

Making extreme computations possible with virtual machines

J Reuter¹, B Chokoufe Nejad¹, T Ohl²

¹ DESY Theory Group, Notkestr. 85, D-22607 Hamburg

² University of Würzburg, Emil-Hilb-Weg 22, D-97074 Würzburg

E-mail: juergen.reuter@desy.de, bijan.chokoufe@desy.de, ohl@physik.uni-wuerzburg.de

Abstract. State-of-the-art algorithms generate scattering amplitudes for high-energy physics at leading order for high-multiplicity processes as compiled code (in Fortran, C or C++). For complicated processes the size of these libraries can become tremendous (many GiB). We show that amplitudes can be translated to byte-code instructions, which even reduce the size by one order of magnitude. The byte-code is interpreted by a Virtual Machine with runtimes comparable to compiled code and a better scaling with additional legs. We study the properties of this algorithm, as an extension of the Optimizing Matrix Element Generator (O'Mega). The bytecode matrix elements are available as alternative input for the event generator WHIZARD. The bytecode interpreter can be implemented very compactly, which will help with a future implementation on massively parallel GPUs.

1. Introduction

Meta-programming is a popular approach in high-energy physics to determine the expression of a cross section in a higher level programming language while the numerical evaluation is performed in high-performance languages. A problem, however, arises when the expression becomes so large that it is impossible to compile and link, and hence to evaluate numerically, due to the sheer size. In **Fortran**, which is known for its excellent numerical performance, we typically encounter this problem for source code of gigabyte sizes irrespective of the available memory. In these proceedings, we sketch the ideas of Ref. [1], how to completely circumvent the tedious compile step by using a virtual machine (VM). To avoid confusions, we note that a VM is in our context a compiled program, an *interpreter*, that is able to read instructions, in the form of *byte-code*, from disk and perform an arbitrary number of operations out of a finite *instruction set*. A VM allows the complexity of the computation to be only set by the available hardware and not limited by software design or intermediate steps. Furthermore, it is easy to implement and makes parallel evaluation obvious.

An important concern is of course whether the VM can still compete with compiled code in terms of speed. The instructions have to be translated by the VM to actual machine code, which is a potential overhead. However, in the computation of matrix elements a typical instruction will correspond to a product of scalar, spinor or vector currents which involves $O(10)$ arithmetical operations on complex numbers. This suggests that the overhead might be small, which has however to be proven by a concrete implementation. In fact, we will show that a VM can even be faster than compiled code for certain processes and compilers since the formulation in terms of a VM has also benefits, especially for large multiplicities, as is discussed in detail below. More importantly, the runtime is in general in the same order of magnitude as the compiled code and as such the VM is very usable for general purpose applications, where the clever use of Monte Carlo (MC) techniques can easily change the number of points needed for convergence by orders of magnitude.

2. General Virtual Machines

One might imagine a VM as a machine, which has a number of registers, and is given instructions, encoded in the byte-code, how to act on them. This picture is quite similar to a CPU, except that we are doing this on a higher level, i.e. our registers are arrays of e.g. wave functions or momenta and the instructions can encode scalar products or more complicated expressions. In order to use VMs for arbitrary processes, a dynamic construction is desirable. For such a construction, it is necessary to include a *header* in the byte-code, which contains the number of objects that have to be allocated. After this the body of instructions follows, whereby each line corresponds to a certain operation that the VM should perform on its registers. The first number of an instruction is the *operation code (opcode)* that specifies which operation will be performed. For illustration, consider the example 1 5 4 3, which could be translated into $\text{momentum}(5) = \text{momentum}(4) + \text{momentum}(3)$, a typical operation to compute the s -channel momentum in a $2 \rightarrow 2$ scattering process. Depending on the context, set by the opcode, the following numbers have different meanings but are typically addresses, i.e. indices of objects, or specify how exactly the function should act on the operands, by what numbers the result should be multiplied, etc. The interpreter is a very simple program that reads the byte-code into memory and then loops over the instruction block with a `decode` function, which is basically a `select/case` statement depending on the opcode. The instructions can be instantly translated to physical machine code, compared to the execution time of the relevant instructions, since the different types of operations are already compiled and only the memory locations of the objects have to be inserted. The byte-code file that is given to the interpreter completely dictates the specific problem, or process in the cross section context, that should be computed. Input data or external parameters are given as arguments to the function call of the VM. The generation of events for collider physics usually parallelizes trivially. Since an integral is in most cases needed, the same code is just evaluated multiple times with different input data. The situation can change, however, for an extreme computation that already uses all caches. Depending on the size of the caches and the scheduler, evaluating such code with multiple data at the same time, can run even slower than the single-threaded execution. Obviously, the computation is then so large, containing numerous objects, that it is worth trying to parallelize the execution with a single set of input data with shared memory.

The byte-code lends itself nicely to parallelization as it can be split into recursion *levels*, whereby in each level all *building blocks* are non-nested and may be computed in parallel. Different levels are separated by necessary synchronization points. It is clear that one should aim to keep the number of synchronization points to the inherent minimum of the computation for optimal performance. This parallelization is straightforward in `Fortran95/2003` and `OpenMP` and shown explicitly in Ref. [1].

As a side note, we mention that the sketched parallelization should be very well suited for an implementation on a graphics processing unit (GPU). A common problem, encountered when trying to do scientific computing on a GPU, is the finite kernel size problem. As noted e.g. in Ref. [2], large source code cannot be processed by the `CUDA` compiler, which is related to the fact that the numerous cores on a GPU are designed to execute simple operations very fast. Dividing an amplitude into smaller pieces, which are computed one by one, introduces more communication overhead and is no ultimate solution since the compilation can still fail for complex amplitudes [2]. The VM on the other hand is a fixed small kernel, no matter how complex the specific computation is. A potential bottleneck might be the availability of the instruction block to all threads, but this question has to be settled by an implementation and might have a quite hardware-dependent answer.

Finally, we note that the phase-space parallelization mentioned in the beginning of this subsection can still be applied. When considering heterogeneous cluster or grid environments, where each node is equipped with multi-core processors, a combination of distributed

memory parallelization for the combination of different phase-space points and shared memory parallelization of a single point seems to be a quite natural and extremely powerful combination.

3. O’Mega Virtual Machine

The concept of a VM can be easily applied to evaluate tree-level matrix elements of arbitrary multiplicity. The Optimizing Matrix Element Generator, O’MEGA [3], avoids the redundant representation of amplitudes in the form of Feynman diagrams by using one-particle off-shell wave functions (1POWs) recursively. O’MEGA tames herewith the computational growth with the number of external particles from a factorial to an exponential one.

The model-independence is achieved in O’MEGA with the meta-programming ansatz mentioned earlier whereby the symbolic representation is determined in OCaml. This abstract expression is then translated to valid Fortran code that is automatically compiled and used in WHIZARD [4–6] for event generation. We replace the last step now with an output module that produces byte-code instead of Fortran code.

The number of distinct operations that have to be performed in the computation of a cross section is related to the Feynman rules and therefore quite limited. As such, these operations are very good candidates for the translation to byte-code. In fact, this results in only about 80 different opcodes for the complete standard model (SM), which have been implemented in the O’Mega virtual machine (OVM). For the parallel execution, we identify the different levels by the number of external momenta a wave function is connected to or equivalently the number of summands in the momentum of the wave function.

4. Speed Benchmarks

All processes shown here, and various others, have been validated against the compiled versions where possible. We stress that every process is computed in its respective model (QCD or SM) to *full* tree-level order including all interferences and we have not restricted e.g. the Drell-Yan amplitudes to only one electroweak propagator. We use the term SM for the full electroweak theory together with QCD and a nontrivial Yukawa matrix but without higher dimensional couplings like $H \rightarrow gg$.

To investigate the compiler dependence of the results, we use two different compilers, `gfortran 4.7.1` and `ifort 14.0.3`. We do not claim that our results are necessarily representative for all Fortran compilers or even compiler versions, but they should still give a good impression of the expected variance in performance. The evaluation time measurements are performed on a computer with two Intel(R) Xeon(R) E5-2440 @ 2.40GHz central processing units (CPUs), having 16 MiB L3 cache on each socket, and 2x 32 GiB RAM running under Scientific Linux 6.5. In Figure 1, we show the measured CPU times for QCD and SM processes with two different optimization levels for the compiled code and the OVM using the GNU and Intel compiler. Since the evaluation times are highly reproducible, we use only three runs to obtain mean and standard deviation. The times are normalized for *each* process to `gfortran-03`, which is why the times are not growing with the number of particles. For `gfortran`, we observe for most processes the fastest performance with `-03` and for `ifort` with `-02`, which is an effect commonly encountered. The fastest performance is given by the source code compiled with `ifort-02` being roughly 0.75 times the time needed by `gfortran-03`.

The crucial point, however, is that `ifort` fails to compile the $n = 7$ gluon and the $u\bar{u} \rightarrow e^+e^-6j$ Drell-Yan process while the OVM immediately starts computing. The GNU compiler is usually able to compile one multiplicity higher compared to the Intel before breaking down. This fits together with the better performance of the compilable processes and longer compile times as `ifort` seems to apply more optimizations.

Another interesting observation is that the OVM gets faster compared to the compiled code with increasing multiplicity of external particles though this feature is more pronounced in SM

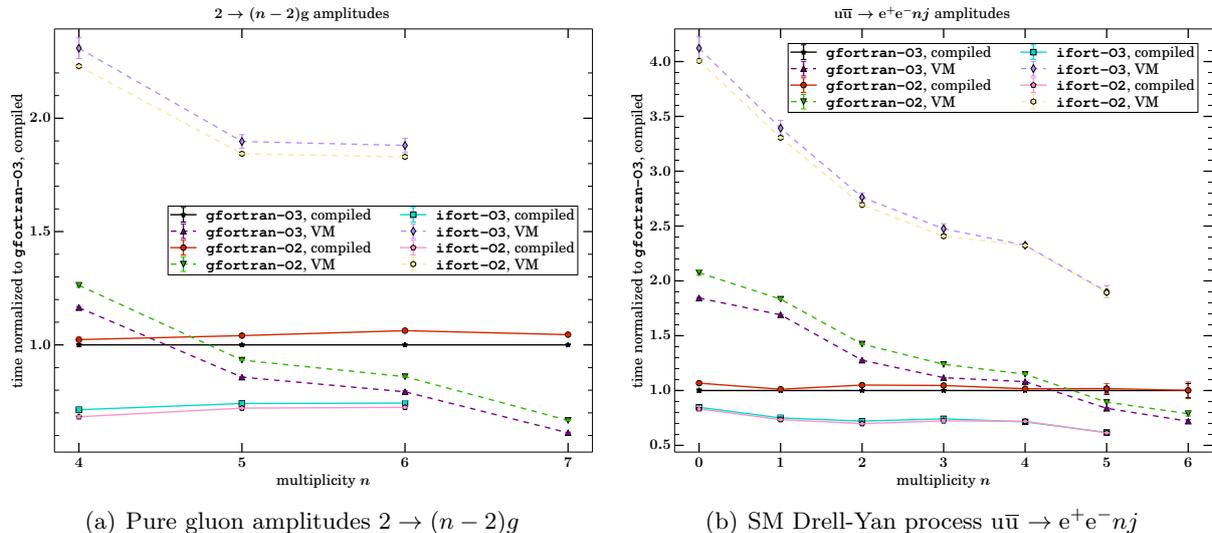


Figure 1. CPU times measured with the Fortran intrinsic `cpu_time` and normalized for each process to the compiled source code using `gfortran -O3`. Dashed (solid) lines represent the OVM (compiled source code).

and QCD processes. As discussed in Ref. [1], this is not due to costs from initialization or virtualization but likely caused by the explicit loop over the instructions in the VM. It gives a higher probability to keep the decode function in the instruction cache compared to the compiled code. We observe roughly the same effect for both compilers, but the OVM compiled with `ifort` is about a factor of two slower than the version with `gfortran`. This is probably solvable with a profile-guided optimization, which allows the compiler to use data instead of heuristics to decide what and how to optimize. Amdahl’s idealized law [7] simply divides an algorithm into parallelizable parts p and strictly serial parts $1 - p$. Therefore, the possible speedup s for a computation with n processors is

$$s(n) \equiv \frac{t(1)}{t(n)} = \frac{1}{(1-p) + \frac{p}{n}}. \quad (1)$$

Communication costs between processors $\mathcal{O}(n)$ have been neglected hereby in the denominator of eq. 1. This means that we have $\lim_{n \rightarrow \infty} s(n) = 1/(1-p)$ in the idealized case and $\lim_{n \rightarrow \infty} s(n) = 0$ including communication costs. In reality, we are interested in high speedups for finite n and also have to care about efficient cache usage.

In Figure 2, we show the speedup with multiple cores, N , by either using the parallelization procedure discussed above to compute one amplitude in parallel or by computing multiple amplitudes for multiple phase-space points in parallel. In Figure 2 a), we can see that the $n = 7$ and $n = 8$ gluon amplitudes parallelize very well with both methods with parallelizable parts above 95%. In the shared memory parallel evaluation of the amplitude (A), the impact of the hardware architecture is quite obvious. For $N = 7$, i.e. when the second socket of the board is activated, we see a drop in efficiency. This relative drop is the stronger the higher the communication costs are compared to the calculation done in the individual threads and can thus be seen most cleanly for the $n = 6$ gluon and the $n = 4$ Drell-Yan processes. This is a cache coherency issue due to the synchronization of the caches of the two CPUs. We observe that the byte-code for the OVM is about one order of magnitude smaller. Some processes together with their compile times are shown in Figure 1. The smaller output format leads to less required

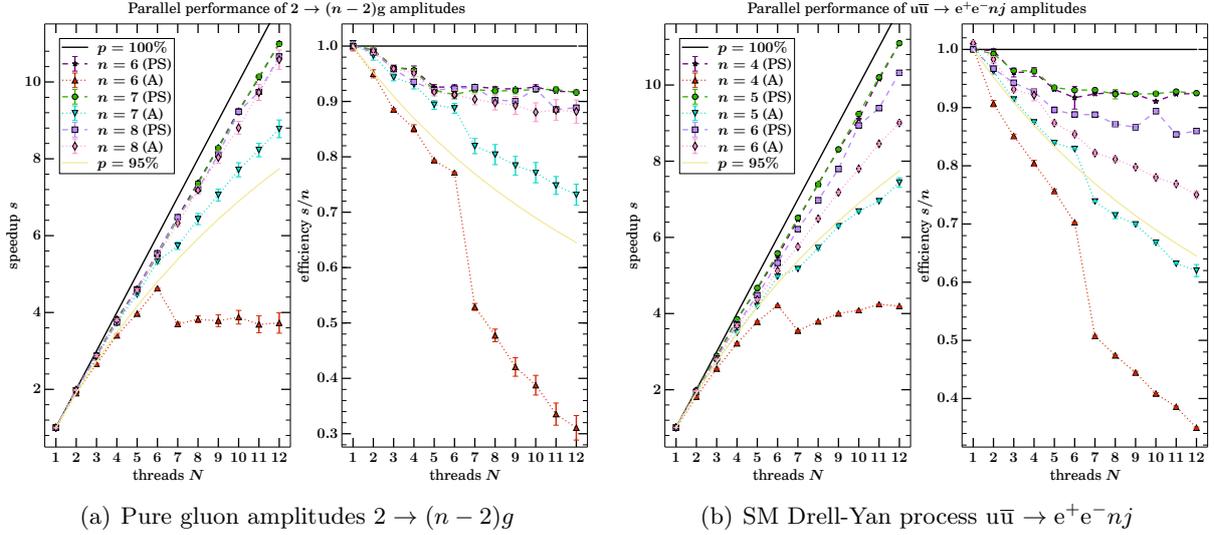


Figure 2. Speedup and efficiency to compute a fixed number of phase-space points for the parallel evaluation of multiple phase-space points (PS) and the parallel evaluation of the amplitude itself (A) are shown as dashed and dotted lines. The solid lines represent Amdahl's law for a fixed value of the parallelizable part p .

RAM and time to produce it. Especially for many color flows, where the generation time of O'MEGA is dominated by the output procedure, we observe e.g. for $gg \rightarrow 6g$ a reduction in memory from 2.17 GiB to 1.34 GiB and in generation time from 11 min 52 s to 3 min 35 s, while staying roughly the same for small processes.

Table 1. Size of the byte-code (BC) compared to the Fortran source code together with the corresponding compile time with gfortran. The compile times were measured on a computer with an i7-2720QM CPU. The $2g \rightarrow 6g$ process fails to compile.

process	BC size	Fortran size	t_{compile}
$gg \rightarrow gggggg$	428 MiB	4.0 GiB	—
$gg \rightarrow gggggg$	9.4 MiB	85 MiB	483(18) s
$gg \rightarrow q\bar{q}q'\bar{q}'q''\bar{q}''g$	3.2 MiB	27 MiB	166(15) s
$e^+e^- \rightarrow 5(e^+e^-)$	0.7 MiB	1.9 MiB	32.46(13) s

5. Summary and Outlook

A VM circumvents the compile and link problems that are associated with huge source code as it emerges from very complex algebraic expressions. VMs are indeed a viable option that is maintaining relatively high performance in the numerical evaluation of these expressions and allows to approach the hardware limits. In practice, a VM saves at least hours of compile time. The concept has been successively applied to construct the OVM that is now an alternative method to compute tree-level matrix elements in the publicly available package O'MEGA and can be chosen in WHIZARD with a simple option since version 2.2.3. Any computation can in principle be performed with a VM though the benefits are clearly in the regime of extreme

computations that run into compiler limits with the conventional method. Here, we have seen that VMs can even perform better than compiled code. Also the parallelization of the amplitude is for very complex processes close to the optimum.

It would be an interesting experiment to remove the virtualization overhead by using dedicated hardware that has the same instruction set as the OVM to compute matrix elements. The number of instructions corresponding to different wave function fusions and propagators is finite for renormalizable theories (including effective theories up to a fixed mass dimension) and implemented similarly in the various matrix element generators. If the authors can agree on a common set of instructions and conventions this machine could therefore be used by all those programs. Field programmable gate arrays (FPGAs) can serve as such a machine as they have comparable if not superior floating-point performance with respect to current microprocessors and the OVM and its instruction set is the first step to test the feasibility and potential gains of computing matrix elements in this environment.

Acknowledgments

JRR wants to thank the ACAT 2016 organizers for the beautiful venue in Valparaíso and a fantastic conference in Chile.

References

- [1] B. Chokoufe Nejad, T. Ohl, and J. Reuter. *Computer Physics Communications*, **196**: (2015), pp. 58–69. [1411.3834].
- [2] K Hagiwara, J Kanzaki, N Okamura, D Rainwater, and T Stelzer. *The European Physical Journal C*, **66**:3-4 (2010), pp. 477–492. [0908.4403].
- [3] M. Moretti, T. Ohl, and J. Reuter (2001). [hep-ph/hep-ph/0102195].
- [4] W. Kilian, T. Ohl, and J. Reuter. *The European Physical Journal C*, **71**:9 (2011), p. 1742. [0708.4233].
- [5] W. Kilian, T. Ohl, J. Reuter, and C. Speckner. *Journal of High Energy Physics*, **2012**:10 (2012), p. 22. [1206.3700v2].
- [6] W. Kilian, J. Reuter, S. Schmidt, and D. Wiesler. *Journal of High Energy Physics*, **2012**:4 (2012), p. 13. [1112.1039].
- [7] G. Amdahl. *AFIPS Conference Proceedings*, **30**: (1967), pp. 483–485.