



博客园
cnblogs.com



阿里云
aliyun.com

抢红包 · 一元购 · 限时1折
#阿里云云市场携众商家双11大返场#为双11
新用户服务，帮助您轻松上云

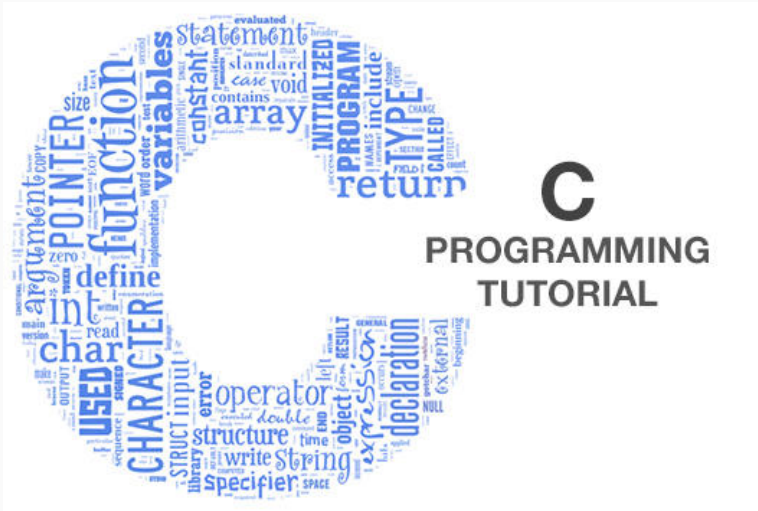


首页 园子 新闻 博文 闪存 收藏 招聘 知识库 反馈问题或建议

博客园 » 新闻 » 程序员

第一个C语言编译器是怎样编写的？

投递人 [itwriter](#) 发布于 2015-11-27 23:07 [评论\(7\)](#) 有2940人阅读 [原文链接](#) [\[收藏\]](#) < >



首先向C语言之父 Dennis Ritchie 致敬！

当今几乎所有的实用的编译器/解释器（以下统称编译器）都是用C语言编写的，有一些语言比如 Clojure, Jython 等是基于 JVM 或者说是用 Java 实现的，IronPython 等是基于 .NET 实现的，但是 Java 和 C# 等本身也要依靠C/C++来实现，等于是间接调用了C。所以衡量某种高级语言的可移植性其实就是在讨论 ANSI/ISO C 的可移植性。

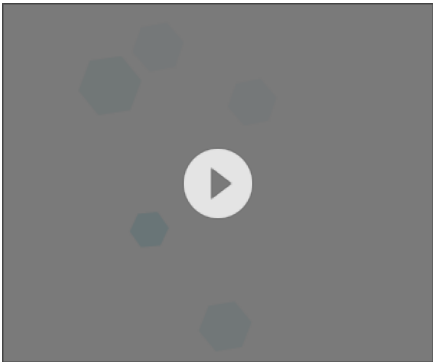
C 语言是很低级的语言，很多方面都近似于汇编语言，在《Intel 32 位汇编语言程序设计》一书中，甚至介绍了手工把简单的C语言翻译成汇编的方法。对于编译器这种系统软件，用C语言来编写是很自然不过的，即使是像 Python 这样的高级语言依然在底层依赖于C语言（举 Python 的例子是因为 Intel 的黑客正在尝试让 Python 不需要操作系统就能运行——实际上是免去了 BIOS 上的一次性 C 代码）。现在的学生，学过编译原理后，只要有点编程能力的都可以实现一个功能简单的类C语言编译器。

可是问题来了，不知道你有没有想过，大家都用C语言或基于C语言的语言来写编译器，那么世界上第一个C语言编译器又是怎么编写的呢？这不是一个“鸡和蛋”的问题.....

还是让我们回顾一下C语言历史：1970 年 Tomphson 和 Ritchie 在 BCPL（一种解释型语言）的基础上开发了B语言，1973 年又在B语言的基础上成功开发出了现在的C语言。在C语言被用作系统编程语言之前，Tomphson 也用过B语言编写过操作系统。可见在C语言实现以前，B语言已经可以投入实用了。因此第一个C语言编译器的原型完全可能是用B语言或者混合B语言与 PDP 汇编语言编写的。我们现在都知道，B语言的执行效率比较低，但是如果全部用汇编语言来编写，不仅开发周期长、维护难度大，更可怕的是失去了高级程序设计语言必需的移植性。所以早期的C语言编译器就采取了一个取巧的办法：先用汇编语言编写一个C语言的一个子集的编译器，再通过这个子集去递推完成完整的C语言编译器。详细的过程如下：

先创建一个只有C语言最基本功能的子集，记作 C0 语言，C0 语言已经足够简单了，可以直接用汇编语言编写出 C0 的编译器。依靠 C0 已有的功能，设计比 C0 复杂，但仍然不完整的C语言的又一个子集 C1 语言，其中 C0 属于 C1，C1 属于C，用 C0 开发出 C1 语言的编译器。在 C1 的基础上设计C语言的又一个子集 C2 语言，C2 语言比 C1 复杂，但是仍然不是完整的C语言，开发出 C2 语言的编译器.....如此直到 CN，CN 已经足够强大了，这时候就足够开发出完整的C语言编译器的实现了。至于这里的N是多少，这取决于你的目标语言（这里是C语言）的复杂程度和程序员的编程能力

搜索新闻



程序员找工作，就在博客园招聘频道

热门评论

today4king 发表于 11-28 11:04
当年想这个问题的时候，觉得应该这样做，没想到就是这样做的。

支持(4) 反对(0)

Stephen Lou 发表于 11-28 00:24
当时人们真有耐心，真厉害

支持(3) 反对(0)

24小时阅读排行

- 63岁老人自学单片机 8年做出机器人：有图有真相
- 谷歌返华：谁会看好，谁在唱衰，谁会恐慌，谁又在...
- 这就是长征5号：亚洲第一重型火箭
- 羽绒服10元一件还包邮 网购衣服超低价卖家图什么
- 诺基亚董董事长：活下去是第一步
- 亚马逊展示新型快递无人机：混合设计 时速近90公里
- 免费才是最贵的
- 更多...

最新新闻

- 央视曝淘宝刷单 暗示国税
- 联合国：手机约会App助长艾滋病在青少年中传播
- 微软3600无线鼠标发布：299元
- NASA公布震撼火星“森林”照 实为沙丘“作怪”
- 北京一男子用100天收集雾霾 灰尘制成板砖
- 北京PM2.5爆表：一加手机不让去上班了
- 世界艾滋病日：关于艾滋病 你该知道的常识
- 更多...

——简单地讲，如果到了某个子集阶段，可以很方便地利用现有功能实现C语言时，那么你就找到N了。下面的图说明了这个抽象过程：

C语言
CN语言
.....
C0语言
汇编语言
机器语言

那么这种大胆的子集简化的方法，是怎么实现的，又有什么理论依据呢？先介绍一个概念，“自编译”Self-Compile，也就是对于某些具有明显自举性质的强类型（所谓强类型就是程序中的每个变量必须声明类型后才能使用，比如C语言，相反有些脚本语言则根本没有类型这一说法）编程语言，可以借助它们的一个有限小子集，通过有限次数的递推来实现对它们自身的表述，这样的语言有 C、Pascal、Ada 等等，至于为什么可以自编译，可以参见清华大学出版社的《编译原理》，书中实现了一个 Pascal 的子集的编译器的。总之，已经有计算机科学家证明了，C语言理论上是可以上面说的 CVM 的方法实现完整的编译器的，那么实际上是怎样做到简化的呢？这张图是不是有点熟悉？对了就是在讲虚拟机的时候见到过，不过这里是 CVM（C Language Virtual Machine），每种语言都是在每个虚拟层上可以独立实现编译的，并且除了C语言外，每一层的输出都将作为下一层的输入（最后一层的输出就是应用程序了），这和滚雪球是一个道理。用手（汇编语言）把一小把雪结合在一起，一点点地滚下去就形成了一个滚雪球，这大概就是所谓的 0 生 1，1 生 C，C 生万物吧？

下面是 C99 的关键字：

```
1.  auto      enum      restrict  unsigned
2.  break     extern    return    void
3.  case      float     short     volatile
4.  char      for       signed    while
5.  const     goto      sizeof    _Bool
6.  continue  if          static    _Complex
7.  default   inline    struct    _Imaginary
8.  do        int       switch
9.  double    long      typedef
10. else     register  union
11. //共37个
```

仔细看看，其实其中有很多关键字是为了帮助编译器进行优化的，还有一些是用来限定变量、函数的作用域、链接性或者生存周期（函数没有）的，这些在编译器实现的早期根本不必加上，于是可以去掉 auto, restrict, extern, volatile, const, sizeof, static, inline, register, typedef，这样就形成了C的子集，C3 语言，C3 语言的关键字如下：

```
1.  enum      unsigned
2.  break     return    void
3.  case      float     short
4.  char      for       signed    while
5.  goto      _Bool
6.  continue  if          _Complex
7.  default   struct    _Imaginary
8.  do        int       switch
9.  double    long
10. else     union
11. //共27个
```

再想一想，发现 C3 中其实有很多类型和类型修饰符是没有必要一次性都加上去的，比如三种整型，只要实现 int 就行了，因此进一步去掉这些关键词，它们是：unsigned, float, short, char (char 是 int)，signed, _Bool, _Complex，_Imaginary，long，这样就形成了我们的 C2 语言，C2 语言关键字如下：

阿里云 aliyun.com

阿里云CPS

15%佣金

一单成交也有
额外奖励

最高2000元

2000元分文不取

2000元

支付1分钱

相关新闻

- 开发者博客：如何用Unity做游戏中寻路导航
- 程序猿也疯狂：一款插件给写代码带来的超爽视觉震撼
- 独立开发者：参加game jam的九大好处
- 成为一名更好的程序员：如何阅读源代码
- 程序员如何在职场中实现“跨越式”成长？

```
1. enum
2. break      return    void
3. case
4. for        while
5. goto
6. continue  if
7. default   struct
8. do        int        switch
9. double
10. else      union
11. //共18个
```

继续思考，即使是只有 18 个关键字的 C2 语言，依然有很多高级的地方，比如基于基本数据类型的复合数据结构，另外我们的关键字表中是没有写运算符的，在C语言中的复合赋值运算符->、运算符的++、- 等过于灵活的表达方式此时也可以完全删除掉，因此可以去掉的关键字有：enum, struct, union，这样我们可以得到 C1 语言的关键字：

```
1. break      return    void
2. case
3. for        while
4. goto
5. continue  if
6. default
7. do        int        switch
8. double
9. else
10. //共15个
```

接近完美了，不过最后一步手笔自然要大一点。这个时候数组和指针也要去掉了，另外 C1 语言其实仍然有很大的冗余度，比如控制循环和分支的都有多种表述方法，其实都可简化成一种，具体的来说，循环语句有 while 循环，do...while 循环和 for 循环，只需要保留 while 循环就够了；分支语句又有 if...{ }，if...{ }...else, if...{ }...else if...，switch，这四种形式，它们都可以通过两个以上的 if...{ } 来实现，因此只需要保留 if,...{ } 就够了。可是再一想，所谓的分支和循环不过是条件跳转语句罢了，函数调用语句也不过是一个压栈和跳转语句罢了，因此只需要 goto（无限制的 goto）。因此大胆去掉所有结构化关键字，连函数也没有，得到的 C0 语言关键字如下：

```
1. break    void
2. goto
3. int
4. double
5. //共5个
```

只有 5 个关键字，已经完全可以汇编语言快速的实现了。通过逆向分析我们还还原了第一个C语言编译器的编写过程，也感受到了前辈科学家们的智慧和勤劳！我们都不过是巨人肩膀上的灰尘罢了！0 生1，1 生C，C生万物，实在巧妙！

来自: CSDN

19

推荐

0

反对

找优秀程序员，就在博客园



标签：程序员

« 上一篇：[人人第三财季亏损扩大至8200万美元](#)(2015-11-27 23:04)

» 下一篇：[微软推出3D音频技术 帮助视障人士在城市当中行走](#)(2015-11-27 23:08)

已经有 7 位网友对此新闻发表了看法。

第1楼 [Stephen Lou](#) 发表于 2015-11-28 00:24

66

当时人们真有耐心，真厉害

支持(3) 反对(0) 回复 引用

第2楼

[today4king](#)

发表于 2015-11-28 11:04

“

当年想这个问题的时候，觉得应该这样做，没想到就是这样做的。

”

支持(4)

反对(0)

回复

引用

第3楼

[bananaplan](#)

发表于 2015-11-28 14:42

“

解惑了

”

支持(0)

反对(0)

回复

引用

第4楼

[X!ao f](#)

发表于 2015-11-28 15:44

“

不奇怪，就好比硬件加工工艺的进化，高精度加工设备本身也是由低一级的设备生产出来的

”

支持(0)

反对(0)

回复

引用

第5楼

[HolleHuang](#)

发表于 2015-11-28 17:01

“

C99 表示写99个子集？还不是1999年定下来的？

”

支持(0)

反对(0)

回复

引用

第6楼

[五星](#)

发表于 2015-11-28 20:17

“

一直在想一个问题：如果第一个编译器有个bug，会不会导致后面的所有编译器都存在bug

”

支持(0)

反对(1)

回复

引用

第7楼

[稻梁诗神](#)

发表于 2015-11-29 12:40

“

赞赞赞

”

支持(0)

反对(0)

回复

引用

注册用户登录才能发表评论，[登录](#)或[注册](#)。

刷新评论

↑ TOP