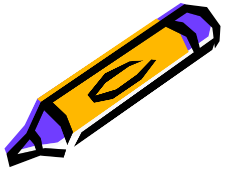
Fiscal Year 2019

Ver. 2019-12-25a

Course number: CSC.T433
School of Computing,
Graduate major in Computer Science

Advanced Computer Architecture

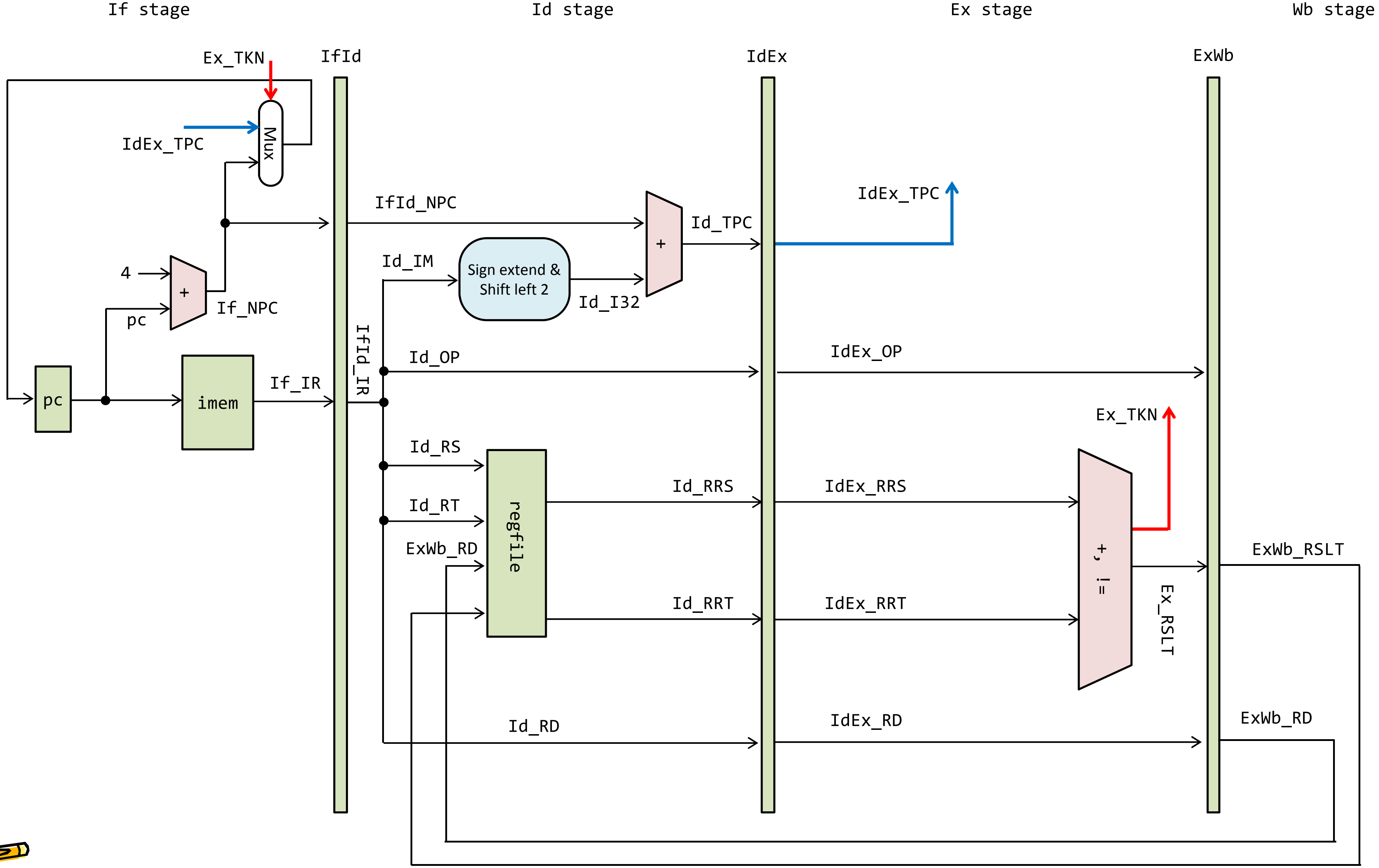
7. Instruction Level Parallelism: Instruction Fetch and Branch Prediction (2)



www.arch.cs.titech.ac.jp/lecture/ACA/
Room No.W936
Mon 13:20-14:50, Thr 13:20-14:50

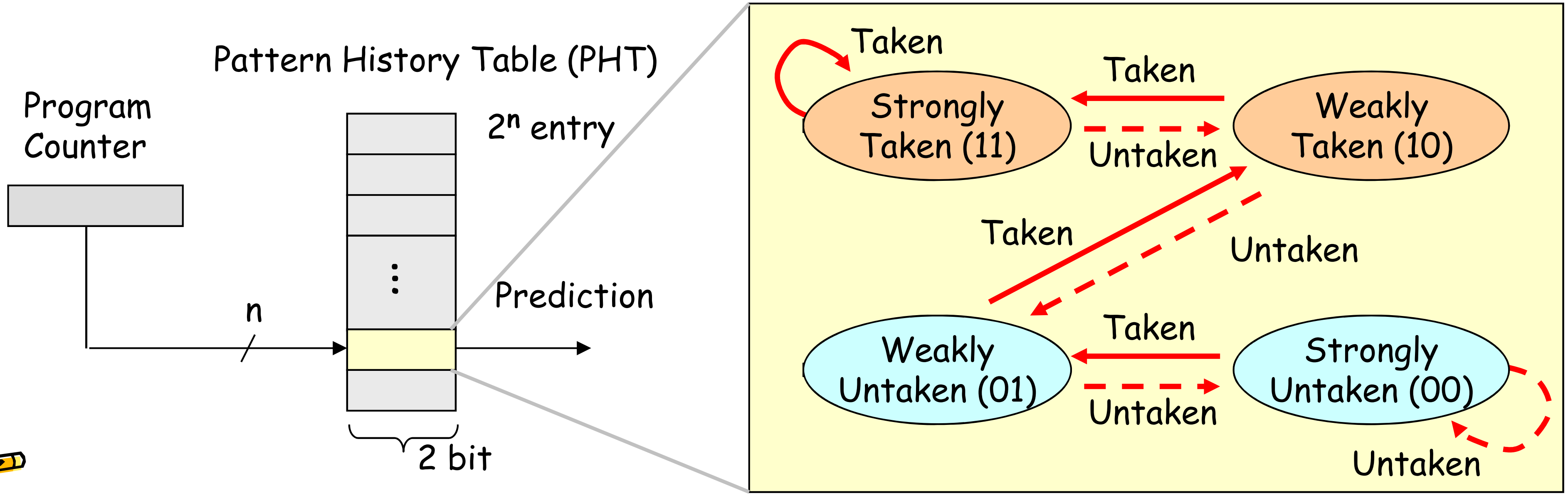
Kenji Kise, Department of Computer Science
kise _at_ c.titech.ac.jp

Four stage pipelined processor supporting ADD and BNE, which does not adopt data forwarding (proc08.v, Assignment 6)



Simple branch predictor: bimodal

- Program has many branch instructions. The behavior may depend on each branch. Use one counter for one branch instruction
- How to predict
  - Select one counter using PC, then it predicts 1 if the MSB of the register is one, otherwise predicts 0.
- How to update
  - Select one counter using PC, then update the counter same manner as 2bit counter.



An innovation in branch predictors in 1993

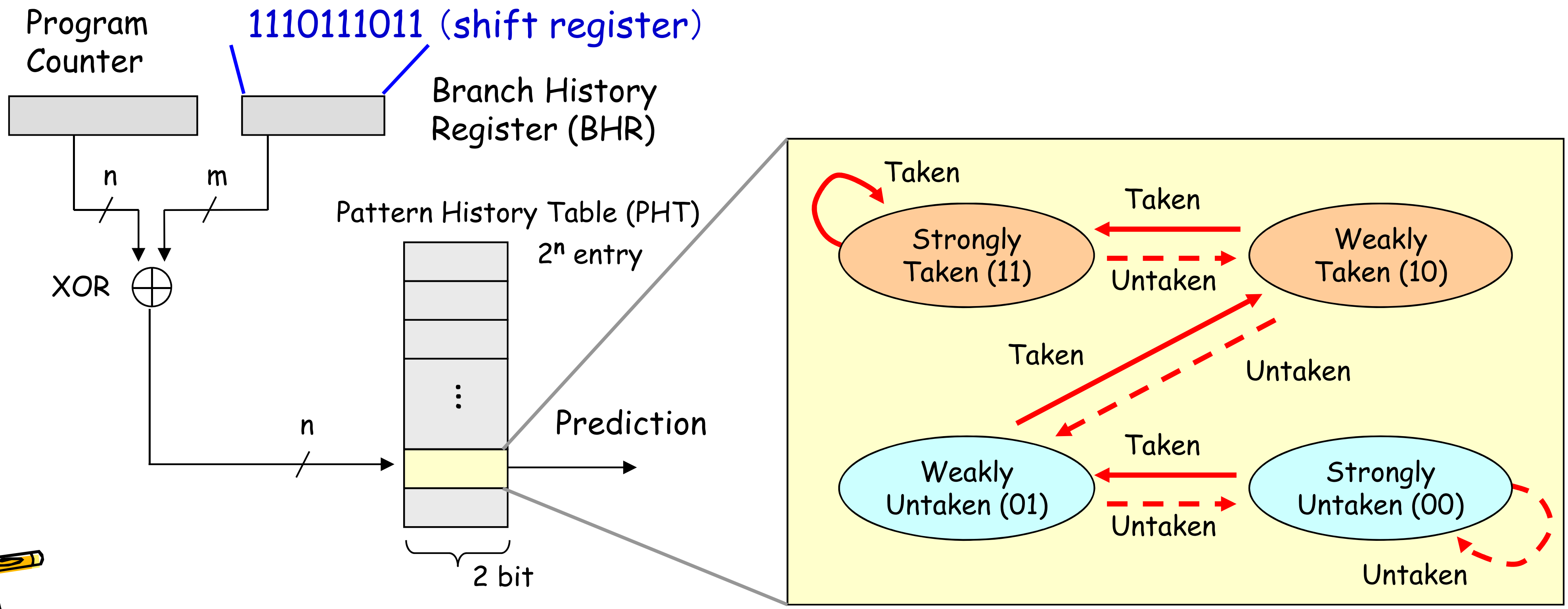
- Using branch history
  - global branch history
  - local branch history
- 2-level branch predictor and Gshare

- *Assume predicting the sequence* 1110 1110 1110 1110 1110 …

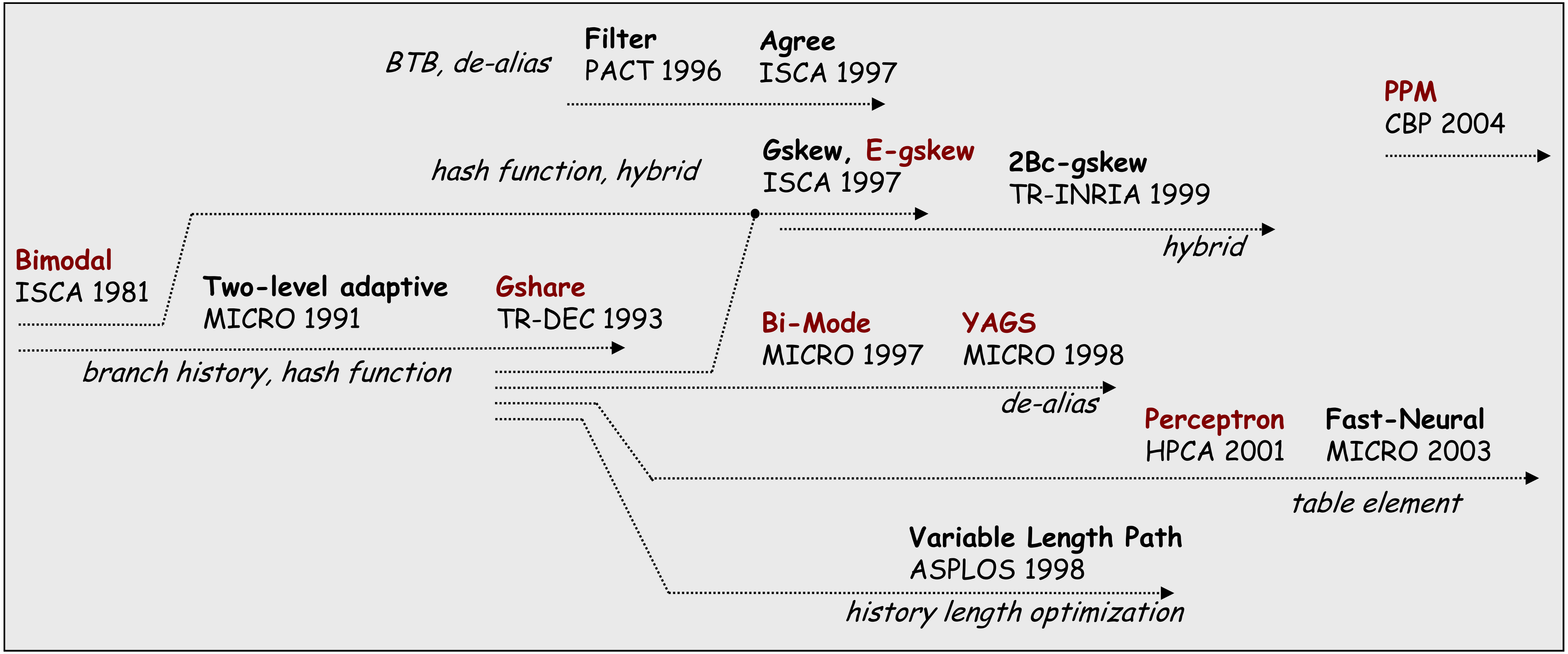
```
11101110 ?
111011101 ?
1110111011 ?
11101110111 ?
111011101110 ?
```

Gshare (TR-DEC 1993)

- How to predict
  - Using the exclusive OR of the global branch history and PC to access PHT, then MSB of the selected counter is the prediction.
- How to update
  - Shifting BHR one bit left and update LSB by branch outcome in IF stage.
  - Update the used counter in the same way as 2BC in WB stage.



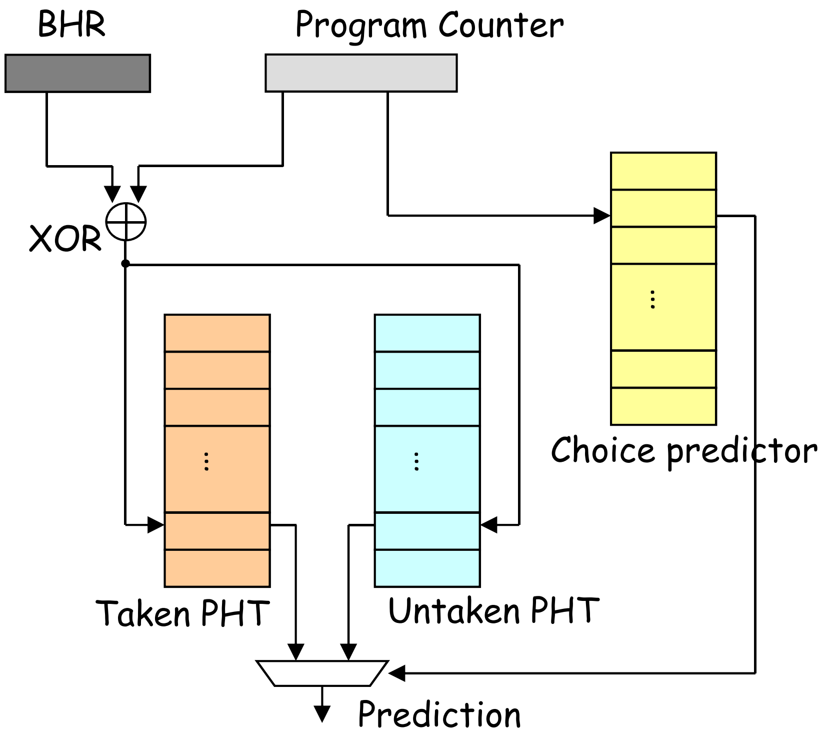
Some typical branch predictors until 2004



- ISCA (International Symposium on Computer Architecture)
- MICRO (International Symposium on Microarchitecture)
- PACT (International Conference on Parallel Architectures and Compilation Techniques)
- ASPLOS (International Conference on Architectural Support for Programming Languages and Operating Systems)

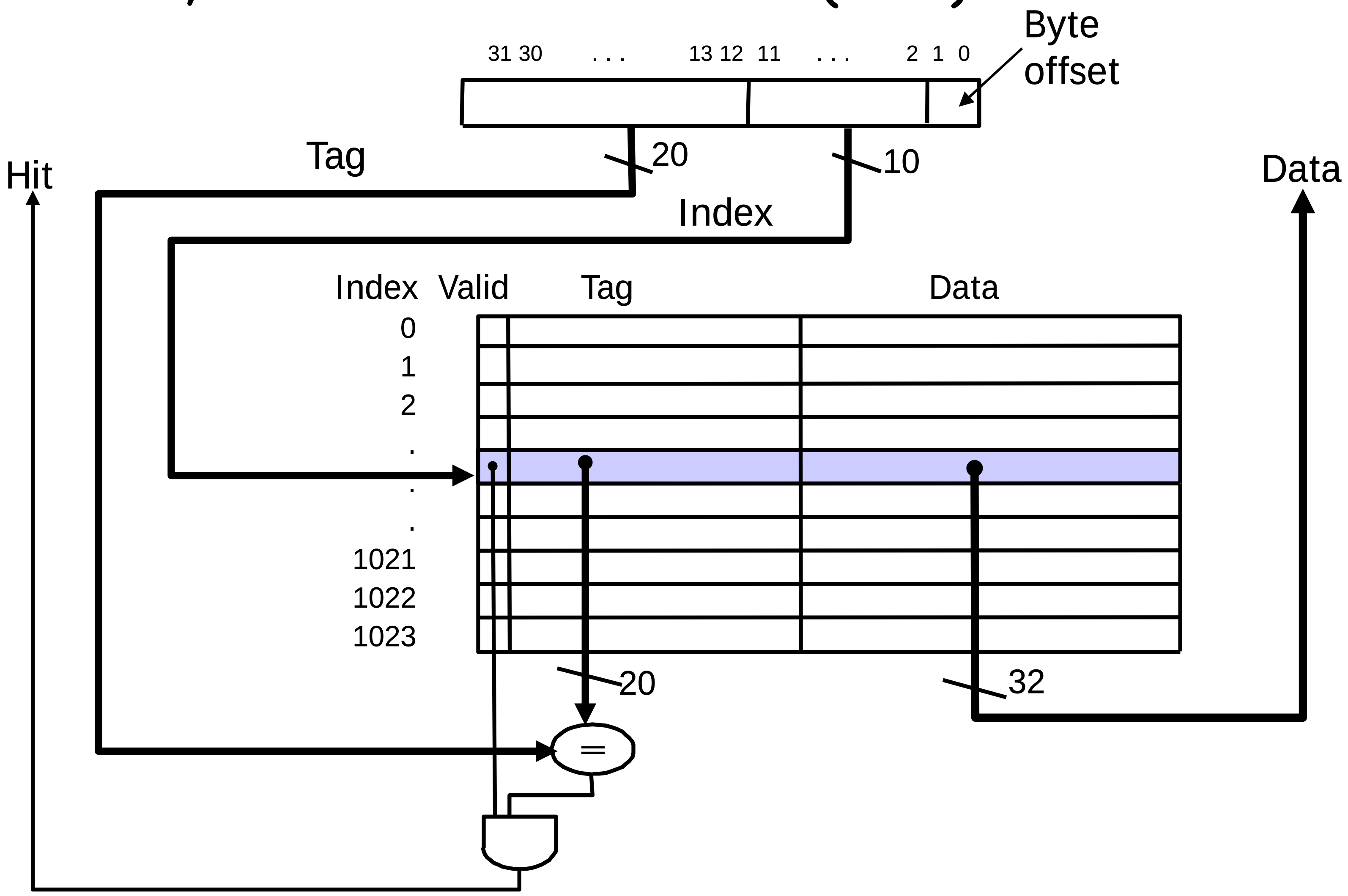
Bi-Mode (MICRO 1997)

- A choice predictor (bimodal) is used as a meta-predictor
- How to predict
  - Like Gshare, both of Taken PHT and Untaken PHT make two predictions.
  - Select one among them by the choice predictor which tracks the global bias of a branch.
- How to update
  - The used PHT is updated in the same way as 2BC.
  - Choice predictor is update in the same way as bimodal



MIPS Direct Mapped Cache Example

- One word/block, cache size = 1K words (4KB)



What kind of locality are we taking advantage of?

YAGS, Yet Another Global Scheme (MICRO 1998)

- Using two tagged PHTs
- When a PHT miss, choice PHT makes a prediction.

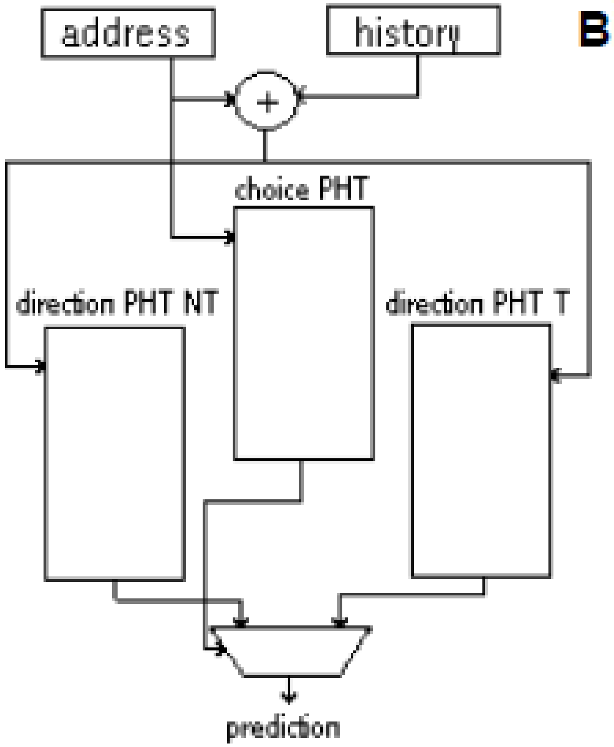


Figure 3.
Bi-Mode

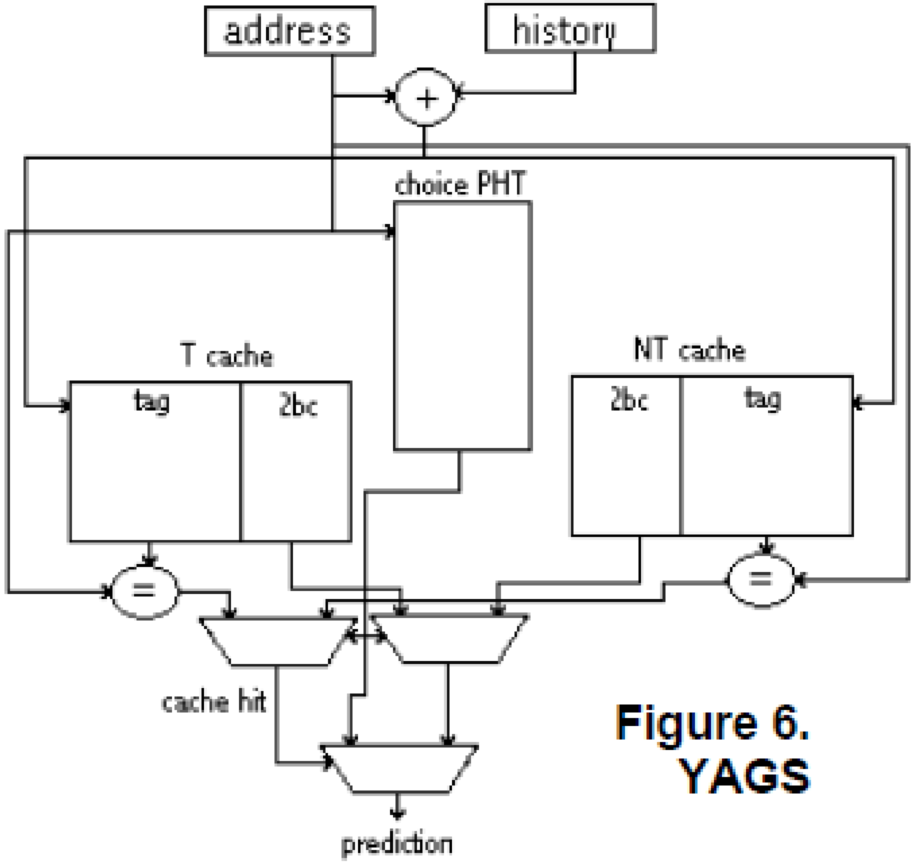
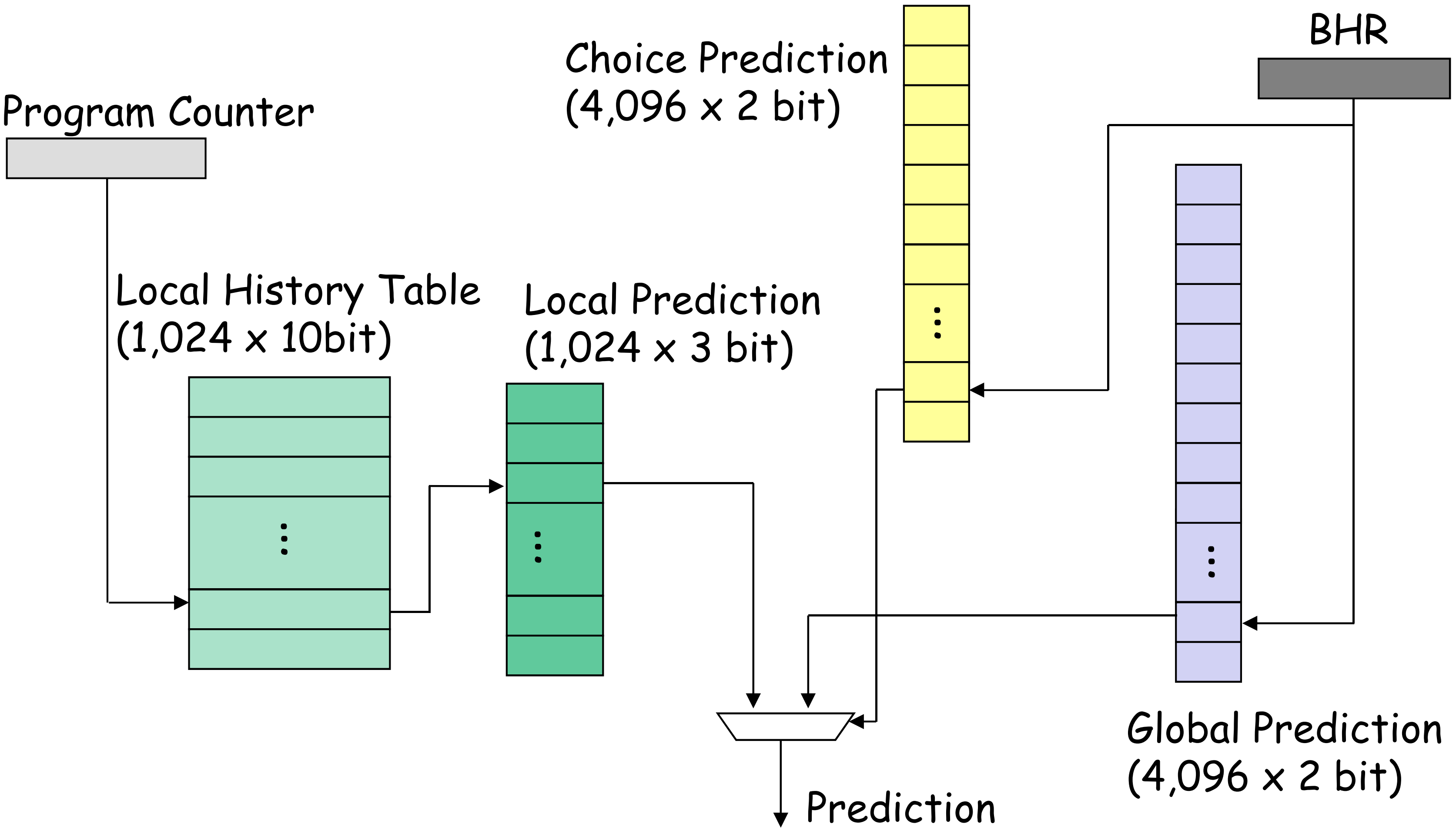


Figure 6.
YAGS

From YAGS paper

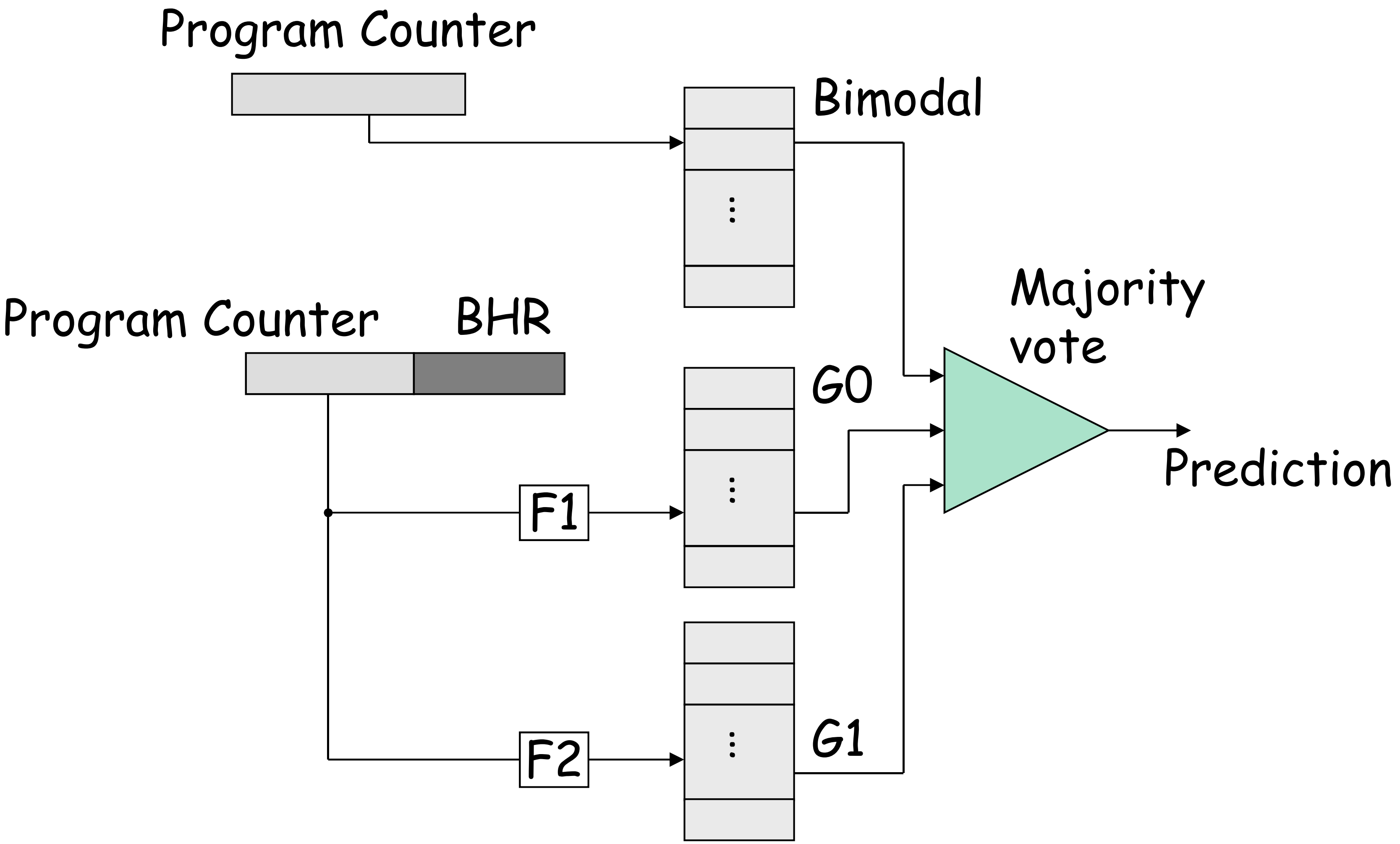
Alpha 21264's hybrid branch predictor in 1996

- A hybrid of local prediction and global prediction implemented in DEC Alpha 21264 which was the state-of-the art commercial processor.
- A choice predictor is used as a meta-predictor



E-gskew (ISCA 1997)

- Using not a meta-predictor but a majority vote



An innovation in branch predictors in 1993

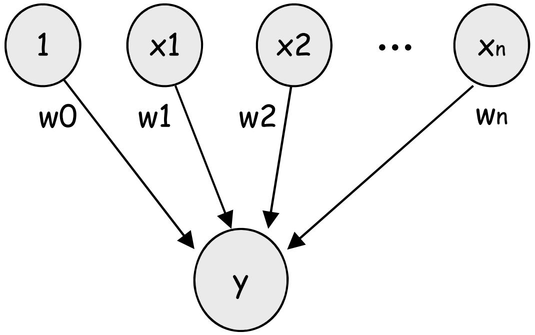
- Using branch history
  - global branch history
  - local branch history
- 2-level branch predictor and Gshare

- Assume predicting the sequence 1110 1110 1110 1110 1110 …

```
11101110 ?               11101110 ?
111011101 ?              111011101 ?
1110111011 ?             1110111011 ?
11101110111 ?            11101110111 ?
111011101110 ?           111011101110 ?
```

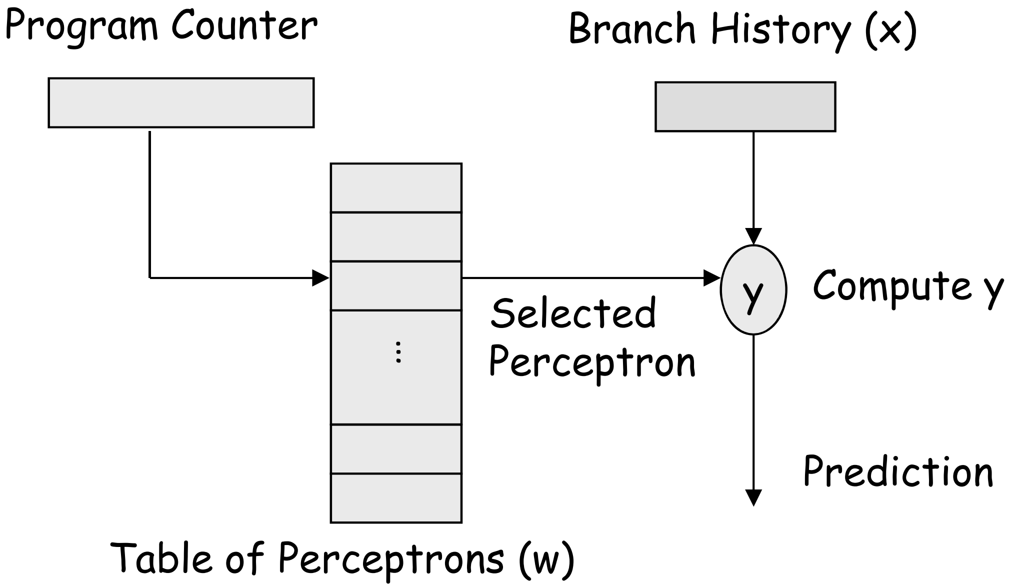
Perceptron (HPCA 2001)

- How to predict
  - Select one perceptron by PC
  - Compute y using the equation. It predicts 1 if y>=0, predicts 0 if y<0
- How to update
  - Train the weights of used perceptron when the prediction miss or |y| < T



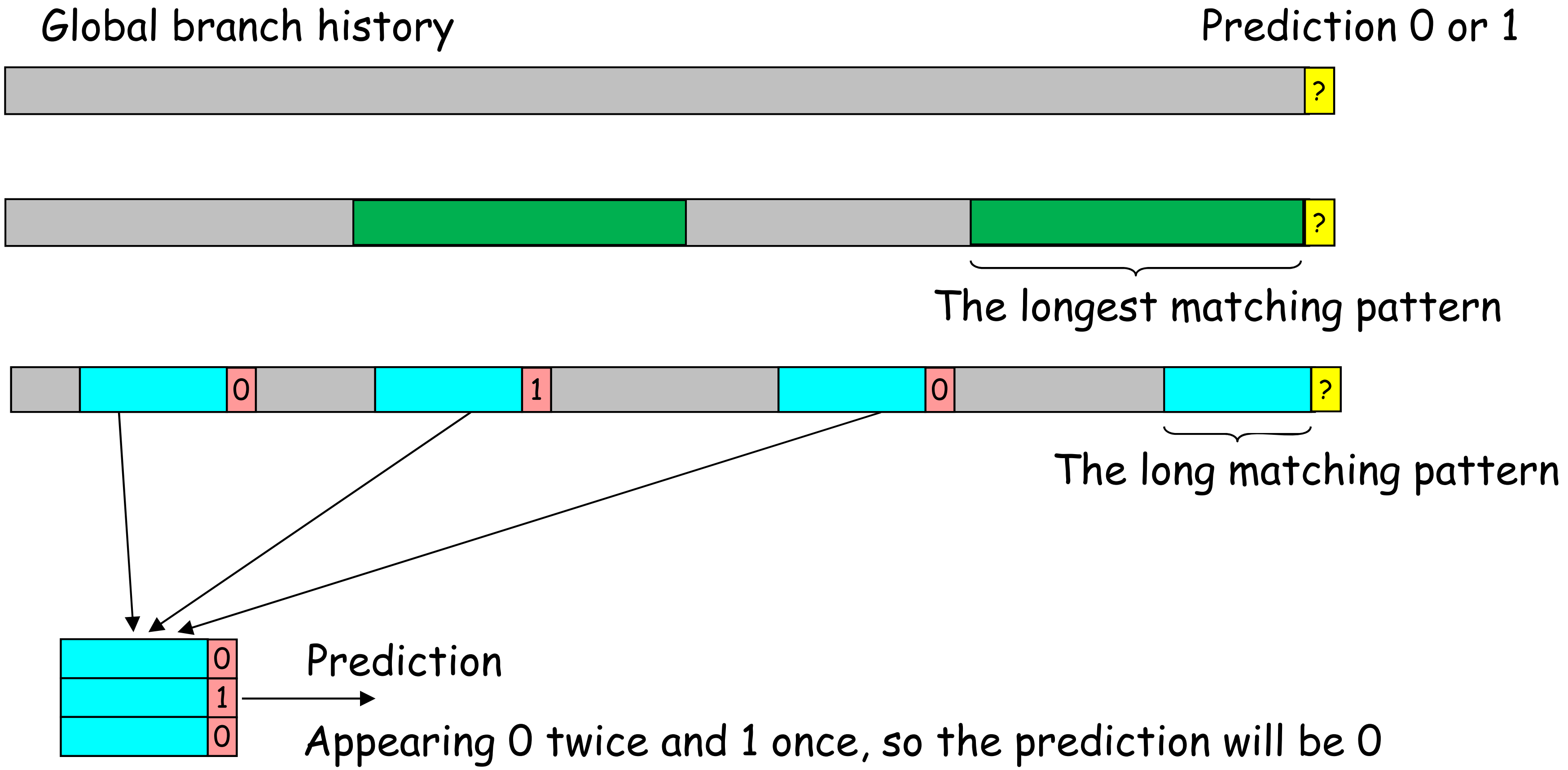
Perceptron Model

$$y = w_0 + \sum_{i=1}^{n} x_i w_i .$$



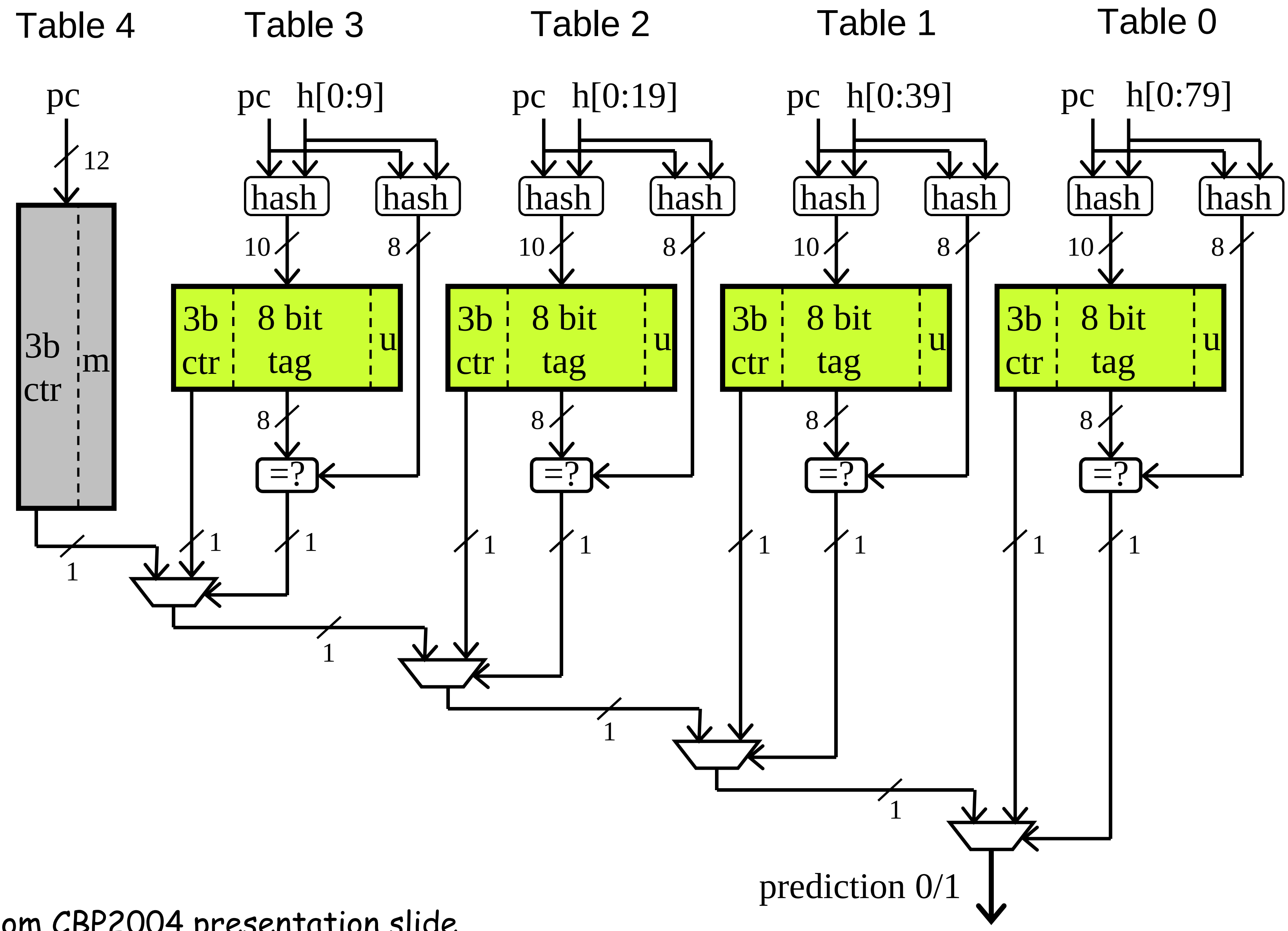
Branch predictors based on pattern matching

- Find the longest matching pattern (green rectangle)
- Select the proper matching length or long matching pattern (blue rectangle)
- Count the number of 0 and the number of 1 after the pattern (red rectangle), then predict.



Appearing 0 twice and 1 once, so the prediction will be 0

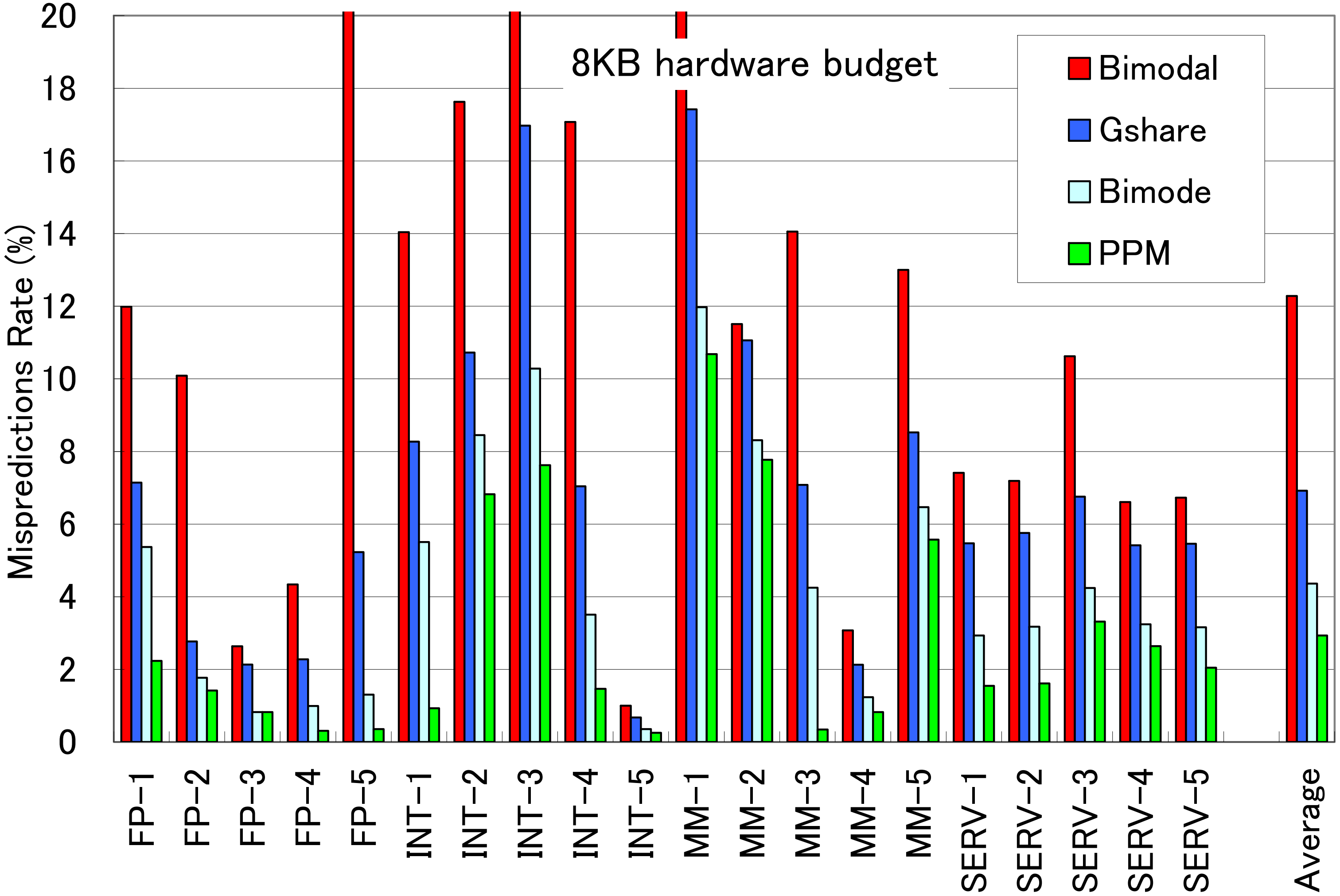
Partial Pattern Matching (CBP 2004)



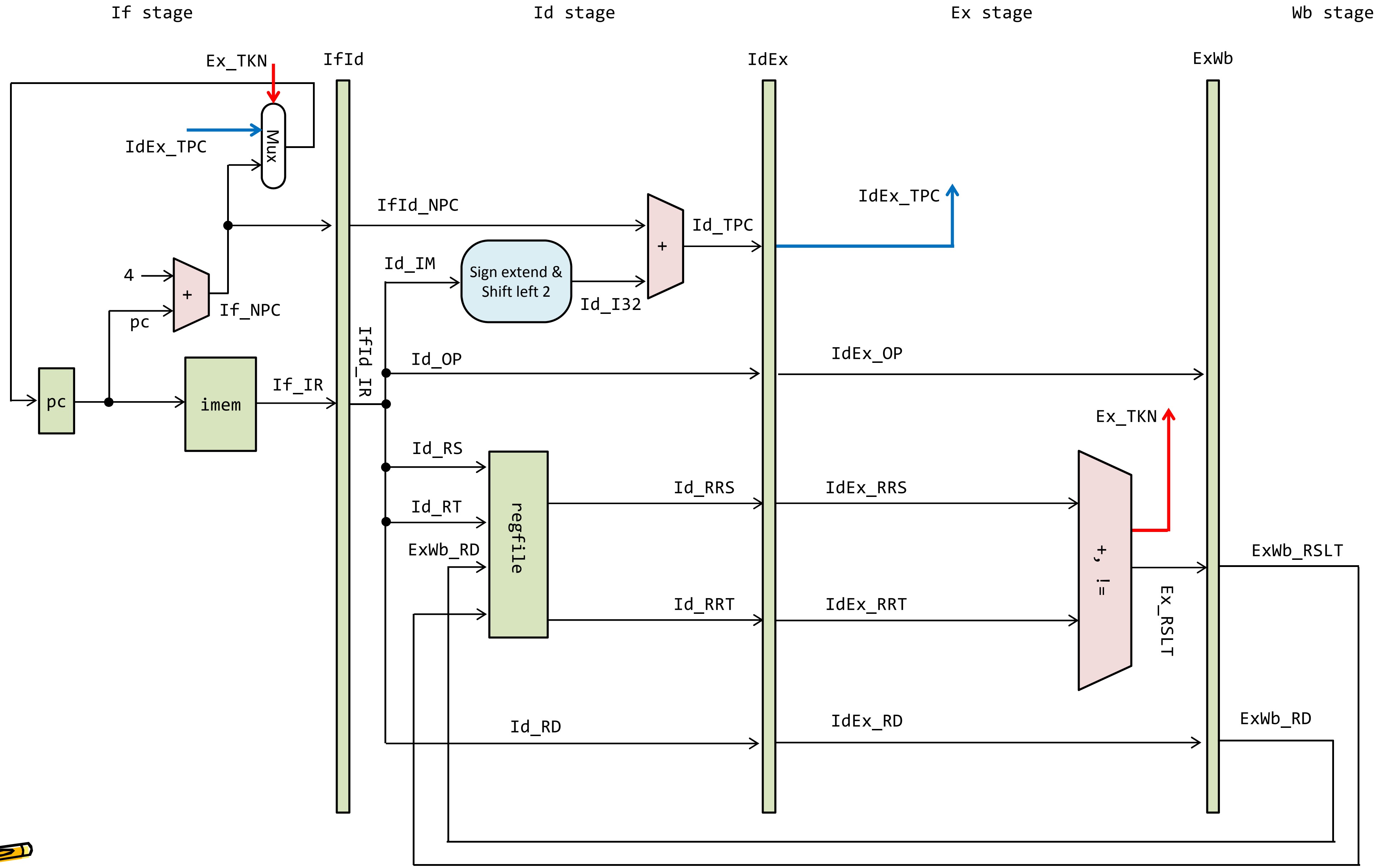
From CBP2004 presentation slide

Prediction accuracy

- The accuracy of 4KB Gshare is about 93%.
- The accuracy of 4KB PPM is about 97%.

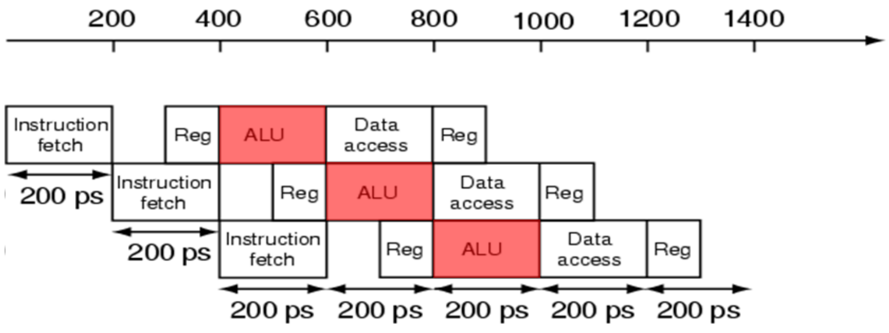


Four stage pipelined processor supporting ADD and BNE, which does not adopt data forwarding (proc08.v, Assignment 6)

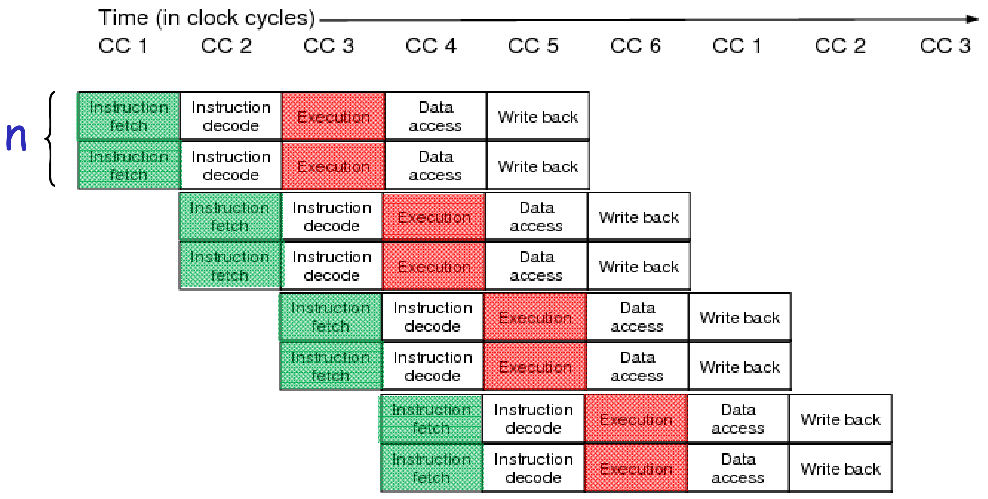


Scalar and Superscalar processors

- Scalar processor can execute at most one single instruction per clock cycle using one ALU.
  - IPC (Executed Instructions Per Cycle) is less than 1.
- Superscalar processor can execute more than one instruction per clock cycle by executing multiple instructions using multiple pipelines.
  - IPC (Executed Instructions Per Cycle) can be more than 1.
  - using n pipelines is called n-way superscalar



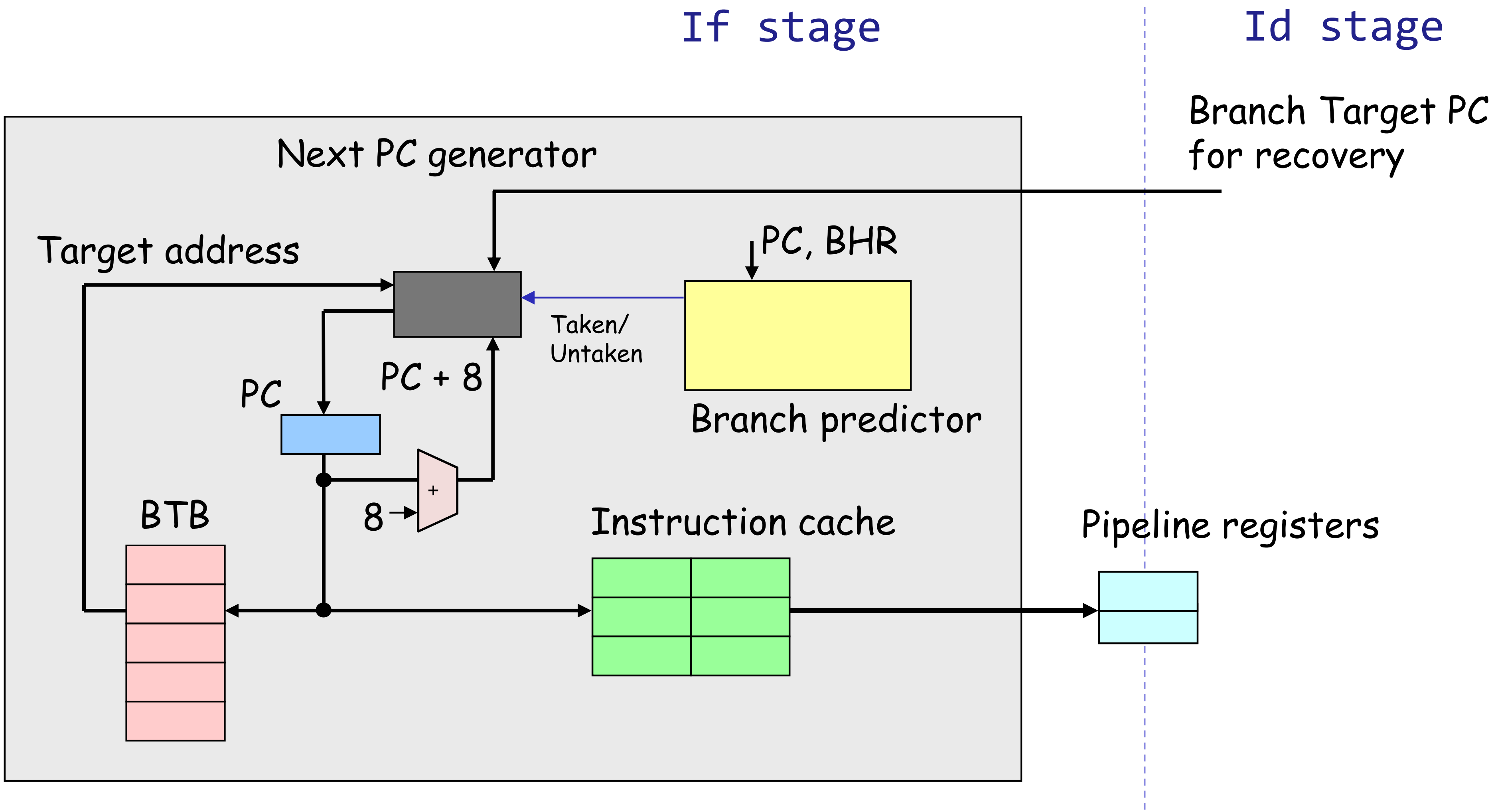
(a) pipeline diagram of scalar processor



(b) pipeline diagram of 2-way superscalar processor

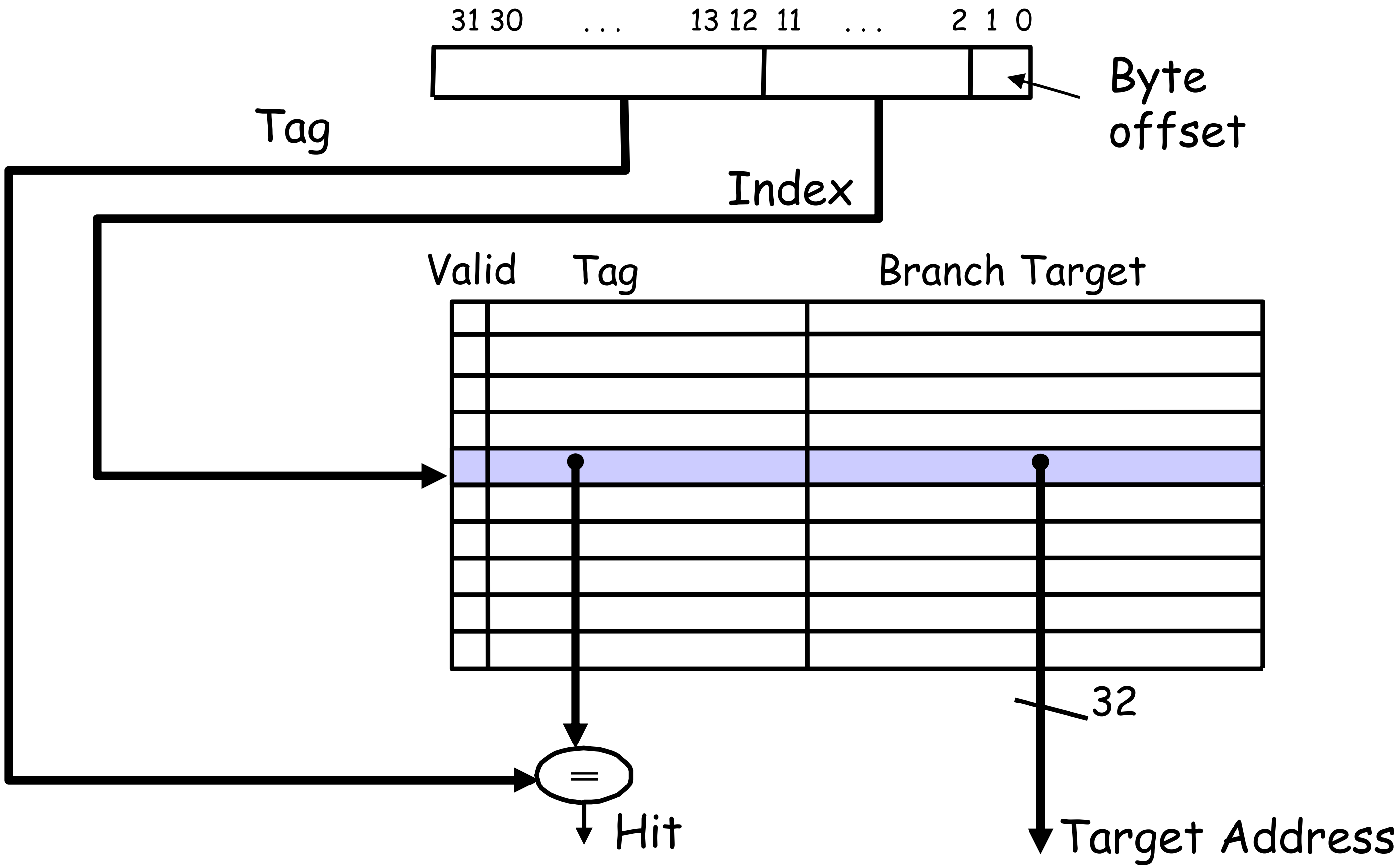
Instruction fetch unit in IF stage

- For high-bandwidth instruction delivery, prediction, and speculation



Branch Target Buffer (BTB)

- A processor must know whether the as-yet-undecoded instruction is a branch and, if so, what the next program counter.
- A branch-prediction cache that stores the predicted address for the next instruction after a branch is called branch target buffer (BTB).



Recommended Reading

- Prophet-Critic Hybrid Branch Prediction
  - Ayose Falcon, UPC, Jared Stark, Intel, Alex Ramirez, UPC, Konrad Lai, Intel, Mateo Valero
  - ISCA-31 pp. 250-261 (2004)

A quote from Introduction (1/2)

Conventional predictors are analogous to a taxi with just one driver.

He gets the passenger to the destination using knowledge of the roads acquired from previous trips; i. e., using history information stored in the predictor's memory structures.

When he reaches an intersection, he uses this knowledge to decide which way to turn.

The driver accesses this knowledge in the context of his current location.

Modern branch predictors access it in the context of the current location (the program counter) plus a history of the most recent decisions that led to the current location.

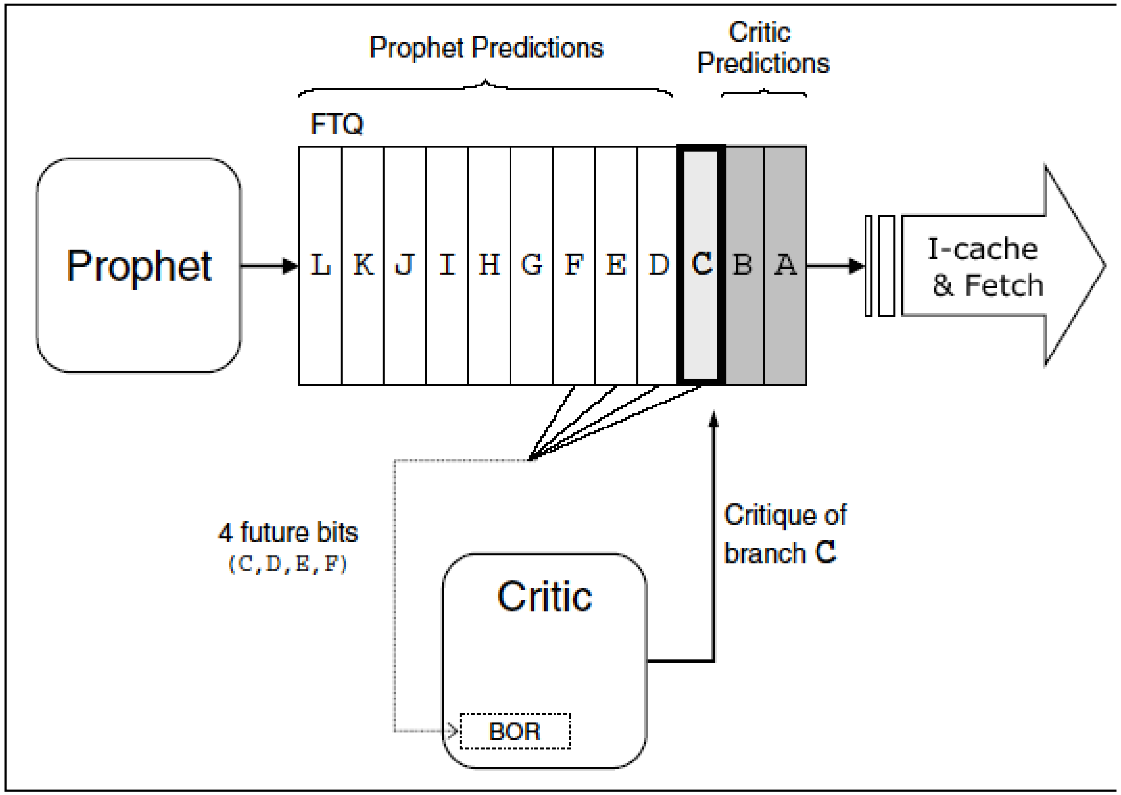
A quote from Introduction (2/2)

Prophet/critic hybrids are analogous to a taxi with two drivers: the front-seat and the back-seat. The front-seat driver has the same role as the driver in the single-driver taxi. This role is called the prophet.

The back-seat driver has the role of critic. She watches the turns the prophet makes at intersections. She doesn't say anything unless she thinks he's made a wrong turn. When she thinks he's made a wrong turn, she waits until he's made a few more turns to be certain they are lost. (Sometimes the prophet makes turns that initially look questionable, but, after he makes a few more turns, in hindsight appear to be correct.) Only when she's certain does she point out the mistake.

To recover, they backtrack to the intersection where she believes the wrong-turn was made and try a different direction.

Prophet-Critic Hybrid Branch Prediction



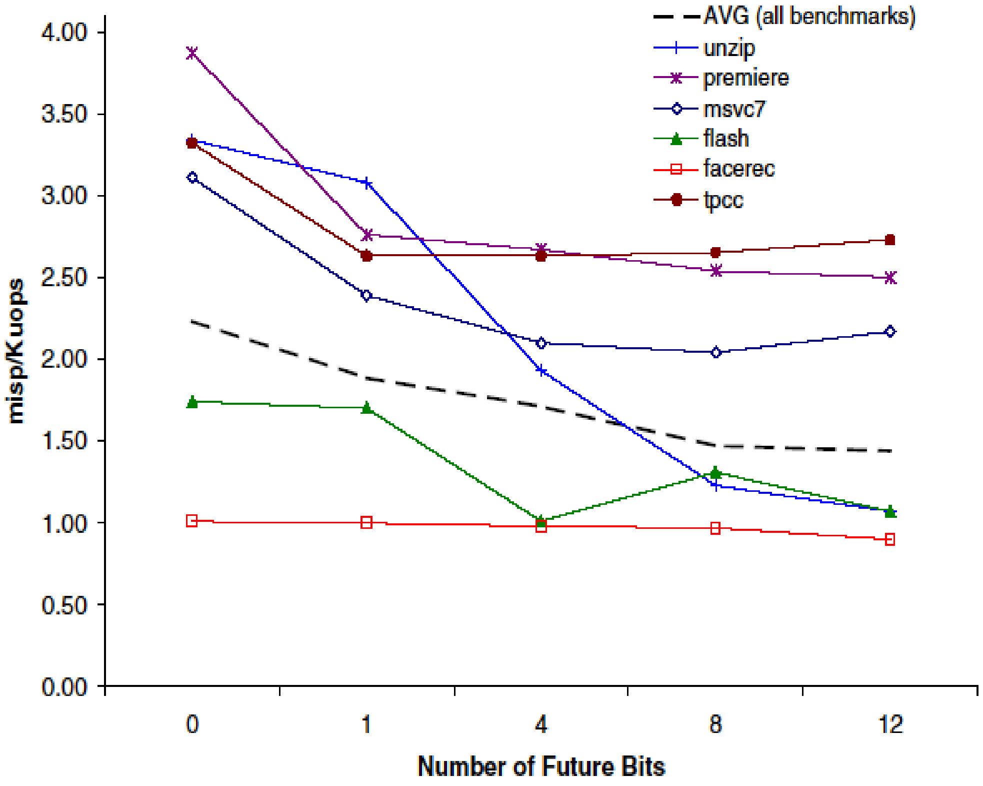


Figure 5. Effect of varying the number of future bits used by the critic on prediction accuracy for selected benchmarks. (prophet: 8KB perceptron; critic: 8KB tagged gshare)