Simple and Practical Amortized Sublinear Private Information Retrieval

Muhammad Haris Mughees Sun I Ling Ren

University of Illinois Urbana-Champaign
{mughees, is16, renling}@illinois.edu

***Abstract*—Recent works in amortized sublinear PIR have demonstrated great potential. Despite the inspiring progress, existing schemes in this new paradigm are still faced with various challenges and bottlenecks, including large client storage, high communication, poor practical efficiency, need for non-colluding servers, or restricted client query sequences. We present simple and lightweight amortized sublinear stateful private information retrieval schemes without these drawbacks using new techniques in hint construction and usage. Our scheme can work with two non-colluding servers or a single server. Our schemes achieve close to optimal amortized or online response overhead, which is only two or four times that of simply fetching the desired entry without privacy. Our schemes have practical efficiency. For an 8 GB database with 32-byte entries, each query of our two-server scheme consumes 34 KB of communication and 2.7 milliseconds of computation, and each query of our single-server scheme consumes amortized 47 KB of communication and 4.5 milliseconds of computation. These results are one or more orders of magnitude better than prior works.**

1. Introduction

Private Information Retrieval (PIR) [14] allows a client to fetch an entry from a public database on a server without revealing which entry the client is interested in [14]. An efficient PIR scheme enables many privacy-preserving applications, such as password check [7], safe browsing [24], anonymous communication [32], [8], private media streaming [21].

Despite decades of research [14], [25], [12], [17], [19], [18], [20], [30], [8], [7], [33], [31], PIR protocols are still quite expensive, especially in the single-server setting that does not assume the existence of non-colluding servers. This is due in large part to a well-known fundamental barrier that limits the practical efficiency of conventional PIR schemes: The amount of server computation will inevitably be linear in the size of the database. Intuitively, a PIR scheme must ask the server to touch every single entry in the database; otherwise, the server learns that the entry the client is looking for is not one of the untouched entries.

Several directions have been explored to circumvent this fundamental barrier. A promising and fruitful recent attempt has been the paradigm of *stateful* PIR, first proposed by Patel, Persiano, and Yeo [36]. In this paradigm, the client stores hints (hence called stateful) and uses these hints to speed up queries. The hints, usually in the form of parities of subsets of database entries, need to be retrieved privately. This is done in an offline phase that can be fairly expensive or may even require downloading the entire database. After an expensive offline phase, the client can make many online queries cheaply before having to rerun the offline phase. With two non-colluding servers, the offline phase can even be one-time, meaning the client can make unlimited online queries afterward. This makes the stateful PIR scheme very efficient in an amortized sense after sufficiently many queries. Although the Patel-Persiano-Yeo scheme still incurred linear server computation per query, the stateful PIR paradigm proves promising.

Corrigan-Gibbs and Kogan [16] give the first stateful PIR scheme with amortized *sublinear* server computation. Follow-up works continue to make further improvements and unlock more potential of this paradigm [37], [24], [15], [26], [39], [27], [40]. Despite the inspiring progress, however, existing amortized sublinear stateful PIR schemes are still faced with various challenges, including large client storage, high communication, and subpar practical efficiency. Many schemes also have to resort to heavy-weight theoretical tools [37], [26], [39], parallel repetition [37], [15], [26], [39], or restricted client query sequences [40]. This paper aims to propose simple and practical amortized sublinear stateful PIR schemes without the aforementioned drawbacks.

Overview of existing amortized sublinear stateful PIR. To better explain our techniques and contributions, let us go over a brief overview of the blueprint of amortized sublinear stateful PIR by Corrigan-Gibbs and Kogan [16]. The client privately retrieves hints in an offline phase. Each hint involves a subset S of $\sqrt{N}$ random distinct indices within $[0, N-1]$ where N is the number of entries in the database. For each hint, the client stores the subset S and the corresponding parity $\bigoplus_{i \in S} \mathsf{DB}[i]$ where $\mathsf{DB}[i]$ is the i-th entry of the database, and $\bigoplus$ represents XOR. In the online phase, if the client wants to retrieve the i-th entry, it finds a subset S that contains i. Since the client stores the parity of entries in S, ideally, it just needs to ask the server for the parity of entries in $S \setminus \{i\}$, from which it can easily recover $\mathsf{DB}[i]$.

TABLE 1. Comparison with recent practical amortized sublinear stateful PIR schemes. Request size and client computation are measured in words of size λ or $\log N$. Response size, client storage, and server computation are measured in database entry size (response is hence a blowup over the insecure baseline). Major performance bottlenecks are marked in red.

Scheme	Number of Servers	Amortized communication Request	Amortized communication Response	Storage Client	Amortized computation Client	Amortized computation Server
Corrigan-Gibbs-Kogan [16]	2	$O(\lambda\sqrt{N})$	$O(\lambda)$	$O(\lambda^2\sqrt{N})$	$O(\lambda\sqrt{N})$	$O(\lambda\sqrt{N})$
Kogan-Corrigan-Gibbs [24]	2	$O(\log N)$	$O(1)$	$O(N)$	$O(\sqrt{N})$	$O(\sqrt{N})$
Lazzaretti-Papamathou [27]	2	$O(\log N)$	$O(\sqrt{N})$ [1]	$O(\lambda\sqrt{N})$	$O(\sqrt{N})$	$O(\sqrt{N})$
This paper	2	$O(\sqrt{N})$	$O(1)$	$O(\lambda\sqrt{N})$	$O(\sqrt{N})$	$O(\sqrt{N})$
Corrigan-Gibbs-Henzinger-Kogan [2] [15]	1	$O(\lambda\sqrt{N})$	$O(\lambda)$	$O(\lambda^2\sqrt{N})$	$O(\lambda\sqrt{N})$	$O(\lambda\sqrt{N})$
Zhou et al. [3] [40]	1	$O(\sqrt{N})$	$O(\sqrt{N})$	$O(\lambda\sqrt{N})$	$O(\sqrt{N})$	$O(\sqrt{N})$
This paper	1	$O(\sqrt{N})$	$O(\sqrt{N}/\lambda)$	$O(\lambda\sqrt{N})$	$O(\sqrt{N})$	$O(\sqrt{N})$

[1] Lazzaretti and Papamathou [27] can invoke an extra single-server PIR to reduce the asymptotic response overhead, but a variant without this second PIR gives better practical efficiency and makes a more fair comparison.
[2] Zhou et al. [40] requires client queries to have no adversarial influence, making it weaker than standard PIR.

However, with the above high-level strategy, the client always sends the server a subset that does not contain the queried index i. This is insecure because the server learns that the queried entry is not one of those in $S \setminus \{i\}$. To fix this problem, Corrigan-Gibbs and Kogan suggest that the client occasionally removes an index other than i from S. However, when the client does so, the client loses the ability to retrieve the queried entry i. To compensate for this loss of correctness, λ instances of their protocol are executed in parallel to achieve an exponentially small (in λ) failure probability. This blows up all efficiency metrics (communication, computation, and client storage) by a factor of λ and renders the scheme impractical.

Kogan and Corrigan-Gibbs [24] and Lazzaretti and Papamathou [27] give two ways to avoid this λ factor blowup. Both schemes have notable drawbacks. First, both schemes require two non-colluding servers, and there is no clear way to extend them to single-server stateful PIR. Zhou et al. [40] adapt the Lazzaretti-Papamathou scheme to a single server but can only handle client queries that are not adversarially influenced. While this restriction may be justifiable in certain scenarios, it is not always valid and is a big departure from the standard PIR model.

Second, both schemes make sacrifices on efficiency. The Kogan-Corrigan-Gibbs scheme [24] requires the client to pay either $\Omega(N)$ storage or $\Omega(N)$ computation *per query*, both of which are clearly undesirable. The Lazzaretti-Papamathou scheme [27] incurs a response overhead of $\Theta(\sqrt{N})$ on every query (and Zhou et al. inherit this response overhead). This large response overhead would be prohibitive for databases with large entries. Though this problem could be mitigated by invoking another regular (i.e., not stateful), most likely lattice-based, single-server PIR scheme, that would not be efficient in practice. We will explain this in more detail in Section 5.

Our results. In this paper, we propose new techniques in hint construction and usage and obtain simple and lightweight amortized sublinear stateful PIR schemes. Our new hint system eliminates the aforementioned leakage associated with removing the queried index, thus obviating the need for parallel repetition. We give both two-server and single-server versions of our stateful PIR scheme. The two-server version has a constant amortized response overhead – to be concrete, four times that of simply fetching the desired entry without privacy – while maintaining sublinear client storage and sublinear client computation. The single-server version achieves $O(\sqrt{N}/\lambda)$ amortized response overhead and a constant online response that is twice that of fetching the desired entry without privacy.

Table 1 gives a comparison with recent practical amortized sublinear PIR schemes in terms of asymptotic efficiency. We exclude schemes that rely on heavy theoretical tools, such as those based on oblivious locally decodable codes [13], [11], [28] or privately puncturable/programmable pseudorandom functions [37], [26], [39]. The three major performance bottlenecks in prior works are marked in red: λ factor repetition, $\Theta(\sqrt{N})$ response overhead, and linear client storage. Our schemes avoid all three bottlenecks.

As a result, our scheme enjoys good concrete efficiency. Take for example an 8 GB database consisting of 2^{28} entries where each entry is 32 bytes. Our two-server scheme requires 60 MB of client storage, and consumes 34 KB of communication and 2.7 milliseconds of computation. In comparison, existing two-server schemes require either over 1 GB of client storage or over 1 MB of communication.

For the same database, our single-server scheme requires 100 MB of client storage, and consumes 47 KB of communication and 4.5 milliseconds of computation, amortized per query. In comparison, a state-of-the-art single-server scheme has to weaken correctness and still needs more than $7\times$ of the communication and $5\times$ of the computation than our scheme. The best prior schemes that do not weaken correctness would be at least two orders of magnitude more expensive.

2. Model and Preliminary

Private Information Retrieval (PIR). Given a database DB of N entries and a query index i, the client wants to privately retrieve the i-th entry in the database. A PIR protocol should satisfy the following two properties.

- **Correctness:** If the client and the server correctly execute the protocol, then the client retrieves the queried entry.
- **Privacy:** The server learns *nothing* about the client's query index.

The privacy requirement of PIR can be more rigorously captured by a game between the server, who is also the adversary, and the client. The game resembles the standard message indistinguishability game for encryption.

1) The server picks two indices i and i', and send them to the client.
2) The client flips a coin $b \leftarrow \{0,1\}$. The client queries index i if $b = 0$ and queries index i' if $b = 1$.
3) The server tries to guess b.

If the server can guess b correctly with $0.5 + \epsilon$ probability where ϵ is non-negligible, then the server wins, and the PIR protocol is insecure.

Stateful PIR. We now extend the above PIR definition with a single query to stateful PIR that deals with a sequence of queries.

Given a database DB of N entries and a sequence of query indices $\mathbf{I} = [i_1, i_2, i_3, \ldots]$, the client makes any (polynomial) number of queries one by one, and privately retrieve the i_j-th entry in the database at the end of the j-th query. A stateful PIR protocol should satisfy the following two properties.

- **Correctness:** If the client and the server correctly execute the protocol, then the client retrieves the i_j-th entry in the database at the end of the j-th query.
- **Privacy:** The server learns *nothing* about the client's sequence of query indices.

Similarly, the privacy requirement of stateful PIR can be more rigorously captured by a game between the client and the server.

1) The server picks two sequences of query indices $\mathbf{I}$ and $\mathbf{I}'$ of equal length and send them to the client.
2) The client flips a coin $b \leftarrow \{0,1\}$. The client queries sequence $\mathbf{I}$ if $b = 0$ and queries index $\mathbf{I}'$ if $b = 1$.
3) The server tries to guess b.

If the server can guess b correctly with $0.5 + \epsilon$ probability where ϵ is non-negligible, then the server wins, and the stateful PIR protocol is insecure.

Note that we let the server choose the two sequences of query indices it wants to distinguish, similar to the indistinguishability game for encryption. Likewise, correctness should also hold for any query sequence, including ones that are chosen by the adversary. We could make the server (adversary) even more powerful by letting it choose the query sequences *adaptively*, i.e., it can choose the next pair of query indices *after* interacting with the client for the previous query in the sequence. Likewise, correctness can also be stated for any adaptively constructed sequence. Most existing stateful PIR schemes, including ours, are correct and secure even for adaptively constructed sequences of queries.

Pseudorandom functions. We assume the server is computationally bounded. We will make use of pseudorandom functions (PRF). PRF is one of the most common cryptographic primitives and can be instantiated from any one-way function, including the standardized and widely used AES block cipher and SHA cryptographic hash functions. A PRF takes a secret key and an input. For convenience, we will omit writing the secret key since there should be no confusion in our schemes that the client holds the secret key (and shares the secret key with one of the servers in the two-server setting). The input to the PRF is often a concatenation of multiple values. For example, a PRF call in our algorithms will be written as $\mathsf{PRF}(x \parallel y \parallel z)$.

3. Algorithms

3.1. Overview of the New Hint System

The key idea is a new type of hint that eliminates the information leakage due to the absence of the queried index. This immediately obviates the need for parallel repetition because there will be no (non-negligible) correctness failure. Our techniques can be applied to the original sublinear scheme of Corrigan-Gibbs and Kogan [16] as well as the partition-based hints of Lazzaretti and Papamathou [27]. Because the partition-based hints offer advantages in compact hint storage and fast membership testing, we will describe our techniques on top of the partition-based hints. In this context, our techniques help avoid the large responses and the need for non-colluding servers of the partition paradigm.

A database of size N is divided into $\sqrt{N}$ partitions each of size $\sqrt{N}$. For convenience, we assume $\sqrt{N}$ is an even integer. The database can always be padded to the square of the next even integer with very small extra overhead. Let $\mathcal{R}$ denote the following distribution: first select $\sqrt{N}/2 + 1$ random distinct partitions (i.e., sample without replacement) out of the $\sqrt{N}$ total partitions; then pick one random index from each of these $\sqrt{N}/2 + 1$ partitions. In other words, a sample from $\mathcal{R}$ consists of $\sqrt{N}/2 + 1$ random indices from $\sqrt{N}/2 + 1$ random partitions, one index per partition.

A hint in our algorithm consists of a sample from $\mathcal{R}$ and its corresponding parity. The client needs to store M hints (M will be specified later). For each $j = 0, 1, 2, \ldots, M - 1$, the client samples $S_j \leftarrow \mathcal{R}$ and stores S_j along with $\sum_{i \in S_j} \mathsf{DB}[i]$ as one hint. Usage of the hint also resembles previous works in principle. When the client makes a query to the i-th entry of the database, the client looks for a hint whose subset S_j contains index i. The client sends $S_j \setminus \{i\}$ to the server. The server returns the parity for $S_j \setminus \{i\}$. The client easily recovers $\mathsf{DB}[i]$ since it has been storing the parity for S_j. We need $M = \lambda\sqrt{N}$ where λ is a security parameter so that a subset containing the queried index can be found with all but exponentially small (in λ) probability.

Eliminating the leakage. Now we tackle the main challenge mentioned in Section 1. With the approach described so far, the subset sent by the client involves $\sqrt{N}/2$ random partitions and contains one random index from each of them. However, since the client always removes the queried index, the subset sent by the client will not contain any index from

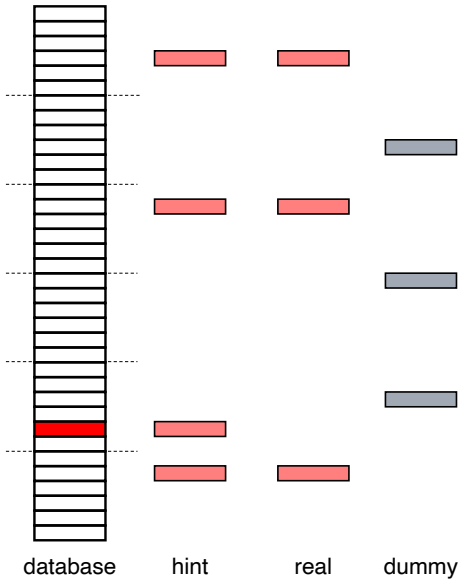

Figure 1. An illustration of the hint system and the client's request. The database has $N = 36$ entries and is divided into $\sqrt{N} = 6$ partitions. Each hint selects $\sqrt{N}/2 + 1 = 4$ random partitions and picks a random index from each. The queried index is removed to produce the real subset. A dummy subset is constructed by picking one random index from each of the remaining three partitions.

the partition that the queried entry belongs to. Thus, the server learns that the queried entry is definitely *not* in any of these $\sqrt{N}/2$ partitions the client sent.

Our main idea to address this leakage is for the client to additionally send a dummy subset that contains one random index from each of the other $\sqrt{N}/2$ partitions. The client also randomly permutes the two subsets, so the server cannot tell apart the real one from the dummy one. This perfectly hides which partition contains the queried entry. In fact, the client's request now reveals no information about the queried entry. The client sends two subsets, each covering $\sqrt{N}/2$ partitions. A random index is picked from each partition, so we only need to ensure that the groupings of the partitions leak no information. To this end, the dummy subset bundles the partition of interest with $\sqrt{N}/2-1$ other random partitions, and the real subset covers the remaining $\sqrt{N}/2$ partitions. This is indistinguishable from a purely random arrangement that would anyway group the partition of interest with $\sqrt{N}/2 - 1$ other random partitions.

The online phase. The online phase of our stateful PIR protocol follows naturally from the above hint system. Upon an input query index i, the client finds a hint whose subset contains the query index i. The client removes i from the subset (this is the real subset). The client then constructs a dummy subset that consists of one random index from each partition not represented in the real subset. The client now sends the two subsets, permuted, to the server. Figure 1 illustrates this process.

The server returns the two parities corresponding to the two subsets. The client discards the dummy parity and uses the real subset parity to recover the desired entry. As a result, the response overhead of our scheme is close to optimal: only twice that of simply fetching the desired entry without privacy.

Hint replenishment and the offline phase. After each query in the online phase, the client needs to replenish one hint since it has just consumed one. Moreover, the replenished hint must follow the same distribution as the one just consumed, i.e., contains index i in the subset. How we carry out the hint replenishment and how we run the offline phase depend on whether we assume a single server or two non-colluding servers. We will describe the two variants later in the section.

3.2. Details of Hints and Online Phase

Sampling a subset of exact size. A step that warrants more clarification is how we sample a subset of size exactly $\sqrt{N}/2 + 1$ out of the $\sqrt{N}$ partitions. For reasons that will become clear later and involve hint replenishment, we will first sample a subset of size exactly $\sqrt{N}/2$ and then supply one extra index.

We start with the first step to sample half of the partitions. We will compute a pseudorandom value using PRF for each hint-partition combination, i.e.,

$$v_{j,k} = \mathsf{PRF}(\text{“select”} \parallel j \parallel k)$$

for the k-th partition of a hint with ID j. The prefix "select" is added because we later need another pseudorandom value for each hint-partition combination. Then, for each hint j, we compute the list of $v_{j,k}$ for all partitions, i.e.,

$$\mathbf{V}_j = [v_{j,0}, v_{j,1}, v_{j,2}, \ldots, v_{j,\sqrt{N}-1}].$$

We then find a cutoff value $\hat{v}_j$ such that exactly $\sqrt{N}/2$ elements in $\mathbf{V}_j$ are smaller than $\hat{v}_j$ and exactly $\sqrt{N}/2$ elements in $\mathbf{V}_j$ are larger than $\hat{v}_j$. A natural choice of $\hat{v}_j$ is the median of $\mathbf{V}_j$. Since we assume $\sqrt{N}$ is an even integer, the median is the average of two elements in $\mathbf{V}_j$. This median $\hat{v}_j$ can be used to divide $\mathbf{V}_j$ into two equal-sized halves. We save this cutoff median value alongside its hint ID for each hint. This will give us an efficient method to check if a partition is selected by the hint, using $O(1)$ time and $O(1)$ client storage per hint.

Next, we need to pick one more index to a hint to make its subset size $\sqrt{N}/2 + 1$. To this end, we need to find a random partition among the $\sqrt{N}/2$ unselected partitions. An easy and effective way to do so is to simply keep picking random partitions and checking if the partition is already selected. Once hitting an unselected partition, a random index is picked from it as the extra index.

Hint storage. Each hint is stored as a tuple $(j, \hat{v}_j, e_j, P_j)$ where j is a unique hint ID, $\hat{v}_j$ is the cutoff median value, e_j is the extra index, and P_j is the parity.

With the hint construction and storage details in place, we can now give more details of the algorithm for the online phase, as is shown in Algorithm 2.

Finding a suitable hint. Upon input query index i, the client computes the partition ℓ that index i belongs to, i.e., $\ell = \lfloor i/\sqrt{N} \rfloor$. The client then goes through the hints to find one whose subset contains i. There are two cases a hint's subset contains i. A straightforward case is that the extra index e_j equals i. The other case is the selection process involving the median cutoff. For each hint j, the client computes $v_{j,\ell}$ and checks if $v_{j,\ell}$ is smaller than $\hat{v}_j$. If so, it means hint j selects partition ℓ, and the client further checks if index i is picked from partition ℓ. This is done by computing a second PRF output as a pseudorandom offset for the partition,

$$r_{j,\ell} = \mathsf{PRF}(\text{“offset”} \parallel j \parallel \ell),$$

and checking if $r_{j,\ell} = i \mod \sqrt{N}$. If two of the two checks above pass, or if the extra index e_j equals i, then index i is included in the subset of hint j.

We remark that this step showcases the benefit of partition-based hints: the partitioning allows us to test in $O(1)$ time whether a hint's subset includes a particular index, as we only need to check the corresponding partition.

Constructing and encoding the two subsets. After finding a hint that contains the query index i, it is straightforward to reconstruct the hint's subset. The client then removes the query index i from the real subset. The client also constructs a dummy subset that contains one random index from each partition that is not in the real subset. Note that a random index (possibly i) will be drawn from the partition of interest. This index will certainly be part of the dummy subset, but the server cannot tell which subset is the dummy one once the client permutes the two subsets.

Although we write our pseudocode to send two subsets following the convention in the literature, we remark that there is an equivalent and more compact way to encode the two subsets. We can use a bit vector $\mathbf{b} = [b_0, b_1, \ldots, b_{\sqrt{N}-1}]$ and an offset vector $\mathbf{r} = [r_0, r_1, \ldots, r_{\sqrt{N}-1}]$. The offset vector encodes which index is picked from each partition. Concretely, $s_k = r_k + k\sqrt{N}$ is the index picked from partition k. The bit vector encodes whether each partition is part of the first or the second subset. In other words, let S_0 and S_1 denote the two subsets of indices that the client would have sent in the pseudocode of Algorithm 2. Then, $s_k \in S_0$ if $b_k = 0$ and $s_k \in S_1$ if $b_k = 1$. We note again that the two subsets are permuted by the client, so the real subset may be either S_0 or S_1, with half-half probability.

It is not hard to see that this encoding is equivalent to sending S_0 and S_1 as done in the pseudocode of Algorithm 2, but is slightly more efficient. Sending S_0 and S_1 directly would cost $\sqrt{N} \log N$ bits. The encoding using $\mathbf{b}$ and $\mathbf{r}$ costs $\sqrt{N} + \sqrt{N} \log \sqrt{N} = (\sqrt{N}/2 + 1) \log N$ bits, roughly reducing the client's request size by half.

3.3. The Two-Server Scheme

When there are two non-colluding servers, we use one server for the offline phase and hint replenishment, and use the other server for the online queries. The offline phase only needs to run once at the beginning of the entire protocol to help the client start with sufficiently many hints. After that, the client invokes hint replenishment on the fly, i.e., during each online query. Pseudocode of the complete two-server stateful PIR protocol is given in Algorithms 1, 2 and 3.

The offline phase. The offline phase is shown in Algorithm 1 and is fairly straightforward. The client initiates the offline phase by sending the offline server its PRF evaluation key. This allows the offline server to fully construct the hints. For each hint, exactly half of the partitions are selected using the cutoff median method described in Section 3.2, and a random index is picked from each selected partition. After that, an extra index is picked from a partition that has not been selected yet. The offline server can easily compute the parity of these entries. Lastly, the offline server sends to the client the cutoff, the extra index, and the parity for each hint. The client stores all these hints. This completes the offline phase.

Hint replenishment. To replenish a hint after querying index i, the client asks the offline server to start Algorithm 3. Since the offline server has the PRF evaluation key, it can construct a new hint using the next available hint ID, similar to what it did in the offline phase. But there are two new catches. First, the offline server does not add the extra index because the client would like to add index i, the index that is just queried, as the extra index, to make sure the replenished hint follows the same distribution as the consumed one, i.e., has i in the subset. Second, we do not want the offline server to learn the new hint's subset, because that would reveal some information to the offline server about the query the client just made. Therefore, we let the offline server compute the parities for both halves, and send both parities to the client, along with the new hint ID and median cutoff.

Upon receiving the above from the offline server, the client chooses the half that does *not* select the partition of i as the real half. To do so, the client may have to redefine the operator $<$ for this hint. In other words, the client stores a bit that indicates whether this hint chooses all partitions whose pseudorandom values ($v_{J,k}$) are smaller or larger than the median cutoff. This essentially permutes the two halves and makes them indistinguishable to the offline server. The client then adds index i to the hint's subset as the extra index and adds $\mathsf{DB}[i]$ (which the client has just retrieved) to the parity. The new hint is now fully constructed and replaces the consumed hint.

3.4. The Single-Server Scheme

Hint replenishment using backup hints. With a single server, we no longer have the luxury of replenishing a hint on the fly. Instead, we will use the idea of backup hints from [15]. The client retrieves additional backup hints in the offline phase, so that the client can replenish a hint during the online phase without contacting the server. Since backup hints will eventually run out, the offline phase needs to be run periodically. Pseudocode of our complete single-server stateful PIR protocol is given in Algorithms 4, 2 and 5.

Algorithm 1 The offline algorithm with two non-colluding servers, run by the offline server

for $j = 0, 1, 2, \ldots, M-1$ **do**
  Initializes parity $P_j = 0$
  Compute $\mathbf{V}_j = [v_{j,0}, v_{j,1}, v_{j,2}, \ldots, v_{j,\sqrt{N}}]$ where $v_{j,k} = \mathsf{PRF}(\text{“select”} \parallel j \parallel k)$
  Find the median $\hat{v}_j$ of $\mathbf{V}_j$ as the cutoff for selection
  $S = \{k \mid v_{j,k} < \hat{v}_j\}$ ▷ the set of partitions selected by this hint
  $P_j = \bigoplus_{k \in S} \mathsf{DB}[r_{j,k} + k\sqrt{N}]$ where $r_{j,k} = \mathsf{PRF}(\text{“offset”} \parallel j \parallel k)$ ▷ one random index per selected partition
  Set the extra index e_j to a random index from a random partition not in S
  $P_j = P_j \oplus \mathsf{DB}[e_j]$
  Send $(j, \hat{v}_j, e_j, P_j)$ to the client to be stored
end for
Set $J = M$, the next available hint ID ▷ J will be strictly increasing

Algorithm 2 The online algorithm, run by the client

Input: queried index i ▷ $v_{j,k}, r_{j,k}, h_j, \hat{v}_j, e_j, P_j$ as defined in Algorithm 1 or 4
$\ell = \lfloor i/\sqrt{N} \rfloor$ ▷ ℓ is the partition that i belongs to
Find main hint j such that $v_{j,\ell} < \hat{v}_j$ and $r_{j,\ell} == i \mod \sqrt{N}$ ▷ hint j contains i
Initialize $S = \emptyset$ and $S' = \emptyset$ ▷ S will be the real subset and S' will be the dummy subset
for $k = 0 : \sqrt{N} - 1$ **do**
  if $v_{j,k} < \hat{v}_j$ **then**
    $S = S \cup \{r_{j,k} + k\sqrt{N}\}$
  else if e_j belongs to partition k **then**
    $S = S \cup \{e_j\}$
  else
    $S' = S' \cup \{\mathsf{rand}() + k\sqrt{N}\}$ ▷ add a random index from partition k to the dummy subset
  end if
end for
$S = S \setminus \{i\}$ ▷ remove the queried index from the real subset
$S' = S' \cup \{\mathsf{rand}() + \ell\sqrt{N}\}$ ▷ add a random index from partition ℓ to the dummy subset
Send (S, S') or (S', S) to the server with half-half probability ▷ permute the real and dummy subsets
Receive the two subset parities P and P' from the server ▷ in the order S and S' are sent
Return $P \oplus P_j$ as $\mathsf{DB}[i]$
Replenish a hint that contains index i from partition ℓ using Algorithm 3 or 5

Algorithm 3 The hint replenish algorithm with two non-colluding servers, run by the offline server and the client

Use the next available hint ID J ▷ The client asks the offline server to start hint replenishment
Initializes parity $P_J = P'_J = 0$
Compute $\mathbf{V}_J = [v_{J,0}, v_{J,1}, v_{J,2}, \ldots, v_{J,\sqrt{N}-1}]$
Find the median $\hat{v}_J$ of $\mathbf{V}_J$
$S = \{k \mid v_{J,k} < \hat{v}_J\}$
$P_J = \bigoplus_{k \in S} \mathsf{DB}[r_{J,k} + k\sqrt{N}]$ ▷ recall $r_{j,k} = \mathsf{PRF}(\text{“offset”} \parallel j \parallel k)$
$P'_J = \bigoplus_{k \notin S} \mathsf{DB}[r_{J,k} + k\sqrt{N}]$
Send $J, \hat{v}_J, P_J, P'_J$ to the client ▷ The rest of the algorithm is run by the client
if $v_{J,\ell} > \hat{v}_J$ **then** ▷ pick the half that does not select partition ℓ
  $P_J = P'_J$
  Set a bit to redefine $<$ to be “greater than” for this hint ▷ Algorithm 2 should check this bit and interpret $<$ accordingly for each hint, but we omitted these details in Algorithm 2 for readability of the pseudocode
end if
Replace hint j with new hint $(J, \hat{v}_J, i, P_J \oplus \mathsf{DB}[i])$ ▷ add i as the extra index to the new hint J

Algorithm 4 The streaming offline algorithm with a single server, run by the client

for $j = 0, 1, 2, \ldots, 1.5M - 1$ **do** ▷ M main hints and $0.5M$ pairs of backup hints
  Initialize parity $P_j = 0$, and additionally initialize $P'_j = 0$ if $j \geq M$ ▷ backup hints come in pairs
  Compute $\mathbf{V}_j = [v_{j,0}, v_{j,1}, v_{j,2}, \ldots, v_{j,\sqrt{N}-1}]$ where $v_{j,k} = \mathsf{PRF}(\text{“select”} \parallel j \parallel k)$
  Find and store the median $\hat{v}_j$ of $\mathbf{V}_j$ as the cutoff for partition selection
  if $j < M$ **then** ▷ main hints
    Set the extra index e_j to a random index from a random partition not in $\{k \mid v_{j,k} < \hat{v}_j\}$
  end if
end for
for $k = 0 : \sqrt{N} - 1$ **do**
  Download $\mathsf{DB}[k\sqrt{N} : (k+1)\sqrt{N} - 1]$ from the server ▷ download partition k
  for $j = 0, 1, 2, \ldots, 1.5M - 1$ **do**
    $x = \mathsf{DB}[r_{j,k} + k\sqrt{N}]$ where $r_{j,k} = \mathsf{PRF}(\text{“offset”} \parallel j \parallel k)$ ▷ a pseudorandom entry is picked from partition k
    if $v_{j,k} < \hat{v}_j$ **then** ▷ partition k is selected by hint j
      $P_j = P_j \oplus x$
    else if $j \geq M$ **then**
      $P'_j = P'_j \oplus x$ ▷ also construct the backup hint in the pair
    end if
    if $\lfloor e_j / \sqrt{N} \rfloor == k$ **then** ▷ the extra index e_j is in partition k
      $P_j = P_j \oplus \mathsf{DB}[e_j]$
    end if
  end for
end for

Algorithm 5 The hint replenish algorithm with a single server, run by the client

Let J be the ID of the next unused pair of backup hints
if $v_{J,\ell} > \hat{v}_J$ **then** ▷ pick the half that does not select partition ℓ
  $P_J = P'_J$
  Set a bit to redefine $<$ to be “greater than” for this hint ▷ Algorithm 2 checks this bit to interpret $<$
end if
$h_j = J$
$e_j = i$
$P_j = P_J \oplus \mathsf{DB}[i]$
Replace hint j with backup hint $(J, \hat{v}_J, i, P_J \oplus \mathsf{DB}[i])$ ▷ add i as the extra index to the new main hint J

In the offline phase, the client retrieves not only the $\lambda\sqrt{N}$ primary hints but also $\lambda\sqrt{N}$ backup hints. A backup hint does not have the extra index and thus contains one fewer index in its subset than a main hint. After the client makes a PIR query for index i, it finds a backup hint that does not select i's partition. The client then adds index i to the subset as the extra index and adds $\mathsf{DB}[i]$ to the parity. The new subset and parity now form a regular main hint that follows the same distribution as the consumed one, i.e., has i in the subset.

A simple strategy is to have $\lambda\sqrt{N}$ independent backup hints. Then, there are in expectation $0.5\lambda\sqrt{N}$ backup hints that skip any given partition. So the client can make close to, but fewer than, $0.5\lambda\sqrt{N}$ (say $0.4\lambda\sqrt{N}$) online queries before having to run the offline phase again. Even if the client keeps querying entries from the same partition, it will not run out of backup hints that skip that partition, except for exponentially small (in $\lambda\sqrt{N}$) probability.

A more clever strategy is to have backup hints in pairs, similar in spirit to the two-server hint replenishment algorithm. This is the strategy taken in the pseudocode of Algorithm 5. From a backup hint ID J, the client computes $\mathbf{V}_J$ as well as the cutoff $\hat{v}_J$. The cutoff $\hat{v}_J$ divides the partitions into two equal-sized halves. The client will store the parities corresponding to both halves. When it is time to replenish a hint that contains index i, the client picks the half that does *not* select the partition ℓ that index i belongs to, and then adds i as the extra index. Similar to the two-server scheme, the client needs to store a bit indicating whether $<$ is redefined to be “greater than” for this hint. This way, the client only needs to use one pair of backup hints per query, as one of the two halves will definitely suffice for a replenishment. The client can now store $\lambda\sqrt{N}/2$ pairs of backup hints, and can make exactly $\lambda\sqrt{N}/2$ online queries before having to run the offline phase again.

Offline phase. In the offline phase, the client needs to retrieve main hints and backup hints in a private manner. This can be done in a few ways. The simplest and most practical way is perhaps to stream the entire database, one partition at a time. The pseudocode of the streaming offline phase is given in Algorithms 4. The extra index of each main hint can be sampled in the same way as described in

Section 3.2: keep picking a random partition and checking if it is already selected. This is now done by the client prior to streaming the database. After downloading a partition, it is straightforward to use $v_{j,k}$ and $r_{j,k}$ to determine, for each main or backup hint j, which index, if any, should be drawn from the current partition k. For each main hint, the client also checks if its extra index is from the current partition. For each backup hint pair, the client updates the parity corresponding to the correct half, based on whether $v_{j,k}$ is smaller or larger than the median cutoff.

3.5. Correctness and Privacy Analysis

We will focus first on the very first query after the offline phase and then extend the analysis to subsequent queries.

Correctness. For correctness, we need to prove that, upon an input query index i, the client will be able to find with overwhelming probability a hint whose subset includes i. To this end, we first observe the following simple fact.

Lemma 1. *Each hint in our construction has at least* $\frac{1}{2\sqrt{N}}$ *probability of containing a particular index.*

Proof. For a hint to contain a particular index i, the hint must select the partition i belongs to and also pick i from that partition. The former happens with $(\sqrt{N}/2+1)/\sqrt{N} > 1/2$ probability (the plus one is due to the extra index), and the latter happens with $1/\sqrt{N}$ probability. □

For correctness to be violated, none of the $\lambda\sqrt{N}$ main hints contains the query index. This happens with less than $(1-\frac{1}{2\sqrt{N}})^{\lambda\sqrt{N}} < e^{-\lambda/2}$ probability.

Privacy. We need to prove that the two subsets sent by the client reveal no information about the query index. We will carry out the proof assuming the PRF is perfectly random. The privacy of our PIR protocol is then reduced to the pseudorandomness of the PRF.

It is more convenient to reason about privacy with the more compact encoding described in Section 3.2. Recall that the client sends a bit vector $\mathbf{b}$ grouping partitions into two subsets along with an offset vector $\mathbf{r}$ encoding the index picked from each partition. First, observe that the offset vector $\mathbf{r}$ consists of pseudorandom values that are independent of the query index.

- For partitions not selected by the hint, a fresh pseudorandom dummy offset is used (Line 11 of Algorithm 2).
- For the partition that contains the query index i, i is removed and is replaced with a fresh pseudorandom dummy offset (Line 15 of Algorithm 2).
- For the remaining partitions that are selected by the hint, the offsets are picked pseudorandomly during the offline phase, and this is the first (and only) time they are revealed to the (online) server.

Thus, from the (online) server's perspective, all $\sqrt{N}$ offsets are fresh, pseudorandom, and independent of the query index.

The crux of the proof is to show that the bit vector $\mathbf{b}$ reveals no information about the query index. Formally, we will prove that the distribution of $\mathbf{b}$ is not affected by, and hence reveals no information about, the query index.

Lemma 2. *For any two query indices* i *and* i', $\Pr(\mathbf{b} \mid i) = \Pr(\mathbf{b} \mid i')$.

Proof. Let ℓ denote the partition index i belongs to and ℓ' denote the partition index i' belongs to. When the query index is i, an index from partition ℓ is added to the dummy subset. For the client to send $\mathbf{b}$, two events must happen. First, the bit b_ℓ represents the dummy subset (as opposed to the opposite bit $1-b_\ell$). This happens with $1/2$ probability. Second, besides partition ℓ, the set of partitions selected by this hint are those marked by the opposite bit, i.e., $T = \{k \mid b_k \neq b_\ell\}$. Since each hint selects $\sqrt{N}/2+1$ partitions at random, the probability for the other $\sqrt{N}/2$ selected partitions to happen to be those in T is $\tau = \binom{\sqrt{N}-1}{\sqrt{N}/2}^{-1}$. These two events are independent, so $\Pr(\mathbf{b} \mid i) = \tau/2$. By the exact same argument, we have

$$\Pr(\mathbf{b} \mid i') = \tau/2 = \Pr(\mathbf{b} \mid i). \qquad \square$$

Lemma 2 is sufficient to establish the privacy of our protocol. But to make things more explicit, we can derive the following simple facts from Lemma 2.

$$\Pr(\mathbf{b}) = \sum_i \Pr(\mathbf{b} \mid i) \cdot \Pr(i) = \tau/2 \cdot \sum_i \Pr(i) = \tau/2.$$

Thus, for any query index i,

$$\Pr(i \mid \mathbf{b}) = \frac{\Pr(i, \mathbf{b})}{\Pr(\mathbf{b})} = \frac{\Pr(\mathbf{b} \mid i) \cdot \Pr(i)}{\Pr(\mathbf{b})} = \Pr(i).$$

The fact that $\Pr(i \mid \mathbf{b}) = \Pr(i)$ for all i means that observing $\mathbf{b}$ does not change an observer's prior on the query index; in other words, $\mathbf{b}$ does not reveal any information about the query index. Therefore, the server will have no advantage in distinguishing the two queries in the privacy game.

Extension to subsequent queries. The above completes the correctness and privacy proofs for the first query after the offline phase. It remains to extend the proofs to subsequent queries. For this step, we need to show that after a query consumes and replenishes a hint, the distribution of the main hints remains the same. Then, our privacy proof above would apply directly to all subsequent queries, and the correctness failure probability over a sequence of queries can be upper bounded by a simple union bound.

Let $H_{j,k}$ be the random variable that represents the index from partition k selected by hint j. If hint j does not select from partition k, $H_{j,k} = \perp$. Then, the main hints are fully described by the following matrix of random variables.

$$\mathbf{H} = \begin{bmatrix} \overrightarrow{H_0} \\ \overrightarrow{H_1} \\ \vdots \\ \overrightarrow{H_{M-1}} \end{bmatrix} = \begin{bmatrix} H_{0,0} & H_{0,1} & \cdots & H_{0,\sqrt{N}-1} \\ H_{1,0} & H_{1,1} & \cdots & H_{1,\sqrt{N}-1} \\ \vdots & \vdots & \ddots & \vdots \\ H_{M-1,0} & H_{M-1,1} & \cdots & H_{M-1,\sqrt{N}-1} \end{bmatrix}$$

Let $\mathbf{H}$ represent the main hints before the current query and let $\mathbf{H}'$ represent the main hints after the current query. We want to show that $\mathbf{H}'$ and $\mathbf{H}$ are identically distributed.

Each hint (row vector) in $\mathbf{H}$ is drawn from the distribution $\mathcal{R}$ described in Section 3.1. Let $\mathcal{R}_i$ be the distribution of a hint *conditioned on* the event that it contains index i. Let $\mathcal{R}_{-i}$ be the distribution of a hint *conditioned on* the event that it *does not* contains index i.

Suppose we scan the main hints from 0 to $M-1$ to look for the query index i. Each hint independently has a probability $q = \frac{\sqrt{N}/2+1}{\sqrt{N}} \cdot \frac{1}{\sqrt{N}}$ to contain i: partition ℓ needs to be selected and i needs to be picked from partition ℓ. Let J be the hint consumed. J follows a geometric distribution with parameter q. (The event that no hint contains i is a negligible one, and for convenience we can assume no hint is consumed or replenished in that case.) We thus have

$$\Pr(J > j) = (1-q)^{j+1},$$
$$\Pr(J = j) = (1-q)^j q,$$
$$\Pr(J < j) = \sum_{l=0}^{j-1} (1-q)^l q = 1 - (1-q)^j.$$

Both the consumed and the replenished hint J follow distribution $\mathcal{R}_i$. All the other hints are unmodified. Moreover, all the hints prior to J follow distribution $\mathcal{R}_{-i}$, and all the hints after J follow distribution $\mathcal{R}$.

Let us now focus on any particular hint j in $\mathbf{H}'$. Given the distribution of J, we can think of hint j in $\mathbf{H}'$ to be sampled in the following manner: with $1-(1-q)^j$ probability, sample from $\mathcal{R}$; for the remaining $(1-q)^j$ probability, sample from $\mathcal{R}_i$ with probability q and sample from $\mathcal{R}_{-i}$ with probability $(1-q)^{j+1}$.

Observe that the q vs. $1-q$ ratio is exactly the likelihood that an original hint in $\mathbf{H}$ does vs. does not contain index i, or equivalently, follows $\mathcal{R}_i$ vs. $\mathcal{R}_{-i}$. Thus, every hint j in $\mathbf{H}'$ follows the same distribution as the hint j in $\mathbf{H}$. This shows that the main hints after a query are identically distributed as they are before the query. Then, by transitivity, the main hints at any point are identically distributed as their original states right after the offline phase. Therefore, our correctness and privacy proofs apply to all subsequent queries.

3.6. Efficiency Analysis

The two-server scheme. The offline phase costs $O(\lambda\sqrt{N})$ communication and $O(\lambda N)$ computation at the offline server. But because the offline phase runs only once, these costs do not factor into the amortized costs after sufficiently many queries are made. Hence, the amortized cost of our two-server scheme only depends on the online phase and the hint replenishment step. The online request size is $(\sqrt{N}/2+1)\log N$ bits using the compact encoding of the two subsets. The online response overhead is $O(1)$, or $4\times$ to be precise since the online server and the offline server both send back two parities.

The expected client computation cost of the client is $O(\sqrt{N})$ due to searching for a hint and reconstructing the hint's subset. Because each hint has at least $\frac{1}{2\sqrt{N}}$ probability of containing a particular index by Lemma 1, the client will find a suitable hint after checking $2\sqrt{N}$ hints in expectation (and each check takes $O(1)$ time). The computation cost of the server is $O(\sqrt{N})$ due to computing the parities. These give the results in Table 1.

The single-server scheme. The online phase is very similar to the two-server scheme: each online query costs $O(\sqrt{N})$ bits in request, $O(1)$ overhead in response, $O(\sqrt{N})$ client computation, and $O(\sqrt{N})$ server computation. The streaming offline phase costs N communication and $O(\lambda N)$ computation, and needs to be run every $0.5\lambda\sqrt{N}$ online queries. This leads to the single-server results in Table 1. The only difference from the two-server case is that the response overhead is $O(\sqrt{N}/\lambda)$, because the $O(N)$ offline communication is amortized over $0.5\lambda\sqrt{N}$ online queries.

4. Evaluation

4.1. Implementation Details

We implemented our scheme in C++. Due to the simplicity of our schemes, the two-server version of our implementation comprises about 600 lines of code and the single-server version comprises about 500 lines of code. We set the parameter λ to 80. We use AES as the pseudorandom function. We use CryptoPP's implementation of AES, which leverages Intel's AES-NI instructions. We break up a single 128-bit AES output into four to eight pseudorandom numbers (i.e., $v_{j,k}$ and $r_{j,k}$ in the algorithms) across different hints or partitions to save computation.

We use 32-bit numbers for elements in $\mathbf{V}_j$ to save client storage and computation. It is worth noting that this gives rise to a corner case where two or more elements in $\mathbf{V}_j$ are equal to the median. (Equal elements occur with negligible probability if we use a full 128-bit PRF output for each element in $\mathbf{V}_j$.) When this happens, the median alone does not give a way to evenly divide $\mathbf{V}_j$ into two equal-sized halves. We could add additional metadata to handle this corner case, but because this corner case happens with a very small probability, we simply consider such a hint invalid and discard it. We have omitted the handling of this corner case from the pseudocode for code readability.

The concrete implementation of the median finding procedure also warrants more explanation. We take advantage of the fact that elements of $\mathbf{V}_j$ are drawn from a uniform random distribution. This allows us to filter out elements that are too large or too small, i.e., outside two heuristic bounds. We keep count of the number of filtered elements. Suppose we filter out X small elements. We then use introselect [35] or a similar linear time selection algorithm to find the $(\sqrt{N}/2-X-1)$-th and $(\sqrt{N}/2-X)$-th smallest elements among the remaining elements. These will be the two middle elements that give the median of $\mathbf{V}_j$. With appropriate bounds, the probability that we filter out one of these two elements is very small. (And when that happens, we simply consider this hint invalid and discard it.) We

think of the random values as 32-bit fixed-point numbers between 0 and 1, and choose the two filtering bounds as $\frac{1}{2} \pm \frac{1}{16}$. In expectation, this filters out $7/8$ of the elements. The probability that one of the middle elements is filtered out is 6×10^{-5} for a database of size 2^{20}, and this probability keeps decreasing with the size of the database.

When $\log N$ does not exceed 32, we use 32-bit integers for the extra indices. The hint IDs in our single-server version can also use 32-bit numbers since they will reset periodically upon offline phases. In the two-server version, however, hint IDs can grow unbounded, so we use 64-bit integers for them.

4.2. Experimental Setup

Baselines. We compare with several practical two-server and single-server schemes. We briefly describe each baseline with its pros and cons below. Two-server baselines include:

- The protocol of Boyle, Gilboa, and Ishai [10] based on distributed point functions (DPF). This is the state-of-the-art two-server PIR scheme that uses linear server computation. It has a logarithmic request size and a constant response overhead. We use their C++ implementation [2].
- The protocol of Kogan and Corrigan-Gibbs for checklists [24]. This is the first two-server amortized sublinear PIR scheme that has been implemented. Their scheme has a logarithmic request size and a constant response overhead but requires either linear client computation or linear client storage. Their implementation in Go [1] uses linear client storage.
- TreePIR by Lazzaretti and Papamathou [27]. This is the state-of-the-art two-server amortized sublinear PIR scheme. Their scheme uses sublinear client storage and client computation and has a logarithmic request size. The downside of their scheme is the $O(\sqrt{N})$ response overhead. We use their implementation in Go [6].

Single-server baseline schemes include:

- Spiral PIR by Menon and Wu [31]. This is the state-of-the-art single-server single-query PIR. It is based on lattice-based leveled FHE and needs to perform a linear amount of homomorphic operations at the server. We use their C++ implementation [5].
- SimplePIR by Henzinger et al. [22]. This is a single-server stateful PIR scheme that still uses linear server computation. We use their implementation in Go [4].
- Piano PIR by Zhou et al. [40]. This is a single-server variant of TreePIR and inherits the $O(\sqrt{N})$ response overhead of TreePIR. It requires the client's queries to have no adversarial influence and thus is weaker than a standard (stateful) PIR scheme defined in Section 2. We use their implementation in Go [3].

Experimental setup. We run all experiments on an AWS m5.8xlarge instance equipped with a 3.1 GHz Intel Xeon processor and 128 GB RAM. Our instance runs Ubuntu 22.04, GCC 11.3, and GO 1.18. We run our scheme and all baselines with a single thread. We analyze the performance of our scheme and the baseline schemes under databases with varying entry counts and entry sizes. We first test databases with 2^{20}, 2^{24}, and 2^{28} entries while fixing the entry size to 32 bytes. We then fix the database to 2^{28} entries and test different entry sizes, specifically 8 and 256 bytes.

4.3. Evaluation Results

Two-server schemes. Table 2 gives a performance comparison of two-server PIR and stateful PIR schemes. The offline phases of the three stateful PIR schemes are run only once, so their amortized per-query costs are simply the online costs after sufficiently many queries are made. The checklist implementation crashed in our last experiment, so their results for the 64 GB database are missing. The DPF implementation does not support 256-byte entries, so its computation result for the 64 GB database is estimated.

DPF-PIR requires no offline phase or client storage. It also has efficient communication ranging from 0.91 KB to 1.52 KB in our tests. Its computation is linear in database size, and grows from 2.5 ms on a 32 MB database to 5960 ms on a 64 GB database. This makes the DPF-PIR very efficient in all aspects for small databases but costly in computation for large databases. In comparison, the three stateful PIR schemes require offline phases and client storage, and in return, achieve orders of magnitude lower per-query computation.

The checklist scheme boasts the lowest communication cost among the schemes we test. It also has a low online computation cost that is comparable to our scheme. Its biggest downside is the linear client storage. This cost is manageable for small databases but becomes prohibitive for large databases. For example, on the 8 GB database, Checklist's client storage is over 1 GB, about one-eighth of the entire database and $\geq 20\times$ of TreePIR and our scheme.

TreePIR requires the smallest client storage among the three but has a large per-query communication cost that is two orders of magnitude larger than our scheme. Its per-query computation is also around $3.8 - 12.8\times$ slower than our scheme. We also test TreePIR with an extra single-server PIR call (not shown in the table). Its communication would improve to around 30 KB but its computation would worsen to hundreds of milliseconds (refer to Spiral result in Table 3 and the discussion in Section 5).

Overall, our scheme achieves a balance of low client storage, low communication, and low computation for all database parameters, by avoiding major bottlenecks in previous schemes such as linear client storage, linear server computation, or high communication.

Single-server schemes. Table 3 gives a performance comparison of single-server PIR and stateful PIR schemes. The amortized per-query cost of Piano and our scheme are calculated as the offline cost divided by the number of queries supported per offline, plus the online cost. Spiral has no offline phase and SimplePIR has a one-time offline phase, so their amortized per-query costs are simply the online

TABLE 2. Comparison of two-servers PIR schemes.

	Database Parameters	Client Storage (MB)	Offline Comm. (MB)	Offline Compute (s)	Online Comm. (KB)	Online Compute (ms)
DPF-PIR		-	-	-	0.91	2.5
Checklist	2^{20} 32-byte entries	7.07	2.88	3.3	0.50	0.17
TreePIR	32 MB in total	2.88	2.88	1.0	65.9	0.45
This paper		3.76	3.76	2.3	2.26	0.12
DPF-PIR		-	-	-	1.1	47
Checklist	2^{24} 32-byte entries	78.60	11.53	73	0.56	0.72
TreePIR	512 MB in total	11.53	11.53	23	262.6	4.9
This paper		15.04	15.04	41	8.64	0.54
DPF-PIR		-	-	-	1.21	182.4
Checklist	2^{28} 8-byte entries	1085.27	11.53	1394	0.52	1.9
TreePIR	2 GB in total	11.53	11.53	398	262.6	20
This paper		30.16	30.16	636	34.0	2.19
DPF-PIR		-	-	-	1.31	745
Checklist	2^{28} 32-byte entries	1119.74	46.14	1141	0.64	1.8
TreePIR	8 GB in total	46.14	46.14	430	1049.6	14
This paper		60.16	60.16	842	34.1	2.7
DPF-PIR	2^{28} 256-byte entries	-	-	-	1.52	5960
TreePIR	64 GB in total	369.09	369.09	1843	8389.6	67
This paper		340.16	340.16	2242	35.0	5.23

TABLE 3. Comparison of the single-server schemes.

	Database Parameters	Client Storage (MB)	Offline Comm. (MB)	Offline Compute (s)	Online Comm. (KB)	Online Compute (ms)	Amortized per query Comm. (KB)	Amortized per query Compute (ms)
Spiral		-	-	-	28	767	28	767
SimplePIR	2^{20} 32-byte entries	20.9	20.9	4.8	40	14	40	14
‘ Piano	32 MB in total	6.64	32	10	20	0.79	22.30	1.5
This paper		6.25	32	4	2.18	0.14	2.99	0.25
Spiral		-	-	-	34.0	3177	34.0	3177
SimplePIR	2^{24} 32-byte entries	86.8	86.6	154	168	103	168	103
Piano	512 MB in total	29.28	512	182	80	2.6	87.69	5.3
This paper		25	512	65	8.56	0.62	11.76	1.0
Spiral		-	-	-	34.5	8427	34.5	8427
SimplePIR	2^{28} 8-byte entries	173.4	173.4	623	338	319.1	338	319.1
Piano	2 GB in total	73.875	2048	3186	128	8.5	134.6	18
This paper		40	2048	989	34.02	2.4	37.22	3.9
Spiral		-	-	-	35.0	30273	35.0	30273
SimplePIR	2^{28} 32-byte entries	352.98	352.98	failed	688	1123	688	1123
Piano	8 GB in total	126.75	8192	3449	320	10	346.38	21
This paper		100	8192	1146	34.06	2.7	46.86	4.5
Piano	2^{28} 256-byte entries	620.25	65536	4495	2112	24	2323.05	38
This paper	64 GB in total	660	65536	2743	34.5	4.2	136.9	8.4

costs after sufficiently many queries are made. Spiral and SimplePIR crashed in our last experiment, so their results for the 64 GB database are missing. The SimplePIR implementation offline phase also failed for the 8 GB database; luckily, their implementation provides a way to test online efficiency without running the offline phase (and naturally, without correctness).

Spiral's communication cost remains relatively stable at different database parameters. Its linear server computation, however, is expensive even for small databases and becomes prohibitive for large databases. Concretely, its per-query computation is over 3 seconds for a 512 MB database and over 30 seconds for an 8 GB database. In comparison, our scheme is thousands of times faster than Spiral in per-query computation, e.g., just 4.5 milliseconds on the same 8 GB database. In terms of per-query communication, our scheme is better than Spiral on small databases but becomes worse on large databases due to the $\Omega(\sqrt{N})$ request size.

SimplePIR's server online computation is a constant factor better than Spiral's, but it is still linear and still very expensive for large databases. Piano, being a sublinear scheme, addresses the server computation bottleneck, but it has to weaken the correctness guarantee of PIR. SimplePIR and Piano also introduce a new bottleneck in communication

since they both have $\Omega(\sqrt{N})$ online response overhead. Concretely, for an 8 GB database, the per-query communication cost is 688 KB for SimplePIR and 346 KB for Piano.

Our scheme achieves much better communication and computation than the other two stateful PIR schemes. Concretely, compared with Piano, our amortized communication is $3.6 - 17\times$ better, our amortized computation is $4.5 - 6\times$ better, and we achieve these improvements while providing a stronger correctness guarantee. Compared with SimplePIR, the state-of-the-art scheme that provides the same standard PIR correctness, our scheme is $9 - 14\times$ better in communication and hundreds of times faster in computation.

Our single-server scheme does have a drawback (shared by Piano): offline communication is very high for large databases due to streaming the whole database. Even though this can be amortized over many online queries, it is still undesirable as it significantly delays the very first query.

5. Related Works

Private Information Retrieval (PIR) is first introduced by Chor et al. [14]. There is an extensive list of works on both multi-server PIR and single-server PIR. Since this work focuses on the two-server and single-server settings, we will focus on these two settings in this section and omit schemes that require three or more servers.

Single-query PIR with linear server computation. Research on PIR started with the simplest and most standard variant: a client has a *single* entry to fetch from the server. We call it single-query PIR. Chor et al. [14] gives the first single-query PIR scheme. Their scheme uses multiple non-colluding servers and provides information-theoretic security. With two servers, the communication cost of their scheme is $O(N^{1/3})$. Subsequent works have improved the communication cost of two-server PIR with both information-theoretic security [18] and computational security [20]. In particular, Gilboa and Ishai give a two-server computational PIR scheme based on *distributed point functions*. Their scheme has a polylogarithmic communication cost and is also reasonably fast in terms of computation.

Kushilevitz and Ostrovsky give the first single-server single-query PIR protocol [25] based on Additively Homomorphic Encryption (AHE). Several subsequent works improve the asymptotic communication cost using various techniques and assumptions [12], [17], [19]. But Sion and Carbunar [38] observe that these schemes in practice often perform worse than downloading the entire database due to the prohibitive cost of applying number-theory-based AHE to the entire database.

Recent practical single-server single-query PIR schemes [30], [8], [7], [33], [31] have switched from AHE to lattice-based leveled Fully Homomorphic Encryption (FHE) to reduce the computation cost. These schemes boast much better server computation than their early AHE counterparts, though server computation can still be a major bottleneck when the database is large. State-of-the-art schemes in this category achieve good communication overhead when the database entry is large (on the order of kilobytes). But for databases with small entries, the communication overhead is very high because the ciphertexts of lattice-based encryption are quite large, usually on the order of tens of kilobytes, no matter how small the underlying plaintext is.

All of the above schemes, multi-server and single-server ones alike, require linear server computation. As mentioned, this is unavoidable in the most standard single-query PIR model. This is formalized by Beimal, Ishai, and Malkin [9] as a lower bound that any PIR scheme on an N-entry database must incur $\Omega(N)$ computation at the server. To our knowledge, three avenues have been explored in an attempt to circumvent the linear server computation barrier. We review them next.

PIR with database preprocessing. In the same paper that established the $\Omega(N)$ server computation lower bound, Beimel, Ishai, and Malkin [9] also show that the lower bound can be circumvented by preprocessing and encoding the database offline. This approach is also taken by a line of works known as doubly efficient PIR [13], [11], [28]. These efforts have so far remained largely theoretical because they have to significantly blow up server storage (superlinearly or by the number of clients), require heavyweight theoretical tools (such as oblivious locally decodable codes or virtual-blackbox obfuscation), or suffer from both drawbacks.

Batch PIR vs. stateful PIR. The other two avenues to circumvent the linear server computation barrier both assume the client has many entries to fetch from the server. One is called batch PIR [23], [8], and the other is what we call stateful PIR [36], [16]. They provide opportunities for substantial efficiency improvements through *amortization*. While the total cost (to fulfill all queries) is still subject to the fundamental barriers, the *per-query* cost may now be made much lower.

The difference between these two variants of PIR is that batch PIR assumes the client has many queries to execute in one go, while in contrast, stateful PIR assumes that the client generates queries sequentially (e.g., the client decides what the next query is after receiving the response for its previous query). This can be formally captured by the adaptive version of the stateful PIR in Section 2. Note that batch PIR is an easier problem than stateful PIR because the client can always send a batch of queries one by one, but it cannot batch chronologically sequential (and potentially causal) queries.

Since the focus of this paper is on stateful PIR, we review batch PIR only briefly for completeness. Ishai et al. [23] propose the first batch PIR scheme (and called it amortized PIR in their paper) using batch codes. Angel et al. [8] gives the first practical batch PIR scheme using cuckoo hashing. The Angel et al. scheme nicely amortizes the linear server computation cost: it costs $O(N)$ server computation to fulfill all the queries in the batch, no matter how large the batch is. But their scheme does not amortize the response overhead: $O(b)$ ciphertexts must be sent back

for a batch of b queries. Mughees and Ren [34] give a batch PIR scheme that further amortizes the response overhead over the batch using vectorized FHE where a single ciphertext can hold as many queried entries as what can fit.

Stateful PIR. Patel, Persiano, and Yeo [36] propose the paradigm of stateful PIR in which the client retrieves hints privately in an offline phase and later uses these hints to speed up online queries. At some level, this offline phase can also be viewed as a preprocessing step. But a crucial difference is that this offline preprocessing step does not alter the server's database in any way, and hence requires no extra server storage. The original work of Patel, Persiano, and Yeo was less ambitious. Their goal was to replace the linear homomorphic encryption operations with linear PRF evaluations, and was not to circumvent the linear server computation bound. But it did not take long for amortized sublinear stateful PIR schemes to emerge.

Corrigan-Gibbs and Kogan [16] give the first amortized sublinear stateful PIR scheme. Their scheme initially works in the two-server setting and is later extended to the single-server setting [15].

The schemes of Corrigan-Gibbs et al. [16], [15] have $\Omega(\lambda\sqrt{N})$ request size. A number of works resort to privately puncturable/programmable pseudorandom functions (PRF) to improve the request size asymptotically, first in the two-server setting [37] and later in the single-server setting [39], [26]. These works are mostly theoretical at the moment because privately puncturable/programmable PRFs are heavyweight theoretical tools and do not have practical instantiations.

It is also worth noting that there is a more pressing performance bottleneck than the request size, namely, the parallel repetitions. All of the above works [16], [15], [37], [39], [26] allow a small probability of a correctness failure. Thus, their scheme must be repeated λ times in parallel to make the correctness failure probability negligible. This will blow up all efficiency metrics, including request size, response size, client storage, client computation, and server computation. To our knowledge, none of the above schemes has been implemented.

Practical amortized sublinear stateful PIR. Two recent works give methods to eliminate this correctness failure and hence avoid the parallel repetitions [24], [27]. Both works make use of (non-private) puncturable PRFs, which do have practical instantiations. As we have mentioned, both schemes only work for the two-server setting and have no clear path to be extended to the single-server setting. Moreover, both schemes come with substantial efficiency losses. The Kogan-Corrigan-Gibbs scheme [24] requires either $\Theta(N)$ client storage or $\Theta(N)$ client computation per query. Since the motivation of stateful PIR is to avoid linear server computation, it is hard to justify shifting a linear cost to the client, which is often more resource-constrained.

The Lazzaretti-Papamathou scheme [27] increases the response overhead to $\Theta(\sqrt{N})$. This large response overhead is usually more problematic than a large request size because requests are measured in $\log N$-sized words (usually less than 32 bits), while the responses are measured in the database entry size, which can be hundreds of bytes or more. Lazzaretti and Papamathou note that the response overhead blowup can be mitigated, in theory, by invoking an extra single-server single-query PIR. However, this would not be efficient in practice. Lazzaretti and Papamathou have already observed in their experiments [27] that this extra PIR can be the performance bottleneck even with a state-of-the-art construction like Spiral [31]. Unfortunately, they still significantly underestimate how costly such a strategy is. First of all, state-of-the-art FHE-based PIR constructions perform a one-time preprocessing of the database using Number-Theory Transform (NTT) [8], [7], [33], [31]. Since the $\Theta(N)$-sized response is computed by the server based on the query, this NTT step will have to be performed on the fly at the end of each query. This will make FHE-based PIR an order of magnitude slower. Secondly, this extra FHE-based PIR would again suffer from the large ciphertext size of lattice encryption. Lastly, we note that FHE-based PIR has other drawbacks such as the typical key-dependent security and several megabytes of server storage per client, both resulting from the substitution keys for query (de-)compression.

Despite the shortcomings in the response overhead and the need for two servers, the Lazzaretti-Papamathou scheme introduces an elegant technique that is crucial for our work. Their scheme uses a more structured hint construction where the database is divided into equal-sized partitions, and each subset consists of one index per partition. These partition-based hints are more amenable to succinct pseudorandom representations and faster membership testing. They hence enable more space-efficient hint storage and offline processing. Our work adopts their partition-based hints.

Zhou et al. [40] adapt the Lazzaretti-Papamathou scheme to a single server. But to do so, they had to weaken the correctness guarantee of PIR and require that the query sequence is not influenced by the adversary. To elaborate, they also use the backup hint technique, but each backup hint is specific to a particular partition. As a result, they require the queries in the sequence to be balanced across all partitions. To do so, they let the server permute the entire database and publish the permutation key. They then require that the client's query sequence is independent of the server's permutation. This assumption may be justifiable in some use cases but not always. It is not hard to conceive scenarios in which the client's query sequence is influenced by an untrusted third party. Because the server's permutation is public, this third party can make their PIR scheme lose correctness, even if the server is honest. If a PIR-based system behaves differently when a query fails, a correctness failure can also lead to a privacy violation.

6. Conclusion

We have presented simple and lightweight stateful PIR schemes with amortized sublinear communication and computation for both the two-server and single-server settings. Our schemes avoid the major performance bottlenecks in

prior works: parallel repetition, linear client storage, and large response overhead.

Our schemes also have drawbacks that call for further studies. An obvious one is the $\Omega(\sqrt{N})$ request size. There exist techniques to reduce the request size, but the challenge is to do so without sacrificing other aspects of the algorithm. A limitation shared by all existing amortized sublinear schemes is that the $O(\lambda\sqrt{N})$ client storage, while sublinear, is still quite large in practice. An indirect consequence is that the single-server offline phase cannot do much better than streaming the whole database when the client needs so many hints. Other general challenges involving stateful PIR include how to handle updates to the database and how to support queries by keywords, and recent works have already started looking into these directions [24], [29].

References

[1] "Checklist-go," Online: https://github.com/dimakogan/checklist.

[2] "Dpf-cpp," Online: https://github.com/dkales/dpf-cpp.

[3] "Pianopir-go," Online: https://github.com/pianopir/Piano-PIR.

[4] "Simplepir-go," Online: https://github.com/ahenzinger/simplepir.

[5] "Spiral-cpp," Online: https://github.com/menonsamir/spiral.

[6] "Treepir-go," Online: https://github.com/alazzaretti/treePIR.

[7] A. Ali, T. Lepoint, S. Patel, M. Raykova, P. Schoppmann, K. Seth, and K. Yeo, "Communication–computation trade-offs in PIR," in *30th USENIX Security Symposium*, 2021, pp. 1811–1828.

[8] S. Angel, H. Chen, K. Laine, and S. Setty, "Pir with compressed queries and amortized query processing," in *2018 IEEE symposium on security and privacy (SP)*. IEEE, 2018, pp. 962–979.

[9] A. Beimel, Y. Ishai, and T. Malkin, "Reducing the servers computation in private information retrieval: Pir with preprocessing," in *20th Annual International Cryptology Conference (CRYPTO)*. Springer, 2000, pp. 55–73.

[10] E. Boyle, N. Gilboa, and Y. Ishai, "Function secret sharing: Improvements and extensions," in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, 2016, pp. 1292–1303.

[11] E. Boyle, Y. Ishai, R. Pass, and M. Wootters, "Can we access a database both locally and privately?" in *Theory of Cryptography: 15th International Conference*. Springer, 2017, pp. 662–693.

[12] C. Cachin, S. Micali, and M. Stadler, "Computationally private information retrieval with polylogarithmic communication," in *International Conference on the Theory and Application of Cryptographic Techniques (EUROCRYPT)*. Springer, 1999, pp. 402–414.

[13] R. Canetti, J. Holmgren, and S. Richelson, "Towards doubly efficient private information retrieval," in *Theory of Cryptography: 15th International Conference*. Springer, 2017, pp. 694–726.

[14] B. Chor, E. Kushilevitz, O. Goldreich, and M. Sudan, "Private information retrieval," *Journal of the ACM (JACM)*, vol. 45, no. 6, pp. 965–981, 1998.

[15] H. Corrigan-Gibbs, A. Henzinger, and D. Kogan, "Single-server private information retrieval with sublinear amortized time," in *Annual International Conference on the Theory and Applications of Cryptographic Techniques*. Springer, 2022, pp. 3–33.

[16] H. Corrigan-Gibbs and D. Kogan, "Private information retrieval with sublinear online time," in *Annual International Conference on the Theory and Applications of Cryptographic Techniques*. Springer, 2020, pp. 44–75.

[17] I. Damgård and M. Jurik, "A generalisation, a simplification and some applications of paillier's probabilistic public-key system," in *4th International Workshop on Practice and Theory in Public Key Cryptosystems (PKC)*. Springer, 2001, pp. 119–136.

[18] Z. Dvir and S. Gopi, "2-server pir with subpolynomial communication," *Journal of the ACM (JACM)*, vol. 63, no. 4, pp. 1–15, 2016.

[19] C. Gentry and Z. Ramzan, "Single-database private information retrieval with constant communication rate," in *International Colloquium on Automata, Languages, and Programming*. Springer, 2005, pp. 803–815.

[20] N. Gilboa and Y. Ishai, "Distributed point functions and their applications," in *33rd Annual International Conference on the Theory and Applications of Cryptographic Techniques (EUROCRYPT)*. Springer, 2014, pp. 640–658.

[21] T. Gupta, N. Crooks, W. Mulhern, S. Setty, L. Alvisi, and M. Walfish, "Scalable and private media consumption with popcorn," in *13th USENIX Symposium on Networked Systems Design and Implementation (NSDI 16)*, 2016, pp. 91–107.

[22] A. Henzinger, M. M. Hong, H. Corrigan-Gibbs, S. Meiklejohn, and V. Vaikuntanathan, "One server for the price of two: Simple and fast single-server private information retrieval," in *Usenix Security*, vol. 23, 2023.

[23] Y. Ishai, E. Kushilevitz, R. Ostrovsky, and A. Sahai, "Batch codes and their applications," in *Proceedings of the thirty-sixth annual ACM symposium on Theory of computing*, 2004, pp. 262–271.

[24] D. Kogan and H. Corrigan-Gibbs, "Private blocklist lookups with checklist," in *30th USENIX Security Symposium*. USENIX Association, 2021.

[25] E. Kushilevitz and R. Ostrovsky, "Replication is not needed: Single database, computationally-private information retrieval," in *Proceedings 38th annual symposium on foundations of computer science*. IEEE, 1997, pp. 364–373.

[26] A. Lazzaretti and C. Papamanthou, "Near-optimal private information retrieval with preprocessing," *Cryptology ePrint Archive*, 2022.

[27] ——, "Treepir: Sublinear-time and polylog-bandwidth private information retrieval from ddh," *Cryptology ePrint Archive*, 2023.

[28] W.-K. Lin, E. Mook, and D. Wichs, "Doubly efficient private information retrieval and fully homomorphic ram computation from ring lwe," in *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*, 2023, pp. 595–608.

[29] Y. Ma, K. Zhong, T. Rabin, and S. Angel, "Incremental offline/online PIR," in *31st USENIX Security Symposium*, 2022, pp. 1741–1758.

[30] C. A. Melchor, J. Barrier, L. Fousse, and M.-O. Killijian, "Xpir: Private information retrieval for everyone," *Proceedings on Privacy Enhancing Technologies*, pp. 155–174, 2016.

[31] S. J. Menon and D. J. Wu, "Spiral: Fast, high-rate single-server pir via fhe composition," in *2022 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2022, pp. 930–947.

[32] P. Mittal, F. Olumofin, C. Troncoso, N. Borisov, and I. Goldberg, "Pir-tor: Scalable anonymous communication using private information retrieval," in *20th USENIX security symposium*, 2011.

[33] M. H. Mughees, H. Chen, and L. Ren, "Onionpir: Response efficient single-server pir," in *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, 2021, pp. 2292–2306.

[34] M. H. Mughees and L. Ren, "Vectorized batch private information retrieval," in *2023 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2023, pp. 437–452.

[35] D. R. Musser, "Introspective sorting and selection algorithms," pp. 983–993, 1997.

[36] S. Patel, G. Persiano, and K. Yeo, "Private stateful information retrieval," in *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*, 2018, pp. 1002–1019.

[37] E. Shi, W. Aqeel, B. Chandrasekaran, and B. Maggs, "Puncturable pseudorandom sets and private information retrieval with near-optimal online bandwidth and time," in *41st Annual International Cryptology Conference (CRYPTO)*. Springer, 2021, pp. 641–669.

[38] R. Sion and B. Carbunar, "On the computational practicality of private information retrieval," in *Proceedings of the Network and Distributed Systems Security Symposium, San Diego, California, USA*, 2007.

[39] M. Zhou, W.-K. Lin, Y. Tselekounis, and E. Shi, "Optimal single-server private information retrieval," in *Annual International Conference on the Theory and Applications of Cryptographic Techniques*. Springer, 2023, pp. 395–425.

[40] M. Zhou, A. Park, E. Shi, and W. Zheng, "Piano: Extremely simple, single-server pir with sublinear server computation," *Cryptology ePrint Archive*, 2023.