WP505 (v1.0) 2018 年 10 月 2 日

Versal: 首款自适应计算加速平台 (ACAP)

正式推出 Versal ACAP，一个完全支持软件编程的异构计算平台，将标量引擎、自适应引擎和智能引擎相结合，实现显著的性能提升，其速度超过当前最高速的 FPGA 20 倍、比当今最快的 CPU 实现快 100 倍，该平台面向数据中心、有线网络、5G 无线和汽车驾驶辅助应用。

摘要

近来涌现的技术挑战迫使业界跳出传统的通用 (one-size-fits-all) 型 CPU 标量处理解决方案，进而探索新的发展方向。大型的矢量处理 (DSP-GPU) 技术能够解决一些问题，但由于其灵活性欠佳及低效率存储器带宽的使用，导致再次陷入了传统的扩展挑战。传统 FPGA 解决方案提供可编程存储器层级，但传统的硬件开发流程一直是阻碍数据中心市场等应用领域广泛、大规模采用FPGA的障碍。

该解决方案将所有这三大要素与一个新的工具流相结合，通过单个自适应计算加速平台 (ACAP)，提供了从框架到 C 到 RTL 级编码的各种不同抽象。赛灵思 Versal™ ACAP 作为一大新器件门类，支持用户利用三大可编程要素定制自己的特定领域专用架构 (DSA)。

介绍

近期在半导体工艺领域涌现的技术挑战阻碍了传统上通用 (one-size-fits-all) 型 CPU 标量计算引擎的扩展。如图 1 所示，半导体工艺频率缩放的变化迫使标准计算单元愈发趋于并行[参考资料 1]。

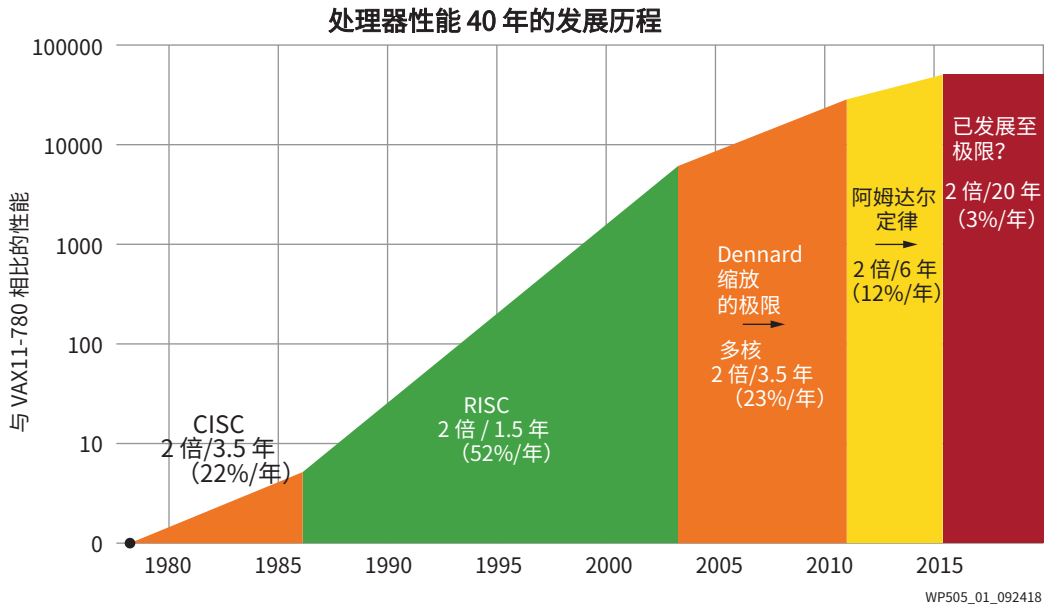

图 1： **处理器性能发展历程**

因此，半导体工业正在探索替代特定领域的架构，包括以往被归入特定极端性能应用的部分，如基于向量的处理 (DSP、GPU) 和完全并行可编程的硬件 (FPGA)。问题在于，哪种架构最适合哪项任务？

- **标量处理单元**（例如 CPU）在具有不同决策树和广泛库的复杂算法中非常有效，但在性能扩展方面受到限制。
- **矢量处理单元**（例如 DSP、GPU）在一组更窄的可并行计算函数集上效率更高，但由于存储器层级结构不灵活，它们会受时延和效率的影响。
- **可编程逻辑**（例如 FPGA）可以精确地根据特定的计算功能定制，这使它们在时延关键型实时应用（例如汽车驾驶辅助）和不规则数据结构（例如基因组测序）方面表现最佳，但算法的更改传统上要花几个小时来编译，而不是几分钟。参见图 2。

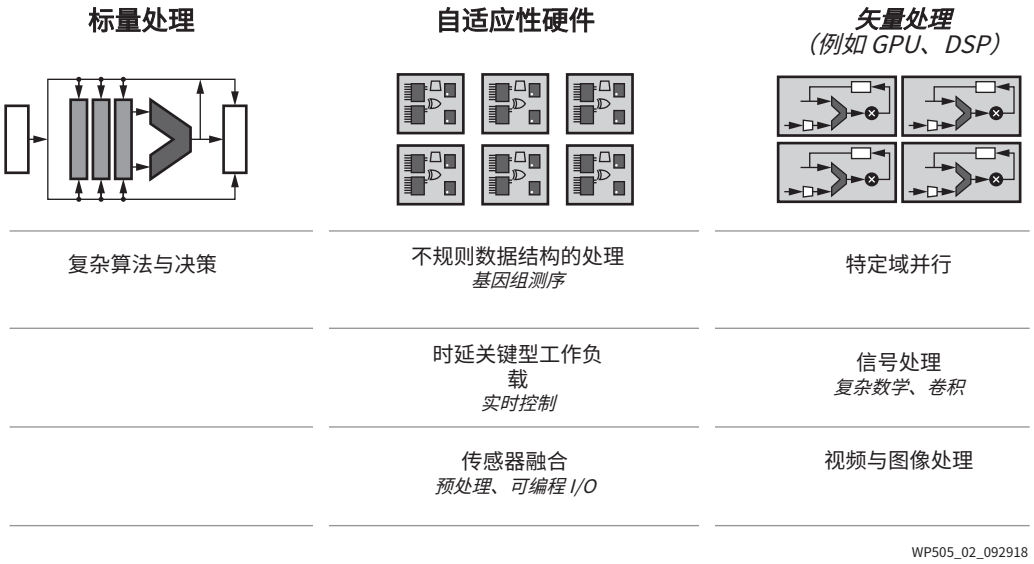

图 2： **计算引擎的类型**

为应对这一问题，赛灵思推出了一种革命性的新异构计算架构，即自适应计算加速平台 (ACAP)，它囊括三大方面优势，提供了与下一代可编程逻辑 (PL) 紧密耦合的世界一流的矢量与标量处理单元，将一切与高带宽片上网络 (NoC) 联通，提供对所有三种处理单元类型的存储器映射访问。这种紧密耦合的混合架构比任何一种单独架构的实现都支持更高的定制水平和性能提升。参见图 3。

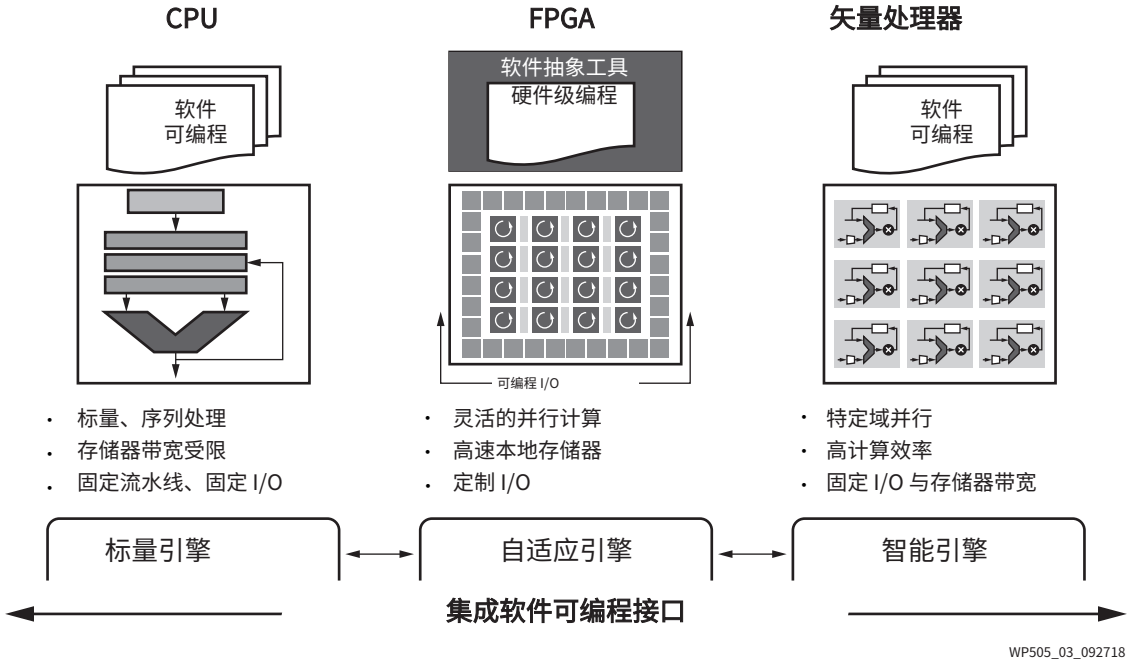

图 3： **异构集成三种类型的可编程引擎**

要想在性能上有如此大的提升，就必须对工具进行类似的大幅改进，并重点关注易用性。ACAP 在设计上不需要 RTL 流，可以开箱即用。ACAP 原生支持软件编程，有助于开展基于 C 和基于框架的设计流程。这些器件具有集成 Shell，包括具有集成型 DMA、NOC 和集成型存储器控制器的高速缓存一致性主机接口（PCIe® 或 CCIX 技术），从而避免了开展 RTL 工作的要求。

新的 ACAP 架构在易用性方面也带来了显著改善。它通过一个统一的工具链为编程提供了一个完全集成的存储器映射平台。赛灵思工具链面向各类开发人员支持多种输入方式。例如，某些应用（如 AI 机器学习推断）可以在框架级别（例如 Caffe、TensorFlow）进行编码；其他应用可以使用预先优化的库（例如 5G 无线电滤波器）用 C 语言进行编码。传统型硬件开发人员仍然可以通过传统的 RTL 输入流将他们现有的 RTL 移植到 ACAP。

本白皮书审视了由传统的基于 CPU 的计算模式开展变革的需要，详细探讨了其他选项，并介绍了赛灵思 Versal ACAP—— 业界首款异构计算平台。

ACAP 的三大主要优势包括：

1. **软件可编程性**——能够通过软件抽象工具链快速开发优化应用。
2. **加速**——指标涵盖广泛的应用，包括人工智能、智能网络接口卡、高密度存储、5G 无线、自动驾驶汽车、高级模块化雷达，以及太比特光网络。
3. **动态自适应重配置**——能够重配置硬件，实现毫秒间加速新的负载。

推出 ACAP：面向并行异构计算开展软硬件优化

ACAP 的特点在于它结合了新一代标量引擎、自适应引擎和智能引擎。NoC 通过存储器映射接口将它们相连，总带宽为 1Tb/s+。除 NoC 之外，可编程逻辑（和集成型 RAM 块）支持的大量存储器带宽支持可编程存储器架构针对单个计算任务进行层级优化（避免了其他基于高速缓存计算单元固有的高时延和时延不确定性)。参见图 4。

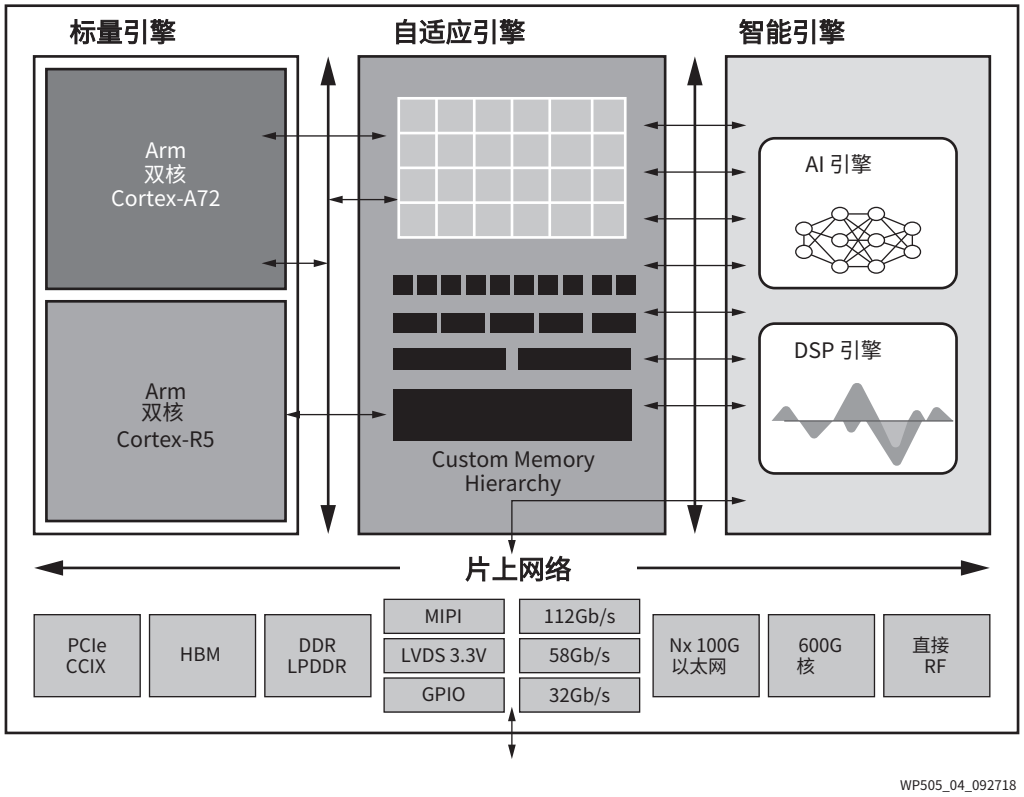

图 4： **赛灵思 Versal ACAP 功能图**

标量引擎基于双核 Arm® Cortex-A72 构建，与赛灵思上一代 Arm Cortex-A53 核相比，每核单线程性能提高了 2 倍。高级的架构和 7nm FinFET 工艺的功耗相结合，DMIPS/WAT 与先前的 16nm 实现方案相比提高了 2 倍。立足赛灵思目前在汽车业大量部署的经验，经 ASIL-C 认证的 (1) UltraScale +™Cortex-R5 标量引擎结合额外的系统级安全特性向 7nm 迁移。

自适应引擎由可编程逻辑和存储器单元组成，与新一代业界最快的可编程逻辑相连。除了支持原有设计之外，还可以重新编程这些结构，以形成针对特定计算任务定制的存储器层级。与最新的 GPU 和 CPU 相比，赛灵思智能引擎可实现 更高的循环效率和更高的单位计算存储器带宽。这是优化边缘时延与功耗，以及优化核心绝对性能的关键。

智能引擎由一组创新的超长指令字 (VLIW) 和单指令、多个数据 (SIMD) 处理引擎以及存储器构成，彼此间的互联速度和存储带宽均为 100Tb/s。这使机器学习和数字信号处理 (DSP) 应用的性能提升了 5-10 倍。

如表 1所示，这些计算函数以不同的比率和大小混合，构成了 Versal 器件产品组合。

1. https://china.xilinx.com/news/press/2018/xilinx-announces-availability-of-automotive-qualified-zynq-ultrascale-mpsoc-family.html

表 1: **Versal 器件产品组合，市场，以及重要特性**

Versal 产品组合	主要市场	重要特性
Versal AI Core	数据中心、无线	最高水平智能引擎计算
Versal AI Edge	汽车、无线、广播、A&D	紧密热度范围下高效智能引擎数降至 5W
Versal AI RF	无线、A&D、有线	直接 RF 转换器与 SD-FEC
Versal Prime	数据中心、有线	带集成型 Shell 的基准平台
Versal Premium	有线、测试与测量	搭载最高水平自适应引擎的高级平台，112G SerDes 和 600G 集成 IP
Versal HBM	数据中心、有线、测试与测量	带 HBM 的高级平台

赛灵思自适应计算加速平台 (ACAP) 结合了矢量、标量和自适应硬件单元，提供了三大引人注目的优势：

- 软件可编程性
- 异构加速
- 灵活应变能力

软件可编程性

由自适应芯片支持的自适应加速

Versal ACAP 提供自适应加速硬件，易于在软件中进行编程。无论任何应用类型，异构引擎都支持软件应用的最佳水平加速。智能引擎能够加速机器学习和常用的经典 DSP 算法。自适应引擎内的新一代可编程逻辑对并行算法进行加速。多核 CPU 为剩余的应用需求提供了全面的嵌入式计算资源。整个 Versal 器件在设计上便于使用软件编程，无需具备硬件专业知识。参见图 5。

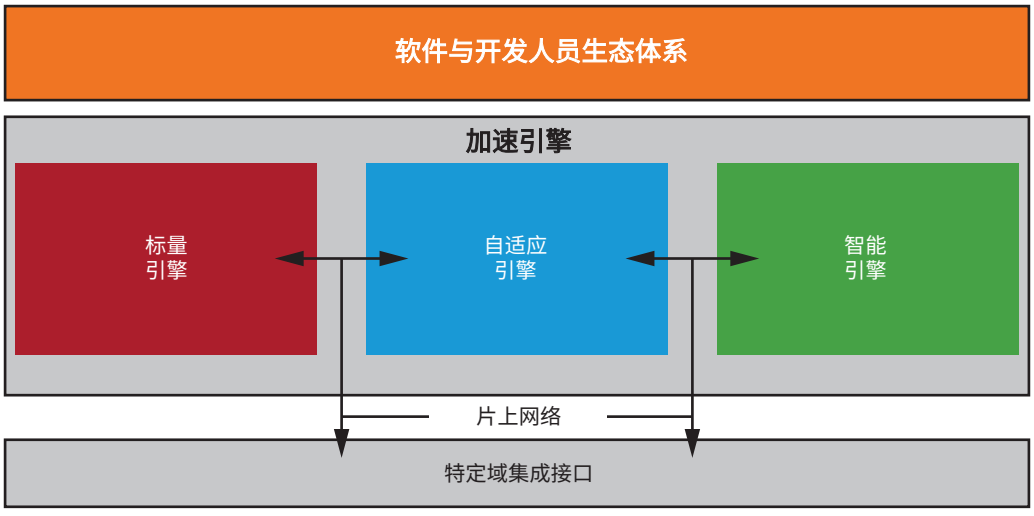

图 5: **Versal ACAP 顶层概念图**

- 数据和 AI 科学家可以部署在标准软件框架中构建的应用，并使用 Versal ACAP 为应用实现数个量级的加速。
- 软件应用开发人员使用赛灵思统一软件开发环境，无需硬件专业知识，就可以使用 Versal ACAP 加速任意软件应用。
- 硬件设计人员可以继续使用 Vivado® Design Suite 进行设计，同时使用 Versa 平台的集成 I/O 接口和 NoC 缩短开发时间。

参见图 6。

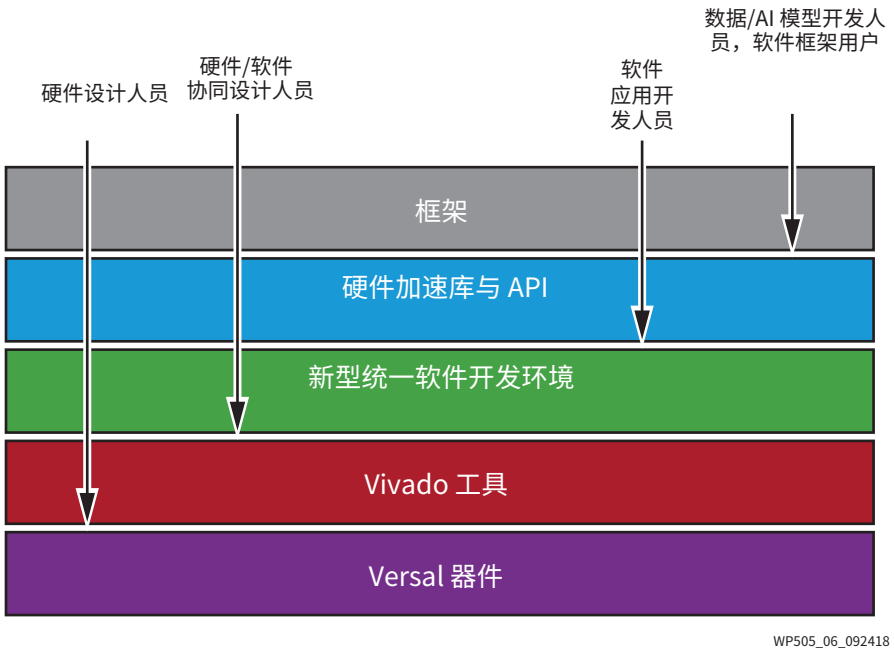

图 6： **Versal 平台软件形象概念**

专用硬件，提高易用性和应用效率

自适应接口逻辑实现了对片外接口的轻松访问。这包括到外部主机处理器的标准接口。在数据中心应用中，软件应用通常驻留于主机 CPU 上，而不是嵌入式微处理器上。连接主机 CPU 和 Versa 平台可编程资源的接口称为 Shell。集成型 Shell 包括完全兼容型高速缓存一致互联，适用于加速器 (CCIX) 或主机 PCIe Gen4x16 接口、DMA 控制器、缓存一致性存储器、集成型存储器控制器、高级功能性安全和安全功能。

NoC 有助于每个硬件组件和软 IP 模块间轻松地相互访问，或通过存储器映射接口访问软件。它提供了一个标准化的、可扩展的硬件框架，使异构引擎和接口逻辑之间能够进行高效通信。

异构加速

虽然可编程逻辑 (FPGA) 和基于矢量的 (DSP、GPU) 近来已展示出明显高于 CPU 的性能提升，但只有当开发人员利用 Veral ACAP 的多个类型计算单元支持紧密耦合的计算模型时，ACAP 架构的真正优势才成为人们关注的重点。在这种架构下，三单元合力可远超仅仅三倍的功效。

表 2 总结了 Versal ACAP 器件为各类市场提供的优势。

表 2: **Versal ACAP 与目标市场**

市场	基准	与 CPU 对比	与 GPU 对比	与 FPGA 对比	注释
数据中心	图像识别（推断）——时延敏感	43 倍	2 倍	5 倍	GoogLeNet v1（不限制批处理大小）
	图像识别（推理）——2ms 时延	不适用	8 倍	5 倍	GoogLeNet v1 (< 2 ms) CPU 时延下线 5ms
	风险分析	89 倍	不适用	>1 倍	用于利率互换 Maxeler 结果的风险价值 (VaR)
	基因组学	90 倍	不适用	>1 倍	人类基因分析 Edico 基因组结果
	弹性搜索	91 倍	不适用	>1 倍	1TB 数据 BlackLynx 结果时延降低 91 倍
无线 5G	16x16 5G 远程无线电	不适用	不适用	>5 倍	为 5G 远程无线电提供 >5 倍的无线电带宽
	波束形成	不适用	不适用	>5 倍	>5 倍的计算能力
A&D 雷达	DSP TMACs	不适用	不适用	>5 倍	超过 27 TMAC
	算法迭代时间	不适用	不适用	>100 倍	软件可编程智能引擎在几分钟内编译完毕
汽车	低时延推断 (<2 ms)	不适用	3 倍	15 倍	ResNet50 Batch=1 AI 引擎能更好地适应低时延、安全关键型 ADAS 和自动驾驶
	外壳类型	1	2	4	ACAP 产品组合是唯一能够高效支持 <10W、20W、30W，以及后备箱安装外壳的器件
有线	加密网络流量	不适用	不适用	4 倍	ACAP 对网络和加密 IP 的集成使多太比特的单芯片实现成为可能。

数据中心人工智能：机器学习推断加速

随着人工智能开始在现代生活中普及，对提高计算效率的需求开始推动半导体领域的创新，但任何单一的实现都很难以开展最大效率的处理。在这方面，矢量处理和可编程硬件之间的紧密耦合具有无可比拟的价值。

计算单元（FP32、FP16、INT16、INT8 等）的精度一直是人们关注的焦点，但对网络类型之间存储器层级需求差异的忽视，导致众多最新的人工智能推断引擎在不同网络上的效率急剧下降。例如，目前业界一流的机器学习推断引擎需要 4 个 HBM 存储器（7.2 Tb/s 的外部存储器带宽）才能达到其最高性能，但它们基于缓存的存储器层级效率仅为 25-30%，并为实时应用带来了显著的时延不确定性。解决方案就是用可编程存储器层级强化智能引擎执行的矢量处理，精确地针对每种网络类型进行优化，并通过 FPGA 逻辑的大规模并行来实现。

例如，GoogLeNet 的 Versal 平台实现为非时延敏感型应用提供了极高性能，比当今最高端的 Skylake Platinum CPU(2) 吞吐量高出 43 倍，比当前的顶级 GPU [参考资料 2]性能高约 3 倍，并且功耗均更低。参见图 7。

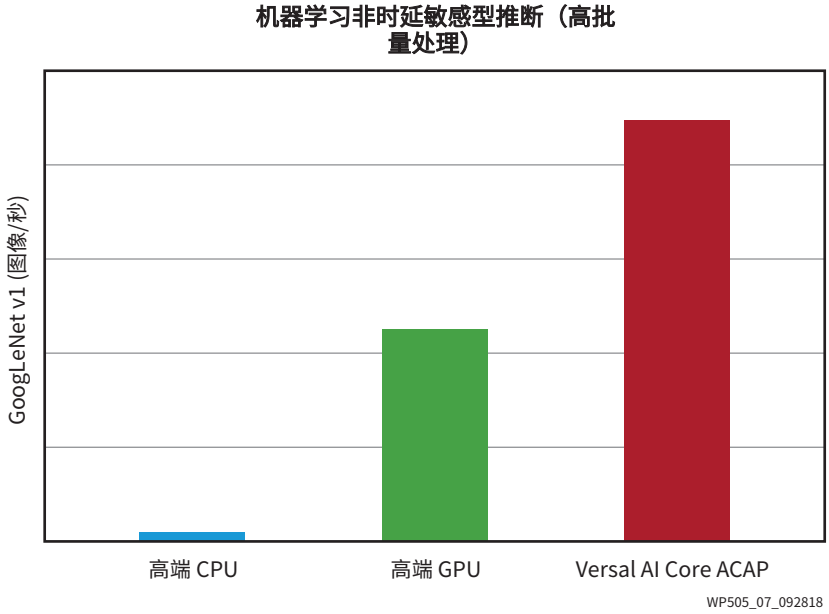

图 7： **GoogLeNet 性能（< 7ms 时延）= 比高端 CPU[1,2]性能高出 43 倍**

1. 测量器件为 Xeon Platinum 8124 Skylake，c5.18xLarge AWS 实例，Intel Caffe： https://github.com/intel/caffe.
2. V100 数据取自 Nvidia 技术概览，“深度学习平台，AI 服务在性能和效率方面的巨大飞跃”。

2. Xeon Platinum 8124 Skylake，c5.18xlarge AWS 实例，Canonical，Ubuntu，16.04LTS，AMD64 Xenial Image 建于 2018 年 8 月 14 日，Intel Caffe。Git 版本：a3d5b02，run_benchmark.pyunmodified。

随着数据中心不断深入地应用于神经网络，多个神经网络可以链接在一起，大大增加了对低时延神经网络的性能需求。例如，实时口语翻译需要语音转换文本，自然语言处理，推荐系统，文本转换语音，然后语音合成[参考资料 2]。这意味着对于该应用，神经网络的总时延预算增加了 5 倍。

随着实时应用数量的不断增加，对数据中心客户而言，选择一种可扩展的技术以满足他们未来的需求极为关键。这就出现了两种趋势：

- 为提高软件设计效率，确定性时延变得愈发重要[参考资料 3]。
- 随着日益复杂的交互建模（人机交互、金融交易）和安全关键型应用（如汽车、工业应用）的增加，神经网络时延要求日益严格。

这两个要求需要消除批处理，这将导致基于 CPU 和基于 GPU 的解决方案的固定的、基于缓存的存储器层级性能显著下降。即使高端 CPU 时延极限也高达 5ms，而一旦时延在 7ms 以下，甚至是高端的 GPU 也会出现显著的性能下降。仅有 Versal ACAP 能够以可接受的性能实现低于 2 ms 时延。参见图 8。

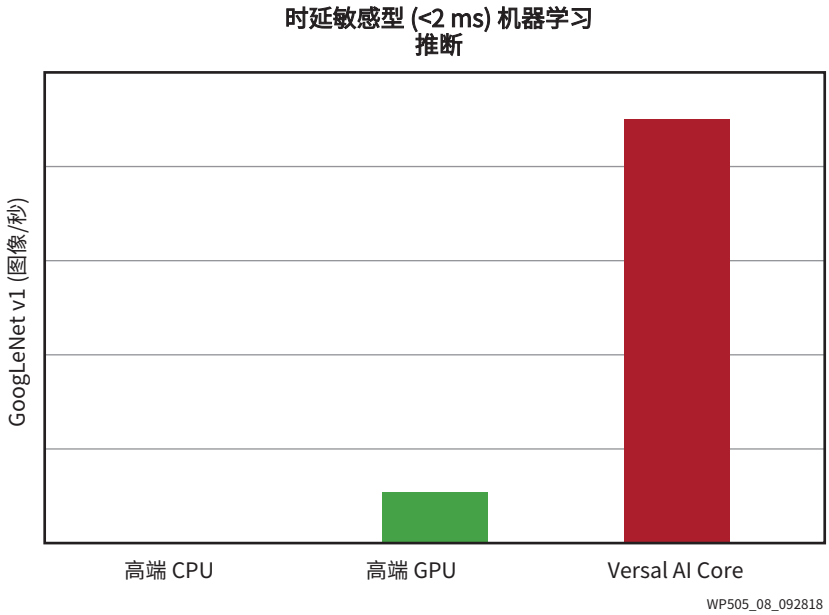

图 8： **GoogLeNet 实时性能（< 2 ms 时延）=高出高端 GPU (Nvidia) 8 倍1, 2**

1. 测量器件为 Xeon Platinum 8124 Skylake，c5.18xLarge AWS 实例，Intel Caffe：https://github.com/intel/caffe.
2. V100 数据取自 Nvidia 技术概览，“深度学习平台，AI 服务在性能和效率方面的巨大飞跃”。

因此，基于 ACAP 的解决方案独有的可编程存储器层级既提供了最高性能的机器学习推断性能，也提供了无与伦比的扩展性，因为未来的应用要求更低和更确定的时延。

数据中心智能 NIC

网络接口卡 (NIC) 起初只是简单的连接。随着时间的推移，它们通过增加额外的网络加速（加密、管理程序网络卸载、虚拟开关）化身为“智能 NIC”。亚马逊在 Annapurna 项目上取得了巨大的成功；它从 CPU 中卸载了所有的程序管理器功能，使 100% 的 CPU 周期都能用于产生收入的计算。

随着智能 NIC 的发展，赛灵思预计将出现三大优势：能够在数据中心以太网逻辑上动态分配和扩展工作负载，能够运行任何计算加速功能的可重配置加速池（最大限度地利用云资源），以及能够与网络数据平面一致运行计算功能。

赛灵思 Versal ACAP 器件支持将 NIC 功能与基于矢量和可编程逻辑的混合计算引擎集成，所有这些功能都由赛灵思的网络 IP 和世界一流的 SerDes 提供深度支持，包括用于新一代 NIC to TOR（机架顶部）链路的单通道 112G SerDes。

此外，可以在新的工作负载上动态地重配置或重新部署这些 NIC 资源。

表 3: **数据中心网络接口卡类型**

	描述	特性	示例
1 类	基础连接性 NIC	· 基础卸载（校验、LSO、RSS） · 单根 I/O 虚拟化 · 某些隧道卸载 (VXLAN、GRE0)	· Fortville · ConnectX · NetExtreme
2 类	用于网络加速的 SmartNIC	· 加密/解密（IP 安全） · 虚拟开关卸载（OVS 等） · 可编程隧道类型	· **赛灵思 2 类** · LiquidIO · Annapurna · Innova
3 类	用于网络计算加速的 SmartNIC	· 内联机器学习 · 内联视频转码 · 数据库分析 · 存储（压缩、加密、Dedupe）	· **赛灵思 3 类** · MSFT (NIC+FPGA)

数据中心存储加速

长期以来，FPGA 一直被用于存储驱动器，执行纠错和写调平任务。它们灵活的 I/O 支持卓越的设计重用，在发展迅速的闪存技术界尤为关键。此外，众多当前的数据库搜索和加速设备都在驱动器附近采用了基于 FPGA 的加速并大获优势。（通过将计算单元直接布局在驱动器旁，可以获得最大限度的效率。）

采用 ACAP 架构，驱动器和数据库加速厂商可以直接在驱动器内（已使用 FPGA）添加机器学习计算，从而将跨数据中心的数据移动（以及相关的时延、功耗和运营开支）减少 10 倍。

5G 无线通信

无线用户对带宽无止境的渴求推动了无线产业“每 10 年 10 倍”的极速创新步伐。在 2020 年奥运会上，业界将开始首次公开演示第五代无线技术，称为“5G”。大多数初始实现将构建于现有的赛灵思器件，特别是极为成功的 16nm RFSoC 器件上，它提供了三个关键优势：

- 集成直接 RF 采样率 ADC 和 DAC
- 集成 LDPC 和 turbo 软决策前向纠错 (SD-FEC) 码块
- 16nm FinFET 工艺技术带来的低功耗 DSP

随着行业的发展涌现出两大挑战：以较低的成本向更宽的频谱迈进，以及在无线电中增加机器学习推断技术，以增强光束引导算法、增强用户交接算法和支持自愈网络。

传统意义上，一些无线厂商通过实现基于矢量 DSP 的 ASIC 来降低成本。Versal ACAP 中加入了一个智能引擎，很大程度上消除了 ASIC 和 FPGA 之间传统的成本差距，因为它提供了超 5 倍的单芯片 TMAC。参见图 9。

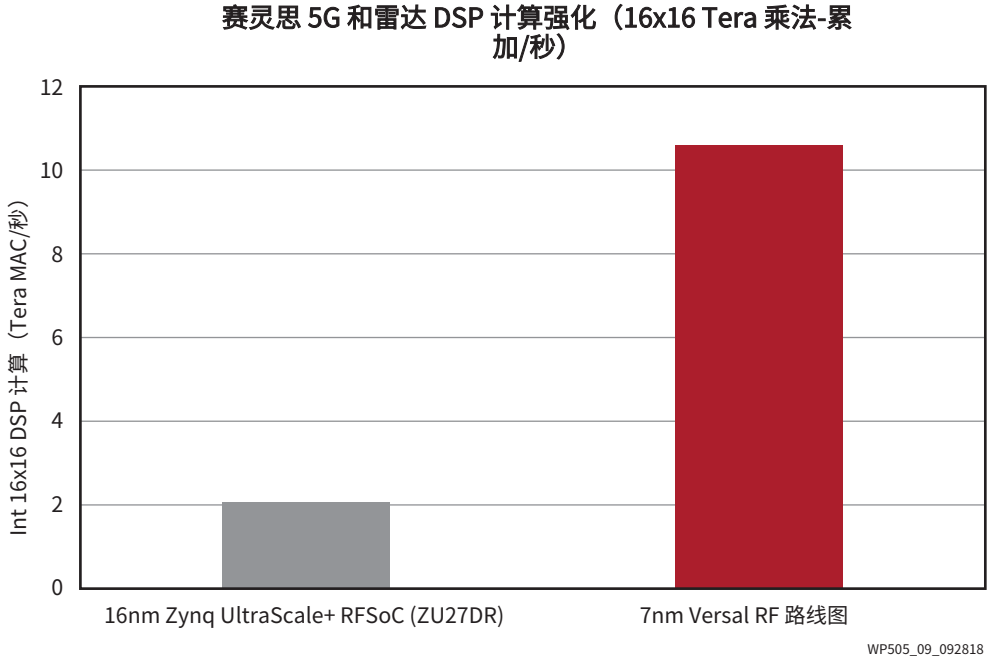

图 9： **Xilinx RF 计算路线图**

因此，虽然 16nm Zynq UltraScale+RFSoC 可实现 200MHz 16x16 远程无线电单元 (RRU)，但 7nm Versal 器件路线图可以实现完整的 800MHz 16x16 RRU。参见图 10。

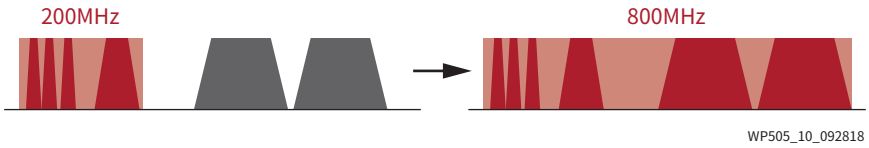

图 10： **16nm 与 7nm 无线电器件的单片频谱**

增加了高效的机器学习（具有框架级的设计流程），为基于 ACAP 的 Versal 产品组合开拓了一个全新的门类。这种技术可以增强光束引导和用户交接算法，比传统的编程定义算法高出两倍，接近理论极限的 85%。参见图 11。

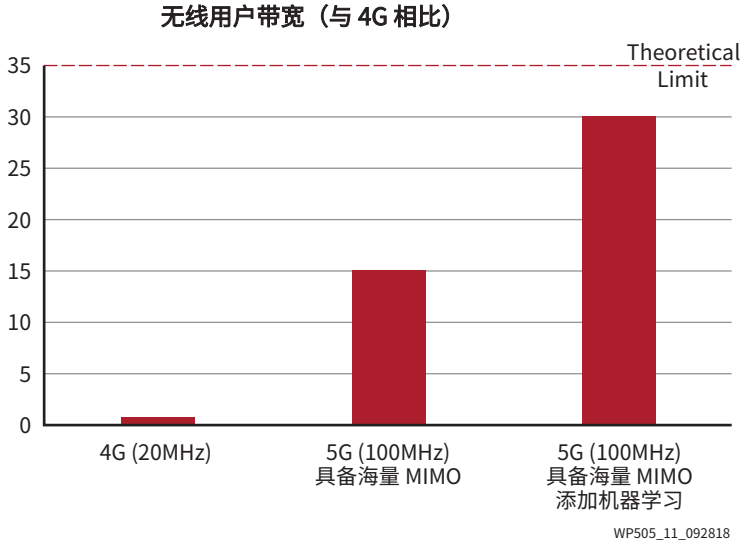

图 11： **无线带宽优化与理论极限的对比**

赛灵思是业界唯一将所有四种关键技术汇聚在单个芯片的供应商：直接 RF 采样 ADC 和 DAC，集成式 SD-FEC 代码，基于高密度矢量的 DSP，以及框架可编程的机器学习推断引擎，打造出业界第一款真正的 5G 片上无线电。

例如，图 12 描述了 ACAP 架构汇聚经典无线需求和紧急 AI/ML 技术的能力的示例。RF 波形分类器在认知无线电应用中将发挥重大作用，有助于提高无线电资源的利用率。使用 AI 机器

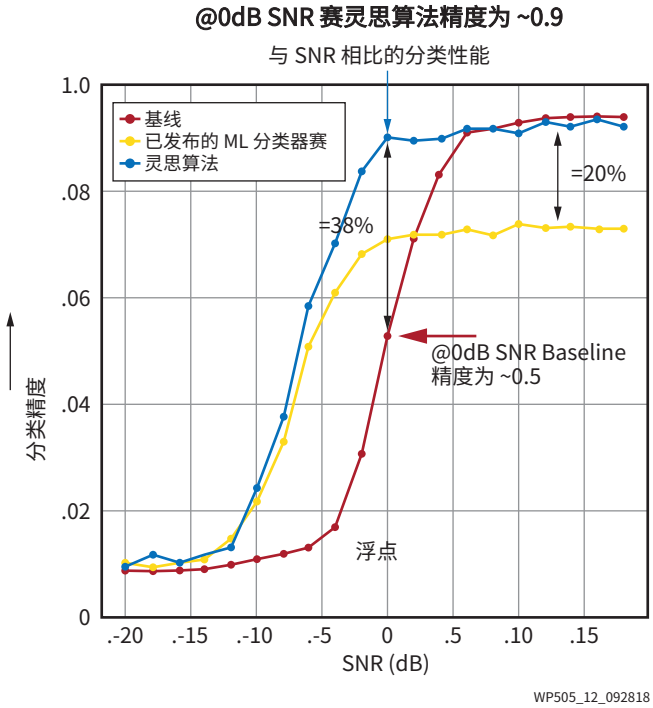

图 12： **机器学习带来的数字预失真 (DPD) 效率提升**

航空航天与国防

FPGA 的大规模并行 DSP 能力长期以来一直是许多国防领域雷达实现的支柱。然而，ADC 技术的最新创新已将 ADC 采样率提高到每秒数万兆次，这要求 DSP 能力也取得相应地提高。

基于矢量的强大 DSP 引擎与 AI 机器学习的融合，使航空航天与国防工业的革命性新产品，如先进的模块化雷达成为可能。由高频波长驱动的天线间距要求采用极小的外形。赛灵思在单一封装器件中，就能提供每秒太比特的天线带宽，以及多达 17 TMAC 的 INT24，或 24 TFLOPS 的 32 位单精度浮点 DSP。

汽车驾驶辅助 (ADAS)

赛灵思在汽车、航空航天、卫星、医疗和商业网络系统领域的高可靠性和热约束系统方面拥有历史悠久的经验。赛灵思技术经专门设计，以减轻 SEU 效应，并能在高达 125°C 的温度下运行，结合对机器视觉和机器学习的关注，可靠性和质量方面的丰富经验意味着赛灵思技术原生适用于汽车驾驶辅助系统 (ADAS) 和未来的自动驾驶汽车技术。迄今为止，赛灵思已经面向各种汽车插槽交付超过 1.5 亿个 FPGA 和 SoC，并专门为 ADAS 应用供货超过 5000 万个器件。汽车业是赛灵思在过去两年中增长最快的细分市场。

赛灵思针对汽车的可扩展 Versal ACAP 内含一个高能效标量引擎，该引擎具有双核 Cortex-R5S、可编程 I/O 和低时延、智能 AI 引擎，该引擎支持节能、功能性安全、AI 强化的自动驾驶解决方案，与当今市面上基于 FPGA，ASIL-C 认证的 ADAS 解决方案相比，INT 8 机器学习性能提高了 15 倍。

此外，通过空中硬件更新对整个器件进行重新编程的能力提高了系统在现场的通用性，从而提高了客户价值。最后，赛灵思可编程 I/O 为厂商提供了变更传感器类型的灵活性和适应性，无需承担等待 ASSP 或 GPU 重设计带来的延误与成本。参见图 13。

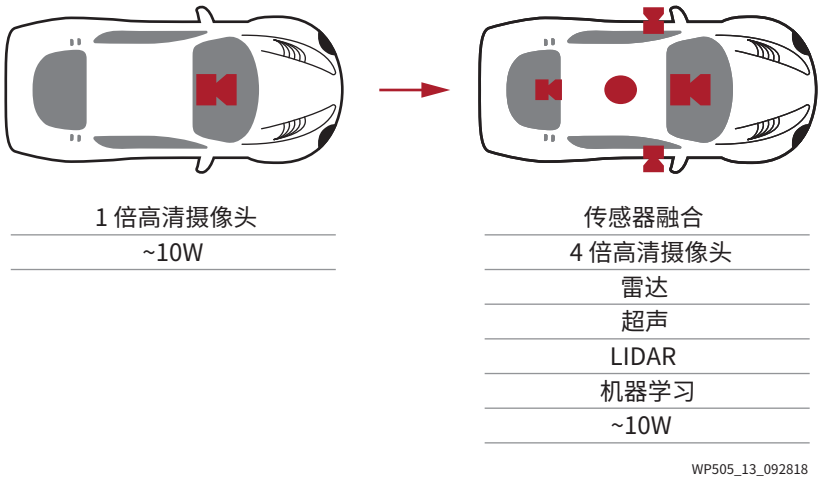

图 13： **赛灵思 ACAP 器件支持低功耗传感器融合**

汽车领域创新频现，重点在于要选择一种可跨多个平台提供代码可移植性和可扩展性的处理器件组合，从 5-10W 风挡安装的前置摄像头设计到 20-30W 座舱中央模块，再到 100W+ 液体冷却后备箱安装的超级计算机，所有这些都具有相同的编程模型。参见表 4。

表 4: **赛灵思汽车产品覆盖面与友商对比（同一编程模型）**

	(10W) 智能端点（例如前置摄像头）	(20W) 中央模块（基本型、无源散热）	(30W) 中央模块（高级型、风冷）	(100W+) 后备箱超级计算机（液体冷却）
赛灵思	•	•	•	•
Nvidia		❍	•	•
Intel MobilEye	•			

在考虑高速行驶的车辆时，时延是一大关键处理性能因素。在 60MPH (100KPH) 的速度下，不同 ADAS 系统的反应时间上几十毫秒的差异会对系统的有效性产生重大影响。随着自动驾驶汽车技术的日益普及，或需将多个神经网络串联执行复杂的任务，这加剧了 GPU 实现依赖大规模批处理的问题。因此，赛灵思优化了 AI Edge 系列，使其能够在低批处理规模下以极高的效率运行。参见图 14。

3. https://china.xilinx.com/news/press/2018/xilinx-announces-availability-of-automotive-qualified-zynq-ultrascale-mpsoc-family.html

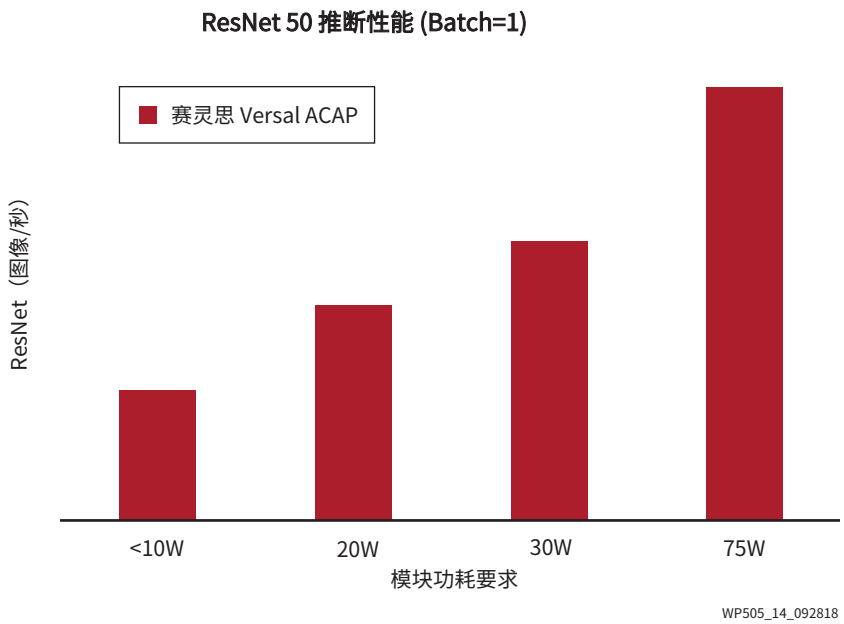

图 14： **低时延安全关键型 Versal 产品组合覆盖面**

当今汽车 ADAS/AD 系统对高分辨率摄像头的要求越来越高。计算需求根据像素进行扩展，这意味着来自高清摄像头 (1080X1920) 的图像比数据中心标准的 224x224 图像明显需要更强大的计算能力。高计算效率的赛灵思 Versal 器件扩展性定位独到，可满足更高的分辨率要求。

有线通信

今天，每一条互联网流量都经过多个赛灵思 FPGA 处理。FPGA 长期以来一直充当“胶水逻辑”，使网络硬件能够适应网络运营商不断变化的需求。赛灵思在最先进的 112G SerDes 技术领域的领先地位使业界能够首次实现新的协议和严格的光、铜电缆和底板标准，以及现有的 58G PAM4 和 32G NRZ 协议，例如标准应用前时期的 PCI Express® Gen5。丰富的 IP 组合使标准化接口的集成成为了可能，并降低了成本和功耗。赛灵思丰富的 IP 组合支持客户进行混合和匹配，从而在硬件级实现差异化。

随着网络运营商不断提出新的功能要求，快速编码和现场更新自适应硬件的能力比依赖原有 ASSP 的硬件更具优势。

赛灵思 Versal ACAP 具有与新一代 600G 波长规划一致的突破性集成 IP 水平，完全支持以太网和 OTN 标准 10G、25G、50G 和 100G SerDes 速率，包括：

- 10/25/40/50/100GE MAC/PCS/FEC，具有 ±1ns IEEE STD 1588 时间戳、eCPRI 和 TSN 支持
- 600G FlexE 核可实现低至 10G 通道和高密度 400GE/200GE/100GE MAC/PCS/FEC

- 600G 线速加密引擎，支持 MACSEC 和 IPSEC，以及批量 AES-GCM 加密
- 集成 FEC 的 600G Interlaken 用于 PAM4 通道
- 用于 DOCSIS 电缆 LDPC 应用的 SD-FEC

这些 SerDes 的显著改善能够支持：

- 用于 OTN 和边缘路由器应用的单芯片 1.0Tb/s+ 网线卡与商用 ASSP 相比具有类似的功率，但灵活性更高
- 单芯片 2.4Tb/s+ 加密数据中心互连 (DCI) 机架安装设备，每个 RU 有多个实例（参见图 15）
- 400Gb/s+ 电缆调制解调器终端系统 (CMTS)，每用户独有加密隧道，面向高级商业和住宅服务。

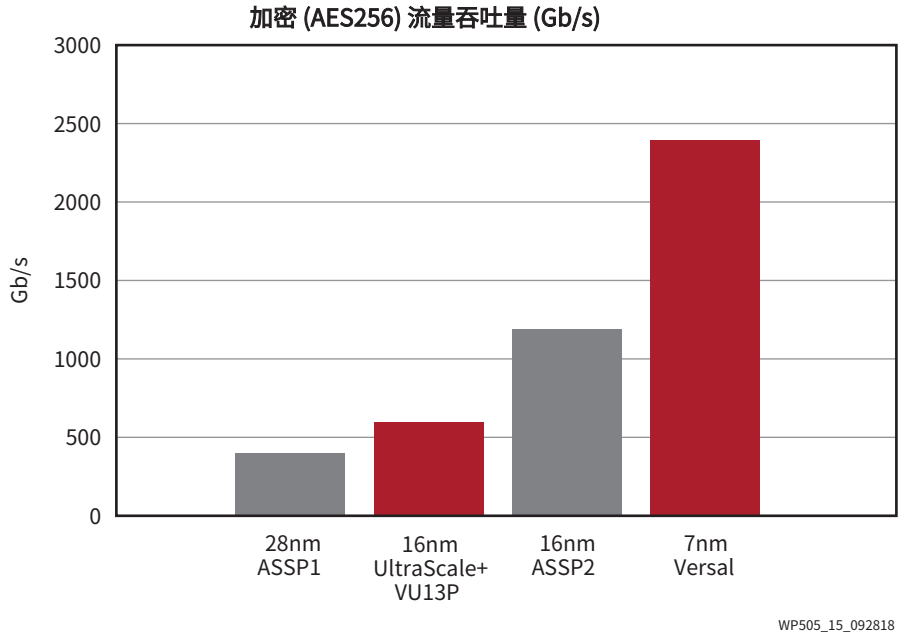

图 15： **有线通信：单芯片加密数据中心流量[1, 2]**

1. Microsemi DIGI-G4 OTN ASSP: https://www.microsemi.com/product-directory/multi-service-otn-processors/4227-pm5990-digi-g4
2. Microsemi DIGI-G5 OTN ASSP: https://www.microsemi.com/product-directory/multi-service-otn-processors/5056-pm6010-digi-g5-otn-processor

自适应性

可编程逻辑技术的一大优势在于现场硬件升级的能力。在今天的 4G 无线和光网络，以及汽车自动驾驶产品中已经广泛部署。

赛灵思 Versal ACAP 通过支持更高级别的抽象（C 或框架级接口）和 8 倍速的部分重配置，实现了更快的内核倒换，从而扩展了这一领域内的升级功能。

自适应硬件

FPGA 的核心价值主张长期以来一直是在现场进行设计变更的能力。无论是纠正错误，优化算法，或添加全新的功能，可编程逻辑提供了所有其他半导体选项不具备的独到灵活性。

赛灵思 Versal ACAP 将这一概念进一步推进，使配置时间加快了近一个数量级，实现了以毫秒为单位的部分比特流的动态倒换，让硬件具有软件的灵活性。

可编程存储器层级

作为一种补充，自适应硬件强化了 Versal ACAP，从而优化了 ACAP 架构新功能的效率。

可编程逻辑的最大优势之一是能够重配置存储器层级，从而针对不同的计算负载进行优化。例如，即使在专注图像识别的神经网络范围内，每幅图像的存储器占用和计算操作也会因算法的不同而带来很大的差异。可编程存储器架构支持对可编程逻辑进行调整，以优化它所支持的每个网络的计算效率。

因此，当结合矢量处理器和可编程逻辑来实现神经网络时，Versal ACAP 可达到领先的 GPU 近 2 倍的计算效率，并实现了固定存储器层级的矢量处理。参见图 16。

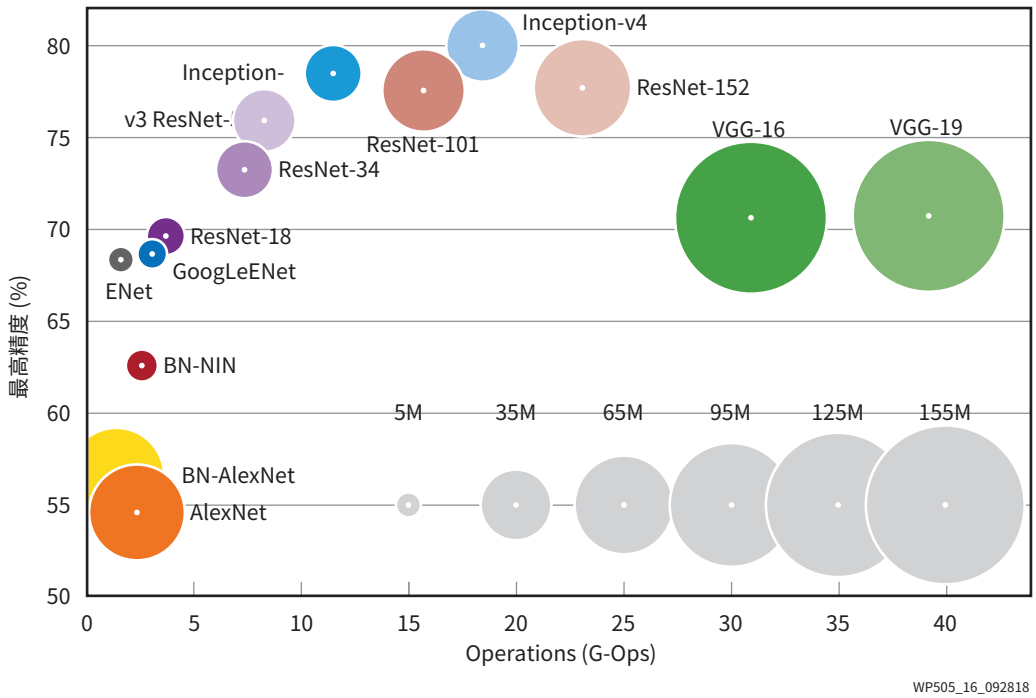

图 16： **基于神经网络类型的计算与存储器利用率**

动态重配置

该器件固有的可编程性将为某些成本敏感的实时应用带来优势，在多个逻辑功能之间复用一组可编程硬件，而且自适应引擎部分重新编程时间低至亚毫秒水平。在数据中心中，与 GPU 这样的更受限的矢量处理器相比，这意味着 Versal ACAP 器件能够执行传统上由 CPU 执行的更广泛的功能。（参见图 17，[参考资料 4]）

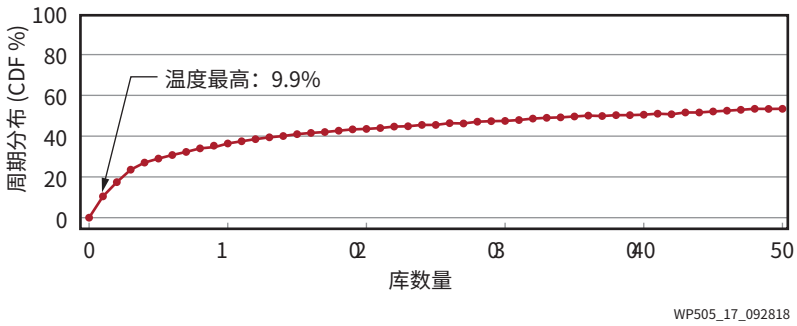

图 17： **由于数据中心工作负载经过广泛分配 (Kanev)，再不会产生“杀手应用”**

总结

近来涌现的技术挑战迫使业界跳出同构通用 (one-size-fits-all) 型 CPU 标量处理解决方案,进而探索新的发展方向。矢量处理 (DSP，GPU) 能够解决部分问题，但由于存储器带宽的使用效率不高，致其在传统的扩展中遭遇挑战。传统的 FPGA 解决方案提供了可编程存储器层级，但传统的硬件流程一直是阻碍推广的障碍。

该解决方案将所有这三大要素与一个新的工具流相结合，通过单个自适应计算加速平台 (ACAP)，提供了从框架到 C 到 RTL 级编码的各种不同抽象。

仅针对可编程逻辑一项，ACAP 架构就显著拓展了其能力。可编程逻辑和矢量处理单元的混合能够支持数据中心、无线网络、汽车驾驶辅助和有线通信中应用的计算量颠覆性的激增。

强大的 AI 机器学习计算、高级网络，以及加密 IP 的结合有助于面向数据中心实现新一类的自适应计算加速引擎以及智能 NIC。

将预制的人工智能机器学习推断与密集 DSP 和直接 RF 采样 ADC/DAC 相结合，与基于 DSP 的自开发 ASIC 相比，能将 5G 无线电的吞吐量翻一番，使 LIDAR、雷达和视觉传感器在汽车驾驶辅助 (ADAS) 应用中的单芯片传感器融合成为可能。

如需了解有关赛灵思 Versal ACAP 器件产品组合的详情,敬请访问赛灵思官网页面：
https://www.author.xilinx.com/products/silicon-devices/acap/versal.html.

参考资料

1. J. Hennessy, D. Patterson, *Computer Architecture: A Quantitative Approach* (6th Edition, 2019).
2. Nvidia, *Nvidia AI Inference Platform: Giant Leaps in Performance and Efficiency for AI Services, from the Data Center to the Network's Edge* (2018). Retrieved from nvidia.com, 2018.
3. N. Jouppi, C. Young, N. Patil, et al., In-Datacenter Performance Analysis of a Tensor Processing Unit™. In *International Symposium on Computer Architecture* (ISCA 2017). Retrieved from arxiv.org, 2018.
4. S. Kanev, J. Darago, K. Hazelwood, et al., *Profiling a warehouse-scale computer* (2015). Retrieved from google.com, 2018.

相关阅读材料

1. H. Esmaeilzadeh, E. Blem, R. St. Amant, et al., Dark Silicon and the End of Multicore Scaling. In *International Symposium on Computer Architecture* (ISCA 2011). Retrieved from gatech.edu, 2018.
2. M. Horowitz, Scaling Power and the Future of CMOS. In *20th International Conference on VLSI Design* (VLSID 2005). Retrieved from semanticscholar.org, 2018.
3. A. Putnam, Large-Scale Reconfigurable Computing in a Microsoft Datacenter. In *IEEE Hot Chips 26 Symposium* (2014). Retrieved from microsoft.com, 2018.

修订历史

下表列出了本文档的修订历史:

日期	版本	修订描述
2018年10月2日	1.0	赛灵思初始版本。

免责声明

本文向贵司/您所提供的信息（下称“资料”）仅在对赛灵思产品进行选择和使用时参考。在适用法律允许的最大范围内：（1）资料均按“现状”提供，且不保证不存在任何瑕疵，赛灵思在此声明对资料及其状况不作任何保证或担保，无论是明示、暗示还是法定的保证，包括但不限于对适销性、非侵权性或任何特定用途的适用性的保证；
且（2）赛灵思对任何因资料发生的或与资料有关的（含对资料的使用）任何损失或赔偿（包括任何直接、间接、特殊、附带或连带损失或赔偿，如数据、利润、商誉的损失或任何因第三方行为造成的任何类型的损失或赔偿），均不承担责任，不论该等损失或者赔偿是何种类或性质，也不论是基于合同、侵权、过失或是其他责任认定原理，即便该损失或赔偿可以合理预见或赛灵思事前被告知有发生该损失或赔偿的可能。赛灵思无义务纠正资料中包含的任何错误，也无义务对资料或产品说明书发生的更新进行通知。未经赛灵思公司的事先书面许可，贵司/您不得复制、修改、分发或公开展示本资料。部分产品受赛灵思有限保证条款的约束，请参阅赛灵思销售条款：http://china.xilinx.com/legal.htm#tos；IP 核可能受赛灵思向贵司/您签发的许可证中所包含的保证与支持条款的约束。赛灵思产品并非为故障安全保护目的而设计，也不具备此故障安全保护功能，不能用于任何需要专门故障安全保护性能的用途。如果把赛灵思产品应用于此类特殊用途，贵司/您将自行承担风险和责任。请参阅赛灵思销售条款：http://china.xilinx.com/legal.htm#tos。

关于与汽车相关用途的免责声明

如将汽车产品（部件编号中含“XA”字样）用于部署安全气囊或用于影响车辆控制的应用（“安全应用”），除非有符合 ISO 26262 汽车安全标准的安全概念或冗余特性（“安全设计”），否则不在质保范围内。客户应在使用或分销任何包含产品的系统之前为了安全的目的全面地测试此类系统。在未采用安全设计的条件下将产品用于安全应用的所有风险，由客户自行承担，并且仅在适用的法律法规对产品责任另有规定的情况下，适用该等法律法规的规定。