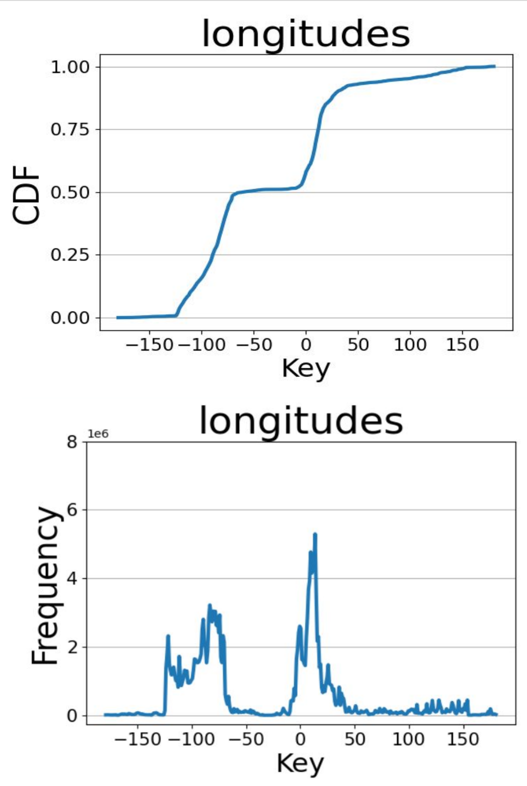
Understanding the Impact of Dynamic Graph Workload on Learned Data Structures

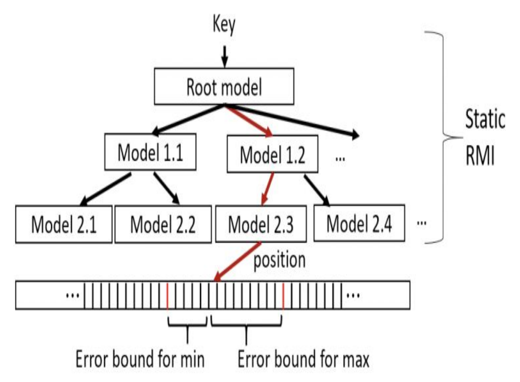
Ishaan Patel and Dhananjay Raghu, UNC Charlotte
Abdullah Al Raqibul Islam, CCI Research

SCAN ME

Introduction

- Learned indexes use machine learning models to predict the position of a piece of data using a key.
- The CDF of the keys is calculated and used to predict the location within a sorted array
- RMI is a hierarchy of models where a higher level model picks the model at the next level, and so on until a prediction is made.
- We used ALEX which is a mutable Learned index implementation.

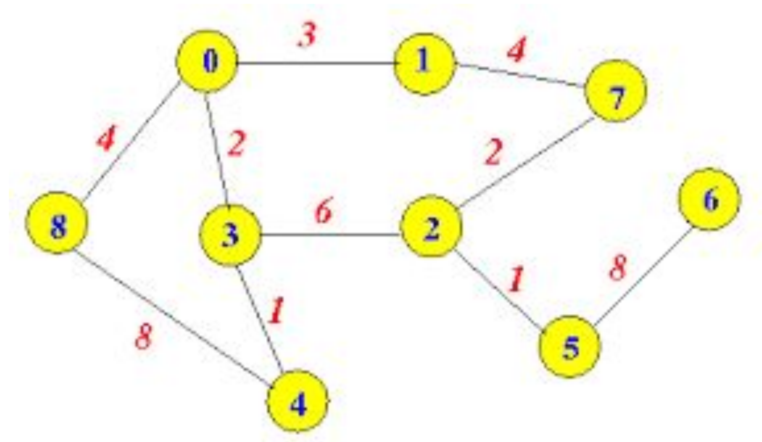




From Kraska, et al

Objectives

- ALEX is currently tested using key-value workloads, our main objective was to test on other workloads, such as graph workloads.
- Graphs are data structures with a collection of nodes filled with data and edges to connect the pairs of nodes, creating a network.
- Our preliminary test results can be used in developing learned graph data structures in the future.



http://www.cs.emory.edu/~cheung/Courses/253/Syllabus/Graph/dijkstra1.html

Method

- Benchmarked ALEX and STX B+ tree with graph datasets.
- Had trouble inputting 3 values (source, destination, weight) into the models
- Combined source and destination (2 32-bit ints) into a 64-bit int using bitwise left shifts and paired the 64-bit combination with the weight to replicate a key-value input.
- Compared collected results from ALEX and STX B+ tree with state of the art graph data structures such as CSR (Compressed Sparse Row), VCSR (Vertex-Centric Compressed Sparse Row), PCSR (Packed Compressed Sparse Row), and BAL (Blocked Adjacency Lists).
- We measured the graph insertion time and graph analysis times for each of the learned data structures.

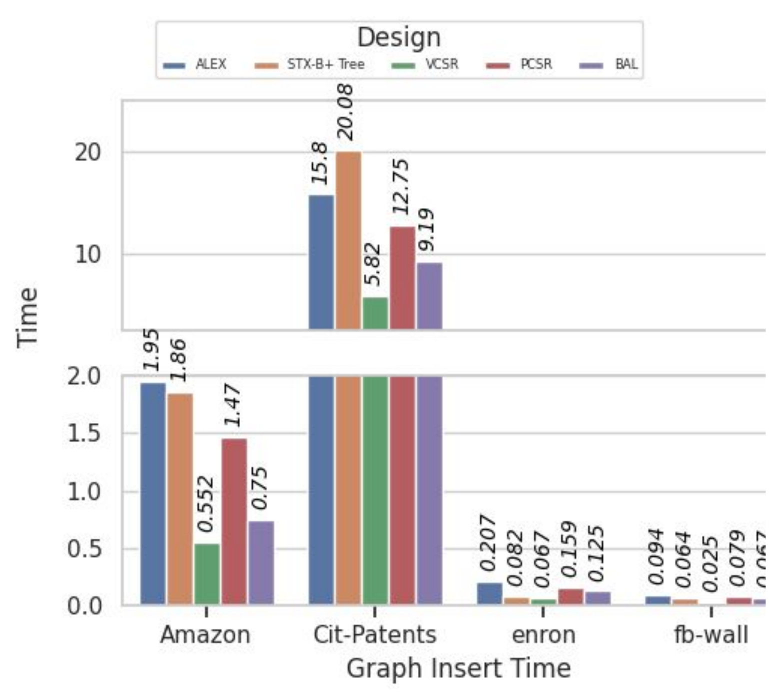
Evaluation Platform

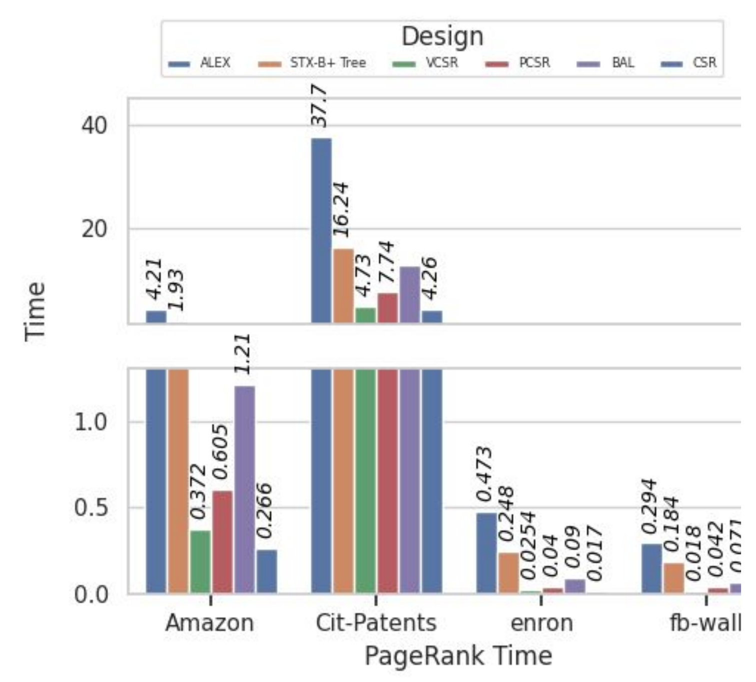
- We conducted the evaluations on UNCC's Galaxy cluster using the Centaurus and GPU partitions, containing 12 general compute nodes with 16 cores and 1 large memory compute node with 16 cores and 768GB RAM.
- We used Graph Insert time and 2 Graph Algorithm Kernels to indicate performance.
  - Pagerank: running on fixed number of interactions (link analysis)
  - SSSP: delta stepping (shortest path)

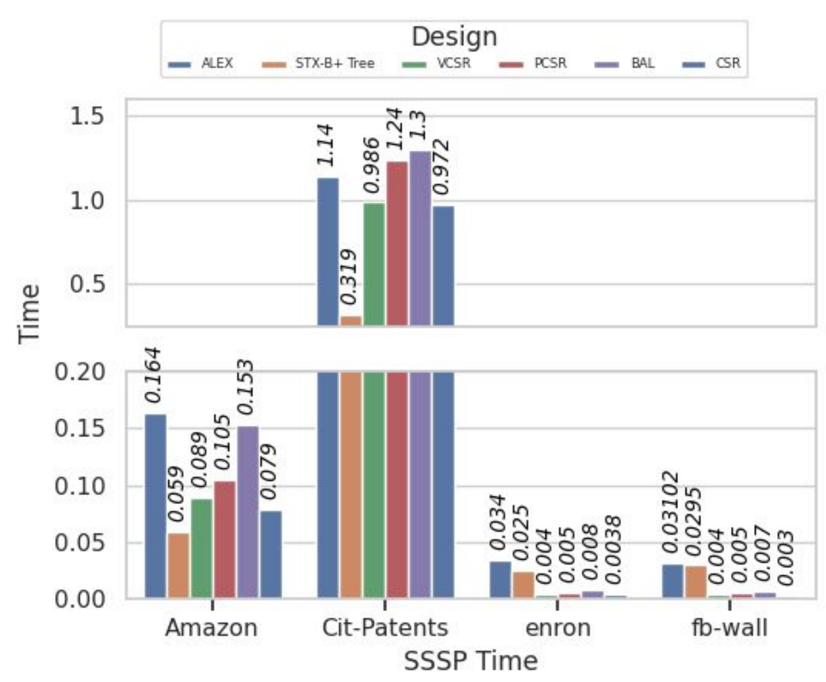
Datasets	Domain	$\|V\|$	$\|E\|$	$\|E\|/\|V\|$
Amazon	purchase	403393	4886816	12
Cit-Patents	citation	6009554	33037894	6
enron	temporal	87273	594912	7
fb-wall	temporal	63891	366824	6

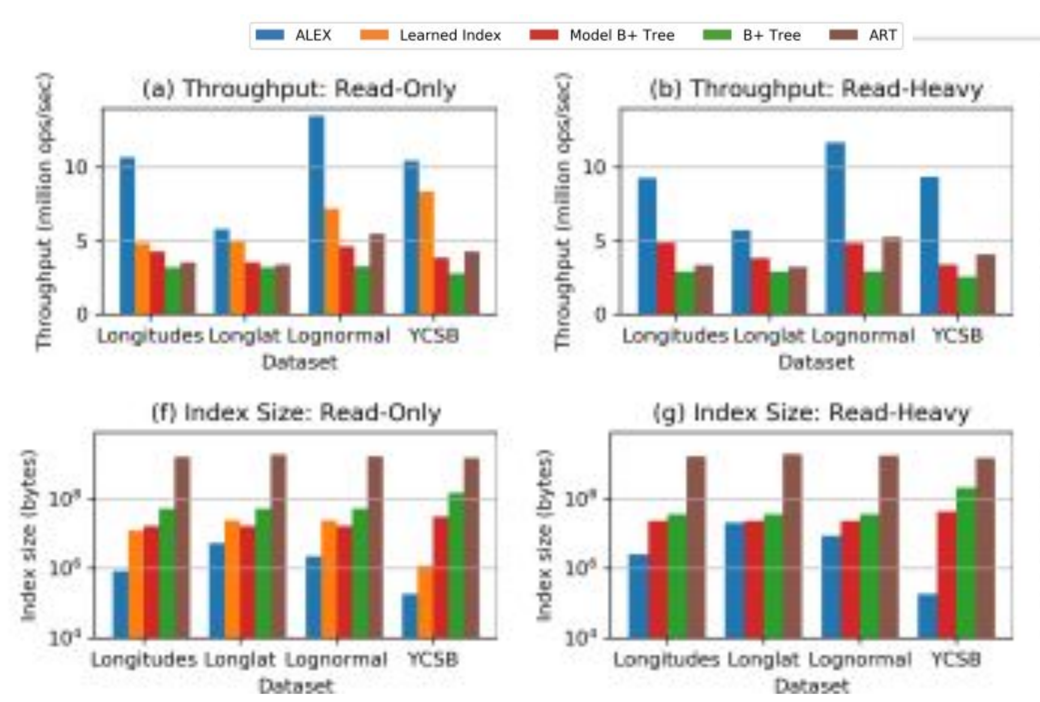
Results

- We discovered that VCSR and CSR performed better than the rest of the implementations in almost all tests
- VCSR and CSR are comparably similar, but CSR is up to 9.88x faster than the average of the rest of the indexes in Pagerank time and up to 4.28x faster in SSSP time. We also found that our implementation of ALEX was outperformed by STX B+ tree in all graph workloads and scenarios except for graph insertions in Cit-Patents.
- In Graph Insert time, VCSR was the fastest index by a minimum of 256.4% compared to 2nd place (STX B+tree in fb-wall)
- ALEX and STX B+ trees are shown to be extremely efficient when it comes to benchmarking simple KV datasets, unlike their graph workload handling implementations.
- As shown in the 4th chart, ALEX excels when dealing with read-only and read-heavy datasets, far more than competing learned indexes and B+ trees.
- This brings up the question of the differences between graph optimized data structures and data structures such as learned indexes like ALEX, or B+ trees like STX.









Conclusions

- We concluded that existing RMIs can be formatted to fit graph workloads, as they are not yet as efficient as graph representations like CSR, and VCSR.
- However, the prospect of using learned indexes to take on dynamic graph workloads is exciting because there are numerous areas for improvement to focus on.
- Future considerations include adding concurrency, implementing other regression techniques instead of just simple linear regression (ex. polynomial regression), and add support for persistence.

References

- Jialin Ding, Umar Farooq Minhas, Jia Yu, Chi Wang, Jaeyoung Do, Yinan Li, Hantian Zhang, Badrish Chandramouli, Johannes Gehrke, Donald Kossmann, David Lomet, and Tim Kraska. 2020. ALEX: An Updatable Adaptive Learned Index. Technical Report. https://www.microsoft.com/en-us/research/publication/msr-alex-techreport/
- A. Al Raqibul Islam, D. Dai and D. Cheng, "VCSR: Mutable CSR Graph Format Using Vertex-Centric Packed Memory Array," 2022 22nd IEEE International Symposium on Cluster, Cloud and Internet Computing (CCGrid), Taormina, Italy, 2022, pp. 71-80, doi: 10.1109/CCGrid54584.2022.00016.