EPiC Series in Computing

Volume 38, 2016, Pages 8–28

Proceedings of the 1st and 2nd Vampire Workshops

Things you Can't do With a Vampire

Geoff Sutcliffe

University of Miami, Coral Gables, USA
geoff@cs.miami.edu

Abstract

The Vampire ATP system has been very successful at proving theorems in first-order logic. Vampire has won the important FOF division of 14 of the last 14 CASCs, and 11 of the last 13 CNF divisions. There have been very many papers about Vampire, the use of Vampire, and results achieved with Vampire. This paper examines the flip side of the Vampire coin ... what kinds of problems are difficult or even impossible for the latest incarnation of Vampire. The talk will help users decide when to use Vampire, and when to use another ATP system, will help the Vampire developers direct their work, and provides the data required to build a portfolio ATP system with Vampire as a component.

1 Introduction

Vampire [13] is automatic theorem prover for first-order logic. It implements ordered binary resolution and superposition, with standard redundancy criteria and simplification techniques. Splitting is controlled by the AVATAR architecture [40]. Vampire produces verifiable proofs/-models in TPTP format [35]. The first implementation of Vampire was completed by Voronkov in Paris in 1993, and was extended to the code tree implementation [39] in Uppsala in 1994. An important early stimulus for further development was an informal competition with the SETHEO ATP system [15] in Munich in 1996 - a precursor to the CADE ATP System Competition (CASC) [32]. The second implementation of Vampire was written by Voronkov and Riazanov in Uppsala, Vienna, and Manchester in 1997, leading to the first win for Vampire in the CNF division of CASC-16 [28] in 1999. The long standing seminal paper on Vampire, "The Design and Implementation of Vampire" appeared in the Artificial Intelligence journal in 2002 [22]. In 2006 Voronkov spent a year at Microsoft Research in Redmond, a period that had a great influence on the further development of Vampire, leading to the third implementation by Voronkov and Hoder in 2007. An important application of Vampire has been for symbol elimination, which started around 2009 [10]. For many years Vampire used a technique of "splitting without backtracking" [21], and in 2011 this was extended to the AVATAR architecture. The most recent implementation of Vampire was completed by Voronkov, Reger, and Suda in Manchester in 2015.

The various implementations of Vampire have achieved significant results, with commensurate fame and glory. Vampire has won the important FOF division of 14 of the last 14 CASCs, and 11 of the last 13 CNF divisions. There have been very many papers about Vampire, the use of Vampire, and results achieved with Vampire.[1] Vampire is embedded as an automatic

[1] I couldn't think of a nice metric for this.

component of many more complex reasoning systems, probably most significantly as an ATP system available in the Sledgehammer module of the widely used Isabelle interactive theorem prover [19]. The results that have been achieved in symbol elimination and interpolation are impressive [10]. Lastly, one cannot ignore the excellent students who have developed their skills working on Vampire, including Alexandre Riazanov, Krystof Hoder, Martin Suda, and Giles Reger.

This paper examines the flip side of the Vampire coin ... what kinds of problems are difficult or even impossible for the latest incarnation of Vampire. The paper will help users decide when to use Vampire, and when to use another ATP system, will help the Vampire developers direct their work, and provides the data required to build a portfolio ATP system with Vampire as a component. Section 2 explains the principles used for evaluating ATP problems and systems, then Section 3 applies those principles to data generated with Vampire 4.0 and recent versions of other ATP systems. Section 4 concludes.

2 Evaluation of ATP Problems and Systems

In order to build more powerful ATP systems, it is important to understand which systems work well for what types of problems. This knowledge is a key to further development, as it precedes any investigation into why the techniques and systems work well or badly. This knowledge is also crucial for users: given a specific problem, a user would like to know which systems are most likely to solve it. This section deals with the empirical evaluation of general purpose ATP systems. This requires also dealing with the issues of assigning ATP problems into classes that are reasonably homogeneous with respect to ATP systems, assigning difficulty ratings to ATP problems, and assigning ratings to ATP systems. Additionally, this section also examines the basic requirements that users have for ATP systems.

2.1 Basic ATP System Requirements

While the ability to solve problems is a key factor in the evaluation of ATP systems, there are other features that come into play, especially from the perspective of a non-expert user from an application domain.

From a theoretical perspective, users (we all) require that ATP systems are sound. Evidence of soundness can be obtained by testing an ATP system over a large set of test problems and checking that none of the results contradict the known/expected status of the problems. Soundness wrt individual solutions is more assured if the system outputs verifiable (and verified!) proofs/models. In contrast, it is understood that while the algorithms implemented in ATP systems may be complete in theory, in practice completeness is impossible due to issues related to the calculus, search control, implementation, and resource limits. Completeness might even be undesirable in terms of problem solving performance.

From a user perspective, ATP systems should be easy to download, unpack, build, and install. To that end it is preferable that ATP systems be developed using commonly available compilers and build tools (and not necessarily the bleeding edge versions). The build process should be supported by automatic configuration tools and compilation support (e.g., `Makefiles`). The built system should be encapsulated within an independent and movable directory/file hierarchy (e.g., no hidden files in the user's home directory).

Once built and installed, ATP systems should be easy to deploy and use. ATP systems should offer a command line interface that allows novice users to obtain immediate results with a simple invocation, but also provide advanced configuration options for power users. In the

world of ATP for classical logics, the ability to input problems in the TPTP language [38], report results using the SZS ontology [30], produce proofs/models in the TPTP language [35], and generally comply with TPTP conventions, is desirable for interoperability with other ATP systems and tools. Error messages output by ATP systems should be meaningful, and systems should react appropriately to signals (e.g., `SIGCPU`, `SIGTERM`). When an ATP system terminates, it should not leave any processes running or intermediate files in the file system. Finally, ATP systems should offer liberal licensing terms, so that users can adopt, adapt, and apply systems without undue constraint.

2.2 Source of ATP Problems

In order to evaluate ATP systems it is necessary to have an appropriate source of ATP problems for the ATP systems to (attempt to) solve. The Thousands of Problems for Theorem Provers (TPTP) problem library is the de facto standard set of test problems for classical ATP systems [31]. The TPTP supplies the ATP community with a comprehensive library of the test problems that are available today, providing an overview and a simple, unambiguous reference mechanism.

The TPTP is large enough to obtain statistical significance, spans a diversity of subject matters, and has an organizational structure designed for evaluating ATP systems. As the real applications of ATP grow, those types of problems are added to the TPTP, so that the TPTP is always a source of relevant problems for evaluating ATP systems. Using the TPTP for the evaluation of ATP systems helps to ensure that the performance results accurately reflect the capabilities of the ATP systems being considered. The TPTP is the best source of problems for the evaluation of general purpose ATP systems.

The TPTP was first released on Friday 12th November 1993. The most recent release of the TPTP, which was used in this work, is v6.2.0. It contains 20654 problems in 51 problems domains, spanning four logical forms: Clause Normal Form (CNF), First-Order Form (FOF), Typed First-order Form (TFF), and Typed Higher-order Form (THF). The TPTP is available online at `http://www.tptp.org`.

2.3 Types of ATP Problems

Various ATP systems and techniques are particularly well suited to problems with certain characteristics, often to the exclusion of problems with other characteristics (e.g., the Waldmeister system [16] can attempt only CNF unit equality problems). Empirical evaluation and comparison of ATP systems must therefore be done in the context of sets of problems that are reasonably accessible and homogeneous with respect to the systems.

ATP problems have easily identifiable logical, language, and syntactic characteristics, which have been used to divide the TPTP problems into homogeneous (wrt ATP systems) Specialist Problem Classes (SPCs). The SPCs take into account the following problem characteristics (the acronyms shown are used in Section 3):

- Logical form: Typed Higher-order Form (THF) vs. Typed First-order form - Polymorphic (TF1) vs. Typed First-order form - Monomorphic (TF0) vs. First-Order Form (FOF) vs. Effectively PRoposition clause normal form (EPR) vs. Clause Normal Form (CNF).
- TF0 arithmetic: With ARIthmetic (ARI) vs. No ARithmetic (NAR).
- CNF reducibility: Real First-Order (RFO) vs. Effectively PRopositional (EPR).
- SZS status: THeoreM (THM) vs. CounterSAtisfiable (CSA) vs. UNSatisfiable (UNS) vs. SATisfiable (SAT).

- Equality: No EQuality (NEQ) vs. Some EQuality (SEQ) vs. Pure EQuality (PEQ), vs. Any EQUality (EQU) - the union of SEQ and PEQ.
- CNF Hornness: HoRN (HRN) vs. Non-HorN (NHN).
- CNF pure equality: Unit EQuality (UEQ) vs. Non-Unit Equality (NUE).

Each path from the top to the bottom of Figure 1 corresponds to an SPC. The homogeneity of these SPCs wrt ATP systems has previously been verified [9]. The evaluation scheme described in Section 2.6 evaluates the ATP within SPCs, and evaluates only those ATP systems that can, in principle, attempt problems with the SPCs characteristics.

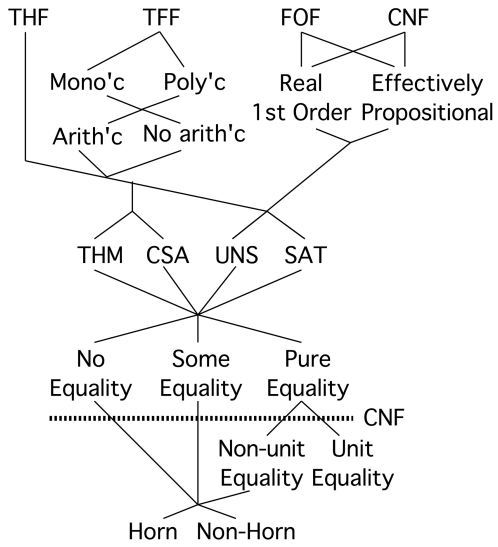

Figure 1: SPC Paths

In addition to being necessary for meaningful evaluation of ATP systems, results in the context of SPCs provides useful information for users, who can identify their problems' SPC, and select an ATP systems based on the corresponding evaluation results. The SystemOnTPTP recommendation tool [27], available at `http://www.tptp.org/cgi-bin/SystemOnTPTP` does this: it takes an ATP problem, determines its SPC, and reports the ratings (see Section 2.6) for the ATP systems that have been evaluated in that SPC. These system recommendations have been leveraged in the SSCPA ATP system [36], which runs a number of the highest rated systems in competition parallel.

2.4 Source of Solution Data

Given the TPTP as the source of problems to be used for evaluating ATP systems, it is necessary to get performance data for the ATP systems on the problems in the SPCs that each of the systems can attempt. The Thousands of Solutions from Theorem Provers (TSTP) solution library is a collection of ATP systems' solutions to TPTP problems. A major use of the TSTP is for ATP system developers to examine solutions to problems, and thus understand how they can be solved, leading to improvements to their own systems. In the context of this work the TSTP provided the performance data necessary for evaluating Vampire and other ATP systems.

The first section of each TSTP solution file is a header that contains information about the TPTP problem, information about the ATP system, characteristics of the computer used, the SZS status and output dataform from the system, and statistics about the solution including the CPU time used. The second section of each TSTP solution file contains the annotated formulae that make up the solution. A key feature of the TSTP is that solutions from many of the ATP systems are written in the TPTP language - the same language as used for TPTP problems. This supports interoperability, e.g., pipelining, of ATP systems and tools that read and write the TPTP language. At the time of writing, the TSTP contained the results of running over 50 ATP systems and system variants on the problems in the appropriate SPCs of the TPTP. This has produced over 200000 files for solved problems, of which over 100000 contain explicit proofs or models (rather than only an assurance of a solution). The TSTP is available online at `http://www.tptp.org/TSTP`.

2.5 Resource Limits

The intuitively acceptable criteria for empirical evaluation of ATP systems are:

- What problems can they solve?
- What computational resources (CPU capability and CPU time) and memory resources do they need to find the solutions?

The first criterion, what problems the systems can solve, measures the completeness of the systems. If no resource limits are imposed then correctly implemented theoretically complete systems solve all problems, providing no differentiation between the systems. In practice however, as was noted in Section 2.1, issues that affect practical completeness are calculus, search control, implementation, and resource limits. The supply of resources is not under the control of the ATP systems, and needs to be factored out of system evaluation. The first criterion therefore apparently needs to be refined to "What problems can they solve, modulo realistic resource limits?". It turns out that adequate evaluation can be achieved without this added qualification.

Figure 2 plots the CPU times taken by several contemporary ATP systems to solve TPTP FOF problems, for each solution found, in increasing order of time taken. The relevant feature of these plots is that each system has a point at which the time taken to find solutions starts to increase dramatically. This is called the system's Peter Principle Point (PPP) [20], as it is the point at which the system has reached its level of incompetence. Evidently a linear increase in the computational resources beyond the PPP would not lead to the solution of significantly more problems. The PPP thus defines a "realistic computational resource limit" for the system. For ATP system evaluation, this insight means that provided enough CPU time and memory are provided for each ATP system to reach its PPP, evaluation is possible using the criterion "What problems can they solve?". Figure 2 indicates that a 300s CPU time limit is adequate. The computers used for generating the TSTP have at least 128GB memory, which is more than adequate for all but the most extreme uses of contemporary ATP.

2.6 The Evaluation Scheme

The evaluation of ATP systems is done using the TPTP evaluation scheme [37], which provides a difficulty rating for each problem, and a rating for each system in each SPC. It thus provides a well-defined measure of how difficult the problems are for the ATP systems, and how effective the ATP systems are for different types of problems. Over time, decreasing ratings for individual problems provide an indication of progress in the field [33].

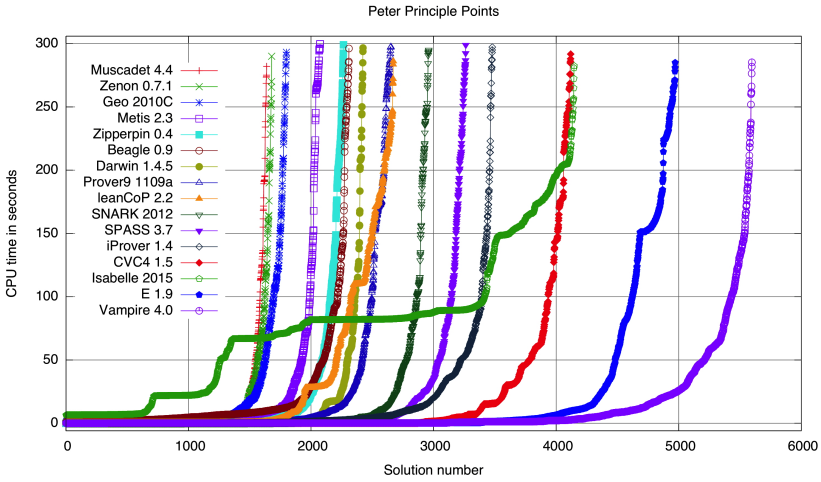

Figure 2: Peter Principle Points

As a preprocessing step, problems in the TPTP that are tagged as "biased", i.e., designed to be well-suited or ill-suited for particular ATP systems or calculi, are excluded. The TPTP problems are then divided into the SPCs, and the TSTP files for each SPC are analyzed. For each SPC, the performances of systems whose set of solved problems is not a subset of that of any other system are used to rate the problems. These systems are called *State-of-the-Art (SOTA) contributors*, because a portfolio of these systems would be able to solve all the problems that any ATP system can solve. The fraction of the SOTA contributors that fail on a problem is the difficulty rating for a problem: problems that are solved by all/some/none of the systems get ratings of 0.00/0.01-0.99/1.00, and are referred to as easy/difficult/hard problems respectively. The fraction of the difficult problems that an ATP system solves in an SPC is the system's rating for that SPC.

This evaluation scheme has been applied to the problems in the TPTP, and the systems that have been used to produce the data in the TSTP. The results in Section 3 are taken from this evaluation.

3 Evaluation of Vampire

This section provides the evaluation results for Vampire 4.0 and and recent versions of other ATP systems, using the evaluation scheme described in Section 2. The evaluation has been limited to the SPCs of the TPTP that have enough problems to draw general conclusions and be significant to users and developers. The SPCs have been grouped according to their language and SZS status characteristics, because there is reasonable consistency between the results for the SPCs with the same values for two characteristics. For each SPC a commentated summary of Vampire's performance is given in the context of the SPCs and other ATP systems that have been evaluated in the SPC. The detailed results are provided in Appendix A.

The TSTP data used for the evaluation was generated on the StarExec cluster [26]. Each computer has

- Two quad-core Intel(R) Xeon(R) E5-2609, 2.40GHz CPUs

- Either 128GB or 256GB memory
- The Red Hat Enterprise Linux Workstation release 6.3 (Santiago) operating system, kernel 2.6.32-431.1.2.el6.x86_64

3.1 Basic Capabilities

This section considers Vampire's status with respect to the basic ATP system requirements described in Section 2.1.

Vampire is (probably) sound - none of Vampire's results in the TSTP contradict the known status of the problem. Vampire outputs refutations for theorems and unsatisfiable formulae, and saturations/finite models for countersatisfiable problems and satisfiable formulae. The proofs/models are in TPTP format, allowing use of the GDV verifier [29] for the proofs, but so far this verification has not been done. Vampire's underlying calculus is complete, but the implementation is naturally incomplete, e.g., due to its limited resource strategy [23].

In terms of deployment, there is no download available from the Vampire web site[2] right now. However, Vampire is written in C++, and should be easy enough to build. The version currently being distributed is a fully encapsulated binary, which is easy to install.

Vampire provides both simple and advanced usage options. In particular, Vampire's auto-mode builds a schedule of strategies suited to the given problem, and implements the necessary strategy scheduling. This makes Vampire easy for non-experts to use. Vampire has is a plethora of advanced options - try running "`vampire --show_options on`"! Vampire is highly TPTP compliant, reading the TPTP's TFF, FOF, and CNF formats, reporting its results using the SZS ontology, and outputting it's proofs/models in TPTP format. The error messages output from Vampire have not been evaluated, because none of the output files seem to have any! Vampire reacts appropriately to signals, and does not leave any dingo poop processes[3] or files. The Vampire licence is quite liberal: it simply disallows modification and distribution of Vampire, or the use of Vampire to compete against Vampire. To obtain a copy of Vampire it is necessary to accept the terms of the licence, but other license options can be negotiated with the developer.

3.2 Types of Problems Vampire Can Solve

In this section, and the subsequent Sections 3.3 and 3.4, each paragraph is headed by the SPC group, and the SPCs that have been analysed are listed. The commentaries identify the top performing SOTA contributors for each SPC, and note the highest system ratings in each SPC.

SPC: CNF_SAT (_EPR, _RFO_NEQ, _RFO_EQU_NUE, _RFO_PEQ_UEQ)
The system of choice for CNF_SAT is Vampire 4.0 (in SAT mode). Vampire is a SOTA contributor in all four SPCs, and in the first three of them it has the highest ratings of 0.92, 0.98, and 0.85. In CNF_SAT_RFO_PEQ_UEQ Mace4 1109a [17] has the highest rating of 0.84. Vampire has a low rating of 0.14, with the well recognized finite model finder Paradox 4.0 [5] also having a high rating of 0.78. It is noteworthy that in CNF_UNS_EPR iProver, which has dominated the EPR division of CASC for several years, has a rating of 0.80.

[2] http://www.vprover.org

[3] Recently there have been problems terminating Vampire on the StarExec cluster, but it is suspected that the problem lies within the StarExec control software rather than in Vampire.

SPC: FOF_THM and FOF_UNS (_EPR, _RFO_NEQ, _RFO_SEQ, _RFO_PEQ)
The system of choice for FOF_THM and FOF_UNS is Vampire 4.0. Vampire is a SOTA contributor in all seven SPCs, and in the FOF_THM_RFO_NEQ, FOF_THM_RFO_SEQ, and FOF_UNS SPCs it has the highest ratings of 0.96, 0.94., 0.77, 1.00, and 1.00. Note that a rating of 1.00 means that Vampire solved all the problems that any system could solve, and is hence the only SOTA contributor. In the case of FOF_UNS_RFO_PEQ Vampire solved all the problems in the SPC. It is interesting to note that the FOF_THM_RFO_SEQ is the largest SPC, with 4974 problems and 24 SOTA contributors - higher precision ratings might be obtained by further dividing this SPC.

In FOF_THM_EPR iProver 1.4 [12] has the highest rating of 0.86. Vampire has a low rating of 0.33, with CVC4 1.5 [1] and Isabelle 2015 [18] having higher ratings of 0.76 and 0.71. In FOF_THM_RFO_PEQ E [25], in its VanHElsing [14], standalone, and ET [11] incarnations, has the highest ratings of 0.92, 0.90, and 0.89. Vampire has a reasonably high rating of 0.82, with SRASS 0.1 [34] also having a high rating of 0.84.

SPC: FOF_CSA and FOF_SAT (_EPR, _RFO_NEQ, _RFO_SEQ, _RFO_NEQ, _RFO_SEQ)
The system of choice for FOF_CSA and FOF_SAT is Vampire 4.0 (in SAT mode). Vampire is a SOTA contributor in all five SPCs, and in all except FOF_CSA_RFO_NEQ it has the highest ratings of 0.91, 0.98, 0.94, and 0.90. In FOF_CSA_RFO_NEQ iProver 1.4 has the highest rating of 0.76. Vampire has a moderate rating of 0.55, and Paradox 4.0 has a rating of 0.47.

3.3 Types of Problems Vampire Can't Solve

This section provides analysis for the TFF and THF SPCs. Vampire cannot attempt these types of problems.

SPC: TF0_CSA (_EQU_ARI) The system of choice for TF0_CSA is Z3 4.4-TPTP [7]. In the one SPC of interest Z3 has the highest rating of 0.86. No other system comes close.

SPC: TF1 (_THM_EQU_NAR)
The system of choice for TF1 is Alt-Ergo 0.95.1 [6]. In fact, at the time when the data for this paper was generated it was the only known ATP system for TF1.[4] In the one SPC of interest Alt-Ergo has rating 1.00.

SPC: TH0_THM (_NEQ, _EQU)
The systems of choice for TH0_THM are Isabelle 2015 and Satallax 2.7 [4]. Both systems are SOTA contributors in the two SPCs. In TH0_THM_NEQ Satallax has the highest rating of 0.88, with Isabelle close behind at 0.87. In TH0_THM_EQU Isabelle has the highest rating of 0.87, with Satallax close behind at 0.85. The only other system with reasonable ratings in LEO-II 1.6.2 [2], with ratings of 0.77 and 0.72.

SPC: TH0_CSA (_NEQ, _EQU)
The system of choice for TH0_CSA is Nitpick 2015 [3]. It is the only SOTA contributor in the two SPCs, with a rating of 1.00 in both.

[4]Since then ZenonModulo [8] has emerged, but has yet to be evaluated.

3.4 Types of Problems Vampire and Other ATP Systems Can Solve

This section provides analysis for SPCs where Vampire performs well, but is not dominant. This is in contrast to the "exceptions" to Vampire's generally dominant performance in the SPCs analysed in Section 3.2.

SPC: CNF_UNS (_EPR, _RFO_NEQ_HRN, _RFO_NEQ_NHN, _RFO_SEQ_HRN, _RFO_SEQ_NHN, _RFO_PEQ_NUE, _RFO_PEQ_UEQ)
The systems of choice for CNF_UNS are E 1.9 and Vampire 4.0. E is a SOTA contributor in all the SPCs except CNF_UNS_EPR, and in the four HRN and PEQ SPCs it has the highest ratings of 0.96, 0.96, 0.92, and 0.89. Vampire is also a SOTA contributor in these four SPCs, with ratings of 0.82, 0.82, 0.82, and 0.84.

Vampire is a SOTA contributor in all the SPCs, and in the EPR and NHN SPCs it has the highest ratings of 0.98, 0.96, and 0.92. E is also a SOTA contributor in the NHN SPCs, with ratings of 0.92 and 0.86.

These results suggest like a combination of E and Vampire would do well for CNF_UNS problems. It is noteworthy that in CNF_UNS_EPR iProver, which has dominated the EPR division of CASC for several years, has a rating of 0.80.

SPC: TF0_THM and TF0_UNS (_NEQ_ARI, _EQU_NAR, _EQU_ARI, _EQU_NAR, _EQU_ARI)
The systems of choice for TF0_THM and TF0_UNS are Vampire 4.0, CVC4 1.5, and Princess 140704 [24]. Vampire is a SOTA contributor in all the SPCs, and in TF0_THM_EQU_NAR and TF0_UNS_EQU_ARI it has the highest ratings of 0.81 and 1.00. CVC4 is also a SOTA contributor in TF0_THM_EQU_NAR, but with a low rating of 0.19. Princess is not a SOTA contributor in these two SPCs.

CVC4 is a SOTA contributor in the first four SPCs, and in TF0_THM_EQU_ARI and TF0_UNS_EQU_NAR it has the highest ratings of 0.88 and 1.00. Vampire is also a SOTA contributor in TF0_THM_EQU_ARI, with a rating of 0.88 (but solving one less problem than CVC4), and in TF0_UNS_EQU_NAR with a rating of 1.00 (both CVC4 and Vampire solved all the problems in TF0_UNS_EQU_NAR). Princess is also SOTA contributor in TF0_THM_EQU_ARI with a rating of 0.83.

Princess is a SOTA contributor in the first and third SPCs, with the highest rating of 0.96 in TF0_THM_NEQ_ARI.

Evidently a portfolio approach would do well for TF0_THM and TF0_UNS problems.

4 Conclusion

The conclusions that can be drawn from this paper are:

- Vampire is good for many things.
- Vampire is bad for some things.
- Other ATP systems are sometimes better than Vampire.
- When in doubt, consult the SystemOnTPTP recommendation tool, or try a few different ATP systems.

References

[1] C. Barrett, C. Conway, M. Deters, L. Hadarean, D. Jovanovic, T. King, A. Reynolds, and C. Tinelli. CVC4. In G. Gopalakrishnan and S. Qadeer, editors, *Proceedings of the 23rd International Conference on Computer Aided Verification*, number 6806 in Lecture Notes in Computer Science, pages 171–177. Springer-Verlag, 2011.

[2] C. Benzmüller, L. Paulson, F. Theiss, and A. Fietzke. LEO-II - A Cooperative Automatic Theorem Prover for Higher-Order Logic. In P. Baumgartner, A. Armando, and D. Gilles, editors, *Proceedings of the 4th International Joint Conference on Automated Reasoning*, number 5195 in Lecture Notes in Artificial Intelligence, pages 162–170. Springer-Verlag, 2008.

[3] J. Blanchette and T. Nipkow. Nitpick: A Counterexample Generator for Higher-Order Logic Based on a Relational Model Finder. In M. Kaufmann and L. Paulson, editors, *Proceedings of the 1st International Conference on Interactive Theorem Proving*, number 6172 in Lecture Notes in Computer Science, pages 131–146. Springer-Verlag, 2010.

[4] C.E. Brown. Satallax: An Automated Higher-Order Prover (System Description). In B. Gramlich, D. Miller, and U. Sattler, editors, *Proceedings of the 6th International Joint Conference on Automated Reasoning*, number 7364 in Lecture Notes in Artificial Intelligence, pages 111–117, 2012.

[5] K. Claessen and N. Sörensson. New Techniques that Improve MACE-style Finite Model Finding. In P. Baumgartner and C. Fermueller, editors, *Proceedings of the CADE-19 Workshop: Model Computation - Principles, Algorithms, Applications*, 2003.

[6] S. Conchon, M. Iguernelala, and A. Mebsout. A Collaborative Framework for Non-Linear Integer Arithmetic Reasoning in Alt-Ergo. In N. Bjorner, editor, *Proceedings of the 15th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing*, pages 161–168. IEEE, 2013.

[7] L. de Moura and N. Bjorner. Z3: An Efficient SMT Solver. In C. Ramakrishnan and J. Rehof, editors, *Proceedings of the 14th International Conference on Tools and Algorithms for the Construction and Analysis of Systems*, number 4963 in Lecture Notes in Artificial Intelligence, pages 337–340. Springer-Verlag, 2008.

[8] D. Delahaye, D. Doligez, F. Gibert, P. Halmagrand, and O. Hermant. Zenon Modulo: When Achilles Outruns the Tortoise using Deduction Modulo. In K. McMillan, A. Middeldorp, and A. Voronkov, editors, *Proceedings of the 19th International Conference on Logic for Programming, Artificial Intelligence, and Reasoning*, number 8312 in Lecture Notes in Computer Science, pages 274–290. Springer-Verlag, 2013.

[9] M. Fuchs and G. Sutcliffe. Homogeneous Sets of ATP Problems. In S. Haller and G. Simmons, editors, *Proceedings of the 15th International FLAIRS Conference*, pages 57–61. AAAI Press, 2002.

[10] K. Hoder, L. Kovacs, and A. Voronkov. Interpolation and Symbol Elimination in Vampire. In J. Giesl and R. Haehnle, editors, *Proceedings of the 5th International Joint Conference on Automated Reasoning*, number 6173 in Lecture Notes in Artificial Intelligence, pages 188–195, 2010.

[11] C. Kaliszyk, S. Schulz, J. Urban, and J. Vyskocil. System Description: E.T. 0.1. In A. Felty and A. Middeldorp, editors, *Proceedings of the 25th International Conference on Automated Deduction*, Lecture Notes in Computer Science, page To appear. Springer-Verlag, 2015.

[12] K. Korovin. iProver - An Instantiation-Based Theorem Prover for First-order Logic (System Description). In P. Baumgartner, A. Armando, and D. Gilles, editors, *Proceedings of the 4th International Joint Conference on Automated Reasoning*, number 5195 in Lecture Notes in Artificial Intelligence, pages 292–298, 2008.

[13] L. Kovacs and A. Voronkov. First-Order Theorem Proving and Vampire. In N. Sharygina and H. Veith, editors, *Proceedings of the 25th International Conference on Computer Aided Verification*, number 8044 in Lecture Notes in Artificial Intelligence, pages 1–35. Springer-Verlag, 2013.

[14] D. Külwein and J. Urban. MaLeS: A Framework for Automatic Tuning of Automated Theorem

Provers. *Journal of Automated Reasoning*, page To appear, 2015.

[15] R. Letz, J. Schumann, S. Bayerl, and W. Bibel. SETHEO: A High-Performance Theorem Prover. *Journal of Automated Reasoning*, 8(2):183–212, 1992.

[16] B. Loechner and T. Hillenbrand. A Phytography of Waldmeister. *AI Communications*, 15(2/3):127–133, 2002.

[17] W.W. McCune. Mace4 Reference Manual and Guide. Technical Report ANL/MCS-TM-264, Argonne National Laboratory, Argonne, USA, 2003.

[18] T. Nipkow, L. Paulson, and M. Wenzel. Isabelle/HOL: A Proof Assistant for Higher-Order Logic. http://www.cl.cam.ac.uk/research/hvg/Isabelle/dist/Isabelle/doc/tutorial.pdf.

[19] L. Paulson and J. Blanchette. Three Years of Experience with Sledgehammer, a Practical Link between Automatic and Interactive Theorem Provers. In G. Sutcliffe, E. Ternovska, and S. Schulz, editors, *Proceedings of the 8th International Workshop on the Implementation of Logics*, number 2 in EPiC, pages 1–11, 2010.

[20] L.J. Peter and R. Hull. *The Peter Principle*. Souvenir Press, 1969.

[21] A. Riazanov and A. Voronkov. Splitting without Backtracking. In B. Nebel, editor, *Proceedings of the 17th International Joint Conference on Artificial Intelligence*, pages 611–617. Morgan Kaufmann, 2001.

[22] A. Riazanov and A. Voronkov. The Design and Implementation of Vampire. *AI Communications*, 15(2-3):91–110, 2002.

[23] A. Riazanov and A. Voronkov. Limited Resource Strategy in Resolution Theorem Proving. *Journal of Symbolic Computation*, 36(1-2):101–115, 2003.

[24] P. Rümmer. A Constraint Sequent Calculus for First-Order Logic with Linear Integer Arithmetic. In I. Cervesato, H. Veith, and A. Voronkov, editors, *Proceedings of the 15th International Conference on Logic for Programming Artificial Intelligence and Reasoning*, number 5330 in Lecture Notes in Artificial Intelligence, pages 274–289. Springer-Verlag, 2008.

[25] S. Schulz. E: A Brainiac Theorem Prover. *AI Communications*, 15(2-3):111–126, 2002.

[26] A. Stump, G. Sutcliffe, and C. Tinelli. StarExec: a Cross-Community Infrastructure for Logic Solving. In S. Demri, D. Kapur, and C. Weidenbach, editors, *Proceedings of the 7th International Joint Conference on Automated Reasoning*, number 8562 in Lecture Notes in Artificial Intelligence, pages 367–373, 2014.

[27] G. Sutcliffe. SystemOnTPTP. In D. McAllester, editor, *Proceedings of the 17th International Conference on Automated Deduction*, number 1831 in Lecture Notes in Artificial Intelligence, pages 406–410. Springer-Verlag, 2000.

[28] G. Sutcliffe. The CADE-16 ATP System Competition. *Journal of Automated Reasoning*, 24(3):371–396, 2000.

[29] G. Sutcliffe. Semantic Derivation Verification. *International Journal on Artificial Intelligence Tools*, 15(6):1053–1070, 2006.

[30] G. Sutcliffe. The SZS Ontologies for Automated Reasoning Software. In G. Sutcliffe, P. Rudnicki, R. Schmidt, B. Konev, and S. Schulz, editors, *Proceedings of the LPAR Workshops: Knowledge Exchange: Automated Provers and Proof Assistants, and The 7th International Workshop on the Implementation of Logics*, number 418 in CEUR Workshop Proceedings, pages 38–49, 2008.

[31] G. Sutcliffe. The TPTP Problem Library and Associated Infrastructure. The FOF and CNF Parts, v3.5.0. *Journal of Automated Reasoning*, 43(4):337–362, 2009.

[32] G. Sutcliffe. The CADE ATP System Competition - CASC. *AI Magazine*, page To appear, 2015.

[33] G. Sutcliffe, M. Fuchs, and C. Suttner. Progress in Automated Theorem Proving, 1997-1999. In H. Hoos and T. Stützle, editors, *Proceedings of the IJCAI'01 Workshop on Empirical Methods in Artificial Intelligence*, pages 53–60, 2001.

[34] G. Sutcliffe and Y. Puzis. SRASS - a Semantic Relevance Axiom Selection System. In F. Pfenning, editor, *Proceedings of the 21st International Conference on Automated Deduction*, number 4603 in

Lecture Notes in Artificial Intelligence, pages 295–310. Springer-Verlag, 2007.

[35] G. Sutcliffe, S. Schulz, K. Claessen, and A. Van Gelder. Using the TPTP Language for Writing Derivations and Finite Interpretations. In U. Furbach and N. Shankar, editors, *Proceedings of the 3rd International Joint Conference on Automated Reasoning*, number 4130 in Lecture Notes in Artificial Intelligence, pages 67–81, 2006.

[36] G. Sutcliffe and D. Seyfang. Smart Selective Competition Parallelism ATP. In A. Kumar and I. Russell, editors, *Proceedings of the 12th International FLAIRS Conference*, pages 341–345. AAAI Press, 1999.

[37] G. Sutcliffe and C.B. Suttner. Evaluating General Purpose Automated Theorem Proving Systems. *Artificial Intelligence*, 131(1-2):39–54, 2001.

[38] A. Van Gelder and G. Sutcliffe. Extending the TPTP Language to Higher-Order Logic with Automated Parser Generation. In U. Furbach and N. Shankar, editors, *Proceedings of the 3rd International Joint Conference on Automated Reasoning*, number 4130 in Lecture Notes in Artificial Intelligence, pages 156–161. Springer-Verlag, 2006.

[39] A. Voronkov. The Anatomy of Vampire. *Journal of Automated Reasoning*, 15(2):237–265, 1995.

[40] A. Voronkov. AVATAR: The New Architecture for First-Order Theorem Provers. In A. Biere and R. Bloem, editors, *Proceedings of the 26th International Conference on Computer Aided Verification*, number 8559 in Lecture Notes in Computer Science, pages 696–710, 2014.

A System Data in SPCs

This appendix provides some details of the performance data for all the SPCs discussed in Section 3. For each SPC the stanza gives:

- The number of unbiased problems in the SPC.
- The combined number of problems solved by the SOTA contributors, with the fraction wrt the total number of problems.
- The number of problems solved by all the SOTA contributors, i.e., the number of problems with rating 0.00 (easy problems), with the fractions wrt to the total number of problems and the number of problems solved by SOTA contributors. Finally the number of difficult problems is given, with its fraction of that number (always 1.00).
- For each SOTA contributor, the number of problems it solved with the fractions wrt to the total number of problems and the number of problems solved by SOTA contributors, and the number of difficult problems solved with the corresponding fraction of difficult problems, i.e., the system rating.

A.1 CNF_SAT

```
CNF_SAT_EPR
Unbiased SPC size                239
SOTA contributors      solved    226 = 0.95 SPC 0.95 attempted
SOTA contributors all  solved    140 = 0.59 SPC 0.62 SOTA    86 =  1.00 SSR
Vampire---SAT-4.0    c solved    219 = 0.92 SPC 0.97 SOTA    79 =  0.92 SSR
Vampire---4.0        C solved    219 = 0.92 SPC 0.97 SOTA    79 =  0.92 SSR
iProver---1.4        C solved    209 = 0.87 SPC 0.92 SOTA    69 =  0.80 SSR
Paradox---4.0        C solved    180 = 0.75 SPC 0.80 SOTA    40 =  0.47 SSR
GrAnDe---1.1         C solved    152 = 0.64 SPC 0.67 SOTA    12 =  0.14 SSR

CNF_SAT_RFO_NEQ
Unbiased SPC size                274
SOTA contributors      solved    274 = 1.00 SPC 1.00 attempted
SOTA contributors all  solved    212 = 0.77 SPC 0.77 SOTA    62 =  1.00 SSR
Vampire---SAT-4.0    C solved    273 = 1.00 SPC 1.00 SOTA    61 =  0.98 SSR
Mace4---1109a        C solved    213 = 0.78 SPC 0.78 SOTA     1 =  0.02 SSR

CNF_SAT_RFO_EQU_NUE
Unbiased SPC size                493
SOTA contributors      solved    402 = 0.82 SPC 0.82 attempted
SOTA contributors all  solved     78 = 0.16 SPC 0.19 SOTA   324 =  1.00 SSR
Vampire---SAT-4.0    C solved    355 = 0.72 SPC 0.88 SOTA   277 =  0.85 SSR
DarwinFM---1.4.5     C solved    241 = 0.49 SPC 0.60 SOTA   163 =  0.50 SSR
CVC4---FNT-1.5pre    C solved    233 = 0.47 SPC 0.58 SOTA   155 =  0.48 SSR
SPASS---3.7          C solved    205 = 0.42 SPC 0.51 SOTA   127 =  0.39 SSR
E---1.9              C solved    202 = 0.41 SPC 0.50 SOTA   124 =  0.38 SSR
Mace4---1109a        C solved    200 = 0.41 SPC 0.50 SOTA   122 =  0.38 SSR
Nitpick---2015       C solved    193 = 0.39 SPC 0.48 SOTA   115 =  0.35 SSR

CNF_SAT_RFO_PEQ_UEQ
Unbiased SPC size                140
SOTA contributors      solved    127 = 0.91 SPC 0.91 attempted
SOTA contributors all  solved     76 = 0.54 SPC 0.60 SOTA    51 =  1.00 SSR
Mace4---1109a        C solved    119 = 0.85 SPC 0.94 SOTA    43 =  0.84 SSR
Paradox---4.0        C solved    116 = 0.83 SPC 0.91 SOTA    40 =  0.78 SSR
Vampire---SAT-4.0    C solved     83 = 0.59 SPC 0.65 SOTA     7 =  0.14 SSR
```

A.2 FOF_THM/UNS

```
FOF_THM_EPR
Unbiased SPC size                290
SOTA contributors      solved    282 = 0.97 SPC 0.97 attempted
SOTA contributors all  solved    261 = 0.90 SPC 0.93 SOTA   21 =  1.00 SSR
iProver---1.4        C solved    279 = 0.96 SPC 0.99 SOTA   18 =  0.86 SSR
CVC4---FOF-1.5pre    C solved    277 = 0.96 SPC 0.98 SOTA   16 =  0.76 SSR
Isabelle---2015      C solved    276 = 0.95 SPC 0.98 SOTA   15 =  0.71 SSR
Vampire---4.0        C solved    268 = 0.92 SPC 0.95 SOTA    7 =  0.33 SSR

FOF_THM_RFO_NEQ
Unbiased SPC size                965
SOTA contributors      solved    944 = 0.98 SPC 0.98 attempted
SOTA contributors all  solved    357 = 0.37 SPC 0.38 SOTA  587 =  1.00 SSR
Vampire---4.0        C solved    922 = 0.96 SPC 0.98 SOTA  565 =  0.96 SSR
E---1.9              C solved    890 = 0.92 SPC 0.94 SOTA  533 =  0.91 SSR
iProver---1.4        C solved    888 = 0.92 SPC 0.94 SOTA  531 =  0.90 SSR
VanHElsing---1.0     C solved    886 = 0.92 SPC 0.94 SOTA  529 =  0.90 SSR
ET---1.0             C solved    885 = 0.92 SPC 0.94 SOTA  528 =  0.90 SSR
iProver-Eq---0.85    C solved    852 = 0.88 SPC 0.90 SOTA  495 =  0.84 SSR
CVC4---FOF-1.5pre    C solved    848 = 0.88 SPC 0.90 SOTA  491 =  0.84 SSR
Z3---4.4-TPTP        C solved    833 = 0.86 SPC 0.88 SOTA  476 =  0.81 SSR
Darwin---1.4.5       C solved    794 = 0.82 SPC 0.84 SOTA  437 =  0.74 SSR
iProverMo---0.7-0.2  C solved    726 = 0.75 SPC 0.77 SOTA  369 =  0.63 SSR
SRASS---0.1          C solved    615 = 0.64 SPC 0.65 SOTA  258 =  0.44 SSR
Geo---2010C          C solved    544 = 0.56 SPC 0.58 SOTA  187 =  0.32 SSR
Equinox---5.0        C solved    447 = 0.46 SPC 0.47 SOTA   90 =  0.15 SSR

FOF_THM_RFO_SEQ
Unbiased SPC size               4974
SOTA contributors      solved   4171 = 0.84 SPC 0.84 attempted
SOTA contributors all  solved     47 = 0.01 SPC 0.01 SOTA 4124 =  1.00 SSR
Vampire---4.0        C solved   3934 = 0.79 SPC 0.94 SOTA 3887 =  0.94 SSR
ET---1.0             C solved   3566 = 0.72 SPC 0.85 SOTA 3519 =  0.85 SSR
E---1.9              C solved   3418 = 0.69 SPC 0.82 SOTA 3371 =  0.82 SSR
VanHElsing---1.0     C solved   3372 = 0.68 SPC 0.81 SOTA 3325 =  0.81 SSR
Isabelle---2015      C solved   2987 = 0.60 SPC 0.72 SOTA 2940 =  0.71 SSR
CVC4---FOF-1.5pre    C solved   2793 = 0.56 SPC 0.67 SOTA 2746 =  0.67 SSR
SRASS---0.1          C solved   2592 = 0.52 SPC 0.62 SOTA 2545 =  0.62 SSR
Fampire---1.3        C solved   2528 = 0.51 SPC 0.61 SOTA 2481 =  0.60 SSR
Princess---140704    C solved   2327 = 0.47 SPC 0.56 SOTA 2280 =  0.55 SSR
SInE---0.4           C solved   2318 = 0.47 SPC 0.56 SOTA 2271 =  0.55 SSR
SPASS---3.7          C solved   2226 = 0.45 SPC 0.53 SOTA 2179 =  0.53 SSR
iProver---1.4        C solved   2153 = 0.43 SPC 0.52 SOTA 2106 =  0.51 SSR
Z3---4.4-TPTP        C solved   1979 = 0.40 SPC 0.47 SOTA 1932 =  0.47 SSR
SNARK---20120808r022 C solved   1796 = 0.36 SPC 0.43 SOTA 1749 =  0.42 SSR
Prover9---1109a      C solved   1781 = 0.36 SPC 0.43 SOTA 1734 =  0.42 SSR
leanCoP---2.2        C solved   1629 = 0.33 SPC 0.39 SOTA 1582 =  0.38 SSR
```

```
Equinox---5.0        C solved  1456 = 0.29 SPC 0.35 SOTA 1409 =  0.34 SSR
Zipperpin---FOF-0.4  C solved  1388 = 0.28 SPC 0.33 SOTA 1341 =  0.33 SSR
Darwin---1.4.5       C solved  1282 = 0.26 SPC 0.31 SOTA 1235 =  0.30 SSR
Muscadet---4.4       C solved  1165 = 0.23 SPC 0.28 SOTA 1118 =  0.27 SSR
Metis---2.3          C solved  1116 = 0.22 SPC 0.27 SOTA 1069 =  0.26 SSR
Geo---2010C          C solved  1006 = 0.20 SPC 0.24 SOTA  959 =  0.23 SSR
iProver-Eq---0.85    C solved   987 = 0.20 SPC 0.24 SOTA  940 =  0.23 SSR
Alt-Ergo---0.95.1    C solved   548 = 0.11 SPC 0.13 SOTA  501 =  0.12 SSR

FOF_THM_RFO_PEQ
Unbiased SPC size               284
SOTA contributors      solved   257 = 0.90 SPC 0.90 attempted
SOTA contributors all  solved    19 = 0.07 SPC 0.07 SOTA  238 =  1.00 SSR
VanHElsing---1.0     C solved   238 = 0.84 SPC 0.93 SOTA  219 =  0.92 SSR
E---1.9              C solved   233 = 0.82 SPC 0.91 SOTA  214 =  0.90 SSR
ET---1.0             C solved   230 = 0.81 SPC 0.89 SOTA  211 =  0.89 SSR
SRASS---0.1          C solved   218 = 0.77 SPC 0.85 SOTA  199 =  0.84 SSR
Vampire---4.0        C solved   215 = 0.76 SPC 0.84 SOTA  196 =  0.82 SSR
Isabelle-HOT---2015  C solved   204 = 0.72 SPC 0.79 SOTA  185 =  0.78 SSR
SPASS---3.7          C solved   191 = 0.67 SPC 0.74 SOTA  172 =  0.72 SSR
Fampire---1.3        C solved   184 = 0.65 SPC 0.72 SOTA  165 =  0.69 SSR
CVC4---FOF-1.5pre    C solved   161 = 0.57 SPC 0.63 SOTA  142 =  0.60 SSR
Princess---140704    C solved   151 = 0.53 SPC 0.59 SOTA  132 =  0.55 SSR
iProver-Eq---0.85    C solved   150 = 0.53 SPC 0.58 SOTA  131 =  0.55 SSR
Metis---2.3          C solved   148 = 0.52 SPC 0.58 SOTA  129 =  0.54 SSR
Prover9---1109a      C solved   142 = 0.50 SPC 0.55 SOTA  123 =  0.52 SSR
Bliksem---1.12       C solved    90 = 0.32 SPC 0.35 SOTA   71 =  0.30 SSR

FOF_UNS_RFO_NEQ
Unbiased SPC size                60
SOTA contributors      solved    29 = 0.48 SPC 0.48 attempted
SOTA contributors all  solved    16 = 0.27 SPC 0.55 SOTA   13 =  1.00 SSR
Vampire---4.0        C solved    26 = 0.43 SPC 0.90 SOTA   10 =  0.77 SSR
CVC4---FOF-1.5pre    C solved    21 = 0.35 SPC 0.72 SOTA    5 =  0.38 SSR

FOF_UNS_RFO_SEQ
Unbiased SPC size                62
SOTA contributors      solved    53 = 0.85 SPC 0.85 attempted
SOTA contributors all  solved    52 = 0.84 SPC 0.98 SOTA    1 =  1.00 SSR
Vampire---4.0        C solved    53 = 0.85 SPC 1.00 SOTA    1 =  1.00 SSR

FOF_UNS_RFO_PEQ
Unbiased SPC size                74
SOTA contributors      solved    74 = 1.00 SPC 1.00 attempted
SOTA contributors all  solved    70 = 0.95 SPC 0.95 SOTA    4 =  1.00 SSR
Vampire---4.0        C solved    74 = 1.00 SPC 1.00 SOTA    4 =  1.00 SSR
```

A.3 FOF_CSA/SAT

```
FOF_CSA_EPR
Unbiased SPC size                156
SOTA contributors      solved    156 = 1.00 SPC 1.00 attempted
SOTA contributors all  solved    133 = 0.85 SPC 0.85 SOTA    23 =  1.00 SSR
Vampire---SAT-4.0    C solved    154 = 0.99 SPC 0.99 SOTA    21 =  0.91 SSR
iProver---1.4        C solved    152 = 0.97 SPC 0.97 SOTA    19 =  0.83 SSR
iProver---SAT-1.4    C solved    146 = 0.94 SPC 0.94 SOTA    13 =  0.57 SSR
E-Darwin---1.5       C solved    137 = 0.88 SPC 0.88 SOTA     4 =  0.17 SSR

FOF_CSA_RFO_NEQ
Unbiased SPC size                321
SOTA contributors      solved    309 = 0.96 SPC 0.96 attempted
SOTA contributors all  solved    191 = 0.60 SPC 0.62 SOTA   118 =  1.00 SSR
iProver---SAT-1.4    C solved    281 = 0.88 SPC 0.91 SOTA    90 =  0.76 SSR
Vampire---SAT-4.0    C solved    256 = 0.80 SPC 0.83 SOTA    65 =  0.55 SSR
Paradox---4.0        C solved    247 = 0.77 SPC 0.80 SOTA    56 =  0.47 SSR
Nitpick---2015       C solved    226 = 0.70 SPC 0.73 SOTA    35 =  0.30 SSR
CVC4---FNT-1.5pre    C solved    218 = 0.68 SPC 0.71 SOTA    27 =  0.23 SSR
Geo---2010C          C solved    211 = 0.66 SPC 0.68 SOTA    20 =  0.17 SSR

FOF_CSA_RFO_SEQ
Unbiased SPC size                253
SOTA contributors      solved    205 = 0.81 SPC 0.81 attempted
SOTA contributors all  solved    144 = 0.57 SPC 0.70 SOTA    61 =  1.00 SSR
Vampire---SAT-4.0    C solved    204 = 0.81 SPC 1.00 SOTA    60 =  0.98 SSR
Vampire---4.0        C solved    149 = 0.59 SPC 0.73 SOTA     5 =  0.08 SSR

FOF_SAT_RFO_NEQ
Unbiased SPC size                 50
SOTA contributors      solved     33 = 0.66 SPC 0.66 attempted
SOTA contributors all  solved     16 = 0.32 SPC 0.48 SOTA    17 =  1.00 SSR
Vampire---SAT-4.0    C solved     32 = 0.64 SPC 0.97 SOTA    16 =  0.94 SSR
iProver---SAT-1.4    C solved     21 = 0.42 SPC 0.64 SOTA     5 =  0.29 SSR

FOF_SAT_RFO_SEQ
SOTA contributors      solved     36 = 0.68 SPC 0.68 attempted
SOTA contributors all  solved     16 = 0.30 SPC 0.44 SOTA    20 =  1.00 SSR
Vampire---SAT-4.0    C solved     34 = 0.64 SPC 0.94 SOTA    18 =  0.90 SSR
E---1.9              C solved     30 = 0.57 SPC 0.83 SOTA    14 =  0.70 SSR
SPASS---3.7          C solved     27 = 0.51 SPC 0.75 SOTA    11 =  0.55 SSR
FIMO---0.3           C solved     24 = 0.45 SPC 0.67 SOTA     8 =  0.40 SSR
Mace4---1109a        C solved     22 = 0.42 SPC 0.61 SOTA     6 =  0.30 SSR
```

A.4 TF0_CSA

```
TF0_CSA_EQU_ARI
Unbiased SPC size                64
SOTA contributors      solved    46 = 0.72 SPC 0.72 attempted
SOTA contributors all  solved    25 = 0.39 SPC 0.54 SOTA   21 =  1.00 SSR
Z3---4.4-TPTP        C solved    43 = 0.67 SPC 0.93 SOTA   18 =  0.86 SSR
H2WO4---11.07        C solved    34 = 0.53 SPC 0.74 SOTA    9 =  0.43 SSR
CVC4---TFA-1.5pre    C solved    29 = 0.45 SPC 0.63 SOTA    4 =  0.19 SSR
```

A.5 TF1

```
TF1_THM_EQU_NAR
Unbiased SPC size               537
SOTA contributors      solved   193 = 0.36 SPC 0.36 attempted
SOTA contributors all  solved   193 = 0.36 SPC 1.00 SOTA    0 =  1.00 SSR
Alt-Ergo---0.95.1    C solved   193 = 0.36 SPC 1.00 SOTA    0 =  0.00 SSR
```

A.6 TH0_THM

```
TH0_THM_NEQ
Unbiased SPC size               548
SOTA contributors      solved   515 = 0.94 SPC 0.94 attempted
SOTA contributors all  solved   209 = 0.38 SPC 0.41 SOTA  306 =  1.00 SSR
Satallax---2.7       C solved   479 = 0.87 SPC 0.93 SOTA  270 =  0.88 SSR
Isabelle---2015      C solved   474 = 0.86 SPC 0.92 SOTA  265 =  0.87 SSR
LEO-II---1.6.2       C solved   445 = 0.81 SPC 0.86 SOTA  236 =  0.77 SSR
TPS---3.120601S1b    C solved   440 = 0.80 SPC 0.85 SOTA  231 =  0.75 SSR
agsyHOL---1.0        C solved   414 = 0.76 SPC 0.80 SOTA  205 =  0.67 SSR
cocATP---0.2.0       C solved   237 = 0.43 SPC 0.46 SOTA   28 =  0.09 SSR

TH0_THM_EQU
Unbiased SPC size              1921
SOTA contributors      solved  1707 = 0.89 SPC 0.89 attempted
SOTA contributors all  solved   399 = 0.21 SPC 0.23 SOTA 1308 =  1.00 SSR
Isabelle---2015      C solved  1537 = 0.80 SPC 0.90 SOTA 1138 =  0.87 SSR
Satallax---2.7       C solved  1509 = 0.79 SPC 0.88 SOTA 1110 =  0.85 SSR
LEO-II---1.6.2       C solved  1339 = 0.70 SPC 0.78 SOTA  940 =  0.72 SSR
agsyHOL---1.0        C solved  1317 = 0.69 SPC 0.77 SOTA  918 =  0.70 SSR
cocATP---0.2.0       C solved   487 = 0.25 SPC 0.29 SOTA   88 =  0.07 SSR
```

A.7 TH0_CSA

```
TH0_CSA_NEQ
Unbiased SPC size                69
SOTA contributors      solved    69 = 1.00 SPC 1.00 attempted
SOTA contributors all  solved    47 = 0.68 SPC 0.68 SOTA   22 =  1.00 SSR
Nitpick---2015       C solved    69 = 1.00 SPC 1.00 SOTA   22 =  1.00 SSR

TH0_CSA_EQU
Unbiased SPC size               289
SOTA contributors      solved   285 = 0.99 SPC 0.99 attempted
SOTA contributors all  solved    62 = 0.21 SPC 0.22 SOTA  223 =  1.00 SSR
Nitpick---2015       C solved   285 = 0.99 SPC 1.00 SOTA  223 =  1.00 SSR
```

A.8 CNF_UNS

```
CNF_UNS_EPR
Unbiased SPC size               649
SOTA contributors      solved   592 = 0.91 SPC 0.91 attempted
SOTA contributors all  solved   412 = 0.63 SPC 0.70 SOTA  180 =  1.00 SSR
Vampire---4.0        C solved   588 = 0.91 SPC 0.99 SOTA  176 =  0.98 SSR
iProver---1.4        C solved   556 = 0.86 SPC 0.94 SOTA  144 =  0.80 SSR
CVC4---FOF-1.5pre    C solved   469 = 0.72 SPC 0.79 SOTA   57 =  0.32 SSR
GrAnDe---1.1         C solved   438 = 0.67 SPC 0.74 SOTA   26 =  0.14 SSR

CNF_UNS_RFO_NEQ_HRN
Unbiased SPC size               543
SOTA contributors      solved   530 = 0.98 SPC 0.98 attempted
SOTA contributors all  solved   406 = 0.75 SPC 0.77 SOTA  124 =  1.00 SSR
E---1.9              C solved   525 = 0.97 SPC 0.99 SOTA  119 =  0.96 SSR
Vampire---4.0        C solved   508 = 0.94 SPC 0.96 SOTA  102 =  0.82 SSR
Darwin---1.4.5       C solved   470 = 0.87 SPC 0.89 SOTA   64 =  0.52 SSR
Prover9---1109a      C solved   457 = 0.84 SPC 0.86 SOTA   51 =  0.41 SSR

CNF_UNS_RFO_NEQ_NHN
Unbiased SPC size               481
SOTA contributors      solved   463 = 0.96 SPC 0.96 attempted
SOTA contributors all  solved   191 = 0.40 SPC 0.41 SOTA  272 =  1.00 SSR
Vampire---4.0        C solved   452 = 0.94 SPC 0.98 SOTA  261 =  0.96 SSR
E---1.9              C solved   441 = 0.92 SPC 0.95 SOTA  250 =  0.92 SSR
ET---1.0             C solved   437 = 0.91 SPC 0.94 SOTA  246 =  0.90 SSR
CVC4---FOF-1.5pre    C solved   397 = 0.83 SPC 0.86 SOTA  206 =  0.76 SSR
Geo---2010C          C solved   389 = 0.81 SPC 0.84 SOTA  198 =  0.73 SSR
Prover9---1109a      C solved   379 = 0.79 SPC 0.82 SOTA  188 =  0.69 SSR
Equinox---5.0        C solved   216 = 0.45 SPC 0.47 SOTA   25 =  0.09 SSR
```

```
CNF_UNS_RFO_SEQ_HRN
Unbiased SPC size                450
SOTA contributors      solved    429 = 0.95 SPC 0.95 attempted
SOTA contributors all  solved    212 = 0.47 SPC 0.49 SOTA  217 =  1.00 SSR
E---1.9              C solved    421 = 0.94 SPC 0.98 SOTA  209 =  0.96 SSR
Vampire---4.0        C solved    389 = 0.86 SPC 0.91 SOTA  177 =  0.82 SSR
Isabelle---2015      C solved    347 = 0.77 SPC 0.81 SOTA  135 =  0.62 SSR
Prover9---1109a      C solved    333 = 0.74 SPC 0.78 SOTA  121 =  0.56 SSR
SNARK---20120808r022 C solved    294 = 0.65 SPC 0.69 SOTA   82 =  0.38 SSR
Geo---2010C          C solved    227 = 0.50 SPC 0.53 SOTA   15 =  0.07 SSR

CNF_UNS_RFO_SEQ_NHN
Unbiased SPC size               2280
SOTA contributors      solved   1910 = 0.84 SPC 0.84 attempted
SOTA contributors all  solved    424 = 0.19 SPC 0.22 SOTA 1486 =  1.00 SSR
Vampire---4.0        C solved   1785 = 0.78 SPC 0.93 SOTA 1361 =  0.92 SSR
ET---1.0             C solved   1720 = 0.75 SPC 0.90 SOTA 1296 =  0.87 SSR
E---1.9              C solved   1709 = 0.75 SPC 0.89 SOTA 1285 =  0.86 SSR
CVC4---FOF-1.5pre    C solved   1448 = 0.64 SPC 0.76 SOTA 1024 =  0.69 SSR
Isabelle---2015      C solved   1436 = 0.63 SPC 0.75 SOTA 1012 =  0.68 SSR
Isabelle-HOT---2015  C solved   1423 = 0.62 SPC 0.75 SOTA  999 =  0.67 SSR
Prover9---1109a      C solved   1074 = 0.47 SPC 0.56 SOTA  650 =  0.44 SSR
iProver---1.4        C solved   1059 = 0.46 SPC 0.55 SOTA  635 =  0.43 SSR
DCTP---1.31          C solved   1011 = 0.44 SPC 0.53 SOTA  587 =  0.40 SSR
SNARK---20120808r022 C solved   1011 = 0.44 SPC 0.53 SOTA  587 =  0.40 SSR
Equinox---5.0        C solved    762 = 0.33 SPC 0.40 SOTA  338 =  0.23 SSR

CNF_UNS_RFO_PEQ_NUE
Unbiased SPC size                541
SOTA contributors      solved    469 = 0.87 SPC 0.87 attempted
SOTA contributors all  solved     11 = 0.02 SPC 0.02 SOTA  458 =  1.00 SSR
E---1.9              C solved    431 = 0.80 SPC 0.92 SOTA  420 =  0.92 SSR
ET---1.0             C solved    423 = 0.78 SPC 0.90 SOTA  412 =  0.90 SSR
Vampire---4.0        C solved    386 = 0.71 SPC 0.82 SOTA  375 =  0.82 SSR
Isabelle---2015      C solved    384 = 0.71 SPC 0.82 SOTA  373 =  0.81 SSR
SPASS---3.7          C solved    339 = 0.63 SPC 0.72 SOTA  328 =  0.72 SSR
Prover9---1109a      C solved    332 = 0.61 SPC 0.71 SOTA  321 =  0.70 SSR
CVC4---FOF-1.5pre    C solved    314 = 0.58 SPC 0.67 SOTA  303 =  0.66 SSR
E-Darwin---1.5       C solved    264 = 0.49 SPC 0.56 SOTA  253 =  0.55 SSR
SNARK---20120808r022 C solved    162 = 0.30 SPC 0.35 SOTA  151 =  0.33 SSR
Bliksem---1.12       C solved     63 = 0.12 SPC 0.13 SOTA   52 =  0.11 SSR
```

```
CNF_UNS_RFO_PEQ_UEQ
Unbiased SPC size               897
SOTA contributors      solved   847 = 0.94 SPC 0.94 attempted
SOTA contributors all  solved    25 = 0.03 SPC 0.03 SOTA  822 =  1.00 SSR
E---1.9              C solved   755 = 0.84 SPC 0.89 SOTA  730 =  0.89 SSR
ET---1.0             C solved   754 = 0.84 SPC 0.89 SOTA  729 =  0.89 SSR
Waldmeister---710    C solved   732 = 0.82 SPC 0.86 SOTA  707 =  0.86 SSR
Vampire---4.0        C solved   718 = 0.80 SPC 0.85 SOTA  693 =  0.84 SSR
Prover9---1109a      C solved   687 = 0.77 SPC 0.81 SOTA  662 =  0.81 SSR
Isabelle---2015      C solved   670 = 0.75 SPC 0.79 SOTA  645 =  0.78 SSR
Fiesta---2           C solved   633 = 0.71 SPC 0.75 SOTA  608 =  0.74 SSR
SPASS---3.7          C solved   600 = 0.67 SPC 0.71 SOTA  575 =  0.70 SSR
LEO-II---1.6.2       C solved   592 = 0.66 SPC 0.70 SOTA  567 =  0.69 SSR
SNARK---20120808r022 C solved   590 = 0.66 SPC 0.70 SOTA  565 =  0.69 SSR
EQP---0.9e           C solved   571 = 0.64 SPC 0.67 SOTA  546 =  0.66 SSR
Metis---2.3          C solved   543 = 0.61 SPC 0.64 SOTA  518 =  0.63 SSR
CiME---2.01          C solved   519 = 0.58 SPC 0.61 SOTA  494 =  0.60 SSR
Bliksem---1.12       C solved   439 = 0.49 SPC 0.52 SOTA  414 =  0.50 SSR
CVC4---FOF-1.5pre    C solved   422 = 0.47 SPC 0.50 SOTA  397 =  0.48 SSR
Geo---2010C          C solved   318 = 0.35 SPC 0.38 SOTA  293 =  0.36 SSR
S-SETHEO---0.0       C solved   118 = 0.13 SPC 0.14 SOTA   93 =  0.11 SSR
```

A.9 TF0_THM/UNS

```
TF0_THM_NEQ_ARI
Unbiased SPC size               282
SOTA contributors      solved   279 = 0.99 SPC 0.99 attempted
SOTA contributors all  solved   224 = 0.79 SPC 0.80 SOTA   55 =  1.00 SSR
Princess---140704    C solved   277 = 0.98 SPC 0.99 SOTA   53 =  0.96 SSR
Beagle---0.9         C solved   276 = 0.98 SPC 0.99 SOTA   52 =  0.95 SSR
CVC4---TFA-1.5pre    C solved   272 = 0.96 SPC 0.97 SOTA   48 =  0.87 SSR
Vampire---4.0          solved   270 = 0.96 SPC 0.97 SOTA   47 =  0.85 SSR
Z3---4.4-TPTP        C solved   229 = 0.81 SPC 0.82 SOTA    5 =  0.09 SSR

TF0_THM_EQU_NAR
Unbiased SPC size               127
SOTA contributors      solved    73 = 0.57 SPC 0.57 attempted
SOTA contributors all  solved    57 = 0.45 SPC 0.78 SOTA   16 =  1.00 SSR
Vampire---4.0        C solved    70 = 0.55 SPC 0.96 SOTA   13 =  0.81 SSR
CVC4---TFA-1.5pre    C solved    60 = 0.47 SPC 0.82 SOTA    3 =  0.19 SSR
```

```
TF0_THM_EQU_ARI
Unbiased SPC size               667
SOTA contributors      solved   616 = 0.92 SPC 0.92 attempted
SOTA contributors all  solved   135 = 0.20 SPC 0.22 SOTA  481 =  1.00 SSR
CVC4---TFA-1.5pre    C solved   557 = 0.84 SPC 0.90 SOTA  422 =  0.88 SSR
Vampire---4.0        C solved   556 = 0.83 SPC 0.90 SOTA  421 =  0.88 SSR
Princess---140704    C solved   533 = 0.80 SPC 0.87 SOTA  398 =  0.83 SSR
Beagle---0.9         C solved   487 = 0.73 SPC 0.79 SOTA  352 =  0.73 SSR
SPASS+T---2.2.22     C solved   469 = 0.70 SPC 0.76 SOTA  334 =  0.69 SSR
SNARK---20120808r022 C solved   348 = 0.52 SPC 0.56 SOTA  213 =  0.44 SSR
Zipperpin---TFF-0.4  C solved   274 = 0.41 SPC 0.44 SOTA  139 =  0.29 SSR

TF0_UNS_EQU_NAR
Unbiased SPC size                20
SOTA contributors      solved    20 = 1.00 SPC 1.00 attempted
SOTA contributors all  solved    10 = 0.50 SPC 0.50 SOTA   10 =  1.00 SSR
CVC4---TFA-1.5pre    C solved    20 = 1.00 SPC 1.00 SOTA   10 =  1.00 SSR
Vampire---4.0        R solved    20 = 1.00 SPC 1.00 SOTA   10 =  1.00 SSR

TF0_UNS_EQU_ARI
Unbiased SPC size                20
SOTA contributors      solved    12 = 0.60 SPC 0.60 attempted
SOTA contributors all  solved    10 = 0.50 SPC 0.83 SOTA    2 =  1.00 SSR
Vampire---4.0        C solved    12 = 0.60 SPC 1.00 SOTA    2 =  1.00 SSR
```