Fast Transient Thermal Simulation Based on Linear System Theory

Yongkui Han, Israel Koren, C. Mani Krishna
ARTS Lab, ECE Dept
University of Massachusetts at Amherst
Thermal Simulation is Important

- Thermal simulation has become a necessity for contemporary microprocessors

- Transient Thermal Simulation Models
  - 1. Finite Elements Method (FEM)
    - Very accurate but computation-intensive
  - 2. Compact thermal RC models
    - Less accurate but faster simulation

- The HotSpot simulator
  - based on the compact thermal RC model - at architecture level
  - developed by LAVA group at Univ. of Virginia
HotSpot is Inefficient

- HotSpot uses the fourth order Runge-Kutta method (rk4) for integration
  - To reduce the truncation error, the step size in rk4 must be small
  - Thousands of rk4 iterations are required when simulating for a 1ms time interval
- HotSpot becomes inefficient when attempting to obtain the temperature profile for a given benchmark

- Our new method improves the simulation speed of HotSpot
Experiments

- Two microprocessors
- SPEC2000 benchmarks
- Most recent HotSpot v3.0.2 is used

Pentium Pro processor:
16 power inputs, 58 nodes.

Alpha 21364 processor:
18 power inputs, 97 nodes.
CPU Chip as a Linear System

\[ \dot{x}(t) = Fx(t) + Gu(t) \quad (1) \]

The response of the linear system is:

\[ x(t) = e^{Ft}x(0) + \int_{0}^{t} e^{F(t-\tau)}Gu(\tau)d\tau \quad (2) \]

During a time interval \([\Delta t, i \Delta t]\), the input power \(u\) is fixed:

\[ x(n) = A x(n-1) + B u(n-1) \quad (3) \]

Let \( A = e^{F\Delta t} \), \( B = \int_{0}^{\Delta t} e^{F(\Delta t-\tau)}Gd\tau \)

\( u(i) \) is the input power during the interval \([(i-1)\Delta t, i \Delta t]\)

The dimensions of \(x, u\) depend on specific processor analyzed.
Time Invariant Linear Thermal System (TILTS) method

- **TILTS**
  - Precompute matrices $A$ and $B$ using rk4 method
  - Compute temperatures using (3)

- Many rk4 iterations are replaced by simple matrix multiplications

- TILTS does not incur any accuracy loss compared to the HotSpot simulator
# Computation Reductions in TILTS

Comparing the number of Floating-Point Multiplications (FPM) in HotSpot and TILTS

<table>
<thead>
<tr>
<th>processor</th>
<th>$\Delta t$</th>
<th>#FPM in HotSpot</th>
<th>#FPM in TILTS</th>
<th>ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pentium Pro</td>
<td>5 $\mu s$</td>
<td>267670</td>
<td>4292</td>
<td>62</td>
</tr>
<tr>
<td>Alpha 21364</td>
<td>3.33 $\mu s$</td>
<td>514197</td>
<td>11155</td>
<td>46</td>
</tr>
</tbody>
</table>
Convolutional TILTS (CTILTS) algorithm
- Precompute matrices $A^j$ and $A^jB$, $j=1,2,...,p$
- Perform the convolution between the input power trace and the step response of the system
- CTILTS reduces the number of floating-point multiplications (FPM) by 5 or 6 times
- Small memory overhead

<table>
<thead>
<tr>
<th>processor</th>
<th>interval</th>
<th>memory size</th>
<th>#FPM in TILTS</th>
<th>#FPM in CTILTS</th>
<th>ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pentium Pro</td>
<td>5.1 ms</td>
<td>7.25 MB</td>
<td>4395008</td>
<td>953636</td>
<td>4.61</td>
</tr>
<tr>
<td>Alpha</td>
<td>3.4 ms</td>
<td>13.64 MB</td>
<td>11422720</td>
<td>1797313</td>
<td>6.36</td>
</tr>
</tbody>
</table>
## Experimental Results

<table>
<thead>
<tr>
<th>processor</th>
<th>prog</th>
<th>HotSpot</th>
<th>TILTS</th>
<th>speedup</th>
<th>CTILTS</th>
<th>speedup</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pentium Pro</td>
<td>gcc</td>
<td>15800s</td>
<td>237s</td>
<td>67x</td>
<td>59s</td>
<td>268x</td>
</tr>
<tr>
<td>Pentium Pro</td>
<td>gzip</td>
<td>15790s</td>
<td>238s</td>
<td>66x</td>
<td>59s</td>
<td>268x</td>
</tr>
<tr>
<td>Pentium Pro</td>
<td>bzip2</td>
<td>15792s</td>
<td>237s</td>
<td>67x</td>
<td>59s</td>
<td>268x</td>
</tr>
<tr>
<td>Pentium Pro</td>
<td>art</td>
<td>15800s</td>
<td>238s</td>
<td>66x</td>
<td>59s</td>
<td>268x</td>
</tr>
<tr>
<td>Pentium Pro</td>
<td>mgrid</td>
<td>15790s</td>
<td>238s</td>
<td>66x</td>
<td>59s</td>
<td>268x</td>
</tr>
<tr>
<td>Alpha</td>
<td>gcc</td>
<td>28010s</td>
<td>592s</td>
<td>47x</td>
<td>129s</td>
<td>217x</td>
</tr>
<tr>
<td>Alpha</td>
<td>gzip</td>
<td>28030s</td>
<td>591s</td>
<td>47x</td>
<td>129s</td>
<td>217x</td>
</tr>
<tr>
<td>Alpha</td>
<td>bzip2</td>
<td>28020s</td>
<td>592s</td>
<td>47x</td>
<td>129s</td>
<td>217x</td>
</tr>
<tr>
<td>Alpha</td>
<td>art</td>
<td>28010s</td>
<td>591s</td>
<td>47x</td>
<td>129s</td>
<td>217x</td>
</tr>
<tr>
<td>Alpha</td>
<td>mgrid</td>
<td>28020s</td>
<td>592s</td>
<td>47x</td>
<td>129s</td>
<td>217x</td>
</tr>
</tbody>
</table>

- **TILTS**
  - 67x for Pentium Pro, 47x for Alpha

- **CTILTS**
  - 268x for Pentium Pro, 217x for Alpha
Conclusions

- Our Time Invariant Linear Thermal System (TILTS) method can greatly improve transient thermal simulation performance:
  - 268 times faster than HotSpot for Pentium Pro processor
  - 217 times faster than HotSpot for Alpha 21364 processor

- TILTS does not incur any accuracy loss compared to the HotSpot simulator