1. Bundled References

In this paper we address the challenge of providing linearizablerange query operations for linked data structures by introducing a new building block; we call bundled references. Using bundled references provides range queries with a path through the data structure consistent with their linearization point. Our design guarantees that range queries only traverse nodes belonging to their snapshot and that they never block writers. With multi-version concurrency control (MVCC) in mind, we implement our technique in three data structures. The experimental evaluation of our bundled linkedlist, skip list, and binary search tree, including their integration as indexes in the DBx1000 in-memory database, shows up to 20% improvement over state-of-the-art techniques to provide linearizable range queries, as well as a more consistent performance profile across a variety of workloads

2. Implementation notes

This work graciously builds on a benchmark developed by Arbel-Raviv and Brown's work (https://bitbucket.org/trbot86/implementations/src/master/cpp/range_queries/) to provide linearizable range queries in linked data structures. We use their codebase as a starting point, and integrate our technique into their benchmark. The core of our bundling implementation is contained in the 'bundle' directory, which implements the global structures as well as necessary functions of bundling. The three data structures we implement are found in 'bundled_*' directories. The scripts necessary to produce the plots found in our paper are included under the 'microbench' directory.

3. Getting Started Guide

The remainder of this document intendeds to support an artifact evaluation for the paper "Bundled References: An Abstraction for Highly-Concurrent Linearizable Range Queries" by Jacob Nelson, Ahmed Hassan and Roberto Palmieri. As stated above, our contributions with respect to the paper are contained in the directories prefixed with "bundle".

a. Important files and directories

Implementation

`./bundle` implements the bundling interface as a linked list of bundle entries. In addition to the linked list bundle, there is an experimental cirular buffer bundle (not included in the paper) as well as an unsafe version that eliminates the overhead of ensuring bundle consistency for comparison.

`./bundle_lazylist`, `./bundle_skiplistlock` and `./bundle_citrus` each implement a data structure to which we apply bundling. Note that we do not apply our technique to the remaining data structures (which are lock-free) because our current bundling implementation would impose blocking.

Experiments

`config.mk` is the primary configuration file and is used across all experiments. Users will need to update this file to match their system (more on this later).

The experiments that we report in the paper are located in the following directories.

- `./microbench` tests each data structure in isolation.
- `./macrobench` ports DBx1000 to include the implemented data structures.

Generated files

`experiment_list.txt` is generated by the microbenchmark, resides in `./microbench` and contains the list of experiments to run.

`./microbench/data` and `./macrobench/data` are the destination directories for all saved data during experiments. This is also where automatically generated .csv files will be saved when plotting the results.

`./figures` stores all generated plots. It is split first into the respective experiments and then into the data structures used.

b. Requirements

The experiments from the paper were executed on a 4-socket machine with Intel Xeon Platinum 8160 processors running Red Hat Enterprise 7.6. However, we also successfully tested on a dual-socket machine with Intel Xeon E5-2630 v3 processors running Ubuntu 18.04. The C++ libraries are requried to build and run the experiments, while the Python libraries are used for plotting results.

C++ Libraries:

- libnuma (e.g., `sudo apt install libnuma-dev`)
- libjemalloc (included in 'lib')

If using an ISA other than x86, **it is necessary to replace the jemalloc library** contained in the `lib` folder. To do so, install the correct version to your system then copy the shared library to `lib` so that the scripts can locate it. Otherwise, update line 82 in `microbench/runscript.sh` and line 73 in `macrobench/runscript.sh` to point to the location of the library.

Python libraries:

- python (v3.6)
- plotly (v4.12.0)
- psutil (v5.7.2)
- requests (v2.24.0)
- pandas (v1.1.3)

The above libraries can be installed using `pip` or Anaconda depending on the users desired level of isolation from the rest of their environment.

c. Configuration

Note: any warnings regarding hardware transactional memory (HTM) can be safely ignored since we do not compare with this competitor.

Once the C++ dependencies have been installed, you can begin to test the microbenchmark. First, configure the build with the `config.mk` file. There are five configuration parameters.

- `maxthreads` is the maximum number of threads to be tested during the experiments
- `maxthreads_powerof2` this is used for bookkeeping and is the next largest power of two from `maxthreads`
- `threadincrement` is the sampling period of threads between 0 and `maxthreads` for each experiment
- `cpu_freq_ghz` is the system's CPU frequency in GHz
- `pinning_policy` is a string that starts with "-bind " (or left blank) and maps threads to cores during execution

Configuration Tips

1. Together, `maxthreads` and `threadincrement` determine the number of samples generated during experiments. For example, on a 44 core machine with `maxthreads=44` and `threadincrement=8` the resulting numbers of threads tested will be [1, 8, 16, 32, 40, 44]. Both 1 and `maxthreads` are always included, regardless of whether `maxthreads` is a multiple of `threadincrement`.

2. The easiest way to determine both `cpu_freq_ghz` and `pinning_policy` is to execute `lspci` on the command line. The first is directly used from the line indicating CPU frequency. The latter is a comma separated list of the NUMA node mappings. Consider a hypothetical machine with NUMA zones of four cores each that has the folling mappings: `NUMA 0: 1,3,5,7` and `NUMA 1: 0,2,4,6`. The pinning policy that mimics our setup would then be `pinning_policy="-bind 1,3,5,7,0,2,4,6`. If `pinning_policy` is left blank then no specific policy is used.

d. Building the Project

Once configured, build the binaries for each of the data structures and range query techniques with the following:

```
cd microbench
make -j lazylist skiplistlock citrus rlu lbundle unsafe
```

The first three arguments to the `make` command (i.e., `lazylist`, `skiplistlock`, `citrus`) build the EBR-based approach from Arbel-Raviv and Brown. The next argument (i.e., `rlu`) builds the RLU-based lazy-list and Citrus tree. The fifth argument (i.e., `lbundle`) builds the bundled lazy-list, optimistic skip-list and Citrus tree. Finally, the last argument (i.e., unsafe) builds the three data structures of interest with no instrumentation for range queries. Unlike the unsafe implementation provided by Arbel-Raviv and Brown, our implementation does not reclaim memory. We chose to do this because the RLU-based data structures do not utilize epoch-based memory reclamation. As such, our unsafe versions are an upper bound on all range query techniques and provides a more general reference.

e. Running Experiments

Finally, run individual tests to obtain results for a given configuration. The following command runs a workload of 5% inserts (`-i 5`), 5% deletes (`-d 5`), 80% gets and 10% range queries (`-rq 10`) on a key range of 100000 (`-k 100000`). Each range query has a range of 50 keys (`-rqsize 50`) and is prefilled (`-p`) based on the ratio of inserts and deletes. The execution lasts for 1s (`-t 1000`). There are no dedicated range query threads (`-nrq 0`) but there are a total of 8 worker threads (`-nwork 8`) and they are bound to cores following the bind policy (`-bind 0-7,16-23,8-15,24-31`). Do not forget to load jemalloc.

```
env LD_PRELOAD=../lib/libjemalloc.so TREE_MALLOC=../lib/libjemalloc.so \
./hostname.skiplistlock.rq_lbundle.out -i 5 -d 5 -k 100000 -rq 10 \
-rqsize 50 -p -t 1000 -nrq 0 -nwork 8 -bind 0-7,16-23,8-15,24-31
```

For more information on the input parameters to the microbenchmark itself see README.txt.old, which was written for the original implementation. We did not change any arguments.

4. Results Validation

a. Microbenchmark

Our results demonstrate that in mixed workload configurations, and in the presence of range queries, our implementation outperforms the competitors. This can be demonstrated by running the full microbenchmark using `microbench/runscript.sh`.

Assuming that the binaries have already been built during the previous steps, let's generate the plots included in the paper (once the dependencies above are installed). From the root directory, run the following:

```
./runscript.sh
cd ..
python plot.py --save_plots --microbench
```

`runscript.sh` will run expeirments based on `experiment_list_generate.sh`, which will write a list of experiments to be run into a file. This generation script can be altered to try out new configurations. `plot.py` pulls the configuration directly from `config.mk` so no changes to it should be necessary.

`experiment_list_generate.sh` includes two experiments. The first, saved under `microbench/data/workloads` fixes the range query size to 50 and tests various workload configurations. This corresponds to Figure 2 in the paper as well as additional experiments for get-only and update-only workloads. The second, whose results will be written to `microbench/data/rq_sizes`, executes a 50%-50% update-rq workload at various range query lengths (i.e., 1, 5, 10, 50, 100, 500). This corresponds to Figure 3.

WARNING: The experiments can take a long time to run because there are many competitors. As was used for our results, have preconfigured the run to execute three trials, run for 3s, and test the optimistic skip-list and Citrus tree. Both `runscript.sh` and `experiment_list_generate.sh` contain some addtional configuration options, but *they are not required*.

- `runscript.sh` contains some additional configuration, that if so desired may be updated. Specifically, lines 9 and 36 are pertinent as they adjust the number of trials per-configuration and the length of each trial. If you do not wish to wait as long for the experiments to terminate, you may adjust these values knowing that the results may differ from those presented in the paper.
- `experiment_list_generate.sh` contains some other configuration options. The current configuration is an abridged version, which limits the nubmer of overall tests to run for the sake of time. Lines 16-18 indicate which competitors to test (`rqtechniques`), which data structures to run them on (`datastructures`), and the key ranges to use (`ksizes`).

Output

As stated previously, the microbenchmark saves data under `./microbench/data`. This raw data is used by the plotting script, but is first translated to a .csv file that is also stored in the subdirectory corresponding to each experiment in `experiment_list_generate.sh`. Upon running `plot.py` with the argument `--save_plots`, the generated graphs will be stored in `./figures` (again, in the corresponding subdirectories).

b. Macrobenchmark

In addition to demonstrating better performance in mixed workloads, we also demonstrate improvements over competitors in index performance when integrated into a database. This can be observed by running the macrobenchmark.

To build and run the DBx1000 integration, run the following from the root directory:

```
cd ./macrobench
./compile.sh
./runscript.sh
cd ..
python plot.py --save_plots --macrobench
```

In comparison to the microbenchmark, this will take longer to run. We suggest going for a long walk, calling a friend, or taking a nap. Two plots will be generated, one for each of the data structures at various numbers of threads.

Output

As with the microbenchmark, the macrobenchmark generates raw output in `./macrobench/data`. The last command in `./runscript.sh` automatically generates the .csv file that is stored in `./macrobench`. This file (i.e., `data.csv`) is then used by the plotting scripts, whose output is saved under `./figures/macrobench`.

c. Memory Reclamation

The initial binaries are built without bundle entry reclamation to match the paper discussion. Whenever a node is deleted its bundle entries are reclaimed but stale bundle entries are not garbage collected for connected nodes. To enable reclamation of bundle entries, uncomment line 11 of `bundle.mk`. The following line defines the number of nanoseconds that elapse between iterations of the cleanup thread. It is currently set to 100ms.

Once `bundle.mk` is updated, remake the the bundled data structures using `make -j lbundle` and rerun the previously described microbenchmark. Be sure to move the original plots so they are not overwritten when regenerating them.