开篇词 | 在真实世界的编译器中游历

2020-06-01 宫文学

编译原理实战课 进入课程 >

讲述：宫文学
时长 14:02　大小 12.86M

你好，我是宫文学，一名技术创业者，现在是北京物演科技的 CEO，很高兴在这里跟你见面。

我在 IT 领域里已经工作有 20 多年了。这其中，我个人比较感兴趣的，也是倾注时间精力最多的，是做基础平台类的软件，比如国内最早一批的 BPM 平台、BI 平台，以及低代码 / 无代码开发平台（那时还没有这个名字）等。这些软件之所以会被称为平台，很重要的原因就是拥有很强的定制能力，比如流程定制、界面定制、业务逻辑定制，等等。而这些定制能力，依托的就是编译技术。

在前几年，我参与了一些与微服务有关的项目。我发现，前些年大家普遍关注那些技术问题，比如有状态的服务（Stateful Service）的横向扩展问题，在云原生、Serverless、FaaS 等新技术满天飞的时代，不但没能被很好地解决，反而更恶化了。究其原因就是，状

态管理还是被简单地交给数据库，而云计算的场景使得数据库的压力更大了，数据库原来在性能和扩展能力上的短板，就更加显著了。

而比较好的解决思路之一，就是大胆采用新的计算范式，发明新的计算机语言，所以我也有意想自己动手搞一下。

我从去年开始做设计，已经鼓捣了一阵了，采用了一些很前卫的理念，比如云原生的并发调度、基于 Actor 的数据管理等。总的目标，是要让开发云原生的、有状态的应用，像开发一个简单的单机应用一样容易。那我们就最好能把云架构和状态管理的细节给抽象掉，从而极大地降低成本、减少错误。而为编程提供更高的抽象层次，从来就是编译技术的职责。

Serverless 和 FaaS 已经把无状态服务的架构细节透明掉了。但针对有状态的服务，目前还没有答案。对我而言，这是个有趣的课题。
在我比较熟悉的企业应用领域，ERP 的鼻祖 SAP、SaaS 的鼻祖 SalesForce，都用自己的语言开发应用，很可惜国内的企业软件厂商还没有做到这一点。而在云计算时代，设计这样一门语言绕不过去的一个问题，就是解决有状态服务的云化问题。我希望能为解决这个问题提供一个新工具。当然，这个工具必须是开源的。

正是因为给自己挖了这么大一个坑，也促使我更关心编译技术的各种前沿动态，也非常想把这些前沿的动态、理念，以及自己的一些实战经验都分享出来。

所以去年呢，我在极客时间上开了一门课程《编译原理之美》，帮你系统梳理了编译技术最核心的概念、理论和算法。不过在做第一季的过程中呢，我发现很多同学都跟我反馈：我确实理解了编译技术的相关原理、概念、算法等，但是有没有更直接的方式，能让我更加深入地把知识与实践相结合呢？

为什么要解析真实编译器？

说到把编译技术的知识与实践相结合，无外乎就是解决以下问题：

> 我已经知道，语法分析有自顶向下的方法和自底向上的方法，但要自己动手实现的话，到底该选择哪个方法呢？是应该自己手写，还是用工具生成呢？
>
> 我已经知道，在语义分析的过程中要做引用消解、类型检查，并且会用到符号表。那具体到自己熟悉的语言，这些工作是如何完成的呢？有什么难点和实现技巧呢？符号表又

被设计成什么样子呢？

我已经知道，编译器中会使用 IR，但实际使用中的 IR 到底是什么样子的呢？使用什么数据结构呢？完成不同的处理任务，是否需要不同的 IR 呢？

我已经知道，编译器要做很多优化工作，但针对自己熟悉的语言，这些优化是如何发生的？哪些优化最重要？又要如何写出便于编译器优化的代码呢？

类似的问题还有很多，但总结起来其实就是：**真实世界的编译器，到底是怎么写出来的？**

那弄明白了这个问题，到底对我们有什么帮助呢？

第一，研究这些语言的编译机制，能直接提高我们的技术水平。

一方面，深入了解自己使用的语言的编译器，会有助于你吃透这门语言的核心特性，更好地运用它，从而让自己向着专家级别的工程师进军。举个例子，国内某互联网公司的员工，就曾经向 Oracle 公司提交了 HotSpot 的高质量补丁，因为他们在工作中发现了 JVM 编译器的一些不足。那么，你是不是也有可能把一门语言吃得这么透呢？

另一方面，IT 技术的进化速度是很快的，作为技术人，我们需要迅速跟上技术更迭的速度。而这些现代语言的编译器，往往就是整合了最前沿的技术。比如，Java 的 JIT 编译器和 JavaScript 的 V8 编译器，它们都不约而同地采用了“Sea of Nodes”的 IR 来做优化，这是为什么呢？这种 IR 有什么优势呢？这些问题我们都需要迅速弄清楚。

第二，阅读语言编译器的源码，是高效学习编译原理的重要路径。

传统上，我们学习编译原理，总是要先学一大堆的理论和算法，理解起来非常困难，让人望而生畏。

这个方法本身没有错，因为我们学习任何知识，都要掌握其中的原理。不过，这样可能离实现一款实用的编译器还有相当的距离。

那么根据我的经验，学习编译原理的一个有效途径，就是阅读真实世界中编译器的源代码，跟踪它的执行过程，弄懂它的运行机制。因为只要你会写程序，就能读懂代码。既然能读懂代码，那为什么不直接去阅读编译器的源代码呢？在开源的时代，源代码就是一个巨大的知

识宝库。面对这个宝库，我们为什么不进去尽情搜刮呢？想带走多少就带走多少，没人拦着。

当然，你可能会犯嘀咕：**编译器的代码一般都比较难吧？以我的水平，能看懂吗？**

是会有这个问题。当我们面对一大堆代码的时候，很容易迷路，抓不住其中的重点和核心逻辑。不过没关系，有我呢。在本课程中，我会给你带路，并把地图准备好，带你走完这次探险之旅。而当你确实把握了编译器的脉络以后，你对自己的技术自信心会提升一大截。这些计算机语言，就被你摘掉了神秘的面纱。

俗话说“读万卷书，行万里路”。如果说了解编译原理的基础理论和算法是读书的过程，那么探索真实世界里的编译器是什么样子，就是行路的过程了。根据我的体会，**当你真正了解了身边的语言的编译器是怎样编写的之后，那些抽象的理论就会变得生动和具体，你也就会在编译技术领域里往前跨出一大步了。**

我们可以解析哪些语言的编译器？

那你可能要问了，在本课程中，**我都选择了哪些语言的编译器呢？选择这些编译器的原因又是什么呢？**

这次，我要带你解析的编译器还真不少，包括了 Java 编译器（javac）、Java 的 JIT 编译器（Graal）、Python 编译器（CPython）、JavaScript 编译器（V8）、Julia 语言的编译器、Go 语言的编译器（gc），以及 MySQL 的编译器，并且在讲并行的时候，还涉及了 Erlang 的编译器。

我选择剖析这些语言的编译器，有三方面的原因：

第一，它们足够有代表性，是你在平时很可能会用到的。这些语言中，除了 Julia 比较小众外，都比较流行。而且，虽然 Julia 没那么有名，但它使用的 LLVM 工具很重要。因为 LLVM 为 Swift、Rust、C++、C 等多种语言提供了优化和后端的支持，所以 Julia 也不缺乏代表性。

第二，它们采用了各种不同的编译技术。这些编译器，有的是编译静态类型的语言，有的是动态类型的语言；有的是即时编译（JIT），有的是提前编译（AOT）；有高级语言，也有 DSL（SQL）；解释执行的话，有的是用栈机（Stack Machine），有的是用

寄存器机，等等。不同的语言特性，就导致了编译器采用的技术会存在各种差异，从而更加有利于你开阔视野。

第三，通过研究多种编译器，你可以多次迭代对编译器的认知过程，并通过分析对比，发现这些编译器之间的异同点，探究其中的原因，激发出更多的思考，从而得到更全面的、更深入的认知。

看到这里，你可能会有所疑虑：**有些语言我没用过，不怎么了解，怎么办？**其实没关系。因为现代的高级语言，其实相似度很高。

一方面，对于不熟悉的语言，虽然你不能熟练地用它们来做项目，但是写一些基本的、试验性的程序，研究它的实现机制，是没有什么问题的。

另一方面，学习编译原理的人会练就一项基本功，那就是更容易掌握一门语言的本质。特别是我这一季的课程，就是要帮你成为钻到了铁扇公主肚子里的孙悟空。研究某一种语言的编译器，当然有助于你通过“捷径”去深入地理解它。

我是如何规划课程模块的？

这门课程的目标，是要让你对现代语言的编译器的结构、所采用的算法以及设计上的权衡，都获得比较真切的认识。其最终结果是，如果要你使用编译技术来完成一个项目，你会心里非常有数，知道应该在什么地方使用什么技术。因为你不仅懂得原理，更有很多实际编译器的设计和实现的思路作为你的决策依据。

为了达到本课程的目标，我仔细规划了课程的内容，将其划分为预备知识篇、真实编译器解析篇和现代语言设计篇三部分。

在**预备知识篇**，我会简明扼要地帮你重温一下编译原理的知识体系，让你对这些关键概念的理解变得更清晰。磨刀不误砍柴工，你学完预备知识篇后，再去看各种语言编译器的源代码和相关文档时，至少不会被各种名词、术语搞晕，也能更好地建立具体实现跟原理之间的关联，能互相印证它们。

在**真实编译器解析篇**，我会带你研究语言编译器的源代码，跟踪它们的运行过程，分析编译过程的每一步是如何实现的，并对有特点的编译技术点加以分析和点评。这样，我们在研究

了 Java、Java JIT、Python、JavaScript、Julia、Go、MySQL 这 7 个编译器以后，就相当于把编译原理印证了 7 遍。

在**现代语言设计篇**，我会带你分析和总结前面已经研究过的编译器，进一步提升你对相关编译技术的认知高度。学完这一模块以后，你对于如何设计编译器的前端、中端、后端、运行时，都会有比较全面的了解，知道如何在不同的技术路线之间做取舍。

好了，以上就是这一季课程的模块划分思路了。你会发现，这次的课程设计，除了以研究真实编译器为主要手段外，会更加致力于扩大你的知识版图、增加你的见识，达到“行万里路”的目的。

可以说，我在设计和组织这一季课程时，花了大量的时间准备。因此这一季课程的内容，不说是独一无二的，也差不多了。你在市面上很少能找到解析实际编译器的书籍和资料，这里面的很多内容，都是在我自己阅读源代码、跟踪源代码执行过程的基础上梳理出来的。

写在最后

近些年，编译技术在全球范围内的进步速度很快。比如，你在学习 Graal 编译器的时候，你可以先去看看，市面上有多少篇围绕它的高质量论文。所以呢，作为老师，我觉得我有责任引导你去看到、理解并抓住这些技术前沿。

我也有一个感觉，在未来 10 年左右，中国在编译技术领域，也会逐步有拿得出手的作品出来，甚至会有我们独特的创新之处，就像我们当前在互联网、5G 等领域中做到的一样。

虽然这个课程不可能涵盖编译技术领域所有的创新点，但我相信，你在其中投入的时间和精力是值得的。你通过我课程中教给你的方法，可以对你所使用的语言产生更加深入的认知，对编译器的内部结构和原理有清晰理解。最重要的是，对于如何采用编译技术来解决实际问题，你也会有能力做出正确的决策。

这样，这个课程就能起到抛砖引玉的作用，让我们能够成为大胆探索、勇于创新的群体的一份子。未来中国在编译技术的进步，就很可能有来自我们的贡献。我们一起加油！

最后，我还想正式认识一下你。你可以在留言区里做个自我介绍，和我聊聊，你目前学习编译原理的最大难点在哪？或者，你也可以聊聊你对编译原理都有哪些独特的思考和体验，欢迎在留言区和我交流讨论。

好了，让我们正式开始编译之旅吧！

下一篇　学习指南 | 如何学习这门编译原理实战课？

精选留言 (24)

写留言

sugar
2020-06-01

终于见到宫老师的第二季啦～ 我来抢个首赞👍 哈哈

展开

作者回复: 又见到sugar!

4

吃鱼
2020-06-01

老师，因为专业要学习二进制安全，所以特别想通过您的课程了解编译方面的知识，我编译原理之前学的不太扎实，您的两个课程我觉得都很硬核，应该先学哪一门比较好呢？

作者回复: 两门课都是围绕编译原理，但讲述方式和侧重点不同。

《编译原理之美》是通过手工实现一款编译器的方式，带你了解这个过程中的知识点，循序渐进地讲解，最后才去介绍难度比较高的那几个算法。对编译器前端工具ANTLR的使用也比较多。如果你想学会如何快速实现一个编译器，可以先从这门课入手。

《编译原理实践课》总体的目标，是带你“行万里路”，扩大你的视野和见识。它讲述的方式是先做一些基础知识的概述（概述部分也会注意扩展你的知识面）、然后是研究多个编译器的实现，最后是总结这些编译器的实现，并探讨现代语言的设计。如果你比较想了解真实世界的编译器的情况，可以先从这门课入手。

你的专业是二进制安全，涉及后端的内容会更多，所以我觉得了解真实世界的编译器的情况会有很大的帮助。

因为编译原理的知识点比较多，所以用不同的课程体现不同的侧面。在我的出书计划中，也发现其实仅用一卷书是不能说完的。比如，有的作者用“编译器DIY”的方式就可以写一本书，但对原理和算法体现得就不够；写算法比较深入的书呢，又没有体现当前真实编译器中所采用的技术。

3

Matrix
2020-06-01

目前是在校研究生，研究方向是二进制的漏洞挖掘与利用。平常在论文、工程实现中多多少少和编译原理相关的知识有交集，如：SSA、AST、LLVM-IR等，相关的理论知识书本上学过，但没有形成较为清晰的知识体系，很多地方有一种雾里看花的感觉，希望能结合实际对编译器的内部结构和原理有更清晰的理解。

展开

作者回复: 你这个专业，肯定要多了解真实语言的实现。二进制漏洞问题本来就是实现编译器时要考虑的一个因素。比如，在使用内存的时候，要让返回的内存地址没有明显的规律，避免出现漏洞。

3

Jacob.C
2020-06-02

上了去年宫老师的课后，我做了一个sql解析器，解决了我司数据仓库字段级血缘分析的难题

作者回复: 恭喜你的成绩！
你说的血缘分析，是Lineage Analysis吗？我以前搞元数据的时候接触过，看上去很亲切:-)

1

至今未来
2020-06-01

编译原理之美 只看了一遍 差不多忘光了 宫老师 我又来了(「·ω·)「嘿

作者回复: 欢迎！学新而忆故，温故而知新，迭代提升认知！

1

王成
2020-06-01

老师好
学习编译的难点:编译原理之美还没有学完，正在努力学习，由于工作学习等多方面原因，学习进度较慢
打算应用编译原理实现的东西:目前工作是实时计算，公司目前关于实时流使用了storm和flink,我想开发一套程序，使得一次开发，可以同时在两个平台运行，同时，可以做到将一...
展开

作者回复: 谢谢你分享自己的使用场景。非常好。我对你的名字有印象:-)
如果有可能，你也可以把自己的设计思路描述一下，我们可以多做一些探讨！
有具体需求推动，学习会更有动力。

1

罗洪涛@融众
2020-06-01

关注前端AST希望落地前端定制化
展开

作者回复: 感谢分享！你这个需求很重要。一个好的软件，一定要能实现定制化、平台化。

另外，当前在流行的NoCode、LowCode开发，跟这个也很有关。
我十年前就做过NoCode的开发平台，当时还没有这个词汇。在云计算条件时代，很多应用的开发成本很高，所以这个方向值得关注。

1 1

Fan
2020-06-01

期待宫老师多讲讲llvm方面相关的。

展开

作者回复: 我在部落里就看到你这个需求啦！

我已经在Julia那部分，放了一些LLVM的内容。

另外，昨天直播的时候，有同学提出是否将一下C++的编译器。C++是LLVM的一个前端。所以可以考虑结合这个需求，再做点加餐。

1

X！！
2020-06-01

第一

展开

作者回复: 欢迎:-)

1

xiaobang
2020-06-09

希望作者能讲一下pypy，一直觉得这个很神奇

展开

作者回复: 我准备的内容是CPython。看明白CPython编译器后，再去看PyPy应该比较简单。毕竟Python的主体是用C写的。

皮特尔
2020-06-07

第一季还没有看完，第二季就来了。得抓紧时间了。

展开

Geek_c34bd6
2020-06-06

我是物联网工程师，这个行业现在主流用的是c语言(由于内存限制以及兼容性等问题，连C++都不太有人用)。但是c的缺点太明显了，开发效率低，需要自己管理内存。所以我想写一门编程语言。这门编程语言有高级语言的一些特性，但是编译出来的代码是c代码。这样只要能用c语言点环境就能用这门语言。以牺牲编译时间为代码换取开发效率和兼容性。所以最近沉迷于编译原理

展开

作者回复: 以前，像Lisp等好几种语言都是编译成c语言，再用c语言的编译器来编译成目标代码。
现在网上也有一些开源的编译器（名字不太记得了），能编译成c代码，你可以搜索一下，作为参考。

Geek_d0aef1
2020-06-05

宫老师，你好

我是从编译原理之美过来的，老师的课很有深度，很喜欢
我自己是用Qt写工具的，是抱着学习的心态来的

展开

作者回复: 用QT写工具，不错。
既然叫做工具，肯定要具备各种自定义能力，使工具具有普适性。这个地方可能就要用到编译技术。

寻回光明
2020-06-04

老师你好，我是一名科班的本科生，但是编译原理这块始终不懂如何学习，希望跟老师进入这个全新的世界。

作者回复: 嗯，知识就在那里，没有拿不下的道理，只要盯住它不放！

灰化黑化肥
2020-06-04

老师，《编译原理之美》我之前没学习，另外我也不是本专业学生，没有系统学习过编译原理，我是否需要在学习这门课之前学习一下《编译原理之美》呢？

展开

作者回复: 这两门课讲述编译原理的方式不大一样，可以互相补充，但没有先后关系。

Sruby
2020-06-03

宫老师，上一门课没有学习过，直接上这门课是否会存在障碍？

作者回复: 这两门课是用不同的方式来讲述编译原理。上一门课主要是用手工实现一门语言为主线。而这门课是以考察真实语言的编译器为主线。这两门课不存在先后关系，可以互相补充。

大马猴
2020-06-03

老师，您好，继续跟您学习

展开

作者回复: 欢迎!
在新的一季共同进步!

Apsaras
2020-06-03

希望老师重点讲讲关于LLVM的部分

展开

作者回复: LLVM确实很有用。它对我也是必不可少的工具。我争取在这方面增加点内容。

ヾ(◍°∇°◍)ﾉﾞ

2020-06-03

谢谢老师，希望这次跟着宫老师坚持下来。编译原理像英语一样，是我大学里是成绩最差的。现在就开始把这些差的东西从新学起来吧。我是做大数据的，非常想把sparksql搞透并自己实现

作者回复: 加油!

黄伟

2020-06-02

果断报名

展开

作者回复: 欢迎!