# 跨架构编程的开放oneAPI是替代CUDA和OpenCL的选项

<https://www.infoq.cn/article/27jwUuoTViVu7ZTkHKZw/>

2019 年 8 月 30 日

本文字数：7197 字

阅读完需：约 24 分钟



在 2018 年底的英特尔架构日上，英特尔首席架构师 Raja Koduri 对外公布了公司正在着力研发的一件“大事”：一个名为 oneAPI 的软件编程框架。oneAPI 是英特尔在通用并行计算架构上的最新尝试，也是英特尔旨在应对硬件芯片架构未来日益多元化问题的前瞻之作。顾名思义，oneAPI 旨在提供一个跨不同硬件架构的统一编程接口，使开发者可以随意在底层硬件之间切换和优化，它不仅支持英特尔的 CPU、GPU、FPGA 以及各种 AI 和其他应用的硬件加速器，也对外部所有硬件厂商开放。oneAPI 的口号是“让每一个晶体管都派上用场”，这也很好地总结了 oneAPI 的终极目标。

当然，支持并行/异构计算的统一编程框架并非英特尔独一家，在此之前开发者群体更为熟悉的是英伟达的 CUDA 和由苹果发起的开放规范 OpenCL。其中 CUDA 完全由英伟达独有，只支持英伟达的GPU，源码和编程语言并不对外开放，但由于完整的配套生态、很好的易用性和高效的更新迭代速度而广受机器学习应用开发者和框架开发者欢迎。OpenCL 相比 CUDA 好的地方在于它是一个开放的编程规范且支持所有硬件架构，但缺点比优点更多，OpenCL 虽有“开放”的美名，但由于实现要靠各大硬件厂商，无法保证质量和更新速度，目前只有 AMD 的支持做的相对成熟，其他架构支持做的不太好；为了支持不同硬件，存在很多冗余代码，硬件利用率不高；在英伟达的卡上性能不如CUDA，在非英伟达的卡上驱动质量参差不齐；异构代码编写复杂且文档不够清晰。

随着越来越多新芯片架构和编程模型的出现，未来的通用并行计算市场仍存在很多不确定因素，CUDA 和 OpenCL 肯定不是终点。按照英特尔规划的技术路线图，oneAPI 测试版将在今年第四季度开放给开发者。在英伟达和 CUDA 已经占据先发优势的前提下，英特尔的 oneAPI 是否能激起新的浪花？同为开放规范，OpenCL 已经遇到的问题是否也会成为 oneAPI 的挑战？这些问题让我们更加期待 oneAPI 测试版本的推出。

在此之前，开发者对于 oneAPI 还有许多好奇：与 CUDA、OpenCL 相比，英特尔的 oneAPI 有何不同之处？具备哪些优势？oneAPI 的架构是如何设计的？怎么做到让不同硬件架构发挥最佳性能？oneAPI 的推出可能会给开发者带来什么变化？在英伟达和 CUDA 已经占据先发优势的前提下，英特尔计划怎么打造 oneAPI 的开发者生态和社区？

近日，InfoQ 记者有幸在英特尔北京总部专访了英特尔架构、图形与软件部副总裁兼计算性能与开发者产品部门总经理 William (Bill) Savage 和英特尔架构、图形与软件部副总裁兼编译器与语言部门总经理 Alice Chan，围绕 oneAPI 的架构设计、定位、与现有同类产品的差异点和未来规划做了深入探讨。

## oneAPI 既是开放规范也是产品

**InfoQ：首先请您对 oneAPI 做一个总体的介绍。**

**Bill Savage：oneAPI 始于硬件架构**，尤其是数据中心。在数据中心里，今天的架构并不只局限于 CPU，还有 GPU、FPGA 以及专用的 AI 芯片。这些分别对应着标量（Scalar）、矢量（Vector）、矩阵（Matrix）和空间（Spatial）的不同计算架构，我们称之为 SVMS 架构。我们提出的 oneAPI 是一种统一的软件架构，它能够跨不同的架构、跨不同的厂商，包括除英特尔之外的其他硬件厂商。总的来讲，**oneAPI 旨在从软件层面简化和统一标量、矢量、矩阵和空间的不同硬件架构。**

oneAPI 包含两部分，第一部分是**跨架构的编程语言** Data Parallel C++（下文简称 DPC++），不同的架构和厂商都可以使用；第二部分是能够满足不同领域需求的**跨架构库的集合**。**无论是编程语言还是架构库，重点都放在性能上**，因为在数据中心里提供最佳性能是重中之重。

**oneAPI 既是一个开放的行业规范，同时也是一个产品。**我们会提供实施参考，以便其他厂商和外部开发者根据参考来进行实施，我们鼓励行业及社区的支持。对于软件开发者来讲，oneAPI 的好处是使他们能够使用同一份代码跨不同架构和厂商，更多地重复利用代码并降低开发成本。**oneAPI 为大家提供除了英伟达 CUDA 之外的另一种选项**。

oneAPI 是基于英特尔的经验以及现有的至强系列产品构建的，从单一架构基础上进行演变，可以支持多架构。英特尔对于多种不同架构都有很多经验，这也成为这一产品背后的支持。目前在行业当中英特尔拥有性能方面最为领先的 C++编译器，英特尔的数学核心库也是使用量上行业排名第一的数学函数库。对于软件开发者来讲，我们拥有性能最优的分析器 VTune Amplifer，我们在这样的基础上建立了英特尔的 oneAPI 产品。

**InfoQ：哪些领域能从 oneAPI 这个项目中受益？**

**Bill Savage：**主要从 oneAPI 当中受益的行业和领域是数据中心，它可以让跨各种架构的计算都得到高性能的实施。但实际上oneAPI给我们带来的很大好处应该说是间接的。现在直接用到 oneAPI 的用户主要有两大领域：第一是高性能计算编程，这方面相关的应用包括人类基因测序、油藏勘察开发建模、金融交易分析、物理原子加速器；第二是人工智能框架开发者。**只要是需要大量自定义代码的应用都会直接使用 oneAPI**。

## 为什么我们还需要一种全新的编程语言？

**InfoQ：能否再详细聊一聊 oneAPI 中的新编程语言 DPC++？**

**Alice Chan：**正如前面 Bill 所说，今天在数据中心拥有大量的多元化硬件架构。如果你希望在这样一个多元化的不同架构中进行编程的话，很可能需要很多种不同工具和不同编程语言。这就意味着在软件开发过程中需要多支团队，他们各自要去学习很多不同的专业技能，这显然不是一种最高效的软件开发方式。**英特尔希望改变这种现状，并不仅仅是为了英特尔自己的硬件去改变，而是为全行业去改变**。

那么我们为什么需要一种全新的语言呢？这个世界上已经有这么多语言了。有一些语言是大家耳熟能详的，例如众所周知的 C++，它是可移植的，而且在底层性能表现非常好。但是今天我们有越来越多并行架构编程的需求，C++本身缺乏一些并行语言的特征，使得用它来做并行编程的目的很难实现。而有一些语言如 MATLAB，可能更多集中在顶层，如果想在底层得到很好的性能也很难。英伟达发明了一种语言 CUDA，它能够进行并行架构的编程，可以把负载转移到加速器，但是英伟达这种语言是专用的，只能用在英伟达自己的硬件上，其他人都无法使用它，如果你没有它的规范就无法实施、也无法编译。还有其他一些语言例如 OpenCL 也能实现类似性能，但是围绕它的社群和整体行业活跃度并不大，所以你也很难获得想要的性能。这就是为什么英特尔需要开发一种新的语言。

正如刚才 Bill 所说，我**们这个全新编程语言的目的就是要实现跨架构和高性能，同时保证是开放的，针对所有软件开发者开放、针对所有的硬件厂商开放**。

首先是跨架构。当前对于数据中心里的硬件，最重要的点之一是要能够进行并行计算，另一点是要能够载入到加速器当中。我们所说的无论是称作流水线还是单指令流、多数据流，还是多线程还是多核，实际上它们的宗旨都是一样的，总之是要能够保证同时进行多项计算。如果你用一个编译器或者运行时工具，可以将 N 个独立的计算映射到统一的并行计算方式，就可以用很多不同的硬件架构了。

DPC++能够完成这些重新编排的工作，从而支持几乎所有的硬件架构。

但是我们为什么自信能够保证性能最优呢？首先我们有世界领先的 C++的编译器，所以我们知道其底层是如何编译的，我们所做的是使用底层虚拟机按照这种格式进行编译。其次，当前已经有有几种可选的技术可以将计算编译成前面提到的有序结构，包括矢量化、单程序多数据（SPMD），还有细粒度 SPMD，**DPC++语言将上述三种技术的思路融会贯通，能够跨架构实现毫不妥协的性能所需的特性和抽象。**

事实上我们已经开始 DPC++这个项目有一段时间了，通过实验我们看到它的性能至少是优于或者等同于之前的这些技术和方式。即便在我们还没有发布测试版本之前，它带来的性能已经超过OpenCL，并且等同于标准的CPU的优化版本。

另一点是开放。我们不希望做一件封闭的事情，不想让这个新的编程语言只属于我们自己，而不能向全行业开放。

DPC++是建立在标准上的，其中 C++是所有人都耳熟能详的国际标准；SYCL本身也是一个标准，是建立在 C++基础上的，它拥有我们寻求的一些特性，比如并行计算以及异构计算。但是我们认为它们现有的特性还不够，所以英特尔将在此基础上开发新特性，来确保实现我们预期的性能。另外对我们来讲很重要的就是要有 Extensions 完整文档，并开放给整个社群。

目前 DPC++已经在 GitHub 上开放，这个 GitHub 项目基于 LLVM 和 Clang。要想找到这个编译器的现有信息，可以查看 https://github.com/intel/llvm/tree/sycl/sycl。现在已经有很多公司参与到了对这个语言及编译器的改进中。未来我们的目标是将内部开发的这些代码最终完全开放给 LLVM。对于在语言方面我们增加的新特性，也将完全开放给基础的标准，也就是 C++以及 SYCL。

**总之，我们的目标是将它提供给所有不同的硬件架构，并且是保证它完全的开放性。**而对于英伟达的 CUDA，现在外部人士完全不知道语言怎么实施的，我们希望能够避免这种现象。

**InfoQ：对于 HPC 编程和 AI 框架开发这两个领域的开发者来说，现在大家已经在使用的编程语言和 oneAPI 的 DPC++之间的关系是什么样的？DPC++对原有的这些编程语言是辅助性的还是替代性的？**

**Bill Savage：**其实他们是相似的，同时更是互补的。作为 AI 框架的开发人员，会更多使用库，像 MKL-DNN 也会少量使用自定义代码；而高性能计算开发人员会更多使用自定义代码，较少量地使用库，总之来讲都是这两者的结合。对于这些开发人员已经投入使用的原有编程语言和库来说，DPC++既是一种互补又是一种替代。为什么这样讲？今天自定义代码所使用的编程语言是针对某一特定架构的，当你选择这一架构你就必须选择这一种语言。**DPC++提供了另外一个选择，你可以选择你想要的那种语言，但是它可以跨任何架构使用**。

**InfoQ：相比 CUDA、C++，DPC++有什么不同之处？**

**Bill Savage：**DPC++是基于 C++以及它的根建立的，所以它实际上也是 C++的一种实施。但是它创造了一种明确的方式去表达我们所说的性能核心程序，它可以进行卸载并传递到核当中。CUDA 本身也有类似特性，它可以写性能核心程序，然后可以得到传递。DPC++的另一个特点是，开发员可以非常明确地确定并行的部分，可以让语言知道这个核就是并行的。但是不同点在于 DPC++的编译器是拿来并行然后非常明确地映射到各种架构上，采取的是全并行的方式，并不是只依赖于 GPU。

**InfoQ：学习 DPC++这种新语言的难度大吗？**

**Bill Savage：**DPC++非常接近于 C++，也非常接近于原来使用 CUDA 程序员所熟悉的语言，对于这些开发人员来讲他们的学习曲线会非常平缓。我们希望能在 Beta 版本发布的时候分享更多的工具，帮助开发人员更方便地进行代码迁移。我们相信 DPC++能够有非常大的使用群体，**如果一个程序员对于 CUDA 非常熟悉的话，那么使用 DPC++进行编程应该没有任何问题。**

**InfoQ：英特尔是否准备了参考设计方案帮助开发者更快的上手 oneAPI 项目？你们采取哪些措施让 oneAPI 更好更快的推广？**

**Bill Savage：**是的。我们首先拥有在语言方面的规范，然后它的各种 Extensions 可以在 GitHub 网站上找到。我们采取的这种方式能够让 oneAPI 得到快速的采纳和实施。今年晚些时候我们会发布 Beta 产品，这是一个非常优秀的基于 oneAPI 的英特尔的参考产品。另外，我们会通过与广泛的合作伙伴建立合作来实现快速采纳。

## 应对芯片架构碎片化挑战的解法

**InfoQ：英特尔指出了 SVMS 这种硬件产品上的架构多元化趋势；同时，业界还有一种声音认为，市面上层出不穷的 AI 芯片带来的架构碎片化会打击开发者在实际软件产品中应用 AI 的积极性。架构多元化跟架构碎片化是同一回事吗？这给 one API 中跨平台统一编程带来了什么样的挑战？**

**Bill Savage：**是一回事。今天友商的 GPU 要求你选用某种特定的、只针对这个 GPU 的专用语言；对 FPGA 的编程也要求采取专用语言；针对人工智能的开发任务，神经网络处理器也要求不同的抽象。最终多种架构都聚集到了数据中心这里，但是每一种架构都有不同的编程要求。**oneAPI 的目的就是要解决这样的问题，使得开发者可以重复利用他在软件上已经做了的投入**。

**InfoQ：如果说出现了一个新的硬件产品（比如下一代 GPU），对于使用 AI 框架的开发者来说是否不需要再针对新的硬件做任何改动了？**

**Bill Savage：**如果开发者本身不直接使用 API 的话，那他不需要做任何事情。但是硬件厂商需要做一些工作来支持 oneAPI。

**InfoQ：那么英特尔打算如何推动不同的硬件厂商来支持 oneAPI？**

**Bill Savage：**我们有三个战术。第一是他们自己找上门来，看到我们提供开源的架构，自然会有人找过来；第二是我们会邀请一些合作伙伴参与进来；第三是由其他开发者来帮助某些硬件厂商实现对 oneAPI 的支持。

**InfoQ：您提到会邀请其他厂商参与到 oneAPI 这个项目中，英特尔会怎么去吸引他们参与其中？**

**Bill Savage：**一个伙伴要参与的话，肯定在业务上是要有一个理由的。**假设他在某些软件采纳上出现了障碍，采用 oneAPI 就可以解决他在软件方面的障碍，从而使他能够专注于硬件的解决方案。**

**InfoQ：相比英伟达的 CUDA，英特尔的 oneAPI 平台有什么独特优势？**

**Bill Savage：oneAPI 是跨不同架构、跨不同厂商的。**英伟达的 CUDA 是封闭的，只支持单一厂商。

**InfoQ：前一段时间**[英伟达开放了一部分GPU硬件接口文档](https://www.infoq.cn/article/1Bp58N-Tkmby4T4O6O8A)**，这会对其封闭的生态带来一些不同吗？**

**Bill Savage：**他们只是开放了很小的一部分东西，他们的语言依然没有开放。

## 如何实现 CPU、GPU、FPGA 整体性能最优？

**InfoQ：如果把 oneAPI 应用在 GPU、FPGA 等其他非英特尔 CPU 硬件上面，oneAPI 如何保证像在 CPU 一样达到最佳性能？**

**Bill Savage：**性能当然是一个挑战。在库这方面不会有什么问题，因为统一接口之后，你可以针对不同的硬件架构使用相同的代码，它能够保证性能。**但难在语言上，这也是英特尔现在投入了最大努力的一个方面，我们基于过去的工作和积累，另外也借鉴观察到的友商实践的方式，并在我们内部做了验证，最终才能够在语言上，通过 oneAPI 统一接口实现毫不妥协的性能。**

但是有一点要声明，虽然源语言可以分享，但是如果不经过对硬件的调优是很难实现最佳性能表现的。所以开发者要基于自己感兴趣的硬件进行调优，现在针对 CPU 他们也都是这样做的。比如至强要想实现最好的性能，虽然可以用以往的源代码，但是依然要进行调优。

还有一点要补充的是，我们并不是说未来有了 oneAPI 之后，就能期待 FPGA 编程起来像 CPU、GPU 那样容易，但是我们有专门提供给 FPGA 的 Extensions，可以基于它使用 DPC++进行编程。所以我们对于 FPGA 的预期是，开发人员通过使用我们的解决方案可以极大地提升基于 FPGA 编程的生产力，但是相对 CPU 和 GPU 来讲还是需要做更多的工作。

**InfoQ：实际使用中 oneAPI 可能一次计算要用到 CPU、GPU、FPGA 等不同硬件上，oneAPI 回怎么规划算力分配？是 oneAPI 自动调整还是由开发者指定？**

**Bill Savage：**我们会有另一个专门的工具来支持 oneAPI，我们称之为 offload advisor。这个工具会研究程序当中的计算及数据，然后判断什么情况下转移到哪个加速器是最为划算的。它能够实现一个完美的平衡，帮助开发者做 GPU、CPU 的分配，然后实现应用的最佳性能。英特尔有 CPU 也有 GPU，所以我们愿意帮助开发者用最聪明的方式使用 CPU、GPU，使之达到最好的状态。

## 围绕 oneAPI 的开源生态

**InfoQ：英特尔的 OpenVINO 深度学习视觉应用软件工具包也是横跨英特尔的不同架构硬件平台，并且已经应用到实际案例当中，是否对 One API 的开发有可以借鉴的经验？**

**Bill Savage：**是的，现在我们已经从 OpenVINO 向 oneAPI 进行相应的库的借鉴，OpenVINO 在底层接口上就支持 oneAPI，而且是在这些库还没有发布之前。虽然说现在 OpenVINO 已经在底层应用 oneAPI，但是 OpenVINO 的用户不需要进行任何改变。

**InfoQ：OpenVINO 和 oneAPI 之间有什么关联，它们是两个分开的工具吗？**

**Bill Savage：**OpenVINO 在 oneAPI 上面一层，我们把它称作 oneAPI 生态系统当中的一部分，但是并不是 oneAPI 规范的一部分。

**InfoQ：英特尔在开源软件方面有长期的投入，也已经有了很多项目。未来 one API 项目如何跟英特尔其他的开源软件项目配合？**

**Bill Savage：**英特尔一直非常大投入在开源软件项目，尤其是操作系统以及系统方面相关的软件开源。最近这几年在我们的开发者软件方面也是非常普遍的。现在英特尔的编译器就是基于 LLVM 的，它也是一个开源项目。我们很多库，包括 MKL-DNN、OpenVINO 都是开源的。总的来讲，oneAPI 也是开源的。其中可能有一些特定的组件未必是开源的，这些领域是因为商业原因所以无法开源。

**InfoQ：对于这部分不开源的组件，只有商业上的合作才能使用吗？**

**Bill Savage：**我们对于不开源的部分，除了付费的选择之外，也会提供免费版本的选择。比如说对于想要持续地拥有企业级支持的那些客户，他们会使用付费版本。

**InfoQ：目前来看 oneAPI 项目是不是在对标英伟达？未来英特尔对 oneAPI 有多高的期待？**

**Bill Savage：**oneAPI 开放给所有硬件厂商，所有硬件厂商都可以基于 oneAPI 开发。oneAPI 也对英伟达开放并且可以提供的。我预期 oneAPI 将会被得到实施和广泛部署，这是我对未来的期待。

对于第二个问题，我对于 oneAPI 的未来预期是非常高的，因为行业要求这样一种开放的、性能卓越的，对于现有解决方案的一种替代选择。英特尔和竞争对手不同，我们致力于向所有的多样的架构提供支持，而我们的竞争对手只对一到两个专用架构感兴趣。我们希望所有架构被支持，并且所有架构能够很好地协作。

**InfoQ：英特尔打算怎么打造 oneAPI 的开发者社区？**

**Bill Savage：**除了我们的合作伙伴项目之外，我们还有大量应用工程师，他们非常密切地和第三方开发者配合，这些第三方开发者和我们的“Intel Developer Zone”将成为 oneAPI 生态系统的一部分。

**采访嘉宾介绍：**

**William (Bill) Savage，**英特尔架构、图形与软件部副总裁兼计算性能与开发者产品部门总经理。

**Alice Chan，**英特尔架构、图形与软件部副总裁兼编译器与语言部门总经理。

2019 年 8 月 30 日 18:547127