Paging and the Address-Translation Problem

MICHAEL A. BENDER, Stony Brook University, Stony Brook, New York, USA
ABHISHEK BHATTACHARJEE, Yale University, New Haven, Connecticut, USA
ALEX CONWAY, Cornell University, New York City, New York, USA
MARTÍN FARACH-COLTON, New York University, New York City, New York, USA
ROB JOHNSON, VMware Research, Palo Alto, California, USA
SUDARSUN KANNAN, Rutgers University, New Brunswick, New Jersey, USA
WILLIAM KUSZMAUL and NIRJHAR MUKHERJEE, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA
DON PORTER, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA
GUIDO TAGLIAVINI, Google, Sunnyvale, California, USA
JANET VOROBYEVA, University of California San Diego, La Jolla, California, USA
EVAN WEST, Stony Brook University, Stony Brook, New York, USA

The classical paging problem, introduced by Sleator and Tarjan in 1985, formalizes the problem of caching pages in RAM in order to minimize IOs. Their online formulation ignores the cost of address translation: Programs refer to data via virtual addresses, and these must be translated into physical locations in RAM. Although the cost of an individual address translation is much smaller than that of an IO, every memory access involves an address translation, whereas IOs can be infrequent. In practice, one can spend money to avoid paging by over-provisioning RAM; in contrast, address translation is effectively unavoidable. Thus address-translation costs can sometimes dominate paging costs, and systems must simultaneously optimize both.

To mitigate the cost of address translation, all modern CPUs have translation lookaside buffers (TLBs), which are hardware caches of common address translations. What makes TLBs interesting is that a single TLB entry can potentially encode the address translation for *many* addresses. This is typically achieved via the use

This research was supported in part by NSF Grants Nos. CCF-2106827, CCF-1725543, CSR-1763680, CCF-1716252, CNS-1938709, CCF-1617618, CCF-1916817, CCF-2106999, CSR-1938180, CCF-1715777, and NRT-HDR 2125295, as well as an NSF GRFP fellowship and a Fannie and John Hertz Fellowship. This research was also partially sponsored by the United States Air Force Research Laboratory and was accomplished under Cooperative Agreement Number FA8750-19-2-1000. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the United States Air Force or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.
Authors' Contact Information: Michael A. Bender, Stony Brook University, Stony Brook, New York, USA; e-mail: bender@cs.stonybrook.edu; Abhishek Bhattacharjee, Yale University, New Haven, Connecticut, USA; e-mail: abhishek.bhattacharjee@yale.edu; Alex Conway, Cornell University, New York City, New York, USA; e-mail: ajc473@cornell.edu; Martín Farach-Colton, New York University, New York City, New York, USA; e-mail: martin@farach-colton.com; Rob Johnson, VMware Research, Palo Alto, California, USA; e-mail: robj@vmware.com; Sudarsun Kannan, Rutgers University, New Brunswick, New Jersey, USA; e-mail: sudarsun.kannan@rutgers.edu; William Kuszmaul, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA; e-mail: kuszmaul@cmu.edu; Nirjhar Mukherjee, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA; e-mail: nirjhar@cmu.edu; Don Porter, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA; e-mail: porter@cs.unc.edu; Guido Tagliavini (corresponding author), Google, Sunnyvale, California, USA; e-mail: guido.tag@gmail.com; Janet Vorobyeva, University of California San Diego, La Jolla, California, USA; e-mail: jvorobyeva@ucsd.edu; Evan West, Stony Brook University, Stony Brook, New York, USA; e-mail: etwest@cs.stonybrook.edu.

ACM 1549-6333/2025/9-ART43
https://doi.org/10.1145/3737700

of huge pages, which translate runs of contiguous virtual addresses to runs of contiguous physical addresses. Huge pages reduce TLB misses at the cost of increasing the IOs needed to maintain contiguity in RAM. This tradeoff between TLB misses and IOs suggests that the classical paging problem does not tell the full story.

This article introduces the Address-Translation Problem, which formalizes the problem of maintaining a TLB, a page table, and RAM in order to minimize the total cost of both TLB misses and IOs. We present an algorithm that achieves the benefits of huge pages for TLB misses without the downsides of huge pages for IOs.

CCS Concepts: • **Software and its engineering → Virtual memory**; • **Theory of computation → Caching and paging algorithms**; **Bloom filters and hashing**;

Additional Key Words and Phrases: virtual memory, address translation, TLB, paging, hashing, iceberg hashing

ACM Reference format:
Michael A. Bender, Abhishek Bhattacharjee, Alex Conway, Martín Farach-Colton, Rob Johnson, Sudarsun Kannan, William Kuszmaul, Nirjhar Mukherjee, Don Porter, Guido Tagliavini, Janet Vorobyeva, and Evan West. 2025. Paging and the Address-Translation Problem. *ACM Trans. Algor.* 21, 4, Article 43 (September 2025), 22 pages.
https://doi.org/10.1145/3737700

1 Introduction

In the classical *paging problem*, a sequence of page requests $p_1, p_2, \ldots$ must be serviced using a memory of size P pages [19, 21, 25, 49]. The cost of servicing a page request is 0 if the page is currently cached in memory. Otherwise there is a *page fault* and an *IO* must be performed, which means that the page must be fetched from disk to RAM (perhaps evicting another page) at a cost of 1.

Paging is critical to *virtual memory* systems, where programs reference pages by *virtual page addresses*. When a page is cached in memory, it also has a *physical page address* in the range $\{1, 2, \ldots, P\}$, specifying the location where it is actually stored. Every virtual address referenced by a program must be translated to a physical address by a process called **address translation (AT)**. If a virtual page does not have a physical address, the page's data must be fetched from external storage, a physical page must be allocated (and potentially freed first), the physical page filled with the contents from storage, and assigned to that virtual address. ATs are stored in an in-RAM dictionary called the *page table*.

AT incurs such a significant cost on real computers that modern CPUs come with specialized hardware accelerators called **translation lookaside buffers (TLBs)** that cache part of the page table. *TLB hits*, that is, successful lookups in the TLB, are fast, typically a single or small number of cycles [47]. In contrast, it can take hundreds or even thousands of CPU cycles to perform an AT in the page table, when there is a *TLB miss* [8, 33].

Although the cost of AT is ignored in the paging problem, the cost can be high—and can even dominate paging costs—because *every memory reference* must undergo AT, whereas page fetches may be rare. Moreover, one can avoid paging by purchasing more RAM, and this is generally considered money well spent. In contrast, TLBs have hit hard physical and power limits, making AT costs effectively unavoidable.

In this article, we address the algorithmic problem of how to organize both the TLB and the physical-address assignment in order to simultaneously optimize the total cost of AT and paging. We show that, by combining ideas from low-associativity paging, recent advances in hashing, and

compression, one can achieve strong, provable guarantees on the costs incurred by both the TLB and the page fetches.

Trends in the Cost of AT. AT overheads are becoming more significant because of several hardware trends. First, TLBs are too small to cache the working sets of modern parallel programs. Second, the access patterns of emerging workloads, such as machine learning and graph analytics, are irregular and difficult to prefetch. The increasing prevalence of parallel programming has led to recent TLBs allowing multiple threads (and even applications) to have entries in the TLB simultaneously [32], meaning that the effective size of the TLB is smaller for each thread. Additionally, in cloud environments, which increase parallelism by using virtual machines, each memory reference undergoes two translations—once in the guest and once in the host—which actually *squares* the cost of a TLB miss in the worst case [7]. Whereas the aforementioned trends result in increased pressure on the TLB and higher TLB-miss costs, trends towards faster storage devices lower the cost of paging, which further increases the relative overhead of AT.

Larger TLBs would have higher hit rates, but the size of TLBs is limited because it is expensive—in terms of time, transistors, and power [12]—to perform (parallel) hardware key-value lookups in tables with many entries. TLBs are so small that some workloads spend as much as 83% of their execution time on ATs [8] (see also [29, 30, 34, 50]).

The Ubiquity of TLBs. TLBs, and hence TLB performance bottlenecks, are also becoming more ubiquitous because of hardware and software trends. Traditionally, peripherals such as GPUs and network cards accessed RAM via physical addresses, and hence had no TLBs. Newer devices are beginning to support virtual memory in order to support safe, concurrent access by mutually distrusting users, as may occur when two virtual machines are sharing a hardware peripheral. For instance, recent GPUs by both Nvidia and AMD include page tables and TLBs so that multiple, unrelated kernels can run concurrently on different compute elements in the GPU. Newer network cards support *remote direct memory access*, in which the network card performs memory reads and writes based on incoming packets without going through the CPU [36, 53]. This is widely used to support concurrent access to memory in a cluster, and these cards have page tables and TLBs (albeit, with different names) in order to ensure the card performs only authorized memory accesses. And, of course, multi-core and multi-CPU systems can have per-core and per-CPU TLBs. The results in this article apply to all these TLBs in a modern computer.

Huge Pages and What Makes TLBs Interesting. What makes TLBs interesting is that, rather than caching data, they cache pointers to data. Notably, this means that a *single pointer* can potentially point to a *very large amount* of data.

Indeed, the main thrust of increasing the effectiveness of TLBs in systems design has been to use *huge pages*, which are runs of pages that are contiguous in the virtual address space [29]. Critically, existing huge-page methods require the run of pages also be placed contiguously in RAM (i.e., physical memory), so that a single TLB entry can translate any address in any page that is included in the huge page.[1] In this case, the TLB is used as a key-value store in which the keys are virtual addresses of huge pages (rather than of standard-size pages) and the values are physical addresses of huge pages (rather than of standard-size pages).

We call the set of page translations that a TLB entry encodes its *coverage*. If the coverage of a TLB entry forms a contiguous run of virtual addresses defined by the high-order bits of the

[1]In our discussion, we elide many details of huge pages and TLBs, such as that most systems that implement huge pages use different TLBs for each size [18, 56] and only between one and three sizes are allowed, depending on the implementation. The algorithmic problems are the same, whether we are considering TLBs in the wild or the semi-domesticated TLBs described here.

virtual addresses so encoded, we say that that entry encodes a *virtual huge page*. If, additionally, the corresponding physical pages are stored contiguously in *physical memory*, then we say those pages form a *physical huge page*.

Virtual and Physical Huge Pages, the Good and the Bad. Virtual huge pages are an effective technique for reducing TLB misses [34, 37], not merely because they increase the coverage of each individual TLB entry, but also because they translate a contiguous run of virtual addresses. Programs that exhibit spatial locality in their memory-access patterns benefit from the large coverage of each huge page.

On the other hand, physical huge pages *increase* IO costs for three reasons:

(1) *Page-Fault Amplification.* In order to represent a collection of pages as a physical huge page, whenever any page within a huge page is fetched from disk, all the constituent pages must also be fetched. This turns what would be an IO for a single block into IOs for many blocks.
(2) *Reduced RAM Utilization.* A physical huge page stores all the pages in its range, even if some are not accessed. This wastes RAM on pages that are not frequently referenced, thus leading to more page faults on pages that are more frequently referenced.
(3) *Fragmentation.* To mitigate these drawbacks, systems generally use a mix of regular and huge pages. Pages in a huge page are stored contiguously in RAM. To make room for them, any (non-huge) pages in the way must be evicted to disk, which can lead to IOs later when those evicted pages are re-accessed.[2]

In summary, huge pages come with a tradeoff: *Virtual* huge pages enable a reduction in TLB misses, but *physical* huge pages cause an increase in IOs. There is a vast architecture and operating systems literature on optimizing the benefits and costs of huge pages [24, 34, 35, 37–39, 44].

The Full Cost of AT. In order to fully quantify the cost of AT, one must consider TLB misses and IOs together. We call the software/hardware algorithm that manages the TLB, the page table, and the layout of pages in RAM a *memory-management algorithm*.

To measure the cost of a memory-management algorithm we introduce the *AT cost model*: Each IO costs 1, each TLB miss costs $\varepsilon \in (0, 1)$, and each TLB hit costs 0.

Huge-Page Decoupling: All the Virtual with None of the Physical. In this article, we are interested in designing memory-management algorithms that enjoy the TLB advantages of virtual huge pages without the IO costs of physical huge pages. A natural approach, which we call *huge-page decoupling*, is to break up the value part of every TLB entry into an array of physical addresses: The ith entry of the array encodes both whether the ith page in the huge page is in RAM, and if so, where the ith page resides in RAM. Huge-page decoupling would mitigate or eliminate the disadvantages of physical huge pages: It would increase RAM utilization by only storing pages that the paging algorithm deems useful; it would reduce page-fault amplification by reducing the footprint of each huge page; and it would eliminate fragmentation by obviating the requirement that the constituent physical pages of a virtual huge page be contiguous.

A priori, huge-page decoupling is not viable because the TLB value does not have enough bits to store such an array. Indeed, current TLBs are designed to store one physical address of $\log P$ bits. In general, we shall use w to denote the number of bits used for each TLB value, and we shall treat w as being set by hardware.

[2]Rather than evicting pages to disk, one could also try to defragment those pages in RAM [34, 35]. The challenge in practice is that the performance overheads of defragmentation, even in memory, can easily exceed the performance benefits of huge pages.

Better Encodings through Low Associativity. We show that huge-page decoupling schemes actually *are* possible by compressing the array of physical addresses as follows. Call a paging algorithm *L-associative* if each page has L possible locations where it can be stored. If L is small, and we use only $O(\log L)$ bits per physical page address, we can store multiple physical page addresses per TLB value.

Intuitively, low associativity may result in all L locations for a page being occupied by pages that the paging algorithm would prefer to keep. If this happens, then the paging algorithm must evict one of the pages, resulting in extra (and otherwise unnecessary) IOs. Hence, this low-associativity approach *also* appears at first sight to be a dead end: We replace the IOs needed for physical huge pages with the IOs needed for low associativity.

This Article. We show how to transform any paging algorithm into a low-associativity paging algorithm without increasing the IOs, while using minimal resource augmentation. Using this transformation, we can implement huge-page decoupling in order to realize the benefits of virtual huge pages without the need for physical huge pages.

Our main theorem is that we can simultaneously match the TLB miss rate of any memory-management algorithm (even one that only cares about minimizing TLB misses) while matching the IOs of any other memory-management algorithm (even one that only cares about minimizing IOs).

Section 2 discusses the results in more depth.

1.1 Impact of This Work

This work is the inspiration and culmination of several lines of algorithmic work. Recent results on hashing, succinct data structures, and low-associativity paging were all directly inspired by the AT problem that we tackle in this article. For example,

- *Hashing*. Iceberg hashing, a new hash table construction with several desirable theoretical properties, was actually a consequence of the current work: Several of the current authors developed Iceberg hashing to resolve the technical problems encountered in Section 4. It turns out that the hash table that one gets from this ends up having several independently desirable properties, both in theory [9] and practice [31]. This is an example of how exploring new models can lead to unexpected and fruitful algorithmic strands.
- *Compressed Data Structures*. Subsequent work introduces the notion of a *tiny pointer* [10], which extends and generalizes the physical-address compression scheme in this article, and can be used to break through several lower bounds in the succinct data structures literature.
- *Low-Associativity Paging*. Finally, recent work on low-associativity paging [11] is inspired in part by the results in this article and shows that it is possible to get low paging costs and low associativity—but provably not as low as in this article—by using power-of-one-choice hashing, the scheme that has been used for decades in real hardware.

This work brings many of the above advances back home into a framework for reasoning simultaneously about AT and paging.

On the practical side, this article has served as the theoretical foundation for the Mosaic Paging System [31], an architectural design for TLBs proposed and evaluated at ASPLOS '23. Whereas the current article formalizes the constraints on a huge-page decoupling scheme and proves the theoretical existence of schemes with strong guarantees, the subsequent work [31] has shown that these techniques are not just theoretical—they can be realized in hardware simulations with significant real-world performance improvements. Mosaic Paging won the Distinguished Paper

Award at ASPLOS'23 (a major computer architecture conference) and was selected as one of IEEE's MICRO Top Picks for 2024.

In summary, the AT problem lies at a surprisingly fertile nexus of hashing, succinct data structures, and external-memory algorithms. And we believe there may be opportunities for more research inspired by this problem. For example, our current model largely ignores the page table, treating it as a block box with uniform access cost ε. In reality the page table is implemented as a data structure, and so it may favor some access patterns over others (e.g., with or without locality). Thus, there may be further opportunities to codesign the TLB and page table so that the TLB's misses form an access pattern that is particularly efficient in the page table.

2 Technical Overview and Outline

This section gives a detailed overview of the results and main technical ideas in the article.

2.1 Technical Overview

Before we describe the contributions of each individual section, we first give a high-level overview of both the problem that we formalize and our algorithmic approach.

Modeling AT and Paging Costs. From a modeling perspective, the main contribution of the article is to formalize a single problem that captures the entire AT pipeline, including both the TLB, and the paging algorithm that assigns virtual pages to physical addresses in RAM. At first, this may seem like a step backwards—shouldn't we try to isolate distinct problems within the pipeline and model them separately? It turns out, however, that by placing the entire pipeline into a single model, we can identify an algorithmic opportunity that has previously been missed. The algorithmic contribution of the article is then to show how to use techniques from the seemingly unrelated succinct-hash-table literature in order to obtain an improved solution to the AT problem.

Rather than starting with the model, let us instead describe the algorithmic opportunity that our model ends up revealing. The basic insight is that, if we modify how TLBs are implemented, then it is possible to decouple the size of pages used in TLB translation from the size of pages stored in RAM. That is, we can use a large page size for the TLB while using another smaller page size for RAM. The TLB gets all the benefits of huge pages (i.e., increased coverage) without passing any of the drawbacks (e.g., fragmentation) on to RAM. We refer to this concept as *huge-page decoupling*.

The basic idea behind huge-page decoupling is that each TLB entry will map a virtual huge page address not just to a single physical page address, but to a bundle of $h_{\max}$ physical page addresses, for some $h_{\max} > 1$. This lets us partition each virtual huge page into multiple physical pages in RAM, all of the addresses of which get stored in the virtual huge page's TLB entry.

If implemented naively, this bundling approach would suffer from a major problem: If we plan to store $h_{\max}$ physical page addresses in each TLB entry, then we will need to significantly increase the number of bits used to implement each entry. An important insight in this article is that, if we design a different part of the system in just the right way (namely, the page-replacement policy used in RAM) then we can actually significantly reduce the amount of information needed to encode each physical page's address. This will allow us to store all $h_{\max}$ physical page addresses within the bits normally used to store a single address.

The Core Algorithmic Problem: Low-Associativity Paging. Thus the task of performing huge-page decoupling reduces to the following question: How can a RAM page-replacement policy allow us to compress physical page addresses? Suppose that the page-replacement policy has *low associativity*: This means that, for a given virtual page address, there are some bounded number s of possible places where the corresponding physical page could be in RAM. Then, if we already know the virtual address of a page, the only information that we need to store in order to recover the physical

address is *which* of these s options are being used. Rather than storing an entire physical page address we can instead store a number from the set $\{1, 2, \ldots, s\}$.

In Section 4, we design a low-associativity page-replacement algorithm using a bucketing scheme in which each page is mapped to one or more buckets and must reside within them. If the buckets are small, then the associativity is small as well. Our bucketing scheme is inspired from results on the construction of succinct hash tables [9], so our main task is to translate it into a caching policy that can be incorporated into our model.

This brings us to the following situation: The good news is that, if our page-replacement policy has low associativity, then we can perform huge-page decoupling. The bad news is that, if we restrict our page-replacement policy to have low associativity, then *a priori*, we may not be able to allocate all of the pages in RAM that we want to place there. Thus, we arrive at the following self-contained algorithmic subproblem: Given an online paging policy that we *want* to use, but that does not have low associativity, is it always possible to design a new policy that is nearly as good, but that does have low associativity? Our final algorithmic contribution is to answer this question in the affirmative with a general-purpose simulation result.

Summary and Discussion. When we put all the pieces together, we get two contributions: a model of page-translation cost that allows for the full spectrum of algorithms that one might consider, including less conventional ones such as those in this article; and a new set of algorithms within the model that allow us to decouple the page sizes used within the TLB from those used within RAM. Our final result (Theorem 4) captures the formal sense in which our techniques allow for the best of two worlds: a world where the TLB gets to use huge pages, but where RAM gets to use small pages.

We remark that there are a few ways in which the model in this article is somewhat unusual for an algorithmic model. First, it captures a problem solved by *hardware*. This means that any new algorithm will also require hardware modifications. Second, the model is quite involved, capturing multiple levels of abstraction (the TLB, RAM, and disk) and how they interact with each other. As we noted earlier, this is actually a necessary evil: Only by understanding the entire pipeline can we reveal the algorithmic opportunities for improvement presented in this article.

2.2 Outline and Results

Section 3: Huge-Page Decoupling. Section 3 formalizes *huge-page decoupling*. Recall that a huge-page decoupling scheme encodes in the TLB value for a virtual huge page, all of the information of *which* of its constituent virtual pages are present in RAM, and *where* those pages are located. The guarantee of a huge-page decoupling scheme is that the TLB can use virtual huge pages of some large *huge-page size* $h_{\max}$ pages, and RAM can be allocated at the granularity of normal-sized pages. That is, the choice of which normal-size pages are in memory can be made independently of the choice of which virtual huge pages are in the TLB.

Our approach to implementing huge-page decoupling is to treat RAM as a low-associative cache; we will avoid increasing IO cost by making use of a small amount of resource augmentation. That is, we equip a huge-page decoupling scheme with a *resource-augmentation parameter* δ, and the huge-page decoupling scheme may assume that there are never more than $(1 - \delta)P$ pages stored in RAM at a time.

The first goal of a huge-page decoupling scheme is to achieve a value of $h_{\max}$ that is as close as possible to the number w of bits that are used for each TLB value. Naturally, $h_{\max}$ cannot be arbitrarily large. There is a natural upper bound of:

$$h_{\max} \leq w, \tag{1}$$

since each TLB value must use $h_{\max}$ bits to encode what subset of the pages in a huge page are present in memory.

The second goal of a huge-page decoupling scheme is to minimize the resource-augmentation parameter δ. If δ is small and the huge-page size $h_{\max}$ is large, then huge-page decoupling allows us to have large virtual huge pages at essentially no cost.

In this article, we are able to come remarkably close to meeting the upper bound (Equation (1)) on $h_{\max}$, achieving:

$$\begin{cases} h_{\max} = \Theta(w/\log\log\log P) \\ \delta = o(1). \end{cases} \tag{2}$$

Recall that P is the number of pages that fit in physical memory; another interpretation of Equation (2) is that, for each of the $h_{\max}$ physical pages that a TLB entry points to, we can encode the physical location of that page using only $\Theta(\log\log\log P)$ bits.

Section 4: Low-Associativity Paging and Compact TLB Encodings. In Section 4, we address the main challenge in designing a huge-page decoupling scheme, which is how to encode all of the information that we wish to store in each TLB value using only w bits. Our huge-page decoupling scheme must be able to use just $w/h_{\max}$ bits to encode the location of each page in a huge page.

A central technical idea in our TLB encodings is to re-purpose a classic technique in caching: low associativity. The idea of low associativity is to break the cache into small bins, hash each cache entry to a random bin, and then manage each bin individually via a paging algorithm, such as LRU. The bin size is referred to as the *associativity* (or *set-associativity*) of the cache. Typically, the purpose of low-associativity caching is to simplify the task of implementing a cache (especially in hardware caches).

In this article, we use low-associativity caching in a starkly different way. By limiting the number of options for where each page can be placed in physical memory, the physical page addresses in the TLB can also be encoded using very few bits. At the same time, the associativity must be large enough that, whenever a page is brought into memory, there is a legal position where it can be placed (i.e., a free position in the right bin). As a warmup result, we show that by setting the associativity to be $\tilde{\Theta}(\log P)$, and using resource augmentation $\delta = o(1)$, we can construct a simple huge-page decoupling scheme with virtual huge pages of size $h_{\max} = \Theta(w/\log\log P)$.

A key insight of this article is that we can use recent advances in the design and analysis of balls-and-bins games to achieve an even smaller associativity. By employing the ICEBERG$[d]$ balls-in-bins strategy, we show how to construct a huge-page decoupling scheme with $h_{\max} = \Theta(w/\log\log\log P)$ and $\delta = o(1)$.

Section 5: Optimizing the Cost of AT. Finally, we consider the task of optimizing the total TLB and IO cost of a memory-management algorithm on a sequence of page requests. For any memory-management algorithm $\mathcal{Z}$ and sequence of page requests $\sigma = (p_1, p_2, \ldots, p_n)$, let $C(\mathcal{Z}, \sigma)$ denote the total cost of $\mathcal{Z}$ in the AT cost model, let $C_{\text{TLB}}(\mathcal{Z}, \sigma)$ denote the total cost incurred due to TLB misses, and let $C_{\text{IO}}(\mathcal{Z}, \sigma)$ denote the total cost incurred due to IOs. To minimize $C(\mathcal{Z}, \sigma)$ we must simultaneously optimize $C_{\text{TLB}}(\mathcal{Z}, \sigma)$ and $C_{\text{IO}}(\mathcal{Z}, \sigma)$.

We prove that the problem of optimizing $C_{\text{TLB}}(\mathcal{Z}, \sigma)$ can be separated from the problem of optimizing $C_{\text{IO}}(\mathcal{Z}, \sigma)$, in the following sense. Let $\mathcal{X}$ and $\mathcal{Y}$ be arbitrary memory-management algorithms, each of which is allowed to use any mixture of huge-page sizes between 1 and $h_{\max}$. The only constraint on $\mathcal{X}$ and $\mathcal{Y}$ is that they operate on a physical memory of size $(1-\delta)P$ (rather than the full physical memory of size P). Using huge-page decoupling, we construct a new memory-management algorithm $\mathcal{Z}$ with the following guarantee. With high probability in P, the total cost

of $\mathcal{Z}$ satisfies:

$$C(\mathcal{Z}, \sigma) \leq C_{\text{TLB}}(\mathcal{X}, \sigma) + C_{\text{IO}}(\mathcal{Y}, \sigma) + \frac{n}{\text{poly}(P)}. \tag{3}$$

Importantly, even if $\mathcal{X}$ minimizes TLB misses (by using huge pages) and $\mathcal{Y}$ minimizes IOs (by not using huge pages), then $\mathcal{Z}$ combines the best performance features of $\mathcal{X}$ and $\mathcal{Y}$. The additive term in Equation (3) says that a vanishingly small fraction $1/\text{poly}(P)$ of page accesses are permitted to be page faults in $\mathcal{Z}$ despite not being page faults in $\mathcal{Y}$.

Related Work. Section 6 discusses related work in depth, both from a theoretical as well as practical standpoint.

3 Huge-Page Decoupling

The idea behind huge-page decoupling is to enable the use of huge pages in the TLB, while letting the paging algorithm operate on normal pages. This will work by encoding the physical page addresses in the TLB entry for a given virtual huge page.

More precisely, a huge-page decoupling scheme takes as input a page replacement policy for RAM (the *RAM-replacement policy*) and a huge-page replacement policy for the TLB (the *TLB-replacement policy*). It consists of a RAM-allocation scheme that reduces the associativity of page placements in RAM, and an encoding/decoding scheme for translating between TLB values and physical addresses. All of these components must interact carefully. For example, the RAM-allocation scheme not only must achieve low associativity but also must be amenable to fast encoding and decoding; and unlike a standard TLB, which only covers virtual addresses that are mapped in physical memory, the encoding/decoding scheme must specify if a page is mapped or not. With so many moving parts, we take this section to carefully define all components of the system and their requirements before moving on to our main theorems.

Recall that the goals of a huge-page decoupling scheme are to maximize the size $h_{\max}$ of huge pages in the TLB, and minimize the resource-augmentation parameter $\delta \in (0, 1)$ for RAM.

The Input Replacement Policies. Let V be the number of pages in virtual memory and P be the number of pages in physical memory. A *virtual page address* is any element of $[V] = \{1, 2, \ldots, V\}$. A *physical page address* is any element of $[P] = \{1, 2, \ldots, P\}$. A *virtual huge-page address* is any element of $[V/h_{\max}]$ (we assume $h_{\max}$ divides V). We use ℓ to denote the number of entries in the TLB, and w to denote the number of bits in each TLB value.

The RAM-replacement policy determines which virtual page addresses are in RAM at any given moment; we refer to the set of such addresses as the *active set* $\mathcal{A} \subseteq [V]$. The only restrictions on the RAM-replacement policy are that $|\mathcal{A}| \leq (1 - \delta)P$ at all times, and that it is oblivious to the state and operation of the huge-page decoupling scheme.

The TLB-replacement policy determines which virtual huge-page addresses are in the TLB at any given moment; we refer to the set of such addresses as $\mathcal{T} \subseteq [V/h_{\max}]$. The only restrictions on the TLB-replacement policy are that $|\mathcal{T}| \leq \ell$ at all times, and that it is oblivious to the state and operation of the huge-page decoupling scheme.

The Huge-Page Decoupling Scheme. A huge-page decoupling scheme is an algorithm with three parts: a *RAM-allocation scheme*, a *TLB-encoding scheme*, and *TLB-decoding scheme*.

The *RAM-allocation scheme* determines the physical address for each page fetched by the RAM-replacement policy. At any given moment in time, we use $\phi : \mathcal{A} \to [P]$ to denote the physical address $\phi(v)$ corresponding to each virtual page address $v \in \mathcal{A}$. The RAM-allocation scheme gets to decide the value of $\phi(v)$ whenever a new page is added to $\mathcal{A}$ by the RAM-replacement policy.

The only restrictions on the RAM-allocation scheme are that ϕ must always be an injection and that ϕ must be *stable*—that is, once a virtual page $v \in \mathcal{A}$ is assigned a physical address $\phi(v)$, that address cannot change until v is removed from $\mathcal{A}$.

In order to establish probability bounds over arbitrarily long sequences of requests for RAM-allocation schemes that have less than full associativity, we must deal with what happens when page v experiences a *paging failure*, that is, when it is added to $\mathcal{A}$ by the RAM-replacement policy but cannot be assigned a physical address by the RAM-allocation scheme. This can happen if, for example, the set of physical addresses in which the RAM-allocation policy can place v are all occupied. Since the RAM-allocation policy is not allowed to move or evict pages, v cannot be placed and a paging failure occurs. The paging failure associated with v lasts until the RAM-replacement policy evicts v.

The *TLB-encoding scheme* determines the w-bit TLB value for each virtual huge page in the TLB. At any given moment in time, we use $\psi : \mathcal{T} \to [2^w]$ to denote the current set of TLB values. The value of $\psi(v)$ is set (resp. unset) by the TLB-encoding scheme when the TLB-replacement policy inserts (resp. removes) v from $\mathcal{T}$. And the value of $\psi(v)$ is updated by the TLB-encoding scheme whenever any of the constituent virtual page addresses u of v are added or removed from $\mathcal{A}$ by the RAM-replacement policy.

The *TLB-decoding scheme* translates TLB values into physical addresses via a *TLB-decoding function* $f : ([V] \times [2^w]) \to ([P] \cup \{-1\})$. The TLB-decoding function must offer the following guarantee: If u is a virtual huge-page address in $\mathcal{T}$, and v is a virtual page address contained in u, then:

$$f(v, \psi(u)) = \begin{cases} \phi(v) & \text{if } v \in \mathcal{A} \\ -1 & \text{otherwise.} \end{cases} \tag{4}$$

In other words, for every virtual page address v that is both contained in u and is in the active page set $\mathcal{A}$, the TLB-decoding function must be able to recover the physical page address $\phi(v)$ associated with v. And for every virtual page address v that is contained in u but is not in the active page set $\mathcal{A}$, the decoding function must encode the fact that v is not in $\mathcal{A}$ by returning the null address -1.

The TLB-decoding function f is determined once at the beginning of time and cannot be subsequently changed. The function f is permitted to be randomized (and thus can read the random bits used by our algorithm).

Constant-Time High-Probability Decoupling Schemes. A huge-page decoupling scheme is said to be *constant time* if, each time that the TLB-replacement policy modifies $\mathcal{T}$ or the RAM-replacement policy modifies $\mathcal{A}$, the huge-page decoupling scheme spends time $O(1)$ updating ϕ and ψ; and if the TLB-decoding function f can be evaluated in time $O(1)$.

Recall that RAM-allocation schemes can experience paging failures when a page v is placed into RAM by the RAM-replacement policy but cannot be placed by the RAM-allocation scheme. Paging failures last until the associated page is removed by the RAM-replacement policy. We use $\mathcal{F} \subseteq \mathcal{A}$ to denote the set of virtual page addresses on which paging failures are occurring at a given moment. A randomized huge-page decoupling scheme is said to succeed *with high probability in P* if, at every point in time, the probability that $|\mathcal{F}| > 0$ is at most $1/\text{poly}(P)$. Later in the article, when we use huge-page decoupling to construct efficient memory-management algorithms, we will handle paging failures by temporarily bringing the affected page into RAM whenever it is needed, and then allowing the page to subsequently be paged back out to disk.

4 Low-Associativity Paging and Compact TLB Encodings

The key challenge in designing a huge-page decoupling scheme is to limit the associativity of the RAM-allocation scheme, so that each page in RAM has only a small number of options for where it can reside. At the same time, we must support an arbitrary RAM-replacement policy (i.e., the paging algorithm for managing which pages are in RAM), whose only constraint is that it never places more than $(1-\delta)P$ pages in RAM at a time.

We limit the associativity of the RAM-allocation scheme by partitioning RAM into n buckets, each one comprising $B = P/n$ consecutive pages. To place a page in RAM, we randomly choose k buckets by computing k hash functions of the virtual page address; we select one of the buckets; and we place the page in some free slot within the chosen bucket. This yields an associativity of kB.

The bucket size B controls a tradeoff between associativity and IO complexity: The smaller the B, the more likely it is for a page to find all of its k chosen buckets already full.

We show that, surprisingly, any (oblivious) RAM-replacement policy can be implemented with a low-associativity RAM-allocation scheme, using a small amount of resource augmentation. Our main theorem in this section is that this can be attained using $k = 3$ hash functions and buckets of size $B = \tilde{\Theta}(\log\log P)$, ultimately leading to a decoupling scheme that achieves $h_{\max} = \Theta(w/\log\log\log P)$ and $\delta = o(1)$.

We begin the section by showing, as a warmup, how to achieve $h_{\max} = \Theta(w/\log\log P)$ using $k = 1$ hash functions. We then extend the result to use $k = 3$ in order to achieve the stronger bound of $h_{\max} = \Theta(w/\log\log\log P)$.

Balls-and-Bins Games. We model RAM-allocation algorithms as *dynamic balls-and-bins games*. In our balls-and-bins game, there are n bins, and there is an adversary that specifies an arbitrary sequence of ball insertions and deletions (and perhaps re-insertions), such that there are never more than m balls in the system. On each insertion, a ball is thrown into some bin according to a rule that randomly chooses k bins and places the ball in one of them. The goal is to design the placement rule, such that it minimizes the maximum load across all bins. Importantly, the adversary is oblivious to the game's randomness; otherwise it could force all balls to go to the same bin.

The relationship between RAM-allocation schemes and balls-and-bins games is as follows. Each bin represents a bucket in RAM, and each ball represents a page. The adversary is the RAM-replacement policy (and the sequence of page requests), and the balls insertions/deletions correspond to page insertions/deletions in $\mathcal{A}$. Based on this analogy, we can use n and m in both contexts. We will use $\lambda = m/n$ to denote the (maximum allowable) average occupancy of the bins.

Observe that not every balls-and-bins game models a RAM-allocation scheme—it has to be *online* (i.e., balls are sequentially placed before seeing future requests) and *stable* (i.e., balls are not moved around once inserted). Both of these features are required in a huge-page decoupling scheme: page requests (and, thus, TLB and paging operations) are served in an online fashion, and the physical address of a page must not be changed until the page is swapped out.

The Difficulty of Reducing Associativity. Suppose that a single hash function is used (i.e., $k = 1$) and that buckets have size $B = 1$, so that the associativity is 1. Then, physical addresses do not require any bits at all—virtual addresses are translated simply by computing their hash value, and thus no translations need to be cached in the TLB. The problem, of course, is that this configuration lends to a prohibitively large number of paging failures (recall that huge-page decoupling schemes must incur *no* paging failures with high probability in P, at any given point in time). To quantify this statement, consider a sequence of P distinct page accesses, starting from an empty RAM. By a standard balls-and-bins argument, where balls represent pages and bins represent page slots in physical memory (the unit-sized buckets), approximately P/e slots remain unused, with high

probability in P. Thus, any paging algorithm that doesn't evict pages during the first $(1-\delta)P$ insertions (e.g., LRU, FIFO) will incur at least $(1/e-\delta)P$ paging failures with high probability in P. For $\delta = o(1)$, this is $\Omega(P)$ paging failures with high probability in P.

Achieving Associativity $\tilde{\Theta}(\log P)$ with $k = 1$. Let $m = (1-\delta)P$ be the maximum number of pages that the RAM-replacement policy can cache simultaneously. We specify the bin size B and δ (and, thus, also m and n) below. For now, we use $k = 1$, which means that each ball is simply assigned to a random bin. To prevent bins from overflowing, we must set the bin size B to be large enough that the maximum load of any bin is at most a $1+\delta$ factor larger than the average load. On the other hand, subject to bins not overflowing, we want B as small as possible to obtain a small associativity.

Since $k = 1$, at any given moment, the maximum load is:

$$\begin{cases} (1+o(1))\frac{\log n}{\log(\log n/\lambda)} & \text{if } 1 \le \lambda = o(\log n) \\ \Theta(\lambda) & \text{if } \lambda = \Theta(\log n) \\ \lambda + O(\sqrt{\lambda \log n}) & \text{if } \lambda = \omega(\log n), \end{cases} \tag{5}$$

with high probability in n [46]. Thus, bin sizes B that allow for a $\delta = o(1)$ are in the third case.

Set the number of bins to be $n = m/(\log P \log\log P)$, so that the average load is $\lambda = \log P \log\log P = \omega(\log n)$. Then, with high probability in n (and thus P), the maximum load is:

$$\begin{aligned} \lambda + O(\sqrt{\lambda \log n}) &= \lambda + O(\log P \sqrt{\log\log P}) \\ &= \lambda(1+\delta), \end{aligned}$$

where $\delta = O(1/\sqrt{\log\log P})$.

Note that the bucket size satisfies:

$$B = \frac{P}{n} = \frac{P}{m} \cdot \lambda = \frac{P}{(1-\delta)P} \cdot \lambda = \frac{1}{1-\delta} \cdot \lambda > (1+\delta)\lambda,$$

which means that B is at least as large as the maximum load of the balls-and-bins game. Thus, every page fits in RAM at any fixed point in time, with high probability in P. Since addresses have size $\log B = \Theta(\log\log P)$, we get a decoupling scheme with huge-page size $h_{\max} = \Theta(w/\log\log P)$.

Theorem 1. *There exists a constant-time huge-page decoupling scheme using resource augmentation $\delta = o(1)$ that supports huge-page size $h_{max} = \Theta(w/\log\log P)$ with high probability in P.*

Proof. Recall that a huge-page decoupling scheme consists of three parts: a RAM-allocation scheme, a TLB-encoding scheme, and a TLB-decoding scheme. By having the RAM-allocation scheme use the balls-and-bins strategy described above, we ensure that each page has at most B positions where it can reside, where $B = \Theta(\log P \log\log P)$.

The TLB-encoding and decoding schemes can treat each TLB value as an array of $\Theta(\log\log P)$-bit elements $a_1, a_2, \ldots, a_{h_{\max}}$. If v is the ith page in the huge page represented by the TLB entry, and v hashes to bin j, then a_i indicates the position in bin j where v resides (or -1 if v is not in $\mathcal{A}$). Note that the huge-page decoupling scheme is easily made constant time by maintaining a hash table that keeps track of what the current value of $\psi(u)$ should be for each virtual huge page u that has at least one constituent page in RAM.

Our final task is to analyze the size of the failure set $\mathcal{F}$. For the sake of analysis, whenever a ball is inserted into a bin, label the ball as *failed* if the ball is inserted into a bin that already has B other balls (that are not labeled as failed). Note that the ball retains its failed label even if subsequently the load of the bin falls below B. From the perspective of the balls-and-bins game, failed balls are like any other balls. On the other hand, for the huge-page decoupling scheme, failed balls correspond to paging failures. That is, $|\mathcal{F}|$ is equal to the number of balls in the system that have the failed

label. At any given moment, there are up to $m = O(P)$ balls $b_1, \ldots, b_m$ present. For each i, when b_i was inserted it had a $1/\text{poly}(P)$ probability of being labeled as failed. By a union bound, it follows that $\Pr[|\mathcal{F}| > 0] \leq m/\text{poly}(P) = 1/\text{poly}(P)$, as desired. □

Achieving Associativity $\tilde{\Theta}(\log\log P)$ with $k = 3$. A natural way to try to improve the associativity further is to use the balls-and-bins rule known as GREEDY[2], in which each ball chooses two bins independently at random, and the ball is placed in the less full bin. With this rule, the maximum load at any moment is at most:

$$O(\lambda) + \log\log n + O(1), \tag{6}$$

with high probability in n [51]. This approach fails because the difference between the average load λ and the bound on the maximum load is $\Omega(\lambda)$, no matter what we choose λ to be.[3] Therefore, this forces the use of $\delta = \Omega(1)$ resource augmentation. Using GREEDY[d] for $d > 2$ doesn't help the situation, because the maximum load still grows as $O(\lambda)$ rather than λ.

Until recently, no balls-and-bins strategy was known to be simultaneously online, stable, and to have a maximum load of $(1 + o(1))\lambda + O(\log\log n)$. The authors of this article have another paper under submission that presents a balls-and-bin rule that has all of these features [9]. The rule, which is called ICEBERG[d], chooses $d + 1$ bins per ball. In the case of $d = 2$, it attains the following bound.

Theorem 2 [9]. *With high probability in n, at any fixed point in time the maximum load of ICEBERG[2] is at most:*

$$(1 + o(1))\lambda + \log\log n + O(1)$$

in the dynamic setting, against any oblivious adversary.

For concreteness, we sketch out ICEBERG[2] here. Balls are placed into bins using three independent hash functions h_1, h_2, h_3. When inserting a ball x, we first look at bin $h_1(x)$ and insert the ball there if the number of balls in the bin is less than $(1 + o(1))\lambda$. Otherwise, we use h_2 and h_3 to insert via GREEDY[2].[4] Intuitively, the reason that ICEBERG[2] works so well is that even though the vast majority of balls get inserted using h_1, their contribution to the maximum load is capped at $(1 + o(1))\lambda$ (because, beyond that point, balls are inserted using GREEDY[2]). Thus, the number of balls managed by GREEDY[2] at any given moment is only $O(n)$; and so the average load of GREEDY[2] is $\lambda = O(1)$ and the known bound (Equation (6)) [51] bounds their contribution to the maximum load as $\log\log n + O(1)$.

We now modify our low-associativity construction to use ICEBERG[2] (with $k = 3$ hash functions) instead of just a single hash function. Set the number of bins to be $m/(\log\log P \log\log\log P)$, so that the average load is $\lambda = \log\log P \log\log\log P = \omega(\log\log n)$. Then, with high probability in n (and thus P), the maximum load is:

$$(1 + o(1))\lambda + \log\log n + O(1) = \lambda(1 + \delta),$$

with $\delta = o(1)$.[5]

[3] Interestingly, it is unknown whether the asymptotic dependence on λ is an artifact of the proof. If one could prove a maximum load of $\lambda + O(\log\log n)$ for GREEDY[2], then one could use GREEDY[2] to achieve the results in this section.

[4] As a minor technical point, the insertions performed using h_1 ignore all balls that were inserted using h_2 and h_3, and, similarly, the GREEDY[2] insertion of balls using h_2 and h_3 ignores all balls that were inserted using h_1.

[5] For our purposes here, we do not make an effort to optimize δ beyond ensuring that $\delta = o(1)$. We point out, however, that if one wanted to optimize δ further, one could set the associativity to $\text{poly}(\log\log P)$ (which only changes $h_{\max}$ by a constant factor), and obtain $\delta = 1/\text{poly}(\log\log P)$, for a polynomial of our choice.

Using this value of δ as the resource-augmentation parameter, it follows that the bin size is (with high probability) at least as large as the maximum load. Since the bin size is $B = \tilde{\Theta}(\log \log P)$, the associativity of the scheme is $3B = \tilde{\Theta}(\log \log P)$. Thus we have a decoupling scheme with huge-page size $h_{\max} = \Theta(w/\log \log \log P)$ and resource augmentation parameter $\delta = o(1)$.

Theorem 3 (The Decoupling Theorem). *There exists a constant-time huge-page decoupling scheme using resource augmentation $\delta = o(1)$ that supports huge-page size $h_{max} = \Theta(w/\log \log \log P)$ with high probability in P.*

Proof. The proof follows just as for Theorem 1, but using Iceberg[2] instead of a single hash function. □

5 Optimizing the Cost of AT

Finally, we consider the task of optimizing the total TLB and IO cost of a memory-management algorithm on a sequence of page requests. More specifically, in this section, we prove that to optimize the cost of a memory-management algorithm, it's enough to *independently* optimize the TLB cost and the paging cost, and combine the two solutions via huge-page decoupling. Moreover, these two separate problems are each equivalent to the classic paging problem [49].

We begin by formalizing the definitions of arbitrary memory-management algorithms and of the AT cost model. These definitions must carefully address several subtleties of the model. First, what is the full range of control that an *arbitrary* memory-management algorithm has? This needs to be carefully specified so that we can prove competitiveness results. Second, how do we define the cost of a memory-management algorithm that sometimes brings pages into RAM even when those pages are not being accessed (e.g., a memory-management algorithm that implements virtual huge pages as physical huge pages, and thus brings entire physical huge pages into RAM at once)? And finally, what types of failures are permitted for a memory-management algorithm? In particular, paging failures (as defined in Section 3) are not acceptable, but we shall see that these types of failures can be handled at a cost of additional IOs.

What a Memory-Management Algorithm Controls. We begin by extending the definitions from Section 3 in order to define what an arbitrary memory-management algorithm controls. A memory-management algorithm controls:

- —which virtual huge-page addresses $\mathcal{T}$ are in the TLB;
- —which virtual page addresses are in the active set $\mathcal{A}$;
- —what the TLB-decoding function f is; and
- —what the virtual-to-physical mapping ϕ is.

In other words, a memory-management algorithm controls not only the features of the system that a huge-page decoupling scheme controls, but also the TLB-replacement policy and the RAM-replacement policy.

Whereas a huge-page decoupling scheme treats $\mathcal{T}$ as consisting of virtual huge pages of size $h_{\max}$, in general, a memory-management algorithm is permitted to use virtual huge pages of any mixture of sizes in $\{1, 2, 4, 8, \ldots, h_{\max}\}$ (we assume $h_{\max}$ is a power of two).[6] Recall from Section 3

[6]The fact that we allow memory-management algorithms to potentially use many different huge-page sizes at the same time will only make our results stronger. In particular, this will allow the memory-management algorithms $\mathcal{X}$ and $\mathcal{Y}$ that are used as inputs to Theorem 4 to be more sophisticated; on the other hand, the output memory-management algorithm $\mathcal{Z}$ produced by Theorem 4 uses only a single size $h_{\max}$ for huge pages.

that, if a huge page is of size 2^r, then it is associated with an address that is an integer multiple of 2^r.

Servicing Page Requests. The purpose of a memory-management algorithm is to service a sequence of virtual-page requests $\sigma = (p_1, p_2, \ldots, p_n)$, where each $p_i \in [V]$.

For the memory-management algorithm to be able to service a page request p_i, it must ensure that virtual page p_i is in RAM (i.e., if $p_i \notin \mathcal{A}$, then p_i must be added to $\mathcal{A}$), and the algorithm must also ensure that a virtual huge page containing p_i is contained in the TLB. Once the page p_i is mapped in both RAM and the TLB, the request p_i can be serviced.[7]

The AT Cost Model. The running time of a memory-management algorithm is evaluated in the *AT cost model*: The cost of adding a new entry to $\mathcal{T}$ is ε and the cost of adding a new element to the active set $\mathcal{A}$ is 1. Evictions (from either the TLB or RAM) are free, and so is updating the TLB value $\psi(u)$ for a virtual huge page address $u \in \mathcal{T}$ when one of u's constituent pages is added or removed from $\mathcal{A}$.

In order to allow for a full range of TLB decoding/encoding schemes, it is necessary to also capture the notion of a *decoding miss*, which costs ε. A decoding miss occurs if a virtual huge page u is in the TLB, and a virtual page v contained in u is in RAM, but the decoding function $f(v, \psi(u))$ incorrectly evaluates to -1 (instead of $\phi(v)$). That is, the huge page u that contains page v is cached within the TLB, but when performing an AT for page v we still experience a TLB miss.

Note that our encoding/decoding schemes used in Section 4 do not allow for the possibility of a decoding miss. In other schemes a decoding miss can occur if, for example, a memory-management algorithm chooses to encode for each virtual huge page u in the TLB only the physical addresses of u's most commonly accessed constituent pages; then the pages that do not get encoded would incur decoding misses when they were accessed. We will use decoding misses in Theorem 4 to capture what happens if a huge-page decoupling scheme experiences a paging failure; we will construct a memory-management algorithm that brings the page experiencing failure into RAM without giving it a TLB encoding.

Recall that for a given memory-management algorithm $\mathcal{X}$, $C(\mathcal{X}, \sigma)$ denotes the total cost of $\mathcal{X}$ (on the request sequence σ), $C_{\text{TLB}}(\mathcal{X}, \sigma)$ denotes the total TLB cost (this does not include decoding misses), and $C_{\text{IO}}(\mathcal{X}, \sigma)$ denotes the total IO cost. Additionally, we define $C_{\text{D}}(\mathcal{X}, \sigma)$ as the total cost incurred due to decoding misses. Then, $C(\mathcal{X}, \sigma) = C_{\text{TLB}}(\mathcal{X}, \sigma) + C_{\text{IO}}(\mathcal{X}, \sigma) + C_{\text{D}}(\mathcal{X}, \sigma)$.

Huge-Page Decoupling as a Technique for Simultaneously Optimizing IO Costs and TLB Costs. Having defined the AT cost model and how it applies to memory-management algorithms, we can now prove Theorem 4.

Theorem 4 (The Simulation Theorem). *Let V and P be the number of pages in the virtual and physical address spaces, respectively. Let ℓ be the number of entries in the TLB, and w be the number of bits in each TLB value.*

Let $\mathcal{D}$ be a huge-page decoupling scheme that uses resource-augmentation $\delta = o(1)$ and supports huge-page size h_{max} with high probability in P.

Let $\sigma = (p_1, \ldots, p_n) \in [V]^n$ be a sequence of virtual page addresses that need to be serviced. Let $\mathcal{X}$ be an arbitrary memory-management algorithm using parameters ℓ, w, V, P, and using huge pages with sizes between 1 and h_{max} pages; let $\mathcal{Y}$ be an arbitrary memory-management algorithm using parameters $\ell, w, V, (1-\delta)P$, and using huge pages with sizes between 1 and h_{max} pages. Using $\mathcal{D}$,

[7]Whereas RAM truly requires that pages be present in order to be accessed, the same requirement isn't strictly necessary for the TLB (since we can always just find the physical address via the page table). On the other hand, in the address-translation cost model, adding an element TLB has the same cost as incurring a TLB miss would. Thus we can assume without loss of generality that every page that is accessed is first added to the TLB's coverage if necessary.

one can construct a new memory-management algorithm $\mathcal{Z}$, using virtual huge pages of size h_{max}, satisfying:

$$C(\mathcal{Z}, \sigma) \leq C_{\text{TLB}}(\mathcal{X}, \sigma) + C_{\text{IO}}(\mathcal{Y}, \sigma) + \frac{n}{\text{poly}(P)}, \tag{7}$$

with high probability in P. Moreover, if $\mathcal{X}$ and $\mathcal{Y}$ are online algorithms, then so is $\mathcal{Z}$.

Proof. To design $\mathcal{Z}$, we combine all three of $\mathcal{X}$, $\mathcal{Y}$, and $\mathcal{D}$. The idea is to use $\mathcal{X}$'s TLB-replacement policy, to use $\mathcal{Y}$'s as the RAM-replacement policy, and then to use $\mathcal{D}$ to combine those two policies into a new memory-management algorithm $\mathcal{Z}$.

We begin by describing how to use $\mathcal{X}$ to determine the TLB-replacement policy for $\mathcal{D}$. Whereas the TLB for $\mathcal{X}$ may use huge pages of many different sizes (up to size $h_{\max}$), the TLB for $\mathcal{Z}$ uses virtual huge pages exclusively of size $h_{\max}$. Let $\mathcal{T}_{\mathcal{Z}}$ denote the set of virtual huge-page addresses in $\mathcal{Z}$'s TLB at any given moment and let $\mathcal{T}_{\mathcal{X}}$ denote the set of virtual huge-page addresses in $\mathcal{X}$'s TLB at any given moment. For each virtual address $v \in V$, let $r(v) = v - (v \pmod{h_{\max}})$ denote the virtual address of the size-$h_{\max}$ virtual huge page containing v. We say that an address $v \in V$ is *covered* by a size-$h_{\max}$ virtual huge-page address u if $r(v) = u$. Note that, if a virtual huge-page address $v \in \mathcal{T}_{\mathcal{X}}$ is covered by a size-$h_{\max}$ virtual huge-page address u, then so are all of the constituent virtual pages v' of v.

We define the TLB-replacement policy for $\mathcal{D}$ so that:

$$\mathcal{T}_{\mathcal{Z}} = \{r(v) \mid v \in \mathcal{T}_{\mathcal{X}}\}.$$

Note that, since $|\mathcal{T}_{\mathcal{X}}| \leq \ell$, we also always have $|\mathcal{T}_{\mathcal{Z}}| \leq \ell$. And, moreover, the TLB-replacement policy only modifies $\mathcal{T}_{\mathcal{Z}}$ if it is also modifying $\mathcal{T}_{\mathcal{X}}$.

Next we describe how to use $\mathcal{Y}$ to determine the RAM-replacement policy for $\mathcal{D}$. We simply have the RAM-replacement policy for $\mathcal{D}$ maintain the active set $\mathcal{A}$ to always match the active set for $\mathcal{Y}$. Note that, since $\mathcal{Y}$ operates on a physical memory of size $(1-\delta)P$, the active set $\mathcal{A}$ is never of size more than $(1-\delta)P$, which in turn meets the resource-augmentation requirement for the huge-page decoupling scheme $\mathcal{D}$.

Although the RAM-replacement policy for $\mathcal{D}$ maintains the active set to match the active set for $\mathcal{Y}$, the huge-page decoupling scheme $\mathcal{D}$ may sometimes experience a paging failure, causing a virtual page v that is in the active set for $\mathcal{Y}$ to not be present in the active set for $\mathcal{Z}$. Whenever a page request p_i is to a page v for which $\mathcal{D}$ is currently experiencing a paging failure, we have the memory-management algorithm $\mathcal{Z}$ handle the page request p_i as follows: (1) the algorithm $\mathcal{Z}$ spends an IO (of cost 1) to temporarily add v to $\mathcal{Z}$'s active set; (2) the algorithm $\mathcal{Z}$ sets $\phi(v)$ to be an arbitrary free physical page address; and (3) the algorithm $\mathcal{Z}$ services the page request to v, incurring an additional cost of ε due to the ensuing decoding miss (note, in particular, that $\mathcal{Z}$ does not make any effort to encode the translation from v to $\phi(v)$ in the TLB). Thus the total cost of servicing a page request p_i to a page v that is experiencing a paging failure is $1 + \varepsilon$. Once the request p_i is serviced, then v may be removed from $\mathcal{A}$ whenever convenient (i.e., whenever $\mathcal{D}$ wishes to assign some other virtual page address $v' \neq v$ to the physical address $\phi(v)$ that v is currently assigned to).

We have now completely defined the memory-management algorithm $\mathcal{Z}$. In summary, $\mathcal{Z}$ is constructed via the huge-page decoupling scheme $\mathcal{D}$ using $\mathcal{X}$ as the TLB-replacement policy and $\mathcal{Y}$ as the RAM-replacement policy; and the only time that $\mathcal{Z}$ departs from the behavior of $\mathcal{D}$ is when $\mathcal{D}$ is experiencing a paging failure on a page request p_i. In this case, $\mathcal{Z}$ serves the page request at a total cost of $1 + \varepsilon$. Note that, if $\mathcal{X}$ and $\mathcal{Y}$ are online algorithms, meaning that at any given moment they only know the value of the next page p_i that will be requested, then $\mathcal{Z}$ is also online.

We conclude by analyzing the cost $C(\mathcal{Z}, \sigma)$. First note that, since $\mathcal{D}$ is a huge-page decoupling scheme that succeeds with high probability, the probability that there is a paging failure during a given page request p_i is at most $1/\text{poly}(P)$ (for a polynomial of our choice). By linearity of expectation, the expected number of page requests p_i at which paging failure is being experienced is at most $n/\text{poly}(P)$. Applying Markov's inequality, it follows that with high probability in P, at most $n/\text{poly}(P)$ page requests p_i occur during paging failures. (Note that the application of Markov's inequality shrinks the $\text{poly}(P)$ term in the denominator, but it nonetheless remains a polynomial of our choice.) The total cost incurred by $\mathcal{Z}$ due to paging failures of $\mathcal{D}$ is therefore at most $(1+\varepsilon)n/\text{poly}(P) \leq n/\text{poly}(P)$ with high probability in P.

To complete the proof, we perform the rest of the analysis ignoring costs incurred by $\mathcal{Z}$ due to paging failures. By using $\mathcal{X}$ as the TLB-replacement policy, we ensure that $\mathcal{Z}$ only ever adds elements to $\mathcal{T}_{\mathcal{Z}}$ when $\mathcal{X}$ also adds an element to $\mathcal{T}_{\mathcal{X}}$. Thus (ignoring requests that experience paging failures), the TLB cost of $\mathcal{Z}$ is at most the TLB cost of $\mathcal{X}$. At the same time, by using $\mathcal{Y}$ as the RAM-replacement policy, we ensure that $\mathcal{X}$ only ever adds elements to its active set when $\mathcal{Y}$ adds the same element to its active set (again, ignoring paging failures). Thus the total IO cost of $\mathcal{Z}$ (ignoring paging failures) is at most the total IO cost for $\mathcal{Y}$. Since $\mathcal{Z}$ does not experience any decoding misses except during paging failures, the cost incurred by $\mathcal{Z}$ due to decoding misses is absorbed by the paging failure cost. This completes the proof. □

Theorem 4 reduces the optimization problem of minimizing $C(\mathcal{Z}, \sigma)$ to the independent (and separate) optimization problems of minimizing $C_{\text{TLB}}(\mathcal{X}, \sigma)$ and $C_{\text{IO}}(\mathcal{Y}, \sigma)$. We conclude the section by observing that these two individual optimization problems are equivalent to the classic paging problem, which counts the number cache misses incurred by a paging algorithm to service a sequence of page requests $p_1, p_2, \ldots, p_n$ on a cache of some size. The paging problem does not have a unique optimal online solution (and thus many algorithms for the problem have been studied [14, 15, 22, 23, 54, 55]). Nonetheless the theoretical and practical properties of the paging problem are well understood.

Lemma 1. *Let $\mathcal{X}$ and $\mathcal{Y}$ be the memory-management algorithms from* Theorem 4*, and let $\sigma = (p_1, p_2, \ldots, p_n) \in [V]^n$. For each virtual page $v \in V$, let $r(v)$ denote the virtual huge page of size h_{max} containing v.*

The problem of minimizing $C_{\text{TLB}}(\mathcal{X}, \sigma)$ in Theorem 4 *is equivalent to the paging problem on the request sequence $r(p_1), r(p_2), \ldots, r(p_n)$ using a cache of size ℓ.*

The problem of minimizing $C_{\text{IO}}(\mathcal{Y}, \sigma)$ in Theorem 4 *is equivalent to the paging problem on the request sequence $p_1, p_2, \ldots, p_n$ using a cache of size $(1-\delta)P$.*

Proof. If we wish to design $\mathcal{X}$ to minimize $C_{\text{TLB}}(\mathcal{X}, \sigma)$, then we can assume without loss of generality that the TLB for $\mathcal{X}$ uses only huge pages of size $h_{\max}$. In particular, any virtual huge page v of size smaller than $h_{\max}$ in the TLB can be substituted with a larger huge page $r(v)$ of size $h_{\max}$ without any increase in TLB cost. If we assume that $\mathcal{X}$ uses huge pages of size $h_{\max}$, then the page request sequence accesses virtual huge pages $r(p_1), r(p_2), \ldots$. We can further assume without loss of generality that $\mathcal{X}$ only adds a virtual huge page v to the TLB when that virtual huge page is about to be accessed in the request sequence (since otherwise, $\mathcal{X}$ could hold off on adding v until v is next accessed). Thus, the problem of minimizing $C_{\text{TLB}}(\mathcal{X}, \sigma)$ is exactly the problem of servicing the virtual huge-page requests $r(p_1), \ldots, r(p_n)$ using the TLB as a size-ℓ cache. Adding a virtual huge page to the TLB in the former problem corresponds exactly to incurring a cache miss in the latter problem.

To see the claim about $C_{\text{IO}}(\mathcal{Y}, \sigma)$, observe that the active set $\mathcal{A}$ for $\mathcal{Y}$ corresponds directly to the cache in the paging problem. Note that, without loss of generality, $\mathcal{Y}$ only adds a page to $\mathcal{A}$

when that page is about to be accessed. Thus, the IOs incurred by $\mathcal{Y}$ correspond exactly to the cache misses incurred in the paging problem. □

6 Related Work

Limited-Associativity Paging. Sleator and Tarjan [49] gave competitive analyses of LRU and FIFO paging algorithms, both with and without resource augmentation. These results were subsequently generalized by many authors [14, 15, 22, 23, 54, 55].

Due to the importance of limited-associativity caches in hardware, there has also been substantial theoretical work on paging algorithms in the low-associativity setting. One direction of work has been to analyze the competitive ratio of low-associativity paging algorithms, where OPT is *also* limited in its associativity [6, 16, 17, 26, 41]. Another direction of work has been to design cache-aware algorithms that interact well with caches of low associativity. Notably, Frigo et al. [27, 28] and Prokop [45] showed how to take any algorithm in the external-memory model [5] and change the algorithm's access patterns in order so that a direct-mapped cache (i.e., a cache with associativity 1) can be used to simulate a fully associative cache up to a constant factor in performance. In a similar direction, Sen and Chatterjee [48] present cache-aware algorithms for several basic problems (e.g., sorting, FFT, and permutations) in a variant of the external-memory model [5, 27, 28] in which cache has limited associativity.

In contrast with past work, our results show how to convert *any* fully associative paging scheme into one with limited associativity at *almost no* overhead. Thus, rather than designing a paging algorithm (or designing an algorithm whose memory accesses play well with a paging algorithm), we are interested deciding *where* pages should reside in memory. And rather than aiming for a constant competitive ratio, our application of AT requires us to be $(1 + o(1))$-competitive with the paging algorithm that we are simulating. On the other hand, whereas past work often treats the associativity as constant (and possibly even 1), our schemes are allowed to use super-constant associativity (although, remarkably, we show that even $\tilde{O}(\log \log P)$-associativity suffices).

Huge Pages. Increasing the granularity of AT is a standard method to amplify TLB coverage. For instance, Linux provides software support for huge pages of size larger than the typical 4 kB [29]. Manufacturers typically manage heterogeneity of page sizes using dedicated TLBs for different sizes. For instance, Intel's Cascade Lake microarchitecture allows 2 MB and 1 GB pages, and provides a 1,536-entry L2 data TLB for 4 kB and 2 MB pages, and a 16-entry L2 data TLB for 1 GB pages [18]. The actual coverage gains are limited by the dedicated TLB size, and are thus much less than the multiplicative blowup in page size.

For some workloads, huge pages are wasteful, creating unnecessary memory pressure that is in turn worsened by the increased swapping cost. Two attempts to overcome this lack of flexibility are Linux's **transparent huge pages (THP)** and *superpages* [35], that work by coalescing areas of virtually and physically contiguous pages into larger blocks (a huge/super page). In these schemes, the OS must either enforce contiguity of physical huge pages or have fallback mechanisms when it cannot allocate a physical huge page. THP attempts to reserve enough space for a huge page and, in case of failure, falls back to allocating typical 4 kB pages that are reallocated later on. Page reallocation incurs large performance penalties, since all applications whose pages are being moved are paused, and in fact a number of commercial databases and other products recommend huge pages should be disabled for optimal performance [1–4]. The superpage system avoids reallocation by always over-allocating memory, and keeps track of unused pages within a superpage so they can be reclaimed by other superpages. The downside is an increased complexity and overhead of OS memory management. Both approaches suffer from increased swapping costs, because once a huge or superpage is created, it is treated as an indivisible mapping unit.

For some workloads huge pages increase page fault latency because Linux has to clear up much larger pages and consolidates fragmented pages to create large continuous physical pages synchronously. Ingens [34] points out that workloads that fragment memory quickly, such as in multi-tenant cloud environments, suffer significant performance penalty, and implements an adaptive policy to promote huge pages and asynchronously defragment memory. HawkEye [37] proposes to synchronously pre-zero freed pages to further reduce latency. GLUE [44] observes that huge pages hurt lightweight system memory management and reduce consolidation in the over-committed cloud deployments which depends on page sharing to share memory. TEMPO [13], a prefetching optimization technique to reduce TLB misses, reports in experiments how much huge pages de-optimize: The more frequent huge pages are used the less effective the optimization becomes.

TLB Encodings. Because of the difficulties in maintaining physical contiguity, a number of research projects have explored practical TLB optimizations that leverage some contiguity when it is present, such as coalescing TLB entries for runs of contiguous translations that are smaller than a huge page [20, 40, 42, 43], or composing a huge page out of "medium" sized frames [24]. Direct Segments [8] allow a programmer to map gigabyte- to terabyte-sized primary segments of memory and making the hardware to represent these segments using a single TLB translation entry. Proposals such as COLT [43] and Translation Ranger [52] identify physically contiguous pages mapped in a process address space, and compress the TLB translation into a single entry. As these examples show, huge paging schemes that require physical contiguity saddle the OS developer with solving the difficult, open problem of efficiently maintaining physical contiguity.

To the best of our knowledge, our work is the first to completely remove the requirement that huge pages be stored physically contiguously.

7 Conclusion

The AT problem lies at a surprisingly fertile nexus of hashing, succinct data structures, and external-memory algorithms. Several lines of theoretical work have been directly inspired by the problem and we believe there are opportunities for future work in this area (both theoretical and hardware). Theoretical work could include more accurately modeling the TLB miss cost. This translation is performed by the page table, a data structure that benefits from temporal locality of access. Thus, it has a more intricate true cost than ε. It could be beneficial to codesign the TLB and page table so that TLB misses are likely to have a low page table translation cost.

Our results suggest how to design new hardware for the encoding and decoding of TLB entries that improves AT. Subsequent work has confirmed that these changes are feasible and have minimal (or even no) impact on energy consumption and latency, while significantly increasing the TLB hit rate [31].

Our results may also change the landscape of some hardware design tradeoffs. For example, it may make sense to modify hardware more aggressively in the future to improve AT. Specifically, this article treats w as a fixed parameter. On the other hand, when designing the hardware of TLBs, there is an opportunity to change the value of w if the payoff is big enough. An interesting feature of our results is that they change the asymptotic relationship between w and the coverage of the TLB: Even small increases in w correspond to potentially large gains in TLB coverage (and, moreover, these gains do not require the storage of additional keys!). Thus larger values of w may make sense using our techniques than was previously the case.

On the other hand, as long as w remains reasonably small, then a hybrid approach may be sensible: One can use both huge-page decoupling and physical huge pages of moderate size. So, for example, if an optimal virtual huge page size is $q \gg h_{\max}$ pages, then we could implement

decoupled huge pages where the *physical* huge pages would have size only $q/h_{\max}$, thus achieving all the coverage of the very large huge pages while mitigating the adverse effects on I/Os.

References

[1] Couchbase. 2021. Disabling Transparent Huge Pages (THP). Retrieved February 11, 2021 from https://docs.couchbase.com/server/current/install/thp-disable.html

[2] MongoDB. 2021. Disable Transparent Huge Pages (THP). Retrieved February 11, 2021 from https://docs.mongodb.com/manual/tutorial/transparent-huge-pages/

[3] Oracle Database. 2021. Disabling Transparent Hugepages. Retrieved February 11, 2021 from https://docs.oracle.com/en/database/oracle/oracle-database/12.2/ladbi/disabling-transparent-hugepages.html

[4] Percona. 2021. Settling the Myth of Transparent Hugepages for Databases. Retrieved February 11, 2021 from https://www.percona.com/blog/2019/03/06/settling-the-myth-of-transparent-hugepages-for-databases/

[5] Alok Aggarwal and Jeffrey S. Vitter. 1988. The input/output complexity of sorting and related problems. *Communications of the ACM* 31, 9 (September 1988), 1116–1127.

[6] Kunal Agrawal, Michael A. Bender, and Jeremy T. Fineman. 2007. The worst page-replacement policy. In *Proceedings of the 4th International Conference on Fun with Algorithms (FUN)*. Springer-Verlag, 135–145.

[7] AMD, Inc. 2008. AMD-V nested paging.

[8] Arkaprava Basu, Jayneel Gandhi, Jichuan Chang, Mark D. Hill, and Michael M. Swift. 2013. Efficient virtual memory for big memory servers. In *Proceedings of the 40th Annual International Symposium on Computer Architecture (ISCA)*. ACM.

[9] Michael A. Bender, Alex Conway, Martín Farach-Colton, William Kuszmaul, and Guido Tagliavini. 2023. Iceberg hashing: Optimizing many hash-table criteria at once. *Journal of the ACM* 70, 6 (November 2023), 1–51.

[10] Michael A. Bender, Alex Conway, Martín Farach-Coltonf, William Kuszmaul, and Guido Tagliavini. 2023. Tiny pointers. In *Proceedings of the 2023 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*. SIAM, 477–508.

[11] Michael A. Bender, Rathish Das, Martín Farach-Colton, and Guido Tagliavini. 2023. An associativity threshold phenomenon in set-associative caches. In *Proceedings of the 35th ACM Symposium on Parallelism in Algorithms and Architectures (SPAA '23)*. ACM, New York, NY, 117–127.

[12] Abhishek Bhattacharjee. 2017. Preserving virtual memory by mitigating the address translation wall. *IEEE Micro* 37, 5 (2017), 6–10.

[13] Abhishek Bhattacharjee. 2017. Translation-triggered prefetching. In *Proceedings of the 22nd International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS)*. ACM, 63–76.

[14] Allan Borodin and Ran El-Yaniv. 1998. *Online Computation and Competitive Analysis*. Cambridge University Press, New York.

[15] Joan Boyar, Lene M. Favrholdt, and Kim S. Larsen. 2007. The relative worst-order ratio applied to paging. *Journal of Computer and System Sciences* 73, 5 (August 2007), 818–843.

[16] Mark Brehob, Richard Enbody, Eric Torng, and Stephen Wagner. 2001. On-line restricted caching. In *Proceedings of the 12th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*. Society for Industrial and Applied Mathematics, 374–383.

[17] Niv Buchbinder, Shahar Chen, and Joseph (Seffi) Naor. 2014. Competitive algorithms for restricted caching and matroid caching. In *Proceedings of the 22nd European Symposium on Algorithms (ESA)*. Springer, Berlin, 209–221.

[18] WikiChip. 2020. Intel's Cascade Lake Microarchitecture. Retrieved February 2, 2020 from https://en.wikichip.org/wiki/intel/microarchitectures/cascade_lake

[19] Fernando J. Corbató. 1969. *A Paging Experiment with the Multics System*. MIT Project MAC Report MAC-M-384.

[20] Guilherme Cox and Abhishek Bhattacharjee. 2017. Efficient address translation for architectures with multiple page sizes. In *Proceedings of the 22nd International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS)*. ACM, 435–448.

[21] Peter J. Denning. 1968. The working set model for program behavior. *Communications of the ACM* 11, 5 (May 1968), 323–333.

[22] Reza Dorrigiv and Alejandro López-Ortiz. 2008. Closing the gap between theory and practice: New measures for on-line algorithm analysis. In *WALCOM: Algorithms and Computation*. Shin-ichi Nakano and Md. Saidur Rahman (Eds.), Springer, Berlin, 13–24.

[23] Reza Dorrigiv, Alejandro López-Ortiz, and J. Ian Munro. 2009. On the relative dominance of paging algorithms. *Theoretical Computer Science* 410, 38–40 (September 2009), 3694–3701.

[24] Y. Du, M. Zhou, B. R. Childers, D. Mossé, and R. Melhem. 2015. Supporting superpages in non-contiguous physical memory. In *Proceedings of the 2015 IEEE 21st International Symposium on High Performance Computer Architecture (HPCA)*, 223–234.

[25] Amos Fiat, Richard M. Karp, Michael Luby, Lyle A. McGeoch, Daniel D. Sleator, and Neal E. Young. 1991. Competitive paging algorithms. *Journal of Algorithms* 12, 4 (1991), 685–699.
[26] Amos Fiat, Manor Mendel, and Steven Seiden. 2002. Online companion caching. *Theoretical Computer Science* 324 (September 2002), 499–511.
[27] Matteo Frigo, Charles E. Leiserson, Harald Prokop, and Sridhar Ramachandran. 1999. Cache-oblivious algorithms. In *Proceedings of the 40th Annual Symposium on Foundations of Computer Science (FOCS)*, 285.
[28] Matteo Frigo, Charles E. Leiserson, Harald Prokop, and Sridhar Ramachandran. 2012. Cache-oblivious algorithms. *ACM Transactions on Algorithms* 8, 1 (2012), 1–22.
[29] Mel Gorman. 2010. Linux Huge Pages. Retrieved from https://lwn.net/Articles/375096/
[30] Mel Gorman. 2018. AMD Zen Architecture. Retrieved from https://en.wikichip.org/wiki/amd/microarchitectures/zen
[31] Krishnan Gosakan, Jaehyun Han, William Kuszmaul, Ibrahim N. Mubarek, Nirjhar Mukherjee, Karthik Sriram, Guido Tagliavini, Evan West, Michael A. Bender, Abhishek Bhattacharjee, et al. 2023. Mosaic pages: Big TLB reach with small pages. In *Proceedings of the 28th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 3 (ASPLOS '23)*. ACM, New York, NY, 433–448.
[32] Intel, Inc. 2016. Intel® 64 and IA-32 architectures software developer's manual volume 3A: System programming guide, Part 1.
[33] V. Karakostas, J. Gandhi, A. Cristal, M. D. Hill, K. S. McKinley, M. Nemirovsky, M. M. Swift, and O. S. Unsal. 2016. Energy-efficient address translation. In *Proceedings of the 2016 IEEE International Symposium on High Performance Computer Architecture (HPCA)*, 631–643.
[34] Youngjin Kwon, Hangchen Yu, Simon Peter, Christopher J. Rossbach, and Emmett Witchel. 2016. Coordinated and efficient huge page management with Ingens. In *Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI)*. USENIX Association, 705–721.
[35] Juan Navarro, Sitaram Iyer, Peter Druschel, and Alan L. Cox. 2002. Practical, transparent operating system support for superpages. In *Proceedings of the 5th Symposium on Operating System Design and Implementation (OSDI)*.
[36] Stanko Novakovic, Yizhou Shan, Aasheesh Kolli, Michael Cui, Yiying Zhang, Haggai Eran, Liran Liss, Michael Wei, Dan Tsafrir, and Marcos K. Aguilera. 2019. Storm: A fast transactional dataplane for remote data structures. arXiv:1902.02411. Retrieved from https://arxiv.org/abs/1902.02411
[37] Ashish Panwar, Sorav Bansal, and K. Gopinath. 2019. HawkEye: Efficient fine-grained OS support for huge pages. In *Proceedings of the 24th International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS)*. ACM, 347–360.
[38] Ashish Panwar, Aravinda Prasad, and K. Gopinath. 2018. Making huge pages actually useful. *ACM SIGPLAN Notices* 53, 2 (March 2018), 679–692.
[39] Chang Hyun Park, Taekyung Heo, Jungi Jeong, and Jaehyuk Huh. 2017. Hybrid TLB coalescing: Improving TLB translation coverage under diverse fragmented memory allocations. *ACM SIGARCH Computer Architecture News* 45, 2 (June 2017), 444–456.
[40] Chang Hyun Park, Taekyung Heo, Jungi Jeong, and Jaehyuk Huh. 2017. Hybrid TLB coalescing: Improving TLB translation coverage under diverse fragmented memory allocations. In *Proceedings of the 44th Annual International Symposium on Computer Architecture (ISCA)*. ACM, 444–456.
[41] Enoch Peserico. 2003. Online paging with arbitrary associativity. In *Proceedings of the 14th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*. Society for Industrial and Applied Mathematics, 555–564.
[42] B. Pham, A. Bhattacharjee, Y. Eckert, and G. H. Loh. 2014. Increasing TLB reach by exploiting clustering in page translations. In *Proceedings of the 2014 IEEE 20th International Symposium on High Performance Computer Architecture (HPCA)*, 558–567.
[43] B. Pham, V. Vaidyanathan, A. Jaleel, and A. Bhattacharjee. 2012. CoLT: Coalesced large-reach TLBs. In *Proceedings of the 2012 45th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO)*, 258–269.
[44] B. Pham, J. Veselý, G. H. Loh, and A. Bhattacharjee. 2015. Large pages and lightweight memory management in virtualized environments: Can you have it both ways? In *Proceedings of the 2015 48th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO)*, 1–12.
[45] H. Prokop. 1999. *Cache Oblivious Algorithms*. Master's thesis. Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology.
[46] Martin Raab and Angelika Steger. 1998. "Balls into bins"—A simple and tight analysis. In *Randomization and Approximation Techniques in Computer Science*. M. Luby, J. D. P. Rolim, and M. Serna (Eds.), Springer, Berlin, 159–170.
[47] 7-Zip LZMA Benchmark. 2021. SandyBridge. Retrieved from https://www.7-cpu.com/cpu/SandyBridge.html
[48] Sandeep Sen and Siddhartha Chatterjee. 2000. Towards a theory of cache-efficient algorithms. In *Proceedings of the 11th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*. Society for Industrial and Applied Mathematics, Philadelphia, 829–838.

[49] Daniel D. Sleator and Robert E. Tarjan. 1985. Amortized efficiency of list update and paging rules. *Communications of the ACM* 28, 2 (February 1985), 202–208.
[50] Michael M. Swift. 2017. Towards $O(1)$ memory. In *Proceedings of the 16th Workshop on Hot Topics in Operating Systems (HotOS)*, 7–11.
[51] Berthold Vöcking. 2003. How asymmetry helps load balancing. *Journal of the ACM* 50, 4 (July 2003), 568–589.
[52] Zi Yan, Daniel Lustig, David Nellans, and Abhishek Bhattacharjee. 2019. Translation ranger: Operating system support for contiguity-aware TLBs. In *Proceedings of the 46th International Symposium on Computer Architecture (ISCA)*. IEEE, New York, 698–710.
[53] Jian Yang, Joseph Izraelevitz, and Steven Swanson. 2020. FileMR: Rethinking RDMA networking for scalable persistent memory. In *Proceedings of the 17th USENIX Symposium on Networked Systems Design and Implementation (NSDI '20)*. USENIX Association, 111–125.
[54] N. Young. 1994. The k-server dual and loose competitiveness for paging. *Algorithmica* 11, 6 (June 1994), 525–541.
[55] Neal E. Young. 1998. On-line file caching. In *Proceedings of the 9th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*. Society for Industrial and Applied Mathematics, Philadelphia, 82–86.
[56] WikiChip. 2020. AMD's Zen Microarchitecture. Retrieved July15, 2020 from https://en.wikichip.org/wiki/amd/microarchitectures/zen

Received 20 December 2021; revised 18 May 2024; accepted 10 May 2025