# 为快速增长的网络边缘人工智能应用提供更高性能的解决方案

莱迪思半导体白皮书

**2019年8月**

存在检测和对象计数等网络边缘人工智能应用越来越受欢迎，但设计人员越来越多地要求在不影响性能的情况下实现低功耗和小尺寸的网络边缘人工智能解决方案。莱迪思的sensAI技术集合的最新版本，适用于ECP5和iCE40 UltraPlus FPGA，为设计人员提供了在网络边缘实现低功耗、高性能AI所需的硬件平台、IP、软件工具、参考设计和设计服务。

**了解更多：**

www.latticesemi.com/zh-CN/sensAI

**在线联络：**

www.latticesemi.com/contact
www.latticesemi.com/buy

# 目录

# 摘要

低成本、高性能的网络边缘解决方案的市场竞争日益激烈。领先的市场研究公司预测，在未来六年内，网络边缘解决方案市场将迎来大爆发。IHS预计到2025年，将有超过400亿台设备在网络边缘运行，而市场情报机构Tractica预测，届时每年将出货超过25亿台网络边缘设备。

随着新一代网络边缘应用的出现，设计人员越来越倾向于开发结合低功耗和小尺寸而不降低性能的解决方案。推动这些新的AI解决方案的是越来越多的网络边缘应用，例如家庭控制中智能门铃和安全摄像头的存在检测，零售应用中用于库存的对象计数，以及工业应用中的物体和存在检测。一方面，市场要求设计人员开发出性能比以往更高的解决方案。另一方面，延迟、带宽、隐私、功耗和成本问题限制了他们依赖云的计算资源来执行分析。

同时，性能、功耗和成本限制因应用而异。随着实时在线网络边缘应用的数据需求不断推动对基于云的服务的需求，设计人员必须解决传统的功耗、电路板面积和成本问题。开发人员如何解决系统对于日益严格的功耗（毫瓦级）和小尺寸（5 $mm^2$到100 $mm^2$）要求。单论各种性能要求就已经很难满足。

## 利用FPGA的优势

莱迪思的FPGA具有独特的优势，可以满足网络边缘设备快速变化的市场需求。设计人员可以在不依赖云端的情况下，快速为网络边缘设备提供更多计算资源的其中一个方法是使用FPGA中本身的并行处理能力来加速神经网络性能。此外，通过使用针对低功耗运行而优化的低密度、小尺寸封装FPGA，设计人员可以满足新的消费和工业应用对功耗和尺寸的严格限制。例如，莱迪思的iCE40 UltraPlus™和ECP5™产品系列支持网络边缘解决方案的开发，功耗低至1 mW到1 W，硬件平台尺寸小至5.5 $mm^2$到100 $mm^2$。通过将超低功耗、高性能和高精度与全面的传统接口支持相结合，这些FPGA为网络边缘设备开发人员提供了满足不断变化的设计要求所需的灵活性。

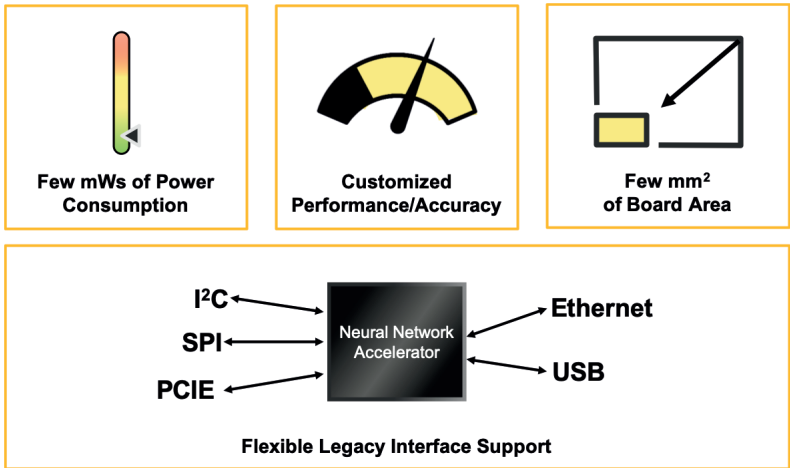


*图1：莱迪思半导体的低功耗、小尺寸FPGA提供适当的性能和功能组合，支持网络边缘人工智能应用*

为了满足这一需求并加速开发，莱迪思推出了业界第一款技术集合sensAI™，为设计人员提供了开发智能家居、智能工厂、智能城市和智能汽车中低功耗、高性能网络边缘设备所需的所有工具。sensAI旨在满足支持AI的网络边缘设备不断增长的需求，提供全面的硬件和软件解决方案，用于在网络边缘运行的智能设备中实现低功耗、实时在线的AI功能。它于2018年推出，旨在无缝创建新设计或更新现有设计，其低功耗AI推理针对这些新应用要求进行了优化。

这个综合设计生态系统中有什么？首先，莱迪思的模块化硬件平台，如带有HM01B0 Shield开发板的iCE40 UPduino 2.0和基于ECP5的嵌入式视觉开发套件（EVDK），为应用开发提供了坚实的基础。UPduino可用于仅需几毫瓦的AI设计，而EVDK支持需要更高功耗但通常工作在1W以下的应用。

软IP可以很容易地实例化到FPGA中，以加速神经网络的开发。因此，sensAI开发包包括CNN加速器IP，能让设计人员在iCE40 UltraPlus FPGA中实现深度学习应用。sensAI还提供完整的CNN可设置参数的加速器IP核，可以在莱迪思的ECP5 FPGA中实现。这些IP支持可变量化。这反过来又使设计人员能够在数据准确性和功耗之间进行权衡。

Lattice的sensAI技术集合允许设计人员通过易于使用的工具流程探索设计选项和权衡。设计人员可以使用Caffe、TensorFlow和Keras等行业标准框架进行网络训练。开发环境还提供神经网络编译器，将训练的网络模型映射为固定点表示，支持权重和激活的可变量化。设计人员可以使用编译器来帮助分析、仿真和编译不同类型的网络，以便在没有RTL经验的情况下在莱迪思的加速器IP核上实现。然后，设计人员可以使用传统的FPGA设计工具，如Lattice Radiant和Diamond来实现整个FPGA设计。

为加快设计实现，sensAI提供了越来越多的参考设计和演示。包括面部识别、手势检测、关键词检测、人员存在检测、面部跟踪、对象计数和速度标志检测。最后，设计团队通常需要一定的专业知识才能完成设计。为满足这一需求，莱迪思与全球各地的众多设计服务合作伙伴建立了合作关系，以便为AI / ML专业知识不足的客户提供支持。

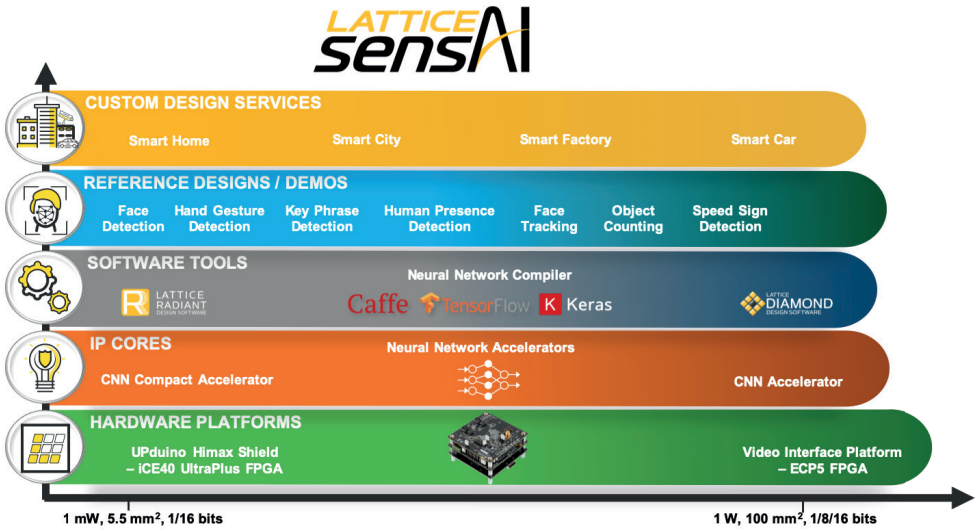


***图2 Lattice sensAI是一整套硬件和软件解决方案，适用于网络边缘人工智能应用的开发***

## 主要更新

为了满足网络边缘AI快速增长的性能要求，莱迪思在2019年发布sensAI更新，增强了其性能并优化了设计流程。更新后的sensAI比上一版本的性能提升了10倍，这是由多个优化促成的，包括通过更新CNN IP和神经网络编译器，新增8位激活量化、智能层合并以及双DSP引擎等特性，优化了存储器的访问。

在最新版本中，由于更新了神经网络编译器，支持8位输入数据，存储器访问序列得到大幅优化。因此不仅外部存储器的访问减少了一半，还支持使用更高分辨率的图像作为数据输入。使用更高分辨率的图像，解决方案自然更为精确。

为进一步加速性能，莱迪思优化了sensAI神经网络中的卷积层，减少了卷积计算耗费的时间。莱迪思将器件中的卷积引擎数量翻倍，减少了约50%的卷积时间。

莱迪思在不增加功耗的情况下提升了sensAI的性能，设计人员因此可以选择ECP5 FPGA产品系列中门数较少的器件。经优化的演示示例可以帮助实现性能提升。例如，针对低功耗运行进行优化、采用CMOS图像传感器的人员侦测演示，通过VGG8网络提供64 x 64 x 3的分辨率。该系统以每秒5帧的速率运行，使用iCE40 UltraPlus FPGA功耗仅为7 mW。第二个性能经优化的演示，针对人员计数应用，同样也使用CMOS图像传感器，通过VGG8网络提供128 x 128 x 3的分辨率。该演示以每秒30帧的速率运行，使用ECP5-85K FPGA功耗为850 mW。


***图3：这些参考设计展示了sensAI提供的功耗与性能可选方案***

与此同时，sensAI给用户带来无缝的设计体验，它支持更多神经网络模型和机器学习框架，从而缩短设计周期。全新可定制化的参考设计可简化对象计数和存在检测等常见的网络边缘解决方案的开发，同时设计合作伙伴生态也在不断拓展，为用户提供重要的设计服务。有了这些，莱迪思能为开发人员提供他们复制或调整其设计所需的全部关键工具。例如，下列框图展示了莱迪思提供的一系列全面的组件，包括训练模型、训练数据集、训练脚本、经过更新的神经网络IP和神经网络编译器。

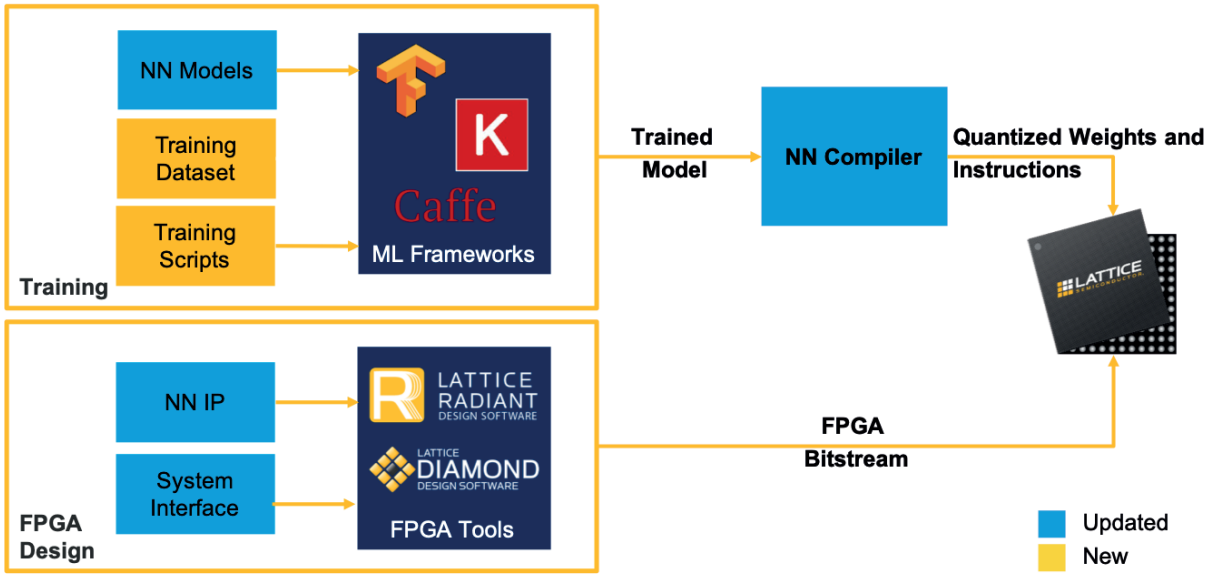


***图4：sensAI的设计流程包括了行业领先的机器学习框架、训练数据和脚本、神经网络IP等设计和训练网络边缘AI设备必需的资源***

莱迪思还拓展了对机器学习框架的支持，致力于提供无缝的用户体验。最初版本的sensAI支持Caffe和TensorFlow，最新版本新增支持Keras，这是用Python编写的开源神经网络，可在TensorFlow、Microsoft Cognition Toolkit或Theano上运行。Keras旨在帮助工程师快速实现深度神经网络，它可以提供用户友好、模块化和可拓展的环境，加速原型设计。Keras最初被当做一种接口而非独立的机器学习框架，它的高度抽象性能让开发人员加速开发深度学习模型。

为进一步简化使用，莱迪思更新了sensAI神经网络编译器工具，它可以在机器学习模型转换为固件文件时自动选择最精确的分数位数。sensAI更新还新增了一个硬件调试工具，让用户可以在神经网络每个层进行读取和写入。进行软件仿真之后，工程师也需要知道他们的网络在实际硬件上的表现。使用此工具，工程师可以在短短几分钟内看到硬件运行的结果。

此外，最新版本的sensAI得到了越来越多公司的支持，他们为莱迪思提供专为低功耗、实时在线的网络边缘设备而优化的设计服务和产品开发技能。这些公司通过无缝更新现有设计或针对特定应用开发完整的解决方案来帮助客户构建网络边缘AI设备。

## sensAI设计案例

莱迪思这一更高性能的全新解决方案可用于下列四种不同的加速器设计案例。在第一个设计案例中（图5）,设计工程师使用sensAI来构建独立运行模式的解决方案。这种系统架构能让设计工程师在莱迪思iCE40 UltraPlus或ECP5 FPGA上开发出实时在线的集成解决方案，具有延迟低、安全性高的特点，其中FPGA资源可用于系统控制。典型的一种应用就是使用独立运行的传感器实现人员侦测和计数。

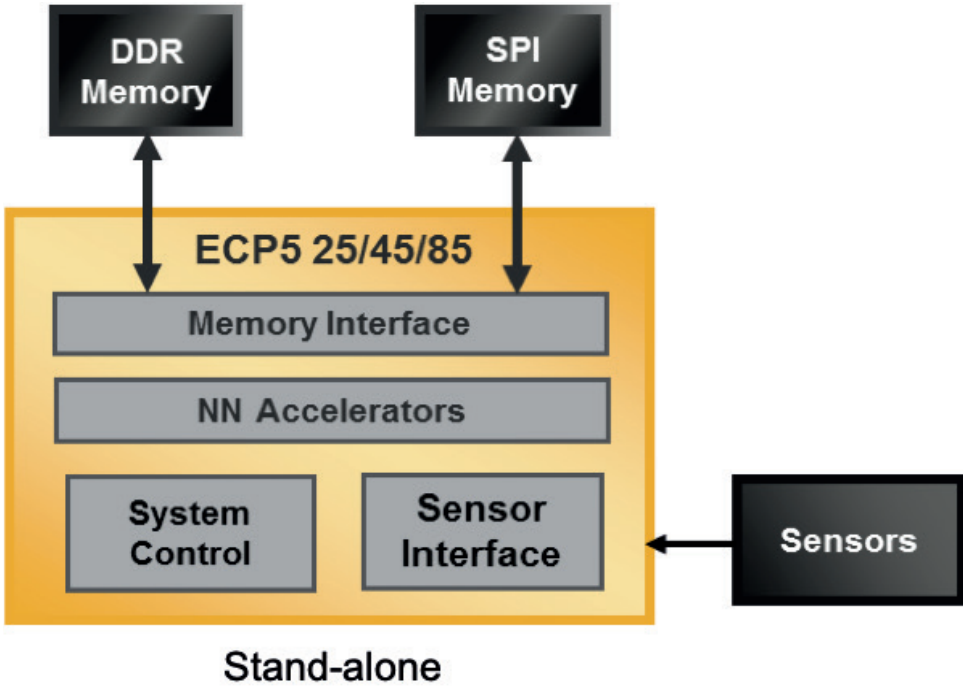


*图5：将sensAI作为独立运行的网络边缘AI处理解决方案*

设计人员还使用sensAI开发两种不同类型的预处理解决方案。第一种情况下（图6），设计人员采用了莱迪思sensAI以及一片低功耗的iCE40 UltraPlus FPGA对传感器数据进行预处理，从而最大程度地降低了向SoC或云端传输数据进行分析的成本。例如，如果是用在智能门铃上，sensAI会初步读取来自图像传感器的数据。如果判断为不是人，比如说是一只猫，那么系统就不会唤醒SoC或连接到云端作进一步处理。因此，这种方法可以最大程度降低数据传输成本和功耗。如果预处理系统判断门口的对象是人，则唤醒SoC作进一步处理。这能极大减少系统需要处理的数据量，同时降低功耗要求，这对于实时在线的网络边缘应用来说至关重要。

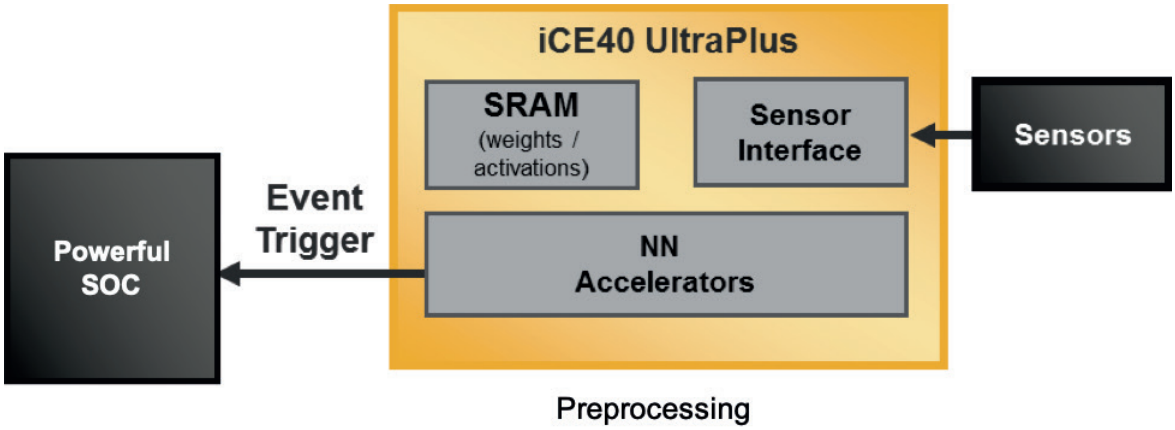


*图6：在此案例中，sensAI会预处理传感器数据来判断该数据是否需要发送到SoC作进一步处理*

在第二个预处理应用中，设计人员可以使用ECP5 FPGA实现神经网络加速（图7）。在此案例中，设计人员利用ECP5 IO的灵活性将各类现有的板载器件（如传感器）连接到低端MCU，实现高度灵活的系统控制。

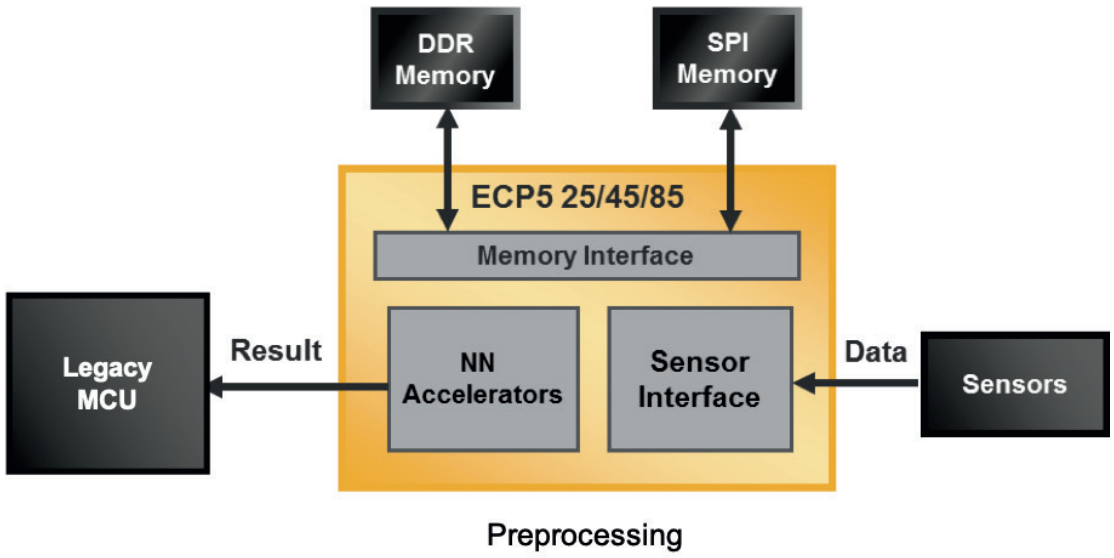


*图7：第二个系统架构也采用了预处理，设计人员可以使用ECP5和sensAI预处理传感器数据，加强神经网络的综合性能*

设计人员还可以在后处理系统中使用sensAI加速器（图8）。越来越多的设计案例表明，很多公司虽然已经开发出经过验证、基于MCU的解决方案，但是他们希望在不更换组件或重新设计的情况下新增某种AI功能。但是在某些情况下，他们的MCU性能相对不足。典型的例子就是智能工业或智能家庭应用，在进行分析之前需要图像滤波。设计人员可以在这里添加另一个MCU，然后经历耗时的设计验证过程，或者也可以在MCU和数据中心之间添加加速器进行后处理，最大限度地减少发送到云端的数据量。这种方法对想要添加AI功能的物联网设备开发人员尤其有吸引力。

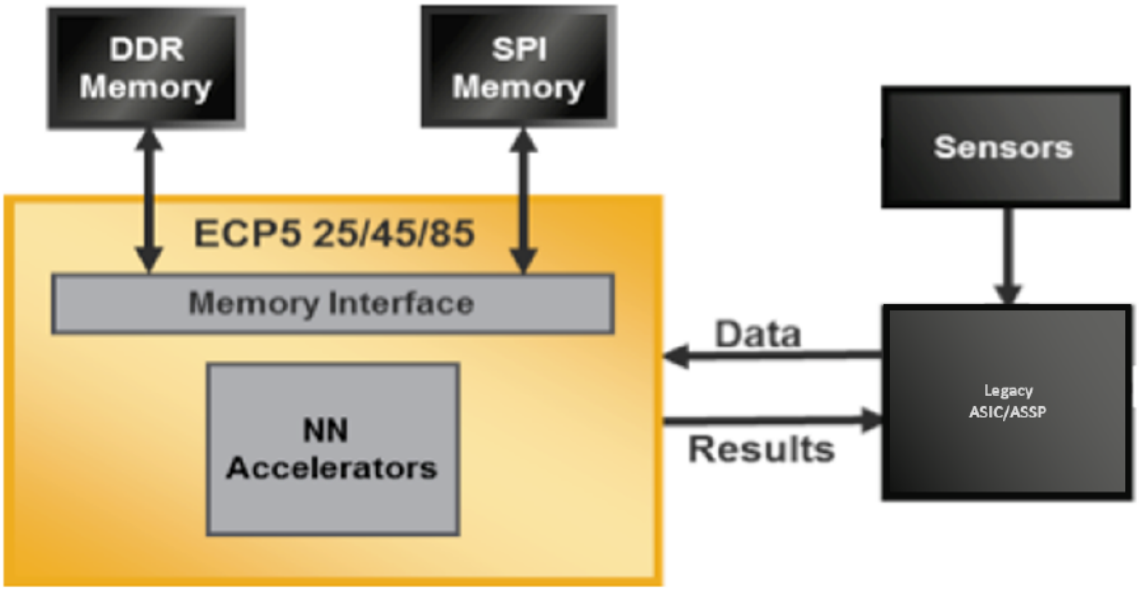


***图8：通过sensAI增强该基于MCU的设计，让现有的设计支持网络边缘AI功能***

## 结论

显然，未来几年将是实时在线的网络边缘智能设备这一市场发展的关键时期。由于应用变得越来越复杂，设计人员将急需能够以低功耗支持更高性能的工具。莱迪思最新版本的sensAI技术配合ECP5和iCE40 UltraPlus FPGA，将为设计人员提供硬件平台、IP、软件工具、参考设计和设计服务，帮助他们战胜竞争对手，快速开发出成功的解决方案。

**了解更多：**

www.latticesemi.com/zh-CN/sensAI

**在线联络：**

www.latticesemi.com/contact
www.latticesemi.com/buy