ECE 571 – Advanced Microprocessor-Based Design Lecture 8

Vince Weaver

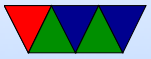
http://web.eece.maine.edu/~vweaver

vincent.weaver@maine.edu

16 February 2017

Announcements

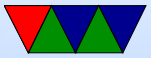
- HW4 Due
- HW5 will be posted



HW#3 Review

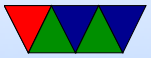
Energy

	sleep	stream	matrix	iozone
cores	0.05	135.1	21.0	0.97
gpu	0.00	0.00	0.00	0.00
pkg	49.37	228.8	32.4	42.4
ram	6.52	25.38	1.23	5.35
time	10.0	9.533	1.38	7.96



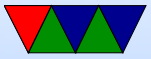
Power

	sleep	stream	matrix	iozone
cores	0.01	14.2	15.0	0.12
gpu	0.00	0.00	0.00	0.00
pkg	4.94	24	23	5.3
ram	0.65	2.6	0.8	0.6



Energy-delay

	1	2	3	4	5	8	16
E	3028	3529	4305	4456	4910	4788	4648
time	168	133	133	132	136	137	141
ED	508k	469k	572k	588k	668k	656k	655k
ED2	85M	62M	76M	78M	91M	90M	92M
P	18W	27W	32W	34W	36W	35W	33W

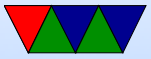


- Does have a GPU. Does seem suspect at 0, but in headless mode. On a similar haswell machine have gotten the GPU to show things but running OpenCL as well as KSP.

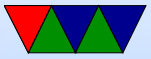
- Energy-delay-squared is E*t*t.
  Not E*E*t, or (E*t)*(E*t)

- Could we show weak scaling? No, problem size is staying same.

- Why did it stop scaling at 4?
  Poorly written benchmark (possible) Not enough

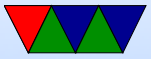


memory (not really, this is a 2001 benchmark) Because even though it has 8 threads, only has 4 cores (likely)



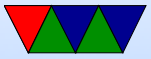
HW#4 Review

- RAPL measurement paper
- EUV paper



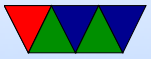
The Branch Problem

- With a pipelined processor, you may have to stall waiting until branch outcome known before you can correctly fetch the next instruction
- Conditional branches are common. On average every 5th instruction [cite?]
- What can you do to speed things up?



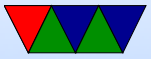
Branch Prediction

- One solution is speculative execution.
  Guess which way branch goes.
- Good branch predictors have a 95% or higher hit rate
- Downsides?
  If wrong, in-flight thrown out, have to replay.
- **Speculation wastes power**



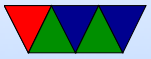
Branch Predictor Implementations

How would you implement a predictor?



Static Prediction of Conditional Branches

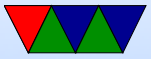
- Backward taken
- Forward not taken
- Can be used as fallback when nothing else is available



Common Access Patterns – For Loop

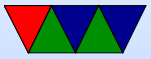
```
for(i=0;i<100;i++) SOMETHING;
```

```
    mov r1,#0
label:
    SOMETHING

    add r1,r1,#1
    cmp r1,#100
    bne label
```



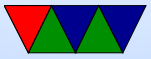
`for` Branch Behavior

- No branch predictor – 100 stalls
- Other way to avoid problem (branch delay slot on MIPS)
- Static prediction BTFN – 99 times predicted right, 1 time wrong (exit)
- 99% correct predict rate (for this particular)
  Depends on iterations, 50% to 99+%



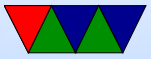
Common Access Patterns – While Loop

```
x=0; while(x<100) { SOMETHING; x++;}

     mov r1,#0
label:
     cmp r1,#100
     bge done

     SOMETHING

     add r1,r1,#1
     b label
done:
```



`while` Branch Behavior

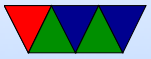
- No branch predictor – 100 stalls (unless branch delay slot)
- Static BTFN prediction – 99 times predicted right, 1 time wrong (exit)
- 99% correct predict rate
- Optimizing compiler may translate this to a for loop (why?)



Common Access Patterns – Do/While Loop

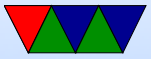
```
x=0; do { SOMETHING; x++;} while(x<100);
```

```
    mov r1,#0
label:
    SOMETHING

    add r1,r1,#1
    cmp r1,#100
    blt label

    b label
done:
```



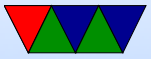
`while` Do/While Behavior

- No branch predictor – 100 stalls (unless branch delay slot)
- Static BTFN prediction – 99 times predicted right, 1 time wrong (exit)
- 99% correct predict rate



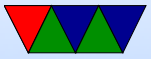
Notes

Optimizing compiler will optimize all above to same for loop (tried it). Why? Because loop unrolling becomes possible?



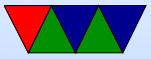
Common Access Patterns – If/Then

```
if (x) { FOO } else { BAR}

   cmp r1,#0
   beq else
then:
   FOO
   b done
else:
   BAR
done:

ARM:
     cmp r1,#0
     FOOeq
     BARne
```

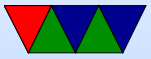


Common Access Patterns – If/Then Behavior

- If x is true, static = 100%, if x is false, 0%
- Assuming completely random, average 50% miss rate
- ARM can use conditional execution/predication to avoid this in simple scenarios

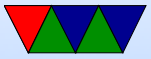


How can we Improve Things?

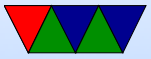


Branch Prediction Hints

- Give compiler (or assembler) hints
- likely() (maps to `__builtin_expect()`)
- unlikely()
- on some processors, (p4) hint for static
- others, just move unlikely blocks out of way for better L1I$ performance

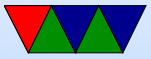


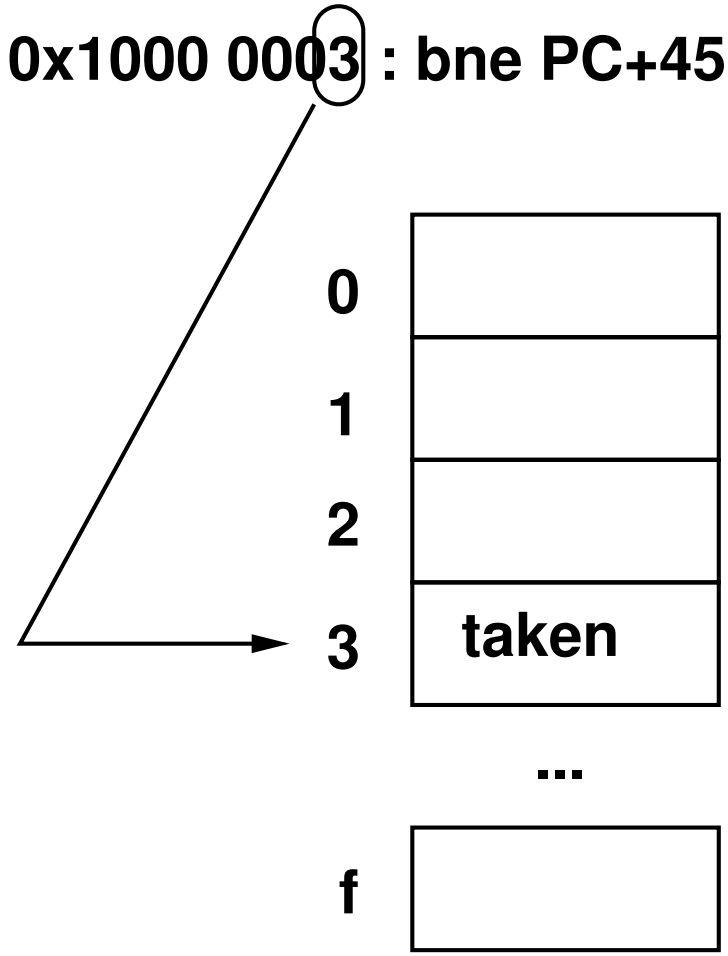
Dynamic Branch Prediction

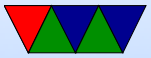


One-bit Branch History Table

- Table, likely indexed by lower bits of instruction
  (Why low bits and not high bits?)
- Can have more advanced indexing to avoid aliasing
  no-tag bits, unlike caches aliasing does not affect correct program execution
- One-bit indicating what the branch did last time
- Update when a branch miss happens

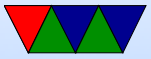



0x1000 0003 : bne PC+45
0
1
2
3
taken
...
f



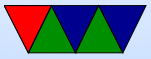
Branch History Table Behavior

- Two misses for each time through loop. Wrong at exit of loop, then wrong again when restarts. (so actually worse than static on loops)
- If/Then potentially better than static if long runs of true/false but can be worse if completely random.



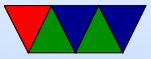
Aliasing

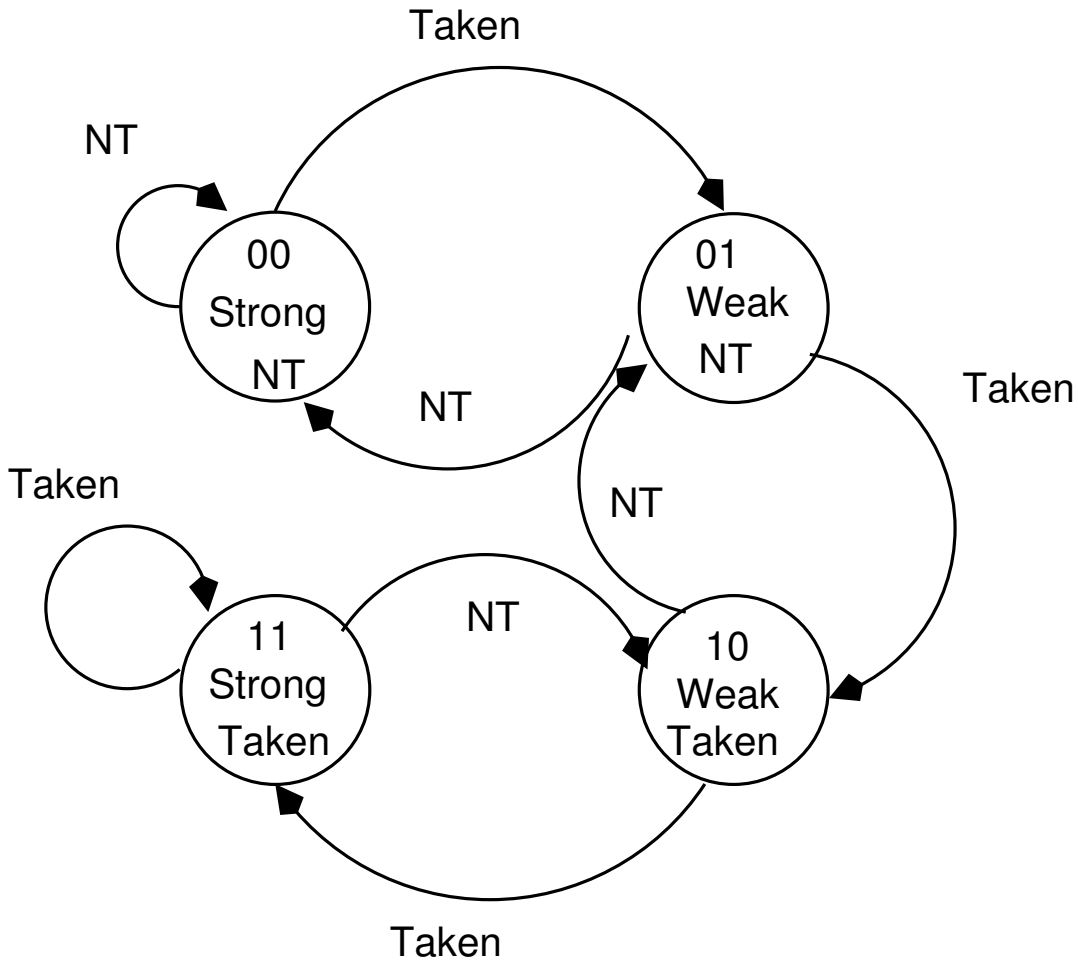
- Is it bad? Good?
- Does the O/S need to save on context switch?
- Do you need to save if entering low-power state?

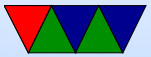


Two-bit saturating counter

- Hysterisis
- Use saturating 2-bit counter
- If 3/2, predict taken, if 1,0 not-taken. Takes two misses or hits to switch from one extreme to the next, letting loops take only one mispredict.
- Needs to be updated on every branch, not just for a mispredict

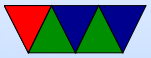



Taken
NT
00
Strong
NT
01
Weak
NT
Taken
NT
NT
Taken
11
Strong
Taken
NT
10
Weak
Taken
Taken

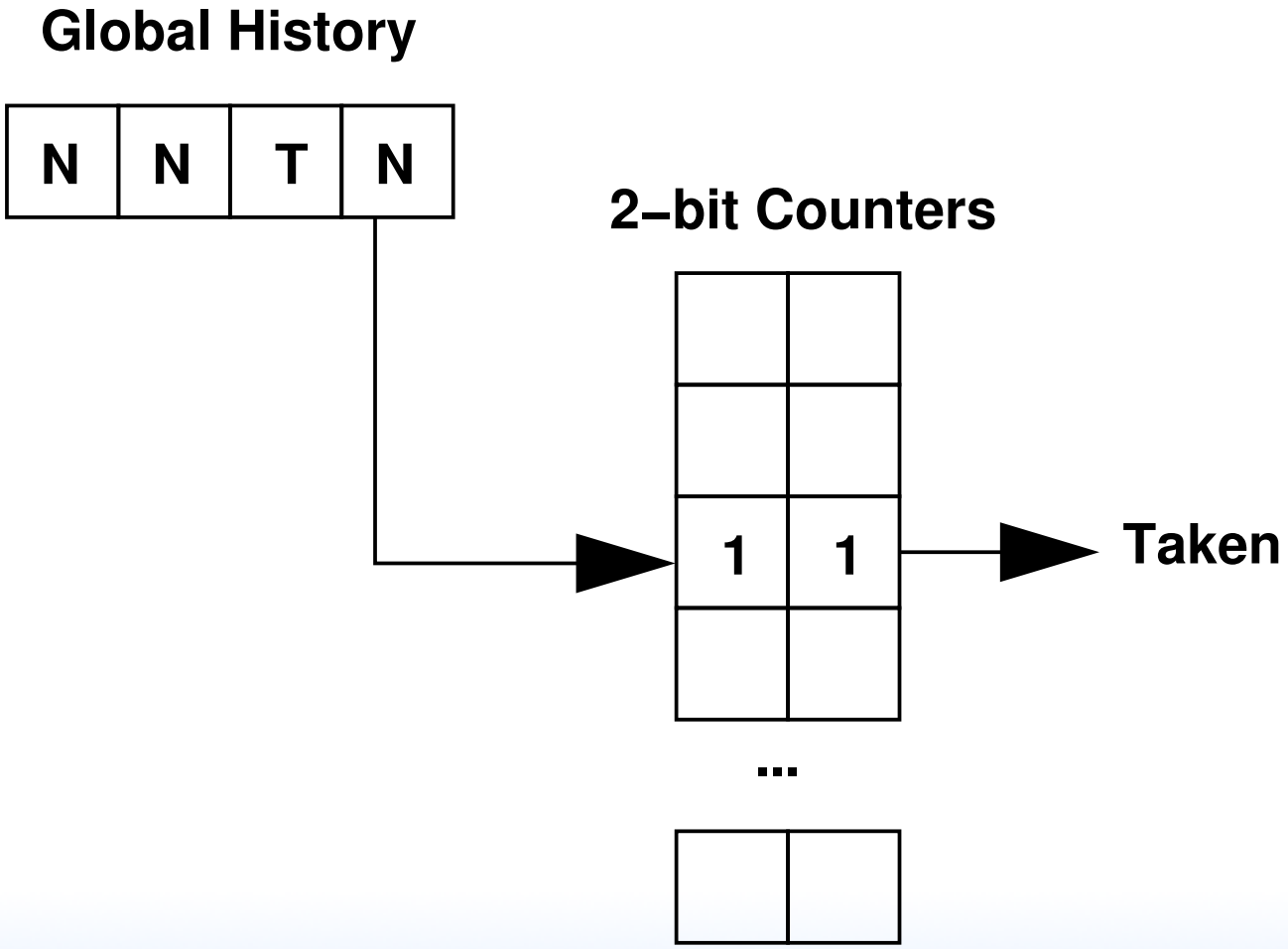


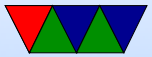
History

- Can use branch history as index into tables
- Use a shift register to hold history
- Global vs Local
  - Global: history is all branches
  - Local: store branch history on a branch by branch basis

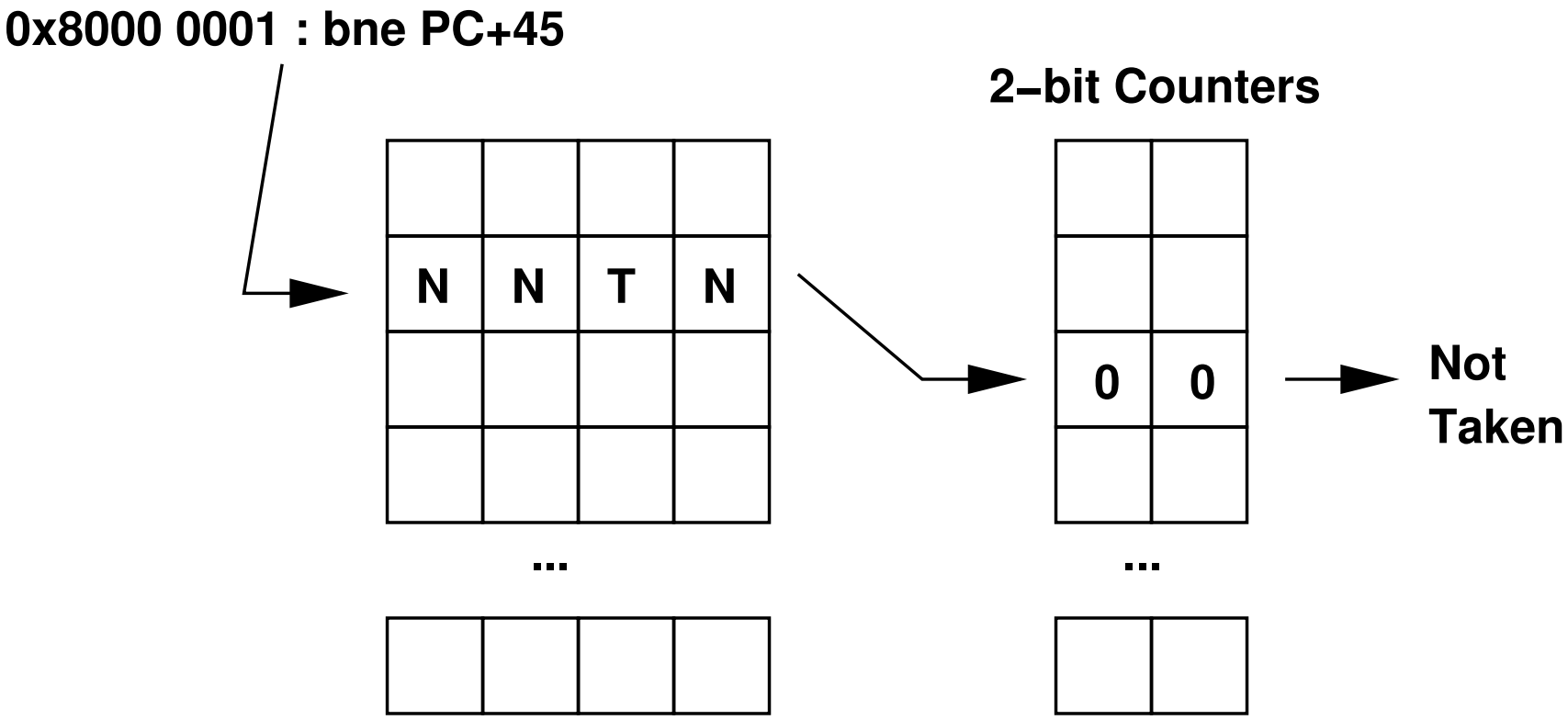


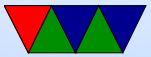
Global Predictor





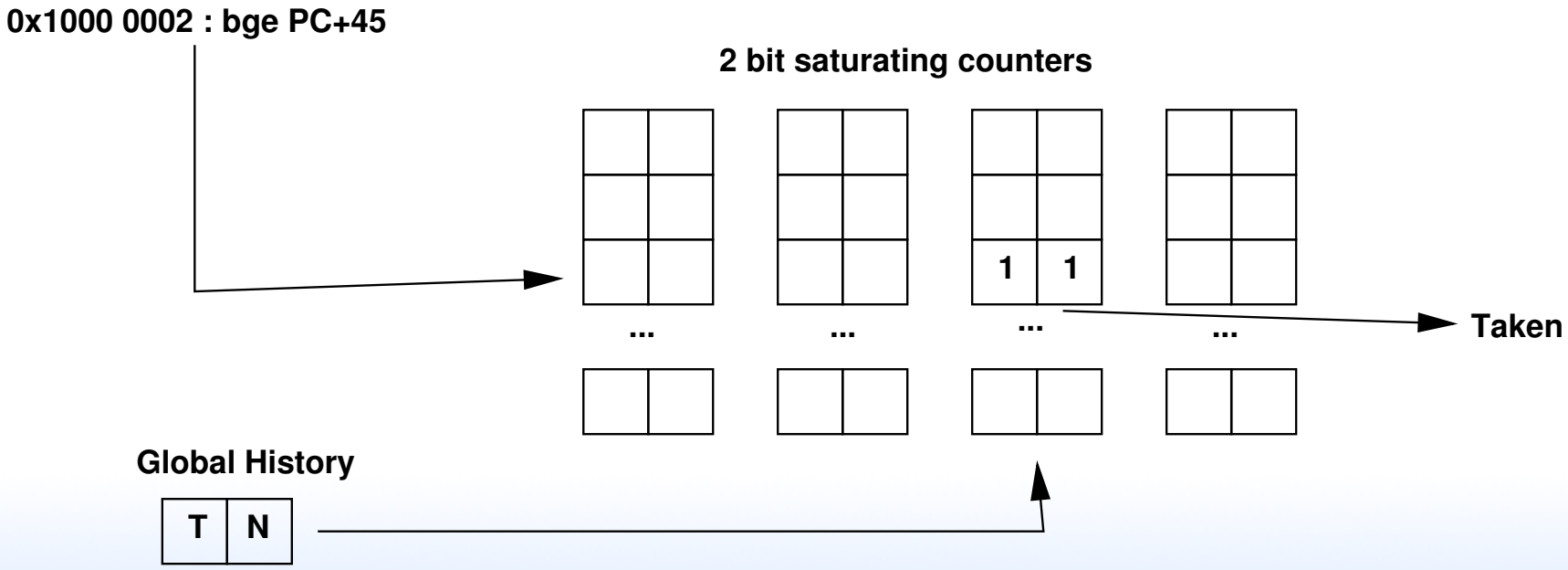
Local Predictor

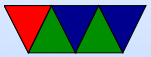




Correlating / Two Level Predictors

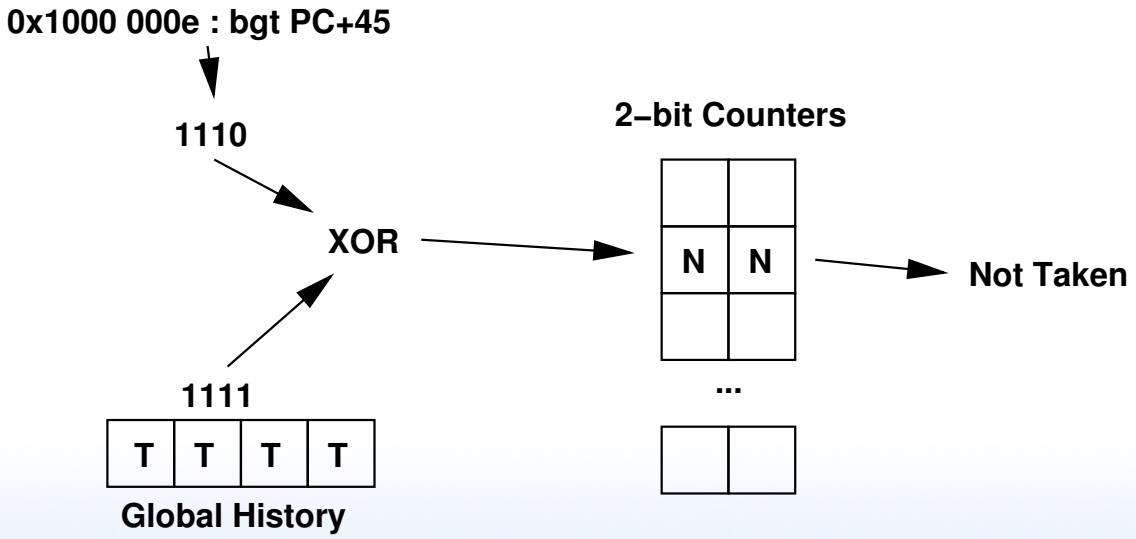
- Take history into account.
  Break branch prediction for a branch out into multiple locations based on history.

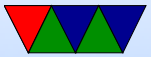




gshare

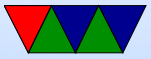
- Xors the global history with the address bits to get which line to use.
- Benefits of 2-level without the extra circuitry





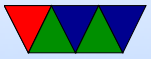
Tournament Predictors

- Which to use? Local or global?
- Have both. How to know which one to use? Predict it!
- 2-bit counter remembers which was best.



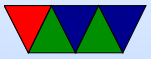
Perceptron

- There are actually Branch Prediction Competitions
- The winner the past few times has been a "Perceptron" predictor
- Neural Networks
- Samsung M1 processor at Hotchips'16 says it has neural-net branch predictor
- AMD Zen uses hashed perceptron



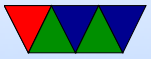
Comparing Predictors

- Branch miss rate not enough
- Usually the total number of bits needed is factored in
- May also need to keep track of logic needed if it is complex.



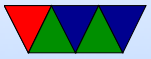
Branch Target Buffer

- Instruction fetcher needs to know next instruction to fetch even before instructions decoded
- Predicts the actual destination of addresses.
- Indexed by whole PC. May be looking up before even know it is a branch instruction.
- Only need to store predicted-taken branches. (Why? Because not-taken fall through as per normal).



Return Address Stack

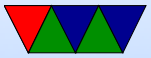
- Function calls can confuse BTB. Multiple locations branching to same spot. Which return address should be predicted?
- Keep a stack of return addresses for function calls
- Playing games with size optimization and fallthrough/tail optimization can confuse.



Adjusting Predictor on the Fly

Some processors let you configure predictor at runtime.
MIPS R12000 let you
ARM possibly does.

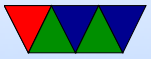
Why is this useful?
In theory if you have a known workload you can pick the one that works best.
Also if realtime you want something that is deterministic, like static prediction.
Also Good for simulator validation



Cortex A9 Branch Predictor

From the Manual:

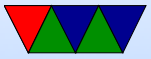
- two-level prediction mechanism, comprising: a two-way BTAC of 512 entries organized as two-way x 256 entries
- a Global History Buffer (GHB) with 4096 2-bit predictors
- a return stack with eight 32-bit entries.
- It is also capable of predicting state changes from ARM to Thumb, and from Thumb to ARM.



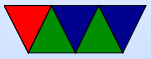
Example

Code in perf_event validation tests for generic events.

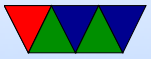
`http://web.eece.maine.edu/~vweaver/projects/perf_events/validation/`



Example Results



```
Part 1
Testing a loop with 1500000 branches (100 times):
On a simple loop like this, miss rate should be very small.
Adjusting domain to 0,0,0 for ARM
Average number of branch misses: 685

Part 2
Adjusting domain to 0,0,0 for ARM

Testing a function that branches based on a random number
   The loop has 7710798 branches.
   500000 are random branches; 250699 of those were taken
Adjusting domain to 0,0,0 for ARM

Out of 7710798 branches, 291081 were mispredicted
Assuming a good random number generator and no freaky luck
The mispredicts should be roughly between 125000 and 375000

Testing ‘‘branch-misses’’ generalized event...                PASSED
```



Value Prediction

- Can we use this mechanism to help other performance issues?
  What about caches?

- Can we predict values loaded from memory?

- Load Value Prediction. You can, sometimes with reasonable success, but apparently not worth trouble as no vendors have ever implemented it.