Online Companion Caching

Manor Mendel
Department of Computer Science
The Hebrew University, Jerusalem

Steven S. Seiden*
Department of Computer Science
Louisiana State University, Baton Rouge

Abstract

This paper is concerned with online caching algorithms for the \((n,k)\)-companion cache, defined by Brehob et. al. In this model the cache is composed of two components: a \(k\)-way set-associative cache and a companion fully-associative cache of size \(n\). We show that the deterministic competitive ratio for this problem is \((n+1)(k+1) - 1\), and the randomized competitive ratio is \(O(\log n \log k)\) and \(\Omega(\log n + \log k)\).

Steve Seiden died in a tragic accident on June 11, 2002. The first named author would like to dedicate this paper to his memory.

1 Introduction

There is a rapidly growing disparity between computer processor speed and computer memory speed. Of prime importance in bridging this gap is the cache, the purpose of which is to allow quick access to memory items that are accessed frequently. Since the cache is so important to system performance, hardware designers have in recent years proposed a sequence of increasingly sophisticated cache designs (see e.g., [8, 12, 5]). Cache designs can be conceptually thought as having two parts: An architecture and a caching algorithm. The architecture describes the physical structure of the cache such as its size and organization. The caching algorithm decides, for a given sequence of requests for memory items, which items are stored in the cache, and how they are organized, at each point in time. While there is a large body of theoretical work on caching algorithms for the simplest types of caches (which we refer to as fully associative), little theoretical work has been done on algorithms for more complicated cache architectures. In this paper, we address this deficiency by providing the first theoretical analysis of the \((n,k)\)-companion cache problem for \(k > 1\).

Problem Description: A popular cache architecture is the set-associative cache. In a \(k\)-way set-associative cache, a cache of size \(s\) is divided into \(m = s/k\) disjoint sets, each of size \(k\). Addresses in main memory are likewise assigned one of \(m\) types, and the \(i\)th associative cache can only store memory cells whose address is type \(i\). Typically, there are \(m = 2^i\) such types, where the \(j\)th \(k\)-wise associative cache is indexed by \(0 \leq j \leq 2^i - 1\) and memory addresses whose last \(i\) bits are equal to \(j\) are mapped to the \(j\)th associative cache. Special cases includes direct-mapped caches, which are

*This research was partially supported by the Louisiana Board of Regents Research Competitiveness Subprogram and by AFOSR grant No. F49620-01-1-0264.
1-way set associative caches, and fully-associative caches, which are \( s \)-way set associative caches. Ideally \( m \) should be small, but in order to maintain the high speed of the cache, \( k \) is usually very small. 1, 2 and 4-way caches are most commonly used.

In order to overcome “hot-spots”, where the same set associative cache is being constantly accessed, computer architects have designed hybrid cache architectures. Typically such a cache has two or more components. A given item can be placed in any of the components of the cache. Brehob et. al. [4, 3] considered the \((n, k)\) companion cache, which consists of two components: A \( k \)-way set associative called the main cache, and a fully-associative cache of size \( n \), called the companion cache (the names stem from the fact that typically \( mk \gg n \)). As argued by Brehob et. al. [4], many of the L1-cache designs suggested in recent years use companion caches as the underlying architecture. Several variations on the basic companion cache structure are possible. These include reorganization/no-reorganization and bypassing/no-bypassing. Reorganization is the ability to move an item from one cache component to another, whereas bypassing is the ability to avoid storing an accessed item in the cache. A schematic view of the companion cache is presented in Fig. 1.

![Figure 1: A schematic description of a companion cache.](image_url)

Since maintenance of the cache must be done online, and this makes it impossible to service requests optimally, we use competitive analysis. The usual assumption is that any referenced item is brought into the cache before it is accessed. Since items in the cache are accessed much more quickly than those outside, we associate costs with servicing items as follows: If the referenced item is already in the cache then we say that the reference is a hit and the cost is zero. Otherwise, we have a fault or miss which costs one. Roughly speaking, an online caching algorithm is called \( r \)-competitive if for any request sequence the number of faults is at most \( r \) times the number of faults of the optimal offline algorithm, allowing a constant additive term.

**Previous Results:** Maintenance of a fully associative cache of size \( k \) is the well known paging problem [2]. Sleator and Tarjan [11] proved that natural algorithms such as Least Recently Used are \( k \)-competitive, and that this is optimal for deterministic online algorithms. Fiat et. al. [6], improved by McGeoch and Sleator [10] and Achlioptas et. al. [1], show a tight \( \approx \ln k \) competitive
randomized algorithm. \( k \)-way set associative caches can be viewed as a collection of independent fully associative caches, each of size \( k \), and therefore they are uninteresting algorithmically.

Brehob et. al. \[3\] study deterministic online algorithms for \((n, 1)\)-companion caches. They investigate the four previously mentioned variants, i.e., bypassing/no-bypassing and reorganization/no-reorganization.

### Previous Results \[3\] (only for main cache of size \( k = 1 \):

| Bypass | Reorg’ | det’/rand’ | Upper Bound | Lower Bound |
|--------|--------|-----------|-------------|-------------|
| −      | −      | det       | 2\(n + 2\)  | \(n + 1\)   |
| √      | −      | det       | 2\(n + 3\)  | 2\(n + 2\)  |
| √      | √      | det       | 2\(n + 3\)  |             |

### New Results (main cache of arbitrary size \( k \):

| Bypass | Reorg’ | det’/rand’ | Upper Bound | Lower Bound |
|--------|--------|-----------|-------------|-------------|
| −/√    | −/√    | det       | \((n + 1)(k + 1) − 1\) | \(\Omega(nk)\) |
| −      | −      | det       | \(O(nk)\)   | \(\Omega(nk)\) |
| −      | −      | rand      | \(O(\log k \log n)\) | \(\Omega(\log k + \log n)\) |

Table 1: Summary of the results in \[3\] and in this paper, for the \((n, k)\) companion cache.

**Our Results:** This paper studies deterministic and randomized caching algorithms for a \((n, k)\)-companion cache. We consider the version where reorganization is allowed but bypassing is not. We show that the deterministic competitive ratio is exactly \((n + 1)(k + 1) − 1\). For randomized algorithms, we present an upper bound of \(O(\log n \log k)\) on the competitive ratio, and a lower bound of \(\Omega(\log n + \log k)\). For the special case of \(k = 1\) that was studied in \[3\], our bounds on the randomized competitive ratio are tight up to a constant factor. The results of \[3\] and those of this paper are summarized and compared in Table 1.

We note that any algorithm for the reorganization model can be implemented (in online fashion) in the no-reorganization model while incurring a cost at most two times larger, and any algorithm for the bypassing model can be implemented (in online fashion) in the no-bypassing model while incurring a cost at most two times larger. Thus, the competitive ratio (both randomized and deterministic) differs by at most a constant factor between the different models.

The techniques we use generalize phase partitioning and marking algorithms \[9, 6\].

2 The Problem

In the \((n, k)\)-companion caching problem, there is a slow main memory and a fast cache. The items in main memory are partitioned into \(m\) types, the set of types is \(T \ (|T| = m)\). The cache consists of two separate components:

- The Main Cache: Consisting of a cache of size \(k\) for each type. \textit{I.e.,} every type \(t, 1 \leq t \leq m\), has its own cache of size \(k\) which can hold only items of type \(t\).
- The Companion Cache: A cache of size \(n\) which can hold items of any type.
We refer to these components collectively simply as the cache. If an item is stored somewhere in the cache, we say it is cached. Our basic assumptions are that there are at least \( k + 1 \) items of every type and that the number of types, \( m \), is greater than the size of the companion cache, \( n \).

A caching algorithm is faced with a sequence of requests for items. When an item is requested it must be cached (i.e., bypassing is not allowed). If the item is not cached, a fault occurs. The goal is to minimize the number of faults. A caching algorithm can swap items of the same type between the main and companion caches without incurring any additional cost (i.e., reorganization is allowed).

We use the competitive ratio to measure the performance of online algorithms. Formally, given an item request sequence \( \sigma \), the cost of an online algorithm \( A \) on \( \sigma \), denoted by \( \text{cost}_A(\sigma) \), is the number of faults incurred by \( A \). An algorithm is called \( r \)-competitive if there exists a constant \( c \), such that for any request sequence \( \sigma \),

\[
E[\text{cost}_A(\sigma)] \leq r \cdot \text{cost}_{\text{Opt}}(\sigma) + c.
\]

To simplify the analysis later, we mention the following fact (attributed to folklore):

**Proposition 1.** We may assume that \( \text{Opt} \) is lazy, i.e., \( \text{Opt} \) evicts an item only when a requested item is not cached.

### 3 Lower bounds on the competitive ratio

Straightforward lower bounds follow from the classical paging problem.

**Theorem 1.** The deterministic competitive ratio for the \((n, k)\)-companion caching problem is at least \((n+1)(k+1)−1\). The randomized competitive ratio is at least \( H_{(k+1)(n+1)−1} = \Omega(\log n + \log k) \).

**Proof.** Consider the situation where there are \((n+1)(k+1)\) items of \( n+1 \) types, \( k+1 \) items of each type. In this case, a caching algorithm has \((n+1)k+n = (n+1)(k+1)−1\) cache slots available. Comparing this situation to the regular paging problem with a main memory of \((n+1)(k+1)\) items and a cache size of \((n+1)(k+1)−1\), we find the two problems are exactly the same. A companion caching algorithm induces a paging algorithm, and the opposite is also true. Hence a lower bound on the competitive ratio for paging implies the same lower bound for companion caching. We conclude there are lower bounds of \((n+1)(k+1)−1\) on the deterministic competitive ratio and \( H_{(n+1)(k+1)−1} = \Omega(\log n + \log k) \) on the randomized competitive ratio for companion caching.

### 4 Phase Partitioning of Request Sequences

In \([9, 6]\) the request sequence for the paging problem is partitioned into phases as follows: A phase begins either at the beginning of the sequence or immediately after the end of the previous phase. A phase ends either at the end of the sequence or immediately before the request for the \((k+1)\)st distinct page in the phase. Similarly, we partition the request sequence for the companion caching problem into phases. However, the more complex nature of our problem implies more complex partition rules.

Let \( \sigma = \sigma_1, \sigma_2, \ldots, \sigma_{|\sigma|} \) denote the request sequence. The indices of the sequence are partitioned into a sequence of disjoint consecutive subsequences \( D_1, D_2, \ldots, D_f \), whose concatenation gives
$P_i$: The indices of the requests associated with phase $i$.

$D_i$: The indices of the requests issued during phase $i$.

$N(t)$: The indices of requests of type $t$ that have not yet been associated with a phase.

$M(t) = \{ \sigma_\ell | \ell \in N(t) \}$

For every type $t \in T$: $M(t) \leftarrow \emptyset$, $N(t) \leftarrow \emptyset$

$P_1 \leftarrow \emptyset$, $D_1 \leftarrow \emptyset$

$i \leftarrow 1$

For $\ell \leftarrow 1, 2, \ldots$ Loop on the requests

Let $\sigma_\ell$ be the current request and $t_0$ its type.

Let $m_t \leftarrow \begin{cases} 
\max\{0, |M(t)| - k\} & t \neq t_0 \\
\max\{0, |M(t_0) \cup \{\sigma_\ell\}| - k\} & t = t_0 
\end{cases}$

If $\sum_{t \in T} m_t > n$ then End of Phase Processing:

For every type $t \in T$ such that $m_t > 0$ do

$P_i \leftarrow P_i \cup N(t)$

$M(t) \leftarrow \emptyset$, $N(t) \leftarrow \emptyset$

$i \leftarrow i + 1$

$P_i \leftarrow \emptyset$, $D_i \leftarrow \emptyset$

$D_i \leftarrow D_i \cup \{\ell\}$

$N(t_0) \leftarrow N(t_0) \cup \{\ell\}$

$M(t_0) \leftarrow M(t_0) \cup \{\sigma_\ell\}$

Figure 2: Phase partition rules described as an algorithm.

The indices are also partitioned into a sequence of disjoint (ascending) subsequences $P_1, P_2, \ldots, P_f$.

In Figure 2 we describe how to generate the sequences $D_i$ and $P_i$. $D_i$ is a consecutive sequence of indices of requests issued during phase $i$. $P_i$ is a (possibly non-consecutive, ascending) sequence of indices of requests associated with phase $i$. Note that $\ell \in D_i$ does not necessarily imply that $\ell \in P_i$ and vice versa. What is true is that $\ell \in D_i$ implies either that $\ell \in P_i$ for some $i' \geq i$, or $\ell \notin P_i$ for all $i'$. Note also that for all $i$, $\max D_i \geq \max P_i$.

Given a set of indices $A$ we denote by $\mathbb{I}(A) = \{ \sigma_\ell | \ell \in A \}$ the set of items requested in $A$, and by $\mathbb{T}(A)$ the set of types of items in $\mathbb{I}(A)$.

Table 2 shows an example of phase partitioning.

In [9] it is shown that any paging algorithm faults at least once in each complete phase. Here we show a similar claim for companion caching.

**Proposition 2.** For any (online or offline) caching algorithm, it is possible to associate with each phase (except maybe the last one) a distinct fault.

**Proof.** Consider the request indices in $P_i$ together with the index $j$ that ends the phase (i.e., $j = \min D_{i+1}$). One of the items in $\mathbb{I}(P_i)$ must be evicted after being requested and before $\sigma_j$ is served. This is simply because the cache cannot hold all these items simultaneously. We associate this eviction with the phase.
Table 2: An example for an \((n,k)\)-companion caching problem where \(n = 3\) and \(k = 2\). The types are denoted by the letters \(a, b, c, d\). The \(i\)th item of type \(\beta \in \{a, b, c, d\}\) is denoted by \(\beta_i\). Note that the requests for items \(d_1\) and \(d_2\) in this example are in \(P_3\), even though they are issued during phases 1 and 2 (i.e., belong to \(D_1\) and \(D_2\)).

We must show that we have not associated the same eviction to two distinct phases. Let \(i_1\) and \(i_2\) be two distinct phases, \(i_1 < i_2\). If the evictions associated with \(i_1\) and \(i_2\) are of different items then they are obviously distinct. Otherwise, the evictions associated with \(i_1\) and \(i_2\) are of the same type \(t\), and \(t \in T(P_{i_1}) \cap T(P_{i_2})\), which means that all indices \(\ell \in P_{i_2}\), where \(\sigma_\ell\) is of type \(t\), must have \(\ell > \max D_{i_1}\). Thus, an eviction associated with phase \(i_2\) cannot be associated with phase \(i_1\).

To help clarify our argument in the proof of Proposition 2, consider the third phase in Table 2. Here \(I(P_3) = \{b_1, b_2, b_3, b_4, b_5, d_1, d_2\}\), and the phase ends because of the request to \(d_3\). It is not possible that all these items reside in the cache simultaneously and thus at least one of the items in \(I(P_3)\) must be evicted before or on the request for item \(d_3\). The item evicted can be either some \(b_i, i \in \{1, 2, 3, 4, 5\}\), or some \(d_i, i \in \{1, 2\}\). If, for example, the item evicted is some \(b_i\), then this eviction must have occurred after \(\max D_1\) — the end of the first phase — and therefore it cannot be an eviction associated with the first phase.

5 Deterministic Marking Algorithms

In a manner similar to [9], based on the phase partitioning of Section 4, we define a class of online algorithms called marking algorithms.

Definition 1. During the request sequence an item \(e \in \bigcup_t M(t)\) is called marked (see Figure 2 for a definition of \(M(t)\)). An online caching algorithm that never evicts marked items is called a marking algorithm.

Remarks:

1. The phase partitioning and dynamic update of the set of marked items can be performed in an online fashion (as given in the algorithm of Fig. 2).

2. At any point in time, the cache can accommodate all marked items.
3. Unlike the marking algorithms of [9], it is not true that immediately after max $D_i$ all marks of the $i$th phase are erased. Only the marked items of types $t \in T(P_i)$ will have their markings erased immediately after max $D_i$.

For a specific algorithm, at any point in time during the request sequence, a type $t$ that has more than $k$ items in the cache is called \textit{represented in the companion cache}. Note that for marking algorithms, a type is in $T(P_i)$ if and only if it is represented in the companion cache at max $D_i$ or it is the type of the item that ended phase $i$.

\textbf{Proposition 3.} The number of faults of any marking algorithm on requests whose indices are in $P_i$ is at most $n(k + 1) + k = (n + 1)(k + 1) - 1$.

\textit{Proof.} Each item $e$ of type $t$ requested in request index $\ell \in P_i$, is marked and is not evicted until after max $D_i$. We note that $|T(P_i)| \leq n + 1$ since at most $n$ types are represented in the companion cache, and the type of the item whose request ends the phase may also be in $T(P_i)$. Thus, $|\mathbb{I}(P_i)| \leq (n + 1)k + n$. \hfill \Box

We conclude from Proposition 3 and Proposition 2.

\textbf{Theorem 2.} Any marking algorithm is $(n + 1)(k + 1) - 1$ competitive.

\textit{Proof.} Immediate from Proposition 3 and Proposition 2. \hfill \Box

Since the marking property can be realized by deterministic algorithms, we conclude

\textbf{Corollary 4.} The deterministic competitive ratio of the $(n,k)$-companion caching problem is $(n + 1)(k + 1) - 1$.

\section{Randomized Marking Algorithms}

In this section we present an $O(\log n \log k)$ competitive randomized marking algorithm. The building blocks of our randomized algorithms are the following three eviction strategies:

On a fault on an item of type $t$:

\textbf{Type Eviction.} Evict an item chosen uniformly at random among all unmarked items of type $t$ in the cache.

\textbf{Cache-wide Eviction.} Let $T$ be the set of types represented in the companion cache, let $U$ be the set of all unmarked items in the cache whose type is in $T \cup \{t\}$. Evict an item chosen uniformly at random from $U$.

\textbf{Skewed cache-wide eviction.} Let $T$ be the set of types represented in the companion cache, let $T' \subset T \cup \{t\}$ be the set of types with at least one unmarked item in the cache. Choose $t'$ uniformly at random from $T'$, let $U$ be the set of all unmarked items of type $t'$, and evict an item chosen uniformly at random from $U$.

Remarks:

1. Type eviction may not be possible as there may be no unmarked items of type $t$ in the cache.
2. Cache-wide eviction and skewed cache-wide eviction are always possible, if there are no un-
marked pages of types represented in the companion cache and no unmarked pages of type $t$
in the cache then the fault would have ended the phase.

The algorithms we use are:

Algorithm TP$_1$. Given a request for item $e$ of type $t$, not in the cache: Update all phase related
status variables (as in the algorithm of Figure 2).

- If $t$ is not represented in the companion cache and there are unmarked items of type $t$, use
type-eviction.
- Otherwise — use cache-wide eviction.

Algorithm TP$_2$. Given a request for item $e$ of type $t$, not in the cache: Update all phase related
status variables (as in the algorithm of Figure 2). Let the current request index be $j \in D_i$, $i \geq 1$.

- If $t$ is not represented in the companion cache and there are unmarked items of type $t$, use
type-eviction.
- If $t$ is represented in the companion cache, $e \in \mathbb{I}(P_{i-1})$, and there are unmarked items of type $t$,
use type eviction.
- Otherwise — use skewed cache-wide eviction.

Algorithm TP. If $k < n$ use TP$_1$, otherwise, use TP$_2$.

In the rest of this section we prove:

**Theorem 3.** Algorithm TP is $O(\log n \log k)$ competitive.

### 6.1 Basic Definitions and Proof Overview

We give an analogue to the definitions of new and stale pages used in the analysis of the randomized
marking paging algorithm of [4].

**Definition 2.** For phase $i$ and type $t$, denote by $i^{-t}$ the largest index $j < i$ such that $t \in \mathbb{T}(P_j)$. If no such $j$ exists we denote $i^{-t} = 0$, and use the convention that $P_0 = \emptyset$. Similarly, $i^{+t}$ is the smallest index $j > i$ such that $t \in \mathbb{T}(P_j)$. If no such index exists, we set $i^{+t} = \infty$, and use the convention that $P_{\infty} = \emptyset$.

**Definition 3.** An item $e$ of type $t$ is called *new* in $P_i$ if $e \in \mathbb{I}(P_i) \setminus \mathbb{I}(P_{i-1})$. We denote by $g_{t,i}$ the number of new items of type $t$ in $P_i$. Note that if $t \notin \mathbb{T}(P_i)$ then $g_{t,i} = 0$.

Let $i_{\text{end}}$ denote the index of the last *completed* phase.

**Definition 4.** For $t \in \mathbb{T}(P_i)$, let $L_{t,i} = \mathbb{I}(P_i) \cap \{\text{items of type } t\}$. Note that $|L_{t,i}| \geq k$. Define

$$
\ell_{t,i} = \begin{cases} 
|L_{t,i}| - k & i < i_{\text{end}} \land t \in \mathbb{T}(P_i) \setminus \mathbb{T}(P_{i+1}), \\
0 & \text{otherwise}.
\end{cases}
$$

8
We will use the above definitions to give an amortized lower bound (see Lemma 7) on the cost of Opt of dealing with the sequence $\sigma$:

$$\text{cost}_{\text{Opt}}(\sigma) \geq \frac{1}{4} \sum_{i \leq \text{end}} \sum_{t \in P_i} (g_{t,i} + \ell_{t,i} - t). \tag{1}$$

Our algorithms belong to a restricted family of randomized algorithms, specifically uniform type preference algorithms defined below. The main advantage of using such algorithms is that their analysis is simplified as they have the property that while dealing with requests $\sigma_j$, $j \in D_i$, the companion cache is restricted to containing only items of types in $T(P_i) \cup T(P_{i-1})$.

**Definition 5.** A type preference algorithm is a marking algorithm such that when a fault occurs on an item of a type that is not represented in the companion cache, it evicts an item of the same type, if this is possible.

**Definition 6.** A uniform type preference algorithm is a randomized type preference algorithm maintaining the invariant that at any point in time between request indices $1 + \max D_i - t$ and $\max D_i$, inclusive, and any type $t \in T(P_i)$, all unmarked items of type $t$ in $I(P_{i-1})$ are equally likely to be in the cache.

Note that both TP$_1$ and TP$_2$ are uniform type preference algorithms.

We use a charge-based amortized analysis to compute the online cost of dealing with a request sequence $\sigma$. We charge the expected cost of all but a constant number of requests in $\sigma$ to at least one of two “charge counts”, charge($D_i$) and/or charge($P_j$) for some $1 \leq i \leq j \leq \text{end}$. The total cost associated with the online algorithm is bounded above by a constant times $\sum_{1 \leq i \leq \text{end}} \text{charge}(D_i) + \sum_{1 \leq i \leq \text{end}} \text{charge}(P_i)$, excluding a constant number of requests.

Other than a constant number of requests, every request $\sigma_\ell \in \sigma$ has $\ell \in D_{i_1} \cup P_{i_2}$ for some $1 \leq i_1 \leq i_2 \leq \text{end}$.

We use the following strategy to charge the cost associated with this request to one (or more) of the charge($D_i$), charge($P_i$):

1. If $\ell \in P_i$ and type($\sigma_\ell$) $\in T(P_i) \setminus T(P_{i-1})$ then we charge the (expected) cost of $\sigma_\ell$ to charge($P_i$). These charges can be amortized against the cost of Opt to deal with $\sigma_\ell$. This amortization is summarized in Proposition 9 (for any uniform type preference algorithm).

2. If $\ell \in D_i$ and type($\sigma_\ell$) $\in T(P_{i-1})$ then we charge the (expected) cost of $\sigma_\ell$ to charge($D_i$). These charges will be amortized against the cost of Opt to within a poly-logarithmic factor. This amortization is summarized in Proposition 13 for algorithm TP$_1$ and Proposition 15 for algorithm TP$_2$.

To compute the expected cost of a request $\sigma_\ell$, $\ell \in D_i$, type($\sigma_\ell$) $\in T(P_{i-1})$, we introduce an analogue to the concept of “holes” used in [6]. In [6] holes were defined to be stale pages that were evicted from the cache.

**Definition 7.** We define the number of holes during $D_i$, $h_i$, to be the maximum over the indices $j \in D_i$ of the total number of items of types in $T(P_{i-1})$ that were requested in $P_{i-1}$ but are not cached when the $j$th request is issued.
6.2 Analysis of the Competitive Ratio for Algorithm TP

6.2.1 Lower Bounds on Opt

Proposition 5. For any request sequence \( \sigma \),

\[
\text{cost}_{\text{Opt}}(\sigma) \geq \frac{1}{2} \sum_{i \leq \text{end}} \sum_{t} g_{t,i}
\]

Proof. We may assume without loss of generality that Opt is lazy. Let \( C_i \) be the items in Opt’s cache at the end of \( P_i \) (\( C_0 = \emptyset \)). For pairs \( i, t \), let \( G_{t,i,i} \) be the set of new items in \( \mathbb{I}(P_i) \) of type \( t \) that do not appear in \( C_{i-1} \), and let \( G_{t,i,i}' \) be the set of new items in \( \mathbb{I}(P_i) \) of type \( t \) that do appear in \( C_{i-1} \). From the definitions, \( |G_{t,i,i}'| + |G_{t,i,i}'| = g_{t,i} \).

First we show that \( \text{cost}_{\text{Opt}}(\sigma) \geq \sum_i \sum_{t \in \mathbb{T}(P_i)} |G_{t,i,i}'| \). For any \( t \in \mathbb{T}(P_i) \) and for any item \( a \in G_{t,i,i}' \) we have \( a \in \mathbb{I}(P_i) \setminus C_{i-1} \). Thus, for any lazy algorithm, the first request for \( a \) in \( P_i \) is a fault. Let the request sequence \( \sigma = \sigma_1, \sigma_2, \ldots, \) define

\[
J(G_{t,i,i}') = \{ j | j = \min \{ \ell | \sigma_{\ell} = a, \ell \in P_i \}, a \in G_{t,i,i}' \}, \quad \text{and} \quad J_i' = \cup_{t \in \mathbb{T}(P_i)} J(G_{t,i,i})
\]

For any lazy algorithm, \( J_i' \) is a set of request indices that result in faults. We are left to argue that \( J_{i_1}' \cap J_{i_2}' = \emptyset \) for \( i_1 \neq i_2 \), but this is obvious, since \( J_i' \subseteq P_i \) and \( P_{i_1} \cap P_{i_2} = \emptyset \).

Next, we show that \( \text{cost}_{\text{Opt}}(\sigma) \geq \sum_i \sum_{t \in \mathbb{T}(P_i)} |G_{t,i,i}'| \). \( G_{t,i,i}' \) denotes the items in Opt’s cache after serving \( C_{i-1} \). As Opt is lazy, all items in \( G_{t,i,i}' \) must reside in the cache continuously since request index \( \text{max} D_{i-1} \). The slots used to store these items will be unavailable to deal with requests whose indices are in \( P_i \). Consider the behavior of Opt on the request sequence \( \sigma \). We claim that Opt must have at least \( \sum_{t \in \mathbb{T}(P_i)} |G_{t,i,i}'| \) evictions of items that were requested in \( P_i \) after their request, and before \( \text{max} D_{i-1} \).

For every type \( t \in \mathbb{T}(P_i) \) there were \( k + \alpha_t \) requests to different items of type \( t \) in \( P_i \), \( \sum_{t \in \mathbb{T}(P_i)} \alpha_t = n \). The total memory that we have available to deal with these \( n + k|\mathbb{T}(P_i)| \) different items is no more than \( n + k|\mathbb{T}(P_i)| \) minus the number of slots that are unavailable, i.e., the number of slots available for requests whose indices are in \( P_i \) is no more than

\[
n + k|\mathbb{T}(P_i)| - \sum_{t \in \mathbb{T}(P_i)} |G_{t,i,i}'|.
\]

Thus, Opt must have evicted at least \( \sum_{t \in \mathbb{T}(P_i)} |G_{t,i,i}'| \) of them by the end of \( \text{max} D_{i} \).

To argue that we do not count the items in \( G_{t,i,i}' \) more than once, we note that if \( t \in \mathbb{T}(P_{i_1}) \cap \mathbb{T}(P_{i_2}) \) for \( i_1 \neq i_2 \) then \( i_1^{+t} \neq i_2^{+t} \).

Proposition 6. For any request sequence \( \sigma \),

\[
\text{cost}_{\text{Opt}}(\sigma) \geq \frac{1}{2} \sum_{i \leq \text{end}} \sum_{t} \ell_{t,i},
\]

Proof. Once again, we can assume Opt is lazy. Fix \( i < i_{\text{end}} \), and a type \( t \in \mathbb{T}(P_i) \) such that \( t \notin P_{i+1} \). Let

\[
L_{t,i}' = L_{t,i} \setminus C_{i+1}; \quad L_{t,i}'' = L_{t,i} \cap C_{i+1}.
\]
Every item \( e \in L'_{t,i} \) has some \( \ell \in P_t \) such that \( \sigma_\ell = e \), and \( e \) was evicted by \( \text{Opt} \) later (but before \( \max D_{t+1} \)). Let \( \ell \) be the largest such \( \ell \in P_t \). We associate one eviction of \( e \) with index \( \ell \). In this way every eviction is associated with at most one index. Indeed, the associated eviction occurs not before \( \min D_i \), and before \( \max D_{i+1} \). At this time frame, only items from \( L_{t,i} \) and \( L_{t,i+1} \) could have been associated with this eviction, but \( t \notin T(P_{i+1}) \). Therefore \( \cos t \text{Opt}(\sigma) \geq \sum_i \sum_{t \in T(P_i)} |L'_{t,i}| \).

Let \( t \in T(P_i) \ \backslash \ T(P_{i+1}) \), and assume \( |L''_{t,i}| > k \). Such items occupy at least

\[
\phi_{i+1} = \sum_{t \in T(P_i) \backslash T(P_{i+1})} \max\{|L''_{t,i}| - k, 0\}
\]

slots in the companion cache at time \( \max D_{i+1} \). In \( P_{i+1} \) there are requests for \( |T(P_{i+1})|k+n \) different items, but considering the cache at time \( \max D_{i+1} \), these items occupy at most \( |T(P_{i+1})|k+n-\phi_{i+1} \) slots. This means that at least \( \phi_{i+1} \) of the items \( I(P_{i+1}) \), were evicted subsequently to being requested at request indices in \( P_{i+1} \) and no later than \( \max D_{i+1} \).

Associate each such eviction of item \( a \in I(P_{i+1}) \) with the largest index \( \ell \in P_{i+1} \) such that \( \sigma_\ell = a \). Note that each such eviction is associated with only one index, and therefore

\[
\cos t \text{Opt}(\sigma) \geq \sum_{i < \ell_{t,i}} \sum_{t \in T(P_i) \ \backslash \ T(P_{i+1})} \max\{|L''_{t,i}| - k, 0\}.
\]

We conclude

\[
\cos t \text{Opt}(\sigma) \geq \max\left\{ \sum_{i < \ell_{t,i}} \sum_{t \in T(P_i) \ \backslash \ T(P_{i+1})} |L'_{t,i}|, \sum_{i < \ell_{t,i}} \sum_{t \in T(P_i) \ \backslash \ T(P_{i+1})} \max\{|L''_{t,i}| - k, 0\} \right\}
\]

\[
\geq \frac{1}{2} \sum_{i < \ell_{t,i}} \sum_{t \in T(P_i) \ \backslash \ T(P_{i+1})} \max\{|L'_{t,i}| + |L''_{t,i}| - k, 0\} \geq \frac{1}{3} \sum_{i < \ell_{t,i}} \ell_{t,i} \quad \Box
\]

By taking a convex combination of the lower bounds of Proposition 5 and Proposition 6, and by algebraic manipulations, we conclude:

**Lemma 7.** For any request sequence \( \sigma \),

\[
\cos t \text{Opt}(\sigma) \geq \frac{1}{4} \sum_{i < \ell_{t,i} \in P_t} (g_{t,i} + \ell_{t,i-\ell}).
\]

### 6.2.2 Upper Bounds on TP

**Proposition 8.** Consider a marking algorithm, a phase \( i \), a type \( t \in T(P_i) \ \backslash \ T(P_{i-1}) \), and a request index \( \max D_{i-\ell} < j \leq \max D_i \). Let \( H \) be the set of items of type \( t \) that were requested in \( P_{i-\ell} \) and evicted afterward without being requested again, up to request index \( j \) (inclusive). Then \( |H| \leq \hat{g}_{t,i} + \ell_{t,i-\ell} \), where \( \hat{g}_{t,i} \leq g_{t,i} \) is the number of new items of type \( t \) requested after \( \max D_{i-\ell} \) and up to time \( j \) (inclusive).

**Proof.** Recall that \( L_{t,i-\ell} \) is the set of marked items of type \( t \) after serving \( \max D_{i-\ell} \). Let \( \hat{G}_{t,i} \) be the set of items requested after request \( \max D_{i-\ell} \) and before request index \( j \) that are not in \( L_{t,i-\ell} \), i.e., \( \hat{G}_{t,i} \) is the set new items of type \( t \) requested up to request index \( j \).
If \( i^{-t} = 0 \) then \( H \subseteq L_{t,i^{-t}} = \emptyset \). Otherwise, as \( H \subseteq L_{t,i^{-t}} \subseteq L_{t,i^{-t}} \cup \hat{G}_{t,i} \), and \( k \) items of \( L_{t,i^{-t}} \cup \hat{G}_{t,i} \) are always in the (main) cache, we conclude

\[
|H| \leq |L_{t,i^{-t}} \cup \hat{G}_{t,i}| - k \leq (|L_{t,i^{-t}}| - k) + |\hat{G}_{t,i}| = \ell_{t,i^{-t}} + \hat{g}_{t,i}. \quad \Box
\]

**Proposition 9.** For a uniform type preference algorithm, the expected number of faults on request indices in \( P_t \), for items of type \( t \in \mathbb{T}(P_i) \setminus \mathbb{T}(P_{i-1}) \) is at most \((1+H_{n+k})(g_{t,i} + \ell_{t,i^{-t}})\). I.e., \( \text{charge}(P_t) \leq (1+H_{n+k})(g_{t,i} + \ell_{t,i^{-t}}) \).

**Proof.** Fix a type \( t \in \mathbb{T}(P_i) \setminus \mathbb{T}(P_{i-1}) \). There are \( g_{t,i} \) faults on new items of type \( t \), the rest of the faults are on items in \( L_{t,i^{-t}} \) that were evicted before being requested again. By Proposition 8 the number of items in \( L_{t,i^{-t}} \) that are not in the cache at any point of time is at most \( g_{t,i} + \ell_{t,i^{-t}} \). For any \( a, b \) in \( L_{t,i^{-t}} \) that have not been requested after max \( D_{i^{-t}} \), the probability that \( a \) has been evicted since \( 1 + \max D_{i^{-t}} \) is equal to the probability that \( b \) has been evicted since \( 1 + \max D_{i^{-t}} \).

Let \( r \) denote the number of items in \( L_{t,i^{-t}} \) that have been requested after max \( D_{i^{-t}} \). There are \(|L_{t,i^{-t}}| - r \) unmarked items of \( L_{t,i^{-t}} \). The probability that an unmarked item of \( L_{t,i^{-t}} \) is not cached is therefore at most \((g_{t,i} + \ell_{t,i^{-t}})/(|L_{t,i^{-t}}| - r)\). Thus, the expected number of faults on requests indices in \( P_t \) for items in \( L_{t,i^{-t}} \) is at most

\[
\sum_{r=0}^{|L_{t,i^{-t}}| - 1} \frac{g_{t,i} + \ell_{t,i^{-t}}}{|L_{t,i^{-t}}| - r} \leq (g_{t,i} + \ell_{t,i^{-t}})H_{|L_{t,i^{-t}}|} \leq (g_{t,i} + \ell_{t,i^{-t}})H_{n+k}. \quad \Box
\]

The following proposition is immediate from the definitions.

**Proposition 10.** A type preference algorithm has the following properties:

1. During \( D_i \), only types in \( \mathbb{T}(P_{i-1}) \cup \mathbb{T}(P_i) \) may be represented in the companion cache.

2. During \( D_i \), when a type \( t \in \mathbb{T}(P_i) \setminus \mathbb{T}(P_{i-1}) \) becomes represented in the companion cache, there are no unmarked cached items of type \( t \), and \( t \) stays represented in the companion cache until \( \max D_i \), inclusive.

Recall the definition of \( h_i \), the “number of holes during \( D_i \)” (Definition 7).

**Proposition 11.** For a type preference algorithm,

\[
h_i \leq \sum_{t \in \mathbb{T}(P_i) \setminus \mathbb{T}(P_{i-1})} (g_{t,i} + \ell_{t,i^{-t}}) + \sum_{t \in \mathbb{T}(P_{i-1})} g_{t,(i-1)^{+t}}. \tag{2}
\]

**Proof.** At time \( \min D_i \), among the types in \( \mathbb{T}(P_{i-1}) \cup \mathbb{T}(P_i) \), only types in \( \mathbb{T}(P_i) \setminus \mathbb{T}(P_{i-1}) \) may have uncached items from \( L_{t,i^{-t}} \). By Proposition 9 the number of such items at the beginning of \( P_i \) is at most \( \sum_{t \in \mathbb{T}(P_i) \setminus \mathbb{T}(P_{i-1})} (\hat{g}_{t,i} + \ell_{t,i^{-t}}) \), where \( \hat{g}_{t,i} \) is the number of new items of type \( t \) requested until \( \max D_{i-1} \) (inclusive).

Consider an eviction of an item of type in \( \mathbb{T}(P_i) \cup \mathbb{T}(P_{i-1}) \) during \( D_i \). The eviction must be caused by a request to an item of either the same type or a type represented in the companion cache. By Proposition 10, the types represented in the companion cache are in \( \mathbb{T}(P_i) \cup \mathbb{T}(P_{i-1}) \), and therefore the type of the requested item is also in \( \mathbb{T}(P_i) \cup \mathbb{T}(P_{i-1}) \). If the requested item is an item of \( L_{t,i^{-t}} \), then the number of uncached items from \( L_{t,i^{-t}} \) has not changed. Otherwise, it is a new item and thus the number of new items increases. In total, we have bounded \( h_i \) as in Inequality (2). \( \Box \)
Definition 8. At any point during $D_i$, call a type $t \in \mathcal{T}(P_{i-1})$ that has unmarked items in the cache and is represented in the companion cache an active type. Call an unmarked item $e \in I(P_{i-1})$ of an active type an active item.

Note that an active item may not be cached.

The following proposition is immediate from the definitions.

Proposition 12. The following properties hold for type preference algorithms:

1. During $D_i$, the set of active types is monotone decreasing w.r.t. containment.

2. During $D_i$, the set of active items is monotone decreasing w.r.t. containment.

Proposition 13. For $\text{TP}_1$, $\text{charge}(D_i) = \text{The expected number of faults on request indices in } D_i$ to types in $\mathcal{T}(P_{i-1}) = \text{is at most } h_i(1 + H_{k+1}(1 + H_{(a+1)(k+1)}))$.

Proof. First, we count the expected number of faults on items in $\bigcup_{t \in \mathcal{T}(P_{i-1})} L_{t,i-1}$. By Proposition 12, the set of active items is monotone decreasing, where an item becomes inactive either by being marked, or because its type is no longer represented in the companion cache. Let $(m_j)_{j=1,...,w_i}$ be the sequence of numbers of active items indexed on the events. An event is either when an active item is requested, or when an active type $t$ becomes inactive by being no longer represented in the companion cache (it is possible that one request generates two events, one from each case).

If the $j$th event is a request for active item, then $m_{j+1} = m_j - 1$. Otherwise, if the $j$th event is the event of type $t$ becoming inactive, and before that event there were $b$ active items of type $t$, then $m_{j+1} = m_j - b$.

In the first case, the expected cost of the request, conditioned on $m_{j}$, is at most $h_i/m_j$.

In the second case, there are $b$ items of type $t$ that became inactive, each had probability at most $h_i/m_j$ of not being in the cache at that moment. This means that the expected number of items among the up-until now active items of type $t$, that are not in the cache, at this point in time, is at most $h_i/m_j$.

Let $g_t$ denote the number of new items of $P_{(i-1)^{+t}}$ (Definition 3) requested during $D_i$ ($g_t \leq g_{L_{t,i-1}}$). After type $t$ becomes inactive, the number of items among $L_{t,i-1}$ that are not in the cache can increase only when a new item of type $t$ is requested. Therefore the expected number of items among $L_{t,i-1}$ that are not in the cache, after the $j$th event (the event when $t$ became inactive), is at most $h_i/m_j + g_t$.

Because of the uniform type eviction property of $\text{TP}_1$, the probability that an item in $L_{t,i-1}$ is not in the cache is the expected number of items among $L_{t,i-1}$, and not in the cache, divided by the number of unmarked items among $L_{t,i-1}$, and therefore the expected number of faults on items of $L_{t,i-1}$ after the $j$th event is at most

$$\sum_{a=1}^{b} \left( \frac{h_i b}{m_j} + g_t \right) \cdot \frac{1}{a} = \left( \frac{h_i b}{m_j} + g_t \right) H_b.$$ 

Note that $b \leq k + 1$, and $\sum_{t \in P_{i-1}} g_t \leq h_i$, and so the expected number of faults on items $e \in \bigcup_{t \in \mathcal{T}(P_{i-1})} L_{t,i-1}$, conditioned on the sequence $(m_j)_{j}$ is at most

$$h_i H_{k+1} + h_i \sum_{j} \left( \frac{m_j - m_{j-1}}{m_j} \right) H_{k+1} \quad (3)$$
Thus, sequences $A$ of items in $L$ of number of faults on items $t$ active item of active type $D$ (faults on request indices in Lemma 14. TP $L$ items among $T$. Each fault is counted by either charge($P$).

Proof. probability that an active item of type $t$ is active, it does not change the numbers of active items not in the cache.

We prove this by induction on the length of the request sequence. Before request index min $D_i$, all items among the active types are in the cache, and the claim trivially holds. A fault on an item of $L_{t,i-1}$, not currently in the cache, of active type $t$, is served by type eviction and therefore the number of items from $L_{t,i-1}$ and not in the cache does not change. A fault on an item of a type not represented in the companion cache that has unmarked items, is served by type eviction, and since that type is not active, it does not change the numbers of active items not in the cache.

If type eviction is not used then the fault is served by increasing the number of items not in the cache among the active types. In this case a skewed cache-wide eviction is used, which chooses a skewing to evict an unmarked page of that type uniformly at random.

We claim that for algorithm TP$_2$, we have similar arguments.

Proposition 15. For TP$_2$, charge($D_i$) — The expected number of faults on request indices in $D_i$ to types in $T(P_{i-1})$ is at most $\sum_{i \in T(P_{i-1})} g_{t,(i-1)+t} \leq h_i$ such faults.

We conclude

Lemma 14. TP$_1$ is $O(\log k \max\{\log n, \log k\})$ competitive.

Proof. Each fault is counted by either charge($P_i$) (Proposition 13) or charge($D_i$) (Proposition 14) (faults on request indices in $D_i$ for items of type in $T(P_{i-1}) \setminus T(P_i)$ are counted twice), and by Lemma 7, we have that the expected number of faults of TP$_1$ is at most

$$(5(1 + H_{n+k}) + 10(1 + H_{k+1}(1 + H_{n+k}(k+1)))) \cdot \text{cost}_{\text{Opt}}.$$
Let $b_i$ be the number of active types immediately before type $t$ became inactive. Thus, $E[u_t] \leq \frac{h_i}{b_t}$, immediately before type $t$ becomes inactive. As of this point of time, $u_t$ could increase only if a new item of type $t$ is requested. We can therefore bound the value $E[u_t] \leq \frac{h_i}{b_t} + g_{t,(i-1)^{+\ell}}$, throughout $D_i$. The expected number of faults on items of type $t$ is at most

$$
\sum_{r=0}^{|L_{t,i-1}| - 1} \frac{h_i}{b_t} + g_{t,(i-1)^{+\ell}} = \left( \frac{h_i}{b_t} + g_{t,(i-1)^{+\ell}} \right) H_{|L_{t,i-1}|}.
$$

Using the facts that $|L_{t,i-1}| \leq n + k$, and $\sum_{t \in \mathbb{T}(P_{t-1})} g_{t,(i-1)^{+\ell}} \leq h_i$, and summing over all $t \in \mathbb{T}(P_{t-1})$, the expected number of faults on types in $\mathbb{T}(P_{t-1})$ is at most

$$
\sum_{t \in \mathbb{T}(P_{t-1})} \left( \frac{h_i}{b_t} + g_{t,(i-1)^{+\ell}} \right) H_{|L_{t,i-1}|} \leq h_i H_{n+k} \sum_{t \in \mathbb{T}(P_{t-1})} b_t^{-1} + h_i H_{n+k} \leq h_i H_{n+k}(1 + H_{n+1}).
$$

We have bounded from above the expected number of faults on items in $\cup_{t \in \mathbb{T}(P_{t-1})} L_{t,i-1}$. We also need to add at most $h_i$ faults on new items of types in $\mathbb{T}(P_{t-1})$.  

We summarize,

**Lemma 16.** $\mathbb{TP}_2$ is $O(\log n \max\{\log n, \log k\})$ competitive.

**Proof.** Each fault is counted by either charge($P_i$) (Proposition 9) or charge($D_i$) (Proposition 15) (faults on request indices in $D_i$ for items of type in $\mathbb{T}(P_{t-1}) \setminus \mathbb{T}(P_t)$ are counted twice), and by Lemma 7 we have that the expected number of faults of $\mathbb{TP}_2$ is at most

$$
(5(1 + H_{n+k}) + 10(1 + (1 + H_{n+1})H_{n+k})) \text{cost}_{\text{Opt}}.
$$

**Proof of Theorem 3.** Follows immediately from Lemma 13 and Lemma 16.

Unfortunately, the competitive ratio of a type preference algorithm is always $\Omega(\log n \log k)$.

**Example 1.** The following example proves that the competitive ratio of a type preference algorithm is always $\Omega(\log n \log k)$. Let $A$ be a type preference algorithm. Let $m = n + 1$ and assume there are exactly $k + 1$ items from each type. In each $P_i$ there is only one new item. At the beginning of phase $i$, $\min D_i$, the adversary requests all the items with the same type as the new item, and $A$ incurs a cost of $H_{k+1}$. After that, $A$ is forced to evict an item of a different type. The adversary chooses a type that has the hole in it with probability at least $\frac{1}{n}$ and requests all the items of this type each time choosing the item with maximum probability of being a hole. This costs $A$

$$
\frac{1}{n(k+1)} + \frac{1}{nk} + \frac{1}{n(k-1)} + \cdots + \frac{1}{n} = \frac{H_{k+1}}{n}.
$$

After that, the hole is in one of $n - 1$ types. Again, the adversary picks a type that has the hole in it with probability at least $\frac{1}{n-1}$ and requests all the items of this type each time choosing the item with maximum probability of being a hole, which costs $A \frac{H_{k+1}}{(n-1)}$, and so on. In total, the expected cost for $A$ for the phase is $H_{k+1}H_n$. 

15
7 Concluding Remarks

We have shown that the deterministic competitive ratio for \((n, k)\)-companion caching is exactly \((n + 1)(k + 1) - 1\). We have also shown a lower bound of \(\Omega(\log n + \log k)\) and an upper bound of \(O(\log n \log k)\) on the randomized competitive ratio. We conjecture that the lower bound we have proven is tight. Specifically, we conjecture that the following algorithm is \(O(\log n + \log k)\) competitive.

**Algorithm CW:** On a fault on item \(e\) of type \(t\): let \(i \geq 1\) be the current phase. If \(t \notin T(P_{i-1})\), use type eviction if possible. Otherwise, use cache-wide eviction.

Acknowledgments

This paper reflects joint work between the listed authors and Amos Fiat. Due to the tragic circumstances, and at Professor Fiat’s insistence, he is not listed as an author.

References

[1] D. Achlioptas, M. Chrobak, and J. Noga. Competitive analysis of randomized paging algorithms. *Theoretical Computer Science*, 234:203–218, 2000.

[2] L.A. Belady. A study of replacement algorithms for virtual storage computers. *IBM Systems Journal*, 5:78–101, 1966.

[3] M. Brehob, R. Enbody, E. Tornq, and S. Wagner. On-line restricted caching. *Journal of Scheduling*, 6:149–166, 2003.

[4] M. Brehob, S. Wagner, E. Tornq, and R. Enbody. Optimal replacement is NP-hard for non-standard caches. Technical Report MSU-CSE-00-14, Department of Computer Science, Michigan State University, June 2000.

[5] K. K. Chan, C. C. Hay, J. R. Keller, G. P. Kurpanek, F. X. Schumacher, and J. Zheng. Design of the HP PA7200. *Hewllet-Packard Journal*, February 1996.

[6] Amos Fiat, Richard Karp, M. Luby, L. A. McGeoch, Daniel D. Sleator, and N. E. Young. Competitive paging algorithms. *Journal of Algorithms*, 12:685–699, 1991.

[7] A. Fiat, M. Mendel, and S. S. Seiden. Online companion caching. *Proc. of the 10th Annual European Symposium on Algorithms*, Springer LNCS 2461, 499–511, 2002.

[8] N. Jouppi. Improving direct-mapped cache by the addition of small fully-associative cache and prefetch buffer. *Proc. of the Seventeenth Annual International Symposium on Computer Architecture*, 18(2):364–373, 1990.

[9] Anna Karlin, M. Manasse, L. Rudolph, and Daniel D. Sleator. Competitive snoopy caching. *Algorithmica*, 3(1):79–119, 1988.

[10] L. McGeoch and D. Sleator. A strongly competitive randomized paging algorithm. *J. Algorithms*, 6:816–825, 1991.
[11] Daniel D. Sleator and Robert E. Tarjan. Amortized efficiency of list update and paging rules. *Communication of the ACM*, 28:202–208, 1985.

[12] A. Seznec. A case for two-way skewed-associative caches. In *Proc. of the 20th International Symposium on Computer Architecture*, 169–178, 1993.