Faster RooFitting: Automated parallel calculation of collaborative statistical models

E G Patrick Bos[1], **Carsten D Burgard**[2], **Vincent A Croft**[3], **Inti Pelupessy**[1], **Jisk J Attema**[1] **and Wouter Verkerke**[2]

[1] *Netherlands eScience Center*, Amsterdam, Netherlands
[2] *ATLAS group*, *Nikhef*, Amsterdam, Netherlands
[3] *Dept. of Physics and Astronomy*, *New York University*, New York, USA

E-mail: `p.bos@esciencecenter.nl`

Abstract. RooFit [1, 2] is the main statistical modeling and fitting package used to extract physical parameters from reduced particle collision data, e.g. the Higgs boson experiments at the LHC [3, 4]. RooFit aims to separate particle physics model building and fitting (the users' goals) from their technical implementation and optimization in the back-end. In this paper, we outline our efforts to further optimize this back-end by automatically running parts of user models in parallel on multi-core machines. A major challenge is that RooFit allows users to define many different types of models, with different types of computational bottlenecks. Our automatic parallelization framework must then be flexible, while still reducing run-time by at least an order of magnitude, preferably more. We have performed extensive benchmarks and identified at least three bottlenecks that will benefit from parallelization. We designed a parallelization layer that allows us to parallelize existing classes with minimal effort, but with high performance and retaining as much of the existing class's interface as possible. The high-level parallelization model is a task-stealing approach. Preliminary results show speed-ups of factor 2 to 20, depending on the exact model and parallelization strategy.

1. Introduction

RooFit is a tool used in large collaborations of hundreds of physicists to fit large statistical models to data coming from particle accelerator experiments. Streamlining the model fitting process is crucial for increasing the productivity of such collaborations. When a model takes only minutes to verify instead of hours, the user can remain focused on the issue at hand instead of having to switch into and out of context again and again. In addition, faster run-times would allow fitting models with much more parameters to larger datasets — necessary to investigate the next generation of particle physics models, like e.g. Effective Field Theory models of the Higgs boson — leading to more precise results, or even completely new findings, like models of dark matter or super-symmetry.

2. RooFit performance bottlenecks

To gauge current performance of RooFit and to identify the most promising optimization targets, we ran a benchmark on both realistic particle physics models and a set of representative toy models[1]. Apart from two key serial optimization opportunities, namely vectorization and memory access pattern optimization [5], no obvious further optimization target was identified without parallelization. In particular, we identified three major bottlenecks that could benefit greatly from parallelization:

(i) Gradient calculation (parameter partial derivatives) in the Minuit2 Migrad minimizer;

(ii) Likelihood evaluation, which is a sum over PDF components evaluated for events; parallelization can happen both over events, scaling with data volumes, and over (unequal) components, scaling with model parameters;

(iii) Integrals (normalization) and other expensive shared components.

Which of these bottlenecks are actually relevant depends very much on the user's specific model. In some cases, parallelization of one type of "bottleneck" may lead to slower run-times due to increased overhead. This calls for the implementation of multiple strategies that can be activated or deactivated depending on the model at hand.

In this paper, we focus on our implementation of the gradient level parallelization strategy. This strategy speeds up fits of likelihoods (or other test statistics) with a large number of parameters, which is the case for the ATLAS and CMS Higgs combination fits. Each fit parameter corresponds to a numerical partial derivative calculation in `Migrad`, and these partial derivatives can be calculated in parallel. In this way, we speed up the most time consuming part of the Migrad minimization procedure [6], the gradient step. For N parameters, this step involves $2N$ test statistic evaluations. The second most expensive item, the line-search step between gradient steps, takes only a few test statistic evaluations. Note that speeding up the test statistic would speed up both steps. However, this is much more complicated due the wide range of possible test statistics. In contrast, given a sufficient number of model parameters (sufficient being a multiple of the number of available CPU cores), the strategy of parallelizing the gradient in the minimizer will always yield performance improvements. This is why we chose to initially focus on this strategy.

3. Parallel design

In order to support multiple strategies for the parallelization of RooFit models, we designed a generic framework, `RooFit::MultiProcess`, that we expect to be close to optimal, flexible and automatic by default. The basis of the framework is a work-stealing approach.[2] This approach performs and scales near optimal in general cases [7, 8]. For each parallelizable task, a number of sub-tasks or jobs is defined and sent to a queue process that handles bookkeeping of these tasks. A pool of workers subsequently "steals" the jobs from the queue process. Each worker only gets one job at a time and returns the result to the queue when it's done. Then the worker will request a new job, until the queue runs out of jobs. This system automatically balances the unequal loads that jobs in the heterogeneous tasks like composite likelihood calculations or partial derivatives creates. Communication between processes is done by message passing using ZeroMQ [9].

To make the implementation flexible and easily extensible to possible new future bottlenecks, we designed the framework in such a way that it is itself independent of existing code, but can be applied as a thin layer over C++ classes in the existing code. This means that the interface of the existing RooFit classes can still be used. Our layer provides the low-level toolkit necessary to easily build a parallelized version of the class.

We aim to provide a smooth transition for users by ensuring that all algorithms implemented in `RooFit::MultiProcess` produce the exact bit-wise identical results as the previous

[1] The benchmarks can be found in our GitHub repositories at `https://github.com/roofit-dev/parallel-roofit-scripts` and `https://github.com/roofit-dev/rootbench`.

[2] Thanks to eScience Center efficient computing expert Jason Maassen.

algorithms. One example is the transformation to `Minuit2` internal parameters, which involves trigonometric functions that cause rounding differences. For more design details, we refer to [10].

4. Results

We next present the results of benchmarks run using our implementation of a gradient-level parallelization strategy in the new `RooFit::MultiProcess` framework. This method was benchmarked on two realistic models:

(i) *Fast* model: a gluon fusion Higgs boson production model on an Asimov data set [11]. This has 13795 likelihood components and 265 parameters. A fit on this model runs in about 20 seconds – our main target is to speed up longer running benchmarks, but we used this model for getting good statistics on the timing data, which inevitably varies due to external influences, like operating system or other background activity.

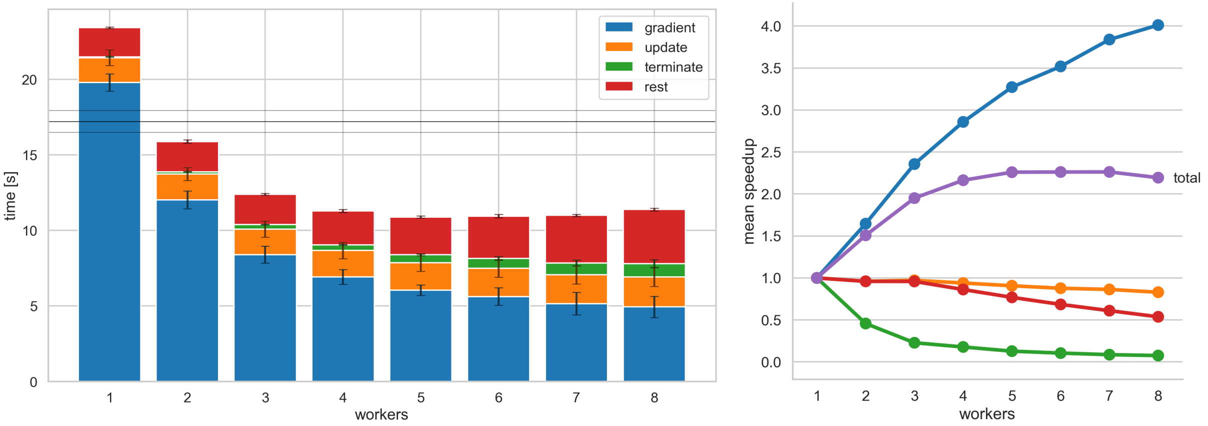
(ii) *Big* model: ATLAS Higgs combination fit [12]. This model has 126883 likelihood components and 1487 parameters. In a realistic scenario, where the starting point of the fitter is not close to the actual minimum, this model fits in a few hours.

We ran the benchmarks on a CentOS 7 node of the Stoomboot cluster at Nikhef. The node runs on an AMD EPYC 7551P 32-core CPU, with 256 GB RAM, which is plenty for our purposes. No other users could use the node at the same time, so the impact of concurrently run programs is minimized to only processes run by the OS.

As per our design (previous section), our fit results using the new parallel framework are exactly the same as those that come out of using the serial RooFit routines. For further physics validation of the models we refer to the respective cited references.

4.1. Fast model results

The fast model fit runs in about 17 seconds with the old `RooFit::RooMinimizer` class that just runs serially in a single process, indicated by the black horizontal line in figure 1a. As figure 1 further shows, the single worker `MultiProcess` run is slower, averaging at 23 seconds. This is in part due to communication (the orange "update" component), which the `RooMinimizer` does not have to bother with, but also clearly the gradient calculation itself was slower in our benchmarks, since it is slower than the entire minimization. We did not measure the `RooMinimizer` separately in terms of these components, so we must partly speculate as to the precise cause of the differences, but we suspect that also the rest term (i.e. mainly the line search step) runs slightly faster in the old situation compared to the single worker situation. We found that this is largely due to the fact that RooFit function calls use a highly efficient memoization mechanism. This mechanism stores calculated parts of the likelihood's expression tree and only recalculates those parts when the parameters that that specific subtree depends on change. However, these cached values are not synchronized between the workers and the master process. Since the main process does the line-search step and the workers do the gradient steps, and parameters change in between these steps, the cache is effectively thrown away each time the work load switches from the master process to the workers and the other way around. Compared to the old `RooMinimizer`, this causes a slight delay both in the master process and in the workers at the start of each step. These effects lead these fast runs to experience a high degree of "overhead", i.e. a lack of perfect scaling. In fact, beyond 8 cores, the lack of further scaling, but growth of the rest term, leads to anti-scaling, i.e. slower wall clock times with increasing number of workers. This can be seen most clearly in figure 1b, specifically in the purple line that represents the speed-up for the total run with respect to the single worker run. Despite this, a speed-up of a factor 2.5 can be achieved with 4 workers on a "fast" run like this.

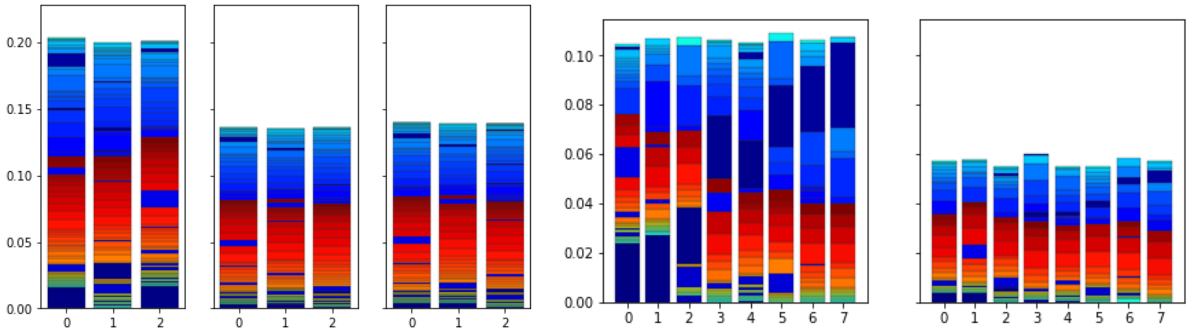


(a) Wall clock time of runs in seconds.

(b) Speed-up of the runs compared to a single worker run.

Figure 1: Fast model wall clock run times for runs with increasing number of workers on the horizontal axis. For each number of workers, the fit was repeated 10 times to get both mean run time — indicated by the height of the bars — and standard deviation — indicated by the black error bars on each histogram bar. Separately measured components of the run time are colored as indicated in the legend: *gradient* calculation time, *update* time of parameters between processes, *terminate* time at the end of a run (shutting down ZeroMQ sockets and context and the forked processes) and the rest of the run time (in these runs this includes the line-search phase). For reference, the black horizontal line at about 17 seconds indicates the mean run time of the old `RooFit::RooMinimizer` class, while the surrounding two grey lines indicate those runs' standard deviation.

4.2. Load balancing



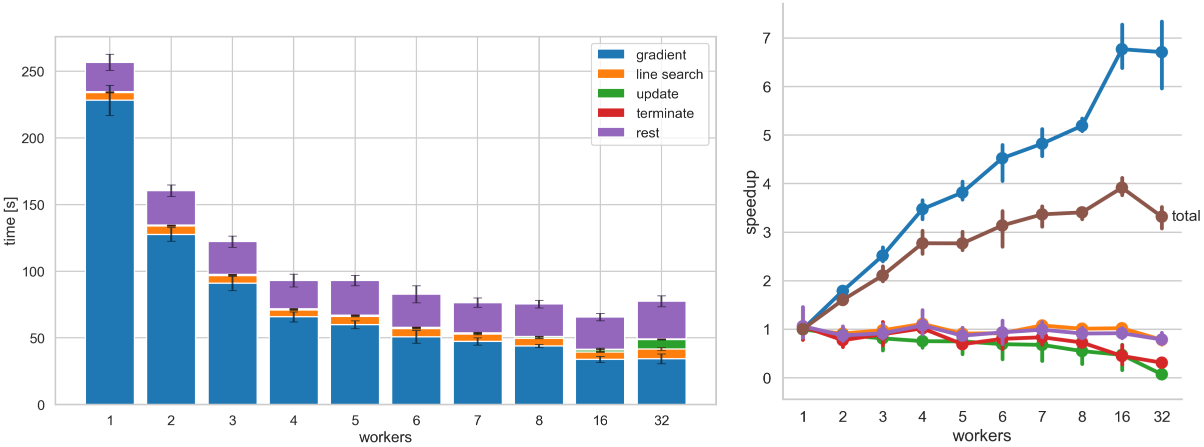
(a) Three iterations of a three worker run.

(b) Two iterations of an eight worker run.

Figure 2: Load balancing of our work stealing algorithm. Each panel represents one gradient calculation. Each gradient calculation consists of 265 partial derivatives, each of which is shown as a differently colored stacked "sub"-bar. The three or eight main bars each represent work done on one of the workers used in that run. Vertical axis shows wall clock time in seconds.

One might suspect that waiting time in-between partial derivative calculations on the workers could be a delaying factor as well, but we confirmed that this was not the case in any significant way. In addition, we investigated whether a sub-optimal load balancing of the partial derivatives over the workers could be causing the sub-optimal scaling. This analysis for a single fast model run is illustrated in figure 2. We see that for three workers (panel 2a), the load for each gradient is, in fact, very well balanced over the workers. In the case of the eight workers (panel 2b), the idle times of some workers that are waiting for the slowest worker becomes more noticeable. We measured that on average this costs about 2% of the run time with 8 workers on the big model run. All in all, we can conclude that the dynamic load balancing of our work stealing approach is efficient.

4.3. Big model results



(a) Timing

(b) Speed-up

Figure 3: Big model benchmark results. In this run, for each number of workers, the fit was repeated only 3 times and we additionally measured the line-search phase separately. See the caption of figure 1 for further details.

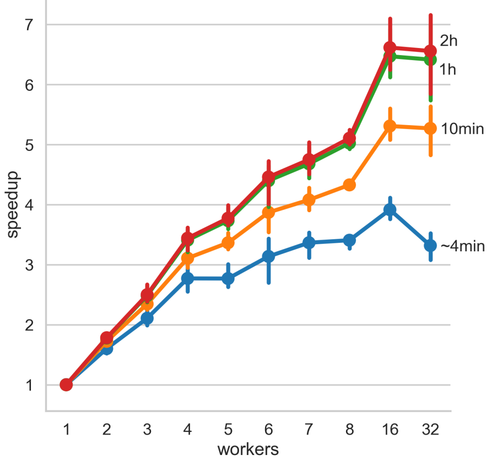


Figure 4: Big model benchmark results for performance models of longer total fitting run times.

Figure 3 shows the main timing results on the big model. Due to time constraints we ran this model with initial starting parameters very close to the actual minimum, leading to only 10 gradient steps per minimization run. We find in this case that a speed-up of a factor 4 can be achieved with 7 workers. The update and termination times seem to have become insignificant in these longer runs. The rest term, on the other hand, plays a major role in keeping the model from scaling. Further analysis revealed, however, that this component happens only once at the beginning of a minimization run. It it caused by the high number of constant terms in this model and the current inefficient implementation of the synchronization of these terms between RooFit and Minuit. Apart from this, the line-search step, which here we do measure independently of the rest term, turns out not to be insignificant either, although at least it remains constant, since it is calculated independently of the workers.

In a typical realistic fit, many more gradient steps would be performed, since the initial guesses of the parameters will not be close to the true minima. To see how our above results generalize to such a more realistic scenario, we also ran several times with starting guesses further from the minimum. We found that the rest and terminate components stay constant within the expected variance. Since these are only one time costs that do not scale with the number of steps (whereas the gradient, update and line search components do), we can easily construct a performance model for longer, more realistic runs. In figure 4 we show these performance model results for three run times: 10 minutes, 1 hour and 2 hours. The latter two are, in fact, the actual realistic range of single core run time using real Run 2 data [12]. We show that using 16 workers (possibly less, since we did not measure any amount of workers between 8 and 16) one can achieve an average total run time speed-up of a factor 6.5.

5. Discussion

Our parallelized gradient method achieves a factor seven speed-up on our main target of big models. The exact speed-up varies slightly, but not significantly from run to run. Communication between the processes causes part of this variable overhead, since we currently synchronize all parameters from and to all nodes after each run, amounting to ~ 1000 numbers being transferred between N processes for each gradient call. This is necessary because the gradient algorithm self-adapts its precision based on the minimizer's search progress. We could reduce the required communication by two orders of magnitude by pinning partial derivatives to specific workers, since the adaptive precision for each derivative component only depends on that component itself. This trade-off of flexibility in dynamic load balancing (which would be lost when pinning gradient components to specific workers) versus reduced communication could be implemented as an alternative strategy. Both strategies may prove useful in different cases.

The framework is currently available in the ROOT fork in the RooFit development GitHub page at `https://github.com/roofit-dev/root/tree/MP_ZeroMQ`. We warn that it should not be considered production-ready. Once ready, it will be included in an upcoming official ROOT release. The authors are in close contact with the ROOT developers team to coordinate this effort.

Acknowledgments

This project was funded by the Netherlands eScience Center. The benchmark results were analyzed and visualized using Python modules NumPy [13], Matplotlib [14], Seaborn [15] and Pandas [16] in Jupyter notebooks [17].

References

[1] Verkerke W and Kirkby D 2003 *ArXiv Physics e-prints* (*Preprint* `physics/0306116`)

[2] Moneta L, Cranmer K, Schott G and Verkerke W 2010 *Proceedings of the 13th International Workshop on Advanced Computing and Analysis Techniques in Physics Research. February 22-27, 2010, Jaipur, India.* p 57 (*Preprint* `1009.1003`)

[3] ATLAS and CMS Collaborations 2015 *Phys. Rev. Lett.* **114** 191803 (*Preprint* `1503.07589`)

[4] ATLAS and CMS Collaborations 2016 *JHEP* **08** 045 (*Preprint* `1606.02266`)

[5] Hageböck S 2019 *J. Phys. Conf. Ser.: ACAT 2019*

[6] James F and Roos M 1975 *Computer Physics Communications* **10** 343–367

[7] Blumofe R D and Leiserson C E 1994 *Proceedings of the 35th Annual Symposium on Foundations of Computer Science* SFCS '94 (Washington, DC, USA: IEEE Computer Society) pp 356–368 ISBN 0-8186-6580-7 URL `https://doi.org/10.1109/SFCS.1994.365680`

[8] Van Nieuwpoort R V, Wrzesińska G, Jacobs C J H and Bal H E 2010 *ACM Trans. Program. Lang. Syst.* **32** 9:1–9:39 ISSN 0164-0925 URL `http://doi.acm.org/10.1145/1709093.1709096`

[9] Hintjens P 2013 *ZeroMQ: Messaging for Many Applications* (O'Reilly Media)

[10] Bos E G P, Pelupessy I, Croft V A, Verkerke W and Burgard C D 2018 *2018 IEEE 14th International Conference on e-Science (e-Science)* pp 345–346

[11] Aaboud M *et al.* 2019 *Physics Letters B* **789** 508 – 529 ISSN 0370-2693 URL `http://www.sciencedirect.com/science/article/pii/S0370269318309936`

[12] ATLAS Collaboration 2019 Combined measurements of Higgs boson production and decay using up to 80 fb^{-1} of proton–proton collision data at $\sqrt{s} = 13$ TeV collected with the ATLAS experiment Tech. Rep. ATLAS-CONF-2019-005 CERN Geneva URL `https://cds.cern.ch/record/2668375`

[13] van der Walt S, Colbert S C and Varoquaux G 2011 *Computing in Science Engineering* **13** 22–30 ISSN 1521-9615

[14] Hunter J D 2007 *Computing in Science Engineering* **9** 90–95 ISSN 1521-9615

[15] Waskom M, Botvinnik O, O'Kane D, Hobson P, Ostblom J, Lukauskas S, Gemperline D C, Augspurger T, Halchenko Y, Cole J B, Warmenhoven J, de Ruiter J, Pye C, Hoyer S, Vanderplas J, Villalba S, Kunter G, Quintero E, Bachant P, Martin M, Meyer K, Miles A, Ram Y, Brunner T, Yarkoni T, Williams M L, Evans C, Fitzgerald C, Brian and Qalieh A 2018 mwaskom/seaborn: v0.9.0 (july 2018) URL `https://doi.org/10.5281/zenodo.1313201`

[16] McKinney W 2010 *Proceedings of the 9th Python in Science Conference* pp 51–56

[17] Kluyver T, Ragan-Kelley B, Pérez F, Granger B, Bussonnier M, Frederic J, Kelley K, Hamrick J, Grout J, Corlay S, Ivanov P, Avila D, Abdalla S and Willing C 2016 *Positioning and Power in Academic Publishing: Players, Agents and Agendas* ed Loizides F and Schmidt B (IOS Press) pp 87 – 90