

18-447

Computer Architecture
Lecture 10: Branch Handling
and Branch Prediction (II)

Prof. Onur Mutlu

Carnegie Mellon University

Spring 2014, 2/5/2014

Readings for Next Few Lectures (I)

- P&H Chapter 4.9-4.11
- Smith and Sohi, “The Microarchitecture of Superscalar Processors,” Proceedings of the IEEE, 1995
 - More advanced pipelining
 - Interrupt and exception handling
 - Out-of-order and superscalar execution concepts
- McFarling, “Combining Branch Predictors,” DEC WRL Technical Report, 1993.
- Kessler, “The Alpha 21264 Microprocessor,” IEEE Micro 1999.

Readings for Next Few Lectures (II)

- Smith and Plezskun, “[Implementing Precise Interrupts in Pipelined Processors](#),” IEEE Trans on Computers 1988 (earlier version in ISCA 1985).

Review: More Sophisticated Direction Prediction

- Compile time (static)
 - ❑ Always not taken
 - ❑ Always taken
 - ❑ BTFN (Backward taken, forward not taken)
 - ❑ Profile based (likely direction)
 - ❑ Program analysis based (likely direction)

- Run time (dynamic)
 - ❑ Last time prediction (single-bit)
 - ❑ Two-bit counter based prediction
 - ❑ Two-level prediction (global vs. local)
 - ❑ Hybrid

Review: Importance of The Branch Problem

- Assume a 5-wide *superscalar* pipeline with 20-cycle branch resolution latency
- How long does it take to fetch 500 instructions?
 - Assume 1 out of 5 instructions is a branch
 - 100% accuracy
 - 100 cycles (all instructions fetched on the correct path)
 - No wasted work
 - 99% accuracy
 - $100 \text{ (correct path)} + 20 \text{ (wrong path)} = 120 \text{ cycles}$
 - 20% extra instructions fetched
 - 98% accuracy
 - $100 \text{ (correct path)} + 20 * 2 \text{ (wrong path)} = 140 \text{ cycles}$
 - 40% extra instructions fetched
 - 95% accuracy
 - $100 \text{ (correct path)} + 20 * 5 \text{ (wrong path)} = 200 \text{ cycles}$
 - 100% extra instructions fetched

Can We Do Better?

- Last-time and 2BC predictors exploit “last-time” predictability
- Realization 1: A branch’s outcome can be correlated with other branches’ outcomes
 - Global branch correlation
- Realization 2: A branch’s outcome can be correlated with past outcomes of the same branch (other than the outcome of the branch “last-time” it was executed)
 - Local branch correlation

Global Branch Correlation (I)

- Recently executed branch outcomes in the execution path is correlated with the outcome of the next branch

```
if (cond1)
...
if (cond1 AND cond2)
```

- If first branch not taken, second also not taken

```
branch Y: if (cond1) a = 2;
...
branch X: if (a == 0)
```

- If first branch taken, second definitely not taken

Global Branch Correlation (II)

branch Y: if (cond1)

...

branch Z: if (cond2)

...

branch X: if (cond1 AND cond2)

- If Y and Z both taken, then X also taken
- If Y or Z not taken, then X also not taken

Global Branch Correlation (III)

■ Eqntott, SPEC 1992

```
if (aa==2)                ;; B1
    aa=0;
if (bb==2)                ;; B2
    bb=0;
if (aa!=bb) {             ;; B3
    ....
}
```

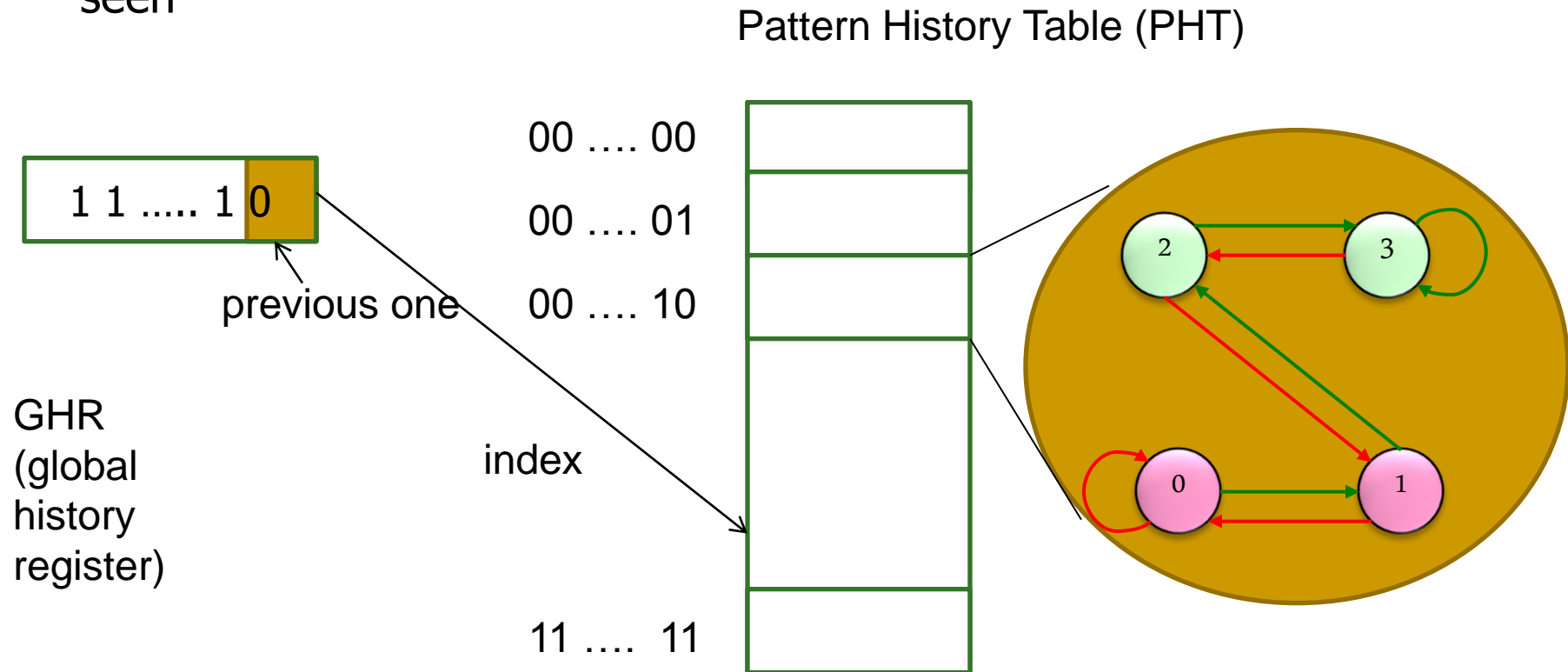
If **B1** is not taken (i.e., $aa==0@B3$) and **B2** is not taken (i.e. $bb=0@B3$)
then **B3** is certainly taken

Capturing Global Branch Correlation

- Idea: Associate branch outcomes with “global T/NT history” of all branches
- Make a prediction based on the outcome of the branch the last time the same global branch history was encountered
- Implementation:
 - Keep track of the “global T/NT history” of all branches in a register → Global History Register (GHR)
 - Use GHR to index into a table that recorded the outcome that was seen for each GHR value in the recent past → Pattern History Table (table of 2-bit counters)
- Global history/branch predictor
- Uses two levels of history (GHR + history at that GHR)

Two Level Global Branch Prediction

- First level: **Global branch history register** (N bits)
 - The direction of last N branches
- Second level: **Table of saturating counters for each history entry**
 - The direction the branch took the last time the same history was seen



How Does the Global Predictor Work?

```
for (i=0; i<100; i++)  
    for (j=0; j<3; j++)
```

After the initial startup time, the conditional branches have the following behavior, assuming GR is shifted to the left:

test	value	GR	result
j<3	j=1	1101	taken
j<3	j=2	1011	taken
j<3	j=3	0111	not taken
i<100		1110	usually taken

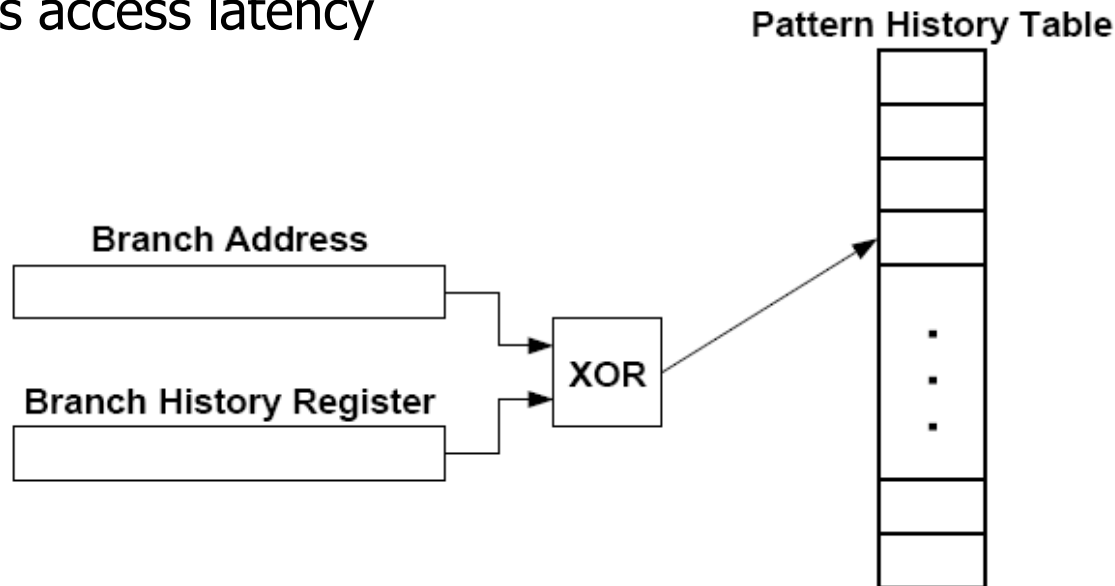
- McFarling, “Combining Branch Predictors,” DEC WRL TR 1993.

Intel Pentium Pro Branch Predictor

- 4-bit global history register
- Multiple pattern history tables (of 2 bit counters)
 - Which pattern history table to use is determined by lower order bits of the branch address

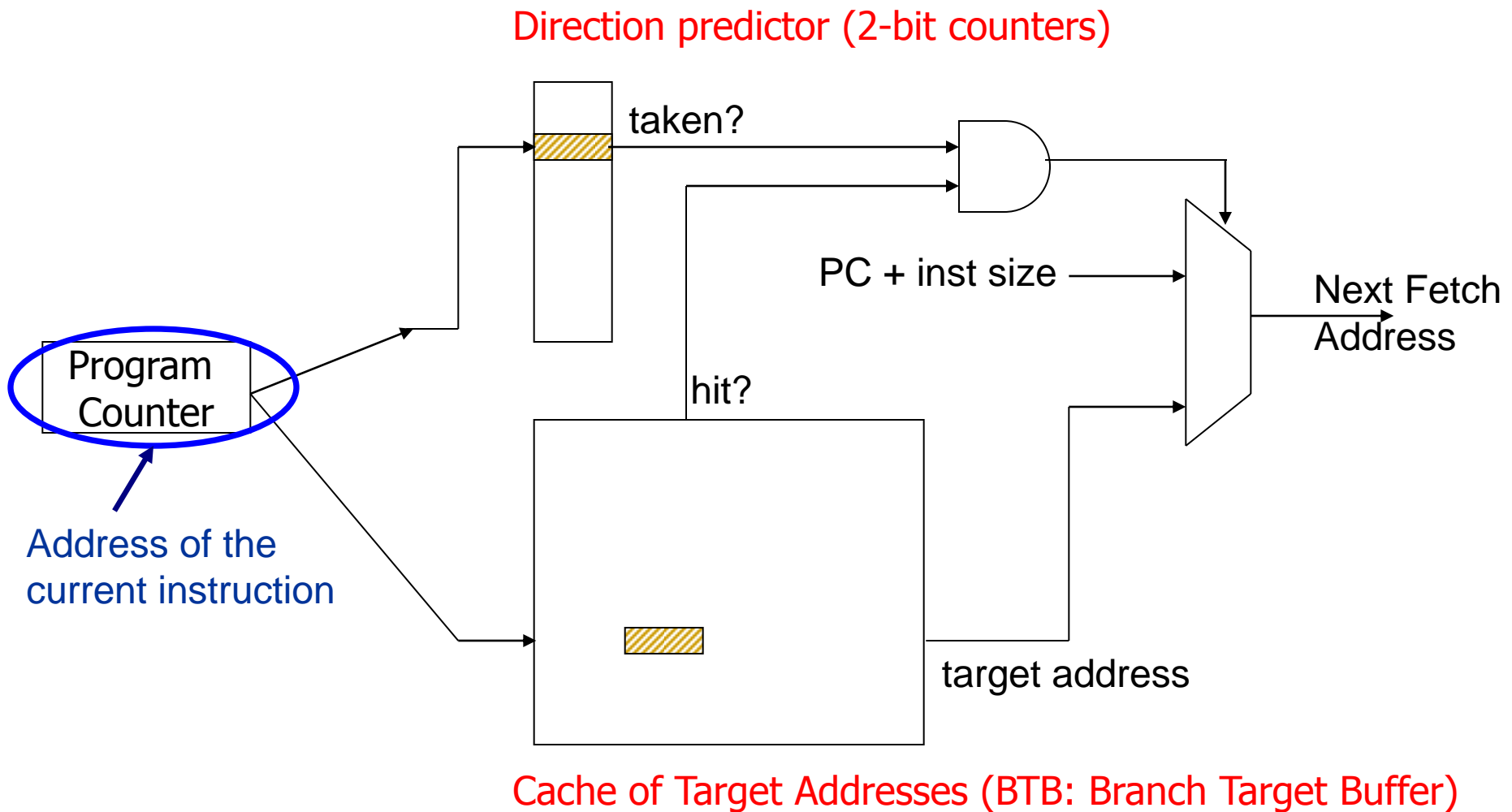
Improving Global Predictor Accuracy

- Idea: Add more context information to the global predictor to take into account which branch is being predicted
 - **Gshare predictor**: GHR hashed with the Branch PC
 - + More context information
 - + Better utilization of PHT
 - Increases access latency

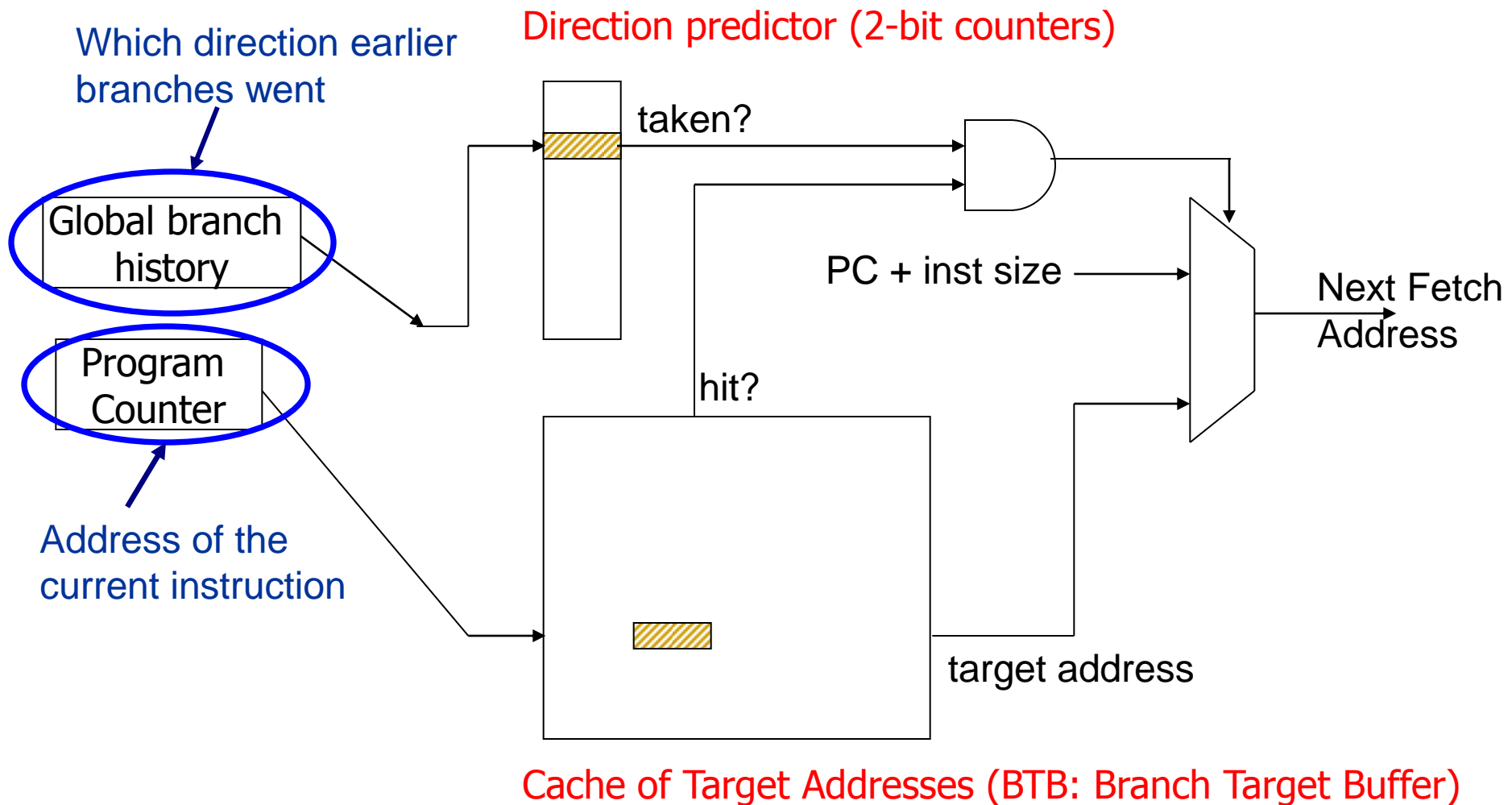


- McFarling, “**Combining Branch Predictors**,” DEC WRL Tech Report, 1993.

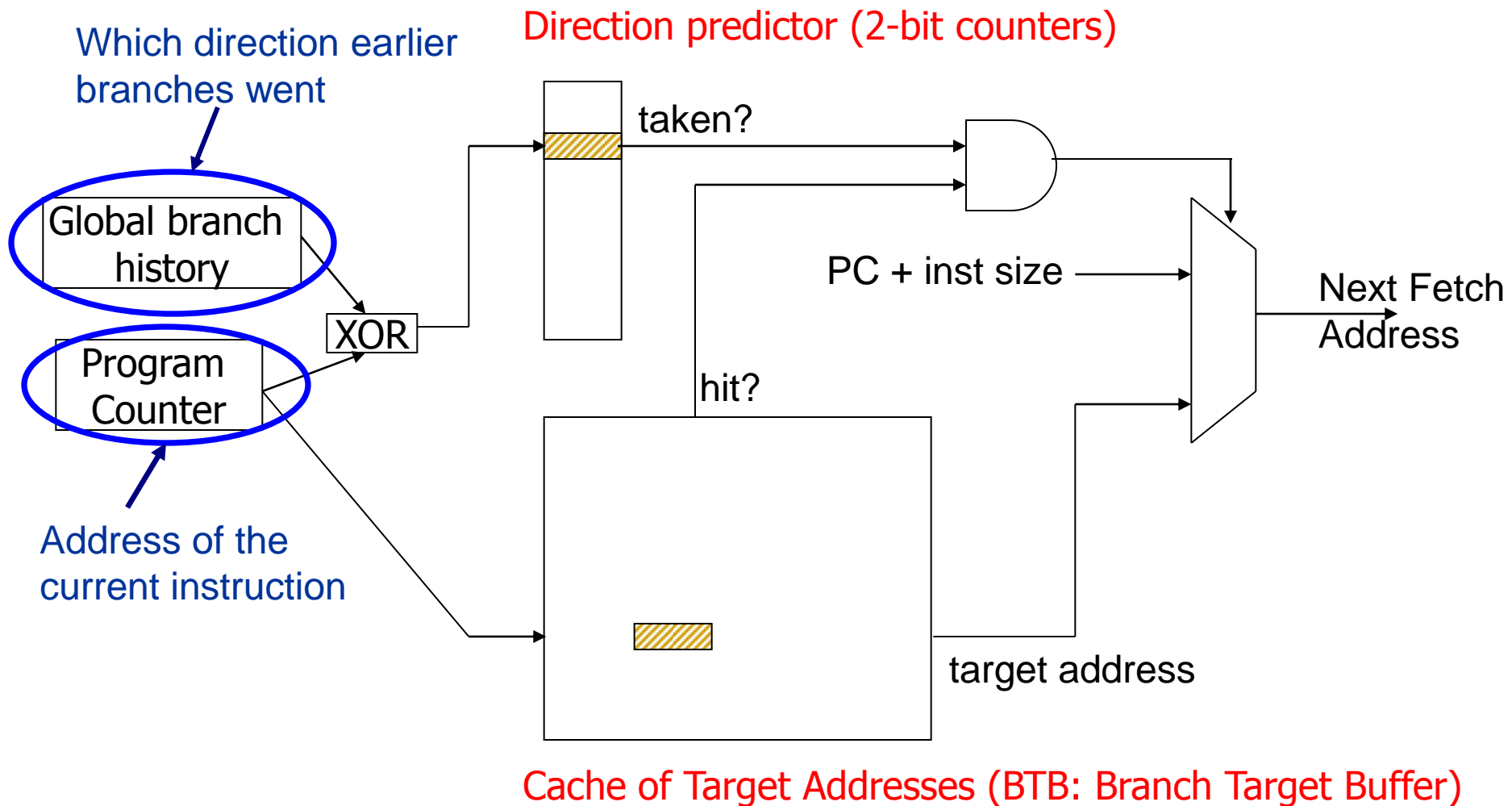
One-Level Branch Predictor



Two-Level Global History Branch Predictor



Two-Level Gshare Branch Predictor



Can We Do Better?

- Last-time and 2BC predictors exploit “last-time” predictability
- Realization 1: A branch’s outcome can be correlated with other branches’ outcomes
 - Global branch correlation
- Realization 2: A branch’s outcome can be correlated with past outcomes of the same branch (other than the outcome of the branch “last-time” it was executed)
 - Local branch correlation

Local Branch Correlation

```
for (i=1; i<=4; i++) { }
```

If the loop test is done at the end of the body, the corresponding branch will execute the pattern $(1110)^n$, where 1 and 0 represent taken and not taken respectively, and n is the number of times the loop is executed. Clearly, if we knew the direction this branch had gone on the previous three executions, then we could always be able to predict the next branch direction.

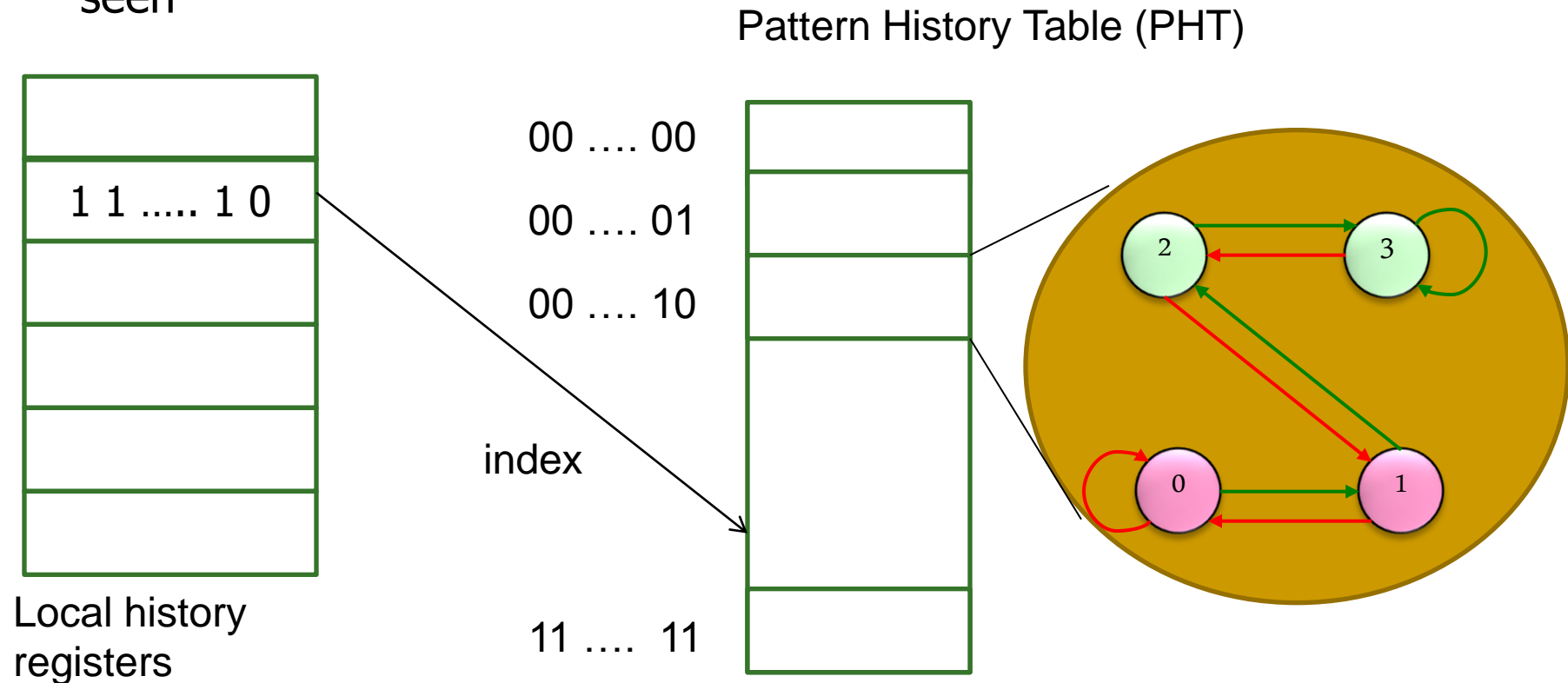
- McFarling, “Combining Branch Predictors,” DEC WRL TR 1993.

Capturing Local Branch Correlation

- Idea: Have a per-branch history register
 - Associate the predicted outcome of a branch with “T/NT history” of the same branch
- Make a prediction based on the outcome of the branch the last time the same local branch history was encountered
- Called the local history/branch predictor
- Uses two levels of history (Per-branch history register + history at that history register value)

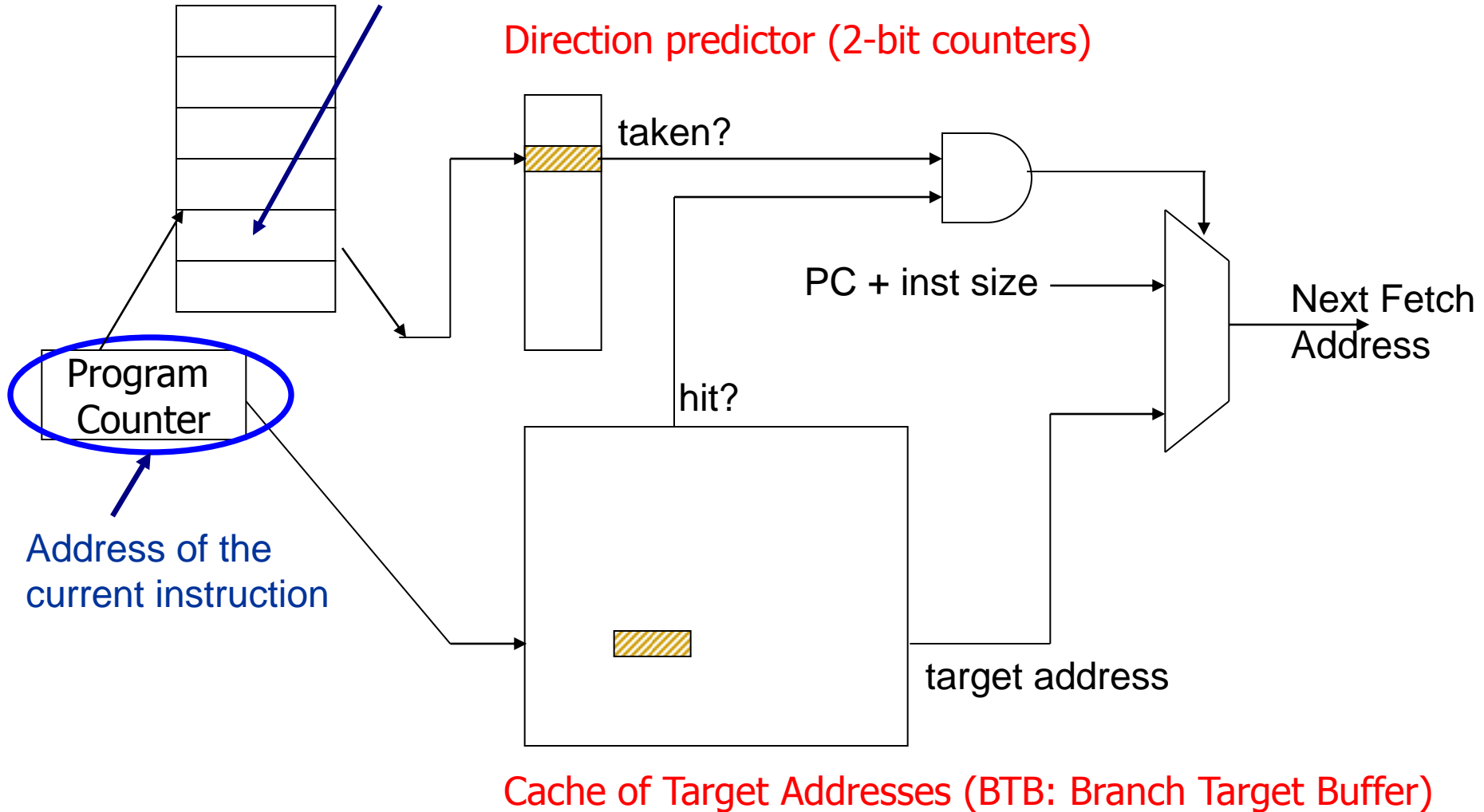
Two Level Local Branch Prediction

- First level: A set of local history registers (N bits each)
 - Select the history register based on the PC of the branch
- Second level: Table of saturating counters for each history entry
 - The direction the branch took the last time the same history was seen



Two-Level Local History Branch Predictor

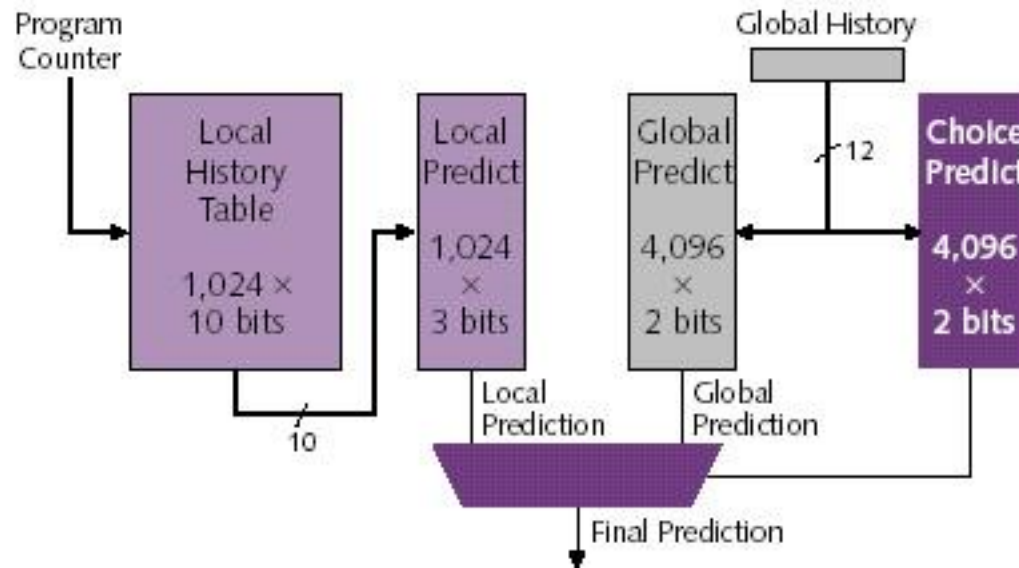
Which directions earlier instances of *this branch* went



Hybrid Branch Predictors

- Idea: Use more than one type of predictor (i.e., multiple algorithms) and select the “best” prediction
 - E.g., hybrid of 2-bit counters and global predictor
- Advantages:
 - + Better accuracy: different predictors are better for different branches
 - + Reduced **warmup** time (faster-warmup predictor used until the slower-warmup predictor warms up)
- Disadvantages:
 - Need “meta-predictor” or “selector”
 - Longer access latency
- McFarling, “**Combining Branch Predictors**,” DEC WRL Tech Report, 1993.

Alpha 21264 Tournament Predictor



- Minimum branch penalty: 7 cycles
- Typical branch penalty: 11+ cycles
- 48K bits of target addresses stored in I-cache
- Predictor tables are reset on a context switch
- Kessler, "The Alpha 21264 Microprocessor," IEEE Micro 1999.

Branch Prediction Accuracy (Example)

- Bimodal: table of 2bc indexed by branch address

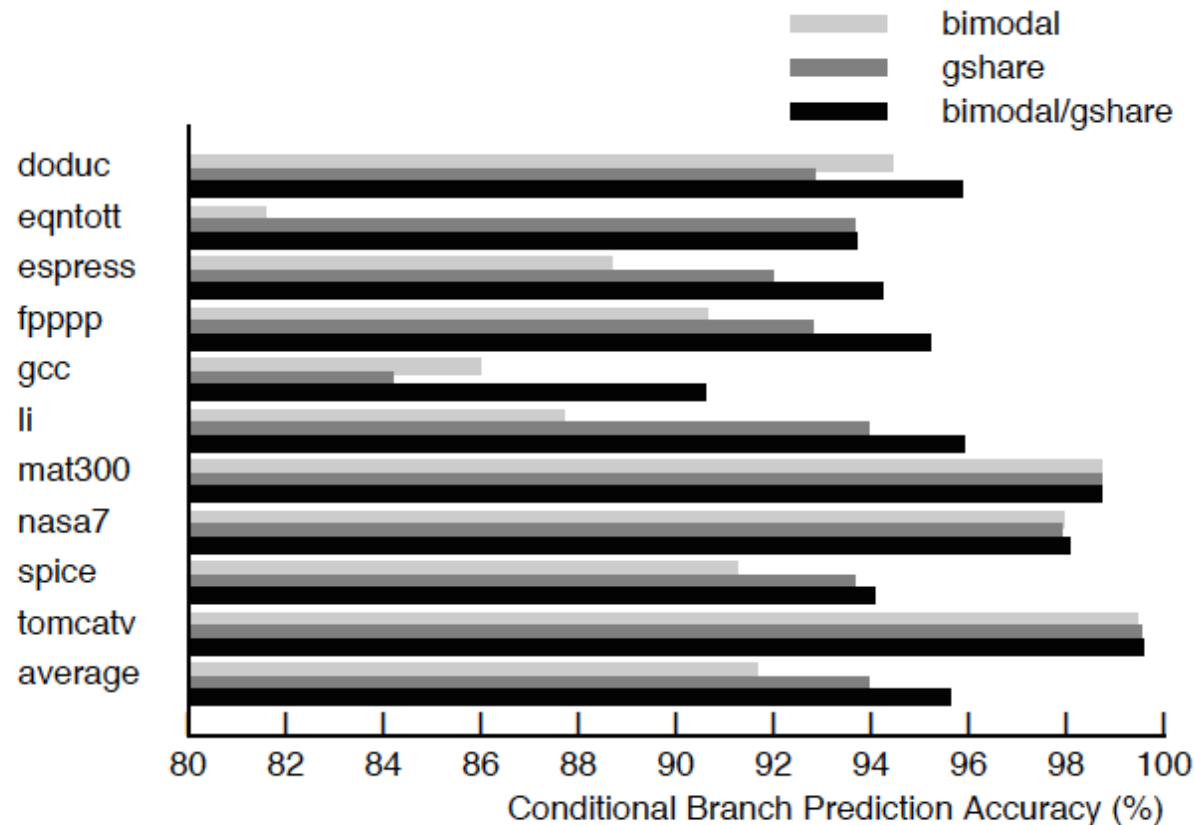


Figure 13: Combined Predictor Performance by Benchmark

Biased Branches

- Observation: Many branches are biased in one direction (e.g., 99% taken)
- Problem: These branches *pollute* the branch prediction structures → make the prediction of other branches difficult by causing “interference” in branch prediction tables and history registers
- Solution: Detect such biased branches, and predict them with a simpler predictor
- Chang et al., “Branch classification: a new mechanism for improving branch predictor performance,” MICRO 1994.

How to Handle Control Dependences

- Critical to keep the pipeline full with correct sequence of dynamic instructions.
- Potential solutions if the instruction is a control-flow instruction:
 - Stall the pipeline until we know the next fetch address
 - Guess the next fetch address (branch prediction)
 - Employ delayed branching (branch delay slot)
 - Do something else (fine-grained multithreading)
 - Eliminate control-flow instructions (predicated execution)
 - Fetch from both possible paths (if you know the addresses of both possible paths) (multipath execution)

Review: Predicate Combining (*not* Predicated Execution)

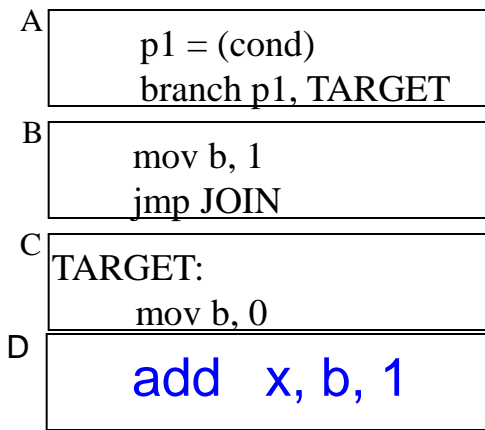
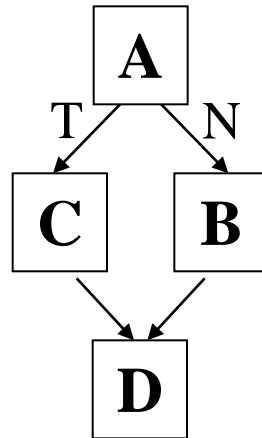
- Complex predicates are converted into multiple branches
 - if ((a == b) && (c < d) && (a > 5000)) { ... }
 - 3 conditional branches
- Problem: This increases the number of control dependencies
- Idea: Combine predicate operations to feed a single branch instruction
 - Predicates stored and operated on using condition registers
 - A single branch checks the value of the combined predicate
- + Fewer branches in code → fewer mipredictions/stalls
- Possibly unnecessary work
 - If the first predicate is false, no need to compute other predicates
- Condition registers exist in IBM RS6000 and the POWER architecture

Predication (Predicated Execution)

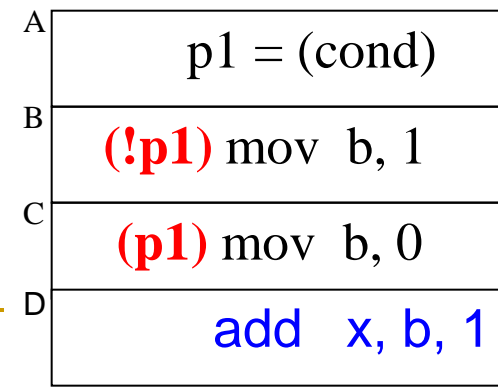
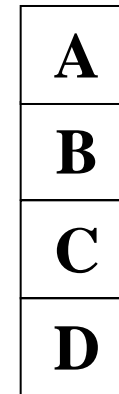
- Idea: Compiler converts control dependence into data dependence → branch is eliminated
 - Each instruction has a predicate bit set based on the predicate computation
 - Only instructions with TRUE predicates are committed (others turned into NOPs)

(normal branch code)

```
if (cond) {  
    b = 0;  
}  
else {  
    b = 1;  
}
```



(predicated code)



Conditional Move Operations

- Very limited form of predicated execution
- CMOV R1 \leftarrow R2
 - R1 = (ConditionCode == true) ? R2 : R1
 - Employed in most modern ISAs (x86, Alpha)

Review: CMOV Operation

- Suppose we had a Conditional Move instruction...
 - CMOV condition, $R1 \leftarrow R2$
 - $R1 = (\text{condition} == \text{true}) ? R2 : R1$
 - Employed in most modern ISAs (x86, Alpha)
- Code example with branches vs. CMOVs
if (a == 5) {b = 4;} else {b = 3;}

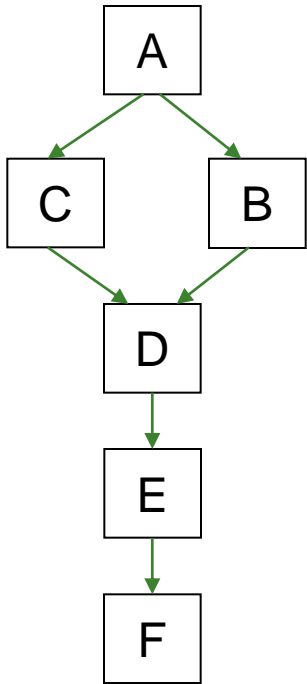
CMPEQ condition, a, 5;

CMOV condition, b \leftarrow 4;

CMOV !condition, b \leftarrow 3;

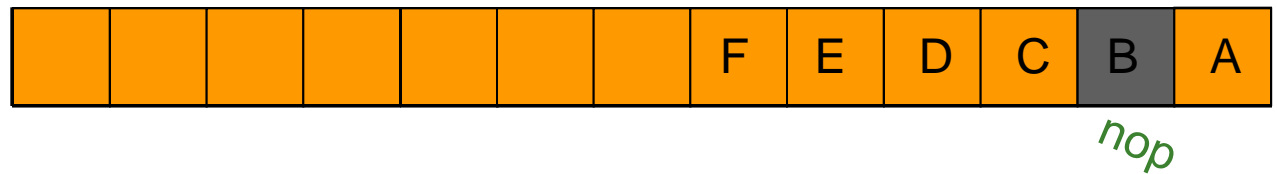
Predicated Execution (II)

- Predicated execution can be high performance and energy-efficient



Predicated Execution

Fetch Decode Rename Schedule RegisterRead Execute



Branch Prediction

Fetch Decode Rename Schedule RegisterRead Execute



Pipeline flush!!

Predicated Execution (III)

■ Advantages:

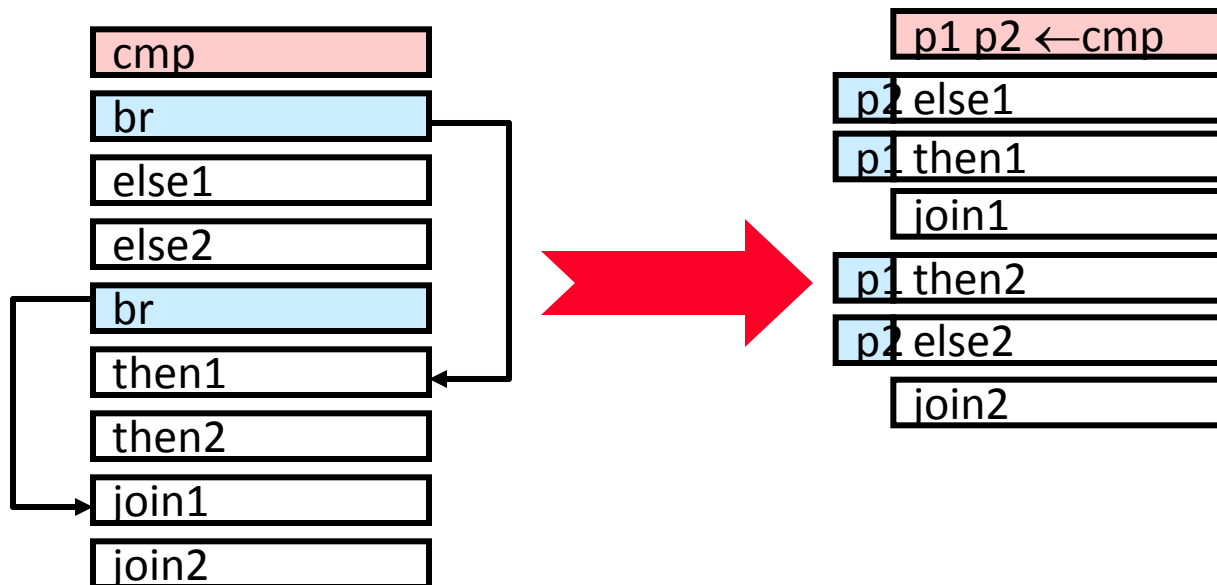
- + Eliminates mispredictions for hard-to-predict branches
 - + No need for branch prediction for some branches
 - + Good if misprediction cost > useless work due to predication
- + Enables code optimizations hindered by the control dependency
 - + Can move instructions more freely within predicated code

■ Disadvantages:

- Causes useless work for branches that are easy to predict
 - Reduces performance if misprediction cost < useless work
 - **Adaptivity**: Static predication is not adaptive to run-time branch behavior. Branch behavior changes based on input set, program phase, control-flow path.
- Additional hardware and ISA support
- Cannot eliminate all hard to predict branches
 - Loop branches?

Predicated Execution in Intel Itanium

- Each instruction can be separately predicated
- 64 one-bit predicate registers
 - each instruction carries a 6-bit predicate field
- An instruction is effectively a NOP if its predicate is false



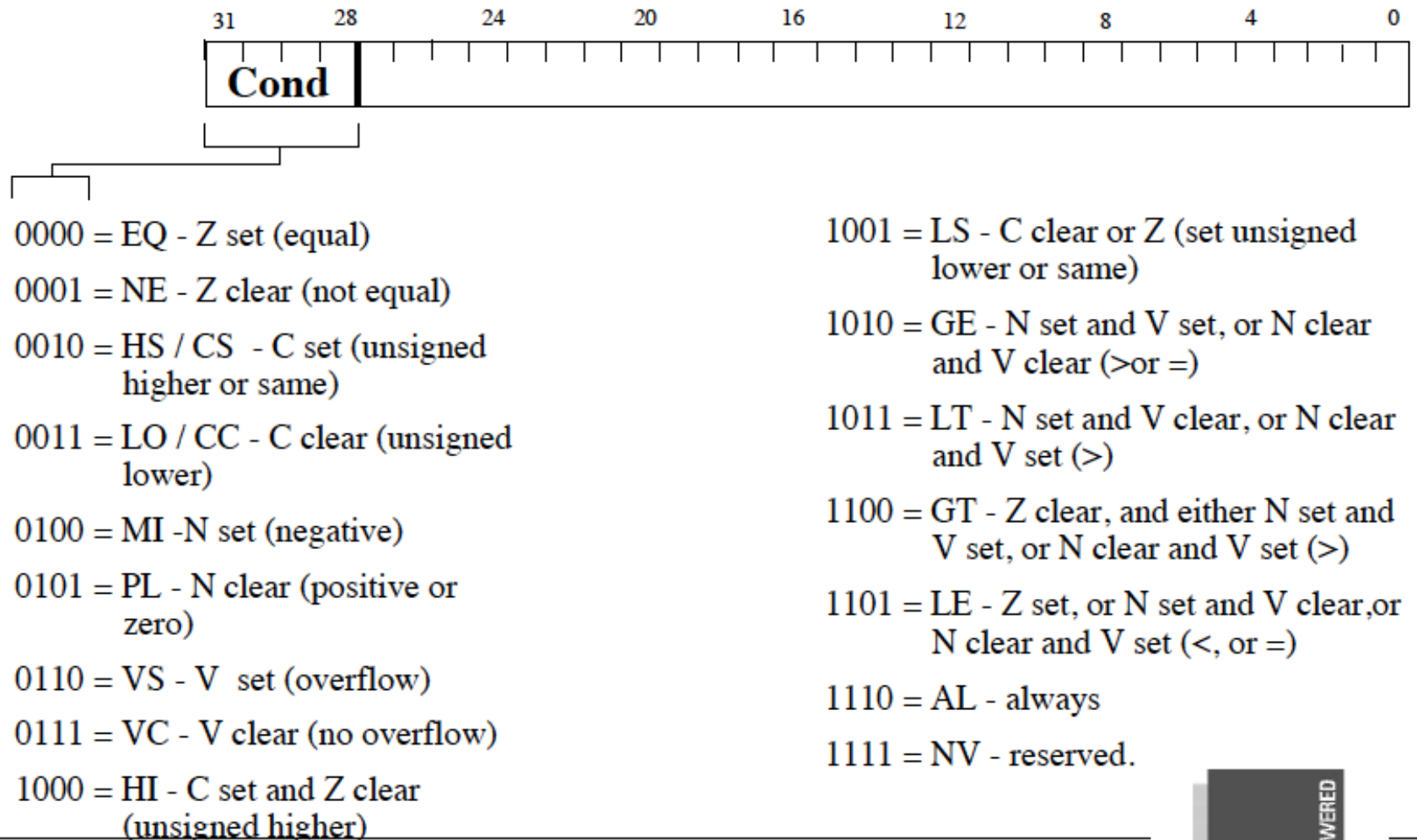
Conditional Execution in the ARM ISA

- Almost all ARM instructions can include an optional condition code.
- An instruction with a condition code is executed only if the condition code flags in the CPSR meet the specified condition.

Conditional Execution in ARM ISA

31	2827				1615				87				0				<u>Instruction type</u>																							
Cond	0	0	I	Opcode				S	Rn				Rd				Operand2				Data processing / PSR Transfer																			
Cond	0	0	0	0	0	0	0	A	S	Rd				Rn				Rs				1	0	0	1	Rm				Multiply										
Cond	0	0	0	0	0	1	U	A	S	RdHi				RdLo				Rs				1	0	0	1	Rm				Long Multiply (v3M / v4 only)										
Cond	0	0	0	1	0	B	0	0	Rn				Rd				0				0	0	0	1	0	0	1	Rm				Swap								
Cond	0	1	I	P	U	B	W	L	Rn				Rd				Offset								Load/Store Byte/Word															
Cond	1	0	0	P	U	S	W	L	Rn				Register List												Load/Store Multiple															
Cond	0	0	0	P	U	1	W	L	Rn				Rd				Offset1				1	S	H	1	Offset2				Halfword transfer : Immediate offset (v4 only)											
Cond	0	0	0	P	U	0	W	L	Rn				Rd				0				0	0	0	1	S	H	1	Rm				Halfword transfer: Register offset (v4 only)								
Cond	1	0	1	L	Offset																							Branch												
Cond	0	0	0	1	0				0	1	0	1				1	1	1	1				1	1	1	0				0	0	1	Rn				Branch Exchange (v4T only)			
Cond	1	1	0	P	U	N	W	L	Rn				CRd				CPNum				Offset								Coprocessor data transfer											
Cond	1	1	1	0	Op1				CRn				CRd				CPNum				Op2				0	CRm				Coprocessor data operation										
Cond	1	1	1	0	Op1				L	CRn				Rd				CPNum				Op2				1	CRm				Coprocessor register transfer									
Cond	1	1	1	1	SWI Number																							Software interrupt												

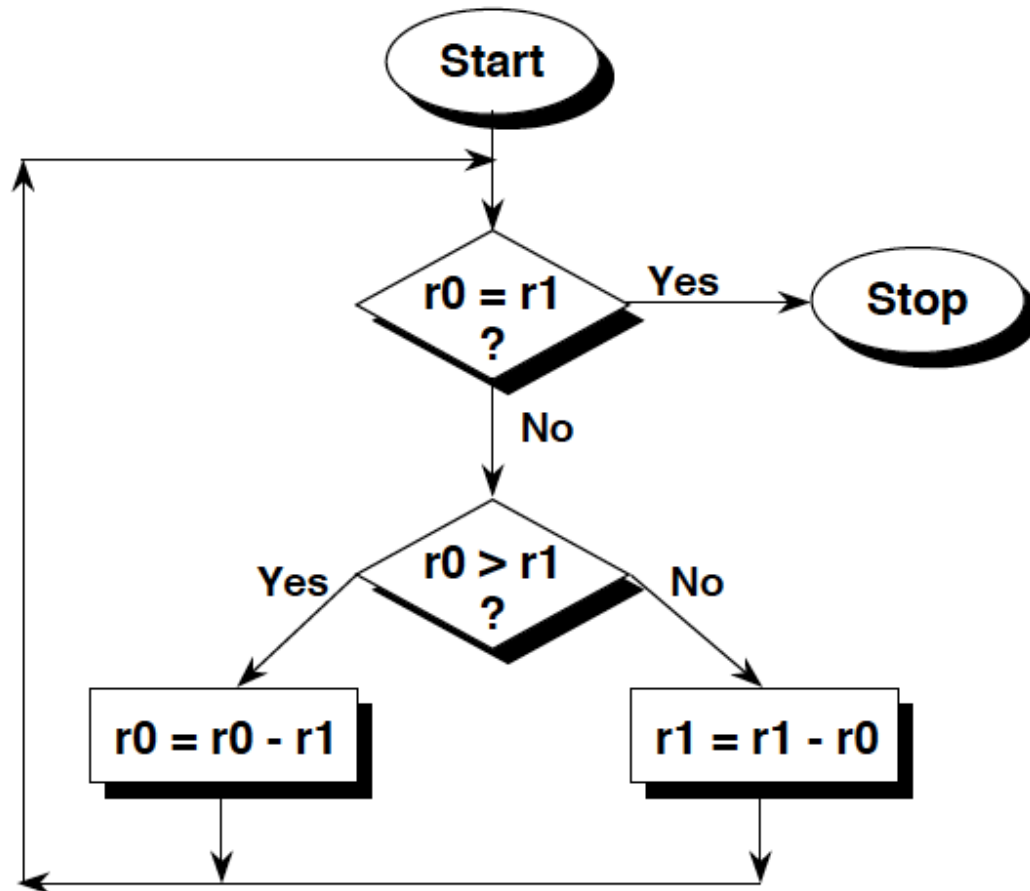
Conditional Execution in ARM ISA



Conditional Execution in ARM ISA

- * **To execute an instruction conditionally, simply postfix it with the appropriate condition:**
 - For example an add instruction takes the form:
 - `ADD r0,r1,r2 ; r0 = r1 + r2 (ADDAL)`
 - To execute this only if the zero flag is set:
 - `ADDEQ r0,r1,r2 ; If zero flag set then...`
`; ... r0 = r1 + r2`
- * **By default, data processing operations do not affect the condition flags (apart from the comparisons where this is the only effect). To cause the condition flags to be updated, the S bit of the instruction needs to be set by postfixing the instruction (and any condition code) with an “S”.**
 - For example to add two numbers and set the condition flags:
 - `ADDS r0,r1,r2 ; r0 = r1 + r2`
`; ... and set flags`

Conditional Execution in ARM ISA



* **Convert the GCD algorithm given in this flowchart into**

- 1) “Normal” assembler, where only branches can be conditional.
- 2) ARM assembler, where all instructions are conditional, thus improving code density.

* **The only instructions you need are **CMP**, **B** and **SUB**.**

Conditional Execution in ARM ISA

“Normal” Assembler

```
gcd    cmp r0, r1      ;reached the end?
        beq stop
        blt less        ;if r0 > r1
        sub r0, r0, r1  ;subtract r1 from r0
        bal gcd
less   sub r1, r1, r0   ;subtract r0 from r1
        bal gcd
stop
```

ARM Conditional Assembler

```
gcd    cmp    r0, r1      ;if r0 > r1
        subgt r0, r0, r1  ;subtract r1 from r0
        sublt r1, r1, r0  ;else subtract r0 from r1
        bne   gcd        ;reached the end?
```

Idealism

- Wouldn't it be nice
 - If the branch is eliminated (predicated) when it would actually be mispredicted
 - If the branch were predicted when it would actually be correctly predicted

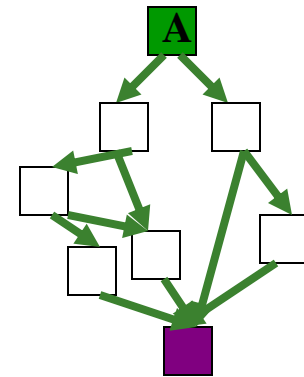
- Wouldn't it be nice
 - If predication did not require ISA support

Improving Predicated Execution

- Three major limitations of predication
 1. **Adaptivity**: non-adaptive to branch behavior
 2. **Complex CFG**: inapplicable to loops/complex control flow graphs
 3. **ISA**: Requires large ISA changes

- **Wish Branches** [Kim+, MICRO 2005]
 - Solve 1 and partially 2 (for loops)

- **Dynamic Predicated Execution**
 - Diverge-Merge Processor [Kim+, MICRO 2006]
 - Solves 1, 2 (partially), 3

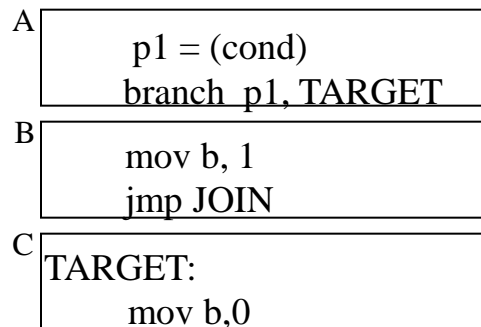
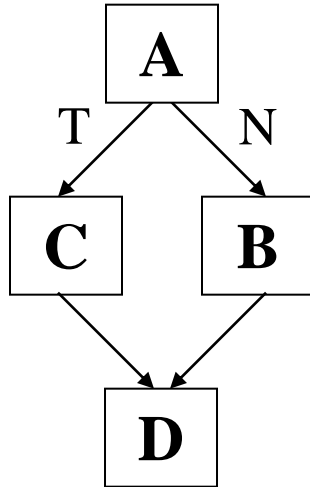


Wish Branches

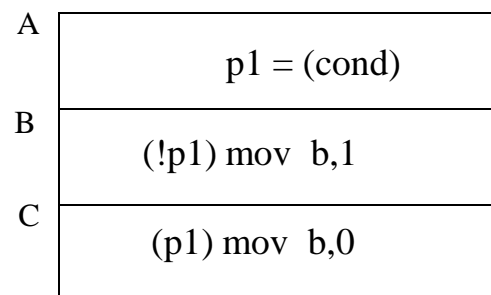
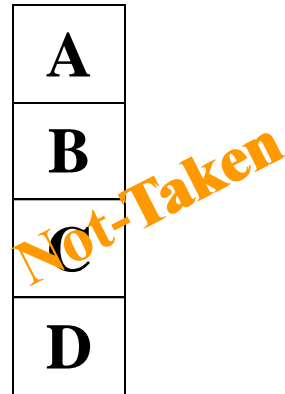
- The **compiler** generates code (with wish branches) that can be executed **either** as predicated code **or** non-predicated code (normal branch code)
- The **hardware decides** to execute predicated code or normal branch code at run-time based on the confidence of branch prediction
- **Easy to predict: normal branch code**
- **Hard to predict: predicated code**
- Kim et al., “**Wish Branches: Enabling Adaptive and Aggressive Predicated Execution**,” MICRO 2006, IEEE Micro Top Picks, Jan/Feb 2006.

Wish Jump/Join

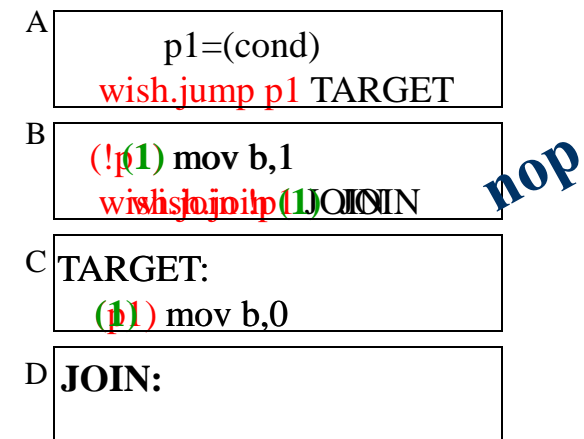
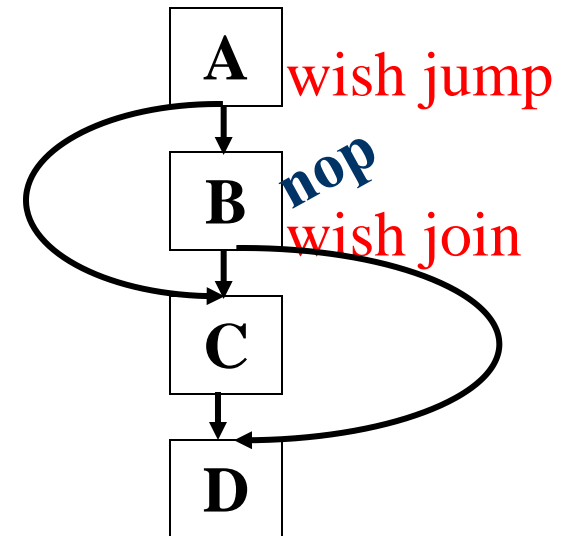
High Confidence



normal branch code



predicated code



wish jump/join code

Wish Branches vs. Predicated Execution

- Advantages compared to predicated execution
 - **Reduces the overhead** of predication
 - Increases the benefits of predicated code by allowing the compiler to generate more **aggressively-predicated code**
 - Makes predicated code less dependent on machine configuration (e.g. branch predictor)
- Disadvantages compared to predicated execution
 - Extra branch instructions use machine resources
 - Extra branch instructions increase the contention for branch predictor table entries
 - **Constrains the compiler's scope for code optimizations**

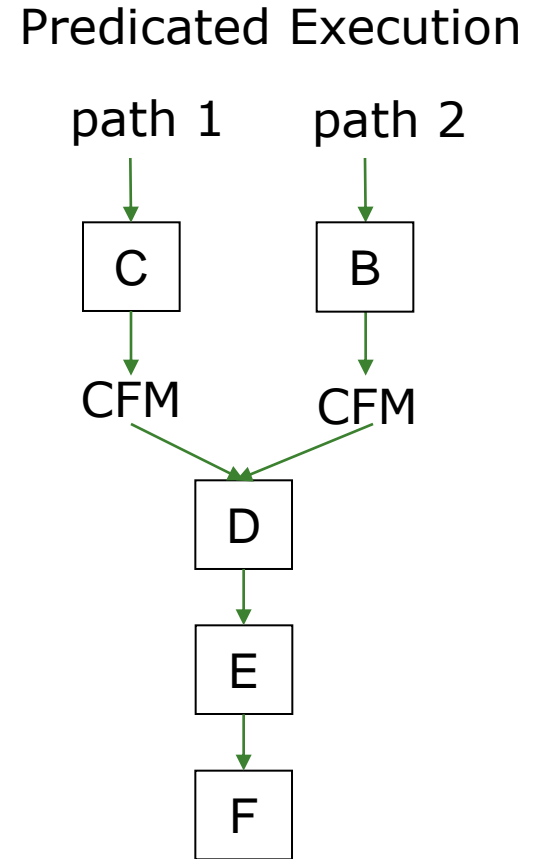
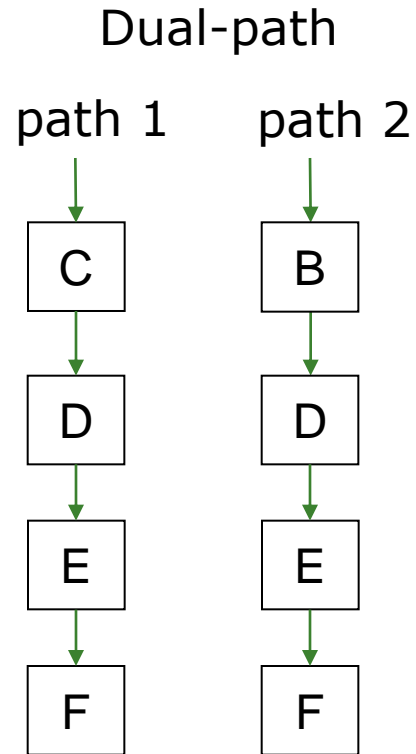
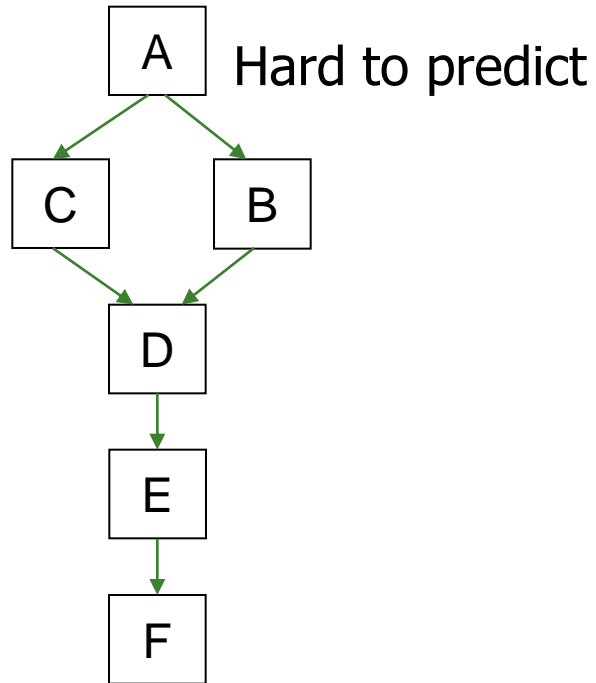
How to Handle Control Dependences

- Critical to keep the pipeline full with correct sequence of dynamic instructions.
- Potential solutions if the instruction is a control-flow instruction:
 - Stall the pipeline until we know the next fetch address
 - Guess the next fetch address (branch prediction)
 - Employ delayed branching (branch delay slot)
 - Do something else (fine-grained multithreading)
 - Eliminate control-flow instructions (predicated execution)
 - Fetch from both possible paths (if you know the addresses of both possible paths) (multipath execution)

Multi-Path Execution

- Idea: Execute both paths after a conditional branch
 - For all branches: Riseman and Foster, “The inhibition of potential parallelism by conditional jumps,” IEEE Transactions on Computers, 1972.
 - For a hard-to-predict branch: Use dynamic confidence estimation
- Advantages:
 - + Improves performance if misprediction cost > useless work
 - + No ISA change needed
- Disadvantages:
 - What happens when the machine encounters another hard-to-predict branch? Execute both paths again?
 - Paths followed quickly become exponential
 - Each followed path requires its own context (registers, PC, GHR)
 - Wasted work (and reduced performance) if paths merge

Dual-Path Execution versus Predication



Remember: Branch Types

Type	Direction at fetch time	Number of possible next fetch addresses?	When is next fetch address resolved?
Conditional	Unknown	2	Execution (register dependent)
Unconditional	Always taken	1	Decode (PC + offset)
Call	Always taken	1	Decode (PC + offset)
Return	Always taken	Many	Execution (register dependent)
Indirect	Always taken	Many	Execution (register dependent)

Different branch types can be handled differently

Call and Return Prediction

■ Direct calls are easy to predict

- Always taken, single target
- Call marked in BTB, target predicted by BTB

■ Returns are indirect branches

- A function can be called from many points in code
- A return instruction can have many target addresses
 - Next instruction after each call point for the same function
- Observation: Usually a return matches a call
- Idea: Use a stack to predict return addresses (Return Address Stack)
 - A fetched call: pushes the return (next instruction) address on the stack
 - A fetched return: pops the stack and uses the address as its predicted target
 - Accurate most of the time: 8-entry stack → > 95% accuracy

Call X

...

Call X

...

Call X

...

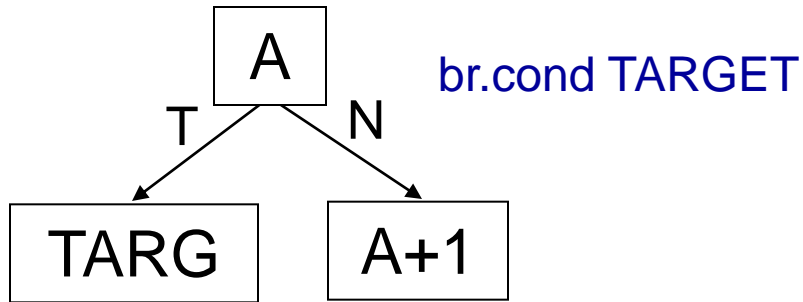
Return

Return

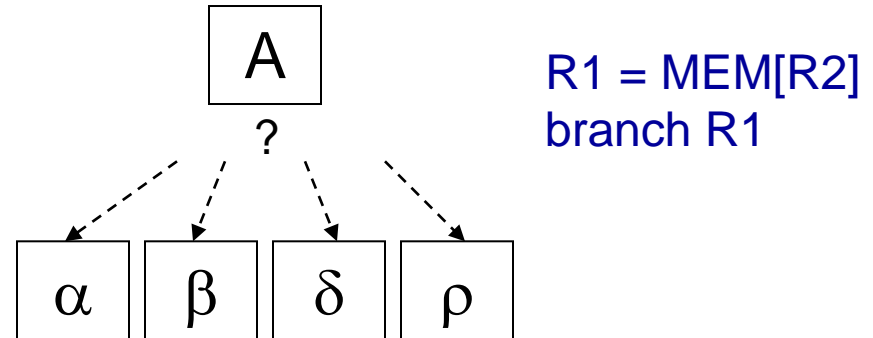
Return

Indirect Branch Prediction (I)

- Register-indirect branches have multiple targets



Conditional (Direct) Branch



Indirect Jump

- Used to implement
 - ❑ Switch-case statements
 - ❑ Virtual function calls
 - ❑ Jump tables (of function pointers)
 - ❑ Interface calls

Indirect Branch Prediction (II)

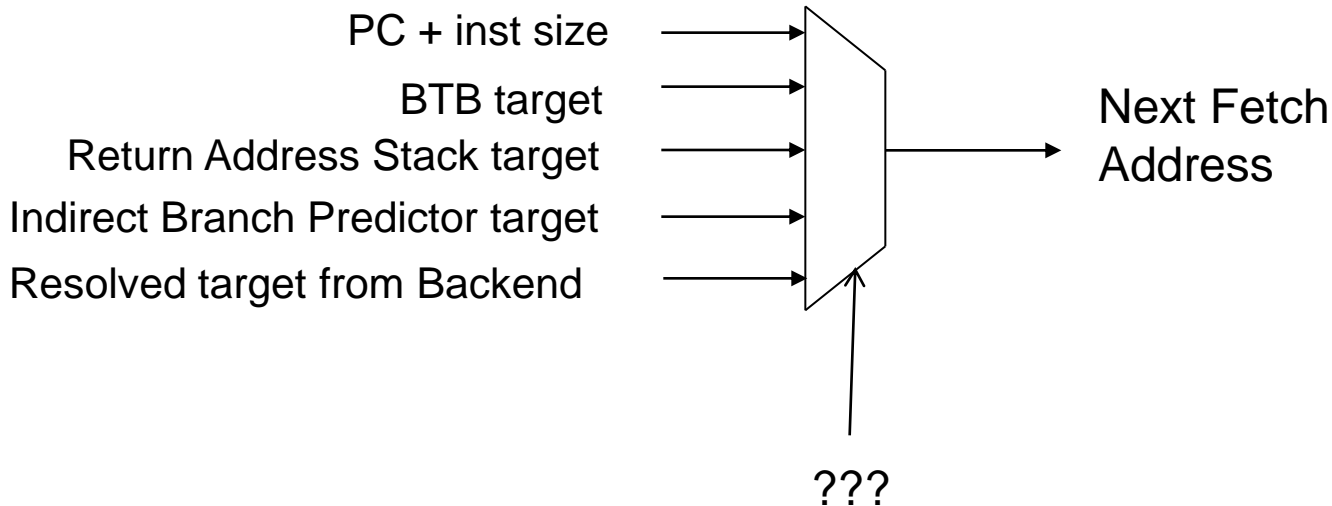
- No direction prediction needed
- Idea 1: Predict the last resolved target as the next fetch address
 - + Simple: Use the BTB to store the target address
 - Inaccurate: 50% accuracy (empirical). Many indirect branches switch between different targets
- Idea 2: Use history based target prediction
 - E.g., Index the BTB with GHR XORed with Indirect Branch PC
 - Chang et al., “Target Prediction for Indirect Jumps,” ISCA 1997.
 - + More accurate
 - An indirect branch maps to (too) many entries in BTB
 - Conflict misses with other branches (direct or indirect)
 - Inefficient use of space if branch has few target addresses

Issues in Branch Prediction (I)

- Need to identify a branch before it is fetched
- How do we do this?
 - BTB hit → indicates that the fetched instruction is a branch
 - BTB entry contains the “type” of the branch
- What if no BTB?
 - Bubble in the pipeline until target address is computed
 - E.g., IBM POWER4

Issues in Branch Prediction (II)

- **Latency:** Prediction is latency critical
 - ❑ Need to generate next fetch address for the next cycle
 - ❑ Bigger, more complex predictors are more accurate but slower

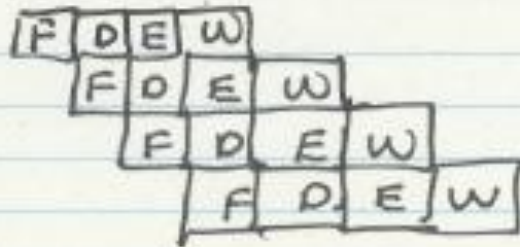


Complications in Superscalar Processors

- “Superscalar” processors
 - ❑ attempt to execute more than 1 instruction-per-cycle
 - ❑ must fetch multiple instructions per cycle

- Consider a 2-way superscalar fetch scenario
 - (case 1) Both insts are not taken control flow inst
 - $nPC = PC + 8$
 - (case 2) One of the insts is a taken control flow inst
 - $nPC = \text{predicted target addr}$
 - *NOTE* both instructions could be control-flow; prediction based on the first one predicted taken
 - If the 1st instruction is the predicted taken branch
 - nullify 2nd instruction fetched

Multiple Instruction Fetch: Concepts



← Fetch 1 inst/cycle

- Downside:

Flynn's bottleneck

If you fetch 1 inst/cycle

you cannot finish > 1 inst/cycle



← Fetch 4 inst/cycle

Two major approaches

1) VLIW

Compiler decides what insts.
can be executed in parallel
→ Simple hardware

2) Superscalar

Hardware detects dependencies
between instructions that
are fetched in the same
cycle.

Review of Last Few Lectures

- Control dependence handling in pipelined machines
 - ❑ Delayed branching
 - ❑ Fine-grained multithreading
 - ❑ Branch prediction
 - Compile time (static)
 - ❑ Always NT, Always T, Backward T Forward NT, Profile based
 - Run time (dynamic)
 - ❑ Last time predictor
 - ❑ Hysteresis: 2BC predictor
 - ❑ Global branch correlation → Two-level global predictor
 - ❑ Local branch correlation → Two-level local predictor
 - ❑ Predicated execution
 - ❑ Multipath execution