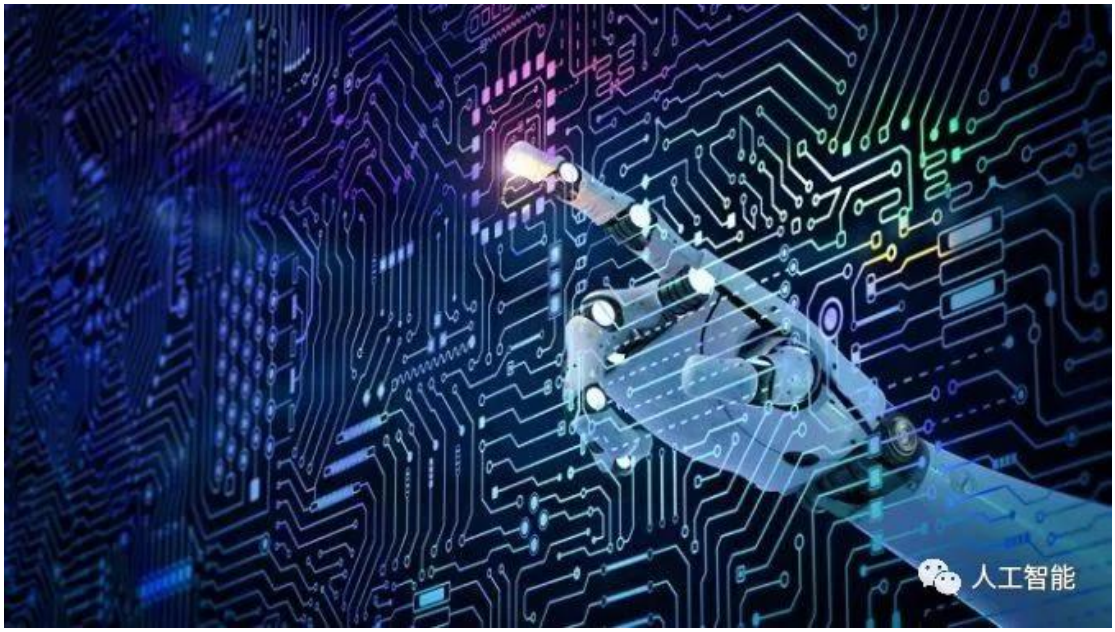
芯片行业30年资深人士：AI为何是高性能计算史上"最大的变革推动者"

人工智能3月29日

人工智能正变得无处不在，全球最快的计算机上也在运行人工智能负载，这也在改变HPC（高性能计算，High Performance Computing）。不过，人工智能将如何影响编程，软硬件以及和训练需求？

本文作者认为，AI可能是HPC历史上最大的变革推动者，至于为什么，他给出了AI在2019年对HPC产生最大影响的十大原因。



10、 Tensors（张量）：人工智能计算的通用语

向量代数的使用催生了为矢量计算设计的计算机。来自 Cray 的早期超级计算机是矢量超级计算机，它带动了应用程序以矢量和矩阵代数问题的方式表示，这反过来又推动了计算机的设计，确保矢量计算能更快运行。多年来，这种循环定义了 HPC。

张量代数可以视为广义矩阵代数，因此它是超级计算机能力的自然演化，而不是一场革命。任何支持矩阵运算的机器都可以进行张量运算。今天，CPU 通过通用编译器，加速 Pythons，增强库和优化框架的支持就可以支持矢量和张量的高性能计算。

正如向量之前对 HPC 的硬件、软件以及想法的影响，张量也正在深刻的改变着我们。

9、语言：高级编程语言

Fortran 编程语言在 HPC 领域占据主导地位 ,再加上 C 和 C ++语言几乎统治了 HPC 市场。通常通过 C 语言接口来扩展来支持加速器。尝试使用新语言来打破现有的格局已经失败，因为现有语言已经形成了一个生态，包括 HPC 的应用程序、用户、代码等。

AI 带来了新的需求，这将扩展与 HPC 相关的语言。他们不会改变使用 Fortran 的大多数物理学家的活动，但使用 MATLAB 和 Python 的数据科学家需要根据他们的需求量身定制解决方案。

Python 以及其它框架和编程语言，似乎正成为 HPC 越来越重要的部分。不过他们实际运行的程序仍将用 C/C++/Fortran 编写，但 AI 程序员既不会知道，也不关心它。

8、以不同方式思考：通过重新思考的方法来替换遗留代码

HPC 非常传统，相对而言人工智能是新的。就目前而言，当两者相互作用时，它将重提有关实现遗留代码的问题，在某些情况下这些代码可能早就该实现了。说法可能是“让我们为这段代码添加一些人工智能功能”，但现实将是努力可能成为浪费时间。还记得 Java 热潮的早期许多“转换为 Java”的努力吗？

就像那些早期疯狂的 Java 时代一样，急于将代码重写为新形式的人既有成功的也有失败的。投资回报率（ROI）将是关键，但预测创新的结果往往是错误的。

7、可移植性和安全性：虚拟化和容器

安全性和可移植性的具体问题是，“我可以在我的机器上安全地运行吗？”和“它能在我的机器上运行吗？”，这是虚拟化和容器试图解决的问题。当然，安全性来自于良好的硬件和软件特性。对于许多人来说，虚拟化和容器似乎能确立这种组合。

容器已引起许多开发人员的关注，因为它们比虚拟机更灵活、可部署、可升级、具备云多功能性，并且可以节省虚拟机授权许可成本。

在任何 HPC 或 AI 的会议上谈论容器似乎只能站着说说。但这正在改变 ,例如 Python 和 Julia 在配置时可以更好地扩展，容器可以帮助部署。

容器为用户提供了良好的环境，2019 年将看到 HPC 领域越来越多的容器使用，部分原因是 AI 用户的对此表现出的兴趣。毫无疑问，这会对 HPC 带来挑战，因为这需要优化的生态系统。如今，这个领域正在进行这方面大量的精细工作，HPC 社区将帮助实现这一目标，满足大家对容器的渴望。

6、规模问题：大数据

只要有人工智能，就有大数据。人工智能的重点是利用数据模型从大量的数据集中找到价值。许多 HPC 中心已经有很多基础设施可以很好地处理大数据问题。

所有 HPC 中心都将大数据作为新系统的主要需求，AI 工作负载是大数据需求的主要动力。

由于存储器的高成本，我们看到存储器大小与 FLOP/s 的比率多年来一直在下降。这对大数据发展不利。与持久内存相关的新功能带来了一些希望，并支持大型机器（包括 HPC）中的大数据模型。这些新的内存技术提供了主内存和本地存储（SSD）的扩展。

我今天写的是人工智能如何影响 HPC，但我还得指出 HPC 对可视化的热爱将对 AI 产生的影响。将数据放在最接近处理器的位置是最适合进行实际数据可视化的处理

器，是 HPC 影响 AI / ML 的最重要的方法之一。当然，使用和理解大数据以及可视化数据和分析是相互交织的。

5、大量计算：云计算

人工智能开发人员可能已经比 HPC 开发人员更多地接受了云计算。虽然 HPC“在云中”已经出现，但 AI 应用的高性能计算需求将加速“云中的 HPC”。

4、硬件：交互式能力，为库和框架提供性能

人工智能的计算量并不大。这意味着少数库接口和框架主宰着“AI 加速器”作为其卖点。

交互能力是一个长期存在的要求，它一直被 HPC 系统“搁置”，现在被 AI 程序员将其放置在“前端和中心”。这种变化对“HPC”的改变速度还有待观察，但 2019 年该领域的创新即使分散且有些隐秘也会引人注目。交互性也可称为“个性化”。

HPC 更多的硬件多样性、交互性支持以及为性能优化的附加库/框架抽象，以支持 AI 工作负载。HPC 社区对性能的关注将有助于说明基础设施的更多融合将有利于数据中心部署。没有人愿意放弃性能，只要他们不必这样做，HPC 社区的专业知识将有助于商业化 AI / ML 的性能，从而带动社区之间更多的硬件技术融合。

3、人员融合：用户多样性和对 HPC 兴趣的增加

AI 将吸引许多具有不同背景的新人才。AI 将以前所未有的规模为 HPC 带来民主化。过去几年，“HPC 民主化”用于描述 HPC（以前只有大型组织的人才可以使用）如何被小的工程师团体和科学家群体使用。数学和物理问题可能推动了早期的超级计算发展，但最近更多的用户发现 HPC 性能在医学、天气预报和风险管理等领域不可或缺。

AI 带来了比 HPC 更广泛的用户群，为 HPC 的民主化带来了全新的应用。将 AI 增加到发展 HPC 的列表中，我们继续为追求世界上最高性能的计算添加更多理由， HPC 专家和 AI 专家正在结合，以产生我们都能感受到的兴奋。

2、新投资：推理

机器学习通常可以被认为是由“训练”的学习阶段和 “推理”的“做”阶段组成。看起来我们需要更多的循环进行推理而不是更多循环进行训练，特别是当我们看到机器学习无处不在地嵌入到身边的解决方案中时。市场分析师估计，推理硬件市场是训练硬件规模的 5-10 倍。

有了这么大的市场机会，毫不奇怪，所有人都希望进入市场更大的推理市场。推理已在 FPGA，GPU，DSP 和众多定制 ASIC 处理器上运行。功耗，延迟和总体成本都是卖点。高性能、低延迟、易于重新编程的 FPGA 似乎是补充当前 CPU 主导的推理市场的合理选择，时间会证明。

跟着市场的选择，您将看到推理工作负载将对包括 HPC 在内的所有计算产生重大影响。

1、应用程序的融合：不是在“重新思考”之后进行替换 ，“融合”两全其美，扩展工作负载多样性并看到不同工作负载的融合

那些有远见的人已经证明，HPC 和 AI 结合时有很多机会。鼓舞人心的研究范围从拥有一个中立的网络学习到“ 像蒙特卡罗模拟一样 ”，具有非常好的结果，只需要一小部分计算需求；将系统整合到能够预测极端天气的模式，如飓风，或天气预报系统。生成对抗网络（GAN）是一类机器学习系统，许多人都非常重视，GAN 无疑有助于融合 HPC 和 AI / ML。

虽然现在很少有应用结合 HPC 算法和 AI 技术，基于早期的结果，我很容易预测这是 HPC 应用的未来，并且将因为 AI 带来 HPC 最大的变化。

理解这十种力量

计算在某种意义上并没有改变：它完全取决于整个系统对用户的作用。虽然需求有变化，但一个完整的系统由硬件起来和软件组成不会改变。实际上，很容易被单一技术（硬件或软件）分散注意力；最好的系统会谨慎地应用最新技术，我非常偏爱地称其为“选择性加速” ，强调在重要时使用加速。当我经常使用 Python 时，我喜欢 Python

加速（一种依赖 CPU 的软件技术）。当我需要低延迟推理时，我喜欢 FPGA 加速。当我只需要一点加速时，我不使用任何一个。这是建立平衡系统的艺术。这前十的名单并没有打破为多用途机器提供最佳整体效果现实的平衡。

结论：AI 将使用 HPC，这将永远改变 HPC

显然 AI 将使用 HPC，这将永远改变 HPC。事实上，AI 可能是 HPC 历史上最大的变革推动者。HPC 随着科技的发展不断进步，工作负载也将随着人工智能的发展而变化。我不认为辩论收敛与交叉给予足够的信任的概念，人工智能用户将加入 HPC 社区，并留下自己的标记。他们也将使用非 HPC 系统，就像其他 HPC 用户一样。

将有专为 AI 工作负载设计和构建的定制高性能机器，其他机器的 AI 工作负载也在可以在更通用的高性能设备上运行。要平衡机器的高性能和灵活才能实现加速。在所有情况下，人工智能将有助于定义未来什么是超级计算，这将永远改变 HPC。

James Reinders 是 HPC 爱好者，也是拥有 8 本书的超过 30 年行业经验的从业者，其中包括在英特尔工作 27 年经验（2016 年 6 月退休）。