Energy-Efficient Transprecision Techniques for Iterative Refinement


Queen's University Belfast - Research Portal:
Link to publication record in Queen's University Belfast Research Portal

Publisher rights
Copyright 2017 IEEE. This work is made available online in accordance with the publisher’s policies. Please refer to any applicable terms of use of the publisher.

General rights
Copyright for the publications made accessible via the Queen's University Belfast Research Portal is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The Research Portal is Queen's institutional repository that provides access to Queen's research output. Every effort has been made to ensure that content in the Research Portal does not infringe any person’s rights, or applicable UK laws. If you discover content in the Research Portal that you believe breaches copyright or violates any law, please contact openaccess@qub.ac.uk.
Energy-Efficient Transprecision Techniques for Iterative Refinement

JunKyu Lee, Hans Vandierendonck, Dimitrios S. Nikolopoulos
Centre for Data Science and Scalable Computing, Queen’s University of Belfast, UK

Abstract

This work presents transprecision techniques for iterative refinement, which utilize various precision arithmetic dynamically according to numeric properties of the algorithm and computational latencies depending on precisions. The transprecision techniques were plugged into a mixed precision iterative refinement on an Intel Xeon E5-2650 2GHz core with MKL 2017 and XBLAS 1.0. The transprecision techniques brought further 2.0 - 3.4 X speedups and 3.0 - 4.1 X energy reductions to a mixed precision iterative refinement when double precision solution accuracy was required for forward error and a matrix size was ranged from 4K to 32K.

Transprecision Techniques

Parallel Computing with $m \times X$ Cores

$m \times$ Speedup $\quad m \times$ Power $\quad 1 \times$ Energy

Need some techniques for energy saving? Mixed Precision Method

Mixed precision Iterative Refinement without increasing cores

$n \times$ Speedup $\quad 1 \times$ Power $\quad 1/n \times$ Energy

Nice, but further energy saving? Transprecision Techniques!

Transprecision Techniques for Mixed precision Iterative Refinement

$(s \times n) \times$ Speedup $\quad 1 \times$ Power $\quad 1/(s \times n) \times$ Energy

$s$ times future energy saving to Mixed precision Iterative Refinement

Impact of Precision

ALU Precision

Lower $\quad$ Higher

Smaller ALUs $\quad$ Larger ALUs

Shorter Wire $\quad$ Higher Clock Rate

Number of ALUs in Fixed Area $\quad$ Number of TRs

Higher Speedup $\quad$ Higher Performance $\quad$ Less Power

Background

Test matrices: Dense uniformly distributed random matrices

Impact of Precision

$\frac{n}{2}$ $\times$ Speedup $\quad n \times$ Power $\quad 1 \times$ Energy

Mixed-IR : Double precision accuracy for forward error

Approximation

Step 1: LUPP $\quad L \times U \times x^{(1)} = \mathbf{b}$

$O(n^3), \varepsilon_0$ for Mixed-IR, $\varepsilon_0$ for Uni-IR

$O(n^2), \varepsilon_0$ for Mixed-IR, $\varepsilon_0$ for Uni-IR

Refinement

Step 2: $x^{(2)} = A \times x^{(1)} - \mathbf{b}$

$O(n^2), \varepsilon_0$ for Mixed-IR, $\varepsilon_0$ for Uni-IR

(TT 1) $\varepsilon^{(\text{dub})}$ = double precision accuracy

Step 3: $L \times U \times x^{(2)} = \mathbf{b}$

$O(n^2), \varepsilon_0$ for Mixed-IR, $\varepsilon_0$ for Uni-IR

(TT 2) if $\|L(L_\text{ref})^{(dub)} \| \times \varepsilon_0 < \varepsilon_0$, refine $x$ using double precision

Step 4: $x^{(11)} = x^{(10)} - x^{(2)}$

$O(n^2), \varepsilon_0$ for Mixed-IR, $\varepsilon_0$ for Uni-IR

(TT 3) if $\varepsilon_0 < \varepsilon_0$, go back to step 2

LUPP: LU factorization with Partial Pivoting, $\varepsilon_0$: Precision for step i

$\varepsilon_0$: Single Precision, $\varepsilon_0$: Double Precision, $\varepsilon_0$: Double-Double Precision

Methodology

Numerical Properties

Transprecision Techniques Plug-In

Transprecision Techniques

Numerical Properties (NP) and Transprecision Techniques (TT)

NP 1:

For Step 1: Finding the solution $x^{(1)}$

Residual accuracy is almost kept with the attachment of cancellation bits

NP 2:

Irreducible rounding errors in $x$ through Step 3

NP 3:

Double precision accuracy guaranteed if $\varepsilon_0 < \varepsilon_0$ and single precision accuracy for $x$ is obtained using TT 2

Acknowledgements

This project has received funding by the European Commission Horizon 2020 research and innovation programme under grant agreement No 732631 (OPRECOMP)

Results

Transprecision Techniques on an Intel Xeon E5-2650 2GHz core with MKL 2017 and XBLAS 1.0 Test matrices: Dense uniformly distributed random matrices

Speedups with TTs

When matrix size $N = 32K$, Mixed-IR Runtime $\propto$ Uni-IR

Mixed-IR Runtime $\propto O(n^2)$

Runtime with TTs $\propto$ Mixed-IR $\propto$ More Energy Saving!

Less Energy with TTs

Transprecision Techniques brought further

2.0 - 3.4 X Speedup

3.0 - 4.1 X Energy Reduction

to Mixed-IR

Efficient Centre for Data Science and Scalable Computing, Queen’s University of Belfast, UK

Irreducible rounding errors in $x$ through Step 3

Back

Back

Back

Back