SMT-EVAL 2013

Progress report
on the 2013 SMT Evaluation

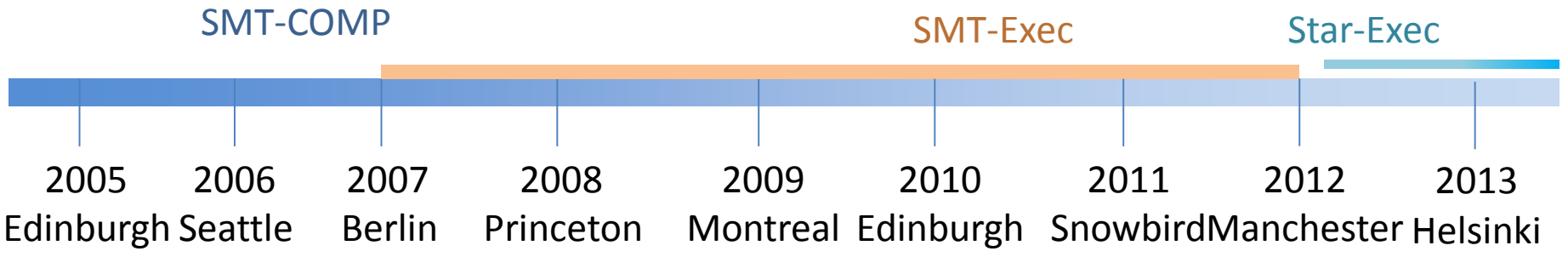
Aaron Stump
U. Iowa,
Iowa City, IA USA

Tjark Weber
Uppsala U.,
Uppsala, Sweden

David Cok
GrammaTech, Inc.
Ithaca, NY USA

A Brief History of Everything

(related to the SMT Competition)



1400 benchmarks
7 divisions
12 solvers

42000 benchmarks
11 divisions
12 solvers

55000 benchmarks
12 divisions
9 solvers

60000 benchmarks
15 divisions
13 solvers

93000 benchmarks
20 divisions
12 solvers

93500 benchmarks
19 divisions
2 tracks
10 solvers

100000+ benchm
21 divisions
3 tracks
11 solvers

Added ~13800 benchmarks
8 divisions
5 tracks
11 solvers

~100,000 main track benchmarks
22 divisions
1 track (so far)
9+32+4 solvers (current and historical)

SMT Competition Goals

(2005-2012)

- Benchmarking research on SMT solvers
- Introduce implementors and users
- Promote standard format (SMT-LIB *v2*)
  - collect additional benchmarks
  - identification/definition of theories for SMT
- Encourage development of industrial-strength solvers for wide-spread use

SMT Evaluation 2013

- Desire by community to slow the pace of competitions
- Take a breath and evaluate where we are
- Enable the transition to Star-Exec without the pressure to have it ready for a competition

Challenges

- Competition results were somewhat predictable
- Need more and better benchmarks, especially from applications
- Less focus on ‘winning’, more on progress
- Have a variety of metrics of progress

Evaluation Goals

- Non-live event, exploring a larger performance space
- Provide a better picture of the state of the art

Timeline

- Late 2012 – Early 2013: discussion
- Jan 2013: Decision to do SMT-EVAL, team formed, announcement and call for comments issued
- 9 Feb 2013: Call for evaluation suggestions, solvers and benchmarks issued
- 7 March 2013: Benchmarks uploaded to StarExec
- ~27 March 2013: Began uploading solvers, as supplied by developers
- Early April: Small sample jobs with supplied solvers
- April, May: Final solvers, improvements to StarExec
- **7 June 2013: Started evaluation runs (with some restarts)**
  - ~ 1.6M job pairs
- **6 July:** First ¼ completed
- **8-9 July: SMT Workshop – status report**
- October 2013: expected completion of evaluation runs and studies

What to Evaluate?

(question set is evolving...)

Questions About Logics

- Which logics are useful in practice? For which applications?
- Which logics have good support in solvers?
- Which logics need implementation work?
- Which logics are no longer relevant?
- We have nearly all combinations of

  QF x A x UF x [BV, LIA, LRA, LIRA]

  Should we consolidate?

Questions About Benchmarks

- What is the source of existing benchmarks? What is the connection between application domains and benchmarks?
- Which application domains need more benchmarks?
- What new application domains could be supported by benchmarks, logic and solver development?
- Do they provide adequate guidance for solver development?
  - Are they adequately representative of the problem space?
  - Do they provide adequate challenge?

Questions About Solvers

- Which capabilities are available?
  - E.g., what features of SMT-LIBv2 are supported?
  - E.g., what additional features are needed in SMT-LIBv2?
- How compliant are existing solvers to SMTLIBv2?
- What implementation techniques are used within solvers?
  - What can be said about which techniques work best?
- As a group,
  - how has solver performance evolved
  - how distributed or competitive are they
- Performance:
  - Does the solver solve hard problems?
  - Does it solve easy problems quickly?

The Big Task

- Performance of all solvers on all benchmarks:
  - «all solvers»: historical and current solvers (since SMT-LIBv2)
  - «all benchmarks»: all benchmarks currently in SMT-LIB
- Year on year comparisons in the past have been muddied by differences in sets of solvers and in benchmark sets.

  - Using Star-Exec
  - Have run about ¼ of the benchmarks so far (~1 month)
  - 25 minute time-out (could go back and run the time-outs longer)
  - SMT-LIB does not yet include the 2012 benchmarks
  - Have not yet organized evaluations of application, parallel, unsat core or proof generation tracks

… about Star-Exec

- A shared logic solving infrastructure
  - A couple years in the making
  - Serves several research communities (SMT, SAT/CASC, TPTP, CoCo, HMC, ...
  - Openly available, web-service front-end
    32 compute nodes, 128GB memory, 22TB storage
    Storage for tools, benchmarks, job management
- ... This was (is) a shake-down cruise
  - Many fixes
  - Lots of work-flow and usability improvements
  - But accomplishing what we need

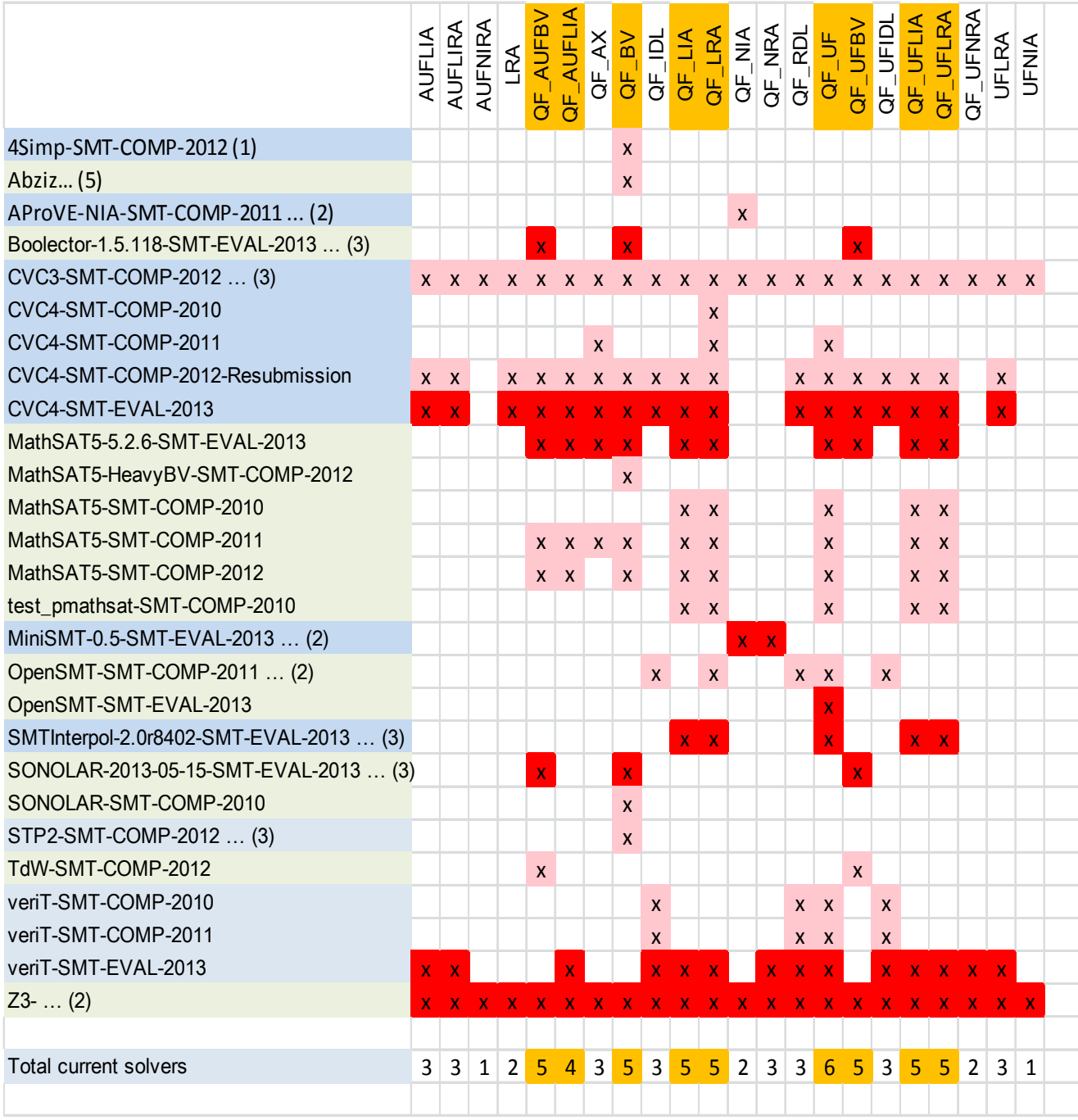
SMT Solver Participation

- All historical solvers since 2010
  - (since SMTLIBv2)
  - 32 total
- + 9 versions that were updated in 2013 + 4 experimental
- All are public (and available in StarExec):
  - Some are simply the current version of the solver
  - Some were prepared knowing this evaluation was planned

Solver	Affiliation	2005	2006	2007	2008	2009	2010	2011	2012	2013
		12	12	9	13	12	10	11	11	10
Abziz…	Cairo U.							X	X	
Boolector	JKU				X	X		X	X	X
CVC/CVCLite/CVC3	NYU	X	X	X	X	X	X	X	X	
CVC4	NYU						X	X	X	X
MathSat-HeavyBV	Trento								X	
MathSAT 3,4,5	FBK	X	X	X	X	X	X	X	X	X
SMTInterpol	U. Freiburg							X	X	X
SONOLAR	U. Bremen						X	X	X	X
STP, simplifyingSTP, STP2	U. Melbourne		X			x	X	X	X	
4Simp	U. Melbourne								X	
Tiffany de Wintermonte	U. Melbourne								X	
opensmt	U. Lugano				X	X	X	X		X
veriT	UFRN					X	X	X		X
Z3	MSR			X	X			X		X
AProVE NIA	RWTH Aachen						X	X		
MiniSMT	U. Innsbruck						X			X
test_pmathsat	FBK-IRST						X			
barcelogic	UPC	X	X	X	X	X				
beaver	UC Berkeley				X	X				
clsat	Washington U.				X	X				
Sateen	U. Col.-Boulder	X	X	X	X	X				
Spear				X	X					
sword	U. Bremen				X	X				
Yices	SRI	X	X	X	X	X				
Alt-Ergo	CNRS				X					
ArgoLib				X						
Fx7				X						
Ario		X	X							
ExtSat			X							
HTP		X	X							
Jat			X							
NuSMV			X							
Sammy		X								
SBT		X								
Simplics		X								
SVC		X								

Logic support

Dark red = current solvers



	AUFLIA	AUFLIRA	AUFNIRA	LRA	QF_AUFBV	QF_AUFLIA	QF_AX	QF_BV	QF_IDL	QF_LIA	QF_LRA	QF_NIA	QF_NRA	QF_RDL	QF_UF	QF_UFBV	QF_UFIDL	QF_UFLIA	QF_UFLRA	QF_UFNRA	UFLRA	UFNIA
4Simp-SMT-COMP-2012 (1)								x														
Abziz... (5)								x														
AProVE-NIA-SMT-COMP-2011 ... (2)												x										
Boolector-1.5.118-SMT-EVAL-2013 ... (3)					x			x								x						
CVC3-SMT-COMP-2012 ... (3)	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
CVC4-SMT-COMP-2010											x											
CVC4-SMT-COMP-2011							x				x				x							
CVC4-SMT-COMP-2012-Resubmission	x	x		x	x	x	x	x	x	x	x			x	x	x	x	x	x		x	
CVC4-SMT-EVAL-2013	x	x		x	x	x	x	x	x	x	x			x	x	x	x	x	x		x	
MathSAT5-5.2.6-SMT-EVAL-2013					x	x	x	x		x	x				x	x		x	x			
MathSAT5-HeavyBV-SMT-COMP-2012								x														
MathSAT5-SMT-COMP-2010										x	x				x			x	x			
MathSAT5-SMT-COMP-2011					x	x	x	x		x	x				x			x	x			
MathSAT5-SMT-COMP-2012					x	x		x		x	x				x			x	x			
test_pmathsat-SMT-COMP-2010										x	x				x			x	x			
MiniSMT-0.5-SMT-EVAL-2013 ... (2)												x	x									
OpenSMT-SMT-COMP-2011 ... (2)									x		x			x	x		x					
OpenSMT-SMT-EVAL-2013															x							
SMTInterpol-2.0r8402-SMT-EVAL-2013 ... (3)										x	x				x			x	x			
SONOLAR-2013-05-15-SMT-EVAL-2013 ... (3)					x			x								x						
SONOLAR-SMT-COMP-2010								x														
STP2-SMT-COMP-2012 ... (3)								x														
TdW-SMT-COMP-2012					x											x						
veriT-SMT-COMP-2010									x					x	x		x					
veriT-SMT-COMP-2011									x					x	x		x					
veriT-SMT-EVAL-2013	x	x				x			x	x	x		x	x	x			x	x	x	x	x
Z3- ... (2)	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Total current solvers	3	3	1	2	5	4	3	5	3	5	5	2	3	3	6	5	3	5	5	2	3	1

Logics & Benchmarks

Most popular logics (solvers):

- QF_BV, QF_UF, QF_UFBV
- QF_AUFBV
- QF_(A)(UF)LIA, QF_(UF)LRA

Benchmarks

- AUFLIRA, QF_AUFBV, QF_BV

Need (IMHO): Quantifier support, theory combinations

- AUFBV
- AUFNIRA

Logic	Solvers	Current	Benchmarks
AUFLIA	7	3	6402
AUFLIRA	6	3	19917
AUFNIRA	5	1	989
LRA	7	2	374
QF_AUFBV	16	5	14335
QF_AUFLIA	10	4	1140
QF_AX	10	3	551
QF_BV	22	5	31747
QF_IDL	11	3	2170
QF_LIA	15	5	5882
QF_LRA	18	5	634
QF_NIA	9	2	530
QF_NRA	7	3	166
QF_RDL	11	3	255
QF_UF	20	6	6647
QF_UFBV	14	5	31
QF_UFIDL	11	3	430
QF_UFLIA	15	5	564
QF_UFLRA	15	5	900
QF_UFNRA	5	2	26
UFLRA	7	3	5
UFNIA	5	1	1796

Everyone (well, 15/45) contributes something unique

Number of benchmarks solved only by the named solver, across all solvers, all years (within the timeout).

Even a later year's solver does not match the earlier year's accomplishment.

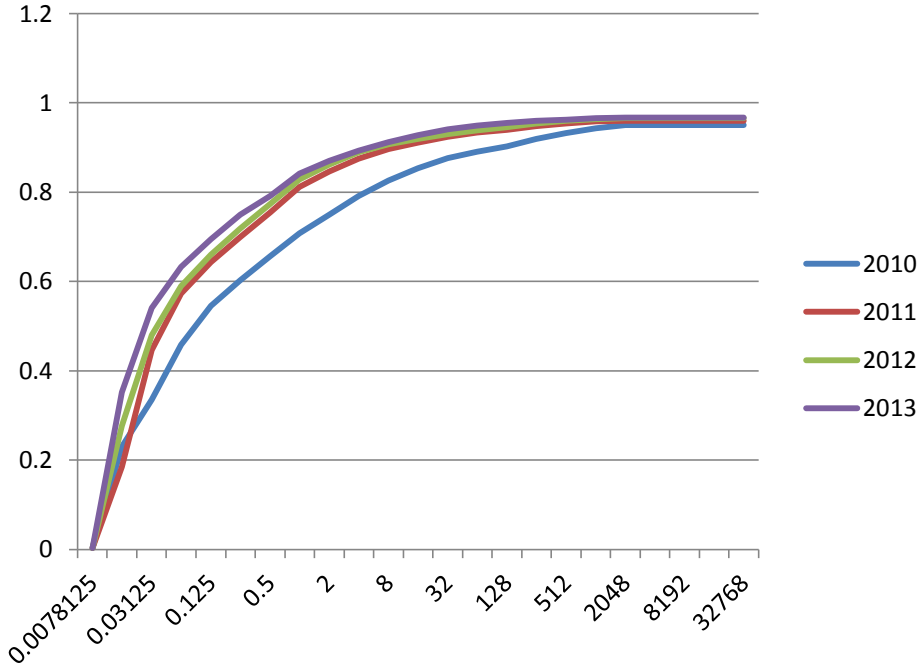
Boolector-1.5.118-SMT-EVAL-2013	1
Boolector-SMT-COMP-2011	1
CVC4-SMT-COMP-2012-Resubmission	2
CVC4-SMT-EVAL-2013	6
MathSAT5-HeavyBV-SMT-COMP-2012	1
MathSAT5-SMT-COMP-2012	1
OpenSMT-SMT-COMP-2010	1
SMTInterpol-2.0r8402-SMT-EVAL-2013	1
SMTInterpol-SMT-COMP-2011	1
SMTInterpol-SMT-COMP-2012	1
STP2-SMT-COMP-2011	3
STP2-SMT-COMP-2012	1
TdW-SMT-COMP-2012	1
Z3-4.3.2.a054b099c1d6-x64-debian-6.0.6-SMT-EVAL-2013	21
Z3-SMT-COMP-2011	15

Preliminary Results

Cumulative distribution of solution times, by year

(all solved benchmarks, all solvers, by year, as fraction of all benchmark-solver pairs for that year)

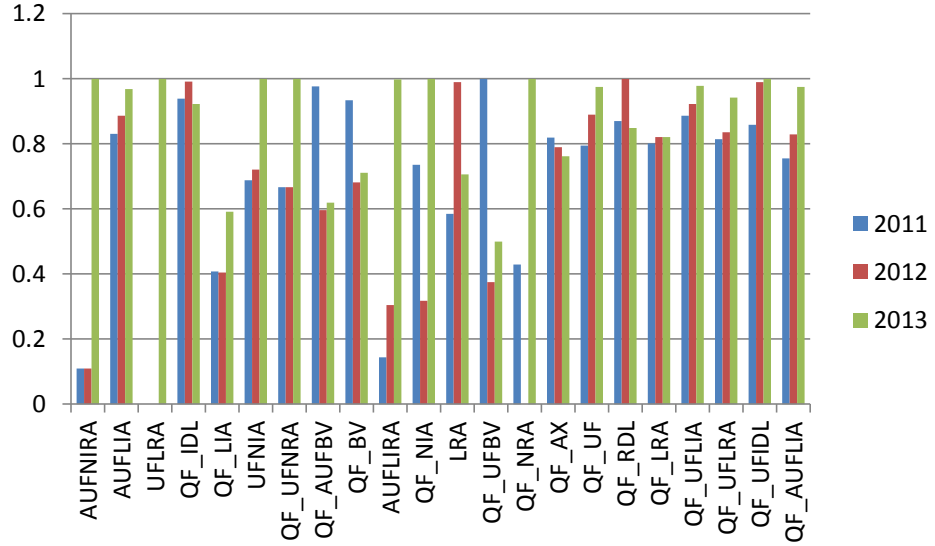
- Overall faster times (70%->84% in under 1 second)



Preliminary Results – Year to year turnover

For each benchmark of a given logic, does the solver family having the best time change from year to year?

- Turnover is typically high
- (perhaps because of a change in set of solvers)

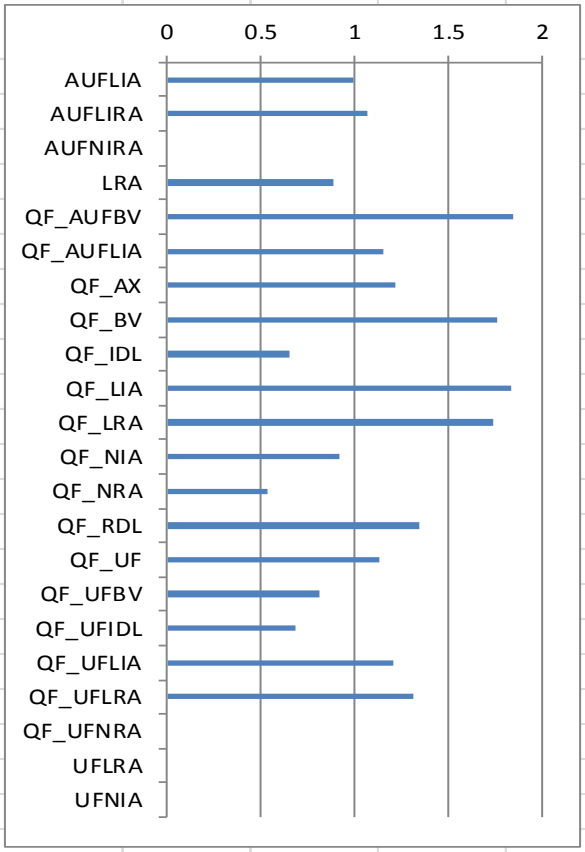


Distribution of winning solvers

Entropy:
0 = always one solver
1 = equally split between 2
2 = equally split among 4 solvers

Within a logic and year (2013), what is the distribution of solvers with best times per benchmark?

Logic	Total 2013 Solvers	Winning 2013 Solvers	Total Benchmarks	Completed Benchmarks
AUFLIA	3	3	1600	1524
AUFLIRA	3	3	4979	4976
AUFNIRA	1	1	248	248
LRA	2	2	93	92
QF_AUFBV	5	5	3584	3575
QF_AUFLIA	4	3	285	285
QF_AX	3	3	138	138
QF_BV	5	5	7936	7902
QF_IDL	3	3	543	483
QF_LIA	5	5	1470	1464
QF_LRA	5	4	159	158
QF_NIA	2	2	132	131
QF_NRA	3	3	42	42
QF_RDL	3	3	64	54
QF_UF	6	6	1661	1661
QF_UFBV	5	2	8	8
QF_UFIDL	3	3	108	106
QF_UFLIA	5	5	141	141
QF_UFLRA	5	5	225	225
QF_UFNRA	2	1	6	6
UFLRA	3	1	1	1
UFNIA	1	1	449	435



Example: QF_NRA – 42 benchmarks: 91%, 7%, 2%
QF_AUFBV – 3575: 43%, 31%, 20%, 4%, 3%

Benchmarks completed by all solvers (in 25 min)

- Results are volatile: new, under-performing solvers can bring the rate down.
- Even so, in most logics, nearly all benchmarks are solved.

	Job-pair completion rate			
	2010	2011	2012	2013
AUFNIRA	1	1	1	1
AUFLIA	0.924375	0.925938	0.922188	0.874583
UFLRA	1	1	0.5	0.333333
QF_LIA	0.917234	0.759694	0.77602	0.846939
QF_IDL	0.587477	0.689227	0.64733	0.780233
QF_UFNRA	1	1	1	0.916667
UFNIA	0.657016	0.864143	0.750557	0.96882
QF_AUFBV	0.952567	0.975893	0.98005	0.990123
QF_BV	0.968414	0.978547	0.978636	0.985761
AUFLIRA	0.998996	0.997289	0.990862	0.985941
QF_NIA	0.921717	0.840909	0.977273	0.848485
LRA	0.956989	0.844086	0.973118	0.790323
QF_UFBV	1	0.71875	0.8	0.75
QF_NRA	0.964286	0.738095	1	0.81746
QF_AX	1	1	1	1
QF_RDL	0.682292	0.734375	0.65625	0.828125
QF_UF	0.987116	0.991227	0.986454	0.988962
QF_LRA	0.927044	0.943396	0.922956	0.971069
QF_UFLIA	0.971631	0.978723	0.978723	0.974468
QF_UFIDL	0.882716	0.905093	0.837963	0.935185
QF_AUFLIA	1	0.983626	0.980117	1
QF_UFLRA	0.955556	0.907778	0.907778	1
TOTAL	0.950736	0.960576	0.965825	0.96705

Benchmarks completed by some solver (in 25 min)

- ~1000 not solved in 25 min
- Could use more difficult benchmarks

	Rate of completion by some solver			
Logic	2010	2011	2012	2013
AUFLIA	92.44%	99.19%	97.88%	95.25%
AUFLIRA	99.90%	99.90%	99.80%	99.94%
AUFNIRA	100.00%	100.00%	100.00%	100.00%
LRA	95.70%	97.85%	100.00%	98.92%
QF_AUFBV	95.26%	99.75%	99.89%	99.75%
QF_AUFLIA	100.00%	100.00%	100.00%	100.00%
QF_AX	100.00%	100.00%	100.00%	100.00%
QF_BV	99.72%	99.61%	99.62%	99.57%
QF_IDL	79.37%	88.40%	82.50%	88.95%
QF_LIA	99.12%	99.25%	98.23%	99.59%
QF_LRA	98.11%	98.74%	98.11%	99.37%
QF_NIA	100.00%	100.00%	97.73%	99.24%
QF_NRA	100.00%	100.00%	100.00%	100.00%
QF_RDL	84.38%	85.94%	82.81%	84.38%
QF_UF	99.88%	100.00%	99.94%	100.00%
QF_UFBV	100.00%	100.00%	100.00%	100.00%
QF_UFIDL	98.15%	98.15%	93.52%	98.15%
QF_UFLIA	100.00%	100.00%	100.00%	100.00%
QF_UFLRA	100.00%	100.00%	100.00%	100.00%
QF_UFNRA	100.00%	100.00%	100.00%	100.00%
UFLRA	100.00%	100.00%	100.00%	100.00%
UFNIA	65.70%	97.10%	75.06%	96.88%
TOTAL	97.41%	99.33%	98.59%	99.09%

Competitiveness

(winning time/runner up time)
(median of distribution across benchmarks for the logic)

- BV logics have times that are close.
- Many others have clear leaders
- There are some drastic changes in time that bear investigation.

Competitiveness (ratio of winning time to runner up)				
	2010	2011	2012	2013
AUFNIRA		0.67		
AUFLIA		0.38	0.68	0.57
UFLRA				
QF_LIA	0.71	0.39	0.16	0.45
QF_IDL	0.49	0.22	0.30	0.28
QF_UFNRA		0.08		0.00
UFNIA		0.33		
QF_AUFBV		0.79	0.94	0.98
QF_BV	0.48	0.90	0.97	0.89
AUFLIRA		0.73	0.79	0.90
QF_NIA	0.36	0.35		0.13
LRA		0.12	0.02	0.30
QF_UFBV		0.96	0.99	0.98
QF_NRA	0.50	0.01		0.44
QF_AX		0.80	0.67	0.74
QF_RDL	0.24	0.37	0.08	0.58
QF_UF	0.89	0.80	0.62	0.71
QF_LRA	0.85	0.80	0.71	0.81
QF_UFLIA	0.70	0.49	0.71	0.50
QF_UFIDL	0.56	0.39	0.60	0.36
QF_AUFLIA		0.53	0.56	0.92
QF_UFLRA	0.77	0.75	0.75	0.80

Questions?

Comments?