

Performance of the bwHPC cluster in the production of $\mu \rightarrow \tau$ embedded events used for the prediction of background for $H \rightarrow \tau\tau$ analyses

Janek Bechtel Sebastian Brommer Artur Gottmann 

Günter Quast

Roger Wolf 

Institut für Experimentelle Teilchenphysik, Karlsruher Institut für Technologie, Karlsruhe, Germany

In high energy physics, a main challenge is the accurate prediction of background events at a particle detector. These events are usually estimated by simulation. As an alternative, data-driven methods use observed events to derive a background prediction and are often less computationally expensive than simulation. The τ lepton embedding method presents a data-driven method to estimate the background from $Z \rightarrow \tau\tau$ events for Higgs boson analyses in the same final state. $Z \rightarrow \mu\mu$ events recorded by the CMS experiment are selected, the muons are removed from the event and replaced with simulated τ leptons with the same kinematic properties as the removed muons. The resulting hybrid event provides an improved description of pile-up and the underlying event compared to the simulation of the full proton-proton collision. In this paper the production of these hybrid events used by the CMS collaboration is described. The production relies on the resources made available by the bwHPC project. The data used for this purpose correspond to 65 million di-muon events collected in 2017 by CMS.

1 The Higgs boson in the Standard Model

In the Standard Model of particle physics, the requirement of a mechanism to achieve electroweak symmetry breaking leads to the prediction of the existence of at least one neutral scalar particle, known as the Higgs boson (Guralnik et al., 1964; Higgs, 1964; Englert et al., 1964). A particle compatible with the required properties of this boson was observed in 2012 by the ATLAS and the CMS experiment at the Large Hadron Collider (LHC) at CERN via its decay into photons, Z bosons and W bosons. (ATLAS collaboration, 2012; CMS collaboration, 2012a; CMS collaboration, 2013). Additionally, a Standard Model Higgs boson is expected to generate mass for particles of half-integer spin, known as fermions. To establish this mechanism known as Yukawa coupling, a measurement of the direct Higgs boson coupling to fermions is necessary. As the coupling of the Higgs bosons to fermions is proportional to its mass, the most promising decay channel is the decay into two oppositely charged τ leptons ($H \rightarrow \tau\tau$) due to the large event rate compared to lighter fermions, and smaller contribution from background events compared to the decay of the Higgs boson into heavy quarks. After being observed already in the combined measurements of the ATLAS and CMS experiments during Run-1, the first observations of this decay into τ leptons by single experiments were presented in 2018 by the CMS and ATLAS collaborations (CMS collaboration, 2018a; ATLAS collaboration, 2018), paving the way for precision measurements of the couplings of the Higgs boson to fermions on the growing dataset collected by the CMS experiment. In addition to these precision measurements, many promising supersymmetric extensions to the Standard Model predict additional heavy neutral Higgs bosons with possibly enhanced couplings to down-type fermions such as the τ lepton.

The challenges of these analyses include an accurate description of background events with the same event signature as decays of the Higgs boson, the most prominent one being the decay of the neutral force carrier of the weak force, the Z boson, into two τ leptons ($Z \rightarrow \tau\tau$). A common way to estimate this background is to simulate proton-proton collisions to obtain samples of simulated $Z \rightarrow \tau\tau$ events. This from-scratch simulation of particle collisions is computationally expensive, and has difficulties in describing many complicated processes at the LHC such as additional jets due to multiple proton-proton collisions or the underlying event of the hard collision. In this paper a data-driven method which provides a computing-efficient way to solve these difficulties is presented.

2 The τ lepton embedding method

The decay of Z bosons into τ leptons ($Z \rightarrow \tau\tau$) constitutes one of the most prominent irreducible backgrounds for the search and analysis of Higgs boson events in the di- τ lepton final state. Instead of purely relying on simulation, this background can mostly be estimated from data. This estimation is done by making use of lepton universality in the Standard Model of particle physics, which in this specific case refers to the equal coupling strength of the Z boson to all leptons such as muons or τ leptons. Z boson events with two muons ($Z \rightarrow \mu\mu$) are selected. The tracker hits and all energy deposits of the initially reconstructed muons are removed from the event and replaced by simulated τ lepton decays with the same kinematic properties as for the removed muons, creating a partially simulated hybrid event. In such hybrid events only the well understood τ lepton decay relies on simulation and parts of the event that are difficult to describe, like the underlying event or the production of additional jets, are estimated from data. Since the simulated τ lepton decays are embedded into the remaining environment of a $Z \rightarrow \mu\mu$ event after removal of the muons, this approach is referred to as τ lepton embedding method. A visualization of this method is given in Figure 1.

The technique has been successfully used in the past by the ATLAS and CMS Collaboration for the search and analysis of Higgs boson events in the context of the Standard Model of particle physics and its minimal supersymmetric extension on the LHC Run-1 dataset (ATLAS collaboration, 2015; CMS collaboration, 2011; CMS collaboration, 2012b; CMS collaboration, 2014a; CMS collaboration, 2014b). Embedded events produced on bwHPC resources using the Run-2 dataset have served as a cross-check of the estimation of the background from $Z \rightarrow \tau\tau$ events from simulation, for the first CMS search for additional heavy Higgs bosons in the $\tau\tau$ final state at 13 TeV in the context of the MSSM (CMS collaboration, 2018b).

With a runtime of $\mathcal{O}(10\text{s})$ per event, the τ lepton embedding method requires considerably less computational effort compared to a from-scratch simulation of proton-proton collisions at the CMS detector, which takes $\mathcal{O}(\text{minutes})$ per event. Still, computational challenges arise due to the large number of events that are needed. The availability of a large amount of computing resources is necessary for coping with these challenges. During the development and production of the embedded events in both 2017 and 2018, the vast majority of events was processed

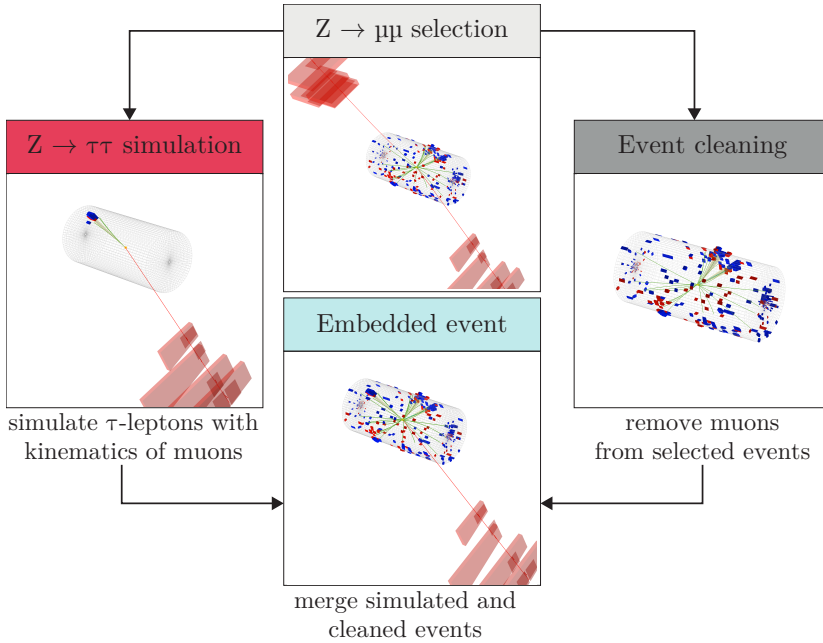


Figure 1: Visualization of the τ lepton embedding method. $Z \rightarrow \mu\mu$ events are selected from data recorded by the CMS experiment. The muon tracks as well as their deposits in the calorimeters are removed from the recorded event. The decay of two τ leptons is now simulated in an empty detector environment and merged into the cleaned event. The resulting hybrid event is used to describe decays of the Z boson into τ leptons. From Ref. (Bechtel, 2017).

on the Cluster NEMO¹ located in Freiburg. In the following, the production of these embedded events is described. In Section 3, an overview of the computational setup used for the production of the embedded events is given. Section 3.1 provides a statistical insight into the performance of the computing resources.

¹<https://www.hpc.uni-freiburg.de/nemo>

3 Production of embedded events at the bwHPC

The production of embedded events starts with the full dataset collected by the CMS experiment in 2017 tagged to contain two muon candidates. This dataset amounts to 219 million events. A full and technical description of the CMS detector can be found in Ref. (CMS collaboration, 2008). For the purpose of the τ lepton embedding method, the uncompressed detector information is used. As this dataset contains all events in which two muon candidates could be identified, a first **pre-selection** step reduces the sample to all events where a Z boson candidate with sufficient invariant mass can be constructed. 3 out of 10 events survive this selection, ultimately stored on disk to be used for the subsequent steps. The storage sites used for this work-flow are both the dedicated high energy physics (HEP) storage at GridKa, located at KIT, Karlsruhe, and the HEP storage at DESY, Hamburg.

On average, the required disk space per uncompressed event is 3 MB, resulting in a total disk usage of 206 TB for 65 million events. These events are then transferred to the computing center in batches of 1000 events for subsequent hybrid event production. The production is split up into four steps executed in sequence, each producing an input file for the following step.

1. **Selection of the di-muon events.** The selection step is repeated to ensure compatibility with changes in the reconstruction software and to adapt the selection criteria. Here, requirements on the energy of the two muon candidates and a requirement of at least one Z boson candidate in the event are set.

Average runtime per event: 5.0 s

2. **Cleaning of the two muons from the event.** All hits in the silicon tracker and muon detection systems of the CMS detector that are used to reconstruct the tracks of the two muons, and all entries in the calorimetry that are related to the muons are deleted from the event.

Average runtime per event: 4.4 s

3. **Simulation step.** The reconstructed kinematics of the two muons are used to create two τ leptons with the same kinematics. The decay of these τ leptons is then simulated in an empty detector environment. In this step, a filter is applied which only selects events in which the τ leptons decay into a predefined final state, the latter being defined by containing an electron, a muon or a harmonically decaying τ lepton. This filter allows to use the selected events

multiple times for each di- τ lepton final state.

Average runtime per event: 4.9 s

4. **Merging of cleaned and simulated event.** The simulated decay of the τ leptons and the underlying event from which the two muon candidates have been selected are merged to form a hybrid *embedded* event. The merging is done on the level of reconstructed objects at the earliest possible stage, which is at the level of reconstructed tracks in the tracking system and hits in the calorimeters, and before the reconstruction of particle candidates by the particle-flow algorithm (CMS collaboration, 2017). The hybrid event now provides a data-driven estimate of $Z \rightarrow \tau\tau$ events in the CMS detector and can be used as background prediction for Higgs analyses.

Average runtime per event: 1.1 s

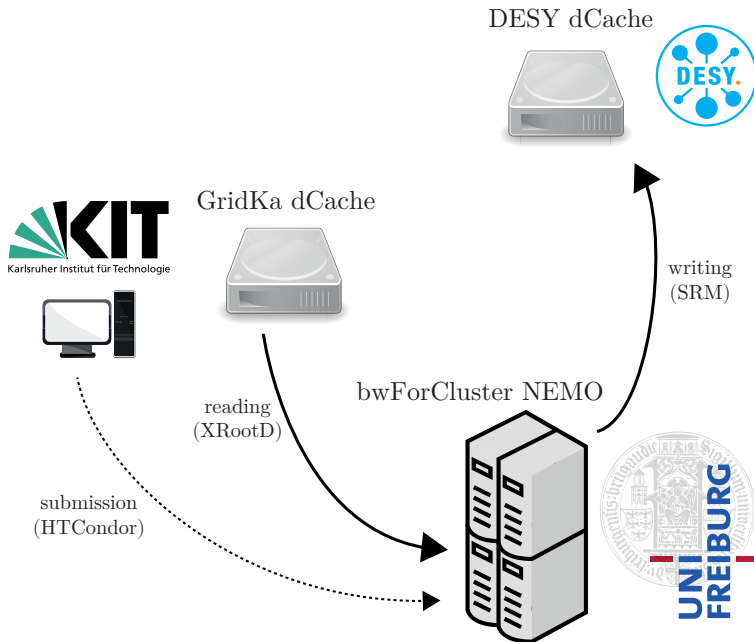


Figure 2: Schematic view of the computational setup used for the production of embedded events at the NEMO cluster. The input files amounting to 206 TB are stored on the GridKa storage located in Karlsruhe. They are then transferred in batches of 1000 events (~ 3 GB) to the computing center NEMO in Freiburg. Finally, a single output file which has been compressed to ~ 20 MB is stored at the HEP storage at DESY. From here, the files are published to the Data Aggregation System of the CMS collaboration to be used by the analysts.

The average CPU-time of an event in the event loop is 15.4s. The CPU efficiency of the setup, measured with a local file, is 92%. Most of the idle time is spent on reading and writing the output files. When running this workflow at a computing center, this efficiency is further decreased by transfer times of the input files as a result of the limited bandwidth between the computing center and the HEP storage. A schematic view of the computational setup is shown in Figure 2.

The four steps are coded as python scripts and submitted via the job submission tool `grid-control` (Stober et al., 2017), using the distributed computing software HTCondor² as a backend. Details on the dynamic integration of resources as well as the specialized software environments that allow the processing of these jobs on opportunistic resources such as NEMO are given in Ref. (Heidecker et al., 2019).

The preselected files are stored on the HEP storage of located at KIT and are read from the computing center via the data access framework XRootD (Dorigo et al., 2005). The files are then processed on NEMO machines and written to the HEP storage at the DESY Tier-2 center via `gridFTP` (Allcock et al., 2005). The output size of all 65 million input events combined lies in the order of 1 TB. This decrease in data size by a factor of 200 is a result of both the compression from the full event description as delivered by the detector to a format suitable and sufficient for physics analyses, as well as a loss of events due to kinematic filters which are implemented in both the selection and the simulation step, removing events that are not relevant for the analysis due to their kinematic properties.

Run label	\mathcal{L} in fb ⁻¹	# files	# events	Size in TB	# jobs
Run2017 B	4.8	2930	5,632,077	15.82	5,633
Run2017 C	9.6	9034	16,627,325	47.25	16,628
Run2017 D	4.2	11684	7,178,226	20.17	7,179
Run2017 E	9.3	25866	15,323,608	51.87	15,324
Run2017 F	13.5	11684	20,408,930	70.94	20,409
Total per final state	41.3	89505	65,170,166	206.10	65,173
Total (6 final states)					391,038

Table 1: Selected $Z \rightarrow \mu\mu$ candidate events used for production of embedded event samples.

The total number of events available for processing are given in Table 1. For the production of $\mu \rightarrow \tau$ embedded events on the 2017 dataset collected by the CMS

²<https://research.cs.wisc.edu/htcondor/>

experiment, a total of 65 million events are available, which are processed six times. First, they are processed for each of the di- τ lepton final states in which the analysis is performed. As a τ lepton decays into an electron (e), muon (μ) or into light hadrons (τ_h), the four final states are $\tau\tau \rightarrow (e\mu, e\tau_h, \mu\tau_h, \tau_h\tau_h)$. The two final states $\tau\tau \rightarrow (ee, \mu\mu)$ are neglected due to their high contamination by Z boson decays into these leptons. Furthermore, the events are processed for $\mu \rightarrow \mu$ and $\mu \rightarrow e$ validation samples in which the initial muons are replaced by muons and electrons respectively. The four final states and two validation samples results in processing all 65 million input events at a total of six times.

For submission, the events are split depending on their run period of the LHC before being further split into input batches of 1000 events. This split by events results in a total number of 65,173 jobs for each validation and final state sample, and therefore 391,038 jobs in total, with a single job being expected to run for around six hours.

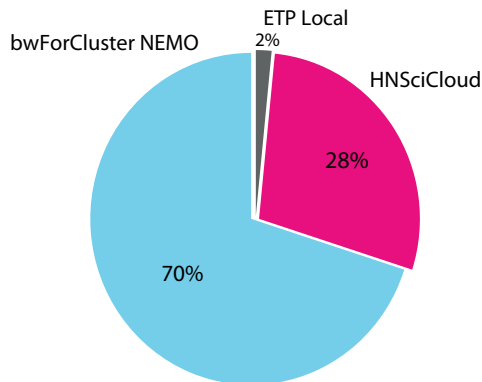


Figure 3: Relative fractions of total number of finished jobs at the available providers. The majority of cores was provided by NEMO.

3.1 Evaluation of site performance

In April 2017, three computing resources were available for the processing of these jobs, and the submission via HTCondor allows to submit them to any cluster, with the job being sent to the next available core. The three resources consisted of

- **bwForCluster NEMO**³ for Elementary Particle Physics, Neuroscience and Microsystems Engineering, supported by the bwHPC project.
- **Helix Nebula Science Cloud**⁴, a partnership between commercial cloud providers and research centers.
- **Local ETP resources** of the computing infrastructure of the Institute for Experimental Particle Physics (ETP) at the Karlsruhe Institute of Technology.

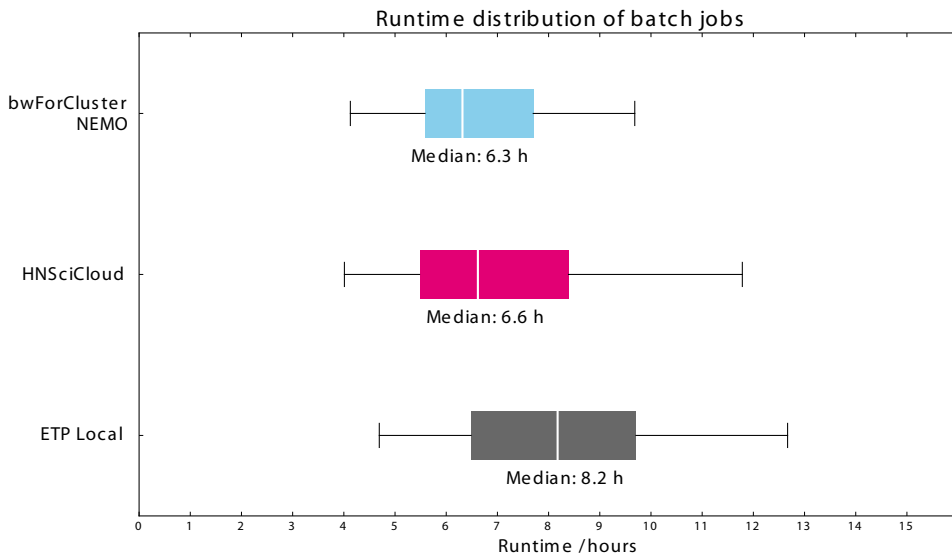


Figure 4: Distribution of job runtime shown in a box-and-whisker representation. The edges of the central box represent the first and third quartile respectively, with the median represented by a white line within the box. This results in 50% of all job runtimes lying within the box. The two outer whiskers represent the 5th and 95th percentiles, meaning only 5% of all values fall below the lower whisker and 5% fall above the upper whisker, and 90% of values fall between them. The job runtimes are compared for the three sites that were utilized for the production of embedded events. The shortest average job runtime was achieved at the machines of NEMO, which also showed the least upward fluctuations, resulting in a reliable processing of events.

The fraction of provided resources of these three is shown in Figure 3. The majority of jobs were completed at NEMO. Without these resources, the production of embedded events would have been prolonged by a factor of four, resulting in a delay

³<https://www.hpc.uni-freiburg.de/nemo>

⁴<http://www.helix-nebula.eu/about-hnscicloud>

between data-taking and the reliable background estimation by embedded events of over four months, which would negate a large benefit of this method. Figure 4 compares the runtime of jobs at the three sites. The cores at NEMO provide the best performance, as indicated by both the shortest average runtime of jobs as well as the small fluctuations in runtime.

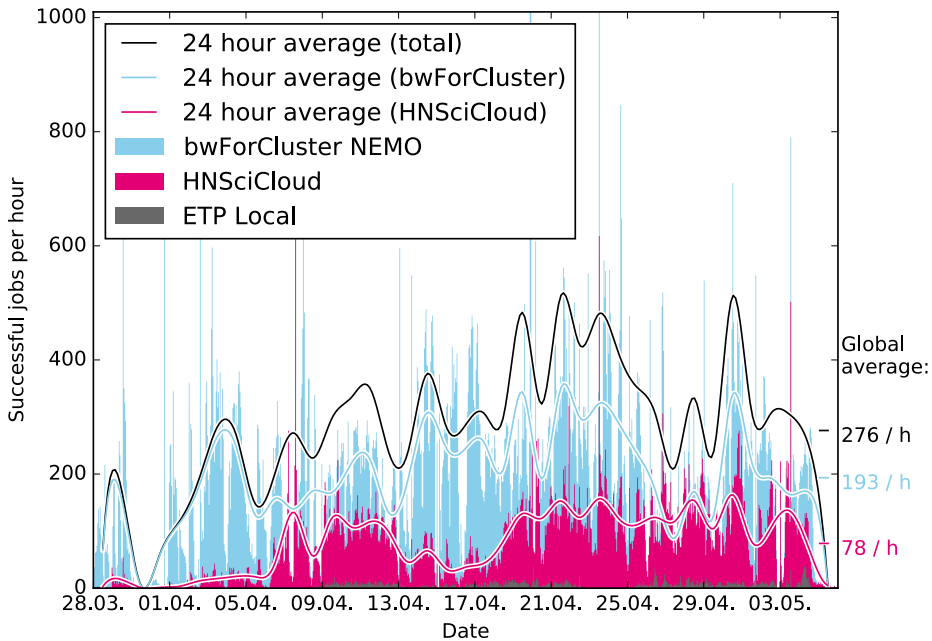


Figure 5: Histogram of successful jobs per hour for the production of embedded events over a timespan of five weeks between 28.03.2018 and 05.05.2018. In addition, the floating 24 hour averages are shown. All sites used have shown a reliable performance and continuous availability during the five weeks in which the $\mu \rightarrow \tau$ embedded samples were produced.

Figure 5 shows a histogram of successful jobs over a timespan of one month at the three sites and in bins of one hour by the time-stamp of job completion. Here, the dependence of the production on NEMO becomes apparent – it supplied the majority of cores, resulting in an average of 193 jobs being successfully completed at each given hour in the five-week timespan.

4 Conclusion

The τ lepton embedding method provides a valuable way to describe the expected background from $Z \rightarrow \tau\tau$ decays. The advantages of the method over the use of simulation lie in the description of complicated event characteristics of proton-proton collisions at the LHC, such as of the underlying event and the production of additional jets. Many corrections that are needed for simulated events become obsolete with the use of embedded events. In this regard, embedded events supplied an important cross-check already during the search for heavy neutral Higgs bosons using 2016 data, an improvement which has only been possible as a result of the opportunistic computing resources supplied by the bwHPC project.

The samples using data collected by the CMS detector in 2017, produced in April 2018 at the bwForCluster NEMO in Freiburg, will ultimately be used for the publication of the upcoming analysis of decays of the Standard Model Higgs boson into two τ leptons.

Acknowledgements

The work presented in this paper would not have been possible without the support by the state of Baden-Württemberg through bwHPC and the German Research Foundation (DFG) through grant no INST 39/963-1 FUGG.


Corresponding Author


Janek Bechtel: janek.bechtel@kit.edu


Institut für Experimentelle Teilchenphysik,


Karlsruher Institut für Technologie, Wolfgang-Gaede-Str. 1, Karlsruhe, Germany

ORCID

Janek Bechtel  <https://orcid.org/0000-0001-5245-7318>

Sebastian Brommer  <https://orcid.org/0000-0001-8988-2035>

Artur Gottmann  <https://orcid.org/0000-0001-6696-349X>

Roger Wolf  <https://orcid.org/0000-0001-9456-383X>

License    4.0 <https://creativecommons.org/licenses/by-sa/4.0>

References

- Allcock, W., J. Bresnahan, R. Kettimuthu and M. Link (2005). »The Globus Striped GridFTP Framework and Server«. In: *Supercomputing, 2005. Proceedings of the ACM/IEEE SC 2005 Conference*, pp. 54–54. DOI: 10.1109/SC.2005.72.
- ATLAS collaboration (2012). »Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC«. In: *Physics Letters B* 716.1, pp. 1–29. ISSN: 0370-2693. DOI: 10.1016/j.physletb.2012.08.020.
- (2015). »Modelling $Z \rightarrow \tau\tau$ processes in ATLAS with τ -embedded $Z \rightarrow \mu\mu$ data«. In: *JINST* 10.09, P09018. DOI: 10.1088/1748-0221/10/09/P09018. arXiv: 1506.05623 [hep-ex].
- (2018). »Cross-section measurements of the Higgs boson decaying to a pair of tau leptons in proton–proton collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector«. In: *Technical Report ATLAS-CONF-2018-021*.
- Bechtel, J. (2017). »Cross-check of the CMS search for additional MSSM Higgs bosons in the di- τ final state using $\mu \rightarrow \tau$ embedded events«. MA thesis. Karlsruhe Institute of Technology. URL: <https://ekp-invenio.physik.uni-karlsruhe.de/record/48943>.
- CMS collaboration (2008). »The CMS experiment at the CERN LHC«. In: *Journal of Instrumentation* 3.08, S08004. URL: <http://stacks.iop.org/1748-0221/3/i=08/a=S08004>.
- (2011). »Search for Neutral Minimal Supersymmetric Standard Model Higgs Bosons Decaying to Tau Pairs in pp Collisions at $\sqrt{s} = 7$ TeV«. In: *Physical Review Letters* 106, p. 231801. DOI: 10.1103/PhysRevLett.106.231801. arXiv: 1104.1619 [hep-ex].
- (2012a). »Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC«. In: *Physics Letters B* 716, p. 30. DOI: 10.1016/j.physletb.2012.08.021. arXiv: 1207.7235 [hep-ex].
- (2012b). »Search for neutral Higgs bosons decaying to tau pairs in pp collisions at $\sqrt{s} = 7$ TeV«. In: *Physics Letters B* 713, p. 68. DOI: 10.1016/j.physletb.2012.05.028. arXiv: 1202.4083 [hep-ex].
- (2013). »Observation of a new boson with mass near 125 GeV in pp collisions at $\sqrt{s} = 7$ and 8 TeV«. In: *JHEP* 06, p. 081. DOI: 10.1007/JHEP06(2013)081. arXiv: 1303.4571 [hep-ex].
- (2014a). »Evidence for the 125 GeV Higgs boson decaying to a pair of τ leptons«. In: *JHEP* 05, p. 104. DOI: 10.1007/JHEP05(2014)104. arXiv: 1401.5041 [hep-ex].
- (2014b). »Search for neutral MSSM Higgs bosons decaying to a pair of tau leptons in pp collisions«. In: *JHEP* 10, p. 160. DOI: 10.1007/JHEP10(2014)160. arXiv: 1408.3316 [hep-ex].

- (2017). »Particle-flow reconstruction and global event description with the CMS detector«. In: *JINST* 12, P10003. DOI: 10.1088/1748-0221/12/10/P10003. arXiv: 1706.04965 [physics.ins-det].
 - (2018a). »Observation of the Higgs boson decay to a pair of τ leptons with the CMS detector«. In: *Physics Letters B* 779, pp. 283–316. DOI: 10.1016/j.physletb.2018.02.004. arXiv: 1708.00373 [hep-ex].
 - (2018b). »Search for additional neutral MSSM Higgs bosons in the $\tau\tau$ final state in proton-proton collisions at $\sqrt{s}=13$ TeV«. In: *Journal of High Energy Physics* 2018.9, p. 7. ISSN: 1029-8479. DOI: 10.1007/JHEP09(2018)007.
- Dorigo, A., P. Elmer, F. Furano and A. Hanushevsky (2005). »XROOTD - A highly scalable architecture for data access«. In: *WSEAS Transactions on Computers* 4, pp. 348–353.
- Englert, F. and R. Brout (1964). »Broken Symmetry and the Mass of Gauge Vector Mesons«. In: *Physical Review Letters* 13, pp. 321–323. DOI: 10.1103/PhysRevLett.13.321.
- Guralnik, G. S., C. R. Hagen and T. W. Kibble (1964). »Global Conservation Laws and Massless Particles«. In: *Physical Review Letters* 13, pp. 585–587. DOI: 10.1103/PhysRevLett.13.585.
- Heidecker, C., M. J. Schnepf, F. von Cube, M. Giffels and G. Quast (2019). »Dynamic Resource Extension for Data Intensive Computing with Specialized Software Environments on HPC systems«. In: *Proceedings of the 5th bwHPC Symposium. HPC Activities in Baden-Württemberg*. Freiburg, September 2018. 5th bwHPC Symposium. TLP, Tübingen, pp. 161–172. DOI: 10.15496/publikation-29051.
- Higgs, P. W. (1964). »Broken Symmetries and the Masses of Gauge Bosons«. In: *Physical Review Letters* 13, pp. 508–509. DOI: 10.1103/PhysRevLett.13.508.
- Stober, F. et al. (2017). »The swiss army knife of job submission tools: grid-control«. In: *CoRR* abs/1707.03198. arXiv: 1707.03198 [cs.DC].