FLEX LOGIX 推出最高性能、最高效率 AI 边缘推理芯片

InferX X1 芯片将 AI 引入更广阔市场空间；

以 1/7 面积和 10-100 倍性价比向 NVIDIA Xavier 发起挑战

美国加州山景城，2020 年 10 月 20 日—Flex Logix 公司今天宣布其 InferX X1 芯片可开始出货。该芯片是 AI 边缘系统领域迄今为止性能最高的芯片。InferX X1 可对目标检测与识别等各类神经网络模型进行加速，其应用范围包括机器人、工业自动化、医学成像、基因测序、银行安全、零售分析、自动驾驶、航天工程等等。与目前业内领先的 NVIDIA Jetson Xavier 相比，InferX X1 在处理 YOLOv3 目标检测识别模型时的性能提高了 30% 。在其他多个用户模型方面，InferX X1 的性能更是高达 NVIDIA Jetson Xavier 的 10 倍。

相比于其他各类目标检测与识别的神经网络模型，YOLOv3 的准确率是最高的。因此，许多机器人、银行安全及零售分析领域的客户都计划在产品中使用 YOLOv3。还有一些客户根据自己的应用需求，开发了一系列 AI 模型，希望可以在较低成本下获得更高的吞吐量。Flex Logix 对这些客户模型进行了基准测试。基准测试的结果显示，InferX X1 不仅可以完全满足用户在吞吐量方面的要求，并且价格更低。Flex Logix 将于近期开始向早期客户出货样品芯片，并计划于明年一季度向更广大客户群体提供样品芯片。批量生产的芯片将于 2021 年下半年开始全面出货。

InferX X1 芯片面积为 54mm^2， 仅为 1 美分硬币的 1/5 大小，更是远远小于 350mm^2 的 NVIDIA Jetson Xavier 芯片。InferX X1 芯片的批量价格仅有 NVIDIA Xavier NX 的约 1/10。这是高质量、高性能的 AI 推理产品第一次真正走向普罗大众。原本昂贵的 AI 推理将不再遥不可及！

InferX X1 配套的软件使其非常易于用户使用。其中的 InferX Compiler 可以将 TensorFlow Lite 或者 ONNX 的模型直接转换为可以在 InferX X1 上运行的程序。

“对于已有边缘推理系统的用户来说，他们需要更高性价比的 AI 推理解决方案。只有这样，他们才能真正将神经网络模型全面应用在其批量化产品中。InferX X1 恰好满足了这类用户的需求。相

比于目前行业的领军产品，InferX X1 可以为用户带来数十倍甚至上百倍的性价比提升。” Flex Logix 的创始人 CEO Geoff Tate 在受访时这样表示。

Gforce Marketing 公司的首席营销主任及 SemiWiki 杂志的撰稿人 Mike Gianfagna 先生认为：“Flex Logix 此次推出的技术将为整个 AI 市场带来革命性的变化。InferX X1 将显著拓宽现有的 AI 应用市场，将 AI 推理带入更为广阔的市场领域。其它推理解决方案在性价比方面无法与 Flex Logix 相匹敌。因此，这次发布的 InferX X1 将颠覆当前的 AI 市场格局。我们相信 Flex Logix 在未来会呈现指数级别的飞速成长。”

TIRIAS Research 的首席分析官 Kevin Krewell 表示：“TIRIAS Research 相信，在数据中心内部和以外的各类应用场景中，针对机器学习模型的大规模部署才刚刚开始。我们是否可以在广阔的工业领域中大规模布局 AI 推理应用的关键性因素就在于芯片解决方案的功耗和成本效率。我们相信，Flex Logix 推出的 InferX X1 加速器和 PCIe 板将带给我们更高的能效比和性价比。”

基于多项 Flex Logix 的专有技术，InferX X1 采用了一种全新的架构，可以在较小面积内实现较高的吞吐量。其中，Flex Logix 专利的 XFLX 可编程互联网络架构，也被应用于嵌入式 FPGA 技术，并在过去数年中被国内外多家知名公司所使用。其中包括 Dialog 半导体、波音、桑迪亚国家实验室、以及大唐电信旗下的辰芯科技。除了 XFLX 以外，InferX X1 还用到了可重配置张量处理器。它由 64 个一维的张量处理器构成，可通过重新配置来高效地支持各种神经网络模型的运算。由于重配置的时间只有几个微秒，所以神经网络模型的每一层都可以拥有最优化的数据路径。

InferX X1 的批量生产芯片和配套软件将于 2021 年第二季度开始全面出货。用户样品及早期软件工具则计划于 2021 年第一季度开始对用户进行供货。目前，Flex Logix 可以向符合以下条件的先期用户提供样品芯片和软件，并进行基准测试支持。这些条件包括：需要有现成的基于 TensorFlow Lite 或者 ONNX 的神经网络模型，并有可在 2021 年进行批量生产的产品项目。相关的具体技术指标及价格指引也都将对这些用户提前提供。

更多相关信息请访问 https://flex-logix.com。

具体技术指标

- 高达 70% MAC 利用率，可使较小面积和较低成本处理高清图像和较大模型。
- 一维张量处理器（1D TPU）即一维脉动阵列
    - 64B 输入张量

    - 64 INT8 MACs
    - 32 BF16 MACs
    - 64Bx256B 权重矩阵
    - 一维脉动阵列每 64 个时钟周期可完成 4096 次乘加运算。
- 每颗 X1 芯片中的可重配置张量处理器由 64 个一维张量处理器（1D TPU）组成
    - 可以将多个 TPU 配置成串联或者并联结构，以实现多种不同的张量运算。这种灵活性可以很有效地支持不断衍生的诸如 3D 卷积等新型运算，并保持较高性能。
    - 可编程互联网络架构可以很好地解决 SRAM 与 TPU 间的数据通路的竞争问题，达到非常高的数据交互速度。
- eFPGA 可编程逻辑可用于实现包括控制 TPU 运行的高性能状态机，以及各种运算符的控制逻辑
- 神经网络模型中的每一层都可被专门进行重配置；每一次重配置只需要几微秒的时间。
- 在处理当前层级的同时，下一层神经网络模型的配置及权重可在后台从 DRAM 中被加载；这可以极大减少由 DRAM 带宽限制所带来的计算的停顿。
- Layer fusion 功能可通过将一个以上的配置文件进行合并来降低 DRAM 延时。
- 仅使用最少的内存资源以降低成本：LPDDR4x DRAM, 总共 14MB SRAM
- x4 PCIe Gen 3 or Gen 4 可提供芯片与主机间的高速通信。
- 在 16nm 制程下芯片面积为 54 mm^2
- 倒装 BGA 封装尺寸为 21 x 21 mm

产品类别及价格

InferX X1 将于近期开始像部分用户进行样品芯片的出货，并计划于 2021 年第二季度开始进行批量产品的出货。

商业报价（1000-100 万颗）

速度	价格
933 MHz	$199-$69
800 MHz	$149-$49
667 MHz	$124-$39
533 MHz	$99-$34

满足工业级或航空级温度要求的 X1 芯片也可向用户出货。

更多价格信息，请访问 *https://flex-logix.com* 或发邮件至 *info@flex-logix.com*。

关于 Flex Logix

Flex Logix 提供业内领先的 eFPGA 及神经网络推理加速解决方案。其研发的 InferX X1 芯片是速度最快的 AI 推理芯片。在边缘系统和边缘服务器领域，InferX X1 与现有的行业领军产品相比有着 10-100 倍的性价比提升。Flex Logix 的 eFPGA 系列产品可以为芯片带来灵活性，使得芯片可以随着行业的协议、标准、算法，以及客户需求的变化不断进行更新和升级。对于一些应用来说，在 eFPGA 上实现比用处理器实现可以提高 30-100 倍的性能。Flex Logix 公司的总部设于加州山景城。更多信息，请访问 *https://flex-logix.com* 。

####

MEDIA CONTACTS
Kelly Karr
Tanis Communications
kelly.karr@taniscomm.com
+408-718-9350

Copyright 2020. All rights reserved. Flex Logix is a registered trademark and InferX and nnMAX are trademarks of Flex Logix, Inc.