In particular, delta updates are performed by prepending the update delta page to the prior page and atomically updating the mapping table, thus avoiding update-in-place which may result in costly cache invalidation especially on multi-socket environment, and preventing the in-use data from being updated simultaneously, enabling latch-free access. The delta update strategy applies to both leaf node update achieved by simply prepending a delta page to the page containing the prior leaf node, and structure modification operations (SMO) (e.g., node split and merge) by a series of non-blocking cooperative and atomic delta updates, which are participated by any worker thread encountering the uncompleted SMO [86]. Delta pages and base page are consolidated in a later pointer, in order to relieve the search efficiency degradation caused by the long chain of delta pages. Replaced pages are reclaimed by the epoch mechanism [234], to protect data potentially used by other threads, from being freed too early.

Siberia in Hekaton. Project Siberia [134], [232], [233] aims to enable Hekaton to automatically and transparently maintain cold data on the cheaper secondary storage, allowing more data fit in Hekaton than the available memory. Instead of maintaining an LRU list like H-Store Anti-Caching [133], Siberia performs offline classification of hot and cold data by logging tuple accesses first, and then analyzing them offline to predict the top K hot tuples with the highest estimated access frequencies, using an efficient parallel classification algorithm based on exponential smoothing [232]. The record access logging method incurs less overhead than an LRU list in terms of both memory and CPU usage. In addition, to relieve the memory overhead caused by the evicted tuples, Siberia does not store any additional information in memory about the evicted tuples (e.g., keys in the index, evicted table) other than the multiple variable-size Bloom filters [235] and adaptive range filters [233] that are used to filter the access to disk. Besides, in order to make it transactional even when a transaction accesses both hot and cold data, it transactionally coordinates between hot and cold stores so as to guarantee consistency, by using a durable update memo to temporarily record notices that specify the current status of cold records [134].

3.1.3 HyPer/ScyPer

HyPer [35], [236], [237] or its distributed version ScyPer [149] is designed as a hybrid OLTP and OLAP high performance in-memory database with utmost utilization of modern hardware features. OLTP transactions are executed sequentially in a lock-less style which is first advocated in [222] and parallelism is achieved by logically partitioning the database and admitting multiple partition-constrained transactions in parallel. It can yield an unprecedentedly high transaction rate, as high as 100,000 per second [35]. The superior performance is attributed to the low latency of data access in in-memory databases, the effectiveness of the space-efficient Adaptive Radix Tree [87] and the use of stored transaction procedures. OLAP queries are conducted on a consistent snapshot achieved by the virtual memory snapshot mechanism based on hardware-supported shadow pages, which is an efficient concurrency control model with low maintenance overhead. In addition, HyPer adopts a dynamic query compilation scheme, i.e., the SQL queries are first compiled into assembly code [112], which can then be executed directly using an optimizing Just-in-Time (JIT) compiler provided by LLVM [167]. This query evaluation follows a data-centric paradigm by applying as many operations on a data object as possible, thus keeping data in the registers as long as possible to achieve register-locality.

The distributed version of HyPer, i.e., ScyPer [149], adopts a primary-secondary architecture, where the primary node is responsible for all the OLTP requests and also acts as the entry point for OLAP queries, while secondary nodes are only used to execute the OLAP queries. To synchronize the updates from the primary node to the secondary nodes, the logical redo log is multicast to all secondary nodes using Pragmatic General Multicast protocol (PGM), where the redo log is replayed to catch up with the primary. Further, the secondary nodes can subscribe to specific partitions, thus allowing the provisioning of secondary nodes for specific partitions and enabling a more flexible multi-tenancy model. In the current version of ScyPer, there is only one primary node, which holds all the data in memory, thus bounding the database size or the transaction processing power to one server. Next, we will elaborate on HyPer's snapshot mechanism, register-conscious compilation scheme and the ART indexing.

Snapshot in HyPer. HyPer constructs a consistent snapshot by *fork*-ing a child process (via *fork()* system call) with its own copied virtual memory space [35], [147], which involves no software concurrency control mechanism but the hardware-assisted virtual memory management with little maintenance overhead. By *fork*-ing a child process, all the data in the parent process is virtually "copied" to the child process. It is however quite light-weight as the copy-on-write mechanism will trigger the real copying only when some process is trying to modify a page, which is achieved by the OS and the memory management unit (MMU). As reported in [236], the page replication is efficient as it can be done in 2 μs. Consequently, a consistent snapshot can be constructed efficiently for the OLAP queries without heavy synchronization cost.

In [147], four snapshot mechanisms were benchmarked: software-based Tuple Shadowing which generates a new version when a tuple is modified, software-based Twin Tuple which always keeps two versions of each tuple, hardware-based Page Shadowing used by HyPer, and HotCold Shadowing which combines Tuple Shadowing and hardware-supported Page Shadowing by clustering update-intensive objects. The study shows that Page Shadowing is superior in terms of OLTP performance, OLAP query response time and memory consumption. The most time-consuming task in the creation of a snapshot in the Page Shadowing mechanism is the copying of a process's page table, which can be reduced by using huge page (2 MB per page on x86) for cold data [135]. The hot or cold data is monitored and clustered with a hardware-assisted approach by reading/resetting the *young* and *dirty* flags of a page. Compression is applied on cold data to further improve the performance of OLAP workload and reduce memory consumption [135].

Snapshot is not only used for OLAP queries, but also for long-running transactions [238], as these long-running transactions will block other short good-natured transactions in the serial execution model. In HyPer, these ill-natured transactions are identified and tentatively executed on a child process with a consistent snapshot, and the changes made by these transactions are effected by issuing a deterministic "apply transaction", back to the main database process. The *apply transaction* validates the execution of the tentative transaction, by checking that all reads performed on the snapshot are identical to what would have been read on the main database if view serializability is required, or by checking the writes on the snapshot are disjoint from the writes by all transactions on the main database after the snapshot was created if the snapshot isolation is required. If the validation succeeds, it applies the writes to the main database state. Otherwise an abort is reported to the client.

Register-conscious compilation. To process a query, HyPer translates it into compact and efficient machine code using the LLVM compiler framework [112], [167], rather than using the classical iterator-based query processing model. The HyPer JIT compilation model is designed to avoid function calls by extending recursive function calls into a code fragment loop, thus resulting in better code locality and data locality (i.e., temporal locality for CPU registers), because each code fragment performs all actions on a tuple within one execution pipeline during which the tuple is kept in the registers, before materializing the result into the memory for the next pipeline.

As an optimized high-level language compiler (e.g., C++) is slow, HyPer uses the LLVM compiler framework to generate portable assembler code for an SQL query. In particular, when processing an SQL query, it is first processed as per normal, i.e., the query is parsed, translated and optimized into an algebraic logical plan. However, the algebraic logical plan is not translated into an executable physical plan as in the conventional scheme, but instead compiled into an imperative program (i.e., LLVM assembler code) which can then be executed directly using the JIT compiler provided by LLVM. Nevertheless, the complex part of query processing (e.g., complex data structure management, sorting) is still written in C++, which is pre-compiled. As the LLVM code can directly call the native C++ method without additional wrapper, C++ and LLVM interact with each other without performance penalty [112]. However, there is a trade-off between defining functions, and inlining code in one compact code fragment, in terms of code cleanness, the size of the executable file, efficiency, etc.

ART Indexing. HyPer uses an adaptive radix tree [87] for efficient indexing. The property of the radix tree guarantees that the keys are ordered bit-wise lexicographically, making it possible for range scan, prefix lookup, etc. Larger span of radix tree can decrease the tree height linearly, thus speeding up the search process, but increase the space consumption exponentially. ART achieves both space and time efficiency by adaptively using different inner node sizes with the same, relatively large span, but different fan-out.

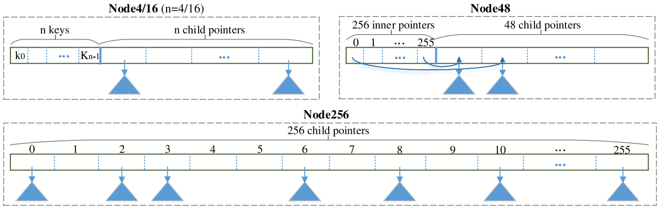

Fig. 5. ART inner node structures.

Specifically, there are four types of inner nodes with a span of 8 bits but different capacities: Node4, Node16, Node48 and Node256, which are named according to their maximum capacity of storing child node pointers. In particular, Node4/Node16 can store up to 4/16 child pointers and uses an array of length 4/16 for sorted keys and another array of the same length for child pointers. Node48 uses a 256-element array to directly index key bits to the pointer array with capacity of 48, while Node256 is simply an array of 256 pointers as normal radix tree node, which is used to store between 49 to 256 entries. Fig. 5 illustrates the structures of Node4, Node16, Node48 and Node256. Lazy expansion and path compression techniques are adopted to further reduce the memory consumption.

3.1.4 SAP HANA

SAP HANA [77], [239], [240] is a distributed in-memory database featured for the integration of OLTP and OLAP [41], and the unification of structured (i.e., relational table) [74], semi-structured (i.e., graph) [241] and unstructured data (i.e., text) processing. All the data is kept in memory as long as there is enough space available, otherwise entire data objects (e.g., tables or partitions) are unloaded from memory and reloaded into memory when they are needed again. HANA has the following features:

- It supports both row- and column-oriented stores for relational data, in order to optimize different query workloads. Furthermore, it exploits columnar data layout for both efficient OLAP and OLTP by adding two levels of delta data structures to alleviate the inefficiency of insertion and deletion operations in columnar data structures [74].
- It provides rich data analytics functionality by offering multiple query language interfaces (e.g., standard SQL, SQLScript, MDX, WIPE, FOX and R), which makes it easy to push down more application semantics into the data management layer, thus avoiding heavy data transfer cost.
- It supports temporal queries based on the *Timeline Index* [242] naturally as data is versioned in HANA.
- It provides snapshot isolation based on multi-version concurrency control, transaction semantics based on optimized two-phase commit protocol (2PC) [243], and fault-tolerance by logging and periodic checkpointing into GPFS file system [148].

We will elaborate only on the first three features, as the other feature is a fairly common technique used in the literature.

Relational stores. SAP HANA supports both row- and column-oriented physical representations of relational tables. Row store is beneficial for heavy updates and inserts, as

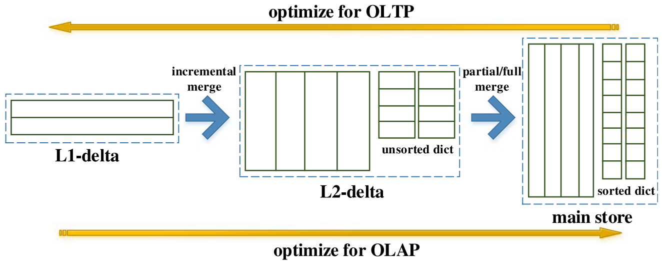

Fig. 6. HANA hybrid store.

well as point queries that are common in OLTP, while column store is ideal for OLAP applications as they usually access all values of a column together, and few columns at a time. Another benefit for column-oriented representation is that it can utilize compression techniques more effectively and efficiently. In HANA, a table/partition can be configured to be either in the row store or in the column store, and it can also be re-structured from one store to the other. HANA also provides a storage advisor [244] to recommend the optimal representation based on data and query characteristics by taking both query cost and compression rate into consideration.

As a table/partition only exists in either a row store or a column store, and both have their own weaknesses, HANA designs a three-level column-oriented unified table structure, consisting of L1-delta, L2-delta and main store, which is illustrated in Fig. 6, to provide efficient support for both OLTP and OLAP workloads, which shows that column store can be deployed efficiently for OLTP as well [41], [74]. In general, a tuple is first stored in L1-delta in row format, then propagated to L2-delta in column format and finally merged with the main store with heavier compression. The whole process of the three stages is called a lifecycle of a tuple in HANA term.

Rich data analytics support. HANA supports various programming interfaces for data analytics (i.e., OLAP), including standard SQL for generic data management functionality, and more specialized languages such as SQL script, MDX, FOX, WIPE [74], [240] and R [245]. While SQL queries are executed in the same manner as in a traditional database, other specialized queries have to be transformed. These queries are first parsed into an intermediate abstract data flow model called “calculation graph model”, where source nodes represent persistent or intermediate tables and inner nodes reflect logical operators performed by these queries, and then transformed into execution plans similar to that of an SQL query. Unlike other systems, HANA supports R scripting as part of the system to enable better optimization of ad-hoc data analytics jobs. Specifically, R scripts can be embedded into a custom operator in the calculation graph [245]. When an R operator is to be executed, a separate R runtime is invoked using the Rserve package [246]. As the column format of HANA column-oriented table is similar to R’s vector-oriented dataframe, there is little overhead in the transformation from table to dataframe. Data transfer is achieved via shared memory, which is an efficient inter-process communication (IPC) mechanism. With the help of *RICE* package [245], it only needs to copy once to make the data available for the R process, i.e., it just copies the data from the database to the shared memory section, and the R runtime can access the data from the shared memory section directly.

Temporal query. HANA supports temporal queries, such as temporal aggregation, time travel and temporal join, based on a unified index structure called the Timeline Index [88], [242], [247]. For every logical table, HANA keeps the *current* version of the table in a *Current Table* and the whole history of previous versions in a *Temporal Table*, accompanied with a *Timeline Index* to facilitate temporal queries. Every tuple of the *Temporal Table* carries a valid interval, from its *commit time* to its *last valid time*, at which some transaction invalidates that value. Transaction Time in HANA is represented by discrete, monotonically increasing *versions*. Basically, the *Timeline Index* maps each *version* to all the write events (i.e., records in the *Temporal Table*) that committed before or at that version. A *Timeline Index* consists of an *Event List* and a *Version Map*, where the *Event List* keeps track of every *invalidation* or *validation* event, and the *Version Map* keeps track of the sequence of events that can be seen by each version of the database. Consequently due to the fact that all visible rows of the Temporal Table at every point in time are tracked, temporal queries can be implemented by scanning *Event List* and *Version Map* concurrently.

To reduce the full scan cost for constructing a temporal view, HANA augments the difference-based *Timeline Index* with a number of complete view representations, called checkpoints, at a specific time in the history. In particular, a checkpoint is a bit vector with length equal to the number of rows in the *Temporal Table*, which represents the visible rows of the *Temporal Table* at a certain time point (i.e., a certain version). With the help of checkpoints, a temporal view at a certain time can be obtained by scanning from the latest checkpoint before that time, rather than scanning from the start of the *Event List* each time.

3.2 In-Memory NoSQL Databases

NoSQL is short for Not Only SQL, and a NoSQL database provides a different mechanism from a relational database for data storage and retrieval. Data in NoSQL databases is usually structured as a tree, graph or key-value rather than a tabular relation, and the query language is usually not SQL as well. NoSQL database is motivated by its simplicity, horizontal scaling and finer control over availability, and it usually compromises consistency in favor of availability and partition tolerance [25], [248].

With the trend of “Memory is the new disk”, in-memory NoSQL databases are flourishing in recent years. There are key-value stores such as Redis [66], RAMCloud [2], MemepiC [60], [138], Masstree [249], MICA [64], Mercury [250], Citrusleaf/Aerospike [34], Kyoto/Tokyo Cabinet [251], Pilaf [252], document stores such as MongoDB [65], Couchbase [253], graph databases such as Trinity [46], Bitsy [254], RDF databases such as OWLIM [255], WhiteDB [50], etc. There are some systems that are partially in-memory, such as MongoDB [65], MonetDB [256], MDB [257], as they use memory-mapped files to store data such that the data can be accessed as if it was in the memory.

In this section, we will introduce some representative in-memory NoSQL databases, including MemepiC [60], [138], MongoDB [65], RAMCloud [2], [75], [126], [258], [259], Redis [66] and some graph databases.

3.2.1 MemepiC

MemepiC [60] is the in-memory version of epiC [23], an extensible and scalable system based on *Actor Concurrent programming model* [260], which has been designed for processing Big Data. It not only provides low latency storage service as a distributed key-value store, but also integrates in-memory data analytics functionality to support online analytics. With an efficient data eviction and fetching mechanism, MemepiC has been designed to maintain data that is much larger than the available memory, without severe performance degradation. We shall elaborate MemepiC in three aspects: system calls reduction, integration of storage service and analytics operations, and virtual memory management.

Less-system-call design. The conventional database design that relies on system calls for communication with hardware or synchronization is no longer suitable for achieving good performance demanded by in-memory systems, as the overhead incurred by system calls is detrimental to the overall performance. Thus, MemepiC subscribes to the less-system-call design principle, and attempts to reduce as much as possible on the use of system calls in the storage access (via memory-mapped file instead), network communication (via RDMA or library-based networking), synchronization (via transactional memory or atomic primitives) and fault-tolerance (via remote logging) [60].

Integration of storage service and analytics operations. In order to meet the requirement of online data analytics, MemepiC also integrates data analytics functionality, to allow analyzing data where it is stored [60]. With the integration of data storage and analytics, it significantly eliminates the data movement cost, which typically dominates in conventional data analytics scenarios, where data is first fetched from the database layer to the application layer, only after which it can be analyzed [172]. The synchronization between data analytics and storage service is achieved based on atomic primitives and *fork*-based virtual snapshot.

User-space virtual memory management (UVMM). The problem of relatively smaller size of main memory is alleviated in MemepiC via an efficient user-space virtual memory management mechanism, by allowing data to be freely evicted to disks when the total data size exceeds the memory size, based on a configurable paging strategy [138]. The adaptability of data storage enables a smooth transition from disk-based to memory-based databases, by utilizing a hybrid of storages. It takes advantage of not only semantics-aware eviction strategy but also hardware-assisted I/O and CPU efficiency, exhibiting a great potential as a more general approach of "Anti-Caching" [138]. In particular, it adopts the following strategies.

- A hybrid of access tracking strategies, including user-supported tuple-level access logging, MMU-assisted page-level access tracking, virtual memory area (VMA)-protection-based method and *malloc*-injection, which achieves light-weight and semantics-aware access tracking.
- Customized WSCLOCK paging strategy based on fine-grained access traces collected by above-mentioned access tracking methods, and other alternative strategies including LRU, aging-based LRU and FIFO, which enables a more accurate and flexible online eviction strategy.
- VMA-protection-based book-keeping method, which incurs less memory overhead for book-keeping the location of data, and tracking the data access in one go.
- Larger data swapping unit with a fast compression technique (i.e., LZ4 [261]) and kernel-supported asynchronous I/O, which can take advantage of the kernel I/O scheduler and block I/O device, and reduce the I/O traffic significantly.

3.2.2 MongoDB

MongoDB [65] is a document-oriented NoSQL database, with few restrictions on the schema of a document (i.e., BSON-style). Specifically, a MongoDB hosts a number of databases, each of which holds a set of collections of documents. MongoDB provides atomicity at the document-level, and indexing and data analytics can only be conducted within a single collection. Thus "cross-collection" queries (such as join in traditional databases) are not supported. It uses primary/secondary replication mechanism to guarantee high availability, and sharding to achieve scalability. In a sense, MongoDB can also act as a cache for documents (e.g., HTML files) since it provides data expiration mechanism by setting TTL (Time-to-Live) for documents.

We will discuss two aspects of MongoDB in detail in the following sections, i.e., the storage and data analytics functionality.

Memory-mapped file. MongoDB utilizes memory-mapped files for managing and interacting with all its data. It can act as a fully in-memory database if the total data can fit into the memory. Otherwise it depends on the virtual-memory manager (VMM) which will decide when and which page to page in or page out. Memory-mapped file offers a way to access the files on disk in the same way we access the dynamic memory—through pointers. Thus we can get the data on disk directly by just providing its pointer (i.e., virtual address), which is achieved by the VMM that has been optimized to make the paging process as fast as possible. It is typically faster to access memory-mapped files than direct file operations because it does not need a system call for normal access operations and it does not require memory copy from kernel space to user space in most operating systems. On the other hand, the VMM is not able to adapt to MongoDB's own specific memory access patterns, especially when multiple tenants reside in the same machine. A more intelligent ad-hoc scheme would be able to manage the memory more effectively by taking specific usage scenarios into consideration.

Data analytics. MongoDB supports two types of data analytics operations: aggregation (i.e., aggregation pipeline and single purpose aggregation operations in MongoDB term) and MapReduce function which should be written in JavaScript language. Data analytics on a sharded cluster that needs central assembly is conducted in two steps:

- The query router divides the job into a set of tasks and distributes the tasks to the appropriate sharded instances, which will return the partial results back to the query router after finishing the dedicated computations.
- The query router will then assemble the partial results and return the final result to the client.

3.2.3 RAMCloud

RAMCloud [2], [75], [126], [258], [259] is a distributed in-memory key-value store, featured for low latency, high availability and high memory utilization. In particular, it can achieve tens of microseconds latency by taking advantage of low-latency networks (e.g., Infiniband and Myrinet), and provide "continuous availability" by harnessing large scale to recover in 1-2 seconds from system failure. In addition, it adopts a log-structured data organization with a two-level cleaning policy to structure the data both in memory and on disks. This results in high memory utilization and a single unified data management strategy. The architecture of RAMCloud consists of a coordinator who maintains the metadata in the cluster such as cluster membership, data distribution, and a number of storage servers, each of which contains two components, a master module which manages the in-memory data and handles read/write requests from clients, and a backup module which uses local disks or flash memory to backup replicas of data owned by other servers.

Data organization. Key-value objects in RAMCloud are grouped into a set of tables, each of which is individually range-partitioned into a number of tablets based on the hash-codes of keys. RAMCloud relies on the uniformity of hash function to distribute objects in a table evenly in proportion to the amount of hash space (i.e., the range) a storage server covers. A storage server uses a single log to store the data, and a hash table for indexing. Data is accessed via the hash table, which directs the access to the current version of objects.

RAMCloud adopts a log-structured approach of memory management rather than traditional memory allocation mechanisms (e.g., C library's *malloc*), allowing 80-90 percent memory utilization by eliminating memory fragmentation. In particular, a log is divided into a set of *segments*. As the log structure is append-only, objects are not allowed to be deleted or updated in place. Thus a periodic clean job should be scheduled to clean up the deleted/stale objects to reclaim free space. RAMCloud designs an efficient two-level cleaning policy.

- It schedules a *segment compaction* job to clean the log segment in memory first whenever the free memory is less than 10 percent, by copying its live data into a smaller *segment* and freeing the original *segment*.
- When the data on disk is larger than that in memory by a threshold, a *combined cleaning* job starts, cleaning both the log in memory and on disk together.

A two-level cleaning policy can achieve a high memory utilization by cleaning the in-memory log more frequently, and meanwhile reduce disk bandwidth requirement by trying to lower the disk utilization (i.e., increase the percentage of deleted/stale data) since this can avoid copying a large percentage of live data on disk during cleaning.

Fast crash recovery. One big challenge for in-memory storage is fault-tolerance, as the data is resident in the volatile DRAM. RAMCloud uses replication to guarantee durability by replicating data in remote disks, and harnesses the large scale of resources (e.g., CPU, disk bandwidth) to speed up recovery process [126], [259]. Specifically, when receiving an update request from a client, the master server appends the new object to the in-memory log, and then forwards the object to R (usually $R = 3$) remote backup servers, which buffer the object in memory first and flush the buffer onto disk in a batch (i.e., in unit of *segment*). The backup servers respond as soon as the object has been copied into the buffer, thus the response time is dominated by the network latency rather than the disk I/O.

To make recovery faster, replicas of the data are scattered across all the backup servers in the cluster in unit of *segment*, thus making more backup servers collaborate for the recovery process. Each master server decides independently where to place a segment replica using a combination of randomization and refinement, which not only eliminates pathological behaviors but also achieves a nearly optimal solution. Furthermore, after a server fails, in addition to all the related backup servers, multiple master servers are involved to share the recovery job (i.e., re-constructing the in-memory log and hash table), and take responsibility for ensuring an even partitioning of the recovered data. The assignment of recovery job is determined by a *will* made by the master before it crashes. The *will* is computed based on *tablet profiles*, each of which maintains a histogram to track the distribution of resource usage (e.g., the number of records and space consumption) within a single table/tablet. The *will* aims to balance the partitions of a recovery job such that they require roughly equal time to recover.

The random replication strategy produces almost uniform allocation of replicas and takes advantage of the large scale, thus preventing data loss and minimizing recovery time. However, this strategy may result in data loss under simultaneous node failures [125]. Although the amount of lost data may be small due to the high dispersability of segment replicas, it is possible that all replicas of certain part of the data may become unavailable. Hence RAMCloud also supports another replication mode based on *Copyset* [125], [258], [259], to reduce the probability of data loss after large, coordinated failures such as power loss. Copyset trades off the amount of lost data for the reduction in the frequency of data loss, by constraining the set of backup servers where all the segments in a master server can be replicated to. However, this can lead to longer recovery time as there are fewer backup servers for reading the replicas from disks. The trade-off can be controlled by the scatter width, which is the number of backup servers that each server's data are allowed to be replicated to. For example, if the scatter width equals the number of all the other servers (except the server that wants to replicate) in the cluster, Copyset then turns to random replication.

3.2.4 Redis

Redis [66] is an in-memory key-value store implemented in C with support for a set of complex data structures, including hash, list, set, sorted set, and some advanced functions such as publish/subscribe messaging, scripting and transactions. It also embeds two persistence mechanisms—snapshotting and append-only logging. Snapshotting will back up all the current data in memory onto disk periodically, which facilitates recovery process, while append-only logging will log every update operation, which guarantees more availability. Redis is single-threaded, but it processes requests asynchronously by utilizing an event notification strategy to overlap the network I/O communication and data storage/retrieval computation.

Redis also maintains a hash-table to structure all the key-value objects, but it uses naive memory allocation (e.g., malloc/free), rather than slab-based memory allocation strategy (i.e., Memcached's), thus making it not very suitable as an LRU cache, because it may incur heavy memory fragmentation. This problem is partially alleviated by adopting the *jemalloc* [262] memory allocator in the later versions.

Scripting. Redis features the server-side scripting functionality (i.e., Lua scripting), which allows applications to perform user-defined functions inside the server, thus avoiding multiple round-trips for a sequence of dependent operations. However, there is an inevitable costly overhead in the communication between the scripting engine and the main storage component. Moreover, a long-running script can degenerate the overall performance of the server as Redis is single-threaded and the long-running script can block all other requests.

Distributed Redis. The first version of distributed Redis is implemented via data sharding on the client-side. Recently, the Redis group introduces a new version of distributed Redis called Redis Cluster, which is an autonomous distributed data store with support for automatic data sharding, master-slave fault-tolerance and online cluster re-organization (e.g., adding/deleting a node, re-sharding the data). Redis Cluster is fully distributed, without a centralized master to monitor the cluster and maintain the metadata. Basically, a Redis Cluster consists of a set of Redis servers, each of which is aware of the others. That is, each Redis server keeps all the metadata information (e.g., partitioning configuration, aliveness status of other nodes) and uses gossip protocol to propagate updates.

Redis Cluster uses a hash slot partition strategy to assign a subset of the total hash slots to each server node. Thus each node is responsible for the key-value objects whose hash code is within its assigned slot subset. A client is free to send requests to any server node, but it will get *redirection* response containing the address of an appropriate server when that particular node cannot answer the request locally. In this case, a single request needs two round-trips. This can be avoided if the client can cache the map between hash slots and server nodes. The current version of Redis Cluster requires manual re-sharding of the data and allocating of slots to a newly-added node. The availability is guaranteed by accompanying a master Redis server with several slave servers which replicate all the data of the master, and it uses asynchronous replication in order to gain good performance, which, however, may introduce inconsistency among primary copy and replicas.

3.2.5 In-Memory Graph Databases

Bitsy. Bitsy [254] is an embeddable in-memory graph database that implements the *Blueprints API*, with ACID guarantees on transactions based on the optimistic concurrency model. Bitsy maintains a copy of the entire graph in memory, but logs every change to the disk during a commit operation, thus enabling recovery from failures. Bitsy is designed to work in multi-threaded OLTP environments. Specifically, it uses multi-level dual-buffer/log to improve the write transaction performance and facilitate log cleaning, and lock-free reading with sequential locks to ameliorate the read performance. Basically, it has three main design principles:

- *No seek*. Bitsy appends all changes to an unordered transaction log, and depends on re-organization process to clean the obsolete vertices and edges.
- *No socket*. Bitsy acts as an embedded database, which is to be integrated into an application (java-based). Thus the application can access the data directly, without the need to transfer through socket-based interface which results in a lot of system calls and serialization/de-serialization overhead.
- *No SQL*. Bitsy implements the Blueprints API, which is oriented for the property graph model, instead of the relational model with SQL.

Trinity. Trinity is an in-memory distributed graph database and computation platform for graph analytics [46], [263], whose graph model is built based on an in-memory key-value store. Specifically, each graph node corresponds to a Trinity cell, which is an (*id*, blob) pair, where *id* represents a node in the graph, and the blob is the adjacent list of the node in serialized binary format, instead of runtime object format, in order to minimize memory overhead and facilitate checkpointing process. However this introduces serialization/deserialization overhead in the analytics computation. A large cell, i.e., a graph node with a large number of neighbors, is represented as a set of small cells, each of which contains only the local edge information, and a central cell that contains cell *id*s for the dispersed cells. Besides, the edge can be tagged with a label (e.g., a predicate) such that it can be extended to an RDF store [263], with both local predicate indexing and global predicate indexing support. In local predicate indexing, the adjacency list for each node is sorted first by predicates and then by neighboring nodes *id*, such as SPO[11] or OPS index in traditional RDF store, while the global predicate index enables the locating of cells with a specific predicate as traditional PSO or POS index.

3.3 In-Memory Cache Systems

Cache plays an important role in enhancing system performance, especially in web applications. Facebook, Twitter, Wikipedia, LiveJournal, et al. are all taking advantage of cache extensively to provide good service. Cache can provide two optimizations for applications: optimization for disk I/O by allowing to access data from memory, and optimization for CPU workload by keeping results without the need for re-computation. Many cache systems have been developed for

[11] S stands for subject, P for predicate, and O for object.

various objectives. There are general cache systems such as Memcached [61] and BigTable Cache [248], systems targeting speeding up analytics jobs such as PACMan [264] and GridGain [51], and more purpose specific systems that have been designed for supporting specific frameworks such as NCache [265] for .NET and Velocity/AppFabric [266] for Windows servers, systems supporting strict transactional semantics such as TxCache [267], and network caching such as HashCache [268].

Nonetheless, cache systems were mainly designed for web applications, as Web 2.0 increases both the complexity of computation and strictness of service-level agreement (SLA). Full-page caching [269], [270], [271] was adopted in the early days, while it becomes appealing to use fine-grained object-level data caching [61], [272] for flexibility. In this section, we will introduce some representative cache systems/libraries and their main techniques in terms of in-memory data management.

3.3.1 Memcached

Memcached [61] is a light-weight in-memory key-value object caching system with strict LRU eviction. Its distributed version is achieved via the client-side library. Thus, it is the client libraries that manage the data partitioning (usually hash-based partitioning) and request routing. Memcached has different versions of client libraries for various languages such as C/C++, PHP, Java, Python, etc. In addition, it provides two main protocols, namely *text protocol* and *binary protocol*, and supports both UDP and TCP connections.

Data organization. Memcached uses a big hash-table to index all the key-value objects, where the key is a text string and the value is an opaque byte block. In particular, the memory space is broken up into slabs of 1 MB, each of which is assigned to a slab class. And slabs do not get reassigned to another class. The slab is further cut into chunks of a specific size. Each slab class has its own chunk size specification and eviction mechanism (i.e., LRU). Key-value objects are stored in the corresponding slabs based on their sizes. The slab-based memory allocation is illustrated in Fig. 7, where the grow factor indicates the chunk size difference ratio between adjacent slab classes. The slab design helps prevent memory fragmentation and optimize memory usage, but also causes slab calcification problems. For example, it may incur unnecessary evictions in scenarios where Memcached tries to insert a 500 KB object when it runs out of 512 KB slabs but has lots of 2 MB slabs. In this case, a 512 KB object will be evicted although there is still a lot of free space. This problem is alleviated in the optimized versions of Facebook and Twitter [273], [274].

Concurrency. Memcached uses *libevent* library to achieve asynchronous request processing. In addition, Memcached is a multi-threaded program, with fine-grained *pthread mutex* lock mechanism. A *static* item lock hash-table is used to control the concurrency of memory accesses. The size of the lock hash-table is determined based on the configured number of threads. And there is a trade-off between the memory usage for the lock hash-table and the degree of parallelism. Even though Memcached provides such a fine-grained locking mechanism, most of operations such as index lookup/update and cache eviction/update still need global locks [63], which prevents current Memcached from scaling up on multi-core CPUs [275].

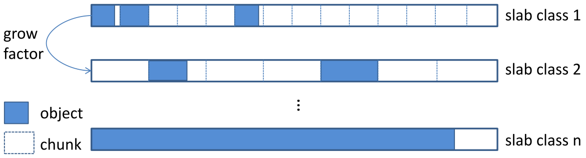

Fig. 7. Slab-based allocation.

Memcached in Production—Facebook's Memcache [273] *and* Twitter's *Twemcache [274]*. Facebook scales Memcached at three different deployment levels (i.e., cluster, region and across regions) from the engineering point of view, by focusing on its specific workload (i.e., read-heavy) and trading off among performance, availability and consistency [273]. Memcache improves the performance of Memcached by designing fine-grained locking mechanism, adaptive slab allocator and a hybrid of lazy and proactive eviction schemes. Besides, Memcache focuses more on the deployment-level optimization. In order to reduce the latency of requests, it adopts parallel requests/batching, uses connection-less UDP for *get* requests and incorporates flow-control mechanisms to limit incast congestion. It also utilizes techniques such as leases [276] and stale reads [277] to achieve high hit rate, and provisioned pools to balance load and handle failures.

Twitter adopts similar optimizations on its distributed version of Memcached, called Twemcache. It alleviates Memcached's slab allocation problem (i.e., slab calcification problem) by random eviction of a whole slab and re-assignment of a desired slab class, when there is not enough space. It also enables a lock-less stat collection via the updater-aggregator model, which is also adopted by Facebook's Memcache. In addition, Twitter also provides a proxy for the Memcached protocol, which can be used to reduce the TCP connections in a huge deployment of Memcached servers.

3.3.2 MemC3

MemC3 [63] optimizes Memcached in terms of both performance and memory efficiency by using optimistic concurrent *cuckoo* hashing and LRU-approximating eviction algorithm based upon CLOCK [279], with the assumption that small and read-only requests dominate in real-world workloads. MemC3 mostly facilitates read-intensive workloads, as the write operations are still serialized in MemC3 and *cuckoo* hashing favors read over write operation. In addition, applications involving a large number of small objects should benefit more from MemC3 in memory efficiency because MemC3 eliminates a lot of pointer overhead embedded in the key-value object. CLOCK-based eviction algorithm takes less memory than list-based strict LRU, and makes it possible to achieve high concurrency as it needs no global synchronization to update LRU.

Optimistic Concurrent Cuckoo *Hashing*. The basic idea of *cuckoo* hashing [278] is to use two hash functions to provide each key two possible buckets in the hash table. When a new key is inserted, it is inserted into one of its two possible buckets. If both buckets are occupied, it will randomly displace the key that already resides in one of these two buckets. The displaced key is then inserted into its alternative bucket, which may further trigger a displacement, until a vacant slot is found or until a maximum number of displacements is reached (at this point, the hash table is rebuilt using new hash functions). This sequence of displacements forms a *cuckoo* displacement path. The collision resolution strategy of *cuckoo* hashing can achieve a high load factor. In addition, it eliminates the pointer field embedded in each key-value object in the chaining-based hashing used by Memcached, which further ameliorates the memory efficiency of MemC3, especially for small objects.

MemC3 optimizes the conventional *cuckoo* hashing by allowing each bucket with four tagged slots (i.e., four-way set-associative), and separating the discovery of a valid *cuckoo* displacement path from the execution of the path for high concurrency. The tag in the slot is used to filter the unmatched requests and help to calculate the alternative bucket in the displacement process. This is done without the need for the access to the exact key (thus no extra pointer de-reference), which makes both look-up and insert operations cache-friendly. By first searching for the *cuckoo* displacement path and then moving keys that need to be displaced backwards along the *cuckoo* displacement path, it facilitates fine-grained optimistic locking mechanism. MemC3 uses lock striping techniques to balance the granularity of locking, and optimistic locking to achieve multiple-reader/single-writer concurrency.

3.3.3 TxCache

TxCache [267] is a snapshot-based transactional cache used to manage the cached results of queries to a transactional database. TxCache ensures that transactions see only consistent snapshots from both the cache and the database, and it also provides a simple programming model where applications simply designate functions/queries as cacheable and the TxCache library handles the caching/invalidating of results.

TxCache uses versioning to guarantee consistency. In particular, each object in the cache and the database is tagged with a version, described by its validity interval, which is a range of timestamps at which the object is valid. A transaction can have a staleness condition to indicate that the transaction can tolerate a consistent snapshot within the past staleness seconds. Thus only records that overlap with the transaction's tolerance range (i.e., the range between its timestamp minus staleness and its timestamp) should be considered in the transaction execution. To increase the cache hit rate, the timestamp of a transaction is chosen lazily by maintaining a set of satisfying timestamps and revising it while querying the cache. In this way, the probability of getting more requested records from the cache increases. Moreover, it still keeps the multi-version consistency at the same time. The cached results are automatically invalidated whenever their dependent records are updated. This is achieved by associating each object in the cache with an invalidation tag, which describes which parts of the database it depends on. When some records in the database are modified, the database identifies the set of invalidation tags affected and passes these tags to the cache nodes.

4 In-Memory Data Processing Systems

In-memory data processing/analytics is becoming more and more important in the Big Data era as it is necessary to analyze a large amount of data in a small amount of time. In general, there are two types of in-memory processing systems: data analytics systems which focus on batch processing such as Spark [55], Piccolo [59], SINGA [280], Pregel [281], GraphLab [47], Mammoth [56], Phoenix [57], GridGain [51], and real-time data processing systems (i.e., stream processing) such as Storm [53], Yahoo! S4 [52], Spark Streaming [54], MapReduce Online [282]. In this section, we will review both types of in-memory data processing systems, but mainly focus on those designed for supporting data analytics.

4.1 In-Memory Big Data Analytics Systems

4.1.1 Main Memory MapReduce (M3R)

M3R [58] is a main memory implementation of MapReduce framework. It is designed for interactive analytics with terabytes of data which can be held in the memory of a small cluster of nodes with high mean time to failure. It provides a backward compatible interfaces with conventional MapReduce [283], and significantly better performance. However, it does not guarantee resilience because it caches the results in memory after map/reduce phase instead of flushing into the local disk or HDFS, making M3R not suitable for long-running jobs. Specifically, M3R optimizes the conventional MapReduce design in two aspects as follows:

- It caches the input/output data in an in-memory key-value store, such that the subsequent jobs can obtain the data directly from the cache and the materialization of output results is eliminated. Basically, the key-value store uses a path as a key, and maps the path to a metadata location where it contains the locations for the data blocks.
- It guarantees partition stability to achieve locality by specifying a partitioner to control how keys are mapped to partitions amongst reducers, thus allowing an iterative job to re-use the cached data.

4.1.2 Piccolo

Piccolo [59] is an in-memory data-centric programming framework for running data analytics computation across multiple nodes with support for data locality specification and data-oriented accumulation. Basically, the analytics program consists of a control function which is executed on the master, and a kernel function which is launched as multiple instances concurrently executing on many worker nodes and sharing distributed mutable key-value tables, which can be updated on the fine-grained key-value object level. Specifically, Piccolo supports the following functionalities:

- A user-defined accumulation function (e.g., max, summation) can be associated with each table, and Piccolo executes the accumulation function during runtime to combine concurrent updates on the same key.
- To achieve data locality during the distributed computation, users are allowed to define a partition

function for a table and co-locate a kernel execution with some table partition or co-locate partitions from different tables.

- Piccolo handles machine failures via a global user-assisted checkpoint/restore mechanism, by explicitly specifying when and what to checkpoint in the control function.
- Load-balance during computation is optimized via work stealing, i.e., a worker that has finished all its assigned tasks is instructed to steal a not-yet-started task from the worker with the most remaining tasks.

4.1.3 Spark/RDD

Spark system [55], [284] presents a data abstraction for big data analytics, called resilient distributed dataset (RDD), which is a coarse-grained deterministic immutable data structure with lineage-based fault-tolerance [285], [286]. On top of Spark, Spark SQL, Spark Streaming, MLlib and GraphX are built for SQL-based manipulation, stream processing, machine learning and graph processing, respectively. It has two main features:

- It uses an elastic persistence model to provide the flexibility to persist the dataset in memory, on disks or both. By persisting the dataset in memory, it favors applications that need to read the dataset multiple times (e.g., iterative algorithms), and enables interactive queries.
- It incorporates a light-weight fault-tolerance mechanism (i.e., lineage), without the need for checkpointing. The lineage of an RDD contains sufficient information such that it can be re-computed based on its lineage and dependent RDDs, which are the input data files in HDFS in the worst case. This idea is also adopted by *Tachyon* [287], which is a distributed file system enabling reliable file sharing via memory.

The ability of persisting data in memory in a fault-tolerance manner makes RDD suitable for many data analytics applications, especially those iterative jobs, since it removes the heavy cost of shuffling data onto disks at every stage as Hadoop does. We elaborate on the following two aspects of RDD: data model and job scheduling.

Data model. RDD provides an abstraction for a read-only distributed dataset. Data modification is achieved by coarse-grained RDD transformations that apply the same operation to all the data items in the RDD, thus generating a new RDD. This abstraction offers opportunities for high consistency and a light-weight fault-tolerance scheme. Specifically, an RDD logs the transformations it depends on (i.e., its lineage), without data replication or checkpointing for fault-tolerance. When a partition of the RDD is lost, it is re-computed from other RDDs based on its lineage. As RDD is updated by coarse-grained transformations, it usually requires much less space and effort to back up the lineage information than the traditional data replication or checkpointing schemes, at the price of a higher re-computation cost for computation-intensive jobs, when there is a failure. Thus, for RDDs with long lineage graphs involving a large re-computation cost, checkpointing is used, which is more beneficial.

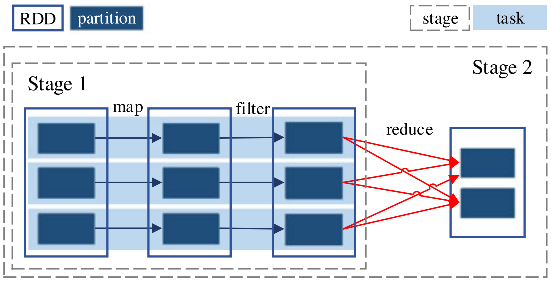

Fig. 8. Spark job scheduler.

The RDD model provides a good caching strategy for "working sets" during computation, but it is not general enough to support traditional data storage functionality for two reasons:

- RDD fault-tolerance scheme is based on the assumption of coarse-grained data manipulation without in-place modification, because it has to guarantee that the program size is much less than the data size. Thus, fine-grained data operations such as updating a single key-value object cannot be supported in this model.
- It assumes that there exists an original dataset persistent on a stable storage, which guarantees the correctness of the fault-tolerance model and the suitability of the block-based organization model. However, in traditional data storage, data is arriving dynamically and the allocation of data cannot be determined beforehand. As a consequence, data objects are dispersed in memory, which results in degraded memory throughput.

Job scheduling. The jobs in Spark are organized into a DAG, which captures job dependencies. RDD uses lazy materialization, i.e., an RDD is not computed unless it is used in an action (e.g., *count()*). When an action is executed on an RDD, the scheduler examines the RDD's lineage to build a DAG of jobs for execution. Spark uses a two-phase job scheduling as illustrated in Fig. 8 [55]:

- It first organizes the jobs into a DAG of stages, each of which may contain a sequence of jobs with only one-to-one dependency on the partition-level. For example, in Fig. 8, Stage 1 consists of two jobs, i.e., map and filter, both of which only have one-to-one dependencies. The boundaries of the stages are the operations with shuffle (e.g., reduce operation in Fig. 8), which have many-to-many dependencies.
- In each stage, a task is formed by a sequence of jobs on a partition, such as the map and filter jobs on the shaded partitions in Fig. 8. Task is the unit of scheduling in the system, which eliminates the materialization of the intermediate states (e.g., the middle RDD of Stage 1 in Fig. 8), and enables a fine-grained scheduling strategy.

4.2 In-Memory Real-Time Processing Systems

4.2.1 Spark Streaming

Spark Streaming [54] is a fault-tolerant stream processing system built based on Spark [55]. It structures a streaming computation as a series of stateless, deterministic batch computations on small time intervals (say 1 s), instead of keeping continuous, stateful operators. Thus it targets applications that tolerate latency of several seconds. Spark Streaming fully utilizes the immutability of RDD and lineage-based fault-tolerance mechanism from Spark, with some extensions and optimizations. Specifically, the incoming stream is divided into a sequence of immutable RDDs based on time intervals, called D-streams, which are the basic units that can be acted on by deterministic transformations, including not only many of the transformations available on normal Spark RDDs (e.g., map, reduce and groupBy), but also windowed computations exclusive for Spark Streaming (e.g., reduceByWindow and countByWindow). RDDs from historical intervals can be automatically merged with the newly-generated RDD as new streams arrive. Stream data is replicated across two worker nodes to guarantee durability of the original data that the lineage-based recovery relies on, and checkpointing is conducted periodically to reduce the recovery time due to long lineage graphs. The determinism and partition-level lineage of D-streams makes it possible to perform parallel recovery after a node fails and mitigate straggler problem by speculative execution.

4.2.2 *Yahoo! S4*

S4 (Simple Scalable Streaming System) [52] is a fully decentralized, distributed stream processing engine inspired by the MapReduce [283] and Actors model [99]. Basically, computation is performed by processing elements (PEs) which are distributed across the cluster, and messages are transmitted among them in the form of data events, which are routed to corresponding PEs based on their identities. In particular, an event is identified by its type and key, while a PE is defined by its functionality and the events that it intends to consume. The incoming stream data is first transformed as a stream of events, which will then be processed by a series of PEs that are defined by users for specific applications. However, S4 does not provide data fault-tolerance by design, since even though automatic PE failover to standby nodes is supported, the states of the failed PEs and messages are lost during the handoff if there is no user-defined state/message backup function inside the PEs.

5 Qualitative Comparison

In this section, we summarize some representative in-memory data management systems elaborated in this paper in terms of data model, supported workloads, indexes, concurrency control, fault-tolerance, memory overflow control, and query processing strategy in Table 3.

In general, in-memory data management systems can also be classified into three categories based on their functionality such as storage and data analytics, namely storage systems, analytics systems, and full-fledged systems that have both capabilities:

- In-memory storage systems have been designed purely for efficient storage service, such as in-memory relational databases only for OLTP (e.g., H-Store[12] [36], Silo [39], Microsoft Hekaton [37]), NoSQL databases without analytics support (e.g., RAMCloud [75], Masstree [249], MICA [64], Mercury [250], Kyoto/Tokyo Cabinet [251], Bitsy [254]), cache systems (e.g., Memcached [61], MemC3 [63], TxCache [267], HashCache [268]), etc. Storage service focuses more on low latency and high throughput for short-running query jobs, and is equipped with a light-weight framework for online queries. It usually acts as the underlying layer for upper-layer applications (e.g., web server, ERP), where fast response is part of the service level agreement.
- In-memory analytics systems are designed for large scale data processing and analytics, such as in-memory big data analytics systems (e.g., Spark/RDD [55], Piccolo [59], Pregel [281], GraphLab [47], Mammoth [56], Phoenix [57]), and real-time in-memory processing systems (e.g., Storm [53], Yahoo! S4 [52], Spark Streaming [54], MapReduce Online [282]). The main optimization objective of these systems is to minimize the runtime of an analytics job, by achieving high parallelism (e.g., multi-core, distribution, SIMD, and pipelining) and batch processing.
- In-memory full-fledged systems include not only in-memory relational databases with support for both OLTP and OLAP (e.g., HyPer [35], Crescando [227], HYRISE [176]), but also data stores with general purpose query language support (e.g., SAP HANA [77], Redis [66], MemepiC [60], [138], Citrusleaf/Aerospike [34], GridGain [51], MongoDB [65], Couchbase [253], MonetDB [256], Trinity [46]). One major challenge for this category of systems is to make a reasonable tradeoff between two different workloads, by making use of appropriate data structures and organization, resource contention, etc.; concurrency control is also very important as it deals with simultaneous mixed workloads.

6 Research Opportunities

In this section, we briefly discuss the research challenges and opportunities for in-memory data management, in the following optimization aspects, which have been introduced earlier in Table 1:

- Indexing. Existing works on indexing for in-memory databases attempt to optimize both time and space efficiency. Hash-based index is simple and easy to implement, and also offers O(1) access time complexity, while tree-based index supports range query naturally and usually has good space efficiency. Trie-based index has bounded O(k) time complexity, where k is the length of the key. There are also other kinds of indexes such as bitmaps and skip-lists, which are amenable to efficient in-memory and distributed processing. For example, the skip-list, which

[12] Based on the H-Store website, it now incorporates a new experimental OLAP engine based on JVM snapshot. Based on its main focus, we put it in the storage category.

TABLE 3
Comparison of In-Memory Data Management Systems

	Systems	**Data Model**	**Workloads**	**Indexes**	**Concurrency Control (CC)**	**Fault Tolerance**	**Memory Overflow**	**Query Processing**
Relational Databases	**H-Store**	relational (row)	OLTP	hashing, B^+-tree, binary tree	partition, serial execution, light-weight locking, speculative CC	command logging, checkpoint, replica	anti-caching	stored procedure
	Hekaton	relational (row)	OLTP	latch-free hashing, Bw-tree	optimistic MVCC	logging, checkpoint, replica	Project Siberia	complied stored procedure
	HyPer/ ScyPer	relational	OLTP, OLAP	hashing, balanced search tree, ART	virtual snapshot, strict timestamp ordering (STO), partition, serial execution for OLTP	logging, checkpoint, replica	compression	JIT, stored procedure
	SAP HANA	relational, graph, text	OLTP, OLAP	timeline index, CSB$^+$-tree, inverted index	MVCC, 2PC	logging, checkpoint, standby server, GPFS	table/partition-level swapping, compression	"calculation graph model"
NoSQL Databases	**MemepiC**	key-value	object operations, analytics	hashing, skip-list	atomic primitives, virtual snapshot	logging, replica	user-space VMM	JIT
	MongoDB	document (bson)	object operations, analytics	B-tree	database-level locking	memory-mapped file	N/A	N/A
	RAMCloud	key-value	object operations	hashing	fine-grained locking	logging, replica	N/A	N/A
	Redis	key-value	object operations	hashing	single-threaded	logging, checkpoint	compression	scripting
Graph Databases	**Bitsy**	graph	OLTP	N/A	optimistic concurrency control (version)	logging, backup	N/A	stored procedure
	Trinity	graph	graph operations	N/A	fine-grained spin-lock	replica, Trinity File System (TFS)	N/A	stored procedure
Cache Systems	**Memcached**	key-value	object operations	hashing	fine-grained locking	N/A	N/A	N/A
	MemC3	key-value	object operations	hashing	lock striping, optimistic locking	N/A	N/A	N/A
	TxCache	key-value	OLTP	hashing	MVCC	N/A	N/A	N/A
Big Data Analytics Systems	**M3R**	key-value	analytics	N/A	partition, locking	N/A	N/A	offline
	Piccolo	key-value	analytics	hashing	locking	checkpoint	N/A	offline
	Spark/ RDD	RDD	analytics	N/A	partition, read/write locking	lineage, checkpoint	block-level swapping	offline
Real-time Processing Systems	**Spark Streaming**	RDD	streaming	N/A	partition, read/write locking	lineage, replica, checkpoint	block-level swapping	N/A
	Yahoo! S4	Event	streaming	hashing	message passing	standby server	N/A	N/A

allows fast point- and range-queries of an ordered sequence of elements with O($log\ n$) complexity, is becoming a desirable alternative to B-trees for in-memory databases, since it can be implemented latch-free easily as a result of its layered structures. Indexes for in-memory databases are different from those for disk-based databases, which focus on I/O efficiency rather than memory and cache utilization. It would be very useful to design an index with constant time complexity for point accesses achieved by hash- and trie-based indexes, efficient support for range accesses achieved by tree- and trie-based indexes, and good space efficiency achieved by hash- and tree-based indexes (like ART index [87]). Lock-free or lock-less index structures are essential to achieve high parallelism without latch-related bottleneck, and index-less design or lossy index is also interesting because of its high throughput and low latency of DRAM [41], [64].

- Data layouts. The data layout or data organization is essential to the performance of an in-memory system as a whole. Cache-conscious design such as columnar structure, cache-line alignment, and space utilization optimization such as compression, data de-fragmentation are the main focuses in the in-memory data organization. The idea of continuous data allocation as log structure has been introduced in main memory systems to eliminate the data fragmentation problem and simplify concurrency control [2]. But it may be better to design an application-independent data allocator with common built-in functionality for in-memory systems such as fault-tolerance, and application-assisted data compression and de-fragmentation.
- Parallelism. Three levels of parallelism should be exploited in order to speed up the processing, which have been detailed in Section 1. It is usually beneficial to increase parallelism at the instruction level (e.g., bit-level parallelism, SIMD) provided in modern architecture, which can achieve nearly optimal speedup, free from concurrency issues and other overhead incurred, but with constraints on the maximum parallelism allowed and data structures to operate on. The instruction-level parallelism may yield a good performance boost, and therefore it should be considered in the design of an efficient in-memory data management system, especially in the design of data structures. With the emergence of

many integrated core (MIC) co-processors (e.g., Intel Xeon Phi), it provides a promising alternative for parallelizing computation, with wider SIMD instructions, many lower-frequency in-order cores and hardware contexts [288].

- Concurrency control/transaction management. For in-memory systems, the overhead of concurrency control significantly affects the overall performance, thus making it perfect if there are no concurrency control at all. Hence, it is worth making the serial execution strategy more efficient for cross-partition transactions and more robust to skewed workloads. Lock-less or lock-free concurrency control mechanism is promising in in-memory data management as a heavy-weight lock-based mechanism can greatly offset the performance improved by the in-memory environment. Atomic primitives provided in most mainstream programming languages are efficient alternatives that can be exploited in designing a lock-free concurrency control mechanism. Besides, HTM provides a hardware-assisted approach for efficient concurrency control protocol, especially under the transactional semantics in databases. Hardware-assisted approaches are good choices in the latency-sensitive in-memory environment, as software solutions usually incur heavy overhead that negates the benefits brought by parallelism and fast data access. But we should take care of its unexpected aborts under certain conditions. A mix of these data protection mechanisms (i.e., HTM, lock, timestamp, atomic primitives) should enable a more efficient concurrency control model. Moreover, the protocol should be data-locality sensitive and cache aware, which matter more for modern machines [192].
- Query processing. Query processing is a widely studied research topic even in traditional disk-based databases. However, traditional query processing framework based on Iterator-/Volcano-style model, although flexible, is no longer suitable for in-memory databases because of its poor code/data locality. The high computing power of modern CPU, and easy-to-use compiler infrastructure such as LLVM [167] enable efficient dynamic compiling [112], which can improve the query processing performance significantly as a result of better code and data locality. SIMD or multi-core boosted processing can be utilized to speed up complex database operations such as join and sort, and NUMA architecture will play a bigger role in the future years.
- Fault tolerance. Fault tolerance is a necessity for an in-memory database in order to guarantee durability; however, it is also a major performance bottleneck caused by I/Os. Thus one design philosophy for fault-tolerance is to make it almost invisible to normal operations by minimizing the I/O cost in the critical path as much as possible. Command logging [131] can reduce the data that needs to be logged, while remote logging used by RAMCloud [2] and 2-Safe visible policy of solidDB [40] can reduce the response time by logging the data in remote nodes and replying back as soon as the data is written into the buffer. Fast recovery can provide high availability upon failure, which may be achievable at the price of more and well-organized backuped files (log/checkpoint). The tradeoff between the interference to the normal performance and the recovery efficiency should be further examined [132]. Hardware/OS-assisted approaches are promising, e.g., NVRAM, memory-mapped file, on top of which optimized algorithms and data structures are required to exert its performance potential.
- Data overflow. In general, approaches to the data overflow problem can be classified into three categories: user-space (e.g., H-Store Anti-caching [133], Hekaton Siberia [134]), kernel-space (e.g., OS Swap, MongoDB memory mapped files [65]) and the hybrid (e.g., Efficient OS Paging [136] and UVMM [138]). The semantics-aware user-space approaches can make more effective decision on the paging strategies, while the hardware-conscious and well-developed kernel-space approaches are able to utilize the I/O efficiency brought by the OS during swapping. Potentially, both the semantics-aware paging strategy and hardware-conscious I/O management can be exploited to boost the performance [136], [138].

In addition to the above, hardware solutions are being increasingly exploited for performance gain. In particular, new hardware/architecture solutions such as HTM, NVM, RDMA, NUMA and SIMD, have been shown to be able to boost the performance of in-memory database systems significantly. Energy efficiency is also becoming attractive in the in-memory systems as DRAM contributes a relatively significant portion of the overall power consumption [289], [290], and distributed computation further exacerbates the problem. Every operational overhead that is considered negligible in disk-based systems, may become the new bottleneck in memory-based systems. Thus the removal of these legacy bottlenecks such as system calls, network stack, and cross-cacheline data layout, would contribute to a significant performance boost for in-memory systems. Furthermore, as exemplified in [117], even the implementation matters a lot in the overhead-sensitive in-memory environment.

7 Conclusions

As memory becomes the new disk, in-memory data management and processing becomes increasingly interesting for both academia and industry. Shifting the data storage layer from disks to main memory can lead to more than 100 × theoretical improvement in terms of response time and throughput. When data access becomes faster, every source of overhead that does not matter in traditional disk-based systems, may degrade the overall performance significantly. The shifting prompts a rethinking of the design of traditional systems, especially for databases, in the aspect of data layouts, indexes, parallelism, concurrency control, query processing, fault-tolerance, etc. Modern CPU utilization and memory-hierarchy-conscious optimization play a significant role in the design of in-memory systems, and new hardware technologies such as HTM and RDMA provide a promising opportunity to resolve problems encountered by software solutions.

In this survey, we have focused on the design principles for in-memory data management and processing, and practical techniques for designing and implementing efficient and high-performance in-memory systems. We reviewed the memory hierarchy and some advanced technologies such as NUMA and transactional memory, which provide the basis for in-memory data management and processing. In addition, we also discussed some pioneering in-memory NewSQL and NoSQL databases including cache systems, batch and online/continuous processing systems. We highlighted some promising design techniques in detail, from which we can learn the practical and concrete system design principles. This survey provides a comprehensive review of important technology in memory management and analysis of related works to date, which hopefully will be a useful resource for further memory-oriented system research.

ACKNOWLEDGMENTS

This work was supported by A*STAR project 1321202073. The authors would like to thank the anonymous reviewers, and also Bingsheng He, Eric Lo, and Bogdan Marius Tudor for their insightful comments and suggestions.

REFERENCES

[1] S. Robbins, “RAM is the new disk,” *InfoQ News*, Jun. 2008.

[2] J. Ousterhout, P. Agrawal, D. Erickson, C. Kozyrakis, J. Leverich, D. Mazières, S. Mitra, A. Narayanan, G. Parulkar, M. Rosenblum, S. M. Rumble, E. Stratmann, and R. Stutsman, “The case for RAMClouds: Scalable high-performance storage entirely in dram,” *ACM SIGOPS Operating Syst. Rev.*, vol. 43, pp. 92–105, 2010.

[3] F. Li, B. C. Ooi, M. T. Özsu, and S. Wu, “Distributed data management using MapReduce,” *ACM Comput. Surv.*, vol. 46, pp. 31:1–31:42, 2014.

[4] HP. (2011). Vertica systems [Online]. Available: http://www.vertica.com

[5] Hadapt Inc.. (2011). Hadapt: Sql on hadoop [Online]. Available: http://hadapt.com/

[6] A. Thusoo, J. S. Sarma, N. Jain, Z. Shao, P. Chakka, S. Anthony, H. Liu, P. Wyckoff, and R. Murthy, “Hive: A warehousing solution over a map-reduce framework,” in *Proc. VLDB Endowment*, vol. 2, pp. 1626–1629, 2009.

[7] Apache. (2008). Apache hbase [Online]. Available: http://hbase.apache.org/

[8] J. C. Corbett, J. Dean, M. Epstein, A. Fikes, C. Frost, J. J. Furman, S. Ghemawat, A. Gubarev, C. Heiser, P. Hochschild, W. Hsieh, S. Kanthak, E. Kogan, H. Li, A. Lloyd, S. Melnik, D. Mwaura, D. Nagle, S. Quinlan, R. Rao, L. Rolig, Y. Saito, M. Szymaniak, C. Taylor, R. Wang, and D. Woodford, “Spanner: Google’s globally-distributed database,” in *Proc. USENIX Symp. Operating Syst. Des. Implementation*, 2012, pp. 251–264.

[9] S. Alsubaiee, Y. Altowim, H. Altwaijry, A. Behm, V. R. Borkar, Y. Bu, M. J. Carey, I. Cetindil, M. Cheelangi, K. Faraaz, E. Gabrielova, R. Grover, Z. Heilbron, Y. Kim, C. Li, G. Li, J. M. Ok, N. Onose, P. Pirzadeh, V. J. Tsotras, R. Vernica, J. Wen, and T. Westmann, “Asterixdb: A scalable, open source BDMS,” in *Proc. Very Large Database*, pp. 1905–1916, 2014.

[10] MySQL AB. (1995). Mysql: The world’s most popular open source database [Online]. Available: http://www.mysql.com/

[11] Apache. (2008). Apache cassandra [Online]. Available: http://cassandra.apache.org/

[12] Oracle. (2013). Oracle database 12c [Online]. Available: https://www.oracle.com/database/index.html

[13] Neo Technology, “Neo4j - the world’s leading graph database,” 2007. [Online]. Available: http://www.neo4j.org/

[14] Aurelius. (2012). Titan—distributed graph database [Online]. Available: http://thinkaurelius.github.io/titan/

[15] A. Kyrola, G. Blelloch, and C. Guestrin, “Graphchi: Large-scale graph computation on just a pc,” in *Proc. 10th USENIX Conf. Operating Syst. Des. Implementation*, 2012, pp. 31–46.

[16] Objectivity Inc. (2010). Infinitegraph [Online]. Available: http://www.objectivity.com/infinitegraph

[17] Apache. (2010). Apache Hama [Online]. Available: https://hama.apache.org

[18] A. Biem, E. Bouillet, H. Feng, A. Ranganathan, A. Riabov, O. Verscheure, H. Koutsopoulos, and C. Moran, “IBM infosphere streams for scalable, real-time, intelligent transportation services,” in *Proc. ACM SIGMOD Int. Conf. Manag. Data*, 2010, pp. 1093–1104.

[19] S. Hoffman, *Apache Flume: Distributed Log Collection for Hadoop*. Birmingham, U.K. Packt Publishing, 2013.

[20] Apache. (2005). Apache hadoop [Online]. Available: http://hadoop.apache.org/

[21] V. Borkar, M. Carey, R. Grover, N. Onose, and R. Vernica, “Hyracks: A flexible and extensible foundation for data-intensive computing,” in *Proc. IEEE 27th Int. Conf. Data Eng.*, 2011, pp. 1151–1162.

[22] M. Isard, M. Budiu, Y. Yu, A. Birrell, and D. Fetterly, “Dryad: Distributed data-parallel programs from sequential building blocks,” in *Proc. 2nd ACM SIGOPS/EuroSys Eur. Conf. Comput. Syst.*, 2007, pp. 59–72.

[23] D. Jiang, G. Chen, B. C. Ooi, K.-L. Tan, and S. Wu, “epiC: An extensible and scalable system for processing big data,” in *Proc. VLDB Endowment*, vol. 7, pp. 541–552, 2014.

[24] H. T. Vo, S. Wang, D. Agrawal, G. Chen, and B. C. Ooi, “LogBase: A scalable log-structured database system in the cloud,” in *Proc. VLDB Endowment*, vol. 5, pp. 1004–1015, 2012.

[25] G. DeCandia, D. Hastorun, M. Jampani, G. Kakulapati, A. Lakshman, A. Pilchin, S. Sivasubramanian, P. Vosshall, and W. Vogels, “Dynamo: Amazon’s highly available key-value store,” *ACM SIGOPS Operating Syst. Rev.*, vol. 41, pp. 205–220, 2007.

[26] S. Ghemawat, H. Gobioff, and S.-T. Leung, “The google file system,” in *Proc. 19th ACM Symp. Operating Syst. Principles*, 2003, pp. 29–43.

[27] K. Shvachko, H. Kuang, S. Radia, and R. Chansler, “The hadoop distributed file system,” in *Proc. IEEE 26th Symp. Mass Storage Syst. Technol.*, 2010, pp. 1–10.

[28] Y. Cao, C. Chen, F. Guo, D. Jiang, Y. Lin, B. C. Ooi, H. T. Vo, S. Wu, and Q. Xu, “ES2: A cloud data storage system for supporting both OLTP and OLAP,” in *Proc. IEEE 27th Int. Conf. Data Eng.*, 2011, pp. 291–302.

[29] FoundationDB. (2013). Foundationdb[Online]. Available: https://foundationdb.com

[30] D. G. Andersen, J. Franklin, M. Kaminsky, A. Phanishayee, L. Tan, and V. Vasudevan, “Fawn: A fast array of wimpy nodes,” in *Proc. ACM SIGOPS 22nd Symp. Operating Syst. Principles*, 2009, pp. 1–14.

[31] H. Lim, B. Fan, D. G. Andersen, and M. Kaminsky, “Silt: A memory-efficient, high-performance key-value store,” in *Proc. 23rd ACM Symp. Operating Syst. Principles*, 2011, pp. 1–13.

[32] B. Debnath, S. Sengupta, and J. Li, “Skimpystash: Ram space skimpy key-value store on flash-based storage,” in *Proc. ACM SIGMOD Int. Conf. Manag. Data*, 2011, pp. 25–36.

[33] Clustrix Inc. (2006). Clustrix [Online]. Available: http://www.clustrix.com/

[34] V. Srinivasan and B. Bulkowski, “Citrusleaf: A real-time NoSQL DB which preserves acid,” in *Proc. Int. Conf. Very Large Data Bases*, 2011, vol. 4, pp. 1340–1350.

[35] A. Kemper and T. Neumann, “HyPer: A hybrid OLTP & OLAP main memory database system based on virtual memory snapshots,” in *IEEE 27th Int. Conf. Data Eng.*, 2011, pp. 195–206.

[36] R. Kallman, H. Kimura, J. Natkins, A. Pavlo, A. Rasin, S. Zdonik, E. P. C. Jones, S. Madden, M. Stonebraker, Y. Zhang, J. Hugg, and D. J. Abadi, “H-store: A high-performance, distributed main memory transaction processing system,” *Proc. VLDB Endowment*, vol. 1, pp. 1496–1499, 2008.

[37] C. Diaconu, C. Freedman, E. Ismert, P.-Å. Larson, P. Mittal, R. Stonecipher, N. Verma, and M. Zwilling, “Hekaton: SQL server’s memory-optimized OLTP engine,” in *Proc. ACM SIGMOD Int. Conf. Manag. Data*, 2013, pp. 1243–1254.

[38] T. Lahiri, M.-A. Neimat, and S. Folkman, “Oracle timesten: An in-memory database for enterprise applications,” *IEEE Data Eng. Bull.*, vol. 36, no. 2, pp. 6–13, Jun. 2013.

[39] S. Tu, W. Zheng, E. Kohler, B. Liskov, and S. Madden, “Speedy transactions in multicore in-memory databases,” in *Proc. ACM Symp. Operating Syst. Principles*, 2013, pp. 18–32.