

## 1 Caches

## 2 Cache Analysis for Least-Recently-Used

## 3 Beyond Least-Recently-Used

- Predictability Metrics
- Relative Competitiveness
- Sensitivity – Caches and Measurement-Based Timing Analysis

## 4 Summary

## 1 Caches

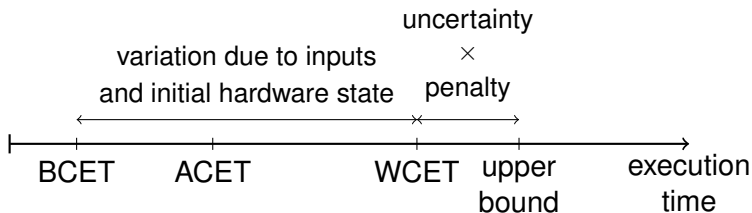
## 2 Cache Analysis for Least-Recently-Used

## 3 Beyond Least-Recently-Used

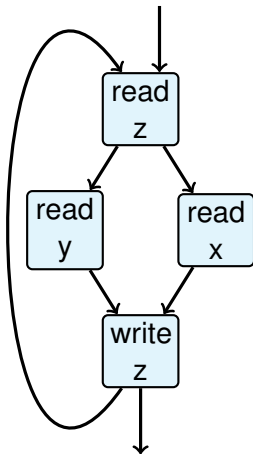
- Predictability Metrics
- Relative Competitiveness
- Sensitivity – Caches and Measurement-Based Timing Analysis

## 4 Summary

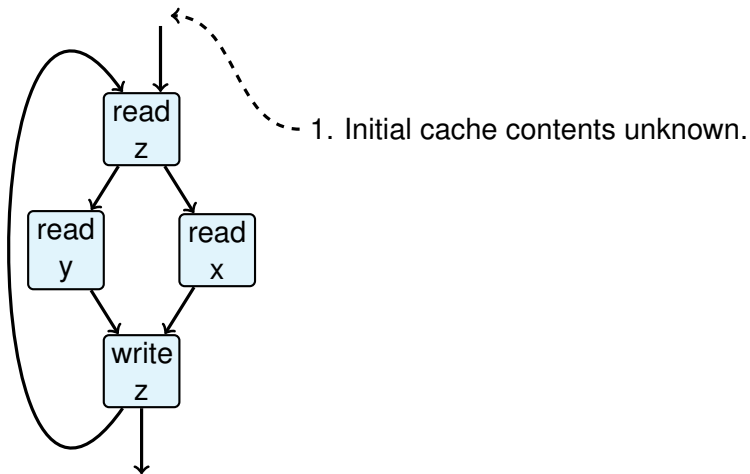
- Amount of uncertainty determines precision of WCET analysis
- Uncertainty in cache analysis depends on replacement policy



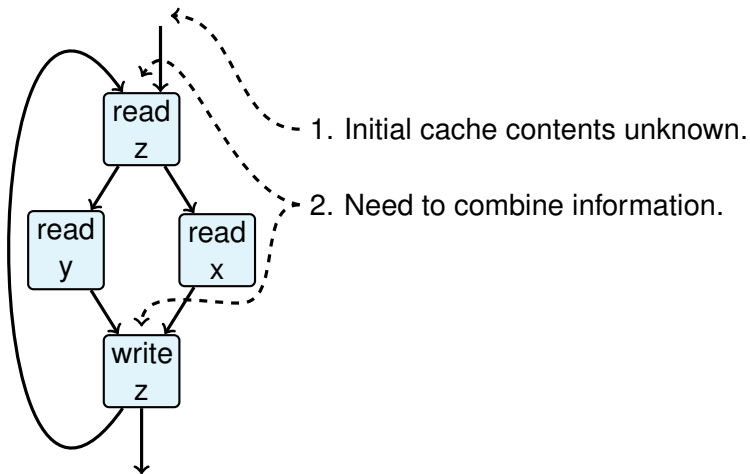
# Uncertainty in Cache Analysis



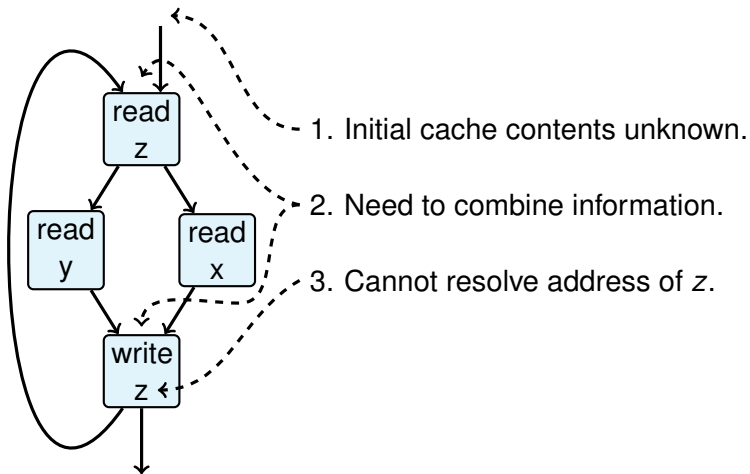
# Uncertainty in Cache Analysis

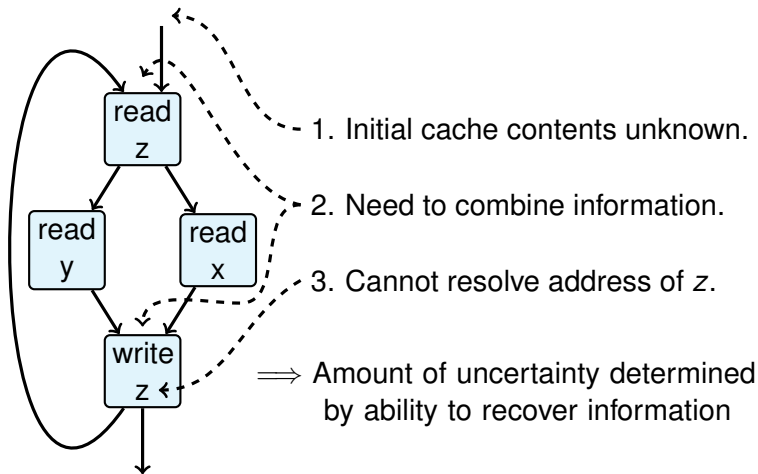


# Uncertainty in Cache Analysis



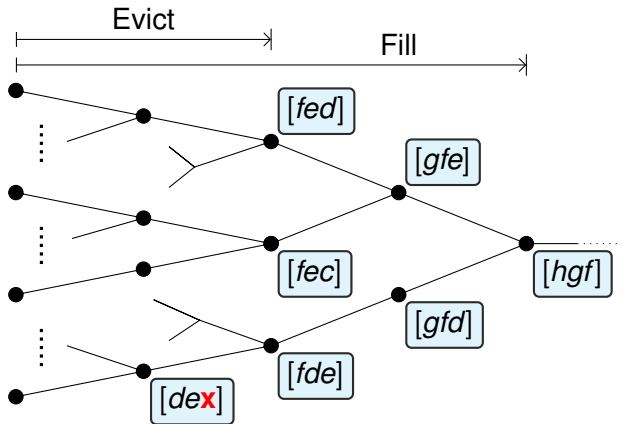
# Uncertainty in Cache Analysis







# Predictability Metrics



Sequence:  $\langle a, \dots, e, f, g, h \rangle$

## ■ Evict

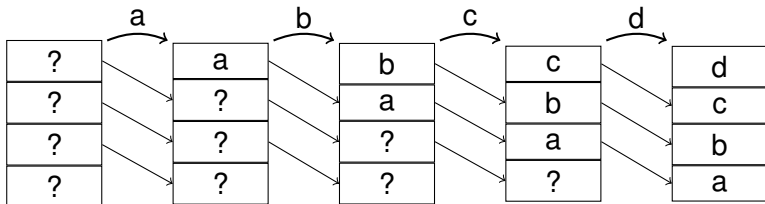
- ▶ Number of accesses to obtain *any may*-information.
- ▶ I.e. when can an analysis predict any cache misses?

## ■ Fill

- ▶ Number of accesses to complete *may*- and *must*-information.
- ▶ I.e. when can an analysis predict each access?

→ Evict and Fill bound the precision of *any* static cache analysis.  
Can thus serve as a benchmark for analyses.

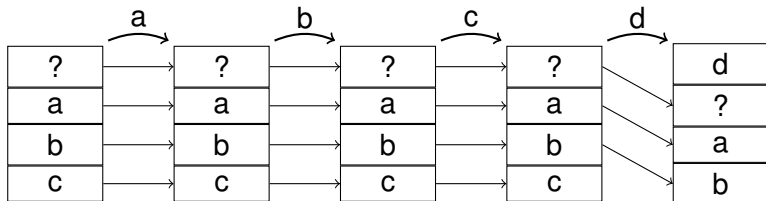
- LRU “forgets” about past quickly:
  - ▶ cares about most-recent access to each block only
  - ▶ order of previous accesses irrelevant



- In the example:  $\text{Evict} = \text{Fill} = 4$
- In general:  $\text{Evict}(k) = \text{Fill}(k) = k$ , where  $k$  is the associativity of the cache

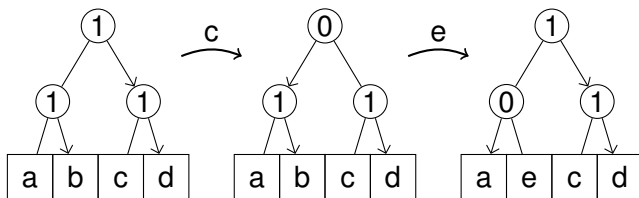
# Evaluation of First-In First-Out (sketch)

- Like LRU in the miss-case
- But: “Ignores” hits



- In the worst-case  $k - 1$  hits and  $k$  misses: ( $k = \text{associativity}$ )  
→  $\text{Evict}(k) = 2k - 1$
- Another  $k$  accesses to obtain complete knowledge:  
→  $\text{Fill}(k) = 3k - 1$

- Tree-bits point to block to be replaced



- Accesses “rejuvenate” neighborhood
  - ▶ Active blocks keep their (inactive) neighborhood in the cache
- Analysis yields:
  - ▶  $\text{Evict}(k) = \frac{k}{2} \log_2 k + 1$
  - ▶  $\text{Fill}(k) = \frac{k}{2} \log_2 k + k - 1$

Policy	Evict( $k$ )	Fill( $k$ )	Evict(8)	Fill(8)
LRU	$k$	$k$	8	8
FIFO	$2k - 1$	$3k - 1$	15	23
MRU	$2k - 2$	$\infty/3k - 4$	14	$\infty/20$
PLRU	$\frac{k}{2} \log_2 k + 1$	$\frac{k}{2} \log_2 k + k - 1$	13	19

- LRU is optimal w.r.t. metrics.
- Other policies are much less predictable.

→ Use LRU if predictability is a concern.

- How to obtain *may*- and *must*-information within the given limits for other policies?

## 1 Caches

## 2 Cache Analysis for Least-Recently-Used

## 3 Beyond Least-Recently-Used

- Predictability Metrics
- **Relative Competitiveness**
- Sensitivity – Caches and Measurement-Based Timing Analysis

## 4 Summary

- **Competitiveness** (Sleator and Tarjan, 1985):  
worst-case performance of an online policy *relative to the optimal offline policy*
  - ▶ used to evaluate online policies
- **Relative competitiveness** (Reineke and Grund, 2008):  
worst-case performance of an online policy *relative to another online policy*
  - ▶ used to derive local and global cache analyses



## Notation

$m_{\mathbf{P}}(p, s)$  = *number of misses that policy  $\mathbf{P}$  incurs on access sequence  $s \in M^*$  starting in state  $p \in C^{\mathbf{P}}$*

## Notation

$m_{\mathbf{P}}(p, s)$  = *number of misses that policy  $\mathbf{P}$  incurs on access sequence  $s \in M^*$  starting in state  $p \in C^{\mathbf{P}}$*

## Definition (Relative miss competitiveness)

Policy  $\mathbf{P}$  is  $(k, c)$ -miss-competitive relative to policy  $\mathbf{Q}$  if

$$m_{\mathbf{P}}(p, s) \leq k \cdot m_{\mathbf{Q}}(q, s) + c$$

for all access sequences  $s \in M^*$  and cache-set states  $p \in C^{\mathbf{P}}, q \in C^{\mathbf{Q}}$  that are compatible  $p \sim q$ .

# Definition – Relative Miss-Competitiveness

## Notation

$m_{\mathbf{P}}(p, s)$  = *number of misses that policy  $\mathbf{P}$  incurs on access sequence  $s \in M^*$  starting in state  $p \in C^{\mathbf{P}}$*

## Definition (Relative miss competitiveness)

Policy  $\mathbf{P}$  is  $(k, c)$ -miss-competitive relative to policy  $\mathbf{Q}$  if

$$m_{\mathbf{P}}(p, s) \leq k \cdot m_{\mathbf{Q}}(q, s) + c$$

for all access sequences  $s \in M^*$  and cache-set states  $p \in C^{\mathbf{P}}, q \in C^{\mathbf{Q}}$  that are compatible  $p \sim q$ .

## Definition (Competitive miss ratio of $\mathbf{P}$ relative to $\mathbf{Q}$ )

The smallest  $k$ , s.t.  $\mathbf{P}$  is  $(k, c)$ -miss-competitive rel. to  $\mathbf{Q}$  for some  $c$ .

# Example – Relative Miss-Competitiveness

**P** is  $(3, 4)$ -miss-competitive relative to **Q**.

If **Q** incurs  $x$  misses, then **P** incurs at most  $3 \cdot x + 4$  misses.

# Example – Relative Miss-Competitiveness

**P** is  $(3, 4)$ -miss-competitive relative to **Q**.

If **Q** incurs  $x$  misses, then **P** incurs at most  $3 \cdot x + 4$  misses.

**Best:** **P** is  $(1, 0)$ -miss-competitive relative to **Q**.

# Example – Relative Miss-Competitiveness

**P** is  $(3, 4)$ -miss-competitive relative to **Q**.

If **Q** incurs  $x$  misses, then **P** incurs at most  $3 \cdot x + 4$  misses.

**Best:** **P** is  $(1, 0)$ -miss-competitive relative to **Q**.

**Worst:** **P** is not-miss-competitive (or  $\infty$ -miss-competitive) relative to **Q**.

# Example – Relative Hit-Competitiveness

**P** is  $(\frac{2}{3}, 3)$ -hit-competitive relative to **Q**.

If **Q** has  $x$  hits, then **P** has at least  $\frac{2}{3} \cdot x - 3$  hits.

# Example – Relative Hit-Competitiveness

**P** is  $(\frac{2}{3}, 3)$ -hit-competitive relative to **Q**.

If **Q** has  $x$  hits, then **P** has at least  $\frac{2}{3} \cdot x - 3$  hits.

**Best:** **P** is  $(1, 0)$ -hit-competitive relative to **Q**.  
Equivalent to  $(1, 0)$ -miss-competitiveness.



# Example – Relative Hit-Competitiveness

**P** is  $(\frac{2}{3}, 3)$ -hit-competitive relative to **Q**.

If **Q** has  $x$  hits, then **P** has at least  $\frac{2}{3} \cdot x - 3$  hits.

**Best:** **P** is  $(1, 0)$ -hit-competitive relative to **Q**.  
Equivalent to  $(1, 0)$ -miss-competitiveness.

**Worst:** **P** is  $(0, 0)$ -hit-competitive relative to **Q**.  
Analogue to  $\infty$ -miss-competitiveness.

Let  $\mathbf{P}$  be  $(1, 0)$ -competitive relative to  $\mathbf{Q}$ :

$$m_{\mathbf{P}}(p, s) \leq 1 \cdot m_{\mathbf{Q}}(q, s) + 0$$

$$\Leftrightarrow m_{\mathbf{P}}(p, s) \leq m_{\mathbf{Q}}(q, s)$$

Let **P** be  $(1, 0)$ -competitive relative to **Q**:

$$m_{\mathbf{P}}(p, s) \leq 1 \cdot m_{\mathbf{Q}}(q, s) + 0$$

$$\Leftrightarrow m_{\mathbf{P}}(p, s) \leq m_{\mathbf{Q}}(q, s)$$

- 1 If **Q** “hits”, so does **P**, and
- 2 if **P** “misses”, so does **Q**.

Let  $\mathbf{P}$  be  $(1, 0)$ -competitive relative to  $\mathbf{Q}$ :

$$m_{\mathbf{P}}(p, s) \leq 1 \cdot m_{\mathbf{Q}}(q, s) + 0$$

$$\Leftrightarrow m_{\mathbf{P}}(p, s) \leq m_{\mathbf{Q}}(q, s)$$

- 1 If  $\mathbf{Q}$  “hits”, so does  $\mathbf{P}$ , and
- 2 if  $\mathbf{P}$  “misses”, so does  $\mathbf{Q}$ .

As a consequence,

- 1 a *must*-analysis for  $\mathbf{Q}$  is also a *must*-analysis for  $\mathbf{P}$ , and
- 2 a *may*-analysis for  $\mathbf{P}$  is also a *may*-analysis for  $\mathbf{Q}$ .

# Global Guarantees: $(k, c)$ -Competitiveness

**Given:** Global guarantees for policy **Q**.

**Wanted:** Global guarantees for policy **P**.

# Global Guarantees: $(k, c)$ -Competitiveness

**Given:** Global guarantees for policy **Q**.

**Wanted:** Global guarantees for policy **P**.

1 Determine competitiveness of policy **P** relative to policy **Q**.

$$m_P \leq k \cdot m_Q + c$$

# Global Guarantees: $(k, c)$ -Competitiveness

**Given:** Global guarantees for policy **Q**.  
**Wanted:** Global guarantees for policy **P**.

- 1 Determine competitiveness of policy **P** relative to policy **Q**.

$$m_P \leq k \cdot m_Q + c$$

- 2 Compute global guarantee for task  $T$  under policy **Q**.

$$m_Q(T)$$

# Global Guarantees: $(k, c)$ -Competitiveness

**Given:** Global guarantees for policy **Q**.  
**Wanted:** Global guarantees for policy **P**.

- 1 Determine competitiveness of policy **P** relative to policy **Q**.

$$m_P \leq k \cdot m_Q + c$$

- 2 Compute global guarantee for task  $T$  under policy **Q**.

$$m_Q(T)$$

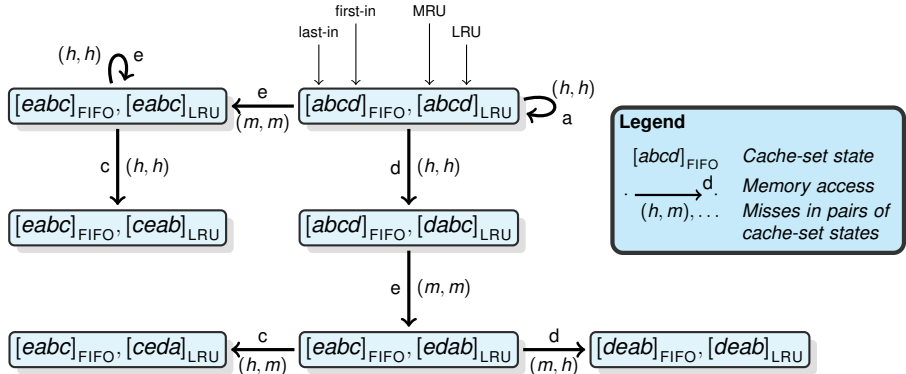
- 3 Calculate global guarantee on the number of misses for **P** using the global guarantee for **Q** and the competitiveness results of **P** relative to **Q**.

$$m_P \leq k \cdot m_Q + c \quad \triangleright \quad m_Q(T) = m_P(T)$$



# Relative Competitiveness: Automatic Computation

**P** and **Q** (here: FIFO and LRU) induce transition system:

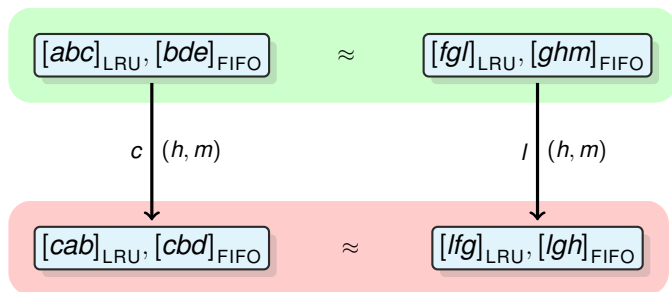


**Competitive miss ratio** = maximum ratio of misses in policy **P** to misses in policy **Q** in transition system

# Transition System is $\infty$ Large

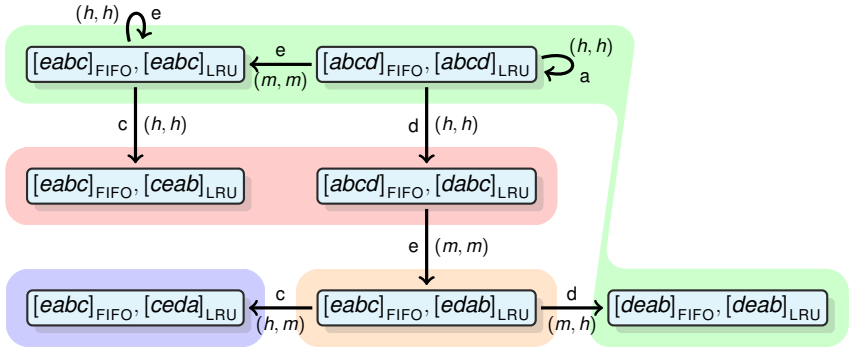
**Problem:** The induced transition system is  $\infty$  large.

**Observation:** Only the *relative positions* of elements matter:

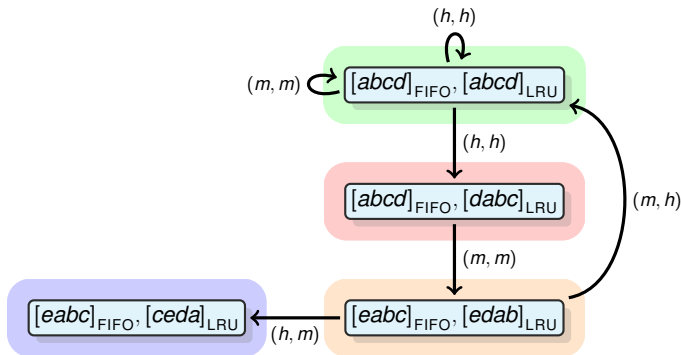


**Solution:** Construct *finite* quotient transition system.

# $\approx$ -Equivalent States in Running Example



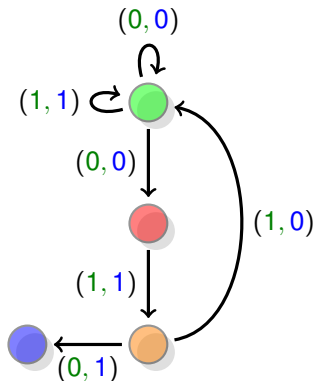
Merging  $\approx$ -equivalent states yields a finite quotient transition system:



# Competitive Ratio = Maximum Cycle Ratio

Competitive miss ratio =

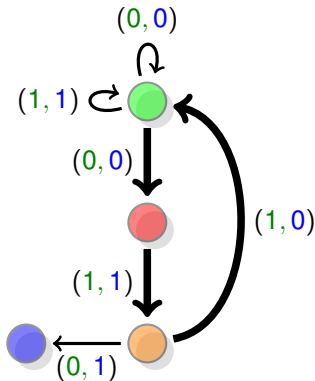
maximum ratio of misses in policy **P** to misses in policy **Q**



# Competitive Ratio = Maximum Cycle Ratio

Competitive miss ratio =

maximum ratio of misses in policy **P** to misses in policy **Q**



$$\text{Maximum cycle ratio} = \frac{0+1+1}{0+1+0} = 2$$

- Implemented in Java, called Relacs
- Interface for replacement policies
- Fully automatic
- Provides example sequences for competitive ratio and constant
- Analysis usually practically feasible up to associativity 8
  - ▶ limited by memory consumption
  - ▶ depends on similarity of replacement policies

Online version:

<http://rw4.cs.uni-sb.de/~reineke/relacs>

# Generalizations

Identified patterns and proved generalizations by hand.  
Aided by example sequences generated by tool.



# Generalizations

Identified patterns and proved generalizations by hand.  
 Aided by example sequences generated by tool.

Previously unknown facts:

$\text{PLRU}(k)$  is  $(1, 0)$  comp. rel. to  $\text{LRU}(1 + \log_2 k)$ ,  
 $\longrightarrow$  LRU-*must*-analysis can be used for PLRU

# Generalizations

Identified patterns and proved generalizations by hand.  
 Aided by example sequences generated by tool.

Previously unknown facts:

$\text{PLRU}(k)$  is  $(1, 0)$  comp. rel. to  $\text{LRU}(1 + \log_2 k)$ ,  
 $\longrightarrow$  LRU-*must*-analysis can be used for PLRU

$\text{FIFO}(k)$  is  $(\frac{1}{2}, \frac{k-1}{2})$  hit-comp. rel. to  $\text{LRU}(k)$ , whereas  
 $\text{LRU}(k)$  is  $(0, 0)$  hit-comp. rel. to  $\text{FIFO}(k)$ , but

# Generalizations

Identified patterns and proved generalizations by hand.  
 Aided by example sequences generated by tool.

Previously unknown facts:

$\text{PLRU}(k)$  is  $(1, 0)$  comp. rel. to  $\text{LRU}(1 + \log_2 k)$ ,  
 $\longrightarrow$  LRU-*must*-analysis can be used for PLRU

$\text{FIFO}(k)$  is  $(\frac{1}{2}, \frac{k-1}{2})$  hit-comp. rel. to  $\text{LRU}(k)$ , whereas  
 $\text{LRU}(k)$  is  $(0, 0)$  hit-comp. rel. to  $\text{FIFO}(k)$ , but

$\text{LRU}(2k - 1)$  is  $(1, 0)$  comp. rel. to  $\text{FIFO}(k)$ , and

$\text{LRU}(2k - 2)$  is  $(1, 0)$  comp. rel. to  $\text{MRU}(k)$ .

$\longrightarrow$  LRU-*may*-analysis can be used for FIFO and MRU

$\longrightarrow$  optimal with respect to predictability metric Evict

# Generalizations

Identified patterns and proved generalizations by hand.  
Aided by example sequences generated by tool.

Previously unknown facts:

PLRU( $k$ ) is  $(1, 0)$  comp. rel. to LRU( $1 + \log_2 k$ ),  
→ LRU-*must*-analysis can be used for PLRU

FIFO( $k$ ) is  $(\frac{1}{2}, \frac{k-1}{2})$  hit-comp. rel. to LRU( $k$ ), whereas  
LRU( $k$ ) is  $(0, 0)$  hit-comp. rel. to FIFO( $k$ ), but

LRU( $2k - 1$ ) is  $(1, 0)$  comp. rel. to FIFO( $k$ ), and  
LRU( $2k - 2$ ) is  $(1, 0)$  comp. rel. to MRU( $k$ ).

→ LRU-*may*-analysis can be used for FIFO and MRU  
→ optimal with respect to predictability metric Evict

FIFO-*may*-analysis used in the analysis of the branch target buffer of the MOTOROLA POWERPC 56X.

## 1 Caches

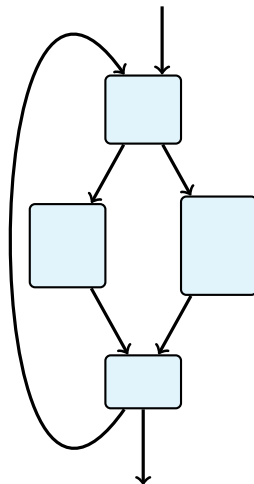
## 2 Cache Analysis for Least-Recently-Used

## 3 Beyond Least-Recently-Used

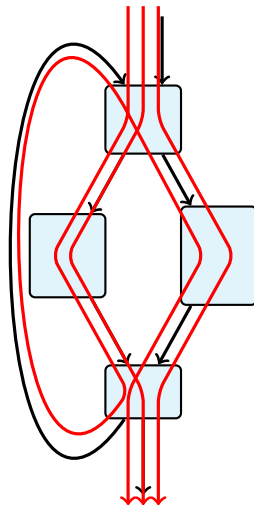
- Predictability Metrics
- Relative Competitiveness
- Sensitivity – Caches and Measurement-Based Timing Analysis

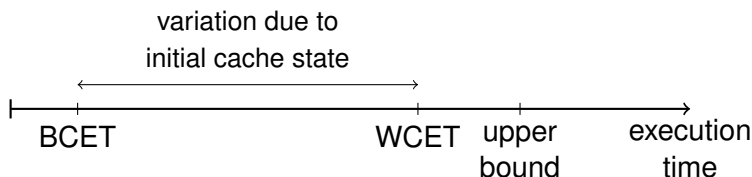
## 4 Summary

- Run program on a number of inputs and initial states.
- Combine measurements for basic blocks to obtain WCET estimation.
- Sensitivity Analysis demonstrates this approach may be dramatically wrong.



- Run program on a number of inputs and initial states.
- Combine measurements for basic blocks to obtain WCET estimation.
- Sensitivity Analysis demonstrates this approach may be dramatically wrong.





## Definition (Miss sensitivity)

Policy  $\mathbf{P}$  is  $(k, c)$ -miss-sensitive if

$$m_{\mathbf{P}}(p, s) \leq k \cdot m_{\mathbf{P}}(p', s) + c$$

for all access sequences  $s \in M^*$  and cache-set states  $p, p' \in C^{\mathbf{P}}$ .



Policy	2	3	4	5	6	7	8
LRU	1,2	1,3	1,4	1,5	1,6	1,7	1,8
FIFO	2,2	3,3	4,4	5,5	6,6	7,7	8,8
PLRU	1,2	—	$\infty$	—	—	—	$\infty$
MRU	1,2	3,4	5,6	7,8	MEM	MEM	MEM

- LRU is optimal. Performance varies in the least possible way.
- For FIFO, PLRU, and MRU the number of misses may vary strongly.
- Case study based on simple model of execution time by Hennessy and Patterson (2003):  
WCET may be 3 times higher than a measured execution time for 4-way FIFO.

## 1 Caches

## 2 Cache Analysis for Least-Recently-Used

## 3 Beyond Least-Recently-Used

- Predictability Metrics
- Relative Competitiveness
- Sensitivity – Caches and Measurement-Based Timing Analysis

## 4 Summary

## Cache Analysis for Least-Recently-Used

- ... efficiently represents sets of cache states by bounding the age of memory blocks from above and below.
- ... requires context-sensitivity for precision.

# Summary

## Cache Analysis for Least-Recently-Used

- ... efficiently represents sets of cache states by bounding the age of memory blocks from above and below.
- ... requires context-sensitivity for precision.

## Predictability Metrics

- ... quantify the predictability of replacement policies.
- LRU is the most predictable policy.

# Summary

## Cache Analysis for Least-Recently-Used

- ... efficiently represents sets of cache states by bounding the age of memory blocks from above and below.
- ... requires context-sensitivity for precision.

## Predictability Metrics

- ... quantify the predictability of replacement policies.
- LRU is the most predictable policy.

## Relative Competitiveness

- ... allows to derive guarantees on cache performance,
- ... yields first *may*-analyses for FIFO and MRU.

# Summary

## Cache Analysis for Least-Recently-Used

- ... efficiently represents sets of cache states by bounding the age of memory blocks from above and below.
- ... requires context-sensitivity for precision.

## Predictability Metrics

- ... quantify the predictability of replacement policies.
- LRU is the most predictable policy.

## Relative Competitiveness

- ... allows to derive guarantees on cache performance,
- ... yields first *may*-analyses for FIFO and MRU.

## Sensitivity Analysis

- ... determines the influence of initial state on cache performance.

# Summary

## Cache Analysis for Least-Recently-Used

- ... efficiently represents sets of cache states by bounding the age of memory blocks from above and below.
- ... requires context-sensitivity for precision.

## Predictability Metrics

- ... quantify the predictability of replacement policies.
- LRU is the most predictable policy.

## Relative Competitiveness

- ... allows to derive guarantees on cache performance,
- ... yields first *may*-analyses for FIFO and MRU.

## Sensitivity Analysis

- ... determines the influence of initial state on cache performance.

Thank you for your attention!

# Summary

## Cache Analysis for Least-Recently-Used

- ... efficiently represents sets of cache states by bounding the age of memory blocks from above and below.
- ... requires context-sensitivity for precision.

## Predictability Metrics

- ... quantify the predictability of replacement policies.
- LRU is the most predictable policy.

## Relative Competitiveness

- ... allows to derive guarantees on cache performance,
- ... yields first *may*-analyses for FIFO and MRU.

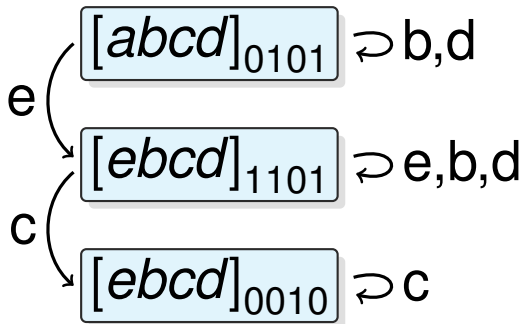
## Sensitivity Analysis

- ... determines the influence of initial state on cache performance.

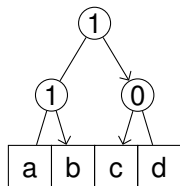
Thank you for your attention!



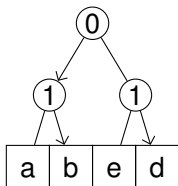
MRU-bits record whether line was recently used



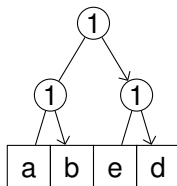
→ Never converges



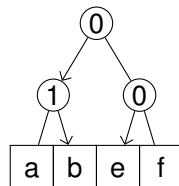
Initial cache-  
set state  
 $[a, b, c, d]_{110}$ .



After a miss  
on  $e$ . State:  
 $[a, b, e, d]_{011}$ .



After a hit  
on  $a$ . State:  
 $[a, b, e, d]_{111}$ .



After a miss  
on  $f$ . State:  
 $[a, b, e, f]_{010}$ .

Hit on  $a$  “rejuvenates” neighborhood; “saves”  $b$  from eviction.

$$May^P(s) := \bigcup_{p \in C^P} CC_P(update_P(p, s))$$

$$Must^P(s) := \bigcap_{p \in C^P} CC_P(update_P(p, s))$$

$$may^P(n) := |May^P(s)|, \text{ where } s \in S^\neq \subsetneq M^*, |s| = n$$

$$must^P(n) := |Must^P(s)|, \text{ where } s \in S^\neq \subsetneq M^*, |s| = n$$

$S^\neq$  : set of finite access sequences with pairwise different accesses

$$\begin{aligned}\text{Evict}^{\mathbf{P}} &:= \min \left\{ n \mid \text{may}^{\mathbf{P}}(n) \leq n \right\}, \\ \text{Fill}^{\mathbf{P}} &:= \min \left\{ n \mid \text{must}^{\mathbf{P}}(n) = k \right\},\end{aligned}$$

where  $k$  is  $\mathbf{P}$ 's associativity.

Let  $P(k)$  be  $(1, 0)$ -miss-competitive relative to policy  $Q(I)$ , then

- (i)  $Evict^P(k) \geq Evict^Q(I)$ ,
- (ii)  $mls^P(k) \geq mls^Q(I)$ .

Let  $I$  be the smallest associativity, such that  $\text{LRU}(I)$  is  $(1, 0)$ -miss-competitive relative to  $P(k)$ . Then

$$\text{Alt-Evict}^P(k) = I.$$

Let  $I$  be the greatest associativity, such that  $P(k)$  is  $(1, 0)$ -miss-competitive relative to  $\text{LRU}(I)$ . Then

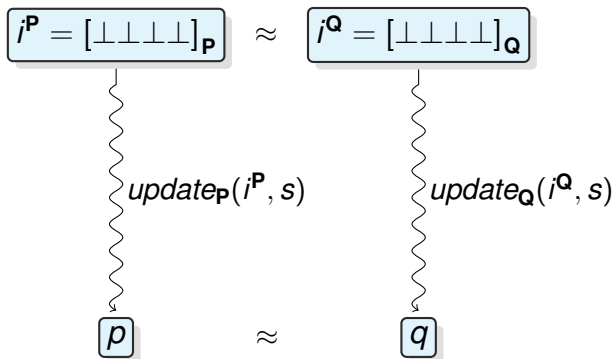
$$\text{Alt-mls}^P(k) = I.$$

$$\underbrace{2^{l+l'}}_{\text{status bits of } \mathbf{P} \text{ and } \mathbf{Q}} \cdot \underbrace{\sum_{i=0}^k \binom{k}{i}}_{\text{non-empty lines in } \mathbf{P}} \cdot \underbrace{\sum_{i'=0}^{k'} \binom{k'}{i'}}_{\text{non-empty lines in } \mathbf{Q}} \cdot \underbrace{\sum_{j=0}^{\min\{i,i'\}} \binom{i}{j} \binom{i'}{j} j!}_{\text{number of overlappings in non-empty lines}}$$

$$\begin{aligned}
 \sum_{j=0}^{\min\{k,k'\}} \binom{k}{j} \binom{k'}{j} j! &\leq k! \cdot k'! \sum_{j=0}^{\min\{k,k'\}} \frac{1}{(k-j)! j! (k'-j)!} \\
 &\leq k! \cdot k'! \sum_{j=0}^{\infty} \frac{1}{j!} = e \cdot k! \cdot k'!
 \end{aligned}$$

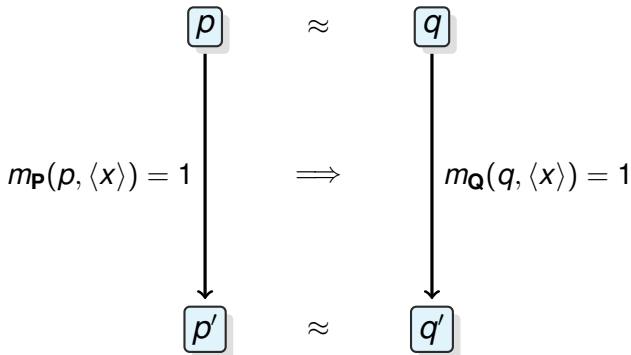
This can be bounded by

$$2^{l+l'+k+k'} \leq |(C_k^l \times C_{k'}^{l'}) / \approx| \leq 2^{l+l'+k+k'} \cdot \underbrace{e \cdot k! \cdot k'!}_{\text{bound on number of overlappings}}$$

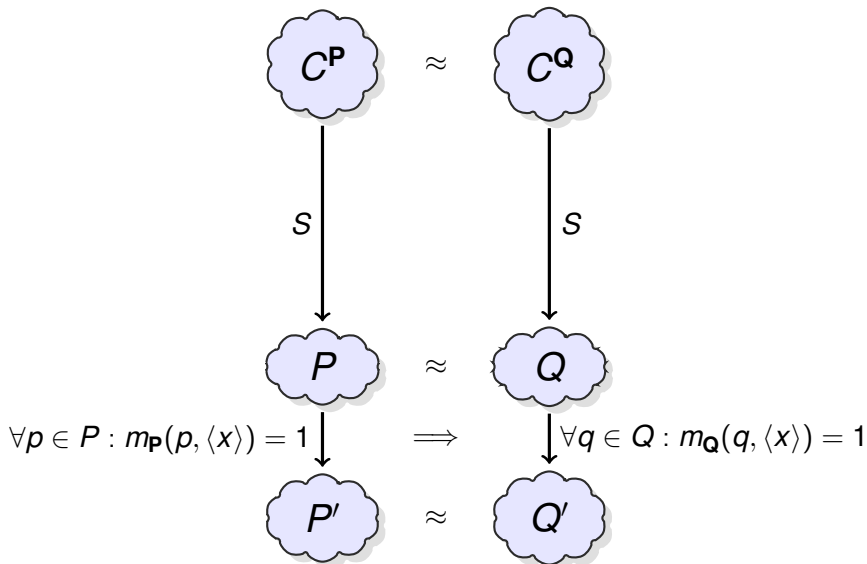




Let  $\mathbf{P}$  be (1, 0)-competitive relative to  $\mathbf{Q}$ , then



# (1, 0)-Competitiveness and May/Must-Analyses



- Simple model of execution time from Hennessy & Patterson (2003)
- $CPI_{hit}$  = Cycles per instruction assuming cache hits only
- $\frac{\text{Memory accesses}}{\text{Instruction}}$  including instruction and data fetches

$$\begin{aligned}\frac{T_{wc}}{T_{meas}} &= \frac{CPI_{hit} + \frac{\text{Memory accesses}}{\text{Instruction}} \times \text{Miss rate}_{wc} \times \text{Miss penalty}}{CPI_{hit} + \frac{\text{Memory accesses}}{\text{Instruction}} \times \text{Miss rate}_{meas} \times \text{Miss penalty}} \\ &= \frac{1.5 + 1.2 \times 0.20 \times 50}{1.5 + 1.2 \times 0.05 \times 50} = \frac{13.5}{4.5} = 3\end{aligned}$$