Hardware-Software Codesign

9. Worst Case Execution Time Analysis

Lothar Thiele
System Design

Swiss Federal Institute of Technology
Computer Engineering
and Networks Laboratory

Specification → System Synthesis

SW-Compilation → Instruction Set → HW-Synthesis

Estimation

Intellectual Prop. Code

Machine Code

Instruction Set

Net lists

Intellectual Prop. Block
Performance Estimation Methods – Illustration

- Real system
- Measurement
- Simulation
- Probabilistic estimation
- Worst case (formal) analysis

→ chapter 9
→ chapter 10-11
Contents

► Introduction
  ▪ problem statement, tool architecture

► Program Path Analysis

► Value Analysis

► Caches
  ▪ must, may analysis

► Pipelines
  ▪ Abstract pipeline models
  ▪ Integrated analyses

► The slides are based on lectures of Reinhard Wilhelm.
Industrial Needs

- **Hard real-time systems**, abound often in safety-critical applications
  - Aeronautics, automotive, train industries, manufacturing control

Sideairbag in car,
Reaction in <10 mSec

Wing vibration of airplane,
sensing every 5 mSec
Hard Real-Time Systems

- Embedded controllers are expected to finish their tasks reliably within time bounds.

- Task scheduling must be performed.

- Essential: *upper bound on the execution times* of all tasks statically known.

- Commonly called the *Worst-Case Execution Time* (WCET)

- Analogously, *Best-Case Execution Time* (BCET)
Measurement – Industry's “best practice”

Works if either
• worst-case input can be determined, or
• exhaustive measurement is performed

Otherwise, determine upper bound from execution times of instructions

Distribution of execution times

Best Case Execution Time

Unsafe:
Execution Time Measurement

Worst Case Execution Time

Upper bound

does this really work?
(Most of) Industry’s Best Practice

**Measurements**: determine execution times directly by observing the execution or a simulation on a set of inputs.
- Does *not guarantee* an upper bound to all executions in general.
- *Exhaustive execution* in general *not possible*! Too large space of (input domain) \( \times \) (set of initial execution states).

**Compute upper bounds** along the *structure* of the program:
- Programs are hierarchically structured.
- Statements are nested inside statements.
- So, compute the upper bound for a statement from the upper bounds of its constituents.
Sequence of Statements

Constituents of $A$:
$A_1$ and $A_2$

Upper bound for $A$ is the sum of the upper bounds for $A_1$ and $A_2$

$\text{ub}(A) = \text{ub}(A_1) + \text{ub}(A_2)$
Conditional Statement

\[ A \equiv \text{if } B \text{ then } A_1 \text{ else } A_2 \]

**Constituents of A:**
1. condition \( B \)
2. statements \( A_1 \) and \( A_2 \)

\[ \text{ub}(A) = \text{ub}(B) + \max(\text{ub}(A_1), \text{ub}(A_2)) \]
Loops

\[ A \equiv \text{for } i \leftarrow 1 \text{ to } 100 \text{ do } \]

\[ A1 \]

\[ i \leftarrow 1 \]

\[ i \leq 100 \]

\[ \text{no} \]

\[ \text{yes} \]

\[ A1 \]

\[ \text{ub}(A) = \]

\[ \text{ub}(i \leftarrow 1) + 100 \times (\text{ub}(i \leq 100) + \text{ub}(A1)) + \text{ub}(i \leq 100) \]
**Assignment**

\[ x \leftarrow a + b \]

\[ ub(x \leftarrow a + b) = \]
\[ \text{cycles} (\text{load } a) + \]
\[ \text{cycles} (\text{load } b) + \]
\[ \text{cycles} (\text{add}) + \]
\[ \text{cycles} (\text{store } x) \]

Assumes constant execution times for instructions

store \( x \)

Not applicable to modern processors!

<table>
<thead>
<tr>
<th>move</th>
<th>cycles</th>
</tr>
</thead>
<tbody>
<tr>
<td>( 1 )</td>
<td></td>
</tr>
</tbody>
</table>
Modern Hardware Features

- Modern processors *increase performance* by using: Caches, Pipelines, Branch Prediction, Speculation

- These features make *WCET computation difficult*: Execution times of instructions vary widely.
  - **Best case** - everything goes smoothly: no cache miss, operands ready, needed resources free, branch correctly predicted.
  - **Worst case** - everything goes wrong: all loads miss the cache, resources needed are occupied, operands are not ready.
  - *Span may be several hundred cycles*.
Access Times

\[ x = a + b; \]

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

**PPC 755**

Execution Time (Clock Cycles)

- **Best Case**
- **Worst Case**

- Clock Cycles
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

- **Micro-architecture Analysis**:  
  - Uses Abstract Interpretation  
  - Excludes as many Timing Accidents as possible  
  - Determines WCET for basic blocks (in contexts)

- **Worst-case Path Determination**  
  - Maps control flow graph to an integer linear program  
  - Determines upper bound and associated path
Overall Structure

Executable program

Control-Flow-Graph
to improve WCET bounds for loops

Static Analyses
Value Analyzer
Cache/Pipeline Analyzer

Micro-architecture Analysis

Micro-architecture Timing Information

Path Analysis
ILP-Generator
LP-Solver
Evaluation

Loop-Bounds

WCET-Visualization

Worst-case Path Determination
Contents

- Introduction
  - problem statement, tool architecture
- Program Path Analysis
- Value Analysis
- Caches
  - must, may analysis
- Pipelines
  - Abstract pipeline models
  - Integrated analyses
what_is_this { 
1     read (a,b);
2     done = FALSE;
3     repeat {
4       if (a>b) 
5          a = a-b;
6       elseif (b>a) 
7          b = b-a;
8       else done = TRUE;
9     } until done;
10    write (a);
}
Program Path Analysis

- **Program Path Analysis**
  - which sequence of instructions is executed in the worst-case (longest runtime)?
  - **problem**: the number of possible program paths grows exponentially with the program length

- **Model**
  - we know the upper bounds (number of cycles) for each basic block from static analysis
  - number of loop iterations must be bounded

- **Concept**
  - transform structure of CFG into a set of (integer) linear equations.
  - solution of the Integer Linear Program (ILP) yields bound on the WCET.
Basic Block

**Definition:** A basic block is a sequence of instructions where the control flow enters at the beginning and exits at the end, without stopping in-between or branching (except at the end).

\[
\begin{align*}
t_1 & := c - d \\
t_2 & := e \times t_1 \\
t_3 & := b \times t_1 \\
t_4 & := t_2 + t_3 \\
\text{if } t_4 & < 10 \text{ goto L}
\end{align*}
\]
Basic Blocks

- *Determine basic blocks of a program:*
  1. *Determine the first instructions of blocks:*
     - the first instruction
     - targets of un/conditional jumps
     - instructions that follow un/conditional jumps
  2. *determine the basic blocks:*
     - there is a basic block for each block beginning
     - the basic block consists of the block beginning and runs until the next block beginning (exclusive) or until the program ends
Control Flow Graph with Basic Blocks

"Degenerated" control flow graph (CFG)
- the nodes are the basic blocks

```
i := 0
L
i := i + 1
if i < 10 goto L
x := t2
```

```
i := 0
t2 := 0
L
  t2 := t2 + i
i := i + 1
if i < 10 goto L
  i >= 10
```
/* k >= 0 */
s = k;
WHILE (k < 10) {
    IF (ok)
        j++;
    ELSE {
        j = 0;
        ok = true;
    }
    k ++;
}
r = j;
Calculation of the WCET

**Definition:** A program consists of $N$ basic blocks, where each basic block $B_i$ has a worst-case execution time $c_i$ and is executed for exactly $x_i$ times. Then, the WCET is given by

$$WCET = \sum_{i=1}^{N} c_i \cdot x_i$$

- the $c_i$ values are determined using the static analysis.
- how to determine $x_i$?
  - structural constraints given by the program structure
  - additional constraints provided by the programmer (bounds for loop counters, etc.; based on knowledge of the program context)
Structural Constraints

Flow equations:

\[
\begin{align*}
    d1 &= d2 = x_1 \\
    d2 + d8 &= d3 + d9 = x_2 \\
    d3 &= d4 + d5 = x_3 \\
    d4 &= d6 = x_4 \\
    d5 &= d7 = x_5 \\
    d6 + d7 &= d8 = x_6 \\
    d9 &= d10 = x_7
\end{align*}
\]
Additional Constraints

- The loop is executed for at most 10 times:
  \[ x_3 \leq 10 \cdot x_1 \]

- B5 is executed for at most one time:
  \[ x_5 \leq 1 \cdot x_1 \]
**WCET - ILP**

*ILP with structural and additional constraints:*

\[ WCET = \max \left\{ \sum_{i=1}^{N} c_i \cdot x_i \mid d_1 = 1 \land \sum_{j \in \text{in}(B_i)} d_j = \sum_{k \in \text{out}(B_i)} d_k = x_i, i = 1 \ldots N \land \text{additional constraints} \right\} \]

program is executed once

structural constraints
Contents

- Introduction
  - problem statement, tool architecture
- Program Path Analysis
- **Value Analysis**
- Caches
  - must, may analysis
- Pipelines
  - Abstract pipeline models
  - Integrated analyses
Overall Structure

Executable program

Control-Flow-Graph

to improve WCET bounds for loops

Micro-architecture Analysis

Static Analyses
- Value Analyzer
- Cache/Pipeline Analyzer

Micro-Architecture

Timing Information

Path Analysis
- ILP-Generator
- LP-Solver
- Evaluation

Worst-case Path Determination

Loop Bounds

WCET Visualization
Abstract Interpretation (AI)

- **Semantics-based method** for static program analysis

- **Basic idea of AI:** Perform the program's computations using value descriptions or *abstract values* in place of the concrete values, start with a description of all possible inputs.

- AI supports *correctness* proofs.
Abstract Interpretation – the Ingredients

- **abstract domain** – related to concrete domain by abstraction and concretization functions, e.g. $L \rightarrow \text{Intervals}$, where $\text{Intervals} = LB \times UB$, $LB = UB = \text{Int} \cup \{-\infty, \infty\}$ instead of $L \rightarrow \text{Int}$

- **abstract transfer functions** for each statement type – abstract versions of their semantics e.g. $+: \text{Intervals} \times \text{Intervals} \rightarrow \text{Intervals}$ where $[a,b] + [c,d] = [a+c, b+d]$ with $+$ extended to $-\infty, \infty$

- **a join function** combining abstract values from different control-flow paths e.g. $\cup: \text{Interval} \times \text{Interval} \rightarrow \text{Interval}$ where $[a,b] \cup [c,d] = [\min(a,c), \max(b,d)]$
Value Analysis

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

**Method:** calculate intervals at all program points, i.e. lower and upper bounds for the set of possible values occurring in the machine program (addresses, register contents, local and global variables).
Value Analysis

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

Which address is accessed here?

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]
Contents

- Introduction
  - problem statement, tool architecture
- Program Path Analysis
- Value Analysis
- Caches
  - must, may analysis
- Pipelines
  - Abstract pipeline models
  - Integrated analyses
Caches: Fast Memory on Chip

- **Caches are used**, because
  - Fast main memory is too expensive
  - The speed gap between CPU and memory is too large and increasing

- Caches work well in the *average case*:
  - Programs access data locally (many hits)
  - Programs reuse items (instructions, data)
  - Access patterns are distributed evenly across the cache
Caches

access takes ~ 1 cycle

access takes ~ 100 cycles

Processor

Cache

Bus

Memory

fast, small, expensive

(relatively)
slow, large, cheap
Caches: How they work

- CPU wants to *read/write at memory address* \( a \), sends a request for \( a \) to the bus.

- **Cases:**
  - Block \( m \) containing \( a \) is in the cache (hit): request for \( a \) is served in the next cycle.
  - Block \( m \) is not in the cache (miss):
    - \( m \) is transferred from main memory to the cache,
    - \( m \) may replace some block in the cache,
    - request for \( a \) is served asap while transfer still continues.

- Several *replacement strategies*: LRU, PLRU, FIFO,... determine which line to replace.
4-Way Set Associative Cache

Address

31 30 29 28 27 ............ 14 13 12 11 10  9 8 7 6 5 4 3 2 1 0

Index
0
1
2

253
254
255

V Tag Data

= = = =

4-to-1 Multiplexor

Hit Data

9 - 39
LRU Strategy

- Each cache set has its own replacement logic => Cache sets are independent. Everything explained in terms of one set

- LRU-Replacement Strategy:
  - Replace the block that has been Least Recently Used
  - Modeled by Ages

- Example: 4-way set associative cache

<table>
<thead>
<tr>
<th>access</th>
<th>age 0</th>
<th>age 1</th>
<th>age 2</th>
<th>age 3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>m₀</td>
<td>m₁</td>
<td>m₂</td>
<td>m₃</td>
</tr>
<tr>
<td>m₄ (miss)</td>
<td>m₄</td>
<td>m₀</td>
<td>m₁</td>
<td>m₂</td>
</tr>
<tr>
<td>m₁ (hit)</td>
<td>m₁</td>
<td>m₄</td>
<td>m₀</td>
<td>m₂</td>
</tr>
<tr>
<td>m₅ (miss)</td>
<td>m₅</td>
<td>m₁</td>
<td>m₄</td>
<td>m₀</td>
</tr>
</tbody>
</table>
Deriving a Cache Analysis

- **Reducing** the semantics (to what concerns caches)
  - e.g. from values to locations,
  - ignoring arithmetic.
  - obtain “auxiliary/instrumented” semantics

- **Abstraction**
  - Changing the domain: sets of memory blocks in single cache lines
Cache Analysis

How to statically precompute cache contents:

- **Must Analysis:**
  For each program point (and calling context), find out which blocks are in the cache. Determines safe information about cache hits. Each predicted cache hit reduces WCET.

- **May Analysis:**
  For each program point (and calling context), find out which blocks may be in the cache. Complement says what is not in the cache.
  
  Determines safe information about cache misses. Each predicted cache miss increases BCET.
Abstract Domain: Must Cache

Abstraction

\[ \alpha \]

\[
\begin{array}{c}
\{ \}\\
\{ \}\\
\{z, x\}\\
\{s\}
\end{array}
\]
Abstract Domain: Must Cache

\[ z, x \in \{ s \in \{ \} \} \]

Concretization

\[ \gamma \]

\[
\begin{align*}
\text{ } & \quad \{} \\
\text{ } & \quad \{} \\
\text{ } & \quad \{z, x\} \\
\text{ } & \quad \{s\}
\end{align*}
\]
Cache with LRU: Transfer for must

concrete
[ access s ]

abstract
[ access s ]

Age
“young”
“old”
Cache Analysis: Join (must)

Join (must)

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

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

“intersection + maximal age”

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

**Interpretation:**
memory block a is definitively in the (concrete) cache => always hit
Abstract Domain: May Cache
Abstract Domain: May Cache

Concretization

\[
\begin{align*}
\gamma & \colon \\
\{m, n, o, p\} & \to \{\{z, s, x\}, \{t\}, \{\}, \{a\}\}
\end{align*}
\]
Cache with LRU: Transfer for may

concrete
[access s]

<table>
<thead>
<tr>
<th></th>
<th>x</th>
<th>t</th>
<th>z</th>
<th>y</th>
</tr>
</thead>
</table>

|   | s  | x  | t  | z  |

“young”
Age
“old”

abstract
[access s]

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

|   | {s}  | {x, t} | {y, z} | {}  |
**Cache Analysis: Join (may)**

*Join (may)*

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

"union + minimal age"

**Interpretation:**

all blocks may be in the cache; none is definitely not in the cache.
Contribution to WCET

Information about cache contents sharpens timings.

ref to s

\[ t_{\text{miss}} \rightarrow t_{\text{hit}} \]

if s is in must-cache:
\[ t_{\text{WCET}} = t_{\text{hit}} \]
otherwise
\[ t_{\text{WCET}} = t_{\text{miss}} \]

if s is in may-cache:
\[ t_{\text{BCET}} = t_{\text{hit}} \]
otherwise
\[ t_{\text{BCET}} = t_{\text{miss}} \]
Contribution to WCET

- Information about cache contents sharpens timings.

```latex
while \ldots \ do [\text{max } n] \\
  \ldots \\
  \text{ref to } s \\
  \ldots \\
  \text{od}
```

**within loop**

\[ n \times \text{t}_{\text{miss}} \]
\[ n \times \text{t}_{\text{hit}} \]
\[ \text{t}_{\text{miss}} + (n - 1) \times \text{t}_{\text{hit}} \]
\[ \text{t}_{\text{hit}} + (n - 1) \times \text{t}_{\text{miss}} \]
\ldots
Contexts

- Cache contents depends on the context, i.e. calls and loops

- First iteration loads the cache:
  - Intersection loses most of the information.

- Distinguish as many contexts as useful:
  - 1 unrolling for caches
  - 1 unrolling for branch prediction (pipeline)

while cond do

join (must)
Contents

► Introduction
  ▪ problem statement, tool architecture

► Program Path Analysis

► Value Analysis

► Caches
  ▪ must, may analysis

► Pipelines
  ▪ Abstract pipeline models
  ▪ Integrated analyses
Comparison of Architectures

- Single cycle
- Multiple cycle
- Pipelining

0 - LW
1 - SW

Einzyklenverarb.
Mehrzyklenverarb.
Pipelineverarb.

single cycle

T1
LW
SW
T2

T1 T2 T3 T4 T5 T6 T7 T8 T9

multiple cycle

IF RF EX MEM WB IF RF EX MEM

pipelining

IF RF EX MEM WB LW
IF RF EX MEM WB SW
Hardware Features: Pipelines

Ideal Case: 1 Instruction per Cycle
Datapath of a Pipeline Architecture
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.*
Pipeline Hazards

Pipeline Hazards:

- **Data Hazards**: Operands not yet available (Data Dependences)

- **Resource Hazards**: Consecutive instructions use same resource

- **Control Hazards**: Conditional branch

- **Instruction-Cache Hazards**: Instruction fetch causes cache miss
Control Hazard

Program execution order (in instructions)

40 beq $1, $3, 28
44 and $12, $2, $5
48 or $13, $6, $2
52 add $14, $2, $2
72 lw $4, 50($7)
## Data Hazard

### Time (in clock cycles)

<table>
<thead>
<tr>
<th>Value of register $2$:</th>
<th>CC 1</th>
<th>CC 2</th>
<th>CC 3</th>
<th>CC 4</th>
<th>CC 5</th>
<th>CC 6</th>
<th>CC 7</th>
<th>CC 8</th>
<th>CC 9</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10/-20</td>
<td>-20</td>
<td>-20</td>
<td>-20</td>
<td>-20</td>
</tr>
</tbody>
</table>

### Program execution order (in instructions)

1. `sub $2, $1, $3`
2. `and $12, $2, $5`
3. `or $13, $6, $2`
4. `add $14, $2, $2`
5. `sw $15, 100($2)`
Static analysis of hazards

**Cache analysis**: prediction of cache hits on instruction or operand fetch or store

```
lwz r4, 20(r1)  Hit
```

**Dependence analysis**: analysis of data/control hazards

```
add r4, r5, r6
lwz r7, 10(r1)
add r8, r4, r4  Operand ready
```

**Resource reservation tables**: analysis of resource hazards

<table>
<thead>
<tr>
<th>IF</th>
<th>EX</th>
<th>M</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

ETH Swiss Federal Institute of Technology

Computer Engineering and Networks Laboratory
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
An Abstract Pipeline for a Basic Block

function \texttt{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 \( t \)
- \( \text{length}(t) \) gives number of cycles

\textbf{What is different?}

- Abstract states may lack information, e.g. about cache contents.
- Assume local worst cases is safe (in the case of no timing anomalies)
- Traces may be longer (but never shorter).
What is different?

- **Starting state** for successor basic block? In particular, if there are several predecessor blocks?

- **Alternatives:**
  - sets of states
  - combine by assuming that local worst case is safe

![Diagram](image)
Summary of Steps

- **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**

- **Input:** an executable program, starting points, loop iteration counts, call targets of indirect function calls, and a description of bus and memory speeds
- **Output:** computes **Worst-Case Execution Time** bounds of tasks