Host Compiled Simulation for Timing and Power Estimation

Gaurav Kukreja
Bo Wang

Intel Mobile Communications

October 29, 2014

Overview

Simulation: Popular Techniques

Cycle Accurate Simulation

- Detailed simulation of processor micro-architecture.
- Cycle Accurate estimation of performance.
- Difficult to develop, Very Slow execution.

Functional Simulation

- High Level of Abstraction.
- Simple to develop, and fast execution.
- Focus is Functional Verification. Cannot be used for performance estimation.

Our Focus

A technique that is,

- Easy to Develop.
- Fast to Execute.
- Accurate.

Host Compiled Simulation

Host Compiled Simulation

- Based on Source Code Instrumentation.
- Instrumented code compiled and run on Host Machine, hence the name.
- Easy to understand, develop and maintain.
- Fast Execution, and accurate results.

Simple Example

```
1   int sum(int array[20])
2   {
3       int i;
4       int sum = 0;
5
6       for (i=0; i<20; i++)
7           sum += array[i];
8
9       return sum;
10  }
```

Simple C Code

```
1   00008068 <sum>:
2   8068:    mov     r3, #0
3   806c:    mov     r2, r3
4   8070:    ldr     r1, [r0, r3]
5   8074:    add     r2, r2, r1
6   8078:    add     r3, r3, #4
7   807c:    cmp     r3, #80 ; 0x50
8   8080:    bne     8070 <sum+0x8>
9   8084:    mov     r0, r2
10  8088:    bx      lr
```

Objdump Code

Simple Example: Mapping Basic Blocks

```
1   int sum(int array[20])
2   {
3       int i;
4       int sum = 0;
5
6       for (i=0; i<20; i++)
7           sum += array[i];
8
9       return sum;
10  }
```

Simple C Code

```
1   00008068 <sum>:
2   8068:    mov     r3, #0
3   806c:    mov     r2, r3
4   8070:    ldr     r1, [r0, r3]
5   8074:    add     r2, r2, r1
6   8078:    add     r3, r3, #4
7   807c:    cmp     r3, #80 ; 0x50
8   8080:    bne     8070 <sum+0x8>
9   8084:    mov     r0, r2
10  8088:    bx      lr
```

Objdump Code

Basic Block in Binary		Matching block in Source	
BlockID	Lines	BlockID	Lines
1	2-3	1	3-4
2	4-8	2	6-7
3	9-10	3	9

Instrumented Code

```
1   unsigned int execCycles;
2   unsigned int memAccessCycles;
3
4   int sum(int array[20])
5   {
6       int i;
7       int sum = 0;
8       execCycles += 2;
9       memAccessCycles += simICache(0x8068, 8);
10
11      for (i=0; i<20; i++)
12      {
13          sum += array[i];
14          memAccessCycles += simDCache(array + i, READ);
15          execCycles += 5;
16          memAccessCycles += simICache(0x8070, 40);
17      }
18
19      execCycles += 2;
20      memAccessCycles += simICache(0x8084, 8);
21      return sum;
22  }
```

```
1   int sum(int array[20])
2   {
3       int i;
4       int sum = 0;
5
6       for (i=0; i<20; i++)
7           sum += array[i];
8
9       return sum;
10  }
```

```
1   00008068 <sum>:
2   8068:    mov     r3, #0
3   806c:    mov     r2, r3
4   8070:    ldr     r1, [r0, r3]
5   8074:    add     r2, r2, r1
6   8078:    add     r3, r3, #4
7   807c:    cmp     r3, #80 ; 0x50
8   8080:    bne     8070 <sum+0x8>
9   8084:    mov     r0, r2
10  8088:    bx      lr
```

Objective: Develop a tool for Automatic Instrumentation

Steps:

- Extract Mapping.
- Instrumentation for cycles spent in Execution.
- Instrumentation for cycles spent in Memory Access.

Output:

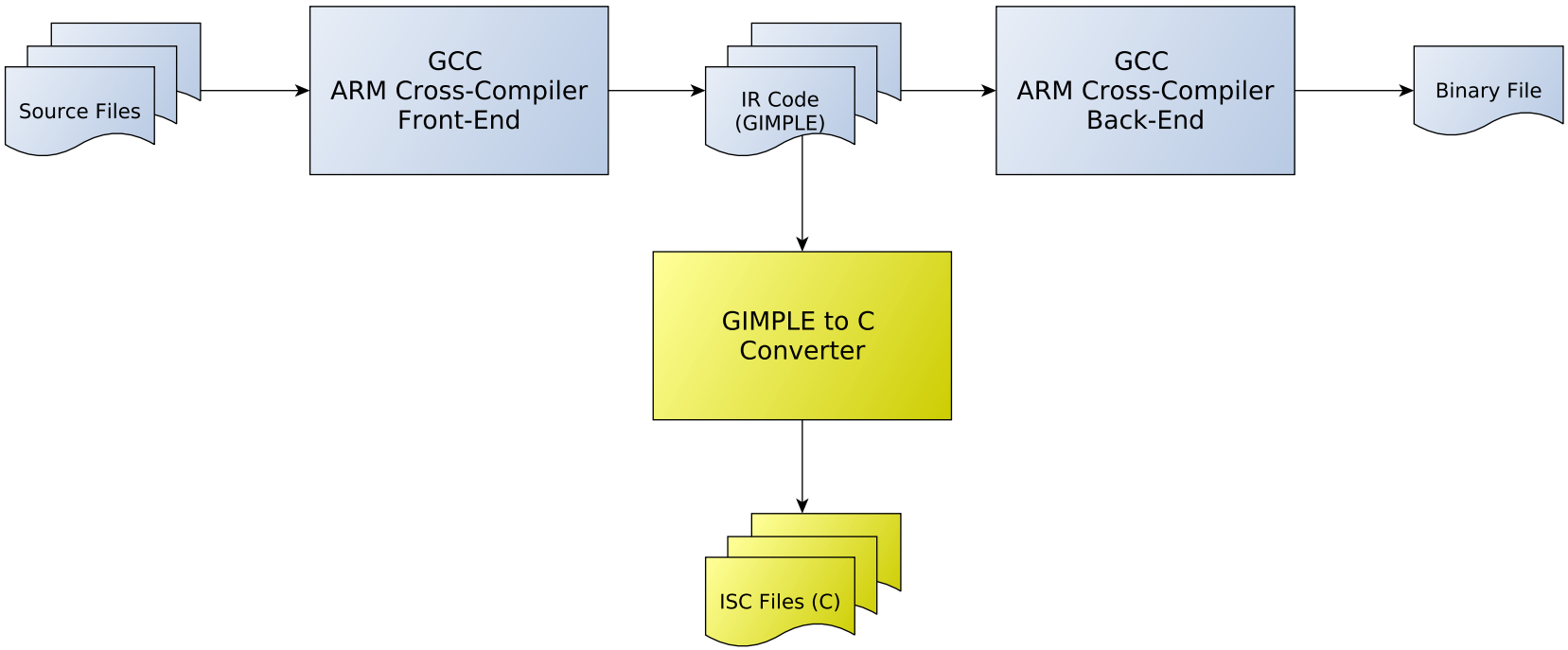
- Instrumented Code.

Extract Mapping between Source Code and Binary Code

Mapping between Source and Binary

- Very important for accurate instrumentation.
- Compiler destroys mapping during optimization phases.
- GDB provides mapping, but ambiguous.
- Control Flow Analysis to generate mapping between Basic Blocks.

Conversion of IR Code to C Code



* Code for GIMPLE to C converter reused from RBA project [1]

Mapping Algorithm

- Extract CFG from ISC and Binary Code.
- Graph Matching Algorithm using Depth First Traversal.
- Special handling for each optimization.
- GDB debug information in corner cases.

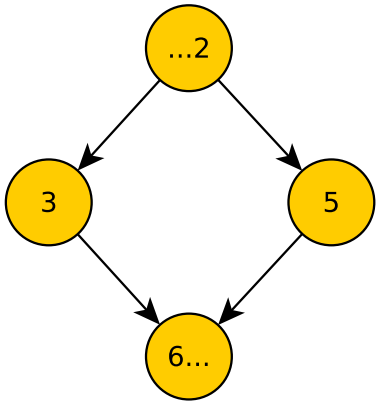
Handling of Compiler Optimization : Conditional Execution

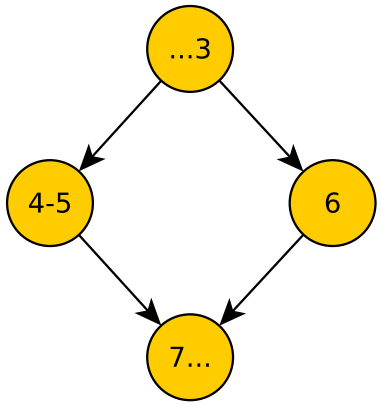
```
1  ...
2  if (a > b)
3      max = a;
4  else
5      max = b;
6  ...
```

Simple C Code

```
1  806c:    ...
2  8070:    cmp     r1, r2
3  8074:    ble     8080
4  8078:    mov     r3, r1
5  807c:    b       8084
6  8080:    mov     r3, r2
7  8084:    ...
```

Unoptimized Assembly Code





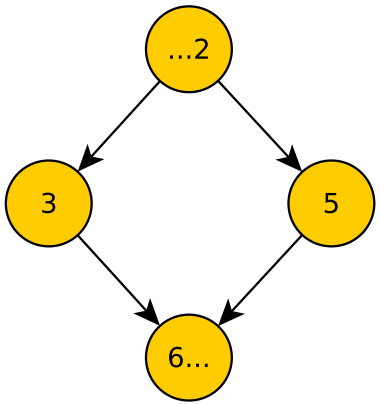
Handling of Compiler Optimization : Conditional Execution

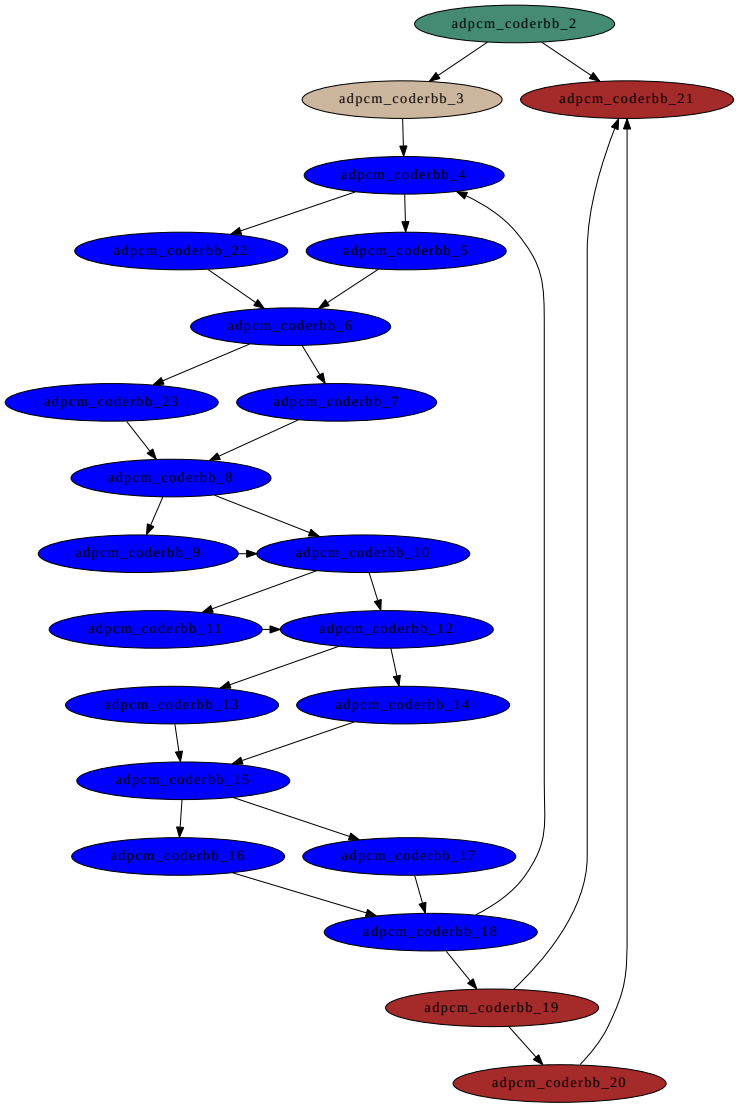
```
1  ...
2  if (a > b)
3      max = a;
4  else
5      max = b;
6  ...
```

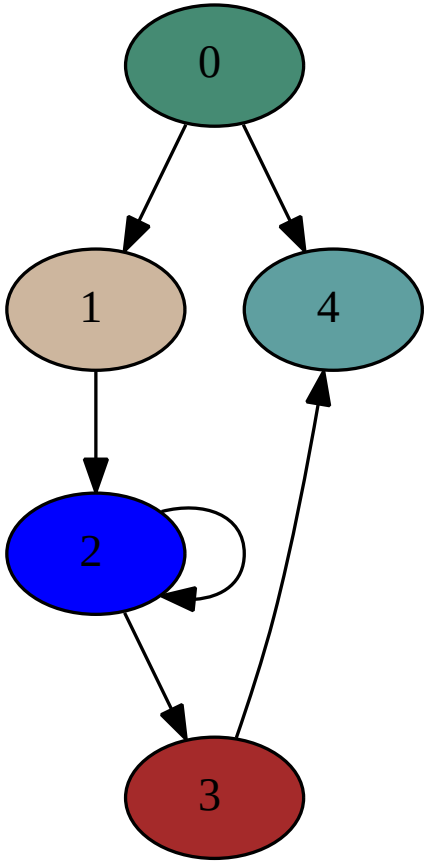
Simple C Code

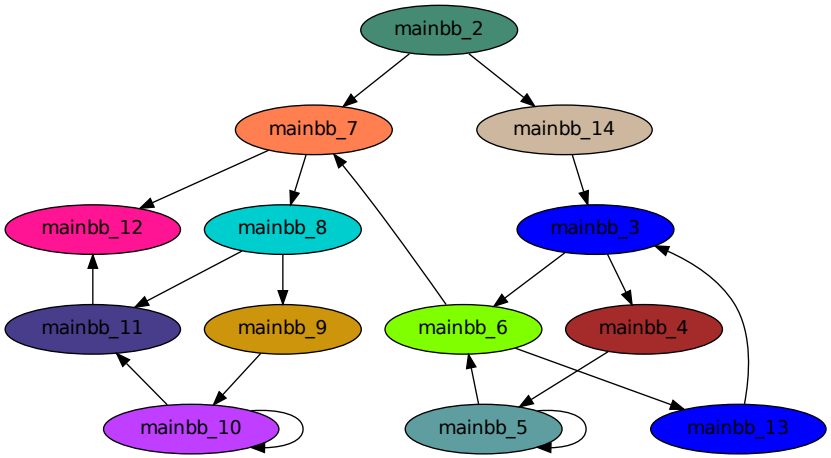
```
1  806c:    ...
2  8070:    cmp     r1, r2
3  8074:    movgt   r3, r1
4  8078:    movle   r3, r2
5  807c:    ...
```

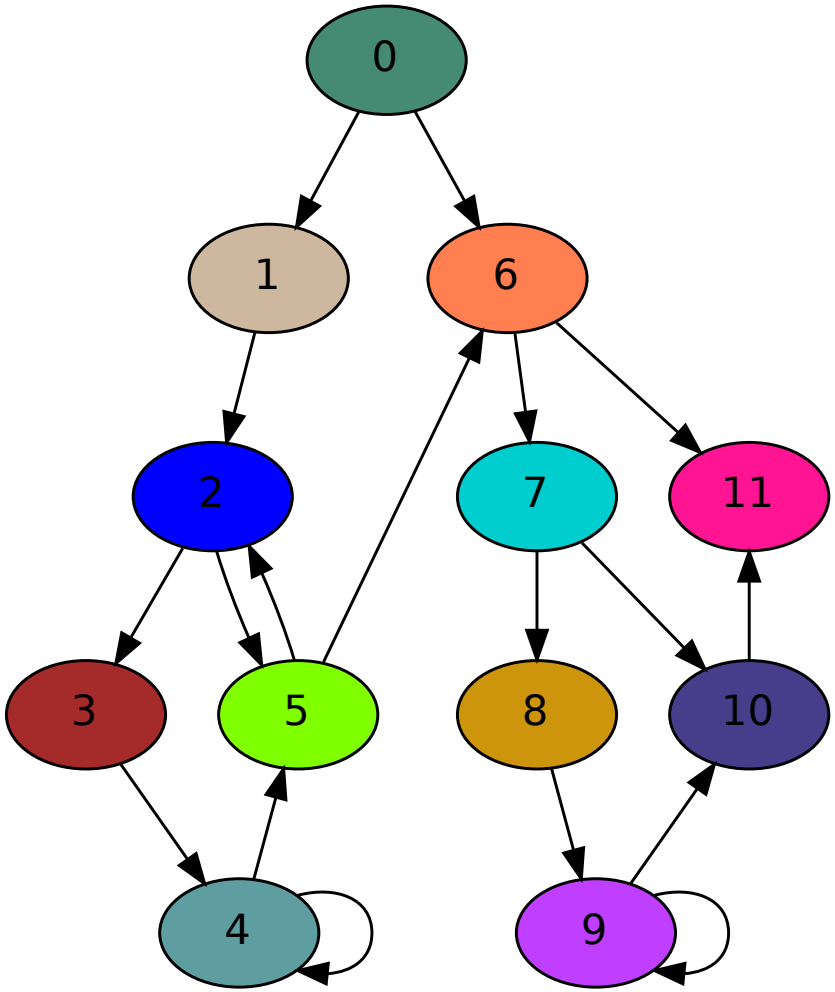
Optimized Assembly Code




adpcm_coderbb_2
adpcm_coderbb_3
adpcm_coderbb_21
adpcm_coderbb_4
adpcm_coderbb_22
adpcm_coderbb_5
adpcm_coderbb_6
adpcm_coderbb_23
adpcm_coderbb_7
adpcm_coderbb_8
adpcm_coderbb_9
adpcm_coderbb_10
adpcm_coderbb_11
adpcm_coderbb_12
adpcm_coderbb_13
adpcm_coderbb_14
adpcm_coderbb_15
adpcm_coderbb_16
adpcm_coderbb_17
adpcm_coderbb_18
adpcm_coderbb_19
adpcm_coderbb_20


0
1
4
2
3


mainbb_2
mainbb_7
mainbb_14
mainbb_12
mainbb_8
mainbb_3
mainbb_11
mainbb_9
mainbb_6
mainbb_4
mainbb_10
mainbb_5
mainbb_13


0
1
6
2
7
11
3
5
8
10
4
9

Annotation for Cycles Spent in Execution

Annotation for Cycles spent in Pipeline

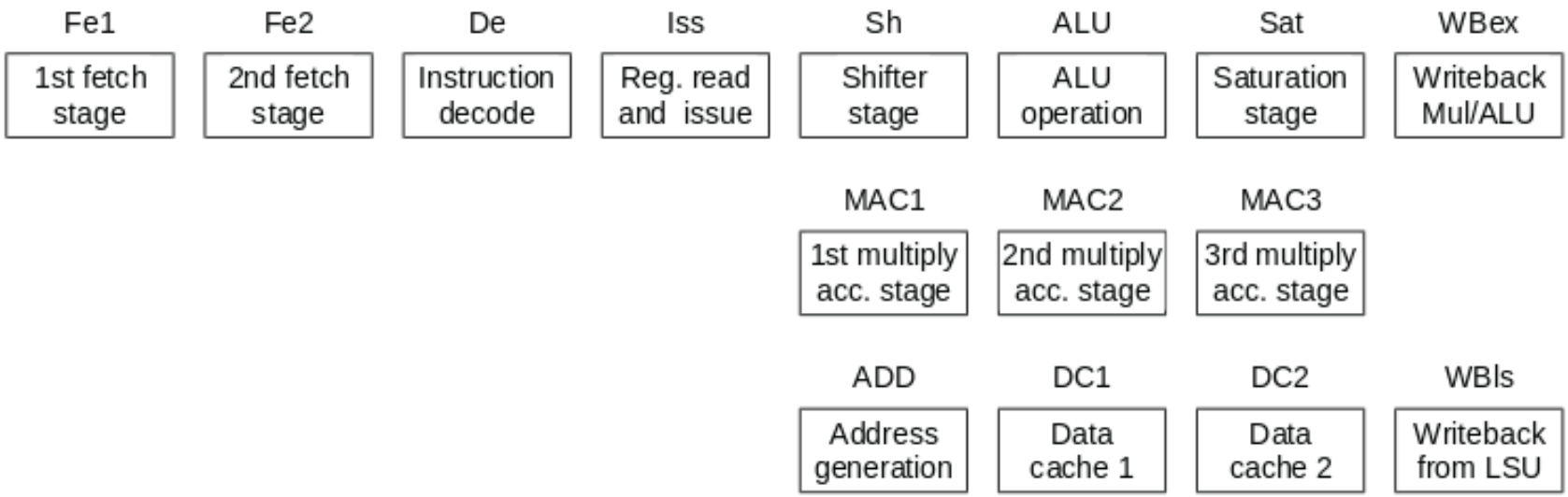
Performed at Basic Block Granularity.

Important to consider

- Pipeline Architecture of the target processor.
- Data and Control Hazards, that leads to pipeline stalls.
- Branch Prediction, that prevents pipeline flushes.

Pipeline architecture of ARM Cortex A5

The ARM Cortex A5 has an 8-stage pipeline.



Lets assume,

- Pipeline is empty at start of Basic Block.
- All data is available without latency.

Effects due to Data and Control Hazards

For each Basic Block from Binary Code,

- Parse instructions sequentially.
- Simulate progress in pipeline.
- Identify interlocking (hazards) and add penalty.
- Annotate to mapped block in Source Code.

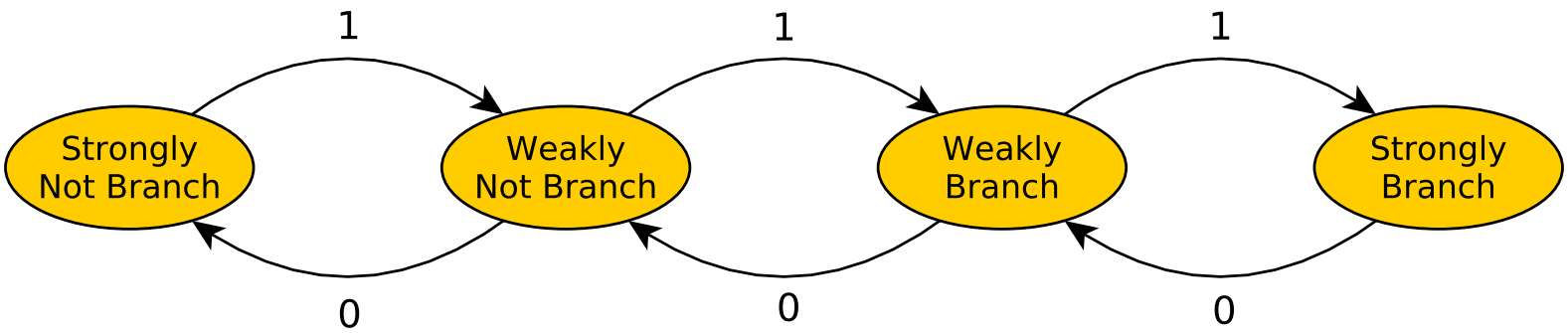
```
...
for(i<0; i<10; i++) {}
    execCycles += 23;
    sum += array[i];
    ...
}
...
```

Branch Prediction

- Reduces number of pipeline flushes.
- Major impact on performance.
- Branch Prediction Unit is simulated to account for this.

Branch Prediction Algorithm

- Simulates the BPU on ARM Cortex A5
- 125 entry Branch History Table.
- State Machine for each Entry. 2 bit state information.



Annotation for Branch Prediction

Branch Prediction Simulator offers following API.

```
unsigned int branchPred_enter{unsigned long startAdd,
                              unsigned long endAdd};
```

```
1  ...
2  for(i=0; i<10; i++) {
3      execCyles += 23;
4      execCycles -= (branchPred_enter(0x348, 0x380) ? 7 : 0);
5      sum += array[i];
6      ...
7  }
8  ...
```

Annotation for Cycles spent in Memory Access

Cache Simulator

Cache Hierarchy on target device.

	Size	N-way	Cache Line Size
L1 D Cache	32 KB	4	32 B
L1 I Cache	32 KB	2	32 B
L2 Cache	256 KB	16	32 B

- Pseudo Random Replacement Policy
- Data Prefetching

Instruction Cache Simulation

Cache Simulator offers API for Instruction Access.

```
unsigned long long simICache(unsigned long address,
                             unsigned long size);
```

```
1  ...
2  for(i=0; i<10; i++) {
3      execCyles += 23;
4      execCycles -= (branchPred_enter(0x348, 0x380) ? 7 : 0);
5      memAccessCycles += simICache(0x348, 56);
6      sum += array[i];
7      ...
8  }
9  ...
```

Data Cache Simulation

Address from host cannot be used for simulation.

Will lead to inaccuracies, because

- Different sizes of Basic Data Types.
- Different Memory Alignment on Host and Target.

Memory Access Reconstruction

- Inspired by [2]
- Resolve address of each load/store instruction on target device.

Memory Access Reconstruction

- Resolve Physical Address of each variable in target memory.
- Identify variable being accessed by each load/store.
- Identify index information from source code.
- Simulate Data Cache.

Resolve Physical Address of Variables

Global Variables

- Bare-Metal Execution. Address fixed at compile time.
- Extracted by Static Analysis of binary.

Annotate address in code as follows,

```
1  unsigned int globVar[100];
2  unsigned long globVar_addr = 0x7c8;
```

Resolve Physical Address of Variables

Local Variable

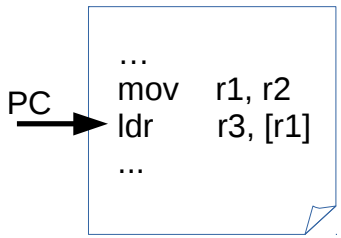
- Stored in Stack, address only known at run-time.
- Address relative to stack, can be known statically.
- Simulate growth of stack at run-time to resolve physical address.

```
1  unsigned long CSIM_SP = 0x1d4c8;          // Initial Value of SP.
2
3  int foo()
4  {
5      CSIM_SP += 88;                        // Size of Stack Frame for foo.
6
7      int localVar;
8      unsigned long localVar_addr = 0x8;    // Address relative to SP.
9      ...
10 }
```

Identify Load/Store Operations in Binary Code

To identify variable accessed by a load/store instruction,

- Binary code is functionally simulated.
- Register state is maintained, and updated as per instruction.
- Branch instructions are ignored.
- Address for each load/store is extracted.
- Variable being accessed is known.



Assembly Code

r0	r1	r2	...	r15
0	7c8	7c8	...	346

Register State

Annotation of Memory Access

To find the line in source code, which causes the memory access operation,

- Each line is parsed using a custom C Parser.
- Index information is extracted.
- Annotation as follows.

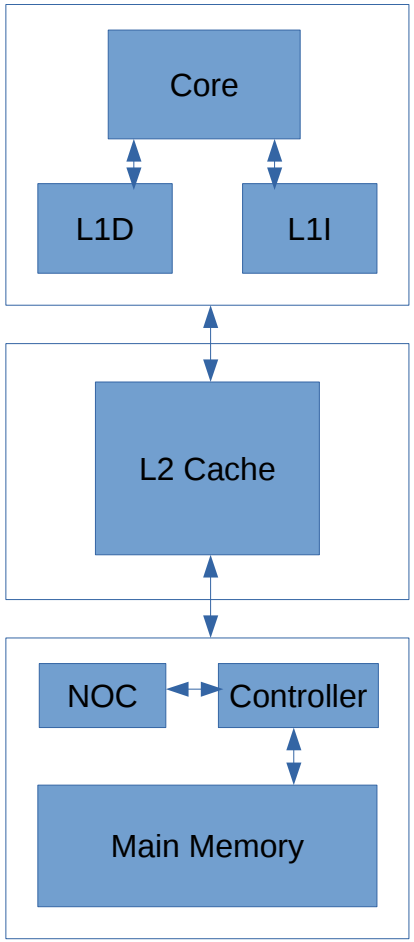
```
1  ...
2  for (i=0; i < 100; i++) {
3      sum += globalVar[i];
4      memAccessCycles += simDCache ( globalVar_addr + i * 4, True );
5  }
6  localVar = sum / 100;
7  memAccessCycles += simDCache ( SP + localVar_addr, False );
8  ...
```

Power Estimation

Power State Model

- Power consumed by each component in active and idle state is known.
- Trace information for each basic block, shows components that were active and the duration.
- Total power being consumed at a time, can be estimated.

Components for Power Estimation



Target System is divided into components as follows

- Core and L1 Caches
- L2 Cache
- External Memory along with controller and NOC.

Limitaitons

- Instrumentation may not succeed in corner cases. Appropriate hints provided in source code as comments to manually perform instrumentation.
- Compiler Optimization Level "O1" only, owing to complexity of Mapping Problem.

Test Setup

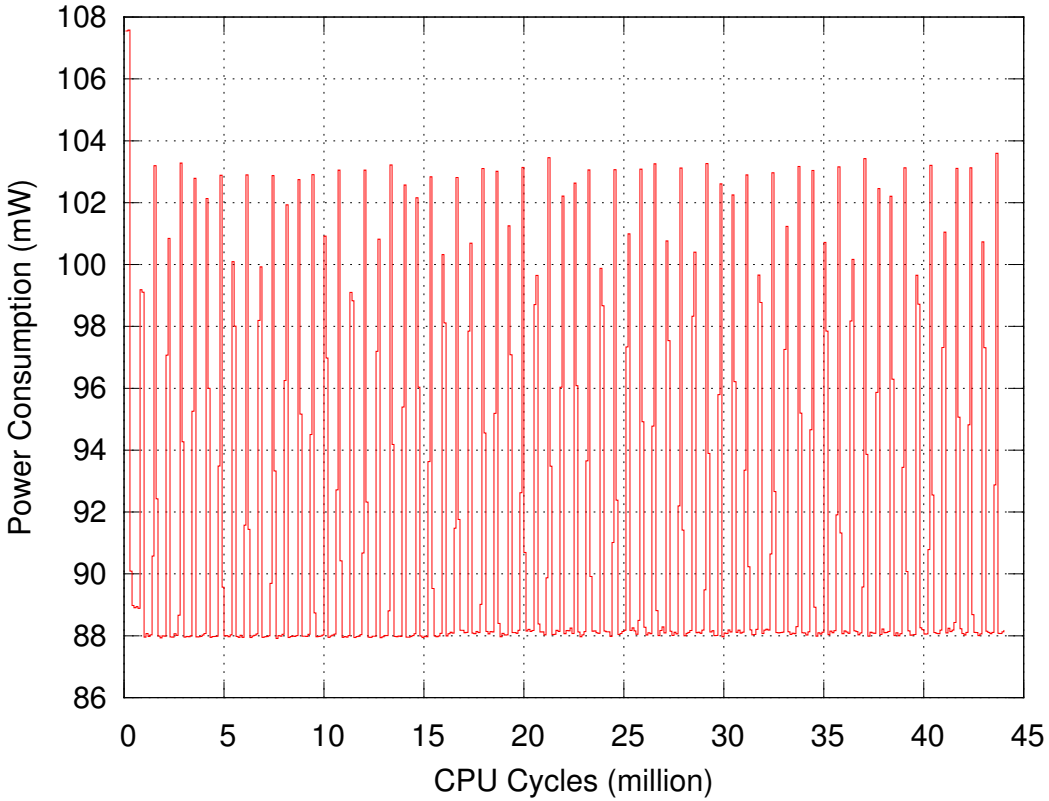
For testing, ARM Cortex A5 core was used. Results from ARM Performance Measurement Unit were compared.

- Number of Cache Misses.
- Total Cycles.

Accurate Prediction of Cache Misses, will verify the correctness of the instrumentation.

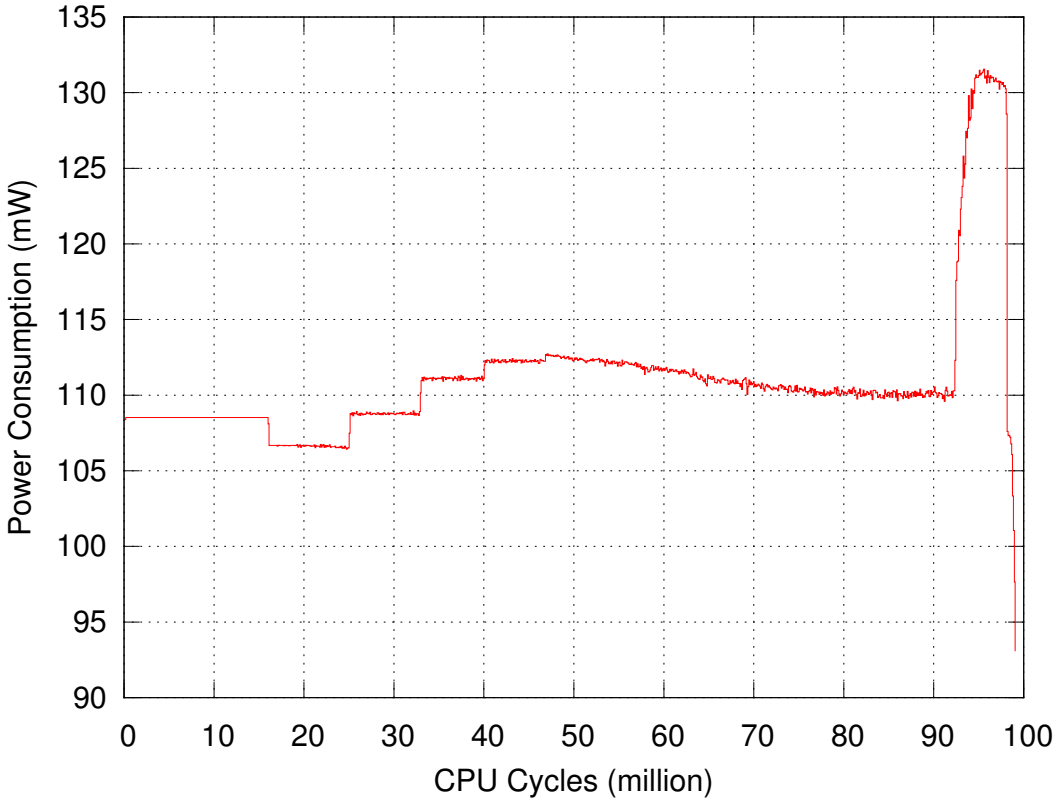
ADPCM Benchmark

	HCS	Actual	Accuracy
Total Cache Miss	91452	91604	99.99%
Total Cycles	46110394	45594862	99.98%



Sieve Benchmark

	HCS	Actual	Accuracy
Total Cache Miss	933109	934250	99.88%
Total Cycles	99068687	100403731	98.67%




S. Chakravarty, Zhuoran Zhao, and A. Gerstlauer.
Automated, retargetable back-annotation for host compiled performance and power modeling.
In *Hardware/Software Codesign and System Synthesis (CODES+ISSS), 2013 International Conference on*, pages 1–10, Sept 2013.


Kun Lu, D. Muller-Gritschneder, and U. Schlichtmann.
Memory access reconstruction based on memory allocation mechanism for source-level simulation of embedded software.
In *Design Automation Conference (ASP-DAC), 2013 18th Asia and South Pacific*, pages 729–734, Jan 2013.