More on weighted servers

or

FIFO is better than LRU[*]

Leah Epstein[†] Csanád Imreh[‡] Rob van Stee[§]

Abstract

We consider a generalized 2-server problem on the uniform space in which servers have different costs. Previous work focused on the case where the ratio between these costs was very large. We give results for varying ratios. For ratios below 2.2, we present a best possible algorithm which is trackless. We present a general lower bound for trackless algorithms depending on the cost ratio, proving that our algorithm is the best possible trackless algorithm up to a constant factor for any cost ratio. The results are extended for the case where we have two sets of servers with different costs.

[*]A preliminary version of this paper appeared in Proc. of 27th Mathematical Foundations of Computer Science (MFCS 2002), LNCS 2420, pages 257-268.

[†]School of Computer Science, The Interdisciplinary Center, P.O.B 167, 46150 Herzliya, Israel. lea@idc.ac.il. Research supported in part by the Israel Science Foundation (grant no. 250/01).

[‡]University of Szeged, Árpád tér, 6701, Szeged, Hungary, most part of this work was done while the author was postdoc at MPI für Informatik, cimreh@amadea.inf.u-szeged.hu. Research supported by Future and Emerging Technologies programme of the EU under contract number IST-1999-14186 (ALCOM-FT) and by the Hungarian National Foundation for Scientific Research, Grant TO30074.

[§]Institut für Informatik, Albert-Ludwigs-Universität, Georges-Köhler-Allee, 79110 Freiburg, Germany. vanstee@uni-freiburg.de. Work supported by the Deutsche Forschungsgemeinschaft, Project AL 464/3-1, and by the European Community, Projects APPOL and APPOL II.

1 Introduction

The weighted k-server problem was introduced by Fiat and Ricklin [8]. In this problem, we are given a metric space $M = (A, d)$ with k mobile servers, and each server s_i has a weight $w_i > 0$. Here A is a set of points and d is a distance function (metric). At each step, a request $r \in M$ is issued that has to be served by one of the servers by moving to r. The cost for server s_i to serve request r is $d(r, s_i) \cdot w_i$. This problem is an on-line problem: each time that there is a request, it needs to be served before the next request becomes known. We denote the cost of an algorithm $\textsc{Alg}$ on a request sequence σ by $\textsc{Alg}(\sigma)$. We denote an optimal off-line algorithm that knows all the input in advance by $\textsc{Opt}$. The goal of an on-line algorithm $\mathcal{A}$ is to minimize its competitive ratio $\mathcal{R}(\mathcal{A})$, which is defined as the smallest value $\mathcal{R}$ that satisfies

$$\mathcal{A}(\sigma) \leq \mathcal{R} \cdot \textsc{Opt}(\sigma) + c,$$

for any request sequence σ and some constant c (independent of σ).

In [8] a doubly exponential upper bound in k is given for uniform spaces. Furthermore, for the special case where only two weights are allowed, a $k^{O(k)}$ competitive algorithm is presented. They also show that the competitive ratio is at least $k^{\Omega(k)}$ in any space with at least $k + 1$ points. In [11], a simple upper bound of kw_{avg}/w_{min} is proven for the general case, where w_{min} is the minimal and w_{avg} is the average weight of the servers.

The special case of two servers and uniform spaces was investigated in [7]. There a 5-competitive version of the Work Function Algorithm and matching lower bound, and a 5-competitive memoryless randomized algorithm and matching lower bound are presented.

All previous work (except the simple result of [11]) focuses on the asymptotic case where the ratio between the weights of the servers tends to ∞. We consider instead the case of smaller ratios and obtain the surprising result that for the weighted 2-server problem in a uniform space, an algorithm that uses both its servers equally is best possible as long as the ratio between the weights is at most 2.2.

We also consider the more general case, where we have κ servers with speed 1 and κ servers with speed w. The total number of servers is 2κ. Since we only investigate uniform spaces, the problem can also be seen as a caching problem where we have two caches of size κ: the cheap cache C, and the expensive cache E. This type of cache is called *a two-level cache* [1, 5]. For this reason, we borrow some terminology from caching theory. We formulate our problem as follows. The algorithm has to serve a sequence of requests for pages. If the requested page is not in either of the caches, then we have to put it into one of the caches, evicting a page from the chosen cache if it is full. This event is called a fault. The set of possible pages is called the slow memory. Moving a page into C has cost 1, and moving a page into E has a cost of $w \geq 1$.

All of the previous algorithms for the weighted server problem store information about most of the previously requested points: the algorithm SAMPLE of [8] has a counter of the points, in the work function algorithm of [7] this information is coded in the work function. This yields that these algorithms might have an extremely large space requirement if there is a large number of different requested points. Moreover, and more importantly, they perform very slowly. For the original k-server problem, a class of algorithms called trackless is introduced in [3] to avoid this problem.

Definition *Trackless* algorithms are algorithms that for each request receive as input only the distances between current server positions and the request point.

A trackless algorithm may memorize such distance values, but it is restricted from explicitly storing any points of the metric space. This also means that it must be *lazy*: only moving one server for each request that occurs.

For the special case of uniform spaces, the trackless property changes into the rule that the algorithm is not allowed to use bookmarks in the slow memory to distinguish between the pages. This means that in the case of a fault, the decision as to which cache is used for the requested page and which page is removed from this cache must be independent of the requested page itself. This restriction seems to be very strong at first look, but we must note that for paging the best possible deterministic marking algorithms are trackless [9], and even for the more general web caching problem the best possible algorithm is trackless [13]. Randomized trackless algorithms for the paging problem are investigated in [2]. It is shown that it is necessary to use bookmarks to reach the best possible $\log k$ competitive ratio.

Our Results:

The results we show in this paper are as follows:

For $\kappa = 1$ we introduce a trackless algorithm which is based on the well-known paging algorithm FIFO. As mentioned before this algorithm has best possible competitive ratio for $w \leq 2.2$. Specifically it has competitive ratio $\max\{2, 3(1+x)/4\}$. We also analyze a modified version of the other well-known paging algorithm LRU, and we obtain that in the weighted case FIFO is better than LRU, which has the competitive ratio $1 + x$. This is in sharp contrast to the intuition from previous literature [6] and practice. This surprising result hints that even in practice it might be the case that two-level caches (of relatively small size) would work more efficiently using FIFO rather than using the standard LRU. A third algorithm we study is an adaptation of the widely known algorithm BALANCE. We show that BALANCE performs even worse than LRU for infinitely many values of w (its competitive ratio is $2w$ for even values of w). All the above results hold for $\kappa = 1$, however the adaptations of FIFO and LRU are defined not only for $\kappa = 1$ but also for general values of κ.

Next we move on to trackless algorithms and $\kappa \geq 1$. We show that for such algorithms the competitive ratio must grow (at least) linearly as a function of κw (it is at least $\kappa(1+w)/2$ for $\kappa \geq 2$ and at least $w/2+1$ for $\kappa = 1$). For the sake of completeness we also show a simple upper bound of $\kappa(1+w)$ for general κ (which is a special case of the upper bound from [11]), pointing out that all marking algorithms (including our version of LRU) are best possible trackless algorithms up to a constant factor 2 (in terms of competitive ratio). The proof also holds for our version of FIFO.

2 Marking Algorithms

To begin with, we give a simple proof for the competitiveness of marking algorithms, which is a special case of the bound from [11].

We first partition the request sequence into phases in the following well-known way: each phase $i \geq 1$ is the maximal subsequence of the request sequence that contains requests to at most 2κ distinct pages, and that starts with the first request after phase $i-1$ ends. Phase 1 starts with the first request of the sequence.

The marking algorithms, which are defined in [9], unmark all the pages in the caches at the beginning of each phase. When a page from one of the caches is requested we mark it. If there is a fault, then we evict some unmarked page (determined by the algorithm), load the requested page and mark it. From the definition of the phase partitioning, we can see that a phase is ended when we have a fault and every page in the caches is marked. We unmark the pages and the new phase is started.

Theorem 2.1 *For any marking algorithm $\mathcal{A}$, $\mathcal{R}(\mathcal{A}) \leq \kappa(1+w)$. Moreover, we also have $\mathcal{R}(\text{FIFO}) \leq \kappa(1+w)$.*

Proof. Every marking algorithm has a cost of at most $\kappa(1+w)$ per phase, whereas OPT has a cost of at least 1 per phase. This also holds for FIFO. $\square$

3 Two Servers

It is possible to adapt the well-known paging algorithms FIFO and LRU for the current problem by considering the two caches as one big cache and ignoring the difference in costs. We denote these adaptations also by FIFO and LRU. We begin by proving the following theorem.

Theorem 3.1 *For $\kappa = 1$,*

$$\mathcal{R}(\text{FIFO}) = \max\left(2, \frac{3}{4}(1+x)\right).$$

Definition 3.1 *A* relevant *request sequence is a request sequence on which* FIFO *faults on every request.*

We can map any request sequence σ onto a relevant request sequence σ' by removing all requests on which FIFO does not fault. This does not affect the cost of FIFO or its decisions: a request in σ' is put into the cheap cache by FIFO if and only if FIFO puts the corresponding request in σ in the cheap cache. Hence $\text{FIFO}(\sigma') = \text{FIFO}(\sigma)$. Moreover, $\text{OPT}(\sigma') \leq \text{OPT}(\sigma)$. Therefore we only need to consider relevant request sequences to determine the competitive ratio of FIFO. Note that for a relevant request sequence, a phase always consists of two consecutive requests. W.l.o.g. we can assume that OPT only loads a page when it is requested and OPT does not have it in a cache.

Lemma 3.1 *Any three consecutive requests in a relevant request sequence are to distinct pages.*

Proof. If this were not the case, FIFO would not fault on every request.

Lemma 3.2 *Suppose* $\{a_i\}_{i=1}^{\ell}$ *is a relevant sequence. For any* $1 \le j \le \ell - 2$, OPT *faults on at least one of the requests* a_{j+1} *and* a_{j+2}.

Proof. We use Lemma 3.1. If OPT does not fault on a_{j+1}, then it has a_j and a_{j+1} in its cache after the request for a_j. Thus OPT must fault on a_{j+2}, since that is a different page. □

Definition 3.2 *A relevant interval is a maximal subsequence* $\{a_i\}_{i=j_1}^{j_2}$ *of a request sequence with the following property:* OPT *faults on* $a_j \iff (j - j_1) \bmod 2 = 0$.

By Lemma 3.2 we can partition any relevant sequence into relevant intervals. Here we assume that both OPT and FIFO start with empty caches. Note that all relevant intervals (except possibly the very last one) end with a fault by OPT, and have odd length.

Lemma 3.3 *Consider a relevant interval* I *of length* $\ell(I) \ge 5$ *and write* $f = \lfloor \ell(I)/4 \rfloor$. *Then* OPT *has at least* f *expensive faults on* I.

Proof. Partition I into f subintervals of length 4, followed by one final subinterval of length at most 3. Consider a subinterval of length 4. It has the form

$$\mathbf{a}b\ \mathbf{c}d$$

where OPT faults only on a and c. Thus d is already in OPT's cache when it is requested. By Lemma 3.1, $d \ne b$ and $d \ne c$. Suppose $d \ne a$. Then after OPT loads a, it must have both a and d in the cache. Since $b \ne a$ by Lemma 3.1, this means

$\textsc{Opt}$ faults on b, a contradiction. Hence $d = a$. Since a remains in the cache when $\textsc{Opt}$ serves c, either a or c must be loaded into the expensive cache by $\textsc{Opt}$.

Therefore $\textsc{Opt}$ has at least one expensive fault per subinterval. $\square$

Lemma 3.4 *For any relevant request sequence σ of length ℓ, we have*

$$\textsc{Opt}(\sigma) \geq \min\left(\frac{2}{3}, \frac{1+w}{4}\right) \cdot \ell.$$

Proof. We partition σ into relevant intervals. Consider any relevant interval I and denote its length by $\ell(I)$. We show that for each such interval, $\textsc{Opt}(I) \geq \min(2/3, (1+w)/4) \cdot \ell(I)$.

If $\ell(I) = 1$, $\textsc{Opt}(I) \geq 1 \geq 2/3$. If $\ell(I) = 3$, $\textsc{Opt}(I) \geq 2 = \frac{2}{3} \cdot 3$.

Suppose $\ell(I) \geq 5$, and write $f = \lfloor \ell(I)/4 \rfloor$. We use Lemma 3.3.

If $\ell(I) = 4f + 1$ then $\textsc{Opt}$ has $2f + 1$ faults, among which at least f are expensive. This is a cost of $(f + 1 + fw)/(4f + 1)$ per request.

If $\ell(I) = 4f + 3$ then $\textsc{Opt}$ has $2f + 2$ faults, among which at least f are expensive. This is a cost of $(f + 2 + fw)/(4f + 3)$ per request.

Both of these are at least $(1 + w)/4$ or at least $2/3$. $\square$

Lemma 3.5 *For $\kappa = 1$, $\mathcal{R}(\textsc{Fifo}) \leq \max(2, 3(1 + x)/4)$.*

Proof. Consider a relevant request sequence σ. We partition σ into phases, and compare the cost of $\textsc{Fifo}$ to the *average* cost of $\textsc{Opt}$ per phase. By Lemma 3.4, we have that $\textsc{Opt}(p) \geq \min(4/3, (1 + w)/2)$ for any phase p on average. We also have $\textsc{Fifo}(p) = 1 + w$ exactly. Therefore, $\mathcal{R}(\textsc{Fifo}) \leq \max(2, 3(1 + x)/4)$. $\square$

We now show a general lower bound for $\textsc{Fifo}$, that holds if both the cheap cache and the expensive cache have some size $\kappa \geq 1$.

Lemma 3.6 *For $\kappa \geq 1$,*

$$\mathcal{R}(\textsc{Fifo}) \geq \max\left(2\kappa, \frac{2\kappa + 1}{2\kappa + 2}\kappa(1 + w)\right).$$

Proof. Consider the following request sequence:

$$(1\ 2\ 3\ \ldots\ 2\kappa + 1)^N .$$

We define two different offline strategies to serve this sequence, and use one or the other depending on x.

For $x \geq (2\kappa + 3)/(2\kappa + 1)$, OPT has $1, \ldots, \kappa$ in the cheap cache at the start, and $\kappa + 1, \ldots, 2\kappa$ in the expensive cache. OPT only faults on pages $1, \ldots, \kappa$ and $2\kappa+1$, each time evicting the page from this set which will be requested the furthest in the future (LFD [4]). In $2\kappa + 1$ phases, there are $2\kappa(2\kappa + 1)$ requests in total and $2\kappa(\kappa + 1)$ requests to pages $1, \ldots, \kappa$ and $2\kappa + 1$. OPT has a fault once every κ requests to these pages, so it has $2(\kappa + 1)$ faults in $2\kappa + 1$ phases.

FIFO pays $\kappa(1 + w)$ per phase, so by letting N grow without bound we find

$$\mathcal{R}(\text{FIFO}) \geq \frac{2\kappa + 1}{2\kappa + 2}\kappa(1 + w) .$$

For $x < (2\kappa + 3)/(2\kappa + 1)$, OPT uses its entire cache in a round-robin fashion. Thus OPT pays on average $(1 + w)/2$ per FIFO phase, so

$$\mathcal{R}(\text{FIFO}) \geq 2\kappa.$$

□

Proof of Theorem 3.1. This follows from Lemma 3.6 for the case $\kappa = 1$ and from Lemma 3.5. □

FIFO cannot have best possible competitive ratio for every w since according to [7] there exists a constant competitive algorithm whereas the competitive ratio of FIFO grows linearly with w. However for relatively small values of w, FIFO has best possible competitive ratio.

Theorem 3.2 FIFO *is a best possible on-line algorithm for the weighted server problem and* $\kappa = 1$ *in the interval* $1 \leq w \leq \sqrt{249}/6 - 1/2 \approx 2.1299556$.

Proof. We prove a general lower bound that matches the upper bound of FIFO from Theorem 3.1. This even holds for a 3-point space.

We construct a sequence of requests, which consists of requests for at most three different pages: 0, 1 and 2. Consequently there are 6 different possible configurations of the cache. Two basic ingredients for the proof are similar to the lower bound for the k-server problem [10]. Many new ingredients are added to the analysis, that becomes more involved for the related case.

As in [10], the sequence of requests is constructed in such a way that the on-line algorithm has a fault in every request. That is, in each step a request is added for the only page that is not present in either of the two caches of the on-line algorithm. In order to give a lower bound on the general (not strict) competitive ratio, we build a long enough sequence. Note that the on-line cost of such a sequence of N requests is at least N. Another similarity to [10] is a comparison of the on-line algorithm to several off-line algorithms. We keep five off-line algorithms that all process the complete sequence along with the on-line algorithm, and consider their average cost. This is a lower bound for the optimal off-line cost.

A sequence of requests that is constructed in this way can be transformed into a sequence of costs that the on-line algorithm pays in every step. We define a cost sequence as a sequence of 1's and w's, where 1 indicates a fault of the cheap cache whereas w indicates a fault of the expensive cache. Given a starting configuration of the on-line algorithm (without loss of generality $C = \{0\}$ and $E = \{1\}$) and a cost sequence, it is easy to recall the request sequence. We define a set of pattern sequences S which consists of a finite number of cost sequences of bounded length s. The patterns in S form a prefix code. In other words, there is a unique way to partition any cost sequence into members of S (a remainder of length at most $s-1$ might be left over). We call the points in the cost sequence where a member of S ends 'breakpoints'. The starting point of the complete sequence is also considered a breakpoint.

Besides having to serve the request sequence, a second requirement from the off-line algorithms is that at every breakpoint, each of the six algorithms (that is, the five off-line algorithms and the on-line algorithm) have a different configuration. This means that at every breakpoint, exactly one algorithm has each possible configuration. Note that we do not require a certain order among the configurations of the off-line algorithms.

For each of the patterns $p \in S$, we compute the cost of the on-line algorithm on p and compute the cost of the five off-line algorithms to serve all requests and end up in a valid configuration (i.e. so that all six final configurations are again all different). Note that this gives room to the off-line algorithms to choose which configuration each of them ends up in, and there are $5! = 120$ possibilities.

As we need to show a lower bound of $\mathcal{R}$, we compare the on-line cost on p, denoted by $\textsc{Onl}(p)$, to the total cost of the off-line algorithms $\textsc{FiveOffs}(p)$, and show $5 \cdot \textsc{Onl}(p) \geq \mathcal{R} \cdot \textsc{FiveOffs}(p)$. This implies that for the total off-line cost we have

$$\textsc{FiveOffs}(\sigma) \leq \frac{5}{\mathcal{R}} \cdot \textsc{Onl}(\sigma) + 5(1+w) + 5(s-1)w.$$

The value $5(1+w)$ is the setup cost of the five off-line algorithms to reach the correct starting configurations, and $5(s-1)w$ is an upper bound on the cost of serving the 'remainder' of the request sequence which is not a pattern in S. As $\textsc{Opt}$ is an optimal off-line algorithm, its cost is at most the average cost of the five off-line algorithms and so $\textsc{Opt}(\sigma) \leq 1/\mathcal{R} \cdot \textsc{Onl}(\sigma) + 1 + sw$. We get that

$$\textsc{Onl}(\sigma) \geq \mathcal{R} \cdot \textsc{Opt}(\sigma) - 1 - sw.$$

It is left to show the sets S that imply the lower bound. We show a pattern set S_1 that proves a lower bound of 2 for every value of w in the interval $[1, 5/3]$, and a pattern set S_2 which proves the lower bound $3(1+w)/4$ for every value of w in the interval $[5/3, (\sqrt{249}-3)/6]$. See Table 1. The upper limit of the second interval is the solution of the equation $\frac{3}{4}(1+w) = 5(3w+4)/(w+20)$ (line 11 in the second table). To check that those patterns are indeed valid (give the correct lower bound on the competitive ratio) it is only required to solve a linear or a quadratic equation for each pattern p. It is also easy to see that both pattern sets form a prefix code. The cost of the on-line algorithm follows directly from the pattern whereas the off-line costs require a short proof. The details of that are omitted. $\square$

Pattern p	$\textsc{Onl}(p)$	$\textsc{FiveOffs}(p)$	Pattern p	$\textsc{Onl}(p)$	$\textsc{FiveOffs}(p)$
111	3	7	w	w	$w+2$
$ww1$	$2w+1$	$2w+5$	$11w$	$w+2$	$w+6$
$11w$	$w+2$	$w+6$	$1ww$	$2w+1$	9
www	$3w$	$2w+5$	1111	4	8
$w111$	$w+3$	$w+9$	$111w$	$w+3$	$w+7$
$w1w1$	$2w+2$	$5w+5$	$1w11w$	$2w+3$	15
$1w1w$	$2w+2$	$5w+5$	$1w1ww$	$3w+2$	$w+14$
$1ww1$	$2w+2$	$2w+8$	$1w1111$	$w+5$	$w+13$
$w11w$	$2w+2$	$2w+8$	$1w111w$	$2w+4$	$2w+12$
$w1ww$	$3w+1$	$4w+6$	$1w1w1w$	$3w+3$	20
$1www$	$3w+1$	$2w+8$	$1w1w11w$	$3w+4$	$w+20$
$1w111$	$w+4$	$3w+8$	$1w1w1111$	$2w+6$	$5w+11$
$1w11w11$	$2w+5$	$4w+13$	$1w1w111w$	$3w+5$	$3w+17$
$1w11ww1$	$3w+4$	$4w+13$			
$1w11w1w$	$3w+4$	$5w+12$			
$1w11www$	$4w+3$	$3w+14$			

Table 1: Left are the patterns for $w \in [1, 5/3]$, right for $w \in [5/3, 2.1299556]$.

Even though all the data was carefully verified, the origin of the pattern sets is a computer program we used. The program performs an exhaustive search on all 120 permutations and finds the cheapest cost of the off-line algorithms for each possible pattern. The result given here is an output of the program when it checked all patterns up to a length of 11. Using this program for patterns of length 13, we also found that FIFO is best possible for $w \leq 2.206$. An extension of the program for patterns of length at most 15 improves the bound by a negligible amount.

Interestingly, it is possible to show that under the current model LRU is strictly worse than FIFO, in contrast to the model in [6]. We have the following result.

Lemma 3.7 $\mathcal{R}(\text{LRU}) = 1 + w$.

Proof. The upper bound follows from Theorem 2.1.

Denote the starting caches of LRU by $C = \{0\}$ and $E = \{1\}$. Without loss of generality, we may assume LRU starts by using the cheap cache. We use the following sequence: $(202101)^N$. For one iteration of this sequence (six requests), LRU pays $2 + 2w$. On the other hand, an offline algorithm can keep 0 in E at all times and only pay 2 per iteration. By letting N grow without bound, we get the desired result. □

Note that the lower bound from Lemma 3.6 also holds for LRU, so for large κ both algorithms have a competitive ratio that tends to $\kappa(1 + w)$. We end this section with a short analysis of two other natural algorithms. When w is large, it is tempting to use an algorithm which uses only the cheap cache, i.e. each time there is a page fault, the algorithm replaces the page in C by the requested page. This algorithm is not competitive at all. Consider the on-line cache in the beginning of the sequence, and denote $C = \{a\}$ and $E = \{b\}$. Let c and d be pages such that all four pages a, b, c, d are distinct. The sequence simply alternates between pages c and d. Let $2N$ be the length of the sequence. Then the cost of the algorithm is

$2N$, whereas OPT can initialize its caches by $C = \{c\}$ and $E = \{d\}$ and pay only $1 + w$. As N grows, the ratio of the costs grows without bound.

Another natural option is to apply the BALANCE algorithm, trying to balance the costs of the two caches. As BALANCE performs as well as LRU and FIFO for $w = 1$ (competitive ratio 2 for paging), it is interesting to see how it performs for other values of w. The modified definition of BALANCE for $w > 1$ is as follows:

Keep a counter for each place in the cache (B_C and B_E). On a page fault, if $B_C + 1 \leq B_E + w$, replace the page in C and increase B_C by 1; otherwise replace the page in E and increase B_E by w.

Lemma 3.8 *The competitive ratio of* BALANCE *is at most* $2w$. *This is tight for infinitely many values of* w.

Proof. We again reduce a request sequence σ to a relevant request sequence and cut it into *parts* determined by the behaviour of BALANCE. Each part contains some number of cheap faults followed by one expensive fault by BALANCE, except for the last part which might not contain an expensive fault. Let N be the number of parts that contain an expensive fault. The cost of BALANCE consists of two amounts; cost for cheap faults and cost for expensive faults. These are stored in B_C and B_E. It is easy to show that after a page is put into E (i.e. at the start of a new part), we have $B_C \leq B_E < B_C + 1$. Hence the cost of BALANCE is $B_E + B_C \leq 2B_E = 2Nw$. To give a bound on OPT, note that there are $N - 1$ subsequences where BALANCE has a cheap fault followed (immediately) by an expensive fault followed by a cheap fault (of the next part). These 3 faults must be on 3 distinct pages and hence OPT must have at least one fault, or $N - 1$ faults in total. Consequently, the competitive ratio approaches $2w$ as the length of the sequence grows.

To show tightness we consider (integer) even values of w, say $w = 2y, y \in \mathbb{N}$. Let $C = \{0\}$ and $E = \{1\}$ the initial configuration of BALANCE. The sequence consists of N parts, each of which is the sequence $(20)^y 2(10)^y 1$. BALANCE has a fault on every request and pays $4w$ for each part. However, an offline algorithm can keep 0 in E at all times and only have two faults per part. This gives the competitive ratio $2w$. $\square$

4 Lower Bounds for Trackless Algorithms

In this part we prove a lower bound for trackless algorithms which is linear in κw. First we prove the lower bound for the case $\kappa = 1$, and then for general κ.

Theorem 4.1 *For $w \geq 2$, any trackless algorithm for $\kappa = 1$ has a competitive ratio of at least $w/2 + 1$.*

Note that for $w < 2$, we can use the general lower bounds of the previous section. Specifically Theorem 3.2 gives a lower bound of 2 for the complete interval $1 \leq w \leq 2$.

Proof. We construct a sequence in pieces, and bound the ratio for the pieces by presenting an offline algorithm OFF to serve each piece. We have a distinguished page denoted by a, this page is kept in E by OFF, and is never placed into E by ONL. The configurations of the caches are called inverse if the caches of OFF are in the configuration $E = \{a\}$, $C = \{y\}$ and the caches of ONL are in $E = \{y\}$, $C = \{a\}$ for some page y. Each piece starts and ends when the algorithms are in an inverse configuration.

The first inverse configuration can be reached as follows. Denote the page that ONL has in cache C by a and the page that it has in E by b. Then, if this is not already the case, OFF loads a into E and b into C. Thus OFF has a startup-cost of at most $1 + w$. We are now ready to start the first piece.

Suppose the caches are ONL : $C = \{a\}, E = \{b\}$; OFF : $C = \{b\}, E = \{a\}$. Denote the requests in the current piece by a_i ($i = 1, \ldots$). We take $a_1 = c$. We use a case analysis. We will consider pieces of the following forms, where the letter denotes the cache used by ONL for the current page:

1. E

2.1. CE^nC and CE^n for $n \geq 1$

2.2.1. CCE

2.2.2. CC followed by a piece of the form 2.1, 2.2.1 or 2.2.2.

Case 1 c is placed by ONL into E. OFF loads c into C and the piece ends. The configuration is inverse and the cost ratio during this piece is $w \geq w/2 + 1$ for $w \geq 2$.

Case 2 c is placed by ONL into C.

Case 2.1 Request a_2 is placed by ONL into E. Then $a_2 = d$. Denote by ℓ_E the number of pages which are placed into E by ONL after this request. These ℓ_E requests are alternating between b and d, and $a_{\ell_E+3} = a$. (If ONL never puts another page into C, then this is the last piece. By extending it arbitrarily long, the ratio for this piece tends to $w \geq w/2 + 1$.) After request a_{ℓ_E+3} we are again in an inverse configuration. During this piece the cost of OFF is $\ell_E + 2$, and the cost of ONL is $(\ell_E + 1)w + 2$. Therefore the ratio is at least $w/2 + 1$.

Case 2.2 Request a_2 is placed by ONL into C. Then $a_2 = a$.

Case 2.2.1 a_3 is placed into E. Then $a_3 = d$, and the piece is ended. OFF has cost 2, and ONL has a cost of $w + 2$.

Case 2.2.2 a_3 is placed into C. Then $a_3 = c$, and we are at the same configuration as we were after the first request for c (the start of case 2). During this loop OFF has cost 0, and ONL has cost 2. We can therefore ignore this loop and continue as at the start of case 2. If ONL never puts another page in E, then this is the last piece. By extending it arbitrarily long, the ratio for this piece tends to ∞.

Since we considered all the possible cases, and the ratio of the costs is at least $w/2+1$ in all cases, we are done. □

Theorem 4.2 *Any trackless algorithm for $\kappa \geq 2$ has a competitive ratio of at least $\kappa(w+1)/2$.*

Proof. Consider an arbitrary trackless online algorithm. Consider two sets of pages, $A = \{a_1, \ldots, a_{\kappa+1}\}$ and $B = \{b_1, \ldots, b_{\kappa+1}\}$. The sequence is constructed in such a way that the online algorithm always has pages from A in C, and pages from B in E. If the algorithm will place the next page into C, then this request is to the page from A which is not currently in C. Otherwise, the request is to the page from B which is not currently in E.

Consider a long sequence of requests produced by this rule, denote by p the number of online faults in C, and by q the number of online faults in E. Then ONL has a cost of $p + qw$.

To estimate the offline cost for the request sequence, we consider two offline algorithms. The first algorithm OFF$_1$ uses E for $A_1, \ldots, A_\kappa$ and one memory cell from C for $A_{\kappa+1}$. Furthermore it uses the other $\kappa - 1$ cells (called active cells) for serving the requests for B, always evicting the page from B which will be requested the furthest in the future (LFD [4]). The other offline algorithm called OFF$_2$ works in the same way, but with the roles of A and B interchanged. Thus OFF$_2$ has the pages from A continuously in its cache.

Consider first the cost of OFF$_1$. It has at most $\kappa w + 1$ cost on the requests from the set A, and a starting cost of $\kappa - 1$, placing the first $\kappa - 1$ requests for the pages from set B. We can bound the remaining cost in a similar way as it is done in [12].

We partition the sequence of requests to pages in B into subsequences of length κ called κ-parts. Consider a κ-part. Suppose there is a request for a page from B, which is not contained in the $\kappa - 1$ active cells of OFF_1. Consider the first such request in the κ-part. Denote the set of active cells after servicing this request by C', the evicted page by b_i, and the remaining element of $B \setminus (C' \cup \{b_i\})$ by b_j. If during the rest of this κ-part there is a request for b_j, then there will be no other fault in the κ-part: by the LFD rule there is no further request for b_i or for the page which was evicted when b_j was placed into the cache. On the other hand, if there is no request for b_j, then there is no fault during the next $\kappa - 2$ requests (there is no request for b_i).

This yields that OFF_1 has at most 2 faults during a κ-part. Therefore we showed that the cost of OFF_1 is at most $\kappa(w+1) + \lceil 2q/\kappa \rceil$. Similarly, OFF_2 has a cost of at most $\kappa(w+1) + \lceil 2p/\kappa \rceil$. Therefore, we obtained that the optimal offline cost is at most $\kappa(w+1) + \lceil 2\min(p,q)/\kappa \rceil$. By considering arbitrarily long sequences, where we have $p + q \to \infty$, we find that the competitive ratio of any on-line algorithm is at least

$$\frac{p + qw}{2\min(p,q)/\kappa} \geq \frac{(1+w)\kappa}{2},$$

which ends the proof. □

Corollary 4.1 *Any trackless algorithm has a competitive ratio at least half that of any marking algorithm.*

5 Open problems

It would be interesting to find out what is the lowest value of w such that FIFO is no longer best possible. There must be such a value since FIFO (and trackless algorithms in general) cannot be best possible for all w.

Very little is known about server problems in more general spaces. It might be interesting to examine the problem of two servers with a small weight ratio also in other spaces.

Also, the complexity of the offline version of this problem is unknown.

References

[1] A. Aggarwal, B. Alpern, A. K. Chandra, and M. Snir. A model for hierarchical memory. In *Proceedings of the 19th ACM Symposium on the Theory of Computing*, pages 305–313. ACM, 1987.

[2] W. W. Bein, R. Fleischer, L. L. Larmore, Limited bookmark randomized online algorithms for the paging problem, *Information Processing letters*, 76: 155–162, 2000.

[3] W. W. Bein, L. L. Larmore, Trackless online algorithms for the server problem, *Information Processing letters*, 74 no 1-2: 73–79, 2000.

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

[5] Marek Chrobak and John Noga. Competitive algorithms for multilevel caching and relaxed list update. *Journal of Algorithms*, 34(2):282–308, 2000.

[6] M. Chrobak and J. Noga. LRU is better than FIFO. *Algorithmica*, 23:180–185, 1999.

[7] M. Chrobak, J. Sgall, The weighted 2-server Problem, in *Proc. of STACS 2000, LNCS 1770*, Springer-Verlag Berlin, 593–604, 2000.

[8] A. Fiat, M. Ricklin, Competitive algorithms for the weighted server problem, *Theoretical Computer Science*, 130: 85–99, 1994.

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

[10] M. Manasse, L. A. McGeoch, and D. Sleator. Competitive algorithms for server problems. *Journal of Algorithms*, 11:208–230, 1990.

[11] L. Newberg, The K-server problem with distinguishable servers, Master's Thesis, Univ. of California at Berkeley, 1991.

[12] D. Sleator, R. E. Tarjan, Amortized efficiency of list update and paging rules, *Communications of the ACM*, 28:202–208, 1985.

[13] Neal E. Young. On-line file caching. *Algorithmica*, 33(3):371–383, 2002.