The Five-minute Rule Thirty Years Later and its Impact on the Storage Hierarchy

Raja Appuswamy
EPFL, Switzerland
raja.appuswamy@epfl.ch

Renata Borovica-Gajic
University of Melbourne, AU
renata.borovica@unimelb.edu.au

Goetz Graefe
Google, Madison WI, USA
goetzg@google.com

Anastasia Ailamaki
EPFL, Switzerland
anastasia.ailamaki@epfl.ch

ABSTRACT

In 1987, Jim Gray and Gianfranco Putzolu put forth the five-minute rule for trading memory to reduce disk-based I/O on the then-current price–performance characteristics of DRAM and HDD. The five-minute rule has gained wide-spread acceptance since its introduction as an important rule-of-thumb in data engineering and has been revisited twice, once in 1997 to account for changes in technology and economic ratio of HDDs and DRAM, and again in 2007 to investigate the impact of NAND flash-based SSDs on the two-tier, DRAM–disk storage hierarchy.

In this paper, we revisit the five-minute rule three decades since its introduction. First, we present changes that have dominated the storage hardware landscape in the last decade and recompute the break-even intervals for today's multi-tiered storage hierarchy. Then, we present recent trends to predict properties of emerging hardware for the next decade, and use the five-minute rule and its methods to predict the implications of these trends on the design of data management systems.

1. INTRODUCTION

The design of computer systems in general, and data management systems in particular, has always been influenced by changes in storage hardware. In the 1980s, database engines used a two-tier storage hierarchy with DRAM as the first level and Hard Disk Drives (HDD) as the second level. Given the speed mismatch between processors and HDD, DRAM was used as a buffer to cache frequently accessed data stored in the HDD. However, given that the amount of DRAM was always only a fraction of available disk capacity, an obvious question that had to be answered was determining when it made economic sense to cache a piece of data in main memory versus storing it on disk.

In 1987, Jim Gray and Gianfranco Putzolu established the now-famous five-minute rule that gave a precise answer to this question– "Pages referenced every five minutes should be memory resident" [9]. They arrived at this value by computing the break-even interval at which the cost of holding a page in memory matches the cost of performing I/O to fetch the page from HDD. Using the then-current price and performance characteristics of DRAM and HDD based on Tandem hardware, they computed the break-even interval to be 400 seconds and rounded it down to five minutes.

If all technologies and prices in the storage hierarchy evolve at the same pace, the five-minute rule would never change. However, different storage technologies, and different aspects even within a particular storage technology (like capacity, latency, and bandwidth), evolve at different rates. These changes in technology and economic factors directly affect the DRAM–HDD five-minute rule as the break-even interval might vary dramatically across generations of storage hardware. Furthermore, the storage hardware landscape is also evolving as new storage media, like NAND flash, with radically different price/performance characteristics, present alternatives to HDD for primary data storage. This results in the original DRAM–HDD rule being augmented with new rules based on how the new storage media is integrated in the conventional two-tier storage hierarchy. Thus, the five-minute rule has been revisited twice, once per decade, since its introduction, considering new metrics in the former case [8] and covering flash memory in the latter case [6].

Today, database engines use storage hierarchies that can span nine different storage media (NVDIMM, 3D XPoint, Memory channel flash, PCIe flash, SATA SSD, 15k RPM HDD, 7k RPM HDD, MAID-based Cold storage devices, tape) with radically different price, performance, and reliability characteristics, grouped across four different tiers (performance, capacity, archival, and backup). Thus, in this paper, we revisit the five-minute rule for a modern, multi-tiered storage hierarchy. In doing so, we present changes to the storage hierarchy in the last decade, highlight trends that will shape the storage hierarchy in the near future, and use guidelines provided by the five-minute rule to identify impending changes in the design of data management engines for emerging hardware.

Contributions of our analysis are as follows:

- Our analysis shows the widening gap between DRAM or SSD-based performance tier and HDD-based capacity tier. What used to be the five-minute rule for DRAM and HDD has now evolved into a four-hour rule, and the break-even interval for SSD–HDD case is one day. This suggests that all performance critical data will soon, if not already, reside only on DRAM and SSD, with HDD being relegated as a high-density storage medium for infrequently accessed data. This also suggests that HDD and SSD vendors should target different tiers and optimization points, namely, IOPS for the performance tier in the case of SSD, and cost/GB for the capacity tier in the case of HDD.

Metric	DRAM				HDD				SATA Flash SSD	
	1987	1997	2007	2017	1987	1997	2007	2017	2007	2017
Unit price($)	5k	15k	48	80	30k	2k	80	49	1k	560
Unit capacity	1MB	1GB	1GB	16GB	180MB	9GB	250GB	2TB	32GB	800GB
$/MB	5k	14.6	0.05	0.005	83.33	0.22	0.0003	0.00002	0.03	0.0007
Random IOPS	-	-	-	-	15	64	83	200	6.2k	67k (r)/20k (w)
Sequential b/w (MB/s)	-	-	-	-	1	10	300	200	66	500 (r)/460 (w)

Table 1: The evolution of DRAM, HDD, and Flash SSD properties

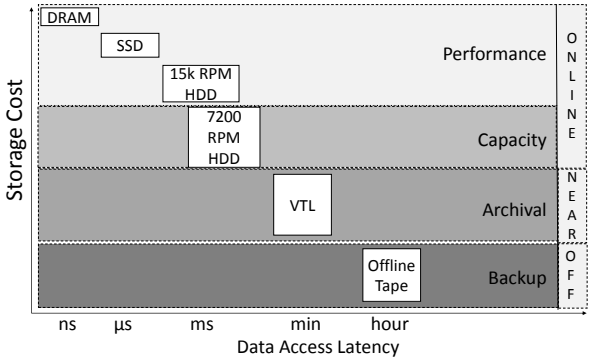

Figure 1: Storage tiering for enterprise databases

- Within the performance tier, emerging PCIe NVMe SSD and new storage technologies like 3D XPoint are quickly closing the gap with DRAM. The break-even interval for DRAM–SSD case has dropped from 15 minutes in 2007 to mere 40 seconds today. This suggests an impending shift from DRAM-based in-memory engines to non-volatile memory-based persistent memory engines.

- Within the capacity tier, as HDD and tape evolve into infinite-capacity storage media, the difference between them with respect to metrics like $/TBScan is also shrinking. Given the popularity of batch analytics and its scan-intensive nature, our analysis reveals that it might be feasible and even economically beneficial to use nearline and tertiary storage devices, that are used only in the archival tier today, as a part of the capacity tier for servicing batch analytics applications.

The rest of this paper is organized as follows. Section 2 provides an overview of the tiered storage used by state-of-the-art database engines. Then, in Section 3, we revisit the five-minute rule three decades since its introduction, compute the break-even intervals using price/performance characteristics typical for storage hardware of today, and compare the intervals with those computed over the last three iterations of the five-minute rule. Following this, we analyze the impact of emerging storage hardware on the five-minute rule by considering the performance tier first in Section 4 and then the capacity tier in Section 5. In each case, we summarize recent trends in hardware technology, compute break-even intervals for emerging hardware, and discuss the implications of our analysis on the appropriate storage tier.

2. BACKGROUND: STORAGE TIERING

Enterprise database engines have long used storage tiering for reducing capital and operational expenses. Traditionally, database engines used a two-tier storage hierarchy. An *online* tier based on enterprise hard disk drives (HDD) provided low-latency (ms) access to data. The *backup* tier based on offline tape cartridges or optical drives, in contrast, provided low-cost, high-latency (hour) storage for storing backups to be restored only during rare failures.

As databases grew in popularity, the necessity to reduce recovery time after failure became important. Further, as regulatory compliance requirements forced enterprises to maintain long-term data archives, the offline nature of the backup tier proved too slow for both storing and retrieving infrequently accessed archival data. This led to the emergence of a new *archival* tier based on nearline storage devices, like virtual tape libraries (VTL), that could store and retrieve data without human intervention in minutes.

Around the early 2000s, NAND flash-based solid-state storage devices (SSD) emerged as an attractive alternative for data storage due to their ability to provide orders of magnitude higher random IOPS compared to HDD. SSDs provided a perfect middle-ground between DRAM and HDD with respect to price and performance. The power consumption of SSDs was also substantially lower compared to HDDs. However, the first-generation flash SSDs suffered from two problems, namely, poor lifetime due to NAND wear out and high cost/GB compared to HDDs. As SSD vendors quickly reduced cost/GB using techniques like multi-level cells and 3D stacking, and improved lifetime using efficient wear leveling and over-provisioning, the traditional HDD-based online tier was bifurcated into two subtiers, namely, a *performance* tier based on RAM or SSD, and a *capacity* tier based on HDD.

Thus, today, enterprise databases typically use a four-tier storage hierarchy as depicted in Figure 1. The DRAM or SSD-based performance tier is used for hosting data accessed by latency-critical applications, like real-time data analytics or high-volume transaction processing. The HDD-based capacity tier is used for hosting data accessed by latency insensitive applications, like batch analytics. The nearline archival tier based on VTL, and the offline backup tier based on tape, are not used for "online" query processing, but for storing data that would be only rarely accessed during disaster recovery or security audits. Database engines classify data based on various characteristics, like the target application, or frequency of accesses, to determine the ideal storage tier. Once determined, Hierarchical Storage Managers (HSM) are used to automatically manage migration of data between online, nearline, and offline storage tiers [14].

Table 1 shows the price, capacity, and performance of DRAM, HDD, and NAND flash-based SSDs across several years. The values reported for 1987, 1997, and 2007 are directly taken from the revised five-minute rule papers published in the respective years [6, 8, 9]. The values listed for 2017 reflect the values that are typical for today's technologies. We obtained these values based on components listed in a server specification from a recent TPC-C report [30]. The table lists performance numbers, obtained from vendor specifications, and unit price as quoted by www.newegg.com on June 1, 2017 for DRAM, SSD, and HDD components specified in the report.

While these values are representative of price/performance characteristics of devices in their respective categories, they are by no means universally accurate. Furthermore, over the past few years,

Tier	1987	1997	2007	2017
DRAM–HDD	5m	5m	1.5h	4h
DRAM–SSD	-	-	15m	7m (r) / 24m (w)
SSD–HDD	-	-	2.25h	1d

Table 2: Evolution of the break-even interval across four decades based on appropriate price, performance, and page size values

HDD and SSD vendors have also started optimizing storage devices for different targets, with HDD vendors focusing on capacity and SSD vendors focusing on performance. This has resulted in a new breed of enterprise-grade, high-density HDDs based on new recording techniques, like Shingled Magnetic Recording (SMR), and high-performance PCIe SSDs that provide an order of magnitude better performance than their SATA counterparts. We will discuss how the five-minute rule is affected by these changes in Section 4, and Section 5. We invite readers to re-evaluate the appropriate formulas for other environments of interest.

3. REVISITING THE FIVE-MINUTE RULE

The five-minute rule explores the trade-off between the cost of DRAM and the cost of disk I/O. Given that caching pages in memory reduces the number of disk I/Os, the five-minute rule provides a formula to predict the optimal *break-even interval*–the time window within which data must be re-accessed in order for it to qualify for being cached in memory. The interval is computed as:

$$\frac{PagesPerMBofDRAM}{AccessesPerSecondPerDisk} \times \frac{PricePerDiskDrive}{PricePerMBofDRAM}$$

The first ratio in the equation was referred to as the *technology ratio*, as both *AccessesPerSecondPerDisk* and *PagesPerMBofDRAM* (or the page size) are determined by the evolution of hardware technology. The second ratio, in contrast, is referred to as the *economic ratio* as pricing is determined by factors other than just hardware technology.

Table 2 presents the break-even interval over the four decades. The values reported for the break-even interval under 1987, 1997, and 2007 are taken verbatim from the original paper [9] and the two follow-up studies that were conducted ten [8] and twenty years after the original paper [6]. The values listed under 2017 are based on metrics listed in Table 1.

DRAM–HDD. First, let us consider the DRAM–HDD case. In 1987, the typical page size used by database engines was 1KB. For 1KB pages, the break-even interval was 400 seconds, which was approximated to five minutes, thus lending the name to this famous rule.

When the study was repeated in 1997, the technology ratio had decreased ten fold due to an improvement in disk IOPS and associated increase in page size from 1KB to 8KB. The economic ratio had increased ten fold due to a drop in DRAM and HDD pricing. As a result, the two ratios balanced each other out and the break-even interval was computed to be 266 seconds, leaving the five-minute rule intact for 8KB pages. However, for 4KB pages, the interval was determined to be nine minutes which was five times longer than the 1987 interval of 100 seconds.

Between 1997 and 2007, DRAM and HDD costs continued to drop further resulting in the economic ratio increasing from 133 ($2k/$15) to 1700 ($80/$0.047). However, the improvement in HDD random IOPS did not keep pace. As a result, the break-even interval increased 10× from nine minutes to 1.5 hours for 4KB pages. The old five-minute rule for DRAM and HDD was determined to apply for a page size of 64KB in 2007.

Over the last decade, the economic ratio has increased further from 1700 to 10,000 ($49/$0.005) due to a further reduction in cost per GB of DRAM and HDD. However, there has only been a 2.5× improvement in HDD performance (200/83), leading to a much lower reduction in technology ratio. As a result, the break-even interval today for DRAM–HDD case is four hours assuming a 4KB page size. The five-minute rule is valid today for 512KB pages.

DRAM–SSD. As we described in Section 2, SSDs are being increasingly used as the storage medium of choice in the latency-critical performance tier. Thus, the five-minute rule can be used to compute a break-even interval for such a scenario where DRAM is used to cache data stored in SSDs. Table 2 shows the interval in 2007, when SSDs were still in their initial stages of adoption, and today, based on SSD price/performance characteristics listed in Table 1. Note that we show two intervals considering both read and write IOPS separately.

We see that the interval has dropped from 15 minutes to seven minutes if we consider 4KB read IOPS. This is in stark contrast with the DRAM–HDD case, where the interval increased 2.7× from 1.5 hours to four hours. In both DRAM–HDD and DRAM–SSD cases, the drop in DRAM cost/MB dominated the economic ratio. However, unlike the 2.5× improvement in random IOPS with HDDs, SSDs have managed to achieve an impressive 11× improvement (67k/6.2k). Thus, the increase in economic ratio was overshadowed by the decrease in technology ratio with SSDs, resulting in the interval shrinking.

SSD–HDD. As SSDs can also be used as a cache for HDDs, the same formulas can also be used to estimate the break-even interval for an SSD–HDD tiering setup. Looking at Table 2, we see that the break-even interval for this case has increased by a factor of 10× from 2.25 hours in 2007 to one day in 2017. The SSD–HDD interval is six times longer than the already-high DRAM–HDD interval (one day/four hours). For the five-minute rule to be valid in the SSD–HDD case, the page size used for HDD access should be 1MB today.

Implications. There are three important consequences of these results. First, the turn-over time in DRAM was six times higher in 2007 if HDD is the second level in the storage hierarchy instead of SSD (1.5h/15m). In 2017, the turn-over time is 34× higher (4h/7m). Thus, in systems tuned using economic considerations, one should replace HDD with SSD, as it would not only improve performance, but also reduce the amount of DRAM required, as only a smaller portion of data needs to be cached to meet required performance goals.

Second, given the four-hour DRAM–HDD and one day SSD–HDD intervals, it is important to keep all active data in the DRAM or SSD-based performance tier and relegate the HDD-based capacity tier to storing only infrequently accessed data.

Third, the large 512KB and 1MB page sizes necessary for keeping the five-minute rule intact in DRAM–HDD and SSD–HDD cases clearly highlight the difference in rate of improvement between random IOPS and sequential bandwidth of HDD. This suggests that HDD should essentially be treated as a sequential access device with data being transferred in large granularities. This is in stark contrast with SSDs, where the shrinking DRAM–SSD break-even interval clearly shows that modern SSDs are optimized for low-latency random accesses.

The growing gap between performance and capacity tiers also implies that SSD and HDD vendors should optimize for different targets, with SSD vendors optimizing for performance and HDD vendors optimizing for cost. In the next two sections, we will explain changes in both the performance and capacity tier that indicate that such targeted optimizations are indeed underway.

Device	Capacity	Price($)	IOPS(k) r/w	B/w (GB/s)
Intel 750	1TB	630	460/290	2.5/1.2
Intel P4800X	384GB	1520	550/500	2.5/2

Table 3: Price/performance metrics for the NAND-based Intel 750 PCIe SSD and 3D-XPoint-based Intel Optane P4800X PCIe SSD

4. FIVE-MINUTE RULE AND THE PERFORMANCE TIER

Over the past few years, there have been several changes in existing storage hardware used for hosting data in the performance tier (DRAM and SSD). In addition, new hardware based on non-volatile memory technologies other than NAND flash, like Intel 3D XPoint, are starting to make inroads in the storage market. In this section, we will provide an overview of these trends in the high-performance storage landscape and use the methodology of the five-minute rule to revisit system design alternatives in the context of emerging hardware.

4.1 Trends in solid-state storage

NAND flash. Solid-state storage has a long history well before NAND flash became the dominating storage media. When NAND flash was introduced in the early 2000s, RAM-based SSDs were the dominating form of enterprise solid-state storage. By the mid 2000s, flash SSD vendors worked out performance and reliability problems with NAND flash and SATA SSDs started gaining widespread popularity as enterprise accelerators.

In the late 2000s, companies like Fusion-IO and Violin introduced a new breed of PCIe flash SSDs that could deliver one to two orders of magnitude higher throughput than their SATA counterparts. Since then, a rapid increase in capacity, drop in pricing, and new low-overhead interfaces like NVMe, have all resulted in PCIe flash SSDs displacing their SATA counterparts as server accelerators of choice. Table 3 (the first row) shows the price/performance characteristics of a representative, state-of-the-art PCIe SSD. Comparing this with Table 1, we can see that the PCIe SSD offers five times higher read IOPS, 12× higher write IOPS, and five times higher sequential access bandwidth than its SATA counterpart.

NVDIMM. As SSD vendors continue to improve throughput and capacity, the bottleneck in storage subsystem has shifted from the device itself to the PCIe bus that is used to interface with the SSD. Thus, over the past few years, NAND flash has started transitioning once again from storage devices that are interfaced via the high-latency (tens of μs), bandwidth-limited (tens of GBps), PCIe bus to *Persistent Memory* (PM) devices that are interfaced via the low-latency, high-bandwidth memory bus used for housing DRAM today. Today, PM devices use a combination of DRAM and flash storage media packaged together as a Dual Inline Memory Module (DIMM)–the standard form factor used for DRAM. Thus, these devices are also referred to as *Non-Volatile DIMMs* (NVDIMM).

Standardization efforts are well underway as SNIA and JEDEC have defined two types of NVDIMMs, namely, *NVDIMM-N*, *NVDIMM-F* that are already available today from several hardware vendors. NVDIMM-F is a flash-only DIMM that is effectively similar to PCIe SSDs on most aspects with the exception of the bus used to interface with the device (memory bus instead of PCIe). Thus, NVDIMM-F devices are block addressed and have access latencies faster than PCIe NAND flash due to the lack of PCIe overhead, but are still considerably slower than DRAM. *Memory1* from Diablo Technologies [29] is an example NVDIMM-F device. NVDIMM-N, in contrast, completely hides flash SSDs from the software stack. Only the on-board DRAM is used during normal operation and flash is used only during power failure as a backup media to checkpoint and restore DRAM content. While performance characteristics of NVDIMM-N are similar to DRAM, the limitations of DRAM power use and packaging issues result in capacities in the gigabyte range. The 8GB NVDIMM from Micron [17] that packages DRAM, flash, and a super-capacitor into a single module is an example of NVDIMM-N devices.

3D Xpoint. Today, NVDIMMs are still niche accelerators compared to PCIe SSDs due to a high cost/GB and relatively lower capacities. Unlike these NVDIMM technologies that rely on NAND flash, Intel's 3D XPoint is a new storage medium that is touted to have better endurance, higher throughput, and lower latency than NAND flash. Intel Optane DC P4800X is a recently announced PCIe SSD based on 3D XPoint technology [11]. Table 3 (the second row) shows the characteristics of this SSD. The cost/GB of 3D XPoint is much higher than NAND flash today as the technology is new and yet to mature. However, compared to NAND flash, 3D XPoint does not exhibit a huge variation between read and write throughput. Preliminary studies have also found that 3D XPoint provides predictable read/write access latencies that are much lower than several state-of-the-art NAND flash devices even under severe load [27].

4.2 Break-even interval and implications

Given these changes, it merits revisiting the five-minute rule for these new storage devices in the performance tier. When we apply the five-minute rule formula using price/performance metrics given in Table 3, the break-even interval we get is 41 seconds/one minute for reads/writes in the DRAM–NAND Flash PCIe SSD case, and 47 seconds/52 seconds for the DRAM–3D XPoint case.

Comparing these results with Table 2, we can see two important trends. First, the break-even interval is shorter when PCIe SSDs or new PM technologies are used as the second tier instead of SATA SSDs. This can be attributed to the drop in technology ratio caused by the improvement in random IOPS. As NAND flash and 3D XPoint-based devices migrate from PCIe to the memory bus, their throughput will increase further, which will result in a proportional reduction in the break-even interval. Second, in traditional SATA SSDs, the interval based on read IOPS is four times lower than the write IOPS-based interval. However, this difference is much smaller with modern PCIe and 3D XPoint-based SSDs. This indicates that random writes are no longer a limiting factor for these high-performance PCIe SSDs.

Implications. Today, in the era of in-memory data management, several database engines are designed based on the assumption that all data is resident in DRAM. However, the dramatic drop in break-even interval computed by the five-minute rule challenges this trend of DRAM-based in-memory data management due to three reasons. First, recent projections indicate that SSD density is expected to increase 40% annually over the next five years, outstripping the rate of increase of HDDs [4]. DRAM, in contrast, is doubling in capacity every three years [15]. As a result, the cost of NAND flash is likely to drop faster than DRAM. This, in turn, will result in the economic ratio dropping further leading to a reduction in the break-even interval.

Second, modern PCIe SSDs are highly parallel devices that can provide very high random I/O throughput by servicing multiple outstanding I/Os concurrently. With the introduction of interfaces like NVMe, the end-to-end latency of accessing data from PCIe NAND flash SSDs is just tens of microseconds. New non-volatile memory technologies like 3D XPoint promise further improvements in both throughput and access latencies over NAND flash. DRAM latency, in contrast, substantially lags behind the improvement in

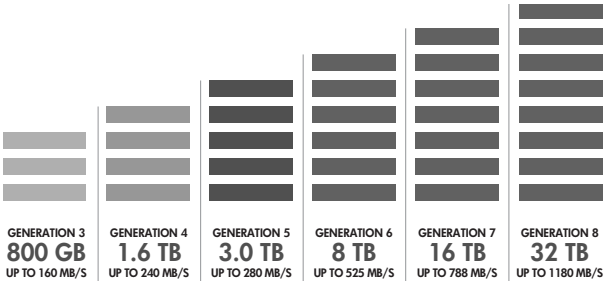

Figure 2: Tape roadmap published by LTO consortium [31]

bandwidth and has actually increased over the past few years. It is also well known that increasing the memory capacity of a system requires using high-density DRAM, and reducing access latency while increasing DRAM bit cell density is a challenging problem [19]. Thus, an improvement in the number of accesses per second of non-volatile solid-state storage media, whether it is based on NAND flash or 3D XPoint, will result in a drop in technology ratio, thereby reducing the break-even interval further.

Third, SSDs consume substantially lower power than DRAM. The 1TB Intel 750 SSD consumes 4W of power when idle and 22W when active. In contrast, 1TB of DRAM in a server would consume 50W when idle and 100W when active [1]. It is also well known that DRAM power consumption increases non-linearly with capacity, as high-density DRAM consumes substantially more power than their low-density counterparts. A recent study that focuses on power consumption in main memory databases showed that in a server equipped with 6TB of memory, the idle power of DRAM would match that of four active CPUs [1].

Such a difference in power consumption between SSD and DRAM directly translates into higher Operational Expenses (OPEX), and hence, higher Total Cost of Ownership (TCO), for DRAM-based database engines compared to their SSD-based counterparts. This is also reflected in the instance pricing of various memory and flash-optimized virtual machines offered by cloud providers. For instance, in the Amazon East (North Virginia) region, a memory-optimized X1 instance equipped with 976GB of DRAM costs $6.66 per hour, while a flash-optimized I3 instance with a 1TB NVMe SSD costs $0.31 per hour. Thus, we see that the cost benefit of SSD over DRAM increases from 8×, when the acquisition cost is used as the metric (Tables 1, 3), to 20× when instance pricing is used as the metric. Computing the break-even interval based on instance pricing would result in a further 2.5× reduction in interval due to a proportional drop in the economic ratio.

Given these three factors, the break-even interval from the five-minute rule seems to suggest an inevitable shift from DRAM-based data management engines to solid-state-storage-based persistent-memory engines. In fact, this change is already well underway, as current database engines are already being updated to fully exploit the performance benefits of PCIe NVMe SSDs. For instance, Oracle Exadata X6 has recently demonstrated 5.6M read IOPS and 5.2M write IOPS OLTP throughput using just NVMe SSDs [32].

5. FIVE-MINUTE RULE AND THE CAPACITY TIER

Unlike the performance tier, where the optimization target for storage devices is high random IOPS, devices in the capacity tier focus on improving cost/GB, as this tier is typically used to house data belonging to latency-insensitive batch analytics applications. In this section, we will present recent trends in the high-density storage hardware and reexamine the five-minute rule.

5.1 Trends in high-density storage

Traditionally, 7,200 RPM HDDs have been the primary storage media used for provisioning the capacity tier. For several years, areal density improvements enabled HDDs to increase capacity at Kryder's rate (40% per year), outstripping the 18-month doubling of transistor count predicted by Moore's law. However, over the past few years, HDD vendors have hit walls in scaling areal density with conventional Perpendicular Magnetic Recording (PMR) techniques. As a result, areal density improvement in HDDs over the past few years has fallen out of sync with Kryder's rate with an annual improvement of only around 16% instead of 40% [18].

HDDs also present another problem when used as the storage medium of choice for building a capacity tier, namely, high idle power consumption. Although enterprises gather vast amounts of data, as one might expect, not all data is accessed frequently. Recent studies estimate that as much as 80% of enterprise data is "cold", meaning infrequently accessed, and that cold data is the largest growing segment with a 60% Cumulative Annual Growth Rate (CAGR) [10, 12, 28]. Unlike tape drives, which consume no power once unmounted, HDDs consume a substantial amount of power even while idle. Such power consumption translates to a proportional increase in operational expenses. These issues associated with traditional PMR HDDs have resulted in a race for designing high-density devices that can substantially reduce the cost/GB of storing data.

5.1.1 Tape-based high-density storage

Unlike HDDs, the areal density of tapes has been increasing steadily at a rate of 33% per year and the LTO roadmap projects continued increase in density for the foreseeable future as shown in Figure 2. Table 4 shows the price/performance metrics of tape storage both in 1997 and today. The 1997 values are based on the corresponding five-minute rule paper [8]. The 2017 values are based on a SpectraLogic T50e tape library [26] using LTO-7 tape cartridges. Note here that we report the compressed capacity and bandwidth values for LTO-7 cartridges. Uncompressed values are half the reported values.

	1997	2017
Tape library cost ($)	10,000	11,000
Number of drives	1	4
Number of slots	14	10
Max capacity per tape	35GB	15TB
Transfer rate per drive (MB/s)	5	750
Access latency	30s	65s

Table 4: Price/performance characteristics of tape

Modern tape libraries use multiple tape drives and the cost varies depending on both the number of drives and the number of slots permissible. High-end tape libraries today can manage up to 50,000 tape cartridges with 144 drives. With individual tape capacity increasing 200× since 1997, the total capacity stored in tape libraries has expanded from hundreds of gigabytes to hundreds of petabytes today. Perhaps more interesting than improvement in cost/GB of tapes is the improvement in bandwidth. Today, a single LTO-7 cartridge is capable of matching, or even outperforming a HDD, with respect to sequential data access bandwidth as shown in Table 5. As tape libraries use multiple drives, the cumulative bandwidth achievable using even low-end tape libraries is 1–2GB/s. High-end libraries can deliver well over 40GB/s. However, random access latency of tape libraries is still 1000× higher than HDDs (minutes vs ms) due to the fact that tape libraries need to load and wind tape cartridges before data can be accessed.

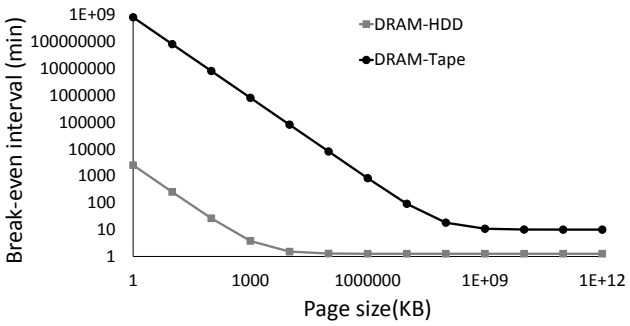

Figure 3: Break-even interval asymptotes for DRAM–HDD and DRAM–tape cases

5.1.2 HDD-based high-density storage

HDD vendors have also started working on several techniques to improve areal density. First, the use of Helium instead of air has allowed HDD vendors to pack more platters tightly and increase areal density. Second, while new magnetic recording techniques like Heat-Assisted Magnetic Recording (HMR), are being researched actively, HDD vendors have resorted to using Shingled-Magnetic Recording (SMR), where adjacent tracks are overlapped, as a stop-gap measure for further boosting density. Third, similar to the way tape libraries use multiple drives for increasing throughput, storage researchers and hardware vendors have recently started developing rack-scale storage devices, also referred to as *Cold storage devices* (CSD), that optimize cost/GB by treating a collection of high-density SMR HDDs as an ensemble.

Each CSD is a Massive Array of Idle Disks (MAID) in which only a small subset of HDDs is spun up and active at any given time [3]. For instance, Pelican CSD packs 1,152 SMR disks in a 52U rack for a total capacity of 5 PB [24]. However, only 8% of disks are spun up at any given time due to restrictions enforced by in-rack cooling (sufficient to cool only two disks per vertical column of disks) and a power budget (enough power to keep only two disks in each tray of disks spinning). Similarly, each OpenVault Knox [33] CSD server stores 30 SMR HDDs in a 2U chassis of which only one can be spun up to minimize the sensitivity of disks to vibration. The net effect of these limitations is that CSD enforce strict restrictions on how many and which disks can be active simultaneously. The set of concurrently active disks is referred to as a *disk group*.

All disks within a group can be spun up or down in parallel. Access to data in any of the disks in the currently spun up storage group can be done with latency and bandwidth comparable to that of the traditional capacity tier. For instance, Pelican, OpenVault Knox, and ArticBlue are all capable of saturating a 10-Gb Ethernet link as they provide between 1-2 GB/s of throughput for reading data from spun-up disks [24, 25, 33]. However, accessing data on a disk that is in a non-active group requires spinning down active disks and loading the next group by spinning up the new set of disks. This operation is referred to as a *group switch*. For instance, Pelican takes eight seconds to perform this group switch. Thus, the best case access latency of CSD is identical to the capacity tier and the worst-case latency is two orders of magnitude higher.

5.2 Break-even interval and implications

We will now revisit the five-minute rule for both DRAM–tape and HDD–tape cases. We do not compute the break-even interval for CSD as information about acquisition cost is not publicly available. However, based on TCO claims from CSD vendors, we will reason about how CSD affects the five-minute rule.

Metric	DRAM	HDD	Tape
Unit capacity	16GB	2TB	10 × 15TB
Unit cost ($)	80	50	11,000
Latency	100ns	5ms	65s
Bandwidth	100 GB/s	200 MB/s	4 × 750MB/s
Kaps	9,000,000	200	0.02
Maps	10,000	100	0.02
Scan time	0.16s	3hours	14hours
$/Kaps	9e-14	5e-09	8e-03
$/Maps	9e-12	8e-09	8e-03
$/TBscan	8e-06	0.003	0.03
$/TBscan (97)	0.32	4.23	296

Table 5: Price/performance metrics of DRAM, HDD, and tape

Using metrics from Table 1, and Table 4 to compute the break-even interval for the DRAM–tape case results in an interval of over 300 years for a page size of 4KB! Jim Gray referred to tape drives as the "data motel" where data checks in and never checks out [7], and this is certainly true today. Figure 3 shows the variation in break-even interval for both HDD and tape for various page sizes. We see that the interval asymptotically approaches one minute in the DRAM–HDD case and ten minutes in the DRAM–tape case. The HDD asymptote is reached at a page size of 100MB and the tape asymptote is reached at a size of 100GB. This clearly shows that: i) randomly accessing data on these devices is extremely expensive, and ii) data transfer sizes with these devices should be large to amortize the cost of random accesses.

However, the primary use of the capacity tier today is not supporting applications that require high-performance random accesses. Rather, it is to reduce the cost/GB of storing data over which latency-insensitive analytics can be performed. This is the reason why Gray and Graefe noted that metrics like KB-accesses-per-second (Kaps) are less relevant for HDD and tape as they grow into infinite-capacity resources [8]. Instead, MB-accesses-per-second (Maps) and time to scan the whole device are more pertinent to these high-density storage devices.

Table 5 shows these new metrics and their values for DRAM, HDD, and tape. In addition to Kaps, Maps, and scan time, the table also shows $/Kaps, $/Maps, and $/TBscan, where costs are amortized over a three-year timeframe as proposed by Gray and Graefe [8]. The scan metric can be considered as a measure of rent for a Terabyte of the media while the media is being scanned.

Looking at $/Kaps, we see that DRAM is five orders of magnitude cheaper than HDD, which, in turn, is six orders of magnitude cheaper than tape. This is expected given the huge disparity in random access latencies and is in accordance with the five-minute rule that favors using DRAM for randomly accessed data. Looking at $/Maps, we see that the difference between DRAM and HDD shrinks to roughly 1,000×. This is due to the fact that HDDs can provide much higher throughput for sequential data accesses over random ones. However, HDDs continue to be six orders of magnitude cheaper than tape even for MB-sized random data accesses. This, also, is in accordance with the HDD/tape asymptote shown in Figure 3. Finally, $/TBscan paints a very different picture. While DRAM remains 300× cheaper than HDD, the difference between HDD and tape shrinks to 10×.

Comparing the $/TBscan values with those reported in 1997, we can see two interesting trends. First, the disparity between DRAM and HDD is growing over time. In 1997, it was 13× cheaper to rent DRAM for a terabyte scan than HDD. Today, it is 300× cheaper. This implies that even for scan-intensive applications, unsurprisingly, optimizing for performance requires avoiding using HDD as the storage medium. Second, the difference between HDD and tape

is following the opposite trend and shrinking over time. In 1997, HDD was 70× cheaper to rent than tape. However, today, it is only 10× cheaper. Unlike HDDs, sequential data transfer bandwidth of tape is predicted to double for the foreseeable future. Hence, this difference is likely to shrink further. Thus, in the near future, it might not make much of a difference whether data is stored in a tape or HDD with respect to the price paid per scan.

Implications. Today, all data generated by an enterprise has to be stored twice, once in the traditional HDD-based capacity tier for enabling batch analytics, and a second time in the tape-based archival tier for meeting regulatory compliance requirements. The shrinking difference in $/TBscan between HDD and tape suggests that it might be worthwhile to revisit the traditional tiering hiearchy where tape is used only in the archival tier and never in an online fashion for supporting batch analytics; it might be economically beneficial to replace the HDD-based capacity and tape-archival tiers with a new *Cold Storage Tier* that subsumes the role of both tiers.

Recent application and hardware trends indicate that such a merge might be feasible. On the application front, batch analytics applications are scan-intensive and latency insensitive unlike interactive analytics applications that are latency sensitive. As interactive applications are already isolated to the performance tier, the Cold Storage Tier only has to cater to the bandwidth demands of batch analytics applications. On the hardware front, improvement in NAND flash density is outstripping that of HDD and tape today. As NAND flash grows in size, the performance tier grows with it proportionately. Thus, in the near future, it might be economically feasible to store all randomly accessed data, in addition to frequently accessed data for which fast access is necessary, on NAND flash in the performance tier. Thus, the capacity tier would be relegated to storing data that is accessed sequentially.

As we described earlier, nearline storage devices like tape libraries and CSD are capable of providing high-throughput access for sequentially accessed data. Further, our cost computation was based on acquisition cost and not TCO. Given the ever-increasing density, low power consumption, and higher longevity of tape, analysts have recently reported that tape-based archival solutions have up to 6× lower TCO per terabyte of data stored compared to HDD archives [21]. Cold Storage Device are touted to offer TCO and sequential data transfer bandwidth comparable to tape while lowering random access latency to seconds instead of minutes [16, 20]. Thus, replacing the capacity and archival tiers with a single Cold Storage tier could result in substantial cost savings for enterprises. Recent research has indeed shown this to be the case [2].

In order for such a Cold Storage Tier to be realized in practice, batch analytics applications should be modified to run directly over tape or CSD. Prior research in the database community on tertiary databases has already investigated techniques for such a storage setup [22, 23]. More recently, researchers have also started investigating extensions to batch processing frameworks for enabling analytics directly over data stored in tape archives and CSD. For instance, Nakshatra implements prefetching and I/O scheduling extensions to Hadoop so that map–reduce jobs can be scheduled to run directly on tape archives [13]. Skipper is a query processing framework that uses adaptive query processing techniques in combination with customized caching and I/O scheduling to enable query execution over CSD [2]. Skipper even shows that for long-running batch queries, using CSD results in query execution time increasing by only 35% compared to a traditional HDD despite the long disk spin-up latency. With such frameworks, it should be possible for installations to switch from the traditional four-tier hierarchy to a two-tier hierarchy with NAND flash or PM-based performance tier and CSD or tape-based Cold Storage Tier.

6. OUTLOOK AND CONCLUSIONS

The design of data management systems has always been driven by changes in application requirements and hardware. In this paper, we showed emerging trends with respect to both these aspects and revisited the five-minute rule.

On the application front, there is a clear distinction today between latency-sensitive interactive analytics and latency-insensitive batch analytics. As data belonging to these two classes of applications are accessed differently, they are stored in different tiers, namely, the performance and capacity tiers.

As NAND flash inches its way closer to the CPU from the PCIe bus to the DRAM bus in the performance tier, both latency and bandwidth have improved dramatically. For state-of-the-art PCIe SSDs, the break-even interval predicted by the five-minute rule is less than a minute. Going forward, further improvements in NAND flash, and the introduction of new PM technologies, will likely result in the break-even interval dropping further. As the data "reuse" window shrinks, it will soon be economically more valuable to store most, if not all, data on SSD or PM devices.

Traditionally, 7,200 RPM HDDs have been used for implementing the capacity tier. However, our analysis showed that bandwidth improvement in tape, and the introduction of CSD as a middle-ground between HDD and tape, may necessitate revisiting the bifurcation between capacity and archival tiers. Given the latency-insensitive nature of batch analytics, and recent work on modifying data analytics frameworks to run on tapes/CSD, it might be economically beneficial to merge the conventional capacity and archival tiers into a single Cold Storage Tier based on tape/CSD.

Both these findings suggest that the introduction of new storage devices in the performance and capacity tier could perhaps have the unexpected effect of shrinking, rather than extending, the storage hierarchy. In order for this to happen, however, several open questions must still be answered on both the hardware and software fronts. For instance, on the hardware front, in order for the Cold Storage Tier to be realized in practice, tape or CSD need to support batch analytics workloads. Given that these devices were traditionally used for archival storage, where data is rarely read, an interesting question is whether the reliability of these storage devices would be affected if they have to support batch analytics where data is read much more frequently, albeit in a sequential fashion.

On the software front, there is a stark difference in design between DRAM-based engines and their HDD-based counterparts. Thus, it is important to revisit the design tradeoffs for both PM devices and high-performance NVMe SSDs. For instance, disk-based engines use HDD as the primary storage tier and DRAM as a buffer cache based on the assumption that DRAM is limited in capacity. In-memory engines, in contrast, eliminate the overhead associated with buffer caching by treating DRAM as the primary storage tier and HDD/SSD as an anti-cache, based on the assumption that most data fits in DRAM [5]. NVMe SSDs and PM are fast enough to expose the overhead of buffer caching. Yet, anti-caching might not be the optimal way of managing them given that most data no longer resides in DRAM.

ACKNOWLEDGMENTS

We would like to thank the anonymous reviewers and the DIAS laboratory members for their constructive feedback that substantially improved the presentation of the paper. This work is funded by the School of Computer and Communication Sciences at EPFL and an industrial-academic collaboration between the DIAS laboratory and Huawei Central Software Institute (CSI).

7. REFERENCES

[1] R. Appuswamy, M. Olma, and A. Ailamaki. Scaling the memory power wall with dram-aware data management. In *DAMON*, 2015.

[2] R. Borovica-Gajic, R. Appuswamy, and A. Ailamaki. Cheap data analytics using cold storage devices. *PVLDB*, 9(12):1029–1040, 2016.

[3] D. Colarelli and D. Grunwald. Massive arrays of idle disks for storage archives. In *Conference on Supercomputing*, 2002.

[4] T. Coughlin. Flash memory areal densities exceed those of hard drives. https://www.forbes.com/sites/tomcoughlin/2016/02/03/flash-memory-areal-densities-exceed-those-of-hard-drives/#113504fe7c72.

[5] J. DeBrabant, A. Pavlo, S. Tu, M. Stonebraker, and S. Zdonik. Anti-caching: A new approach to database management system architecture. *PVLDB*, 6(14):1942–1953, 2013.

[6] G. Graefe. The five-minute rule 20 years later (and how flash memory changes the rules). *Commun. ACM*, 52(7):48–59, 2009.

[7] J. Gray. The five-minute rule. research.microsoft.com/en-us/um/people/gray/talks/fiveminuterule.ppt.

[8] J. Gray and G. Graefe. The five-minute rule ten years later, and other computer storage rules of thumb. *SIGMOD Rec.*, 26(4), 1997.

[9] J. Gray and F. Putzolu. The 5 minute rule for trading memory for disc accesses and the 10 byte rule for trading memory for cpu time. In *SIGMOD*, 1987.

[10] IDC. Technology assessment: Cold storage is hot again finding the frost point. http://www.storiant.com/resources/Cold-Storage-Is-Hot-Again.pdf.

[11] Intel. Optane technology. https://www.intel.com/content/www/us/en/architecture-and-technology/intel-optane-technology.html.

[12] Intel. Cold Storage in the Cloud: Trends, Challenges, and Solutions. http://www.intel.com/content/dam/www/public/us/en/documents/white-papers/cold-storage-atom-xeon-paper.pdf, 2013.

[13] A. Kathpal and G. A. N. Yasa. Nakshatra: Towards running batch analytics on an archive. In *MASCOTS*, 2014.

[14] B. Laliberte. Automate and optimize a tiered storage environment fast! ESG White Paper, 2009.

[15] K. Lim, J. Chang, T. Mudge, P. Ranganathan, S. K. Reinhardt, and T. F. Wenisch. Disaggregated memory for expansion and sharing in blade servers. In *ISCA*, 2009.

[16] S. Logic. Arctic blue pricing calculator. https://www.spectralogic.com/arcticblue-pricing-calculator/.

[17] Micron. Memory1. https://www.micron.com/products/dram-modules/nvdimm#/.

[18] F. Moore. Storage outlook 2016. https://horison.com/publications/storage-outlook-2016.

[19] O. Mutlu. Rethinking memory system design (along with interconnects). In *MEMCON*, 2015.

[20] S. Newsletter. Costs as barrier to realizing value big data can deliver. http://www.storagenewsletter.com/rubriques/market-reportsresearch/37-of-cios-storing-between-500tb-and-1pb-storiantresearch-now/.

[21] D. Reine and M. Kahn. Continuing the search for the right mix of long-term storage infrastructure a tco analysis of disk and tape solutions, 2015.

[22] S. Sarawagi. Query Processing in Tertiary Memory Databases. In *VLDB*, 1995.

[23] S. Sarawagi and M. Stonebraker. Reordering Query Execution in Tertiary Memory Databases. In *VLDB*, 1996.

[24] R. B. Shobana Balakrishnan, A. Donnelly, P. England, A. Glass, D. Harper, S. Legtchenko, A. Ogus, E. Peterson, and A. Rowstron. Pelican: A building block for exascale cold data storage. In *OSDI*, 2014.

[25] Spectra. Arcticblue deep storage disk. Product, https://www.spectralogic.com/products/arcticblue/.

[26] SpectraLogic. Spectralogic t50e. http://www.pcm.com/p/Spectralogic-Tape-Drives/product~dpno~8247348~pdp.gchecjj.

[27] StorageReview. Intel optane memory review. http://www.storagereview.com/intel_optane_memory_review.

[28] H. I. Strategies. Tiered storage takes center stage. Report, 2015.

[29] D. Technologies. Memory1. http://www.diablo-technologies.com/memory1/.

[30] TPC-C. Dell-microsoft sql server tpc-c executive summary. http://www.tpc.org/tpcc/results/tpcc_result_detail.asp?id=114112501, 2014.

[31] L. Ultrium. Lto ultrium roadmap. http://www.ltoultrium.com/lto-ultrium-roadmap/.

[32] K. Umamageswaran and G. Goindi. Exadata: Delivering memory performance with shared flash. http://www.oracle.com/technetwork/database/exadata/exadatadeliveringmemoryperformance-3518310.pdf.

[33] M. Yan. Open compute project: Cold storage hardware v0.5. http://www.opencompute.org/wp/wp-content/uploads/2013/01/Open_Compute_Project_Cold_Storage_Specification_v0.5.pdf, 2013.