Worst-Case Execution Time Analysis

Prof. Dr. Jian-Jia Chen

LS 12, TU Dortmund

29, 30 April, 2019
Most Essential Assumptions for Real-Time Systems

Upper bound on the execution times:

- Commonly, called the *Worst-Case Execution Time (WCET)*
What does Execution Time Depend on

• Input parameters
  • Algorithm parameters
  • Problem size
  • etc.

• Initial states and intermediate states of the system while executing
  • Cache configuration, replacement policies
  • Pipelines
  • Speculations
  • etc.

• Interferences from the environment
  • Scheduling
  • Interrupts
  • etc.
How to Derive the Worst-Case Execution Time (WCET)

- Most of industry’s best practice
  - Measure it: determine WCET directly by running or simulating a set of inputs.
    - There is no guarantee to give an upper bound of the WCET.
    - The derived WCET could be too optimistic.
  - Exhaustive execution: by considering the set of all the possible inputs
    - In general, not possible
    - The inputs have to cover all the possible initial states and intermediate states of the system, which is also usually not possible.

- Compute it
  - In general, not possible neither, as computing (tight) WCET for a program is *uncomputable* by Turing machines.
  - Based on some structures, it is possible and the derived solution is a safe upper bound of the WCET.
Why is It Uncomputable?

Halting Problem

Given the description of a Turing machine \( m \) and its input \( x \), the problem is to answer the question whether the machine halts on \( x \).

Theorem

The Halting Problem is undecidable (uncomputable). In other words, one cannot use an algorithm to decide whether another algorithm \( m \) halts on a specific input.

WCET is undecidable

It is even undecidable if it terminates at all. Deriving the WCET is of course undecidable.

Please refer to the textbook of Computational Complexity by Prof. Papadimitriou.
Our objectives:

- **Upper bound** of the execution time as tightly as possible.
- All control-flow paths, by considering all possible inputs.
- All paths through the architecture, resulting from the potential initial and assumed intermediate architectural states.
Timing Analysis

By considering systems, in general, with

- finite architectural configurations, finite input domains, and bounded loops and recursion,

WCET is computable.
Timing Analysis

By considering systems, in general, with
- finite architectural configurations, finite input domains, and bounded loops and recursion,

WCET is computable.

But....... the search space is too large to explore it exhaustively!
Why is It Hard for Analyzing WCET?

Execution time \( e(i) \) of machine instruction \( i \)

- In the good old time:
  \( e(i) \) is a constant \( c \), which could be found in the data sheet
- Nowadays, especially for high-performance processors:
  \( e(i) \) also depends on the (architectural) execution state \( s \).

\[
\min\{e(i, s) | s \in S\} \leq e(i) \leq \max\{e(i, s) | s \in S\},
\]

where \( S \) is the set of all states.

- Using \( \max\{e(i, s) | s \in S\} \) is safe for WCET, but might be not tight since some states in \( S \) might not be possible to reach by some inputs.

- Execution history, resulting in a smaller set of reachable execution states, has to be enforced to improve the tightness of the analysis.
Variability of Execution Times

\[ x = a + b; \]

![Assembly Code](image)

```
LOAD r2, _a
LOAD r1, _b
ADD r3, r2, r1
```

PPC 755

![Execution Time Graph](image)

Execution Time (Clock Cycles)

- Best Case
- Worst Case
Timing Accidents and Penalties

- **Timing Accident:** cause for an increase of the execution time of an instruction
- **Timing Penalty:** the associated increase
- **Types of timing accidents**
  - Cache misses
  - Pipeline stalls
  - Branch mispredictions
  - Bus collisions
  - Memory refresh of DRAM
  - TLB miss
Overall Approach: Modularization

- Architecture Analysis:
  - Use Abstract Interpretation.
  - Exclude as many Timing Accidents as possible during analysis.
    - Certain timing accidents will never happen, e.g., at a certain program point, instruction fetch will never cause a cache miss.
    - The more accidents excluded, the lower (better) the upper bound.
  - Determine WCET for basic blocks, based on context information.
- Worst-Case Path Determination:
  - Map control flow graph to an Integer Linear Program (ILP).
  - Determine upper bound and associated path.

High-Level Objectives: Upper Bound of WCET

- It must be safe, i.e., not underestimate.
- It should be tight, i.e., not far away from real WCET.
- The analysis effort must be tolerable.
Overall Structure

Executable Binary
Program

Control-Flow-Graph (CFG)
Reconstruction

Loop Analysis and Unfolding

Loop Bounds

Static Analysis
Value Analyzer
Cache/Pipeline Analyzer

Micro-Architecture Abstraction
Timing Information

Path Analysis
ILP-Generator
ILP-Solver
Evaluation

WCET Visualization and Analysis

Micro-architecture Analysis
Worst-Case Path Analysis
Outline

Introduction

Program Path Analysis

Static Analysis
  Value Analysis
  Cache Analysis
  Pipeline Analysis
```
what_is_this {
  read (a,b);
  done = FALSE;
  repeat {
    if (a>b) 
      a = a-b;
    elseif (b>a)
      b = b-a;
    else done = TRUE;
  } until done;
  write (a);
}
```
Basic Blocks

Definition: A basic block is a sequence of instructions where the control flow enters at the beginning and exits at the end, in which it is highly amenable to analysis.

\[
\begin{align*}
a[0] & := b[0] + c[0] \\
d & := a[0] \ast a[1] \\
e & := d / a[2] \\
\text{if } e < 10 \text{ goto } L
\end{align*}
\]
Basic Blocks

Definition: A basic block is a sequence of instructions where the control flow enters at the beginning and exits at the end, in which it is highly amenable to analysis.

Determining the basic blocks

- **Beginning:**
  - the first instruction
  - targets of un/conditional jumps
  - instructions that follow un/conditional jumps

- **Ending:**
  - the basic block consists of the block beginning and runs until the next block beginning (exclusive) or until the program ends

```
a[0] := b[0] + c[0]
d := a[0] * a[1]
e := d / a[2]
if e < 10 goto L
```
Program Path Analysis

- Problem: Which sequence of instructions is executed in the worst case (i.e., the longest execution time)?
- Input:
  - Timing information for each basic block, derived from static analysis (value/cache/pipeline analysis)
  - Loop bounds by specification
  - CFG derived from the executable binary program
- Basic Concept:
  - Transform structure of CFG into a set of (integer) linear equations
  - Solution of the Integer Linear Program (ILP) yields bound on the WCET.
Program Path Analysis: Formal Definition

<table>
<thead>
<tr>
<th>Input</th>
</tr>
</thead>
<tbody>
<tr>
<td>A CFG with $N$ basic blocks, in which each basic block $B_i$ has a worst-case execution time $c_i$, given by static analysis.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>Suppose that each block $B_i$ is executed exactly $x_i$ times. What is the worst-case execution time $\text{WCET} = \sum_{i=1}^{N} c_i \cdot x_i$, such that the values of $x_i$s satisfy the structural constraints in the CFG?</td>
</tr>
</tbody>
</table>

Note that additional constraints provided by the programmer (bounds for loop counters, etc.) can also be included.
Example for CFG Constraints

Flow equations: ($x_i$ is a variable)

\[ d_1 = d_2 = x_1 \]
\[ d_3 + d_9 = d_2 + d_8 = x_2 \]
\[ d_4 + d_5 = d_3 = x_3 \]
\[ d_6 + d_7 = d_8 = x_4 \]
\[ d_4 = d_6 = x_5 \]
\[ d_5 = d_7 = x_6 \]

\[ s = k; \]

\[ \text{while} (k < 20) \]

\[ \text{if} (ok) \]

\[ j++; \]
\[ j=0; \text{ok=true}; \]

\[ k++; \]

\[ r=j \]
Example for Additional Constraints

The loop is executed for at most 20 times when \( k \) is initialized with a non-negative number:

\[ x_3 \leq 20x_1. \]

The basic block for \( j = 0; ok = true; \) is executed for at most one time:

\[ x_6 \leq x_1. \]
WCET: ILP Formulation

\[
WCET = \max \left\{ \sum_{i=1}^{N} c_i \cdot x_i \right\}
\]

\[
|d_1 = 1
\]

and

\[
\sum_{j \in in(B_i)} d_j = \sum_{k \in out(B_i)} d_k = x_i, \ i = 1, \ldots, N
\]

and additional constraints
Outline

Introduction

Program Path Analysis

Static Analysis
  Value Analysis
  Cache Analysis
  Pipeline Analysis
Overall Structure

Executable Binary

Control–Flow–Graph (CFG) Reconstruction

Loop Analysis and Unfolding

Loop Bounds

Static Analysis

Value Analyzer

Cache/Pipeline Analyzer

Micro–Architecture Abstraction

Timing Information

WCET Visualization and Analysis

Path Analysis

ILP–Generator

ILP–Solver

Evaluation

Micro–architecture Analysis

Worst–Case Path Analysis
Outline

Introduction

Program Path Analysis

Static Analysis
  Value Analysis
  Cache Analysis
  Pipeline Analysis
Value Analysis: Motivation and Method

- **Motivation**
  - Provide access information to data-cache/pipeline analysis
  - Detect infeasible paths
  - Derive loop bounds

- **Method**
  - Calculate intervals at all program points
  - By considering addresses, register contents, local and global variables.

**Abstract Interpretation**

Perform the program’s computation using value descriptions or abstract values in place of the concrete values.
Abstract Interpretation: Manipulation

- abstract domain - related to concrete domain by abstraction and concretization functions
  - Replace an integer/double operator by using intervals
  - e.g., \( L = [3, 5] \) stands for \( L \) is a value between 3 and 5
- abstract transfer functions for each statement type
  - e.g., operator \(+\): \([3, 5] + [2, 6] = [5, 11]\)
  - e.g., operator \(-\): \([3, 5] - [2, 6] = [-3, 3]\)
- a join function combining intervals from different paths
  - \([a, b] \) join \([c, d]\) becomes \([\min\{a, c\}, \max\{b, d\}]\)
  - \([3, 5] \) join \([2, 4]\) becomes \([2, 5]\).
Value Analysis

D1:[-4,4], A0:[0x1000,0x1000]

move #4,D0

D0:[4,4], D1:[-4,4], A0:[0x1000,0x1000]

add D1,D0

D0:[0,8], D1:[-4,4], A0:[0x1000,0x1000]

move (A0,D0),D1

access [0x1000,0x1008]

- Intervals are computed along the CFG edges
- At joins, intervals are "unioned"

D1: [-2,+2]

D1: [-4,0]

D1: [-4,+2]
Outline

Introduction

Program Path Analysis

Static Analysis
  Value Analysis
  Cache Analysis
  Pipeline Analysis
Overall Structure

Executable Binary Program

Control-Flow-Graph (CFG) Reconstruction

Loop Analysis and Unfolding

Loop Bounds

Static Analysis

Value Analyzer

Cache/Pipeline Analyzer

Micro-Architecture Abstraction

Timing Information

Path Analysis

ILP-Generator

ILP-Solver

Evaluation

WCET Visualization and Analysis

Micro-architecture Analysis

Worst-Case Path Analysis
Caches: Fast Memory to Deal with the Memory Wall

- How they work:
  - dynamically
  - managed by replacement policy

![Diagram of CPU, Cache, and Main Memory]

- Why they work: principle of locality
  - spatial
  - temporal

Capacity:
CPU: 32 KB
Cache: 3 MB
Main Memory: 2 MB

Latency:
CPU: 3 cycles
Cache: 100 cycles
Caches: Fast Memory to Deal with the Memory Wall

- **How they work:**
  - dynamically
  - managed by replacement policy

- **Why they work:** principle of locality
  - spatial
  - temporal

**Diagram:**
- CPU
- Cache
- Main Memory

**Capacity:**
- Cache: 32 KB
- Main Memory: 2 MB

**Latency:**
- Cache: 3 cycles
- Main Memory: 100 cycles

“hit” [ab]
Caches: Fast Memory to Deal with the Memory Wall

- **How they work:**
  - dynamically
  - managed by replacement policy

![Diagram of cache system](image)

- **Why they work:** *principle of locality*
  - spatial
  - temporal

- **CPU**
  - Capacity: 32 KB
  - Latency: 3 cycles
- **Cache**
  - "hit" [ab]
  - \( a_1 \)!
- **Main Memory**
  - 2 MB
  - Latency: 100 cycles
Caches: Fast Memory to Deal with the Memory Wall

• How they work:
  • dynamically
  • managed by replacement policy

![Diagram of CPU, Cache, and Main Memory]

Capacity: 32 KB
Latency: 3 cycles

“miss” [ab]

Why they work: principle of locality
  • spatial
  • temporal
Caches: Fast Memory to Deal with the Memory Wall

- **How they work:**
  - dynamically
  - managed by replacement policy

```
CPU                   Cache                   Main Memory

Capacity: 32 KB
Latency: 3 cycles
```

- **Why they work:** *principle of locality*
  - spatial
  - temporal

"miss" [ab]
Caches: Fast Memory to Deal with the Memory Wall

- **How they work:**
  - dynamically
  - managed by replacement policy

![Diagram of CPU, Cache, and Main Memory]

- **Why they work:** *principle of locality*
  - spatial
  - temporal

- Capacity: 32 KB
- Latency: 3 cycles
- Main Memory: 2 MB
- Latency: 100 cycles

\[ c = \langle c_1, c_2, c_3, c_4 \rangle ! \]
Caches: Fast Memory to Deal with the Memory Wall

- How they work:
  - dynamically
  - managed by replacement policy

```
CPU  Cache  Main Memory
```

- Why they work: principle of locality
  - spatial
  - temporal

```
Capacity: 32 KB
Latency: 3 cycles
```

```
Capacity: 2 MB
Latency: 100 cycles
```

"miss" [ac]
Caches: Fast Memory to Deal with the Memory Wall

- **How they work:**
  - dynamically
  - managed by replacement policy

  ![Diagram of CPU, Cache, and Main Memory]

  - **CPU**
  - **Cache**
  - **Main Memory**

  - Capacity: 32 KB
  - Latency: 3 cycles
  - Capacity: 2 MB
  - Latency: 100 cycles

- **Why they work:** *principle of locality*
  - spatial
  - temporal
Caches: Fast Memory to Deal with the Memory Wall

- How they work:
  - dynamically
  - managed by replacement policy

![Diagram of cache memory system]

- Why they work: *principle of locality*
  - spatial
  - temporal

Capacity:
- Cache: 32 KB
- Main Memory: 2 MB

Latency:
- Cache: 3 cycles
- Main Memory: 100 cycles
Fully Associative Caches

Address:

Tag

Block offset

$\log_2(8 \times b)$

= ?

No: Miss!

Yes: Hit!

MUX

Data

$\hat{k}$ = associativity

Tag

Data Block

Tag

Data Block

Tag

Data Block

$\log_2(s)$

$\log_2(8 \times b)$

$\times$
Set-Associative Caches

Special cases:

- direct-mapped cache: only one line per cache set
- fully-associative cache: only one cache set
Replacement Policies

- Least-Recently-Used (LRU) used in Intel Pentium I and MIPS 24K/34K
- First-In First-Out (FIFO or Round-Robin) used in Motorola PowerPC 56x, Intel XScale, ARM9, ARM11
- Pseudo-LRU (PLRU) used in Intel Pentium II-IV and PowerPC 75x
- Most Recently Used (MRU) as described in literature

Each cache set is treated independently:

\[\rightarrow\text{Set-associative caches are compositions of fully-associative caches.}\]
Cache Analysis for LRU

Two types of cache analyses:

1. Local guarantees: classification of individual accesses
   - Must-Analysis \(\rightarrow\) Underapproximates cache contents
   - May-Analysis \(\rightarrow\) Overapproximates cache contents

2. Global guarantees: bounds on cache hits/misses
Challenges for Cache Analysis

Always a cache hit/always a miss?

read z

read y

read x

write z
Challenges for Cache Analysis

Always a cache hit/always a miss?

1. Initial cache contents unknown.
2. Different paths lead to these points.
3. Cannot resolve address of z.
Using Abstract Interpretation

**Collecting Semantics** = set of states at each program point that any execution may encounter there

Two approximations:

- Collecting Semantics uncomputable
- \( \subseteq \) Cache Semantics computable
- \( \subseteq \gamma\) (Abstract Cache Sem.) efficiently computable
Using Abstract Interpretation

Collecting Semantics $\equiv$
set of states at each program point that any execution may encounter there

Two approximations:

Collecting Semantics uncomputable
\subseteq Cache Semantics computable
\subseteq $\gamma$(Abstract Cache Sem.) efficiently computable
Using Abstract Interpretation

**Collecting Semantics** = set of states at each program point that any execution may encounter there

Two approximations:

- Collecting Semantics uncomputable
- ⊆ Cache Semantics computable
- ⊆ γ(Abstract Cache Sem.) efficiently computable
Using Abstract Interpretation

Collecting Semantics = set of states at each program point that any execution may encounter there

Two approximations:
- Collecting Semantics uncomputable
- Cache Semantics computable
- $\subseteq \gamma$(Abstract Cache Sem.) efficiently computable
Using Abstract Interpretation

Collecting Semantics = set of states at each program point that any execution may encounter there

Two approximations:

- Collecting Semantics uncomputable
- Cache Semantics computable
- $\gamma$(Abstract Cache Sem.) efficiently computable
Least-Recently-Used (LRU): Concrete Behavior

“Cache Miss”:

“Cache Hit”:

LRU has notion of age
LRU: Must-Analysis: Abstract Domain

- Used to predict cache hits.
- Maintains upper bounds on ages of memory blocks.
- Upper bound $\leq$ associativity $\rightarrow$ memory block definitely cached.

Example

Describes the set of all concrete cache states in which $x$, $s$, and $t$ occur,

$\gamma([\{x\}, \{\}, \{s, t\}, \{\}]) = \{[x, s, t, a], [x, t, s, a], [x, s, t, b], \ldots\}$
Sound Update – Local Consistency

Abstract Update

\((must)\) \[\rightarrow\] \((must')\)

\[\gamma\]

Lifted Concrete Update

concrete cache states

concrete cache states
LRU: Must-Analysis: Update

“Potential Cache Miss”:

“Definite Cache Hit”:

Why does \( t \) remain in the same set in Abstract Interpretation in the second case?
LRU: Must-Analysis: Join

Need to combine information where control-flow merges.

Join should be conservative:

- $\gamma(A) \subseteq \gamma(A \sqcup B)$
- $\gamma(B) \subseteq \gamma(A \sqcup B)$

```
{a}  \{c\}  \{e\}  \{d\}
{d}  \{a\}  \{a,c\}  \{d\}
{c,f} \{e\}  \{a\}  
{d}  
{a,c}  
{d}  
```

“Intersection + Maximal Age”
LRU: Must-Analysis: Join

Need to combine information where control-flow merges.

Join should be conservative:

- \( \gamma(A) \subseteq \gamma(A \sqcup B) \)
- \( \gamma(B) \subseteq \gamma(A \sqcup B) \)

```
\[
\begin{array}{c}
\{a\} \\
\{\} \\
\{c,f\} \\
\{d\}
\end{array}
\quad \sqcup \\
\begin{array}{c}
\{c\} \\
\{e\} \\
\{a\} \\
\{d\}
\end{array}
\quad \begin{array}{c}
\{\} \\
\{\} \\
\{a,c\} \\
\{d\}
\end{array}
\]
```

“Intersection + Maximal Age”
LRU: Must-Analysis: Join

Need to combine information where control-flow merges.

Join should be conservative:

- $\gamma(A) \subseteq \gamma(A \sqcup B)$
- $\gamma(B) \subseteq \gamma(A \sqcup B)$

```
{a}
{ }
{c,f}
{d}

{c}
{e}
{a}
{d}

{}  
{}  
{a,c}  
{d}
```

“Intersection + Maximal Age”
LRU: Must-Analysis: Join

Need to combine information where control-flow merges.

Join should be conservative:

- $\gamma(A) \subseteq \gamma(A \sqcup B)$
- $\gamma(B) \subseteq \gamma(A \sqcup B)$

```
\begin{array}{lll}
\{a\} & \{c\} & \{\}\ \\
\{\} & \{e\} & \{\} \\
\{c,f\} & \{a\} & \{a,c\} \\
\{d\} & \{d\} & \{d\}
\end{array}
```

“Intersection + Maximal Age”

Prof. Dr. Jian-Jia Chen (LS 12, TU Dortmund)
LRU: Must-Analysis: Join

Need to combine information where control-flow merges.

Join should be conservative:

- \( \gamma(A) \subseteq \gamma(A \sqcup B) \)
- \( \gamma(B) \subseteq \gamma(A \sqcup B) \)

```
\{a\}  \{c\}  \{\}
\{\}   \{e\}  \{\}
\{c,f\}\{a\}  \{a,c\}
\{d\}  \{d\}  \{d\}
```

“Intersection + Maximal Age”
LRU: Must-Analysis: Join

Need to combine information where control-flow merges.

Join should be conservative:

- $\gamma(A) \subseteq \gamma(A \sqcup B)$
- $\gamma(B) \subseteq \gamma(A \sqcup B)$

```
{a}  {c}  {}
{}   {e}  {}
{c,f} {a}  {a,c}
{d}   {d}  {d}
```

“Intersection + Maximal Age”
LRU: Must-Analysis: Join

Need to combine information where control-flow merges.

Join should be conservative:

- $\gamma(A) \subseteq \gamma(A \sqcup B)$
- $\gamma(B) \subseteq \gamma(A \sqcup B)$

```
\begin{array}{c|c|c}
  & \{c\} & \{c\} \\
\hline
\{a\} & \{e\} & \{\}
\end{array}
```

```
\begin{array}{c|c|c}
  & \{d\} & \{d\} \\
\hline
\{\} & \{a,c\} & \{\}
\end{array}
```

“Intersection + Maximal Age”

How many memory blocks can be in the must-cache?
Example: Must-Analysis

entry \[\{\}, \{\}, \{\}, \{\}\]

A \perp B \perp C \perp D \perp exit

Prof. Dr. Jian-Jia Chen (LS 12, TU Dortmund) 41 / 53
Example: Must-Analysis

entry \[\{\}, \{\}, \{\}, \{\}\]

\[\bot \sqcup [\{\}, \{\}, \{\}, \{\}] = [\{\}, \{\}, \{\}, \{\}]\]

exit \[\top\]
Example: Must-Analysis

\[ \{\}, \{\}, \{\}, \{\} \]

\[ \perp \sqcup \{\}, \{\}, \{\}, \{\} \] = \{\}, \{\}, \{\}, \{\}

\[ \{A\}, \{\}, \{\}, \{\} \]

\[ \{A\}, \{\}, \{\}, \{\} \]

\[ \{\}, \{\}, \{\}, \{\} \]

\[ \{\}, \{\}, \{\}, \{\} \]

\[ \{\}, \{\}, \{\}, \{\} \]

\[ \{\}, \{\}, \{\}, \{\} \]

exit \[ \perp \]
Example: Must-Analysis

\[
\begin{align*}
\text{entry} & \quad [\{\}, \{\}, \{\}, \{\}, \{\}] \\
\perp \sqcup [\{\}, \{\}, \{\}, \{\}, \{\}] &= [\{\}, \{\}, \{\}, \{\}, \{\}]
\end{align*}
\]

\[
\begin{align*}
[\{A\}, \{\}, \{\}, \{\}, \{\}] & \quad [\{A\}, \{\}, \{\}, \{\}, \{\}]
\end{align*}
\]

\[
\begin{align*}
[\{B\}, \{A\}, \{\}, \{\}, \{\}] \sqcup [\{C\}, \{A\}, \{\}, \{\}, \{\}] &= [\{\}, \{A\}, \{\}, \{\}, \{\}]
\end{align*}
\]

exit \quad \perp
Example: Must-Analysis

entry \[[\{\}, \{\}, \{\}, \{\}\]\]

\[
\{D\}, \{\}, \{A\}, \{\} \sqcup \{\}, \{\}, \{\}, \{\} = \\
\{\}, \{\}, \{\}, \{\}
\]

\[[\{A\}, \{\}, \{\}, \{\}\]\]

D

\[[\{B\}, \{A\}, \{\}, \{\}\] \sqcup \{\}, \{C\}, \{A\}, \{\}, \{\} = \\
\{\}, \{A\}, \{\}, \{\}\]

B

A

C

\[[\{A\}, \{\}, \{\}, \{\}\]\]

\[[\{D\}, \{\}, \{A\}, \{\}\]

exit

No cache hits can be predicted!
Context-Sensitive Analysis/Virtual Loop-Unrolling

- **Problem:**
  - The first iteration of a loop will always result in cache misses.
  - Similarly for the first execution of a function.

- **Solution:**
  - Virtually Unroll Loops: Distinguish the first iteration from others
  - Distinguish function calls by calling context.

Virtually unrolling the loop once:

- **Accesses to** $A$ **and** $D$ **are provably hits after the first iteration**
- **Accesses to** $B$ **and** $C$ **can still not be classified. Within each execution of the loop, they may only miss once.**

$\rightarrow$ Persistence Analysis
LRU: May-Analysis: Abstract Domain

- Used to predict *cache misses*.
- Maintains *lower bounds on ages* of memory blocks.
- Lower bound $\geq$ associativity
  $\rightarrow$ memory block definitely *not* cached.

**Abstract state:**

<table>
<thead>
<tr>
<th>-aged state</th>
<th>age 0</th>
<th>age 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>${x, y}$</td>
<td>${}$</td>
<td>${}$</td>
</tr>
<tr>
<td>${s, t}$</td>
<td>${}$</td>
<td>${}$</td>
</tr>
<tr>
<td>${u}$</td>
<td>${}$</td>
<td>${}$</td>
</tr>
</tbody>
</table>

and its interpretation:

Describes a set of all concrete cache states, where no memory blocks except $x$, $y$, $s$, $t$, and $u$ occur,

- $x$ and $y$ with an age of at least 0,
- $s$ and $t$ with an age of at least 2,
- $u$ with an age of at least 3.

$$\gamma([\{x, y\}, \{}, \{s, t\}, \{u\}]) = \{[x, y, s, t], [y, x, s, t], [x, y, s, u], \ldots\}$$
LRU: May-Analysis: Update

“Definite Cache Miss”:

```
{x}
{}  {z}
{y}  {x}
{s,t}  {s,t}
```

“Potential Cache Hit”:

```
{x}
{}  {s}
{y}  {x}
{s,t}  {x}
{y,t}  {y,t}
```

Why does $t$ move to an older set in Abstract Interpretation in the second case?
LRU: May-Analysis: Join

Need to combine information where control-flow merges.

Join should be conservative:

- $\gamma(A) \subseteq \gamma(A \sqcup B)$
- $\gamma(B) \subseteq \gamma(A \sqcup B)$

\[
\begin{array}{c}
\{a\} \\
\{} \\
\{c,f\} \\
\{d\}
\end{array}
\sqcup
\begin{array}{c}
\{c\} \\
\{e\} \\
\{a\} \\
\{d\}
\end{array}
= \begin{array}{c}
\{a,c\} \\
\{e\} \\
\{f\} \\
\{d\}
\end{array}
\]

"Union + Minimal Age"
LRU: May-Analysis: Join

Need to combine information where control-flow merges.

Join should be conservative:

- $\gamma(A) \subseteq \gamma(A \sqcup B)$
- $\gamma(B) \subseteq \gamma(A \sqcup B)$

```
\begin{array}{c}
\{a\} \\
\{} \\
\{c,f\} \\
\{d\}
\end{array}
\sqcup
\begin{array}{c}
\{c\} \\
\{e\} \\
\{a\} \\
\{d\}
\end{array}
=
\begin{array}{c}
\{a,c\} \\
\{e\} \\
\{f\} \\
\{d\}
\end{array}
```

“Union + Minimal Age”
LRU: May-Analysis: Join

Need to combine information where control-flow merges.

Join should be conservative:

- $\gamma(A) \subseteq \gamma(A \sqcup B)$
- $\gamma(B) \subseteq \gamma(A \sqcup B)$

```
{a}  \sqcup  {c}  =  {a,c}
{c,f}  \sqcup  {e}  =  {e}
{d}    \sqcup  {f}  =  {f}
{d}          \sqcup  {d}  =  {d}
```

“Union + Minimal Age”
LRU: May-Analysis: Join

Need to combine information where control-flow merges.

Join should be conservative:

- $\gamma(A) \subseteq \gamma(A \sqcup B)$
- $\gamma(B) \subseteq \gamma(A \sqcup B)$

```
 sighting
 \{a\}   \{c\}   \{a,c\}
 \{}      \{e\}      \{e\}
 \{c,f\}  \{a\}      \{f\}
 \{d\}    \{d\}      \{d\}
```
LRU: May-Analysis: Join

Need to combine information where control-flow merges.

Join should be conservative:

- $\gamma(A) \subseteq \gamma(A \sqcup B)$
- $\gamma(B) \subseteq \gamma(A \sqcup B)$

\[
\begin{align*}
\{a\} & \quad \{c\} & \quad \{a,c\} \\
\{\} & \quad \{e\} & \quad \{e\} \\
\{c,f\} & \quad \{a\} & \quad \{f\} \\
\{d\} & \quad \{d\} & \quad \{d\}
\end{align*}
\]

“Union + Minimal Age”
Need to combine information where control-flow merges.

Join should be conservative:

- $\gamma(A) \subseteq \gamma(A \sqcup B)$
- $\gamma(B) \subseteq \gamma(A \sqcup B)$

```
\begin{array}{c}
\{a\} \\
\{} \\
\{c, f\} \\
\{d\}
\end{array} \sqcup \begin{array}{c}
\{c\} \\
\{e\} \\
\{a\} \\
\{d\}
\end{array} = \begin{array}{c}
\{a, c\} \\
\{e\} \\
\{f\} \\
\{d\}
\end{array}
```

“Union + Minimal Age”
Overall Structure

Executable Binary
Program

Control−Flow−Graph (CFG)
Reconstruction

Loop Analysis and Unfolding

Loop Bounds

Static Analysis

Value Analyzer

Cache/Pipeline Analyzer

Path Analysis

ILP−Generator

ILP−Solver

Evaluation

Micro−Architecture Abstraction

Timing Information

WCET Visualization and Analysis

Micro−architecture Analysis

Worst−Case Path Analysis
Pipelines

An instruction execution consists of several sequential phases, e.g.,

- Fetch
- Decode
- Execute
- Write Back

<table>
<thead>
<tr>
<th>Inst 1</th>
<th>Inst 2</th>
<th>Inst 3</th>
<th>Inst 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fetch</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Decode</td>
<td>Fetch</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Execute</td>
<td>Decode</td>
<td>Fetch</td>
<td></td>
</tr>
<tr>
<td>Write Back</td>
<td>Execute</td>
<td>Decode</td>
<td>Fetch</td>
</tr>
<tr>
<td></td>
<td>Write Back</td>
<td>Execute</td>
<td>Decode</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Write Back</td>
<td>Execute</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Write Back</td>
</tr>
</tbody>
</table>
Hardware Features: Pipelines

- Instruction execution is split into several stages.
- Several instructions can be executed in parallel.
- Some pipelines can begin more than one instruction per cycle: VLIW, Superscalar.
- Some CPUs can execute instructions out-of-order.

**Practical Problems: Hazards and cache misses.**
- Data Hazards: Operands not yet available (Data Dependences)
  - By applying dependence analysis
- Control Hazards: Conditional branch
  - By applying dependence analysis
- Resource Hazards: Consecutive instructions use same resource
  - By applying analysis on the resource reservation tables
- Instruction-Cache Hazards: Instruction fetch causes cache miss
  - By applying static cache analysis
CPU as a (Concrete) State Machine

- Processor (pipeline, cache, memory, inputs) viewed as a *big state machine*, performing transitions every clock cycle.
- Starting in an initial state for an instruction, transitions are performed, until a final state is reached:
  - end state: instruction has left the pipeline
  - \# transitions: execution time of instruction
- function `exec (b: basic block, s: concrete pipeline state) t: trace`
  - interprets instruction stream of `b` starting in state `s` producing trace `t`
  - successor basic block is interpreted starting in initial state `last(t)`
  - `length(t)` gives number of cycles
- function `exec (b: basic block, s: abstract pipeline state) t: trace`
  - interprets instruction stream of `b` (annotated with cache information) starting in state `s` producing trace `t`
  - `length(t)` gives number of cycles
- Alternatives: use a set of states for an instruction instead of one starting state. This is useful when there are multiple predecessors.
Summary of WCET Analysis

- Value analysis
- Cache analysis
  - using statically computed effective addresses and loop bounds
- Pipeline analysis
  - assume cache hits where predicted,
  - assume cache misses where predicted or not excluded.
  - Only the “worst” result states of an instruction need to be considered as input states for successor instructions!
aiT-Tool