Playing with Vampire: the dark art of theorem proving

Laura Kovács and Martin Suda

Vienna University of Technology

RiSE & LogiCS Spring School on Logic and Verification

Vampire is . . .

. . . an automated theorem prover for first-order logic

Vampire is . . .

. . . an automated theorem prover for first-order logic

What is special about Vampire?

- very fast (35 trophies from CASC over the last 16 years)
- simple to use, but also many ways to configure
- detailed proofs
- limited resource strategy (LRS)
- AVATAR architecture for clause splitting
- reasoning with theories (arithmetic, arrays, . . .) and quantifiers (!)
- finite model building
- fully TPTP compliant; also understands SMTLIB2
- question answering, interpolants, consequence elimination, . . .

CASC 2015 results[1]

Higher-order Theorems	Satallax 2.8	LEO-II 1.6.2	*Satallax -* 1.3	Isabelle 2015
Solved/400	271/400	195/400	285/400	267/400
Av. CPU Time	14.96	12.25	21.67	61.02
Solutions	268/400	191/400	0/400	0/400

Higher-order Non-theorems	Nitpick 2015	Refute 2015	Satallax 2.8
Solved/200	200/200	74/200	49/200
Av. CPU Time	7.92	24.34	0.05

Typed First-order Theorems +*-/	VampireZ 1.0	CVC4 TFA-1.5	Vampire 4.0	Beagle 0.9.22	SPASS+T 2.2.22	ZenonAri 0.1.0	Princess 20150706	*CVC4 1.4-TFF*
Solved/200	172/200	163/200	160/200	131/200	108/200	60/200	143/200	131/200
Av. CPU Time	11.85	17.27	10.75	21.76	10.04	2.86	17.38	10.67
Solutions	172/200	163/200	160/200	131/200	108/200	60/200	0/200	0/200

Typed First-order Non-theorems +*-/	CVC4 TFN-1.5	Princess 20150706	Beagle 0.9.22
Solved/20	10/20	6/20	6/20
Av. CPU Time	0.00	0.97	1.33

First-order Theorems	Vampire 4.0	*Vampire 2.6*	E 1.9.1	ET 0.2	CVC4 FOF-1.5	iProver 2.0	leanCoP 2.2	iProverM 0.7-0.3	Prover9 1109a	ePrincess 1.0	Muscadet 4.5	Geo-III 2015E
Solved/400	380/400	371/400	316/400	303/400	257/400	222/400	159/400	127/400	111/400	113/400	37/400	37/400
Av. CPU Time	12.20	14.86	20.18	20.96	33.40	21.12	46.76	30.15	28.01	48.39	7.32	38.47
Solutions	374/400	368/400	316/400	303/400	256/400	217/400	159/400	127/400	111/400	88/400	37/400	37/400

First-order Non-theorems	Vampire SAT-4.0	iProver SAT-2.0	*iProver SAT-1.0*	CVC4 FNT-1.5	E FNT-1.9.1	Geo-III 2015E
Solved/200	195/200	163/200	134/200	71/200	51/200	38/200
Av. CPU Time	38.95	44.11	79.93	57.78	9.62	21.89
Solutions	195/200	163/200	134/200	71/200	51/200	38/200

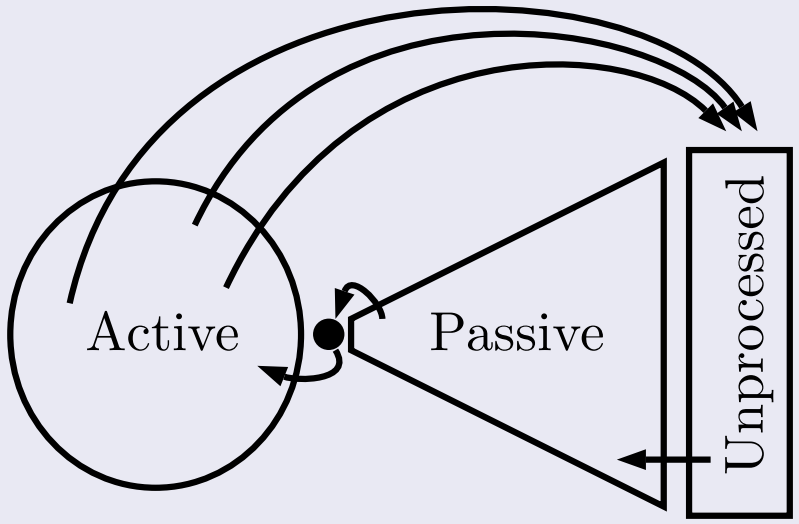
Effectively Propositional CNF	Vampire 4.0	*iProver 0.9*	iProver 2.0	E 1.9.1	Geo-III 2015E
Solved/200	192/200	161/200	153/200	101/200	9/200
Av. CPU Time	27.61	27.91	36.57	11.09	86.71

Large Theory Batch Problems	Vampire 4.0-LTB	MaLARea 0.5	E 1.9.1-LTB	iProver 2.0-LTB
Solved/1600	1208/1600	837/1600	799/1600	352/1600
Av. WC Time	6.51	8.30	7.40	13.04
Solutions	1208/1600	837/1600	799/1600	352/1600

[1] http://www.cs.miami.edu/~tptp/CASC/25/WWWFiles/DivisionSummary1.html

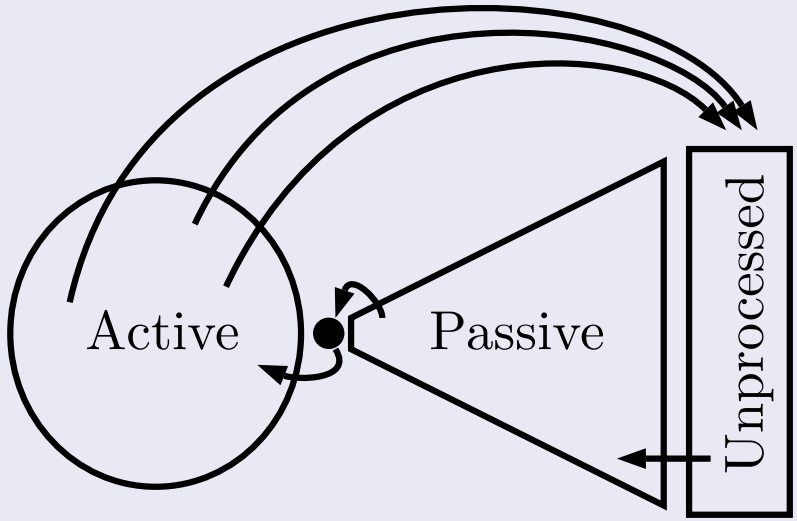
Saturation

Selecting the given clause



Saturation

Selecting the given clause

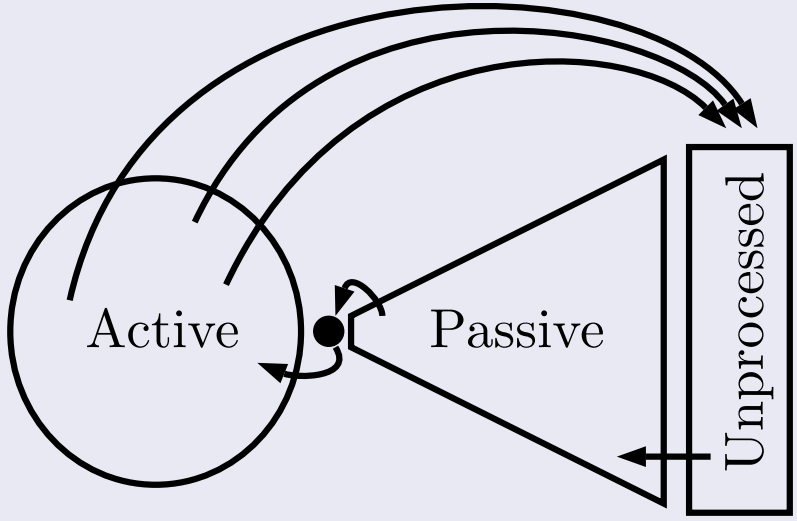


Saturation algorithms in Vampire

- the Discount loop
- the Otter loop
- Limited Resource Strategy [RV03]

Saturation

Selecting the given clause



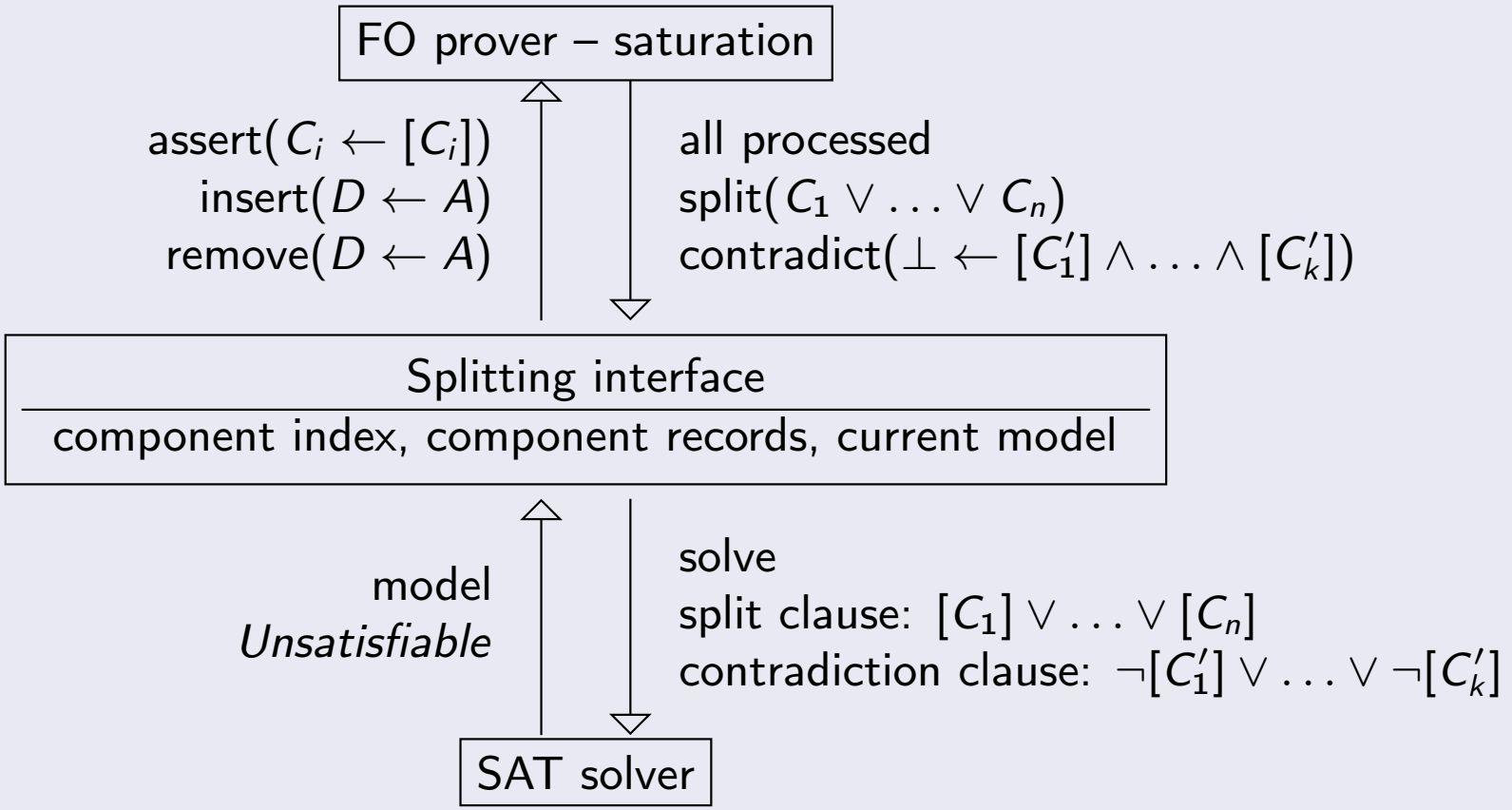
Saturation algorithms in Vampire

- the Discount loop
- the Otter loop
- Limited Resource Strategy [RV03]

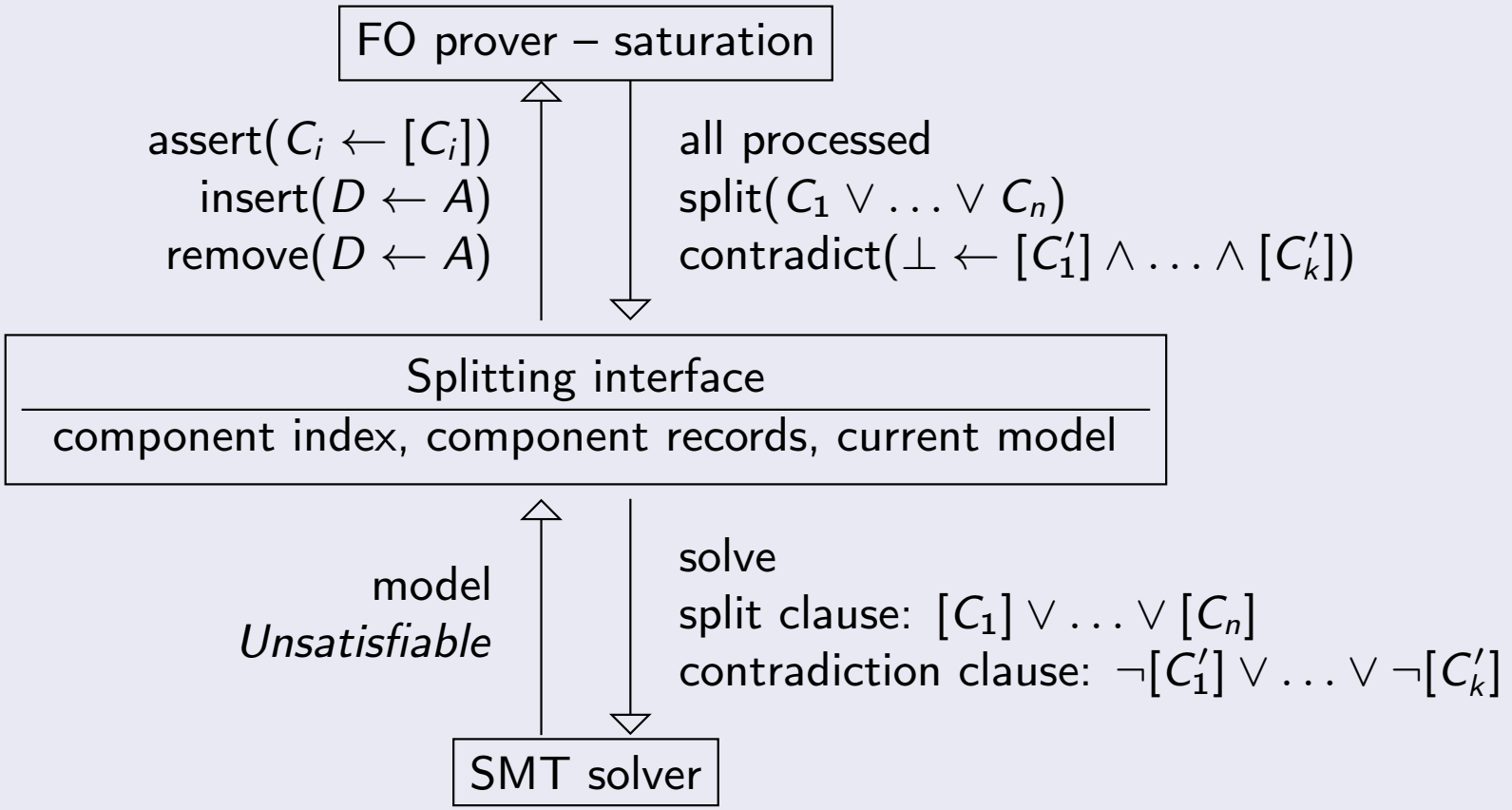
```
./vampire -awr 5:1 -fsr off Problems/GRP140-1.p
```


V

AVATAR – architecture overview



AVATAR – architecture overview



Reasoning with theories

Arithmetic reasoning in Vampire

- Evaluation of ground interpreted terms ($1 + X \longrightarrow X$)
- Theory axioms
  - hand-crafted set
  - either all added or none added (based on an option)
- VampireZ3 = AVATAR with an SMT solver
  - current implementation for Z3
  - Idea: Vampire only explores theory-consistent ground sub-problems
- ...

Reasoning with theories

Arithmetic reasoning in Vampire

- Evaluation of ground interpreted terms ($1 + X \longrightarrow X$)
- Theory axioms
  - hand-crafted set
  - either all added or none added (based on an option)
- VampireZ3 = AVATAR with an SMT solver
  - current implementation for Z3
  - Idea: Vampire only explores theory-consistent ground sub-problems
- ...

```
tff(sum_something_0_samething,conjecture,(
    ! [X: $int] :
      ( ( $less(-1,X)
        & $less(X,1) )
     => $sum(21,X) = 21 ) )).

./vampire Problems/ARI163\=1.p
```

Finite model building

MACE/Paradox-style model finding

- try looking for a finite counterexample to an invalid conjecture
- iterate model sizes $n = 1, 2, \ldots$
- pose the question "$\exists M, |M| = n \;\&\; M \models Ax \wedge \neg C$" as a SAT problem

Finite model building

MACE/Paradox-style model finding

- try looking for a finite counterexample to an invalid conjecture
- iterate model sizes $n = 1, 2, \ldots$
- pose the question "$\exists M, |M| = n$ & $M \models Ax \wedge \neg C$" as a SAT problem

```
%---- 1 * x = x
fof(left_identity,axiom, ! [X] : mult(e,X) = X).
%---- i(x) * x = 1
fof(left_inverse,axiom, ! [X] : ?[Y] : mult(Y, X) = e).
%---- (x * y) * z = x * (y * z)
fof(associativity,axiom,
    ! [X,Y,Z] : mult(mult(X,Y),Z) = mult(X,mult(Y,Z))).

./vampire Problems/grp_ord2.p
```

Dark art mini-CHALLENGE

```
./vampire --show_options on
```

Dark art mini-CHALLENGE

`./vampire --show_options on`

Some options to play with

1. Set of support (`-sos on`)
2. AVATAR turned off (`-spl off`)
   default: 8675; sploff: 8112, also 290 new
3. Discount saturation loop and the age-weight ratio
   (`-sa discount -awr 10`)
   discount only: 8519; with awr10: 8707
4. Lookahead literal selection (`-s 1011`)
   default: 8675; lookahead: 8114 but 838 new
5. Backward subsumption (`-bs on`)

Ready made solution from the Vizzard

CASC mode

- a conditional portfolio mode
- a cocktail of a strategies optimized for good general performance
- incomplete strategies in the mix; complementarity for coverage
- `--mode casc` (there is also `--mode casc_sat`)
- The schedule is 5+ minutes long (use with `-t 5m`)

Ready made solution from the Vizzard

CASC mode

- a conditional portfolio mode
- a cocktail of a strategies optimized for good general performance
- incomplete strategies in the mix; complementarity for coverage
- `--mode casc` (there is also `--mode casc_sat`)
- The schedule is 5+ minutes long (use with `-t 5m`)

A small experiment (5 minutes time limit)

TPTP 6.3.0 total:	20762	
Discarded (hol, tff, ...):	4904	
Eligible (cnf, fof):	15858	
casc:	11878	74.9 %
casc_sat:	10568	66.6 %
union:	12457	78.6 %

Leftovers

```
--mode clausify
-sa inst_gen
-stat on
--show_active on
```