45

Computing for Perturbative QCD

S. Höche, L. Reina, M. Wobisch

C. Bauer, Z. Bern, R. Boughezal, J. Campbell, N. Christensen, L. Dixon, T. Gehrmann, J. Kanzaki, A. Mitov, P. Nadolsky, F. Olness, M. Peskin, F. Petriello, S. Pozzorini, F. Siegert, D. Wackeroth, J. Walsh, C. Williams

45.1 Introduction

One of the main challenges facing the particle-physics community to date is the interpretation of LHC measurements on the basis of accurate and robust theoretical predictions. The discovery of a Higgs-like particle in Summer 2012 [1, 29] serves as a remarkable example of the level of detail and accuracy that must be achieved in order to enable a discovery [33, 34, 47]. Signals for the Higgs boson of the Standard Model (SM) are orders of magnitude smaller than their backgrounds at the LHC, and they are determined by quantum effects. Detailed calculations are therefore mandatory, and they will become even more necessary as we further explore the Terascale at the full LHC design energy.

Providing precise theoretical predictions has been a priority of the US theoretical particle-physics community for many years, and has seen an unprecedented boost of activity during the last ten years. With the aim of extracting evidence of new physics from the data, theorists have focused on reducing the systematic uncertainty of their predictions by including strong (QCD) and electroweak (EW) effects at higher orders in the perturbative expansion. This is particularly important as beyond Standard Model effects are expected roughly at the TeV scale. Typical decay chains of potential new particles would involve many decay products, several of which can be massive. The SM backgrounds are complex processes which call for highly sophisticated calculational tools in order to provide realistic predictions.

We have reached a time when no conceptual problems stay in the way of breaking Next-to-Leading Order (NLO) perturbative QCD calculations into standard modular steps and automate them, making them available to the worldwide LHC community. The same is true for matching NLO calculations and parton showers. It is implicit that automation will benefit greatly from a unified environment in which calculations can be performed and data can be exchanged freely between theorists and experimentalists, as well as from the availability of adequate computational means for extensive multiple analyses.

We nowadays see the frontier of perturbative calculations for collider phenomenology being both in the development and optimization of Next-to-Next-to-Leading Order (NNLO) QCD calculations, sometimes combined with EW corrections, and in the study of more exclusive signatures that requires resummation of logarithmically enhanced higher-order corrections to all orders. It is also conceivable that techniques for matching NNLO fixed-order calculations to parton-shower simulations will be constructed in the near to mid-term future. In all cases, the availability of extensive computational resources could be instrumental in boosting the exploration of new techniques as well as in obtaining very accurate theoretical predictions at a pace and in a format that is immediately useful to the experiments.

Type of calculation	CPU hours per project	projects per year
NLO parton level	50,000 - 600,000	10-12
NNLO parton level	50,000 - 1,000,000	5-6
Hadronic event generation	50,000 - 250,000	5-8
Matrix Element Method	$\sim$ 200,000	3-5
Exclusive jet cross sections	$\sim$ 300,000	1-2
Parton Distributions	$\sim$ 50,000	5-6

Table 45-1. *Summary of computing requirements for the projects in Sec. 45.4*

45.2 Main results and recommendations

This workshop provided a framework for implementing higher order calculations in a standardized computing environment made available by DOE at the National Energy Research Scientific Computing Center (NERSC) [62]. Resource requirements were determined for the calculation of important background and signal reactions at the LHC, including higher order QCD and EW effects. Prototypical results have been summarized in a white paper [9]. Resource requirements are also listed in Tab. 45-1.

Different High Performance Computing (HPC) environments were tested during this workshop and their suitability for perturbative QCD calculations was assessed. We find that it would be beneficial to make the national HPC facilities ALCF [7], OLCF [63] and NERSC [62] generally accessible to particle theorists and experimentalists in order to enable the use of existing calculational tools for experimental studies involving extensive multiple runs without depending on the computer power and manpower available to the code authors. Access to these facilities will also allow prototyping the next generation of parallel computer programs for QCD phenomenology and precision calculations.

The computation of NLO corrections in perturbative QCD has been automatized entirely. Resource requirements for NLO calculations determined during this workshop can thus be seen as a baseline that enables phenomenology during the LHC era. NNLO calculations are still performed on a case-by-case basis, and their computing needs can only be projected with a large uncertainty. It seems clear, however, that cutting edge calculations will require access to leadership class computing facilities.

The use of HPC in perturbative QCD applications is currently in an exploratory phase. We expect that the demand for access to HPC facilities will continue to grow as more researchers realize the potential of parallel computing in accelerating scientific progress. Long-term support of senior personnel, providing an interface between advanced computing research and application, may be required to fully exploit the potential of new technologies. At the same time, we expect growing demand for educating young researchers in cutting edge computing technology. It would be highly beneficial to provide a series of topical schools and workshops related to HPC in HEP. They may be co-organized with experimentalists to foster the creation of a knowledge base.

Large-scale distributed computing in Grid environments may become relevant for perturbative QCD applications in the near future. This development will be accelerated if Computing Grids can also provide access to HPC facilities and clusters where parallel computing is possible on a smaller scale. The Open Science Grid (OSG) [64] has taken first steps in this direction, and we have successfully used their existing interface. The amount of training for new users could be minimized if the OSG were to act as a front-end to the national HPC facilities as well as conventional computing facilities.

45.3 Technology review

Most state-of-the art perturbative QCD calculations are performed using Monte-Carlo (MC) methods. The advantage of this technique is the dimension-independent $1/\sqrt{N}$ convergence of the integral (where N denotes the number of sample points, or *events*, used in the integration), and the possibility to generate events with kinematics distributed according to the integrand itself. A further advantage is that the calculation can be parallelized trivially, by generating multiple sets of events with different random seeds and combining them. This possibility is exploited by experiments when running Monte Carlo event generators on the Worldwide LHC Computing Grid [69].

45.3.1 Parallel computing using MPI

Typical event generators need an initialization phase. During this stage, the a priori weights of adaptive MC algorithms like VEGAS [56] or the Multi Channel [54] are optimized for the particular calculation of interest. True parallel computing on tightly coupled systems is highly advantageous in this phase in order to quickly exchange data needed for the optimization between different compute nodes.

The benefits of Message-Passing-Interface (MPI) parallelization in this context has been assessed during this workshop by implementing MPI communication into a representative Monte-Carlo event generation framework. Similar strategies can be employed in other programs. For the purpose of this study we have chosen the parton-level event generator BlackHat+Sherpa [13, 43, 41]. On the Cray XE6™ System "Hopper" at NERSC we observe strong scaling up to 1,024 cores. This makes it possible to attempt calculations considered prohibitively time-consuming previously, and has been used in the computation of $pp \to W^{\pm}$+5 jets at NLO [15].

45.3.2 Parallel computing using multi-threading

In addition to MPI, multi-threading can be used to reduce the memory footprint of executables on large-scale parallel computing systems, which currently often suffer from a small memory per compute core. In this case, executables can be designed such that various independent parts of the calculation, like the evaluation of hard matrix elements and the corresponding phase space, are performed in parallel.

Multi-threading can also be used efficiently in recursive algorithms to compute matrix elements [10] and phase space weights [23]. This has been implemented and tested in various event generators previously [42, 40]. In the context of this workshop we found that parallelization using MPI can be more efficient than multi-threading, as it scales with the number of nodes, while multi-threaded applications are limited by the number of compute cores on a single machine. Hybrid approaches are very promising.

45.3.3 Parallel computing using accelerators

Several algorithms have been proposed, which allow to accelerate the calculation even further using GPUs [45, 40, 44]. Proof-of-concept implementations of these methods have shown great potential and may be used on a larger scale in the near future.

n-gluons	integration (BASES)	generation (SPRING)
0	95	24
1	84	44
2	67	70
3	39	>1000
4	18	n.a.

Table 45-2. *Ratios of total execution times of CPU and GPU programs for MC integration and parton-level event generation in $u\bar{d} \to W^+ + n$gluons with BASES/SPRING [53]. We used an NVidia Tesla C2075 GPU with CUDA™ 4.2 and an Intel® Core i7 2.67 GHz.*

For the purpose of this study we use a GPU-implementation of a $u\bar{d} \to W^+ + n$ gluons tree-level calculation [44] as the benchmark process. Table 45-2 shows ratios of total execution times between CPU and GPU programs for MC integration and parton-level event generation with BASES/SPRING [53]. We parallelized the event generation in the sense that multiple phase-space points are produced at the same time, in a load balanced approach between the different CUDA kernels. The CPU program was not parallelized, hence the numbers shown in Tab. 45-2 are indicative only of the gain when using GPUs alone. Newer GPU architectures should further enhance performance.

We have also tested MPI communication on an Intel®Xeon Phi™coprocessor, which was made available to us by CERN OpenLab [28]. The clock frequency of its compute cores was 1.238 GHz, and the total memory of the system was 16 GB. The total number of compute cores was 244 (61×4). We find that the execution speed of the executable can be reduced, depending on the complexity of the problem. In simple MPI mode we observe strong scaling up to about 32 coprocessor cores. The ratio of computation time when run on a single coprocessor core, compared to a single CPU core ranges from 4.67 for $e^+e^- \to$ 2jets over 7.52 for $e^+e^- \to$ 4jets to 19.2 for $pp \to W^\pm +$ 5jets. It therefore seems more promising to use the coprocessor in offload and multi-threaded mode, rather than as an additional many-core processor. Current limitations in this context are entirely due to the structure of our executable, which has not yet been optimized for coprocessors. Substantially higher gain may therefore be expected in the future.

Developments for the application of GPUs and coprocessors to more complex and time consuming problems are ongoing. Massively parallel computations using accelerators will likely become an important technique for perturbative QCD calculations and should be combined with other HPC methods.

45.3.4 Suitability of existing HPC resources

During this workshop, several HPC resources under supervision of the U.S. Department of Energy (DOE) were made available for tests of automated NLO calculations and Monte-Carlo event generators. We have benchmarked code performance at the following three facilities:

- Cray XE6™ "Hopper" at NERSC [62]
  24 AMD Opteron™ 2.1 GHz cores per node (153,216 total cores)
  32/64 GB RAM per node (6,000/384 nodes)
  Cray Gemini 3D Torus Network

- Cray XK7™ "Titan" at OLCF [63]
  16 AMD Opteron™ 2.2 GHz cores per node (299,008 total cores)
  32 GB RAM per node (all nodes)
  NVidia®K20 GPU accelerators (18,688 total GPUs)
  Cray Gemini 3D Torus Network

- IBM®BlueGene®/Q test system "Vesta" at ALCF [7]
  16 1.6 GHz PowerPC®A2 cores per node (32,768 total cores)
  16 GB RAM per node (all nodes)
  IBM 5D Torus Network

The two Cray systems resemble standard Linux environments, which makes porting of existing codes convenient. Standard software like the GNU compiler collection is available on all three systems. It was used for compiling our benchmark applications. It is to some extent simpler to test and debug code on the Cray architecture, as the environment includes interactive nodes which run a full fledged Linux kernel.

We found that runtimes for the BlackHat+Sherpa event generator are identical to within 5% on a Hopper node and on a Titan node. On these systems we have tested weak scaling up to 8,192 nodes and shown strong scaling up to 1,024 nodes. These numbers are not necessarily representative for other applications, and they strongly depend on the process under consideration. It is likely that they will improve substantially over the next few years.

During our tests on Vesta we encountered a larger I/O latency, which could be fatal for applications designed to perform I/O operations on a per-node basis. The lower clock frequency of the IBM BlueGene®system does not allow direct comparisons between single compute cores on Vesta and the two Cray machines. In order to reduce turnover time, parallel codes would be favored strongly. To achieve similar performance in our benchmark, the number of compute cores had to be increased by a factor 2.2 compared to Hopper and Titan. However, the large number of cores on the test system Vesta and the corresponding production system Mira (786,432 total cores) may compensate this. We have tested weak scaling up to 16,000 cores on Vesta.

45.3.5 Suitability of the Open Science Grid

We have tested the performance of Monte-Carlo event generation frameworks (both parton-level and particle level) on the Open Science Grid [5, 64]. We found excellent usability, combined with very strong user support. The Condor-based glidein Workflow Management System, used by OSG, proved to be a very convenient tool for Monte Carlo event production as larger jobs can easily be split into multiple subjobs.

In this manner we have carried out a Standard Model background study for Supersymmetry searches in the phenomenological MSSM within several days. The project required the use of 150,000 CPU hours and produced ∼1 TB of data. We used custom scripts to transfer the data from each worker node to the Storage Element (SE) at the site where the job ran. Later we transferred the data from the SEs to SLAC. We have started looking into using the OSG public storage service to automate this data handling.

We have explored MPI parallelization on the OSG by running small-scale HPC jobs. The number of cores accessible through the system is currently limited to the maximum number of cores per node (between 4 and 64). It will be highly beneficial to break this limitation in the future and provide access to larger HPC facilities through the OSG as a common interface. This will reduce the training requirements for new

Process	Ref.	Requirements	
		CPU [core h]	Storage [GB]
$pp \to W^{\pm} + 5jets$	[15]	600,000	1,500
$pp \to W^{\pm} + 4jets$	[12]	100,000	200
$pp \to Z + 4jets$	[52]	200,000	200
$pp \to Z + 3jets$	[11]	50,000	100
$pp \to 4jets$	[14]	200,000	150

Table 45-3. *CPU and Storage requirements for calculations on the list of important processes identified during the LesHouches series of workshops [2]. Numbers assume cross-checks using at least two independent runs to guarantee the correctness of results and are reported for AMD Opteron™ processors running at 2.1 GHz. Storage requirements are reported for Root NTuple files which can be used to replicate the entire event analysis.*

researchers entering the field, who want to make use of both, large-scale distributed computing as well as HPC.

45.4 Resource requirements

One of the aims of this workshop has been to provide a quantitative estimate of the computational resources needed to continue and expand the scientific program of the theory community working at providing phenomenological predictions for the LHC experiments. In this section we summarize the results obtained in this context. The discussion focuses on three main building blocks: higher order QCD and EW corrections to the parton level cross sections, parton distribution functions, and event generators.

45.4.1 Higher-order perturbative calculations

Collider events with large final-state particle multiplicity, and in particular with many jets, constitute the main backgrounds to many new physics searches. The need to describe such events with the best possible theoretical precision has triggered substantial improvements in NLO calculations. On the one hand, Monte-Carlo programs have become available, which can compute all but the virtual corrections at NLO automatically and generate parton-level events. On the other hand, "One-Loop Providers" (OLP) have emerged, programs that efficiently compute the one-loop amplitudes entering the virtual corrections [13, 46, 30, 48, 27, 65]. Combination of the two developments yields powerful and fully automated tools for LHC phenomenology. Hence, a standard was proposed in 2009 to combine MC and OLP in a unified manner [16], which is used by all major projects.

Table 45-3 shows CPU and storage requirements for typical NLO calculations required for LHC phenomenology. The numbers assume at least one cross-check of the result to be performed with an entirely independent run, and they include all parts of the NLO calculation. The calculation is required to reach a Monte-Carlo uncertainty which makes the comparison with experimental measurements meaningful, see the references for details. All numbers provided include cross section calculations and the production of event samples, which are stored in Root [19] NTuple format and analyzed to produce the histograms shown in the respective publications.

Process	Ref.	Requirements CPU [core h]	CPU clock [GHz]
$pp \to W/Z$	[60, 57]	50,000	2.67
$pp \to H$	[6]	50,000	2.67
$pp \to t\bar{t}$	[8, 32]	1,000,000	2.27
$pp \to$ jets (g only)	[66]	85,000	2.20
$pp \to H$+jet (g only)	[17]	500,000	2.67

Table 45-4. *Summary of computing requirements for NNLO calculations. Numbers were obtained on Intel® Xeon® CPU's with varying clock frequency and are therefore not directly comparable.*

NLO accuracy has been recently reached also in techniques like the matrix-element method, which efficiently separates signal and background events in searches at the LHC [24]. The increased precision will greatly enhance the potential of analyses focused on, for example, Higgs boson decays into WW pairs, but it requires substantially larger computing power compared to analogous LO techniques. Indeed, in addition to unobserved kinematic variables in the event, at NLO one must integrate over the extra degrees of freedom corresponding to emission of real radiation. Including integration over detector response functions and experimentally unobserved variables, this is equivalent to performing an entire cross section calculation for each event included in the analysis. Studying the existing $pp \to H \to WW$ candidate events in the 20 fb^{-1} LHC data set requires 200,000 CPU hours in this approach.

The precision being achieved in numerous benchmark measurements at the LHC is imposing ever-increasing demands upon theoretical predictions. NLO QCD calculations are no longer sufficient to match the level of accuracy, for example, in measurements of W- and Z-boson properties. Perturbative QCD calculations at the NNLO, sometimes combined with NLO electroweak corrections, have become available in numerical programs like FEWZ [59, 39, 57] and FEHiP [6], and will become available for several other benchmark processes, including top-quark pair production [8, 32] and Higgs plus jet production [17] in the near future. A summary of current requirements is given in Tab. 45-4.

Some programs like FEWZ are specifically designed to run in high throughput mode on parallel computing systems. The NNLO QCD corrections are thereby split into independent regions according to their underlying singularity structure, and are integrated independently on separate grids. In this manner, a full comparison with measurements of the $d^2\sigma/dM/dY$ distribution in lepton-pair production divided into 150 bins requires approximately 50,000 CPU hours on Intel®Xeon®2.67 GHz CPUs. New subtraction schemes that provide a framework for NNLO calculations to arbitrarily-complicated processes also rely upon a splitting of the final-state phase space [31, 18].

Cutting-edge NNLO calculations that have recently become available, like top-quark pair production [8, 32], jet production [66] and Higgs plus jet production [17], are much more demanding in terms of computational resources, as can be seen in Table 45-4. Due to the increased demand for NNLO predictions, we can expect substantially larger computing resources to be needed in the near future. At the same time, the development and adoption of standard techniques that are optimized to provide public NNLO codes, is still very preliminary. As a result, the numbers in Table 45-4 have been obtained using different methods and with the goal to reach a theoretical accuracy that could efficiently compare with experiments. This makes the estimate of needed computational resources, for the time being, very process/calculation dependent. The access to HPC facilities will shorten this exploratory phase and will allow a more rapid convergence towards the selection of efficient techniques to be implemented in a broad range of future NNLO calculations.

While fixed-order perturbative QCD calculations can predict inclusive jet observables very well, they cannot be applied to the analysis of exclusive jet bins due to the reduced accuracy associated with a veto on the transverse momentum of real emissions. Resummed calculations are required in this context, which have been in the focus of interest recently due to their importance in the W^+W^- decay channel of the Higgs boson. They require computational resources to run publicly available software for NLO and NNLO cross section calculations in order to numerically extract coefficients needed for the resummation, and also in order to match to the fixed-order result. Such studies have been performed for Higgs plus jet cross sections on the Carver cluster at NERSC (Intel®Xeon®CPU at 2.67 GHz) and required approximately 300,000 CPU hours [68]. The demand for similar predictions is rising.

45.4.2 Parton Distribution Functions

Uncertainty in parton distribution functions (PDFs) constitutes the leading theoretical uncertainty in many collider processes. It needs to be reduced to realize the potential of multi-loop calculations for QCD hard cross sections.

Computing needs of the future PDF analysis will be determined by a number of trends. First, implementation of fast NLO and NNLO computations using the methods of ApplGrid [25] and FastNLO [55] will speed up PDF fits. These methods replace point-by-point K-factor lookup tables that have been used in PDF fits to rapidly estimate higher-order radiative contributions with some loss in the accuracy. Without fast computations or K factor tables, the CPU time needed for fitting the PDFs increases by 1-2 orders of magnitude.

Second, many numerical approximations in the computation of QCD cross sections that speed up the current PDF analyses will need to be eliminated in the future to match the accuracy of the NNLO and N^{3}LO hard cross sections.

The global PDF analysis involves repetitive minimization, integration, and interpolation in the space of many parameters. The CPU time expenses for every such step will need to be raised by up to an order of magnitude, leading to nonlinear growth in the CPU time spent on the whole fit.

For example, in the current CTEQ and nCTEQ fits, scattering cross sections are computed with numerical accuracy below a fraction of percent. A typical recent study, such as production of CT10 eigenvector sets, nCTEQ nuclear PDF sets, or investigation of charm mass dependence in the PDF analysis [38, 37, 67], required 3,000-10,000 CPU hours for 20-30 PDF parameters and using the K factor tables or fast (N)NLO cross sections. This time is spent on finding the parametrizations of PDFs describing the data, exploring the PDF parameter space in order to determine uncertainties in the PDF parameters, and validating the results. Higher accuracy and more free parameters and fitted processes can quickly increase the CPU time demands, possibly by a factor 5-10 in the foreseeable future.

45.4.3 Particle level event generators

Fully exclusive event generators [21] are a crucial tool to compare theoretical calculations to experimental observables including realistic experimental cuts and detector resolution effects. There has been a large effort worldwide to increase the precision of the theoretical calculations these generators are based on, typically by including information from fixed order perturbation theory at NLO. The two matching methods MC@NLO [36] and POWHEG [61], which allow to combine NLO calculations with parton showers, provide

Process	N_{jet}		Ref.	CPU [core h]
	NLO	LO		
$pp \to W^{\pm} + jets$	$\leq$2	$\leq$4	[49]	100,000
$pp \to h + jets$	$\leq$2	$\leq$3	[50]	150,000
$pp \to t\bar{t} + jets$	$\leq$1	$\leq$2	[51]	250,000
$pp \to l\bar{\nu}\bar{l}'\nu'$	$\leq$1	$\leq$2	[26]	50,000

Table 45-5. *Computing requirements for NLO-merged predictions in various benchmark processes, using the Sherpa event generator. Numbers assume cross-checks using at least two independent runs to guarantee the correctness of results.*

the theoretical basis. In addition, three different methods have been introduced recently to combine multiple NLO-matched calculations for varying jet multiplicity with each other. They produce inclusive event samples, which can be reduced to NLO-accurate predictions at arbitrary jet multiplicity [49, 58, 35]. Such calculations are very demanding, since they rely on NLO calculations as an input to the simulation. The challenge of performing these NLO calculations at high multiplicity can be appreciated by inspecting Tab. 45-3. Table 45-5 lists some exemplary computing requirements for the most challenging calculations performed with MC event generators.

The Monte-Carlo framework Geneva [4] aims to go further and improve the formal accuracy of the resummation of large logarithms, which are typically only resummed at leading or next-to-leading logarithmic order by the parton shower. Geneva also allows the combination of multiple NLO calculations with each other. The validation of the first results in $e^+e^- \to$hadrons [4] required of the order of 200,000 CPU hours. CPU time at the same order of magnitude is being used for the validation of results in hadronic collisions [3].

An important aspect in the construction of event generators is the validation for new types of processes [20], the preparation of public releases and their tuning [22]. All three aspects require substantial computing resources. Typically, of the order of 50,000 CPU hours are spent for a full set of tests of the generator. The tuning process involves substantially more resources, typically 150,000 - 300,000 CPU hours.

45.5 Conclusions

Groups of particle theorists at both US universities and DOE laboratories have been playing a leading role in each of the research areas outlined in this report. To keep their impact and momentum at a time when the LHC is putting both precision SM studies and broad studies of physics beyond the SM at high demand, these groups will need to have access to extensive computing resources that could, to some extent, efficiently be provided within the framework of national supercomputing facilities. At the same time, local resources must remain available for prototyping and testing of new applications.

We see two major benefits arising from the use of HPC facilities. First, a variety of existing calculations/software can be made public within a common well-tested framework, and in this way can be used for experimental studies involving extensive multiple runs without depending on the computer power and manpower available to their authors. At the same time, new sophisticated calculations can fully exploit the technological advantage of the facility to provide cutting-edge results that would not otherwise be within reach. Specific examples of both uses have been given in this report.

Local computing resources provide a steady basis for small-scale testing as well as prototyping and development of new calculations. Efficient code development can only be guaranteed in an environment which

allows unrestricted access to computing time, which is often possible only with a resource partially or entirely under the management of the researcher and its home institution. However, temporarily idle resources could efficiently be harvested by the Open Science Grid and thus made available to other researchers nationwide.

45.6 Acknowledgments

We are indebted to the Department of Energy, Office of Science, for initiating this study. We are especially grateful to Lali Chatterjee and Larry Price for stimulating discussions and for providing the required resources at NERSC. We thank Richard Gerber, Tom LeCompte and Richard Mount for their support in testing different LCFs. We are indebted to Salman Habib for many stimulating and fruitful discussions on code performance and optimization. We thank Gabriele Garzoglio, Tanya Levshina and Marko Slyz for their help in using the OSG. We are grateful to Andrea Dotti and Paweł Szostek for help in testing the Intel Phi.

R. Boughezal is supported by the U.S. Department of Energy under Contract No. DE-AC02-06CH11357. J. Campbell and C. Williams are supported by the U.S. Department of Energy under Contract No. DE-AC02-06CH11359. L. Dixon and S. Höche are supported by the U.S. Department of Energy under Contract No. DE-AC02-76SF00515. A. Mitov is supported by ERC grant 291377 "LHCtheory: Theoretical predictions and analyses of LHC physics: advancing the precision frontier". P. Nadolsky and F. Olness are supported by the U.S. Department of Energy under grant DE-FG02-04ER41299 and Early Career Research Award DE-SC0003870, and by the Lightner Sams Foundation. F. Petriello is supported by the U.S. Department of Energy under Grant No. DE-SC0010143. L. Reina is supported by the U.S. Department of Energy under grant DE-FG02-13ER41942.

References

[1] Georges Aad et al. Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC. *Phys.Lett.*, B716:1–29, 2012.

[2] J. Alcaraz Maestre et al. The SM and NLO Multileg and SM MC Working Groups: Summary Report. 2012.

[3] Simone Alioli, Christian W. Bauer, Calvin Berggren, Andrew Hornig, Frank J. Tackmann, et al. Combining Higher-Order Resummation with Multiple NLO Calculations and Parton Showers in the GENEVA Monte Carlo Framework. 2013.

[4] Simone Alioli, Christian W. Bauer, Calvin J. Berggren, Andrew Hornig, Frank J. Tackmann, et al. Combining Higher-Order Resummation with Multiple NLO Calculations and Parton Showers in GENEVA. 2012.

[5] Mine Altunay et al. A Science driven production cyberinfrastructure: the Open Science Grid. *J.Grid Comput.*, 9:201–218, 2011.

[6] Charalampos Anastasiou, Kirill Melnikov, and Frank Petriello. Fully differential Higgs boson production and the di-photon signal through next-to-next-to-leading order. *Nucl. Phys.*, B724:197–246, 2005.

[7] Argonne Leadership Computing Facility. http://www.alcf.anl.gov.

[8] Peter Bärnreuther, Michal Czakon, and Alexander Mitov. Percent level precision physics at the Tevatron: first genuine NNLO QCD corrections to $q\bar{q} \to t\bar{t} + X$. *Phys.Rev.Lett.*, 109:132001, 2012.

[9] C. Bauer et al. The computing needs of theoretical high energy physics at the Energy Frontier. http://snowmass2013.org/tiki-index.php?page=HEP+Theory+and+High-Performance-Computing.

[10] Frits A. Berends and W. T. Giele. Recursive calculations for processes with n gluons. *Nucl. Phys.*, B306:759, 1988.

[11] C. F. Berger, Z. Bern, L. J. Dixon, F. Febres-Cordero, D. Forde, T. Gleisberg, H. Ita, D. A. Kosower, and D. Maître. Next-to-leading order QCD predictions for Z, γ^*+3-Jet distributions at the Tevatron. *Phys. Rev.*, D82:074002, 2010.

[12] C. F. Berger, Z. Bern, L. J. Dixon, F. Febres-Cordero, D. Forde, T. Gleisberg, H. Ita, D. A. Kosower, and D. Maître. Precise Predictions for W + 4-Jet Production at the Large Hadron Collider. *Phys. Rev. Lett.*, 106:092001, 2011.

[13] C. F. Berger, Z. Bern, L. J. Dixon, F. Febres-Cordero, D. Forde, H. Ita, D. A. Kosower, and D. Maître. Automated implementation of on-shell methods for one-loop amplitudes. *Phys. Rev.*, D78:036003, 2008.

[14] Z. Bern, G. Diana, L.J. Dixon, F. Febres Cordero, S. Höche, et al. Four-Jet Production at the Large Hadron Collider at Next-to-Leading Order in QCD. *Phys.Rev.Lett.*, 109:042001, 2012.

[15] Z. Bern, L.J. Dixon, F. Febres Cordero, S. Hoeche, H. Ita, et al. Next-to-Leading Order W + 5-Jet Production at the LHC. 2013.

[16] T. Binoth et al. A proposal for a standard interface between Monte Carlo tools and one-loop programs. *Comput. Phys. Commun.*, 181:1612–1622, 2010.

[17] Radja Boughezal, Fabrizio Caola, Kirill Melnikov, Frank Petriello, and Markus Schulze. Higgs boson production in association with a jet at next-to-next-to-leading order in perturbative QCD. *JHEP*, 1306:072, 2013.

[18] Radja Boughezal, Kirill Melnikov, and Frank Petriello. A subtraction scheme for NNLO computations. *Phys.Rev.*, D85:034025, 2012.

[19] R. Brun and F. Rademakers. ROOT: An object oriented data analysis framework. *Nucl.Instrum.Meth.*, A389:81–86, 1997.

[20] Andy Buckley et al. Rivet user manual. 2010.

[21] Andy Buckley et al. General-purpose event generators for LHC physics. *Phys. Rept.*, 504:145–233, 2011.

[22] Andy Buckley, Hendrik Hoeth, Heiko Lacker, Holger Schulz, and Jan Eike von Seggern. Systematic event generator tuning for the LHC. *Eur. Phys. J.*, C65:331–357, 2010.

[23] E. Byckling and K. Kajantie. N-particle phase space in terms of invariant momentum transfers. *Nucl. Phys.*, B9:568–576, 1969.

[24] John M. Campbell, Walter T. Giele, and Ciaran Williams. The Matrix Element Method at Next-to-Leading Order. *JHEP*, 1211:043, 2012.

[25] Tancredi Carli et al. A posteriori inclusion of parton density functions in NLO QCD final-state calculations at hadron colliders: The APPLGRID Project. *Eur. Phys. J.*, C:503–524, 2010.

[26] Fabio Cascioli, Stefan Hoeche, Frank Krauss, Philipp Maierhofer, Stefano Pozzorini, and Frank Siegert. Precise Higgs-background predictions: merging NLO QCD and squared quark-loop corrections to four-lepton + 0,1 jet production. 2013.

[27] Fabio Cascioli, Philipp Maierhofer, and Stefano Pozzorini. Scattering Amplitudes with Open Loops. *Eur.Phys.J.*, C72:1889, 2012.

[28] CERN openlab. http://openlab.web.cern.ch.

[29] Serguei Chatrchyan et al. Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC. *Phys.Lett.*, B716:30–61, 2012.

[30] Gavin Cullen, Nicolas Greiner, Gudrun Heinrich, Gionata Luisoni, Pierpaolo Mastrolia, Giovanni Ossola, Thomas Reiter, and Francesco Tramontano. Automated One-Loop Calculations with GoSam. *Eur.Phys.J.*, C72:1889, 2012.

[31] M. Czakon. A novel subtraction scheme for double-real radiation at NNLO. *Phys.Lett.*, B693:259–268, 2010.

[32] Michal Czakon, Paul Fiedler, and Alexander Mitov. The total top quark pair production cross-section at hadron colliders through $O(\alpha_S^4)$. 2013.

[33] S. Dittmaier et al. Handbook of LHC Higgs Cross Sections: 1. Inclusive Observables. 2011.

[34] S. Dittmaier et al. Handbook of LHC Higgs Cross Sections: 2. Differential Distributions. 2012.

[35] Rikkert Frederix and Stefano Frixione. Merging meets matching in MC@NLO. *JHEP*, 1212:061, 2012.

[36] Stefano Frixione and Bryan R. Webber. Matching NLO QCD computations and parton shower simulations. *JHEP*, 06:029, 2002.

[37] Jun Gao, Marco Guzzi, Joey Huston, Hung-Liang Lai, Zhao Li, et al. The CT10 NNLO Global Analysis of QCD. 2013.

[38] Jun Gao, Marco Guzzi, and Pavel M. Nadolsky. Charm quark mass dependence in a global QCD analysis. 2013.

[39] Ryan Gavin, Ye Li, Frank Petriello, and Seth Quackenbush. FEWZ 2.0: A code for hadronic Z production at next-to-next-to-leading order. *Comput.Phys.Commun.*, 182:2388–2403, 2011.

[40] Walter Giele, Gerben Stavenga, and Jan-Christopher Winter. Thread-Scalable Evaluation of Multi-Jet Observables. *Eur.Phys.J.*, C71:1703, 2011.

[41] T. Gleisberg, S. Höche, F. Krauss, M. Schönherr, S. Schumann, F Siegert, and J. Winter. Event generation with Sherpa 1.1. *JHEP*, 02:007, 2009.

[42] Tanju Gleisberg and Stefan Höche. Comix, a new matrix element generator. *JHEP*, 12:039, 2008.

[43] Tanju Gleisberg, Stefan Höche, Frank Krauss, Andreas Schälicke, Steffen Schumann, and Jan Winter. Sherpa 1.α, a proof-of-concept version. *JHEP*, 02:056, 2004.

[44] K. Hagiwara, J. Kanzaki, Q. Li, N. Okamura, and T. Stelzer. Fast computation of MadGraph amplitudes on graphics processing unit (GPU). 2013.

[45] K. Hagiwara, J. Kanzaki, N. Okamura, D. Rainwater, and T. Stelzer. Fast calculation of HELAS amplitudes using graphics processing unit (GPU). *Eur.Phys.J.*, C66:477–492, 2010.

[46] Thomas Hahn. Automatic loop calculations with FeynArts, FormCalc, and LoopTools. *Nucl.Phys.Proc.Suppl.*, 89:231–236, 2000.

[47] S. Heinemeyer et al. Handbook of LHC Higgs Cross Sections: 3. Higgs Properties. 2013.

[48] Valentin Hirschi, Rikkert Frederix, Stefano Frixione, Maria Vittoria Garzelli, Fabio Maltoni, and Roberto Pittau. Automation of one-loop QCD corrections. *JHEP*, 1105:044, 2011.

[49] Stefan Höche, Frank Krauss, Marek Schönherr, and Frank Siegert. QCD matrix elements + parton showers: The NLO case. 2012.

[50] Stefan Höche, Frank Krauss, Marek Schönherr, and Frank Siegert. to appear. 2013.

[51] Stefan Hoeche, Junwu Huang, Gionata Luisoni, Marek Schoenherr, and Jan Winter. Zero and one jet combined NLO analysis of the top quark forward-backward asymmetry. 2013.

[52] Harald Ita, Zvi Bern, Lance J. Dixon, Fernando Febres-Cordero, David A. Kosower, and Daniel Maître. Precise Predictions for Z + 4 Jets at Hadron Colliders. *Phys.Rev.*, D85:031501, 2012.

[53] Setsuya Kawabata. A New version of the multidimensional integration and event generation package BASES/SPRING. *Comput.Phys.Commun.*, 88:309–326, 1995.

[54] Ronald Kleiss and Roberto Pittau. Weight optimization in multichannel Monte Carlo. *Comput. Phys. Commun.*, 83:141–146, 1994.

[55] T. Kluge, K. Rabbertz, and M. Wobisch. FastNLO: Fast pQCD calculations for PDF fits. pages 483–486, 2006.

[56] G. Peter Lepage. A New Algorithm for Adaptive Multidimensional Integration. *J. Comput. Phys.*, 27:192, 1978.

[57] Ye Li and Frank Petriello. Combining QCD and electroweak corrections to dilepton production in FEWZ. *Phys.Rev.*, D86:094034, 2012.

[58] Leif Lönnblad and Stefan Prestel. Merging Multi-leg NLO Matrix Elements with Parton Showers. *JHEP*, 1303:166, 2013.

[59] Kirill Melnikov and Frank Petriello. Electroweak gauge boson production at hadron colliders through O(alpha(s)**2). *Phys.Rev.*, D74:114017, 2006.

[60] Kirill Melnikov and Frank Petriello. The W boson production cross section at the LHC through $O(\alpha_s^2)$. *Phys.Rev.Lett.*, 96:231803, 2006.

[61] Paolo Nason. A new method for combining NLO QCD with shower Monte Carlo algorithms. *JHEP*, 11:040, 2004.

[62] National Energy Research Scientific Computing Center. http://www.nersc.gov.

[63] Oak Ridge Leadership Computing Facility. http://www.nccs.gov.

[64] Open Science Grid. http://www.opensciencegrid.org.

[65] Laura Reina and Thomas Schutzmeier. Towards W b bbar + j at NLO with an Automatized Approach to One-Loop Computations. *JHEP*, 1209:119, 2012.

[66] Aude Gehrmann-De Ridder, Thomas Gehrmann, E.W.N. Glover, and Joao Pires. Second order QCD corrections to jet production at hadron colliders: the all-gluon contribution. *Phys.Rev.Lett.*, 110:162003, 2013.

[67] I. Schienbein, J.Y. Yu, K. Kovarik, C. Keppel, J.G. Morfin, et al. PDF Nuclear Corrections for Charged and Neutral Current Processes. *Phys.Rev.*, D80:094004, 2009.

[68] Iain W. Stewart, Frank J. Tackmann, Jonathan R. Walsh, and Saba Zuberi. Jet p_T Resummation in Higgs Production at NNLL'+NNLO. 2013.

[69] Worldwide LHC Computing Grid. http://wlcg.web.cern.ch.