Optimized range queries for large scale networks

Nicolas Hidalgo*, Erika Rosas*, Luciana Arantes*, Olivier Marin*, Pierre Sens* and Xavier Bonnaire†
*Université Pierre et Marie Curie, CNRS - INRIA - REGAL, Paris, France
E-mail: [firstname.lastname]@lip6.fr

†Department of Computer Science, Universidad Técnica Federico Santa María, Valparaíso, Chile
E-mail: firstname.lastname@inf.utfsm.cl

***Abstract*—Distributed Hash Tables (DHTs) provide the substrate to build scalable and efficient Peer-to-Peer (P2P) networks: distributed systems with the potential to handle massive amounts of data on a very large scale. However, traditional DHTs provide very poor support for range queries.**

In this article we present a search mechanism that efficiently supports range queries over a ring-like DHT structure using a prefix tree index. Load balancing is improved by delegating the routing of queries to the nodes that store data, and by updating neighbor information through an optimistic approach. Our solution reduces latency and message traffic in environments where queries are more frequent than data insertion operations. We evaluate the performance of the system through simulations and show that our solution in not affected by data skewness.

***Keywords*-Peer-to-Peer, DHT, Information Retrieval, Range Queries.**

I. Introduction

Supporting range queries in large-scale, dynamic networks and acquiring an exhaustive list of results in a timely manner is now a common issue for a wide variety of distributed applications such as music/movie storage, peer-to-peer persistent games, scientific computation, data mining, and many types of data warehouses. A range query retrieves all the objects with values within a given range.

Peer-to-peer (P2P) systems offer an abundance of resources, with a huge data storage capacity [1]. They are autonomous, fully decentralized, self-organized and highly scalable with the potential to grow up to millions of nodes. In particular, Distributed Hash Tables (DHTs) provide an efficient lookup functionality. They are scalable, fault tolerant, and offer load balancing. Some well-known DHT-based overlays found in the literature are Pastry [2], Chord [3], and Kademlia [4] on top of which a lookup operation is performed in $\mathcal{O}(\log(N))$ hops, where N is the total number of nodes of the network. However, DHTs do not support range queries efficiently since the use of uniform hashing destroys data locality.

Several solutions in the literature attempt to provide efficient support of range queries over DHT-based P2P networks. However, none of them manage to combine load balancing, low message overhead, and low search latency simultaneously while preserving the scalability of the system. The existing solutions mainly fall into two categories: *over-DHT* and *overlay-Dependent* solutions. *Over-DHT* approaches build an additional structure or index, which preserves data locality and thus allows to support range queries [5]–[10]. Several of these approaches rely on a tree-based indexing structure (e.g. PHT [5], RST [7], DST [6], and LIGHT [10]). On one hand, as *over-DHT* solutions are built over a generic DHT, they preserve the good properties of the latter, are portable, and are easy to implement. On the other hand, their main drawbacks lie in the extra maintenance cost, in the significant query latency due to the top down nature of traversal searches over the structure, and in the load balancing problems associated to the search process. *Overlay-Dependent* approaches directly support range queries by adapting the DHT (e.g. Yarqs [11], Skipnet [12], Skipgraph [13], Hivory [14]), by using non-uniform mapping techniques which preserve data locality (e.g. MANN [15], AR [16], Mercury [17]) or by adopting a tree-based approach (e.g. Pgrid [9], BATON [18], Skip tree graph [19], Hyperring [20], Ptree [21]). Nevertheless, *overlay-Dependent* solutions present load balancing and portability problems while some of them induce high latency and substantial maintenance costs.

In this paper we aim at improving range queries search in environments where queries significantly outnumber data insertion operations; resource discovery and data mining belong to this category of applications. We thus propose a solution which focuses on the search process; it works independently of the data distribution and can perform range queries in the order of $\mathcal{O}(\log(L))$ steps, where L is the number of nodes that store data. Based on the indexing structure of Prefix Hash Tree (PHT) [5], our approach provides a high performance range query management: it reduces search latency, achieves good load balancing among the nodes, and produces low message traffic overhead by diverting the queries to the nodes that store the data. We conducted an extensive set of simulations which demonstrates a significant improvement when compared to the original PHT approach with a reduction of traffic overhead by at least 65%.

The rest of this paper is organized as follows. Section II briefly describes the overlay and the indexing approach on top of which we build our solution, while Section III presents the structure and search operations. In Section IV we present a performance evaluation conducted through simulation on top of PeerSim [22]. Finally, Section V concludes this paper.

II. Background

A DHT is a self-organizing structured substrate built over P2P overlay networks in which any data can be located within

a bounded number of routing hops. Several existing DHT overlays like Chord [3], Pastry [2] or Tapestry [23] logically organize the nodeID space of nodes into a ring.

Our work uses Pastry [2], a ring-like structured DHT. Every node in Pastry is associated to a unique nodeID in an m-bit space using the SHA-1 hash function. Numeric keys represent application data (objects) and are chosen from the same name-space as the node identifiers. Pastry assigns each key k to the node whose nodeID is numerically closest to k.

While lookup operations can be efficiently executed over DHTs, the latter do not provide support for complex queries, such as range queries. To this end, several indexing schemes have been proposed in the literature [5]–[7], [10], [15], [17]. We base our work on Prefix Hash Tree (PHT) [5], which provides an indexing data structure built over DHT-based P2P networks.

In PHT, keys of objects to be indexed are within the domain $\{0,1\}^D$, where D is the length of the string. PHT is thus a binary prefix tree (binary trie) where the left branch of a node is labeled 0 and the right branch is labeled 1. Each node n of the trie is identified with a chain of P bits (prefix) produced by the concatenation of the labels of all branches in the path from the root to n. PHT builds a prefix tree in which objects are stored at *leaf nodes*. Hence, an object with key k is stored at a leaf node whose label is a prefix of k. Fig. 1 illustrates an example of a PHT.

Compared to other tree-linked solutions [6], [7] where a B-tree with O objects requires $\log(O)$ lookups, PHT requires only $\log(D)$ lookups, where D is the key size. Additionally, replication over the DHT can prevent data loss.

The PHT trie is completely distributed among the peers of the network by *hashing* the prefix labels of the PHT nodes over the underlying DHT identifier space. We call *internal nodes* those nodes that belong to the PHT trie but are not leaves while those nodes that do not participate to the PHT trie we denote *external nodes*.

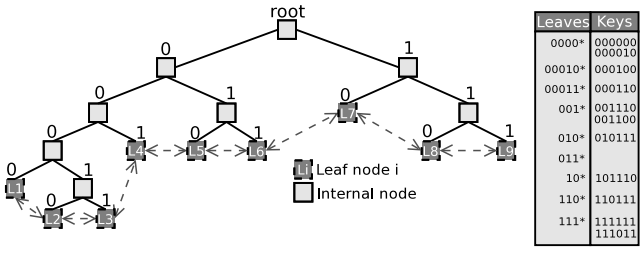


Fig. 1. PHT indexing structure.

Whenever a leaf node X reaches its maximum storage capacity, it splits into two children so as to distribute its storage load. The left child will have the prefix label of its parent concatenated with 0 and the right child with 1.

Classic searches in PHT follow either a linear or a binary method. A linear search starts with the node that corresponds to the smallest possible prefix of a given key k. The next step consists in a DHT lookup operation in order to find the next node in the trie whose prefix is one bit more similar to k. The lookup operation is repeated until a leaf node is reached. This search method produces a number of *DHT-lookups* of order D. A binary search is a half-interval process that starts by querying a middle prefix of D. If the prefix corresponds to an internal node of the PHT, the search discards the lower half of the interval and continues querying a new middle prefix of the remaining interval. If the prefix corresponds to an external node, the search discards the upper half of the interval. This search method produces a number of *DHT-lookups* of order $\log(D)$. However, in dynamic scenarios the binary search can fail, i.e. be unable to correctly locate the leaf node [5].

PHT maintains a double list which links all leaf nodes (*Threaded leaves*), as shown in Fig. 1 by the dashed lines. This list allows to search all data of the requested range sequentially, starting from the lower bound. The sequential scan of the list can be avoided by parallelizing the search within the range. PHT [5] proposes to start from the node that corresponds to the prefix that completely covers the requested range and then continue making a number of searches that correspond to the number of children which overlap the range, until reaching all the leaf nodes. However, if the trie is highly unbalanced, the number of messages will grow without decreasing the latency when retrieving all data. In Section IV, we refer to this type of search as *PHT-parallel*.

III. Optimizing Range Queries

This section details PORQUE (Peer Optimized Range QUEries), a two-layer ring structure that effectively supports range queries. A node exploits information provided by range queries that it has previously issued. When performing a new query, such information is used to contact the leaf nodes directly, thus reducing search latency. In order to further optimize search latency, we take into account range coverage. PORQUE is an extension of our previous work DRing [24] which proposes a new method to improve the retrieval process.

A. Building a Double-ring Architecture

PHT nodes are uniformly distributed among the nodes of the DHT, based on the prefix of their key. In order to improve the performance of range queries, our approach maintains a second ring structure over the DHT overlay. This second ring, denoted *Leaf Ring*, has a structure based on Skipnet [12] and comprises only the *leaf nodes*, i.e. nodes that store data. The identifier of a node in the Leaf Ring is its corresponding label identifier in the trie. Such a choice improves sequential data search, and therefore facilitates range querying.

In the Leaf Ring, every node maintains a reference both to its successor and predecessor leaf in the trie, similarly to the double list structure of PHT. Notice that PHT exploits this list for data retrieval only, not for searching. Fig. 2 presents the double-ring proposed architecture, where nodes in grey represent leaf nodes.

In addition to successor and predecessor references, every node in the *Leaf Ring* maintains a *Search Table* which stores references to other leaf nodes in the *Leaf Ring*. Fig. 3 shows an example of a Search Table. The maximum number of entries

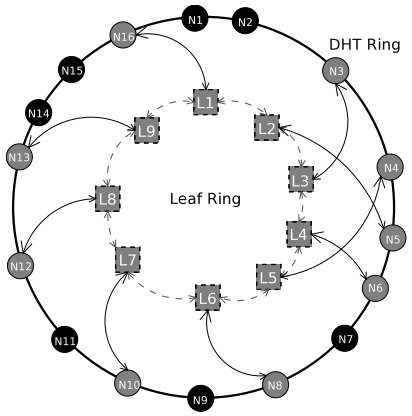


Fig. 2. Double structure: Leaf Ring over the overlay ring

is the maximum object key size D. The ith row of the Search Table contains two references to nodes that are at a *distance* of 2^i hops in both directions (forward and backward) to a prefix label identifier. For example, for node $L5$ in Fig. 2, the nodes that are 2 (2^1) hops forward and backward are $L7$ and $L3$ respectively.

Each reference comprises both the prefix label of a leaf node, and a static reference (IP Address) to this leaf node. Maintaining the prefix label is paramount in dynamic environments, where static references become obsolete very fast. When the access to a static reference fails, PORQUE contacts the leaf node by hashing the corresponding prefix label and looking it up in the DHT.

In order to optimize access the Leaf Ring structure, information from past queries is stored on every node in the P2P network. This information allows any DHT node to contact the Leaf Ring directly, thus skipping the internal level of the PHT trie-structure (see Section III-D).

Within the Leaf Ring layer, searching follows the routing technique of the underlying DHT structure: using a greedy technique that forwards the search to the node which is closest to the key.

SEARCH TABLE
Node L5

	Distance from L5	Fwd_Prefix	Fwd_IP Address	Back_Prefix	Back_IP Address
1	2^1	10	190.168.0.15:99	00011	200.50.0.15:99
2	2^2	111	200.1.19.85:99	0000	81.168.0.15:99
3	2^3	001	81.160.5.1:99	011	132.227.64.15:99
⋮					
i	2^i	...	...	...	...

Fig. 3. A Search Table of a node

A node iteratively fills its Search Table the first time it joins the Leaf Ring. To complete the forward reference at distance 2^{i+1} in the Search Table, a node X asks the node at distance 2^i for the ith forward reference in its table. For example, to find the node at forward distance 2^1, X can obtain the information from its successor (since it is at distance 2^0); and to find the node at distance 2^5, X can ask the node at distance 2^4 to return the 4th entry of its Search table. The same goes for backward references, with the predecessor set as the node at a backward distance of 1.

The approach described above performs efficiently if the trie is balanced. Indeed, join and leave operations induce a global change in the tables of the leaf nodes, which grow linearly with the number of nodes. This happens because we take into account distances instead of making a static partition of the name-space. It is worth noticing that a static partition of the index space is not efficient either: when the trie is unbalanced, the load distribution follows the disequilibrium. In order to overcome this balancing problem and repair the Leaf Tables of nodes, we propose an optimistic approach which consists in updating tables only when performing a range query (see Section III-C). We argue that this suffices to achieve good performance in scenarios where queries significantly outnumber data insertion operations.

B. Split Operation

Every split operation in the logical PHT trie induces two joins and one leave operation in the Leaf Ring. The split node is removed from the Leaf Ring and both its children join the Leaf Ring.

Whenever a node splits, this node – denoted parent – must inform its children about the identifier of its predecessor and successor nodes in the Leaf Ring: the predecessor of the left child is its parent's predecessor in the Leaf Ring and its successor is its right brother; similarly, the predecessor of the right child is its left brother and the successor is its parent's successor in the Leaf Ring. The children Search Tables are thus initialized with the information provided by their parent node and they are further updated in an optimistic way as described in Section III-C.

Fig. 4 shows an example of a split operation in the Leaf Ring. The node with identifier 10 reaches its storage capacity and is replaced by two other nodes with prefixes 100 and 101. The information from the node with prefix 10 is used by its children in order to update the successor and predecessor links as well as the Search Table.

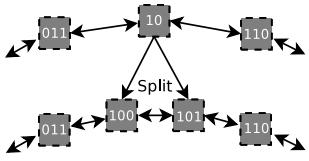


Fig. 4. A split operation in the Leaf Ring

In case of concurrent splits by neighbor nodes in the Leaf Ring, the obtained information that a child gets from its parent might be out of date. Thus if the parent's successor Y is no longer in the Leaf Ring, the right child will contact the left child of Y. Notice that the prefix label of Y's left child can be obtained by the concatenation of Y's prefix label with '0'; similarly, if the parent's predecessor Z is no longer in the Leaf Ring, the left child will contact Z's right child. The nodeID

of the latter is obtained by concatenating of Z's prefix label with '1'.

C. Optimistic Table Maintenance

We propose to repair the Search Tables optimistically: reference updates only occur when a range query takes place.

In Pastry, data corresponding to a leaf node with prefix label A is stored on the node whose identifier is closest to $SHA(A)$. Since DHT nodes can crash or leave the network, Pastry replicates data on several nodes in the numerical vicinity of $SHA(A)$ so as to avoid information loss. Therefore in dynamic environments, the node responsible for another node can change. When a leaf node leaves the network or crashes, its static reference is no longer valid. In this case, Pastry falls back on the new current node whose identifier is numerically closest to $SHA(A)$ as the holder of the data. Given the prefix label, this new node can easily be contacted through a DHT lookup operation and the static reference to its IP address is updated. It is worth pointing out that no other static reference in the Search Table requires an update.

Every split of a leaf node of the trie produces inconsistency in its Search Table. Similarly, this table is only updated when a range query takes place. As will be described in Section III-D, a range query is forwarded to a leaf node which is numerically closest to one of the bounds of the range query. Upon receiving the query, this node looks into its Search Table references and forwards a message to the i reference which is numerically closest to the searched node, also including its table reference $i+1$ in the message (`maintenance`). The node that receives the message checks if this reference corresponds to its own at row i in its table. If such is not the case, it passes on its entry reference to the sender which will then update its table.

Since queries are necessary to keep the Search Table updated, PORQUE performs optimally when range queries outnumber insertions of data significantly. In the opposite case, the nodes in PORQUE cannot update their tables efficiently enough and the performance diminishes (see Section IV for details). No matter how many splits occur, the number of maintenance messages will be the same: one for each query transmission. If a query reaches its target in d hops, PORQUE may generate d maintenance messages if every entry is out of date. Since these messages are at the leaf level and they are one-hop messages, the impact on performance remains low.

D. Range Query Strategies over the Leaf Ring

Every node in the DHT maintains a *Cache List* which contains the leaf nodes *contacted* during previous range query requests. *Contact nodes* are thus potential starting points for direct access searches over the Leaf Ring.

The size of the Cache List is a parameter of the system. Every entry stores both the static IP address reference and the prefix label that identifies a leaf node in the trie (and in the Leaf Ring). If an access first fails, a DHT lookup operation finds the new contact node by applying the hash function over the prefix label kept by the entry. When the Cache List is full, new entries replace old entries following a Least Recently Used (LRU) strategy.

If a node has never performed any query yet, its Cache List is empty: it knows no *contact node*. In order to fill its Cache List, a node first asks for the Cache List of its neighbors. However, if such an information is not available either, it applies one of the PHT search mechanisms (linear or binary), as explained in Section II.

A range query corresponds to an interval of data to be searched. It has the form $R_q = [l, u]$, where l is the lower bound and u is the upper bound. To collect the data in the range, we propose two search protocols:

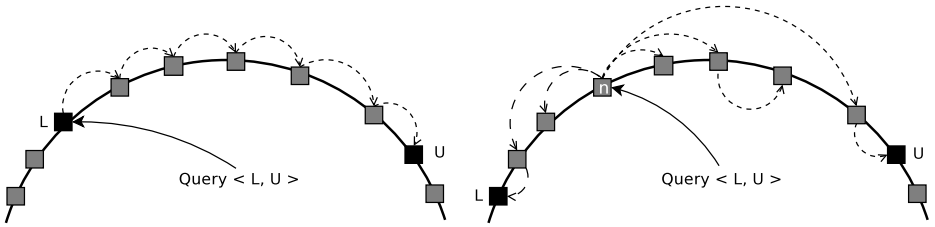
Algorithm 1: Route Operation: Sequential Search

```
   input : Query Q, the query
1  maintenance (Q.maintenanceEntry, Q.direction )
2  if iamClosest to Q.targetBound then
3      getRange (Q)
4  else
5      n = getClosest in Search Table to Q.targetBound
6      Q.maintenanceEntry = getEntry(n.index + 1)
7      Q.direction = getDirection(n.index)
8      send: route (Q) to n.address
```

Algorithm 2: GetRange: Sequential Search

```
   input : Query Q, the query
1  sendData
2  if Q.targetBound = Q.lowerBound then
3      if successor =< Q.upperBound then
4          forward: Q to successor
5  else
6      if predecessor => Q.lowerBound then
7          forward: Q to predecessor
```

- *Sequential Search*: The simplest mechanism is to start the search at one of the bounds. Therefore, the requester must select a node in its Cache List (*contact node*) with a prefix label numerically close either to the lower or to the upper bound. The search will start from the bound value that has the greatest common prefix label with the nodes referenced in the Cache List of the requester. If both bound values satisfy this condition or if the Cache List is empty, either of the bound values can work as the *starting search node*. From the contact node, the target bound can be easily found through the Leaf Ring. The routing process followed to retrieve the query information is presented in Algorithm 1. The query is then routed to the closest node including entry $i+1$ of its Search Table (line 6). Firstly, the contact node must check if it is in charge of the query target by invoking the function `iamClosest`. If it is the case, the `getRange` sequential function is invoked to start the retrieval process (line 3). To retrieve all the data comprised in the range query, the search can continue along the successor or predecessor links of the Leaf Ring. On the other hand, the `getRange` sequential is presented in Algorithm 2. Fig. 5(a) presents an example



(a) Sequential Search (b) Optimized Search

Fig. 5. Search Schemes

where the lower bound is the starting point of the range search.

Algorithm 3: Route Operation: Optimised Search

```
input : Query Q, the query
1  maintenance (Q.maintenanceEntry, Q.direction)
2  if iamWithinRange then
3      Q.direction = null
4      Q.bound = null
5      getRangeOp (Q)
6  else
7      n = getClosest in Search Table to Q.targetBound
8      Q.maintenanceEntry = getEntry(n.index + 1)
9      Q.direction = getDirection(n.index)
10     send: route (Q) to n.address
```

- *Optimized Search*: In order to reduce search latency, the information of the entries of the Search Table allows concurrent access to the leaf nodes that contain data within the range of the query. In this case, the starting search node is the one that stores the keys associated to $(l + u)/2$, i.e. the node in the middle of the requested range. However, as the routing forwards the query to the node which is numerically closest to this value, it is highly possible that the query will reach a node within the range before it reaches the middle of the requested range. The routing process is presented in Algorithm 3. In the optimized case, when the query reaches any node within the lower and upper bound the data can be retrieved. Therefore, when the request arrives at any leaf node contained in the searched range, this node forwards/backwards the search messages to all the nodes of its Search Table able to answer the query. All receiving nodes will apply this same approach. The retrieval process is presented in Algorithm 4.
  So as to avoid redundant forwarding/backwarding messages, the query message collects information about previously visited leaf nodes. Fig. 5(b) illustrates this scheme. Node n, the initial receiver of the query, forwards the query to its successor and two other nodes of its Search Table. In parallel, it also backwards it to its predecessor and one other node of its Search Table. All these nodes have a prefix that covers part of the range query.

Our approach also supports other complex queries in a simple, efficient way. For instance *Min/max queries* look for the smallest and largest value of the indexed data. PORQUE carries them out with a single lookup operation. *K-NN queries* return the K nearest data values to a given key. The successor and predecessor links of the Leaf Ring allow to find the required data efficiently.

Algorithm 4: GetRangeOpt: Optimised Search

```
input : Query Q, the query
1  sendData
2  foreach Entry e in Search Table do
3      if Q.direction = forward or null then
4          if e.fwdPrefix < Q.upperBound then
5              if Q.bound = null or e.fwdPrefix < Q.bound then
6                  Q.bound = next_fwd_entry_in_table
7                  Q.direction = forward
8                  forward: Q to e.fwdAddress
9      if Q.direction = backward or null then
10         if e.bwdPrefix > Q.lowerBound then
11             if Q.bound = null or e.bwdPrefix > Q.bound then
12                 Q.bound = next_bwd_entry_in_table
13                 Q.direction = backward
14                 forward: Q to e.bwdAddress
```

IV. Performance Evaluation

In this section we discuss performance evaluation results when using PORQUE and PHT to perform queries. Our experiments were conducted on top of Peersim [22] using an event-driven model.

Queries are generated over data sets with different distributions: Uniform, Gaussian, and the real dataset of DBLP[1]. The indexed attribute in DBLP dataset is the author name. Nodes that generate queries are selected randomly.

We consider a network of $10,000$ peers, where each peer can store at most 100 objects. During every simulation we perform a total of $50,000$ queries, and take a snapshot of the simulation every $1,000$ queries. The range span of the queries is minimal in the first experiments so as to give a separate analysis. We present our most interesting results in Fig. 6.

Data distributions: The trie structure shape is tightly related to the indexed data keys. Uniform distributions generate uniform tries where all leaf nodes are located at almost the same level. On the other hand, skewed distributions (eg. Gaussian

[1]DBLP Dataset http://dblp.uni-trier.de/xml/

and DBLP) generate unbalanced tries where clustered data (ie. frequent data keys within a given range) are stored on the deeper leaf nodes of the trie.

Gaussian and DBLP dataset present skewed data distributions where keys are clustered on small parts of the whole key range. The DBLP dataset distribution is different from the Gaussian in that keys are clustered in many spots. A comparison between the Gaussian and DBLP dataset distributions used for our simulations is given in Fig. 6(r).

Traffic Message Overhead: We have compared the total number of messages generated in the search processes. Furthermore, each simulation starts by inserting $20,000$ objects in the network. In the case of PORQUE the results also include the maintenance messages, which actually represent only 0.1% of the total traffic.

Fig. 6(a), 6(b), and 6(c) show the message traffic for Uniform, Gaussian, and DBLP dataset distribution respectively. These figures illustrate how strongly the performance of PHT searches depends on the distribution of data, which defines the shape of the index trie. Linear PHT generates more messages on deep tries. This occurs in the case of DBLP (Fig. 6(c)), where data distribution is skewed.

The traffic associated to the binary PHT search also depends on the shape of the index trie structure: for the Gaussian and DBLP distributions, PHT binary search can find a leaf node very fast. We should point out that this result relies heavily on the maximum key prefix size (which in our experiments is set to 40) and on the size of the searched branch.

Comparatively, the number of messages generated by our PORQUE search process is far smaller than for both PHT searches. In addition, PORQUE search performance does not depend on data distribution, as it avoids access to internal nodes of the indexing structure and performs the search only over the leaf nodes. At the beginning of the experiment peers do not know any leaf node, so nodes will use the default PHT linear search. Every node quickly starts adding leaf nodes to its Cache List as it handles queries. After 10,000 queries PHT search is not longer used, and the beneficial impact of PORQUE in terms of messages becomes obvious. In the case of the Uniform distribution, the number of messages is reduced by over 80% and the improvement reaches almost 70% for the Gaussian and DBLP distributions.

Latency: In highly dynamic environments exhibiting significant churn, static references are useless and DHT-lookup allows to find data. Therefore, in order to assess latency, we measure the number of hops necessary to reach a leaf node during DHT-lookup operations. We then compare the results for PHT and PORQUE searches.

Fig. 6(d), 6(e), and 6(f) show the number of hops relative to the search process for all three distributions when the cumulative number of queries grows till 50,000. The gain in performance of PORQUE over PHT binary search is not significant in this context since it can no longer use direct static references. In the case of Gaussian and DBLP distributions (Fig. 6(e) and 6(f)), PHT binary search performance is very similar to PORQUE: the former is barely better with Gaussian distribution, and worse than the latter with DBLP distribution. Since it diverts the search directly to the leaf node, our approach reduces by at least 50% the number of hops in comparison with PHT linear search.

Data insertions make the trie structure grow, so we also compare the impact of data size on the search processes with respect to different distributions. Fig. 6(g) and 6(h) show the results with Uniform and DBLP distributions respectively. In the case of Uniform distribution, the cost of a PHT linear search increases faster than PORQUE. Indeed in our system the number of hops grows logarithmically with the number of leaf nodes. PHT binary search, on the other hand, exhibits a fluctuating but overall good performance. The fluctuation of the curve can be explained since the number of iterations to find a leaf node can vary depending on the size of the searched branch of the trie and the maximum key prefix size.

Query Rate: One direct consequence of the latency reduction is the increment on the number of queries processed per time unity (Query Rate). Fig. 6(m), 6(n), 6(o) show that PORQUE outperforms both PHT search methods allowing to process a higher number of queries per simulation window. PORQUE can process up to 5 times more queries than PHT. Such a result is reached after performing $10,000$ when PORQUE search performance remains stable. Fig. 6(p) presents the average results obtained for the three data distributions.

Note that the results obtained are different from those of the previous section. Latency experiments count hops, and one hop is a DHT-lookup operation in the case of the PHT search. PORQUE, on the other hand, normally uses direct messages.

Load Balancing: One of the main advantages of the PORQUE approach is that traffic is spread over the leaf nodes of the trie. Generally, search over tree-based approaches traverse part of the trie structure to find data. The upper levels of the trie are thus highly overloaded, since queries are distributed among a small amount of nodes. This introduces bottleneck problems into the upper levels of the trie. PORQUE, however, diverts queries from the upper levels directly to the leaf nodes. Therefore, searches occur at the leaf level only and messages are distributed over a greater number of nodes. Our approach avoids bottlenecks by distributing these messages at the leaf level. In the case of PHT binary search, even if there are fewer accesses to overloaded levels, bottleneck problems still persist.

Fig. 6(i) shows how the load is distributed among the levels of the trie in the case of the DBLP distribution for a total of $50,000$ queries. The lowest levels of the trie have very similar loads, but PHT linear search introduces much more load on the 5 first levels. This has a significant impact on the average load per node. In Fig. 6(j) we can observe that the first levels incur a huge load of messages in the case of linear search, reaching $25,655$, $12,922$, and $7,148$ messages per node in the first, second and third levels respectively. PHT

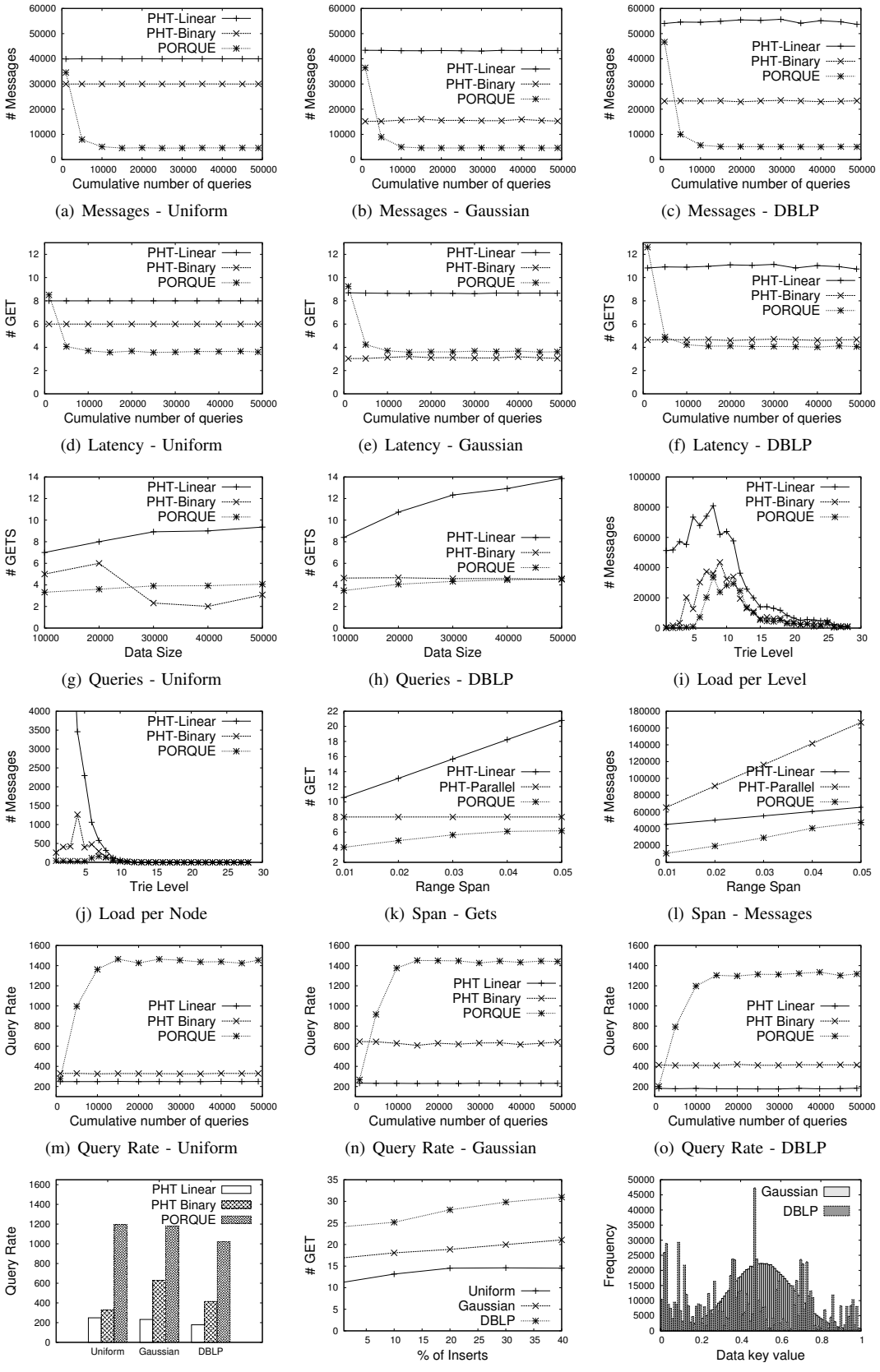


Fig. 6. Performance evaluation

binary search, on the other hand, only suffers a bottleneck at level 4 (i.e., some middle level of the trie).

Range Span analysis: We have compared different strategies for gathering the results of a range query: PHT linear, PHT binary, PHT-parallel, and PORQUE searches.

Fig. 6(k) and 6(l) show respectively the number of hops and messages for PHT linear, PHT-parallel and optimized PORQUE searches for different range sizes (spans), starting from 1% up to 5% of the data space for the Uniform distribution. Linear search has the highest latency since it accesses all the nodes sequentially. Latency increases linearly with the range span. The PHT-parallel search, however, has to perform several linear searches. Since it does so concurrently, it incurs no extra time for traversing nodes that store the data: the number of hops will be equal to the longest branch that overlap the range span. Nevertheless, the traffic in terms of messages grows linearly with the range span: this can strongly degrade the performance of the system.

We can observe that the optimized search approach of PORQUE is more efficient when compared to PHT searches, reducing latency and generating lower traffic. It is worth pointing out that the slope is less steep than for the linear search and the PHT parallel search in Fig. 6(k) and 6(l) respectively. Such an improvement is the result of parallelizing the search over the leaf nodes only, and also of starting data retrieval from every node encountered inside the range as the search progresses.

Impact of Data Insertion Operation: When data insertion operations are performed in PORQUE, neither the Search Table nor the Cache List get updated. We use range queries to disseminate table information and update their entries. However, if the insertion operation rate is similar to or greater than the query rate, the performance of PORQUE degrades. Fig. 6(q) shows the results of experiments performed where the percentage of the insertion operations increases with regard to range query rate. The range span is 1% of the data space. The figure shows a continuous but slow degradation of the performance of the PORQUE search algorithm.

V. Conclusion

This paper presents PORQUE, a new approach for range queries on top of ring-like DHT systems. By exploiting a prefix index scheme, PORQUE maintains a second ring overlay, called a Leaf Ring, composed of all the nodes that store data. We also propose an efficient method to access this Leaf Ring through information obtained from past range queries. Our solution efficiently reduces query latencies and message traffic.

Simulation results show that our system outperforms PHT [5] searches, reducing the traffic of messages by at least 50% in environments where queries significantly outnumber data insertion operations. Latency is reduced by up to 50% compared to PHT. As a direct consequence of latency reduction, PORQUE can process up to 5 times more queries than PHT. Load balancing improvements introduced in PORQUE minimize bottleneck occurrences: the load among the storage nodes is balanced, hence reducing bottleneck issues in the upper levels of the indexing structure. Simulation results using different data distributions confirm that our approach ensures good performances, even in the presence of skewed data. Although our solution is limited to scenarios where insert operations are less frequent than queries.

References

[1] R. Rodrigues and P. Druschel, "Peer-to-peer systems," *Communications of the ACM*, vol. 53, no. 10, pp. 72–82, 2010.

[2] A. I. T. Rowstron and P. Druschel, "Pastry: Scalable, decentralized object location, and routing for large-scale peer-to-peer systems," in *Middleware*, 2001, pp. 329–350.

[3] I. Stoica, R. Morris, D. Karger, M. Kaashoek, and H. Balakrishnan, "Chord: A scalable p2p lookup service for internet applications," in *SIGCOMM*, 2001, pp. 149–160.

[4] P. Maymounkov and D. Mazières, "Kademlia: A peer-to-peer information system based on the xor metric," in *IPTPS '01*, 2002, pp. 53–65.

[5] S. Ramabhadran, S. Ratnasamy, J. M. Hellerstein, and S. Shenker, "Brief announcement: prefix hash tree," in *PODC'04*, 2004, pp. 368–368.

[6] C. Zheng, G. Shen, S. Li, and S. Shenker, "Distributed segment tree: Support of range query and cover query over dht," in *IPTPS'05*, 2006.

[7] J. Gao and P. Steenkiste, "An adaptive protocol for efficient support of range queries in dht-based systems," in *Proc. of the 12th IEEE International Conference on Network Protocols*, 2004, pp. 239–250.

[8] W. Huijin and L. Yongting, "Cone: A topology-aware structured p2p system with proximity neighbor selection," *FGCN*, vol. 1, pp. 43–49, 2007.

[9] A. Datta, M. Hauswirth, R. John, R. Schmidt, and K. Aberer, "Range queries in trie-structured overlays," in *P2P'05*, 2005, pp. 57–66.

[10] Y. Tang, S. Zhou, and J. Xu, "Light: A query-efficient yet low-maintenance indexing scheme over dhts," *IEEE TKDE*, vol. 22, no. 1, pp. 59 –75, Jan 2010.

[11] Z. Hao, J. Hai, and Z. Qin, "Yarqs: Yet another range queries schema in dht based p2p network," in *Volume 02*, ser. CIT '09, 2009, pp. 51–56.

[12] N. J. A. Harvey, M. B. Jones, S. Saroiu, M. Theimer, and A. Wolman, "Skipnet: a scalable overlay network with practical locality properties," in *USITS'03*, 2003, pp. 9–9.

[13] J. Aspnes and G. Shah, "Skip graphs," in *SODA '03*, 2003, pp. 384–393.

[14] M. Mordacchini, L. Ricci, L. Ferrucci, M. Albano, and R. Baraglia, in *P2P'10*, 2010, pp. 1–10.

[15] M. Cai, M. Frank, J. Chen, and P. Szekely, "Maan: a multi-attribute addressable network for grid information services," in *The 4th Int. Workshop on Grid Computing*, 2003, pp. 184–191.

[16] A. Gupta, D. Agrawal, and A. E. Abbadi, "Approximate range selection queries in peer-to-peer systems." in *CIDR'03*, January 2003.

[17] A. R. Bharambe, M. Agrawal, and S. Seshan, "Mercury: supporting scalable multi-attribute range queries," *SIGCOMM*, vol. 34, no. 4, pp. 353–366, 2004.

[18] H. V. Jagadish, B. C. Ooi, and Q. H. Vu, "Baton: a balanced tree structure for peer-to-peer networks," in *VLDB '05*, 2005, pp. 661–672.

[19] A. GonzalezBeltran, P. Sage, and P. Milligan, "Skip tree graph: a distributed and balanced search tree for peer-to-peer networks," in *ICC '07*, June 2007, pp. 1881–1886.

[20] B. Awerbuch and C. Scheideler, "The hyperring: a low-congestion deterministic data structure for distributed environments," in *SODA '04*, 2004, pp. 318–327.

[21] A. Crainiceanu, P. Linga, J. Gehrke, and J. Shanmugasundaram, "Querying peer-to-peer networks using p-trees," in *Proc. of the 7th International Workshop on the Web and Databases*, 2004, pp. 25–30.

[22] M. Jelasity, A. Montresor, G. P. Jesi, and S. Voulgaris, "The Peersim simulator," `http://peersim.sf.net`.

[23] B. Zhao, L. Huang, J. Stribling, S. Rhea, A. Joseph, and J. Kubiatowicz, "Tapestry: a resilient global-scale overlay for service deployment," *IEEE Journal on Selected Areas in Communications*, vol. 22, no. 1, pp. 41–53, 2004.

[24] N. Hidalgo, E. Rosas, L. Arantes, O. Marin, P. Sens, and X. Bonnaire, "Dring: A layered scheme for range queries over dhts," in *Computer and Information Technology (CIT), 2011 IEEE 11th International Conference on*, 31 2011-sept. 2 2011, pp. 29 –34.