Cache Simulator

Usage

To build the program, use `g++ cache_sim.cpp`.
Then to run, use `./a.out {input_file}`.

Implementation

The addresses given are assumed to be block addresses, which makes the cache and main memory block addressable.

Replacement policy

- As described, each Cache Set is divided into two groups:

1. One group contains the HIGH PRIORITY lines of the set
2. The other group contains the LOW PRIORITY lines of the set

   The sizes of the groups within a set is not fixed. If there are `N` blocks in a set and `H` blocks are in high priority group, the remaining `N-H` blocks will comprise the low priority group. Now if a block is moved from high priority to low priority, then the high priority group will have `H-1` blocks and low priority group will have `N-H+1` blocks.

- If a line is accessed again after the initial access that fetches it into the cache, it is promoted to the HIGH PRIORITY group. If a line is not accessed for sufficiently long (T cache accesses) after being moved to the HIGH PRIORITY group, it is moved to the LOW PRIORITY group.

- Replacement is always done in LOW PRIORITY group. If there are no blocks in LOW PRIORITY group, then replacement is done in HIGH PRIORITY group. Within a priority group, the **Least Recently Used policy** is used to manage the lines.

- Within a set, the HIGH PRIORITY group physically comes before the LOW PRIORITY group i.e. If there are 3 high priority blocks and 1 low priority block in a set, then block #0 to block #2 are high priority and block #3 is low priority. This makes searching in HIGH PRIORITY group more effective as HIGH PIORITY blocks are traversed befor the LOW PRIORITY blocks and as blocks in HIGH PRIORITY group are more likely to be accessed again, they increase the overall efficiency of searching in a set. The high and low priority groups are separated by a *set divider*.

- Initially all blocks belong to low priority group.

Reads & Writes

- **Write back on data-write hit:** On write hits, write to the cache and set dirty bit to 1. Write back to the main memory whenever a dirty block is replaced.
- **Write allocate on data-write miss:** On write miss, update the main memory and load to the cache from the main memory.

Note: The main memory & test memory are initialized with values equal to the block number/address.

Testing

Set the `DEBUG` macro in `cache_sim.cpp` to 1 to print the state of the cache after each request and additional information useful for testing.

A *dummy test memory* which has no cache component is used to verify the results of all read requests. All `W` requests simply write to the test memory and all `R` requests simply read from the test memory which are then compared with the results of the `R` request given by the cache simulation. If matched, "Correct read!!" is printed and "Wrong read!!" is printed otherwise.

Generating test inputs

`test_input_generator.cpp` can be used to generate large random test inputs which can be checked with the dummy test memory for correctness.

The test inputs generated can be customized as follows:

```
int access_requests = 10000; // number of cache access requests to generate
int distinct_mem_address = 20; // number of distinct memory addresses in
requests. These are generated with a random frequency.

// read or writes generated with frequency as specified in read_or_write_freq
vector
vector<int> read_or_write_vector{R, W};
vector<int> read_or_write_freq{65, 35}; // Reads comprise 65% of total
requests, and Writes comprise 35%.
```

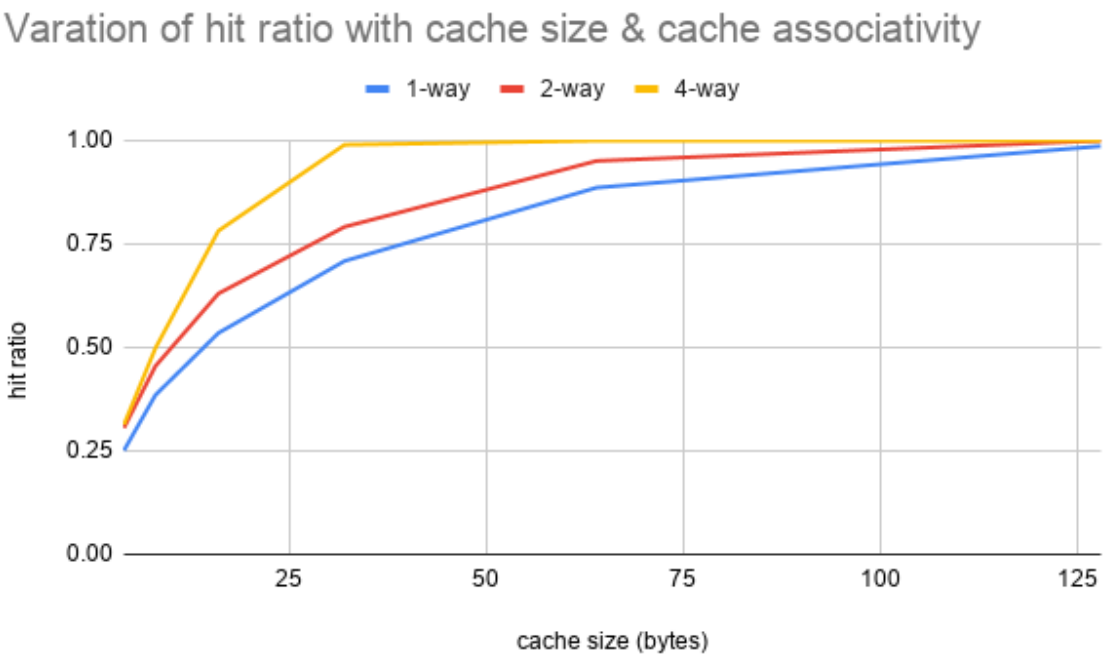
Test files are provided in `input/` folder and they were all verified by testing against the dummy test memory. The test files have access requests ranging upto **50,000** and even more can be generated using `test_input_generator.cpp`.

Observations

These observations were made using `input/inp_gen_obs.txt`.

- Variation of hit ratio with cache size and cache associativity

cache size (bytes)	1-way	2-way	4-way
4	0.251033	0.305533	0.313533
8	0.385033	0.4549	0.498267
16	0.5349	0.6301	0.781667
32	0.7086	0.791133	0.9894
64	0.885833	0.950367	0.999333
128	0.986067	0.999333	0.999333



- Variation of hit ratio with T and cache size (at 4-way cache associativity)

cache size (bytes)	T = 4	T = 8	T = 32
4	0.305533	0.312	0.317833
8	0.4549	0.462033	0.469867
16	0.6301	0.6371	0.640667
32	0.791133	0.794267	0.795767
64	0.950367	0.9506	0.9528
128	0.999333	0.999333	0.999333

- Variation of hit ratio with T and cache associativity (constant cache size = 32 bytes)

cache associativity	T=4	T=8	T=32
1-way	0.7086	0.7086	0.7086
2-way	0.791133	0.794267	0.795767
4-way	0.9894	0.9897	0.992267

Conclusions

Division of sets into HIGH and LOW priority groups increases the hit ratio of the cache in general as HIGH priority blocks are likely to be accessed again following the principle of temporal & spacial locality. But this will require some additional hardware overheads for maintaining the division. Also, it can be seen that increasing the cache block size and cache associativity for the same test inputs will result in higher hit ratios but will require more time for searching within a set, as expected.