intel

边缘 AI 驱动，助力新质生产力

英特尔® 工业人工智能白皮书 2025 年版

Intel® Industrial AI Playbook 2025 Edition

编委会:

主编：刘　俊、马小龙、朱永佳

编委：方辛月、高　畅、高杨帆、胡　杨、刘　波、吕晓峰、邱丽颖、单　娜、张　恒、张心宇

* 编辑按姓名首字母排序

前言

人工智能（AI）技术的快速发展掀起了新一轮工业革命浪潮，通用大模型的出现让 AI 技术从专用化迈向了通用化。AI 技术正在步入工业领域的千行百业，帮助企业实现从传统的劳动密集型、资源密集型企业，向技术密集型、知识密集型的高端化、智能化、绿色化方向转型升级，打造依托于人工智能、大数据、云计算等现代信息技术的新质生产力。

工业 AI 和大模型的应用，已经渗透到工业生产的产品设计、企业流程管理规划、智能化生产、设备预测性维护、供应链优化、创新服务、绿色制造、智能客服等众多环节，它通过处理和分析海量工业数据，帮助企业在上述各个环节中做出最优的智能化决策，从而在多个环节全方位实现提质、增效、降本，增强竞争力。

在日趋激烈的工业市场竞争中，寻求部署新技术来提升综合竞争力，是企业的生存之道。而引领工业革命浪潮的 AI 技术和大模型，是企业从多维度重塑自身生产方式、实现新质生产力的关键。

通过这本白皮书，工业领域的企业和合作伙伴可以更系统、更全面地了解 AI 技术如何为工业制造的各个环节赋予怎样的智能化能力，以及英特尔在帮助企业落地部署 AI 技术方面所能提供的产品、平台和系统性支持与服务以及成功案例。

本白皮书中包括了工业 AI 和工业大模型的概念介绍、当前的市场规模与市场增长潜力、工业 AI 和工业大模型能为汽车、消费电子、新能源锂电、半导体制造等重点行业所带来的赋能创新机会，以及当前大模型在工业领域落地应用所面临的挑战和英特尔针对工业 AI 和大模型落地部署从硬件，到软件，到整体方案的技术赋能。

英特尔希望通过本白皮书，促进工业 AI 技术的广泛应用，并与行业伙伴共同探讨和制定工业 AI 的标准化流程和最佳实践，共同构建开放、协同的工业 AI 生态系统，推动制造业向智能制造转型升级，赋能新质生产力。

—张宇博士

英特尔中国区网络与边缘事业部首席技术官

目录

01 工业人工智能（AI）行业观察 01

1.1 工业 AI — 市场规模与增长潜力 02

1.2 工业 AI 的应用范畴 03

1.3 工业大模型 05

1.3.1 大模型 05

1.3.2 工业大模型 06

1.3.3 工业大模型的应用范畴 06

1.4 行业应用 07

1.4.1 汽车行业 07

1.4.2 消费电子行业 08

1.4.3 新能源锂电行业 08

1.4.4 半导体行业 09

1.5 工业 AI 与工业大模型落地应用面临的挑战 10

02 英特尔® 技术方案 12

2.1 硬件 13

2.1.1 第 12 代英特尔® 酷睿™ 移动处理器 13

2.1.2 英特尔® 酷睿™ Ultra 处理器 16

2.1.3 英特尔® 至强® Max 系列 & 英特尔® 至强® 6 处理器 20

2.1.4 英特尔锐炫™ 显卡 26

2.2 软件 29

2.2.1 英特尔® oneAPI 工具包 — 跨架构性能加速 29

2.2.2 OpenVINO™ 工具套件 32

2.2.3 英特尔® Geti™ 平台 35

2.2.4 英特尔® CVOI（工业机器视觉优化参考实现） 36

目录

2.3 创新技术方案 38

2.3.1 大语言模型赋能工业机械手臂 38

2.3.2 基于视觉大模型的零样本或少样本异常检测 40

2.3.3 RAG 检索增强生成模型实现 42

2.3.4 人形机器人 44

03 成功案例 46

3.1 英特尔：智能晶圆视觉检测 47

3.2 美的楼宇科技美控：楼宇 AI 节能解决方案 49

3.3 利珀：晶硅电池隐裂检测产品 51

3.4 诺达佳：基于 AI 的在线式视觉随动同步点胶机应用 53

3.5 新松：智能巡检机器人 55

3.6 华泰软件：智能化图纸生成管家 57

3.7 联想：基于 AI 的设备维护解决方案 58

04 合作伙伴加速项目和产品推荐 60

4.1 AI 硬件产品推荐 61

4.2 PIPC 工业电脑优选项目介绍 64

4.3 PIPC 机器视觉产品推荐 69

01

工业人工智能（AI）行业观察

1.1 工业 AI — 市场规模与增长潜力

工业 AI，是 AI 技术在工业领域的应用，它通过机器学习、深度学习、计算机视觉等先进的计算智能方法，实现对工业生产过程的优化和智能化，最终帮助企业提高生产效率、降低成本、提升产品质量，实现数字化转型。

2023 年 12 月，由信通院牵头、多家单位联合编制的《工业大模型技术应用与发展报告》指出，AI 与大模型将加速赋能新型工业化，预计从 2022 年至 2032 年，工业 AI 市场规模将以 46% 的年均复合增长率高速成长。

相较于发达国家，中国制造企业的 AI 应用率相对较低，大约在 11% 左右。Gartner 预测，到 2027 年，中国制造业的 AI 使用渗透率将以 10% 的年复合增长率上升。

随着技术的不断进步和应用场景的拓展，我们认为工业 AI 有望成为推动工业 4.0 和智能制造发展的关键力量。

1.2 工业 AI 的应用范畴

AI 技术在工业领域的应用，已经贯穿于产品设计、生产、管理、服务等众多环节，它主要通过各种方式收集海量数据，然后利用机器学习和统计模型对数据进行分析，并依据数据分析结果辅助决策，帮助企业优化资源配置，提质增效，节省成本。

具体来看，AI 技术在工业领域的应用主要在以下几大方面：

研发与规划

- **需求分析与预测：**基于历史数据和机器学习算法，构建预测模型，通过分析大量用户数据和市场趋势，洞察市场需求，预测未来趋势，精准定位产品的设计与迭代方向。
- **优化研发流程管理：**基于当前项目状态和历史数据建立预测模型，预测每项任务的完成时间，并评估整个项目的完成时间，有助于提前发现潜在延迟风险，让团队合理分配时间和其他资源，保证项目按时或提前完成。
- **自动化代码编写与优化：**AI 编程助手利用深度学习算法和大量代码数据训练模型，通过分析代码的结构和模式，并根据开发者的需求，自动生成函数、类、模块等代码，甚至优化现有代码，从而帮助开发者加速代码生成，减少错误。
- **优化产品结构与应用模拟：**通过形态识别技术，将产品外形及特征转化为数据，辅助设计师不断优化迭代。利用收集到数据构建数字孪生产品模型，模拟产品的各种实际应用场景，如正常操作、极限性能、潜在故障等，预测产品性能表现，进一步指导设计改进。

生产过程管控

在生产过程管控方面，AI 技术的应用主要集中在提高生产效率、优化资源配置、增强质量控制和实现生产过程的自动化与智能化。具体包括：

- **设备管理：**

  在设备入库管理方面，AI 通过深度学习识别设备上的条形码、二维码或设备特征，自动读取设备信息如型号、序列号等；AI 的自然语言处理功能，可以自动提取设备手册或标签上的文字信息，获取设备规格、性能指标等关键参数。这些都能显著提升设备入库管理的效率和准确性。

  在设备运维管理方面，利用机器学习算法，对部署在设备上的温度、压力、振动等各种传感器给出的监测数据进行处理分析，实时监控设备运行状态，并可通过模式识别算法检测数据中的异常，预测可能出现的故障或发现故障甚至给出修复建议，便于运维人员及时实施预测性维护或故障修复，减少停机时间，提高设备的可靠性和生产效率。

- **质量管理：**产品缺陷检测是质量管理的重要一环，尤其是对于金属等高反光产品、薄膜产品的划痕、裂纹、凹坑、气孔、污染等非常难检出的外观缺陷，利用传统视觉算法，对工业相机采集到的图像经过预处理，基于图像分割等深度学习模型，高效且较为准确地检出缺陷，为传统的视觉检测技术赋予高度智能化。质量检测也是目前 AI 技术在工业领域落地应用较多、较为成功的一个方向。

- **智能生产管理：**

  在生产计划和排程方面，AI 算法可以优化生产计划和排程，最大程度地减少产线空闲时间，提高产品交付准时率。

  在生产资源分配方面，通过深度学习和大数据分析，AI 系统能够根据实时数据预测生产任务，自动调整生产参数，并合理地分配人力、设备、物料等生产资源，提高资源利用率，确保生产线始终保持在最佳工作状态，提高生产效率。

  在生产过程监控和优化方面，AI 算法通过分析生产线上的各种运行状态反馈数据和工艺参数，能够预测及发现潜在问题，并自动调整参数，优化产线运行状态。

- **生产安全管理：**通过智能视频分析技术分析从生产现场采集的视频，进行行为识别与违规监测，如自动识别生产线上的工人是否穿了防护服、佩戴安全帽，是否进入违禁区等，并立即给出违规报警。还可以在仓库等重点防火区域部署智能视频分析系统，实时检测烟雾、火焰等火灾迹象，并快速触发报警。

此外，AI 技术在生产过程管控方面还可用于排产与调度优化、资源与物料管理、能耗与排放管理等环节，推动制造业向更高效、智能的方向发展。

经营管理优化

- **库存管理：**利用深度学习和大数据分析，分析历史销售数据、季节性变化、市场趋势等因素，预测库存需求、实时监控库存水平、自动调整补货策略、精准管理库存品类、优化库存地域布局等，提高库存周转率，降低库存成本。AI 聊天机器人可以随时了解 ERP 库存系统、跟踪订单和其他更新。

- **物流配送与运输管理：**机器人在深度学习算法和 3D 相机的加持下，可以识别被配送货物的形状、尺寸和条形码，自动分拣和归类，提高仓库分拣效率和准确性。利用大数据分析和机器学习优化配送路线，实时监控物流配送过程，提高配送效率、降低成本。

- **财务与人力管理：**通过训练模型，可以自动读取发票和收据，将其转换为数字格式，直接导入会计系统，减少了数据录入和处理的时间和错误。使用自然语言处理（NLP）和机器学习算法，能快速分析候选人简历，识别出与职位相关的教育背景、工作经历等关键信息，快速筛选出符合条件的候选人，提高招聘效率。

1.3 工业大模型

1.3.1 大模型

大模型（Large Model，也称基座模型，即 Foundation Model），是指具有大量参数和复杂结构的机器学习模型，能够处理海量数据、完成各种复杂的任务，如自然语言处理、计算机视觉、语音识别等。大模型通常包括大语言模型（LLM）、视觉大模型（CV）、多模态大模型等各种类型。

大模型通过训练海量数据来学习复杂的模式和特征，具有更强大的泛化能力，可以对未见过的数据做出准确的预测，能够处理更加复杂的任务和数据。

展开来讲，大模型技术有以下几项基本特征：

1. **普遍基于 Transformer 架构。**Transformer 架构通过引入自注意力（Self-Attention）机制，在处理序列数据时，能同时关注输入序列的所有元素，并直接建立任意两个元素之间的联系，从而捕捉序列中的长距离依赖关系，实现对输入序列的高效处理和理解。由于不依赖序列顺序，Transformer 架构在模型训练和推理时的并行处理能力更强，效率更高。

2. **参数规模大。**大模型通常包含数千万、数亿甚至更多参数；巨大的参数规模使大模型能够处理更加复杂和多样的任务。

3. **强大的泛化能力。**大模型通过在大规模数据集上进行训练，学习到了丰富的知识和特征表示，从而具有强大的泛化能力，能够有效处理多种从未见过的数据或新任务，甚至能处理一些与训练数据截然不同的任务。这使得大模型能应用于多种任务和场景，具有广泛的适用性。

4. **灵活性和可定制性。**大模型通常具有灵活的架构和可定制的参数，可以根据特定需求对通用大模型进行定制和优化。通过微调（Fine-tuning）技术，预训练的大模型可以快速适应新的任务和数据集，而无需从头开始训练。此外，还可以通过添加新的层或修改现有层的结构，来扩展大模型的功能和性能。

1.3.2 工业大模型

工业大模型，是指在工业生产中使用的大型模型。工业大模型在满足大模型技术基本特征的同时，具备在各个工业领域及工业各环节进行应用的能力，或在工业装备、软件等融合中赋能的模型。

相较于工业专用小模型而言，工业大模型泛化性强，可以单模型应对多任务，更适合长尾落地。另外，从工程层面来讲，工业大模型的开发成本及维护成本，低于工业专用小模型。

1.3.3 工业大模型的应用范畴

具体来看，工业大模型主要通过以下四种核心能力，为工业应用赋能：

第一，语言理解与知识问答能力。

利用大模型对于自然语言的理解能力，能理解和识别用户意图，使员工能通过自然语言就能与机器进行交互；另外通过为大模型外挂知识库，增强知识检索能力，可以提升知识获取和共享效率。这些能力在工业领域可普遍应用于智能客服、知识管理、教学与培训、工业文档检索与统计等场景中，大幅提升工作效率，减少人力劳动和成本。

还可以基于行业大模型提供知识问答/异常诊断/产线维护/排产建议，大幅提升制造效率，降低运维成本。

第二，创作与内容生成能力，如工业运控软件代码、设计模型、应用文档的生成。

在模型具备语言理解的基础之上，工业大模型具备了内容创作与生成的能力，这种内容生成的能力可大幅提高内容生成效率，提升员工工作效率。其与工业设备及系统的自然交互及推理的能力，可助力基于 LLM 工业代码的快速生成、优化与调试，大大促进工业应用的生成与落地。

第三，识别/模拟/预测能力。

在工业质检环节，用大量数据训练视觉大模型（CV），使模型具备更强的场景泛化识别能力，可用于产品质检，安全监测复判等流程，助力实现零样本或少样本缺陷检测。

在生产制造环节之外，工业大模型的仿真与模拟能力，亦可助力工业产品研发与设计环节。例如实时仿真模型的建立与仿真环境的创建。

在预测方面，工业大模型助力由原先局部建模预测至基于全局信息、更高效、高精度预测的转换与优化。

第四，多模态分析能力，由传统单一格式的工业数据处理，转化为多格式数据综合转换分析。

大模型不仅能够处理单一类型的工业数据，还能够综合分析多种格式的数据，实现跨格式的信息转换与分析。在工业应用中，大模型能同时处理包括设备运行数据、业务数据和管理决策数据在内的多种数据类型，为企业的运营和决策提供更为全面和精确的数据支持。

尽管目前工业大模型的应用已经渗透到工业的多个环节，应用场景较多，但碎片化明显。其中，知识管理/知识问答、数据助手/数据问答、专业内容生产以及视觉检测四个方向，是目前应用探索最多的领域。工业大模型经过一年多的发展，目前总体处于小规模商业应用落地阶段。

工业大模型凭借其卓越的理解、生成和泛化能力，通过与工业领域的深度融合，有望为工业领域带来“基础模型 + 各类应用”的新范式。因此，工业大模型的成功落地，离不开针对特定行业的丰富现场经验和深厚的行业 know-how 能力。

1.4 行业应用

1.4.1 汽车行业

汽车制造作为制造业皇冠上的明珠，也是 AI 技术落地应用的重要领域。目前，AI 技术已经渗透到汽车制造中繁多复杂的生产流程中，从汽车零部件的质量检测、到生产物流运输、装配生产线的自动化、再到整车质量检测等众多环节，AI 技术的使用都显著提高了生产效率和产品质量。

汽车造型辅助设计	工业大模型可广泛应用于汽车造型设计等领域。例如，在汽车造型设计中，设计师可通过对话、画图等方式与大模型交互，完善创意灵感，生成 3D 汽车数字模型，并能对模型进行风格调整、零部件编辑及颜色更换等操作。这能使原本需要 1-2 年的设计周期大幅缩短。
零部件及整车智能制造	汽车零部件和整车的性能，不仅关乎驾驶性能和体验，更关乎生命安全。因此，必须保证汽车零部件完好无缺陷，整车装配高度精准可靠，确保每一个部件都符合严格的安全标准。 例如，轮毂是汽车的重要组成部分，其质量直接关系到汽车的安全性和使用寿命。在轮毂的生产制造中，容易产生划痕、擦伤、气孔、毛刺、喷涂不到位、黑点等外观缺陷。缺陷的多样性、表面反光的干扰以及生产线上的实时检测要求，使得效率和准确率低下且容易漏检的人工质检和容易受复杂环境光干扰的传统机器视觉检测方法无法胜任。将 AI 视觉算法技术与机器视觉成像技术相结合，利用经过缺陷图像训练的深度学习模型识别工业相机捕获的缺陷图像，满足终端检测节拍要求 24 秒/轮毂，提高检测精度和生产线效率。
车身漆面质量检测	车身表面的涂漆质量是衡量整车品质的重要指标之一，它不仅关系到车辆的美观性，更事关车辆的防腐性、耐久性等问题。漆面喷涂环节工艺繁多复杂，易出现颗粒、缩孔、焊渣、脏污等各类缺陷，进而影响整车外观甚至漆面的耐久性。 传统的人工漆面缺陷检测方法，受检测人员自身状态及长时间工作易疲劳等因素的影响，无法精确检出各类缺陷，很难满足现代汽车生产需求。 在 AI 算法赋能下的 3D 成像技术，与机器人手臂协同作业，能够在线采集整车漆面数据进行并行计算，实现车身漆面缺陷的精准检测与定位，缺陷测量精度需达 0.15mm，检出率高达 99%，缺陷分类准确率 > 85%，需能够实现每车 60s 的检测节拍。还能支持多颜色、多车型在线混检，支持超过 20 余种漆面缺陷，实现多角度在线检测。AI 赋能的方案，大幅提升了车身漆面缺陷的检出率和检测效率，满足生产线的快速节拍需求。

1.4.2 消费电子行业

以智能手机、平板电脑、笔记本电脑等为主导的消费电子产品以及生产制造，也是 AI 技术和工业大模型落地应用的一个重点行业。

精准高效的缺陷检测

消费电子产品对品质要求极高，过检指标和漏检指标严格，且产线速度快。很多产品缺陷种类复杂、缺陷细小、区分度低，传统的人工检测和机器视觉方案，检出率低，速度慢，无法满足生产质量和高速产线的节拍要求。AI 技术与机器视觉检测方案相结合，为这类难检缺陷提供有效解决方案。

以手机玻璃盖板为例，手机玻璃盖板在生产过程中可能会出现划痕、蹭伤、崩边、气泡、手指纹、水迹、水印等多种微小且不易察觉的缺陷，缺陷种类最多可达 30 多种。必须精准、高效地检出这些缺陷以保证产品质量，检测精度一般要求达到 10μm，检测节拍根据盖板尺寸大小通常在 6 秒到 1 秒/件之间，甚至更快。

传统的人眼检测，不但无法达到微小缺陷的检测精度要求，而且人眼容易疲劳，存在效率低、误检漏检偏高等问题，无法满足生产的精度和节拍要求。将深度学习算法与高精度成像系统相结合，更快速地识别出产品图像中的缺陷及种类，满足生产线对检测精度和速度的要求。

智能化功能增强

更加个性化、智能化、功能强大的手机、PC 等消费电子产品，是驱动消费电子产品更新换代和市场复苏的关键因素。

消费电子产品将是大模型部署的新阵地。围绕用户的个性化需求，包括不同的使用场景和使用习惯等，大模型的部署需要根据用户特征对模型进行差异化增强。为了保护数据隐私，与用户隐私相关的应用模型的训练，将在端侧而非云上进行，这也对边缘端的算力提出了更高要求。

加速产品的更新换代

消费电子产品的特征之一是快速更新迭代，快速上市新产品意味着抢占市场先机。

在新产品的设计生产方面，基于 AI 的市场需求预测模型能快速分析消费者需求趋势，辅助设计/生产软件能基于历史数据和现有数据加速新产品设计，优化生产管理流程，快速上市新产品。

1.4.3 新能源锂电行业

AI 技术强大的计算和分析能力，已经为锂电制造行业带来巨大变革，从材料选型、器件设计和优化生产保障质量方面，帮助锂电制造企业缩短开发周期，提升检测效率，控制成本投入。

锂电池质量检测

锂电池的质量直接关乎电动车的安全性，因此锂电对质检要求严苛。锂电生产过程中的检测工序繁多，包括原料生产中的隔膜缺陷检测，前段工序中的极片表面缺陷检测、涂布外观缺陷检测，中段工序中的密封钉焊道缺陷检测、电池包蓝膜后缺陷检测，后段工序中的 Busbar 焊后检测等。目前锂电检测的主要痛点在于：如何以接近 100% 的检测良率精准地检测出多种复杂难检的缺陷；同时质检速度还要跟上生产节拍，以保证甚至提升产能。

以电芯顶盖板焊接质量检测为例，在将电芯顶盖焊接到电池壳体的过程中，很容易出现爆点、焊坑、孔洞、断焊、漏焊、翻边等缺陷，导致漏液、短路等安全风险。将 AI 技术与 3D 成像技术相结合，利用数据样本自适应扩充训练技术，缩短模型训练时间，通过针对性的缺陷检测算法，提高了缺陷检测效率和准确率，降低了工人检测的过杀、漏杀情况，实现缺陷检测无人化，降低人力成本。

新材料的快速筛选	锂电池未来的技术核心竞争点在于材料。快速筛选出高能效的材料，是掌握竞争优势的关键。大模型通过高通量计算与数据库构建、分子生成模型和高通量筛选策略等步骤，能从数百万种材料中，快速筛选出具有高能效潜力的材料，缩短新材料的发现周期。 高效能材料的发现，直接关系着电池的能量密度、性能表现、使用寿命、安全性和成本等关键指标。电池企业正在材料筛选及研发上积极探索 AI 技术的深入应用。
加速设计	在锂电池设计方面，利用 AI 高效仿真模型，可以在原子、分子、颗粒、电极和电芯等多个尺度上进行仿真模拟，让研发人员更深入地理解电池内部的作用机理，并在此基础上快速优化材料和结构设计，缩短设计时长。

1.4.4 半导体行业

半导体制造作为一个高度复杂、技术密集、资本密集的行业，如何实现产品的快速设计、确保生产过程的精度和良率，以保障研发和生产成本的良性投入，最终满足市场对芯片产品的快速更新迭代需求，是半导体行业面临的痛点问题。

加速集成电路芯片设计流程	随着制造工艺提升，集成电路芯片制造的工艺线宽不断缩小，这将带来更复杂和更大规模的电路设计，传统 EDA 设计流程在应对设计规则复杂度、功耗及热管理、信号完整性等方面面临一系列挑战。 将 AI 技术与 EDA 工具相结合，在电路设计阶段，AI 可以自动识别和优化电路拓扑结构，通过深度学习模型预测不同电路设计的性能指标（如功耗、速度、面积等），从而快速筛选出最优设计方案。这种方法大大减少了人工试错的时间，加速了设计迭代过程。在布局布线阶段，优化布局布线是集成电路设计中最为耗时的步骤之一，涉及到芯片上数百万甚至数十亿个元器件的物理位置和连接。AI 技术可以在此阶段通过强化学习等方法，自动学习最优的布局策略，实现快速而高效的布局布线，同时优化信号完整性、功耗和热管理等关键指标。
晶圆缺陷检测	半导体晶圆制造过程极为复杂、精密，任何微小缺陷都可能影响芯片性能。晶圆中常见的缺陷包括表面的划痕、裂纹、污染物、凸起，表面翘曲，切割瑕疵、晶体缺陷等。这些缺陷大多细微不易察觉，通常需要微米级甚至更小的检测精度。人工检测效率低下，易出错，无法满足大规模生产的效率需求；传统的机器视觉检测算法，无法满足对多种缺陷的检测需求。 采用大模型结合机器视觉成像技术，首先使用大规模无标注图像对大模型预训练，然后再针对晶圆缺陷检测任务，在标注的晶圆缺陷图像数据集上进行微调，优化模型对微小缺陷的识别能力。最终经过优化的大模型，在晶圆缺陷检测任务上，最小能检出 0.1 微米级别的缺陷尺寸，检测精准度需高于 99.5%，检测节拍大多需达 300 片/分钟以上，检测精度、检出率和检测效率都比传统方法有大幅提升，满足大规范生产需求。

1.5 工业 AI 与工业大模型落地应用面临的挑战

毋庸置疑，AI 技术的应用正为工业领域带来前所未有的创新性变革。而且，工业领域对 AI 技术的部署，正在随着 AI 技术本身的发展和工业应用复杂性的增加，日渐从传统的 AI 技术向更加复杂的工业大模型过渡。

传统 AI 技术则主要基于规则和知识库实现智能工作，它通常使用神经网络结构，通过大量数据进行训练，来获得较好的性能。传统 AI 具有较强的实时性，能在特定场景下快速解决问题。但是，对于更加复杂的多样化应用场景，比如需要处理文本、图像、音频等多模态数据时，传统 AI 的落地还是有差距；而大模型凭借强大的自学习能力和泛化能力，以及与具体行业数据的结合调优，优势明显。

大模型的出现，将 AI 技术在工业领域的应用推向了新的发展阶段。其具体落地将会以基础大模型为技术底座，融合工业细分行业的数据和专家经验，形成垂直化、场景化、专业化的工业大模型。工业大模型相对基础大模型具有参数量少、专业度高、落地性强等优势，可以为工业垂直领域的技术突破、产品创新、生产变革等提供低成本解决方案。

尽管传统 AI 技术和大模型在解决各种工业问题方面，从理论上讲存在诸多明显优势，但是要将 AI 技术和大模型真正成功落地应用，依然有很多具有挑战性的问题亟待解决。

第一，数据问题。

无论是传统工业 AI 技术，还是工业大模型的落地应用，数据都是首要问题。首先是数据的数量问题，如何从应用场景中收集到大量的数据作为训练算法或模型，是算法或大模型具备更智能化分析和决策能力的基础。而往往很多时候来自工业现场的数据量非常有限甚至极少。其次是数据质量问题，即数据的清洁性，并非所有来自工业现场的数据都是有用的，需要对数据进行清洁。如何从实际应用场景中采集或生成丰富且有价值的可用数据，是 AI 及工业大模型成功落地应用的挑战之一。再次是数据的标注和处理，即便有了足够的数据，对这些数据进行标注和处理也在难度和工作量方面面临极大挑战。最后是数据安全和隐私问题，数据是 AI 技术及工业大模型应用的基础，这些来自应用端的数据，其中包含着技术、工艺机密信息或个人隐私信息。如何在数据传输、训练、处理过程中保用户数据的安全性和防止数据滥用，也是工业 AI 乃至工业大模型成功落地应用的挑战之一。

第二，算力问题。

无论是训练 AI 算法还是各种工业大模型，都需要强大的算力支撑。工业大模型动辄参数规模都在十亿、百亿甚至千亿级别，需要庞大的计算资源进行训练。这种训练过程涉及海量的数据运算，对 CPU、GPU 或 NPU 等加速计算硬件提出了极高的要求。

第三，实时响应问题。

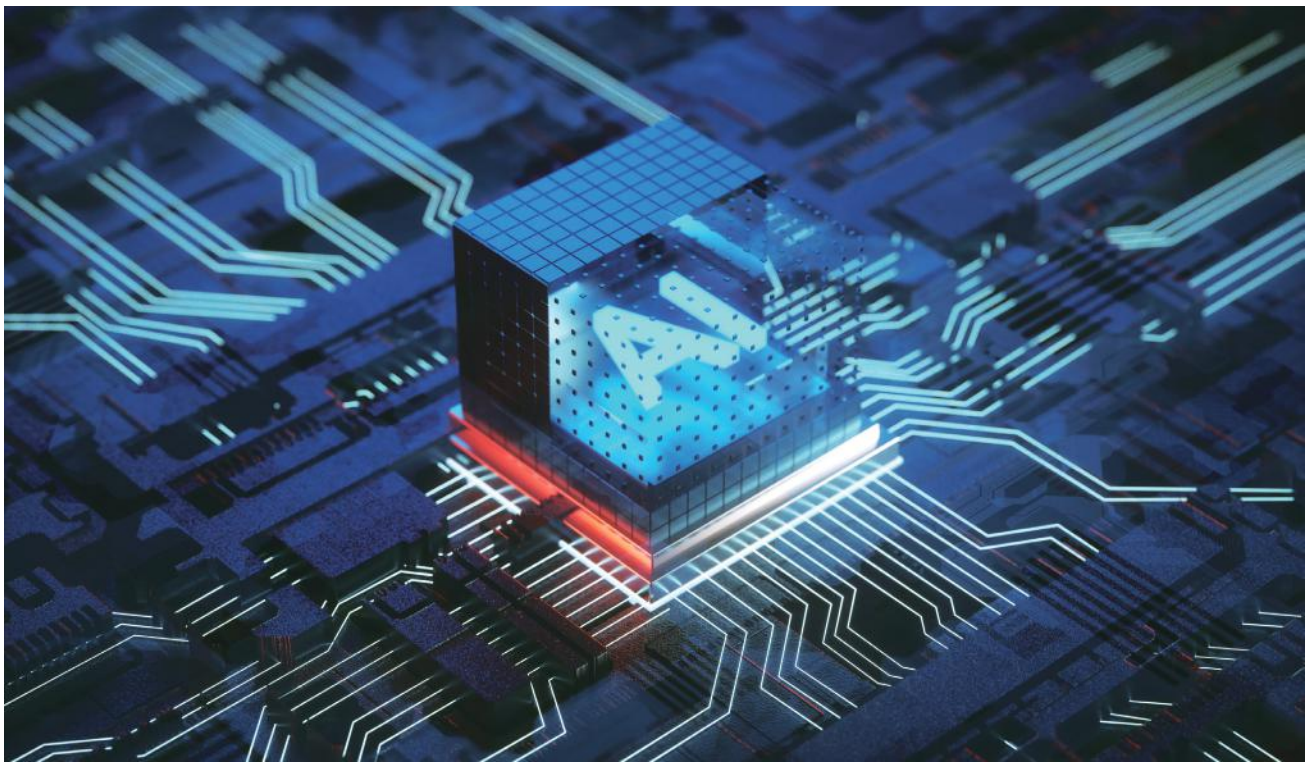
工厂在线检测、智能驾驶等应用，需要系统实时做出响应的情况下，需要模型能够实时处理输入数据并快速做出响应。将场景应用端的数据再传输到云端处理，庞大的数据量会造成带宽拥挤，影响处理的时效性。采用边缘计算方案来缓解时效性问题，但是这对边缘端计算硬件的实时处理能力提出了挑战。

第四，模型应用准确性问题。

工业大模型在实际应用中的准确度尚不尽人意。目前大模型比较擅长知识问答、文档生成、数据分析等场景应用，但在面向实际工程的代码生成能力仍有很大提升空间，尤其在实用算法、科学计算和数据结构等领域能力偏弱。另外，针对缺陷样本极少的工业质检应用场景，工业大模型基于真实缺陷图生成仿真缺陷图的能力，目前在准确性方面依然有待提升。

第五，成本和技术问题。

工业 AI 技术尤其是工业大模型的部署，要面对高昂的训练成本和技术壁垒，这往往令中小企业望而却步。工业大模型的应用不仅需要大量的资金投入，还需要专业的人才支持。包括技术研发、算力资源投入、数据采集与标注，以及市场推广与商业化扩展等方面都需要专业的人才进行操作和管理。在技术壁垒方面，数据处理难题、算力基础设施部署、商业落地的可靠性等方面，都面临挑战。前期需求高昂的投入成本，而项目的投入产出却难以清晰测量，也在阻碍了落地应用的实施。



02

英特尔®
技术方案

2.1 硬件

2.1.1 第 12 代英特尔® 酷睿™ 移动处理器

第 12 代英特尔® 酷睿™ 移动处理器为物联网部署创造更多价值，采用全新高性能混合架构，大幅提升单线程和多线程性能，其高性能小尺寸的设计兼顾了图形密度和 AI 加速功能。

首款采用高性能混合架构的英特尔® 酷睿™ 处理器

创新的芯片设计将专注于主要工作负载的 P-core（性能核）与专为多任务处理而建构的 E-core（能效核）相结合。英特尔® 硬件线程调度器可智能指示操作系统将适当的工作负载与合适的内核相匹配。

硬件加速核英特尔® 锐炬® X 显卡成就出色的 AI 功能

大量的图形 EU 同样便于 AI 推理，可提高 AI 工作负载常用数学运算的并行程度。该平台还通过英特尔® 深度学习加速技术（英特尔® DL Boost）和 VNNI 指令支持基于硬件的 AI 加速，通过 Int8 量化实现强大的 AI 性能。平台支持英特尔® 发行版 OpenVINO™ 工具套件，可提供优化的性能，同时帮助开发人员对常见用例进行 AI 模型预训练，从而加快上市时间。

第 12 代英特尔® 酷睿™ 移动处理器	高达	高达	高达	高达
性能测量结果基于同第 11 代英特尔® 酷睿™ 处理器的比较[1]	1.07 倍	1.29 倍	2.47 倍	2.77 倍
	单线程性能提升[1]	**多线程性能提升**[1]	**显卡性能提升**[1]	**GPU 图像分类推理性能提升**[1]

主要特性

性能和效率

- 英特尔® 7 制程工艺
- 多达 14 个核心和 20 个线程，具有高性能混合架构
- Intel® Thread Director6 使您的核心与工作负载相匹配
- 高达 24 MB Intel® 智能缓存

确定性实时性

- 利用英特尔® TCC 进行实时计算
- 支持时间敏感型网络（TSN）
- 通过英特尔® PLL 锁相环技术，可锁单 P 核或者 4 个一组 E 核作为实时任务，而其他核按需动态调整频率

工业特性

- IBECC 内存
- 处理器基本功率范围为 15W 至 45W，低功耗 SKU 支持无风扇设计
- 工业级 SKU 支持宽温运行

AI 加速

- 英特尔® 锐炬 X 显卡拥有多达 96 个执行单元（EU），便于视觉识别、测量以及视觉引导等应用中高度并行的 AI 工作负载处理
- 通过在 CPU 上运行包含 VNNI 指令的英特尔® DLBoost 在 GPU 上运行 DP4a（int8）指令，以及采用英特尔® 发行版 OpenVINO™ 工具套件，加速 AI 推理工作负载

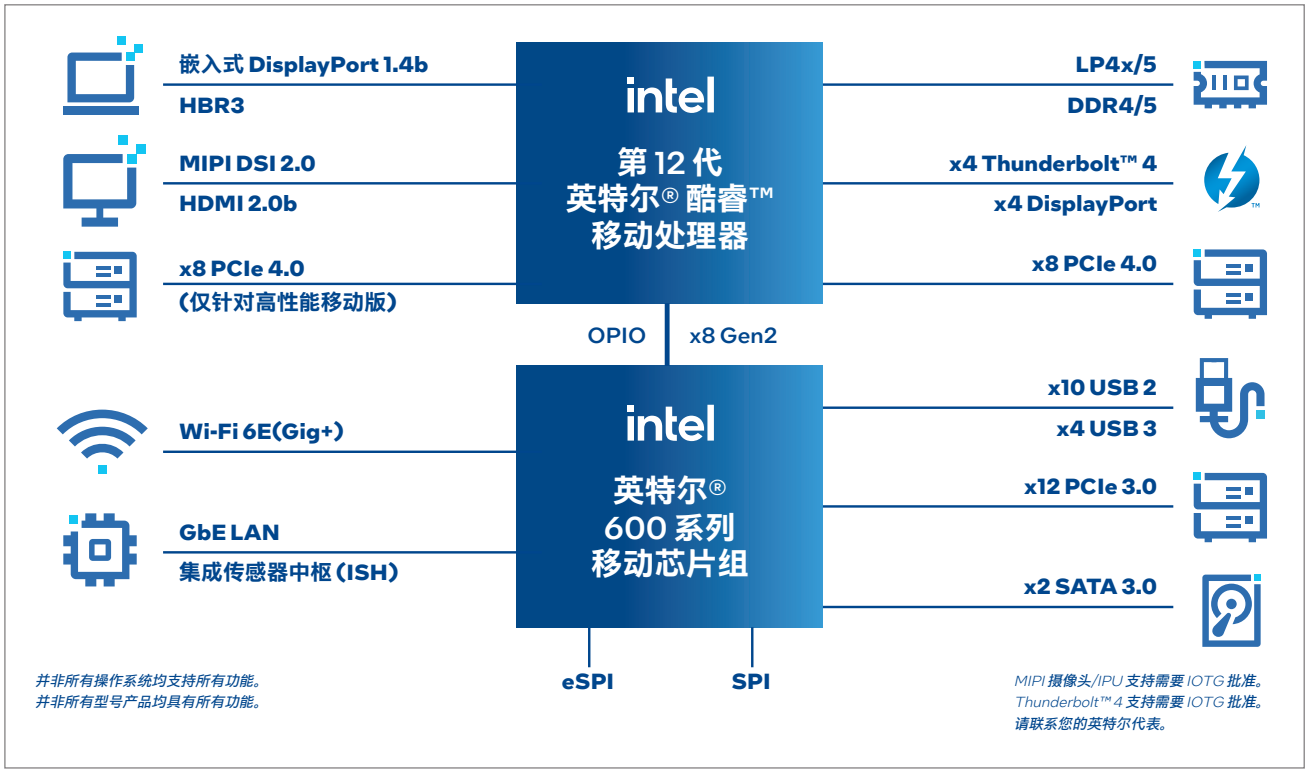
管理与安全

- 英特尔® vPro 平台适用于特定 SKU
- 英特尔® 融合安全管理引擎（Intel® CSME）版本 16

操作系统支持

- Windows 10 IoT 企业版 2021 长期服务频道（LTSC）
- 支持 EFLOW
- Linux 内核覆盖，可轻松采用物联网功能
- Celadon（Android）（社区支持）
- 支持 Ubuntu、Red Hat Enterprise、Wind River Linux 和 Wind River VxWorks 7

第 12 代英特尔® 酷睿™ 移动处理器示意图



并非所有操作系统均支持所有功能。
并非所有型号产品均具有所有功能。

MIPI 摄像头/IPU 支持需要 IOTG 批准。
Thunderbolt™ 4 支持需要 IOTG 批准。
请联系您的英特尔代表。

第 12 代英特尔® 酷睿™ 移动处理器产品线

第 12 代英特尔® 酷睿™ 处理器（高性能移动版 45W）

处理器编号	处理器内核数	P-core 数	E-core 数	线程数	英特尔® 智能高速缓存 (L3)	最大睿频频率 (GHz)[A]		处理器基础频率 (GHz)		最大显卡频率 (GHz)	英特尔® 平台	固件支持的版本和类型		处理器显卡	执行单元 (EU) 数	视频解码器	PCIe 通道总数	最大内存速度	最大内存容量	处理器基础功率 (W)
						P-core	E-core	P-core	E-core		英特尔® vPro® Enterprise[B]	ME16	ME16							
英特尔® 酷睿™ i7-12800HE 处理器	14	6	8	20	24 MB	高达 4.6	高达 3.5	2.4 (@45W) 1.6 (@35W)	1.8	1.35	是	企业	消费者	英特尔锐炬® X[e] 显卡[D]	96	2	16 (CPU) 12 (PCH)	DDR5-4800 LPDDR5-5200 DDR4-3200 LPDDR4x-4267	64 GB	45W（基础功率） 35W（最小保证功率）
英特尔® 酷睿™ i5-12600HE 处理器	12	4	8	16	18 MB	高达 4.5	高达 3.3	2.5 (@45W) 1.7 (@35W)	1.8	1.3	是	企业	消费者		80	2				
英特尔® 酷睿™ i3-12300HE 处理器	8	4	4	12	12 MB	高达 4.3	高达 3.3	1.9 (@45W) 1.1 (@35W)	1.5	1.15	否	企业[C]	消费者	英特尔® 超核芯显卡	48	1				

了解有关第 12 代英特尔® 酷睿™ 移动处理器的更多信息，请访问：
https://www.intel.cn/content/www/cn/zh/products/platforms/details/alder-lake-p.html

1. 性能测试结果基于配置信息中显示的日期进行的测试，且可能并未反映所有公开可用的安全更新。预测或模拟结果使用英特尔内部分析或架构模拟或建模，该等结果仅供您参考。系统硬件、软件或配置中的任何差异将可能影响您的实际性能。关于性能和基准测试程序结果的更多信息，请访问：intel.cn/PerformanceIndex

2.1.2 英特尔® 酷睿™ Ultra 处理器

新特性

- 基于极紫外（EUV）光刻技术的英特尔 4 制程工艺
- 单个 SoC 内配备众多计算引擎：P-core（性能核）、E-core（能效核）、英特尔锐炫™ GPU[2] 以及 AI 专用的内置神经处理单元（NPU）英特尔® AI Boost[3] 单个 SoC 内配备众多计算引擎：P-core（性能核）、E-core（能效核）、英特尔锐炫™ GPU[2] 以及 AI 专用的内置神经处理单元（NPU）英特尔® AI Boost[3]
- 内置英特尔锐炫™ GPU[2]，提供多达 8 个 X^e 内核（多达 128 个图形执行单元）
- 硬件加速 AV1 编码、内置 DisplayPort2.1（USB-C）和 HDMI 2.1，以及全新图形系统控制器

英特尔® 酷睿™ Ultra 处理器

人工智能	能效	图形处理
高达	高达	高达
1.5 倍	2.56 倍	1.81 倍
AI 性能提升	每瓦 AI 性能提升	图形处理性能提升
与上一代产品比较[1]	与上一代产品比较[1]	与上一代产品比较[1]

实际性能受使用情况、配置和其他因素的差异影响。更多信息请见 intel.com/processorclaims（英特尔® 酷睿™ Ultra 处理器 — 边缘）。结果可能不同。

采用高能效 BGA 封装，以先进的 AI 和图形处理性能，助力部署边缘解决方案

即使在空间和功耗受限的环境中，也能快速轻松地在边缘部署 AI 和图形处理功能，以满足视觉和自动化等用例的需求。英特尔® 酷睿™ Ultra 处理器配备众多计算引擎，采用高能效 BGA 封装，能够为创新设计提供更大的灵活性，是应对边缘严苛工作负载的理想选择。这些功能强大的边缘处理器可以加速从 AI 获取结果，为每台设备提供更多媒体流，并提供长期供货保证[2]，以提升长期价值。

单个封装内部署更多 AI 引擎

利用英特尔® 酷睿™ Ultra 处理器提升竞争优势，部署客户迫切需要的先进 AI 工作负载。P-core（性能核）、E-core（能效核）、英特尔锐炫™ GPU[3] 以及英特尔® AI Boost[4] 等众多计算引擎协同加速边缘 AI 推理，同时减少对独立加速器的需求，帮助降低系统复杂性和成本。

此外，该款处理器支持英特尔® 发行版 OpenVINO™ 工具套件，可为工作负载匹配合适的计算引擎，从而提高 AI 性能，并能够通过跨架构编程功能和自动计算引擎检测，帮助简化 AI 工作流程。OpenVINO™ 还为 TensorFlow*、PyTorch* 和 ONNX 等主流 AI 框架提供支持和优化，以帮助提高性能并简化开发工作。另外，英特尔® Gaussian & Neural Accelerator（英特尔® GNA）3.5 可用于改善音频降噪和语音识别。

提升图形密集型应用性能，无需入门级独立 GPU

为自助服务终端、终端以及细节丰富的界面整合系统并降低硬件成本。英特尔® 酷睿™ Ultra 处理器配备内置英特尔锐炫™ GPU[3]，提供多达 8 个 X^e 内核（多达 128 个图形执行单元），有助于减少对入门级独立 GPU 的需求。这一代处理器支持多达 50 个 HDR 视频流，可提供细节更加丰富的视效，支持在硬件加速主流 AV1 编解码器，可实现比 H.265 更高效的压缩。对于高级视频墙应用，英特尔® 酷睿™ Ultra 处理器支持多达 4x 4K 显示器或 2x 8K 显示器、通道锁定同步和边框校正功能。

降低要求严苛的 AI 和视频工作负载的能耗

借助能效优于上一代产品的平台简化边缘 AI 部署[5]。英特尔® 酷睿™ Ultra 处理器采用 BGA 封装，在相同功耗水平下，可提供比上一代产品更高的 AI 性能，让终端客户能够在空间有限的环境中灵活运行更多的工作负载。这一平台非常适合需要无风扇或较少散热的边缘设计，同时还优化了电源设计，可通过控制活动较少时段的能耗，帮助降低能耗成本。

英特尔® 酷睿™ Ultra 处理器还配备英特尔® 硬件线程调度器[5]，可以对 CPU 内核间的并行工作负载进行智能优化。它通过识别每个工作负载的类别并使用能效核和性能核评分机制，帮助操作系统合理调度内核线程，以提高性能或能效。

主要特性

性能

- 基于 EUV 光刻技术的英特尔 4 制程工艺
- 采用英特尔® 酷睿™ 处理器的高性能混合架构，配备英特尔® 硬件线程调度器[5]
- 多达 16 个内核和 22 条线程
- 多达 24 MB 的英特尔® 智能高速缓存
- 15 W 至 45 W 的处理器基础功耗范围

加速 AI

- 单个 SoC 内配备众多计算引擎：P-core（性能核）、E-core（能效核）、英特尔锐炫™ GPU[2] 以及英特尔® AI Boost[4]
- 经优化的英特尔® Gaussian & Neural Accelerator（英特尔® GNA）3.5
- 英特尔® 深度学习加速技术（Intel® Deep Learning Boost，英特尔® DL Boost）与 DP4a 指令
- 受 OpenVINO™ 工具套件全面支持

能效

- 2 个低功耗嵌入式 DisplayPort 接口

图形处理

- 内置英特尔锐炫™ GPU[3]，提供多达 8 个 Xᵉ 内核（多达 128 个图形执行单元）
- 硬件加速 AV1 编码
- 集成的 DisplayPort 2.1（USB-C）和 HDMI 2.1
- 图形系统控制器（GSC）
- 集成的英特尔® 图像处理单元
- Windows 通道锁定视频同步，带边框校正功能和 EDID 管理/显示锁定
- 多达 50 个同步 HEVC HDR 10b 1080p30 视频流
- 多达 4 个并发 4K60 HDR 显示器或 2 个 8K 显示器
- 基于 SR-IOV 的 GPU 虚拟化

内存和 I/O

- 高达 LPDDR5-6400、LPDDR5x-7467（Type 4 载板）、DDR5-5600
- 8 条 PCIe 5.0 通道[7]
- 多达 20 条 PCIe 4.0 通道

高能效部署与长期支持

- 焊入式（soldered-down）BGA 封装
- 长达 10 年的长期供货保证[2]

安全性与可管理性

- Elemental security engine（ESE）
- NIST 800-88r1（存储介质清理）

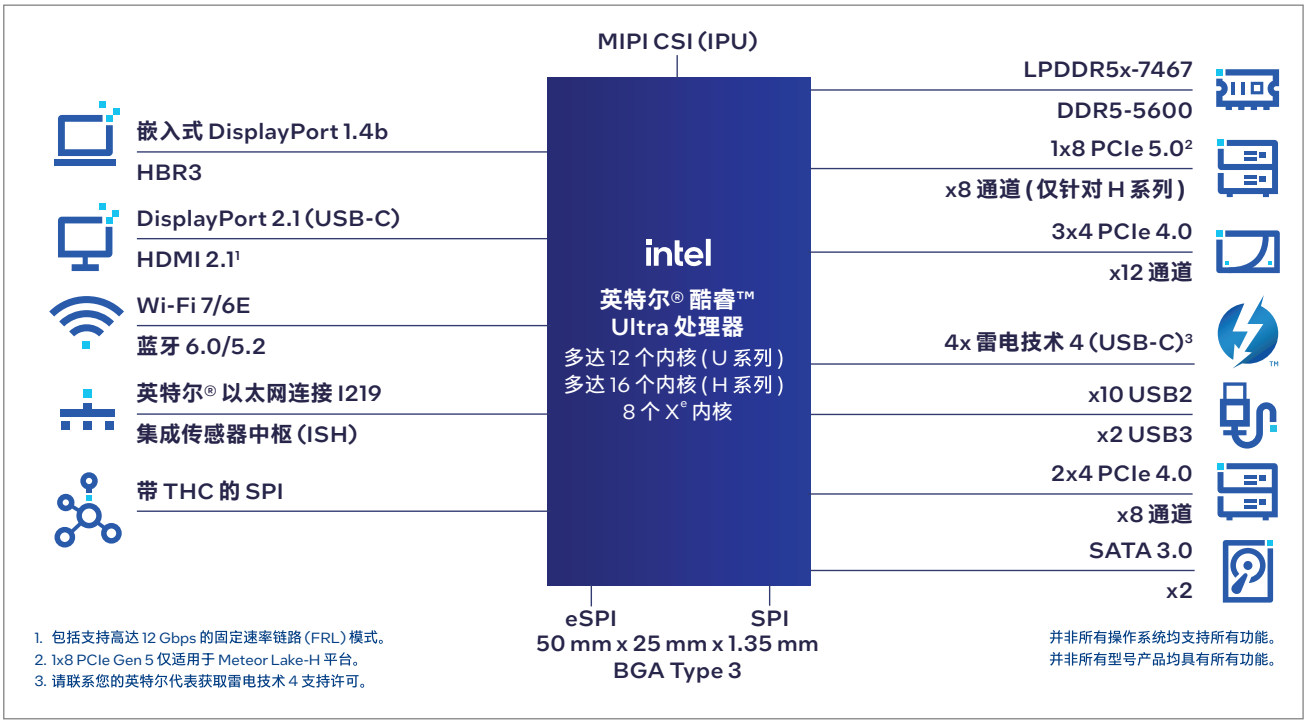
连接

- USB4/英特尔® 雷电技术 4[6]
- 经过英特尔® 独立 Wi-Fi 7（英特尔® Wi-Fi 7 BE200、英特尔® Wi-Fi6E AX210）验证
- 蓝牙 5.4/5.3

软件和操作系统支持

- OpenVINO™ 工具套件、英特尔® oneAPI 工具套件、英特尔® oneAPI Video Processing Library（英特尔® oneVPL）
- Windows 10 IoT Enterprise 2021 LTSC 和 Windows 11 IoT Enterprise 2024 LTSC（2H'24）
- Ubuntu、Red Hat Enterprise Linux、Wind River Linux
- Azure IoT EFLOW、Yocto Project 和基于 Linux 内核的虚拟机（KVM）
- UEFI/BIOS 和英特尔® 固件支持软件包（Intel® Firmware Support Package，英特尔® FSP）以及 Slim Bootloader 和英特尔® FSP

英特尔® 酷睿™ Ultra 处理器示意图



1. 包括支持高达 12 Gbps 的固定速率链路 (FRL) 模式。
2. 1x8 PCIe Gen 5 仅适用于 Meteor Lake-H 平台。
3. 请联系您的英特尔代表获取雷电技术 4 支持许可。

并非所有操作系统均支持所有功能。
并非所有型号产品均具有所有功能。

英特尔® 酷睿™ Ultra 处理器产品线

英特尔® 酷睿™ Ultra 处理器 (H 系列，28 W)

处理器名称	处理器内核数	P-core 数	E-core 数	LPE 内核数	线程数	英特尔® 智能高速缓存 (L3)	最大睿频频率 (GHz)[A]		处理器基础频率 (GHz)		最大显卡频率 (GHz)	处理器显卡	执行单元 (EU) 数	视频解码器	PCIe 通道总数	最大内存速度	最大内存容量	TCC/ TSN	宽温支持	处理器基础功耗 (W)
							P-core	E-core	P-core	E-core										
英特尔® 酷睿™ Ultra 7 处理器 165H	16	6	8	2	22	24 MB	5.0	3.8	1.4 (@28W)	0.9	2.3	英特尔锐炫™ 显卡[B]	128	2	8 (CPU: 1x8 PCIe 5.0) 20 (PCIe 4.0)	DDR5-5600 LPDDR5-6400 LPDDR5x-6400 LPDDR5x-7467	64 GB	否	否	65 W (最大保证功耗) 28 W (基础功耗) 20 W (最小保证功耗)
英特尔® 酷睿™ Ultra 7 处理器 155H	16	6	8	2	22	24 MB	4.8	3.8	1.4 (@28W)	0.9	2.25		128	2				否	否	
英特尔® 酷睿™ Ultra 5 处理器 135H	14	4	8	2	18	18 MB	4.6	3.6	1.7 (@28W)	1.2	2.2		128	2				否	否	
英特尔® 酷睿™ Ultra 5 处理器 125H	14	4	8	2	18	18 MB	4.5	3.6	1.2 (@28W)	0.7	2.2		112	2				否	否	

英特尔® 酷睿™ Ultra 处理器（U 系列，15 W）

处理器名称	处理器内核数	P-core 数	E-core 数	LPE 内核数	线程数	英特尔® 智能高速缓存（L3）	最大睿频频率（GHz）[A]		处理器基础频率（GHz）		最大显卡频率（GHz）	处理器显卡	执行单元（EU）数	视频解码器	PCIe 通道总数	最大内存速度	最大内存容量	TCC/TSN	宽温支持	处理器基础功耗（W）
							P-core	E-core	P-core	E-core										
英特尔® 酷睿™ Ultra 7 处理器 165U	12	2	8	2	14	12 MB	4.9	3.8	1.7（@15W）	1.2	2	英特尔® 显卡	64	2	20 PCIe 4.0	DDR5-5600 LPDDR5-6400 LPDDR5x-6400 LPDDR5x-7467	64 GB	否	否	28 W（最大保证功耗） 15 W（基础功耗） 12 W（最小保证功耗）
英特尔® 酷睿™ Ultra 7 处理器 155U	12	2	8	2	14	12 MB	4.9	3.8	1.7（@15W）	1.2	2		64	2				否	否	
英特尔® 酷睿™ Ultra 5 处理器 135U	12	2	8	2	14	12 MB	4.4	3.6	1.6（@15W）	1.1	1.9		64	2				否	否	
英特尔® 酷睿™ Ultra 5 处理器 125U	12	2	8	2	14	12 MB	4.3	3.6	1.3（@15W）	0.8	1.85		64	2				否	否	

A. 内核频率和内核类型因工作负载、功耗和其他因素而异。更多信息请见 https://www.intel.cn/content/www/cn/zh/architecture-and-technology/turbo-boost/intel-turbo-boost-technology.html

B. 英特尔锐炫™ GPU 仅适用于部分搭载 H 系列英特尔® 酷睿™ Ultra 处理器的系统，且系统内存为至少 16 GB 的双通道配置。需要 OEM 支持；请咨询 OEM 以了解系统配置详细信息。

产品规格请参阅 https://ark.intel.com/content/www/cn/zh/ark.html

了解更多有关英特尔® 酷睿™ Ultra 处理器的信息，请访问：
https://www.intel.cn/content/www/cn/zh/products/details/embedded-processors/core-ultra.html

1. 实际性能受使用情况、配置和其他因素的差异影响。更多信息请见 intel.com/processorclaims（英特尔® 酷睿™ Ultra 处理器 — 边缘）。结果可能不同。
2. 英特尔® 不以路线图指导的方式承诺或保证产品可用性或软件支持。英特尔® 保留通过标准 EOL/PDN 流程更改路线图，或是中止产品、软件和软件支持服务的权利。有关更多信息，请联系您的英特尔® 客户代表。
3. 英特尔锐炫™ GPU 仅适用于部分搭载 H 系列英特尔® 酷睿™ Ultra 处理器的系统，且系统内存为至少 16 GB 的双通道配置。需要 OEM 支持；请咨询 OEM 以了解系统配置详细信息。
4. 发布时提供的英特尔® AI Boost 支持有限。
5. Windows 11 IoT Enterprise LTSC 和 Linux 6.x 将支持英特尔® 硬件线程调度器。
6. 请联系您的英特尔® 代表获取雷电技术 4 支持许可。
7. 1x8 PCIe Gen 5 仅适用于 Meteor Lake-H 平台。

* 文中涉及的其他名称及商标属于各自所有者的资产。

2.1.3 英特尔® 至强® Max 系列 & 英特尔® 至强® 6 处理器

英特尔® 至强® Max 系列处理器

作为唯一一款基于 x86 的高带宽内存（HBM）处理器，英特尔® 至强® Max 系列处理器可最大程度提高带宽。英特尔® Max 系列 CPU 在架构设计上大幅增强采用 HBM 的英特尔® 至强® 平台的性能，相较于竞品，其针对实际工作负载的性能提升了 4.8 倍[1]，比如建模、人工智能、深度学习、高性能计算（HPC）和数据分析。

最大限度提高带宽

英特尔® 至强® Max 系列处理器旨在加速需求最严苛的工作负载，实现了：

提升高达

5 倍

与竞品和前几代产品相比，内存带宽性能显著提升。[1,2]

提升高达

20 倍

采用 HBM 时，用于 NLP 的 Numenta AI 技术与其他 CPU 相比实现的速度提升。[3]

提升高达

8.6 倍

洛斯阿拉莫斯国家实验室在不更改当前 HPC 系统代码的情况下获得的性能增益。[4]

通过提高带宽来最大化性能

英特尔® 至强® Max 系列处理器采用了新的微架构，并支持丰富的平台增强功能，包括更多的核心数量、先进的 I/O 和内存子系统，以及内置加速器。英特尔® 至强® Max 系列处理器特点包括：

多达 56 个性能核	英特尔® 至强® Max 系列处理器利用由四个小芯片构成的多达 56 个性能核的强大功能，通过英特尔® 的嵌入式多芯片互联桥接（EMIB）技术在 350 瓦封装中互相连接。
64 GB 高带宽内存	英特尔® 至强® Max 系列处理器通过 4 个 HBM2e 堆栈、64 GB 超高带宽封装内存和每核超过 1 GB 的 HBM 容量，最大限度提升性能。
节省成本和时间	对于适合 64 GB 容量且每核心需要 1-2 GB 内存的工作负载的应用，使用英特尔® 至强® Max 系列处理器无需 DDR 和耗时的代码更改，实现了成本节省。

适用于不同 AI 工作负载的灵活性

英特尔® 至强® Max 系列处理器提供了不同的内存模式，可根据工作负载的需求灵活配置：

仅 HBM 模式	支持在 64 GB 容量内的工作负载，并且能够按每核心 1-2 GB 的内存进行扩展，在系统启动时，仅 HBM 模式不需要任何代码更改且不需要 DDR。
HBM 平坦模式	为需要大内存容量的应用程序提供灵活性，HBM 平坦模式提供了一个包含 HBM 和 DRAM 的平坦内存区域，可用于每核心需要 > 2 GB 内存的工作负载。该模式可能需要进行代码更改。
HBM 缓存模式	旨在提高 > 64 GB 容量或每核心需要 > 2 GB 内存的工作负载的性能。该模式无需代码更改，HBM 会缓存 DDR。

跨不同架构加速 AI 应用程序

整个英特尔® 至强® Max 系列产品通过英特尔® oneAPI 统一，为一个共通的、开放的、基于标准的编程模型，释放生产力和性能。开发者可以使用英特尔® oneAPI 基础工具包和英特尔® oneAPI 高性能计算工具包，更容易地构建、分析、优化和扩展通用计算、高性能计算和 AI 应用程序，跨越多种类型的架构，并使用包括在向量化、多线程、多节点并行化和内存优化方面的最先进技术。使用英特尔® 至强® Max 系列处理器和英特尔® oneAPI，开发者可以轻松构建高性能、多架构软件，为高性能计算做好准备。

英特尔® 至强® Max 系列处理器产品

处理器名称	处理器内核数	基频频率 (GHz)	全部核心睿频频率 (GHz)	最高睿频频率 (GHz)	高速缓存 (MB)	TDP (W)	可扩展性	DDR5 内存速度	英特尔® SGX 的最大 Enclave Page Cache (每个核心)
英特尔® 至强® CPU Max 9480 处理器	56	1.9	2.6	3.5	112.5	350	2S	4800	512 GB
英特尔® 至强® CPU Max 9470 处理器	52	2	2.7	3.5	105	350	2S	4800	512 GB
英特尔® 至强® CPU Max 9468 处理器	48	2.1	2.6	3.5	105	350	2S	4800	512 GB
英特尔® 至强® CPU Max 9460 处理器	40	2.2	2.7	3.5	97.5	350	2S	4800	128 GB
英特尔® 至强® CPU Max 9462 处理器	32	2.7	3.1	3.5	75	350	2S	4800	128 GB

了解更多有关英特尔® 至强® Max 系列处理器的信息，请访问：
https://www.intel.cn/content/www/cn/zh/products/details/processors/Xeon/max-series.html

1. 有关工作负载和配置的信息，请访问：intel.com/performanceindex（活动：Supercomputing 22）。结果可能有所差异。
2. 2S 英特尔® 至强 Max CPU 对比 2S AMD EPYC 7773X 和 2S 第三代英特尔® 至强® 8380。
3. Numenta BERT-Large AMD Milan：由 Numenta 测试，截至 2022 年 11 月 28 日。1 个节点，AWS m6a.48xlarge 上的 2 个 AMD EPYC 7R13，768 GB DDR4-3200，Ubuntu 20.04 内核 5.15，OpenVINO™ 2022.3，BERT-Large，序列长度 512，批大小为 1。英特尔® 至强® 8480+：由 Numenta 测试，截至 2022 年 11 月 28 日。1 个节点，2 个英特尔® 至强® 8480+，512 GB DDR5-4800，Ubuntu 22.04 内核 5.17，OpenVINO™ 2022.3，Numenta 优化的 BERT-Large，序列长度 512，批大小为 1。英特尔® 至强® Max 9468：由 Numenta 测试，截至 2022 年 11 月 30 日。1 个节点，2 个英特尔® 至强® Max 9468，128 GB HBM2e 3200 MT/s，Ubuntu 22.04 内核 5.15，OpenVINO™ 2022.3，Numenta 优化的 BERT-Large，序列长度 512，批大小为 1。
4. Crossroads 上的预生产芯片。有关工作负载和配置的信息，请查看 https://arxiv.org/abs/2211.05712。结果可能有所差异。
5. 性能因用途、配置和其他因素而异。性能结果基于截至配置中所示日期的测试，可能无法反映所有公开发布的更新。没有任何产品或组件能够做到绝对安全。请访问：www.Intel.cn/PerformanceIndex 了解详情。

英特尔® 至强® 6 性能核处理器

经过精心优化，英特尔® 至强® 6 性能核处理器的单核性能非常出色，拥有比其他通用 CPU 更好的 AI 性能，能够从容应对广泛的工作负载。与第五代英特尔® 至强® 可扩展处理器（常用于较新的计算密集型解决方案）相比，英特尔® 至强® 6 性能核处理器的性能提升高达 2 倍。

英特尔® 至强® 6 性能核处理器

通用计算
高达
2 倍
整数和浮点吞吐量提升[1]

AI
高达
2 倍
GenAI 性能提升（采用 BF16 数据类型）[2]

科学计算
高达
2.3 倍
科学计算性能提升（基于行业标准 HPCG 基准测试）[3]

与第五代英特尔® 至强® 可扩展处理器比较

为广泛的工作负载实现高性能

采用性能核的英特尔® 至强® 6 处理器，每个插槽可灵活扩展至 128 个内核、12 个内存通道和 96 个 PCIe 通道，帮助企业满足不同的应用需求。对于希望缓解内存带宽瓶颈的 IT 团队来说，创新的多路合并阵列双列直插内存模组（MCR DIMM）可提供高达每秒 8,800 兆次（MT/s）的传输速度，同时通过快速完成工作来降低总体拥有成本。内置加速器为目标工作负载提供额外的提升，实现更高的性能和效率。

利用 CPU 的强大 AI 性能

英特尔® 至强® 6 性能核处理器旨在支持许多要求严苛的 AI 用例。P-core（性能核）通过英特尔® Advanced Matrix Extensions（英特尔® AMX）等加速功能，INT8、BF16 和 FP16（新）等数据类型。因此，性能核可帮助满足从目标检测到中型 GenAI 等多种人工智能模型的服务级别协议（SLA），同时提供开放标准、高性能、RAS 功能，并根据需要支持其他加速器。由于配备了增强的内核、更大的内存带宽和强大的矩阵引擎，采用性能核的英特尔® 至强® 6 处理器可提供充足的算力，以支持中小规模生成式人工智能模型的推理、微调和检索增强生成（RAG）用例。此外，针对英特尔® 至强® 处理器的优化已集成到 TensorFlow* 和 PyTorch* 等在内的流行深度学习框架的主流发行版。

优化通用工作负载的性能

采用性能核的英特尔® 至强® 6 处理器在全范围工作负载上表现出色，其主流系列产品拥有 8-86 个内核，在基于双 CPU 的系统中，网络和存储外接卡拥有多达 176 个 PCIe 5.0 通道，而基于单 CPU 的系统中，单插槽产品则拥有 136 个 PCIe 通道。所有英特尔® 至强® 6 处理器都能随着服务器利用率的增加而提供可扩展的每瓦性能，在整个负载线路上提供近乎线性的功耗-性能消耗，这凸显了所有英特尔® 至强® 6 处理器的高效性。对于性能要求苛刻的工作负载，这意味着平台在高负载下有效地利用能耗，以帮助快速完成工作。

利用增强的安全功能跟上业务增长的步伐

在本地、边缘和云服务器上追求新的业务模式和数据共享，即使在处理敏感数据或受监管数据时也是如此。基于可信执行环境（TEE）的机密计算能够帮助在使用过程中保护数据和 AI 模型。采用性能核的英特尔® 至强® 6 处理器允许客户选择最符合其业务和监管要求的机密计算技术。

应用程序隔离	英特尔® 软件防护扩展（英特尔® SGX）提供旨在保护使用中数据的应用程序隔离。英特尔® SGX 是目前市场上经过深入研究和多次更新的数据中心级机密计算技术。
虚拟机（VM）级隔离	英特尔® 信任域扩展（英特尔® TDX）在虚拟机级别提供隔离和机密性。在基于英特尔® TDX 的机密虚拟机中，客户机操作系统和虚拟机应用程序被隔离开来，无法被云端主机、虚拟机管理程序和平台的其他虚拟机访问。

亮点技术

AI 计算能力

- 单路英特尔® 至强® 6 性能核处理器拥有多达 128 个内核，实现了更高密度计算性能和可扩展性。
- 对于基于 BF16 和 FP16 的模型，英特尔® AMX 的乘法累加（MAC）运算速度比英特尔® 高级矢量扩展 512（Intel® Advanced Vector Extensions 512，英特尔® AVX-512）提升高达 16 倍，AI 性能显著增强。
- 英特尔® AVX-512 包含特有的指令，每个内核拥有两个 512 位融合乘加（FMA）单元，大幅提高了 AI、科学计算和数据库工作负载常见的矢量计算速度。
- 支持 VNNI 指令的英特尔® AVX2 以及将精度快速转换为 BF16和 FP16 的能力为英特尔® 至强® 6 能效核处理器提供了更好的 AI 兼容性。

内存

- 与标准 DDR5 DIMM 相比，MCR DIMM 能够提供超过 37% 的额外内存带宽，可支持 AI 和科学计算中的带宽受限用例。
- 多达 12 条内存通道，进一步支持更高的内存带宽。
- 当使用低成本内存（如支持 CXL 2.0 的 DDR4）时，“Flat” 内存模式可帮助扩展系统内存并优化 TCO。

连接与 I/O

- 英特尔® 超级通道互联（Intel® Ultra Path Interconnect，英特尔® UPI）2.0 的跨插槽内带宽速度高达 24 GT/s，与上一代产品相比提升高达 20%。
- 双路服务器拥有多达 178 条 PCIe Gen 5 通道，单路服务器则多达 136 条，可以支持重要的 I/O 附加组件，包括加速器、网络适配器、存储控制器和存储。
- 多达 64 条 Compute Express Link（CXL）2.0 通道，每条通道的数据传输速率高达 32 GT/s，支持 CXL 功能，包括内存扩展和共享（包括 Type 3 设备）。

数据

- 英特尔® 数据保护与压缩加速技术（Intel® QuickAssist Technology，英特尔® QAT）支持卸载批量加密和压缩，以加速网络和存储。
- 英特尔® 数据流加速器（Intel® Data Streaming Accelerator，英特尔® DSA）2.0 能够卸载数据传输和转换操作，例如移动、填充、比较、循环冗余校验（CRC）、数据完整性字段（DIF）、增量和刷新。
- 英特尔® 存内分析加速器（Intel® In-Memory Analytics Accelerator，英特尔® IAA）可以卸载内存压缩和解压缩、扫描和过滤功能以及循环冗余校验。
- 英特尔® 动态负载均衡器（Intel® Dynamic Load Balancer，英特尔® DLB）支持动态分配网络数据包处理和卸载重排序操作。
- 英特尔® TDX 用 AES-256 和 2,048 个加密密钥进行了升级，机密计算能力进一步增强，能够更好地保护敏感的企业数据。
- 英特尔® On Demand 服务使硬件提供商可以启用部分基于 CPU 的特性和功能。它通过两种模式提供服务：基于一次性许可证激活功能，以及基于用量付费。

英特尔® 至强® 6 性能核处理器概览

英特尔® 至强® 6 性能核处理器		
英特尔® 至强® 6900 系列处理器	**旗舰级** 采用全新的英特尔® 服务器平台设计，非常适合云计算、AI、科学计算、软件即服务 (SaaS) 和基础设施即服务 (IaaS) 等工作负载。	• 每个 CPU 拥有多达 128 个内核（256 个线程） • 每个 CPU 高达 500W • 单路或双路服务器 • 12 条内存通道 • 高达 6,400 MT/s DDR5 • 8,800 MT/s MCR DIMM • 多达 96 条 PCIe 5.0 通道 • 6 条英特尔® UPI 2.0 链路 **即将推出**
英特尔® 至强® 6500 系列/ 6700 系列处理器	**高端级** 对现有的英特尔® 服务器平台进行了大幅升级。面向企业 IT、数字服务提供商和电信的主流边缘协同服务器。非常适合 AI、科学计算、网络和媒体、数字服务、基础设施和存储、Web、应用以及微服务等工作负载。	• 每个 CPU 拥有多达 86 个内核（172 个线程） • 每个 CPU 高达 350 W • 单路、双路、四路或八路服务器 • 8 条内存通道 • 高达 6,400 MT/s DDR5 • 高达 8,000 MT/s MCR DIMM • 多达 88 条 PCIe 5.0 通道，其中单路设计最高可达 136 条 • 4 条英特尔® UPI 2.0 链路 **即将推出**

了解更多有关英特尔® 至强® 6 性能核处理器的信息，请访问：
https://www.intel.cn/content/www/cn/zh/products/details/processors/Xeon/Xeon6-p-cores.html

1. 详情请见以下网址的：intel.com/processorclaims（英特尔® 至强® 6 处理器）。结果可能不同
2. 详情请见以下网址的：intel.com/processorclaims（英特尔® 至强® 6 处理器）。结果可能不同。
3. 详情请见以下网址的：intel.com/processorclaims（英特尔® 至强® 6 处理器）。结果可能不同。

2.1.4 英特尔锐炫™ 显卡

英特尔锐炫™ 显卡是独立显卡产品线，为网络和物联网边缘应用大幅提升 AI、图形和媒体处理性能。锐炫™ 系列显卡提供了不同系列产品，可以高中低级别的 AI 工作负载。其中，高端显卡锐炫™ 7 系列，可拥有最高 28 颗 X^e 核心和 16 GB GDDR6 显存，为重型 AI 工作负载和广泛的用例，提供高性能的支持；中端显卡锐炫™ 5 系列，可拥有最高 24 颗 X^e 核心和 16 GB GDDR6 显存，完美满足边缘对 AI 推理能力、性能以及性价比的需求；低端显卡锐炫™ 3 系列，可拥有最高 8 颗 X^e 核心和 6 GB GDDR6 显存，满足了边缘应用对于低功耗和小尺寸形态的要求，满足 AI 推理能力的需求。

英特尔® X^e -HPG 微架构驱动边缘 AI 工作负载

英特尔锐炫™ 显卡采用了英特尔® X^e -HPG 微架构，凭借其全新的 X^e 内核，满足边缘 AI 工作负载对计算效率与性能的要求。X^e 内核是英特尔® GPU 产品中新的基础计算异构模块，针对特定的工作负载进行优化。每个 X^e 内核配备 AI 引擎，利用英特尔® X^e 矩阵扩展（XMX）技术，加速 AI 工作负载。与传统的 GPU 矢量单元相比，XMX AI 引擎完成 AI 推理操作的计算能力是其 16 倍，可为大幅提升边缘 AI 应用的生产力。

开放和基于标准的 GPU 编程工具 OpenVINO™

英特尔® 提供了开源的 OpenVINO™ 工具包，为 AI 工作负载提供了在英特尔锐炫™ GPU 上最大的加速和优化。同时，OpenVINO™ 可简化和优化跨不同平台运行的 AI 推理代码开发。一次编码，即可在 GPU、CPU 和其他硬件加速器上运行代码，可以使开发者摆脱昂贵、不灵活的 GPU 编程接口和工具的束缚，消除供应商锁定。

使用英特尔为边缘设计的 GPU 构建 AI 应用

英特尔通过强大的 ODM 生态系统提供多种 GPU 卡片，用于解决边缘 AI 应用复杂、多样的需求。这些卡片具有多种形态、功率和性能水平。结合英特尔以及合作伙伴在特定垂直市场中的专业知识，基于锐炫™ 显卡可以打造强大的边缘 AI 应用。

主要特性

X^e -HPG 微架构

- 每颗英特尔® X^e 核心配备了 16 个矩阵引擎和 16 个矢量引擎（最高 28 核搭载于 A750E）
- GDDR6 显存（最高 16 GB，搭载于 A580E 和 A750E）
- 最多 448 个执行单元
- 体引擎支持 HDR、VP9、H.264/AVC、H.265/HEVC 和 AV1
- HDMI 2.0b 和 DisplayPort 1.4a、2.0 接口，支持最多 2x 8k60 HDR、4x 4k120 HDR、1080p360 或 1440p360 分辨率

AI 推理

- 搭载了 XMX AI 引擎，加速 AI 工作负载
- 在边缘使用 OpenVINO™ 工具包支持的 AI 加速
  - 提高 AI 推理的性能和效率
  - 兼容多种硬件架构以及最流行的 AI 框架，包括 TensorFlow* 和 PyTorch*
  - 使用先进的 LLM 权重压缩技术，降低内存消耗，简化生成式 AI 应用

专为边缘设计

- 英特尔 ODM 合作伙伴生态系统提供的广泛产品，涵盖不同的形态和性能要求
- 满足边缘的低功耗（25-75W）和小尺寸形态要求的 SKU
- 五年产品供应和软件支持
- 多种操作系统支持，包括 Linux、Windows client、Windows 10 LTSC
- 转为嵌入式使用设计的 SKU 满足更高的可靠性要求
- 定期、稳定的驱动程序发布
- 支持 PCIe 4.0，适用于 PCIe 和 MxM
- 提供带有现代软件用户界面的控制软件
- 高性能选项（功率 195W – 225W）可提供最高 236 INT8 TOPS 算力（A750E）

视觉计算和图形处理

- 支持高达 8K 分辨率，4 个显示信道
- 支持 DirectX、OpenGL 和 Vulkan
- 支持视频墙和数字标牌软件功能；用于增强扩展显示的 Genlock/pipelock，EDID 管理，边框补偿及多 GPU 单一大屏幕能力

媒体处理

- 英特尔® VPL 提供对专用媒体硬件的高级访问权限，以及英特尔® GPU 上的编码、解码和视频处理功能
- 媒体引擎支持大多数现代编解码器，高带宽 8K HDR 工作流程执行和多 GPU 同步
- 支持下一代 AV1 媒体标准的硬件加速编码

英特尔锐炫™ 显卡示意图

英特尔锐炫™ 显卡产品线

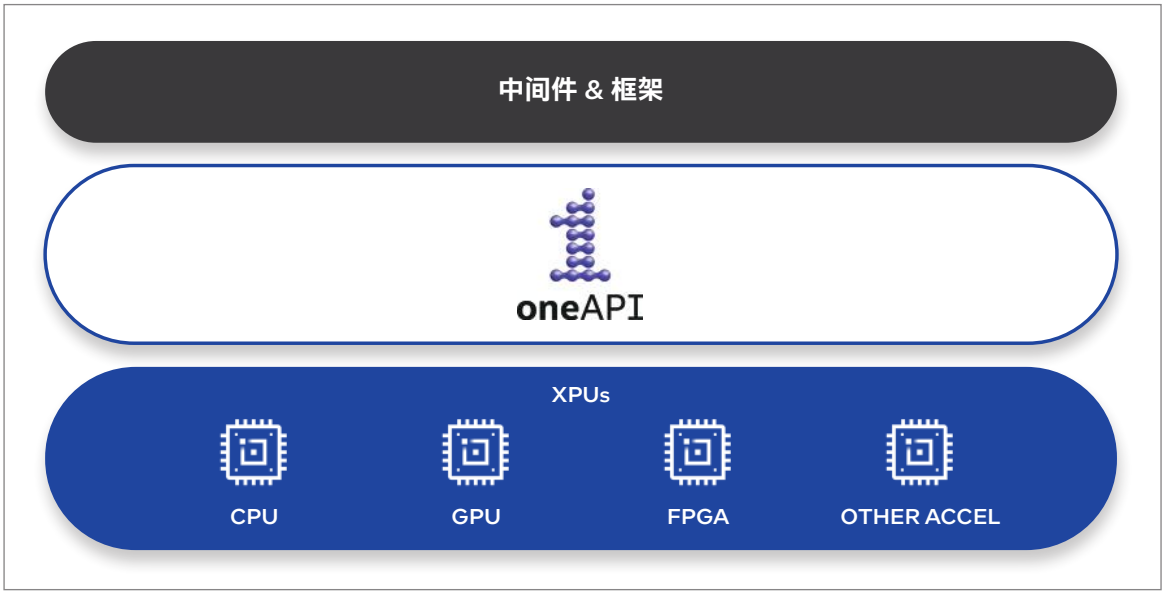
型号	微架构	Xe-core	Xe 矢量引擎	显卡时钟频率	GPU 峰值 TOPS (INT8)	专用高带宽内存	内存类型	显卡内存接口	显卡内存带宽	显卡内存速度	多种格式编解码器引擎数	垂直市场	发行时间
英特尔锐炫™ A310 显卡	Xe HPG	6	96	2000 MHz	52	4 GB	GDDR6	64 bit	124 GB/s	15.5 Gbps	2	桌面端	Q3'22
英特尔锐炫™ A380 显卡		8	128	2000 MHz	66	6 GB	GDDR6	96 bit	186 GB/s	15.5 Gbps	2		Q2'22
英特尔锐炫™ A580 显卡		24	384	1700 MHz	197	8 GB	GDDR6	256 bit	512 GB/s	16 Gbps	2		Q4'23
英特尔锐炫™ A750 显卡		28	448	2050 MHz	229	8 GB	GDDR6	256 bit	512 GB/s	16 Gbps	2		Q3'22
英特尔锐炫™ A770 显卡 (8 GB)		32	512	2100 MHz	262	8 GB	GDDR6	256 bit	512 GB/s	16 Gbps	2		Q3'22
英特尔锐炫™ A770 显卡 (16 GB)		32	512	2100 MHz	262	16 GB	GDDR6	256 bit	560 GB/s	17.5 Gbps	2		Q3'22
英特尔锐炫™ A310E 显卡		6	96	2000 MHz	49	4 GB	GDDR6	64 bit	124 GB/s	15.5 Gbps	2	嵌入式	Q2'24
英特尔锐炫™ A350E 显卡		6	96	1150 MHz	28	4 GB	GDDR6	64 bit	112 GB/s	14 Gbps	2		Q2'24
英特尔锐炫™ A370E 显卡		8	128	1550 MHz	38	4 GB	GDDR6	64 bit	112 GB/s	14 Gbps	2		Q2'24
英特尔锐炫™ A380E 显卡		8	128	2000 MHz	66	6 GB	GDDR6	96 bit	186 GB/s	15.5 Gbps	2		Q2'24

了解更多有关英特尔锐炫™ 显卡的信息，请访问：
https://www.intel.cn/content/www/cn/zh/architecture-and-technology/visual-technology/arc-discrete-graphics.html

2.2 软件

2.2.1 英特尔® oneAPI 工具包 — 跨架构性能加速

oneAPI 是一种跨行业、开放、基于标准的统一编程模型。它定义了一个通用、统一和开放的多架构和多供应商软件平台，确保在不同硬件供应商和加速器技术之间的功能代码可移植性和性能可移植性。oneAPI 的核心语言是 SYCL，它可以被用于编程加速器和多种处理器。SYCL 允许开发者在不同的硬件平台上（CPU、GPU、FPGA 和其他加速器）之间重用代码，并为特定架构进行优化。基于 SYCL，oneAPI 定义了一套广泛的规范和库 API，以满足跨行业和计算以及 AI 使用案例的编程领域需求。此外，oneAPI 提供一个开发者社区和开放论坛，以推动统一的 API，为统一的行业宽多架构软件开发平台，并鼓励生态系统合作。



图：oneAPI 图示（来源：https://www.oneapi.io/）

作为 oneAPI 指导委员会的重要成员，英特尔® 根据 oneAPI 规范推出了英特尔® oneAPI 工具包，旨在帮助开发者使用英特尔® 优化的一流的编译器、性能库、框架以及分析和调试工具，构建、分析并优化在 CPU 和 XPU 上的高性能、跨架构应用程序。在英特尔® oneAPI 工具包的加持下，开发者可以自由选择架构以解决他们所面临的问题，无需为了新的架构和平台而重写软件。针对不同领域的开发者，英特尔® oneAPI 工具包提供了不同的工具包来满足他们不同的需求。

英特尔® oneAPI 基础工具包

针对一般开发者，英特尔® oneAPI 基础工具包是一套核心工具和库，用于跨不同架构开发高性能、以数据为中心的应用程序。它包含一个行业领先的 SYCL 编译器，以及为特定领域优化的库和英特尔® Python 发行版，提供了针对不同架构的即插即用加速。此外，这套工具包还包含增强的性能分析、设计辅助和调试工具，帮助开发者更好地开发应用程序。英特尔® oneAPI 基础工具包提供了以下工具：

- 英特尔® oneAPI DPC++/C++ 编译器
- 英特尔® DPC++ 兼容性工具
- 英特尔® oneAPI DPC++ 库
- 英特尔® oneAPI 数学核心库（oneMKL）
- 英特尔® oneAPI 多线程构件（oneTBB）
- 英特尔® oneAPI 集合通信库（oneCCL）
- 英特尔® oneAPI 数据分析库（oneDAL）
- 英特尔® oneAPI 深度神经网络库（oneDNN）
- 英特尔® 集成性能原语（Intel® IPP）
- 英特尔® VTune™ 性能分析器
- 英特尔® Advisor
- 英特尔® GDB 发行版
- 英特尔® Python* 发行版
- 用于英特尔® oneAPI DPC++ / C++ 编译器的 FPGA 支持包

英特尔® 高性能计算工具包

针对高性能计算应用程序开发者，英特尔® 高性能计算工具包提供了所需的优化、分析和扩展应用程序所需的技术，包括向量化、多线程、多节点并行化和内存优化等。这个工具包是对英特尔® oneAPI 基础工具包的补充，包括以下工具：

- 英特尔® Fortran 编译器
- 英特尔® Fortran 编译器经典版
- 英特尔® MPI 库

英特尔® 渲染工具包

英特尔® 渲染工具包是一套强大的开源渲染、光线追踪、去噪和路径引导库，用于 AI 合成数据生成、数字孪生、高保真和高性能可视化，以及沉浸式内容创作。利用这些库和英特尔® CPU 与 GPU 硬件，实现优化的渲染性能，构成一个可扩展的解决方案。英特尔® 渲染工具包提供了以下工具：

- 英特尔® Embree
- 英特尔® 开放体积核心库（Intel® Open VKL）
- 英特尔® 开放图像去噪（Intel® Open Image Denoise）
- 英特尔® OSPRay
- 英特尔® OSPRay Studio
- 英特尔® 开放路径引导库（Intel® Open PGL）
- 渲染工具包实用程序

AI 工具包

英特尔® AI 工具包（原名英特尔® AI 分析工具包）针对数据科学家、AI 开发者和研究者提供了他们所熟悉的 Python 工具和框架，以加速在英特尔® 架构上端到端的数据科学和分析流程。该工具包中的组件的底层计算优化是由 oneAPI 库构建的。英特尔® AI 工具包从预处理到机器学习，提供最大化性能优化，并提供高效模型开发的互操作性，包含以下工具：

- 包含高度优化的 scikit-learn 的英特尔® Python* 发行版
- 英特尔® Pytorch* 扩展
- 英特尔® TensorFlow* 扩展
- 英特尔® XGBoost 优化
- 英特尔® 神经网络压缩器
- 英特尔® AI 参考模型
- Modin（pandas 的即插即用替代品）

英特尔® OpenVINO™ 工具套件发行版

英特尔® OpenVINO™ 工具套件发行版一个开源工具包，它加速了 AI 推理，降低了延迟，提高了吞吐量，同时保持了准确性，减少了模型占用空间，并优化了硬件使用。该工具包用于简化了 AI 开发和深度学习在计算机视觉、大型语言模型（LLM）、生成性 AI 等领域的集成，包含以下工具：

- 模型优化器
- 深度学习工作台
- 推理引擎
- 部署管理引擎
- OpenCV DL 流媒体处理器
- 训练后优化工具

英特尔® 系统启动工具包

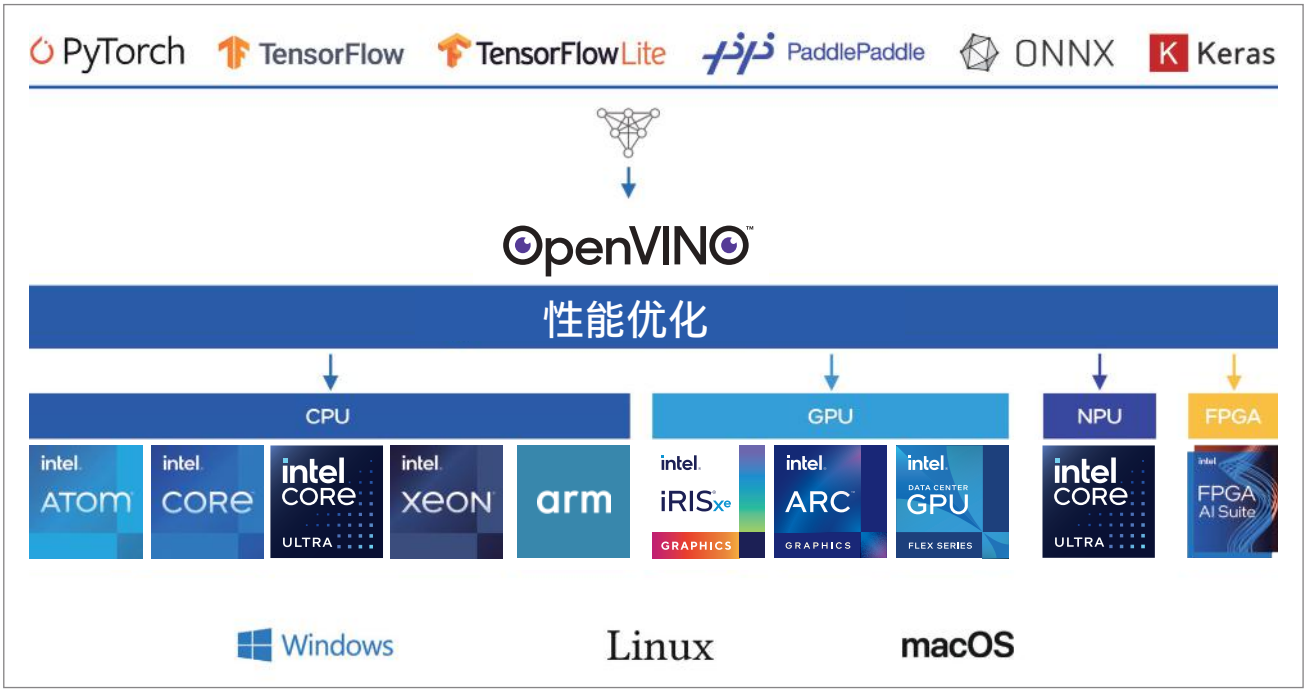
英特尔® 系统启动工具包旨在帮助系统制造商和开发者在硬件和软件层面上启动和优化新系统。这套包括调试、跟踪以及功率和性能分析工具的集合，允许开发者快速调试和分析整个平台中的硬件、固件、UEFI、BIOS、操作系统内核、设备驱动程序等。英特尔® 系统启动工具包包含以下工具：

- 英特尔® SoC Watch
- 英特尔® 系统 Debugger
- 英特尔® VTune™ 性能分析器

了解更多有关英特尔® oneAPI 工具包的信息，请访问：
https://www.intel.com/content/www/us/en/developer/tools/oneapi/overview.html

2.2.2 OpenVINO™ 工具套件

OpenVINO™ 是一个开源工具套件，主要用于优化和部署从云到边缘的深度学习模型。它加速了在各种应用场景中使用到的深度学习推理，例如大语言模型、生成式 AI、视频、音频和语言处理等等。OpenVINO™ 支持使用 PyTorch*、TensorFlow*、ONNX 等流行框架搭建的模型模型，开发者可以使用他们喜欢的框架搭建深度学习模型。同时，OpenVINO™ 提供了一系列模型优化和转换工具，帮助开发者优化他们搭建的模型的性能，例如提高推理速度、降低运行内存等等。最后，OpenVINO™ 支持将模型部署在各种各样的环境上，无论是云端、浏览器还是本地设备、英特尔® 或是第三方硬件、CPU、GPU、NPU 或 FPGA。



图：OpenVINO™ 图示（来源：https://docs.openvino.ai/2024/index.html）

模型搭建

OpenVINO™ 支持多种第三方模型格式，包括 PyTorch*、TensorFlow*、TensorFlow Lite*、ONNX 和 PaddlePaddle*，通过这种方式允许开发者使用他们熟悉的框架搭建 AI 模型。同时，OpenVINO™ 还支持开发者使用类似 Kaggle*、Hugging Face* 或 Torchvision models* 等模型库中的预训练的模型。当开发者搭建好模型，他们可以选择直接用原始格式在 OpenVINO™ 上运行这些模型，这种情况下，OpenVINO™ 会自动对这些模型进行转换并优化。如果开发者需要更好的性能和高级优化，更推荐的方法是将他们搭建好的模型转换成 OpenVINO™ IR 格式。OpenVINO™ 提供了模型转换工具，将前面提到支持的格式的模型，通过 Python* API 或命令行工具转换成 OpenVINO™ IR 格式。

模型优化

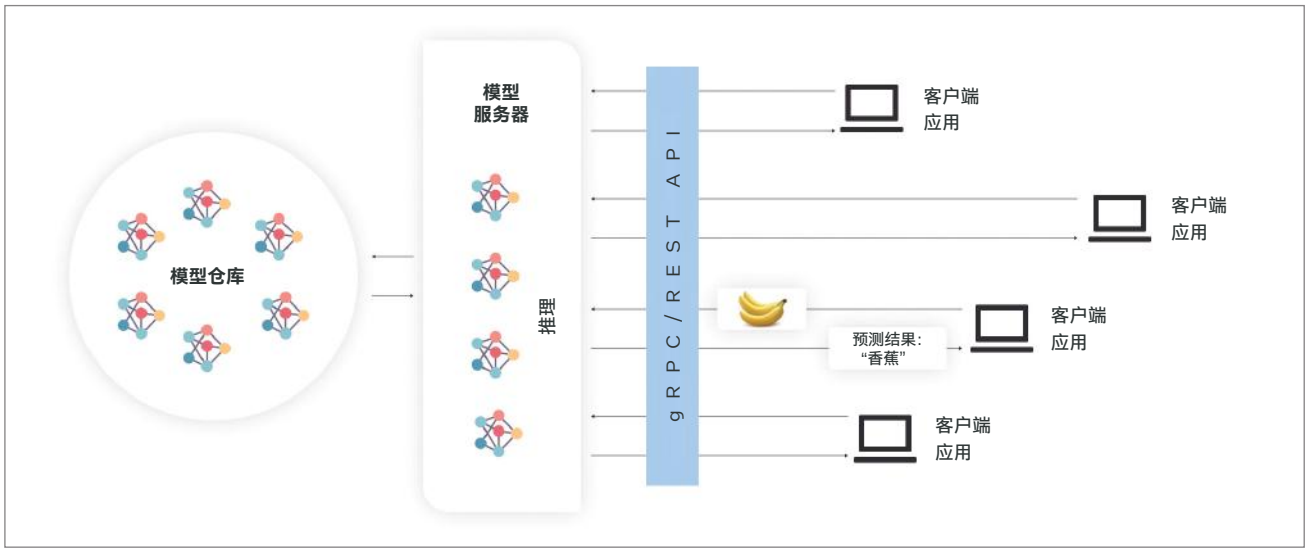
OpenVINO™ 支持多种模型优化方法来提高模型的大小和在推理时的性能，在 OpenVINO™ 神经网络压缩框架（NNCF）中，提供了三种优化路径：

- 训练后量化，通过应用训练后的 8 位整数量化来优化深度学习模型的推理，这优化方法不需要模型重新训练或微调。
- 训练时优化，这是一套在例如 PyTorch* 和 TensorFlow* 2.x 这样的深度学习框架内进行训练时模型优化的高级方法，支持诸如量化感知训练、结构化和非结构化剪枝等方法。
- 权重压缩，这是是一种用于减少 AI 大模型大小并加速推理的方法。

模型部署

使用 OpenVINO™ 运行模型以来 OpenVINO™ 运行时，一组带有 C 和 Python 绑定的 C++ 库，提供了一个通用的 API，在开发者选择的平台上部署和运行推理，无论是 CPU、GPU、NPU 还是 FPGA。OpenVINO™ 运行时时通过插件架构实现跨平台能力的，其插件包含了在各种特定硬件平台上运行模型所需的软件组件。每个插件都实现了统一的 API，并提供了用于配置设备的额外硬件特定 API。

通过 OpenVINO™，开发者可以将模型部署在本地，同时 OpenVINO™ 还提供了模型服务器。OpenVINO™ 模型服务器用于托管模型，并通过标准网络协议使它们能够被客户端软件访问：客户端向模型服务器发送请求，模型服务器执行模型推理并将响应发送回客户端。模型服务器具有很多优势。轻量级边缘 AI 应用只需要具备执行 API 调用的必要功能，通过网络调用远程推理；而模型服务器端可以基于微服务的应用程序和在云环境中部署的理想架构，并通过水平和垂直推理扩展实现高效的资源利用。此外，通过模型服务器的部署方式，模型的拓扑结构和权重不会直接暴露给客户端应用程序，这使得控制对模型的访问变得更加容易。



图：OpenVINO™ 模型服务器图示（来源：https://docs.openvino.ai/2024/ovms_what_is_openvino_model_server.html）

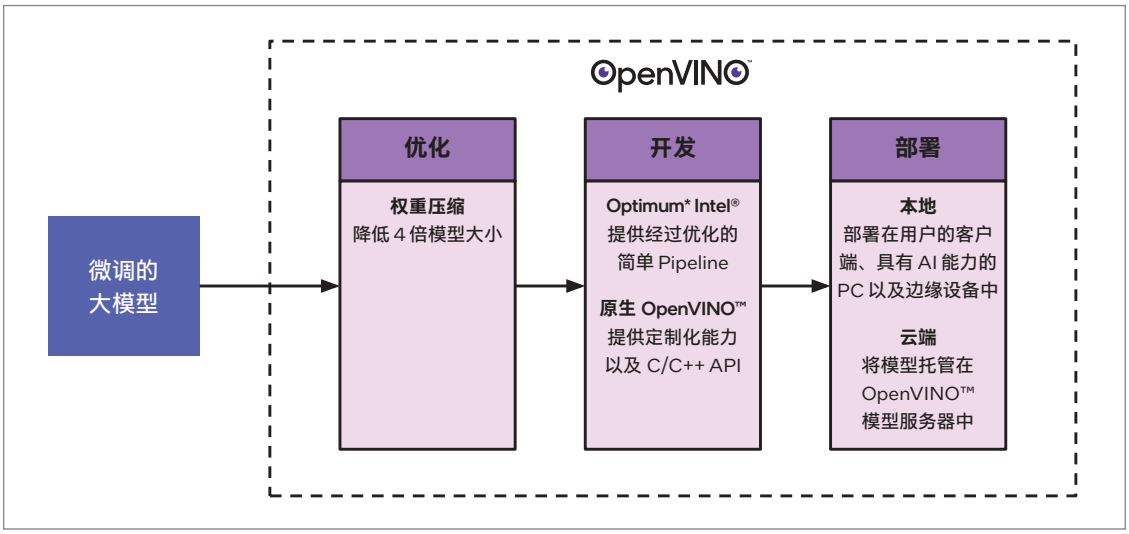
OpenVINO™ 与 AI 大模型

在最新版本的 2024.3 中，OpenVINO™ 增加了需要针对 AI 大模型的新功能，包括：

- 扩大了对生成式 AI 和大模型框架的覆盖和支持，并在 Hugging Face* 上提供 OpenVINO™ 预优化模型，帮助开发者更容易地开始使用这些模型。

- 支持更广泛的模型压缩技术，通过添加动态量化、多头注意力（MHA）和 OneDNN 增强，显著提高了在英特尔® 集成和独立显卡上的大模型性能。
- 提高大模型的可移植性和性能，并在 OpenVINO™ 模型服务器中支持了 vLLM 和连续批处理，帮助开发者更好地在边缘、云端或本地运行大模型推理。

基于这些新功能，OpenVINO™ 可以提供一套用于优化和部署 AI 大模型到最终用户的系统和设备中的领先的解决方案。开发者可以使用 OpenVINO™ 来压缩大模型，将它们集成到 AI 助手应用程序中，并以最大性能将它们部署到边缘设备或云端。



图：使用 OpenVINO™ 优化和部署 AI 大模型的流程示意图

OpenVINO™ 为部署大模型提供了一个灵活高效的运行时环境。其优势包括部署大小、速度、支持、灵活性，以及在多种硬件上运行的能力。

- 在模型简化方面，由于 OpenVINO™ 是一个独立的软件包，相比 Hugging Face*、PyTorch* 和其他机器学习框架相比，它需要的依赖更少。因此使用 OpenVINO™ 运行和部署大模型时，其更精简的二进制大小和内存占用减少了对硬件及存储的需求。同时，较少的依赖也意味着在部署环境中进行包和版本管理时的麻烦更少。
- 在运行速度方面，大多数大模型运行时库都依赖于通过 Python* 解释器执行的 Python* 代码，而 OpenVINO™ 是一个专门为资源优化的生产环境而设计的运行时库，并提供完整 C/C++ API。当然，OpenVINO™ 也提供了 Python* API，这允许开发者更快地开发算法和程序。开发者可以在 Python* 中原型化解决方案，然后使用 OpenVINO™ 用 C++ 进行优化。

了解更多有关 OpenVINO™ 工具套件的信息，请访问：
https://www.intel.com/content/www/us/en/developer/tools/openvino-toolkit/overview.html

2.2.3 英特尔® Geti™ 平台

英特尔® Geti™ 平台使企业团队能够快速构建计算机视觉 AI 模型。通过直观的图形界面，用户可以添加图像或视频数据、进行标注、训练、重新训练、导出以及优化 AI 模型以便部署。配备了最先进的技术，如主动学习、任务链和智能标注，英特尔® Geti™ 平台减少了劳动密集型任务，实现了协作模型开发，并加快了模型创建的速度。

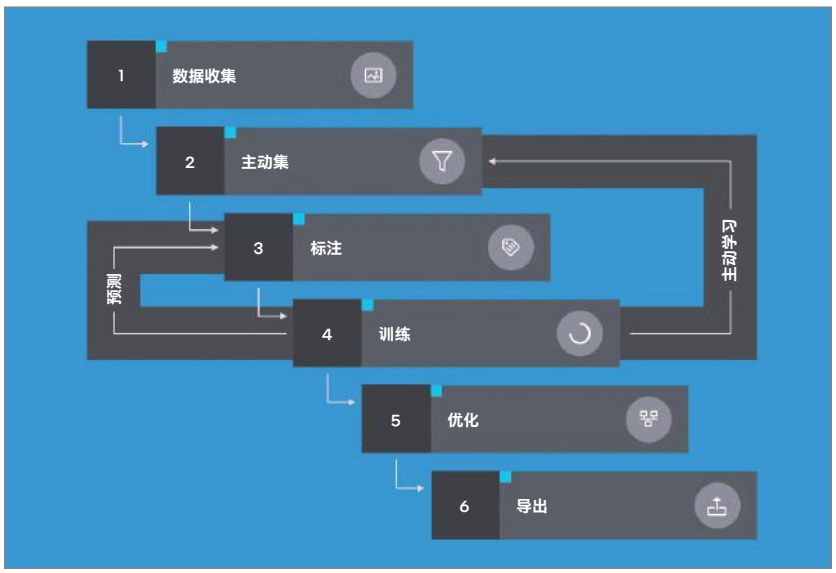
英特尔® Geti™ 平台提供的功能包括：

- **定制化的计算机视觉任务：**英特尔® Geti™ 平台加速了 AI 任务的模型创建，如分类、对象检测、语义分割或异常检测。
- **直观的用户界面：**英特尔® Geti™ 平台提供了便捷的图形用户界面和交互功能，如标注助手，允许几乎没有 AI 经验的团队成员协助计算机视觉模型训练。
- **任务链：**使用英特尔® Geti™ 平台，用户可以通过链接两个或更多任务来训练模型，而无需编写额外的代码，从而使用多步骤的智能应用程序。
- **智能标注：**英特尔® Geti™ 平台提供了使用铅笔、多边形工具和 OpenCV GrabCut 等绘图功能加速数据标注和图像分割，提高用户体验。
- **超参数优化：**英特尔® Geti™ 平台内置的优化方法使数据科学家的工作变得更加轻松，通过精炼对模型学习过程至关重要的超参数。
- **生产就绪的模型：**英特尔® Geti™ 平台输出的深度学习模型可以是 TensorFlow* 或 PyTorch* 格式。该平台还可以输出为 OpenVINO™ 工具包优化的模型，以便在英特尔® 架构的 CPU、GPU 和 VPU 上运行。

英特尔® Geti™ 平台的优势

- **节省时间：**英特尔® Geti™ 平台通过提供一个无需编码的直观平台，帮助缩短开发 AI 应用计算机视觉模型所需的时间。这加快了通常需要长时间训练计划的自定义 AI 模型创建的工作流程。
- **节省成本：**英特尔® Geti™ 平台有助于消除进入 AI 领域的障碍（如培训或支付专业服务费用），使计算机视觉 AI 建模对各种组织更加容易和经济高效，特别是对于较小的公司来说。降低与人员培训相关的成本进一步增加了整体价值。
- **灵活部署：**英特尔® Geti™ 平台可以在本地部署或通过云虚拟机部署，为那些使用这两种基础设施之一或两者的组织提供灵活性。

英特尔® Geti™ 平台的使用流程示例



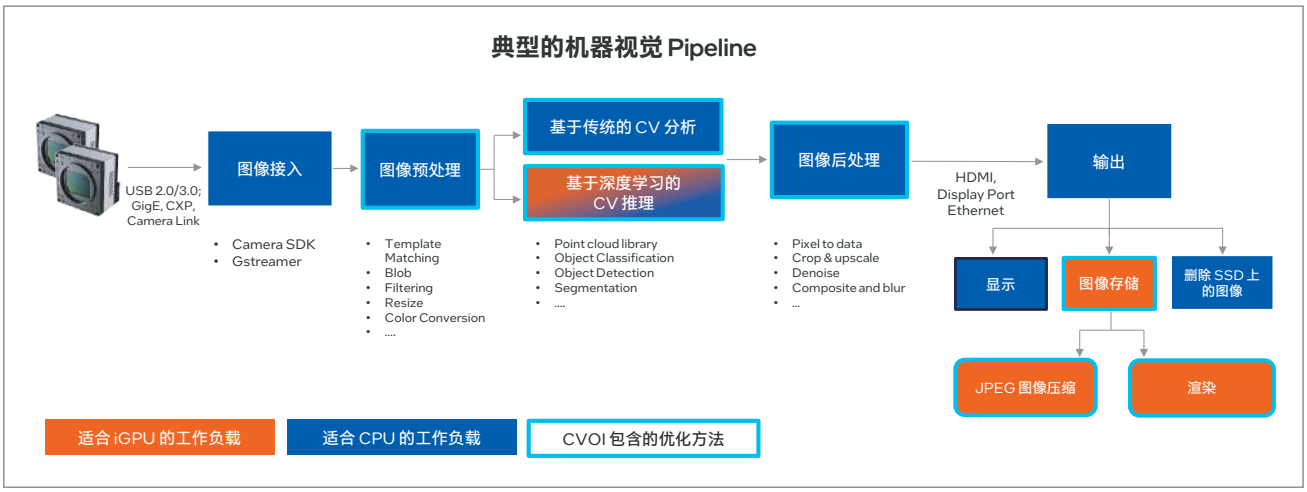
图：使用英特尔® Geti™ 平台训练 AI 模型的流程示意图

1. **数据收集** — 首先，您需要构建数据集，该数据集将用于训练您的模型。英特尔® Geti™ 平台提供了一个方便的机制，在上传多媒体数据（图像或视频）时进行标注。上传后，英特尔® Geti™ 平台会存储所有数据集。
2. **主动集** — 这个功能会自动选择多媒体数据进行最优化的训练会话。
3. **标注** — 这是您开始教机器如何思考的阶段。英特尔® Geti™ 平台提供了一套工具来促进标注工作。UI 中可用的标注工具会根据您选择的项目类型而有所不同。由于这是您将花费大部分时间的地方，英特尔® Geti™ 平台确保了简化的流程，并在您选择标签的方式上给予了您一定的自由。
4. **训练** — 在标注了预定义数量的多媒体数据后，英特尔® Geti™ 平台会自动启动基于这些标注好的数据的模型训练。完成第一轮训练后，英特尔® Geti™ 平台将自动开始对新的多媒体数据进行预测。
5. **优化** — 英特尔® Geti™ 平台使用 OpenVINO™ 工具包来优化模型，并通过在英特尔® 硬件上一次编写、随处部署的方法来提高它们的性能。您还可以随时使用新参数重新训练每个模型版本。
6. **导出** — 您可以导出模型并将其集成到您的应用程序中或与他人共享。

了解更多有关英特尔® Geti™ 平台的信息，请访问：
https://geti.intel.com/

2.2.4 英特尔® CVOI（工业机器视觉优化参考实现）

英特尔® Computer Vision Optimization Implementation（英特尔® COVI）是一个一站式资源库，其中包括最佳实践方法（BKMs）、指导手册和样例代码，专为全面优化英特尔® 平台上工业机器视觉的性能和稳定性而设计。该平台整合了英特尔® 的多种软件技术，如 OneAPI、OneVPL 等，以支持客户在英特尔® 产品上部署机器视觉解决方案。



图：典型的机器视觉 Pipeline 示意图

典型的机器视觉 Pipeline 包括若干子任务，如图像摄取、图像预处理、图像分析（传统的计算机视觉分析和/或深度学习推理）、后处理以及输出（显示、图像存储和图像删除）。在这些任务中，图像预处理、传统的计算机视觉分析和后处理最适合在 CPU 设备上运行。另一方面，深度学习推理、图像编码和渲染具有良好的并行性，使它们适合卸载到 GPU 加速器上。CVOI 基于这样的 Pipeline，提供了是一套全面的指南和示例代码，旨在优化英特尔® 平台上计算机视觉算子和整体系统 Pipeline 的性能。CVOI 是一个强大的工具，可以提高计算机视觉软件和系统的效率和可靠性，释放它们的全部潜力。

英特尔® CVOI 包含：

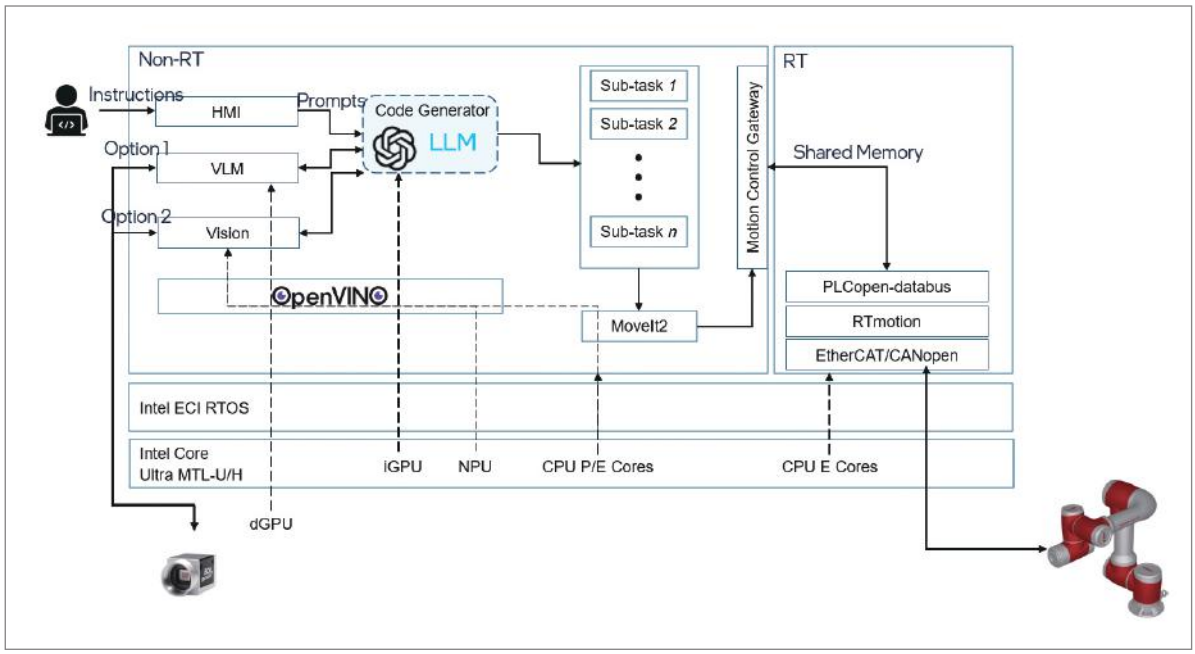
- 适用于英特尔® 第 12/13 代及以后的平台的性能优化最佳实践方法（BKMs）。客户可以参考该流程和方法论，自行进行优化。
- 在 2D 领域，包括均值滤波、模板匹配等在内的 10 多个加速的 OpenCV 算子参考示例代码。
- 在 3D 领域，超过 25 个加速的点云算子参考示例代码，集成到 PCL 和 FLANN 库中。
- 基于实际使用案例中机器视觉 Pipeline 的优化示例，例如 PCB 缺陷检测。
- 针对更高效利用混合核心架构（P/E Core）的最佳实践方法（BKMs）。

图：英特尔® CVOI 架构图

2.3 创新技术方案

2.3.1 大语言模型赋能工业机械手臂

大语言模型操控机械臂的技术解决方案架构如下图所示：



整体架构从任务的角度整个任务可以分解成三个阶段：

第一阶段是外部数据收集	第二阶段是任务理解和分解	第三阶段就是执行的阶段
这一阶段主要依赖两个外部输入，分别是人类的语音输入和摄像头的图像/视频信息，语音输入。运行中文分析模型分解匹配预先设置的提示词（prompts），视频和图像信息则是直接给到后一阶段。	有了上一阶段的提示词输入，大模型（目前在 Qwen 和 Phi3 上验证）会将其拆解为一连串的子任务序列，子任务序列和图像视觉信息结合就构成了更加准确的执行目标，例如一个子任务，移动到红色盘子上方，结合视觉信息找到的红色盘子，计算得出上方的坐标信息，经过 interpreter 时也会做代码级别的验证，之后就可以给到 MoveIt2 路径规划，来规划出中间的一个一个路点。	有了路点的信息后，通过共享内存机制，实时系统将会得到路点数据，通过 RTMotion 运动控制功能块，驱动机械臂上电机执行对应的加减速控制，来完成最终机械臂的整体运动，从而整体实现用过自然语言对机械臂的操控。

从系统的角度，这个架构中 Intel 的软硬件扮演着非常重要的角色：

硬件层面	软件层面
基于 MTL-H 的算力，其 NPU 和 iGPU 在语言的解析和图像/视频的处理上起到了关键作用，CPU 中性能核和部分能效核用于计算非实时域内的部分负载，而少部分能效核被单独隔离出来执行实时任务，配合专属的 Intel 网卡运行 EtherCAT 或 CANopen 的总线协议，来达到机械臂运动确定性的要求。	Intel 工业边缘软件平台（ECI）毫无疑问充当了整个系统的底座，实时运动控制部分运行在 PreeemptRT/Xᵉ nomai 环境下，而非实时部分通过叠加 OpenVINO™ 对大语言模型的推理提供了加速，其中 FastSAM 起到分割图像作用，而 CLIP 满足识别的功能，同时系统也对视频和图像处理提供了效率上的提升。

目前大模型的方案相对于传统机械臂编程方案有以下两点：

1. 任务拆解规划能力，sub task 体现出相对于传统机械臂靠人来分析拆解运动步骤的优势
2. 视觉模型均为 Zero Shot，不需要巨大的 ground truth 数据集支持，模型的泛化能力更强

2.3.2 基于视觉大模型的零样本或少样本异常检测

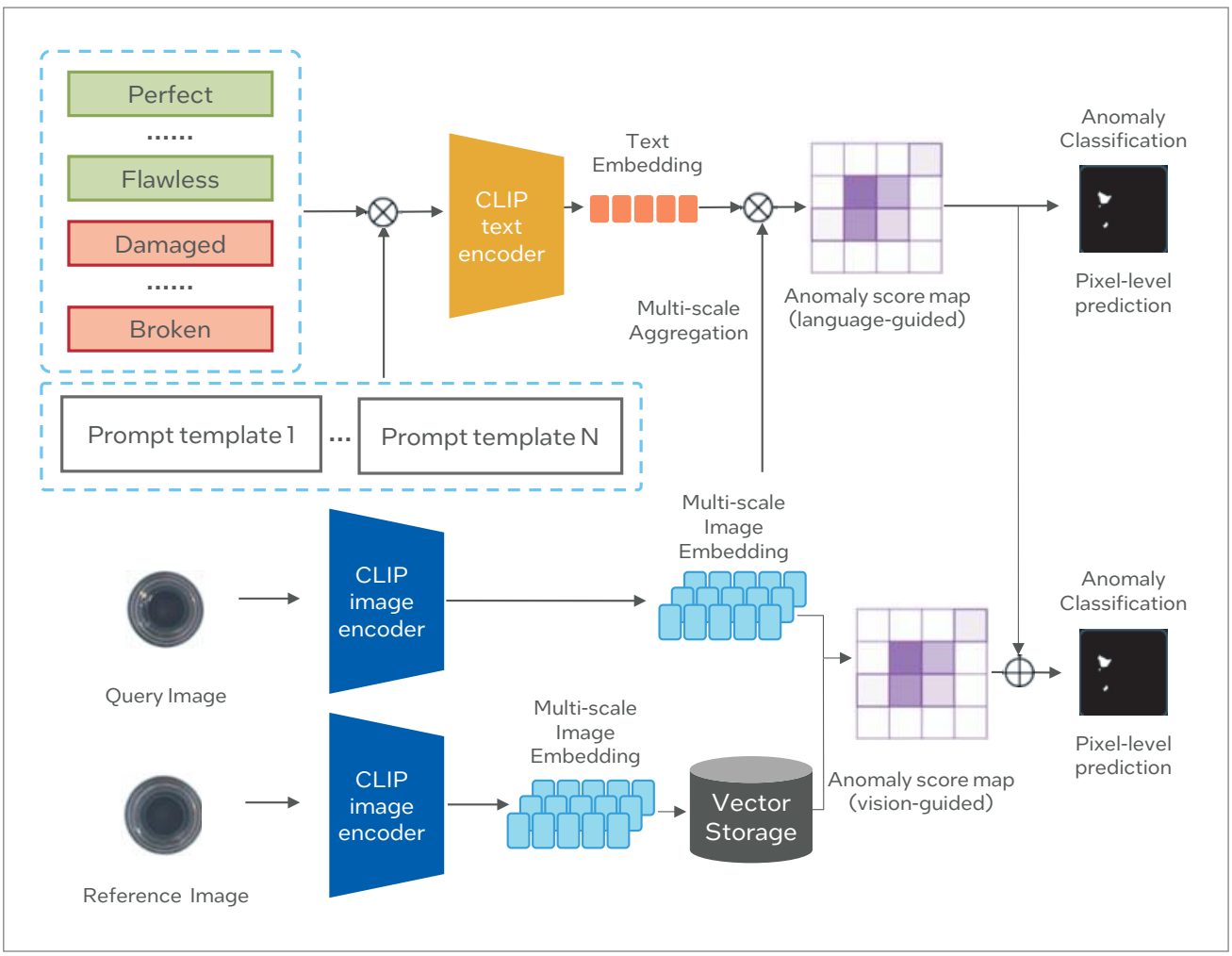
工业异常检测任务是制造业中的一项关键技术，它通过计算机视觉来识别和定位工业产品中的异常区域。这一过程对于确保产品的质量、避免潜在的安全风险以及减少经济损失具有重要意义。在传统的生产线上，产品质量控制和安全性监测往往依赖于人工检查或简单的自动化工具，但这些方法效率低下且容易出错。

随着深度学习技术的发展，基于卷积神经网络（CNN）的异常检测方法已经被广泛研究和应用。这些方法通过训练深度模型来自动识别产品中的缺陷或异常。然而，这些基于 CNN 的方法通常需要大量的标注样本来训练模型，尤其是需要大量的正常样本来学习正常情况下的产品特征。在实际应用中，这种对大量样本的依赖往往是不现实的，特别是在涉及用户数据隐私保护或新生产线快速部署的场景中。

为了解决这一问题，零样本或少样本异常检测（ZSAD or FSAD）目标是在没有或仅有极少量目标类别样本的情况下，依然能够有效地执行异常检测任务。这要求模型具备一定的泛化能力，能够在没有先验知识的情况下识别未知的异常类型。

具体来说，可以通过将产品的正常特征与异常特征用自然语言描述，并将这些描述与产品图像相结合，来训练模型。在预训练阶段，模型学习到了如何将图像内容与文本描述相匹配的能力。在实际应用中，即使没有异常样本，模型也可以利用其 Zero-shot 学习的能力，通过比较产品图像与正常情况的描述，来识别和定位异常。

这种方法的优势在于，它不仅减少了对大量标注样本的依赖，而且能够更好地适应新的生产环境和产品类型。此外，由于多模态预训练模型能够理解自然语言，因此可以更容易地通过语言指令来调整或优化检测任务，使得模型更加灵活和易于部署。



基于预训练的 CLIP 模型的零样本/少样本异常检测算法

如图所示，展示了一种基于 CLIP 模型（Contrastive Language-Image Pre-training 对比学习语言-图像预训练）的异常检测方法，该方法利用利用预训练的 CLIP 模型，其强大的文本和图像的理解能力来进行异常分类。

针对零样本异常检测的场景。首先，使用预先设计的一系列提示模板（Prompt template）通过 CLIP 文本编码器进行处理，生成对应的文本嵌入。同时，输入的查询图像通过 CLIP 图像编码器进行处理，生成多尺度图像嵌入。多尺度的图像嵌入通过聚合和文本嵌入进行相似度对比，形成语言引导的异常得分映射，异常得分映射上采样获得异常区域分割结果。同时，通过阈值判定整张图像的是否异常的类别。

虽然大模型具有更强的场景迁移能力，能够实现零样本异常检测，但是对于工业场景来说，通常对于检测的准确度有较高的要求，零样本异常检测很难达到工业客户的要求。少样本异常检测方案能够通过提供少量的参考样本显著提升检测准确度，更适合工业场景。

少样本异常检测是在零样本异常检测方案的基础上要求用户额外提供一组参考图像，参考图像和查询图像一样，通过 CLIP 图像编码器进行处理，并将生成的多尺度图像嵌入聚合，同查询图像的图像聚合特征进行比对，生成视觉引导的异常得分映射。视觉引导的异常得分映射和文本引导的异常得分映射加权平均获得最终的异常区域分割结果和类别判别结果。

除了上面介绍的算法参考方案的实现，英特尔® 还提供了从硬件算力到软件工具的全方位支持。首先英特尔® 为工业用户提供可以支持不同级别 AI 工作负载的独立显卡-锐炫™ 显卡系列和集成显卡两种不同的硬件算力平台。

同时，在软件工具方面，英特尔® 提供了开源的 OpenVINO™ 工具包，为 AI 工作负载提供了在英特尔® 硬件平台最大的加速和优化。同时，OpenVINO™ 可简化和优化跨不同平台运行的 AI 推理代码开发。利用 OpenVINO™ 工具套件的优化技术，例如模型量化、层融合和硬件级优化，用户可以显著提高神经网络推理的效率。部署到独立 GPU 上时，这些经优化的模型可以利用 GPU 的并行处理能力，从而加快推理。最新版本的 OpenVINO™ 2024.3 通过增加更广泛的模型支持、减少内存占用以及为大型模型引入其他压缩技术进一步提升推理性能。

丰富的算力平台和软件工具包支持灵活的将异常检测方案部署在英特尔® 集成显卡和独立显卡上，满足用户满足客户对于不同部署场景、不同性能和成本的需求。

2.3.3 RAG 检索增强生成模型实现

大语言模型（Large Language Models，LLMs）在自然语言处理领域取得了显著的进展，但它们在实际应用中在准确性，知识更新速度以及答案透明度上都有挑战。

为了解决这些挑战，检索增强生成（Retrieval-Augmented Generation，RAG）技术应运而生。RAG 在优化 LLM 方面，相较于其他方法具有显著的优势【Shuster et al., 2021; Yasunaga et al., 2022; Wang et al., 2023c; Borgeaud et al., 2022】：

RAG 作为一种关键的方法，通过有效地结合了大模型的能力和外部知识库的丰富性，提高了大语言模型在各种任务上的表现，尤其是在需要最新信息和专业知识的场景中，有效减少语言模型中出现的虚假信息。RAG 具备高度的定制化能力。通过索引与特定领域相关的文本语料库，RAG 能够为不同领域提供专业的知识支持，通过引用信息来源，用户可以核实答案的准确性，这增强了人们对模型输出结果的信任，使得生成的回答更加准确可信。随着技术的不断发展，RAG 有潜力进一步提升大语言模型的实用性和可靠性。

RAG 适合的场景：

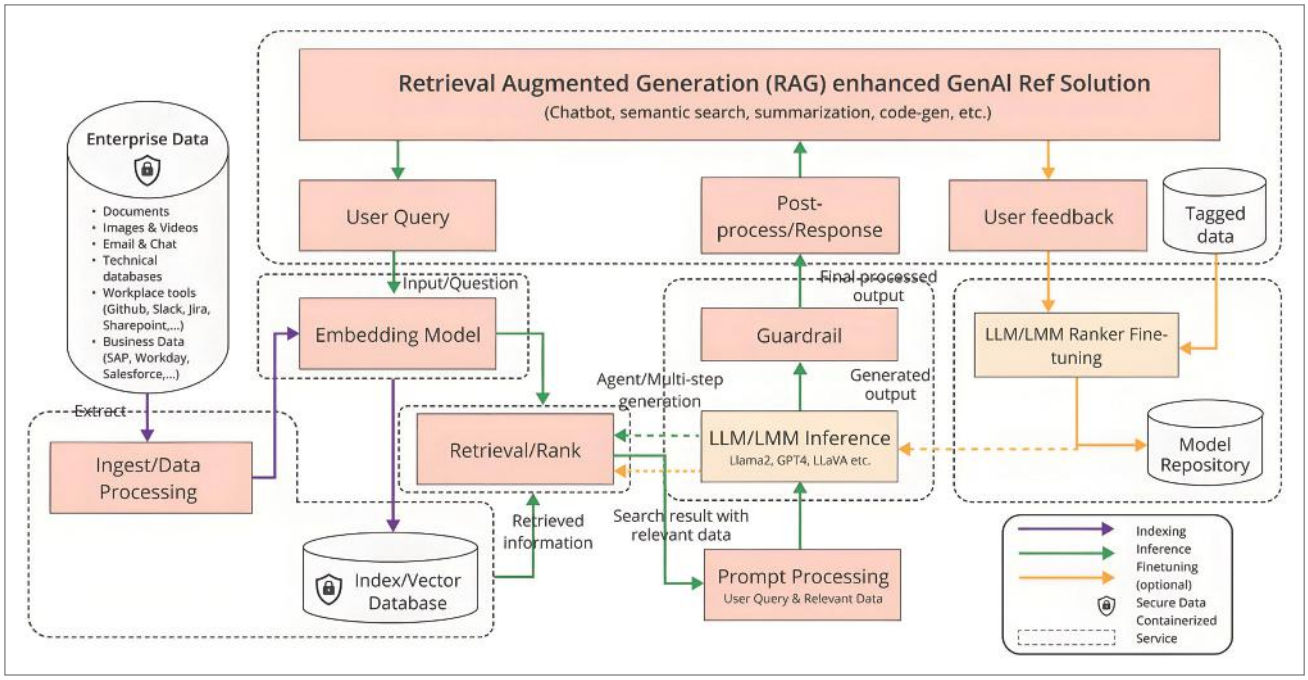
1. 知识密集型任务，比如客服，医疗，法律等特定领域。
2. 动态内容生成，需要实时或频繁更新的内容生成场景。
3. 文档辅助的任务，如合同审阅、法律建议、科学论文分析等。
4. 减少大模型的幻觉问题，因为模型生成的内容是基于检索到的真实信息。
5. 相对微调来说，不需要训练模型，更经济高效。

英特尔® Open Platform for Enterprise AI（OPEA）是一个开放平台项目，通过整个解决方案生态系统的多合作伙伴组件来实现可组合和可配置的生成式 AI 云原生解决方案。

OPEA 具有以下特性：

- **高效：**利用现有的基础设施，如 AI 加速器或其他您选择的硬件。
- **无缝：**与企业软件集成，支持系统和网络的异构性，提供稳定性。
- **开放：**汇集了最佳的创新成果，并且不受专有供应商锁定的限制。
- **无处不在：**通过为云、数据中心、边缘和 PC 构建的灵活架构，实现无处不在的运行。
- **可信：**具有安全的企业级流程和工具，支持责任、透明度和可追溯性。
- **可扩展：**提供访问活跃的合作伙伴生态系统，帮助构建和扩展您的解决方案。

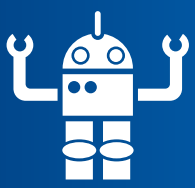
OPEA 为客户提供了 RAG 参考 pipeline，其架构如下图所示。



OPEA 使用微服务为企业创建高质量的 GenAI 应用程序，简化生产环境中的扩展和部署流程。这些微服务利用一个服务组合器将它们组装成一个巨大的服务，从而创建出真实世界的企业人工智能应用程序。OPEA 为客户提供了 RAG 参考 pipeline，其架构如下图所示。该参考结构旨在通过高效的数据检索和生成式 AI 技术，提高企业数据处理的效率和准确性，同时确保数据的安全性和合规性。该架构以可组合微服务模组的方式提供包含数据存储，提示引擎，检索优化器，以及 LLM 等构建 RAG 服务的核心功能。图中展示了建立索引，问答推理，RAG 微调的流程。在基础架构的基础上，英特尔® 还对 RAG 进行了全链条的优化，包括从数据增强，分块优化，embedding 微调，向量化数据库的优化，优化查询，重排序以及内容压缩等，帮助用户降低成本提升效率。

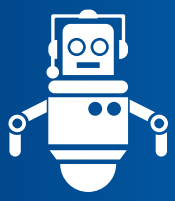
2.3.4 人形机器人

在工业领域，人形机器人的出现标志着自动化技术的一次飞跃。这些机器人集成了先进的传感器、控制系统和人工智能算法，使得它们能够在复杂的工业环境中执行精密作业，提高生产效率，同时降低人力成本。通过模仿人类的动作和决策过程，人形机器人能够无缝地融入现有的工作流程，执行从组装、焊接到质量检验等多样化任务。随着机器学习和认知计算技术的进步，人形机器人正在变得更加智能和自适应。它们能够实时分析环境数据，优化工作策略，甚至在遇到未知情况时进行自主学习和决策。这种技术革新不仅提升了工业生产的灵活性和响应速度，也为未来工厂的智能化和数字化转型奠定了基础。



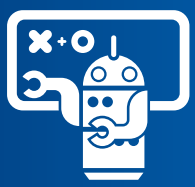
基础人形机器人负载

以满足需求的运动执行能力为核心，任务相对固定，以传统运控算法为主。



标准人形机器人负载

借助强化学习增强运动执行能力，借助本地及云端大模型实现覆盖场景需求的感知泛化能力与任务生成能力。



旗舰人形机器人负载

在智能性，自主性层面增强，在技术路径层面通过端到端模型代替分层决策模型，整体负载可能收敛至以 AI 为主。

在人形机器人内部，CPU、GPU 和 NPU（神经处理单元）各自承担着不同的任务，它们的设计和优化针对的是不同类型的计算需求。以下是它们各自的特点和处理任务的差异:

CPU 负责小脑:

- 负载：实时控制，轨迹规划，运控算法，步态算法，实时控制。
- CPU 是通用处理器，设计用于处理各种类型的计算任务。具有较少的核心，但每个核心的计算能力较强，适合执行顺序性强的复杂任务。
- CPU 通常负责机器人的高级决策逻辑、任务规划、运动规划、传感器数据的融合处理等。
- CPU 也负责协调其他处理器的工作，如分配任务给 GPU 或 NPU。

GPU/NPU 负责大脑:

- 负载：VSLAM，环境感知，任务编排，自主规划，模仿学习，强化学习。
- 在人形机器人中，GPU 常用于视觉处理任务，如图像识别、视频分析、3D 建模和环境映射。
- 随着深度学习的发展，GPU 也被广泛用于加速神经网络的训练和推理过程。
- NPU 是专门为神经网络计算和机器学习任务设计的处理器。通常用于执行机器人的感知任务，如物体识别、语音识别、自然语言理解等。

在实际应用中，这三种处理器可能会协同工作，各自处理它们擅长的任务。例如，CPU 可能会处理传感器数据融合和决策逻辑，GPU 负责图像处理和深度学习模型的并行计算，而 NPU 则专注于快速高效地执行神经网络推理任务。这种分工可以使人形机器人在执行复杂任务时更加高效和智能。在选择处理器时，机器人的开发者需要考虑到机器人的任务类型、实时性要求、能耗限制、散热能力、成本预算等因素。

利用英特尔® 酷睿™ Ultra 处理器提升竞争优势，在单个 SOC 中，P-core（性能核）、E-core（能效核）、英特尔锐炫™ GPU[3]，英特尔 NPU 以及英特尔® AI Boost[4] 等众多计算引擎协同加速边缘 AI 推理。GPU 支持 2D 视觉，大模型运算、2D/3D 视觉其他并行计算，CPU 支持 VLSAM 计算。AVX2，核数/线程提升，增强的 GPU，Intel Math Kernal Library 优化 CPU 对传统算法运算性能。以上功能在 SOC 实现，减少对独立加速器的需求，帮助降低系统复杂性和成本。

03

成功案例

3.1 英特尔：智能晶圆视觉检测

背景与挑战

传统的质量检测是在某些间隔“离线”进行的 — 通常是在两个或更多添加步骤完成后。例如，在晶圆研磨和膜应用后检查一定比例的晶圆。但也因此，离线检测带来了几个挑战：

- 高风险产生废品和缺陷逃逸。在对一个晶圆批次进行检查时，可能已经处理了多达九个更多的批次。如果机器或过程错误引入了缺陷，很可能会损坏更多的晶圆，导致高废品和低质量产品的风险。
- 检测受阻。因为在离线检查之前会发生多个添加过程，直接检查晶圆表面可能是不可能的。例如，在检查之前将膜应用于晶圆的背面，但这阻止了对研磨错误的直接检测。
- 过程依赖人工。检查员使用显微镜手动检查晶圆。但随着产品变得更小，人类发现微米级缺陷更加困难。例如，相较于整个晶圆 300 毫米的直径，研磨缺陷可能只有 5 微米长，找到缺陷就像在足球场上找到一粒米。
- 必须在有限的空间内运行，不干扰研磨工具的操作；且不需要对研磨工具进行任何修改，能够与研磨工具通信（例如停止其操作）。

在保证质量的同时，随着产量加大，缺陷检测工作将需要增加大量的工程资源；即便如此也依然可能存在无法跟上生产速度的问题。此外，因为所有新一代的英特尔产品都在向高级封装转型，1 个单一的缺陷可能会导致大量废品。不仅如此，在微小的产品上，电路空间非常有限；一个逃逸的缺陷可能会在客户现场导致关键故障，可能对客户的业务和英特尔的质量和可靠性声誉造成损害。

解决方案

通过高分辨率摄像头每秒拍摄多张图像，同时研磨工具对晶圆进行薄化处理，并安装保护性聚酯膜。将收集到的图像由边缘的机器学习模型分析处理。如果检测到缺陷，解决方案可以发出警报甚至停止设备。该解决方案包括英特尔® 酷睿™ i9 处理器、英特尔® 至强® 可扩展处理器和英特尔® ARC A770 独立 GPU，正在部署到英特尔® 组装和测试工厂。

方案组件

基于第 12 代英特尔® 酷睿™ i9 边缘视觉控制器和英特尔® ARC A770 独立 GPU 以加速数据处理。数据传输速度约为 200 Gbit/秒。该系统可以存储约 40 TB（相当于三周的量）的原始图像和检查结果。

机器学习模型部署在英特尔® 私有云中的高性能计算服务器上，使用英特尔® 至强® 可扩展处理器进行训练。模型在数十万张晶圆图像上训练之后，部署到生产线的边缘位置。边缘上的模型推理工作负载（在摄像头控制器上）可以卸载到英特尔® ARC A770 独立 GPU 上，以加速图像分析。英特尔® 至强® 可扩展处理器提供强大的计算能力来处理繁重的工作负载，加快训练过程。根据我们的经验，模型训练的速度比使用其他处理器快达 50%，现阶段每季度进行一次完整迭代训练流程（如下图）。

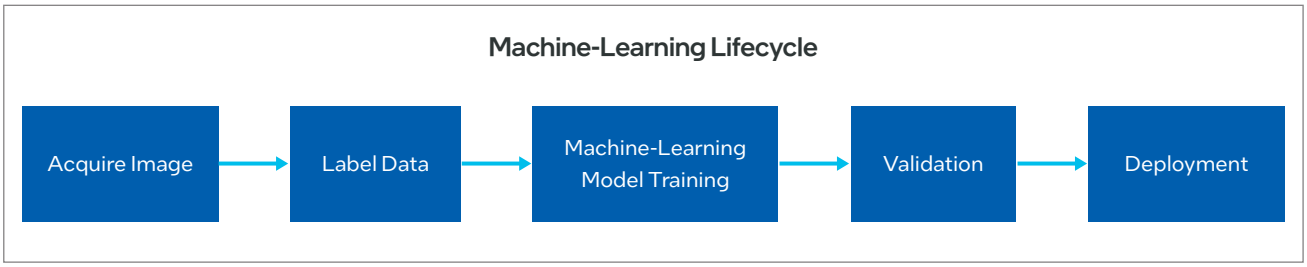


Figure. Our solution spans the entire machine-learning lifecycle, from raw data acquisition and labeling to training and validation to deployment.

收集的图像由包含 50 个底层核心算法的混合机器学习模型分析，包括各种如基于 ResNet50 的检查、语义剪辑分割（以实现像素级的缺陷识别）以及识别和测量（自动识别图像中的边缘类型和自动测量距离）等功能。模型经英特尔® OpenVINO™ 工具包优化，提高在英特尔® 硬件平台上的效率。工业视觉控制软件支持图像审查和模拟，所有功能以模块形式集成，便于扩展应用，支持低代码操作和流程向导。

方案优势

标准化设计实现了快速部署和易于扩展到新的用例。该解决方案能够准确识别晶圆研磨过程中的多种类型缺陷，包括凹痕、各种大小的划痕、研磨痕迹、污点、裂纹、气泡、晶圆偏移和安装偏移。与离线检查相比，使用在线检查可以更早地检测到多达 50% 的晶圆研磨问题。当警报响起时，可以对晶圆进行返工，这避免了下游过程中的整片晶圆分层。更重要的是，该解决方案超越了离线计量所能实现的：在问题发生时检测偏差，检测过程缺陷并迅速关闭工具，实现了检查框架清洁度和内环的新能力。

帮助工厂实现业务效益：

- 避免废品，每年为工厂节省高达 200 万美元。
- 降低了业务风险。
- 更高的产品质量。
- 使工程师免于繁琐的手动离线检查。

3.2 美的楼宇科技美控：楼宇 AI 节能解决方案

背景与挑战

随着全球对可持续发展的呼声日益高涨，建筑行业正面临前所未有的转型挑战。据《2023 中国建筑与城市基础设施碳排放研究报告》显示，建筑运行阶段的碳排放占据了全国碳排放总量的 21.9%，其中暖通空调系统能耗占据了建筑能耗的近半壁江山。而在暖通空调系统能耗中，基于暖通空调业务数据测算，制冷机房系统能耗占据 60%，仍具有较大的节能空间。

目前，常见的制冷机房节能改造措施包括更换高效设备如主机、水泵、冷却塔，以及优化系统设计。此外，通过升级或优化制冷机房的自动控制系统，可以进一步提高运行效率。尽管许多机房已配备自动控制系统，但节能潜力仍然巨大。这主要是因为现有的自控系统采用基于规则的简单逻辑控制，难以适应暖通空调系统的动态和非线性特性，以及设备间的相互影响，导致系统无法持续高效运行。因此，如何在确保舒适性和稳定性的前提下，实现暖通空调系统的经济性最优运行，成为了行业面临的一大挑战。这不仅需要智能化的技术支持，以优化系统运行效率，还需要为终端用户持续带来价值，实现经济与环境效益的双重提升。

解决方案

针对现有问题，美的楼宇科技推出了面向暖通空调水系统的智慧控制系统 — Smart Control。该系统利用 “数据驱动 + 物理约束” 的 AI 建模技术，构建高精度模型，通过模型仿真预测与全局寻优，实时调整系统参数，如供水温度和设备组合，实现系统级的优化节能服务。

Smart Control 的快捷部署能力允许客户无感改造，方便切换模式，方案设计精准满足节能需求，且投入成本低，风险小，运营费用低。作为美控脑机 II 代 Smart Control，采用自研边缘工控机硬件，搭载了英特尔® 的强大处理器和计算架构，能够快速处理大量数据，实时执行 AI 优化算法。这种硬件加持使得 Smart Control 算法引擎能够更加高效地进行数据驱动的决策，精确调整控制策略以适应不同环境条件和使用需求。通过这种硬件和软件的结合，美的不仅显著提升了系统的运行效率，还大幅降低了能耗和碳排放，为公共建筑的节能改造提供了一种经济高效的解决方案。

名称	KONG Smart Control
操作系统	支持 Windows/Ubuntu 操作系统
CPU	Intel® Core™ i7 1255U
系统内存	DDR4 3200Hz 内存模组 16G *2
硬盘	2T SATA3 固态硬盘
电源要求	DC12~28V，3pin 端子，过流保护，过压保护
网络	2*Intel® i210 千兆网口
尺寸	(L*W*H): 228.5*160*75mm

方案优势

深圳某办公楼的典型案例展示了美的楼宇科技在机房节能改造方面的成功应用。该项目位于深圳市南山区，机房自 2011 年建成后自控系统基本失效。美的楼宇科技对机房进行了升级改造，首先将原有失效的自控系统升级为基于美的 KONG DDC 的系统，然后进一步安装了 Smart Control 智慧控制系统，并利用 AI 算法实时优化运行参数，实现了 BA（Building Automation 楼宇自控）控制模式和 AI 控制模式的无缝切换。

通过这次改造，机房的节能率超过了 30%，并在最热月份实现了平均能效达到 4.32 的卓越表现，其中 Smart Control 的 AI 控制相比 BA 自控基础上额外提供了 18% 的节能效果。在 6 月份的高温季节，通过对比测试显示，新一代智慧控制系统相较于常规 BA 控制系统能效提升了 14.89%。Smart Control 方案凭借其三大核心能力 — 系统仿真预测、全局实时优化以及健康诊断分析 — 为暖通空调系统带来了革命性的智能化管理。

Smart Control 方案的快捷部署特性，允许客户无感改造，即在不干扰现有运营的前提下轻松部署，同时提供了灵活的模式切换功能，以适应不同的运行需求。基于脑机的边缘端服务进一步强化了方案的执行效率，英特尔的技术支持使得数据处理和分析更加迅速和精确。方案的小而精设计，专注于节能服务，实施起来简便且能精准满足客户需求。投入轻量化的特点，意味着 Smart Control 在资金投入、项目风险和运营费用上都大大优于传统节能改造项目，为客户提供了一种经济高效的节能解决方案。

图：Smart Control 脑机

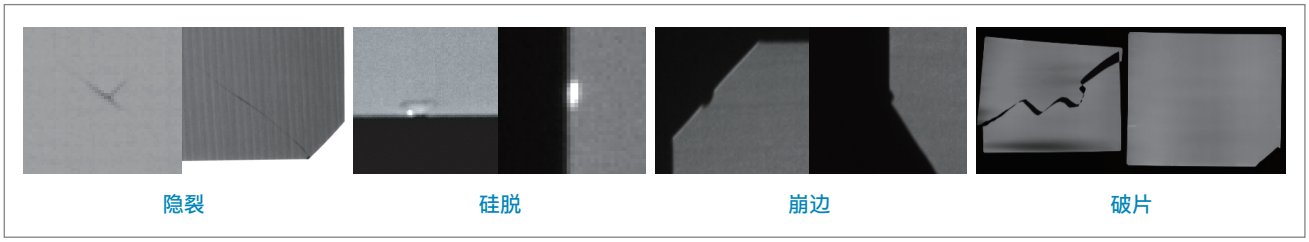
图：Smart Control 软件界面

图：Smart Control 系统架构图

3.3 利珀：晶硅电池隐裂检测产品

背景与挑战

可见光波段的 AOI 技术已经在各行业的诸多工厂中得以广泛应用，以光伏行业为例，从最初的终品分选检测，拓展到 PE 色差分选和丝印缺陷检测，直至当前 AOI 技术在花篮、石英舟、石墨舟等设备中的普及，利用可见光波段成像方案，已日益成为光伏行业保障产品品质与提升生产工艺流程的标配。然而，可见光成像技术在应用上存在着一个固有的局限性，这一局限性恰好构成了光伏产业中电池片工厂实现全制程生产质量控制的一个关键环节，即隐裂检测。



图：SC 系列产品能够检测典型缺陷外观

隐裂（Micro Crack）即肉眼难以直接在被测物体表面观察到的细微裂纹。在光伏行业的自动化生产流程中，存在隐裂的电池片，在经历自动搬运作业时，极易受到外力的影响，导致隐裂扩展并演变为明裂，进而引发碎片现象。这些碎片在自动化流水线的传输皮带上或存储于各类容器中时，可能因相互接触或机械作用，致使邻近的完好电池片也遭受损害，转化为碎片，从而对整个生产线的效率与产品质量构成严重影响。隐裂可能产生于电池片自动化流水线的任何工艺段，鉴于此，为了有效管控工艺流程与产品品质，各工艺段的上料及下料工位均需配备隐裂自动光学检测（AOI）设备。这一需求不仅凸显了隐裂检测在光伏生产中的重要地位，还极大地推动了隐裂 AOI 设备市场的扩张，使得其需求量远超光伏行业内其他单一类型的 AOI 设备，展现出广阔的市场前景。

解决方案与优势

利珀在光伏行业深耕多年，针对隐裂检测这一行业刚需，设计研发了 SC 系列晶硅电池隐裂检测产品。该系列产品兼容的硅片规格包括 156mm~230mm 方片或准方片，硅片厚度覆盖 100~220um，不仅能够检测隐裂，还能够同时对硅脱、崩边、破片等缺陷进行检测，不仅能够适用于传统的 PERC 工艺，还能兼容目前最新的 HJT、BC 和 TOPCON 工艺。目前，该系列产品已在众多光伏行业领先企业中得到应用，并逐渐成为行业内普遍采用的标准解决方案。该产品基于英特尔® 酷睿™ 系列 CPU 和利珀自研的机器视觉平台软件灵闪（Intelliblink）及底层算法库 Leaper Vision Toolkit（LPV）进行打造，从早期的 4 代 i5-4570 和 6 代 i7-6700 到现在的 12 代 i5-12400F，Intelliblink 和 LPV 为尽可能提升运算速度所使用的 CPU 指令集也从 SSE 4.2 升级到了 AVX2。同时，为了充分利用英特尔® 12 代酷睿 CPU 的多核性能，利珀还针对不同的底层算法进行了最优的多线程并行化设计优化。目前，通过充分利用 AVX2 指令集提供的 256 位宽指令以及针对特定英特尔® CPU 实施的精细多线程优化策略，利珀显著提升了隐裂检测产品中各类图像处理算法的执行效率，实现了对英特尔® CPU 计算潜能的深度挖掘与高效利用。

面对隐裂检测面临对比度低及外观形态高度多样化的技术挑战，利珀公司依托多年累积的丰富图像样本库，还成功训练出了一套光伏、半导体等多行业通用的隐裂缺陷智能检测 AI 模型。在终端工厂的部署过程中，该模型将依托于 OpenVINO™ 框架进行推理运算，同时结合锐炫™ A380 显卡的强大性能对模型推理过程进行加速处理，从而构建出一套既高效又极具成本效益的 AI 推理解决方案。

图：利珀 SC 系列晶硅电池隐裂检测产品示意图

3.4 诺达佳：基于 AI 的在线式视觉随动同步点胶机应用

背景与挑战

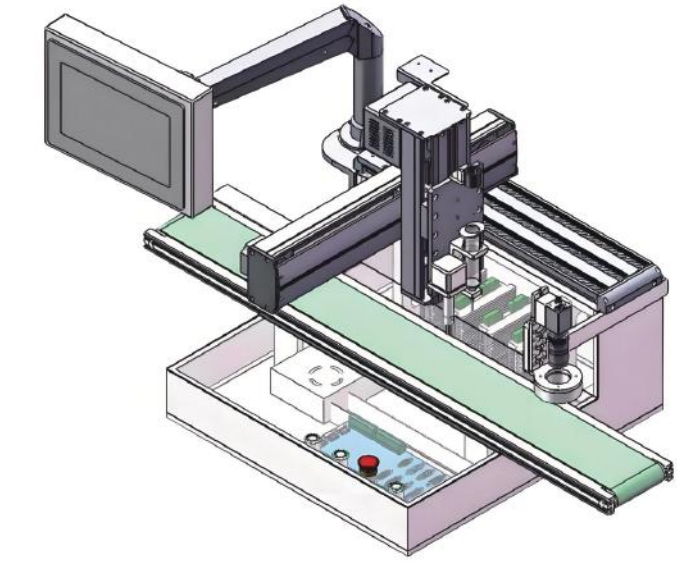
传统的点胶系统方案中通常通过 PLC + 运动控制板卡构成控制系统，根据预先设定的路径进行点胶的动作，效率比较低，功能相对比较固定，不易实现扩展性和灵活的配置，多种硬件的耦合，不但数据交互效率不高，而且技术升级需要更换设备或进行生产线改造，成本高昂且耗时长，给设备生产厂家带来成本控制和换料操作的困扰。

在线式视觉随动同步点胶机系统能应对这一挑战。在工业生产中具有广泛的应用，特别是在需要高精度和高效率的生产环境中。这种设备通过集成先进的机器视觉技术，能够实现快速准确的点胶操作，大大提高了作业效率和产品质量。其应用领域包括但不限于电子产品制造中确保产品的电气性能和结构稳定性，汽车制造中确保密封和固定工作的准确性，提高车辆的安全性和性能，在医疗器械制造中，可以精确地涂抹胶水，确保设备的密封性和耐用性，保障患者的安全，航空航天领域中，能够提供必要的支持，确保组件的牢固和安全。

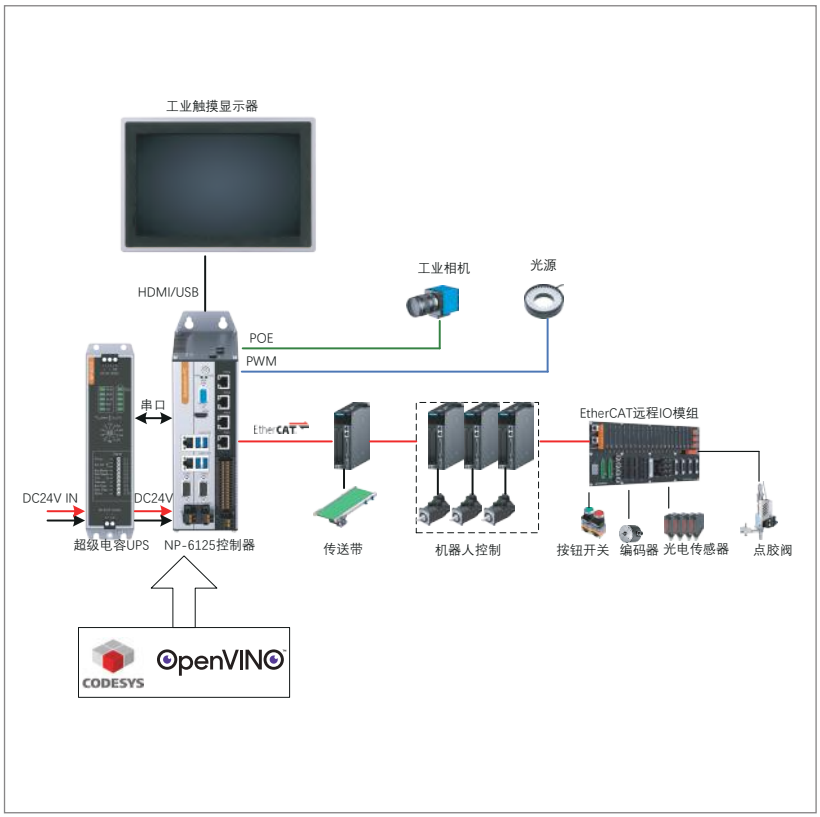
解决方案

基于 NP-612x 系列工控机集成 CODESYS 和 OpenVINO™ 工具包的 AI 视觉运动控制解决方案，集成了高性能工控计算能力、IEC61131-3 的开发环境以及尖端的深度学习工具，为机器视觉应用带来了一个性能优异、稳定可靠、易于集成及扩展的平台。

系统机械结构由直角坐标系机器人，传送带以及点胶运动结构组成，最终的结果为当物体在传送带上匀速移动的过程中，通过视觉可以采集测量出物体的轮廓并生成点胶路径，机器人可以在线跟随物体的移动，并进行精准的点胶的动作。



NP-612x 系列控制器搭载第 12 代英特尔® 酷睿™ i3/i5/i7 高性能桌面级多核 CPU，小体积，大算力，同时集成 POE 网卡、光源控制、串口、USB 以及 DIO 等多种 IO 功能，与传统工控机搭配板卡的方案相比，不但体积小，精巧的结构设计，保证了系统的连接可靠性，丰富的 IO 接口可适应于多种应用场景。通过搭配 CodeSys 软 PLC 的环境，极大发挥了 CPU 的多核处理能力和运动控制性能，硬件功能软件化，进一步实现了传统的硬件组合的解耦，通过共享内存来实现功能组件之间的通讯，解决了大数据交互的速度瓶颈。通过 EtherCAT 总线扩展 IO 或者执行单元，不但有利于设备安装，而且在后期的维护和升级过程中带来更多的灵活性。控制系统搭配超级电容 UPS 守护系统与数据安全，断电无忧。OpenVINO™ 提供性能优化工具，如模型优化器、推理引擎等。支持 OpenCV 库，方便图像处理和显示。



方案优势

NP-612x-H1 是一款高性能的工业计算机或嵌入式系统，专为需要复杂计算和高效数据处理的应用场景设计。其主要功能特点包括：

- **高性能图像处理：**通过集成 OpenVINO™ 加速深度学习模型在各种硬件平台上的推理速度，AI 视觉运动控制器能够快速、高效地处理图像识别、物体检测、人脸识别等多种计算机视觉任务，适用于工业检测、自动化引导车辆（AGV）、智能安防等场景。
- **实时控制能力：**CODESYS 是一个国际领先的跨平台自动化软件开发环境，支持 IEC 61131-3 编程标准。它允许开发者使用高级语言（如 Structured Text，Ladder Diagram 等）编写控制逻辑，实现对机械设备的精准控制。结合 NP-6122-H1 的高性能处理器，可以确保控制指令的实时响应，满足工业自动化中对时间敏感的控制需求。
- **灵活的硬件接口：**NP-612x-H1 通常配备有丰富的 I/O 接口（如 Ethernet、USB、串口等），便于连接各种传感器、执行器、相机等设备，为 AI 视觉系统提供全面的硬件支持。这使得系统能够采集多样化的输入信息，并根据 AI 分析结果迅速作出反应，执行相应的动作控制。
- **支持多种点胶工艺需求：**如单点、直线、不规则多段线、弧形、画圆等，能够满足不同产品的特定需求。其操作简便，可以与各种自动化设备进行连接和控制，进一步提高了生产效率和产品质量。

3.5 新松：智能巡检机器人

背景与挑战

移动机器人凭借着在效率、场景适应性、经济性等方面的优势，已经日趋广泛地应用于工业巡检、安防巡逻、园区服务等诸多场景之中，并展现出了巨大的发展潜力。作为移动机器人的关键模块，移动机器人控制器承载着传感数据集成、数据处理、控制等重要负载，需要在算法、算力、稳定性、易用性等方面克服严峻的挑战，以加速移动机器人方案在不同场景的落地应用。

要加速移动机器人的场景化落地，满足用户对于移动机器人日渐增长的需求，机器人产品与方案提供商需要在算力、稳定性、经济性等方面入手，化解如下挑战：

1. **复杂的负载带来较高的算力要求：**为了满足更复杂、更广泛的场景应用所需，实现更高的任务精确度、智能性，移动机器人正在强化 3D 点云 + 视觉多传感器融合、深度学习推理等技术的应用，这些应用负载带来了较高的算力要求。
2. **技术门槛较高带来产品开发困境：**移动机器人是一种较为复杂的机器平台，巡检、物料操作等场景对于控制实时性要求较高，在技术实现难度、算力平台复杂性等方面，远超普通机器人，这就带来了较高的技术门槛。
3. **繁多的模块与外设带来成本、可靠性、运维等多重挑战：**在移动机器人开发过程中，各种模块、外设数量繁多且种类多样，不同组件的部署会导致机器人设计面临较大的挑战。

解决方案

基于英特尔® 架构的移动机器人控制器方案，该控制器搭载了英特尔® 酷睿™ Ultra 处理器，结合英特尔® oneAPI 工具包、英特尔® 发行版 OpenVINO™ 工具套件、英特尔® Robotics SDK 等软件，能够灵活、充分地贴合新松机器人产品对算力和 I/O 的需求，可缩短研发周期、降低计算平台投入成本。同时，该方案集成了英特尔® 与新松联合推出的 3D 点云 + 视觉多传感器融合技术，能够满足移动巡检等场景的应用需求，帮助客户实现数字化、智能化转型。

该控制器不仅具备硬件模块，还集成了导航、避障等算法及软件，能够集中处理人机交互、充能储能、运动控制、环境感知、无线通讯等负载，加速负载的运行，同时满足移动机器人在稳定性、扩展性等方面的要求。目前，该控制器已经广泛应用于新松智能巡检机器人、电力无人值守机器人、安防巡逻机器人、园区无人值守机器人、轮式井下无人值守机器人等多种产品之中，满足不同细分场景的应用需求。

图：新松移动机器人

方案优势

新松移动机器人控制器在硬件层面上，采用了支持英特尔® 酷睿™ Ultra 处理器的模块化设计，提供了卓越的扩展性和灵活性。这款处理器采用了革命性的混合集成片上系统架构，通过英特尔® 的 Foveros 3D 封装技术实现高效连接，配备了强大的英特尔锐炫™ GPU 和首次集成的神经网络处理单元（NPU），专为 AI 加速而设计，满足了移动机器人在 AI 推理和复杂工作负载方面的高性能要求。未来几代的英特尔® 酷睿™ Ultra 处理器将进一步提升 AI 计算能力，保持与现有平台的兼容性，推动移动机器人技术的创新和应用。

在软件及算法层面，新松移动机器人控制器支持先进的 3D 点云和多传感器融合定位技术，以及固态激光雷达和超声波雷达的避障技术，确保了多维度的安全保护。新松还计划采用英特尔® Robotics SDK，加速移动机器人应用程序的开发和部署。英特尔® Robotics SDK 提供了基于 ROS*2 的库和工具，支持跨多个硬件配置的部署，加快了客户应用程序的上市时间，构建了从设备到边缘的完整端到端解决方案，推动了移动机器人在多种场景中的广泛应用和技术演进。

SIASUN

3.6 华泰软件：智能化图纸生成管家

背景与挑战

在数字化浪潮中，工业软件领域的计算机辅助设计（CAD）软件对产品设计的创新和生产流程的效率起着决定性作用。面对市场快速变化，企业亟需适应新的设计理念和客户需求，这对 CAD 软件提出了更高的挑战。传统的设计方法已难以满足高效率和高质量的双重要求，企业急需新技术的支持以提升竞争力。

解决方案

华泰软件结合英特尔® 至强® MAX 和酷睿™ Ultra 处理器的强大计算能力，开发了《智能化图纸生成管家》产品，该产品通过自然语言和表格数据与大语言模型进行交互，理解用户的设计需求，并通过 CAD 软件的指令接口完成图纸绘制。产品集成了 CAD 图纸数据学习、自然语言解析处理、图元高速绘制和图纸布局优化等核心技术，利用 DL Boost、AI Boost 和 OpenVINO™ 等技术框架，显著提升了 AI 运算效率和设计质量。

《智能化图纸生成管家》利用机器学习算法，从历史工程数据中学习提取特征，为设计师提供智能建议，辅助做出合理设计决策。自然语言解析技术使得设计师能够通过简单的语言描述直接生成 CAD 图纸，而自动布局功能则解决了图形元素摆放和连线的问题，使图纸更有组织、易于理解。此外，产品以 CAD 插件形式嵌入，实现了图形元素的实时输出，极大节约了设计时间，提升了工程效率，让设计师能够专注于产品创新。

方案优势

在实际应用中，《智能化图纸生成管家》凭借与英特尔® 新一代处理器的深度优化，展现出了显著提升企业竞争力与工作效率的能力。硬件效能方面，产品操作响应迅速，执行任务周期短，能够实现 7*24 小时不间断运行，并且电力消耗更低。知识传承上，它能够提炼和再利用海量设计数据，将优质工程经验转化为企业核心竞争力。此外，产品操作简便，即使是初中级设计师也能轻松应对复杂设计任务，同时针对工程设计行业的特定需求进行业务数据匹配，确保高度适应性。

安全性和可靠性方面，通过私有化部署，所有数据在企业内部加密存储和处理，严格权限控制有效消除信息安全风险。在制图效率上，产品大幅提升了绘图速度，原本需要 3 至 5 人天完成的工作量，现在能够在半小时内完成，极大节约了时间和人力成本。

Huatek

3.7 联想：基于 AI 的设备维护解决方案

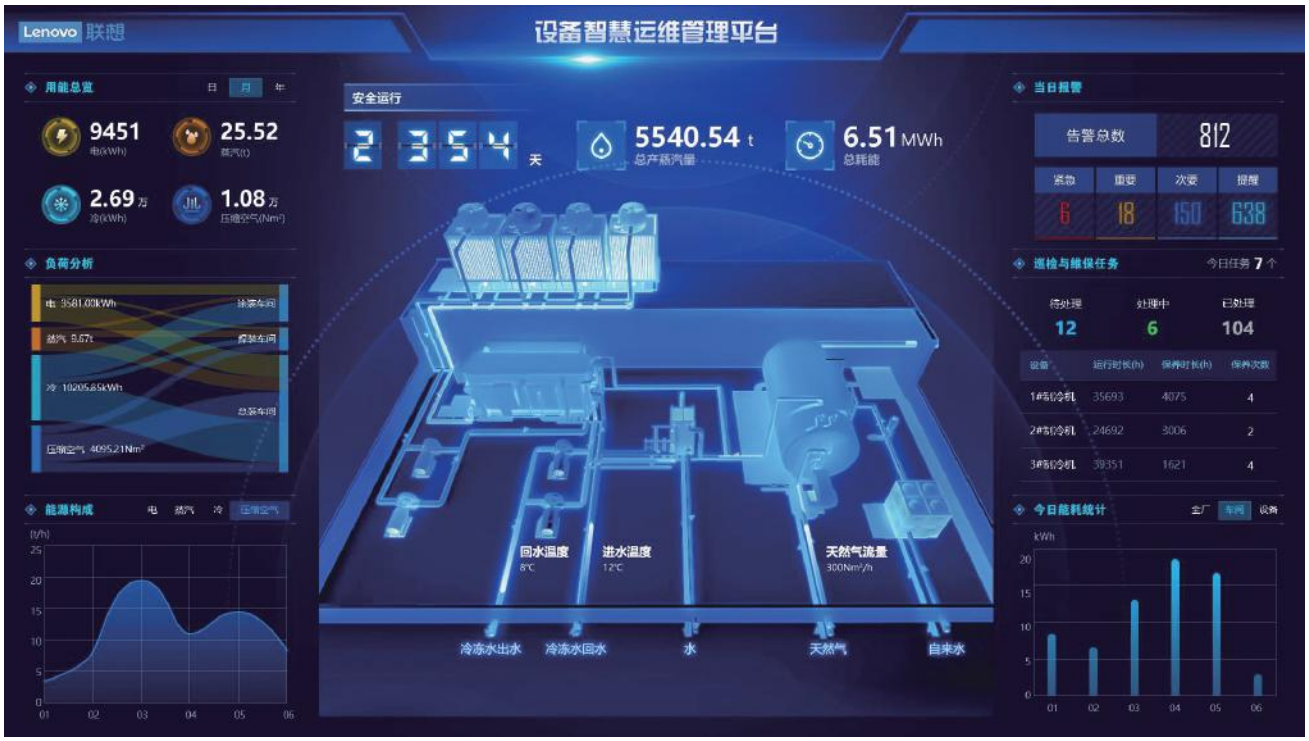
背景与挑战

在现代工业运营中，企业面临着多方面的设备维护管理挑战。首先，设备运行工况无法实时掌握，这导致无法及时响应设备的维护需求，增加了设备故障的风险。其次，缺乏数据支撑的设备服务和运营服务使得维护决策难以做到精准和高效，难以实现资源的最优配置。此外，随着技术的发展和设备的复杂性增加，运维成本不断攀升，给企业带来了经济压力。最后，传统的被动服务模式导致运维服务体验度差，无法满足客户对于高质量服务的期待。面对这些挑战，企业迫切需要一种能够提前预测设备潜在问题并进行预防性维护的智能解决方案，以提高设备的可靠性和生产效率，降低维护成本。

解决方案

联想基于 AIoT 的设备预防性维护解决方案，通过在边缘进行计算乃至 AI 算法，实现高效的本地数据处理和决策。该方案提供了本地化的边缘计算产品，不仅支持主流工业协议，还能接入海量多类型的设备，满足不同行业和场景的需求。以物联网的广泛连接为基础，该方案构建了一个统一的平台，实现了设备数据的集中管理和分析，从而提高了运维效率和服务质量。

此外，解决方案中的高效数据通道能够支撑算法加持的智能诊断，使得设备的运行状态可以实时监控，故障可以快速预测和诊断。这种智能化的维护方式不仅提高了设备的可靠性，还降低了运维成本，提升了服务体验。通过这些先进的技术和方法，联想的解决方案能够帮助企业实现设备维护的智能化和自动化，从而在激烈的市场竞争中保持领先。



方案优势

联想的 AIoT 设备预防性维护解决方案在硬件层面上采用了英特尔® 的先进 CPU，确保了高性能和可靠性。边缘工控机搭载了英特尔® Alder Lake –S 系列 CPU，边缘网关则使用了英特尔® Elkhart Lake。这些硬件的选择不仅提供了强大的计算能力，而且保证了系统的兼容性和易于集成到各种工业环境中。

解决方案的优势还体现在其能够打通现场生产设备，实现跨系统和设备的数据全面接入。这种无缝的数据整合为企业提供了一个面向多类角色、覆盖多种业务场景的统一平台，从而使得设备管理和维护更加高效和灵活。

此外，该解决方案基于丰富的知识库和 AI 算法来预测潜在问题，提前采取预防措施，从而减少了意外停机的风险并延长了设备的使用寿命。通过智能化的数据分析和故障预测，企业能够实现更加主动的维护策略，优化资源配置，提升整体运营效率。

04

合作伙伴加速项目和产品推荐

4.1 AI 硬件产品推荐

FUTUREROBOT
卓信创驰®

深圳市卓信创驰技术有限公司是一家专注于工业控制、机器视觉、自动化等领域的国家级高新技术企业。坚持以市场需求为导向，以创新技术为基础，以快速定制化解决方案为核心，聚焦于嵌入式计算设备的研发、生产和销售，致力于工业领域的自动化、数字化和智能化，为客户提供技术全面、稳定可靠、灵活便捷的硬件产品及系统解决方案。

E500-M 是其推出的搭载英特尔® 酷睿™ Ultra 处理器的工业计算机。使用英特尔® 酷睿™ Ultra 处理器 (Meteor Lake)，支持 MXM GPU 模块，旨在进行图像处理和 AI 推理。可扩展多种协议接口，是高可靠性工业设计。

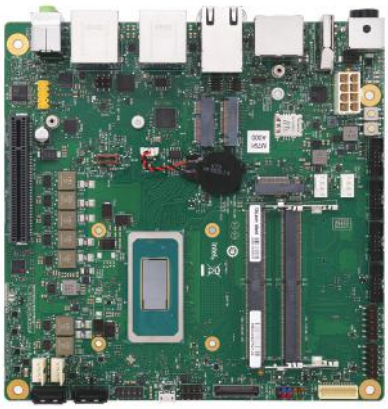
产品特性：

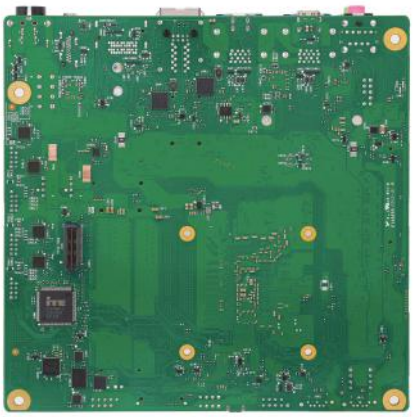
- 英特尔® 酷睿™ Ultra 处理器 (Meteor Lake)，高达 14C/20T
- Intel X^{e} LPG 图形显示，高达 128EU
- 支持 MXM GPU 模块，旨在进行图像处理和 AI 推理
- 独立三显 (2 x HDMI，1 x DP)
- 双通道 DDR5 5600MT/s，高达 96 GB SODIMM
- 丰富 I/O 扩展：3GbE LAN，5USB3.0，2COM，PCIex4
- 可扩展多种协议接口，支持 Wi-Fi，4G，5G 无线通讯模组
- 高可靠性工业设计

* 排序依照公司英文首字母排序

SEAVO
| 信 | 步 | 主 | 板 |
中国工控主板领导者

深圳市信步科技有限公司，专注工控主板研发和创新 32 年，并屡获殊荣：中国大陆第一家研发出 x86 主板，第一家把品质标准提高到最高万分之二，第一家承担 Intel 高性能新平台参考板设计。目前信步推出 1000 余款主板，应用覆盖 AI、机器视觉、机器人、边缘计算、智慧医疗、智能零售等重点领域，服务全球 100 多家上市公司，是中国大陆最大的工控主板和 ODM 产品提供商。





McLaren Island 是其推出的一款采用 Intel 最新一代 CPU，搭载 11T NPU 提供充沛算力，高速 DDR5 内存，可应用于运动控制、机器人等各种高性能，大数据处理的工业主板系列设计。

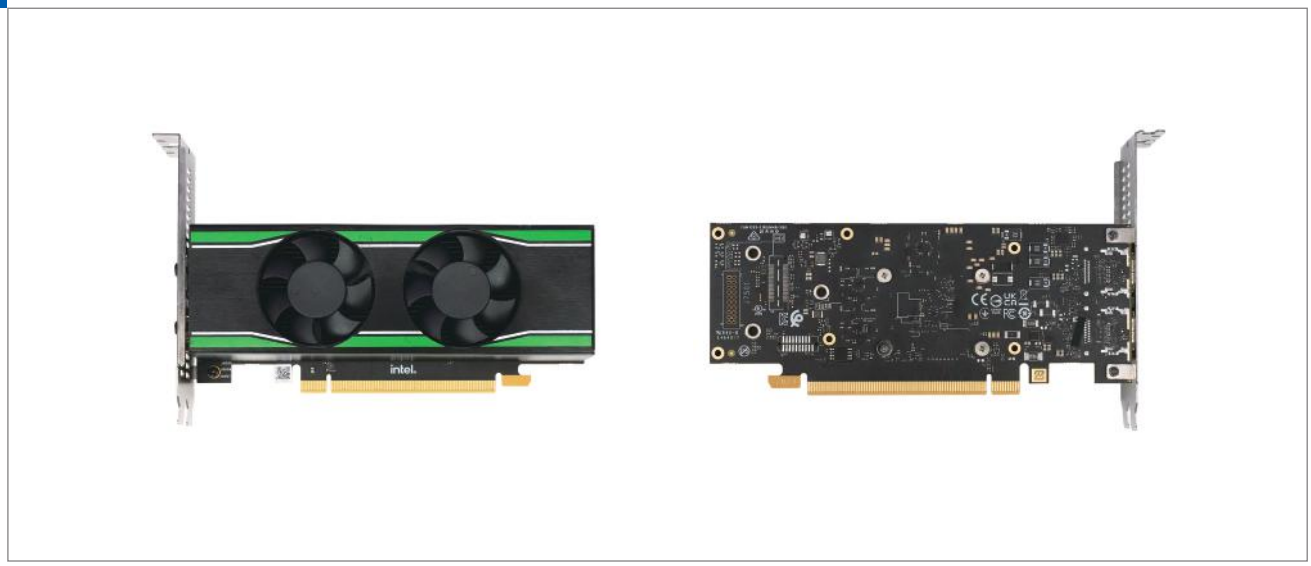
产品特性:

- 采用 Intel® Core™ Ultra 处理器 (Meteor Lake-U/H)
- 支持 DDR5 内存，1*SO-DIMM，最大 32 GB
- 提供 1* 千兆网卡，4*USB 3.2，6*USB 2.0
- 3*M.2，支持 NVMe SSD，WiFi+BT
- 2*HDMI，1*eDP/LVDS，1*VGA
- Mini-ITX，DC 12V 供电

* 排序依照公司英文首字母排序

ZRT 智锐通

深圳智锐通科技有限公司成立于 2014 年，是一家集研发、生产、销售于一体的国家级高新技术企业，是英特尔® 物联网解决方案联盟会员及国内多家 AI 芯片头部品牌的生态合作伙伴。产品以 AI 加速和工业显示技术为核心，重点致力于医疗影像 AI 分析、医疗信息化终端以及高端的机器视觉等提供解决方案。产品涵括 AI 加速卡、工业显示、AI 终端等硬件产品。



工业 AI 视觉检测设备是智能制造的关键工具，利用先进的人工智能算法与高清成像技术对生产线上的产品进行非接触式、高速的表面缺陷检测。精准识别产品表面的划痕、污渍等瑕疵，确保产品质量。

智锐通英特尔® Arc™ A380 6GD6 4H 显卡搭载了强大的英特尔® Arc™ A380 芯片和 6 GB GDDR6 显存，192bit 宽带宽提供了充足的数据处理速度，确保在复杂的工业视觉处理任务中的高效数据处理能力。四个 HDMI 2.0 接口支持 4K 高清输出，能够满足工业自动化、质量检测和机器视觉系统中对多显示器高分辨率输出的需求。

此外，紧凑的 SFF/2 Slot 设计使得这款显卡能够轻松适配各种工业环境中的控制柜和机器人控制器，而 PCI-E 16x 接口则保证了与工业计算机的高速稳定连接。在工业自动化领域，智锐通英特尔® Arc™ A380 6GD6 4H 显卡能够加速图形处理流程，提升生产线的视觉检测速度和精度，同时支持高级的图形界面和实时监控系统，不仅提高了工业系统的响应速度和运行效率，也为工业 4.0 和智能制造的实现提供了强有力的硬件支持。

* 排序依照公司英文首字母排序

4.2 PIPC 工业电脑优选项目介绍

4.2.1 英特尔® 工业电脑优选项目介绍

软件调优与工控稳定性验证，加速系统集成商与终端用户优选产品

英特尔® 工业电脑优选项目是英特尔针对中国工业用户的本地开发习惯，利用英特尔在工业级芯片、工业边缘节点参考架构、工业边缘软件平台的多年积累，**挑选适合不同应用的工业电脑产品进行测试及优选验证的项目**。该项目针对不同工业应用和需求，分别设立软件调优等级，具体为：

典型工业应用

- 工业控制
- 机器视觉
- 显控一体机 Panel PC NEW

软件调优等级

- **优选级 Elementary Select**
- **甄选级 Advanced Select**

工业级芯片及模块化设计*

项目入选机型配置均符合：

- 新款主流 X86 平台
- I/O 端口已适配不同工业应用需求

针对工业级工况提高可靠性要求

项目入选机型配置均提供：

- 基于标准（如 IEC 等）的可靠性测试报告
- 权威第三方机构（如 CE、FCC 等）出具系统安规证书

不同工业应用软件调优：

项目入选机型配置均完成：

- 该机型已针对不同工业应用进行差异化调优
- 调优结果均通过了英特尔® ECI 或 CVOI 内置工具验证

4.2.2 工业电脑优选项目测试函及加速计划

英特尔® 工业电脑优选项目测试函由英特尔签发（中英文双语测试函），提供给通过测试机型的生产厂家，载明的信息包括:

机型信息

- 生产厂家以及 Logo
- 型号以及机器外观图片
  - 硬件配置
  - 软件环境及工具

测试信息

- 适用工业应用
- 通过条件及等级
- 测试函有效期

项目加速 & 营销计划

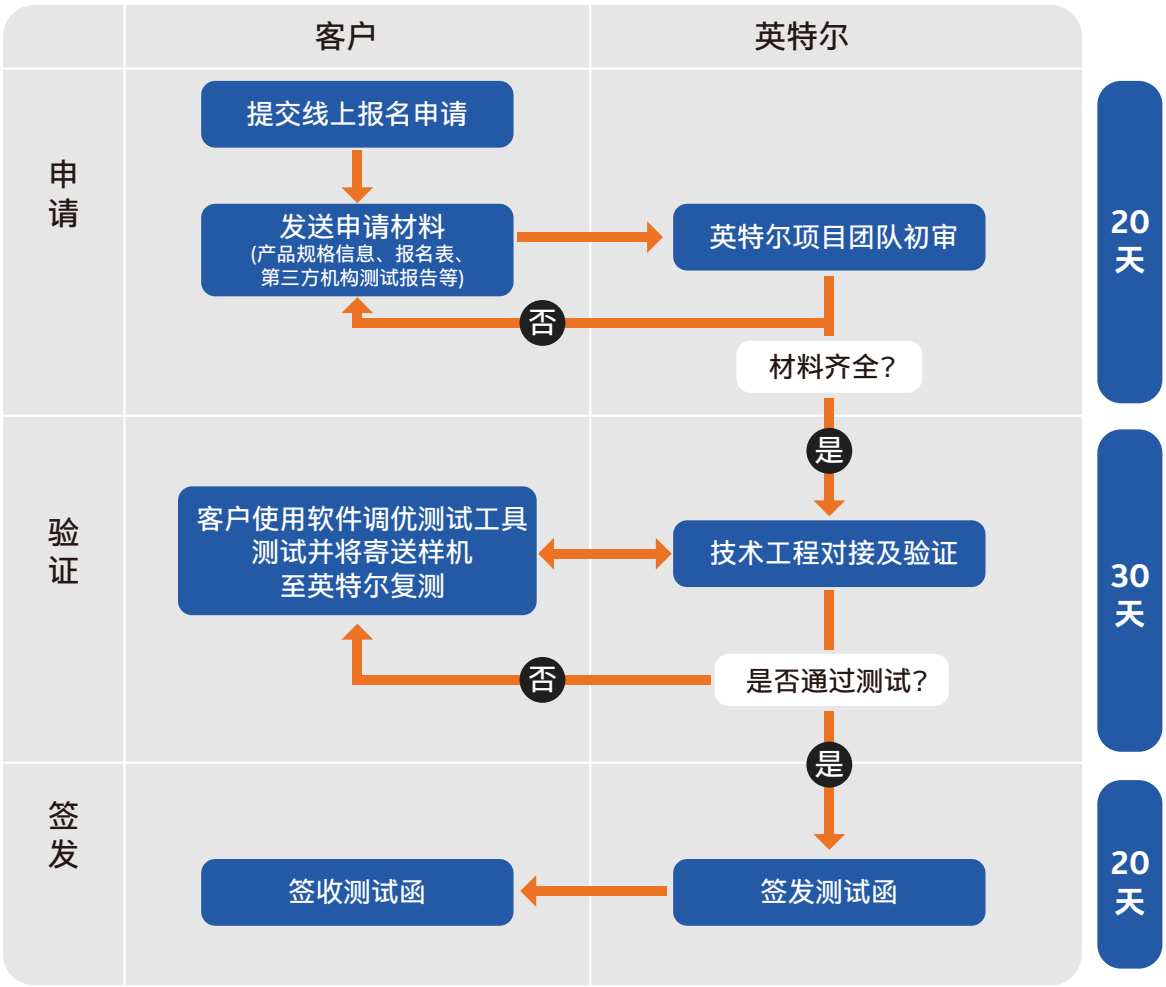
- 软硬件技术支持
- 市场联合营销机会
- 上下游业务对接机会

* 英特尔保留在发出通知或不发出通知的情况下对上述要求和条件进行临时的或永久的修改或停止提供的权利。具体信息以测试函所载内容为准。

4.2.3 工业电脑优选项目申请流程与节点安排

项目时间线

项目流程



* 每家客户每个应用 x 等级最多只能申请一台机器

** 每批次开放时间与名额有限，请您与英特尔客户经理联系锁定意向

*** 锁定报名意向后，英特尔将向您发送线上报名链接，届时正式启动项目参与

**** 流程中的天数为自然天天数

4.2.4 联系方式

如果您对本项目的内容感兴趣或想进一步了解项目，欢迎您与您公司所对应的英特尔客户经理联系。

若您对项目申请条件、流程有任何疑问、意见或建议，您可以联系下方项目联系人。

英特尔® 工业电脑优选项目 —
邱丽颖 Alice Chiu alice.l.chiu@intel.com

英特尔智能制造
扫码下载年度精选案例手册！
机器视觉/运动控制/机器人/新能源专区

4.3 PIPC 机器视觉产品推荐

阿普奇	阿普奇®
康士达	CSTIPC 康士达科技
卓信创驰	FUTUREROBOT 卓信创驰®
诺达佳	NODKA 诺达佳
信步科技	SEAVO 信步主板 中国工控主板领导者

*排序依照公司英文首字母排序

阿普奇

快速联系阿普奇

阿普奇成立于 2009 年，总部位于苏州，专注于工业 AI 边缘计算领域。公司提供多种 IPC 产品，包括传统工业电脑、一体机、显示器、主板和控制器。阿普奇同时开发了 IPC 小助手和 IPC 大管家等软件产品，领先推出 E-Smart IPC。这些创新广泛应用于视觉、机器人、运动控制和数字化领域，为工业边缘智能计算提供可靠解决方案。

阿普奇在苏州、成都和深圳设有三大研发基地，在华东、华南、华北和华西设有四个销售中心，并拥有 34 个以上服务渠道。公司在全国十多个地点设有子公司和办事处，增强研发和客户服务能力。阿普奇为 100 多个行业和 3000 多客户提供定制解决方案，累计出货量超过 60 万台。

工业电脑甄选

E7DS-Q670

阿普奇嵌入式工控机 E7DS 系列 Q670 平台是一款功能强大的嵌入式工控机，支持 Intel® 12/13th Gen Core / Pentium/ Celeron-S 处理器，提供 2 个 Intel 网络接口（1GbE & 2.5GbE），确保高速稳定的网络连接，具有 3 路显示输出，最高支持 4K@60Hz 分辨率，提供丰富的 USB、串口扩展接口和 PCIe、mini PCIe、M.2 扩展插槽，可根据具体应用需求进行自由选择，满足各种复杂的工业自动化需求，同时采用智能风扇主动散热设计，确保系统在高负载下的稳定运行。

特性：

- 采用 Intel® Q670/H610 芯片组
- 支持 Intel® Alder Lake-S/ Raptor Lake-S 系列处理器
- 2*DDR4 SO-DIMM 插槽，最大支持 64GB
- 1*M.2 Key-M（PCIe4.0*4，2280）硬盘接口
- 1*Intel® 2.5GbE + 1*Intel® 1GbE
- 8*USB 接口
- 支持 WiFi/4G 无线扩展
- 支持 DC18-36V 宽压电源输入
- 可选多槽位 PCIe/PCII 扩展

工业电脑优选

AK6215A2-2A1E

阿普奇弹匣式智能控制器 AK6 系列是专为机器视觉和边缘计算应用而设计的超紧凑型工业计算机，搭载英特尔 ADL-U/RPL-P 平台处理器，板载两口千兆网卡，6 个 USB 高速接口，支持大容量高速 DDR5 内存，可支持 PCIE 扩展，确保流畅的多任务处理，提供高效计算能力，同时可自由增加、更换高速扩展的主弹匣或多 I/O 扩展的辅弹匣，既满足通用性需求，也能适应不同的行业性需求。

特性：

- 采用 Intel® Alder Lake-U/ Raptor Lake-P 系列处理器
- 1*DDR5 SO-DIMM 插槽，最大支持 32GB
- 1*M.2 Key-M（PCIe4.0*4，2280）硬盘接口
- 2*Intel® 千兆网口
- 支持 WiFi/4G 无线扩展
- 支持 DC12-28V 宽压电源输入
- 可以根据具体情况增加、更换高速扩展的主弹匣的或是多 I/O 扩展的辅弹匣

康士达

CSTIPC
康士达科技

快速联系康士达科技

深圳市康士达科技有限公司成立于 2009 年，一直致力于工控电脑板卡、整机、工业平板、智能系统的设计开发。是一家集研发、设计、生产、销售、定制化服务为一体的国家高新技术企业、专精特新企业。作为专业的智能系统开发商，公司全力为各行业客户提供个性化软硬件服务，包括 CPU 控制器、视觉处理卡、AI 加速卡、底层驱动技术、中间开发包、AI SDK 及 APP 开发指导等 OEM&ODM 服务。康士达始终以 “让设备更智能” 的发展使命服务行业客户，立足于自主创新、自主研发，所有的产品拥有完全的自主知识产权。产品广泛应用于工业控制、机器视觉、机器人、边缘计算、自动驾驶等行业。

工业电脑甄选

R6AMV-BD1

本产品基于 Intel 酷睿 12 代处理器平台 Alder Lake P 而设计，采用 Core i3-1215U/i5-1235U/i5-1335U/i7-1260P 处理器，支持 Win、Linux 操作系统。整机尺寸为 200（长）× 150（宽）× 72.2（高）mm。产品散热件采用散热性能优良的铝型材为主体，并做表面喷砂铁灰色阳极氧化处理，壳体采用厚度 T=1mm 的钣金材料打造，表面处以铁灰色烤漆；本产品结构简洁，外形美观，采用 Alder Lake P 系列高能效处理器平台，具备丰富的 IO 扩展，是一款为机器视觉、工业网关等应用而设计的工控电脑产品。

特性：

- 32GB DDR4-3200MT/s SO-DIMM
- 128GB SSD SATA3.0 (M.2 2280 Key B+M)
- 2 个 SATA3.0 2.5 寸硬盘扩展位
- 6 个 RJ45 千兆网口 — 其中 LAN3~6 支持 POE 802.3AF 标准
- 1*DP，1*HDMI2.0b，6 个 USB3.0 接口
- 1 个复合 IO 接口（8DI&8DO、4 路光源输出、4 路外触发输入）
- 1 个 M.2 3042/52 B-Key（支持 4G/5G 模块扩展），1 个 M.2 2230 E-Key（支持 WIFI+ 蓝牙扩展），2 个 PCIE X16 扩展槽

工业电脑优选

U12MV-BA1

本产品基于 Intel 酷睿 12 代处理器平台 Alder Lake P 而设计，采用 Core i3-1215U/i5-1235U/i5-1335U/i7-1260P 处理器，支持 Win、Linux 操作系统。整机尺寸为 200（长）× 150（宽）× 72.2（高）mm。产品散热件采用散热性能优良的铝型材为主体，并做表面喷砂铁灰色阳极氧化处理，壳体采用厚度 T=1mm 的钣金材料打造，表面处以铁灰色烤漆；本产品结构简洁，外形美观，采用 Alder Lake P 系列高能效处理器平台，具备丰富的 IO 扩展，是一款为机器视觉、工业网关等应用而设计的工控电脑产品。

特性：

- 支持 2*DDR5-4800MT/s 笔记本内存，Max 64GB
- 1* 标准的 SATA3.0 接口，1*M.2 2280 M-Key 插槽，可 BOM 选 NVMe，1 个 MINI-PCIE 插槽，支持 4G/WIFI
- 1 个 HDMI 2.0b，1 个 VGA
- 4 个 USB3.0 接口，5 个 RJ45 千兆网口
- 2 个 COM 接口，16*DI & 16*DO
- 9-36V 直流输入，运行温度：-20℃~+60℃

卓信创驰

FUTUREROBOT
卓信创驰®

快速联系卓信创驰

深圳市卓信创驰技术有限公司是一家专注于工业控制、机器视觉、自动化等领域的国家级高新技术企业。坚持以市场需求为导向，以创新技术为基础，以快速定制化解决方案为核心，聚焦于嵌入式计算设备的研发、生产和销售，致力于工业领域的自动化、数字化和智能化，为客户提供技术全面、稳定可靠、灵活便捷的硬件产品及系统解决方案。

工业电脑优选

E223

E223 系列通过 Intel 工业电脑优选项目机器视觉优选级测试，是一款针对机器视觉打造的工业计算机，搭载 Intel Atom® 平台 Alder Lake-N 系列处理器，支持 DI/DO、Encode、2LED、Trigger 和 RS-232/485 等 IO 接口，采用无风扇紧凑式结构设计，小尺寸，低功耗，性能均衡，满足负载整合的市场需求。

特性:

- Intel Alder Lake-N 系列处理器
- 单通道 DDR5-4800，最高 16 GB 内存
- DP 和 HDMI 独立双显
- 3 LAN（2 x Intel i210 支持 PoE，1 x Realtek RTL8111H）
- 1 RS-232/422/485，1 RS-232
- 3 USB 3.0，1 USB 2.0
- 2 触发，2 光源
- 8 DI，8 DO，2 编码器
- 支持 4G/5G，Wi-Fi 6 无线通讯模块扩展
- DC 9-36V，可选支持 UPS 模块扩展
- 172 x 125 x 62.5mm (LxWx H)，支持壁挂/背挂/导轨安装
- 模块化、无线缆高可靠设计
- CE/FCC Class B，ESD 8KV/15KV
- Windows 10/11，centos 8，ubuntu 20.04/22.04

诺达佳

快速联系诺达佳

诺达佳（NODKA）创立于 2001 年，致力于工业 PC 和 HMI 系统平台的研发创新，为自动化、测量、通讯等领域的客户提供全面的产品解决方案，产品线涵盖 X86/ARM 主板及核心模块、嵌入式计算机、工业平板电脑、工业显示器、工业操作面板、Automation PC、EtherCAT 从站 IO、网络安全硬件平台等，配套有高低温、湿度、振动、跌落、EMC、ESD、EFT、雷击浪涌等完善的 DQA 测试实验室，保障工业产品设计的可靠性；建立系统整机组装、SMT/DIP、精密钣金、CNC 加工、烤漆和阳极氧化表面处理等全流程的智能制造中心，工厂制程与管控皆符合 ISO9001、ISO14001 认证规范，满足为客户提供高质量的少量多样化和大批量生产制造，兼具 OEM/ODM 客制化服务能力。产品广泛应用于工业自动化、交通、电力、石化、钢铁、新能源、环保、医疗及商业自助终端等行业。

工业电脑优选

NP-6113-161160

NP-6113 及其系列产品为诺达佳 Automation PC 系列中的一款低功耗无风扇书本式工控机，秉承功能完备、高性价比的理念，紧凑美观的外形和无线缆设计，机壳采用高精铝合金型材，采用大面积铝鳍作为 CPU 散热器，结构紧凑、外形小巧兼具强固性，无风扇设计，全封闭的结构防止粉尘进入，保障产品的可靠性和使用寿命。

特性：

- 2 x Intel® 千兆 PoE 网口
- 1 x RTL8111H 千兆网口
- 2 x USB3.0 接口，2 x USB2.0 接口，板载内置 1 个 USB 口可安装硬件加密狗
- 1 x RS232/RS485，RS485，支持自动流控；1 x RS232
- 支持 VGA 和 HDMI 双显示接口
- 1 x miniPCIe 扩展槽，可扩展 Wifi，3G/4G 模块
- 16 x 隔离 DI，16 x 隔离 DO
- 支持 mSATA 固态盘存储

信步科技

SEAVO
| 信 | 步 | 主 | 板 |
中国工控主板领导者

快速联系信步科技

信步科技，自 1992 年起，专注包括 Intel x86 在内的多个平台的工控主板研发和创新，并屡获殊荣：中国大陆第一家研发出 x86 主板，第一家把品质标准提高到最高万分之二，第一家承担 Intel 高性能新平台参考板设计……时至今日，信步科技拥有全球范围内综合素质领先的研发团队，推出 900 多款产品，服务全球十多个行业的 100 多家上市公司，年销量遥遥领先同行，是中国大陆最大的工控主板和 ODM 产品提供商，成为人类 “把机器变成人” 这一时代进程的关键力量！

工业电脑甄选

NAV-700

基于工业计算机的高可靠性需求与挑战，信步科技推出一款全新设计的工业计算机。整体采用无风扇散热设计，提供 4 个扩展插槽，通过多项高可靠设计测试，性能强劲，算力充沛，可应用于机器视觉、运动控制、机器人、边缘计算等领域。

特性:

- 采用第 12 代 Intel® Core™ 处理器，支持高性能显卡
- 2 * DDR4，4 * SATA，10 * GbE，10 * USB，2 * HDMI
- 1 * PCIe x16，1 * PCIe x4，2 * PCI
- 全机身一体式无风扇散热设计
- 20 多项抗干扰可靠性设计，MTBF 5 万小时
- 9-36V DC 供电
- 232 * 274 * 212mm

英特尔致力于尊重人权，坚决不参与谋划践踏人权的行为。参见英特尔的《全球人权原则》。英特尔的产品和软件仅限用于不会导致或有助于违反国际公认人权的应用。

实际性能受使用情况、配置和其他因素的差异影响。更多信息请见 www.Intel.com/PerformanceIndex

性能测试结果基于配置信息中显示的日期进行测试，且可能并未反映所有公开可用的安全更新。详情请参阅配置信息披露。没有任何产品或组件是绝对安全的。

具体成本和结果可能不同。

英特尔技术可能需要启用硬件、软件或激活服务。

英特尔未做出任何明示和默示的保证，包括但不限于，关于适销性、适合特定目的及不侵权的默示保证，以及在履约过程、交易过程或贸易惯例中引起的任何保证。

英特尔并不控制或审计第三方数据。请您审查该内容，咨询其他来源，并确认提及数据是否准确。