The IJCAR ATP System Competition: All the Details

Geoff Sutcliffe
Department of Computer Science
University of Miami, USA
Email: geoff@cs.miami.edu

Christian Suttner
Antfactory
München, Germany
Email: csuttner@antfactory.com

Jeff Pelletier
Department of Computing Science
University of Alberta, Canada
Email: jeffp@cs.ualberta.ca

UM-CSC-2001-001

Abstract

The IJCAR ATP System Competition (CASC-JC) was held on 21st June 2001, as part of the International Joint Conference on Automated Reasoning (IJCAR), in Siena, Italy. CASC-JC evaluated the performance of fully automatic, first-order ATP systems. The evaluation was in terms of the number of problems solved, the number of proofs output, and the average runtime for problems solved. The evaluation was done in the context of a bounded number of eligible problems chosen from the TPTP Problem Library, and a specified time limit for each solution attempt. CASC-JC was the sixth such ATP system competition, following the successful competitions at CADEs-13 to -17. This report presents the results of CASC-JC.

1 Introduction

The IJCAR ATP System Competition (CASC-JC) was held on 21st June 2001, as part of the International Joint Conference on Automated Reasoning (IJCAR), in Siena, Italy. CASC-JC evaluated the performance of fully automatic, first-order ATP systems. The evaluation was in terms of the number of problems solved, the number of proofs output, and the average runtime for problems solved. The evaluation was done in the context of a bounded number of eligible problems chosen from the TPTP Problem Library [SS98c], and a specified time limit for each solution attempt. CASC-JC was the sixth such ATP system competition, following the successful competitions at CADEs-13 to -17 [SS97a, SS98d, SS99, Sut00b, Sut01b]. This paper presents the results of CASC-JC.[1]

Twenty three ATP systems, listed in Table 1, competed in the various competition and demonstration divisions. The winners of the CASC-17 divisions were automatically entered into those divisions, to provide benchmarks against which progress can be judged. System descriptions for the entered systems are in [Sut01a] and are also available from the WWW site given below. Short descriptions of the division winners are given in Section 5.

The competition was organized by Geoff Sutcliffe and Christian Suttner. The competition was overseen by a panel consisting of Maria Paola Bonacina, Claude Kirchner, Jeff Pelletier, and Toby Walsh. The competition was run on SUN Ultras supplied by Technische Universität München. The CASC-JC WWW site provides access to information and data used before, during, and after the event:

`http://www.cs.miami.edu/~tptp/CASC/JC/`

1.1 Design Changes

The design and procedures of CASC-JC evolved from those of CASCs-13 to -17 [SS97c, SS97b, SS98a, SS98b, Sut99, Sut00a]. The CASC-JC design improved on the previous competitions' designs in several ways:

- The SEM (semantic) division, introduced in CASC-17, was motivated by the idea that entrants would be able to demonstate the extent to which their systems could be tuned to an application domain. However, despite strong expressions of interest at CASC-16, at CASC-17 no systems were especially tuned to the chosen application domain of set theory. Due to this lack of interest, and overlap of the SEM division with the syntactically defined divisions, the SEM division was not continued in CASC-JC.

[1] CASC is an acronym for the "CADE ATP System Competition", and the CASC numbering -13 to -17 corresponds to those CADEs. In 2001 CADE was part of IJCAR, hence the break in the numbering sequence, and instead "JC" for "Joint Conference".

Table 1: The ATP systems and entrants

ATP System	Entrants
Bliksem 1.12	Hans de Nivelle *Max-Planck-Institut für Informatik, Germany*
DCTP 0.1	Gernot Stenz, Reinhold Letz *Technische Universität München, Germany*
E 0.62	Stephan Schulz *Technische Universität München, Germany*
EP 0.62	A composition of E 0.62 and a proof presentation tool, for the MIX division Proof class
E 0.6	Entered as winner of the CASC-17 MIX division
E-SETHEO csp01	Gernot Stenz, Reinhold Letz, Stephan Schulz *Technische Universität München, Germany*
Gandalf c-2.3	Tanel Tammet *Tallinn Technical University, Estonia* and *Safelogic, Sweden*
GandalfFOF c-2.3	A variant of Gandalf c-2.3 using Otter for conversion to CNF, for the FOF division
GandalfSat 1.1	A variant of Gandalf, for the SAT division
GandalfSat 1.0	Entered as winner of the CASC-17 SAT division
MACE 2.0	William McCune, Larry Wos, Bob Veroff *Argonne National Laboratory, USA*
MUSCADET 2.3	Dominique Pastre *Université René Descartes, France*
Otter 3.2	William McCune, Larry Wos, Bob Veroff *Argonne National Laboratory, USA*
Otter-MACE 3.2-2.0	A composition of Otter 3.2 and MACE 2.0, for the EPR division
PizEAndSATO 0.2	Geoff Sutcliffe, Stephan Schulz *University of Miami, USA* and *Technische Universität München, Germany*
SCOTT 6.0.0	John Slaney, Kal Hodgson *Australian National University, Australia*
Vampire 2.0	Alexandre Riazanov, Andrei Voronkov *University of Manchester, England*
VampireEPR 2.0	A variant of Vampire 2.0, for the EPR division
VampireFOF 2.0	A variant of Vampire 2.0, for the FOF division
VampireJC 2.0	A variant of Vampire 2.0, for the MIX division
VampireFOF 1.0	Entered as winner of the CASC-17 FOF division
Waldmeister 601	Thomas Hillenbrand, Bernd Loechner, Andreas Jaeger, Arnim Buch *Max-Planck-Institut für Informatik* and *Universität Kaiserslautern, Germany*
Waldmeister 600	Entered as winner of the CASC-17 UEQ division

- There are many CNF problems, often from "real world" applications, that have a finite Herbrand universe. These problems are effectively propositional, and can be solved using techniques quite different to those used for problems with an infinite Herbrand universe. Prior to CASC-17 these problems were eligible in the MIX division, but because of their distinct nature they were excluded in CASC-17. However, these problems are of interest, and therefore a new EPR (effectively propositional) division, for these problems, was added to the CASC-JC competition divisions.
- There is evidence that for some applications of ATP there is a need for proof or model output. As a first step towards evaluating systems' abilities to produce such output, and to stimulate research in this direction, the MIX division was ranked in two classes. The first class ranked the systems according to the number of problems solved, and the second class ranked the systems according to the number of problems solved with a proof output.
- Analysis of ATP system performances on problems in the SAT division has shown that there is system specialization between problems with equality and problems without equality [FS00]. It is appropriate to separate the evaluation of systems on these two types of problems, so that the specialist capabilities can be observed. Therefore the SAT division was divided into two problem categories, one with equality and one without.
- There has been concern in the ATP community about the extent to which the systems entered into CASC have been tuned to the TPTP problems that have been likely to be eligible for use. Such tuning may improve only the systems' abilities to solve TPTP problems, and not produce generally applicable advances (it may even degrade the general capability of a system). In order to make such overtuning undesirable, the CASC-JC problems were taken from an unreleased version of the TPTP, so that the systems could not be tuned for the new problems in that TPTP version. Overtuning for the old problems in the TPTP was hence potentially disadvantageous, because it could degrade performance on the new problems, with a consequent degradation in overall performance.
- When a problem is added to the TPTP, it is labelled as incomplete or augmented if the clauses were designed to make the problem solvable using an ATP system. Up to CASC-17, incomplete and augmented problems were excluded from CASC, as there was a perceived danger that the problems might be biased towards a particular ATP system. It has since been concluded that such modifications are generally effective for all ATP systems. Therefore incomplete and augmented problems were made eligible in CASC-JC.

Further details and motivations for these changes are given in [Sut01a].

2 Divisions

CASC-JC was divided into divisions according to problem and system characteristics. There were five competition divisions in which the systems were explicitly ranked, and one demonstration division in which systems could demonstrate their abilities without being formally ranked. Entry into the competition divisions was subject to the following rules:

- ATP systems could be entered at only the division level.
- ATP systems could be entered into more than one division. A system that was not entered into a division is assumed to perform worse than the entered systems, for that type of problem.
- The ATP systems had to run on a single locally provided standard UNIX workstation (the *general hardware* - see Section 3). ATP systems that could not run on the general hardware could be entered into the Demonstration division.

2.1 Competition Divisions

The **MIX** division: Mixed CNF really-non-propositional theorems.
Mixed means Horn and non-Horn problems, with or without equality, but not unit equality problems (see the UEQ division below). *Really-non-propositional* means with an infinite Herbrand universe (so that the problems cannot be solved by finite saturation methods). The MIX division had five problem categories:

- The **HNE** category: Horn with No Equality
- The **HEQ** category: Horn with some (not pure) Equality

- The **NNE** category: Non-Horn with No Equality
- The **NEQ** category: Non-Horn with some (not pure) Equality
- The **PEQ** category: Pure Equality

The MIX division had two ranking classes:

- The **Assurance** class: Ranked according to the number of problems solved (a "yes" output, giving an *assurance* of the existence of a proof).
- The **Proof** class: Ranked according to the number of problems solved with an acceptable proof output. The competition panel judged whether or not each system's proof format is *acceptable*.

Eleven systems competed in the MIX division. They were Bliksem, DCTP, E 0.6 (the CASC-17 winner), E 0.62, EP, E-SETHEO, Gandalf, Otter, SCOTT, Vampire, and VampireJC. All systems that competed in the MIX division were ranked in the Assurance class. Systems that output acceptable proofs were also ranked in the Proof class. They were Bliksem, EP, Otter, SCOTT, Vampire, and VampireJC.

The **UEQ** division: Unit equality CNF really-non-propositional theorems.
Unit equality means that each clause consists of a single equality literal.

Eight systems competed in the UEQ division. They were Bliksem, E 0.62, Gandalf, Otter, SCOTT, Vampire, Waldmeister 600 (the CASC-17 winner), and Waldmeister 601.

The **SAT** division: CNF really-non-propositional non-theorems.
The SAT Division had two problem categories:

- The **SNE** category: SAT with No Equality
- The **SEQ** category: SAT with Equality

Six systems competed in the SAT division. They were DCTP, E-SETHEO, GandalfSat 1.0 (the CASC-17 winner), GandalfSat 1.1, MACE, and SCOTT.

The **FOF** division: FOF non-propositional theorems.
FOF means "natural" First Order Form, including quantifiers. The FOF division had two problem categories:

- The **FNE** category: FOF with No Equality
- The **FEQ** category: FOF with Equality

Eight systems competed in the FOF division. They were Bliksem, E-SETHEO, GandalfFOF, MUSCADET, Otter, SCOTT, VampireFOF 1.0 (the CASC-17 winner), and VampireFOF 2.0.

The **EPR** division: CNF effectively propositional theorems and non-theorems.
Effectively propositional means syntactically non-propositional but with a finite Herbrand universe (and hence semantically propositional). The EPR Division had two problem categories:

- The **EPT** category: Effectively Propositional Theorems (unsatisfiable clauses)
- The **EPS** category: Effectively Propositional non-theorems (Satisfiable clauses)

Seven systems competed in the EPR division. They were DCTP, E 0.62, E-SETHEO, Gandalf, Otter-MACE, SCOTT, and VampireEPR. As the EPR division was new in CASC-JC, there was no CASC-17 winner to be entered.

2.2 Demonstration Division

ATP systems that could not run on the general hardware, or could not be entered into the competition divisions for any other reason, could be entered into the Demonstration division. Demonstration division systems could run on the general hardware, or the hardware could be supplied by the entrant. The entry specified which competition divisions' problems were to be used. The results are presented along with the competition divisions' results, but may not be comparable with those results. Only one system was entered into the Demonstration division. It was PizEAndSATO, which used the EPR division problems. PizEAndSATO was entered into the Demonstration division because one of the entrants was also a competition organizer.

3 Organization

For CASC-JC, the *general hardware* was 25 SUN UltraSparc IIi workstations, each having a 440 MHz UltraSparc II CPU, 256MB memory, and the SunOS 5.8 operating system. The machines were connected in a network with no other users having access to the machines during the competition.

The problems were taken from the TPTP Problem Library, v2.4.0. TPTP v2.4.0 was not released until after the competiton, so that the systems could not be tuned for the new problems in TPTP v2.4.0. Unbiased TPTP problems with a TPTP difficulty rating in the range 0.21 to 0.99 were eligible for selection in all divisions. In addition, in order to make sufficient problems eligible, in the UEQ division problems with difficulty 1.00 (i.e., not yet solved by any system in normal testing) were also eligible, and in the EPR division problems with difficulty down to 0.16 were also eligible. The problems used were randomly selected from the eligible problems at the start of the competition, based on a seed supplied by the competition panel. A limiting procedure [Sut00b] was used to prevent the selection of an excessive number of very similar problems for any division or category. The selection mechanism was biased to select new problems until 50% of the problems in each category had been selected, after which random selection (from old and new problems) continued. The actual percentage of new problems used was dependent on how many new problems were eligible and the limitation on very similar problems. Table 2 gives the numbers of eligible problems, the numbers of new eligible problems, the maximal numbers that could be used after taking into account the limitation on very similar problems, the numbers of problems used, and the numbers of new problems used, in each division and category. Due to the small maximal number of usable problems in the EPS category, the limitation on the number of very similar problems could not be imposed. To ensure that no system received an advantage or disadvantage due to the specific presentation of the problems in the TPTP, the `tptp2X` utility was used to replace all predicate and function symbols with new symbols, randomly reorder the formulae and clauses' literals, and randomly reverse the unit equalities in the UEQ problems.

Table 2: Numbers of eligible and used problems

Division			MIX			UEQ
Category	HNE	HEQ	NNE	NEQ	PEQ	
Eligible	95	72	47	565	64	114
New eligible	2	4	3	263	0	1
Max usable	36	69	27	565	64	114
Used	20	30	20	30	20	90
New used	2	4	3	10	0	1
Division	SAT		FOF		EPR	
Category	SNE	SEQ	FNE	FEQ	EPT	EPS
Eligible	80	145	151	463	25	78
New eligible	59	105	1	246	0	5
Max usable	55	97	151	463	25	8
Used	40	50	40	50	25	25
New used	25	31	1	16	0	5

The ATP systems were required to be sound and fully automatic. The organizers tested for soundness by submitting non-theorems to the systems participating in the MIX, UEQ, FOF, and EPR divisions, and theorems to the systems participating in the SAT and EPR divisions. Claiming to have found a proof of a non-theorem or a disproof of a theorem indicates unsoundness. One system failed this test and was repaired. Fully automatic operation meant that any command line switches had to be the same for all problems. With the exception of the MIX division Proof class, the ATP systems were not required to output solutions (proofs or models), but systems that did output solutions are highlighted in the presentation of results. A 300 second CPU time limit was imposed on each solution attempt.

4 Results

For each ATP system, for each problem attempted, three items of data were recorded: whether or not a solution was found, the CPU time taken, and whether or not a solution (proof or model) was output. In the MIX division Proof class the systems were ranked according to the number of problems solved with a proof output. In the MIX division Assurance class, and all other divisions, the systems were ranked according to the numbers of problems solved. If there was a tie according to these rankings, then the tied systems were ranked according to their average CPU times over problems solved. This section presents the recorded data, and provides some analysis of the results.

It is important to be aware that the results obtained are modulo the competition design. There are many parameters that affect the results, and therefore the results do not necessarily determine the best overall ATP systems. Rather the results should be viewed as providing interesting insights into the ATP systems.

4.1 The MIX Division

Tables 3 and 4 summarize the results in the MIX division and categories. The tables show the numbers of proofs found, the average CPU times over problems solved, and whether or not proofs were output.

E-SETHEO and VampireJC both solved the most problems, with very close average CPU times. Due to the very close performances of these two systems, the competition panel declared a tie between these two systems in the Assurance class. E-SETHEO does not output proofs, therefore VampireJC is the winner of the Proof class.[2] E 0.62 solved the third most problems, slightly ahead of the CASC-17 winner E 0.6. It thus seems that substantial progress has been made in the MIX division since CASC-17. Six of the eleven systems output proofs.

The rankings in the HNE, NNE, and PEQ categories align quite closely with the division ranking. The NEQ category ranking aligns least with the division ranking, with Gandalf and Vampire performing much better in this category than in the division as a whole. The HEQ category is the only category in which E-SETHEO and VampireJC are not the two top systems, and the NEQ category is the only category in which VampireJC outperforms E-SETHEO. These variations in performance indicate some specialization of the systems.

Table 5 shows, for each system, the fractions of old problems (in the TPTP before the competition, and thus available for system tuning), new problems, and all problems solved in the MIX division. The new problems have only a slightly lower average difficulty rating than the old ones, and all the systems except VampireJC solved a higher fraction of new problems than old problems and all problems. VampireJC (and Vampire) both had an internal limit of 31 literals per clause (a legacy from early versions of Vampire), hence preventing them from attempting four of the new `NLP` problems. This strong performance on the new problems counters the concern that systems have been overtuned to TPTP problems: if the systems were tuned using TPTP problems, then that tuning also worked for the new problems, and therefore seems likely to be effective in general.

Table 3: MIX division results

ATP System	MIX /120	Average time	Proof output?
E-SETHEO	93	38.7	no
VampireJC	93	43.4	yes
E 0.62	84	36.6	no
E 0.6	81	34.7	no
Vampire	76	40.5	yes
EP	73	36.2	yes
Gandalf	61	56.2	yes
Otter	31	34.2	no
SCOTT	30	77.7	yes
Bliksem	29	66.9	yes
DCTP	14	18.8	no

4.1.1 Times Taken for Each Problem

Tables 6 to 10 show the CPU times taken by the systems, for each problem in each category. The systems are listed in order of number of problems solved in the category, with the system that solved the most problems leftmost. The problems are listed using their TPTP names in alphabetical order (not the order that was used in the competition). A * after a problem name indicates that it was new in TPTP v2.4.0. A "T" result indicates that the system timed out, an "N" indicates that the system abandoned the proof attempt before the time limit was reached, and a "U" indicates that the system terminated before the time limit for some unknown reason.

No problems were solved by all the systems. Eleven problems, spread across the HEQ, NNE, and NEQ categories, were unsolved. Of those, the problems in the HEQ category have all been solved by E in normal testing. E's failure

[2] Some systems, including VampireJC and Bliksem, may have solved some problems within the CPU time limit but exceeded the CPU time limit while building a proof. As the CPU time limit was exceeded, these solutions were not counted towards the totals for the rankings. Variants of these systems that do not attempt to build a proof might have performed better in the Assurance class, as, e.g., E did, relative to EP. In future CASCs it is planned to record separately the times when the problem has been solved and when the proof has been output.

Table 4: MIX category results

ATP System	HNE /20	Avg time	HEQ /30	Avg time	NNE /20	Avg time	NEQ /30	Avg time	PEQ /20	Avg time
E-SETHEO	18	47.4	19	38.9	15	38.0	22	22.8	19	49.1
VampireJC	17	56.6	19	49.2	13	26.7	25	33.3	19	50.2
E 0.62	16	29.9	20	40.2	11	69.5	19	24.9	18	30.7
E 0.6	15	38.7	20	47.1	10	47.6	18	19.4	18	25.8
Vampire 2.0	15	56.0	13	57.5	10	33.2	21	29.3	17	32.0
EP	15	46.4	18	51.9	8	62.3	17	20.5	15	11.2
Gandalf	15	92.0	9	15.3	10	80.4	22	42.7	5	33.0
Otter	2	14.5	7	12.1	6	24.1	8	64.7	8	35.5
SCOTT	2	217.3	8	114.1	7	46.2	6	43.3	7	57.4
Bliksem	5	143.9	1	224.4	4	1.4	6	51.6	13	52.3
DCTP	2	1.8	0	-	5	4.7	7	33.7	0	-

Table 5: Fractions of old, new, and all problems solved in the MIX division

	Old	New	All
Number	101	19	120
Av. rating	0.50	0.46	0.49
E-SETHEO	0.77	0.79	0.78
VampireJC	0.79	0.68	0.78
E 0.62	0.69	0.74	0.70
E 0.6	0.67	0.68	0.68
Vampire	0.62	0.68	0.63
EP	0.60	0.63	0.61
Gandalf	0.47	0.74	0.51
Otter	0.21	0.53	0.26
SCOTT	0.22	0.42	0.25
Bliksem	0.22	0.37	0.24
DCTP	0.07	0.37	0.12

to solve them in CASC-JC confirms a previous observation [Sut01b] that E is sensitive to the reordering done to the problems for CASC. Two of the three unsolved problems in the NNE category have been solved by Vampire in normal testing, and again the reordering appears to be the cause of failure in the competition. The remaining five problems have all been solved by systems that were not entered into CASC-JC.

Although the overall performances of E-SETHEO and VampireJC were very close, they did not solve very similar sets of problems. On the other hand, as might be expected, E 0.62 and E 0.6 solved almost the same problems, with E 0.62 subsuming (solving a superset of the problems solved by) E 0.6 and EP. E-SETHEO subsumed EP and DCTP, and VampireJC subsumed DCTP. No other systems subsumed another system, indicating that most of the systems have some unique abilities, and therefore contribute to the "state-of-the-art" [SS01].

It is pleasant to note that the lowly ranked DCTP, which is a new system based on the disconnection calculus [LS01], solved some problems much faster than the more established resolution and superposition based systems, e.g., DCTP solved the NEQ problem `SET019-4` in 0.4 seconds, while VampireJC, the top NEQ system, took 60.6 seconds.

Table 6: CPU times in seconds for the HNE category

Problem	VampireJC		E 0.6		Vampire		Bliksem		Otter	
	E-S'O	E 0.62		EP		Gandalf		DCTP		SCOTT
LCL002-1	T	90.8	T	T	T	280.5	160.8	T	T	T	261.6
LCL070-1	35.2	13.7	127.4	33.8	264.5	55.0	4.6	T	T	22.7	T
LCL105-1	15.7	T	0.8	14.8	2.6	T	N	T	T	T	T
LCL221-1	106.0	T	35.2	34.3	75.1	T	U	T	T	N	T
LCL224-1	41.0	0.3	40.1	128.0	82.1	0.9	128.8	290.5	T	N	T
LCL225-1	40.3	3.2	39.4	93.5	76.9	3.2	128.9	284.2	T	N	T
LCL231-1	2.0	33.3	1.3	1.4	3.3	41.9	128.6	136.5	T	N	T
LCL253-1	264.4	187.9	T	T	T	T	U	T	T	N	T
LCL391-1	151.4	204.9	136.7	163.7	T	T	N	T	T	T	T
LCL394-1	84.1	T	82.9	106.7	159.7	T	N	T	T	T	T
NUM017-1	12.5	58.6	9.9	T	22.4	88.8	178.0	4.2	T	T	173.0
PLA008-1	0.9	0.9	0.6	0.4	1.2	0.8	116.8	T	T	T	T
PLA009-1	1.3	0.2	0.6	0.7	1.3	0.2	41.2	T	T	T	T
PLA010-1	3.4	0.6	3.5	2.8	4.9	0.6	113.2	T	T	T	T
PLA012-1	1.0	0.5	0.2	0.3	0.8	0.4	48.0	T	T	T	T
PLA014-1	0.6	0.2	0.0	0.0	0.3	0.2	44.8	T	3.5	T	T
PUZ042-1*	T	221.9	N	N	U	188.1	0.2	4.0	T	6.3	N
RNG001-2	92.4	144.5	T	T	T	178.8	139.9	T	T	T	T
SWV014-1*	0.6	0.1	0.1	0.0	0.5	0.1	101.1	T	T	T	T
SYN311-1	0.5	0.1	0.0	0.0	0.3	0.1	45.4	T	0.0	N	T

4.1.2 Proofs as a Function of Time

Figure 1 plots the CPU times taken by each system for proofs found, in increasing order of CPU time taken. The highest Solution number coordinate of each plot gives the number of proofs found by that system.

The plots divide the systems into five groups. The first group contains E-SETHEO and Vampire JC, which outperform the other systems. The second group contains the two versions of E, the third group contains Vampire and EP, Gandalf is alone in the fourth group, and the remaining systems make up the final group. All the systems except Otter and DCTP continue to solve problems right up to the CPU time limit, suggesting that they may be able to solve more problems if given a higher time limit. This is in contrast to previous CASCs, where almost all systems appeared to have reached their performance limits within the CPU time allowed, and suggests that a higher time limit may be appropriate for the MIX division in future competitions.

A line drawn parallel to the X-axis in Figure 1 shows how many problems would be solved by each system within a time limit given by the Y-axis value of the line. The system ranking is stable for time limits from about 200 seconds, so the final competition ranking appears to be reasonably time limit independent. The flattening off of E-SETHEO's plot, compared to the steepening of VampireJC's plot, suggests that with a higher time limit E-SETHEO may solve more problems than VampireJC.

4.2 The UEQ Division

Table 11 summarizes the results in the UEQ division. As was the case in CASCs-14 to -17, Waldmeister is the winner. Waldmeister 600 solved the same number of problems (and, as is shown in Tables 12 and 13, the same problems as) the new Waldmeister 601, but with a slightly higher average CPU time. It seems that not much progress has been made in the UEQ division. Six of the eight systems output proofs.

The problems had an average difficulty rating of 0.54.

Table 7: CPU times in seconds for the HEQ category

Problem	E 0.62	E 0.6	VampireJC E-S'O		Vampire EP		Gandalf	SCOTT	Otter	Bliksem DCTP	
ANA004-1	T	T	T	T	T	T	N	T	T	T	T
BOO008-1	20.7	19.5	26.7	2.1	40.2	2.3	N	212.1	14.9	T	T
BOO014-1	25.2	19.2	26.7	3.9	68.0	3.8	N	184.5	28.0	T	T
BOO014-3	17.1	38.7	47.6	5.1	43.1	5.2	N	T	26.1	T	T
COL003-7	69.6	50.6	2.4	29.9	152.3	29.8	24.6	T	N	T	T
COL006-3	T	T	T	19.0	T	T	N	T	T	T	T
COL043-2	2.0	1.8	8.9	0.6	5.9	0.6	29.2	T	T	T	T
COL044-2	1.1	1.0	2.9	1.3	3.6	1.3	25.2	T	T	T	T
COL044-3	1.8	2.2	3.5	1.5	5.5	1.5	25.3	T	T	T	T
COL044-5	4.3	59.5	111.6	1.4	12.5	1.4	25.3	T	T	T	T
HEN011-3	12.1	12.9	18.9	1.8	41.4	1.8	N	45.2	T	T	T
HWV002-1	T	T	T	163.7	T	T	U	T	T	224.4	T
LAT005-3	T	T	T	T	T	T	N	T	T	T	T
LCL109-4	65.7	61.3	72.5	259.0	271.3	T	N	131.7	T	T	U
LCL152-1	T	T	T	25.0	T	T	U	T	T	T	U
LCL230-3	T	T	T	1.4	T	139.9	U	T	T	T	T
LCL243-3	40.1	30.2	47.5	T	88.2	T	U	T	T	T	T
LCL252-3	T	T	T	T	T	T	N	T	T	T	T
LCL299-3	T	T	T	196.3	T	T	N	T	T	T	T
LCL303-3	T	T	T	T	T	286.8	N	T	T	T	T
LCL306-3	26.5	31.5	32.4	4.2	59.6	257.6	U	T	T	T	T
LCL312-3	5.5	3.7	12.4	23.7	14.0	T	N	T	T	T	T
LCL326-3	245.6	169.6	246.5	12.0	T	15.1	U	T	T	T	T
LCL330-3	54.6	239.7	64.2	183.2	121.9	T	U	T	T	T	T
LCL349-3	T	T	T	T	T	T	U	T	T	T	T
NLP251-1*	1.0	1.1	5.6	N	1.7	T	2.6	140.1	4.1	T	T
NLP252-1*	0.8	0.8	1.5	T	1.5	T	1.7	31.9	4.2	T	T
NLP257-1*	0.6	0.8	3.8	N	1.2	T	2.0	107.0	3.9	T	T
NLP258-1*	1.0	1.1	3.2	T	1.8	T	1.5	60.2	3.8	T	T
ROB015-2	208.9	196.4	T	T	T	T	T	T	T	T	T

Table 8: CPU times in seconds for the NNE category

Problem	VampireJC E-S'O	Vampire E 0.62		Gandalf E 0.6		SCOTT EP		DCTP Otter	Bliksem	
ANA002-2	276.8	T	T	T	T	45.1	T	T	T	T	T
ANA002-3	37.6	T	T	T	T	66.9	T	T	T	T	T
ANA002-4	3.4	T	T	T	T	60.6	T	T	T	T	T
FLD011-3	T	74.2	T	161.0	T	273.4	T	T	T	U	T
FLD014-3	2.6	25.8	113.0	25.6	101.7	1.5	T	39.6	117.6	23.2	T
FLD025-1	0.7	19.7	251.5	19.7	N	N	T	T	T	0.1	T
FLD038-1	T	T	T	T	T	N	T	T	T	T	T
FLD042-3	T	49.9	T	50.0	T	0.6	T	40.6	5.8	T	T
FLD050-2	T	T	T	T	T	N	T	T	T	T	T
FLD061-1	0.6	1.3	181.1	1.3	171.8	157.8	T	T	T	T	T
FLD061-4	6.4	41.0	87.9	40.8	79.0	88.1	188.8	143.2	T	T	T
FLD066-3	14.7	32.5	T	32.5	T	4.8	T	34.5	21.3	T	T
FLD072-4	T	T	T	T	T	N	T	T	T	U	T
NLP079-1*	0.6	0.2	0.2	0.2	0.2	U	0.7	23.3	0.0	0.1	1.7
NLP080-1*	0.8	0.1	0.2	0.2	0.2	U	0.9	21.2	0.0	0.2	2.2
NLP081-1*	0.6	0.2	0.2	0.2	0.2	U	0.8	21.0	0.0	0.1	1.4
SYN036-1	27.3	T	2.7	T	9.9	104.9	6.9	T	T	T	0.2
SYN067-1	98.8	34.2	75.4	N	61.7	U	163.3	T	T	T	T
SYN067-2	61.1	1.2	37.9	T	37.0	U	90.5	T	T	T	T
SYN067-3	37.8	66.5	14.7	T	13.9	U	46.6	T	T	T	T

Table 9: CPU times in seconds for the NEQ category

Problem	VampireJC	E-S'O	Gandalf	Vampire	E 0.62	E 0.6	EP	Otter	DCTP	SCOTT	Bliksem
CAT002-3	34.7	1.4	0.4	1.3	17.1	17.8	55.0	175.1	T	32.1	T
CAT004-3	39.2	1.1	0.7	T	T	T	T	T	T	61.7	T
GEO006-2	1.5	28.9	86.8	1.4	0.1	0.1	0.7	T	T	T	T
GEO007-1	0.4	33.4	11.1	16.9	3.9	4.4	10.8	T	T	82.5	T
GEO009-3	18.7	57.5	3.9	18.5	24.1	167.0	58.3	176.6	T	28.9	T
GEO028-2	T	T	N	T	T	T	T	T	U	T	T
GEO044-2	T	T	N	248.3	T	T	T	T	T	T	T
GEO049-2	165.0	102.0	22.1	164.0	75.7	70.6	131.2	T	T	T	T
GEO111-1*	30.0	2.0	34.4	29.9	T	T	T	T	U	T	123.5
GRP008-1	0.2	0.9	0.7	0.2	0.0	0.0	0.3	0.9	5.3	T	T
MGT025-1*	0.1	0.9	0.2	0.2	0.1	0.1	0.5	T	86.2	N	3.8
MGT039-1*	0.3	0.9	10.5	0.3	0.0	0.2	0.4	1.1	T	N	T
MGT042-1*	40.4	T	30.6	40.2	T	T	T	T	U	4.7	T
MGT060-1*	54.9	2.4	34.2	55.1	157.9	T	T	0.1	0.4	T	T
MSC007-2.005	101.3	45.7	U	T	17.1	3.2	58.9	T	T	N	T
SET015-3	T	T	N	T	T	T	T	T	T	T	T
SET016-1	0.1	0.9	64.9	0.2	0.0	0.0	0.3	1.7	U	T	T
SET019-4	60.6	1.9	26.2	T	T	T	T	3.5	0.4	T	T
SET025-9	1.0	1.8	35.0	0.9	0.2	0.2	0.7	T	U	T	37.4
SET095-6	0.9	30.4	76.7	0.9	0.1	0.2	0.5	T	U	T	T
SET097-7	120.1	T	N	T	T	T	T	T	U	T	T
SET217-6	127.3	T	N	T	T	T	T	T	U	T	T
SET236-6	29.7	1.3	63.4	29.4	14.6	10.5	26.5	T	U	T	37.4
SET261-6	0.5	24.0	123.9	0.5	0.4	0.4	1.2	158.8	T	T	17.0
SWC054-1*	2.8	113.6	136.5	2.7	160.7	72.4	T	T	37.1	T	T
SWC118-1*	0.5	25.0	60.5	0.6	1.0	1.2	2.4	T	105.7	T	T
SWC190-1*	T	T	N	T	T	T	T	T	U	T	T
SWC227-1*	T	T	N	T	T	T	T	T	T	T	T
SWC277-1*	2.7	24.1	106.4	2.7	0.2	0.2	0.7	T	U	T	90.7
SYN013-1	0.1	0.9	10.0	0.1	0.2	0.1	0.8	T	0.5	50.0	T

Table 10: CPU times in seconds for the PEQ category

Problem	VampireJC	E-S'O	E 0.6	E 0.62	Vampire	EP	Bliksem	Otter	SCOTT	Gandalf	DCTP
BOO020-1	159.2	15.3	152.8	0.7	15.2	3.7	T	T	T	N	U
COL074-3	0.5	0.1	0.0	0.0	0.2	0.3	T	132.0	T	75.1	T
COL076-1	T	233.4	T	T	T	T	T	T	76.6	N	T
COL079-2	270.7	126.5	T	T	T	T	T	T	26.0	28.3	T
COL080-2	1.3	22.7	1.6	1.7	22.8	4.9	T	T	56.7	29.1	T
COL081-2	13.0	0.3	0.1	0.1	0.4	0.5	65.0	106.1	54.8	29.1	T
GRP051-1	128.7	74.0	43.2	114.0	4.3	T	12.6	N	U	N	T
GRP052-1	80.8	T	38.0	156.3	11.7	T	3.2	N	16.5	N	T
GRP054-1	4.9	76.0	5.1	13.2	T	28.2	U	T	N	N	T
GRP056-1	13.7	92.6	2.7	9.0	216.0	25.6	8.5	N	N	N	T
GRP061-1	8.8	80.1	4.4	12.7	18.4	35.5	3.2	N	12.2	N	T
GRP064-1	0.5	0.3	0.0	0.0	0.2	0.7	2.3	5.1	T	N	T
GRP065-1	0.7	0.2	0.2	0.2	0.2	0.3	1.4	0.4	N	T	N
GRP069-1	0.5	0.1	0.0	0.0	0.1	0.5	0.0	4.7	T	N	T
GRP073-1	6.8	57.2	3.1	14.2	121.2	47.8	7.9	T	T	N	T
GRP095-1	0.6	0.3	0.0	0.0	0.3	0.7	0.9	0.7	T	N	T
GRP100-1	1.1	0.5	1.4	0.6	0.5	3.7	294.8	13.9	T	N	T
GRP104-1	2.0	40.3	1.7	1.3	40.2	6.1	281.1	19.2	T	N	T
GRP108-1	2.2	39.6	0.4	1.7	39.5	7.9	0.1	2.4	T	3.3	T
LAT005-4	237.6	95.0	209.7	227.3	52.8	T	T	T	158.7	T	T

Figure 1: Solution number vs CPU time, for the MIX division

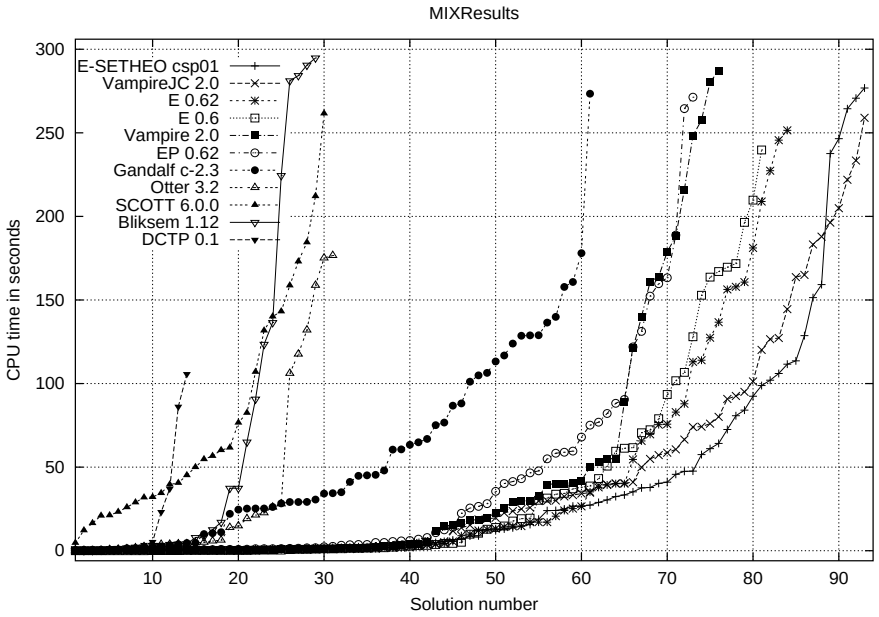


Table 11: UEQ division results

ATP System	UEQ /90	Average time	Proof output?
Waldmeister 601	69	9.6	yes
Waldmeister 600	69	12.0	yes
E	43	34.3	no
SCOTT	23	93.5	yes
Otter	22	22.2	no
Bliksem	13	40.8	yes
Vampire	8	66.8	yes
Gandalf	7	108.1	yes

4.2.1 Times Taken for Each Problem

Tables 12 and 13 shows the CPU times taken by the systems, for each problem, in seconds. No problems were solved by all the systems. None of the twenty problems of rating 1.00 were solved, but all of the other problems were solved by at least one system. Nineteen problems were solved by only the Waldmeister systems. The Waldmeisters subsume Gandalf and Vampire, and E also subsumes Vampire. It is interesting to note that the one new problem, `LAT038-1`, was solved by four systems, but not by either of the Waldmeister systems.

4.2.2 Proofs as a Function of Time

Figure 2 plots the CPU times taken by each system for proofs found, in increasing order of CPU time taken. The dominance of the Waldmeister systems is clear both in terms of problems solved and times taken, and only for some harder problems does Waldmeister 601 outperform the older version. As was the case in CASC-17, it seems likely that SCOTT may have benefited from a higher time limit, but this would not have changed the outcome of the division.

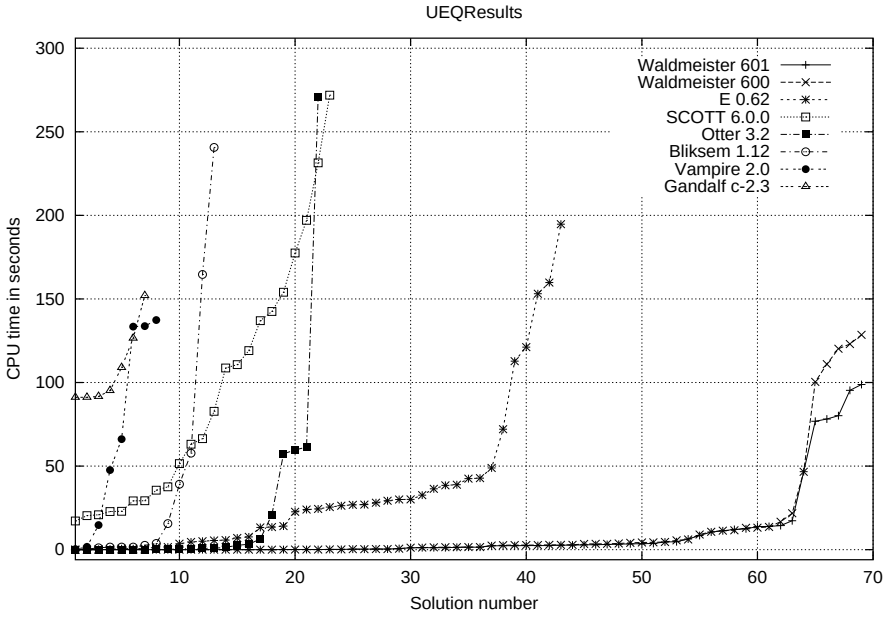
Table 12: CPU times in seconds for the UEQ division

Problem	Wald'r 600	Wald'r 601	SCOTT	E	Bliksem	Otter	Vampire	Gandalf
COL002-5	0.0	0.0	T	22.8	0.1	0.0	T	N
COL003-1	2.8	2.8	24.0	T	N	T	T	N
COL004-1	0.2	0.2	7.7	T	T	T	T	N
COL006-1	1.2	1.2	T	T	T	T	T	126.5
COL006-7	0.0	0.0	T	T	T	T	T	N
COL011-1	0.4	0.4	153.1	17.2	0.6	T	T	91.2
COL033-1	0.0	0.0	0.2	110.7	0.0	T	0.1	91.7
COL034-1	0.0	0.0	1.6	66.4	0.1	T	T	95.2
COL037-1	0.1	0.1	1.5	23.0	2.6	T	T	91.2
COL038-1	1.1	1.5	T	T	T	T	T	N
COL042-1	1.5	1.6	T	T	T	T	T	N
COL042-6	0.0	0.0	13.3	20.9	0.1	1.6	133.7	N
COL042-8	0.0	0.0	13.5	29.3	3.3	1.6	133.4	N
COL042-9	0.0	0.0	14.1	35.6	0.1	1.6	137.4	N
COL043-1	2.7	2.5	T	T	T	T	T	N
COL043-3	0.0	0.0	T	T	T	T	T	N
COL044-6	0.0	0.0	T	T	T	T	T	N
COL044-7	0.0	0.0	T	T	T	T	T	N
COL044-8	0.0	0.0	T	T	T	T	T	N
COL044-9	0.0	0.0	T	T	T	T	T	N
COL046-1	1.2	1.3	48.9	T	T	T	T	N
COL049-1	0.3	0.3	32.6	82.8	0.3	T	1.7	109.0
COL066-1	3.7	4.3	112.8	T	T	T	T	T
COL067-1	T	T	T	T	T	T	T	T
COL068-1	T	T	T	T	T	T	T	N
GRP164-1	98.8	100.4	T	T	T	T	T	N
GRP167-3	2.9	3.5	22.7	231.4	T	T	T	N
GRP167-4	3.0	3.4	42.8	T	T	T	T	N
GRP177-1	T	T	T	T	T	T	T	N
GRP177-2	6.4	6.1	T	T	1.4	T	T	T
GRP178-1	1.4	1.4	121.2	154.0	T	T	T	T
GRP178-2	1.5	1.5	26.3	T	T	T	T	N
GRP179-1	2.5	2.9	24.4	T	T	T	T	N
GRP179-2	2.3	2.8	27.0	T	T	T	T	N
GRP179-3	2.4	2.7	30.1	T	T	T	T	T
GRP180-1	3.3	3.7	30.0	T	T	T	T	N
GRP181-3	10.6	21.8	7.0	20.4	T	39.2	47.6	N
GRP183-2	2.5	2.9	42.5	T	T	T	T	T
GRP183-3	2.4	2.8	36.4	271.9	T	T	T	N
GRP183-4	2.5	2.9	29.3	T	T	T	T	T
GRP185-1	0.3	0.3	T	37.6	T	15.6	T	152.0
GRP185-2	0.2	0.2	T	T	T	3.7	T	T
GRP185-4	0.1	0.1	T	T	270.6	T	T	T
GRP186-2	3.1	3.5	38.5	T	T	T	T	T
GRP187-1	46.7	46.6	T	T	T	T	T	N
GRP196-1	U	U	T	T	T	T	T	T
LAT010-1	8.5	9.1	5.1	177.5	N	240.6	T	T
LAT017-1	1.2	1.6	25.5	T	T	T	T	N
LAT018-1	T	T	T	T	T	T	T	N
LAT019-1	0.4	0.4	0.3	51.5	6.4	0.3	14.7	T
LAT020-1	76.8	111.0	72.1	T	T	164.6	T	T
LAT022-1	3.6	5.6	159.8	108.7	61.4	T	T	T
LAT023-1	3.5	4.1	38.9	63.1	57.4	T	T	T
LAT038-1*	T	T	26.8	29.2	0.8	2.6	T	T
RNG009-5	4.6	4.4	28.1	T	T	57.7	T	T
RNG009-7	4.8	4.6	4.6	T	T	1.1	66.1	T
RNG010-5	T	T	T	T	T	T	T	T
RNG010-6	T	T	T	T	T	T	T	T
RNG010-7	T	T	T	T	T	T	T	T
RNG019-6	0.0	0.0	3.6	137.0	0.0	T	T	N
RNG020-6	0.0	0.0	0.7	119.1	0.4	T	T	N
RNG021-6	0.0	0.0	0.6	197.0	0.0	T	T	N
RNG021-7	0.0	0.0	0.6	T	0.2	T	T	N
RNG025-5	0.7	0.6	T	T	59.7	T	T	N
RNG025-6	0.1	0.0	5.7	142.5	0.8	T	T	N
RNG025-7	0.1	0.1	5.5	T	20.9	T	T	N
RNG026-6	0.0	0.0	0.0	T	T	T	T	N
RNG026-7	0.0	0.0	0.0	T	T	T	T	N
RNG027-7	11.3	10.7	T	T	T	T	T	N
RNG027-8	12.1	11.4	T	T	T	T	T	N
RNG027-9	12.5	11.6	T	T	T	T	T	N
RNG028-5	13.9	13.2	T	T	T	T	T	N
RNG028-7	13.8	13.1	T	T	T	T	T	N
RNG028-9	14.4	13.6	T	T	T	T	T	N
RNG030-6	T	T	T	T	T	T	T	N
RNG030-7	T	T	T	T	T	T	T	N

Table 13: CPU times in seconds for the UEQ division, continued

Problem	Wald'r 600	Wald'r 601	SCOTT E	Otter	Bliksem	Vampire	Gandalf
RNG031-6	T	T	T	T	T	T	T	N
RNG032-6	T	T	T	T	N	T	T	N
RNG032-7	T	T	T	T	T	T	T	N
RNG033-6	T	T	T	N	T	T	T	N
RNG033-8	T	T	T	T	T	T	T	T
RNG035-7	17.3	16.7	194.7	T	T	T	T	T
RNG036-7	U	U	T	T	N	T	T	T
ROB006-1	80.2	120.2	T	T	T	T	T	T
ROB006-2	95.3	128.5	T	T	T	T	T	N
ROB007-2	T	U	T	T	T	T	T	N
ROB020-2	T	U	T	T	T	T	T	N
ROB024-1	U	U	T	T	T	T	T	T
ROB026-1	78.2	123.0	T	T	T	T	T	T
ROB027-1	U	U	T	T	T	T	T	N

Figure 2: Solution number vs CPU time, for the UEQ division



4.3 The SAT Division

Table 14 summarizes the results in the SAT division. For the first time ever, in any division, the previous CASC's division winner, here GandalfSat 1.0, outperformed the new systems, including the new version of the same system.[3] Therefore no winner was announced. Only SCOTT and MACE output models.

In the SNE category DCTP performed much better than it does in the division as a whole, and in the SEQ category SCOTT outperformed the other systems. GandalfSat 1.1 and E-SETHEO performed consistently well across the two categories. An interesting feature is the extremely low average CPU times of SCOTT.

Table 15 shows, for each system, the fractions of old problems, new problems, and all problems solved. In contrast to the other divisions, where the winners all solved more than 75% of the problems, GandalfSat 1.0 solved only 53% of the problems. This difference in success rate may be partially attributable to "harder" problems in the SAT division - they had an average difficulty rating of 0.57, as opposed to 0.49, 0.54, 0.47, and 0.31 in the MIX, UEQ, FOF, and EPR divisions, respectively. It is interesting that E-SETHEO and DCTP did significantly worse on the old problems than the new problems, while all the other systems did better on the old problems. This skewing can be attributed to the large fraction of new problems (62% of the division) with low diversity - 45 of the new problems (80%) were `NLP` problems. DCTP was well suited to the 20 `NLP` problems in the SNE category, solving 15 of them. E-SETHEO also solved these 15 problems using its DCTP component, and overall was the only system to solve more than half of the new problems.

Table 14: SAT division and category results

ATP System	SAT /90	Avg time	SNE /40	Avg time	SEQ /50	Avg time	Model output?
GandalfSat 1.0	48	15.9	27	10.8	21	22.5	no
GandalfSat 1.1	46	26.7	22	14.0	24	38.3	no
E-SETHEO	44	35.5	21	54.7	23	18.0	no
SCOTT	41	1.5	17	2.1	24	1.2	yes
MACE	25	20.4	8	47.3	17	7.8	yes
DCTP	20	10.1	19	10.7	1	0.0	no

Table 15: Fractions of old, new, and all problems solved in the SAT division

	Old	New	All
Number	34	56	90
Av. rating	0.54	0.59	0.57
GandalfSat 1.0	0.62	0.48	0.53
GandalfSat 1.1	0.68	0.41	0.51
E-SETHEO	0.15	0.70	0.49
SCOTT	0.77	0.27	0.46
MACE	0.47	0.16	0.28
DCTP	0.03	0.34	0.22

4.3.1 Times Taken for Each Problem

Tables 16 and 17 show the CPU times taken by the systems, for each problem, in seconds. No problems were solved by all the systems, and twelve problems were unsolved. Of those, nine have been solved by SPASS [WGR96], which was not entered into CASC-JC, two have been solved by an older version of MACE, and one has been solved by SCOTT (the clause reordering is assumed to be the cause of SCOTT's failure in CASC-JC). No system subsumes another. The low CPU times of SCOTT are again evident: over problems solved by both SCOTT and GandalfSat 1.0, SCOTT is an order of magnitude faster.

4.3.2 Disproofs as a Function of Time

Figure 3 plots the CPU times taken by each system for disproofs found, in increasing order of CPU time taken. The stepping effect in the plots for GandalfSat and E-SETHEO clearly reflect the nature of these two systems, which both

[3] Subsequent email from the system developer suggests that he had neglected to enable certain features in the new version, so the new system did not perform as well as it was intended to.

Table 16: CPU times in seconds for the SNE category

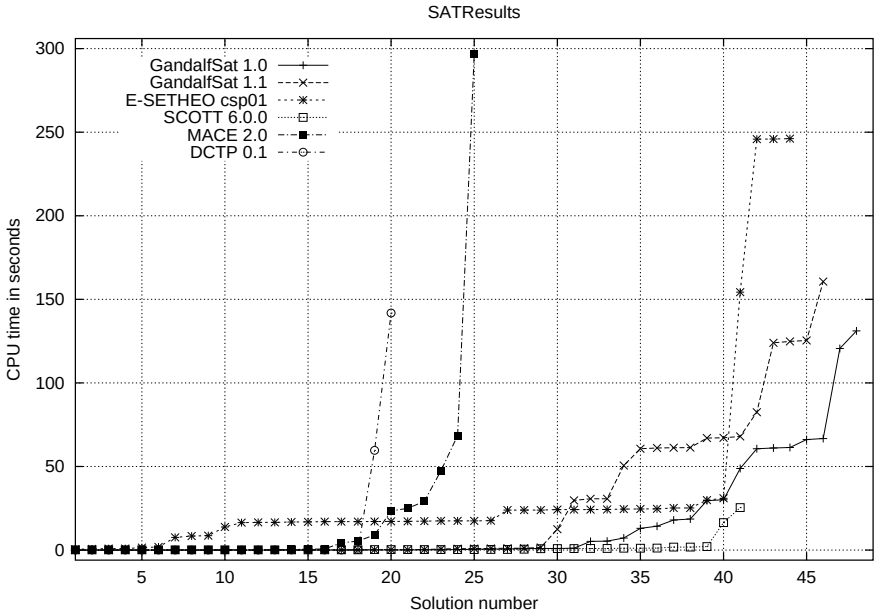
Problem	Gand'Sat 1.1		DCTP			MACE
	Gand'Sat 1.0		E-S'O	SCOTT		
LCL078-1	0.3	0.3	T	T	0.2	0.0
LCL168-1	N	N	T	T	T	T
NLP027-1*	0.0	0.0	T	T	0.3	N
NLP031-1*	0.0	0.0	16.6	0.1	0.3	N
NLP044-1*	0.0	0.1	16.9	0.0	N	T
NLP045-1*	0.1	0.1	16.8	0.0	N	0.1
NLP047-1*	0.1	0.0	17.0	0.0	N	25.0
NLP060-1*	0.0	0.0	T	T	0.3	N
NLP138-1*	0.1	0.2	17.0	0.0	N	T
NLP160-1*	0.5	0.7	16.6	0.0	T	47.5
NLP164-1*	0.7	0.9	T	141.8	T	T
NLP165-1*	0.9	1.1	T	59.7	T	9.1
NLP167-1*	0.6	0.8	T	T	T	0.4
NLP169-1*	1.0	1.6	16.4	0.0	T	T
NLP192-1*	5.4	T	17.3	0.0	T	N
NLP194-1*	5.2	82.5	17.1	0.0	T	N
NLP195-1*	7.2	T	17.2	0.0	T	N
NLP199-1*	14.3	T	17.2	0.1	T	N
NLP231-1*	18.0	T	17.5	0.1	T	N
NLP232-1*	T	T	17.5	0.3	T	N
NLP234-1*	T	T	17.5	0.1	T	N
NLP239-1*	18.6	T	17.6	0.1	T	N
SWV012-1*	0.0	0.0	245.9	0.0	0.2	N
SWV013-1*	0.0	0.0	245.8	0.0	0.2	N
SWV015-1*	U	U	246.1	T	0.2	N
SWV016-1*	U	U	T	T	0.4	N
SWV018-1*	U	U	T	N	T	N
SYN303-1	60.6	60.6	0.5	T	N	T
SYN330-1	66.7	67.0	0.7	T	T	296.6
SYN335-1	61.4	61.3	T	T	T	T
TOP001-1	N	N	T	T	25.4	N
TOP003-1	N	N	T	T	1.1	N
TOP003-2	0.0	0.0	154.3	T	0.2	0.0
TOP006-1	N	N	T	U	T	N
TOP008-1	N	N	T	T	1.0	N
TOP013-1	N	N	T	U	1.1	N
TOP016-1	N	T	T	U	1.0	N
TOP017-1	0.2	0.2	T	U	1.2	N
TOP018-1	30.1	30.7	T	U	0.9	N
TOP019-1	N	N	T	T	0.9	N

Table 17: CPU times in seconds for the SEQ category

Problem	Gand'Sat 1.1		Gand'Sat 1.0			DCTP
	SCOTT		E-S'O	MACE		
ALG008-1	0.5	30.7	T	0.0	0.1	T
BOO008-3	0.2	0.0	7.6	0.0	0.0	U
BOO030-1	0.2	0.0	T	0.0	0.0	T
BOO033-1	0.2	0.0	T	0.0	0.1	T
CAT015-3	N	0.0	29.9	0.0	N	0.0
COL005-1	0.2	0.0	T	0.0	0.2	T
COL047-1	0.3	68.0	T	U	0.7	T
COL071-1	0.4	160.6	T	U	4.3	T
COL073-1	0.8	U	T	U	5.4	T
GRP025-4	T	U	T	U	T	T
GRP081-1	0.2	1.1	T	1.1	N	T
GRP112-1	0.2	0.1	T	0.1	N	T
GRP207-1*	0.2	0.0	T	0.0	0.0	T
LCL142-1	0.2	1.0	T	1.0	0.0	T
LCL165-1	T	U	T	U	29.1	T
LCL206-3	0.2	0.2	T	0.2	0.0	T
LCL267-3	0.3	29.7	T	29.6	N	T
LCL280-3	0.3	67.2	T	66.1	0.1	T
LCL288-3	0.2	12.5	T	13.0	N	T
LCL291-3	0.2	50.6	T	48.8	0.0	T
MGT033-1*	16.4	125.4	0.9	131.1	23.4	T
MGT033-2*	0.2	124.0	31.0	120.6	68.5	U
MGT037-2*	N	T	0.9	T	T	T
MGT038-2*	0.2	0.0	T	0.0	0.5	U
NLP040-1*	0.3	0.0	1.9	0.0	N	T
NLP055-1*	T	T	T	T	N	T
NLP058-1*	T	T	24.2	T	N	T
NLP074-1*	T	T	17.0	T	N	T
NLP075-1*	T	T	13.8	T	N	T
NLP082-1*	T	T	25.2	T	N	T
NLP085-1*	T	T	25.2	T	N	T
NLP089-1*	T	T	T	T	N	U
NLP090-1*	T	T	T	T	N	U
NLP093-1*	T	T	T	T	N	T
NLP150-1*	1.8	61.1	23.9	61.1	N	U
NLP155-1*	2.1	124.8	23.9	T	N	T
NLP157-1*	1.8	61.2	23.9	0.0	N	T
NLP170-1*	N	U	24.2	T	N	T
NLP172-1*	N	U	24.2	T	N	T
NLP173-1*	N	U	24.3	T	N	T
NLP181-1*	T	U	T	T	N	U
NLP184-1*	T	T	T	T	N	U
NLP188-1*	T	T	T	T	N	U
NLP189-1*	T	U	T	T	N	U
NLP203-1*	N	U	24.7	T	N	U
NLP206-1*	N	U	24.6	T	N	U
NLP209-1*	N	U	24.5	T	N	T
NLP250-1*	N	U	8.4	T	N	T
NLP259-1*	N	U	8.6	U	N	T
SWC128-1*	T	T	1.5	T	N	U

try a sequence of strategies, allowing a fraction of the CPU time limit for each strategy, until some strategy solves the problem or the CPU time limit is exhausted. Each step in the plot corresponds to problems solved by a certain strategy. Only one problem was solved in more that 250 seconds, indicating that the CPU time limit was adequate for this division. The system ranking is stable for time limits from about 75 seconds, so the final competition ranking appears to be reasonably time limit independent.

Figure 3: Solution number vs CPU time, for the SAT division



4.4 The FOF Division

Table 18 summarizes the results in the FOF division. All the systems except MUSCADET work by converting to CNF and producing a CNF refutation. The winner is E-SETHEO. The second place system, VampireFOF 1.0, is the winner of the FOF division from CASC-17, suggesting that only modest progress has been made in the FOF division in the last year (although it must be noted that VampireFOF 1.0 solves 80% of the problems, making it quite hard for another system to solve many more, compared to the MIX division where the CASC-17 winner solved only 68% of the problems). VampireFOF 1.0 outperformed VampireFOF 2.0, apparently because the new version uses a new experimental clausifier, while the old version uses FLOTTER [WGR96]. The new clausifier has some advanced optimizations on the formula level that FLOTTER does not have, but has a very primitive clausification algorithm as compared to FLOTTER's. Five of the seven CNF based systems output a refutation for the CNF of the problem, and MUSCADET does not output a proof.

The ranking in the FNE category is almost the same as for the division, but with the top three systems very close together. In the FEQ category GandalfFOF performs significantly better than it does in the division as a whole. MUSCADET is also clearly specialized to the FEQ category.

Table 19 shows, for each system, the fractions of old problems, new problems, and all problems solved. The top four systems all solved more than 75% of all the problems, and did comparably well on the harder new problems. Of the 17 new problems, 16 were in the FEQ category. The best performance on these came from VampireFOF 1.0, which solved 15 new problems, all from the FEQ category. In contast, the four weaker systems did very poorly on the new problems

4.4.1 Times Taken for Each Problem

Tables 20 and 21 show the CPU times taken by the systems, for each problem, in seconds. Three problems were solved by all the systems, and one problem was unsolved. E-SETHEO subsumes SCOTT, GandalfFOF subsumes Otter, and VampireFOF subsumes Bliksem.

In the FNE category, all except `SWV014+1` are problems with a finite Herbrand universe. The difference between the CNF based systems and MUSCADET on these problems is highlighted by the fact that most of them were

Table 18: FOF division and category results

ATP System	FOF /90	Avg time	FNE /40	Avg time	FEQ /50	Avg time	Proof output?
E-SETHEO	75	17.6	40	2.7	35	34.7	no
VampireFOF 1.0	72	8.6	39	4.0	33	14.0	yes
VampireFOF 2.0	71	27.3	39	12.7	32	45.5	yes
GandalfFOF	68	32.5	30	7.2	38	52.4	yes
Otter	43	21.6	27	0.1	16	58.0	no
SCOTT	39	17.6	28	13.0	11	29.2	yes
Bliksem	34	9.4	26	0.7	8	37.7	yes
MUSCADET	18	0.9	2	0.3	16	1.0	no

Table 19: Fractions of old, new, and all problems solved in the FOF division

	Old	New	All
Number	73	17	90
Av. rating	0.45	0.53	0.47
E-SETHEO	0.86	0.71	0.83
VampireFOF 1.0	0.78	0.88	0.80
VampireFOF 2.0	0.80	0.77	0.79
GandalfFOF	0.77	0.71	0.76
Otter	0.55	0.18	0.48
SCOTT	0.49	0.18	0.43
Bliksem	0.44	0.12	0.38
MUSCADET	0.25	0.00	0.20

solved quickly by all systems except MUSCADET. MUSCADET's specialization is highlighted by its solution of five FEQ set theory problems that were not solved by any other system. Another interesting aspect of MUSCADET's performance is its consistently very low average CPU times, on the problems that it can solve.

4.4.2 Proofs as a Function of Time

Figure 4 plots the CPU times taken by each system for proofs found, in increasing order of CPU time taken. As in the MIX division, the plots divide the systems into groups, here into four groups. The first group contains E-SETHEO and VampireFOF 1.0. VampireFOF 2.0 is alone the second group, starting out with a performance similar to the first group, then falling back to a performance similar to the third group, which contains GandalfFOF. The final group contains the remaining three systems. The two systems in the first group use the FLOTTER clausifier. E-SETHEO and VampireFOF 2.0 each solved two problems in more than 250 seconds, suggesting that they may have solved a few more problems with a higher time limit. The system ranking is however stable for time limits from about 130 seconds, so the final competition ranking appears to be reasonably time limit independent.

4.5 The EPR Division

Table 22 summarizes the results in the EPR division. The division was won by E-SETHEO, by a large margin. It is interesting to note that for this type of problem E-SETHEO relies largely on grounding and invoking a propositional decision procedure, using first-order techniques only when the grounding program fails because there are too many ground instances. PizEAndSATO, although not competing, also ran on the general hardware, and solved the second most problems with an average CPU time much lower than the other systems. PizEAndSATO uses only the grounding approach. Evidently the grounding approach is effective for this type of problem. Only SCOTT produces both proofs and models.

In the EPT category VampireEPR and Gandalf perform better than they do in the division as a whole, and in the EPS category DCTP performs better than it does in the division. Note that E-SETHEO solved all the problems in the EPS category.

The problems had an average difficulty rating of 0.31. The five new problems in the division were all in the EPS category, and were solved by all systems that could attempt them (the five problems exceeded VampireEPR's internal limit of 31 literals per clause, and Gandalf was designed only to prove theorems).

Table 20: CPU times in seconds for the FNE category

Problem	Vamp'FOF 1.0 E-S'O	Vamp'FOF 2.0	GandalfFOF	SCOTT	Otter	Bliksem	MUSCADET
COM003+1	17.3	4.7	170.7	111.1	T	T	T	N
PUZ031+1	1.2	0.2	0.2	0.0	83.7	0.0	3.2	N
SWV014+1*	1.5	T	0.2	101.5	T	T	T	N
SYN048+1	0.9	0.2	0.1	0.0	0.2	0.0	0.0	N
SYN049+1	1.0	0.2	0.2	0.0	0.2	0.0	0.0	N
SYN050+1	1.0	0.2	0.2	0.0	0.3	0.0	0.0	N
SYN051+1	1.0	0.2	0.1	0.0	0.2	0.0	0.0	N
SYN053+1	1.1	0.3	0.2	0.0	0.2	0.0	0.0	N
SYN057+1	0.9	0.2	0.1	0.0	1.7	0.0	0.0	N
SYN063+1	1.0	0.2	0.1	0.0	0.2	0.0	0.0	N
SYN070+1	1.0	0.2	0.1	0.0	1.2	0.0	0.0	N
SYN317+1	1.0	0.2	0.1	0.0	0.2	0.0	0.0	N
SYN318+1	1.0	0.2	0.1	0.0	0.1	0.0	0.0	N
SYN323+1	1.1	0.2	0.1	0.0	0.2	0.0	0.0	N
SYN332+1	1.2	0.2	0.1	U	U	N	T	U
SYN341+1	1.0	0.1	0.2	0.0	2.1	0.0	0.0	N
SYN346+1	1.0	0.2	0.2	0.0	0.2	0.0	0.0	N
SYN347+1	1.1	0.3	0.2	0.0	10.2	0.0	T	N
SYN353+1	1.5	0.3	0.2	0.7	T	T	T	N
SYN360+1	1.0	0.2	0.1	0.0	0.2	0.0	0.0	N
SYN361+1	1.2	0.2	0.1	0.0	0.2	0.0	0.0	0.3
SYN363+1	1.0	0.2	0.2	0.0	0.2	0.0	0.0	N
SYN367+1	1.0	0.2	0.1	0.0	0.2	0.0	0.0	N
SYN368+1	1.0	0.2	0.2	0.0	0.2	0.0	0.0	N
SYN373+1	1.1	0.2	0.2	0.0	0.1	0.0	0.0	N
SYN383+1	0.9	0.2	0.1	0.0	0.2	0.0	0.0	N
SYN385+1	1.1	0.2	0.1	0.0	0.2	0.0	0.0	N
SYN396+1	1.1	0.2	0.1	0.1	0.2	0.0	0.0	0.2
SYN407+1	1.0	0.1	0.1	0.0	0.2	0.0	0.0	N
SYN410+1	0.9	0.1	0.1	0.0	0.2	0.0	0.0	N
SYN451+1	3.1	13.0	3.2	U	T	N	T	N
SYN457+1	30.5	14.4	5.0	U	T	N	T	N
SYN458+1	3.1	21.2	2.5	U	T	N	T	N
SYN469+1	3.1	26.1	3.5	U	T	N	T	N
SYN472+1	3.8	14.0	4.3	U	T	N	T	N
SYN482+1	4.2	24.8	4.3	U	T	N	T	N
SYN508+1	3.9	18.1	4.2	U	T	N	T	N
SYN512+1	3.2	11.5	4.3	U	T	N	T	N
SYN548+1	1.8	0.3	T	N	259.3	T	14.4	N
SYN549+1	1.2	0.3	276.0	1.5	1.4	1.9	T	N

Figure 4: Solution number vs CPU time, for the FOF division

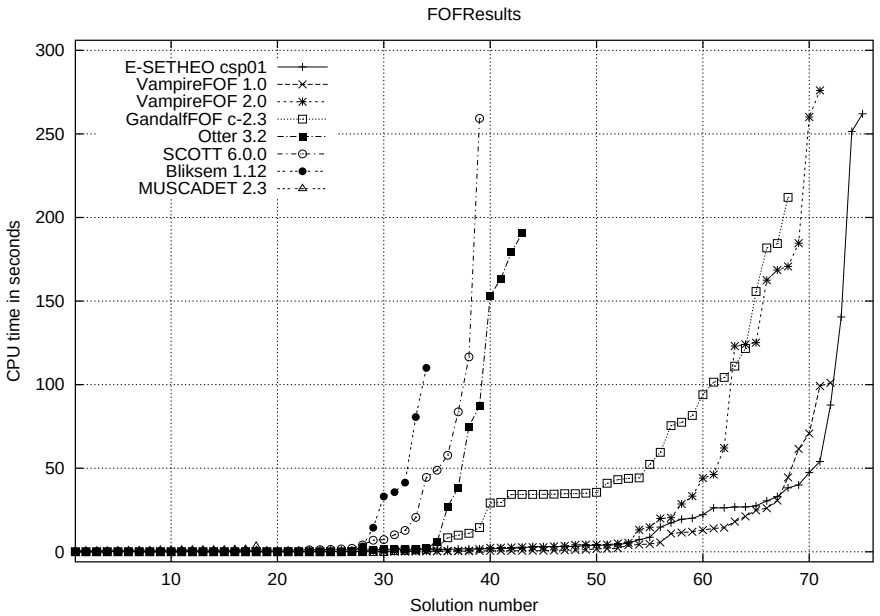


Table 21: CPU times in seconds for the FEQ category

Problem	Gandalf	E-S'O FOF	Vamp'FOF 2.0 Vamp'FOF 1.0	MUSCADET		Otter	SCOTT	Bliksem
GEO112+1*	0.2	22.2	99.2	162.5	N	2.2	0.4	T
GEO148+1*	14.5	3.2	1.9	19.9	N	163.4	T	T
MGT005+2	2.8	1.5	0.6	125.1	0.9	5.9	1.3	T
MGT023+1	0.0	1.5	0.3	0.2	0.5	0.8	N	0.0
MGT034+2	10.0	87.8	0.4	0.6	N	179.1	N	33.1
MGT035+1	U	40.0	5.6	28.6	N	T	N	T
MGT051+1*	34.5	140.5	10.9	168.6	N	T	20.6	T
MGT055+1*	34.3	T	1.1	2.8	N	T	T	T
MGT060+1*	34.3	1.5	101.1	46.2	N	1.9	T	T
MGT062+1*	34.3	2.3	0.4	0.3	N	T	6.9	T
SET055+1	8.4	1.8	U	T	N	0.1	T	T
SET063+4	43.9	1.4	61.5	0.5	0.5	1.8	44.5	0.6
SET094+1	155.7	54.0	0.6	0.4	3.3	74.9	T	T
SET108+1	34.4	1.7	0.5	0.2	N	T	T	0.0
SET200+3	29.2	1.5	T	33.3	0.3	T	116.6	T
SET579+3	44.2	1.8	T	0.4	0.3	T	T	T
SET608+3	59.5	20.1	T	T	0.4	T	T	T
SET610+3	77.5	19.5	T	T	N	T	T	T
SET612+3	81.6	262.0	T	T	N	T	T	T
SET644+3	121.6	8.7	0.4	14.7	N	38.4	T	T
SET649+3	35.1	14.8	70.8	124.1	N	26.8	7.3	T
SET652+3	35.6	251.5	T	184.6	N	T	4.1	T
SET656+3	34.8	5.4	4.0	T	N	153.1	12.8	T
SET660+3	75.5	38.3	1.3	20.4	N	T	T	T
SET661+3	N	T	T	T	N	T	T	T
SET666+3	29.6	2.1	12.0	5.4	N	87.3	48.8	41.4
SET671+3	34.8	4.2	T	T	N	T	T	T
SET686+3	0.8	33.0	0.4	123.1	N	T	T	T
SET694+4	94.0	T	T	T	0.5	T	T	T
SET722+4	N	T	T	T	1.0	N	T	T
SET734+4	N	T	T	T	1.0	T	T	T
SET737+4	U	T	T	T	1.4	N	T	T
SET746+4	N	T	T	T	1.2	T	T	T
SET763+4	43.2	T	T	260.1	1.0	T	T	T
SET764+4	11.1	7.2	2.2	0.7	0.9	T	T	T
SET770+4	N	T	U	T	1.0	T	T	T
SET776+4	0.1	T	4.3	T	1.0	1.9	T	35.7
SWC010+1*	181.8	27.5	1.7	2.8	N	T	T	T
SWC094+1*	184.4	26.8	0.6	1.6	N	T	T	T
SWC206+1*	1.1	26.3	0.9	2.4	N	T	T	110.0
SWC211+1*	N	T	30.7	T	N	T	T	T
SWC227+1*	N	T	44.5	T	N	T	T	T
SWC258+1*	104.3	26.3	0.6	2.8	N	T	T	80.6
SWC281+1*	N	T	T	T	N	T	T	T
SWC302+1*	N	T	0.8	T	N	T	T	T
SWC330+1*	N	47.4	0.7	44.1	N	T	T	T
SWC380+1*	212.0	26.8	0.7	2.6	N	T	T	T
SYN075+1	0.1	1.2	0.2	0.2	N	0.1	57.7	T
SYN417+1	52.3	2.0	0.2	62.1	N	190.5	N	T
SYN551+1	41.0	T	0.3	13.2	N	T	N	T

Table 22: EPR division and category results

ATP System	EPR /50	Avg time	EPT /25	Avg time	EPS /25	Avg time	Proof output?	Model output?
E-SETHEO	49	20.5	24	32.4	25	9.1	no	no
Otter-MACE	28	25.9	11	57.1	17	5.6	no	yes
VampireEPR	27	35.9	19	26.3	8	58.7	yes	no
DCTP	20	14.4	4	28.6	16	10.9	no	no
E	17	24.7	8	38.5	9	12.5	no	no
SCOTT	15	10.7	8	19.4	7	0.8	yes	yes
Gandalf	14	67.9	14	67.9	0	-	no	-
Demonstration division								
PizEAndSATO	41	4.8	19	1.8	22	7.4	no	no

Table 23: CPU times in seconds for the EPT category

Problem	Vamp'EPR	E-S'O	Otter-MACE	Gandalf	SCOTT	E	DCTP	PizE'SATO
GRP125-2.005	0.6	1.5	161.1	121.1	T	T	T	0.1
GRP127-2.006	0.7	T	31.0	73.7	T	T	T	0.2
GRP128-2.006	4.6	T	160.8	149.4	T	T	T	0.2
GRP128-3.005	1.2	T	168.1	T	T	148.4	T	0.5
GRP129-3.004	0.7	T	206.4	125.5	6.7	14.8	T	0.1
GRP129-4.004	0.7	88.3	3.1	0.8	81.3	6.9	T	0.1
LAT005-1	89.6	3.1	6.1	0.5	42.7	16.0	89.3	N
LAT005-2	25.7	2.0	0.0	0.3	0.6	19.6	24.9	N
PUZ010-1	1.1	T	U	T	T	T	T	0.3
PUZ017-1	T	0.2	T	T	T	T	T	25.7
PUZ018-1	2.3	T	0.0	153.9	20.2	9.5	0.0	1.5
PUZ036-1.005	188.9	0.3	0.6	0.4	0.5	92.6	T	N
PUZ037-1	0.6	0.1	0.1	0.1	0.2	0.0	0.0	N
PUZ037-2	189.3	0.5	1.3	2.7	2.8	T	T	N
PUZ037-3	221.5	87.6	47.1	N	T	T	T	N
SYN436-1	1.1	37.8	U	T	T	T	T	0.2
SYN439-1	12.4	28.4	U	T	T	T	T	0.6
SYN440-1	17.0	19.0	U	T	T	T	T	0.6
SYN447-1	2.3	32.4	165.1	T	T	T	T	0.6
SYN457-1	4.7	42.3	U	T	T	T	T	0.7
SYN460-1	2.7	13.2	U	T	T	T	T	0.5
SYN466-1	3.7	17.0	U	T	T	T	T	0.3
SYN467-1	2.1	38.9	U	T	T	T	T	0.3
SYN472-1	1.7	45.4	U	T	T	T	T	0.4
SYN482-1	3.3	41.4	U	T	T	T	T	0.4

4.5.1 Times Taken for Each Problem

Tables 23 and 24 show the CPU times taken by the systems, for each problem, in seconds. No problems were solved by all the systems (including PizEAndSATO), and no problems were unsolved. E-SETHEO subsumes DCTP, E, Gandalf, and Otter-MACE.

Table 24: CPU times in seconds for the EPS category

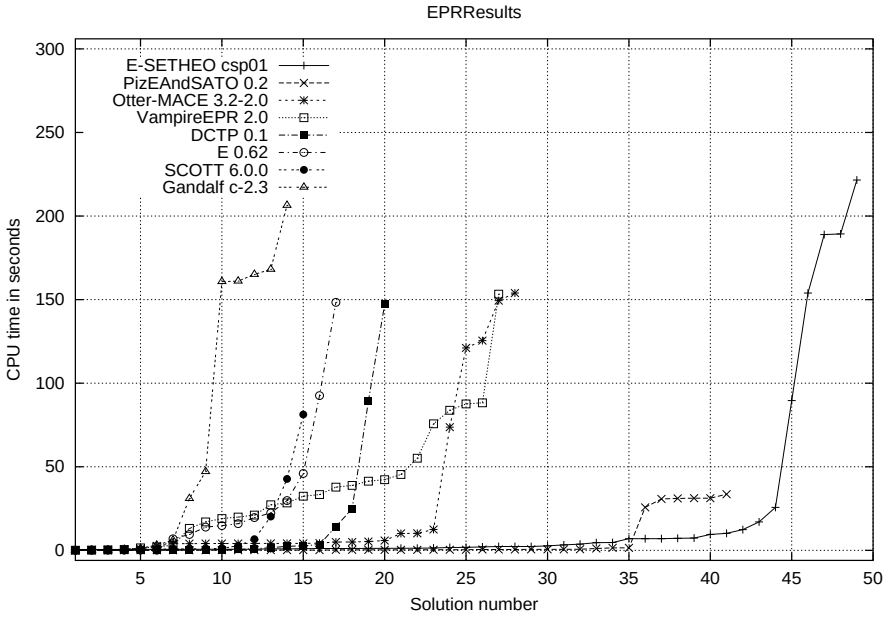
Problem	Otter-MACE	E-S'O	DCTP	E	Vamp'EPR	SCOTT	Gandalf	PizE'SATO
GRP123-2.005	1.8	5.3	1.6	T	T	T	U	1.1
GRP124-7.005	0.6	5.9	1.0	T	T	T	U	0.1
GRP126-2.005	0.6	5.0	14.3	T	T	T	U	0.1
GRP126-3.005	0.9	10.1	T	T	T	T	U	0.4
GRP127-3.005	0.9	10.2	T	T	21.1	T	U	0.3
GRP128-1.004	0.5	4.2	3.5	29.8	75.7	T	U	0.0
GRP129-3.005	1.2	12.5	T	T	T	T	U	0.5
GRP130-3.004	0.7	5.0	T	22.4	153.3	N	U	0.2
GRP130-4.004	0.5	4.3	0.3	46.0	T	N	U	0.0
GRP133-2.004	0.5	4.1	0.1	13.9	27.3	N	U	0.0
NLP005-1*	7.0	4.0	0.0	0.1	T	0.6	U	31.2
NLP006-1*	7.2	4.1	0.3	0.0	T	1.2	N	30.8
NLP008-1*	7.4	4.2	0.2	0.1	T	0.8	N	33.6
NLP012-1*	7.0	4.2	0.0	0.1	T	0.7	U	31.2
NLP013-1*	7.0	4.1	0.0	0.1	T	0.9	N	31.1
PUZ018-2	2.3	T	0.5	T	T	T	U	1.5
SYN307-1	0.5	4.0	0.0	T	T	0.9	U	0.1
SYN423-1	9.5	T	2.5	T	T	T	T	N
SYN428-1	10.2	T	2.7	T	T	T	T	N
SYN434-1	1.3	T	T	T	T	T	U	0.2
SYN437-1	1.3	T	T	T	55.1	T	U	0.3
SYN438-1	1.4	4.2	T	T	83.8	0.5	U	0.2
SYN446-1	1.1	T	T	T	19.9	T	U	0.3
SYN463-1	1.0	T	T	T	33.4	T	U	0.2
SYN544-1	153.9	T	147.2	T	T	T	T	N

4.5.2 Solutions as a Function of Time

Figure 5 plots the CPU times taken by each system for solutions found, in increasing order of CPU time taken. The plots contain the points for EPT problems and for EPS problems. In most cases the points at higher CPU times correspond to either only EPT problems, e.g., the E-SETHEO and Otter-MACE plots, or only EPS problems, e.g., the PizEAndSATO plot. A noteworthy exception is the VampireEPR plot, in which the points at higher CPU times

come from both types of problem. The plots clearly show that E-SETHEO outperformed the other systems. Only two problems were solved in more than 200 seconds, one each by E-SETHEO (the top system in the ranking) and Gandalf (the last system in the ranking), indicating that the CPU time limit was adequate for this division.

Figure 5: Solution number vs CPU time, for the EPR division



5 System Descriptions

This section provides short descriptions of the division winners.

VampireJC, the MIX division Proof class winner and a MIX division Assurance class co-winner, is a system for first-order classical logic. It implements the calculi of ordered binary resolution, hyperresolution, and superposition for handling equality. The splitting rule is simulated by introducing new predicate symbols. A number of standard redundancy criteria and simplification techniques are used for pruning the search space: subsumption, tautology deletion, subsumption resolution, and rewriting by ordered unit equalities. A number of efficient indexing techniques are used to implement all major operations on sets of terms and clauses, such as an improved version [RV00] of code trees [Vor95] for forward subsumption, and a combination of path indexing [Sti89] and database joins for backward subsumption. In the preprocessing stage it exploits a number of primitive techniques, such as elimination of simple predicate and function definitions. Compared to Vampire 1.0 that participated in the previous competition, this version has many more literal selection functions, more flexible splitting without backtracking, and improved memory management. VampireJC adjusts its search strategy based on some some syntactic properties of the problem, such as presence of multiliteral, non-Horn and ground clauses, equations and non-equational literals. Additionally some quantitative characteristics, such as the number of axioms, literals, small and large terms, are considered. Vampire is implemented in C++, and is available through `http://www.cs.man.ac.uk/~riazanoa/Vampire`

E-SETHEO, a MIX division Assurance class co-winner, the FOF division winner, and the EPR division winner, is a compositional theorem prover for formulae in first-order logic. Its principal components are the superposition prover E [Sch01], the model-elimination prover SETHEO [MIL+97], and the disconnection prover DCTP [LS01]. It also includes a grounding procedure and a propositional prover for near-propositional and propositional proof tasks, and uses FLOTTER [WGR96] to transform FOF problems into CNF. E-SETHEO first classifies the given proof problem into one of a set of predetermined categories, and selects a corresponding *schedule* which assigns resources to the different component systems. The components are then invoked sequentially with the predetermined resource limits, and try to solve the proof tasks individually. Schedules are computed automatically (using a combination of genetic algorithms and hill climbing) from results of the different strategies on a test set. E-SETHEO and all its components except for FLOTTER are developed at the Automated Reasoning Group at the Technische Universtität München.

WALDMEISTER 601, the UEQ division winner, is a system for unit equational deduction. Its theoretical basis is unfailing completion, in the sense of [BDP89], with refinements towards ordered completion. The prover saturates the input axiomatization in a repeated cycle that works on a set of active and passive facts. The selection of the reduction ordering and the heuristic guidance of the proof search are described in [HJL99]. Recently, stronger redundancy criteria have been integrated, including ground joinability tests with ordering constraints on variables [AHL00]. In several problem domains this technique is helpful especially for harder proof tasks, as can be seen in the competition results when comparing the system with last year's version. Some restructuring of the prover is in progress, including an implementation of confluence trees with full ordering constraints. However, further work is necessary to have them speed up the proof search. The Waldmeister WWW page is located at `http://www-avenhaus.informatik.uni-kl.de/waldmeister`

6 Conclusion

The IJCAR ATP System Competition was the sixth large scale competition for first order ATP systems. The MIX division Proof class was won by VampireJC, the Assurance class was tied by VampireJC and E-SETHEO, the UEQ division was again won by Waldmeister, the SAT division had no winner because the CASC-17 winner outperformed the newly entered systems, and E-SETHEO won the FOF and EPR divisions.

Two significant changes to the competition design produced positive outcomes for general purpose ATP. First, the use of unseen problems has provided evidence that any tuning done by entrants using TPTP problems seems likely to be effective in general. The evidence is most compelling in the MIX division, where concern about such tuning has been strongest. Second, the ranking of systems in the MIX division according to the number of proofs output has further stimulated interest and research into proof production.

A positive aspect of CASC-JC, in contrast to CASCs-16 and -17, was the level of enthusiasm and interest from both entrants and observers. The entrants made significant efforts to meet the requirements imposed by the competition design, and as a result the systems were more robust and usable than in the past. The small number of system subsumptions confirms that most of the systems are now independently useful. In the environment of the combined IJCAR conference, observers with a broad range of perspectives were evidently interested in the competition and its outcomes. In particular, it was pleasing to see some commercial interest in the best performing systems.

The organizers believe that the competition fulfilled its main motivations: evaluation of relative capabilities of ATP systems, stimulation of research, motivation for improving implementations, and providing an exciting event. For the entrants, their research groups, and their systems, there has been substantial publicity both within and outside the ATP community. The significant efforts that have gone into developing the ATP systems have received public recognition. The competition has provided an overview of which researchers and research groups have decent, running, fully automatic ATP systems.

6.1 Future ATP System Competitions

The major changes planned for CASC-18 are to extend the Proof class of the MIX division to include some form of automated proof checking, and to rank systems in more divisions by both number of problems solved and number of solutions output (proofs and models). Minor changes include an increased CPU time limit in the MIX division, and some mechanism to acknowledge when a system has solved a problem but then runs past the CPU time limit while producing a solution.

As is always the case, it is hoped that the TPTP will grow throughout the year, so that many new problems will be eligible for use in CASC-18.

References

[AHL00] J. Avenhaus, T. Hillenbrand, and B. Löchner. On Using Ground Joinable Equations in Equational Theorem Proving. In P. Baumgartner and H. Zhang, editors, *Proceedings of the 3rd International Workshop on First Order Theorem Proving*, pages 33–43, 2000.

[BDP89] L. Bachmair, N. Dershowitz, and D.A. Plaisted. Completion Without Failure. In H. Ait-Kaci and M. Nivat, editors, *Resolution of Equations in Algebraic Structures*, pages 1–30. Academic Press, 1989.

[FS00] M. Fuchs and G. Sutcliffe. Homogeneous Sets of ATP Problems. Technical Report TR-ARP-09-00, Automated Reasoning Project, Australian National University, Canberra, Australia, 2000.

[HJL99] T. Hillenbrand, A. Jaeger, and B. Löchner. Waldmeister - Improvements in Performance and Ease of Use. In H. Ganzinger, editor, *Proceedings of the 16th International Conference on Automated Deduction*, number 1632 in Lecture Notes in Artificial Intelligence, pages 232–236. Springer-Verlag, 1999.

[LS01] R. Letz and G. Stenz. System Description: DCTP - A Disconnection Calculus Theorem Prover. In R. Gore, A. Leitsch, and T. Nipkow, editors, *Proceedings of the International Joint Conference on Automated Reasoning*, number 2083 in Lecture Notes in Artificial Intelligence, pages 381–385. Springer-Verlag, 2001.

[MIL+97] M. Moser, O. Ibens, R. Letz, J. Steinbach, C. Goller, J. Schumann, and K. Mayr. SETHEO and E-SETHEO: The CADE-13 Systems. *Journal of Automated Reasoning*, 18(2):237–246, 1997.

[RV00] A. Riazanov and A. Voronkov. Partiallly Adaptive Code Trees. In M. Ojeda-Aciego, I. de Guzman, G. Brewka, and L. Pereira, editors, *Proceedings of the 7th European Workshop on Logics in Artificial Inteligence*, number 1919 in Lecture Notes in Artificial Intelligence, pages 209–223. Springer-Verlag, 2000.

[Sch01] S. Schulz. System Abstract: E 0.61. In R. Gore, A. Leitsch, and T. Nipkow, editors, *Proceedings of the International Joint Conference on Automated Reasoning*, number 2083 in Lecture Notes in Artificial Intelligence, pages 370–375. Springer-Verlag, 2001.

[SS97a] G. Sutcliffe and C.B. Suttner. Special Issue: The CADE-13 ATP System Competition. *Journal of Automated Reasoning*, 18(2), 1997.

[SS97b] G. Sutcliffe and C.B. Suttner. The Procedures of the CADE-13 ATP System Competition. *Journal of Automated Reasoning*, 18(2):163–169, 1997.

[SS97c] C.B. Suttner and G. Sutcliffe. The Design of the CADE-13 ATP System Competition. *Journal of Automated Reasoning*, 18(2):139–162, 1997.

[SS98a] G. Sutcliffe and C. Suttner. The CADE-14 ATP System Competition. Technical Report 98/01, Department of Computer Science, James Cook University, Townsville, Australia, 1998.

[SS98b] G. Sutcliffe and C.B. Suttner. Proceedings of the CADE-15 ATP System Competition. Lindau, Germany, 1998.

[SS98c] G. Sutcliffe and C.B. Suttner. The TPTP Problem Library: CNF Release v1.2.1. *Journal of Automated Reasoning*, 21(2):177–203, 1998.

[SS98d] C.B. Suttner and G. Sutcliffe. The CADE-14 ATP System Competition. *Journal of Automated Reasoning*, 21(1):99–134, 1998.

[SS99] G. Sutcliffe and C.B. Suttner. The CADE-15 ATP System Competition. *Journal of Automated Reasoning*, 23(1):1–23, 1999.

[SS01] G. Sutcliffe and C.B. Suttner. Evaluating General Purpose Automated Theorem Proving Systems. *Artificial Intelligence*, 131(1-2):39–54, 2001.

[Sti89] M.E. Stickel. A Prolog Technology Theorem Prover: A New Exposition and Implementation in Prolog. Technical Report Technical Note 464, SRI International, Menlo Park, USA, 1989.

[Sut99] G. Sutcliffe. Proceedings of the CADE-16 ATP System Competition. Trento, Italy, 1999.

[Sut00a] G. Sutcliffe. Proceedings of the CADE-17 ATP System Competition. Pittsburgh, USA, 2000.

[Sut00b] G. Sutcliffe. The CADE-16 ATP System Competition. *Journal of Automated Reasoning*, 24(3):371–396, 2000.

[Sut01a] G. Sutcliffe. Proceedings of the IJCAR ATP System Competition. Siena, Italy, 2001.

[Sut01b] G. Sutcliffe. The CADE-17 ATP System Competition. *Journal of Automated Reasoning*, To appear, 2001.

[Vor95] A. Voronkov. The Anatomy of Vampire. *Journal of Automated Reasoning*, 15(2):237–265, 1995.

[WGR96] C. Weidenbach, B. Gaede, and G. Rock. SPASS and FLOTTER. In M. McRobbie and J.K. Slaney, editors, *Proceedings of the 13th International Conference on Automated Deduction*, number 1104 in Lecture Notes in Artificial Intelligence, pages 141–145. Springer-Verlag, 1996.