Moving beyond Transaction-Level Modeling

*Sandro Penolazzi (sandrop@kth.se)
Ahmed Hemani (hemani@kth.se)
Outline

- Motivations and goal
- Methodology description
  - A - IP Level
  - B - Algorithmic Level
  - C - Refinement Level
- Methodology validation
  - Accuracy against gate level
  - Speed against TLM PV
- Conclusions and future work
Motivations and goal

Predict performance & energy early in the design cycle is crucial to be able to take the best architectural decisions.

A high-level methodology is needed for fast and accurate system-level estimation and exploration.

FUNTIME
( Functional UNTIMEEd)
What does “high-level methodology” mean?

Model at a few levels that target the “pain” and risk in your D&V flow
Main claim
Implement a methodology for rapid and accurate:

- system estimation (energy & performance)
- system exploration at Functional Level

Challenge
Relate the Functional Level to the actual architecture

Novelty and advantages
Speed: no architectural-level simulation (RTL,TLM). All information inferred from functional level → faster
Accuracy: detailed pre-characterization of IPs allows high accuracy
Low engineering effort: no architectural model required
The Funtime Methodology
A 3-layer approach

C – Refinement Level

B – Algorithmic Level

A – IP Level
Goal: characterize each IP transaction for energy and performance

A transaction expresses the main functionalities of an IP (ALT = Architectural Level Transaction)

Why?
- Promoting IP reuse is a common practice
- Sensible to model each IP transaction for energy and cycles
  - One-time effort done by IP vendor
  - IP energy/performance model can be used from higher levels of abstraction to do estimation

Accurate IP characterization is crucial for overall Funtime accuracy
- Characterization is done on layout
A 3-layer approach: IP Level

[ALT(energy, cycles)_1, ..., ALT(energy, cycles)_N]
B – Algorithmic Level: Overview

- **Goal:** infer the number of transactions
- **How?**
  - NOT by simulating an architecture model (TLM, RTL)
  - By instrumenting the algorithmic model of the application to emit target transactions (Primary ALTs) while running on a common PC
- **Advantages**
  - Algorithmic development is anyhow a necessary step
  - Reduced engineering effort: no time for coding hardware simulation models
  - Very fast as it runs natively
B – Algorithmic Level

Example of Instrumentation

Instrumented source code (runs on a PC)

```c
:- 19:int fibonacci(int i)
57313: 20:{
57313: 21: if (i < 0) return -1; /* F(i) non e' definit
57313: 21:block 0
57313: 22:
57313: 23: if (i == 0) return 0;
57313: 23:block 0
57313: 24:block 1
46367: 24: else if (i == 1) return 1;
46367: 24:block 0
17711: 25:block 1
57313: 25: else return fibonacci(i-1) + fibonacci(i-2);
57313: 26:block 0
57313: 27:block 1
:- 27;
:-
```

Output

<table>
<thead>
<tr>
<th>Total report = 1517196</th>
</tr>
</thead>
<tbody>
<tr>
<td>save 57314</td>
</tr>
<tr>
<td>st 189650</td>
</tr>
<tr>
<td>ld 218306</td>
</tr>
<tr>
<td>cmp 190993</td>
</tr>
<tr>
<td>bge 57313</td>
</tr>
<tr>
<td>bne 321986</td>
</tr>
<tr>
<td>mov 150993</td>
</tr>
<tr>
<td>b 193680</td>
</tr>
<tr>
<td>bne 193680</td>
</tr>
<tr>
<td>add 85968</td>
</tr>
<tr>
<td>call 57313</td>
</tr>
</tbody>
</table>

Mixed source-assembly for target

```c
20:fibonacci.c **** {
127 .stbn 68,0,20,.LLM1.fibonacci
128 .LLM1:
129 .PROLOGUE# 0
130 0000 9DE3BF90 save %sp, -112, %sp
131 .PROLOGUE# 1
132 0004 F027A044 st %0, [%fp+68]
21:fibonacci.c **** if (i < 0) return -1; /* F(i) non e'
133 .stbn 68,0,21,.LLM2.fibonacci
134 .LLM2:
135 0008 C207A044 ld [%fp+68], %g1
136 000c 80A00000 cmp %g1, 0
137 0010 16800006 bge .LL2
138 0014 01000000 nop
139 .stbn 68,0,21,.LLM3.fibonacci
140 .LLM3:
141 0018 82103FFF mov -1, %g1
142 001c C227EFF4 st %g1, [%fp-12]
143 0020 1080000F b .LL1
144 0024 01000000 nop
145 .LL2:
22:fibonacci.c ****
23:fibonacci.c **** if (i == 0) return 0;
146 .stbn 68,0,23,.LLM4.fibonacci
147 .LLM4:
148 0028 C207A044 ld [%fp+68], %g1
```

ICT - Internal Workshop
B – Algorithmic Level: limitations

- Library functions
  - Calls to functions that are not implemented in any of the C files that we are profiling are excluded from the count (ex. printf, malloc, sin, pow, etc.)

- Order of execution
  - GCOV instrumentation only considers the final total source lines count
  - The order of execution of the C source lines is ignored
  - Only a partial order of execution can be inferred with the granularity of a basic block
B – Algorithmic Level: Block Diagram

AS = Architecture Specification
UCS = Use-Case Specification
C – Refinement Level: Overview

- Energy and performance figures from Level A and B are likely to be inaccurate
  1. Interdependency among ALTs (investigated)
  2. Caches (investigated)
  3. Bus arbitration (partially investigated)
  4. Presence of OS (investigated – in progress)
  5. Power management (not yet investigated)

- [1-5] can be factored in as a trace of extra ALTs (Secondary ALTs) which refine the original trace inferred at Level B
C - Refinement Level: Block Diagram

In

C - Refinement Level
- Refinements
  - ALTs interdependency
  - Hw/sw optimizations (caches, etc.)
  - Operating System
  - Bus arbitration

# Secondary ALTs (S-ALTs)

UCS

B - Algorithmic Level
- Instrumentation
  ![Instrumentation](Instrumentation.png)
- Instrumented Algorithmic Models
  ![Instrumented Algorithmic Models](InstrumentedAlgorithmicModels.png)

# Primary ALTs (P-ALTs)

AS

A - IP Level
- IP Energy & Performance Models
  ![IP Energy & Performance Models](IPModels.png)

[ALT(energy, cycles)_1, ..., ALT(energy, cycles)_N]

# ALTs

Out

Exec. time

Energy

10/10/09
Reference SoC Platform

- Leon3
- SRAM
- AHB/APB Bridge
- Timer
- Irq Ctrl
- AHB Ctrl
- APB Ctrl
Funtime Validation: Accuracy Interdependency Refinement

- Gate Level estimation is our reference
- 1 source of error
  - Leon3 IP model (no cache)
- Nr. of instructions and cycles is measured

<table>
<thead>
<tr>
<th></th>
<th>FFT</th>
<th>Quicksort</th>
<th>Fibonacci</th>
<th>Viterbi</th>
<th>Queens</th>
<th>Dhrystone</th>
<th>MD5</th>
</tr>
</thead>
<tbody>
<tr>
<td># Instructions</td>
<td>1371584</td>
<td>4008651</td>
<td>1585313</td>
<td>2452903</td>
<td>4266783</td>
<td>1950813</td>
<td>1802380</td>
</tr>
<tr>
<td># Cycles</td>
<td>6485094</td>
<td>31649232</td>
<td>13902019</td>
<td>17511989</td>
<td>24193885</td>
<td>15938144</td>
<td>14428588</td>
</tr>
<tr>
<td>Measured energy (mJ)</td>
<td>0.2396</td>
<td>1.1508</td>
<td>0.4906</td>
<td>0.6250</td>
<td>0.8571</td>
<td>0.5641</td>
<td>0.5228</td>
</tr>
<tr>
<td>1-instr-based model</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Estimated instructions energy (mJ)</td>
<td>0.1388</td>
<td>0.5209</td>
<td>0.1917</td>
<td>0.3009</td>
<td>0.4389</td>
<td>0.2581</td>
<td>0.2192</td>
</tr>
<tr>
<td>Estimated stalls energy (mJ)</td>
<td>0.1319</td>
<td>0.6766</td>
<td>0.3218</td>
<td>0.3871</td>
<td>0.5166</td>
<td>0.3529</td>
<td>0.3302</td>
</tr>
<tr>
<td>Estimated total energy (mJ)</td>
<td>0.2706</td>
<td>1.1975</td>
<td>0.5135</td>
<td>0.6880</td>
<td>0.9555</td>
<td>0.6110</td>
<td>0.5494</td>
</tr>
<tr>
<td>Error (%)</td>
<td>+12.96</td>
<td>+4.05</td>
<td>+4.67</td>
<td>+10.08</td>
<td>+11.48</td>
<td>+8.31</td>
<td>+4.90</td>
</tr>
<tr>
<td>2-instr-based model</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Estimated instructions energy (mJ)</td>
<td>0.1187</td>
<td>0.4116</td>
<td>0.1504</td>
<td>0.2458</td>
<td>0.3622</td>
<td>0.2060</td>
<td>0.1809</td>
</tr>
<tr>
<td>Estimated stalls energy (mJ)</td>
<td>0.1319</td>
<td>0.6766</td>
<td>0.3218</td>
<td>0.3871</td>
<td>0.5166</td>
<td>0.3529</td>
<td>0.3302</td>
</tr>
<tr>
<td>Estimated total energy (mJ)</td>
<td>0.2506</td>
<td>1.0882</td>
<td>0.4722</td>
<td>0.6329</td>
<td>0.8788</td>
<td>0.5589</td>
<td>0.5111</td>
</tr>
<tr>
<td>Error (%)</td>
<td>+4.59</td>
<td>-5.45</td>
<td>-3.75</td>
<td>+1.25</td>
<td>+2.53</td>
<td>-0.92</td>
<td>-2.41</td>
</tr>
</tbody>
</table>
Funtime Validation: Accuracy Interdependency Refinement

- 2 sources of error
  - Leon3 IP model (no cache)
  - Nr. of cycles
- Nr. of instructions is measured

Table XI. 2 sources of inaccuracy: Leon3 energy and performance model + estimated number of cycles (stalls) (no cache).

<table>
<thead>
<tr>
<th>Calibration benchmarks</th>
<th>FFT</th>
<th>Quicksort</th>
<th>Fibonacci</th>
<th>Viterbi</th>
<th>Queens</th>
<th>Dhrystone</th>
<th>MD5</th>
</tr>
</thead>
<tbody>
<tr>
<td># Instructions</td>
<td>1 371 584</td>
<td>4 008 651</td>
<td>1 585 313</td>
<td>2 452 903</td>
<td>4 266 783</td>
<td>1 950 813</td>
<td>1 802 380</td>
</tr>
<tr>
<td>Measured # cycles</td>
<td>6 485 094</td>
<td>31 649 232</td>
<td>13 902 019</td>
<td>17 511 989</td>
<td>24 193 885</td>
<td>15 938 144</td>
<td>14 428 588</td>
</tr>
<tr>
<td>Estimated # cycles</td>
<td>7 217 187</td>
<td>31 153 927</td>
<td>13 251 730</td>
<td>18 484 266</td>
<td>24 878 824</td>
<td>15 660 388</td>
<td>14 082 568</td>
</tr>
<tr>
<td>Error (%)</td>
<td>+11.29</td>
<td>−1.56</td>
<td>−4.68</td>
<td>+5.55</td>
<td>+2.83</td>
<td>−1.74</td>
<td>−2.40</td>
</tr>
<tr>
<td>Measured energy (mJ)</td>
<td>0.2396</td>
<td>1.1508</td>
<td>0.4906</td>
<td>0.6250</td>
<td>0.8571</td>
<td>0.5641</td>
<td>0.5228</td>
</tr>
<tr>
<td>Estimated energy (mJ)</td>
<td>0.2746</td>
<td>1.0719</td>
<td>0.4509</td>
<td>0.6648</td>
<td>0.9013</td>
<td>0.5498</td>
<td>0.4989</td>
</tr>
<tr>
<td>Error (%)</td>
<td>+14.62</td>
<td>−6.86</td>
<td>−8.10</td>
<td>+6.36</td>
<td>+5.15</td>
<td>−2.54</td>
<td>−4.74</td>
</tr>
</tbody>
</table>
Funtime Validation: Accuracy Interdependency Refinement

- 3 sources of error
  - Leon3 IP model (no cache)
  - Nr. of cycles
  - Nr. of instructions

<table>
<thead>
<tr>
<th>Calibration benchmarks</th>
<th>FFT</th>
<th>Quicksort</th>
<th>Fibonacci</th>
<th>Viterbi</th>
<th>Queens</th>
<th>Dhrystone</th>
<th>MD5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Measured # instructions</td>
<td>1371584</td>
<td>4008651</td>
<td>1585313</td>
<td>2452903</td>
<td>4266783</td>
<td>1950813</td>
<td>1802380</td>
</tr>
<tr>
<td>Estimated # instructions</td>
<td>1273927</td>
<td>3824253</td>
<td>1517196</td>
<td>2398390</td>
<td>4006083</td>
<td>1880389</td>
<td>1707214</td>
</tr>
<tr>
<td>Error (%)</td>
<td>−7.12</td>
<td>−4.6</td>
<td>−4.3</td>
<td>−2.22</td>
<td>−6.11</td>
<td>−3.61</td>
<td>−5.28</td>
</tr>
<tr>
<td>Measured # cycles</td>
<td>6485094</td>
<td>31649232</td>
<td>13902019</td>
<td>17511989</td>
<td>24193885</td>
<td>15938144</td>
<td>14428588</td>
</tr>
<tr>
<td>Estimated # cycles</td>
<td>6731528</td>
<td>29427456</td>
<td>12185120</td>
<td>17783529</td>
<td>23680975</td>
<td>15131674</td>
<td>13233901</td>
</tr>
<tr>
<td>Error (%)</td>
<td>+3.8</td>
<td>−7.02</td>
<td>−12.35</td>
<td>+3.55</td>
<td>−2.12</td>
<td>−5.06</td>
<td>−8.28</td>
</tr>
<tr>
<td>Measured energy (mJ)</td>
<td>0.2396</td>
<td>1.1508</td>
<td>0.4906</td>
<td>0.6250</td>
<td>0.8571</td>
<td>0.5641</td>
<td>0.5228</td>
</tr>
<tr>
<td>Estimated energy (mJ)</td>
<td>0.2666</td>
<td>1.0417</td>
<td>0.4061</td>
<td>0.6611</td>
<td>0.9054</td>
<td>0.5265</td>
<td>0.4588</td>
</tr>
<tr>
<td>Error (%)</td>
<td>+11.29</td>
<td>−9.48</td>
<td>−17.23</td>
<td>+5.78</td>
<td>+5.64</td>
<td>−6.67</td>
<td>−12.25</td>
</tr>
</tbody>
</table>
# Funtime Validation: Speed

**Without OS**

<table>
<thead>
<tr>
<th>Image</th>
<th>#ALTs</th>
<th>TLM-PV</th>
<th>FUNTHERE</th>
<th>Speedup</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jpeg2k_128x128</td>
<td>82M</td>
<td>3.00 sec</td>
<td>0.62 sec</td>
<td>5</td>
</tr>
<tr>
<td>Jpeg2k_256x256</td>
<td>272M</td>
<td>10.30 sec</td>
<td>0.69 sec</td>
<td>15</td>
</tr>
<tr>
<td>Jpeg2k_512x512</td>
<td>986M</td>
<td>38.66 sec</td>
<td>1.04 sec</td>
<td>37</td>
</tr>
<tr>
<td>H264_176x144</td>
<td>1.68B</td>
<td>50.75 sec</td>
<td>1.41 sec</td>
<td>36</td>
</tr>
<tr>
<td>H264_352x288</td>
<td>2.55B</td>
<td>79.80 sec</td>
<td>1.58 sec</td>
<td>51</td>
</tr>
</tbody>
</table>

**With RTEMS**

<table>
<thead>
<tr>
<th>Tasks</th>
<th>TLM[s]</th>
<th>Fun[s]</th>
<th>Speedup</th>
</tr>
</thead>
<tbody>
<tr>
<td>$T_1 \rightarrow$ Jpeg2k_128x128</td>
<td>46.66</td>
<td>1.86</td>
<td>25</td>
</tr>
<tr>
<td>$T_2 \rightarrow$ Jpeg2k_512x512</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$T_1 \rightarrow$ H264_176x144</td>
<td>147.52</td>
<td>3.19</td>
<td>46</td>
</tr>
<tr>
<td>$T_2 \rightarrow$ H264_352x288</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$T_1 \rightarrow$ Jpeg2k_128x128</td>
<td>93.56</td>
<td>2.40</td>
<td>39</td>
</tr>
<tr>
<td>$T_2 \rightarrow$ H264_352x288</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Future Work

- Extend RTOS refinement
  - Priority-based scheduling policy
  - Resource contention
- Complete the bus arbitration refinement
- Investigate the Power management refinement
- Understand in which order the different refinement should be applied
Thank you for your attention!

Questions?