SaGe: Web Preemption for Public SPARQL Query Services
Thomas Minier, Hala Skaf-Molli, Pascal Molli

To cite this version:

HAL Id: hal-02017155
https://hal.archives-ouvertes.fr/hal-02017155
Submitted on 27 Feb 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
SAGe: Web Preemption for Public SPARQL Query Services

Thomas Minier
LS2N, University of Nantes
Nantes, France
thomas.minier@univ-nantes.fr

Hala Skaf-Molli
LS2N, University of Nantes
Nantes, France
hala.skaf@univ-nantes.fr

Pascal Molli
LS2N, University of Nantes
Nantes, France
pascal.molli@univ-nantes.fr

ABSTRACT
To provide stable and responsive public SPARQL query services, data providers enforce quotas on server usage. Queries which exceed these quotas are interrupted and deliver partial results. Such interruption is not an issue if it is possible to resume queries execution afterward. Unfortunately, there is no preemption model for the Web that allows for suspending and resuming SPARQL queries. In this paper, we propose SaGe: a SPARQL query engine based on Web preemption. SaGe allows SPARQL queries to be suspended by the Web server after a fixed time quantum and resumed upon client request. Web preemption is tractable only if its cost in time is negligible compared to the time quantum. The challenge is to support the full SPARQL query language while keeping the cost of preemption negligible. Experimental results demonstrate that SaGe outperforms existing SPARQL query processing approaches by several orders of magnitude in terms of execution time and the time for first results.

ACM Reference Format:

1 INTRODUCTION
Context and motivation: Following the Linked Open Data principles (LOD), data providers published billions of RDF triples [5, 24]. However, providing a public service that allows anyone to execute any SPARQL query at any time is still an open issue. As public SPARQL query services are exposed to an unpredictable load of arbitrary SPARQL queries, the challenge is to ensure that the service remains available despite variation in terms of arrival rate of queries and resources required to process queries.

To overcome this problem, most public LOD providers enforce a fair use service policy based on quotas [3]. According to DBpedia administrators: “A Fair Use Policy is in place in order to provide a stable and responsive endpoint for the community.” The public DBpedia SPARQL endpoint 2 Fair Use Policy prevents the execution of SPARQL longer than 120 seconds or that return more than 10000 results, with a limit of 50 concurrent connections and 100 requests per second per IP address. Quotas aim to share fairly server resources among Web clients. Quotas on communications limit the arrival rate of queries per IP. Quotas on space prevent one query to consume all the memory of the server. Quotas on time aim to avoid the convoy phenomenon [6], i.e., a long-running query will slow down a short-running one, in analogy with a truck on a single-lane road that creates a convoy of cars. The main drawback of quotas is that interrupted queries can only deliver partial results, as they cannot be resumed. This is a serious limitation for Linked Data consumers, that want to execute long-running queries [22].

Related works: Existing approaches address this issue by decomposing SPARQL queries into subqueries that can be executed under the quotas and produce complete results [4]. Finding such decomposition is hard in the general case, as quotas can be different from one server to another, both in terms of values and nature [4]. The Linked Data Fragments (LDF) approach [17, 27] tackles this issue by restricting the SPARQL operators supported by the server. For example, in the Triple Pattern Fragments (TPF) approach [27], a TPF server only evaluates triple patterns. However, LDF approaches generate a large number of subqueries and substantial data transfer.

Approach and Contributions: We believe that the issue related to time quotas is not interrupting a query, but the impossibility for the client to resume the query execution afterwards. In this paper, we propose SaGe, a SPARQL query engine based on Web preemption. Web preemption is the capacity of a Web server to suspend a running query after a time quantum with the intention to resume it later. When suspended, the state Si of the query is returned to the Web client. Then, the client can resume query execution by sending Si back to the Web server.

Web preemption adds an overhead for the Web server to suspend the running query and resume the next waiting query. Consequently, the main scientific challenge here is to keep this overhead marginal whatever the running queries, to ensure good query execution performance. The contributions of this paper are as follows:

• We define and formalize a Web preemption model that allows to suspend and resume SPARQL queries.
• We define a set of preemptable query operators for which we bound the complexity of suspending and resuming of these operations, both in time and space. This allows to build a preemptive Web server that supports a large fragment of the SPARQL query language.
• We propose SaGe, a SPARQL query engine, composed of a preemptive Web server and a smart Web client that allows executing full SPARQL queries.
• We compare the performance of the SaGe engine with existing approaches used for hosting public SPARQL services. Experimental results demonstrate that SaGe outperforms existing approaches.

1http://wiki.dbpedia.org/public-sparql-endpoint
2http://dbpedia.org/sparql

WWW ’19, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
This is the author’s version of the work. It is posted here for your personal use. Not for redistribution. The definitive Version of Record was published in Proceedings of the 2019 World Wide Web Conference (WWW’19), May 13–17, 2019, San Francisco, CA, USA, https://doi.org/10.1145/3308558.3313652.

The SaGe software and a demonstration are available at http://sage.univ-nantes.fr
by several orders of magnitude in term of the average total query execution time and the time for first results.

This paper is organized as follows. Section 2 summarizes related works. Section 3 defines the Web preemption execution model and details the SaGe server and the SaGe client. Section 4 presents our experimental results. Finally, conclusions and future work are outlined in Section 5.

2 RELATED WORKS

SPARQL endpoints. SPARQL endpoints follow the SPARQL protocol\(^1\), which “describes a means for conveying SPARQL queries and updates to a SPARQL processing service and returning the results via HTTP to the entity that requested them”. Without quotas, SPARQL endpoints execute queries using a First-Come First-Served (FCFS) execution policy [10]. Thus, by design, they can suffer from convoy effect [6]: one long-running query occupies the server resources and prevents other queries from executing, leading to long waiting time and degraded average completion time for queries.

To prevent convoy effect and ensure a fair sharing of resources among end-users, most SPARQL endpoints configure quotas on their servers. They mainly restrict the arrival rate per IP address and limit the execution time of queries. Restricting the arrival rate allows end-users to retry later, however, limiting the execution time leads some queries to deliver only partial results. To illustrate, consider the SPARQL query \(Q_1\) of Figure 1. Without any quota, the total number of results of \(Q_1\) is 35 215, however, when executed against the DBpedia SPARQL endpoint, we found only 10 000 results out of 35 215 \(^2\).

Delivering partial results is a serious limitation for a public SPARQL service. In SaGe, we deliver complete results whatever the query. In some way, quotas interrupt queries without giving the possibility to resume their execution. SaGe also interrupts queries, but allows data consumers to resume their execution later on.

Decomposing queries and restricting server interfaces. Evaluation strategies [4] have been studied for federated SPARQL queries evaluated under quotas. Queries are decomposed into a set of sub-queries that can be fully executed under quotas. The main drawbacks of these strategies are: (i) They need to know which quotas are configured. Knowing all quotas that a data provider can implement is not always possible. (ii) They can only be applied to a specific class of SPARQL queries, strongly bounded SPARQL queries, to ensure complete and correct evaluation results.

The Linked Data Fragments (LDF) [17, 27] restrict the server interface to a fragment of the SPARQL algebra, to reduce the complexity of queries evaluated by the server. LDP servers are no more compliant with the W3C SPARQL protocol, and SPARQL query processing is distributed between smart clients and LDF servers. Hartig et al. [17] formalized this approach using Linked Data Fragment machines (LDFMs). The Triple Pattern Fragments (TPF) approach [27] is one implementation of LDF where the server only evaluates paginated triple pattern queries. As paginated triple pattern queries can be evaluated in bounded time [18], the server does not suffer from the convoy effect. However, as joins are performed on the client, the intensive transfer of intermediate results leads to poor SPARQL query execution performance. For example, the evaluation of the query \(Q_1\), of Figure 1, using the TPF approach generates 507 156 subqueries and transfers 2Gb of intermediate results in more than 2 hours. The Bindings-Restricted Triple Pattern Fragments (brTPF) approach [15] improves the TPF approach by using the bind-join algorithm [14] to reduce transferred data but joins still executed by the client. In this paper, we explore how Web preemption allows the server to execute a larger fragment of the SPARQL algebra, including joins, without generating convoy effects. Processing joins on server side allow to drastically reduce transferred data between client and server and improve significantly performance. For example, SaGe executes the query \(Q_1\) of Figure 1 in less than 53s, with 553 requests and 2.1Mb transferred.

Preemption and Web preemption. FCFS scheduling policies and the convoy effect [6] have been heavily studied in operating systems. In a system where the duration of tasks vary, a long-running task can block all other tasks, deteriorating the average completion time for all tasks. The Round-Robin (RR) algorithm [19] provides a fair allocation of CPU between tasks, avoids convoy effect, reduces the waiting time and provides good responsiveness. RR runs a task for a given time quantum, then suspends it and switches to the next task. It repeatedly does so until all tasks are finished. The value of this time quantum is critical for performance: when too high, RR behaves like FCFS with the same issues, and when too low, the overhead of context switching dominates the overall performance. The action of suspending a task with the intention of resuming it later is called preemption. In public Web servers, preemption is already provided by the operating systems, but only between running tasks, excluding those waiting in the server’s queue. If we want to build a fully preemptive Web server, we need to consider the queries sent to the server as the tasks and the Web server as the resource that tasks competed to get access to. In this paper, we explore how preemption can be defined at the Web level to execute SPARQL queries.

3 WEB PREEMPTION APPROACH

We define Web preemption as the capacity of a Web server to suspend a running query after a fixed quantum of time and resume the next waiting query. When suspended, partial results and the state of the suspended query \(S_i\) are returned to the Web client \(^3\). The client can resume query execution by sending \(S_i\) back to the Web server. Compared to a First-Come First-Served (FCFS) scheduling policy, Web preemption provides a fair allocation of Web server resources across queries, a better average query completion time per query and a better time for first results [2]. To illustrate, consider three SPARQL

\(^1\)[https://www.w3.org/TR/2013/REC-sparql11-protocol-20130321/]

\(^2\)[All results were obtained on DBpedia version 2016-04.]

\(^3\)[\(S_i\) can be returned to the client or saved server-side and returned by reference.]
Web preemption

Figure 2: First-Come First-Served (FCFS) policy compared to Web Preemption (time quantum of 30s and overhead of 3s).

queries \(Q_1\), \(Q_2\), and \(Q_3\) submitted concurrently by three different Web clients. \(Q_1\), \(Q_2\), \(Q_3\) execution times are respectively 60s, 5s and 5s. Figure 2 presents a possible execution of these queries with a FCFS policy on the first line. In this case, the throughput of FCFS is \(\frac{3}{60} = 0.042\) queries per second, the average completion time per query is \(\frac{76 + 38 + 65 + 3}{4} = 65s\) and the average time for first results is also 65s. The second line describes the execution of \(Q_1 - Q_3\) using Web preemption, with a time quantum of 30s. We consider a preemption overhead of 3s (10% of the quantum). In this case, the throughput is \(\frac{3}{30} = 0.033\) query per second but the average completion time per query is \(\frac{2 + 3 \times 38 + 43}{3} = 52.3s\) and the average time for first results is approximately \(\frac{76 + 38 + 65 + 3}{4} = 37s\). If the quantum is set to 60s, then Web preemption is equivalent to FCFS. If the quantum is too low, then the throughput and the average completion time are deteriorated.

Consequently, the challenges with Web preemption are to bound the preemption overhead in time and space and determine the time quantum to amortize the overhead.

### 3.1 Web Preemption Model

We consider a preemptive Web server, hosting RDF datasets, and a smart Web client, that evaluates SPARQL queries using the server. For the sake of simplicity, we only consider read-only queries\(^7\) in this paper. The server has a pool of server workers parameterized with a fixed time quantum. Workers are in charge of queries execution. The server has also a query queue to store incoming queries

\(^7\)Preemption and concurrent updates raise issues on correctness of results.

when all workers are busy. We consider an infinite population of clients, a finite server queue and a finite number of Web workers.

The preemptive Web server suspends and resumes queries after the time quantum. A running query \(Q_1\) is represented by its physical query execution plan, denoted \(P_{Q_1}\). Suspending the execution of \(Q_1\) is an operation applied on \(P_{Q_1}\) that produces a saved state \(S_1\): \(\text{Suspend}(P_{Q_1}) = S_1\). Resuming the execution of a query \(Q_1\) is the inverse operation, it takes \(S_1\) as parameter and restores the physical query execution plan in its suspended state. Therefore, the preemption is correct if \(\text{Resume}(\text{Suspend}(P_{Q_1})) = P_{Q_1}\).

Figure 3 presents possible states of a query. The transitions are executed either by the Web server or by the client.

The Web server accepts, in its waiting queue, Web requests containing either SPARQL queries \(Q_i\) or suspended queries \(S_i\). If a worker is available, it picks a query in the waiting queue. For \(Q_i\), the worker produces a physical query execution plan \(P_{Q_i}\) using the optimize-then-execute [12] paradigm and starts its execution for the time quantum. For \(S_i\), the server resumes the execution of \(Q_i\). The time to produce or resume the physical query execution plan for a query is not deducted from the quantum.

If a query terminates before the time quantum, then results are returned to the Web client. If the time quantum is exhausted and the query is still running, then the Suspend operation is triggered, producing a state \(S_i\) which is returned to the Web client with partial results. The time to suspend the query is not deducted from the quantum. Finally, the Web client is free to continue the query execution by sending \(S_i\) back to the Web server.

The main source of overhead in this model is the time and space complexity of the Suspend and Resume operations, i.e., time to stop and save the running query followed by the time to resume the next waiting query. Our objective is to bound these complexities such that they depend only on the query complexity, i.e., the number of operations in the query plan. Consequently, the problem is to determine which physical query plans \(P_{Q_i}\) have a preemption overhead bounded in \(O(|Q_i|)\), where \(|Q|\) denotes the number of operators in the expression tree of \(P_{Q_i}\).

### 3.2 Suspending and Resuming Physical Query Execution Plans

The Suspend operation is applied to a running physical query execution plan \(P_{Q_i}\). It is obtained from the logical query execution plan of \(Q_i\) by selecting physical operators that implement operators in the logical plan [11]. A running \(P_{Q_i}\) can be represented as an expression tree of physical operators where each physical operator has a state, i.e., all internal data structures allocated by the operator. Suspending a plan \(P_{Q_i}\) requires to traverse the expression tree of the plan and save the state of each physical operator. To illustrate, consider the execution of query \(Q_2\) in Figure 4a. The state of \(P_{Q_2}\) can be saved as described in Figure 4b: the state of the Scan operator is represented by the id of the last triple read \(t_i\), the state of the Index Loop Join operator is represented by mappings pulled from the previous operator and the state of the inner scan of \(\text{tp}_1\). Suspending and resuming a physical query execution plan raise several major issues.

Suspending physical operators in constant time. Bounding the time complexity of \(\text{Suspend}(P_{Q_i})\) to \(O(|Q_i|)\) requires to save

---

**Figure 3: Possible states of a query execution in a preemptive Web Server.**

- **Smart Web Client**
  - New query
  - Saved
  - Done

- **Preemptive Web Server**
  - Worker available
  - Client sends \(Q_i\) to server
  - Waiting in queue
  - Quantum exhausted
  - Suspend
  - Executed completed
  - Server sends \(S_i\)

- **Web Server**
  - Running
  - Waiting
  - Quantum exhausted
  - Suspended
  - Server sends \(S_i\)
the state of all physical operators in constant time. However, some physical operators need to materialize data, i.e., build collections of mappings collected from other operators to perform their actions. We call these operators full-mappings operators, in opposition to mapping-at-a-time operators that only need to consume one mapping at a time from child operators\(^8\). Saving the state of full-mappings operators cannot be done in constant time. To overcome this problem, we distribute operators between a SaGe server and a SaGe smart client as follows:

- **Mapping-at-a-time** operators are suitable for Web preemption, so they are supported by the SaGe server. These operators are a subset of CompxSPARQL [17], composed of Triple Patterns, AND, UNION, FILTER and SELECT. We explain how to implement these operators server-side in Section 3.3.
- **Full-mappings** operators do not support Web preemption, so they are implemented in the SaGe smart client. These operators are: OPTIONAL, SERVICE, ORDER BY, GROUP BY, DISTINCT, MINUS, FILTER EXIST and aggregations (COUNT, AVG, SUM, MIN, MAX). We explain how to implement these operators in the smart client in Section 3.4.

As proposed in LDF [17, 27], the collaboration of the SaGe smart client and the SaGe server allows to support full SPARQL queries.

\(^8\)This is a clear reference to full-relation and tuple-at-a-time operators in database systems [11][Chapter 15.2]
Table 1: Complexities of preemption of physical query iterators. \(|id|\) and \(|tp|\) denote the size of encoding an index key and a triple pattern, respectively.

<table>
<thead>
<tr>
<th>Preemptable iterator</th>
<th>Space complexity of local state</th>
<th>Time complexity of loading local state</th>
<th>Remarks</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\pi_{\varphi}(P))</td>
<td>(O(k +</td>
<td>\varphi(P)</td>
<td>))</td>
</tr>
<tr>
<td>Index Scan (tp)</td>
<td>(O(</td>
<td>tp</td>
<td>+</td>
</tr>
<tr>
<td>Merge Join (P_1 \leftarrow P_2)</td>
<td>(O(</td>
<td>\varphi(P_1)</td>
<td>+</td>
</tr>
<tr>
<td>Index Loop Join (P \leftarrow tp)</td>
<td>(O(</td>
<td>\varphi(P)</td>
<td>+</td>
</tr>
<tr>
<td>(P_1 \cup P_2)</td>
<td>(O(1))</td>
<td>(O(1))</td>
<td>Multi-set Union</td>
</tr>
<tr>
<td>(P) Filter (W)</td>
<td>(O(</td>
<td>\varphi(P)</td>
<td>+</td>
</tr>
<tr>
<td>Server physical plan</td>
<td>(O(</td>
<td>D</td>
<td>\times \log_2(</td>
</tr>
</tbody>
</table>

Algorithm 1 presents the implementation of the Suspend and Resume functions for a pipeline of preemptable query iterators. Suspend simply stops and saves recursively each iterator in the pipeline and Resume reloads the pipeline in the suspended state using a serialized state. To illustrate, consider the SaGe page of Figure 4b. This page contains the plan that evaluates the SPARQL query \(Q_2\) of size \(|Q_2| = 4\), using index loop joins [12]. The evaluation of \(tp_2\) has been preempted after scanning 224 solution mappings. The index loop join with \(tp_1\) has been preempted after scanning two triples from the inner loop.

In the following, we review SPARQL operators implemented by the SaGe server as preemptable query iterators. We modify the regular implementations of these operators to include non interruptible section when needed. Operations left unchanged are not detailed. Table 1 resumes the complexities related to the preemption of server operators, where \(\varphi(P)\) denotes the set of variables in \(P\) as defined in [21].

Algorithm 2: A Preemptable Select Iterator

Require: \(V = \{v_1, \ldots, v_k\}\) projection variables, \(I\): predecessor in the pipeline

<table>
<thead>
<tr>
<th>Data: (\mu \leftarrow \text{nul})</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Function GetNext():</td>
</tr>
<tr>
<td>2 if (\mu = \text{nul}) then</td>
</tr>
<tr>
<td>3 (\mu \leftarrow I).GetNext()</td>
</tr>
<tr>
<td>4 {non interruptible}</td>
</tr>
<tr>
<td>5 let (\mu' \leftarrow \text{Proj}(\mu, V))</td>
</tr>
<tr>
<td>6 (\mu \leftarrow \text{nul})</td>
</tr>
<tr>
<td>7 return (\mu')</td>
</tr>
</tbody>
</table>

SELECT operator: The projection operator \(\pi_V(P)\) [25], also called SELECT, performs the projection of mappings obtained from its predecessor \(P\) according to a set of projection variables \(V = \{v_1, \ldots, v_k\}\). Algorithm 2 gives the implementation of this iterator. In this algorithm, Lines 4-7 are non-interruptible. This is necessary to ensure that the operator does not discard mappings without applying projection to them. Consequently, after the termination of \(\text{Suspend}(P)\), \(\pi_V(P)\) can be suspended in \(O(k + |\varphi(P)|)\) in space. The projection iterator reloads its local state in \(O(1)\).

Triple Pattern: The evaluation of a triple pattern \(tp\) over \(D\) \(\|tp\| \geq 21\) [21] is the set of solution mappings of \(tp\). \(\mu\) is a solution mapping if \(\mu(tp) = t\) where \(t\) is the triple obtained by replacing the variables in \(tp\) according to \(\mu\), such that \(t \in D\). The evaluation of \(\|tp\| \geq 21\) requires to read \(D\) and produce the corresponding set of solutions mappings.

The triple pattern operator supposes that data are accessed using index scans and a clustered index [11]. Algorithm 3 gives the implementation of a preemptable index scan for evaluating a triple pattern. We omitted the \(\text{Stop}\) operation, as the iterator does not need to do any action to stop. The \(\text{Save}\) operation stores the triple pattern \(tp\) and the index’s \(id(t)\) of the last triple \(t\) read. Thus, suspending an index scan is done in constant time and the size of the operator state is in \(O(|tp| + |id(t)|)\).

The \(\text{Load}\) operation uses the saved index’s \(id(t)\) on \(tp\) to find the last matching RDF triple read. The time complexity of this operation predominates the time complexity of \(\text{Resume}\). This complexity depends on the type of index used. With a B+-tree [7], the time complexity of resuming is bound to \(O(\log_b(|D|))\). B+-tree indexes are offered by many RDF data storage systems [8, 20, 28].

Algorithm 3: A Preemptable Index Scan Iterator, evaluating a triple pattern \(tp\) using an index

Require: \(tp\): triple pattern, \(D\): RDF dataset queried, \(I_{tp}\): clustered index over \(tp\)

<table>
<thead>
<tr>
<th>Data: (t): last matching RDF triple read</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Function GetNext():</td>
</tr>
<tr>
<td>2 {non interruptible}</td>
</tr>
<tr>
<td>3 (t \leftarrow \text{next RDF triple matching } tp \text{ in } D)</td>
</tr>
<tr>
<td>4 let (\mu \leftarrow \text{set of solutions mappings such as } \mu(tp) = t)</td>
</tr>
<tr>
<td>5 return (\mu)</td>
</tr>
<tr>
<td>6 Function Save():</td>
</tr>
<tr>
<td>7 return ((tp, Id(t)))</td>
</tr>
<tr>
<td>8 Function Load(id):</td>
</tr>
<tr>
<td>9 (t \leftarrow I_{tp}.\text{Locate}(id)) // Locate the last triple read</td>
</tr>
</tbody>
</table>

Basic Graph patterns (AND): A Basic Graph pattern (BGP) evaluation corresponds to the natural join of a set of triple patterns. Join operators fall into three categories [11]: (1) Hash-based joins e.g., Symmetric Hash join or XJoin, (2) Sort-based joins, e.g., (Sort) Merge join, (3) Loop joins, e.g., index loop or nested loop joins.

Hash-based joins [12, 29] operators are not suitable for preemption. As they build an in-memory hash table on one or more inputs to evaluate the join, they are considered as full-mappings operators. The sort-based joins and loop joins are mapping-at-a-time operators; consequently, they could be preempted with low overhead. However, for sort-based joins, we can only consider merge joins, i.e., joins inputs are already sorted on the join attribute. Otherwise, this will require to perform an in-memory sort on the inputs. In the following, we present algorithms for building preemptable Merge join and preemptable Index Loop join operators:
Preemptable merge join: The Merge join algorithm merges the solutions mappings read from two join inputs, i.e., others operators sorted on the join attribute. SAGe extends the classic merge join [12] to the Preemptable Merge join iterator as shown in Algorithm 4. Basically, its Stop, Save and Load functions recursively stop, save or load the joins inputs, respectively. Thus, the only internal data structures holds by the iterator are the two inputs and the last sets of solution mappings read from them. As described in Table 1, the local state of a merge join is resumable in constant time, while its space complexity depends on the size of two sets of solution mappings saved.

Preemptable Index Loop join: The Index Loop join algorithm [12] exploits indexes on the inner triple pattern for efficient join processing. This algorithm has already been used for evaluating BGP s in [16]. SAGe extends the classic Index Loop joins to a Preemptable Index join iterator (PIJ-Iterator) presented in Algorithm 5. To produce solutions, the iterator executes the same steps repeatedly until all solutions are produced: (1) It pulls solutions mappings \( \mu_c \) from its predecessor. (2) It applies \( \mu_c \) to \( tp_l \) to generate a bound pattern \( b = \mu_c(tp_l) \). (3) If \( b \) has no solution mappings in \( D \), it tries to read again from its predecessor (jump back to Step 1). (4) Otherwise, it reads triple matching \( b \) in \( D \), produces the associated set of solution mappings and then goes back to Step 1. A PIJ-Iterator supports preemption through the Stop, Save and Load functions. The non-interruptible section of GetNext() only concerns a scan in the dataset. Therefore, in the worse case, the iterator has to wait for one index scan to complete before being interruptible. As with the merge join, the Save and Load functions need to stop and resume the left join input. The latter also needs to resume an Index Scan, which can be resumed in \( O(\log_b(|D|)) \).

Regarding the saved state, the iterator saves the triple pattern joined with the last set of solution mappings pulled from the predecessor and an index scan. For instance, in Figure 4b, the saved state of the join of \( tp_2 \) and \( tp_1 \) is \( \mu_c = \{ ?v1 \mapsto \text{wm:Product12}, ?v3 \mapsto "4356" \} \).

UNION fragment: The UNION fragment is defined as the union of solution mappings from two graph patterns. We consider a multi-set union semantic, as set unions require to save intermediate results to remove duplicates and thus cannot be implemented as mapping-wise evaluation. A multi-set union is equivalent to the sequential evaluation of all graph patterns in the union. When preemption occurs, all graph patterns in the union are saved recursively. So the union has no local state on its own. Similarly, to resume an union evaluation, each graph pattern is reloaded recursively.

FILTER fragment: A SPARQL FILTER is denoted \( F = \sigma_R(P) \), where \( P \) is a graph pattern and \( R \) is a built-in filter condition. The evaluation of \( F \) yields the solutions mappings of \( P \) that verify \( R \). Some filter conditions require collection of mappings to be evaluated, like the EXISTS filter which requires the evaluation of a graph pattern. Consequently, we limit the filter condition to pure logical expressions (\( =, <, \geq, \land \), etc) as defined in [21, 25]. The preemption of a FILTER is similar to those of a projection. We only need to suspend or resume \( P \), respectively. For the local state, we need to save the last solution mappings pulled from \( P \) and \( R \).

Table 1 summarizes the complexities of Suspend and Resume in time and space. The space complexity to Save a physical plan is established to \( O(|Q| \times \log_b(|D|)) \). We supposed that \( |\var{P}| \) and the number of \( |tp| \) to save are close to \(|Q|\). However, the size of index IDs are integers that can be as big as the number of RDF triples in \( D \). Hence, they can be encoded in at most \( \log_b(|D|) \) bits. The time complexity to Resume a physical plan is \( O(|Q| \times \log_b(|D|)) \). It is higher than the time complexity of Suspend, as resuming index

---

**Algorithm 4: A Preemptable Merge Join Iterator \( I_l \): a preemptable join operator used by SAGe for BGP evaluation**

Require: \( I_{left} \): the outer join input, \( I_{right} \): the inner join input, \( \mu_l \): the last element read from \( I_{left} \), \( \mu_r \): the last element read from \( I_{right} \). RDF dataset queried.

1. **Function Stop:**
   1. \( I_{left}.\text{Stop()} \)
   2. \( I_{right}.\text{Stop()} \)

2. **Function Save:**
   1. \( I_{left}.\text{Save()} \)
   2. \( I_{right}.\text{Save()} \)

3. **Function Load:**
   1. \( I_{left}.\text{Load}(s_l) \)
   2. \( I_{right}.\text{Load}(s_r) \)

4. **Return** \( \langle s_l, s_r, \mu_l, \mu_r \rangle \)

---

**Algorithm 5: A Preemptable Index Join Iterator \( I_l \): a preemptable index join operator used by SAGe for BGP evaluation**

Require: \( I_{left} \): the iterator responsible for the evaluation of the outer join input, \( tp_1 \): the inner join input, \( D \): RDF dataset queried.

1. **Function Open:**
   1. \( I_{left}.\text{Open()} \)
   2. \( \mu_c \leftarrow \text{nil} \)
   3. \( \text{ift} \leftarrow \text{PreemptableIndexScanIterator over } \emptyset \)

2. **Function GetNext:**
   1. \( \text{while } \neg \text{ift.\text{HasNext}}() \text{ do} \)
   2. \( \mu_c \leftarrow \text{ift.\text{GetNext()}} \)
   3. \( \text{ift} \leftarrow \text{PreemptableIndexScanIterator over } \mu_c(tp_r) \)
   4. \( \mu_r \leftarrow \text{ift.\text{GetNext()}} \)
   5. \( \mu \leftarrow \mu \cup \mu_r \)

6. **Function Load:**
   1. \( tp_r \leftarrow tp_r \)
   2. \( \mu_r \leftarrow \mu_r \)
   3. \( \text{ift} \leftarrow \text{PreemptableIndexScanIterator over } \mu_c(tp_r) \)
   4. \( \text{ift.\text{Load()}} \)

7. **Function Save:**
   1. \( \text{ift} \leftarrow \text{the last triple read} \)
   2. \( \mu_c \leftarrow \text{ift.\text{Save()}} \)

8. **Function Stop:**
   1. \( I_{left}.\text{Stop()} \)
   2. \( I_{right}.\text{Stop()} \)

9. **Return** \( \langle tp_r, \mu_c, t \rangle \)
The client needs to pay the network latency twice per request. (2) The query is not fully supported by the SaGe server. Then, the client decomposes the query into a set of subqueries supported by the server, evaluate each subquery as in the first case, and recombine intermediate results to produce query results.

Data transfer: We also distinguish two cases: (1) If the query is fully supported by the SaGe server, the only overhead is the size of the saved plan $S_1$ multiplied by the number of requests. Notice that the saved plan $S_1$ can be returned by reference or by value, i.e., saved server-side or client-side. (2) If the query is not fully supported by the SaGe server, then the client decomposes the query and recombine results of subqueries. Among these results, some are intermediate results and part of the client overhead.

Consequently, the challenge for the smart client is to minimize the overhead in terms of the number of requests and transferred data. The data transfer could be reduced by saving the plans server-side rather than returning it to clients.\(^\text{10}\) However, the main source of client overhead comes from the decomposition made to support full SPARQL queries, as these queries increase the number of requests sent to the server. Some decompositions are more costly than others. To illustrate, consider the evaluation of $(P_1 \ {\text{OPTIONAL \ } P_2})$ where $P_1$ and $P_2$ are expressions supported by the SaGe server. A possible approach is to evaluate $[P_1 \ ]_D$ and $[P_2 \ ]_D$ on the server, and then perform the left outer join on the client. This strategy generates only two subqueries but materializes $[P_2 \ ]_D$ on client. If there is no join results, $[P_2 \ ]_D$ are just useless intermediate results.

Another approach is to rely on a nested loop join approach: evaluates $[P_1 \ ]_D$ and for each $\mu_1 \in [P_1 \ ]_D$, if $\mu_2 \in \mu_1 [P_2]_D$ then $\mu_1 \cup \mu_2$ are solutions to $P_1 \Join \ast \mu_2$. Otherwise, only $\mu_1$ is a solution to the left-join. This approach sends at least as many subqueries to the server than there is solutions to $[P_1 \ ]_D$.

To reduce the communication, the SaGe client implements BindLEFTJOINS to process local join in a block fashion, by sending unions of BGPs to the server. This technique is already used in federated SPARQL query processing [26] with bound joins and in BrTPF [15]. Consequently, the number of request to the SaGe server is reduced by a factor equivalent to the size of a block of mappings. However, the number of requests sent still depends on the cardinality of $P_1$.

Consequently, we propose a new technique, called OptJOIN, for optimizing the evaluation of a subclass of left-joins. The approach relies on the fact that:

$$\{P_1 \Join \ast \mu_2 \} \subseteq \{P_1 \Join \ast \mu_2 \} \cup \{\mu_1 \cup \mu_2 \}$$

So we can deduce that: $\{P_1 \Join \ast \mu_2 \} \subseteq \{P_1 \Join \ast \mu_2 \} \cup \{P_1 \Join \ast \mu_2 \}$. If both $P_1$ and $P_2$ are evaluable by the SaGe server, then the smart client computes the left-join as follows. First, it sends the query $(P_1 \Join \ast \mu_2) \cup P_1$ to the server. Then, for each mapping $\mu_2$ received, it builds local materialized views for $[P_1 \Join \ast \mu_2 \ ]_D$ and $[P_1 \ ]_D$. The client knows that $\mu \in [P_1 \Join \ast \mu_2 \ ]_D$ if $\text{dom}(\mu) \subseteq \text{var}(P_1 \Join \ast \mu_2)$, otherwise $\mu \in [P_1 \ ]_D$. Finally, the client uses the views to process the left-join locally. With this technique, the client only use one subquery to evaluate the left-join and, in the worst case, it transfers $[P_1 \Join \ast \mu_2 \ ]_D$ as additional intermediate results.

To illustrate, consider query $Q_3$ from Figure 5a. $Q_3$, with 88334 solutions. The cardinality of $tp1 \Join \ast tp2$ is also of 88334, as every

---

\(^{10}\)In our implementation, we choose to save plans client-side to tolerate server failures.
actor has a birth place. This is the worse case for a BindLeftJoin, which will require additional requests to evaluate the left join. However, with an OptJoin, the client is able to evaluate Q3 in approximately 500 requests.

We implement both BindLeftJoin and OptJoin as physical operators to evaluate OPTIONS, depending on the query. We also implement regular BindJoin for processing SERVICE queries.

4 Experimental Study

We want to empirically answer the following questions: What is the overhead of Web preemption in time and space? Does Web preemption improves the average workload completion time? Does Web preemption improves the time for first results? What are the client overheads in terms of numbers of requests and data transfer? We use Virtuoso as the baseline for comparing with SPARQL endpoints, with TPF and BrTPF as the baselines for LDF approaches.

We implemented the SaGe client in Java, using Apache Jena\textsuperscript{11}. As an extension of Jena, SaGe is just as compliant with SPARQL 1.1. The SaGe server is implemented as a Python Web service and uses HDT\textsuperscript{9} for storing data. Notice that the current implementation of HDT cannot ensure \( \log_2(n) \) access time for all triple patterns, like \((a \sim p \circ b)\). This impacts negatively the performance of SaGe when resuming some queries. The code and the experimental setup are available on the companion website\textsuperscript{12}.

4.1 Experimental setup

Dataset and Queries: We use the Waterloo SPARQL Diversity Benchmark (WatDiv\textsuperscript{13}) \textsuperscript{1}. We re-use the RDF dataset and the SPARQL queries from the BrTPF \textsuperscript{15} experimental study\textsuperscript{14}. The dataset contains \( 10^7 \) triples and queries are arranged in 50 workloads of 193 queries each. They are SPARQL conjunctive queries with STAR, PATH and SNOWFLAKE shapes. They vary in complexity, up to 10 joins per query with very high and very low selectivity. 20% of queries require more than \( \approx 30s \) to be executed using the virtuoso server. All workloads follow nearly the same distribution of query execution times as presented in Figure 6. The execution times are measured for one workload of 193 queries with SaGe and an infinite time quantum.

Approaches: We compare the following approaches:

\begin{itemize}
  \item \textbf{SaGe:} We run the SaGe query engine with various time quantum: 75ms and 1s, denoted SaGe-75ms and SaGe-1s respectively. HDT indexes are loaded in memory while HDT data is stored on disk.
  \item \textbf{Virtuoso:} We run the Virtuoso SPARQL endpoint \textsuperscript{8} (v7.2.4), without any quotas or limitations.
  \item \textbf{TPF:} We run the standard TPF client (v2.0.5) and TPF server (v2.2.3) with HDT files as backend (same settings as SaGe).
  \item \textbf{BrTPF:} We run the BrTPF client and server used in \textsuperscript{15}, with HDT files as backend (same settings as SaGe). BrTPF is currently the LDF approach with the lowest data transfer \textsuperscript{15}.
\end{itemize}

Servers configurations: We run all the servers on a machine with Intel(R) Xeon(R) CPU E7-8870@2.10GHz and 1.5TB RAM.

Clients configurations: In order to generate load over servers, we rely on 50 clients, each one executing a different workload of queries. All clients start executing their workload simultaneously. The clients access servers through HTTP proxies to ensure that client-server latency is kept around 50ms.

Evaluation Metrics: Presented results correspond to the average obtained of three successive execution of the queries workloads. (1) Workload completion time (WCT): is the total time taken by a client to evaluate a set of SPARQL queries, measured as the time between the first query starting and the last query completing. (2) Time for first results (TFR) for a query: is the time between the query starting and the production of the first query’s results. (3) Time preemption overhead: is the total time taken by the server’s Suspend and Resume operations. (4) Number of HTTP requests and data transfer: is the total number of HTTP requests sent by a client to a server and the number of transferred data when executing a SPARQL query.

4.2 Experimental results

We first ensure that the SaGe approach yield complete results. We run both Virtuoso and SaGe and verify that, for each query, SaGe delivers complete results using Virtuoso results as ground truth.

What is the overhead in time of Web preemption? The overhead in time of Web preemption is the time spent by the Web server
for suspending a running query and the time spent for resuming the next waiting query. To measure the overhead, we run one workload of queries using SaGe-75ms and measure time elapsed for the Suspend and Resume operations. Figure 7 shows the overhead in time for SaGe: Suspend and Resume operations using different sizes of the WatDiv dataset. Generally, the size of the dataset does not impact the overhead, which is around 1ms for Suspend and 1.5ms for Resume. As expected, the overhead of the Resume operation is greater than the one of the Suspend operation, due to the cost of resuming the Index scans in the plan. With a quantum of 75ms, the overhead is ≈ 3% of the quantum, which is negligible.

**What is the overhead in space of Web preemption?** The overhead in space of the Web preemption is the size of saved plans produced by the Suspend operation. According to Section 3.3, we determined that the SaGe physical query plans can be saved in \(O(|Q|)\). To measure the overhead, we run a workload of queries using SaGe-75ms and measure the size of saved plans. Saved plans are compressed using Google Protocol Buffers 15. Table 2 shows the overhead in space for SaGe. As we can see the size of a saved plan remains very small, with no more than 6 kb for a query with ten joins. Hence, this space is indeed proportional to the size of the plan suspended.

**Does Web preemption improve the average workload completion time?** To enable Web preemption, the SaGe server has a restricted choice of physical query operators, thus physical query execution plans generated by the SaGe server should be less performant than those generated by Virtuoso. This tradeoff only make sense if the Web preemption compensates the loss in performance. Compensation is possible only if the workload alternates long-running and short running queries. In the setup, each client runs a different workload of 193 queries that vary from 0 to 100 queries, following an exponential distribution. All clients execute their workload concurrently and start simultaneously. We experiment up to 50 concurrent clients by step of 5 clients. As there is only one worker on the Web server, and queries execution time vary, this setup is the worst case for Virtuoso.

Figure 8 shows the average workload completion time obtained for all approaches, with a logarithmic scale. As expected, Virtuoso is significantly impacted by the convoy effect and delivers the worse WTC after 20 concurrent clients. TPF and BrTPF avoid the convoy and behave similarly. BrTPF is more performant thanks to its bind-join technique that group requests to the server. SaGe-75ms and SaGe-1s avoid the convoy effect and delivers better WTC than TPF and BrTPF. As expected, increasing the time quantum also increases the probability of convoy effect and, globally, SaGe-75ms offers the best WTC. We rerun the same experiment with 4 workers for SaGe-75ms and Virtuoso. Figure 9 shows the average workload completion time obtained for both approaches. As we can see, both approaches benefit of the four workers. However, Virtuoso still suffers from the convoy effect.

**Does Web preemption improve the time for the first results?** The Time for first results (TFR) for a query is the time between the query starting and the production of the first query’s results. Web preemption should provides better time for first results. Avoiding convoy effect allows to start queries earlier and then get results earlier. We rerun the same setup as in the previous section and measure the time for first results (TFR). Figure 10 shows the results with a linear scale. As expected, Virtuoso suffers from the convoy effect that degrades significantly the TFR when the concurrency

---

15https://developers.google.com/protocol-buffers/
We generate new WatDiv queries from one set of 193 queries, using the following protocol. For each query \(Q = \{t_{p_1}, \ldots, t_{p_k}\}\), we select \(t_{p_k} \in Q\) with the highest cardinality, then we generate \(Q' = t_{p_k} \cup (Q \setminus t_{p_k})\). Such queries verify conditions for using `BindLeftJoin` and `OptJoin`. They are challenging for the `BindLeftJoin` as they generate many mappings; they are also challenging for the `OptJoin` as all joins yield results. Table 3(b) shows the results when evaluating `OPTIONAL` queries with `OptJoin` and `BindLeftJoin` approaches. We observe that, in general, `OptJoin` outperforms `BindLeftJoin`. Furthermore, `OptJoin` is improved when using a higher quantum, as the single subquery sent is evaluated more quickly. This is not the case for `BindLeftJoin`, as the number of requests still depends on the number of intermediate results.

Finally, we re-use 166 SELECT queries from `Feasible` [23] with the DBpedia 3.5.1 dataset, which were generated from real-users queries. We excluded queries that time-out, identified in [23], and measure the number of requests sent to the server. Table 3(a) shows the results. Of course, Virtuoso just sends 166 requests to the server. As we can see the ratio of requests between SaGe-75 and TPF is nearly the same as the previous experiment. However, the difference between SaGe-1s and SaGe-75ms has clearly decreased, because most requests sent are produced by the decomposition of `OPTIONAL` and `FILTER` and not by the evaluation of BGPs.

### 5 Conclusion and Future Work

In this paper, we demonstrated how Web preemption allows not only to suspend SPARQL queries, but also to resume them. This opens the possibility to efficiently execute long-running queries with complete results. The scientific challenge was to keep the Web preemption overhead as low as possible; we demonstrated that the overhead of large a fragment of SPARQL can be kept under 2ms.

Compared to SPARQL endpoint approaches without quotas, SaGe avoids the convoy effect and is a winning bet as soon as the queries of the workload vary in execution time. Compared to LDF approaches, SaGe offers a more expressive server interface with joins, union and filter evaluated server side. Consequently, this considerably reduces data transfer and the communication cost, improving the execution time of SPARQL queries.

SaGe opens several perspectives. First, for the sake of simplicity, we made the hypothesis of read-only datasets in this paper. If we allow concurrent updates, then a query can be suspended with a version of a RDF dataset and resumed with another version. The main challenge is then to determine which consistency criteria can be ensured by the preemptive Web server. As SaGe only accesses the RDF dataset when scanning triple patterns, it could be possible to compensate concurrent updates and deliver, at least, eventual consistency for query results. Second, we aim to explore how the interface of the server can be extended to support named graphs and how the optimizer of the smart client can be tuned to produce better query decompositions, especially in the case of federated SPARQL query processing. Third, we plan to explore if more elaborated scheduling policies could increase performance. Finally, the Web preemption model is not restricted to SPARQL. An interesting perspective is to design a similar approach for SQL or GraphQL.

### Acknowledgments

This work has been partially funded through the FaBuLA project, part of the AtlanSTIC 2020 program.