Applying the Execution-Cache-Memory Performance Model: Current State of Practice

G. Hager, J. Eitzinger, J. Hornich, F. Cremonesi, C. A. Alappat, T. Röhl, G. Wellein

1 Erlangen Regional Computing Center (RRZE), Erlangen, Germany
2 École Polytechnique Fédérale de Lausanne (EPFL), Geneva, Switzerland

Problem: Saturation

Plain ECM model too optimistic. ϵ is sensitive to the tight kernel with small \( \theta_{\text{L}} \).

Problem: \( \text{L1} \) saturation: Latency penalty depends on bus utilization \( u_b \).

Preliminary: Adaptive latency penalty, depends on bus utilization \( u_b \).

Putting the model together: Overlap assumptions

Notation for model contributions:

- \( \theta_{\text{L}} \) summary data
- \( \theta_{\text{L}} \) full overlap of data-related contributions
- \( \theta_{\text{L}} \) most optimistic assumption: full overlap of data-related contributions
- \( \theta_{\text{L}} \) most pessimistic assumption: no overlap of data-related contributions

Notation for model predictions:

- \( \theta_{\text{L}} \) full overlap of data-related contributions
- \( \theta_{\text{L}} \) most optimistic assumption: full overlap of data-related contributions
- \( \theta_{\text{L}} \) most pessimistic assumption: no overlap of data-related contributions

Application: Complex stencils

FP32 code solving Navier-Stokes Equations via TMM for thin-film solar cells, AVX vectorization via C intrinsics, complex arithmetic:

\[
\text{mggate}[_{iml}] = \frac{1}{1 + \exp (-0.062 \cdot _v/(3.57))};
\]

Optimistic assumption: Performance scaling is linear until a bandwidth bottleneck

Example: No overlapping model

\[
\text{max}(\text{coremax}, \text{coremin})
\]

Example: No overlapping model

\[
\alpha = \alpha + 1(s)
\]

Embedding MM content

Support by

\[ \text{https://www.rci.uci.edu} \]

Application: Conjugate Gradient solver

CMB [cy/iter]

\[
\text{CPU} = \frac{\text{max} (\text{coremax}, \text{coremin})}{\text{coremin}}
\]

Software: Intel Xeon E5-2600v3

\[ \Rightarrow \text{CPU} = \text{coremin}/\text{coremax} \]

Optimistic assumption: Performance scaling is linear until a bandwidth bottleneck (e.g., \( \theta_{\text{L}} \)).

Mixed model: partial overlap of data-related contributions

Example: No overlapping model

\[ \text{max} (\text{coremax}, \text{coremin}) \]

0.5 GB/s

Throughput analysis

\[ \text{L2} \] case analysis only

Example: Non-overlapping CPU with on-chip clock domain down to \( \text{L3} \), \( \text{fcpu} \), independent of \( \text{f} \)

Application: Blue Brain Project kernels

What is the computational cost of a synapse?

- Study performance of brain simulation code via mini-apps
- SODA execution by core

Case 1: Synthetic current kernel

- \( \text{fcpu} \) measured
- \( \text{fcpu} \) speed

Case 2: Sodium ion channel (Na, Na2, Na3)

- \( \text{fcpu} \) measured
- \( \text{fcpu} \) speed