新一代高能效 AI 加速器 (DRP-AI3)：将自主系统的高级 AI 的嵌入式处理速度提高 10 倍

概述

由于生育率下降和老年人口比例上升导致劳动人口减少，社会中的各个方面，包括工厂、物流、医疗保健、市内服务机器人和安保摄像头在内，都将需要使用高级人工智能 (AI) 处理，例如，周围环境识别、行动决策和运动控制。 系统将需要在各种类型的程序中实时进行高级人工智能 (AI) 处理。 具体而言，系统必须嵌入到设备中才能对不断变化的环境做出快速反应。在嵌入式设备中，AI 芯片需要在保持更低功耗的同时执行高级 AI 处理，对于发热有着严格的限制。

为了满足这些市场需求，瑞萨电子开发出了 DRP-AI（用于 AI 的动态可配置处理器）作为 AI 加速器，将高速 AI 推理处理与边缘设备所需的低功耗和灵活性相结合。 这种经过多年培养的可配置 AI 加速器处理器技术被嵌入到针对 AI 应用的 RZ/V 系列 MPU 中。 DRP-AI3 是新一代 DRP-AI，与上一代相比，电源能效高出约 10 倍。 DRP-AI3 能够应对 AI 的未来发展需要以及机器人等应用的复杂需求。 此白皮书介绍了专为 DRP-AI3 开发的关键技术，说明了 DRP-AI3 如何克服发热挑战、实现高速实时处理，并使 AI 产品实现更高的性能和更低的功耗。

DRP-AI3 加速器特性

(1) 用于实现 AI 模型轻量化（剪枝）的硬件 (H/W)-软件(S/W) 协同技术使电源能效达到了传统模型的 10 倍。

- AI 加速器（DRP-AI3 硬件）引入了高速低功耗技术剪枝模型
- 软件可轻松生成适合 DRP-AI3 的剪枝模型，并以最优方式将其实施到 H/W 中

(2) DRP-AI、DRP 和 CPU 在异构架构中协同工作，可以加快包括 AI 在内的各种算法。

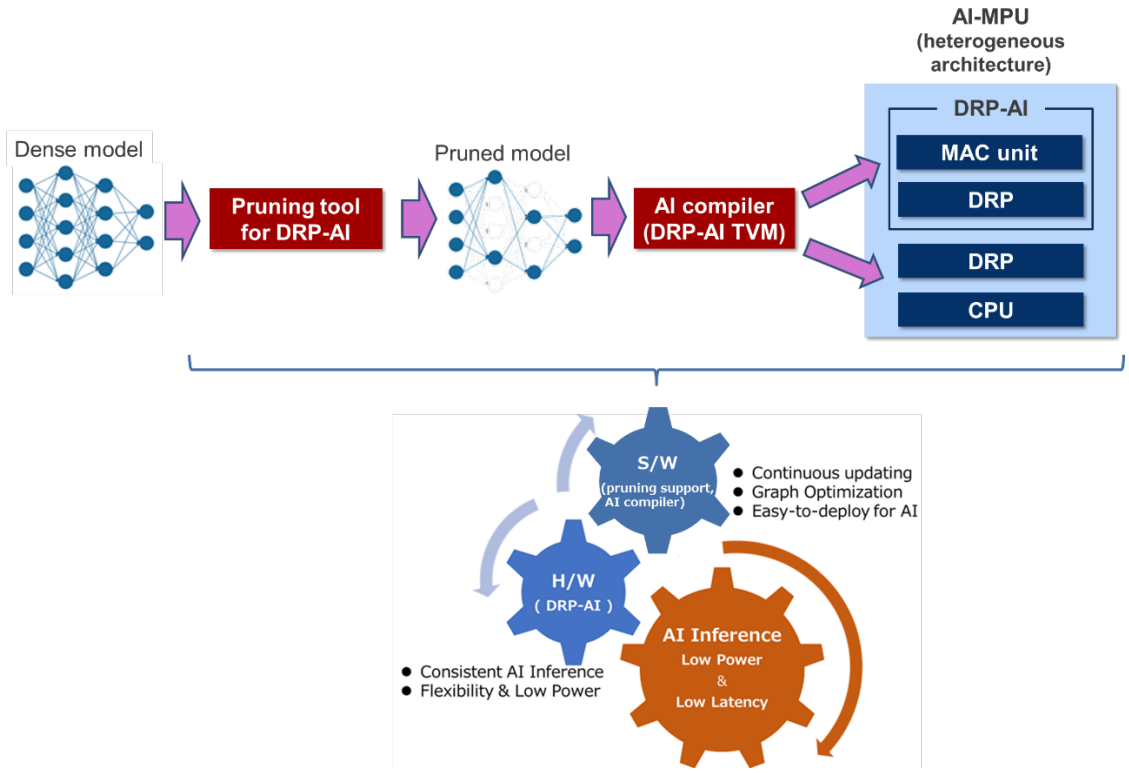


图 1： 用于 DRP-AI3 的硬件-软件协同设计

剪枝 AI 模型的高速、低功耗特性

- 硬件架构支持减少位数 (INT8) 的主流轻量化技术以及剪枝技术
- DRP-AI3 的灵活性能够实现更快的随机模型剪枝，现有硬件难以实现这一点。
- 与未进行剪枝时相比，处理时间缩短至原来的 1/16，功耗降低至原来的 1/8 左右。

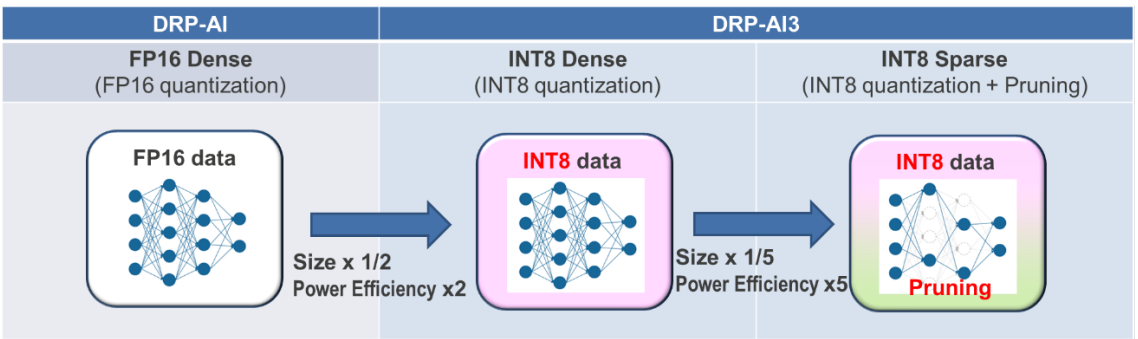


图 2：应用于 DRP-AI3 的轻量化技术

DRP-AI3 引入了支持主流 AI 模型轻量化方法的高速低功耗方案。 具体而言，它支持以下轻量化方法：

1) 量化：降低各层的神经网络权重信息（权重）和输入/输出数据（特征图）的位权重。 将传统 DRP-AI 中的 16 位浮点运算变为 8 位整数运算 (INT8)。
2) 剪枝：一种将不影响识别准确度的权重信息（枝）设为零，以此来跳过计算的技术。

(1) 理想情况下，量化所产生的功耗预计比传统 DRP-AI（16 位处理）低约 2 倍以上，这是因为算术单元大小和数据访问量相比于位数来说权重更轻。 (2) 此外，剪枝取决于 AI 模型可以保留多少权重信息，例如，如果可以实现 90% 的剪枝，那么速度和功耗将分别有望提高和降低约 10 倍。

当前的 AI 硬件所面临的一个主要挑战在于，它们无法高效处理 AI 模型，特别是 (2) 剪枝 AI 模型。 AI 硬件通常基于 SIMD（单指令多数据）架构，该架构可同时执行大量的乘积累加运算，进而高效处理神经网络的大型乘积累加矩阵运算。 由于不影响识别准确度的权重在矩阵中随机分布，因此即使部分权重在并行积分累加运算中变为零，也仍然会对这些权重和非零权重执行并行计算。 因此，剪枝无法减少计算次数 （图 3）。

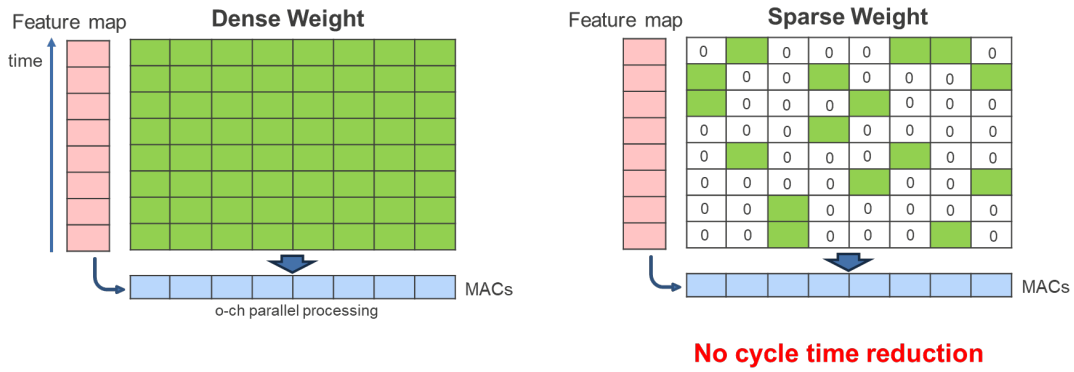


图 3：通过通用并行架构进行轻量化（剪枝）模型处理

结构化剪枝中是 AI 硬件中的常用剪枝方法，其中的值被设为零（例如，权重矩阵每一列为零），并且不影响并行性。 然而，这种方法无法实现高剪枝率，因为对于本质随机的权重信息来说，条件存在明显的不同。 另一种方法是选择并计算权重矩阵 (1) 中两个相邻权重中的一个。 这种方法最多也仅可将权重中的信息量减少 1/2，所以它在提升速度、降低功耗方面的效果有限。

因此，瑞萨电子开发出了灵活的 N:M 剪枝方法，这种方法可以灵活跳过运算，即使是在面临更加随机的剪枝时也是如此。

如图 4 所示，此技术的基本概念是将原始权重矩阵分解成 M 列的权重矩阵组，将它们重构为更小的 N 列权重矩阵组，仅从中提取出每一组的重要权重。 然后，对新的权重矩阵组进行并行运算。 在这个过

程中，DRP-AI3 的一项新功能可以通过自由切换每个权重矩阵组的 N 值来调整运算周期数，使其能够以最优方式，对实际 AI 模型中的局部变化剪枝率执行跳过运算处理，如图 4 所示。 通过这种微调 N 值的能力，还可以具体设置整个权重矩阵的剪枝率，根据用户所需的功耗、运算速度和识别准确度要求实现最优的剪枝处理。

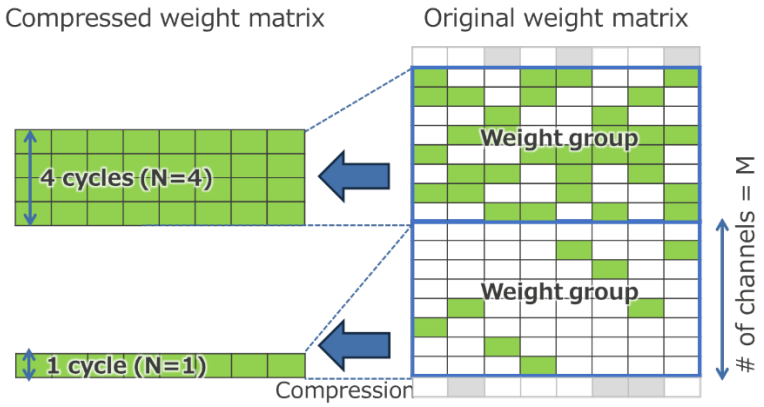


图 4：使用 DRP-AI3 压缩轻量化（剪枝）模型

通过分析实际 AI 模型的剪枝方法和识别准确度之间的关系，我们确定了平衡准确度、面积、功耗提升的最优参数，并将它们反映在了硬件架构中。

相比于传统的 AI 加速器配置，这项技术将 AI 模型的处理周期数减少了至少 1/16，同时还将功耗降低了至少 1/8（图 5）。 这解决了传统 AI 处理器存在的发热问题，并且该技术可以实施到机器人和小型 AI 设备内部，无需冷却机制。

** M=16, N=variable

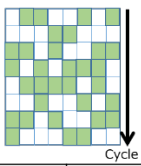
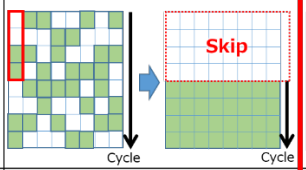
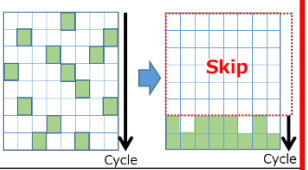
	General AI accelerators		50% pruning arch. (ref [1])	This work (DRP-AI)**
zero weight non-zero weight				
Pruning structure	structured	unstructured	unstructured	unstructured
pruning rate	0~30%	0~90%	0 / 50%	0 ~ 93%
weight data size	~1/3x	~1/10x	~5/8x	~1/10x
Cycle time	~1/3x	x1x	1/2x	~1/16x

图 5：不同加速器的剪枝模型处理性能对比

生成和实施剪枝方法的软件特性

• 基于 DRP-AI3 架构特性的硬件/软件协同设计，能够考虑可高效提升性能的剪枝位置

• 开发 DRP-AI 扩展包（剪枝工具）

• 开发 DRP-AI TVM（INT8 量化工具和编译器）

如前所述，剪枝是一种将不影响识别准确度的权重信息（枝）设为零，从而跳过计算的方法。 AI 模型剪枝越多，预期的性能就越快，功耗也越低，但识别准确度也会随之降低。

如图 6 所示，为了提升剪枝率，同时防止出现识别准确度下降的问题，通常会采取剪枝流程。 在初始训练后，通过选择剪枝点来执行剪枝，并在剪枝后使用权重信息执行再训练。 再训练有望阻止剪枝导致识别准确度下降的问题。

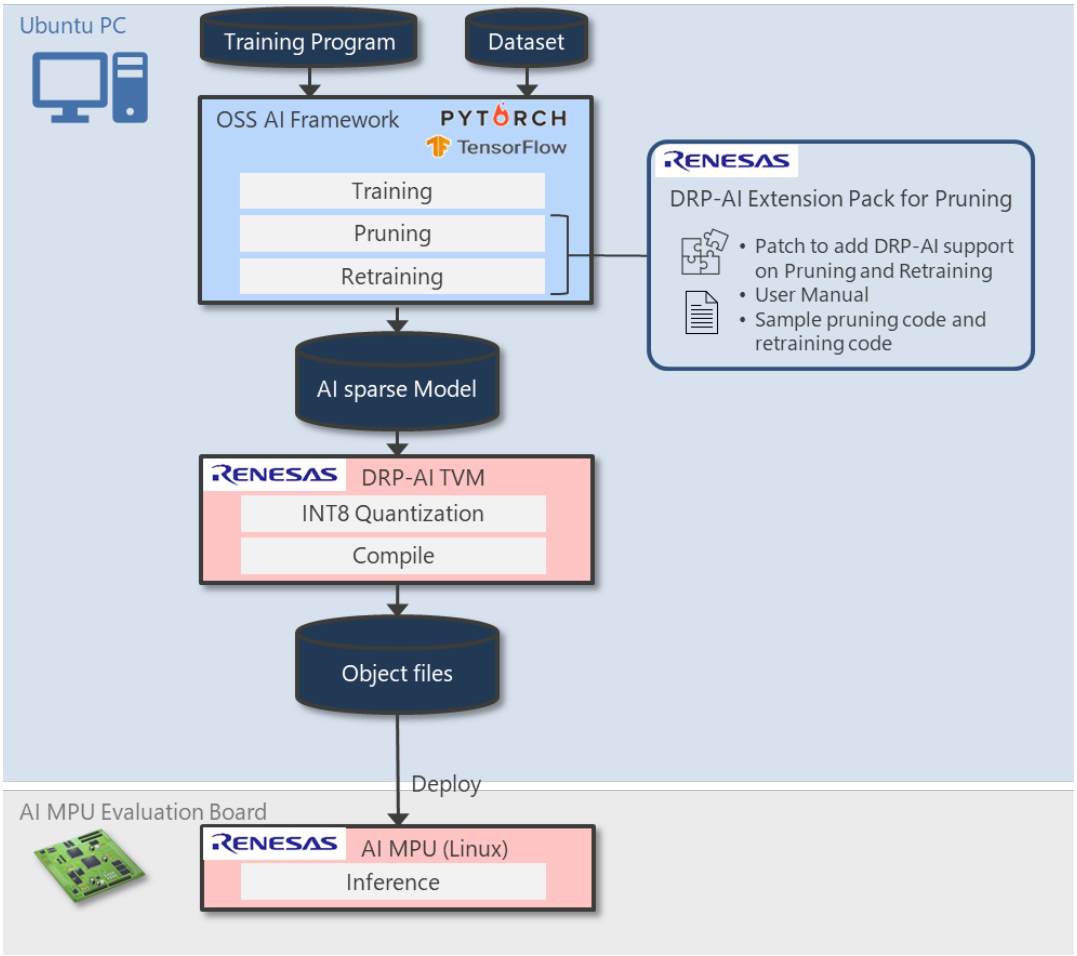


图 6：DRP-AI3 AI 模型轻量化和实施流程

一般来说，在初始训练后，应该选择对识别准确度影响最少的剪枝点，即训练中权重绝对值较小或梯度较小的点。 除了这些因素之外，瑞萨电子还开发了一款剪枝工具（DRP-AI 扩展包），该工具可以选择剪枝点来满足上述 DRP-AI3 剪枝硬件架构限制。 用户仅需指定剪枝率即可应用 DRP-AI3 的“灵活 N:M 剪枝”特性。

为了进一步降低应用剪枝的难度，以上工具以库的形式集成在 OSS AI 框架（Pytorch、Tensorflow）中，仅需在用户现有的训练脚本中添加几行代码即可实现剪枝和再训练（剪枝和再训练如图 6 所示）。

此外，剪枝工具生成的剪枝后 AI 模型可以由 DRP-AI TVM 进行转换，从而同时进行 INT8 量化和编译。 DRP-AI TVM 可以轻松生成可执行的目标文件。

在这里，DRP-AI TVM 作为一种工具，可将已训练的 AI 模型转换为可在瑞萨电子 Renesas AI MPU 中执行的格式。 它基于 OSS ML 编译器框架 Apache TVM，能够将 AI 模型各层中可由 DRP-AI 执行的运算以及无法由 DRP-AI 执行的运算分配到 CPU 进行处理。 这种合并使用多个处理器的计算被称为异构计算。 DRP-AI 是一款强大的 AI 加速器，但它能够执行的运算次数有限。 通过引入异构计算机制，可以大幅扩展受支持的 AI 模型种类（AI 模型覆盖范围）。

通过这种方式，瑞萨电子提供了多种软件环境，例如 DRP-AI 扩展包，它充分减少了用户应用剪枝所需的时间和精力，同时通过硬件-软件协同设计提高剪枝效率，进而充分利用 DRP-AI 硬件架构，此外还有 DRP-AI TVM，它通过异构计算实现了 AI 模型覆盖范围的大幅扩展，由此改善了 UX。

DRP-AI、DRP 和 CPU 协同运行的异构架构特性

• 使用 AI 加速器、DRP 和 CPU 的多线程流水线处理

• 搭载 DRP（动态可配置有线逻辑硬件）的低抖动高速机器人应用

以服务机器人为例，它需要高级 AI 处理才能识别周围环境。 另一方面，不使用 AI 的算法型处理对于决定和控制机器人的行为也必不可少。 然而，当今的嵌入式处理器 (CPU) 缺少足够的资源来实时执行这些不同类型的处理。 瑞萨电子通过开发异构架构技术解决了此问题，能够使动态可配置处理器 (DRP)、AI 加速器 (DRP-AI) 和 CPU 协同工作。

如图 7 所示，动态可配置处理器 (DRP) 可以在执行应用的同时，根据待处理的内容，在各个运算时钟周期动态切换芯片上算术单元的电路连接配置。 由于仅会用到必要的算术电路，因此相比于 CPU 处理，DRP 消耗的功耗更少，所能达到的速度也更高。 此外，在 CPU 中，由于缓存缺失和其他原因，外部频繁读取内存会削弱性能，而与此相比，DRP 可以提前在硬件中建立必要的数据路径，从而降低因内存读取而导致的性能弱化以及运算速度波动（抖动）幅度。

DRP 还具有动态可配置功能，可以在每次算法变化时切换电路连接信息，即使是在需要多算法处理的机器人应用中，也能利用有限的硬件资源完成处理。

DRP 在处理流式数据方面（例如图像识别）十分有效，其中的并行化和流水线设计可以直接提高性能。 另一方面，机器人行为决策和控制等程序需要在更改条件和处理细节的同时进行处理，从而对周围环境的变化做出反应。 相比于硬件处理（例如在 DRP 中），CPU 软件处理可能更适合这种应用。 重要的是，将处理分配到正确的位置并以协调的方式进行运算。 瑞萨电子的异构架构技术使 DRP 和 CPU 可以协同工作。

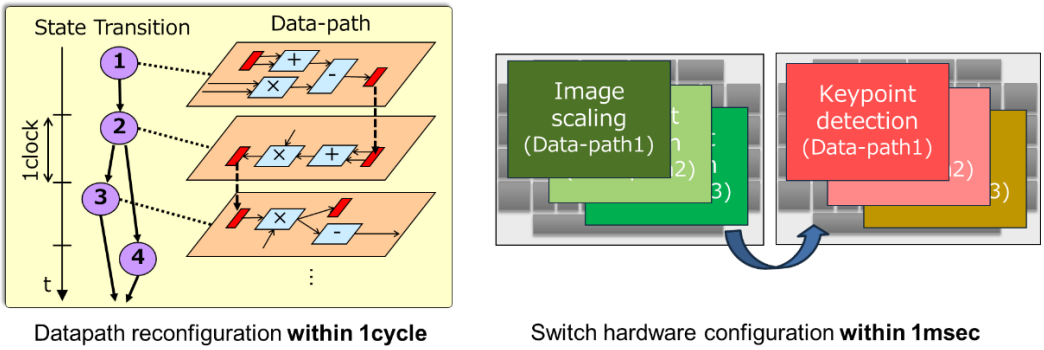


图 7： 灵活的动态可配置处理器 (DRP) 特性

MPU 和 AI 加速器 (DRP-AI) 架构的概览如图 8 所示。 机器人应用实现了 AI 型图像识别和非 AI 型决策及控制算法的复杂组合。 因此，兼具用于 AI 处理的 DRP (DRP-AI) 和用于非 AI 算法的 DRP 的配置将大幅提升机器人应用吞吐量。

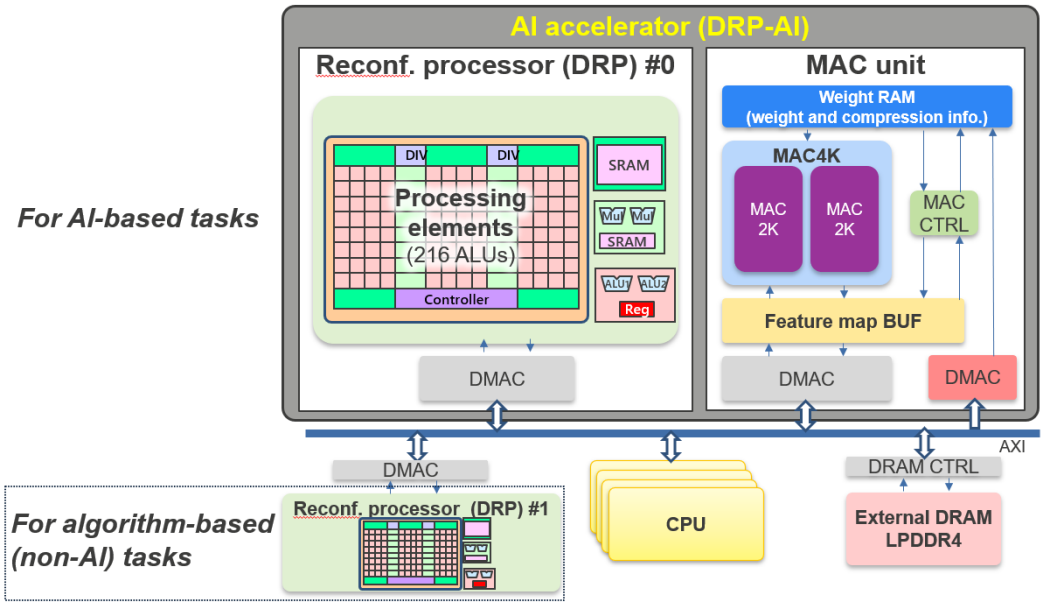


图 8：基于 DRP-AI3 的异构架构配置

评估结果

(1) AI 模型处理性能评估

搭载此技术的原型测试芯片在加速器处理性能方面可实现高达每秒 8 万亿次乘积累加运算 (8 TOPS)。此外，对于已剪枝的 AI 模型，可以根据剪枝量信息成比例地减少运算周期次数，从而使 AI 模型运算性能达到剪枝前模型的峰值水平 (80 TOPS)（注释 1）。 这相当于高出传统 DRP-AI 处理性能约 80 倍，这样的大幅度性能提升足以跟上 AI 快速发展的步伐（图 9）。

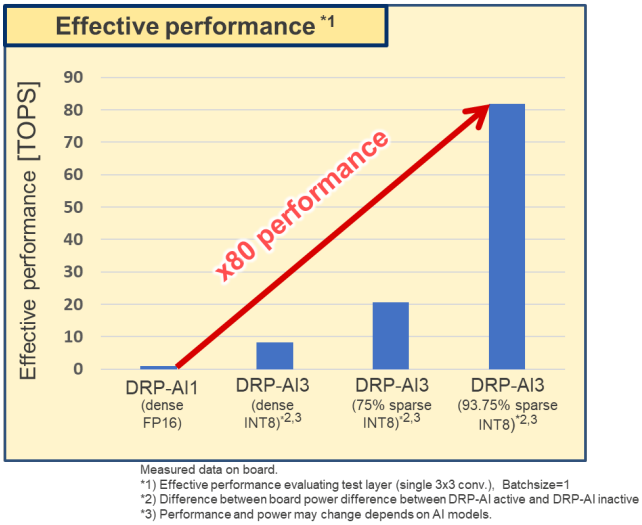


图 9：DRP-AI 的峰值性能测量对比

一方面，随着 AI 处理速度的提升，不使用 AI 的算法型图像处理（例如 AI 前期和后期处理）所用的处理时间相对较长，正日益成为一个制约因素。 在 AI-MPU 中，一部分图像处理程序被转移到 DRP，从而帮助缩短了整体系统处理时间。 （图 10）

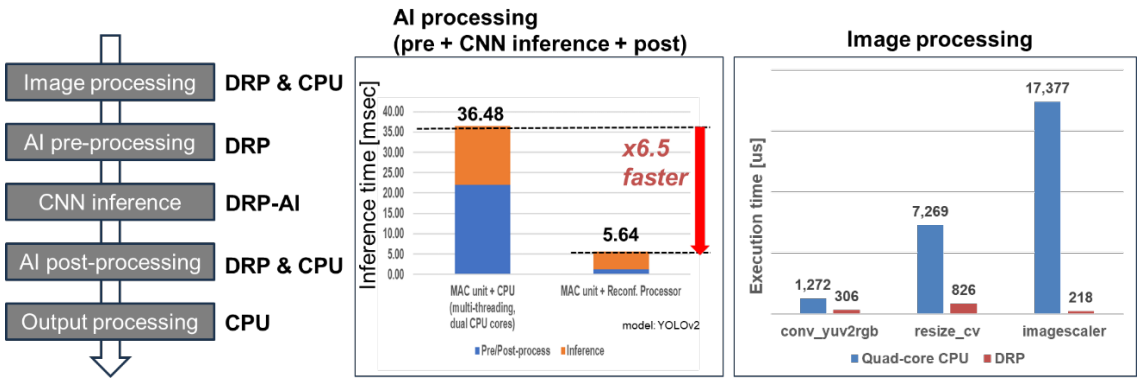


图 10：异构架构加速图像识别处理

在电源能效方面，仅 AI 加速器的性能评估显示，其最高理论性能约为 23 TOPS/W，运行主流 AI 模型时的电源能效达到世界顶尖水平（约为 10 TOPS/W）。（图 11）

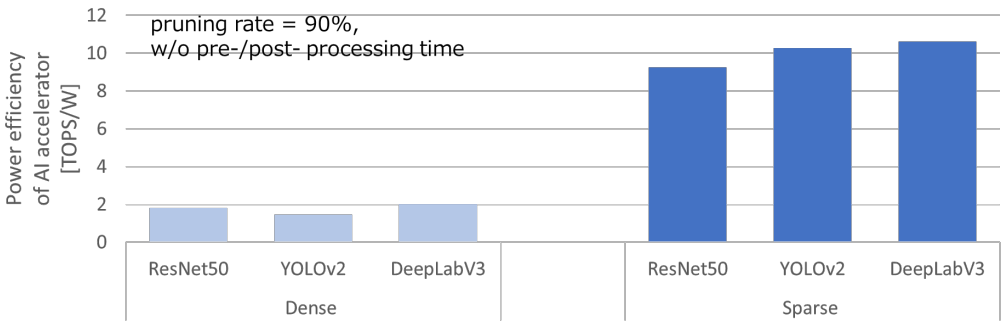


图 11：真实 AI 模型的电源能效

我们还证明了即使是在搭载原型芯片且未配备风扇的评估板上，也能执行相同的 AI 实时处理，并且其温度与配备风扇的竞品相当。（图 12）

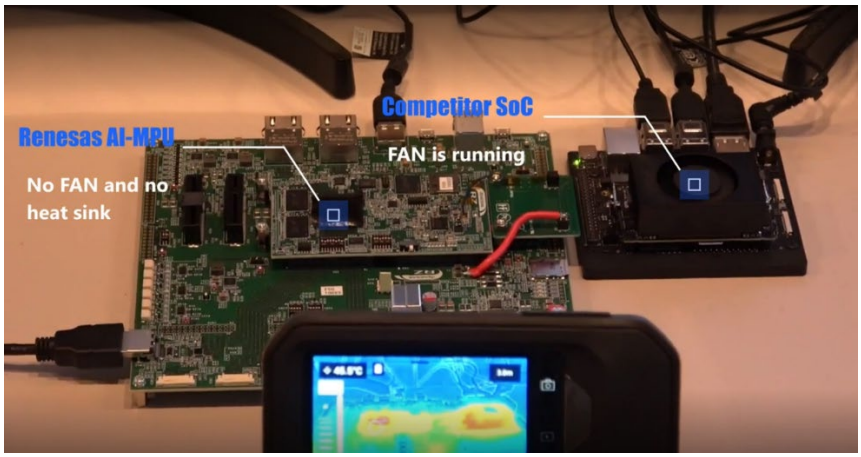


图 12：无风扇 DRP-AI 测试板与带风扇 GPU 的发热对比

(2) 机器人应用示例

例如，常见机器人应用之一的 SLAM（同步定位与建图）具有十分复杂的配置，需要多程序进程并行识别机器人位置，并通过 AI 处理进行环境识别。 瑞萨电子的 DRP 使机器人可以瞬时切换程序，并且事实证明，AI 加速器与 CPU 的并行运算速度约为单 CPU 运算的 17 倍，功耗低至单 CPU 运算水平的 1/12。

总结

瑞萨电子开发了 DRP-AI3，它是 DRP-AI（用于 AI 的动态可配置处理器）的高级版本。这款独特的 AI 加速器将端点所需的低功耗与灵活性相结合，具有轻量化 AI 模型的处理能力，并且电源能效比以往型号高出 10 倍 (10 TOPS/W)。 瑞萨电子将不断扩大 MPU 阵容，提供更多搭载了这款强大 AI 加速器的可扩展 MPU 产品（RZ/V 系列）。

AI 发展预计将日渐复杂，而瑞萨电子将紧跟发展步伐，及时推出产品，推动部署能够智能实时地对端点产品做出反应的系统。

2024 年 2 月 18 日举行的国际顶尖半导体电路会议 ISSCC 2024（2024 年国际固态电路会议）更为详细地介绍了相关技术内容。

相关信息

- RZ/V2H：四核视觉 AI MPU，配备 DRP-AI 加速器和高性能实时处理器
- DRP-AI：瑞萨电子独创的 AI 加速器，兼具高 AI 推理性能和低功耗

公司总部

135-0061, 日本东京江东区

豊洲 3-2-24, TOYOSU FORESIA

https://www.renesas.com

联系信息

有关产品、技术的更多信息，文档的最新版本，或

离您最近的销售办公室，请访问：

https://www.renesas.com/contact-us

Doc Number: R01WP0022CC0100