Jeffrey Lau, Giovanni Lupo, Zacharia Hammad
11/08/2024
ECEC 412
Prof. Anup Das
Project 3

ECE 412 Project 3

LRU: 531.deepjseng_r_llc.mem_trace

cache_size	assoc	Hit rate
128	4	76.939192%
256	4	88.972509%
512	8	94.859542%
1024	16	95.751690%
2048	16	95.958330%
2048	4	95.906338%
128	16	78.541262%

LFU: 531.deepjseng_r_llc.mem_trace

cache_size	assoc	Hit rate
128	4	75.574815%
256	4	88.382205%
512	8	94.017041%
1024	16	94.542108%
2048	16	95.262287%
2048	4	95.820579%
128	16	72.702271%

LRU: 541.leela_r_llc.mem_trace

cache_size	assoc	Hit rate
128	4	42.554451%

256	**4**	**70.743799%**
512	**8**	**94.443418%**
1024	**16**	**99.011727%**
2048	**16**	**99.627833%**
2048	**4**	**99.455119%**
128	**16**	**55.102674%**

LFU: 541.leela_r_llc.mem_trace

cache_size	assoc	Hit rate
128	**4**	**48.440187%**
256	**4**	**70.096169%**
512	**8**	**87.869778%**
1024	**16**	**94.520435%**
2048	**16**	**99.132343%**
2048	**4**	**99.136207%**
128	**16**	**51.064270%**

LRU: 548.exchange2_r_llc.mem_trace

cache_size	assoc	Hit rate
128	**4**	**99.926982%**
256	**4**	**99.984771%**
512	**8**	**99.985094%**
1024	**16**	**99.985094%**
2048	**16**	**99.985094%**
2048	**4**	**99.985094%**
128	**16**	**99.990552%**

LFU: 548.exchange2_r_llc.mem_trace

cache_size	assoc	Hit rate
128	**4**	**72.732198%**
256	**4**	**99.984948%**
512	**8**	**99.985094%**
1024	**16**	**99.985094%**
2048	**16**	**99.985094%**
2048	**4**	**99.985094%**
128	**16**	**99.990552%**

When evaluating cache performance for AI applications, various cache sizes and levels of associativity were tested using both LRU (Least Recently Used) and LFU (Least Frequently Used) replacement policies. In most cases, the LRU policy yielded higher hit rates. For example, the 531.deepjseng_r workload achieved a hit/miss ratio of 95.96% with a cache capacity of 2048 and 16-way set associativity, while the 541.leela_r workload reached an impressive 99.63%. The 548.exchange2_r workload consistently maintained high hit rates across all configurations, with a hit rate of 99.99% for cache sizes of 256 and above.

On the other hand, the LFU policy generally resulted in slightly lower hit ratios, regardless of the associative capacity. However, it performed better as cache sizes increased. This indicates that the LRU policy, which is based on recency, outperforms the LFU policy based on frequency for these specific AI workloads, mainly when using larger caches.