

国内编译技术进展

董渊 2010

dongyuan@tsinghua.edu.cn

清华大学

计算机科学与技术系

董渊, 冯晓兵*, 王生原, 陈文光, 编译技术年度进展报告, 中国计算机学会, 2010。

* 中国科学院计算技术研究所

报告提纲

- 一. 关于我们
- 二. 背景介绍
- 三. 编译技术面临的问题
- 四. 前沿研究进展
- 五. 开源平台
- 六. 产品编译系统
- 七. 大学教育
- 八. 几点建议
- 九. 致谢
- 十. Q&A

一、关于我们：计算机科学与技术系



清华大学计算机科学与技术系
Department of Computer Science and Technology, Tsinghua University

2010年09月17日 星期五 您所在的时间: 科 研

计算机研究所

计算机研究所是国家二级重点学科单位, 研究方向包括智能工程、知识工程、计算机与VLSI设计自动化、可视化技术与计算机图形学、软件工程及系统软件等研究领域。

目前, 有教授、副教授等研究人员16名, 其中: 教授6名, 博士生导师6名, 副教授6名, 讲师0名, 青博士2名, 硕士生47名, 博士生50名。

在研的留学973、863、自然科学基金及国际合作等重大项目近50项, 每年在国内外学术杂志、学术会议上发表论文百余篇, 承担20余门计算机系本科生和研究生的课程教学。

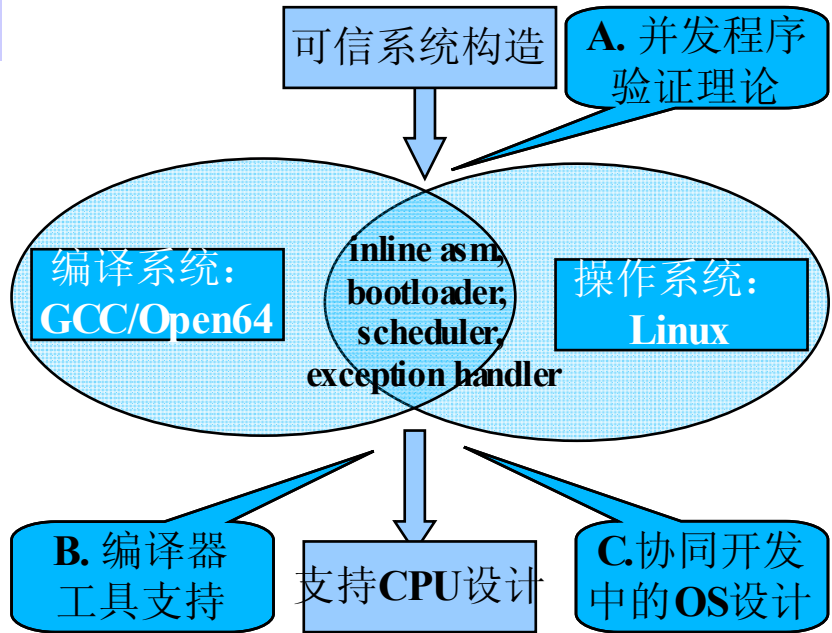
研究所的主要研究方向有:

- 数据工程**
研究内容包括数据压缩、XML数据库管理系统、海量数据处理技术、Web数据整理、基于场景的数据挖掘、个性化信息定制服务、半结构化文本信息抽取机制、数字电子装备的件库的设计与实现、电子政务。
- 知识工程**
研究内容包括智能推理模式下知识推理机制的研究, 包括语义Web和语义Web服务、本文与社会网络挖掘、基于XML的推理处理技术关键技术。
- 计算机与VLSI设计自动化**
研究内容包括系统级描述及高层次综合、物理版图设计自动化、设计正确性验证及寄生参数提取、VLSI计算机辅助设计软件系统等。
- 可视化技术与计算机图形学**
研究内容包括数字影像数据可视化、地形数据可视化、基于图形的建模与绘制、基于图形的光照与遮挡控制、小波及其在信号处理、图像图像中的应用、计算几何。
- 软件工程与系统软件**
研究内容包括测试自动化技术、模型驱动的开发与测试、面向对象计算、嵌入式系统软件、交叉编译系统、编译优化技术等。

Copyright ©2009-2010 清华大学计算机科学与技术系 All Rights Reserved

一、关于我们：个人科研重点

基础软件 核心代码



二、背景介绍：Unix后门难题

- 汤普森攻击(Thompson hack)
 - Unix/C之父，图灵奖/美国国家技术奖获得者Ken Thompson
 - 1983年图灵奖获奖报告中揭开谜底
 - 在编译过程中恶意注入，在Unix系统中潜伏~15年

Awards


[\[edit\]](#)

Turing Award

[\[edit\]](#)

In 1983, Thompson and Ritchie jointly received the [Turing Award](#) for their development of generic operating systems theory and specifically for the implementation of the UNIX operating system. His acceptance speech, "Reflections on Trusting Trust"^[4] presented the backdoor attack now known as the Thompson hack or trusting trust attack, and is widely considered a seminal computer security work in its own right.



Thompson (left) and Ritchie (center) receiving the National Medal of Technology from President Clinton. 

National Medal of Technology

[\[edit\]](#)

On April 27, 1999, Thompson and Ritchie jointly received the 1998 [National Medal of Technology](#) from President [Bill Clinton](#) for co-inventing the UNIX operating system and the C programming language which together have led to enormous advances in computer hardware, software, and networking systems and stimulated growth of an entire industry, thereby enhancing American leadership in the Information Age.^{[5][6]}

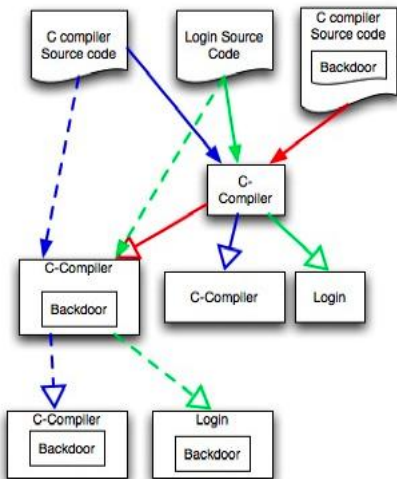
Tsutomu Kanai Award

[\[edit\]](#)

In 1999, the [Institute of Electrical and Electronics Engineers](#) chose Thompson to receive the first [Tsutomu Kanai Award](#) for his role in creating the UNIX operating system, which for decades has been a key platform for distributed systems work.^[7]

二、背景介绍：Unix后门难题

- David A. Wheeler of IDA (2009)
 - IDA, The Institute for Defense Analysis 国防分析研究所
 - 单纯的编译系统源代码分析无法检测该问题
 - 提出两次分离编译方案



History & Mission

两次分离编译方案基本思想

- 使用可信编译器/原有编译器编译同一份无注入的编译器源代码，得到两个新的可用编译器
- 用这两个新编译器编译无注入的编译器源代码得到可执行程序，进行结果对比

二、背景介绍

信息时代软件系统的核心基石：

- “图灵奖”：截止2010年，43年历史中约1/3的获奖在“编译技术和程序设计语言”领域，占绝对领先地位
- 本领域顶极会议PLDI，引用数量得到的CiteSeer全部计算机学科会议和期刊影响力排行榜中居第三

未来十年应对挑战的关键技术：

- 多核/众核处理器编程
- 复杂软件系统安全性/可靠性

二、背景介绍

具有战略价值：

- 编译系统是信息产业链中沟通处理器和应用程序之间最为关键的环节，是基础软件中的基础，属于战略必争领域，是信息产业的核心竞争力之一，对于我国信息产业的可持续发展具有重要的意义。
- “核心电子器件、高端通用芯片及基础软件产品”等国家科技重大专项中已经部署

国内相关学术活动：

- CCF“系统软件专业委员会”
- “多核环境下的编译技术研讨会”
- “编译课程教学研讨会”

三、编译技术面临的问题

时代特征:

- 计算无处不在: 云计算时代, 服务器+移动设备
主要面临三个方面的问题:

- 如何高效开发应用程序?
 - 编程模型+开发工具
- 如何提高程序运行效能?
 - 优化技术: 性能、体积、功耗
- 如何保证程序值得信赖?
 - 检测工具、形式化验证

第四个问题:

- 如何快速有效地推广技术?
 - 开源系统+大学教育

四、研究进展1：多核编程模型

计算的主流架构

- 异构多核：多核/众核/专用加速设备
- 高性能/嵌入系统
- 语言 vs 编程模型？

编程模型

- 消息/共享：MPI/OpenMP/OpenTM/UPC
- 新兴模型：CUDA/OpenCL

主要工作

- 中科院计算所
- 北京科技大学
- 江南计算所
- 国防科技大学
- 清华大学

四、研究进展2: 计算管理

基本问题

- 系统规模扩大、计算规模扩大

三个方面的工作

- 调度管理
 - 不同计算节点间的负载均衡
 - 不同计算部件GPU/CPU之间的负载均衡
- 容错管理
 - 系统级容错
 - 应用级容错**
- 功耗管理
 - 绿色计算/高效能计算/低碳生活
 - 功耗模型+优化调度

四、研究进展3：程序性能优化

性能优化是编译的持久话题

重要方法：

- 静态优化
 - 经典方向：针对特定应用类型，获得较高性能提升
- 动态优化
 - JVM及时编译，二进制翻译，脚本语言
- 迭代优化
 - 近期热点，进展突出：PLDI'10
- 反馈优化
 - Profile（剖析）based 方法
 - 插桩/PMU
- 其他方法？

四、研究进展4：检测与分析

面临的问题

- 并发程序 **meet** 多核系统

重要进展

- 指针（别名）分析
 - 精度 vs 开销
 - 超过100万行程序的分析 CGO'10
- 违反顺序一致性检测
 - MySQL/Apache
- 高危整数溢出检测
 - WINS (2000/2003) BaiDuHi

方兴未艾的重要方向

- 并发程序难以调试的特性

四、研究进展5: 验证与可信编译 清华大学

面临的问题

- 并发程序 **meet** 多核系统
- 可信软件的需求: 航空/航天/核电/高铁 等等

彻底的解决方案?

- **Tony Hoare**, 验证式编译: 计算研究的伟大挑战
- 软件自动化验证是未来五十年重要挑战

重要进展

- 携带证明方法
 - 底层代码和中间代码验证
 - 出具证明编译技术
- 携带模型代码方法

四、研究进展6: 量子计算支持



未来技术

- 2008PLDI会议特邀报告
- Alfred Aho (龙书作者) “量子计算机的编译器”

国内南京大学徐家福先生领导的团队

- 分析几种典型量子程序语言
- 设计量子程序语言NDQJava
- 给出一个处理系统
 - 编译-解释方式
 - 代码转换程序+量子汇编+解释程序

五、开源平台 1: GCC

The GNU Compiler Collection (usually shortened to GCC) is a set of programming language compilers produced by the GNU Project. It is free software distributed by the Free Software Foundation (FSF) under the GNU GPL and GNU LGPL, and is a key component of the GNU toolchain. It is the standard compiler for the free software Unix-like operating systems, and several proprietary operating systems, notably Apple Mac OS X.



Originally named the GNU C Compiler, because it only handled the C programming language, GCC was later extended to compile C++, Objective-C, Java, Fortran, and Ada among others.

GCC 技术特点

多语言支持

- C/C++/Fortran/JAVA/ ...

可移植性强：主要采用C语言编写

交叉支持能力

- build, host and target

处理器支持多：

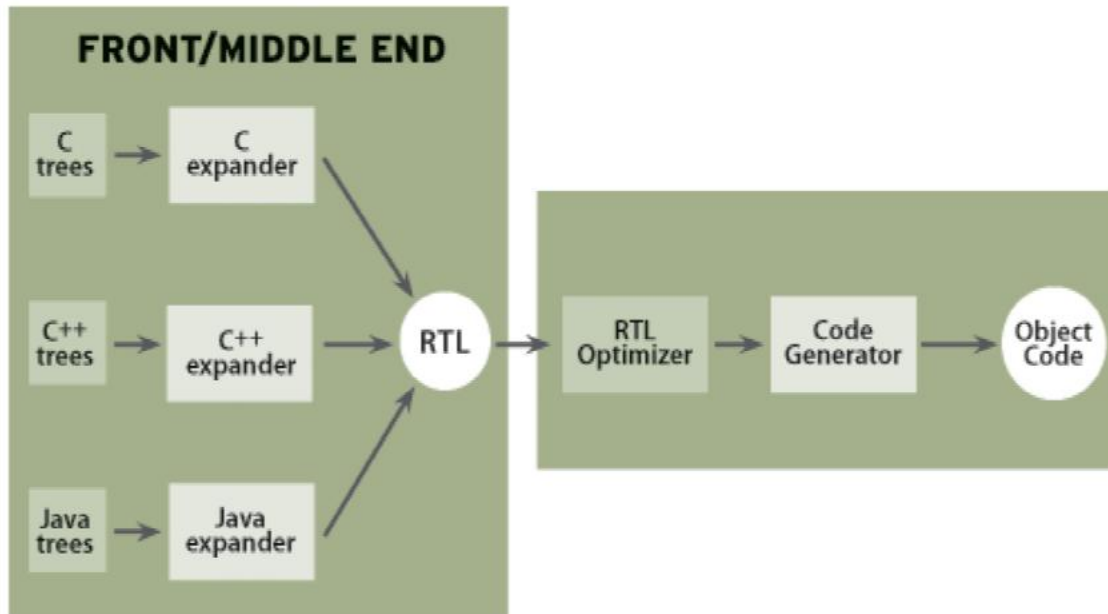
- x86/ia64/MIPS/ARM/SPARC/ ...
- 8/16/32/64 bit Processor

应用范围广

- 覆盖：高性能计算、商用服务器、PC、嵌入系统

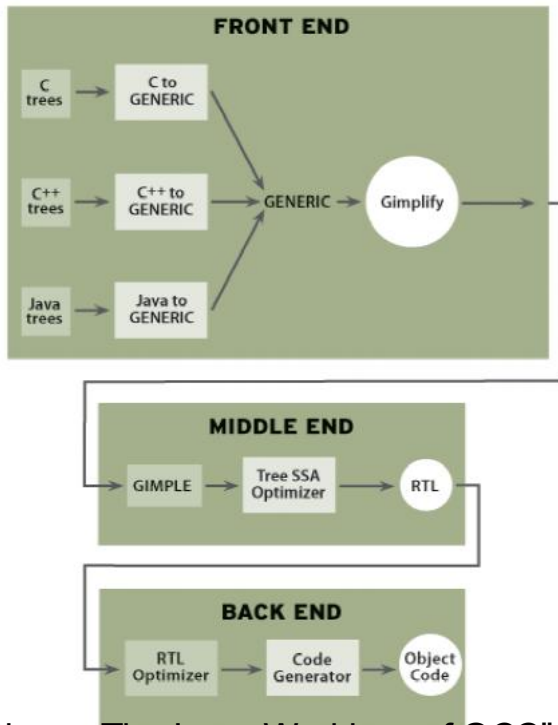
模块化设计：前端+中间表示+后端，松耦合

GCC 3.* 结构



“From Source to Binary: The Inner Workings of GCC” by Diego Novillo

GCC 4.* 结构



“From Source to Binary: The Inner Workings of GCC” by Diego Novillo

五、开源平台2: Open64

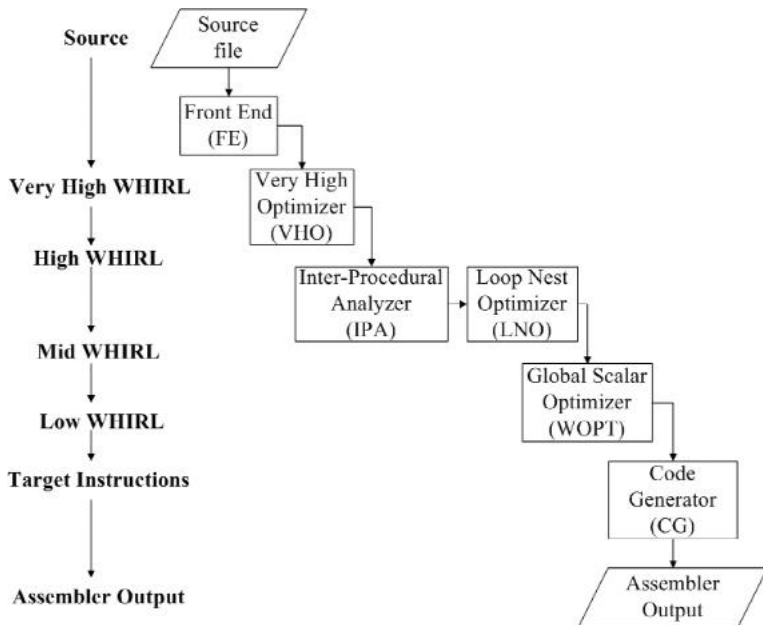
基本情况:

- 前身是**SGI**公司的一款商业高性能编译器
- 开源之后被研究机构广泛使用
 - 截止2010年10月12日, 61,446份下载, sourceforge
- 重定向到多种体系结构
 - MIPS (包括龙芯), IA64, X86, CUDA...
 - PowerPC (我们的工作)
- 在性能测试中有非常优异的表现
 - 全球性能最好的编译器

和GCC比较:

- GCC使用范围非常广泛
- GCC设计之初优化能力有限

五、开源平台2: Open64编译流程



五、开源平台2：国内贡献

安腾处理器IA64:

- 重定向和性能优化
- OpenMP支持
- C++支持/Fortran支持
- GCC扩展和Kernel编译

国产处理器支持:

- 龙芯/神威
- 在性能测试中有非常优异的表现

PowerPC处理器:

- 重定向和性能优化
- GCC扩展和kernel编译

小结:

- 队伍强、代码多、水平高

六、产品编译器

龙芯：

- 基于GCC的产品编译器
- 基于Open64的产品编译器
- DigitalBridge 进程级二进制翻译系统（X86）

神威睿智：

- 基于GCC的产品编译器
- 基于Open64的产品编译器
- 针对国产主机的二进制翻译系统（PowerPC/X86）

银河：

- 基于GCC的产品编译器
- YH-CVT 编译测试验证工具

嵌入系统：

- SimpleLight/Tsinghua, ATLAS测试语言, 流处理器

七、大学教育

编译课程的特征

- 计算机要不要开设编译原理课程？
- 非常重要，专业的标志性课程之一
- 联系计算机科学与计算机系统的典范

主要问题是如何教学

- 实例教学
- 多层次-多目标-多效果
- 资源共享
- 结合科学前沿
-

八、未来发展的几点建议

夯实理论基础：

- PLDI vs POPL

掌握开源平台：

- 游戏规则：贡献 == 发言权
- GCC vs Open64

建设评估体系：

- 反映中国信息发展的基准测试程序
- 共建、共享、共赢

培育人才队伍：

- 人才是核心竞争力
- 继续编译学术和教学研讨会
- 人员交流、国际交流、开源培训

九、致谢

报告编写得到国家自然科学基金(90818019), 国家863(2008AA01Z102)和国家科技重大专项(2009ZX01036-002)项目资助:

- 北京大学梅宏教授提出编写本报告的最初设想
- 台湾中央研究院资讯所游本中教授报告内容组织
- 复旦大学臧斌宇教授报告撰写工作机制提出意见
- 英特尔中国研究院陈兴中先生给出开源编译器建议
- 资料提供:
 - 国防科技大学杨灿群教授、吉林大学金英教授
 - 中国科技大学陈意云教授、复旦大学陈海波博士
 - 北京科技大学胡长军教授、江南计算所李中升高工
 - 北京工业大学蒋宗礼教授审阅初稿并给出重要建议

Q&A

Thank you...

