XILINX®

自适应计算

技术概览

作者：Greg Martin
赛灵思公司战略市场营销总监

内容

执行摘要

自适应计算构建在现有 FPGA 技术之上，但它比以往任何时候都更易于为更加广泛的开发者和应用所使用。这项技术的基本功能基于自适应硬件，这种硬件具备在制造后进行修改的独特能力。每个硬件块阵列都可以配置并按需进行连接，从而可为任意应用构建高效的特定领域架构。

这种灵活应变能力是有异于 CPU、GPU 和 ASSP 的独特差异化因素，因为它们都采用固定硬件架构。灵活应变能力的益处因应用而异，但与相同功能的 CPU 实现相比，性能提升 20 倍的情况并不鲜见。

自适应计算可以为各种硬件、软件和 AI 开发者所使用。自适应平台则涵盖了综合全面的开发工具在内的自适应平台，为高效开发差异化的最终产品奠定了坚实基础。在量产系统中部署自适应平台的效益包括缩短上市时间、降低运营成本，以及系统与生俱来的面向未来的能力。

从数据中心到网络，再到边缘乃至终端，自适应平台适用于各种类型的最终应用。从自动驾驶汽车到火星探测车，自适应计算正赋能新一代智能、高效应用。

自适应计算简介

1984 年，Ross Freeman 将其关于现场可编程门阵列 (FPGA) 的绝妙构想付诸实践并创立了赛灵思公司，从而确立了自适应计算的原理。从那时起，这项技术及其满足众多应用需求的能力，已历经漫长的发展道路。尽管自适应计算构建在 FPGA 技术的基础上，但它已经发展至足以覆盖更为广阔的应用类型。

自适应计算的核心由能够针对特定应用进行高度优化的芯片硬件构成。该优化发生在硬件制造完成后，而且能以几乎无限的次数反复进行。这种独特的灵活性，使得硬件可以在器件完全部署到量产设置后进行更改。正如量产 CPU 可以运行新程序一样，自适应平台也能灵活适应给定的全新硬件配置，即使处于实时量产设置中。

在部署后改变硬件功能的能力就是 FPGA 的“FP”或“现场可编程”，它意味着硬件可以在部署到量产环境后于现场进行编程。FPGA 中的“GA”指“门阵列”。自门阵列诞生后，自适应计算平台经历了漫长的发展，但其概念依然是解释底层技术运作方式的有效途径。我们将在下一节中进一步探讨这个话题。

自适应硬件

自适应硬件允许在制造后对硬件的底层功能进行配置。之所以能做到这一点，源于自适应硬件的两个独特功能。

第一个独特功能是可配置硬件块的规则结构。在 FPGA 中，这些被称为“可配置逻辑块”，并且每个块都能进行配置，以执行众多可能的算术函数之一，这些算术函数对多个输入开展运算并产生输出。在较新的自适应计算平台上，这些块的复杂性显著提高，构成了矢量处理器等高度复杂的可配置功能。现在我们将重点了解之前的版本，较新的版本将在后续章节中介绍。

第二个独特功能是块与块之间的可配置连接。这种可配置互联使自适应硬件能够在块与块之间按需进行连接，从而通过连接特定的块构建更加复杂的功能。此外，可配置互联还允许终端系统可能需要的非可配置块之间进行连接，例如存储器、嵌入式 CPU、数字信号处理器 (DSP) 和硬件器件的输入与输出（ I/O 或“引脚”）。

下图是未配置块的示意图（左侧），以及为某个应用配置后的相同块（右侧）。在该示例中，已经实现了 A 和 B 两个独立的功能。每个功能都有级别，其中部分功能可以并行实现（AF5a 和 AF5b）。A 和 B 本身也完全并行。箭头显示了已被配置到自适应硬件中的连接，而白色文本则表示已完成唯一配置以实现特定功能的块。本示例中有四个块未进行配置（仍为灰色），但如果未来需要进行更改，则仍可以使用。

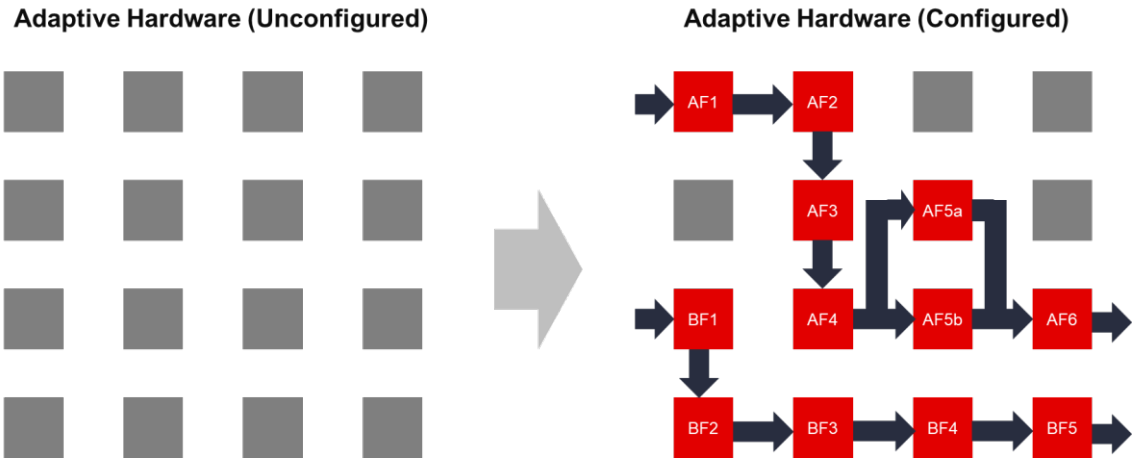

图 1：未配置和已配置的自适应硬件示意图

自适应硬件与 CPU 对比

尽管 CPU 高度灵活，但其底层硬件是固定的。一旦 CPU 制造完成，硬件便不能更改。其依赖于软件告知需要执行哪些特定运算（算术函数），以及需要在存储器中调用哪些数据。CPU 每次从固定数量的预定运算中选择。硬件必须能够执行所有可能的运算，称之为使用软件指令，不过 CPU 的原子处理组件（称为算术逻辑单元，或 ALU）一次通常只能执行一条指令。现代 CPU 可以有 32 个或更多数量的 ALU，但从根本上说，每个 ALU 都要以顺序方式依次执行运算——一次一条指令。

如前文所述，编写软件是为了指示 CPU 需要执行的确切运算、从何处获取进行运算的数据，以及在何处存储结果数据。CPU 仍然在基本的冯·诺依曼架构（或者更确切地说是存储程序计算机）上运行。在这种架构中，数据从存储器读取到处理器，进行运算，然后写回到存储器。从根本上讲，该架构以 ALU 为中心，并要求每次运算过程中将数据移入和移出 ALU。

CPU 能够提供极大的灵活性，因为它们是软件可编程的。然而，无论 CPU 架构有多么先进，仍有部分应用无法有效地适应冯·诺依曼架构。

例如，道路收费站上用于监测车牌的智能视觉应用。它必须接收视频数据流、解码数据（预处理），然后将其传递给识别车牌字符的 AI 推断模型。它将字符信息馈送到后处理块，后处理块将字符信息进行组合，并通过网络向集中式数据中心报告车牌号码，以执行计费流程。在这样的流式数据应用中，冯·诺依曼架构远非最佳方案。

相比于将数据移动到处理器，更高效的做法是构建一条处理流水线，在数据流经上文描述的处理阶段时处理数据。这种架构可以先接收视频数据流，接着执行预处理，从而为 AI 模型做好准备。然后数据直接流入 AI 模型，完成车牌识别。在对之前一帧开展 AI 处理的同时，下一帧正在流经预处理器。

使用 CPU 实现相同的应用，需要首先处理数据，即从存储器读取数据、处理，接着将数据写回到存储器。然后必须读入新数据以执行 AI 功能。这与流式架构相比效率较低，因为它必须持续向处理器传递数据，而不是构建处理流水线，直接在流数据上运行。

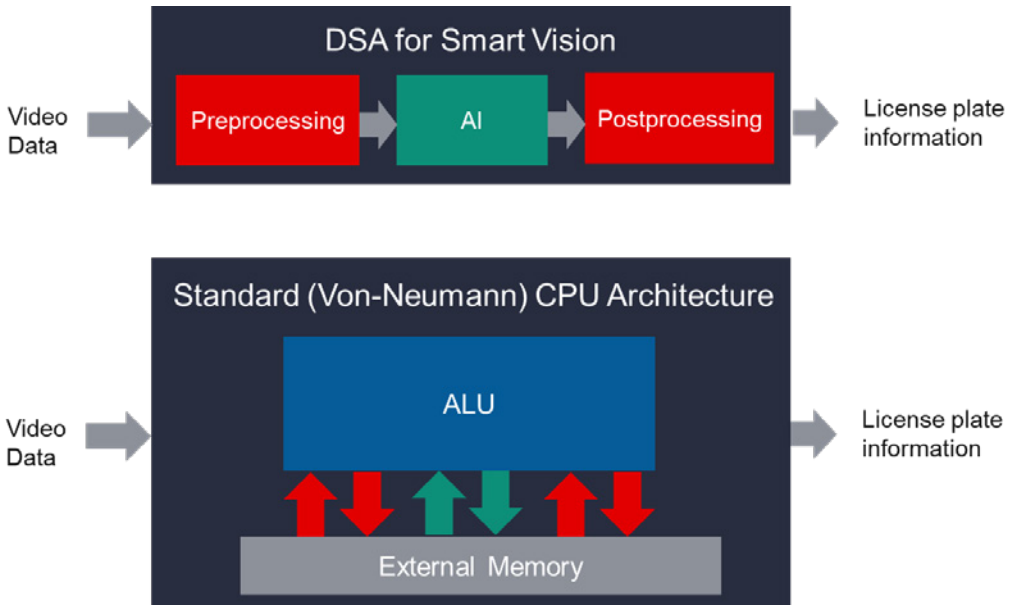

图 2：数据流架构与冯·诺依曼架构对比

左侧的架构是特定领域架构 (DSA) 的示例——本示例中显示的是为智能视觉领域构建的专用架构。DSA 将计算、存储器带宽、数据路径和 I/O 与领域的具体要求紧密匹配。对于某些领域，这种方法与通用 CPU 架构相比，能大幅提升处理效率水平。

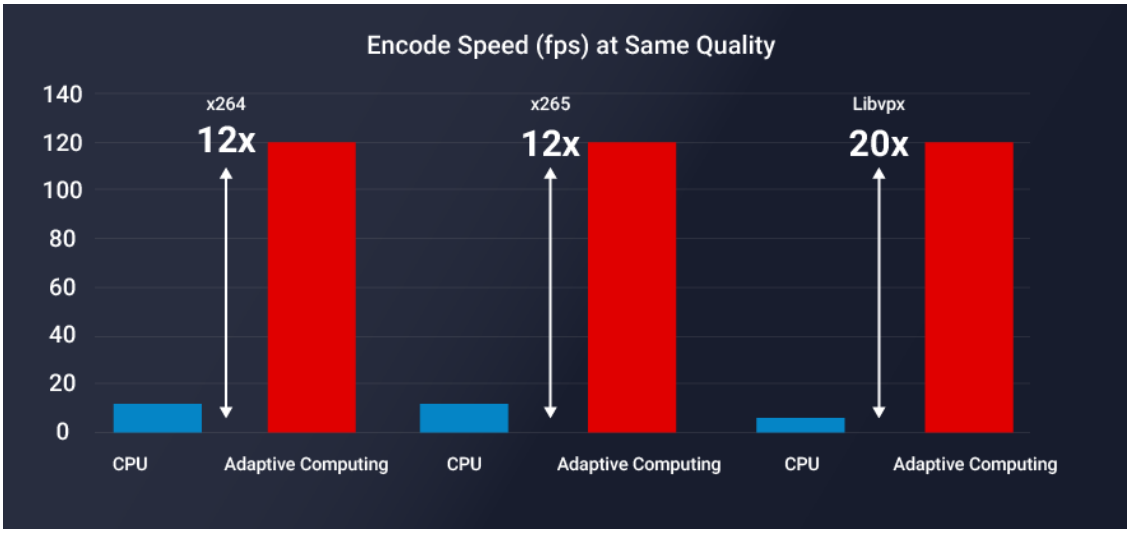

图 3：视频编解码效率 - 自适应计算与通用 CPU 对比（来源：赛灵思）

尽管存在这些局限性，但大量应用仍然可以从部分使用 CPU 的实现方案中受益匪浅。大多数复杂应用都可以受益于异构硬件阵列，每种类型的底层硬件为应用的不同部分提供最佳实现方案。

自适应硬件与 GPU 对比

通用 GPU (GPGPU)，我们简称为 GPU，克服了 CPU 的主要不足，能够并行处理大量数据。

从根本上来讲 GPU 与 CPU 类似，因为它们有固定的硬件，并使用软件指令运行。它们与 CPU 的不同之处在于能在非常宽泛的数据集上运行。一条指令就能处理上千条或者更多数量的数据，尽管通常必须对同时处理的每一条数据进行相同的运算。现代 GPU 的架构极其复杂，不过并行数据运算是与 CPU 的根本区别。

尽管存在这种区别，GPU 的核心仍然包含某种类型的冯·诺依曼处理器。原子处理元在数据矢量上运行，但仍然是每个 ALU 执行一条固定指令。数据还必须通过固定路径从存储器传递给这些处理单元。

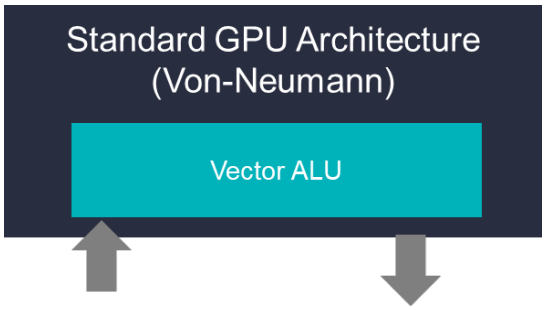

图 4：从根本上来讲，GPU 仍然采用冯·诺依曼架构

类似于 CPU，GPU 也极为适用于特定类型的应用。然而，GPU 与 CPU 一样，仍然采用固定硬件构建——基本架构和数据流在制造前就已固定。它不能适应特定应用，而且一旦构建完成就肯定无法更改。GPU 不具备实现特定领域架构的能力，因而也不能在特定领域内以最佳方式实现特定应用。

自适应硬件与 ASSP 对比

DSA 可使用专用（固定）芯片器件构建，通常称之为特定应用标准产品或 ASSP。这种说法属于过度简化，不过总体上 ASSP 为给定应用或者给定领域实现经优化的硬件，因此它们一般可以被视为 DSA。

在固定的 ASSP 中实现 DSA 既有优势，也有劣势。我们介绍两种主要的劣势。

首先是创新步伐。为了跟上创新步伐，制造商被期望在比以往更短的时间内打造和提供新的服务。更具体来讲，这个时间短于设计开发新的固定芯片 DSA 所需的时间。这就造成了市场的创新需求与企业设计制造 ASSP 所需时间之间的根本性市场错位。行业标准变或其他需求波动会很快导致这些器件过时。

第二个考量因素是定制芯片的成本。设计与制造独特的芯片设计（如复杂的 7nm ASIC）的一次性成本可能导致数亿美元的非重复性工程 (NRE) 成本。随着器件工艺缩小到 5nm 及更小，预计成本还将进一步上升。在您购买 ASSP 时，前期 NRE 会分摊在每个销售的器件上。ASSP 只有销量很大时才具有成本效益。成本上升正在延缓 ASSP 采用先进节点制造的进程，导致其用户固守过时低效的技术。

自适应平台

自适应平台一词指以自适应硬件为核心的任意类型产品或解决方案。从应用开发的角度来看，平台的作用是为构建产品提供一套功能。它为开发者构建应用奠定了坚实的基础。通过为应用奠定基础，平台让开发者能够专注于应用的具体差异化。

自适应平台完全基于相同的自适应硬件基础。但它们包含的内容远不止芯片硬件或器件。自适应平台涵盖了综合全面的成套设计和运行时软件。通过结合软硬件提供了开发高灵活性、高效率应用的独特功能。

自适应平台的优势

自适应平台使得自适应计算能够为广泛的软件和系统开发者所使用。这些平台也能为众多产品奠定基础。采用自适应平台的优势包括：

加快上市进程。使用 Alveo™ 数据中心加速器卡这样的平台构建的应用，能为特定应用利用加速硬件，但不需要硬件定制。将 PCIe 卡添加到服务器，就可以从现有软件应用直接调用加速库。

降低运营成本。与基于 CPU 的解决方案相比，由于计算密度的提升，基于自适应平台的优化应用能在每节点提供大幅提高的效率。

灵活且动态的工作负载。自适应平台可根据当前需求重新配置。开发者可以在自适应平台内轻松切换已部署应用，使用相同设备即可满足不断变化的工作负载需求。

兼容未来。自适应平台能不断进行调整。如果现有应用需要新的功能，则可以对硬件重新编程，以最佳方式实现这些功能，减少硬件升级需求，进而延长系统使用寿命。

加速整体应用。AI 推断很少单独存在。它是更大的数据分析与处理链条的组成部分，往往与使用传统（非 AI）实现方案的多个上游级和下游级并存。这些系统中的嵌入式 AI 部分得益于 AI 加速，而非 AI 部分也能从加速中获益。自适应计算的天然灵活性适合为 AI 和非 AI 处理任务进行加速，称之为“整体应用加速”。随着计算密集型 AI 推断渗透到更多应用中，其重要性也在日益提升。

自适应平台的适用性

在过去，运用 FPGA 技术需要开发者构建自己的硬件板，并用硬件描述语言 (HDL) 配置 FPGA。相比之下，自适应平台允许开发者使用自己熟悉的软件框架和语言（例如C++、Python、TensorFlow 等），直接发挥自适应计算的效能。软件和 AI 开发者现在无需构建电路板或成为硬件专家，就能运用自适应计算。

例如，视频应用开发者可以通过 FFmpeg 等行业标准框架，直接使用自适应计算。他们无需成为 FPGA 专家或硬件实现视频编解码器专家，同样能开发出最为先进、高效的应用。此外，如果他们拥有现成的软件代码，则可以通过直接的 API 调用更新，将其进行移植以使用自适应计算。

自适应平台使得自适应硬件能在多个设计抽象层面使用。独立软件供应商 (ISV) 生态系统和厂商提供的开源库已提供了大量加速 API。此外，较大的设计团队也有自己的硬件工程师负责打造定制加速 API，供其软件团队在最终应用中使用。

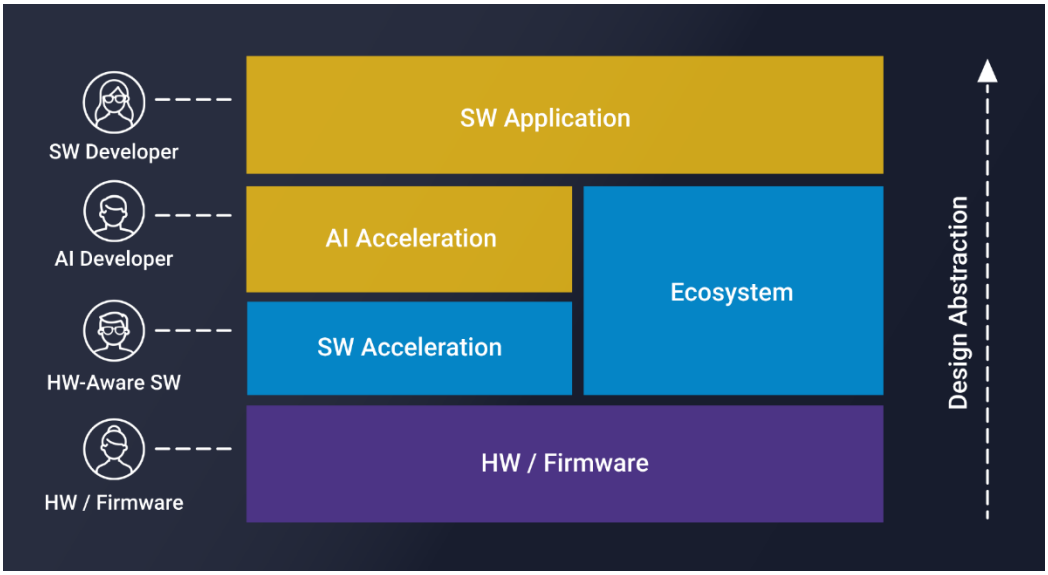

图 5：自适应开发堆栈示例（来源：赛灵思）

自适应平台类型

根据应用和需求，存在多种类型的自适应平台，包括数据中心加速器卡和标准化边缘模块。大量平台的存在旨在为开发所需应用提供尽可能最佳的起点。而应用类型则十分广泛，既有自动驾驶和实时视频流等时延敏感型应用，也有高度复杂的 5G 信号处理和非结构化数据库的数据处理。

自适应计算能够部署到云端、网络、边缘甚至终端，将最新的架构创新带到单独及端到端的应用。鉴于存在各种自适应平台，部署位置也可以是多样化的——从数据中心内 PCIe 加速器卡上的大容量器件，到适用于物联网设备所需终端处理的小型低功耗器件。

边缘端的自适应平台包括赛灵思最新 Kria™ 自适应系统模块 (SOM)。Kria 自适应 SOM 围绕赛灵思 Zynq® UltraScale+™ MPSoC 架构而构建，使得开发者能够在交钥匙自适应平台上开发边缘应用。通过将系统的核心部分标准化，开发者有更多时间专注于打造使其技术差异于竞争对手的功能特性。

图 6：Kria 自适应 SOM

数据中心中的自适应平台包括 Alveo 加速器卡。Alveo 加速器卡采用行业标准的 PCI-express，为任意数据中心应用提供了硬件卸载能力。自适应计算并不只是在数据中心内卸载计算，其还可用于 SmartSSD 存储，在存储访问点上进行加速；此外还能用于 SmartNIC，直接在网络流量上提供加速。

图 7：Alveo PCIe 加速卡

自适应计算的近期发展

我们在前文中讨论了自适应硬件的两个独特特性，即可配置块和可配置互联。这些构成了自适应硬件与非自适应（固定）硬件的根本差异。

自适应计算的最大创新之一在于赛灵思推出的 AI 引擎。AI 引擎是一种革命性的新方法，为计算密集型应用提供了前所未有的计算密度。

AI 引擎从根本上讲是一个可配置块，但它也具有类似 CPU 的可编程性。AI 引擎并非由标准 FPGA 处理硬件构成，而是包含高性能标量和单指令多数据 (SIMD) 矢量处理器。这些处理器经过优化，能高效地实现 AI 推断和无线通信中常见的计算密集功能。

AI 引擎阵列的连接使用类似 FPGA 的灵活应变数据互联，为构建目标应用提供了高效、优化的数据路径。这种计算密集（数学丰富）、类似 CPU 的处理元与类似 FPGA 的互联相结合，正引领新一代 AI 和通信产品。

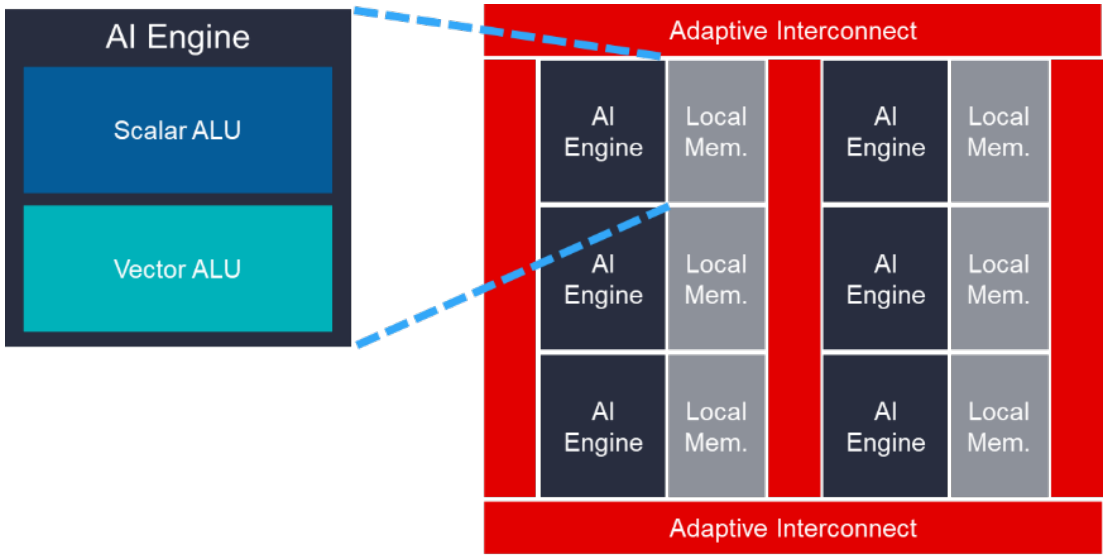

图 8：先进的自适应硬件 —— AI 引擎阵列

自适应计算在行动

正如我们已经看到的，自适应计算赋予应用动态更新的能力。它不仅支持软件无线 (OTA) 更新，也同时支持硬件的无线更新。随着处理变得更加分散，并且应用部署在难以抵达的位置，这一功能显得尤为重要。在这一点上，没有比火星更难抵达的地方了。美国航空航天局的“毅力号”火星车现正探索火星表面，其采用了自适应计算技术。

“毅力号”将自适应计算用于其综合视觉处理器。它使用 FPGA 平台为基础构建，能够加速 AI 和非 AI 视觉任务，包括图像校正、过滤、检测以及匹配。“毅力号”发回给美国航空航天局的图像都已经自适应计算处理。

如果在“毅力号”前往火星的八个月内开发出了新算法，或者发现了硬件缺陷，那么自适应计算会允许通过无线或空间远程发送硬件更新。这些更新如同软件更新一样快速简便。在进行远程部署时，这种远程硬件更新不止是便利，更是必需。

图 9：自适应计算第三次抵达火星（图片：美国航空航天局）

结论

自适应计算构建在现有 FPGA 技术之上，但它比以往任何时候都更易于为更加广泛的开发者和应用所采用。对于软件和 AI 开发者而言，现在能够运用自适应计算（一种过去他们无法使用的技术）构建优化型应用。

硬件能够根据特定应用进行配置，是与 CPU、GPU 和 ASSP 相比的独特差异，因为它们均采用固定硬件架构。自适应计算允许硬件为特定应用进行定制，从而实现了高效率，而如果工作负载或标准发生演进，还可以实现未来适应。

自适应平台是数据中心、网络、边缘和终端中各种最终应用的理想选择。自适应平台可以加快上市进程，并且交付高效、可升级的解决方案。

自适应计算作为一种先进技术，使得无需硬件专业知识就能打造优化的硬件加速应用。自适应计算现已广泛应用于众多行业，并已在全球乃至太空的大量应用中证明了自身价值。

随着世界进一步迈向互联互通和万物智能，自适应计算将继续居于优化、加速应用的最前沿，从而赋能全体开发者创造更加美好的未来。

进一步了解自适应计算：china.xilinx.com/adaptivecomputing

修订版 1.0，2021 年 6 月

公司总部
Xilinx, Inc.
2100 Logic Drive
San Jose, CA 95124
USA
Tel: 408-559-7778
www.xilinx.com

赛灵思欧洲
Bianconi Avenue
Citywest Business Campus
Saggart, County Dublin
Ireland
Tel: +353-1-464-0311
www.xilinx.com

日本
Xilinx K.K.
Art Village Osaki Central Tower 4F
1-2-2 Osaki, Shinagawa-ku
Tokyo 141-0032 Japan
Tel: +81-3-6744-7777
japan.xilinx.com

Asia Pacific Pte. Ltd.
赛灵思亚太地区
5 Changi Business Park
Singapore 486040
Tel: +65-6407-3000
www.xilinx.com

印度
Xilinx India Technology Services Pvt. Ltd.
Block A, B, C, 8th & 13th floors,
Meenakshi Tech Park, Survey No. 39
Gachibowli(V), Seri Lingampally (M),
Hyderabad -500 084
Tel: +91-40-6721-4747
www.xilinx.com

在美国印刷 WW6/8/21