Approximate Computing for Energy Efficiency and Resource-Constrained Embedded Systems

Stefano Cherubin – s.cherubin@napier.ac.uk

19 – 21 November 2021

Advanced Photoelectric Sensor Integration and Bioimaging Analysis

The trend of machine-assisted medicine is increasingly adopting machine-learning techniques to detect patterns of interest in sensors data. In the computer architecture world, at the same time, hardware manufacturers started developing hardware accelerators to speed up the execution of machine-learning software. The main innovation brought in by these products is the use of reduced precision data types used during the computation. Here is where **approximate computing** steps in. Measuring the functional impact on the output (numerical accuracy of the result) and estimating the non-functional effect (achieved speedup or energy saving) is not an easy. This is why researchers in the field of computer architectures and software engineering developed tools to automatically derive this profile for each specific application, and thus to guide application experts towards an optimal use of the hardware they have available.

TAFFO [1] is an open source toolchain that enables smart data type conversion on numerically-intensive applications. It is capable to leverage floating point and fixed point data types. TAFFO can be configured to dynamically switch to faster floating point value, to replace them with fixed point (integer) numbers, or to mix these representation in the most efficient way. It can optimise the code for time-to-solution, for energy consumption, and other metrics can be implemented as needed.

TAFFO is based on the cutting-edge LLVM compiler technology. It is designed for efficiency and modularity, so it can be adapted to different programming environments with little effort. It has been tested in embedded systems [2, 3] as well as in high-performance computing scenarios [4, 5]. In particular, it showed great potential in machine-learning applications running on resource-constrained systems [6].

While there are many works in the field of precision tuning [7], very few of them actually implement the technology to automatically translate the application requirements into optimised code, even fewer of them are able to achieve it without a dedicated programming environment. TAFFO does all of that, plus it is the only one with a catchy logo! – see Figure 1.

If you want to know more about how TAFFO can improve your code, feel free to contact the research team behind the tool. We are based in Edinburgh Napier University and Politecnico di Milano, Italy – see taffo-org.github.io.

Figure 1: TAFFO: a Tuning Assistant for Floating point and Fixed point Optimisations – with a very catchy logo

References

[1] Stefano Cherubin, Daniele Cattaneo, Michele Chiari, Antonio Di Bello, and Giovanni Agosta. TAFFO: Tuning assistant for floating to fixed point optimization. *IEEE Embedded Systems Letters*, 12(1):5–8, March 2020.

[2] Gabriele Magnani, Daniele Cattaneo, Michele Chiari, and Giovanni Agosta. The Impact of Precision Tuning on Embedded Systems Performance: A Case Study on Field-Oriented Control. In João Bispo, Stefano Cherubin, and José Flich, editors, *12th Workshop on Parallel Programming and Run-Time Management Techniques for Many-core Architectures and 10th Workshop on Design Tools and Architectures for Multicore Embedded Computing Platforms (PARMA-DITAM 2021)*, volume 88 of *Open Access Series in Informatics (OASIcs)*, pages 3:1–3:13, Dagstuhl, Germany, 2021. Schloss Dagstuhl – Leibniz-Zentrum für Informatik.

[3] Daniele Cattaneo, Michele Chiari, Gabriele Magnani, Nicola Fossati, Stefano Cherubin, and Giovanni Agosta. FixM: Code generation of fixed point mathematical functions. *Sustainable Computing: Informatics and Systems*, 29, March 2021.

[4] Stefano Cherubin, Daniele Cattaneo, Michele Chiari, and Agosta Giovanni. Dynamic precision autotuning with TAFFO. *ACM Transaction on Architecture and Code Optimization*, 17(2), may 2020.

[5] Daniele Cattaneo, Michele Chiari, Stefano Cherubin, and Giovanni Agosta. Feedback-driven performance and precision tuning for automatic fixed point exploitation. In *International Conference on Parallel Computing*, volume 36 of *Advances in Parallel Computing*, pages 299–308, Sep 2019.

[6] Nicola Fossati, Daniele Cattaneo, Michele Chiari, Stefano Cherubin, and Giovanni Agosta. Automated precision tuning in activity classification systems: A case study. In *Proceedings of the 11th Workshop on Parallel Programming and Run-Time Management Techniques for Many-core Architectures and the 9th Workshop on Design Tools and Architectures For Multicore Embedded Computing Platforms*, PARMA-DITAM '20, pages 1–6, Jan 2020.

[7] Stefano Cherubin and Giovanni Agosta. Tools for reduced precision computation: a survey. *ACM Computing Surveys*, 53(2), Apr 2020.

Approximate Computing for Energy Efficiency and Resource-Constrained Embedded Systems

Stefano Cherubin
<s.cherubin@napier.ac.uk>
www.taffo-org.github.io

Edinburgh Napier UNIVERSITY
in collaboration with
POLITECNICO MILANO 1863

improve your software using just the numerical precision you need

Save energy

Save run time

via precision tuning techniques

Make your machine learning software fully exploit the hardware capabilities

Success Stories

We delivered improvements in

+ AI classification algorithms
+ Real-time operating systems
+ Control software
+ HPC benchmarks

up to: 860% 180% 258% 170%

	Stm32				Raspberry				Intel				AMD			
	Precise	Balanced	Fast	TAFFO	Precise	Balanced	Fast	TAFFO	Precise	Balanced	Fast	TAFFO	Precise	Balanced	Fast	TAFFO
2mm	4.6	16.3	140.8	355.2	0.0	-0.0	102.4	96.8	0.0	-4.2	-20.1	-39.9	-3.6	-2.3	7.0	-6.3
3mm	3.3	16.5	177.6	334.0	-0.3	-0.3	53.1	95.0	-0.5	-1.7	4.1	-19.8	-0.7	-0.5	-5.4	-9.3
adi	0.6	42.8	165.8	486.6	0.2	0.1	-55.7	-67.4	-0.6	0.1	19.9	-57.0	5.6	5.9	15.1	-71.6
atax	5.2	73.2	231.6	393.2	0.0	-0.2	43.3	90.6	-0.9	3.5	140.7	62.3	0.6	-0.3	90.2	81.4
bicg	6.4	72.8	147.4	411.4	0.5	0.6	21.9	87.1	-1.1	-3.5	-16.8	100.9	-0.3	-1.3	81.1	
cholesky	5.6	8.2	15.9	277.3	-0.9	-1.0	62.6	78.4	-1.3	-0.5	-0.8	79.5	-0.2	-0.2	76.2	30.6
corr	60.1	12.0	2.3	195.2	0.1	0.4	83.4	104.2	4.0	6.5	87.6	18.1	2.1	1.8	18.6	12.4
covariance	2.8	17.2	79.2	327.8	0.3	0.6	94.6	98.1	-1.0	-0.6	72.0	11.3	-1.1	0.2	14.7	-0.4
deriche	0.1	4.0	120.2	174.9	-19.7	-22.4	60.0	63.0	-35.7	-33.6	30.9	45.1	-29.9	-32.7	72.7	101.0
doitgen	4.7	17.8	94.7	254.3	-1.0	-1.3	42.9	68.7	2.8	1.2	17.0	-24.3	0.1	-0.6	3.6	19.3
durbin	5.8	4.7	303.3	291.3	-0.6	0.4	21.6	16.1	21.6	29.3	28.7	-27.4	-1.3	-1.3	20.2	-28.2
fdtd-2d	1.3	5.7	27.7		-1.6	-1.6	98.4	88.1	-5.7	-0.8	30.1	-14.5	4.4	2.7		48.1
f oyd-warshall	16.6	-22.2	344.0	282.1	-0.1	-0.1	47.7	46.9	2.3	-16.8	13.4	111.5	1.5	3.2	3.3	-30.1
gemm	4.6	-7.4	166.5	158.8	0.1	0.2	61.5	75.4	1.5	0.4	1.4	-53.8	-4.4	-0.9	-56.4	-55.4
gemmver	7.3	36.5	133.7	399.6	-1.6	-1.1	64.5	60.7	-12.7	0.1		63.7	0.1	-1.8	79.3	51.6
gesummv	6.5	110.1	409.7	420.4	-59.5	-59.6	-0.0	1.5	0.3	7.6	150.4	86.0	31.2	26.7		
gramschmidt	-7.5	-0.0	-8.9	287.8	1.5	1.6			-43.7	-44.6	-44.9	-32.9	-31.3	-32.1	-6.7	-27.6
heat-3d	1.6	2.0			-1.5	-1.4	62.9	45.0	-1.4	-1.0	96.7	95.8	-0.2	-0.2	23.2	18.9
jacobi-1d	-0.6	1.6			0.1	0.1	92.4	80.0	-14.1	-14.1	-3.6	-49.0	-6.9	-6.9	-4.1	-41.6
jacobi-2d	1.8	5.2			-0.5	-0.8	71.5	88.5	-1.3	-1.2	9.4	-35.5	0.3	-1.1	11.6	-34.3
lu	5.6	7.8	12.3	304.0	0.2	0.5	74.7	88.7	-4.3	-3.8		64.3	-0.3	0.2	74.7	29.3
ludcmp	0.6	5.5	29.6	272.1	0.8	1.1	83.9	83.8	1.8	4.4		88.7	0.1	-0.0	70.9	34.0
mvt	7.0	62.2	112.2	308.7	0.0	0.0		119.0	2.3	-15.6	155.0	121.5	-1.1	2.0	75.7	
nussinov	-5.5	-42.4	229.3	224.1	-0.1	0.0	118.5	112.9	-1.7	12.6	13.9	-20.2	-0.1	-44.8	15.7	75.5
seidel-2d	2.4	40.1			-13.1	-13.1			-1.5	-2.1	23.3	53.2	-27.1	-27.1	-17.0	76.9
symm	1.5	16.6	432.3	434.0	0.5	0.4	111.7	112.1	0.3	-0.4	-2.8	5.6	-2.2	-1.1	2.7	-0.1
syr2k	2.1	16.8			0.1	0.3	65.4		0.5	-2.0	-2.3	-36.1	3.2	3.5	-41.4	-22.5
syrk	4.0	23.1	385.0	400.1	0.3	0.4			-1.9	1.5	-1.2	-40.0	5.1	0.6	60.3	-5.9
trisolv	7.0	78.3	156.2	346.9	-0.0	-0.0		101.9	-1.5	9.2	6.0	123.7	0.4	1.4		
trmm	5.0	27.1	355.3	352.1	-0.3	-17.8			-0.0	1.8	0.3	-12.0	-0.8	-3.7	16.5	-6.9

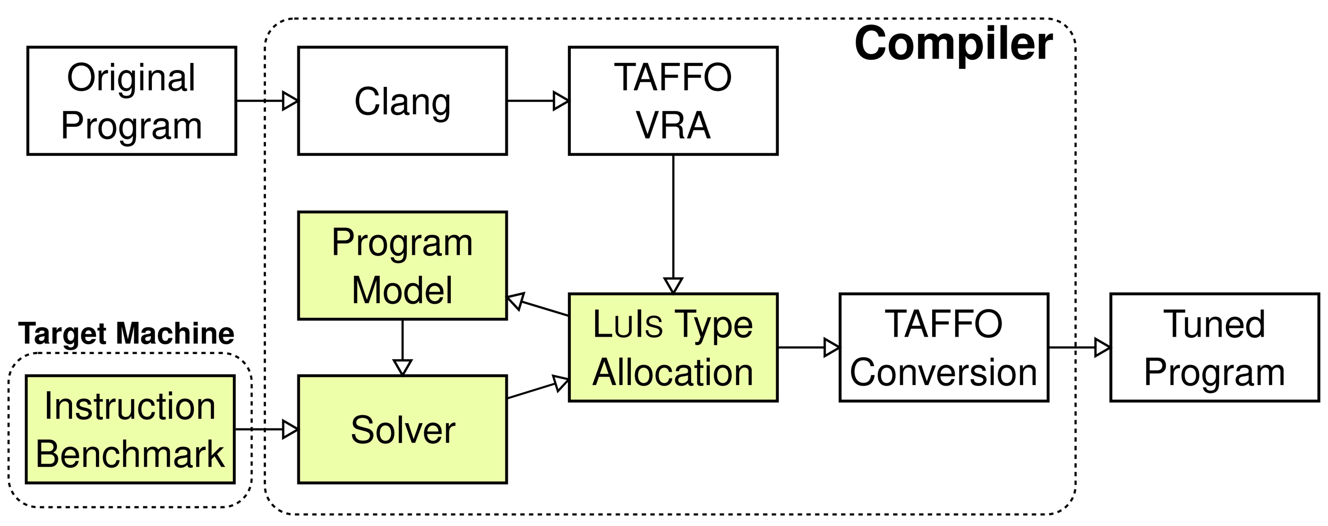
.... and we are looking for new, interesting, and challenging application scenarios

01011
10110

Our Solution

Tuning Assistant for Floating point and Fixed point Optimizations

TAFFO



based on industry-grade LLVM compiler technology

+ Modular
+ Efficient
+ Extendable

Works for all C-like languages

Read text abstract online