MADALGO
CENTER FOR MASSIVE DATA ALGORITHMICS


Qin Zhang
Aarhus University

Danmarks Grundforskningsfond
Danish National Research Foundation

A Tight Lower Bound for Dynamic Membership in the External Memory Model

Problems and Motivations

Membership: Maintain a set S in the universe U with |S| ≤ n. Given an x in U, answer whether x is in S or not?

Dictionary: Same as Membership. Except that if x is in S, then we have to return x's associate information.

Goal: Tradeoff between the (amortized) update cost t_u and the query cost t_q in the external memory (EM) model.

Motivations:

- Two of the most fundamental data structure problems in computer science!
- Used extensively in databases, routers and search engines.

The Conjecture

A long-time folklore conjecture in the external memory community (explicitly stated by Jensen and Pagh, 2007):

t_u must be Ω(1) if t_q is required to be O(1)

t_u : expected amortized update cost.
t_q : expected average query cost.

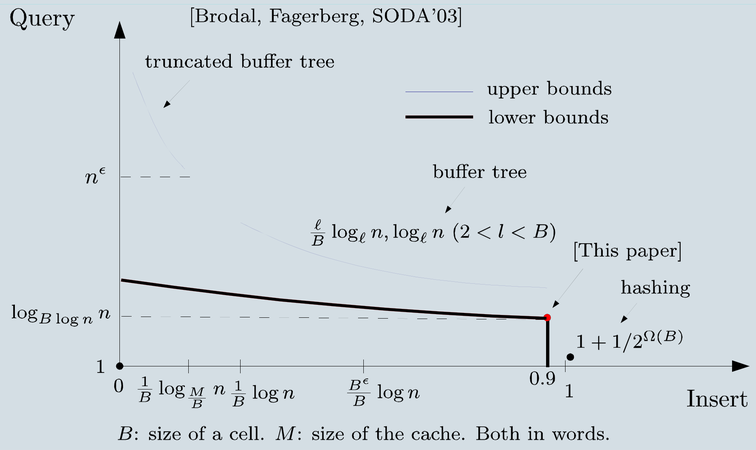
Intuitively speaking, we cannot buffer the updates without sacrificing the query cost much.

Selected Previous Results

- Knuth [2] analyzed *external hashing*: Expected average cost of an operation is $1 + \frac{1}{2}^{\Omega(B)}$, provided that the load factor is less than a constant smaller than 1. (Assuming truly random hash function exists.)
- Arge [3] presented a data structure, *buffer tree*, supporting amortized updates cost $O(1/B^{1-\varepsilon} \log n)$ and query cost $O(\log_B n)$.

Two results in two extremes.

Our Results



B: size of a cell. M: size of the cache. Both in words.

Striking Implications

- A sharp threshold result: for the external dictionary/membership problem, either use external hashing, or use the buffer tree, there is nothing in between!

  Buffering is impossible in the EM model with sublogarithmic query time.

- By reduction we can show that the query complexities of many problems such as 1D-range reporting, predecessor, partial-sum, etc., are all the same in the regime where the update time is less than 1.

  EM model is a "clearner" model than RAM in certain perspectives.

References

[1] E. Verbin and Q. Zhang. *A Tight Lower Bound for Dynamic Membership in the External Memory Model.* STOC 2010.

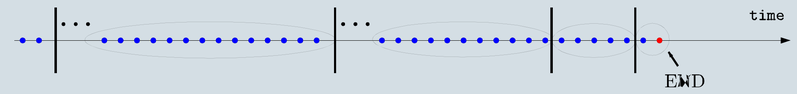
[2] D. E. Knuth. *Sorting and Searching. Volume 3 of The Art of Computer Programming.* Addison-Wesley, Reading MA, second edition,1998.

[3] L. Arge. *The Buffer Tree: A Technique for Designing Batched External Data Structures.* Algorithmica 2003.

Technical Contributions

Framework:

Insert n random items. Pick a random snapshot "END" in the insertion sequence, and divide previous insertions into blocks with exponentially increasing size. We find a set of cells C_i for each block. C_i are disjoint. Finally, we argue that a random query needs to read one bit from at least $\Omega(\log_{B \log n} n)$ blocks.



Proof ideas:

- For each block i, we first prove that for 0.99 fraction of items inserted in block i, their query paths will intersect C_i . The proof is by encoding arguments.
- Next, we introduce a lemma called LOSI, and then use it to prove that if 0.99 fraction of newly inserted items having query paths intersecting C_i , then at least a constant fraction of items in the universe have query paths intersecting C_i . We are done.

Illustration of the LOSI