Trace-driven simulation of multithreaded applications in gem5
Gilles Sassatelli, Alejandro Nocua, Florent Bruguier, Anastasiia Butko

To cite this version:

HAL Id: lirmm-01723755
https://hal-lirmm.ccsd.cnrs.fr/lirmm-01723755
Submitted on 18 Jun 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Trace-driven simulation of multithreaded applications in gem5

Gilles SASSATELLI
sassatelli@lirmm.fr

Alejandro NOCUA, Florent BRUGUIER, Anastasiia BUTKO

Cambridge, UK – September 11, 2017
Background motivations

Mont-Blanc 1, 2 & 3 projects (FP7, H2020)
- Getting ARM technology ready for HPC: HW, SW & Apps
- Advances in energy efficiency towards Exascale

Initial effort: using gem5 for performance prediction (2011)
- STE Nova A9500 SoC (dual-core Cortex A9)
- Fed publicly available parameters into a gem5 FS model
- 1.5% - 18% error, due to rough DRAM model and interconnect

Background motivations cont’d

- **Calibration against real hardware**
  - Using **gem5** for performance prediction
  - And **McPAT** for power estimation

- **Then onto exploring fancy architecture configurations**
  - Heterogeneous single-ISA multicores *à la* big.LITTLE
  - Assymetric, 3 levels of heterogeneity etc.
Background motivations cont’d

- Not ready for manycores, too slow!
  - 1K-1M (simulated) IPS
  - Scales bad with system size
  - Already much better than RTL though

- Trading speed for accuracy? Any sweet spot?

---

Trading speed for accuracy? Any sweet spot?

---

Accuracy

Low

High

Speed

Low

High

RTL

Cycle-accurate

Cycle-approximate

Distributed

Trace-driven

JIT

Ideal
Background motivations cont’d

- Not ready for manycores, too slow!
  - 1K-1M (simulated) IPS
  - Scales bad with system size
  - Already much better than RTL though

- Trading speed for accuracy?

---

1 core FS simulation

Accuracy

- Cycle-accurate
- Cycle-approximate
- Distributed
- Trace-driven
- JIT

Speed

- Low
- High

Accuracy

- Low
- High

10s MIPS
~ 70% simulation effort goes into CPU

- Abstracting away CPU cores sounds like a good idea
- Between 2 consecutive L1 cache misses (in-order) CPU cores perform « consistently »
SimMATE: 2-stage process

- Trace collection: tracing only L1 miss related transactions
- Trace replay: Using trace injectors that initiate transactions as previously recorded
SimMATE for faster DSE

- **Trace collection = freezing**
  - CPU parameters alongside application SW
  - Private caches sizes, speed etc.

- **Trace replay allows exploring the rest**
  - L2 size, policy etc.
  - Interconnect type & speed
  - Main memory speed

![Diagram of Trace Collection and Simulation](image-url)
SimMATE for faster DSE cont’d

- Trace replay performs event (re-) scheduling
  - Simple « time shifting » approach
  - Maintaining constant compute phases

---

Trace-Driven Approach

`Trace Injector`

Events scheduling

Memory & Interconnect

$\text{TI}$

$T_{R1}$

$T_{A1}$

$T_{A1}'$

$T_{R2}'$

Communication

$\text{Time}$
SimMATE Benchmarking

- Tuning DRAM latency
  - Collection performed with 30ns
  - TD simulation from 5ns to 55ns
  - FS used as reference

![Diagram showing execution time error comparison between TD Simulation and FS Simulation across different latencies.](image-url)
## Tuning L2 size

- Collection performed without L2
- TD simulation from 0 to 16MB L2
- Errors originate from Cold-start bias / cache warmup

### Execution time error [%]

**Traces:**
- ET
- ET+init
- ET+OS boot

<table>
<thead>
<tr>
<th>L2 cache size</th>
<th>ET</th>
<th>ET+init</th>
<th>ET+OS boot</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original</td>
<td>15%</td>
<td>8%</td>
<td>7.5%</td>
</tr>
<tr>
<td>256kB</td>
<td>18%</td>
<td>10%</td>
<td>8%</td>
</tr>
<tr>
<td>1MB</td>
<td>15%</td>
<td>8%</td>
<td>8%</td>
</tr>
<tr>
<td>8MB</td>
<td>18%</td>
<td>10%</td>
<td>8%</td>
</tr>
<tr>
<td>16MB</td>
<td>15%</td>
<td>8%</td>
<td>7%</td>
</tr>
</tbody>
</table>
Multithreaded applications

- Having these traces collected makes it easy to:
  - Perform « Trace replication » i.e. emulate more CPU cores for scalability study
  - This corresponds to weak scaling experiments, i.e. per-core workload remains same

![Trace Collection Diagram](image.png)

![Trace Simulation Diagram](image.png)

**Number of Cache Misses**

- Core 0/Thread 0
- Core 1/Thread 1
Yet synchronizations must be accounted for!

- Using whatever API: POSIX threads, OpenMP 3.0 …
- Approach: embed synchronizations into traces
- Have an arbiter that takes care of locking (when barrier reached) and unlocking TIs
- And most ARM AP are OoO (Out-of-Order)
  - Meaning multiple outstanding memory transactions
  - The assumption of constant time btw. 2 misses does not hold

- big-LITTLE & other heterogeneous friends everywhere
  - And there microarchitecture details cannot be overlooked
Modeling micro-architecture timing & dependencies

- Tracing with O3 model + probes, without L2 cache
- Replay done in a smart « elastic » fashion

Elastic Traces: Trace-driven simulation for OoO

http://gem5.org/TraceCPU
Smart TraceCPU

- Updating a dependency graph pushing ready instructions into a queue for issue

Elastic Traces: Trace-driven simulation for OoO

http://gem5.org/TraceCPU
SimMATE + Elastic Traces = ElasticSimMATE

- Enabling both OoO + multithreaded applications
- Key: embed synchronization information @ tracing time.
Proper tracing of synchronizations

- API-dependant: **OpenMP 3.x**
- Tracing whenever entering or leaving parallel region, barrier etc.

```c
#pragma omp parallel
for(i=0;i<n;i++) {
    /* do_some work */
}
```

```c
for(i=0;i<n;i++) {
    OMP_runtime_call()
    /* do_some work */
}
```

<table>
<thead>
<tr>
<th>Tick</th>
<th>PC</th>
<th>th_id</th>
<th>type</th>
<th>IC</th>
<th>DC</th>
</tr>
</thead>
<tbody>
<tr>
<td>389178</td>
<td>177216</td>
<td>0</td>
<td>1</td>
<td>1081157</td>
<td>165738</td>
</tr>
</tbody>
</table>

...
ESW wrapup

- ESM flow wrapup
  - Using BSC Mercurium compiler / Nanos++ runtime
  - Tweaked runtime such that custom m5 pseudo instructions produce trace records
**Benchmarking**

- **Two main use cases:**
  - Fast parameter exploration
  - Scalability study: « trace replication »

- **Speedup & accuracy?**
  - Experiments on Rodinia application kernels

---

**Graphs:**

- **Execution Time [s]**
  - L2 Cache Size: 0MB, 16kB, 128kB, 512kB, 1MB, 2MB
  - FS, ET, ESM

- **Error Percentage [%]**
  - L2 Cache Size: 0MB, 16kB, 128kB, 512kB, 1MB, 2MB
  - FS vs ET, FS vs ESM
Two main use cases:

- Fast parameter exploration
- Scalability study: «trace replication»

Speedup & accuracy?

- Experiments on Rodinia application kernels

**Benchmarking**

- **Execution Time [ms]**
  - FS
  - ET
  - ESM

- **Error Percentage [%]**
  - FS vs ET
  - FS vs ESM

<table>
<thead>
<tr>
<th>L2 Cache Size</th>
<th>Execution Time [ms]</th>
<th>Error Percentage [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>0MB</td>
<td>140</td>
<td>0</td>
</tr>
<tr>
<td>16kB</td>
<td>120</td>
<td>2</td>
</tr>
<tr>
<td>128kB</td>
<td>100</td>
<td>4</td>
</tr>
<tr>
<td>512kB</td>
<td>80</td>
<td>6</td>
</tr>
<tr>
<td>1MB</td>
<td>60</td>
<td>8</td>
</tr>
<tr>
<td>2MB</td>
<td>40</td>
<td>10</td>
</tr>
</tbody>
</table>

1 core  KMEANS
Two main use cases:
- Fast parameter exploration
- Scalability study: « trace replication »

Speedup & accuracy?
- Experiments on Rodinia application kernels

Benchmarking

Execution Time [s]

<table>
<thead>
<tr>
<th>L2 Cache Size</th>
<th>0MB</th>
<th>16kB</th>
<th>128kB</th>
<th>512kB</th>
<th>1MB</th>
<th>2MB</th>
</tr>
</thead>
<tbody>
<tr>
<td>FS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ET</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ESM</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Error Percentage [%]

<table>
<thead>
<tr>
<th>L2 Cache Size</th>
<th>0MB</th>
<th>16kB</th>
<th>128kB</th>
<th>512kB</th>
<th>1MB</th>
<th>2MB</th>
</tr>
</thead>
<tbody>
<tr>
<td>FS vs ET</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FS vs ESM</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
**Benchmarking**

- **Two main use cases:**
  - Fast parameter exploration
  - Scalability study: « trace replication »

- **Speedup & accuracy?**
  - Experiments on Rodinia application kernels

![Graphs showing L2 cache miss rate and overall miss latency for FS and ESM across different cache sizes](image)
Two main use cases:

- Fast parameter exploration
- Scalability study: «trace replication»

Speedup & accuracy?

- Experiments on Rodinia application kernels

Benchmarking

<table>
<thead>
<tr>
<th>Number of Cores</th>
<th>Blackscholes</th>
<th>Hotspot</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>2</td>
<td>200</td>
<td>200</td>
</tr>
<tr>
<td>4</td>
<td>400</td>
<td>400</td>
</tr>
<tr>
<td>8</td>
<td>800</td>
<td>800</td>
</tr>
<tr>
<td>16</td>
<td>1600</td>
<td>1600</td>
</tr>
<tr>
<td>32</td>
<td>3200</td>
<td>3200</td>
</tr>
<tr>
<td>64</td>
<td>6400</td>
<td>6400</td>
</tr>
<tr>
<td>128</td>
<td>12800</td>
<td>12800</td>
</tr>
</tbody>
</table>

Simulation Time [min]

<table>
<thead>
<tr>
<th>Number of Cores</th>
<th>Blackscholes</th>
<th>Hotspot</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>2</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td>4</td>
<td>40</td>
<td>40</td>
</tr>
<tr>
<td>8</td>
<td>80</td>
<td>80</td>
</tr>
<tr>
<td>16</td>
<td>160</td>
<td>160</td>
</tr>
<tr>
<td>32</td>
<td>320</td>
<td>320</td>
</tr>
<tr>
<td>64</td>
<td>640</td>
<td>640</td>
</tr>
<tr>
<td>128</td>
<td>1280</td>
<td>1280</td>
</tr>
</tbody>
</table>
Scalability analysis has limits

- Requires additional features s.a. **address offsetting**
- **Weak scaling** only (replicated per-core workloads)

Programming models moving from loops to *tasks*

- OpenMP 4.0, OmpSs
- Still pragma-based
- More parallelisms available at run-time ... more opportunities for smart job scheduling

---

Perspectives

BSC OmpSs: Cholesky decomposition
Unbinding traces from cores

- One trace per Task, not per core!
- Assign traces to cores by emulating runtime behaviour in trace replay
- This is real strong scaling

Trace collection

- Ready Task queue
- Scheduler
- Task 1 trace
- Task 2 trace
- Task N trace
- Task dependencies

Trace simulation

- Ready Task queue
- Task 2 trace
- Scheduler / runtime (emulated)
- Task 1 trace
- Task 1 trace
- Task 1 trace
- Task 2 trace
- Task 2 trace
- Interconnect
- Memory

Perspectives
■ Current ESM prototype
  ● ~5x - 10x speedup for low core count, probably more for tens / thousands
  ● Nice solution for fast DSE
  ● Remaining accuracy issues for some applications
    - Common to ESM & Elastic Traces
    - Under investigation with ARM

■ Use cases
  ● Exploration of memory subsystem
  ● Some microarchitecture parameters (Elastic Traces)
  ● …

■ Future directions
  ● Ruby compatibility
  ● Could be combined with other initiatives (dist-gem5)
  ● Can be extended to other PM / APIs (Tasking, MPI…)

http://montblanc-project.eu