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
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

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.

---

**Figure 4: Performance and energy trade-offs.**

- Baseline (4A7/4A15)
- SMP
- big.LITTLE/big.Medium
- big.Medium.LITTLE
- Energy Delay Product
- Energy Delay Product

---

**Diagram:**

- 'big' cluster
- 'LITTLE' cluster
- Assymetric Big.LITTLE
- Big.medium.LITTLE

---

**Notes:**

- Accuracy assessments
- Results
- Execution time error ~20%
- Power error ~12%
- 2 L2 caches, Crossbar Bus
- Independent clock per cluster
- 8 heterogeneous cores
- Octa
- Exynos
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?

![Diagram showing trade-off between speed and accuracy](image-url)
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?

![Diagram showing trade-off between accuracy and speed for different simulation modes: RTL, Cycle-accurate, Cycle-approximate, Distributed, Trace-driven, JIT.](image-url)
~ 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]

- **Trace Collection**
  - Cores
  - Interconnect
  - Memory
  - Hits & Misses

- **Trace Simulation**
  - TIs
  - Interconnect
  - Memory & L2
  - Misses

1 time

Multiple times
- Trace replay performs event (re-) scheduling
  - Simple « time shifting » approach
  - Maintaining constant compute phases
Tuning DRAM latency

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

Execution time error [%]

- MJPEG
- FFT

<table>
<thead>
<tr>
<th>Execution time</th>
<th>MJPEG</th>
<th>FFT</th>
</tr>
</thead>
<tbody>
<tr>
<td>5ns</td>
<td>0,6%</td>
<td>2,8%</td>
</tr>
<tr>
<td>15ns</td>
<td>0,5%</td>
<td>4,5%</td>
</tr>
<tr>
<td>30ns</td>
<td>0,1%</td>
<td>5,8%</td>
</tr>
<tr>
<td>45ns</td>
<td>0,4%</td>
<td>5,1%</td>
</tr>
<tr>
<td>55ns</td>
<td>0,4%</td>
<td>6%</td>
</tr>
</tbody>
</table>

Original
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 Original</th>
<th>ET+init Original</th>
<th>ET+OS boot Original</th>
<th>ET 256kB</th>
<th>ET+init 256kB</th>
<th>ET+OS boot 256kB</th>
<th>ET 1MB</th>
<th>ET+init 1MB</th>
<th>ET+OS boot 1MB</th>
<th>ET 8MB</th>
<th>ET+init 8MB</th>
<th>ET+OS boot 8MB</th>
<th>ET 16MB</th>
<th>ET+init 16MB</th>
<th>ET+OS boot 16MB</th>
</tr>
</thead>
<tbody>
<tr>
<td>0,1%</td>
<td>15%</td>
<td>8%</td>
<td>8%</td>
<td>7,5%</td>
<td>1%</td>
<td>0,3%</td>
<td>14%</td>
<td>10%</td>
<td>1%</td>
<td>15%</td>
<td>8%</td>
<td>7%</td>
<td>8%</td>
<td>7%</td>
<td></td>
</tr>
</tbody>
</table>
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

![Diagram of Trace Collection and Simulation]

**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
Elastic Traces: Trace-driven simulation for OoO

- Modeling micro-architecture timing & dependencies
  - Tracing with O3 model + probes, without L2 cache
  - Replay done in a smart « elastic » fashion

![Diagram showing One simulation run: capture trace and Design space exploration: replay same trace]
Elastic Traces: Trace-driven simulation for OoO

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

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

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

ElasticSimMATE (ESM)
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 */
}

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

```plaintext
1.bin
OMP runtime + scheduler

MS

MH

Tick  PC  th_id  type  IC     DC
389178 177216 0   1   1081157 165738
...
```
ESM flow wrapup

- Using BSC Mercurium compiler / Nanos++ runtime
- Tweaked runtime such that custom m5 pseudo instructions produce trace records
Two main use cases:

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

Speedup & accuracy?

- Experiments on Rodinia application kernels
Two main use cases:

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

Speedup & accuracy?

- Experiments on Rodinia application kernels

---

### Execution Time [ms]

<table>
<thead>
<tr>
<th>L2 Cache Size</th>
<th>FS</th>
<th>ET</th>
<th>ESM</th>
</tr>
</thead>
<tbody>
<tr>
<td>0MB</td>
<td>120</td>
<td>120</td>
<td>120</td>
</tr>
<tr>
<td>16kB</td>
<td>120</td>
<td>120</td>
<td>120</td>
</tr>
<tr>
<td>128kB</td>
<td>120</td>
<td>120</td>
<td>120</td>
</tr>
<tr>
<td>512kB</td>
<td>120</td>
<td>120</td>
<td>120</td>
</tr>
<tr>
<td>1MB</td>
<td>120</td>
<td>120</td>
<td>120</td>
</tr>
<tr>
<td>2MB</td>
<td>120</td>
<td>120</td>
<td>120</td>
</tr>
</tbody>
</table>

### Error Percentage [%]

<table>
<thead>
<tr>
<th>L2 Cache Size</th>
<th>FS vs ET</th>
<th>FS vs ESM</th>
</tr>
</thead>
<tbody>
<tr>
<td>0MB</td>
<td>12</td>
<td>12</td>
</tr>
<tr>
<td>16kB</td>
<td>12</td>
<td>12</td>
</tr>
<tr>
<td>128kB</td>
<td>12</td>
<td>12</td>
</tr>
<tr>
<td>512kB</td>
<td>12</td>
<td>12</td>
</tr>
<tr>
<td>1MB</td>
<td>12</td>
<td>12</td>
</tr>
<tr>
<td>2MB</td>
<td>12</td>
<td>12</td>
</tr>
</tbody>
</table>
Two main use cases:

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

Speedup & accuracy?

- Experiments on Rodinia application kernels

Benchmarking

<table>
<thead>
<tr>
<th>L2 Cache Size</th>
<th>Execution Time [s]</th>
<th>Error Percentage [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>0MB</td>
<td>FS</td>
<td>ET</td>
</tr>
<tr>
<td>16kB</td>
<td></td>
<td></td>
</tr>
<tr>
<td>128kB</td>
<td></td>
<td></td>
</tr>
<tr>
<td>512kB</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1MB</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2MB</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

1 core CANNEAL

FS vs ET
FS vs ESM
Two main use cases:

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

Speedup & accuracy?

- Experiments on Rodinia application kernels

---

**L2**

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

Speedup & accuracy?
- Experiments on Rodinia application kernels
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 parallelims available at run-time … more opportunities for smart job scheduling

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

Trace simulation

Scheduler / runtime (emulated)
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…)

Conclusion