Department of Informatics, University of Zürich

BSc Thesis

An Adaptive Index for Hierarchical Distributed Database Systems

Rafael Kallis

Matrikelnummer: 14-708-887

Email: rk@rafaelkallis.com

February 1, 2018
supervised by Prof. Dr. Michael Böhlen and Kevin Wellenzohn

University of Zurich[UZH]
Department of Informatics

1 Introduction

Frequently adding and removing data from hierarchical indexes causes them to repeatedly grow and shrink. A single insertion or deletion can trigger a sequence of structural index modifications (node insertions/deletions) in a hierarchical index. Skewed and update-heavy workloads trigger repeated structural index updates over a small subset of nodes to the index.

Informally, a frequently added or removed node is called *volatile*. Volatile nodes deteriorate index update performance due to two reasons. First, frequent structural index modifications are expensive since they cause many disk accesses. Second, frequent structural index modifications also increase the likelihood of conflicting index updates by concurrent transactions. Conflicting index updates further deteriorate update performance since concurrency control protocols need to resolve the conflict.

Wellenzohn et al. [4] propose the Workload-Aware Property Index (WAPI). The WAPI exploits the workloads' skewness by identifying and not removing volatile nodes from the index, thus significantly reducing the number of expensive structural index modifications. Since fewer nodes are inserted/deleted, the likelihood of conflicting index updates by concurrent transactions is reduced.

When the workload characteristics change, new index nodes can become volatile while others cease to be volatile and become *unproductive*. Unproductive index nodes slow down queries as traversing an unproductive node is useless, because neither the node itself nor any of its descendants contain an indexed property and thus cannot yield a query match. Additionally, unproductive nodes occupy storage space that could otherwise be reclaimed. lf the workload changes frequently, unproductive nodes quickly accumulate in the index and the query performance deteriorates over time. Therefore, unproductive nodes must be cleaned up to keep query performance stable over time and reclaim disk space as the workload changes.

Wellenzohn et al. [4] propose periodic Garbage Collection (GC), which traverses the entire index subtree and prunes all unproductive index nodes at once. Additionally we propose Query-Time Pruning (QTP), an incremental approach to cleaning up unproductive nodes in the index. The idea is to turn queries into updates. Since Oak already traverses unproductive nodes as part of query processing, these nodes could be pruned at the same time. In comparison to GC, with QTP only one query has to traverse an unproductive node, while subsequent queries can skip this overhead and thus perform better.

The goal of this BSc thesis is to study, implement, and empirically compare GC and QTP as proposed by [4] in Apache Jackrabbit Oak (Oak).

2 Background

2.1 Apache Jackrabbit Oak (Oak)

Oak is a hierarchical distributed database system which makes use of a hierarchical index. Multiple transactions can work concurrently by making use of Multiversion Concurrency Control (MVCC) [3], a commonly used optimistic concurrency control technique [2].

Figure 1 depicts Oak's multi-tier architecture. Oak embodies the *Database Tier*. Whilst Oak is responsible for handling the database logic, it stores the actual data on MongoDB[1], labeled as *Persistence Tier*. On the other end, applications can make use of Oak as shown in Figure 1 under *Application Tier*. w/One such application is Adobe's enterprise content management system (CMS), the Adobe Experience Manager[2].

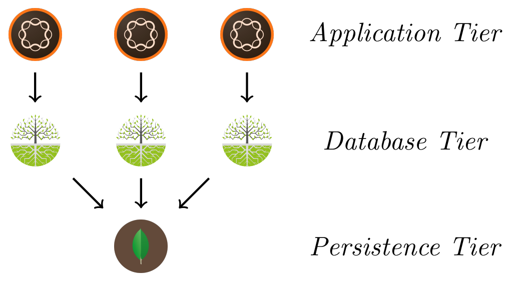


Figure 1: Apache Jackrabbit Oak's system architecture.

2.2 Workload Aware Property Index (WAPI)

Oak mostly executes content-and-structure (CAS) queries [1], defined as follows.

Definition 1. (CAS-Query): Given node m, property k and value v, a CAS query $Q(k, v, m)$ returns all descendants of m which have k set to v, i.e

$$Q(k, v, m) = \{\ n \mid n[k] = v \wedge n \in desc(m)\ \}$$

The WAPI is a hierarchical index and indexes the properties of nodes in order to answer CAS-queries efficiently. Additionally, it takes into account if an index node is volatile before performing structural index modifications. If a node is considered volatile, we do not remove it from the index.

Volatility is the measure which is used by the WAPI in order to distinguish when to remove a node or not from the index.

Wellenzohn et al. [4] propose to look at the recent transactional workload to check whether a node n is volatile. The workload on Oak instance O_i is represented by a sequence $H_i = \langle \ldots, G^a, G^b, G^c \rangle$ of snapshots, called a history. Let t_n be the current time

[1]https://www.mongodb.com/what-is-mongodb
[2]http://www.adobe.com/marketing-cloud/experience-manager.html

and $t(G^b)$ be the point in time snapshot G^b was committed, $N(G^a)$ is the set of nodes which are members of snapshot G^a. $pre(G^b)$ is the predecessor of snapshot G^b in H_i.

Node n is volatile iff n's volatility count is at least τ, called volatility threshold. The volatility count of n is defined as the number of times n was added or removed from snapshots in a sliding window of length L over history H_i. Let n^i denote version i of node n that belongs to the node set $N(G^i)$ of snapshot G^i. Given two snapshots G^a and G^b we write n^a and n^b to emphasize that nodes n^a and n^b are two versions of the same node n, i.e, they have the same absolute path from the root node.

Definition 2. (Volatility Count): The volatility count $vol(n)$ of node n is the number of times node n was added or removed from snapshots contained in a sliding window with length L over history H_i.

$$\begin{aligned} vol(n) = |\{G^b | G^b \in H_i \wedge t(G^b) \in [t_{n-L+1}, t_n] \wedge \exists G^a [\\ G^a = pre(G^b) \wedge ([n^a \notin N(G^a) \wedge n^b \in N(G^b)] \vee \\ [n^a \in N(G^a) \wedge n^b \notin N(G^b)])]\}| \end{aligned} \tag{1}$$

Definition 3. (Volatile Node): Node n is volatile iff n's volatility count (see Definition 2) is greater or equal than the volatility threshold τ, i.e

$$volatile(n) \iff vol(n) \geq \tau$$

3 Unproductive Nodes

When time passes and the database workload changes, volatile nodes cease to be volatile and they become unproductive.

Definition 4. (Unproductive Node): Node n is unproductive iff n, and any descendant of n, is neither matching nor volatile, i.e

$$unproductive(n) \iff \nexists m(m \in (\{n\} \cup desc(n)) \wedge (volatile(m) \vee *m[k] \neq v))$$

Example 1. Consider the snapshots depicted in Figure 2. Assume $H_h = \langle G^0, G^1, G^2, G^3, G^4, G^5, G^6 \rangle$. O_h executes transactions $T_1, T_2, T_3, T_4, T_5, T_6$. Snapshot G^0 was committed at time $t(G^0) = t$. Given snapshot G^0, transaction T_1 adds property $x = 1$ `/a/b/d` and commits snapshot G^1 at time $t(G^1) = t + 1$. Next, transaction T_2 removes property x from `/a/b/d` given snapshot G^1 and commits snapshot G^2 at time $t(G^2) = t + 2$. The index nodes are not pruned during T^2 since they are volatile. Transaction T_3 adds property $x = 1$ to `/a/c/e` given G^2 and commits G^3 at time $t(G^3) = t + 3$. Notice how `/i/x/1/a/b/d` and `/i/x/1/a/b` are the only unproductive index nodes and `/i/x/1/a/c/e` as well as `/i/x/1/a/c/e` are the only volatile index nodes in G^3. Transaction T^4 again adds property $x = 1$ to `/a/b/d` given snapshot G^3 and commits snapshot G^4 at time $t(G^4) = t + 4$. In G^4, nodes `/i/x/1/a/b/d` and `/i/x/1/a/b` are not unproductive anymore since `/i/x/1/a/b/d`'s content node is matching. Transaction T^5 again removes property x from `/a/b/d` given snapshot G^4 and commits snapshot G^5 at time $t(G^5) = t + 5$. Since nodes `/i/x/1/a/b/d` and `/i/x/1/a/b` are not volatile, they are pruned from the property index during T^5. Finally transaction T^6 removes property $x = 1$ from `/a/c/e` given snapshot G^5 and commits snapshot G^6 at time $t(G^6) = t + 6$. Index nodes `/i/x/1/a/c/e` and `/i/x/1/a/c` are pruned from the property index during T^6 because they are not volatile.

Unproductive nodes slow down queries. During query execution, traversing an unproductive node is useless, because neither the node itself nor any of its descendants contains an indexed property and therefore cannot contribute a query match. We formalize the statements above into the following scientific questions:

Q_1 Query execution runtime increases over time.

Q_2 Unproductive nodes account for the increase in query execution runtime.

We recorded the query execution runtime throughout the experiment and present the data below. Figures 3a and 3b show the recorded query runtime over time as observed from the synthetic and AEM dataset respectively. Figures 3c and 3d show the recorded query runtime over update operations from the synthetic and AEM dataset respectively.

Next, we present data regarding the type of index nodes being traversed while executing a query. Figures 4a to 4d depict the number of traversed volatile and unproductive index nodes during query execution with respect to time and update operations from our datasets. Additionally, Figures 4e to 4h show the density of volatile and unproductive nodes over time and update operations from our datasets.

The collected data shows a significant increase in query execution runtime over the recorded time. The number of unproductive nodes grows over time while the number of volatile nodes seem to stall after 30 seconds. We also observe that the majority of traversed nodes during query execution are unproductivea and therefore must likely account for the increase in query runtime. Interestingly, we also observe the volatile and unproductive node density to converge below 0.2 and above 0.8 respectively.

3.1 Volatility threshold (τ)

3.2 Sliding window length (L)

4 Periodic Garbage Collection (GC)

5 Query Time Pruning (QTP)

6 Experimental Evaluation

References

[1] C. Mathis, T. Härder, K. Schmidt, and S. Bächle. XML indexing and storage: fulfilling the wish list. *Computer Science - R&D*, 30(1):51–68, 2015.

[2] M. T. Özsu and P. Valduriez. *Principles of Distributed Database Systems, Third Edition*. Springer, 2011.

[3] G. Weikum and G. Vossen. *Transactional Information Systems: Theory, Algorithms, and the Practice of Concurrency Control and Recovery*. Morgan Kaufmann, 2002.

[4] K. Wellenzohn, M. Böhlen, S. Helmer, M. Reutegger, and S. Sakr. A Workload-Aware Index for Tree-Structured Data. To be published.

$$\begin{tikzcd} G^0 \arrow[r, "T_1"] & G^1 \arrow[r, "T_2"] & G^2 \arrow[r, "T_3"] & G^3 \arrow[r, "T_4"] & G^4 \arrow[r, "T_5"] & G^5 \arrow[r, "T_6"] & G^6 \end{tikzcd}$$

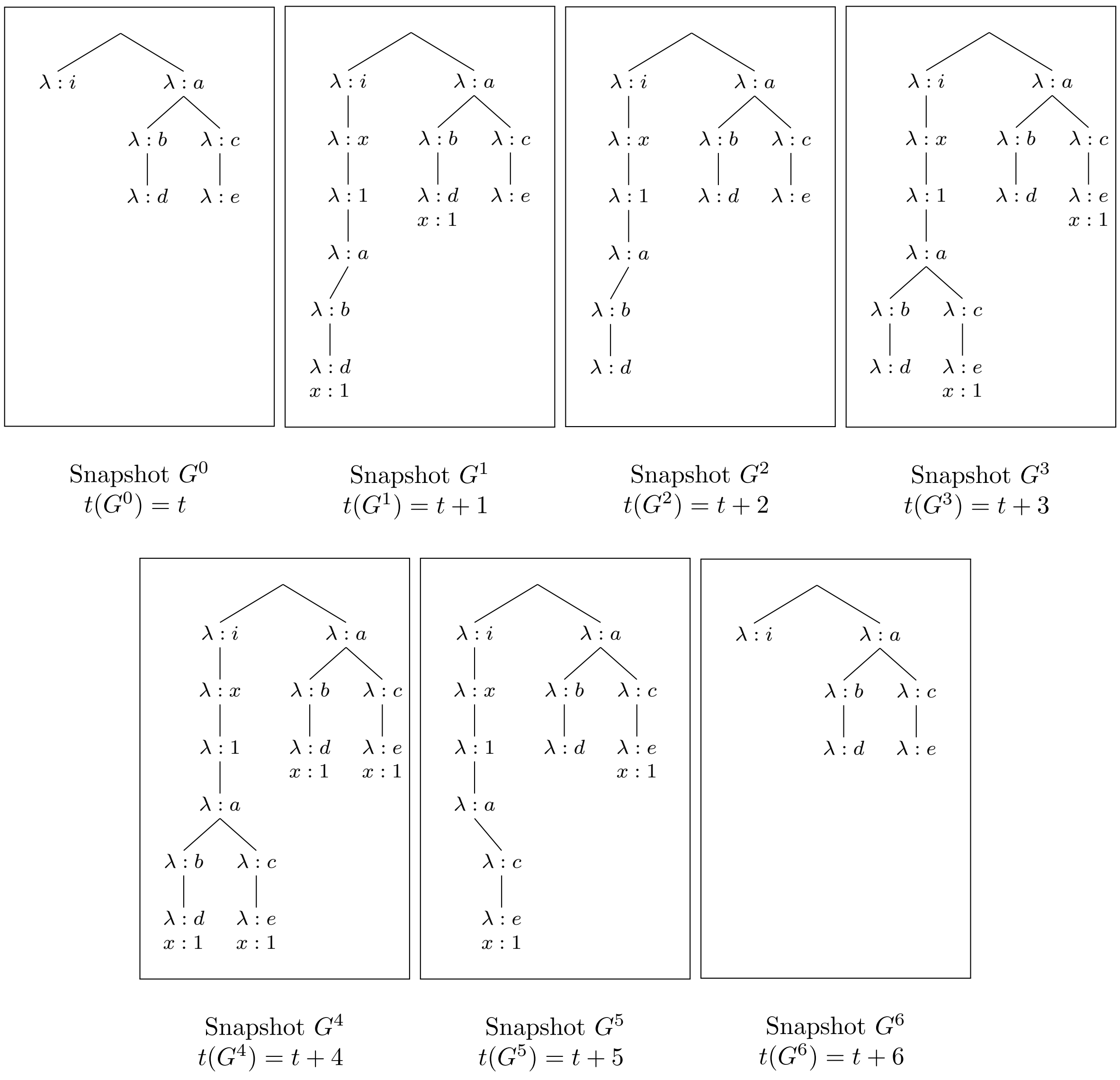


Figure 2: Unproductive Nodes. `/i/x/1/a/b/d` and `/i/x/1/b` are unproductive in snapshot G^3. They are not volatile and don't match either.

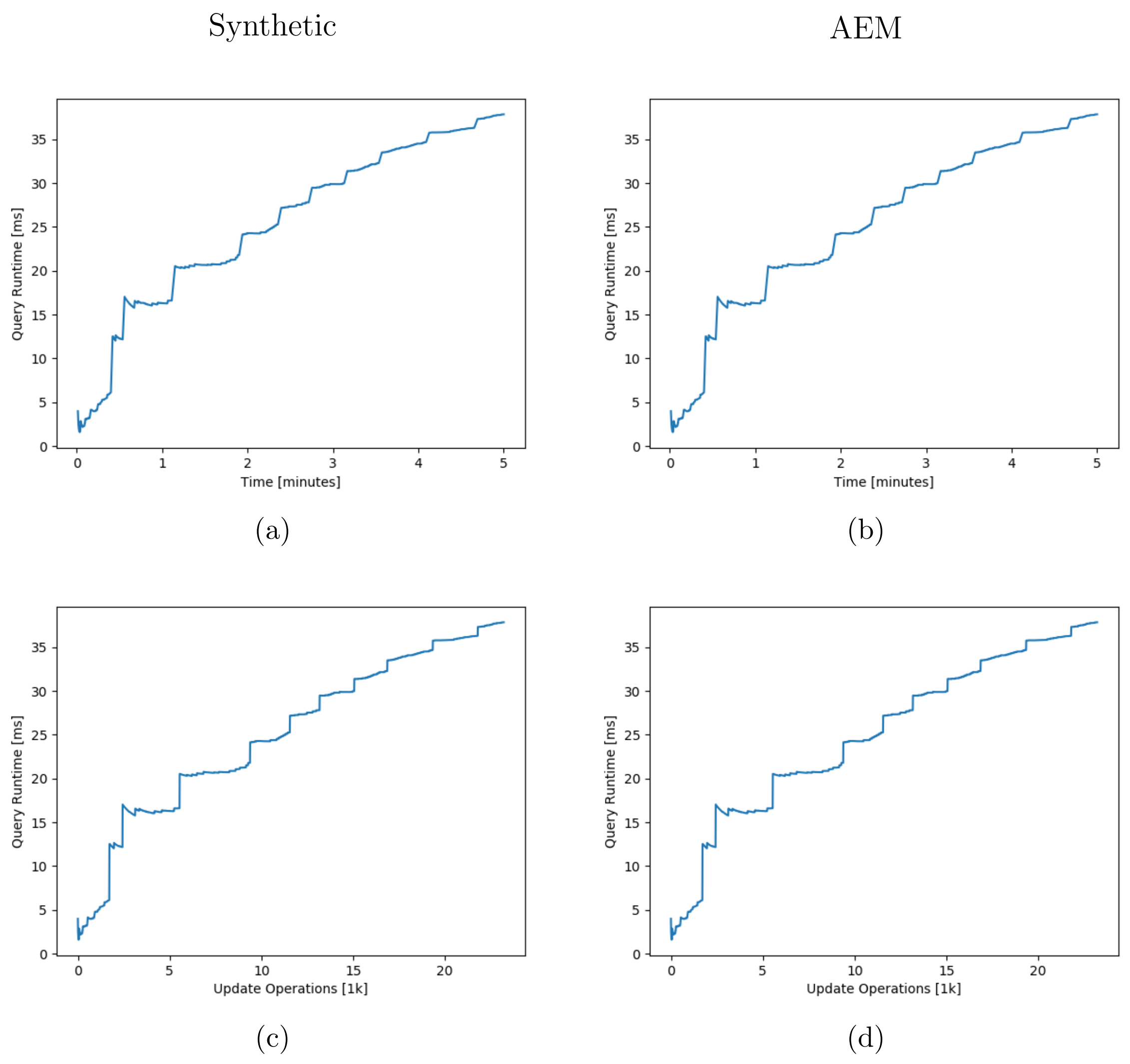
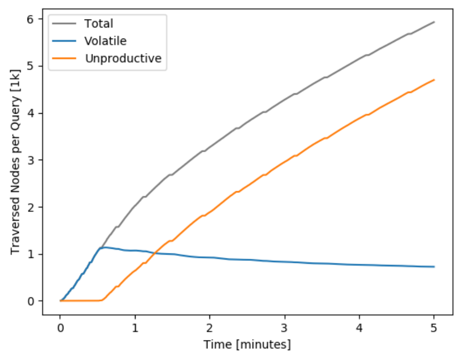


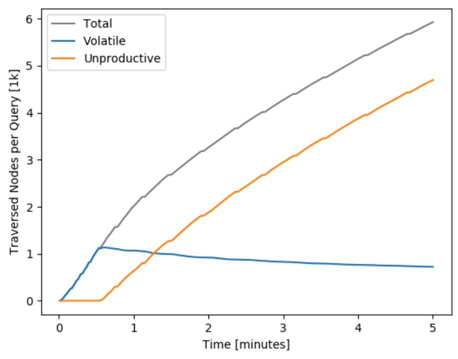
Figure 3: Query Execution Runtime

Synthetic

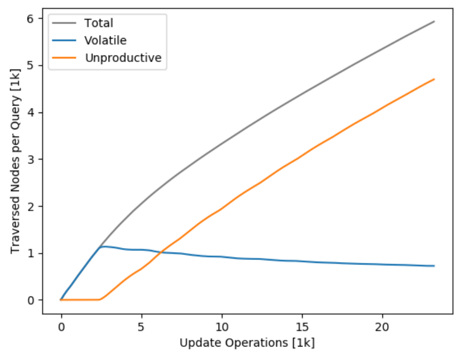
AEM



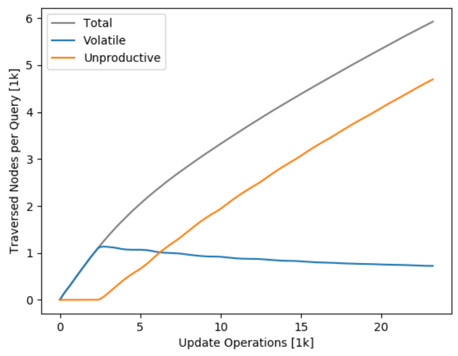
(a)



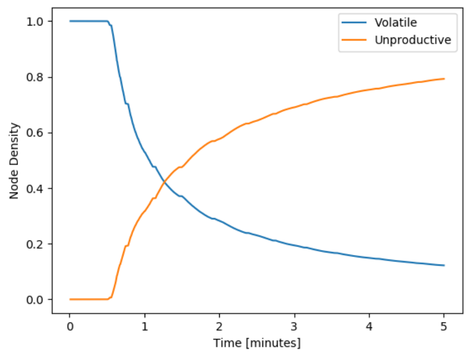
(b)



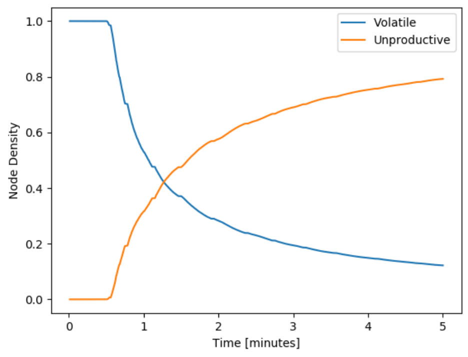
(c)



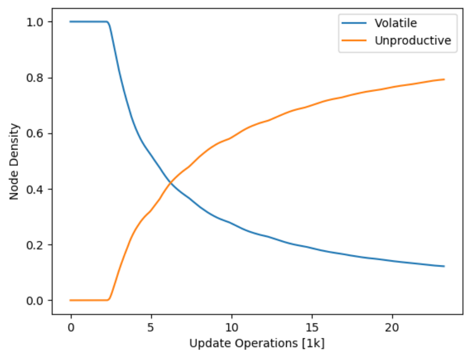
(d)



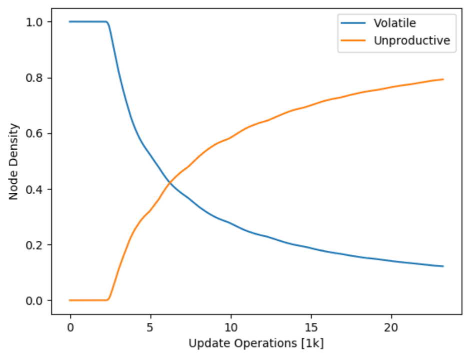
(e)



(f)



(g)



(h)

Figure 4: Index nodes encountered while executing a query