Modeling Performance through Memory-Stalls

Roman Iakymchuk and Paolo Bientinesi

AICES, RWTH Aachen University
iakymchuk@aices.rwth-aachen.de

Facing the Multicore-Challenge II
September 28-30, 2011
Karlsruhe, Germany
Performance = \frac{\#FLOPS}{\text{Execution\_Time}}

\#FLOPS is known \text{ a priori}
Performance \(= \frac{\#FLOPS}{Execution\_Time}\)

\(\#FLOPS\) is known \textbf{a priori}

**Modeling through time measurements**

- **Timing** the kernels (BLAS subroutines)
Performance = \frac{\#FLOPS}{Execution\_Time}

\#FLOPS is known \textbf{a priori}

Modeling through time measurements

- **Timing** the kernels (BLAS subroutines)

Modeling through memory-stalls

- \textbf{Without executing} either the \textit{algorithm} or \textit{parts of it}
\[ \text{Execution time} = \]
\[ \alpha \times \text{L1 cache misses} \times \text{L1 cache miss time} + \]
\[ \beta \times \text{L2 cache misses} \times \text{L2 cache miss time} + \]
\[ \gamma \times \text{TLB misses} \times \text{TLB miss time} + \]
\[ \eta \times \#\text{FLOPS} \times \text{FLOP time} \]
Execution_time =
\[\alpha \times L1\_cache\_misses \times L1\_cache\_miss\_time +\]
\[\beta \times L2\_cache\_misses \times L2\_cache\_miss\_time +\]
\[\gamma \times TLB\_misses \times TLB\_miss\_time +\]
\[\eta \times \#FLOPS \times FLOP\_time\]

- \(\alpha\), \(\beta\), \(\gamma\), and \(\eta\) are used to model the overlap between computations and data movement
\[ \text{Execution\_time} = \]
\[ \alpha \times L1\_cache\_misses \times L1\_cache\_miss\_time + \]
\[ \beta \times L2\_cache\_misses \times L2\_cache\_miss\_time + \]
\[ \gamma \times TLB\_misses \times TLB\_miss\_time + \]
\[ \eta \times \#FLOPS \times FLOP\_time \]

- $\alpha$, $\beta$, $\gamma$, and $\eta$ are used to model the overlap between computations and data movement

Our goal is to model L1 misses.
Unblocked LU Factorization

\[ a_i := \frac{a_i}{\alpha_{ii}} \quad \text{SCAL} \]
\[ A_i := A_i - a_i a_j^T \quad \text{GER} \]
Unblocked LU Factorization

\[ a_i := a_i / \alpha_{ii} \quad \text{SCAL} \]
\[ A_i := A_i - a_i a_j^T \quad \text{GER} \]

- **GER** performs more than **95 %** of the **#FLOPS** in the LU
Unblocked LU Factorization

\[ a_i := a_i / \alpha_{ii} \quad \text{SCAL} \]
\[ A_i := A_i - a_i a_j^T \quad \text{GER} \]

- **GER** performs more than **95%** of the \#FLOPS in the LU
- Model L1 misses in **LU through** modeling **GER**
Unblocked LU Factorization

\[ a_i := a_i / \alpha_{ii} \quad \text{SCAL} \]
\[ A_i := A_i - a_i a_j^T \quad \text{GER} \]

- **GER** performs more than 95% of the \#FLOPS in the LU
- Model L1 misses in LU through modeling GER

**Assumptions**

- Data reside in the L2 cache → only L1 misses occur
- \( A_i, a_i, \) and \( a_j^T \) are aligned
- \( a_i, a_j^T \), and a portion of \( A_i \) can be kept in the L1 cache
Figure: \( m - \left\lfloor \frac{p}{d} \right\rfloor d < d \)

\[
L1\_misses = \left\lfloor \frac{mq}{d} \right\rfloor + \left\lceil \frac{p}{d} \right\rceil + \left\lceil \frac{q}{d} \right\rceil
\]
Modeling L1 Misses

\[
L1_{\text{misses}} = \begin{cases} 
\left\lfloor \frac{mq}{d} \right\rfloor + \left\lceil \frac{p}{d} \right\rceil + \left\lceil \frac{q}{d} \right\rceil, & \text{if } m - \left\lfloor \frac{p}{d} \right\rfloor d < d \\
\zeta + \left\lceil \frac{p}{d} \right\rceil + \left\lceil \frac{q}{d} \right\rceil, & \text{otherwise}
\end{cases}
\]

where

\[
\zeta = \left\lfloor \frac{p}{d} \right\rfloor + \sum_{i=1}^{n-1} \left\lfloor \frac{p + (mi \mod d)}{d} \right\rfloor
\]
Each core has L1 (32 KB) and L2 (3072 KB) caches

- GER from the GotoBLAS library is used
- The deviation is less than 2%

Figure: Deviation on Intel Penryn; $LDA = 512; p \geq q$. 
Figure: Modeling L1 misses on Intel Penryn; $LDA = 512$.

- Closer to origin the deviation is higher
- When $p = q$ increases the deviation $\to 0$
Conclusions and Future Work

\[ \text{Execution_time} = \alpha \times L1\text{\_cache\_misses} \times L1\text{\_cache\_miss\_time} + \]
\[ \beta \times L2\text{\_cache\_misses} \times L2\text{\_cache\_miss\_time} + \]
\[ \gamma \times TLB\text{\_misses} \times TLB\text{\_miss\_time} + \]
\[ \eta \times \#FLOPS \times FLOP\_time \]

Conclusions

- The \textbf{model} was \textbf{validated} by predicting L1 misses of the LU factorization on AMD Barcelona and Intel Penryn
- The \textbf{deviation} is mostly less than 2-3 %
Conclusions and Future Work

\[ \text{Execution\_time} = \]
\[ \alpha \times \text{L1\_cache\_misses} \times \text{L1\_cache\_miss\_time} + \]
\[ \beta \times \text{L2\_cache\_misses} \times \text{L2\_cache\_miss\_time} + \]
\[ \gamma \times \text{TLB\_misses} \times \text{TLB\_miss\_time} + \]
\[ \eta \times \#\text{FLOPS} \times \text{FLOP\_time} \]

Conclusions

- The **model** was **validated** by predicting L1 misses of the LU factorization on AMD Barcelona and Intel Penryn
- The **deviation** is mostly less than 2-3%

Future Work

- From **L1 misses** to **execution time**
- **Model** for **L2 misses**
Financial support from the Deutsche Forschungsgemeinschaft through grant GSC 111 is gratefully acknowledged.