ISCAS 中国科学院软件研究所学术年会'2021
暨计算机科学国家重点实验室开放周

关键技术

基于自动机理论的字符串约束判定算法

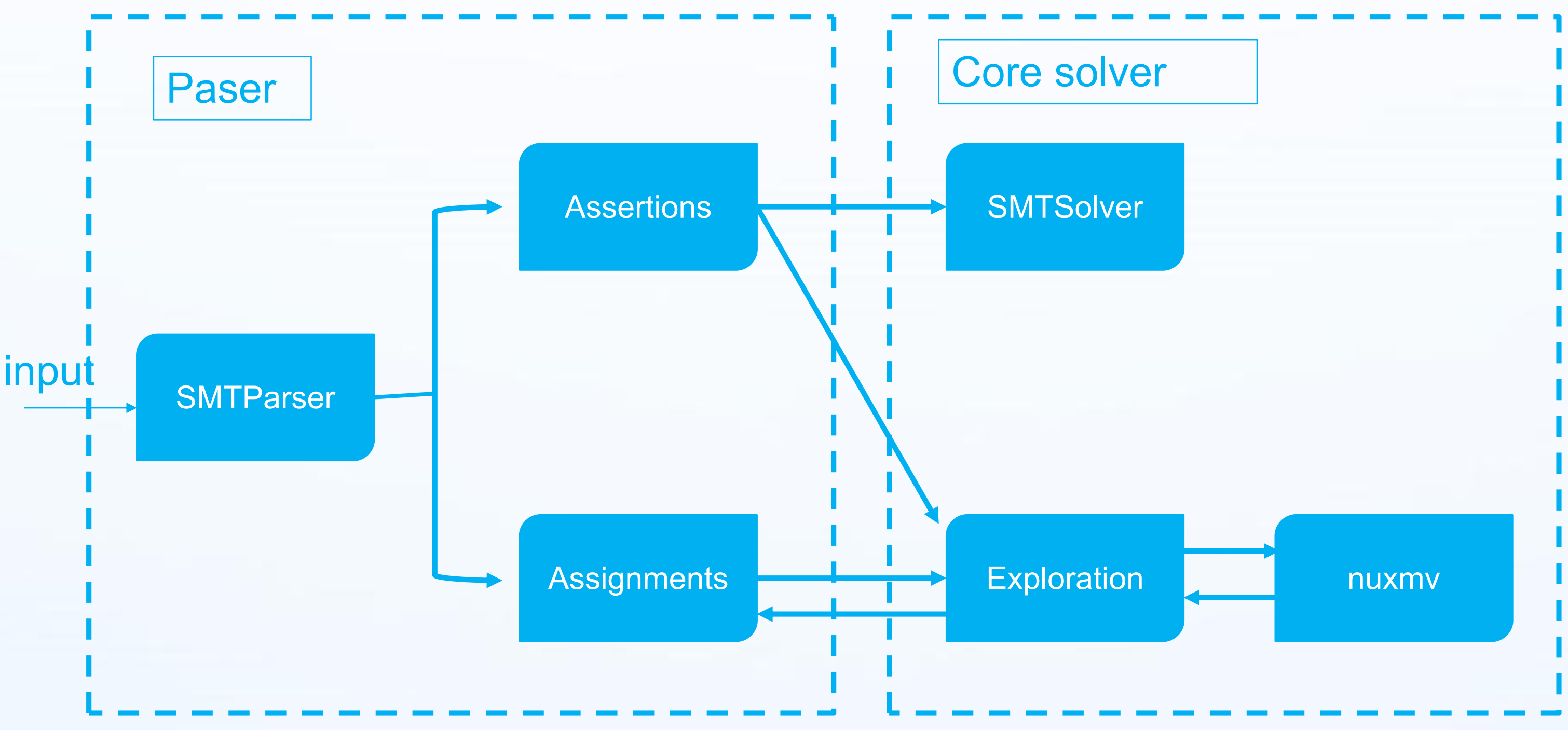
吴志林，陈韬略，Alejandro Flores Lamas, Matthew Hague，陈艳, 何锦龙,
胡登杭，韩志磊，Anthony W. Lin, 阚双龙，Philipp Ruemmer
联系方式：吴志林，13810643403，wuzl@ios.ac.cn

关键技术介绍

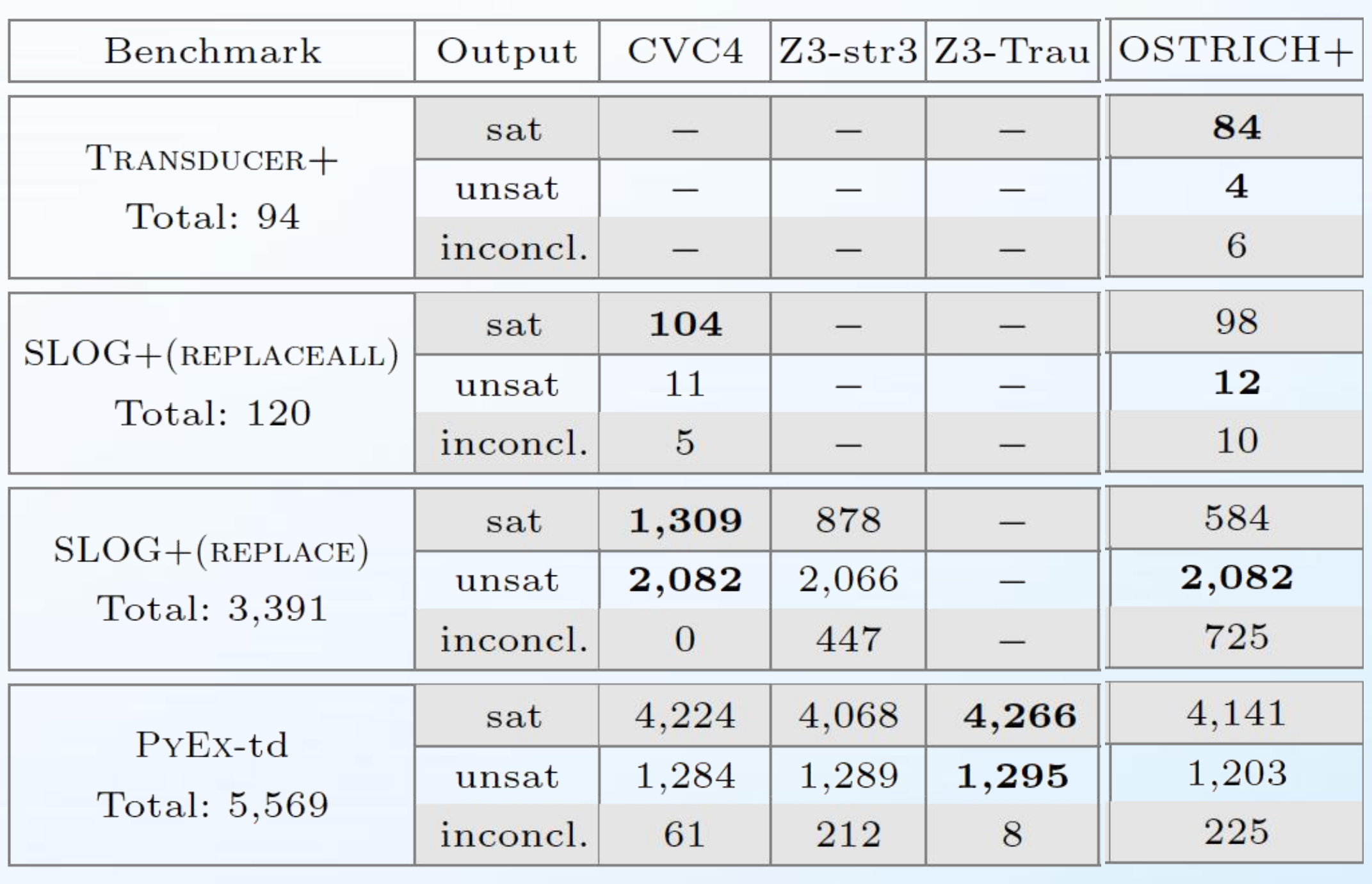
- **字符串是几乎每一种编程语言都支持的数据类型，其在Web应用、Android应用等中被广泛使用。字符串约束求解是对字符串程序进行分析与验证的基础。**
- **由于字符串约束求解一般来说不可判定，目前的字符串约束求解器基本使用启发式算法进行求解。**
- **我们针对字符串程序的符号执行，考虑了字符串约束的直线子集，基于自动机理论提出了字符串约束直线子集可判定的语义条件，并对满足语义条件的字符串约束提出了一般性的判定算法，实现了字符串求解器OSTRICH (https://github.com/uuverifiers/ostrich)。**

字符串约束	$\varphi \quad ::= \quad z = x \circ y \mid z = \mathit{replaceAll}_e(x, y) \mid y = \mathit{reverse}(x) \mid$ $y = T(x) \mid x \in e \mid \varphi \wedge \varphi$ 这里*e*是正则表达式，*T*是一个有限状态转换器
可判定语义条件	字符串为直线子集（即变量之间不存在相互依赖的子集），且每个字符串操作*f*都满足: 对于每个正则语言*L*, $f^{-1}(L)$，即*L*关于*f*的前象，是一个可识别的关系(recognizable relations)，且可以从*L*能行地计算出来。
判定算法	对于每一个等式 $y = f(\vec{x})$ 和正则约束 $y \in L$，计算 $f^{-1}(L) = \bigcup_{i}^{n} L_{i,1} \times \cdots \times L_{i,k}$ 选取*i*, 添加正则约束 $x_1 \in L_{i,1} \wedge \ldots \wedge x_k \in L_{i,k}$，然后去掉该等式 最终得到一个不含有等式的正则约束的合取进行求解

该判定算法可以扩展到支持**整数数据类型**(length, substring, indexof)和
编程语言中的正则表达式特性(比如greedy/lazy Kleene star/plus和capturing groups)



OSTRICH字符串约束求解器架构



Benchmark	Output	CVC4	Z3-str3	Z3-Trau	OSTRICH+
TRANSDUCER+ Total: 94	sat	–	–	–	**84**
	unsat	–	–	–	**4**
	inconcl.	–	–	–	6
SLOG+(REPLACEALL) Total: 120	sat	**104**	–	–	98
	unsat	11	–	–	**12**
	inconcl.	5	–	–	10
SLOG+(REPLACE) Total: 3,391	sat	**1,309**	878	–	584
	unsat	**2,082**	2,066	–	**2,082**
	inconcl.	0	447	–	725
PYEX-td Total: 5,569	sat	4,224	4,068	**4,266**	4,141
	unsat	1,284	1,289	**1,295**	1,203
	inconcl.	61	212	8	225

部分实验结果：能够在30s之内求解的测试用例个数，横线表示求解器不支持该种类型的约束，粗体表示最好结果

更多细节请参考我们的POPL 18、POPL 19、ATVA 20文章

技术指标

- 支持以下字符串操作：
  - ❖ *concatenation*
  - ❖ $\mathit{extract}_{i,e}(x)$: 提取*e*中第*i*个capturing group的值
  - ❖ $\mathit{replaceAll}_e(x,\ y)$
  - ❖ *reverse*
  - ❖ *finite transducers*
  - ❖ *length*
  - ❖ $\mathit{substring}(x,\ i,\ j)$,
  - ❖ $\mathit{charAt}(x,\ i)$
  - ❖ $\mathit{indexOf}_u\ (x,\ i)$
  - ❖ *regular constraints* $x \in e$
- 平均求解时间：每个测试用例10s
- 30s之内可求解的字符串约束最大规模（文件大小）：18KB

标志性技术进步

- 统一和扩展了目前已有的字符串约束可判定子集；
- 可高效实现的字符串约束判定算法；
- 支持字符串操作最多的字符串约束求解器
- 第一个支持编程语言正则表达式特性的字符串约束求解器

可应用领域

- Javascript程序的符号执行
- 跨站点脚本攻击（XSS）检测
- 针对正则表达式的DOS（Denial of Service）攻击检测