****



**研 究 生 毕 业 论 文**

**（申请硕士学位）**

**论文题目** 循环不变式生成方法研究

**作者姓名** 刘自恒

**学科、专业名称**  计算机软件与理论

**研 究 方 向**  信息安全

**指导教师** 曾庆凯 教授

二零一二年 五月 二十日

**学 号：MG0933021**

**论文答辩日期： 年 月 日**

**指 导 教 师： （签字）**

**Nanjing University**

(Master Degree Dissertation)

**Research on Static Software Security**

**Flaw Detection**

By

**Zi-heng Liu**

Supervised by

Professor **Qing-kai Zeng**

Department of Computer Science and Technology

Nanjing University

May 2012

Nanjing, P.R.China

# 摘要

软件作为信息系统的实现载体，广泛应用在各个领域，软件中的任何安全漏洞或错误的实现都可能导致非常严重的后果。通过大量的测试可以提高软件的可靠性，但成本较高，也不能完全保证软件的可靠。形式化证明方法基于严密的数学和逻辑基础，经过正确性证明的程序可以保证软件符合指定的程序规范，从而大大提高软件可靠性。

程序正确性证明的一个关键问题是发现充分的循环不变式来辅助、支撑证明过程的完成。由人工寻找循环不变式工作繁琐，容易出错。因此，研究如何自动地有效分析和生成循环不变式，使得软件证明过程顺利完成，成为形式化方法研究领域的一个重要研究课题。本文主要针对软件验证技术中的循环不变式生成技术进行研究和实现：

（1）对形式化方法主要理论进行总结。本文根据形式化用到的主要理论的不同对形式化技术进行总结。对各自的理论核心、主要解决的问题、典型工具进行了阐述。在此基础上，指出形式化证明过程的难点。

（2）介绍和总结了目前求解循环不变式的主要方法。根据他们解决问题范围的不同进行阐述，指出了目前工作存在的不足。

（3）本文提出了一种基于条件赋值转换和自适应模板技术的循环不变式生成方法。根据程序控制流结构、依赖分析结果和值分析结果，生成可能的循环不变式，然后由定理证明器求解。该方法可以自动地运行，并将结果以注释的形式添加到程序中，辅助完成程序属性的证明过程。

（4）基于上述方法，在Frama-C平台和APRON库的基础上，实现了一个分析插件。选取了一些程序作为分析对象，与其他方法作了比较，结果表明可以发现更多的不变式，使一些程序规范顺利得到证明。

**关键字：验证、程序规范、循环不变式、自动化、插件**

# Abstract

Software has been used widely in various areas as the carrier of information systems and any vulnerability or incorrect development in software will cause highly serious failures. It can increase the confidence with enough testing which will be very expensive. But unfortunately it still cannot guarantee that the software is fully reliability. Formal methods are based on strict mathematically and logically basis. It can ensure that the software meets its given specifications through proving of correctness, which greatly improves the reliability of software.

A key issue in proving is finding sufficient loop invariants to assist the automatic program proving. However, writing loop invariants manually is boring and error-prone. Thus, the research on how to analysis and generate loop invariants automatically becomes an important issue in the area of formal methods. This dissertation focused on the generation of loop invariants and mainly covers the subjects of theory of formal methods and generation technologies for invariants.

(1)We summarized the main theories of formal methods. It is focused on the core theory, the main problem to resolve and classical tools. Then we point out the difficulty of formal proof.

(2) We summarized the methods for generating loop invariants, including the kind of program to be analyzed, the way how they work and the shortcomings.

(3) In this thesis, we propose an approach based on conditional assignment and adaptive template. It generates potential loop invariants according to CFG, PDG and value analysis which are then resolved by theory prover. It can run automatically and insert invariants as ACSL comments.

(4)Based on the method above, we designed and implemented an plugin named loopInv with Frama-C and APRON. According to the experiments on some programs, it can generate more invariants comparing with other tools.

**Keywords：verify, program specification, loop invariant, value analysis, plugin**

# 目录

[摘要 I](#_Toc318212116)

[Abstract II](#_Toc318212117)

[目录 III](#_Toc318212118)

[图片目录 IV](#_Toc318212119)

[表格目录 IV](#_Toc318212120)

[第一章 引言 1](#_Toc318212121)

[1.1 研究背景 1](#_Toc318212122)

[1.2 研究内容 2](#_Toc318212123)

[1.3 本文组织 3](#_Toc318212124)

[第二章 形式化方法基础 4](#_Toc318212125)

[第三章 循环不变式生成技术研究 6](#_Toc318212126)

[第四章 循环不变式生成技术改进 7](#_Toc318212127)

[第五章 loopInv插件的设计与实现 7](#_Toc318212128)

[第六章 实验过程与结果 9](#_Toc318212129)

[第七章 结论与展望 9](#_Toc318212130)

[参考文献 9](#_Toc318212131)

[攻读硕士学位期间参与的科研项目 11](#_Toc318212132)

[攻读硕士学位期间完成的论文 11](#_Toc318212133)

[致谢 11](#_Toc318212134)

# 图片目录

[**图 1.1 CVE 数据库近10年来的脆弱性数量统计** 1](#_Toc317862373)

[**图 5.2 Frama-C结构图** 8](#_Toc317862374)

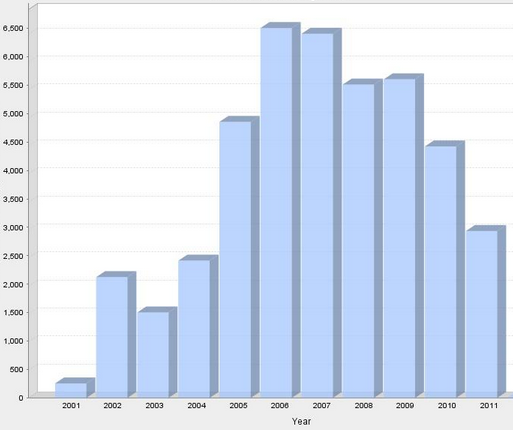
# 表格目录

[**表 6.1 比较** 9](#_Toc317860539)

# 第一章 引言

## 研究背景

随着计算机科学的迅速发展和信息化时代的深刻变革，信息系统已经深入到人类生产生活的各个领域，软件作为信息系统的实现载体，广泛应用在各个领域，软件的任何漏洞都可能导致非常严重的后果。无论是软件设计人员还是软件的用户，均希望在软件系统投入正式使用之前，能得到软件正确性的保证，避免造成软件后期维护成本的上升和使用有缺陷软件造成的损失。软件的安全性一直是系统安全性、可靠性、稳定性的一个重要方面。已经有很多系统由于软件缺陷的存在，造成了巨大的损失，CVE（[cve.mitre.org](http://cve.mitre.org/)）不断发布最新的软件安全漏洞报告，由图 1.1（[nvd.nist.gov](http://nvd.nist.gov/)对CVE数据的统计）可以看出，软件的脆弱性形势一直十分严峻，软件的安全性和可靠性一直困扰着软件开发和使用人员。



**图 1.1 CVE 数据库近10年来的脆弱性数量统计**

另一方面，软件的实现是否和设计需求和规范是否一致决定了软件的正确性，不正确的软件开发实现同样会带来重大损失。

首先用来试图得到无错误软件的方法是测试；通过在程序上运行一组输入集合并检查软件的行为是否符合预期；对于关键性软件，必须达到很高的测试覆盖度才可以获得相对较高的可信度，但是这样的测试输入有时候甚至是无穷的，因而代价很高[[Mauborgne 2004](#_ENREF_9)]。

为了最大限度地保证软件的正确性和无脆弱性，从另一个角度考虑，如果可以从理论上证明程序符合预期的行为或者没有某种错误，同样可以认为软件是可靠的。这需要开发人员对已编码软件的正确性进行分析证明，同时证明软件中没有某些导致软件安全缺陷的代码实现。正确性是指代码的实现和设计的算法目标一致，代码的执行与算法的预期是相同的，代码没有错误地实现算法。没有安全缺陷是指代码的实现过程没有引入可能导致软件崩溃、被其他恶意软件利用篡改等的部分。

证明的过程需要对算法、编码和安全漏洞有着深刻的认识，需要严密的逻辑演算过程，同时也会有一些琐碎的证明过程。这样的过程如果全部由人工来完成，无疑需要大量的人力、物力、财力，而且人工实施难免会出差错。

最早从50年代开始出现了形式化方法（formal methods）的程序设计思想，即J.Backus提出BNF描述Algol60语言的语法，以严格的数学和逻辑理论为基础，适合于软件和硬件系统的描述、开发和验证。经过多年的研究，取得了大量、重要的成果，逐渐融入软件开发周期的各个阶段。在编码后期的检查阶段，使用形式化方法可以自动地帮助分析人员发现其它方法不容易发现的系统描述不一致性或脆弱性，增加软件开发人员对系统的理解，实现开发可靠的软件产品的目标。形式化方法并不排斥测试，但可以认为是一种更好的测试框架[[Airchinnigh 1995](#_ENREF_1)]。形式方法的主要研究内容包括:形式规范说明 (formal specification)和形式验证(formal verification)。基于模型的；基于公理系统和过程代数的

定理证明 (theorem proving)是一种形式验证方法，它是根据已构造的规格说明生成反映该规格说明应具有的性质，将其表示成定理形式，并加以证明，从而达到对系统规格说明验证的目的。形式化的方式主要思路是对程序源代码、程序规范说明进行建模、分析，以判断程序是否符合程序规范的要求，并对不符合要求的程序给出反例路径，供开发人员分析修正。这一过程可以不需要或较少需要人工干预，因而自动化程度高。

## 研究内容

本文主要包括研究内容如下

（1）阐述了形式化方法的主要理论基础，总结了现有的形式化方法的主要种类。

主要从前置条件、后置条件、不变式、状态机等方面阐述，比较了常用的形式化方法的特点、适用范围、典型工具、局限等内容。

（2）分析、总结、比较了现有的生成循环不变式的方法。

本文列出了目前使用的方法的针对情形，指出了用到的关键算法，指出了各自的特点和不足，并提出了可能的解决方法。

（3）针对数值型计算程序和含有数组的程序，基于条件赋值转换和自适应模板生成技术，提出了改进的循环不变式生成方法。

（4）基于Frama-C和APRON库，将上述方法实现为Frama-C的一个插件loopInv。使用该插件可以自动地完成循环不变式的生成和插桩工作，辅助后续的证明过程。

（5）选取若干样例程序，与其他不变式生成工具做对比。

选取了若干简单示例程序和实际部署使用的程序，计算其中存在的不变式，与其他工具对比，并人工统计分析生成的不变式的质量。

## 本文组织

本文的组织结构如下：

第二章阐述了形式化方法的主要理论基础，给出了目前常用的形式化方法，

第三章探讨生成循环不变式的方法，给出了目前常用的形式化方法的理论核心、主要解决的问题、典型工具。

第四章提出了一种自动生成循环不变式的方法，该方法基于条件赋值转换和自适应模板生成技术，自动化程度高，生成尽可能多的不变式。

第五章基于Frama-C和APRON库，将上述方法加以实现，成为Frama-C的一个插件。

简要介绍了Frama-C的主要功能和特点，介绍了APRON库的作用和效果。遵从Frama-C的开发要求，设计了一个Frama-C的插件。描述了该插件的主要组成、工作原理和实现的主要算法。

第六章选取了一些样例程序，将上述方法和现有方法的效果加以对比；选取了一些实际中使用的程序，对生成的结果进行分析，评价生成的不变式的质量。

第七章总结了目前的工作，并对下一步的研究方向进行了展望。

# 第二章 形式化方法基础

软件开发自动化技术是提高软件生产率的根本途径之一。软件自动化的前提是形式化，形式化不仅仅是对用户需求，而且也是对整个软件系统的严格定义。传统的软件开发方法由于大量的使用自然语言和多种图形符号,结果是尽管经历了仔细地复审，最后的系统规约说明中仍然包含歧义的、含糊的、矛盾的、不完整的需求描述及混乱的抽象层次。使用形式化方法可以克服这些缺点，有理论优势。本章首先以形式化的方式定义相关概念和操作，然后阐述形式化分析方法的理论，阐述理论核心、主要解决的问题和典型工具。

符号约定

程序P使用抽象语法树（Abstract Syntax Tree，AST）[[V.Aho, S.Lam et al. 2007](#_ENREF_12)]表示，是[源代码](http://zh.wikipedia.org/wiki/%E6%BA%90%E4%BB%A3%E7%A0%81)的抽象[语法](http://zh.wikipedia.org/wiki/%E8%AF%AD%E6%B3%95%E5%AD%A6)结构的[树](http://zh.wikipedia.org/wiki/%E6%A0%91_(%E5%9B%BE%E8%AE%BA))状表现形式。

控制流图（Control Flow Graph, CFG）一个程序（函数）的控制流图可定义为



其中，是程序控制点集合，是程序变量集合，是边的集合，是初始位置，是终止位置，每条边由一个转换关系标记，是程序变量的当前状态，是程序变量的下一状态，代表程序变量的初始值。程序从初始位置以初始值开始执行，到达终止位置时结束。程序P的控制流图可以从其抽象语法树计算得到。

是一个简单的程序，它的CFG如所示。



void phase()

{

int x=0,y=10;

while(x<=y)

{

x++;

}

}

转换系统

程序规范

最弱前置条件：保证断言有效，为安全属性、终止生成最通用反例

最强后置条件：得到程序的可达状态集，当没有断言时不能保证精度。

不变式

归纳不变式

形式化的程序分析主要由程序规范说明、程序和属性建模、模型分析过程几个阶段组成。

形式化证明的主要方法有。模型检测。定理证明。谓词抽象。

嵌套循环，多路径程序，递归程序。

形式化证明的一个关键难点在于程序中循环结构的存在。数组的存在。归纳不变式。目前对循环的处理方法主要有。

循环展开

综上，循环不变式是循环的后置条件的一种弱化形式[[Furia and Meyer 2010](#_ENREF_6)]。某个程序规范未得到证明，不一定就是因为程序和规范不一定，而是因为证明需要的信息不充分，没有给出足够的程序中包含的循环不变式。这表明，循环不变式的重要性[[Moy and Marché 2010](#_ENREF_10)]

# 第三章 循环不变式生成技术研究

上一章阐述了形式化方法的基础理论，并指出程序证明过程的一个关键问题是寻找循环结构中存在的不变式。本章将从几个具体程序出发，根据其具体特点，分析了现有的生成循环不变式的方法，指出了它们的优缺点。在这些现有方法的基础上，有针对性地提出了改进目标。

基于迭代的不动点计算：数据流分析，模型检测，抽象解释等

基于约束的不变式生成：目标指向性明确，效率高。不需要放大，精确。[[刘自恒报告20110301, Gulwani et al. 2008](#_ENREF_13)]

针对线性算术式的不变式模板求系数法；[[刘自恒报告20110301, Gulwani et al. 2008](#_ENREF_13); [Gupta and Rybalchenko 2009](#_ENREF_8)]由用户指定不变式模版（程序有默认的可以使用），是基于数值约束求解[[Colon, Sankaranarayanan et al. 2003](#_ENREF_3)]的。对于此，用户指定的模板形式固定，适应性不强，可以通过要证明的规范或断言，启发式地得到可能的不变式模板形式，然后逐一进行约束求解。约束求解库可以使用APRON，定义的抽象域包括Box,Oct，NewPolka，PPL等，不同的抽象域求解的效率和精度不同。[[Gupta and Rybalchenko 2009](#_ENREF_8)]将自己集成到Blast中，使Blast的分析效果得到提升。

Interproc使用APRON和fixpoint库实现了一个简单语言程序的不变式生成，fixpoint是按照[[Gopan and Reps 2007](#_ENREF_7)]中的思想实现的，不同的转换系统实例可以得到不同的分析结果，[[Gopan and Reps 2007](#_ENREF_7)]中给出了两个转换系统实例。可以将Interproc看作一个抽象解释器，invGen即在前端使用了Interproc作为解释器，并跟踪它生成的不变式。

多阶段循环的展开法(循环展开，控制流精化)或条件赋值法；当循环体中有if语句时，程序的路径有多条，由于循环控制变量的变化，可以把循环看成一个多阶段的执行过程。可以将循环体做语句展开，做执行路径的具体化，或者将if控制下的语句转换为条件赋值语句[[刘自恒报告20110426, Kovács et al. 2009](#_ENREF_14)]，或者通过分割谓词将一个循环展开为多个基本形式的循环[[Sharma, Dillig et al. 2011](#_ENREF_11)]。[[Beyer, Henzinger et al. 2007](#_ENREF_2)]将这样的循环连同其他部分看做多个路径程序分别分析。将if条件直接转换到不变式中，同时做相应的变量转换，类似[[Furia and Meyer 2010](#_ENREF_6)]，之前黄老师说不加改变的是无意义的，因为在推理的时候，推理证明器可以根据if条件进行推理。

将数组根据状态划分为小数组的分割法；有数组参与时，可以将所有数组元素用互不同的变量代替，也可以用一个变量代替所有元素，或根据循环的执行进程，将数组分为不同状态的子数组[[刘自恒报告20110426, Kovács et al. 2009](#_ENREF_14); [Cousot, Cousot et al. 2011](#_ENREF_4)]，根据状态得到数组变量的不变式性质。有字符串操作的语句也可以看成对数组进行操作，一般通过系统函数如memcpy等完成，这时候可以对系统函数加上规范说明得到关于字符数组的不变式性质。

[[Flanagan and Qadeer 2002](#_ENREF_5)]使用了预定义原子谓词来辅助程序证明过程，使用谓词抽象技术猜测后面可能使用到的谓词。

# 第四章 循环不变式生成技术改进

# 第五章 loopInv插件的设计与实现

本章介绍第四章设计的循环不变式生成技术的具体实现。该实现是基于Frama-C平台和APRON库的，作为Frama-C的一个插件而集成到Frama-C平台中，命名为loopInv，并提供编程接口供其他开发使用。本章首先简要介绍相关的Frama-C平台和APRON库的相关内容，然后给出loopInv插件的结构组成和处理流程。

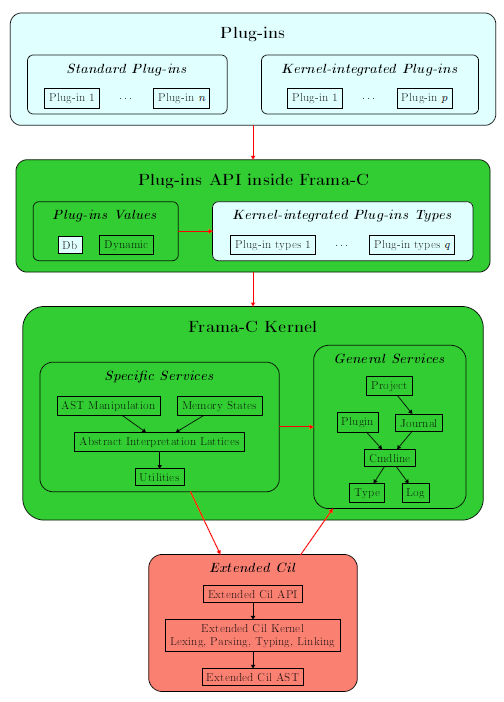
ACSL

ACSL的全称是ANSI C Specification Language，是一种在Frama-C框架下实现的行为接口类规范描述语言（Behavioral Interface Specification Language，BISL），用来描述C语言代码的行为属性。不同于JML（Java Modeling Language）既用于动态检测断言又用于静态验证，ACSL用于静态验证和归纳验证。

更多内容见<http://frama-c.com/acsl.html>。

Frama-C

Frama-C是一个执行C程序静态分析的开发平台工具，收集了多种静态分析技术，将它们作为插件整合在一个统一的框架中，并对外提供开发接口，允许用户使用既有插件的结果快速开发自己的功能插件。**图 5.2**是Frama-C的结构设计图。



**图 5.2 Frama-C结构图**

APRON

获取循环中可能的变量值更新变化的步长集合

获取循环中更新变量值的常数集合

获取循环之后程序中存在的判断表达式集合Adj

提取Adj中的变量集合。通过插件PDG进行依赖关系计算，得到循环中使用每个变量的结点集合。提取其中的常数，作为步长候选值。

根据循环中用的变量，生成模板，使用wp验证是否满足，不确定时加入转换系统，由APRON再次检查。

成立的不变式用ACSL注释表示，加入程序中，输出得到分析后的含有循环不变式的程序。使用WHY等工具来验证程序中的属性是否可以得到满足。

# 第六章 实验过程与结果

**表 6.1 比较**

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

# 第七章 结论与展望

由于理论尚不够成熟，形式化方法还难以证明大型软件系统的正确性[[Airchinnigh 1995](#_ENREF_1)]。希望我们的工作有助于形式化证明的发展。

# 参考文献

Airchinnigh, M. a. (1995). Formal Methods & Testing. Tutorials of the Sixth International Software Quality Week, 625 Third Street, San Francisco, CA 94107-1997.

Beyer, D., T. A. Henzinger, et al. (2007). Path invariants. PLDI '07, ACM New York, NY, USA ©2007.

Colon, M., S. Sankaranarayanan, et al. (2003). Linear invariant generation using non-linear constraint solving. CAV'03.

Cousot, P., R. Cousot, et al. (2011). A parametric segmentation functor for fully automatic and scalable array content analysis. POPL'11, ACM.

Flanagan, C. and S. Qadeer (2002). Predicate Abstraction for Software Verification. POPL '02.

Furia, C. A. and B. Meyer (2010). Inferring Loop Invariants Using Postconditions, Springer-Verlag Berlin, Heidelberg ©2010

Gopan, D. and h. Reps (2007). Guided Static Analysis. 14th International Static Analysis Symposium, Kongens Lyngby, Denmark.

Gupta, A. and A. Rybalchenko (2009). InvGen: An Efficient Invariant Generator. CAV '09.

Mauborgne, L. (2004). Astrée: verification of absence of run-time error. Building the Information Society, Kluwer Academic.

Moy, Y. and C. Marché (2010). "Modular inference of subprogram contracts for safety checking☆." Journal Of Symbolic Computation **45**(11): 1184-1211.

Sharma, R., I. Dillig, et al. (2011). Simplifying Loop Invariant Generation Using Splitter Predicates. CAV'11.

V.Aho, A., M. S.Lam, et al. (2007). Compilers Principles Techniques and Tools. America, Addison Wesley.

刘自恒报告20110301, S. Gulwani, et al. (2008). Program analysis as constraint solving. PLDI '08, Tucson, Arizona, USA, ACM New York, NY, USA ©2008.

刘自恒报告20110426, L. Kovács, et al. (2009). Finding Loop Invariants for Programs over Arrays Using a Theorem Prover. FASE 2009, Springer-Verlag Berlin Heidelberg 2009.

# 攻读硕士学位期间参与的科研项目

# 攻读硕士学位期间完成的论文

# 致谢