CO262. Tutorial 7. 13 – April – 2018

1. An (m,n) correlating branch predictor uses the behavior of the most recent *m* executed branches to choose from *2^m* predictors, each of which is an *n*- bit predictor. A two-level local predictor works in a similar fashion, but only keeps track of the past behavior of each individual branch to predict future behavior.

There is a design trade-off involved with such predictors: Correlating predictors require little memory for history which allows them to maintain 2-bit predictors for a large number of individual branches (reducing the probability of branch instructions reusing the same predictor), while local predictors require substantially more memory to keep history and are thus limited to tracking a relatively small number of branch instructions. Consider a (1,2) correlating predictor that can track four branches (requiring 16 bits) versus a (1,2) local predictor that can track two branches using the same amount of memory. For the following branch outcomes (P.T.O. for the table of branch outcomes), provide each prediction, the table entry used to make the prediction, any updates to the table as a result of the prediction, and the final misprediction rate of each predictor. Assume that all branches up to this point have been taken. Initialize each predictor to the following:

Correlating predictor

Entry	Branch	Last 2 outcomes	Prediction
0	0	T	T with one misprediction
1	0	NT	NT
2	1	T	NT
3	1	NT	T
4	2	T	T
5	2	NT	T
6	3	T	NT with one misprediction
7	3	NT	NT

Local predictor

Entry	Branch	Last 2 outcomes	Prediction
0	0	T,T	T with one misprediction
1	0	T,NT	NT
2	0	NT,T	NT
3	0	NT	T
4	1	T,T	T
5	1	T,NT	T with one misprediction
6	1	NT,T	NT
7	1	NT,NT	NT

Correlating Predictor Behavior

Branch PC (word address)	Outcome	Prediction	Table entry used to make the prediction	Changes to the predictor table
454	T			
543	NT			
777	NT			
543	NT			
777	NT			
454	T			
777	NT			
454	T			
543	T			

Local Predictor Behavior

Branch PC (word address)	Outcome	Prediction	Table entry used to make the prediction	Changes to the predictor table
454	T			
543	NT			
777	NT			
543	NT			
777	NT			
454	T			
777	NT			
454	T			
543	T			

2. Given data: Clock Speed: 2 GHz . Peak Floating Point Rate: 4 GFLOPs . Sustained Memory Bandwidth: 2 GB/s . Sustained L1 Cache Bandwidth: 16 GB/s . GFLOPs = 10^9 floating point operations per second . GB/s = GigaByte per second = 10^9 bytes per second . The following code fragment calculates the inner product of two vectors.

```
Sum = 0
for i=0 to N-1
      Sum = Sum + A[i] * B[i]
endfor
```

(a) What is the maximum rate, in GFLOPs, at which this can perform if the data fits in L1 cache?

(b) What is the maximum rate, in GLOPSs, at which this can perform if the data does not fit in cache?

(c) What is the fraction of peak performance that can be obtained with this code in these two cases?