解决方案简介

基于英特尔® CRI-RM CPU 亲和性调度技术 浪潮 AIStation 加速释放算力性能

概述

施行人工智能（AI）战略，在企业部署 AI 训练与推理应用已经成为众多企业推进数字化转型、激发业务创新活力的重要选择。为了降低资源配置的复杂性，实现计算、内存、存储等关键资源的敏捷扩展与统一管理，越来越多的用户开始使用 Kubernetes（K8s）来搭建 AI 训练的基础平台，以支持技术人员自由地在不同平台中调用不同的资源进行 AI 训练。

作为专为企业级 AI 生产环境打造的人工智能开发平台，浪潮 AIStation V3 使用了基于 K8s 容器引擎优化资源的管理与调度策略，为 AI 研发提供了强大算力支撑。为了在利用英特尔® 至强® 可扩展处理器进行 AI 训练加速的情况下，突破使用 K8s 原生 CPU 管理机制所带来的性能瓶颈，浪潮利用英特尔® CRI-RM（基于容器运行时接口的资源管理器）组件进行了 AI 训练加速实践，可以在 K8s 集群上，按照拓扑资源实现物理主机的最优分配，从而能够大幅提升 AI 计算的性能。

挑战：K8s 原生 CPU 管理机制带来 AI 性能瓶颈

在互联网、自动驾驶、智能制造、智慧金融等场景，AI 已经成为企业产品与服务的核心竞争力来源，越来越多的企业选择构建 AI 平台，以推进 AI 技术创新与实践落地。但是，AI 模型从开发进入到生产部署阶段面临着多重困难和挑战。一般而言，AI 模型需要经过大量的调试和测试才能部署上线，这一过程通常需要 2-3 天；AI 线上服务计算资源通常较固定，对于突发需求资源响应慢，业务扩展难。此外，由于 AI 模型来源不同，企业难以实现对于多种 AI 模型以及资源请求的统一管理。

K8s 为解决上述问题提供了一个重要途径。作为云原生的核心技术之一，K8s 能够管理云平台中多个主机上的容器化应用，实现了 AI 资源的统一部署、规划、更新、维护，能够有效提高用户的 AI 资源管理率，并提升平台的可管理性、可扩展性、弹性与可用性。

在基于 K8s 技术的 AI 开发平台建设实践中，CPU 作为通用算力提供者，有着无可替代的地位。CPU 无论在采购成本，抑或使用难度等方面有着重要优势，同时 CPU 因为其算力的通用性，不光支持 AI 运算，亦可用于其他应用负载。使用 CPU 服务器可有效利用空置资源以及空闲时间，通过 K8s 的弹性资源调度分配给其它应用。

虽然使用 CPU 进行 AI 训练是一个颇具可行性的选择，但是在 K8s 集群上，用户会在性能方面遇到一定的瓶颈。这是因为 K8s 原生的 CPU 管理机制没有考虑 CPU 绑定与

NUMA 亲和性，高版本的 K8s 只会对 QOS 为 Guaranteed 的 Pod 生效，这可能会导致 CPU 在 AI 训练中的性能无法充分发挥。这一原生 CPU 管理机制存在的问题包括：

- 对于无法升级 K8s 版本的某些场景，不能使用 CPU 绑定和 NUMA 亲和性。
- CPU Manager 和 Topology Manager 代码集成在 Kubelet 组件中，不易进行扩展定制化开发。
- 在 AI 场景下，大部分用户都不希望在 Pod 中进行内存限制。而高版本的 K8s 只会对 QOS 为 Guaranteed 的 Pod 生效，这会导致无法使用 Kubernetes 默认的特性。

解决方案：使用英特尔 CRI-RM 组件，浪潮 AIStation 对 CPU 核进行智能化调度

浪潮 AIStation 面向人工智能企业训练开发与服务部署场景，提供完整的模型开发、训练、部署全流程，可视化开发、集中化管理等特性，为用户提供极致高性能的 AI 计算资源，实现高效的计算力支撑、精准的资源管理和调度、敏捷的数据整合及加速、流程化的 AI 场景及业务整合，有效打通开发环境、计算资源与数据资源，提升开发效率。

AIStation V3 实现了对于 K8s 容器引擎的支持，可以更便捷地实现 AI 容器化部署并提供智能化任务调度，提高了集群资源利用率和深度学习训练性能。具体而言，AIStation V3 资源调度更亲和，可智能化实现最优节点分配策略，充分利用闲暇时间训练任务，最大化发挥计算资源的性能；训练数据分层缓存预读机制可大幅提高训练速度。此外，AIStation V3 还可支持多种模式的单机和多机训练任务调度，并实现 batch 类型的训练作业的提交和稳定训练。

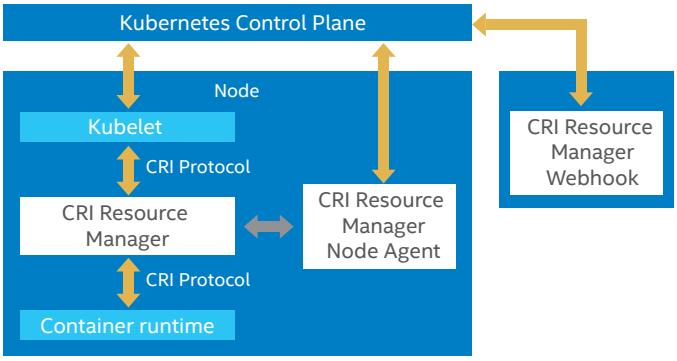

图 2. CRI-RM 组件架构图

为了解决使用 K8s 原生 CPU 管理机制所带来的性能瓶颈，浪潮在 AIStation V3 中应用了英特尔 CRI-RM 组件，该组件可以插在 Kubelet 和 CR 之间，截取来自 Kubelet CRI 协议的请求，扮演 CR 的非透明代理，跟踪所有集群节点容器状态，能够更好地将处理器、内存、IO 外设、内存控制器等资源分配给应用负载。在 K8s 环境中，英特尔 CRI-RM 组件能够用于拓扑感知资源对齐、容器防吵闹、延迟关键工作负载、CPU 时钟速度限制等场景。

图 1. 浪潮 AIStation V3 架构图

K8s 集群可以配置为 CRI-RM 节点和非 CRI-RM 节点，在 CRI-RM 节点上的必要组件可以通过 CRI-RM 部署过程进行安装和配置。当应用程序请求某些资源时，K8s 调度器根据设定的调度策略将任务调度到满足资源需求的节点上。

如图 3 所示，Worke1 节点上的 cri-remgr 应用程序将根据预配置的策略（例如拓扑感知）和 pod 注释处理请求，为 pod 分配最佳资源。如果配置发生变化，例如从选择隔离核心到共享核心，cri-resmgr-agent 会获取配置并通知 cri-resource-manager。CRI Resouce Manager 会存储配置信息以创建应用程序容器，并调用运行时来调整容器的原始资源配置。

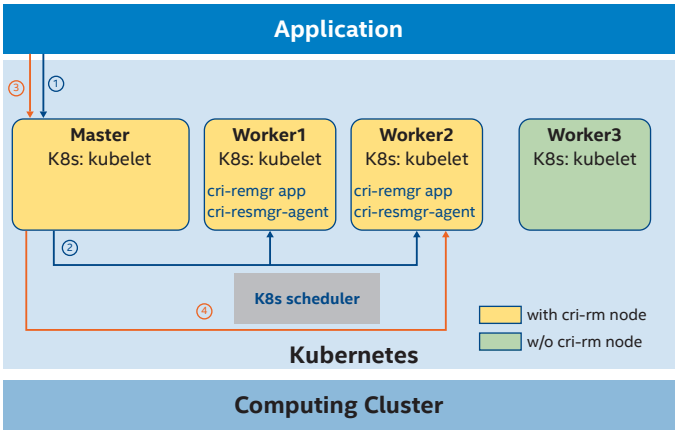

图 3. CRI-RM 组件调用流程

浪潮在 AIStation V3 的 K8s 集群中集成了 cri-resource-manager 组件，可以在 K8s 集群中，按照拓扑资源，实现物理主机的最优分配，从而大幅提升 AI 计算的性能，其主要优势包括：

- 方案与 K8s 源码解耦，可以按需进行自定义开发，在不升级 K8s 版本的情况下，通过简单配置就能实现 CPU 绑定、NUMA 亲和性。
- 充分考虑到 AI 开发者的使用习惯，在 Pod 不进行内存限制时，也可以进行 CPU 绑定和 NUMA 亲和性限制。
- 丰富物理主机的资源拓扑视图，CRI-RM 组件会收集有更多的拓扑信息，用于资源的调度。
- 基于 CRI-RM 组件，开发者可以获得更加丰富的物理主机拓扑信息，基于 CRI-RM 与 K8s 的松耦合性，开发者可以自定义节点内的资源分配策略。

对于行业用户来说，通过在 Kubernetes 集群中的 AI 训练场景引入 CRI-RM 组件，他们能够提供一种在 Kubernetes 集群内 AI 计算加速的实践方案。

验证：实现显著性能提升

本方案可以部署在基于高性能计算（HPC）集群的 AIStation V3 中，充分利用 HPC 集群强大的 CPU 算力，并通过 AIStation V3 高效的资源统一管理与灵活配置能力，为多种 AI 训练任务提供支持。

为了证实 AIStation V3 在集成英特尔 CRI-RM 组件之后，运行 AI 训练负载的性能表现，浪潮选择 Tensorflow | model: resnet50 以及 Tensorflow | model: customized cnn 两个 AI 训练用例进行了测试。测试环境如表 1 所示：

CPU	英特尔至强金牌 6132 处理器 @ 2.60GHz，28 核，56 线程
内存	192G
OS	Centos 7.8.2003
Kubernetes	1.14.8
Docker	19.03
AIStation	3.1

表 1： 测试环境

在 Tensorflow | model: resnet50 测试用例中，测试人员发现，AIStation V3 创建容器的资源规格为 14 个 CPU 内核，容器创建成功后，查询实际分配的 CPU 内核为：0-27，说明未做

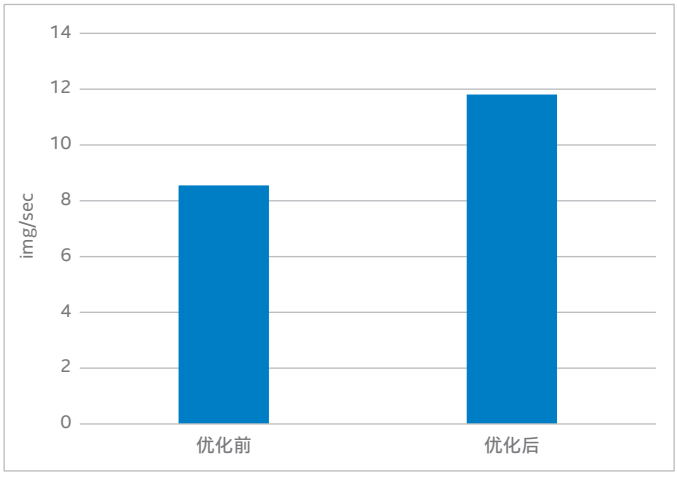

图 4. 测试用例 1 — 使用 CRI-RM 进行优化前后的性能对比（越高越好）

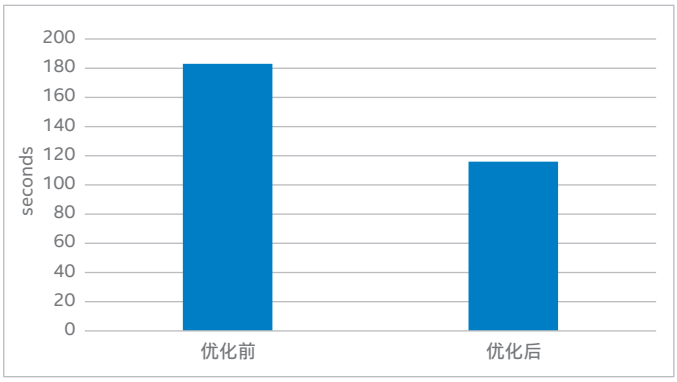

图 5. 测试用例 2 — 使用 CRI-RM 进行优化前后的性能对比（越低越好）

CPU pinning，但 k8s 只允许容器使用 14 个 CPU 内核的资源，此时的性能为 8.55 img/sec。在使用 CRI-RM 并成功创建容器之后，查询实际分配的 CPU 内核为：cpuset.cpus: 14-27，即 CRI-RM 全部使用 NUMA1 内的 CPU 内核，此时的性能为 11.81 img/sec。[1]

在 Tensorflow | model: customized cnn 用例的测试中，应用 CRI-RM 之后，使用了 AIStation 创建容器的资源规格为 14 个 CPU core，容器创建成功后，查询实际分配的 CPU core 为：cpuset.cpus: 14-27，即 CRI-RM 全部使用 NUMA1 内的 CPU core。此时，实例的运行时间 116 seconds，相比使用之前的 183 seconds，性能提升 57.76%。[2]

这意味着，在未对硬件配置进行更新的前提下，CRI-RM 的应用会带来大幅度的性能提升，使得用户无需在硬件进行投入就能够获得可观的 AI 训练性能提升，从而提高了基础设施的利用效率，并节约了总体拥有成本（TCO）。

展望：加速基于 CPU 的 AI 训练任务赋能智慧变革

本次测试不仅验证了英特尔 CRI-RM 技术能够给基于 K8s 的 AI 训练任务带来的提升，还证明了 CPU 在 AI 工作负载中的巨大潜力。双方还在探索在基于第三代英特尔至强可扩展处理器上的 HPC 集群上进行进一步的性能验证。

与上代处理器相比，第三代英特尔至强可扩展处理器提供了 8 个插槽配置的多插槽内核计数密度，每个处理器最多可达 40 个核心，在性能、吞吐量和 CPU 频率都实现了显著提高，这为其处理 AI 负载提供了关键的性能基础。值得一提的是，第三代英特尔至强可扩展处理器搭载了英特尔® 深度学习加速技术，还在业界首次实现 16 位脑浮点（blfoat16）的 x86 支持，带来增强的人工智能推理和训练性能。

未来，浪潮与英特尔计划在利用 CPU 进行人工智能推理和训练方面进行更为广泛的合作，通过硬件选型、软件优化、系统集成等多种不同的方式，加速从云端到边缘基础设施上的人工智能性能表现，赋能金融、互联网、制造、电信、公共事业、政府、零售等行业的智慧化转型。

intel®

[1,2] 数据援引自浪潮内部测试结果；测试配置：英特尔至强金牌 6132 处理器 @ 2.60GHz，28 核，56 线程，192 GB 内存，Centos 7.8.2003，Kubernetes 1.14.8，Docker 19.03，AIStation 3.1。

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 www.Intel.com/PerformanceIndex

具体成本和结果可能不同。

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。

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。

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。

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。